

# Test statistics and 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 August 26, 2019.

Project web site:  
<http://alrw.net>

## 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.

A version of this paper (which is updated less often) is published as an arXiv technical report: <https://arxiv.org/abs/1702.02590>. The conference version is published in the Proceedings of COPA 2019 (Proceedings of Machine Learning Research 105:89–104).

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Test statistics</b>	<b>2</b>
<b>3</b>	<b>Induced test statistics, p-functions, and p-values</b>	<b>8</b>
<b>4</b>	<b>Diffuse test statistics and exact p-functions</b>	<b>10</b>
<b>5</b>	<b>Randomized p-values</b>	<b>13</b>
<b>6</b>	<b>An alternative to randomizing p-values</b>	<b>15</b>
<b>7</b>	<b>A summary</b>	<b>18</b>
	<b>References</b>	<b>18</b>
<b>A</b>	<b>History of p-values</b>	<b>21</b>

... 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 Introduction

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

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

where  $[f \leq f(x)] = \{y : f(y) \leq f(x)\}$ . 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

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

of the validity of p-values.

**Example 1.1.** The sample space  $\Omega$  is the collection (known to set theorists as  $\omega_1$ ) 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  $\omega_0$  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  $\omega_1$  is the set of countable ordinals.

The  $\sigma$ -algebra  $\Sigma$  consists of all countable subsets of  $\Omega$  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

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

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 and not necessarily in combination with randomization.

## 2 Test statistics

### 2.1 Nominal test statistics

**Definition 2.1.** 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  $f : \Omega \rightarrow R$  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 1.1 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  $\Sigma_1, \Sigma_2$  are  $\sigma$ -algebras on any set  $R$  and  $\Sigma_1 \subseteq \Sigma_2$  then every  $R$ -valued function that is measurable with respect to  $\Sigma_2$  is measurable with respect to  $\Sigma_1$ . In other words, the smaller the  $\sigma$ -algebra, the greater the collection of measurable functions. This motivates the following definition.

**Definition 2.2.** The *p-minimal*  $\sigma$ -algebra on an ordered set  $R$  is the least  $\sigma$ -algebra on  $R$  that contains every initial segment of the form  $(-\infty, r]$ .

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

**Definition 2.3.** 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 2.4.** A nominal test statistic is a *traditional test statistic* if its codomain is the real line  $\mathbb{R}$  with the Borel  $\sigma$ -algebra.

The real line is of course the set of real numbers with the standard order. Its Borel  $\sigma$ -algebra coincides with its *p-minimal*  $\sigma$ -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  $(x, y)$ ,  $(-\infty, x)$ ,  $(y, \infty)$ , 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  $(R, \leq)$  is a pair  $(x, y)$  of points of  $(R, \leq)$  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., [5, Theorem II and Lemma 3] and (in a somewhat less explicit form) in [4, Theorem 24 in Section VIII.11].

**Proposition 2.5.** *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 2.6.** A *nearly traditional test statistic* is a nominal test statistic whose codomain is second-countable.

*Remark 2.7* (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 2.4 is a major exception. As far as Definition 2.9 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 2.5, 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  $\Omega$ :  $f(x) \leq f(y)$  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 2.8.** Let  $f$  be a traditional test statistic on a discrete measurable space  $(\Omega, \Sigma)$  with  $\Omega$  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)  $\sigma$ -algebra. Set  $\Omega'$  to the real segment  $[0, 1]$  and  $\Sigma'$  to the p-minimal (i.e., Borel)  $\sigma$ -algebra on  $[0, 1]$ . (In the context of randomized p-values,  $\Omega'$  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.  $\triangleleft$

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 2.9.** 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 1.1 we mentioned that  $\omega_1$  is the set of all countable ordinals and that ordinals are naturally ordered by inclusion. Think of any ordinal  $\alpha$  as the linear order of the smaller ordinals, e.g.,  $\omega_1$  as the linear order of countable ordinals.

We are particularly interested in  $\omega_1$ -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 2.10.** *Let  $R$  be an ordered set. The following claims are equivalent.*

1.  $R$  is  $\omega_1$ -short.
2. Every nonempty subset  $X$  of  $R$  includes a sequence  $x_1 \leq x_2 \leq \dots$  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

$$\mathbf{P} \left( \bigcup_{x \in X} (-\infty, x] \right) = \sup_{x \in X} \mathbf{P}(-\infty, x]. \quad (3)$$

Here and below, presenting a sequence in the form

$$\begin{aligned} &x_1 < x_2 < \dots, \text{ or } x_1 \leq x_2 \leq \dots, \text{ or} \\ &x_1 > x_2 > \dots, \text{ or } x_1 \geq x_2 \geq \dots, \end{aligned}$$

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  $x_1 \leq x_2 \leq \dots$  in  $X$  is cofinal with  $X$ . We construct an embedding of  $\omega_1$  into  $R$ . Choose  $\eta(0)$  arbitrarily in  $X$ . Suppose that  $\beta$  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 (3) are zero. Suppose that  $X \neq \emptyset$ . By 2, there is a sequence  $x_1 \leq x_2 \leq \dots$  of points in  $X$  cofinal with  $X$ . Accordingly, it suffices to prove that  $\mathbf{P}(\bigcup_n (-\infty, x_n]) = \sup_n \mathbf{P}(-\infty, x_n]$ , 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  $\eta$  of  $\omega_1$  into  $R$  and let  $X = \text{Range } \eta$ . We construct a probability measure  $\mathbf{P}$  for which Equation (3) fails. Without loss of generality,  $X$  is cofinal in  $R$ . Let  $\Sigma$  be the p-minimal  $\sigma$ -algebra of  $R$ . By Carathéodory's extension theorem, to define  $\mathbf{P}$  on  $\Sigma$ , 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 (3) 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  $\alpha$ ,  $\mathbf{P}(-\infty, x] = 0$ ; by definition, this implies  $\mathbf{P}(x, y] = 1$ . It remains to check that  $\mathbf{P}$  is  $\sigma$ -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  $n$ .  $\square$

*Remark 2.11.* 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 [18, 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  $\beta$  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  $\omega_1$  is known as  $\omega_1^*$ . In other words,  $\omega_1^*$  is the set of all countable ordinals with the reverse order  $\alpha < \beta \iff \beta \in \alpha$ . By symmetry, Proposition 2.10 has the following corollary.

**Corollary 2.12.** *Let  $R$  be an ordered set. The following two claims are equivalent.*

1.  $R$  is  $\omega_1^*$ -short.
2. Every nonempty subset  $X$  of  $R$  includes a sequence  $x_1 \geq x_2 \geq \dots$  coinitial with  $X$ .

**Proposition 2.13.** *Suppose  $R$  is  $\omega_1$ -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  $R$ ,*

$$\mathbf{P}\left(\bigcup_{x \in X} [x, \infty)\right) = \sup_{x \in X} \mathbf{P}[x, \infty) \quad (4)$$

(which includes the existence of all these probabilities).



*Proof.* By Proposition 2.10, 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 2.10, for any  $x$  there exists a sequence  $x_1 \leq x_2 \leq \dots$  cofinal with  $(-\infty, x)$ , so that  $(-\infty, x) = \bigcup_n (-\infty, x_n]$ . Hence  $(-\infty, x)$  is measurable.  $\square$

Since

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

and

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

(4) 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  $\omega_1$  and  $\omega_1^*$  is limited.  $\omega_1^*$  is less dangerous than  $\omega_1$ . Replacing the order  $\omega_1$  in Example 1.1 by the reverse order  $\omega_1^*$  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.  $\omega_1^*$ -short orders have some desirable properties. Corollary 2.12 and Proposition 2.13 indicate some of them. Others will be pointed out later. The following definition is borrowed from the literature on linear orders; see [16] and [25, p. 88].

**Definition 2.14.** A linear order is *short* if it is  $\omega_1$ -short and  $\omega_1^*$ -short.

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

**Definition 2.15.** A *test statistic* is a nominal test statistic whose codomain is short.

**Proposition 2.16.** *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 2.5,  $R$  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  $\eta$  is an order-preserving or order-reversing map from  $\omega_1$  to  $R$ . We obtain uncountably many disjoint open nonempty intervals  $(\eta(\alpha), \eta(\alpha + 2))$  where  $\alpha$  is a limit ordinal. Each of these intervals contains a point from  $C$ , which is impossible.  $\square$

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

#### 3.1 Induced test statistics

**Definition 3.1.** 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 3.2.** 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 2.3,  $f$  induces a probability distribution  $\mathbf{P}_f(X) = \mathbf{P}(f^{-1}(X))$  on  $R$ . To simplify notation, we omit the subscript  $f$ .

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

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

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  $\omega_1$ -short and so there is a sequence  $s_1 \leq s_2 \leq \dots$  in  $S_r$  cofinal with  $S_r$  so that  $S_r = \bigcup_n (r, s_n]$  and thus is measurable. Further,  $\mathbf{P}(\bigcup_n (r, s_n]) = \lim_n \mathbf{P}(r, s_n] = 0$ . 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  $\hat{\hat{f}}(x) = \hat{f}(x)$ .

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

**Theorem 3.3.** 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.  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for every  $\varepsilon \in \text{Range } f$ .

*Proof.* Obviously 2 implies 1. By Lemma 3.2, 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

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

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

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

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  $\omega_1$ -short.

**Proposition 3.4.** *Let  $R$  be an  $\omega_1$ -long ordered set. There exists an  $R$ -valued nominal test statistic  $f$  such that  $\mathbf{P}[\hat{f} \leq 0] = 1$  even though  $0 \in \text{Range } \hat{f}$ .*

*Proof.* Fix an embedding  $\eta$  of  $\omega_1$  into  $R$ , and let

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

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

For every member  $X$  of  $\Sigma$ , either  $X$  or  $R - X$  contains at most a countable subset of  $\text{Range } \eta$ . 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  $\Sigma$  as follows: If  $X$  contains at most a countable subset of  $\text{Range } \eta$  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  $\hat{f}(x) = \mathbf{P}[f \leq f(x)] = 0$ . So  $\mathbf{P}[\hat{f} \leq 0] = \mathbf{P}(L) = 1$ .  $\square$

### 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 3.3, we have  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for any  $\varepsilon \in \text{Range } f$ . But in practice people also use conservative p-values. To accommodate this practice, we give a more general definition of p-functions.

#### Definition 3.5.

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

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

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

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

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

**Theorem 3.6.** *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 3.3. To prove the other implication, assume 1. By Theorem 3.3,  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for any  $\varepsilon \in \text{Range } f$ . It remains to prove that  $\mathbf{P}[f \leq \varepsilon] \leq \varepsilon$  for every  $\varepsilon \in [0, 1] - \text{Range } f$ .

Let  $\varepsilon \in [0, 1] - \text{Range } f$ ,  $S = \{s \in \text{Range } f : s < \varepsilon\}$  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  $s_1 \leq s_2 \leq \dots$  of reals in  $S$  converging to  $\varepsilon_0$ . Then

$$\begin{aligned} \mathbf{P}[f \leq \varepsilon] &= \mathbf{P}[f \leq \varepsilon_0] = \mathbf{P}\left(\bigcup_{n=1}^{\infty} [f \leq s_n]\right) \\ &= \lim_{n \rightarrow \infty} \mathbf{P}([f \leq s_n]) = \lim_{n \rightarrow \infty} s_n = \varepsilon_0 \leq \varepsilon. \quad \square \end{aligned}$$

**Definition 3.7.** 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 1.1, the complement of every countable set is an atom.

**Lemma 4.1.** *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  $\Sigma$  is the  $\sigma$ -algebra of  $R$ , and  $\mathbf{P}$  is the probability measure on  $\Sigma$ , consider the  $\sigma$ -algebra  $\Sigma_A = \{A \cap X : X \in \Sigma\}$  and restrict  $\mathbf{P}$  to  $\Sigma_A$ . 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 2.10, there is a sequence  $x_1 \leq x_2 \leq \dots$  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  $y \leq x$  for some  $x \in I$ ). By the continuity from below, see Equation (3),

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

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 (4),

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

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

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 4.1, 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 4.2.** *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 4.1 it suffices to prove that  $\mathbf{P}[\hat{f} = \varepsilon] = 0$  for every  $\varepsilon \in \text{Range } \hat{f}$ . Fix such a number  $\varepsilon$ . Since  $\varepsilon \in \text{Range } \hat{f}$ , 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 3.2. If  $x \in X$  and  $r = f(x)$ , we have  $\varepsilon = \hat{f}(x) = \mathbf{P}[f \leq r] = \eta(r)$  and so  $[\hat{f} = \varepsilon] = \{y :$

$\hat{f}(y) = \hat{f}(x)\} = \{y : \eta(f(y)) = \eta(r)\}$ . Let  $U(r) = \{y : f(y) \geq r \text{ and } \eta(f(y)) = \eta(r)\}$  and  $U'(r) = \{y : f(y) > r \text{ and } \eta(f(y)) = \eta(r)\}$ . In the proof of Lemma 3.2 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  $r_1 > r_2 > \dots$  in  $R_0$  that is coinitial with  $R_0$ . We have

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

**Theorem 4.3.** *Let  $f$  be a diffuse test statistic. Then the induced test statistic  $\hat{f}$  is an exact p-function.*

*Proof.* By Theorem 3.6,  $\hat{f}$  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  $\varepsilon \in \text{Range } \hat{f}$ . It remains to prove that  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for every  $\varepsilon \in [0, 1] - \text{Range } \hat{f}$ . Fix such a number  $\varepsilon$ .

Let  $\varepsilon_0 = \sup\{\delta \in \text{Range } \hat{f} : \delta < \varepsilon\}$  and  $\varepsilon_1 = \inf\{\delta \in \text{Range } \hat{f} : \delta > \varepsilon\}$ . Our convention here is that  $\sup \emptyset = 0$  and  $\inf \emptyset = 1$ . There exists a sequence  $\delta_1 \leq \delta_2 \leq \dots$  in  $\text{Range } \hat{f}$  that converges to  $\varepsilon_0$ , 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  $\delta_1 \geq \delta_2 \geq \dots$  in  $\text{Range } \hat{f}$  that converges to  $\varepsilon_1$ , 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  $\mathbf{P}[\hat{f} \leq \varepsilon] = \mathbf{P}[\hat{f} \leq \varepsilon_0] = \varepsilon_0 = \varepsilon$ .  $\square$

**Proposition 4.4.** *If  $R$  is an  $\omega_1^*$ -long (and  $\omega_1$ -short) linearly ordered set endowed with the p-minimal  $\sigma$ -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.

- $\Omega$  is  $\omega_1^*$ , i.e., the set of countable ordinals with the reverse order  $\alpha < \beta \iff \beta \in \alpha$ .
- $\Sigma$  is the p-minimal  $\sigma$ -algebra on  $\Omega$ . It consists of the countable subsets of  $\Omega$  and their complements.
- $\mathbf{P}(X) = 0$  if  $X$  is countable. In particular, every  $\mathbf{P}(-\infty, \alpha] = 1$ .

Since  $R$  is  $\omega_1^*$ -long, there is an order isomorphism  $\eta$  from  $\omega_1^*$  into  $R$ ;  $\eta$  is an  $R$ -valued nominal test statistic on  $\mathcal{T}$ . Indeed, since the  $\sigma$ -algebra of  $R$  is  $\mathfrak{p}$ -minimal, it suffices to show that every  $\eta^{-1}(-\infty, r] \in \Sigma$ . Since  $\omega_1$  is well-ordered, every  $X \subseteq \omega_1$  has a minimal point in  $\omega_1$ ; accordingly every  $X \subseteq \omega_1^*$  has a maximal point in  $\omega_1^*$ . 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  $\mathfrak{p}$ -function  $\hat{\eta}$  takes only two values and thus is not exact.  $\square$

## 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 5.3 below).

As discussed in Appendix A, Egon Pearson [22] 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 applied statistics and machine learning where randomized p-values are essential: multiple hypothesis testing in bioinformatics (see, e.g., [12]) and on-line testing the hypothesis of exchangeability using conformal martingales (see, e.g., [32, Section 7.1.] and [34]). 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. [32, Theorem 8.1], which uses randomized p-values, with its predecessor, [31, Theorem 1], which avoids randomization).

Recall that the product  $\mathbf{P}_1 \times \mathbf{P}_2$  of probability measures  $\mathbf{P}_1$  and  $\mathbf{P}_2$  on measurable spaces  $(\Omega_1, \Sigma_1)$  and  $(\Omega_2, \Sigma_2)$  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 2.8, except that the assumption that the codomain of  $f$  is countable is dropped.

By Theorem 3.3, the induced  $\mathfrak{p}$ -function  $f$  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., take any  $\varepsilon \in (0, 1)$  when the sample space is a singleton. Randomized p-values are a way of making  $f$  exact, i.e., achieving the equality  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for all  $\varepsilon \in [0, 1]$ . 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 5.1.** *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  $\omega_1$ -short, i.e., that for every well-ordered set  $A$ , if there exists an order preserving map  $\eta : A \rightarrow R \times S$ , 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.  $\square$

**Lemma 5.2.** *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  $\Omega$  has the form  $\{x_1\} \times \{x_2\}$  and this is measurable. The second claim follows from the fact that  $\mathbf{P}(\{x_1\} \times \{x_2\}) = \mathbf{P}_1\{x_1\} \cdot \mathbf{P}_2\{x_2\}$ .  $\square$

Now we are ready to address the issue of randomized p-values. We start from our usual setting of a given traditional test statistic  $f : \Omega \rightarrow \mathbb{R}$  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  $\sigma$ -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  $\Omega$  is replaced by a finer test statistic

$$F(x, r) = (f(x), r) \tag{5}$$

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 5.3.** *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  $\varepsilon \in [0, 1]$ .*



*Proof.* By Lemma 5.1, 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 4.1, the  $\mathbf{P}_F$ -atoms, if any, of  $\mathbb{R} \times [0, 1]$  are singletons. By Lemma 5.2 and  $\mathbf{P}_F = \mathbf{P}_f \times \mathcal{U}$  (cf. (5)),  $\mathbf{P}_F$  does not have singleton atoms and thus is diffuse. It remains to apply Theorem 4.3.  $\square$

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 2.6 and Proposition 2.5, the function  $F$  defined by (5) 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 2.15, as Lemma 5.1 shows.

*Remark 5.4.* 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., [12] and [32]) is

$$\hat{F}(x, r) = \mathbf{P}[f < f(x)] + r\mathbf{P}[f = f(x)], \quad (6)$$

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)].

## 6 An alternative to randomizing p-values

Many practical statisticians dislike the idea of randomized p-values. A pioneer of randomized p-values (Stevens on the last page of [26]) 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 [22] 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 [19], 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 (6):

$$\mathbf{P}[f < f(x)] + \frac{1}{2}\mathbf{P}[f = f(x)]; \quad (7)$$

in other words it is defined to be the arithmetic mean of  $\mathbf{P}[f < f(x)]$  and  $\mathbf{P}[f \leq f(x)]$  (whereas the randomized p-value is distributed uniformly between

$\mathbf{P}[f < f(x)]$  and  $\mathbf{P}[f \leq f(x)]$ ). The corresponding *mid-p-function* maps each outcome  $x \in \Omega$  to the mid-p-value (7). 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. Although mathematically less elegant than randomizing p-values, it reduces the conservativeness of p-functions while maintaining their validity.

Wilcoxon's rank-sum test [33, 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  $R'_x \leq R_x$ . 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  $x$ s tend to be smaller than the  $y$ s.

Wilcoxon's rank-sum test is remarkably efficient (see, e.g., [21, 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 [24, 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 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  $(R'_x, T'_2) \leq (R_x, T_2)$ , 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

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

as the same table in [24] 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  $k/924$ ,  $k = 1, 2, \dots$ , 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 non-traditional 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  $\{1/924, 2/924, \dots, 1\}$  (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  $\{1/924, 2/924, \dots, 1\}$  (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

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

to the list  $(R_x, T_2, \dots)$ , with the order on  $(R_x, T_2, \dots, t)$  still being lexicographic. The p-value corresponding to a given value of  $(R_x, T_2, \dots, t)$  is the probability that  $(R'_x, T'_2, \dots, t') \leq (R_x, T_2, \dots, t)$ , where  $(R'_x, T'_2, \dots, t')$  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  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for any  $\varepsilon \in \{1/924, 2/924, \dots, 1\}$  and the power  $\mathbf{P} = P^{m+n}$  of any continuous probability distribution  $P$  on  $\mathbb{R}$ . On the other hand, we will have  $\mathbf{P}[f \leq \varepsilon] = \varepsilon$  for any  $\varepsilon \in [0, 1]$  and the power  $\mathbf{P} = P^{m+n}$  of any Gaussian distribution  $P$  on the real line.

*Remark 6.1.* In principle, we could have used only traditional test statistics in this section since even the most complicated of our test statistics,  $(R_x, T_2, \dots, t)$  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 test statistic, such as  $(R_x, T_2, \dots, t)$  (cf. Example 2.8). 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  $\mathbb{R}^k$ ).

## 7 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 3.3 and Proposition 3.4, this will lead to valid p-values (but possibly conservative p-functions). By Theorem 4.3, the corresponding p-functions will be exact in the case of diffuse test statistics. Finally, Theorem 5.3 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

- [1] Francis J. Anscombe. The validity of comparative experiments (with discussion). *Journal of the Royal Statistical Society A*, 111:181–211, 1948.
- [2] John Arbuthnott. 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, 1710.

- [3] Daniel Bernoulli. Recherches physiques et astronomiques sur le problème proposé pour la seconde fois par l'Académie Royale des Sciences de Paris: Quelle est la cause physique de l'inclinaison des plans des orbites des planetes par rapport au plan de l'équateur de la révolution du soleil autour de son axe; Et d'où vient que les inclinaisons de ces orbites sont différentes entre elles. *Recueil des pièces qui ont remporté les prix de l'Académie Royale des Sciences*, 3:95–122, 1735. The original Latin text occupies pages 125–144 of this volume; according to the author's preface, his French translation on pages 95–122 contains small additions and clarifications.
- [4] Garrett Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, RI, third edition, 1967. First edition: 1940. Second edition: 1948.
- [5] Frank S. Cater. On order topologies and the real line. *Real Analysis Exchange*, 25:771–780, 1999.
- [6] 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.
- [7] Herbert A. David. First (?) occurrence of common terms in statistics and probability. *American Statistician*, 49:121–133, 1995.
- [8] Herbert A. David. First (?) occurrence of common terms in statistics and probability—a second list, with corrections. *American Statistician*, 52:36–40, 1998.
- [9] Herbert A. David. First (?) occurrence of common terms in statistics and probability. In Herbert A. David and Anthony W. F. Edwards, editors, *Annotated Readings in the History of Statistics*, pages 209–246 (Appendix B). Springer, New York, 2001. This list subsumes the 1995 and 1998 lists [7, 8].
- [10] W. Edwards Deming. *Statistical Adjustment of Data*. Wiley, New York, 1943.
- [11] Gustav Deuchler. 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, 1914.
- [12] 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.
- [13] Mark W. Eudey. *On the treatment of discontinuous random variables*. PhD thesis, Statistical Laboratory, University of California, Berkeley, CA, 1949. PhD thesis. Supervised by Jerzy Neyman.

- [14] Ronald A. Fisher. *Statistical Methods for Research Workers*. Hafner, New York, 1925.
- [15] William S. Gosset [Student]. The probable error of a mean. *Biometrika*, 6:1–25, 1908.
- [16] Yuri Gurevich and Saharon Shelah. Modest theory of short chains. II. *Journal of Symbolic Logic*, 44:491–502, 1979.
- [17] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York, 1998.
- [18] Paul R. Halmos. *Naive Set Theory*. Van Nostrand, New York, 1960.
- [19] H. Oliver Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56:223–234, 1961.
- [20] Pierre-Simon Laplace. De l’action de la lune sur l’atmosphère. *Annales de Chimie et de Physique*, 24:280–294, 1823.
- [21] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York, 2005.
- [22] Egon S. Pearson. On questions raised by the combination of tests based on discontinuous distributions. *Biometrika*, 37:383–398, 1950.
- [23] Karl Pearson. 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, 1900.
- [24] John W. Pratt and Jean D. Gibbons. *Concepts of Nonparametric Theory*. Springer, New York, 1981.
- [25] Joseph G. Rosenstein. *Linear Orderings*. Academic Press, New York, 1982.
- [26] Wilfred L. Stevens. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, 37:117–129, 1950.
- [27] Stephen M. Stigler. Studies in the history of probability and statistics xxv. napoleonic statistics: the work of laplace. *Biometrika*, 62:503–517, 1975.
- [28] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press, Cambridge, MA, 1986.
- [29] Keith D. Tocher. Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika*, 37:130–144, 1950.
- [30] Isaac Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Macmillan, London, 1865.

- [31] 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.
- [32] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.
- [33] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.
- [34] 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>.

## A 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 ( $3 \times 2^2$ ) 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 clean calculation of a p-value corresponding to a non-extreme value of a test statistic was that given by Daniel Bernoulli in his

1735 paper [3]. In 1732, the Academy of Sciences of Paris had set a prize for the following problem: “What is the physical cause of the inclination of the planes of the planetary orbits in relation to the plane of revolution of the Sun about its axis; and what is the reason for the inclinations of these orbits to differ among themselves” (translation by Hald [17]). None of the memoirs submitted was deemed to be worthy of the prize, and in 1734 the Academy proposed the same subject again with a double prize, which was awarded to John Bernoulli and his son Daniel. The key problem to explain was the small inclinations of the planetary orbits of all six known planets (Mercury, Venus, Earth, Mars, Jupiter, and Saturn) to the plane of the Sun’s equator. Before looking for a physical cause, Daniel Bernoulli set out to investigate whether chance alone would be a feasible explanation. In modern terminology, his null hypothesis was that the six inclinations were chosen randomly (from the uniform distribution on the interval  $[0, 90^\circ]$ ), and his test statistic was their maximum. He computes the actual value of the test statistic as  $7^\circ 30'$ , achieved for the Earth (according to 1701 data that he used), and finds the p-value as

$$(7^\circ 30' / 90^\circ)^6 = 1/12^6 = 1/2,985,984$$

([3], pp. 98–99, [30], p. 223, [17], p. 69). The way he states this result, however, is reminiscent of the modern confusion between p-values and probabilities: “if all the orbits [orbital planes] were placed randomly with respect to the Sun’s equator, I would bet 2985983 against 1 that they would not be so close.” We can criticize, on physical grounds, his choice of the null hypothesis, but from the statistical point of view his calculation is sound (which cannot be said about two other, more Earth-centered, p-values calculated in the same paper).

Another influential (albeit less clean) calculation of a p-value was that in Laplace’s 1823 paper [20] devoted to atmospheric tides, a more difficult object of study than ocean tides. It is described in detail in, e.g., [27, Section 4] and [28, Chapter 4], and it 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 [23] about the  $\chi^2$  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  $\chi^2$  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  $\chi$  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 [26, 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 [22] 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 [29]. Randomized p-values were later used in the well-known book [24] (Section 1.5.5).