p-values

Yuri Gurevich and Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #16

First posted January 11, 2017. Last revised February 10, 2017.

Project web site:
http://alrw.net

Abstract

While p-values are widely used and discussed, we do not know a careful definition of the concept in the literature. We sought to fill in the gap. In the process we discovered that the traditional notion of test statistic is too narrow, and this article indicates a natural generalization.

An older version of this paper is published as an arXiv technical report: `https://arxiv.org/abs/1702.02590`.

Contents

1 Preamble **1**

2 Introduction **3**

3 Preliminaries **7**

4 Preorder topology **9**

5 Traditional test statistics **13**

6 Test preorders **14**

7 Tightening **16**

8 Exact p-values **19**

9 p-functions **20**

10 ω_1-short preorders **22**

11 Randomized p-values **24**

12 An alternative to randomizing p-values **26**

References **29**

> ...the enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious...
>
> Eugene P. Wigner, 1960

> ...statistics is a branch of applied mathematics, like symbolic logic or hydrodynamics. Examination of foundations is desirable, but it must be remembered that undue emphasis on niceties is a disease to which persons with mathematical training are specially prone.
>
> George A. Barnard, 1947

1 Preamble

This article is aimed at two audiences. One audience is non-statisticians and especially computer scientists who might have seen our previous elementary article [17] on the subject. This audience may benefit from the gentle introduction to p-values in Section 2.

The other audience is statisticians, and in this section we try to explain to statisticians that there is still something interesting to say about such a classical notion as p-value (and so our article is not vacuous).

The standard definition of the p-value associated with a given test statistic f and outcome x is

$$\hat{f}(x) = \mathbf{P}[f \leq f(x)] \tag{1}$$

where $[f \leq f(x)] = \{y : f(y) \leq f(x)\}$. It is not always realized by statisticians, however, that this definition is somewhat precarious. The standard textbook convention is that the test statistic f takes values in the real line. But the definition of $\hat{f}$ requires only that the codomain of f be an ordered measurable space and the initial segments $(-\infty, f(x)]$ be measurable. Should we generalize the notion of test statistics appropriately? Would any linearly ordered measurable space work as the codomain of a statistic f provided that the initial segments $(-\infty, f(x)]$ are measurable?

Our answer to the former question is "yes": a useful generalization is possible and sometimes even used in applied statistics, and in this article we will argue for its wider use. But the answer to the latter question turns out to be an emphatic no. There exist a probability space $(\Omega, \Sigma, \mathbf{P})$ and a test statistic f with values in a linearly ordered set with measurable initial segments such that every $\hat{f}(x) = 0$. Informally, this makes no sense: we are entitled to reject the null hypothesis whatever happens. Formally, this contradicts the standard property of validity

$$\mathbf{P}[\hat{f} \leq \epsilon] \leq \epsilon \qquad \text{for every nonnegative } \varepsilon < 1 \tag{2}$$

of p-values.

Example 1.1. The sample space Ω is the collection (known to set-theorists as ω_1) of countable (that is finite and infinite countable) ordinals. In set-theory, every ordinal is the set of smaller ordinals: 0 is the empty set, $1 = \{0\}$, $2 = \{0, 1\}$, $3 = \{0, 1, 2\}$, the first infinite ordinal $\omega_0 = \{0, 1, \ldots\}$ is the set of natural numbers, $\omega_0 + 1 = \omega_0 \cup \{\omega_0\}$, $\omega_0 + 2 = \omega_0 \cup \{\omega_0, \omega_0 + 1\}$, $\ldots$, $\omega_0 + \omega_0 = \omega_0 \cup \{\omega_0 + n : n \in \omega_0\}$, and so on. The first uncountable ordinal ω_1 is the set of countable ordinals.

The σ-algebra Σ consists of all countable subsets of Ω and their complements, and

$$\mathbf{P}(X) = \begin{cases} 0 & \text{if } X \text{ is countable,} \\ 1 & \text{if } \Omega - X \text{ is countable.} \end{cases}$$

Finally the test statistic f is the identity function: $f(x) = x$. The order on the codomain is natural:

$$x < y \iff x \in y,$$

so that

$$0 < 1 < \cdots < \omega_0 < \omega_0 + 1 < \cdots < \omega_0 + \omega_0 < \cdots .$$

For every countable ordinal x, the initial segment $[0, x]$ is countable. Accordingly

$$\hat{f}(x) = \mathbf{P}[f \leq f(x)] = \mathbf{P}[0, x] = 0. \qquad \square$$

It is interesting that Example 1.1 is, in a sense, the only thing that can go wrong with generalized test statistics: one of our results is that anomalies of this kind can arise only when ω_1 can be embedded into f's codomain. The resulting generality can be, and implicitly has been, useful in practical statistics; e.g., even to incorporate the humble notion of randomized p-values into the standard definition (1) we need generalized test statistics with codomains that are richer than the real line, as we discuss later in the article. And we argue that generalized test statistics should be used more widely and not necessarily in combination with randomization.

The paper includes a careful comparison of two natural approaches to define and work with p-values:

- One approach, standard in statistics, starts with test statistics.
- The other approach starts with a linear preorder on the sample space. The idea is that the main role of a test statistic f is to order the sample points x according to their "strangeness"; the precise values $f(x)$ are not essential. This approach has been instrumental in our exploration of proper generalizations of traditional p-values.

The two approaches are believed to be equivalent but we are not aware of any systematic study. (There is a third approach, standard in the algorithmic theory of randomness [26], which focuses on nested families of rejection regions $[f \leq c] = \{x : f(x) \leq c\}$, but it is outside the scope of this article, and we address it elsewhere [18].)

In this journey, we will encounter surprisingly many hurdles and various interesting phenomena.

2 Introduction

2.1 Separation of concerns

It is practically certain that a predicted event of small probability will not happen. This maxim, sometimes called the Cournot Principle, is universally accepted. A question arises how small should the small probability be.

Before we proceed, let's make things a little more precise. Let $\mathcal{T}$ be a probability trial $(\Omega, \Sigma, \mathbf{P})$ where

> Ω is a nonempty set, the set of possible outcomes of $\mathcal{T}$;
>
> Σ is a σ-algebra of subsets of Ω, the set of events of $\mathcal{T}$; and
>
> $\mathbf{P}$ is a σ-additive function from Σ to the real segment $[0, 1]$ with $\mathbf{P}(\Omega) = 1$, the probability distribution of $\mathcal{T}$.

Think of $\mathbf{P}$ as an alleged probability distribution or the *null hypothesis* of the trial. The purpose of the trial is to test the null hypothesis.

An outcome x in a predicted event of small probability, if and when it occurs, provides impugning, or falsifying, evidence against the null hypothesis. A sticky point is whether the impugning power of x is sufficient to impugn the null hypothesis. At this point it makes sense to separate two distinct concerns: what the impugning power of x is, and whether the impugning power of x is sufficient.

The concept of p-value allows us to address the first concern. The p-value of x is a measure of the impugning power of x. The second concern is most interesting as well but we do not address it in this article.

It is surprising that the notion of p-value, so popular in applications and discussions, is not defined carefully in the literature as far as we know. Our best reference on p-values has been Chapter 3 of the 1974 book [5] where the term is not even used. Hence this article.

2.2 Background information

The discussion above may give a false impression that the p-value of an outcome x is determined by the probability trial itself. In the real world, a trial comes with background information.

Example 2.1. Consider the following story from a 1985 New York Times article [34].

> "The New Jersey Supreme Court today caught up with the Essex County Clerk and a Democrat who has conducted drawings for decades that have given Democrats the top ballot line in the county 40 out of 41 times."

The probability trial of the story can be easily formalized. For simplicity, we assume that, in all 41 elections, the number of Democratic candidates was equal to

that of the other candidates. Then the null hypothesis of the trial asserts that all 2^{41} possible outcomes are equally probable. But not all outcomes are equally disturbing. For an outcome x, let $d(x)$ be the number of elections with a Democrat on the top ballot line. If we restrict attention to outcomes x with $d(x) \geq \frac{1}{2} \cdot 41$ then the story suggests that $\text{ImpugningPower}(x_1) \geq \text{ImpugningPower}(x_2)$ if and only if $d(x_1) \geq d(x_2)$. But too small number of Democrats on the top ballot lines would also be suspicious. Even though the clerk is a Democrat, he may be working against the party. Let $m(x) = \min\{d(x), 41 - d(x)\}$. Arguably the story suggests that $\text{ImpugningPower}(x_1) \geq \text{ImpugningPower}(x_2)$ if and only if $m(x_1) \leq m(x_2)$.

Suppose that our probability trial $\mathcal{T}$ is furnished with background information $\mathcal{B}$ that determines the impugning power of the outcomes of $\mathcal{T}$. What is background information exactly? That problem seems intractable. There are no limits on the kinds of background information that occur in real world. Fortunately, for our purposes, we don't have to solve that problem.

The term "impugning power" suggests that outcomes are linearly ordered by their impugning power except that some outcomes may have the same impugning power. In other words, the outcomes are linearly preordered by their impugning power. Recall that a linear preorder of Ω is a binary relation $\leq$ that is reflexive, transitive and quasi-linear, so that for all $x, y, z \in \Omega$ we have

$$x \leq x, \text{ and } (x \leq y \leq z \implies x \leq z), \text{ and } (x \leq y \text{ or } y \leq x).$$

Instead of defining background information $\mathcal{B}$, we postulate that it determines a unique linear preorder on Ω. We do not care about the distinction between background information $\mathcal{B}_1$ and background information $\mathcal{B}_2$ if they determine the same linear preorder. For mathematical purposes, we can — and will — assume that the given background information $\mathcal{B}$ *is* a linear preorder. We return to this issue in [18].

2.3 Statistics

The given linear preorder $\leq$ allows us to assign a numerical value, namely $\mathbf{P}\{y : y \leq x\} = \mathbf{P}(-\infty, x]$, to the impugning power of any outcome x. The greater the impugning power, the smaller the numerical value; we will address this issue shortly.

Definition 2.2. Given a probabilistic trial $\mathcal{T}$ furnished with a test preorder $\leq$, the *exact p-value* of an outcome x is the probability $\mathbf{P}(-\infty, x]$.

Fixing a test preorder is typically done by means of a test statistic. For example, the election story in Example 2.1 focuses our attention on the discrepancy between the number $d(x)$ of Democrats and the number $r(x)$ of non-Democrats in the first ballot lines of x. This calls for a test statistic $m(x)$ defined to be $\min\{d(x), r(x)\}$, which induces the test preorder $x \leq y \iff m(x) \leq m(y)$.

Remark 2.3. The reverse test preorder

$$x \geq y \iff \text{ the impugning power of } x \geq \text{ the impugning power of } y$$

may seem more natural. On the other hand, we definitely want the exact p-value $\mathbf{P}(-\infty, x]$ (a natural and important for our purposes statistic) to assign smaller numerical values to outcomes of greater impugning power: smaller p-values having greater impugning power is a universally accepted convention.

2.4 History of p-values

In the applications of statistics, the results of hypothesis testing are almost invariably packaged as p-values. However, it's difficult to pinpoint when exactly the term was introduced; it developed slowly and informally. The expression "p-value" has a whiff of barbarity about it, and there are at least a dozen (3×2^2) different ways of writing it. Do you capitalize "p"? (If not in general, do you capitalize it at the start of a sentence?) Do you set it in italics? Do you put a hyphen after it? We will ignore such differences in this discussion. But even then, the list of alternative expressions is bewildering. Herbert A. David's list [9, p. 211] includes 9 such expressions: probability level, sample level of significance, observed significance level, significance probability, descriptive level of significance, critical level, significance level, prob-value, and associated probability. And that list is clearly incomplete: e.g., "achieved significance level" (usually abbreviated to ASL) and "attained significance level" are also popular.

The situation with the notion (rather than the term) of p-value is different, and it has a long and venerable history. To our knowledge, the first p-value was computed by John Arbuthnott in 1710 [2]; having observed that the number of male births in London exceeded the number of female births during each of the 82 years from 1629 to 1710, he rejected the null hypothesis of even chances for the birth of male or female. He evaluated the p-value, which he referred to using the generic term "lot", as 2^{-82}. This was, however, a very special case as the value attained by his chosen test statistic (the number of years in which more males were born) was extreme (82 out of 82). The p-value was one-sided, which was justified because, in Arbuthnott's words, "the external accidents to which are males subject (who must seek their food with danger) do make a great havock of them."

As far as we know, the first p-value corresponding to a non-extreme value of a test statistic was computed by Laplace in his 1823 paper [23] devoted to atmospheric tides, a more difficult object of study than ocean tides. In that paper he tested the null hypothesis that changes in atmospheric pressure are not affected by the phase of the Moon. His test statistic was the difference between the mean value of the change (between 9 a.m. and 3 p.m.) in the atmospheric pressure at the Paris Observatory on 792 days near syzygies (times when the Moon, Earth, and Sun are nearly aligned, in either order) and on 792 days near quadratures (times when the Moon, Earth, and Sun form a right angle at the Earth). The p-value that he computed, using a Gaussian approximation and

several more or less convincing assumptions, was 15.7%, which he interpreted as being insufficiently small to reject the null hypothesis. For further details, see [32, Section 4] and [33, Chapter 4]. Laplace's 1823 paper is given by David [9] as the first (?) appearance of the notion of p-value ("(?)" standing for "to his knowledge").

Karl Pearson in his famous 1900 paper [27] about the χ^2 test and Ronald A. Fisher in his 1925 textbook [14] initiated the large-scale use of p-values. Pearson used P as his notation for p-values, and on three occasions referred to it as "the value of P" (crucially, the caption of his table of p-values for χ^2 given at the end of the paper is "Values of P for..."). Fisher's textbook used "value of P" and simply "P" interchangeably. According to David [8], the "value of P" first (?) morphed into "P value" in [10, Remark on p. 30] (which is a great progress since his previous paper [7], according to which this happened in a 1960 book). Deming's use of the form "P values" is consistent with his use of the notation $P(f)$ to refer to the p-values produced by a test statistic f (such as Karl Pearson's χ or Fisher's z).

Randomized p-values might have been first introduced in print explicitly and in a fairly general form (for integer-valued test statistics) by Stevens in 1950 [31, Section 4]. Shortly before that, Anscombe in his discussion paper [1] had introduced randomized p-values in the special cases of Fisher's exact test (Section 5.07) and confidence bounds for the parameter of the Bernoulli model (Section 5.17, slightly less explicitly). But even at the time, this was not a novel idea: e.g., Egon Pearson in his 1950 paper [28] defending randomized p-values says: "The possibility of this conversion has been recognized by statisticians for a number of years" (Section 1); here "conversion" is his term for complementing the test statistic by a separate "random experiment." At about the same time, but less explicitly, randomized p-values were used by Eudey [13] and Tocher [35]. Randomized p-values were later used in the well-known book [29] (Section 1.5.5).

2.5 Contents of the rest of this article

Section 3 recalls several basic definitions which the reader can skim through and then consult if and when needed. Section 4 goes into details of preorder topologies which are important in the following sections.

The next two sections, 5–6, introduce two of the three main languages for talking about p-values that we mentioned in Section 1, although there we stop short of actually defining p-values and only lay the groundwork by formalizing (in two different ways) the impugning power of different outcomes of the trial. Section 5 is devoted to the most traditional method, going back to Arbuthnott and Laplace (see Subsection 2.4), which starts from choosing a test statistic; this method is used in Cox and Hinkley's textbook [5] and countless other textbooks. Section 6 describes a language, conceptually the simplest of the three languages, which starts from the notion of preorder and directly models the impugning power of various outcomes. A third language, that of nested rejection regions, is standard in the algorithmic theory of randomness [26] and was adopted in the newest edition [25, Section 3.3] of Lehmann's classical textbook; it was

discussed in [17] in the simplest discrete case (and is discussed in [18] in general), but in this article we concentrate on the first two languages. We discuss the assumptions under which the two languages are equivalent.

The following three sections, 7–9 give details of how p-values emerge using the two languages and give a more abstract definition of "p-functions", i.e., functions producing p-values (with Sections 7–8 concentrating on the former and Section 9 on the latter).

While Sections 5–9 deal with the standard picture, in which test statistics are real-valued, the following three sections, 10–12, generalize the notion of test statistic, show the usefulness of the generalization in the case of randomized p-values, and argue for a wider use of the generalization. (In Section 12, we may lose many of our readers lacking statistics background.)

2.6 Final remark

Article [17] was an appetizer for this article. There we restricted attention to the discrete case with no outcomes of zero probability. That article was addressed to computer science audience. It required only baby probability theory. Here again we address a rather general audience. We have tried to make this article as self-contained as possible. A modest mathematical background should be sufficient for understanding.

3 Preliminaries

For readers' convenience, we recall here various definitions and terms.

3.1 Preorders

A *preorder* of a set Ω is a binary relation $\leq$ on Ω that is reflexive and transitive, so that for all $x, y, z \in \Omega$ we have

$$\begin{aligned} &x \leq x, \\ &\text{if } x \leq y \text{ and } y \leq z \text{ then } x \leq z. \end{aligned}$$

A preorder $\leq$ of Ω is *linear* if for all $x, y \in \Omega$ we have

$$x \leq y \text{ or } y \leq x.$$

For linear preorders we write $x < y$ to mean that $y \leq x$ is false. A preorder $\leq$ of Ω is called a *linear order* or *total order* or just *order* if for all $x, y \in \Omega$ we have

$$x < y \text{ or } x = y \text{ or } x > y.$$

Proviso 3.1. In the rest of the article, preorders are by default linear.

Let $(\Omega, \leq)$ be a preordered set. Points (or elements) x, y are *quasi-equal* if $x \leq y$ and $y \leq x$. The point set $\mathrm{QE}(x) = \{y : x \leq y \leq x\}$ is the quasi-equality class of point x. The induced relation

$$\mathrm{QE}(x) \leq \mathrm{QE}(y) \iff x \leq y$$

on the quasi-equality classes is an order called the *quotient order* of preorder $\leq$.

A point set of a preordered set $(\Omega, \leq)$ is *convex* if, together with any points $x \leq y$, it contains every point z with $x \leq z \leq y$. Convex point sets are also known as *segments*. Notice that segments are closed under quasi-equality; if a segment contains a point x then it contains all points quasi-equal to x. Now consider a point set X closed under quasi-equality. A segment S is a maximal segment (of $\leq$) in X if $S \subseteq X$ and there is no segment S' such that $S \subset S' \subseteq X$. For every point $x \in X$, there is a unique maximal segment S in X that contains x; of course S could be just the equality class of x. Thus every point set closed under quasi-equality is the disjoint union (possibly uncountable) of its maximal segments.

An *initial segment* of $(\Omega, \leq)$ is a segment that, together with every point x, contains every $y \leq x$. A *final segment* of $(\Omega, \leq)$ is a segment that, together with every point x, contains every $y \geq x$. We use the standard notation:

$$\begin{aligned}
(x, y) &= \{z : x < z < y\}, & [x, y) &= \{z : x \leq z < y\}, \\
(x, y] &= \{z : x < z \leq y\}, & [x, y] &= \{z : x \leq z \leq y\}, \\
(-\infty, x) &= \{y : y < x\}, & (-\infty, x] &= \{y : y \leq x\}, \\
(x, \infty) &= \{y : x < y\}, & [x, \infty) &= \{y : x \leq y\}.
\end{aligned}$$

But not every segment has one of these 8 forms. Think for example of the segment of rational numbers q with $q^2 < 2$ in the usual order of rationals.

3.2 Measurable spaces

Proviso 3.2. Countable means finite or infinite countable. So when we say that a set is countable, we do not exclude that it is finite or even empty.

A *σ-algebra* on a set Ω is any collection of subsets of Ω that contains the empty set and is closed under complements and countable unions. The *tensor product* $\Sigma_1 \otimes \Sigma_2$ of σ-algebras Σ_1, Σ_2 on some sets Ω_1, Ω_2 respectively is the least σ-algebra on the direct product $\Omega_1 \times \Omega_2$ containing the sets $X_1 \times X_2$ where $X_1 \in \Sigma_1$ and $X_2 \in \Sigma_2$.

In a topological space, the *Borel σ-algebra* is the least σ-algebra that contains the open sets. Members of the Borel σ-algebra are *Borel sets*.

A *measurable space* is a pair (Ω, Σ) where Ω is a set and Σ is a σ-algebra on Ω. In this context, Ω is the sample space, and its elements are the sample points. One example of a measurable space is the real line $\mathbb{R}$ with the Borel σ-algebra.

The *product measurable space* of measurable spaces (Ω_1, Σ_1) and (Ω_2, Σ_2) is the measurable space $\big(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2\big)$.

3.3 Measure spaces

A *measure* μ on a measurable space (Ω, Σ) is a countably additive function from Σ to the extended real segment $[0, \infty]$. A *measure space* is a triple (Ω, Σ, μ) where (Ω, Σ) is a measurable space and μ is a measure on (Ω, Σ).

On the real line with Borel σ-algebra, the *standard Borel measure* is the unique measure with $\mu[a, b] = b - a$ for any reals $a \leq b$.

A *null set* in a measure space (Ω, Σ, μ) is a measurable set of measure zero. Subsets of null sets are *negligible sets*. The measure μ and the measure space (Ω, Σ, μ) are *complete* if every negligible set is measurable. Every measure space (Ω, Σ, μ) can be extended to a complete measure space (Ω, Σ', μ') where $\Sigma' = \{S \cup N : S \in \Sigma \text{ and } N \text{ is negligible}\}$ and $\mu'(S \cup N) = \mu(S)$. The measure μ' is the *completion* of measure μ. On the real line, the Lebesgue measure is the completion of the standard Borel measure.

3.4 Probability spaces and trials

A *probability space* is a measure space $(\Omega, \Sigma, \mathbf{P})$ where $\mathbf{P}(\Omega) = 1$. $\mathbf{P}$ is the *probability measure* or the *probability distribution* of the probability space. Probability measures are continuous from above, in the sense that $P(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$ for any nested sequence $A_1 \supseteq A_2 \supseteq \cdots$ of elements of Σ, and continuous from below, in the sense that $P(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$ for any nested sequence $A_1 \subseteq A_2 \subseteq \cdots$ of elements of Σ.

Think of the probability space $(\Omega, \Sigma, \mathbf{P})$ as a probability trial. Sample points are outcomes, measurable point sets are events, and $\mathbf{P}$ assigns probabilities to events. $\mathbf{P}$ is the *null hypothesis* of the trial. Think of it as an alleged probability distribution. The purpose of the trial is to test $\mathbf{P}$.

The *product measure* $\mathbf{P}_1 \times \mathbf{P}_2$ of probability measures $\mathbf{P}_1$, $\mathbf{P}_2$ on measurable spaces (Ω_1, Σ_1), (Ω_2, Σ_2) respectively is the unique probability measure on $(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2)$ such that $\mathbf{P}(X_1 \times X_2) = \mathbf{P}_1(X_1) \cdot \mathbf{P}_2(X_2)$ for $X_1 \in \Sigma_1$ and $X_2 \in \Sigma_2$. (The existence and uniqueness of the product measure is often regarded as part of Fubini's theorem.)

4 Preorder topology

For future use, we need some facts about preorder topologies. Recall that our preorders are linear and that a set is countable if it is finite or infinite countable. We need yet another convention.

Proviso 4.1. By default, our sequences are countable (unless we say explicitly that a sequence in question is transfinite).

A preorder $\leq$ of a nonempty set Ω induces a *preorder topology* on Ω where the open sets are unions of segments of the form (x, y), $(-\infty, x)$, (x, ∞) and $(-\infty, \infty)$. In other words, such segments of $\leq$ form the base of the preorder topology. In the case where $\leq$ is a linear order, one speaks about order topology.

A topology is *second-countable* [38, Section 16] if it has a countable base of open sets. In other words, a topology is second-countable if there is a countable collection $\mathcal{W}$ of open sets such that, for any open set U and any point $x \in U$, we have $x \in W \subseteq U$ for some $W \in \mathcal{W}$. The countable base $\mathcal{W}$ *witnesses* the second-countability of the topology. For example, the set of real segments (a, b) with rational end-points a, b witnesses the second-countability of the real line $\mathbb{R}$.

We say that a preorder $\leq$ is second-countable if the preorder topology that $\leq$ induces is so. An *open segment* is a set that is both open and a segment.

Lemma 4.2. *For any preorder $\leq$, the following claims are equivalent.*

1. *The preorder is second-countable.*
2. *There is a countable set $I_1, I_2, \ldots$ of open segments such that, for any open segment J and point $x \in J$, we have $x \in I_n \subseteq J$ for some I_n.*
3. *There are countable families $L_1, L_2, \ldots$ and $R_1, R_2, \ldots$ of initial and final open segments respectively such that*
   - *for any initial open segment X and point $x \in X$, we have $x \in L_n \subseteq X$ for some L_n and*
   - *for any final open segment X and point $x \in X$, we have $x \in R_n \subseteq X$ for some R_n.*

Proof. Claim 2 obviously implies Claim 1. To prove that Claim 1 implies Claim 2, let open sets $U_1, U_2, \ldots$ witness the second countability of $\leq$. Every U_n splits into countably many disjoint open segments. Indeed, U_n is closed under quasi-equality and thus is the disjoint union of its maximal segments. If all these segments are open then each of them contains a separate U_i and so the number of these segments is countable. It remains to check that every maximal segment S in U_n is open. Suppose that $x \in S$. It suffices to prove that S includes some open segment $(a, b) \ni x$ (where we allow $a = -\infty$ or $b = \infty$). Since U_n is open, it contains an open segment $(a, b) \ni x$. Since $S \cup (a, b)$ is convex and S is maximal in U_n, S includes (a, b).

The desired collection $I_1, I_2, \ldots$ comprises all the component open segments of all U_n. To finish the proof, we check that Claim 2 is equivalent to Claim 3.

$2 \Longrightarrow 3$. Let $I_1, I_2, \ldots$ be as in Claim 2, and define

$$L_n = \bigcup_{x \in I_n} (-\infty, x],$$
$$R_n = \bigcup_{x \in I_n} [x, \infty).$$

If X is an initial (resp. final) open segment and $x \in X$ then, by Claim 2, $x \in I_n \subseteq X$ for some I_n. It follows that $x \in L_n \subseteq X$ (resp. $x \in R_n \subseteq X$).

$3 \Longrightarrow 2$. Let L_n, R_n be as in Claim 3. The desired base $I_1, I_2, \ldots$ consists of the intersections $L_m \cap R_n$. Indeed, if J is an open segment containing a point

x, let $L = \bigcup_{y \in J}(-\infty, y]$ and $R = \bigcup_{y \in J}[y, \infty)$. By Claim 3, there exist L_m, R_n such that $x \in L_m \subseteq L$ and $x \in R_n \subseteq R$. Then $x \in L_m \cap R_n \subseteq J$. $\square$

A topological space is *separable* if it has a countable subset that is dense, i.e., intersects every nonempty open set. A pair of points $x < y$ form a *jump* in a preorder $\leq$ if there is no point z with $x < z < y$, so that x is an immediate predecessor of y, and y is an immediate successor of x. We identify jumps $x_1 < y_1$ and $x_2 < y_2$ if x_1, x_2 are quasi-equal (and therefore y_1, y_2 are quasi-equal). So jumps $x_1 < y_1$ and $x_2 < y_2$ are distinct if and only if $x_1 < x_2$ or $x_2 < x_1$.

To keep this article self-contained, we prove the following three lemmas here, even though the facts are known for orders (see, e.g., [3, Section VIII.11], although our references will be to [4]), and the case of preorders is very similar.

Lemma 4.3 (cf. [4, Lemma 3])**.** *A preorder $\leq$ is second-countable if and only it is separable and has only countably many distinct jumps.*

Proof. First suppose that $\leq$ is second-countable, and let open sets $U_1, U_2, \ldots$ witness that $\leq$ is second countable. Pick any point $x_n \in U_n$. The points x_n witness that $\leq$ is separable.

To prove that there are only countably many jumps, let $L_1, L_2, \ldots$ be as in Claim 3 of Lemma 4.2. For every jump $x < x'$, there is a positive integer n_x such that $x \in L_{n_x} \subseteq (-\infty, x]$. Note that x is a maximal element in L_{n_x}. It follows that, if jumps $x < x'$ and $y < y'$ are distinct then $n_x \neq n_y$. Since there are only countably many positive integers, there are only countably many jumps.

Second suppose $\leq$ is separable and has only countably many jumps. Then there is a countable point set $C = \{a_1, a_2, \ldots\}$ such that

- it is dense in $\leq$;
- for every jump $x < y$, the set C contains points quasi-equal to x and to y,
- if x is a minimal or maximal point in $\leq$ then C contains a point quasi-equal to x.

Let $L_n = (-\infty, a_n)$ and $R_n = (a_n, \infty)$, and let $L_0 = R_0 = (-\infty, \infty)$. By Lemma 4.2 and the initial/final symmetry, it suffices to prove that, for any initial open segment X and any $x \in X$, there exists n such that $x \in L_n \subseteq X$.

If the open set $(x, \infty) \cap X$ is nonempty, some $a_n \in (x, \infty) \cap X$ so that $x < a_n \in X$ and therefore $x \in L_n \subseteq X$.

If $(x, \infty) \cap X$ is empty then x is a maximal element of X. But X is open. It follows that either x is maximal in $\leq$ or else x has a successor x', so that $x < x'$ is a jump. In the first case, $x \in L_0 \subseteq X$. In the second case, there is an n such that a_n is quasi equal to x' and therefore $x \in L_n \subseteq X$. $\square$

If $(A, \leq)$ is a preordered set and $\emptyset \neq B \subseteq A$ then there are two natural topologies on B. One is the preorder topology of B, or more exactly the preorder topology of $(B, \leq_B)$ where $\leq_B$ is the restriction of $\leq$ to B; the open sets of the

preorder topology of B are unions of segments (x, y) of $\leq_B$ (where we allow $x = -\infty$ or $y = \infty$). The second one is the subspace topology of B; the open sets of the subspace topology have the form $B \cap U$ where U is an open set in the preorder topology of $(A, \leq)$. To demonstrate the difference between the two topologies, let $(A, \leq)$ be the real line, and let B consist of the negative reals and the number 1. The singleton set $\{1\}$ is open in the subspace topology of B but not in the preorder topology of B.

Lemma 4.4 (cf. [4, Theorem II]). *Suppose that a preordered set $(A, \leq)$ is second-countable, and let $\emptyset \neq B \subseteq A$. Both the preorder topology of B and the subspace topology of B are second-countable.*

Proof. The subspace topology is second-countable with the base sets of the form $B \cap U$ where U belongs to a countable base of A.

By Lemma 4.3, the subspace topology is separable. If a countable point set C witnesses the separability of the subspace topology, it also witnesses the separability of the preorder topology because every open segment (b_1, b_2) of B has the form $B \cap U$ where U is the open segment (b_1, b_2) in A.

A jump (b_1, b_2) of B may give rise to a nonempty open segment (b_1, b_2) of A. But for different jumps of B these segments of A are disjoint, and therefore there are only countably many of the former. By Lemma 4.3, the preorder topology is second countable. □

Lemma 4.5 (cf. [4, Theorem II]). *An ordered set $(\Omega, \leq)$ is second-countable if and only if it is order isomorphic to a subset S of the real line $\mathbb{R}$ with the standard order of reals.*

Proof. The if direction follows from the previous lemma. We prove the only-if direction.

Suppose that $(\Omega, \leq)$ is second-countable. By Lemma 4.3, $(\Omega, \leq)$ is separable and has only countably many jumps. Therefore there is a countable point set $C = \{a_1, a_2, \ldots\}$ such that C is dense in $\leq$, and C contains every jump in $\leq$, and if x is minimal or maximal in $\leq$ then $x \in C$.

It was known already to Cantor that every countable order is order isomorphic to a set of rational numbers with the standard order. In particular this applies to our C with the inherited order. The desired order isomorphism f is easy to construct. Suppose that we have already constructed $f(a_1), \ldots, f(a_n)$, and let $A_n = \{a_1, \ldots, a_n\} \subset C$, $B_n = \{f(a_1), \ldots, f(a_n)\} \subset \mathbb{Q}$. If $a_{n+1} < \min(A_n)$, set $f(a_{n+1}) = \min(B_n) - 1$. If $a_{n+1} > \max(A_n)$, set $f(a_{n+1}) = \max(B_n) + 1$. Otherwise, let $a_i = \max\{a \in A_n : a < a_{n+1}\}$ and $a_j = \min\{a \in A_n : a > a_{n+1}\}$. Set $f(a_{n+1}) = (f(a_i) + f(a_j))/2$.

We extend the order isomorphism f to the whole $(\Omega, \leq)$ as follows: If x is a point in $\leq$ that is outside of C then $x = \sup\{a \in C : a < x\}$, and we set $f(x) = \sup\{f(a) : a \in C \text{ and } a < x\}$. □

Lemma 4.6. *Let $(R, \leq_R)$ be a second-countable ordered set and $f : \Omega \to R$. The preorder*

$$x \leq y \iff f(x) \leq_R f(y)$$

induced by f on Ω is second-countable.

Proof. If $\leq$ is a linear order, then f is an order embedding of $(\Omega, \leq)$ into $(R, \leq_R)$. By Lemma 4.4, $\leq$ is second-countable.

If $\leq$ is only a linear preorder, form the quotient order $(\Omega', \leq')$, lift f to work on Ω' and conclude that $(\Omega', \leq')$ is second-countable. Any witness of the second-countability of $(\Omega', \leq')$ gives rise to a witness of the second-countability of $(\Omega, \leq)$. Here's the argument in detail. Let

- x' be the quasi-equality class of $x \in \Omega$,
- $X' = \{x' : x \in X\}$ for any $X \subseteq \Omega$,
- $f'(x') = f(x)$,
- $x' \leq' y' \iff f'(x') \leq_R f'(y')$

so that $(\Omega', \leq')$ is the quotient of $(\Omega, \leq)$ over quasi-equality and $f' : \Omega' \to R$. By the first part of the proof, $\leq'$ is second-countable. Let open segments $J_1, J_2, \ldots$ of $\leq'$ witness the second-countability of $\leq'$, and let $I_n = \{x \in \Omega : x' \in J_n\}$. The open segments $\{I_1, I_2, \ldots\}$ witness the second-countability of $\leq$. □

5 Traditional test statistics

Convention 5.1. By default, a set of real numbers is measurable if it is Borel measurable.

Definition 5.2. A *traditional test statistic* for a measurable space (Ω, Σ) is a real-valued function f on Ω that is measurable, so that Σ contains the preimage $f^{-1}(B) = \{x \in \Omega : f(x) \in B\}$ of every Borel set B of reals.

Since the Borel sigma-algebra of the real line is generated by the initial segments $(-\infty, r]$, the measurability requirement in Definition 5.2 can be simplified to this: Every $[f \leq r]$ belongs to Σ.

In applied statistics, mostly traditional test statistics are used. In the next section, the notion of test statistic is slightly generalized. A more radical generalization is discussed in Sections 10–12.

How do we justify Convention 5.1? Well,

(a) we want that for any traditional test statistic f, all sets $[f \leq r]$ be measurable, and

(b) we are reluctant to impose additional constraints, i.e., constraints that do not follow from (a).

This brings us to the notion of universally measurable sets.

Universally measurable sets of reals are sets measurable with respect to every complete probability measure on $\mathbb{R}$ that measures all Borel sets. There exist universally measurable sets that are not Borel [21, Theorems 14.2 and 21.10].

Proposition 5.3. *If the probability measure* $\mathbf{P}$ *of our trial* $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$ *is complete then every traditional test statistic* f *on* $\mathcal{T}$ *respects (in the sense of having measurable pre-images) not only Borel sets but also all universally measurable sets.*

Proof. Consider a trial $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$ where $\mathbf{P}$ is complete. Any traditional test statistic f for $\mathcal{T}$ induces a complete measure $\mu(X) = \mathbf{P}[f^{-1}(X)]$ on $\mathbb{R}$. If a real set Y is universally measurable then Y is μ-measurable, and thus $f^{-1}(Y) \in \Sigma$. □

Since the operation of measure completion is rather innocent, the version of Convention 5.1 obtained by replacing "Borel measurable" with "universally measurable" is a defensible alternative convention.

6 Test preorders

Recall that our preorders are linear, and that the default σ-algebra on the real line $\mathbb{R}$ is the Borel σ-algebra. Let (Ω, Σ) be a measurable space.

Theorem 6.1. *For any real-valued function* f *on* Ω*, the preorder*

$$x \leq y \iff f(x) \leq f(y)$$

induced by f *on* Ω *is second-countable. Furthermore, if* f *is measurable then* $(-\infty, x] \in \Sigma$ *for all* $x \in \Omega$*.*

Proof. The first claim follows from Lemma 4.6.

Every $(-\infty, f(x)]$ is Borel. Since f is measurable,

$$(-\infty, x] = f^{-1}(-\infty, f(x)] \in \Sigma. \qquad \square$$

Theorem 6.2. *Any second-countable preorder* $\leq$ *on* Ω *is induced by some real-valued function* f*, so that* $x \leq y \iff f(x) \leq f(y)$*. Furthermore, if* $(-\infty, x] \in \Sigma$ *for all* $x \in \Omega$ *then such* f *is measurable.*

Proof. We assume that $\leq$ is a second-countable preorder on Ω, and we construct a function $f : \Omega \to \mathbb{R}$ that induces $\leq$ and which is measurable when all $(-\infty, x] \in \Sigma$.

To this end, we use notation of Lemma 4.6: x' is the quasi-equality class of x, and $(\Omega', \leq')$ is the quotient of $(\Omega, \leq)$ over quasi-equality. By Lemma 4.5, there is a function $g : \Omega' \to \mathbb{R}$ such that $x' \leq y' \iff g(x') \leq g(y')$. The desired $f : \Omega \to \mathbb{R}$ is given by $f(x) = g(x')$.

It remains to check that, if every $(-\infty, x] \in \Sigma$, then f is measurable. Since the Borel σ-algebra on $\mathbb{R}$ is the least σ-algebra containing all initial segments $(-\infty, r]$, it suffices to check that every initial segment $[f \leq r] = f^{-1}(-\infty, r] \in \Sigma$.

If $[f \leq r]$ is empty then it belongs to Σ. Suppose that it is nonempty. If $[f \leq r]$ has a maximal point x then $[f \leq r] = (-\infty, x] \in \Sigma$. Suppose there is no maximal point. By Lemma 4.3, $\leq$ is separable. It follows that there is an increasing sequence $x_1 < x_2 < \cdots$ in $[f \leq r]$ such that $[f \leq r] = \bigcup_n (-\infty, x_n] \in \Sigma$. □

Remark 6.3. The notion of separability is better known than that of second-countability, and one may be tempted to weaken the conditions of Theorem 6.2 by replacing second-countability with separability. The following counterexample shows that separability is insufficient. Let $\Omega = \mathbb{R} \times \{0, 1\}$ and $\leq$ be the lexicographic order on Ω: $(r, i) \leq (r', i')$ if and only if $r < r'$ or both $r = r'$ and $i \leq i'$. The condition of separability is clearly satisfied: the points (q, i) with rational q are dense in $\leq$. Yet, $\leq$ is not induced by any real-valued function. For, suppose that a real-valued function f induces $\leq$. Then $f(r, 0) < f(r, 1)$ for all real r. Thus $\mathbb{R}$ has uncountably many disjoint open segments, which is false. Remember that, according to Lemma 4.3, second countability is equivalent to the conjunction of separability and the countability of the number of jumps.

Since traditional test statistics are measurable real-valued functions, we have the following corollary.

Corollary 6.4. *A preorder $\leq$ on (Ω, Σ) is induced by a traditional test statistic if and only if it is second-countable and every initial segment $(-\infty, x]$ of $\leq$ belongs to Σ.*

Let us check that the initial segments $(-\infty, x]$ of a second countable $\leq$ generate the Borel σ-algebra; remember that the latter is defined as the smallest σ-algebra containing all open sets.

Lemma 6.5. *The Borel σ-algebra on a second countable preordered set $(R, \leq)$ is generated by the initial segments $(-\infty, x]$, $x \in R$.*

Proof. Since each $(-\infty, x]$ is a Borel set (being the complement of an open set), the σ-algebra generated by the initial segments $(-\infty, x]$, $x \in R$, is included in the Borel σ-algebra. Therefore, we are only required to prove that any open set in R is measurable with respect to the σ-algebra generated by the initial segments $(-\infty, x]$. By Lemma 4.2 (together with its proof) and the initial/final symmetry, it suffices to prove that any initial open segment X is measurable with respect to the σ-algebra generated by the initial segments $(-\infty, x]$. We apply the argument that we used in the last paragraph of the proof of Theorem 6.2. If X is empty then it belongs to Σ. If X has a maximal point x then $X = (-\infty, x] \in \Sigma$. Suppose X is non-empty without a maximal point. By Lemma 4.3, $\leq$ is separable. It follows that there is an increasing sequence $x_1 < x_2 < \cdots$ in X such that $X = \bigcup_n (-\infty, x_n] \in \Sigma$. □

Definition 6.6.

1. A *test preorder* on measurable space (Ω, Σ) is a second-countable preorder of Ω such that all initial segments $(-\infty, x]$ belong to Σ.

2. A *test statistic* on measurable space (Ω, Σ) is a measurable function from Ω to the real line $\mathbb{R}$ (in which case, f is a *traditional test statistic*) or to any other second-countable ordered codomain with Borel σ-algebra.

Traditionalists may argue that the generalization of the traditional test statistic in the last definition is vacuous, and in a sense it is vacuous. By Theorem 6.2, any test preorder on (Ω, Σ) is induced by a traditional test statistic. It follows that any nontraditional test statistic f can be replaced by a traditional test statistic f'. However, as the following schematic example shows, the nontraditional test statistic f may be much more convenient to work with. The example is inspired by the literature on randomized p-values but it does not presuppose the knowledge of randomized p-values, which will be introduced in Section 11.

Example 6.7. Let f be a traditional test statistic on a measurable space (Ω, Σ). Assume that f takes values in a countable subset R of the real line $\mathbb{R}$ with its usual order and Borel σ-algebra. (In the context of randomized p-values this may correspond to the case of a discrete sample space Ω, when randomizing p-values becomes particularly useful.) Set Ω' to the real segment $[0, 1]$ and Σ' to the Borel σ-algebra on $[0, 1]$. (In the context of randomized p-values, Ω' is interpreted as the range of random numbers generated by a random number generator.) Order $R \times [0, 1]$ lexicographically, so that

$$(p, r) \leq (q, s) \iff p < q \text{ or } (p = q \text{ and } r \leq s).$$

It is easy to check that the lexicographic order is second-countable (this uses the countability of R) and that the function $F(p, r) = (f(p), r)$ is a test statistic on the product measurable space $(\Omega \times \Omega', \Sigma \otimes \Sigma')$. While F is rather natural, a traditional test statistic inducing the same test preorder on $(\Omega \times \Omega', \Sigma \otimes \Sigma')$ may be rather involved: think, e.g., of the case $R = \mathbb{Q}$ (the set of all rationals).

Radicals may argue that our generalization of traditional test statistics is way too timid, that there are natural examples with more general test preorders. We agree with the radicals and return to the issue in Section 11.

7 Tightening

In this section we introduce p-values though without using the term for the time being. In the following two sections we will discuss the intuition behind our definitions.

A test preorder $\leq$ on a probability space $(\Omega, \Sigma, \mathbf{P})$ is *slack* if there are points $x < y$ with $\mathbf{P}(-\infty, x] = \mathbf{P}(-\infty, y]$, where $(-\infty, x]$, $(-\infty, y]$ are initial segments of $\leq$. In other words, $\leq$ is slack if and only if there are points $x < y$ with $\mathbf{P}(x, y] = 0$.

Definition 7.1. The *tightening* of a test preorder $\leq$ is the test preorder

$$x \mathrel{\widehat{\leq}} y \iff \mathbf{P}(-\infty, x] \leq \mathbf{P}(-\infty, y].$$

Two test preorders are *equivalent* if they have the same tightening. A test preorder is *tight* if coincides with its tightening.

Corollary 7.2.

- *A test preorder is tight if and only if it is not slack.*
- *A test preorder $\leq$ is tight if and only if $x \leq y \iff \mathbf{P}(-\infty, x] \leq \mathbf{P}(-\infty, y]$.*

Lemma 7.3. *Let $\leq$ be a test preorder.*

1. $\mathbf{P}(-\infty, x]_{\widehat{\leq}} = \mathbf{P}(-\infty, x]_{\leq}$.
2. $\widehat{\widehat{\leq}} = \widehat{\leq}$, *so that* $\widehat{\leq}$ *is tight.*
3. $\leq$ *and* $\widehat{\leq}$ *are equivalent.*
4. *Equivalent tight test preorders are identical.*

Proof. For brevity, $(-\infty, x]_{\leq}$ will be denoted $(-\infty, x]$, and subscript $\widehat{\leq}$ will be denoted by subscript T.

1. Clearly $(-\infty, x] \subseteq (-\infty, x]_T$ and so $\mathbf{P}(-\infty, x] \leq \mathbf{P}(-\infty, x]_T$. To prove the opposite inequality, $\mathbf{P}(-\infty, x]_T \leq \mathbf{P}(-\infty, x]$, we consider two cases.
   - If $(-\infty, x]_T = (-\infty, y]$ for some y, we have
   $$\mathbf{P}(-\infty, x]_T = \mathbf{P}(-\infty, y] = \mathbf{P}(-\infty, x],$$
   the last equality holding by the definition of $\widehat{\leq}$.
   - Now suppose $(-\infty, x]_T \neq (-\infty, y]$ for any $y \in \Omega$. Since $\leq$ is separable, there is a sequence of points $y_1 \leq y_2 \leq \cdots$ in $(-\infty, x]_T$ such that the union of $(-\infty, y_k]$ over k is $(-\infty, x]_T$. Since $\mathbf{P}$ is continuous from below,
   $$\mathbf{P}(-\infty, x]_T = \lim_k \mathbf{P}(-\infty, y_k] \leq \mathbf{P}(-\infty, x].$$
   Here we have used the inequality $\mathbf{P}(-\infty, y_k] \leq \mathbf{P}(-\infty, x]$, which follows from the definitions:
   $$y_k \in (-\infty, x]_T \iff y_k \widehat{\leq} x \iff \mathbf{P}(-\infty, y_k] \leq \mathbf{P}(-\infty, x].$$
2. We have
   $$\begin{aligned} x \widehat{\widehat{\leq}} y &\iff \mathbf{P}(-\infty, x]_T \leq \mathbf{P}(-\infty, y]_T && \text{by def. of } \widehat{\widehat{\leq}} \\ &\iff \mathbf{P}(-\infty, x] \leq \mathbf{P}(-\infty, y] && \text{by Claim 1} \\ &\iff x \widehat{\leq} y && \text{by def. of } \widehat{\leq} \end{aligned}$$
3. The claim follows from Claim 2.
4. The claim follows from Claim 2. $\square$

Definition 7.4. A test preorder $\leq$ *induces* the test statistic

$$\mathrm{TS}_{\leq}(x) = \mathbf{P}(-\infty, x]_{\leq}.$$

A test statistic f *induces* the test preorder $x \leq y \iff f(x) \leq f(y)$.

Lemma 7.5. *Equivalent test preorders induce the same test statistic.*

Proof. It suffices to prove that a test preorder $\leq$ and its tight version $\widehat{\leq}$ induce the same test statistic. To this end, use Claim 1 of Lemma 7.3. $\square$

A natural question arises what the tight version of a test statistic f is. The test preorder $\leq_f$ induced by a test statistic f contains all essential information about f even though particular values of f may be more convenient in some applications. The tightening of $\leq_f$ preserves all the essential information. So we can define the tightening of a test statistic f as the test statistic induced by $\leq_f$ (or, equivalently by the tightening of $\leq_f$). But there is an intrinsic way to tighten a test statistic. Recall that $[f \leq r]$ abbreviates $\{y : f(y) \leq r\}$.

Definition 7.6. The *tightening* of a test statistic f is the test statistic

$$\hat{f}(x) = \mathbf{P}[f \leq f(x)] \tag{3}$$

(which is the same expression as (1), for a good reason). Two test statistics are *equivalent* if they have the same tightening. A test statistic is *tight* if it is its own tightening.

Lemma 7.7. *Let $\leq$ be the test preorder induced by a test statistic f.*

1. $(-\infty, x] = [f \leq f(x)]$ *for all* x.
2. $\hat{f} = \mathrm{TS}_{\leq} = \mathrm{TS}_{\widehat{\leq}}$.
3. $\hat{f}$ *induces* $\widehat{\leq}$, *so that if* f *is tight then so is* $\leq$.
4. $\hat{\hat{f}} = \hat{f}$, *so that* $\hat{f}$ *is tight.*
5. f *and* $\hat{f}$ *are equivalent.*
6. *Equivalent tight test statistics are identical.*

Proof.

1. Since f induces $\leq$, we have $y \leq x \iff f(y) \leq f(x)$, so that $(-\infty, x] = [f \leq f(x)]$.
2. By Lemma 7.5 and Lemma 7.3(3), it suffices to prove that $\hat{f} = \mathrm{TS}_{\leq}$. We have $\mathrm{TS}_{\leq}(x) = \mathbf{P}(-\infty, x] = \mathbf{P}[f \leq f(x)] = \hat{f}(x)$ where the second equality follows from Claim 1.

3. We have

$$\begin{aligned}
\hat{f}(x) \le \hat{f}(y) &\iff \mathbf{P}[f \le f(x)] \le \mathbf{P}[f \le f(y)] && \text{by def. of } \hat{f} \\
&\iff \mathbf{P}(-\infty, x] \le \mathbf{P}(-\infty, y] && \text{by Claim 1} \\
&\iff x \mathrel{\widehat{\le}} y && \text{by def. of } \widehat{\le}
\end{aligned}$$

4. By Claim 3, $\hat{f}$ induces $\widehat{\le}$. By the first equality of Claim 2, with $\hat{f}$ and $\widehat{\le}$ playing the roles of f and $\le$ respectively, $\hat{\hat{f}} = \mathrm{TS}_{\widehat{\le}}$. But, again by Claim 2, $\mathrm{TS}_{\widehat{\le}} = \hat{f}$.

5. The claim follows from Claim 4.

6. The claim follows from Claim 4. □

Corollary 7.8.

1. *If a test preorder $\le$ induces a test statistic f then f induces $\widehat{\le}$ and is tight.*

2. *If a test statistic f induces a test preorder $\le$ then $\le$ induces the tight version $\hat{f}$ of f.*

Proof.

1. Suppose that $\le$ induces a test statistic f, so that $f(x) = \mathbf{P}(-\infty, x]$. The preorder induced by f is $f(x) \le f(y)$, i.e., $\mathbf{P}(-\infty, x] \le \mathbf{P}(-\infty, y]$, which coincides with $\widehat{\le}$. The tightening of f is, by Claim 1 of Lemma 7.3,

$$\hat{f}(x) = \mathbf{P}[f \le f(x)] = \mathbf{P}(-\infty, x]_{\widehat{\le}} = \mathbf{P}(-\infty, x]_{\le} = f(x).$$

2. Suppose that f induces $\le$. By Claim 1 of Lemma 7.7, $(-\infty, x] = [f \le f(x)]$. Hence $\mathbf{P}(-\infty, x] = \mathbf{P}[f \le f(x)]$; in other words, the test statistic induced by $\le$ coincides with $\hat{f}$. □

8 Exact p-values

In this section we introduce a different name for values taken by tight statistics, exact p-values. A more general notion of p-values will be given in the next section, where we will allow conservative p-values. (In this article we never consider approximate p-values, which are sometimes used in statistics, both theoretical and applied.)

Definition 8.1. Given a probability trial $(\Omega, \Sigma, \mathbf{P})$ furnished with a test preorder $\le$, the *exact p-value* of an outcome x is the probability $\mathbf{P}(-\infty, x]$.

Notice the dependence on the test preorder. Of course, the test preorder can also be given by means of a test statistic. Test statistics are especially popular in applications. In this connection, let us give an alternative definition of exact p-values.

Definition 8.2. Given a probability trial $(\Omega, \Sigma, \mathbf{P})$ furnished with a test statistic f, the *exact p-value* of an outcome x is the number $\hat{f}(x) = \mathbf{P}[f \leq f(x)]$.

Of course the two definitions are consistent.

Lemma 8.3.

1. *If a test statistic f induces a test preorder $\leq$ then $\mathbf{P}(-\infty, x] = \mathbf{P}[f \leq f(x)]$.*
2. *If a test preorder $\leq$ induces a test statistic f then $\mathbf{P}[f \leq f(x)] = \mathbf{P}(-\infty, x]$.*

Proof.

1. The events $(-\infty, x]$ and $[f \leq f(x)]$ coincide.
2. The equality $\mathbf{P}[f \leq f(x)] = \mathbf{P}(-\infty, x]$ can be written as $\mathbf{P}[f \leq f(x)] = f(x)$ and follows from f being tight (by Claim 2 of Corollary 7.8 in combination with every test preorder being induced by some test statistic, asserted by Theorem 6.2). □

Let us now compute the exact p-value for Example 2.1.

Example 8.4 (Example 2.1 continued)**.** With respect to the test statistic m, the exact p-value of the actual outcome is

$$\mathbf{P}[m \leq 1] = \frac{2(1+41)}{2^{41}} < \frac{2 \times 2^6}{2^{41}} = 2^{-34}.$$

(Our statistical readers will appreciate that while this p-value is formally correct, in real life it should be taken with a grain of salt because of the likely implicit multiple hypothesis testing: e.g., the Essex county is not the only county in the USA where such an anomaly could have occurred.)

9 p-functions

In this section we will introduce and discuss the fundamental property of validity for p-values (cf. (2) in the Preamble). Let ε range over the segment $[0, 1]$

Lemma 9.1. *Let f be a tight test statistic.*

1. $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ *if* $\varepsilon \in \mathrm{Range}(f)$.
2. $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ *if* $\varepsilon = \sup S$ *for some* $S \subseteq \mathrm{Range}(f)$.
3. $\mathbf{P}[f \leq \varepsilon] \leq \varepsilon$ *for every* ε.

Proof.

1. If $\varepsilon = f(x)$ then

$$\begin{aligned}\mathbf{P}[f \le \varepsilon] &= \mathbf{P}[f \le f(x)] \\ &= f(x) && \text{as } f \text{ is tight} \\ &= \varepsilon.\end{aligned}$$

2. If $\varepsilon = \sup S$ for some $S \subseteq \text{Range}(f)$, then there is a sequence $s_1 < s_2 < \cdots$ of elements of S converging to ε. Then, assuming $\varepsilon \notin \text{Range}(f)$ (the case $\varepsilon \in \text{Range}(f)$ was covered by Claim 1),

$$\begin{aligned}\mathbf{P}[f \le \varepsilon] = \mathbf{P}[f < \varepsilon] = \mathbf{P}\left(\cup_{n=1}^{\infty}[f \le s_n]\right) \\ = \lim_{n\to\infty} \mathbf{P}\left([f \le s_n]\right) = \lim_{n\to\infty} s_n = \varepsilon.\end{aligned}$$

3. Let $\varepsilon_0 = \sup\{s \in \text{Range}(f) : s \le \varepsilon\}$. Then $[f \le \varepsilon] = [f \le \varepsilon_0]$ and therefore, by Claim 2, $\mathbf{P}[f \le \varepsilon] = \mathbf{P}[f \le \varepsilon_0] = \varepsilon_0 \le \varepsilon$. □

Lemma 9.1 is the fundamental property of validity of p-values. It means that, for a small significance level $\varepsilon > 0$, we are entitled, by the Cournot principle, to reject the null hypothesis once we observe a p-value $\le \varepsilon$. In particular, in Example 8.4 we are entitled to reject the null hypothesis since the p-value that we obtained is so small, smaller than any reasonable significance level. In the next definition we isolate this property of validity.

Definition 9.2.

- A *p-function* is a test statistic f such that $\mathbf{P}[f \le \varepsilon] \le \varepsilon$ for every ε.
- A p-function f is *exact* if $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for every $\varepsilon \in \text{Range}(f)$; otherwise f is *conservative*.
- Values of a p-function are *p-values*.

If f is a p-function then cf is a p-function for every $c \ge 1$. Indeed

$$\mathbf{P}[cf \le \varepsilon] = \mathbf{P}[f \le \varepsilon/c] \le \varepsilon/c \le \varepsilon.$$

If $c \in (0,1)$ then cf may not be a p-function. In particular if $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for at least one $\varepsilon > 0$ then cf is not a p-function because, for that ε, we have

$$\mathbf{P}[cf \le c\varepsilon] = \mathbf{P}[f \le \varepsilon] = \varepsilon > c\varepsilon.$$

Theorem 9.3. *For any test statistic f, the following two claims are equivalent:*

1. *f is tight.*
2. *f is an exact p-function.*

Proof. The implication $1 \implies 2$ is proven in Lemma 9.1. To prove the implication $2 \implies 1$, assume that f is an exact p-function, $x \in \Omega$ is an arbitrary outcome and $\varepsilon = f(x)$. We have

$$\begin{aligned} \hat{f}(x) = \mathbf{P}[f \le f(x)] &= \mathbf{P}[f \le \varepsilon] \\ &= \varepsilon && \text{as } f \text{ is exact} \\ &= f(x). \end{aligned}$$

□

Corollary 9.4.

- *Every test preorder induces a unique exact p-function.*
- *Every test statistic f is equivalent to a unique exact p-function.*

Remark 9.5. Conservative p-functions appear naturally when adjusting p-values for multiple hypothesis testing. Consider, e.g., a case where the same outcome of a probability trial is analyzed n times, using test statistics $f_1, \ldots, f_n$ and getting p-values $p_1, \ldots, p_n$. Of course one would love to report the p-value $\min\{p_1, \ldots, p_n\}$ but it needs to be adjusted for the multiple testing of the null hypothesis. The simplest Bonferroni correction is one way to adjust that p-value, by multiplying it by n (see, e.g., [20]). The p-value $n \min\{p_1, \ldots, p_n\}$ is valid though usually conservative. More exactly the test statistic

$$f = n \min(f_1, \ldots, f_n)$$

is a p-function, typically conservative. Indeed,

$$\mathbf{P}[f \le \varepsilon] = \mathbf{P}\big(\cup_{i=1}^n [f_i \le \varepsilon/n] \big) \le \sum_{i=1}^n \mathbf{P}[f_i \le \varepsilon/n] \le \sum_{i=1}^n \varepsilon/n = \varepsilon.$$

10 ω_1-short preorders

In Example 1.1 we mentioned that ω_1 is the set of all countable ordinals and that ordinals are naturally ordered by inclusion. Think of ω_1 as the linear order of countable ordinals. A mapping f whose domain and codomain are both preordered sets is *order-preserving* (or an *embedding*) if $x < y$ implies $f(x) < f(y)$ for all x and y in its domain.

Definition 10.1. A preorder $\le$ of a set R is ω_1*-short* if there is no order-preserving map from ω_1 to $(R, \le)$.

Remark 10.2. A standard notion of shortness (not used in this article) is a two-sided version of our definition (the order is called short if neither ω_1 nor its reverse ω_1^* can be embedded into it): see [16] and [30, p. 88]. It is interesting that in the context of p-values the symmetry between ω_1 and ω_1^* is lost, and ω_1^* is not nearly as dangerous as ω_1. Replacing in Example 1.1 the order ω_1 by the reverse order ω_1^*, we will obtain a p-value that is identically equal to 1; it's valid albeit very conservative (and this can easily happen even for very simple test statistics: e.g., whenever the size of f's codomain is 1).

There is a useful positive characterization of ω_1-short preorders. We start with recalling the notion of cofinality. Sets X, Y of points of preorder $\leq$ are *cofinal* if for every $x \in X$ there is $y \geq x$ in Y and for every $y \in Y$ there is $x \geq y$ in X.

Lemma 10.3. *A preorder $\leq$ is ω_1-short if and only if every initial segment without a maximal element includes a cofinal sequence $x_1 < x_2 < \cdots$.*

Proof. It suffices to prove that ω_1 can be embedded into $\leq$ if and only if there is an initial segment L of $\leq$ without a maximal element such that no sequence $x_1 < x_2 < \cdots$ is cofinal with L. If η is an order-preserving map from ω_1 to the underlying set of $\leq$, the desired L is the least initial segment that includes the range of η. Indeed, consider any sequence $x_1 < x_2 < \cdots$ in L. For every x_n there is a countable ordinal α_n such that $x_n \leq \eta(\alpha_n)$. Assuming $\alpha_1 < \alpha_2 < \cdots$, $\sup_n \alpha_n$ is another countable ordinal β so that $\eta(\beta) \in L$ and $\eta(\beta) > \eta(\alpha_n) \geq x_n$ for all n.

In the opposite direction, suppose that L is an initial segment of $\leq$ without a maximal element and such that no sequence $x_1 < x_2 < \cdots$ is cofinal with L. We choose y_0 arbitrarily in L. For any countable β, if we have constructed a transfinite sequence $\langle y_\alpha : \alpha < \beta \rangle$, that transfinite sequence (containing countably many elements) is not cofinal with L, and so we can choose y_β in L greater than all y_α with $\alpha < \beta$. This way we construct a transfinite sequence $\langle y_\alpha : \alpha < \omega_1 \rangle$ and an embedding $\alpha \mapsto y_\alpha$ of ω_1 to $\leq$. □

Remark 10.4. The argument in the proof of Lemma 10.3 (second paragraph) is an instance of what is known as definition by transfinite induction (in our case, over the countable ordinals). For details, see the transfinite recursion theorem in [19], Section 18. It becomes applicable if we fix a choice function that maps every transfinite sequence $\langle y_\alpha : \alpha < \beta \rangle$ for every countable ordinal β to y_β satisfying our desideratum (in this particular case, y_β being greater than all y_α with $\alpha < \beta$).

Definition 10.5. A *generalized test statistic* is a measurable function $f : \Omega \to R$ taking values in an ordered set R (perhaps different from the real line $\mathbb{R}$) equipped with a σ-algebra making all initial segments of the form $(-\infty, r]$, $r \in R$, measurable. Its *tightening* $\hat{f}$ is defined, as usual, by (3):

$$\hat{f}(x) = \mathbf{P}[f \leq f(x)].$$

Theorem 10.6. *If the codomain of f is ω_1-short, the tightening $\hat{f}$ of f is a p-function. If an ordered set R is not ω_1-short, there exists a trial $(\Omega, \Sigma, \mathbf{P})$ and a generalized test statistic $f : \Omega \to R$ whose tightening is not a p-function.*

Proof. Our proof of the first statement will be by contrapositive. Suppose $\hat{f}$ is not a p-function. Fix an $\varepsilon \in [0, 1)$ such that $\mathbf{P}[\hat{f} \leq \varepsilon] > \varepsilon$. Set $E = [\hat{f} \leq \varepsilon]$, so that $\mathbf{P}(E) > \varepsilon$ whereas $\mathbf{P}[f \leq f(x)] \leq \epsilon$ for all $x \in E$. An *embedding* of an ordinal β into E is a family $\langle x_\alpha : \alpha < \beta \rangle$ of points of E such that $f(x_\alpha) < f(x_{\alpha'})$ for all $\alpha < \alpha' < \beta$. Our goal is to define an embedding $\langle x_\alpha : \alpha < \omega_1 \rangle$ of ω_1 into

E; we will do so by transfinite induction over the countable ordinals. Choose $x_0 \in E$ arbitrarily; this can be done since $\mathbf{P}(E) > \varepsilon$ and so E is non-empty. Suppose, for a countable β, we have constructed an embedding $\langle x_\alpha : \alpha < \beta \rangle$ of β into E. Since $\mathbf{P}(\cup_{\alpha<\beta}[f \le f(x_\alpha)]) \le \varepsilon$ (by the continuity of $\mathbf{P}$ from below and the countability of β) and $\mathbf{P}(E) > \varepsilon$, we can extend the family $\langle x_\alpha : \alpha < \beta \rangle$ by adding another point x_β of E with $f(x_\beta) > f(x_\alpha)$ for all $\alpha < \beta$. This way we construct an embedding of the whole of ω_1 into E.

On the other hand, suppose an ordered set R is not ω_1-short. We proceed as in Example 1.1. Set $\Omega = \omega_1$ with the same σ-algebra Σ and probability measure $\mathbf{P}$ as before, and let f be an order-preserving map from Ω to R (which exists by our assumption). We still have $\mathbf{P}[f \le f(x)] = 0$ for all $x \in \Omega$. □

Theorem 10.6 shows that the class of generalized test statistics taking values in an ω_1-short ordered set is particularly important; let's check directly that this class includes the traditional test statistics.

Proposition 10.7. *The set of real numbers $\mathbb{R}$ is ω_1-short.*

Proof. Suppose toward a contradiction that η is an order-preserving map from ω_1 to $\mathbb{R}$. Since there is a rational number q_α between any $\eta(\alpha)$ and $\eta(\alpha) + 1$, we obtain uncountably many distinct rationals q_α, which is impossible. □

11 Randomized p-values

In this section we discuss randomized p-values, which we already mentioned in Subsection 2.4, as a natural example of generalized test statistics taking values in ω_1-short orders. Randomized p-values arise naturally in situations where the distribution of the test statistic is not continuous. They are produced by generalized test statistics whose codomain is $\mathbb{R} \times [0, 1]$ with the lexicographic order; intuitively, we add a random number to the test statistic to break ties if there are any. This makes the distribution of a randomized p-value uniform over the segment $[0, 1]$ (as shown in Proposition 11.1 below).

As discussed earlier, Egon Pearson [28] defended randomized p-values in 1950, but he was mainly writing about the abstract notion. At this time there are at least two (and probably many more) fields of relatively practical statistics where randomized p-values are essential: multiple hypothesis testing in bioinformatics (see, e.g., [12]) and in on-line testing the hypothesis of exchangeability using conformal martingales (see, e.g., [36, Section 7.1]). In both cases p-values are used repeatedly a large number of times, and any non-uniformity of their distribution quickly accumulates and destroys the power of the overall procedure.

Recall that the product $\mathbf{P}_1 \times \mathbf{P}_2$ of probability measures $\mathbf{P}_1$ and $\mathbf{P}_2$ on measurable spaces (Ω_1, Σ_1) and (Ω_2, Σ_2) respectively is the unique probability measure on $(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2)$ with $\mathbf{P}(X_1 \times X_2) = \mathbf{P}_1(X_1) \cdot \mathbf{P}_2(X_2)$ for $X_1 \in \Sigma_1$ and $X_2 \in \Sigma_2$. We are essentially in the situation of Example 6.7, except that the assumption that the codomain of f is countable is dropped.

By Lemma 9.1, the tightening f of a test statistic is only guaranteed to satisfy the inequality $\mathbf{P}[f \leq \varepsilon] \leq \varepsilon$, and very simple examples show that $\mathbf{P}[f \leq \varepsilon] < \varepsilon$ is indeed possible: e.g., setting f to the tightening of the test statistic m in Examples 2.1 and 8.4 and ϵ to 2^{-34}, we have $\mathbf{P}[f \leq \varepsilon] = 2(1+41)2^{-41} < \varepsilon$. Randomized p-values are a way of achieving the equality $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ for all $\epsilon \in [0,1]$. Informally, we enrich our probability space by adding a random number generator and using its output for breaking ties in values of the test statistic for different outcomes.

Formally, we start from our usual setting of a given traditional test statistic $f : \Omega \to \mathbb{R}$ on a trial $(\Omega, \Sigma, \mathbf{P})$. The output of a random number generator is modelled as the trial $([0,1], \mathcal{B}, \mathcal{U})$, where $\mathcal{B}$ is the Borel σ-algebra on $[0,1]$ and $\mathcal{U}$ is the uniform probability measure on $([0,1], \mathcal{B})$. The overall trial is now the product

$$(\bar{\Omega}, \bar{\Sigma}, \bar{\mathbf{P}}) = (\Omega \times [0,1], \Sigma \otimes \mathcal{B}, \mathbf{P} \times \mathcal{U})$$

and the test statistic f on Ω is replaced by the finer test statistic

$$F(x,r) = (f(x), r)$$

on $\bar{\Omega}$. The order on the codomain $\mathbb{R} \times [0,1]$ of F is lexicographic,

$$(p,r) \leq (q,s) \iff p < q \text{ or } (p = q \text{ and } r \leq s).$$

Intuitively, this means that the impugning power of our test statistic is determined by f, and the outcome of the random number generator is only used for tie breaking. Let us call all functions F that can be obtained in this way *randomized test statistics*. They have the following useful property.

Proposition 11.1. *The tightening $\hat{F}$ of any randomized test statistic F satisfies $\bar{\mathbf{P}}[\hat{F} \leq \varepsilon] = \varepsilon$ for any $\epsilon \in [0,1]$.*

Proof. Fix any $\varepsilon \in (0,1)$ (it is clear that it suffices to consider this range of ε). The subset

$$\{t \in \mathbb{R} \times [0,1] : \bar{\mathbf{P}}[F \leq t] \leq \varepsilon\}$$

of $\mathbb{R} \times [0,1]$ is bounded from above, and we let (p,r) be its supremum. By the continuity of probability measures from below, $\bar{\mathbf{P}}[F \leq (p,r)] \leq \varepsilon$; let us check that in fact $\bar{\mathbf{P}}[F \leq (p,r)] = \varepsilon$. Indeed, suppose $\bar{\mathbf{P}}[F \leq (p,r)] = \varepsilon' < \varepsilon$. There are two cases:

- If $r < 1$, there exists $r' > r$ such that it is still true that $\bar{\mathbf{P}}[F \leq (p,r')] < \varepsilon$, which contradicts the definition of (p,r).
- If $r = 1$, by the continuity of probability measures from above there is a $p' > p$ such that $\mathbf{P}[f \in (p,p']] < \varepsilon - \varepsilon'$ (where f is first component of F, as above; remember that p and p' are real numbers). Therefore, $\bar{\mathbf{P}}[F \leq (p',0)] < \varepsilon$, and we again obtain a contradiction.

The equality $\bar{\mathbf{P}}[F \le (p,r)] = \varepsilon$ and the chain of equivalences

$$\hat{F}(x) \le \varepsilon \iff \mathbf{P}[F \le F(x)] \le \varepsilon \iff F(x) \le (p,r)$$

now imply

$$\bar{\mathbf{P}}[\hat{F} \le \varepsilon] = \bar{\mathbf{P}}[F \le (p,r)] = \varepsilon. \qquad \square$$

The value $\hat{F}(x)$ is the randomized p-value corresponding to an outcome x. According to Definition 6.6 and the example in Remark 6.3, the function F is not a test statistic (not a traditional test statistic and not even a test statistic in the wider sense of Definition 6.6). However, we are still on a relatively safe ground since its codomain is ω_1-short, as the following proposition says.

Proposition 11.2. *If R and S are ω_1-short orders, then the product $R \times S$ (ordered lexicographically) is an ω_1-short order.*

Proof. Let $\langle (r_\alpha, s_\alpha) : \alpha < \omega_1 \rangle$ be a strictly increasing transfinite sequence of elements of $R \times S$; we will arrive at a contradiction. Define a transfinite sequence r'_β of elements of R by transfinite induction as follows: $r'_0 = r_0$ and $r'_\beta = r_\alpha$ for the smallest α such that $r_\alpha > r'_\gamma$ for all $\gamma < \beta$. Since R is ω_1-short, this transfinite sequence is only defined on countably many ordinals; therefore, there are only countably many distinct r_α. And since for each r_α there are only countably many s_β such that $s_\beta = s_\alpha$, we obtain a contradiction. $\square$

Remark 11.3. When using randomized p-values, statisticians (and computer scientists in related areas) do not usually emphasize the use of generalized test statistics, which remain implicit. They prefer to define randomized p-values from scratch rather than using the standard definition (1). Namely, the usual definition (as given in, e.g., [12] and [36]) is

$$\hat{F}(x,r) = \mathbf{P}[f < f(x)] + r\mathbf{P}[f = f(x)], \tag{4}$$

where r is a random number in $[0,1]$. The only explicit use of the lexicographic order in connection with randomized p-values that we are aware of is in [6, p. 91, (3.4)].

12 An alternative to randomizing p-values

Many practical statisticians dislike the idea of random p-values. A pioneer of randomized p-values (Stevens on the last page of [31]) says: "We suppose that most people will find repugnant the idea of adding yet another random element to a result which is already subject to the errors of random sampling." When reporting on Anscombe's previous work [1] he says that the method "was there dismissed rather briefly as being unsatisfactory". Egon Pearson [28] was more positive but still admitted that there are "a number of objections" to the use of the method, "which many statisticians would regard as decisive."

A popular alternative to randomized p-values is mid-p-values, introduced by Lancaster in 1961 [22], his motivation being that in some cases computing

randomized p-values may be "time-consuming and even embarrassing to the statistician." The *mid-p-value* is defined to be the following modification of (4):

$$\mathbf{P}[f < f(x)] + \frac{1}{2}\mathbf{P}[f = f(x)]; \tag{5}$$

in other words it is defined to be the arithmetic mean of $\mathbf{P}[f < f(x)]$ and $\mathbf{P}[f \le f(x)]$ (whereas the randomized p-value is distributed uniformly between $\mathbf{P}[f < f(x)]$ and $\mathbf{P}[f \le f(x)]$). The corresponding *mid-p-function* maps each outcome $x \in \Omega$ to the mid-p-value (5). The definition of mid-p-values is natural, but the main problem with it is that mid-p-functions are not guaranteed to be p-functions (and they are not p-functions in interesting cases).

In randomized p-values we complement a given test statistic by a random number to break ties. We can easily imagine less repugnant (to use Stevens's expression) ways of tie-breaking using the lexicographic order on $\mathbb{R}^2$ (or $\mathbb{R}^k$ for $k > 2$). In this section we will discuss several specific examples, but it will be clear that the approach is general.

Wilcoxon's rank-sum test [37, 11] is the workhorse of nonparametric statistics. It is used for comparing two groups of observations (real numbers),

$$x_1, \dots, x_m \text{ and } y_1, \dots, y_n;$$

for simplicity we will assume that these observations are all different. The test statistic R_x is the sum of the ranks of the first m observations, in this case the sum of the ranks of $x_1, \dots, x_m$, where all $m + n$ observations are ranked from 1 (the smallest) to $m + n$ (the largest). The p-value corresponding to the given value R_x of the test statistic is the probability that $R'_x \le R_x$. Here R'_x is obtained by applying the test statistic to a random permutation $z_1, \dots, z_{m+n}$ of the $m + n$ observations; in other words R'_x is the sum of the ranks of observations $z_1, \dots, z_m$. Intuitively, we are testing the null hypothesis that all $m+n$ observations are drawn independently from the same continuous distribution on the real line against the alternative that the xs tend to be smaller than the ys.

Wilcoxon's rank-sum test is surprisingly efficient (see, e.g., [24, Section II.4]), but its p-values have a discrete distribution; first of all, it is clear that this distribution is concentrated on the set

$$\left\{ \frac{1}{\binom{m+n}{m}}, \frac{2}{\binom{m+n}{m}}, \dots, 1 \right\},$$

as for any permutation test. For example, when $m = n = 6$, it is concentrated on the set $\{1/924, 2/924, \dots, 1\}$. However, Table 5.1 in [29, Chapter 5] shows that Wilcoxon's test statistic is much cruder: it takes values in the set

$$\{1/924, 2/924, 4/924, 7/924, 12/924, 19/924, 30/924, 43/924, 61/924, \dots, 1\}$$

(where the ellipsis does not imply that the reader is supposed to be able to fill in the missing values).

To partially break the ties between the values of the test statistic (this is the only kind of ties we are interested in since all observations were assumed to be different), we can complement R_x by the value T_2 of the Fisher–Yates–Terry statistic, which is computed similarly to R_x but which applies a monotonic transformation to all the ranks changing the sums of ranks accordingly. The new test statistic, (R_x, T_2), takes values in $\mathbb{R}^2$ equipped with the lexicographic order. The p-value corresponding to the given value of (R_x, T_2) is the probability that $(R'_x, T'_2) \le (R_x, T_2)$, where (R'_x, T'_2) is computed by applying the test statistic to a random permutation of our $m+n$ observations. The range of the new p-function will be extended by adding the points in

$$\{5/924, 8/924, 10/924, 14/924, 15/924, 17/924, 21/924, \\ 22/924, 24/924, 26/924, 28/924, 32/924, 34/924, 35/924, \\ 37/924, 39/924, 40/924, 42/924, 48/924, 49/924, \ldots\},$$

as the same table in [29] shows (although we suspect that Pratt and Gibbons's results may be affected by the limited numeric accuracy of their calculations).

Lots of numbers of the form $k/924$, $k = 1, 2, \ldots$, are still missing, so we might add the van der Waerden statistic T_3, which is similar to T_2 but uses a slightly different monotonic transformation. Now the combined test statistic (R_x, T_2, T_3) takes values in $\mathbb{R}^3$ with the lexicographic order. However, adding T_3 will only add one number, $41/924$, to the intersection of the range of the p-function and $[0, 49/924]$. The reason for this poor tie-breaking performance of T_3 is that the test statistics T_2 and T_3 are so similar: both are based on monotonic transformations of ranks defined in terms of the Gaussian distribution. One way to break the ties more efficiently is to replace T_3 by a test statistic analogous to T_2 or T_3 but based on, e.g., monotonic transformations of ranks defined in terms of the Laplace distribution (popular in robust statistics).

An advantage of all these generalized test statistics is that the corresponding p-function will be valid (in the sense of (2)) whenever the observations are generated independently from the same continuous probability distributions on $\mathbb{R}$, and no parametric assumptions are required. A disadvantage is that even in the ideal situation (from the point of view of breaking ties) the range of the test statistic is $\{1/924, 2/924, \ldots, 1\}$ (where we revert to the normal use of the ellipsis: the reader is expected to fill it in). Therefore, the distribution is still not uniform on $[0, 1]$, although it is uniform on $\{1/924, 2/924, \ldots, 1\}$ (and so "almost uniform" on $[0, 1]$). In the rest of this section we will assume that we are in this ideal situation.

A drastic step perfectly breaking all ties (with probability one) but partly sacrificing the non-parametric character of the test is to add Student's [15] t-statistic

$$t = \frac{\bar{x} - \bar{y}}{S},$$

where we ignore an irrelevant constant factor and use the notation

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S = \sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2},$$

to the list $(R_x, T_2, \ldots)$, with the order on $(R_x, T_2, \ldots, t)$ still being lexicographic. The p-value corresponding to a given value of $(R_x, T_2, \ldots, t)$ is the probability that $(R'_x, T'_2, \ldots, t') \le (R_x, T_2, \ldots, t)$, where $(R'_x, T'_2, \ldots, t')$ is computed by applying the test statistic to a random sample of size $m+n$ drawn independently from the standard Gaussian distribution. Because of the nonparametric nature of the test statistics preceding t, the resulting p-function f will satisfy $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for any $\varepsilon \in \{1/924, 2/924, \ldots, 1\}$ and the power $\mathbf{P} = P^{m+n}$ of any continuous probability distribution P on $\mathbb{R}$. On the other hand, we will have $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for any $\varepsilon \in [0,1]$ and the power $\mathbf{P} = P^{m+n}$ of any Gaussian distribution P on the real line.

Remark 12.1. In principle, we could have used only traditional test statistics in this section since even the most complicated of our test statistics, $(R_x, T_2, \ldots, t)$ used in the last paragraph, had all components but one taking values in discrete sets. As we know, such orders can be embedded in the real line. However, the resulting traditional test statistic would be awkward, and it is much more natural to think in terms of the original statistic, such as $(R_x, T_2, \ldots, t)$ (cf. Example 6.7). And even for a practical statistician, it may be reassuring to know that she is on safe ground when using any $\mathbb{R}^k$-valued test statistics (with the lexicographic order on $\mathbb{R}^k$).

Acknowledgments

We thank Andreas Blass and Steffen Lauritzen for useful comments on a draft of this article.

References

[1] Anscombe, Francis J. (1948). The validity of comparative experiments (with discussion). *Journal of the Royal Statistical Society A* **111** 181–211.

[2] Arbuthnott, John (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society* **27** 186–190.

[3] Birkhoff, Garrett (1967). *Lattice Theory*. American Mathematical Society, Providence, RI, third edition. First edition: 1940. Second edition: 1948.

[4] Cater, Frank S. (1999). On order topologies and the real line. *Real Analysis Exchange* **25** 771–780.

[5] Cox, David R. and Hinkley, David V. (1974). *Theoretical Statistics*. Chapman & Hall, Boca Raton, FL.

[6] Coudin, Elise (2007). Inférence exacte et non paramétrique dans les modèles de régression et les modèles structurels en présence d'hétéroscédasticité de forme arbitraire. PhD thesis, University of Montreal.

[7] David, Herbert A. (1995). First (?) occurrence of common terms in statistics and probability. *American Statistician* **49** 121–133.

[8] David, Herbert A. (1998). First (?) occurrence of common terms in statistics and probability—a second list, with corrections. *American Statistician* **52** 36–40.

[9] David, Herbert A. (2001). First (?) occurrence of common terms in statistics and probability. In *Annotated Readings in the History of Statistics* (Herbert A. David and Anthony W. F. Edwards, eds.) 209–246 (Appendix B). Springer, New York. This list subsumes the 1995 and 1998 lists [7, 8].

[10] Deming, W. Edwards (1943). *Statistical Adjustment of Data*. Wiley, New York.

[11] Deuchler, Gustav (1914). Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. *Zeitschrift für Pädagogische Psychologie und Experimentelle Pädagogik* **15** 114–131, 145–159, and 229–242.

[12] Dickhaus, Thorsten, Strassburger, Klaus, Schunk, Daniel, Morcillo-Suarez, Carlos, Illig, Thomas, and Navarro, Arcadi (2012). How to analyze many contingency tables simultaneously in genetic association studies. *Statistical Applications in Genetics and Molecular Biology* **11**:4 Article 12.

[13] Eudey, Mark W. (1949). On the treatment of discontinuous random variables. PhD thesis, Statistical Laboratory, University of California, Berkeley, CA. Supervised by Jerzy Neyman.

[14] Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*. Hafner, New York.

[15] Gosset, William S. [Student] (1908). The probable error of a mean. *Biometrika* **6** 1–25.

[16] Gurevich, Yuri and Shelah, Saharon (1979). Modest theory of short chains. II. *J. Symbolic Logic* **44** 491–502.

[17] Gurevich, Yuri and Vovk, Vladimir (2016). Fundamentals of p-values: introduction. *Bulletin of the European Association for Theoretical Computer Science* **119** 79–92. `http://arxiv.org/abs/1603.03952` (version 1).

[18] Gurevich, Yuri and Vovk, Vladimir (2017). Test pyramids and p-values. Unpublished manuscript.

[19] Halmos, Paul R. (1960). *Naive Set Theory*. Van Nostrand, New York.

[20] Holm, Sture (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6** 65–70.

[21] Kechris, Alexander S. (1995). *Classical Descriptive Set Theory*. Springer, New York.

[22] Lancaster, H. Oliver (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association* **56** 223–234.

[23] Laplace, Pierre-Simon (1823). De l'action de la lune sur l'atmosphere. *Annales de Chimie et de Physique* **24** 280–294.

[24] Lehmann, Erich L. (2005). *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York.

[25] Lehmann, Erich L. and Romano, Joseph P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

[26] Martin-Löf, Per (1966). The definition of random sequences. *Information and Control* **9** 602–619.

[27] Pearson, Karl (1900). On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **50** 157–175.

[28] Pearson, Egon S. (1950). On questions raised by the combination of tests based on discontinuous distributions. *Biometrika* **37** 383–398.

[29] Pratt, John W. and Gibbons, Jean D. (1981). *Concepts of Nonparametric Theory*. Springer, New York.

[30] Rosenstein, Joseph G. (1982). *Linear orderings*. Academic Press, New York.

[31] Stevens, Wilfred L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37** 117–129.

[32] Stigler, Stephen M. (1975). Studies in the history of probability and statistics XXV. Napoleonic statistics: the work of Laplace. *Biometrika* **62** 503–517.

[33] Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press, Cambridge, MA.

[34] Sullivan, Joseph F. (1985). Court in Jersey upholds equal odds for all. *New York Times*, July 23.

[35] Tocher, Keith D. (1950). Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* **37** 130–144.

[36] Vovk, Vladimir, Gammerman, Alex, and Shafer, Glenn (2005). *Algorithmic Learning in a Random World*. Springer, New York.

[37] Wilcoxon, Frank (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1** 80–83.

[38] Willard, Stephen (1970). *General topology*. Addison-Wesley, Reading, MA.