Hypothesis Testing with E-values

Aaditya Ramdas[1] and Ruodu Wang[2]
aramdas@cmu.edu, wang@uwaterloo.ca

November 4, 2024

[1]Department of Statistics and Data Science, and Machine Learning Department, Carnegie Mellon University
[2]Department of Statistics and Actuarial Science, University of Waterloo

Contents

Preface

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, that are both under-studied and under-utilized; (c) the application domains in the natural and social sciences would benefit from knowing and adapting methodologies of e-values in certain contexts to address various pressing issues related to scientific reproducibility; (d) there has been an explosion of exciting research over the past few years, and we think the time is right 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 less than 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 four points to justify this, which will be clear from the content of the book. First, we will show that a nontrivial test exists if and only if a nontrivial e-value exists. Second, for many simple-to-state and fundamental composite testing problems (for example, testing if your univariate data comes from a mixture of two Gaussians), the *only* test we are aware of proceeds by constructing an e-value. This technique is called *universal inference*, and the e-value is the *split likelihood ratio*. Third, 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 arbitrarily 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.

(c) The recent crisis of scientific reproducibility — largely related to the use and misuse of p-values (especially peeking at p-values and optional stopping and continuation of experiments) — 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 2020-2024, 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, 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 places, sometimes with different notations and terminologies, due to the fast development. Thus, there is a 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 safe and valid statistical inference in the natural and social sciences. From the 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, but which is intimately connected to e-values. Further, except for Chapter 6, 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 books dedicated to these topics in future years.

Acknowledgments

We acknowledge our many wonderful coauthors over the years on these topics, and our family members, especially our spouses, for their unwavering support.

We have received, and continue to receive, many useful comments and corrections from our readers and colleagues, who we will properly acknowledge in a later version.

Lists of notation, conventions, and examples

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 $\{0, 1, 2, \dots\}$
$[K]$	$K \in \mathbb{N}$	$\{1, \dots, K\}$
$a \vee b$	$a, b \in \mathbb{R}$	$\max(a, b)$
$a \wedge b$	$a, b \in \mathbb{R}$	$\min(a, b)$
a_+	$a \in \mathbb{R}$	$a \vee 0$
a_-	$a \in \mathbb{R}$	$(-a) \vee 0$
e		$\exp(1)$
$\log$		natural logarithm to the base 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)
$x_{(k)}$	$x_1, \dots, x_K \in \mathbb{R}$, $k \in [K]$	kth ascending order statistic of $x_1, \dots, x_K$
$x_{[k]}$	$x_1, \dots, x_K \in \mathbb{R}$, $k \in [K]$	kth 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 $(x_1, \dots, x_K) \mapsto \sum_{k=1}^K \phi_k(x_k)$

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 $\mathbb{R}$
$\mathbb{P}$, $\mathbb{Q}$		probability measures on $(\Omega, \mathcal{F})$
$\mathbb{P}^n$	$\mathbb{P} \in \mathcal{M}_1$, $n \in \mathbb{N}$	the n-fold product measure of $\mathbb{P}$ on $\Omega \times \cdots \times \Omega$
$\mathbb{Q} \ll \mathbb{P}$	$\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1$	$\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$
$\mathbb{E}^{\mathbb{P}}[X]$	$\mathbb{P} \in \mathcal{M}_+$, $X \in \mathcal{X}$	$\int X \mathrm{d}\mathbb{P}$; it is the expectation under $\mathbb{P}$ if $\mathbb{P} \in \mathcal{M}_1$
$\mathrm{var}^{\mathbb{P}}(X)$	$\mathbb{P} \in \mathcal{M}_1$, $X \in \mathcal{X}$	variance of X under $\mathbb{P}$
$\mathcal{P}$, $\mathcal{Q}$		sets of probability measures on $(\Omega, \mathcal{F})$
$\mathrm{Conv}(\mathcal{P})$	$\mathcal{P} \subseteq \mathcal{M}_1$	convex hull of $\mathcal{P}$
$X \overset{\mathrm{d}}{\sim} \mu$	$\mu \in \mathcal{M}_1(\mathbb{R})$, $X \in \mathcal{X}$	X is distributed as μ (under some $\mathbb{P}$)
$\mathrm{U}[a, b]$	$a < b$	uniform distribution on $[a, b]$
$\mathrm{N}(\mu, \sigma^2)$	$\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$	normal distribution with parameter (μ, σ^2)
$\mathbb{1}_A$	$A \in \mathcal{F}$	binary-valued indicator function of event A
δ_x	$x \in \mathbb{R}$	Dirac delta distribution at x
Φ		standard normal cumulative distribution function

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

Abbreviations

Abbreviation	**Meaning**
iid	independent and identically distributed
pdf	probability density function
cdf	cumulative distribution function
pmf	probability mass function
wrt	with respect to
KL	Kullback-Leibler (divergence), also called the 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 like a, p, e, f, or x.

2. The constants e (often a realized e-value) and $\mathtt{e}$ (exponential of 1) are different, and we avoid using them in the same place.

3. Bolded constants like $\mathbf{x}$ refer to vectors, and we refer to its k-th entry as x_k (not $\mathbf{x}_k$).

4. Capital letters like X and E are usually reserved for random variables and functions, but there are exceptions made for important constants (like the number of hypotheses K in multiple testing). Random vectors are denoted like $\mathbf{X}$ 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. Terms like "positive" and "negative" are in the strict sense.

6. Inequalities between vectors 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. When $\mathbb{E}^{\mathbb{P}}[F(X)]$ appears, the suitable measurability of F is tacitly assumed.

9. We sometimes omit the domain of integrals and it should be understood as the natural domain (e.g., $\mathbb{R}$ or Ω), which should be clear from context.

10. When we do not provide the proof of a nontrival result, we will point to its proof in the literature in the bibliographical note section of the corresponding chapter.

11. 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 $\infty/\infty = 1$.

12. 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.

13. If $\mathbb{E}[\cdot]$ or $\mathrm{var}(\cdot)$ appears without the probability specified, it is with respect to the true data generating probability.

List of examples of e-values

Context	Null	Alternative	Location
Likelihood ratio	$\mathbb{P}$	$\mathbb{Q}$	Section 1.3
Soft-rank e-value	exchangeable data	otherwise	Section 1.4
Normal two-sided	$\mathrm{N}(0,1)$	$\mathrm{N}(\mu,1),\ \mu \neq 0$	Section 2.1
Mean with bounded variance	$\mathbb{E}[X]=0,\ \mathrm{var}(X) \leq \sigma^2$	otherwise	Section 2.9
Mixture method	$\mathbb{P}$	$\mathcal{Q}$	Section 3.4
T-test (unknown σ)	$\mathrm{N}(0,\sigma^2)$, σ unknown	$\mathrm{N}(\mu,\sigma^2);\ \mu \neq 0$	Section 3.4
Plug-in method	$\mathbb{P}$	$\mathcal{Q}$	Section 3.4
Universal inference e-value	$\mathcal{P}$	$\mathcal{Q}$	Section 4.2
Numeraire	$\mathcal{P}$	$\mathbb{Q}$	Section 5.1
Exponential family	$p_\theta : \theta \geq \theta_0$	p_{θ_1}	Section 5.6
Symmetry	symmetry around zero	otherwise	Section 5.6
Mean with support $[0,1]$	mean $\leq \mu$	mean μ_1	Section 5.6
SubGaussian distributions	mean ≤ 0	mean μ	Section 5.6
Likelihood ratio bounds	$\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}_0 \leq \gamma$ for given $\mathbb{P}_0$	$\mathbb{Q}$	Section 5.6
Likelihood ratio process	$\mathbb{P}$	$\mathbb{Q}$	Section 6.4
Universal inference e-process	$\mathcal{P}$	$\mathcal{Q}$	Section 6.6
Empirically adaptive e-process	$\mathcal{P}$	$\mathcal{Q}$	Section 6.7
Time-mixture e-processes	$\mathcal{P}$	$\mathcal{Q}$	Section 6.8
Multi-armed bandit testing	$\mathbb{E}[X_{k,j} \vert \mathcal{F}_{j-1}] \leq 1\ \forall j$	$\mathbb{E}[X_{k,j} \vert \mathcal{F}_{j-1}] > 1\ \exists j$	Section 8.1
Compound t-test (unknown σ_k)	$\mathrm{N}(0,\sigma_k^2);\ k \in [K]$	$\mathrm{N}(\mu_k,\sigma_k^2);\ k \in [K]$	Section 8.3
Testing the β-quantile (VaR_β)	$\mathrm{VaR}_\beta(X) \leq r$	$\mathrm{VaR}_\beta(X) > r$	Section 13.3
Testing an expected loss	$\mathbb{E}[\ell(X)] \leq r$	$\mathbb{E}[\ell(X)] > r$	Section 13.3
Testing the risk measure ES	$\mathrm{ES}_\beta(X) \leq r,\ \mathrm{VaR}_\beta(X) = z$	$\mathrm{ES}_\beta(X) > r$	Section 13.3

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. For a formal definition see Section 1.2.

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); see Chapter 3. 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 (Chapters 4 and 5).

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 (also defined in Section 1.2). The naming analogy is that under the null hypothesis,

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

since p-values must be uniformly distributed (or stochastically larger) under the null. Another analogy, treated in Section 2.8, 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.

Before diving into technical definitions, we briefly mention some major differences between e-values and p-values, and situations in which one might prefer the former to the latter.

1.1 Standing notation and conventions

Throughout, $\mathbb{N} = \{1, 2, \dots\}$ is the set of all positive integers and $\mathbb{R} = (-\infty, \infty)$ is the set of all real numbers. 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,

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

We begin with a sample space Ω equipped with a σ-algebra $\mathcal{F}$, and the set $\mathcal{M}_1$ of all probability measures (called distributions) on $(\Omega, \mathcal{F})$. We assume that some distribution $\mathbb{P}^* \in \mathcal{M}_1$ governs our data X. 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 $\mathbb{P}^*$, 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, like $\{\mathbb{P}\}$ 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 $\mathbb{P}^* \in \mathcal{P}$, and H_1 the alternative hypothesis that $\mathbb{P}^* \in \mathcal{Q}$; 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 $\mathrm{Conv}(\mathcal{P})$ 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.

The Euler constant $\exp(1)$ is denoted by $\mathtt{e}$ (please note that this is different from the variable e, but of course we will avoid using both in the same equation whenever possible). We generally rely on standard measure-theoretic 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, the last part of which is somewhat unusual but simplifies notation:

$$x/\infty = 0 \text{ and } \infty/x = \infty \text{ for } x \in [0, \infty), \text{ and } \infty/\infty = 1.$$

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 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 $\mathbb{P}(X \le x) \le F(x)$ for all $x \in \mathbb{R}$. We write $X \overset{\mathrm{d}}{\sim} F$ if the distribution of X is F under a probability measure $\mathbb{P}$ that should be clear from context.

A distribution $\mathbb{Q}$ is said to be absolutely continuous with respect to another distribution $\mathbb{P}$, 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 $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ 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 $(\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P})(x) = q(x)/p(x)$.

We will introduce some more specific notation in different chapters, and a list of commonly used notation and convention is presented separately in the beginning of the book.

1.2 E-variables, p-variables, and tests

We first define the most important stochastic quantities underlying the ideas of e-values and p-values. As above, let $\mathcal{P} \subseteq \mathcal{M}_1$ be a set of probability measures.

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 $\mathbb{E}^{\mathbb{P}}[E] \le 1$ 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 that satisfies $\mathbb{P}(P \le \alpha) \le \alpha$ for all $\alpha \in (0, 1)$ and all $\mathbb{P} \in \mathcal{P}$. A p-variable P is *exact* if the preceding inequality holds with equality.

(iii) A *test* is a measurable function $\phi : \Omega \to [0, 1]$. A test is *binary* if its range is $\{0, 1\}$. The *type-I error* 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 $\mathfrak{E} = \mathfrak{E}(\mathcal{P})$ the set of all e-variables for $\mathcal{P}$ and by $\mathfrak{U} = \mathfrak{U}(\mathcal{P})$ the set of all p-variables for $\mathcal{P}$, with $\mathcal{P}$ often omitted.

Above, we use $\mathfrak{U}$ (hinting at "uniform") instead of $\mathfrak{P}$ to avoid overloading the letter "P", which are already used for p-variables (often P), probability measures (often $\mathbb{P}$), and sets of probability measures (often $\mathcal{P}$).

Some remarks are below:

1. A p-variable is often truncated at 1; this usually does not affect our discussions. We occasionally mention that a random variable X is a p-variable even if it can take the value ∞, but it should be understood as saying that $X \wedge 1$ is a p-variable.

2. 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. Theorems 2.25 and 2.26 in Chapter 2 will show that e-variables are naturally represented by conditional expectations, whereas p-variables are naturally represented by conditional probabilities.

3. 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 $\mathbb{P}^* \in \mathcal{P}$. If $\mathcal{P}$ is a singleton $\{\mathbb{P}\}$, then we also say "an e-variable for $\mathbb{P}$".

4. 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 in $\mathcal{P}$, and we hope to reject H_0 when the data are generated by elements in $\mathcal{Q}$.

5. 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.

The value taken by the e-variable or p-variable on 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 the context.

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 8, 9, and 11, 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 6.

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 "V") 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 require both "e" in e-values and "p" in p-values to be italic, so it is possible to see "*e*-value" in some journal publications.

Next, we first discuss two natural examples of e-values, and many more examples for various hypotheses will be given throughout the book.

1.3 A first example: likelihood ratio

Suppose that we are testing a simple hypothesis $\mathbb{P}$ versus a simple hypothesis $\mathbb{Q}$, where $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$. For this setting, a natural e-variable is the likelihood ratio $E = (\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P})(X)$ where X is the observed data. It is straightforward to verify that $E \geq 0$ and it satisfies $\mathbb{E}^{\mathbb{P}}[E] = 1$.

We specialize in two simple settings two help the reader to keep concrete examples in mind. In what follows, $X_1, X_2, \dots$ are iid data points from $\mathbb{P}$ or $\mathbb{Q}$, and we denote by $S_n := \sum_{i=1}^n X_i$ the sum of the first n data points.

The most standard example is testing $\mathbb{P} = \mathrm{N}(0,1)$ against $\mathbb{Q} = \mathrm{N}(\mu, 1)$ for a known $\mu > 0$. Rigorously, $\mathbb{P}$ and $\mathbb{Q}$ are the (infinite) product measure on the data sample $X_1, X_2, \dots$ (which are iid normal), but we slightly abuse the notation here and later, which should be clear from the context. In this setting, a natural e-variable for a sample of size $n \in \mathbb{N}$ is given by the likelihood ratio

$$E_n = \exp\left(\mu S_n - n\mu^2/2\right). \tag{1.1}$$

It is straightforward to check that $\mathbb{E}^{\mathbb{P}}[E_n] = 1$, and hence the e-variable E_n is exact. 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 $X = S_n/\sqrt{n}$, which is $\mathrm{N}(0,1)$-distributed under the null hypothesis. The e-variable in (1.1) can be rewritten as

$$E_n = \exp\left(\delta X - \delta^2/2\right). \tag{1.2}$$

The Neyman-Pearson p-variable is computed as $P_n = \Phi(-S_n/\sqrt{n}) = \Phi(-X)$, 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 $\mathbb{P}$.

Note that E_n is an e-variable for any choice of μ, and hence the alternative hypothesis does not matter for the validity of E_n, and we can use it to test $\mathrm{N}(0,1)$ against $\mathrm{N}(\mu,1)$ for an unknown $\mu > 0$. The choice of δ in (1.2) affects the power of e-variable, as discussed in Chapters 2 and 3. We can also test the two-sided alternative hypothesis: $\mathrm{N}(\mu,1)$ for an unknown $\mu \neq 0$. In this case, a natural e-variable is

$$E_n = \frac{\exp\left(\delta X - \delta^2/2\right) + \exp\left(-\delta X - \delta^2/2\right)}{2}, \tag{1.3}$$

for a parameter $\delta > 0$ free to choose. The corresponding p-variable is given by

$$P_n = 2\Phi(-|X|). \tag{1.4}$$

Another simple example is testing $\mathbb{P} = \mathrm{Bernoulli}(p)$ against $\mathbb{Q} = \mathrm{Bernoulli}(q)$, 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

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

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, as introduced earlier.

1.4 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 explain how to build a natural e-value in this context.

Random variables $Z_1, \dots, Z_n$ 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 $\{1, \dots, n\}$.

Consider for now a single hypothesis $\mathcal{P}$ and let L_0 be the test statistic calculated from the original data. Let $L_1, \dots, L_B$ be B statistics that are constructed to be exchangeable together with L_0 under the null, and $r > 0$ 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. The limiting case of $r = 0$ yields $R_b = L_b - L_*$. Note that the L_i's are not assumed to be positive. If they were, we could choose $R_b = L_b/L_*$, or simply set $R_b = L_b$. Most importantly, $\{R_b\}_{b=0}^B$ are also exchangeable under the null. Now, define

$$E = (B+1)\frac{R_0}{\sum_{i=0}^B R_i}.$$

The fact that E is an e-variable is guaranteed by $\mathbb{E}^{\mathbb{P}}[R_0 | \sum_{b=0}^B R_i] = \sum_{b=0}^B R_b/(B+1)$ for $\mathbb{P} \in \mathcal{P}$ due to exchangeability. Contrasting this with the usual p-variable P defined by

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

we find that

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

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). 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 7–9, we will see that e-values enjoy several advantages over p-values in the contexts of multiple testing and constructing confidence intervals.

1.5 Some statistical advantages of e-values

Below, we list some situations when one may potentially prefer to use e-values over p-values or other methods for non-philosophical, purely statistical, reasons. As the reader will soon find out in Chapter 2, the reciprocal of an e-value is a p-value, so the p-values referred to below are those that are not obtained as reciprocals of e-values but are instead constructed using more classical techniques.

The advantages below require knowledge of some sub-areas of statistical testing and materials in this book to understand. The reader may skip this subsection (as well as the remaining ones in this chapter on history and literature) if they would like to start to get their hands on e-values immediately.

1. **Robustness to dependence (across data for a single hypothesis)**. To construct a single valid p-value, it is often assumed (for convenience of deriving limiting distributions) that the underlying observations are independent, and indeed their validity is often hurt if this assumption is violated. However, we can often construct an e-value quite easily in settings where the observations are dependent, and this is because we are requiring less of an e-value (just bounded in expectation) than of a p-value (knowledge of its whole distribution). As one example, suppose we observe non-negative data $X_1, \dots, X_n$ that have the same marginal distribution $\mathbb{P}$ which is assumed to have at least one moment. Suppose we wish to test whether $\mathbb{P}$ has a mean at most μ, that is $H_0 : \mathbb{E}^{\mathbb{P}}[X] \leq \mu$. Then, $(X_1 + \cdots + X_n)/(n\mu)$ is an e-value for any dependence structure of $X_1, \dots, X_n$, and does not require making any distributional or independence assumption.

2. **Robustness to dependence (across multiple hypotheses).** Our ability to combine multiple p-values relies heavily on dependence assumptions made between the p-values, and for each dependence assumption there is no single prototypical way to combine p-values. This is quite unlike the situation with e-values, where averages of arbitrarily dependent e-values are e-values, and products of independent e-values are e-values. This is useful, for example, in the field of risk management, analysis of risks with unknown or complicated dependence has recently been an active research topic. This robustness to dependence is exploited and explored in several chapters on multiple testing in our monograph, and in particular it gives rise to the e-BH procedure in Chapter 8, the e-BY procedure in Chapter 11, and valid methods for merging p-values in Chapter 9.

3. **Post-hoc inference.** When testing with p-values, the type-I error level must be chosen in advance of seeing the data. In contrast, Section 2.4 shows that that e-values satisfy a natural post-hoc notion of validity, and they are unique in doing so. This phenomenon can be extended to other settings involving decision making, where the decision task is determined after observing the data.

4. **Irregular (composite) models**. There are many composite null hypothesis testing problems for which we know of no direct way to construct a valid p-value even under low-dimensional asymptotics—this could happen because the model is singular or irregular and Wilks' theorem fails to hold and in such cases the validity of the bootstrap is also typically unknown. As discussed in Chapter 4, the universal inference methodology based on the *split* likelihood ratio statistic yields an e-value under no assumptions on the composite null, or on d and n. Examples of new settings in which one can now construct e-values under no regularity assumptions include mixtures (e.g., testing if data comes from a mixture of $\leq k$ versus $> k$ components), shape-constraints (e.g., log-concavity), dependence structures (e.g., multivariate total positivity), and several latent variable models.

5. **Avoiding high-dimensional asymptotics**. One of the most classical ways to compute p-values is to use the asymptotic distribution of the likelihood ratio test statistic, as given by Wilks' theorem. However, the correctness of Wilks' theorem is typically justified when the dimensionality d of the data remains fixed, and the sample size n tends to infinity. There are several results on the high-dimensional asymptotics of likelihood ratios, but the resulting p-value often relies on the practitioner making assumptions on the relative scaling of n and d. In contrast, the likelihood ratio for a point null hypothesis is a valid e-value in finite samples, meaning that its expectation equals one under the null regardless of d or n. The same holds for mixtures (over the alternative) of likelihood ratios as well. In fact, Chapter 5 will show that an optimal e-variable exists for any composite null.

6. **Robustness to misspecfication**. In genetics, it is not uncommon to encounter p-values with astronomically small values (like 10^{-20}), or sometimes point masses near the value one, even though the sample sizes may not intuitively support such extreme evidence. This is often (but not always) reflective of utilizing a model that is not perfectly specified. The validity of p-values is quite sensitive to model misspecification, because they utilize the entire (hypothesized) distribution of the test statistic. In contrast, e-values can be constructed without over-reliance on fine-grained tail information, and so they are typically more robust than p-values to misspecification (but less powerful under perfect specification). For example, instead of assuming that the data X is Gaussian to build a p-value, we may instead assume that it is symmetric about the origin (under no further moment constraints), in

which case $\exp(\lambda X - \lambda^2 X^2/2)$ is a valid e-value for any $\lambda \in \mathbb{R}$. See Sections 5.6 for more details and improvements of this e-value.

The aforementioned advantages hold for collecting and analyzing data of any fixed sample size n. However, some of the primary advantages of working with e-values stems from their flexibility with regards to extending experiments and analyzing data at stopping times. We mention a few such examples below. While they are all related, there are subtle differences amongst the advantages presented below.

7. **Accumulation of information and evidence**. Suppose the ultimate goal of a scientist is to either reject (with level α) or not reject a given hypothesis *in a single run*. Then, the most powerful method according to the Neyman–Pearson lemma is to reject when a certain p-value p is no larger than α. This testing procedure can also be achieved by a simple e-value $e := \mathbb{1}_{\{p \le \alpha\}}/\alpha$, using a threshold $1/\alpha$ (called an all-or-nothing e-value in Chapter 2); in fact, no other e-value would be more powerful in general. The situation becomes quite different when the hypothesis (if seen as promising by the first experiment) will be tested with future evidence and possibly by other scientists: the all-or-nothing e-value carries little information for the next studies, whereas a generic e-value can provide a continuum of evidence strength. Consider a setting where the first e-value is moderately large (e.g., $e_1 = 4$). A different agent may view this as promising but not conclusive, and choose to collect their own data and form their own e-value (say $e_2 = 6$). This by itself is also promising but not conclusive. However the two e-values can easily be merged, for example by simply multiplying the two e-values. This is valid if $\mathbb{E}[E_2|E_1 = e_1] \le 1$ under the null, which it would be in the case that fresh data was collected to form e_2). On the other hand, a p-value from the first experiment is difficult to merge with those from future studies, because the very existence of those studies may depend on the p-value obtained from the first experiment. Thus, the number of p-values is itself random, and this is a very insidious and subtle form of dependence that is often ignored in meta-analyses. Such sequentially-dependent p-values are difficult to merge, resulting either in the discarding of earlier evidence (where the later study ignores earlier evidence) or incorrect combinations (assuming independence of studies). Hence, for a *dynamic flow* of experiments, common in modern sciences, it appears beneficial to track and report e-values instead of p-values.

8. **Sequential inference**. For a simple null and alternative, the likelihood ratio process (of the alternative to the null) is a nonnegative martingale (under the null) with initial value one. The optional stopping theorem implies that at *any* stopping time, the stopped process is an e-value. We thus say that the likelihood ratio process is an example of an *e-process* (see Section 1.2 for a formal definition). Moving beyond parametric settings, open-ended sequential inference is enabled by designing nonparametric *nonnegative supermartingales* (called *test supermartingales* later) that immediately yield e-values at the stopping time. For point nulls, e-processes can always be dominated by test martingales (meaning they can be increased to a martingale without threat to validity). But for more general composite nulls, there is a big and important difference between e-processes, supermartingales and martingales, and it is (powerful) e-processes that always exist for any sequentially testable problem. In summary, e-values arise very naturally in sequential inference as stopped e-processes, and, in the other direction, e-values are the building blocks of test supermartingales and e-processes.

9. **Analysis of data collected via an 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 may often be able to calculate an e-value. To explain this issue, consider a simple problem of testing fairness of a coin from independent and identically distributed observations $X_1, X_2, \dots$ where $X_i = 1$ indicates a head and $X_i = 0$ indicates a tail. The null hypothesis $\mathbb{P}$ is Bernoulli(1/2) and the alternative hypothesis $\mathbb{Q}$ is Bernoulli(θ) for a known $\theta > 1/2$. Let $N_n = \sum_{i=1}^n X_i$. Suppose that the scientist is presented with the dataset $(1, 1, 0, 1, 1, 1, 1, 1)$ without explaining how the data are collected. 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 $\mathbb{P}(N_8 \ge 7) = 0.035$, considered as significant in some areas of science. In case (b), the standard p-value is $\mathbb{P}$(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. As explained in Section 1.3, a natural e-value for a sample $X_1, \dots, X_n$ of size n is the likelihood ratio $E_n = \theta^{N_n}(1-\theta)^{n-N_n}/2^n$ where $N_n = \sum_{i=1}^n X_i$. This is a valid e-value regardless of how the data collecting procedure is designed, due to the fact that $(E_n)_{n\geq 1}$ is an e-process.

10. **Indefinite sample size and optional continuation.** Consider an alteration of the previous example, where now the scientist plays the role of both the experiment designer and the analyst. It is very natural that the scientist would like to, after seeing data of size n, potentially request more data to be collected. This may happen more than once. In such a setting, it is very hard to calculate a traditional p-value (or worse, *p-hacking* occurs by ignoring the effect of sequential data collection), because the data-dependent reason for wanting more data is vague and unspecified in the beginning. However, when evidence is measured using an e-process E, the scientist can continue or terminate the experiment at any time τ and for any data-dependent reason. In any such setting, E_τ remains an e-variable, allowing for valid inference. Such e-processes are partially motivated by and intimately linked to game-theoretic statistics, and in particular through the powerful methodology of testing by betting. This will be explained in Chapter 6.

Thus, there exist many settings when one can, and should, use e-values to quantify evidence against a null. Of course, there of course remain innumerable situations in which p-values are perfectly reasonable choices.

To summarize, there are several reasons to work with e-values: they arise naturally in sequential settings, we know how to construct e-values in settings where we do not know how to construct p-values, and e-values can be more robust to misspecification or uncertain asymptotics in high-dimensional settings. Of course, there are also many reasons *not* to work with e-values: in particular, p-values may yield more "powerful" (single/multiple) tests when the underlying modeling assumptions are true and we wish to analyze a fixed sample size dataset collected in a known manner, and wish to make a clear accept-reject decision at the end. Thus 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.

Overall, we view the current monograph as providing the building blocks of *a theory of evidence* (where measuring evidence using e-values provides many benefits). 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, and our theory of evidence does not aim to output a 0 or 1 decision, but simply to measure the current amount of evidence against a null hypothesis.

1.6 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 section, 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 $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$, 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 4 and 5. 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 (like 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 understanding 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 in these contexts, but also because there is much basic work to be done in the sequential context before a book can be written about it, but needless to say, such a book should be written in due course. Nevertheless, Chapter 6 is dedicated to the sequential setting, giving readers a glimpse of the broader framework.

1.7 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 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). Gammerman, 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 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 about ten 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 confidence sequences, extending and generalizing Robbins' extensive work on the topic.

- 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.

1.9 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.

On the other hand, 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 statistically 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.

1.10 Road map

The rest of this book is organized into three parts: Fundamental Concepts, Core Ideas, and Advanced Topics.

The first part includes three chapters (including the current one) that introduce the basic concepts. Chapter 2 presents key concepts such as Markov’s inequality, calibrators, e-power, a certain duality between e-variables and p-variables, and randomized tests. Chapter 3 proves that the log-optimal e-variable for a simple null against a simple alternative is their likelihood ratio, and discusses extensions to composite alternatives using the mixture and plug-in methods.

The second part includes five chapters. Chapter 4 introduces the general method of universal inference for constructing e-variables for general composite nulls and alternatives. Chapter 5 presents 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. Chapter 6 presents the key concepts in sequential anytime-valid inference (SAVI) like e-processes, testing by betting and Ville's inequality. Chapter 7 discusses merging e-values via averages (arbitrary dependence) or products (independence or "sequential" dependence). Chapter 8 presents the notion of a compound e-value 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. It also derives the log-optimal simple separate compound e-variable as a ratio of mixture likelihoods.

The third part presents five advanced topics. Chapter 9 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, and presents an e-weighted BH procedure. Chapter 10 derives methods for combining dependent p-values via e-values, and also ways to improve on such combination rules by using random and exchangeable improvements of Markov's inequality. Chapter 11 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 12 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 13 studies the connection between e-values and risk measures, and offer some nonparametric methods for testing the forecast of general statistical functions (such as the mean, the variance, or a quantile) using e-values.

Chapter 2

Fundamental properties of e-values

In this chapter, we study fundamental issues on e-values, in particular, on their connection to p-values, validity, and tests. For this reason, the alternative hypothesis $\mathcal{Q}$ does not appear and plays no role, except for the concepts of nontrivial tests and nontrivial e-values in Section 2.3.

2.1 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

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, αE 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

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

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 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 2.6 and Chapters 4 and 12. A sequential generalization, called Ville's inequality, is presented in Chapter 6.

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.[1] To understand how these two tests differ, we consider the two-sided normal test in Section 1.3, 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 x taken by the statistic X. Figure 2.1 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

[1] However, if one desires, one can always make the sets $\{P \leq \alpha\}$ and $\{E \geq 1/\alpha\}$ identical, by choosing $E = \mathbb{1}_{\{P \leq \alpha\}}/\alpha$. One can easily verify that E is indeed an e-variable. Section 2.3 has more details on such e-variables.

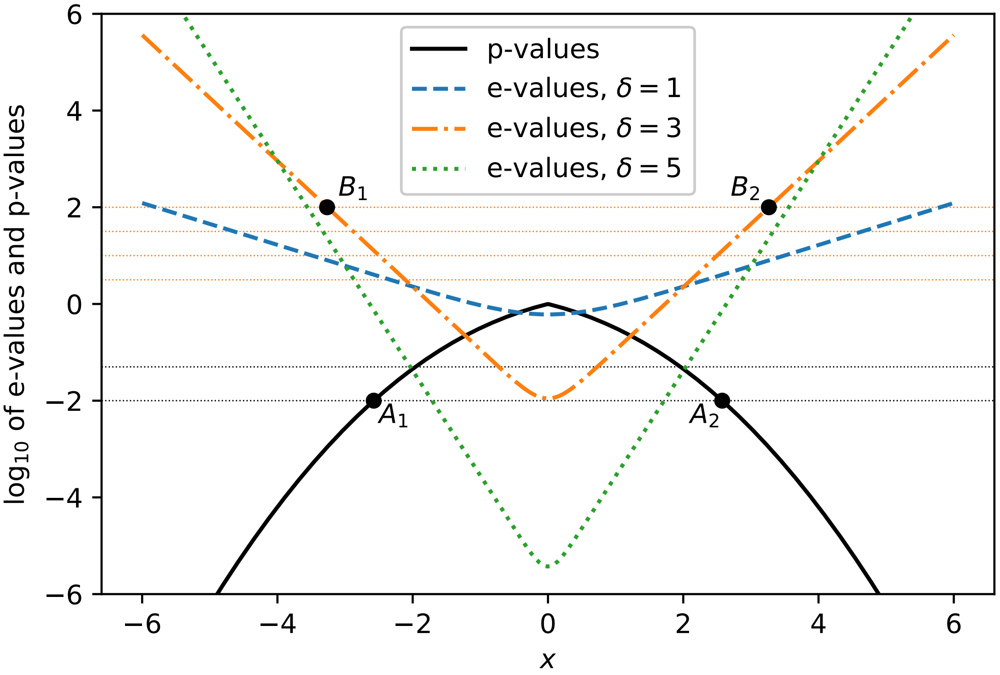

Figure 2.1: A comparison of e-values and p-values for the two-sided normal test in Section 1.3. The curves represent the e-values or p-values as a function of the test statistic X. The top four horizontal dotted lines correspond to 10^β with $\beta \in \{2, 1.5, 1, 0.5\}$ for e-values. The bottom two 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 formally explained in Section 2.10).

$x = A_2$ 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 2.1. 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 2.10 for more discussions on this point.

2.2 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 2.2: Calibrators

(i) A *p-to-e calibrator* is a decreasing function $f : [0, \infty) \to [0, \infty]$ 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 $g : [0, \infty] \to [0, \infty)$ 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 2.3 and 2.4.

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.

In Definition 2.2, the property of calibrators needs to hold for any hypothesis. Nevertheless, it suffices to consider one atomless probability $\mathbb{P}$ 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 2.3

The function $f : [0, \infty] \to [0,1]$ 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 f is another e-to-p calibrator. It suffices to check that f is dominated by $t \mapsto \min(1, 1/t)$. Suppose $f(t) < \min(1, 1/t)$ for some $t \in [0, \infty]$. Fix a simple atomless probability $\mathbb{P}$. Consider two cases:

(a) If $f(t) < \min(1, 1/t) = 1/t$ 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 $f(E)$ is $f(t) < 1/t$ with probability $1/t$, whereas it would have satisfied $\mathbb{P}(f(E) \le f(t)) \le f(t) < 1/t$ had it been a p-variable.

(b) If $f(t) < \min(1, 1/t) = 1$ for some $t \in [0,1]$, fix such t and consider the e-variable $E = t$. Then $f(E) = f(t) < 1$, 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 12.

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

Proposition 2.4

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

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 $\int_0^1 f(p) \mathrm{d}p \le 1$, P is a p-variable for an atomless probability $\mathbb{P}$, and P' is uniformly distributed on $[0,1]$. Since $\mathbb{P}(P \le x) \le x \vee 1 = \mathbb{P}(P' \le x)$ for all $x \ge 0$ and f is decreasing, we have

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

for all $x \geq 0$, which implies

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

Both necessity and sufficiency in the second statement of Proposition 2.4 are straightforward.

Simple examples of admissible p-to-e 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),

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

and special simple forms

$$f(p) = p^{-1/2} - 1 \tag{2.3}$$

and

$$f(p) = -\log p. \tag{2.4}$$

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 $p' = 1/9$.

A possible interpretation of results on the 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 2.3. 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; this follows from Proposition 2.4. 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 what follows, we omit "p-to-e" when mentioning calibrators. Although all calibrators in (2.1)–(2.4) are admissible, some are more useful. In hypothesis testing, very small p-values and very large e-values are interesting. Calibrators in (2.3) and (2.4) 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.

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 10.

Proposition 2.5

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

Proof.

Denote by $g : x \mapsto af(ax)$. We need to check a few properties in Proposition 2.4. First, it follows from $f(p) = 0$ for $p > 1$ that $g(p) = af(ap) = 0$ for $p > 1$, and further

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

Upper semicontinuity and $g(0) = \infty$ follow directly from those properties of f.

The next result states that, any e-variable E for an atomless probability $\mathbb{P}$ 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 2.6

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.9 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 2.4 and the fact that the right quantile function is upper semicontinuous.

2.3 Nontrivial tests, e-values and e-power

While it may appear from the previous sections that there may be some loss in translating between e-values and tests, this is not entirely true. 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 2.7: 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 2.8: Duality of tests and e-variables

There is a one-to-one correspondence between 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 $\mathbb{1}_{\{P\leq\alpha\}}/\alpha$.

Therefore, all binary tests based on p-values — routinely used in the sciences — can be replicated using e-values.

Observe that 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, then it is testable with e-values. We expand below.

Definition 2.9: Power, nontrivial tests and e-values

The power of a test ϕ against an alternative $\mathbb{Q}$ is simply $\mathbb{E}^{\mathbb{Q}}[\phi]$. When testing a null hypothesis $\mathcal{P}$, a level-α test ϕ is called nontrivial against $\mathcal{Q}$ if $\mathbb{E}^{\mathbb{Q}}[\phi] > \alpha$ for every $\mathbb{Q} \in \mathcal{Q}$. An e-variable E is called nontrivial against $\mathcal{Q}$ if $\mathbb{E}^{\mathbb{Q}}[E] > 1$ for every $\mathbb{Q} \in \mathcal{Q}$.

If ϕ is binary, the power of ϕ against $\mathbb{Q}$ is the probability of rejecting the null hypothesis under the alternative $\mathbb{Q}$. We can now present a simple yet important observation connecting tests and e-values.

Proposition 2.10

For testing $\mathcal{P}$ against $\mathcal{Q}$, there exists a nontrivial level-α test if and only if there exists a nontrivial e-variable. Also, there exists a nontrivial binary test if and only if there exists a nontrivial all-or-nothing e-variable.

The proof is straightforward: Given a nontrivial (binary) ϕ, we can define the (all-or-nothing) e-variable $E = \phi/\alpha$. Given a nontrivial (all-or-nothing) E, we can define the (binary) test $\phi = \alpha E$.

Given this high-level equivalence between tests and e-values, 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 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 2.11: 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 is said to have *positive e-power* (against $\mathbb{Q}$) if $\mathbb{E}^{\mathbb{Q}}[\log E] > 0$.

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 $-\infty$. However, whenever there is a nontrivial level-α test, we can design e-variables with positive e-power.

Jensen's inequality applied to the logarithm function implies that if E has positive e-power, then E is a nontrivial e-variable, but the reverse implication is not true in general. However, from any nontrivial 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.

Proposition 2.12

An e-variable with positive e-power exists if and only if a nontrivial e-variable exists. In fact, for any random variable $X \geq 0$ and any $\mathbb{Q} \in \mathcal{M}_1$,

$$\mathbb{E}^{\mathbb{Q}}[X] > 1 \iff \mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda + \lambda X)] > 0 \text{ for some } \lambda \in [0, 1]. \tag{2.5}$$

Proof.

The backward direction in (2.5) is a simple application of Jensen's inequality, which yields

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

To show the forward direction in (2.5), it suffices to verify $\mathbb{E}^{\mathbb{Q}}[\log(1-\lambda+\lambda X)] > 1$ for $\lambda > 0$ small enough. Note that $\mathbb{E}^{\mathbb{Q}}[X] > 1$ implies $\mathbb{E}^{\mathbb{Q}}[X \wedge K] > 1$ for some $K \geq 1$. We denote by $Y = X \wedge K$

and let $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$ for $x \in \mathbb{R}$. Since $\mathbb{E}^{\mathbb{Q}}[(Y-1)_+] - \mathbb{E}^{\mathbb{Q}}[(Y-1)_-] = \mathbb{E}^{\mathbb{Q}}[Y-1] > 0$, there exists some $\varepsilon \in (0,1)$ such that

$$\frac{1}{1+\varepsilon}\mathbb{E}^{\mathbb{Q}}[(Y-1)_+] - \frac{1}{1-\varepsilon}\mathbb{E}^{\mathbb{Q}}[(Y-1)_-] > 0.$$

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 $\lambda(Y-1) \in (-\varepsilon, \varepsilon)$, we have

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}}[\log(1-\lambda+\lambda X)] &\geq \mathbb{E}^{\mathbb{Q}}[\log(1+\lambda(Y-1))] \\ &\geq \frac{1}{1+\varepsilon}\mathbb{E}^{\mathbb{Q}}[\lambda(Y-1)_+] - \frac{1}{1-\varepsilon}\mathbb{E}^{\mathbb{Q}}[\lambda(Y-1)_-] > 0, \end{aligned}$$

thus showing the desired inequality.

Remark 2.13. As we can see from the proof of Proposition 2.12, the existence of a nontrivial e-variable further guarantees the existence of a *bounded* e-variable with positive e-power.

We end with a brief remark about the role of randomization. A non-binary test ϕ is usually provided with the interpretation that it outputs a rejection with probability ϕ. In other words, one can convert a non-binary test ϕ to a binary test ϕ' as $\phi' = \mathbb{1}_{\{U \leq \phi\}}$, where U is an independent uniform random variable on $[0,1]$. This is closely related to the randomized Markov's inequality discussed in Section 2.6.

2.4 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 generalization of tests to random $\hat{\alpha}$ is as follows.

Definition 2.14: 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 $\mathbb{E}^{\mathbb{P}}\left[\phi_{\hat{\alpha}}/\hat{\alpha}\right] \leq 1$. 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 $\mathbb{E}^{\mathbb{P}}\left[\mathbb{1}_{\{P \leq \hat{\alpha}\}}/\hat{\alpha}\right] \leq 1$. Equivalently, we require that

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

Clearly, in Definition 2.14, 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.

We will see that testing with e-variables yields post-hoc family of tests. Moreover, any post-hoc p-variable is intimately linked to an e-variable.

Proposition 2.15

If E is an e-variable for $\mathcal{P}$, the collection $(\mathbb{1}_{\{E\geq 1/\alpha\}})_{\alpha\in(0,1)}$ is a post-hoc valid family of tests. Indeed, every post-hoc valid family of tests must take the form $(\mathbb{1}_{\{E_\alpha\geq 1/\alpha\}})_{\alpha\in(0,1)}$, where E_α is an all-or-nothing e-variable. Finally, 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.

The proof of the first claim mimics Markov's inequality. Recalling that $\mathbb{1}_{\{x\geq 1\}} \leq x$ for $x \geq 0$, we have that for any $\hat{\alpha}$ taking values in $(0,1)$,

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

The second claim follows from the duality of tests and e-variables (Fact 2.8).

Next, we show the equivalence between statements (i), (ii) and (iii) on post-hoc p-variables. To see (i)⇒(ii), first note that $\mathbb{P}(P = 0)$ 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 $\hat{\alpha} = P\mathbb{1}_{\{P<1\}} + \beta\mathbb{1}_{\{P\geq 1\}}$ for some $\beta \in (0,1)$. We have that $\hat{\alpha}$ is $(0,1)$-valued and $\{P < 1\} = \{P \leq \hat{\alpha}\}$. Therefore,

$$\mathbb{E}^{\mathbb{P}}[E] = \mathbb{E}^{\mathbb{P}}\left[\frac{1}{P}\mathbb{1}_{\{P<1\}}\right] = \mathbb{E}^{\mathbb{P}}\left[\frac{1}{P\mathbb{1}_{\{P<1\}}}\mathbb{1}_{\{P\leq\hat{\alpha}\}}\right] = \mathbb{E}^{\mathbb{P}}\left[\frac{1}{\hat{\alpha}}\mathbb{1}_{\{P\leq\hat{\alpha}\}}\right] \leq 1 \quad \text{for all } \mathbb{P} \in \mathcal{P},$$

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)⇒(iii) is immediate. Finally, the direction (iii)⇒(i) follows from

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

for all $\mathbb{P} \in \mathcal{P}$, where the penultimate inequality is identical to the first claim.

From Proposition 2.15, 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).

There is a small variation of the result in Proposition 2.15, 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

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

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 2.15.

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 2.14.

2.5 A duality between e-values and p-values

We next 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 smaller value of X represents evidence against the null hypothesis. Simple examples are (1.1) and (1.5) in Section 1.3, 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 2.16

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,

$$P \geq S(X|\mathbb{P}) \quad \text{and} \quad E \leq 1/S(X|\mathbb{P}) \quad \mathbb{P}\text{-almost surely.}$$

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

$$\mathbb{P}(T(X') \leq T(X) \mid X) \geq \mathbb{P}(X' \geq X \mid X) = S(X|\mathbb{P}) \quad \mathbb{P}\text{-almost surely.}$$

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 2.3, and then applying the result in the first part.

In fact, in Proposition 2.16, $S(X|\mathbb{P})$ 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 2.17: Duality between e-values and p-values

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,

$$\sup_{E \in \mathcal{E}^X} E = \inf_{\mathbb{P} \in \mathcal{P}} \frac{1}{S(X|\mathbb{P})} = \frac{1}{\inf_{P \in \mathcal{U}^X} P} \quad \mathbb{P}\text{-almost surely for every } \mathbb{P} \in \mathcal{P}.$$

In particular, if $\mathcal{P} = \{\mathbb{P}\}$, then $\sup_{E \in \mathcal{E}^X} E = 1/S(X|\mathbb{P})$.

Proof.

Proposition 2.16 and the fact that $1/E$ for an e-variable E is a p-variable guarantee

$$\sup_{E \in \mathcal{E}^X} E \leq \sup_{P \in \mathcal{U}^X} \frac{1}{P} \leq \inf_{\mathbb{P} \in \mathcal{P}} \frac{1}{S(X|\mathbb{P})} \quad \mathbb{P}\text{-almost surely for every } \mathbb{P} \in \mathcal{P}. \tag{2.6}$$

Below we will show the equalities. For $s \in \mathbb{R}$, let

$$E_s = \frac{1}{\sup_{\mathbb{P}\in\mathcal{P}} S(s|\mathbb{P})} \mathbb{1}_{\{X \geq s\}} \quad \text{with the convention } 0/0 = 1,$$

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 (2.6) hold.

The quantity $\sup_{\mathbb{P}\in\mathcal{P}} S(X|\mathbb{P})$ in Theorem 2.17 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 2.17 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 2.17, which is the maximal set).

The next corollary strengthens the Markov inequality for e-variables, which is immediate from Theorem 2.17.

Corollary 2.18

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 2.18 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 12.2 has a more general treatment of this property.

2.6 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 2.19: Randomized Markov inequality

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

$$\mathbb{P}(Y \geq U/\alpha) = \mathbb{E}^{\mathbb{P}}[\min(\alpha Y, 1)] \leq \alpha\mathbb{E}^{\mathbb{P}}[Y]. \tag{2.7}$$

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 X, 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 (2.7) 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 (2.7) holds with equality for any $\alpha \le 1/C$. 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 2.19, which offers additional interpretations of the relation between p-variables and e-variables and randomization.

Corollary 2.20

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 $U \overset{\mathrm{d}}{\sim} \mathrm{Unif}(0,1)$ 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* that dominates $e \mapsto (1/e) \wedge 1$, the latter only being admissible among non-randomized e-to-p calibrators (Proposition 2.3).

A more general treatment of combining a p-value and an e-value is presented in Section 9.3.

2.7 Axiomatic justification of the e-power

In this section we justify the definition of e-power in Definition 2.11 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 $\Lambda^{\mathbb{Q}} : \mathcal{X} \to \mathbb{R}$, without specifying that e-variables need to satisfy $\mathbb{E}^{\mathbb{P}}[E] \le 1$ 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 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 2.21: Homogeneity

If $\Lambda^{\mathbb{Q}}(E_1) \ge \Lambda^{\mathbb{Q}}(E_2)$, then $\Lambda^{\mathbb{Q}}(e_0 E_1) \ge \Lambda^{\mathbb{Q}}(e_0 E_2)$ 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 has asymptotic power 1.

Axiom 2.22: 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 $\alpha \in (0,1)$, as $n \to \infty$,

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

It turns out that Axioms 2.21 and 2.22 are sufficient to jointly characterize the e-power, which has the form $\mathbb{E}^{\mathbb{Q}}[\log E]$ as we saw in Definition 2.11.

Theorem 2.23

Given an atomless probability measure $\mathbb{Q} \in \mathcal{M}_1$, a mapping $\Lambda^{\mathbb{Q}} : \mathcal{X} \to \mathbb{R}$ satisfies Axioms 2.21 and 2.22 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 $E \mapsto \mathbb{E}^{\mathbb{Q}}[\log E]$ satisfies the two axioms; Axiom 2.22 can be checked with the strong law of large numbers. Clearly, a strictly increasing transform does not matter for the two axioms. Hence, $E \mapsto f(\mathbb{E}^{\mathbb{Q}}[\log E])$ 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

$$\Lambda^{\mathbb{Q}}(E) > \Lambda^{\mathbb{Q}}(E') \text{ and } \mathbb{E}^{\mathbb{Q}}[\log E] < \mathbb{E}^{\mathbb{Q}}[\log E']. \tag{2.8}$$

Using Axiom 2.21, we can find some $c > 0$ such that

$$\Lambda^{\mathbb{Q}}(cE) > \Lambda^{\mathbb{Q}}(cE') \text{ and } \mathbb{E}^{\mathbb{Q}}[\log(cE)] < 0 < \mathbb{E}^{\mathbb{Q}}[\log(cE')]. \tag{2.9}$$

Using Axiom 2.22, 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 (2.9) together imply $\log \prod_{t=1}^{n} E_t \to -\infty$ and $\log \prod_{t=1}^{n} E'_t \to \infty$ in $\mathbb{Q}$. This violates Axiom 2.22. Hence, (2.8) does not hold. This implies

$$\Lambda^{\mathbb{Q}}(E) > \Lambda^{\mathbb{Q}}(E') \implies \mathbb{E}^{\mathbb{Q}}[\log E] \geq \mathbb{E}^{\mathbb{Q}}[\log E']. \tag{2.10}$$

Equivalently,

$$\Lambda^{\mathbb{Q}}(E) \leq \Lambda^{\mathbb{Q}}(E') \impliedby \mathbb{E}^{\mathbb{Q}}[\log E] < \mathbb{E}^{\mathbb{Q}}[\log E']. \tag{2.11}$$

Using (2.11), we know that $c \mapsto \Lambda^{\mathbb{Q}}(cE)$ 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

$$\Lambda^{\mathbb{Q}}(E) > \Lambda^{\mathbb{Q}}(E') \text{ and } \mathbb{E}^{\mathbb{Q}}[\log E] = \mathbb{E}^{\mathbb{Q}}[\log E']. \tag{2.12}$$

Using Axiom 2.21, we have from (2.12), for any $c > 0$, $\Lambda^{\mathbb{Q}}(cE) > \Lambda^{\mathbb{Q}}(cE')$. Note that for any $\lambda < 1$, we have, from (2.11), $\Lambda^{\mathbb{Q}}(\lambda cE) \leq \Lambda^{\mathbb{Q}}(cE')$, because $\mathbb{E}^{\mathbb{Q}}[\log(\lambda cE)] < \mathbb{E}^{\mathbb{Q}}[\log(cE')]$. Hence, $\lambda \mapsto \Lambda^{\mathbb{Q}}(\lambda cE)$ is discontinuous at $\lambda = 1$ for all $c > 0$. This implies that $c \mapsto \Lambda^{\mathbb{Q}}(cE)$ is discontinuous everywhere, a contradiction to the fact that this mapping is monotone and hence has bounded variation. Therefore, (2.12) does not hold, and by (2.10), we have

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

This, together with (2.10), shows that $\Lambda^{\mathbb{Q}}$ and $X \mapsto \mathbb{E}^{\mathbb{Q}}[\log X]$ generate the same ordinal relation, and hence the representation holds.

Theorem 2.23 implies that if Axioms 2.21 and 2.22 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 $\mathbb{E}^{\mathbb{Q}}[\log E]$, which generates the same order as using $\mathbb{E}^{\mathbb{Q}}[\log E]$ directly. Therefore, this result theoretically justifies the use of $\mathbb{E}^{\mathbb{Q}}[\log E]$ as e-power in Definition 2.11. It also shows what the e-power essentially guarantees: It is the most suitable when considering the product e-value with asymptotic consistency.

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

In what follows, we let

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

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 2.24

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).

Other properties of the e-power, including its maximizers, will be presented in the next chapters, and we will directly use $\mathbb{E}^{\mathbb{Q}}[\log E]$ instead of $\Lambda^{\mathbb{Q}}(E)$ in most places.

2.8 E-variables as conditional expectations

In this section, we present two representation results on p-variables and e-variables, which helps to justify "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. (If the reader is familiar with measure theory, this means that Y is measurable with respect to the σ-algebra of X.)

Theorem 2.25: 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

$$E \leq \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mid X\right] \qquad \mathbb{P}\text{-almost surely.} \tag{2.13}$$

If all $\mathbb{P} \in \mathcal{P}$ are equivalent to $\mathbb{L}$, then every e-variable built on X is dominated by (meaning $\leq$)

$$\inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}_{\mathbb{P}}}{\mathrm{d}\mathbb{P}} \mid X\right], \quad \text{which is an e-variable for } \mathcal{P},$$

for some class $(\mathbb{Q}_{\mathbb{P}})_{\mathbb{P}\in\mathcal{P}}$ in $\mathcal{M}_1$ absolutely continuous with respect to $\mathbb{L}$.

Proof.

The direction (ii)⇒(i) is straightforward, because for all $\mathbb{P} \in \mathcal{P}$, the tower property of conditional expectation gives

$$\mathbb{E}^{\mathbb{P}}[E] \leq \mathbb{E}^{\mathbb{P}}\left[\mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mid X\right]\right] = \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = 1.$$

Next we show (i)⇒(ii). For each $\mathbb{P} \in \mathcal{P}$, let $\mathbb{Q}$ be a probability measure defined by $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} = E/\mathbb{E}^{\mathbb{P}}[E]$ 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 $\mathbb{E}^{\mathbb{P}}[E] > 0$,

$$\mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mid X\right] = \mathbb{E}^{\mathbb{P}}\left[\frac{E}{\mathbb{E}^{\mathbb{P}}[E]} \mid X\right] = \frac{E}{\mathbb{E}^{\mathbb{P}}[E]} \geq E \qquad \mathbb{P}\text{-almost surely.}$$

For $\mathbb{P}$ with $\mathbb{E}^{\mathbb{P}}[E] = 0$, we have $E = 0$ $\mathbb{P}$-almost surely, which trivially satisfies (2.13).

The last statement follows by taking an infimum over $\mathbb{P}$ for (2.13).

Theorem 2.25 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.* The next result shows that *p-variables are conditional probabilities on the data.*

Theorem 2.26: 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 $T : \mathbb{R}^n \to \mathbb{R}$ such that for every $\mathbb{P} \in \mathcal{P}$,

$$P \geq \mathbb{P}\left(T(X') \leq T(X) \mid X\right) \qquad \mathbb{P}\text{-almost surely,} \tag{2.14}$$

where X' is an iid copy of X under $\mathbb{P}$.

If all $\mathbb{P} \in \mathcal{P}$ are equivalent to $\mathbb{L}$, then every p-variable built on X is dominated by (meaning $\geq$)

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\left(T(X'_{\mathbb{P}}) \leq T(X) \mid X\right) \quad \text{which is a p-variable for } \mathcal{P},$$

for some measurable function $T : \mathbb{R}^n \to \mathbb{R}$, where $X'_{\mathbb{P}}$ is an iid copy of X under $\mathbb{P}$.

Proof.

We first show (ii)⇒(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)⇒(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 statement follows by taking a supremum over $\mathbb{P}$ for (2.14).

2.9 Asymptotic e-variables

Many commonly used p-values are only asymptotically valid, meaning that they satisfy the condition $\mathbb{P}(P \leq \alpha) \leq \alpha + o(1)$, weaker than $\mathbb{P}(P \leq \alpha) \leq \alpha$, where the p-value P is calculated using n data points, and the $o(1)$ term vanishes in probability when n goes to infinity.

Similarly, we can define asymptotic e-values. In some settings, it is impossible or difficult to construct a nondegenerate e-variable from a finite sample, but one can still define an *asymptotic* e-variable that grows to infinity under any alternative.

Definition 2.27

Given a sequence of nonnegative extended random variables $(E_n)_{n \in \mathbb{N}}$, we say its components are asymptotic e-variables for $\mathcal{P}$ if $\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}[E_n] \leq 1$ as $n \to \infty$, and they are pointwise asymptotic e-variables for $\mathcal{P}$ if $\limsup_{n\to\infty} \mathbb{E}^{\mathbb{P}}[E_n] \leq 1$ for all $\mathbb{P} \in \mathcal{P}$.

We would also say “asymptotic e-values" in non-mathematical statements, and they should be interpreted as asymptotic e-variables or their realizations.

As one example, let $X_1, \dots, X_n$ be drawn iid from $\mathbb{P}$. We assume that $\sigma^2 = \mathrm{Var}(X_i) \in (0, \infty)$ is unknown, and we seek to test $H : \mu = 0$, where $\mu = \mathbb{E}^{\mathbb{P}}[X_i]$. Define $S_n = \sum_{i=1}^n X_i$, $V_n^2 = \sum_{i=1}^n X_i^2$, and for fixed $\lambda \in \mathbb{R}$:

$$E = E_n = \exp\left(\lambda \frac{S_n}{V_n} - \frac{\lambda^2}{2}\right). \tag{2.15}$$

Proposition 2.28

The random variable $E = E_n$ is a pointwise asymptotic e-variable for

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[X] = 0;\ \mathrm{Var}(X) \in (0, \infty)\}.$$

Proof.

First, note that by the central limit theorem, the law of large numbers and Slutsky's theorem, S_n/V_n converges in distribution to $\mathrm{N}(0,1)$ as $n \to \infty$. We may directly verify that,

$$\mathbb{E}^{\mathbb{P}}\left[\exp\left(\lambda Z - \frac{t^2}{2}\right)\right] = 1,$$

for $Z \sim \mathrm{N}(0,1)$. Next, we note that by the aforementioned convergence in distribution, it holds that $(S_n/V_n)_n$ is $O_{\mathbb{P}}(1)$ (i.e., stochastically bounded). Thus, 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 so we can conclude.

The above conclusion extends to a broader class of asymptotic e-values obtained by mixing in (2.15) 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,

$$\int \exp\left(\lambda \frac{S_n}{V_n} - \frac{\lambda^2}{2}\right) \mathrm{d}G(\lambda),$$

is a pointwise asymptotic e-variable.

One central purpose of defining asymptotic e-variables is to handle cases where one does not know how to construct a non-asymptotic e-variable. Indeed, for the above $\mathcal{P}$, we do not know of non-asymptotic e-variables that are not degenerate (like constants). Had a bound on the variance been known, such non-asymptotic e-variables do exist.

Example 2.29: Testing the mean with bounded variance

Given some $\sigma > 0$, consider the null hypothesis

$$\mathcal{P}_\sigma = \left\{\mathbb{P} \in \mathcal{M}_1 : \mathbb{E}^{\mathbb{P}}[X] = 0;\ \mathrm{var}^{\mathbb{P}}(X) \le \sigma^2\right\}.$$

Then X^2/σ^2 is an e-variable for $\mathcal{P}_\sigma$, 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 $x \ge 0$ 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.

We leave the proof of the facts in Example 2.29 for the reader.

2.10 Informal interpretation of thresholds for 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 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 is 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. We summarize this rule of thumb in Table 2.1. Our rough recommendation is to use $e > 4$ in place of $p < 0.05$ and $e > 10$ in place of $p < 0.01$, but one should keep in mind that these choices are quite arbitrary since p-values and e-values are do not have a one-to-one correspondence with each other.

e-value	level of evidence	Shafer's p-value
$0 \le e < 1$	null hypothesis is supported	$0.25 < p \le 1$
$1 < e < 3.16$	no more than a bare mention	$0.0577 < p < 0.25$
$3.16 < e < 10$	substantial	$8.3\times10^{-3} < p < 0.0577$
$10 < e < 31.6$	strong	$9.4\times10^{-4} < p < 8.3\times10^{-3}$
$31.6 < e < 100$	very strong	$9.8\times10^{-5} < p < 9.4\times10^{-4}$
$100 < e$	decisive	$0 \le p < 9.8\times10^{-5}$

Table 2.1: 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 $e = p^{-1/2} - 1$. The boundary values can be put in either of the two adjacent categories.

Bibliographical note

This chapter largely resembles results and concepts in various important papers on e-values, but also it contains many results that are newly obtained 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 2.2 dates back, at least, to Vovk [1993]; the admissibility results are in Vovk and Wang [2021]. Section 2.3 contains many observations from Zhang et al. [2024], with e-power formally defined in Vovk and Wang [2024b]. Section 2.4 combines results that are explicitly or implicitly found in Wang and Ramdas [2022], Xu et al. [2024], Grünwald [2024], Koning [2023a], in the context of post-hoc α for testing or controlling the false discovery rate or false coverage rate. Section 2.6 is based on Ignatiadis et al. [2024b] and Ramdas and Manole [2024]. Section 2.9 is based on Ignatiadis et al. [2024a] and Wang and Ramdas [2023a]. Section 2.10 can be found in [Jeffreys, 1961, Appendix B], also recapped in Kass and Raftery [1995].

Chapter 3

Log-optimality for a simple null

This section deals primarily with simple null hypothesis $\mathcal{P} = \{\mathbb{P}\}$ and simple alternatives $\mathcal{Q} = \{\mathbb{Q}\}$, but some definitions and results also deal with composite nulls or alternatives. We will assume that $\mathbb{Q} \ll \mathbb{P}$, and recall that 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 $(\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P})(x) = q(x)/p(x)$.

3.1 Admissible e-variables for $\mathbb{P}$ are likelihood ratios

We begin with a straightforward but important observation: For simple nulls, every admissible e-variable can be written as a likelihood ratio. We often omit "for $\mathbb{P}$" when mentioning e-variables in this chapter.

An e-variable E is said to be *admissible* if $E' \geq E$ implies that $E' = E$ for any other e-variable E' (where the equalities or inequalities are interpreted $\mathbb{P}$-almost surely).

Said differently, we say an e-variable E_1 dominates another e-value E_2 if $E_1 \geq E_2$ and $\mathbb{P}(E_1 > E_2) > 0$. An e-variable is admissible if it is not dominated by another e-value.

Proposition 3.1

For testing a simple null $\mathbb{P}$, an e-variable E is admissible if and only if it is exact. Further, any exact e-variable can be written in the form of a likelihood ratio $\mathrm{d}\mathbb{S}/\mathrm{d}\mathbb{P}$ for some distribution $\mathbb{S} \ll \mathbb{P}$.

Proof.

Assume, for the sake of contradiction, that an admissible E is not exact, meaning that $\mathbb{E}^{\mathbb{P}}[E] < 1$. If $\mathbb{E}^{\mathbb{P}}[E] = 0$ then it is dominated by the constant 1. Otherwise, defining $E' = E/\mathbb{E}^{\mathbb{P}}[E]$, we see that $E' > E$, while E' is still an e-value. Thus, E is inadmissible, contradicting the assumption. Thus, admissible e-values must be exact. To show the converse, it suffices to note that if E_2 dominates an exact e-variable E_1, then $\mathbb{E}^{\mathbb{P}}[E_2] > 1$. For the last statement, since $\int E(x)p(x)\mathrm{d}x = 1$, and $s(x) = E(x)p(x)$ is nonnegative, we infer that $s(x)$ is a probability density (and that $\mathbb{S} \ll \mathbb{P}$). Thus, we get that $E(x) = s(x)/p(x)$, as claimed.

The reader may note that the above definition and result ignores any alternative $\mathbb{Q}$, but for a stronger test it is desirable to improve an e-variable under $\mathbb{Q}$ instead of $\mathbb{P}$. However, if $\mathbb{P}$ and $\mathbb{Q}$ are mutually absolutely continuous, then the condition that $\mathbb{P}(E_1 > E_2) > 0$ is equivalent to the condition that $\mathbb{Q}(E_1 > E_2) > 0$.

Proposition 3.1 is similar to Theorem 2.25, which also connects e-variables to likelihood ratios, in a composite null setting.

3.2 Log-optimality

Next, we move to consider the behaviour of an e-variable under the alternative probability measure $\mathbb{Q}$.

Definition 3.2: Log-optimality

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

Recall that the expected logarithm of E under $\mathbb{Q}$ is called the e-power against $\mathbb{Q}$ in Definition 2.11. 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 like testing a Gaussian against a Cauchy because the tails of the Cauchy are much thicker than that of the Gaussian.

Proposition 3.3

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

We do not prove the first part in this monograph, since the proof is very technical, and defer the second part to Chapter 5. 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 3.4.

Note that $\mathbb{E}^{\mathbb{Q}}[\log E] > 0$ is equivalent to saying that the geometric mean of X is larger than one: $\exp \mathbb{E}^{\mathbb{Q}}[\log E] > 1$, which of course is a stronger constraint than requiring the (arithmetic) mean of E to be larger than 1: $\mathbb{E}^{\mathbb{Q}}[E] > 1$. These two conditions are closely related; recall Proposition 2.12.

3.3 Likelihood ratios are log-optimal for testing $\mathbb{P}$ against $\mathbb{Q}$

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

$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}) := \mathbb{E}^{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = \int q(x) \log(q(x)/p(x)) \mathbb{L}(\mathrm{d}x), \tag{3.1}$$

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: $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) \geq 0$. Further, equality holds if and only if $\mathbb{P} = \mathbb{Q}$.

Theorem 3.4

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

Proof.

Let $\mathbb{S} \ll \mathbb{P}$ be an arbitrary distribution. By Proposition 3.1, $E := \mathrm{d}\mathbb{S}/\mathrm{d}\mathbb{P}$ covers all exact e-variables. For our setting of a point null $\mathbb{P}$, it suffices to consider exact e-variables E' in Definition 2.11, 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 $E^* := \mathrm{d}\mathbb{S}^*/\mathrm{d}\mathbb{P}$. 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 $E = \mathrm{d}\mathbb{S}/\mathrm{d}\mathbb{P}$, where $\mathbb{S} \ll \mathbb{Q} \ll \mathbb{P}$. By direct substitution,

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

where the inequality follows from the fact that $\mathrm{KL}(\mathbb{Q}, \mathbb{S}) \geq 0$. This shows that $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ is log-optimal. Uniqueness follows from the fact that the inequality holds as an equality only when $\mathbb{S} = \mathbb{Q}$. 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.

In Chapter 5, we present a more general result encompassing Theorem 3.4, which even relaxes the assumption $\mathbb{Q} \ll \mathbb{P}$.

3.4 Composite alternatives: 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 to define an e-variable $E_n = E(X^n)$ such that for any $\mathbb{Q} \in \mathcal{Q}$,

$$\lim_{n\to\infty} \frac{\mathbb{E}^{\mathbb{Q}^n}[\log E_n]}{n} \to \mathrm{KL}(\mathbb{Q}, \mathbb{P}). \tag{3.2}$$

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 inequality 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

$$E_n = \int \prod_{i=1}^{n} \frac{q(X_i)}{p(X_i)} \nu(\mathrm{d}q). \tag{3.3}$$

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 compute.

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

$$E_n = \prod_{i=1}^{n} \frac{q_{i-1}(X_i)}{p(X_i)}. \tag{3.4}$$

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

It is possible to prove the asymptotic log-optimality (3.2) 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 may very well expect the aforementioned plug-in method to also have the same property. However, results of such generality remain conjectures.

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

Example 3.5: T-test

Consider the standard t-testing situation with $\mathcal{P} = \{\mathrm{N}(0, \sigma^2) : \sigma > 0\}$ and $\mathcal{Q} = \{\mathrm{N}(\mu, \sigma^2) : \sigma > 0,\ \mu \neq 0\}$ (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 (3.2) for $\mathcal{Q}$.

3.5 Existence of nontrivial e-variables for finite composite nulls

In this section, we address the existence of nontrivial e-variables and p-variables by means of two characterization results. In this section, L is a positive integer.

For a finite collection of probability measures $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_L\}$, we say that $\mathcal{P}$ is *jointly atomless* if there exists $\mathbb{P} \gg \sum_{i=1}^L \mathbb{P}_i$ and a random variable X such that under $\mathbb{P}$, X is continuously distributed and independent of $(\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}, \dots, \mathrm{d}\mathbb{P}_L/\mathrm{d}\mathbb{P})$. When $L = 1$, this is the usual definition of an atomless probability measure $\mathbb{P}_1$ (by choosing $\mathbb{P} = \mathbb{P}_1$). We will assume the following condition

$$(\mathcal{P}, \mathcal{Q}) \text{ is jointly atomless,} \tag{JA}$$

which allows us to generate a uniformly distributed random variable independent of all the probability measures and their Radon-Nikodym derivatives.

For a given $\mathcal{P}$, a random variable X is *pivotal* if it has the same distribution under all $P \in \mathcal{P}$.

Theorem 3.6

Consider testing $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_L\}$ against $\mathcal{Q} = \{\mathbb{Q}\}$. If (JA) holds, the following are equivalent:

(i) there exists a nontrivial p-variable;

(ii) there exists a bounded e-variable that has nontrivial e-power against $\mathcal{Q}$;

(iii) there exists an e-variable that is nontrivial for $\mathcal{Q}$;

(iv) it holds that $\mathbb{Q} \notin \mathrm{Conv}(\mathbb{P}_1, \dots, \mathbb{P}_L)$.

The fourth condition may appear to be a natural one, but it is in fact not sufficient beyond the finite $\mathcal{P}$ case. For general $\mathcal{P}$, the appropriate condition in fact involves an object called the *bipolar* of $\mathcal{P}$, and is formally presented in Theorem 5.8.

Theorem 3.7

Consider testing $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_L\}$ against $\mathcal{Q} = \{\mathbb{Q}\}$. If (JA) holds, the following are equivalent:

(i) there exists an exact (hence pivotal) and nontrivial p-variable;

(ii) there exists a pivotal, exact, bounded e-variable that has nontrivial e-power against $\mathcal{Q}$;

(iii) there exists an exact e-variable that is nontrivial against $\mathcal{Q}$;

(iv) there exists a random variable X that is pivotal for $\mathcal{P}$ but has a different distribution under $\mathbb{Q}$, where the laws of X under both are atomless;

(v) it holds that $\mathbb{Q} \notin \mathrm{Span}(\mathbb{P}_1, \dots, \mathbb{P}_L)$.

While the above results only consider existence, in the following sections, we will see how to construct nontrivial (and even log-optimal) e-variables for composite nulls.

Bibliographical note

The notion of log-optimality was proposed by Kelly [1956]. Larsson et al. [2024] proved the results in Section 3.2 showing the existence and uniqueness of a log-optimal e-variable under absolutely no conditions. The result in Section 3.3 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 3.4, 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 [2023b]. The existence results in Section 3.5 were proved in Zhang et al. [2024] using tools from optimal transport theory.

Part II

Core Ideas

Chapter 4

Universal inference

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.

4.1 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 or is drawn from a mixture of two Gaussians. Defining p as the density of $(1-\lambda)\mathrm{N}(\mu_1, 1) + \lambda\mathrm{N}(\mu_2, 1)$, suppose we want to test the null hypothesis represented by the set $\{(\mu_1, \mu_2, \lambda) : \lambda = 0\}$, which also equals $\{(\mu_1, \mu_2, \lambda) : \mu_1 = \mu_2\}$. Under the alternative, the triplet (μ_1, μ_2, λ) is unrestricted (except for excluding the null). In the first representation, μ_2 is not identified, and in the second, λ is not identified. This is an irregular testing problem, with a composite null and alternative, and the methodology presented here 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, like 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 methods presented in this chapter will be applicable to this problem primarily because the maximum likelihood log-concave density can be solved in polynomial time using a (relatively) efficient optimization algorithm.

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 drawn from $\mathbb{P}^*$ is a vector $X^n = (X_1, \ldots, X_n)$ with iid entries and 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.

4.2 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 randomly into two groups D_0 and D_1 (independently of X^n). One may think of these as being roughly equally sized sets as a default choice; the validity of the e-variable does not depend on the relative sizes, but the e-power does.

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 $\hat{p}_0$:

$$E = \prod_{i \in D_0} \frac{\hat{q}_1(X_i)}{\hat{p}_0(X_i)}. \tag{4.1}$$

We call this the *split 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).

Theorem 4.1

The split likelihood ratio statistic defined in (4.1) 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 conditional on D_1, $\hat{q}_1$ is a fixed density, and we recall from Section 1.3 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.

Remark 4.2. Recalling the discussion from Section 3.4 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. To elaborate, one can use D_1 to pick a distribution ν over $\mathcal{Q}$, and then define

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

which is also an e-variable, with an almost identical proof, while noting that integrals are just averages, and averages of e-variables are e-variables (more generally treated in Chapter 7).

4.3 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 (4.1) (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 $E^{(1)}, \ldots, E^{(B)}$ that can be combined by averaging:

$$\bar{E} = \frac{E^{(1)} + \cdots + E^{(B)}}{B}.$$

We call this the *subsampled likelihood ratio e-variable.* One can show that this (usually strictly) improves the e-power of the method. This relies on noting that $E^{(1)}, \ldots, E^{(B)}$ are *exchangeable.*

Recall that $Z_1, \ldots, Z_n$ are called exchangeable if the joint distribution of $(Z_1, \ldots, Z_n)$ equals that of $(Z_{\pi(1)}, \ldots, Z_{\pi(n)})$ for any permutation π of $\{1, \ldots, n\}$. A sequence $Z_1, Z_2, \ldots,$ is called exchangeable if $Z_1, \ldots, Z_n$ are exchangeable for any $n \geq 1$. Of course, iid random variables are exchangeable, but so are identical copies of the same random variable. Also, exchangeable random variables have the same marginal distribution and the same expectation.

Proposition 4.3: Improving e-power by averaging

Let $E^{(1)}, \ldots, E^{(B)}$ be exchangeable e-variables and let $\bar{E}$ denote their average. Then for any $\mathbb{Q}$,

$$\mathbb{E}^{\mathbb{Q}}[\log \bar{E}] \geq \mathbb{E}^{\mathbb{Q}}[\log E^{(1)}],$$

assuming that the right-hand side is well-defined. In particular, the e-power of the subsampled likelihood ratio statistic is at least as large as that of the split likelihood ratio statistic.

This proof follows immediately from concavity of the logarithm function and noting that $E^{(1)}, \ldots, E^{(B)}$ have the same e-power because they are identically distributed. Further, as $B \to \infty$, $\bar{E}$ converges to a fixed random variable.

The above result can be informally viewed as a form of Rao-Blackwellization. We introduced 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 $(E^{(1)} + \cdots + E^{(b)})/b$. 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, produce $E^{(1)}, E^{(2)}, \ldots$ one by one, and thus calculating their running averages $\bar{E}^{(1)}, \bar{E}^{(2)}, \ldots$ one be one, and rejecting the null at the first time any of these running averages is $\geq 1/\alpha$. One can also terminate this procedure at any large, data-dependent B, without violating the type-I error guarantee.

Proposition 4.4: 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.

4.4 Exchangeable Markov's inequality

We now state the generalization of the Markov inequality under exchangeability.

Theorem 4.5: Exchangeable Markov inequality (EMI)

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

$$\mathbb{P}\left(\sup_{n\geq 1}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq \frac{1}{\alpha}\right) \leq \mathbb{P}\left(\sup_{n\geq 1}\frac{1}{n}\sum_{i=1}^{n}|Z_i| \geq \frac{1}{\alpha}\right) \leq \alpha\ \mathbb{E}^{\mathbb{P}}[|Z_1|]. \tag{4.2}$$

In particular, if $Z_1, Z_2, \ldots$ 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 second inequality is trivially true. If Z_1 is integrable, $(\sum_{i=1}^{n}|Z_i|/n)_{n\geq 1}$ is a nonnegative reverse martingale with respect to the "exchangeable filtration" (a reverse filtration), and this allows us to infer the second inequality as a consequence of a time-reversed variant of Ville's inequality.

A useful corollary of the EMI is as follows. Let $Z_1, \ldots, Z_n$ be any set of (potentially nonexchangeable) arbitrarily dependent random variables. Let π be a uniformly random permutation of $\{1, \ldots, n\}$. Then, for any $\alpha > 0$,

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

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 $Z_{\pi(1)}, \ldots, Z_{\pi(n)}$ are n samples drawn uniformly at random with or without replacement from this bag. Then, we have

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

This holds because the sampling process induces the exchangeability required for (4.2) to be invoked on the otherwise non-exchangeable random variables.

4.5 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^*. Such a confidence set, without requiring any regularity conditions, is immediately given by inverting the universal inference test. Meaning that 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 4.6

Given any predefined error level $\alpha \in (0,1)$, the set $C(X^n)$ defined above satisfies

$$\mathbb{P}_{\theta^*}(\theta^* \in C(X^n)) \geq 1 - \alpha,$$

for any $\theta^* \in \Theta$.

The proof is immediate, relying only on the observation that $E = \prod_{i \in D_0} \frac{p_{\hat{\theta}}(X_i)}{p_{\theta^*}(X_i)}$ 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.

4.6 Extensions

Profile likelihood

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

$$B(X^n) = \{\psi : C(X^n) \cap g^{-1}(\psi) \neq \varnothing\}.$$

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

$$\mathcal{L}_0^\dagger(\psi) := \sup_{\theta : g(\theta) = \psi} \mathcal{L}_0(\theta),$$

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.

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 $\int k(x,y)\mathrm{d}y = 1$ for any x. For any density p_θ, let its smoothed version be denoted

$$\widetilde{p}_\theta(y) := \int k(x,y) p_\theta(x) \mathrm{d}x,$$

Note that $\widetilde{p}_\theta$ is also a probability density. Let the smoothed empirical density based on D_0 be defined as

$$\widetilde{p}_n := \frac{1}{|D_0|} \sum_{i \in D_0} k(X_i, \cdot).$$

Define the smoothed maximum likelihood estimator as the KL projection of $\widetilde{p}_n$ onto $\{\widetilde{p}_\theta\}_{\theta\in\Theta_0}$:

$$\widetilde{\theta}_0 := \arg\min_{\theta\in\Theta_0} \mathrm{KL}(\widetilde{p}_n, \widetilde{p}_\theta).$$

If we define the smoothed likelihood on the first half of the data D_0 as

$$\widetilde{\mathcal{L}}_0(\theta) := \prod_{i\in D_0} \exp \int k(X_i, y) \log \widetilde{p}_\theta(y)\mathrm{d}y,$$

then it can be checked that $\widetilde{\theta}_0$ maximizes the smoothed likelihood, that is $\widetilde{\theta}_0 = \arg\max_{\theta\in\Theta_0} \widetilde{\mathcal{L}}_0(\theta)$. As before, let $\widehat{\theta}_1 \in \Theta$ be any estimator based on D_1. The smoothed split LRT is defined as:

$$\text{reject } H_0 \text{ if } \widetilde{U}_n > 1/\alpha, \text{ where } \widetilde{U}_n = \frac{\widetilde{\mathcal{L}}_0(\widehat{\theta}_1)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}.$$

We now verify that the smoothed split LRT controls type-1 error, by simply checking that $\widetilde{U}_n$ is an e-variable. Indeed, for any fixed $\psi \in \Theta$, we have

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left[\frac{\widetilde{\mathcal{L}}_0(\psi)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}\right] &\overset{(i)}{\leq} \mathbb{E}_{\theta^*}\left[\frac{\widetilde{\mathcal{L}}_0(\psi)}{\widetilde{\mathcal{L}}_0(\theta^*)}\right] \\
&= \prod_{i\in D_0} \int \exp\left(\int k(x,y)\log\frac{\widetilde{p}_\psi(y)}{\widetilde{p}_{\theta^*}(y)}\mathrm{d}y\right) p_{\theta^*}(x)\mathrm{d}x \\
&\overset{(ii)}{\leq} \int \left(\int k(x,y)\frac{\widetilde{p}_\psi(y)}{\widetilde{p}_{\theta^*}(y)}\mathrm{d}y\right) p_{\theta^*}(x)\mathrm{d}x \\
&= \int \left(\frac{\int k(x,y)p_{\theta^*}(x)\mathrm{d}x}{\widetilde{p}_{\theta^*}(y)}\right)\widetilde{p}_\psi(y)\mathrm{d}y \ = \int \widetilde{p}_\psi(y)\mathrm{d}y = 1,
\end{aligned}$$

where inequality (i) is because $\widetilde{\theta}_0$ maximizes the smoothed likelihood, and step (ii) follows by Jensen's inequality.

Relaxed 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

$$\widehat{\theta}_0^F := \arg\max_\theta F_0(\theta).$$

If we instead define the test statistic

$$T_n' := \frac{\mathcal{L}_0(\widehat{\theta}_1)}{F_0(\widehat{\theta}_0^F)},$$

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 $F_0(\widehat{\theta}_0^F) \geq \mathcal{L}_0(\widehat{\theta}_0)$, 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.*

Powered likelihood

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

$$\mathbb{E}_\theta\left[\left(\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta)}\right)^\eta \,\middle|\, D_1\right] = \prod_{i \in D_0} \int p^\eta_{\widehat{\theta}_1}(y_i) p_\theta^{1-\eta}(y_i) \mathrm{d}y_i \leq 1,$$

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.)

Conditional likelihood

Our presentation so far 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 one subset of 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.

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 mentioned in Manole and Ramdas [2023], but its implications for testing with e-variables was first detailed in Ramdas and Manole [2024]. Nguyen [2020] derived an extension to composite likelihood settings.

Dunn et al. [2022] studied the efficiency of universal confidence sets when the data are multivariate Gaussian with identity covariance — here standard likelihood methods work perfectly well, so it serves as good benchmark. 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). They also studied the optimal splitting ratio in this setting, guided by the metric of squared radius of the confidence set, deriving an expression which approaches $1/2$ as d increases. 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. [2024] applied the method to the problem of testing log-concavity, for which no other valid test is known. Shi and Drton [2024] recently studied universal inference for Gaussian mixtures. They prove that 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.

Chapter 5

The numeraire e-variable and the 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 (subprobability) distribution called the reverse information projection (RIPr).

Setting: $\mathcal{P}$ versus $\mathbb{Q}$.

As elaborated in Chapter 3.4, 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 phrase e-variable below, we always mean e-variable *for* $\mathcal{P}$. We let $\mathfrak{E}$ denote the set of e-variables:

$$\mathfrak{E} = \{E \geq 0 : \mathbb{E}^{\mathbb{P}}[E] \leq 1 \text{ for all } \mathbb{P} \in \mathcal{P}\}.$$

5.1 Numeraire

Definition 5.1

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 $E^* = \mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$. 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 5.2: 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* in the sense that

$$\mathbb{E}^{\mathbb{Q}}\left[\log \frac{E}{E^*}\right] \leq 0 \text{ for every e-variable } E \in \mathfrak{E}, \tag{5.1}$$

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 (5.1) even if $\mathbb{E}^{\mathbb{Q}}[\log E^*]$ happens to be infinite.

Proposition 5.3: 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, the theory simplifies if the alternative $\mathbb{Q}$ does not assign positive probability to any event that is null under every $\mathbb{P} \in \mathcal{P}$. In this case we say that $\mathbb{Q}$ *is absolutely continuous with respect to* $\mathcal{P}$, written $\mathbb{Q} \ll \mathcal{P}$. 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.

Theorem 5.4

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 $Q \ll \mathcal{P}$.

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 $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$. 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 (3.1) 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 $\mathbb{Q} \in \pi$, define

$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}) = -\mathbb{E}^{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{P}^a}{\mathrm{d}\mathbb{Q}}\right],$$

where $\mathbb{P}^a$ is the absolutely continuous part of $\mathbb{P}$ with respect to $\mathbb{Q}$. When $\mathbb{P} \in \pi$, this definition coincides with (3.1).

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 $\mathbb{P} \in \mathrm{Conv}(\mathcal{P})$,

$$\mathbb{E}^{\mathbb{Q}}\left[\log\left(E\frac{\mathrm{d}\mathbb{P}^a}{\mathrm{d}\mathbb{Q}}\right)\right] \leq \mathbb{E}^{\mathbb{P}}[E] - 1 \leq 0.$$

This immediately implies the following fact that we record for later use.

Proposition 5.5: Weak duality

For all e-variables $E \in \mathfrak{E}$ and distributions $\mathbb{P} \in \mathrm{Conv}(\mathcal{P})$, we have $\mathbb{E}^{\mathbb{Q}}[\log E] \leq \mathrm{KL}(\mathbb{Q}, \mathbb{P})$ (where if $\mathbb{E}^{\mathbb{Q}}[\log E]$ is undefined, we treat it as $-\infty$), and thus

$$\sup_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E] \leq \inf_{P \in \mathrm{Conv}(\mathcal{P})} \mathrm{KL}(\mathbb{Q}, \mathbb{P}). \tag{5.2}$$

Note that both sides of (5.2) are nonnegative and well defined, since the constant 1 is always an e-variable. It turns out that the numeraire E^* satisfies

$$\mathbb{E}^{\mathbb{Q}}[\log E^*] = \mathrm{KL}(\mathbb{Q}, \mathbb{P}^*),$$

for an appropriately defined $\mathbb{P}^*$ that we introduce and discuss next.

5.2 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 5.6

Suppose $\mathcal{P} = \{\mathbb{P}_1, \dots, \mathbb{P}_K\}$ with $\mathbb{P}_k \ll \mathbb{Q}$ for all $k \in [K]$, and $\min_{\mathbb{P} \in \mathrm{Conv}(\mathcal{P})} \mathrm{KL}(\mathbb{Q}, \mathbb{P}) < \infty$. Then,

$$\mathbb{E}^{\mathbb{Q}}[\log E^*] = \max_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E] = \min_{\mathbb{P} \in \mathrm{Conv}(\mathcal{P})} \mathrm{KL}(\mathbb{Q}, \mathbb{P}) = \mathrm{KL}(\mathbb{Q}, \mathbb{P}^*),$$

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 $\min_{\lambda \in \Delta_K} \mathrm{KL}(\mathbb{Q}, \sum_{k=1}^K \lambda_k \mathbb{P}_k)$. 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. We now prove the strong duality result in the middle equality.

Proof of Proposition 5.6

Recall from the earlier weak duality result of Proposition 5.5 that if we find any e-variable whose e-power equals $\mathrm{KL}(\mathbb{Q}, \mathbb{P}^*)$, 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 $\mathbb{P} \in \mathcal{P}$,

$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}^*) = \min_{\alpha \in [0,1]} \mathrm{KL}(\mathbb{Q}, \alpha \mathbb{P}^* + (1-\alpha)\mathbb{P})$$

The optimality criterion (Karusch-Kuhn-Tucker condition) for the above minimization problem is that the gradient with respect to α is nonpositive at $\alpha = 1$ (going from the optimizer 1 towards 0 increases the objective):

$$-\mathbb{E}^{\mathbb{Q}}\left[\frac{\mathrm{d}(\mathbb{P}^* - \mathbb{P})}{\mathrm{d}\mathbb{P}^*}\right] \leq 0 \iff \mathbb{E}^{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^*}\right] \leq 1 \iff \mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}^*}\right] \leq 1.$$

The last inequality means that $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}^*$ is an e-variable. By definition of e-power, the e-power of this e-variable is $\mathrm{KL}(\mathbb{Q}, \mathbb{P}^*)$. Thus, the numeraire E^* is $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}^*$ and the proof is complete.

5.3 Reverse information projection

Remarkably, all assumptions in Proposition 5.6 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 $\mathrm{Conv}(\mathcal{P})$ to $\mathbb{Q}$ as measured by KL. However, when the infimum KL divergence equals infinity, it is unclear how do define the "closest" distribution to $\mathbb{Q}$.

Second, in the previous section, we used the compactness of $\mathrm{Conv}(\mathcal{P})$ when $\mathcal{P}$ is finite to argue that the infimum is attained by a distribution in $\mathrm{Conv}(\mathcal{P})$. However, when $\mathcal{P}$ is uncountable (as would typically be the case in statistical applications), the set is no longer compact (in 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 $\mathrm{Conv}(\mathcal{P})$, and in fact may be a *sub*distribution 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 5.7: Reverse information projection (RIPr)

Let E^* be the numeraire for $\mathcal{P}$ against $\mathbb{Q}$. Define the reverse information projection of $\mathbb{Q}$ onto $\mathcal{P}$, denoted $\mathbb{P}^*$, via

$$\frac{\mathrm{d}\mathbb{P}^*}{\mathrm{d}\mathbb{Q}} = \frac{1}{E^*}.$$

In words, $\mathbb{P}^*$ is a nonnegative measure whose likelihood ratio with respect to $\mathbb{Q}$ equals $1/E^*$, which is understood to equal zero on $\{E^* = \infty\}$. Since E^* exists, 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.

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

$$\mathcal{P}^\circ = \{E \in \mathcal{X}_+ : \mathbb{E}^{\mathbb{P}}[E] \le 1 \text{ for all } \mathbb{P} \in \mathcal{P}\}.$$

This is precisely the set of e-variables, i.e., $\mathcal{P}^\circ = \mathfrak{E}$. The *bipolar* of $\mathcal{P}$ is

$$\mathcal{P}^{\circ\circ} = \{\mathbb{P} \in \mathcal{M}_+ : \mathbb{E}^{\mathbb{P}}[E] \le 1 \text{ for all } E \in \mathcal{P}^\circ\}.$$

It is easy to see that

$$\mathcal{P}^{\circ\circ} \cap \mathcal{M}_1 \supseteq \mathrm{Conv}(\mathcal{P}),$$

and there are examples where the inclusion is 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 that $\mathbb{P}'(A) \le \mathbb{P}(A)$ for all A.

The bipolar 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 nontrivial e-variable from Definition 2.9.

Theorem 5.8

A nontrivial 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 2.12 that a nontrivial test exists if and only if a nontrivial e-variable exists. The latter condition is possible if and only if $\mathbb{Q} \notin \mathcal{P}^{\circ\circ}$ by definition of $\mathcal{P}^{\circ\circ}$.

We now claim that even though it is possible that $\mathbb{P}^* \notin \mathrm{Conv}(\mathcal{P})$ in general, it does like in $\mathcal{P}^{\circ\circ}$.

Proposition 5.9

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,

$$\mathbb{E}^{\mathbb{P}^*}[E] = \mathbb{E}^{\mathbb{Q}}[\mathbb{1}_{\{E^*<\infty\}} E/E^*] \leq \mathbb{E}^{\mathbb{Q}}[E/E^*] \leq 1,$$

because E^* is the numeraire.

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 $\mathbb{E}^{\mathbb{Q}}[(\log E)_-] = \infty$.

Theorem 5.10

Assume $Q \ll \mathcal{P}$. Letting E^* be the numeraire and $\mathbb{P}^*$ the RIPr, one has the strong duality relation

$$\mathbb{E}^{\mathbb{P}}[\log E^*] = \sup_{E \in \mathfrak{E}} \mathbb{E}^{\mathbb{Q}}[\log E] = \inf_{\mathbb{P} \in \mathcal{P}^{\circ\circ}} \mathrm{KL}(\mathbb{Q}, \mathbb{P}) = \mathrm{KL}(\mathbb{Q}, \mathbb{P}^*),$$

where these quantities may equal ∞.

We also provide some characterizing properties of the RIPr.

Theorem 5.11

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 q. The following statements imply each other:

1. $\mathbb{P}^*$ is the RIPr.
2. $\mathbb{E}^{\mathbb{Q}}\left[\frac{d\mathbb{P}^a}{d\mathbb{P}^*}\right] \leq 1$ for all $\mathbb{P} \in \mathcal{P}^{\circ\circ}$
3. $\mathbb{E}^{\mathbb{Q}}\left[\log \frac{d\mathbb{P}^a}{d\mathbb{P}^*}\right] \leq 0$ 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

4. $\mathbb{E}^{\mathbb{Q}}[p/p^*] \leq 1$ 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 like exponential families, as we will see in Section 5.6.

5.4 Finding the numeraire and RIPr

We first present a verification theorem which simplifies the task of checking that candidates E^* and $\mathbb{P}^*$ (for the numeraire and RIPr respectively) are in fact optimal.

Theorem 5.12

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 $\mathrm{d}\mathbb{P}^*/\mathrm{d}\mathbb{Q} = 1/E^*$. 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 5.4 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 5.9 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 $1 = \mathbb{E}^{\mathbb{P}^*}[E^*] < \mathbb{E}^{\mathbb{P}}[E^*] \leq 1$.

Corollary 5.13

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

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{M}_1 \colon \mathbb{E}^{\mathbb{P}}[E] \leq 1 \text{ for all } E \in \mathfrak{E}_0\}. \tag{5.3}$$

Let E^* be a $\mathbb{Q}$-almost surely strictly positive e-variable such that

$$\mathbb{E}^{\mathbb{Q}}\left[\frac{1}{E^*}\right] = 1 \text{ and } \mathbb{E}^{\mathbb{Q}}\left[\frac{E}{E^*}\right] \leq 1 \text{ for all } E \in \mathfrak{E}_0. \tag{5.4}$$

Then E^* is the numeraire and the RIPr belongs to $\mathcal{P}$.

Proof

Let $\mathbb{P}^*$ be given by $\mathrm{d}\mathbb{P}^*/\mathrm{d}\mathbb{Q} = 1/E^*$. Then (5.4) 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 (5.3), $\mathbb{P}^*$ belongs to $\mathcal{P}$, hence to $\mathcal{P}^{\circ\circ}$. The result now follows from Theorem 5.12.

Thanks to Lemma 5.2, Corollary 5.13 also holds if one has '$\supseteq$' instead of '$=$' in (5.3). 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 (5.3) can become smaller than $\mathcal{P}$. In that case, (5.4) only gets harder to satisfy, and if we find an e-variable E^* satisfying it, it surely must still be the numeraire.

5.5 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 $E^{\mathrm{UI}} = q(Z)/p_{\max}(Z)$ 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 5.14

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 $E^* \geq E^{\mathrm{UI}}$, 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 3.4. A reasonable goal is to derive an e-variable $E_n = E(X^n)$ for $\mathcal{P}$ such that for any $\mathbb{Q} \in \mathcal{Q}$,

$$\lim_{n\to\infty} \frac{\mathbb{E}^{\mathbb{Q}}[\log E_n]}{n} \to \mathrm{KL}(\mathbb{Q}, \mathbb{P}^*), \tag{5.5}$$

where $\mathbb{P}^*$ is the RIPr of $\mathbb{Q}$ 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 3.4. 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 $\widetilde{\mathbb{Q}}^{(n)} = \int \mathbb{Q}^n \nu(\mathrm{d}\mathbb{Q})$. We conjecture that this is indeed asymptotically log-optimal under weak assumptions, but such a general result is currently not known.

5.6 Examples

We now turn to examples. In these examples we consider a single random observation Z in a measurable space $\mathcal{Z}$. For concreteness we take $(\Omega, \mathcal{F})$ to be $\mathcal{Z}$ with its σ-algebra, and Z the canonical random variable. 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 let $\mathcal{X}_+$ and $\mathcal{M}_+$ denote the set of all $[0,\infty]$-valued measurable functions and nonnegative measures on $\mathcal{F}$.

Exponential family with one-dimensional sufficient statistic

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 $\mathcal{P} = \{p_\theta \colon \theta \in \Theta_0\}$ 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 smallest element θ^* and that $\theta_1 < \theta^*$. It is natural to conjecture that p_{θ^*} is the RIPr. To confirm this, it suffices to verify the fourth condition in Theorem 5.11. Using the standard formula for the moment generating function of the sufficient statistic we get

$$\mathbb{E}^{\mathbb{Q}}\left[\frac{p_\theta}{p_{\theta^*}}\right] = \mathbb{E}^{\mathbb{Q}}\left[e^{(\theta-\theta^*)T(Z) - A(\theta) + A(\theta^*)}\right] = e^{A(\theta_1+\theta-\theta^*) - A(\theta_1) - A(\theta) + A(\theta^*)}$$

for all $\theta \in \Theta_0$. Since A is convex, its derivative A' is increasing. Together with the fundamental theorem of calculus, as well as $\theta_1 < \theta^*$ and $\theta - \theta^* \geq 0$ for $\theta \in \Theta_0$, this gives

$$\begin{aligned} A(\theta_1 + \theta - \theta^*) - A(\theta_1) &= \int_0^1 (\theta - \theta^*) A'(\theta_1 + t(\theta - \theta^*)) dt \\ &\leq \int_0^1 (\theta - \theta^*) A'(\theta^* + t(\theta - \theta^*)) dt \\ &= A(\theta) - A(\theta^*). \end{aligned}$$

We conclude that $\mathbb{E}^{\mathbb{Q}}[p_\theta / p_{\theta^*}] \leq 1$ for all $\theta \in \Theta_0$ so that the fourth condition in Theorem 5.11 holds and p_{θ^*} is indeed the RIPr. As a result the numeraire is the likelihood ratio

$$E^* = \frac{p_{\theta_1}(Z)}{p_{\theta^*}(Z)} = e^{(\theta_1 - \theta^*) T(Z) - A(\theta_1) + A(\theta^*)}.$$

Testing symmetry

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

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

which is again 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 $\widetilde{\mathbb{P}}$ of $\mathbb{Q}$, whose density is $\widetilde{p}(z) = (q(z) + q(-z))/2$. However, this cannot quite be true in general because $\widetilde{\mathbb{P}}$ 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 $\widetilde{\mathbb{P}}$ 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 $\widetilde{\mathbb{P}}$ belongs to $\mathcal{P}$, which in particular shows that $\mathbb{Q} \ll \mathcal{P}$ since $\mathbb{Q} \ll \widetilde{\mathbb{P}}$. (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

$$E^* = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}^*} = \frac{2q(Z)}{q(Z) + q(-Z)}.$$

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. Next, for any e-variable $E \leq 1 + \phi(Z)$ where ϕ is odd we get

$$\mathbb{E}^{\mathbb{Q}}\left[\frac{E}{E^*}\right] \leq \mathbb{E}^{\mathbb{Q}}\left[(1 + \phi(Z)) \frac{q(Z) + q(-Z)}{2q(Z)}\right] = \mathbb{E}^{\widetilde{\mathbb{P}}}\left[(1 + \phi(Z)) \mathbb{1}\{\{q(Z) > 0\}\}\right] \leq \mathbb{E}^{\widetilde{\mathbb{P}}}[1 + \phi(Z)] = 1.$$

This confirms the numeraire property.

Finally, let us remark that it is not really necessary that $\mathbb{Q}$ admit a density. The symmetrization of $\mathbb{Q}$ is still well-defined as $\widetilde{\mathbb{P}} = \frac{1}{2}(\mathbb{Q}+\widetilde{\mathbb{Q}})$ where $\widetilde{\mathbb{Q}}$ is the distribution of $-Z$ under $\mathbb{Q}$, or equivalently, the pushforward of $\mathbb{Q}$ under the reflection map $z \mapsto -z$. The RIPr is then the absolutely continuous part $\mathbb{P}^* = \widetilde{\mathbb{P}}^a$, and the numeraire is any nonnegative version of the Radon–Nikodym derivative $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}^*$ such that $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}^* - 1$ is odd.

Testing bounded means

Let $\mathcal{Z} = [0,1]$ so that the random variable Z takes values in the unit interval. Fix $\mu \in (0, \frac{1}{2})$ and consider the null hypothesis that the mean of Z is at most μ,

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

There is no single dominating measure for $\mathcal{P}$ since it contains the uncountable non-dominated family $\{\delta_z \colon z \in [0,\mu]\}$. However, $\mathcal{P}$ is generated by $\mathfrak{E}_0 = \{Z/\mu\}$ in the sense of Corollary 5.13, 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 $\lambda \in [0, \mu^{-1}]$. 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

$$\mathbb{E}^{\mathbb{Q}}\left[\frac{Z-\mu}{1+\lambda^*(Z-\mu)}\right] = 0. \tag{5.6}$$

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

$$\int_0^1 \frac{\lambda^*(z-\mu)}{1+\lambda^*(z-\mu)}\mathrm{d}z = 1 - \int_0^1 \frac{1}{1+\lambda^*(z-\mu)}\mathrm{d}z = 1 - \frac{1}{\lambda^*}\log\left(\frac{1+\lambda^*(1-\mu)}{1-\lambda^*\mu}\right).$$

Thus (5.6) 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 (5.6) 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 (5.4) of Corollary 5.13 is satisfied and, hence, that E^* is the numeraire and the RIPr $\mathbb{P}^*$ belongs to $\mathcal{P}$.

Testing subGaussian 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. That is, we set

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

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 $\mathfrak{E}_0 = \{e^{\lambda Z - \lambda^2/2} \colon \lambda \in [0,\infty)\}$, so we will again look for a candidate numeraire and verify it using Corollary 5.13. 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, (5.4) of Corollary 5.13 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.

Testing likelihood ratio bounds

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

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

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 $\{\mathbb{P}_0\}$.

For a 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 $\mathbb{P}_0$. We claim that the numeraire for testing $\mathcal{P}$ against $\mathbb{Q}$ is given by

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

where $Z = \mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}_0$ and $z_0 \geq 0$ is the largest constant such that

$$\int_0^{1/\gamma} (q_t(Z) \vee z_0)\mathrm{d}t = 1.$$

To show this claim, we first verify that E^* is an e-variable for $\mathcal{P}$. This follows by noting that

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

where the first equality is a classic result on risk measures; see McNeil et al. [2015, Theorem 8.14] and (13.6) in Chapter 13. Next, we verify that E^* is optimal. Let $\mathbb{P}^* \in \mathcal{M}_+$ be specified by

$$\frac{\mathrm{d}\mathbb{P}^*}{\mathrm{d}\mathbb{P}_0} = \gamma \wedge \frac{\gamma Z}{z_0},$$

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 $\mathrm{d}\mathbb{P}^*/\mathrm{d}\mathbb{P}_0 \leq \gamma$. Moreover,

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}^*} = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}^*}\mathbb{1}_{\{Z>0\}} = \frac{Z \vee z_0}{\gamma} = E^*.$$

Therefore, by Theorem 3.4 (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 $\int X \mathrm{d}\mathbb{P}^* \leq 1$. Since the an e-variable for $\mathcal{P}$ always satisfies the constraint $\int X \mathrm{d}\mathbb{P}^* \leq 1$, we know that E^* is also log-optimal for testing $\mathcal{P}$ against $\mathbb{Q}$, and thus it is the numeiraire.

The special case of $\gamma = 1$ and $\mathbb{Q} \ll \mathbb{P}_0$ in the above example is precisely Theorem 3.4, with the log-optimal e-variable $E^* = Z = \mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}_0$ and z_0 being the largest constant such that $z_0 \leq Z$ almost surely (typically $z_0 = 0$).

Bibliographical note

This chapter is largely shaped by treatment in Larsson et al. [2024]. Historically, the RIPr 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. [2024] 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 (5.5).

The proof of Lemma 5.14 can be found in Larsson et al. [2024]. The symmetry example has been studied in Larsson et al. [2024], Ramdas et al. [2020] and Koning [2023b]. The numeraire for more general group invariant nulls has been derived in Pérez-Ortiz et al. [2024]. The crucial example of the t-test (omitted here) has been studied in the previous paper, but also in Lai [1976] and Wang and Ramdas [2023b]. The numeraire for one-parameter exponential families was derived in Grünwald et al. [2024b], and then in Larsson et al. [2024] 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 5.10 is closely related to, yet different from, the $\mathrm{KL}_{\inf}$ 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.

Chapter 6

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.

6.1 Preliminaries

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 $(\mathcal{F}_t)_{t\in T}$ (a nested sequence of σ-algebras), where T is a finite or infinite set of indices starting from 0. We deal mainly with discrete time processes, with $T = \mathbb{N}_0$.

A sequence of random variables $Y = (Y_t)_{t\geq 0}$ 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. Y is called *predictable* if Y_t is measurable with respect to $\mathcal{F}_{t-1}$.

Often $\mathcal{F}$ is chosen as the natural filtration of data, that is, $\mathcal{F}_t := \sigma(X^t)$, where $X^t = (X_1, \ldots, X_t)$, with $\mathcal{F}_0$ being trivial ($\mathcal{F}_0 = \{\varnothing, \Omega\}$). 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.

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

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\exists t \geq 1 : \phi_t = 1) \leq \alpha. \tag{6.1}$$

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. In the literature, 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 $\phi_\infty = \lim_{t\to\infty} \phi_t = \sup_{t\in\mathbb{N}} \phi_t$, 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 $\inf \varnothing = \infty$. 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 (6.1) as:

$$\mathbb{E}^{\mathbb{P}}[\phi_\tau] \leq 1, \tag{6.2}$$

for all $\mathbb{P} \in \mathcal{P}$ and all $\tau \in \mathcal{T}$.

6.2 Test (super)martingales and e-processes

In the discrete-time setting of this chapter, an integrable process M is a *martingale* for $\mathbb{P}$ if

$$\mathbb{E}^{\mathbb{P}}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1} \tag{6.3}$$

for all $t \geq 1$, where all such equalities between random variables are understood to hold $\mathbb{P}$-almost surely. M is a *supermartingale* for $\mathbb{P}$ if it satisfies (6.3) with "=" relaxed to "$\leq$".

Definition 6.1: Test (super)martingales and e-processes

(i) A process M is called a *test (super)martingale* for $\mathcal{P}$ if for every $\mathbb{P} \in \mathcal{P}$, it satisfies three properties: (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 nonnegative process E is called an *e-process* it 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 $\mathbb{P}$.

The fact that the above two definitions (ii.a) and (ii.b) for an e-process are equivalent is nontrivial, but its proof is omitted. In definition (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 6.2: Asymptotic growth rate and log-optimality

An e-process M is said to be consistent against $\mathbb{Q}$ if $M_t \to \infty$, $\mathbb{Q}$-almost surely as $t \to \infty$. The asymptotic growth rate of an e-process M is defined as

$$\liminf_{t\to\infty} \frac{\log M_t}{t}.$$

This limit is typically $\mathbb{Q}$-almost surely a constant (which depends on $\mathbb{Q}$), and we call that constant the *asymptotic growth rate* of M against $\mathbb{Q}$. The e-process M is asymptotically log-optimal if for any other e-process M', we have

$$\liminf_{t\to\infty} \frac{1}{t} (\log M_t - \log M'_t) \geq 0 \quad \text{in } L^1(\mathbb{Q}).$$

Note that the asymptotic growth rate of any e-process for $\mathcal{P}$ against $\mathbb{P} \in \mathcal{P}$ is always nonpositive. An asymptotically positive 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.

6.3 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 6.3: The optional stopping theorem

If M is a test (super)martingale for $\mathbb{P}$, then for any $\mathcal{F}$-stopping time τ (possibly infinite), $\mathbb{E}^{\mathbb{P}}[M_\tau] \leq 1$. 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 6.4: Ville's inequality

If M is an e-process for $\mathcal{P}$, then the following three equivalent statements hold:

$$\mathbb{P}\left(\exists t \in \mathbb{N} : M_t \geq \frac{1}{\alpha}\right) \leq \alpha \text{ for every } \mathbb{P} \in \mathcal{P} \text{ and } \alpha \in [0,1];$$

$$\iff \quad \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\sup_{t\in\mathbb{N}} M_t \geq \frac{1}{\alpha}\right) \leq \alpha \text{ for every } \alpha \in [0,1]; \tag{6.4a}$$

$$\iff \quad \sup_{\mathbb{P}\in\mathcal{P},\tau\geq 0} \mathbb{P}\left(M_\tau \geq \frac{1}{\alpha}\right) \leq \alpha \text{ for every } \alpha \in [0,1]. \tag{6.4b}$$

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 (6.4a) and (6.4b) usually only hold with inequality (for example, for the singleton $\mathcal{P} = \{\mathbb{P}\}$), but it can hold with equality for larger nontrivial nonparametric classes $\mathcal{P}$. 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 $M^\lambda = (\exp(\lambda B_t - \lambda^2 t/2))_{t\in\mathbb{R}_+}$ 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. The same holds true for many continuous-time and continuous-path martingales.

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. A converse fact is also true, as presented next.

Proposition 6.5

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 (6.2). 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.

A practical use of e-processes without thresholding concerns the property of *optional continuation*, 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 6.6

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 $t \in T$. 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 6.1 (ii.b). Clearly, $M_t \leq M_{\tau\wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}}$. Below we show that $M^{\mathbb{P}} := (M_{\tau\wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}})_{t\in T}$ 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

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}}[M_t^{\mathbb{P}}|\mathcal{F}_{t-1}] &= \mathbb{E}^{\mathbb{P}}[M_{\tau\wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}}|\mathcal{F}_{t-1}] \\
&= \mathbb{E}^{\mathbb{P}}[M_{\tau\wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}} \mathbb{1}_{\{\tau<t\}}|\mathcal{F}_{t-1}] + \mathbb{E}^{\mathbb{P}}[M_{\tau\wedge t}^{A,\mathbb{P}} M_t^{B,\mathbb{P}} \mathbb{1}_{\{\tau\geq t\}}|\mathcal{F}_{t-1}] \\
&= \mathbb{1}_{\{\tau<t\}} M_\tau^{A,\mathbb{P}} \mathbb{E}^{\mathbb{P}}[M_t^{B,\mathbb{P}}|\mathcal{F}_{t-1}] + \mathbb{1}_{\{\tau\geq t\}} \mathbb{E}^{\mathbb{P}}[M_t^{A,\mathbb{P}}|\mathcal{F}_{t-1}] \\
&\leq \mathbb{1}_{\{\tau<t\}} M_\tau^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} + \mathbb{1}_{\{\tau\geq t\}} M_{t-1}^{A,\mathbb{P}} \\
&= \mathbb{1}_{\{\tau\leq t-1\}} M_\tau^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} + \mathbb{1}_{\{\tau> t-1\}} M_{t-1}^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} \\
&= M_{\tau\wedge(t-1)}^{A,\mathbb{P}} M_{t-1}^{B,\mathbb{P}} = M_{t-1}^{\mathbb{P}}.
\end{aligned}$$

Hence, $M^{\mathbb{P}}$ is a test supermartingale, implying that M is an e-process according to Definition 6.1 (ii.b).

We make some more observations on the result in Proposition 6.6: 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 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 6.6 repeatedly.

We end this section by noting that Ville's inequality (6.4a) immediately implies that if M is an e-process, then $(\inf_{s\le t} 1/M_s)_{t\in T}$ is a p-process. 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. The latter may appear more powerful, but it is also perhaps 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 6.5.

6.4 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

$$M_0 = 1 \quad \text{and} \quad M_t = \prod_{i=1}^{t} \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(X_i) \text{ for } t = 1, 2, \dots \tag{6.5}$$

is an e-process adapted to the filtration generated by the data. Moreover, we can easily see that M is a test martingale for $\mathbb{P}$.

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.

Fact 6.7: The likelihood ratio process is log-optimal

In the setting of testing the simple null hypothesis $\mathbb{P}$ against $\mathbb{Q} \ll \mathbb{P}$ with iid data, the likelihood ratio process M_t in (6.5) is log-optimal among all e-variables using the first t data points for any $t \in \mathbb{N}$. This claim follows from Theorem 3.4, by noting that M_t is precisely the likelihood ratio of $\mathbb{Q}$ to $\mathbb{P}$ based on the first t data points.

In fact, even if the data is not iid, the likelihood ratio based on the first n data points,

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(X_1, \dots, X_n),$$

is still a test martingale for $\mathbb{P}$, and indeed the log-optimal one against $\mathbb{Q}$.

Revisiting the example in Section 1.3 with the e-variable $E_n = \exp\left(\mu S_n - n\mu^2/2\right)$ in (1.1), we can check that

$$\mathbb{E}^{\mathbb{P}}[E_n|E_{n-1}] = E_{n-1}\mathbb{E}^{\mathbb{P}}[\exp(\mu X_n - \mu^2/2)] = E_{n-1} \quad \text{for } n \in \mathbb{N} \text{ with } E_0 = 1,$$

and hence E is not only an e-process, but also a test martingale. This is true for all likelihood ratio processes in (6.5), including the example in (1.5).

For testing a simple null $\mathbb{P}$ against a composite alternative $\mathcal{Q}$, there essentially only two alternatives, and they were both discussed in Section 3.4. Both the mixture likelihood ratio in (3.3) and the plug-in likelihood ratio in (3.4) are actual test supermartingales and hence e-processes.

For testing composite nulls, the situation is much more complicated, as non-trivial composite martingales may not exist while non-trivial e-processes may exist. We discuss this case in Section 6.6 and later.

6.5 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 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 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 1.

Algorithm 1 Testing by betting: simple null $\mathbb{P}$

```
Statistician's initial wealth is $W_0 = 1$
for $t = 1, 2, \dots$ do
    Statistician declares bet $S_t : \mathcal{X} \to [0, \infty]$ such that $\mathbb{E}^{\mathbb{P}}[S_t(X) \mid X_1, \dots, X_{t-1}] \leq 1$
    Observe $X_t$
    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 3.4: 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 (6.5).

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 2.

Algorithm 2 Testing by betting: composite null $\mathcal{P}$

```
for $\mathbb{P} \in \mathcal{P}$ do
    Statistician's initial wealth in game $\mathbb{P}$ is $W_0^{\mathbb{P}} = 1$
for $t = 1, 2, \dots$ do
    Statistician declares bet $S_t^{\mathbb{P}} : \mathcal{X} \to [0, \infty]$ such that $\mathbb{E}^{\mathbb{P}}[S_t^{\mathbb{P}}(X) \mid X_1, \dots, X_{t-1}] \leq 1$
    Observe $X_t$
    Statistician's wealth in game $\mathbb{P}$ is updated as $W_t^{\mathbb{P}} = W_{t-1}^{\mathbb{P}} \cdot S_t^{\mathbb{P}}(X_t)$
    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.

6.6 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

$$E_t = \prod_{i=1}^{t} \frac{\hat{q}_{i-1}(X_i)}{\hat{p}_t(X_i)}, \tag{6.6}$$

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 6.8

The process E in (6.6) 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, $\hat{p}_n$ 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 $\mathcal{Q}$,

$$E_t = \int_{\mathcal{Q}} \prod_{i=1}^{t} \frac{q(X_i)}{\hat{p}_t(X_i)} \mathrm{d}\nu(q)$$

is an e-process. Without too much ambiguity, we refer to this also as being the universal inference e-process.

6.7 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 T is a finite or infinite set of indices starting from 0, and we further denote by T_+ the set $T \setminus \{0\}$.

Definition 6.9: Sequential e-values

The e-variables E_t with $t \in T_+$ for $\mathcal{P}$, adapted to the filtration $\mathcal{F}$, are *sequential* if $\mathbb{E}^{\mathbb{P}}[E_t|\mathcal{F}_{t-1}] \le 1$ for all $t \in T_+$ and $\mathbb{P} \in \mathcal{P}$. In case the filtration $\mathcal{F}$ is not specified, by default we meant the natural filtration; in other words, the required condition is $\mathbb{E}^{\mathbb{P}}[E_t \mid E_1, \dots, E_{t-1}] \le 1$ for all $t \in T_+$ 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, $t \in T_+$ for $\mathbb{P}$ are exact, then they must satisfy $\mathbb{E}^{\mathbb{P}}[E_t|\mathcal{F}_0] = 1$ for each $t \in T_+$ 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 6.10

Let $(E_t)_{t \in T_+}$ be a sequence (finite or infinite) of sequential e-variables for $\mathcal{P}$. Define

$$M_t = \prod_{s=1}^{t} E_s \text{ for } t \in T_+ \text{ and } M_0 = 1.$$

Then M is an e-process for $\mathcal{P}$.

The proof follows by noting that for $\mathbb{P} \in \mathcal{P}$, $\mathbb{E}^{\mathbb{P}}[M_t|\mathcal{F}_{t-1}] = M_{t-1}\mathbb{E}^{\mathbb{P}}[E_t|\mathcal{F}_t] \le M_{t-1}$, and thus $(M_t)_{t \in T}$ is a test supermartingale, sufficient for an e-process.

More generally, e-processes can be constructed from sequential e-variables through testing by betting discussed in Section 6.5. In Chapter 7, 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 6.11: Log-optimal and empirically adaptive e-processes

Let $(E_t)_{t \in T_+}$ be a sequence (finite or infinite) of sequential e-variables.

(i) An *e-process built on* $(E_t)_{t \in T_+}$ is $M = (M_t)_{t \in T}$ defined by

$$M_t = \prod_{s=1}^{t} \left((1 - \lambda_s) + \lambda_s E_s\right) \text{ for } t \in T_+ \text{ and } M_0 = 1, \tag{6.7}$$

where for $t \in T_+$, λ_t is any measurable function of $E_1, \dots, E_{t-1}$ (i.e., $(\lambda_t)_{t \in T_+}$ is predictable wrt $\mathcal{F}$) and takes value in $[0, 1]$.

(ii) For an alternative probability measure $\mathbb{Q}$, the $\mathbb{Q}$-*log-optimal e-process built on* $(E_t)_{t \in T_+}$ is the e-process in (6.7) such that λ_t for $t \in T_+$ solves the following equation

$$\lambda_t = \operatorname*{arg\,max}_{\lambda \in [0,1]} \mathbb{E}^{\mathbb{Q}} \left[\log\left((1 - \lambda) + \lambda E_t\right) \mid \mathcal{F}_{t-1}\right]. \tag{6.8}$$

(iii) For a parameter $\gamma \in (0, 1]$, the *empirically adaptive e-process* is the e-process in (6.7) such that $\lambda_1 = 0$ and λ_t for $t \ge 2$ solves the following equation

$$\lambda_t = \operatorname*{arg\,max}_{\lambda \in [0,\gamma]} \frac{1}{t-1} \sum_{s=1}^{t-1} \log\left((1 - \lambda) + \lambda E_s\right).$$

Similarly to the product process in Proposition 6.10, the construction (6.7) guarantees that M_t is a supermartingale.

To interpret the $\mathbb{Q}$-log-optimal e-process built on $(E_t)_{t\in T}$, at every step t, λ_t is chosen to maximize the e-power in Definition 2.11, 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 (6.7). If E_t is independent of $\mathcal{F}_{t-1}$ under $\mathbb{Q}$, then λ_t in (6.8) is given by maximizing the unconditional expectation,

$$\lambda_t = \underset{\lambda\in[0,1]}{\arg\max}\, \mathbb{E}^{\mathbb{Q}}\left[\log\left((1-\lambda)+\lambda E_t\right)\right],$$

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 $t\ge 2$, by Proposition 2.12, $\lambda_t = 0$ if and only if the empirical mean of $E_1,\dots,E_{t-1}$ is less than or equal to 1, that is, $\sum_{s=1}^{t} E_s \le t$. 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 6.12

Let $(E_t)_{t\in\mathbb{N}}$ be an infinite sequence of e-variables that is iid under the alternative probability $\mathbb{Q}$ such that $\mathbb{E}^{\mathbb{Q}}[\log E_1]$ is finite. The empirically adaptive martingale M with $\gamma = 1$ satisfies the following:

(i) Asymptotic log-optimality of M holds: for any e-process M' built on $(E_t)_{t\in\mathbb{N}}$, we have

$$\lim_{t\to\infty} \frac{1}{t}\left(\log M_t - \log M'_t\right) \ge 0 \quad \text{in } L^1(\mathbb{Q}).$$

(ii) Consistency of M holds: if $\mathbb{E}^{\mathbb{Q}}[E_1] > 1$, then $M_t \to \infty$ $\mathbb{Q}$-almost surely as $t\to\infty$.

We omit a formal proof of Theorem 6.12, 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 the $\mathbb{Q}$-log-optimal e-process built on $(E_t)_{t\in\mathbb{N}}$ by the law of large numbers. Thus, M has an asymptotic growth rate that is the same as the $\mathbb{Q}$-log-optimal e-process, which is optimal among all choices of e-processes built on $(E_t)_{t\in\mathbb{N}}$. Part (ii) follows by using Proposition 2.12 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 $t\to\infty$.

It should be clear that both statements in Theorem 6.12 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 $\mathbb{Q}$.

6.8 Obtaining an e-process from e-variables using a 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 $n\ge 1$, 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 $\mathbb{N} = \{1, 2, \ldots\}$ and define

$$M_n = \sum_{j \leq n} w(j) E^{(j)}. \tag{6.9}$$

with $M_0 = 1$ by default. Note that M is an increasing process, which is unusual for e-processes.

Proposition 6.13

The time-mixed process M defined in (6.9) 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),

$$\mathbb{E}^{\mathbb{P}}[M_\tau] = \mathbb{E}^{\mathbb{P}}\left[\sum_{n=1}^{\infty} \mathbb{1}_{\{\tau=n\}} \sum_{j \leq n} w(j) E^{(j)}\right] = \mathbb{E}^{\mathbb{P}}\left[\sum_{j \geq 1} w(j) E^{(j)} \sum_{n=j}^{\infty} \mathbb{1}_{\{\tau=n\}}\right].$$

Since $\sum_{n=j}^{\infty} \mathbb{1}_{\{\tau=n\}} \leq 1$ and each $E^{(j)}$ is an e-variable, we get

$$\mathbb{E}^{\mathbb{P}}[M_\tau] \leq \sum_{j \geq 1} w(j) \mathbb{E}^{\mathbb{P}}\left[E^{(j)}\right] \leq 1,$$

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 $X_1, \ldots, X_n$ (more precisely, that means testing $\mathcal{P}$ against $\mathbb{Q}$ conditioned on the σ-algebra generated by $X_1, \ldots, X_n$). In general, $E^{(1)}, E^{(2)}, \ldots$ 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,

$$\lim_{n \to \infty} \frac{1}{n} \log M_n - \frac{1}{n} \log E^{(n)} = 0$$

meaning that the asymptotic growth rate of our e-process M matches that of the sequence of numeraire e-variables.

Bibliographical note

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 [2023] 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].

The predictable process $(\lambda_t)_{t \in T_+}$ 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 6.12 is based on Wang et al. [2022] and its proof can be found in Wang et al. [2022, Theorem 3].

Chapter 7

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. The presence of multiple e-values arises in multiple testing as in Chapter 8 but it also can arise in single testing with e-values computed from different parts of the data or data splitting as in Chapters 4 and 6. In this chapter, we do not distinguish how they are computed, but focus on handling e-variables for some common hypothesis.

7.1 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. The cases of independent e-variables or those with some specific dependence structures are considered in the next section. In what follows, inequalities between functions are understood to hold everywhere.

Definition 7.1: Merging functions for e-values

(i) An *e-merging function* (of K e-values) is an increasing Borel function $F : [0, \infty)^K \to [0, \infty)$ 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 $\mathbf{e} \in [0, \infty)^K$,

$$G(\mathbf{e}) > 1 \Longrightarrow F(\mathbf{e}) \geq G(\mathbf{e}).$$

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 2.2, the defining property of F in Definition 7.1 (i) is required to hold for any hypothesis, but it suffices for F to satisfy the property for one atomless probability $\mathbb{P}$ 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 like $0 \times \infty$ 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 7.2. 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

$$\mathbb{M}_K(e_1, \ldots, e_K) = \frac{e_1 + \cdots + e_K}{K}, \qquad e_1, \ldots, e_K \in [0, \infty).$$

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,

$$\Delta_n = \left\{ (x_1, \ldots, x_n) \in [0,1]^n : \sum_{k=1}^{n} x_k = 1 \right\}.$$

It is straightforward to verify that for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{K+1}) \in \Delta_{K+1}$, the function

$$\mathbb{M}_{\boldsymbol{\lambda}} : (e_1, \ldots, e_K) \mapsto \sum_{k=1}^{K} \lambda_k e_k + \lambda_{K+1}, \tag{7.1}$$

is an e-merging function. We will refer to $\mathbb{M}_{\boldsymbol{\lambda}}$ as the $\boldsymbol{\lambda}$-weighted arithmetic mean, although strictly speaking it is a weighted average of the input vector $(e_1, \ldots, e_K)$ and the constant 1.

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 $\mathbb{M}_{\boldsymbol{\lambda}}$.

Proposition 7.3

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 $(e_1, \ldots, e_K) \in [0, \infty)^K$ such that

$$b := F(e_1, \ldots, e_K) > \max\left(\frac{e_1 + \cdots + e_K}{K}, 1\right) =: a. \tag{7.2}$$

Let Π_K be the set of all permutations of $\{1, \ldots, K\}$, π be randomly and uniformly drawn from Π_K, and $(D_1, \ldots, D_K) := (e_{\pi(1)}, \ldots, e_{\pi(K)})$. Further, let $(D'_1, \ldots, D'_K) := (D_1, \ldots, D_K)\mathbb{1}_A$, where A is an event independent of π and satisfying $P(A) = 1/a$ (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 $e_1, \ldots, e_K$ with equal probability, we have $\mathbb{E}^{\mathbb{P}}[D_k] = (e_1 + \cdots + e_K)/K$, which implies $\mathbb{E}^{\mathbb{P}}[D'_k] = (e_1 + \cdots + e_K)/(Ka) \leq 1$. Together with the fact that D'_k is nonnegative, we know D'_k is an e-varaible for $\mathbb{P}$. Moreover, by symmetry,

$$\mathbb{E}[F(D'_1, \ldots, D'_K)] = Q(A)F(e_1, \ldots, e_K) + (1 - Q(A))F(0, \ldots, 0) \geq b/a > 1,$$

a contradiction. Therefore, we conclude that there is no $(e_1, \ldots, e_K)$ such that (7.2) holds.

In particular, Proposition 7.3 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., the geometric mean or the minimum), then F is dominated by $\mathbb{M}_K$.

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 $\mathbb{M}_{\boldsymbol{\lambda}}$ in (7.1) is admissible and that the class (7.1) is precisely the class of all admissible e-merging functions. Moreover, every e-merging function is dominated by one of (7.1).

Theorem 7.4

For a function $F : \mathbb{R}_+^K \to \mathbb{R}_+$,

(i) if F is an e-merging function, then $F \le \mathbb{M}_{\boldsymbol{\lambda}}$ for some $\boldsymbol{\lambda} \in \Delta_{K+1}$;

(ii) F is an admissible e-merging function if and only if $F = \mathbb{M}_{\boldsymbol{\lambda}}$ for some $\boldsymbol{\lambda} \in \Delta_{K+1}$.

The proof of Theorem 7.4 is put in Chapter 9, as it requires several other advanced technical results.

7.2 Independent e-values and their products

In this section we consider merging functions for independent e-values.

Definition 7.5: Merging functions for independent e-values

(i) An *ie-merging function* is a Borel function $F : [0,\infty)^K \to [0,\infty)$ 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 $e_1,\dots,e_K$,

$$(e_1,\dots,e_K) \in [1,\infty)^K \implies F(e_1,\dots,e_K) \ge G(e_1,\dots,e_K).$$

The corresponding definitions of domination, strict domination, and admissibility are obtained from Definition 7.1 by replacing “e-merging” with “ie-merging”; this also applies to the later definition of se-merging functions.

Similarly to the arithmetic mean for arbitrarily dependent e-values, there is an important ie-merging function: the *product* Π_K, defined by

$$\Pi_K(e_1,\dots,e_K) = \prod_{k=1}^K e_k, \qquad e_1,\dots,e_K \in [0,\infty). \tag{7.3}$$

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, we are also interested in the case where some of the input e-values are below 1.

The following proposition gives a simple condition for admissibility.

Proposition 7.6

Fix an atomless probability measure $\mathbb{P} \in \Pi$. For a Borel function $F : [0,\infty)^K \to [0,\infty)$, 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).

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

$$\mathbb{E}^{\mathbb{P}}[G(E_1, \dots, E_K)] > \mathbb{E}^{\mathbb{P}}[G(E_1, \dots, E_K)] = 1,$$

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 2.2. Fisher's method will be discussed in Chapter 9. The ie-merging function Π_K is admissible by Proposition 7.6. More generally, we can see that the U-statistic functions U_n, defined by

$$U_n(e_1, \dots, e_K) = \frac{1}{\binom{K}{n}} \sum_{\{k_1, \dots, k_n\} \subseteq \{1, \dots, K\}} e_{k_1} \dots e_{k_n}, \quad n \in \{0, 1, \dots, K\}, \tag{7.4}$$

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 7.6 implies that the U-statistic functions in (7.4) and their convex combinations are admissible ie-merging functions.

Let us now establish a very weak counterpart of Proposition 7.3 for independent e-values; on the positive side it will not require the assumption of symmetry needed for Proposition 7.3.

Proposition 7.7

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 $k \in \{1, \dots, K\}$ 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 (7.4) form a complete class of admissible ie-merging functions. They do not, as shown by the following example.

Example 7.8

Define a function $f : [0,\infty)^2 \to [0,\infty)$ by

$$f(e_1,e_2) = \frac{1}{2}\left(\frac{e_1}{1+e_1} + \frac{e_2}{1+e_2}\right)(1+e_1e_2.)$$

Using Proposition 7.6, one can check that f is an admissible ie-merging function. Nevertheless, it is different from any convex mixture of (7.4).

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

$$\mathbb{P}\left(\bigcap_{k\in[K]}\{E_k \le x_k\}\right) \le \prod_{k\in[K]} \mathbb{P}(E_k \le x_k) \quad \text{for all } (x_1,\dots,x_K) \in \mathbb{R}^K, \tag{7.5}$$

then the product Π_K and any U-statistic functions in (7.4) of these e-variables are also e-variables. Clearly, if the central inequality in (7.5) 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.

7.3 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 7.9: Sequential e-values and their merging functions

(i) The e-variables $E_1,\dots,E_K$ for $\mathcal{P}$ are *sequential* if $\mathbb{E}^{\mathbb{P}}[E_k \mid E_1,\dots,E_{k-1}] \le 1$ for all $k\in[K]$ and $\mathbb{P}\in\mathcal{P}$.

(ii) An *se-merging function* is a Borel function $F : [0,\infty)^K \to [0,\infty)$ 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, for any hypothesis, $F(E_1,\dots,E_K)$ is an exact e-variable for all sequential exact e-variables $E_1,\dots,E_K$.

One subtle distinction of the above definition from those in the previous sections is that we no longer require increasing monotonicity in all arguments of the merging function. This relaxation is quite natural under the betting interpretation (see Chapter 6): 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.

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 6.

It is straightforward to check that all convex mixtures of (7.4) are se-merging functions. Since independent e-variables are sequential, se-merging functions also produce a valid e-variable for independent e-variables. Example 7.8 gives an ie-merging function but not an se-merging function.

Among se-merging functions, the convex mixtures of (7.4) are admissible. They are also admissible in the class of ie-merging functions by Proposition 7.6. We also note that it suffices for $E_1,\dots,E_K$ to be sequential in any order for these merging methods to be valid.

A particular class of se-merging functions, larger than (7.4), is defined below.

Definition 7.10: Martingale merging functions

An function $F : [0,\infty)^K \to [0,\infty)$ is a *martingale merging function* if

$$F(e_1,\dots,e_K) = \prod_{k=1}^{K}(1-\lambda_k+\lambda_k e_k) \tag{7.6}$$

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 (7.4) are indeed all martingale merging functions. We leave these as exercises for the reader.

A martingale merging function in (7.6) 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 $(E_t)_{t\in T_+}$ in Definition 6.11 is constructed by merging e-values with a martingale merging function in Definition 7.10 for each fixed t and also in a consistent way across different t.

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 7.11

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

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[F(E_1,\dots,E_K) \mid \mathcal{F}_{K-1}] &= F(E_1,\dots,E_{K-1},1)(1-L_K+L_K\mathbb{E}^{\mathbb{P}}[E_K|\mathcal{F}_{K-1}]) \\ &\leq F(E_1,\dots,E_{K-1},1).\end{aligned}$$

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 $t_k : [0,\infty)^k \to [0,\infty)$, $k \in [K]$, that for each $k \in [K]$,

$$F(e_1,\dots,e_k,1,\dots,1) = F(e_1,\dots,e_{k-1},1,\dots,1) + t_k(e_1,\dots,e_{k-1})(e_k-1),$$

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 $(t_k)_{k\in[K]}$.

The next result shows that the class of martingale merging functions is precisely that of all admissible se-merging functions.

Theorem 7.12

Any se-merging function is dominated by a martingale merging function.

The proof of Theorem 7.12 is quite technical and not pursued in this book.

Different from the merging functions obtained in Sections 7.1 and 7.2, 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 $k \in [K]$ 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 7.13: 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 $k \in [K]$, 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 (7.6) that F is not increasing. This strategy corresponds to stopping an e-process in Chapter 6 as soon as it can make a rejection at level α by up-shooting $1/\alpha$.

Example 7.14: Merging via the empirically adaptive e-process

The empirically adaptive e-process in Definition 6.11 induces a martingale merging function. For a parameter $\gamma \in (0, 1]$, the martingale merging function is defined as by (7.6) with $\lambda_1 = 0$ and λ_k for $k \in \{2, \dots, K-1\}$ are given by

$$\lambda_k(e_1, \dots, e_{k-1}) = \operatorname*{arg\,max}_{\lambda \in [0,\gamma]} \frac{1}{k-1} \sum_{s=1}^{k-1} \log\left((1-\lambda) + \lambda e_s\right).$$

This martingale merging function is uniquely determined by its parameter γ.

Advantages of the empirically adaptive martingale merging function in Example 7.14 are justified in Theorem 6.12 via its corresponding e-process. This function can be used as a default choice without prior information among se-merging functions.

7.4 Mean-variance trade-off

We now pay special attention to the product function Π_K in (7.3), which is both an ie-merging function and an se-merging function. Proposition 7.6 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 7.1, 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 7.15

For any ie-merging function $F : [0, \infty)^K \to [0, \infty)$ 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 $\mathbb{E}^{\mathbb{P}}[G(E_1, \dots, E_K)] \le 1$ 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

$$\mathbb{E}^{\mathbb{P}}[G(E_1, \dots, E_K)] = \frac{1}{a}\mathbb{E}^{\mathbb{P}}[F(aE_1, E_2, \dots, E_K)] \le \mathbb{E}^{\mathbb{P}}[F(E_1', E_2, \dots, E_K)] \le 1,$$

where the first inequality following from $F \ge 0$. 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 nontrivial against $\mathbb{Q}$.

Proposition 7.16

Let $F : [0, \infty)^K \to [0, \infty)$ be an se-merging function. For independent nonnegative random variables $E_1, \dots, E_K$ under $\mathbb{Q}$ with $\mathbb{E}^{\mathbb{Q}}[E_k] \ge 1$, $k \in [K]$, we have, for every $m \in \mathbb{N}$,

$$\mathbb{E}^{\mathbb{Q}}\left[F(E_1, \dots, E_K)^m\right] \le \prod_{k=1}^{K} \mathbb{E}^{\mathbb{Q}}[E_k^m] = \mathbb{E}^{\mathbb{Q}}\left[\Pi_K(E_1, \dots, E_K)^m\right]. \tag{7.7}$$

In particular, if F is exact and $E_1, \dots, E_K$ are independent exact e-variables for $\mathbb{P}$, then

$$\mathrm{Var}^{\mathbb{P}}(F(E_1, \dots, E_K)) \le \mathrm{Var}^{\mathbb{P}}(\Pi_K(E_1, \dots, E_K)). \tag{7.8}$$

Proof.

First, we argue that it suffices to show (7.7) for $E_1, \dots, E_K$ being exact e-variables for $\mathbb{Q}$. Suppose that this condition holds true. For independent nonnegative $X_1, \dots, X_K$ with mean larger than or equal to 1, set $a_k := \mathbb{E}^{\mathbb{P}}[X_k] \ge 1$, $E_k := X_k/a_k$ for $k \in [K]$, and $a := \prod_{k=1}^{K} a_k$. Let $G : (e_1, \dots, e_K) \mapsto F(a_1e_1, \dots, a_Ke_K)/a$. Clearly, $E_1, \dots, E_K$ are independent exact e-variables. Using Lemma 7.15 repeatedly, we deduce that G is an se-merging function. For $m \in \mathbb{N}$, if (7.7) holds for all exact e-variables, then

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}}\left[F(X_1, \dots, X_K)^m\right] &= a^m\mathbb{E}^{\mathbb{P}}\left[\left(\frac{F(a_1E_1, \dots, a_KE_K)}{a}\right)^m\right] \\
&= a^m\mathbb{E}^{\mathbb{P}}\left[G(E_1, \dots, E_K)^m\right] \\
&\le a^m\mathbb{E}^{\mathbb{P}}\left[\Pi_K(E_1, \dots, E_K)^m\right] \\
&= \mathbb{E}^{\mathbb{P}}\left[\Pi_K(X_1, \dots, X_K)^m\right].
\end{aligned}$$

Therefore, the general case of (7.7) follows from the case of exact e-variables.

Let $E_1, \dots, E_K$ be independent exact e-variables; we will show

$$\mathbb{E}^{\mathbb{Q}}\left[F(E_1, \dots, E_K)^m\right] \le \prod_{k=1}^{K} \mathbb{E}^{\mathbb{Q}}[E_k^m]. \tag{7.9}$$

Using Theorem 7.12, 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 (7.9) 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

$$\mathbb{E}^{\mathbb{Q}}[G(E_1,\dots,E_{K-1})^m] \leq \prod_{k=1}^{K-1} \mathbb{E}^{\mathbb{Q}}[E_k^m] \tag{7.10}$$

for every se-merging function $G : [0,\infty)^{K-1} \to [0,\infty)$. Since F is a martingale merging function, we can write

$$F(e_1,\dots,e_K) = G(e_1,\dots,e_{K-1})\left(1-\lambda_K(e_1,\dots,e_{K-1}) + \lambda_K(e_1,\dots,e_{K-1})e_K\right),$$

for some $\lambda_K : [0,\infty)^{K-1} \to [0,1]$. Let us write $\mathbf{Y} = (E_1,\dots,E_{K-1})$ and $L = \lambda_K(E_1,\dots,E_{K-1})$. We have

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}[F(E_1,\dots,E_K)^m \mid \mathbf{Y}] &= \mathbb{E}^{\mathbb{Q}}[G(\mathbf{Y})^m(1-L+LE_K)^m \mid \mathbf{Y}] \\
&\leq G(\mathbf{Y})^m \left(1-L+L\mathbb{E}^{\mathbb{Q}}[E_K^m|\mathbf{Y}]\right) \\
&\leq G(\mathbf{Y})^m \mathbb{E}^{\mathbb{Q}}[E_K^m].
\end{aligned}$$

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 (7.10). Therefore, we obtain (7.7). If F is an exact se-merging function, we obtain (7.8) from (7.9) 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 7.16 has two implications. First, under the alternative $\mathbb{Q}$, if the e-variables are nontrivial against $\mathbb{Q}$, 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 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 (which is likely the case for a large second moment) 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 6.7, 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 7.17 below.

Example 7.17: All-in versus log-optimal

We consider the e-processes built on $(E_t)_{t\in\mathbb{N}}$ as in Definition 6.11, allowing an infinite number of e-variables. Suppose that under $\mathbb{Q}$, each E_t for $t \in \mathbb{N}$ is distributed as $\mathbb{Q}(E_t = 4) = \mathbb{Q}(E_t = 0) = 1/2$ and they are independent.

(a) The all-in strategy is to choose $\lambda_t = 1$ for all t, 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 built on $(E_t)_{t\in[k]}$ in Definition 6.11 M_k^* has $\lambda_t = 1/3$ for all t, leading to $\mathbb{E}^{\mathbb{Q}}[\log(1 - \lambda_t + \lambda_t E_t)] = \log(4/3)$. Then, by the law of large numbers, $M_k^* \sim (4/3)^k$ $\mathbb{Q}$-almost surely as $k \to \infty$. (This asymptotic growth rate is also achieved by the empirically adaptive e-process by Theorem 6.12.)

One can see that, although $\mathbb{E}^{\mathbb{Q}}[M_k^{\text{all-in}}] \ge \mathbb{E}^{\mathbb{Q}}[M_k^*]$ for all $k \in \mathbb{N}$ (indeed, $M_k^{\text{all-in}}$ has a largest expectation among all e-processes built on $(E_t)_{t\in[k]}$ by Proposition 7.16), its distribution has a large chance of getting 0, and moreover $M_k^{\text{all-in}} \to 0$ $\mathbb{Q}$-almost surely. These are highly undesirable features of all-in betting strategy.

7.5 Summary

The following points summarize the results in the previous few sections.

(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 7.14 can be used as a default method without prior information on the e-values' behaviour 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.

Bibliographical note

The content in this chapter is mainly based on Vovk and Wang [2021] and Vovk and Wang [2024a]. E-merging functions were introduced by Vovk and Wang [2021]. The proof of Theorem 7.12 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].

Chapter 8

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 as follows.

Suppose we observe data X drawn according to some unknown probability distribution $\mathbb{P}^*$. (Here, X denotes the entire dataset available to the researcher.) Let $\mathcal{P}_1, \ldots, \mathcal{P}_K$ be K sets of probability distributions (subsets of $\mathcal{M}_1$). The k-th null hypothesis that is being tested is defined as

$$H_k : \mathbb{P}^* \in \mathcal{P}_k.$$

Some of these nulls are true (meaning that $\mathbb{P}^*$ is indeed in $\mathcal{P}_k$), while the others are false, and apriori any configuration of true and false nulls is possible. We write $\mathcal{N} = \mathcal{N}(\mathbb{P}^*)$, the unknown set of true null hypotheses. We denote by K_0 the number of true null hypotheses, that is, $K_0 = |\mathcal{N}|$. Finally, let $\mathcal{K} = [K]$ and

$$\mathcal{P} = \bigcup_{k \in \mathcal{K}} \mathcal{P}_k.$$

Note that it is possible $\mathbb{P}^* \in \mathcal{M}_1 \backslash \mathcal{P}$, meaning that the hypotheses could all be false (or all be true, or anything in between).

8.1 Compound e-variables and the e-BH procedure

Compound e-variables

Each hypothesis is often 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 $(E_1, \ldots, E_K)$ to satisfy a more relaxed definition, given below.

Definition 8.1: Compound e-variables

For $k \in \mathcal{K}$, let E_k be a $[0, \infty]$-valued random variable. We say that $E_1, \ldots, E_K$ are compound e-variables for $(\mathcal{P}_1, \ldots, \mathcal{P}_K)$ if for every $\mathbb{P} \in \mathcal{P}$, we have

$$\sum_{k : \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k] \leq K.$$

They are called *tight* if the supremum of the left hand side over $\mathbb{P} \in \mathcal{P}$ equals K.

When $K = 1$, this 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 8.2: Weighted e-values

If $E_1, \dots, E_K$ are e-values, then they are also compound e-values. 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-values.

Note that the class of compound e-values is much richer than the class of weighted e-values in Example 8.2, because the later satisfies the stronger constraint

$$\sum_{k \in \mathcal{K}} \sup_{\mathbb{P} \in H_k} \mathbb{E}^{\mathbb{P}}[\tilde{E}_k] \leq K,$$

compared to the condition for the compound e-values

$$\sup_{\mathbb{P} \in \bigcup_{k \in \mathcal{K}} H_k} \sum_{k \in \mathcal{N}(\mathbb{P})} \mathbb{E}^{\mathbb{P}}[E_k] \leq K.$$

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 8.3

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 $k = 1, \dots, K$, 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; $X_{k,j} > 1$ presents a profit and $X_{k,j} < 1$ means a loss) of agent k from month 1 to month n. The no-skill null hypothesis is $\mathbb{E}[X_{k,j} \mid \mathcal{F}_{j-1}] \leq 1$ for $j = 1, \dots, n$, 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 $k = 1, \dots, K$. 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, $E_k = \prod_{j=1}^{n} X_{k,j}$ 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 6.9, and they can be combined using merging methods in Chapter 7. 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 6.

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 $\bar{c} = (c_1 + \cdots + c_K)/K$, 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.

The e-BH procedure

We first define the FDR. We denote a multiple testing procedure by $\mathcal{D}$, that is, a Borel function of X that produces a subset of $\mathcal{K}$ representing the indices of rejected hypotheses. The hypotheses that are rejected by $\mathcal{D}$, given by $\mathcal{D}(X) \in 2^{\mathcal{K}}$, are called discoveries. We overload notation and call 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 write $F_{\mathcal{D}} = |\mathcal{N} \cap \mathcal{D}|$ as the number of true null hypotheses that are rejected (i.e., false discoveries), and $R_{\mathcal{D}} := |\mathcal{D}|$ as the total number of discoveries. The FDR is the expected value of the false discovery proportion (FDP), that is,

$$\mathrm{FDR}_{\mathcal{D}} := \mathbb{E}^{\mathbb{P}^*}[F_{\mathcal{D}}/R_{\mathcal{D}}]$$

with the specification $0/0 = 0$. Clearly, $\mathrm{FDR}_{\mathcal{D}}$ depends on the unknown $\mathbb{P}^*$, but we will study procedures that control the FDR for all possible $\mathbb{P}^*$, either inside $\mathcal{P}$ or outside it. A testing procedure $\mathcal{D}$ that takes e-values or compound e-values as input is called an e-testing procedure.

The FDR has been the most popular metric for false discovery 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 8.4: The e-BH procedure

Suppose that $e_1, \dots, e_K$ are realized compound e-values for the hypotheses $H_1, \dots, H_K$. 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

$$k^* := \max\left\{k \in \mathcal{K} : \frac{k e_{[k]}}{K} \geq \frac{1}{\alpha}\right\},$$

with the convention $\max(\varnothing) = 0$.

Before showing the FDR guarantee of the e-BH procedure, we slightly enlarge the scope of our methodology. The e-BH procedure belongs to be larger class of self-consistent e-testing procedures.

Definition 8.5: 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 satisfying

$$e_k \geq \frac{K}{\alpha R_{\mathcal{D}}}.$$

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 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 8.6: 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 $\alpha K_0/K$. 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

$$\frac{F_{\mathcal{D}_\alpha}}{\alpha R_{\mathcal{D}_\alpha}} = \frac{|\mathcal{D}_\alpha(\mathbf{E}) \cap \mathcal{N}|}{\alpha(R_{\mathcal{D}_\alpha} \vee 1)} = \sum_{k\in\mathcal{N}} \frac{\mathbb{1}_{\{k\in\mathcal{D}_\alpha(\mathbf{E})\}}}{\alpha(R_{\mathcal{D}_\alpha} \vee 1)} \leq \sum_{k\in\mathcal{N}} \frac{\mathbb{1}_{\{k\in\mathcal{D}_\alpha(\mathbf{E})\}} E_k}{K} \leq \sum_{k\in\mathcal{N}} \frac{E_k}{K},$$

where the first inequality is due to self-consistency. By the definition of compound e-variables, we immediately obtain that

$$\mathbb{E}^{\mathbb{P}^*}\left[\sup_{\alpha\in(0,1)} \frac{F_{\mathcal{D}_\alpha}}{\alpha R_{\mathcal{D}_\alpha}}\right] \leq 1.$$

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 8.1.

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 8.1.

Proposition 8.7

Let $e_1, \dots, e_k \in \mathbb{R}_+$ and $\alpha \in (0,1)$. Define

$$R(t) = |\{k \in \mathcal{K} : e_k \geq t\}| \vee 1 \quad \text{and} \quad t_\alpha = \inf\{t \in [0,\infty) : tR(t) \geq K/\alpha\}.$$

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 $e'_k \geq t_\alpha$.

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 $T : [0, \infty] \to [0, K]$ by letting $T(x)$ be the largest number in $K/\mathcal{K} \cup \{0\}$ that is no larger than x. In other words,

$$T(x) = \frac{K}{\lceil K/x \rceil} \mathbb{1}_{\{x \geq 1\}} \quad \text{with } T(\infty) = K. \tag{8.1}$$

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* $b_k \geq 1$ is such that

$$\mathbb{E}^{\mathbb{P}}[T(\alpha b_k E_k)] \leq \alpha \quad \text{for } \mathbb{P} \in \mathcal{P}_k, \tag{8.2}$$

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 (8.2) because $T(x) \leq x$ for all $x \in \mathbb{R}_+$. One would try to choose the largest b_k possible subject to (8.2). 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 (8.2). 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 8.8

Applied to arbitrary non-negative random variables $E'_1, \dots, E'_K$, the e-BH procedure $\mathcal{D}$ at level $\alpha \in (0, 1)$ satisfies

$$\mathbb{E}^{\mathbb{P}^*}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right] \leq \frac{1}{K} \sum_{k \in \mathcal{N}} \mathbb{E}^{\mathbb{P}^*}\left[T(\alpha E'_k)\right],$$

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 (8.2), then the FDR is controlled at $K_0\alpha/K$.

The proof of Theorem 8.8 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 8.7, e-values are rejected only at thresholds in $\alpha^{-1}(K/\mathcal{K})$. Therefore, for the desired FDR control, it suffices to verify that $T(\alpha E_k)/\alpha$ are e-values. This leads to the condition (8.2), which justifies 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.

Connection between the e-BH and the BH procedures

In this section, we illustrate an important point that connects the e-BH procedure to the BH procedure with arbitrary dependence correction.

Definition 8.9: The BH procedure

Suppose that $p_1, \dots, p_K$ are realized p-values for the hypotheses $H_1, \dots, H_K$. 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

$$k^* = \max\left\{k \in \mathcal{K} : \frac{K p_{(k)}}{k} \le \alpha\right\},$$

with the convention $\max(\varnothing) = 0$. The *BH procedure with arbitrary dependence correction* 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 on the FDR control of the BH procedure.

(i) 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.

(ii) The BH procedure at level α with arbitrary dependence correction controls FDR at level $(K_0/K)\alpha$.

Definition 8.10: 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 $\mathbf{y} \ge \mathbf{x}$.

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 $x \mapsto \mathbb{P}\{(P_\ell)_{\ell \in \mathcal{K}} \in A \mid P_k \le x\}$ is increasing on $[0,1]$.

A caveat of PRDS in Definition 8.10 is that it enforces certain positive dependence between the nulls and non-nulls.

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 arbitrary dependence 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 H_k (Proposition 2.3), 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 arbitrary dependence correction under such a setting.

The main result in this section shows that the BH procedure with arbitrary dependence correction can be recovered by applying the e-BH procedure to some e-values.

Proposition 8.11

(i) With T in (8.1) 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 $H_1, \dots, H_K$. The BH procedure with arbitrary dependence correction 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

$$\mathbb{E}^{\mathbb{P}}[f(U)] = \mathbb{E}^{\mathbb{P}}\left[\frac{1}{\alpha}T\left(\frac{\alpha}{\ell_K U}\right)\right] = \frac{1}{\alpha}\sum_{k=1}^{K}\frac{K}{k}\mathbb{P}\left(\frac{\alpha(k-1)}{K\ell_K} \le U < \frac{\alpha k}{K\ell_K}\right) = \frac{1}{\alpha}\sum_{k=1}^{K}\frac{\alpha}{\ell_K k} = 1.$$

This shows that f is a calibrator.

(ii) As we explained in Section 8.1, 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 $1/(\ell_K P_1), \dots, 1/(\ell_K P_K)$. This is further equivalent to applying the BH procedure to $\ell_K P_1, \dots, \ell_K P_K$, thus the BH procedure with arbitrary dependence correction.

As a direct consequence of Proposition 8.11, the FDR guarantee of the BH procedure with arbitrary dependence correction directly follows from Theorem 8.6.

There is a different boosting method for the e-BH procedure under the PRDS condition, which we omit here. This boosting method allows one to convert the e-BH procedure into the BH procedure with the same FDR guarantee under PRDS, similarly to Proposition 8.11.

8.2 Universality of compound e-values and e-BH

From FDR control to compound e-values

The following result enables the construction of (approximate/asymptotic) compound e-values from any (approximate/asymptotic) FDR controlling procedure.

Theorem 8.12: From FDR control to compound e-values

Let $\mathcal{D}$ be any procedure that controls the FDR at a known level $\alpha \in (0, 1)$ for the hypotheses $H_1, \dots, H_K$ for every choice of $\mathbb{P}^* \in \mathcal{P}$. Let $V_k \in \{0, 1\}$ be an indicator of whether 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

$$E_k = \frac{K}{\alpha}\frac{V_k}{R_{\mathcal{D}} \vee 1}. \tag{8.3}$$

Then, $E_1, \dots, E_K$ are compound e-variables for $(\mathcal{P}_1, \dots, \mathcal{P}_K)$. Further, if FDR control only holds (ε, δ)-approximately or asymptotically (as the size of X or K grows to infinity), then $E_1, \dots, E_K$ are (ε, δ)-approximate or asymptotic compound e-values.

Proof.

The result can be seen from, for every $\mathbb{P} \in \mathcal{P}$,

$$\sum_{k:\mathbb{P}\in\mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k] = \sum_{k:\mathbb{P}\in\mathcal{P}_k} \mathbb{E}^{\mathbb{P}}\left[\frac{K}{\alpha}\frac{V_k}{R_{\mathcal{D}} \vee 1}\right] = \frac{K}{\alpha}\mathbb{E}^{\mathbb{P}}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}} \vee 1}\right] \le \frac{K}{\alpha}\alpha = K,$$

where the inequality follows from the definition of FDR guarantee. The proof of the asymptotic claim follows analogously.

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 8.13: Universality of e-BH with compound e-values

Let $\mathcal{D}$ be any procedure that controls the FDR at level α for the hypotheses $H_1, \dots, H_K$ for every choice of $\mathbb{P}^* \in \mathcal{P}$. 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 Theorems 8.12 and 8.22. 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 (8.3). 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 $E'_1 = \cdots = E'_K = 1$ 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 $H_1, \dots, H_K$, 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 8.9, it controls FDR if we assume 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 (8.3), but these e-variables are not necessarily independent or PRDS.

Combination and derandomization

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 $H_1, \dots, H_K$, suppose we run L different multiple testing procedures $\mathcal{D}_1, \dots, \mathcal{D}_L$ (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 for any $\mathbb{P}^* \in \mathcal{P}$ (and hence for any $\mathbb{P}^* \in \mathcal{M}_1$, because for $\mathbb{P}^* \in \mathcal{M}_1 \backslash \mathcal{P}$, all nulls are false, so the FDR equals zero). Then we may proceed as follows:

1. For the ℓ-th multiple testing procedure, form $E_1^{(\ell)}, E_K^{(\ell)}$ which are compound e-variables for $\mathcal{P}_1, \dots, \mathcal{P}_K$ (which always exist by Theorem 8.12). They could be — but need not necessarily be — the implied ones formed in Theorem 8.12.
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 8.21, 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 suffice 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 to the new compound e-values $E_1, \dots, E_K$ at level α.

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 8.12).

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 $L \to \infty$).

8.3 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 8.14: 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 8.3) and then we provide two constructions of (approximate) compound e-values motivated by compound decision theory (Sections 8.3 and 8.3).

Background: compound decision theory

Let $\mathbb{P}_{\boldsymbol{\mu}}$ be the probability measure governing a Gaussian sequence model:

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K) \text{ fixed}, \quad X_k \sim \mathrm{N}(\mu_k, 1) \text{ for } k \in \mathcal{K}. \tag{8.4}$$

Suppose we are interested in constructing estimators $\hat{\mu}_k$ of μ_k such that the following expected compound loss would be small:

$$\frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\mu}}[(\hat{\mu}_k - \mu_k)^2], \tag{8.5}$$

where $\mathbb{E}^{\boldsymbol{\mu}}$ means $\mathbb{E}^{\mathbb{P}_{\boldsymbol{\mu}}}$. If μ_k has a prior distribution B, then, by denoting $\mathbb{P}_B$ the joint distribution of $\boldsymbol{\mu}$ 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 8.14.

Now suppose $\boldsymbol{\mu}$ is deterministic as in (8.4). Given knowledge of $\boldsymbol{\mu}$, which function $s_{\boldsymbol{\mu}} : \mathbb{R} \to \mathbb{R}$ leads to a simple separable estimator $\hat{\mu}_k^{s_{\boldsymbol{\mu}}} = s_{\boldsymbol{\mu}}(X_k)$ that minimizes the risk in (8.5)? The answer lies in the fundamental theorem of compound decisions, which formally connects (8.4) to the univariate Bayesian problem with prior $M = \sum_{k \in \mathcal{K}} \delta_{\mu_k}/K$ equal to the empirical distribution of $\boldsymbol{\mu}$ (with δ_{μ_k} denoting the Dirac measure at μ_k):

$$\mu' \sim M, \quad X' \mid \mu' \sim \mathrm{N}(\mu', 1). \tag{8.6}$$

The fundamental theorem of compound decision theory states the following. Given any fixed $s : \mathbb{R} \to \mathbb{R}$, we have that

$$\frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\mu}}[(s(X_k) - \mu_k)^2] = \sum_{k \in \mathcal{K}} \frac{1}{K} \mathbb{E}^{\mu_k}[(s(X_k) - \mu_k)^2] = \mathbb{E}^{\mathbb{P}_M}[(s(X') - \mu')^2]. \tag{8.7}$$

In the left-hand side display, $\boldsymbol{\mu} \in \mathbb{R}^K$ is treated as fixed (as in (8.4)), 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 (8.6). From (8.7) it immediately follows that the optimal simple separable estimator is the Bayes estimator under M, that is $s(x) = \mathbb{E}^{\mathbb{P}_M}[\mu' \mid X' = x]$.

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 $\boldsymbol{\mu}$, while the feasible estimator is non-separable but must work without knowledge of $\boldsymbol{\mu}$.

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 dν-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 $\mathbb{P}^* = \bigtimes_{k \in \mathcal{K}} \mathbb{P}_k$).

For $k \in \mathcal{K}$, we let $\mathbb{Q}_k$ be distributions on $\mathcal{Y}$ with dν-densities q_k, representing the alternative distributions. We seek to solve the following optimization problem:

$$\begin{aligned} \underset{s(\cdot):\mathcal{Y} \to [0,\infty]}{\text{maximize}} \quad & \frac{1}{K} \sum_{k \in \mathcal{K}} \mathbb{E}^{\mathbb{Q}_k}[\log(s(X_k))] \\ \text{subject to} \quad & s(X_1), \ldots, s(X_K) \text{ are compound e-variables for } \mathcal{P}_1 \ldots, \mathcal{P}_K. \end{aligned} \tag{8.8}$$

Theorem 8.15: Log-optimal simple separable compound e-values

The optimal solution to optimization problem (8.8) is given by the likelihood ratio of mixtures over the null and alternative:

$$s(x) = \frac{\sum_{j \in \mathcal{K}} q_j(x)}{\sum_{j \in \mathcal{K}} p_j(x)}. \tag{8.9}$$

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 $s(X_1), \ldots, s(X_K)$ are compound e-values for $\mathbb{P}_1, \ldots, \mathbb{P}_K$.

Define the mixture distributions $\overline{\mathbb{P}} = \sum_{k\in\mathcal{K}} \mathbb{P}_k/K$ and $\overline{\mathbb{Q}} = \sum_{k\in\mathcal{K}} \mathbb{Q}_k/K$. Then, for any fixed $s : \mathcal{Y} \to [0,\infty]$, we have the following equalities:

$$\frac{1}{K}\sum_{k\in\mathcal{K}} \mathbb{E}^{\mathbb{Q}_k}[\log(s(X_k))] = \mathbb{E}^{\overline{\mathbb{Q}}}[\log(s(X))], \quad \frac{1}{K}\sum_{k\in\mathcal{K}} \mathbb{E}^{\mathbb{P}_k}[s(X_k)] = \mathbb{E}^{\overline{\mathbb{P}}}[s(X)],$$

where X is a generic random variable following the distribution specified with the expectation. Thus, it suffices to solve the following optimization problem:

$$\begin{aligned} &\underset{s:\mathcal{Y}\to[0,\infty]}{\text{maximize}} && \mathbb{E}^{\overline{\mathbb{Q}}}[\log(s(X))] \\ &\text{subject to} && \mathbb{E}^{\overline{\mathbb{P}}}[s(X)] \leq 1. \end{aligned}$$

As seen in Chapter 3, this has solution given by (8.9). Finally we may verify in that this s indeed yields compound e-values for $\mathbb{P}_1, \ldots, \mathbb{P}_K$. To this end, write $\overline{q}$ for the dν density of $\overline{\mathbb{Q}}$. Then, we have the following:

$$\sum_{k:\mathbb{P}^*\in\mathcal{P}_k} \mathbb{E}^{\mathbb{P}^*}[E_k] = \sum_{k:\mathbb{P}^*\in\mathcal{P}_k} \int \frac{K\overline{q}(x)}{\sum_{j\in\mathcal{K}} p_j(x)} p_k(x)\nu(\mathrm{d}x) \leq \int K\overline{q}(x)\nu(\mathrm{d}x) = K,$$

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 (8.9) for some densities $q_1, \ldots, q_K$.

Proposition 8.16

Suppose that $E_1, \ldots, E_K$ are tight simple separable compound e-values (for the same nulls defined as in the previous parts of this section). Then there exist dν-densities $q_1, \ldots, q_K$ such that E_k may be represented as in (8.9) on the support of $\sum_{k\in\mathcal{K}} \mathbb{P}_k/K$ and it is without loss of generality to take $q_1 = \cdots = q_K$.

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 $\overline{p}$ as the dν density of $\overline{\mathbb{P}}$, where the latter object is defined as in Theorem 8.15. Then the above may be written as:

$$\int s(x)\overline{p}(x)\mathrm{d}\nu(x) = 1.$$

Thus, if we define $q_k(x) = s(x)\overline{p}(x)$ for all $k \in \mathcal{K}$, we find that q_k is indeed a dν-density and $s(x)$ may be represented as in (8.9).

The objective value of the optimization problem (8.8) is given by $\mathrm{KL}(\overline{\mathbb{Q}}, \overline{\mathbb{P}})$. By convexity, $\mathrm{KL}(\overline{\mathbb{Q}}, \overline{\mathbb{P}}) \leq \sum_{k\in\mathcal{K}} \mathrm{KL}(\mathbb{Q}_k, \mathbb{P}_k)$, 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.

The simultaneous t-test sequence model

We end this chapter with an example of a construction of an approximate compound e-value (as in Definition 8.20) in a setting with composite null hypotheses. For the k-th hypothesis we observe

$$X_{kj} \sim \mathrm{N}(\mu_k, \sigma_k^2), \text{ for } j = 1, \dots, n, \ \ \mu_k \in \mathbb{R}, \ \ \sigma_k > 0, \tag{8.10}$$

and we assume that all X_{kj}, $1 \le k \le K$, $1 \le j \le n$ are mutually independent and $n \ge 2$. We seek to test $H_k : \mu_k = 0$. The data in (8.10) may be summarized via the sufficient statistics $\hat{\mu}_k = \sum_{j=1}^n X_{kj}/n$ and $\hat{\sigma}_k^2 = \sum_{j=1}^n (X_{kj} - \hat{\mu}_k)^2/(n-1)$. Further let $S_k^2 := \sum_{j=1}^n X_{kj}^2/n$ be the mean of squares of the observations for the k-th hypothesis. Observe that $\mathbb{E}[S_k^2] = \sigma_k^2 + \mu_k^2$ and so $\mathbb{E}[S_k^2] = \sigma_k^2$ for $k \in \mathcal{N}$. We conservatively estimate the sum $\sum_{k \in \mathcal{N}} \sigma_k^2$ via $\hat{c} = \sum_{k \in \mathcal{K}} \hat{\sigma}_k^2$, where the expectation is taken under the hypothesis (8.10). This leads to the following construction:

$$E_k = \frac{K S_k^2}{\hat{c}} = \frac{K S_k^2}{\sum_{j \in \mathcal{K}} \hat{\sigma}_j^2}.$$

One can find explicit values for $\varepsilon, \delta \approx 0$ such that $E_1, \dots, E_K$ are (ε, δ)-approximate compound e-values and such that the FDR inflation in Theorem 8.22 becomes negligible, $\alpha(1+\varepsilon) + \delta \approx \alpha$, in some practically relevant settings.

8.4 Extensions of the e-BH procedure and compound e-values

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 $K \le 10$. We will focus on e-values and boosted e-values.

First, choose an e-merging function $F : [0,\infty)^K \to [0,\infty)$ in Chapter 7. We allow for a general choice of F other than $\mathbb{M}_K$ 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 $\bigcap_{k=1}^K H_k$ via the rejection condition $F(e_1, \dots, e_K) \ge 1/\alpha$, 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 $\mathcal{D}_\alpha^F = \varnothing$;
2. if $F(e_1, \dots, e_K) \ge 1/\alpha$, 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 8.17

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 the event that $F(E_1, \dots, E_K) \geq 1/\alpha$. If $K_0 < K$, then by using Theorem 8.6,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}^*}\left[\frac{F_{\mathcal{D}_\alpha^F}}{R_{\mathcal{D}_\alpha^F}}\right] &= \mathbb{E}^{\mathbb{P}^*}\left[\frac{F_{\mathcal{D}_{\alpha'}}}{R_{\mathcal{D}_{\alpha'}}}\mathbb{1}_A\right] + \mathbb{E}^{\mathbb{P}^*}\left[\frac{F_\varnothing}{R_\varnothing}(1-\mathbb{1}_A)\right] \\ &= \mathbb{E}^{\mathbb{P}^*}\left[\frac{F_{\mathcal{D}_{\alpha'}}}{R_{\mathcal{D}_{\alpha'}}}\mathbb{1}_A\right] \leq \mathbb{E}^{\mathbb{P}^*}\left[\frac{F_{\mathcal{D}_{\alpha'}}}{R_{\mathcal{D}_{\alpha'}}}\right] \leq \frac{K_0}{K}\alpha' \leq \alpha.\end{aligned}$$

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

$$S : (e_1, \dots, e_K) \mapsto \max_{k=1,\dots,K} \frac{k e_{[k]}}{K}. \tag{8.11}$$

The function S is an e-merging function and it is dominated by $\mathbb{M}_K$ because $\sum_{k=1}^K e_k \geq k e_{[k]}$. By definition of the e-BH procedure, $S(e_1, \dots, e_K) < 1/\alpha$ implies $\mathcal{D}_\alpha = \varnothing$. Therefore, if $\mathbb{M}_K(e_1, \dots, e_K) < 1/\alpha$, then $\mathcal{D}_\alpha = \varnothing = \mathcal{D}_\alpha^{\mathbb{M}_K}$. 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 (8.11) on the boosted e-values in Section 8.1. 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 $\varnothing$.
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 8.17.

Stochastic rounding of e-values and the e-BH procedure

We next discuss an improvement of the e-BH procedure with randomization. The randomized Markov's inequality in Theorem 2.19 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 given 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

$$x^+ := \inf\{y \in \mathcal{G} : y \geq x\}, \qquad x_- := \sup\{y \in \mathcal{G} : y \leq x\},$$

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 $\mathcal{G}$, denote $S_{\mathcal{G}}(x)$, 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 $S_{\mathcal{G}}$.

Proposition 8.18

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. Fix $\alpha \in (0, 1)$, and let

$$\alpha_i := \alpha i / K, \qquad \mathcal{K} := \{\alpha_i^{-1} : i \in [K]\} \cup \{0, \infty\}$$

denote the set of possible levels that e-BH may reject e-values at and in addition to 0 and ∞. If $X \geq K/(\alpha k)$ for some $i \in [K]$, then $S_{\mathcal{K}}(X) \geq K/(\alpha k)$ as well. Thus, an e-value that is stochastically rounded to $\mathcal{K}$ can only improve power when used in conjunction with e-BH. We elaborate and then improve on the above idea in the next section.

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^+.

One can view the e-BH procedure as selecting a data-dependent threshold

$$\widehat{\alpha}^* := \alpha(k^* + 1)/K$$

and rejecting the ith hypothesis if and only if $X_i \geq 1/\widehat{\alpha}^*$ (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 $1/\widehat{\alpha}^*$ can only take values from the grid $\{K/(k\alpha) : k \in [K]\}$. If we "rounded" down each X_i to the closest value in the grid that is less than X_i (or 0 if X_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^* is the number of discoveries made by e-BH applied to e-values $X_1, \ldots, X_K$. The *stochastically-rounded R_1-eBH procedure* simply applies the e-BH procedure to the set of e-values $\{S_{\mathcal{K}}(X_k)\}_{i \in [K]}$. Let $\mathcal{D}_1$ be the set of rejections made by R$_1$-eBH, and $\mathcal{D}_{\text{eBH}}$ be the set of rejections made by e-BH at level α.

Theorem 8.19

For any arbitrarily dependent e-values $(X_1, \dots, X_K)$, the R_1-eBH procedure ensures FDR $\leq \alpha$ and $\mathcal{D}_1 \supseteq \mathcal{D}_{\mathrm{eBH}}$. Further, for any $\mathbb{P} \in \mathcal{M}_1$, $\mathbb{P}(\mathcal{D}_1 \supsetneq \mathcal{D}_{\mathrm{eBH}}) > 0$, i.e., the probability that R_1-eBH makes extra discoveries over e-BH is positive, if and only if

$$\mathbb{P}\left(\exists k \in [K - k^*] : X_{[k^*+j]} > \frac{K}{\alpha(k^* + k + 1)} \text{ for all } j \in [k]\right) > 0. \tag{8.12}$$

The proof follows from the fact that if X_i ever takes on a value that is between levels in $\mathcal{K}$, the e-BH procedure will reject at the same level (and make the same rejections) as if $(X_i)_-$ were substituted in its place. Stochastic rounding guarantees that $S_{\mathcal{K}}(X_i) \geq (X_i)_-$ almost surely, so it can only increase the number of rejections. Further, when X_i is between two levels in $\mathcal{K}$, then $S_{\mathcal{K}}(X_i) = (X_i)_+ > X_i$ with positive probability, which leads to rejecting hypotheses that e-BH did not reject. This intuition leads us to the condition in (8.12) for which R_1-eBH has strictly more power than e-BH. We omit the proof.

It is worth remarking that (8.12) 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 (8.12) 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 (8.12) is not pathologically equal to zero due to some awkward worst-case dependence structure.

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 like the BH procedure under positive dependence, and indeed we conjecture that it does not. Both the BH and BY procedures have their FDR upper bounds being achieved with equality in a particular setting; yet it appears as if the power of BY 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.

Asymptotic compound e-values and FDR control

The next definition introduces compound e-variables in the approximate or asymptotic sense.

Definition 8.20: Approximate or asymptotic compound e-values

For $k \in \mathcal{K}$, let E_k be a $[0, \infty]$-valued random variable. Also let $\varepsilon, \delta \in [0, 1)$. We say that $E_1, \dots, E_K$ are (ε, δ)-approximate compound e-variables for $\mathcal{P}_1, \dots, \mathcal{P}_K$ if, for every $\mathbb{P} \in \mathcal{P}$, there exists an event A such that:

$$\sum_{k : \mathbb{P} \in \mathcal{P}_k} \mathbb{E}^{\mathbb{P}}[E_k \mathbb{1}_A] \leq K(1 + \varepsilon), \quad \mathbb{P}(A) \geq 1 - \delta.$$

When allowing the size of the data X or the number of hypotheses K to vary, we call $E_1, \dots, E_K$ asymptotic compound e-values if they are $(\varepsilon, 0)$-approximate compound e-values for $\varepsilon = o(1)$ as the size of X or the number K grows to ∞ (or both).

Both notions of asymptotic error control 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.

As a minor remark, let $E_1, \ldots, E_K$ be (ε, δ)-approximate compound e-values. Suppose that there exists a uniformly distributed random variable U on $[0,1]$ and independent of both $(E_k : k \in \mathcal{K})$ and A (the conditioning event in Definition 8.20) under each $\mathbb{P}$. Then $E_1, \ldots, E_K$ are also $(0, \delta')$-approximate compound e-values, where $\delta' = (\delta + \varepsilon)/(1+\varepsilon)$, by choosing $A \cap \{U \leq 1/(1+\varepsilon)\}$ as the conditioning event.

Example 8.21: Convex combinations of compound e-values

Suppose that $E_1^{(\ell)}, \ldots, E_K^{(\ell)}$ are compound e-values for $\ell = 1, \ldots, L$. Let $w_1, \ldots, w_L \geq 0$ 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 $E_1, \ldots, E_K$ are compound e-values. If $E_1^{(\ell)}, \ldots, E_K^{(\ell)}$ are asymptotic compound e-values for $\ell = 1, \ldots, L$ (with L fixed), then $E_1, \ldots, E_K$ are also asymptotic compound e-values.

Let $\mathbb{P}^*$ be the true data generating distribution. A procedure is said to have (ε, δ)-approximate FDR control at level α, if there exists a set A of probability $1-\delta$ under $\mathbb{P}^*$, for which $\mathbb{E}^{\mathbb{P}^*}[\mathbb{1}_A F_{\mathcal{D}}/R_{\mathcal{D}}] \leq \alpha(1+\varepsilon)$ holds true, which implies that $\mathrm{FDR} \leq \alpha(1+\varepsilon)+\delta$. It is said to have asymptotic FDR control (as the number of hypotheses or size of data grows to infinity) if we can take $\delta = 0$ and $\varepsilon = o(1)$.

Theorem 8.22

Suppose that we apply the e-BH procedure at level α to $E_1, \ldots, E_K$. Then:

- If $E_1, \ldots, E_K$ are compound e-values, then e-BH controls the false discovery rate at level α.
- If $E_1, \ldots, E_K$ are (ε, δ)-approximate compound e-values, then e-BH satisfies (ε, δ)-approximate FDR control, and so it controls the FDR at level $\alpha(1+\varepsilon)+\delta$.
- If $E_1, \ldots, E_K$ are asymptotic compound e-values, then e-BH controls the false discovery rate at level α asymptotically.

The same conclusions also hold for any self-consistent e-testing procedures at level α (Definition 8.5).

Notably, the above result requires no assumption whatsoever on the dependence across the E_k.

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 BH procedure equals $\alpha\ell_K$, a fact mentioned in Section 8.1.

The e-BH procedure was proposed by Wang and Ramdas [2022]. The proof of Theorem 8.8 can be found in Wang and Ramdas [2022], where other boosting methods for e-values than the one in Section 8.1 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, 2024, Lee and Ren, 2024] or relaxed e-values [Ren and Barber, 2024, Gablenz and Sabatti, 2024]. A unified treatment of compound e-values was provided in Ignatiadis et al. [2024a], from which this chapter draws many results.

The term compound e-values pays tribute and connect the definition 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 (8.4) as $K \to \infty$.

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 8.4 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 8.4 was developed in Xu and Ramdas [2023].

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.

Part III

Advanced Topics

Chapter 9

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 7.4, 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 $\mathbb{P}$ on some $(\Omega, \mathcal{F})$ as in Chapter 7.

9.1 Optimal transport duality

A technical tool that will be useful in two main proofs in Sections 9.2 and 10.1 is optimal transport duality. We state a classic version of multi-marginal optimal transport duality without proving it.

Let $F : [0,\infty)^K \to \mathbb{R}$ be a bounded Borel function. Denote by $\mathcal{B}$ the set of Borel functions on $[0,\infty)$, and define the operator $\bigoplus$ as

$$\left(\bigoplus_{k=1}^{K} \phi_k\right)(x_1,\dots,x_K) := \sum_{k=1}^{K} \phi_k(x_k), \quad (\phi_1,\dots,\phi_K) \in \mathcal{B}^K, \quad (x_1,\dots,x_K) \in [0,\infty)^K.$$

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 $\boldsymbol{\phi} = (\phi_1,\dots,\phi_K)$ and $\boldsymbol{\mu} = (\mu_1,\dots,\mu_K)$.

Lemma 9.1: Optimal transport duality

For any bounded Borel function $F : [0,\infty)^K \to \mathbb{R}$ and distributions $\mu_1,\dots,\mu_K$ on $\mathbb{R}$,

$$\sup_{\pi \in \Gamma(\boldsymbol{\mu})} \int F \mathrm{d}\pi = \inf_{\boldsymbol{\phi} \in D_F} \sum_{k=1}^{K} \int \phi_k \mathrm{d}\mu_k, \tag{9.1}$$

where in the infimum we only consider those with $\sum_{k=1}^{K} \int \phi_k \mathrm{d}\mu_k$ well-defined. Moreover, if F is upper semicontinuous, then the infimum is attainable, and the functions $\phi_1,\dots,\phi_K$ in (9.1) 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.

9.2 Admissible e-merging functions

The main content of this section is to prove Theorem 7.4, recalled here.

Recalling Theorem 7.4

For a function $F : \mathbb{R}_+^K \to \mathbb{R}_+$,

(i) if F is an e-merging function, then $F \le \mathbb{M}_{\boldsymbol{\lambda}}$ for some $\boldsymbol{\lambda} \in \Delta_{K+1}$;

(ii) F is an admissible e-merging function if and only if $F = \mathbb{M}_{\boldsymbol{\lambda}}$ for some $\boldsymbol{\lambda} \in \Delta_{K+1}$.

It suffices to show statement (i). Statement (ii) follows directly from (i) and the simple fact that $\mathbb{M}_{\boldsymbol{\lambda}}$ is an e-merging function.

First, we characterize the simplest case $K = 1$, which will be used in the proof of Theorem 7.4. All e-variables in this section are for $\mathbb{P}$.

Lemma 9.2

Let $\theta \in [1, \infty]$ and $r \in \mathbb{R}_+$. If a function $g : [0, \theta] \to \mathbb{R}_+$ satisfies $\mathbb{E}^{\mathbb{P}}[g(E)] \le r$ for all e-variables E taking values in $[0, \theta]$, then there exists $\lambda \in [0, 1]$ such that $g(x) \le r(1 - \lambda + \lambda x)$ for $x \in [0, \theta] \cap \mathbb{R}$.

As a particular case of Lemma 9.2 with $r = 1$, if g is an e-merging function of dimension 1, then there exists $\lambda \in [0, 1]$ such that $g(x) \le (1 - \lambda) + \lambda x$ for $x \in \mathbb{R}_+$. Hence, the desired statement in Theorem 7.4 holds for $K = 1$.

Lemma 9.2 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 9.2 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 $(1 - \lambda_t + \lambda E_t)$ in Section 6.7 to build an e-process from given sequential e-variables $(E_t)_{t \in T_+}$.

Proof of Lemma 9.2.

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 $g(x) \le 1$ 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 $g(y) \le y$ 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, $g(y) \le 1$ for $y \le 1$ is also clear, which in particular implies $g(1) \le 1$.

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 $g(y) \le y$ for $y > 1$ and $g(x) \le 1$ 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

$$\Lambda_0 = \{\lambda \in [0,1] : g(y) > (1-\lambda) + \lambda y \text{ for some } y > 1\};$$
$$\Lambda_1 = \{\lambda \in [0,1] : g(x) > (1-\lambda) + \lambda x \text{ for some } x < 1\}.$$

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.

The next lemma gives an upper bound on F that allows us to apply optimal transport duality in Lemma 9.1. In what follows, $\mathbf{0}$ and $\mathbf{1}$ represent a vector of zeros and a vector of ones of the appropriate dimension, respectively.

Lemma 9.3

Any e-merging function $F : \mathbb{R}_+^K \to \mathbb{R}_+$ satisfies $F(\mathbf{e}) \le 1 \vee \max(\mathbf{e})$ for all $\mathbf{e} \in \mathbb{R}_+^K$.

Proof.

Suppose that there exists $\mathbf{e} \in \mathbb{R}_+^K$ such that $F(\mathbf{e}) > 1 \vee \max(\mathbf{e})$. Let $\bar{e} = \max\{\mathbf{e}\}$. If $\bar{e} \le 1$, then $\mathbf{e}$ is a vector of constant e-variables, but $F(\mathbf{e}) > 1$ is not an e-variable, a contradiction. Next, consider $\bar{e} > 1$. Consider e-variables $E_1, \dots, E_K$ defined by $\mathbb{P}((E_1, \dots, E_K) = \mathbf{e}) = 1/\bar{e}$ and $\mathbb{P}((E_1, \dots, E_K) = \mathbf{0}) = 1 - 1/\bar{e}$. It is straightforward to see that $E_1, \dots, E_K$ are e-variables. Moreover, $\mathbb{E}[F(E_1, \dots, E_K)] > \bar{e}/\bar{e} = 1$, a contradiction to the assumption that F is an e-merging function.

The next lemma allows us to only consider upper semicontinuous e-merging functions.

Lemma 9.4

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.

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 $\mathbf{E}_\varepsilon = (\mathbf{E} + \varepsilon \mathbf{1})\mathbb{1}_{A_\varepsilon}$ (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

$$1 \geq \mathbb{E}[F(\mathbf{E}_\varepsilon)] = (1-\varepsilon)\mathbb{E}\left[F(\mathbf{E} + \varepsilon \mathbf{1})\right] + \varepsilon F(\mathbf{0}),$$

which implies

$$\mathbb{E}\left[F(\mathbf{E} + \varepsilon \mathbf{1})\right] \leq \frac{1 - \varepsilon F(\mathbf{0})}{1 - \varepsilon}.$$

Fatou's lemma yields

$$\mathbb{E}[F^*(\mathbf{E})] = \mathbb{E}\left[\lim_{\varepsilon \downarrow 0} F(\mathbf{E} + \varepsilon \mathbf{1})\right] \leq \lim_{\varepsilon \downarrow 0} \mathbb{E}\left[F(\mathbf{E} + \varepsilon \mathbf{1})\right] \leq \lim_{\varepsilon \downarrow 0} \frac{1 - \varepsilon F(\mathbf{0})}{1 - \varepsilon} = 1.$$

Therefore, F^* is an e-merging function.

Proof of Theorem 7.4

By Lemma 9.4, 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 9.1 we get

$$\sup_{\boldsymbol{\mu} \in (\mathcal{M}_\mathcal{E}^\theta)^K} \sup_{\pi \in \Gamma(\boldsymbol{\mu})} \int F \mathrm{d}\pi = \sup_{\boldsymbol{\mu} \in (\mathcal{M}_\mathcal{E}^\theta)^K} \inf_{\boldsymbol{\phi} \in D_F} \sum_{k=1}^K \int \phi_k \mathrm{d}\mu_k = \sup_{\boldsymbol{\mu} \in (\mathcal{M}_\mathcal{E}^\theta)^K} \inf_{\boldsymbol{\phi} \in D_F^+} \sum_{k=1}^K \int \phi_k \mathrm{d}\mu_k, \qquad (9.2)$$

where D_F^+ is the subset of D_F containing all $\boldsymbol{\phi}$ with nonnegative and upper semicontinuous components. Define the mapping

$$J : (\boldsymbol{\mu}, \boldsymbol{\phi}) \mapsto \sum_{k=1}^K \int \phi_k \mathrm{d}\mu_k.$$

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

$$\sup_{\boldsymbol{\mu}\in(\mathcal{M}^\theta_\varepsilon)^K}\ \inf_{\boldsymbol{\phi}\in D_F^+}\sum_{k=1}^K\int\phi_k\mathrm{d}\mu_k=\inf_{\boldsymbol{\phi}\in D_F^+}\ \sup_{\boldsymbol{\mu}\in(\mathcal{M}^\theta_\varepsilon)^K}\sum_{k=1}^K\int\phi_k\mathrm{d}\mu_k, \tag{9.3}$$

Since F is an e-merging function, each term in (9.2) is bounded by 1. Using (9.3), we get

$$1\geq\inf_{\boldsymbol{\phi}\in D_F^+}\ \sup_{\boldsymbol{\mu}\in(\mathcal{M}^\theta_\varepsilon)^K}\sum_{k=1}^K\int\phi_k\mathrm{d}\mu_k=\inf_{\boldsymbol{\phi}\in D_F^+}\sum_{k=1}^K\ \sup_{\mu_k\in\mathcal{M}^\theta_\varepsilon}\int\phi_k\mathrm{d}\mu_k. \tag{9.4}$$

Denote by $T_\phi=\sup_{\mu\in\mathcal{M}^\theta_\varepsilon}\int\phi\mathrm{d}\mu$ for $\phi\in\mathcal{B}$. For any $\varepsilon>0$, by (9.4), 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 9.2, for each $k\in[K]$, there exists a constant $h_{\phi_k}\in[0,1]$ such that

$$\phi_k(x)\leq T_{\phi_k}\left(1-h_{\phi_k}+h_{\phi_k}x\right)\ \text{ for all } x\in[0,\theta].$$

Since $(\phi_1,\dots,\phi_K)\in D_F^+$, we have

$$F(x_1,\dots,x_K)\leq\sum_{k=1}^K T_{\phi_k}\left(1-h_{\phi_k}+h_{\phi_k}x_k\right)\ \text{ for all } x_1,\dots,x_K\in[0,\theta].$$

This means that there exists $\boldsymbol{\lambda}_{\varepsilon,\theta}\in\Delta_{K+1}$ such that $F\leq(1+\varepsilon)\mathbb{M}_{\boldsymbol{\lambda}_{\varepsilon,\theta}}$ on $[0,\theta]^K$. Since Δ_{K+1} is compact, we can find a limit $\boldsymbol{\lambda}_0\in\Delta_{K+1}$ of some subsequence of $\boldsymbol{\lambda}_{\varepsilon,\theta}$ as $\varepsilon\downarrow 0$ and $\theta\to\infty$. Continuity of $\boldsymbol{\lambda}\mapsto\mathbb{M}_{\boldsymbol{\lambda}}$ yields $F\leq\mathbb{M}_{\boldsymbol{\lambda}_0}$ on $\mathbb{R}_+^K$, thus the desired statement.

9.3 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 9.4.

Propositions 2.3 and 2.4 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 9.5

A function $f:[0,1]\times[0,\infty]\to[0,\infty]$ 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 $C_h^{\mathrm{ie}}:(p,e)\mapsto h(p)e$. The convention here is $0\times\infty=\infty$.

[ip] Return P/E, capped at 1, by using the function $C^{\mathrm{ip}}:(p,e)\mapsto(p/e)\wedge 1$.

[e] Return $\lambda h(P)+(1-\lambda)E$ by using the function $C^{\mathrm{e}}_{\lambda,h}:(p,e)\mapsto\lambda h(p)+(1-\lambda)e$ for some $\lambda\in(0,1)$.

[p] Return $2\min(P, 1/E)$, capped at 1, by using the function $C^{\mathrm{p}} : (p, e) \mapsto (2(p \wedge e^{-1})) \wedge 1$.

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^{ie}_h and $C^{\mathrm{e}}_{\lambda,h}$ 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 $C^{\mathrm{e}}_{\lambda,h}$, it may be convenient to choose $\lambda = 1/2$, so that $C^{\mathrm{e}}_{\lambda,h}(P, E)$ 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. Indeed, we have already seen the validity of C^{ip} in Corollary 2.20 in the context of randomized tests. Admissibility of the above combiners is obtained in the following result.

Theorem 9.6

For an admissible calibrator h and $\lambda \in (0, 1)$, C_h is an admissible i-pe/e combiner, C^{ip} is an admissible i-pe/p combiner, $C^{\mathrm{e}}_{\lambda,h}$ 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 omit the proof of Theorem 9.6, but checking that C^{ip} is a valid i-pe/p combiner is simple: For $t \in (0, 1)$ and any independent pair (P, E) of p-variable and e-variable for $\mathbb{P}$, we have

$$\mathbb{P}(C^{\mathrm{ip}}(P, E) \le t) = \mathbb{P}(P/E \le t) = \mathbb{P}(P \le tE) = \mathbb{E}^{\mathbb{P}}[\mathbb{P}(P \le tE|E)] \le \mathbb{E}^{\mathbb{P}}[tE] \le t.$$

Certainly, both $(P/E) \wedge 1$ and P/E are p-variables. 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 9.4.

Example 9.7: Randomization

Consider an e-variable E for $\mathbb{P}$ and generate an independent uniform variable $U \sim \mathrm{U}[0, 1]$. Then, using the quotient combiner, U/E is a p-variable that satisfies $U/E \le 1/E$. Hence, the unique admissible e-to-p calibrator $f : e \mapsto (1/e) \wedge 1$ in Proposition 2.3 can be improved if we allow for randomization. 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 $P_{\mathrm{F}} := 1 - \chi_4(-2\log(PP'))$, 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.

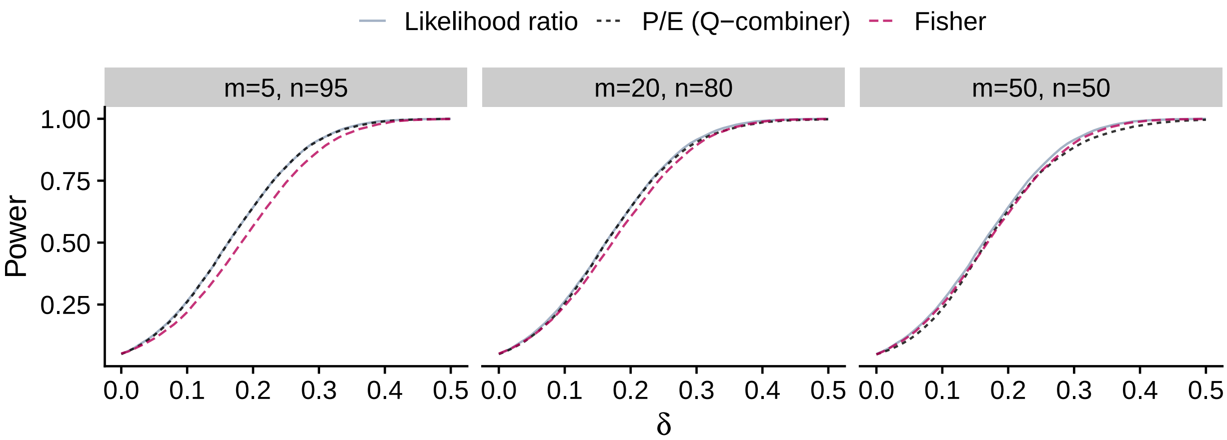

Figure 9.1: **Simulation study for a meta-analysis combining two samples**: We compare the likelihood ratio test, the quotient combiner (Q-combiner), and the Fisher combination test, plotting power against signal strength δ. The panels correspond to different choices of the two sample sizes m and n. The quotient combiner is visibly more powerful on the left (matching the likelihood ratio), and Fisher's combination is marginally better on the right.

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 9.8: A stylized example illustrating the rule of thumb

Suppose we have two independent samples of iid data points, $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, both from a distribution $\mathbb{P}$, where $n \geq m \geq 1$. We seek to test $H_0 : \mathbb{P} = \mathrm{N}(0, 1)$ against $H_1 : \mathbb{P} = \mathrm{N}(\delta, 1)$, 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 $\mathrm{N}(\delta, 1)^m$ over $\mathrm{N}(0, 1)^m$.

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) $\alpha \to 0$, that is, when the type-I error is very stringent, and (b) $m/n \to 0$, that is, when $\mathbf{Y}$ is substantially more informative than $\mathbf{X}$.

(ii) A small simulation study (Fig. 9.1) 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, in case of using one p-value and one e-value, the primary dataset should be used to generate the p-value.

9.4 E-values as weights in p-value based multiple testing

In this section, we study how the cross-merging methods in Section 9.3 can be useful in the context of multiple testing in Chapter 8.

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 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 9.8 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 9.9: E-weighted p-value procedure (ep-$\mathcal{D}$)

Let p-$\mathcal{D}$ be a multiple testing procedure based on p-values. We define the e-weighted p-value procedure ep-$\mathcal{D}$ 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 p-$\mathcal{D}$.

We mainly focus on the e-weighted version of the BH procedure in Section 8.1, 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 8.1. 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 9.10

Suppose that $\mathbf{P}$ is independent of $\mathbf{E}$. and that $\mathbf{P}$ satisfies PRDS (Definition 8.10). Then, the ep-BH procedure has FDR at most $K_0\alpha/K$.

Proof.

Let 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

$$\mathbb{E}\left[\frac{F_{\text{ep-}\mathcal{D}}}{R_{\text{ep-}\mathcal{D}}} \,\middle|\, \mathbf{E}\right] \leq \frac{1}{K}\sum_{k\in\mathcal{N}} E_k\alpha.$$

Hence, by iterated expectation, the FDR of ep-$\mathcal{D}$ is at most $\mathbb{E}[\sum_{k\in\mathcal{N}} E_k\alpha/K] \leq \pi_0\alpha$.

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 whatsoever 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 9.1 can be verified with several results in Kellerer [1984]; the formula (9.1) 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 7.4 presented in Section 9.2 was obtained by Wang [2024].

The content of Sections 9.3–9.4 is based on Ignatiadis et al. [2024b]. The proof of Theorem 9.6 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 9.4 was introduced by Ignatiadis et al. [2024b], along with results on many other procedures weighted by e-values.

Chapter 10

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. We first consider deterministic methods for arbitrarily dependent p-values. Then we turn to methods based on the assumption of exchangeability among p-values, and methods based on randomization.

As in Chapter 9, we fix a positive integer $K \geq 2$ and an atomless probability $\mathbb{P}$ on some $(\Omega, \mathcal{F})$ throughout this chapter.

10.1 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 9.1, which converts constraints on probability into constraints on expectation.

P-merging functions and examples

We first give the definition of p-merging functions, which is analogous to e-merging functions in Definition 7.1.

Definition 10.1

(i) An *p-merging function* (of K p-values) is an increasing Borel function $F : [0, \infty)^K \to [0, \infty)$ such that for any hypothesis, $F(P_1, \ldots, P_K)$ is a p-variable for any p-variables $P_1, \ldots, P_K$.

(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 7 and 9, it suffices to consider p-variables for the fixed atomless probability $\mathbb{P}$. We let $\mathfrak{U}$ be the set of all p-variables for $\mathbb{P}$.

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 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 $(ax) \wedge (by)$.

Finally, 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.

We collect some basic properties of admissible p-merging functions, which will be useful in our analysis later. We skip the proofs.

Proposition 10.2

(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.

We will always write $\mathbf{p} = (p_1, \ldots, p_K)$. 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 10.3: Order-family

The order-family is parameterized by $k \in [K]$, and its kth element is the function

$$G_{k,K} : \mathbf{p} \mapsto \frac{K}{k} p_{(k)} \wedge 1,$$

where $p_{(k)}$ is the kth ascending order statistic of $p_1, \ldots, p_K$.

Example 10.4: Mean-family

The mean-family is parameterized by $r \in [-\infty, \infty]$, and its element with index r has the form

$$F_{r,K} : \mathbf{p} \mapsto b_{r,K}\mathbb{M}_{r,K}(\mathbf{p}) \wedge 1, \tag{10.1}$$

where $\mathbb{M}_{r,K}$ is given by

$$\mathbb{M}_{r,K}(\mathbf{p}) = \left(\frac{p_1^r + \cdots + p_K^r}{K}\right)^{1/r}$$

and $b_{r,K} \geq 1$ is a suitable constant making $F_{r,K}$ a p-merging function. The average $\mathbb{M}_{r,K}$ is also defined for $r \in \{0, \infty, -\infty\}$ as the limiting cases of (10.1), 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 10.5: 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 10.6: 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, called the *Hommel function*,

$$H_K := \left(\sum_{k=1}^{K} \frac{1}{k} \right) \bigwedge_{k=1}^{K} G_{k,K} = \ell_K S_K.$$

Using e-values to merge p-values

We first explain a simple way of merging p-values through e-values. Recall that a calibrator transforms a p-variable to an e-variable. Let Δ_K be the standard K-simplex, that is,

$$\Delta_K := \{(\lambda_1, \ldots, \lambda_K) \in [0,1]^K : \lambda_1 + \cdots + \lambda_K = 1\}.$$

For any calibrators $f_1, \ldots, f_K$ and any $(\lambda_1, \ldots, \lambda_K) \in \Delta_K$, define the function

$$G(\mathbf{p}) := \lambda_1 f_1(p_1) + \cdots + \lambda_K f_K(p_K).$$

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 $f_1(P_1), \ldots, f_K(P_K)$. 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 2.3 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 10.7 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

$$R_\varepsilon(F) := \left\{ \mathbf{p} \in [0,\infty)^K : F(\mathbf{p}) \leq \varepsilon \right\}.$$

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 $F : [0,\infty)^K \to [0,1]$ by the equation

$$F(\mathbf{p}) = \inf\{\varepsilon \in (0,1) : \mathbf{p} \in R_\varepsilon\}, \tag{10.2}$$

with the convention $\inf \varnothing = 1$. It is immediate that F is a p-merging function if and only if $\mathbb{P}(\mathbf{P} \in R_\varepsilon) \le \varepsilon$ for all $\varepsilon \in (0,1)$ and $\mathbf{P} \in \mathfrak{U}^K$.

Theorem 10.7

(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

$$R_\varepsilon(F) \subseteq \left\{ \mathbf{p} \in [0,\infty)^K : \sum_{k=1}^K \lambda_k g_k(p_k) \ge \frac{1}{\varepsilon} \right\}. \tag{10.3}$$

(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

$$R_\varepsilon(F) = \left\{ \mathbf{p} \in [0,\infty)^K : \sum_{k=1}^K \lambda_k f_k\left(\frac{p_k}{\varepsilon}\right) \ge 1 \right\} \qquad \text{for each } \varepsilon \in (0,1). \tag{10.4}$$

(iv) For any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and calibrators $f_1, \dots, f_K$, (10.4) determines a homogeneous p-merging function.

As a consequence of Theorem 10.7 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 10.7 are connected via $g_k(x) = f_k(x/\varepsilon)/\varepsilon$, $x \in [0,\infty)$, for each $k \in [K]$. This choice yields admissible calibrators by Proposition 2.5. 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 2.5.

The p-merging function in (10.4) can be explicitly expressed via (10.2) as

$$F(\mathbf{p}) = \left\{ \varepsilon \in (0,1] : \sum_{k=1}^K \lambda_k f_k\left(\frac{p_k}{\varepsilon}\right) \ge 1 \right\} \qquad \text{for } \mathbf{p} \in [0,\infty)^K. \tag{10.5}$$

To prove Theorem 10.7, 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 $\mathbf{y} \le \mathbf{x}$ (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

$$\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in L) = \sup_{\mathbf{P} \in \mathfrak{U}^K} \mathbb{P}(\mathbf{P} \in L), \tag{10.6}$$

because replacing an p-variable with a smaller exact p-variable does not reduce the above probability. The fact (10.6) will be repeatedly used in the proof below.

Lemma 10.8

Let $R \subseteq [0,\infty)^K$ be a decreasing Borel set. For any $\beta \in (0,1)$, we have

$$\sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in \beta R) \ge \beta \iff \sup_{\mathbf{P} \in \mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R) = 1.$$

Proof.

We first prove the ⇐ 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

$$\mathbf{P}^* = \beta\mathbf{P} \times \mathbb{1}_A + \mathbf{1} \times \mathbb{1}_{A^c}.$$

It is straightforward to check $\mathbf{P}^* \in \mathfrak{U}^K$. Hence, by (10.6),

$$\beta\mathbb{P}\left(\mathbf{P} \in R\right) = \mathbb{P}(A)\mathbb{P}\left(\mathbf{P} \in R\right) \leq \mathbb{P}(\mathbf{P}^* \in \beta R) \leq \gamma,$$

and thus $\mathbb{P}(\mathbf{P} \in R) \leq \gamma/\beta$. Since $\gamma/\beta < 1$, this yields $\sup_{\mathbf{P}\in\mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R) < 1$ and completes the ⇐ direction.

Next we show the ⇒ 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 (10.6).

Proof of Theorem 10.7.

For part (i), since any p-merging function has an admissible p-merging function dominating it (Proposition 10.2), 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 10.2. Hence $\mathbb{1}_{R_\varepsilon(F)}$ is upper semicontinuous. Using optimal transport duality, Lemma 9.1,

$$\min_{(h_1,\dots,h_K)\in\mathcal{B}^K} \left\{ \sum_{k=1}^{K} \int_0^1 h_k(x)\mathrm{d}x : \bigoplus_{k=1}^{K} h_k \geq \mathbb{1}_{R_\varepsilon(F)} \right\} = \max_{\mathbf{P}\in\mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R_\varepsilon(F)) = \varepsilon,$$

where the last equality holds because F is precise (Proposition 10.2). Take $(h_1^\varepsilon, \dots, h_K^\varepsilon) \in \mathcal{B}^K$ such that $\bigoplus_{k=1}^K h_k^\varepsilon \geq \mathbb{1}_{R_\varepsilon(F)}$ and $\sum_{k=1}^K \int_0^1 h_k^\varepsilon(x)\mathrm{d}x = \varepsilon$. By the corresponding conditions on $\mathbb{1}_{R_\varepsilon(F)}$ in Lemma 9.1, we can choose each h_k^ε to be non-negative, decreasing and left-continuous. Moreover, one can also require $h_k^\varepsilon(0) = \infty$ for each k.

Lemma 10.8 gives

$$\max_{\mathbf{P}\in\mathfrak{U}_0^K} \mathbb{P}(\mathbf{P} \in R_\varepsilon(F)) = \varepsilon \implies \max_{\mathbf{P}\in\mathfrak{U}_0^K} \mathbb{P}(\varepsilon\mathbf{P} \in R_\varepsilon(F)) = 1. \tag{10.7}$$

Therefore, using duality in Lemma 9.1 again,

$$\min_{(h_1,\dots,h_K)\in\mathcal{B}^K}\left\{\sum_{k=1}^K \frac{1}{\varepsilon}\int_0^\varepsilon h_k(x)\mathrm{d}x : \bigoplus_{k=1}^K h_k \geq \mathbb{1}_{R_\varepsilon(F)}\right\} = 1,$$

implying $\sum_{k=1}^K \int_0^\varepsilon h_k^\varepsilon(x)\mathrm{d}x \geq \varepsilon$. As $\sum_{k=1}^K \int_0^1 h_k^\varepsilon(x)\mathrm{d}x = \varepsilon$ 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 $G : [0,\infty)^K \to \mathbb{R}$ via $R_\delta(G) = \delta\varepsilon^{-1}A_\varepsilon$ for all $\delta \in (0,1)$. By the above properties of A_ε and Lemma 10.8, 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 $\lambda_k = \varepsilon^{-1}\int_0^\varepsilon g^\varepsilon(x)\mathrm{d}x$ and $g_k : [0,\infty) \to \mathbb{R}$, $x \mapsto \varepsilon^{-1}h_k^\varepsilon(x)/\lambda_k$ 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 (10.5), the condition $\mathbb{P}(F(\mathbf{P}) \leq \varepsilon) \leq \varepsilon$ for $\mathbf{P} \in \mathfrak{U}_0^K$ follows from Proposition 2.5 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 (10.4), 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 the homogeneous p-merging function F is symmetric, then $f_1,\dots,f_K$, as well as $\lambda_1,\dots,\lambda_K$, in Theorem 10.7 can be chosen identical.

Theorem 10.9

For any F that is admissible within the family of homogeneous symmetric p-merging functions, there exists an admissible calibrator f such that

$$R_\varepsilon(F) = \left\{\mathbf{p} \in [0,\infty)^K : \frac{1}{K}\sum_{k=1}^K f\left(\frac{p_k}{\varepsilon}\right) \geq 1\right\} \qquad \text{for each } \varepsilon \in (0,1). \tag{10.8}$$

Conversely, for any calibrator f, (10.8) determines a homogeneous symmetric p-merging function.

The p-merging function in (10.8) can be explicitly expressed as

$$F(\mathbf{p}) = \left\{\varepsilon \in (0,1] : \frac{1}{K}\sum_{k=1}^K f\left(\frac{p_k}{\varepsilon}\right) \geq 1\right\} \qquad \text{for } \mathbf{p} \in [0,\infty)^K.$$

Proof.

The proof is similar to that of Theorem 10.7 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 10.7 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 10.7 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 10.7 by noting that (10.8) defines a symmetric rejection region.

The requirement $f(0) = \infty$ for an admissible calibrator f implies that the combined test (10.8) gives a rejection as soon as one of the input p-values is 0, which is obviously necessary for admissibility (Proposition 10.2). 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

$$\widetilde{F}(\mathbf{p}) := \begin{cases} F(\mathbf{p} \wedge \mathbf{1}) \wedge 1 & \text{if } \mathbf{p} \in (0, \infty)^K \\ 0 & \text{otherwise,} \end{cases}$$

which does not affect the validity of the p-merging function.

For a decreasing function $f : [0, \infty) \to [0, \infty]$ and a p-merging function F taking values in $[0, 1]$, we say that f *induces* F if (10.8) holds; similarly, we may say that $\lambda_1, \dots, \lambda_K$ and $f_1, \dots, f_K$ *induce* F if (10.3) holds. Theorems 10.7 and 10.9 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 (10.7) 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 (10.8) is

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} f(P_k) \geq 1\right) = 1 \quad \text{for some } (P_1, \dots, P_K) \in \mathfrak{U}_0^K. \tag{10.9}$$

Condition (10.9) may be difficult to check for a given f in general. A special known as is that f is convex. In this case, (10.9) 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 10.9. The proof of Theorem 10.10 is advanced and omitted here.

Theorem 10.10

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 (10.8) 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 $(0, \eta]$, $f(\eta+) \leq K$, f is strictly convex on $(\eta, \tau]$, and $f(\tau) = 0$.

Remark 10.11. There is a statement for the case of concavity instead of convexity. The p-merging function induced by f via (10.8) 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 $f(\tau) = 0$.

Examples

We present a few examples of p-merging functions.

Example 10.12: Order-family

The p-merging function $F := G_{k,K}$, $k \in [K]$, is induced by the calibrator $(K/k)\mathbb{1}_{[0,k/K]}$. Its zero-one adjusted version is admissible for $k \in [K-1]$.

Example 10.13: Twice the arithmetic mean

The function $\mathbf{p} \mapsto 2\mathbb{M}_{1,K}(\mathbf{p} \wedge \mathbf{1}) \wedge \mathbb{1}_{\{\min \mathbf{p}>0\}}$, which is the zero-one adjustment of twice the arithmetic mean, is a precise p-merging function but not admissible.

Example 10.14: $\mathtt{e}$ times the geometric mean

The function $\mathtt{e}\mathbb{M}_{0,K}$, that is, $\mathtt{e}$ times the geometric mean, is a p-merging function that is not precise, but approximately precise.

Example 10.15: $\log K$ 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.

Example 10.16: 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

$$f(x) = \frac{K\mathbb{1}_{\{\ell_K x \leq 1\}}}{\lceil K\ell_K x\rceil}, \quad x \geq 0,$$

and H_K^* is admissible when K is not a prime number (this is an intriguing fact; for example, if $K = 2$ or 3, H_K^* is shown to be not admissible).

10.2 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 10.1 (like twice the average in Example 10.13) 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 10.17

An *ex-p-merging function* is an increasing Borel function $F : [0,\infty)^K \to [0,\infty)$ 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 10.18

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 σ be a random permutation of $\{1, \dots, K\}$, uniformly drawn from all permutations of $\{1, \dots, K\}$ and independent of $\mathbf{P}$. Let $\mathbf{P}^\sigma = (P_{\sigma(1)}, \dots, P_{\sigma(K)})$. Note that $\mathbf{P}^\sigma$ is exchangeable by construction. If F is a symmetric ex-p-merging function, it must satisfy $F(\mathbf{P}^\sigma) = F(\mathbf{P})$. Because $F(\mathbf{P}^\sigma)$ 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 4.5, the p-merging method from Theorem 10.9 and the formula (10.5).

Theorem 10.19: Combination of exchangeable p-values

Define the function $F : [0, \infty)^K$ by

$$F(\mathbf{p}) = \left\{ \varepsilon \in (0,1] : \bigvee_{\ell=1}^{K} \left(\frac{1}{\ell} \sum_{k=1}^{\ell} f\left(\frac{p_k}{\varepsilon}\right) \right) \geq 1 \right\},$$

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 10.19, 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.

10.3 Randomized combinations of p-values

In Section 10.1, we obtained methods to merge p-values using e-values in Theorems 10.7 and 10.9, and they are admissible with the condition in Theorem 10.10. When randomization is allowed, the Markov inequality can be enhanced to the randomized Markov inequality, presented in Theorem 2.19. 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 10.1. The following result describes a method that uses randomization to improve p-merging functions. This result can be seen as a combinations of Theorem 2.19 and 10.7 and the formula (10.5).

Theorem 10.20: Randomized combination of p-values

Define the function $F : [0,\infty)^K \times [0,\infty)$ by

$$F(\mathbf{p}, u) = \left\{ \varepsilon \in (0,1] : \sum_{k=1}^{K} \lambda_k f_k \left(\frac{p_k}{\varepsilon} \right) \geq u \right\}, \tag{10.10}$$

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 U independent of $\mathbf{P}$, $F(\mathbf{P}, U)$ is a p-variable.

Proof.

If $U \overset{\mathrm{d}}{\sim} \mathrm{U}[0,1]$, then 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 2.19. For a general p-variable U, it suffices to notice that $F(\mathbf{p}, u)$ is increasing in u, and hence $\mathbb{P}(F(\mathbf{P}, U) \leq \alpha) \leq \mathbb{P}(F(\mathbf{P}, V) \leq \alpha) \leq \alpha$, where $V \overset{\mathrm{d}}{\sim} \mathrm{U}[0,1]$ is independent of $\mathbf{P}$.

A simple subclass of (10.10) is

$$F(\mathbf{p}, u) = \left\{ \varepsilon \in (0,1] : \frac{1}{K} \sum_{k=1}^{K} f \left(\frac{p_k}{\varepsilon} \right) \geq u \right\},$$

for a calibrator f, similarly to (10.5).

For the deterministic choice $U = 1$, Theorem 10.20 gives the validity statement in Theorem 10.7 (iv). In practice, when randomization is allowed, U should be chosen as uniformly distributed on $[0,1]$. This merging method produces a smaller p-value than the deterministic p-merging function in (10.5).

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 $U = P_k$ and apply the merging function to the rest of the p-variables with randomization through U.

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 10.3 applied to a random permutation of the p-values.

Table 10.1 summarizes some examples of p-merging methods under arbitrary dependence, randomization, or exchangeability. They are derived from Theorems 10.9, 10.19 and 10.20 with different calibrators f, and correspond to Examples 10.12–10.15.

Bibliographical note

The content of Section 10.1 is mainly based on Vovk et al. [2022], and the content of Sections 10.2–10.3 is based on Gasparin et al. [2024].

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 10.6 were proposed by Simes [1986] and Hommel [1983], respectively. The order-family was proposed by Rüger [1978]. The result on arithmetic mean was derived by Rüschendorf [1982]. The harmonic mean was proposed by Wilson [2019]. The mean-family was formally studied by Vovk and Wang [2020]. In some 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 necessary and sufficient condition for (10.9) to hold for convex f is given in Theorem 3.2 of Wang and Wang [2016]. This result is also needed to prove Theorem 10.10. The proof of Theorem 10.10 and justifications for statements in Section 10.1 are found in Section 6 of Vovk et al. [2022].

Context	Arbitrary dependence (Section 10.1)	Exchangeability (Section 10.2)	Arbitrary dependence, randomized (Section 10.3)
Result	Theorem 10.9	Theorem 10.19	Theorem 10.20
Order statistics	$\frac{K}{k} p_{(k)}$	$\frac{K}{k} \bigwedge_{m=1}^{K} p^{m}_{(\lambda_m)}$	$\frac{K}{k} p_{(\lceil Uk \rceil)}$
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	$e\mathbb{M}_0(\mathbf{p})$	$e \bigwedge_{m=1}^{K} \mathbb{M}_0(\mathbf{p}_m)$	$e^U \mathbb{M}_0(\mathbf{p})$
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 10.1: Some combination rules corresponding to Examples 10.12–10.15. 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 $p^m_{(\lambda_m)}$ 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 $K \geq 2$.

Chapter 11

E-confidence intervals

This chapter studies confidence intervals, or confidence regions, formulated by e-values. We will see that these 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 $\mathbb{P}^* \in \mathcal{P}$, and we will be interested in estimating $\theta^* = \vartheta(\mathbb{P}^*)$, where $\vartheta : \mathcal{P} \to \Theta$ is a predefined functional of interest (like the mean, median, and so on). Let $\mathcal{P}_\theta := \{\mathbb{P} \in \mathcal{P} : \vartheta(\mathbb{P}) = \theta\}$ denote the set of all distributions whose functional equals θ.

11.1 Defining and constructing E-CIs

Definition 11.1: 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 wrt $\mathcal{P}_\theta$.

For a fixed $\alpha \in [0, 1]$, a set $C(\alpha) \subset \Theta$ 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 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 say that $\{C(\alpha)\}_{\alpha\in[0,1]}$ is a family of *e-confidence intervals (e-CIs)* if $C(\alpha)$ is a $(1 - \alpha)$-CI for every 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 $C(0) = \Theta$, $C(1) = \varnothing$. 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 $\mathbb{E}^{\mathbb{P}}[E(\theta, \alpha)] \leq 1$ 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 2.3. Therefore, the more interesting object to us is the level-free e-CI families.

We have already seen one general construction of e-CIs: those obtained via universal inference in Section 4.5. Here, we develop two more examples: via stopped confidence sequences and calibrated CIs.

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 6. 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 11.2

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 $C^\tau(\alpha) \subset \Theta$ 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 6.

For any fixed α, let $E(\theta,\alpha) \equiv \{E_t(\theta,\alpha)\}_{t\geq 0}$ denote an e-process wrt $\mathcal{P}_\theta$. 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 11.3

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 11.3 directly follows from Ville's inequality stated in in Fact 6.4.

E-CIs by calibrating CIs

Recall the notion of a calibrator from Definition 2.2. 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 11.4

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 $C^{\mathrm{cal}}(\alpha)$ is a $(1-\alpha)$-e-CI:

$$C^{\mathrm{cal}}(\alpha) = C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right) \supseteq C(\alpha).$$

Moreover, $\{C^{\mathrm{cal}}(\alpha)\}_{\alpha\in[0,1]}$ is a level-free family of e-CIs.

Theorem 11.4 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 11.4.

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 2.2), 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.

11.2 False coverage rate and the Benjamini-Yekutieli procedure

Consider a scientist who observes some data, $\mathbf{X} = (X_1, \ldots, X_K)$ drawn from an unknown $\mathbb{P}^* \in \mathcal{P}$, and we are potentially interested in the values of K of its functionals[1] $\vartheta_1, \ldots, \vartheta_K$, but our interest in them depends on the unknown parameter values $\boldsymbol{\theta}^* := (\theta_1^*, \ldots, \theta_K^*)$, where $\theta_i^* := \vartheta_i(\mathbb{P}^*)$. For example, the scientist may only 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 $i \in [K]$, 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, $S \subseteq [K]$, 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)

[1] Technically these functionals could each lie in different sets Θ but this complicates notation, and we will anyway not explicitly need these sets later on. Note that each ϑ_i need not be bijective; for example, it could capture the median of a distribution.

and *false coverage rate* (FCR) of such a procedure are:

$$\mathrm{FCP} = \mathrm{FCP}(S, \{\alpha_i\}_{i\in S}) := \frac{\sum_{i\in S} \mathbb{1}_{\{\theta_i^* \notin C_i(\alpha_i)\}}}{|S| \vee 1}, \qquad \mathrm{FCR} := \mathbb{E}^{\mathbb{P}^*}\left[\mathrm{FCP}\right].$$

Note the obvious similarity between the concepts of FCR and FCP and the concepts of FDR and FDP in Chapter 8. If every α_i equals the same constant (say γ), we use the more succinct notation $\mathrm{FCP}(S, \gamma)$ to abbreviate $\mathrm{FCP}(S, \{\alpha_i\}_{i\in S})$.

Our goal is to design a method for choosing $\{\alpha_i\}_{i\in S}$ which guarantees $\mathrm{FCR} \leq \delta$ for 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 $S = \mathcal{S}(\mathbf{X})$.*

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 $X_1, \ldots X_K$. 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 Kth 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 8.1, and one may, analogously to the situation of e-BH versus BH procedures in Chapter 8, 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.

11.3 The e-BY procedure

Now, we formally define the e-BY procedure as follows.

Definition 11.5

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 11.6

Let $\{C_i(\alpha)\}_{\alpha\in[0,1]}$ be a level-free family of e-CIs for each $i \in [K]$. Then, the e-BY procedure ensures $\mathrm{FCR} \leq \delta$ for any $\delta \in (0, 1)$ under any dependence structure between $X_1, \ldots, X_K$, and for any selection rule $\mathcal{S}$. In fact, the e-BY procedure satisfies the stronger guarantee:

$$\mathbb{E}^{\mathbb{P}^*}\left[\sup_{S\in 2^{[K]}} \sup_{\delta\in(0,1)} \frac{\mathrm{FCP}(S, \delta|S|/K)}{\delta}\right] \leq 1.$$

Proof of Theorem 11.6

We directly show an upper bound for the FCR as follows:

$$\begin{aligned}
\mathrm{FCR} &= \mathbb{E}\left[\frac{\sum_{i\in S} \mathbb{1}_{\{\theta_i^* \notin C_i(\delta|S|/K)\}}}{|S| \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i\in [K]} \mathbb{1}_{\{E_i(\theta_i^*)|S|\delta/K>1\}} \cdot \mathbb{1}_{i\in S}}{|S| \vee 1}\right] \\
&\le \sum_{i\in[K]} \mathbb{E}\left[\frac{E_i(\theta_i^*)|S|\delta}{K(|S|\vee 1)}\right] = \sum_{i\in[K]} \frac{\delta}{K}\mathbb{E}\left[E_i(\theta_i^*)\cdot \frac{|S|}{|S|\vee 1}\right] \le \delta,
\end{aligned}$$

where the first inequality is because $\mathbb{1}_{\{x>1\}} \le x$ for all $x \ge 0$. 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.

11.4 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 $\alpha \in (0,1)$:

$$\mathbb{P}\big(c \in C_k\big) \ge 1-\alpha, \quad k = 1,\dots,K, \tag{11.1}$$

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 (11.1), 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 $k \in [K]$. 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.

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 measurable 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 $C_k = C_k(z) \subseteq \mathcal{S}$, $k = 1, \dots, K$, based on the observed data, follow the property (11.1), 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:

$$C^M := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{\{s \in C_k\}} > \frac{1}{2} \right\}. \tag{11.2}$$

Theorem 11.7

Let $C_1, \dots, C_K$ be $K \geq 2$ different sets based on the observed data z, satisfying property (11.1). Then, the set C^M defined in (11.2) has coverage of at least $1 - 2\alpha$:

$$\mathbb{P}\big(c \in C^M\big) \geq 1 - 2\alpha. \tag{11.3}$$

Proof

Let $\phi_k = \phi_k(Z, c) = \mathbb{1}_{\{c \notin C_k\}}$ be a Bernoulli random variable such that $\mathbb{E}[\phi_k] \leq \alpha$, $k = 1, \dots, K$. Thus $E_k = \phi_k / \alpha$ is an e-value, meaning that $\bar{E} = (E_1 + \cdots + E_K)/K$ 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 11.7, the probability of miscoverage is equal to $\alpha K / \lceil K/2 \rceil = 2\alpha K/(K+1)$, approaching (11.3) for large K.

This result 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 11.8 (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 11.7: 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 remark, 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

$$C^\tau := \left\{ s \in \mathcal{S} : \frac{1}{K}\sum_{k=1}^K \mathbb{1}_{\{s\in C_k\}} > \tau \right\}. \tag{11.4}$$

Theorem 11.9

Let $C_1,\dots,C_K$ be $K \geq 2$ different sets satisfying property (11.1). Then,

$$\mathbb{P}\big(c \in C^\tau\big) \geq 1-\alpha/(1-\tau).$$

The proof follows the same lines as the original results outlined in Theorem 11.7 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 $\bigcap_{k=1}^K C_k \neq \varnothing$, then C^τ is an interval for any τ.

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 the largest interval.

Theorem 11.10

Let $\nu(C^\tau)$ be the measure (size) of C^τ in (11.4). Then, for all $\tau \in [0,1)$,

$$\nu(C^\tau) \leq \frac{1}{K\tau}\sum_{k=1}^{K}\nu(C_k). \tag{11.5}$$

If the input sets are K one-dimensional intervals, for all $\tau \in [\frac{1}{2}, 1)$, we have that

$$\nu(C^\tau) \leq \max_k \nu(C_k). \tag{11.6}$$

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 (11.5) 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, (11.5) 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 $K \geq 2$ confidence intervals such that $C_1 = \cdots = C_{\lceil K/2 \rceil} = (0,2)$ and $C_{\lceil K/2 \rceil + 1} = \cdots = C_K = (1,3)$. It is possible to see that if $\tau = (K-2)/(2K)$, then the set C^τ coincides with the union of the initial sets, which is larger than the input intervals. Furthermore, for finite K we have $\tau = (K-2)/(2K) < 1/2$, which implies that the bound in (11.6) is also tight. This fact provides a practical justification for choosing $1/2$ as value of τ: it is the smallest τ for which the combined set cannot be larger than the input sets. In particular, a simple majority vote seems to offer a good compromise between coverage and size.

Exchangeable, randomized and weighted variants

Surprisingly, when $C_1, \ldots, C_K$ 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 $C_1, \ldots, C_k$. Now define

$$C^E := \bigcap_{k=1}^{K} C^M(1:k),$$

which can be equivalently represented as

$$C^E = \left\{ s \in \mathcal{S} : \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{s \in C_j\}} > \frac{1}{2} \text{ holds for all } k = \{1, \ldots, K\} \right\}.$$

Essentially, C^E is formed by the intersection of sets obtained through sequential processing of the sets derived from the majority vote.

Theorem 11.11

If $C_1, \ldots, C_K$ 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 11.7, but uses the exchangeable Markov inequality from Chapter 4 in place of Markov's inequality. This method shares a same idea as in designing the ex-p-merging functions in Section 10.2.

This result points at a simple way at improving majority vote for arbitrarily dependent sets: process them in a random order. To elaborate, let π be a uniformly random permutation of $\{1, 2, \ldots, K\}$ 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 11.12

If $C_1, \ldots, C_K$ 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 11.11 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:

$$C^R := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{\{s \in C_k\}} > \frac{1}{2} + \frac{u}{2} \right\}. \tag{11.7}$$

As a small variant, define

$$C^U := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{\{s \in C_k\}} > u \right\}.$$

Theorem 11.13

Let $C_1, \ldots, C_K$ be $K \geq 2$ different sets satisfying the property (11.1). 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 11.7, but uses the randomized Markov inequality in Theorem 2.19 from Chapter 2 in place of Markov's inequality. This is related to randomized p-merging functions in Section 10.3, 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.

Assume as before that the sets $C_1, \ldots, C_K$ based on the observed data follow the property (11.1). In addition, let $w = (w_1, \ldots, w_K)$ 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:

$$C^W := \left\{ s \in \mathcal{S} : \sum_{k=1}^{K} w_k \mathbb{1}_{\{s \in C_k\}} > \frac{1}{2} + u/2 \right\}. \tag{11.8}$$

Theorem 11.14

Let $C_1, \ldots, C_K$ be $K \geq 2$ different sets satisfying property (11.1). Then, the set C^W defined in (11.8) 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

$$\nu(C^W) \leq 2 \sum_{k=1}^{K} w_k \nu(C_k). \tag{11.9}$$

If the weights are equal to $w_k = \frac{1}{K}$, for all $k = 1, \ldots, K$, then the set C^W coincides with the set C^R defined in (11.7) and it is a subset of that in (11.2). 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, (11.9) 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 11.15

If $C_1, \ldots, C_K$ are $K \geq 2$ different sets having coverage $1 - \alpha_1, \ldots, 1 - \alpha_K$ (possibly unknown), then the set C^W defined in (11.8) 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 11.14, 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 $(w_1, \ldots, w_K)$ that achieve this as long as $\alpha/2$ is in the convex hull of $(\alpha_1, \ldots, \alpha_K)$.

Since it is desirable to have as small an interval as possible if coverage (11.1) 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.

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 11.16

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). $(C_k^{(t)})_{t\geq 1}$ 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 $k = 1, \ldots, K$. 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

$$C^E(1:t) := \bigcap_{i=1}^{t} C^M(1:i).$$

We claim that the above sequence of sets is actually a $1-2\alpha$ confidence *sequence* for c:

Theorem 11.17

Given an exchangeable sequence of confidence sets $C_1, C_2, \ldots$ (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:

$$\mathbb{P}\left(\exists t \geq 1 : c \notin C^E(1:t)\right) \leq 2\alpha.$$

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.

Bibliographical note

E-confidence intervals were studied by Vovk and Wang [2023] and Xu et al. [2024]. 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 and studied 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. 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.

Chapter 12

Improving the threshold for e-values 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 7 and 9, 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 is 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.

12.1 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 Section 5.6), 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.6 in Appendix A.2.

Lemma 12.1

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 $\gamma \in (0,1]$, we have that $\{t \geq 1 : R_{1/t}(\mathcal{E}) \leq \alpha\}$ is not empty. By using Lemma A.6, 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

$$T_\alpha(\mathcal{E}) = \min\{t \geq 1 : R_{1/t}(\mathcal{E}) \leq \alpha\} = \min\{t \geq 1 : \mathbb{P}(E \geq t) \leq \alpha \text{ for all } E \in \mathcal{E}\},$$

showing the second statement.

In Section 2.2 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 $f : [0,\infty] \to [0,\infty)$ is an *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 2.3, 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 12.2

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 $\mathbb{P}(g(E) \le \alpha) \le \alpha$ for all $\alpha \in (0,1)$ and $E \in \mathcal{E}$. By definition, for $x \in [0,\infty]$ and $E \in \mathcal{E}$,

$$g(x) \ge \mathbb{P}(E \ge x) = \mathbb{P}(-E \le -x) = \ F_{-E}(-x),$$

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 $\mathbb{P}(F_{-E}(-E) \le \alpha) \le \alpha$ for all $E \in \mathcal{E}$. Therefore, $\mathbb{P}(g(E) \le \mathbb{P}(F_{-E}(-E) \le \alpha) \le \alpha$ for all $\alpha \in (0,1)$.

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 $f(x) < \mathbb{P}(E \ge x)$. This implies $\mathbb{P}(f(E) \le f(x)) \ge \mathbb{P}(E \ge x) > f(x)$, which violates the definition of $f(E)$ as a p-variable.

Proposition 12.2 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.

12.2 Comonotonic e-variables

In this section, we discuss $R_\gamma(\mathcal{E})$ for a set $\mathcal{E}$ of comonotonic e-values. Our results here can be seen as a generalized version of Proposition 2.16.

Consider the simple example in Section 1.4, that is, to test $\mathrm{N}(0,1)$ against $\mathrm{N}(\mu,1)$ with $\mu > 0$ for iid observations $X_1,\dots,X_n$. A natural e-variable E_μ is the likelihood ratio

$$E_\mu = \exp(\mu S_n - n\mu^2/2), \tag{12.1}$$

where $S_n = \sum_{i=1}^n X_i$. An interesting observation is that E_μ, $\mu > 0$ are *comonotonic* random variables. Random variables E_θ, $\theta \in \Theta$ are comonotonic if there exists a common random variable Z such that E_θ is an increasing function of Z for each $\theta \in \Theta$. The following lemma allows us to analyze R_γ for a set of comonotonic e-varibales.

Lemma 12.3

Let $x \in \mathbb{R}$ and $\gamma \in [0,1]$. Suppose that $\mathcal{X}_C$ is a collection of comonotonic random variables satisfying $\mathbb{P}(X \ge x) \le \gamma$ for each $X \in \mathcal{X}_C$. Then $\mathbb{P}(\sup_{X \in \mathcal{X}_C} X \ge x) \le \gamma$.

Proof.

Suppose that all elements of $\mathcal{X}_C$ are increasing functions of Z. Note that the sets $\{X \ge x\}$ 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 $\{Z \ge z\}$ or $\{Z > z\}$ for different z and a common Z. Therefore, we can move the supremum outside the probability and get $\mathbb{P}(\sup_{X \in \mathcal{X}_C} X \ge x) = \sup_{X \in \mathcal{X}_C} \mathbb{P}(X \ge x) \le \gamma$.

The next result follows immediately.

Proposition 12.4

For any collection $\mathcal{E} \subseteq \mathfrak{E}$ of comonotonic e-variables for $\mathbb{P}$, we have $\mathbb{P}(\sup_{E \in \mathcal{E}} E \ge T_\alpha(\mathcal{E})) \le \alpha$.

In particular, Proposition 12.4 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 2.18 with the help of a statistic X. In the setting where e-variables are computed from likelihood ratios, we discussed using the mixture of likelihood ratios in Section 3.4. Proposition 12.4 suggests that, under comonotonicity, we can use the supremum of e-variables instead of their mixture, and the supremum has a higher power.

Example 12.5: Testing normal mean

In the example of testing $\mathrm{N}(0,1)$ against $\{\mathrm{N}(\mu,1) : \mu > 0\}$ with n data points, instead of using E_μ in (12.1) 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 $\{\mathrm{N}(\mu,1) : \mu \leq 0\}$ against $\{\mathrm{N}(\mu,1) : \mu > 0\}$, the same type-I error guarantee holds true, following from Corollary 2.18.

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 $\mathbb{1}_{\{E\geq 1/\alpha\}}$ 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 4 (which yield 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.

12.3 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 modelled by the following sets of e-variables:

$$\begin{aligned}
\mathcal{E}_{\mathrm{D}} &= \{E \in \mathfrak{E} : E \text{ has a decreasing density on its support}\},\\
\mathcal{E}_{\mathrm{D}>1} &= \{E \in \mathfrak{E} : E \text{ has a decreasing density over } [1,\infty)\},\\
\mathcal{E}_{\mathrm{U}} &= \{E \in \mathfrak{E} : E \text{ has a unimodal density on } [0,\infty)\}.
\end{aligned}$$

A random variable on $[0,\infty)$ is said to have 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 $\mathcal{E}_{\mathrm{D}}$ is satisfied by, for instance, exponential and Pareto distributions. Whenever we consider a density, we include its limit case; that is, we allow $E \in \mathcal{E}_{\mathrm{D}}$ 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 $\mathcal{E}_{\mathrm{D}>1}$ is useful and can be observed from some e-variables obtained from universal inference discussed in Chapter 4. The assumption of unimodal density in $\mathcal{E}_{\mathrm{U}}$ is satisfied by, for instance, log-normal and gamma distributions. The set of all modes of $E \in \mathcal{E}_{\mathrm{U}}$ may be a singleton or an interval. Note that $\mathcal{E}_{\mathrm{D}} \subseteq \mathcal{E}_{\mathrm{U}}$, and hence $R_\gamma(\mathcal{E}_{\mathrm{U}}) \geq R_\gamma(\mathcal{E}_{\mathrm{D}})$ for all $\gamma \in (0,1]$.

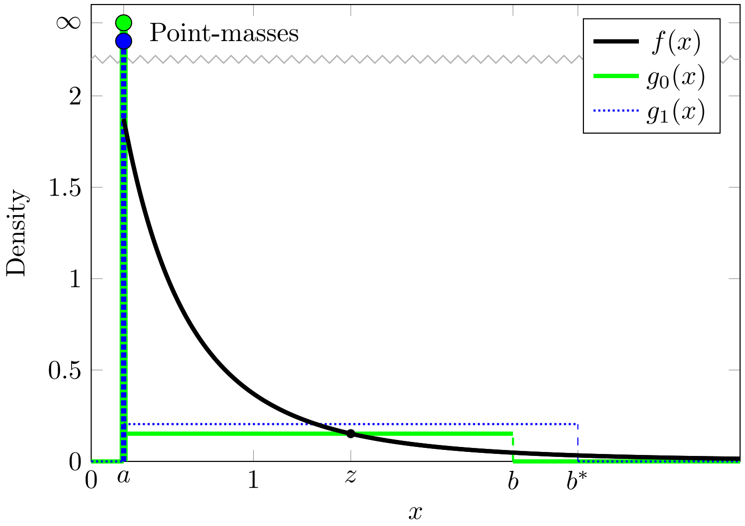

Figure 12.1: Proof sketch in part (i) of Theorem 12.6. Here, f and g_0 have the same area (probability) exceeding z, with g_0 having a smaller mean and a point mass at 0; g_1 has mean 1 and larger probability of exceeding z than g_0.

Theorem 12.6: Decreasing and unimodal densities

For $\gamma \in (0,1]$,

(i) $R_\gamma(\mathcal{E}_{\mathrm{D}}) = \gamma/2$ if $\gamma \neq 1$ and $R_1(\mathcal{E}_{\mathrm{D}}) = 1$;

(ii) $R_\gamma(\mathcal{E}_{\mathrm{D}>1}) = \gamma/(1+\sqrt{1-\gamma^2})$;

(iii) $R_\gamma(\mathcal{E}_{\mathrm{U}}) = (\gamma/2) \vee (2\gamma - 1)$.

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 $E \in \mathcal{E}_{\mathrm{D}}$ 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 $1 - \int_a^\infty g_0(x)\mathrm{d}x$.

We illustrate in Figure 12.1 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 $E \in \mathcal{E}_{\mathrm{D}}$ with left end-point of support a, there exists an e-variable $E' \in \mathcal{E}_{\mathrm{D}}$ 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 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 $\mathbb{E}^{\mathbb{P}}[E'] = 1$.

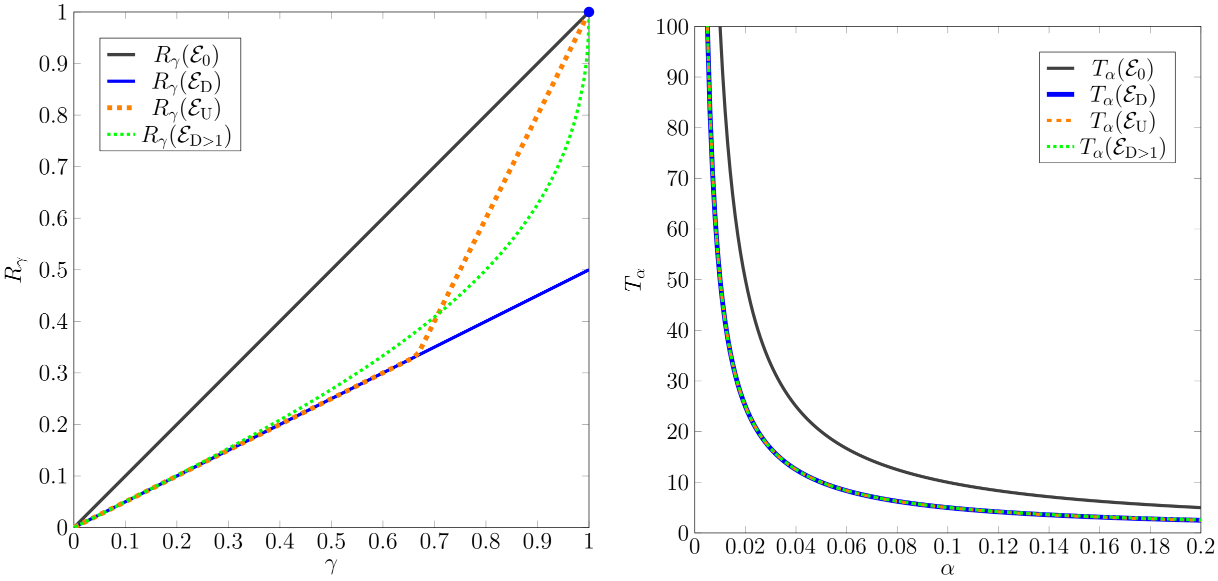

Figure 12.2: Comparison of worst-case type-I errors and improved thresholds for decreasing and unimodal densities. In the left panel, the diagonal line is $R_\gamma(\mathfrak{E})$, corresponding two the Markov inequality, and all other curves have a slope of approximately $1/2$ for small γ.

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 $\mathbb{E}[E'] = 1$ 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 12.1. 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 $E \in \mathcal{E}_\mathrm{D}$. 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 12.6 is that when an e-variable has a decreasing density, its threshold can be boosted by a factor of 2. That is, $T_\alpha(\mathcal{E}_\mathrm{D}) = 1/(2\alpha)$. Further, $T_\alpha(\mathcal{E}_{\mathrm{D}>1}) = 1/(2\alpha) + \alpha/2 = T_\alpha(\mathcal{E}_\mathrm{D}) + \alpha/2$. Therefore, for small values of α, we can boost the threshold by a factor slightly less than 2 when $E \in \mathcal{E}_{\mathrm{D}>1}$. 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 12.6, $R_\gamma(\mathcal{E}_\mathrm{U}) = \gamma/2$, for $\gamma \in (0, 2/3]$ and $R_\gamma(\mathcal{E}_\mathrm{U}) = 2\gamma - 1$ 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 12.2 a comparison of different worst-case type-I errors and improved thresholds for $\mathcal{E}_0, \mathcal{E}_\mathrm{D}, \mathcal{E}_{\mathrm{D}>1}$ and $\mathcal{E}_\mathrm{U}$.

12.4 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 6. 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:

$$\begin{aligned}
\mathcal{E}_{\mathrm{LS}} &= \{E \in \mathfrak{E} : \log E \text{ has a symmetric distribution}\},\\
\mathcal{E}_{\mathrm{LU}} &= \{E \in \mathfrak{E} : \log E \text{ has a unimodal density}\},\\
\mathcal{E}_{\mathrm{LD}>0} &= \{E \in \mathfrak{E} : \log E \text{ has a decreasing density over } [0,\infty)\},\\
\mathcal{E}_{\mathrm{LD}} &= \{E \in \mathfrak{E} : \log E \text{ has a decreasing density}\},\\
\mathcal{E}_{\mathrm{LUS}} &= \{E \in \mathfrak{E} : \log E \text{ has a unimodal and symmetric distribution}\},\\
\mathcal{E}_{\mathrm{LN}} &= \{E \in \mathfrak{E} : E \text{ has a log-normal distribution}\}.
\end{aligned}$$

In all sets above, we require $\mathbb{P}(E = 0) = 0$, so that $\log E$ is a real-valued random variable. Note that $\mathcal{E}_{\mathrm{LN}} \subseteq \mathcal{E}_{\mathrm{LUS}} \subseteq \mathcal{E}_{\mathrm{LS}}$ and $\mathcal{E}_{\mathrm{LD}} \subseteq \mathcal{E}_{\mathrm{LD}>0}$. 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 $\log E \overset{\mathrm{d}}{\sim} \mathrm{N}(\mu, \sigma^2)$ with $\mu + \sigma^2/2 \le 0$.

Theorem 12.7

For $\gamma \in (0, 1]$,

(i) $R_\gamma(\mathcal{E}_{\mathrm{LS}}) = \gamma \wedge (1/2)$ if $\gamma \ne 1$ and $R_1(\mathcal{E}_{\mathrm{LS}}) = 1$;

(ii) $R_\gamma(\mathcal{E}_{\mathrm{LU}}) = \gamma$;

(iii) $\mathcal{E}_{\mathrm{LD}>0} \subseteq \mathcal{E}_{\mathrm{D}>1}$ and

$$\frac{\gamma}{\mathsf{e} - \gamma} \le R_\gamma(\mathcal{E}_{\mathrm{LD}>0}) = \mathsf{e}^{-a_\gamma} \le \frac{\gamma}{1 + \sqrt{1 - \gamma^2}},$$

where a_γ is the unique solution of $\mathsf{e}^a(1 - a - \log \gamma) = 1$ for $a \in (-\log \gamma, \infty)$;

(iv) $R_\gamma(\mathcal{E}_{\mathrm{LD}}) = R_\gamma(\mathcal{E}_{\mathrm{LUS}})$ and

$$\frac{\gamma}{e} \le R_\gamma(\mathcal{E}_{\mathrm{LD}}) = R_\gamma(\mathcal{E}_{\mathrm{LUS}}) \le \frac{\gamma}{\mathsf{e}(1 - \gamma^2)} \wedge \left(\frac{\gamma}{1 + \sqrt{1 - \gamma^2}}\right);$$

(v) if $\gamma \ne 1$ then

$$R_\gamma(\mathcal{E}_{\mathrm{LN}}) = \Phi\left(-\sqrt{-2 \log \gamma}\right) \le \frac{\gamma}{2\sqrt{-\pi \log \gamma}},$$

where Φ is the standard normal cdf, and $R_1(\mathcal{E}_{\mathrm{LN}}) = 1$.

We omit the proof of Theorem 12.7. To summarize the results, for log-transformed symmetric distributions and log-transformed unimodal distributions, the standard Markov's bound cannot be improved for $\gamma \le 1/2$. An improvement of roughly the order of $1/\mathsf{e}$ 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 12.7 in Figure 12.3. Table 12.1 summarizes some numerical values for results in Theorems 12.6 and 12.7.

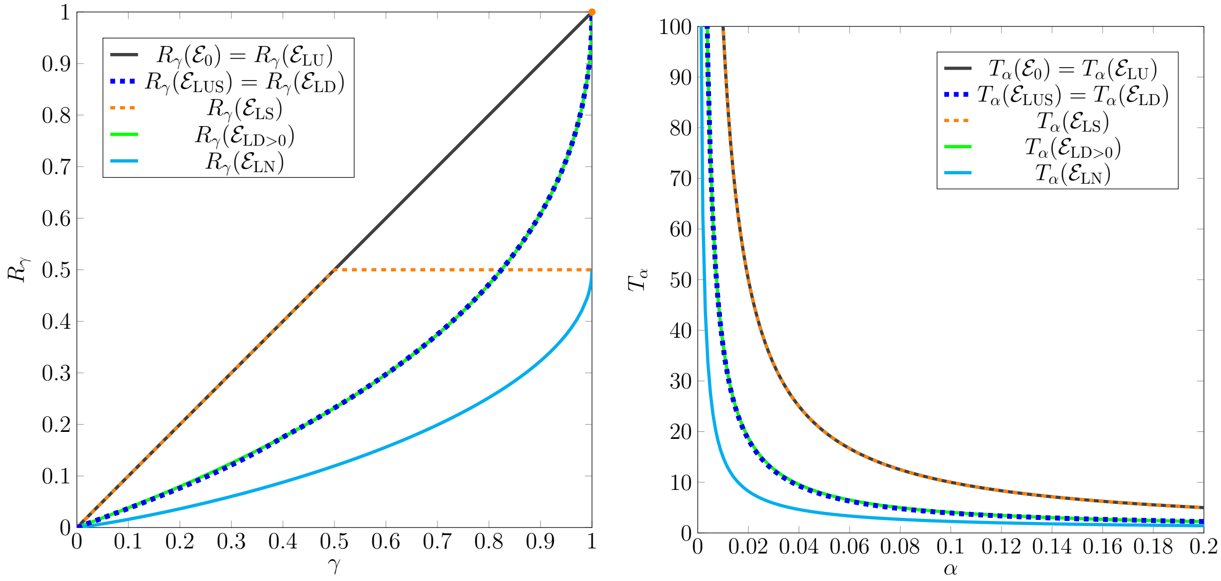

Figure 12.3: Comparison of worst-case type-I errors and improved thresholds for log-transformed e-variables. In the left panel, the diagonal line is $R_\gamma(\mathfrak{E})$, corresponding two the Markov inequality, and the curves for LUS, LD and LD$_{>0}$ have a slope of approximately $\mathtt{e}^{-1}$ for small γ. The curve for LN is the smallest. Results for LUS are conservative bounds since we only find an upper bound for $R_\gamma(\mathcal{E}_{\mathrm{LUS}})$ and $T_\alpha(\mathcal{E}_{\mathrm{LUS}})$.

Table 12.1: Improved thresholds $T_\alpha(\mathcal{E})$ for different distributional hypothesis.

	α					
	0.001	0.01	0.02	0.05	0.1	0.2
$\mathcal{E}_0, \mathcal{E}_{\mathrm{LS}}$	1000	100	50	20	10	5
$\mathcal{E}_{\mathrm{D}}, \mathcal{E}_{\mathrm{U}}$	500	50	25	10	5	2.5
$\mathcal{E}_{\mathrm{D}>1}$	500	50.01	25.01	10.03	5.05	2.60
$\mathcal{E}_{\mathrm{LUS}}, \mathcal{E}_{\mathrm{LD}}$	367.88	36.82	18.45	7.49	3.93	2.25
$\mathcal{E}_{\mathrm{LD}>0}$	368.25	37.16	18.77	7.73	4.07	2.25
$\mathcal{E}_{\mathrm{LN}}$	118	14.97	8.24	3.87	2.27	1.42

Bibliographical note

The content of this paper is largely based on Blier-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 12.2 is a central concept in dependence modeling. It is also central to decision theory under ambiguity via its connection to Choquet integrals, studied by Schmeidler [1989]. A good reference on comonotonicity is Denneberg [1994].

Chapter 13

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.

13.1 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 $X \overset{\mathrm{d}}{\sim} F$ (under $\mathbb{P}_0$). This correspondence is one-to-one if $\mathcal{R}$ satisfies *law invariance*: $\mathcal{R}(X) = \mathcal{R}(Y)$ if $X \overset{\mathrm{d}}{=} Y$. 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.

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 is 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 lower β-quantile:

$$\mathrm{VaR}_\beta(F) = \inf\{x \in \mathbb{R} : F(x) \geq \beta\}, \quad F \in \mathcal{M}_1(\mathbb{R}),$$

and the ES (also called CVaR and TVaR) is defined as

$$\mathrm{ES}_\beta(F) = \frac{1}{1-\beta}\int_\beta^1 \mathrm{VaR}_\gamma(F)\mathrm{d}\gamma, \quad F \in \mathcal{M}_1(\mathbb{R}).$$

Here, $\mathrm{ES}_\beta(F)$ may take the value ∞ if F does not have a finite mean. We also write $\mathrm{VaR}_\beta(X)$ and $\mathrm{ES}_\beta(X)$ for $X \in \mathcal{X}$ to convert between versions in (a) and (b). Note that $\mathrm{VaR}_\beta(X) = q_{1-\beta}(X)$ for the quantile function defined in Chapters 5 and 12. A well-known connection between ES and VaR is

$$\mathrm{VaR}_\beta(X) \in \arg\min_{z\in\mathbb{R}} \left\{ z + \frac{1}{1-\beta}\mathbb{E}^{\mathbb{P}_0}[(X-z)_+] \right\}, \tag{13.1}$$

$$\mathrm{ES}_\beta(X) = \min_{z\in\mathbb{R}} \left\{ z + \frac{1}{1-\beta}\mathbb{E}^{\mathbb{P}_0}[(X-z)_+] \right\}. \tag{13.2}$$

As of 2024, ES at level $\beta = 0.975$ is the standard measure for market risk in the current global banking regulation. 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.

The most important 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 discussions here.

For $\beta \in (0,1)$, $\mathrm{ES}_\beta : \mathcal{X} \to (-\infty,\infty]$ satisfies all of the above, and $\mathrm{VaR}_\beta : \mathcal{X} \to \mathbb{R}$ satisfies all but subadditivity. Hence, ES_β is coherent, but VaR_β is not.

Finally, a risk measure $\mathcal{R} : \mathcal{X}_\mathcal{R} \to (-\infty,\infty]$ is said to be *continuous from above (↓-continuous)* if

$$\mathcal{R}(X_n) \to \mathcal{R}(X) \text{ for any sequence } (X_n)_{n\in\mathbb{N}} \text{ of bounded random variables with } (X_n)_{n\in\mathbb{N}} \downarrow X \text{ point-wise.}$$

In the next section, we will consider the set $\mathcal{X}_+$ of all nonnegative random variables, and connect (↓-continuous) coherent risk measures on $\mathcal{X}_+$ to e-values. We summarize the domains of $\mathcal{R}$ that we will encounter in Table 13.1.

$\mathcal{X}$	the set of all random variables
$\mathcal{X}_B$	the set of bounded random variables
$\mathcal{X}_0$	the set of bounded nonnegative random variables
$\mathcal{X}_+$	the set of nonnegative random variables
$\mathcal{X}_\mathcal{R}$	a generic domain of $\mathcal{R}$, assumed to be a convex cone containing $\mathbb{R}$

Table 13.1: Domains of risk measures $\mathcal{R}$.

13.2 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

$$\mathcal{R}(X) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[X] \tag{13.3}$$

for X in a set such that the above expectations are well-defined.

The mapping $\mathcal{R}$ in (13.3), 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 coherent risk measure, and it has essentially the only form of coherent risk measures.

Theorem 13.1

Let $\mathcal{X}_0$ be the set of all bounded nonnegative random variables. A mapping $\mathcal{R} : \mathcal{X}_0 \to \mathbb{R}$ is a $\downarrow$-continuous coherent risk measure if and only if

$$\mathcal{R}(X) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[X], \qquad X \in \mathcal{X}_0 \tag{13.4}$$

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 $\mathcal{R}' : \mathcal{X}_B \to \mathbb{R}$ is a $\downarrow$-continuous coherent risk measure if and only if it has a representation in (13.4) on $\mathcal{X}_B$. For $\mathcal{R}$ on $\mathcal{X}_{\mathcal{R}}$, it suffices to extend it to $\mathcal{X}_B$ by letting $\mathcal{R}(X) = \mathcal{R}(X - t) + t$ for $X \in \mathcal{X}_B \setminus \mathcal{X}_{\mathcal{R}}$ 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 (13.4) on $\mathcal{X}_{\mathcal{R}}$.

The set $\mathcal{X}_0$ of bounded and nonnegative random variables in Theorem 13.1 can be replaced by larger sets, but usually characterization results like Theorem 13.1 are formulated with small domains for the greatest strength. A useful remark is that if $\mathcal{R} : \mathcal{X}_B \to \mathbb{R}$ is law invariant, then the representation (13.4) holds without requiring $\downarrow$-continuity. Moreover, on any convex cone of random variables, (13.4) always defines a coherent risk measure, possibly taking the value ∞. This in particular holds on $\mathcal{X}_+$, leading to the next observation.

Corollary 13.2

For any $\mathcal{P}$, there exists a coherent risk measure $\mathcal{R} : \mathcal{X}_+ \to [0, \infty]$ 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

$$\begin{aligned} \text{to maximize} \quad & \mathbb{E}^{\mathbb{Q}}[\log X] \\ \text{subject to} \quad & \mathcal{R}(X) \leq 1;\ X \geq 0, \end{aligned} \tag{13.5}$$

where $\mathcal{R}$ is a coherent risk measure with representation (13.3) on $\mathcal{X}_+$. The problem (13.5) has a clear financial interpretation: We aim to maximize our expected utility of the 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 6.

We have seen one example of (13.5) in Section 5.6. The dual representation of ES_β in (13.4) is given by

$$\mathrm{ES}_\beta(X) = \sup\left\{\mathbb{E}^{\mathbb{P}}[X] : \mathbb{P} \in \mathcal{M}_1,\ \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_0} \le \frac{1}{1-\beta}\right\}, \quad X \in \mathcal{X}. \tag{13.6}$$

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 $\mathcal{Q}$:

$$\begin{aligned} \text{to maximize} \quad & \int_{\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log X]\mathrm{d}\nu \\ \text{subject to} \quad & \mathcal{R}(X) \le 1;\ X \ge 0, \end{aligned} \tag{13.7}$$

where and $\mathcal{R}$ is again the coherent risk measure in (13.3). The second is to maximize the worst-case e-power

$$\begin{aligned} \text{to maximize} \quad & \inf_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[\log X] \\ \text{subject to} \quad & \mathcal{R}(X) \le 1;\ X \ge 0. \end{aligned} \tag{13.8}$$

The third is to maximize the relative e-power compared with an oracle e-variable,

$$\begin{aligned} \text{to maximize} \quad & \inf_{\mathbb{Q}\in\mathcal{Q}} \left(\mathbb{E}^{\mathbb{Q}}[\log X] - \mathbb{E}^{\mathbb{Q}}[\log E^*_{\mathbb{Q}}]\right) \\ \text{subject to} \quad & \mathcal{R}(X) \le 1;\ X \ge 0, \end{aligned} \tag{13.9}$$

where $E^*_{\mathbb{Q}}$ is the numeraire for $\mathcal{P}$ against $\mathbb{Q}$ (from Chapter 5).

When there are multiple observations available for the testing problem, the solution to the third problem (13.9) is able to learn the true $\mathbb{P}^*$ with data as discussed in Section 3.4, whereas the solutions to the first two problems do not.

We end this section by noting that the problems (13.7), (13.8), and (13.9) 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.

13.3 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 $\mathbb{R}$.

E-statistics

The general goal is to test the value of $\rho(X_t)$ through a sequence $(X_t)_{t\in T_+}$ of data points, which are not assumed to be iid. Here, the index set T_+ may be finite or infinite. For instance, the null hypothesis may be

$$H_0 : X_t \text{ has a distribution } F_t \text{ with } \rho(F_t) \le r_t \text{ for each } t \in T_+. \tag{13.10}$$

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 6, 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 6.7.

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

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) = r, \tag{13.11}$$

and to test

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) \le r. \tag{13.12}$$

In finance. we are mainly interested in the one-sided hypothesis in (13.12) and its sequential version (13.10), 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 $r \ge \rho(F)$ means that the capital reserve is safe and passing the regulatory requirement. Two-sided tests for (13.11) can be easily constructed by combining one e-variable for testing $\rho(F) \le r$ and one e-variable for testing $\rho(F) \ge r$ 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 $(\mathrm{var}, \mathrm{mean})$, where we are only interested in testing variance, but the information on the mean is available. The corresponding hypotheses are

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) = r \text{ and } \phi(F) = z, \tag{13.13}$$

and

$$H_0 : X \text{ follows any } F \in \mathcal{M}_* \text{ with } \rho(F) \le r \text{ and } \phi(F) = z. \tag{13.14}$$

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 (13.11)–(13.14). Below, $\rho(\mathcal{M}_*)$ is the set of values taken by $\rho(F)$ for $F \in \mathcal{M}_*$.

Definition 13.3: E-statistics for ρ

Fix $\mathcal{M}_* \subseteq \mathcal{M}_\rho$ and $\rho : \mathcal{M}_\rho \to \mathbb{R}$, and let $e : \mathbb{R}^2 \to [0, \infty]$ be a measurable function.

(i) The function e is an *point e-statistic for* ρ if $e(X, r)$ is an e-variable for H_0 in (13.11), and it is a *one-sided e-statistic for* ρ if $e(X, r)$ is an e-variable for H_0 in (13.12).

(ii) A one-sided e-statistic e for ρ is a *backtest e-statistic for* ρ if $\int_{\mathbb{R}} e(x, r) \mathrm{d}F(x) > 1$ 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 13.3 (see Example 13.5 below). The property of having a mean larger than 1 under the alternative is crucial for the e-statistic. By Proposition 2.12, 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 6.12.

We analogously define e-statistics for (ρ, ϕ), which has one more variable.

Definition 13.4: E-statistics for (ρ, ϕ)

Fix $\mathcal{M}_* \subseteq \mathcal{M}_\rho$ and $(\rho, \psi) : \mathcal{M}_\rho \to \mathbb{R}^2$, and let $e : \mathbb{R}^3 \to [0, \infty]$ be a measurable function.

(i) The function e is an *point e-statistic for* (ρ, ψ) if $e(X, r)$ is an e-variable for H_0 in (13.13), and it is a *one-sided e-statistic for* (ρ, ψ) if $e(X, r)$ is an e-variable for H_0 in (13.14).

(ii) A one-sided e-statistic e for (ρ, ϕ) is a *backtest e-statistic for* (ρ, ϕ) if $\int_{\mathbb{R}} e(x, r, z)\mathrm{d}F(x) > 1$ 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 (13.10), as well as other similar hypotheses, by letting $E_t = e(X_t, r_t)$ for $t \in T_+$ where r_t is the forecast for the risk of X_t, and build an e-process from the e-variables $E_1, E_2, \dots$. These e-variables are sequential as long as r_t is the conditional forecast of $\rho(X_t)$ given the information up to $t-1$, which is the typical situation in financial risk management (e.g., firms provide forecast for the next-day risk of their portfolio). Therefore, even if $X_1, X_2, \dots$ are not iid and possibly dependent, we can still use an e-process, e.g., via the empirically adaptive e-process in Section 6.7, to test the null hypothesis. In what follows, we will focus on e-statistics for some commonly used statistical functions.

Mean, variance, quantiles, and expected losses

We derive backtest e-statistics for some common statistical functions. Throughout, we use the convention that $0/0 = 1$ and $1/0 = \infty$, and let $\mathbb{R}_+ = [0, \infty)$. Let $\mathcal{M}^k \subseteq \mathcal{M}_1(\mathbb{R})$ be the set of all distributions with finite kth-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 13.5: 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 13.6: Backtest e-statistic for $(\mathrm{var}, \mathrm{mean})$

Consider $(\mathrm{var}, \mathrm{mean}) : \mathcal{M}^2 \to \mathbb{R}^2$, 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 $(\mathrm{var}, \mathrm{mean})$. 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 $r \geq \mathrm{var}(X)$. Moreover, since $z = \mathbb{E}[X]$ minimizes $\mathbb{E}[(X-z)^2]$ over $z \in \mathbb{R}$ and $\mathrm{var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2]$, $\mathrm{var}(X) > r \geq 0$ implies

$$\mathbb{E}[e(X, r, z)] = \frac{\mathbb{E}[(X - z)^2]}{r} \geq \frac{\mathrm{var}(X)}{r} > 1.$$

Example 13.7: Backtest e-statistic for a quantile

Take $\beta \in (0, 1)$. Define the function

$$e^q_\beta(x, r) = \frac{1}{1 - \beta} \mathbb{1}_{\{x>r\}}, \quad x, r \in \mathbb{R}.$$

We have $\mathbb{E}[e^q_\beta(X, r)] \leq 1$ for any random variable X with distribution F and $r \geq \mathrm{VaR}_\beta(F)$. Moreover, $\mathrm{VaR}_\beta(F) > r$ implies $\mathbb{P}(X > r) > 1 - \beta$, and hence $\mathbb{E}[e^q_\beta(X, r)] > 1$. Therefore, e^q_β is a monotone backtest e-statistic for the β-quantile.

Example 13.8: Backtest e-statistic for an expected loss

For some $a \in \mathbb{R}$, let $\ell : \mathbb{R} \to [a, \infty)$ 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 13.5, e is a monotone backtest e-statistic for the expected loss $F \mapsto \int \ell \mathrm{d}F$ 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 with 1 with the weight between 0 and 1 is also a backtest e-statistic for ψ. Depending on the specific situation, either e-statistic may be useful in practice.

E-statistics for testing ES

The functional $(\rho, \phi) = (\mathrm{var}, \mathrm{mean})$ in Example 13.6 is an example of a *Bayes pair*; that is, there exists a measurable function $L : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\phi(F) \in \arg\min_{z\in\mathbb{R}} \int L(z, x)\mathrm{d}F(x) \quad \text{and} \quad \rho(F) = \min_{z\in\mathbb{R}} \int L(z, x)\mathrm{d}F(x), \quad F \in \mathcal{M}_*, \tag{13.15}$$

where $\int L(z, x)\mathrm{d}F(x)$ 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 $(\mathrm{var}, \mathrm{mean})$. Bayes pairs often admit backtest e-statistics. A typical example commonly used in risk management practice is $(\mathrm{ES}_\beta, \mathrm{VaR}_\beta) : \mathcal{M}^1 \to \mathbb{R}^2$ 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

$$e^{\mathrm{ES}}_\beta(x, r, z) = \frac{(x - z)_+}{(1 - \beta)(r - z)}, \quad x \in \mathbb{R}, \ z \leq r.$$

This defines a backtest e-statistic for $(\mathrm{ES}_\beta, \mathrm{VaR}_\beta)$. Recall the convention that $0/0 = 1$ and $1/0 = \infty$, and set $e^{\mathrm{ES}}_\beta(x, r, z) = \infty$ if $r < z$, which is a case of no relevance since $\mathrm{ES}_\beta(F) \geq \mathrm{VaR}_\beta(F)$ for any $F \in \mathcal{M}^1$.

Theorem 13.9

For $\beta \in (0,1)$, the function e_β^{ES} is a monotone backtest e-statistic for $(\mathrm{ES}_\beta, \mathrm{VaR}_\beta)$.

Proof.

By (13.1) and (13.2), the pair $(\mathrm{ES}_\beta, \mathrm{VaR}_\beta)$ is a Bayes pair in (13.15) with $L : (z, x) \mapsto z + (x-z)_+/(1-\beta)$. Since $L(z,x) \geq z$, for $r \geq z$, we have that $e_\beta^{\mathrm{ES}}(x,r,z) = (L(z,x)-z)/(r-z) \geq 0$, and it is decreasing in r. We have for $r \geq \mathrm{ES}_\beta(X)$ that

$$\mathbb{E}\left[\frac{L(\mathrm{VaR}_\beta(X), X) - \mathrm{VaR}_\beta(X)}{r - \mathrm{VaR}_\beta(X)}\right] = \frac{\mathrm{ES}_\beta(X) - \mathrm{VaR}_\beta(X)}{r - \mathrm{VaR}_\beta(X)} \leq 1.$$

Furthermore, for $z < r \leq \mathrm{ES}_\beta(X)$,

$$\mathbb{E}\left[\frac{L(z, X) - z}{r - z}\right] \geq \frac{\mathrm{ES}_\beta(X) - z}{r - z} \geq 1$$

with equality if and only if $r = \mathrm{ES}_\beta(X)$.

As a consequence of Theorem 13.9, $e_\beta^{\mathrm{ES}}(X, r, z)$ is an e-variable for the null hypothesis

$$H_0 : \mathrm{ES}_\beta(X) \leq r \text{ and } \mathrm{VaR}_\beta(X) = z.$$

A generalization of this result is that $e_\beta^{\mathrm{ES}}(X, r, z)$ is also an e-variable for the larger null hypothesis

$$H_0 : \mathrm{ES}_\beta(X) - \mathrm{VaR}_\beta(X) \leq r - z \text{ and } \mathrm{VaR}_\beta(X) \leq z.$$

Nevertheless, we note that $e_\beta^{\mathrm{ES}}(X, r, z)$ is not an e-variable for H_0: $\mathrm{ES}_\beta(X) \leq r$ and $\mathrm{VaR}_\beta(X) \leq z$.

While Examples 13.5-13.8 and Theorem 13.9 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 13.1 was obtained by Delbaen [2002]. The representation under law invariance, without assuming continuity, was obtained by Jouini et al. [2006]. 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 (13.1) and (13.2) were obtained by Rockafellar and Uryasev [2002, Theorem 10], where ES is called a CVaR, which is common in the optimization literature.

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 13.3 is largely based on Wang et al. [2022], where they showed that most of the backtest e-statistics in Sections 13.3 and 13.3 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 13.9 is closely connected to the joint elicitability 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. [2024].

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.

Remember that a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is *atomless* if it has no *atoms*, i.e., sets $A \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) \in \{0, \mathbb{P}(A)\}$ for any $B \in \mathcal{F}$ such that $B \subseteq A$. For the same concept, we also say that the probability measure $\mathbb{P}$ is atomless.

We start our discussion from a 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 finite sets.

Proof.

The equivalence between (i) and (ii) is stated in Föllmer and Schied [2016, Proposition A.27]. 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.27] again.

If $(\Omega, \mathcal{F})$ is a measurable space and $\mathcal{P}$ is a collection of probability measures on $(\Omega,)$, 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 $\mathbb{P} \times U$, where $([0, 1], \mathcal{B}, \mathbb{L})$ is the standard measurable space $[0, 1]$ 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 $E : \Omega \to [0, \infty]$ 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 $F : [0, \infty)^K \to [0, \infty)$ such that, for all $E_1, \ldots, E_K$,

$$(E_1, \ldots, E_K) \in \mathfrak{E}_{\mathcal{P}}^K \implies F(E_1, \ldots, E_K) \in \mathfrak{E}_{\mathcal{P}}.$$

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 $F : [0, \infty)^K \to [0, \infty)$ 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 7.1);

(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

$$\sup\left\{\mathbb{E}^{\mathbb{P}}[F(\mathbf{E})] : \mathbb{P} \in \mathcal{P},\ \ \mathbf{E} \in \mathfrak{E}_{\mathcal{P}}^K\right\} = \sup\left\{\mathbb{E}^{\mathbb{P}'}[F(\mathbf{E}')] : \mathbb{P}' \in \mathcal{P}',\ \ \mathbf{E}' \in \mathfrak{E}_{\mathcal{P}'}^K\right\}. \tag{16}$$

Suppose

$$\sup\left\{\mathbb{E}^{\mathbb{P}}[F(\mathbf{E})] : \mathbb{P} \in \mathcal{P},\ \ \mathbf{E} \in \mathfrak{E}_{\mathcal{P}}^K\right\} > c$$

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 $\mathbf{E}' = (E_1', \ldots, E_K')$ 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 $E^{\mathbb{P}'}[F(\mathbf{E}')] > c$ for all $\mathbb{P}' \in \mathcal{P}'$. This shows

$$\sup\left\{\mathbb{E}^{\mathbb{P}}[F(\mathbf{E})] : \mathbb{P} \in \mathcal{P},\ \ \mathbf{E} \in \mathfrak{E}_{\mathcal{P}}^K\right\} \leq \sup\left\{\mathbb{E}^{\mathbb{P}'}[F(\mathbf{E}')] : \mathbb{P}' \in \mathcal{P}',\ \ \mathbf{E}' \in \mathfrak{E}_{\mathcal{P}'}^K\right\},$$

and we obtain equality by symmetry.

The implications (ii) ⇒ (iii) and (iii) ⇒ (i) are obvious. To check (i) ⇒ (ii), suppose 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 (16) 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 7.4.

Proposition A.3 shows that in the definition of an e-merging function it suffices to require 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 7.

We can state a similar proposition for calibrators. A *p-variable for a statistical model* $(\Omega, \mathcal{F}, \mathcal{P})$ is a random variable $P : \Omega \to [0, \infty)$ satisfying

$$\mathbb{P}(P \le \varepsilon) \le \varepsilon \text{ for all } \varepsilon \in (0, 1) \text{ and } \mathbb{P} \in \mathcal{P}.$$

The set of p-variables for $(\Omega, \mathcal{F}, \mathcal{P})$ is denoted by $\mathfrak{U}_{\mathcal{P}}$. A decreasing function $f : [0, 1] \to [0, \infty]$ 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 $f : [0, 1] \to [0, \infty]$ 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 2.2);

(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 2.2 and combiners in Chapter 9, which we omit. The content of this section is based on the appendices of Vovk and Wang [2021].

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, 5, 12 and 13.

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

$$Q^-_\alpha(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \le x) \ge \alpha\} \quad \text{for } \alpha \in (0, 1]$$

and its right quantile function is defined as

$$Q^+_\alpha(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \le x) > \alpha\} \quad \text{for } \alpha \in [0, 1).$$

Note that $Q^-_\alpha(X) = q_{1-\alpha}(X)$ for $\alpha \in (0, 1)$ with $q_{1-\alpha}(X)$ in Sections 5.6 and 12.1, but we slightly generalize the definition by including the possibility of $\alpha = 1$.

The first result is a list of translation rules between probability statements and quantile statements.

Lemma A.6

For $\alpha \in [0, 1]$, $t \in \mathbb{R}$ and any random variable X, the following equivalence statements hold:

(i) $Q^-_\alpha(X) > t \iff \mathbb{P}(X \le t) < \alpha$;

(ii) $Q^-_\alpha(X) \le t \iff \mathbb{P}(X \le t) \ge \alpha$;

(iii) $Q^-_\alpha(X) \ge t \iff \mathbb{P}(X \le s) < \alpha$ for all $s < t$;

(iv) $Q^-_\alpha(X) < t \iff \mathbb{P}(X \le s) \ge \alpha$ for some $s < t$;

(v) $Q^+_\alpha(X) < t \iff \mathbb{P}(X < t) > \alpha$;

(vi) $Q^+_\alpha(X) \ge t \iff \mathbb{P}(X < t) \le \alpha$;

(vii) $Q^+_\alpha(X) \le t \iff \mathbb{P}(X < s) > \alpha$ for all $s > t$;

(viii) $Q^+_\alpha(X) > t \iff \mathbb{P}(X < s) \le \alpha$ for some $s > t$.

Proof.

To show (i), denote by $A_\alpha = \{t \in \mathbb{R} : \mathbb{P}(X \le t) \ge \alpha\}$. Note that A_α is closed in $\mathbb{R}$ since $t \mapsto \mathbb{P}(X \le t)$ is upper semicontinuous. This gives $Q^-_\alpha(X) = \min A_\alpha$. Hence, $\alpha > \mathbb{P}(X \le t) \iff t \notin A_\alpha \iff Q^-_\alpha(X) > t$. Part (ii) follows from (i) directly; (iii) follows from (i) and (iv) follows from (ii).

To show (v), denote by $B_\alpha = \{t \in \mathbb{R} : \mathbb{P}(X < t) \le \alpha\}$ 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.7

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 $Q^-_\alpha(X) \ge Q^+_\beta(X) \ge Q^-_\beta(X)$. 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 $Q^-_0(X)$ an $Q^+_1(X)$ are undefined, but this does not matter since U takes the value 0 or 1 with zero probability.

Lemma A.8

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.7, $Q^-_U(X) = Q^+_U(X)$ almost surely. Therefore, it suffices to consider $Q^-_U(X)$. By Lemma A.6, we have, for any $\alpha \in (0,1)$ and $t \in \mathbb{R}$,

$$\mathbb{P}(X \le t) < \alpha \iff Q^-_\alpha(X) > t \implies \mathbb{P}(Q^-_U(X) > t) \ge 1 - \alpha \iff \mathbb{P}(Q^-_U(X) \le t) \le \alpha, \tag{17}$$

and

$$\mathbb{P}(X \le t) \ge \alpha \iff Q^-_\alpha(X) \le t \implies \mathbb{P}(Q^-_U(X) \le t) \ge \alpha. \tag{18}$$

Note that (17) implies

$$\begin{aligned} \mathbb{P}(X \le t) \le \alpha &\iff \mathbb{P}(X \le t) < \beta \text{ for all } \beta > \alpha \\ &\implies \mathbb{P}(Q^-_U(X) \le t) \le \beta \text{ for all } \beta > \alpha \iff \mathbb{P}(Q^-_U(X) \le t) \le \alpha. \end{aligned}$$

Putting this and (18) together we get $X \overset{\mathrm{d}}{=} Q^-_U(X)$.

Lemma A.8 implies, in particular, the well known formula

$$\mathbb{E}^{\mathbb{P}}[X] = \int_0^1 Q_\alpha^-(X)\mathrm{d}\alpha = \int_0^1 Q_\alpha^+(X)\mathrm{d}\alpha.$$

The next result is due to Ryff [1965].

Lemma A.9

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 most intuitive simple 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.9 by standard probabilistic arguments. 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.7, $Q_U^-(X) = Q_U^+(X)$ almost surely, so we will only show the first equality. First, suppose that there exists a uniformly distributed random variable V independent of X. Define

$$U = F(X) - V(F(X) - F(X-)), \tag{19}$$

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 $V_b = 0$ ouside 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 (19), 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 (19), 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*, 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.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Trambak Banerjee, Bowen Gang, and Jianliang He. Harnessing the collective wisdom: Fusion learning using decision sequences from diverse sources. *arXiv preprint*, 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.

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.

Christopher Blier-Wong and Ruodu Wang. Improved thresholds for e-values. *arXiv preprint 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.

Ziyu Chi, Aaditya Ramdas, and Ruodu Wang. Multiple testing under negative dependence. *Bernoulli*, 2024.

J. B. 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.

I. Csiszar and F. Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6): 1474–1490, 2003. doi: 10.1109/TIT.2003.810633.

I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statist. Decisions*, pages 205–237, 1984. ISSN 0721-2631. Recent Results in Estimation Theory and Related Topics.

Donald A. Darling and Herbert Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences of the United States of America*, 57(5):1188–1192, 1967.

Freddy Delbaen. Coherent risk measures on general probability spaces. *Advances in Finance and Stochastics*, pages 1–37, 2002.

Dieter Denneberg. *Non-additive measure and integral*, volume 27. Springer Science & Business Media, 1994.

Vaidehi Dixit and Ryan Martin. Anytime valid and asymptotically optimal statistical inference driven by predictive recursion. *arXiv preprint arXiv:2309.13441*, 2023.

Joseph L. Doob. *Stochastic processes.* Wiley, New York, 1953.

Robin Dunn, Aaditya Ramdas, Sivaraman Balakrishnan, and Larry Wasserman. Gaussian universal likelihood ratio testing. *Biometrika*, 2022.

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*, 2024.

Paul Embrechts, Tiantian Mao, Qiuqi Wang, and Ruodu Wang. Bayes risk, elicitability, 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*, page asae049, 2024.

Lasse Fischer, Ziyu Xu, and Aaditya Ramdas. An online generalization of the (e-)Benjamini-Hochberg procedure. *arXiv preprint 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 elicitability 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 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*, 2024.

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 preprint arXiv:2403.15527*, 2024a.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024b.

Matteo Gasparin, Ruodu Wang, and Aaditya Ramdas. Combining exchangeable p-values. *arXiv preprint arXiv:2404.03484*, 2024.

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.

Benjamin T. Graham and Geoffrey R. Grimmett. Influence and sharp-threshold theorems for monotonic measures. *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*, 2024.

Peter Grünwald, Rianne De Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society, Series B (with discussion)*, 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 preprint arXiv:2404.19465*, 2024b.

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, Ruodu Wang, and Aaditya Ramdas. Compound e-values 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.

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.

J.L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, pages 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 preprint arXiv:2310.01153*, 2023b.

Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. The HulC: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 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.

Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory (to appear)*, 2024.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *arXiv preprint arXiv:2402.18810*, 2024.

Junu Lee and Zhimei Ren. Boosting e-BH via conditional calibration. *arXiv:2404.17562*, 2024.

Guanxun Li and Xianyang Zhang. A Note on E-values and Multiple Testing. *Biometrika*, page asae050, 10 2024. ISSN 1464-3510. doi: 10.1093/biomet/asae050.

J.Q. 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.

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

Hien Duy Nguyen. Universal inference with composite likelihoods. *arXiv preprint arXiv:2009.00848*, 2020.

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. *Annals of Statistics*, 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*, 2024.

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.

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, Heidelberg*, 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*, pages 571–587, 1989.

Glenn Shafer. Testing by betting: a strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistic Society A*, 184(2):407–478, 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 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 preprint arXiv:2407.19361*, 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 preprint 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.

Leonard H. C. Tippett. *The Methods of Statistics: An Introduction Mainly for Experimentalists*. Williams and Norgate, London, 1931.

Timmy Tse and Anthony C Davison. A note on universal inference. *Stat*, 11(1):e501, 2022.

Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

Cédric Villani. *Optimal transport: old and new*. Springer, 2009.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, 1939.

V.G. Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society, series B*, 55:317–351, 1993. (with discussion).

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, 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 (Methodological)*, 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.

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 Applications*, 2023a.

Hongjian Wang and Aaditya Ramdas. Anytime-valid t-tests and confidence sequences for gaussian means with unknown variance. *arXiv preprint arXiv:2310.03722*, 2023b.

Qiuqi Wang, Ruodu Wang, and Johanna Ziegel. E-backtesting. *arXiv preprint arXiv:2209.00991*, 2022.

Ruodu Wang. The only admissible way of merging e-values. *arXiv preprint arXiv:2409.19888*, 2024.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022.

Shaun S Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of risk and insurance*, pages 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 (Statistical Methodology)*, 2024. (with discussion).

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 preprint 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*, 2024.

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 (to appear)*, 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*, pages 1–13, 2024.