Proceedings of Machine Learning Research 105:1–16, 2019 Conformal and Probabilistic Prediction and Applications

Test statistics and p-values

Yuri Gurevich GUREVICH@UMICH.EDU
University of Michigan, Ann Arbor, MI, USA

Vladimir Vovk V.VOVK@RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Evgueni Smirnov

Abstract

We point out that the traditional notion of test statistic is too narrow, even for the purpose of conformal prediction. The most natural generalization of the traditional notion happens to be too wide. We propose another natural generalization which is arguably the widest reasonable generalization. The study is restricted to simple statistical hypotheses.

Keywords: Exact p-value, randomized p-value, test statistic, lexicographic order.

1. Introduction

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

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

where $[f \le f(x)] = \{y : f(y) \le f(x)\}$. 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 that the initial segments $(-\infty, f(x)]$ be measurable. Should we generalize the notion of test statistics? Would any linearly ordered measurable space work as the codomain of a test statistic f provided that the initial segments $(-\infty, f(x)]$ are measurable?

Our answer to the first question is "yes". Generalized test statistics have been, albeit implicitly, used in conformal prediction and applied statistics. For example, to incorporate the notion of randomized p-values into definition (1), one needs generalized test statistics with codomains that are richer than the real line, as we discuss later in this article (Section 5). Our analysis turns up a more radical generalization, which is more natural and arguably the widest generalization that makes sense.

The answer to the second question is an emphatic "no". There exist a probability space $(\Omega, \Sigma, \mathbf{P})$ and a statistic f with values in a linearly ordered set with measurable initial segments such that every $\hat{f}(x) = 0$. Intuitively, this makes no sense: we are entitled to reject the null hypothesis whatever happens. Formally, this contradicts the standard property

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

of the validity of p-values.

Example 1 The sample space Ω is the collection (known to set theorists as ω_1) of countable (that is finite or 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 ω_0 is the set $\{0, 1, \dots\}$ of natural numbers, $\omega_0 + 1 = \omega_0 \cup \{\omega_0\}$, $\omega_0 + 2 = \omega_0 \cup \{\omega_0, \omega_0 + 1\}$, $\dots$, $\omega_0 + \omega_0 = \omega_0 \cup \{\omega_0 + n : n \in \omega_0\}$, 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 statistic f is the identity function: $f(x) = x$. The order on the codomain is natural:

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

so that

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

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

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

As we already mentioned, a modest but useful generalization of the traditional p-value is already used in conformal prediction and applied statistics (Section 5). We analyze what can go wrong with generalized test statistics and arrive (in Subsection 2.3) at a generalization that is arguably the right one. And we argue that generalized test statistics should be used more widely.

2. Test statistics

2.1. Nominal test statistics

Definition 1 Let $\mathcal{T}$ be a probability space $(\Omega, \Sigma, \mathbf{P})$ and R any ordered measurable space with all initial segments $(-\infty, r]$ measurable. Any measurable function $f : \Omega \to R$ is a *nominal test statistic* for $\mathcal{T}$.

Notation $\mathcal{T}$ alludes to "probability trial". An ordered measurable space is a measurable space endowed with a linear order; in this article, "ordered" always means "linearly ordered".

The notion of nominal test statistic is auxiliary. As Example 1 shows, a nominal test statistic may be unreasonable.

To comply with the traditional definition of p-values, Equation (1) above, we restrict attention to ordered measurable spaces where every initial segment of the form $(-\infty, r]$ is measurable. Note that if Σ_1, Σ_2 are σ-algebras on any set R and $\Sigma_1 \subseteq \Sigma_2$ then every R-valued function that is measurable with respect to Σ_2 is measurable with respect to Σ_1. In other words, the smaller the σ-algebra, the greater the collection of measurable functions. This motivates the following definition.

Definition 2 The *p-minimal* σ-algebra on an ordered set R is the least σ-algebra on R that contains every initial segment of the form $(-\infty, r]$.

If R is an ordered measurable space whose σ-algebra is p-minimal then the measurability requirement for R-valued nominal test statistics simplifies to this: Every set $[f \le r]$ is measurable.

Definition 3 Let f be a nominal test statistic for a probability trial $(\Omega, \Sigma, \mathbf{P})$ with codomain R. The nominal test statistic f *induces* the probability measure $\mathbf{P}_f(S) = \mathbf{P}(f^{-1}(S))$ on the measurable subsets of R.

2.2. Traditional and nearly traditional test statistics

Definition 4 A nominal test statistic is a *traditional test statistic* if its codomain is the real line $\mathbb{R}$ with the Borel σ-algebra.

The real line is of course the set of real numbers with the standard order. Its Borel σ-algebra coincides with its p-minimal σ-algebra. In the applications of conformal prediction and statistics, mostly traditional test statistics are used.

We introduce a slight generalization of traditional test statistics that may be convenient. (That is not the radical generalization mentioned in the Introduction.) To this end, we recall a few definitions and facts.

Every ordered set (of size at least 1) is equipped with its *order topology*. The segments (x, y), $(-\infty, x)$, (y, ∞), and $(-\infty, \infty)$ form a base of the order topology. In this paper, the order topology is the default topology on ordered sets.

A topology is *second-countable* if it has a countable base of open sets. A *jump* in an ordered set $(R, \le)$ is a pair (x, y) of points of $(R, \le)$ such that $x < y$ and there is no $z \in R$ with $x < z < y$. A mapping f whose domain and codomain are both ordered sets is *order-preserving* (or an *embedding*) if $x < y$ implies $f(x) < f(y)$ for all x and y in its domain. The following proposition is proved in, e.g., Cater (1999, Theorem II and Lemma 3) and (in a somewhat less explicit form) in Birkhoff (1967, Theorem 24 in Section VIII.11).

Proposition 5 *A linear order can be embedded into the real line $\mathbb{R}$ if and only if its order topology is second countable. The order topology of a linear order is second countable if and only if the topology is separable and the order has at most countably many jumps.*

Notice that any embedding of a linear order with second-countable topology into the real line will be measurable: the pre-image of an initial interval $(-\infty, r]$ of the real line can be represented as a countable union of initial intervals of this form and is, therefore, measurable. Now we are ready to introduce the slight generalization of traditional test statistics.

Definition 6 A *nearly traditional test statistic* is a nominal test statistic whose codomain is second-countable.

Remark 7 (codomains vs ranges) Consider a nominal test statistic f with a codomain R, and let R' be a nonempty subset of R endowed with the order inherited from R. If R' includes the range of f, we may view f as a nominal test statistic with codomain R'. In particular, we may always view f as a nominal test statistic whose codomain coincides with its range. This is especially relevant in contexts where the range inherits the assumed properties of the codomain, which is the case in most of our definitions and statements. (Definition 4 is a major exception. As far as Definition 8 is concerned, the property "long" may not be inherited but the more relevant property "short" is inherited.)

Traditionalists may argue that the generalization to the nearly traditional test statistics is vacuous, and in a sense it is. By Proposition 5, any nearly traditional test statistic can be composed with an order embedding to obtain a traditional test statistic. This implies that any nearly traditional test statistic f can be replaced by a traditional test statistic f'; f and f' will be equivalent in the sense of inducing the same order on Ω: $f(x) \leq f(y)$ if and only if $f'(x) \leq f'(y)$. However, as the following schematic example shows, the nearly traditional test statistic f may be more convenient to work with. This 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 5).

Example 2 Let f be a traditional test statistic on a discrete measurable space (Ω, Σ) with Ω countable (in which case randomizing p-values becomes particularly useful). Let R be the range of f and equip R with its natural order and the p-minimal (i.e., discrete in this case) σ-algebra. Set Ω' to the real segment $[0, 1]$ and Σ' to the p-minimal (i.e., 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 nearly traditional test statistic on the product measurable space $(\Omega \times \Omega', \Sigma \otimes \Sigma')$. While F is rather natural, an equivalent traditional test statistic may be rather involved; think, e.g., of the case where R is the set $\mathbb{Q}$ of rationals. ◁

Radicals may argue that the generalization to the nearly traditional test statistics is too timid, that there are natural examples of nominal test statistics with more general ordered codomains. We agree.

2.3. A general notion of test statistic

Definition 8 Let O and R be ordered sets. R is *O-long* if there is an embedding of (i.e., an order-preserving map from) O into R; otherwise R is *O-short*.

In Example 1 we mentioned that ω_1 is the set of all countable ordinals and that ordinals are naturally ordered by inclusion. Think of any ordinal α as the linear order of the smaller ordinals, e.g., ω_1 as the linear order of countable ordinals.

We are particularly interested in ω_1-short ordered sets. There is a useful positive characterization of such ordered sets. Recall that sets X, Y of points of an order $\leq$ are *cofinal* if for every $x \in X$ there is $y \geq x$ in Y and if for every $y \in Y$ there is $x \geq y$ in X (and *coinitial* is defined symmetrically).

Proposition 9 *Let R be an ordered set. The following claims are equivalent.*

1. *R is ω_1-short.*
2. *Every nonempty subset X of R includes a sequence $x_1 \leq x_2 \leq \cdots$ cofinal with X.*
3. *Any probability measure $\mathbf{P}$ on R measuring all initial segments $(-\infty, x]$ is continuous from below in the following sense. For every subset X of R, the initial segment $\bigcup_{x \in X}(-\infty, x]$ is measured by $\mathbf{P}$ and*

$$\mathbf{P}\left(\bigcup_{x \in X}(-\infty, x]\right) = \sup_{x \in X} \mathbf{P}(-\infty, x]. \tag{2}$$

Here and below, presenting a sequence in the form

$$x_1 < x_2 < \cdots, \text{ or } x_1 \leq x_2 \leq \cdots, \text{ or}$$
$$x_1 > x_2 > \cdots, \text{ or } x_1 \geq x_2 \geq \cdots,$$

we presume that the indices range over the positive integers. Also, we use the convention that the supremum of the empty set of probabilities is zero and the infimum of the empty set of probabilities is one.

Proof

1 $\implies$ 2 We prove the implication 1 $\implies$ 2 by contrapositive. Assume that X is a nonempty subset of R such that no sequence $x_1 \leq x_2 \leq \cdots$ in X is cofinal with X. We construct an embedding of ω_1 into R. Choose $\eta(0)$ arbitrarily in X. Suppose that β is a countable ordinal and a (possibly transfinite) sequence $\langle \eta(\alpha) : \alpha < \beta \rangle$ has been constructed. The sequence contains only countably many elements and thus cannot be cofinal with X. Choose $\eta(\beta)$ in X greater than all $\eta(\alpha)$ with $\alpha < \beta$. This way we construct the desired embedding $\langle \eta(\alpha) : \alpha < \omega_1 \rangle$.

2 $\implies$ 3 Assume 2 and fix an arbitrary subset X of R. If $X = \emptyset$ then both sides of Equation (2) are zero. Suppose that $X \neq \emptyset$. By 2, there is a sequence $x_1 \leq x_2 \leq \cdots$ of points in X cofinal with X. Accordingly, it suffices to prove that $\mathbf{P}(\cup_n(-\infty, x_n]) = \sup_n \mathbf{P}(-\infty, x_n]$, which follows from the countable additivity of probability measures.

3 $\implies$ 1 Again we prove the desired implication by contrapositive. Assume that there is an embedding η of ω_1 into R and let $X = \operatorname{Range} \eta$. We construct a probability measure $\mathbf{P}$ for which Equation (2) fails. Without loss of generality, X is cofinal in R. Let Σ be the p-minimal σ-algebra of R. By Carathéodory's extension theorem, to define $\mathbf{P}$ on Σ, it suffices to define $\mathbf{P}$ on the initial segments $(-\infty, x]$ (in such a way that the conditions of the theorem are satisfied, which we will check carefully

later). Define $\mathbf{P}(-\infty, x] = 0$ if $X \cap (-\infty, x]$ is countable and $\mathbf{P}(-\infty, x] = 1$ otherwise. Then the left side of Equation (2) is $\mathbf{P}(R) = 1$ while the right side is 0. It remains to check the applicability of Carathéodory's theorem. To make sure $\mathbf{P}$ is defined on a semi-ring, set $\mathbf{P}(\emptyset) = 0$ and $\mathbf{P}(x, y] = \mathbf{P}(-\infty, y] - \mathbf{P}(-\infty, x]$ for all $x < y$. Let us first check that $\mathbf{P}(x, y] = 0$ if $X \cap (x, y]$ is countable and $\mathbf{P}(x, y] = 1$ otherwise. The case where $X \cap (x, y]$ is countable is trivial, so let us assume that $X \cap (x, y]$ is uncountable. In this case we have $\mathbf{P}(-\infty, y] = 1$ and, since $(x, y]$ contains $\eta(\alpha)$ for a countable ordinal α, $\mathbf{P}(-\infty, x] = 0$; by definition, this implies $\mathbf{P}(x, y] = 1$. It remains to check that $\mathbf{P}$ is σ-additive: if $(x, y] = \cup_{n=1}^{\infty}(x_n, y_n]$, where the union is disjoint, then $\mathbf{P}(x, y] = \sum_{n=1}^{\infty} \mathbf{P}(x_n, y_n]$. This follows immediately from $X \cap (x_n, y_n]$ being uncountable for at most one n.

■

Remark 10 The argument in the proof of $1 \implies 2$ is an instance of definition by transfinite induction (in our case, over the countable ordinals). For details, see the transfinite recursion theorem in Halmos (1960, Section 18). It becomes applicable if we fix a choice function that maps every transfinite sequence $\langle \eta(\alpha) : \alpha < \beta \rangle$ in X for every countable ordinal β to $\eta(\beta) \in X$ satisfying our desideratum (in this particular case, $\eta(\beta)$ being greater than all $\eta(\alpha)$ with $\alpha < \beta$).

The reverse of the ordered set ω_1 is known as ω_1^*. In other words, ω_1^* is the set of all countable ordinals with the reverse order $\alpha < \beta \iff \beta \in \alpha$. By symmetry, Proposition 9 has the following corollary.

Corollary 11 *Let R be an ordered set. The following two claims are equivalent.*

1. *R is ω_1^*-short.*
2. *Every nonempty subset X of R includes a sequence $x_1 \geq x_2 \geq \cdots$ coinitial with X.*

Proposition 12 *Suppose R is ω_1-short. Any probability measure $\mathbf{P}$ on R measuring all initial segments $(-\infty, x]$ is continuous from above in the following sense. For any subset X of R,*

$$\mathbf{P}\left(\bigcup_{x \in X} [x, \infty)\right) = \sup_{x \in X} \mathbf{P}[x, \infty) \tag{3}$$

(which includes the existence of all these probabilities).

Proof By Proposition 9, it suffices to check that the measurability of all initial segments $(-\infty, x]$ implies the measurability of all final segments $[x, \infty)$, i.e., the measurability of every initial segment $(-\infty, x)$. Suppose that every $(-\infty, x]$ is measurable. By Proposition 9, for any x there exists a sequence $x_1 \leq x_2 \leq \cdots$ cofinal with $(-\infty, x)$, so that $(-\infty, x) = \bigcup_n (-\infty, x_n]$. Hence $(-\infty, x)$ is measurable. ■

Since

$$1 - \mathbf{P}\left(\bigcup_{x \in X} [x, \infty)\right) = \mathbf{P}\left(\bigcap_{x \in X} (-\infty, x)\right)$$

and

$$1 - \sup_{x \in X} \mathbf{P}[x, \infty) = \inf_{x \in X} \mathbf{P}(-\infty, x),$$

(3) is equivalent to

$$\mathbf{P}\left(\bigcap_{x \in X} (-\infty, x)\right) = \inf_{x \in X} \mathbf{P}(-\infty, x).$$

The definition of p-value is not symmetric with respect to order reversal. Accordingly, in our context, the symmetry between ω_1 and ω_1^* is limited. ω_1^* is less dangerous than ω_1. Replacing the order ω_1 in Example 1 by the reverse order ω_1^* leads to a p-value that is identically equal to 1, which is a valid p-value (and occurs for very simple test statistics, e.g., whenever the size of f's codomain is 1) albeit not useful. ω_1^*-short orders have some desirable properties. Corollary 11 and Proposition 12 indicate some of them. Others will will be pointed out later. The following definition is borrowed from the literature on linear orders; see Gurevich and Shelah (1979) and Rosenstein (1982, p. 88).

Definition 13 A linear order is *short* if it is ω_1-short and ω_1^*-short.

Now we are ready to introduce our general notion of test statistic.

Definition 14 A *test statistic* is a nominal test statistic whose codomain is short.

Proposition 15 *Any ordered set with second-countable order topology is short. Thus every nearly traditional test statistic is a test statistic.*

In particular, the real line $\mathbb{R}$ is short.

Proof Let R be an ordered set with second-countable order topology. By Proposition 5, R is separable and has at most countably many jumps. Let C be a countable set that is dense in R and contains all points involved in jumps. Suppose toward a contradiction that η is an order-preserving or order-reversing map from ω_1 to R. We obtain uncountably many disjoint open nonempty intervals $(\eta(\alpha), \eta(\alpha + 2))$ where α is a limit ordinal. Each of these intervals contains a point from C, which is impossible. ■

3. Induced test statistics, p-functions, and p-values

3.1. Induced test statistics

Definition 16 Any test statistic f for a probability trial $(\Omega, \Sigma, \mathbf{P})$ *induces* a traditional test statistic $\hat{f}(x) = \mathbf{P}[f \leq f(x)]$ on $(\Omega, \Sigma, \mathbf{P})$.

Lemma 17 *If f is a test statistic then $\hat{\hat{f}} = \hat{f}$. In other words, any induced test statistic is self-induced.*

Proof Let f be a test statistic for a probability trial $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$ and R be the codomain of f. By Definition 3, f induces a probability distribution $\mathbf{P}_f(X) = \mathbf{P}(f^{-1}(X))$ on R. To simplify notation, we omit the subscript f.

Let $\eta(r) = \mathbf{P}[f \le r]$, so that $\eta(f(x)) = \mathbf{P}[f \le f(x)] = \hat{f}(x)$. If $s \le r$ then $\eta(s) \le \eta(r)$. On the other hand,

$$\begin{aligned} \eta(s) \le \eta(r) &\iff \text{either } s \le r \text{ or else } (s > r \text{ and } \eta(s) = \eta(r)) \\ &\iff \text{either } s \le r \text{ or else } (s > r \text{ and } \mathbf{P}(r, s] = 0). \end{aligned}$$

For any $r \in R$, the set $S_r = \{s \in R : s > r \text{ and } \mathbf{P}(r, s] = 0\}$ is measurable in R and $\mathbf{P}(S_r) = 0$. Indeed, by the definition of nominal test statistics, the initial segments $(-\infty, t]$ of R are measurable. Since f is a genuine test statistic (rather than just nominal), R is ω_1-short and so there is a sequence $s_1 \le s_2 \le \cdots$ in S_r cofinal with S_r so that $S_r = \bigcup_n (r, s_n]$ and thus is measurable. Further, $\mathbf{P}(\cup_n (r, s_n]) = \lim_n \mathbf{P}(r, s_n] = 0$. Thus

$$0 = \mathbf{P}(S_r) = \mathbf{P}(f^{-1}(S_r)) = \mathbf{P}\{y : f(y) > r \text{ and } \eta(f(y)) = \eta(r)\}.$$

Now we are ready to prove $\hat{\hat{f}}(x) = \hat{f}(x)$.

$$\begin{aligned} \hat{\hat{f}}(x) &= \mathbf{P}[\hat{f} \le \hat{f}(x)] = \mathbf{P}\{y : \eta(f(y)) \le \eta(f(x))\} \\ &= \mathbf{P}\{y : f(y) \le f(x)\} + \mathbf{P}\{y : f(y) > f(x) \text{ and } \eta(f(y)) = \eta(f(x))\} \\ &= \mathbf{P}\{y : f(y) \le f(x)\} = \mathbf{P}[f \le f(x)] = \hat{f}(x). \end{aligned}$$

■

Theorem 18 *Let f be a traditional test statistic with values in the real segment $[0, 1]$. The following claims are equivalent.*

1. *f is induced by some test statistic.*
2. *f is self-inducing, i.e., $\hat{f} = f$.*
3. *$\mathbf{P}[f \le \varepsilon] = \varepsilon$ for every $\varepsilon \in \operatorname{Range} f$.*

Proof Obviously 2 implies 1. By Lemma 17, 1 implies 2. It suffices to prove that 2 and 3 are equivalent.

2 $\implies$ 3 Assume 2 and let $\varepsilon = f(x)$ for some x. We have

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

3 $\implies$ 2 Assume 3. Given a sample point x, let $\varepsilon = f(x)$. We have

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

■

The definition of the induced test statistic $\hat{f}$ can be extended to the case where f is only a nominal test statistic. But the following proposition emphasizes the important role of the property of being ω_1-short.

Proposition 19 *Let R be an ω_1-long ordered set. There exists an R-valued nominal test statistic f such that $\mathbf{P}[\hat{f} \le 0] = 1$ even though $0 \in \operatorname{Range} \hat{f}$.*

Proof Fix an embedding η of ω_1 into R, and let

$$L = \{x \in R : x \le \eta(\alpha) \text{ for some } \alpha \in \omega_1\}.$$

Let Σ be the least σ-algebra on R that contains the initial segments $(-\infty, x]$ and also contains L.

For every member X of Σ, either X or $R - X$ contains at most a countable subset of $\operatorname{Range} \eta$. Indeed, every initial segment of the form $(-\infty, x]$ and L have this property, and the property is preserved by complementation and countable unions.

Define a probability measure $\mathbf{P}$ on Σ as follows: If X contains at most a countable subset of $\operatorname{Range} \eta$ then $\mathbf{P}(X) = 0$; otherwise $\mathbf{P}(X) = 1$.

The desired nominal test statistic f is the identity function on R. It is easy to see that $x \in L$ if and only if $\hat{f}(x) = \mathbf{P}[f \le f(x)] = 0$. So $\mathbf{P}[\hat{f} \le 0] = \mathbf{P}(L) = 1$. ■

3.2. p-functions and p-values

We want to define p-functions and p-values in such a way that p-values are the values of p-functions. It is tempting to define a p-function as a traditional test statistic $\hat{f}$ induced by some test statistic f. By Theorem 18, we have $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for any $\varepsilon \in \operatorname{Range} f$. But in practice people also use conservative p-values. To accommodate this practice, we give a more general definition of p-functions.

Definition 20

- A *p-function* is a traditional test statistic f with values in the real segment $[0, 1]$ such that $\mathbf{P}[f \le \varepsilon] \le \varepsilon$ for every $\varepsilon \in [0, 1]$.
- A p-function f is *range-exact* if $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for every $\varepsilon \in \operatorname{Range} f$; otherwise f is *conservative*.
- A p-function f is *everywhere exact* or simply *exact* if $\mathbf{P}[f \le \varepsilon] = \varepsilon$ for every $\varepsilon \in [0, 1]$.

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 21 *Let f be a traditional test statistic with values in the real segment $[0, 1]$. The following claims are equivalent.*

1. *f is an induced test statistic.*

2. *f is a range exact p-function.*

Proof 2 implies 1 by Theorem 18. To prove the other implication, assume 1. By Theorem 18, $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ for any $\varepsilon \in \operatorname{Range} f$. It remains to prove that $\mathbf{P}[f \leq \varepsilon] \leq \varepsilon$ for every $\varepsilon \in [0,1] - \operatorname{Range} f$.

Let $\varepsilon \in [0,1] - \operatorname{Range} f$, $S = \{s \in \operatorname{Range} f : s < \varepsilon\}$ and $\varepsilon_0 = \sup S$. If $S = \emptyset$ then $[f \leq \varepsilon] = \emptyset$ and $\mathbf{P}[f \leq \varepsilon] = 0 \leq \varepsilon$. Otherwise there is a sequence $s_1 \leq s_2 \leq \cdots$ of reals in S converging to ε_0. Then

$$\mathbf{P}[f \leq \varepsilon] = \mathbf{P}[f \leq \varepsilon_0] = \mathbf{P}\left(\bigcup_{n=1}^{\infty} [f \leq s_n]\right) = \lim_{n\to\infty} \mathbf{P}\left([f \leq s_n]\right) = \lim_{n\to\infty} s_n = \varepsilon_0 \leq \varepsilon.$$

■

Definition 22 Let F be a p-function for some probability trial $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$. For any outcome $x \in \Omega$, the number $F(x)$ is the *p-value* associated with the test statistic F and the outcome x. If F is an induced test statistic and f is any test statistic for $\mathcal{T}$ inducing F then $F(x)$ is also the p-value associated with f and x. If $\mathbf{P}[F \leq F(x)] = F(x)$ then the p-value $F(x)$ is *exact*; otherwise it is *conservative*.

4. Diffuse test statistics and exact p-functions

We are particularly interested in exact p-functions $f(x)$, the p-functions with $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ for all $\varepsilon \in [0,1]$. Classical parametric statistics is a rich source of exact p-functions; randomized p-values (discussed in the next section) is another important example.

Recall that an *atom* in a probability space $\mathcal{T}$ is an event (i.e., a measurable set) of positive probability that cannot be split into a disjoint union of two events of positive probability. $\mathcal{T}$ is *diffuse* if it has no atoms. Every singleton event of positive probability is an atom. In Example 1, the complement of every countable set is an atom.

Lemma 23 *Let R be a short ordered set as well as a probability space where all initial segments $(-\infty, r]$ are measurable. R is diffuse if and only if it has no singleton atoms.*

Proof The "only if" implication is trivial. To prove the "if" implication, suppose toward a contradiction that R has no singleton atoms and yet it does have an atom A. The subset A by itself is a short ordered set. If Σ is the σ-algebra of R, and $\mathbf{P}$ is the probability measure on Σ, consider the σ-algebra $\Sigma_A = \{A \cap X : X \in \Sigma\}$ and restrict $\mathbf{P}$ to Σ_A. In the rest of the proof, we work with A. Without loss of generality, we may assume that A is the whole R. Accordingly R itself is an atom in R.

Obviously $\mathbf{P}(R) = 1$. Since R is an atom, for any $X \in \Sigma$, the probability $\mathbf{P}(X)$ is either 0 or 1. Notice that every singleton set $\{r\}$ is measurable in R. Indeed, $(-\infty, r]$ is measurable, so it suffices to prove that $(-\infty, r)$ is measurable. By Proposition 9, there is a

sequence $x_1 \le x_2 \le \cdots$ converging to r, and so $(-\infty, r)$ is a countable union of measurable sets.

Let I be the set of points $x \in R$ with $\mathbf{P}(-\infty, x] = 0$ (it is an initial segment, in the sense of containing any y such that $y \le x$ for some $x \in I$). By the continuity from below, see Equation (2),

$$\mathbf{P}(I) = \mathbf{P}\left(\bigcup_{x \in I} (-\infty, x]\right) = \sup_{x \in I} \mathbf{P}(-\infty, x] = 0.$$

Let F be the final segment $R - I$. For any $y \in F$, $\mathbf{P}(-\infty, y] = 1$ and therefore $\mathbf{P}(y, \infty) = 0$. Since singleton sets are measurable in R and R has no singleton atoms, every $\mathbf{P}[y, \infty) = \mathbf{P}\{y\} + \mathbf{P}(y, \infty) = 0$. By the continuity from above, see Equation (3),

$$\mathbf{P}(F) = \mathbf{P}\left(\bigcup_{x \in F} [x, \infty)\right) = \sup_{x \in F} \mathbf{P}[x, \infty) = 0.$$

Thus $\mathbf{P}(R) = \mathbf{P}(I) + \mathbf{P}(F) = 0$, which gives us the desired contradiction. ■

Call a nominal test statistic f on a probability trial $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$ *diffuse* if the probability distribution $\mathbf{P}_f$ that f induces on its codomain R is diffuse. If R is short then, by Lemma 23, the test statistic f is diffuse if and only if $\mathbf{P}(f^{-1}(r)) = \mathbf{P}_f(r) = 0$ for every point $r \in R$.

Proposition 24 *Let f be a diffuse test statistic. Then the induced test statistic $\hat{f}$ is diffuse.*

Proof The codomain $[0, 1]$ of $\hat{f}$ is short, and every singleton set in $[0, 1]$ is measurable. By Lemma 23 it suffices to prove that $\mathbf{P}[\hat{f} = \varepsilon] = 0$ for every $\varepsilon \in \operatorname{Range} \hat{f}$. Fix such a number ε. Since $\varepsilon \in \operatorname{Range} \hat{f}$, the set $X = \{x : \hat{f}(x) = \varepsilon\} \neq \emptyset$. Let $R_0 = \{f(x) : x \in X\}$.

We use the notation and results established in the proof of Lemma 17. If $x \in X$ and $r = f(x)$, we have $\varepsilon = \hat{f}(x) = \mathbf{P}[f \le r] = \eta(r)$ and so $[\hat{f} = \varepsilon] = \{y : \hat{f}(y) = \hat{f}(x)\} = \{y : \eta(f(y)) = \eta(r)\}$. Let $U(r) = \{y : f(y) \ge r \text{ and } \eta(f(y)) = \eta(r)\}$ and $U'(r) = \{y : f(y) > r \text{ and } \eta(f(y)) = \eta(r)\}$. In the proof of Lemma 17 we established that $\mathbf{P}(U'(r)) = 0$. But $U(r) = U'(r) \cup f^{-1}(r)$. Since f is diffuse, $\mathbf{P}(f^{-1}(r)) = \mathbf{P}_f(r) = 0$ and so $\mathbf{P}(U(r)) = 0$.

If $r = \min R_0$ then $\mathbf{P}[\hat{f} = \varepsilon] = \mathbf{P}[\hat{f} = \eta(r)] = \mathbf{P}(U(r)) = 0$. Suppose that R_0 does not have a minimal element. Since R is short, there exists a sequence $r_1 > r_2 > \ldots$ in R_0 that is coinitial with R_0. We have

$$\mathbf{P}[\hat{f} = \varepsilon] = \mathbf{P}\left(\bigcup_n U(r_n)\right) = 0.$$

■

Theorem 25 *Let f be a diffuse test statistic. Then the induced test statistic $\hat{f}$ is an exact p-function.*

Proof By Theorem 21, $\hat{f}$ is a range-exact p-function, so that $\mathbf{P}[f \leq \varepsilon] \leq \varepsilon$ for $\varepsilon \in [0,1]$ and $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ for all $\varepsilon \in \operatorname{Range} \hat{f}$. It remains to prove that $\mathbf{P}[f \leq \varepsilon] = \varepsilon$ for every $\varepsilon \in [0,1] - \operatorname{Range} \hat{f}$. Fix such a number ε.

Let $\varepsilon_0 = \sup\{\delta \in \operatorname{Range} \hat{f} : \delta < \varepsilon\}$ and $\varepsilon_1 = \inf\{\delta \in \operatorname{Range} \hat{f} : \delta > \varepsilon\}$. Our convention here is that $\sup \emptyset = 0$ and $\inf \emptyset = 1$. There exists a sequence $\delta_1 \leq \delta_2 \leq \cdots$ in $\operatorname{Range} \hat{f}$ that converges to ε_0, so that

$$\mathbf{P}[\hat{f} \leq \varepsilon_0] = \lim_n \mathbf{P}[\hat{f} \leq \delta_n] + \mathbf{P}[\hat{f} = \varepsilon_0] = \lim_n \delta_n + 0 = \varepsilon_0.$$

Similarly, there is a sequence $\delta_1 \geq \delta_2 \geq \cdots$ in $\operatorname{Range} \hat{f}$ that converges to ε_1, so that

$$\mathbf{P}[\hat{f} \geq \varepsilon_1] = \mathbf{P}[\hat{f} = \varepsilon_1] + \lim_n \mathbf{P}[\hat{f} > \delta_n] = \lim_n (1 - \delta_n) = 1 - \varepsilon_1.$$

We have

$$1 = \mathbf{P}[\hat{f} \leq \varepsilon_0] + \mathbf{P}[\hat{f} \geq \varepsilon_1] = \varepsilon_0 + (1 - \varepsilon_1),$$

so that $\varepsilon_0 = \varepsilon_1 = \varepsilon$ and $\mathbf{P}[\hat{f} \leq \varepsilon] = \mathbf{P}[\hat{f} \leq \varepsilon_0] = \varepsilon_0 = \varepsilon$. ■

Proposition 26 *If R is an ω_1^*-long (and ω_1-short) linearly ordered set endowed with the p-minimal σ-algebra then there is an R-valued diffuse nominal test statistic f such that $\hat{f}$ is not an exact p-function.*

Proof Consider a trial $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$ where $\Omega, \Sigma, \mathbf{P}$ are as follows.

- Ω is ω_1^*, i.e., the set of countable ordinals with the reverse order $\alpha < \beta \iff \beta \in \alpha$.
- Σ is the p-minimal σ-algebra on Ω. It consists of the countable subsets of Ω and their complements.
- $\mathbf{P}(X) = 0$ if X is countable. In particular, every $\mathbf{P}(-\infty, \alpha] = 1$.

Since R is ω_1^*-long, there is an order isomorphism η from ω_1^* into R; η is an R-valued nominal test statistic on $\mathcal{T}$. Indeed, since the σ-algebra of R is p-minimal, it suffices to show that every $\eta^{-1}(-\infty, r] \in \Sigma$. Since ω_1 is well-ordered, every $X \subseteq \omega_1$ has a minimal point in ω_1; accordingly every $X \subseteq \omega_1^*$ has a maximal point in ω_1^*. In particular, let $y = \max\{\alpha \in \omega_1^* : \eta(\alpha) \leq r\}$. Accordingly, $\eta^{-1}(-\infty, r] = \eta^{-1}(-\infty, \eta(y)] = (-\infty, y] \in \Sigma$.

Since $\mathbf{P}$ takes only two values, the induced p-function $\hat{\eta}$ takes only two values and thus is not exact. ■

5. Randomized p-values

In this section we discuss randomized p-values as a natural application of non-traditional test statistics. Randomized p-values arise naturally in situations where the distribution of the test statistic is not continuous. They are produced by test statistics whose codomain is $\mathbb{R} \times [0,1]$ with the lexicographic order; intuitively, we add a random number to a traditional

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 Theorem 29 below).

At this time there are at least two (and probably many more) fields of applied statistics and machine learning where randomized p-values are essential: multiple hypothesis testing in bioinformatics (see, e.g., Dickhaus et al. 2012) and on-line testing the hypothesis of exchangeability using conformal martingales (see, e.g., Vovk et al. 2005, Section 7.1, and Zhang et al. 2019). 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. In the theory of conformal prediction, allowing randomized p-values greatly simplifies and strengthens the main property of validity (cf. Vovk et al. 2005, Theorem 8.1, which uses randomized p-values, with its predecessor, Vovk 2002, Theorem 1, which avoids randomization).

We are essentially in the situation of Example 2, except that the assumption that the codomain of f is countable is dropped.

By Theorem 18, the induced p-function f of a test statistic is only guaranteed to satisfy the inequality $\mathbf{P}[f \le \varepsilon] \le \varepsilon$, and very simple examples show that $\mathbf{P}[f \le \varepsilon] < \varepsilon$ is indeed possible: e.g., take any $\epsilon \in (0,1)$ when the sample space is a singleton. Randomized p-values are a way of making f exact, i.e., achieving the equality $\mathbf{P}[f \le \varepsilon] = \varepsilon$ 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.

We need a couple of auxiliary results.

Lemma 27 *If R and S are short orders, then the product $R \times S$, ordered lexicographically, is short.*

Proof By symmetry it suffices to prove that $R \times S$ is ω_1-short, i.e., that for every well-ordered set A, if there exists an order preserving map $\eta : A \to R \times S$, then A is countable. Each $\eta(a)$ has the form (r_a, s_a), and $(r_a, s_a) < (r_b, s_b)$ if and only if either $r_a < r_b$ or else both $r_a = r_b$ and $s_a < s_b$.

Since S is short, the set $A_r = \{a : r_a = r\}$ is countable for every $r \in R$. Since A is well-ordered, the subset $\{\min A_r : A_r \neq \emptyset\}$ of A is well ordered. Since R is short, the well-ordered subset $\{\eta(\min A_r) : A_r \neq \emptyset\}$ of R is countable, so that there are only countable many nonempty sets A_r. Therefore A is a countable union of countable sets, so that A is countable. ■

Lemma 28 *Let $(\Omega_1, \Sigma_1, \mathbf{P}_1)$ and $(\Omega_2, \Sigma_2, \mathbf{P}_2)$ be probability spaces where every singleton set is measurable, and form the product probability space*

$$(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2, \mathbf{P}_1 \times \mathbf{P}_2).$$

Every singleton set is measurable in the product space. Furthermore, the product space has the property that the probability of every singleton event is zero if at least one of the factors has this property.

Proof By the definition of the product, $\mathbf{P}(X_1 \times X_2) = \mathbf{P}_1(X_1) \cdot \mathbf{P}_2(X_2)$ for any $X_1 \in \Sigma_1$ and $X_2 \in \Sigma_2$. A singleton set in Ω has the form $\{x_1\} \times \{x_2\}$ and this is measurable. The second claim follows from the fact that $\mathbf{P}(\{x_1\} \times \{x_2\}) = \mathbf{P}_1\{x_1\} \cdot \mathbf{P}_2\{x_2\}$. ■

Now we are ready to address the issue of randomized p-values. We start from our usual setting of a given traditional test statistic $f : \Omega \to \mathbb{R}$ on a trial $\mathcal{T} = (\Omega, \Sigma, \mathbf{P})$. 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

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

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

$$F(x, r) = (f(x), r) \tag{4}$$

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 traditional test statistics*. They have the following useful property.

Theorem 29 *The induced p-function $\hat{F}$ of any randomized traditional test statistic F is exact, so that $\bar{\mathbf{P}}[\hat{F} \leq \varepsilon] = \varepsilon$ for any $\epsilon \in [0,1]$.*

Proof By Lemma 27, the codomain $\mathbb{R} \times [0,1]$ of F is short. It is easy to see that all initial segments $(-\infty, x]$ of $\mathbb{R} \times [0,1]$ are measurable. By Lemma 23, the $\mathbf{P}_F$-atoms, if any, of $\mathbb{R} \times [0,1]$ are singletons. By Lemma 28 and $\mathbf{P}_F = \mathbf{P}_f \times \mathcal{U}$ (cf. (4)), $\mathbf{P}_F$ does not have singleton atoms and thus is diffuse. It remains to apply Theorem 25. ■

It is easy to see that the theorem generalizes to the case where the component f of test statistic F is any test statistic.

The value $\hat{F}(x, r)$ is the randomized p-value corresponding to an outcome x and random number r. But it is easy to see that the topology of the lexicographically ordered $\mathbb{R} \times [0,1]$ is not second-countable (not even separable). According to Definition 6 and Proposition 5, the function F defined by (4) is not a traditional or even nearly-traditional test statistic. But $\mathbb{R}$ and $[0,1]$ are short, and so F is a test statistic according to Definition 14, as Lemma 27 shows.

Remark 30 When using randomized p-values, statisticians (and computer scientists in related areas) do not usually emphasize the use of non-traditional test statistics, which remain implicit. They prefer to define randomized p-values from scratch rather than using the traditional definition (1). Namely, the usual definition (as given in, e.g., Dickhaus et al. 2012 and Vovk et al. 2005) is

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

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 Coudin (2007, p. 91, (3.4)).

6. A summary

This paper's aim has been to investigate advantages and drawbacks of various classes of nominal test statistics. If forced to choose one class, our recommendation would be to use the class of test statistics (i.e., nominal test statistics with short codomains). In view of Theorem 18 and Proposition 19, this will lead to valid p-values (but possibly conservative p-functions). By Theorem 25, the corresponding p-functions will be exact in the case of diffuse test statistics. Finally, Theorem 29 is applicable to any test statistic (as we say after its proof) and allows us to define exact p-functions by using the device of randomization.

Acknowledgments

We thank Andreas Blass, Steffen Lauritzen, and Glenn Shafer for useful comments on drafts of this article. Comments by the COPA 2019 reviewers are also gratefully appreciated.

References

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

Frank S. Cater. On order topologies and the real line. *Real Analysis Exchange*, 25:771–780, 1999.

Elise Coudin. *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, 2007.

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

Yuri Gurevich and Saharon Shelah. Modest theory of short chains. II. *Journal of Symbolic Logic*, 44:491–502, 1979.

Paul R. Halmos. *Naive Set Theory*. Van Nostrand, New York, 1960.

Joseph G. Rosenstein. *Linear Orderings*. Academic Press, New York, 1982.

Vladimir Vovk. On-line confidence machines are well-calibrated. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science*, pages 187–196, Los Alamitos, CA, 2002. IEEE Computer Society.

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

Xiao Zhang, Peter Lu, Josée Martens, Gary Ericson, and Kent Sharkey. *Time Series Anomaly Detection module in Microsoft Azure*. Microsoft, Seattle, WA, May 2019. Online documentation: https://docs.microsoft.com/en-gb/azure/machine-learning/studio-module-reference/time-series-anomaly-detection.