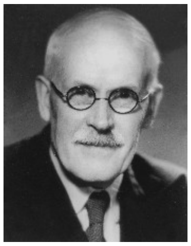
Nonparametric e-tests of symmetry

Vladimir Vovk Ruodu Wang



Users of these tests speak of the
5 per cent. point [p-value of 5%]
in much the same way as I should
speak of the $K = 10^{-1/2}$ point
[e-value of $10^{1/2}$], and of the 1
per cent. point [p-value of 1%]
as I should speak of the
$K = 10^{-1}$ point [e-value of 10].

Project "Hypothesis testing with e-values"

Working Paper #9

First posted August 18, 2022. Last revised January 26, 2024.

Project web site:
http://alrw.net/e

Abstract

The notion of an e-value has been recently proposed as a possible alternative to critical regions and p-values in statistical hypothesis testing. In this paper we consider testing the nonparametric hypothesis of symmetry, introduce analogues for e-values of three popular nonparametric tests, define an analogue for e-values of Pitman's asymptotic relative efficiency, and apply it to the three nonparametric tests. We discuss limitations of our simple definition of asymptotic relative efficiency and list directions of further research.

Contents

1 Introduction

The study of the efficiency of nonparametric tests that started in the late 1940s is often regarded as a success story in statistics. Some nonparametric tests, such as Wilcoxon's signed-rank and rank-sum tests, are highly efficient even when used in the framework of popular parametric models, such as the Gaussian model. Theoretical results mostly concern asymptotic efficiency of those tests, but there is also empirical evidence for their finite-sample efficiency. While some nonparametric tests (such as Wilcoxon's) became very popular after their high efficiency had been discovered, others (such as Wald and Wolfowitz's run test) were gradually discarded from the statistical literature after their low efficiency had been demonstrated [16, Introduction].

The usual approach to hypothesis testing is based on critical regions or p-values, but in this paper we replace them with their alternative, e-values (see, e.g., [23, 20, 7]). We show that some of the old results about the efficiency of nonparametric tests carry over to hypothesis testing based on e-values. To distinguish our notions of power, tests, etc., from the standard notions, we add the prefix "e-". (The prefix "p-" is sometimes added to signify standard notions based on p-values, but in this paper we rarely need it since the key notion that we are interested in, Pitman's asymptotic relative efficiency, is defined in terms of critical regions rather than p-values.)

We explain basics of e-testing in Sect. 2, and in particular, we state an analogue of the Neyman–Pearson lemma in e-testing. In the following section, Sect. 3, we give a simple example of a parametric e-test, one for testing the null hypothesis $N(0,1)$ against an alternative $N(\theta,1)$ in an IID situation.

In Sect. 4 we give the first, and in some sense most powerful, of the three examples of nonparametric e-tests that we discuss in this paper. It was introduced by Fisher in his 1935 book [5]. Our nonparametric null hypothesis is that of symmetry around 0 (and for simplicity we consider independent observations coming from a continuous distribution).

The material of Sects. 2–4 is standard. After that (Sect. 5) we define the asymptotic relative efficiency of e-tests in the spirit of Pitman's definition [17]. We regard our definition of asymptotic relative efficiency as a direct translation of the classical definition. Then in Sect. 6 we compute the Pitman-type asymptotic relative efficiency of the Fisher-type test discussed in Sect. 4. This is complemented by similar computations for e-versions of the sign test in Sect. 7 and Wilcoxon's signed-rank test in Sect. 8. Our results for all three tests agree perfectly with the classical results. This is just a first step, and in Sect. 9 we discuss limitations of our approach (which are considerable) and list natural directions of further research.

2 General principles of e-testing

Let P be a given probability measure on a sample space Ω (a measurable space). Our *null hypothesis* is $\{P\}$; it is simple in the sense of containing a single probability measure. (We will sometimes also refer to P as our null hypothesis.)

We observe $\omega \in \Omega$ and are interested in whether ω was generated from P. An *e-variable* for testing P is an $[0,\infty]$-valued random variable E such that $\int E \,\mathrm{d}P \le 1$. In order to be used for testing, we need to choose E before we observe ω. By Markov's inequality, E can be large only with a small probability (for any threshold $c > 1$, $P(E \ge c) \le 1/c$); therefore, observing a large E casts doubt on ω being generated from P.

In the classical Neyman–Pearson approach to hypothesis testing, in addition to P we also have an alternative hypothesis Q. The *e-power* of an e-variable E is then defined as $\int \log E \,\mathrm{d}Q$. This is an analogue of the usual notion of power, but it only works in regular cases. One of such regular cases will be discussed in the next section. The following lemma is very well known (see, e.g., [20, Sect. 2.2.1] and the references therein), and we provide a simple proof.

Lemma 2.1. *For given null and alternative hypotheses P and Q, respectively, such that $Q \ll P$, the largest e-power is attained by the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$: for any e-variable E,*

$$\int \log E \,\mathrm{d}Q \le \int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \,\mathrm{d}Q. \tag{1}$$

And if $Q \ll P$ is violated, the largest e-power is ∞.

The likelihood ratio $\mathrm{d}Q/\mathrm{d}P$ in Lemma 2.1 is understood to be the Radon–Nikodym derivative of Q w.r.t. to P.

Proof of Lemma 2.1. If $Q \ll P$ is violated, there is an event $A \subseteq \Omega$ such that $P(A) = 0$ and $Q(A) > 0$. Then the e-power of the e-variable

$$E(\omega) := \begin{cases} \infty & \text{if } \omega \in A \\ 1 & \text{otherwise} \end{cases}$$

is ∞.

It remains to consider the case $Q \ll P$. In this case, let q be a probability density function of Q w.r.t. to P. In terms of q, we can rewrite (1) as

$$\int q \log E \,\mathrm{d}P \le \int q \log q \,\mathrm{d}P, \quad \text{i.e.,} \quad \int q \log \frac{E}{q} \,\mathrm{d}P \le 0.$$

The last inequality follows from $\log x \le x - 1$. □

According to Lemma 2.1, which is an analogue for e-values of the Neyman–Pearson lemma, the optimal e-variable for testing a null hypothesis P against an alternative $Q \ll P$ is the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$. The maximum e-power is

$$\mathrm{KL}(Q \parallel P) := \int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \,\mathrm{d}Q$$

(cf. [20, Sect. 2.3] and [7, Theorem 1]). This is simply the Kullback–Leibler divergence [12] of the alternative Q from the null hypothesis P; we will call it the *optimal e-power*.

We will sometimes refer to $\log E$ as the *observed e-power* of E; the e-power is then the expectation of the observed e-power w.r. to the alternative hypothesis Q.

The notion of e-power is very close to Shafer's [20] implied target, the main difference being that the implied target only depends on the null hypothesis P and the e-variable E.

As a short detour, let us check that our notion of e-power enjoys a natural property in testing with multiple e-values. Denote by Π^Q the function

$$\Pi^Q : E \mapsto \int \log E \, \mathrm{d}Q \tag{2}$$

that maps an e-variable to its e-power. Independent e-variables $E_1, \dots, E_K$ can be combined into one e-variable using a merging function, the most common choices being convex mixtures of the product functions

$$F_M : (e_1, \dots, e_K) \mapsto \prod_{k \in M} e_k,$$

where M is a subset of $\{1, \dots, K\}$, with $F_\varnothing$ set to 1. Denote by $\mathcal{M}$ the convex hull of all functions F_M. Useful elements of the class $\mathcal{M}$ are U-statistics, symmetric merging functions discussed in [23, Sect. 4].

Proposition 2.2. *Let $\mathbf{E} = (E_1, \dots, E_K)$ be a vector of independent e-variables.*

(i) For all $F \in \mathcal{M}$, $F(\mathbf{E})$ is an e-variable.

(ii) If $\Pi^Q(E_k) > 0$ for each $k = 1, \dots, K$, then $\Pi^Q(F(\mathbf{E})) > 0$ for all $F \in \mathcal{M} \setminus \{F_\varnothing\}$.

(iii) If $\Pi^Q(E_k) \geq 0$ for each $k = 1, \dots, K$, then $\Pi^Q(F(\mathbf{E})) \geq 0$ for all $F \in \mathcal{M}$.

Proof. Part (i) follows from the fact that the product of independent e-variables is an e-variable, and a convex mixture of e-variables is an e-variable. Next we prove (ii). For all M other than $M = \varnothing$, we have

$$\Pi^Q(F_M(\mathbf{E})) = \sum_{k \in M} \Pi^Q(E_k) > 0,$$

and $\Pi^Q(F_\varnothing(\mathbf{E})) = 0$. Note that the mapping (2) is concave on the set of nonnegative random variables. Since $F(\mathbf{E})$ is a convex mixture of $F_M(\mathbf{E})$ for $M \subseteq \{1, \dots, K\}$, we get $\Pi^Q(F(\mathbf{E})) \geq 0$, and the inequality is strict unless $F = F_\varnothing$. This proves (ii). The case (iii) is similar to (ii). □

Proposition 2.2 shows that e-power remains positive when combining independent e-values with positive e-power using a large class of merging functions. As a special case of Proposition 2.2 applied to only one e-variable, if $\Pi^Q(E) > 0$, then $\Pi^Q(1 - \lambda + \lambda E) > 0$ for all $\lambda \in (0, 1]$. The operation of changing E to $1 - \lambda + \lambda E$ is common in building e-processes; see, e.g., [24].

3 A parametric e-test

We start our discussion of specific e-tests from a very simple parametric case, that of the Gaussian statistical model $Q_\theta := N(\theta, 1)$, $\theta \in \mathbb{R}$, with the variance known to be 1. We observe realizations of independent $Z_1, \dots, Z_n \sim N(\theta, 1)$. The null hypothesis P is $N(0, 1)$, and we are interested in the alternatives $Q = Q_\theta = N(\theta, 1)$ for $\theta \neq 0$.

For observations $z_1, \dots, z_n$ and a given alternative $N(\theta, 1)$, the likelihood ratio of the alternative to the null hypothesis is

$$E_\theta(z_1, \dots, z_n) := \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (z_i - \theta)^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^2\right)} = \exp\left(\theta \sum_{i=1}^n z_i - \frac{1}{2} n\theta^2\right). \tag{3}$$

The corresponding optimal e-power is

$$\int \log E_\theta \, \mathrm{d}Q_\theta = \theta n \theta - \frac{1}{2} n\theta^2 = \frac{1}{2} n\theta^2. \tag{4}$$

The interpretation of the optimal e-power (4) usually depends on the law of large numbers and its refinements (such as the central limit theorem and large deviation inequalities). The presence of log in the definition $\int \log E \, \mathrm{d}Q$ of the e-power of E under the alternative Q reflects the fact that a typical e-value is obtained by multiplying components coming from the individual observations z_i. This can be seen from (3) (and also expressions (9), (15), and (19) below, which are typical). Taking the logarithm leads to a much more regular distribution, which is, e.g., approximately Gaussian under standard regularity conditions. In the case of (3), the key component of the logarithm is $\sum_{i=1}^n z_i$, and we can apply, e.g., the central limit theorem to see that the observed e-power is between the narrow limits $\frac{1}{2} n\theta^2 \pm c\sqrt{n}\theta$ with probability close (in this particular case, even exactly equal) to $\Phi(c) - \Phi(-c)$, where $c > 0$ and Φ is the standard Gaussian cumulative distribution function.

Remark 3.1. To get the full idea of the power of E under Q, we need the whole distribution of the observed e-power $\log E$ under Q, and replacing it by its expectation is a crude step. (The next step might be, e.g., complementing the expectation with the standard deviation of $\log E$ under Q.) We leave such more realistic notions of power for future research.

We regard the family (3) of e-variables as a test (an *e-test*) of the null hypothesis $N(0, 1)$. While for several important statistical models there are uniformly most powerful p-tests (see, e.g., [14, Chap. 3]), this is not the case for e-tests, and the e-tests considered in this paper are always families of e-variables.

The fact that the e-variable (3) depends on the unknown alternative parameter θ is a disadvantage. A natural way out is to integrate it under the prior distribution $N(0, 1)$ over θ, which gives us the e-variable

$$\frac{1}{\sqrt{2\pi}} \int \exp\left(\theta \sum_{i=1}^n z_i - \frac{1}{2} n\theta^2 - \frac{1}{2}\theta^2\right) \mathrm{d}\theta$$

$$= \sqrt{\frac{1}{n+1}} \exp\left(\frac{1}{2n+2}\left(\sum_{i=1}^{n} z_i\right)^2\right) \tag{5}$$

(cf. Remark 3.2 below). Notice that the operation of integration makes the e-variable "two-sided": while (3) is monotone in $\sum_i z_i$, (5) is monotone in $|\sum_i z_i|$. The remaining disadvantage of the e-variable (5) is that it is valid only under the simple Gaussian null hypothesis $N(0,1)$. In the following sections we will replace this simple null hypothesis with a composite nonparametric one.

Remark 3.2. In our computations in this paper we often use the formula

$$\int \exp\left(-Ax^2 + Bx\right) \mathrm{d}x = \sqrt{\frac{\pi}{A}} \exp\left(\frac{B^2}{4A}\right),$$

where $A > 0$ and $B \in \mathbb{R}$. Equations (3) and (5) are simple calculations, and they appear in the context of mixture martingales, which date back to, at least, the work of Robbins (e.g., [19]); see also the more recent [10] and the references therein.

4 Fisher-type nonparametric e-test of symmetry

Let $Z_1, \dots, Z_n$ be continuous IID random variables. We are interested in the null hypothesis that their distribution is symmetric around 0. This is an example of a nonparametric hypothesis, since the distribution of $Z_1, \dots, Z_n$ is not described in a natural way by finitely many real-valued parameters. Intuitively, we are interested in two alternatives: the one-sided alternative that Z_i, even though IID, are not symmetric but shifted to the right; and the two-sided alternative that Z_i are shifted to the right or to the left.

A typical case in applications is where $Z_i := Y_i - X_i$, X_i is a pre-treatment measurement, and Y_i is a post-treatment measurement, and we are interested in whether the treatment has any effect. Assuming that raising X_i is desirable, the one-sided alternative is that the treatment is beneficial.

We will formalize our null hypothesis in a way similar to repetitive and one-off structures [22, Sects. 11.2.4 and 11.2.5]. However, we will not need general definitions and will adapt them to our special case.

The *symmetry model* for a sample size n is the pair (t, b), where $t : \mathbb{R}^n \to \Sigma$ is the mapping

$$t : (z_1, \dots, z_n) \mapsto (|z_1|, \dots, |z_n|)$$

from the *sample space* $\mathbb{R}^n$ to the *summary space* $[0, \infty)^n$, and b is the Markov kernel that maps each summary $(z_1, \dots, z_n) \in [0, \infty)^n$ to the uniform probability measure on the set

$$t^{-1}(z_1, \dots, z_n) = \{(j_1 z_1, \dots, j_n z_n) \mid (j_1, \dots, j_n) \in \{-1, 1\}^n\}. \tag{6}$$

An *e-variable* for testing the null hypothesis of symmetry is a function $E : \mathbb{R}^n \to [0,\infty]$ such that $\int E \, \mathrm{d}b(t(z_1,\dots,z_n)) \le 1$ for all $z_1,\dots,z_n$. It is *admissible* if $\le$ holds as $=$ for all $z_1,\dots,z_n$; in other words, if it ceases to be an e-variable (w.r.t. to the symmetry model) as soon as its value is increased at any point.

Remark 4.1. The definition of admissibility that we give is adapted to our current context; see [18, Sect. 9] for a more general discussion.

In this section we define the first of our three e-tests for testing symmetry. We are interested in the e-variables of the form

$$E_\lambda(z_1,\dots,z_n) := \exp\left(\lambda S(z_1,\dots,z_n) - C\right), \tag{7}$$

where $S(z_1,\dots,z_n) := \sum_{i=1}^n z_i$, $\lambda > 0$ is a positive parameter, and C is chosen to make E an admissible e-variable, i.e.,

$$C = C(\lambda, t(z_1,\dots,z_n)) := \log \int \exp(\lambda S) \mathrm{d}b(t(z_1,\dots,z_n))$$

(in other words, $C := \log \mathbb{E} \exp(\lambda S)$, the expectation being under the null hypothesis, i.e., under the symmetry model). Lemma 4.2 will give a convenient formula for computing C.

The form (7) for our e-variables can be justified by the analogy with the e-variable (3) that we obtained in the Gaussian case. The expression for the normalizing constant C will, however, be different and will be derived momentarily.

The justification of the symmetry model from the point of view of standard statistical modelling is that, under the null hypothesis of symmetry, t is a sufficient statistic giving rise to b as conditional distribution.

For simplicity, we will assume that $z_1,\dots,z_n$ are all different (under our assumption that the random variables $Z_1,\dots,Z_n$ are continuous, the realizations will be all different almost surely).

Lemma 4.2. *The value of C in (7) is given by*

$$C = \sum_{i=1}^n \log \frac{e^{\lambda z_i} + e^{-\lambda z_i}}{2}. \tag{8}$$

Proof. We find

$$e^C = 2^{-n} \sum_{j_1=0}^{1} \cdots \sum_{j_n=0}^{1} e^{\lambda j_1 z_1 + \cdots + \lambda j_n z_n} = 2^{-n} \prod_{i=1}^n \left(e^{\lambda z_i} + e^{-\lambda z_i}\right).$$

(Alternatively, we can see straight away that the average of (9) below w.r.t. to $b(t(z_1,\dots,z_n))$ is 1.) □

Plugging (8) into (7) gives the e-variable

$$E_\lambda(z_1,\dots,z_n) = e^{-C} \prod_{i=1}^n e^{\lambda z_i} = \prod_{i=1}^n \frac{e^{\lambda z_i}}{\frac{1}{2}\left(e^{\lambda z_i} + e^{-\lambda z_i}\right)}. \tag{9}$$

This is an e-version of Fisher's permutation test, which he introduced and applied to Charles Darwin's data [3, Chap. 1] in his 1935 book [5, Sects. 21 and 21.1] on experimental design.

Again, since there is no uniformly most powerful e-test, we consider a family of e-variables. The e-variable (9) is, of course, admissible.

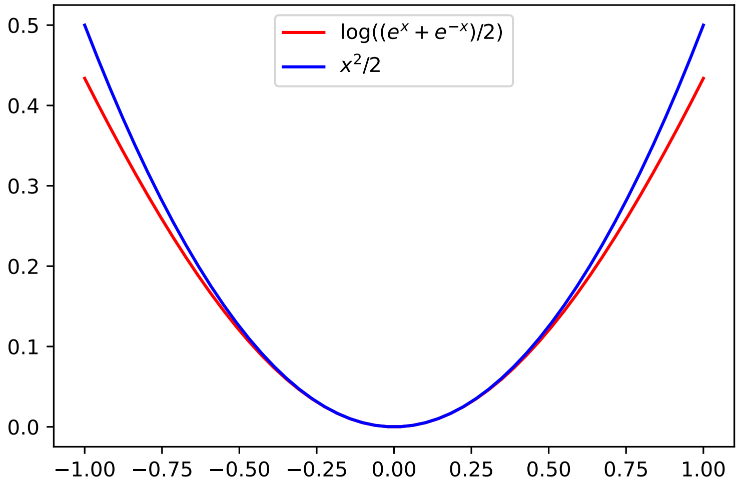


Figure 1: The inequality (11) on the log scale

The e-variable (9) dominates

$$E'_\lambda(z_1,\dots,z_n) := \prod_{i=1}^{n} e^{\lambda z_i - \lambda^2 z_i^2/2}, \tag{10}$$

in the sense $E' \le E$. Therefore, E' is also an e-variable, albeit inadmissible in general. To check the inequality $E' \le E$, it suffices to check that

$$\frac{1}{2}\left(e^x + e^{-x}\right) \le e^{x^2/2}. \tag{11}$$

Expanding both sides into Taylor's series shows that this inequality indeed holds for all x. The inequality is not excessively loose, especially for small values of x (which will be the case that we will be interested in when computing the Pitman efficiencies): cf. Figure 1.

Remark 4.3. The fact that (10) is an e-variable was established by de la Peña [4, Lemma 6.1]. Ramdas et al. [18, Sect. 10] point out that it is inadmissible, and they define several natural admissible alternatives to (9). Investigating the asymptotic relative efficiency of those admissible alternatives is an interesting direction of further research.

49	23	56
−67	28	24
8	41	75
16	14	60
6	29	−48

Table 1: Differences in eighths of an inch between cross- and self-fertilised plants of the same pair (Table 3 in [5, Sect. 17])

In order to get rid of the dependence of (9) or (10) on λ, we can integrate these expression over a prior distribution on λ. This can be easily done explicitly (see Remark 3.2) in the case of (10) and the prior distribution $N(0,1)$ on λ:

$$\frac{1}{\sqrt{2\pi}} \int \prod_{i=1}^{n} e^{\lambda z_i - \lambda^2 z_i^2/2 - \lambda^2/2} \, \mathrm{d}\lambda = \sqrt{\frac{1}{1+\sum_{i=1}^{n} z_i^2}} \exp\left(\frac{\left(\sum_{i=1}^{n} z_i\right)^2}{2+2\sum_{i=1}^{n} z_i^2}\right). \quad (12)$$

The right-hand side of (12) is close to the right-hand side of (5) under $N(0,1)$ as the null hypothesis: this follows from $\sum_{i=1}^{n} z_i^2 \approx n$ (for large n and with high probability). However (as noticed in [4]), this relatively small change drastically changes the property of validity of the e-test: while the right-hand side of (5) is an e-test of $N(0,1)$ only, the right-hand side of (12) is an e-test of the nonparametric hypothesis of symmetry.

Results for Charles Darwin's data

In this subsection we will compute Fisher-type nonparametric e-values for data used by Darwin [3, Chap. 1] to test whether cross-fertilization of plants was advantageous to the progeny as compared with self-fertilization. This was an important question from the evolutionary point of view, and Darwin's preliminary work had convinced him that cross-fertilization was indeed advantageous; in particular, nature went to great lengths to prevent self-fertilization [2].

Table 1 reports results for a small subset of Darwin's data, those for maize. This subset was analyzed for Darwin by Francis Galton (as Darwin describes in detail in [3, Chap. 1]) and was reanalyzed by Fisher in [5, Chap. 3]. Fisher offered both parametric analysis (assuming the Gaussian distribution) and novel nonparametric analysis, and his finding was that Student's t-test and Fisher's nonparametric test produce remarkably similar results.

Table 1 lists the differences in height between 15 pairs of matched plants, with a cross- and self-fertilized plant in each pair (meaning a plant grown from a cross- or self-fertilized seed, respectively). A positive difference means that the cross-fertilized plant is taller, which we *a priori* expect to happen more often. Fisher was interested in two alternatives to the null hypothesis of symmetry: the one-sided alternative of positive observations being more common than negative ones and the two-sided alternative of asymmetry (with positive observations being either more or less common than negative ones).

Fisher's p-value for testing the one-sided hypothesis is 2.634%, and his p-value for testing the two-sided hypothesis is twice as large, 5.267%. Therefore, the one-sided p-value is significant but not highly significant, whereas the two-sided p-value is not even significant.

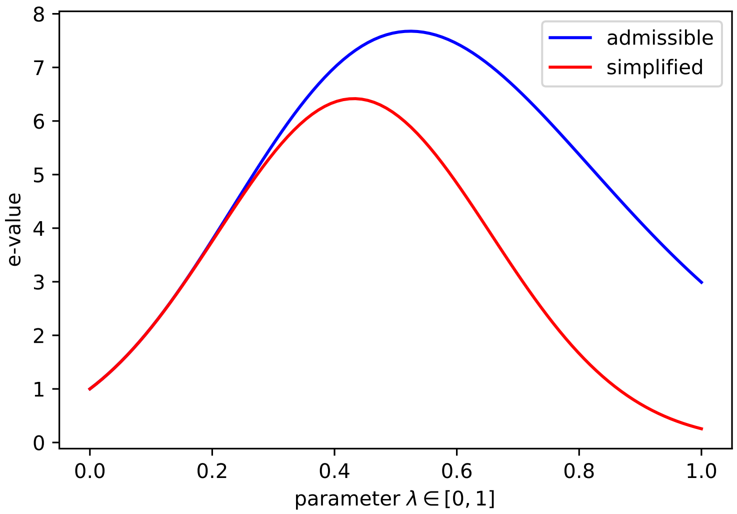


Figure 2: Results for the Fisher-type e-test applied to Darwin's data

Figure 2 plots the Fisher-type admissible e-values (9) (in blue) and the simplified e-values (10) (in red) for the parameter λ in the range $[0, 1]$. The meaning of λ depends on the scale of the numbers $z_1, \ldots, z_{15}$ in Table 1, and in order to make λ less arbitrary we normalize $z_1, \ldots, z_{15}$ by dividing them by the standard deviation of these 15 numbers. Jeffreys's [11, Appendix B] rule of thumb is to consider an e-value of 10 as being analogous to a p-value of 1% and to consider an e-value of $\sqrt{10} \approx 3.162$ as being analogous to a p-value of 5%. (See [23, Sect. 2] for a more detailed discussion of relations between e-values and p-values.) This makes Figure 2 roughly comparable to Fisher's p-values, especially if we ignore the inadmissible simplified e-values. If we guess in advance that $\lambda := 0.5$ is a good parameter value, we will get an e-value of 7.651. More realistically, averaging the e-values for $\lambda \in [0, 1]$ will give the one-sided e-value 5.149. Replacing $\lambda \in [0, 1]$ by $\lambda \in [-1, 1]$ gives the two-sided e-value 2.633 not reaching the threshold of $\sqrt{10}$.

5 Pitman-type asymptotic relative efficiency

The following definition is in the spirit of Pitman's definition, which can be found in, e.g., [21, Sect. 14.3]. Let $(Q_\theta \mid \theta \in \Theta)$ be a statistical model, i.e., a set of probability measures on the real line $\mathbb{R}$, with the observations generated from one of those probability measures in the IID fashion. We assume, for simplicity, that $\Theta = \mathbb{R}$ and regard Q_0 as the null hypothesis; informally, the alternative is either one-sided, $\theta > 0$, or two-sided, $\theta \neq 0$ (for specific e-tests, we will have the same results for one-sided and two-sided Pitman efficiency). By an e-variable we mean an e-variable w.r.t. to Q_0^n. In our asymptotic framework we consider sequences of parameter values θ_ν that depend on the "difficulty" $\nu = 1, 2, \ldots$ of our testing problem; in the one-sided case we will assume $\theta_\nu \downarrow 0$ (the sequence is strictly decreasing and converges to 0), and in the two-sided case we will assume $\theta_\nu \to 0$.

Let $\mathcal{E}_1^n$ and $\mathcal{E}_2^n$ be families of e-variables on $\mathbb{R}^n$; we are interested in the case where $\mathcal{E}_1^n$ is a family of interest to us (a nonparametric e-test such as (9) above, or (16) or (17) below) and $\mathcal{E}_2^n$ is the baseline family of all e-variables on $\mathbb{R}^n$. The *asymptotic relative efficiency* of $\mathcal{E}_1^n$ w.r.t. to $\mathcal{E}_2^n$ is c if, for any $\beta > 0$ and any $\theta_\nu \downarrow 0$ (one-sided case) or $\theta_\nu \to 0$ (two-sided case), we have $n_{\nu,2}/n_{\nu,1} \to c$, where $n_{\nu,j}$, $j = 1, 2$, is the minimal number of observations n such that

$$\exists E \in \mathcal{E}_j^n : \int \log E \, \mathrm{d}Q_{\theta_\nu}^n \geq \beta.$$

For example, if the asymptotic relative efficiency is 0.5, the best e-test in $(\mathcal{E}_1^n)$ requires twice as many observations n as the best test in $(\mathcal{E}_2^n)$ to achieve the same e-power (if the best e-tests exist).

The idea of using an auxiliary parametric statistical model (Q_θ), such as the Gaussian model, to assay the efficiency of nonparametric e-tests is illustrated in Figure 3. We are testing a nonparametric null hypothesis (the hypothesis of symmetry in this paper), but we are afraid that for a popular parametric model (the Gaussian model $Q_\theta := N(\theta, 1)$ in this paper, which plays the role of an *assay statistical model*) our testing method loses a lot. We are interested in the case where the intersection between the nonparametric null hypothesis and the assay model contains only one probability measure; we refer to this intersection as the *parametric null hypothesis* in Figure 3 (in this paper, it is $\{N(0,1)\}$). For a given simple alternative hypothesis $Q = Q_\theta$ in the assay model (shown as the red dot in Figure 3), we are hoping to show that the best e-power achieved for testing the simple parametric null hypothesis vs Q is not much better than the best e-power achieved for testing the composite (and usually massive) nonparametric null hypothesis. Or, if Pitman-type notion of efficiency is to be used (as in this paper), that the same e-power is attained for numbers of observations that are not wildly different.

Our use of the Gaussian model with variance 1 as assay model motivates using (7) with $S(z_1, \ldots, z_n) := z_1 + \cdots + z_n$ as a nonparametric e-test. The sign and Wilcoxon versions will be natural modifications (corresponding to relaxing the symmetry assumption, as explained in Remark 7.1 below).

For all three nonparametric e-tests considered in this paper (Sects. 6–8 below) we will need the number $n_{\nu,2}$ of observations required by our baseline, which is, by Lemma 2.1, the likelihood ratio $\mathrm{d}N(\theta_\nu, 1)/\mathrm{d}N(0,1)$. By (4), achieving an

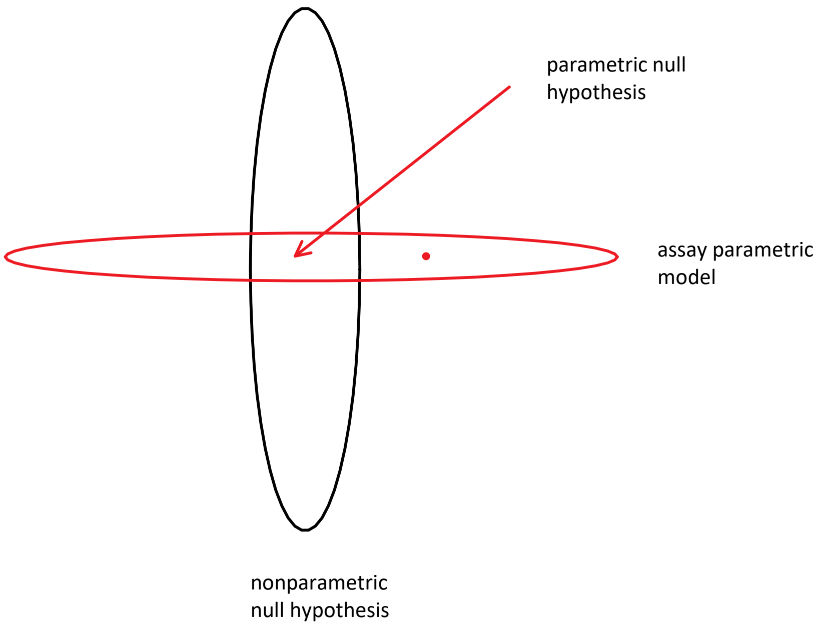


Figure 3: Assaying a non-parametric e-test

e-power of β requires approximately

$$2\beta\theta_\nu^{-2} \tag{13}$$

observations (namely, $\lceil 2\beta\theta_\nu^{-2} \rceil$ observations).

Remark 5.1. In the context of regular statistical models such as Gaussian, it is natural to set $\theta_\nu := c\nu^{-1/2}$. In this case the "difficulty" ν (referred to as "time" in [21, Sect. 14.3]) becomes proportional to the number of observations required to achieve a given e-power.

6 Asymptotic efficiency of the Fisher-type e-test

In the classical case, the relative efficiency of Fisher's test is 1 [6, Chapter 7, Example 4.1], as first shown by Hoeffding [9] (according to Mood [15]). Let us check that this remains true for the e-version as well.

First we find informally a suitable e-variable in the family (9) and then show that it requires the optimal number (13) of observations to achieve an e-power of β. Under the symmetry model, each observation z_i is split into its magnitude $m_i := |z_i|$ and sign $s_i := \operatorname{sign}(z_i)$. Given the magnitudes, the signs are independent and $\mathbb{P}(s_i = 1) = 1/2$ under the null hypothesis $N(0, 1)$ and

$$\begin{aligned}\mathbb{P}(s_i = 1) &= \frac{\exp\left(-\frac{1}{2}(m_i - \theta_\nu)^2\right)}{\exp\left(-\frac{1}{2}(m_i - \theta_\nu)^2\right) + \exp\left(-\frac{1}{2}(-m_i - \theta_\nu)^2\right)}\\ &= \frac{\exp(\theta_\nu m_i)}{\exp(\theta_\nu m_i) + \exp(-\theta_\nu m_i)}\end{aligned}$$

under the alternative hypothesis $N(\theta_\nu, 1)$. The conditional likelihood ratio for the signs is

$$\prod_{i=1}^{n} \frac{2\exp(\theta_\nu z_i)}{\exp(\theta_\nu m_i) + \exp(-\theta_\nu m_i)} = \prod_{i=1}^{n} \frac{\exp(\theta_\nu z_i)}{1 + \theta_\nu^2 m_i^2/2 + o(\theta_\nu^2 m_i^2)}.$$

This is Fisher's e-test (9) corresponding to $\lambda := \theta_\nu$. Its observed e-power is

$$\sum_{i=1}^{n} \left(\theta_\nu z_i - \theta_\nu^2 m_i^2/2 + o(\theta_\nu^2 m_i^2)\right) = \theta_\nu \sum_{i=1}^{n} z_i - (1 + o(1))\frac{\theta_\nu^2}{2} \sum_{i=1}^{n} m_i^2.$$

Since, under the alternative hypothesis $N(\theta_\nu, 1)$,

$$\mathbb{E} \sum_{i=1}^{n} z_i = n\theta_\nu$$

and

$$\mathbb{E} \sum_{i=1}^{n} m_i^2 = \mathbb{E} \sum_{i=1}^{n} z_i^2 = n + n\theta_\nu^2 = (1 + o(1))n,$$

the e-power is

$$n\theta_\nu^2 - (1 + o(1))\frac{\theta_\nu^2}{2} n \sim \frac{1}{2} n\theta_\nu^2.$$

We obtain the optimal e-power (4) with $\theta = \theta_\nu$, and so the asymptotic relative efficiency of Fisher's e-test is 1.

7 Sign e-test

In this and following sections we use (7) for different statistics S, and with C still chosen to make E_λ an admissible e-variable. In this section we make the simplest choice of $S(z_1, \dots, z_n)$ in (7), which is the number k of positive z_i among $z_1, \dots, z_n$. This gives the *sign e-test* with parameter $\lambda > 0$. The use of the signs for hypothesis testing goes back to [1].

To obtain a useful alternative representation of the sign e-test, let $p \in (0, 1)$ be defined by the equation

$$\frac{p}{1 - p} = e^\lambda$$

(so that λ becomes the log-odds ratio). The e-variable (7) then becomes

$$E_\lambda(z_1,\dots,z_n) = e^{\lambda k - C} = p^k(1-p)^{-k}e^{-C} = \frac{p^k(1-p)^{n-k}}{2^{-n}}. \tag{14}$$

The last expression is the likelihood ratio of an alternative to the null hypothesis, and so is an admissible e-variable. This gives us the representation

$$E_p(z_1,\dots,z_n) := \frac{p^k(1-p)^{n-k}}{2^{-n}} \tag{15}$$

of the sign e-test.

The equality between the last two terms in (14) gives an explicit expression for C,

$$C = -n\log(2(1-p)) = n\log\frac{1+e^\lambda}{2},$$

which in turn gives the alternative representation

$$E_\lambda(z_1,\dots,z_n) = e^{\lambda k}\left(\frac{2}{1+e^\lambda}\right)^n \tag{16}$$

of the sign e-test.

In view of our informal alternative hypothesis, we are often interested in $\lambda > 0$, i.e., $p > 1/2$.

Remark 7.1. Notice that in this section we are actually testing a wider null hypothesis than the symmetry model, since the magnitudes of z_i do not matter. Namely, the sign e-test is valid for testing the hypothesis that the signs of $Z_1,\dots,Z_n$ are ± 1 independently. A similar remark can also be made about the nonparametric e-test discussed in the following section, which in fact tests an intermediate null hypothesis.

As before, we have a dependence of the sign e-test (15) on a parameter, p. To get rid of this dependence, we can, e.g., integrate (15) over $p \in [0,1]$, obtaining

$$E(z_1,\dots,z_n) := 2^n \mathrm{B}(k+1, n-k+1),$$

where B is the beta function. For testing the one-sided hypothesis we can integrate (15) over the uniform probability measure on $[0.5, 1]$, which gives

$$E(z_1,\dots,z_n) := 2^{n+1}\bigl(\mathrm{B}(k+1,n-k+1) - \mathrm{B}(0.5; k+1, n-k+1)\bigr),$$

where the second entry of B stands for the incomplete beta function.

Efficiency of the sign test

In this and next sections we consider the same assay parametric model and still assume that the null hypothesis is $N(0,1)$ and the alternative is $N(\theta_\nu, 1)$. Suppose we only observe the signs s_i of z_i, which is sufficient when testing the null hypothesis with the sign e-test. By Lemma 2.1 the largest e-power for an e-variable of this kind will be achieved by the likelihood ratio for the signs.

The sign of Z_i is 1 with probability 1/2 under the null hypothesis and $1/2 + \tilde{\theta}_\nu/\sqrt{2\pi}$ under the alternative for $\tilde{\theta}_\nu \sim \theta_\nu$, due to the first-order Taylor approximation of the standard Gaussian cumulative distribution function Φ. With k being the number of positive z_i, the likelihood ratio for the signs is

$$\frac{\left(\frac{1}{2}+\frac{\tilde{\theta}_\nu}{\sqrt{2\pi}}\right)^k \left(\frac{1}{2}-\frac{\tilde{\theta}_\nu}{\sqrt{2\pi}}\right)^{n-k}}{(1/2)^n} = \left(1+\sqrt{\frac{2}{\pi}}\tilde{\theta}_\nu\right)^k \left(1-\sqrt{\frac{2}{\pi}}\tilde{\theta}_\nu\right)^{n-k}.$$

This is an instance of the sign e-test (15), corresponding to $p = 1/2 + \tilde{\theta}_\nu/\sqrt{2\pi}$. The observed e-power of this e-test is

$$k \log\left(1+\sqrt{\frac{2}{\pi}}\tilde{\theta}_\nu\right) + (n-k)\log\left(1-\sqrt{\frac{2}{\pi}}\tilde{\theta}_\nu\right) \\ = (2k-n)\sqrt{\frac{2}{\pi}}\tilde{\theta}_\nu - \frac{1}{\pi}n\tilde{\theta}_\nu^2 + o(n\tilde{\theta}_\nu^2)$$

(we have used the second-order Taylor approximation). This gives the e-power

$$\left(2\left(\frac{1}{2}+\frac{\tilde{\theta}_\nu}{\sqrt{2\pi}}\right)n - n\right)\sqrt{\frac{2}{\pi}}\tilde{\theta}_\nu - \frac{1}{\pi}n\tilde{\theta}_\nu^2 + o(n\tilde{\theta}_\nu^2) = \frac{1}{\pi}n\tilde{\theta}_\nu^2 + o(n\tilde{\theta}_\nu^2) \sim \frac{1}{\pi}n\theta_\nu^2.$$

To achieve an e-power of β, the sign e-test needs $\sim \pi\beta\theta_\nu^{-2}$ observations. Therefore, the asymptotic efficiency of the sign e-test is $2/\pi \approx 0.64$, exactly the same as in the standard case [6, Example 3.1]. (In the standard case the sign test is usually compared with the t-test, but in this paper we use an even more basic assay parametric model; namely, we assume that the variance is known to be 1.)

Since the asymptotic efficiency is approximately 2/3, we can say that the sign test wastes every third observation in our Gaussian setting. This is the least efficient of the three nonparametric e-tests considered in this paper when efficiency is measured using the Gaussian assay model as yardstick.

Sign test for Darwin's data

It is interesting that the sign test gives the one-sided p-value of 0.00369 and the two-sided p-value of 0.00739. In contrast with Fisher's p-test, both p-values are highly significant, the reason being that the two negative numbers in Table 1 are so large in absolute value.

Figure 4 is an analogue of Figure 2 for the sign test. The attainable e-values are now much larger, and the average over all $p \in [0,1]$ is 19.310. To use Jeffreys's [11, Appendix B] expressions, we have strong evidence against the null hypothesis of cross- and self-fertilization being equally efficient. The

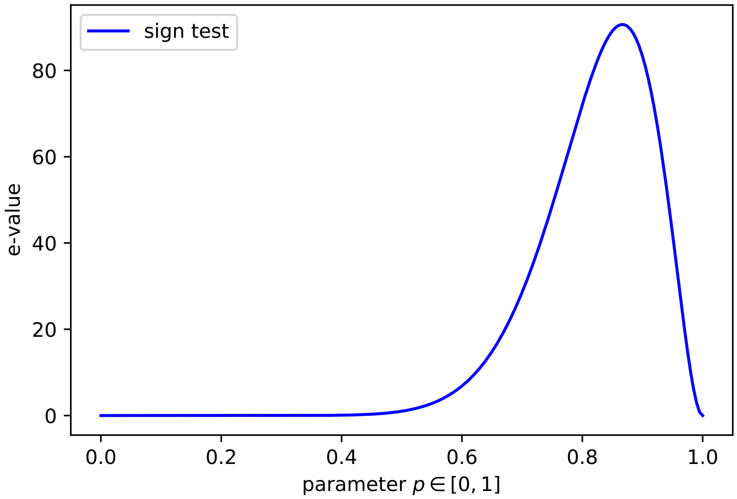


Figure 4: Results for the sign e-test applied to Darwin's data

corresponding one-sided e-value, found as the average over all $p \in [0.5, 1]$, is 38.544, and in Jeffreys's terminology it provides very strong evidence (for cross-fertilization tending to produce taller plants, in this context).

Table 1 comprises only small part of the overwhelming evidence in favour of cross-fertilization collected by Darwin over 11 years. Darwin chose maize to illustrate his and Galton's statistical methods in [3, Chap. 1], but in [3, Chaps. 2–6] he has 99 similar tables (with our Table 1 corresponding to Darwin's Table 97). With this amount of evidence, statistics is hardly needed to see that the evidence is really overwhelming.

8 Wilcoxon's signed-rank e-tests

Wilcoxon's signed-rank test [25] is based on arranging the magnitudes $|z_i|$ of the observations in the ascending order and assigning to each its *rank*, which is a number in the range $\{1, \dots, n\}$: the observation z_i with the smallest $|z_i|$ gets rank 1, the one with the second smallest $|z_i|$ gets rank 2, etc. Notice that the symmetry model (i.e., the uniform probability measure on (6)) implies that for any set $A \subseteq \{1, \dots, n\}$, the probability is 2^{-n} that the observations with the ranks in A will be positive and all other observations will be negative. This determines the distribution (conditional on the magnitudes $|z_i|$) of Wilcoxon's statistic V_n defined as the sum of the ranks of the positive observations.

We will be interested in the nonparametric e-test (7) with $S := V_n$, i.e.,

$$E_\lambda(z_1, \dots, z_n) := \exp(\lambda V_n - C) . \tag{17}$$

The following lemma gives a convenient formula for computing C.

Lemma 8.1. *The value of C in (17) is given by*

$$C = \sum_{i=1}^{n} \log \frac{1+e^{\lambda i}}{2}. \tag{18}$$

Proof. Using Fisher's conditional distribution (the uniform probability measure on (6)), we can write C in the form

$$C = \log \left(2^{-n} \sum_{A \subseteq \{1,\dots,n\}} \exp(\lambda \operatorname{sum}(A)) \right),$$

where $\operatorname{sum}(A)$ is the sum of all elements of A. Setting

$$\Sigma_i := \sum_{A \subseteq \{1,\dots,i\}} \Lambda^{\operatorname{sum}(A)},$$

where $\Lambda := \exp(\lambda)$, and using the recursion

$$\Sigma_i = \Sigma_{i-1} + \Lambda^i \Sigma_{i-1}$$

(obtained by splitting all subsets of $\{1,\dots,i\}$ into those that do not contain i and those that do), we obtain

$$\Sigma_n = \prod_{i=1}^{n} (1+\Lambda^i). \qquad \square$$

Plugging (18) into (17), we obtain *Wilcoxon's signed-rank e-test*

$$E_\lambda(z_1,\dots,z_n) := \exp(\lambda V_n) \prod_{i=1}^{n} \frac{2}{1+e^{\lambda i}}. \tag{19}$$

Efficiency of Wilcoxon's signed-rank e-test

Our derivation in this subsection will follow [13, Example 3.3.6]. The statistic

$$T_n := V_n / \binom{n}{2}, \tag{20}$$

V_n being Wilcoxon's signed-rank statistic defined at the beginning of this section, is asymptotically normal both under the null hypothesis $N(0,1)$,

$$T_n \sim N\left(\frac{1}{2}, \frac{1}{3n}\right), \tag{21}$$

and under the alternative hypothesis $N(\theta_\nu, 1)$,

$$T_n \sim N\left(\frac{1}{2} + \frac{\theta_\nu}{\sqrt{\pi}}, \frac{1}{3n}\right). \tag{22}$$

The mean value $1/2+\theta_\nu/\sqrt{\pi}$ in (22) is found as the first-order approximation to the probability of $Z_1+Z_2>0$, where Z_1 and Z_2 are independent and distributed according to the alternative hypothesis $N(\theta_\nu,1)$ (see [13, (3.3.40)]). Namely, it is obtained from $Z_1+Z_2\sim N(2\theta_\nu,2)$ and from the standard Gaussian density being $1/\sqrt{2\pi}$ at 0.

From (21) and (22) we obtain the asymptotic likelihood ratio

$$\frac{\exp\left(-\frac{1}{2}\left(T_n-\frac{1}{2}-\frac{\theta_\nu}{\sqrt{\pi}}\right)^2/\frac{1}{3n}\right)}{\exp\left(-\frac{1}{2}\left(T_n-\frac{1}{2}\right)^2/\frac{1}{3n}\right)}=\exp\left(3n\left(T_n-\frac{1}{2}\right)\frac{\theta_\nu}{\sqrt{\pi}}-\frac{3n}{2}\frac{\theta_\nu^2}{\pi}\right) \qquad (23)$$

(of the form (17); see below). The observed e-power is obtained by removing the exp, and then the e-power is obtained by taking the expectation w.r.t. to T_n distributed as (22). Therefore, the e-power is, asymptotically,

$$3n\frac{\theta_\nu}{\sqrt{\pi}}\frac{\theta_\nu}{\sqrt{\pi}}-\frac{3n}{2}\frac{\theta_\nu^2}{\pi}=\frac{3n}{2}\frac{\theta_\nu^2}{\pi}.$$

The number of observations required for achieving an e-power of β is, asymptotically,

$$\frac{2\pi}{3}\beta\theta_\nu^{-2}.$$

Comparing this with the baseline (13) gives the asymptotic relative efficiency of $3/\pi\approx 0.955$, as in the classical case. Wilcoxon's test wastes one observation out of about 22 (under the Gaussian model as compared with the e-test optimized for that model).

The approximate e-test used in this calculation (given by the right-hand side of (23)) is of the form (17) with

$$\lambda := \frac{3n\theta_\nu}{\binom{n}{2}\sqrt{\pi}}$$

(obtained by expressing (23) in terms of V_n using (20)). This, however, ignores the definition of C in (17). In practical application we should use, of course, the precise expression (19).

9 Directions of further research

In the previous sections we mentioned several limitations of our definitions. In this concluding section we will add further details.

The notion of e-power as used in the definition of efficiency

Our notion of e-power for an e-variable E is crude in that it depends only on the expectation of $\log E$, as explained in Remark 3.1. This crudeness is inherited by our definition of the asymptotic relative efficiency of e-tests. According to our definition in Sect. 5, the asymptotic relative efficiency is c if $n_{\nu,2} \sim c n_{\nu,1}$. This statement will be particularly useful if, under the alternative hypothesis, the full distribution of the original likelihood ratio, such as (3) for $\theta = \theta_\nu$ and $n_{\nu,2}$ observations, is close, in a suitable sense, to the full distribution of the e-test, such as (9), (16), or (19) (with $n_{\nu,1}$ observations and the corresponding value of the parameter). Therefore, a fuller treatment of asymptotic relative efficiency will not use e-power directly (which will make it more complicated).

Definition of efficiency in terms of mixtures

Our definition of Pitman-type efficiency is close to being a direct translation of the classical one. It considers the alternatives $N(0, \theta_\nu)$ that approach the null hypothesis $N(0,1)$ as the difficulty ν increases. In the classical case, this works perfectly for many popular assay models because of the existence of a uniformly most powerful test: the optimal size α critical region does not depend on ν (assuming $\theta_\nu > 0$). In the e-case, on the contrary, the optimal e-variable does depend on ν.

A possible alternative definition would be to replace $N(\theta_\nu, 1)$ by a mixture $\int N(\theta, 1)\mu_\nu(\mathrm{d}\theta)$ of $N(\theta, 1)$ w.r.t. to a probability measure $\mu_\nu(\mathrm{d}\theta)$ that is increasingly concentrated around $\theta = 0$ as $\nu \to \infty$. In a sense, the assay statistical model considered in this paper is "pure" in that it consists of pure Gaussian distributions. Considering mixtures $\int N(\theta, 1)\mu_\nu(\mathrm{d}\theta)$ would make the results more realistic but would also make the definitions more complicated.

Other assay models

In our efficiency results, the Gaussian model can be replaced by other statistical models. It is particularly interesting to compare nonparametric e-tests with the optimal e-tests under those models; nowadays, comparison with the t-test, which was done in many of the classical papers (e.g., [8]), looks less convincing for non-Gaussian assay models.

Our choice of the form (7) of the nonparametric e-tests considered in this paper was motivated by the Gaussian assay model: see the likelihood ratio (3). Using other assay models would lead to other nonparametric e-tests. Therefore, varying the assay model may be a useful design tool for nonparametric e-tests.

Other notions of efficiency

The Pitman-type notion of efficiency is "local", in the sense of being defined in terms of progressively more difficult alternatives that tend to the null hypothesis as $\nu \to \infty$. It is the most popular notion of efficiency for nonparametric tests, but it would be interesting to develop e-versions of other, non-local, notions of asymptotic relative efficiency (see, e.g., [16, Chap. 1]).

Acknowledgements

Many thanks to the anonymous reviewers of the journal version of this paper. Vladimir Vovk's research has been supported by Mitie. Ruodu Wang acknowledges financial support by grants CRC-2022-00141 and RGPIN-2018-03823 from the Natural Sciences and Engineering Research Council of Canada.

References

[1] 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.

[2] Charles Darwin. *On the Various Contrivances by Which British and Foreign Orchids Are Fertilised by Insects, and On the Good Effects of Intercrossing*. John Murray, London, 1862.

[3] Charles Darwin. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. John Murray, London, 1876.

[4] Victor H. de la Peña. A general class of exponential inequalities for martingales and ratios. *Annals of Probability*, 27:537–564, 1999.

[5] Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935. Section 21.1 appeared in the 7th edition (1960).

[6] Donald A. S. Fraser. *Nonparametric Methods in Statistics*. Wiley, New York, 1957.

[7] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801 [math.ST], arXiv.org e-Print archive, June 2020. Journal version is to appear in *Journal of the Royal Statistical Society B* (with discussion).

[8] Joseph L. Hodges and Erich L. Lehmann. The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, 27:324–335, 1956.

[9] Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23:169–192, 1952.

[10] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49:1055–1080, 2021.

[11] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961.

[12] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[13] Erich L. Lehmann. *Elements of Large-Sample Theory*. Springer, New York, 1999.

[14] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, Cham, fourth edition, 2022.

[15] A. M. Mood. On the asymptotic efficiency of certain nonparametric two-sample tests. *Annals of Mathematical Statistics*, 25:514–522, 1954.

[16] Yakov Nikitin. *Asymptotic Efficiency of Nonparametric Tests*. Cambridge, Cambridge University Press, 1995.

[17] Edwin J. G. Pitman. Lecture notes on nonparametric statistical inference. Columbia University, New York, 1948.

[18] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. Technical Report arXiv:2009.03167 [math.ST], arXiv.org e-Print archive, September 2020 (version 2).

[19] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41:1397–1409, 1970.

[20] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society A*, 184:407–478, 2021.

[21] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

[22] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

[23] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49:1736–1754, 2021.

[24] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting (with discussion). Technical Report arXiv: 2010.09686 [math.ST], arXiv.org e-Print archive, August 2022. Journal version is to appear in *Journal of the Royal Statistical Society B* (with discussion).

[25] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.