

Hypothesis Testing with E-values

Aaditya Ramdas¹ and Ruodu Wang²
`aramdas@cmu.edu`, `wang@uwaterloo.ca`

September 10, 2025

¹Department of Statistics and Data Science, and Machine Learning Department, Carnegie Mellon University

²Department of Statistics and Actuarial Science, University of Waterloo

Contents

Preface	vi
Lists of notation, conventions, and examples	viii
I Fundamental Concepts	1
1 Introduction	2
1.1 The three roles of e-values	3
1.2 P-values versus e-values	4
1.3 Mathematical notation and conventions	7
1.4 Definitions of e-variables, p-variables, and tests	8
1.5 Betting interpretation of e-values	9
1.6 A first example: likelihood ratio	10
1.7 A second example: the soft-rank e-value	13
1.8 Historical context	14
1.9 Road map	18
2 Validity: E-values under the null hypothesis	20
2.1 Testing using Markov's inequality	20
2.2 Testing using distributional knowledge	21
2.3 Relating p-values and e-values using calibrators	22
2.4 Randomized tests and calibrators	25
2.5 Admissible e-variables	26
2.6 More examples of e-values	29
2.7 Informal grids for realized e-values	34
3 Efficiency: E-values under the alternative hypothesis	36
3.1 Powered tests, e-values, and p-values	36
3.2 A unified existence result	38
3.3 E-power and powered e-values	40
3.4 Conditions for the existence of powered e-variables	43
3.5 Log-optimality and likelihood ratios	44
3.6 Conditional log-optimality and log-optimal calibrators	46
3.7 Asymptotic log-optimality: methods of mixtures and plug-in	47

3.8	Axiomatic justification of the e-power	48
4	Post-hoc testing and decision making with e-values	52
4.1	Testing at data-dependent error levels	52
4.2	Post-hoc decisions using p-values and e-values	55
4.3	Optional continuation of experiments	59
4.4	Incentive-compatible principal-agent hypothesis testing	61
II	Core Ideas	64
5	The universal inference e-value	65
5.1	Motivation: Irregular models	65
5.2	Split likelihood ratio e-variable	66
5.3	Subsampled likelihood ratio e-variable	67
5.4	Exchangeable Markov's inequality	68
5.5	Universal confidence sets	69
5.6	Extensions	70
5.7	Numerical studies	73
6	The log-optimal e-value and reverse information projection	77
6.1	The numeraire e-variable	77
6.2	Strong duality for finite $\mathcal{P}$	80
6.3	Reverse information projection	81
6.4	Numeraire is the only e-variable that is a likelihood ratio	84
6.5	Numeraire dominates universal inference	85
6.6	Connections between e-values and Bayes factors	86
6.7	Examples	88
7	Sequential anytime-valid inference using e-processes	93
7.1	The problematic practice of peeking at p-values	93
7.2	Wald's sequential probability ratio test	94
7.3	Test supermartingales, sequential tests, and e-processes	96
7.4	Optional stopping and Ville's inequality	97
7.5	E-processes constructed from likelihood ratios	100
7.6	Testing by betting	104
7.7	The universal inference e-process	105
7.8	Sequential e-values and empirically adaptive e-processes	106
7.9	Obtaining an e-process from e-values using time-mixture	108
7.10	E-processes avoid reasoning about hypothetical worlds	109
7.11	A pathological example for the e-power of e-processes	110
8	Handling multiple e-values	113
8.1	Merging e-values under arbitrary dependence	113
8.2	Independent e-values and their products	115
8.3	Sequential e-values and martingale merging functions	117

8.4	Mean-variance trade-off	120
8.5	Family-wise error rate and merging e-values	122
9	False discovery rate control using compound e-values	125
9.1	Compound e-values and the e-BH procedure	125
9.2	Universality of compound e-values and e-BH	134
9.3	Log-optimal simple separable compound e-values	135
9.4	A minimally adaptive e-BH procedure	138
9.5	Stochastic rounding of e-values and randomized e-BH	140
9.6	The closed e-BH procedure	143
III	Advanced Topics	152
10	Approximate and asymptotic e-values	153
10.1	Approximate p-values and e-values	153
10.2	Asymptotic p-values and e-values	156
10.3	An asymptotic e-value via the central limit theorem	160
10.4	Approximate/asymptotic compound p-values and e-values	162
11	Combining e-values and using e-values as weights	172
11.1	Optimal transport duality	172
11.2	Admissible e-merging functions	173
11.3	Cross-merging: merging an e-value and a p-value	176
11.4	E-values as weights in p-value-based multiple testing	178
12	Using e-values to combine dependent p-values	181
12.1	The arithmetic mean of p-values	181
12.2	Merging p-values under arbitrary dependence	187
12.3	Combining exchangeable p-values	194
12.4	Randomized combinations of p-values	195
13	False coverage rate control using e-confidence intervals	198
13.1	Defining and constructing e-CIs	198
13.2	The e-BY procedure for false coverage rate	201
13.3	Combining CIs via majority vote	202
14	Further contrasts between e-values and p-values	214
14.1	A duality between e-values and p-values	214
14.2	Representations as conditional expectations and probabilities	216
14.3	Existence of powered exact e-values and p-values	217
15	Improved e-value thresholds under additional conditions	219
15.1	Conditional e-to-p calibrators on a subset of e-values	219
15.2	Comonotonic e-variables	221
15.3	Density conditions	223

15.4	Conditions on log-transformed e-variables	225
16	E-values and risk measures	228
16.1	Risk measures and statistical functions	228
16.2	Connecting e-values to coherent risk measures	230
16.3	E-values for testing risk measures	231
Appendix		237
A.1	Atomless probability spaces	237
A.2	Quantile functions	239

To Leila and Fangda
for their unwavering support and patience
and to Zoya and Xuelu
for the joy and wonder they brought into our lives

The book is published as the inaugural issue of
Foundations and Trends[®] in Statistics, Vol. 1: No. 1-2, pp 1-390 (2025)
<http://dx.doi.org/10.1561/36000000002>

Preface

An e-value is a nonnegative test statistic whose expected value is at most one under the null hypothesis. This book is written to offer a humble, but unified, treatment of e-values for education and research in statistics and its applications. In our opinion, the need for such a book at this time can be explained by at least four reasons: (a) e-values have been named, utilized, and studied as a stand-alone concept only in the last few years, and a large body of its potential users do not know what they are; (b) e-values are fundamental objects at the core of hypothesis testing and estimation, and they are both under-studied and under-utilized; (c) several application domains in the natural and social sciences would benefit from knowing and adopting methodologies based on e-values in certain contexts to improve statistical efficiency and scientific reproducibility; (d) there has been an explosion of exciting research over the past few years, and we think the time is ripe to collate resources in a self-contained and concise manner. We expand on these reasons below.

(a) As we discuss in more detail in the introduction, e-values (and their sequential extensions: nonnegative supermartingales and e-processes) have been around, either explicitly or implicitly, in the statistics and probability literature for the better part of a century. However, they have not been studied under a unified umbrella or terminology until recent years — indeed, even the terms e-value and e-process are only roughly five years old. As a result, the e-value is not a ‘household concept’ and statisticians often do not know much beyond its definition (if that).

(b) E-values are a fundamental concept/tool in hypothesis testing. We make five points to justify this, which will be clear from the content of the book. First, we will show that a *powered* test exists if and only if a *powered* e-value exists. Second, when moving beyond binary accept/reject decisions as encountered in testing, if we have more than two actions, or the losses are data-dependent, then all decisions that guarantee risk control must be based on thresholding e-values. Third, for many simple-to-state and fundamental composite testing problems (for example, testing if data come from a mixture of Gaussians), the only test we are aware of proceeds by constructing an e-value; this technique is called *universal inference*. Further, there is a unique and well-defined notion of a *log-optimal* e-value — called the *numeraire* — even when testing any arbitrary composite null hypothesis, with no restriction. Fourth, when it comes to anytime-valid inference (in a sequential testing context) and combining dependent test statistics (in a multiple testing context), methods based on e-values and e-processes are the *only admissible ones* to maintain statistical validity. Fifth, treating e-values as abstract objects leads to a deeper understanding and improvement of procedures that do not concern e-values by themselves, such as false discovery rate controlling procedures and methods to combine dependent p-values.

(c) It is well known the sequential procedures can yield improved statistical efficiency over ones that need to fix the sample size in advance of seeing the data. However, the recent crisis of scientific reproducibility is largely related to the use and misuse of p-values in such contexts (especially peeking at p-values and uncorrected optional stopping and continuation of experiments). This calls for methodologies that are statistically justifiable under various new and complicated environments. E-values are one tool (though certainly not the only one) to address this challenge, because they benefit from their simple definition, natural connections to game-theoretic probability and statistics, flexibility and robustness in multiple testing under dependence, and their central role in anytime-valid statistical inference. Equipping applied statisticians with the knowledge of e-values has visible benefits to the sciences and for information technology companies, and for that purpose an accessible technical book is useful.

(d) Within the first half of this decade (2020s), research on e-values has been published at all the major

venues in statistics, machine learning, information theory, and related fields — the Annals of Statistics, the Journal of the American Statistical Association, the Journal of the Royal Statistical Society, Biometrika, Proceedings of the National Academy of Sciences, IEEE Transactions on Information Theory, Bernoulli, Statistical Science, Management Science, Operations Research, Journal of Machine Learning Research, Neural Information Processing Systems, International Conference on Machine Learning, the International Conference on Artificial Intelligence and Statistics, and many more. There have been 3 discussion papers on e-values in the Journal of the Royal Statistical Society within this period. These high-quality modern materials are scattered in different venues, sometimes with different notations and terminologies, due to the fast development. Thus, there can be significant value in collating a good portion of them in one place.

We hope that, by putting the materials together in this book, the concept of e-values becomes more accessible for educational, research, and practical use. It is thus also the hope to stimulate both young and established researchers, especially the next generation of statisticians, to contribute to this exciting and rapidly developing field, and to benefit the application domains of sequential anytime-valid statistical inference in the natural and social sciences. From an educational perspective, the book should be teachable to graduate students in statistics, probability, or computer science programs with good mathematical background, or strong undergraduate students who have a good knowledge of probability theory and statistics.

We do not anticipate this book to be the last word on the topic. In fact, we made the intentional choice to focus on hypothesis testing, and not on estimation. There is a rich and historied literature on *confidence sequences* that we only briefly touch on in this book (Chapter 13), but which is intimately connected to e-values. Further, except for Chapter 7, we also do not get into *sequential* hypothesis testing. This is for many reasons, part of which is that the optimality theory there is less well understood, and many fundamental questions related to filtrations remain open. We anticipate that books dedicated to these topics will be written in future years.

Acknowledgments

We acknowledge our many wonderful coauthors and other scholars over the years on these topics, and our family members, especially our spouses, for their unwavering support. Some parts of the book are based on joint papers with our coauthors, which will be acknowledged in the bibliographic notes in each chapter.

We have received many useful comments and corrections from our readers and colleagues. We would in particular like to thank Christopher Blier-Wong, Nick Koning, Peter Grünwald, Romain Périer, Vladimir Vovk, Will Fithian, and Zhenyuan Zhang for comprehensive comments on various parts of the book.

We also acknowledge the detailed feedback by Aytijhya Saha, Ben Chugg, David Bickel, Diego Martinez Taboada, Hongjian Wang, Isaac Gibbs, Johannes Ruf, Kory Johnson, Lasse Fischer, Matteo Gasparin, Michael Wieck-Sosa, Neil Xu, Nikos Ignatiadis, Pratik Patil, Shubhada Agrawal, Tomas Gonzalez Lara, Yo Joong Choe, and Yusuf Sale, who helped eliminate many typos and inconsistencies in earlier versions of the book. Last, we thank Anastasios Angelopoulos for the LaTeX formatting used for preprints of this book.

Finally, we thank the editors and editorial managers of Foundations and Trends[®] in Statistics and the publisher for excellent editorial assistance.

July 2025
Pittsburgh and Waterloo

Aaditya Ramdas
Ruodu Wang

Lists of notation, conventions, and examples

Standard Probability Notation

Symbol	for ...	Meaning
Ω		sample space
$\mathcal{F}$		a σ -algebra over Ω , or a filtration over Ω
$\mathcal{X}$		the set of all random variables on $(\Omega, \mathcal{F})$
$\mathcal{M}_1$		the set of all probability measures on $(\Omega, \mathcal{F})$
$\mathcal{M}_+$		the set of all nonnegative measures on $(\Omega, \mathcal{F})$
$\mathcal{M}_1(\mathbb{R})$		the set of all Borel probability measures on $\mathbb{R}$
$\mathbb{P}^n$	$\mathbb{P} \in \mathcal{M}_1, n \in \mathbb{N}$	the n -fold product measure of $\mathbb{P}$ on $\Omega \times \cdots \times \Omega$
$\mathbb{Q} \ll \mathbb{P}$	$\mathbb{P}, \mathbb{Q} \in \mathcal{M}_+$	$\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$
$\mathbb{E}^{\mathbb{P}}[X]$	$\mathbb{P} \in \mathcal{M}_+, X \in \mathcal{X}$	$\int X d\mathbb{P}$; it is the expectation under $\mathbb{P}$ if $\mathbb{P} \in \mathcal{M}_1$
$\text{var}^{\mathbb{P}}(X)$	$\mathbb{P} \in \mathcal{M}_1, X \in \mathcal{X}$	variance of X under $\mathbb{P}$
$\sigma(X)$	$X \in \mathcal{X}$	the σ -algebra generated by X
$\mathcal{P}, \mathcal{Q}$		sets of probability measures on $(\Omega, \mathcal{F})$
$\text{Conv}(\mathcal{P})$	$\mathcal{P} \subseteq \mathcal{M}_1$	convex hull of $\mathcal{P}$
$X \stackrel{d}{\sim} \mu$	$\mu \in \mathcal{M}_1(\mathbb{R}), X \in \mathcal{X}$	X is distributed as μ (under some $\mathbb{P}$ from context)
$U[a, b]$	$a < b$	uniform distribution on $[a, b]$
$N(\mu, \sigma^2)$	$\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$	normal distribution with parameter (μ, σ^2)
$\mathbb{1}_A$	$A \in \mathcal{F}$	binary-valued indicator function of event A
δ_x	$x \in \mathbb{R}$ or $x \in \mathbb{R}^n$	Dirac delta distribution at x
Φ		standard normal cumulative distribution function

Standard Mathematics Notation

Symbol	for ...	Meaning
$\mathbb{R}$		set of real numbers $(-\infty, \infty)$
$\mathbb{R}_+$		set of nonnegative real numbers $[0, \infty)$
$\mathbb{N}$		set of positive integers $\{1, 2, \dots\}$
$\mathbb{N}_0$		set of nonnegative integers $\{0, 1, 2, \dots\}$
$[K]$	$K \in \mathbb{N}$	$\{1, \dots, K\}$
$ S $	countable set S	cardinality of S
$a \vee b$	$a, b \in \mathbb{R}$	$\max(a, b)$
$a \wedge b$	$a, b \in \mathbb{R}$	$\min(a, b)$
a_+	$a \in \mathbb{R}$	$a \vee 0$
a_-	$a \in \mathbb{R}$	$(-a) \vee 0$
$\mathbf{e}$		$\exp(1)$
$\log$		natural logarithm to the base $\mathbf{e}$
Δ_K	$K \in \mathbb{N}$	$\{(\lambda_1, \dots, \lambda_K) \in [0, 1]^K : \sum_{k=1}^K \lambda_k = 1\}$
ℓ_K	$K \in \mathbb{N}$	$\sum_{k=1}^K k^{-1}$
2^S	set S	powerset of S (the set of all subsets of S)
$x_{(k)}$	$x_1, \dots, x_K \in \mathbb{R}, k \in [K]$	k -th ascending order statistic of $x_1, \dots, x_K$
$x_{[k]}$	$x_1, \dots, x_K \in \mathbb{R}, k \in [K]$	k -th descending order statistic of $x_1, \dots, x_K$
$\lceil x \rceil$	$x \in \mathbb{R}$	smallest integer n with $n \geq x$
$\lfloor x \rfloor$	$x \in \mathbb{R}$	largest integer n with $n \leq x$
$\mathbb{M}_K$	$K \in \mathbb{N}$	the arithmetic average function
$\bigoplus_{k=1}^K \phi_k$	functions $\phi_1, \dots, \phi_K$	the function $(x_1, \dots, x_K) \mapsto \sum_{k=1}^K \phi_k(x_k)$
$\binom{K}{n}$	$K, n \in \mathbb{N}, n \leq K$	K choose n , equal to $(K!)/(n!(K-n)!)$
$\forall$		for all
$\exists$		there exists
$\mathbf{0}$		a vector of zeros (dimension from context)
$\mathbf{1}$		a vector of ones (dimension from context)

Notation Specific to the Book

Symbol	for ...	Meaning
$\mathfrak{E}, \mathfrak{E}(\mathcal{P})$	$\mathcal{P} \subseteq \mathcal{M}_1$	the set of all e-variables for $\mathcal{P}$
$\mathfrak{U}, \mathfrak{U}(\mathcal{P})$	$\mathcal{P} \subseteq \mathcal{M}_1$	the set of all p-variables for $\mathcal{P}$
$\mathcal{E}$	$\mathcal{E} \subseteq \mathfrak{E}$	a generic set of e-variables
$\mathcal{U}$	$\mathcal{U} \subseteq \mathfrak{U}$	a generic set of p-variables
E	$E \in \mathfrak{E}$	a generic e-variable
P	$P \in \mathfrak{U}$	a generic p-variable
$\mathcal{K}$		the set $\{1, \dots, K\}$, indexing multiple hypotheses

Abbreviations

Abbreviation	Meaning
iid	independent and identically distributed
pdf	probability density function
cdf	cumulative distribution function
pmf	probability mass function
KL	Kullback–Leibler (divergence), relative entropy
BH	Benjamini–Hochberg (procedure)
CI	confidence interval
CS	confidence sequence
BY	Benjamini–Yekutieli (procedure)
SAVI	sequential anytime-valid inference

Conventions

1. Scalar constants and functions are usually represented by small letters, such as a , p , e , f , or x .
2. The constants e (often a realized e-value) and $\mathbf{e}$ (exponential of 1) are different, and we avoid using them in the same place.
3. Boldface letters like $\mathbf{x}$ refer to vectors, and its k -th entry is x_k (not $\mathbf{x}_k$). All vectors are by default column vectors.
4. Capital letters such as X , P and E are usually reserved for random variables and functions, but there are exceptions made for important constants (such as the number of hypotheses K in multiple testing). Random vectors are denoted like $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{E}$. We sometimes also use X to represent the entire data, which could be one-dimensional or multi-dimensional.
5. Terms like “increasing” and “decreasing” are in the non-strict (weak) sense, but “positive” and “negative” are in the strict sense.

6. Inequalities between vectors and functions are always understood componentwise.
7. The summation $\sum$ over the empty set is 0, and the product $\prod$ over the empty set is 1.
8. The set inclusion is written as $\subseteq$ and the strict set inclusion is written as $\subsetneq$.
9. When $\mathbb{E}^{\mathbb{P}}[F(X)]$ appears, the suitable measurability of F is tacitly assumed.
10. We sometimes omit the domain of integrals and it is understood as the natural domain (e.g., $\mathbb{R}$ or Ω), which should be clear from the context.
11. When we do not provide the proof of a nontrivial result, we will point to its proof in the literature in the bibliographical note section of the corresponding chapter.
12. We generally rely on standard conventions such as $0 \cdot x = 0$ for $x \in [-\infty, \infty]$ and $x \cdot \infty = \infty$ for $x \in (0, \infty]$, as well as $\log(0) = -\infty$ and $\log(\infty) = \infty$. When evaluating ratios of $[0, \infty]$ -valued random variables we often use the following convention: $x/\infty = 0$, $\infty/x = \infty$ for $x \in [0, \infty)$, and $\infty/\infty = 1$.
13. As standard in probability theory, we remind the reader that conditional expectations like $\mathbb{E}^{\mathbb{P}}[Y|X]$ for random variables X, Y and $\mathbb{E}^{\mathbb{P}}[Y|\mathcal{F}]$ for a σ -algebra $\mathcal{F}$ are defined only in the $\mathbb{P}$ -almost sure sense.
14. If $\mathbb{E}[\cdot]$ or $\text{var}(\cdot)$ appears without the probability specified, it is with respect to the true data-generating probability measure.

List of examples of e-values

Context	Null	Alternative	Section
Likelihood ratio	$\mathbb{P}$	$\mathbb{Q}$	1.6
Soft-rank e-value	exchangeable data	otherwise	1.7
Normal two-sided	$N(0, 1)$	$N(\mu, 1), \mu \neq 0$	2.1
Monotone likelihood ratio	$p_\theta : \theta \leq \theta_0$	$p_\theta : \theta \geq \theta_1$	2.6.1; 6.7.2
Symmetry	symmetry around zero	otherwise	2.6.2; 6.7.3
Mean with nonnegative data	$\mathbb{E}[X] \leq \mu, X \geq 0$	$\mathbb{E}[X] > \mu$	2.6.3
Mean and variance	$\mathbb{E}[X] = \mu, \text{var}(X) \leq \sigma^2$,	$\text{var}(X) > \sigma^2$	2.6.3; 10.3
Sub-Gaussian distributions	sub-Gaussian, mean ≤ 0	otherwise	2.6.4; 6.7.5
Sub-Gaussian distributions	sub-Gaussian, mean $= 0$	otherwise	2.6.4
Likelihood ratio bounds	$d\mathbb{P}/d\mathbb{P}_0 \leq \gamma$ for given $\mathbb{P}_0$	$\mathbb{Q}$	2.6.6; 6.7.6
Exchangeability	exchangeable data	$\mathbb{Q}$	2.6.7; 6.7.7
Mixture method	$\mathbb{P}$	$\mathcal{Q}$	3.7
Plug-in method	$\mathbb{P}$	$\mathcal{Q}$	3.7
T-test (unknown σ)	$N(0, \sigma^2), \sigma$ unknown	$N(\mu, \sigma^2), \mu \neq 0$	3.7
Changepoint (unknown location)	$\mathbb{P}^n$	$\{\mathbb{P}^{n-k}\mathbb{Q}^k\}_{k \in [n]}$	3.7
Universal inference e-value	$\mathcal{P}$	$\mathcal{Q}$	5.2
Nomeraire	$\mathcal{P}$	$\mathbb{Q}$	6.1
Exponential family	$p_\theta : \theta \leq \theta_0$	p_{θ_1} with $\theta_1 > \theta_0$	6.7.1
Mean with support $[0, 1]$	$\mathbb{E}[X] \leq \mu$	$\mathbb{E}[X] > \mu$	6.7.4
Likelihood ratio process	$\mathbb{P}$	$\mathbb{Q}$	7.5
Universal inference e-process	$\mathcal{P}$	$\mathcal{Q}$	7.7
Empirically adaptive e-process	$\mathcal{P}$	$\mathcal{Q}$	7.8
Time-mixture e-processes	$\mathcal{P}$	$\mathcal{Q}$	7.9
Multi-armed bandit testing	$\mathbb{E}[X_{k,j} \mathcal{F}_{j-1}] \leq 1 \ \forall j$	$\mathbb{E}[X_{k,j} \mathcal{F}_{j-1}] > 1 \ \exists j$	9.1.1
Separable compound e-values	$\{\mathbb{S} : \mathbb{S}_k = \mathbb{P}_k\}, k \in [K]$	$\mathbb{Q}_k, k \in [K]$	9.3.2
Asymptotic zero-mean	$\mathbb{E}[X] = 0, \text{var}(X) \in (0, \infty)$	otherwise	10.3
Compound t-test (unknown σ_k)	$N(0, \sigma_k^2), k \in [K]$	$N(\mu_k, \sigma_k^2), k \in [K]$	10.4.4
Testing the β -quantile (VaR_β)	$\text{VaR}_\beta(X) \leq r$	$\text{VaR}_\beta(X) > r$	16.3.2
Testing an expected loss	$\mathbb{E}[\ell(X)] \leq r$	$\mathbb{E}[\ell(X)] > r$	16.3.2
Testing the risk measure ES	$\text{ES}_\beta(X) \leq r, \text{VaR}_\beta(X) = z$	$\text{ES}_\beta(X) > r$	16.3.3

Part I

Fundamental Concepts

Chapter 1

Introduction

An e-value is a nonnegative test statistic whose expected value is at most one under the null hypothesis.

We begin this chapter by offering high-level discussions on the concept of e-values and its comparison to p-values in Sections 1.1 and 1.2. Formal mathematical notation and conventions are introduced in Section 1.3, and formal definitions of e-values, p-values, and tests are given in Section 1.4. The first examples are then discussed in Sections 1.6 and 1.7. Historical context on e-values is provided in Section 1.8, and a road map for the rest of the book is described in Section 1.9.

Below are some simple messages, all formally treated in the book. E-values, or e-variables as we usually refer to the underlying random variables, are a fundamental tool for hypothesis testing. E-values are directly interpreted as evidence against the null: the larger the e-value, the more evidence that the null is false. The “e” in e-value could stand for *expectation* (their defining property) or for *evidence* (their interpretation).

The most straightforward way to view e-values is that they are composite generalizations of likelihood ratios. For testing a simple null hypothesis, all e-values are likelihood ratios (or are dominated by them). For testing a composite null hypothesis, building such “generalized likelihood ratios” (which are e-values) is always possible but not simple, and we will spend several chapters of the book discussing methodology and theory in this case.

E-values are named as such in order to evoke a comparison to p-values, which are a much more classical statistical tool within hypothesis testing. The naming analogy is that under the null hypothesis,

e-values are *expectation* constrained
versus
p-values are *probability* constrained,

since p-values must be uniformly distributed (or stochastically larger) under the null. Another analogy is that e-values can be seen as certain *conditional expectations* on data, and p-values can be seen as certain *conditional probabilities* on data. One important interpretational difference is that for p-values, smaller values suggest that the null hypothesis is false, but it is the opposite for e-values.

Writing conventions.

Similarly to the situation of “p-value”, which has several variants, it is possible that different authors might write “e-value” differently.

Our default choice of writing convention, also our recommendation to other authors, is to write “e-value” (as well as related terms) with the small letter “e”, and without making it italic. When it is the initial letter of a title or a sentence, “E” (instead of writing “e-Value”) should be capitalized just like in a usual English word; otherwise we do not capitalize it. This convention is adapted by the authors of this book and most of their co-authors.

Some journals have specific styles. For instance, some may require both “e” in e-values and “p” in p-values to be italic, so it is possible to see “*e*-value” in journal publications.

1.1 The three roles of e-values

Our book will attempt to elaborate on three somewhat distinct roles that e-values play in statistical inference: as methodological tools with advantages in particular applications, as technical tools to prove other results of independent interest, and as fundamental tools lying at the very heart of statistics. We elaborate on these now, pointing forward to which angles are elaborated in which chapters.

1. **Methodological.** E-values easily yield the more common inferential tools of p-values and hypothesis tests: indeed, the inverse of an e-value is a p-value (e-to-p calibration; Chapter 2), and rejecting the null when the e-value exceeds $1/\alpha$ is a valid level- α test (Markov’s inequality; Chapter 2). In fact, sometimes e-values yield the only known path to performing statistical inference: Chapter 5 presents a general technique called universal inference to derive e-values for (possibly) irregular testing problems, for which no other inferential methods are known. Chapter 6 then shows that for a fixed alternative, universal inference can be improved to a unique “log-optimal” e-value, which always exists without regularity assumptions. Chapter 7 demonstrates how e-values enable sequential anytime-valid inference. E-values play a key role in selective inference by providing simple methods to control the false discovery rate (Chapter 9) and false coverage rate (Chapter 13), and they also yield nonparametric tests for risk measure forecasts (Chapter 16).
2. **Technical.** E-values appear as technical tools to solve problems that may appear to have nothing to do with e-values in their problem statement. For example, we will show that every procedure that controls the false discovery rate is actually an instance of the e-Benjamini–Hochberg procedure applied to certain *compound* e-values (Chapter 9). There are no such results available for p-values or the original Benjamini–Hochberg procedure. These (compound) e-values may not be of direct interest in and of themselves, but appear as intermediate technical tools towards some other aim, such as combining inferences across many multiple testing procedures or derandomizing multiple testing procedures that utilize external randomness. Similarly, when combining arbitrarily dependent p-values into a single combined p-value, all admissible merging functions must employ e-values under the hood (Chapter 12). Methods based on e-values are intimately connected to other theories in modern mathematics, notably the reverse information projection (Chapter 6) and nonnegative (super)martingales (Chapter 7); once more, e-values arise centrally in both settings, whether or not one is interested in them.
3. **Fundamental.** Chapter 3 points out that for testing any composite null $\mathcal{P}$ against any composite alternative $\mathcal{Q}$, without restriction, there exists a powered test (a test with a prespecified level under any null, and power greater than that level against any alternative) if and only if there exists a bounded powered e-variable (a variable with expected value at most one under any null, and expected value larger than one under any alternative). Chapter 4 shows that for general decision-making problems with more than two decisions, or with data-dependent loss functions, any decision that guarantees risk control must be (implicitly) based on e-values.

Overall, we view the current book as providing the building blocks, namely e-values, of *a theory of evidence*. This is contrast to, say, *a theory of decision making* as developed by Neyman, Pearson, Wald and Savage. While the two are not at complete odds, they are also not in complete agreement.

Given that one might argue that p-values also play similar roles in the literature as written above, we next address the relationship of e-values to p-values.

1.2 P-values versus e-values

The book does not seek to displace or eliminate the use of p-values. Instead, we argue that e-values are complementary to p-values, and indeed a critical part of the modern statistician’s toolkit. We also argue that e-values and p-values should be studied on an equal footing, and that they may each have advantages in different settings.

1.2.1 E-values and p-values are statistical cousins

P-values have a long and rich history in statistics, long considered a standard tool in theory, methodology and applications. E-values, by contrast, have been relatively late to enter the scene, and are not currently considered by common scientists to be as fundamental on all three aforementioned fronts. However, we argue that e-values and p-values should in fact be mentioned in the same breath — they are closely related, equally fundamental, and we would like both to occupy an equal footing in the literature.

This is a tall ask, but e-values are up to the challenge. It is immediately apparent from their definitions that p-values and e-values are rather different objects: they are both test statistics of the data but with entirely different properties. Thus, we will focus here on the commonalities between e-values and p-values, in order to make the case that one should view them as statistical “cousins”:

- First, as classes of objects, p-variables, e-variables and tests exist under exactly the same conditions. Theorem 3.9 proves that there exists a bounded powered e-variable if and only if there exists a powered level- α test for some (or every) α . This holds without restriction on the (possibly composite) null hypothesis being tested, or on the composite alternative. Further, under a very mild assumption (U) that there exists a random variable that is uniformly distributed under every null and alternative, independent of all other random variables, the preceding existence statements are also equivalent to the claim that a powered p-variable exists.
- Second, as mentioned earlier, e-variables and p-variables can be transformed to each other by “calibration” (Section 2.3). While e-values and p-values are in general just real-valued measures of evidence, they can be thresholded to yield rejection decisions of the null hypothesis with guaranteed type-I error control. While e-to-p calibration with $e \mapsto 1/e$ preserves the final decision (for any α , the calibrated p-value is below α if and only if the original e-value is above $1/\alpha$), this is in general not true for all p-to-e calibrators. Nevertheless, the particular p-to-e calibrator $p \mapsto \mathbb{1}_{\{p \leq \alpha\}}/\alpha$ results in the all-or-nothing e-value that preserves the final decision at a fixed level α .

Despite the above similarities, some key differences are as follows:

- **E-values allow for post-hoc decision making.** For testing, the level α can be chosen to depend on the data while maintaining a nontrivial guarantee (Section 4.1). More generally, if one needs to make decisions beyond just accept/reject, e-values allow handling of data-dependent loss functions (Section 4.2). Critically, this is true for *all* e-values, and while this may be true for *some* p-values (like those derived from e-values), it is provably not true for all p-values.
- **E-values allow for optional continuation of experiments.** Suppose a scientist collects some data that appear promising but inconclusive to make a clear decision of rejecting a null hypothesis. Seeing this result, another scientist may then be inspired to run their own follow-up experiment, and their e-values can be merged by a simple multiplication. On the other hand, sequentially obtained p-values are difficult to merge, because the very existence of the latter studies may depend on the p-value obtained from the first experiment, creating a subtle form of dependence that could lead to incorrect meta-analyses. These issues are discussed in detail in Sections 4.3. This is also related to “p-hacking” (Section 7.1), for which e-processes provide a compelling solution.
- **E-values are suitable for unknown sampling scheme.** The construction and validity of p-values rely crucially on a completely specified design of the experiment, in particular the data collecting procedure. If a dataset is simply handed to us without specifying how it was collected, one may not be

able to correctly calculate a p-value, but we are often able to calculate an e-value, for a wide range of reasonable sampling schemes. This is illustrated in Example 1.4.

- **Compound e-values are central to multiple hypothesis testing.** Chapter 9 shows that every multiple testing procedure that controls the false discovery rate can be written an instance of the e-Benjamini–Hochberg procedure with certain compound e-values, while Chapter 11 shows that every admissible way for combining dependent p-values must proceed via converting them to e-values and averaging those e-values. Chapter 13 illustrates how e-values are useful in constructing and combining dependent confidence intervals.

There are other differences one can point out in more specific settings. For example, when testing whether a Gaussian mean is nonpositive, there exists a uniformly most powerful p-value which does not depend on the (positive) alternative mean. This is not true for e-values, whose *e-power* typically depends on the alternative, even in special situations such as the one considered. In contrast, for a fixed alternative, a log-optimal (“uniformly most e-powerful”) e-value always exists for any composite null (Chapter 6), but uniformly most powerful p-values do not exist outside of very special cases.

Thus, p-values and e-values are very different test statistics with several contrasting properties; Chapter 14 contains more such contrasts. At a high level, p-values may be better choices in certain rigid and constrained settings, but e-values may be better choices in more open-ended or flexible settings. Yet, they are strongly related as families of objects, and can be converted to each other. This justifies our calling them “statistical cousins”.

1.2.2 Power and e-power

A question that arises very often, perhaps the first question of many users, is

Are e-values more or less powerful than p-values?

This question requires some pause to carefully interpret and answer. E-values and p-values are summary statistics and do not have “power” by themselves. Power refers to the probability of rejection under some specific alternative hypothesis. The questioner often means *if I construct a level- α test by rejecting an e-value when it exceeds $1/\alpha$, does it have lower power than the level- α test that rejects when a p-value is smaller than α* . Or, *if I embed an e-value in a decision making (e.g., multiple testing) procedure, will I make better decisions (e.g., more discoveries)?* This question does not specify which p-values and e-values one is using. So our answer can be split into four cases:

1. If the p-value P was itself derived from the e-value E (since $1/E$ is a p-value; see Chapter 2), then clearly the two tests are identical and there is no difference in power.
2. If one starts with a p-value P , and uses it to define the e-value $E = \mathbf{1}_{\{P \leq \alpha\}}/\alpha$, then clearly the resulting test has the same power at level- α as that based on the original p-value.
3. If a given p-value P is converted to an e-value E using a different p-to-e calibrator than above (Chapter 2), then the probability that E exceeds $1/\alpha$ will be smaller than the probability that P is below α (see also Section 2.7).
4. When the p-value and e-value are not derived from each other in the above ways, no general comparisons can be made.

The second point above can be generalized to an important, though pedantic, observation — any level- α test (even those that may or may not be based on p-values) can be recovered by thresholding an appropriate e-variable at $1/\alpha$ (Chapter 3). Indeed, if A is the rejection region of some level- α test, meaning that we reject when the data lies in A , then $E = \mathbf{1}_A/\alpha$ is an “all-or-nothing” e-value that equals or exceeds α exactly when the original test rejects. Thus, clearly, using e-values does not by itself lose expressibility or power in any way.

It is worth emphasizing that while we repeatedly use $1/\alpha$ as the “default” threshold for converting e-values to level- α tests, one must be somewhat more careful, keeping the following points in mind:

- **Full information, possible randomization.** If the distribution of the e-value under the null hypothesis is known (and it could be different under each element of the null), then it is possible to design a threshold smaller than $1/\alpha$ (Section 2.2). If the null hypothesis is simple or if the e-value is *pivotal*, meaning that its distribution is invariant under the composite null hypothesis, then this type-I error can be made exactly α (possibly using randomization, as in the classical Neyman–Pearson lemma).
- **No information, randomization.** It is always possible to reduce the threshold for rejection (from $1/\alpha$) by a random amount using the randomized Markov’s inequality (Section 2.4). One could equivalently view this as converting the e-value to a smaller randomized p-value, or equivalently stochastically rounding the e-value to an ‘all-or-nothing’ e-value that has larger probability of exceeding $1/\alpha$ (power) than the original (Section 9.5).
- **Partial information, no randomization.** If partial information is available about the e-value, for example if the null distribution of the e-value satisfies certain shape constraints, then one can utilize thresholds that are below $1/\alpha$ (Chapter 15).
- **Multiple e-values, possible randomization.** If multiple exchangeable e-values must be combined (which could occur when the e-values are calculated using a randomized procedure, such as sample splitting), then simply averaging them and thresholding at $1/\alpha$ is suboptimal, and one can improve on it by using an exchangeable improvement of Markov’s inequality (Chapter 5). If the e-values are arbitrarily dependent, they can be made exchangeable by applying a uniformly random permutation, and then the previous sentence applies.

The above points may not alleviate all concerns. Indeed, users who directly plug in specific e-values from this book into their favorite decision procedures (that are currently based on p-values) may indeed find that the e-value procedures are less powerful. Why is that? Here is the main reason.

The typical e-values in this book are optimizing a different criterion than power. Indeed, the preferred criterion in this book is called “e-power” (Chapter 3), and corresponds to the expected logarithm of the e-value under the alternative $\mathbb{Q}$, that is, $\mathbb{E}^{\mathbb{Q}}[\log(E)]$. On that note, we make the following two observations:

- If we were to optimize for power at level α , we would simply get an all-or-nothing e-value $\mathbf{1}_A/\alpha$ for some set A . These have e-power equal to minus infinity, because they take on the value zero with non-negligible probability, which is infinitely penalized by the logarithm.
- Log-optimal e-values, in contrast, are those which maximize the e-power against $\mathbb{Q}$ (these always exist, even for composite nulls; Chapter 6). Log-optimal e-values obviously avoid the value zero: they are $\mathbb{Q}$ -almost surely strictly positive.

Since our “preferred” e-values are log-optimal, they are rather different from the ones that maximize power. Thus, directly plugging them into the reader’s favorite procedures (which may be judging the results by more traditional metrics like power), may lead to disappointment. This is not a fault of e-values, but of *which e-values one chooses to use*. Carefully aligning the optimality criterion of the e-value with that of the procedure within which they are used is likely to return more favorable results. Indeed, we already pointed out that you certainly cannot *lose* anything by thinking in terms of e-values instead of p-values.

The reader may then wonder: what makes us prefer e-power (and log-optimality) as the criterion of choice? Indeed, one could take any increasing function f and ask for the e-value E that optimizes $\mathbb{E}^{\mathbb{Q}}[f(E)]$; power at level α and the logarithm are only two such choices. Nevertheless, the logarithm is a *natural* choice if you take into account that e-values are interpreted as measures of evidence against the null, and one that is particularly easy to update when more evidence is collected.

For example, consider the optional continuation setting in Section 1.2.1. Suppose a sequence of scientists collect a sequence of e-values $E_1, E_2, \dots$ against the null; then $\prod_{i=1}^n E_i$ is also an e-value, even if the e-values are sequentially dependent, meaning that the decision to collect E_i was made after observing $E_1, \dots, E_{i-1}$. Now it may be immediately apparent why one wants to avoid zeros: if $E_3 = 0$, no future multiplication of large e-values will ever result in large evidence.

Further, note that $\prod_{i=1}^n E_i = \exp(n \cdot \overline{\log E}_{[n]})$, where $\overline{\log E}_{[n]} = \sum_{i=1}^n \log E_i / n$, so that the product increases to infinity whenever $\overline{\log E}_{[n]}$ converges to a positive constant. Further, in that case, the product

converges to infinity exponentially fast, and choosing the e-values that maximize $\mathbb{E}^{\mathbb{Q}}[\overline{\log E_{[n]}}]$ (under the alternative) also maximizes the *rate* at which the product approaches infinity. Here, the logarithm was not simply a “convenient choice of a utility function”, but arises naturally in reasoning about a sequence of products. We also offer an axiomatic derivation of the logarithm in Chapter 3.

We are now ready to begin with the formal mathematical framework of concepts in the book.

1.3 Mathematical notation and conventions

We begin with a sample space Ω equipped with a σ -algebra $\mathcal{F}$, and the set $\mathcal{M}_1$ of all probability measures on $(\Omega, \mathcal{F})$. Elements of $\mathcal{M}_1$ are also called distributions. Formally, a random variable is a measurable function from $(\Omega, \mathcal{F})$ to $\mathbb{R}$ or $[-\infty, \infty]$. When we allow for a random variable to take infinite values, we will write it explicitly; otherwise it takes values in $\mathbb{R}$. We assume that some distribution $\mathbb{P}^\dagger \in \mathcal{M}_1$ governs our data X (which is a random variable). Note that X could be a vector $(X_1, \dots, X_n)$. The variables $X_1, \dots, X_n$ could be independent and identically distributed (iid) under $\mathbb{P}^\dagger$, but they need not necessarily be so. We use X^t as a shorthand for the first t data points $(X_1, \dots, X_t)$.

Definition 1.1

A *hypothesis* is a set of probability measures in $\mathcal{M}_1$. A hypothesis is *simple* if it is a singleton, such as $\{\mathbb{P}\}$ and $\{\mathbb{Q}\}$. Otherwise it is *composite*.

We will always use $\mathcal{P}$ for the null hypothesis and $\mathcal{Q}$ for the alternative hypothesis. We will also write H_0 as the null hypothesis that $\mathbb{P}^\dagger \in \mathcal{P}$, and H_1 the alternative hypothesis that $\mathbb{P}^\dagger \in \mathcal{Q}$; that is, we interpret a hypothesis both as a mathematical object $\mathcal{P}$ and as a statistical statement H_0 . For a simple hypothesis $\mathcal{P} = \{\mathbb{P}\}$, we say “testing H_0 ”, “testing $\mathcal{P}$ ”, and “testing $\mathbb{P}$ ” interchangeably. The sets $\mathcal{P}$ and $\mathcal{Q}$ are always assumed to be non-intersecting. Let $\text{Conv}(\mathcal{P})$ denote the convex hull of $\mathcal{P}$.

As alluded above, we always use $\mathcal{P}, \mathcal{Q}$ for sets of distributions, and $\mathbb{P}, \mathbb{Q}$ for a single distribution. In the context of a hypothesis $\mathcal{P}$, statements on independence, distribution, and almost-sure equalities are meant to hold under each member. We say that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, denoted $\mathbb{Q} \ll \mathbb{P}$, if $\mathbb{P}(A) = 0$ implies that $\mathbb{Q}(A) = 0$ for any measurable set A . In this case, we use $d\mathbb{Q}/d\mathbb{P}$ to refer to the likelihood ratio (or, more generally, the Radon–Nikodym derivative) between $\mathbb{Q}$ and $\mathbb{P}$. Letting q and p represent the densities of $\mathbb{Q}$ and $\mathbb{P}$ with respect to some reference measure $\mathbb{L}$ (often the Lebesgue measure, hence the symbol), we can write $(d\mathbb{Q}/d\mathbb{P})(x) = q(x)/p(x)$. The expectation of a random variable X under $\mathbb{P}$ is denoted by $\mathbb{E}^\mathbb{P}[X] = \int X d\mathbb{P}$.

Some notation will be used throughout. The set $\mathbb{N} = \{1, 2, \dots\}$ is set of all positive integers, and $\mathbb{R} = (-\infty, \infty)$ is the set of all real numbers. We say that a random variable X is stochastically larger (under $\mathbb{P}$, sometimes omitted if clear from the context) than a distribution on $\mathbb{R}$, represented by its cumulative distribution function (cdf) F , if $\mathbb{P}(X \leq x) \leq F(x)$ for all $x \in \mathbb{R}$. We write $X \stackrel{d}{\sim} F$ if the distribution of X is F (either a cdf or a distribution) under a probability measure $\mathbb{P}$ that should be clear from context; the notation $\stackrel{\text{iid}}{\sim}$ is similar. For a positive integer K , we denote by $[K] = \{1, \dots, K\}$ and by Δ_K the standard simplex in $\mathbb{R}^K$, that is,

$$\Delta_K = \left\{ (v_1, \dots, v_K) \in [0, 1]^K : \sum_{k=1}^K v_k = 1 \right\}.$$

The Euler number $\exp(1)$ is denoted by $\mathbf{e}$ (please note that this is different from the variable e , but of course we will avoid using both in the same equation whenever possible).

All terms like “increasing” and “decreasing” are in the non-strict sense. We write $x \wedge y$ and $x \vee y$ for the minimum and the maximum of real numbers x, y respectively.

We will introduce some more specific notation in different chapters. Lists of commonly used notation and convention are collected separately in the beginning of the book, for readers navigating non-sequentially.

1.4 Definitions of e-variables, p-variables, and tests

As above, let $\mathcal{P} \subseteq \mathcal{M}_1$ be a set of probability measures representing the null hypothesis.

Definition 1.2: E-variables, p-variables, and tests

- (i) An *e-variable* E for $\mathcal{P}$ is a $[0, \infty]$ -valued random variable satisfying $\mathbb{E}^{\mathbb{P}}[E] \leq 1$ for all $\mathbb{P} \in \mathcal{P}$. An *e-variable* E is *exact* if $\mathbb{E}^{\mathbb{P}}[E] = 1$ for all $\mathbb{P} \in \mathcal{P}$.
- (ii) A *p-variable* P for $\mathcal{P}$ is a $[0, \infty)$ -valued random variable satisfying $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ and all $\mathbb{P} \in \mathcal{P}$. A *p-variable* P is *exact* if $\mathbb{P}(P \leq \alpha) = \alpha$ for all $\alpha \in (0, 1)$ and all $\mathbb{P} \in \mathcal{P}$.
- (iii) A *test* ϕ is a $[0, 1]$ -valued random variable. A test is *binary* if its range is $\{0, 1\}$. The *type-I error (rate)* of a test ϕ for $\mathbb{P}$ is $\mathbb{E}^{\mathbb{P}}[\phi]$. A test ϕ has *level* $\alpha \in [0, 1]$ for $\mathcal{P}$ if its type-I error is at most α for every $\mathbb{P} \in \mathcal{P}$.

We denote by $\mathfrak{E} = \mathfrak{E}(\mathcal{P})$ (Fraktur E) the set of all e-variables for $\mathcal{P}$ and by $\mathfrak{U} = \mathfrak{U}(\mathcal{P})$ (Fraktur U) the set of all p-variables for $\mathcal{P}$, with $\mathcal{P}$ often omitted.

Above, we use $\mathfrak{U}$ (hinting at “uniform”) instead of Fraktur P to avoid overloading the letter “P”, which is already used for p-variables (P), probability measures ($\mathbb{P}$), and null hypotheses ($\mathcal{P}$).

The value taken by the e-variable or p-variable upon observing the data is called an *e-value* or a *p-value*. As such, we can see that e/p-values are actually not mathematically rigorously defined objects, although they are common terms in the statistical literature. In this book, we always use “e/p-variables” when making a mathematical statement or result about the corresponding random variables, and mention “e/p-values” when their statistical interpretation is more important. In the latter case, an e/p-value may refer to the random variable or its realized value, and this should be clear from context.

Some immediate remarks are made below, but the reader may directly jump to Section 1.6 to see a simple example of an e-variable, which is understandable without reading the following remarks.

1. Our definition of a test allows for external randomization: upon observing the data, the realized value of a test ϕ is interpreted as the probability of rejecting the null hypothesis. More explicitly, one imagines drawing a uniform random variable U independent of the data, and reject the null if $U \leq \phi$. With this randomization mechanism, the type-I error for $\mathbb{P}$, $\mathbb{E}^{\mathbb{P}}[\phi]$, is precisely the unconditional probability of rejecting the null hypothesis $\mathbb{P}$, when the data is drawn from $\mathbb{P}$. A binary test rejects or does not reject without using randomization.
2. A p-variable defined as a random variable that is stochastically larger than a uniform random variable U on $[0, 1]$ under all $\mathbb{P} \in \mathcal{P}$ (an equivalent formulation is $\mathbb{E}^{\mathbb{P}}[f(P)] \geq \mathbb{E}^{\mathbb{P}}[f(U)]$ for all increasing functions f and $\mathbb{P} \in \mathcal{P}$). A p-variable is often truncated at 1; this usually does not affect our discussions. We occasionally mention that a random variable P is a p-variable even if it can take the value ∞ (such as one over an e-variable; see Proposition 2.1), but it should be understood as saying that $P \wedge 1$ is a p-variable.
3. A large realized e-variable is interpreted as evidence against the null hypothesis (because their expected value is no larger than 1 under the null). Similarly, if we observe a small realized p-variable, then we have evidence against the null hypothesis. This interpretation will be kept throughout the book. Two results in Chapter 14 will show that e-variables are naturally represented by conditional expectations, whereas p-variables are naturally represented by conditional probabilities.
4. We sometimes use the phrase “an e-variable for H_0 ” to mean “an e-variable for $\mathcal{P}$ ” where the hypothesis H_0 is identified by $\mathcal{P}$ (as in Definition 1.1). If $\mathcal{P} = \{\mathbb{P}\}$, then we also say “an e-variable for $\mathbb{P}$ ”.
5. Most tests that we will deal with are binary, with 1 representing rejection and 0 representing no rejection. Any level-0 test ϕ_0 must satisfy $\phi_0 = 0$ $\mathbb{P}$ -almost surely for every $\mathbb{P} \in \mathcal{P}$, and we typically choose $\phi_0 = 0$. As standard in hypothesis testing, the general aim is that we hope to not reject the null hypothesis H_0 when the data are generated by some element of $\mathcal{P}$, and we hope to reject H_0 when the data are generated by some element of $\mathcal{Q}$.

6. An e-variable is allowed to take the value ∞ ; observing $E = \infty$ for an e-variable E means that we are entitled to reject the null hypothesis; this corresponds to observing 0 for a p-variable. However, E must be $\mathbb{P}$ -almost surely finite for every $\mathbb{P} \in \mathcal{P}$ (otherwise its expectation would not be bounded by 1 under $\mathbb{P}$). In particular, E is allowed to be infinite only on $\mathcal{P}$ -nullsets (which are sets of measure zero under every $\mathbb{P} \in \mathcal{P}$), which correspond to events that are impossible under the null (but may be possible under some alternative). At least one e-variable always exists for any $\mathcal{P}$: the trivial e-variable that always equals 1 ($\mathbb{P}$ -almost surely for every $\mathbb{P} \in \mathcal{P}$). Of course, any constant smaller than one also yields a valid e-variable, but such an e-variable would be dominated by 1.

We summarize some useful properties of p-variables and e-variables below, which can be immediately checked from Definition 1.2.

Fact 1.3

For a random variable X and a distribution $\mathbb{P} \in \mathcal{M}_1$, let $F_{X|\mathbb{P}}$ be the cdf of X under $\mathbb{P}$.

- (i) For any random variable X and distribution $\mathbb{P}$, $F_{X|\mathbb{P}}(X)$ is a p-variable for $\mathbb{P}$.
- (ii) For any random variable X and hypothesis $\mathcal{P}$, $\sup_{\mathbb{P} \in \mathcal{P}} F_{X|\mathbb{P}}(X)$ is a p-variable for $\mathcal{P}$.
- (iii) If $E_1, \dots, E_K$ are e-variables for $\mathcal{P}$, then their arithmetic average is an e-variable for $\mathcal{P}$.
- (iv) If $E_1, \dots, E_K$ are e-variables for $\mathcal{P}$ and they are independent under each $\mathbb{P} \in \mathcal{P}$, then their product is an e-variable for $\mathcal{P}$.
- (v) If P is a p-variable for $\mathcal{P}$ and $P' \geq P$, then P' is a p-variable for $\mathcal{P}$.
- (vi) If E is an e-variable for $\mathcal{P}$ and $0 \leq E' \leq E$, then E' is an e-variable for $\mathcal{P}$.

In standard statistical practice, p-values are often computed from (i)–(ii) of Fact 1.3, where X is a test statistic. The combination properties of e-variables in (iii)–(iv) of Fact 1.3 are not shared by p-variables. The combination of e-values and that of p-values are formally treated in Chapters 8 and 12, respectively.

Although we defined e-variables and p-variables together, the concept and methodologies of e-values are not dependent on those of p-values. We will formally connect the two concepts in Chapter 2, due to the familiarity of p-values to the statistical community, while we should keep in mind that testing with e-values without relying on p-values is the main purpose of this book. Moreover, in Chapters 9, 11, and 13, we will see how e-values are fundamental for multiple testing, both with and without p-values.

E-variables are also useful, or arguably even central, to sequential hypothesis testing, as we will define and discuss formally later. Informally, a finite or infinite sequence $E_1, E_2, \dots$ of e-variables (adapted to an underlying filtration) is called an *e-process* if for every stopping time τ (with respect to that filtration), E_τ is also an e-value. This definition, despite its simplicity, is rather nontrivial and hides some deep connections to the theory of nonnegative (super)martingales that we discuss and explore in Chapter 7.

Next, we discuss the betting interpretation of e-values, followed by some simple examples.

1.5 Betting interpretation of e-values

E-variables have a clear interpretation in terms of betting. An e-variable for $\mathcal{P}$ is a valid bet against $\mathcal{P}$. To elaborate, imagine two rational parties called Forecaster and Skeptic (here, by “rational” we mean that they are risk-neutral, i.e., expected value maximizers). Forecaster claims that some $\mathbb{P} \in \mathcal{P}$ describes the distribution of a particular (currently unseen) outcome in a satisfactory way. Skeptic does not believe it and would like to bet against Forecaster. The process is formalized using a statistical contract: Skeptic pays one dollar up front, with the promise of getting $E(\omega)$ back if outcome $\omega \in \Omega$ is observed (such a function $E : \Omega \rightarrow \mathbb{R}_+$ must be specified by Skeptic; this is the bet). If the contract is accepted by Forecaster, they receive a dollar, then the outcome ω is observed, and they pay Skeptic $E(\omega)$ dollars. When would Forecaster accept the contract? When its terms are favorable to Forecaster. This happens when E is an e-variable,

because then Forecaster believes that they will make money by partaking in the contract. Naturally, if Skeptic believes that $\mathbb{Q}$ describes the world better than $\mathbb{P}$, they will suggest E only if $\mathbb{E}^{\mathbb{Q}}[E] > 1$. Here E , the e-variable, is the bet, while its realization $E(\omega)$ is an e-value, also called a betting score.

P-values also have a betting interpretation, albeit a flawed one. The setup, continuing from above, goes as follows. For every $r \in (0, 1]$, Forecaster offers a contract priced at a dollar, which pays $1/r$ dollars if some statistic P of the data is smaller than r , and nothing otherwise. When would Forecaster offer such a contract? Only when P is a p-variable for $\mathbb{P}$. Equivalently, we are really saying that if P is a p-variable, then $\mathbb{1}_{\{P \leq r\}}/r$ is an e-variable for any $r \in (0, 1]$; indeed $p \mapsto \mathbb{1}_{\{p \leq r\}}/r$ is a p-to-e calibrator (Section 2.3). In any case, the allowable bets are indexed by r . Now, in this game, the realized p-value equals the smallest value of r at which Skeptic would have received a profit, *had they picked this particular, best, bet in advance*.

The above betting interpretations demonstrates why e-values and reciprocals of p-values should not be compared on the same scale: an e-variable is genuinely a valid bet, but one over a p-variable is not—the latter is effectively “double dipping”. Wouldn’t we all like to have picked the best bet we could have in hindsight? P-values are effectively doing that, but e-values are not. In short, p-values are more akin to hindsight-optimal bets, potentially exaggerating evidence against the null. This fact helps to motivate the grids in Section 2.7.

1.6 A first example: likelihood ratio

Suppose that we are testing a simple null hypothesis $\mathbb{P}$ versus a simple alternative hypothesis $\mathbb{Q}$, assuming $\mathbb{Q} \ll \mathbb{P}$. For this setting, a natural e-variable is the likelihood ratio $E = (d\mathbb{Q}/d\mathbb{P})(X)$ where we treat the likelihood ratio as a function of the observed data X . For any $\mathbb{Q} \ll \mathbb{P}$, the likelihood ratio $d\mathbb{Q}/d\mathbb{P}$ is always an e-variable for $\mathbb{P}$.

We specialize in two simple settings to help the reader to keep concrete examples in mind. In what follows, $X_1, \dots, X_n$ are real-valued iid data points from $\mathbb{P}$ or $\mathbb{Q}$, and we denote by $S_n = \sum_{i=1}^n X_i$ their sum.

The most standard example is testing $\mathbb{P} = N(0, 1)$ against $\mathbb{Q} = N(\mu, 1)$ for a known $\mu > 0$. Rigorously, $\mathbb{P}$ and $\mathbb{Q}$ are the product measure on the data sample $X_1, \dots, X_n$ (which are iid normal), but we slightly abuse the notation here and later, which should be clear from the context. In this setting, the likelihood ratio e-variable is given by

$$E_n = \exp(\mu S_n - n\mu^2/2) = \prod_{i=1}^n \exp(\mu X_i - \mu^2/2). \quad (1.1)$$

It is straightforward to check that $\mathbb{E}^{\mathbb{P}}[E_n] = 1$, and hence the e-variable E_n is exact (in fact, each of its building blocks $\exp(\mu X_i - \mu^2/2)$ is an exact e-variable). This e-variable is not an ad-hoc choice; it has certain optimality that will be formally studied in Chapter 3. Write $\delta = \mu\sqrt{n}$, which measures the strength of the signal under the alternative hypothesis, and $Z = S_n/\sqrt{n}$, which is $N(0, 1)$ -distributed under the null hypothesis. The e-variable in (1.1) can be rewritten as

$$E_n = \exp(\delta Z - \delta^2/2). \quad (1.2)$$

The Neyman–Pearson p-variable is computed as $P_n = \Phi(-S_n/\sqrt{n}) = \Phi(-Z)$, where Φ is the standard normal cdf. Clearly, the e-value is increasing in S_n and the p-value is decreasing in S_n , both confirming the intuition that a larger observed S_n favours $\mathbb{Q}$ over $\mathbb{P}$.

A simple calculation shows that E_n is an e-variable for any choice of $\delta \in \mathbb{R}$ (not just for $\delta = \mu\sqrt{n}$ used earlier). Suppose that we are now testing $N(0, 1)$ against $N(\mu, 1)$ for an unknown $\mu > 0$. We can choose a parameter $\delta > 0$, which affects the power of the e-variable, as discussed in Chapters 2 and 3. We can also test the two-sided alternative hypothesis: $N(\mu, 1)$ for an unknown $\mu \neq 0$. In this case, a simple e-variable is

$$E_n = \frac{\exp(\delta Z - \delta^2/2) + \exp(-\delta Z - \delta^2/2)}{2}, \quad (1.3)$$

for a parameter $\delta > 0$ free to choose. We discuss how to pick δ (in a more general context) in Section 3.7. The corresponding p-variable is given by

$$P_n = 2\Phi(-|Z|). \quad (1.4)$$

Another simple example is testing $\mathbb{P} = \text{Bernoulli}(p)$ against $\mathbb{Q} = \text{Bernoulli}(q)$, where the parameters $p, q \in (0, 1)$ are known. It is without loss of generality to only consider the case $q > p$. In this setting, a natural e-variable for a sample of size $n \in \mathbb{N}$ is given by the likelihood ratio

$$E_n = \left(\frac{q}{p}\right)^{S_n} \left(\frac{1-q}{1-p}\right)^{n-S_n}. \quad (1.5)$$

The Neyman–Pearson p-variable is computed as $P_n = 1 - F_{n,p}(S_n)$ where $F_{n,p}$ is the cdf of the binomial distribution with parameters (n, p) . Again, the e-variable E_n is exact, and noting that $q > p$, the e-value is increasing in S_n and the p-value is decreasing in S_n . The above e-variables $(E_n)_{n \in \mathbb{N}}$ in fact also form e-processes, introduced in Chapter 7.

The point that e-values can be used in unknown sampling scheme, mentioned in Section 1.2.1, is illustrated in the following example, based on the e-value in (1.5).

Example 1.4: Unknown sampling scheme

Consider a simple setting of testing $\mathbb{P} = \text{Bernoulli}(1/2)$ against $\mathbb{Q} = \text{Bernoulli}(q)$, where $q > 1/2$ with n iid data points. Suppose that a scientist is presented with the dataset $(1, 1, 0, 1, 1, 1, 1)$ without explaining how the data collection was designed (e.g., from older experiments). Consider two possibilities: (a) it was designed such that 8 data points are collected; (b) it was designed such that data points are collected until 5 heads in a roll are observed. In case (a), the standard p-value is computed by $\mathbb{P}(N_8 \geq 7) = 0.035$, considered as significant in some areas of science. In case (b), the standard p-value is $\mathbb{P}(\text{observing 5 heads in a roll at or before } n = 8) = 0.078$, often considered as insignificant. The subtle point here is that if only the data are presented to the scientist but not the design (the data collector may not even have a design in mind when collecting data, and decided to stop without a plan), there is no way for the scientist to distinguish between the cases (a), (b), and other possible cases. This is a possible problematic situation in scientific practice. On the other hand, $E_n = q^{S_n}(1-q)^{n-S_n}/2^n$ in (1.5) is a valid e-variable, regardless of how the data collecting procedure is designed as long as it collects all obtained data from the beginning. This is due to the fact that $(E_n)_{n \geq 1}$ is an e-process (formally treated in Chapter 7).

One may keep in mind that not all sampling schemes are allowed for e-processes either. Some obviously problematic situations include that the data were cherry-picked so that only 1s are reported, the data collector removed the first few 0s, or the data collector reported the “best” batch of data after several trials. In these, arguably malicious, situations, any statistical methodology purely based on the collected data fails to work, including those based on e-values.

An illustration

Below we illustrate in a simple case how the e-values from likelihood ratios look like and how they vary with n . Suppose that a scientist is interested in a parameter $\theta^\dagger \in \Theta$, and iid observations $X_1, \dots, X_n$ from a distribution parameterized by $\theta^\dagger$ are available and sequentially revealed to her one by one. She tests $H_0 : \theta^\dagger = 0$ against $H_1 : \theta^\dagger \in \Theta_1$ where $0 \notin \Theta_1 \subseteq \Theta$. Let ℓ be the likelihood ratio function given by

$$\ell(x; \theta) = \frac{d\mathbb{Q}_\theta}{d\mathbb{P}}(x),$$

where $\mathbb{Q}_\theta$ is the probability measure for one observation that corresponds to $\theta \in \Theta$, and $\mathbb{P}$ corresponds to $\theta = 0$. The fundamental observation, explained above, is that $\ell(X_i; \theta)$ for any $\theta \in \Theta$ and $i \in [n]$ is an e-variable for H_0 . Since H_1 is a composite hypothesis, there is some freedom in choosing the parameter θ . Consider the following strategies:

- (a) Fix $\theta_1, \dots, \theta_n \in \Theta_1$, and define the e-variables $E_i := \ell(X_i; \theta_i)$ for $i \in [n]$. One may simply choose all θ_i to be the same. This includes the e-variable in (1.1).

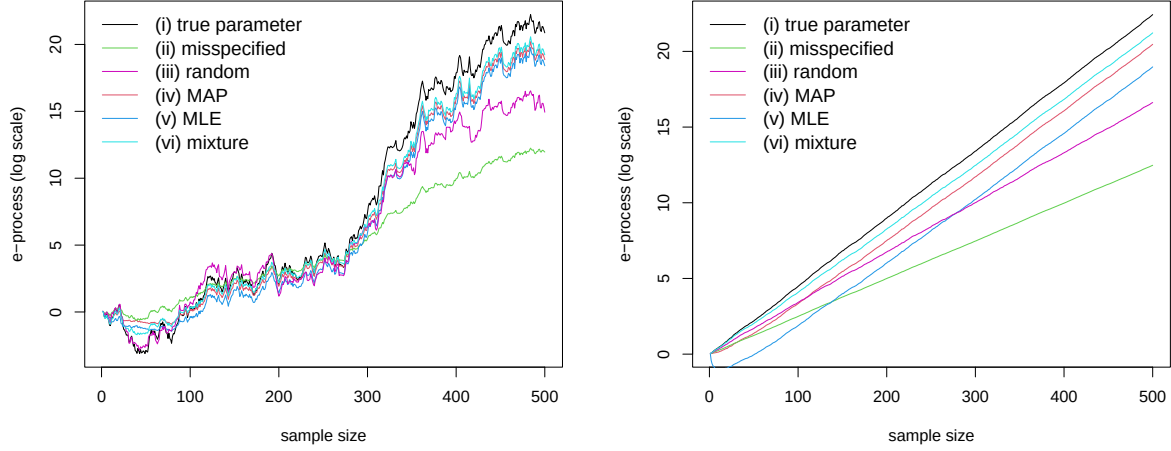


Figure 1.6: A few ways of constructing e-values from likelihood ratio processes. Left: one run. Right: the average of 1000 runs (with average taken on the log values).

- (b) Choose $\theta_1, \dots, \theta_n$ adaptively, where for each k , θ_k is estimated from $(X_1, \dots, X_{k-1})$. This can be obtained, e.g., as a point estimate of θ through maximum likelihood estimation (MLE) or maximum a posteriori (MAP) estimate using some prior.

The e-variables E_i , $i \in [n]$ are combined to form a final e-variable $M_n := \prod_{i=1}^n E_i$. All methods produce a valid final e-variable; indeed, it is straightforward to verify that the running product of $(E_i)_{i \in [n]}$ forms a nonnegative martingale $M = (M_i)_{i \in [n]}$ under $\mathbb{P}$ (called an e-process in Chapter 7), because, with $M_0 = 1$,

$$\mathbb{E}^{\mathbb{P}}[M_i | X_1, \dots, X_{i-1}] = M_{i-1} \mathbb{E}^{\mathbb{P}}[\ell(X_i; \theta_i) | X_1, \dots, X_{i-1}] = M_{i-1}$$

for each $i \in [n]$, as long as θ_i is computed based on $X_1, \dots, X_{i-1}$. A further possible strategy is to take a mixture over M for different strategies above (studied in Chapter 3). The example below gives a concrete illustration.

Example 1.5

We illustrate the processes M obtained from different methods for testing normal distributions. Suppose that an iid sample $(X_1, \dots, X_n)$ from $N(\theta^\dagger, 1)$ is available. The null hypothesis is $\theta^\dagger = 0$, and the alternative is $\theta^\dagger > 0$. We set $\theta^\dagger = 0.3$, which governs the data. We consider five different ways of constructing $E_i := \ell(X_i; \theta_i) = \theta_i X_i - \theta_i^2/2$ for $i \in [n]$, and a further method of mixture of the processes: (i) choose $\theta_i := \theta^\dagger = 0.3$ (knowing the true alternative), which grows optimally (this is called log-optimality; see Chapter 3) (ii) choose $\theta_i := 0.1$, which is a misspecified alternative; (iii) random: take iid θ_i following the uniform distribution on $[0, 0.5]$; (iv) choose θ_i by the MAP estimator with prior $\theta \sim N(0.1, 0.2^2)$; (v) choose θ_i with $\theta_1 := 0.1$ and θ_i the MLE of θ based on $X_1, \dots, X_{i-1}$; (vi) compute a mixture of M each with a fixed $\theta_i \in [0, 0.6]$, uniformly weighted via a discrete grid.

The resulting M curves are plotted in the logarithmic scale in Figure 1.6 with $n = 500$. For sufficiently large sample sizes, the methods (iv) and (v) based on adaptively choosing θ_i and the method (vi) based on the mixture are more powerful than (ii) and (iii) based on misspecified or random alternatives, and they have asymptotically the same slope (in the log-scale) as the best performer (i); these will be theoretically justified in Chapters 3 and 7. The scientist may reject the null hypothesis based on a large value of M_n (see Chapter 2), but she can also reject the null hypothesis based on the maximum of M or a stopping time applied to it (formally studied in Chapter 7).

1.7 A second example: the soft-rank e-value

There is a large class of classical and modern testing procedures that use some form of Monte Carlo sampling in order to produce test statistics that are exchangeable under the null, and use the rank of the original test statistic as a corresponding p-value. Below we give an abstract problem setup, then an explicit example—a permutation test for independence—and then describe how to build a natural e-value in this context.

Consider testing some hypothesis $\mathcal{P}$ and let L_0 be the test statistic calculated from the original data (a larger value of L_0 indicates evidence against the null). Further, suppose we can calculate B new statistics $L_1, \dots, L_B$, that are constructed to be exchangeable together with L_0 under the null $\mathcal{P}$. Recall that random variables $L_0, \dots, L_B$ are called *exchangeable* if the joint distribution of $(L_0, \dots, L_B)$ equals that of $(L_{\sigma(0)}, \dots, L_{\sigma(B)})$ for any permutation σ of $\{0, \dots, B\}$.

For example, suppose that our observed data consist of pairs (X_i, Y_i) , $i \in [n]$, which are assumed to be iid from some joint distribution. Consider testing whether X is independent of Y ($H_0 : X \perp Y$; $H_1 : X \not\perp Y$). We can reframe this as testing whether $X_1, \dots, X_n$ are iid conditional on $Y_1, \dots, Y_n$. Consider a simple test statistic defined as the sample correlation between $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$, denoted by L_0 . In order to see whether the sample correlation is sufficiently large to be statistically significant, we would compare against the sample correlation L_σ computed on the permuted data, $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ and $(Y_1, \dots, Y_n)$. We draw B such independent random permutations $\sigma_1, \dots, \sigma_B$, and the corresponding sample correlations are denoted by $L_1, \dots, L_B$. The key property is that under the null hypothesis, $L_0, L_1, \dots, L_B$ are exchangeable.

Returning to our abstract formulation in order to build an e-value, we transform the statistics $L_0, L_1, \dots, L_B$ into some nonnegative scores $R_0, R_1, \dots, R_B$ while preserving their order and exchangeability. An explicit transformation is given in Example 1.7 below. Define

$$E = \frac{(B+1)R_0}{\sum_{b=0}^B R_b},$$

where we define $E = 1$ if $\sum_{b=0}^B R_b = 0$. The fact that E is an e-variable is guaranteed by $\mathbb{E}[R_0 | \sum_{b=0}^B R_b] = \sum_{b=0}^B R_b / (B+1)$ for $\mathbb{P} \in \mathcal{P}$ due to the assumed exchangeability of $R_0, R_1, \dots, R_B$ under $\mathcal{P}$. We will later see in Section 6.7.7 that the *log-optimal* e-variable against a simple alternative (called the numeraire) also takes the above form.

Example 1.7

Let $r > 0$ be a prespecified constant. Define $L_* = \min_{b=0, \dots, B} L_b$ to be the smallest of the $(B+1)$ test statistics. For $b = 0, \dots, B$, define the transformed statistic

$$R_b = \frac{\exp(rL_b) - \exp(rL_*)}{r}.$$

This transformation is performed to ensure that R_b is nonnegative while the ordering amongst the test statistics is preserved. Note that L_* can be replaced by a constant lower bound on L_b . We also allow the limiting case of $r = 0$, which yields $R_b = L_b - L_* \geq 0$. If $L_1, \dots, L_B$ are assumed nonnegative, we can choose $R_b = L_b$, i.e., using 0 as the lower bound with $r = 0$; all of these choices guarantee exchangeability and nonnegativity of $\{R_b\}_{b=0}^B$.

Contrasting E with the usual p-variable P defined by

$$P = \frac{\sum_{b=0}^B \mathbf{1}_{\{L_b \geq L_0\}}}{B+1},$$

we find that

$$P = \frac{\sum_{b=0}^B \mathbf{1}_{\{R_b \geq R_0\}}}{B+1} \leq \frac{\sum_{b=0}^B R_b / R_0}{B+1} = 1/E.$$

(If $R_0 = 0$ and $R_b > 0$ for some b , then the above $1/E$ is interpreted as ∞ .) Since P quantifies the rank of L_0 amongst $L_0, \dots, L_B$, we see that $1/E$ can be seen as a smoothed notion of rank, or “soft-rank” for short

(much like the “soft-max” is achieved by exponentiation in Example 1.7). Hence, we call E as the soft-rank e-value and $1/E$ as the soft-rank p-value. Since the direct p-value P is always smaller than the soft-rank p-value $1/E$, there is apparently no advantage to using the latter for testing a single hypothesis with the same threshold α . Nevertheless, later in Chapters 8–11, we will see that e-values enjoy several advantages over p-values in the contexts of multiple testing and constructing confidence intervals.

1.8 Historical context

This section summarizes our historical view on hypothesis testing with e-values for the interested reader.

1.8.1 Who invented e-values?

In some sense, the question of who invented (or, discovered?) the concept of an e-value or e-variable is impossible to answer. As already mentioned in brief at the start of the chapter, and as we will discuss in later chapters, if $\mathcal{P} = \{\mathbb{P}\}$ is a singleton, then all optimal e-values take the form of $d\mathbb{Q}/d\mathbb{P}$, that is likelihood ratios of $\mathbb{Q}$ against $\mathbb{P}$, for some (implicit or explicit) alternative $\mathbb{Q}$. So, technically e-values have been around for 100 years, in the form of likelihood ratios. Similarly, for simple nulls and composite alternatives, Bayes factors are also likelihood ratios, and hence e-values. Likelihood ratios are central objects in parametric and nonparametric statistics, in both frequentist and Bayesian as well as other paradigms. Therefore, in some sense, e-values have always been everywhere in modern statistics. But this is only a part of the story.

The recent coining and usage of the term “e-value” is simply to recognize the importance of a more general concept that has utility much beyond point/simple nulls. Indeed, beyond the case of simple hypotheses, e-values can be viewed as nonparametric/composite generalizations of likelihood ratios to complex settings involving nonparametric and composite nulls and alternatives. Frequentists often use the generalized likelihood ratio (the ratio between the maximum likelihoods over the alternative and the null), which is not an e-value. Bayesians (or at least those who consider testing hypotheses) typically generalize Bayes factors by using the ratio of mixture likelihoods, picking a different distribution over the alternative and the null, and this is also not always an e-value. As we will later see, constructing e-values for composite nulls and alternatives is a key and central topic in the literature, and we dedicate two chapters to it in this book: Chapters 5 and 6. Informally, the former chapter points out that the ratio of the mixture likelihood over alternative to the maximum likelihood over the null (a curious mixture of frequentist and Bayesian approaches) is always an e-variable. The latter chapter effectively (if only implicitly) points out that sometimes Bayes factors are indeed e-variables, if the mixture distributions are chosen in a very particular manner.

There are perhaps five central authors who we would like to centrally credit for the development of e-values, either explicitly or implicitly through their work on sequential testing, on betting, on the very related theory of nonnegative (super)martingales, or on the theory of log-optimality. These are Ville, Wald, Kelly, Robbins and Cover (in that historical order). They each had fundamental works that are intimately connected to the topics in this book, even if our motivation, treatment, and usage may substantially differ from all of them.

In its modern avatar, there have been several other central contributors to the philosophy, theory, methodology, and applications of e-values, and more generally to the subfield now called *game-theoretic statistics and sequential anytime-valid inference (SAVI)*: Vovk, Shafer, Grünwald, and if we may say so, the authors of this book. It was in the 2018-20 period that several independent works by these authors (and their collaborators) were posted to arXiv, each with highly aligned motivation, philosophy and techniques, but using different terminology. This instigated a joint videoconferencing call in 2020 where these 5 authors jointly decided on using the term “e-value” whenever it is meant in a measure-theoretic context (as in this book). Of course, all of these authors have had many collaborators who have made key contributions, and they will be appropriately cited in place.

Research on e-values has blossomed recently, often without acknowledgment of its roots. We hope that the above historical context, and the bibliographical accompaniments below are helpful for the reader.

This particular book will focus mainly on the non-sequential context, primarily because there is already

so much to say, but also because there is much more basic work to be done in the sequential context before a book can be written about it. Needless to say, such a book will be written in due course. Nevertheless, Chapter 7 is dedicated to the sequential setting, giving readers a glimpse of the broader framework.

1.8.2 Ville, Wald, Kelly, Robbins, Cover, and Vovk

We now discuss some of the key contributions of the aforementioned authors within the context of e-values and game-theoretic statistics (and its SAVI applications). The references are not intended to be comprehensive, but are often just exemplary of an influential line of work. We hope they serve as starting points for the reader who wishes to dig more deeply.

Ville brought martingales to the front and center of modern measure-theoretic probability theory [Ville, 1939]. He proved that for any event of measure zero, there exists a nonnegative martingale (with respect to that measure) which explodes to infinity on that event. Ville designed betting strategies that test the law of large numbers or the law of the iterated logarithm. We will encounter some of Ville’s contributions in this book. While our literature is certainly inspired by Ville, he was not actually motivated by statistical inference (either testing or estimation), issues around optional stopping and continuation, and so on. As the title of his thesis suggests, he was interested in a completely different goal: pointing out some basic flaws in Richard von Mises’ theory of collectives. The book chapter by Shafer [2022] provides a much broader historical context for his work.

Wald invented sequential analysis [Wald, 1945]. Nonnegative martingales, in particular the likelihood ratio process, play a central role in Wald’s methodology and theory (though he originally did not use that terminology, since the modern theory of martingales was also being developed in parallel to his work by Doob, who was aware of Ville’s work). Wald emphasized his sequential test, specified in terms of a single stopping time, that optimally trades off type-I and type-II errors. In contrast, we emphasize the underlying process (the nonnegative supermartingale, or more generally, the e-process), and focus on guarantees that hold at *any* stopping time, possibly not defined or anticipated in advance.

Kelly proposed the key notion of log-optimality [Kelly, 1956]. A few years after Shannon had developed information theory, Kelly’s paper pointed out a fundamental connection between gambling and information theory. Kelly pointed out that when playing a repeated game with binary outcomes and favorable odds, it is possible to bet in such a way that your wealth grows exponentially fast, and thus it is natural to wish to optimize the rate of growth of that wealth. This immediately leads to the criterion of wanting to optimize the expected logarithm of the wealth, which we call log-optimality in this book. Breiman [1961] made a key contribution in generalizing Kelly’s work beyond the binary case, and pointing out that the same criterion also asymptotically optimizes two other related criteria: minimizing the expected time needed to reach a threshold wealth (as the threshold goes to infinity), and maximizing the expected limiting wealth (as the time goes to infinity); the latter is often called *competitive optimality*. While their works were not written in terms of statistical hypothesis testing, we borrow their intuition when setting up the “testing by betting” framework. There, a testing problem is transformed into a game with unfavorable odds under the null but favorable odds under the alternative, and thus a natural goal becomes to aim for log-optimality under the alternative.

Robbins defined the fundamental concept of confidence sequences [Darling and Robbins, 1967], the natural sequential extension of confidence intervals, and developed its duality to “power-one” tests Robbins [1970]. He recognized and utilized that the product of (independent or sequentially dependent) e-values yields nonnegative supermartingales (despite not using the term “e-value”), and constructed these in some nonparametric settings. Robbins’ philosophical motivations were most directly aligned with ours, in that he moved away from Wald’s design of a single stopping rule, to the design of anytime-valid tests and confidence sequences. However, there are still some subtle philosophical differences. Robbins focused on level- α concepts, whether tests or confidence sequences, while our theory emphasizes the underlying objects (e-values and e-processes), which by themselves are measures of evidence, even if they are not explicitly thresholded to Robbins’ tools.

Finally, Cover developed the framework of log-optimal portfolios for financial trading [Cover, 1984]. Without making distributional assumptions, he developed regret bounds that compare the log-wealth of a gambler that can adaptively rebalance their portfolio in every round (reassign their current wealth to the set

of available stocks) to the best constant rebalanced portfolio in hindsight (in turn which offers better returns than the best stock in hindsight). Cover’s work does not show up as centrally as the others in this book, but this is because we only provide glimpses of the sequential setting.

Ville, Wald, Robbins and Cover all implicitly or explicitly employed mixture or plug-in betting strategies (or likelihood ratios, or nonnegative supermartingales). The plug-in principle and the method of mixtures will be our standard go-to methods for handling composite alternatives in this work, and they can both be traced back to ideas presented by all of these authors.

Vovk spearheaded the modern development of these ideas. The core ideas underlying what is now known as *game-theoretic probability* were presented in Vovk [1993]. Vovk and V’yugin [1993] defined and used essentially e-values in the context of algorithmic complexity and randomness (though in this context e-values had been introduced earlier by Levin and Gacs, which can then be traced back to Martin-Löf and then to Ville). Gammernan, Vovk, and Vapnik [1998] had an explicit instance of an e-value for the composite null hypothesis of testing exchangeability in the context of conformal prediction. Following that, Shafer and Vovk [2001] wrote a book on viewing probability and finance game-theoretically. Many central ideas for game-theoretic statistics were contained in Shafer, Shen, Vereshchagin, and Vovk [2011].

1.8.3 Recent progress on e-values

As mentioned earlier, the current resurgence of interest in this topic can be attributed to a series of works that were first made public the 2018–2020 period. We discuss a dozen such works in their order of appearance (typically on arXiv) below. We also omit a dozen of our own papers that appeared during this period, as well as several papers by other authors. This has resulted in an incomplete and biased list. But we still think that this is a list that will retrospectively be viewed as containing works that developed key ideas in this space in a short time span of around two years.

- In Aug–Oct 2018, Howard, Ramdas, McAuliffe, and Sekhon [2020, 2021] identified one of the definitions of an e-process for composite nulls: a nonnegative process that is upper bounded by a family of supermartingales. They combined these with Ville’s inequality to unify and improve a host of nonparametric Chernoff-style concentration inequalities. They developed several variants of the method which yielded computationally and statistically efficient nonparametric sequential tests and confidence sequences, extending and generalizing Robbins’ extensive work on the topic.
- In Nov 2018, Kaufmann and Koolen [2021] also used Robbins’ method of mixtures applied to likelihood ratios arising from one-parameter exponential families. They then also used Ville’s inequality to derive time-uniform concentration bounds for the Kullback–Leibler divergence (relative entropy) between the distributions parameterized by the empirical and true means. This yielded sequential tests and confidence sequences with applications to parametric multi-armed bandit problems.
- In Mar 2019, Shafer [2021] wrote an expository article promoting the use of betting in scientific communication and arguing for the adoption of the log-optimality criterion. The paper thoroughly develops the story for testing a simple null against a simple alternative, proving that the log-optimal bet is their likelihood ratio.
- In Mar 2019, Shafer and Vovk [2019] published a book on the game-theoretic foundations of probability and finance, providing a solid philosophical and mathematical basis for these fields. Many ideas in game-theoretic statistics owe their philosophical or methodological roots to ideas presented in this book. Certainly, this book predates all papers mentioned here as Mar 2019 is its publication time.
- In Jun 2019, Grünwald, De Heide, and Koolen [2024a] identified what we now recognize as the simpler of the two definitions of an e-process for composite nulls: a nonnegative process that is an e-value at any stopping time. They also argued for the log-optimality criterion, pointed out the key role played by the reverse information projection in the design of log-optimal e-variables, and argued that meta-analysis (of an increasing number of studies) is a particularly sensible application area for these concepts.

- In Dec 2019, [Vovk and Wang \[2021\]](#) developed fundamental results related to combining e-values, and transforming e-values to p-values or vice versa (calibrators). For instance, they prove that the arithmetic average is essentially the only admissible symmetric function that can combine dependent e-values, and that inverting an e-value is the only admissible way to convert it to a p-value. Before this paper, authors used different names for the same concept, but following this paper, there was a joint agreement amongst authors to use the term “e-value”.
- In Dec 2019, [Wasserman, Ramdas, and Balakrishnan \[2020\]](#) developed particularly simple and broadly applicable e-variables and e-processes within their framework of “universal inference”. They proved that whenever the maximum likelihood under the (composite) null hypothesis can be calculated, an e-variable can be constructed, leading to a valid (fixed-time or sequential) test by Markov’s inequality.
- In Jul 2020, [Vovk, Wang, and Wang \[2022\]](#) showed that admissible ways of merging p-values under arbitrary dependence have to go through e-values. This result is based on a classic optimal transport duality that converts probability constraints under arbitrary dependence into expectation constraints. Using this connection, many admissible methods of merging p-values are constructed from merging e-values.
- In Sep 2020, [Wang and Ramdas \[2022\]](#) developed an analog of the Benjamini–Hochberg (BH) procedure that employed e-values instead of p-values (e-BH). They showed that unlike BH, the e-BH procedure controlled the false discovery rate under arbitrary dependence between the e-values.
- In Sep 2020, [Ramdas, Ruf, Larsson, and Koolen \[2020\]](#) identified necessary and sufficient conditions for admissibility for SAVI tools (e-processes, p-processes, confidence sequences, sequential tests), and prove that all admissible SAVI methods must be based on nonnegative martingales. They also prove that the two definitions of e-processes mentioned above are actually equivalent.
- In Oct 2020, [Waudby-Smith and Ramdas \[2024\]](#) employed a nonparametric testing by betting approach to produce state-of-the-art solutions to the classic problem of estimating a bounded mean (when sampling with or without replacement). In doing so, they provide a universal representation for all composite nonnegative martingales, and a simple betting strategy that easily extends to other problems.

The pace and quality of progress did not stop here, and this book aims to collect many of the results presented in these above works, and many others that followed them. The bibliographic notes after each chapter mention many relevant papers in addition to the above, but they are by no means exhaustive.

1.8.4 Unfortunate name clashes: other e-values in the literature

Given the popularity of p-values, many other terms have been created in the literature by swapping the p in p-value for other letters. All the above authors had already repeatedly contemplated other options, such as the s-value (s for safe), v-value (v to honor Ville and Vovk), b-value (b for bet), m-value (m for martingale), and i-value (i for integral). In fact, the name i-value was used informally by a few individuals in the 1990s to describe a concept that is essentially an e-value in a different context, where the constraint is an integral (more general than an expectation) bounded by 1.

In the end, we as a community stuck with e-value, because it was defined by its expectation, and is directly interpreted as evidence against the null. There was also a consideration of its pronunciation and contrast to p-values. However, as time has passed, we have discovered many other e-values in the literature that we were initially unaware of. We mention them below to avoid confusions for the casual reader, who may search for the term e-value and may mistakenly wander into a different literature with other goals and definitions. We apologize for any unintentional misrepresentations of their concepts, on which we are not experts.

Importantly, terms like “e-variable” and “e-process” do not have known clashes, and the reader is reassured when these terms accompany “e-value”, so we recommend mentioning them in papers wherever suitable.

1. **BLAST E-value.** BLAST [\[Altschul et al., 1990\]](#) is an extremely popular software tool used in bioinformatics. In the context of sequence alignment given a database of sequences, the BLAST

E-value (E for Expect) is the number of expected hits of similar quality (score) that could be found just by chance. Their E-value appears to be a particular calibrator applied to a particular p-value: $E = -\log(1 - P)$. This appears to keep E-values and p-values on the same scale (lower E-values point towards statistical significance). This is related to Greenland’s definition of an S-value (S for surprisal or Shannon information), which is $-\log(P)$, which would be a special case of our e-value.

2. **Sensitivity E-value.** VanderWeele and Ding [2017] introduced E-values in the particular context of sensitivity analysis in causal inference. The E-value is defined as the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates. A large E-value implies that considerable unmeasured confounding would be needed to explain away an effect estimate. A small E-value implies little unmeasured confounding would be needed to explain away an effect estimate. This notion has become extremely popular in applied causal inference.
3. **Bayesian E-value.** The e-value $ev(H|X)$ — the *epistemic*-value of hypothesis H given observations X , or the evidence-value of observations X in favor (or in support) of hypothesis H — is a Bayesian statistical significance measure introduced together with the Full Bayesian Significance Test [Pereira and Stern, 1999]. Concise entries about the e-value and the FBST are available at the International Encyclopedia of Statistical Science and online at Wiley’s StatsRef. The recent survey by Stern et al. [2022] points to rich and growing literature on the topic over the last 25 years.

1.9 Road map

The rest of this book is organized into three parts. The first part, **Fundamental Concepts**, includes four chapters (including the current one) that introduce the basic concepts.

- Chapter 2 concerns the properties of e-values under the null hypothesis, including key concepts such as Markov’s inequality, calibrators, and randomized tests. It presents tools to convert e-values to p-values and tests, and vice versa. It also presents explicit e-variables in several simple examples. This chapter focuses on *validity* (under the null), leaving questions of *statistical efficiency* to the following chapter.
- Chapter 3 concerns the properties of e-values under the alternative hypothesis, including topics of e-power, log-optimality, existence of powered e-values, and mixture and plug-in methods. As an important result, the log-optimal e-variable for a simple null against a simple alternative is their likelihood ratio. It provides methods to derive “good” or “useful” e-values, and shows how e-values, p-values and tests are fundamentally connected via a unified existence result.
- Chapter 4 demonstrates that e-values play a fundamental role in post-hoc testing and decision making, where some parts of the testing problem, available decisions, or data collection depend on the current data. While the previous chapters focused on putting e-values and p-values on a similar footing, this is the first chapter to clearly separate their utility: what one can do with e-values versus what one can (or cannot) do with p-values.

The second part, **Core Ideas**, includes five chapters.

- Chapter 5 introduces the general method of universal inference for constructing e-variables for general composite nulls and alternatives. This method is very simple to apply, and provides the first known test for many “irregular” testing problems. However, it is inadmissible in a strong sense: it is always dominated by the e-variable constructed in the next chapter.
- Chapter 6 contains a key result: for a given composite null and simple alternative, there always exists a unique log-optimal e-variable (called the numeraire), and presents a strong duality result with the reverse information projection (RIPr) of that alternative onto the null. Remarkably, the numeraire is also the only e-variable that is a likelihood ratio between the alternative and some element of the (effective) null.

- Chapter 7 presents the framework of sequential anytime-valid inference (SAVI), emphasizing key concepts such as e-processes, testing by betting, optional stopping and Ville’s inequality. It also discusses key differences between SAVI and Wald’s sequential testing paradigm, as well as the problematic practice of peeking at p-values.
- Chapter 8 discusses merging e-values via averages (arbitrary dependence) or products (independence or “sequential” dependence), as well as the trade-offs in choosing merging methods.
- Chapter 9 presents the notion of compound e-values in the context of multiple hypothesis testing, introduces the e-Benjamini–Hochberg (e-BH) procedure and proves that every FDR controlling procedure can be seen as applying e-BH to some set of compound e-values.

The third part, **Advanced Topics**, has seven chapters.

- Chapter 10 studies approximate and asymptotic p-values and e-values, as well as their counterparts in multiple testing. This chapter provides a bridge between classical statistics, which often relies on asymptotics, and most of our book, which is nonasymptotic.
- Chapter 11 shows that the only admissible way to combine dependent e-values is by (weighted) averaging. It also develops methods to combine a p-value and an e-value, which give rise to an e-weighted BH procedure.
- Chapter 12 derives methods for combining multiple dependent p-values, which are proved to rely on e-values through optimal transport duality, and also ways to improve on such combination rules by using random and exchangeable improvements of Markov’s inequality.
- Chapter 13 defines e-confidence intervals (e-CI), and presents the e-Benjamini–Yekutieli (e-BY) procedure to control the false coverage rate after selecting a subset of e-CIs to report. It also discusses how to merge dependent confidence intervals using (various versions of) majority vote.
- Chapter 14 offers further contrasts between p-values and e-values, including a duality between them, a representation of e-variables and p-variables as conditional expectations and probabilities, and an existence result on powered exact e-values and p-values.
- Chapter 15 shows how Markov’s inequality can be improved under additional conditions on the shape of the distribution of a single e-variable or on the dependence between multiple e-variables.
- Chapter 16 studies the connection between e-values and risk measures, and offers some nonparametric methods for testing the forecast of general statistical functions (such as the mean, the variance, or a quantile) using e-values.

In terms of the goal of each chapter, all chapters can be grouped in the following three categories.

1. Understanding the foundation of e-values: Chapters 2, 4, 10, and 14.
2. Constructing good and powerful e-values in single hypothesis testing: Chapters 3, 6, and 15.
3. Designing inference methods with e-values: Chapters 5, 7, and 16.
4. Designing methods for multiple testing with e-values: Chapters 8, 9, 11, 12, and 13.

In the first two categories, we are interested in e-values themselves, whereas in the remaining categories, we are primarily interested in the specific statistical problem.

Chapter 2

Validity: E-values under the null hypothesis

In this chapter, we study fundamental issues of e-values, in particular, on their connection to p-values, validity, and tests. These properties concern the behavior of the e-variables under the null hypothesis $\mathcal{P}$. For this reason, the alternative hypothesis $\mathcal{Q}$ does not appear and plays no role.

2.1 Testing using Markov's inequality

As mentioned before, e-values can be interpreted in their own right as evidence against the null hypothesis without resort to other concepts. However, the reader may naturally be curious how exactly e-values may be transformed (perhaps with some loss of efficiency) into the more classical concepts of level- α tests and p-values. It is in this context that a central role will be played by Markov's inequality (though it apparently was discovered by Chebyshev, Markov's advisor).

Markov's inequality guarantees that a test rejecting an e-value larger than $1/\alpha$ yields a level- α test.

Proposition 2.1: Markov's inequality for e-values

Let E be an e-variable for $\mathcal{P}$. We have $\mathbb{P}(E \geq 1/\alpha) \leq \alpha$ for all $\mathbb{P} \in \mathcal{P}$ and $\alpha \in (0, 1]$. Hence, $1/E$ is a p-variable, $(\alpha E) \wedge 1$ is a level- α test, and $\mathbb{1}_{\{E \geq 1/\alpha\}}$ is a level- α binary test.

The proof follows directly from the simple inequality $\mathbb{1}_{\{x \geq 1\}} \leq x$ for all $x \geq 0$, by observing that

$$\mathbb{P}(E \geq 1/\alpha) = \mathbb{E}^{\mathbb{P}}[\mathbb{1}_{\{E \geq 1/\alpha\}}] \leq \alpha \mathbb{E}^{\mathbb{P}}[E] \leq \alpha.$$

For an e-variable E and a p-variable P , even though $1/E$ is a p-variable, $1/P$ is generally not an e-variable; in fact $1/P$ has infinite mean if P is exact. The p-variable $1/E$ is conservative since it cannot be uniformly distributed on $[0, 1]$, and it is typically far away from being so. As a consequence, the type-I error of the test $\mathbb{1}_{\{E \geq 1/\alpha\}}$ for $\mathbb{P}$ is usually smaller than α unless E is distributed on the two points 0 and $1/\alpha$ with mean 1 under $\mathbb{P}$.

Proposition 2.1 can be generalized in several different directions relevant for testing with e-values; some are presented in Section 2.4 and Chapters 5 and 15. A sequential generalization, called Ville's inequality, is presented in Chapter 7.

With a p-variable P and an e-variable E for the same hypothesis, the tests $\mathbb{1}_{\{E \geq 1/\alpha\}}$ and $\mathbb{1}_{\{P \leq \alpha\}}$ are in general different, except for the special case $E = \mathbb{1}_{\{P \leq \alpha\}}/\alpha$ (see Section 3.1 for more details on such e-variables). To understand how these two tests differ, we consider the two-sided normal test in Section 1.6, where a family of e-variables E_n is given by (1.3), indexed by $\delta > 0$, and the p-variable P_n is given by (1.4). We treat both the e-variables and the p-variable as functions of the value x taken by the statistic Z . Figure 2.2

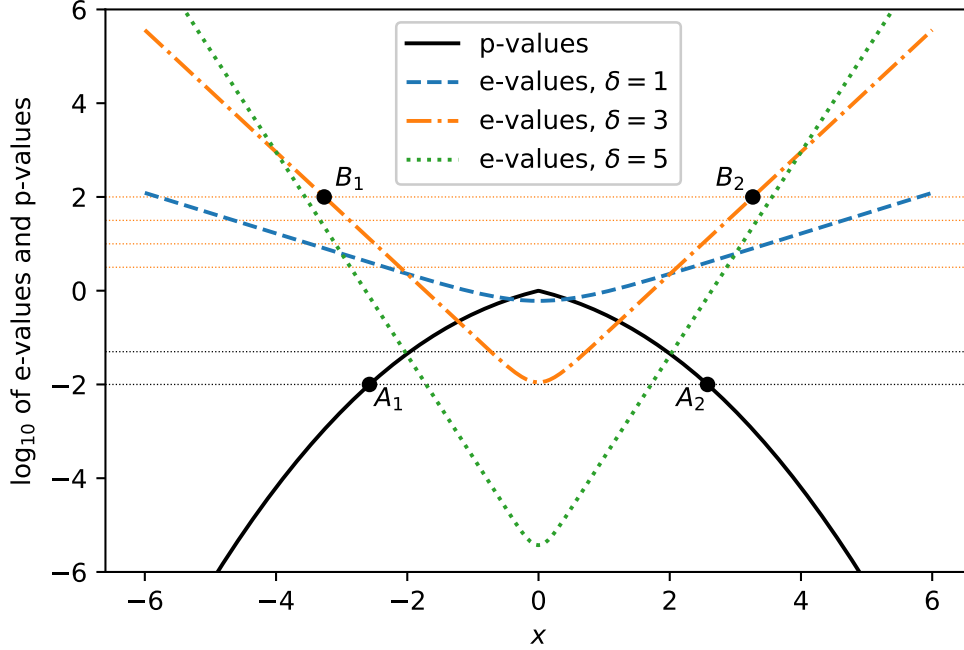


Figure 2.2: A comparison of e-values and p-values for the two-sided normal test in Section 1.6. The curves represent the e-values or p-values as a function of the test statistic $Z = x$. The top four horizontal dotted lines correspond to 10^β with $\beta \in \{2, 1.5, 1, 0.5\}$ for e-values. The bottom two horizontal dotted lines correspond to $\alpha \in \{0.05, 0.01\}$ for p-values. (The reason why e-values are compared with levels 10^β with $\beta \in \{2, 1.5, 1, 0.5\}$ is explained in Section 2.7.)

illustrates two intervals $[A_1, A_2]$ and $[B_1, B_2]$, which correspond to $P_n(x) = 0.01$ for $x = A_1$ or $x = A_2$ and $E_n(x) = 100$ for $x = B_1$ or $x = B_2$, with $\delta = 3$. Here, $A_2 = |A_1| \approx 2.58$ and $B_2 = |B_1| \approx 3.27$. Other choices of δ lead to different intervals, as shown in Figure 2.2. We can see that it is harder for an e-value to reach 100 compared to a p-value to reach 0.01, and this is because the Markov inequality is generally conservative. It is generally unfair to directly compare the values of E and $1/P$. See Section 2.7 for more discussions on this point.

2.2 Testing using distributional knowledge

Proposition 2.1 implies that for an e-variable E and $\alpha \in (0, 1)$, we can always reject the null hypothesis by using the level- α test $\mathbb{1}_{\{E \geq 1/\alpha\}}$. If the distribution of E under the null hypothesis is known or computable (there may be multiple distributions if the null is composite), then we can find thresholds that are smaller than $1/\alpha$ to test based on E . Denote by $S(\cdot; \mathbb{P})$ the left-continuous survival function of E under $\mathbb{P} \in \mathcal{P}$, that is,

$$S(x; \mathbb{P}) = \mathbb{P}(E \geq x)$$

for $x \in \mathbb{R}$. If $S(\cdot; \mathbb{P})$ is the same across $\mathbb{P} \in \mathcal{P}$, then E is called pivotal (this will appear again in Section 14.3). The function $S(\cdot; \mathbb{P})$ can be continuous or discontinuous, depending on both E and $\mathbb{P}$.

For a threshold $t > 0$, we would like the test $\mathbb{1}_{\{E \geq t\}}$ to have level α . Clearly, choosing $t \geq 1/\alpha$ is sufficient, but we can do better. Indeed, by definition, as long as $\sup_{\mathbb{P} \in \mathcal{P}} S(t|\mathbb{P}) \leq \alpha$, t is a valid threshold for the test $\mathbb{1}_{\{E \geq t\}}$ with level α . If $\sup_{\mathbb{P} \in \mathcal{P}} S(t|\mathbb{P})$ is continuous, then the smallest t that satisfies this condition is

$$t^* = \inf \left\{ t > 0 : \sup_{\mathbb{P} \in \mathcal{P}} S(t|\mathbb{P}) \leq \alpha \right\},$$

which is also equal to the largest left α -quantile of E over $\mathbb{P} \in \mathcal{P}$ (Lemma 15.1 in Section 15.1). In particular, if $\mathcal{P} = \{\mathbb{P}\}$ is simple, then we can easily see that

$$t^* = \inf \{t > 0 : \mathbb{P}(E \geq t) \leq \alpha\} = \inf \{t > 0 : \mathbb{P}(E > t) \leq \alpha\},$$

which is the left α -quantile of E under $\mathbb{P}$. The level- α guarantee of $\mathbb{1}_{\{E \geq t^*\}}$ requires $t \mapsto \sup_{\mathbb{P} \in \mathcal{P}} S(t|\mathbb{P})$ to be continuous at t^* ; otherwise we can use the test $\mathbb{1}_{\{E > t^*\}}$, or a suitable randomization between $\mathbb{1}_{\{E > t^*\}}$ and $\mathbb{1}_{\{E \geq t^*\}}$ (analogously to the Neyman–Pearson lemma).

Reducing the threshold in the above manner is in fact no different from treating the e-value as any other test statistic and picking the threshold appropriately to derive a level- α test, as is standard in classical statistics.

Even in situations where $S(\cdot; \mathbb{P})$ is difficult to fully specify or compute, bounds on t^* may be obtained based on partial information of the distribution of the e-value. Several concrete cases will be studied in Chapter 15.

2.3 Relating p-values and e-values using calibrators

P-values and e-values can be converted between each other, via calibrators. Below, when we say “any hypothesis”, we mean any specification of the sample space $(\Omega, \mathcal{F})$ and hypothesis $\mathcal{P}$ on the space.

Definition 2.3: Calibrators

- (i) A *p-to-e calibrator* is a decreasing function $f : [0, \infty) \rightarrow [0, \infty]$ satisfying $f = 0$ on $(1, \infty)$ such that for any hypothesis, $f(P)$ is an e-variable for any p-variable P .
- (ii) An *e-to-p calibrator* is a decreasing function $g : [0, \infty] \rightarrow [0, \infty)$ such that for any hypothesis, $g(E)$ is a p-variable for any e-variable E .
- (iii) A calibrator f is said to *dominate* a calibrator g if $f \geq g$ (p-to-e) or $f \leq g$ (e-to-p), and the domination is *strict* if further $f \neq g$. A calibrator is *admissible* if it is not strictly dominated by any other calibrator.

We often omit “p-to-e” in “p-to-e calibrators” and simply call them calibrators, for a reason that will be made clear soon in Propositions 2.4 and 2.5.

The interval $[0, \infty)$ can be safely replaced by $[0, 1]$ in the above definitions since p-values with values larger than 1 are not interesting for testing.

A probability measure $\mathbb{P}$ (or the corresponding probability space) is called *atomless* if there exists a continuously distributed random variable (e.g., uniform, normal) under $\mathbb{P}$. Hence, when we encounter a continuously distributed random variable, implicitly the probability measure is assumed atomless. In Definition 2.3, the property of calibrators needs to hold for any hypothesis. Nevertheless, it suffices to consider one atomless probability measure $\mathbb{P}$ on some $(\Omega, \mathcal{F})$, because if for the simple hypothesis $\mathbb{P}$, $f(P)$ is an e-variable for any p-variable P , then it holds also for any other, simple or composite, hypotheses. This observation will be useful in several places throughout. A formal justification of this simplification is presented in Appendix A.1.

In the next two results we characterize all admissible calibrators in both directions. We first look at the simpler direction. The following proposition says that there is, essentially, only one e-to-p calibrator, $f : t \mapsto \min(1, 1/t)$, that is offered by Markov’s inequality.

Proposition 2.4

The function $f : [0, \infty] \rightarrow [0, 1]$ defined by $f(t) = \min(1, 1/t)$ is an e-to-p calibrator. It dominates every other e-to-p calibrator. In particular, it is the only admissible e-to-p calibrator.

Proof.

The fact that $t \mapsto \min(1, 1/t)$ is an e-to-p calibrator follows from Markov's inequality in Proposition 2.1. On the other hand, suppose that g is another e-to-p calibrator. It suffices to check that g is dominated by $t \mapsto \min(1, 1/t)$. Suppose $g(t) < \min(1, 1/t)$ for some $t \in [0, \infty]$. Fix an atomless probability measure $\mathbb{P}$. Consider two cases:

- (a) If $g(t) < \min(1, 1/t) = 1/t$ for some $t > 1$, fix such t and consider an e-variable E that is equal to t with probability $1/t$ and 0 otherwise under $\mathbb{P}$. Then $g(E)$ is $g(t) < 1/t$ with probability $1/t$, whereas it would have satisfied $\mathbb{P}(g(E) \leq g(t)) \leq g(t) < 1/t$ had it been a p-variable.
- (b) If $g(t) < \min(1, 1/t) = 1$ for some $t \in [0, 1]$, fix such t and consider the e-variable $E = t$. Then $g(E) = g(t) < 1$, and so it is not a p-variable.

Instead of working with the set of all e-variables, we can also consider *conditional* e-to-p calibrators that are valid only on some subset $\mathcal{E}$ of e-variables, which is a weaker requirement, and such calibrators depend on the choice of $\mathcal{E}$. This topic is treated in Chapter 15.

In sharp contrast to the uniqueness of the admissible e-to-p calibrator, the class of p-to-e calibrators is much richer.

Proposition 2.5

A decreasing function $f : [0, \infty) \rightarrow [0, \infty]$ with $f = 0$ on $(1, \infty)$ is a p-to-e calibrator if and only if $\int_0^1 f(p) dp \leq 1$. It is admissible if and only if f is upper semicontinuous, $f(0) = \infty$, and $\int_0^1 f(p) dp = 1$.

In the context of this proposition, being upper semicontinuous is equivalent to being left-continuous.

Proof.

The first “only if” statement is obvious. To show the first “if” statement, suppose that $\int_0^1 f(p) dp \leq 1$, P is a p-variable for an atomless probability measure $\mathbb{P}$, and P' is uniformly distributed on $[0, 1]$. Since $\mathbb{P}(P \leq x) \leq x \wedge 1 = \mathbb{P}(P' \leq x)$ for all $x \geq 0$ and f is decreasing, we have

$$\mathbb{P}(f(P) > x) \leq \mathbb{P}(f(P') > x)$$

for all $x \geq 0$, which implies

$$\mathbb{E}^{\mathbb{P}}[f(P)] \leq \mathbb{E}^{\mathbb{P}}[f(P')] = \int_0^1 f(p) dp \leq 1.$$

Both necessity and sufficiency in the second statement of Proposition 2.5 are straightforward.

Below we present some simple examples of admissible p-to-e calibrators. We only specify them on $(0, 1]$, as they are ∞ at 0 and 0 on $(1, \infty)$. These calibrators include the power form

$$f(p) = \kappa p^{\kappa-1} \quad \text{for some } \kappa \in (0, 1), \tag{2.1}$$

the mixture of (2.1),

$$f(p) = \int_0^1 \kappa p^{\kappa-1} d\kappa = \frac{1 - p + p \log p}{p(-\log p)^2}, \tag{2.2}$$

and some special simple forms

$$f(p) = 2(1 - p); \quad (2.3)$$

$$f(p) = p^{-1/2} - 1; \quad (2.4)$$

$$f(p) = -\log p. \quad (2.5)$$

Converting a p-value to an e-value using a p-to-e calibrator and then back to p-value using an e-to-p calibrator generally loses quite a lot of evidence. For instance, starting with $p = 0.01$, a conversion with the p-to-e calibrator $p \mapsto p^{-1/2} - 1$ gives $e = 9$, and another conversion with the e-to-p calibrator $e \mapsto \min(1/e, 1)$ yields $p' = 1/9$. Therefore, it is not recommended in practice to calibrate iteratively in both directions.

A possible interpretation of the above results on calibrators is that e-variables and p-variables are connected via a very rough relation $1/e \sim p$. In one direction, the statement is precise: the reciprocal (truncated to 1 if needed) of an e-variable is a p-variable by Proposition 2.4. On the other hand, using a calibrator $f(p) = \kappa p^{\kappa-1}$ in (2.1) with a small $\kappa > 0$ and ignoring positive constant factors, we can see that the reciprocal of a p-variable is approximately an e-variable; this is also the case of (2.2). In fact, $f(p) \leq 1/p$ for all p when f is a calibrator; otherwise $\int_0^1 f(q) dq \geq pf(p) > 1$, conflicting the condition in Proposition 2.5. However, $f(p) = 1/p$ for a fixed p is only possible in the extreme case $f : q \mapsto \mathbb{1}_{\{q \in [0, p]\}}/p$.

In the rest of the book, we omit “p-to-e” when mentioning calibrators, unless we want to stress the direction. Although all calibrators in (2.1)–(2.5) are admissible, some are more useful. In hypothesis testing, very small p-values and very large e-values are interesting. Calibrators in (2.3), (2.4) and (2.5) penalize very small p-values more than (2.2), which can be seen from their asymptotic behaviour as $p \downarrow 0$, and therefore they are less likely to produce very large e-values; the calibrator in (2.3) is even bounded by 2 for $p \neq 0$. Nevertheless, they may be useful when we take the product of many independent e-values.

We end this section with a simple property of admissible calibrators. It also explains why we choose the domain of the calibrators as $[0, \infty)$ instead of $[0, 1]$, although its value is set to 0 for input p-values larger than 1. This result will be useful for later results in Chapter 12.

Proposition 2.6

If f is a calibrator or admissible calibrator, then so is the mapping $x \mapsto af(ax)$ for any $a \geq 1$.

Proof.

Suppose that f is an admissible calibrator. Denote by $g : x \mapsto af(ax)$. We need to check a few properties in Proposition 2.5. First, it follows from $f(p) = 0$ for $p > 1$ that $g(p) = af(ap) = 0$ for $p > 1$, and further

$$\int_0^1 g(x) dx = \int_0^1 af(ax) dx = \int_0^{1/a} af(ax) dx = \int_0^1 f(p) dp = 1.$$

Upper semicontinuity and $g(0) = \infty$ follow directly from those properties of f . Therefore, g is an admissible calibrator. The same argument above shows that g is also a calibrator when f is a calibrator that is not necessarily admissible.

The next result states that any e-variable E for an atomless probability measure $\mathbb{P}$ can be seen as some calibrator applied to an exact p-variable. Indeed, as we see from the proof of the result below, this calibrator is a transform of the quantile function of E .

Proposition 2.7

Let E be an e-variable for an atomless probability measure $\mathbb{P}$. Then, there exists a calibrator f and an exact p-variable P for $\mathbb{P}$ such that $E = f(P)$, $\mathbb{P}$ -almost surely. Moreover, if E is exact, then we can require the calibrator f to be admissible.

Proof.

For $p \in (0, 1]$, let $f(p)$ be given by $f(p) = \inf\{x \in \mathbb{R} : \mathbb{P}(E \leq x) > 1 - p\}$, that is, the right quantile of E at level $1 - p$, and further we set $f(0) = \infty$ and $f(x) = 0$ for $x > 1$. By a standard property of the quantile function (formally stated in Lemma A.10 in Appendix A.2), there exists an exact p-variable P such that $E = f(P)$ almost surely. In case E is exact, the admissibility of this calibrator follows from Proposition 2.5 and the fact that the right quantile function is upper semicontinuous.

Example 2.8

In Section 1.6, for testing $\mathbb{P} = N(0, 1)$ against $\mathbb{Q} = N(\mu, 1)$ with n observations, the likelihood ratio e-value in (1.2) is $E = \exp(\delta Z - \delta^2/2)$, and the Neyman–Pearson p-variable is $P = \Phi(-Z)$, where Z is $N(0, 1)$ -distributed, Φ is the standard normal cdf, and $\delta = \mu\sqrt{n}$. Both E and P are exact. It is straightforward to see $E = f(P)$ with the admissible calibrator f given by

$$f(p) = \exp(-\delta\Phi^{-1}(p) - \delta^2/2).$$

2.4 Randomized tests and calibrators

A randomized test is a test that utilizes independent external randomization. It is possible to improve Markov’s inequality using external randomization, and this has direct implications for converting e-variables into tests or p-variables.

Theorem 2.9: Randomized Markov inequality

Fix $\mathbb{P} \in \mathcal{M}_1$. Let Y be a nonnegative random variable and $U \stackrel{d}{\sim} U[0, 1]$ be independent of Y . Then, for any $\alpha > 0$,

$$\mathbb{P}(Y \geq U/\alpha) = \mathbb{E}^{\mathbb{P}}[\min(\alpha Y, 1)] \leq \alpha \mathbb{E}^{\mathbb{P}}[Y]. \quad (2.6)$$

In particular, if Y is an e-variable for $\mathcal{P}$, then $\mathbb{P}(Y \geq U/\alpha) \leq \alpha$ for any $\mathbb{P} \in \mathcal{P}$.

The proof is simple. Since U is uniform and independent of Y , we get

$$\mathbb{P}(Y \geq U/\alpha) = \mathbb{E}^{\mathbb{P}}[\mathbb{P}(U \leq \alpha Y \mid Y)] = \mathbb{E}^{\mathbb{P}}[\min(\alpha Y, 1)],$$

yielding the claim. The first equality in (2.6) becomes an inequality $\leq$ if U is stochastically larger than standard uniform, and in particular $U = 1$ yields an alternative proof of Markov’s inequality. Also, if Y is bounded, that is $Y \in [0, C]$ almost surely for some $C > 0$, then the inequality in (2.6) holds with equality for any $\alpha \leq 1/C$. In contrast, Markov’s inequality holds with equality only for distributions supported on $\{0, 1/\alpha\}$.

The following corollary is a direct consequence of Theorem 2.9, which offers additional interpretations of the relation between p-variables, e-variables, and randomization.

Corollary 2.10

For any hypothesis $\mathcal{P}$, if a p-variable P and an e-variable E are independent, then P/E is a p-variable. In particular, if E is an e-variable, then U/E is a p-variable, where $U \stackrel{d}{\sim} U[0, 1]$ is independent of E .

Since $U < 1$ almost surely, U/E is a strictly smaller p-value than $1/E$ in all but degenerate cases. Thus, the corresponding randomized test is typically strictly more powerful than the non-randomized one. Said differently, $e \mapsto (U/e) \wedge 1$ can be seen as a *randomized e-to-p calibrator*. In fact, this randomized e-to-p

calibrator is admissible (treated in Section 11.3) and it dominates $e \mapsto (1/e) \wedge 1$, the latter only being admissible among non-randomized e-to-p calibrators (Proposition 2.4).

A more general treatment of combining a p-value and an e-value is presented in Section 11.3.

2.5 Admissible e-variables

In this section, we consider admissible e-variables, which are, intuitively, e-variables that cannot be enlarged, formally defined below.

Definition 2.11

An e-variable E for $\mathcal{P}$ is *admissible* if for any other e-variable E' for $\mathcal{P}$ satisfying $E' \geq E$, we have $\mathbb{P}(E' = E) = 1$ for each $\mathbb{P} \in \mathcal{P}$.

Equivalently, we may say that an e-variable is admissible if it is not dominated by another e-variable, where by domination of E' over E , we mean $E' \geq E$ and $\mathbb{P}(E' > E) > 0$ for some $\mathbb{P} \in \mathcal{P}$.

A trivial example is that the constant 1 is an admissible e-variable for $\mathcal{P}$.

The above definition does not use any alternative $\mathbb{Q}$, but if all $\mathbb{P} \in \mathcal{P}$ and $\mathbb{Q}$ are mutually absolutely continuous, then E being admissible means that one cannot find another e-variable E' such that $\mathbb{Q}(E' \geq E) = 1$ and $\mathbb{Q}(E' > E) > 0$. Indeed, if $\mathbb{Q}(E' \geq E) = 1$ and $\mathbb{Q}(E' > E) > 0$ hold for an e-variable E' , then $E' \vee E$ is an e-variable that dominates E .

2.5.1 Admissible e-variables for a simple null are likelihood ratios

We first make a straightforward but important observation: For simple nulls, every admissible e-variable can be written as a likelihood ratio.

Proposition 2.12

For a simple null $\mathbb{P}$, an e-variable E is admissible if and only if it is exact. Further, any exact e-variable for $\mathbb{P}$ equals a likelihood ratio $d\mathbb{S}/d\mathbb{P}$ $\mathbb{P}$ -almost surely for some distribution $\mathbb{S} \ll \mathbb{P}$.

Proof.

If an e-variable E is not exact, meaning that $\mathbb{E}^{\mathbb{P}}[E] < 1$, then $E + (1 - \mathbb{E}^{\mathbb{P}}[E])$ is an e-variable dominating E , and hence E cannot be admissible. To show the converse, it suffices to note that if E' dominates an exact e-variable E , then $\mathbb{E}^{\mathbb{P}}[E'] > 1$; hence, exact e-variables for $\mathbb{P}$ are admissible. For the last statement, we can define $\mathbb{S}$ via the equality $d\mathbb{S}/d\mathbb{P} = E$, and because E is nonnegative and exact, we know that $\mathbb{S}$ is a distribution that is absolutely continuous with respect to $\mathbb{P}$.

2.5.2 Characterization of admissibility

We now characterize admissible e-variables for general $\mathcal{P}$. For $A \in \mathcal{F}$ and $\delta > 0$, let $\mathcal{P}_A^\delta = \{\mathbb{P} \in \mathcal{P} : \mathbb{P}(A) > \delta\}$ and $\mathcal{P}_A = \{\mathbb{P} \in \mathcal{P} : \mathbb{P}(A) > 0\}$.

The first result gives a simple necessary condition for admissible e-variables, which boils down to the exactness condition in Proposition 2.12 when $\mathcal{P} = \{\mathbb{P}\}$.

Proposition 2.13

If E is an admissible e-variable for $\mathcal{P}$, then for any $A \in \mathcal{F}$ with $\mathcal{P}_A \neq \emptyset$, we have $\sup_{\mathbb{P} \in \mathcal{P}_A} \mathbb{E}^{\mathbb{P}}[E] = 1$.

Proof.

Suppose $\sup_{\mathbb{P} \in \mathcal{P}_A} \mathbb{E}^{\mathbb{P}}[E] < 1$ for some $A \in \mathcal{F}$. Let $\varepsilon = 1 - \sup_{\mathbb{P} \in \mathcal{P}_A} \mathbb{E}^{\mathbb{P}}[E] > 0$ and $E' = E + \varepsilon \mathbf{1}_A$. Clearly $\mathbb{E}^{\mathbb{P}}[E'] \leq 1 - \varepsilon + \varepsilon = 1$ for $\mathbb{P} \in \mathcal{P}_A$, and $\mathbb{E}^{\mathbb{P}}[E'] = \mathbb{E}^{\mathbb{P}}[E] \leq 1$ for $\mathbb{P} \in \mathcal{P} \setminus \mathcal{P}_A$. Therefore, E' is an e-variable for $\mathcal{P}$ dominating E , a contradiction.

Then next result gives a full characterization of admissible e-variables using its expectations under $\mathbb{P} \in \mathcal{P}_A^\delta$ for $A \in \mathcal{F}$. In this result, we use the convention $\inf \emptyset = 0$.

Theorem 2.14

A random variable $E : \Omega \rightarrow [0, \infty]$ is an admissible e-variable for $\mathcal{P}$ if and only if

$$\liminf_{\delta \downarrow 0} \inf_{\mathbb{P} \in \mathcal{P}_A^\delta} \frac{1 - \mathbb{E}^{\mathbb{P}}[E]}{\delta} = 0 \text{ for all } A \in \mathcal{F}. \quad (2.7)$$

Proof.

We first show the “if” direction. First, by choosing $A = \Omega$, (2.7) implies that $(1 - \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[E])/\delta \rightarrow 0$ as $\delta \downarrow 0$, which implies that $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[E] = 1$, and thus E is an e-variable. Now, suppose that there exists an e-variable E' with $E' \geq E$ and $\mathbb{P}'(E' > E) > 0$ for some $\mathbb{P}' \in \mathcal{P}$. Then, there exists $\varepsilon > 0$ such that the set $A := \{E' - E \geq \varepsilon\}$ satisfies $\mathbb{P}'(A) > 0$. Clearly, $E' \geq E + \varepsilon \mathbf{1}_A$. By (2.7), there exists $\delta > 0$ such that $\sup_{\mathbb{P} \in \mathcal{P}_A^\delta} \mathbb{E}^{\mathbb{P}}[E] > 1 - \varepsilon\delta$. It follows that

$$\sup_{\mathbb{P} \in \mathcal{P}_A^\delta} \mathbb{E}^{\mathbb{P}}[E'] \geq \sup_{\mathbb{P} \in \mathcal{P}_A^\delta} \mathbb{E}^{\mathbb{P}}[E + \varepsilon \mathbf{1}_A] > 1 - \varepsilon\delta + \varepsilon\delta = 1.$$

This shows that E' cannot be an e-variable. Hence, E is an admissible e-variable.

Next, we show the “only if” direction. Suppose that (2.7) does not hold; that is, there exist an event $A \in \mathcal{F}$ and a constant $c > 0$ such that

$$\sup_{\mathbb{P} \in \mathcal{P}_A^{\delta_n}} \mathbb{E}^{\mathbb{P}}[E] < 1 - c\delta_n$$

for n large enough, where we take $\delta_n = 2^{-n}$. Let $E' = E + c\mathbf{1}_A/2$. For $\mathbb{P} \in \mathcal{P}_A^{\delta_n} \setminus \mathcal{P}_A^{\delta_{n-1}}$, we have

$$\mathbb{E}^{\mathbb{P}}[E'] \leq 1 - c\delta_n + \frac{c}{2}\mathbb{P}(A) \leq 1 - c\delta_n + c\delta_n \leq 1.$$

For $\mathbb{P} \in \mathcal{P} \setminus \mathcal{P}_A$, $\mathbb{E}^{\mathbb{P}}[E'] = \mathbb{E}^{\mathbb{P}}[E] \leq 1$. Therefore, for all $\mathbb{P} \in \mathcal{P}$, $\mathbb{E}^{\mathbb{P}}[E'] \leq 1$, and E' dominates E . This contradicts the assumption that E is an admissible e-variable for $\mathcal{P}$.

Several sufficient conditions for (2.7) are easier to check, summarized in the next corollary.

Corollary 2.15

An e-variable E for $\mathcal{P}$ is admissible if any of the following conditions holds:

- (a) for each $A \in \mathcal{F}$ with $\mathcal{P}_A \neq \emptyset$, there exists $\mathbb{P} \in \mathcal{P}_A$ such that $\mathbb{E}^{\mathbb{P}}[E] = 1$;
- (b) for each $\mathbb{P}' \in \mathcal{P}$, there exists $\mathbb{P} \in \mathcal{P}$ such that $\mathbb{P}' \ll \mathbb{P}$ and $\mathbb{E}^{\mathbb{P}}[E] = 1$;
- (c) E is exact.

Proof.

(a) For each $A \in \mathcal{F}$ with $\mathcal{P}_A \neq \emptyset$, $\inf_{\mathbb{P} \in \mathcal{P}_A^\delta} (1 - \mathbb{E}^\mathbb{P}[E]) = 0$ for $\delta > 0$ small enough, and hence (2.7) holds. It is straightforward to check that (b) implies (a), and (c) implies (b).

The condition (a) in Corollary 2.15 deserves some further attention. We will say that an e-variable is *strongly admissible* if condition (a) in Corollary 2.15 holds; note that (a) is implied by either of (b) and (c). This condition is strictly stronger than admissibility for infinite $\mathcal{P}$ as characterized in Theorem 2.14 (for which there need not be any $\mathbb{P} \in \mathcal{P}$ with $\mathbb{E}^\mathbb{P}[E] = 1$). When $\mathcal{P}$ is finite, strong admissibility is equivalent to admissibility, as shown in the next result.

Proposition 2.16

Suppose that $\mathcal{P}$ is finite. Then, E is an admissible e-variable for $\mathcal{P}$ if and only if $\sup_{\mathbb{P} \in \mathcal{P}_A} \mathbb{E}^\mathbb{P}[E] = 1$ for all $A \in \mathcal{F}$ with $\mathcal{P}_A \neq \emptyset$.

Proof.

The “only-if” statement follows from Proposition 2.13. To show the “if” statement, first observe that E is an e-variable by choosing $A = \Omega$. Then, admissibility of E follows by using condition (a) in Corollary 2.15.

2.5.3 Admissible e-variables on product spaces

Next, we consider e-variables defined on product spaces. Let $\mathcal{P}'$ be a set of probabilities on another measurable space $(\Omega', \mathcal{F}')$. We write $\mathcal{P} \times \mathcal{P}'$ for the set $\{\mathbb{P} \times \mathbb{P}' : \mathbb{P} \in \mathcal{P}, \mathbb{P}' \in \mathcal{P}'\}$ of product probabilities. For E on Ω and E' on Ω' , let $E \times E'$ be defined by $(\omega, \omega') \mapsto E(\omega)E'(\omega')$. Clearly, if E is an e-variable for $\mathcal{P}$ and E' is an e-variable for $\mathcal{P}'$, then $E \times E'$ is an e-variable for $\mathcal{P} \times \mathcal{P}'$, because

$$\mathbb{E}^{\mathbb{P} \times \mathbb{P}'}[E \times E'] = \mathbb{E}^\mathbb{P}[E] \mathbb{E}^{\mathbb{P}'}[E'] \leq 1$$

for all $\mathbb{P} \in \mathcal{P}$ and $\mathbb{P}' \in \mathcal{P}'$. The next result further shows that $E \times E'$ is strongly admissible when E and E' are strongly admissible.

Proposition 2.17

Let E and E' be strongly admissible e-variables for $\mathcal{P}$ and $\mathcal{P}'$, respectively. Then $E \times E'$ is a strongly admissible e-variable for $\mathcal{P} \times \mathcal{P}'$.

Proof.

As we explained above, $E \times E'$ is an e-variable for $\mathcal{P} \times \mathcal{P}'$. To show its strong admissibility, take any B with $(\mathcal{P} \times \mathcal{P}')_B \neq \emptyset$. It follows from standard real analysis that there exist $A \in \mathcal{F}$ and $A' \in \mathcal{F}'$ such that $A \times A' \subseteq B$ and $(\mathcal{P} \times \mathcal{P}')_{A \times A'} \neq \emptyset$. By strong admissibility of E and E' , there exist $\mathbb{P} \in \mathcal{P}_A$ and $\mathbb{P}' \in \mathcal{P}'_{A'}$ such that $\mathbb{E}^\mathbb{P}[E] = \mathbb{E}^{\mathbb{P}'}[E'] = 1$. This gives $(\mathbb{P} \times \mathbb{P}')(B) > 0$ and $\mathbb{E}^{\mathbb{P} \times \mathbb{P}'}[E \times E'] = 1$, and therefore $E \times E'$ is strongly admissible.

Combining Propositions 2.16 and 2.17, we obtain that $E \times E'$ is an admissible e-variable for admissible e-variables E and E' when $\mathcal{P}$ and $\mathcal{P}'$ are finite.

2.6 More examples of e-values

Continuing from the examples in Chapter 1, we now provide several simple examples of e-values for different testing problems. Some of these e-values also enjoy a notion of optimality, introduced in Chapter 3 and analyzed in detail in Chapter 6, so these examples will be revisited later in the book.

In most examples below, we consider a single random observation (statistic) Z in a measurable space $\mathcal{Z}$ and the sample space is chosen as $(\Omega, \mathcal{F}) = (\mathcal{Z}, \sigma(Z))$; thus $\mathcal{M}_1$ is the set of all probability measures on $\sigma(Z)$.

2.6.1 Monotone likelihood ratio

Let $(\mathbb{P}_\theta)_{\theta \in \Theta}$ be a parametric family of distributions with their density functions of the test statistic Z given by $(p_\theta)_{\theta \in \Theta}$, where $\Theta \subseteq \mathbb{R}$ and $\mathcal{Z} = \mathbb{R}$. Assume that monotone likelihood ratio holds: for all $\theta < \theta'$, the likelihood ratio $p_{\theta'}(z)/p_\theta(z)$ exists and is increasing (or decreasing) in z .

Proposition 2.18

Consider testing $\mathcal{P} = \{\mathbb{P}_\theta : \theta \leq \theta_0\}$ against $\mathcal{Q} = \{\mathbb{P}_\theta : \theta \geq \theta_1\}$, where $\theta_1 > \theta_0$. Under the monotone likelihood ratio assumption, for any $\theta' > \theta_0$, the random variable

$$E = \frac{p_{\theta'}(Z)}{p_{\theta_0}(Z)}$$

is an e-variable for $\mathcal{P}$.

The following short proof uses Chebyshev's association inequality, which states, assuming integrability, $\mathbb{E}^\mathbb{P}[f(X)g(X)] \geq \mathbb{E}^\mathbb{P}[f(X)]\mathbb{E}^\mathbb{P}[g(X)]$ for any increasing functions f, g and random variable X .

Proof.

Write $\mathbb{E}^\theta$ for $\mathbb{E}^{\mathbb{P}_\theta}$. For any $\theta \leq \theta_0 < \theta'$, we have

$$\begin{aligned} 1 &= \mathbb{E}^{\theta_0} \left[\frac{p_{\theta'}(Z)}{p_{\theta_0}(Z)} \right] = \mathbb{E}^\theta \left[\frac{p_{\theta'}(Z)}{p_{\theta_0}(Z)} \frac{p_{\theta_0}(Z)}{p_\theta(Z)} \right] \\ &\geq \mathbb{E}^\theta \left[\frac{p_{\theta'}(Z)}{p_{\theta_0}(Z)} \right] \mathbb{E}^\theta \left[\frac{p_{\theta_0}(Z)}{p_\theta(Z)} \right] = \mathbb{E}^\theta \left[\frac{p_{\theta'}(Z)}{p_\theta(Z)} \right], \end{aligned}$$

where the inequality follows from Chebyshev's association inequality. This gives $\mathbb{E}^\theta[E] \leq 1$ for all $\theta \leq \theta_0$, as required.

The same conclusion in Proposition 2.18 holds true for testing $\mathcal{P} = \{\mathbb{P}_\theta : \theta \geq \theta_0\}$ against $\mathcal{Q} = \{\mathbb{P}_\theta : \theta \leq \theta_1\}$ with $\theta_0 > \theta_1$; moreover, either inequality (or both) in $\mathcal{P}$ and $\mathcal{Q}$ can be replaced by a strict inequality. The same e-variable can also be used to test $\mathcal{P} = \{\mathbb{P}_\theta : \theta \leq \theta_0\}$ against $\mathcal{Q} = \{\mathbb{P}_\theta : \theta > \theta_0\}$.

The example of normal distributions in Section 1.6 satisfies the monotone likelihood ratio assumption. Therefore, the e-variable in (1.2) for any $\mu > 0$ is an e-variable for testing $\{N(\theta, 1) : \theta \leq 0\}$ against $\{N(\theta, 1) : \theta > 0\}$.

An important special case is given by one-parameter exponential families, which we discuss further in Section 6.7.1. There, and in Section 6.7.2, we will argue that the above e-variable is actually *log-optimal* for $\mathcal{Q} = \mathbb{P}_{\theta_1}$ for any $\theta_1 > \theta_0$, simply because $p_{\theta_1}/p_{\theta_0}$ is an e-variable for $\mathcal{P}$.

2.6.2 Testing symmetry

Take $\mathcal{Z} = \mathbb{R}$. We now consider the null hypothesis that Z is symmetrically distributed around 0, that is,

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : Z \text{ and } -Z \text{ have the same distribution under } \mathbb{P}\}.$$

An example in clinical trials is to set $Z = Y_1 - Y_0$, where Y_1 is a post-treatment measurement and Y_0 is pre-treatment measurement, and we are interested in whether the treatment has any effect. Such an application typically has more than one data point, which will be discussed later, but for now we first consider the case of one summary statistic Z .

Fix any alternative hypothesis $\mathbb{Q}$ that admits a Lebesgue density q . Let $\mathbb{P}^*$ be given by its density

$$p^*(z) = \frac{1}{2} (q(z) + q(-z)) \mathbb{1}_{\{q(z) > 0\}},$$

which is absolutely continuous with respect to $\mathbb{Q}$. Define

$$E = \frac{d\mathbb{Q}}{d\mathbb{P}^*} = \frac{2q(Z)}{q(Z) + q(-Z)} \mathbb{1}_{\{q(Z) > 0\}}. \quad (2.8)$$

We can check that E is an e-variable for $\mathcal{P}$ because, using the convention $0/0 = 0$,

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[E] &= \mathbb{E}^{\mathbb{P}} \left[\frac{q(Z)}{q(Z) + q(-Z)} \right] + \mathbb{E}^{\mathbb{P}} \left[\frac{q(Z)}{q(Z) + q(-Z)} \right] \\ &= \mathbb{E}^{\mathbb{P}} \left[\frac{q(Z)}{q(Z) + q(-Z)} \right] + \mathbb{E}^{\mathbb{P}} \left[\frac{q(-Z)}{q(Z) + q(-Z)} \right] = 1. \end{aligned}$$

Corollary 2.15 implies that E is admissible because it is exact. The optimality of this e-variable against $\mathbb{Q}$ is studied in Section 6.7.3.

The e-variable in (2.8) based on one realization is at most 2, and in applications we need to rely on multiple data points to accumulate more evidence against the null. Suppose that multiple iid data points sampled from X are available, summarized by $Z = (X_1, \dots, X_n)$ with $\mathcal{Z} = \mathbb{R}^n$. We consider the hypothesis $\mathcal{P}$ of symmetry, that is, X and $-X$ have the same distribution under $\mathbb{P}$, and for simplicity, assume further that X is continuously distributed. There are many e-variables to test $\mathcal{P}$, and we give a few. First, let

$$E_\lambda = \prod_{i=1}^n \frac{2 \exp(\lambda X_i)}{\exp(\lambda X_i) + \exp(-\lambda X_i)}, \quad \lambda \in \mathbb{R},$$

Corollary 2.15 implies E_λ is admissible since it is exact. This is the product of several e-variables in (2.8) with a specifically chosen q . Another class of e-variables is

$$E'_\lambda = \prod_{i=1}^n \exp(\lambda X_i - \lambda^2 X_i^2 / 2), \quad \lambda \in \mathbb{R}.$$

We have $E'_\lambda \leq E_\lambda$, because

$$\exp(x) + \exp(-x) \leq 2 \exp(x^2 / 2)$$

for all $x \in \mathbb{R}$. To remove the reliance on λ , we can take $E = \int_{\mathbb{R}} E_\lambda d\mu(\lambda)$, where μ is any distribution on $\mathbb{R}$. Thus E'_λ is inadmissible.

Note that the signs of $X_1, \dots, X_n$ are iid and uniformly distributed on $\{-1, 1\}$ under the null hypothesis. Therefore, a class of e-variables is given by

$$E_p = \frac{p^k (1-p)^{n-k}}{2^n}, \quad p \in (0, 1),$$

where k is the number of 1s in $\{X_1, \dots, X_n\}$. Similarly to E_λ above, E_p is also a product of several e-variables in (2.8), but now this class is indexed by $p \in (0, 1)$.

A different approach is to use the signed-rank e-variables, based on Wilcoxon's signed-rank test. The idea is to arrange the magnitudes $|X_i|$ of the observations in the ascending order and assigning to each its rank, which is a number in $[n]$. Under the null hypothesis, any subset of $[n]$ has the same probability 2^{-n} to be the set of ranks of the positive observations. This determines the distribution of Wilcoxon's statistic V_n defined as the sum of the ranks of the positive observations. The class of signed-rank e-variables is defined by

$$E_\beta = \exp(\beta V_n) \prod_{i=1}^n \frac{2}{1 + \exp(\beta i)}, \quad \beta \in \mathbb{R}.$$

It is a standard exercise to check that E_β is an e-variable.

2.6.3 Testing the mean and the variance

Let $\mathcal{Z} \subseteq \mathbb{R}_+$ so that the statistic Z takes nonnegative values. Fix $\mu > 0$ and consider the null hypothesis that the mean of Z is at most μ ,

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[Z] \leq \mu\}.$$

The alternative hypothesis $\mathcal{Q}$ is that mean of Z is larger than μ . Two natural e-variables are Z/μ and the constant one. All convex combinations of these are also e-variables; therefore,

$$(1 - \lambda) + \lambda \frac{Z}{\mu}, \quad \lambda \in [0, 1], \quad (2.9)$$

is an e-variable for $\mathcal{P}$. This construction works outside $\mathbb{R}_+$ as long as Z is bounded from below, by using $(Z - z)/(\mu - z)$ in place of Z/μ , where z is a lower bound on Z . An optimal choice of λ against a specific $\mathbb{Q}$ is discussed in Section 6.7.4 in the case that $\mathcal{Z}$ is a bounded interval. Corollary 2.15(b) can be used to deduce that the above e-variable is admissible, despite not being exact.

We can also consider the null hypothesis that the mean of Z is at most μ and the variance is at most σ^2 ,

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[Z] \leq \mu; \text{var}^{\mathbb{P}}(Z) \leq \sigma^2\}.$$

In this case, because $\mathbb{E}^{\mathbb{P}}[Z^2] \leq \mu^2 + \sigma^2$ for $\mathbb{P} \in \mathcal{P}$, each of

$$(1 - \lambda) + \lambda \frac{Z^2}{\mu^2 + \sigma^2}, \quad \lambda \in [0, 1], \quad (2.10)$$

is an e-variable for $\mathcal{P}$, in addition to those in (2.9), as well as the mixtures of (2.9) and (2.10). If $\mathcal{Z} = \mathbb{R}$, meaning that Z may take negative values, then (2.10) is an e-variable for

$$\{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[Z] = \mu; \text{var}^{\mathbb{P}}(Z) \leq \sigma^2\}. \quad (2.11)$$

As a further example, suppose that we observe nonnegative data $Z = (X_1, \dots, X_n)$ with the same marginal distribution $\mathbb{P}$ and possibly dependent components, and we wish to test whether $\mathbb{P}$ has a mean at most μ . Then, $(X_1 + \dots + X_n)/(n\mu)$ is an e-value for any dependence structure of $(X_1, \dots, X_n)$, and does not require making any distributional or independence assumption. An e-variable can be constructed similarly for the case of testing the mean and variance. This example shows the advantage of e-values when it comes to dependent data, and this will be further explored in Chapter 9.

2.6.4 Testing sub-Gaussian means

Now take $\mathcal{Z} = \mathbb{R}$ and consider the null hypothesis that the observation Z has a 1-sub-Gaussian distribution with nonpositive mean. This class is a nonparametric generalization of Gaussians with nonpositive mean and unit variance. Formally, we set

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[e^{\lambda Z - \lambda^2/2}] \leq 1 \text{ for all } \lambda \in [0, \infty) \right\}.$$

Here the sub-Gaussian restriction is only placed on the right tail. For $\mathbb{P} \in \mathcal{P}$ and $\lambda \in \mathbb{R}$, let $M_{\mathbb{P}}(\lambda) = \mathbb{E}^{\mathbb{P}}[e^{\lambda Z}]$, which is the moment generating function of Z . The constraint in $\mathcal{P}$ means

$$M_{\mathbb{P}}(\lambda) \leq e^{\lambda^2/2}, \quad \lambda \geq 0.$$

A Taylor expansion in a neighborhood of $\lambda = 0$ yields,

$$\begin{aligned} M_{\mathbb{P}}(\lambda) &= 1 + \lambda \mathbb{E}^{\mathbb{P}}[Z] + \frac{\lambda^2}{2} \mathbb{E}^{\mathbb{P}}[Z^2] + o(\lambda^2); \\ e^{\lambda^2/2} &= 1 + \frac{\lambda^2}{2} + o(\lambda^2). \end{aligned}$$

Therefore, for $\lambda \geq 0$ small enough,

$$\lambda \mathbb{E}^\mathbb{P}[Z] + \frac{\lambda^2}{2} \mathbb{E}^\mathbb{P}[Z^2] \leq \frac{\lambda^2}{2}, \quad (2.12)$$

and this implies $\mathbb{E}^\mathbb{P}[Z] \leq 0$.

Since the standard Gaussian distribution (zero mean, unit variance) satisfy the above with equality (inferred via their moment generating function) for every real λ , the constraint expands the class to account for two aspects: the mean can be negative, and the right tails of the random variable only have to be lighter than that of a unit-variance Gaussian. It is well-known that any bounded distribution on $[-1, 1]$ with zero mean is 1-sub-Gaussian; so the class $\mathcal{P}$ clearly does not have any common reference measure, containing discrete distributions with uncountable support as well as continuous unbounded distributions.

A class of e-variables for $\mathcal{P}$ is given by:

$$\{e^{\lambda Z - \lambda^2/2} : \lambda \in [0, \infty)\},$$

and certainly any convex combinations (mixtures) of its elements are also e-variables. The determination of the optimal e-variable against a specific $\mathbb{Q}$ is studied in Section 6.7.5. The determination of all possible e-variables for $\mathcal{P}$ is a nontrivial task.

We also consider the closely related null hypothesis that the observation Z has a 1-sub-Gaussian distribution with zero (instead of nonpositive) mean:

$$\mathcal{P}^0 = \left\{ \mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^\mathbb{P}[e^{\lambda Z - \lambda^2/2}] \leq 1 \text{ for all } \lambda \in \mathbb{R} \right\}.$$

Here, the sub-Gaussian restriction is placed on both the left and the right tails. Every $\mathbb{P} \in \mathcal{P}^0$ has zero mean, because now (2.12) holds for both positive and negative λ . Again, a natural class of e-variables for $\mathcal{P}^0$ is

$$\{e^{\lambda Z - \lambda^2/2} : \lambda \in \mathbb{R}\},$$

and certainly any convex combinations (mixtures) of its elements are also e-variables.

It turns out that every $\mathbb{P} \in \mathcal{P}^0$ also has variance bounded by one, and this follows from (2.12) with $\mathbb{E}^\mathbb{P}[Z] = 0$. Thus, Z^2 is also an e-variable, as is any convex combination with the aforementioned class. Indeed, $\mathcal{P}^0$ is a subset of the hypothesis in (2.11) with $\mu = 0$ and $\sigma = 1$, and also a subset of the one-sided sub-Gaussian class $\mathcal{P}$. However, Z^2 is not an e-variable for $\mathcal{P}$, because those could potentially have very heavy left tails, meaning that their variance is not necessarily bounded.

We note in passing that both the above classes of e-variables are admissible, but the rather technical proof is omitted.

2.6.5 Constrained hypotheses

Many of the above examples of $\mathcal{P}$, and many others we encounter in this book, can be called “constrained hypotheses”, meaning that they are defined via integral constraints. For such classes, we can sometimes cleanly characterize the set of *all* e-variables (and all admissible ones) in terms of those constraints.

Let $\mathcal{L}$ denote the set of all real-valued measurable functions on the underlying probability space. A constraint set is any nonempty set of functions $\Psi \subseteq \mathcal{L}$. Define the null hypothesis generated by Ψ as

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{M}_1 : \int |f| d\mathbb{P} < \infty \text{ and } \int f d\mathbb{P} \leq 0 \text{ for all } f \in \Psi \right\}. \quad (2.13)$$

We say that a property holds $\mathcal{P}$ -quasi surely if it holds everywhere outside of a set A such that $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(A) = 0$.

We start with the case where $\Psi = \{g_1, \dots, g_K\}$ is finite, as in the case of testing a mean, or mean and variance. One can show that the set of all e-variables $\mathfrak{E}$ for $\mathcal{P}$ is given by all $[0, \infty]$ -valued measurable functions which are $\mathcal{P}$ -quasi surely dominated by $1 + \sum_{k=1}^K \pi_k g_k$ for some $\pi \in \mathbb{R}_+^K$. As a corollary, every admissible e-variable for $\mathcal{P}$ is $\mathcal{P}$ -quasi surely equal to $1 + \sum_{k=1}^K \pi_k g_k$ for some π in the set

$$\Pi^\Psi := \left\{ \pi \in \mathbb{R}_+^K : 1 + \sum_{k=1}^K \pi_k g_k \geq 0, \text{ } \mathcal{P}\text{-quasi surely} \right\}.$$

Clearly, the zero vector $\mathbf{0}$ is in Π^Ψ . But for some choices of Ψ , Π^Ψ may be equal to $\{\mathbf{0}\}$, in which case all e-variables for $\mathcal{P}$ are dominated by the constant 1.

In settings where Ψ is infinite, such as the sub-Gaussian example in the previous subsection, the corresponding statement is more complex. Define

$$\mathcal{C} = \{f \in \mathcal{L} : f \leq g \text{ } \mathcal{P}\text{-quasi surely for some } g \in \text{cone}(\Psi)\},$$

where $\text{cone}(\Psi)$ is the conic hull of Ψ . Then, the set of all e-variables $\mathfrak{E}$ for $\mathcal{P}$ is given by all $[0, \infty]$ -valued measurable functions that are $\mathcal{P}$ -quasi surely equal to $1 + f$ for some $f \in \bar{\mathcal{C}}$, where $\bar{\mathcal{C}}$ is the weak closure of $\mathcal{C}$ (which is the closure with respect to a particular weak topology that is omitted here for brevity).

The constraints in (2.13) are formulated via integrals of $f \in \Psi$ over $\mathbb{P}$. In some applications, the constraints are formulated via other functionals, such as risk measures, and some of them are studied in Chapter 16.

2.6.6 Testing likelihood ratio bounds

Let $\mathcal{Z}$ be a general sample space. Let $\mathbb{P}_0$ be a fixed probability measure and $\gamma \in [1, \infty)$, and consider the null hypothesis

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : d\mathbb{P}/d\mathbb{P}_0 \leq \gamma\}$$

against an alternative $\mathbb{Q} \ll \mathbb{P}_0$. To interpret, testing $\mathcal{P}$ means to test whether the true data-generating distribution is close to a given benchmark distribution $\mathbb{P}_0$ in terms of likelihood ratio being bounded by γ . The case $\gamma = 1$ corresponds to the simple null hypothesis $\{\mathbb{P}_0\}$.

Clearly, the likelihood ratio $Z^* = d\mathbb{Q}/d\mathbb{P}_0$ is not an e-variable for $\mathcal{P}$ unless $\gamma = 1$, but if we modify it to $E = Z^*/\gamma$, then $\mathbb{E}^{\mathbb{P}}[E] \leq \gamma \mathbb{E}^{\mathbb{P}_0}[Z^*/\gamma] \leq 1$. Hence, $E = Z^*/\gamma$ is an e-variable for $\mathcal{P}$. We can actually improve this e-variable, as described below.

For any random variable Z , denote by $q_t(Z) = \inf\{x \in \mathbb{R} : \mathbb{P}_0(Z \leq x) \geq 1 - t\}$ for $t \in (0, 1)$; that is, $t \mapsto q_{1-t}(Z)$ is the left quantile function of Z under $\mathbb{P}_0$. Let E^* be a transform of the likelihood ratio given by

$$E^* = \frac{Z^* \vee z_0}{\gamma},$$

where $z_0 \geq 0$ is the largest constant such that

$$\int_0^{1/\gamma} (q_t(Z^*) \vee z_0) dt = 1.$$

Clearly $E^* \geq E$. Let us verify that E^* is an e-variable for $\mathcal{P}$. This follows by noting that

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[E^*] = \gamma \int_0^{1/\gamma} q_t(E^*) dt = \int_0^{1/\gamma} (q_t(Z^*) \vee z_0) dt = 1,$$

where the first equality is a classic result on risk measures; see (16.6) in Chapter 16. The optimality of the e-variable E^* is proved in Section 6.7.6.

2.6.7 Testing exchangeability

Let $\mathcal{Z}$ be arbitrary, and suppose $Z = (Z_1, \dots, Z_n)$. Denote by $\mathcal{S}_n$ the set of all permutations of the set $[n]$. For any $\sigma \in \mathcal{S}_n$, let Z_σ denote the permuted vector $(Z_{\sigma(1)}, \dots, Z_{\sigma(n)})$. Consider testing the null hypothesis of exchangeability

$$\mathcal{P} = \{\mathbb{P} : Z \stackrel{d}{=} Z_\sigma \text{ under } \mathbb{P} \text{ for all permutations } \sigma\},$$

where $\stackrel{d}{=}$ denotes equality of distribution. Let $\mathbb{Q}$ be an alternative distribution of Z with density q , which is not exchangeable, i.e., not in $\mathcal{P}$.

Let π be uniformly distributed on $\mathcal{S}_n$ independent of Z , and let

$$E = \frac{q(Z)}{q(Z_\pi)} = \frac{q(Z)}{\frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} q(Z_\sigma)}.$$

By exchangeability, $\mathbb{E}[E] = \mathbb{E}[q(Z)/q(Z_\pi)] = 1$ for all $\mathbb{P} \in \mathcal{P}$, so E is an exact e-variable, which is admissible by Corollary 2.15. The e-variable E has the form of the soft-rank e-value in Section 1.7, where we use $r = 0$ and $R_b = L_b = q(Z_\sigma)$ for $\sigma \in \mathcal{S}_n$. The optimality of the e-variable E is proved in Section 6.7.7.

2.7 Informal grids for realized e-values

In testing scientific hypotheses, thresholds for p-values are often chosen as 0.01 or 0.05 which correspond to type-I errors controlled at these levels. The e-to-p calibrator $e \mapsto \min(1/e, 1)$ implies that thresholds of 100 and 20 for e-values also have the above type-I error control. However, it is not recommended in practice to directly use these thresholds as the conversion $e \mapsto \min(1/e, 1)$ is typically wasteful. In Section 1.5, and also in Section 4.1 later, we see that for $a \in (0, 1)$, an e-value of $1/a$ carries more information than a p-value of a .

In order to judge how significant results of testing using e-values are, the type-I error, based on which p-values are defined, may not be the desirable metric. Although there are no universally agreed thresholds to use for e-values, the rule of thumb of Jeffreys, originally designed for likelihood ratios, may offer some insight, as e-values are generalizations of likelihood ratios (for example, recall the result presented in Section 2.5.1). We summarize this rule of thumb in Table 2.19. For instance, if $e > 3.16$, we recommend to associate it with “substantial” (similar to $p < 0.05$), and if $e > 10$, we recommend to associate it with “strong” (similar to $p < 0.01$). One should keep in mind that p-values and e-values do not have a one-to-one correspondence with each other, and the interpretation of the informal grid for e-values does not rely on converting into p-values.

Because Jeffreys focused on likelihood ratios (or Bayes factors), the above rule of thumb applies well to e-values that are obtained “naturally”, such as likelihood ratios or their generalizations in the subsequent chapters. At the other extreme, for all-or-nothing e-variables (see Definition 3.5), such as $E = \mathbf{1}_{\{P \leq \alpha\}}/\alpha$ for some p-variable P , one should not refer to the grids in Table 2.19, because for such an e-variable, $E = 10$ is equivalent to $P \leq 0.1$. Finally, Section 6.6 draws deeper connections between e-values and Bayes factors, but it builds on some more advanced material.

e-value	level of evidence	Shafer’s p-value
$[0, 1)$	null hypothesis is supported	$(0.25, 1]$
$(1, 3.16)$	no more than a bare mention	$(0.0577, 0.25)$
$(3.16, 10)$	substantial	$(8.3 \times 10^{-3}, 0.0577)$
$(10, 31.6)$	strong	$(9.4 \times 10^{-4}, 8.3 \times 10^{-3})$
$(31.6, 100)$	very strong	$(9.8 \times 10^{-5}, 9.4 \times 10^{-4})$
$(100, \infty]$	decisive	$[0, 9.8 \times 10^{-5})$

Table 2.19: Applying Jeffreys’s rule of thumb for likelihood ratios to e-values. For comparison, we also reported Shafer’s p-value, which corresponds to the range of p via $e = p^{-1/2} - 1$ in (2.4). The boundary values can be put in either of the two adjacent categories.

It is important to note that in general, p-values and e-values need not even “point in the same direction”. We can easily have situations where the p-value appears to provide strong evidence against the null, but the e-value supports the null, or vice versa. Let us elaborate with a simple example. Consider testing $\mu = 0$ against $\mu = \delta$ (for some $\delta > 0$) when we see data $Z \sim N(\mu, 1)$. Since the classic z-test is uniformly most powerful in this setting, we would get the same p-value $P = 1 - \Phi(Z)$, regardless of the alternative mean δ , while the (arguably natural) likelihood ratio e-value $E = \exp(\delta Z - \delta^2/2)$ depends strongly on δ . For instance, with $\delta = 5$ and $X = 2$, the p-value $P \approx 0.023$ (suggesting evidence against the null), but the e-value $E \approx 0.08$ (suggesting evidence for the null as a likelihood ratio). Thus the realized p-value and e-value could “point in different directions” depending on δ and the realized data. This is another reason that we do not like to

in general compare e-value and p-value based methods: while calibration can often preserve the direction (large/small e-values yield small/large p-values and vice versa), natural p-values and e-values that we may directly construct from the data need not have any natural concordance with each other, and so the resulting methods are not always directly comparable.

Bibliographical note

This chapter largely resembles results and concepts in various important papers on e-values, but also it contains many results that are derived for the book to make a coherent piece.

The example in Section 2.1 is based on [Vovk and Wang \[2023\]](#). The concept of a calibrator in Section 2.3 dates back, at least, to [Vovk \[1993\]](#); the admissibility results are in [Vovk and Wang \[2021\]](#). [Bickel \[2025\]](#) discussed p-to-e calibrators that depend on the sample size based on Bayes factors. Section 2.4 is based on [Ignatiadis et al. \[2024b\]](#) and [Ramdas and Manole \[2025\]](#).

The definition of admissibility in Section 2.5 is derived from [Ramdas et al. \[2020\]](#), which studies it more generally in sequential settings. Some sufficient conditions in Corollary 2.15 also follow from the aforementioned paper.

Among the examples in Section 2.6, exponential families and monotone likelihood ratios were studied by [Grünwald et al. \[2024a,b\]](#). The testing symmetry example was first studied by [Ramdas et al. \[2020\]](#), with follow-up work in [Koning \[2023b\]](#), [Vovk and Wang \[2024b\]](#), [Larsson et al. \[2025a\]](#). The bounded mean example is a variant of one studied in [Waudby-Smith and Ramdas \[2024\]](#), revisited by [Larsson et al. \[2025a\]](#). The mean-and-variance example was studied by [Fan et al. \[2025\]](#). Sub-Gaussian e-values were studied by [Howard et al. \[2020\]](#), [Ramdas et al. \[2020\]](#), [Larsson et al. \[2025a\]](#). The example on constrained hypotheses was studied by [Clerico \[2024\]](#), [Larsson et al. \[2025b\]](#); our presentation is based on the latter. The exchangeability example was studied in [Vovk \[2021\]](#) and [Ramdas et al. \[2022\]](#) in a sequential context. In the batch context, the soft-rank e-value was proposed in the first arXiv preprint of [Wang and Ramdas \[2022\]](#) (but omitted in the journal version), and revisited by [Koning \[2023b\]](#).

The grids in Section 2.7 can be found in [[Jeffreys, 1961](#), Appendix B], with a variant proposed in [Kass and Raftery \[1995\]](#).

Chapter 3

Efficiency: E-values under the alternative hypothesis

In Chapter 2 we focused only on the properties of the e-values under the null hypothesis. In this chapter, we describe its desired properties under the alternative hypothesis. This requires the notions of powered e-values, e-power, and log-optimality.

3.1 Powered tests, e-values, and p-values

We first introduce the concept of power for tests and e-values. The power of a binary test ϕ against $\mathbb{Q}$ is defined as the probability of rejecting the null hypothesis under the alternative $\mathbb{Q}$. This concept can be naturally generalized to non-binary tests as we present below.

Definition 3.1: Powered tests and e-values

The *power* of a test ϕ against an alternative $\mathbb{Q}$ is simply $\mathbb{E}^{\mathbb{Q}}[\phi]$. The *type-II error (rate)* of ϕ equals one minus its power. When testing a null hypothesis $\mathcal{P}$, a test ϕ is *powered* at level $\alpha \in (0, 1)$ against $\mathcal{Q}$ if

$$\mathbb{E}^{\mathbb{P}}[\phi] \leq \alpha \text{ for every } \mathbb{P} \in \mathcal{P} \text{ and } \mathbb{E}^{\mathbb{Q}}[\phi] > \alpha \text{ for every } \mathbb{Q} \in \mathcal{Q},$$

and it is *uniformly* powered if further $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\phi] > \alpha$. An e-variable E is *powered* against $\mathcal{Q}$ if $\mathbb{E}^{\mathbb{Q}}[E] > 1$ for every $\mathbb{Q} \in \mathcal{Q}$, and it is *uniformly* powered against $\mathcal{Q}$ if $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[E] > 1$.

Remark 3.2. A closely related concept is that of level- α *unbiased* tests, i.e., tests ϕ satisfying $\mathbb{E}^{\mathbb{P}}[\phi] \leq \alpha$ for every $\mathbb{P} \in \mathcal{P}$ and $\mathbb{E}^{\mathbb{Q}}[\phi] \geq \alpha$ for every $\mathbb{Q} \in \mathcal{Q}$. For the purpose of this chapter, our definition of powered tests is needed, and it excludes trivial cases such as the constant test $\phi = \alpha$.

When we mention a powered test ϕ without specifying the level, it should be understood as $\alpha = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[\phi]$. Tests that are not powered can still be practically useful. For instance, the existence of powered tests and e-variables requires that the alternative $\mathcal{Q}$ does not contain a “boundary point” of the null $\mathcal{P}$, as shown by Example 3.3 below. Nevertheless, although powered tests do not exist for testing $H_0 : \mu < 0$ against $H_1 : \mu \geq 0$ in that example, any powered test for testing H_0 against $H'_1 : \mu > 0$ can still be useful to test H_0 against H_1 in practical applications.

Example 3.3

Suppose that iid data $X_1, \dots, X_n$ are sampled from $N(\mu, 1)$ with unknown μ . If we test $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, then $\exp(X_1 + \dots + X_n - n/2)$ is a powered e-variable (but not uniformly e-powered). This e-variable is uniformly powered against $H_1 : \mu \geq \varepsilon$ for any $\varepsilon > 0$.

If we test $H_0 : \mu < 0$ against $H_1 : \mu \geq 0$, then there does not exist a powered e-variable (or test) because it is impossible to have power against the boundary point 0.

Proof of the claim in Example 3.3

Let p_m be the probability density function of $X^n = (X_1, \dots, X_n)$ sampled from $N(\mu, 1)$ with $\mu = -1/m$, and let p_0 be that of the case $\mu = 0$. For any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, we have $f(\mathbf{x})p_m(\mathbf{x}) \rightarrow f(\mathbf{x})p_0(\mathbf{x})$ pointwise in $\mathbf{x}$ as $m \rightarrow \infty$. So Fatou's lemma yields that $\int f(\mathbf{x})p_0(\mathbf{x})d\mathbf{x} \leq \liminf_m \int f(\mathbf{x})p_m(\mathbf{x})d\mathbf{x}$. When $f(X^n)$ is chosen to be an e-variable (resp. level- α test) for $H_0 : \mu < 0$, the right-hand side equals 1 (resp. α), and the inequality above means that f is also an e-variable (resp. level- α test) for the boundary point of $\mu = 0$. Thus, every e-variable (resp. level- α test) for $H_0 : \mu < 0$ is also an e-variable (resp. level- α test) for $H_0 : \mu \leq 0$, or in other words, it is powerless against the point $\mu = 0$. However, one does exist if we omit 0 from the alternative — the aforementioned e-variable is again powered for $H_0 : \mu < 0$ against $H_1 : \mu > 0$.

Remark 3.4. Example 3.3 and its proof demonstrate a more general phenomenon. When testing a parametric model (like the one in the example), if the null hypothesis is an open subset of the parameter space, the alternative is its complement, and densities exist and are continuous in the parameter, then powered tests and e-variables do not exist.

Next, we discuss the natural connection between e-values and tests. Given any level- α binary test, one can always reproduce its decision by combining an “all-or-nothing e-value” with Markov's inequality, as we explain (and generalize) below.

Definition 3.5: All-or-nothing e-variable

An e-variable E is called an *all-or-nothing e-variable* if $E = \mathbb{1}_A/c$ for some event A and $c > 0$.

All-or-nothing e-variables take on only two values: zero or $1/c$ for some c . Further, it is clear that A must satisfy $\mathbb{P}(A) \leq c$ for all $\mathbb{P} \in \mathcal{P}$. For such e-variables, Markov's inequality at level c holds with equality, so there is no loss in translating between e-values, tests, and p-values.

Fact 3.6: Equivalence between tests and bounded e-variables

For any $\alpha \in (0, 1)$, there is a one-to-one correspondence between $[0, 1/\alpha]$ -valued e-variables and level- α tests. Specifically, for any $\alpha \in (0, 1]$, we can use the relationship

$$E = \phi/\alpha$$

to move between a level- α test ϕ and a corresponding e-variable E .

If ϕ is binary, then E is an all-or-nothing e-variable, which takes values in $\{0, 1/\alpha\}$. In particular, for any binary test ϕ that rejects the null hypothesis when a realized p-value is smaller than or equal to α , ϕ can be expressed in terms of the all-or-nothing e-variable $\mathbb{1}_{\{P \leq \alpha\}}/\alpha$.

Thus, there exists a powered level- α binary test if and only if there exists a powered e-variable taking values in $\{0, 1/\alpha\}$. We will expand this observation to include p-variables in Theorem 3.9.

As a consequence of Fact 3.6, all binary tests based on p-values — routinely used in the sciences — can be replicated using e-values. A non-binary test ϕ is usually provided with the interpretation that it outputs a

rejection with probability ϕ , realized as a number in $[0, 1]$. In other words, one can convert a non-binary test ϕ to a binary test ϕ' as $\phi' = \mathbf{1}_{\{U \leq \phi\}}$, where U is an independent uniform random variable on $[0, 1]$. This is closely related to the randomized Markov's inequality discussed in Section 2.4.

An e-variable E taking values in $[0, 1/\alpha]$ is associated with a level- α test $\phi = \alpha E$; in fact, it also yields a level- β test for any $\beta < \alpha$. A single e-variable yields a family of level- α tests (indexed by α), but in the other direction, every level- α test yields a different e-variable.

The above discussion highlights an important conceptual point — e-values suffice for testing: If a problem is testable (i.e., a powered test exists), then it is testable with e-values. This also applies to a notion of powered p-variables, although it is conceptually slightly more complicated, as we define below.

Definition 3.7: Powered p-values

For $\beta \in (0, 1]$, a p-variable P is *powered on* $[0, \beta]$ against $\mathcal{Q}$ if for every $\mathbb{Q} \in \mathcal{Q}$, we have $\mathbb{Q}(P \leq t) \geq t$ for all $t \in [0, \beta]$ and strict inequality holds for some $t \in [0, \beta]$. We say P is *powered* if the above holds for $\beta = 1$; this means that P is strictly stochastically smaller than a uniform random variable on $[0, 1]$ under each $\mathbb{Q} \in \mathcal{Q}$.

For practical purposes, it is more than sufficient to consider $\beta = 0.25$, for example, as rejections via p-values are always done at small levels. In the next section, we will connect powered tests, powered e-variables, and powered p-variables.

3.2 A unified existence result

We sometimes make the following simple assumption:

$$\begin{aligned} \text{There exists } U \stackrel{d}{\sim} U[0, 1] \text{ under every } \mathbb{P} \in \mathcal{P} \text{ and } \mathbb{Q} \in \mathcal{Q}, \\ \text{independent of any other random variables considered.} \end{aligned} \tag{U}$$

This assumption is very weak and is typically satisfied in all statistical applications.

Remark 3.8. To satisfy Assumption (U), it may be helpful to consider a lifted sample space $\Omega \times [0, 1]$ from Ω and product probability measures $\mathbb{P} \times \lambda$ and $\mathbb{Q} \times \lambda$ from $\mathbb{P}$ and $\mathbb{Q}$, where λ is the Lebesgue measure. Then, choose U with $U(\omega_1, \omega_2) = \omega_2$, and extend random variables X in the original space to the mappings $(\omega_1, \omega_2) \mapsto X(\omega_1)$, so that U is independent of the other random variables that we consider.

With this assumption, we can formally connect the existence conditions of powered e-variables, powered p-variables, and powered tests.

Theorem 3.9

Suppose that Assumption (U) holds. For testing $\mathcal{P}$ against $\mathcal{Q}$ (without restriction), the following are equivalent:

- (i) there exists a bounded powered e-variable;
- (ii) there exists a powered level- α test for some (or every) $\alpha \in (0, 1)$;
- (iii) there exists a powered level- α binary test for some (or every) $\alpha \in (0, 1)$;
- (iv) there exists a powered on $[0, \beta]$ p-variable for some (or every) $\beta \in (0, 1)$;
- (v) there exists a powered p-variable.

In the directions (i) $\Leftrightarrow$ (ii), (iii) $\Rightarrow$ (i) and (v) $\Rightarrow$ (i), we do not need (U).

Proof.

In items (ii)–(iv), there is a “some” version and an “every” version. When we prove such an item as a consequent, we use the “every” version. When we take such an item as a premise, we use the “some” version.

The equivalence (i) $\Leftrightarrow$ (ii) follows from Fact 3.6, noting that from a bounded powered e-variable E we can construct another powered e-variable $\lambda E + (1 - \lambda)$ for some $\lambda \in (0, 1)$ that has a smaller upper bound. The direction (iii) $\Rightarrow$ (i) is also straightforward from Fact 3.6.

Next, we will prove in the route that (i) $\Rightarrow$ (iv) $\Rightarrow$ (v) $\Rightarrow$ (i) $\Rightarrow$ (iii).

(i) $\Rightarrow$ (iv) and (i) $\Rightarrow$ (iii): Let E_0 be a bounded powered e-variable and take $\beta \in (0, 1)$. Since E_0 is bounded, there exists $\lambda \in (0, 1)$ such that $E := \lambda E_0 + (1 - \lambda)$ is bounded above by $1/\beta$. Clearly, E is an e-variable, and $\mathbb{E}^{\mathbb{Q}}[E] = \mathbb{E}^{\mathbb{Q}}[\lambda E_0 + (1 - \lambda)] > 1$ for every $\mathbb{Q} \in \mathcal{Q}$. Let $P = U/E$, where U is in (U) and independent of E . By Corollary 2.10, P is a p-variable. To see that it is powered on $[0, \beta]$, we can check that, for $t \in (0, \beta]$ and $\mathbb{Q} \in \mathcal{Q}$,

$$\mathbb{Q}(P \leq t) = \mathbb{Q}(U \leq tE) = \mathbb{E}^{\mathbb{Q}}[(tE) \wedge 1] = \mathbb{E}^{\mathbb{Q}}[tE] > t.$$

Thus, (iv) holds. To see (iii), a powered level- β binary test is $\mathbb{1}_{\{P \leq \beta\}}$ for every $\beta \in (0, 1)$.

(iv) $\Rightarrow$ (v): Let P be a powered on $[0, \beta]$ p-variable, U be in (U) independent of P , and P' be given by

$$P' = P\mathbb{1}_{\{P \leq \beta\}} + (\beta + (1 - \beta)U)\mathbb{1}_{\{P > \beta\}}.$$

Let $\mathbb{S} \in \mathcal{P} \cup \mathcal{Q}$. For $t \in (0, \beta]$, we have $\mathbb{S}(P' \leq t) = \mathbb{S}(P \leq t)$. For $t \in (\beta, 1)$, we have

$$\begin{aligned} \mathbb{S}(P' \leq t) &= \mathbb{S}(P \leq t \wedge \beta) + \mathbb{S}(\beta + (1 - \beta)U \leq t) \mathbb{S}(P > \beta) \\ &= \mathbb{S}(P \leq \beta) + \frac{t - \beta}{1 - \beta} (1 - \mathbb{S}(P \leq \beta)) \\ &= \mathbb{S}(P \leq \beta) \left(1 - \frac{t - \beta}{1 - \beta}\right) + \frac{t - \beta}{1 - \beta}. \end{aligned}$$

Note that if $\mathbb{S}(P \leq \beta) = \beta$ then $\mathbb{S}(P' \leq t) = t$ for $t \geq \beta$. Putting these together, we get that for $t \in (0, 1)$, $\mathbb{S}(P' \leq t) \leq t$ if $\mathbb{S} \in \mathcal{P}$, and $\mathbb{S}(P' \leq t) \geq t$ if $\mathbb{S} \in \mathcal{Q}$. Moreover, $\mathbb{S}(P' \leq t) > t$ for some $t \in (0, \beta]$ if $\mathbb{S} \in \mathcal{Q}$. Therefore, P' is a powered p-variable for $\mathcal{P}$ against $\mathcal{Q}$.

(v) $\Rightarrow$ (i): Let P be a powered p-variable, and

$$f : \mathbb{R}_+ \rightarrow \mathbb{R}_+, \quad f(x) = (2 - 2x)_+.$$

Note that $\int_0^1 f(x)dx = 1$ and f is a calibrator in Section 2.3. Define $E = f(P)$, which is an e-variable for $\mathcal{P}$. Moreover, for each $\mathbb{Q} \in \mathcal{Q}$, we have $\mathbb{E}^{\mathbb{Q}}[E] > 1$ by the definition of the powered p-variable P and the fact that f is strictly decreasing on $[0, 1]$. Therefore, E is bounded and powered against $\mathcal{Q}$.

Condition (U) is needed for implications involving constructing powered p-variables, because without (U), e.g., in the case of discrete sample spaces (see Example 3.10 below), one may not be able to construct a p-variable P that satisfies $\mathbb{P}(P \leq \alpha) \leq \alpha \leq \mathbb{Q}(P \leq \alpha)$ for all $\alpha \in (0, 1)$ and all $\mathbb{P}, \mathbb{Q} \in \mathcal{Q}$ even when powered e-variables exist. Basically, finite sample spaces are not suitable for constructing powered p-variables, but there could still be potentially useful (discrete) p-variables such as the one for the Bernoulli distributions in Section 1.6.

Example 3.10

Consider a finite sample space Ω and testing $\mathbb{P}$ against $\mathbb{Q}$ with $\mathbb{P} \neq \mathbb{Q}$ and $\mathbb{P}(\{\omega\}) > 0$ for all $\omega \in \Omega$ (as in the Bernoulli setting in Section 1.6). A powered e-variable is the likelihood ratio $\mathbb{Q}/\mathbb{P}$. Any p-variable P defined on Ω takes finitely many positive values. Therefore, $\mathbb{Q}(P \leq \alpha) = 0$ for $\alpha > 0$ small enough, showing that P cannot be $[0, \beta]$ powered against $\mathbb{Q}$ for any $\beta \in (0, 1]$.

Theorem 3.9 offers an equivalence between the existence conditions, but it is silent about how to check such existence for given $\mathcal{P}$ and $\mathcal{Q}$. This issue will be addressed in Section 3.4.

3.3 E-power and powered e-values

Given this high-level equivalence between tests and bounded e-values discussed in Fact 3.6, one may be tempted to ask: What is new here? Are we just simply working with tests in disguise? The answer is this: While all-or-nothing e-values are interesting conceptually in order to relate them to tests and p-values, these are not typically how one would construct e-values in practice. Recall the example from Section 1.6 on likelihood ratios. They will commonly have support on the entire positive real line (not just on two points), and never take on the value zero if $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to each other. To explain this, the following concept of e-power is useful.

Definition 3.11: E-power

The *e-power* of an e-variable E against an alternative $\mathbb{Q}$ is $\mathbb{E}^{\mathbb{Q}}[\log E]$, assuming that the expectation is well-defined. An e-variable E has *positive e-power* against $\mathcal{Q}$ if $\mathbb{E}^{\mathbb{Q}}[\log E] > 0$ for each $\mathbb{Q} \in \mathcal{Q}$. It has *uniformly positive e-power* against $\mathcal{Q}$ if $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log E] > 0$.

The e-power of an e-variable could equal ∞ or $-\infty$ (these still count as being well-defined). Indeed, whenever we make statements involving expected logarithms, we require them to be well-defined. The e-power (higher is better) of all-or-nothing e-variables is $-\infty$.

The concept of e-power is closely related to the Kelly criterion in betting and investment. The logarithm reflects the fact that in many applications, e-values are multiplied together. For a simplest example, consider a sequence of iid e-variables $E_1, E_2, \dots$ with finite e-power and the process $M = (M_t)_{t \in \mathbb{N}}$ defined by

$$M_t = \prod_{s=1}^t E_s, \quad t \in \mathbb{N}. \quad (3.1)$$

The e-power of E_1 against $\mathbb{Q}$ equals the *asymptotic growth rate* of M , that is, the $\mathbb{Q}$ -almost sure limit of $(\log M_t)/t$ as $t \rightarrow \infty$ (see Definition 7.4), justified by the strong law of large numbers. The asymptotic growth rate will be important in the methodology of testing by betting, formally treated in Chapter 7. When converting M_t into the level- α test $\mathbb{1}_{\{M_t \geq 1/\alpha\}}$ for $\alpha \in (0, 1)$, the role of e-power is clear from the following proposition.

Proposition 3.12: Relating e-power to power

Let $E_1, E_2, \dots$ be a sequence of e-variables that are iid under $\mathbb{Q}$, and M be given by (3.1). If $\mathbb{E}^{\mathbb{Q}}[\log E_1] > 0$, then $\mathbb{Q}(M_t > 1/\alpha) \rightarrow 1$ as $t \rightarrow \infty$ for every $\alpha > 0$. If $\mathbb{E}^{\mathbb{Q}}[\log E_1] < 0$, then $\mathbb{Q}(M_t > 1/\alpha) \rightarrow 0$ as $t \rightarrow \infty$ for every $\alpha > 0$.

The proof of Proposition 3.12 is a simple application of the strong law of large numbers.

Note that for each $t \in \mathbb{N}$, E_1 has positive e-power against $\mathbb{Q}$ if and only if M_t has positive e-power against $\mathbb{Q}$. Proposition 3.12 implies that in the simple iid setting, a positive e-power leads to asymptotic power-1 tests, and a negative e-power leads to asymptotic power-0 tests. An axiomatic justification of e-power is presented in Section 3.8, which uses an axiom based on an idea that is similar to Proposition 3.12.

Example 3.13

For testing $\mathbb{P} = N(0, 1)$ against $\mathbb{Q} = N(\mu, 1)$ with n observations, the likelihood ratio e-value in (1.1) is $E_n = \prod_{i=1}^n \exp(\mu X_i - \mu^2/2)$. Its e-power is thus $\mathbb{E}^{\mathbb{Q}}[\log E_n] = n\mathbb{E}^{\mathbb{Q}}[\mu X_1 - \mu^2/2] = n\mu^2/2$. This e-variable has positive e-power against $\mathbb{Q}$ and the e-power grows linearly in n . Indeed, these two properties hold for all likelihood ratio e-values with iid observations.

Let $\mathbb{Q}_\theta = N(\theta, 1)$ for $\theta \in \mathbb{R}$. Noting that $\mathbb{E}^{\mathbb{Q}_\theta}[\log E_n] = n(\mu\theta - \mu^2/2)$, the e-variable E_n has positive e-power against $\{\mathbb{Q}_\theta : \theta > \nu\}$ for $\nu \geq \mu/2$ (uniformly if $\nu > \mu/2$), and it does not have positive e-power against any $\mathbb{Q}_\theta$ with $\theta \leq \mu/2$.

The condition $\mathbb{E}^{\mathbb{Q}}[\log E] > 0$ in Definition 3.11 is equivalent to saying that the geometric mean of X is larger than one: $\exp \mathbb{E}^{\mathbb{Q}}[\log E] > 1$, which is a stronger constraint than requiring the (arithmetic) mean of E to be larger than 1, that is, $\mathbb{E}^{\mathbb{Q}}[E] > 1$ in Definition 3.1 (this can be seen from Jensen's inequality). Hence, if E has positive e-power, then E is a powered e-variable, but the reverse implication is not true in general. These two conditions are closely related. Indeed, for a simple alternative $\mathbb{Q}$, from any powered e-variable one can easily find an e-variable with positive e-power, through the transform $E \mapsto 1 - \lambda + \lambda E$ for some small $\lambda > 0$ that depends on E and $\mathbb{Q}$. This is formalized in the following result.

Theorem 3.14

Let $X \geq 0$ be a nonnegative random variable. For any $\mathbb{Q} \in \mathcal{M}_1$,

$$\mathbb{E}^{\mathbb{Q}}[X] > 1 \iff \mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda + \lambda X)] > 0 \text{ for some } \lambda \in [0, 1]. \quad (3.2)$$

In fact, either condition in (3.2) implies the existence of a *bounded* nonnegative Y with $\mathbb{E}^{\mathbb{Q}}[Y] > 1$ and $\mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda + \lambda Y)] > 0$. For a general $\mathcal{Q} \subseteq \mathcal{M}_1$, if X is bounded or $\mathcal{Q}$ is finite, then

$$\begin{aligned} \inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X] > 1 \\ \iff \inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda + \lambda X)] > 0 \text{ for some } \lambda \in [0, 1]. \end{aligned} \quad (3.3)$$

As a consequence, for a finite $\mathcal{Q}$, an e-variable with positive e-power against $\mathcal{Q}$ exists if and only if a powered e-variable against $\mathcal{Q}$ exists.

Proof.

The backward direction in all statements does not require any conditions, and it follows from a simple application of Jensen's inequality, which yields

$$0 < \mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda + \lambda X)] \leq \log \mathbb{E}^{\mathbb{Q}}[1 - \lambda + \lambda X] \implies \mathbb{E}^{\mathbb{Q}}[X] > 1.$$

For the forward direction, we first note that if $\mathcal{Q}$ is finite, then $\min_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X] > 1$ implies $\min_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X \wedge K] > 1$ for some $K \geq 1$. Hence, we can without loss of generality assume that X is bounded. It then remains to prove (3.3) in case X is upper bounded by some $K > 1$.

Write $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$ for $x \in \mathbb{R}$ and denote by $\delta = \inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X] - 1 > 0$. Next, fix any $\mathbb{Q} \in \mathcal{Q}$. Since $\mathbb{E}^{\mathbb{Q}}[(X-1)_+] - \mathbb{E}^{\mathbb{Q}}[(X-1)_-] = \mathbb{E}^{\mathbb{Q}}[X-1] \geq \delta > 0$ and $\mathbb{E}^{\mathbb{Q}}[(X-1)_-] \in [0, 1]$, we have

$$\mathbb{E}^{\mathbb{Q}}[(X-1)_+] \geq \mathbb{E}^{\mathbb{Q}}[(X-1)_-] + \delta \geq (1 + \delta)\mathbb{E}^{\mathbb{Q}}[(X-1)_-].$$

Fix $\varepsilon \in (0, 1)$ such that $(1 + \varepsilon)/(1 - \varepsilon) < 1 + \delta$, and let

$$\eta = \left(\frac{1}{1 + \varepsilon} - \frac{1}{(1 + \delta)(1 - \varepsilon)} \right) \delta > 0.$$

Since $\mathbb{E}^{\mathbb{Q}}[(X-1)_+] \geq \delta > 0$, we have

$$\begin{aligned} & \frac{\mathbb{E}^{\mathbb{Q}}[(X-1)_+]}{1+\varepsilon} - \frac{\mathbb{E}^{\mathbb{Q}}[(X-1)_-]}{1-\varepsilon} \\ & \geq \left(\frac{1}{1+\varepsilon} - \frac{1}{(1+\delta)(1-\varepsilon)} \right) \mathbb{E}^{\mathbb{Q}}[(X-1)_+] \geq \eta > 0. \end{aligned}$$

Note that $\log(1+x) \geq x/(1+\varepsilon)$ for $x \in [0, \varepsilon]$ and $\log(1+x) \geq x/(1-\varepsilon)$ for $x \in (-\varepsilon, 0)$, that is,

$$\log(1+x) \geq \frac{x_+}{1+\varepsilon} - \frac{x_-}{1-\varepsilon} \quad \text{for } x \in (-\varepsilon, \varepsilon).$$

Hence, for $\lambda \in (0, \varepsilon/K)$, implying $\lambda(X-1) \in (-\varepsilon, \varepsilon)$, we have

$$\mathbb{E}^{\mathbb{Q}}[\log(1+\lambda(X-1))] \geq \frac{\mathbb{E}^{\mathbb{Q}}[\lambda(X-1)_+]}{1+\varepsilon} - \frac{\mathbb{E}^{\mathbb{Q}}[\lambda(X-1)_-]}{1-\varepsilon} \geq \lambda\eta > 0.$$

Note that the choice of $(\lambda, \varepsilon, \eta)$ does not depend on $\mathbb{Q} \in \mathcal{Q}$. Hence, by taking an infimum over $\mathbb{Q} \in \mathcal{Q}$ we complete the proof.

The forward direction in (3.3) fails if $\mathcal{Q}$ is infinite and X is unbounded. Below, we give an example of an unbounded e-variable E satisfying $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[E] > 1$ (that is, uniformly powered against $\mathcal{Q}$), but

$$\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log(1-\lambda+\lambda E)] < 0$$

for all $\lambda \in (0, 1]$. Moreover, this example also illustrates that a bounded powered e-variable may not exist even if a uniformly powered e-variable exists.

Example 3.15

Let $\Omega = \mathbb{N}_0$, $\mathbb{P} = (1/2)\delta_0 + (1/2)\delta_1$ where δ_x is the point mass at x , and $\mathcal{Q} = \{\mathbb{Q}_n : n \in \mathbb{N}\}$, where $\mathbb{Q}_n = (1/n)\delta_n + (1-1/n)\delta_0$.

- (i) Let E be given by $E(n) = 2n$ for $n \in \mathbb{N}_0$. This gives $\mathbb{E}^{\mathbb{P}}[E] = 1$ and $\mathbb{E}^{\mathbb{Q}}[E] = 2$ for all $\mathbb{Q} \in \mathcal{Q}$. Thus, E is a uniformly powered e-variable, and it is unbounded. For any $\lambda \in (0, 1]$, we have

$$\mathbb{E}^{\mathbb{Q}_n}[\log(1-\lambda+\lambda E)] = \frac{1}{n} \log((1-\lambda)^{n-1}(1+(2n-1)\lambda)).$$

Since $(1-\lambda)^{n-1}(1+(2n-1)\lambda) < 1$ for n large enough, we know that $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log(1-\lambda+\lambda E)] < 0$ for all $\lambda \in (0, 1]$.

- (ii) There does not exist any bounded e-variable that is powered against $\mathcal{Q}$. To see this, for an e-variable E to be powered against $\mathbb{Q}_1$, we have $E(1) > 1$. For E to be powered against $\mathbb{Q}_n$, we need $(1/n)E(n) + (1-1/n)E(0) > 1$ for all $n \in \mathbb{N}$. Since E is bounded, this means $E(0) \geq 1$. Together with $E(1) > 1$, we have $\mathbb{E}^{\mathbb{P}}[E] > 1$, conflicting the fact that E is an e-variable for $\mathbb{P}$.

Note that here $\mathbb{P}$ does not lie in $\mathcal{Q}$, or even in its convex hull, but it does lie within its closed convex hull, where the closure is taken with respect to the total variation (TV) distance (see Theorem 3.17 below). In particular, by sending $n \rightarrow \infty$, $\mathbb{Q}_n$ approximates δ_0 arbitrarily well in TV. Averaging δ_0 with $\mathbb{Q}_1$ yields $\mathbb{P}$.

As hinted above, the e-power $\mathbb{E}^{\mathbb{Q}}[\log E]$ may not always be well-defined, due to the unbounded nature of the logarithm function. Moreover, there are some special (perhaps pathological) examples in which the

e-power, although well-defined, may be infinite, and does not have a useful interpretation. We give a stylized, unrealistic, example here.

Example 3.16

Let Y be a Pareto(1) random variable under $\mathbb{Q}$, that is, $\mathbb{Q}(Y > x) = 1/x$ for $x \geq 1$ (this distribution has an infinite mean), and the distribution of Y under $\mathbb{P}$ be such that $\mathbb{E}^{\mathbb{P}}[\exp(6 - Y/100)] \leq 1$. Note that for any two distributions F, G , we can always find $\mathbb{P}, \mathbb{Q}$ such that a random variable Y is distributed as F under $\mathbb{P}$ and G under $\mathbb{Q}$.

The e-variable E for $\mathbb{P}$ given by

$$E = \exp(6 - Y/100)$$

satisfies $\mathbb{E}^{\mathbb{Q}}[\log E] = \mathbb{E}^{\mathbb{Q}}[6 - Y/100] = -\infty$. We can easily compute

$$\mathbb{Q}(E > 100) > \mathbb{Q}(E > \exp(5)) = \mathbb{Q}(Y < 100) = 0.99.$$

Therefore, E has a high probability under $\mathbb{Q}$ to take large numbers (providing very strong evidence against the null), although its e-power is $-\infty$. If another e-variable E' is identically distributed to $1/E$ under $\mathbb{Q}$, then $\mathbb{Q}(E' > 1) = 1/600$ but $\mathbb{E}^{\mathbb{Q}}[\log E'] = \infty$. Thus, E' has more e-power than E (and any e-variables with finite e-power), which may be regarded as counter-intuitive. Nevertheless, if we have an infinite iid sequence $E'_1, E'_2, \dots$ following the same distribution as E' under $\mathbb{Q}$, then the product $M_n = \prod_{i=1}^n E'_i$ goes to ∞ $\mathbb{Q}$ -almost surely as $n \rightarrow \infty$, due to the law of large numbers applied to $(\log M_n)_{n \geq 1}$, and this is consistent with Kelly's criterion. A similar example in the sequential setting will be discussed in Section 7.11.

3.4 Conditions for the existence of powered e-variables

In this section, we address the existence of powered e-variables and p-variables by means of two characterization results. These results connect the existence of p-variables and e-variables to the structure of the null and alternative hypotheses, different from Theorem 3.9, which does not discuss how to check the existence. In these results, let L and M be two positive integers and we continue to assume (U) in Section 3.2.

The following result uses the total variation distance between sets of distributions, which we now define. For any two distributions $\mathbb{P}, \mathbb{Q}$, define

$$d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where the supremum is taken over all measurable sets A . Then, given two sets of distributions $\mathcal{P}, \mathcal{Q}$, we can define

$$d_{\text{TV}}(\mathcal{P}, \mathcal{Q}) = \inf_{\mathbb{P} \in \mathcal{P}, \mathbb{Q} \in \mathcal{Q}} d_{\text{TV}}(\mathbb{P}, \mathbb{Q}).$$

The following result is closely connected to a classic theorem of Kraft [1955], proved by Le Cam in private communication. The Kraft–Le Cam theorem states that for general $\mathcal{P}$ and $\mathcal{Q}$ (not necessarily finite) that have a common reference measure, the following holds: for any $\varepsilon \geq 0$, there exists a test ϕ such that $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\phi] - \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[\phi] > \varepsilon$ if and only if the total variation distance between the closed convex hulls of $\mathcal{P}$ and $\mathcal{Q}$ is greater than 2ε .

Theorem 3.17

For testing $\mathcal{P}$ against $\mathcal{Q}$ with a common reference measure, the following are equivalent:

- (i) there exists a test ϕ such that $\inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\phi] > \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[\phi]$;
- (ii) there exists a uniformly powered bounded e-variable E for $\mathcal{P}$ against $\mathcal{Q}$;
- (iii) $d_{\text{TV}}(\text{Conv}(\mathcal{P}), \text{Conv}(\mathcal{Q})) > 0$.

Proof.

The equivalence of (i) and (iii) is by the Kraft-Le Cam theorem. Statement (ii) immediately implies (i) because the definition of an e-variable implies that the right-hand side of (i) is bounded by one; since the e-variable is uniformly bounded, one can transform it to a test by dividing by this bound. So it only remains to establish that (i) implies (ii). First note that the right-hand side of (i) must be strictly larger than zero. This is because if ϕ equals zero $\mathbb{P}$ almost surely for any $\mathbb{P} \in \mathcal{P}$, then it also equals zero almost surely for any $\mathbb{Q} \in \mathcal{Q}$ because of the common reference measure assumption, making the strict inequality in (i) impossible. Thus, using the ϕ in (i), one can define

$$E = \frac{\phi}{\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[\phi]},$$

for which we see that (ii) holds, completing the proof.

If there is no common reference measure, then condition (iii) is not equivalent to the earlier ones. In fact, even when $\mathcal{Q} = \{\mathbb{Q}\}$ is simple, $d_{TV}(\text{Conv}(\mathcal{P}), \mathbb{Q}) > 0$ does not imply the existence of a powered e-variable. The appropriate condition actually involves an object called the *bipolar* of $\mathcal{P}$, and is formally presented in Theorem 6.10. Separately, note that when $\mathcal{P}$ and $\mathcal{Q}$ are finite, condition (iii) above reduces to $\text{Conv}(\mathbb{Q}_1, \dots, \mathbb{Q}_M) \cap \text{Conv}(\mathbb{P}_1, \dots, \mathbb{P}_L) = \emptyset$.

The existence of exact p-variables and e-variables is technically more advanced, and studied in Section 14.3, where we obtain a similar result (Theorem 14.6) to Theorem 3.17.

3.5 Log-optimality and likelihood ratios

Next, we turn to the key concept of log-optimality for a simple alternative $\{\mathbb{Q}\}$, which is defined essentially via the comparison of e-power.

Definition 3.18: Log-optimality

When comparing two e-variables E_1 and E_2 for $\mathcal{P}$ against $\mathbb{Q}$, E_1 is said to have greater (resp. equal) e-power than E_2 if $\mathbb{E}^{\mathbb{Q}}[\log(E_2/E_1)] \leq 0$ (resp. $= 0$). Moreover, E is *log-optimal* if $\mathbb{E}^{\mathbb{Q}}[\log(E'/E)] \leq 0$ for any other e-variable E' .

Definitions 3.11 and 3.18 are obviously connected. If the maximum e-power, defined as

$$\max_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E],$$

is finite, then the log-optimal e-variable maximizes the e-power, and the definition of log-optimality reduces to saying $\mathbb{E}^{\mathbb{Q}}[\log E'] \leq \mathbb{E}^{\mathbb{Q}}[\log E]$. The current definition allows the maximum e-power to be infinite, which could occur even in simple cases such as testing a Gaussian against a Cauchy because the tails of the Cauchy are much heavier than that of the Gaussian; see also Example 3.16. Note that $\mathbb{E}^{\mathbb{Q}}[\log(E'/E)] \leq 0$ and $\mathbb{E}^{\mathbb{Q}}[\log(E')] \leq \mathbb{E}^{\mathbb{Q}}[\log(E)]$ are equivalent when E and E' have finite e-power, but the two conditions are not equivalent in general.

Proposition 3.19

For testing any $\mathcal{P}$, a log-optimal e-variable against $\mathbb{Q}$ always exists and is $\mathbb{Q}$ -almost surely unique.

We do not prove the first part in this book, since the proof is quite technical, and defer the second part to Chapter 6. However, in the case of a simple null hypothesis $\mathbb{P}$ with $\mathbb{Q} \ll \mathbb{P}$, existence and uniqueness of the log-optimal e-variable can be easily argued, as we will see in Theorem 3.20.

In this section, we present an important result on the log-optimal e-values for a simple null hypothesis $\mathcal{P} = \{\mathbb{P}\}$ and a simple alternative $\mathcal{Q} = \{\mathbb{Q}\}$. We will assume that $\mathbb{Q} \ll \mathbb{P}$. Let q and p represent the densities of $\mathbb{Q}$ and $\mathbb{P}$ with respect to some reference measure $\mathbb{L}$, by which we can write $(d\mathbb{Q}/d\mathbb{P})(x) = q(x)/p(x)$.

The classic Neyman–Pearson lemma for hypothesis testing states that when testing a simple null $\mathbb{P}$ against a simple alternative $\mathbb{Q}$, the most powerful level- α test involves thresholding the likelihood ratio. We now present a result that can be seen as the e-analog of the Neyman–Pearson lemma. Importantly, we are not interested in designing tests (which can be obtained via Markov’s inequality as discussed in Chapter 2), but simply in designing e-variable that is log-optimal (as e-powerful as possible).

To prepare for the statement, define the Kullback–Leibler divergence between $\mathbb{Q}$ and $\mathbb{P}$ as

$$\text{KL}(\mathbb{Q}, \mathbb{P}) := \mathbb{E}^{\mathbb{Q}} \left[\log \frac{d\mathbb{Q}}{d\mathbb{P}} \right] = \int q(x) \log(q(x)/p(x)) \mathbb{L}(dx), \quad (3.4)$$

where it is defined as ∞ if $\mathbb{Q}$ is not absolutely continuous with respect to $\mathbb{P}$. Gibbs’ inequality, or the Shannon–Kolmogorov information inequality, states that the Kullback–Leibler divergence between any two distributions is nonnegative: $\text{KL}(\mathbb{Q}, \mathbb{P}) \geq 0$. Further, equality holds if and only if $\mathbb{P} = \mathbb{Q}$.

Theorem 3.20

For testing $\mathbb{P}$ against $\mathbb{Q}$ with $\mathbb{Q} \ll \mathbb{P}$, the ($\mathbb{P}$ -almost surely) unique log-optimal e-variable is the likelihood ratio $d\mathbb{Q}/d\mathbb{P}$. Thus, the optimal e-power equals $\text{KL}(\mathbb{Q}, \mathbb{P})$, which could be infinite.

Proof.

Let $\mathbb{S} \ll \mathbb{P}$ be an arbitrary distribution. By Proposition 2.12, $E := d\mathbb{S}/d\mathbb{P}$ covers all exact e-variables. For our setting of a point null $\mathbb{P}$, it suffices to consider exact e-variables E' in Definition 3.11, because any non-exact e-variable can be made strictly larger by multiplying it by a constant larger than 1 (the inverse of its expected value under $\mathbb{P}$).

We now argue that it suffices to consider $\mathbb{S} \ll \mathbb{Q}$. Indeed, if $\mathbb{S}$ is not absolutely continuous with respect to $\mathbb{Q}$, then we can take its absolutely continuous part $\mathbb{S}^*$, and define an e-variable $E^* := d\mathbb{S}^*/d\mathbb{P}$. Note that $E = E^*$ $\mathbb{Q}$ -almost surely, and thus they have the same e-power, but E^* is not exact, so we do not need to consider either E or E^* .

Thus it suffices to consider e-variables $E = d\mathbb{S}/d\mathbb{P}$, where $\mathbb{S} \ll \mathbb{Q} \ll \mathbb{P}$. By direct substitution,

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} \left[\log \frac{E}{d\mathbb{Q}/d\mathbb{P}} \right] &= \int q(x) \log \left(\frac{s(x)/p(x)}{q(x)/p(x)} \right) \mathbb{L}(dx) \\ &= - \int q(x) \log \left(\frac{q(x)}{s(x)} \right) \mathbb{L}(dx) \leq 0, \end{aligned}$$

where the inequality follows from the fact that $\text{KL}(\mathbb{Q}, \mathbb{S}) \geq 0$. This shows that $d\mathbb{Q}/d\mathbb{P}$ is log-optimal. Uniqueness follows from the fact that the inequality holds as an equality only when $\mathbb{S} = \mathbb{Q}$, so $E = d\mathbb{Q}/d\mathbb{P}$ $\mathbb{Q}$ -almost surely, and $\mathbb{E}^{\mathbb{P}}[d\mathbb{Q}/d\mathbb{P}] \leq \mathbb{E}^{\mathbb{P}}[E] \leq 1 = \mathbb{E}^{\mathbb{P}}[d\mathbb{Q}/d\mathbb{P}]$ forces E to be unique $\mathbb{P}$ -almost surely. This completes the proof of the first claim. The second claim follows by simply plugging in $q(x)/p(x)$ into the definition of e-power.

Example 3.21

Continue the setting of testing $\mathbb{P} = N(0, 1)$ against $\mathbb{Q} = N(\mu, 1)$ with n observations in Example 3.13. Consider the e-variable $E_n(\theta) = \prod_{i=1}^n \exp(\theta X_i - \theta^2/2)$ for some $\theta \in \mathbb{R}$ that is not necessarily equal to μ . Its e-power is $\mathbb{E}^{\mathbb{Q}}[\log E_n(\theta)] = n\mathbb{E}^{\mathbb{Q}}[\theta X_1 - \theta^2/2] = n(\theta\mu - \theta^2/2)$. Simple algebra yields that $\theta = \mu$ gives the largest e-power, confirming the implication of Theorem 3.20.

In Chapter 6, we present a more general result encompassing Theorem 3.20, which even relaxes the assumption $\mathbb{Q} \ll \mathbb{P}$.

3.6 Conditional log-optimality and log-optimal calibrators

In statistical applications, e-variables and p-variables are constructed based on available data, and thus they belong to a σ -algebra $\mathcal{G} \subseteq \mathcal{F}$ generated by the data. With this restriction, we need to consider the *log-optimal e-variable with respect to $\mathcal{G}$* , that is, an e-variable E measurable with respect to $\mathcal{G}$ that satisfies $\mathbb{E}^{\mathbb{Q}}[\log(E'/E)] \leq 0$ for all e-variables E' that are measurable with respect to $\mathcal{G}$. In what follows, for $\mathbb{P} \in \mathcal{M}_1$, $\mathbb{P}|_{\mathcal{G}}$ is the restriction of $\mathbb{P}$ to $\mathcal{G}$, that is, $\mathbb{P}|_{\mathcal{G}}(A) = \mathbb{P}(A)$ for $A \in \mathcal{G}$.

Proposition 3.22

Let $\mathcal{G}$ be a sub- σ -algebra of $\mathcal{F}$. If $\mathbb{Q} \ll \mathbb{P}$, then the log-optimal e-variable for testing $\mathbb{P}$ against $\mathbb{Q}$ with respect to $\mathcal{G}$ exists and is given ($\mathbb{P}$ -almost surely uniquely) by

$$E = \frac{d\mathbb{Q}|_{\mathcal{G}}}{d\mathbb{P}|_{\mathcal{G}}} = \mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid \mathcal{G} \right]. \quad (3.5)$$

The first equality in (3.5) remains true if we assume $\mathbb{Q}|_{\mathcal{G}} \ll \mathbb{P}|_{\mathcal{G}}$ instead of $\mathbb{Q} \ll \mathbb{P}$.

Proof.

The first equality in (3.5) follows directly from Theorem 3.20, by restricting the problem to the smaller measurable space $(\Omega, \mathcal{G})$. The second equality in (3.5) follows from, for any $A \in \mathcal{G}$,

$$\int_A \frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{P} = \mathbb{Q}(A) = \mathbb{Q}|_{\mathcal{G}}(A) = \int_A \frac{d\mathbb{Q}|_{\mathcal{G}}}{d\mathbb{P}|_{\mathcal{G}}} d\mathbb{P}|_{\mathcal{G}} = \int_A \frac{d\mathbb{Q}|_{\mathcal{G}}}{d\mathbb{P}|_{\mathcal{G}}} d\mathbb{P},$$

and therefore $d\mathbb{Q}|_{\mathcal{G}}/d\mathbb{P}|_{\mathcal{G}}$ is the conditional expectation of $d\mathbb{Q}/d\mathbb{P}$ on $\mathcal{G}$.

A direct consequence of Proposition 3.22 is the log-optimal calibrator. Suppose that we know a p-variable P is uniformly distributed on $[0, 1]$ under $\mathbb{P}$; thus it is exact. For a given simple alternative $\mathbb{Q}$, we can define the *log-optimal calibrator*, that is, a calibrator f such that

$$\mathbb{E}^{\mathbb{Q}}[\log(f(P)/g(P))] \leq 0$$

for any other calibrator g . In other words, $f(P)$ is the log-optimal e-variable among all possible e-variables that are calibrated from P .

Corollary 3.23

Let P be an exact p-variable testing $\mathbb{P}$ against $\mathbb{Q} \ll \mathbb{P}$. Assume that $\mathbb{E}^{\mathbb{P}}[d\mathbb{Q}/d\mathbb{P} \mid P]$ is decreasing in P . Then the log-optimal calibrator exists and it is given (almost everywhere uniquely) by

$$f(P) = \mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid P \right] \text{ and } f = 0 \text{ on } (1, \infty).$$

Proof.

Clearly, $\int_0^1 f(u) du = \mathbb{E}^{\mathbb{P}}[f(P)] = \mathbb{E}^{\mathbb{P}}[d\mathbb{Q}/d\mathbb{P}] = 1$, and f is decreasing, so it is a calibrator. Its optimality follows from Proposition 3.22.

3.7 Asymptotic log-optimality: methods of mixtures and plug-in

Consider now the case of testing a simple null hypothesis $\mathbb{P}$ against a composite alternative $\mathcal{Q}$. It is easiest to think about the case where all the distributions involved are iid product probability measures, in which case, with only a slight overload of notation, we can let $\mathbb{P}$ and $\mathbb{Q} \in \mathcal{Q}$ refer to the distributions of each individual observation, and let these be associated with densities p and q .

Let the data be represented by a vector X^n . A reasonable objective is that of *asymptotic log-optimality*. Here, the goal is to define a sequence of e-variables $E_n = E(X^n)$ such that for any $\mathbb{Q} \in \mathcal{Q}$,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}^{\mathbb{Q}^n}[\log E_n]}{n} \rightarrow \text{KL}(\mathbb{Q}, \mathbb{P}). \quad (3.6)$$

The right-hand side of the above expression is the e-power of an oracle that knows which alternative distribution is generating the data. If the above limit holds, our e-variable is able to asymptotically achieve the same e-power without this knowledge.

There are two broad classes of methods that do achieve these goals under weak assumptions (that we do not expand on here): the *mixture* and the *plug-in* methods.

Mixture method. Pick a mixture distribution ν over $\mathcal{Q}$, and define the e-variable

$$E_n = \int \prod_{i=1}^n \frac{q(X_i)}{p(X_i)} \nu(dq). \quad (3.7)$$

Note the order of integral and product: we are calculating the mixture over likelihood ratios. The key limitation to applicability of this method is that the mixture integral should be analytically or computationally easy to evaluate.

Plug-in method. Let $\mathbb{Q}_{i-1} \in \mathcal{Q}$ be a distribution of the i -th observation, picked based on the first $i-1$ observations, and let q_{i-1} be its density. Define

$$E_n = \prod_{i=1}^n \frac{q_{i-1}(X_i)}{p(X_i)}. \quad (3.8)$$

It is not hard to check that E_n is an e-variable.

It is possible to prove the asymptotic log-optimality (3.6) of the above methods under some rather weak assumptions, but we omit these details. We only make some informal remarks here, which are still useful for the practitioner. For asymptotic log-optimality of the mixture method, it is necessary that ν has full support over $\mathcal{Q}$; for parametric problems with finite dimensional alternatives, this condition is often sufficient as well. For the plug-in method, a sensible choice of $\mathbb{Q}_{i-1}$ is the posterior mean of $\mathbb{Q} \in \mathcal{Q}$ after seeing X^{i-1} , when starting with prior ν . If the mixture method with ν is asymptotically log-optimal, then one would typically expect the aforementioned plug-in method to also have the same property.

The mixture method is also useful for composite nulls. Here we give an explicit example without derivation.

Example 3.24: T-test

Consider the standard t-testing situation with $\mathcal{P} = \{N(0, \sigma^2) : \sigma > 0\}$ and $\mathcal{Q} = \{N(\mu, \sigma^2) : \sigma > 0, \mu \neq 0\}$ (elements of $\mathcal{P}$ and $\mathcal{Q}$ are the marginal distributions). Suppose we observe n iid observations $X_1, \dots, X_n$ from some $\mathbb{P} \in \mathcal{P} \cup \mathcal{Q}$. Let $S_n = \sum_{i=1}^n X_i$ and $V_n = \sum_{i=1}^n X_i^2$. Then for any $c > 0$,

$$\sqrt{\frac{c^2}{n+c^2}} \left(\frac{(n+c^2)V_n}{(n+c^2)V_n - S_n^2} \right)^{n/2}$$

is an e-variable for $\mathcal{P}$ that satisfies (3.6) for $\mathcal{Q}$. Note that this e-variable is scale-invariant, meaning that rescaling the data by a constant does not change the e-variable, and when $n = 1$, it equals 1 and has zero e-power.

While the following observation is not about the method of mixtures, it feels appropriate to point out here the surprising fact that there do exist non-constant e-variables that have positive e-power against some alternatives, even when $n = 1$. For example, for any $a \in \mathbb{R}$,

$$E_1 = |X_1| \exp((1 - (X_1 - a)^2)/2)$$

is one such e-variable for $\mathcal{P}$; note however that it is not scale-invariant. To see that E_1 is an e-variable, note that for all σ , the inequality $|x| \leq \sigma \exp(x^2/(2\sigma^2) - 1/2)$ holds. Thus, we see that under any $\mathbb{P} \in \mathcal{P}$, which equals $N(0, \sigma^2)$ for some unknown σ , we have

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[E_1] &\leq \mathbb{E}^{\mathbb{P}}[\sigma \exp(X_1^2/(2\sigma^2) - 1/2) \exp((1 - (X_1 - a)^2)/2)] \\ &= \mathbb{E}^{\mathbb{P}} \left[\frac{\exp(-(X_1 - a)^2/2)}{\sigma^{-1} \exp(-X_1^2/(2\sigma^2))} \right] = 1, \end{aligned}$$

where the final inequality holds by recognizing that we are evaluating the expectation of a likelihood ratio of $N(a, 1)$ to $\mathbb{P}$. For the distribution $\mathbb{Q} = N(a, 1)$, we can see $\mathbb{E}^{\mathbb{Q}}[E_1] = \mathbb{E}^{\mathbb{Q}}[|Z + a| \exp(1 - Z^2/2)]$, where Z follows a standard normal distribution under $\mathbb{Q}$, and hence $\mathbb{E}^{\mathbb{Q}}[E_1] > 1$ when a is large enough.

Next, we demonstrate a setting with non-iid data where the mixture method is helpful and natural.

Example 3.25: Changepoint e-value

Suppose we observe n ordered observations $X_1, \dots, X_n$. The null hypothesis equals the singleton $\{\mathbb{P}^n\}$ for some given $\mathbb{P}$. The alternative hypothesis equals $\mathcal{Q} = \{\mathbb{P}^{n-k} \mathbb{Q}^k\}$ for some unknown $k \in [n]$ and known $\mathbb{Q}$, meaning that $X_1, \dots, X_{n-k}$ are iid from $\mathbb{P}$, and the rest are iid from $\mathbb{Q}$. It is natural to consider a mixture over the unknown k , in particular forming the e-value

$$\frac{1}{n} \sum_{k=1}^n \prod_{i=n-k+1}^n \frac{q(X_i)}{p(X_i)},$$

based on uniform-mixture likelihood ratios.

In fact, the above example underlies the much more general, nonparametric, *sequential* change detection methodology of *e-detectors*, which is out of the scope of this book.

3.8 Axiomatic justification of the e-power

In this section we justify the definition of e-power in Definition 3.11 through an axiomatic approach.

Let $\mathcal{X}$ be the set of all nonnegative random variables that are bounded from above and away from 0. We will restrict our consideration of e-power on $\mathcal{X}$ to avoid issues with infinity, and it can be naturally extended to the set of all nonnegative random variables. As standard in the axiomatic approach in decision

theory, axioms become weaker when they are formulated on a smaller set of objects, making the axiomatic characterization results stronger. Therefore, our restriction to $\mathcal{X}$ is harmless.

Let $\mathcal{M}_1$ be the set of all probability measures on $(\Omega, \mathcal{F})$. We will fix the alternative hypothesis $\mathbb{Q} \in \mathcal{M}_1$, which is an atomless probability measure. The requirement of $\mathbb{Q}$ being atomless is only used to justify the existence of a non-degenerate iid sequence.

In general, e-variables for the null hypothesis $\mathcal{P}$ may be arbitrarily distributed under the alternative hypothesis $\mathbb{Q}$, and the e-power is a concept concerning the performance of an e-variable under the alternative hypothesis. Therefore, the choice of the null hypothesis $\mathcal{P}$ is irrelevant for defining e-power.

For this reason, without loss of generality, we directly consider the candidates for the e-power as mappings $\Lambda^{\mathbb{Q}} : \mathcal{X} \rightarrow \mathbb{R}$, without specifying that e-variables need to satisfy $\mathbb{E}^{\mathbb{P}}[E] \leq 1$ for $\mathbb{P} \in \mathcal{P}$.

For $\Lambda^{\mathbb{Q}}$ to represent the e-power under $\mathbb{Q}$, the natural interpretation of $\Lambda^{\mathbb{Q}}(E_1) > \Lambda^{\mathbb{Q}}(E_2)$ is that E_1 is considered as more powerful than E_2 . We keep this important interpretation in mind in the following discussions.

Suppose that a statistician has designed a two-stage experiment, and observes an e-value of e_0 at the first stage. She needs to choose between two e-variables E_1 and E_2 at second stage. If E_1 is more powerful than E_2 , then she should choose E_1 over E_2 . However, the effective e-variables to compare in this procedure are $e_0 E_1$ and $e_0 E_2$. Hence, there should be a consistency between the pair (E_1, E_2) and the pair $(e_0 E_1, e_0 E_2)$, which is formalized by the following axiom.

Axiom 3.26: Homogeneity

If $\Lambda^{\mathbb{Q}}(E_1) \geq \Lambda^{\mathbb{Q}}(E_2)$, then $\Lambda^{\mathbb{Q}}(e_0 E_1) \geq \Lambda^{\mathbb{Q}}(e_0 E_2)$ for any constant $e_0 > 0$.

The second axiom concerns the design of e-variables in an asymptotic sense. It states that, to choose between two configurations of iid e-variables, if the product e-process of less powerful configuration has asymptotic power 1, then that from the more powerful configuration should also have asymptotic power 1. This axiom reflects the consideration of the Kelly criterion, which we discussed in Section 3.3.

Axiom 3.27: Asymptotic consistency

Suppose that $(E_n)_{n \in \mathbb{N}}$ and $(E'_n)_{n \in \mathbb{N}}$ are two iid sequences of nonnegative random variables under $\mathbb{Q}$. If $\Lambda^{\mathbb{Q}}(E_1) \geq \Lambda^{\mathbb{Q}}(E'_1)$, then for any $\alpha \in (0, 1)$,

$$\mathbb{Q} \left(\prod_{t=1}^n E'_t \geq \frac{1}{\alpha} \right) \rightarrow 1 \implies \mathbb{Q} \left(\prod_{t=1}^n E_t \geq \frac{1}{\alpha} \right) \rightarrow 1,$$

both as $n \rightarrow \infty$.

It turns out that Axioms 3.26 and 3.27 are sufficient to jointly characterize the e-power, which has the form $\mathbb{E}^{\mathbb{Q}}[\log E]$ as we saw in Definition 3.11.

Theorem 3.28

Given an atomless probability measure $\mathbb{Q} \in \mathcal{M}_1$, a mapping $\Lambda^{\mathbb{Q}} : \mathcal{X} \rightarrow \mathbb{R}$ satisfies Axioms 3.26 and 3.27 if and only if it can be represented by

$$\Lambda^{\mathbb{Q}}(E) = f(\mathbb{E}^{\mathbb{Q}}[\log E]), \quad E \in \mathcal{X}$$

for some strictly increasing function f .

Proof.

It is straightforward to verify that $E \mapsto \mathbb{E}^Q[\log E]$ satisfies the two axioms; Axiom 3.27 can be checked with the strong law of large numbers. Clearly, a strictly increasing transform does not matter for the two axioms. Hence, $E \mapsto f(\mathbb{E}^Q[\log E])$ also satisfies the two axioms.

Now we show the “only if” direction. We first suppose for the purpose of contradiction that there exist E, E' such that

$$\Lambda^Q(E) > \Lambda^Q(E') \text{ and } \mathbb{E}^Q[\log E] < \mathbb{E}^Q[\log E']. \quad (3.9)$$

Using Axiom 3.26, we can find some $c > 0$ such that

$$\Lambda^Q(cE) > \Lambda^Q(cE') \text{ and } \mathbb{E}^Q[\log(cE)] < 0 < \mathbb{E}^Q[\log(cE')]. \quad (3.10)$$

Using Axiom 3.27, for the iid sequence $(E_n)_{n \in \mathbb{N}}$ distributed as cE and the iid sequence $(E'_n)_{n \in \mathbb{N}}$ distributed as cE' , the strong law of large numbers and (3.10) together imply $\log \prod_{t=1}^n E_t \rightarrow -\infty$ and $\log \prod_{t=1}^n E'_t \rightarrow \infty$ in $\mathbb{Q}$. This violates Axiom 3.27. Hence, (3.9) does not hold. This implies

$$\Lambda^Q(E) > \Lambda^Q(E') \implies \mathbb{E}^Q[\log E] \geq \mathbb{E}^Q[\log E']. \quad (3.11)$$

Equivalently,

$$\Lambda^Q(E) \leq \Lambda^Q(E') \iff \mathbb{E}^Q[\log E] < \mathbb{E}^Q[\log E']. \quad (3.12)$$

Using (3.12), we know that $c \mapsto \Lambda^Q(cE)$ is increasing on $(0, \infty)$ for each $E \in \mathcal{X}$.

Suppose for the purpose of contradiction that there exist $E, E' \in \mathcal{X}$ such that

$$\Lambda^Q(E) > \Lambda^Q(E') \text{ and } \mathbb{E}^Q[\log E] = \mathbb{E}^Q[\log E']. \quad (3.13)$$

Using Axiom 3.26, we have from (3.13), for any $c > 0$, $\Lambda^Q(cE) > \Lambda^Q(cE')$. Note that for any $\lambda < 1$, we have, from (3.12), $\Lambda^Q(\lambda cE) \leq \Lambda^Q(cE')$, because $\mathbb{E}^Q[\log(\lambda cE)] < \mathbb{E}^Q[\log(cE')]$. Hence, $\lambda \mapsto \Lambda^Q(\lambda cE)$ is discontinuous at $\lambda = 1$ for all $c > 0$. This implies that $c \mapsto \Lambda^Q(cE)$ is discontinuous everywhere, a contradiction to the fact that this mapping is monotone and hence has bounded variation. Therefore, (3.13) does not hold, and by (3.11), we have

$$\Lambda^Q(E) > \Lambda^Q(E') \implies \mathbb{E}^Q[\log E] > \mathbb{E}^Q[\log E'].$$

This, together with (3.11), shows that Λ^Q and $X \mapsto \mathbb{E}^Q[\log X]$ generate the same ordinal relation, and hence the representation holds.

Theorem 3.28 implies that if Axioms 3.26 and 3.27 are reasonable for a notion of e-power, then the only possible way to quantify the e-power is to use a strictly increasing transform of $\mathbb{E}^Q[\log E]$, which generates the same ordinal structure as using $\mathbb{E}^Q[\log E]$ directly. It also shows what the e-power essentially guarantees: It is the most suitable when considering the product e-value with asymptotic consistency (also discussed in Section 3.3). A simple example of a strictly increasing transform of the e-power is the geometric mean of the e-variable E , given by $\exp \mathbb{E}^Q[\log E]$. It is nonnegative, and it is larger than 1 if the e-power is positive.

In the above result, the e-power $\mathbb{E}^Q[\log E]$ in Definition 3.11 is justified as a comparative notion. This serves a very similar purpose to the classic notion of power of tests (Definition 3.1) when we compare e-variables to determine which ones are more “powerful”. When it comes to the meaning of the actual value, its natural interpretation is the asymptotic growth rate in a multiplicative setting, which is different from the classic notion of power, whose interpretation is the probability of rejection under the alternative.

We can easily extend the mapping Λ^Q to the set of all nonnegative random variables, provided that $\mathbb{E}^Q[\log E]$ is well-defined, that is, at least one of $\mathbb{E}^Q[(\log E)_+]$ and $\mathbb{E}^Q[(\log E)_-]$ is finite.

In what follows, we let

$$\Lambda^Q : X \mapsto \mathbb{E}^Q[\log X]$$

for any nonnegative random variable X . We treat $\Lambda^{\mathbb{Q}}(X)$ as undefined if $\mathbb{E}^{\mathbb{Q}}[\log X]$ is not well-defined. The next proposition gives some convenient properties of $\Lambda^{\mathbb{Q}}$.

Proposition 3.29

Let E be an e-variable.

- (i) If $\Lambda^{\mathbb{Q}}(E) > 0$, then $\Lambda^{\mathbb{Q}}(1 - \lambda + \lambda E) > 0$ for all $\lambda \in (0, 1]$.
- (ii) If $\Lambda^{\mathbb{Q}}(E) \geq 0$, then $\Lambda^{\mathbb{Q}}(1 - \lambda + \lambda E) \geq 0$ for all $\lambda \in [0, 1]$.
- (iii) $\Lambda^{\mathbb{Q}}(1 - \lambda + \lambda E)$ is always well-defined for $\lambda \in [0, 1)$.
- (iv) If $\Lambda^{\mathbb{Q}}(E) < \infty$, then $\Lambda^{\mathbb{Q}}(1 - \lambda + \lambda E) \in \mathbb{R}$ for all $\lambda \in [0, 1)$.

Proof.

For (i), it suffices to note that $f : \lambda \mapsto \Lambda^{\mathbb{Q}}(1 + \lambda(E - 1))$ is concave with $f(0) = 0$ and $f(1) > 1$. This implies $f(\lambda) > 0$ for all $\lambda \in (0, 1]$. The case of (ii) is similar. Part (iii) follows by noticing that $1 - \lambda + \lambda E$ is bounded away from 0, and hence $\mathbb{E}^{\mathbb{Q}}[(\log(1 - \lambda + \lambda E))_-] < \infty$. Part (iv) follows from $\mathbb{E}^{\mathbb{Q}}[(\log(1 - \lambda + \lambda E))_+] \leq \mathbb{E}^{\mathbb{Q}}[(\log E)_+] < \infty$, which holds by noting $\mathbb{Q}(\log(1 - \lambda + \lambda E) > x) \leq \mathbb{Q}(\log E > x)$ for $x > 0$, together with the observation $\mathbb{E}^{\mathbb{Q}}[(\log(1 - \lambda + \lambda E))_-] < \infty$ in (iii).

Bibliographical note

The connection between tests and bounded e-variables described in Fact 3.6 was discussed in Koning [2024] in detail. The e-power in Section 3.3 is formally defined in Vovk and Wang [2024b]. Example 3.16 was discussed in Wang [2024a]. The notion of log-optimality was proposed by Kelly [1956] in the context of betting and investment. Larsson et al. [2025a] showed the existence and uniqueness of a log-optimal e-variable with no conditions on $\mathcal{P}, \mathbb{Q}$. The result in Section 3.5 essentially dates back to Breiman [1961], though its presentation here mirrors that in Shafer [2021]. The mixture and plug-in methods to handle composite alternatives, as presented in Section 3.7, are very well known; in fact, both ideas were already mentioned by Wald [1947]. Later, Robbins and Siegmund [1974] derived some strong connections between these methods. The t-test e-variable is derived and studied in, for example, Lai [1976], Pérez-Ortiz et al. [2024], and Wang and Ramdas [2025], with the latter paper having the e-variable for $n = 1$. The asymptotic optimality of the method of mixtures is strongly related to information-theoretic topics such as redundancy and description length, and many more such optimality results can be inferred using a plethora of results on the latter topic in Grünwald [2007]. The changepoint e-value in Example 3.25 is borrowed from Shin et al. [2024].

Chapter 4

Post-hoc testing and decision making with e-values

In this chapter, we study e-values in the context of post-hoc testing and decision making and compare them with p-values. The post-hoc setting allows for either the risk level, the possible set of decisions, or the risk constraint to depend on the data that are used for the decision. It also allows for future data collection and experiment design to depend on previously collected data. The main message is that e-values play a central role in the post-hoc testing and decision problems. We also note that several of the ideas presented here have been extended to *multiple* hypothesis testing in Section 9.6.

4.1 Testing at data-dependent error levels

Recall that a level- α test for $\mathcal{P}$ is defined to be a function ϕ such that $\mathbb{E}^{\mathbb{P}}[\phi] \leq \alpha$ for every $\mathbb{P} \in \mathcal{P}$. As presented so far, the type-I error level α has to be a predefined constant. However, e-variables do yield a generalized form of error control when testing at data-dependent error levels $\hat{\alpha}$ that could be arbitrarily dependent on the e-variable. A natural way to incorporate random $\hat{\alpha}$ is through a uniform generalization of the constraint $\mathbb{E}^{\mathbb{P}}[\phi/\alpha] \leq 1$, although there may be other ways to define post-hoc validity.

Definition 4.1: Post-hoc valid tests and p-values

A *post-hoc valid family of tests* for $\mathcal{P}$ is a collection of tests $(\phi_{\alpha})_{\alpha \in (0,1)}$ such that for every $\mathbb{P} \in \mathcal{P}$, and for every random variable $\hat{\alpha}$ taking values in $(0, 1)$, we have $\mathbb{E}^{\mathbb{P}}[\phi_{\hat{\alpha}}/\hat{\alpha}] \leq 1$. Equivalently, we require that

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{\phi_{\alpha}}{\alpha} \right] \leq 1 \quad \text{for all } \mathbb{P} \in \mathcal{P}.$$

A *post-hoc p-variable* for $\mathcal{P}$ is a nonnegative random variable P such that for every $\mathbb{P} \in \mathcal{P}$, and for every random variable $\hat{\alpha}$ taking values in $(0, 1)$, we have

$$\mathbb{E}^{\mathbb{P}} \left[\frac{\mathbb{1}_{\{P \leq \hat{\alpha}\}}}{\hat{\alpha}} \right] \leq 1.$$

Equivalently, we require that

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{\mathbb{1}_{\{P \leq \alpha\}}}{\alpha} \right] \leq 1, \tag{4.1}$$

for all $\mathbb{P} \in \mathcal{P}$.

For a p-variable P , a weaker condition than (4.1) holds:

$$\sup_{\alpha \in (0,1)} \mathbb{E}^{\mathbb{P}} \left[\frac{\mathbb{1}_{\{P \leq \alpha\}}}{\alpha} \right] \leq 1 \quad \text{for all } \mathbb{P} \in \mathcal{P}. \quad (4.2)$$

Indeed, (4.2) is precisely the definition of a p-variable P .

Remark 4.2. Let P be a p-variable for $\mathbb{P}$. The condition

$$\mathbb{E}^{\mathbb{P}} \left[\frac{\mathbb{1}_{\{P \leq T\}}}{T} \right] \leq 1 \quad (4.3)$$

also holds for some nonnegative random variables T , where $0/0$ is understood as 0. A sufficient condition for (4.3) on T is

$$p \mapsto \mathbb{P}(T \leq t \mid P \leq p) \text{ is increasing for all } t \in \mathbb{R}; \quad (4.4)$$

this defines a form of negative dependence between P and T . Some examples satisfying (4.4) include (a) T is a constant; (b) T is independent of P ; (c) T is a decreasing function of P . Note that condition (4.3) means that $\mathbb{1}_{\{P \leq T\}}/T$ is an e-variable. Moreover, (4.3) is implied by

$$p \mapsto \mathbb{P}(T \leq t \mid P = p) \text{ is (almost everywhere) increasing for all } t \in \mathbb{R},$$

which is a stronger condition.

Clearly, in Definition 4.1, each member ϕ_α of a post-hoc valid family of tests is itself a level- α test for any fixed $\alpha \in (0,1)$, and P is itself a p-variable. It is clear that for a post-hoc p-variable P , $(\mathbb{1}_{\{P \leq \alpha\}})_{\alpha \in (0,1)}$ is a post-hoc valid family of binary tests; this is not true for a usual p-variable. Moreover, it does not hurt to require a post-hoc valid family $(\phi_\alpha)_{\alpha \in (0,1)}$ of tests to be increasing (that is, ϕ_α is increasing in α), because replacing ϕ_α with $\sup_{\gamma \geq \alpha} \phi_\gamma$ does not affect its post-hoc validity.

We say that a post-hoc valid family $(\phi_\alpha)_{\alpha \in (0,1)}$ of tests is dominated by another family $(\phi'_\alpha)_{\alpha \in (0,1)}$ if $\phi_\alpha \leq \phi'_\alpha$ for all $\alpha \in (0,1)$. Intuitively, a dominated family is less useful than a dominating family, because the latter rejects at least as often as the former. In the next result, we will see testing with e-variables is essentially the only useful way of constructing post-hoc valid families of (binary) tests.

Proposition 4.3

Let E be an e-variable for $\mathcal{P}$. Then, $((\alpha E) \wedge 1)_{\alpha \in (0,1)}$ is a post-hoc valid family of tests and $(\mathbb{1}_{\{E \geq 1/\alpha\}})_{\alpha \in (0,1)}$ is a post-hoc valid family of binary tests.

Conversely, every post-hoc valid family of tests is dominated by $((\alpha E) \wedge 1)_{\alpha \in (0,1)}$ for some e-variable E for $\mathcal{P}$, and every post-hoc valid family of binary tests is dominated by $(\mathbb{1}_{\{E \geq 1/\alpha\}})_{\alpha \in (0,1)}$ for some e-variable E for $\mathcal{P}$.

Proof.

It is elementary to verify

$$\mathbb{1}_{\{x \geq 1\}} \leq x \wedge 1 \leq x \quad \text{for } x \geq 0, \quad (4.5)$$

which can be used to show Markov's inequality. Using (4.5), for any $\hat{\alpha}$ taking values in $(0,1)$, we have

$$\mathbb{E}^{\mathbb{P}} \left[\frac{\mathbb{1}_{\{E \geq 1/\hat{\alpha}\}}}{\hat{\alpha}} \right] \leq \mathbb{E}^{\mathbb{P}} \left[\frac{(\hat{\alpha} E) \wedge 1}{\hat{\alpha}} \right] \leq \mathbb{E}^{\mathbb{P}} \left[\frac{\hat{\alpha} E}{\hat{\alpha}} \right] = \mathbb{E}^{\mathbb{P}} [E] \leq 1.$$

This shows that both $(\mathbb{1}_{\{E \geq 1/\alpha\}})_{\alpha \in (0,1)}$ and $((\alpha E) \wedge 1)_{\alpha \in (0,1)}$ are post-hoc valid families of tests.

For the converse claims, let $(\phi_\alpha)_{\alpha \in (0,1)}$ be a post-hoc valid family of tests, and let $E = \sup_{\alpha \in (0,1)} \phi_\alpha / \alpha$. By the definition of post-hoc validity, E is an e-variable for $\mathcal{P}$. Moreover,

$\phi_\alpha \leq (\alpha E) \wedge 1$ by definition, showing the first converse claim. If members of $(\phi_\alpha)_{\alpha \in (0,1)}$ are binary, then $\{\phi_\alpha = 1\} \subseteq \{E \geq 1/\alpha\}$ for every $\alpha \in (0,1)$. Therefore, $\phi_\alpha \leq \mathbb{1}_{\{E \geq 1/\alpha\}}$, showing the desired dominance in the second converse claim.

It is important that in Proposition 4.3, the same e-variable E is used to construct all members in a post-hoc valid family of tests. The next result gives an intimate connection between post-hoc p-variables and e-variables.

Proposition 4.4

For a random variable P , the following are equivalent:

- (i) P is a post-hoc p-variable for $\mathcal{P}$;
- (ii) $P = 1/E$ on the event $P < 1$ for some e-variable E for $\mathcal{P}$;
- (iii) $P \geq (1/E) \wedge 1$ for some e-variable E for $\mathcal{P}$.

Proof.

To see (i) $\Rightarrow$ (ii), first note that $\mathbb{P}(P = 0) = 0$ for all $\mathbb{P} \in \mathcal{P}$, since P is a p-variable for $\mathcal{P}$. Let $E = (1/P)\mathbb{1}_{\{P < 1\}}$ and $\hat{\alpha} = P\mathbb{1}_{\{P < 1\}} + \gamma\mathbb{1}_{\{P \geq 1\}}$ for some $\gamma \in (0,1)$. We have that $\hat{\alpha}$ is $(0,1)$ -valued ($\mathbb{P}$ -almost surely for every $\mathbb{P} \in \mathcal{P}$) and $\{P < 1\} = \{P \leq \hat{\alpha}\}$. Therefore, for all $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}^{\mathbb{P}}[E] = \mathbb{E}^{\mathbb{P}} \left[\frac{1}{P} \mathbb{1}_{\{P < 1\}} \right] = \mathbb{E}^{\mathbb{P}} \left[\frac{1}{P\mathbb{1}_{\{P < 1\}}} \mathbb{1}_{\{P \leq \hat{\alpha}\}} \right] = \mathbb{E}^{\mathbb{P}} \left[\frac{1}{\hat{\alpha}} \mathbb{1}_{\{P \leq \hat{\alpha}\}} \right] \leq 1,$$

where the last inequality follows from the definition of the post-hoc p-variable. Hence, E is an e-variable, and clearly $P = 1/E$ when $P < 1$. The direction (ii) $\Rightarrow$ (iii) is immediate. Finally, the direction (iii) $\Rightarrow$ (i) follows from

$$\begin{aligned} \mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{\mathbb{1}_{\{P \leq \alpha\}}}{\alpha} \right] &\leq \mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{\mathbb{1}_{\{(1/E) \wedge 1 \leq \alpha\}}}{\alpha} \right] \\ &= \mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{\mathbb{1}_{\{1/E \leq \alpha\}}}{\alpha} \right] \leq \mathbb{E}^{\mathbb{P}}[E] \leq 1 \end{aligned}$$

for all $\mathbb{P} \in \mathcal{P}$, where the penultimate inequality is due to (4.5).

There is a small variation of the result in Proposition 4.4, further connecting post-hoc p-values and e-values. If we change the range of α from $(0,1)$ to $(0,\infty)$, that is, using the stronger requirement

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,\infty)} \frac{\mathbb{1}_{\{P \leq \alpha\}}}{\alpha} \right] \leq 1 \quad \text{for all } \mathbb{P} \in \mathcal{P} \quad (4.6)$$

in the definition of a post-hoc p-variable P , then a post-hoc p-variable would be precisely $1/E$ for some e-variable E , following the same argument in the proof of Proposition 4.4.

From Proposition 4.4, it is immediate that $(1/E) \wedge 1$ is a post-hoc p-variable; so is $1/E$, which may take value larger than 1 (recall that we allow p-variables and post-hoc p-variables to take values larger than 1). Using the same result, an admissible (i.e., not dominated by another) post-hoc valid family $(\phi_\alpha)_{\alpha \in (0,1)}$ of binary tests can always be written as

$$\phi_\alpha = \mathbb{1}_{\{P \leq \alpha\}} \quad \text{for } \alpha \in (0,1),$$

for some post-hoc p-variable P . In other words, a post-hoc p-variable is the smallest level that the associated admissible post-hoc family of binary tests can reject the null hypothesis. P-values are sometimes said to equal the smallest level at which one could have rejected the null, *had that level been fixed in advance*. Our preceding discussion implies that the italicized text can be ignored for post-hoc p-variables, but in some scientific research it is also ignored for usual p-variables, leading to misuse or misinterpretation of p-values.

The connection between $1/E$ and post-hoc p-variables also illustrates that it is unfair to directly compare the value of $1/E$ for an e-variable with the value of a usual p-variable P , as the former offers a stronger type of guarantee: post-hoc validity in the sense of Definition 4.1.

To interpret this stronger validity, compare the situations that we observe an e-value $E = 25$ versus a p-value $P = 0.04$ on the same hypothesis. We can safely say (in the sense of Definition 4.1) that we reject the null hypothesis based on E at level 0.05 (and even at 0.04), but it would not be statistically valid to say that we can reject based on P at level 0.05, unless this level was prespecified before collecting and seeing data. Indeed, one cannot claim rejection at level 0.05 even if one observes $P = 0.0001$, if the level 0.05 is not prespecified. The procedure of first seeing P and then specifying the significance level is a statistical malpractice, but it appears often in published studies. For instance, in many papers in the social sciences, first p-values are computed (sometimes on many different configurations of regression analysis) and then they are marked with asterisk signs to indicate at which levels they can reject the corresponding null hypothesis. For such a purpose, e-values (or post-hoc p-values) are the right tool to use, instead of the usual p-values. To re-emphasize, an e-value equal to $1/a$ carries more information than a p-value equal to a for $a \in (0, 1)$; see Section 2.7 for more discussions on comparing these values.

4.2 Post-hoc decisions using p-values and e-values

In this section, we show that e-value-based decision rules naturally appear in the context of post-hoc decisions, in a similar way to Section 4.1 but with a different motivation and interpretation. We consider a framework of decision making that is more general than standard hypothesis testing, which concerns binary decisions of acceptance and rejection.

Consider a class of loss functions $L_b : \{0, 1\} \times D_b \rightarrow \mathbb{R}$ indexed by scenarios $b \in \mathcal{B} \subseteq \mathbb{R}$, where $D_b \subseteq \mathbb{N}_0$ represents the set of possible decisions. (It may be instructive to first consider $\mathcal{B} \subseteq \mathbb{N}_0$ for simplicity.) A scenario b refers to a possible set of decisions that may be known after seeing the data. For instance, a scientist may have a different set of possible recommendations to the policy makers when they see different data results (see Example 4.8 below). For the moment, we focus on a fixed scenario b , and the general case is treated in Section 4.2.3.

The value $L_b(0, d)$ represents the loss to the decision maker with decision $d \in D_b$ when the null hypothesis $\mathcal{P}$ is true; the value $L_b(1, d)$ represents the loss to the decision maker with decision $d \in D_b$ when the alternative hypothesis is true. The decision maker knows the loss functions and the scenario.

Without loss of generality, we assume that

1. $d \mapsto L_b(0, d)$ is strictly increasing, and
2. $d \mapsto L_b(1, d)$ is strictly decreasing.

This assumption is harmless because we can discard all decisions d' with $L_b(0, d') \geq L_b(0, d)$ and $L_b(1, d') \geq L_b(1, d)$ for some d , because d' will never be better than d . After discarding these decisions, we can rearrange the decisions according to an increasing order of $L_b(0, d)$. (When b varies, we can do this for each b .) Intuitively, one can interpret a larger d as a more aggressive decision, and the above assumption means that as d increases, there is more loss to the decision maker if the null is true, and there is more gain to the decision maker if the alternative is true (see Example 4.5 below).

The dataset is represented by a random vector X taking values in a set $\mathcal{Z}$ (e.g., $\mathbb{R}^n$), and $\delta(X) \in D_b$ is the decision made based on the data X . When we allow b to vary, δ will also depend on b . The rest of this section will show:

- For binary decision spaces and fixed scenarios, p-values can safely be used to make decisions.

- In non-binary decision problems or in settings where the scenario could be data-dependent, decisions based on p-values do not yield the desired risk guarantee, but those based on e-values do.

For the sake of illustration, we assume $\mathcal{Q} = \{\mathbb{Q}\}$ in this section, that is, a simple alternative. The problem can be easily generalized using a composite $\mathcal{Q}$, by considering a more general objective in the optimization, which is not essential to our discussion below.

4.2.1 Decision rules based on p-values and e-values

Given a fixed scenario $b \in \mathcal{B}$, the decision maker's problem is

$$\begin{aligned} & \text{to minimize} && \mathbb{E}^{\mathbb{Q}}[L_b(1, \delta(X))] \\ & \text{over} && \delta : \mathcal{Z} \rightarrow D_b \\ & \text{subject to} && \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[L_b(0, \delta(X))] \leq \gamma, \end{aligned} \tag{4.7}$$

where $\gamma > 0$ is a prespecified risk budget, and the constraint in (4.7) is called a risk constraint. The scenario b will be allowed to depend on the data later.

Example 4.5: Ternary significance decisions are common scientific practice

A practical example in scientific research in the above framework is to assign a label from “non-significant (NS), significant (S), very significant (VS)” to the test result after seeing data. In our notation, it is $D_b = \{0, 1, 2\}$, with 0 representing NS, 1 representing S, and 2 representing VS. If the null hypothesis is true, then making the decision NS has the least loss. If the alternative hypothesis is true, then making the decision VS has the least loss. A common choice is to set $\gamma = 1$, $L_b(0, 0) = 0$, $L_b(0, 1) = 1/\alpha_1$ and $L_b(0, 2) = 1/\alpha_2$ for some $0 < \alpha_2 < \alpha_1 < 1$, and the values $L_b(1, \delta)$ depend on the actual problem.

We define two potential approaches based on p-values and e-values. Suppose that a p-variable $P(X)$ and an e-variable $E(X)$ for $\mathcal{P}$ are computed from the data. The decision δ_P based on the p-value is given by

$$\delta_P(X) = \max\{d \in D_b : P(X)L_b(0, d) \leq \gamma\}. \tag{4.8}$$

The decision δ_E based on the e-value is given by

$$\delta_E(X) = \max\{d \in D_b : L_b(0, d) \leq \gamma E(X)\}. \tag{4.9}$$

Example 4.6: Significance decisions based on p-values and e-values

In the context of Example 4.5, the p-value-based decision rule in (4.8) yields $\delta_P(X) = 1$ (significant) if $\alpha_2 < P(X) \leq \alpha_1$ and $\delta_P(X) = 2$ if $P(X) \leq \alpha_2$ (very significant); otherwise $\delta_P(X) = 0$ (non-significant). This is a common practice in scientific research, but it does not satisfy the risk control in (4.7) as shown in Example 4.10 below.

The e-value-based decision rule in (4.9) yields $\delta_E(X) = 1$ if $1/\alpha_1 \leq E(X) < 1/\alpha_2$ and $\delta_E(X) = 2$ if $E(X) \geq 1/\alpha_2$; otherwise $\delta_E(X) = 0$. This decision rule always satisfies the risk control in (4.7); see Theorem 4.9.

4.2.2 Binary decisions in fixed scenarios

For now, consider the simplest case $D_b = \{0, 1\}$, that is, the case of binary decisions, where 0 represents the decision of accepting the null hypothesis, and 1 represents the decision of rejecting the null hypothesis. Without loss of generality, the loss function can be specified by $L_b(0, 0) = L_b(1, 1) = 0$ and $L_b(0, 1) = L_b(1, 0) = 1$ (this is without loss of generality because of the monotonicity assumed above). In this case, (4.7) becomes

$$\begin{aligned} & \text{to maximize} && \mathbb{Q}(\delta(X) = 1) \\ & \text{over} && \delta : \mathcal{Z} \rightarrow \{0, 1\} \\ & \text{subject to} && \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\delta(X) = 1) \leq \gamma, \end{aligned} \tag{4.10}$$

which is the standard problem of maximizing power while keeping type-I error control γ .

In the context of binary decisions in (4.10), the p-value-based decision rule means choosing $\delta_P(X) = 1$ if $P(X) \leq \gamma$ and choosing $\delta_P(X) = 0$ if $P(X) > \gamma$; thus, it is the standard way in which p-values are used in decision making (see also Example 4.6). The e-value-based decision rule means choosing $\delta_E(X) = 1$ if $E(X) \geq 1/\gamma$ and choosing $\delta_E(X) = 0$ if $E(X) < 1/\gamma$.

Proposition 4.7

Given a fixed scenario b with $D_b = \{0, 1\}$, both δ_P in (4.8) and δ_E in (4.9) satisfy the risk constraint in (4.10).

Proof.

To see that δ_P is feasible for problem (4.10), we note that for $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}^{\mathbb{P}}[L_b(0, \delta_P(X))] = \mathbb{P}(\delta_P(X) = 1) = \mathbb{P}(P(X) \leq \gamma) \leq \gamma.$$

To see that δ_E is feasible for problem (4.10), we note that for $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}^{\mathbb{P}}[L_b(0, \delta_E(X))] \leq \mathbb{E}^{\mathbb{P}}[\gamma E(X)] \leq \gamma,$$

as desired.

Since $1/E(X)$ is a p-variable (Section 2.1), the e-value-based approach is typically more conservative than the p-value-based approach. In Section 4.2.3, we will see that when D_b is not binary, δ_E maintains the risk control, whereas δ_P does not.

4.2.3 Non-binary decisions or data-dependent scenarios

In many real-world applications, the testing problem, specified by (L_b, D_b) , depends on the scenario b , which itself depends on the data through, for example, the result of a preliminary exploratory experiment, the evidence provided by the first half of the data, or the amount of collected data (which the scientist may not know before collecting them). A simple example of data-dependent scenarios with $\mathcal{B} = \{1, 2\}$ is given below.

As the scenario b depends on the data X , we represent it by a random variable B (which is a function of X), whose realization is observable to the decision maker. The decision rule δ now takes both realizations of X and B as its inputs, written as $\delta(X, B)$.

Example 4.8

After seeing data from a clinical trial experiment, if the data is promising, the scientist finds themselves in the situation $B = 1$, in which case they can choose not to vaccinate anyone ($\delta(X, 1) = 0$) or to vaccinate the highest-risk population ($\delta(X, 1) = 1$). If the vaccine turns out (unexpectedly) to be extremely effective (the case $B = 2$), then the second option ($\delta(X, 2) = 1$) changes to vaccinating the *full* population. Note that the decision sets D_1 and D_2 have different meanings, and in reality the scientist must choose from the decision set D_B . In this example, we may set $L_b(0, 0) = L_b(1, 1) = 0$ for $b \in \{1, 2\}$, $L_1(0, 1) = L_1(1, 0) = L_2(1, 0) = 1$ and $L_2(0, 1) = 10$. (To be pedantic, the first argument of the loss function still indicates whether the null is true or not (unknown to the scientist), while the second argument indicates the decision taken in the scenario in the subscript.) The scientist actually receives the loss $L_B(0, \delta(X, B))$ or $L_B(1, \delta(X, B))$, depending on the unknown state of nature.

The problem (4.7) becomes

$$\begin{aligned}
& \text{to minimize} && \mathbb{E}^{\mathbb{Q}}[L_B(1, \delta(X, B))] \\
& \text{over} && \delta : \mathcal{Z} \times \mathcal{B} \rightarrow \mathbb{N}_0; \delta(x, b) \in D_b \\
& \text{subject to} && \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{Q}}[L_B(0, \delta(X, B))] \leq \gamma,
\end{aligned} \tag{4.11}$$

The constraint and objective in (4.11) are formulated for the given random variable B . In case how B depends on the data is not known to the decision maker, one may consider a stronger constraint and a worst-case objective, that is,

$$\begin{aligned}
& \text{to minimize} && \mathbb{E}^{\mathbb{Q}} \left[\sup_{b \in \mathcal{B}} L_b(1, \delta(X, b)) \right] \\
& \text{over} && \delta : \mathcal{Z} \times \mathcal{B} \rightarrow \mathbb{N}_0; \delta(x, b) \in D_b \\
& \text{subject to} && \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}} \left[\sup_{b \in \mathcal{B}} L_b(0, \delta(X, b)) \right] \leq \gamma.
\end{aligned} \tag{4.12}$$

The objectives to minimize in (4.11) and (4.12) are not important in our discussions, and they can be replaced by any objective that is decreasing in $\delta(X, b)$. The risk constraint is our focus.

In the setting of this section, the p-value-based decision δ_P and the e-value-based decision δ_E are given by (4.8) and (4.9) while treating b as an additional input variable; so we write them as $\delta_P(X, b)$ and $\delta_E(X, b)$ going forward. Note that for each $b \in \mathcal{B}$, the definition of δ_E guarantees

$$L_b(0, \delta_E(X, b)) \leq \gamma E(X). \tag{4.13}$$

Theorem 4.9

For the problems (4.11) and (4.12), the following hold.

- (i) The e-value-based decision δ_E for any $E(X)$ satisfies the risk constraints in (4.11) and (4.12).
- (ii) Any decision rule δ^* satisfying the risk constraint in (4.11) can be written as $\delta^*(X, B) = \delta_E(X, B)$ for some e-variable $E(X)$.
- (iii) Any decision rule δ^* satisfying the risk constraint in (4.12) satisfies $\delta^*(X, b) \leq \delta_E(X, b)$ for some e-variable $E(X)$ and all $b \in \mathcal{B}$, meaning that δ^* is weakly improved by utilizing e-variables.
- (iv) The p-value-based decision δ_P does not satisfy the risk constraint in (4.11) or (4.12) in general when D_b has at least 3 elements or $\mathcal{B}$ has at least 2 elements.

Implied by (ii) and (iii), if an optimizer δ^* of (4.11) or (4.12) exists, then it is dominated by some δ_E .

Proof.

- (i): This follows directly follows from (4.13), which yields, for $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{b \in \mathcal{B}} L_b(0, \delta_E(X, b)) \right] \leq \mathbb{E}^{\mathbb{P}}[\gamma E(X)] \leq \gamma,$$

which guarantees the risk constraints in (4.11) and (4.12).

- (ii): Take $E(X) = L_B(0, \delta^*(X, B))/\gamma$. Then $E(X)$ is an e-variable because δ^* satisfies the risk constraint in (4.11). We have

$$\delta_E(X, b) = \max \{d \in D_b : L_b(0, d) \leq L_B(0, \delta^*(X, B))\}.$$

This gives $\delta_E(X, B) = \delta^*(X, B)$ under the assumption of a strictly increasing $d \mapsto L_b(0, d)$.

(iii): Take $E(X) = \sup_{b \in \mathcal{B}} L_b(0, \delta^*(X, b)) / \gamma$. Then $E(X)$ is an e-variable because δ^* satisfies the risk constraint in (4.12). We have

$$\delta_E(X, b) = \max \left\{ d \in D_b : L_b(0, d) \leq \sup_{b' \in \mathcal{B}} L_{b'}(0, \delta^*(X, b')) \right\}.$$

This gives $\delta_E(X, b) \geq \delta^*(X, b)$ under the assumption of a strictly increasing $d \mapsto L_b(0, d)$.

(iv): Counter-examples are provided in Example 4.10 below.

Example 4.10: The risk constraint when $|D_b| \geq 3$ or $|\mathcal{B}| \geq 2$

We only show how δ_P breaks the constraint in (4.11), as that in (4.12) is more strict. Fix $\mathbb{P} \in \mathcal{P}$ and let $P(X)$ be uniformly distributed on $[0, 1]$ under $\mathbb{P}$. First, consider $\mathcal{B} = \{1, 2\}$ and $D_b = \{0, 1\}$ for $b \in \mathcal{B}$. Set $L_1(0, 1) = \ell_1$ and $L_2(0, 1) = \ell_2$ with $\ell_1 < \ell_2$, let $B = 2$ when $P(X) \leq \gamma/\ell_2$. We can compute

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[L_B(0, \delta_P(X))] &= \ell_1 \mathbb{P}(\delta_P(X) = 1; B = 1) + \ell_2 \mathbb{P}(\delta_P(X) = 1; B = 2) \\ &= \ell_1 \mathbb{P}(\gamma/\ell_2 \leq P(X) \leq \gamma/\ell_1) + \ell_2 \mathbb{P}(P(X) \leq \gamma/\ell_2) \\ &= \gamma \left(1 - \frac{\ell_1}{\ell_2} \right) + \gamma > \gamma. \end{aligned} \tag{4.14}$$

For an example in case $D_b = \{0, 1, 2\}$ and $\mathcal{B} = \{1\}$, set $L_1(0, 1) = \ell_1$ and $L_1(0, 2) = \ell_2$ with $\ell_1 < \ell_2$, as in Example 4.6. Following the same calculation as in (4.14), we have

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[L_1(0, \delta_P(X))] &= \ell_1 \mathbb{P}(\delta_P(X) = 1) + \ell_2 \mathbb{P}(\delta_P(X) = 2) \\ &= \ell_1 \mathbb{P}(\gamma/\ell_2 \leq P(X) \leq \gamma/\ell_1) + \ell_2 \mathbb{P}(P(X) \leq \gamma/\ell_2) \\ &> \gamma. \end{aligned}$$

Although the p-value-based decision does not control the risk level in (4.11), it is still often employed (misused) in real-world situations; see Examples 4.5 and 4.6. Sometimes the misuse is implicit or unconscious, as the decision maker may not realize that the decision problem has changed due to the collected data, or the fact that there are non-binary decisions (or both). Therefore, when using p-value-based methods such as the ones described in Examples 4.5 and 4.6, one should be cautious and mindful about what type of statistical guarantee they actually provide. This problem does not exist for the e-value-based rule.

To summarize the message from this section, when decisions are non-binary or the decision problem depends on the data, e-value-based decision rules are needed to satisfy the risk control, and the optimizers are based on e-values.

In the simple case of binary decisions, the frameworks of Sections 4.1 and 4.2 can be connected via setting $\gamma = 1$, and $L_b(0, \delta) = \mathbb{1}_{\{\delta \leq b\}}/b$ with $D_b = \mathcal{B} = (0, 1)$, where we slightly generalize the assumption $D_b \subseteq \mathbb{N}_0$; the latter assumption was needed to make sure that the maximum in (4.9) is attained. A post-hoc p-variable δ satisfies the risk constraint for all data-dependent b . Section 4.1 does not specify the loss function under the alternative hypothesis.

4.3 Optional continuation of experiments

Suppose that a scientist tests whether a coin is fair (the null hypothesis) or not. It is anticipated that if the coin is not fair, the bias could be around $\theta = 0.6$ (the alternative hypothesis). A short calculation shows that to guarantee type-I and type-II errors at 0.05, the optimal (by the Neyman–Pearson lemma)

likelihood ratio test would toss the coin 280 times and reject the null if at least 154 heads appear. However, on performing those coin tosses, suppose that they do not reject the null. This could happen by chance under the alternative hypothesis (a type-II error), or because the scientist had overestimated the strength of the alternative. Perhaps, they should have planned for a coin bias of $\theta = 0.55$, but it is too late: Such thoughts are futile, because the type-I error budget of the test has been exhausted and the null was not rejected. However, a conceptually different approach using e-values is possible, as explained next.

The original test can be rewritten as rejecting the null when the p-value from the likelihood ratio test drops below 0.05. Suppose they had instead summarized the data as an e-value from the very beginning; the likelihood ratio would be a natural one in this case.

Based on the realized value of E_1 , the scientist (or another scientist) is permitted to toss the coin some more times, say collecting 50 more tosses. Suppose they calculate an e-value E_2 , based only on the fresh batch of data. In this case, note the important property of conditional expectation $\mathbb{E}^\mathbb{P}[E_2 \mid E_1] = 1$ under the null hypothesis. Then, $E^{(2)} = E_2 E_1$ is also a valid e-value for the same null, which can be verified by the law of iterated expectation. Note that both the option of running the second experiment and the design of that experiment can depend on the first batch of data, and thus it is a post-hoc decision.

The scientist may still not be happy. Perhaps $E^{(2)}$ is still promising and not conclusive. They have the right to continue their experiment and collect one more batch of data, summarize it as an e-value E_3 , and calculate $E^{(3)} = E_3 E^{(2)}$, which is itself an e-value for the same null.

The scientist is allowed to repeat this process indefinitely. They can continually monitor the evidence process $E^{(1)}, E^{(2)}, \dots$ and can reject the null as soon as their current level of evidence exceeds $1/\alpha$. This controls type-I error by Ville’s inequality (Chapter 7). Alternatively, they may simply stop at any data-dependent stopping time τ and report their evidence against the null as $E^{(\tau)}$, which is an e-value due to the optional stopping theorem (Chapter 7). When the size of each batch is reduced to one, one ends up with a fully sequential test, in this case Wald’s sequential probability ratio test that we revisit in Section 7.2.

However, despite the clear relationship to optional stopping and sequential testing above, the two settings (the one discussed above, and the one in Chapter 7) are motivated differently. Above, the scientists who may be continuing the study may be different from the first scientist that collected the first batch of data. This could happen in settings where data collection is more expensive than a coin toss; for example, the first scientist may simply have run out of money to collect more data, or may have given up if the evidence looked far from promising. Thus, we call the above setting as “optional continuation”, as opposed to the setting in Chapter 7, which we call “optional stopping”.

Such optional continuation is particularly prevalent in the meta-analysis of clinical trials. Suppose that Country B is hesitant to run an expensive trial on some drug (e.g., a vaccine for a pandemic). Country A goes ahead with such a trial, and the findings are impressive, but Country B deems it still too risky to launch the drug without its own trial. Piggybacking on the prior success of the drug, now Country B runs its own trial, and appends its evidence to that collected by Country A. Now Country C, which perhaps was not even considering the vaccine, may get convinced to run its own trial. The body of evidence against the null hypothesis (the drug being a placebo) is growing as more countries optionally continue this “living meta-analysis”. At this meta-level across countries, it does seem like a sequential meta-analysis is proceeding, but not one that was planned or anticipated in advance. Indeed, each country may have only intended to collect its own batch of data. As long as the evidence from each trial was summarized as an e-value, the overall evidence can easily be updated and sequentially monitored. For example, while wealthier countries may launch a vaccine program if the e-value exceeds 20, there may be several poorer countries that will wait till the overall e-value exceeds 100 before launching a vaccine program without their own dedicated trial.

There are three important caveats:

- No participants can discard or hide their findings, even if their findings lead to a small e-value (less than 1). After the decision of running the experiment, the resulting e-value needs to be incorporated into the evidence.
- Building on the above point, it is clear that a single scientist with an e-value equal to zero can permanently destroy the evidence accumulated thus far. This can be accounted for in two ways. Perhaps the scientists agree beforehand on using (e.g., log-optimal) e-values that are not equal to zero by design,

or they agree to an accumulation of evidence of the form $\prod_{i=1}^n (1 + E_i)/2$, so that each individual batch of data cannot lose more than half of the existing evidence; we can of course generalize this to $\prod_{i=1}^n (\lambda_i + (1 - \lambda_i)E_i)$, where $\lambda_i \in [0, 1]$ must be specified before seeing E_i . Either way, both the e-value being used in the i -th round and how to combine the i -th e-value with the previous ones must be specified before seeing the i -th batch of data.

- The decision to continue the collection of data at the second stage cannot be based on the first p-value not being smaller than 0.05 (because testing at the first stage with this p-value has used up all the type-I error budget). Either one commits to using a p-value with a single batch of data, or one commits to using an e-value with the possibility of extending the analysis.

It is worth emphasizing that it appears difficult to merge p-values that have been collected using the above protocol. While it may appear to be the case that second batch of data (say the trial data from Country B) is independent of the first batch of data (the trial data from Country A), it would be a mistake to summarize these as p-values and attempt to merge them using a standard method for combining independent p-values, like Fisher’s rule. The reason is that the second p-value is technically not independent of the first, despite being calculated on “separate” data. The second p-value would not even have existed had the first p-value not been promising. So the *number* of p-values is itself random, and the p-values are *sequentially dependent* in a complex way.

Such post-hoc decisions to run trials, or to make decisions based on accumulated data from different trials, is a key feature of e-value-based evidence collection and analysis.

4.4 Incentive-compatible principal-agent hypothesis testing

In this section, we consider another setup where e-values play a key role in enabling decision making with constraints.

Consider two parties, the *principal* and the *agent*. As a running example for the section, the principal can be thought of as a regulator, like a government agency that grants licenses, while the agent can be thought of as the experimenter, like a pharmaceutical company developing and testing various drugs. This setting has two unique features:

- The principal and agent have different amounts of information about the true state of the world/nature. For example, the agent may have a clearer picture of whether a drug is effective or not, or at least how likely it is to be effective.
- The principal and agent have different, typically conflicting, incentives. For example, the principal may not want ineffective drugs (placebos) to be falsely marketed as cures for some disease, while the agent may like to be granted licenses to make money (whether or not the drugs are effective).

Usually, a pharmaceutical company declares to the regulator that it is performing a clinical trial in order to determine the efficacy of its drug. For simplicity, let us ignore the multi-phase aspects and other details of such trials, and assume that if the p-value is smaller than some threshold (such as 0.05), then the regulating agency grants permission (a *license*) to the company to market the drug and profit from its sales.

The problem with the usual hypothesis testing paradigm is that it is not *incentive compatible*. A testing paradigm is incentive compatible if it discourages malicious agents — those who know the null hypothesis is true, for example, those who have a placebo drug and know it — from attempting to run such a trial. Indeed, suppose that a trial costs 10 million dollars, then consider the following calculation by a malicious agent:

- If the malicious agent estimates that they can make 100 million dollars by selling the drug, then they will not run the trial, because they will expect to lose 5 million dollars (100 million times 0.05, minus 10 million).
- If their anticipated profit is one billion dollars, then a malicious agent will certainly run such a trial, since their expected profit is now 40 million dollars.

If there are twenty times as many ineffective drugs as effective ones, and 5 percent of ineffective drugs are approved, then a quick calculation shows that at least half of all approved drugs will be ineffective.

Another problem with the above protocol is that the profit margin available to a malicious agent is only limited by the market size, and not by the strength of their statistical evidence. It seems reasonable to suggest a modification in which the more (less) evidence there is that a drug is effective, the more (less) an agent should be able to sell its drugs (or profit from its sale). We assume for simplicity that the profit/sales is determined by the regulator, not by the true effectiveness of the drug (which will be unknown to the market when sales begin, at least for a reasonable period of time, during which the market will assume that the drug is effective because it was approved by the regulator).

Thus, we may be interested in designing other interaction protocols between the principal and the agent, which have two natural properties:

- the protocol should be incentive compatible, in the sense that malicious agents should not have any incentive to participate.
- if an agent participates in the protocol, the sales/profit license increases with the strength of the statistical evidence.

We now show how this can be easily achieved with the use of a *statistical contract* based on e-values. We assume that the effectiveness of a drug can be captured by a quality parameter $\theta \in \Theta$ which is known to the agent, while the regulator deems that drugs with $\Theta_0 \subsetneq \Theta$ characterizes the placebos and would prefer not to have them be sold in the marketplace. The statistical contract is a protocol in which the agent follows the following steps in order:

1. Choose a nonnegative license function f from a menu of options $\mathcal{F}$.
2. Incur a fixed cost C to run a trial in order to collect data $Z \stackrel{d}{\sim} \mathbb{P}_\theta$.
3. Receive a license to profit at most $L = f(Z)$.

Under such a statistical contract, it is clear that the expected profit for the agent is $\mathbb{E}^{\mathbb{P}_\theta}[f(Z)] - C$, and we may for simplicity assume that they will participate in the trial if and only if this is positive. Thus, the regulator would like to ensure that this quantity is negative if $\theta \in \Theta_0$.

The above setup leads to the following result.

Proposition 4.11

The statistical contract is incentive compatible if and only if $f(Z)/C$ is an e-variable for $\{\mathbb{P}_\theta : \theta \in \Theta_0\}$ for every $f \in \mathcal{F}$.

The proof is immediate, from realizing that the statistical contract is incentive compatible if and only if $\mathbb{E}^{\mathbb{P}_\theta}[f(Z)] \leq C$ for all $\theta \in \Theta_0$.

As a particular example, if the evidence is going to be summarized as a p-value P , then it may be tempting to promise a license that profit margin that is inversely proportional to P . However, this would lead to an infinite expected profit under the null, and is not incentive compatible. In contrast, it is incentive compatible to award a license of $C/(2\sqrt{P})$, whose expected value is no more than C under the null. Indeed, $p \mapsto 1/(2\sqrt{p})$ is a p-to-e calibrator.

The setting in this section is connected to the one in Section 4.2 in the following way. There are two decision makers, making different types of decisions. The regulator is choosing the menu $\mathcal{F}$ to offer, and then the agent must choose whether to participate in the contract (run a trial).

1. Going first, the regulator has no particular objective to maximize, but they would like to meet their constraint of incentive compatibility, meaning that the menu must discourage malicious agents from participating. This is, as we have seen, achieved by using a menu composed of (scaled) e-values. Of course, if they hope to encourage agents with effective drugs to participate, the menu $\mathcal{F}$ should not be restricted in any other way: it should be the set of all (scaled) e-values. Assume this is how the regulator plays.

2. Then, the agent is choosing whether to participate in the statistical contract, given that their eventual payout is data-dependent (that is, they face a data-dependent loss function). They do not need to worry about incentive compatibility (that constraint has already been enforced by the regulator’s choice of $\mathcal{F}$), and they have to simply decide to participate or not before running the trial. Assuming that they know the ground truth and there exists some $f \in \mathcal{F}$ that is profitable (in expectation under the alternative), their decision is clear: they will not participate if they are malicious, and otherwise they may (depending on their assessment of the efficacy of the drug, and thus of their expected profit).

In summary, the decision problem for both agents considerably simplifies via a smart use of data-dependent loss functions based on the menu $\mathcal{F}$ of scaled e-values.

Bibliographical note

Section 4.1 combines results that are explicitly or implicitly found in Wang and Ramdas [2022], Xu et al. [2024], Koning [2023a], in the context of post-hoc α for testing or controlling the false discovery rate or false coverage rate. A notion of post-hoc p-values is defined in Koning [2023a] through (4.6), which is slightly different from our definition and coincides with the reciprocal of e-values. The sufficient condition in Remark 4.2 was proved by Blanchard and Roquain [2008].

Section 4.2 is based on Grünwald [2024], where some omitted details, admissibility of the e-value-based decision rules, and many examples can be found, but our results and presentation differ slightly.

Grünwald et al. [2024a] coined the term optional continuation in Section 4.3. This theme was further explored by Shafer [2023], who argues that it is easier to justify optional continuation using game-theoretic probability than using measure-theoretic probability (the latter relying on the Ionescu-Tulcea theorem for justifying extending the probability space from its initial specification). Ter Schure and Grünwald [2022] detailed an application of these ideas to meta-analysis.

The incentive-compatible principal-agent hypothesis testing problem in Section 4.4 was formulated and studied by Bates et al. [2022].

Part II

Core Ideas

Chapter 5

The universal inference e-value

Universal inference is a general method (or set of methods) to construct an e-variable (or an e-process) for a composite null $\mathcal{P}$ against a composite alternative $\mathcal{Q}$ under no *regularity conditions*. The simple takeaway message is: if we can efficiently calculate the maximum likelihood under the null, then we can construct an e-variable, and hence a test.

5.1 Motivation: Irregular models

The development of universal inference was motivated by the difficulty of constructing tests for *irregular problems*, for which we often know no (even asymptotically) valid test.

Without getting into formal definitions, because they will not be relevant later, irregular problems are usually plagued by many issues all at once: the maximum likelihood estimator may not be asymptotically normal, the Fisher information matrix may not always be invertible (especially on the boundary between null and alternative), the bootstrap may not have provable guarantees, Wilks' theorem may not hold (the log-likelihood ratio may not be asymptotically chi-squared), and so on. We provide two motivating examples here, for which, to the best of our knowledge, universal inference provides the first valid test.

We begin with the (conceptually, not technically) simple example of testing whether our data is Gaussian (null) or is drawn from a mixture of two Gaussians (alternative). Defining p as the density of $(1 - \lambda)\mathcal{N}(\mu_1, 1) + \lambda\mathcal{N}(\mu_2, 1)$, suppose we want to test the null hypothesis represented by the set $\{(\mu_1, \mu_2, \lambda) : \lambda = 0\}$, which also equals $\{(\mu_1, \mu_2, \lambda) : \mu_1 = \mu_2\}$. The alternative set of distributions is given by triplets (μ_1, μ_2, λ) that are unrestricted (except for excluding the null). For the null, μ_2 is not identified, and for the alternative, λ is not identified. This is a prototypical example of an irregular testing problem, with a composite null and alternative, and it is hard to find classical asymptotic procedures that can provably control the type I error. In contrast, the methodology presented here does work without any further assumptions, and has straightforward extensions to multidimensional settings, mixtures of more than two components, non-Gaussian distributions, and so on.

A second example involves testing shape constraints such as log-concavity. A density p on $\mathbb{R}$ is log-concave if $p = e^g$ for some concave function g . Consider testing $H_0 : p$ is log-concave versus $H_1 : p$ is not log-concave. The method presented in this chapter will be applicable to this problem, while we know of no other nontrivial test. Further, our method is also computationally efficient because the maximum likelihood log-concave density can be solved in polynomial time using a (relatively) efficient optimization algorithm.

As mentioned earlier, universal inference can be summarized as

a *reduction* from hypothesis testing to maximum likelihood estimation under the null.

Here, the word “reduction” means that the original problem of testing is reduced to an arguably simpler problem, that of (maximum likelihood) estimation.

We make two assumptions in this chapter:

1. $\mathcal{P}, \mathcal{Q}$ have a common reference measure $\mathbb{L}$, so we will associate distributions with their densities with respect to $\mathbb{L}$.
2. The data X are represented by a vector $X^n = (X_1, \dots, X_n)$ with iid entries. We identify distributions $\mathbb{P} \in \mathcal{P}$ and $\mathbb{Q} \in \mathcal{Q}$ by the corresponding densities of a single entry, denoted $p(x)$ or $q(x)$. For instance, we use $p \in \mathcal{P}$ and $\mathbb{P} \in \mathcal{P}$ interchangeably.

5.2 Split likelihood ratio e-variable

As suggested by its name, the key idea of the split likelihood ratio method is based on sample splitting. We divide the n data points into two groups D_0 and D_1 such that D_0 and D_1 are independent. For example, take $D_0 = \{X_i : i \in I\}$ and $D_1 = \{X_i : i \notin I\}$ where I is a random subset of $[n]$ independent from X^n . Even more precisely, for each point X_i , one could flip an independent coin (of some prespecified bias) to decide whether to put it in D_0 or in D_1 . One may think of these as being roughly equally sized sets as a default choice (i.e. the coin bias equals half). The validity of the e-variable does not depend on the relative sizes, just on their independence, but the e-power does. We also note that the split does not have to be random; one could define $D_0 = \{X_1, \dots, X_m\}$ for some $1 < m < n$ and $D_1 = \{X_{m+1}, \dots, X_n\}$.

The two sets have complementary roles. First, one uses the data in D_1 to pick an alternative $\hat{q}_1 \in \mathcal{Q}$. This choice is unrestricted: The maximum likelihood estimator, or a Bayes estimator (a posterior mean or mode based on some prior), or a robust one; the choice does not affect validity, but it does affect e-power. Our default recommendation in practice is to use a Bayes estimator with a prior that puts mass everywhere in $\mathcal{Q}$; this comes with certain asymptotic optimality properties, not discussed here. No particular properties of Bayesian inference are important here; one can simply view it as a smoothed maximum likelihood estimator where the smoothing particularly helps for small sample sizes.

The second step is to use D_0 to calculate the maximum likelihood estimator under the null, denoted $\hat{p}_0 = \arg \max_{p \in \mathcal{P}} \prod_{i \in D_0} p(X_i)$.

The final step uses D_0 again to calculate the likelihood ratio between $\hat{q}_1$ and $\hat{p}_0$:

$$E = \prod_{i \in D_0} \frac{\hat{q}_1(X_i)}{\hat{p}_0(X_i)}. \quad (5.1)$$

We call this the *split likelihood ratio e-variable*. When the roles of the two data splits are swapped, the split likelihood ratio E^{swap} is recomputed, then $E^{\text{cross-fit}} = (E + E^{\text{swap}})/2$ is called the *cross-fit likelihood ratio e-variable*. When combined with (some variant of) Markov's inequality (Chapter 2) to yield a level- α test, we call it the *split likelihood ratio test* (split LRT) or *cross-fit LRT*.

Theorem 5.1

The split likelihood ratio statistic defined in (5.1) is an e-variable for $\mathcal{P}$.

Proof.

By definition of $\hat{p}_0$, it holds that $\prod_{i \in D_0} \hat{p}_0(X_i) \geq \prod_{i \in D_0} p(X_i)$ for any $\mathbb{P} \in \mathcal{P}$, and thus

$$\mathbb{E}^{\mathbb{P}} \left[\prod_{i \in D_0} \frac{\hat{q}_1(X_i)}{\hat{p}_0(X_i)} \mid D_1 \right] \leq \mathbb{E}^{\mathbb{P}} \left[\prod_{i \in D_0} \frac{\hat{q}_1(X_i)}{p(X_i)} \mid D_1 \right] = 1,$$

where the final equality holds simply because D_0 and D_1 are independent, and conditional on D_1 , $\hat{q}_1$ is a fixed density, and we recall from Section 1.6 that a likelihood ratio of any alternative to p is an exact e-variable for p . The proof is completed by taking a further expectation with respect to D_1 and the tower property of conditional expectation.

Recalling the discussion from Section 3.7 about mixture and plug-in methods for handling composite alternatives, one may notice that the split likelihood ratio e-variable is based on the plug-in method. However, the techniques for handling composite nulls and composite alternatives are modular, in the sense that one can swap in the mixture method for the plug-in method in the numerator of these statistics, while leaving the denominator identical.

Remark 5.2 (Mixture universal inference). A variant of the split LRT uses D_1 to pick a distribution ν over $\mathcal{Q}$, and then defines

$$E = \int_{\mathcal{Q}} \prod_{i \in D_0} \frac{q(X_i)}{\hat{p}_0(X_i)} \nu(dq).$$

This is also an e-variable, which can be seen with an almost identical proof to above, by noting that integrals are just averages, and averages of e-variables are e-variables (more generally treated in Chapter 8).

5.3 Subsampled likelihood ratio e-variable

The split likelihood ratio statistic has an extra source of variability (beyond the randomness of the data) that was introduced algorithmically by sample splitting. It is possible to effectively remove this extra randomness by averaging, as we describe next. We could swap the roles of D_0 and D_1 , recalculate the e-variable using (5.1) (call it E'), and use $(E + E')/2$ as the e-variable. We call this the *crossfit likelihood ratio e-variable*. A better idea is to recalculate the split likelihood ratio e-variable B times, each time involving an identical and independent random split of X^n (these can be done in parallel). This yields e-variables $E^{(1)}, \dots, E^{(B)}$ that can be combined by averaging:

$$\bar{E}^{(B)} = \frac{E^{(1)} + \dots + E^{(B)}}{B}. \quad (5.2)$$

We call this the *subsampled likelihood ratio e-variable*. One can show that this (usually strictly) improves the e-power of the method, as proved below. This relies on noting that $E^{(1)}, \dots, E^{(B)}$ have the same e-power (since they are identically distributed).

Proposition 5.3: Improving e-power by averaging

Let $E^{(1)}, \dots, E^{(B)}$ be arbitrary e-variables, and let $\bar{E}^{(B)}$ denote their average. For any $\mathbb{Q}$,

$$\mathbb{E}^{\mathbb{Q}}[\log \bar{E}^{(B)}] \geq \frac{1}{B} \sum_{b=1}^B \mathbb{E}^{\mathbb{Q}}[\log E^{(b)}],$$

assuming that the right-hand side is well-defined. In particular, the e-power (against any $\mathbb{Q}$) of the subsampled likelihood ratio statistic (5.2) is at least as large as that of the split likelihood ratio statistic (5.1).

This proof follows immediately from concavity of the logarithm function. Further, as $B \rightarrow \infty$, $\bar{E}^{(B)}$ converges to a fixed random variable (see Fact 5.5).

Proposition 5.3 can be viewed as a form of Rao–Blackwellization. The split likelihood ratio statistic utilized an external source of randomness (“algorithmic randomness”) by data splitting, thus artificially introducing a source of variance, which is then removed by averaging.

Above, B has to be fixed in advance and cannot depend on the data. If one wants to obtain a p-value or a test from the subsampled likelihood ratio statistic, then it is possible for B to be data-dependent, as we now describe.

Subsampled LRT. For $b \geq 1$, let $\bar{E}^{(b)}$ denote $(E^{(1)} + \dots + E^{(b)})/b$. The subsampled LRT rejects the null as soon as any $\bar{E}^{(b)}$ exceeds $1/\alpha$ for any $b \geq 1$. Thus, the subsampled LRT can be run sequentially to produce $E^{(1)}, E^{(2)}, \dots$ one by one, from which one can calculate their running averages $\bar{E}^{(1)}, \bar{E}^{(2)}, \dots$ one by one, rejecting the null at the first time any of these exceeds $1/\alpha$. One can also terminate this procedure at any large, data-dependent B , without violating the type-I error guarantee.

Proposition 5.4: Improving power by averaging

The subsampled LRT described above controls type-I error at level α and is at least as powerful as the split LRT procedure that is based on $E^{(1)}$ and Markov's inequality.

The second claim about larger power is evident because the first step of the subsampled LRT corresponds exactly to the split LRT. The first claim about type-I error is less obvious, and is a direct consequence of the *exchangeable Markov inequality* introduced next.

5.4 Exchangeable Markov's inequality

We now state the generalization of the Markov inequality under exchangeability. All statements are made under a probability measure $\mathbb{P}$ that is implicit.

Recall that $Z_1, \dots, Z_n$ are called exchangeable if the joint distribution of $(Z_1, \dots, Z_n)$ equals that of $(Z_{\pi(1)}, \dots, Z_{\pi(n)})$ for any permutation π of $\{1, \dots, n\}$. A sequence $Z_1, Z_2, \dots$ is called exchangeable if $Z_1, \dots, Z_n$ are exchangeable for any $n \geq 1$.

For instance, iid random variables are exchangeable, and so are identical copies of the same random variable. Also, exchangeable random variables have the same marginal distribution and the same expectation.

Fact 5.5

For any exchangeable sequence $Z_1, Z_2, \dots$ of integrable random variables, their sample mean process $\bar{Z} = (\bar{Z}_n)_{n \in \mathbb{N}}$ given by $\bar{Z}_n := (Z_1 + \dots + Z_n)/n$ is a *backward martingale* with respect to the natural backward filtration in the sense that

$$\mathbb{E}[\bar{Z}_n | \mathcal{G}_{n+1}] = \bar{Z}_{n+1} \text{ for } n \in \mathbb{N}, \text{ where } \mathcal{G}_{n+1} = \sigma(\bar{Z}_{n+1}, \bar{Z}_{n+2}, \dots).$$

Further, $(\bar{Z}_n)_{n \in \mathbb{N}}$ converges almost surely to an integrable random variable.

The proof of the martingale property in Fact 5.5 follows by first noting by direct calculation that $\mathbb{E}[\bar{Z}_n | \bar{Z}_{n+1}] = \bar{Z}_{n+1}$, and the limiting claim is a consequence of the backward martingale convergence theorem of Doob [1953]. We are now in a position to state a useful nonasymptotic inequality.

Theorem 5.6: Exchangeable Markov inequality (EMI)

For any exchangeable sequence $Z_1, Z_2, \dots$ of random variables and any $\alpha > 0$,

$$\mathbb{P} \left(\sup_{n \geq 1} \left| \frac{1}{n} \sum_{i=1}^n Z_i \right| \geq \frac{1}{\alpha} \right) \leq \alpha \mathbb{E}[|Z_1|]. \quad (5.3)$$

In particular, if $Z_1, Z_2, \dots$ are e-variables,

$$\mathbb{P} \left(\exists n \geq 1 : \frac{1}{n} \sum_{i=1}^n Z_i \geq \frac{1}{\alpha} \right) \leq \alpha.$$

The above claims also hold for finite sets of exchangeable random variables.

The proof is short, albeit technical, and we only briefly mention it here without details. The first inequality is trivial. We next note that if Z_1 is not integrable, the desired inequality in (5.3) is trivially true. If Z_1 is integrable, $(\sum_{i=1}^n |Z_i|/n)_{n \geq 1}$ is a nonnegative backward martingale by Fact 5.5, and this allows us to infer the desired inequality by invoking a time-reversed variant of Ville's inequality (Fact 7.6 in Chapter 7).

A useful corollary of the EMI is as follows. Let $Z_1, \dots, Z_n$ be any set of (potentially nonexchangeable) arbitrarily dependent random variables. Let π be a uniformly random permutation of $\{1, \dots, n\}$. Then, for

any $\alpha > 0$,

$$\mathbb{P} \left(\sup_{1 \leq m \leq n} \frac{1}{m} \sum_{i=1}^m |Z_{\pi(i)}| \geq \frac{1}{\alpha} \right) \leq \alpha \frac{\mathbb{E} \mathbb{P}[|Z_1| + \dots + |Z_n|]}{n}.$$

Here, the original random variables are effectively made exchangeable by the random permutation, thus allowing us to invoke the original inequality.

We state another variant of the inequality. Suppose we take N arbitrarily dependent random variables and put them in a bag. Suppose $Z_{\pi(1)}, \dots, Z_{\pi(n)}$ are n samples drawn uniformly at random with or without replacement from this bag. Then, we have

$$\mathbb{P} \left(\sup_{1 \leq m \leq n} \frac{1}{m} \sum_{i=1}^m |Z_{\pi(i)}| \geq \frac{1}{\alpha} \right) \leq \alpha \frac{\mathbb{E} \mathbb{P}[|Z_1| + \dots + |Z_N|]}{N}.$$

This holds because the sampling process induces the exchangeability required for (5.3) to be invoked on the otherwise non-exchangeable random variables.

We end with a result that combines Theorem 2.9 with Theorem 5.6.

Theorem 5.7: Exchangeable Uniformly Randomized Markov's Inequality (EUMI)

Let $Z_1, \dots, Z_n$ be a set of exchangeable random variables. Then, for any $a \in (0, 1)$,

$$\mathbb{P} \left(Z_1 \geq U/a \text{ or } \exists t \leq n : \left| \sum_{i=1}^t Z_i/t \right| \geq 1/a \right) \leq a \cdot \mathbb{E} \mathbb{P}[Z_1],$$

where U is a uniform random variable on $[0, 1]$ that is independent of $Z_1, \dots, Z_n$.

Its proof is similar to that of EMI, and is omitted for brevity.

5.5 Universal confidence sets

It is useful, but not necessary, to adopt a parametric framework for what follows. Assume that the set of distributions under consideration can be represented as $\{p_\theta\}_{\theta \in \Theta}$. If we were testing, then the null and alternative hypotheses would correspond to certain subsets $\Theta_0 \subsetneq \Theta$ and $\Theta_1 \subsetneq \Theta$ respectively. But suppose instead that the underlying distribution of X^n is p_{θ^*} and we want to design a $(1 - \alpha)$ confidence set for θ^* , which is a set $C(X^n)$ such that for any $\theta^* \in \Theta$,

$$\mathbb{P}_{\theta^*}(\theta^* \in C(X^n)) \geq 1 - \alpha.$$

Such a confidence set, without requiring any regularity conditions, is immediately given by inverting the universal inference test. To elaborate, we simply test each point null hypothesis $\theta \in \Theta$ at level α , and retain those we failed to reject. Let us describe the set more formally by inverting the split LRT.

We first randomly split the data X^n into D_0 and D_1 . Let $\hat{\theta}$ be any estimator based only on D_1 . Define

$$C(X^n) := \left\{ \theta \in \Theta : \prod_{i \in D_0} \frac{p_{\hat{\theta}}(X_i)}{p_\theta(X_i)} < \frac{1}{\alpha} \right\}.$$

Denoting

$$\mathcal{L}_0(\theta) := \prod_{i \in D_0} p_\theta(X_i),$$

we see that $C(X^n) = \{\theta \in \Theta : \mathcal{L}_0(\hat{\theta})/\mathcal{L}_0(\theta) < 1/\alpha\}$. Then, we have the following guarantee.

Proposition 5.8

For any $\theta^* \in \Theta$ and any predefined error level $\alpha \in (0, 1)$, the set $C(X^n)$ defined above is a $(1 - \alpha)$ confidence set for θ^* .

The proof is immediate, relying only on the observation that $E = \prod_{i \in D_0} (p_{\hat{\theta}}(X_i)/p_{\theta^*}(X_i))$ is an e-variable, or more accurately that $\mathbb{1}_{\{E \geq 1/\alpha\}}$ is a level- α test. Sometimes, for instance for simply estimating a multivariate normal mean with a known covariance matrix, the above set is in closed form. Typically, the set can be made smaller (in terms of expected diameter) by replacing the split LRT by the subsampled LRT.

An example: multivariate Gaussian data with identity covariance

It is not obvious what the optimal splitting proportion is for split LRT based confidence sets. For iid Gaussian data with identity covariance in d dimensions, these sets are spherical, and one can derive explicit expressions for the proportion p_0^* of the data in $\mathcal{D}_0$ that minimizes the squared radius r^2 of $C(X^n)$.

Theorem 5.9

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta^*, I_d)$ under $\mathbb{P}$. The splitting proportion that minimizes $\mathbb{E}^\mathbb{P}[r^2(C(X^n))]$ is

$$p_0^* = 1 - \frac{\sqrt{4d^2 + 8d \log(1/\alpha)} - 2d}{4 \log(1/\alpha)}.$$

The proof is omitted.

As $d \rightarrow \infty$ for fixed α , the optimal split proportion p_0^* converges to 0.5. Alternatively, as $\alpha \rightarrow 0$ for fixed d , the proportion p_0^* converges to 1, suggesting that one should use nearly all data for the likelihood estimation. This is not an issue for reasonable α levels, though. For instance, at $d = 1$, one would need to set $\alpha < \exp(-40)$ to produce an optimal split proportion p_0^* that exceeds 0.90.

Let $C^{\text{LRT}}(X^n)$ denote the confidence set obtained by inverting the classical LRT (whose distribution is known exactly in this Gaussian setting). With an equal split, $\mathbb{E}[r^2(C(X^n))] = (4/n) \log(1/\alpha) + 4d/n$, and one can easily prove the following two bounds:

- As $d \rightarrow \infty$, $\mathbb{E}[r^2(C(X^n))]/r^2(C^{\text{LRT}}(X^n)) \rightarrow 4$,
- As $\alpha \rightarrow 0$, $\mathbb{E}[r^2(C(X^n))]/r^2(C^{\text{LRT}}(X^n)) \rightarrow 2$;

a more explicit expression is available for any d and α , that we visualize in Figure 5.10. One can also derive high probability bounds in the ratio of radii, that deliver the same message as above.

5.6 Extensions

5.6.1 Handling nuisance variables

Continuing with the theme of estimation, now consider the case where there is a nuisance parameter in both the null and the alternative. For example, suppose $\theta^* = (\theta_0^*, \theta_1^*)$ but we are only interested in a confidence set for θ_0^* . More generally, suppose we are interested in a confidence set for $\psi^* = g(\theta^*)$ for some given function g . Define $g^{-1}(\psi) = \{\theta : g(\theta) = \psi\}$. Now define

$$B(X^n) = \{\psi : C(X^n) \cap g^{-1}(\psi) \neq \emptyset\}.$$

Since C is a $(1 - \alpha)$ confidence set for θ^* , it is straightforward to check that B is a $(1 - \alpha)$ confidence set for ψ^* . But the above notation is a bit cumbersome. There is a more straightforward way to write this set. Define the *profile likelihood* on D_0 as

$$\mathcal{L}_0^\dagger(\psi) := \sup_{\theta: g(\theta) = \psi} \mathcal{L}_0(\theta),$$

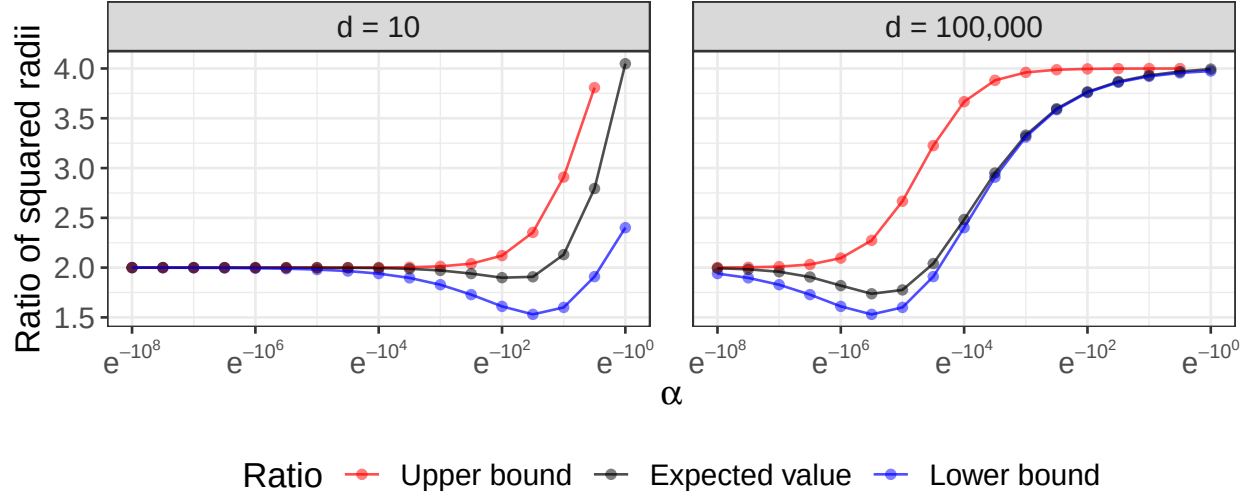


Figure 5.10: Expectation (black), lower bound (blue), and upper bound (red) of $\mathbb{E} [r^2(C(X^n))/r^2(C^{\text{LRT}}(X^n))]$. Data points correspond to values at $\alpha = \exp(-10^x)$ for x from 8 to 0 in increments of -0.5 .

Then, we can write

$$B(X^n) = \left\{ \psi : \frac{\mathcal{L}_0^\dagger(\psi)}{\mathcal{L}_0(\hat{\theta})} \geq \alpha \right\},$$

an expression which is possibly easier to work with.

5.6.2 Smoothed likelihood

Sometimes the maximum likelihood under the null may be infinite since the likelihood function is unbounded. One can then use a smoothed likelihood in its place. Consider a kernel $k(x, y)$ such that $\int k(x, y)dy = 1$ for any x . For any density p_θ , let its smoothed version be denoted

$$\tilde{p}_\theta(y) := \int k(x, y)p_\theta(x)dx,$$

Note that $\tilde{p}_\theta$ is also a probability density. Let the smoothed empirical density based on D_0 be defined as

$$\tilde{p}_n := \frac{1}{|D_0|} \sum_{i \in D_0} k(X_i, \cdot).$$

Define the smoothed maximum likelihood estimator as the KL projection of $\tilde{p}_n$ onto $\{\tilde{p}_\theta\}_{\theta \in \Theta_0}$:

$$\tilde{\theta}_0 := \arg \min_{\theta \in \Theta_0} \text{KL}(\tilde{p}_n, \tilde{p}_\theta).$$

If we define the smoothed likelihood on the first half of the data D_0 as

$$\tilde{\mathcal{L}}_0(\theta) := \prod_{i \in D_0} \exp \int k(X_i, y) \log \tilde{p}_\theta(y) dy,$$

then it can be checked that $\tilde{\theta}_0$ maximizes the smoothed likelihood, that is $\tilde{\theta}_0 = \arg \max_{\theta \in \Theta_0} \tilde{\mathcal{L}}_0(\theta)$. As before, let $\hat{\theta}_1 \in \Theta$ be any estimator based on D_1 . The smoothed split LRT is defined as:

$$\text{reject } H_0 \text{ if } \tilde{U}_n > 1/\alpha, \text{ where } \tilde{U}_n = \frac{\tilde{\mathcal{L}}_0(\hat{\theta}_1)}{\tilde{\mathcal{L}}_0(\tilde{\theta}_0)}.$$

We now verify that the smoothed split LRT controls type-I error, by simply checking that $\tilde{U}_n$ is an e-variable. Indeed, for any fixed $\psi \in \Theta$, we have (where $\mathbb{E}^\theta$ is evaluated with respect to p_θ)

$$\begin{aligned} \mathbb{E}^{\theta^*} \left[\frac{\tilde{\mathcal{L}}_0(\psi)}{\tilde{\mathcal{L}}_0(\tilde{\theta}_0)} \right] &\stackrel{(i)}{\leq} \mathbb{E}^{\theta^*} \left[\frac{\tilde{\mathcal{L}}_0(\psi)}{\tilde{\mathcal{L}}_0(\theta^*)} \right] \\ &= \prod_{i \in D_0} \int \exp \left(\int k(x, y) \log \frac{\tilde{p}_\psi(y)}{\tilde{p}_{\theta^*}(y)} dy \right) p_{\theta^*}(x) dx \\ &\stackrel{(ii)}{\leq} \prod_{i \in D_0} \int \left(\int k(x, y) \frac{\tilde{p}_\psi(y)}{\tilde{p}_{\theta^*}(y)} dy \right) p_{\theta^*}(x) dx \\ &= \prod_{i \in D_0} \int \left(\frac{\int k(x, y) p_{\theta^*}(x) dx}{\tilde{p}_{\theta^*}(y)} \right) \tilde{p}_\psi(y) dy \\ &= \prod_{i \in D_0} \int \tilde{p}_\psi(y) dy = 1, \end{aligned}$$

where inequality (i) is because $\tilde{\theta}_0$ maximizes the smoothed likelihood, and step (ii) follows by Jensen's inequality.

5.6.3 Relaxed, powered and conditional likelihood

There may be settings where computing the MLE and/or the maximum likelihood (under the null) is computationally hard. Suppose one could come up with a relaxation F_0 of the null likelihood $\mathcal{L}_0$ (or of the log-likelihood) in the sense that

$$\max_{\theta} F_0(\theta) \geq \max_{\theta} \mathcal{L}_0(\theta).$$

For example, $\mathcal{L}_0$ may be defined as $-\infty$ outside its domain, but F could extend the domain. As another example, instead of minimizing the negative log-likelihood which could be nonconvex and hence hard to minimize, we could instead minimize a convex relaxation. In such settings, define

$$\hat{\theta}_0^F := \arg \max_{\theta} F_0(\theta).$$

If we instead define the test statistic

$$T'_n := \frac{\mathcal{L}_0(\hat{\theta}_1)}{F_0(\hat{\theta}_0^F)},$$

then inference may proceed using T'_n instead of T_n in the split or crossfit LRT. This is simply because of the aforementioned property of the relaxation that $F_0(\hat{\theta}_0^F) \geq \mathcal{L}_0(\hat{\theta}_0)$, and hence $T'_n \leq T_n$.

One particular case when this would be useful is the following. Testing the sparsity level in a high-dimensional linear model corresponds to solving the best subset selection problem, which is NP-hard in the worst case (integer programming). There exist well-known quadratic programming relaxations that are far more computationally tractable.

The takeaway message is that *it suffices to upper bound the maximum likelihood under the null in order to perform inference*.

When model misspecification is a concern, inferences can be made robust by replacing the likelihood $\mathcal{L}$ with the power likelihood $\mathcal{L}^\eta$ for some $0 < \eta < 1$. Note that

$$\mathbb{E}^\theta \left[\left(\frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} \right)^\eta \mid D_1 \right] = \prod_{i \in D_0} \int p_{\theta_1}^\eta(y_i) p_\theta^{1-\eta}(y_i) dy_i \leq 1,$$

and hence all the aforementioned methods can be used with the power-robustified likelihood as well. (The last inequality follows because the η -Rényi divergence is nonnegative.)

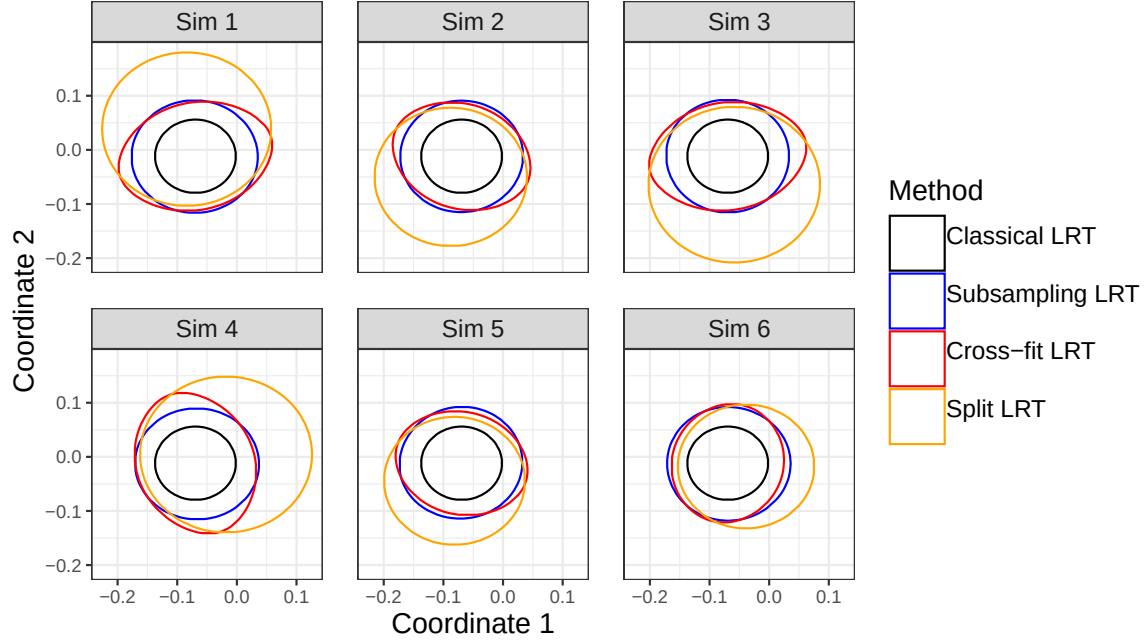


Figure 5.11: Coverage regions of classical LRT (black), subsampling LRT (blue), cross-fit LRT (red), and split LRT (orange) at $\alpha = 0.1$. The six simulations use the same 1000 observations from $N((0,0), I_2)$.

Our presentation has assumed that the data are drawn iid from some distribution under the null. However, this is not really required (even under the null), and was assumed for expositional simplicity. All that is needed is that we can calculate the likelihood on D_0 conditional on D_1 (or vice versa). For example, this could be tractable in models involving sampling without replacement from an urn with $M \gg n$ balls. Here θ could represent the unknown number of balls of different colors. Such hypergeometric sampling schemes obviously result in non-iid sampling, but conditional on the first half of the data (for example how many red, green and blue balls were sampled from the urn in that subset), one can potentially still evaluate the conditional likelihood of the second half of the data and maximize it, rendering it possible to apply our universal tests and confidence sets.

5.7 Numerical studies

Since it is easier to visualize confidence sets, Figure 5.11 shows coverage regions of the classical LRT, split LRT, cross-fit LRT, and subsampling LRT at $B = 100$ from six simulations with $\theta^* = (0,0)$. We generate 1000 observations from $N(\theta^*, I_2)$ once, and we use this sample for all simulations. Hence, the variation in the split, cross-fit, and subsampling LRTs across simulations is due to algorithmic randomness.

The coverage regions in Figure 5.11 suggest several relationships that can be formally proven. We see that the classical LRT provides the smallest confidence regions. This is not surprising since, even in finite samples, the classical LRT statistic follows a chi-squared distribution under the null in the Gaussian case. It can be proven that the volume of the cross-fit LRT set is less than or equal to the volume of the split LRT set, although the cross-fit set is not entirely contained within the split set. The split and cross-fit approaches both use a single split of the data, but there is a notable improvement from cross-fitting. One can prove that the subsampling set also has less volume than the split LRT set. While any individual split LRT region is guaranteed to be spherical, the subsampling set is not necessarily a spherical region. For large B , however, we see that the subsampling region is approximately spherical.

Next, Figure 5.12 shows the average squared radius of the split LRT confidence sets at p_0^* and at surrounding choices of p_0 . The expected squared radius, given by the red curve, is more sensitive to changes

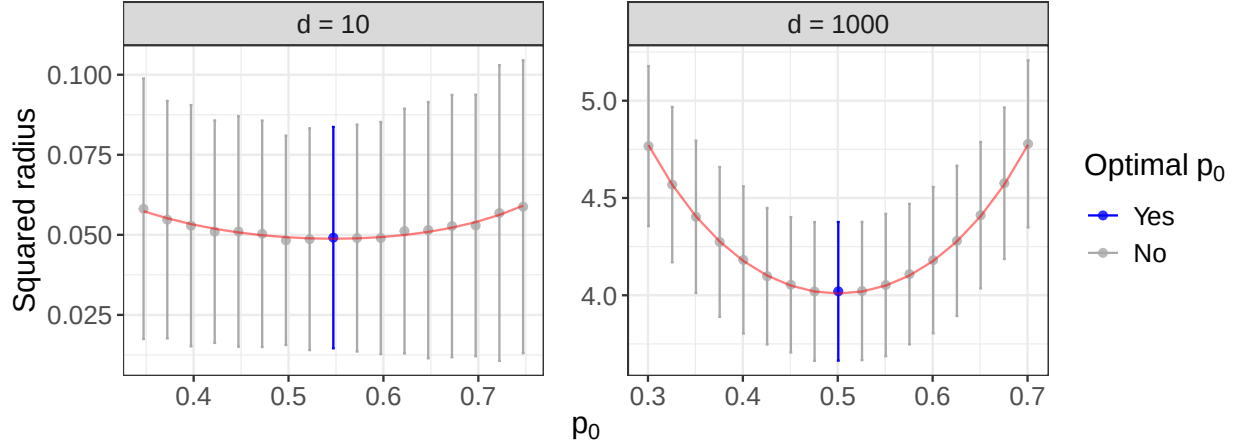


Figure 5.12: Squared radius of multivariate normal split LRT with varying p_0 . We simulate $X_1, \dots, X_{1000} \stackrel{d}{\sim} N(0, I_d)$ and compute the split LRT region at $\alpha = 0.10$ and varying p_0 . We repeat this simulation 1000 times. At each p_0 , the circular point is the mean squared radius and the error bar represents the mean squared radius ± 1.96 standard deviations. Hence, the error bars represent a typical range of squared radius values for each d and p_0 . Blue points/lines correspond to p_0^* . The red curve is the expected squared radius. See Theorem 5.9 proof in the supplement for a derivation of the expected squared radius at p_0 .

in p_0 at higher values of d . That is, use of the optimal p_0^* has a greater effect on the split LRT squared radius in higher dimensions, where p_0^* is close to 0.5.

Figure 5.13 plots the power of the LRTs against $\|\theta^*\|^2$. Each vector θ^* has the form $c\mathbf{1}$. To plot the classical and subsampling LRT power, this figure uses the standard normal cumulative distribution function approximation to the non-central χ^2 cumulative distribution function. We use simulations to approximate the power of the split and cross-fit LRTs. For a given value of θ^* , we simulate $n = 1000$ observations $Y_1, \dots, Y_n \stackrel{d}{\sim} N(\theta^*, I_d)$. We construct split LRT and cross-fit LRT confidence sets from this sample. Then we test whether $\theta = 0$ is in each confidence set. We repeat this procedure 5000 times at each θ^* , and each procedure's estimated power at θ^* is the proportion of times that $0 \notin C_n(\alpha)$.

As we would expect, the power is higher when θ^* is farther away from 0. In addition, the classical LRT has the highest power, followed in order by the subsampling LRT, the cross-fit LRT, and the split LRT. Interestingly, at $d = 1$ the subsampling and cross-fit LRT have nearly identical approximate power. As d increases, the difference between the subsampling and cross-fit LRT power increases.

Universal Inference for model selection. We now compare various methods for the problem of testing the number of components in a Gaussian mixture model. It is well-known that the parametric family of Gaussian mixtures does not satisfy the regularity conditions required for (twice the negative logarithm of) the likelihood ratio statistic to admit its traditional χ^2 limiting distribution. In contrast, the method of universal inference is valid without any regularity conditions, and is therefore a natural candidate for this problem.

Although the limiting distribution of the likelihood ratio statistic is unknown or intractable for general Gaussian mixtures, it admits a simple expression when the underlying mixing proportions are known. We will assume this to be the case so that we can use the likelihood ratio test (LRT) as a benchmark, but we emphasize that the LRT cannot easily be used to derive a valid test for more general Gaussian mixtures, where the universal inference method would remain valid.

Let

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} 0.25 \cdot N(\mu_1, 1) + 0.75 \cdot N(\mu_2, 1), \quad (5.4)$$

where the only unknown parameters are $\mu_1, \mu_2 \in \mathbb{R}$, and consider the problem of testing whether the above

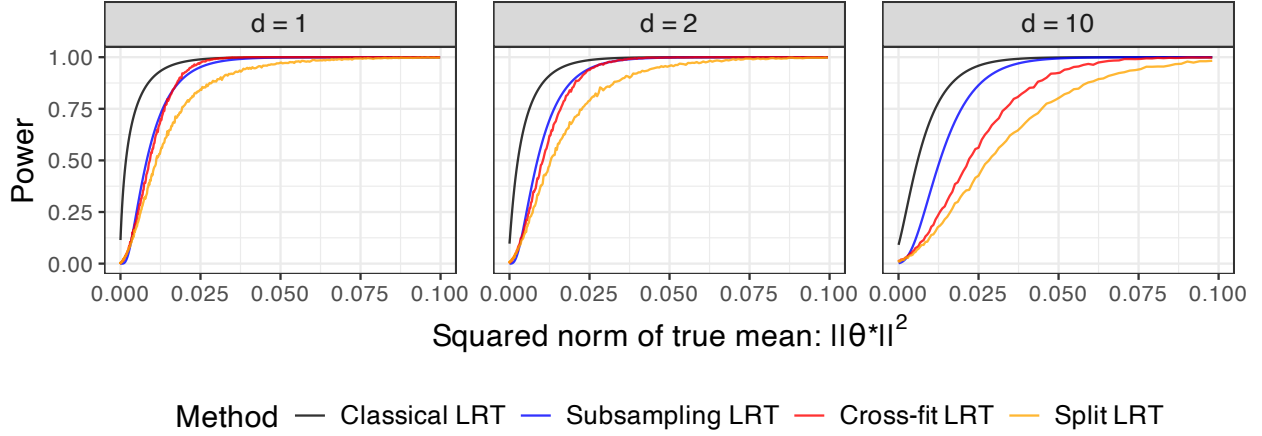


Figure 5.13: Estimated power of classical LRT (black), limiting subsampling LRT (blue), cross-fit LRT (red), and split LRT (orange). We are testing $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ across varying true $\|\theta^*\|^2$. We use the standard normal cumulative distribution function approximation for the classical and subsampling LRT power calculations, and we use simulations to estimate the cross-fit and split LRT power.

mixture has one versus two components, i.e.,

$$H_0 : \mu_1 = \mu_2, \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2. \quad (5.5)$$

If λ denotes the likelihood-ratio statistic for these hypotheses, then $-2 \log \lambda$ admits the limiting distribution $\max(0, Z)^2$, for $Z \stackrel{d}{\sim} N(0, 1)$, thus a valid level- α test for H_0 is to reject if $-2 \log \lambda > q_{1-2\alpha}$, the latter being the $1 - 2\alpha$ quantile of the χ_1^2 distribution. We will refer to this as the LRT test. We compare its numerical performance to that of the following universal inference tests:

- UI: split LR e-variable with Markov's inequality;
- UMI-UI: split LR e-variable with the uniformly randomized Markov's inequality;
- SUI: subsampling LR e-variable with Markov's inequality;
- UMI-SUI: subsampling LR e-variable with uniformly randomized Markov's inequality;
- EMI-SUI: subsampling LR e-variable with exchangeable Markov's inequality;
- EUMI-SUI: subsampling LR e-variable with exchangeable, uniformly-randomized Markov's inequality;

Figure 5.14 reports the empirical power of these procedures based on 500 samples of size $n = 500$ from model (5.4), under ten equally-spaced values of $\mu := -\mu_1 = \mu_2 \in [0, 1]$. We observe that the power of the UMI-UI method uniformly dominates that of the original UI method, and similarly, the UMI-SUI, EMI-SUI, and EUMI-SUI methods dominate their SUI counterpart. The methods UMI-SUI and EUMI-SUI exhibit similar performance, and their absolute increase in power compared to SUI is on the order of 15% for some values of μ . Although all methods based on the split LRT are markedly more conservative than the LRT, we recall that they can be used in arbitrary mixture models, while the LRT cannot.

Bibliographical note

The universal inference method was developed by Wasserman et al. [2020]. These authors also showed that for regular models, the universal confidence sets had width scaling as $\sqrt{d/n}$, where d is the dimensionality of the data, as would be expected. The exchangeable Markov inequality was first derived in Manole and

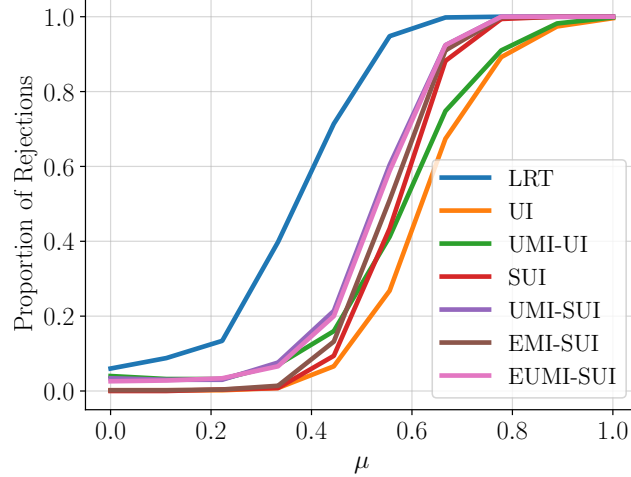


Figure 5.14: Empirical power of the seven tests for the null hypothesis H_0 in equation (5.5). While the benchmark test provided by the LRT is most powerful, the variants of Universal Inference based on the EMI, UMI, and EUMI are more powerful than their MI counterparts.

Ramdas [2023], but its implications for testing with e-variables was first detailed in Ramdas and Manole [2025]. The smoothed MLE idea is related to the method in Seo and Lindsay [2013].

Theorem 5.9 was proved in Dunn et al. [2023]. The authors find that the universal sets match the standard ones in terms of scaling with respect to d, n, α and the signal to noise ratio, their radii being larger by a small constant factor (smaller than 2). Strieder and Drton [2022] also study the optimal splitting ratio for the split LRT, but by examining the power of the test. They derive the (complicated) limiting distribution of the split LRT, and Monte-Carlo simulations show that their proposed split ratio also converges to $1/2$ when testing a fixed number of parameters, but otherwise varies with the dimension of the tested hypothesis.

Dunn et al. [2025] applied the method to the problem of testing log-concavity, for which no other valid test is known. The Gaussian mixture model selection problem was previously studied by [Ghosh and Sen, 1984, Chen and Li, 2009], and the limiting distributions we employed can be found in Goffinet et al. [1992]. Shi and Drton [2024] prove that for this problem, the split likelihood ratio statistic is asymptotically normal with increasing mean and variance, and that it achieves the optimal detection rate of $\sqrt{(\log \log n)/n}$. Tse and Davison [2022] study some asymptotic approaches to improving the power of universal inference for high-dimensional problems with a large number of nuisance parameters.

The code to reproduce plots is at

https://github.com/RobinMDunn/GaussianUniv_sims

for the first four figures, and at

<https://github.com/tmanole/Randomized-Markov>

for the final figure.

Chapter 6

The log-optimal e-value and reverse information projection

While the universal inference e-variable (or e-process) exists in quite some generality and is relatively straightforward to construct, it is only known to be *asymptotically* log-optimal but does not have a finite-sample log-optimality guarantee. It turns out that in even more generality, a log-optimal e-variable exists and is unique. In fact, it always dominates universal inference, rendering the latter method inadmissible. The catch, of course, is that it may not be numerically as easy to construct, but we will be able to characterize it analytically in many cases.

This chapter is dedicated to describing a duality theory that is central to designing log-optimal e-variables. The two main characters in this duality are a special e-variable, called the numeraire, and a special (sub-probability) distribution called the reverse information projection (RIPr).

Setting: $\mathcal{P}$ versus $\mathbb{Q}$.

As elaborated in Chapter 3.7, there are two central techniques for dealing with composite alternatives — mixtures and plug-in. For this reason in the current chapter, we consider only a point alternative $\mathbb{Q}$ and a composite null $\mathcal{P}$.

Our most general results make no assumptions on $\mathbb{Q}$ and $\mathcal{P}$ whatsoever, not even the existence of a common reference measure or the minimum KL divergence between $\mathbb{Q}$ and $\mathcal{P}$ being finite. However, some of these results have sophisticated proofs that are omitted. We will instead prove some of the results in simpler special cases, making some simplifying assumptions.

When we use the term e-variable below, we always mean e-variable *for* $\mathcal{P}$. We let $\mathfrak{E}$ denote the set of e-variables:

$$\mathfrak{E} = \{E \geq 0 : \mathbb{E}^{\mathbb{P}}[E] \leq 1 \text{ for all } \mathbb{P} \in \mathcal{P}\}.$$

6.1 The numeraire e-variable

We first give the definition of the numeraire e-variable, which is closely connected to the log-optimal e-variable in Section 3.5.

Definition 6.1

A *numeraire e-variable* (or just *numeraire* for short) is a $\mathbb{Q}$ -almost surely strictly positive e-variable E^* such that $\mathbb{E}^{\mathbb{Q}}[E/E^*] \leq 1$ for every e-variable $E \in \mathfrak{E}$.

Numeraires are unique up to $\mathbb{Q}$ -nullsets. Indeed, if E_1^* and E_2^* are numeraires, then the ratio $Y = E_2^*/E_1^*$ satisfies $1 \leq 1/\mathbb{E}^{\mathbb{Q}}[Y] \leq \mathbb{E}^{\mathbb{Q}}[1/Y] \leq 1$ thanks to the numeraire property of E_1^* , Jensen's inequality, and the numeraire property of E_2^* . Thus Jensen's inequality holds with equality, so Y is $\mathbb{Q}$ -almost surely equal to a constant which must be one. It follows that E_1^* and E_2^* are $\mathbb{Q}$ -almost surely equal. In view of this uniqueness, we often speak of 'the' numeraire.

In the case of a simple null $\mathcal{P} = \{\mathbb{P}\}$ with $\mathbb{Q} \ll \mathbb{P}$, the numeraire is just the likelihood ratio $E^* = d\mathbb{Q}/d\mathbb{P}$. Indeed, E^* is $\mathbb{Q}$ -almost surely strictly positive, it is an e-variable, and for any other e-variable $E \in \mathfrak{E}$ we have by a simple change of measure that $\mathbb{E}^{\mathbb{Q}}[E/E^*] = \mathbb{E}^{\mathbb{P}}[E] \leq 1$.

Even for composite nulls, the numeraire is a likelihood ratio of $\mathbb{Q}$ to the 'reverse information projection'. Before getting there, we present a few other simple results.

Lemma 6.2: Lifting

Let E^* denote the numeraire for $\mathcal{P}$, and consider a larger null hypothesis $\mathcal{P}' \supseteq \mathcal{P}$. If E^* is still an e-variable for $\mathcal{P}'$, then it is also the numeraire for $\mathcal{P}'$.

The proof is simple. If E is an e-variable for $\mathcal{P}'$, it is also an e-variable for $\mathcal{P}$ and one has $\mathbb{E}^{\mathbb{Q}}[E/E^*] \leq 1$ by assumption, proving the lemma.

A numeraire E^* is also log-optimal (Section 3.5) in the sense that

$$\mathbb{E}^{\mathbb{Q}} \left[\log \frac{E}{E^*} \right] \leq 0 \text{ for every e-variable } E \in \mathfrak{E}, \quad (6.1)$$

where the left-hand side may be $-\infty$. This follows directly from Jensen's inequality and the numeraire property. The converse is also true, and we record the equivalence in the following proposition. The proof shows that the numeraire property is really the first-order condition for log-optimality. Note also that a numeraire E^* is the $\mathbb{Q}$ -almost surely unique log-optimal e-variable in the sense of (6.1) even if $\mathbb{E}^{\mathbb{Q}}[\log E^*]$ happens to be infinite.

Proposition 6.3: Log-optimality of numeraire

Let E^* be a $\mathbb{Q}$ -almost surely strictly positive e-variable. Then E^* is a numeraire if and only if it is log-optimal. In particular, a log-optimal e-variable is unique up to $\mathbb{Q}$ -nullsets.

Proof.

The forward direction was argued above. To prove the converse we assume E^* is log-optimal. For any e-variable E and $t \in (0, 1)$, $E(t) = (1-t)E^* + tE$ is an e-variable. Thus by log-optimality, $\mathbb{E}^{\mathbb{Q}}[t^{-1} \log(E(t)/E^*)] \leq 0$. The expression inside the expectation equals $t^{-1} \log(1-t + tE/E^*)$, which converges to $E/E^* - 1$ as t tends to zero and is bounded below by $t^{-1} \log(1-t)$, hence by $-2 \log 2$ for $t \in (0, 1/2)$. Fatou's lemma thus yields $\mathbb{E}^{\mathbb{Q}}[E/E^* - 1] \leq 0$, showing that E^* is a numeraire. Finally, the uniqueness statement follows from the equivalence just proved together with uniqueness of the numeraire up to $\mathbb{Q}$ -nullsets.

The following result says that a numeraire always exists without any assumptions whatsoever on $\mathcal{P}$ or $\mathbb{Q}$. Although no assumptions on $\mathcal{P}$ or $\mathbb{Q}$ are needed in general, we do not present this more general theory for reasons of brevity. It turns out that the theory simplifies under the following condition, which, despite being nontrivial, is still very weak.

We say that $\mathbb{Q}$ is *absolutely continuous with respect to* $\mathcal{P}$, written $\mathbb{Q} \ll \mathcal{P}$, if $\mathbb{Q}$ does not assign positive probability to any event that is null under every $\mathbb{P} \in \mathcal{P}$. Equivalently,

$$\mathbb{Q} \ll \mathcal{P} \iff \mathbb{Q}(A) > 0 \text{ implies } \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(A) > 0 \text{ for every } A. \quad (6.2)$$

This natural generalization of absolute continuity for pairs of measures is satisfied in most statistically relevant settings, while being significantly weaker than requiring a common dominating reference measure. It is also weaker than requiring that $\mathbb{Q} \ll \mathbb{P}$ for every $\mathbb{P} \in \mathcal{P}$, which would require that $\mathbb{Q}(A) = 0$ as long as any (as opposed to every) $\mathbb{P} \in \mathcal{P}$ assigns $\mathbb{P}(A) = 0$ (or contrapositively, that $\mathbb{Q}(A) > 0$ implies $\mathbb{P}(A) > 0$ for all $\mathbb{P} \in \mathcal{P}$).

Theorem 6.4

A numeraire always exists. Moreover, the following conditions are equivalent:

- (a) The numeraire is $\mathbb{Q}$ -almost surely finite.
- (b) Every e-variable is $\mathbb{Q}$ -almost surely finite.
- (c) $\mathbb{Q}$ is absolutely continuous with respect to $\mathcal{P}$, that is $\mathbb{Q} \ll \mathcal{P}$.

Proof

The proof that the numeraire exists is sophisticated and technical, thus omitted here. We focus on the equivalence of the other three claims. Clearly (b) implies (a). To see that (a) implies (b), let E^* be a $\mathbb{Q}$ -almost surely finite numeraire and note that the numeraire property implies that any e-variable E must be $\mathbb{Q}$ -almost surely finite too. Next, we prove that (b) is equivalent to (c). First, suppose (c) fails. Then there is an event A with $\mathbb{Q}(A) > 0$ and $\mathbb{P}(A) = 0$ for all $\mathbb{P} \in \mathcal{P}$, in which case $E = \infty \mathbb{1}_A$ is an e-variable that is not $\mathbb{Q}$ -almost surely finite. Thus (b) fails also. The converse direction follows by noting that every e-variable is finite $\mathbb{P}$ -almost surely for all $\mathbb{P} \in \mathcal{P}$.

We end with a few additional properties of the numeraire. Since the constant one is always an e-variable, the numeraire E^* satisfies the aesthetically pleasing property that

$$\mathbb{E}^{\mathbb{P}}[E^*] \leq 1 \text{ for all } \mathbb{P} \in \mathcal{P} \text{ and } \mathbb{E}^{\mathbb{Q}}\left[\frac{1}{E^*}\right] \leq 1.$$

When $\mathcal{P} = \{\mathbb{P}\}$ is simple and $\mathbb{Q} \ll \mathbb{P}$, both inequalities above hold with equality, attained by the numeraire $d\mathbb{Q}/d\mathbb{P}$. Another pleasing property is obtained by Jensen's inequality: The numeraire E^* satisfies

$$\mathbb{E}^{\mathbb{Q}}[E^*/E] \geq 1 \text{ for all } E \in \mathfrak{E}.$$

Finally, note that the definition of KL divergence in (3.4) can be rewritten and generalized to nonnegative measures $\mathbb{P}$ in the following manner. Let $\mathcal{M}_+$ denote the set of all nonnegative measures on Ω . Then, for any $\mathbb{P} \in \mathcal{M}_+$ and $\mathbb{Q} \in \mathcal{M}_1$, define

$$\text{KL}(\mathbb{Q}, \mathbb{P}) = -\mathbb{E}^{\mathbb{Q}}\left[\log \frac{d\mathbb{P}^a}{d\mathbb{Q}}\right],$$

where $\mathbb{P}^a$ is the absolutely continuous part of $\mathbb{P}$ with respect to $\mathbb{Q}$. When $\mathbb{P} \in \mathcal{M}_1$, this definition coincides with (3.4).

We first begin with the simple observation that since $\log y \leq y - 1$, we have by a change of measure that for any $E \in \mathfrak{E}$ and $\mathbb{P} \in \text{Conv}(\mathcal{P})$,

$$\mathbb{E}^{\mathbb{Q}}\left[\log\left(E \frac{d\mathbb{P}^a}{d\mathbb{Q}}\right)\right] \leq \mathbb{E}^{\mathbb{P}}[E] - 1 \leq 0.$$

This immediately implies the following fact that we record for later use.

Proposition 6.5: Weak duality

For all e-variables $E \in \mathfrak{E}$ and distributions $\mathbb{P} \in \text{Conv}(\mathcal{P})$, we have $\mathbb{E}^{\mathbb{Q}}[\log E] \leq \text{KL}(\mathbb{Q}, \mathbb{P})$ (where if $\mathbb{E}^{\mathbb{Q}}[\log E]$ is undefined, we treat it as $-\infty$), and thus

$$\sup_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E] \leq \inf_{\mathbb{P} \in \text{Conv}(\mathcal{P})} \text{KL}(\mathbb{Q}, \mathbb{P}). \quad (6.3)$$

Note that both sides of (6.3) are nonnegative and well defined, since the constant 1 is always an e-variable. It turns out that the numeraire E^* satisfies

$$\mathbb{E}^{\mathbb{Q}}[\log E^*] = \text{KL}(\mathbb{Q}, \mathbb{P}^*),$$

for an appropriately defined $\mathbb{P}^*$ that we introduce and discuss next.

6.2 Strong duality for finite $\mathcal{P}$

We start here with the case that $\mathcal{P}$ is finite. Then, it is possible to show the following result. Recall that Δ_K denotes the probability simplex.

Proposition 6.6

Suppose $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_K\}$ with $\mathbb{P}_k \ll \mathbb{Q}$ for all $k \in [K]$, and $\min_{\mathbb{P} \in \text{Conv}(\mathcal{P})} \text{KL}(\mathbb{Q}, \mathbb{P}) < \infty$. Then,

$$\mathbb{E}^{\mathbb{Q}}[\log E^*] = \max_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E] = \min_{\mathbb{P} \in \text{Conv}(\mathcal{P})} \text{KL}(\mathbb{Q}, \mathbb{P}) = \text{KL}(\mathbb{Q}, \mathbb{P}^*),$$

where E^* is the numeraire and $\mathbb{P}^*$ achieves the minimum.

To elaborate, we know that the numeraire achieves the left hand side ($\mathbb{Q}$ -almost surely uniquely), so the first equality holds by its definition. The minimization problem can be rewritten as

$$\min_{\lambda \in \Delta_K} \text{KL} \left(\mathbb{Q}, \sum_{k=1}^K \lambda_k \mathbb{P}_k \right).$$

This is a (strictly) convex minimization problem over the compact convex set Δ_K , so it achieves its minimum uniquely at some $\lambda^* \in \Delta_K$, and we define the reverse information projection (RIPr) to be

$$\mathbb{P}^* = \sum_{k=1}^K \lambda_k^* \mathbb{P}_k,$$

so the last equality also holds by definition. Since $\text{KL}(\mathbb{Q}, \mathbb{P}^*) < \infty$ by assumption, we must have $\mathbb{Q} \ll \mathbb{P}^*$. We now prove the strong duality result in the middle equality.

Proof of Proposition 6.6

Recall from the earlier weak duality result of Proposition 6.5 that if we find any e-variable whose e-power equals $\text{KL}(\mathbb{Q}, \mathbb{P}^*)$, then the proof is complete, and further, that e-variable must be the numeraire. By the optimality of $\mathbb{P}^*$ and convexity of $\{\sum_{k=1}^K \lambda_k \mathbb{P}_k : \lambda \in \Delta_K\}$, we that have for any $\mathbb{P} \in \mathcal{P}$,

$$\text{KL}(\mathbb{Q}, \mathbb{P}^*) = \min_{\alpha \in [0,1]} \text{KL}(\mathbb{Q}, \alpha \mathbb{P}^* + (1 - \alpha) \mathbb{P})$$

The optimality criterion (Karush–Kuhn–Tucker condition) for the above minimization problem

is that the gradient with respect to α is nonpositive at $\alpha = 1$ (going from 1 towards 0 increases the objective):

$$-\mathbb{E}^{\mathbb{Q}} \left[\frac{d(\mathbb{P}^* - \mathbb{P})}{d\mathbb{P}^*} \right] \leq 0 \iff \mathbb{E}^{\mathbb{Q}} \left[\frac{d\mathbb{P}}{d\mathbb{P}^*} \right] \leq 1 \iff \mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}^*} \right] \leq 1.$$

The last inequality means that $d\mathbb{Q}/d\mathbb{P}^*$ is an e-variable. By definition of e-power, the e-power of this e-variable is $\text{KL}(\mathbb{Q}, \mathbb{P}^*)$. Thus, the numeraire E^* is $d\mathbb{Q}/d\mathbb{P}^*$ and the proof is complete.

6.3 Reverse information projection

Remarkably, all assumptions in Proposition 6.6 can be dropped. But doing this is very challenging, as explained below.

First, it is clear that in the previous section, $\mathbb{P}^*$ is the closest distribution in $\text{Conv}(\mathcal{P})$ to $\mathbb{Q}$ as measured by KL. However, when the infimum KL divergence equals infinity, it is unclear how to define the “closest” distribution to $\mathbb{Q}$.

Second, in the previous section, we used the compactness of $\text{Conv}(\mathcal{P})$ when $\mathcal{P}$ is finite to argue that the infimum is attained by a distribution in $\text{Conv}(\mathcal{P})$. However, when $\mathcal{P}$ is uncountable (as would typically be the case in statistical applications), the set is no longer compact (with respect to the weak topology) and it is not clear that the infimum is achieved inside the set. Indeed, it is typically the case that the minimizer lies outside of $\text{Conv}(\mathcal{P})$, and in fact may be a *subdistribution* in general (a nonnegative measure that integrates to less than one). In that case, where does it lie, and how do we characterize it?

Finally, when all absolute continuity requirements are dropped, all arguments involving likelihood ratios (Radon–Nikodym derivatives) and KL divergences become much more subtle, and one cannot proceed as easily as in the previous section.

Since formally handling the above subtleties requires some technical measure-theoretic care, we only state the results here but do not prove them. The first step is to define the reverse information projection (RIPr) appropriately.

Definition 6.7: Reverse information projection (RIPr)

Let E^* be the numeraire for $\mathcal{P}$ against $\mathbb{Q}$. The reverse information projection of $\mathbb{Q}$ onto $\mathcal{P}$, denoted $\mathbb{P}^*$, is defined via

$$\frac{d\mathbb{P}^*}{d\mathbb{Q}} = \frac{1}{E^*},$$

where the right-hand side is understood to equal zero on $\{E^* = \infty\}$.

In words, $\mathbb{P}^*$ is a nonnegative measure whose likelihood ratio with respect to $\mathbb{Q}$ equals $1/E^*$. Since E^* exists, and is $\mathbb{Q}$ -almost surely unique and positive, we see that $\mathbb{P}^*$ is uniquely defined up to $\mathbb{Q}$ -nullsets. Also, $\mathbb{P}^* \ll \mathbb{Q}$ by definition. Moreover, $\mathbb{P}^*(\Omega) \leq 1$ because $\mathbb{E}^{\mathbb{Q}}[1/E^*] \leq 1$, and hence $\mathbb{P}^*$ is in general a sub-probability measure, though it is a probability measure in some special cases.

6.3.1 The effective null hypothesis is the bipolar of $\mathcal{P}$

Let $\mathcal{X}_+$ denote the set of all $[0, \infty]$ -valued measurable functions on Ω and recall that $\mathcal{M}_+$ denotes the set of all nonnegative measures on Ω . The *polar* of $\mathcal{P}$ is defined by

$$\mathcal{P}^\circ = \{E \in \mathcal{X}_+ : \mathbb{E}^{\mathbb{P}}[E] \leq 1 \text{ for all } \mathbb{P} \in \mathcal{P}\}.$$

This is precisely the set of e-variables, i.e., $\mathcal{P}^\circ = \mathfrak{E}$. The *bipolar* of $\mathcal{P}$ is

$$\mathcal{P}^{\circ\circ} = \{\mathbb{P} \in \mathcal{M}_+ : \mathbb{E}^{\mathbb{P}}[E] \leq 1 \text{ for all } E \in \mathcal{P}^\circ\}.$$

It is easy to see that

$$\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1 \supseteq \text{Conv}(\mathcal{P}).$$

Equality holds when $\mathcal{P}$ is finite, but the inclusion can be strict. Because the constant $E = 1$ is always an e-variable, elements of $\mathcal{P}^{\circ\circ}$ have total mass at most one, and sometimes strictly less than one. For example, if $\mathbb{P}$ belongs to $\mathcal{P}^{\circ\circ}$, then so does any $\mathbb{P}' \in \mathcal{M}_+$ that is set-wise dominated by $\mathbb{P}$, meaning $\mathbb{P}'(A) \leq \mathbb{P}(A)$ for all A . Below we give two examples where $\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1$ is a strict superset of $\text{Conv}(\mathcal{P})$.

Example 6.8

Suppose the null $\mathcal{P}$ consists of all unit variance Gaussians with mean less than 0. Then the standard Gaussian is not in the convex hull of $\mathcal{P}$, but it is in $\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1$, which contains all unit variance Gaussians with nonnegative mean (and their mixtures).

The reader may note that if we had considered the total variation closure of the convex hull of $\mathcal{P}$ (motivated by Theorem 3.17) in Example 6.8, it would have coincided with $\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1$. However, more generally, this is not the case.

Example 6.9

Let $\Omega = [0, 1]$, and let

$$\mathcal{P} = \{\delta_x : x \in [0, 0.5]\},$$

where δ_x is a point mass at x . Clearly, $\mathcal{P}$ has no common reference measure. The total variation closure of the convex hull of $\mathcal{P}$ still contains only atomic measures. However, $\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1$ contains every bounded distribution on $[0, 0.5]$.

In the above example, if, instead of total variation closure, we had taken a weak closure of the convex hull of $\mathcal{P}$, then again we would have obtained $\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1$. However, one can construct yet other examples to show that this is not true in general, but this gets rather deep into topological measure theory, and is out of the scope of the current book.

The bipolar of $\mathcal{P}$ can be interpreted as the *effective null hypothesis*: it is the largest family of distributions whose set of e-variables is exactly $\mathfrak{E}$. Thus it consists of those distributions against which e-variables in $\mathfrak{E}$ do not provide any evidence. In particular, we note the following simple yet important consequence, recalling the definition of a powered e-variable from Definition 3.1.

Theorem 6.10

A powered test (or e-variable) for $\mathcal{P}$ against $\mathbb{Q}$ exists if and only if $\mathbb{Q} \notin \mathcal{P}^{\circ\circ}$.

Proof.

The proof follows in two simple steps. It follows from Proposition 3.9 that a powered test exists if and only if a powered e-variable exists. The latter condition is possible if and only if $\mathbb{Q} \notin \mathcal{P}^{\circ\circ}$ by definition of $\mathcal{P}^{\circ\circ}$.

It turns out that considering only bounded e-variables suffices. Let $\mathfrak{E}^b \subsetneq \mathcal{P}^\circ = \mathfrak{E}$ be defined as $\mathfrak{E}^b = \{E \in \mathcal{X}_+ : \mathbb{E}^\mathbb{P}[E] \leq 1 \text{ for all } \mathbb{P} \in \mathcal{P}, \sup_{\omega \in \Omega} E(\omega) < \infty\}$. Then an application of the monotone convergence theorem yields that

$$\mathcal{P}^{\circ\circ} = \{\mathbb{P} \in \mathcal{M}_+ : \mathbb{E}^\mathbb{P}[E] \leq 1 \text{ for all } E \in \mathfrak{E}^b\}.$$

We now claim that even though it is possible that $\mathbb{P}^* \notin \text{Conv}(\mathcal{P})$ in general, it does lie in $\mathcal{P}^{\circ\circ}$.

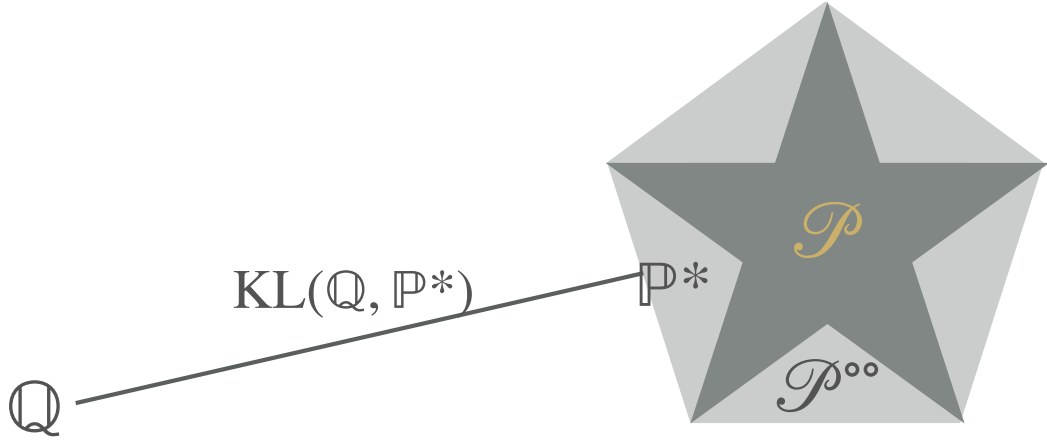


Figure 6.12: A visualization of the RPr $\mathbb{P}^*$ of $\mathbb{Q}$ onto $\mathcal{P}$, which lies in $\mathcal{P}^{\circ\circ}$, but not necessarily in $\mathcal{M}_1$.

Proposition 6.11

The reverse information projection of $\mathbb{Q}$ onto $\mathcal{P}$ lies in the bipolar of the null: $\mathbb{P}^* \in \mathcal{P}^{\circ\circ}$.

The proof follows because for any e-variable E ,

$$\mathbb{E}^{\mathbb{P}^*}[E] = \mathbb{E}^{\mathbb{Q}}[\mathbf{1}_{\{E^* < \infty\}} E/E^*] \leq \mathbb{E}^{\mathbb{Q}}[E/E^*] \leq 1,$$

because E^* is the numeraire. The above result can be visualized as Figure 6.12.

6.3.2 A general strong duality result

We now state a more general strong duality result removing almost all measure-theoretic assumptions. As before, $\mathbb{E}^{\mathbb{Q}}[\log E]$ is understood as $-\infty$ whenever $\mathbb{E}^{\mathbb{Q}}[(\log E)_-] = \infty$. Recall the condition $\mathbb{Q} \ll \mathcal{P}$ from (6.2).

Theorem 6.13

Assume $\mathbb{Q} \ll \mathcal{P}$. Letting E^* be the numeraire and $\mathbb{P}^*$ the RPr, one has the strong duality relation

$$\mathbb{E}^{\mathbb{P}^*}[\log E^*] = \sup_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E] = \inf_{\mathbb{P} \in \mathcal{P}^{\circ\circ}} \text{KL}(\mathbb{Q}, \mathbb{P}) = \text{KL}(\mathbb{Q}, \mathbb{P}^*),$$

where these quantities may equal ∞ .

We also provide some characterizing properties of the RPr.

Theorem 6.14

Assume $\mathbb{Q} \ll \mathcal{P}$ and let $\mathbb{P}^*$ be an element of $\mathcal{P}^{\circ\circ}$ equivalent to $\mathbb{Q}$. Let $\mathbb{P}^a$ denote the absolutely continuous part of $\mathbb{P}$ with respect to $\mathbb{Q}$. The following statements imply each other:

- (i) $\mathbb{P}^*$ is the RIPr.
- (ii) $\mathbb{E}^{\mathbb{Q}} \left[\frac{d\mathbb{P}^a}{d\mathbb{P}^*} \right] \leq 1$ for all $\mathbb{P} \in \mathcal{P}^{\circ\circ}$.
- (iii) $\mathbb{E}^{\mathbb{Q}} \left[\log \frac{d\mathbb{P}^a}{d\mathbb{P}^*} \right] \leq 0$ for all $\mathbb{P} \in \mathcal{P}^{\circ\circ}$.

If we assume further that all elements of $\mathcal{P}$ are absolutely continuous with respect to a probability measure $\mathbb{L}$, then all elements of $\mathcal{P}^{\circ\circ}$ are also absolutely continuous with respect to $\mathbb{L}$, so that we can identify any $\mathbb{P} \in \mathcal{P}^{\circ\circ}$ with its density $p = d\mathbb{P}/d\mathbb{L}$. Further, each of the above conditions is equivalent to

- (iv) $\mathbb{E}^{\mathbb{Q}} [p/p^*] \leq 1$ for all densities p of $\mathbb{P} \in \mathcal{P}$.

We do not prove the above theorems here. Crucially, note that the fourth condition above only needs to hold over $\mathcal{P}$ and not over the bipolar, making it easier to verify in settings with a reference measure such as exponential families, as we will see in Section 6.7.1.

6.4 Numeraire is the only e-variable that is a likelihood ratio

We first present a verification theorem that simplifies the task of checking that candidates E^* and $\mathbb{P}^*$ (for the numeraire and RIPr respectively) are in fact optimal.

Theorem 6.15

Assume that $\mathbb{Q} \ll \mathcal{P}$. Let E^* be a $\mathbb{Q}$ -almost surely strictly positive e-variable and let $\mathbb{P}^*$ be the measure given by $d\mathbb{P}^*/d\mathbb{Q} = 1/E^*$. Then E^* is a numeraire if and only if $\mathbb{P}^*$ belongs to the effective null $\mathcal{P}^{\circ\circ}$. Thus the numeraire is the only e-variable that is also a likelihood ratio between $\mathbb{Q}$ and some equivalent element of $\mathcal{P}^{\circ\circ}$. Finally, $\mathbb{E}^{\mathbb{P}^*}[E^*] = 1$ and $\mathbb{P}^*$ is maximal in the sense that no other $\mathbb{P} \in \mathcal{P}^{\circ\circ}$ equivalent to $\mathbb{Q}$ can satisfy $\mathbb{P}(A) \geq \mathbb{P}^*(A)$ for all A with strict inequality for some A .

Proof

Recall that all e-variables are $\mathbb{Q}$ -almost surely finite thanks to Theorem 6.4 and the assumption that $\mathbb{Q} \ll \mathcal{P}$; in particular, $\mathbb{P}^*$ is equivalent to $\mathbb{Q}$. Now, if E^* is a numeraire, then Proposition 6.11 yields that $\mathbb{P}^* \in \mathcal{P}^{\circ\circ}$. For the converse, the definition of $\mathbb{P}^*$ and the fact that it belongs to $\mathcal{P}^{\circ\circ}$ yield $\mathbb{E}^{\mathbb{Q}}[E/E^*] = \mathbb{E}^{\mathbb{P}^*}[E] \leq 1$ for any e-variable E . This is the numeraire property. Finally, the definition of $\mathbb{P}^*$ immediately gives $\mathbb{E}^{\mathbb{P}^*}[E^*] = 1$, and $\mathbb{P}^*$ must be maximal because if a ‘dominating’ equivalent $\mathbb{P} \in \mathcal{P}^{\circ\circ}$ existed we would get the contradiction $1 = \mathbb{E}^{\mathbb{P}^*}[E^*] < \mathbb{E}^{\mathbb{P}}[E^*] \leq 1$.

We stress that the above result is quite remarkable: the numeraire is clearly a likelihood ratio between $\mathbb{Q}$ and some equivalent element of $\mathcal{P}^{\circ\circ}$, but it is also the unique e-variable with such a representation! This property helps identify the numeraire in several simple examples (eg: Sections 6.7.1 and 6.7.2), where we guess the numeraire as being a particular simple likelihood ratio of the alternative to some ‘extreme’ point of the null, verify that it is indeed an e-variable, and conclude that it is in fact the log-optimal e-variable.

Corollary 6.16

Assume $\mathbb{Q} \ll \mathcal{P}$. Let $\mathfrak{E}_0$ be a family of $[0, \infty]$ -valued random variables that generates the null hypothesis in the sense that

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[E] \leq 1 \text{ for all } E \in \mathfrak{E}_0\}. \quad (6.4)$$

Let E^* be a $\mathbb{Q}$ -almost surely strictly positive e-variable such that

$$\mathbb{E}^{\mathbb{Q}} \left[\frac{1}{E^*} \right] = 1 \text{ and } \mathbb{E}^{\mathbb{Q}} \left[\frac{E}{E^*} \right] \leq 1 \text{ for all } E \in \mathfrak{E}_0. \quad (6.5)$$

Then E^* is the numeraire and the RPr belongs to $\mathcal{P}$.

Proof

Let $\mathbb{P}^*$ be given by $d\mathbb{P}^*/d\mathbb{Q} = 1/E^*$. Then (6.5) says that $\mathbb{P}^*$ is a probability measure such that $\mathbb{E}^{\mathbb{P}^*}[E] \leq 1$ for all $E \in \mathfrak{E}_0$. Thus by (6.4), $\mathbb{P}^*$ belongs to $\mathcal{P}$, hence to $\mathcal{P}^{\circ\circ}$. The result now follows from Theorem 6.15.

Thanks to Lemma 6.2, Corollary 6.16 also holds if one has ‘ $\supseteq$ ’ instead of ‘ $=$ ’ in (6.4). This is intuitive: if one enlarges the generating set $\mathfrak{E}_0$ beyond what is necessary to specify $\mathcal{P}$, then the right-hand side of (6.4) can become smaller than $\mathcal{P}$. In that case, (6.5) only gets harder to satisfy, and if we find an e-variable E^* satisfying it, it surely must still be the numeraire.

Another useful corollary is the following *shrinking lemma*, a dual result to the *lifting lemma* (Lemma 6.2).

Lemma 6.17: Shrinking

Assume that $\mathbb{Q} \ll \mathcal{P}$, and let $\mathbb{P}^*$ denote the RPr of $\mathbb{Q}$ onto $\mathcal{P}$ and let $X^* = d\mathbb{Q}/d\mathbb{P}^*$ denote the numeraire for $\mathcal{P}$ against $\mathbb{Q}$. For any $\overline{\mathcal{P}} \subseteq \mathcal{P}$, if $\mathbb{P}^* \in \overline{\mathcal{P}}^{\circ\circ}$, then $\mathbb{P}^*$ is the RPr of $\mathbb{Q}$ onto $\overline{\mathcal{P}}$, so X^* is still the numeraire for $\overline{\mathcal{P}}$ against $\mathbb{Q}$.

Proof.

Clearly, $d\mathbb{Q}/d\mathbb{P}^*$ is an e-variable for $\mathcal{P}$ and thus also for $\overline{\mathcal{P}} \subseteq \mathcal{P}$. Since $\mathbb{P}^* \in \overline{\mathcal{P}}^{\circ\circ}$, and noting that the numeraire is the only e-variable that is a likelihood ratio between $\mathbb{Q}$ and some equivalent element of $\overline{\mathcal{P}}^{\circ\circ}$, it must be the case that X^* is the numeraire for $\overline{\mathcal{P}}$ against $\mathbb{Q}$ and hence $\mathbb{P}^*$ is the RPr of $\mathbb{Q}$ onto $\overline{\mathcal{P}}$.

As an example of how one might apply the shrinking and lifting lemmata referenced above, consider $\mathcal{P}$ being unit-variance Gaussians with nonpositive mean, and $\mathbb{Q}$ being $N(1, 1)$. It can be easily checked that the RPr is the standard Gaussian, and the numeraire is the likelihood ratio of $N(1, 1)$ to $N(0, 1)$. The shrinking lemma implies that if we consider any subset of $\mathcal{P}$ that still contains $N(0, 1)$, the RPr and numeraire remain identical. We could have also gone in the other direction (from the singleton null of $N(0, 1)$ to the one-sided null $\mathcal{P}$), and used the lifting lemma instead to reach the same conclusions.

6.5 Numeraire dominates universal inference

Assume now that there exists a common reference measure $\mathbb{L}$, and identify distributions with their respective densities, written in lowercase. For a point alternative q over the data Z , the method of universal inference boils down to constructing the e-variable $E^{\text{UI}} = q(Z)/p_{\max}(Z)$ where $p_{\max}(Z) = \sup_{p \in \mathcal{P}} p(Z)$ is the maximum

likelihood. (Strictly speaking, the supremum here should be understood as an essential supremum under the reference measure.)

To compare the universal inference e-variable with the numeraire $E^* = q(Z)/p^*(Z)$, we first claim that the RIPr satisfies $p^*(Z) \leq p_{\max}(Z)$ up to $\mathbb{L}$ -nullsets. This is obvious if the RIPr belongs to $\mathcal{P}$, but in general it only belongs to $\mathcal{P}^{\circ\circ}$. In this case the claim follows by applying the following lemma with $f = p_{\max}$.

Lemma 6.18

Let f be a (potentially random) function such that $f \geq p$, $\mathbb{L}$ -almost surely, for all $p \in \mathcal{P}$. Then $f \geq p$, $\mathbb{L}$ -almost surely, for all $p \in \mathcal{P}^{\circ\circ}$.

We omit the technical proof. As an immediate consequence of the lemma, we see that $p^*(Z) \leq p_{\max}(Z)$, and hence $E^* \geq E^{\text{UI}}$, up to $\mathbb{L}$ -nullsets. In nondegenerate situations involving a composite null hypothesis, the inequality will be strict with positive $\mathbb{L}$ -probability. Thus, in such cases, the relatively general method of universal inference is in fact inadmissible. We end by noting that the numeraire imposes weaker assumptions than universal inference (and thus is, in some sense, even more universal), since the latter needs a reference measure to define likelihoods.

Composite alternatives

Recall the composite alternative setup and notation described in Section 3.7. A reasonable goal is to derive an e-variable $E_n = E(X^n)$ for $\mathcal{P}$ such that for any $\mathbb{Q} \in \mathcal{Q}$,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}^{\mathbb{Q}}[\log E_n]}{n} \rightarrow \text{KL}(\mathbb{Q}, \mathbb{P}^*), \quad (6.6)$$

where $\mathbb{P}^*$ is the RIPr of $\mathcal{Q}$ onto $\mathcal{P}$. As we have seen in this chapter, the right-hand side is the growth rate of an oracle that knows the data-generating distribution (when it lies in the alternative). One can derive appropriate extensions of the mixture and plug-in e-variables described in Section 3.7. For example, in the iid setting considered there, consider the numeraire for testing $\mathcal{P}^{(n)} = \{\mathbb{P}^n : \mathbb{P} \in \mathcal{P}\}$ against the alternative $\tilde{\mathcal{Q}}^{(n)} = \int \mathbb{Q}^n \nu(d\mathbb{Q})$. We conjecture that this is indeed asymptotically log-optimal under weak assumptions, but such a general result is currently not known.

We end with a note on logical coherence. Saying that a composite null $\mathcal{P}$ is true is equivalent to saying that a simple null $\mathbb{P}$ is true for some $\mathbb{P} \in \mathcal{P}$. Thus, one may expect to only reject $\mathcal{P}$ if one can also reject every $\mathbb{P} \in \mathcal{P}$ separately. A method that satisfies this property is called logically coherent. We can extend this definition to e-variables as well. Let $E^{\mathcal{P}}$ denote an e-variable for testing $\mathcal{P}$. The collection $(E^{\mathcal{P}})_{\mathcal{P} \subseteq \mathcal{M}_1}$ can be called coherent if $E^{\mathcal{P}} \leq E^{\mathcal{P}'}$ whenever $\mathcal{P}' \subseteq \mathcal{P}$. Fixing an alternative $\mathbb{Q}$ for simplicity, we now make two simple observations, stated here without proof:

- If $E^{\mathcal{P}}$ equals the universal inference e-variable for $\mathcal{P}$ against $\mathbb{Q}$, the collection $(E^{\mathcal{P}})_{\mathcal{P} \subseteq \mathcal{M}_1}$ is always logically coherent. (This point is easy to verify by definition.)
- If $E^{\mathcal{P}}$ equals the numeraire e-variable for $\mathcal{P}$ against $\mathbb{Q}$, the collection $(E^{\mathcal{P}})_{\mathcal{P} \subseteq \mathcal{M}_1}$ is generally not logically coherent, except for some special cases.

Incoherent families of e-variables may be harder to interpret, because enlarging the null may counterintuitively increase the evidence against the null. Thus, there appears to be a tradeoff between e-power and coherence/interpretability when comparing universal inference and the numeraire.

6.6 Connections between e-values and Bayes factors

For those Bayesians who are willing to consider testing hypotheses, one of the standard quantities to examine is the *Bayes factor*. Suppose we observe n iid data points $X_1, \dots, X_n$, and we are interested in testing

whether their marginal distribution lies in a composite null $\mathcal{P}$ or in a composite alternative $\mathcal{Q}$. Also suppose that there is a common reference measure, allowing us to identify distributions with their densities. Then the Bayes factor is defined as

$$B_n = \frac{\int_{\mathcal{Q}} \prod_{i=1}^n q(X_i) \mu(dq)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i) \nu(dp)},$$

where μ and ν are (often subjective) *priors* over $\mathcal{Q}$ and $\mathcal{P}$. Bayesians often refer to tables like in Section 2.7 in order to interpret the realized value, and typical Bayesians do not threshold B_n to make clear decisions of accepting or rejecting the null, and consequently no effort is made to control the type-I error. The law of iterated expectation yields that B_n is an e-variable *if nature draws a random null from the scientist's prior*— $\mathbb{E}^{p \sim \nu(\text{dp})}[B_n] \leq 1$ —but it remains unclear to many why this is a reasonable assumption. Indeed, this is often the problem with Bayesian methods: the scientist using them is perfectly convinced of its correctness (internal consistency) but their analysis fails to convince all others (lack of external consistency). Thus the main drawbacks of this approach are: (a) it is unclear how to generalize these ideas to nonparametric settings without densities and likelihoods, (b) it is unclear how to formally make decisions with B_n with (frequentist) error guarantees.

In contrast, the typical frequentist approach in the above setting would be to consider the generalized likelihood ratio:

$$G_n = \frac{\sup_{q \in \mathcal{Q}} \prod_{i=1}^n q(X_i)}{\sup_{p \in \mathcal{P}} \prod_{i=1}^n p(X_i)}.$$

Under some regularity conditions, Wilks' theorem implies that $\log G_n$ has a chi-squared limiting distribution as $n \rightarrow \infty$, with degrees of freedom determined by some notion of parametric dimensionality difference between $\mathcal{Q}$ and $\mathcal{P}$ (for the discussion here we allow $\mathcal{P}$ and $\mathcal{Q}$ to be nested, which is different from the main setting of the book). This is then used to construct an asymptotic p-value. However, the main drawback here is once more (a) it is unclear how to generalize these ideas to nonparametric settings without densities and likelihoods, and (b) what to do when the regularity conditions for Wilks' theorem fails, and the asymptotics are unwieldy.

In contrast, we note that universal inference takes a middle ground between the above two approaches, using a mixture (i.e. a prior without Bayesian connotations) over the alternative, but a maximum likelihood over the null:

$$U_n = \frac{\int_{\mathcal{Q}} \prod_{i=1}^n q(X_i) \mu(dq)}{\sup_{p \in \mathcal{P}} \prod_{i=1}^n p(X_i)},$$

which is always an e-variable, under no regularity conditions. However, there are still two drawbacks (a) the previous section shows that this is not optimal, and (b) we still require reference measures and densities to define likelihoods, so nonparametric extensions are not immediately obvious.

The numeraire resolves these issues. This section shows that one should calculate the RIPr $\mathbb{P}^*$ of $\mathbb{Q}^{\text{eff}} := \int_{\mathcal{Q}} \mathbb{Q}^n \mu(d\mathbb{Q})$ onto (the bipolar of) $\{\mathbb{P}^n : \mathbb{P} \in \mathcal{P}\}$. Since $\mathbb{P}^* \ll \mathbb{Q}^{\text{eff}}$ by definition, we can define

$$E_n = \frac{d\mathbb{Q}^{\text{eff}}}{d\mathbb{P}^*}(X_1, \dots, X_n).$$

This is always well defined, even in nonparametric settings with no reference measures for $\mathcal{P}, \mathcal{Q}$, it is always an e-variable (thus guaranteeing the frequentist error control), and it dominates U_n (Section 6.5).

In some situations, the RIPr will actually lie in the convex hull (set of mixtures) of $\{\mathbb{P}^n : \mathbb{P} \in \mathcal{P}\}$, and in this case, the numeraire can be written in the form of B_n . However, for every choice of alternative mixture, there is a unique mixture over the nulls that would yield such a representation (Section 6.4), meaning that one cannot decouple the pair of mixtures as the Bayesian did for B_n . In these cases, the numeraire can be seen as the unique Bayes factor which is also an e-variable.

Of course, when the null and alternative are simple, all the above methods coincide. When only the null is simple, the Bayes factor, the universal inference e-variable, and the numeraire coincide. When neither is simple, the Bayes factor is sometimes (but usually not) an e-variable, the numeraire is sometimes (but usually not) a Bayes factor, and the generalized likelihood ratio is neither an e-variable nor a Bayes factor in general.

6.7 Examples

We now turn to examples of numeraire e-variables, revisiting several examples in Section 2.6. Consider a single random observation (statistic) Z in a measurable space $\mathcal{Z}$, and we take $(\Omega, \mathcal{F})$ to be $\mathcal{Z}$ with its σ -algebra $\sigma(Z)$. We will always have $\mathbb{Q} \ll \mathcal{P}$. Recall that $\mathcal{M}_1$ is the set of all probability measures on $\mathcal{F}$, and denote by $\mathcal{X}_+$ the set of all $[0, \infty]$ -valued measurable functions and by $\mathcal{M}_+$ that of nonnegative measures on $\mathcal{F}$.

6.7.1 Exponential family with one-dimensional sufficient statistic

The example here will be subsumed by the (much shorter) Section 6.7.2. Nevertheless, we provide a self-contained treatment below that does not explicitly exploit the monotone likelihood ratio property, which may be less familiar to some. Here $\mathcal{Z}$ can be general and is equipped with a reference measure $\mathbb{L}$. Consider an exponential family of densities

$$p_\theta(z) = e^{\theta T(z) - A(\theta)}$$

with respect to $\mathbb{L}$, where A is convex and differentiable, the sufficient statistic $T(z)$ is one-dimensional, and the natural parameter θ ranges in some interval $\Theta \subseteq \mathbb{R}$. The null hypothesis is $\mathcal{P} = \{p_\theta : \theta \in \Theta_0\}$ for some closed subset $\Theta_0 \subseteq \Theta$, and the alternative is $q = p_{\theta_1}$ for some $\theta_1 \in \Theta$. We suppose that Θ_0 has a largest element θ_0 and that $\theta_1 > \theta_0$, as in Section 2.6.1. It is natural to conjecture that p_{θ_0} is the RIPr. To confirm this, it suffices to verify the fourth condition in Theorem 6.14. Using the standard formula for the moment generating function of the sufficient statistic we get

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} \left[\frac{p_\theta(Z)}{p_{\theta_0}(Z)} \right] &= \mathbb{E}^{\mathbb{Q}} \left[e^{(\theta - \theta_0)T(Z) - A(\theta) + A(\theta_0)} \right] \\ &= e^{A(\theta_1 + \theta - \theta_0) - A(\theta_1) - A(\theta) + A(\theta_0)} \end{aligned}$$

for all $\theta \in \Theta_0$. Any real convex function f satisfies $f(a + b + c) - f(a + c) \geq f(b + c) - f(c)$ for $a, b \geq 0$ and $c \in \mathbb{R}$. Therefore, $\mathbb{E}^{\mathbb{Q}}[p_\theta(Z)/p_{\theta_0}(Z)] \leq 1$ for all $\theta \in \Theta_0$, so that the fourth condition in Theorem 6.14 holds and p_{θ_0} is indeed the RIPr. As a result the numeraire is the likelihood ratio

$$E^* = \frac{p_{\theta_1}(Z)}{p_{\theta_0}(Z)} = e^{(\theta_1 - \theta_0)T(Z) - A(\theta_1) + A(\theta_0)}.$$

6.7.2 Monotone likelihood ratio

Returning to the example described in Section 2.6.1, we now point out that for a point alternative p_{θ_1} , the e-variable described there is in fact the numeraire. This generalizes the previous example which is well-known to satisfy the monotone likelihood ratio property (in the sufficient statistic). The proof follows by the same argument as above: since $p_{\theta_1}/p_{\theta_0}$ is an e-variable for $\mathcal{P}$, then the fourth condition in Theorem 6.14 implies that p_{θ_0} is the RIPr, and thus $p_{\theta_1}/p_{\theta_0}$ is the numeraire.

6.7.3 Testing symmetry

Take $\mathcal{Z} = \mathbb{R}$. As in Section 2.6.2, we consider the null hypothesis that Z is symmetric,

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : Z \text{ and } -Z \text{ have the same distribution under } \mathbb{P}\},$$

which is a non-dominated family. We also fix an alternative hypothesis $\mathbb{Q}$ that admits a Lebesgue density q . It is natural to conjecture that the RIPr is given by the symmetrization $\tilde{\mathbb{P}}$ of $\mathbb{Q}$, whose density is $\tilde{p}(z) = (q(z) + q(-z))/2$. However, this cannot quite be true in general because $\tilde{\mathbb{P}}$ need not be equivalent to $\mathbb{Q}$. Instead, we claim that the RIPr is the measure $\mathbb{P}^*$ with density

$$p^*(z) = \frac{1}{2} (q(z) + q(-z)) \mathbb{1}_{\{q(z) > 0\}}.$$

This is the absolutely continuous part of $\tilde{\mathbb{P}}$ with respect to $\mathbb{Q}$. It is a probability measure if $\mathbb{Q}$ has symmetric support, and otherwise a proper sub-probability measure. Note that $\tilde{\mathbb{P}}$ belongs to $\mathcal{P}$, which in particular shows that $\mathbb{Q} \ll \mathcal{P}$ since $\mathbb{Q} \ll \tilde{\mathbb{P}}$. (Alternatively, we again have that $\mathbb{P}(A) = 0$ for all $\mathbb{P} \in \mathcal{P}$ implies that A is empty, which also yields $\mathbb{Q} \ll \mathcal{P}$.)

To check that $\mathbb{P}^*$ is the RIPr we will show that the implied candidate numeraire is, in fact, the numeraire. It is given by

$$E^* = \frac{d\mathbb{Q}}{d\mathbb{P}^*} = \frac{2q(Z)}{q(Z) + q(-Z)} \mathbf{1}_{\{q(Z) > 0\}}.$$

(We recall that the numeraire is only defined up to $\mathbb{Q}$ -nullsets.) We claim that the set of all e-variables is

$$\mathfrak{E} = \{E \in \mathcal{X}_+ : E \leq 1 + \phi(Z) \text{ for some odd function } \phi\},$$

where we recall that a function ϕ is odd if $\phi(z) + \phi(-z) = 0$ for all $z \in \mathbb{R}$. Indeed, any E of this form satisfies $\mathbb{E}^{\mathbb{P}}[E] \leq 1 + \mathbb{E}^{\mathbb{P}}[\phi(Z)] = 1$, where the symmetry of Z and the oddness of ϕ were used to get $\mathbb{E}^{\mathbb{P}}[\phi(Z)] = \mathbb{E}^{\mathbb{P}}[\phi(-Z)] = -\mathbb{E}^{\mathbb{P}}[\phi(Z)]$ and hence $\mathbb{E}^{\mathbb{P}}[\phi(Z)] = 0$. Conversely, for any e-variable $E = f(Z)$ we may write $E = \frac{1}{2}(f(Z) + f(-Z)) + \phi(Z)$ where $\phi(z) = \frac{1}{2}(f(z) - f(-z))$ is the odd part of $f(z)$. For any $z \in \mathbb{R}$ we use the symmetric distribution $\mathbb{P} = \frac{1}{2}(\delta_z + \delta_{-z})$ and the fact that E is an e-variable to get $\frac{1}{2}(f(z) + f(-z)) = \mathbb{E}^{\mathbb{P}}[f(Z)] \leq 1$. Hence $E \leq 1 + \phi(Z)$ as required.

We can now verify the numeraire property. First, E^* is $\mathbb{Q}$ -almost surely strictly positive and finite, and it is an e-variable because it can be written as $E^* = 1 + \phi^*(Z)$ where

$$\phi^*(z) = \frac{q(z) - q(-z)}{q(z) + q(-z)}$$

is odd (we also verified that E^* is an e-variable in Section 2.6.2). Next, for any e-variable $E \leq 1 + \phi(Z)$ where ϕ is odd we get

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} \left[\frac{E}{E^*} \right] &\leq \mathbb{E}^{\mathbb{Q}} \left[(1 + \phi(Z)) \frac{q(Z) + q(-Z)}{2q(Z)} \right] \\ &= \mathbb{E}^{\tilde{\mathbb{P}}} [(1 + \phi(Z)) \mathbf{1}_{\{q(Z) > 0\}}] \leq \mathbb{E}^{\tilde{\mathbb{P}}} [1 + \phi(Z)] = 1. \end{aligned}$$

This confirms the numeraire property.

Finally, note that it is not really necessary that $\mathbb{Q}$ admit a density. The symmetrization of $\mathbb{Q}$ is still well-defined as $\tilde{\mathbb{P}} = \frac{1}{2}(\mathbb{Q} + \tilde{\mathbb{Q}})$ where $\tilde{\mathbb{Q}}$ is the distribution of $-Z$ under $\mathbb{Q}$, or equivalently, the push-forward of $\mathbb{Q}$ under the reflection map $z \mapsto -z$. The RIPr is then the absolutely continuous part $\mathbb{P}^* = \tilde{\mathbb{P}}^a$, and the numeraire is any nonnegative version of the Radon–Nikodym derivative $d\mathbb{Q}/d\mathbb{P}^*$ such that $d\mathbb{Q}/d\mathbb{P}^* - 1$ is odd.

6.7.4 Testing bounded means

Let $\mathcal{Z} = [0, 1]$ so that the random variable Z takes values in the unit interval. Fix $\mu \in (0, 1/2)$ and consider (as in Section 2.6.3) the null hypothesis that the mean of Z is at most μ ,

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[Z] \leq \mu\}.$$

There is no single dominating measure for $\mathcal{P}$ since it contains the uncountable non-dominated family $\{\delta_z : z \in [0, \mu]\}$. However, $\mathcal{P}$ is generated by $\mathfrak{E}_0 = \{Z/\mu\}$ in the sense of Corollary 6.16, so our strategy will be to locate a candidate numeraire and then apply the corollary to verify that the candidate is, in fact, the numeraire. The alternative hypothesis $\mathbb{Q}$ is the uniform distribution on $[0, 1]$, and we have $\mathbb{Q} \ll \mathcal{P}$ for the simple reason that $\mathbb{P}(A) = 0$ for all $\mathbb{P} \in \mathcal{P}$ implies that A must actually be empty.

To find a candidate numeraire, we observe that two natural e-variables are Z/μ and the constant one. All convex combinations of these are also e-variables; equivalently, $1 + \lambda(Z - \mu)$ is an e-variable for each $\lambda \in [0, 1/\mu]$. We now look for a log-optimal e-variable in this class by directly maximizing $f(\lambda) = \mathbb{E}^{\mathbb{Q}}[\log(1 + \lambda(Z - \mu))]$ over $\lambda \in [0, \mu^{-1}]$. This is a strictly concave function whose derivative $f'(\lambda) = \mathbb{E}^{\mathbb{Q}}[(Z - \mu)/(1 + \lambda(Z - \mu))]$

satisfies $f'(0) = \mathbb{E}^{\mathbb{Q}}[Z] - \mu > 0$ and $f'(\mu^{-1}) = \mathbb{E}^{\mathbb{Q}}[\mu - \mu^2/Z] = -\infty$. Thus there is a unique interior maximizer $\lambda^* \in (0, \mu^{-1})$, which is characterized by the first-order condition

$$\mathbb{E}^{\mathbb{Q}} \left[\frac{Z - \mu}{1 + \lambda^*(Z - \mu)} \right] = 0. \quad (6.7)$$

Since $\mathbb{Q}$ is the standard uniform distribution we can be more explicit. Nothing changes if we first multiply both sides by λ^* , and then the left-hand side becomes

$$\begin{aligned} \int_0^1 \frac{\lambda^*(z - \mu)}{1 + \lambda^*(z - \mu)} dz &= 1 - \int_0^1 \frac{1}{1 + \lambda^*(z - \mu)} dz \\ &= 1 - \frac{1}{\lambda^*} \log \left(\frac{1 + \lambda^*(1 - \mu)}{1 - \lambda^*\mu} \right). \end{aligned}$$

Thus (6.7) for $\lambda^* \in (0, \mu^{-1})$ is equivalent to

$$\frac{1 + \lambda^*(1 - \mu)}{1 - \lambda^*\mu} = e^{\lambda^*},$$

which is easily solved numerically. This leads us to the candidate numeraire

$$E^* = 1 + \lambda^*(Z - \mu),$$

which is strictly positive and finite. To verify that this is indeed the numeraire, we use (6.7) to get, for any e-variable of the form $X = 1 + \lambda(Z - \mu) = E^* + (\lambda - \lambda^*)(Z - \mu)$, that

$$\mathbb{E}^{\mathbb{Q}} \left[\frac{X}{E^*} \right] = 1 + (\lambda - \lambda^*) \mathbb{E}^{\mathbb{Q}} \left[\frac{Z - \mu}{1 + \lambda^*(Z - \mu)} \right] = 1.$$

Taking $\lambda = 0$ and $\lambda = 1/\mu$ we see that (6.5) of Corollary 6.16 is satisfied and, hence, that E^* is the numeraire and the RIPr $\mathbb{P}^*$ belongs to $\mathcal{P}$.

6.7.5 Testing sub-Gaussian means

Now take $\mathcal{Z} = \mathbb{R}$ and consider the null hypothesis that the observation Z has a 1-sub-Gaussian distribution with nonpositive mean (as in Section 2.6.4). That is, we set

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[e^{\lambda Z - \lambda^2/2}] \leq 1 \text{ for all } \lambda \in [0, \infty) \right\}.$$

The “one-sided” restriction $\lambda \in [0, \infty)$ implies that Z has nonpositive (potentially nonzero and even infinite) mean under any $\mathbb{P} \in \mathcal{P}$. Indeed, monotone convergence and the definition of $\mathcal{P}$ yield $\mathbb{E}^{\mathbb{P}}[Z] = \lim_{\lambda \downarrow 0} \mathbb{E}^{\mathbb{P}}[(e^{\lambda Z} - 1)/\lambda] \leq \lim_{\lambda \downarrow 0} (e^{\lambda^2/2} - 1)/\lambda = 0$. As in the previous example, $\mathcal{P}$ does not admit any dominating measure. It is generated by the family $\mathfrak{E}_0 = \{e^{\lambda Z - \lambda^2/2} : \lambda \in [0, \infty)\}$, so we will again look for a candidate numeraire and verify it using Corollary 6.16. We let the alternative hypothesis $\mathbb{Q}$ be normal with mean $\mu > 0$ and unit variance. As in the previous example, and for the same reason, we have $\mathbb{Q} \ll \mathcal{P}$.

To find a candidate numeraire we maximize $\mathbb{E}^{\mathbb{Q}}[\log X]$ over $X \in \mathfrak{E}_0$. That is, we maximize $\mathbb{E}^{\mathbb{Q}}[\lambda Z - \lambda^2/2] = \lambda\mu - \lambda^2/2$ over $\lambda \in [0, \infty)$. The maximizer is $\lambda^* = \mu$, which yields the candidate

$$E^* = e^{\mu Z - \mu^2/2}.$$

This is finite and strictly positive. Moreover, (6.5) of Corollary 6.16 is satisfied because

$$\mathbb{E}^{\mathbb{Q}} \left[\frac{e^{\lambda Z - \lambda^2/2}}{e^{\mu Z - \mu^2/2}} \right] = \mathbb{E}^{\mathbb{Q}} \left[e^{(\lambda - \mu)(Z - \mu) - (\lambda - \mu)^2/2} \right] = 1.$$

We conclude that E^* is the numeraire and, consequently, that the RIPr $\mathbb{P}^*$ is the standard normal distribution. This example can be easily generalized to σ -sub-Gaussian distributions for $\sigma \neq 1$, but we omit this for brevity.

6.7.6 Testing likelihood ratio bounds

Consider the setting in Section 2.6.6, where for a fixed probability measure $\mathbb{P}_0$ and $\gamma \in [1, \infty)$, we test the null hypothesis

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : d\mathbb{P}/d\mathbb{P}_0 \leq \gamma\}$$

against an alternative $\mathbb{Q} \ll \mathbb{P}_0$. We have seen in Section 2.6.6 that the random variable E^* is an e-variable for $\mathcal{P}$, specified by

$$E^* = \frac{Z^* \vee z_0}{\gamma},$$

where $Z^* = d\mathbb{Q}/d\mathbb{P}_0$ and $z_0 \geq 0$ is the largest constant such that $\int_0^{1/\gamma} (q_t(Z^*) \vee z_0) dt = 1$. Recall our notation of the left quantile function $q_t(Z) = \inf\{x \in \mathbb{R} : \mathbb{P}_0(Z \leq x) \geq 1 - t\}$ for $t \in (0, 1)$ and any random variable Z .

We next show that it is indeed the numeraire for testing $\mathcal{P}$ against $\mathbb{Q}$. Let $\mathbb{P}^* \in \mathcal{M}_+$ be specified by

$$\frac{d\mathbb{P}^*}{d\mathbb{P}_0} = \gamma \wedge \frac{\gamma Z^*}{z_0},$$

where Z^*/z_0 is set to 0 if $z_0 = 0$ and $Z^* = 0$, and it is set to ∞ if $z_0 = 0$ and $Z^* > 0$. By straightforward computations, we can check that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}^*$, $\mathbb{P}^*(\Omega) \leq 1$, and $d\mathbb{P}^*/d\mathbb{P}_0 \leq \gamma$. Moreover,

$$\frac{d\mathbb{Q}}{d\mathbb{P}^*} = \frac{d\mathbb{Q}}{d\mathbb{P}} \frac{d\mathbb{P}}{d\mathbb{P}^*} \mathbf{1}_{\{Z^* > 0\}} = \frac{Z^* \vee z_0}{\gamma} = E^*.$$

Therefore, by Theorem 3.20 (it is easy to see that the result holds true even if $\mathbb{P} \in \mathcal{M}_+$ is not a probability measure), E^* maximizes $\mathbb{E}^{\mathbb{Q}}[\log X]$ under the constraint $\int X d\mathbb{P}^* \leq 1$. Since an e-variable for $\mathcal{P}$ always satisfies the constraint $\int X d\mathbb{P}^* \leq 1$, we know that E^* is also log-optimal for testing $\mathcal{P}$ against $\mathbb{Q}$, and thus it is the numeraire.

The special case of $\gamma = 1$ and $\mathbb{Q} \ll \mathbb{P}_0$ in the above example is precisely Theorem 3.20, with the log-optimal e-variable $E^* = Z^* = d\mathbb{Q}/d\mathbb{P}_0$ and z_0 being the largest constant such that $z_0 \leq Z^*$ almost surely (typically $z_0 = 0$).

6.7.7 Testing exchangeability

Consider the problem of testing exchangeability in Section 2.6.7. For the data $Z = (Z_1, \dots, Z_n)$, the null hypothesis of exchangeability is

$$\mathcal{P} = \{\mathbb{P} : Z \stackrel{d}{=} Z_\sigma \text{ under } \mathbb{P} \text{ for all permutations } \sigma\},$$

and $\mathbb{Q}$ is an alternative distribution of Z with density q , which is not in $\mathcal{P}$. Let π be uniformly drawn from $\mathcal{S}_n$. In Section 2.6.7, we have verified that the random variable

$$E = \frac{q(Z)}{q(Z_\pi)} = \frac{q(Z)}{\frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} q(Z_\sigma)}$$

is an e-variable for $\mathcal{P}$. To see that it is the numeraire for $\mathcal{P}$ against $\mathbb{Q}$ and further, $q(Z_\pi)$ is the RPr of q onto $\mathcal{P}$, we note that the distribution $\mathbb{P}^*$ with density $q(Z_\pi)$ is exchangeable, hence in $\mathcal{P}$. This then yields the above claim by Theorem 6.15, since $d\mathbb{Q}/d\mathbb{P}^*$ is the only e-variable which can be represented as a likelihood ratio of $\mathbb{Q}$ to some element of the bipolar $\mathcal{P}^{\circ\circ} \supseteq \mathcal{P}$.

Bibliographical note

This chapter is largely shaped by treatment in Larsson et al. [2025a]. Historically, the RPr first appeared in Csiszár and Tusnády [1984]. Foundational work was done by Li [1999], who proved (under some assumptions) its uniqueness and an inequality that (in hindsight) yields the e-variable property of the numeraire, amongst

several other facts. Grünwald et al. [2024a] first described its key role in constructing optimal e-variables. Lardy et al. [2024] subsequently extended the original definition to more general settings by relaxing some conditions. Finally, Larsson et al. [2025a] eliminated all conditions necessary to define the RIPr, but doing so required defining it via the numeraire e-variable, as done in this chapter.

Grünwald et al. [2024a] also consider composite alternatives. They first consider a worst-case optimality criterion that they call GROW: finding an e-variable for $\mathcal{P}$ that maximizes the worst-case e-power over $\mathcal{Q}$; we find this to be a rather pessimistic objective (such an e-variable tuned for the worst case would not adapt to simpler alternatives that could admit larger growth rates), but it does involve some nice mathematics such as a *joint information projection*. This motivates their REGROW criterion, which is similar to our asymptotic log-optimality objective (6.6).

The proof of Lemma 6.18 can be found in Larsson et al. [2025a]. The numeraire for testing exchangeability in Section 6.7.7 was obtained by Koning [2023b] and Larsson et al. [2025a] using different techniques. The numeraire for null hypotheses that satisfy a different notion of group invariance has been derived in Pérez-Ortiz et al. [2024]. The numeraire for one-parameter exponential families was derived in Grünwald et al. [2024b], and then in Larsson et al. [2025a] using different techniques; we follow the latter presentation. Exponential families are treated in much more generality and detail by Grünwald et al. [2024b]. Csiszar and Matus [2003] has more details on the reverse information projection for exponential families. The bounded mean example is a variant of one studied in Waudby-Smith and Ramdas [2024], with the corresponding duality theory derived explicitly in Honda and Takemura [2010].

The minimum KL divergence in Theorem 6.13 is closely related to, yet different from, the KL_{inf} metric found in the modern multi-armed bandit literature; see Agrawal et al. [2021] for a recent example. The exact relationships are yet to be fully worked out. The strong duality in Theorem 6.13 also holds for some other divergences than Kullback–Leibler (corresponding to $\phi(E)$ instead of $\log E$ for some concave ϕ). See [Larsson et al., 2025a, Section 6] for an example involving Rényi divergences, with $\phi(x) = x^{1-\gamma}/(1-\gamma)$ for $\gamma > 1$.

The fact that numeraire e-variables are in general logically incoherent (in contrast to universal inference) was first observed by Bickel [2024].

Chapter 7

Sequential anytime-valid inference using e-processes

Classical sequential testing, as developed by Wald, involves specifying a particular stopping rule that achieves a prespecified bound on the type-I and type-II errors. In contrast, this chapter deals with several central concepts for sequential *anytime-valid* inference (SAVI). The latter is a new paradigm for sequential testing (or estimation) that allows the statistician to stop at any arbitrary stopping time, possibly not anticipated or specified in advance.

7.1 The problematic practice of peeking at p-values

Let us motivate SAVI procedures by describing the not-so-uncommon practice of *peeking at p-values*, and identify the heart of the problem.

Consider a scientist in a psychology laboratory in a university who painstakingly collects data, subject by subject, where each subject has to undergo an hour-long experiment. It appears entirely reasonable that they may be interested in examining the data after every few subjects, for example by calculating a p-value. For simplicity, assume that the scientist calculates a p-value P_n after seeing n subjects, so that $P_1, P_2, \dots$ is a sequence of p-values calculated on increasing amounts of data. (Assume for now that the statistical models are well specified so that if the null hypothesis is true, P_n is indeed uniformly distributed.)

Peeking at p-values refers to the practice of stopping data collection when $P_n \leq \alpha$ (such as 0.05, or some other appropriate threshold, e.g., 0.04, if the scientist wants to avoid the appearance of stopping just after the data appears significant at the 0.05 level). Why is this an issue? Because the final sample size of the experiment is random, and in particular it is a stopping time

$$\tau := \inf\{n \geq 1 : P_n \leq \alpha\}.$$

The heart of the issue is that even though

$$\mathbb{P}(P_n \leq \alpha) \leq \alpha, \text{ for each } n \in \mathbb{N},$$

it is to be typically expected that

$$\mathbb{P}(P_\tau \leq \alpha) > \alpha.$$

To elaborate, even under the null hypothesis, it would be expected under most typical situations (like a t-test p-value) that τ is almost surely finite, meaning that one will eventually stop and reject the null, and $\mathbb{P}(P_\tau \leq \alpha) = 1$, meaning that we are predestined to make a type-I error! This is sometimes called “sampling to a foregone conclusion”.

Of course, the catch is that when the data are reported in a scientific article, it is rarely reported that the data were collected in this fashion (perhaps because the scientist may not even realize that this could

create issues). Thus, it may be hard to detect and correct for such issues from the data alone: one really needs to know the statistical protocol for collecting the data to identify this issue. This issue is well known to be a key factor (of many) underlying the “reproducibility crisis” in many disciplines, such as psychology and biomedical and social sciences, that was recognized and heavily discussed in the second decade of the twenty-first century [Baker, 2016].

Since the desire to peek and analyze the data as it is being collected is fairly natural (who does not want to potentially stop early and save time and money if that is possible?), it is sensible to provide statistical techniques that can account for such peeking (without really knowing any of its details), rather than seeking to ban it (the current solution, which cannot be enforced). That will be the goal of this chapter: to develop sequential anytime-valid inferential (SAVI) procedures.

We begin by summarizing the dominant paradigm for sequential inference introduced by Wald, which demonstrates the allure of sequential methods and dispels a myth that they are less effective than fixed sample-size procedures because one has to account for the multiple looks at the data.

7.2 Wald’s sequential probability ratio test

Before we get into formal definitions, we present a numerical example that can be quite surprising to those unfamiliar with sequential analysis. It convincingly brings out the allure of sequential methods compared to batch methods (those that collect and analyze data as a single batch).

Suppose we have a coin and wish to test whether it is biased or not. Mathematically, we have access to data $X_1, X_2, \dots$ which are iid Bernoulli(p) random variables with unknown $p \in (0, 1)$. Under the null hypothesis, we suppose $p = 1/2$. For the sake of simplicity, under the alternative hypothesis it is $p = 0.6$. Let the null and alternative densities be represented by f_0, f_1 respectively. For a fixed sample size test, the classic Neyman–Pearson lemma dictates that the optimal test is the likelihood ratio test (LRT), which thresholds the likelihood ratio

$$L_n = \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)},$$

meaning that we reject the null hypothesis when L_n is larger than some threshold γ . Different choices of n and γ yield different type-I and type-II errors achieved by the test, but the classic Neyman–Pearson lemma promises that no other test can achieve a better tradeoff of the two errors. If we desire a type-I error of $\alpha = 0.05$ and a type-II error of $\beta = 0.05$, a numerical calculation shows that one requires a sample of size $n = 280$ for our problem setup, and the LRT rejects when L_{280} exceeds $\gamma = 0.962$ (equivalently, when at least 154 heads are observed). Note that L_n is an e-variable, but the Neyman–Pearson threshold is *much* smaller than $1/\alpha = 20$, but is rather a quantile of L_n . It is indeed odd that we can reject the null for a threshold smaller than 1, meaning even when the null has greater likelihood than the alternative.

In contrast, Wald’s sequential probability ratio test (SPRT) works by sampling the data one by one, continually updating L_t . Indeed, $(L_t)_{t \in \mathbb{N}}$ is an e-process, which will be formally studied in this chapter, so the SPRT has a lot in common with the SAVI and e-value methodology, but also several differences that are highlighted at the end of this section.

We describe the SPRT below. If L_t falls below some threshold γ_0 , we give up and accept the null (more rigorously, we do not reject the null); if it goes above some other threshold γ_1 , we reject the null. Else, we keep sampling. It turns out that $\gamma_0 = \beta$ and $\gamma_1 = 1/\alpha$ suffices to keep the type-I and type-II errors below α and β respectively, as a consequence of Ville’s inequality, to be introduced later in this chapter. (These choices are more conservative than the approximate thresholds of $\beta/(1 - \alpha)$ and $(1 - \beta)/\alpha$ often recommended for the SPRT.) It can be shown that this procedure will stop almost surely under the null or alternative.

In summary, our experiments pitch the Neyman–Pearson LRT rule:

$$\text{reject if } L_{280} \geq 0.963,$$

against a conservative variant of Wald’s SPRT rule, which does for $t \geq 1$:

$$\text{reject if } L_t \geq 20, \text{ accept if } L_t \leq 0.05, \text{ continue sampling otherwise.}$$

Data-generating distribution Setting	$p = 0.5$ null	$p = 0.55$ misspecified	$p = 0.6$ alternative	$p = 0.65$ misspecified
SPRT expected sample size	137.7	234.5	139.3	76.1
LRT pre-specified sample size	280	280	280	280
Power of SPRT	0.04	0.52	0.96	1.00
Power of LRT	0.05	0.48	0.95	1.00

Table 7.1: A comparison of SPRT and LRT for testing $p = 0.5$ against $p = 0.6$, with target type-I error rate at 0.05 and target type-II error rate at 0.05. All numbers are produced with Monte Carlo simulations using 10,000 repetitions.

Numerical experiments show that for the SPRT, the expected sample size (equivalently, expected stopping time) under the null is 137.7, while under the alternative it is 139.3. This is about half the number of samples needed by the LRT, a remarkable improvement over the optimal batch test in such a simple setting!

We can also test the methods out when the data is neither drawn from the null or alternative; this is often called the *misspecified* setting. When the bias of the coin is smaller than anticipated, e.g. $p = 0.55$, the LRT has power 0.52 while the SPRT has power 0.48, which is quite similar, but the expected sample size of SPRT is 234.5. When the bias is larger than anticipated, e.g., $p = 0.65$, both tests have power equal to one, but the SPRT expected sample size is just 76.1. The numbers are summarized in Table 7.1.

Wald proved that for simple nulls and alternatives, his SPRT provided the optimal tradeoff between type-I error, type-II error and stopping time. He showed that any other stopping rule with lower type-I and type-II errors must have larger expected stopping time. In light of this result, the aforementioned improvement of the SPRT over the LRT is not a coincidence. We record this below for ease of reference.

Theorem 7.2: Wald’s SPRT over Neyman–Pearson’s LRT

Let α, β be the type-I and type-II errors of the LRT with n samples (using some threshold). Let α', β' be the type-I and type-II errors of the SPRT (using some thresholds), and τ be its stopping time. If $\alpha \leq \alpha'$ and $\beta \leq \beta'$, then $\max(\mathbb{E}^{\mathbb{P}}[\tau], \mathbb{E}^{\mathbb{Q}}[\tau]) \leq n$.

Wald’s paradigm for sequential testing has been hugely successful: the literature on sequential analysis is large and mature. Although using the same process and similar stopping rules, the SAVI goals deviate from Wald’s goals in four significant ways:

1. **Decisions versus evidence:** Wald focused on decision theory, that is making a decision quickly with predefined error rates. Our focus here is on designing measures of evidence that can be sequentially updated. Of course, our evidence measures can be used to make decisions, but that is a secondary and not primary goal for us. Said differently, we do not necessarily make binary decisions (accept or reject). When using SAVI methods to make binary decisions, our rules will resemble Wald’s.
2. **Single stopping time versus all stopping times:** Wald focused on designing a single stopping time (or rule) at which the scientist would reject or accept the null. We will instead provide a notion of evidence (the e-process) that is anytime-valid in the sense that it has a guarantee (under the null) at any stopping time. When SAVI techniques are applied to the setting of Wald’s SPRT, the likelihood ratio e-process gives us a way of continually examining and updating evidence before (and leading up to) Wald’s stopping time.
3. **Differing optimality criteria:** Our notion of optimality — log-optimality — is very different from Wald’s notion of optimal tradeoffs of the two errors. For simple nulls and alternatives, it is remarkable that the same statistic (the likelihood ratio martingale) simultaneously optimizes both objectives, but this will in general not be true.
4. **Beyond likelihood ratios:** While the SPRT techniques presented above can be extended to certain composite nulls and alternatives, it has typically always dealt with settings where likelihood ratios exist

(e.g., parametric settings, with a common reference measure). However, e-processes provide us with a universal language of handling *any* sequential testing problem: we will show that every sequential test (for any composite, possibly nonparametric, null) can be obtained using e-processes.

We now move on to defining the basic concepts in the paradigm of e-processes.

7.3 Test supermartingales, sequential tests, and e-processes

There are four main SAVI tools: e-processes, p-processes, $(1 - \alpha)$ -confidence sequences and level- α sequential tests; it is immediately apparent that the latter two are associated with a predefined level $\alpha \in [0, 1]$ but the former two are not. These are respectively the sequential analogs of e-variables, p-variables, $(1 - \alpha)$ -confidence intervals and level- α tests. We will primarily focus on e-processes because they are the central object; the others can be easily obtained as a consequence of Ville's inequality applied to one or more e-processes.

We begin with the technical background necessary for the sequential setting. We assume that the reader is familiar with terminologies in stochastic processes, recapped below.

We work in a fixed filtered probability space, meaning that $\mathcal{F}$ refers to a filtration $(\mathcal{F}_t)_{t \in \mathbb{T}}$ (a nested sequence of σ -algebras), where $\mathbb{T}$ is a finite or infinite set of indices starting from 0. We only deal with discrete-time processes, so $\mathbb{T} = \mathbb{N}_0$ or $\{0, 1, \dots, n\}$ for some $n \in \mathbb{N}$. Let $\mathcal{F}_\infty = \bigcup_{t \in \mathbb{T}} \mathcal{F}_t$. Further denote by $\mathbb{T}_+$ the set $\mathbb{T} \setminus \{0\}$.

A sequence of random variables $Y = (Y_t)_{t \in \mathbb{T}}$ is called a *process* if it is adapted to $\mathcal{F}$ —i.e., if Y_t is measurable with respect to $\mathcal{F}_t$ for every t . A process Y is called *predictable* if Y_t is measurable with respect to $\mathcal{F}_{t-1}$ for each $t \geq 1$. Inequalities comparing processes are understood to hold at each $t \in \mathbb{T}$.

Often $\mathcal{F}$ is chosen as the natural filtration of data, that is, $\mathcal{F}_t := \sigma(X^t)$, where $X^t = (X_1, \dots, X_t)$, with $\mathcal{F}_0$ being trivial ($\mathcal{F}_0 = \{\emptyset, \Omega\}$). In this case Y_t being measurable with respect to $\mathcal{F}_t$ means that Y_t is a measurable function of X^t . But $\mathcal{F}$ is sometimes a coarser filtration (we discard information) or a richer one (for example allowing $\mathcal{F}_0 = \sigma(U_0)$ for a uniformly distributed random variable U_0 that is independent of all other randomness).

A stopping time (or rule) τ is a nonnegative integer-valued random variable such that $\{\tau \leq t\} \in \mathcal{F}_t$ for each $t \geq 0$. In words: we know at each time whether the rule is telling us to stop or keep going. Denote by $\mathcal{T}$ the set of all stopping times, implicitly depending on $\mathcal{F}$, including ones that may never stop. For a stopping time τ , the σ -algebra $\mathcal{F}_\tau$ is defined by $\{A \in \mathcal{F}_\infty : A \cap \{\tau \leq t\} \in \mathcal{F}_t \text{ for every } t \in \mathbb{T}\}$.

As in previous chapters, the set of distributions $\mathcal{P}$ represents our null hypothesis. A level- α sequential test for $\mathcal{P}$ is a binary process ϕ such that

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\exists t \geq 1 : \phi_t = 1) \leq \alpha. \quad (7.1)$$

Interpreting $\phi_t = 1$ as rejecting $\mathcal{P}$, the above condition means that the probability that we will *ever* falsely reject the null is at most α , meaning that with probability $1 - \alpha$, ϕ equals the sequences of all zeros. These are also called one-sided or open-ended or power-one tests.

Without loss of generality, we can assume that if $\phi_t = 1$, then $\phi_s = 1$ for all $s \geq t$. Further, we will define $\phi_\infty = \lim_{t \rightarrow \infty} \phi_t = \sup_{t \in \mathbb{N}} \phi_t$, meaning that if a test did not reject at any finite time, then it also does not reject at infinity.

Then, we can more easily identify a sequential test by the stopping time $\tau := \inf\{t \geq 1 : \phi_t = 1\}$ with the convention $\inf \emptyset = \infty$. Then, a level- α test requires that $\tau < \infty$ with probability at most α under every $\mathbb{P} \in \mathcal{P}$. We can then rephrase the requirement in (7.1) as:

$$\mathbb{E}^\mathbb{P}[\phi_\tau] \leq \alpha, \quad (7.2)$$

for all $\mathbb{P} \in \mathcal{P}$ and all $\tau \in \mathcal{T}$.

In the discrete-time setting of this book, an integrable process M is a *martingale* for $\mathbb{P}$ if

$$\mathbb{E}^\mathbb{P}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1} \quad (7.3)$$

for all $t \in \mathbb{T}_+$, where all such relations between random variables are understood to hold $\mathbb{P}$ -almost surely. M is a *supermartingale* for $\mathbb{P}$ if it satisfies (7.3) with “=” relaxed to “ $\leq$ ”.

Definition 7.3: Test (super)martingales and e-processes

- (i) A process $M = (M_t)_{t \in \mathbb{T}}$ is called a *test (super)martingale* for $\mathcal{P}$ if for every $\mathbb{P} \in \mathcal{P}$, it satisfies: (a) M is $\mathbb{P}$ -almost surely nonnegative, (b) M is a (super)martingale under $\mathbb{P}$, and (c) $\mathbb{E}^{\mathbb{P}}[M_0] \leq 1$. A family of processes $(M^{\mathbb{P}})_{\mathbb{P} \in \mathcal{P}}$ is called a *test (super)martingale family* if each $M^{\mathbb{P}}$ is a test (super)martingale for $\mathbb{P}$.
- (ii) A sequence of e-values $E = (E_t)_{t \in \mathbb{T}}$ that is adapted to $\mathcal{F}$ is called an *e-process* if it satisfies one of two equivalent conditions: (a) $\mathbb{E}^{\mathbb{P}}[E_{\tau}] \leq 1$ for any stopping time $\tau \in \mathcal{T}$ and any $\mathbb{P} \in \mathcal{P}$; (b) there exists a test martingale family $(M^{\mathbb{P}})_{\mathbb{P} \in \mathcal{P}}$ such that $E \leq M^{\mathbb{P}}$, $\mathbb{P}$ -almost surely for each $\mathbb{P} \in \mathcal{P}$.

The fact that the above two definitions (ii.a) and (ii.b) for an e-process are equivalent is nontrivial, and its proof is omitted. In item (ii.b), one can require $(M^{\mathbb{P}})_{\mathbb{P} \in \mathcal{P}}$ to be a test supermartingale family instead of a test martingale family, without altering the definition. This is a consequence of Doob’s decomposition theorem; we invite the reader to fill in the details. It is clear that test (super)martingales are e-processes, but not vice versa. We sometimes omit “for $\mathcal{P}$ ” (as we did above) and it should be clear from context.

Definition 7.4: Asymptotic growth rate and log-optimality

An e-process M is said to be *consistent* against $\mathbb{Q}$ if $M_t \rightarrow \infty$, $\mathbb{Q}$ -almost surely as $t \rightarrow \infty$. The lower limit

$$\liminf_{t \rightarrow \infty} \frac{\log M_t}{t}$$

is typically $\mathbb{Q}$ -almost surely a constant (which depends on $\mathbb{Q}$ and may be infinite), and in that case, we call that constant the *asymptotic growth rate of M against $\mathbb{Q}$* .

Note that the asymptotic growth rate of any e-process for $\mathcal{P}$ against $\mathbb{P} \in \mathcal{P}$ is always nonpositive. A positive asymptotic growth rate implies consistency, but not vice versa; when an e-process is consistent against $\mathbb{Q}$, it need not have a positive asymptotic growth rate against $\mathbb{Q}$. But in many simple examples (say involving iid data), the asymptotic growth rate will indeed be positive.

E-processes can be converted to sequential tests using Ville’s inequality, which we introduce next.

7.4 Optional stopping and Ville’s inequality

We first state a cornerstone fact in probability theory, the optional stopping theorem, usually attributed to Doob. A precursor to the optional stopping theorem is the following fact, called the supermartingale convergence theorem: Any test supermartingale M for $\mathbb{P}$ has a well-defined limit, meaning that the random variable M_{∞} exists as the (finite) $\mathbb{P}$ -almost sure limit of M_n .

Fact 7.5: Optional stopping theorem

If M is a test (super)martingale for $\mathbb{P}$, then for any $\mathcal{F}$ -stopping times τ and $\sigma \leq \tau$,

$$\mathbb{E}^{\mathbb{P}}[M_{\tau} \mid \mathcal{F}_{\sigma}] \leq M_{\sigma}.$$

In particular, by taking $\sigma = 0$, we have

$$\mathbb{E}^{\mathbb{P}}[M_{\tau}] \leq 1.$$

Consequently, if M is an e-process for $\mathcal{P}$, then $\mathbb{E}^{\mathbb{P}}[M_{\tau}] \leq 1$ for all $\mathbb{P} \in \mathcal{P}$.

It is important to note that this particular version of the optional stopping theorem does not have any restrictions placed on the stopping time, and this is primarily due to the nonnegativity of the underlying

test supermartingale or e-process. If M were not nonnegative, conditions would have to be placed on the stopping time or boundedness of the process for the same result to hold, but we do not concern ourselves with such variants.

For our purposes, the relevant version of Ville's inequality can be stated as follows.

Fact 7.6: Ville's inequality

If M is an e-process for $\mathcal{P}$, then the following three equivalent statements hold:

$$\begin{aligned} & \mathbb{P} \left(\exists t \in \mathbb{N} : M_t \geq \frac{1}{\alpha} \right) \leq \alpha \text{ for every } \mathbb{P} \in \mathcal{P} \text{ and } \alpha \in [0, 1]; \\ \iff & \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P} \left(\sup_{t \in \mathbb{N}} M_t \geq \frac{1}{\alpha} \right) \leq \alpha \text{ for every } \alpha \in [0, 1]; \end{aligned} \quad (7.4a)$$

$$\iff \sup_{\mathbb{P} \in \mathcal{P}, \tau \geq 0} \mathbb{P} \left(M_\tau \geq \frac{1}{\alpha} \right) \leq \alpha \text{ for every } \alpha \in [0, 1], \quad (7.4b)$$

where τ is any $\mathcal{F}$ -stopping time taking values in $[0, \infty]$.

Clearly, for any fixed t , M_t is an e-variable. Thus, $\mathbb{P}(M_t \geq 1/\alpha) \leq \alpha$ holds by Markov's inequality. As such, Ville's inequality is best seen as a time-uniform generalization of Markov's inequality that applies to e-processes rather than e-variables.

Note that (7.4a) and (7.4b) often only holds with inequality, but they can sometimes hold with equality. For example, if we were in continuous rather than discrete time, and $(B_t)_{t \in \mathbb{R}_+}$ is a standard (centered) Brownian motion, then for any nonzero λ , the process $M = (\exp(\lambda B_t - \lambda^2 t/2))_{t \in \mathbb{R}_+}$ can be checked to be a nonnegative continuous-time martingale. (A continuous-time martingale $M = (M_t)_{t \in \mathbb{R}_+}$ satisfies $\mathbb{E}[M_t | \mathcal{F}_s] = M_s$ for all $0 \leq s < t$, where $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ is an increasing chain of σ -algebras.) In this case, Ville's inequality holds with equality. In general, whether or not equality holds depends on whether M *overshoots* $1/\alpha$ at the stopping time $\tau := \inf\{t \geq 1 : M_t \geq 1/\alpha\}$, where infimum of an empty set equals infinity. A more detailed discussion of the tightness of Ville's inequality is out of the scope of this book.

We also remark that a conditional version of Ville's inequality is also true, though we do not utilize it much. Specifically, if M is a test supermartingale for $\mathcal{P}$, then

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P} \left(\exists t \geq s : M_t \geq \frac{M_s}{\alpha} \middle| \mathcal{F}_s \right) \leq \alpha \text{ for every } \alpha \in [0, 1].$$

Ville's inequality effectively states that $\phi_t = \mathbb{1}_{\{M_t \geq 1/\alpha\}}$ yields a level- α sequential test; see Figure 7.7 for a simple illustration. A converse fact is also true, as presented next.

Proposition 7.8

Every level- α sequential test ϕ for $\mathcal{P}$ can be obtained by thresholding some e-process M for $\mathcal{P}$ at $1/\alpha$.

Proof.

Define $M_t = \phi_t/\alpha$. Then M_t only takes on two values: 0 and $1/\alpha$. So $M_t \geq 1/\alpha$ if and only if $\phi_t = 1$. The only thing left to check is that M is an e-process for $\mathcal{P}$. Indeed, for any stopping time τ and any $\mathbb{P} \in \mathcal{P}$, we have $\mathbb{E}^\mathbb{P}[M_\tau] = \mathbb{E}^\mathbb{P}[\phi_\tau/\alpha] \leq 1$ by (7.2). This completes the proof.

Thus, e-processes provide a complete framework for sequential testing. Instead of focusing on the test, we can instead focus on constructing e-processes. However, we do not only think of e-processes as a tool to derive tests. We think of e-processes as measures of evidence against the null: the larger they are (or become), the more evidence we have against the null. We may threshold them to make decisions if we wish to, but they retain interpretability without thresholding as well.

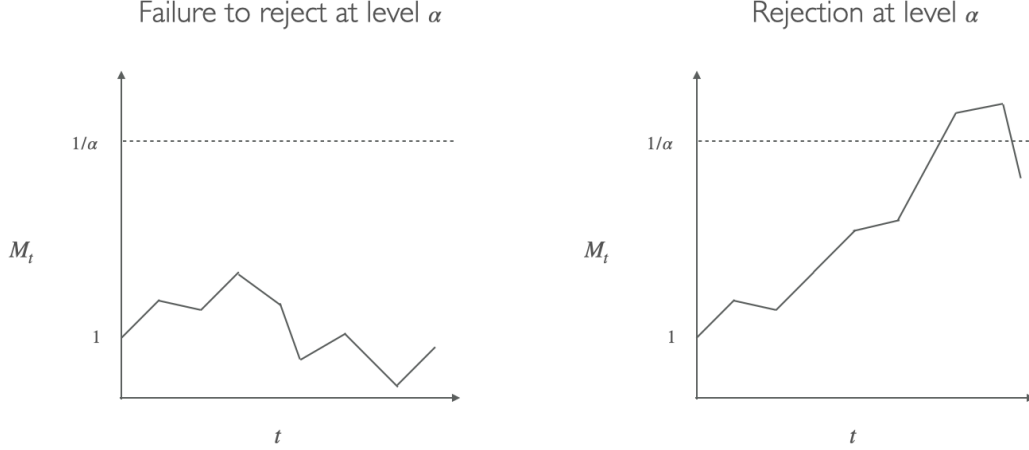


Figure 7.7: A visualization of sequential testing with Ville's inequality applied to an e-process M .

A practical use of e-processes without thresholding concerns the property of optional continuation (discussed in Section 4.3), meaning that another scientist can continue an e-process obtained by a previous scientist (peeking is allowed) by designing new experiments, and multiplying the new e-process to the value of the previous e-process, to build a new e-process. Mathematically, this means, if M_s^A is the time- s value of an e-process M^A , and M^B is another e-process with $M_t^B = 1$ for all $t \leq s$, then M defined by $M_t = M_{s \wedge t}^A M_t^B$ is an e-process. Intuitively, up to time s this process coincides with M^A , and after time s the process carries the evidence obtained from M^A and continues with M^B . The same statement holds true if s is not a fixed time, but a stopping time. We put this part formally in the next proposition.

Proposition 7.9

Let M^A and M^B be two e-processes for $\mathcal{P}$ and τ a stopping time, such that $M_t^B = 1$ on $\{t \leq \tau\}$. Define a process M by $M_t = M_{\tau \wedge t}^A M_t^B$ for $t \in \mathbb{T}$. Then M is an e-process for $\mathcal{P}$.

Proof.

For any $\mathbb{P} \in \mathcal{P}$ let $M^{A,\mathbb{P}}$ and $M^{B,\mathbb{P}}$ be two test supermartingales that dominate M^A and M^B , respectively, as in Definition 7.3 (ii.b). Clearly, $M_t \leq M_{\tau \wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}}$. Below we show that $M^\mathbb{P} := (M_{\tau \wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}})_{t \in \mathbb{T}}$ is a test supermartingale for $\mathbb{P}$, which is sufficient to justify that M is an e-process. For $t \geq 1$, noting that $\{\tau < t\} \in \mathcal{F}_{t-1}$ and $\mathbb{1}_{\{\tau < t\}} M_{\tau}^{A,\mathbb{P}}$ is $\mathcal{F}_{t-1}$ -measurable, we get

$$\begin{aligned}
\mathbb{E}^\mathbb{P}[M_t^\mathbb{P} \mid \mathcal{F}_{t-1}] &= \mathbb{E}^\mathbb{P}[M_{\tau \wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}} \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}^\mathbb{P}[M_{\tau \wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}} \mathbb{1}_{\{\tau < t\}} \mid \mathcal{F}_{t-1}] + \mathbb{E}^\mathbb{P}[M_{\tau \wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}} \mathbb{1}_{\{\tau \geq t\}} \mid \mathcal{F}_{t-1}] \\
&= \mathbb{1}_{\{\tau < t\}} M_{\tau}^{A,\mathbb{P}} \mathbb{E}^\mathbb{P}[M_t^{B,\mathbb{P}} \mid \mathcal{F}_{t-1}] + \mathbb{1}_{\{\tau \geq t\}} \mathbb{E}^\mathbb{P}[M_t^{A,\mathbb{P}} \mid \mathcal{F}_{t-1}] \\
&\leq \mathbb{1}_{\{\tau < t\}} M_{\tau}^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} + \mathbb{1}_{\{\tau \geq t\}} M_{t-1}^{A,\mathbb{P}} \\
&= \mathbb{1}_{\{\tau \leq t-1\}} M_{\tau}^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} + \mathbb{1}_{\{\tau > t-1\}} M_{t-1}^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} \\
&= M_{\tau \wedge (t-1)}^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} = M_{t-1}^\mathbb{P}.
\end{aligned}$$

Hence, $M^\mathbb{P}$ is a test supermartingale, implying that M is an e-process according to Definition 7.3 (ii.b).

We make some more observations on the result in Proposition 7.9: First, whether and when the process switches from M^A to M^B depends on data as well as other sources such as personal judgment after seeing data, which are all included in the filtration $\mathcal{F}$. Second, the process M^B does not need to be specified before τ ; it only needs to be designed after τ happens. Third, switching from one e-process to another can happen multiple times, and the output is always an e-process; this follows by applying Proposition 7.9 repeatedly.

We end this section by noting that Ville's inequality (7.4a) immediately implies that if M is an e-process, then $(\inf_{s \leq t} 1/M_s)_{t \in \mathbb{T}}$ is a p-process, as defined next.

Definition 7.10

A nonnegative process $P = (P_t)_{t \in \mathbb{T}}$ is called a *p-process* if P_τ is a p-variable for any stopping time $\tau \in \mathcal{T}$ and any $\mathbb{P} \in \mathcal{P}$.

Thus an e-process measures evidence based on what we currently have, while a p-process measures evidence based on the best evidence we ever accumulated in the past. Even though all p-processes are not of the form $(\inf_{s \leq t} 1/M_s)_{t \in \mathbb{T}}$ for an e-process M , this is certainly a common way to derive p-processes. In this case, it may be clear why p-processes perhaps provide an exaggerated sense of evidence: If only the quality, performance and promise of companies could be judged by the maximum wealth they ever had! We present the financial interpretation of e-processes in Section 7.6.

7.5 E-processes constructed from likelihood ratios

7.5.1 Log-optimality of likelihood ratio processes

Suppose that we are testing a simple hypothesis $\mathbb{P}$ versus a simple hypothesis $\mathbb{Q} \ll \mathbb{P}$. If we observe iid data $X_1, X_2, \dots$ sequentially, then the likelihood ratio process M given by

$$M_0 = 1 \quad \text{and} \quad M_t = \prod_{i=1}^t \frac{d\mathbb{Q}}{d\mathbb{P}}(X_i) \quad \text{for } t \in \mathbb{N} \quad (7.5)$$

is an e-process adapted to the filtration generated by the data, where $d\mathbb{Q}/d\mathbb{P}(X)$ is the likelihood ratio of $\mathbb{Q}$ to $\mathbb{P}$ observing X . We have encountered this e-process in Section 7.2, used in the SPRT. Moreover, we can easily see that M is a test martingale for $\mathbb{P}$. In fact, even if the data is not iid, the likelihood ratio based on the first n data points,

$$M_n = \frac{d\mathbb{Q}}{d\mathbb{P}}(X_1, \dots, X_n), \quad (7.6)$$

is still a test martingale for $\mathbb{P}$ (which can be checked by conditioning). Revisiting the example in Section 1.6 with the e-variable $E_n = \exp(\mu S_n - n\mu^2/2)$ in (1.1), we can check that with $E_0 = 1$,

$$\mathbb{E}^\mathbb{P}[E_n | E_{n-1}] = E_{n-1} \mathbb{E}^\mathbb{P}[\exp(\mu X_n - \mu^2/2)] = E_{n-1} \quad \text{for } n \in \mathbb{N},$$

and hence E is not only an e-process, but also a test martingale. This is true for all likelihood ratio processes in (7.5), including the example in (1.5).

When testing a simple hypothesis $\mathbb{P}$, it is optimal to use a test martingale to construct e-processes: Any e-process for $\mathbb{P}$ can be upper bounded by its Snell envelope under $\mathbb{P}$ (which is a test supermartingale), which in turn can be dominated by the test martingale for $\mathbb{P}$ that is given by its Doob decomposition; we leave the details to the reader.

In the setting of testing the simple null hypothesis $\mathbb{P}$ against $\mathbb{Q} \ll \mathbb{P}$, the likelihood ratio M_t in (7.5) is log-optimal among all e-variables using the first t data points for any $t \in \mathbb{N}$. This claim follows from Theorem 3.20, by noting that M_t is precisely the likelihood ratio of $\mathbb{Q}$ to $\mathbb{P}$ based on the first t data points. But a stronger claim can be made about the optimality of the entire *process*, which we do next.

Theorem 7.11: The likelihood ratio process is log-optimal

Suppose that for testing a simple hypothesis $\mathbb{P}$ versus a simple hypothesis $\mathbb{Q}$, the likelihood ratio process M given by (7.6) exists for every n . Then, for any stopping time τ that is finite $\mathbb{Q}$ -almost surely, and any e-process M' for $\mathbb{P}$,

$$\mathbb{E}^{\mathbb{Q}}[\log M_{\tau}] \geq \mathbb{E}^{\mathbb{Q}}[\log M'_{\tau}]. \quad (7.7)$$

Proof.

Note that M'_{τ} is an e-variable because M' is an e-process, and M'_{τ} is measurable with respect to $\mathcal{F}_{\tau}$. Take any event $A \in \mathcal{F}_{\tau}$ with $\mathbb{P}(A) = 0$. Then,

$$\begin{aligned} \mathbb{Q}(A) &= \mathbb{E}^{\mathbb{Q}}[\mathbf{1}_A] = \mathbb{E}^{\mathbb{Q}}[\mathbf{1}_A \mathbf{1}_{\{\tau < \infty\}}] \\ &= \lim_{n \rightarrow \infty} \mathbb{E}^{\mathbb{Q}}[\mathbf{1}_A \mathbf{1}_{\{\tau \leq n\}}] = \lim_{n \rightarrow \infty} \mathbb{E}^{\mathbb{P}}[\mathbf{1}_A \mathbf{1}_{\{\tau \leq n\}} M_n] = 0, \end{aligned}$$

where the last step follows because $\mathbb{P}(A) = 0$ and M_n is finite. Thus, $\mathbb{Q} \ll \mathbb{P}$ on $\mathcal{F}_{\tau}$. Invoking Proposition 3.22 completes the proof.

For testing a simple null $\mathbb{P}$ against a composite alternative $\mathcal{Q}$, there are two options discussed in Section 3.7. Both the mixture likelihood ratio in (3.7) and the plug-in likelihood ratio in (3.8) are actual test martingales and hence e-processes.

For testing composite nulls, the situation is much more complicated, as nontrivial composite martingales may not exist while nontrivial e-processes may exist. We discuss this case in Section 7.7 and later.

7.5.2 Asymptotic log-optimality and the plug-in method

We now focus on the setting of observing iid data $X_1, X_2, \dots$ sequentially. In this subsection, we use $\mathbb{P}$ for the distribution of X_1 under the null, and $\mathbb{Q}$ for the distribution of X_1 under the alternative ($\mathbb{Q} \neq \mathbb{P}$). Recall that $\text{KL}(\mathbb{Q}, \mathbb{P})$ is the Kullback–Leibler divergence between $\mathbb{Q}$ and $\mathbb{P}$ defined in (3.4). As in some other parts of the book, we will slightly abuse the notation here by using $\mathbb{P}$ and $\mathbb{Q}$ also for their product measures; phrases like “testing $\mathbb{P}$ against $\mathbb{Q}$ ” and “ $\mathbb{Q}$ -almost surely” refer to the product measures, and it should be clear from the context.

Proposition 7.12

Consider testing $\mathbb{P}$ against $\mathbb{Q}$ in the iid data setting. The likelihood ratio process has an asymptotic growth rate equal to $\text{KL}(\mathbb{Q}, \mathbb{P}) > 0$, and hence it is consistent; any other e-process cannot have a larger asymptotic growth rate.

Proof.

Let M^* be the likelihood ratio process and M be another e-process, both for $\mathbb{P}$ against $\mathbb{Q}$. The first statement follows from the law of large numbers: as $t \rightarrow \infty$, $\mathbb{Q}$ -almost surely,

$$\frac{\log M_t^*}{t} = \frac{1}{t} \sum_{i=1}^t \log \left(\frac{d\mathbb{Q}}{d\mathbb{P}}(X_i) \right) \rightarrow \mathbb{E}^{\mathbb{Q}} \left[\log \frac{d\mathbb{Q}}{d\mathbb{P}} \right] = \text{KL}(\mathbb{Q}, \mathbb{P}).$$

To show the second statement, suppose that

$$\liminf_{t \rightarrow \infty} \frac{\log M_t}{t} > \text{KL}(\mathbb{Q}, \mathbb{P}) \quad \mathbb{Q}\text{-almost surely.}$$

Let $\varepsilon > 0$ and τ be the first time t such that $\log M_t - \log M_t^* > \varepsilon$, which is $\mathbb{Q}$ -almost surely finite by the assumed limits. It follows that $\mathbb{E}^{\mathbb{Q}}[\log M_\tau] > \mathbb{E}^{\mathbb{Q}}[\log M_\tau^*]$, which conflicts the conclusion of Theorem 7.11.

Since the likelihood ratio process has the largest asymptotic growth rate, we can define a notion of asymptotically log-optimality using the likelihood ratio process as a benchmark.

Definition 7.13

An e-process M is *asymptotically log-optimal* for $\mathbb{P}$ against $\mathcal{Q}$ if for every $\mathbb{Q} \in \mathcal{Q}$, we have

$$\lim_{t \rightarrow \infty} \frac{1}{t} (\log M_t - \log M_t^{\mathbb{Q}}) \geq 0 \quad \text{in } L^1\text{-convergence under } \mathbb{Q},$$

where $M^{\mathbb{Q}}$ is the likelihood ratio process of $\mathbb{Q}$ to $\mathbb{P}$.

The definition of asymptotic log-optimality does not require the iid assumption, although the iid setting is natural because in this case $\log M_t^{\mathbb{Q}}/t$ has a constant and positive L^1 -limit $\text{KL}(\mathbb{Q}, \mathbb{P})$.

Because $\text{KL}(\mathbb{Q}, \mathbb{P}) > 0$ for every $\mathbb{Q} \in \mathcal{Q}$, any asymptotically log-optimal e-process against $\mathcal{Q}$ is consistent against every $\mathbb{Q} \in \mathcal{Q}$. Moreover, a mixture of an asymptotically log-optimal e-process with any other e-process is also asymptotically log-optimal.

In the previous chapters, log-optimality is defined against single $\mathbb{Q}$ (the null can be composite), and that is because the log-optimal e-variables (in the sense of Definition 3.18) or e-processes (in the sense of (7.7)) are determined by $\mathbb{Q}$. For *asymptotic* log-optimality, we are able to find nontrivial e-processes that work against a composite alternative $\mathcal{Q}$, as demonstrated below. On the other hand, we present these methods for a simple null in order to define likelihood ratio processes as the benchmark.

To illustrate asymptotic log-optimality, we now analyze the problem of testing parametric models with likelihood ratios discussed in Section 1.6, but with an infinite sequence of iid data $X_1, X_2, \dots$ (we considered finite data in Section 1.6). Suppose that data follow a distribution parameterized by $\theta \in \Theta$ with true unknown value $\theta^\dagger$. We assume Θ is an Euclidean space. The null hypothesis $\mathbb{P}$ corresponds to $\theta = \theta_0$ and the alternative hypothesis is $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta_1\}$, where $\Theta_1 \subseteq \Theta$ does not include θ_0 . Let ℓ be the likelihood ratio function given by

$$\ell(x; \theta) = \frac{d\mathbb{Q}_\theta}{d\mathbb{P}}(x),$$

and let θ_i be measurable to $\sigma(X_j : j \in [i-1])$ for each $i \in \mathbb{N}$. As we see in Section 1.6, the process $M = (M_i)_{i \in \mathbb{N}_0}$ given by

$$M_0 = 1 \quad \text{and} \quad M_t = \prod_{i=1}^t \ell(X_i; \theta_i) \quad \text{for } t \in \mathbb{N} \quad (7.8)$$

is an e-process for $\mathbb{P}$. If $\theta_i = \theta$ for all $i \in \mathbb{N}$, then M is the likelihood ratio process, which is log-optimal against $\mathbb{Q}_\theta$ in the sense of Theorem 7.11.

For the problem of testing $\mathbb{P}$ against $\mathcal{Q}$ above, it may not be feasible to choose θ_i deterministically because the tester does not know the true value $\theta^\dagger$. Therefore, it is desirable to design methods that are adaptive, hoping for asymptotic log-optimality against the composite hypothesis $\mathcal{Q}$. This is indeed possible, as long as $\theta_i \rightarrow \theta$ under $\mathbb{Q}_\theta$ in a suitable sense that depends on the parametric model. In the next result, we give a sufficient condition based on L^2 convergence.

Proposition 7.14

For testing $\mathbb{P}$ against $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta_1\}$, suppose that for each x , $\theta \mapsto \log \ell(x; \theta)$ is concave, and it has a gradient denoted by $\eta(x; \theta)$. If for each $\theta \in \Theta_1$, $\mathbb{E}^{\mathbb{Q}_\theta} [\|\theta_i - \theta\|^2] \rightarrow 0$ as $i \rightarrow \infty$ and

$$\sup_{i \in \mathbb{N}} \mathbb{E}^{\mathbb{Q}_\theta} [\|\eta(X_i; \theta_i)\|^2] < \infty,$$

then M in (7.8) is asymptotically log-optimal against $\mathcal{Q}$.

Proof.

Take $\theta \in \Theta_1$, and let M^* be the likelihood ratio process for $\mathbb{P}$ against $\mathbb{Q}_\theta$. Using the assumed concavity of the log-likelihood function, we get

$$\begin{aligned} \frac{1}{t} (\log M_t - \log M_t^*) &= \frac{1}{t} \sum_{i=1}^t (\log \ell(X_i; \theta_i) - \log \ell(X_i; \theta)) \\ &\geq \frac{1}{t} \sum_{i=1}^t (\theta_i - \theta) \eta(X_i; \theta_i). \end{aligned}$$

By Cauchy-Schwarz, we have

$$(\mathbb{E}^{\mathbb{Q}_\theta} [(\theta_i - \theta) \eta(X_i; \theta_i)])^2 \leq \mathbb{E}^{\mathbb{Q}_\theta} [\|\theta_i - \theta\|^2] \mathbb{E}^{\mathbb{Q}_\theta} [\|\eta(X_i; \theta_i)\|^2] \rightarrow 0.$$

Using

$$\mathbb{E}^{\mathbb{Q}_\theta} \left[\left| \frac{1}{t} \sum_{i=1}^t (\theta_i - \theta) \eta(X_i; \theta_i) \right| \right] \leq \frac{1}{t} \sum_{i=1}^t \mathbb{E}^{\mathbb{Q}_\theta} [(\theta_i - \theta) \eta(X_i; \theta_i)],$$

and the fact that the average of any vanishing real sequence converges to 0, we conclude

$$\liminf_{t \rightarrow \infty} \mathbb{E}^{\mathbb{Q}_\theta} \left[\frac{\log(M_t/M_t^*)}{t} \right] \geq \lim_{t \rightarrow \infty} \mathbb{E}^{\mathbb{Q}_\theta} \left[\frac{1}{t} \sum_{i=1}^t (\theta_i - \theta) \eta(X_i; \theta_i) \right] = 0,$$

which yields the desired asymptotic log-optimality against $\mathcal{Q}$.

The concavity of the log-likelihood ratio function in Proposition 7.14 is satisfied by many distributions in the exponential family.

In the specific case of normal distributions with mean $\theta \in \mathbb{R}$ as in Example 1.5, we have $\log \ell(x; \theta) = x\theta - \theta^2/2$, which is concave in θ , and $\eta(x; \theta) = x - \theta$. In this case, the conditions in Proposition 7.14 hold as soon as $\theta_i \rightarrow \theta$ in L^2 -convergence under each $\mathbb{Q}_\theta$. Therefore, the following corollary is obtained.

Corollary 7.15

For testing $\mathbb{P} = \mathcal{N}(\theta_0, 1)$ against $\mathcal{Q} = \{\mathbb{Q}_\theta = \mathcal{N}(\theta, 1) : \theta \in \Theta_1\}$, if $\mathbb{E}^{\mathbb{Q}_\theta} [(\theta_i - \theta)^2] \rightarrow 0$ as $i \rightarrow \infty$ for each $\theta \in \Theta_1$, then M in (7.8) is asymptotically log-optimal against $\mathcal{Q}$.

Corollary 7.15 justifies some observations made in Figure 1.6. The condition $\mathbb{E}^{\mathbb{Q}_\theta} [(\theta_i - \theta)^2] \rightarrow 0$ holds for the MLE, that is, when θ_i is the sample mean of $X_1, \dots, X_{i-1}$ for the normal model, since $\mathbb{E}^{\mathbb{Q}_\theta} [(\theta_i - \theta)^2] = (i-1)^{-1}$ for $i \geq 2$. Similarly, it also holds for MAP with a normal prior.

Another sufficient condition for asymptotic log-optimality with consistent estimators is when both Θ_1 and the range of the data X_1 are compact regions and the derivative η of $\theta \mapsto \log \ell(x; \theta)$ is continuous (hence bounded). In this case, if $\theta_i \rightarrow \theta$, $\mathbb{Q}_\theta$ -almost surely, then we have asymptotic log-optimality. This follows by

noting two facts: first,

$$\frac{1}{t}(\log M_t - \log M_t^*) = \frac{1}{t} \sum_{i=1}^t (\theta_i - \theta) \eta(X_i; \tilde{\theta}_i)$$

for some $\tilde{\theta}_i$ between θ and θ_i by the mean-value theorem, and second, $\mathbb{E}^{\mathbb{Q}_\theta}[(\theta_i - \theta) \eta(X_i; \tilde{\theta}_i)]$ vanishes.

7.6 Testing by betting

There is a straightforward way to view test (super)martingales as wealth processes arising from games. The key idea is as follows:

In order to test a null hypothesis $\mathcal{P}$, we set up a game of chance, where the statistician (who plays the role of Skeptic from Section 1.5) bets on each observation before observing it. This game must satisfy two properties:

- (a) If the null is true, meaning $\mathbb{P} \in \mathcal{P}$, then no betting strategy should be able to systematically make money (that is, it may make money by chance, but cannot guarantee making money);
- (b) if the null is false, then there should exist “good” betting strategies that can make money (while one may never be certain of short term profit, there is a way to guarantee making money in the long run).

If such a game can be designed, then the wealth of the statistician in the game is directly a measure of evidence against the null: the more the wealth (or more accurately, the larger the ratio of final to initial wealth), the more evidence that the null is likely false.

Let us begin with the simplest example, as discussed in the previous section. We show how to interpret testing $\mathbb{P}$ as a game in Algorithm 7.16.

Algorithm 7.16 Testing by betting: simple null $\mathbb{P}$

- 1: Statistician’s initial wealth is $W_0 = 1$
- 2: **for** $t = 1, 2, \dots$ **do**
- 3: Statistician declares bet $S_t : \mathcal{X} \rightarrow [0, \infty]$ such that

$$\mathbb{E}^{\mathbb{P}}[S_t(X) \mid X_1, \dots, X_{t-1}] \leq 1$$

- 4: Observe X_t
 - 5: Statistician’s wealth is updated as $W_t = W_{t-1} \cdot S_t(X_t)$
-

There are three immediate points worthy of comment. First, in the above game, the constraint on S_t is simply that it is an e-variable for $\mathbb{P}$ conditioned on the past. This immediately implies that for any betting strategy, the wealth process W is a test supermartingale for $\mathbb{P}$.

Second, one may ask: what is the *optimal* bet? Assume for simplicity that the alternative $\mathbb{Q}$ is simple, and that $\mathbb{P} \ll \mathbb{Q}$. Then, the wealth can grow exponentially fast under $\mathbb{Q}$, and so it is natural to optimize the exponent. This means choosing S_t to optimize $\mathbb{E}^{\mathbb{Q}}[\log S_t(X) \mid X_1, \dots, X_{t-1}]$ subject to S_t being an e-variable for $\mathbb{P}$. The answer was already derived in Theorem 3.20: it is the likelihood ratio of $\mathbb{Q}$ to $\mathbb{P}$, conditioned on $X_1, \dots, X_{t-1}$ (this conditioning is irrelevant if $\mathbb{P}$ and $\mathbb{Q}$ are iid product distributions on $\mathcal{X}^\infty$, but it is relevant if they are not). Thus, the log-optimal wealth process in this game is given precisely by (7.5).

The last point is that if one wants to test a composite null $\mathcal{P}$, the only alteration to the protocol is to require that S_t is an e-variable for $\mathcal{P}$, meaning that the constraint in the above protocol must hold for all $\mathbb{P} \in \mathcal{P}$. Then, the wealth process W is a test supermartingale for $\mathcal{P}$.

The reader may now naturally wonder how e-processes can be interpreted in this testing by betting framework, as opposed to simply test supermartingales. The answer is that their game-theoretic interpretation is slightly more involved. The statistician does not play a single game against $\mathcal{P}$, but instead plays a separate game against each $\mathbb{P} \in \mathcal{P}$, and accepts its net wealth as the worst wealth amongst all these games. This is made explicit in Algorithm 7.17.

Algorithm 7.17 Testing by betting: composite null $\mathcal{P}$

- 1: **for** $\mathbb{P} \in \mathcal{P}$ **do**
- 2: Statistician's initial wealth in game $\mathbb{P}$ is $W_0^{\mathbb{P}} = 1$
- 3: **for** $t = 1, 2, \dots$ **do**
- 4: Statistician declares bet $S_t^{\mathbb{P}} : \mathcal{X} \rightarrow [0, \infty]$ such that

$$\mathbb{E}^{\mathbb{P}}[S_t^{\mathbb{P}}(X) \mid X_1, \dots, X_{t-1}] \leq 1$$

- 5: Observe X_t
 - 6: Statistician's wealth in $\mathbb{P}$ is updated as $W_t^{\mathbb{P}} = W_{t-1}^{\mathbb{P}} \cdot S_t^{\mathbb{P}}(X_t)$
 - 7: Statistician's overall wealth is updated as $W_t = \inf_{\mathbb{P} \in \mathcal{P}} W_t^{\mathbb{P}}$
-

The above interpretation stems directly from the second of the two equivalent definitions of e-processes. There are some difficulties in the above game: how do we instantiate and play (potentially uncountably) infinitely many games? How do we separately specify a betting strategy in each game? Further, there may be measure-theoretic subtleties associated with the $\inf_{\mathbb{P} \in \mathcal{P}}$ operation. We sidestep these issues here. For us, the game-theoretic interpretation of e-processes is just that: an interpretation. It helps us understand the relationship to test supermartingales, and indeed recognize why e-processes are a more general concept. However, we will never actually set up and play such a game explicitly in order to derive and use sensible e-processes in practice. Those can be derived more directly, as we shall see next.

7.7 The universal inference e-process

The previous sections have emphasized the fixed- n view, in that the split and subsampled likelihood ratio e-variables were not e-processes. We now note that there is a variant of universal inference that directly yields an e-process. Define $E_0 = 1$ and

$$E_t = \prod_{i=1}^t \frac{\hat{q}_{i-1}(X_i)}{\hat{p}_t(X_i)}, \quad (7.9)$$

where $\hat{q}_{i-1} \in \mathcal{Q}$ is allowed to depend on $X_1, \dots, X_{i-1}$ and $\hat{p}_t \in \mathcal{P}$ is the maximum likelihood estimator in $\mathcal{P}$ based on $X_1, \dots, X_t$. Note that the numerator can be updated in an online fashion, but all terms in the denominator need to be recalculated after observing each new data point.

Proposition 7.18

The process E in (7.9) is an e-process for $\mathcal{P}$.

The proof is simple and mimics the argument for the split likelihood ratio e-variable. By definition of $\hat{p}_n$ being the maximum likelihood estimator for $\mathcal{P}$, we see that for each $\mathbb{P}$ in $\mathcal{P}$,

$$E_t \leq \prod_{i=1}^t \frac{\hat{q}_{i-1}(X_i)}{p(X_i)},$$

where the right-hand side is a test martingale for $\mathcal{P}$, thus verifying the defining property of an e-process.

One may view this, and other variants of universal inference, as calculating the ratio

$$\frac{\text{out-of-sample alternative likelihood}}{\text{in-sample null likelihood}}.$$

To clarify, each $\hat{q}_{i-1}$ is evaluated on an independent data point X_i , which is outside the sample that was used to choose $\hat{q}_{i-1}$. In contrast, $\hat{p}_t$ is evaluated on $X_1, \dots, X_t$, which are all inside the sample that was used to choose it. Using the language of overfitting in machine learning, we overfit under the null by looking at its “training likelihood”, but we do not overfit under the alternative, by looking at its “test likelihood”. This

is why the threshold of $1/\alpha$ does not depend on $\mathcal{P}$ and $\mathcal{Q}$. This is contrast to generalized likelihood ratio methods, which take ratios of maximum likelihoods under alternative and null, but the threshold must then be adjusted (made larger) to take into account the difference in complexities of $\mathcal{P}$ and $\mathcal{Q}$.

We remark briefly again that instead of the plug-in method for handling composite $\mathcal{Q}$, we can use the mixture method instead. In particular, for any distribution ν over $\mathcal{Q}$,

$$E_t = \int_{\mathcal{Q}} \prod_{i=1}^t \frac{q(X_i)}{\hat{p}_t(X_i)} d\nu(q)$$

is an e-process. Without too much ambiguity, we refer to this also as the universal inference e-process.

7.8 Sequential e-values and empirically adaptive e-processes

In this section, we formally define sequential e-variables (e-values), and a specific form of e-processes built from these sequential e-variables. As in the previous section, we will use $\mathbb{T}$ is a finite or infinite set of indices starting from 0, and $\mathbb{T}_+$ is the set $\mathbb{T} \setminus \{0\}$.

Definition 7.19: Sequential e-values

The e-variables E_t with $t \in \mathbb{T}_+$ for $\mathcal{P}$, adapted to the filtration $\mathcal{F}$, are *sequential* if $\mathbb{E}^{\mathbb{P}}[E_t | \mathcal{F}_{t-1}] \leq 1$ for all $t \in \mathbb{T}_+$ and $\mathbb{P} \in \mathcal{P}$. In case the filtration $\mathcal{F}$ is not specified, by default we mean the natural filtration; in other words, the required condition is $\mathbb{E}^{\mathbb{P}}[E_t | E_1, \dots, E_{t-1}] \leq 1$ for all $t \in \mathbb{T}_+$ and $\mathbb{P} \in \mathcal{P}$.

A possible interpretation of sequential e-variables is that $E_1, E_2, \dots$ are obtained by laboratories 1, 2, $\dots$ in this order, and laboratory t makes sure that its result E_t is a valid e-variable given the previous results $E_1, \dots, E_{t-1}$. Note that if sequential e-variables E_t , $t \in \mathbb{T}_+$ for $\mathbb{P}$ are exact, then they must satisfy $\mathbb{E}^{\mathbb{P}}[E_t | \mathcal{F}_0] = 1$ for each $t \in \mathbb{T}_+$ because of the tower property of conditional expectation.

We can build e-processes directly based on sequential e-variables, unspecific to the underlying data or hypothesis. The first simple observation is that the product process of sequential e-values is always an e-process.

Proposition 7.20

Let $(E_t)_{t \in \mathbb{T}_+}$ be a sequence (finite or infinite) of sequential e-variables for $\mathcal{P}$. Define

$$M_t = \prod_{s=1}^t E_s \text{ for } t \in \mathbb{T}_+ \text{ and } M_0 = 1.$$

Then M is an e-process for $\mathcal{P}$.

The proof follows by noting that for $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}^{\mathbb{P}}[M_t | \mathcal{F}_{t-1}] = M_{t-1} \mathbb{E}^{\mathbb{P}}[E_t | \mathcal{F}_{t-1}] \leq M_{t-1},$$

and thus $(M_t)_{t \in \mathbb{T}}$ is a test supermartingale, and thus an e-process.

More generally, e-processes can be constructed from sequential e-variables through testing by betting discussed in Section 7.6. In Chapter 8, we will see that this is actually the only admissible way of constructing e-processes based only on the sequential e-variables. In what follows, the filtration $\mathcal{F}$ is the natural filtration of these e-variables.

Definition 7.21: Log-optimal and empirically adaptive e-processes

Let $(E_t)_{t \in \mathbb{T}_+}$ be a sequence (finite or infinite) of sequential e-variables.

- (i) An *e-process built on $(E_t)_{t \in \mathbb{T}_+}$* is $M = (M_t)_{t \in \mathbb{T}}$ defined by

$$M_t = \prod_{s=1}^t ((1 - \lambda_s) + \lambda_s E_s) \text{ for } t \in \mathbb{T}_+ \text{ and } M_0 = 1, \quad (7.10)$$

where for $t \in \mathbb{T}_+$, λ_t is any measurable function of $E_1, \dots, E_{t-1}$ (i.e., $(\lambda_t)_{t \in \mathbb{T}_+}$ is predictable with respect to $\mathcal{F}$) and takes value in $[0, 1]$.

- (ii) For an alternative probability measure $\mathbb{Q}$, the *$\mathbb{Q}$ -log-optimal e-process built on $(E_t)_{t \in \mathbb{T}_+}$* is the e-process in (7.10) such that λ_t for $t \in \mathbb{T}_+$ solves the following equation

$$\lambda_t \in \arg \max_{\lambda \in [0, 1]} \mathbb{E}^{\mathbb{Q}} [\log ((1 - \lambda) + \lambda E_t) \mid \mathcal{F}_{t-1}]. \quad (7.11)$$

- (iii) For a parameter $\gamma \in (0, 1]$, the *empirically adaptive e-process* is the e-process in (7.10) such that λ_t for $t \geq 2$ solves the following equation

$$\lambda_t \in \arg \max_{\lambda \in [0, \gamma]} \frac{1}{t-1} \sum_{s=1}^{t-1} \log ((1 - \lambda) + \lambda E_s),$$

and $\lambda_1 = 0$.

Note that similarly to the product process in Proposition 7.20, the construction (7.10) guarantees that M_t is a supermartingale.

To interpret the $\mathbb{Q}$ -log-optimal e-process built on $(E_t)_{t \in \mathbb{T}}$, at every step t , λ_t is chosen to maximize the e-power in Definition 3.11, which is allowed to use all past observed e-values. It is not necessarily log-optimal among all e-processes testing the alternative $\mathbb{Q}$; it is only log-optimal within the class in (7.10). If E_t is independent of $\mathcal{F}_{t-1}$ under $\mathbb{Q}$, then λ_t in (7.11) is given by maximizing the unconditional expectation,

$$\lambda_t \in \arg \max_{\lambda \in [0, 1]} \mathbb{E}^{\mathbb{Q}} [\log ((1 - \lambda) + \lambda E_t)],$$

which is deterministic.

The empirically adaptive martingale does not have a specific alternative, and at every step t , λ_t is chosen to maximize the e-power with respect to the empirical distribution of $E_1, \dots, E_{t-1}$; this also explains the choice of the name. The parameter γ controls an upper bound for λ_t , and an uninformative default choice can be $\gamma = 1/2$.

In the construction of the empirically adaptive e-process M , for $t \geq 2$, by Theorem 3.14, $\lambda_t = 0$ if and only if the empirical mean of $E_1, \dots, E_{t-1}$ is less than or equal to 1, that is, $\sum_{s=1}^{t-1} E_s \leq t$. It is straightforward to verify that M is a martingale with respect to the natural filtration of the e-variables if these e-variables are exact; otherwise it is a supermartingale. In particular, M_τ is an e-variable for any stopping time τ . The main advantage of the empirically adaptive martingale is the simple observation that if the e-variables are iid under the alternative hypothesis, then it will have good e-power against the null hypothesis. It generally has good e-power if the iid assumption holds approximately, and it always has a valid type-I error control at level α for the standard threshold of rejection $1/\alpha$. The following result formalizes the claim about good e-power in an asymptotic setting.

Theorem 7.22

Let $(E_t)_{t \in \mathbb{N}}$ be an infinite sequence of e-variables that is iid under the alternative distribution $\mathbb{Q}$ such that $\mathbb{E}^{\mathbb{Q}}[\log E_1]$ is finite. The empirically adaptive e-process M with $\gamma = 1$ satisfies the following:

(i) Asymptotic log-optimality of M holds in the following sense:

$$\lim_{t \rightarrow \infty} \frac{1}{t} (\log M_t - \log M'_t) \geq 0 \quad \text{in } L^1\text{-convergence under } \mathbb{Q}$$

for the $\mathbb{Q}$ -log-optimal e-process M' built on $(E_t)_{t \in \mathbb{N}}$.

(ii) Consistency of M holds: if $\mathbb{E}^{\mathbb{Q}}[E_1] > 1$, then $M_t \rightarrow \infty$ $\mathbb{Q}$ -almost surely as $t \rightarrow \infty$.

Theorem 7.22 (i) shows that M has an asymptotic growth rate that is equal to the $\mathbb{Q}$ -log-optimal e-process, which is the largest among all choices of e-processes built on $(E_t)_{t \in \mathbb{N}}$.

We omit a formal proof of Theorem 7.22, but its intuition is very simple. Part (i) follows essentially by proving that asymptotically, by the iid assumption, λ_t in the empirically adaptive martingale converges in probability to the optimal value λ^* that maximizes (over λ) the e-power of $1 - \lambda + \lambda E_1$. The e-power of $1 - \lambda^* + \lambda^* E_1$ is the asymptotic growth rate of M'_t by the law of large numbers, so the two e-processes share the same asymptotic growth rate. Part (ii) follows by using Theorem 3.14 to show that the optimal growth rate is positive under the assumption $\mathbb{E}^{\mathbb{Q}}[E_1] > 1$, and hence M_t , which has the optimal growth rate asymptotically, grows to infinity as $t \rightarrow \infty$.

It should be clear that both statements in Theorem 7.22 also holds for the $\mathbb{Q}$ -log-optimal e-process. The advantage of the empirically adaptive e-process is that one does not need to know $\mathbb{Q}$.

7.9 Obtaining an e-process from e-values using time-mixture

While the numeraire has been presented as a method for obtaining a single e-variable, one can convert a sequence of numeraires (calculated at different sample sizes n) into an e-process using the following simple black-box construction.

For all $n \geq 1$, let $E^{(n)}$ be an e-variable for $\mathcal{P}$ based on the first n data points $X_1, \dots, X_n$. Take any probability mass function w on the natural numbers $\mathbb{N} = \{1, 2, \dots\}$ and define

$$M_n = \sum_{j \leq n} w(j) E^{(j)}. \quad (7.12)$$

with $M_0 = 1$ by default. Note that M is an increasing process, which is unusual for e-processes.

Proposition 7.23

The time-mixed process M defined in (7.12) is an e-process for $\mathcal{P}$.

Proof.

For any $\mathbb{P} \in \mathcal{P}$ and any random time τ (possibly, but not necessarily, a stopping time),

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[M_\tau] &= \mathbb{E}^{\mathbb{P}} \left[\sum_{n=1}^{\infty} \mathbb{1}_{\{\tau=n\}} \sum_{j \leq n} w(j) E^{(j)} \right] \\ &= \mathbb{E}^{\mathbb{P}} \left[\sum_{j \geq 1} w(j) E^{(j)} \sum_{n=j}^{\infty} \mathbb{1}_{\{\tau=n\}} \right]. \end{aligned}$$

Since $\sum_{n=j}^{\infty} \mathbb{1}_{\{\tau=n\}} \leq 1$ and each $E^{(j)}$ is an e-variable, we get

$$\mathbb{E}^{\mathbb{P}}[M_{\tau}] \leq \sum_{j \geq 1} w(j) \mathbb{E}^{\mathbb{P}}[E^{(j)}] \leq 1,$$

concluding the proof.

Referring again to the motivation at the start of this section, let $E^{(n)}$ be the numeraire for $\mathcal{P}$ against some $\mathbb{Q}$ based on $X_1, \dots, X_n$ (more precisely, that means testing $\mathcal{P}$ against $\mathbb{Q}$ conditioned on the σ -algebra generated by $X_1, \dots, X_n$). In general, $E^{(1)}, E^{(2)}, \dots$ is not an e-process. Now, choose

$$w(n) = \frac{c}{n(\log n)^2},$$

for some normalizing constant $c > 0$ that ensures $\sum_{n \in \mathbb{N}} w(n) = 1$. we get that

$$\log M_n \geq \log E^{(n)} - \log n - 2 \log \log n - O(1).$$

Thus,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M_n - \frac{1}{n} \log E^{(n)} = 0$$

meaning that the asymptotic growth rate of our e-process M matches that of the sequence of numeraire e-variables.

7.10 E-processes avoid reasoning about hypothetical worlds

We end the chapter with a somewhat bizarre example involving a very subtle (mis)use of p-values with optional stopping, whose solution is provided by employing e-processes.

Example 7.24: Quantum-like interference in ordinary experiments

Consider a data scientist, Bob, at an IT company. Bob runs an A/B test, which is a two-sample test intended to compare two options A and B of commercial design. Bob fully intends to collect 10,000 observations from the very start. Each observation involves a person choosing between options A and B. The data can thus be seen as a sequence of heads and tails, and let us assume that the null hypothesis is that the coin is fair. At the end of the experiment, Bob finds that his p-value is 0.04, and he is satisfied with that decision, and rejects the null. Later that day, he reports these results to his boss, Alice. Alice is impressed with Bob's statistical rigor, as well as patience. Alice asks Bob how he managed to control his curiosity to look at the data before the experiment was completed. Bob replied that he indeed looked at the data once, after 5000 points were collected, but he simply let the experiment go on. Alice proceeds to ask Bob: "If you had seen over 4800 tails in the first 5000 tosses, what would you have done?". Bob replies that he would have of course stopped the experiment. Alice is disappointed, and tells Bob that his reported p-value of 0.04 is incorrect. Bob protests and asks why this thought experiment in a hypothetical world affects the interpretation of the data what actually happened in the real world. Alice responds that Bob's answer demonstrates that he did not actually have a fixed sample size: he had a data-dependent decision rule, because the sample size across different (hypothetical) runs of the experiment is sometimes 5000 and sometimes 10,000. This changes the p-value calculation, which assumed a fixed sample size of 10,000. A dejected Bob returns to his office to attempt a recalculation of the p-value based on an examination of when he would have stopped in these hypothetical worlds (or to re-run the experiment again, and not peek at the data).

Similarly, consider an alternate world (in the same setting as above) in which Bob fully intends to collect 5000 samples. He does not peek at the data, and after collecting 5000 samples, he calculates a p-value of 0.04. Alice is once more pleased with Bob's patience and rigor, but she asks him "Would you have collected more data had the p-value been 0.055?". If Bob says yes, then Alice immediately says that Bob's actual realized p-value of 0.04 is incorrect. Once more, Bob's answer to a question about a hypothetical world changes the validity of the p-value in the real world.

While the numbers used in the above example were intentionally extreme in order to make the responses seem believable, what the above example really demonstrates is that reasoning about the validity of p-values requires reasoning about (all possible) "hypothetical worlds". Bob's answers to what-if questions in hypothetical worlds affects the validity of his p-value in the real world that did occur. This is clearly not a desirable property, since it is practically infeasible, and indeed one that is often ignored in theory and practice. Despite the small effect on the final p-value that this reasoning may have had in the current experiment (due to the extreme numbers), it points to a fundamental issue that could have a larger effect in other settings, where it may also be more difficult to address.

The effect described above is quantum-like in the sense that merely observing the experiment once before its completion changes its properties. The mere *possibility* of interference in the experiment seems to have actually interfered in its validity.

If the final evidence reported was $E_{10,000}$, where $(E_n)_{n \in \mathbb{N}}$ is an e-process, then this issue would have been avoided. Indeed, because 5000 and 10,000 are both valid stopping times, it does not matter that in some other hypothetical world, Bob might have stopped at time 5000. Whenever Bob stops in any world, the reported e-value is a valid e-value. Moreover, it is fine to continue sampling if $E_{10,000}$ is not satisfactory to Bob, or to a later researcher. (Note that we really need $(E_n)_{n \in \mathbb{N}}$ to be an e-process and not just a sequence of e-values.)

7.11 A pathological example for the e-power of e-processes

The e-power has been a central criterion in the considerations of e-processes so far. We close this chapter with two examples, showing that the e-power has some limitations, similar to the ones discussed in Example 3.16, but in a sequential setting. The takeaway message from these examples is

- (a) it is possible to find e-processes with positive infinite e-power at each fixed time but declines to 0 with

probability 1 under $\mathbb{Q}$;

- (b) it is possible to find e-processes with negative infinite e-power at each fixed time but increases to ∞ with probability 1 under $\mathbb{Q}$.

Therefore, extra caution should be taken when infinity is involved in the calculation of e-power.

Example 7.25

Let $E_1, E_2, \dots$, be e-variables for $\mathbb{P}$ given by

$$E_t = \exp(t^2 - Y_t), \quad t \in \mathbb{N},$$

where $Y_1, Y_2, \dots$, are iid Pareto(1) random variables under $\mathbb{Q}$ (that is, $\mathbb{Q}(Y_1 > x) = 1/x$ for $x \geq 1$), and they are independent under $\mathbb{P}$ such that $\mathbb{E}^{\mathbb{P}}[E_t] \leq 1$ for each $t \in \mathbb{N}$. Let $M = (M_t)_{t \in \mathbb{N}}$ be given by $M_t = \prod_{s=1}^t E_s$, which is an e-process. Simple calculation shows, for any $x > 0$,

$$\begin{aligned} \mathbb{Q}(M_t \leq x) &= \mathbb{Q}\left(\sum_{s=1}^t Y_s \geq \frac{t(t+1)(2t+1)}{6} - \log x\right) \\ &\leq t \mathbb{Q}\left(Y_1 \geq \frac{(t+1)(2t+1)}{6} - \frac{\log x}{t}\right) \rightarrow 0 \quad \text{as } t \rightarrow \infty. \end{aligned}$$

Therefore, $M_t \rightarrow \infty$ in probability under $\mathbb{Q}$, and the probability of rejecting the null by thresholding M_t approaches 1. However, because $\mathbb{E}^{\mathbb{Q}}[Y_1] = \infty$, we have $\mathbb{E}^{\mathbb{Q}}[\log E_t] = \mathbb{E}^{\mathbb{Q}}[\log M_t] = -\infty$ for each $t \in \mathbb{N}$, and the e-process M has the same e-power $-\infty$ as the zero process.

Example 7.26

Suppose that $E_1, E_2, \dots$ are independent e-variables for $\mathbb{P}$. If E_1 satisfies $\mathbb{E}^{\mathbb{Q}}[\log E_1] = -\infty$ like in (a) above and E_t has finite e-power for all $t \geq 2$, then the process M given by $M_t = \prod_{s=1}^t E_s$, satisfies $\mathbb{E}^{\mathbb{Q}}[\log M_t] = -\infty$ for all $t \in \mathbb{N}$, no matter how fast it grows after $t = 1$; in particular, one can easily build M with $M_t \rightarrow \infty$ in probability under $\mathbb{Q}$.

In either example above, we find an e-process that is intuitively powerful but has negative infinite e-power. Taking a reciprocal in Examples 7.25–7.26, we also find e-processes with positive infinite e-power but declines to 0 with probability 1 under $\mathbb{Q}$.

Bibliographical note

The sequential probability ratio test described in Section 7.2 was introduced in Wald [1945, 1947], with its optimality proved in Wald and Wolfowitz [1948]. The log-optimality of the likelihood ratio process (Theorem 7.11) appears in a more restricted form in Koolen and Grünwald [2021]; our result is more general.

Without the term e-process being used, the second definition of an e-process was employed in Howard et al. [2020], while the first was later proposed in Grünwald et al. [2024a]. Despite having these new definitions, both papers still primarily focused on test supermartingales obtained as products of e-variables. The equivalence of the two definitions of e-processes was proved in Ramdas et al. [2020] using Snell envelopes and Doob decompositions. Ramdas et al. [2022] showed the fundamental difference between e-processes and test supermartingales in their study of testing exchangeability. Certain key geometric concepts like fork-convexity play a key role: test supermartingales for $\mathcal{P}$ are also test supermartingales for the closed fork-convex hull of $\mathcal{P}$ but this is not true for e-processes.

The universal inference e-process appears in Wasserman et al. [2020, Chapter 8], but it is a well known statistic in sequential inference. For example it is called the adaptive likelihood ratio in Tartakovsky et al. [2014, §5.4.2], but they do not appear to define or study e-processes as a more general concept beyond this

example. [Dixit and Martin \[2024\]](#) proved that when using a particular nonparametric mixture technique (predictive recursion), universal inference is asymptotically log-optimal.

The modern framework of testing by betting appears in [Shafer and Vovk \[2019\]](#), but its roots can be traced all the way back to [Ville \[1939\]](#). The optional stopping theorem was proved by [Doob \[1953\]](#). A modern self-contained proof of Ville’s inequality [[Ville, 1939](#)] can be found in [Howard et al. \[2020\]](#). Results on the tightness of Ville’s inequality can be found in [Ramdas et al. \[2020\]](#) and [[Howard et al., 2021](#), Proposition 4].

The predictable process $(\lambda_t)_{t \in \mathbb{T}_+}$ in the empirically adaptive e-process is also called the plug-in betting strategy, or the optimal strategy for the growth rate adaptive to the particular alternative (GRAPA), which was studied by [Waudby-Smith and Ramdas \[2024\]](#). Sequential e-variables were first defined in [Vovk and Wang \[2021\]](#) with respect to their natural filtration. Theorem 7.22 is based on [Wang et al. \[2025\]](#) and its proof can be found in [Wang et al. \[2025, Theorem 3\]](#). Example 7.25 was discussed in [Wang \[2024a\]](#).

Chapter 8

Handling multiple e-values

In this chapter, we discuss several methods handling multiple e-values. Throughout the chapter, let $E_1, \dots, E_K$ be K e-variables, where $K \geq 2$ is a fixed positive integer. We write $\mathcal{K} = [K]$, which will be used in this and the subsequent chapters, wherever we deal with K e-variables. The presence of multiple e-values arises in multiple testing as in Chapter 9 but it also can arise in single testing with e-values computed from different parts of the data or data splitting as in Chapters 5 and 7. Here, we do not distinguish how they are computed, but focus on handling e-variables for some common hypothesis.

8.1 Merging e-values under arbitrary dependence

An important advantage of e-values over p-values is that they are easy to combine. This is the topic of this section, in which we consider the general case, without any assumptions on the joint distribution of the input e-variables. Thus, our aim is to design methods to combine arbitrarily dependent e-values. The cases of independent e-variables or those with some specific dependence structures are considered in the next sections. In what follows, inequalities between functions are understood to hold everywhere.

Definition 8.1: E-merging functions

- (i) An *e-merging function* (of K e-values) is an increasing Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$ such that for any hypothesis, $F(E_1, \dots, E_K)$ is an e-variable for any e-variables $E_1, \dots, E_K$.
- (ii) An e-merging function F is *symmetric* if $F(\mathbf{e})$ is invariant under any permutation of $\mathbf{e}$.
- (iii) An e-merging function F *dominates* an e-merging function G if $F \geq G$. The domination is *strict* if $F \geq G$ and $F(\mathbf{e}) > G(\mathbf{e})$ for some $\mathbf{e} \in [0, \infty)^K$. An e-merging function is *admissible* if it is not strictly dominated by any e-merging function.
- (iv) An e-merging function F *essentially dominates* an e-merging function G if, for all $\mathbf{e} \in [0, \infty)^K$,

$$G(\mathbf{e}) > 1 \implies F(\mathbf{e}) \geq G(\mathbf{e}).$$

We will treat K as fixed and will omit the dimension for an e-merging function, which should be clear from the context. Similarly to the situation of Section 2.3, the defining property of F in Definition 8.1 (i) is required to hold for any hypothesis, but it suffices for F to satisfy the property for one atomless probability measure $\mathbb{P}$ on some $(\Omega, \mathcal{F})$; this is formally justified in Appendix A.1.

Although an e-variable may take the value ∞ , this occurs with 0 probability under the null hypothesis. Hence, it is safe to set the value of F to ∞ if any of its input is ∞ . Therefore, without loss of generality, all e-merging functions in this text are defined on $[0, \infty)^K$, and this spares us from specifying terms such as $0 \times \infty$ in our analysis.

The increasing monotonicity assumed for e-merging functions reflects the natural assumption that larger

individual e-values represent stronger evidence against the null hypothesis, and thus producing a larger e-value.

Remark 8.2. An increasing function on $\mathbb{R}^K$ for $K \geq 2$ is not automatically Borel, although it is when $K = 1$. An example is given in the discussion after Theorem 4.4 of [Graham and Grimmett \[2006\]](#).

The interpretation of domination is self-evident: We would hope an e-merging function to be large, providing useful e-values. The notion of essential domination weakens that of domination in a natural way: We require that F is not worse than G only in cases where G carries evidence. A fact about admissibility is that any e-merging function is dominated by an admissible e-merging function. For this reason, we are only interested in admissible e-merging functions.

An important example of e-merging function is the *arithmetic mean* $\mathbb{M}_K$, defined by

$$\mathbb{M}_K(e_1, \dots, e_K) = \frac{e_1 + \dots + e_K}{K}, \quad e_1, \dots, e_K \in [0, \infty).$$

Another example is the weighted arithmetic mean. Let Δ_n be the standard simplex in $\mathbb{R}^n$ for $n \in \mathbb{N}$, that is,

$$\Delta_n = \left\{ (x_1, \dots, x_n) \in [0, 1]^n : \sum_{k=1}^n x_k = 1 \right\}.$$

It is straightforward to verify that for $\lambda = (\lambda_1, \dots, \lambda_{K+1}) \in \Delta_{K+1}$, the function

$$\mathbb{M}_\lambda : (e_1, \dots, e_K) \mapsto \sum_{k=1}^K \lambda_k e_k + \lambda_{K+1}, \quad (8.1)$$

is an e-merging function. We will refer to $\mathbb{M}_\lambda$ as the λ -weighted arithmetic mean, although strictly speaking it is a weighted average of the input vector $(e_1, \dots, e_K)$ and the constant 1. On the other hand, the geometric mean is also an e-merging function, but it is dominated by the arithmetic average.

We present two main results in this section. First, the arithmetic mean essentially dominates all other symmetric e-merging functions. Second, all admissible e-merging functions take the form of $\mathbb{M}_\lambda$.

Proposition 8.3

The arithmetic mean $\mathbb{M}_K$ essentially dominates any symmetric e-merging function.

Proof.

Fix an atomless probability measure $\mathbb{P}$. Let F be a symmetric e-merging function. Suppose for the purpose of contradiction that there exists $(e_1, \dots, e_K) \in [0, \infty)^K$ such that

$$b := F(e_1, \dots, e_K) > \max\left(\frac{e_1 + \dots + e_K}{K}, 1\right) =: a. \quad (8.2)$$

Let $\mathcal{S}_K$ be the set of all permutations of $\mathcal{K}$, π be randomly and uniformly drawn from $\mathcal{S}_K$, and $(D_1, \dots, D_K) := (e_{\pi(1)}, \dots, e_{\pi(K)})$. Further, let $(D'_1, \dots, D'_K) := (D_1, \dots, D_K)\mathbf{1}_A$, where A is an event independent of π and satisfying $\mathbb{P}(A) = 1/a$ (the existence of such random π and A is guaranteed for any atomless probability space by Lemma A.1 in the Appendix).

For each k , since D_k takes the values $e_1, \dots, e_K$ with equal probability, we have $\mathbb{E}^\mathbb{P}[D_k] = (e_1 + \dots + e_K)/K$, which implies $\mathbb{E}^\mathbb{P}[D'_k] = (e_1 + \dots + e_K)/(Ka) \leq 1$. Together with the fact that D'_k is nonnegative, we know D'_k is an e-variable for $\mathbb{P}$. Moreover, by symmetry,

$$\begin{aligned} \mathbb{E}^\mathbb{P}[F(D'_1, \dots, D'_K)] &= \mathbb{P}(A)F(e_1, \dots, e_K) + (1 - \mathbb{P}(A))F(0, \dots, 0) \\ &\geq b/a > 1, \end{aligned}$$

a contradiction. Therefore, we conclude that there is no $(e_1, \dots, e_K)$ such that (8.2) holds.

In particular, Proposition 8.3 implies that if F is an e-merging function that is symmetric and positively homogeneous (i.e., $F(\lambda \mathbf{e}) = \lambda F(\mathbf{e})$ for all $\lambda > 0$; this is satisfied by e.g., $\mathbb{M}_K$, the geometric mean, the harmonic average, or the minimum), then F is dominated by $\mathbb{M}_K$, because for any $\mathbf{e} \in [0, \infty)^K$ with $F(\mathbf{e}) > 0$, we have $F(\lambda \mathbf{e}) \leq \mathbb{M}_K(\lambda \mathbf{e})$ for λ large enough, leading to $F(\mathbf{e}) \leq \mathbb{M}_K(\mathbf{e})$.

It is clear that the arithmetic mean $\mathbb{M}_K$ does not dominate every symmetric e-merging function; for example, the convex mixtures of the trivial e-merging function 1 and $\mathbb{M}_K$, that is, $\lambda + (1 - \lambda)\mathbb{M}_K$ for $\lambda \in [0, 1]$, are pairwise non-comparable with respect to the relation of domination.

In the theorem below, we show that each function $\mathbb{M}_\lambda$ in (8.1) is admissible and that the class (8.1) is precisely the class of all admissible e-merging functions. Moreover, every e-merging function is dominated by one of (8.1).

Theorem 8.4

For a function $F : \mathbb{R}_+^K \rightarrow \mathbb{R}_+$,

- (i) if F is an e-merging function, then $F \leq \mathbb{M}_\lambda$ for some $\lambda \in \Delta_{K+1}$;
- (ii) F is an admissible e-merging function if and only if $F = \mathbb{M}_\lambda$ for some $\lambda \in \Delta_{K+1}$.

The proof of Theorem 8.4 is put in Chapter 11, as it requires several other advanced technical results.

8.2 Independent e-values and their products

In this section we consider merging functions for independent e-values.

Definition 8.5: Ie-merging functions

- (i) An *ie-merging function* is a Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$ such that for any hypothesis, $F(E_1, \dots, E_K)$ is an e-variable for any independent e-variables $E_1, \dots, E_K$.
- (ii) An ie-merging function F *weakly dominates* an ie-merging function G if, for all $e_1, \dots, e_K$,

$$(e_1, \dots, e_K) \in [1, \infty)^K \implies F(e_1, \dots, e_K) \geq G(e_1, \dots, e_K).$$

The corresponding definitions of domination, strict domination, and admissibility are obtained from Definition 8.1 by replacing “e-merging” with “ie-merging”; this also applies to the later definition of se-merging functions.

One subtle distinction of the above definition from e-merging functions in the previous section is that we do not require increasing monotonicity in all arguments for the ie-merging function. This relaxation is useful in the broader class of se-merging functions, which will be explained in Section 8.3.

Similarly to the arithmetic mean for arbitrarily dependent e-values, there is an important ie-merging function: the *product* Π_K , defined by

$$\Pi_K(e_1, \dots, e_K) = \prod_{k=1}^K e_k, \quad e_1, \dots, e_K \in [0, \infty). \quad (8.3)$$

For weak dominance, we require that F is not worse than G if all input e-values carry evidence. This requirement is very weak because, especially for a large K , it is rarely the case to have all input e-values larger than 1.

The following proposition gives a simple condition for admissibility.

Proposition 8.6

Fix an atomless probability measure $\mathbb{P}$. For an increasing Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$, if $\mathbb{E}\mathbb{P}[F(\mathbf{E})] = 1$ for all vectors $\mathbf{E}$ of exact e-variables (resp., for all vectors $\mathbf{E}$ of independent exact e-variables), then F is an admissible e-merging function (resp., an admissible ie-merging function). For the statement on the ie-merging function, increasing monotonicity of F is not needed.

Proof.

It is obvious that F is an e-merging function (resp., ie-merging function). Next we show that F is admissible. Suppose for the purpose of contradiction that there exists an ie-merging function G such that $G \geq F$ and $G(e_1, \dots, e_K) > F(e_1, \dots, e_K)$ for some $(e_1, \dots, e_K) \in [0, \infty)^K$. Take a vector $(E_1, \dots, E_K)$ of independent exact e-variables such that $\mathbb{P}((E_1, \dots, E_K) = (e_1, \dots, e_K)) > 0$. Such a random vector is easy to construct by considering any distribution with a positive mass on each of $e_1, \dots, e_K$. Then we have

$$\mathbb{P}(G(E_1, \dots, E_K) > F(E_1, \dots, E_K)) > 0,$$

which implies

$$\mathbb{E}\mathbb{P}[G(E_1, \dots, E_K)] > \mathbb{E}\mathbb{P}[F(E_1, \dots, E_K)] = 1,$$

contradicting the assumption that G is an ie-merging function. Therefore, no ie-merging function strictly dominates F . Noting that an e-merging function is also an ie-merging function, admissibility of F is guaranteed under both settings.

If $E_1, \dots, E_K$ are independent e-variables, their product $E_1 \dots E_K$ will also be an e-variable. This is the analogue of Fisher's method for p-values, according to the rough relation $e \sim 1/p$ mentioned in Section 2.3. Fisher's method will be discussed in Chapter 11. The ie-merging function Π_K is admissible by Proposition 8.6. More generally, we can see that the U-statistic functions U_n , defined by

$$U_n(e_1, \dots, e_K) = \frac{1}{\binom{K}{n}} \sum_{A \subseteq \mathcal{K}: |A|=n} \left(\prod_{k \in A} e_k \right), \quad n \in \{0, 1, \dots, K\}, \quad (8.4)$$

and their convex mixtures are ie-merging functions. Notice that this class includes product (for $n = K$), arithmetic average $\mathbb{M}_K$ (for $n = 1$), and constant 1 (for $n = 0$). Proposition 8.6 implies that the U-statistic functions in (8.4) and their convex combinations are admissible ie-merging functions.

Let us now establish a very weak counterpart of Proposition 8.3 for independent e-values; on the positive side it will not require the assumption of symmetry needed for Proposition 8.3.

Proposition 8.7

The product ie-merging function Π_K weakly dominates any ie-merging function.

Proof.

Indeed, suppose that there exists $(e_1, \dots, e_K) \in [1, \infty)^K$ such that

$$F(e_1, \dots, e_K) > \Pi_K(e_1, \dots, e_K) = e_1 \cdots e_K.$$

Fix an atomless $\mathbb{P} \in \Pi$. Let $E_1, \dots, E_K$ be independent random variables such that each E_k for $k \in \mathcal{K}$ takes values in the two-element set $\{0, e_k\}$ and $E_k = e_k$ with probability $1/e_k$. Then

each E_k is an e-variable but

$$\begin{aligned}\mathbb{E}^\mathbb{P}[F(E_1, \dots, E_K)] &\geq F(e_1, \dots, e_K) \mathbb{P}(E_1 = e_1, \dots, E_K = e_K) \\ &> e_1 \cdots e_K (1/e_1) \cdots (1/e_K) = 1,\end{aligned}$$

which contradicts F being an ie-merging function.

A natural question is whether the convex mixtures of (8.4) form a complete class of admissible ie-merging functions. They do not, as shown by the following example.

Example 8.8

Define a function $f : [0, \infty)^2 \rightarrow [0, \infty)$ by

$$f(e_1, e_2) = \frac{1}{2} \left(\frac{e_1}{1 + e_1} + \frac{e_2}{1 + e_2} \right) (1 + e_1 e_2).$$

Using Proposition 8.6, one can check that f is an admissible ie-merging function. Nevertheless, it is different from any convex mixture of (8.4).

The assumption of the independence of e-variables $E_1, \dots, E_K$ is not necessary for the product $E_1 \cdots E_K$ to be an e-variable. For instance, if the e-variables $E_1, \dots, E_K$ are *negative lower orthant dependent*, meaning that for all $(x_1, \dots, x_K) \in \mathbb{R}^K$,

$$\mathbb{P} \left(\bigcap_{k \in \mathcal{K}} \{E_k \leq x_k\} \right) \leq \prod_{k \in \mathcal{K}} \mathbb{P}(E_k \leq x_k), \quad (8.5)$$

then the product Π_K and any U-statistic functions in (8.4) of these e-variables are also e-variables. Clearly, if the central inequality in (8.5) is replaced by an equality, then we get independence. In the next section, we consider a particularly useful class of e-variables for which the product and the U-statistic functions are valid merging functions.

8.3 Sequential e-values and martingale merging functions

The notion of sequential e-variables generalizes that of independent e-variables, and it appears in several natural contexts that we will explain later. For conditional expectations, since they are defined in an almost-sure sense, we omit “almost surely” in their statements.

Definition 8.9: Sequential e-values and se-merging functions

- (i) The e-variables $E_1, \dots, E_K$ for $\mathcal{P}$ are *sequential* if

$$\mathbb{E}^\mathbb{P}[E_k \mid E_1, \dots, E_{k-1}] \leq 1$$

for all $k \in \mathcal{K}$ and $\mathbb{P} \in \mathcal{P}$.

- (ii) An *se-merging function* is a Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$ such that for any hypothesis, $F(E_1, \dots, E_K)$ is an e-variable for any sequential e-variables $E_1, \dots, E_K$.
- (iii) An se-merging function F is *exact* if $F(E_1, \dots, E_K)$ is an exact e-variable for all sequential exact e-variables $E_1, \dots, E_K$.

Sequential e-variables can also be defined with a general filtration $(\mathcal{F}_t)_{t \in \{0,1,\dots,K\}}$ instead of the one generated by $E_1, \dots, E_K$; this is treated in Chapter 7.

It is straightforward to check that all convex mixtures of (8.4) are se-merging functions. Since independent e-variables are sequential, se-merging functions also produce a valid e-variable for independent e-variables. Example 8.8 gives an ie-merging function but not an se-merging function.

Among se-merging functions, the convex mixtures of (8.4) are admissible. They are also admissible in the class of ie-merging functions by Proposition 8.6. We also note that it suffices for $E_1, \dots, E_K$ to be sequential in any one of the $K!$ possible orders for these merging methods to be valid.

A particular class of se-merging functions, larger than (8.4), is defined below.

Definition 8.10: Martingale merging functions

A function $F : [0, \infty)^K \rightarrow [0, \infty)$ is a *martingale merging function* if

$$F(e_1, \dots, e_K) = \prod_{k=1}^K (1 - \lambda_k + \lambda_k e_k), \quad (8.6)$$

where $\lambda_k := \lambda_k(e_1, \dots, e_{k-1})$ for $k \geq 2$ is a function of $e_1, \dots, e_{k-1}$ taking values in $[0, 1]$, and $\lambda_1 \in [0, 1]$ is a constant.

One can verify that, although not completely obvious, the arithmetic mean, the product and other U-statistic functions in (8.4) are indeed all martingale merging functions. We leave these as exercises for the reader.

A martingale merging function in (8.6) has a clear betting interpretation: At each round k , one uses λ_k proportion of the capital to bet on the next e-value. With this interpretation, the arithmetic mean corresponds to betting a fixed amount $1/K$ (but not a fixed proportion) at each round. The e-process $(M_t)_{t \in T}$ built on $(E_t)_{t \in \mathbb{T}_+}$ in Definition 7.21 is constructed by merging e-values with a martingale merging function in Definition 8.10 for each fixed t and also in a consistent way across different t .

The above connection to betting in Chapter 7 also reveals that non-monotonicity is quite natural under the betting interpretation: If we gain evidence in an early round, then we may reduce our bet in the next round, which leads to non-monotonicity of the resulting merging function (see Example 8.13 below).

It turns out that martingale merging functions play a special role among all se-merging functions. But let us first verify that martingale merging functions are exact se-merging functions.

Proposition 8.11

Any martingale merging function is an exact se-merging function. Moreover, a convex combination of martingale merging functions is again a martingale merging function.

Proof.

Let F be a martingale merging function. For sequential exact e-variables $E_1, \dots, E_K$ for $\mathbb{P}$, by writing $L_K = \lambda_K(E_1, \dots, E_{K-1})$ and $\mathcal{F}_{K-1} = \sigma(E_1, \dots, E_{K-1})$, we have

$$\begin{aligned} \mathbb{E}^\mathbb{P}[F(E_1, \dots, E_K) \mid \mathcal{F}_{K-1}] \\ &= F(E_1, \dots, E_{K-1}, 1)(1 - L_K + L_K \mathbb{E}^\mathbb{P}[E_K \mid \mathcal{F}_{K-1}]) \\ &\leq F(E_1, \dots, E_{K-1}, 1). \end{aligned}$$

Therefore, $\mathbb{E}^\mathbb{P}[F(E_1, \dots, E_K)] \leq \mathbb{E}^\mathbb{P}[F(E_1, \dots, E_{K-1}, 1)]$. Using an induction, we get

$$\mathbb{E}^\mathbb{P}[F(E_1, \dots, E_K)] \leq \mathbb{E}^\mathbb{P}[F(1, \dots, 1)] = 1.$$

This shows that the martingale merging function F is an se-merging function. If the e-variables

are exact, then the above inequalities are equalities, which shows that F is exact.

The last statement follows by noting the following equivalent formulation of a martingale merging function: $F \geq 0$ satisfies, for some measurable functions $t_k : [0, \infty)^k \rightarrow [0, \infty)$, $k \in \mathcal{K}$, that for each $k \in \mathcal{K}$,

$$\begin{aligned} F(e_1, \dots, e_k, 1, \dots, 1) \\ = F(e_1, \dots, e_{k-1}, 1, \dots, 1) + t_k(e_1, \dots, e_{k-1})(e_k - 1), \end{aligned}$$

and $F(1, \dots, 1) = 1$. A convex combination of several choices of F is the same as a convex combination of several choices of $(t_k)_{k \in \mathcal{K}}$.

The next result shows that the class of martingale merging functions is precisely that of all admissible se-merging functions.

Theorem 8.12

Any se-merging function is dominated by a martingale merging function.

The proof of Theorem 8.12 is quite technical and not pursued in this book.

Different from the merging functions obtained in Sections 8.1 and 8.2, a useful martingale merging function need not be increasing in all arguments. This is because λ_k is generally not increasing or decreasing in its arguments. Although monotonicity does not hold, any martingale merging function F satisfies a property of *sequential monotonicity*: for fixed $k \in \mathcal{K}$ and $(e_1, \dots, e_{k-1}) \in [0, \infty)^{k-1}$, the function $e_k \mapsto F(e_1, \dots, e_{k-1}, e_k, 1, \dots, 1)$ is increasing. Intuitively it means that, for given $k-1$ e-values, the overall combined e-value is larger if the next observed e-value e_k is larger, assuming that all future e-values are 1. Treating future e-values as 1 is equivalent to using $\lambda_j = 0$ for all $j > k$, or not seeing these e-values at all.

Example 8.13: Hit and stop

The non-monotonicity of F appears naturally in a “hit-and-stop” strategy: for a fixed $\alpha \in (0, 1)$ and for each $k \in \mathcal{K}$, if $F(e_1, \dots, e_k, 1, \dots, 1) \geq 1/\alpha$, then we choose $\lambda_{k+1} = 0$ (which implies $\lambda_j = 0$ for all $j \geq k+1$); otherwise we choose $\lambda_{k+1} > 0$. It is clear from (8.6) that F is not increasing. This strategy corresponds to stopping an e-process in Chapter 7 as soon as it can make a rejection at level α by up-shooting $1/\alpha$.

Example 8.14: Merging via the empirically adaptive e-process

The empirically adaptive e-process in Definition 7.21 induces a martingale merging function. For a parameter $\gamma \in (0, 1]$, the martingale merging function is defined as by (8.6) with $\lambda_1 = 0$ and λ_k for $k \in \{2, \dots, K-1\}$ are given by

$$\lambda_k(e_1, \dots, e_{k-1}) = \arg \max_{\lambda \in [0, \gamma]} \frac{1}{k-1} \sum_{s=1}^{k-1} \log((1-\lambda) + \lambda e_s).$$

This martingale merging function is uniquely determined by its parameter γ .

Advantages of the empirically adaptive martingale merging function in Example 8.14 are justified in Theorem 7.22 via its corresponding e-process. This function can be used as a default choice without prior information among se-merging functions.

8.4 Mean-variance trade-off

We now pay special attention to the product function Π_K in (8.3), which is both an ie-merging function and an se-merging function. Proposition 8.6 shows that this function is optimal in a weak sense among all ie-merging functions (hence among all se-merging functions). However, different from the case of the arithmetic mean for arbitrarily dependent e-values treated in Section 8.1, this does not suggest that Π_K is the best merging function to use for independent or sequential e-values. Indeed, Π_K has an undesirable property, which also highlights the usefulness of other martingale merging functions.

We first present a simple lemma that is useful in the proof of the next result.

Lemma 8.15

For any ie-merging function $F : [0, \infty)^K \rightarrow [0, \infty)$ and $a \geq 1$, the function $G : (e_1, \dots, e_K) \mapsto F(ae_1, e_2, \dots, e_K)/a$ is an ie-merging function.

Proof.

Fix an atomless probability measure $\mathbb{P}$. Take an event A with $\mathbb{P}(A) = 1/a$. For any independent e-variables $E'_1, \dots, E'_K$, we can take independent e-variables $E_1, \dots, E_K$ independent of A such that $(E_1, \dots, E_K)$ is distributed identically to $(E'_1, \dots, E'_K)$, since the probability space is atomless (guaranteed by Lemma A.1). Therefore, it suffices to show $\mathbb{E}^\mathbb{P}[G(E_1, \dots, E_K)] \leq 1$ for independent e-variables $E_1, \dots, E_K$ independent of A . Let $E'_1 := aE_1\mathbb{1}_A$; this is an e-variable. We can compute

$$\begin{aligned} \mathbb{E}^\mathbb{P}[G(E_1, \dots, E_K)] &= \frac{1}{a} \mathbb{E}^\mathbb{P}[F(aE_1, E_2, \dots, E_K)] \\ &\leq \mathbb{E}^\mathbb{P}[F(E'_1, E_2, \dots, E_K)] \leq 1, \end{aligned}$$

where the first inequality follows from $F \geq 0$. Hence, G is an ie-merging function.

Now we can compare the product function and other se-merging functions with respect to the moments of the resulting e-variable. Recall that $\mathbb{E}^\mathbb{Q}[E] > 1$ is the defining condition that E is powered against $\mathbb{Q}$.

Proposition 8.16

Let $F : [0, \infty)^K \rightarrow [0, \infty)$ be an se-merging function. For independent nonnegative random variables $E_1, \dots, E_K$ under $\mathbb{Q}$ with $\mathbb{E}^\mathbb{Q}[E_k] \geq 1$, $k \in \mathcal{K}$, we have, for every $m \in [1, \infty)$,

$$\mathbb{E}^\mathbb{Q}[F(E_1, \dots, E_K)^m] \leq \prod_{k=1}^K \mathbb{E}^\mathbb{Q}[E_k^m] = \mathbb{E}^\mathbb{Q}[\Pi_K(E_1, \dots, E_K)^m]. \quad (8.7)$$

In particular, if F is exact and $E_1, \dots, E_K$ are independent exact e-variables for $\mathbb{P}$, then

$$\text{var}^\mathbb{P}(F(E_1, \dots, E_K)) \leq \text{var}^\mathbb{P}(\Pi_K(E_1, \dots, E_K)). \quad (8.8)$$

Proof.

First, we argue that it suffices to show (8.7) for $E_1, \dots, E_K$ being exact e-variables for $\mathbb{Q}$. For independent nonnegative $X_1, \dots, X_K$ with mean larger than or equal to 1, set $a_k := \mathbb{E}^\mathbb{P}[X_k] \geq 1$, $E_k := X_k/a_k$ for $k \in \mathcal{K}$, and $a := \prod_{k=1}^K a_k$. Let $G : (e_1, \dots, e_K) \mapsto F(a_1 e_1, \dots, a_K e_K)/a$. Clearly, $E_1, \dots, E_K$ are independent exact e-variables. Using Lemma 8.15 repeatedly, we deduce

that G is an se-merging function. For $m \in [1, \infty)$, if (8.7) holds for all exact e-variables, then

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[F(X_1, \dots, X_K)^m] &= a^m \mathbb{E}^{\mathbb{P}} \left[\left(\frac{F(a_1 E_1, \dots, a_K E_K)}{a} \right)^m \right] \\ &= a^m \mathbb{E}^{\mathbb{P}}[G(E_1, \dots, E_K)^m] \\ &\leq a^m \mathbb{E}^{\mathbb{P}}[\Pi_K(E_1, \dots, E_K)^m] \\ &= \mathbb{E}^{\mathbb{P}}[\Pi_K(X_1, \dots, X_K)^m].\end{aligned}$$

Therefore, the general case of (8.7) follows from the case of exact e-variables.

Let $E_1, \dots, E_K$ be independent exact e-variables; we will show

$$\mathbb{E}^{\mathbb{Q}}[F(E_1, \dots, E_K)^m] \leq \prod_{k=1}^K \mathbb{E}^{\mathbb{Q}}[E_k^m]. \quad (8.9)$$

Using Theorem 8.12, F is dominated by a martingale merging function. Hence, it suffices to show the proposition for a martingale merging function F . We show the proposition by induction. Note that $\mathbb{E}^{\mathbb{Q}}[E_1^m] \geq \mathbb{E}^{\mathbb{Q}}[E_1]^m \geq 1$. Moreover, this expectation is increasing in m because for $m > t$, $\mathbb{E}^{\mathbb{Q}}[E_1^m] \geq (\mathbb{E}^{\mathbb{Q}}[E_1^t])^{m/t} \geq \mathbb{E}^{\mathbb{Q}}[E_1^t]$. The inequality (8.9) holds for $K = 1$ since, by convexity of $x \mapsto x^m$, we have

$$\mathbb{E}^{\mathbb{Q}}[(1 - \lambda + \lambda E_1)^m] \leq (1 - \lambda) + \lambda \mathbb{E}^{\mathbb{Q}}[E_1^m] \leq \mathbb{E}^{\mathbb{Q}}[E_1^m]$$

for all $\lambda \in [0, 1]$. To argue by induction, suppose that

$$\mathbb{E}^{\mathbb{Q}}[G(E_1, \dots, E_{K-1})^m] \leq \prod_{k=1}^{K-1} \mathbb{E}^{\mathbb{Q}}[E_k^m] \quad (8.10)$$

for every se-merging function $G : [0, \infty)^{K-1} \rightarrow [0, \infty)$. Since F is a martingale merging function, we can write

$$F(e_1, \dots, e_K) = G(\mathbf{e}_{[K-1]}) (1 - \lambda_K(\mathbf{e}_{[K-1]}) + \lambda_K(\mathbf{e}_{[K-1]}) e_K),$$

for some $\lambda_K : [0, \infty)^{K-1} \rightarrow [0, 1]$, where $\mathbf{e}_{[K-1]} = (e_1, \dots, e_{K-1})$. Let us write $\mathbf{Y} = (E_1, \dots, E_{K-1})$ and $L = \lambda_K(E_1, \dots, E_{K-1})$. We have

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}[F(E_1, \dots, E_K)^m \mid \mathbf{Y}] &= \mathbb{E}^{\mathbb{Q}}[G(\mathbf{Y})^m (1 - L + L E_K)^m \mid \mathbf{Y}] \\ &\leq G(\mathbf{Y})^m (1 - L + L \mathbb{E}^{\mathbb{Q}}[E_K^m \mid \mathbf{Y}]) \\ &\leq G(\mathbf{Y})^m \mathbb{E}^{\mathbb{Q}}[E_K^m].\end{aligned}$$

As a consequence,

$$\mathbb{E}^{\mathbb{P}}[F(E_1, \dots, E_K)^m] \leq \mathbb{E}^{\mathbb{P}}[G(\mathbf{Y})^m] \mathbb{E}^{\mathbb{P}}[E_K^m] \leq \prod_{k=1}^K \mathbb{E}^{\mathbb{P}}[E_k^m],$$

where the last inequality follows by the inductive assumption (8.10). Therefore, we obtain (8.7). If F is an exact se-merging function, we obtain (8.8) from (8.9) with $m = 2$ and $\mathbb{P} = \mathbb{Q}$ since $\mathbb{E}^{\mathbb{P}}[F(E_1, \dots, E_K)] = \mathbb{E}^{\mathbb{P}}[\Pi_K(E_1, \dots, E_K)] = 1$.

Proposition 8.16 has two implications. First, under the alternative $\mathbb{Q}$, if the e-variables are powered, then all moments of the resulting combined e-variable are maximized by the product function. Second, under the

null, the product function Π_K results in the largest variance among all exact se-merging functions if the e-variables are exact and independent.

Although having a large mean under $\mathbb{Q}$ can be seen as desirable, having a large variance is generally not a desirable property. Hence the product function can be seen as an extreme choice of ie-merging function.

In particular, if the mean under $\mathbb{Q}$, $\mathbb{E}^{\mathbb{Q}}[F(E_1, \dots, E_K)]$, is chosen as an objective to optimize, then the product function Π_K dominates all other se-merging functions F under the assumption that $\mathbb{E}^{\mathbb{Q}}[E_k] \geq 1$ for all k . Putting this into the context of testing by betting as in Section 7.8, this means the “all-in” strategy of choosing $\lambda_t = 1$ for all t yields an e-process with the largest expected value at any given time, if each sequential e-value has mean larger than 1 under $\mathbb{Q}$. Nevertheless, we remind the reader that e-power is the right objective, instead of the expected value. Recall that the e-power is measured by the expected logarithm under $\mathbb{Q}$, not any of the moments. Generally, $\Pi_K(E_1, \dots, E_K)$ does not have the largest e-power. A comparison is provided in Example 8.17 below.

Example 8.17: All-in versus log-optimal

We consider the e-processes built on $(E_k)_{k \in \mathbb{N}}$ as in Definition 7.21, allowing an infinite number of e-variables. Suppose that under $\mathbb{Q}$, each E_k for $k \in \mathbb{N}$ is distributed as $\mathbb{Q}(E_k = 4) = \mathbb{Q}(E_k = 0) = 1/2$ and they are independent.

- (a) The all-in strategy is to choose $\lambda_k = 1$ for all k , leading to $M_k^{\text{all-in}} = \Pi_k(E_1, \dots, E_k)$ for any $k \in \mathbb{N}$, which has two-point distribution with probability 2^{-k} taking the value 4^k and with probability $1 - 2^{-k}$ taking the value 0.
- (b) On the other hand, we can compute that the $\mathbb{Q}$ -log-optimal e-process $(M_k^*)_{k \in \mathbb{N}}$ built on $(E_t)_{t \in [k]}$ in Definition 7.21 has $\lambda_k = 1/3$ for all $k \in \mathbb{N}$, leading to $\mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda_k + \lambda_k E_t)] = \log(4/3)$. Then, by the law of large numbers, $M_k^* \sim (4/3)^k$ $\mathbb{Q}$ -almost surely as $k \rightarrow \infty$. (This asymptotic growth rate is also achieved by the empirically adaptive e-process by Theorem 7.22.)

One can see that, although $\mathbb{E}^{\mathbb{Q}}[M_k^{\text{all-in}}] \geq \mathbb{E}^{\mathbb{Q}}[M_k^*]$ for all $k \in \mathbb{N}$ (indeed, $M_k^{\text{all-in}}$ has the largest expectation among all e-processes built on $(E_t)_{t \in [k]}$ by Proposition 8.16), its distribution has a large chance of getting 0, and moreover $M_k^{\text{all-in}} \rightarrow 0$ $\mathbb{Q}$ -almost surely. These are highly undesirable features of the all-in betting strategy.

Summary

The following points summarize the results in the previous few sections on choosing merging functions.

- (i) For merging arbitrarily dependent e-values, one needs to use the arithmetic mean or a weighted arithmetic mean.
- (ii) For merging sequential e-values, one needs to use martingale merging functions. In particular, the empirically adaptive martingale merging function in Example 8.14 can be used as a default method without prior information on the e-values’ behavior under the alternative hypothesis.
- (iii) For merging independent e-values, one can use, among many choices, the product or the U-statistic functions. Although the product e-merging function has a weak optimality, it also suffers from an undesirable property of having large variance.
- (iv) For merging iid e-values, the empirically adaptive martingale merging function in Example 8.14 has asymptotic log-optimality in the sense of Theorem 7.22.

8.5 Family-wise error rate and merging e-values

The merging methods developed in this chapter can be used to design multiple testing procedures based on e-values. In Chapter 9, we will study procedures that control the false discovery rate in great detail. In this

Algorithm 8.18 Adjusting e-values for FWER using the arithmetic mean

Require: A sequence of e-values $e_1, \dots, e_K$ (realizations of $E_1, \dots, E_K$).

```

1: Find a permutation  $\pi : \mathcal{K} \rightarrow \mathcal{K}$  such that  $e_{\pi(1)} \leq \dots \leq e_{\pi(K)}$ .
2: Set  $e_{(k)} := e_{\pi(k)}$ ,  $k \in \mathcal{K}$  (these are the ascending order statistics).
3:  $S_0 := 0$ 
4: for  $i = 1, \dots, K$  do
5:    $S_i := S_{i-1} + e_{(i)}$ 
6: for  $k = 1, \dots, K$  do
7:    $e_{\pi(k)}^* := e_{\pi(k)}$ 
8:   for  $i = 1, \dots, k - 1$  do
9:      $e := \frac{e_{\pi(k)} + S_i}{i+1}$ 
10:    if  $e < e_{\pi(k)}^*$  then
11:       $e_{\pi(k)}^* := e$ 

```

section, we briefly explain how e-values can be used for testing multiple hypotheses with family-wise error rate (FWER) control, and we omit procedures for other error metrics.

We give a minimal set of definitions here, and they will be re-elaborated in Chapter 9. Consider K different null hypotheses, described by sets of probability measures $\mathcal{P}_1, \dots, \mathcal{P}_K$. For a probability measure $\mathbb{P} \in \mathcal{M}_1$, the null set of hypotheses is given by $\mathcal{N} = \{k \in \mathcal{K} : \mathbb{P} \in \mathcal{P}_k\}$ (which depends on $\mathbb{P}$).

Suppose that E_k is an e-variable for $\mathcal{P}_k$, $k \in \mathcal{K}$. For any set $A \subseteq \mathcal{K}$, we can define an e-variable for the intersection of $\mathcal{P}_k$, $k \in A$, via

$$E_A = F(E_k : k \in A),$$

where F is an e-merging function of $|A|$ e-values (or an ie-merging function if we assume that the e-variables are independent). For instance, taking F as the arithmetic mean $\mathbb{M}_{|A|}$, we get the e-variable

$$E_A = \frac{1}{|A|} \sum_{k \in A} E_k. \quad (8.11)$$

The collection $(E_A)_{A \subseteq \mathcal{K}}$ in (8.11) is called the *mean e-collection*.

With any e-merging function F , E_A is an e-variable for $\bigcap_{k \in A} \mathcal{P}_k$. An e-testing procedure takes K e-values as input and output a set of indices $\mathcal{D} \subseteq \mathcal{K}$ (more formal definitions are given in Chapter 9). We would like the procedure to control; FWER at level $\alpha \in (0, 1)$, meaning the guarantee

$$\mathbb{P}(\mathcal{D} \cap \mathcal{N} \neq \emptyset) \leq \alpha$$

for all $\mathbb{P} \in \mathcal{M}_1$. In other words, the probability of rejecting any null hypothesis is at most α , for every possible distribution $\mathbb{P}$.

A simple procedure to control FWER based on $(E_A)_{A \subseteq \mathcal{K}}$ is described below. Let

$$E_k^* = \min_{A \subseteq \mathcal{K} : k \in A} E_A.$$

The e-testing procedure $\mathcal{D}$ rejects all indices $k \in \mathcal{K}$ with $E_k^* \geq 1/\alpha$. To see its FWER control, note that for any $\mathbb{P} \in \mathcal{M}_1$,

$$\max_{k \in \mathcal{N}} E_k^* = \max_{k \in \mathcal{N}} \min_{A : k \in A} E_A \leq \max_{k \in \mathcal{N}} E_{\mathcal{N}} = E_{\mathcal{N}},$$

and

$$\mathbb{P}(\mathcal{D} \cap \mathcal{N} \neq \emptyset) = \mathbb{P}\left(\max_{k \in \mathcal{N}} E_k^* \geq 1/\alpha\right) \leq \mathbb{P}(E_{\mathcal{N}} \geq 1/\alpha) \leq \alpha.$$

Note that for the FWER control, we only require that each E_A is an e-variable for $\bigcap_{k \in A} \mathcal{P}_k$, and it need not be computed by merging e-values from the individual hypotheses. Such $(E_A)_{A \subseteq \mathcal{K}}$ is called an e-collection; see Section 9.6.

For the mean e-collection in (8.11), a simple $O(K^2)$ algorithm to compute E_k^* from the original e-values $E_1, \dots, E_K$ is described in Algorithm 8.18. If we use the product ie-merging function instead, then a simple $O(K)$ algorithm is available, omitted here. Algorithm 8.18 effectively computes

$$E_{\pi(k)}^* = \min_{i \in [k-1]} \frac{E_{\pi(k)} + E_{(1)} + \dots + E_{(i)}}{i + 1}$$

for $k \in \mathcal{K}$, where π is the ordering permutation such that $E_{(j)} = E_{\pi(j)}$ is the j th ascending order statistic among $E_1, \dots, E_K$.

Bibliographical note

The content in this chapter is mainly based on Vovk and Wang [2021, 2024a]. E-merging functions were introduced by Vovk and Wang [2021]. The proof of Theorem 8.12 is in Vovk and Wang [2024a]. Negative lower orthant dependence was introduced by Block et al. [1982], and the corresponding results on merging e-values are in Chi et al. [2024]. The FWER controlling procedures based on e-values were studied by Vovk and Wang [2021] and Hartog and Lei [2025], while recent connections to closed testing were explored in Xu et al. [2025] and Goeman et al. [2025].

Chapter 9

False discovery rate control using compound e-values

Our purpose in this chapter is to show how e-values are (a) useful for multiple testing while controlling the false discovery rate (FDR), (b) inherent and central to all FDR controlling procedures. The main setting is explained next, part of which is briefly mentioned in Section 8.5.

Suppose we observe data X drawn according to some unknown distribution $\mathbb{P}^\dagger$. Here, X denotes the entire dataset available to the researcher. We consider testing K different hypotheses. The null hypotheses are described by sets of probability measures $\mathcal{P}_1, \dots, \mathcal{P}_K$. Recall the notation $\mathcal{K} = [K]$, which avoids a clash with certain decreasing order statistics used in this chapter. For $k \in \mathcal{K}$, the k -th null hypothesis H_k is $\mathbb{P}^\dagger \in \mathcal{P}_k$, corresponding to $H_0 : \mathbb{P}^\dagger \in \mathcal{P}$ in Chapter 1. Note that it is possible that $\mathbb{P}^\dagger \notin \bigcup_{k \in \mathcal{K}} \mathcal{P}_k$, meaning that all null hypotheses could be false. As in the setting of single hypothesis testing, the unknown data-generating $\mathbb{P}^\dagger$ is used to formally state the multiple hypothesis testing problem and will not appear in the subsequent analysis.

We write

$$\mathcal{N}(\mathbb{P}) = \{k \in \mathcal{K} : \mathbb{P} \in \mathcal{P}_k\} \subseteq \mathcal{K}, \quad \mathbb{P} \in \mathcal{M}_1,$$

which is the index set of true null hypotheses if $\mathbb{P}$ is the data generating distribution. Denote by $\mathcal{N} = \mathcal{N}(\mathbb{P})$ and $K_0 = |\mathcal{N}(\mathbb{P})|$, where we omit the reliance on $\mathbb{P}$. In the setting of multiple testing, we will let $\mathcal{P}$ denote a generic set of distributions. Unless otherwise specified, we simply take (explained in Section 9.1.1)

$$\mathcal{P} = \bigcup_{k \in \mathcal{K}} \mathcal{P}_k.$$

Other specifications of $\mathcal{P}$ may represent some assumptions or restrictions on the distributions in the testing problem, which will be useful in Section 9.1.4 and Chapter 10. A central component of this chapter is the e-Benjamini–Hochberg (e-BH) procedure, introduced in Section 9.1.

9.1 Compound e-values and the e-BH procedure

9.1.1 Compound e-variables

A typical setting for our study here is that each hypothesis is associated with an e-variable E_k . Sometimes, we also consider and compare with the classic setting, where each hypothesis is associated with a p-value P_k . However, more generally, we will only require $(E_1, \dots, E_K)$ to satisfy a more relaxed definition, given below.

Definition 9.1: Compound e-variables

Fix the null hypotheses $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ and a set $\mathcal{P}$ of distributions. Let $E_1, \dots, E_K$ be $[0, \infty]$ -valued random variables. We say that $E_1, \dots, E_K$ are *compound* e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$ if

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k] \leq K \quad \text{for all } \mathbb{P} \in \mathcal{P}.$$

They are called *tight* compound e-variables if the supremum of the left hand side over $\mathbb{P} \in \mathcal{P}$ equals K .

We omit “under $\mathcal{P}$ ” in case $\mathcal{P}$ is the set of all distributions. For all conditions to be checked in the above definition, it suffices to consider $\mathbb{P} \in \bigcup_{k \in \mathcal{K}} \mathcal{P}_k$. Therefore, it is without loss of generality to assume $\mathcal{P} \subseteq \bigcup_{k \in \mathcal{K}} \mathcal{P}_k$, and by default (without any restrictions) it is

$$\mathcal{P} = \bigcup_{k \in \mathcal{K}} \mathcal{P}_k.$$

When $K = 1$, Definition 9.1 coincides with the definition of an e-variable. Of course, a vector of e-variables is a trivial but important special case of compound e-variables. As is clear from the definition, no restriction is placed on the dependence structure of the e-variables.

Example 9.2: Weighted e-values

If each E_k is an e-variable for $\mathcal{P}_k$, then $E_1, \dots, E_K$ are compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$. Further, let $w_1, \dots, w_K \geq 0$ be deterministic nonnegative numbers such that $\sum_{k \in \mathcal{K}} w_k \leq K$ and define $\tilde{E}_k = E_k w_k$. Then $\tilde{E}_1, \dots, \tilde{E}_K$ are compound e-variables.

Note that the class of compound e-values is much richer than the class of weighted e-values in Example 9.2, because the later satisfies the stronger constraint

$$\sum_{k \in \mathcal{K}} \sup_{\mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[\tilde{E}_k] \leq K,$$

compared to the condition for the compound e-values

$$\sup_{\mathbb{P} \in \bigcup_{k \in \mathcal{K}} \mathcal{P}_k} \sum_{k \in \mathcal{N}(\mathbb{P})} \mathbb{E}^{\mathbb{P}}[E_k] \leq K.$$

We now give a more elaborate example of a nonparametric setting where one naturally encounters compound e-variables. More importantly, these e-variables are dependent in a complicated way, making the usual assumptions on the dependence structure, such as independence, inapplicable.

Example 9.3: Multi-armed bandit testing

Suppose there are K traders (or machines), and a researcher is interested in knowing which ones are skillful (or useful). This is a classic problem in finance. For $k \in \mathcal{K}$, the null hypothesis H_k is that trader k is not skillful, meaning that they make no profit on average (without loss of generality we can assume the market risk-free return rate is 0). The nonnegative random variables $X_{k,1}, \dots, X_{k,n}$ are the monthly realized performance (i.e., the ratio of payoff to investment; $X_{k,t} > 1$ presents a profit and $X_{k,t} < 1$ means a loss) of agent k from month 1 to month n . The no-skill null hypothesis is $\mathbb{E}[X_{k,t} | \mathcal{F}_{t-1}] \leq 1$ for $t \in [n]$, where the σ -field $\mathcal{F}_t$ represents the available market information up to time $t \in \{0, \dots, n\}$, and we naturally assume that $(X_{k,t})_t$ is adapted to $(\mathcal{F}_t)_t$. Since the agents are changing investment strategies over time and all strategies depend on the financial market evolution, there is complicated serial dependence within $(X_{k,1}, \dots, X_{k,n})$ for single k , as well as cross dependence among agents $k \in \mathcal{K}$. Because of the complicated serial dependence and the lack of distributional assumptions of the performance data, it is difficult to obtain useful p-values for these agents. Nevertheless, we can easily obtain useful e-values: For instance, $E_k = \prod_{t=1}^n X_{k,t}$ is a valid e-value, as well as any mixture of U-statistics of $X_{k,1}, \dots, X_{k,n}$, including the mean and the product; indeed, $X_{k,1}, \dots, X_{k,n}$ are sequential e-variables in Definition 7.19, and they can be combined using merging methods in Chapter 8. Moreover, the obtained e-values $E_1, \dots, E_K$ are dependent in a complicated way. Even if these e-values are not very large, they can be useful for other studies on these traders. For this problem, other sophisticated e-values can also be constructed such as the ones from the empirically adaptive e-processes; see Chapter 7.

In the above setting, we implicitly assumed that each agent has the same initial wealth 1. If they have different initial wealth values $c_1, \dots, c_K$ with $\bar{c} = (c_1 + \dots + c_K)/K$, then the wealth of the k -th trader is $E'_k = c_k E_k$, and it is easy to check that $E'_1/\bar{c}, \dots, E'_K/\bar{c}$ are compound e-variables, while $E'_1/c_1, \dots, E'_K/c_K$ are e-variables.

We conclude this section with two general constructions of compound e-variables.

Example 9.4: Convex combinations of compound e-values

Suppose that $E_1^{(\ell)}, \dots, E_K^{(\ell)}$ are compound e-variables for $\ell = 1, \dots, L$. Let $w_1, \dots, w_L \geq 0$ be deterministic nonnegative numbers such that $\sum_{\ell=1}^L w_\ell = 1$. Define $E_k = \sum_{\ell=1}^L w_\ell E_k^{(\ell)}$. Then $E_1, \dots, E_K$ are compound e-variables.

Example 9.5: Combining evidence from follow-up studies

Suppose we summarize data X into compound e-variables $E_1, \dots, E_K$. Based on these, suppose that we select a subset of indices $\mathcal{S} \subseteq \mathcal{K}$ of interest, for which we seek additional evidence. With this goal in mind, suppose we collect additional data Y and summarize it in the form of e-variables $\{E'_j\}_{j \in \mathcal{S}}$ that are conditionally valid given the past data; in particular, $\mathbb{E}^\mathbb{P}[E'_j | E_j] \leq 1$ for all $j \in \mathcal{S}$ with $\mathbb{P} \in \mathcal{P}_j$. Then defining $E'_j = 1$ for $j \notin \mathcal{S}$, we have that $E_1 E'_1, \dots, E_K E'_K$ are also compound e-variables.

9.1.2 FDR and the e-BH procedure

We first define the concept of FDR, central to modern multiple hypotheses testing.

Definition 9.6: Multiple testing procedures

Let X represent the available data used for testing.

- (i) A *multiple testing procedure* $\mathcal{D}$ is a Borel function of X that produces a subset of $\mathcal{K}$ representing the indices of rejected hypotheses. A testing procedure $\mathcal{D}$ that takes e-values or compound e-values as input is called an *e-testing procedure*.
- (ii) The hypotheses that are rejected by $\mathcal{D}$, given by $\mathcal{D}(X) \in 2^\mathcal{K}$, are called *discoveries*.

We overload notation and write both the procedure (the mapping from X to $\mathcal{K}$) and the realized set of discoveries (a random subset of $\mathcal{K}$) as $\mathcal{D}$, because it is typically clear which one is meant from context. We use the convention

$$\frac{0}{0} = 0$$

in the calculation of error metrics in multiple hypothesis testing.

Definition 9.7: FDP and FDR

In the following objects, $\mathcal{D}$ should be interpreted as the random subset $\mathcal{D}(X)$ of $\mathcal{K}$.

- (i) A *false discovery* is a true null hypothesis rejected by $\mathcal{D}$. Let

$$F_{\mathcal{D}} = \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbf{1}_{\{k \in \mathcal{D}\}} = |\mathcal{N}(\mathbb{P}) \cap \mathcal{D}| \quad \text{and} \quad R_{\mathcal{D}} = |\mathcal{D}|,$$

which are the number of false discoveries and the total number of discoveries, respectively.

- (ii) The *false discovery proportion (FDP)* is $F_{\mathcal{D}}/R_{\mathcal{D}}$.

- (iii) The *false discovery rate (FDR)* under $\mathbb{P}$ is the expected value of the FDP, that is,

$$\text{FDR}_{\mathcal{D}} = \mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \right] = \mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}} \vee 1} \right].$$

- (iv) A multiple testing procedure $\mathcal{D}$ has *FDR control* (at most) $\alpha \in (0, 1)$ under $\mathcal{P}$ if $\text{FDR}_{\mathcal{D}} \leq \alpha$ for all $\mathbb{P} \in \mathcal{P}$. We omit “under $\mathcal{P}$ ” when $\mathcal{P} = \mathcal{M}_1$ or $\mathcal{P} = \bigcup_{k \in \mathcal{K}} \mathcal{P}_k$.

Clearly, both $\text{FDR}_{\mathcal{D}}$ and $F_{\mathcal{D}}$ depend on the unknown $\mathbb{P}$, and we omit this reliance for brevity. Similarly to the case of compound e-values, FDR is trivially 0 if $\mathbb{P}$ is outside $\bigcup_{k \in \mathcal{K}} \mathcal{P}_k$, so we do not need to check such $\mathbb{P}$.

The FDR has been the most popular error metric in multiple testing, with wide applications in many scientific disciplines. There are many FDR-controlling procedures with p-values as the input. The Benjamini–Hochberg (BH) procedure of [Benjamini and Hochberg \[1995\]](#) is arguably the most popular and successful in this context (which is, to our knowledge, the most cited paper in statistics at the time of writing of this book). We will study a corresponding e-testing procedure called the e-BH procedure, which as an intimate connection to the BH procedure that will be explained later.

Definition 9.8: The e-BH procedure

Suppose that $e_1, \dots, e_K$ are realized compound e-values for the hypotheses $(\mathcal{P}_1, \dots, \mathcal{P}_K)$. For $k \in \mathcal{K}$, let $e_{[k]}$ be the k -th order statistic of $e_1, \dots, e_K$, from the largest to the smallest. The *e-BH procedure at level* $\alpha \in (0, 1)$ rejects all hypotheses with the largest k^* e-values, where

$$k^* := \max \left\{ k \in \mathcal{K} : \frac{ke_{[k]}}{K} \geq \frac{1}{\alpha} \right\},$$

with the convention $\max(\emptyset) = 0$.

A stylized illustration of the e-BH procedure is depicted in [Figure 9.10](#).

Before showing the FDR guarantee of the e-BH procedure, we slightly enlarge the scope of our methodology.

Definition 9.9: Self-consistent e-testing procedures

An e-testing procedure $\mathcal{D}$ is *self-consistent at level $\alpha \in (0, 1)$* if every rejected hypothesis H_k has a realized compound e-value e_k that satisfies

$$e_k \geq \frac{K}{\alpha R_{\mathcal{D}}}.$$

This class includes the e-BH procedure as a special case.

The e-BH procedure dominates all other self-consistent e-testing procedures by definition, meaning that it maximizes the number of discoveries within this class. Nevertheless, self-consistent testing procedures that reject a smaller set than that of the e-BH procedure may be useful if we want the set of discoveries to also satisfy some additional structural or logical constraint. For example, if the hypotheses are structured as a graph, then we may want to reject a set of hypotheses that form connected subgraphs.

The main result in this section is the FDR control of self-consistent e-testing procedures for any compound e-variables.

Theorem 9.11: FDR control and post-hoc FDR control

Any self-consistent e-testing procedure at level $\alpha \in (0, 1)$, including the e-BH procedure, has FDR at most α . This claim holds for any $\mathbb{P} \in \mathcal{M}_1$ and for arbitrary compound e-variables, regardless of the dependence structure that $\mathbb{P}$ induces amongst them.

In fact, for any compound e-variables, and for any class of self-consistent e-testing procedures $(\mathcal{D}_\alpha)_{\alpha \in (0, 1)}$ indexed by their level α , the post-hoc FDR guarantee holds for any $\mathbb{P} \in \mathcal{M}_1$:

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0, 1)} \frac{F_{\mathcal{D}_\alpha}}{\alpha R_{\mathcal{D}_\alpha}} \right] \leq 1.$$

Further, if the compound e-variables are in fact e-variables, then the upper bound above can be improved to K_0/K .

The last claim above allows for a formal guarantee even when α is itself a function of the data. Omitting the supremum in the final statement yields the preceding FDR claim.

Proof.

Let $\mathbf{E} = (E_1, \dots, E_K)$ be an arbitrary vector of compound e-variables fed to the testing procedure $\mathcal{D}_\alpha$ and $\mathcal{D}_\alpha(\mathbf{E})$ be the set of rejected hypotheses. By direct substitution, we have

$$\begin{aligned} \frac{F_{\mathcal{D}_\alpha}}{\alpha R_{\mathcal{D}_\alpha}} &= \frac{|\mathcal{D}_\alpha(\mathbf{E}) \cap \mathcal{N}|}{\alpha(R_{\mathcal{D}_\alpha} \vee 1)} = \sum_{k \in \mathcal{N}} \frac{\mathbb{1}_{\{k \in \mathcal{D}_\alpha(\mathbf{E})\}}}{\alpha(R_{\mathcal{D}_\alpha} \vee 1)} \\ &\leq \sum_{k \in \mathcal{N}} \frac{\mathbb{1}_{\{k \in \mathcal{D}_\alpha(\mathbf{E})\}} E_k}{K} \leq \sum_{k \in \mathcal{N}} \frac{E_k}{K}, \end{aligned}$$

where the first inequality is due to self-consistency. By the definition of compound e-variables, we immediately obtain for arbitrary $\mathbb{P} \in \mathcal{M}_1$ that

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0, 1)} \frac{F_{\mathcal{D}_\alpha}}{\alpha R_{\mathcal{D}_\alpha}} \right] \leq 1.$$

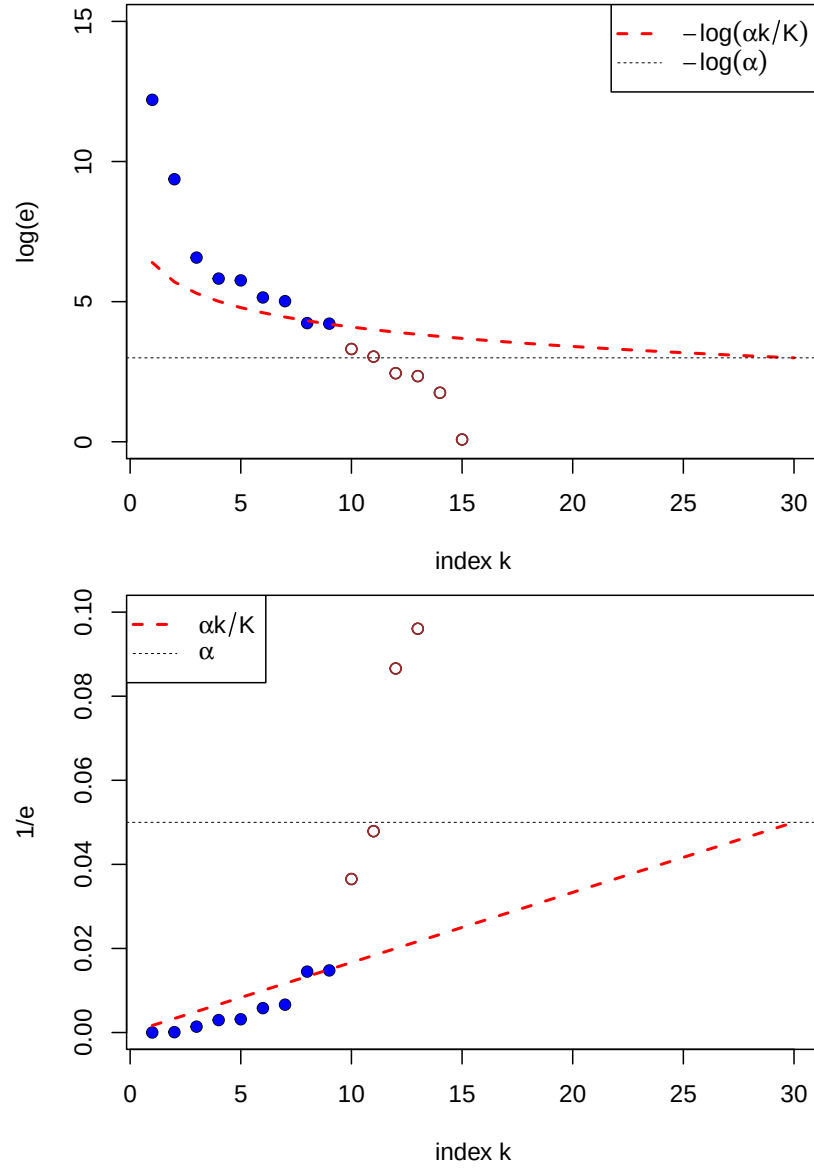


Figure 9.10: An illustration of the e-BH procedure in Definition 9.8 applied to a simulated sample with $K = 30$ and $\alpha = 0.05$. We compare $e_{[k]}$ with $K/(\alpha k)$ for $k \in \mathcal{K}$ on the log scale in the top panel and on the reciprocal scale in the bottom panel (some small e-values are outside the plotted relevant range). Hypotheses associated with e-values no less than the last point above the red dashed curve in the top panel are rejected, including one that lies below the curve. The bottom panel also illustrates the BH procedure in Definition 9.14, with the reciprocals of e-values as p-values.

If $E_1, \dots, E_K$ are e-variables, then we have

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{F_{\mathcal{D}_\alpha}}{\alpha R_{\mathcal{D}_\alpha}} \right] \leq \sum_{k \in \mathcal{N}} \mathbb{E}^{\mathbb{P}} \left[\frac{E_k}{K} \right] \leq \frac{K_0}{K},$$

as claimed.

Most notably, the e-BH procedure controls FDR regardless of how the e-variables are dependent. This key feature distinguishes it from the usual BH procedure with input p-values, for which the FDR guarantee requires certain dependence assumptions, a topic we return to later in Section 9.1.4.

In the next proposition we provide an alternative description of the e-BH procedure, which illustrates that the e-BH procedure rejects e-values according to a threshold t_α that depends on all other input e-values. This alternative description is useful in Section 9.1.3.

Proposition 9.12

Let $e_1, \dots, e_K \in \mathbb{R}_+$ and $\alpha \in (0, 1)$. Define

$$R(t) = |\{k \in \mathcal{K} : e_k \geq t\}| \vee 1; \quad t_\alpha = \inf\{t \in [0, \infty) : tR(t) \geq K/\alpha\}.$$

For each k , the e-BH procedure at level α applied to the realizations $e_1, \dots, e_K$ and rejects H_k if and only if $e_k \geq t_\alpha$. Moreover, $t_\alpha R(t_\alpha) = K/\alpha$, and t_α takes values in $\{K/(k\alpha) : k \in \mathcal{K}\}$.

Proof.

For $t \in [0, \infty)$, let $f(t)$ be the number of true null hypotheses with an e-value e_k larger than or equal to t . Define the quantity $g(t) = tR(t)/K$, and by definition $t_\alpha = \inf\{t \in [0, \infty) : g(t) \geq 1/\alpha\}$. Clearly $t_\alpha \in [1/\alpha, K/\alpha]$ since $g(t) \leq t$ and $R(t) \geq 1$. Since g only has downside jumps and $g(0) = 0$, we know $g(t_\alpha) = 1/\alpha$, and thus $t_\alpha R(t_\alpha) = K/\alpha$.

If $e_k \geq t_\alpha$, then H_k is rejected by the definition of e-BH. If $e_{[k]} < t_\alpha$, then by definition of g , we have

$$\frac{ke_{[k]}}{K} \leq \frac{R(e_{[k]})e_{[k]}}{K} = g(e_{[k]}) < 1/\alpha.$$

Thus, each H_k is rejected by the e-BH procedure if and only if $e_k \geq t_\alpha$.

9.1.3 Boosting the e-values with distributional information

If we know some information of the null distribution of the e-variables $E_1, \dots, E_K$ fed to the e-BH procedure, then we can enhance the power of the e-BH procedure by a mechanism called *boosting e-values*.

Let

$$K/\mathcal{K} := \{K/k : k \in \mathcal{K}\},$$

and define a truncation function $T : [0, \infty) \rightarrow [0, K]$ by letting $T(x)$ be the largest number in $K/\mathcal{K} \cup \{0\}$ that is no larger than x . In other words,

$$T(x) = \frac{K}{\lceil K/x \rceil} \mathbf{1}_{\{x \geq 1\}} \quad \text{with } T(\infty) = K. \quad (9.1)$$

Note that T truncates x to take only values in $K/\mathcal{K} \cup \{0\}$. For each $k \in \mathcal{K}$, take a *boosting factor* $b_k \geq 1$ such that

$$\mathbb{E}^{\mathbb{P}}[T(\alpha b_k E_k)] \leq \alpha \quad \text{for } \mathbb{P} \in \mathcal{P}_k, \quad (9.2)$$

and let $E'_k = b_k E_k$. We call $E'_1, \dots, E'_K$ the boosted e-values. Note that choosing $b_k = 1$ always satisfies (9.2) because $T(x) \leq x$ for all $x \in \mathbb{R}_+$. One would try to choose the largest b_k possible subject to (9.2). Computing the precise value of b_k depends on our knowledge of the distribution of E_k under $\mathbb{P}$. In some situations, one may have partial but not full distributional information of E_k under $\mathbb{P}$, leading to a smaller b_k than the one that attains equality in (9.2). If no distributional information is available, i.e., one only knows that E_k is an e-variable for $\mathcal{P}_k$ but nothing more, then one has to set $b_k = 1$.

Theorem 9.13

Applied to arbitrary nonnegative random variables $E'_1, \dots, E'_K$, the e-BH procedure $\mathcal{D}$ at level $\alpha \in (0, 1)$ satisfies

$$\mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \right] \leq \frac{1}{K} \sum_{k \in \mathcal{N}} \mathbb{E}^{\mathbb{P}} [T(\alpha E'_k)],$$

for any $\mathbb{P} \in \mathcal{M}_1$. In particular, if $E'_1, \dots, E'_K$ are e-variables for $\mathcal{P}_1, \dots, \mathcal{P}_K$ or the boosted e-values via (9.2), then the FDR is controlled at $K_0 \alpha / K$.

The proof of Theorem 9.13 can be described with a very simple intuition: If e-values $e_1, \dots, e_K$ are replaced by $T(\alpha e_1)/\alpha, \dots, T(\alpha e_K)/\alpha$, then they will lead to the same set of rejected hypotheses, because by Proposition 9.12, e-values are rejected only at thresholds in $(\alpha^{-1}K)/K$. Therefore, for the desired FDR control, it suffices to verify that $T(\alpha E_k)/\alpha$ are e-values. This leads to condition (9.2), justifying the use of $b_k E_k$.

We can also boost compound e-values by using $b_1, \dots, b_K$ satisfying

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}} [T(\alpha b_k E_k)] \leq K \alpha \quad \text{for } \mathbb{P} \in \mathcal{P},$$

following the same argument.

9.1.4 Connection between the e-BH and the BH procedures

There are two main connections we wish to emphasize in this subsection. First, the BH procedure and its guarantees under PRDS can be recovered as an instance of the e-BH procedure via a certain *boosting* operation, and second, the BH procedure with the Benjamini–Yekutieli correction (which we call the BHY procedure for brevity) can be recovered by applying the e-BH procedure to some e-values. We do not expand on the first point here, but do prove the second in Proposition 9.16. In sum, BH is a special case of a *boosted* e-BH, and BHY is a special case of e-BH.

Definition 9.14: BH and BHY procedures

Suppose that $p_1, \dots, p_K$ are realized p-values for the hypotheses $(\mathcal{P}_1, \dots, \mathcal{P}_K)$. For $k \in \mathcal{K}$, let $p_{(k)}$ be the k -th order statistics of $p_1, \dots, p_K$, from the smallest to the largest. The *BH procedure* at level $\alpha \in (0, 1)$ rejects all hypotheses with the smallest k^* p-values, where

$$k^* = \max \left\{ k \in \mathcal{K} : \frac{K p_{(k)}}{k} \leq \alpha \right\},$$

with the convention $\max(\emptyset) = 0$. The *BHY procedure* is to apply the BH procedure to $\ell_K p_1, \dots, \ell_K p_K$, where $\ell_K = \sum_{k=1}^K k^{-1}$.

The following are classic results:

- (i) The BHY procedure at level α controls FDR at level $(K_0/K)\alpha$.
- (ii) The BH procedure at level α controls FDR at level $(K_0/K)\alpha$ if the p-variables are independent or satisfy the PRDS condition defined below (that is, under the set $\mathcal{P}$ of distributions satisfying the specified conditions); see Remark 9.17. A generalization of this is Theorem 11.10.

Definition 9.15: Positive regression dependence on a subset

A set $A \subseteq \mathbb{R}^K$ is said to be *increasing* if $\mathbf{x} \in A$ implies $\mathbf{y} \in A$ for all $\mathbf{y} \geq \mathbf{x}$.

A vector $(P_1, \dots, P_K)$ of p-variables satisfies positive regression dependence on a subset (PRDS) if for any null index $k \in \mathcal{N}$ and increasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}\{(P_\ell)_{\ell \in \mathcal{K}} \in A \mid P_k \leq x\}$ is increasing on $[0, 1]$.

A caveat of PRDS in Definition 9.15 is that it enforces certain positive dependence between the nulls and non-nulls, and it is often difficult to verify in many applications, such as genome-wise association studies and finance (Example 9.3).

We can see that the price to pay to get validity under arbitrary dependence for the BH procedure is a factor of ℓ_K , which is close to $\log K$. Since p-values are less useful when they are large, this correction makes the BH procedure to reject less, and typically much less, hypotheses.

With e-variables $E_1, \dots, E_K$, it is immediate that the e-BH procedure is precisely the BH procedure applied to $1/E_1, \dots, 1/E_K$. Since $P_k := 1/E_k$ is a p-variable for $\mathcal{P}_k$ (Proposition 2.4), if the BH procedure controls FDR with p-values $P_1, \dots, P_K$, then the e-BH procedure controls FDR with e-values $E_1, \dots, E_K$. However, the e-BH procedure guarantees something more: the FDR control is valid under arbitrary dependence among the e-values, whereas the BH procedure needs the ℓ_K correction under such a setting.

Proposition 9.16

- (i) With T in (9.1) and $\alpha \in (0, 1)$, the function $f : p \mapsto T(\alpha/(\ell_K p))/\alpha$ is a calibrator.
- (ii) Let $P_1, \dots, P_K$ be p-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$. The BHY procedure applied to $P_1, \dots, P_K$ is identical to the e-BH procedure to the e-variables $f(P_1), \dots, f(P_K)$.

Proof.

(i) Clearly f is nonnegative and decreasing, and it takes value 0 on $(1, \infty)$. Let U be a uniformly distributed random variable on $[0, 1]$ (under some $\mathbb{P}$). We can compute

$$\begin{aligned} \mathbb{E}^\mathbb{P}[f(U)] &= \mathbb{E}^\mathbb{P} \left[\frac{1}{\alpha} T \left(\frac{\alpha}{\ell_K U} \right) \right] \\ &= \frac{1}{\alpha} \sum_{k=1}^K \frac{K}{k} \mathbb{P} \left(\frac{\alpha(k-1)}{K\ell_K} \leq U < \frac{\alpha k}{K\ell_K} \right) = \frac{1}{\alpha} \sum_{k=1}^K \frac{\alpha}{\ell_K k} = 1. \end{aligned}$$

This shows that f is a calibrator.

(ii) As we explained in Section 9.1.3, applying the e-BH procedure to

$$\frac{1}{\alpha} T \left(\frac{\alpha}{\ell_K P_1} \right), \dots, \frac{1}{\alpha} T \left(\frac{\alpha}{\ell_K P_K} \right)$$

is the same as applying the e-BH procedure to $1/(\ell_K P_1), \dots, 1/(\ell_K P_K)$, which is equivalent to applying the BH procedure to $\ell_K P_1, \dots, \ell_K P_K$, the latter being the BHY procedure.

As a direct consequence of Proposition 9.16, the FDR guarantee of the BHY procedure directly follows from Theorem 9.11.

Remark 9.17. There is a short proof of FDR control of the BH procedure under PRDS that utilizes e-values. First, note that $\hat{\alpha} := \alpha k^*/K$ is a decreasing function of all $P_1, \dots, P_K$, where k^* is the number of discoveries of the BH procedure. Using this, the PRDS condition implies that $T = \hat{\alpha}$ and $P = P_k$ satisfy (4.4) in Remark 4.2 for each null p-value P_k , and therefore (as explained in Remark 4.2)

$$\frac{\mathbb{1}_{\{P_k \leq \hat{\alpha}\}}}{\hat{\alpha}}$$

is an e-variable. Then, for any $\mathbb{P} \in \mathcal{M}_1$, we have by the definition of BH and FDR that

$$\text{FDR} = \mathbb{E}^{\mathbb{P}} \left[\frac{\sum_{k \in \mathcal{N}} \mathbb{1}_{\{P_k \leq \hat{\alpha}\}}}{k^*} \right] = \frac{\alpha}{K} \sum_{k \in \mathcal{N}} \mathbb{E}^{\mathbb{P}} \left[\frac{\mathbb{1}_{\{P_k \leq \hat{\alpha}\}}}{\hat{\alpha}} \right] \leq \frac{\alpha |\mathcal{N}|}{K},$$

as required.

9.2 Universality of compound e-values and e-BH

In this section, we present the (perhaps surprising) fact that all FDR controlling procedures can be written as the e-BH procedure applied to some compound e-values, through two results that are simple to prove. Then, we discuss their implications in combining and derandomizing multiple testing procedures.

The first result enables the construction of compound e-values from any FDR controlling procedure.

Theorem 9.18: From FDR control to compound e-values

Let $\mathcal{D}$ be any procedure that controls the FDR at a given level $\alpha \in (0, 1)$. Let $V_k \in \{0, 1\}$ be an indicator of whether hypothesis H_k is rejected by $\mathcal{D}$, so that $F_{\mathcal{D}} = \sum_{k \in \mathcal{N}} V_k$ and $R_{\mathcal{D}} = \sum_{k \in \mathcal{K}} V_k$. Define

$$E_k = \frac{K}{\alpha} \frac{V_k}{R_{\mathcal{D}} \vee 1}. \quad (9.3)$$

Then, $E_1, \dots, E_K$ are compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$.

Proof.

The result can be seen from, for every $\mathbb{P} \in \mathcal{P}$,

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k] = \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}} \left[\frac{K}{\alpha} \frac{V_k}{R_{\mathcal{D}} \vee 1} \right] = \frac{K}{\alpha} \mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}} \vee 1} \right] \leq \frac{K}{\alpha} \alpha = K,$$

where the inequality follows from the definition of FDR guarantee.

In (9.3), the random variables $E_1, \dots, E_K$ are well-defined, noting that V_k and $R_{\mathcal{D}}$ do not depend on $\mathbb{P}$, although $F_{\mathcal{D}}$ depends on $\mathbb{P}$.

The second result shows that every FDR procedure is e-BH. In what follows, for a given set $\mathcal{P}$ of hypotheses, an FDR procedure $\mathcal{D}$ at level α is admissible if for any FDR procedure $\mathcal{D}'$ at level α such that $\mathcal{D} \subseteq \mathcal{D}'$, we have $\mathcal{D} = \mathcal{D}'$.

Theorem 9.19: Universality of e-BH with compound e-values

Let $\mathcal{D}$ be any procedure that controls the FDR at level α . Then, there exists a choice of compound e-values such that the e-BH procedure yields identical discoveries as $\mathcal{D}$. Further, if $\mathcal{D}$ is admissible, then these compound e-values can be chosen to be tight.

Proof.

The first part directly follows from Theorem 9.18. To show the last statement, consider an FDR procedure $\mathcal{D}$ at level α and take the compound e-values $E_1, \dots, E_K$ from (9.3). Define

$$K^* := \sup_{\mathbb{P} \in \mathcal{P}} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k].$$

If K^* is equal to K , then there is nothing to show as $E_1, \dots, E_K$ are tight compound e-values. Otherwise, $K^* < K$. The case $K^* = 0$ means one never rejects any hypotheses, for which choosing $E'_1 = \dots = E'_K = 1$ would suffice as tight compound e-values that produce $\mathcal{D}$ via e-BH. For $K^* > 0$, we let

$$E'_k = \frac{K}{K^*} E_k \quad \text{for } k \in \mathcal{K},$$

and apply the e-BH procedure to $E'_1, \dots, E'_K$ (which are tight compound e-values). This new procedure controls FDR at level α and produces at least as many discoveries as $\mathcal{D}$. Since $\mathcal{D}$ is admissible, this procedure must coincide with $\mathcal{D}$, and hence $\mathcal{D}$ is e-BH applied to the tight compound e-values $E'_1, \dots, E'_K$.

It is worth noting that the above result holds for every fixed $(\mathcal{P}_1, \dots, \mathcal{P}_K)$, meaning that the procedure $\mathcal{D}$ that we consider could be tuned to this particular set of hypotheses and does not need to control FDR for any other settings. If this tuned procedure $\mathcal{D}$ is admissible, it must still be recoverable via e-BH with tight compound e-values.

For the classic BH procedure in Definition 9.14, it controls FDR if we assume that the probability measures in $\mathcal{P}_1, \dots, \mathcal{P}_K$ guarantee that the p-values are independent or PRDS. It can be identified with the e-BH procedure with e-variables given in (9.3), but these e-variables are not necessarily independent or PRDS.

The connections between FDR controlling procedures, e-BH, and compound e-values laid out above motivate the following general and practical mechanism for combining discoveries across multiple testing procedures. To be concrete, in order to test the hypotheses $\mathcal{P}_1, \dots, \mathcal{P}_K$, suppose we run L different multiple testing procedures $\mathcal{D}_{\text{GeBH}}, \dots, \mathcal{D}_L$ (this can be easily generalized to the case where each procedure tests a different subset of hypotheses). Assume that the ℓ -th procedure controls the FDR at level α_ℓ . Then we may proceed as follows:

1. For the ℓ -th multiple testing procedure, form $E_1^{(\ell)}, \dots, E_K^{(\ell)}$, which are compound e-variables for $\mathcal{P}_1, \dots, \mathcal{P}_K$ (which always exist by Theorem 9.18). They could be — but need not necessarily be — the implied ones formed in Theorem 9.18.
2. Fix weights $w_1, \dots, w_L \geq 0$ with $\sum_{\ell=1}^L w_\ell = 1$. Then construct $E_1, \dots, E_K$ by convex combination as in Example 9.4, i.e., $E_k = \sum_{\ell=1}^L w_\ell E_k^{(\ell)}$ for all $k \in \mathcal{K}$. These will be compound e-values for $\mathcal{P}_1, \dots, \mathcal{P}_K$. We also allow $w_1, \dots, w_L \geq 0$ to be random, as long as they are independent of all e-values used in the procedures, and in that case it suffices to require $\sum_{\ell=1}^L \mathbb{E}^\mathbb{P}[w_\ell] \leq 1$ for any $\mathbb{P} \in \mathcal{P}$ (this condition is similar to the condition for compound e-values).
3. Apply the e-BH procedure at level α to the new compound e-values $E_1, \dots, E_K$.

The above construction is guaranteed to control the FDR at level α . Note that the value of α_ℓ does not matter, as it is only used in the construction of $E_1^{(\ell)}, \dots, E_K^{(\ell)}$, possibly implicitly (e.g., in Theorem 9.18).

One important application of the above recipe is derandomization. Suppose that $\mathcal{D}_\ell$ is a randomized multiple testing procedure, that is, it is a function of both the data X as well as a random variable U_ℓ generated during the analysis. Such randomness may not be desirable, since different random number generation seeds will lead to different sets of discoveries. In such cases, the above recipe can be used to construct a new derandomized procedure $\mathcal{D}$ that is less sensitive to $U_1, \dots, U_L$ (formally, full derandomization would occur if U_ℓ are iid and $L \rightarrow \infty$).

9.3 Log-optimal simple separable compound e-values

In this section we focus on sequence models, wherein the data X may be written as $X = (X_k : k \in \mathcal{K})$ and in principle we may test each hypothesis H_k using only X_k . We record the following definition, which we will call upon throughout this section.

Definition 9.20: Simple separable e-variables

$E_1, \dots, E_K$ are called separable if for all k , E_k is X_k -measurable, that is $E_k = E_k(X_k)$. They are called simple separable if $E_k = f(X_k)$ for some function f (which is the same for all k).

In what follows, we provide a brief summary of compound decision theory, which motivates the nomenclature “compound e-values” (Section 9.3.1) and then we provide a construction of compound e-values motivated by compound decision theory in Section 9.3.2 (see also Section 10.4.4 for another related construction).

9.3.1 Background: Robbins’ compound decision theory

Let $\mathbb{P}_\mu$ be the probability measure governing a Gaussian sequence model:

$$\mu = (\mu_1, \dots, \mu_K) \text{ fixed, } X_k \sim N(\mu_k, 1) \text{ for } k \in \mathcal{K}. \quad (9.4)$$

Suppose we are interested in constructing estimators $\hat{\mu}_k$ of μ_k such that the following expected compound loss would be small:

$$\frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^\mu [(\hat{\mu}_k - \mu_k)^2], \quad (9.5)$$

where $\mathbb{E}^\mu$ means $\mathbb{E}^{\mathbb{P}^\mu}$. If μ_k has a prior distribution B , then, by denoting $\mathbb{P}_B$ the joint distribution of μ and $\mathbf{X} = (X_1, \dots, X_K)$, the best estimator is the Bayes estimator $\hat{\mu}_k^B = \mathbb{E}^{\mathbb{P}_B}[\mu_k | \mathbf{X}] = \mathbb{E}^{\mathbb{P}_B}[\mu_k | X_k]$, which is simple separable in the sense of Definition 9.20.

Now suppose μ is deterministic as in (9.4). Given knowledge of μ , which function $s_\mu : \mathbb{R} \rightarrow \mathbb{R}$ leads to a simple separable estimator $\hat{\mu}_k^{s_\mu} = s_\mu(X_k)$ that minimizes the risk in (9.5)? The answer lies in the fundamental theorem of compound decisions, which formally connects (9.4) to the univariate Bayesian problem with prior $M = \sum_{k \in \mathcal{K}} \delta_{\mu_k} / K$ equal to the empirical distribution of μ (with δ_{μ_k} denoting the Dirac measure at μ_k):

$$\mu' \sim M, \quad X' | \mu' \sim N(\mu', 1). \quad (9.6)$$

The fundamental theorem of compound decision theory states the following. Given any fixed $s : \mathbb{R} \rightarrow \mathbb{R}$, we have that

$$\begin{aligned} \frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^\mu [(s(X_k) - \mu_k)^2] &= \sum_{k \in \mathcal{K}} \frac{1}{K} \mathbb{E}^{\mu_k} [(s(X_k) - \mu_k)^2] \\ &= \mathbb{E}^{\mathbb{P}_M} [(s(X') - \mu')^2]. \end{aligned} \quad (9.7)$$

In the left-hand side display, $\mu \in \mathbb{R}^K$ is treated as fixed (as in (9.4)), while in the right-hand side display, we only have a single random $\mu' \in \mathbb{R}$ that is randomly drawn from M as in (9.6). From (9.7) it immediately follows that the optimal simple separable estimator is the Bayes estimator under M , that is $s(x) = \mathbb{E}^{\mathbb{P}_M}[\mu' | X' = x]$.

This construction motivates the construction of feasible non-separable estimators $\hat{\mu}_k = \hat{\mu}_k(X)$ that have risk close to the optimal simple separable estimator. Thus the optimal simple separable estimator defines an oracle benchmark. Just as in classical decision theory, one often restricts attention to subclasses of estimators, e.g., equivariant or unbiased, one sets a benchmark defined by simple separable estimators. What is slightly unconventional here is a fundamental asymmetry: the oracle estimator must be simple separable but has access to the true μ , while the feasible estimator is non-separable but must work without knowledge of μ .

9.3.2 The log-optimal simple separable compound e-values

The connection of compound e-values to the fundamental theorem of compound decisions is born from the relaxation of the requirement of having a vector of e-variables to that of requiring only compound e-variables.

To further clarify the connection, we provide a construction of optimal simple separable compound e-values in sequence models with dominated null marginals via the fundamental theorem of compound decisions.

Suppose that $X = (X_k : k \in \mathcal{K})$, where X_k takes values in a space $\mathcal{Y}$. For any distribution $\mathbb{S} \in \mathcal{M}_1$, let $\mathbb{S}_k$ be the k -th marginal of $\mathbb{S}$, that is, the distribution of X_k when $X \sim \mathbb{S}$.

Suppose that for all $k \in \mathcal{K}$, the k -th null hypothesis is specified as a simple point null hypothesis on the k -th marginal, i.e.,

$$\mathcal{P}_k = \{\mathbb{S} : \mathbb{S}_k = \mathbb{P}_k\}$$

for some prespecified distribution $\mathbb{P}_k$ with $d\nu$ -density equal to p_k , where ν is a common dominating measure. We assume that there exists at least one possible true data-generating distribution $\mathbb{P}$ such that $\mathcal{N} = \mathcal{K}$ (as would be the case, e.g., if the X_k are independent and we take $\mathbb{P} = \times_{k \in \mathcal{K}} \mathbb{P}_k$).

For $k \in \mathcal{K}$, we let $\mathbb{Q}_k$ be distributions on $\mathcal{Y}$ with $d\nu$ -densities q_k , representing the alternative distributions. We seek to solve the following optimization problem:

$$\begin{aligned} & \underset{s: \mathcal{Y} \rightarrow [0, \infty]}{\text{maximize}} && \frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^{\mathbb{Q}_k} [\log(s(X_k))] \\ & \text{subject to} && s(X_1), \dots, s(X_K) \text{ are compound} \\ & && \text{e-variables for } \mathcal{P}_1, \dots, \mathcal{P}_K. \end{aligned} \tag{9.8}$$

Theorem 9.21: Log-optimal simple separable compound e-values

The optimal solution to optimization problem (9.8) is given by the likelihood ratio of mixtures over the null and alternative:

$$s(x) = \frac{\sum_{j \in \mathcal{K}} q_j(x)}{\sum_{j \in \mathcal{K}} p_j(x)}. \tag{9.9}$$

Proof.

Let $\mathbb{P}$ be any probability measure with k -th marginal given by $\mathbb{P}_k$ for all $k \in \mathcal{K}$. We will first solve the optimization problem subject to the weaker constraint that $s(X_1), \dots, s(X_K)$ are compound e-values for $\mathbb{P}_1, \dots, \mathbb{P}_K$.

Define the mixtures $\bar{\mathbb{P}} = \sum_{k \in \mathcal{K}} \mathbb{P}_k / K$ and $\bar{\mathbb{Q}} = \sum_{k \in \mathcal{K}} \mathbb{Q}_k / K$. Then, for any fixed $s : \mathcal{Y} \rightarrow [0, \infty]$, we have the following equalities:

$$\begin{aligned} \frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^{\mathbb{Q}_k} [\log(s(X_k))] &= \mathbb{E}^{\bar{\mathbb{Q}}} [\log(s(X))], \\ \frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^{\mathbb{P}_k} [s(X_k)] &= \mathbb{E}^{\bar{\mathbb{P}}} [s(X)], \end{aligned}$$

where X is a generic random variable following the distribution specified with the expectation. Thus, it suffices to solve the following optimization problem:

$$\begin{aligned} & \underset{s: \mathcal{Y} \rightarrow [0, \infty]}{\text{maximize}} && \mathbb{E}^{\bar{\mathbb{Q}}} [\log(s(X))] \\ & \text{subject to} && \mathbb{E}^{\bar{\mathbb{P}}} [s(X)] \leq 1. \end{aligned}$$

As seen in Chapter 3, this has a unique solution given by (9.9). Finally we may verify that this s indeed yields compound e-values for $\mathbb{P}_1, \dots, \mathbb{P}_K$. To this end, write $\bar{q}$ for the $d\nu$ -density of $\bar{\mathbb{Q}}$. Then, we have the following:

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}} [E_k] = \sum_{k: \mathbb{P} \in \mathcal{P}_k} \int \frac{K \bar{q}(x)}{\sum_{j \in \mathcal{K}} p_j(x)} p_k(x) \nu(dx) \leq \int K \bar{q}(x) \nu(dx) = K,$$

where the first equality holds because for any $\mathbb{P} \in \mathcal{P}_k$, its k -th marginal equals $\mathbb{P}_k$, and the inequality holds by dropping the constraint on k in the sum. This completes the proof.

Note that any tight simple separable compound e-value for testing the nulls in the above problem must be of the form (9.9) for some densities $q_1, \dots, q_K$.

Proposition 9.22

Suppose that $E_1, \dots, E_K$ are tight simple separable compound e-values (for the same nulls defined as in the previous parts of this section). Then there exist $d\nu$ -densities $q_1, \dots, q_K$ such that E_k may be represented as in (9.9) on the support of $\sum_{k \in \mathcal{K}} \mathbb{P}_k/K$ and it is without loss of generality to take $q_1 = \dots = q_K$.

Proof.

First, by definition of simple and separable, there exists a function s such that $E_k = s(X_k)$. The supremum in the definition of tightness must be attained when $\mathcal{N} = K$, and so:

$$\sum_{k \in \mathcal{K}} \mathbb{E}^{\mathbb{P}_k}[s(X_k)] = K.$$

Now define $\bar{p}$ as the $d\nu$ density of $\bar{\mathbb{P}}$, where the latter object is defined as in Theorem 9.21. Then the above may be written as:

$$\int s(x) \bar{p}(x) d\nu(x) = 1.$$

Thus, if we define $q_k(x) = s(x) \bar{p}(x)$ for all $k \in \mathcal{K}$, we find that q_k is indeed a $d\nu$ -density and $s(x)$ may be represented as in (9.9).

The objective value of the optimization problem (9.8) is given by $\text{KL}(\bar{\mathbb{Q}}, \bar{\mathbb{P}})$. By convexity, $\text{KL}(\bar{\mathbb{Q}}, \bar{\mathbb{P}}) \leq \sum_{k \in \mathcal{K}} \text{KL}(\mathbb{Q}_k, \mathbb{P}_k)$, which means that the optimal simple separable compound e-value has a worse objective than using the optimal separable e-value for each individual testing problem of $\mathbb{P}_k$ vs $\mathbb{Q}_k$ given by $E_k^* = q_k(X_k)/p_k(X_k)$.

In view of the above, why is the optimal simple separable compound e-value of interest? The reason is that if K is large (but the individual statistical problems have limited sample sizes), then it may be possible to mimic the log-optimal simple separable compound e-value without exact knowledge of the null $\mathbb{P}_k$, e.g., via empirical Bayes methods.

To elaborate, suppose we have a “nuisance parameter” $\theta \in \Theta$ that is common to all hypotheses, meaning that all null hypotheses are actually indexed by some θ : $\mathcal{P}_1^\theta, \dots, \mathcal{P}_K^\theta$, but θ is unknown. If θ were known, we could simply find an e-value separately for each hypothesis. Since θ is unknown, we cannot do that, but since θ is common across all hypotheses, we could potentially learn θ by pooling data from all hypotheses, and then form (approximate) compound e-values from the data that asymptotically match the log-optimal simple separable compound e-values.

Let us look at an explicit example that is in the spirit of the above discussion, but the resulting compound e-values are not of the above form.

9.4 A minimally adaptive e-BH procedure

We describe a minimally adaptive procedure by proposing a tiny but uniform improvement of the e-BH procedure. This improvement is negligible for large values of K and it may only be practically interesting for small K such as $K \leq 10$. We will focus on e-values and boosted e-values.

First, choose an e-merging function $F : [0, \infty)^K \rightarrow [0, \infty)$ in Chapter 8. We allow for a general choice of F other than the arithmetic mean $\mathbb{M}_K$ as it will be useful for the case of boosted e-values.

With a chosen e-merging function F and a level $\alpha \in (0, 1)$, the improved e-BH procedure, denoted by $\mathcal{D}_\alpha^F$, is designed as follows. We first test the global null $\bigcap_{k=1}^K \mathcal{P}_k$ via the rejection condition $F(e_1, \dots, e_K) \geq 1/\alpha$,

which has a type-I error of at most α , and if the global null is rejected, we then apply the e-BH procedure at level $\alpha' = K\alpha/(K-1)$. In other words,

1. if $F(e_1, \dots, e_K) < 1/\alpha$, then $\mathcal{D}_\alpha^F = \emptyset$;
2. if $F(e_1, \dots, e_K) \geq 1/\alpha$, then $\mathcal{D}_\alpha^F = \mathcal{D}_{\alpha'}$ where $\alpha' = K\alpha/(K-1)$ and $\mathcal{D}_{\alpha'}$ is the e-BH procedure at level α' .

The next proposition shows that by choosing $F = \mathbb{M}_K$, the resulting improved BH procedure dominates the base BH procedure.

Proposition 9.23

For $\alpha \in (0, 1)$, the improved e-BH procedure $\mathcal{D}_\alpha^F$ applied to arbitrary e-values has false discovery rate at most α . In case $F = \mathbb{M}_K$, $\mathcal{D}_\alpha^{\mathbb{M}_K}$ dominates the e-BH procedure $\mathcal{D}_\alpha$, that is, $\mathcal{D}_\alpha \subseteq \mathcal{D}_\alpha^{\mathbb{M}_K}$.

Proof.

Let A be the event that $F(E_1, \dots, E_K) \geq 1/\alpha$. If $K_0 < K$, then by using Theorem 9.11,

$$\begin{aligned} \mathbb{E} \mathbb{P} \left[\frac{F_{\mathcal{D}_\alpha^F}}{R_{\mathcal{D}_\alpha^F}} \right] &= \mathbb{E} \mathbb{P} \left[\frac{F_{\mathcal{D}_{\alpha'}}}{R_{\mathcal{D}_{\alpha'}}} \mathbf{1}_A \right] + \mathbb{E} \mathbb{P} \left[\frac{F_\emptyset}{R_\emptyset} (1 - \mathbf{1}_A) \right] \\ &= \mathbb{E} \mathbb{P} \left[\frac{F_{\mathcal{D}_{\alpha'}}}{R_{\mathcal{D}_{\alpha'}}} \mathbf{1}_A \right] \leq \mathbb{E} \mathbb{P} \left[\frac{F_{\mathcal{D}_{\alpha'}}}{R_{\mathcal{D}_{\alpha'}}} \right] \leq \frac{K_0}{K} \alpha' \leq \alpha. \end{aligned}$$

If $K_0 = K$, then the false discovery rate of $\mathcal{D}_\alpha^F$ is at most the probability $\mathbb{P}(A)$ of rejecting the global null via $F(E_1, \dots, E_K) \geq 1/\alpha$. In this case, $\mathbb{P}(A) \leq \alpha$ by Markov's inequality and the fact that F is an e-merging function. Hence, in either case, the FDR of $\mathcal{D}_\alpha^F$ is at most α .

To show the statement on dominance, let

$$S : (e_1, \dots, e_K) \mapsto \max_{k \in \mathcal{K}} \frac{ke_{[k]}}{K}. \quad (9.10)$$

The function S is an e-merging function and it is dominated by $\mathbb{M}_K$ because $\sum_{k=1}^K e_k \geq ke_{[k]}$. By definition of the e-BH procedure, we have $S(e_1, \dots, e_K) < 1/\alpha$ implies $\mathcal{D}_\alpha = \emptyset$. Therefore, if $\mathbb{M}_K(e_1, \dots, e_K) < 1/\alpha$, then $\mathcal{D}_\alpha = \emptyset = \mathcal{D}_\alpha^{\mathbb{M}_K}$. Moreover, since $\alpha < \alpha'$, we always have $\mathcal{D}_\alpha \subseteq \mathcal{D}_{\alpha'}$. Hence, $\mathcal{D}_\alpha \subseteq \mathcal{D}_\alpha^{\mathbb{M}_K}$.

Next, we briefly discuss the case of boosted e-values. The arithmetic average of boosted e-values is not necessarily a valid e-value, so one must be a bit more careful. Nevertheless, it turns out that we can use the function S in (9.10) on the boosted e-values in Section 9.1.3. The new procedure can be described as the following steps.

1. Boost the raw e-values with level α .
2. If $S(e'_1, \dots, e'_K) < 1/\alpha$ where $e'_1, \dots, e'_K$ are the boosted e-values in step 1, then return $\emptyset$.
3. Else: boost the raw e-values with level $\alpha' = K\alpha/(K-1)$.
4. Return the discoveries by applying the base e-BH procedure to the boosted e-values in step 3.

This new procedure dominates the e-BH procedure, and it has FDR at most α . The proof is similar to that of Proposition 9.23.

9.5 Stochastic rounding of e-values and randomized e-BH

We next discuss how to improve the e-BH procedure by utilizing randomization. The randomized Markov's inequality in Theorem 2.9 gives rise to a randomized test based on e-values, and a randomized e-to-p calibrator. Here, we show how to use external randomization to convert e-values to other (stochastically rounded) e-values.

Consider any closed set $\mathcal{G} \subseteq [0, \infty]$ (where $\mathcal{G}$ stands for “grid” since $\mathcal{G}$ will often be countable or finite below). Denote $g_* := \inf\{x : x \in \mathcal{G}\}$ and $g^* := \sup\{x : x \in \mathcal{G}\}$, and note that $g_*, g^* \in \mathcal{G}$ since $\mathcal{G}$ is closed. Further, for any $x \in [g_*, g^*]$, let

$$x^+ := \inf\{y \in \mathcal{G} : y \geq x\}, \quad x_- := \sup\{y \in \mathcal{G} : y \leq x\},$$

and note that $x^+, x_- \in \mathcal{G}$ with $x^+ \geq x$ and $x_- \leq x$, and if $x \notin \mathcal{G}$, then $x_- < x < x^+$.

Now define the stochastic rounding of $x \in [0, \infty]$ onto $\mathcal{G}$, denoted by $S_{\mathcal{G}}(x)$, as follows. If $x < g_*$, $x > g^*$, $x \in \mathcal{G}$, or $x^+ = \infty$ then define $S_{\mathcal{G}}(x) = x$. Otherwise, define

$$S_{\mathcal{G}}(x) = \begin{cases} x_- & \text{with probability } \frac{x^+ - x}{x^+ - x_-}, \\ x^+ & \text{with probability } \frac{x - x_-}{x^+ - x_-}. \end{cases}$$

Note that $S_{\mathcal{G}}(x)$ need not lie in $\mathcal{G}$, because if x lies outside the range of $\mathcal{G}$ then it is left unchanged. Also note that when $\mathcal{G} = [0, \infty]$, we have $S_{\mathcal{G}}(x) = x$ for all $x \in [0, \infty]$. In what follows, $\mathbb{P}$ denotes the joint distribution of the random variable X and the stochastic rounding $S_{\mathcal{G}}$.

One can also use a generalized version of rounding, where one stochastically rounds an input x by sampling from a mixture distribution over all values in $\mathcal{G}$ that are larger than x , instead of only x^+ , but we do not discuss this idea further here.

Proposition 9.24

For any grid $\mathcal{G}$, and any integrable random variable X , $\mathbb{E}^{\mathbb{P}}[X] = \mathbb{E}^{\mathbb{P}}[S_{\mathcal{G}}(X)]$. In particular, if X is an e-variable, then $S_{\mathcal{G}}(X)$ is also an e-variable.

The proof is simple: by design, $\mathbb{E}^{\mathbb{P}}[S_{\mathcal{G}}(x)] = x$ for any real x , and thus when applied to any random variable, it leaves the expectation unchanged.

A key property is that $S_{\mathcal{G}}(X)$ can be larger than X with (usually) positive probability, since it can get rounded up to X^+ , and is at least X_- , even when rounded down.

Remarkably, elements of the grid can also be random. For a concrete demonstration, consider the grid

$$\mathcal{G}(\hat{\alpha}) = \{0, \hat{\alpha}^{-1}, \infty\}$$

and let $S_{\hat{\alpha}}$ be shorthand for $S_{\mathcal{G}(\hat{\alpha})}$.

Proposition 9.25

If X is an e-value for $\mathbb{P}$, and $\hat{\alpha}$ taking values in $[0, 1]$ possibly depending on X , $S_{\hat{\alpha}}(X)$ is also an e-value and satisfies $\mathbb{E}^{\mathbb{P}}[S_{\hat{\alpha}}(X)] = \mathbb{E}^{\mathbb{P}}[X]$.

To prove the above result, notice that another way of writing $S_{\hat{\alpha}}$ is as follows:

$$S_{\hat{\alpha}}(X) = (X \cdot \mathbf{1}_{\{X \geq \hat{\alpha}^{-1}\}}) \vee (\hat{\alpha}^{-1} \cdot \mathbf{1}_{\{X \geq U\hat{\alpha}^{-1}\}}),$$

where U is a uniform random variable over $[0, 1]$ that is independent of X and $\hat{\alpha}$, and we treat $0/0 = 0$. Indeed, we can check that $\mathbb{E}^{\mathbb{P}}[S_{\hat{\alpha}}(X)] = \mathbb{E}^{\mathbb{P}}[X]$ by first taking expectation with respect to U (while conditioning on X) and then with respect to X .

Proof.

We rewrite $S_{\hat{\alpha}}(X)$ into the following equivalent form:

$$S_{\hat{\alpha}}(X) = X \cdot \mathbf{1}_{\{X \geq \hat{\alpha}^{-1}\}} + \underbrace{\hat{\alpha}^{-1} \cdot \mathbf{1}_{\{\hat{\alpha}^{-1} > X \geq U\hat{\alpha}^{-1}\}}}_A, \quad (9.11)$$

where $U \stackrel{d}{\sim} \mathcal{U}[0, 1]$ is independent of X and $\hat{\alpha}$. First, we derive the following equality for the expectation of A as indicated in (9.11):

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[A] &= \mathbb{E}^{\mathbb{P}}[\mathbb{E}^{\mathbb{P}}[\hat{\alpha}^{-1} \cdot \mathbf{1}_{\{\hat{\alpha}^{-1} > X \geq U\hat{\alpha}^{-1}\}} \mid X, \hat{\alpha}]] \\ &= \mathbb{E}^{\mathbb{P}}[\hat{\alpha}^{-1} \cdot \mathbf{1}_{\{X < \hat{\alpha}^{-1}\}} \mathbb{P}(U \leq \hat{\alpha}X \mid X, \hat{\alpha})] \\ &= \mathbb{E}^{\mathbb{P}}[\hat{\alpha}^{-1} \cdot \mathbf{1}_{\{X < \hat{\alpha}^{-1}\}} \cdot (\hat{\alpha}X \wedge 1)] \\ &= \mathbb{E}^{\mathbb{P}}[X \cdot \mathbf{1}_{\{X < \hat{\alpha}^{-1}\}}]. \end{aligned} \quad (9.12)$$

Then, we can simply upper bound the expectation of $S_{\hat{\alpha}}(X)$ as follows.

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[S_{\hat{\alpha}}(X)] &= \mathbb{E}^{\mathbb{P}}[X \cdot \mathbf{1}_{\{X \geq \hat{\alpha}^{-1}\}}] + \mathbb{E}^{\mathbb{P}}[X \cdot \mathbf{1}_{\{X < \hat{\alpha}^{-1}\}}] \\ &= \mathbb{E}^{\mathbb{P}}[X] \leq 1, \end{aligned} \quad (9.13)$$

where the first equality is by application of (9.11) and (9.12). Hence, we have shown our desired result. Note that if U were to be a superuniform (stochastically larger than uniform) random variable, the last equality in (9.13) would be an inequality.

We make a final important observation.

Proposition 9.26

If $X \geq \hat{\alpha}^{-1}$, then $S_{\hat{\alpha}}(X) \geq \hat{\alpha}^{-1}$.

The proof is omitted, since it is similar to that of the preceding result.

Above, we introduced stochastic rounding above for any random variable X . Next, we apply the stochastic rounding to e-values and the e-BH procedure. In this context, stochastic rounding trades off e-power for power. Indeed, it is not hard to check, by Jensen's inequality, that for an e-variable E ,

$$\mathbb{E}^{\mathbb{P}}[\log(S_{\hat{\alpha}}(E))] \leq \mathbb{E}^{\mathbb{P}}[\log E].$$

Thus, the e-power of a stochastically rounded e-value is always at most that of the original e-value, while its power (at level $\hat{\alpha}$) is at least that of the original e-value.

Fix $\alpha \in (0, 1)$, and let

$$\alpha_i = \alpha i / K, \quad \mathcal{G}_{\alpha} = \{\alpha_i^{-1} : i \in \mathcal{K}\} \cup \{0, \infty\}$$

denote the set of possible levels that e-BH may reject e-values at and in addition to 0 and ∞ . If $E \geq K/(\alpha k)$ for some $k \in \mathcal{K}$, then $S_{\mathcal{G}_{\alpha}}(E) \geq K/(\alpha k)$ as well. Thus, an e-value that is stochastically rounded to $\mathcal{G}_{\alpha}$ can only improve power when used in conjunction with e-BH.

One can view the e-BH procedure as selecting a data-dependent threshold

$$\hat{\alpha}^* := \alpha(k^* + 1)/K \quad (9.14)$$

and rejecting the i th hypothesis if and only if $E_i \geq 1/\hat{\alpha}^*$ (indeed, if $k^* = K$, the claim is trivially true, and if $k^* < K$, then there can only be k^* hypotheses with e-values larger than $K/(\alpha(k^* + 1))$, otherwise we violate the maximality of k^* in its definition). Now, observe that $1/\hat{\alpha}^*$ can only take values from the grid

Algorithm 9.28 The De-BH procedure on input e-values $(E_1, \dots, E_K)$.

Apply the e-BH procedure to $(S_{\mathcal{G}_\alpha}(E_1), \dots, S_{\mathcal{G}_\alpha}(E_K))$.

Let k^* denote the number of rejections above. Define $\hat{\alpha}^* = \alpha(k^* + 1)/K$.

Reject the i th hypothesis if $S_{\hat{\alpha}^*}(S_{\mathcal{G}_\alpha}(E_i)) \geq (\hat{\alpha}^*)^{-1}$.

$\{K/(k\alpha) : k \in \mathcal{K}\}$. If we “rounded” down each e-value E_i to the closest value in the grid that is less than E_i (or 0 if E_i is smaller than any value in the grid), the discovery set that is output by e-BH would be identical to the one where no rounding had occurred. But if we could round *up* the e-value, we would potentially gain power; however, this would inflate its expectation and it would no longer be an e-value. Our key insight is that if we *randomly* rounded every e-value up or down — appropriately so that its expectation is unchanged — then we could increase power with positive probability.

The fact that e-values are often continuous and will typically lie between grid points provides a broad opportunity to significantly increase the power. In what follows, we use independent external randomness to stochastically round e-values, as introduced previously, and increase the number of discoveries made by e-BH.

Let k^* be the number of discoveries made by e-BH applied to e-values $E_1, \dots, E_K$. The *grid-rounded e-BH (Ge-BH) procedure* simply applies the e-BH procedure to the set of e-values $\{S_{\mathcal{G}_\alpha}(E_k)\}_{k \in \mathcal{K}}$. Let $\mathcal{D}_{\text{GeBH}}$ be the set of rejections made by Ge-BH, and $\mathcal{D}_{\text{eBH}}$ be the set of rejections made by e-BH at level α .

Theorem 9.27

For any arbitrarily dependent e-values $(E_1, \dots, E_K)$, the Ge-BH procedure ensures $\text{FDR} \leq \alpha$ and $\mathcal{D}_{\text{GeBH}} \supseteq \mathcal{D}_{\text{eBH}}$. Further, for any $\mathbb{P} \in \mathcal{M}_1$, $\mathbb{P}(\mathcal{D}_{\text{GeBH}} \supsetneq \mathcal{D}_{\text{eBH}}) > 0$, i.e., the probability that Ge-BH makes extra discoveries over e-BH is positive, if and only if

$$\mathbb{P}\left(\exists k \in [K - k^*] : E_{[k^*+j]} > \frac{K}{\alpha(k^* + k + 1)}, \forall j \in [k]\right) > 0. \quad (9.15)$$

The proof follows from the fact that if E_i ever takes on a value that is between levels in $\mathcal{G}_\alpha$, the e-BH procedure will make the same rejections as if $(E_i)_-$ were substituted in its place. Stochastic rounding guarantees that $S_{\mathcal{G}_\alpha}(E_i) \geq (E_i)_-$ almost surely, so it can only increase the number of rejections. Further, when E_i is between two levels in $\mathcal{G}_\alpha$, then $S_{\mathcal{G}_\alpha}(E_i) = (E_i)_+ > E_i$ with positive probability, which leads to rejecting hypotheses that e-BH did not reject. This intuition leads us to the condition in (9.15) for which Ge-BH has strictly more power than e-BH. We omit the proof.

It is worth remarking that (9.15) is an extremely weak condition that would be very frequently satisfied. For example, if the e-values are independent and continuously distributed over $[0, K/\alpha]$ (or a larger interval), then (9.15) will hold. As an explicit example, if the data Z_i for testing the i -th hypothesis are Gaussian with variance σ^2 , and we are testing whether the mean of Z_i is nonpositive (or equal to zero) against the alternative that the mean is positive, all admissible e-values are mixtures of likelihood ratios between a positive mean Gaussian and a zero mean Gaussian: these likelihood ratio e-values take the form $\exp(\lambda Z_i - \lambda^2 \sigma^2 / 2)$ for $\lambda > 0$, which are clearly continuous and unbounded, as are many other e-values for testing parametric and nonparametric hypotheses. The independence mentioned above is far from necessary, but it is sufficient to ensure that the probability in (9.15) is not pathologically equal to zero due to some awkward worst-case dependence structure.

One can further improve the above procedure with more randomization. We now describe a more powerful procedure below, as applied to input e-values $(E_1, \dots, E_K)$. This procedure makes two iterations of rounding, first with the grid $\mathcal{G}_\alpha$, and then with the data-dependent grid $\{0, (\hat{\alpha}^*)^{-1}, \infty\}$, where $\hat{\alpha}^*$ is in (9.14). This procedure is explained in Algorithm 9.28, and it is called the *doubly-rounded e-BH (De-BH) procedure*, with $\mathcal{D}_{\text{DeBH}}$ denoting its discovery set resulting.

Finally, we mention the *uniformly-randomized e-BH (Ue-BH) procedure*: this runs the e-BH procedure on $(E_1/U, \dots, E_K/U)$, where U is an independent $U[0, 1]$ random variable. Let $\mathcal{D}_{\text{UeBH}}$ be the resulting discovery set.

Theorem 9.29

For any arbitrarily dependent e-values $(E_1, \dots, E_K)$, the De-BH and Ue-BH procedures both control FDR at level α . Further, the discoveries of De-BH is a superset of that of Ge-BH and e-BH, and the discoveries of Ue-BH is a superset of that of e-BH. Further, $\mathbb{P}(\mathcal{D}_{\text{DeBH}} \supsetneq \mathcal{D}_{\text{GeBH}}) > 0$ and $\mathbb{P}(\mathcal{D}_{\text{UeBH}} \supsetneq \mathcal{D}_{\text{eBH}}) > 0$ if and only if

$$\mathbb{P}(\exists k \in \mathcal{K} : E_k \in (0, \hat{\alpha}^{*-1})) > 0.$$

The proof is omitted, but follows from earlier observations on stochastic rounding.

As an important special case, since the BHY procedure for FDR control with arbitrarily dependent p-values (Definition 9.14) can be reinterpreted as the e-BH procedure applied to e-values obtained via a particular p-to-e calibrator, the above results also mean that the BHY procedure can be dominated by a randomized variant, which always rejects a superset of discoveries and yet controls the FDR at the same level.

A note on randomization. While the set of randomized multiple testing procedures contains all deterministic ones, it is far from clear when the most powerful randomized procedure is strictly more powerful (in at least some situations) than the most powerful deterministic one. It could be the case that randomization confers no additional benefit in any situation. As an important counterpoint, we do not know of any way for randomization to improve the power of other procedures such as the BH procedure under positive dependence, and indeed we conjecture that it does not. Both the BH and BHY procedures have their FDR upper bounds being achieved with equality in a particular setting; yet it appears as if the power of BHY can be essentially improved in almost all other situations, while we conjecture that without further knowledge, the BH procedure cannot be strictly improved in any situation without worsening its power in other situations. Thus, it is not the case that randomization should indeed always or “naturally” help in multiple testing.

Numerical experiments. In order to gauge the amount of improvement that randomization delivers, we run simulations of the e-BH procedure, where we test $K = 100$ hypotheses. Let $\pi_0 = (K - |\mathcal{N}|)/K = 0.3$, i.e., the proportion of hypotheses where the null is false. For each hypothesis, we sample $Z_i \stackrel{d}{\sim} N(\mu_i, 1)$, and perform the following one-sided test:

$$H_0 : \mu_i = 0 \text{ vs. } H_1 : \mu_i \geq 0,$$

for each $i \in \mathcal{K}$. We consider two dependence settings:

1. Positive dependence: $\text{cov}(Z_i, Z_j) = \rho^{|i-j|}$ for each $i, j \in \mathcal{K}$, i.e., the covariance matrix is Toeplitz.
2. Negative dependence: $\text{cov}(Z_i, Z_j) = -\rho/(K-1)$, i.e., the covariance is equal among all pairs of samples.

We let μ range in $[1, 4]$, and ρ range in $[0, 0.9]$. For each $i \in \mathcal{K}$, where the null hypothesis is false, we let $\mu_i = \mu$, i.e., all non-null hypotheses have a single mean. Power is defined as the expected proportion of non-null hypotheses that are discovered, i.e., $\mathbb{E}[|\mathcal{D} \setminus \mathcal{N}|/(K - |\mathcal{N}|)]$. We calculate the power averaged over 500 trials. The e-variable used is $E_i = \exp(\lambda Z_i - \lambda^2/2)$, with $\lambda = \sqrt{2 \log(1/\alpha)}$ and $\alpha = 0.05$. The results are summarized in Figure 9.30.

9.6 The closed e-BH procedure

We end this chapter by describing a procedure that (usually strictly) dominates the e-BH procedure. Due to strong connections with the classical closure principle in family-wise error rate control, we call it the *closed* e-BH procedure. This section has four goals:

- To improve the e-BH procedure, when starting with K e-values, one for each hypothesis.
- To recover any given FDR procedure for any problem as an instance of the closed e-BH procedure.

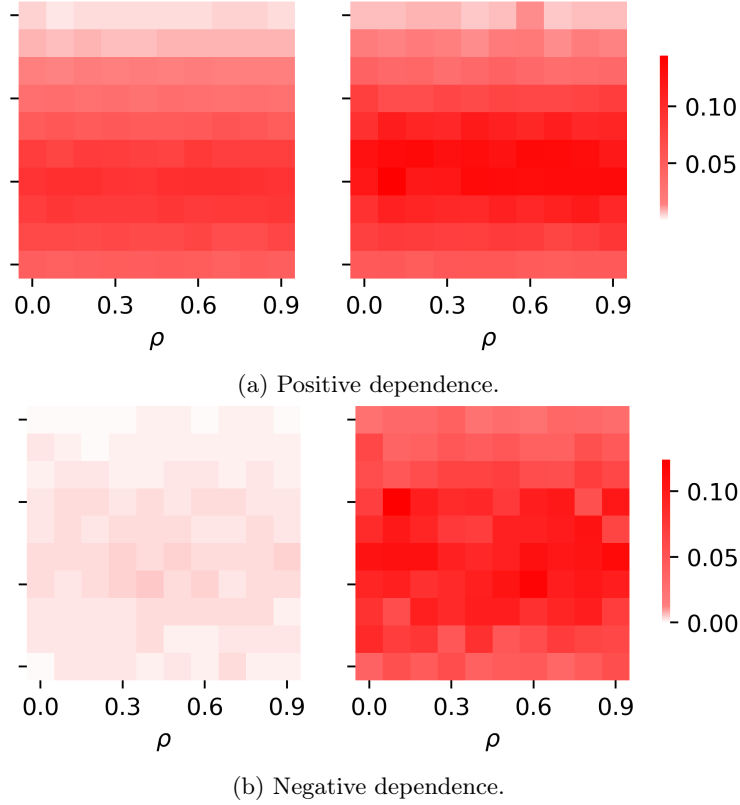


Figure 9.30: Heatmaps of the power improvement of De-BH over e-BH (left) and Ue-BH over e-BH (right), across various values of μ (y-axis) and ρ (x-axis).

- To recover and/or improve any given post-hoc FDR procedure using the closed e-BH procedure.
- To point out that the techniques developed are applicable to many error metrics beyond FDR.

First, let us show that the closed e-BH procedure is both universal and unimprovable for FDR control.

9.6.1 Improving the e-BH procedure

For any $A \subseteq \mathcal{K}$, define the intersection hypothesis $\mathcal{P}_A := \bigcap_{k \in A} \mathcal{P}_k$. The set $\{E_A\}_{A \subseteq \mathcal{K}}$ is called an *e-collection* if E_A is an e-variable for $\mathcal{P}_A$ for every $A \subseteq \mathcal{K}$ (see Section 8.5). We take $E_\emptyset = 1$.

If the e-collection is not explicitly specified, then we assume that we only have access to one e-variable E_k for each base hypothesis $\mathcal{P}_k$, $k \in \mathcal{K}$, which induces an e-collection by defining

$$E_A = \frac{1}{|A|} \sum_{k \in A} E_k, \quad A \subseteq \mathcal{K}.$$

We refer to the above as *the mean e-collection*. For any sets $A \subseteq \mathcal{K}$ and $R \subseteq \mathcal{K}$, define

$$\text{FDP}_A(R) = \frac{|A \cap R|}{|R| \vee 1},$$

which is the false discovery proportion of R if (hypothetically) the set of all true nulls were A . Using this notation, the actual FDP of a procedure that outputs R is $\text{FDP}_{\mathcal{N}}(R)$, which is unknown since $\mathcal{N}$ is unknown.

Definition 9.31: Closed e-BH procedure

Let $\{E_A\}_{A \subseteq \mathcal{K}}$ be an e-collection. Define the *candidate discovery sets* as

$$\mathcal{C}_\alpha = \left\{ R \subseteq \mathcal{K} : E_A \geq \frac{\text{FDP}_A(R)}{\alpha} \text{ for all } A \subseteq \mathcal{K} \right\}. \quad (9.16)$$

Define an *e-closed* procedure as a procedure that outputs some set $R \in \mathcal{C}_\alpha$. Finally, define the *closed e-BH procedure* ($\overline{\text{e-BH}}$) at level α as the procedure that rejects a largest set in $\mathcal{C}_\alpha$ (if there are many largest sets of equal size, we may pick any of them).

When working with the mean e-collection derived from $\mathcal{K}$ base e-values, let $R_{[1:k]}$ denote the indices corresponding to the largest k base e-values, with $R_0 = \emptyset$. One can easily check that in this case, the largest set in $\mathcal{C}_\alpha$ equals $R_{[1:\bar{k}]}$, where

$$\bar{k} = \max\{k \geq 0 : R_{[1:k]} \in \mathcal{C}_\alpha\}.$$

For a comparison to e-BH, recall that the set of all self-consistent discovery sets is given by:

$$\mathcal{C}'_\alpha = \left\{ R \subseteq \mathcal{K} : \min_{k \in R} E_k \geq \frac{K}{\alpha|R|} \right\},$$

where we insist that $\emptyset \in \mathcal{C}_\alpha, \mathcal{C}'_\alpha$ by convention, and note that the e-BH procedure rejects the largest self-consistent set $R_{[1:k^*]}$. Also, let $k^\#$ represent the number of rejections of the minimally adaptive e-BH procedure in Section 9.4.

Theorem 9.32

Any procedure that reports some $R \in \mathcal{C}_\alpha$, and in particular the closed e-BH procedure, controls the FDR at level α . Moreover, the level post-hoc guarantee holds for any $\mathbb{P} \in \mathcal{M}_1$:

$$\mathbb{E}^\mathbb{P} \left[\sup_{\alpha \in (0,1)} \sup_{R \in \mathcal{C}_\alpha} \frac{\text{FDP}_\mathcal{N}(R)}{\alpha} \right] \leq 1.$$

Further, for the mean e-collection, $\mathcal{C}'_\alpha \subseteq \mathcal{C}_\alpha$; thus, $\overline{\text{e-BH}}$ always rejects a superset of e-BH. In fact, $\overline{\text{e-BH}}$ also rejects a superset of the minimally adaptive e-BH procedure, and thus $\bar{k} \geq k^\# \geq k^*$.

Proof.

The proof is straightforward. First, fix $\mathbb{P}$ and thus $\mathcal{N} = \mathcal{N}(\mathbb{P})$, which we assume is nonempty; otherwise the proof is trivial. Choosing $A = \mathcal{N}$, we see that every set $R \in \mathcal{C}_\alpha$ satisfies $\text{FDP}_\mathcal{N}(R) \leq \alpha E_\mathcal{N}$, and thus

$$\mathbb{E}^\mathbb{P} \left[\sup_{\alpha \in (0,1)} \sup_{R \in \mathcal{C}_\alpha} \frac{\text{FDP}_\mathcal{N}(R)}{\alpha} \right] \leq \mathbb{E}^\mathbb{P}[E_\mathcal{N}] \leq 1,$$

completing the proof of the first claim.

To prove $\mathcal{C}'_\alpha \subseteq \mathcal{C}_\alpha$, note the deterministic inequality for any non-empty $A, R \subseteq \mathcal{K}$:

$$E_A = \frac{1}{|A|} \sum_{k \in A} E_k \geq \frac{|A \cap R|}{|A|} \min_{k \in A \cap R} E_k \geq \frac{|A \cap R|}{K} \min_{k \in R} E_k. \quad (9.17)$$

For any non-empty self-consistent set $R \in \mathcal{C}'_\alpha$, we have $\min_{k \in R} E_k \geq K/(\alpha|R|)$, and so the above inequality simplifies to

$$E_A \geq \frac{|A \cap R|}{\alpha|R|},$$

which implies that $R \in \mathcal{C}_\alpha$, completing the proof of the second claim.

We now argue that $R_{[1:k^\#]}$, the rejection set of minimally adaptive e-BH, is also a candidate set, that is $R_{[1:k^\#]} \in \mathcal{C}_\alpha$, so that $R_{[1:\bar{k}]} \supseteq R_{[1:k^\#]}$. We naturally consider the case where $k^\# \geq 1$, implying that $E_K \geq 1/\alpha = \text{FDP}_K(R_{[1:k^\#]})/\alpha$. For any other A such that $|A| \leq K - 1$. Altering the last step of (9.17) to take $|A| \leq K - 1$ into account yields

$$E_A \geq \frac{1}{K-1} |A \cap R_{[1:k^\#]}| \min_{k \in R_{[1:k^\#]}} E_k \geq \frac{|A \cap R_{[1:k^\#]}|}{\alpha |R_{[1:k^\#]}|} = \text{FDP}_A(R_{[1:k^\#]})/\alpha,$$

where the second inequality follows because the minimally adaptive e-BH satisfies $\min_{k \in R_{[1:k^\#]}} E_k \geq (K-1)/(|R_{[1:k^\#]}|\alpha)$. This completes the proof.

Though we do not present it here, the closed e-BH procedure can be run in $O(K^3)$ time when using the mean e-collection. For other e-collections, as with the usual closed testing procedure, the rejection set of closed e-BH may take exponential time to compute.

To better illustrate the features of the closed e-BH procedure, we take a close look at the thresholds on the individual e-values for rejection. Choose $A = \{j\}$ for each $j \in R_{[1:\bar{k}]}$, we see that (9.16) enforces the individual constraint that

$$E_j \geq \frac{1}{\alpha \bar{k}} \text{ for every } j \in R_{[1:\bar{k}]}.$$

(Of course, it imposes other “joint constraints” on averages of these e-values as well: for instance, the average of the smallest m rejected e-values must exceed $m/(\alpha \bar{k})$.) In contrast, the e-BH procedure satisfies

$$E_j \geq \frac{K}{\alpha k^*} \geq \frac{1}{\alpha}, \text{ for every } j \in R_{[1:k^*]},$$

which is a significantly harsher threshold. Meanwhile, the minimally adaptive e-BH procedure that makes $k^\# \geq k^*$ rejections satisfies

$$E_j \geq \frac{K-1}{k^\# \alpha} \text{ for every } j \in R_{[1:k^\#]},$$

which again is a much more stringent condition.

Example 9.33: Closed e-BH improves minimally adaptive e-BH

We give a simple example with $K = 3$ to show that the closed e-BH rule can reject strictly more than the minimally adaptive e-BH rule. For the set of input e-values (60, 39, 11), the e-BH procedure only makes one rejection at $\alpha = 0.05$, narrowly missing out on the second rejection. The minimally adaptive e-BH procedure effectively runs e-BH at level $3\alpha/2$, and makes two rejections. The closed e-BH procedure rejects all 3 hypotheses, which can be checked by verifying that the set $\{1, 2, 3\}$ lies in $\mathcal{C}$ in (9.16), since each e-value is larger than $20/3$, all pairwise averages exceed $40/3$ and the overall average exceeds 20.

Why is it even possible to reject e-values below $1/\alpha$, indeed as small as $1/(3\alpha)$ in the above example, while still controlling FDR at α ? The informal explanation is as follows. Assume for the moment that the first two e-values are so large that it is almost without doubt that they are non-nulls, while the third hypothesis is indeed null. Then, the FDP can only be 0 or $1/3$; it is $1/3$ only when the third e-value exceeds $1/(3\alpha)$, which happens with probability at most 3α by Markov’s inequality. Thus, the FDR is still bounded by α in this case. This intuition may be helpful to understand the more abstract proof of Theorem 9.32.

Our results also have implications for the BHY procedure in Definition 9.14. Recall from Proposition 9.16 that the BHY procedure is an instance of the e-BH procedure with e-values formed via calibrating the input p-values using a particular p-to-e calibrator. Define the *closed BHY* procedure as the one obtained by using the closed e-BH procedure with the mean e-collection formed by the aforementioned e-values. Both procedures output a set of the form $R_{[1:k]}$.

Corollary 9.34: Improving the BH procedure under arbitrary dependence

The discoveries made by the closed BHY procedure are a (usually strict) superset of those made by the BHY procedure, while still controlling the FDR at level α . Thus, the BHY procedure is not admissible in the sense that it can be improved while maintaining the same FDR control.

9.6.2 Recovering any FDR procedure

This subsection will show that every FDR controlling procedure can be written as the closed e-BH procedure applied to an appropriate e-collection.

Theorem 9.35

Consider an arbitrary FDR controlling procedure $\mathcal{D}$. Then, for each α , there exists an e-collection $\{E_A\}_{A \subseteq \mathcal{K}}$ for which the corresponding closed e-BH procedure recovers $\mathcal{D}$ at level α .

Proof.

Suppose that running $\mathcal{D}$ at level α on some dataset produces a rejection set $R = R(\alpha)$. Now, define the e-collection $\{E_A\}_{A \subseteq \mathcal{K}}$ given by

$$E_A = \text{FDP}_A(R)/\alpha. \quad (9.18)$$

First note that each E_A as defined above is an e-value for $\mathcal{P}_A$ because $\mathcal{D}$ is an FDR controlling procedure. Defining the candidates $\mathcal{C}_\alpha$ by (9.16), we observe that $R \in \mathcal{C}_\alpha$ by construction. Indeed, one can check that $\mathcal{C}_\alpha = \{\emptyset, R\}$: no other nonempty set $R' \neq R$ can lie in $\mathcal{C}_\alpha$ because the inequality $E_A \geq \text{FDP}_A(R')/\alpha$ cannot hold for all A , noting that $|A \cap R'|/|R'| > |A \cap R|/|R|$ for $A = R'$ when $R' \subseteq R$ and $A = R' \setminus R$ otherwise. Thus the only set in $\mathcal{C}_\alpha$ is R itself, and the closed e-BH procedure rejects it.

The e-collection in the above proof is worth further examination. If $R = \emptyset$, then all e-values equal zero, so we will assume for simplicity that R is nonempty. For every $j \in R$, we have $E_j = 1/(\alpha|R|)$, and for every $j \notin R$, we have $E_j = 0$. Meanwhile, $E_{\mathcal{K}} = E_R = 1/\alpha$ and $E_{\mathcal{K} \setminus R} = 0$. Consider the directed acyclic graph formed by $2^{\mathcal{K}} - 1$ nodes, indexed by nonempty sets $A \subseteq \mathcal{K}$, where the K leaves are the singletons, the root is the set $\mathcal{K}$, and a node A is a parent of node B if $A \supsetneq B$ and $|A \setminus B| = 1$. Then these e-values are increasing from leaves to the root. Adding nonrejected elements to any A does not change its assigned e-value, but adding rejected elements to A increases its e-value (from $1/(|R|\alpha)$ in the leaf to $1/\alpha$ in the root).

We end by noting that there is no contradiction that every FDR procedure can be written both as the e-BH procedure with compound e-values and as the closed e-BH procedure with an e-collection.

In fact, we can also write down a version of the closed e-BH procedure that works with compound e-values. Indeed, given an input vector $\tilde{\mathbf{E}}$ of compound e-values, we can define an e-collection using $E_A = K^{-1} \sum_{k \in A} \tilde{E}_k$. Then, one can show that the corresponding closed e-BH procedure makes more discoveries than the e-BH procedure, using a proof nearly identical to that of Theorem 9.32. We omit the details.

9.6.3 Recovering or improving any post-hoc FDR procedure

Recall that a post-hoc FDR procedure $\mathcal{D}$ is one that produces a family of rejection sets $\{R_\alpha\}_{\alpha \in (0,1)}$ that satisfies

$$\sup_{\mathbb{P} \in \mathcal{M}_1} \mathbb{E}^{\mathbb{P}} \left[\sup_{\alpha \in (0,1)} \frac{\text{FDP}_{\mathcal{N}}(R_\alpha)}{\alpha} \right] \leq 1.$$

(By default, take $R_0 = \emptyset$ and $R_1 = \mathcal{K}$ without loss of generality.) We now claim that we can recover any such method using the closed e-BH procedure.

Theorem 9.36

Consider an arbitrary post-hoc FDR procedure $\mathcal{D}$. Then, there exists an e-collection $\{E_A\}_{A \subseteq \mathcal{K}}$ for which the corresponding closed e-BH procedure (at any α) recovers or improves $\mathcal{D}$ (at that α).

Proof.

For each $A \subseteq \mathcal{K}$, define the e-value

$$\bar{E}_A = \sup_{\beta \in (0,1]} \frac{\text{FDP}_A(R_\beta)}{\beta}. \quad (9.19)$$

Then $(\bar{E}_A)_{A \subseteq \mathcal{K}}$ form an e-collection because $\mathcal{D}$ is a post-hoc FDR procedure.

At any (potentially data-dependent) $\alpha \in (0,1)$, we form $\mathcal{C}_\alpha$ using (9.16). Observe that $R_\alpha \in \mathcal{C}_\alpha$ by construction. Thus, the closed e-BH procedure either rejects R_α or a larger set.

Let us briefly compare the construction of the e-collection in (9.19) to (9.18) in the preceding subsection. First, in the former and unlike the latter, the e-collection does not depend on α (E_A depends on α , but $\bar{E}_A$ does not). Second, if the e-BH procedure (that takes $\mathcal{K}$ base e-values as input) is plugged into Theorem 9.35, we get back the e-BH procedure itself (as Theorem 9.35 claims). Every e-value in the above e-collection $\{\bar{E}_A\}$ is at least as large as the preceding e-collection $\{E_A\}$, typically strictly larger, and so the above construction will yield a procedure strictly improving e-BH, at least at some level α . Note also that the previous construction sets $\bar{E}_j = 0$ every $j \notin R_{k^*}$. However, the above construction has no e-values equal to 0 because every index is eventually included in R_β for some β .

9.6.4 A general e-closure principle for arbitrary error metrics

What we have implicitly developed in this section is a much more far-reaching “e-closure” principle that applies to any error metric, not just the FDR. To elaborate, let $L_A(R)$ represent an arbitrary error metric that quantifies the quality of the selected set R , when the true set of null hypotheses happens to equal A . Suppose we would like to design procedures $\mathcal{D}$ whose rejected sets R satisfy

$$\sup_{\mathbb{P} \in \mathcal{M}_1} \mathbb{E}^\mathbb{P}[L_{\mathcal{N}}(R)] \leq \alpha, \quad (9.20)$$

for any α in some range; usually the losses L are nonnegative and bounded, and so the range can often be taken to be $[0,1]$, as we will do below for simplicity. In the case of FDR control, we had $L_A = \text{FDP}_A$. But we could consider alternatives such as

- k -familywise error rate (k -FWER): $L_A(R) = \mathbf{1}_{\{|A \cap R| \geq k\}}$,
- per-family error rate (PFER): $L_A(R) = |A \cap R|/K$ (the division by K is to rescale the range to $[0,1]$).
- false discovery exceedence (FDX): $\mathbf{1}_{\{\text{FDP}_A(R) > \gamma\}}$ for some $\gamma \in (0,1)$.

Let $\mathcal{L}$ denote the set of all such loss functions.

Analogous to what was done in this section, given an e-collection $\{E_A\}_{A \subseteq \mathcal{K}}$, we could define the candidate set for a particular L as

$$\mathcal{C}_\alpha(L) = \left\{ R \subseteq \mathcal{K} : E_A \geq \frac{L_A(R)}{\alpha} \text{ for all } A \subseteq \mathcal{K} \right\}.$$

Then, any set $R \in \mathcal{C}_\alpha(L)$ satisfies (9.20), and we could, in particular, choose the largest such set to reject. For example, for FWER control (i.e. k -FWER control with $k = 1$), e-closure with the mean e-collection recovers the e-Holm procedure, whose details are omitted here for brevity.

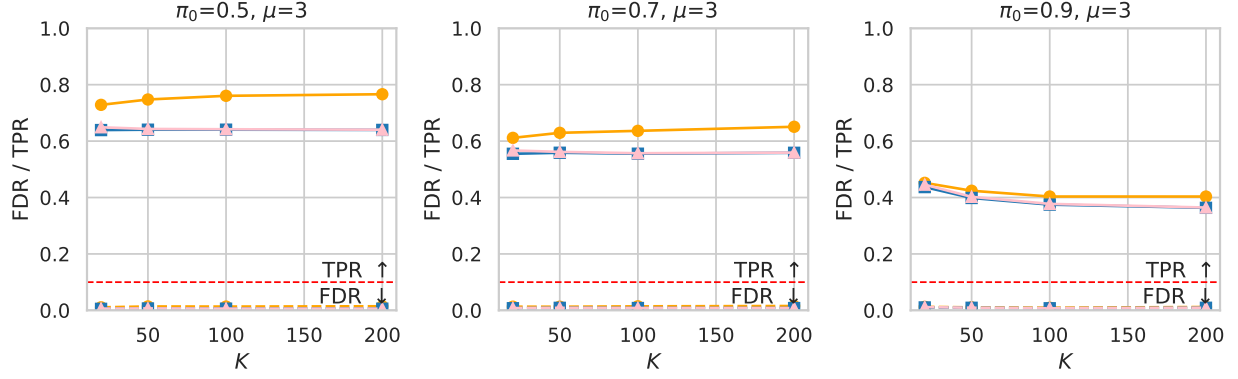


Figure 9.37: Plot of FDR and TPR for different null proportions π_0 and alternative mean $\mu = 3$. All differences are significant; the closed e-BH (orange circle, upper curve in all plots) dominates both e-BH and minimally adaptive e-BH (eBHm; almost identical lower curves in all plots).

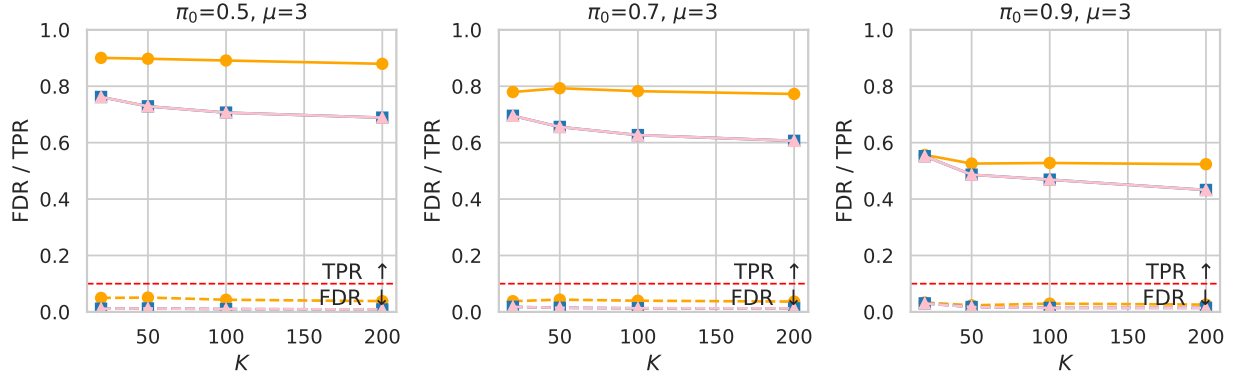


Figure 9.38: Plot of FDR and TPR for different null proportions π_0 and alternative mean $\mu = 3$. All differences are significant; the closed BHY procedure (orange circle, upper curve in all plots) dominates the BHY procedure and its minimally adaptive variant (almost identical lower curves in all plots).

Further, if we begin with any procedure $\mathcal{D}$ that guarantees (9.20), it can be recovered or improved by applying our e-closure technique on the e-collection defined by $E_A = L_A(R)/\alpha$. In fact, we would get the stronger claim that

$$\sup_{\mathbb{P} \in \mathcal{M}_1} \mathbb{E}^{\mathbb{P}} \left[\sup_{L \in \mathcal{L}} \sup_{\alpha \in [0,1]} \sup_{R \in \mathcal{C}_\alpha(L)} \frac{L_{\mathcal{N}}(R)}{\alpha} \right] \leq 1.$$

In particular, it means that both the loss function L and the level α can be chosen post-hoc. For example, if the scientist does not like the rejection set achieved with FWER control at level 0.01, they can report the results when requiring FDR control at level 0.05 (and indeed, can do through all such loss functions and error levels before deciding what to report).

Further developments of e-closure are omitted here.

9.6.5 Numerical experiments

We perform some simple simulations for testing one-sided Gaussian means. Let $X_k \sim N(\mu_k, 1)$, $k \in \mathcal{K}$, and let the k -th null hypothesis be given by $H_k : \mu_k \leq 0$ against the alternative that $\mu_k > 0$. For our first simulation, we let the data be independent across k . Let $\pi_0 := |\mathcal{N}|/K$ be the null proportion, μ be the signal

of all alternatives; so let $X_k \sim N(0, 1)$ for $k \in \mathcal{N}$ and $X_k \sim N(\mu, 1)$ otherwise. Define $E_k := \exp(\lambda X_k - \lambda^2/2)$ for $\lambda = \mu = 3$, which is the log-optimal choice for H_k . We let $\alpha = 0.1$ in our simulations. We plot the empirical average FDR and true positive rate (TPR) $:= \mathbb{E}[|\mathcal{N}^c \cap R|/|\mathcal{N}^c|]$ of e-BH, minimally adaptive e-BH, and closed e-BH over 1000 trials. We plot the results in Figure 9.37, and can see that the closed e-BH procedure improves noticeably over the standard e-BH and the minimally adaptive e-BH procedures across all settings, while controlling the FDR below α .

We also evaluate the BHY procedure, the application of minimally adaptive BHY procedure (which applies minimally adaptive e-BH to the calibrated p-values from Proposition 9.16), and the closed BHY procedure from Corollary 9.34, where we set $P_i := 1 - \Phi(X_i)$. Here, we draw $(X_1, \dots, X_K)$ from a multivariate normal with the same component means as described above and with Toeplitz-like covariance matrix where $|\text{Cov}(X_i, X_j)| = \exp(-|i - j|/10)/5$ with the covariance being positive if $|i - j|$ is even and negative otherwise. Similarly, our results in Figure 9.38 show that our closed BHY procedure has substantial power increase over the original BHY procedure.

Comparing Figures 9.37 and 9.38 also presents a frequently observed empirical phenomenon: the e-value-based procedures tend to be less powerful than p-value-based procedures. But this observation comes with three caveats:

- e-value-based procedures offer a much stronger type of post-hoc type-I error control;
- as the BHY example shows, it is sometimes possible to improve p-value-based procedures by employing e-values;
- the empirical observations depend heavily on which e-values are used, especially since it is possible to recover all p-value-based procedures using e-value-based procedures.

We remind the reader of the more detailed discussion in Section 1.2.2.

Bibliographical notes

Benjamini and Hochberg [1995] introduced the FDR error metric and developed the BH procedure for FDR control for independent p-values. Benjamini and Yekutieli [2001] extended the result to p-values under the PRDS condition. That paper also showed that under arbitrary dependence, the FDR control achieved by the BHY procedure equals $\alpha \ell_K$ (the Benjamini–Yekutieli correction), a fact mentioned in Section 9.1.4.

The e-BH procedure was proposed by Wang and Ramdas [2022], with some simple codes for the e-BH procedure available at

<https://github.com/ruoduwang/e-BH>.

The proof of Theorem 9.13 is in Wang and Ramdas [2022], where other boosting methods for e-values than the one in Section 9.1.3 are also studied, including the boosting methods under the PRDS condition.

Compound e-values were first used (without being given a name) in Wang and Ramdas [2022, Theorem 3]. The term compound e-values first appeared in Ignatiadis et al. [2024b]. Other authors have used the term generalized e-values [Banerjee et al., 2023, Bashari et al., 2023, Zhao and Sun, 2025, Lee and Ren, 2024] or relaxed e-values [Ren and Barber, 2024, Gablenz and Sabatti, 2025]. A unified treatment of compound e-values was provided in Ignatiadis et al. [2024a], from which this chapter draws many results. The term “self-consistent procedures” originated from the literature on multiple testing procedures based on p-values such as Blanchard and Roquain [2008], who also developed the proof technique in Remark 9.17 (despite not using the terminology of e-values).

The term compound e-values pays tribute to Robbins’ compound decision theory [Robbins, 1951] in which multiple statistical problems are connected through a (compound) loss function that averages over individual losses. We refer the reader to e.g., Copas [1969], Zhang [2003], Jiang and Zhang [2009] for more comprehensive accounts and present a brief overview here. As a very concrete example (which was studied in detail already in Robbins [1951]), consider the Gaussian sequence model. Empirical Bayes [Robbins, 1956] ideas work

well for this task, for example, [Jiang and Zhang \[2009\]](#) provide sharp guarantees for the performance of a nonparametric empirical Bayes method in mimicking the best simple separable estimator in the Gaussian sequence model in (9.4) as $K \rightarrow \infty$.

The combination and derandomization recipe was used by [Ren and Barber \[2024\]](#) to derandomize the model-X knockoff filter [[Candès et al., 2018](#)], a flexible set of methods for variable selection in regression with finite-sample FDR control that previously relied on additional randomness. Some further applications include the following: [Banerjee et al. \[2023\]](#) use the same recipe for meta-analysis in which the ℓ -th study only reports the set of tested hypotheses, the set of discoveries, as well as the targeted FDR level. [Li and Zhang \[2024\]](#) also discuss designing multiple testing procedures in various contexts by aggregating e-values from different procedures. The current chapter unifies and extends several of these observations.

The minimally adaptive e-BH procedure in Section 9.4 was proposed by [Ignatiadis et al. \[2024b\]](#), and it was inspired by a similar improvement of the BH procedure in [Solari and Goeman \[2017\]](#). The stochastic rounding of e-values in Section 9.5 was developed in [Xu and Ramdas \[2023\]](#). The intersection of closure and e-values has been tackled in many papers including [Vovk and Wang \[2021, 2023\]](#), [Fischer and Ramdas \[2024\]](#), [Hartog and Lei \[2025\]](#), and the closed e-BH procedure in Section 9.6 is based on [Xu et al. \[2025\]](#) and [Goeman et al. \[2025\]](#). Online multiple testing with e-values was studied in [Xu and Ramdas \[2024\]](#). Recently, [Fischer et al. \[2024\]](#) proposed an online analog of the BH and e-BH procedures.

The code to reproduce the randomized e-BH figures is at

`https://github.com/neilzxu/rand-ebh,`

and the code to reproduce the closed e-BH figures is at

`https://github.com/neilzxu/closed-ebh.`

Part III

Advanced Topics

Chapter 10

Approximate and asymptotic e-values

We have been studying e-values and p-values that are valid under the null. In many statistical contexts, one may have approximate or asymptotic p-values and e-values, that satisfy relaxed conditions, and offer relaxed statistical guarantees. This chapter formalizes the notions of approximate and asymptotic e-values, compound e-values, and asymptotic compound e-values, show their versatility and usefulness for multiple testing, as well as provide certain optimal constructions and examples.

10.1 Approximate p-values and e-values

We first define approximate p-variables and e-variables. The basic idea is that, for these approximate objects, the conditions for p-variables and e-variables in Definition 1.2 should be satisfied approximately. It turns out that two different forms of approximation error often exist in statistical applications:

1. One, represented by ε below, measures how much the value of the random variable deviates from a p-variable or e-variable. For instance, the random variable $(1 + \varepsilon)P$ for $P \stackrel{d}{\sim} U(0, 1)$ can be seen as “ ε -away” from a p-variable.
2. The other one, represented by δ below measures with how much probability the random variable deviates from a p-variable or e-variable. For instance, the random variable $P\mathbb{1}_A$ for $P \stackrel{d}{\sim} U[0, 1]$ and an event A with probability $\mathbb{P}(A) = 1 - \delta$ can be seen as “ δ -away” from a p-variable.

This consideration leads to the following formal definition of approximate p-variables and e-variables, and their connections to the above two forms of approximation error is justified in Proposition 10.2.

Definition 10.1: Approximate p-variables and e-variables

Fix the null hypothesis $\mathcal{P}$ and two functions $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$, and write $\varepsilon_{\mathbb{P}} = \varepsilon(\mathbb{P})$ and $\delta_{\mathbb{P}} = \delta(\mathbb{P})$ for $\mathbb{P} \in \mathcal{P}$.

- (i) An $\mathbb{R}_+$ -valued random variable P is an (ε, δ) -*approximate* p-variable for $\mathcal{P}$ if

$$\mathbb{P}(P \leq t) \leq (1 + \varepsilon_{\mathbb{P}})t + \delta_{\mathbb{P}} \quad \text{for all } t \in (0, 1) \text{ and all } \mathbb{P} \in \mathcal{P}.$$

- (ii) A $[0, \infty]$ -valued random variable E is an (ε, δ) -*approximate* e-variable for $\mathcal{P}$ if

$$\mathbb{E}^{\mathbb{P}}[E \wedge t] \leq 1 + \varepsilon_{\mathbb{P}} + \delta_{\mathbb{P}}t \quad \text{for all } t \in \mathbb{R}_+ \text{ and all } \mathbb{P} \in \mathcal{P}.$$

For two constants $\varepsilon \in \mathbb{R}_+$ and $\delta \in [0, 1]$, (ε, δ) -approximation is understood as those in Definition 10.1 with the corresponding constant functions. Similarly to “e-values” and “p-values”, we also speak of “approximate e-values” and “approximate p-values” when we do not emphasize them as random variables.

We make some quick observations. Clearly, $(0, 0)$ -approximate p-variables (e-variables) are just p-variables (e-variables). For a constant $\varepsilon \in \mathbb{R}_+$, P is an $(\varepsilon, 0)$ -approximate p-variable if and only if $(1 + \varepsilon)P$ is a p-variable, and E is an $(\varepsilon, 0)$ -approximate e-variable if and only if $E/(1 + \varepsilon)$ is an e-variable. The δ -parameter is more lenient than the ε -parameter; for instance, (ε, δ) -approximation implies $(0, \varepsilon + \delta)$ -approximation in both classes (see Proposition 10.5 below for a sharper result). If $\delta = 1$, then (ε, δ) -approximation puts no requirements on P or E .

As mentioned above, the definitions of approximate p-variables and approximate e-variables can be interpreted as allowing for a wiggle room in the definitions of p-variables and e-variables in two ways: (a) the probability and expectation requirements have a multiplicative factor $1 + \varepsilon_{\mathbb{P}}$; (b) the requirements are only required to hold on an event of probability $1 - \delta_{\mathbb{P}}$. This will be made transparent through the following proposition.

Proposition 10.2

Let $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$ and P and E be nonnegative random variables ($[0, \infty]$ -valued for E). For the following statements,

(P1) P is an (ε, δ) -approximate p-variable for $\mathcal{P}$;

(P2) for every $\mathbb{P} \in \mathcal{P}$, there exists an event A such that

$$\mathbb{P}(P \leq t, A) \leq (1 + \varepsilon_{\mathbb{P}})t \text{ for all } t \in (0, 1) \quad \text{and} \quad \mathbb{P}(A) \geq 1 - \delta_{\mathbb{P}};$$

(E1) E is an (ε, δ) -approximate e-variable for $\mathcal{P}$;

(E2) for every $\mathbb{P} \in \mathcal{P}$, there exists an event A such that

$$\mathbb{E}^{\mathbb{P}}[E\mathbf{1}_A] \leq 1 + \varepsilon_{\mathbb{P}} \quad \text{and} \quad \mathbb{P}(A) \geq 1 - \delta_{\mathbb{P}};$$

we have that (P2) implies (P1) and (E2) implies (E1). Moreover, if each $\mathbb{P} \in \mathcal{P}$ is atomless, then (P1)–(P2) are equivalent, and (E1)–(E2) are equivalent.

Proof.

When $\mathbb{P}$ is atomless, for any random variable X , there exists an event A with $\mathbb{P}(A) = 1 - \delta_{\mathbb{P}}$ such that $\{X < t\} \subseteq A \subseteq \{X \leq t\}$ for some $t \in \mathbb{R}$; a formal statement is presented in Lemma A.6. This fact will be used in the proof below. If $\delta_{\mathbb{P}} = 1$ there is nothing to prove, so we assume $\delta_{\mathbb{P}} < 1$ below.

(P2) $\Rightarrow$ (P1): It suffices to notice

$$\mathbb{P}(P \leq t) \leq \mathbb{P}(P \leq t, A) + \mathbb{P}(A^c) \leq (1 + \varepsilon_{\mathbb{P}})t + \delta_{\mathbb{P}}$$

for $\mathbb{P} \in \mathcal{P}$ and $t \in [0, 1]$.

(P1) $\Rightarrow$ (P2) in case each $\mathbb{P}$ is atomless: Let A be an event with $\mathbb{P}(A) = 1 - \delta_{\mathbb{P}}$ such that $\{P > s\} \subseteq A \subseteq \{P \geq s\}$ for some $s \in [0, 1]$. Note that $\mathbb{P}(P \leq t, A^c) = \min\{\mathbb{P}(P \leq t), \mathbb{P}(A^c)\}$. Therefore,

$$\mathbb{P}(P \leq t, A) = (\mathbb{P}(P \leq t) - \mathbb{P}(A^c))_+ \leq ((1 + \varepsilon_{\mathbb{P}})t + \delta_{\mathbb{P}} - \delta_{\mathbb{P}})_+ = (1 + \varepsilon_{\mathbb{P}})t$$

for all $t \in (0, 1)$, as desired.

(E2) $\Rightarrow$ (E1): It suffices to notice

$$\begin{aligned} \mathbb{E}^{\mathbb{P}}[E \wedge t] &= \mathbb{E}^{\mathbb{P}}[(E \wedge t)\mathbf{1}_A] + \mathbb{E}^{\mathbb{P}}[(E \wedge t)\mathbf{1}_{A^c}] \leq \mathbb{E}^{\mathbb{P}}[E\mathbf{1}_A] + \mathbb{E}^{\mathbb{P}}[t\mathbf{1}_{A^c}] \\ &\leq (1 + \varepsilon_{\mathbb{P}}) + \delta_{\mathbb{P}}t \end{aligned}$$

for $\mathbb{P} \in \mathcal{P}$ and $t \in \mathbb{R}_+$.

(E1) $\Rightarrow$ (E2) in case each $\mathbb{P}$ is atomless: Let A be an event with $\mathbb{P}(A) = 1 - \delta_{\mathbb{P}}$ such that $\{E < t\} \subseteq A \subseteq \{E \leq t\}$ for some $t \in \mathbb{R}_+$. Note that $E \geq t$ on A^c . We have,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[E\mathbf{1}_A] &= \mathbb{E}^{\mathbb{P}}[(E \wedge t)\mathbf{1}_A] = \mathbb{E}^{\mathbb{P}}[E \wedge t] - \mathbb{E}^{\mathbb{P}}[(E \wedge t)\mathbf{1}_{A^c}] \\ &\leq 1 + \varepsilon_{\mathbb{P}} + t\delta_{\mathbb{P}} - \mathbb{E}^{\mathbb{P}}[t\mathbf{1}_{A^c}] = 1 + \varepsilon_{\mathbb{P}},\end{aligned}$$

as desired.

Without the atomless assumption on each $\mathbb{P} \in \mathcal{P}$, the implications (P1) $\Rightarrow$ (P2) and (E1) $\Rightarrow$ (E2) in Proposition 10.2 do not hold, because the event A with desired probability may not exist (counterexamples can be found on discrete spaces). We emphasize that the atomless assumption does not pertain to the distribution of P or E (i.e., P or E can be discrete). For instance, this atomless assumption is automatic if external randomization is permitted.

The two formulations (P1) and (P2) in the above proposition are obviously equivalent when $\delta = 0$ without requiring any condition; the same holds for (E1) and (E2).

In Chapter 2, we described calibrators from p-values to e-values and those from e-values to p-values. Recall from Chapter 2 that a (p-to-e) calibrator is a decreasing function $f : [0, \infty) \rightarrow [0, \infty]$ satisfying $\int_0^1 f(x)dx \leq 1$ and $f(x) = 0$ for $x \in (1, \infty)$. An e-to-p calibrator is a decreasing function $g : [0, \infty] \rightarrow [0, \infty)$ such that $g(x) \geq (1/x) \wedge 1$ for all $x \in [0, \infty]$, because the only admissible e-to-p calibrator is $x \mapsto (1/x) \wedge 1$. The next theorem demonstrates that we can calibrate approximate p-values and approximate e-values in exactly the same way as in calibrating p-values and e-values.

Theorem 10.3

Let $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$. Using any calibrators, calibrating an (ε, δ) -approximate p-variable yields an (ε, δ) -approximate e-variable, and vice versa.

Proof.

Fix any $\mathbb{P} \in \mathcal{P}$. If $\delta_{\mathbb{P}} = 1$ there is nothing to prove, so we assume $\delta_{\mathbb{P}} < 1$ below. For the first statement, let P be an (ε, δ) -approximate p-variable for $\mathbb{P}$ and f be a (p-to-e) calibrator and $b = (1 + \varepsilon_{\mathbb{P}})/(1 - \delta_{\mathbb{P}}) \geq 1$. Let F be a cdf on $[0, 1/b]$ given by $F(s) = (1 + \varepsilon_{\mathbb{P}})s + \delta_{\mathbb{P}}$ for $s \in [0, 1/b]$. We have $\mathbb{P}(P \leq s) \leq F(s)$ for $s \in [0, 1]$ and $F(0) = \delta_{\mathbb{P}}$. Hence, noting that the calibrator f is decreasing, we have, for all $t \in \mathbb{R}_+$,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[f(P) \wedge t] &\leq \int_0^{1/b} (f(s) \wedge t) dF(s) \\ &\leq (1 + \varepsilon_{\mathbb{P}}) \int_0^{1/b} f(s) ds + tF(0) \\ &\leq (1 + \varepsilon_{\mathbb{P}}) \int_0^1 f(s) ds + \delta_{\mathbb{P}}t \leq 1 + \varepsilon_{\mathbb{P}} + \delta_{\mathbb{P}}t.\end{aligned}$$

For the second statement, it suffices to consider the calibrator $x \mapsto (1/x) \wedge 1$, and we define $P = (1/E) \wedge 1$. We have, for $t \in (0, 1)$,

$$\begin{aligned}\mathbb{P}(P \leq t) &= \mathbb{P}(E \geq 1/t) \\ &= \mathbb{P}(E \wedge (1/t) \geq 1/t) \\ &\leq t\mathbb{E}^{\mathbb{P}}[E \wedge (1/t)] \leq t \left(1 + \varepsilon_{\mathbb{P}} + \frac{\delta_{\mathbb{P}}}{t} \right) = (1 + \varepsilon_{\mathbb{P}})t + \delta_{\mathbb{P}},\end{aligned}$$

as desired.

Remark 10.4. Following a similar proof, the statements in Theorem 10.3 hold true also if we use the alternative formulations (P2) and (E2) in Proposition 10.2 for both approximate p-variables and approximate e-variables.

The roles of ε and δ are asymmetric and, as mentioned earlier, the wiggle room created by δ is more lenient. The following result shows that an (ε, δ) -approximate p-variable (or e-variable) with $\varepsilon > 0$ is always also an $(0, \delta')$ -approximate p-variable (or e-variable) where δ' can be chosen strictly smaller than $\varepsilon + \delta$.

Proposition 10.5

Let $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$, $\delta : \mathcal{P} \rightarrow [0, 1]$, $\delta' = (\varepsilon + \delta)/(1 + \varepsilon)$.

- (i) An (ε, δ) -approximate p-variable for $\mathcal{P}$ is also an $(0, \delta')$ -approximate p-variable for $\mathcal{P}$.
- (ii) An (ε, δ) -approximate e-variable for $\mathcal{P}$ is also an $(0, \delta')$ -approximate e-variable for $\mathcal{P}$.

Proof.

Let P be an (ε, δ) -approximate p-variable. Fix $\mathbb{P} \in \mathcal{P}$. Then, for any $t \in (0, (1 - \delta_{\mathbb{P}})/(1 + \varepsilon_{\mathbb{P}}))$, we have that:

$$\mathbb{P}(P \leq t) \leq (1 + \varepsilon_{\mathbb{P}})t + \delta \leq t + \varepsilon_{\mathbb{P}} \frac{1 - \delta_{\mathbb{P}}}{1 + \varepsilon_{\mathbb{P}}} + \delta_{\mathbb{P}} = t + \delta'_{\mathbb{P}}.$$

Meanwhile the inequality $\mathbb{P}(P \leq t) \leq t + \delta'_{\mathbb{P}}$ is clearly true for $t \geq (1 - \delta_{\mathbb{P}})/(1 + \varepsilon_{\mathbb{P}})$ (since the right-hand side is ≥ 1). Thus P is an $(0, \delta')$ -approximate p-variable for $\mathcal{P}$.

The argument for the approximate e-variable is analogous (noting that only $t \geq 1$ needs to be checked in the condition for approximate e-variables) and omitted.

Proposition 10.5 also suggests that controlling δ is a stronger requirement, leading to two different notions of asymptotic e-values and p-values in Section 10.2.

10.2 Asymptotic p-values and e-values

10.2.1 Definitions and conditions

We consider a sequence of datasets $X = X^{(n)}$ of growing size as $n \rightarrow \infty$. Based on $X^{(n)}$ we compute a $[0, 1]$ -valued statistic $P^{(n)} = P^{(n)}(X^{(n)})$. Similarly, we can compute a nonnegative extended random variable $E^{(n)} = E^{(n)}(X^{(n)})$. The following definitions give some intuitive conditions on $P^{(n)}$ and $E^{(n)}$ for them to be considered as p-variables and e-variables in an asymptotic sense.

Definition 10.6: Asymptotic p-variables and e-variables

A sequence $(E^{(n)})_{n \in \mathbb{N}}$ of $[0, \infty]$ -valued random variables is called a sequence of:

- (i) *asymptotic e-variables* for $\mathcal{P}$, if for each $n \in \mathbb{N}$, $E^{(n)}$ is $(\varepsilon_n, \delta_n)$ -approximate for $\mathcal{P}$ for some $(\varepsilon_n, \delta_n)$, satisfying

$$\text{for each } \mathbb{P} \in \mathcal{P}, \quad \lim_{n \rightarrow \infty} \varepsilon_n(\mathbb{P}) = \lim_{n \rightarrow \infty} \delta_n(\mathbb{P}) = 0; \quad (10.1)$$

- (ii) *strongly asymptotic e-variables* for $\mathcal{P}$, if (10.1) is replaced by

$$\text{for each } \mathbb{P} \in \mathcal{P}, \quad \lim_{n \rightarrow \infty} \varepsilon_n(\mathbb{P}) = 0 \text{ and } \delta_n(\mathbb{P}) = 0 \text{ for } n \text{ large enough};$$

- (iii) *uniformly asymptotic e-variables* for $\mathcal{P}$, if (10.1) is replaced by

$$\lim_{n \rightarrow \infty} \sup_{\mathbb{P} \in \mathcal{P}} \varepsilon_n(\mathbb{P}) = \lim_{n \rightarrow \infty} \sup_{\mathbb{P} \in \mathcal{P}} \delta_n(\mathbb{P}) = 0;$$

- (iv) *uniformly strongly asymptotic e-variables* for $\mathcal{P}$, if condition (10.1) is replaced by

$$\lim_{n \rightarrow \infty} \sup_{\mathbb{P} \in \mathcal{P}} \varepsilon_n(\mathbb{P}) = 0 \quad \text{and} \quad \sup_{\mathbb{P} \in \mathcal{P}} \delta_n(\mathbb{P}) = 0 \text{ for } n \text{ large enough};$$

Sequences $(P^{(n)})_{n \in \mathbb{N}}$ of (uniformly, strongly) asymptotic p-variables for $\mathcal{P}$ are defined analogously.

We sometimes simply say that a nonnegative sequence is asymptotic for $\mathcal{P}$, and it should be clear from context whether we mean a sequence of asymptotic e-variables (allowed to be $[0, \infty]$ -valued) or that of asymptotic p-variables ($\mathbb{R}_+$ -valued). Clearly, if $\mathcal{P}$ is finite, then asymptotic e-variables or p-variables are also uniformly asymptotic.

Several conditions for asymptotic p-values and e-values are often encountered in typical statistical contexts. We summarize them together in the following proposition.

Proposition 10.7

- (i) If under each $\mathbb{P} \in \mathcal{P}$, a nonnegative sequence converges to an e-variable (resp. a p-variable) in distribution, then it is a sequence of asymptotic e-variables (resp. p-variables) for $\mathcal{P}$.
- (ii) The nonnegative sequence $(E^{(n)})_{n \in \mathbb{N}}$ is uniformly strongly asymptotic for $\mathcal{P}$ if and only if

$$\limsup_{n \rightarrow \infty} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[E^{(n)}] \leq 1.$$

Similarly, it is strongly asymptotic for $\mathcal{P}$ if and only if

$$\limsup_{n \rightarrow \infty} \mathbb{E}^{\mathbb{P}}[E^{(n)}] \leq 1 \text{ for all } \mathbb{P} \in \mathcal{P}.$$

- (iii) If a sequence of asymptotic e-variables is uniformly integrable for each $\mathbb{P} \in \mathcal{P}$, then it is strongly asymptotic.
- (iv) The nonnegative sequence $(P^{(n)})_{n \in \mathbb{N}}$ is uniformly asymptotic for $\mathcal{P}$ if and only if

$$\limsup_{n \rightarrow \infty} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(P^{(n)} \leq t) \leq t$$

for all $t \in (0, 1)$. Similarly, it is asymptotic for $\mathcal{P}$ if and only if $\limsup_{n \rightarrow \infty} \mathbb{P}(P^{(n)} \leq t) \leq t$ for all $\mathbb{P} \in \mathcal{P}$ and $t \in (0, 1)$.

Proof.

(i) Fix $\mathbb{P} \in \mathcal{P}$. Suppose that $(P^{(n)})_{n \in \mathbb{N}}$ converges in distribution to a p-variable P . Note that

$$\limsup_{n \rightarrow \infty} \mathbb{P}(P_n \leq t) \leq \mathbb{P}(P \leq t) \leq t$$

for each $t \in [0, 1]$. This implies $\lim_{n \rightarrow \infty} \mathbb{P}(P_n \leq t) \vee t = t$ for each $t \in [0, 1]$. Since point-wise convergence of increasing functions to a continuous function on a compact interval implies uniform convergence, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \sup_{t \in [0, 1]} (\mathbb{P}(P_n \leq t) \vee t - t) &= 0 \\ \implies \lim_{n \rightarrow \infty} \sup_{t \in [0, 1]} (\mathbb{P}(P_n \leq t) - t)_+ &= 0. \end{aligned}$$

This shows that $(P^{(n)})_{n \in \mathbb{N}}$ is a sequence of asymptotic p-variables.

Next we show the statement on asymptotic e-variables. For $\delta \in [0, 1]$, and a nonnegative random variable E , consider the function

$$f_\delta(E) = \sup_{t \in \mathbb{R}_+} (\mathbb{E}^\mathbb{P}[E \wedge t] - \delta t) - 1.$$

By definition, if $\varepsilon := f_\delta(E) \vee 0$ is finite, then E is an (ε, δ) -approximate e-variable for $\{\mathbb{P}\}$. Let E be the limit of the nonnegative sequence $(E^{(n)})_{n \in \mathbb{N}}$ in distribution, and take $m \in \mathbb{R}_+$ be such that $\mathbb{P}(E \geq m) < \delta$. Since $E^{(n)} \rightarrow E$ in distribution, there exists $N \in \mathbb{N}$ such that $\mathbb{P}(E^{(n)} \geq m) < \delta$ for $n > N$ by the portmanteau theorem. It follows that, for all $t \in \mathbb{R}_+$,

$$\begin{aligned} \mathbb{E}^\mathbb{P}[E^{(n)} \wedge t] - \mathbb{E}^\mathbb{P}[E^{(n)} \wedge m] - \delta t &\leq \mathbb{P}(E^{(n)} \geq m)(t - m)_+ - \delta t \\ &\leq \delta(t - m)_+ - \delta t \leq 0, \end{aligned}$$

and hence $f_\delta(E^{(n)}) \leq \mathbb{E}^\mathbb{P}[E^{(n)} \wedge m] - 1 \rightarrow \mathbb{E}^\mathbb{P}[E \wedge m] - 1 \leq 0$ as $n \rightarrow \infty$. This implies, in particular, that for any $\delta > 0$, there exists $N' \in \mathbb{N}$ such that for $n > N'$, $E^{(n)}$ is an (δ, δ) -approximate e-variable. Since $\mathbb{P} \in \mathcal{P}$ is arbitrary, this shows that $(E^{(n)})_{n \in \mathbb{N}}$ is a sequence of asymptotic e-variables for $\mathcal{P}$.

(ii) We prove this result for the non-uniform case. The other argument is analogous. First suppose that $(E^{(n)})_{n \in \mathbb{N}}$ is a sequence of strongly asymptotic e-variables for $\mathcal{P}$. Fix $\mathbb{P} \in \mathcal{P}$. Then:

$$\mathbb{E}^\mathbb{P}[E^{(n)}] \leq 1 + \varepsilon_n(\mathbb{P}) \rightarrow 1 \text{ as } n \rightarrow \infty,$$

where the first inequality holds for large enough n . For the other direction, note that any nonnegative extended random variable $E^{(n)}$ is an $(\varepsilon_n, 0)$ -approximate e-variable for $\mathcal{P}$ with the choice $\varepsilon_n(\mathbb{P}) = (\mathbb{E}^\mathbb{P}[E^{(n)}] - 1)_+$ as long as $\mathbb{E}^\mathbb{P}[E^{(n)}]$ is finite. If $\limsup_{n \rightarrow \infty} \mathbb{E}^\mathbb{P}[E^{(n)}] \leq 1$, this implies that $\mathbb{E}^\mathbb{P}[E^{(n)}] < \infty$ for n large enough and the above $\varepsilon_n(\mathbb{P})$ satisfies $\lim_{n \rightarrow \infty} \varepsilon_n(\mathbb{P}) = 0$, that is, $(E^{(n)})_{n \in \mathbb{N}}$ is a sequence of strongly asymptotic e-variables.

(iii) Fix $\mathbb{P} \in \mathcal{P}$. Note that for $t \in \mathbb{R}_+$,

$$\mathbb{E}^\mathbb{P}[E^{(n)}] \leq \mathbb{E}^\mathbb{P}[E^{(n)} \wedge t] + \mathbb{E}^\mathbb{P}[E^{(n)} \mathbf{1}_{\{E^{(n)} \geq t\}}].$$

By uniform integrability, for any $\eta > 0$ there exists $T > 0$ such that

$$\sup_{n \in \mathbb{N}} \mathbb{E}^\mathbb{P} \left[E^{(n)} \mathbf{1}_{\{E^{(n)} \geq T\}} \right] \leq \eta.$$

Thus,

$$\mathbb{E}^{\mathbb{P}}[E^{(n)}] \leq 1 + \varepsilon_n(\mathbb{P}) + \delta_n(\mathbb{P})T + \eta,$$

and hence,

$$\limsup_{n \rightarrow \infty} \mathbb{E}^{\mathbb{P}}[E^{(n)}] \leq 1 + \eta.$$

Since $\eta > 0$ is arbitrary, we conclude by using part (ii).

(iv) The proof is similar to that of part (i) and omitted here.

In Proposition 10.7, part (i) gives a sufficient condition that best describes the intuition behind asymptotic e-variables and p-variables, via convergence in distribution. Converging to an e-variable or p-variable in distribution is a sufficient condition for a sequence to be asymptotic e-variables or asymptotic p-variables, but this condition is not necessary. For instance, any sequence that is smaller than an e-variable is a sequence of asymptotic e-variables, and it does not need to converge. Part (ii) gives an intuitive condition on the expectation for strongly asymptotic e-variables. Part (iii) connects asymptotic e-variables and strongly asymptotic e-variables via uniform integrability, and part (iv) gives equivalent conditions for uniformly (but not strongly) asymptotic p-variables.

We can naturally calibrate asymptotic p-variables and e-variables by applying a (possibly different) calibrator to each element in the sequence, in the sense of Theorem 10.3.

Proposition 10.8

Calibrating a sequence of (uniformly, strongly) asymptotic p-variables yields a sequence of (uniformly, strongly) asymptotic e-variables, and vice versa.

The above result follows directly from Theorem 10.3. The most common way of constructing asymptotic p-variables and e-variables is via limit theorems such as the central limit theorem.

Proposition 10.9

Let $Z^{(n)}$ be a statistic computed based on $X^{(n)}$ for $n \in \mathbb{N}$. Suppose that

$$Z^{(n)} \rightarrow Z \text{ in distribution as } n \rightarrow \infty \text{ under every } \mathbb{P} \in \mathcal{P},$$

where Z is a random variable with a continuous cumulative distribution function F .

- (i) Let $P^{(n)} = F(Z^{(n)})$ or $P^{(n)} = 1 - F(Z^{(n)})$ for $n \in \mathbb{N}$. Then $(P^{(n)})_{n \in \mathbb{N}}$ is a sequence of asymptotic p-variables for $\mathcal{P}$.
- (ii) Let $h : \mathbb{R} \rightarrow [0, \infty]$ be a function with $\mathbb{E}[h(Z)] \leq 1$ that is increasing, decreasing, or upper semicontinuous, and $E^{(n)} = h(Z^{(n)})$ for $n \in \mathbb{N}$. Then $(E^{(n)})_{n \in \mathbb{N}}$ is a sequence of asymptotic e-variables for $\mathcal{P}$. If h is further bounded, then $(E^{(n)})_{n \in \mathbb{N}}$ is strongly asymptotic for $\mathcal{P}$.

Proof.

By the continuous mapping theorem, $P^{(n)}$ converges weakly to the uniform distribution, and hence item (i) directly follows from Proposition 10.7, part (i).

For (ii) we may argue as follows. If h would have been continuous, then the result would follow from the continuous mapping theorem and Proposition 10.7, part (i). For upper semicontinuity, note that $\limsup_{n \rightarrow \infty} \mathbb{P}(E^{(n)} \geq x) \leq \mathbb{P}(h(Z) \geq x)$ for all $x \in \mathbb{R}$ by convergence in distribution. This gives $\limsup_{n \rightarrow \infty} f_{\delta}(E^{(n)}) \leq f_{\delta}(h(Z)) \leq 0$ using the notation and argument in the proof

of Proposition 10.7, part (i). The last statement follows from Proposition 10.7, part (iii). When h is increasing or decreasing, replacing it with its upper semicontinuous version, which does not change $\mathbb{E}[h(Z)]$, yields the desired statements.

Some common cases where we can apply Proposition 10.9 include:

1. $Z \stackrel{d}{\sim} N(0, 1)$ as would follow from the central limit theorem.
2. $Z \stackrel{d}{\sim} |N(0, 1)|$, where $|N(0, 1)|$ is the folded normal distribution (i.e., the distribution of $|X|$ where $X \stackrel{d}{\sim} N(0, 1)$).
3. $Z \stackrel{d}{\sim} \chi_\nu^2$, where χ_ν^2 denotes the chi-square distribution with ν degrees of freedom. Such a limit occurs e.g., for likelihood ratio or chi-square tests.

10.3 An asymptotic e-value via the central limit theorem

We next provide a prototypical example of asymptotic e-variables along the lines of the central limit theorem.

Example 10.10: Testing the mean with unknown variance

Suppose we observe the dataset $X^{(n)} = (X^1, \dots, X^n)$ where $X^i \stackrel{\text{iid}}{\sim} \mathbb{P}$ are real-valued. We assume that their variance, $\text{var}^{\mathbb{P}}(X^i)$, is positive and finite, but is unknown. We seek to test the hypothesis

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[X^i] = 0, \text{var}^{\mathbb{P}}(X^i) \in (0, \infty)\}, \quad (10.2)$$

where $\mathcal{M}_1$ denotes the set of all probability measures on $\mathbb{R}$. Define

$$\begin{aligned} \bar{X}^{(n)} &= \frac{1}{n} \sum_{i=1}^n X^i, \quad S^{(n)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X^i)^2}, \\ \hat{\sigma}^{(n)} &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X^i - \bar{X}^{(n)})^2}, \end{aligned} \quad (10.3)$$

and for fixed $\lambda \in \mathbb{R}$, let

$$E^{(n)} = \exp\left(\lambda \frac{\sqrt{n} \bar{X}^{(n)}}{S^{(n)}} - \frac{\lambda^2}{2}\right), \quad (10.4)$$

and

$$\tilde{E}^{(n)} = \frac{1}{2} \exp\left(\lambda \frac{\sqrt{n} \bar{X}^{(n)}}{S^{(n)}} - \frac{\lambda^2}{2}\right) + \frac{1}{2} \exp\left(-\lambda \frac{\sqrt{n} \bar{X}^{(n)}}{S^{(n)}} - \frac{\lambda^2}{2}\right). \quad (10.5)$$

The next theorem justifies that (10.4) defines a sequence of strongly asymptotic e-variables.

Theorem 10.11

The sequences of random variables $E^{(n)}$ in (10.4) and $\tilde{E}^{(n)}$ in (10.5) are strongly asymptotic e-variables for $\mathcal{P}$ in (10.2). If $\hat{\sigma}^{(n)}$ is used in place of $S^{(n)}$ in (10.4) and (10.5), then $E^{(n)}$ and $\tilde{E}^{(n)}$ are asymptotic e-variables for $\mathcal{P}$. In either case, for $\lambda > 0$, $E^{(n)}$ (resp. $\tilde{E}^{(n)}$) grows to infinity with probability 1 under the alternative that $X^1, \dots, X^n$ are drawn iid from $\mathbb{Q}$ with $\mathbb{E}^{\mathbb{Q}}[X^i] > 0$ (resp. $\mathbb{E}^{\mathbb{Q}}[X^i] \neq 0$) and $\text{var}^{\mathbb{Q}}(X^i) < \infty$.

Proof.

Take any $\mathbb{P} \in \mathcal{P}$. By the central limit theorem, the law of large numbers and Slutsky's theorem, $\sqrt{n}\bar{X}^{(n)}/S^{(n)}$ converges in distribution to $N(0, 1)$ as $n \rightarrow \infty$. We may verify that

$$\mathbb{E}^{\mathbb{P}} \left[\exp \left(\lambda Z - \frac{\lambda^2}{2} \right) \right] = 1,$$

for $Z \sim N(0, 1)$. By Proposition 10.9 it follows that $E^{(n)}$ are asymptotic e-variables. To upgrade this result to strongly asymptotic, we can apply Theorem 2.5 of Giné et al. [1997] to show that the sequence $(E^{(n)})_{n \in \mathbb{N}}$ is uniformly integrable and then conclude the proof by Proposition 10.7, part (iii).

Next take an alternative $\mathbb{Q}$ with $\mathbb{E}^{\mathbb{Q}}[X^i] > 0$, $\text{var}^{\mathbb{Q}}(X^i) < \infty$, and suppose that $\lambda > 0$. Observe the following. By the strong law of large numbers

$$\bar{X}^{(n)} \rightarrow \mathbb{E}^{\mathbb{Q}}[X^i] > 0 \quad \text{almost surely as } n \rightarrow \infty,$$

and

$$S^{(n)} \rightarrow \sqrt{\text{var}^{\mathbb{Q}}(X^i) + \mathbb{E}^{\mathbb{Q}}[X^i]^2} > 0 \quad \text{almost surely as } n \rightarrow \infty.$$

The above imply that

$$\lambda \frac{\sqrt{n}\bar{X}^{(n)}}{S^{(n)}} \rightarrow \infty \quad \text{almost surely as } n \rightarrow \infty,$$

which also means that $E^{(n)}$ goes to ∞ with probability 1. The argument for $\tilde{E}^{(n)}$ is analogous and omitted. The argument for $\hat{\sigma}^{(n)}$ follows from Proposition 10.9.

The above conclusion extends to a broader class of asymptotic e-variables obtained by mixing in (10.4) over λ with respect to a distribution with bounded support. Concretely, let G be any distribution on $\mathbb{R}$ with compact support. Then, under the assumptions of this section,

$$\int \exp \left(\lambda \frac{\sqrt{n}\bar{X}^{(n)}}{S^{(n)}} - \frac{\lambda^2}{2} \right) dG(\lambda),$$

is a strongly asymptotic e-variable.

One central purpose of defining asymptotic e-variables is to handle cases where one does not know how to construct a nonasymptotic e-variable. Indeed, for the above $\mathcal{P}$, we do not know of nonasymptotic e-variables (other than constants). Had a bound on the variance been known, such nonasymptotic e-variables do exist, as shown next.

Example 10.12: Testing the mean with bounded variance

Given some $\sigma > 0$, consider the null hypothesis

$$\mathcal{P}_\sigma = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^\mathbb{P}[X] = 0, \text{var}^\mathbb{P}(X) \leq \sigma^2\}.$$

Then X^2/σ^2 is an e-variable for $\mathcal{P}_\sigma$ (see Section 2.6.3), and for any $\lambda \in \mathbb{R}$ the following are e-variables for $\mathcal{P}_\sigma$:

$$\exp\left(\lambda X - \frac{\lambda^2}{6}(X^2 + 2\sigma^2)\right)$$

and

$$\exp\left(\phi(\lambda X) - \frac{\lambda^2}{2}\sigma^2\right),$$

where $\phi(x) = \log(1 + x + x^2/2)$ for $x \geq 0$ and $\phi(x) = -\log(1 - x + x^2/2)$ for $x < 0$. The different choices of e-variables have large e-power against different alternatives.

10.4 Approximate/asymptotic compound p-values and e-values

In this section we extend the setting of Sections 10.1 and 10.2 to the multiple hypothesis testing setting, introduced in Chapter 9.

10.4.1 Compound p-values and e-values

Compound e-variables are defined in Definition 9.1, and they satisfy

$$\sum_{k:\mathbb{P} \in \mathcal{P}_k} \mathbb{E}^\mathbb{P}[E_k] \leq K \quad \text{for all } \mathbb{P} \in \mathcal{P}.$$

We next define their counterpart, compound p-variables.

Definition 10.13: Compound p-variables

Fix the null hypotheses $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ and a set $\mathcal{P}$ of distributions. Let $P_1, \dots, P_K$ be nonnegative random variables. We say that $P_1, \dots, P_K$ are *compound* p-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$ if

$$\sum_{k:\mathbb{P} \in \mathcal{P}_k} \mathbb{P}(P_k \leq t) \leq Kt$$

for all $t \in (0, 1)$ and all $\mathbb{P} \in \mathcal{P}$.

Similar to Theorem 10.3, compound p-variables and compound e-variables can be transformed via calibration.

Theorem 10.14

Let h be any (p-to-e) calibrator. If $P_1, \dots, P_K$ are compound p-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$, then $E_k = h(P_k)$, $k \in \mathcal{K}$, are compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$. Similarly, calibration in the e-to-p direction is also valid.

Proof.

Fix an atomless $\mathbb{P}$ and it suffices to verify the statement for this $\mathbb{P}$. Let $K_0 = |\mathcal{N}(\mathbb{P})|$ and

take $L \stackrel{d}{\sim} U(\mathcal{N}(\mathbb{P}))$ and $U \stackrel{d}{\sim} U[0, 1]$, mutually independent and independent of everything else. Then

$$\tilde{P} := P_L \mathbb{1}_{\{U \leq K_0/K\}} + \mathbb{1}_{\{U > K_0/K\}}$$

is a bona fide p-value because for $t \in (0, 1)$,

$$\mathbb{P}(\tilde{P} \leq t) = \mathbb{P}(P_L \leq t) \mathbb{P}(U \leq K_0/K) = \frac{K_0}{K} \frac{1}{K_0} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(P_k \leq t) \leq t.$$

Therefore,

$$\begin{aligned} \frac{1}{K} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[h(P_k)] &= \frac{K_0}{K} \mathbb{E}^{\mathbb{P}}[h(P_L)] \\ &\leq \mathbb{E}^{\mathbb{P}}[h(P_L) \mathbb{1}_{\{U \leq K_0/K\}} + h(1) \mathbb{1}_{\{U > K_0/K\}}] \\ &= \mathbb{E}^{\mathbb{P}}[h(\tilde{P})] \leq 1, \end{aligned}$$

where the last inequality holds since h is a calibrator. The other direction of calibration can be checked directly.

Similarly to Example 9.2, an example of compound p-variables is given by weighted p-variables.

Example 10.15: Weighted p-values

Let $w_1, \dots, w_K \geq 0$ be deterministic nonnegative numbers such that $\sum_{k \in \mathcal{K}} w_k \leq K$. Let $P'_1, \dots, P'_K$ be p-variables and define $\tilde{P}_k = P'_k/w_k$ for $k \in \mathcal{K}$ with the convention $0/0 = 0$. Then, $\tilde{P}_1, \dots, \tilde{P}_K$ are compound p-variables.

Example 10.16: Combining p-variables and compound e-variables

Suppose that P_k is a p-variable for $\mathcal{P}_k$ for all $k \in \mathcal{K}$; $P_1, \dots, P_K$ can be arbitrarily dependent. Moreover, let $E_1, \dots, E_K$ be compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$. If $(E_1, \dots, E_K)$ is independent of $(P_1, \dots, P_K)$, then $P_1/E_1, \dots, P_K/E_K$ are compound p-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$. This can be seen from Theorem 2.9.

10.4.2 Approximate compound p/e-values

We next generalize the approximate p-variables and e-variables in Definition 10.1 to introduce approximate compound p-variables and e-variables.

Throughout this subsection, we assume that each $\mathbb{P}$ is atomless; this means that we can simulate a uniformly distributed random variable under $\mathbb{P}$. This enables us to formulate the definition of approximate compound p-variables and e-variables following the equivalent formulation for approximate p-variables/e-variables in Proposition 10.2. This formulation is helpful for our later results on the approximate and asymptotic FDR control (Theorem 10.24).

Definition 10.17: Approximate compound p-variables and e-variables

Fix the null hypotheses $\mathcal{P}_1, \dots, \mathcal{P}_K$, a set of distributions $\mathcal{P}$, and two functions $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$, and write $\varepsilon_{\mathbb{P}} = \varepsilon(\mathbb{P})$ and $\delta_{\mathbb{P}} = \delta(\mathbb{P})$ for $\mathbb{P} \in \mathcal{P}$.

- (i) Nonnegative random variables $P_1, \dots, P_K$ are (ε, δ) -approximate compound p-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$ if for every $\mathbb{P} \in \mathcal{P}$, there exists an event A such that $\mathbb{P}(A) \geq 1 - \delta_{\mathbb{P}}$ and

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(P_k \leq t, A) \leq K(1 + \varepsilon_{\mathbb{P}})t \quad \text{for all } t \in (0, 1);$$

- (ii) $[0, \infty]$ -valued random variables $E_1, \dots, E_K$ are (ε, δ) -approximate compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$ if for every $\mathbb{P} \in \mathcal{P}$, there exists an event A such that $\mathbb{P}(A) \geq 1 - \delta_{\mathbb{P}}$ and

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k \mathbf{1}_A] \leq K(1 + \varepsilon_{\mathbb{P}}).$$

We sometimes omit “under $\mathcal{P}$ ” when it is clear.

From per-hypothesis approximate p-variables (or e-variables), we can recover approximate compound p-variables (or e-variables) for the multiple testing problem.

Proposition 10.18

Suppose that for all $k \in \mathcal{K}$, P_k (resp. E_k) is an $(\varepsilon_k, \delta_k)$ -approximate p-variable (resp. e-variable) for $\mathcal{P}_k$. Then, $P_1, \dots, P_K$ (resp. $E_1, \dots, E_K$) are (ε, δ) -approximate compound p-variables (resp. e-variables) for the choice:

$$\varepsilon(\mathbb{P}) = \frac{1}{K} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \varepsilon_k(\mathbb{P}), \quad \delta(\mathbb{P}) = \sum_{k: \mathbb{P} \in \mathcal{P}_k} \delta_k(\mathbb{P}).$$

Proof.

We carry out the argument only for compound e-variables (the argument for p-variables is analogous). Take any $\mathbb{P} \in \mathcal{P}$ and let A_k be the event provided by Proposition 10.2 for the k -th $(\varepsilon_k, \delta_k)$ -approximate e-variable. Then, letting $A := \bigcap_{k: \mathbb{P} \in \mathcal{P}_k} A_k$, it follows that,

$$\mathbb{P}(A^c) = \mathbb{P}\left(\bigcup_{k: \mathbb{P} \in \mathcal{P}_k} A_k^c\right) \leq \sum_{k: \mathbb{P} \in \mathcal{P}_k} \delta_k(\mathbb{P}) = \delta(\mathbb{P}).$$

We get

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k \mathbf{1}_A] \leq \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k \mathbf{1}_{A_k}] \leq \sum_{k: \mathbb{P} \in \mathcal{P}_k} (1 + \varepsilon_k(\mathbb{P})) \leq K(1 + \varepsilon(\mathbb{P})).$$

This proves the desired statement for compound e-variables.

We next provide an analogous statement to that of Proposition 10.2 for approximate compound e-variables. A corresponding representation for approximate compound p-variables seems to be more complicated, and that is the reason why we chose to formulate Definition 10.17 through the event A instead of using (E1) below.

Proposition 10.19

Let $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$ and $E_1, \dots, E_K$ be $[0, \infty]$ -valued random variables. The following two statements are equivalent (assuming that $\mathbb{P}$ is atomless for all $\mathbb{P} \in \mathcal{P}$):

(E1) For all $t \in \mathbb{R}_+$ and $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}^{\mathbb{P}} \left[\left(\sum_{k: \mathbb{P} \in \mathcal{P}_k} E_k \right) \wedge t \right] \leq K(1 + \varepsilon_{\mathbb{P}}) + \delta_{\mathbb{P}} t;$$

(E2) $E_1, \dots, E_K$ are (ε, δ) -approximate compound e-variables for the hypotheses $\mathcal{P}_1, \dots, \mathcal{P}_K$ under $\mathcal{P}$.

Proof.

We start with (E1). By direct manipulation we see that (E1) is equivalent to the following statement:

$$\mathbb{E}^{\mathbb{P}} \left[\left(\frac{1}{K} \sum_{k: \mathbb{P} \in \mathcal{P}_k} E_k \right) \wedge t \right] \leq 1 + \varepsilon_{\mathbb{P}} + \delta_{\mathbb{P}} t \text{ for all } t \in \mathbb{R}_+ \text{ and all } \mathbb{P} \in \mathcal{P}.$$

Now fix $\mathbb{P} \in \mathcal{P}$. According to the above statement,

$$E := \frac{1}{K} \sum_{k: \mathbb{P} \in \mathcal{P}_k} E_k \quad \text{is an } (\varepsilon, \delta)\text{-approximate e-variable for } \{\mathbb{P}\}.$$

By Proposition 10.2 there exists an event A with $\mathbb{P}(A) \geq 1 - \delta$ such that $\mathbb{E}^{\mathbb{P}}[E \mathbf{1}_A] \leq 1 + \varepsilon_{\mathbb{P}}$, that is,

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k \mathbf{1}_A] \leq K(1 + \varepsilon_{\mathbb{P}}).$$

Since $\mathbb{P} \in \mathcal{P}$ was arbitrary, the above establishes the implication (E1) $\Rightarrow$ (E2).

For the other direction, fix an arbitrary $\mathbb{P} \in \mathcal{P}$. Suppose (E2) holds, that is, $\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k \mathbf{1}_A] \leq K(1 + \varepsilon_{\mathbb{P}})$ for some event A with $\mathbb{P}(A) \geq 1 - \delta$. Define $E := (1/K) \sum_{k: \mathbb{P} \in \mathcal{P}_k} E_k$, and by applying the direction (E2) $\Rightarrow$ (E1) in Proposition 10.2 with the hypothesis $\{\mathbb{P}\}$, we obtain the condition in (E1), completing the proof.

We can calibrate approximate compound p-variables to approximate compound e-variables and vice versa as per usual if we use *the same* calibrator for all hypotheses. For instance, given a p-to-e calibrator h , if $P_1, \dots, P_K$ are approximate compound p-variables for $\mathcal{P}_1, \dots, \mathcal{P}_K$ under $\mathcal{P}$, then $h(P_1), \dots, h(P_K)$ are approximate compound e-variables for the same hypotheses.

Proposition 10.20

Let $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$. Calibrating (ε, δ) -approximate compound p-variables yields (ε, δ) -approximate compound e-variables, and vice versa.

Proof.

For the first statement, let $P_1, \dots, P_K$ be (ε, δ) -approximate compound p-variables. Fix any $\mathbb{P} \in \mathcal{P}$ and let h be a (p-to-e) calibrator. Let A be an event with $\mathbb{P}(A) \geq 1 - \delta_{\mathbb{P}}$ such that $\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(P_k \leq t, A) \leq K(1 + \varepsilon_{\mathbb{P}})t$ for $t \in (0, 1)$, and let $b = (1 + \varepsilon_{\mathbb{P}})/\mathbb{P}(A) - 1 \geq 0$. Let

$\mathbb{P}_A(\cdot) = \mathbb{P}(\cdot \mid A)$. We have

$$\frac{1}{K} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}_A(P_k \leq t) \leq (1+b)t \quad \text{for } t \in (0, 1).$$

Define U, L, P_L , and $\tilde{P}$ as in the proof of Theorem 10.14. Then, $\tilde{P}$ satisfies

$$\begin{aligned} \mathbb{P}_A(\tilde{P} \leq t) &= \mathbb{P}_A(P_L \leq t) \mathbb{P}_A(U \leq K_0/K) \\ &= \frac{K_0}{K} \frac{1}{K_0} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}_A(P_k \leq t) \leq (1+b)t. \end{aligned}$$

This means that $\tilde{P}$ is a $(b, 0)$ -approximate p-variable for $\{\mathbb{P}_A\}$. By Theorem 10.3, $h(\tilde{P})$ is a $(b, 0)$ -approximate e-variable for $\{\mathbb{P}_A\}$. Thus,

$$\begin{aligned} \frac{1}{K} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}_A} [h(P_k)] &= \frac{K_0}{K} \mathbb{E}^{\mathbb{P}_A} [h(P_L)] \\ &\leq \mathbb{E}^{\mathbb{P}_A} [h(\tilde{P})] \leq 1+b = \frac{1+\varepsilon_{\mathbb{P}}}{\mathbb{P}(A)}. \end{aligned}$$

It follows that $\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}} [h(P_k) \mathbf{1}_A] \leq K(1+\varepsilon_{\mathbb{P}})$, and therefore, we have that $h(P_1), \dots, h(P_K)$ are (ε, δ) -approximate compound e-variables.

The second claim follows by Markov's inequality. Fix $\mathbb{P} \in \mathcal{P}$ and let A be the event given by the definition of approximate compound p-variables. Let $P_k = (1/E_k) \wedge 1$. Then, for any $t \in (0, 1)$,

$$\begin{aligned} \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(P_k \leq t, A) &= \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(E_k \geq 1/t, A) \\ &= \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(E_k \mathbf{1}_A \geq 1/t) \\ &\leq \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}} [E_k \mathbf{1}_A] t \leq K(1+\varepsilon_{\mathbb{P}})t, \end{aligned}$$

as desired.

Remark 10.21. We may define an alternative notion of $(\varepsilon, \delta)^*$ -approximate compound e-variables and p-variables as follows. Fix the null hypotheses $\mathcal{P}_1, \dots, \mathcal{P}_K$, a set of distributions $\mathcal{P}$, and two functions $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$, and write $\varepsilon_{\mathbb{P}} = \varepsilon(\mathbb{P})$ and $\delta_{\mathbb{P}} = \delta(\mathbb{P})$ for $\mathbb{P} \in \mathcal{P}$.

- (i) Nonnegative random variables $P_1, \dots, P_K$ are $(\varepsilon, \delta)^*$ -approximate compound p-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$ if

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{P}(P_k \leq t) \leq K((1+\varepsilon_{\mathbb{P}})t + \delta_{\mathbb{P}}) \quad \text{for all } t \in (0, 1) \text{ and } \mathbb{P} \in \mathcal{P}.$$

- (ii) $[0, \infty]$ -valued random variables $E_1, \dots, E_K$ are $(\varepsilon, \delta)^*$ -approximate compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$ if

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}} [E_k \wedge t] \leq K(1+\varepsilon_{\mathbb{P}} + \delta_{\mathbb{P}}t) \quad \text{for all } t \in \mathbb{R}_+ \text{ and } \mathbb{P} \in \mathcal{P}.$$

This definition is also a natural generalization of Definition 10.1, but it is different from Definition 10.17. The difference is transparent from Proposition 10.19. We stick to Definition 10.17 because it fits better with the approximate and asymptotic FDR control that we study below.

10.4.3 Asymptotic compound p/e-values

In the context of multiple testing procedures in Chapter 9, compound e-values are the essential quantities for their FDR control. The next definition introduces compound e-variables in the approximate or asymptotic sense. The reader may directly anticipate the definitions that follow. However, what is most important, is to clarify the asymptotic setup. As before, we seek to test hypotheses $\mathcal{P}_1, \dots, \mathcal{P}_K$ based on data $X^{(n)}$. We are interested in asymptotics in which either $n \rightarrow \infty$ (with K fixed) or $K \rightarrow \infty$ (with n fixed) or both $K, n \rightarrow \infty$.

Formally, we consider triangular-array type of asymptotics indexed by a single parameter $m \in \mathbb{N}$ that determines the dataset size, $n = n(m)$, and the number of hypotheses, $K = K(m)$. We require that $m \rightarrow \infty$ and that $\max\{K, n\} \rightarrow \infty$ as $m \rightarrow \infty$.

All three asymptotic regimes are frequently encountered in the multiple testing literature. When using Gaussian approximations to construct Z-test or t-test p-values, one is relying on asymptotics as the size of X grows to infinity. When using empirical Bayes arguments to share strength and learn nuisance parameters across hypotheses, one is relying on asymptotics as K grows to infinity.

Definition 10.22: Asymptotic compound p-variables and e-variables

A sequence

$$\left(E_1^{(n(m))}, \dots, E_{K(m)}^{(n(m))} \right)_{m \in \mathbb{N}}$$

of tuples of $[0, \infty]$ -valued random variables is called a sequence of:

- (a) *asymptotic compound e-variables* for $\mathcal{P}_1, \dots, \mathcal{P}_{K(m)}$ under $\mathcal{P}$, if for each m , $E_1^{(n(m))}, \dots, E_{K(m)}^{(n(m))}$ are $(\varepsilon_m, \delta_m)$ -approximate compound e-variables for $\mathcal{P}_1, \dots, \mathcal{P}_{K(m)}$ under $\mathcal{P}$ for some $(\varepsilon_m, \delta_m)$, satisfying

$$\text{for each } \mathbb{P} \in \mathcal{P}, \quad \lim_{m \rightarrow \infty} \varepsilon_m(\mathbb{P}) = \lim_{m \rightarrow \infty} \delta_m(\mathbb{P}) = 0; \quad (10.6)$$

- (b) *strongly asymptotic compound e-variables* for $\mathcal{P}_1, \dots, \mathcal{P}_{K(m)}$ under $\mathcal{P}$, if condition (10.6) is replaced by

$$\begin{aligned} \text{for each } \mathbb{P} \in \mathcal{P}: \quad & \lim_{m \rightarrow \infty} \varepsilon_m(\mathbb{P}) = 0 \\ & \text{and } \delta_m(\mathbb{P}) = 0 \text{ for } m \text{ large enough.} \end{aligned}$$

Uniformly (strongly) asymptotic compound e-variables are defined analogously.

Sequences of (uniformly, strongly) asymptotic compound p-variables can be defined analogously.

For the remainder of this section, we assume that each $\mathbb{P}$ is atomless; this means that we can simulate a uniformly distributed random variable under $\mathbb{P}$. This allows us to use the equivalent definition in Proposition 10.2 of approximate e-variables through an event A with a constraint on its probability. As in Definition 9.7, we denote by $F_{\mathcal{D}}$ and $R_{\mathcal{D}}$ the number of false discoveries and the total number of discoveries, that is,

$$F_{\mathcal{D}} = \sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbf{1}_{\{k \in \mathcal{D}\}} = |\mathcal{N}(\mathbb{P}) \cap \mathcal{D}| \quad \text{and} \quad R_{\mathcal{D}} = |\mathcal{D}|.$$

Definition 10.23: Approximate and asymptotic FDR control

Let $\mathcal{P}$ be a set of distributions, $\mathcal{D}$ be a multiple testing procedure, and let $\varepsilon : \mathcal{P} \rightarrow \mathbb{R}_+$ and $\delta : \mathcal{P} \rightarrow [0, 1]$. The procedure $\mathcal{D}$ is said to have

- (a) (ε, δ) -approximate FDR control at level α for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}$, if for every $\mathbb{P} \in \mathcal{P}$, there exists an event A such that

$$\mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \mathbb{1}_A \right] \leq \alpha(1 + \varepsilon_{\mathbb{P}}) \quad \text{and} \quad \mathbb{P}(A) \geq 1 - \delta_{\mathbb{P}};$$

- (b) asymptotic FDR control at level α for $(\mathcal{P}_1, \dots, \mathcal{P}_{K(m)})$ under $\mathcal{P}$ if for each $m \in \mathbb{N}$ it has $(\varepsilon_m, \delta_m)$ -approximate FDR control at level α for some $(\varepsilon_m, \delta_m)$, satisfying

$$\text{for each } \mathbb{P} \in \mathcal{P}, \quad \lim_{m \rightarrow \infty} \varepsilon_m(\mathbb{P}) = \lim_{m \rightarrow \infty} \delta_m(\mathbb{P}) = 0.$$

It is clear that the $(0, 0)$ -approximate FDR control is equivalent to the FDR control defined in Definition 9.7. A procedure with (ε, δ) -approximate FDR control also satisfies

$$\text{FDR}_{\mathcal{D}} = \mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \right] \leq \alpha(1 + \varepsilon_{\mathbb{P}}) + \delta_{\mathbb{P}} \text{ for any } \mathbb{P} \in \mathcal{P}.$$

Conversely, any procedure satisfying the above inequality has $(\varepsilon, \delta/\alpha)$ -approximate FDR control, by invoking the equivalence in Proposition 10.2; more precisely, the above condition implies, for all $t \geq 1$,

$$\mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{\alpha R_{\mathcal{D}}} \wedge t \right] \leq \mathbb{E}^{\mathbb{P}} \left[\frac{F_{\mathcal{D}}}{\alpha R_{\mathcal{D}}} \right] \leq (1 + \varepsilon_{\mathbb{P}}) + \frac{\delta_{\mathbb{P}}}{\alpha} \leq (1 + \varepsilon_{\mathbb{P}}) + \frac{\delta_{\mathbb{P}}}{\alpha} t.$$

With Definitions 10.22 and 10.23, we can formally state that the e-BH procedure applied to asymptotic compound e-values has asymptotic FDR control.

Theorem 10.24

Suppose that we apply the e-BH procedure at level α to $E_1, \dots, E_K$. Then:

- (i) If $E_1, \dots, E_K$ are compound e-variables, then e-BH controls the FDR at level α .
- (ii) If $E_1, \dots, E_K$ are (ε, δ) -approximate compound e-variables, then e-BH satisfies (ε, δ) -approximate FDR control, and so it controls the FDR at level $\alpha(1 + \varepsilon) + \delta$.
- (iii) If $E_1, \dots, E_K$ are asymptotic compound e-variables, then e-BH has asymptotic FDR control at level α .

The same conclusions also hold for any self-consistent e-testing procedures at level α (Definition 9.9).

Theorem 10.24 follows directly from the definitions and the previous results on the e-BH procedure.

10.4.4 The simultaneous t-test sequence model

We next describe a prototypical model of constructing asymptotic compound e-variables in a setting with composite null hypotheses.

In our setting, we observe K groups of data of size n , that is, we observe $X = (X_k^1, \dots, X_k^n)_{k \in \mathcal{K}}$. We consider distributions $\mathbb{P}$ according to which all observations are jointly independent and the data within each group are iid. The universe of distributions $\mathcal{P}$ is understood to obey the above assumptions. Given any $\mathbb{P} \in \mathcal{P}$, we write $\mathbb{P}_k$ for its k -th group-wise marginal so that $X \sim \mathbb{P}$ entails that

$$X_k^i \stackrel{\text{iid}}{\sim} \mathbb{P}_k \quad \text{for } i = 1, \dots, n.$$

We seek to test the hypotheses,

$$\mathcal{P}_k = \{\mathbb{P} \in \mathcal{P} : \mathbb{E}^{\mathbb{P}_k}[X_k^i] = 0, \text{ var}^{\mathbb{P}_k}(X_k^i) \in (0, \infty)\}.$$

To this end, we start by constructing summary statistics in a group-wise manner, similar to (10.3):

$$\begin{aligned} \bar{X}_k^{(n)} &= \frac{1}{n} \sum_{i=1}^n X_k^i, \quad S_k^{(n)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_k^i)^2}, \\ \hat{\sigma}_k^{(n)} &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_k^i - \bar{X}_k^{(n)})^2}. \end{aligned}$$

To motivate the construction that follows, fix any $\mathbb{P} \in \mathcal{P}$ and write $\mu_k(\mathbb{P}) = \mathbb{E}^{\mathbb{P}_k}[X_k^i]$ and $\sigma_k^2(\mathbb{P}) = \text{var}^{\mathbb{P}_k}(X_k^i)$ (supposing these exist). Observe that for any $k \in \mathcal{K}$,

$$\mathbb{E}^{\mathbb{P}_k}[(S_k^{(n)})^2] = \sigma_k^2(\mathbb{P}) + \mu_k^2(\mathbb{P}) \text{ for all } k \in \mathcal{K},$$

and so

$$\mathbb{E}^{\mathbb{P}_k}[(S_k^{(n)})^2] = \sigma_k^2(\mathbb{P}) \text{ for } \mathbb{P} \in \mathcal{P}_k.$$

One can check that for any $\mathbb{P} \in (\bigcup_{k \in \mathcal{K}} \mathcal{P}_k) \cap \mathcal{P}$,

$$\sum_{k: \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[\tilde{E}_k^{(n)}(\mathbb{P})] = K, \quad \text{where} \quad \tilde{E}_k^{(n)}(\mathbb{P}) = \frac{K(S_k^{(n)})^2}{\sum_{j: \mathbb{P} \in \mathcal{P}_j} \sigma_j^2(\mathbb{P})}.$$

The above construction is not directly usable in practice because of the dependence of $\tilde{E}_k^{(n)}(\mathbb{P})$ on $\mathbb{P}$. However, it motivates a construction in which we conservatively estimate the $\mathbb{P}$ -dependent quantity $\sum_{j: \mathbb{P} \in \mathcal{P}_j} \sigma_j^2(\mathbb{P})$ by $\sum_{j \in \mathcal{K}} (\hat{\sigma}_j^{(n)})^2$ and then let

$$E_k^{(n)} = \frac{K(S_k^{(n)})^2}{\sum_{j \in \mathcal{K}} (\hat{\sigma}_j^{(n)})^2}, \quad k \in \mathcal{K}.$$

We will show that the above are asymptotic compound e-variables under an asymptotic setup indexed by $m \in \mathbb{N}$ such that $n = n(m)$, $K = K(m)$ and $\max\{n, K\} \rightarrow \infty$ as $m \rightarrow \infty$. In case $K \rightarrow \infty$, we further restrict the class $\mathcal{P}$ to require that for any $\mathbb{P} \in \mathcal{P}$, $|\{k \in \mathbb{N} : \mathbb{P} \in \mathcal{P}_k\}| = \infty$. In addition to this restriction (which remains implicit in our notation), our formal result will be stated in terms of the following two sets of distributions. The first set imposes some restrictions on moments of the marginals for null groups,

$$\mathcal{P}^B = \{\mathbb{P} \in \mathcal{P} : \mathbb{E}^{\mathbb{P}_k}[|X_k^i|^{2(1+\delta)}] \leq C, \text{ var}^{\mathbb{P}_k}(X_k^i) \geq c, \forall k \text{ with } \mathbb{P} \in \mathcal{P}_k\},$$

where $c, C \in (0, \infty)$ and $\delta \in (0, 1]$ are fixed constants. The second set also imposes normality of the marginals for null groups,

$$\mathcal{P}^N = \{\mathbb{P} \in \mathcal{P} : \mathbb{P}_k = N(0, \sigma_k^2(\mathbb{P})), c \leq \sigma_k^2(\mathbb{P}) \leq C, \forall k \text{ with } \mathbb{P} \in \mathcal{P}_k\},$$

for some constants $c, C \in (0, \infty)$.

Proposition 10.25

Suppose that $n = n(m) \geq 2$ and $K = K(m) \geq 1$ are such that $\max\{n, K\} \rightarrow \infty$ as $m \rightarrow \infty$. Then:

- (i) $(E_1^{(n)}, \dots, E_K^{(n)})_{m \in \mathbb{N}}$ is a sequence of asymptotic compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}^B$.
- (ii) $(E_1^{(n)}, \dots, E_K^{(n)})_{m \in \mathbb{N}}$ is a sequence of strongly asymptotic compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$ under $\mathcal{P}^N$.

Proof.

Fix an arbitrary distribution $\mathbb{P} \in (\bigcup_k \mathcal{P}_k) \cap \mathcal{P}^{\mathcal{B}}$ and write $\sigma_k^2 = \sigma_k^2(\mathbb{P})$, omitting the dependence on $\mathbb{P}$. We write $K_0(\mathbb{P}) = |\mathcal{N}(\mathbb{P})| = \{k \in \mathcal{K} : \mathbb{P} \in \mathcal{P}_k\}$. Note that $\max\{K_0(\mathbb{P}), n\} \rightarrow \infty$ and that

$$\frac{1}{K} \sum_{k:\mathbb{P} \in \mathcal{P}_k} E_k^{(n)} \leq \frac{\sum_{k:\mathbb{P} \in \mathcal{P}_k} (S_k^{(n)})^2}{\sum_{k:\mathbb{P} \in \mathcal{P}_k} (\hat{\sigma}_k^{(n)})^2}. \quad (10.7)$$

We will argue that the right-hand side of the above expression converges in probability (and thus also in distribution) to the constant 1. Arguing as in Proposition 10.7, part (i), it will then follow that $E_1^{(n)}, \dots, E_K^{(n)}$ are asymptotic compound e-variables. We claim that

$$\frac{n \sum_{k:\mathbb{P} \in \mathcal{P}_k} (S_k^{(n)})^2}{n \sum_{k:\mathbb{P} \in \mathcal{P}_k} \sigma_k^2} \xrightarrow{\mathbb{P}} 1 \quad \text{as } m \rightarrow \infty.$$

Writing $U^{(m)} = n \sum_{k:\mathbb{P} \in \mathcal{P}_k} (S_k^{(n)})^2 = \sum_{k:\mathbb{P} \in \mathcal{P}_k} \sum_{i=1}^n (X_k^i)^2$, we see that $\mathbb{E}^{\mathbb{P}}[U^{(m)}] = n \sum_{k:\mathbb{P} \in \mathcal{P}_k} \sigma_k^2$. Moreover, by our assumptions and using the Marcinkiewicz–Zygmund inequality, there exists another constant C' that depends only on $\delta > 0$ such that:

$$\mathbb{E}^{\mathbb{P}}[U^{(m)}] \geq cnK_0(\mathbb{P}), \quad \mathbb{E}^{\mathbb{P}}[|U^{(m)} - \mathbb{E}^{\mathbb{P}}[U^{(m)}]|^{1+\delta}] \leq C'nK_0(\mathbb{P}).$$

Thus, since $nK_0(\mathbb{P}) \rightarrow \infty$ and using Markov's inequality, it follows that $U^{(m)}/\mathbb{E}^{\mathbb{P}}[U^{(m)}]$ converges in probability to 1 as desired. We can analogously argue that:

$$\frac{n \sum_{k:\mathbb{P} \in \mathcal{P}_k} (\hat{\sigma}_k^{(n)})^2}{n \sum_{k:\mathbb{P} \in \mathcal{P}_k} \sigma_k^2} \xrightarrow{\mathbb{P}} 1 \quad \text{as } m \rightarrow \infty.$$

Thus implies that $E_1^{(n)}, \dots, E_K^{(n)}$ are asymptotic compound e-variables.

Next fix $\mathbb{P} \in (\bigcup_k \mathcal{P}_k) \cap \mathcal{P}^{\mathcal{N}}$. It suffices to prove uniform integrability of the right-hand side in (10.7) for all m such that $(n-1)K_0(\mathbb{P}) \geq 5$ (see Proposition 10.7, part (iii)). We prove uniform integrability by controlling the following η -th moment for $\eta \in (1, 1.25)$:

$$\begin{aligned} & \mathbb{E}^{\mathbb{P}} \left[\left(\frac{\sum_{k:\mathbb{P} \in \mathcal{P}_k} (S_k^{(n)})^2}{\sum_{k:\mathbb{P} \in \mathcal{P}_k} (\hat{\sigma}_k^{(n)})^2} \right)^{\eta} \right] \\ & \leq \mathbb{E}^{\mathbb{P}} \left[\left(\sum_{k:\mathbb{P} \in \mathcal{P}_k} (S_k^{(n)})^2 \right)^{2\eta} \right]^{1/2} \mathbb{E}^{\mathbb{P}} \left[\left(\frac{1}{\sum_{k:\mathbb{P} \in \mathcal{P}_k} (\hat{\sigma}_k^{(n)})^2} \right)^{2\eta} \right]^{1/2}. \end{aligned}$$

Notice that by our assumptions, $\sigma_k^2 = \sigma_k^2(\mathbb{P})$ is both uniformly lower bounded and upper bounded for all k such that $\mathbb{P} \in \mathcal{P}_k$. Ignoring multiplicative constants, it suffices to bound:

$$\mathbb{E}^{\mathbb{P}} \left[\left(\sum_{k:\mathbb{P} \in \mathcal{P}_k} n \left(\frac{S_k^{(n)}}{\sigma_k} \right)^2 \right)^{2\eta} \right]^{1/2} \mathbb{E}^{\mathbb{P}} \left[\left(\frac{(n-1)^{-1}}{\sum_{k:\mathbb{P} \in \mathcal{P}_k} (\hat{\sigma}_k^{(n)}/\sigma_k^2)^2} \right)^{2\eta} \right]^{1/2}.$$

This can be accomplished by noting that,

$$\sum_{k:\mathbb{P} \in \mathcal{P}_k} n (S_k^{(n)}/\sigma_k)^2 \stackrel{d}{\sim} \chi_{nK_0(\mathbb{P})}^2, \quad \sum_{k:\mathbb{P} \in \mathcal{P}_k} (n-1) (\hat{\sigma}_k^{(n)}/\sigma_k)^2 \stackrel{d}{\sim} \chi_{(n-1)K_0(\mathbb{P})}^2,$$

where χ_d^2 denotes the chi-square distribution with d degrees of freedom. The remainder of the calculation amounts to direct evaluation of the (2η) -th moment (resp. (-2η) -th moment) of chi-squared random variables.

Bibliographical note

Materials in this chapter are largely based on [Ignatiadis et al. \[2024a\]](#). The notion of $(\varepsilon, 0)$ -approximate e-variable (not using this terminology) was considered by [Wang and Ramdas \[2022, Section 6.2\]](#). The standard examples following Proposition [10.9](#) in Section [10.2.1](#) can be found in standard textbooks [Van Der Vaart \[1998\]](#), [Lehmann and Romano \[2005\]](#). Example [10.12](#) is found in [Wang and Ramdas \[2023\]](#). The notion of compound p-variables was introduced by [Armstrong \[2022\]](#) under a different name (average significance controlling p-values and tests) and further studied in an empirical Bayes setting by [Ignatiadis and Sen \[2025\]](#). The notion of approximate compound e-variables appears implicitly in [Ignatiadis et al. \[2024b\]](#). A notion related to approximate FDR control was used as proof technique (without being given that name) by [Blanchard and Roquain \[2008\]](#). The construction in Section [10.4.4](#) appeared in [Ignatiadis et al. \[2024a,b\]](#) and pertains to the problem of conducting multiple t-tests.

Chapter 11

Combining e-values and using e-values as weights

In this chapter, we first present a classic result in optimal transport. Using this result, we prove Theorem 8.4, which characterizes the class of all e-merging functions. Then, we discuss how to combine an e-value and a p-value. Based on this combination, we illustrate how e-values can be used as weights in multiple testing procedures.

Throughout this chapter, $K \geq 2$ is a positive integer. We continue to fix one atomless probability measure $\mathbb{P}$ on some $(\Omega, \mathcal{F})$ as in Chapter 8.

11.1 Optimal transport duality

A technical tool that will be useful in two main proofs in Sections 11.2 and 12.2 is optimal transport duality. We state a classic version of multi-marginal optimal transport duality without proving it.

Let $F : [0, \infty)^K \rightarrow \mathbb{R}$ be a bounded Borel function. Denote by $\mathcal{B}$ the set of Borel functions on $[0, \infty)$, and define the operator $\bigoplus$ on $(\phi_1, \dots, \phi_K) \in \mathcal{B}^K$ by

$$\left(\bigoplus_{k=1}^K \phi_k \right) (x_1, \dots, x_K) := \sum_{k=1}^K \phi_k(x_k), \quad (x_1, \dots, x_K) \in [0, \infty)^K.$$

Define

$$D_F = \left\{ (\phi_1, \dots, \phi_K) \in \mathcal{B}^K : \bigoplus_{k=1}^K \phi_k \geq F \right\}.$$

Let $\Gamma(\mu_1, \dots, \mu_K)$ be the set of all Borel measures on $\mathbb{R}^K$ with marginal distributions $\mu_1, \dots, \mu_K$ on $\mathbb{R}$. In what follows, we always write $\phi = (\phi_1, \dots, \phi_K)$ and $\mu = (\mu_1, \dots, \mu_K)$.

Lemma 11.1: Optimal transport duality

For any bounded Borel function $F : [0, \infty)^K \rightarrow \mathbb{R}$ and distributions $\mu_1, \dots, \mu_K$ on $\mathbb{R}$,

$$\sup_{\pi \in \Gamma(\mu)} \int F d\pi = \inf_{\phi \in D_F} \sum_{k=1}^K \int \phi_k d\mu_k, \quad (11.1)$$

where in the infimum we only consider those with $\sum_{k=1}^K \int \phi_k d\mu_k$ well-defined. Moreover, if F is upper semicontinuous, then the infimum is attainable, and the functions $\phi_1, \dots, \phi_K$ in (11.1) can be chosen as upper semicontinuous. If F is decreasing (nonnegative), then $\phi_1, \dots, \phi_K$ can be chosen as decreasing (nonnegative). If F is bounded in $[0, 1]$, then $\phi_1, \dots, \phi_K$ can be chosen as bounded in $[0, 1]$.

At an abstract level, the reason why optimal transport duality appears in this chapter and the next one is that merging functions both for p-values and for e-values need to produce an output under arbitrary dependence, and optimization under arbitrary dependence is precisely the topic of optimal transport theory. In our context, it concerns multi-marginal optimal transport, to be precise.

11.2 Admissible e-merging functions

The main content of this section is to prove Theorem 8.4, recalled here.

Recalling Theorem 8.4

For a function $F : \mathbb{R}_+^K \rightarrow \mathbb{R}_+$,

- (i) if F is an e-merging function, then $F \leq \mathbb{M}_\lambda$ for some $\lambda \in \Delta_{K+1}$;
- (ii) F is an admissible e-merging function if and only if $F = \mathbb{M}_\lambda$ for some $\lambda \in \Delta_{K+1}$.

It suffices to show statement (i). Statement (ii) follows directly from (i) and the simple fact that $\mathbb{M}_\lambda$ is an e-merging function.

First, we characterize the simplest case $K = 1$, which will be used in the proof of Theorem 8.4. All e-variables in this section are for $\mathbb{P}$.

Lemma 11.2

Let $\theta \in [1, \infty]$ and $r \in \mathbb{R}_+$. If a function $g : [0, \theta] \rightarrow \mathbb{R}_+$ satisfies $\mathbb{E}^\mathbb{P}[g(E)] \leq r$ for all e-variables E taking values in $[0, \theta]$, then there exists $\lambda \in [0, 1]$ such that $g(x) \leq r(1 - \lambda + \lambda x)$ for $x \in [0, \theta] \cap \mathbb{R}$.

As a particular case of Lemma 11.2 with $r = 1$, if g is an e-merging function of dimension 1, then there exists $\lambda \in [0, 1]$ such that $g(x) \leq (1 - \lambda) + \lambda x$ for $x \in \mathbb{R}_+$. Hence, the desired statement in Theorem 8.4 holds for $K = 1$.

Lemma 11.2 has the following useful implication. If one is supplied with an e-variable E , but wants another e-value (hoping that it would work better than E in some context), then one has to use $(1 - \lambda) + \lambda E$ for some $\lambda \in [0, 1]$, as all other ways are dominated by this class. For instance, $\sqrt{E}$ is a valid e-variable for any $E \in \mathfrak{E}$, but Lemma 11.2 says that it is dominated by $(1 - \lambda) + \lambda E$ for some λ , which turns out to be $1/2$ in this case. This is the reason why there is no way of betting in each round other than using $(1 - \lambda_t + \lambda_t E_t)$ in Section 7.8 to build an e-process from given sequential e-variables $(E_t)_{t \in T_+}$.

Proof of Lemma 11.2.

The case $r = 0$ is trivial as g is the constant function 0. Otherwise, $r > 0$, and without loss of generality we can assume $r = 1$.

If $\theta = 1$, then $g(x) \leq 1$ for all x , and taking $\lambda = 0$ gives the desired inequality. In what follows, we assume $\theta \in (1, \infty]$, and all points x, y that appear below are in $[0, \theta] \cap \mathbb{R}$.

First, it is easy to note that $g(y) \leq y$ for $y > 1$; indeed, if $g(y) > y$ then taking a random variable X with $\mathbb{P}(X = y) = 1/y$ and 0 otherwise gives $\mathbb{E}^\mathbb{P}[g(X)] > 1$ and breaks the assumption. Moreover, $g(y) \leq 1$ for $y \leq 1$ is also clear, which in particular implies $g(1) \leq 1$.

Suppose for the purpose of contradiction that the statement in the lemma does not hold. This means that for each $\lambda \in [0, 1]$, either (a) $g(x) > (1 - \lambda) + \lambda x$ for some $x < 1$ or (b) $g(y) > (1 - \lambda) + \lambda y$ for some $y > 1$ (or both). Since $g(y) \leq y$ for $y > 1$ and $g(x) \leq 1$ for $x < 1$, we know that $\lambda = 1$ implies (a) and $\lambda = 0$ implies (b).

We claim that there exists $\lambda_0 \in (0, 1)$ for which both (a) and (b) happen. To show this claim, let

$$\begin{aligned}\Lambda_0 &= \{\lambda \in [0, 1] : g(y) > (1 - \lambda) + \lambda y \text{ for some } y > 1\}; \\ \Lambda_1 &= \{\lambda \in [0, 1] : g(x) > (1 - \lambda) + \lambda x \text{ for some } x < 1\}.\end{aligned}$$

Clearly, the above arguments show $\Lambda_0 \cup \Lambda_1 = [0, 1]$, $0 \in \Lambda_0$, and $1 \in \Lambda_1$. Moreover, since the function $\lambda \mapsto (1 - \lambda) + \lambda x$ is monotone for either $x < 1$ or $x > 1$, we know that both Λ_0 and Λ_1 are intervals. Let $\lambda_* = \sup \Lambda_0$ and $\lambda^* = \inf \Lambda_1$. We will argue $\lambda_* \notin \Lambda_0$ and $\lambda^* \notin \Lambda_1$. If $\lambda_* \in \Lambda_0$, then there exists $y > 1$ such that $g(y) > (1 - \lambda_*) + \lambda_* y$. By continuity, there exists $\hat{\lambda}_* > \lambda_*$ such that $g(y) > (1 - \hat{\lambda}_*) + \hat{\lambda}_* y$, showing that $\hat{\lambda}_* \in \Lambda_0$, contradicting the definition of λ_* . Therefore, $\lambda_* \notin \Lambda_0$. Similarly, $\lambda^* \notin \Lambda_1$, following the same argument. If $\lambda_* = \lambda^*$, then this point is not contained in $\Lambda_0 \cup \Lambda_1$, a contradiction to $\Lambda_0 \cup \Lambda_1 = [0, 1]$. Hence, it must be $\lambda_* > \lambda^*$, which implies that $\Lambda_0 \cap \Lambda_1$ is not empty.

Let $x_0 < 1$ and $y_0 > 1$ be such that

$$g(x_0) > 1 - \lambda_0 + \lambda_0 x_0 \quad \text{and} \quad g(y_0) > 1 - \lambda_0 + \lambda_0 y_0.$$

Let X be distributed as $\mathbb{P}(X = y_0) = (1 - x_0)/(y_0 - x_0)$ and $\mathbb{P}(X = x_0) = (y_0 - 1)/(y_0 - x_0)$, which clearly satisfies $\mathbb{E}^\mathbb{P}[X] = 1$ and is binary. Moreover,

$$\begin{aligned}\mathbb{E}^\mathbb{P}[g(X)] &= \frac{1 - x_0}{y_0 - x_0} g(y_0) + \frac{y_0 - 1}{y_0 - x_0} g(x_0) \\ &> \frac{1 - x_0}{y_0 - x_0} (1 - \lambda_0 + \lambda_0 y_0) + \frac{y_0 - 1}{y_0 - x_0} (1 - \lambda_0 + \lambda_0 x_0) = 1.\end{aligned}$$

This yields a contradiction.

In what follows, $\mathbf{0}$ and $\mathbf{1}$ represent a vector of zeros and a vector of ones of the appropriate dimension, respectively. The next lemma allows us to only consider upper semicontinuous e-merging functions.

Lemma 11.3

If F is an e-merging function, then its upper semicontinuous version F^* is given by

$$F^*(\mathbf{e}) = \lim_{\varepsilon \downarrow 0} F(\mathbf{e} + \varepsilon \mathbf{1}), \quad \mathbf{e} \in [0, \infty)^K;$$

is also an e-merging function, and $F^* \geq F$.

Proof.

Take any vector $\mathbf{E}$ of e-variables. For every rational $\varepsilon \in (0, 1)$, let A_ε be an event independent of $\mathbf{E}$ with $\mathbb{P}(A_\varepsilon) = 1 - \varepsilon$, and $\mathbf{E}_\varepsilon = (\mathbf{E} + \varepsilon \mathbf{1}) \mathbf{1}_{A_\varepsilon}$ (here we use the convention that $\mathbf{E}_\varepsilon = \mathbf{0}$ if the event A_ε does not occur). For each ε , $\mathbb{E}[\mathbf{E}_\varepsilon] \leq (1 - \varepsilon)(\mathbf{1} + \varepsilon \mathbf{1}) \leq \mathbf{1}$. Therefore, $\mathbf{E}_\varepsilon$ is a vector of e-variables, and hence

$$1 \geq \mathbb{E}[F(\mathbf{E}_\varepsilon)] = (1 - \varepsilon) \mathbb{E}[F(\mathbf{E} + \varepsilon \mathbf{1})] + \varepsilon F(\mathbf{0}),$$

which implies

$$\mathbb{E}[F(\mathbf{E} + \varepsilon \mathbf{1})] \leq \frac{1 - \varepsilon F(\mathbf{0})}{1 - \varepsilon}.$$

Fatou's lemma yields

$$\begin{aligned}\mathbb{E}[F^*(\mathbf{E})] &= \mathbb{E}\left[\lim_{\varepsilon \downarrow 0} F(\mathbf{E} + \varepsilon \mathbf{1})\right] \\ &\leq \lim_{\varepsilon \downarrow 0} \mathbb{E}[F(\mathbf{E} + \varepsilon \mathbf{1})] \leq \lim_{\varepsilon \downarrow 0} \frac{1 - \varepsilon F(\mathbf{0})}{1 - \varepsilon} = 1.\end{aligned}$$

Therefore, F^* is an e-merging function. The domination $F^* \geq F$ follows from the fact that F is increasing.

Proof of Theorem 8.4

By Lemma 11.3, it suffices to consider an upper semicontinuous e-merging function F . Let $\mathcal{M}_{\mathcal{E}}$ be the set of all distributions on $\mathbb{R}_+$ with mean no larger than 1, i.e., the set of all distributions of e-variables, and $\mathcal{M}_{\mathcal{E}}^{\theta}$ be the subset of $\mathcal{M}_{\mathcal{E}}$ containing all distributions on $[0, \theta]$ for $\theta \geq 1$.

Fix $\theta \geq 1$. Since F is bounded and nonnegative on $[0, \theta]^K$ and upper semicontinuous, using Lemma 11.1 we get

$$\begin{aligned}\sup_{\boldsymbol{\mu} \in (\mathcal{M}_{\mathcal{E}}^{\theta})^K} \sup_{\pi \in \Gamma(\boldsymbol{\mu})} \int F d\pi &= \sup_{\boldsymbol{\mu} \in (\mathcal{M}_{\mathcal{E}}^{\theta})^K} \inf_{\boldsymbol{\phi} \in D_F} \sum_{k=1}^K \int \phi_k d\mu_k \\ &= \sup_{\boldsymbol{\mu} \in (\mathcal{M}_{\mathcal{E}}^{\theta})^K} \inf_{\boldsymbol{\phi} \in D_F^+} \sum_{k=1}^K \int \phi_k d\mu_k,\end{aligned}\tag{11.2}$$

where D_F^+ is the subset of D_F containing all $\boldsymbol{\phi}$ with nonnegative and upper semicontinuous components. Define the mapping

$$J : (\boldsymbol{\mu}, \boldsymbol{\phi}) \mapsto \sum_{k=1}^K \int \phi_k d\mu_k.$$

We will verify a few conditions, which allows us to apply a minimax theorem.

- (i) The mapping J is bilinear, and therefore both convex and concave.
- (ii) Since $[0, \theta]$ is compact, the set $(\mathcal{M}_{\mathcal{E}}^{\theta})^K$ equipped with the weak topology is tight. By Prokhorov's theorem, it is sequentially compact, and hence compact.
- (iii) Since each component of $\boldsymbol{\phi}$ is upper semicontinuous, $\boldsymbol{\mu} \mapsto J(\boldsymbol{\mu}, \boldsymbol{\phi})$ is upper semicontinuous with respect to weak convergence in $\boldsymbol{\mu}$ for each $\boldsymbol{\phi}$.

The above conditions allow us to use Fan [1953, Theorem 2] to conclude

$$\sup_{\boldsymbol{\mu} \in (\mathcal{M}_{\mathcal{E}}^{\theta})^K} \inf_{\boldsymbol{\phi} \in D_F^+} \sum_{k=1}^K \int \phi_k d\mu_k = \inf_{\boldsymbol{\phi} \in D_F^+} \sup_{\boldsymbol{\mu} \in (\mathcal{M}_{\mathcal{E}}^{\theta})^K} \sum_{k=1}^K \int \phi_k d\mu_k.\tag{11.3}$$

Since F is an e-merging function, each term in (11.2) is bounded by 1. Using (11.3), we get

$$1 \geq \inf_{\boldsymbol{\phi} \in D_F^+} \sup_{\boldsymbol{\mu} \in (\mathcal{M}_{\mathcal{E}}^{\theta})^K} \sum_{k=1}^K \int \phi_k d\mu_k = \inf_{\boldsymbol{\phi} \in D_F^+} \sum_{k=1}^K \sup_{\mu_k \in \mathcal{M}_{\mathcal{E}}^{\theta}} \int \phi_k d\mu_k.\tag{11.4}$$

Denote by $T_\phi = \sup_{\mu \in \mathcal{M}_\varepsilon^e} \int \phi d\mu$ for $\phi \in \mathcal{B}$. For any $\varepsilon > 0$, by (11.4), we can find $(\phi_1, \dots, \phi_K) \in D_F^+$ such that $\sum_{k=1}^K T_{\phi_k} \leq 1 + \varepsilon$. Using Lemma 11.2, for each $k \in \mathcal{K}$, there exists a constant $h_{\phi_k} \in [0, 1]$ such that

$$\phi_k(x) \leq T_{\phi_k} (1 - h_{\phi_k} + h_{\phi_k} x) \text{ for all } x \in [0, \theta].$$

Since $(\phi_1, \dots, \phi_K) \in D_F^+$, we have

$$F(x_1, \dots, x_K) \leq \sum_{k=1}^K T_{\phi_k} (1 - h_{\phi_k} + h_{\phi_k} x_k) \text{ for all } x_1, \dots, x_K \in [0, \theta].$$

This means that there exists $\lambda_{\varepsilon, \theta} \in \Delta_{K+1}$ such that $F \leq (1 + \varepsilon) \mathbb{M}_{\lambda_{\varepsilon, \theta}}$ on $[0, \theta]^K$. Since Δ_{K+1} is compact, we can find a limit $\lambda_0 \in \Delta_{K+1}$ of some subsequence of $\lambda_{\varepsilon, \theta}$ as $\varepsilon \downarrow 0$ and $\theta \rightarrow \infty$. Continuity of $\lambda \mapsto \mathbb{M}_\lambda$ yields $F \leq \mathbb{M}_{\lambda_0}$ on $\mathbb{R}_+^K$, thus the desired statement.

11.3 Cross-merging: merging an e-value and a p-value

Here we consider the setting where we have one e-variable E and one p-variable P for the same hypothesis. This will help to design and understand procedures when both p-values and e-values are available, treated in Section 11.4.

Propositions 2.4 and 2.5 in Chapter 2 identify all admissible calibrators between p-values and e-values. We consider four cases of merging a p-value P and an e-value E : whether P and E are independent, and whether the output is a p-value or an e-value.

To discuss these cases, we define combiners, in a similar way as calibrators and merging functions.

Definition 11.4

A function $f : [0, 1] \times [0, \infty] \rightarrow [0, \infty]$ is called an i-pe/e combiner if $f(P, E)$ is an e-value for any independent p-value P and e-value E , and $(p, e) \mapsto f(p, e)$ is decreasing in p and increasing in e . Similarly, we define i-pe/p, pe/p, and pe/e combiners, where i indicates independence, and p and e are self-explanatory. If the output is a p-value, the combiner is increasing in p and decreasing in e .

In what follows we omit “pe” in “i-pe/e” and other terms to be concise. We provide four natural combiners to the above four cases, some relying on an admissible calibrator h .

- [ie] Return $h(P)E$ by using the function $C_h^{\text{ie}} : (p, e) \mapsto h(p)e$. The convention here is $0 \times \infty = \infty$.
- [ip] Return P/E , capped at 1, by using the function $C^{\text{ip}} : (p, e) \mapsto (p/e) \wedge 1$.
- [e] Return $\lambda h(P) + (1 - \lambda)E$ by using the function $C_{\lambda, h}^e : (p, e) \mapsto \lambda h(p) + (1 - \lambda)e$ for some $\lambda \in (0, 1)$.
- [p] Return $2 \min(P, 1/E)$, capped at 1, by using the function $C^p : (p, e) \mapsto (2(p \wedge e^{-1})) \wedge 1$.

The superscript in the notation C^{ie} suggests that the combiner assumes independence and outputs an e-value; the other cases are similar. The combiners C_h^{ie} and $C_{\lambda, h}^e$ depend on h whereas C^{ip} and C^p do not, similarly to the situation of calibrators: There are many more choices when the output is an e-value, compared to the case when the output is a p-value. For the combiner $C_{\lambda, h}^e$, it may be convenient to choose $\lambda = 1/2$, so that $C_{\lambda, h}^e(P, E)$ is the arithmetic average of two e-values $h(P)$ and E , although this choice has no special advantage, since the positions of $h(P)$ and E are not symmetric.

Validity of these combiners is straightforward to verify. Admissibility of the above combiners is obtained in the following result.

Theorem 11.5

For an admissible calibrator h and $\lambda \in (0, 1)$, C_h^{ie} is an admissible i-pe/e combiner, C^{ip} is an admissible i-pe/p combiner, $C_{\lambda, h}^{\text{e}}$ is an admissible pe/e combiner, and C^{p} is an admissible pe/p combiner.

The most useful combiner is the i-pe/p combiner C^{ip} that produces a p-value based on independent P and E . We provide a proof for the statement on C^{ip} in Theorem 11.5 and omit the rest of the proof.

Proof of the [ip] case in Theorem 11.5.

Checking that C^{ip} is a valid i-pe/p combiner is simple (seen already in Corollary 2.10 in Section 2.4): For $t \in (0, 1)$ and any independent pair (P, E) of p-variable and e-variable for $\mathbb{P}$, we have

$$\begin{aligned}\mathbb{P}(C^{\text{ip}}(P, E) \leq t) &= \mathbb{P}(P/E \leq t) \\ &= \mathbb{P}(P \leq tE) = \mathbb{E}^{\mathbb{P}}[\mathbb{P}(P \leq tE \mid E)] \leq \mathbb{E}^{\mathbb{P}}[tE] \leq t.\end{aligned}$$

To show its admissibility, suppose for the purpose of contradiction that an i-pe/p combiner f satisfies $f \leq C^{\text{ip}}$ and $f(p, e) < (p/e) \wedge 1$ for some $(p, e) \in [0, 1] \times [0, \infty)$. Since $a \mapsto f(p, a)$ is decreasing, we can assume $e \in [p, \infty)$ by replacing e with p if $e < p$. Take $P \stackrel{\text{d}}{\sim} \text{U}[0, 1]$ and E distributed as $\mathbb{P}(E = e) = 1/e = 1 - \mathbb{P}(E = 0)$ if $e \geq 1$ and $\mathbb{P}(E = e) = 1/2 = \mathbb{P}(E = 2 - e)$ if $e < 1$. It is clear that $\mathbb{E}^{\mathbb{P}}[E] = 1$.

Since $q \mapsto f(q, e)$ is increasing, there exists $p' < p$ such that $f(q, e) \leq f(p, e) < p'/e$ for all $q \in [0, p]$. This gives

$$\mathbb{P}(f(P, e) \leq p'/e) \geq \mathbb{P}(P \leq p) = p.$$

For $\alpha = p'/e \in (0, 1)$, if $e \geq 1$, then

$$\begin{aligned}\mathbb{P}(f(P, E) \leq \alpha) &= \mathbb{P}(f(P, e) \leq p'/e)e^{-1} + \mathbb{P}(f(P, 0) \leq \alpha)(1 - e^{-1}) \\ &\geq p/e > \alpha.\end{aligned}$$

If $e < 1$, then

$$\begin{aligned}\mathbb{P}(f(P, E) \leq \alpha) &= \frac{1}{2}\mathbb{P}(f(P, e) \leq p'/e) + \frac{1}{2}\mathbb{P}(f(P, 2 - e) \leq \alpha) \\ &\geq \frac{1}{2}p + \frac{1}{2}\mathbb{P}(P \leq \alpha(2 - e)) \\ &= \frac{1}{2}(p + (2 - e)p'/e) = \frac{1}{2}(p - p') + p'/e > \alpha.\end{aligned}$$

Hence, $f(P, E)$ is not a p-value, and this contradicts the fact that f is an i-pe/p combiner. This contradiction shows that C^{ip} is admissible.

We will call C^{ip} the *quotient combiner* because it outputs the quotient of P and E (capped at 1). This combiner typically leads to more powerful procedures compared to the other combiners and provides the foundation for our insight that e-values can act as unnormalized weights in multiple testing in Section 11.4.

Example 11.6: Randomization

Consider an e-variable E for $\mathbb{P}$ and generate an independent uniform variable $U \sim \text{U}[0, 1]$. Then, using the quotient combiner, U/E is a p-variable that satisfies $U/E \leq 1/E$. This implies that the unique admissible e-to-p calibrator $f : e \mapsto (1/e) \wedge 1$ can be improved if we allow for randomization, as studied already in Section 2.4. Although U/E may not be practical in general due to external randomization, it becomes practical when applied to a p-value P (computed from data) that is independent of E .

The quotient combiner is a useful general-purpose method for meta-analysis from two independent datasets. The quotient combiner is immediately applicable when the researcher summarizes the first dataset as a single p-value, and the second dataset as a single e-value. Such a situation could occur when the second dataset is collected in such a way, e.g., with optional stopping and continuation, that inference is more natural with e-values. It could also be the case that one dataset comprises of a large sample size, allowing for asymptotic approximations to compute p-values, while the second dataset is smaller and may require finite-sample inference methods.

The quotient combiner is also useful when the above data constraints are not in place and the researcher can in principle compute both a p-value P' and an e-value E on the second dataset, both of which are independent of the p-value P computed on the first dataset. In that case, the researcher could apply a p-value combination method based on P and P' , e.g., Fisher's combination $P_F := 1 - \chi_4(-2\log(PP'))$, where χ_4 is the cdf of the chi-square distribution with 4 degrees of freedom. However, the researcher may still prefer to proceed with the quotient combiner P/E (we omit the cap at 1 for simplicity, which does not affect testing results). We suggest the following rule of thumb:

The Fisher combination is preferable to the quotient combiner under dataset exchangeability: Suppose that the analyst considers the two datasets as a-priori exchangeable. In that case, it may be undesirable to use an asymmetric combination rule such as P/E , and Fisher's combination P_F is preferable on conceptual grounds. If the two datasets are also exchangeable in terms of their statistical properties (i.e., they have similar power), then P_F will typically have higher power than P/E .

The quotient combiner is preferable to the Fisher combination for imbalanced datasets: When one dataset (the “primary” dataset) is substantially more well-powered (larger anticipated signal or sample size) than the secondary dataset, and the investigator knows which dataset is more well-powered, then the P/E combiner can often outperform Fisher's combination test in terms of power. A proviso is that the p-value is computed on the primary (more well-powered) dataset and the e-value on the secondary dataset.

Example 11.7: A stylized example illustrating the rule of thumb

Suppose we have two independent samples of iid data points, $\mathbf{X} = (X_1, \dots, X_m)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$, both from a distribution $\mathbb{P}$, where $n \geq m \geq 1$. We seek to test $H_0 : \mathbb{P} = N(0, 1)$ against $H_1 : \mathbb{P} = N(\delta, 1)$, where $\delta > 0$ is known. The optimal (Neyman–Pearson) p-value based on $\mathbf{X}$ is given by $P_{\mathbf{X}} := 1 - \Phi(T_{\mathbf{X}})$, where Φ is the standard normal distribution function and $T_{\mathbf{X}} := m^{-1/2} \sum_{i=1}^m X_i$. Analogously we may compute p-values $P_{\mathbf{Y}}$ based on $\mathbf{Y}$, as well as $P_{\mathbf{Z}}$, where $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ is the full dataset. The optimal e-value $E_{\mathbf{X}}$ based on $\mathbf{X}$ is the likelihood ratio of $N(\delta, 1)^m$ over $N(0, 1)^m$.

By the Neyman–Pearson Lemma, the most powerful test is the likelihood ratio test, which coincides with a test based on the p-value $P_{\mathbf{Z}}$. We compare $P_{\mathbf{Z}}$ against the quotient combiner $P_E := P_{\mathbf{Y}}/E_{\mathbf{X}}$ and P_F by considering the hypothesis tests that reject H_0 when the p-value is smaller than or equal to some $\alpha > 0$. Omitting details, we report the following conclusions.

(i) Theoretically, using the notion of asymptotic relative efficiency, we can prove that P_E is as efficient as $P_{\mathbf{Z}}$ in two asymptotically settings: (a) $\alpha \rightarrow 0$, that is, when the type-I error is very stringent, and (b) $m/n \rightarrow 0$, that is, when $\mathbf{Y}$ is substantially more informative than $\mathbf{X}$.

(ii) A small simulation study (Fig. 11.8) yields the following observations: (a) if m/n is small, then P_E has almost the same power as $P_{\mathbf{Z}}$, and both outperform the P_F ; (b) when $m = n$, P_F has slightly more power than P_E , and both have slightly less power than the $P_{\mathbf{Z}}$.

These observations also confirm that, when using one p-value and one e-value, the primary dataset should be used to generate the p-value.

11.4 E-values as weights in p-value-based multiple testing

In this section, we study how the cross-merging methods in Section 11.3 can be useful in the context of multiple testing in Chapter 9.

The basic setting is that a vector $\mathbf{P} = (P_1, \dots, P_K)$ of p-values and a vector $\mathbf{E} = (E_1, \dots, E_K)$ of e-values

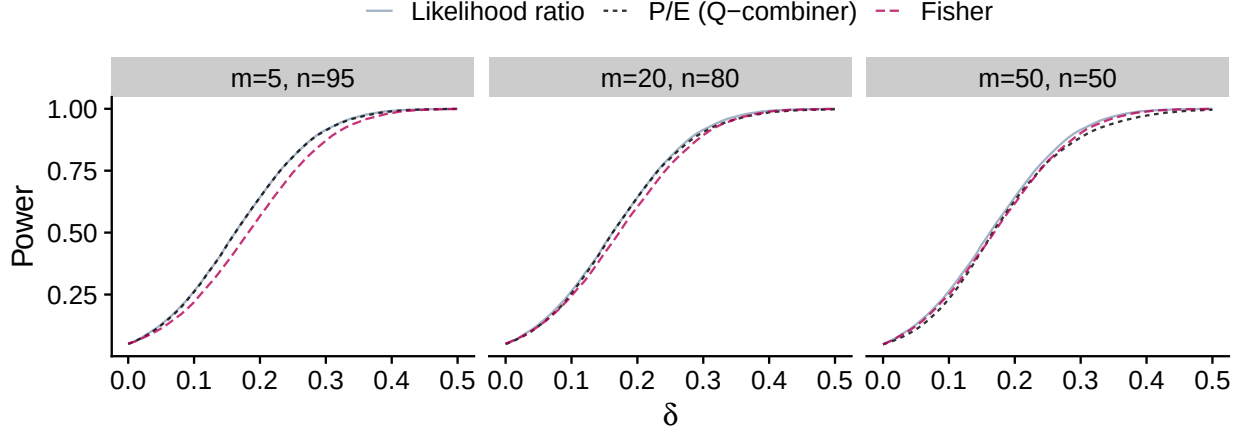


Figure 11.8: **Simulation study for a meta-analysis combining two samples:** We compare the likelihood ratio test, the quotient combiner (Q-combiner), and the Fisher combination test, plotting power against signal strength δ . The panels correspond to different choices of the two sample sizes m and n . The quotient combiner is visibly more powerful on the left (matching the likelihood ratio), and Fisher's combination is marginally better on the right.

are available for the hypotheses $H_1, \dots, H_K$. More precisely, P_k is a p-variable for H_k , and E_k is an e-variable for H_k for each $k \in \mathcal{K}$. A useful context could be that e-values are obtained from preliminary data analysis and p-values are obtained from follow-up confirmatory research on these hypotheses (recall from Example 11.7 that, in the context of testing a single hypothesis, we would like to choose the primary dataset to generate the p-value). We allow for some hypotheses with missing p-value or e-value, and in these cases P_k and E_k are set to 1. The interpretation is that these hypotheses miss either a preliminary or a follow-up analysis.

Definition 11.9: E-weighted p-value procedure (ep- $\mathcal{D}$)

Let $\mathbf{p}\text{-}\mathcal{D}$ be a multiple testing procedure based on p-values. We define the e-weighted p-value procedure $\mathbf{ep}\text{-}\mathcal{D}$ which proceeds as follows: for $k \in \mathcal{K}$, compute the quotient combiner $P_k^* := (P_k/E_k) \wedge 1$, and then supply $\mathbf{P}^*$ to $\mathbf{p}\text{-}\mathcal{D}$.

We mainly focus on the e-weighted version of the BH procedure in Section 9.1.4, which we call the ep-BH procedure. Recall that the BH procedure at level α has valid FDR control in case the p-values are PRDS, as mentioned in Section 9.1.4. In the e-BH procedure, we interpret the e-values as weights for the p-values. Intuitively, if $E_k > 1$, then there is evidence against H_k being a null, and we have $P_k/E_k < P_k$ (assuming $P_k \neq 0$), that is, the weight strengthens the signal of P_k . Conversely, if $E_k < 1$, then there is no evidence against H_k being a null, and we have $P_k/E_k > P_k$. The above interpretation of e-values as weights is quite natural, and the perspective is useful in deriving guarantees. Below we prove that ep-BH controls the FDR under the assumption that $\mathbf{P}$ is PRDS and $\mathbf{P}$ and $\mathbf{E}$ are independent. Note that these two conditions do not imply PRDS of $\mathbf{P}^*$, and hence some arguments are needed to establish the FDR control of the ep-BH procedure. The following result integrates over the randomness in the weights, i.e., the e-values.

Theorem 11.10

Suppose that $\mathbf{P}$ is independent of $\mathbf{E}$ and that $\mathbf{P}$ satisfies PRDS (Definition 9.15). Then, the ep-BH procedure has FDR at most $K_0\alpha/K$.

Proof.

Let $\text{ep-}\mathcal{D}$ be the ep-BH procedure at level α . Since $\mathbf{E}$ is independent of $\mathbf{P}$, conditional on $\mathbf{E}$, the ep-BH procedure becomes a weighted p-BH procedure with weight vector $\mathbf{E}$ applied to the p-values $\mathbf{P}$ that are PRDS. Using well-known existing results on the false discovery rate of the weighted p-BH procedure (e.g., [Ramdas et al., 2019](#), Theorem 1), we get

$$\mathbb{E} \left[\frac{F_{\text{ep-}\mathcal{D}}}{R_{\text{ep-}\mathcal{D}}} \mid \mathbf{E} \right] \leq \frac{1}{K} \sum_{k \in \mathcal{N}} E_k \alpha.$$

Hence, by iterated expectation, the FDR of $\text{ep-}\mathcal{D}$ is guaranteed to be at most $\mathbb{E}[\sum_{k \in \mathcal{N}} E_k \alpha / K] \leq K_0 \alpha / K$.

Compared with the weighted BH procedure using constant weights, the ep-BH procedure uses weights that do not necessarily add up to K , i.e., they are not normalized. This allows us to save evidence obtained from the preliminary analysis using e-values. Moreover, the above result does not require any assumption about the dependence within $\mathbf{E}$.

There are many other types of procedures that take p-values as input, and they can be generalized with e-values as weights.

Bibliographical note

The optimal transport duality statements in Lemma 11.1 can be verified with several results in [Kellerer \[1984\]](#); the formula (11.1) can be found in [Rachev and Rüschendorf \[2006, Theorem 2.1.1\]](#) and [Rüschendorf \[2013, Theorem 2.3\]](#). A standard textbook on optimal transport theory is [Villani \[2009\]](#). [Vovk and Wang \[2021\]](#) identified all symmetric admissible e-merging functions, but not the non-symmetric ones. Theorem 8.4 as presented in Section 11.2 was obtained by [Wang \[2025\]](#).

The content of Sections 11.3–11.4 is based on [Ignatiadis et al. \[2024b\]](#). The full proof of Theorem 11.5 can be found in [Ignatiadis et al. \[2024b, Appendix S1\]](#). Using deterministic weights for the BH procedure to control FDR was studied by [Benjamini and Hochberg \[1997\]](#). The e-weighted BH procedure in Section 11.4 was introduced by [Ignatiadis et al. \[2024b\]](#), along with results on many other procedures weighted by e-values.

Chapter 12

Using e-values to combine dependent p-values

In this chapter we study merging functions of p-values, and show how e-values appear as an essential step in such merging functions. The simplest form of merging is to take an average, which we study first. We turn to the more general deterministic methods for arbitrarily dependent p-values, followed by methods based on the assumption of exchangeability among p-values, and methods based on randomization.

As in Chapter 11, in most parts of this chapter, we consider a fixed positive integer $K \geq 2$ and an atomless probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ without further specification. In a few instances, we mention “for all $\mathbb{P} \in \mathcal{P}$ ” or “for all $K \geq 2$ ”, and thus temporarily allowing $\mathbb{P}$ and K to vary, but it should be clear from context.

12.1 The arithmetic mean of p-values

12.1.1 Average p-values

Intuitively, an average p-value should be the arithmetic mean of several p-values. The distribution of such an arithmetic mean may depend on how the underlying null hypothesis $\mathcal{P}$ is specified, for instance, simple or composite. We would like to define an average p-value based on its distribution under each $\mathbb{P}$, as in the cases of p-values and e-values. This issue is clarified in the next proposition. We will use the notation P for average p-variables, as they share many properties with p-variables.

Proposition 12.1

For a random variable P and an atomless probability measure $\mathbb{P}$, the following are equivalent:

- (i) P is the arithmetic mean of finitely many p-variables for $\mathbb{P}$;
- (ii) P is a (possibly continuous) mixture of p-variables for $\mathbb{P}$;
- (iii) $P \geq \mathbb{E}^{\mathbb{P}}[U|f(U)]$ for some measurable function f and a random variable U whose distribution is standard uniform under $\mathbb{P}$;
- (iv) $\mathbb{E}^{\mathbb{P}}[g(P)] \geq \int_0^1 g(u)du$ for all increasing concave functions $g : \mathbb{R} \rightarrow \mathbb{R}$.
- (v) $\mathbb{E}^{\mathbb{P}}[(v - P)_+] \leq v^2/2$ for all $v \in (0, 1)$.

Without assuming that $\mathbb{P}$ is atomless, we have (i) $\Rightarrow$ (ii) $\Rightarrow$ (iv) $\Leftrightarrow$ (v).

Proof.

We will show (i) $\Rightarrow$ (ii) $\Rightarrow$ (iv) $\Leftrightarrow$ (v) without assuming that $\mathbb{P}$ is atomless and (iii) $\Rightarrow$ (iv). The two remaining directions (v) $\Rightarrow$ (i) and (v) $\Rightarrow$ (iii) involve advanced results and are briefly explained.

(i) $\Rightarrow$ (ii) follows by definition. (ii) $\Rightarrow$ (iv) follows from Jensen's inequality: For a probability measure ν on a space Θ and a set $\{P_\theta : \theta \in \Theta\}$ of p-variables, we have

$$\mathbb{E}^\mathbb{P} \left[g \left(\int_\Theta P_\theta \nu(d\theta) \right) \right] \geq \mathbb{E}^\mathbb{P} \left[\int_\Theta g(P_\theta) \nu(d\theta) \right] \geq \mathbb{E}^\mathbb{P}[g(U)],$$

where we interchanged the expectation and the integral by Fubini's theorem, noting that g is bounded from below on $[0, \infty)$. (iv) $\Rightarrow$ (v) is straightforward by taking $g(x) = -(v - x)_+$, and (v) $\Rightarrow$ (iv) is due to the well-known fact that any increasing concave functions on $(0, 1)$ belong to the convex hull of $-(v - x)_+ + b$ for $v \in \mathbb{R}$ and $b \in \mathbb{R}$. (iii) $\Rightarrow$ (iv) follows from conditional Jensen's inequality:

$$\mathbb{E}^\mathbb{P}[g(P)] \geq \mathbb{E}^\mathbb{P}[g(\mathbb{E}^\mathbb{P}[U|f(U)])] \geq \mathbb{E}^\mathbb{P}[g(U)].$$

The implication (v) $\Rightarrow$ (i) relies on the following nontrivial fact: the average of three standard uniform random variables can attain any distribution satisfying (v) with mean $1/2$. The implication (v) $\Rightarrow$ (iii) relies on a refinement of Strassen's theorem. For these two directions, see the bibliographic note.

Proposition 12.1 leads to the following definition of average p-values purely based on their distributions under the null.

Definition 12.2: Average p-values

A random variable P is an *average p-variable* for $\mathcal{P}$ if $\mathbb{E}^\mathbb{P}[(v - P)_+] \leq v^2/2$ for all $v \in (0, 1)$ for all $\mathbb{P} \in \mathcal{P}$.

It is straightforward to check that a standard uniform random variable satisfies each condition in Proposition 12.1, and hence a p-variable is also an average p-variable. Moreover, for any p-variables $P_1, \dots, P_K$ for $\mathcal{P}$, their arithmetic mean is an average p-variable, regardless of whether $\mathcal{P}$ is simple or composite. Similarly to the case of p-variables, a truncation of an average p-variable at 1 is still an average p-variable.

Remark 12.3. Condition (iv) of Proposition 12.1 is known as second-order stochastic dominance in economics. Indeed, p-variables can be defined using the so-called first-order stochastic dominance; that is, P is a p-variable for $\mathbb{P}$ if and only if $\mathbb{E}^\mathbb{P}[g(P)] \geq \mathbb{E}^\mathbb{P}[g(U)]$ for all increasing functions $g : \mathbb{R} \rightarrow \mathbb{R}$ (this was briefly mentioned in Section 1.4). Clearly, this condition is stronger than condition (iv).

All results in this and the next few subsections hold for both simple hypotheses and composite hypotheses, and we will focus on the case of a fixed atomless $\mathbb{P}$ from now on.

As a simple consequence of Proposition 12.1, the set of average p-variables for $\mathbb{P}$ is convex and closed under convergence in distribution, whereas the set of p-variables is not convex. Indeed, the set of average p-variables is the convex hull of the set of p-variables.

Although a p-variable is also an average p-variable, the converse is not true. Indeed, even the arithmetic mean of iid standard uniform p-variables is not a p-variable, as it is asymptotically concentrated around $1/2$ (by the law of large numbers), conflicting the definition of a p-variable. It turns out that a factor of 2 will convert an average p-variable into a p-variable.

Proposition 12.4

An average p-variable is not necessarily a p-variable, but twice an average p-variable is a p-variable. Moreover, the factor 2 cannot be improved; that is, for any $\gamma \in [0, 2)$, γP is not a p-variable for some average p-variable P .

Proof.

The first statement is explained above. For the second statement, using Proposition 12.1, we can write an average p-variable P for $\mathbb{P}$ as $P = (P_1 + \dots + P_K)/K$, where $P_1, \dots, P_K$ are p-variables for $\mathbb{P}$ and $K \geq 2$. Take any $\alpha \in (0, 1)$. Note that $p \mapsto (2 - 2p)_+$ is a calibrator. By Proposition 2.6, $p \mapsto \gamma(2 - 2\gamma p)_+$ is a calibrator for any $\gamma \geq 1$. This implies that $E_k := (2 - 2P_k/\alpha)_+/\alpha$ is an e-variable for each $k \in \mathcal{K}$ and $\alpha \in (0, 1)$. We have

$$\begin{aligned} \mathbb{P}\left(\frac{2(P_1 + \dots + P_K)}{K} \leq \alpha\right) &= \mathbb{P}\left(\frac{1}{K} \sum_{k=1}^K \left(2 - \frac{2P_k}{\alpha}\right) \geq 1\right) \\ &= \mathbb{P}\left(\frac{1}{K} \sum_{k=1}^K \frac{1}{\alpha} \left(2 - \frac{2P_k}{\alpha}\right) \geq \frac{1}{\alpha}\right) \\ &\leq \mathbb{P}\left(\frac{1}{K} \sum_{k=1}^K E_k \geq \frac{1}{\alpha}\right) \leq \alpha, \end{aligned}$$

where the last inequality is due to Markov's inequality.

For the last statement, take a random variable U that is uniformly distributed on $[0, 1]$ under $\mathbb{P}$. Clearly, U and $1 - U$ are p-variables, and hence $1/2$ is an average p-variable. The constant $\gamma/2$ is not a p-variable for any $\gamma < 2$.

Proposition 12.4 implies that for any collection of p-variables, twice their arithmetic mean is a p-variable. Similarly, twice their median is a p-variable, which we will study later (see Example 12.11).

12.1.2 Special cases: mid p-values and posterior predictive p-values

In addition to the natural appearance as the arithmetic mean of p-values, another useful class of average p-values is that of *mid p-values*, which are computed based on discrete test statistics. We first recall the usual practice to obtain p-values for a simple hypothesis $\mathbb{P}$, mentioned in Fact 1.3 in Section 1.4. Let T be a test statistic which is a function of the observed data. Here, a smaller value of T represents stronger evidence against the null hypothesis. The p-variable P is usually computed from the conditional probability

$$P = \mathbb{P}(T' \leq T|T) = F(T), \quad (12.1)$$

where F is the distribution of T , and T' is an independent copy of T ; here and below, a copy of T is a random variable identically distributed as T .

If T has a continuous distribution under $\mathbb{P}$, then the p-variable P defined by (12.1) has a standard uniform distribution. If the test statistic T is discrete, e.g., when testing a binomial model, P is strictly stochastically larger than a uniform random variable on $[0, 1]$.

The discreteness of T leads to a conservative p-value; in particular, $\mathbb{E}[P] > 1/2$. One way to address this is to randomize the p-value to make it uniform on $[0, 1]$; however, randomization is generally undesirable in testing. As the most natural alternative, *mid p-values* arise in the presence of discrete test statistics. For the test statistic T , the mid p-value is given by

$$P_T = \frac{1}{2} \mathbb{P}(T' \leq T|T) + \frac{1}{2} \mathbb{P}(T' < T|T) = \frac{1}{2} F(T-) + \frac{1}{2} F(T),$$

where $F(t-) = \lim_{s \uparrow t} F(s)$ for $t \in \mathbb{R}$. Clearly, $P_T \leq F(T)$ and $\mathbb{E}[P_T] = 1/2$. If T is continuously distributed, then $P_T = F(T)$ is uniform on $[0, 1]$. In case T is discrete, P_T is not a p-variable. In Figure 12.5 we present some examples of quantile functions of p-, mid p- and average p-variables.

A mid p-value is an average p-value for any atomless probability measure $\mathbb{P}$. Suppose that there exists a standard uniform random variable V independent of T . Let $U = F(T-) + V(F(T) - F(T-))$, which is

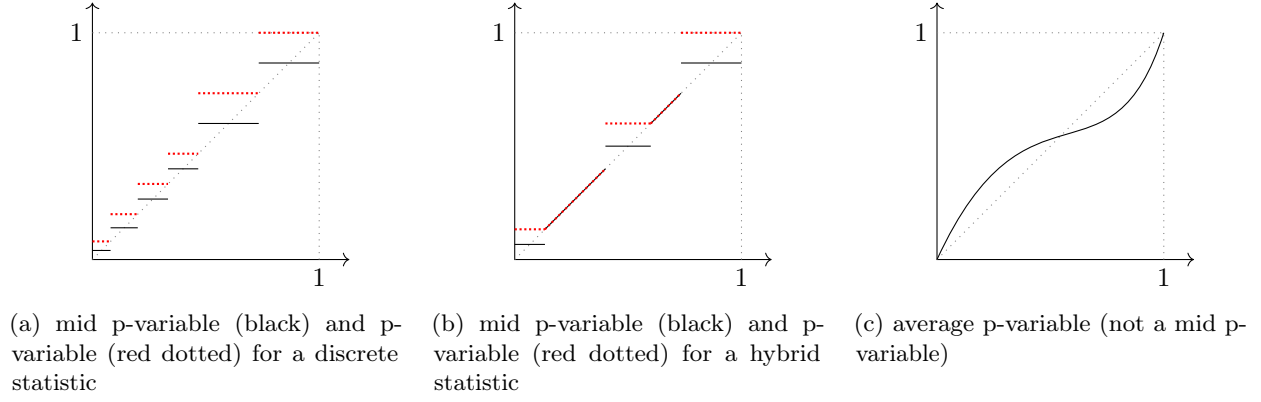


Figure 12.5: Examples of quantile functions of p-, mid p- and average p-variables

uniformly distributed on $[0, 1]$. It follows that

$$\frac{1}{2}F(T-) + \frac{1}{2}F(T) = \mathbb{E}^{\mathbb{P}}[U|T],$$

and then we conclude by Jensen's inequality that a mid p-value is an average p-value. (The above U exists even without assuming the existence of V independent of T ; see Lemma A.10.)

Another example of average p-values is the class of posterior predictive p-values in Bayesian analysis, which we explain briefly below.

Let X be a vector of data. The null hypothesis H_0 is given by $\{\psi \in \Psi_0\}$ where Ψ_0 is a subset of the parameter space Ψ on which a prior distribution is specified. The posterior predictive p-value is defined as the realization of the random variable

$$P_B := \mathbb{P}(D(X', \psi) \geq D(X, \psi) \mid X),$$

where D is a function (taking a similar role as test statistics), X' and X are iid conditional on ψ , and the probability is computed under the joint posterior distribution of (X', ψ) . Note that P_B can be rewritten as

$$P_B = \int \mathbb{P}(D(X', y) \geq D(X, y) \mid X, y) \Pi(dy \mid X),$$

where Π is the posterior distribution of ψ given the data X . The fact that P_B is an average p-variable follows from Proposition 12.1, condition (ii). In this context, average p-variables are obtained by integrating p-variables over the posterior distribution of some unobservable parameter. We omit a detailed discussion of the Bayesian interpretation, which is not the focus on the book, but simply make the point that average p-values have a natural appearance in the Bayesian context.

12.1.3 Calibration between average p-values and e-values

In the next result, we show that average p-values can be calibrated into an e-value through p-to-e calibrators in Section 2.3, under the additional assumption that the calibrator is convex.

Theorem 12.6

For an average p-variable P , $f(P)$ is an e-variable for any calibrator f that is convex on $[0, 1]$. For any e-variable E , $(2E)^{-1} \wedge 1$ is an average p-variable.

Proof.

To show the first statement, note that it suffices to consider p -variables $P_1, \dots, P_K$ taking values in $[0, 1]$. Let $P = (P_1 + \dots + P_K)/K$ and $E = f(P)$. By the convexity of the calibrator f ,

$$\mathbb{E}^{\mathbb{P}}[E] \leq \mathbb{E}^{\mathbb{P}} \left[\frac{1}{K} \sum_{k=1}^K f(P_k) \right] \leq 1,$$

and hence E is an e -variable.

Next, we show the second statement. Clearly, it suffices to show that $1/(2E)$ is an average p -variable for any e -variable E with mean 1, since any e -variable with mean less than 1 is dominated by an e -variable with mean 1. Let δ_x be the point mass at $x \in \mathbb{R}$.

Assume that E has a two-point distribution (including the point mass δ_1). With $\mathbb{E}^{\mathbb{P}}[E] = 1$, the distribution of E can be characterized with two parameters $p \in (0, 1)$ and $a \in [0, 1/p]$ as

$$p\delta_{1+(1-p)a} + (1-p)\delta_{1-pa}.$$

Write $x_1 = 1/(2 + 2(1-p)a) \in (0, 1/2]$ and $x_2 = 1/(2 - 2pa) \in [1/2, \infty)$. The distribution of $P = 1/(2E)$ (we allow P to take the value ∞ in case $a = 1/p$, which is harmless) is given by

$$p\delta_{x_1} + (1-p)\delta_{x_2}.$$

To show that P is an average p -variable, let $U \stackrel{d}{=} U[0, 1]$ and $V = \mathbb{E}^{\mathbb{P}}[U | \mathbb{1}_{\{U \leq x_1\}}]$. We will show $\mathbb{E}^{\mathbb{P}}[g(P)] \geq \mathbb{E}^{\mathbb{P}}[g(V)] \geq \mathbb{E}^{\mathbb{P}}[g(U)]$ for all increasing concave functions g . The second inequality follows from Jensen's inequality, and it suffices to prove the first one. The distribution of V is given by

$$2x_1\delta_{x_1} + (1 - 2x_1)\delta_{x_3},$$

where $x_3 = x_1 + 1/2$. Note that $2x_1 \geq 1/(2-p) \geq p$. Moreover, by Jensen's inequality,

$$\mathbb{E}^{\mathbb{P}}[P] = \mathbb{E}^{\mathbb{P}} \left[\frac{1}{2E} \right] \geq \frac{1}{2\mathbb{E}^{\mathbb{P}}[E]} = \frac{1}{2}.$$

Note that

$$\mathbb{E}^{\mathbb{P}}[g(P)] - \mathbb{E}^{\mathbb{P}}[g(V)] = (1-p)g(x_2) - (2x_1-p)g(x_1) - (1-2x_1)g(x_3).$$

Since $\mathbb{E}^{\mathbb{P}}[P] \geq \mathbb{E}^{\mathbb{P}}[V] = 1/2$, we know that

$$\frac{2x_1-p}{1-p}\delta_{x_1} - \frac{1-2x_1}{1-p}\delta_{x_3}$$

has a mean at most x_2 . Jensen's inequality then gives $\mathbb{E}^{\mathbb{P}}[g(P)] - \mathbb{E}^{\mathbb{P}}[g(V)] \geq 0$. Hence, P is an average p -variable.

We use two other known facts that can be checked directly: (a) For a general e -variable E with mean 1, its distribution can be rewritten as a mixture of two-point distributions with mean 1. (b) The set of average p -variables is closed under mixtures of distributions. Using these two facts, we conclude that $f(E)$ is an average p -variable.

Convex calibrators are quite common. Indeed, all calibrators (2.1)–(2.5) in Section 2.3 are convex on $[0, 1]$. Practically, average p -values can be calibrated into e -values using any calibrators introduced before, except for the all-or-nothing one $p \mapsto \mathbb{1}_{\{p \leq \alpha\}}/\alpha$ for some $\alpha \in (0, 1)$.

Figure 12.7 summarizes the calibrators among p -values, average p -values and e -values, including the ones

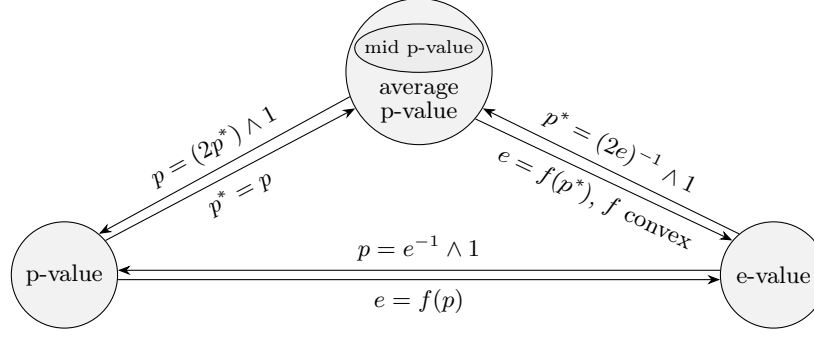


Figure 12.7: Calibration among p-values, average p-values and e-values, where the calibrator $f : [0, 1] \rightarrow [0, \infty]$ is left-continuous and decreasing with $f(0) = \infty$ and $\int_0^1 f(t)dt = 1$. All these calibrators are admissible in the sense of Section 2.3.

in Section 2.3. As implied by Theorem 12.6, the calibrator $e \mapsto (2e)^{-1}$ to an average p-value has an extra factor of $1/2$ compared to the standard calibrator $e \mapsto e^{-1}$ to a p-value. A composition of $e \mapsto (2e)^{-1} \wedge 1$ (from an e-value to an average p-value) and $p \mapsto (2p) \wedge 1$ (from an average p-value to a p-value) leads to the unique admissible e-to-p calibration $e \mapsto e^{-1} \wedge 1$ in Section 2.3, thus showing that average p-values serve as an intermediate object between e-values and p-values.

12.1.4 Tests based on average p-values

For an average p-variable P , the test $\mathbb{1}_{\{P \leq \alpha\}}$ is not a level- α test in general, because P is not necessarily a p-variable. Nevertheless, $\mathbb{1}_{\{P \leq V\}}$ for a random variable V with mean α independent of P often gives a level- α test, as shown in the next result. In what follows, independence is understood to hold for all $\mathbb{P} \in \mathcal{P}$.

Theorem 12.8

Let $\alpha \in (0, 1)$, $V \geq 0$ be independent of P such that $\mathbb{E}^{\mathbb{P}}[V \wedge 1] \leq \alpha$ for each $\mathbb{P} \in \mathcal{P}$.

- (i) If P is a p-variable, then $\mathbb{1}_{\{P \leq V\}}$ is a level- α test for $\mathcal{P}$.
- (ii) If P is an average p-variable and V has a decreasing density function on $(0, \infty)$ under each $\mathbb{P} \in \mathcal{P}$, then $\mathbb{1}_{\{P \leq V\}}$ is a level- α test for $\mathcal{P}$.

In particular, $\mathbb{1}_{\{P \leq 2\alpha U\}}$ is a level- α test for $\mathcal{P}$, where U is independent of P and uniformly distributed on $[0, 1]$ under each $\mathbb{P} \in \mathcal{P}$.

Proof.

Part (i) follows from $\mathbb{E}^{\mathbb{P}}[\mathbb{1}_{\{P \leq V\}}] = \mathbb{E}^{\mathbb{P}}[V \wedge 1] \leq \alpha$ for $\mathbb{P} \in \mathcal{P}$.

Part (ii) follows from the following argument. For $\mathbb{P} \in \mathcal{P}$, since V has a decreasing density, the function $g : x \mapsto \mathbb{P}(V \geq x)$ is decreasing convex. Hence, using Proposition 12.1, we have

$$\mathbb{E}^{\mathbb{P}}[\mathbb{1}_{\{P \leq V\}}] = \mathbb{E}^{\mathbb{P}}[g(P)] \leq \int_0^1 g(u)du = \mathbb{E}^{\mathbb{P}}[V \wedge 1] \leq \alpha,$$

as desired.

Theorem 12.8 gives rise to randomized tests based on the arithmetic mean of several p-values regardless

of their dependence structure; i.e.

$$\mathbb{1}_{\{(P_1 + \dots + P_K)/K \leq 2\alpha U\}} \quad (12.2)$$

is a level- α test for any p-variables $P_1, \dots, P_K$ and $U \stackrel{d}{\sim} U[0, 1]$ independent of them. If randomization is not allowed, then we can use

$$\mathbb{1}_{\{(P_1 + \dots + P_K)/K \leq 2\alpha\}},$$

which is weaker than (12.2) in terms of power. This is a method of testing by combining arbitrarily dependent p-values. The next few sections are dedicated to a formal theory of general methods for combining p-values, including an exchangeable improvement and another randomized test based on the average p-value.

12.2 Merging p-values under arbitrary dependence

The structure of merging functions for p-values is much richer than that of e-merging functions. It turns out that admissible ways of merging p-values under arbitrary dependence have to go through merging e-values. At a high level, the reason why merging p-values relies on e-values is due to optimal transport duality in Lemma 11.1, which converts constraints on probability into constraints on expectation.

12.2.1 P-merging functions and examples

We first give the definition of p-merging functions, which is analogous to e-merging functions in Definition 8.1.

Definition 12.9: P-merging functions

- (i) A *p-merging function* (of K p-values) is an increasing Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$ such that for any hypothesis, $F(P_1, \dots, P_K)$ is a p-variable for any p-variables $P_1, \dots, P_K$.
- (ii) A p-merging function F is *symmetric* if $F(\mathbf{p})$ is invariant under any permutation of $\mathbf{p}$.
- (ii) A p-merging function F is *homogeneous* if $F(\lambda \mathbf{p}) = \lambda F(\mathbf{p})$ for all $\lambda \in (0, 1]$ and $\mathbf{p}$ with $F(\mathbf{p}) \leq 1$.
- (iii) A p-merging function F *dominates* a p-merging function G if $F \leq G$. The domination is *strict* if $F \leq G$ and $F(\mathbf{p}) < G(\mathbf{p})$ for some $\mathbf{p} \in [0, \infty)^K$. A p-merging function is *admissible* if it is not strictly dominated by any p-merging function.

Similarly to the situation in Chapters 8 and 11, it suffices to consider p-variables for the fixed atomless probability measure $\mathbb{P}$. Let $\mathfrak{U}$ be the set of all p-variables for $\mathbb{P}$. A p-merging function F is *precise* if

$$\sup_{\mathbf{P} \in \mathfrak{U}^K} \mathbb{P}(F(\mathbf{P}) \leq \varepsilon) = \varepsilon \text{ for all } \varepsilon \in (0, 1).$$

In other words, ε by ε , F attains the largest possible probability allowed for $F(\mathbf{P})$ to be a p-value.

Note that p-values are more useful to be small, and hence domination between p-merging functions is defined via an opposite direction to that between e-merging functions. We will also speak of admissibility within smaller classes of p-merging functions, such as the class of symmetric p-merging functions.

All p-merging functions that we encounter in this chapter are homogeneous and symmetric. Although we allow the domain of F to be $[0, \infty)^K$ in order to simplify presentation, the informative part of F is its restriction to $[0, 1]^K$.

We have some specific notation for this section. Denote by $\mathbf{0}$ the K -vector of zeros, and $\mathbf{1}$ the K -vector of ones. All vector inequalities and the operation $\wedge$ of taking the minimum of two vectors are component-wise. For $a, b, x, y \in \mathbb{R}$, $ax \wedge by$ should be understood as $(ax) \wedge (by)$.

We collect some basic properties of admissible p-merging functions, which will be useful in our analysis later. We skip the proofs.

Proposition 12.10

- (i) Any admissible p-merging function is precise and lower semicontinuous, takes value 0 on $[0, \infty)^K \setminus (0, \infty)^K$, and satisfies $F(\mathbf{p}) = F(\mathbf{p} \wedge \mathbf{1}) \wedge 1$ for all $\mathbf{p} \in [0, \infty)^K$.
- (ii) The point-wise limit of a sequence of p-merging functions is a p-merging function.
- (iii) Any p-merging function is dominated by an admissible p-merging function.

In what follows, we will always write $\mathbf{p} = (p_1, \dots, p_K)$. Two natural families of p-merging functions are the *order-family* based on order statistics and the *mean-family* based on generalized mean. All functions that we see below are homogeneous and symmetric.

Example 12.11: Order-family

The order-family is parameterized by $k \in \mathcal{K}$, and its k -th element is the function

$$G_{k,K} : \mathbf{p} \mapsto \frac{K}{k} p_{(k)} \wedge 1,$$

where $p_{(k)}$ is the k -th ascending order statistic of $p_1, \dots, p_K$. One can verify that the functions $G_{k,K}$, $k \in \mathcal{K}$ are precise p-merging functions; see also Example 12.20.

Example 12.12: Mean-family

The mean-family is parameterized by $r \in [-\infty, \infty]$, and its element with index r has the form

$$F_{r,K} : \mathbf{p} \mapsto b_{r,K} \mathbb{M}_{r,K}(\mathbf{p}) \wedge 1, \tag{12.3}$$

where $\mathbb{M}_{r,K}$ is given by

$$\mathbb{M}_{r,K}(\mathbf{p}) = \left(\frac{p_1^r + \dots + p_K^r}{K} \right)^{1/r}$$

and $b_{r,K} \geq 1$ is a suitable constant making $F_{r,K}$ a p-merging function. The arithmetic mean $\mathbb{M}_{1,K}$ of p-values has been studied in Section 12.1.

The average $\mathbb{M}_{r,K}$ is also defined for $r \in \{0, \infty, -\infty\}$ as the limiting cases of (12.3), which correspond to the geometric mean, the maximum, and the minimum, respectively. Another useful member of the mean-family is the multiple $F_{-1,K}$ of the harmonic mean $\mathbb{M}_{-1,K}$.

Example 12.13: Bonferroni correction and maximum

The initial and final elements of the M- and O-families coincide: the initial element is the Bonferroni p-merging function

$$G_{1,K} = F_{-\infty,K} : \mathbf{p} \mapsto K \min(\mathbf{p}) \wedge 1,$$

and the final element is the maximum p-merging function

$$G_{K,K} = F_{\infty,K} : \mathbf{p} \mapsto \max(\mathbf{p}).$$

Example 12.14: Simes and Hommel functions

By minimizing over k in the order-family, we get the *Simes function*

$$S_K := \bigwedge_{k=1}^K G_{k,K}$$

which is not a p-merging function, but it produces a p-variable if the input p-variables are independent or PRDS. Moreover, one can show that the Simes function is the minimum of all symmetric p-merging functions. Multiplied by a constant $\ell_K := \sum_{k=1}^K k^{-1}$ representing a factor of penalization, we get a precise p-merging function

$$H_K := \left(\sum_{k=1}^K \frac{1}{k} \right) \bigwedge_{k=1}^K G_{k,K} = \ell_K S_K,$$

which we call the *Hommel function*.

12.2.2 Using e-values to merge p-values

We first explain a simple way of merging p-values through e-values. Recall that a calibrator $f : [0, \infty) \rightarrow [0, \infty]$ transforms a p-variable to an e-variable, and they satisfy $\int_0^1 f(p) dp \leq 1$ and $f(p) = 0$ for $p > 1$. Let Δ_K be the standard K -simplex, that is,

$$\Delta_K := \{(\lambda_1, \dots, \lambda_K) \in [0, 1]^K : \lambda_1 + \dots + \lambda_K = 1\}.$$

For any calibrators $f_1, \dots, f_K$ and any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$, define the function

$$G(\mathbf{p}) := \lambda_1 f_1(p_1) + \dots + \lambda_K f_K(p_K).$$

For any vector $\mathbf{P}$ of p-variables, the function G produces an e-variable $G(\mathbf{P})$, since it is a convex combination of e-variables $f_1(P_1), \dots, f_K(P_K)$. Markov's inequality implies

$$\mathbb{P}\left(G(\mathbf{P}) \geq \frac{1}{\varepsilon}\right) \leq \varepsilon \quad \text{for all } \varepsilon \in (0, 1).$$

Thus, $1/G$ is a p-merging function, and this can also be seen as applying the e-to-p calibrator $t \mapsto (1/t) \wedge 1$ in Proposition 2.4 to the e-variable $G(\mathbf{P})$.

Such a “naive” (but valid) procedure of merging p-values through merging e-values, producing $1/G$, is generally not admissible, as the last conversion step is wasteful. However, perhaps surprisingly, all admissible and homogeneous p-merging functions can be obtained via the above procedure, choosing different functions G for each ε . This is shown in Theorem 12.15 below, the main result of this section.

To explain this result, we reformulate p-merging functions by their rejection regions. The *rejection region* of a p-merging function F at level $\varepsilon \in (0, 1)$ is defined as

$$R_\varepsilon(F) := \{\mathbf{p} \in [0, \infty)^K : F(\mathbf{p}) \leq \varepsilon\}.$$

If F is homogeneous, then $R_\varepsilon(F)$, $\varepsilon \in (0, 1)$, takes the form $R_\varepsilon(F) = \varepsilon A$ for some $A \subseteq [0, \infty)^K$. Conversely, any increasing collection of Borel lower sets $\{R_\varepsilon \subseteq [0, \infty)^K : \varepsilon \in (0, 1)\}$ determines an increasing Borel function $F : [0, \infty)^K \rightarrow [0, 1]$ by the equation

$$F(\mathbf{p}) = \inf\{\varepsilon \in (0, 1) : \mathbf{p} \in R_\varepsilon\}, \tag{12.4}$$

with the convention $\inf \emptyset = 1$. It is immediate that F is a p-merging function if and only if $\mathbb{P}(\mathbf{P} \in R_\varepsilon) \leq \varepsilon$ for all $\varepsilon \in (0, 1)$ and $\mathbf{P} \in \mathfrak{U}^K$.

Theorem 12.15

- (i) For any p-merging function F and any $\varepsilon \in (0, 1)$, there exists a homogeneous p-merging function G such that $R_\varepsilon(F) \subseteq R_\varepsilon(G)$.
- (ii) For any admissible p-merging function F and any $\varepsilon \in (0, 1)$, there exist $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and admissible calibrators $g_1, \dots, g_K$ such that

$$R_\varepsilon(F) \subseteq \left\{ \mathbf{p} \in [0, \infty)^K : \sum_{k=1}^K \lambda_k g_k(p_k) \geq \frac{1}{\varepsilon} \right\}. \quad (12.5)$$

- (iii) If F is an admissible homogeneous p-merging function, then there exist $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and admissible calibrators $f_1, \dots, f_K$ such that for each $\varepsilon \in (0, 1)$,

$$R_\varepsilon(F) = \left\{ \mathbf{p} \in [0, \infty)^K : \sum_{k=1}^K \lambda_k f_k\left(\frac{p_k}{\varepsilon}\right) \geq 1 \right\}. \quad (12.6)$$

- (iv) For any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and calibrators $f_1, \dots, f_K$, (12.6) determines a homogeneous p-merging function.

As a consequence of Theorem 12.15 part (i), if the level α of type-I error control is determined before choosing the p-merging function, then it suffices to consider homogeneous ones, since their rejection sets are at least as larger as those of other p-merging functions. Note that this does not imply that there exists a homogeneous p-merging function G dominating F in general, because the construction of G depends on the given α .

The calibrators in (ii) and (iii) of Theorem 12.15 are connected via $g_k(x) = f_k(x/\varepsilon)/\varepsilon$, $x \in [0, \infty)$, for each $k \in \mathcal{K}$. This choice yields admissible calibrators by Proposition 2.6. Therefore, when F is an admissible homogeneous p-merging function, the calibrators for different ε -levels are generated by one tuple of calibrators through rescaling in the sense of Proposition 2.6.

The p-merging function in (12.6) can be explicitly expressed via (12.4) as

$$F(\mathbf{p}) = \inf \left\{ \varepsilon \in (0, 1] : \sum_{k=1}^K \lambda_k f_k\left(\frac{p_k}{\varepsilon}\right) \geq 1 \right\} \quad \text{for } \mathbf{p} \in [0, \infty)^K. \quad (12.7)$$

To prove Theorem 12.15, we need an additional lemma. We say that a set $R \subseteq [0, \infty)^K$ is a decreasing set if $\mathbf{x} \in R$ implies $\mathbf{y} \in R$ for all $\mathbf{y} \in [0, \infty)^K$ with $\mathbf{y} \leq \mathbf{x}$ (componentwise). Let $\mathfrak{U}_0$ be the set of uniformly distributed random variables on $[0, 1]$ under $\mathbb{P}$ (i.e., they are exact p-variables). Clearly, $\mathfrak{U}_0 \subseteq \mathfrak{U}$. For any decreasing set L , we have

$$\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in L) = \sup_{\mathbf{P} \in \mathfrak{U}^K} \mathbb{P}(\mathbf{P} \in L), \quad (12.8)$$

because replacing a p-variable with a smaller exact p-variable does not reduce the above probability. The fact (12.8) will be repeatedly used in the proof below.

Lemma 12.16

Let $R \subseteq [0, \infty)^K$ be a decreasing Borel set. For any $\beta \in (0, 1)$, we have

$$\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in \beta R) \geq \beta \iff \sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R) = 1.$$

Proof.

We first prove the $\Leftarrow$ direction by contraposition. Suppose

$$\gamma := \sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in \beta R) < \beta.$$

Take an event A with probability β and any $\mathbf{P} \in \mathfrak{U}_0^K$ independent of A . Define $\mathbf{P}^*$ by

$$\mathbf{P}^* = \beta \mathbf{P} \times \mathbf{1}_A + \mathbf{1} \times \mathbf{1}_{A^c}.$$

It is straightforward to check $\mathbf{P}^* \in \mathfrak{U}^K$. Hence, by (12.8),

$$\beta \mathbb{P}(\mathbf{P} \in R) = \mathbb{P}(A) \mathbb{P}(\mathbf{P} \in R) \leq \mathbb{P}(\mathbf{P}^* \in \beta R) \leq \gamma,$$

and thus $\mathbb{P}(\mathbf{P} \in R) \leq \gamma/\beta$. Since $\gamma/\beta < 1$, this then yields that $\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R) < 1$ and completes the $\Leftarrow$ direction.

Next we show the $\Rightarrow$ direction. Suppose $\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in \beta R) \geq \beta$. For any $\varepsilon \in (0, \beta)$, there exists $\mathbf{P} = (P_1, \dots, P_K) \in \mathfrak{U}_0^K$ such that $\mathbb{P}(\mathbf{P} \in \beta R) > \beta - \varepsilon$. Let $A = \{\mathbf{P} \in \beta R\}$, $\gamma = \mathbb{P}(A)$, and B be an event containing A with $\mathbb{P}(B) = \beta \vee \gamma$. Let $\mathbf{P}^* = (P_1^*, \dots, P_K^*)$ follow the conditional distribution of $\mathbf{P}/\beta$ given B . We have

$$\mathbb{P}(\mathbf{P}^* \in R) = \mathbb{P}(\mathbf{P} \in \beta R \mid B) = \mathbb{P}(A \mid B) = \frac{\gamma}{\beta \vee \gamma}.$$

Note that for $k \in \{1, \dots, K\}$,

$$\mathbb{P}(P_k^* \leq p) = \mathbb{P}(P_k/\beta \leq p \mid B) \leq \frac{\mathbb{P}(P_k \leq \beta p)}{\mathbb{P}(B)} = \frac{\beta p}{\beta \vee \gamma} \leq p,$$

and hence $\mathbf{P}^* \in \mathfrak{U}^K$. Since $\gamma > \beta - \varepsilon$ and $\varepsilon \in (0, \beta)$ is arbitrary, we can conclude $\sup_{\mathbf{P} \in \mathfrak{U}^K} \mathbb{P}(\mathbf{P} \in R) = 1$, yielding $\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R) = 1$ via (12.8).

Proof of Theorem 12.15.

For part (i), since any p-merging function has an admissible p-merging function dominating it (Proposition 12.10), it suffices to prove (i) for admissible p-merging functions.

Let F be an admissible p-merging function and fix $\varepsilon \in (0, 1)$. Note that the set $R_\varepsilon(F)$ is a lower set, and it is closed due to Proposition 12.10. Hence $\mathbf{1}_{R_\varepsilon(F)}$ is upper semicontinuous. Using optimal transport duality, Lemma 11.1,

$$\begin{aligned} & \min_{(h_1, \dots, h_K) \in \mathcal{B}^K} \left\{ \sum_{k=1}^K \int_0^1 h_k(x) dx : \bigoplus_{k=1}^K h_k \geq \mathbf{1}_{R_\varepsilon(F)} \right\} \\ &= \max_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R_\varepsilon(F)) = \varepsilon, \end{aligned}$$

where the last equality holds because F is precise (Proposition 12.10). Take $(h_1^\varepsilon, \dots, h_K^\varepsilon) \in \mathcal{B}^K$ such that

$$\bigoplus_{k=1}^K h_k^\varepsilon \geq \mathbf{1}_{R_\varepsilon(F)} \quad \text{and} \quad \sum_{k=1}^K \int_0^1 h_k^\varepsilon(x) dx = \varepsilon.$$

By the corresponding conditions on $\mathbf{1}_{R_\varepsilon(F)}$ in Lemma 11.1, we can choose each h_k^ε to be

nonnegative, decreasing and left-continuous. Moreover, one can also require $h_k^\varepsilon(0) = \infty$ for each k .

Lemma 12.16 gives

$$\max_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R_\varepsilon(F)) = \varepsilon \implies \max_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\varepsilon \mathbf{P} \in R_\varepsilon(F)) = 1. \quad (12.9)$$

Therefore, using duality in Lemma 11.1 again,

$$\min_{(h_1, \dots, h_K) \in \mathcal{B}^K} \left\{ \sum_{k=1}^K \frac{1}{\varepsilon} \int_0^\varepsilon h_k(x) dx : \bigoplus_{k=1}^K h_k \geq \mathbb{1}_{R_\varepsilon(F)} \right\} = 1,$$

implying $\sum_{k=1}^K \int_0^\varepsilon h_k^\varepsilon(x) dx \geq \varepsilon$. As $\sum_{k=1}^K \int_0^1 h_k^\varepsilon(x) dx = \varepsilon$ and each h_k^ε is nonnegative, we know $h_k^\varepsilon(x) = 0$ for $x > \varepsilon$.

Define the set $A_\varepsilon := \{\mathbf{p} \in [0, \infty)^K : \sum_{k=1}^K h_k^\varepsilon(p_k) \geq 1\}$. Since $\bigoplus_{k=1}^K h_k^\varepsilon \geq \mathbb{1}_{R_\varepsilon(F)}$, we have $R_\varepsilon(F) \subseteq A_\varepsilon$. Note that A_ε is a closed lower set. By Markov's inequality,

$$\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P} \left(\bigoplus_{k=1}^K h_k^\varepsilon(\mathbf{P}) \geq 1 \right) \leq \sup_{P \in \mathfrak{U}_0} \sum_{k=1}^K \mathbb{E}^\mathbb{P}[h_k^\varepsilon(P)] = \varepsilon.$$

Hence, we can define a function $G : [0, \infty)^K \rightarrow \mathbb{R}$ via $R_\delta(G) = \delta \varepsilon^{-1} A_\varepsilon$ for all $\delta \in (0, 1)$. By the above properties of A_ε and Lemma 12.16, we can check that G is a valid homogeneous p-merging function, and $R_\varepsilon(F) \subseteq R_\varepsilon(G)$. This proves part (i). To see the result in part (ii), it suffices to take $\lambda_k = \varepsilon^{-1} \int_0^\varepsilon g^\varepsilon(x) dx$ and $g_k : [0, \infty) \rightarrow \mathbb{R}$, $x \mapsto \varepsilon^{-1} h_k^\varepsilon(x)/\lambda_k$ for each $k = 1, \dots, K$, with $f_k = 1$ if $\lambda_k = 0$, which gives that $g_1, \dots, g_K$ are admissible calibrators, and

$$A_\varepsilon = \left\{ \mathbf{p} \in [0, \infty)^K : \sum_{k=1}^K \lambda_k g_k(p_k) \geq \frac{1}{\varepsilon} \right\}.$$

Before proving part (iii), we first prove part (iv). For a function F defined by (12.7), the condition $\mathbb{P}(F(\mathbf{P}) \leq \varepsilon) \leq \varepsilon$ for $\mathbf{P} \in \mathfrak{U}_0^K$ follows from Proposition 2.6 and Markov's inequality. Its increasing monotonicity and homogeneity are straightforward to check. Therefore, it is a homogeneous p-merging function.

Now we can prove part (iii). Let $f_k(x) = \varepsilon g_k(\varepsilon x)$ for $x \in [0, \infty)$, and define H via the rejection regions in (12.6), which is a homogeneous p-merging function by part (iv). Then we have $R_\varepsilon(F) \subseteq A_\varepsilon = R_\varepsilon(H)$. Since F and H are both homogeneous, this means $R_\delta(F) \subseteq R_\delta(H)$ for all $\delta \in (0, 1)$. The admissibility of F now gives $F = H$, and this proves part (iii).

If a homogeneous p-merging function F is symmetric, then $f_1, \dots, f_K$ in Theorem 12.15 can be chosen identical, and the same applies to $\lambda_1, \dots, \lambda_K$.

Theorem 12.17

For any F that is admissible within the family of homogeneous symmetric p-merging functions, there exists an admissible calibrator f such that for each $\varepsilon \in (0, 1)$,

$$R_\varepsilon(F) = \left\{ \mathbf{p} \in [0, \infty)^K : \frac{1}{K} \sum_{k=1}^K f\left(\frac{p_k}{\varepsilon}\right) \geq 1 \right\}. \quad (12.10)$$

Conversely, for any calibrator f , (12.10) determines a homogeneous symmetric p-merging function.

The p-merging function in (12.10) can be explicitly expressed as

$$F(\mathbf{p}) = \inf \left\{ \varepsilon \in (0, 1] : \frac{1}{K} \sum_{k=1}^K f\left(\frac{p_k}{\varepsilon}\right) \geq 1 \right\} \text{ for } \mathbf{p} \in [0, \infty)^K. \quad (12.11)$$

Proof.

The proof is similar to that of Theorem 12.15 and we only mention the differences. For the first statement, it suffices to notice two facts. First, if R_ε is symmetric, then $h_1^\varepsilon, \dots, h_K^\varepsilon$ in the proof of Theorem 12.15 can be chosen as identical; for instance, one can choose the average of them. Second, the symmetry of $R_\varepsilon(F)$ guarantees that G in the proof of Theorem 12.15 is symmetric, and hence it is sufficient to require the admissibility of F within homogeneous symmetric p-merging functions in this proposition. The last statement in the proposition follows from Theorem 12.15 by noting that (12.10) defines a symmetric rejection region.

The requirement $f(0) = \infty$ for an admissible calibrator f implies that the combined test (12.10) gives a rejection as soon as one of the input p-values is 0, which is obviously necessary for admissibility (Proposition 12.10). Although many examples in the M- and O-families, in particular $F_{r,K}$ for $r > 0$ and $G_{k,K}$ for $k > 1$, do not satisfy this, we can make the zero-one adjustment

$$\tilde{F}(\mathbf{p}) := \begin{cases} F(\mathbf{p} \wedge \mathbf{1}) \wedge 1 & \text{if } \mathbf{p} \in (0, \infty)^K \\ 0 & \text{otherwise,} \end{cases}$$

which does not affect the validity of the p-merging function.

For a decreasing function $f : [0, \infty) \rightarrow [0, \infty]$ and a p-merging function F taking values in $[0, 1]$, we say that f *induces* F if (12.10) holds; similarly, we may say that $\lambda_1, \dots, \lambda_K$ and $f_1, \dots, f_K$ *induce* F if (12.5) holds. Theorems 12.15 and 12.17 imply that any admissible p-merging function F is induced by some admissible calibrators (but they need not be uniquely determined by F).

The converse direction, constructing an admissible p-merging function, is more delicate. In general, a p-merging function induced by admissible calibrators is not necessarily admissible. Using (12.9) and a compactness argument, a necessary and sufficient condition for a calibrator f to induce a precise p-merging function (a weaker requirement than admissibility) via (12.10) is

$$\mathbb{P} \left(\frac{1}{K} \sum_{k=1}^K f(P_k) \geq 1 \right) = 1 \quad \text{for some } (P_1, \dots, P_K) \in \mathfrak{U}_0^K. \quad (12.12)$$

Condition (12.12) may be difficult to check for a given f in general. A special known case is that f is convex. In this case, (12.12) holds if and only if $f \leq K$ on $(0, 1]$, which is a highly nontrivial mathematical result. This condition turns out to be also sufficient for admissibility of F in Theorem 12.17. The proof of Theorem 12.18 is advanced and omitted here.

Theorem 12.18

If f is an admissible calibrator that is strictly convex on $(0, 1]$, $f \leq K$ on $(0, 1]$, and $f(1) = 0$, then the p-merging function induced by f via (12.10) is admissible.

More generally, the same holds true if f is an admissible calibrator satisfying the following condition: for some $\eta \in [0, 1/K]$ and $\tau = 1 - (K - 1)\eta$, $f = K$ on $(0, \eta]$, $f(\eta+) \leq K$ (where $f(\eta+)$ is the right limit of f at η), f is strictly convex on $(\eta, \tau]$, and $f(\tau) = 0$.

Remark 12.19. There is a corresponding statement for the case of concavity instead of convexity in Theorem 12.18. The p-merging function induced by f via (12.10) is admissible if f satisfies the following condition: for some $\eta \in [0, 1/K]$ and $\tau = 1 - (K - 1)\eta$, $f = K$ on $(0, \eta]$, $f(\eta+) \geq K/(K - 1)$, f is strictly concave on $(\eta, \tau]$, and $f(\tau) = 0$.

12.2.3 Examples

We present a few examples of p-merging functions, continuing those in Section 12.2.1, through the constructions Theorem 12.15 using calibrators.

Example 12.20: Order-family

The p-merging function $F := G_{k,K}$, $k \in \mathcal{K}$, is induced by the calibrator $f(p) = (K/k)\mathbb{1}_{p \in [0, k/K]}$. Its zero-one adjusted version is admissible for $k \in [K-1]$.

Example 12.21: Twice the arithmetic mean

The function $2\mathbb{M}_{1,K}$, which is twice the arithmetic mean, is a precise p-merging function but not admissible. We have verified that it is a p-merging function in Proposition 12.4 in Section 12.1.

This p-merging function can be written in the form (12.11) with $f(p) = 2 - 2p$ for $p \geq 0$, but this f is not a calibrator because $f(p)$ is not 0 for $p > 1$. Using the calibrator $p \mapsto (2 - 2p)_+$ in (12.11) gives a p-merging function that dominates $2\mathbb{M}_{1,K}$, a slight improvement.

Example 12.22: e times the geometric mean

The function $e\mathbb{M}_{0,K}$, that is, e times the geometric mean, is a p-merging function that is not precise, but approximately precise for large K .

This p-merging function can be written in the form (12.11) with $f(p) = -\log(p)$ for $p \geq 0$. Using the calibrator $p \mapsto (-\log(p))_+$ in (12.11) gives a slight improvement.

Example 12.23: $\log K$ times the harmonic mean

Let $T_K = \log K + \log \log K + 1$. The function $(T_K + 1)\mathbb{M}_{-1,K}$, that is, $T_K + 1$ times the harmonic mean, is a p-merging function that is not precise, but approximately precise for large K .

This p-merging function can be written in the form (12.11) with

$$f(p) = \left(\frac{1}{T_K p} - \frac{1}{T_K} \right) \wedge K$$

for $p \geq 0$, and it is an elementary exercise in calculus to check that $p \mapsto (f(p))_+$ is a calibrator, and using it in (12.11) gives a slight improvement.

Example 12.24: Hommel function

The Hommel function H_K is a precise p-merging function but not admissible. For $K \geq 4$, it is strictly dominated by the p-merging function H_K^* induced by the calibrator f given by

$$f(p) = \frac{K \mathbb{1}_{\{\ell_K p \leq 1\}}}{\lceil K \ell_K p \rceil}, \quad p \geq 0,$$

and H_K^* is admissible when K is not a prime number. The latter point is an intriguing fact; for example, if $K = 2$ or 3 , H_K^* is shown to be not admissible.

12.3 Combining exchangeable p-values

In this section, we additionally assume that the input p-variables are exchangeable. Exchangeability among p-values is encountered, for example, in statistical testing via sample splitting. Reasons for sample splitting

include to relax the assumptions needed to obtain theoretical guarantees and to reduce computational costs. The drawback of methods based on sample-splitting is that the obtained p-values are affected by the randomness of the split, and thus repeated re-sampling can be performed, resulting in exchangeable p-values.

One can of course combine the obtained p-values by using the rules obtained in Section 12.2 (such as twice the average in Example 12.21) for arbitrarily dependent p-values. But we can do better by exploiting exchangeability. However, the constant 2 in front of the arithmetic average cannot be directly improved. Indeed, we show that if we stick to symmetric merging functions, there is no hope to improve even assuming exchangeability.

Definition 12.25

An *ex-p-merging function* is an increasing Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$ such that $\mathbb{P}(F(\mathbf{P}) \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ and $\mathbf{P} \in \mathfrak{U}^K$ that is exchangeable. It is *homogeneous* if $F(\lambda \mathbf{p}) = \lambda F(\mathbf{p})$ for all $\lambda \in (0, 1]$ and $\mathbf{p}$ with $F(\mathbf{p}) \leq 1$.

Proposition 12.26

A symmetric ex-p-merging function is necessarily a p-merging function. Hence, for an ex-p-merging function to strictly dominate an admissible p-merging function, it cannot be symmetric.

Proof.

Let $\mathbf{P} \in \mathfrak{U}^K$, and let π be a random permutation of $\{1, \dots, K\}$, uniformly drawn from all permutations of $\{1, \dots, K\}$ and independent of $\mathbf{P}$. Let $\mathbf{P}^\pi = (P_{\pi(1)}, \dots, P_{\pi(K)})$. Note that $\mathbf{P}^\pi$ is exchangeable by construction. If F is a symmetric ex-p-merging function, it must satisfy $F(\mathbf{P}^\pi) = F(\mathbf{P})$. Because $F(\mathbf{P}^\pi)$ is a p-variable, so is $F(\mathbf{P})$, showing that F is a p-merging function.

In many practical settings, the exchangeable p-values can be generated one by one by repeating the same randomized procedure many times, generating a stream of p-values. We introduce combination rules that would simply process these p-values in the order that they are generated.

The next result follows by combining the exchangeable Markov inequality in Theorem 5.6, the p-merging method from Theorem 12.17 and the formula (12.7).

Theorem 12.27: Combination of exchangeable p-values

Define the function $F : [0, \infty)^K \rightarrow [0, 1]$ by

$$F(\mathbf{p}) = \inf \left\{ \varepsilon \in (0, 1] : \bigvee_{\ell=1}^K \left(\frac{1}{\ell} \sum_{k=1}^{\ell} f\left(\frac{p_k}{\varepsilon}\right) \right) \geq 1 \right\},$$

where f is a calibrator. Then, for any vector $\mathbf{P}$ of p-variables that is exchangeable, $F(\mathbf{P})$ is a p-variable. That is, F is an ex-p-merging function.

In fact, for the method in Theorem 12.27, we do not need to fix the number of p-values ahead of time, they can just be processed online, yielding a p-value whenever this procedure is stopped. This makes our merging rules particularly simple and practical. However, note that the online procedure requires that the calibrator f does not depend on K .

12.4 Randomized combinations of p-values

In Section 12.2, we obtained methods to merge p-values using e-values in Theorems 12.15 and 12.17, and they are admissible with the condition in Theorem 12.18. When randomization is allowed, the Markov inequality

can be enhanced to the randomized Markov inequality, presented in Theorem 2.9. This allows us to design more powerful way of merging p-values under randomization. We continue to work under the assumption that the p-values are arbitrarily dependent, as in Section 12.2. The following result describes a method that uses randomization to improve p-merging functions. This result can be seen as a combination of Theorem 2.9 and 12.15 and the formula (12.7).

Theorem 12.28: Randomized combination of p-values

Define the function $F : [0, \infty)^K \times [0, \infty) \rightarrow [0, 1]$ by

$$F(\mathbf{p}, u) = \inf \left\{ \varepsilon \in (0, 1] : \sum_{k=1}^K \lambda_k f_k \left(\frac{p_k}{\varepsilon} \right) \geq u \right\}, \quad (12.13)$$

where $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and $f_1, \dots, f_K$ are calibrators. Then, for any vector $\mathbf{P}$ of p-variables and another p-variable V independent of $\mathbf{P}$, $F(\mathbf{P}, V)$ is a p-variable.

Proof.

Let $U \stackrel{d}{\sim} U[0, 1]$. The condition $\mathbb{P}(F(\mathbf{P}, U) \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ follows from the randomized Markov inequality in Theorem 2.9. For a general p-variable V , it suffices to notice that $F(\mathbf{p}, u)$ is increasing in u , and hence $\mathbb{P}(F(\mathbf{P}, V) \leq \alpha) \leq \mathbb{P}(F(\mathbf{P}, U) \leq \alpha) \leq \alpha$.

A simple subclass of (12.13) is

$$F(\mathbf{p}, u) = \inf \left\{ \varepsilon \in (0, 1] : \frac{1}{K} \sum_{k=1}^K f \left(\frac{p_k}{\varepsilon} \right) \geq u \right\},$$

for a calibrator f , similarly to (12.7).

For the deterministic choice $V = 1$, Theorem 12.28 gives the validity statement in Theorem 12.15 (iv). In practice, when randomization is allowed, V should be chosen as uniformly distributed on $[0, 1]$. This merging method produces a smaller p-value than the deterministic p-merging function in (12.7).

Randomization is undesirable in many statistical applications. One deterministic remedy is that, when one of the p-variables P_k in $\mathbf{P}$ is known to be independent of the others, one can use $V = P_k$ and apply the merging function to the rest of the p-variables with randomization through V .

If randomization is permitted, one can also improve the existing rules for combining arbitrarily dependent p-values by using the exchangeable combination rule in Section 12.4 applied to a random permutation of the p-values.

Table 12.29 summarizes some examples of p-merging methods under arbitrary dependence, randomization, or exchangeability. They are derived from Theorems 12.17, 12.27 and 12.28 with different calibrators f , and correspond to Examples 12.20–12.23.

Bibliographical note

Merging p-values has a long history, and some early works include Tippett [1931], Pearson [1933] and Fisher [1948]. The Simes and Hommel functions in Example 12.14 were proposed by Simes [1986] and Hommel [1983], respectively. The order-family was proposed by Rüger [1978]. The harmonic mean was proposed by Wilson [2019]. The mean-family was formally studied by Vovk and Wang [2020]. In some of the above works, often the p-values are assumed to be independent or follow a certain dependence structure; key exceptions are Rüger [1978], Rüschendorf [1982] and Vovk and Wang [2020].

The content of Section 12.1 is based on Wang [2024b], where average p-values are called p*-values. Proposition 12.4 was first shown by Rüschendorf [1982] based on optimal transport duality and then by

Context	Arbitrary dependence (Section 12.2)	Exchangeability (Section 12.3)	Arbitrary dependence, randomized (Section 12.4)
Result	Theorem 12.17	Theorem 12.27	Theorem 12.28
Order statistics	$\frac{K}{k} p_{(k)}$	$\frac{K}{k} \bigwedge_{m=1}^K p_{(\lambda_m)}^m$	$\frac{K}{k} p_{(\lceil Uk \rceil)}$
Arithmetic mean	$2\mathbb{M}_1(\mathbf{p})$	$2 \bigwedge_{m=1}^K \mathbb{M}_1(\mathbf{p}_m)$	$\frac{2}{2-U} \mathbb{M}_1(\mathbf{p})$
Geometric mean	$\mathbf{e} \mathbb{M}_0(\mathbf{p})$	$\mathbf{e} \bigwedge_{m=1}^K \mathbb{M}_0(\mathbf{p}_m)$	$\mathbf{e}^U \mathbb{M}_0(\mathbf{p})$
Harmonic mean	$(T_K + 1) \mathbb{M}_{-1}(\mathbf{p})$	$(T_K + 1) \bigwedge_{m=1}^K \mathbb{M}_{-1}(\mathbf{p}_m)$	$(T_K U + 1) \mathbb{M}_{-1}(\mathbf{p})$

Table 12.29: Some combination rules corresponding to Examples 12.20–12.23. The order-statistics family is indexed by $k \in \{1, \dots, K\}$. Here, $\mathbf{p} = (p_1, \dots, p_K)$ denotes the vector of p-values, and $\mathbf{p}_m$ represents the vector containing the first m values of $\mathbf{p}$. In the table, $p_{(k)}$ is the k -th smallest value of $\mathbf{p}$, while $p_{(\lambda_m)}^m$ is the $\lambda_m = \lceil mk/K \rceil$ ordered value of $\mathbf{p}_m$. The random variable U is uniformly distributed on the interval $[0, 1]$ independent of the p-values. The functions $\mathbb{M}_1$, $\mathbb{M}_0$ and $\mathbb{M}_{-1}$ respectively denote the arithmetic mean, the geometric mean, and the harmonic mean of the suitable dimension (we omit the dimension from the subscript). The value T_K is given by $T_K = \log K + \log \log K + 1$ for $K \geq 2$. Another randomized combination rule $\frac{1}{2-2U} \mathbb{M}_1(\mathbf{p})$ can be derived from (12.2).

Meng [1994] based on convex ordering inequalities; our proof is shorter and simpler. The full proof of Proposition 12.1 can be found in Wang [2024b], where (iv) $\Leftrightarrow$ (v) follows from Shaked and Shantikumar [2007, Theorem 4.A.2], (iv) $\Rightarrow$ (i) is based on Mao et al. [2019, Theorem 5] and (iv) $\Rightarrow$ (iii) is based on the refinement of Strassen’s theorem in Nutz et al. [2024, Theorem 3.1]. Mid p-values were studied by Lancaster [1952], and some methods of combining independent mid p-values were proposed by Rubin-Delanchy et al. [2019]. The posterior predictive p-values were proposed by Meng [1994]. The content of Section 12.2 is mainly based on Vovk et al. [2022], and the content of Sections 12.3–12.4 is based on Gasparin et al. [2025].

The necessary and sufficient condition for (12.12) to hold for convex f is given in Theorem 3.2 of Wang and Wang [2016]. This result is also needed to prove Theorem 12.18. The proof of Theorem 12.18 and justifications for statements in Section 12.2.3 are found in Section 6 of Vovk et al. [2022].

Chapter 13

False coverage rate control using e-confidence intervals

This chapter studies confidence intervals formulated by e-values, that we call e-confidence intervals. These will often not be intervals, but confidence sets (regions), but we still call them confidence intervals for simplicity (and to enable the common abbreviation CI).

We will see that these e-confidence intervals have additional properties compared to the usually constructed confidence intervals. In this chapter, we let $\mathcal{P}$ denote the set of all possible data distributions. The data will be drawn from some $\mathbb{P}^\dagger \in \mathcal{P}$, and we will be interested in estimating $\vartheta(\mathbb{P}^\dagger)$, where $\vartheta : \mathcal{P} \rightarrow \Theta$ is a predefined functional of interest (such as the mean, median, and so on). Since $\mathbb{P}^\dagger$ is unknown, we will consider generic distributions $\mathbb{P} \in \mathcal{P}$, as in Chapter 9. Let

$$\mathcal{P}_\theta := \{\mathbb{P} \in \mathcal{P} : \vartheta(\mathbb{P}) = \theta\}$$

denote the set of all distributions whose functional equals θ . We continue to write $\mathcal{K} = [K]$ as in the previous chapters.

13.1 Defining and constructing e-CIs

We first define e-confidence intervals as a generalization of the usual confidence intervals.

Definition 13.1: Families of e-variables and e-confidence intervals (e-CIs)

We say that $\{E(\theta)\}_{\theta \in \Theta}$ is a *family of e-variables* for $\{\mathcal{P}_\theta\}_{\theta \in \Theta}$ if for each $\theta \in \Theta$, $E(\theta)$ is an e-variable for $\mathcal{P}_\theta$.

For a fixed $\alpha \in [0, 1]$, a set $C(\alpha) \subseteq \Theta$ is called a $(1 - \alpha)$ -confidence interval (CI) for a functional ϑ if $\mathbb{P}(\vartheta(\mathbb{P}) \in C(\alpha)) \geq 1 - \alpha$ for all $\mathbb{P} \in \mathcal{P}$. We say that $\{C(\alpha)\}_{\alpha \in [0, 1]}$ is a *family of confidence intervals* for ϑ if for every $\alpha \in [0, 1]$, $C(\alpha)$ is a $(1 - \alpha)$ -CI for ϑ .

For a fixed $\alpha \in [0, 1]$, we say that $C(\alpha)$ is an $(1 - \alpha)$ -e-CI if there exists a family of e-variables $\{E(\theta, \alpha)\}_{\theta \in \Theta}$ for $\{\mathcal{P}_\theta\}_{\theta \in \Theta}$ such that

$$C(\alpha) = \left\{ \theta \in \Theta : E(\theta, \alpha) < \frac{1}{\alpha} \right\}.$$

We often omit the level $(1 - \alpha)$ for CIs and e-CIs, which should be clear from the notation $C(\alpha)$ or from the context.

We say that $\{C(\alpha)\}_{\alpha \in [0, 1]}$ is a family of *e-confidence intervals (e-CIs)* if $C(\alpha)$ is an e-CI for every $\alpha \in [0, 1]$. Finally, a family of e-CIs called *level-free* if the above e-variable $E(\theta, \alpha)$ does not depend on α ; in this case we simply write it $E(\theta)$.

Without loss of generality, we can take $C(0) = \Theta$, $C(1) = \emptyset$. Moreover, we usually would assume $C(\alpha) \supseteq C(\alpha')$ if $\alpha \leq \alpha'$; this is automatically true for a level-free family of e-CIs.

It is straightforward to observe that every e-CI is a CI and every family of e-CIs is a family of CIs: for the latter case, observe that for any $\mathbb{P} \in \mathcal{P}_\theta$, we have $\mathbb{P}(\theta \notin C(\alpha)) = \mathbb{P}(E(\theta, \alpha) \geq 1/\alpha) \leq \alpha$ by Markov's inequality, since $\mathbb{E}[E(\theta, \alpha)] \leq 1$ for any $\mathbb{P} \in \mathcal{P}_\theta$. (This also explains why we formulated $C(\alpha)$ using $< 1/\alpha$ instead of $\leq 1/\alpha$; our choice leads to a smaller CI.)

Moreover, every CI is actually an e-CI by using

$$E(\theta, \alpha) = \mathbb{1}_{\{\theta \in C(\alpha)\}},$$

which is an all-or-nothing e-variable in Section 3.1. Therefore, the more interesting object to us is the level-free e-CI families.

Example 13.2: e-CI for Gaussian mean with unknown variance

We continue from Example 3.24 to provide an intriguing e-CI for a Gaussian mean μ , with an unknown variance σ^2 . The following is an e-CI that can be constructed with a single observation, i.e. $n = 1$:

$$X_1 \pm \alpha^{-1} \exp((X_1^2 - 1)/2).$$

It is formed via the family of e-variables $\{|X_1 - \mu| \exp((1 - X_1^2)/2)\}_{\mu \in \mathbb{R}}$ for testing $\{\bigcup_{\sigma > 0} N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}}$. In contrast, the classical t-CI for μ needs at least two observations to yield a nontrivial interval.

We have already seen one general construction of e-CIs: those obtained via universal inference in Section 5.5. Here, we develop two more examples: via stopped confidence sequences and calibrated CIs.

13.1.1 E-CIs from stopped confidence sequences

Let us first define the central concept of a *confidence sequence*. The setup here is inherently sequential, as in Chapter 7. One observes an increasing amount of data from an unknown distribution $\mathbb{P}$, and desires to estimate its functional ϑ . Rather than a single σ -algebra $\mathcal{F}$, one must now consider a filtration $\{\mathcal{F}_t\}_{t \geq 0}$, which is a nested sequence of σ -algebras, indicating obtaining an increasing amount of information with the passage of time. A stopping time τ is a random variable such that $\{\tau \leq t\}$ is $\mathcal{F}_t$ -measurable for all $t \geq 0$.

Definition 13.3

Given $\alpha \in [0, 1]$, a $(1 - \alpha)$ -confidence sequence for a functional ϑ is a sequence of sets $(C^t(\alpha))_{t \in \mathbb{N}}$ such that $C^\tau(\alpha) \subseteq \Theta$ is an $(1 - \alpha)$ -CI for any stopping time τ . It also has the following equivalent definition: for any $\mathbb{P} \in \mathcal{P}$, we have

$$\mathbb{P}(\vartheta(\mathbb{P}) \in C^t(\alpha) \text{ for all } t \in \mathbb{N}) \geq 1 - \alpha.$$

A universal way to construct such an object is using a family of e-processes, which are sequential versions of e-values, as introduced in Chapter 7.

For any fixed α , let $E(\theta, \alpha) \equiv \{E_t(\theta, \alpha)\}_{t \geq 0}$ denote an e-process for $\mathcal{P}_\theta$. Then, Ville's inequality implies that the set

$$C^t(\alpha) = \left\{ \theta \in \Theta : E_t(\theta, \alpha) < \frac{1}{\alpha} \right\}$$

is a confidence sequence for ϑ .

Proposition 13.4

If $\{C^t(\alpha)\}_{t \in \mathbb{N}}$ is a confidence sequence as constructed above, then for any stopping τ , $C^\tau(\alpha)$ is an e-CI. Further, if $E(\theta, \alpha)$ does not depend on α , then $\{C^\tau(\alpha)\}_{\alpha \in [0, 1]}$ is a level-free family of e-CIs.

Proposition 13.4 directly follows from Ville's inequality stated in Fact 7.6.

13.1.2 E-CIs by calibrating CIs

Recall the notion of a calibrator from Definition 2.3. Define f^{-1} to be the right inverse of the calibrator f , i.e., $f^{-1}(x) := \sup\{p : f(p) \geq x\}$. When f is invertible, f^{-1} is the inverse of f . Using f^{-1} , we can convert any CI to an e-CI. Before we do so, define two properties about any set-valued CI function $C : [0, 1] \mapsto 2^\Theta$. Let C be *decreasing* if, for any $\alpha, \beta \in [0, 1]$, $\alpha \leq \beta$ implies that $C(\alpha) \supseteq C(\beta)$. Further, define *continuous from below* to be the property that for any $\alpha \in [0, 1]$, $C(\alpha) = \bigcup_{\beta > \alpha} C(\beta)$.

Theorem 13.5

Let $C : [0, 1] \mapsto 2^\Theta$ be a decreasing function that is continuous from below such that $C(\alpha)$ produces a $(1 - \alpha)$ -CI, and f be a calibrator with right inverse f^{-1} . Then, the following set $C^{\text{cal}}(\alpha)$ is a $(1 - \alpha)$ -e-CI:

$$C^{\text{cal}}(\alpha) = C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right) \supseteq C(\alpha).$$

Moreover, $\{C^{\text{cal}}(\alpha)\}_{\alpha \in [0, 1]}$ is a level-free family of e-CIs.

Theorem 13.5 exploits the duality between CIs and families of tests (or families of p-values). For each $\theta \in \Theta$, we can test the null hypothesis $\mathcal{P}_\theta$. The CI then corresponds to the values of θ for which the aforementioned test does not reject at level α . We can invert this process to use the CI to define a p-value for every null hypothesis $\mathcal{P}_\theta$. We can then calibrate the implicit p-value to yield an e-value, from which we can produce our e-CI. This is formalized in the proof below.

Proof of Theorem 13.5.

The inclusion $C^{\text{cal}}(\alpha) \supseteq C(\alpha)$ follows directly from the fact that $f(p) \leq 1/p$ for all $p \in [0, 1]$ (explained in Section 2.3), and hence $f^{-1}(1/\alpha) \geq \alpha$, and thus $C(f^{-1}(1/\alpha)) \supseteq C(\alpha)$. We next prove that $\{C^{\text{cal}}(\alpha)\}_{\alpha \in [0, 1]}$ is a level-free family of e-CIs.

For any $\theta \in \Theta$, note that the following is a p-variable for $\mathcal{P}_\theta$:

$$P^{\text{dual}}(\theta) := \inf\{\alpha \in [0, 1] : \theta \notin C(\alpha)\}.$$

Consequently, $E^{\text{cal}}(\theta) := f(P^{\text{dual}}(\theta))$ is an e-variable. Hence,

$$\left\{\theta \in \Theta : E^{\text{cal}}(\theta) < \frac{1}{\alpha}\right\}$$

yields a $(1 - \alpha)$ -e-CI. The theorem now follows because

$$\begin{aligned} \left\{\theta \in \Theta : E^{\text{cal}}(\theta) < \frac{1}{\alpha}\right\} &= \left\{\theta \in \Theta : f(P^{\text{dual}}(\theta)) < \frac{1}{\alpha}\right\} \\ &= \left\{\theta \in \Theta : P^{\text{dual}}(\theta) > \max\left\{p : f(p) \geq \frac{1}{\alpha}\right\}\right\} \\ &= \left\{\theta \in \Theta : P^{\text{dual}}(\theta) > f^{-1}\left(\frac{1}{\alpha}\right)\right\} \\ &= C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right). \end{aligned}$$

Above, the third equality is a result of f being decreasing and upper semicontinuous at $1/\alpha$; hence the supremum is achieved and the equality holds. The final equality is because C is decreasing and continuous from below at $f^{-1}(1/\alpha)$; if $P^{\text{dual}}(\theta) = f^{-1}(1/\alpha)$, then $\theta \notin C_i(P^{\text{dual}}(\theta))$ by C being continuous from below.

13.2 The e-BY procedure for false coverage rate

13.2.1 The goal of FCR control

Consider a scientist who observes some data, $\mathbf{X} = (X_1, \dots, X_K)$ drawn from some distribution $\mathbb{P} \in \mathcal{P}$. The scientist is potentially interested in the values of K of its functionals $\vartheta_1, \dots, \vartheta_K$ through the parameter values $\boldsymbol{\theta}^* := (\theta_1^*, \dots, \theta_K^*)$, where $\theta_i^* := \vartheta_i(\mathbb{P})$. The functionals $\vartheta_1, \dots, \vartheta_K$ could each take values in different spaces, but this complicates notation, and we will not explicitly need these sets later on. Each ϑ_i need not be bijective; for example, it could capture the median of a distribution.

The scientist may be interested in identifying the $L \geq 1$ indices corresponding to the largest L values of $\boldsymbol{\theta}^*$, assuming they are real-valued. Or they may be interested in any index k such that θ_k^* is larger than some prespecified (or data dependent) threshold. In short, the scientist may not know which indices are of interest to them before observing the data, and it is quite likely that after observing the data, only a small fraction of indices are actually of interest.

For each $i \in \mathcal{K}$, we assume that from X_i , the scientist can construct a $(1 - \alpha)$ -confidence interval for θ_i^* .

The scientist uses the data $\mathbf{X}$ to select a subset of “interesting” parameters, $S \subseteq \mathcal{K}$, using some potentially complex data-dependent selection rule $\mathcal{S} : \mathbf{X} \mapsto S$. The scientist must then devise confidence levels for the CI of each selected parameter, $\{\alpha_i\}_{i \in S}$, that *can depend on the data* $\mathbf{X}$. The *false coverage proportion* (FCP) and *false coverage rate* (FCR) of such a procedure are defined by

$$\text{FCP} = \text{FCP}(S, \{\alpha_i\}_{i \in S}) := \frac{\sum_{i \in S} \mathbb{1}_{\{\theta_i^* \notin C_i(\alpha_i)\}}}{|S| \vee 1}, \quad \text{FCR} = \mathbb{E}^{\mathbb{P}} [\text{FCP}].$$

Note the obvious similarity between the concepts of FCR and FCP and between the concepts of FDR and FDP in Chapter 9; all of these terms depend on $\mathbb{P} \in \mathcal{P}$ through $\theta_i^* = \vartheta_i(\mathbb{P})$ and $\mathbb{E}^{\mathbb{P}}$. If every α_i equals the same constant (say γ), we use the more succinct notation $\text{FCP}(S, \gamma)$ to abbreviate $\text{FCP}(S, \{\alpha_i\}_{i \in S})$.

Our goal is to design a method for choosing $\{\alpha_i\}_{i \in S}$ that guarantees $\text{FCR} \leq \delta$ (for all $\mathbb{P} \in \mathcal{P}$) at a predefined level $\delta \in [0, 1]$ provided by the scientists in advance, *regardless of what the selection rule $\mathcal{S}$ is, and in particular even if the rule is unknown and we only observe the selected set $S = \mathcal{S}(\mathbf{X})$.*

13.2.2 The e-BY procedure

Our primary point of comparison is the so-called Benjamini–Yekutieli (BY) procedure. The BY procedure’s choice of $\{\alpha_i\}_{i \in S}$ and resulting guarantees depend upon assumptions (or knowledge) of the dependence structure in $\mathbf{X}$ and the selection algorithm $\mathcal{S}$. Under certain restrictions (omitted here for brevity) on $\mathcal{S}$, the BY procedure sets $\alpha_i = \delta|S|/K$ to ensure that the FCR is controlled at level δ for mutually independent $X_1, \dots, X_K$. However, when no such assumptions can be made (i.e., under arbitrary dependence and an unknown selection rule) the BY procedure sets $\alpha_i = \delta|S|/(K\ell_K)$, where $\ell_K := \sum_{i=1}^K i^{-1} \approx \log K$ is the K -th harmonic number. Clearly, the BY procedure produces much more conservative CIs when no assumptions can be made about dependence or selection.

The above facts are reminiscent of the BH procedure discussed in Section 9.1.4, and one may, analogously to the situation of e-BH versus BH procedures in Chapter 9, hope that an e-CI based procedure has FCR control without the additional assumptions and the multiplicative penalty ℓ_K . This will be addressed next.

Now, we formally define the e-BY procedure as follows.

Definition 13.6

The *e-BY procedure* at level $\delta \in [0, 1]$ sets $\alpha_i = \delta|S|/K$ for each $i \in S$.

We show that a FCR bound can be proven quite simply given the fact that e-CIs are constructed for each selected parameter.

Theorem 13.7

Let $\{C_i(\alpha)\}_{\alpha \in [0,1]}$ be a level-free family of e-CIs for each $i \in \mathcal{K}$. Then, the e-BY procedure ensures $\text{FCR} \leq \delta$ for any $\delta \in (0, 1)$ under any dependence structure between $X_1, \dots, X_K$, and for any selection rule $\mathcal{S}$. In fact, the e-BY procedure satisfies the stronger guarantee:

$$\mathbb{E}^{\mathbb{P}} \left[\sup_{S \in 2^{\mathcal{K}}} \sup_{\delta \in (0,1)} \frac{\text{FCP}(S, \delta |S|/K)}{\delta} \right] \leq 1 \quad \text{for all } \mathbb{P} \in \mathcal{P}.$$

Proof of Theorem 13.7

We directly show an upper bound for the FCR as follows:

$$\begin{aligned} \text{FCR} &= \mathbb{E}^{\mathbb{P}} \left[\frac{\sum_{i \in S} \mathbb{1}_{\{\theta_i^* \notin C_i(\delta |S|/K)\}}}{|S| \vee 1} \right] \\ &= \mathbb{E}^{\mathbb{P}} \left[\frac{\sum_{i \in \mathcal{K}} \mathbb{1}_{\{E_i(\theta_i^*) |S| \delta / K > 1\}} \cdot \mathbb{1}_{\{i \in S\}}}{|S| \vee 1} \right] \\ &\leq \sum_{i \in \mathcal{K}} \mathbb{E}^{\mathbb{P}} \left[\frac{E_i(\theta_i^*) |S| \delta}{K(|S| \vee 1)} \right] = \sum_{i \in \mathcal{K}} \frac{\delta}{K} \mathbb{E}^{\mathbb{P}} \left[E_i(\theta_i^*) \cdot \frac{|S|}{|S| \vee 1} \right] \leq \delta, \end{aligned}$$

where the first inequality is because $\mathbb{1}_{\{x > 1\}} \leq x$ for all $x \geq 0$. The second inequality is a result of the definition of the e-value for θ_i^* having its expectation under $\mathbb{P}$ be upper bounded by 1. This achieves our desired bound.

The proof of the more general claim follows by an easy amendment of the above proof.

We note in passing that FCR control of the e-BY procedure implies FDR control of the e-BH procedure, while the converse is not true.

13.3 Combining CIs via majority vote

Given K uncertainty sets that are arbitrarily dependent — for example, confidence intervals for an unknown parameter obtained with K different estimators, or prediction sets obtained via conformal prediction based on K different algorithms on shared data — we now address the question of how to efficiently combine them in a black-box manner to produce a single uncertainty set. We present a simple and broadly applicable majority vote procedure that produces a merged set with nearly the same error guarantee as the input sets. We then extend this core idea in a few ways: we show that weighted averaging can be a powerful way to incorporate prior information, and a simple randomization trick produces strictly smaller merged sets without altering the coverage guarantee. Further improvements can be obtained if the sets are exchangeable. Underlying all of these methods are e-values and various versions of Markov’s inequality.

Formally, we start with a collection of K different sets C_k (one from each “agent”) for the same target quantity c , each having a confidence level $1 - \alpha$ for some $\alpha \in (0, 1)$:

$$\mathbb{P}(c \in C_k) \geq 1 - \alpha, \quad k \in \mathcal{K}. \quad (13.1)$$

The probability measure $\mathbb{P}$ in this section is the one that generates the data. We say that C_k has *exact* coverage if $\mathbb{P}(c \in C_k) = 1 - \alpha$. Since the sets C_k are based on data, they are random quantities by definition, but c can be either fixed or random; for example, in the case of confidence sets for a target functional of a distribution it is fixed, but it is random in the case of prediction sets for an outcome (e.g., conformal prediction). Our method will be agnostic to such details.

Our objective as the “aggregator” of uncertainty is to combine the sets in a black-box manner in order to create a new set that exhibits favorable properties in both coverage and size. A first (trivial) solution is to define the set C^J as the union of the others:

$$C^J = \bigcup_{k=1}^K C_k.$$

Clearly, C^J respects the property defined in (13.1), but the resulting set is typically too large and has significantly inflated coverage. On the other hand, the set resulting from the intersection $C^I = \bigcap_{k=1}^K C_k$ is narrower, but typically has inadequate coverage — it guarantees at least $1 - K\alpha$ coverage by the Bonferroni inequality, but this is uninformative when K is large.

If the aggregator knows the $(1 - \alpha)$ -confidence intervals not just for a single α but for every $\alpha \in (0, 1)$, using the duality between CIs and tests, one can calculate a p-value for each $\theta \in \Theta$, each agent $k \in \mathcal{K}$. We have already seen many ways to combine dependent p-values, for example, by averaging them and multiplying by two, and these can be used to combine these p-values across agents into a single one and then obtain a $(1 - \alpha)$ -confidence interval for any α of the aggregator’s choice. The current section addresses the setting where only a single interval is known from each agent, ruling out the above averaging schemes.

13.3.1 The majority vote procedure

Let the observed data Z lie in a sample space $\mathcal{Z}$, while our target c is a point in a measure space $(\mathcal{S}, \mathcal{A}, \nu)$, where $\mathcal{A}$ is a σ -algebra on $\mathcal{S}$ and ν is a measure on $\mathcal{S}$. As mentioned earlier, it is important to note that c can itself be a random variable. The sets $C_k = C_k(z) \subseteq \mathcal{S}$, $k \in \mathcal{K}$, based on the observed data, follow the property (13.1), where the probability refers to the joint distribution (Z, c) (or only Z if c is fixed). Let us define a new set C^M that includes all points *voted* by at least half of the sets:

$$C^M := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\{s \in C_k\}} > \frac{1}{2} \right\}. \quad (13.2)$$

Theorem 13.8

Let $C_1, \dots, C_K$ be $K \geq 2$ different sets based on the observed data z , satisfying property (13.1). Then, the set C^M defined in (13.2) has coverage of at least $1 - 2\alpha$:

$$\mathbb{P}(c \in C^M) \geq 1 - 2\alpha. \quad (13.3)$$

Proof

Let $\phi_k = \phi_k(Z, c) = \mathbb{1}_{\{c \notin C_k\}}$ be a Bernoulli random variable such that $\mathbb{E}[\phi_k] \leq \alpha$, $k \in \mathcal{K}$. Thus $E_k = \phi_k/\alpha$ is an e-value, implying that $\bar{E} = (E_1 + \dots + E_K)/K$ is an e-value. Thus,

$$\mathbb{P}(c \notin C^M) = \mathbb{P}\left(\frac{1}{K} \sum_{k=1}^K \phi_k \geq \frac{1}{2}\right) = \mathbb{P}\left(\bar{E} \geq \frac{1}{2\alpha}\right) \leq 2\alpha,$$

by Markov’s inequality, concluding the proof.

Actually, a slightly tighter bound can be obtained if K is odd. In this case, for a point to be contained in C^M , it must be voted for by at least $\lceil K/2 \rceil$ of the other sets. This implies that, with the same arguments as used in Theorem 13.8, the probability of miscoverage is equal to

$$\alpha K / \lceil K/2 \rceil = 2\alpha K / (K + 1),$$

approaching (13.3) for large K .

The conservative threshold of $1/2$ is needed to handle arbitrary dependence between the input sets. If they were independent, the right-hand side of (13.2) can be replaced by the α -quantile of a Binomial distribution with K trials of probability $1 - \alpha$.

Theorem 13.8 is known to be tight in a worst-case sense; a simple example shows that if K is odd and if the sets have a particular joint distribution, then the error will equal $(\alpha K)/\lceil K/2 \rceil$. This worst-case distribution allows for only two types of cases: either all agents provide the same set that contains c (so majority vote is correct), or $\lfloor K/2 \rfloor$ sets contain c but the others do not (so majority vote is incorrect). Each of the latter cases happens with some probability p , so the probability that majority vote makes an error is $\binom{K}{\lfloor K/2 \rfloor + 1} p$. The probability that any particular agent makes an error is $\binom{K-1}{\lfloor K/2 \rfloor} p$, which we set as our choice of α , and then we see that the probability of error for majority vote simplifies to $\alpha K / \lceil K/2 \rceil$. Despite the apparent tightness of majority vote in the worst-case, we will develop several ways to improve this procedure in non-worst-case instances, while retaining the same worst-case performance.

Remark 13.9 (When does majority vote overcover and when does it undercover?). While the worst-case theoretical guarantee for majority vote is a coverage level of $1 - 2\alpha$, sometimes it will get close to the desired $1 - \alpha$ coverage, and sometimes it may even overcover, achieving coverage closer to one. Here, we provide some intuition for when to expect each type of behavior in practice assuming $\alpha < 1/2$, foreshadowing many results to come. If the sets are actually independent (or nearly so), we should expect the method to have coverage more than $1 - \alpha$. This can be seen via an application of Hoeffding's inequality in place of Markov's inequality in the proof of Theorem 13.8: since each ϕ_k has expectation (at most) α , we should expect $\frac{1}{K} \sum_{k=1}^K \phi_k$ to concentrate around α , and the probability that this average exceeds $1/2$ is exponentially small (as opposed to 2α), being at most $\exp(-2K(1/2 - \alpha)^2)$ by Hoeffding's inequality. In contrast, if the sets are identical (the opposite extreme of independence), clearly the method has coverage $1 - \alpha$. As argued in the previous paragraph, there is a worst-case dependence structure that forces majority to vote to have an error of (essentially) 2α . Finally, if the sets are exchangeable, it appears more likely that the coverage will be closer to $1 - \alpha$ than $1 - 2\alpha$. While one informal reason may be that exchangeability connects the two extremes of independence and being identical (with coverages close to 1 and $1 - \alpha$), a slightly more formal reason is that under exchangeability, we will later in this section actually devise a strictly tighter set C^E than C^M which also achieves the same coverage guarantee of $1 - 2\alpha$, thus making C^M itself likely to have a substantially higher coverage.

The above method and result can be easily generalized beyond the threshold value of $1/2$. We record it as a result for easier reference. For any $\tau \in [0, 1)$, let

$$C^\tau := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\{s \in C_k\}} > \tau \right\}. \quad (13.4)$$

Theorem 13.10

Let $C_1, \dots, C_K$ be $K \geq 2$ different sets satisfying property (13.1). Then,

$$\mathbb{P}(c \in C^\tau) \geq 1 - \alpha/(1 - \tau).$$

The proof follows the same lines as the original result outlined in Theorem 13.8 and is thus omitted. As expected, it can be noted that the obtained bounds decrease as τ increases. In fact, for larger values of τ , smaller sets will be obtained. One can check that this result also yields the right bound for the intersection ($\tau = 1 - 1/K$) and the union ($\tau = 0$). In certain situations, it is possible to identify an upper bound to the coverage of the set resulting from the majority vote.

Even if the input sets are intervals, the majority vote set may be a union of intervals. One can easily construct a simple aggregation algorithm to find this set quickly by sorting the endpoints of the input intervals and checking some simple conditions.

Frequently, the majority vote set is indeed an interval. In fact, it is easy to check that if $C_1, \dots, C_K$ are one-dimensional intervals and $\bigcap_{k=1}^K C_k \neq \emptyset$, then C^τ is an interval for any τ . See Figure 13.11.

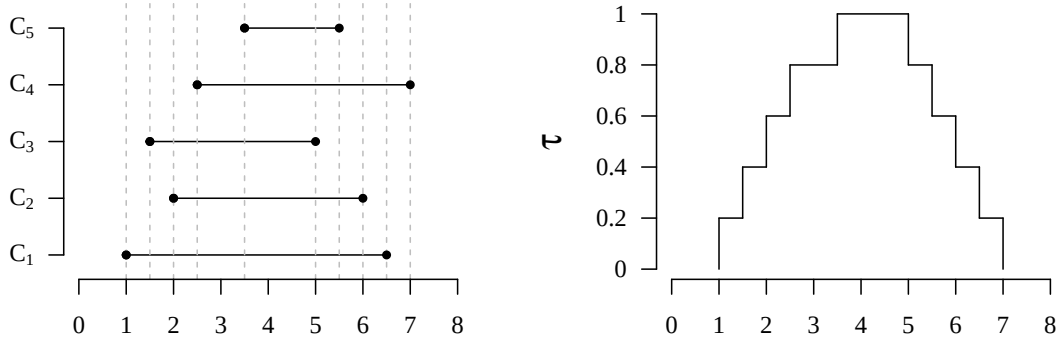


Figure 13.11: Visualization of the majority vote procedure when $\bigcap_{k=1}^K C_k \neq \emptyset$.

How large can the majority vote set be? One naive way to combine the K sets is to randomly select one of them as the final set; this method clearly has coverage $1 - \alpha$, and its length is in between their union and intersection, so it seems reasonable to ask how it compares to majority vote. Surprisingly, majority vote is not always strictly better than this approach in terms of the expected length of the set: consider, for example, three nested intervals C_1, C_2, C_3 of width 10, 8 and 3, respectively. The majority vote set is C_2 , with a length of 8, but randomly selecting an interval results in an average length of 7. However, we show next that the majority vote set cannot be more than twice as large. In addition, when the input sets are intervals, it is never wider than the largest interval.

Theorem 13.12

Let $\nu(C^\tau)$ be the measure (size) of C^τ in (13.4). Then, for all $\tau \in [0, 1)$,

$$\nu(C^\tau) \leq \frac{1}{K\tau} \sum_{k=1}^K \nu(C_k). \quad (13.5)$$

If the input sets are K one-dimensional intervals, then for all $\tau \in [\frac{1}{2}, 1)$,

$$\nu(C^\tau) \leq \max_k \nu(C_k).$$

Above, ν could be the Lebesgue measure (for intervals), or the counting measure (for discrete, categorical sets), for example. The proof uses the fact that the majority vote set is elementwise monotonic in its input sets, meaning that if any of the input sets gets larger, the majority vote set can never get smaller. From (13.5) we have that if $\tau = 1/2$, then the measure of the majority vote set is never larger than 2 times the average of the measure of the initial sets. This result is essentially tight as can be seen in the following example involving one-dimensional intervals. For odd K , let $(K + 1)/2$ intervals have a length L , while the rest have a length of nearly 0. The average length is then $(K + 1)L/(2K)$, and the majority vote has length L , whose ratio approaches $1/2$ for large K . In addition, (13.5) gives the right bound for the intersection ($\tau \uparrow 1$) and for the union ($\tau \uparrow 1/K$).

Consider a scenario similar to that of the last example. Suppose we have $K > 2$ confidence intervals, with odd K , such that $C_1 = \dots = C_{\lceil K/2 \rceil} = (-\varepsilon, 1)$ and $C_{\lceil K/2 \rceil + 1} = \dots = C_K = (0, 1 + \varepsilon)$, for some $\varepsilon > 0$. By definition, if $\tau = (K - 2)/(2K) < 1/2$, then the set C^τ coincides with the union of the initial sets, which is larger than the input intervals and has width $1 + 2\varepsilon$. On the other hand, choosing $\tau = 1/2$ we obtain a set

whose size is $1 + \varepsilon$, matching the width of the input sets.

13.3.2 Exchangeable, randomized and weighted variants

Surprisingly, when $C_1, \dots, C_K$ are not independent, but are exchangeable, something better than a naive majority vote can be accomplished. To describe the method, let $C^M(1 : k)$ denote the majority vote of sets $C_1, \dots, C_k$. Now define

$$C^E := \bigcap_{k=1}^K C^M(1 : k),$$

which can be equivalently represented as

$$C^E = \left\{ s \in \mathcal{S} : \frac{1}{k} \sum_{j=1}^k \mathbb{1}_{\{s \in C_j\}} > \frac{1}{2} \text{ holds for all } k \in \mathcal{K} \right\}.$$

Essentially, C^E is formed by the intersection of sets obtained through sequential processing of the sets derived from the majority vote.

Theorem 13.13

If $C_1, \dots, C_K$ are $K \geq 2$ exchangeable sets having coverage $1 - \alpha$, then C^E is a $1 - 2\alpha$ uncertainty set, and it is never worse than majority vote ($C^E \subseteq C^M$).

The proof mimics that of Theorem 13.8, but uses the exchangeable Markov inequality from Chapter 5 in place of Markov's inequality. This method shares a same idea as in designing the ex-p-merging functions in Section 12.3.

This result points at a simple way of improving majority vote for arbitrarily dependent sets: process them in a random order. To elaborate, let π be a uniformly random permutation of $\{1, \dots, K\}$ that is independent of the K sets, and define

$$C^\pi := \bigcap_{k=1}^K C^M(\pi(1) : \pi(k)).$$

Since $C^M(\pi(1) : \pi(K)) = C^M(1 : K)$, C^π is also never worse than majority vote despite satisfying the same coverage guarantee:

Corollary 13.14

If $C_1, \dots, C_K$ are $K \geq 2$ arbitrarily dependent uncertainty sets having coverage $1 - \alpha$, and π is a uniformly random permutation independent of them, then C^π is a $1 - 2\alpha$ uncertainty set, and it is never worse than majority vote ($C^\pi \subseteq C^M$).

The proof follows as a direct corollary of Theorem 13.13 by noting that the random permutation π induces exchangeability of the sets (the joint distribution of every permutation of sets is the same, due to the random permutation). Of course, if the sets were already “randomly labeled” 1 to K (for example, to make sure there was no special significance to the labels), then the aggregator does not need to perform an extra random permutation.

Moving in a different direction below, we demonstrate that the majority vote can be improved with the aim of achieving a tighter set through the use of independent randomization, while maintaining the same coverage level.

Let U be an independent random variable that is distributed uniformly on $[0, 1]$, and let u be a realization. We then define a new set C^R as

$$C^R := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\{s \in C_k\}} > \frac{1}{2} + \frac{u}{2} \right\}. \quad (13.6)$$

As a small variant, define

$$C^U := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\{s \in C_k\}} > u \right\}. \quad (13.7)$$

Theorem 13.15

Let $C_1, \dots, C_K$ be $K \geq 2$ different sets satisfying the property (13.1). Then, the set C^R has coverage at least $1 - 2\alpha$ and is never larger than C^M , while the set C^U has coverage at least $1 - \alpha$ and is never smaller than C^R .

The proof follows that of Theorem 13.8, but uses the randomized Markov inequality in Theorem 2.9 from Chapter 2 in place of Markov's inequality. This is related to randomized p-merging functions in Section 12.4, in particular to the randomized order-family combination. Even though C^U does not improve on C^M , we include it since it involves random thresholding and delivers the same coverage as the input sets, a feature that we do not know how to obtain without randomization.

It is not unusual for each interval to be assigned distinct “weights” (importances) in the voting procedure. This can occur, for instance, when prior studies empirically demonstrate that specific methods for constructing uncertainty sets consistently outperform others. Alternatively, a researcher might assign varying weights to the sets based on their own prior insights.

Since it is desirable to have as small an interval as possible if coverage (13.1) is respected, we would like to assign a higher weight to intervals of smaller size. However, the weights must be assigned before seeing the sets.

Assume as before that the sets $C_1, \dots, C_K$ based on the observed data follow the property (13.1). In addition, let $w = (w_1, \dots, w_K)$ be a set of weights, such that

$$w_k \in [0, 1] \quad \text{and} \quad \sum_{k=1}^K w_k = 1.$$

These weights can be interpreted as the aggregator's prior belief in the quality of the received sets. A higher weight signifies that we attribute greater importance to that specific interval. As before, let U be an independent random variable that is distributed uniformly on $[0, 1]$, and let u be a realization. We then define a new set C^W as:

$$C^W := \left\{ s \in \mathcal{S} : \sum_{k=1}^K w_k \mathbb{1}_{\{s \in C_k\}} > \frac{1}{2} + \frac{u}{2} \right\}. \quad (13.8)$$

Theorem 13.16

Let $C_1, \dots, C_K$ be $K \geq 2$ different sets satisfying property (13.1). Then, the set C^W defined in (13.8) has coverage of at least $1 - 2\alpha$:

$$\mathbb{P}(c \in C^W) \geq 1 - 2\alpha.$$

In addition, let $\nu(C^W)$ be the measure associated with the set C^W , then

$$\nu(C^W) \leq 2 \sum_{k=1}^K w_k \nu(C_k). \quad (13.9)$$

If the weights are equal to $w_k = 1/K$ for all $k \in \mathcal{K}$, then the set C^W coincides with the set C^R defined in (13.6) and it is a subset of that in (13.2). This means that in the case of a democratic (equal-weighted) vote $C^R \subseteq C^M$, since C^M is obtained by choosing $u = 0$. Furthermore, (13.9) says that the measure of the set obtained using the weighted majority method cannot be more than twice the average measure obtained by randomly selecting one of the intervals with probabilities proportional to w . When there are only two sets and $u = 0$, the weighted majority vote set will correspond to the set with the greater weight.

Proposition 13.17

If $C_1, \dots, C_K$ are $K \geq 2$ different sets having coverage $1 - \alpha_1, \dots, 1 - \alpha_K$ (possibly unknown), then the set C^W defined in (13.8) has coverage

$$\mathbb{P}(c \in C^W) \geq 1 - 2 \sum_{k=1}^K w_k \alpha_k.$$

In particular, this implies that the majority vote of asymptotic $(1 - \alpha)$ intervals has asymptotic coverage at least $(1 - 2\alpha)$.

The proof is identical to that of Theorem 13.16, with the exception that the expected value for the variables ϕ_k is equal to α_k , and is thus omitted. If the α_k levels are known (which they may not be, unless the agents report it and are accurate), and if one in particular wishes to achieve a target level $1 - \alpha$, then it is always possible to find weights $(w_1, \dots, w_K)$ that achieve this as long as $\alpha/2$ is in the convex hull of $(\alpha_1, \dots, \alpha_K)$.

Suppose now that we have K different *confidence sequences* for a parameter that need to be combined into a single confidence sequence. For this setting we show a simple result:

Proposition 13.18

Given K different $1 - \alpha$ confidence sequences for the same parameter that are being tracked in parallel, their majority vote set is a $1 - 2\alpha$ confidence sequence.

It may not be initially apparent how to deal with the time-uniformity in the definitions. The proof proceeds by first observing that an equivalent definition of a confidence sequence is a confidence interval that is valid at any arbitrary stopping time τ (here the underlying filtration is implicitly that generated by the data itself). The sequence $(C_k^{(t)})_{t \geq 1}$ is a confidence sequence if and only if for every stopping time τ , $\mathbb{P}(c \in C_k^{(\tau)}) \geq 1 - \alpha$, for all $k \in \mathcal{K}$. Now, the proof follows by applying our earlier results.

Now, let the data be fixed, but consider the setting where an unknown number of confidence sets arrive one at a time in a random order, and need to be combined on the fly. Now, we propose to simply take a majority vote of the sequences we have seen thus far. Borrowing terminology from earlier, denote by

$$C^E(1 : t) := \bigcap_{i=1}^t C^M(1 : i).$$

We claim that the above sequence of sets is actually a $(1 - 2\alpha)$ confidence sequence for c .

Theorem 13.19

Given an exchangeable sequence of confidence sets $C_1, C_2, \dots$ (or confidence sets arriving in a uniformly random order), the sequence of sets formed by their “running majority vote” $(C^E(1 : t))_{t \geq 1}$ is a $(1 - 2\alpha)$ confidence sequence:

$$\mathbb{P}(\exists t \geq 1 : c \notin C^E(1 : t)) \leq 2\alpha.$$

Such a result is useful when derandomizing a statistical procedure, by repeating it many times one by one, and attempting to combine the results of these repetitions on the fly.

13.3.3 Median of Midpoints and Median of Median of Means

A special case for the majority vote set in (13.2) occurs if the input sets are (or can be bounded by) intervals of the same width, and can thus each can be represented as their midpoint plus/minus a half-width, then the following rule results in a more computationally efficient procedure. We call this method the “Median of Midpoints”.

Theorem 13.20

Suppose the input sets $C_1, \dots, C_K$ are intervals having the same width, and let C^M be their majority vote set. Let the midpoint of C_k be denoted by c_k . If K is odd, let $c_{(\lceil K/2 \rceil)}$ denote their median, and let $C^{(K/2)}$ denote the interval whose midpoint is $c_{(\lceil K/2 \rceil)}$. If K is even, let $C^{(K/2)}$ be defined by the intersection of the sets whose midpoints are $c_{(K/2)}$ and $c_{(1+K/2)}$. Then, $C^{(K/2)} \supseteq C^M$ and is hence also a $1 - 2\alpha$ uncertainty set. Further, $C^{(K/2)}$ has at most the same width as the input sets. If $\bigcap_{k=1}^K C_k \neq \emptyset$, then $C^{(K/2)} = C^M$.

Since the paper has so far focused on uncertainty sets, we first describe a general result that derandomizes point estimators “directly”.

Corollary 13.21

Suppose $\hat{\theta}_1, \dots, \hat{\theta}_K$ are K univariate point estimators of θ that are each built using n data points and satisfy a high probability concentration bound:

$$\mathbb{P}(|\hat{\theta}_k - \theta| \leq w(n, \alpha)) \geq 1 - \alpha,$$

for some function w . Then, their median $\theta_{(\lceil K/2 \rceil)}$ satisfies

$$\mathbb{P}(|\hat{\theta}_{(\lceil K/2 \rceil)} - \theta| \leq w(n, \alpha)) \geq 1 - 2\alpha.$$

Further, if $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K, \hat{\theta}_{K+1} \dots$ are exchangeable, then

$$\mathbb{P}(\forall K \geq 1 : |\hat{\theta}_{(\lceil K/2 \rceil)} - \theta| \leq w(n, \alpha)) \geq 1 - 2\alpha.$$

We present an application of this idea to derandomizing the median of means estimator. The aim of the celebrated median-of-means procedure is to produce estimators of the mean that have sub-Gaussian tails, despite the underlying unbounded data having only a finite variance. The setup assumes n iid data points from an unknown univariate distribution P with unknown finite variance, whose mean μ we seek to estimate. The method works by randomly dividing the n points into B buckets of roughly n/B points each. One takes the mean within each bucket, and then calculates the median $\hat{\mu}^{\text{MoM}}$ of those numbers.

The method does not produce confidence intervals for the mean; indeed, one needs to know a bound on the variance σ^2 for any nonasymptotic CI to exist. But the point estimator $\hat{\mu}^{\text{MoM}}$ satisfies the following sub-Gaussian tail bound: for any $t \geq 0$,

$$\mathbb{P}(|\hat{\mu}^{\text{MoM}} - \mu| \geq C\sigma\sqrt{t/n}) \leq 2e^{-t},$$

for a constant $C = \sqrt{\pi} + o(1)$, where the $o(1)$ term vanishes as $B, n/B \rightarrow \infty$. This is a much faster rate than the $1/t^2$ on the right-hand side obtained for the sample mean via Chebyshev’s inequality. However, one drawback of the method is that it is randomized, meaning that it depends on the random split of the data into buckets.

Here, we point out that Theorem 13.21 yields a simple way to derandomize $\hat{\mu}^{\text{MoM}}$. We simply repeat the median-of-means procedure independently K times to obtain exchangeable estimators $\{\hat{\mu}_k^{\text{MoM}}\}_{k=1}^K$, and then report the “median of median of means” (MoMoM):

$$\hat{\mu}_K^{\text{MoMoM}} := \hat{\mu}_{\lceil K/2 \rceil}^{\text{MoM}}.$$

Clearly, $\{\hat{\mu}_k^{\text{MoM}}\}_{k=1}^K$ form an exchangeable set, whose empirical distribution will stabilize as $K \rightarrow \infty$, and thus whose median will also be stabilize for large K .

Corollary 13.22

Under the setup described above, the *median of median of means* satisfies a nearly identical sub-Gaussian behavior as the original:

$$\mathbb{P}\left(\exists K \geq 1 : |\hat{\mu}_K^{\text{MoMoM}} - \mu| \geq C\sigma\sqrt{t/n}\right) \leq 4e^{-t}.$$

The simultaneity over K allows us to choose any sequence of constants $k_1 \leq k_2 \leq \dots$ and produce the sequence of estimators $\hat{\mu}_{k_1}^{\text{MoMoM}}, \hat{\mu}_{k_2}^{\text{MoMoM}}, \dots$, and plot this sequence as it is being generated. We can then stop the derandomization whenever the plot appears to stabilize, with the guarantee now holding for whatever data-dependent random K we stopped at. Clearly, one can use the same technique to derandomize other random estimators by taking their median while “importing” their nonasymptotic bounds. We will see an empirical example in the next section.

13.3.4 Empirical examples

We now present a simulation to investigate the relative performance of our proposed methods. We apply the majority vote procedure on simulated high-dimensional data with $n = 100$ observations and $p = 120$ regressors. Specifically, we simulate the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where each column is independent of the others and contains standard normal entries. The outcome vector is equal to $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a sparse vector with only the first $m = 10$ elements different from 0 (generated independently from $N(0, 4)$) while $\boldsymbol{\varepsilon} \stackrel{d}{\sim} N(0, I_n)$. A test point $(\mathbf{x}_{n+1}, y_{n+1})$ is generated with the same data-generating mechanism. At each iteration we estimate the regression function using the Lasso algorithm with penalty parameter λ varying over a fixed sequence of values $K = 20$ and then construct a conformal prediction interval for each λ in $\mathbf{x}_{n+1}$ using the split conformal method presented in package R `ConformalInference`. We run 10,000 iterations at $\alpha = 0.05$. We then merge the K different sets using the method described in (13.8) with $w_k = 1/K$ for $k \in \mathcal{K}$.

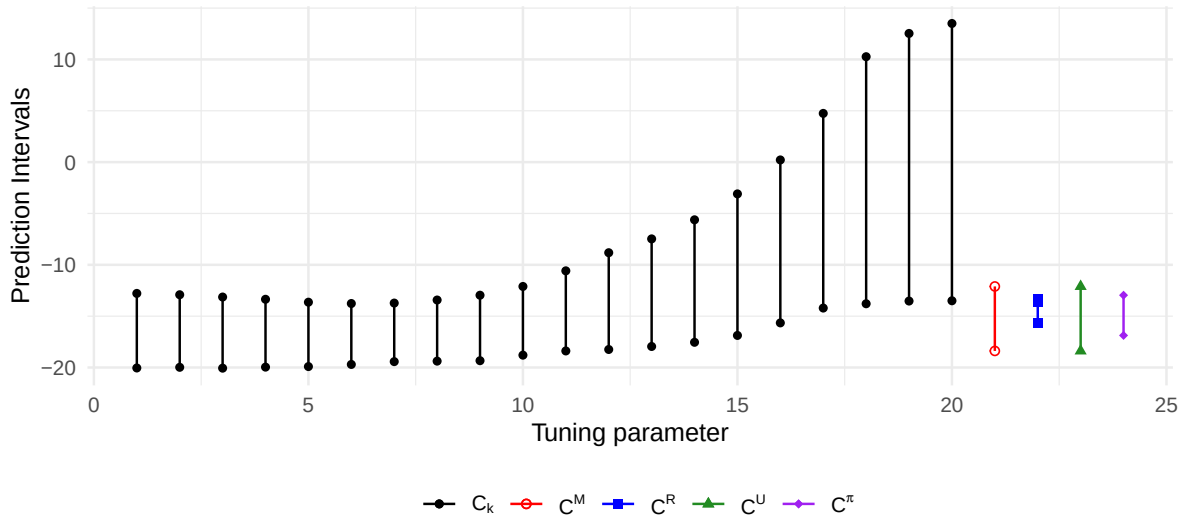


Figure 13.23: Intervals obtained using different values of λ , $C^M(x)$, $C^R(x)$, $C^U(x)$ and $C^\pi(x)$. For standardization, the value of U in the randomized thresholds is set to $1/2$. The smallest set C^R . Since $U = 1/2$, the sets C^M and C^U coincides.

An example of the result is shown in Figure 13.23. The empirical coverages of the intervals C^M and C^R

are

$$\frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\{y_{n+1}^b \in C_b^M(\mathbf{x}_{n+1})\}} = 0.97 \quad \text{and} \quad \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\{y_{n+1}^b \in C_b^R(\mathbf{x}_{n+1})\}} = 0.92.$$

By definition, the second method produces narrower intervals while maintaining the coverage level $1 - 2\alpha$. As explained in the previous sections, by inducing exchangeability through permutation, it may be possible to enhance the majority vote results. In fact, the empirical coverage of the sets C^π is 0.93 while the sets are smaller than the ones produced by the simple majority vote. Furthermore, we tested the sets C^U defined in (13.7) and obtained an empirical coverage equal to 0.96, which is very close to the nominal level $1 - \alpha$. In all five cases, the occurrence of obtaining an union of intervals as output is very low, specifically less than 1% of the iterations.

We now present the results on a real dataset on Parkinson’s disease [Tsanas and Little, 2009]. The goal is to predict the total UPDRS (Unified Parkinson’s Disease Rating Scale) score using a range of biomedical voice measurements from people suffering early-stage Parkinson’s disease. We used split conformal prediction and $K = 4$ different algorithms (linear model, lasso, random forest, neural net) to obtain the conformal prediction sets. In particular, we choose $n = 5000$ random observations to construct our intervals and the others $n_0 = 875$ observations as test points. Prior weights also in this case are uniform over the K models that represent the different agents, so a priori all methods are of the same importance. If previous studies had been carried out, one could, for example, put more weight on methods with better performance. Otherwise, one can assign a higher weight to more flexible algorithms such as random forest or neural net.

The results are reported in Table 13.23 where it is possible to note that all merging procedures obtain good results in terms of length and coverage. In addition, also the randomized vote obtains good results in terms of coverage, with an empirical length that is slightly larger than the one obtained by the neural net. The percentage of times that a union of intervals or an empty set is outputted is nearly zero for all methods. In this situation, the intervals produced by the random forest outperform the others in terms of size of the sets; as a consequence, one may wish to put more weight into the method, which results in smaller intervals on average.

Methods	LM	Lasso	RF	NN	C^M	C^R	C^U	C^π
Coverage	0.958	0.960	0.949	0.961	0.951	0.923	0.961	0.918
Width	40.143	40.150	13.286	32.533	29.508	20.620	32.544	20.710

Table 13.23: Coverage and length of the methods for the Parkinson’s dataset.

As previously mentioned, a potential method to derandomize the median of means estimator is to repeat the procedure K times and subsequently take the median of the estimators. The natural question that arises is: how large should K be in practice to achieve a derandomized result? Indeed, considering computational and time resources, one would prefer a small value for K . We conducted a simulation study to study the impact of K .

We simulated observations from a standard Student’s t-distribution with 3 degrees of freedom and from a skew-t distribution with 3 degrees of freedom and 1 as a skewness parameter. The number of batches is equal to $B = 21$ and the number of data points generated are $n = 210$. For each iteration, we computed $\hat{\mu}_1^{\text{MoMoM}}, \dots, \hat{\mu}_K^{\text{MoMoM}}$ with $K = 70$ and considered the absolute value of the difference $\hat{\mu}_k^{\text{MoMoM}} - \hat{\mu}_{k-1}^{\text{MoMoM}}$ as a measure of stability.

In the first two columns of Figure 13.24, we present two examples of the procedure, where it can be observed that in the first case, $\hat{\mu}^{\text{MoMoM}}$ tends to stabilize after 50 iterations, while in the second case, where the data are generated from a skew-t distribution, 25 iterations are sufficient. In the third column of

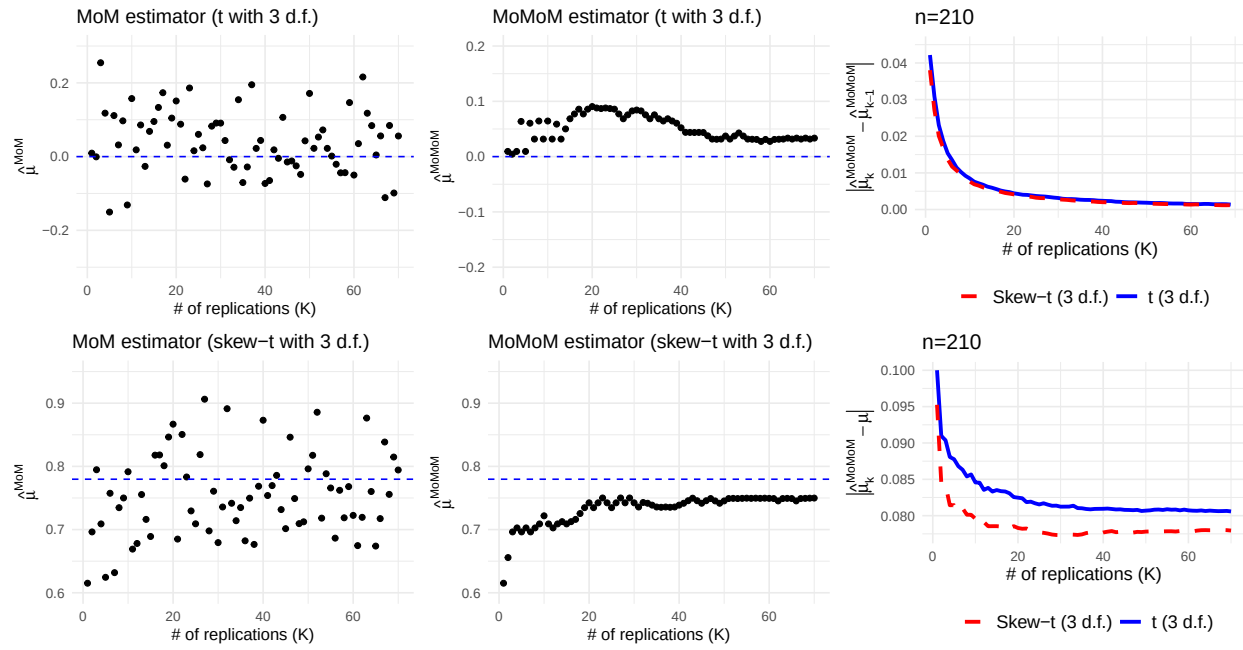


Figure 13.24: First column: MoM estimator obtained during various replications in a *single run of the procedure* using data generated from the t-distribution (above) and the skew-t distribution (below). The dashed line is the true mean μ . Second column: MoMoM estimator obtained during the replications; both series stabilize for $k > 40$. Third column: average over 1000 runs of the absolute difference $|\hat{\mu}_k^{\text{MoMoM}} - \hat{\mu}_{k-1}^{\text{MoMoM}}|$ against k (above) and average over 1000 runs of the absolute difference $|\hat{\mu}_k^{\text{MoMoM}} - \mu|$ against K . Note that the third column is only available to us in the simulation, but plots like the second column can be constructed on the fly as K increases, and it can be adaptively tracked and stopped, while retaining the same statistical guarantee at the stopped K .

Figure 13.24, we report the empirical average over 1000 replications of $|\hat{\mu}_k^{\text{MoMoM}} - \hat{\mu}_{k-1}^{\text{MoMoM}}|$. We observe an inflection point around 25 for both distributions, and it seems that once this threshold is reached, the gain becomes negligible.

Bibliographical note

E-confidence intervals were studied by Vovk and Wang [2023] and Xu et al. [2024]. The t-CI example is from Wang and Ramdas [2025]. Confidence sequences were introduced by Darling and Robbins [1967]. The original BY procedure for false coverage rate control was proposed by Benjamini and Yekutieli [2005], while the e-BY procedure was proposed by in Xu et al. [2024].

Kuncheva et al. [2003] studied majority voting for classifiers and derived its miscoverage rate. Solari and Djordjilović [2022] used majority vote to derandomize and stabilize split conformal prediction [Vovk et al., 2005]. Both their results can in turn be improved using the randomized and exchangeable extensions of majority vote, which were proposed by Gasparin and Ramdas [2024b]. The latter paper also presents several applications to derandomization of other statistical procedures based on sample splitting such as median of means [Lugosi and Mendelson, 2019] and HulC [Kuchibhotla et al., 2024]. The weighted variants of majority vote were exploited by Gasparin and Ramdas [2024a] in the development of an online conformal prediction algorithm that aggregates the outputs of many models in an online manner, upweighting the good models over time.

The median of means estimator stems back to, at least, a book by Nemirovskij and Yudin [1983], but some modern references include Devroye et al. [2016], Lugosi and Mendelson [2019]. Minsker [2023] proposed a derandomized variant that is computationally intensive: it involves taking the median of all possible sets of size n/B . The code to reproduce experiments on derandomizing median of means (and conformal prediction) is at

<https://github.com/matteogaspa/MergingUncertaintySets>.

Chapter 14

Further contrasts between e-values and p-values

In this chapter, we present a few results strengthening the theoretical foundation of e-values. This includes a duality between e-values and p-values, representations of e-variables and p-variables as conditional expectations and probabilities, and equivalent conditions for the existence of powered exact p-values and e-values.

14.1 A duality between e-values and p-values

We present a duality between e-values and p-values, further clarifying their intimate connection.

We consider a simple setting where a real-valued test statistic X is available, and a larger value of X represents evidence against the null hypothesis. Simple examples are (1.1) and (1.5) in Section 1.6, where the test statistic X is the sample sum. In this context, an e-variable should be an increasing function of X , and a p-variable should be a decreasing function of X . Denote by

$$S(x; \mathbb{P}) = \mathbb{P}(X \geq x)$$

for $x \in \mathbb{R}$, which is the left-continuous survival function of X under $\mathbb{P}$.

The next proposition explains that under the monotonicity restriction, $S(X; \mathbb{P})$ provides natural bounds on both p-variables and e-variables.

Proposition 14.1

Suppose that E is an e-variable for $\mathbb{P}$ that is an increasing function of X , and P is a p-variable for $\mathbb{P}$ that is a decreasing function of X . Then,

$$P \geq S(X; \mathbb{P}) \text{ and } E \leq 1/S(X; \mathbb{P}) \text{ } \mathbb{P}\text{-almost surely.}$$

Proof.

For the statement on the p-variable P , first note that for any decreasing function T and an iid copy X' of X , we have $\{T(X') \leq T(X)\} \supseteq \{X' \geq X\}$, which implies

$$\mathbb{P}(T(X') \leq T(X) \mid X) \geq \mathbb{P}(X' \geq X \mid X) = S(X; \mathbb{P}) \text{ } \mathbb{P}\text{-almost surely.}$$

Using this and the fact that the cdf F_P of P under $\mathbb{P}$ is smaller than or equal to the identity on $[0, 1]$ (this is the defining property of a p-variable), we get $P \geq F_P(P) \geq S(X; \mathbb{P})$.

The statement on the e-variable E follows directly by observing that $1/E$ is a p-variable,

guaranteed by Proposition 2.4, and then applying the result in the first part.

In fact, in Proposition 14.1, $S(X; \mathbb{P})$ is a p-variable for $\mathbb{P}$, but $1/S(X; \mathbb{P})$ is not an e-variable for $\mathbb{P}$. Nevertheless, in the next result we will see that, $1/S(X; \mathbb{P})$ is the supremum of all e-variables under the monotonicity condition.

Theorem 14.2: Duality between e-values and p-values

Suppose that probability measures in $\mathcal{P}$ are absolutely continuous with respect to a reference measure $\mathbb{L}$. Let $\mathcal{E}^X$ be the set of all e-variables for $\mathcal{P}$ that are increasing functions of X , and $\mathcal{U}^X$ be the set of all p-variables for $\mathcal{P}$ that are decreasing functions of X . Then,

$$\sup_{E \in \mathcal{E}^X} E = \inf_{\mathbb{P} \in \mathcal{P}} \frac{1}{S(X; \mathbb{P})} = \frac{1}{\inf_{P \in \mathcal{U}^X} P} \quad \mathbb{L}\text{-almost surely.}$$

In particular, if $\mathcal{P} = \{\mathbb{P}\}$, then $\sup_{E \in \mathcal{E}^X} E = 1/S(X; \mathbb{P})$, $\mathbb{P}$ -almost surely.

Proof.

Proposition 14.1 and the fact that $1/E$ for an e-variable E is a p-variable guarantee

$$\sup_{E \in \mathcal{E}^X} E \leq \sup_{P \in \mathcal{U}^X} \frac{1}{P} \leq \inf_{\mathbb{P} \in \mathcal{P}} \frac{1}{S(X; \mathbb{P})} \quad \mathbb{L}\text{-almost surely.} \quad (14.1)$$

Below we will show the equalities. For $s \in \mathbb{R}$, let

$$E_s = \frac{1}{\sup_{\mathbb{P} \in \mathcal{P}} S(s; \mathbb{P})} \mathbb{1}_{\{X \geq s\}} \quad \text{where we set } E_s = 1 \text{ in case of } 0/0,$$

which is increasing in X . Moreover, E_s is an e-variable, because $\mathbb{E}^{\mathbb{P}}[E_s] = S(s; \mathbb{P}) / \sup_{\mathbb{P}' \in \mathcal{P}} S(s; \mathbb{P}') \leq 1$ for all $\mathbb{P} \in \mathcal{P}$. It is straightforward to see that the supremum over $s \in \mathbb{R}$ for E_s is achieved at $s = X$, that is,

$$\sup_{s \in \mathbb{R}} E_s = \frac{1}{\sup_{\mathbb{P} \in \mathcal{P}} S(X; \mathbb{P})}$$

and hence the equalities in (14.1) hold.

The quantity $\sup_{\mathbb{P} \in \mathcal{P}} S(X; \mathbb{P})$ in Theorem 14.2 is a p-variable for $\mathcal{P}$, again the best one under the monotonicity restriction, as it is equal to $\inf_{P \in \mathcal{U}^X} P$. Theorem 14.2 suggests that the supremum of e-values over a set $\mathcal{E} \subseteq \mathcal{E}^X$ can be seen as a conceptual middle ground between an e-value ($\mathcal{E}$ is a singleton) and the reciprocal of a p-value ($\mathcal{E} = \mathcal{E}^X$ in Theorem 14.2, which is the maximal set).

The next corollary strengthens the Markov inequality for e-variables, which is immediate from Theorem 14.2.

Corollary 14.3

For any set $\mathcal{E}$ of e-variables for $\mathcal{P}$ that are increasing functions of X , we have

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P} \left(\sup_{E \in \mathcal{E}} E \geq \frac{1}{\alpha} \right) \leq \alpha \quad \text{for all } \alpha \in (0, 1).$$

If increasing monotonicity of E in X is replaced by decreasing monotonicity (and analogously for P), similar conclusions in this section hold true, with the left-continuous survival function of X replaced by its

cdf. Indeed, what matters for the type-I error control in Corollary 14.3 is that e-variables are *comonotonic*, meaning that they are increasing functions of the same random variable (in our context, it is X or $-X$, but it can be anything). Section 15.2 has a more general treatment of this property.

14.2 Representations as conditional expectations and probabilities

In this section, we present two representation results on p-variables and e-variables, which helps to further explain “e” in “e-values”.

In what follows, X represents a generic data sample, which takes values in $\mathbb{R}^n$; formally X is a measurable mapping from Ω to $\mathbb{R}^n$. In fact, the space $\mathbb{R}^n$ does not matter in this section. We say that a p-variable or an e-variable Y is built on X if it is a measurable function of X . In the language of measure theory, this means that Y is measurable with respect to the σ -algebra of X .

Theorem 14.4: Representation theorem for e-variables

For a random variable E , the following are equivalent:

- (i) E is an e-variable for $\mathcal{P}$ built on X ;
- (ii) For every $\mathbb{P} \in \mathcal{P}$, there exists $\mathbb{Q} \in \mathcal{M}_1$ with $\mathbb{Q} \ll \mathbb{P}$ such that

$$E \leq \mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid X \right] \quad \mathbb{P}\text{-almost surely.} \quad (14.2)$$

If all $\mathbb{P} \in \mathcal{P}$ are equivalent to some measure $\mathbb{L}$, then every e-variable built on X is dominated by (meaning $\leq$)

$$\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}_{\mathbb{P}}}{d\mathbb{P}} \mid X \right], \quad \text{which is an e-variable for } \mathcal{P},$$

for some class $(\mathbb{Q}_{\mathbb{P}})_{\mathbb{P} \in \mathcal{P}}$ in $\mathcal{M}_1$ absolutely continuous with respect to $\mathbb{L}$.

For given $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1$, the conditional expectation in (14.2) often takes the form

$$\mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid X \right] = \frac{p(X)}{q(X)},$$

where p and q are the density functions of the data X under $\mathbb{P}$ and $\mathbb{Q}$, when they exist, as in Chapter 5.

Proof.

The direction (ii) $\Rightarrow$ (i) is straightforward, because for all $\mathbb{P} \in \mathcal{P}$, the tower property of conditional expectation gives

$$\mathbb{E}^{\mathbb{P}}[E] \leq \mathbb{E}^{\mathbb{P}} \left[\mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid X \right] \right] = \mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \right] = 1.$$

Next we show (i) $\Rightarrow$ (ii). For each $\mathbb{P} \in \mathcal{P}$, let $\mathbb{Q}$ be a probability measure defined by $d\mathbb{Q}/d\mathbb{P} = E/\mathbb{E}^{\mathbb{P}}[E]$ if $\mathbb{E}^{\mathbb{P}}[E] > 0$, and $\mathbb{Q} = \mathbb{P}$ if $\mathbb{E}^{\mathbb{P}}[E] = 0$. We have, for $\mathbb{P}$ with $\mathbb{E}^{\mathbb{P}}[E] > 0$,

$$\mathbb{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid X \right] = \mathbb{E}^{\mathbb{P}} \left[\frac{E}{\mathbb{E}^{\mathbb{P}}[E]} \mid X \right] = \frac{E}{\mathbb{E}^{\mathbb{P}}[E]} \geq E \quad \mathbb{P}\text{-almost surely.}$$

For $\mathbb{P}$ with $\mathbb{E}^{\mathbb{P}}[E] = 0$, we have $E = 0$ $\mathbb{P}$ -almost surely, which trivially satisfies (14.2).

The last claim follows by taking an infimum over $\mathbb{P}$ in (14.2).

Theorem 14.4 generalizes Proposition 2.12, which connects e-variables to likelihood ratios in a simple null setting. Theorem 14.4 implies that all e-variables built from a dataset are indeed the conditional expectation of some likelihood ratio. In this sense, e-values are not only defined by constraints on expectations, but *they are indeed conditional expectations on the data*. Moreover, Proposition 3.22 gives the optimality of this construction when testing a simple null $\mathbb{P}$ against $\mathbb{Q}$, where $\mathbb{Q}$ in (14.2) is the alternative hypothesis.

The next result shows that *p-variables are conditional probabilities on the data*.

Theorem 14.5: Representation theorem for p-variables

For a random variable P , the following are equivalent:

- (i) P is a p-variable for $\mathcal{P}$ built on X ;
- (ii) There exists a measurable function $T : \mathbb{R}^n \rightarrow \mathbb{R}$ such that for every $\mathbb{P} \in \mathcal{P}$,

$$P \geq \mathbb{P}(T(X') \leq T(X) \mid X) \quad \mathbb{P}\text{-almost surely,} \quad (14.3)$$

where X' is an iid copy of X under $\mathbb{P}$.

If all $\mathbb{P} \in \mathcal{P}$ are equivalent to some measure $\mathbb{L}$, then every p-variable built on X is dominated by (meaning $\geq$)

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(T(X'_{\mathbb{P}}) \leq T(X) \mid X) \quad \text{which is a p-variable for } \mathcal{P},$$

for some measurable function $T : \mathbb{R}^n \rightarrow \mathbb{R}$, where $X'_{\mathbb{P}}$ is an iid copy of X under $\mathbb{P}$.

Proof.

We first show (ii) $\Rightarrow$ (i). Fix any $\mathbb{P} \in \mathcal{P}$, and denote by F_T the cdf of $T(X)$ under $\mathbb{P}$. Note that $\mathbb{P}(T(X') \leq T(X) \mid X) = F_T(T(X))$. Hence, $\mathbb{P}(P \leq \alpha) \leq \mathbb{P}(F_T(T(X)) \leq \alpha) \leq \alpha$, a well-known fact in probability. This shows that P is an e-variable for $\mathcal{P}$.

Next, we show (i) $\Rightarrow$ (ii). Take $T(X) = P$, treating P as a function of X . Fix any $\mathbb{P} \in \mathcal{P}$, and denote by F_P the cdf of P under $\mathbb{P}$. Since P is a p-variable, F_P is by definition smaller than or equal to the identity on $[0, 1]$. It follows that $P \geq F_P(P) = \mathbb{P}(T(X') \leq T(X) \mid X)$.

The last claim follows by taking a supremum over $\mathbb{P}$ in (14.3).

14.3 Existence of powered exact e-values and p-values

By considering finite hypotheses, we can derive another existence result, similar to Section 3.4, in the case that we require the e-variable to be *exact* (meaning that its expectation equals one under every $\mathbb{P} \in \mathcal{P}$), which turns out to need the same geometric condition as requiring that the p-variable is exact (meaning it is uniformly distributed for every $\mathbb{P} \in \mathcal{P}$). For a given $\mathcal{P}$, a random variable X is *pivotal* if it has the same distribution under all $\mathbb{P} \in \mathcal{P}$.

Theorem 14.6

Consider testing a finite $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_L\}$ against a finite $\mathcal{Q} = \{\mathbb{Q}_1, \dots, \mathbb{Q}_M\}$. If (U) holds, the following are equivalent:

- (i) there exists an exact (hence pivotal) p-variable that is powered against $\mathcal{Q}$;
- (ii) there exists a pivotal, exact, bounded e-variable that has positive e-power against $\mathcal{Q}$;
- (iii) there exists an exact e-variable that is powered against $\mathcal{Q}$;
- (iv) it holds that $\text{Conv}(\mathbb{Q}_1, \dots, \mathbb{Q}_M) \cap \text{Span}(\mathbb{P}_1, \dots, \mathbb{P}_L) = \emptyset$.

Compared to the Theorem 3.9, the convex hull of the null has been replaced by the span, defined as

$$\text{Span}(\mathbb{P}_1, \dots, \mathbb{P}_L) = \left\{ \sum_{\ell=1}^L \lambda_{\ell} \mathbb{P}_{\ell} : \lambda_{\ell} \in \mathbb{R} \text{ for each } \ell \in [L] \right\},$$

which is a set of signed measures on the sample space (the convex hull would only allow each $\lambda_{\ell} \geq 0$ and $\sum_{\ell=1}^L \lambda_{\ell} = 1$).

Though we do not provide a full proof of the theorem, which requires modern tools from optimal transport, the appearance of the span can be intuitively justified. If an exact e-variable E is powered against $\mathcal{Q}$, then for every $\mathbb{Q} \in \text{Conv}(\mathcal{Q})$, $\mathbb{E}^{\mathbb{Q}}[E] > 1$. But if such a $\mathbb{Q}$ can be written as a linear combination $\sum_{\ell=1}^L \lambda_{\ell} \mathbb{P}_{\ell}$ of elements of $\mathcal{P}$ (which implies $\sum_{\ell=1}^L \lambda_{\ell} = 1$ because $\mathcal{Q}$ is a probability), we get a contradiction, because we know that $\mathbb{E}^{\mathbb{P}}[E] = 1$ for every $\mathbb{P} \in \mathcal{P}$. This proves the direction (iii) $\Rightarrow$ (iv) above.

Bibliographical note

The duality in Section 14.1 is based on some results in [Blier-Wong and Wang \[2024\]](#). The observations in Section 14.2 were made in [Wang \[2021\]](#). Section 14.3 summarizes results from [Zhang et al. \[2024\]](#). In particular, Theorem 14.6 was proved in [Zhang et al. \[2024\]](#) using tools from optimal transport theory, and with a weaker assumption than (U), called joint non-atomicity, which does not require randomization.

Chapter 15

Improved e-value thresholds under additional conditions

Fix an atomless probability measure $\mathbb{P}$ for which all e-variables in this chapter are defined. Similarly to the case of Chapters 8 and 11, it suffices to consider the null $\{\mathbb{P}\}$, and all results can be easily translated into the case of general composite nulls.

The Markov inequality in Proposition 2.1 gives $\mathbb{P}(E \geq 1/\alpha) \leq \alpha$ for any E in the set $\mathfrak{E}$ of all e-variables, and it cannot be improved without further assumptions. Nevertheless, the probability $\mathbb{P}(E > 1/\alpha)$ can be improved if the e-variable E is constrained in a subset $\mathcal{E} \subseteq \mathfrak{E}$ of e-variables, which we will study in this section. The set $\mathcal{E}$ will be chosen to satisfy some distributional conditions to be specified later.

The application context of results in this section is when the tester does not know the distribution of the e-value (e.g., due to its complicated or black-box design), but have some shape information about the e-value.

We consider a simple hypothesis in this section, but if a composite null hypothesis is considered, then it suffices to assume the corresponding shape information for all distributions in the null hypothesis.

15.1 Conditional e-to-p calibrators on a subset of e-values

We are interested in the quantity $R_\gamma(\mathcal{E})$ for $\gamma > 0$, defined by

$$R_\gamma(\mathcal{E}) = \sup_{E \in \mathcal{E}} \mathbb{P}(E \geq 1/\gamma),$$

that is, the largest probability that $\mathbb{P}(E \geq 1/\gamma)$ can attain for $E \in \mathcal{E}$. Hence, for any set $\mathcal{E}$ of e-variables, it holds that $R_\gamma(\mathcal{E}) \leq R_\gamma(\mathfrak{E}) = \gamma$ for $\gamma \in (0, 1]$.

We are interested only in the case $\gamma \in (0, 1]$. For $\gamma > 1$, it is usually the case that $R_\gamma(\mathcal{E}) = 1$ because for most classes $\mathcal{E}$ that we consider, either $1 \in \mathcal{E}$ or 1 is the limit of elements of $\mathcal{E}$.

To build a level- α test using the e-variable E , one needs to find a threshold $t > 0$, better to be smaller, such that $\mathbb{P}(E \geq t) \leq \alpha$. For this, we intuitively should use the smallest t such that $R_{1/t}(\mathcal{E}) \leq \alpha$, which satisfies $R_{1/t}(\mathcal{E}) = \alpha$ in case $\gamma \mapsto R_\gamma(\mathcal{E})$ is continuous. Because of the usual interpretation of an e-variable E that $E \leq 1$ carries no evidence against the null hypothesis, we consider $t \geq 1$.

The following result computes the smallest threshold t for the e-test. Denote by $q_\beta(X)$ the left $(1 - \beta)$ -quantile function of X (as in Sections 2.6.6 and 6.7.6), that is,

$$q_\beta(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq 1 - \beta\} \quad \text{for } \beta \in (0, 1).$$

Conditions on the quantile function can be translated into those on the cdf. A summary of these translations is provided in Lemma A.7 in Appendix A.2.

Lemma 15.1

For $\alpha \in (0, 1)$, the quantity $T_\alpha(\mathcal{E}) := \inf\{t \geq 1 : R_{1/t}(\mathcal{E}) \leq \alpha\}$ satisfies

$$T_\alpha(\mathcal{E}) = \left(\sup_{E \in \mathcal{E}} q_\alpha(E) \right) \vee 1.$$

If $\gamma \mapsto R_\gamma(\mathcal{E})$ is continuous, then $T_\alpha(\mathcal{E})$ is the smallest real number $t \geq 1$ such that $\mathbb{P}(E \geq t) \leq \alpha$ for all $E \in \mathcal{E}$.

Proof.

Since $R_\gamma(\mathcal{E}) \leq \gamma$ for $\gamma \in (0, 1]$, the set $\{t \geq 1 : R_{1/t}(\mathcal{E}) \leq \alpha\}$ is not empty. By using Lemma A.7, we have

$$\begin{aligned} T_\alpha(\mathcal{E}) &\geq \inf\{t \geq 1 : \mathbb{P}(E > t) \leq \alpha \text{ for all } E \in \mathcal{E}\} \\ &= \inf\{t \geq 1 : q_\alpha(E) \leq t \text{ for all } E \in \mathcal{E}\} = \left(\sup_{E \in \mathcal{E}} q_\alpha(E) \right) \vee 1. \end{aligned}$$

Take any $\varepsilon \in (0, 1)$. We have

$$\begin{aligned} T_\alpha(\mathcal{E}) - \varepsilon &= \inf\{t \geq 1 - \varepsilon : R_{1/(t+\varepsilon)}(\mathcal{E}) \leq \alpha\} \\ &= \inf\left\{t \geq 1 - \varepsilon : \sup_{E \in \mathcal{E}} \mathbb{P}(E \geq t + \varepsilon) \leq \alpha\right\} \\ &= \inf\{t \geq 1 - \varepsilon : \mathbb{P}(E \geq t + \varepsilon) \leq \alpha \text{ for all } E \in \mathcal{E}\} \\ &\leq \inf\{t \geq 1 - \varepsilon : \mathbb{P}(E > t) \leq \alpha \text{ for all } E \in \mathcal{E}\} \\ &= \left(\sup_{E \in \mathcal{E}} q_\alpha(E) \right) \vee (1 - \varepsilon) \leq \left(\sup_{E \in \mathcal{E}} q_\alpha(E) \right) \vee 1 \leq T_\alpha(\mathcal{E}). \end{aligned}$$

Since $\varepsilon \in (0, 1)$ is arbitrary, we have $T_\alpha(\mathcal{E}) = (\sup_{E \in \mathcal{E}} q_\alpha(E)) \vee 1$, showing the first statement. If $\gamma \mapsto R_\gamma(\mathcal{E})$ is continuous, then

$$\begin{aligned} T_\alpha(\mathcal{E}) &= \min\{t \geq 1 : R_{1/t}(\mathcal{E}) \leq \alpha\} \\ &= \min\{t \geq 1 : \mathbb{P}(E \geq t) \leq \alpha \text{ for all } E \in \mathcal{E}\}, \end{aligned}$$

showing the second statement.

In Section 2.3 we have seen e-to-p calibrators, which are mappings that convert any e-value into a p-value. The function $\gamma \mapsto R_{1/\gamma}(\mathcal{E})$ serves to refine this concept. For a subset $\mathcal{E}$ of e-variables, we say that a function $f : [0, \infty] \rightarrow [0, \infty)$ is a *conditional e-to-p calibrator on $\mathcal{E}$* if f is decreasing, and $f(E)$ is a p-variable for all $E \in \mathcal{E}$. Different from Proposition 2.4, which gives that $x \mapsto (1/x) \wedge 1$ is the only useful e-to-p calibrator on $\mathfrak{E}$, we can find better conditional e-to-p calibrators based on $R_{1/\gamma}(\mathcal{E})$ than $x \mapsto (1/x) \wedge 1$ for various subsets $\mathcal{E}$ of $\mathfrak{E}$. Moreover, the following result implies that any class $\mathcal{E}$ admits a smallest conditional e-to-p calibrator. This is in sharp contrast to the set of calibrators from p-values to e-values, which does not admit a smallest element.

Proposition 15.2

The function $x \mapsto R_{1/x}(\mathcal{E})$ on $[0, \infty]$ is a conditional e-to-p calibrator on $\mathcal{E}$, and it is the smallest such calibrator.

Proof.

We first show that the function $g : x \mapsto R_{1/x}(\mathcal{E})$ is a conditional e-to-p calibrator. This means $\mathbb{P}(g(E) \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ and $E \in \mathcal{E}$. By definition, for $x \in [0, \infty]$ and $E \in \mathcal{E}$,

$$g(x) \geq \mathbb{P}(E \geq x) = \mathbb{P}(-E \leq -x) = F_{-E}(-x),$$

where F_X is the cdf of a random variable X . Since $F_X(X)$ is a p-variable for any random variable X , we have $\mathbb{P}(F_{-E}(-E) \leq \alpha) \leq \alpha$ for all $E \in \mathcal{E}$. Therefore, for all $\alpha \in (0, 1)$,

$$\mathbb{P}(g(E) \leq \alpha) \leq \mathbb{P}(F_{-E}(-E) \leq \alpha) \leq \alpha.$$

We now prove that g is the smallest calibrator on $\mathcal{E}$. Suppose that f is another conditional e-to-p calibrator on $\mathcal{E}$ such that $f(x) < g(x)$ for at least one $x \in [1, \infty]$. By definition, there exists $E \in \mathcal{E}$ such that $f(x) < \mathbb{P}(E \geq x)$. This implies

$$\mathbb{P}(f(E) \leq f(x)) \geq \mathbb{P}(E \geq x) > f(x),$$

which violates the definition of $f(E)$ as a p-variable.

Proposition 15.2 implies that by computing $R_\gamma(\mathcal{E})$ or an upper bound on it, we can find conditional e-to-p calibrators to convert e-values realized by elements of $\mathcal{E}$ into p-values, which work better than the e-to-p calibrator $x \mapsto (1/x) \wedge 1$ on $\mathfrak{E}$. This can be useful in procedures that take p-values as input, such as the Benjamini–Hochberg procedure.

We make a further observation on the probability bound $R_\gamma(\mathcal{E})$ and the threshold $T_\alpha(\mathcal{E})$. For a class $(E_\theta)_{\theta \in \Theta}$ of e-variables in $\mathcal{E}$ and a probability measure π on Θ , let E_π follow the mixture distribution, that is, $\mathbb{P}(E_\pi \leq x) = \int_\Theta \mathbb{P}(E_\theta \leq x) \pi(d\theta)$ for $x \in \mathbb{R}$. It is straightforward to see $\mathbb{P}(E_\pi \geq 1/\gamma) \leq R_\gamma(\mathcal{E})$. In other words, the probability bounds and thresholds obtained for $\mathcal{E}$ also apply to the convex hull of $\mathcal{E}$ with respect to distribution mixtures.

15.2 Comonotonic e-variables

In this section, we discuss $R_\gamma(\mathcal{E})$ for a set $\mathcal{E}$ of comonotonic e-variables. Our results here can be seen as a generalized version of Proposition 14.1.

Consider the simple example in Section 1.7, that is, to test $N(0, 1)$ against $N(\mu, 1)$ with $\mu > 0$ for iid observations $X_1, \dots, X_n$. A natural e-variable E_μ is the likelihood ratio

$$E_\mu = \exp(\mu S_n - n\mu^2/2), \tag{15.1}$$

where $S_n = \sum_{i=1}^n X_i$. An interesting observation is that E_μ , $\mu > 0$ are *comonotonic* random variables.

Definition 15.3: Comonotonicity

A set $\mathcal{X}_C$ of random variables is *comonotonic* if there exists a common random variable Z such that each $X \in \mathcal{X}_C$ is an increasing function of Z . We also say that the random variables in $\mathcal{X}_C$ are comonotonic.

The following lemma allows us to analyze R_γ for a set of comonotonic e-variables.

Lemma 15.4

Let $x \in \mathbb{R}$ and $\gamma \in [0, 1]$. Suppose that $\mathcal{X}_C$ is a collection of comonotonic random variables satisfying $\mathbb{P}(X \geq x) \leq \gamma$ for each $X \in \mathcal{X}_C$. Then $\mathbb{P}(\sup_{X \in \mathcal{X}_C} X \geq x) \leq \gamma$.

Proof.

Suppose that all elements of $\mathcal{X}_C$ are increasing functions of Z . Note that the sets $\{X \geq x\}$ for $X \in \mathcal{X}$ are nested since $\mathcal{X}_C$ is a set of comonotonic random variables, i.e., each of the set is of the form $\{Z \geq z\}$ or $\{Z > z\}$ for different z and a common Z . Therefore, we can move the supremum outside the probability and get

$$\mathbb{P}\left(\sup_{X \in \mathcal{X}_C} X \geq x\right) = \sup_{X \in \mathcal{X}_C} \mathbb{P}(X \geq x) \leq \gamma,$$

as claimed.

The next result follows immediately.

Proposition 15.5

For any collection $\mathcal{E} \subseteq \mathfrak{E}$ of comonotonic e-variables for $\mathbb{P}$, we have $\mathbb{P}(\sup_{E \in \mathcal{E}} E \geq T_\alpha(\mathcal{E})) \leq \alpha$.

In particular, Proposition 15.5 implies $\mathbb{P}(\sup_{E \in \mathcal{E}} E \geq 1/\alpha) \leq \alpha$ for any comonotonic set $\mathcal{E}$, which we have seen in Corollary 14.3 with the help of a statistic X . As a consequence, $(\sup_{E \in \mathcal{E}} E)^{-1}$ is a p-variable.

In the setting where e-variables are computed from likelihood ratios, we discussed using the mixture of likelihood ratios in Section 3.7. Proposition 15.5 suggests that, under comonotonicity, we can use the supremum of e-variables instead of their mixture, and the supremum has a higher power.

Example 15.6: Testing normal mean

In the example of testing $N(0, 1)$ against $\{N(\mu, 1) : \mu > 0\}$ with n data points, instead of using E_μ in (15.1) for a fixed $\mu > 0$ in or its mixture, we can use

$$E = \sup_{\mu > 0} E_\mu = \sup_{\mu > 0} \exp(\mu S_n - n\mu^2/2) = \exp\left(\frac{(S_n)_+^2}{2n}\right),$$

which, although not being an e-variable, gives $\mathbb{P}(E \geq 1/\alpha) \leq \alpha$ under the null. If we are testing $\{N(\mu, 1) : \mu \leq 0\}$ against $\{N(\mu, 1) : \mu > 0\}$, the same type-I error guarantee holds true, following from Corollary 14.3.

The above observation on the normal distributions can be generalized to other families. Suppose that we test $\mathbb{P} = \mathbb{Q}_{\theta_0}$ against $\{\mathbb{Q}_\theta : \theta \in \Theta\}$ with a class of likelihood ratio e-variables

$$E_\theta = \prod_{i=1}^n \frac{q_\theta(X_i)}{p(X_i)}.$$

As long the e-variables are increasing or decreasing functions of the same test statistic, we can take the supremum over these e-variables, and particular this yields $E = E_{\hat{\theta}}$, where $\hat{\theta}$ is the maximum likelihood estimator. Although E is not an e-variable, the test using $\mathbf{1}_{\{E \geq 1/\alpha\}}$ has level α . Note that here the maximum likelihood estimator $\hat{\theta}$ appears in the numerator, instead of the denominator as in the universal inference in Chapter 5 (which yields e-variables).

In addition to the normal distributions, testing several other members of the exponential family yields a class of likelihood ratio functions that are monotonic in a common test statistic, and we omit a detailed discussion here.

Other than likelihood ratios, several examples in Section 2.6 also involve parametric classes $(E_\theta)_{\theta \in \Theta}$ of comonotonic e-variables, and $\mathbb{P}(\sup_{\theta \in \Theta} E_\theta \geq 1/\alpha) \leq \alpha$ holds for these examples.

15.3 Density conditions

We next study some common conditions on the density function of the e-variables in $\mathcal{E}$ to compute $R_\gamma(\mathcal{E})$ or an upper bound on it.

Three conditions, motivated by different applications, are modeled by the following sets of e-variables:

$$\begin{aligned}\mathcal{E}_D &= \{E \in \mathfrak{E} : E \text{ has a decreasing density on its support}\}, \\ \mathcal{E}_{D>1} &= \{E \in \mathfrak{E} : E \text{ has a decreasing density over } [1, \infty)\}, \\ \mathcal{E}_U &= \{E \in \mathfrak{E} : E \text{ has a unimodal density on } [0, \infty)\}.\end{aligned}$$

A random variable on $[0, \infty)$ has a unimodal density, or a unimodal distribution, if it has a density function that is increasing on $[0, x]$ and decreasing on $[x, \infty)$ for some constant $x \in [0, \infty)$, and it can possibly have a point mass at x . Such x is called a mode of the random variable.

The assumption of decreasing densities in $\mathcal{E}_D$ is satisfied by, for instance, exponential and Pareto distributions. Note that here assumptions are made on the densities of the e-variables themselves, not on those in the hypotheses of the underlying testing problem.

Whenever we consider a density, we include its limit case; that is, we allow $E \in \mathcal{E}_D$ to have a point mass at the left end-point of its support. This convention does not change any results but simplifies many arguments in the proofs. The assumption of decreasing density over $[1, \infty)$ but arbitrary elsewhere in $\mathcal{E}_{D>1}$ is useful and can be observed from some e-variables obtained from universal inference discussed in Chapter 5. The assumption of unimodal density in $\mathcal{E}_U$ is satisfied by, for instance, log-normal and gamma distributions. Moreover, as explained in Section 15.1, our obtained bounds also apply to the mixtures of the distributions within each class. The set of all modes of $E \in \mathcal{E}_U$ may be a singleton or an interval. Note that $\mathcal{E}_D \subseteq \mathcal{E}_U$, and hence $R_\gamma(\mathcal{E}_U) \geq R_\gamma(\mathcal{E}_D)$ for all $\gamma \in (0, 1]$.

Theorem 15.7: Decreasing and unimodal densities

For $\gamma \in (0, 1]$,

- (i) $R_\gamma(\mathcal{E}_D) = \gamma/2$ if $\gamma \neq 1$ and $R_1(\mathcal{E}_D) = 1$;
- (ii) $R_\gamma(\mathcal{E}_{D>1}) = \gamma/(1 + \sqrt{1 - \gamma^2})$;
- (iii) $R_\gamma(\mathcal{E}_U) = (\gamma/2) \vee (2\gamma - 1)$.

Proof.

We only present the proof of part (i) here, which is the simplest, but it illustrates the basic techniques of proving more complicated statements. The proof of parts (ii) and (iii) are omitted. We first show $\mathbb{P}(E \geq 1/\gamma) \leq \gamma/2$ for each $E \in \mathcal{E}_D$ and $\gamma \in (0, 1)$. Let f be the density function of E and let a be the left end-point of the support of E . For a fixed $z > 1$, set $b := z + \mathbb{P}(E \geq z)/f(z)$. We construct an e-variable E_0 with density g_0 such that (a) $g_0(x) = f(z)$ for $a < x < b$; (b) E_0 has a point mass at a with probability $1 - \int_a^\infty g_0(x)dx$.

We illustrate in Figure 15.8 how to construct g_0 from f . Since f has a decreasing density over $[a, \infty)$, we will have $f(x) \geq g_0(x)$ for $x \in [a, z)$ and $f(x) \leq g_0(x)$ for $x > z$. We can construct g_0 by shifting the area between $f(x)$ and $f(z)$, for $x \in (a, z)$ to a point mass at a and shifting the area between $f(x)$ and 0, for $x > b$ to the area between $f(x)$ and $f(z)$, for $x \in (z, b)$. Note that b is chosen so that $\mathbb{P}(E_0 \in [z, b)) = \mathbb{P}(E \geq z)$, so $\mathbb{P}(E_0 \geq z) = \mathbb{P}(E \geq z)$. Further, since we construct g_0 by shifting the density of f to the left, we have $\mathbb{E}^\mathbb{P}[E_0] \leq \mathbb{E}^\mathbb{P}[E]$.

Therefore, for any $E \in \mathcal{E}_D$ with left end-point of support a , there exists an e-variable $E' \in \mathcal{E}_D$ such that E' has a point mass at a and a uniform density on $[a, b]$ for some constant b , and $\mathbb{P}(E' \geq z) = \mathbb{P}(E \geq z)$. To show $\mathbb{P}(E \geq z) \leq 1/(2z)$, it suffices to look at the collection of

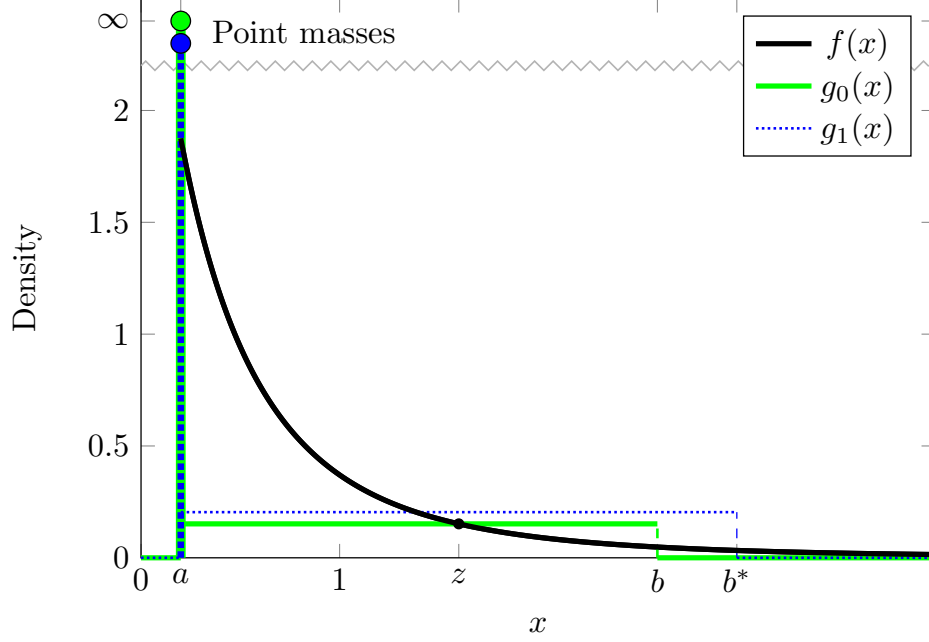


Figure 15.8: Proof sketch in part (i) of Theorem 15.7. Here, f and g_0 have the same area (probability) exceeding z , with g_0 having a smaller mean and a point mass at 0; g_1 has mean 1 and larger probability of exceeding z than g_0 .

e-variables that have a point mass at a and a uniform density on $[a, b]$ for any $b > 1$. Moreover, it suffices to consider the case $\mathbb{E}^{\mathbb{P}}[E'] = 1$.

Note that $\mathbb{E}^{\mathbb{P}}[E'] = (1 - w)a + w(a + b)/2$ for some $w \in [0, 1]$ and $b \geq 1$. The requirement $\mathbb{E}^{\mathbb{P}}[E'] = 1$ implies that $w = 2(1 - a)/(b - a)$. Now, we seek b that maximizes $\mathbb{P}(E' \geq z)$. We have $\mathbb{P}(E' \geq z) = w(b - z)/(b - a) = 2(1 - a)(b - z)/(b - a)^2$. Taking the derivative with respect to b and setting the result equal to zero, we find that $b^* := 2z - a$ for $z > 1$. Substituting this value of b^* back into $\mathbb{P}(E' \geq z)$, we find that $\mathbb{P}(E' \geq z) = (1 - a)/(2(z - a))$. We illustrate the density of E' that maximizes $\mathbb{P}(E' \geq z)$, denoted by g_1 , in Figure 15.8. Since $z > a$, the latter probability is a decreasing function of a , hence its supremum occurs for $a = 0$ and we have $\mathbb{P}(E \geq z) \leq 1/(2z)$.

It remains to show $\mathbb{P}(E \geq 1/\gamma) = \gamma/2$ for some $E \in \mathcal{E}_{\mathcal{D}}$. Take E following a mixture distribution of a uniform distribution on $[0, 2/\gamma]$ with weight γ and a point mass at 0. It is easy to see that $\mathbb{P}(E \geq 1/\gamma) = \gamma/2$ and $\mathbb{E}^{\mathbb{P}}[E] = 1$, showing the desired result. The last statement of part (i) is trivial by checking with $E = 1$.

The main message from Theorem 15.7 is that when an e-variable has a decreasing density, its threshold can be boosted by a factor of 2. That is, $T_{\alpha}(\mathcal{E}_{\mathcal{D}}) = 1/(2\alpha)$. Further, $T_{\alpha}(\mathcal{E}_{\mathcal{D}>1}) = 1/(2\alpha) + \alpha/2 = T_{\alpha}(\mathcal{E}_{\mathcal{D}}) + \alpha/2$. Therefore, for small values of α , we can boost the threshold by a factor slightly less than 2 when $E \in \mathcal{E}_{\mathcal{D}>1}$. This validates our message that, for small α , the distribution of uninformative e-values (between 0 and 1) has little impact on the rejection region; the shape of the distribution of large e-values has the most influence on the rejection threshold.

In part (iii) of Theorem 15.7, $R_{\gamma}(\mathcal{E}_{\mathcal{U}}) = \gamma/2$, for $\gamma \in (0, 2/3]$ and $R_{\gamma}(\mathcal{E}_{\mathcal{U}}) = 2\gamma - 1$ for $\gamma \in (2/3, 1]$. Since $\gamma \leq 2/3$ is the most practical situation (meaning a type-I error control of $1/3$), the main message from part (iii) is similar to that from part (i): when an e-variable has a unimodal density, its threshold can be boosted by a factor of 2 in practice. We provide in Figure 15.9 a comparison of different worst-case type-I errors and

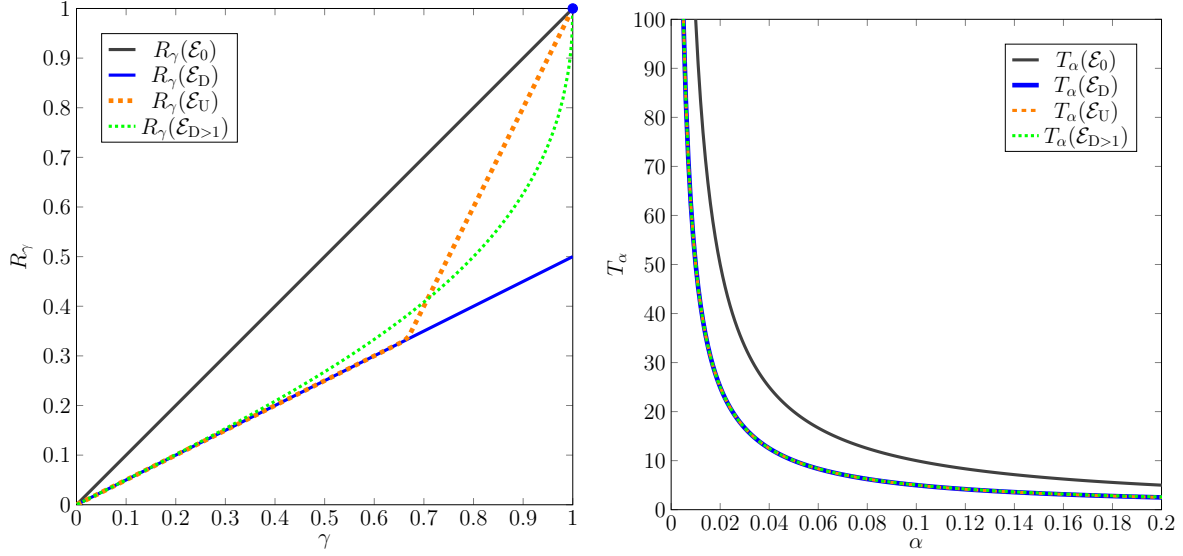


Figure 15.9: Comparison of worst-case type-I errors and improved thresholds for decreasing and unimodal densities. In the left panel, the diagonal line is $R_\gamma(\mathfrak{E})$, corresponding to Markov's inequality, and all other curves have a slope of approximately 1/2 for small γ .

improved thresholds for $\mathcal{E}_0, \mathcal{E}_D, \mathcal{E}_{D>1}$ and $\mathcal{E}_U$, where we use $\mathcal{E}_0$ for the full set $\mathfrak{E}$.

15.4 Conditions on log-transformed e-variables

In many applications of e-values, the final e-variable E is the product of many e-variables, especially in the context of e-processes in Chapter 7. In such a setting, assuming $\log E$ has some simple distributional properties may be reasonable, due to effects of central limit theorems. We say that a random variable X has a symmetric distribution if there exists $c \in \mathbb{R}$ such that X and $c - X$ are identically distributed. We consider six different sets in this section:

$$\begin{aligned} \mathcal{E}_{LS} &= \{E \in \mathfrak{E} : \log E \text{ has a symmetric distribution}\}, \\ \mathcal{E}_{LU} &= \{E \in \mathfrak{E} : \log E \text{ has a unimodal density}\}, \\ \mathcal{E}_{LD>0} &= \{E \in \mathfrak{E} : \log E \text{ has a decreasing density over } [0, \infty)\}, \\ \mathcal{E}_{LD} &= \{E \in \mathfrak{E} : \log E \text{ has a decreasing density}\}, \\ \mathcal{E}_{LUS} &= \{E \in \mathfrak{E} : \log E \text{ has a unimodal and symmetric distribution}\}, \\ \mathcal{E}_{LN} &= \{E \in \mathfrak{E} : E \text{ has a log-normal distribution}\}. \end{aligned}$$

In all sets above, we require $\mathbb{P}(E = 0) = 0$, so that $\log E$ is a real-valued random variable. Note that

$$\mathcal{E}_{LN} \subseteq \mathcal{E}_{LUS} \subseteq \mathcal{E}_{LS} \quad \text{and} \quad \mathcal{E}_{LD} \subseteq \mathcal{E}_{LD>0}.$$

The point mass distributions $x \in (0, 1]$ are included in all sets above, which are degenerate cases of log-normal distributions. Note that $E \in \mathfrak{E}$ has a log-normal distribution if and only if $\log E \stackrel{d}{\sim} N(\mu, \sigma^2)$ with $\mu + \sigma^2/2 \leq 0$.

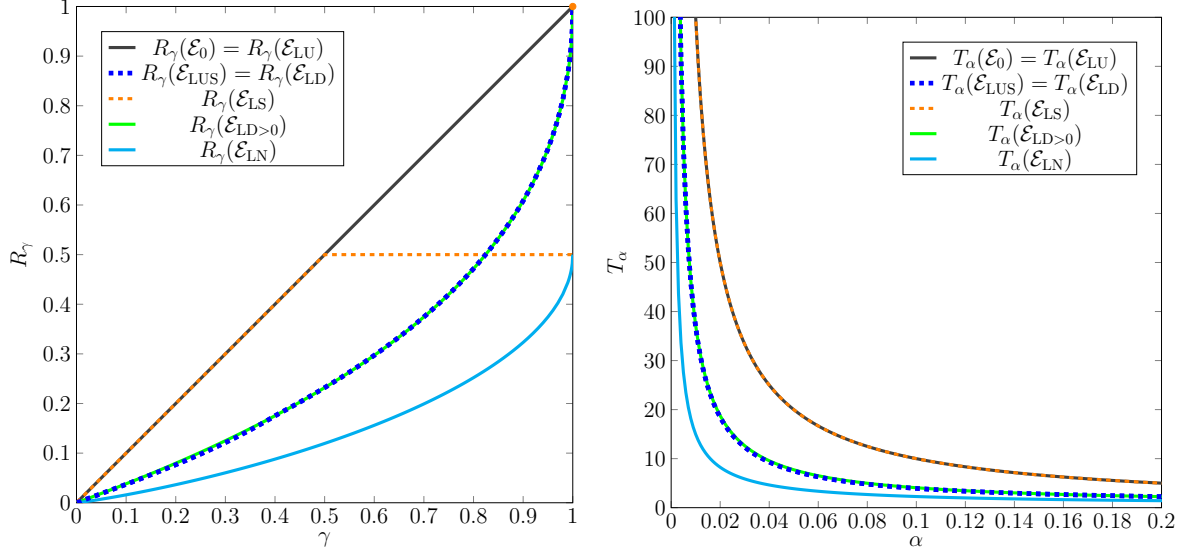


Figure 15.11: Comparison of worst-case type-I errors and improved thresholds for log-transformed e-variables. In the left panel, the diagonal line is $R_\gamma(\mathfrak{E})$, corresponding to the Markov inequality, and the curves for LUS, LD and LD>0 have a slope of approximately e^{-1} for small γ . The curve for LN is the smallest. Results for LUS are conservative bounds since we only find an upper bound for $R_\gamma(\mathcal{E}_{LUS})$ and $T_\alpha(\mathcal{E}_{LUS})$.

Theorem 15.10

For $\gamma \in (0, 1]$,

(i) $R_\gamma(\mathcal{E}_{LS}) = \gamma \wedge (1/2)$ if $\gamma \neq 1$ and $R_1(\mathcal{E}_{LS}) = 1$;

(ii) $R_\gamma(\mathcal{E}_{LU}) = \gamma$;

(iii) $\mathcal{E}_{LD>0} \subseteq \mathcal{E}_{D>1}$ and

$$\frac{\gamma}{e - \gamma} \leq R_\gamma(\mathcal{E}_{LD>0}) = e^{-a_\gamma} \leq \frac{\gamma}{1 + \sqrt{1 - \gamma^2}},$$

where a_γ is the unique solution of $e^a(1 - a - \log \gamma) = 1$ for $a \in (-\log \gamma, \infty)$;

(iv) $R_\gamma(\mathcal{E}_{LD}) = R_\gamma(\mathcal{E}_{LUS})$ and

$$\frac{\gamma}{e} \leq R_\gamma(\mathcal{E}_{LD}) = R_\gamma(\mathcal{E}_{LUS}) \leq \frac{\gamma}{e(1 - \gamma^2)} \wedge \left(\frac{\gamma}{1 + \sqrt{1 - \gamma^2}} \right);$$

(v) if $\gamma \neq 1$ then

$$R_\gamma(\mathcal{E}_{LN}) = \Phi\left(-\sqrt{-2 \log \gamma}\right) \leq \frac{\gamma}{2\sqrt{-\pi \log \gamma}},$$

where Φ is the standard normal cdf, and $R_1(\mathcal{E}_{LN}) = 1$.

We omit the proof of Theorem 15.10. To summarize the results, for log-transformed symmetric distributions and log-transformed unimodal distributions, the standard Markov's bound cannot be improved for $\gamma \leq 1/2$. An improvement of roughly the order of $1/e$ for small γ is possible for log-transformed decreasing densities on the positive real line, log-transformed decreasing densities and log-transformed unimodal-symmetric densities.

We plot the bounds on worst-case type-I errors and improved thresholds from Theorem 15.10 in Figure 15.11. Table 15.12 summarizes some numerical values for results in Theorems 15.7 and 15.10 (with $\mathcal{E}_0 = \mathfrak{E}$).

	α					
	0.001	0.01	0.02	0.05	0.1	0.2
$\mathcal{E}_0, \mathcal{E}_{\text{LS}}$	1000	100	50	20	10	5
$\mathcal{E}_{\text{D}}, \mathcal{E}_{\text{U}}$	500	50	25	10	5	2.5
$\mathcal{E}_{\text{D}>1}$	500	50.01	25.01	10.03	5.05	2.60
$\mathcal{E}_{\text{LUS}}, \mathcal{E}_{\text{LD}}$	367.88	36.82	18.45	7.49	3.93	2.25
$\mathcal{E}_{\text{LD}>0}$	368.25	37.16	18.77	7.73	4.07	2.25
$\mathcal{E}_{\text{LN}}$	118	14.97	8.24	3.87	2.27	1.42

Table 15.12: Improved thresholds $T_\alpha(\mathcal{E})$ for different distributional hypothesis.

Bibliographical note

The content of this chapter is largely based on [Blie-Wong and Wang \[2024\]](#), where proofs of all results can be found, as well as some improved thresholds for combined e-values and the e-BH procedure. Computing probability bounds with distributional information is a well-studied topic in operations research and risk management; some general references on related techniques are [Shaked and Shantikumar \[2007\]](#) and [Rüschendorf et al. \[2024\]](#). Comonotonicity in Section 15.2 is a central concept in dependence modeling. It is also central to decision theory under ambiguity via its connection to Choquet integrals [[Schmeidler, 1989](#)]. A good reference on comonotonicity is [Denneberg \[1994\]](#).

Chapter 16

E-values and risk measures

This chapter studies the connection between e-values and risk measures. By risk measures, we mean general mappings from random variables or distributions to real numbers, which include the usual statistical functions such as the mean, the variance, and the quantiles.

16.1 Risk measures and statistical functions

Risk measures are quantitative tools used to quantify the riskiness level associated with financial positions, such as asset returns and losses, both at the individual level and at the portfolio level, in a fixed time period. Recall that $\mathcal{X}$ is the set of all random variables on a given measurable space $(\Omega, \mathcal{F})$, and $\mathcal{M}_1(\mathbb{R})$ is the set of all distributions on $\mathbb{R}$. There are two common formulations of risk measures (just like the expected value, which can be seen as either a mapping from random variables or one from distributions, to the extended real line).

- (a) A risk measure can be formulated as a mapping from a subset of $\mathcal{X}$ to $(-\infty, \infty]$. We will use $\mathcal{R}$ for such mappings. In this setting, $\mathcal{R}(X)$ represents the evaluation of the riskiness of a random variable X , representing a financial loss or gain.
- (b) A risk measure can be formulated as a mapping from a subset of $\mathcal{M}_1(\mathbb{R})$ to $(-\infty, \infty]$. We will use ρ, ϕ for such mappings. In this setting, $\rho(F)$ represents the evaluation of the riskiness of the distribution F of some financial loss or gain.

Mappings in both (a) and (b) will be called risk measures. When a reference probability measure $\mathbb{P}_0$ is fixed, the two formulations above can be connected via $\mathcal{R}(X) = \rho(F)$ where $F \in \mathcal{M}_1(\mathbb{R})$ and $X \stackrel{d}{\sim} F$ (under $\mathbb{P}_0$). This correspondence is one-to-one if $\mathcal{R}$ satisfies *law invariance*: $\mathcal{R}(X) = \mathcal{R}(Y)$ if $X \stackrel{d}{=} Y$. In this case, a risk measure $\mathcal{R}$ or ρ is also called a *statistical function*.

Certainly, common statistical functions, such as the mean, the variance, or the median, can also be naturally defined for distributions on $\mathbb{R}^d$ for $d \in \mathbb{N}$, and in this chapter we focus on the case of distributions on $\mathbb{R}$. We keep our discussions on risk measures minimal and omit the generalizations such as set-valued, vector-valued, systemic, and dynamic risk measures, and many other variants, that are actively studied in the finance literature.

A main interpretation of a risk measure in finance is that its value $\mathcal{R}(X)$ for $X \in \mathcal{X}$ or $\rho(F)$ for $F \in \mathcal{M}_1$ represents the regulatory capital requirement of the random loss X or distribution F . When X is treated as a random asset price instead of loss, $\mathcal{R}(X)$ can be interpreted as the price to purchase X (the “ask price”) in an incomplete financial market. The price $\mathcal{R}$ would have been linear if the financial market were complete, and the nonlinearity is due to market frictions such as hedging limitations, bid-ask spreads, and transaction fees.

With the capital requirement interpretation, the most commonly used risk measures in the banking and insurance industries are the *Value-at-Risk* (VaR) and the *Expected Shortfall* (ES), defined as the following

two classes of risk measures. Since they are law invariant, we will use the same notation for their version in both (a) and (b). We identify all distributions in $\mathcal{M}_1(\mathbb{R})$ with their cdf in this chapter.

At level $\beta \in (0, 1)$, the VaR is defined as the left β -quantile:

$$\text{VaR}_\beta(F) = \inf\{x \in \mathbb{R} : F(x) \geq \beta\}, \quad F \in \mathcal{M}_1(\mathbb{R}),$$

and the ES is defined as

$$\text{ES}_\beta(F) = \frac{1}{1-\beta} \int_\beta^1 \text{VaR}_\gamma(F) d\gamma, \quad F \in \mathcal{M}_1(\mathbb{R}).$$

ES is also called Conditional VaR (CVaR) and Tail VaR (TVaR) in different parts of the literature. Here, $\text{ES}_\beta(F)$ may take the value ∞ if F does not have a finite mean. We also write $\text{VaR}_\beta(X)$ and $\text{ES}_\beta(X)$ for $X \in \mathcal{X}$ to convert between versions in (a) and (b). Note that $\text{VaR}_\beta(X) = q_{1-\beta}(X)$ for the quantile function defined in Chapters 2, 6 and 15. A well-known connection between ES and VaR is, for X with finite mean,

$$\text{VaR}_\beta(X) \in \arg \min_{z \in \mathbb{R}} \left\{ z + \frac{1}{1-\beta} \mathbb{E}^{\mathbb{P}^0}[(X-z)_+] \right\}, \quad (16.1)$$

$$\text{ES}_\beta(X) = \min_{z \in \mathbb{R}} \left\{ z + \frac{1}{1-\beta} \mathbb{E}^{\mathbb{P}^0}[(X-z)_+] \right\}. \quad (16.2)$$

As of 2025, ES at level $\beta = 0.975$ is the standard measure for market risk in the current global banking regulation provided by the Basel Committee on Banking Supervision. VaR at various levels is the standard measure in insurance regulation and operational risk, and it is closely connected to the probability of default criterion used to measure credit risk.

A popular class of risk measures is that of *coherent risk measures*. Let $\mathcal{X}_{\mathcal{R}} \subseteq \mathcal{X}$ be a convex cone in $\mathcal{X}$ containing all constant random variables (identified with constants in $\mathbb{R}$). A coherent risk measure $\mathcal{R}$ is a mapping from $\mathcal{X}_{\mathcal{R}}$ to $(-\infty, \infty]$ satisfying four economic axioms:

- (i) Monotonicity: $\mathcal{R}(X) \leq \mathcal{R}(Y)$ for all $X, Y \in \mathcal{X}_{\mathcal{R}}$ with $X \leq Y$;
- (ii) Cash invariance: $\mathcal{R}(X + c) = \mathcal{R}(X) + c$ for all $X \in \mathcal{X}_{\mathcal{R}}$ and $c \in \mathbb{R}$;
- (iii) Subadditivity: $\mathcal{R}(X + Y) \leq \mathcal{R}(X) + \mathcal{R}(Y)$ for all $X, Y \in \mathcal{X}_{\mathcal{R}}$;
- (iv) Positive homogeneity: $\mathcal{R}(\lambda X) = \lambda \mathcal{R}(X)$ for all $X \in \mathcal{X}_{\mathcal{R}}$ and $\lambda > 0$.

The four axioms have clear financial meanings. For instance, monotonicity means that a larger loss has bigger risk; subadditivity means that a portfolio is not riskier than the individual risks summed due to possible diversification effects. We omit the detailed interpretations here. Note also that subadditivity and positive homogeneity together imply convexity. For $\beta \in (0, 1)$, $\text{ES}_\beta : \mathcal{X} \rightarrow (-\infty, \infty]$ satisfies all of the above, and $\text{VaR}_\beta : \mathcal{X} \rightarrow \mathbb{R}$ satisfies all but subadditivity. Hence, ES_β is coherent, but VaR_β is not.

Finally, a risk measure $\mathcal{R} : \mathcal{X}_{\mathcal{R}} \rightarrow (-\infty, \infty]$ is said to be *continuous from above* ($\downarrow$ -continuous) if

$$\mathcal{R}(X_n) \rightarrow \mathcal{R}(X) \text{ for any bounded sequence } (X_n)_{n \in \mathbb{N}} \downarrow X,$$

where $(X_n)_{n \in \mathbb{N}} \downarrow X$ is understood point-wise. In the next section, we will consider the set $\mathcal{X}_+$ of all nonnegative random variables, and connect ($\downarrow$ -continuous) coherent risk measures on $\mathcal{X}_+$ to e-values. We summarize the domains of $\mathcal{R}$ that we will encounter in Table 16.1.

Remark 16.2. The convexity of ES_β on $\mathcal{X}$ yields a simple alternative proof of the result that twice an average p-variable is a p-variable, presented in Proposition 12.4. Let $P = (P_1 + \dots + P_K)/K$, where $P_1, \dots, P_K$ are (arbitrarily dependent) p-variables. Without loss of generality we can assume that each of $P_1, \dots, P_K$ is uniformly distributed on $[0, 1]$. It suffices to show $\text{VaR}_\alpha(2P) \geq \alpha$ for all $\alpha \in (0, 1)$ (see Lemma A.7 for conversion between probability and quantile conditions), which is equivalent to $\text{VaR}_{1-\alpha}(-2P) \leq -\alpha$ for all $\alpha \in (0, 1)$. This follows from

$$\text{VaR}_{1-\alpha}(-2P) \leq \text{ES}_{1-\alpha}(-2P) \leq \frac{2}{K} \sum_{k=1}^K \text{ES}_{1-\alpha}(-P_k) = -\alpha,$$

where we used the convexity of $\text{ES}_{1-\alpha}$.

$\mathcal{X}$	the set of all random variables
$\mathcal{X}_B$	the set of bounded random variables
$\mathcal{X}_{B+}$	the set of bounded nonnegative random variables
$\mathcal{X}_+$	the set of nonnegative random variables
$\mathcal{X}_{\mathcal{R}}$	a generic domain of $\mathcal{R}$ that is a convex cone containing $\mathbb{R}$

Table 16.1: Domains of risk measures $\mathcal{R}$.

16.2 Connecting e-values to coherent risk measures

Let us first notice that, for any null hypothesis $\mathcal{P}$, the requirement of a random variable $E \geq 0$ to be an e-variable is simply a risk measure constraint $\mathcal{R}(E) \leq 1$. This risk measure $\mathcal{R}$ is defined as

$$\mathcal{R}(X) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[X] \quad (16.3)$$

for X in a set such that the above expectations are well-defined (such as the set of nonnegative random variables).

The mapping $\mathcal{R}$ in (16.3), sometimes called a super-expectation, has a long history as an important objects in different fields (and with different names), such as robust statistics, decision theory, optimization, finance, imprecise probability, and game-theoretic probability. As we see next, $\mathcal{R}$ is a coherent risk measure, and it has essentially the only form of coherent risk measures.

Theorem 16.3

Let $\mathcal{X}_{B+}$ be the set of all bounded nonnegative random variables. A mapping $\mathcal{R} : \mathcal{X}_{B+} \rightarrow \mathbb{R}$ is a $\downarrow$ -continuous coherent risk measure if and only if

$$\mathcal{R}(X) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[X], \quad X \in \mathcal{X}_{B+} \quad (16.4)$$

for some set $\mathcal{P} \subseteq \mathcal{M}_1$.

Proof.

It is straightforward to verify the “if” statement. For the “only if” statement, let $\mathcal{X}_B$ be the set of all bounded random variables. By Theorem 4.22 of Föllmer and Schied [2016], any $\mathcal{R}' : \mathcal{X}_B \rightarrow \mathbb{R}$ is a $\downarrow$ -continuous coherent risk measure if and only if it has a representation in (16.4) on $\mathcal{X}_B$. For $\mathcal{R}$ on $\mathcal{X}_{B+}$, it suffices to extend it to $\mathcal{X}_B$ by letting $\mathcal{R}(X) = \mathcal{R}(X - t) + t$ for $X \in \mathcal{X}_B \setminus \mathcal{X}_{B+}$ where t is the infimum value of X . One can easily check that with this extension, $\mathcal{R}$ is a coherent risk measure on $\mathcal{X}_B$, and hence the aforementioned result guarantees (16.4) on $\mathcal{X}_{B+}$.

The set $\mathcal{X}_{B+}$ of bounded and nonnegative random variables in Theorem 16.3 can be replaced by larger sets, but usually characterization results such as Theorem 16.3 are formulated with small domains for the greatest strength. A useful remark is that if $\mathcal{R} : \mathcal{X}_B \rightarrow \mathbb{R}$ is law invariant, then the representation (16.4) holds without requiring $\downarrow$ -continuity. Moreover, on any convex cone of random variables, (16.4) always defines a coherent risk measure, possibly taking the value ∞ . This in particular holds on $\mathcal{X}_+$, leading to the next observation.

Corollary 16.4

For any $\mathcal{P}$, there exists a coherent risk measure $\mathcal{R} : \mathcal{X}_+ \rightarrow [0, \infty]$ such that all e-variables for $\mathcal{P}$ are precisely those $E \geq 0$ that satisfy $\mathcal{R}(E) \leq 1$.

For any $\mathcal{P} \subseteq \mathcal{M}_1$, the log-optimal e-variable for $\mathcal{Q}$ can be reformulated as the solution to the following problem

$$\begin{aligned} & \text{to maximize} && \mathbb{E}^{\mathbb{Q}}[\log X] \\ & \text{subject to} && \mathcal{R}(X) \leq 1; X \geq 0, \end{aligned} \tag{16.5}$$

where $\mathcal{R}$ is a coherent risk measure with representation (16.3) on $\mathcal{X}_+$. The problem (16.5) has a clear financial interpretation: We aim to maximize our expected utility of the random wealth X under $\mathbb{Q}$ with a log utility function, and the constraint is that our wealth X cannot be negative, and its price $\mathcal{R}(X)$ is bounded by our initial budget 1. Recall that when X represents a random financial wealth, the risk measure value $\mathcal{R}(X)$ can be interpreted as the price of X in an incomplete financial market. This interpretation is consistent with testing by betting treated in Chapter 7.

We have seen one example of (16.5) in Section 6.7.6. The dual representation of ES_β in (16.4) is given by

$$\text{ES}_\beta(X) = \sup \left\{ \mathbb{E}^{\mathbb{P}}[X] : \mathbb{P} \in \mathcal{M}_1, \frac{d\mathbb{P}}{d\mathbb{P}_0} \leq \frac{1}{1-\beta} \right\}, \quad X \in \mathcal{X}. \tag{16.6}$$

In other words, the constraint of ES_β formulates all e-variables for the null hypothesis $\mathcal{P}$ that the true data-generating distribution is close to $\mathbb{P}_0$ in terms of likelihood ratio being bounded by $1/(1-\beta)$. Similarly, many other families of coherent risk measures correspond to other testing problems.

Finally, when the alternative hypothesis $\mathcal{Q}$ is composite, one has a few options to pick. The first is to consider the problem of maximizing the mixture e-power for some probability measure ν over $\mathcal{Q}$:

$$\begin{aligned} & \text{to maximize} && \int_{\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log X] d\nu \\ & \text{subject to} && \mathcal{R}(X) \leq 1; X \geq 0, \end{aligned} \tag{16.7}$$

where $\mathcal{R}$ is again the coherent risk measure in (16.3). The second is to maximize the worst-case e-power

$$\begin{aligned} & \text{to maximize} && \inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log X] \\ & \text{subject to} && \mathcal{R}(X) \leq 1; X \geq 0. \end{aligned} \tag{16.8}$$

The third is to maximize the relative e-power compared with an oracle e-variable,

$$\begin{aligned} & \text{to maximize} && \inf_{\mathbb{Q} \in \mathcal{Q}} (\mathbb{E}^{\mathbb{Q}}[\log X] - \mathbb{E}^{\mathbb{Q}}[\log E_{\mathbb{Q}}^*]) \\ & \text{subject to} && \mathcal{R}(X) \leq 1; X \geq 0, \end{aligned} \tag{16.9}$$

where $E_{\mathbb{Q}}^*$ is the numeraire for $\mathcal{P}$ against $\mathbb{Q}$ (from Chapter 6).

When there are multiple observations available for the testing problem, the solution to the third problem (16.9) is able to learn the true data-generating distribution with data as discussed in Section 3.7, whereas the solutions to the first two problems do not.

We end this section by noting that the problems (16.7), (16.8), and (16.9) all have concave objectives to maximize, subject to (possibly infinitely many) linear constraints, and therefore it is a convex program and can be solved efficiently in many cases, in particular when the state space, $\mathcal{Q}$ and $\mathcal{P}$ are all finite.

16.3 E-values for testing risk measures

In this section, we discuss how to design e-values to test the value of a given law-invariant risk measure. We will mainly use the formulation ρ that maps a subset of $\mathcal{M}_1(\mathbb{R})$ to $\mathbb{R}$.

16.3.1 E-statistics

The general goal is to test the value of $\rho(X_t)$ through a sequence $(X_t)_{t \in \mathbb{T}_+}$ of data points, which are not assumed to be iid. Here, the index set $\mathbb{T}_+$ may be finite or infinite. For instance, the null hypothesis may be

$$H_0 : X_t \text{ has distribution } F_t \text{ with } \rho(F_t) \leq r_t \text{ for } t \in \mathbb{T}_+. \quad (16.10)$$

Here, the values r_t can be interpreted as the risk forecast for X_t evaluated with ρ .

Our general testing procedure is to follow the idea of testing by betting described in Chapter 7, in the following two steps.

1. Compute an e-variable E_t from each data point X_t , such that $E_1, E_2, \dots$ are sequential e-variables.
2. Build an e-process via testing by betting, such as the empirically adaptive e-process in Section 7.8.

To follow the above procedure, we need to build an e-variable from each data point. Below, let X be a random variable representing one data point. We will first study how we can build an e-variable from this data point. Let $\mathcal{M}_* \subseteq \mathcal{M}_\rho$ represent the set of distributions of interest (it may be smaller than the domain of ρ). The simplest problems are to test

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) = r, \quad (16.11)$$

and to test

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) \leq r. \quad (16.12)$$

In finance, we are mainly interested in the one-sided hypothesis in (16.12) and its sequential version (16.10), as they have a clear interpretation: We are testing whether the value r is sufficiently conservative for the risk measure ρ of financial risks. In banking, r usually represents regulatory capital charge for a risky investment, and $r \geq \rho(F)$ means that the capital reserve is safe and passing the regulatory requirement. Two-sided tests for (16.11) can be easily constructed by combining one e-variable for testing $\rho(F) \leq r$ and one e-variable for testing $\rho(F) \geq r$ via taking an average.

We also consider the situation in which we have another statistical function ϕ , for which we have information such as $\phi(F) = z$, but we are not interested in testing it. We assume the domain of ϕ contains $\mathcal{M}_\rho$. An example of (ρ, ϕ) is (var, mean), where we are only interested in testing variance, but the information on the mean is available. The corresponding hypotheses are

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) = r \text{ and } \phi(F) = z, \quad (16.13)$$

and

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) \leq r \text{ and } \phi(F) = z. \quad (16.14)$$

We can easily allow ϕ to take values in $\mathbb{R}^d$ instead of $\mathbb{R}$, and we omit such a generalization.

To test the above hypotheses, we are interested in finding a function $e : (x, r, z) \mapsto \mathbb{R}$ such that $e(X, r, z)$ is an e-variable for the hypotheses in (16.11)–(16.14). Below, $\rho(\mathcal{M}_*)$ is the set of values taken by $\rho(F)$ for $F \in \mathcal{M}_*$.

Definition 16.5: E-statistics for ρ

Fix $\mathcal{M}_* \subseteq \mathcal{M}_\rho$ and $\rho : \mathcal{M}_\rho \rightarrow \mathbb{R}$, and let $e : \mathbb{R}^2 \rightarrow [0, \infty]$ be a measurable function.

- (i) The function e is a *point e-statistic* for ρ if $e(X, r)$ is an e-variable for H_0 in (16.11), and it is a *one-sided e-statistic* for ρ if $e(X, r)$ is an e-variable for H_0 in (16.12).
- (ii) A one-sided e-statistic e for ρ is a *backtest e-statistic* for ρ if $\int_{\mathbb{R}} e(x, r) dF(x) > 1$ for all $r \in \rho(\mathcal{M}_*)$ and $F \in \mathcal{M}_*$ with $\rho(F) > r$.
- (iii) A backtest e-statistic e is *monotone* if $r \mapsto e(x, r)$ is decreasing for each x .

Note that $\mathcal{M}_*$ is implicit in the above definition. Among the above concepts, it is clear that

$$\text{backtest e-statistic} \implies \text{one-sided e-statistic} \implies \text{point e-statistic}.$$

For example, if ψ is the mean and $\mathcal{M}_*$ is the set of distributions on $(0, \infty)$, then $e : (x, r) \mapsto x/r$ is a monotone backtest e-statistic, satisfying all requirements in Definition 16.5 (see Example 16.7 below). The property of having a mean larger than 1 under the alternative is crucial for the e-statistic. By Theorem 3.14, if e is a backtest e-statistic and $\rho(F) > r$, then there exists $\lambda \in (0, 1)$ such that $(1 - \lambda) + \lambda e(X, r)$ has positive e-power. In the context of an infinite sequence of observations, this condition is sufficient for establishing consistency of the empirically adaptive e-process by Theorem 7.22.

We analogously define e-statistics for (ρ, ϕ) , which has one more variable.

Definition 16.6: E-statistics for (ρ, ϕ)

Fix $\mathcal{M}_* \subseteq \mathcal{M}_\rho$ and $(\rho, \psi) : \mathcal{M}_\rho \rightarrow \mathbb{R}^2$, and let $e : \mathbb{R}^3 \rightarrow [0, \infty]$ be a measurable function.

- (i) The function e is a *point e-statistic* for (ρ, ψ) if $e(X, r)$ is an e-variable for H_0 in (16.13), and it is a *one-sided e-statistic* for (ρ, ψ) if $e(X, r)$ is an e-variable for H_0 in (16.14).
- (ii) A one-sided e-statistic e for (ρ, ϕ) is a *backtest e-statistic* for (ρ, ϕ) if $\int_{\mathbb{R}} e(x, r, z) dF(x) > 1$ for all $(r, z) \in (\rho, \phi)(\mathcal{M}_*)$ and $F \in \mathcal{M}_*$ with $\rho(F) > r$.
- (iii) A backtest e-statistic e is *monotone* if $r \mapsto e(x, r, z)$ is decreasing for each (x, z) .

The crucial interpretation of a backtest e-statistic for (ρ, ϕ) is that, if the risk measure ρ is underestimated, then a backtest e-statistic has power against the null hypothesis, regardless of whether the prediction of the auxiliary risk measure ϕ is truthful.

To explain the motivation behind monotone backtest e-statistics, we consider a financial context. In banking regulation, the risk measure ρ is used to compute regulatory capital, and ϕ has no direct financial consequence. Therefore, a firm may have incentive to forecast ϕ arbitrarily (correctly or incorrectly), but forecasting a large ρ would result in financial cost. This motivation explains the term “backtest”, which is a testing procedure in finance for risk models. If a backtest e-statistic is monotone, then an overestimation of the risk is rewarded: A firm being scrutinized by the regulator can deliberately report a higher risk value (which means a higher capital reserve) to pass to the regulatory test, thus rewarding prudence. We omit a detailed discussion on the financial interpretation here. In a non-financial context, a statistician may be only interested in point or one-sided e-statistics.

As mentioned above, after choosing a suitable e-statistic e , we can test H_0 in (16.10), as well as other similar hypotheses, by constructing the e-variables $E_t = e(X_t, r_t)$ for $t \in \mathbb{T}_+$, where r_t is the forecast for the risk of X_t conditional on $\mathcal{F}_{t-1}$, which is the typical situation in financial risk management (e.g., firms provide forecast for the next-day risk of their portfolio). These e-variables are sequential, because $\mathbb{E}^\mathbb{P}[e(X_t, r_t) | \mathcal{F}_{t-1}] \leq 1$ under any distribution $\mathbb{P}$ in the null hypothesis. Therefore, even if $X_1, X_2, \dots$ are not iid and possibly dependent, we can build an e-process based on $(E_t)_{t \in \mathbb{T}_+}$ as in Section 7.8 to test the null hypothesis, that is,

$$M_t = \prod_{s=1}^t ((1 - \lambda_s) + \lambda_s E_s) \text{ for } t \in \mathbb{T}_+ \text{ and } M_0 = 1, \quad (16.15)$$

The predictable process $(\lambda_t)_{t \in \mathbb{T}_+}$ in (16.15) can be chosen in various ways. For instance, it may be computed through the empirically adaptive e-process (Definition 7.21), or computed based on the empirical distribution of the sample $(X_t)_{t \in \mathbb{T}_+}$ without using the forecasts $(r_t)_{t \in \mathbb{T}_+}$. In what follows, we will focus on e-statistics for some commonly used statistical functions.

16.3.2 Mean, variance, quantiles, and expected losses

We derive backtest e-statistics for some common statistical functions. In this chapter, we use the convention that $0/0 = 1$ and $1/0 = \infty$. Let $\mathcal{M}^k \subseteq \mathcal{M}_1(\mathbb{R})$ be the set of all distributions with finite k -th-moment for

$k > 0$. All expectations below only concern the distribution of X , and we omit the probability $\mathbb{P}$ in $\mathbb{E}[\cdot]$.

Example 16.7: Backtest e-statistic for the mean

Let $\mathcal{M}_*$ be the set of distributions on $\mathbb{R}_+$ in $\mathcal{M}^1$. Define the function $e(x, r) = x/r$ for $x, r \geq 0$. In this case, we have $\mathbb{E}[e(X, r)] \leq 1$ for all random variables X with distribution in $\mathcal{M}_*$ and $r \geq \mathbb{E}[X]$. Moreover, for any such X , $\mathbb{E}[X] > r \geq 0$ implies $\mathbb{E}[e(X, r)] > 1$. Therefore, e is a monotone $\mathcal{M}_*$ -backtest e-statistic for the mean.

Example 16.8: Backtest e-statistic for (var, mean)

Consider $(\text{var}, \text{mean}) : \mathcal{M}^2 \rightarrow \mathbb{R}^2$, which is the pair of the variance and the mean. The function $e(x, r, z) = (x - z)^2/r$ for $x, z \in \mathbb{R}$ and $r \geq 0$ is a monotone backtest e-statistic for $(\text{var}, \text{mean})$. To see this, for all random variables X with distribution in $\mathcal{M}^2$, we have

$$\mathbb{E}[e(X, r, \mathbb{E}[X])] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{r} \leq 1$$

for $r \geq \text{var}(X)$. Moreover, since $z = \mathbb{E}[X]$ minimizes $\mathbb{E}[(X - z)^2]$ over $z \in \mathbb{R}$ and $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, $\text{var}(X) > r \geq 0$ implies

$$\mathbb{E}[e(X, r, z)] = \frac{\mathbb{E}[(X - z)^2]}{r} \geq \frac{\text{var}(X)}{r} > 1.$$

Example 16.9: Backtest e-statistic for a quantile

Take $\beta \in (0, 1)$. Define the function

$$e_\beta^q(x, r) = \frac{1}{1 - \beta} \mathbb{1}_{\{x > r\}}, \quad x, r \in \mathbb{R}.$$

We have $\mathbb{E}[e_\beta^q(X, r)] \leq 1$ for any random variable X with distribution F and $r \geq \text{VaR}_\beta(F)$. Moreover, $\text{VaR}_\beta(F) > r$ implies $\mathbb{P}(X > r) > 1 - \beta$, and hence $\mathbb{E}[e_\beta^q(X, r)] > 1$. Therefore, e_β^q is a monotone backtest e-statistic for the β -quantile.

Example 16.10: Backtest e-statistic for an expected loss

For some $a \in \mathbb{R}$, let $\ell : \mathbb{R} \rightarrow [a, \infty)$ be a function that is interpreted as a loss. Define the function $e(x, r) = (\ell(x) - a)/(r - a)$ for $x \in \mathbb{R}$ and $r \geq a$. Analogously to Example 16.7, e is a monotone backtest e-statistic for the expected loss $F \mapsto \int \ell dF$ on its natural domain.

For fixed ρ or (ρ, ϕ) , the choice of a backtest e-statistic e is not unique. For instance, a linear combination of e and 1 with weights straightly between 0 and 1 is also a backtest e-statistic for ρ or (ρ, ϕ) . Depending on the specific situation, either e-statistic may be useful in practice.

16.3.3 E-statistics for testing ES

The functional $(\rho, \phi) = (\text{var}, \text{mean})$ in Example 16.8 is an example of a *Bayes pair*; that is, there exists a measurable function $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that

$$\begin{aligned} \phi(F) &\in \arg \min_{z \in \mathbb{R}} \int L(z, x) dF(x) \text{ and} \\ \rho(F) &= \min_{z \in \mathbb{R}} \int L(z, x) dF(x) \quad \text{for } F \in \mathcal{M}_*, \end{aligned} \tag{16.16}$$

where $\int L(z, x) dF(x)$ is assumed to be well-defined for each $z \in \mathbb{R}$ and $F \in \mathcal{M}_*$. The function L is the square loss $(z - x)^2$ in the case of (var, mean). Bayes pairs often admit backtest e-statistics. A typical example commonly used in risk management practice is $(\text{ES}_\beta, \text{VaR}_\beta) : \mathcal{M}^1 \rightarrow \mathbb{R}^2$ treated below. Note that we restrict the domain to $\mathcal{M}^1$ so that ES_β takes real values.

For $\beta \in (0, 1)$, define the function

$$e_\beta^{\text{ES}}(x, r, z) = \frac{(x - z)_+}{(1 - \beta)(r - z)}, \quad x \in \mathbb{R}, \quad z \leq r.$$

This defines a backtest e-statistic for $(\text{ES}_\beta, \text{VaR}_\beta)$. Recall the convention that $0/0 = 1$ and $1/0 = \infty$, and set $e_\beta^{\text{ES}}(x, r, z) = \infty$ if $r < z$, which is a case of no relevance since $\text{ES}_\beta(F) \geq \text{VaR}_\beta(F)$ for any $F \in \mathcal{M}^1$.

Theorem 16.11

For $\beta \in (0, 1)$, the function e_β^{ES} is a monotone backtest e-statistic for $(\text{ES}_\beta, \text{VaR}_\beta)$.

Proof.

By (16.1) and (16.2), the pair $(\text{ES}_\beta, \text{VaR}_\beta)$ is a Bayes pair in (16.16) with $L : (z, x) \mapsto z + (x - z)_+ / (1 - \beta)$. Since $L(z, x) \geq z$, for $r \geq z$, we have that $e_\beta^{\text{ES}}(x, r, z) = (L(z, x) - z) / (r - z) \geq 0$, and it is decreasing in r . We have for $r \geq \text{ES}_\beta(X)$ that

$$\mathbb{E} \left[\frac{L(\text{VaR}_\beta(X), X) - \text{VaR}_\beta(X)}{r - \text{VaR}_\beta(X)} \right] = \frac{\text{ES}_\beta(X) - \text{VaR}_\beta(X)}{r - \text{VaR}_\beta(X)} \leq 1.$$

Furthermore, for $z < r \leq \text{ES}_\beta(X)$,

$$\mathbb{E} \left[\frac{L(z, X) - z}{r - z} \right] \geq \frac{\text{ES}_\beta(X) - z}{r - z} \geq 1$$

with equality if and only if $r = \text{ES}_\beta(X)$.

As a consequence of Theorem 16.11, $e_\beta^{\text{ES}}(X, r, z)$ is an e-variable for the null hypothesis

$$H_0 : \text{ES}_\beta(X) \leq r \text{ and } \text{VaR}_\beta(X) = z.$$

A generalization of this result is that $e_\beta^{\text{ES}}(X, r, z)$ is also an e-variable for the larger null hypothesis

$$H_0 : \text{ES}_\beta(X) - \text{VaR}_\beta(X) \leq r - z \text{ and } \text{VaR}_\beta(X) \leq z.$$

Nevertheless, we note that $e_\beta^{\text{ES}}(X, r, z)$ is not an e-variable for $H_0 : \text{ES}_\beta(X) \leq r$ and $\text{VaR}_\beta(X) \leq z$.

While Examples 16.7-16.10 and Theorem 16.11 show that interesting backtest e-statistics exist, much more can be said about their general structure, in particular, they are shown to be essentially the only choices, up to some transforms.

Bibliographical note

Coherent risk measures were introduced by Artzner et al. [1999], and the representation in a form similar to Theorem 16.3 was obtained by Delbaen [2002]. The representation under law invariance, without assuming continuity, was obtained by Jouini et al. [2006], and generalized by Delbaen [2012]. Coherent risk measures were used as pricing formulas by Wang [2000]. There are many more general classes of risk measures than the coherent risk measures; a good reference is Föllmer and Schied [2016], which also discusses pricing in incomplete financial markets. A comprehensive treatment of the use of risk measures in risk management and

financial regulation is [McNeil et al. \[2015\]](#). In the finance literature, many authors formulate risk measures on random gains instead of random losses, and their X is our $-X$ in that case (for instance, monotonicity would be formulated in an opposite direction). The formulas (16.1) and (16.2) were obtained by [Rockafellar and Uryasev \[2002, Theorem 10\]](#), where ES is called a CVaR, common in the optimization literature. For the dual representation of ES in (16.6), see e.g., [McNeil et al. \[2015, Theorem 8.14\]](#). Seven pedagogical proofs of the subadditivity of ES are provided by [Embrechts and Wang \[2015\]](#).

Super-expectations have a very long history. For instance, they were axiomatized by [Huber \[1981, Chapter 10\]](#) in the context of robust statistics. In addition to statistics and finance, super-expectations are also fundamentally important objects in the areas of decision theory under ambiguity ([Schmeidler \[1989\]](#), [Gilboa and Schmeidler \[1989\]](#)), imprecise probabilities ([Walley \[1991\]](#)), non-linear expectations ([Peng \[2019\]](#)), and game-theoretic probability ([Shafer and Vovk \[2001, 2019\]](#)).

The content of Section 16.3 is largely based on [Wang et al. \[2025\]](#), where they showed that most of the backtest e-statistics in Sections 16.3.2 and 16.3.3 are essentially the only possible choices in a suitable sense. They also studied backtesting methods for risk measures based on e-statistics with financial data. The use of e-values and e-processes for forecast comparison was first studied by [Henzi and Ziegel \[2022\]](#) on probability forecasts. Sequential testing and estimation of quantiles were also studied in [Howard and Ramdas \[2022\]](#). Bayes pairs were introduced by [Embrechts et al. \[2021\]](#). The backtest e-statistic for the pair of ES and VaR in Theorem 16.11 is closely connected to the joint elicibility of ES and VaR, studied by [Fissler and Ziegel \[2016\]](#). Building e-statistics for the mean and variance and testing them via e-processes was also studied by [Fan et al. \[2025\]](#).

Appendix

A.1 Atomless probability spaces

In several of our definitions, such as those of a calibrator or a merging function, we have a universal quantifier over probability spaces. In this section, we justify our claim in the main text that, when studying concepts like calibrators, e-merging function and p-merging functions, we can safely work with only one atomless probability measure.

Formally, a probability measure $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is *atomless* if it has no *atoms*; an atom is a set $A \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) \in \{0, \mathbb{P}(A)\}$ for any $B \in \mathcal{F}$ with $B \subseteq A$. With the same idea, we also say that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is atomless. We used an alternative definition (point (ii) below) of atomless probability measures in the main text, which is justified by the following well-known lemma.

Lemma A.1

The following three statements are equivalent for any probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

- (i) $(\Omega, \mathcal{F}, \mathbb{P})$ is atomless;
- (ii) there is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ that is uniformly distributed on $[0, 1]$;
- (iii) for any Polish space S and any probability measure R on S , there is a random element on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in S that is distributed as R .

Typical examples of a Polish space in item (iii) that are useful for us are $\mathbb{R}^K$ and countable sets.

Proof.

The equivalence between (i) and (ii) is stated in [Föllmer and Schied \[2016, Proposition A.31\]](#). It remains to prove that (ii) implies (iii). According to Kuratowski's isomorphism theorem [\[Kechris, 1995, Theorem 15.6\]](#), S is Borel isomorphic to $\mathbb{R}$, $\mathbb{N}$, or a finite set (the last two equipped with the discrete topology). The only nontrivial case is where S is Borel isomorphic to $\mathbb{R}$, in which case we can assume $S = \mathbb{R}$. It remains to apply [Föllmer and Schied \[2016, Proposition A.31\]](#) again.

If $(\Omega, \mathcal{F})$ is a measurable space and $\mathcal{P}$ is a collection of probability measures on $(\Omega, \mathcal{F})$, we refer to $(\Omega, \mathcal{F}, \mathcal{P})$ as a *statistical model*. We say that it is *rich* if there exists a random variable on $(\Omega, \mathcal{F})$ that is uniformly distributed on $[0, 1]$ under any $\mathbb{P} \in \mathcal{P}$.

Remark A.2. Intuitively, any statistical model $(\Omega, \mathcal{F}, \mathcal{P})$ can be made rich by complementing it with a random number generator producing a uniform random value in $[0, 1]$: we replace Ω by $\Omega \times [0, 1]$, $\mathcal{F}$ by $\mathcal{F} \times \mathcal{B}$, and each $\mathbb{P} \in \mathcal{P}$ by $\mathbb{P} \times \mathbb{L}$, where $([0, 1], \mathcal{B}, \mathbb{L})$ is the standard Borel space $([0, 1], \mathcal{B})$ equipped with the uniform probability measure $\mathbb{L}$. If $\mathcal{P} = \{\mathbb{P}\}$ contains a single probability measure $\mathbb{P}$, being rich is equivalent to being atomless by Lemma [A.1](#).

For a statistical model $(\Omega, \mathcal{F}, \mathcal{P})$, an *e-variable* is a random variable $E : \Omega \rightarrow [0, \infty]$ satisfying

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[E] \leq 1,$$

as in Chapter 1. The set of e-variables is denoted by $\mathfrak{E}_{\mathcal{P}}$. The set $\mathfrak{E}_{\mathcal{P}}^K$ is the K -fold product set $\mathfrak{E}_{\mathcal{P}} \times \cdots \times \mathfrak{E}_{\mathcal{P}}$.

An *e-merging function* for $(\Omega, \mathcal{F}, \mathcal{P})$, is an increasing Borel function $F : [0, \infty)^K \rightarrow [0, \infty)$ such that, for all $E_1, \dots, E_K$,

$$(E_1, \dots, E_K) \in \mathfrak{E}_{\mathcal{P}}^K \implies F(E_1, \dots, E_K) \in \mathfrak{E}_{\mathcal{P}}.$$

This definition requires that K e-values for $(\Omega, \mathcal{F}, \mathcal{P})$ be transformed into an e-value for $(\Omega, \mathcal{F}, \mathcal{P})$.

Proposition A.3

Let $F : [0, \infty)^K \rightarrow [0, \infty)$ be an increasing Borel function. The following statements are equivalent:

- (i) F is an e-merging function for some rich statistical model;
- (ii) F is an e-merging function for all statistical models (i.e., an e-merging function in Definition 8.1);
- (iii) F is an e-merging function for all statistical models $(\Omega, \mathcal{F}, \mathcal{P})$ with a singleton $\mathcal{P}$.

Proof.

Let us first check that, for any two rich statistical models $(\Omega, \mathcal{F}, \mathcal{P})$ and $(\Omega', \mathcal{F}', \mathcal{P}')$, we always have

$$\begin{aligned} & \sup \{ \mathbb{E}^{\mathbb{P}}[F(\mathbf{E})] : \mathbb{P} \in \mathcal{P}, \mathbf{E} \in \mathfrak{E}_{\mathcal{P}}^K \} \\ &= \sup \{ \mathbb{E}^{\mathbb{P}'}[F(\mathbf{E}')] : \mathbb{P}' \in \mathcal{P}', \mathbf{E}' \in \mathfrak{E}_{\mathcal{P}'}^K \}. \end{aligned} \quad (17)$$

Suppose $\sup \{ \mathbb{E}^{\mathbb{P}}[F(\mathbf{E})] : \mathbb{P} \in \mathcal{P}, \mathbf{E} \in \mathfrak{E}_{\mathcal{P}}^K \} > c$ for some constant c . Then there exist $\mathbf{E} \in \mathfrak{E}_{\mathcal{P}}^K$ and $\mathbb{P} \in \mathcal{P}$ such that $\mathbb{E}^{\mathbb{P}}[F(\mathbf{E})] > c$. Take a random vector $\mathbf{E}' = (E'_1, \dots, E'_K)$ on $(\Omega', \mathcal{F}')$ such that $\mathbf{E}'$ is distributed under each $\mathbb{P}' \in \mathcal{P}'$ identically to the distribution of $\mathbf{E}$ under $\mathbb{P}$. This is possible as $\mathcal{P}'$ is rich by Lemma A.1 applied to the probability space $([0, 1], \mathcal{B}, \mathbb{L})$, where $\mathbb{L}$ is the uniform (Lebesgue) measure and $\mathcal{B}$ is the Borel σ -algebra on $[0, 1]$. By construction, $\mathbf{E}' \in \mathfrak{E}_{\mathcal{P}'}^K$, and $\mathbb{E}^{\mathbb{P}'}[F(\mathbf{E}')] > c$ for all $\mathbb{P}' \in \mathcal{P}'$. This shows that (17) holds with “=” replaced by “ $\leq$ ”, and we obtain the other direction by symmetry.

The implications (ii) $\Rightarrow$ (iii) and (iii) $\Rightarrow$ (i) are obvious. To check (i) $\Rightarrow$ (ii), suppose that F is an e-merging function for some rich statistical model. Consider any statistical model. Its product with the uniform probability measure on $[0, 1]$ will be a rich statistical model (see Remark A.2). It follows from (17) that F will be an e-merging function for the product. Therefore, it will be an e-merging function for the original statistical model.

Remark A.4. The assumption of being rich is essential in item (i) of Proposition A.3. For instance, if we take $\mathcal{P} := \{\delta_{\omega} \mid \omega \in \Omega\}$, where δ_{ω} is the point mass at ω , then $\mathfrak{E}_{\mathcal{P}}$ is the set of all random variables taking values in $[0, 1]$. In this case, the maximum of e-variables is still an e-variable, but the maximum function is not a valid e-merging function as seen from Theorem 8.4.

Proposition A.3 shows that in the definition of an e-merging function it suffices to consider a fixed atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Following the same idea, similar statements can be made for ie-merging functions and se-merging functions in Chapter 8.

We can state a similar proposition for calibrators. A *p-variable for a statistical model* $(\Omega, \mathcal{F}, \mathcal{P})$ is a random variable $P : \Omega \rightarrow [0, \infty)$ satisfying

$$\mathbb{P}(P \leq \varepsilon) \leq \varepsilon \text{ for all } \varepsilon \in (0, 1) \text{ and } \mathbb{P} \in \mathcal{P}.$$

The set of p-variables for $(\Omega, \mathcal{F}, \mathcal{P})$ is denoted by $\mathfrak{U}_{\mathcal{P}}$. A decreasing function $f : [0, 1] \rightarrow [0, \infty]$ is a (*p-to-e*) *calibrator* for $(\Omega, \mathcal{F}, \mathcal{P})$ if $f(P) \in \mathfrak{E}_{\mathcal{P}}$ for any p-variable $P \in \mathfrak{U}_{\mathcal{P}}$.

Proposition A.5

Let $f : [0, 1] \rightarrow [0, \infty]$ be a decreasing Borel function. The following statements are equivalent:

- (i) f is a calibrator for some rich statistical model;
- (ii) f is a calibrator for all statistical models (i.e., a p-to-e calibrator in Definition 2.3);
- (iii) f is a calibrator for all statistical models $(\Omega, \mathcal{F}, \mathcal{P})$ with a singleton $\mathcal{P}$.

The proof is similar to that of Proposition A.3. There is an obvious analogue of Proposition A.5 for e-to-p calibrators in Definition 2.3 and combiners in Chapter 11, which we omit. The content of this section is based on the appendices of Vovk and Wang [2021].

The next lemma is useful for some proofs in Chapter 14, and it is a restatement of Wang and Zitikis [2021, Lemma A.3].

Lemma A.6

Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is an atomless probability space. For any random variable X and $p \in (0, 1)$, there exists an event $A \in \mathcal{F}$ with $\mathbb{P}(A) = p$ such that $\{X > x\} \subseteq A \subseteq \{X \geq x\}$ for some $x \in \mathbb{R}$, which can be chosen as $x = \inf\{t \in \mathbb{R} : \mathbb{P}(X \leq t) \geq 1 - p\}$.

Proof.

Note that the chosen x is the left $(1 - p)$ -quantile of X , which satisfies $\mathbb{P}(X > x) \leq p \leq \mathbb{P}(X \geq x)$; see Lemma A.7. If either of the above inequalities is an equality, then we can take $A = \{X > x\}$ or $A = \{X \geq x\}$. By Lemma A.1, there exists a random variable U uniformly distributed on $[0, 1]$. For $s \in [0, 1]$, let $A_s = \{X > x\} \cup (\{X \geq x\} \cap \{U \leq s\})$. Clearly, $\mathbb{P}(A_0) = \mathbb{P}(X > x) \leq p$ and $\mathbb{P}(A_1) = \mathbb{P}(X \geq x) \geq p$. Moreover, for any $s, t \in [0, 1]$ with $s \leq t$, we have $\mathbb{P}(A_t) - \mathbb{P}(A_s) = \mathbb{P}(A_t \setminus A_s) \leq \mathbb{P}(s \leq U \leq t) = t - s$. Hence, $s \mapsto \mathbb{P}(A_s)$ is continuous. As a consequence, there exists $s \in [0, 1]$ such that $\mathbb{P}(A_s) = p$. The condition $\{X > x\} \subseteq A_s \subseteq \{X \geq x\}$ holds by construction.

A.2 Quantile functions

In this section, we provide a few fundamental facts about quantile functions, which are helpful to understand some results in the book using quantiles in Chapters 2, 6, 15 and 16.

Fix a probability measure $\mathbb{P}$. We define two versions of the quantile function under $\mathbb{P}$. For a random variable X , its left quantile function is defined as

$$Q_{\alpha}^{-}(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq \alpha\} \quad \text{for } \alpha \in (0, 1]$$

and its right quantile function is defined as

$$Q_{\alpha}^{+}(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) > \alpha\} \quad \text{for } \alpha \in [0, 1).$$

Note that $Q_{\alpha}^{-} = q_{1-\alpha}$ for $\alpha \in (0, 1)$ with $q_{1-\alpha}$ in Sections 2.6.6, 6.7.6 and 15.1, but we slightly generalize the definition here by including the possibility of $\alpha = 1$.

The first result is a list of translation rules between probability statements and quantile statements.

Lemma A.7

For $\alpha \in [0, 1]$, $t \in \mathbb{R}$ and any random variable X , the following equivalence statements hold:

- (i) $Q_\alpha^-(X) > t \iff \mathbb{P}(X \leq t) < \alpha$;
- (ii) $Q_\alpha^-(X) \leq t \iff \mathbb{P}(X \leq t) \geq \alpha$;
- (iii) $Q_\alpha^-(X) \geq t \iff \mathbb{P}(X \leq s) < \alpha$ for all $s < t$;
- (iv) $Q_\alpha^-(X) < t \iff \mathbb{P}(X \leq s) \geq \alpha$ for some $s < t$;
- (v) $Q_\alpha^+(X) < t \iff \mathbb{P}(X < t) > \alpha$;
- (vi) $Q_\alpha^+(X) \geq t \iff \mathbb{P}(X < t) \leq \alpha$;
- (vii) $Q_\alpha^+(X) \leq t \iff \mathbb{P}(X < s) > \alpha$ for all $s > t$;
- (viii) $Q_\alpha^+(X) > t \iff \mathbb{P}(X < s) \leq \alpha$ for some $s > t$.

Proof.

To show (i), denote by $A_\alpha = \{t \in \mathbb{R} : \mathbb{P}(X \leq t) \geq \alpha\}$. Note that A_α is closed in $\mathbb{R}$ since $t \mapsto \mathbb{P}(X \leq t)$ is upper semicontinuous. This gives $Q_\alpha^-(X) = \min A_\alpha$. Hence, $\alpha > \mathbb{P}(X \leq t) \iff t \notin A_\alpha \iff Q_\alpha^-(X) > t$. Part (ii) follows from (i) directly; (iii) follows from (i) and (iv) follows from (ii).

To show (v), denote by $B_\alpha = \{t \in \mathbb{R} : \mathbb{P}(X < t) \leq \alpha\}$ which is closed in $\mathbb{R}$ since $t \mapsto \mathbb{P}(X < t)$ is lower semicontinuous. This gives $Q_\alpha^+(X) = \max B_\alpha$. It follows that $\alpha < \mathbb{P}(X < t) \iff t \notin B_\alpha \iff Q_\alpha^+(X) < t$. Part (vi) follows from (v) directly; (vii) follows from (v) and (viii) follows from (vi).

The next lemma clarifies that the two quantile functions are almost everywhere equal.

Lemma A.8

For any random variable X , $Q_\alpha^-(X) = Q_\alpha^+(X)$ for almost every $\alpha \in (0, 1)$.

Proof.

Note that, for any $\beta < \alpha$, we have $Q_\alpha^-(X) \geq Q_\beta^+(X) \geq Q_\beta^-(X)$. Hence, by sending $\alpha \downarrow \beta$, we see that the two values $Q_\beta^+(X)$ and $Q_\beta^-(X)$ can only differ at discontinuous points of the curve $\alpha \mapsto Q_\alpha^-(X)$. Since $\alpha \mapsto Q_\alpha^-(X)$ is increasing, its jump points are countable, and thus the statement holds true.

The next lemma shows that if the quantile level α is replaced by a uniformly distributed random variable U on $[0, 1]$, then $Q_U^-(X)$ is identically distributed as X , where $Q_U^-(X)$ is understood as the random variable given by $\omega \mapsto Q_{U(\omega)}(X)$. This is true also for $Q_U^+(X)$. Note that $Q_0^-(X)$ and $Q_1^+(X)$ are undefined, but this does not matter since U takes the value 0 or 1 with zero probability.

Lemma A.9

For any random variable X and any uniformly distributed random variable U on $[0, 1]$, the random variables X , $Q_U^-(X)$ and $Q_U^+(X)$ are identically distributed.

Proof.

By Lemma A.8, $Q_U^-(X) = Q_U^+(X)$ almost surely. Therefore, it suffices to consider $Q_U^-(X)$. By Lemma A.7, we have, for any $\alpha \in (0, 1)$ and $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(X \leq t) < \alpha &\iff Q_\alpha^-(X) > t \\ &\implies \mathbb{P}(Q_U^-(X) > t) \geq 1 - \alpha \iff \mathbb{P}(Q_U^-(X) \leq t) \leq \alpha, \end{aligned} \quad (18)$$

and

$$\mathbb{P}(X \leq t) \geq \alpha \iff Q_\alpha^-(X) \leq t \implies \mathbb{P}(Q_U^-(X) \leq t) \geq \alpha. \quad (19)$$

Note that (18) implies

$$\begin{aligned} \mathbb{P}(X \leq t) \leq \alpha &\iff \mathbb{P}(X \leq t) < \beta \text{ for all } \beta > \alpha \\ &\implies \mathbb{P}(Q_U^-(X) \leq t) \leq \beta \text{ for all } \beta > \alpha \\ &\iff \mathbb{P}(Q_U^-(X) \leq t) \leq \alpha. \end{aligned}$$

Putting this and (19) together we get $X \stackrel{d}{=} Q_U^-(X)$.

Lemma A.9 implies, in particular, the well known formula

$$\mathbb{E}^\mathbb{P}[X] = \int_0^1 Q_\alpha^-(X) d\alpha = \int_0^1 Q_\alpha^+(X) d\alpha.$$

The next result is due to [Ryff \[1965, Theorem 1\]](#), although presented in a different form.

Lemma A.10

Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is an atomless probability space. For any random variable X , there exists a uniformly distributed random variable U on $[0, 1]$ such that $X = Q_U^-(X) = Q_U^+(X)$ almost surely.

We first explain the simplest and most intuitive case. Let F be the cdf of X . If X is continuously distributed, then it suffices to choose $U = F(X)$, which satisfies the conditions in Lemma A.10 by a standard probabilistic exercise. The complicated case is when X is not continuously distributed and there does not exist a $U[0, 1]$ random variable independent of X .

Proof.

By Lemma A.8, $Q_U^-(X) = Q_U^+(X)$ almost surely, so we will only show the first equality. First, suppose that there exists a standard uniform random variable V independent of X . Define

$$U = F(X) - V(F(X) - F(X-)), \quad (20)$$

where $F(x-) = \lim_{y \uparrow x} F(y)$ for any $x \in \mathbb{R}$. It is a standard exercise in probability theory to check that U is uniformly distributed on $[0, 1]$ and $X = Q_U^-(X)$ almost surely.

Next, we consider the more difficult case that there does not exist any uniformly distributed random variable independent of X . In this case, let B be the set of discontinuity points of F , which is a countable set. If B is empty then X is continuously distributed, and U can be chosen as $F(X)$. Next, suppose that B is not empty. Denote by $A_b = \{X = b\}$ for $b \in B$. We have $\mathbb{P}(A_b) > 0$, because $\mathbb{P}(A_b) = F(b) - F(b-)$. Since the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is atomless and $\mathbb{P}(A_b) > 0$, the space $(A_b, \mathcal{F}|_{A_b}, \mathbb{P}|_{A_b})$ is also atomless. Hence, there exists a $U[0, 1]$ -distributed random variable V_b on this space. We extend V_b to $(\Omega, \mathcal{F}, \mathbb{P})$ by setting $V_b = 0$ outside A_b .

Finally, define

$$U^* = F(X) - V_X(F(X) - F(X-)),$$

where V_X is understood as the random variable $V_X(\omega) = V_b(\omega)$ if $X(\omega) = b \in B$, and otherwise $V_X(\omega) = 0$. Note that V_X is $\mathcal{F}$ -measurable since B is countable. Moreover, U^* has the same distribution as U in (20), because conditional on A_b for each $b \in B$, V_b and V are identically distributed, and outside $\bigcup_{b \in B} A_b$ the random variables V and V_X do not matter. Following the same statement for U in (20), we obtain that U^* satisfies the desired conditions in the lemma.

Bibliography

- Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, volume 134, pages 26–62. PMLR, 2021.
- Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- Timothy B. Armstrong. False discovery rate adjustments for average significance level controlling tests. *arXiv:2209.13686*, 2022.
- Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.
- Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.
- Trambak Banerjee, Bowen Gang, and Jianliang He. Harnessing the collective wisdom: Fusion learning using decision sequences from diverse sources. *arXiv:2308.11026*, 2023.
- Meshi Bashari, Amir Epstein, Yaniv Romano, and Matteo Sesia. Derandomized novelty detection with FDR control via conformal e-values. In *Advances in Neural Information Processing Systems*, volume 36, pages 65585–65596, 2023.
- Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Principal-agent hypothesis testing. *arXiv:2205.06812*, 2022.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300, 1995.
- Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.
- David R Bickel. David r. bickel’s contribution to the discussion of ‘safe testing’ by grünwald, de heide, and koolen. *Journal of the Royal Statistical Society Series B*, 86(5):1133–1134, 08 2024.
- David R Bickel. A small-sample bayesian information criterion that does not overstate the evidence, with an application to calibrating p-values from likelihood-ratio tests. *Statistical Papers*, 66(3):1–17, 2025.
- Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.
- Christopher Blier-Wong and Ruodu Wang. Improved thresholds for e-values. *arXiv:2408.11307*, 2024.
- Henry W Block, Thomas H Savits, and Moshe Shaked. Some concepts of negative dependence. *The Annals of Probability*, 10(3):765–772, 1982.

- Leo Breiman. Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.
- Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: ‘Model-X’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, 80(3): 551–577, 2018.
- Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542, 2009.
- Ziyu Chi, Aaditya Ramdas, and Ruodu Wang. Multiple testing under negative dependence. *Bernoulli*, 31(2): 1230–1255, 2024.
- Eugenio Clerico. Optimal e-value testing for properly constrained hypotheses. *arXiv:2412.21125*, 2024.
- John Copas. Compound decisions and empirical Bayes. *Journal of the Royal Statistical Society Series B (with Discussion)*, 31(3):397–417, 1969.
- Thomas Cover. An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory*, 30(2):369–373, 1984.
- Imre Csiszar and Frantisek Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.
- Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions*, pages 205–237, 1984. Recent Results in Estimation Theory and Related Topics.
- Donald A. Darling and Herbert Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967.
- Freddy Delbaen. Coherent risk measures on general probability spaces. In *Advances in Finance and Stochastics*, pages 1–37. Springer, 2002.
- Freddy Delbaen. *Monetary utility functions*. Osaka University Press, 2012.
- Dieter Denneberg. *Non-additive measure and integral*, volume 27. Springer Science & Business Media, 1994.
- Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.
- Vaidehi Dixit and Ryan Martin. Anytime-valid and asymptotically efficient inference driven by predictive recursion. *Biometrika*, 112(2), 11 2024.
- Joseph L. Doob. *Stochastic processes*. Wiley, New York, 1953.
- Robin Dunn, Aaditya Ramdas, Sivaraman Balakrishnan, and Larry Wasserman. Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337, 2023.
- Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal inference meets random projections: a scalable test for log-concavity. *Journal of Computational and Graphical Statistics*, 34(1): 267–279, 2025.
- Paul Embrechts and Ruodu Wang. Seven proofs for the subadditivity of expected shortfall. *Dependence Modeling*, 3(1):126–140, 2015.
- Paul Embrechts, Tiantian Mao, Qiuqi Wang, and Ruodu Wang. Bayes risk, elicibility, and the expected shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.
- Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.
- Yixuan Fan, Zhanyi Jiao, and Ruodu Wang. Testing the mean and variance by e-processes. *Biometrika*, 112 (1):asae049, 2025.

- Lasse Fischer and Aaditya Ramdas. Admissible online closed testing must employ e-values. *arXiv:2407.15733*, 2024.
- Lasse Fischer, Ziyu Xu, and Aaditya Ramdas. An online generalization of the (e-)Benjamini-Hochberg procedure. *arXiv:2407.20683*, 2024.
- Ronald A. Fisher. Combining independent tests of significance. *American Statistician*, 2:30, 1948.
- Tobias Fissler and Johanna F Ziegel. Higher order elicibility and osband’s principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.
- Hans Föllmer and Alexander Schied. *Stochastic finance: An introduction in discrete time*. De Gruyter, Berlin, fourth ed edition, 2016.
- Paula Gablenz and Chiara Sabatti. Catch me if you can: Signal localization with knockoff e-values. *Journal of the Royal Statistical Society Series B*, 87(1):56—73, 2025.
- Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1998.
- Matteo Gasparin and Aaditya Ramdas. Conformal online model aggregation. *arXiv:2403.15527*, 2024a.
- Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv:2401.09379*, 2024b.
- Matteo Gasparin, Ruodu Wang, and Aaditya Ramdas. Combining exchangeable p-values. *Proceedings of the National Academy of Sciences*, 122(11):e2410849122, 2025.
- Jayanta K Ghosh and Pranab Kumar Sen. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. Technical report, North Carolina State University. Dept. of Statistics, 1984.
- Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.
- Evarist Giné, Friedrich Götze, and David M. Mason. When is the Student t-statistic asymptotically standard normal? *The Annals of Probability*, 25(3):1514–1531, 1997.
- Jelle Goeman, Rianne de Heide, and Aldo Solari. The e-partitioning principle of false discovery rate control. *arXiv:2504.15946*, 2025.
- Bruno Goffinet, Patrice Loisel, and Beatrice Laurent. Testing in normal mixture models when the proportions are known. *Biometrika*, 79(4):842–846, 1992.
- Benjamin T. Graham and Geoffrey R. Grimmett. Influence and sharp-threshold theorems for monotonic measures. *The Annals of Probability*, 34:1726–1745, 2006.
- Peter Grünwald. Beyond Neyman-Pearson: e-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024.
- Peter Grünwald, Rianne De Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B (with Discussion)*, 86(5):1091–1128, 2024a.
- Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K Bar-Lev, and Martijn de Jong. Optimal e-values for exponential families: the simple case. *arXiv:2404.19465*, 2024b.
- Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.
- Will Hartog and Lihua Lei. Family-wise error rate control with e-values. *arXiv:2501.09015*, 2025.
- Alexander Henzi and Johanna F Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663, 2022.

- Gerhard Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, 25:423–430, 1983.
- Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Conference on Learning Theory*, pages 67–79. Citeseer, 2010.
- Steven R Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022.
- Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.
- Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.
- Peter J Huber. *Robust statistics*. Wiley Series in Probability and Mathematical Statistics, 1981.
- Nikolaos Ignatiadis and Bodhisattva Sen. Empirical partially Bayes multiple testing and compound χ^2 decisions. *The Annals of Statistics*, 53(1):1–36, 2025.
- Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. Asymptotic and compound e-values: multiple testing and empirical Bayes. *arXiv:2409.19812*, 2024a.
- Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. E-values as unnormalized weights in multiple testing. *Biometrika*, 111(2):417–439, 2024b.
- Harold Jeffreys. *Theory of probability*. Oxford University Press, London, 3rd edition, 1961.
- Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.
- Elyès Jouini, Walter Schachermayer, and Nizar Touzi. Law invariant risk measures have the Fatou property. In *Advances in Mathematical Economics*, pages 49–71. Springer, 2006.
- Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.
- Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22:246–1, 2021.
- Alexander S. Kechris. *Classical descriptive set theory*. Springer, New York, 1995.
- Hans G Kellerer. Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 67:399–432, 1984.
- John Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926, 1956.
- Nick W Koning. Post-hoc α hypothesis testing and the post-hoc p-value. *arXiv:2312.08040*, 2023a.
- Nick W Koning. Post-hoc and anytime valid permutation and group invariance testing. *arXiv:2310.01153*, 2023b.
- Nick W Koning. A continuous generalization of hypothesis testing. *arXiv:2409.05654*, 2024.
- W. Koolen and P. Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82, 2021. Festschrift for G. Shafer’s 75th Birthday.
- Charles Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publication in Statistics*, 2:125–141, 1955.
- Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. The HulC: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B*, 86(3):586–622, 2024.

- Ludmila I Kuncheva, Christopher J Whitaker, Catherine A Shipp, and Robert PW Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6:22–31, 2003.
- Tze Leung Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976.
- HO Lancaster. Statistical control of counting experiments. *Biometrika*, 39(3/4):419–422, 1952.
- Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 2024.
- Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025a.
- Martin Larsson, Aaditya Ramdas, and Johannes Ruf. Testing hypotheses generated by constraints. *arXiv:2504.02974*, 2025b.
- Junu Lee and Zhimei Ren. Boosting e-BH via conditional calibration. *arXiv:2404.17562*, 2024.
- Erich L. Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer New York, New York, NY, 3 edition, 2005.
- Guanxun Li and Xianyang Zhang. A note on e-values and multiple testing. *Biometrika*, 10 2024.
- Jonathan Li. *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.
- Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.
- Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023.
- Tiantian Mao, Bin Wang, and Ruodu Wang. Sums of standard uniform random variables. *Journal of Applied Probability*, 56(3):918–936, 2019.
- Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: Concepts, techniques and tools. Revised edition*. Princeton University Press, 2015.
- Xiao-Li Meng. Posterior predictive p -values. *The Annals of Statistics*, 22(3):1142–1160, 1994.
- Stanislav Minsker. U-statistics of growing order and sub-Gaussian mean estimators with sharp constants. *Mathematical Statistics and Learning*, 2023.
- Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.
- Marcel Nutz, Ruodu Wang, and Zhenyuan Zhang. Martingale transports and monge maps. *The Annals of Applied Probability*, 34(6):5556–5577, 2024.
- Karl Pearson. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 25:379–410, 1933.
- Shige Peng. *Nonlinear expectations and stochastic calculus under uncertainty*. Springer, 2019.
- Carlos Alberto de Bragança Pereira and Julio Michael Stern. Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110, 1999.
- Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne De Heide, and Peter Grünwald. E-statistics, group invariance and anytime valid testing. *The Annals of Statistics*, 52(4):1410–1432, 2024.
- Svetlozar T Rachev and Ludger Rüschendorf. *Mass transportation problems volume 1: Theory*. Springer Science & Business Media, 2006.

- Aaditya Ramdas and Tudor Manole. Randomized and exchangeable improvements of Markov’s, Chebyshev’s and Chernoff’s inequalities. *Statistical Science*, 2025.
- Aaditya Ramdas, Rina F. Barber, Martin J. Wainwright, and Michael I. Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47:2790–2821, 2019.
- Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020.
- Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.
- Zhimei Ren and Rina Foygel Barber. Derandomised knockoffs: Leveraging e -values for false discovery rate control. *Journal of the Royal Statistical Society Series B*, 86(1):122–154, 2024.
- Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 131–149. University of California Press, 1951.
- Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. The Regents of the University of California, 1956.
- Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.
- Herbert Robbins and David Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436, 1974.
- R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.
- Patrick Rubin-Delanchy, Nicholas A Heard, and Daniel J Lawson. Meta-analysis of mid-p-values: some new results based on the convex order. *Journal of the American Statistical Association*, 114(527):1105–1112, 2019.
- Bernhard Rüger. Das maximale Signifikanzniveau des Tests “Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen”. *Metrika*, 25:171–178, 1978.
- Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.
- Ludger Rüschendorf. *Mathematical risk analysis*. Springer, 2013.
- Ludger Rüschendorf, Steven Vanduffel, and Carole Bernard. *Model risk management: Risk bounds under uncertainty*. Cambridge University Press, 2024.
- John V Ryff. Orbits of l^1 -functions under doubly stochastic transformation. *Transactions of the American Mathematical Society*, 117:92–100, 1965.
- David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.
- Byungtae Seo and Bruce G Lindsay. A universally consistent modification of maximum likelihood. *Statistica Sinica*, 23(2):467–487, 2013.
- Glenn Shafer. Testing by betting: a strategy for statistical and scientific communication. *Journal of the Royal Statistic Society A (with Discussion)*, 184(2):407–431, 2021.

- Glenn Shafer. Did Jean Ville invent martingales? In *The Splendors and Miseries of Martingales: Their History from the Casino to Mathematics*, pages 107–122. Springer, 2022.
- Glenn Shafer. Improving data analysis by testing by betting: optional continuation and descriptive statistics. *The New England Journal of Statistics in Data Science*, 2(2):215–228, 2023.
- Glenn Shafer and Vladimir Vovk. *Probability and finance: It’s only a game*. Wiley, New York, 2001.
- Glenn Shafer and Vladimir Vovk. *Game-theoretic foundations for probability and finance*. Wiley, Hoboken, New Jersey, 2019.
- Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.
- Moshe Shaked and J. George Shantikumar. *Stochastic orders*. Springer Science & Business Media, 2007.
- Hongjian Shi and Mathias Drton. On universal inference in Gaussian mixture models. *arXiv:2407.19361*, 2024.
- Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: a nonparametric framework for online changepoint detection. *New England Journal of Statistics in Data Science*, 2(2):229–260, 2024.
- R. John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.
- Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184: 109395, 2022.
- Aldo Solari and Jelle J Goeman. Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. *Biometrical Journal*, 59(4):776–780, 2017.
- Julio Michael Stern, Carlos Alberto de Braganca Pereira, Marcelo de Souza Lauretto, Luis Gustavo Esteves, Rafael Izbicki, Rafael Bassi Stern, Marcio Alves Diniz, and Wagner de Souza Borges. The e-value and the full Bayesian significance test: logical properties and philosophical consequences. *arXiv:2205.08010*, 2022.
- David Strieder and Mathias Drton. On the choice of the splitting ratio for the split likelihood ratio test. *Electronic Journal of Statistics*, 16(2):6631–6650, 2022.
- Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2014.
- J. Ter Schure and P. Grünwald. ALL-IN meta-analysis: breathing life into living systematic reviews. *F1000Research*, 11(549), 2022.
- Leonard H. C. Tippett. *The methods of statistics: An introduction mainly for experimentalists*. Williams and Norgate, London, 1931.
- Athanasios Tsanas and Max Little. Parkinsons telemonitoring. UCI Machine Learning Repository, 2009. DOI: <https://doi.org/10.24432/C5ZS3N>.
- Timmy Tse and Anthony C Davison. A note on universal inference. *Stat*, 11(1):e501, 2022.
- A. W. Van Der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK; New York, NY, USA, 1998.
- Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *The Annals of Internal Medicine*, 167(4):268–274, 2017.
- Cédric Villani. *Optimal transport: old and new*. Springer, 2009.
- Jean Ville. *Étude critique de la notion de collectif (Doctoral thesis)*. Gauthier-Villars, 1939.

- Vladimir Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society Series B (with Discussion)*, 55:317–351, 1993.
- Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021.
- Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.
- Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49:1736–1754, 2021.
- Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 38(2):329–354, 2023.
- Vladimir Vovk and Ruodu Wang. Merging sequential e-values via martingales. *Electronic Journal of Statistics*, 18:1185–1205, 2024a.
- Vladimir Vovk and Ruodu Wang. Nonparametric e-tests of symmetry. *The New England Journal of Statistics in Data Science*, 2(2):261–270, 2024b.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.
- Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.
- Vladimir G Vovk and Vladimir V V'yugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society Series B*, 55(1):253–266, 1993.
- Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
- Abraham Wald. *Sequential analysis*. Wiley, New York, 1947.
- Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.
- Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman & Hall, 1991.
- Bin Wang and Ruodu Wang. Joint mixability. *Mathematics of Operations Research*, 41(3):808–826, 2016.
- Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and Their Applications*, 163:168–202, 2023.
- Hongjian Wang and Aaditya Ramdas. Anytime-valid t-tests and confidence sequences for Gaussian means with unknown variance. *Sequential Analysis*, 44(1):56–110, 2025.
- Qiuqi Wang, Ruodu Wang, and Johanna Ziegel. E-backtesting. *Management Science*, 2025.
- Ruodu Wang. Ruodu Wang’s contribution to the discussion of ‘Testing by betting: A strategy for statistical and scientific communication’ by Glenn Shafer. *Journal of the Royal Statistical Society Series A*, 184(2):463–464, 2021.
- Ruodu Wang. Proposer of the vote of thanks to Grünwald, de Heide, and Koolen and contribution to the discussion of ‘Safe testing’. *Journal of the Royal Statistical Society Series B*, 86(5):1129–1131, 2024a.
- Ruodu Wang. Testing with p*-values: Between p-values, mid p-values, and e-values. *Bernoulli*, 30(2):1313–1346, 2024b.
- Ruodu Wang. The only admissible way of merging arbitrary e-values. *Biometrika*, 112(2):asaf020, 2025.
- Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B*, 84(3):822–852, 2022.

- Ruodu Wang and Ričardas Zitikis. An axiomatic foundation for the expected shortfall. *Management Science*, 67(3):1413–1429, 2021.
- Shaun S Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, 67(1):15–36, 2000.
- Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.
- Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B (with Discussion)*, 86(1):1–27, 2024.
- Daniel J Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.
- Ziyu Xu and Aaditya Ramdas. More powerful multiple testing under dependence via randomization. *arXiv:2305.11126*, 2023.
- Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics*, pages 3997–4005. PMLR, 2024.
- Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. *Electronic Journal of Statistics*, 18(1):2292–2338, 2024.
- Ziyu Xu, Lasse Fischer, and Aaditya Ramdas. Bringing closure to FDR control: beating the e-Benjamini-Hochberg procedure. *arXiv:2504.11759*, 2025.
- Cun-Hui Zhang. Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, 31(2):379–390, 2003.
- Zhenyuan Zhang, Aaditya Ramdas, and Ruodu Wang. On the existence of powerful p-values and e-values for composite hypotheses. *The Annals of Statistics*, 52(5):2241–2267, 2024.
- Zinan Zhao and Wenguang Sun. False discovery rate control for structured multiple testing: Asymmetric rules and conformal Q-values. *Journal of the American Statistical Association*, 120(550):805–817, 2025.