Statistics Surveys
Vol. 4 (2010) 1–39
ISSN: 1935-7516
DOI: 10.1214/09-SS051

Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules*

Michael P. Fay and Michael A. Proschan

6700B Rockledge Drive, Bethesda, MD 20892-7609
e-mail: mfay@niaid.nih.gov; proscham@niaid.nih.gov

Abstract: In a mathematical approach to hypothesis tests, we start with a clearly defined set of hypotheses and choose the test with the best properties for those hypotheses. In practice, we often start with less precise hypotheses. For example, often a researcher wants to know which of two groups generally has the larger responses, and either a t-test or a Wilcoxon-Mann-Whitney (WMW) test could be acceptable. Although both t-tests and WMW tests are usually associated with quite different hypotheses, the decision rule and p-value from either test could be associated with many different sets of assumptions, which we call perspectives. It is useful to have many of the different perspectives to which a decision rule may be applied collected in one place, since each perspective allows a different interpretation of the associated p-value. Here we collect many such perspectives for the two-sample t-test, the WMW test and other related tests. We discuss validity and consistency under each perspective and discuss recommendations between the tests in light of these many different perspectives. Finally, we briefly discuss a decision rule for testing genetic neutrality where knowledge of the many perspectives is vital to the proper interpretation of the decision rule.

Keywords and phrases: Behrens-Fisher problem, interval censored data, nonparametric Behrens-Fisher problem, Tajima's D, t-test, Wilcoxon rank sum test.

Received July 2009.

Contents

*This paper was accepted by Peter J. Bickel, the Associate Editor for the IMS.

1. Introduction

In this paper we explore assumptions for statistical hypothesis tests and how several sets of assumptions may relate to the interpretation of a single decision rule (DR). Often statistical hypothesis tests are developed under one set of assumptions, then subsequently the DR is shown to remain valid after relaxing those original assumptions. In other situations the later conditions that the DR is studied under are not a relaxing of original assumptions, but an exploration of an entirely different pair of probability models which neither completely contain the original probability models nor are contained within them. In either case, both the original interpretations of the DR and the later interpretations are always available to the user each time the DR is applied, and the major point of this paper is that it may make sense to package a single DR with several sets of assumptions.

We use the term 'hypothesis test' to denote a DR coupled with a set of assumptions that delineate the null and alternative hypotheses. Some DRs will be approximately valid for several sets of assumptions, and we can package these assumptions together with the DR as a *multiple perspective DR* (MPDR), where each perspective is a different hypothesis test using the same DR.

The MPDR outlook is a way of looking at the assumptions of the statistical DR and how the DR is interpreted, so we start in Section 2 discussing assumptions in scientific research in general, showing how the MPDR assumptions (i.e., statistical assumptions) fit into scientific inferences. In Section 3 we

formalize our notation and terminology surrounding MPDRs. In Section 4 we define the MPDR and discuss some useful properties. In Section 5 we detail some MPDRs for two sample tests of central tendency, formally stating many of the perspectives and the associated properties. This is the primary example of this paper and it fleshes out cases where some perspectives are subsets of other perspectives within the same MPDR. In Section 6.1 we discuss tests for interval censored data as an example of a different use of the MPDR. In this case, the MPDR outlook takes two different DRs developed under two different sets of assumptions and shows that either DR may be applied under the other set of assumptions, and we can compare the two decisions by looking at them from the same perspective (i.e., from the same set of assumptions). In Section 6.2 we discuss genetic tests of neutrality as an example of how having several perspectives on a DR may be vital to the proper interpretation of the decision.

2. Assumptions in scientific research

In order to show that the MPDR framework has practical value, we need to first outline how statistical assumptions fit into scientific research. Thus, although non-statistical assumptions for scientific research are not our main focus, we briefly discuss them in this section. Throughout this section we refer to the Physician's Health Study (PHS) (Hennekens, Eberlein for PHS Research Group, 1985; Sterring Committee of the PHS Research Group, 1988, 1989), as a specific example to clarify general concepts.

The Physician's Health Study was a randomized, 2 by 2 factorial, double blind, placebo-controlled clinical trial of male physicians in the US between the ages of 40 and 84. An invitational questionnaire was mailed to 261,248 individuals, and 33,211 were willing, eligible and started on a run-in phase of the study. Of these individuals, 22,071 adhered to their regimen sufficiently well to be enrolled in the randomized portion of the study, where each subject was randomized to either aspirin or aspirin-placebo and either β-carotene or β-carotene-placebo. We will focus on the aspirin aspect of the study which was designed to detect whether alternate day consumption of aspirin would reduce total cardiac mortality and death from all causes.

We begin our scientific assumption review with the influential book by Mayo (Mayo, 1996, see also Mayo and Spanos, 2004, 2006), which describes how hypothesis tests are used in scientific research, and it "successfully knots together ideas from the philosophy of science, statistical concepts, and accounts of scientific practice" (Lehmann, 1997). Mayo (1996) starts with a "primary model" which in her examples is often a fairly specific framing of the problem in a statistical model. In this aspect, Mayo (1996) matches closely with the typical presentation of a statistical model by a mathematical statistician. For example, Mayo (1996) goes into great detail on Jean Perrin's experiments on Brownian motion, describing the primary model as (Table 7.3):

> Hypothesis: $\mathcal{H}$: the displacement of a Brownian particle over time t, S_t follows the Normal distribution with $\mu = 0$ and variance$= 2Dt$.

This is a wonderful example of a statistical model that describes a scientific phenomenon. Although it is not framed as a null and alternative set of assumptions, Mayo's (1996) "primary model" is nonetheless framed as a statistical model. Mayo (1996) then defines the other models and assumptions which make up the particular scientific inquiry (models of experiment, models of data, experimental design, data generation). It is not helpful to go into the details of those models here, particularly since:

> [Mayo does] not want to be too firm about how to break down an inquiry into different models since it can be done in many different ways (Mayo, 1996, p. 222).

The point of this paper is that although there are examples where the primary hypothesis that drives the experimental design can be stated within a clear statistical model (e.g., Perrin's experiments), often in the biological sciences the motivating primary theory is much more vague and perhaps many different statistical models could equally well describe the primary scientific theory. In other words, often the null hypothesis is meant to express that the data are somehow random, but that randomness may be formalized by many different statistical models. This vagueness and lack of focus on one specific statistical model is inherent in biological phenomena, since unlike the physical sciences, often there are several statistical models that can equally well describe for example, a male physician's response to aspirin or aspirin-placebo. Thus, Mayo's (1996) notion that the design of an experiment begins with a primary model which can be represented as a statistical model does not appear to describe many of the biological experiments we encounter in our work. If possible, a statistical model based on the science of the application is preferred; however, often so little is known about the mechanism of action of the effect to be measured that any of several different statistical models could be applied.

Note that although this paper emphasizes hypothesis testing and p-values, we are not implying that other statistics should not supplement the p-value. For example, if a meaningful confidence interval is available then it can add valuable information. Similarly, power calculations or severity calculations (see e.g., Mayo, 2003, or Mayo and Spanos, 2006) could also supplement the hypothesis test. The problem with all three of these statistics is that often more structure is required of the hypothesis test assumptions in order to define these statistics.

Let us outline a typical experiment in the biological sciences. Here, we will start with what we will call a *scientific theory*, which is less connected to a particular statistical model than Mayo's (1996) "primary model". In the PHS a scientific theory is that prolonged low-dose aspirin will decrease cardiovascular mortality. This scientific theory is not attached to a particular statistical model; for example, that theory is not that low-dose aspirin will decrease cardiovascular mortality by the same relative risk for all people by the same parameter which is to be estimated from the study. The scientific theory is more vague than a detailed statistical model. Next the researchers make some assumptions to be able to study the scientific theory. Since these assumptions relate the study being done to some external theory that motivated the study, we will call them the *external validity* assumptions. In the PHS, the external validity assumptions

include the assumption that people who are eligible, elect to participate, and sufficiently adhere to a regimen (i.e., those who actually could end up in the study) will be similar to prospective patients to whom we would like to suggest a prolonged low-dose aspirin regimen. Since the PHS is restricted to male physicians aged 40 to 84, an external validity assumption is that this population will tell us something about future prospective patients.

Assumptions related to the statistical hypothesis test should be kept separate from these external validity assumptions. This position was stated in a different way by Kempthorne and Doerfler (1969):

> [T]here are two aspects of experimental inference. The first is to form an opinion about what would happen with repetitions of the experiment with the same experimental units, such repetitions being unrealizable because the experiment 'destroys' the experimental units. The second is to extend this 'inference' to some real population of experimental material that is of interest. (p. 235)

The classic example in which this dichotomy arises is with randomization tests (see e.g., Ludbrook and Dudley, 1998; Mallows, 2000). Ludbrook and Dudley (1998) emphasize the first point of view, arguing that randomization tests are preferred to t or F tests because they not only perform better for small sample sizes, but because "randomization rather than random sampling is the norm in biomedical research". They point out, in keeping with the sentiments of Kempthorne and Doerfler (1969) and others, that the subsequent generalization to a larger population is separate and not statistical: "However, this need not deter experimenters from inferring that their results are applicable to similar patients, animals, tissues, or cells, though their arguments must be verbal rather than statistical. (p. 129)"

External validity assumptions are important, and they must be reasonable, otherwise one may design a very repeatable study which gives not very useful results. In the PHS the external validity assumptions seem reasonable since one would expect that although male physicians are different from the general population (especially females), it is not unreasonable to expect that results from the study would tell us something about non-physicians – at least for males of the same ages. The external validity assumption that allows us to apply the results to females is a bigger one, but this issue is separate from the internal results of the study and whether there was a significant effect for the PHS. In this paper we are emphasizing statistical hypothesis tests, which are tools to make inferences about the study that was performed and cannot tell us anything about the external validity of the study. For example, statistical assumptions have nothing to do with whether we can generalize the results of the PHS to females.

The next step in the process is that the researchers design an ideal study (either an experiment or an observational study) to test their theory. The term ideal refers to the study being done exactly according to the design. In the PHS the ideal study is a double-blind placebo-controlled study on male physicians aged 40 to 84. It is ideal in the sense that all inclusion criteria and randomization are carried out exactly as designed and all instruments are accurate to the specified degree, and the data are recorded correctly, etc. We call the assumptions

needed to treat the actual study as the ideal study, the *study implementation assumptions*.

The study is carried out and data are observed. Then the researchers make some statistical assumptions in order to perform a statistical hypothesis test and calculate related statistics such as p-values and confidence intervals. For the PHS some statistical assumptions were that the individuals were independent, the randomization was a true and fair one, and that the rate of myocardial infarction (MI) events (heart attacks) under the null hypothesis was the same for the group randomized to the aspirin as the group randomized to aspirin-placebo. The PHS showed that the relative risk of aspirin to placebo for MI was significantly lower than 1. If the null hypothesis is rejected, then *if* all the assumptions are correct, chance alone is not a reasonable explanation of the results. What this statistical decision says about the scientific theory depends on all of the assumptions, the statistical assumptions, the study implementation assumptions, and the external validity assumptions.

In this paper we will focus on statistical assumptions. This focus should not be interpreted to imply that the other types of assumptions are not vital for the scientific process. However, often times the non-statistical assumptions are separated enough from the statistical ones that the statistical assumptions may be changed without modifying the non-statistical ones. Here we treat each possible set of statistical assumptions as a different lens through which we can look at the data. Through the lenses of the statistical assumptions and the accompanying DRs, we can see how randomness may play its role in the data. Each DR in our toolbox is associated with many lenses, and for any particular study some of those lenses may be more useful than others. It is even possible that using multiple lenses on the same study may clarify its interpretation. The usefulness of the DR depends on how clearly the lenses of the statistical assumptions help us see reality.

3. Terminology and properties for hypothesis tests and decision rules

Since we are associating a DR with many different hypotheses, we need to be clear about terminology and about what properties of hypothesis tests are associated with the DRs apart from the assumptions about the hypotheses.

Consider a study where we have observed some data, x. In order to perform a hypothesis test we need to make several assumptions. We assume the sample space, $\mathcal{X}$, the (possibly uncountable) set of all possible values of new realizations of the data if the study could be repeated. Let X be an arbitrary member of $\mathcal{X}$.

In the usual hypothesis testing framework, we assume that the data were generated from one of a set of probability models, $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, and we partition that set into two disjoint sets, the null hypothesis, $H = \{P_\theta | \theta \in \Theta_H\}$ and the alternative hypothesis, $K = \{P_\theta | \theta \in \Theta_K\}$, where θ may be an infinite dimensional parameter. Since we will compare many sets of assumptions, we bundle all the assumptions together as $A = (\mathcal{X}, H, K)$. We often assign subscripts to different sets of assumptions.

Let α be the predetermined significance level, and let $\delta(\cdot, \alpha) = \delta$ denote a DR (also called the critical function, see Lehmann and Romano, 2005) which is a function of α and either x or X. The function δ takes on values in $[0, 1]$ representing the probability of rejecting the null hypothesis. In this paper we only consider non-randomized DRs, where $\delta(X, \alpha) \in \{0, 1\}$, for all $X \in \mathcal{X}$. We call the set (δ, A) a hypothesis test. This terminology is a more formal statement of the standard usage for the term 'hypothesis test' where the assumptions are often implied or left unstated. For example, the Wilcoxon-Mann-Whitney (WMW) DR is often used without any explicit statement of what hypotheses are being tested.

Let the power of a DR under P_θ be denoted $Pow[\delta(X, \alpha); \theta] = Pr[\delta(X, \alpha) = 1; \theta]$. A test is a valid test (or an **α-level** test) if for any $0 < \alpha < .5$ the size of the test is less than or equal to α, where the **size** is $\alpha_n^* = \sup_{\theta \in \Theta_H} Pow\left[\delta(X_n, \alpha); \theta\right]$, where $X = X_n$ with n indexing the sample size. There are two types of asymptotic validity (see Lehmann and Romano, 2005, p. 422). A test is pointwise asymptotically valid (PAV) (or pointwise asymptotically level α) if for any $\theta \in \Theta_H$, $\limsup_{n\to\infty} Pow\left[\delta(X_n, \alpha); \theta\right] \leq \alpha$. Note that when a test is PAV, this does not mean that the size of the test will necessarily converge to a value less than α. This latter, more stringent property is the following: a test is uniformly asymptotically valid (UAV) (or uniformly asymptotically level α) if $\limsup_{n\to\infty} \alpha_n^* \leq \alpha$. We give examples of these asymptotic validity properties with respect to two-sample tests in Section 5.2.1.

The classical approach to developing a hypothesis test is to set the assumptions, then choose the decision rule which produces a valid test with the largest power under the alternative. If the null and alternative hypotheses are simple (i.e., represent one probability distribution each) then the Neyman-Pearson fundamental lemma (see Lehmann and Romano, 2005, Section 3.2) provides a method for producing the (possibly randomized) most powerful test. For applications, the hypotheses are often composite (i.e., represent more than one probability distribution), and ideally we desire one decision rule which is most powerful for all $\theta \in \Theta_K$. If such a decision rule exists, the resulting hypothesis test is said to be **uniformly most powerful** (UMP).

For some assumptions, A, there does not exist a UMP test, and extra conditions may be added so that among those tests which meet that added condition, a most powerful test exists. We discuss two such added conditions next, unbiasedness and invariance. A test is **unbiased** (or strictly unbiased) if the power under any alternative model is greater than or equal to (or strictly greater than) the size. A test that is uniformly most powerful among all unbiased tests is called **UMP unbiased**. Often we will wish the hypothesis test to be invariant to certain transformations of the sample space. A general statement of invariance is beyond the scope of this paper (see Lehmann and Romano, 2005, Chapter 6); however, one important case is invariance to monotonic transformations of the responses. A test invariant to monotonic transformations would, for example, give the same results regardless of whether we log transform the responses or not. Tests based on the ranks of the responses are invariant to monotonic transformations. Additionally, we may want to restrict our optimality consideration

to the behavior of the tests close to the boundary between the null and alternative and study the *locally most powerful tests*, which are defined as the UMP tests within a region of the alternative space that is infinitesimally close to the null space (see Hájek and Šidák, 1967, p. 63).

A test is pointwise consistent in power (or simply *consistent*) if the power goes to one for all $\theta \in \Theta_K$. Because many tests may be consistent, in order to differentiate between them using asymptotic power results, we consider a sequence of tests that begin in the alternative space and approach the null hypothesis space. Consider the parametric case where θ is k dimensional. Let $\theta_n = \theta_0 + hn^{1/2}$, where h is a k dimensional constant with $|h| > 0$, $\theta_n \in \Theta_K$ and $\theta_0 \in \Theta_H$. The test (δ, A) is **asymptotically most powerful** (AMP) if it is PAV and for any other sequence of PAV tests, say (δ^*, A), we have $\limsup_{n\to\infty} \{Pow[\delta(X_n, \alpha); \theta_n] - Pow[\delta^*(X_n, \alpha); \theta_n]\} \geq 0$ (see Lehmann and Romano, 2005, p. 541).

For the applied statistician, usually the asymptotic optimality criteria are not as important as the finite sample properties of power. At a minimum we want the power to be larger than α for some $\theta \in \Theta_K$, and practically we want the power to be larger than some large pre-specified level, say $1 - \beta$, for some $\theta \in \Theta_K$.

Now consider some properties of DRs that require only the sample space assumption, $\mathcal{X}$, and need not be interpreted in reference to the rest of the assumptions in A. First we call a DR **monotonic** if for any $0 < \alpha' < \alpha < .5$

$$\delta(X, \alpha') \leq \delta(X, \alpha) \qquad \text{for all } X \in \mathcal{X} \ .$$

Non-monotonic tests have been proposed as a way to increase power for unbiased tests, but Perlman and Wu (1999) (see especially the discussion of McDermott and Wang, 1999) argue against using these tests. Although there are cases where the most powerful test is non-monotonic (see Lehmann and Romano, 2005, p. 96 prob 3.17, p. 105, prob 3.58), non-monotonic tests are rarely if ever needed in applied statistics.

The p-value is

$$p(X) = \inf\{\alpha : \delta(X, \alpha) = 1\},$$

and is a function of the DR and $\mathcal{X}$ only. The validity of the p-value depends on the assumptions. We say a p-value is valid if (see e.g., Berger and Boos, 1994)

$$\sup_{\theta \in \Theta_H} Pr_\theta[p(X) \leq \alpha] \leq \alpha$$

We will say a hypothesis test is valid if the p-value is valid. For non-randomized monotonic DRs, a valid hypothesis test (i.e., valid p-value) implies an α-level hypothesis test, and in this paper we will use the terms interchangeably. Most reasonable tests are at least PAV, asymptotically strictly unbiased, and monotonic.

4. Multiple perspective decision rules

We call the set $(\delta, A_1, \ldots, A_k)$ a multiple perspective DR, with each of the hypothesis tests, $(\delta, A_i), i = 1, \ldots, k$ called a perspective. We only consider perspectives which are either valid or at least PAV.

We say that A_i is *more restrictive* than A_j if $\mathcal{X}_i \subseteq \mathcal{X}_j$, $H_i \subseteq H_j$, $K_i \subseteq K_j$, and either $H_i \cap H_j \neq H_j$ or $K_i \cap K_j \neq K_j$, and we denote this as $A_i \sqsubset A_j$. In other words, if both (δ, A_i) and (δ, A_j) are valid tests and $A_i \sqsubset A_j$ then (δ, A_j) is a less parametric test than (δ, A_i).

We state some simple properties of MPDR which are obvious by inspection of the definitions. If $A_i \sqsubset A_j$ then:

- if (δ, A_j) is valid then (δ, A_i) is valid (and if (δ, A_i) is invalid then (δ, A_j) is invalid),
- if (δ, A_j) is PAV then (δ, A_i) is PAV,
- if (δ, A_j) is UAV then (δ, A_i) is UAV,
- if (δ, A_j) is unbiased then (δ, A_i) is unbiased, and
- if (δ, A_j) is consistent then (δ, A_i) is consistent.

Note that the statements about validity only require $H_i \subseteq H_j$.

We briefly mention broad types of perspectives. First, an *optimal* perspective shows how the DR has some optimal property (e.g., uniformly most powerful test) under that perspective. Second, a consistent perspective delineates an alternative space whereby for each $P_\theta \in K$, the DR has asymptotic power going to one. Sometimes the full probability space for consistent perspectives (i.e., $\mathcal{P} = H \cup K$) is not a "natural" probability space in that it is hard to justify the assumption that the true model exists within the full probability space but not within some smaller subset of that space. This vague concept will become more clear through the examples given in Section 5. In cases where the full probability space is not natural in this sense, we call the perspective a *focusing* perspective, since it focuses the attention on certain alternatives. We now consider some real examples to clarify these broad types of perspectives and show the usefulness of the MPDR framework.

5. MPDRs for two-sample tests of central tendency

Consider the case where the researcher wants to know if group A has generally larger responses than group B. The researcher may think the choice between the Wilcoxon rank sum/Mann-Whitney U test (WMW test) and the t-test depends on the results of a test of normality (see e.g., Figure 8.5 of Dowdy, Weardon and Chilko, 2004). In fact, the issue is not so simple. In this section we explore these two DRs and the choice between them and some related DRs.

Let the data be $x = x_n = \{y_1, y_2, \ldots, y_n, z_1, z_2 \ldots, z_n\}$ where y_i represents the ordinal (either discrete or continuous) response for the ith individual, and z_i is either 0 (for n_0 individuals) or 1 (for $n_1 = n - n_0$ individuals) representing either of the two groups. Let $X = X_n = \{Y_1, Y_2, \ldots, Y_n, Z_1, Z_2 \ldots, Z_n\}$ be

another possible realization of the experiment. Throughout this section, for all asymptotic results we will assume that $n_0/n \to \lambda_0$, where $0 < \lambda_0 < 1$.

For all of the perspectives discussed in the following except Perspective 9, we assume the sample space is

$$\mathcal{X} = \{z, Y : Y \in \mathcal{Y}\},$$

where $\mathcal{Y}$ is a set of possible values of Y if the experiment were repeated. The z vector does not change for this sample space. Further, for all perspectives except Perspective 9, we assume that the Y_i are independent with $Y_i \sim F$ if $Z_i = 1$ and $Y_i \sim G$ if $Z_i = 0$. Failure of the independence assumption can have a large effect on the validity of some if not all the hypothesis tests to be mentioned. This problem exists for all kinds of distributions, but incidental correlation is not a serious problem for randomized trials (see Box, Hunter and Hunter, 2005, Table 3A.2, p. 118, and Proschan and Follmann, 2008).

5.1. Wilcoxon-Mann-Whitney and related decision rules

5.1.1. Valid perspectives for the WMW decision rule

Wilcoxon (1945) proposed the exact WMW DR allowing for ties presenting the test as a permutation test on the sum of the ranks in one of the two groups. Let δ_W be this exact DR, which can be calculated using network algorithms (see e.g., Mehta, Patel and Tsiatis, 1984). Wilcoxon (1945) does not explicitly give hypotheses which are being tested but talks about comparing the differences in means. We do not have a valid test if we only assume that the means of F and G are equal, since difference in variances can also cause the test statistic to be significant. Thus, for our first perspective we assume $F = G$ under the null to ensure validity, and make no assumptions about the discreteness or the continuousness of the responses:

Perspective 1. *Difference in Means, Same Null Distributions*

$$\begin{aligned} H_1 &= \{F, G : F = G\} \\ K_1 &= \{F, G : E_F(Y) \neq E_G(Y)\} \end{aligned}$$

This is a focusing perspective, since $\mathcal{P} = H \cup K$ is hard to justify in an applied situation because $\mathcal{P}$ is the strange set of all distributions F and G except those that have equal means but are not equal. This is not a consistent perspective.

Mann and Whitney (1947) assumed continuous responses and tested for stochastic ordering. Letting Ψ_C be the set of continuous distributions, the perspective is:

Perspective 2. *Stochastic Ordering:*

$$\begin{aligned} H_2 &= \{F, G : F = G; F \in \Psi_C\} \\ K_2 &= \{F, G : F <_{st} G \text{ or } G <_{st} F; F, G \in \Psi_C\} \end{aligned}$$

where $F <_{st} G$ denotes that G is stochastically larger than F, which is equivalent to $G(y) \leq F(y)$ for all y with strict inequality for at least some y.

Mann and Whitney (1947) showed the consistency of the WMW DR under the Stochastic Ordering (SO) perspective. Lehmann (1951) shows the unbiasedness under the SO perspective, and his result holds even without the continuity assumption. Lehmann (1951) also notes that the Mann and Whitney (1947) consistency proof shows the consistency for all alternatives under the following perspective:

Perspective 3. *Mann-Whitney Functional (continuous, equal null distributions):*

$$\begin{aligned} H_3 &= \{F, G : F = G; F \in \Psi_C\} \\ K_3 &= \left\{F, G : \phi(F, G) \neq \frac{1}{2}; F, G \in \Psi_C\right\} \end{aligned}$$

where ϕ is called the Mann-Whitney functional, defined (to make sense for discrete data as well) as

$$\begin{aligned} \phi(F, G) &= \frac{1}{2}\left(1 + \int G(y)dF(y) - \int F(y)dG(y)\right) \\ &= Pr[Y_F > Y_G] + \frac{1}{2}Pr[Y_F = Y_G] \end{aligned}$$

where $Y_F \sim F$ and $Y_G \sim G$.

Similar to Perspective 1, the Mann-Whitney functional perspective is a focusing one since the full probability set, $\mathcal{P}$, is created more for mathematical necessity than by any scientific justification for modeling the data, which in this case does not include distributions with both $\phi(F, G) = 1/2$ and $F \neq G$. It is hard to imagine a situation where this complete set of allowable models, $\mathcal{P}$, and only that set of models is justified scientifically; the definition K_3 acts more to focus on where the WMW procedure is consistent. A more realistic perspective in terms of $\mathcal{P} = H \cup K$ is the following:

Perspective 4. *Distributions Equal or Not*

$$\begin{aligned} H_4 &= \{F = G\} \\ K_4 &= \{F \neq G\} \end{aligned}$$

Later we will introduce another realistic perspective (Perspective 10), for which the WMW procedure is invalid, but first we list a few more perspectives for which δ_W is valid.

Under the following optimal perspective, δ_W is the locally most powerful rank test (see Hettmansperger, 1984, section 3.3) as well as an AMP test (van der Vaart, 1998, p. 225):

Perspective 5. *Shift in logistic distribution*

$$\begin{aligned} H_5 &= \{F, G : F = G; F \in \Psi_L\} \\ K_5 &= \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_L\} \end{aligned}$$

where Ψ_L is the set of logistic distributions.

Hodges and Lehmann (1963) showed how to invert the WMW DR to create a confidence interval for the shift in location under the following perspective:

Perspective 6. *Location Shift (continuous)*

$$\begin{aligned} H_6 &= \{F, G : F = G; F \in \Psi_C\} \\ K_6 &= \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_C\} \end{aligned}$$

Sometimes we observe responses, say Y_i^*, where $Y_i^* \in (0, \infty)$ and there are extreme right tails in the distribution. A general case is the gamma distribution, which has chi squared and exponential distributions as special cases. For the gamma distribution a shift in location does not conserve the support of the distribution (i.e., $\mathcal{Y}$ changes with location shifts), and a better model is a scale change. This is equivalent to a location shift after taking the log transformation. Let $Y_i = \log_e(Y_i^*)$; then the scale change for the random variables Y_i^* is equivalent to the location shift for Y_i, which has a log-gamma distribution.

Perspective 7. *Shift in log-gamma distribution*

$$\begin{aligned} H_7 &= \{F, G : F = G; F \in \Psi_{LG}\} \\ K_7 &= \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_{LG}\} \end{aligned}$$

where Ψ_{LG} is the set of log gamma distributions.

Consider a perspective useful for discrete responses. Assume there exists a latent unobserved continuous variable for the ith observation, and denote it here as Y_i^*. Let F^* and G^* be the associated distributions for each group. All we observe is some coarsening of that variable, $Y_i = c(Y_i^*)$, where $c(\cdot)$ is an unknown non-decreasing function that takes the continuous responses and assigns them into k ordered categories (without loss of generality let the sample space for Y_i be $\{1, \ldots, k\}$). We assume that for $j = 1, \ldots, k$, $c(Y^*) = j$ if $\xi_{j-1} < Y^* \leq \xi_j$ for some unknown cutpoints $-\infty \equiv \xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k \equiv \infty$. Then $F^*(\xi_j) = F(j)$. Then a perspective that allows easy interpretation of this type of data is the proportional odds model, where

$$\frac{G^*(y^*)}{1 - G^*(y^*)} = \left(\frac{F^*(y^*)}{1 - F^*(y^*)} \right) \Delta^* \text{ for all } y^* \text{ with } \Delta^* > 0.$$

The observed data then also follows a proportional odds model regardless of the unknown values of ξ_j:

Perspective 8. *Proportional Odds*

$$\begin{aligned} H_8 &= \{F, G : F = G; F \in \Psi_{D_k}\} \\ K_8 &= \left\{F, G : \frac{G(y)}{1-G(y)} = \left(\frac{F(y)}{1-F(y)}\right)\Delta \text{ for all } y; \Delta \neq 1; F \in \Psi_{D_k}\right\} \end{aligned}$$

where Ψ_{D_k} *is a set of discrete distributions with sample space* $1, \ldots, k$.

Under the proportional odds model the WMW test is the permutation test on the efficient score, which is the locally most powerful similar test for one-sided alternatives (McCullagh, 1980, p. 117). Note that if we could observe Y_i^* then the logistic shift perspective on Y_i^* follows the proportional odds model, since the shifting of a logistic distribution by Δ_L, say, is equivalent to multiplying the odds by $\exp(\Delta_L)$.

All of the above perspectives use a *population model*, i.e., postulate a distribution for the $Y_i, i = 1, \ldots, n$. Another model is the *randomization model* that assumes that each time the experiment is repeated the responses $y_i, i = 1, \ldots, n$ will be the same, but the group assignments $Z_i, i = 1, \ldots, n$ may change. Notice how this perspective is very different from the other perspectives, which are all different types of population models (see Lehmann, 1975, for a comparison of the randomization and population models):

Perspective 9. *Randomization Model*

$$\mathcal{X} = \{y, Z : Z \in \Pi(z)\},$$

where $\Pi(z)$ *is the set of all permutations of the ordering of* z, *which has* $N = n!/(n_1!(n-n_1)!)$ *unique elements. Let* $\Pi(z) = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_N\}$.

$$H_9 = \{Pr[Z = \mathcal{Z}_a] = N^{-1} \text{ for all } a\}$$

The randomization model does not explicitly state a set of alternative probability models for the data, i.e., it is a *pure* significance test (see Cox and Hinkley, 1974, Chapter 3), although it is possible to define the alternative as any probability model not in H_9.

5.1.2. An invalid perspective and some modified decision rules

Here is one perspective for which the WMW procedure is not valid:

Perspective 10. *Mann-Whitney Functional (Different null distributions):*

$$\begin{aligned} H_{10} &= \left\{F, G : \phi(F, G) = \frac{1}{2}; Var(F(Y_G)) > 0, Var(G(Y_F)) > 0\right\} \\ K_{10} &= \left\{F, G : \phi(F, G) \neq \frac{1}{2}; Var(F(Y_G)) > 0, Var(G(Y_F)) > 0\right\}. \end{aligned}$$

Note that the conditions $Var(F(Y_G)) > 0$ and $Var(G(Y_F)) > 0$ are not very restrictive, allowing both discrete or continuous distributions. This perspective

has also been called the nonparametric Behrens-Fisher problem (Brunner and Munzel, 2000), and under this perspective δ_W is invalid (see e.g., Pratt, 1964). Brunner and Munzel (2000) gave a variance estimator for $\phi(\hat{F},\hat{G})$ which allows for ties, where $\hat{F}$ and $\hat{G}$ are the empirical distributions. Since for continuous data the variance estimator can be shown equivalent to Sen's jackknife estimator (Sen, 1967, see e.g., Mee, 1990), we denote this V_J. Brunner and Munzel (2000) showed that comparing

$$T_{NBF} = \frac{\phi(\hat{F},\hat{G}) - \frac{1}{2}}{\sqrt{V_J}}$$

to a t-distribution with degrees of freedom estimated using a method similar to Satterthwaite's gives a PAV test under Perspective 10. We give a slight modification of that DR in Appendix A, and denote it as δ_{NBFa}.

For the continuous random variable case, Janssen (1999) showed that if we perform a permutation test using a statistic equivalent to T_{NBF} as the test statistic, then the test is PAV under Perspective 10 with the added condition of $F, G \in \Psi_C$. Neubert and Brunner (2007) generalized this to allow for ties, i.e., they created a permutation test based on Brunner and Munzel's T_{NBF}. We denote that DR as δ_{NBFp}. Since it is a permutation test, δ_{NBFp} is valid for finite samples for perspectives where $F = G$.

5.2. Decision rules for two-sample difference in means (t-tests)

5.2.1. Some valid t-tests

To talk about DRs for t-tests we introduce added notation. Let the sample means and variances for the first group be $\hat{\mu}_1 = n_1^{-1}\sum_{i=1}^n Z_i Y_i$ and $\hat{\sigma}_1^2 = (n_1 - 1)^{-1}\sum_{i=1}^n Z_i(Y_i - \hat{\mu}_1)^2$ and similarly define the sample mean $(\hat{\mu}_0)$ and variance $(\hat{\sigma}_0^2)$ for the second group. Let $\hat{\sigma}_t^2 = (n-2)^{-1}\left((n_1-1)\hat{\sigma}_1^2 + (n_0-1)\hat{\sigma}_0^2\right)$ be the pooled sample variance used in the usual t-test DR, say δ_t. Let

$$T_t(X) = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_t\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}},$$

so the DR is

$$\delta_t(X) = \begin{cases} 1 & \text{if } |T_t(X)| > t_{n-2}^{-1}(1-\alpha/2) \\ 0 & \text{otherwise} \end{cases}$$

where $t_{n-2}^{-1}(q)$ is the qth quantile of the t-distribution with $n-2$ degrees of freedom. The standard perspective for this DR is:

Perspective 11. *Shift in Normal Distribution*

$$\begin{aligned} H_{11} &= \{F, G : F = G; F \in \Psi_N\} \\ K_{11} &= \{F, G : G(y) = F(y + \Delta); \Delta \neq 0; F \in \Psi_N\} \end{aligned}$$

where Ψ_N is the set of normal distributions.

Under this perspective, the associated test, (δ_t, A_{11}), is the uniformly most powerful unbiased test and can be shown to be asymptotically most powerful (see Lehmann and Romano, 2005, Section 5.3 and Chapter 13). Under the following perspective, the standard t-test is PAV (see Lehmann and Romano, 2005, p. 446):

Perspective 12. *Difference in Means, (Finite Variances, Equal Null distributions)*

$$\begin{aligned} H_{12} &= \{F, G : F = G; F \in \Psi_{fv}\} \\ K_{12} &= \{F, G : E(Y_G) \neq E(Y_F), F, G \in \Psi_{fv}\} \end{aligned}$$

where Ψ_{fv} is the class of distributions with finite variances.

Note that the fact that a test is PAV does not guarantee that for sufficiently large n, the size of the test is close to α; for this we require UAV. In fact, (δ_t, A_{12}) is pointwise but not uniformly asymptotically valid (PNUAV). To see this, take $F_n = G_n$ to be Bernoulli(p_n) such that $np_n \to \lambda$. The numerator of the t-statistic behaves like a difference of Poissons, and the type I error rate can be inflated (see Appendix B for details).

We show in Appendix C that if we restrict the distributions to the following perspective then the t-test is UAV:

Perspective 13. *Difference in Means, (Variance $\geq \epsilon > 0$ and $E(Y^4) \leq B$ with $0 < B < \infty$, Equal Null distributions)*

$$\begin{aligned} H_{12} &= \{F, G : F = G; F \in \Psi_{B\epsilon}\} \\ K_{12} &= \{F, G : E(Y_G) \neq E(Y_F), F, G \in \Psi_{B\epsilon}\} \end{aligned}$$

where $\Psi_{B\epsilon}$ is the class of distributions with $Var(Y) \geq \epsilon > 0$ and $E(Y^4) \leq B$, with $0 < B < \infty$.

Using different methods, Cao (2007) appears to show that only finite third moments are needed to show that the t-test is UAV.

Consider the exact permutation test for the difference in means. This DR is defined by enumerating all permutations of indices of the z_i, recalculating $\hat{\mu}_1 - \hat{\mu}_0$ for each permutation, and using the permutation distribution of that statistic to define the DR. Denote this DR as δ_{tp}. It is equivalent to the permutation test on the standard t-test (Lehmann and Romano, 2005, p. 180). Note that although δ_{tp} is asymptotically equivalent to δ_t under H_{fv} (Lehmann and Romano, 2005, p. 642), δ_{tp} is valid (and hence UAV) while δ_t is not UAV. This issue is discussed in Appendix D. Note that because δ_{tp} is a permutation test, it is valid for any n for any perspective for which $F = G$ under the null.

Now consider a perspective where neither δ_t nor δ_{tp} is valid:

Perspective 14. *Behrens-Fisher: Difference in Normal Means, Different Variances*

$$\begin{aligned} H_{14} &= \{F, G : E_F(Y) = E_G(Y); F, G \in \Psi_N\} \\ K_{14} &= \{F, G : E_F(Y) \neq E_G(Y); F, G \in \Psi_N\} \end{aligned}$$

For this perspective, Welch's modification to the t-test is often used. Let the Behrens-Fisher statistic (see e.g., Dudewicz and Mishra, 1988, p. 501) be

$$T_{BF}(X) = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}}.$$

Welch's DR is

$$\delta_{tW}(X) = \begin{cases} 1 & \text{if } |T_{BF}(X)| > t_{d_W}^{-1}(1-\alpha/2) \\ 0 & \text{otherwise} \end{cases},$$

where

$$d_W = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}\right)^2}{\frac{\left(\hat{\sigma}_1^2/n_1\right)^2}{n_1-1} + \frac{\left(\hat{\sigma}_0^2/n_0\right)^2}{n_0-1}}. \tag{5.1}$$

Welch's DR, δ_{tW}, is approximately valid for the Behrens-Fisher perspective. If we replace d_W with $\min(n_1, n_0)-1$ then the associated DR is Hsu's, δ_{tH}, which is a valid test under the Behrens-Fisher perspective (see e.g., Dudewicz and Mishra, 1988, p. 502). Both (δ_{tW}, A_{14}) and (δ_{tH}, A_{14}) are asymptotically most powerful tests (Lehmann and Romano, 2005, p. 558).

Finally, as with the rank test, we may permute T_{BF} to obtain a PAV test under Perspective 14 and a valid test whenever $F = G$; we denote this statistic δ_{BFp}. Janssen (1997) showed that δ_{BFp} is PAV under less restrictive assumptions than Perspective 14, replacing Ψ_N with the set of distributions with finite means and variances.

5.2.2. An invalid perspective

Consider the seemingly natural perspective:

Perspective 15. *Difference in Means, Different Null Distributions*

$$\begin{aligned} H_{15} &= \{F, G : E_F(Y) = E_G(Y)\} \\ K_{15} &= \{F, G : E_F(Y) \neq E_G(Y)\} \end{aligned}$$

For this perspective, for any DR that has a power of $1-\beta > \alpha$ for some probability model in the alternative space, the size of the resulting hypothesis test is at least $1-\beta$. To show this, take any two distributions, F^* and G^*, for which the DR has a probability of rejection equal to $1-\beta$. Now create F and G as mixture distributions, where $F = (1-\epsilon)F^* + \epsilon F^\epsilon$, $G = (1-\epsilon)G^* + \epsilon G^\epsilon$ and F^ϵ and G^ϵ are distributions which depend on ϵ that can be chosen to make $E_F(Y) = E_G(Y)$ for all ϵ. More explicitly, let $\mu(\cdot)$ be the mean functional, so that the mean of the distribution F is $\mu(F)$. Then for any constant u, we can create a F^ϵ which meets the above condition as long as $\mu(F^\epsilon) = \epsilon^{-1}\{u - (1-\epsilon)\mu(F^*)\}$, and similarly for $\mu(G^\epsilon)$. For any fixed n we can make ϵ close enough to zero, so

that the the power of the test under this probability model in the null hypothesis space approaches $1-\beta$, and the resulting test is not an $\alpha-$level test. Note that the above result holds even if we restrict the perspective so that F and G have finite variances. See Lehmann and Romano (2005), Section 11.4, for similar ideas but which focuses mostly on the one-sample case.

5.3. Comparing assumptions and decision rules

5.3.1. Relationships between assumptions

In Figure 1 we depict the cases where we can say that one set of assumptions is more restrictive than another. Most of these relationships are immediately apparent from the definitions; however, the relationships related to stochastic ordering (assumptions A_2) are not obvious. If G is stochastically larger than F (i.e., $F <_{st} G$) then for all nondecreasing real-valued functions h, $E(h(Y_F)) < E(h(Y_G))$ (see e.g., Whitt, 1988). Thus, stochastic ordering implies both ordering of the means (letting h be the identity function) and ordering of the Mann-Whitney function for continuous data (letting $h = F$), so that $A_2 \sqsubset A_1$ and $A_2 \sqsubset A_3$. We can use Figure 1 to show validity and consistency of DRs under various perspectives.

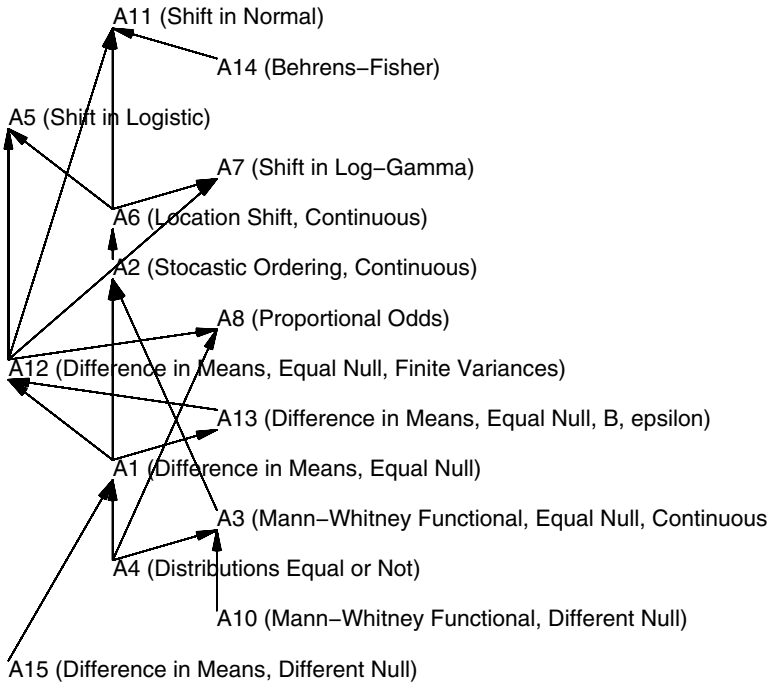

FIG 1. *Relationship between assumptions. $A_i \leftarrow A_j$ denotes that $A_i \sqsubset A_j$ (i.e., A_i are more restrictive assumptions than A_j).*

TABLE 1
Validity and Consistency of Two Sample MPDRs

		Decision Rules							
		WMW	=Wilcoxon-Mann-Whitney (exact)						
		NBF_a	=Nonparametric Behrens-Fisher (asymptotic)						
		NBF_p	=Nonparametric Behrens-Fisher (permutation)						
		t	=t-test (pooled variance)						
		t_W	=Welch's t-test (Satterthwaite's df)						
		t_H	=Hsu's t-test (df=min(n_i)-1)						
		t_p	=permutation t-test						
		t_{BFp}	=permutation test on Behrens-Fisher statistic						
	Perspective	WMW	NBF_a	NBF_p	t	t_W	t_H	t_p	t_{BFp}
11.	Normal Shift	yy	uy	yy	yy	uy	yy	yy	yy
14.	Behrens-Fisher	n-	ay	ay	n-	uy	yy	n-	ay
5.	Shift in Logistic	yy	uy	yy	ay	ay	ay	yy	yy
7.	Shift in Log-Gamma	yy	uy	yy	ay	ay	ay	yy	yy
6*.	Location Shift, fv	yy	uy	yy	ay	ay	ay	yy	yy
2*.	Stochastic Ordering, SN, fv	yy	uy	yy	ay	ay	ay	yy	yy
8.	Proportional Odds, SN	yy	uy	yy	ay	ay	ay	yy	yy
12.	Diff in Means, SN,fv	yn	un	yn	py	py	py	yy	yy
13.	Diff in Means, SN, $B\epsilon$	yn	un	yn	uy	uy	uy	yy	yy
3*.	Mann-Whitney Func., SN, fv	yy	uy	yy	an	an	an	yn	yn
4*.	Distributions Equal or Not, fv	yn	un	yn	an	an	an	yn	yn
15*.	Diff in Means, DN, fv	n-	n-	n-	n-	n-	n-	n-	n-
10*.	Mann-Whitney Func., DN, fv	n-	ay	ay	n-	n-	n-	n-	n-
9.	Randomization Model	y-	- -	y-	- -	- -	- -	y-	y-

Perspective numbers with * have the additional assumption that $F, G \in \Psi_{fv}$ in both H and K.
SN=Same Null Distns., DN=Different Null Distns., fv=Finite Var.,
$B\epsilon = \left\{E(Y^4) \leq B \text{ and } Var(Y) \geq \epsilon\right\}$

Each hypothesis test is represented by 2 sets of symbols representing the 2 properties:
(i) validity, and (ii) (pointwise) consistency, where each character answers the question,
This test has this property: y=yes, n=no, and $-$ = not applicable.
For validity we also have the symbols: u=UAV, a = PAV, p =PNUAV.

5.3.2. Validity and consistency

Table 1 summarizes the validity and consistency of the DRs introduced under different perspectives. For this table we assume finite variances for F and G (so Perspectives 6, 2, 3, 4, and 10 have this additional restriction), although both validity and consistency results hold for the rank tests without this additional assumption. The first symbol for each test answers the question "Is this test valid?" with either: y=yes (for all n), u=UAV, a = PAV, p = pointwise but not uniformly asymptotically valid (PNUAV), n=no (not even asymptotically). The symbol a for PAV, denotes an unsolved answer to the question of whether the perspective is UAV or PNUAV. For Perspective 9 there is no probability model for the responses so only permutation based DRs will be valid. Others will be marked as '-' for undefined. Justifications for the validity symbols of Table 1 not previously discussed are given in Appendix D.1.

The second symbol in Table 1 denotes consistency: y=yes or n=no, but will only be given for tests that are at least PAV, otherwise we use a '-' symbol. Justification for the consistency results is given in Appendix D.2.

Besides the asymptotic results, there are many papers which simulate the size of the t-test for different situations. For example, for many practical situations when $F = G$ (e.g., lumpy multimodal distributions, and distributions with digital preference), Sawilowsky and Blair (1992) show by simulation that the t-test is approximately valid for a range of finite samples. These simulations agree with the above.

5.3.3. Power and small sample sizes

We consider first the minimum sample size needed to have any possibility of rejecting the null. For simplicity we deal only with the case where there are equal numbers in both groups. It is straightforward to show that when $\alpha = 0.05$, the sample size should be at least 4 per group (i.e., $n = 8$) in order for the WMW exact or the permutation t procedures to have a possibility of rejecting the null. In contrast, we only need 2 per group (i.e., $n = 4$) for Student's t and Welch's t. There are subtle issues on using tests with such small sample sizes. For example, a t-test with only 2 people per treatment group could be highly statistically significant, but the 2 treated patients might have been male and the two controls female so that sex may have explained the observed difference.

5.3.4. Relative efficiency of WMW vs. t-test

Since the permutation t-test is valid under the fairly loose assumptions of Perspectives 4 or 9, some might argue that it is preferred over the WMW because the WMW uses ranks which is like throwing some data away. For example, Edgington (1995), p. 85, states:

> When more precise measurements are available, it is unwise to degrade the precision by transforming the measurements into ranks for conducting a statistical test. This transformation has sometimes been made to permit the use of a nonparametric test because of the doubtful validity of the parametric test, but it is unnecessary for that purpose since a randomization test provides a significance value whose validity is independent of parametric assumptions, while using the raw data.

Although this position appears to make sense on the surface, it is misleading because there are many situations where the WMW test has more power and is more efficient. For example, for distributions with heavy tails or very skewed distributions, we can get better power by using the WMW procedure rather than the t procedure. Previously, Blair and Higgins (1980) carried out some extensive simulations showing that in most of the situations studied, the WMW is more powerful than the t-test. Here we just present two classes of distributions: for skewed distributions we will consider the location tests on the log-gamma distribution (equivalent to scale tests on the gamma distribution) and for heavy tails we consider location tests on the t-distribution.

Consider first the location shift on the log gamma distribution (Perspective 7). Here is the pdf of the log transformed gamma distribution with scale=1

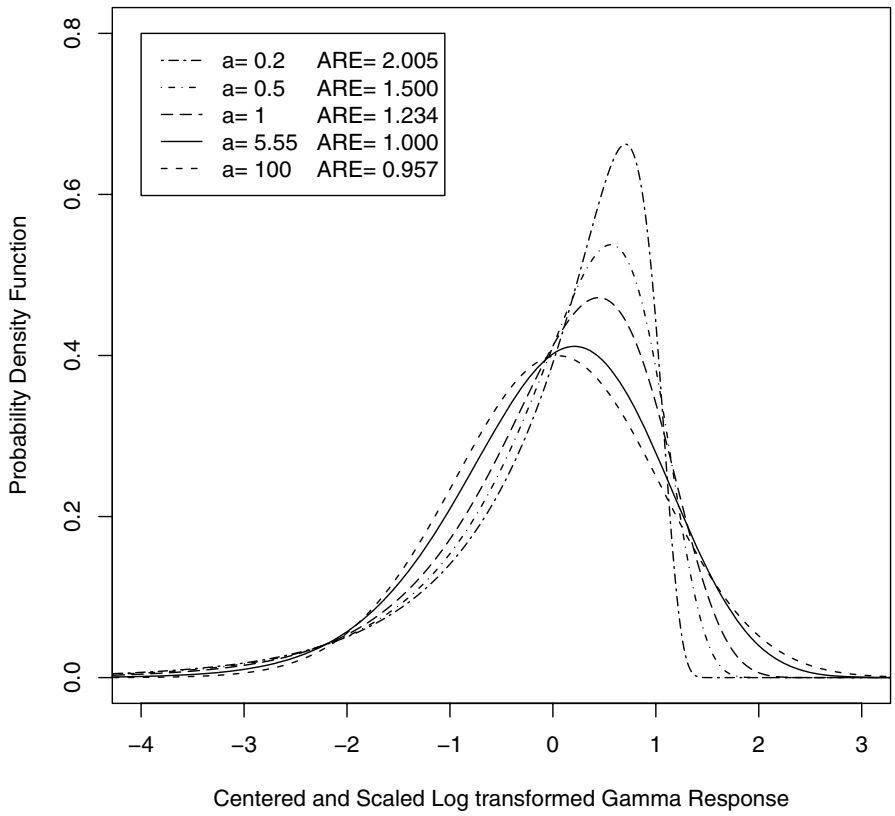

FIG 2. *The probability density functions for some log transformed gamma distributions. All distributions are scaled and shifted to have mean 0 and variance 1. The value a is the shape parameter, and ARE is asymptotic relative efficiency. An ARE of 2 denotes that it will take twice as many observations to obtain the same asymptotic power for the t-test compared to the WMW-test.*

and shape=a.

$$f(y) = \frac{1}{\Gamma(a)} \exp\left(ay - e^y\right) \tag{5.2}$$

By changing the a parameter we can change the extent of the skewness. We plot some probability density functions standardized to have a mean 0 and variance 1 in Figure 2. Although all the tails are more heavy on the left, the results below would be identical if we had used $-y$ so that the tails are skewed to the right. In the legend, we give the asymptotic relative efficiency (ARE) of the WMW compared to the t-test for different distributions. This ARE is given by (see e.g., Lehmann, 1999, p. 176)

$$ARE = 12\sigma^2 \left(\int f^2(y) dy \right)^2$$

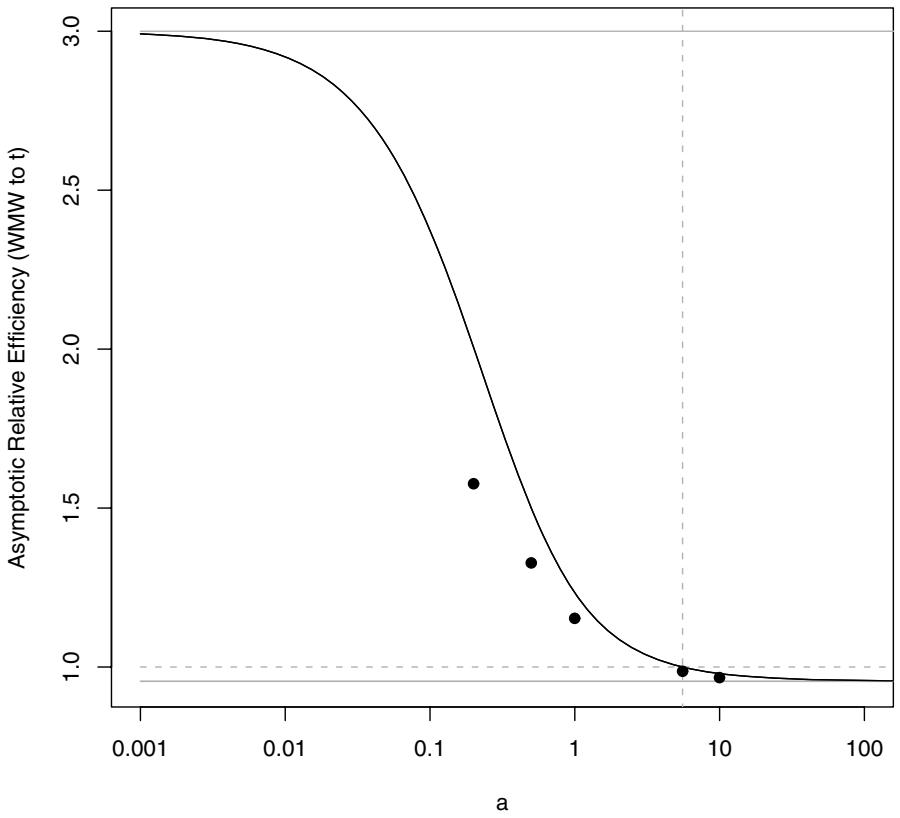

FIG 3. *Relative efficiency of WMW test to t-test for testing for a location shift in log-gamma distribution. The value a is the shape parameter. The solid black line is the ARE. The dotted grey horizontal line is at 1, and is where both tests are equally asymptotically efficient, which occurs at the dotted grey vertical line at* $a = 5.55$. *The solid grey horizontal lines are at* 3 *and* $3/\pi = .955$, *which are the limits as* $a \to 0$ *and* $a \to \infty$. *Points are simulated relative efficiency for shifts which give about 80% power for the WMW DR when there are about 20 in each group.*

where σ^2 is the variance associated with the distribution $f(y)$. Now using f from equation 5.2 we get

$$ARE = \frac{12\psi'(a)\Gamma^2(2a)}{\Gamma(a)2^{4a}}$$

where $\psi'(a)$ is the trigamma function, $\psi'(a) = \frac{\partial^2\{\log(\Gamma(a))\}}{\partial a}$. We plot the ARE for different values of a in Figure 3.

To show that these AREs work well enough for finite samples, we plot additionally the simulated relative efficiency for several values of a where the associated shift for each a is chosen to give about 80% power for a WMW DR with sample sizes of 20 in each group. The simulated relative efficiency (SRE) is the ratio of the expected sample size for the t (pooled variance) over that for the

exact WMW, where the expected sample size for each test randomizes between the sample size, say n, that gives power higher than .80 and $n-1$ that gives power lower than .80 such that the expected power is .80 (see Lehmann, 1999, p. 178). The powers are estimated by a local linear kernel smoother on a series of simulations at different sample sizes (with up to 10^5 replications for sample sizes close to the power of .80).

Note that from Figure 2 the distribution where the ARE=1 looks almost symmetric. Thus, histograms for moderate sample sizes that look symmetric may still have some small indiscernible asymmetry which causes the WMW DR to be more powerful.

Now suppose the underlying data have a t-distribution, which highlights the heavy tailed case. The ARE of the WMW test to the t-test when the distribution is t with d degrees of freedom ($d > 2$) is

$$ARE = \frac{12\Gamma^4\left(\frac{d+1}{2}\right)\Gamma^2\left(\frac{2d+1}{2}\right)}{\pi(d-2)\Gamma^4\left(\frac{d}{2}\right)\Gamma^2(d+1)} \tag{5.3}$$

We can show that as d approaches 2 the ARE approaches infinity, and since the t-distribution approaches the normal distribution as d gets large, $\lim_{d\to\infty} ARE = 3/\pi$. From equation 5.3, if $d \leq 18$ then the WMW test is more efficient, while if $d \geq 19$ then the t-test is more efficient. We plot the scaled t-distribution with $d = 18$ (scaled to have variance equal to 1), and the standard normal distribution in Figure 4. In Figure 4a, we can barely see that the tails of the t-distribution are larger than that of the normal distribution, but on the log-scale (Figure 4b) we can see the larger tails. Further, there is a distribution for which the ARE is minimized (see e.g., Lehmann, 1975, p. 377) at $108/125 = .864$, given by $f(Y) = \frac{3}{20\sqrt{5}}(5-y^2)$ for $y \in (-\sqrt{5}, \sqrt{5})$. We plot this distribution in Figure 4 as well.

In Figure 5 we plot the ARE and simulated relative efficiency for the case where 20 in each group give a power of about 80% for the WMW DR. Note that the ARE gives a fairly good picture of the efficiency even for small samples (although when $d \to 2$ and the ARE $\to \infty$, the SRE at $d = 2$ is only 2.3).

In Figure 4 we see that the distribution with ARE=1 looks very similar to a normal distribution. If the tails of the distribution are much less heavy than the t with 18 degrees of freedom, then the t-test is recommended. This matches with the simulation of Blair and Higgins (1980) who showed that the uniform distribution had slightly better power for the t procedure than for the WMW.

5.3.5. Robustness

Robustness is a very general term that is used in many ways in statistics. Some traditional ways the term is used we have already discussed. We have seen that the classical t-test, i.e., (δ_t, A_{11}), is a UMP unbiased test, yet δ_t retains asymptotic validity (specifically UAV) when the normality assumption does not hold (i.e., it is asymptotically robust to the normality assumption), and all we

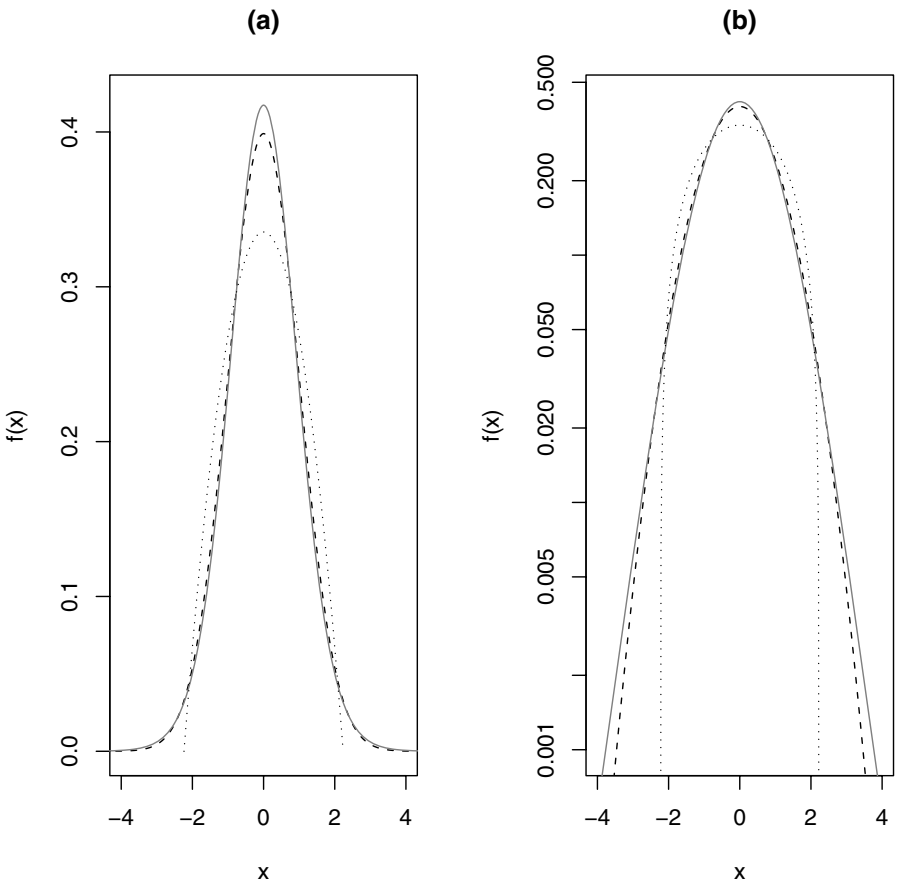

FIG 4. *Standard normal distribution (black dashed) and scaled t-distribution with 18 degrees of freedom (grey solid), and the distribution with the minimum ARE (black dotted), where all distributions have mean 0 and variance 1. The plots are the same except the right plot (b), has the $f(x)$ plotted on the log scale to be able to see the difference in the extremities of the tails.*

require for this UAV is second and fourth moment bounds on the distributions (see Perspective 13 and Table 1). Similarly we can say that although the test given by (δ_W, A_6) is AMP, the validity of that test is robust to violations of the logistic distribution assumption (see Table 1).

We have addressed robustness of efficiency indirectly by focusing on the efficiency comparisons between the t-test and the WMW test with respect to location shift models as discussed in Section 5.3.4. We can make statements about the robustness of efficiency of the WMW test from those results. Since we know that (δ_t, A_{11}) is asymptotically most powerful, and we know that the t-distribution approaches the normal distribution as $n \to \infty$, we can say that for large n the WMW test retains 95.5% efficiency compared to the AMP test against the normal shift.

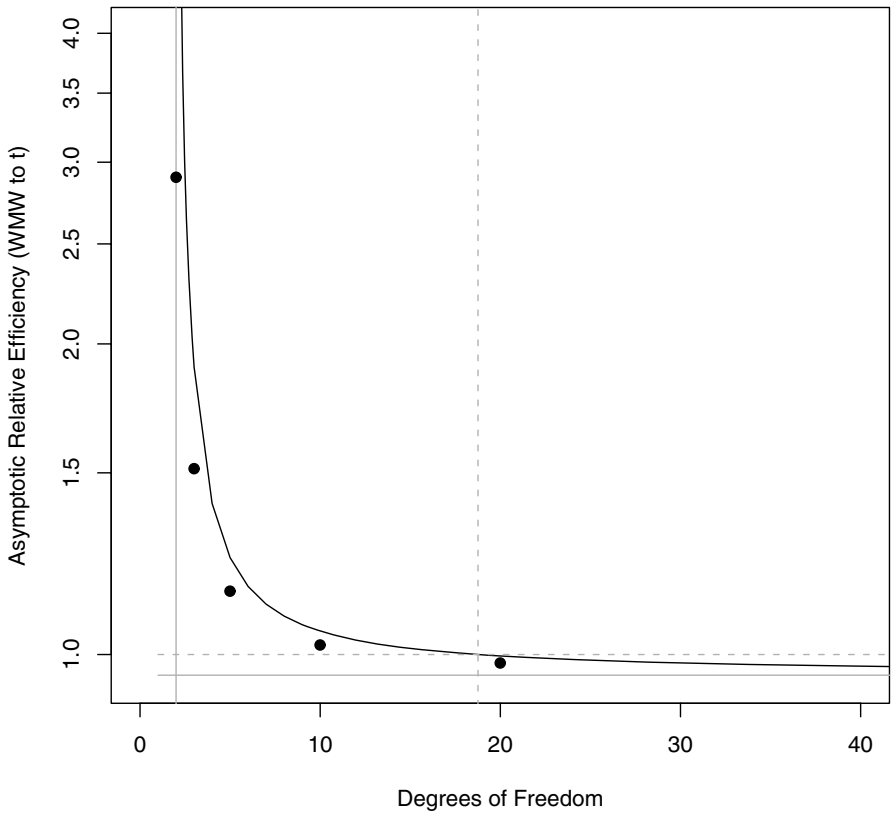

FIG 5. *Relative efficiency of WMW test to t-test for testing for location shift in t-distributions. The dotted grey horizontal line is at 1, and is where both tests are equally asymptotically efficient, which occurs at the dotted grey vertical line at* 18.76. *The solid grey lines denote limits, the vertical line shows ARE goes to infinity at* $df = 2$, *the horizontal line shows ARE goes to* $3/\pi = .955$ *as* $df \to \infty$. *Points are simulated relative efficiency for shifts which give about 80% power for the WMW DR when there are about 20 in each group.*

Consider another type of robustness. Instead of wishing to make inferences about the entire distribution of the data, we may wish to make inferences about the bulk of the data. For example, consider a contamination model where the distribution F (G) may be written in terms of a primary distribution, F_p (G_p) and an outlier distribution, F_o (G_o) with ϵ_f (ϵ_g) of the data following the outlier distribution. In other words,

$$\begin{aligned} F &= (1-\epsilon_f)F_p + \epsilon_f F_o \\ &\textit{and} \\ G &= (1-\epsilon_g)G_p + \epsilon_g G_o \end{aligned} \tag{5.4}$$

In this setup, we want to make inferences about F_p and G_p, **not** about F and G, and the distributions F_o and G_o represent gross errors that we do not wish

to overly influence our results. In this setup, the WMW decision rule outperforms the t-test in terms of robustness of efficiency. We can perform a simple simulation to demonstrate this point. Consider $X_1, \ldots, X_{100} \sim N(0,1)$ and $Y_1, \ldots, Y_{99} \sim N(1,1)$ and $Y_{100} = 1000$ is an outlier caused by perhaps an error in data collection. When we simulate the scenario excluding Y_{100}, then all p-values for δ_t and δ_{tW} are less than 5×10^{-6} and all for δ_W are less than 3×10^{-5}, while if we include Y_{100} we get simulated p-values for δ_t and δ_{tW} between 0.26 and 0.29 and p-values for δ_W between 10^{-15} and 10^{-4}. Clearly the WMW decision rule has much better power to detect differences between F_p and G_p in the presence of the outlier. Here we see that only one very gross error in the data may totally "break down" the power of the t-test, even when the outlier is in the direction away from the null hypothesis. A formal statement of this property is given in He, Simpson and Portnoy (1990).

There is an extensive literature on robust methods in which many more aspects of robustness are described in very precise mathematics, and although not a focus, robustness for testing is addressed within this literature (see e.g., Hampel, Ronchetti, Rousseeuw and Stahel, 1986; Huber and Ronchetti, 2009; Jureckova and Sen, 1996). Besides the power breakdown function previously mentioned, an important theoretical idea for limiting the influence of outliers is to find the *maximin* test, the test which maximizes the minimum power after defining the null and alternative hypotheses as neighborhoods around simple hypotheses (e.g., using equation 5.4 with F_p and G_p representing two distinct single distributions). Huber (1965) showed that maximin tests (also called minimax tests) are censored likelihood ratio-type tests (see also Lehmann and Romano, 2005, Section 8.3, or Huber and Ronchetti, 2009, Chapter 10). The problem with this framework is that it is not too convenient for composite hypotheses (Jureckova and Sen, 1996, p. 407). An alternative more general framework is to work out asymptotic robustness based on the influence function (see Huber and Ronchetti, 2009, Chapter 13). A thorough review of those robust methods and related methods and properties are beyond the scope of this paper.

5.3.6. Recommendations on choosing decision rules

The choice of a decision rule for an application should be based on knowledge of the application, and ideally should be done before looking at the data to avoid the appearance of choosing the DR to give the lowest p-value. To keep this section short, we focus on choosing primarily between δ_t (or δ_{tW}) and δ_W, although for any particular application one of the other tests presented in Table 1 may be appropriate. When using the less well known or more complicated DRs, one should decide whether their added complexity is worth the gains in robustness of validity or some other property.

The choice between t- and WMW DRs should not be based on a test of normality. We have seen that under quite general conditions the t-test decision rules are asymptotically valid (see Table 1), so even if we reject the normality assumption, we may be justified in using a t-test decision rule. Further, when the data

are close to normal or the sample size is small it may be very difficult to reject normality. Hampel et al. (1986, p.23), reviewed some research on high-quality data and the departures from normality of that data. They found that usually the tails of the distribution are larger than the normal tails and t-distributions with degrees of freedom from 3 to 10 often fit real data better than the normal distribution. In light of the difficulty in distinguishing between normality and those t-distributions with moderate sample sizes, and in light of the relative efficiency results that showed that the WMW is asymptotically more powerful for t-distributions with degrees of freedom less than 18 (see Section 5.3.4), it seems that in general the WMW test will often be asymptotically more powerful than the t-test for real high quality data. Additionally, the WMW DR has better power properties than the classical t-test when the data are contaminated by gross errors (see Section 5.3.5).

In a similar vein to the recommendation not to test for normality, it is not recommended to use a test of homogeneity of variances to decide between the classical (pooled variance) t-test DR (δ_t) and Welch's DR (δ_{tW}), since this procedure can inflate the type I errors (Moser, Stevens and Watts, 1989).

One case where a t-test procedure may be clearly preferred over the WMW DR is when there are too few observations to produce significance for the WMW DR (see Section 5.3.3). Also, if there are differences in variance, then δ_{tW} (or some of the other decision rules, see Table 1) may be used while δ_W is not valid. In general, whenever the difference in means is desired for interpretation of the data, then the t-tests are preferred. Nonetheless, if there is a small possibility of gross errors in the data (see Section 5.3.5), then there may be better robust estimators of the difference in means which will have better properties (see References in Section 5.3.5).

6. Other examples and uses of MPDRs

In Section 5 we went into much detail on how some common tests may be viewed under different perspectives. In this section, we present without details two examples that show different ways that the MPDR framework can be useful.

6.1. Comparing decision rules: Tests for interval censored data

Another use of the MPDR outlook is to compare DRs developed under different assumptions. Sun (1996) developed a test for interval censored data under the assumption of discrete failure times. In the discussion of that paper, Sun states that his test "is for the situation in which the underlying variable is discrete" and "if the underlying variable is continuous and one can assume proportional hazards model, one could use Finkelstein's [1986] score test". Although there can be subtle issues in differentiating between continuous and discrete models especially as applied to censored data (see e.g., Andersen, Borgan, Gill and Keiding, 1993, Section IV.1.5), in this case Sun's (1996) test can be applied if the underlying variable is continuous. If we look at Sun's (1996) DR as a MPDR then

this extends the usefulness and applicability of his test, since it can be applied to both continuous and discrete data. In fact, under the MPDR outlook Finkelstein's (1986) test can be applied to discrete data as well. See Fay (1999) for details.

6.2. Interpreting rejection: Genetic tests of neutrality

In the examples of Section 5, the MPDR outlook was helpful in interpreting the scope of the decision. Some perspectives provide a fairly narrow scope with perhaps some optimal property (e.g., t-test of difference in normal means with the same variances is uniformly most powerful unbiased), while other perspectives provide a much broader scope for interpreting similar effects (e.g., the difference in means from the t-test can be asymptotically interpreted as a shift in location for any distribution with finite variance). In this section, we provide an example where the different perspectives do not just provide a difference between a broad or narrow scope of the same general tendency, but the different perspectives highlight totally different effects. In other words, from one perspective rejecting the null hypothesis means one thing, and from another perspective rejecting the null hypothesis means something else entirely.

The example is a test of genetic neutrality (Tajima's [1989] D statistic), and the original perspective on rejection is that evolution of the population has not been neutral (e.g., natural selection has taken place). This perspective requires many assumptions. Before mentioning these we first briefly describe the DR.

The data for this problem are n sequences of DNA, where each sequence is from a different member of a population of n individuals from the same species. The sequences have been aligned so that each sequence is an ordered list of w letters, where each letter represents one of the four nucleotides of the genetic code (A,T,C, and G). We call each position in the list, a *site*. Let S be the number of sites where not all n sequences are equal to the same letter. Let $\hat{k}$ be the average number of pairwise differences between the n sequences. Tajima's D statistic is

$$D = \frac{\hat{k} - \frac{S}{a_1}}{\sqrt{\hat{V}}}$$

where $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$, and $\hat{V}$ is an estimate of the variance of $\hat{k} - \frac{S}{a_1}$ which is a function of S and some constants which are functions of n only (see Tajima, 1989, equation 38). We reject if D is extreme compared to a generalized beta distribution over the range D_{min} to D_{max} with mean equal to 0 and variance 1, and both range parameters are also functions of n only (see Tajima, 1989).

To create a probability model for D Tajima assumed (under the null hypothesis) that:

1. there is no selection (i.e., there is genetic neutrality),
2. the population size is not changing over time,
3. there is random mating in the population,
4. the species is diploid (has two copies of the genetic material),

5. there is no recombination (i.e., a parent passes along either his/her mother's or his/her father's genetic material in its entirety instead of picking out some from the mother and some from the father),
6. any new mutation happens at a new site where no other mutations have happened,
7. the mutation rate is constant over time.

If all the assumptions hold then Tajima's D has expectation 0 and the associated DR is an approximately valid test. The problem is that when we observe an extreme value of D, then it could be either due to (1) chance (but this is unlikely because it is an extreme value), (2) selection has taken place in that population (i.e., assumption 1 is not true), or (3) one of the other assumptions may not be true. This interpretation may seem obvious, but unfortunately, according to Ewens (2004), p. 348, the theory related to tests of genetic neutrality is often applied "without any substantial assessment of whether [the assumptions] are reasonable for the case at hand".

The MPDR framework applied to this problem could define the same null hypothesis as listed above, but have a different alternative hypothesis for each perspective according to whether one of the assumptions does not hold. For intuition into the following, recall that D will be negative if each site on average has a lower frequency of pairwise nucleotide differences than would be expected. Now consider alternatives where one and only one of the assumptions of the null is false.

Selection: If Assumption 1 is false, then the associated alternative creates a perspective that is Tajima's original one, and that is why the test is called a test of genetic neutrality. When we reject the null hypothesis, then this is seen as implying that there is selection (i.e., there is not genetic neutrality). Specifically, if there has recently been an advantageous mutation such that variability is severely decreased in the population, this is a selective sweep and the expectation of D would be negative. Conversely, if there is balancing selection, then the expectation of D would be positive (see e.g., Durrett, 2002).

Non-constant Population Size: Consider when Assumption 2 is false under the alternative.

Growth of Population: If the population is growing exponentially then we would expect D to be negative (see e.g., Durrett, 2002, p. 154).

Recent Bottleneck: A related alternative view is that the genetic variation in the population happened within a fairly large population, but then the population size was suddenly reduced dramatically and the small remaining population grew into a larger one again. This is known as bottle-necking (see e.g., Winter, Hickey and Fletcher, 2002). Tajima (1989) warned that rejection of the null hypothesis could be caused by recent bottlenecking, and Simonsen, Churchill and Aquadro (1995) showed that Tajima's D has reasonable power to reject under the alternative hypothesis of a recent bottleneck.

Random Mating: Consider the alternative where Assumption 3 is false. If the mating is more common (but still random) within subgroups, then this can lead to positive expected values of D (see e.g., Durrett, 2002, p. 154, Section 2.3).

These results for Tajima's D are now 'well known', and a user of the method should be aware of all the possible alternative interpretations (different perspectives) when the null hypothesis is rejected. As with other MPDRs the p-value is calculated the same way, but the interpretation has very real differences depending on the perspective. But unlike the previous examples of Section 5 and 6.1, the different interpretations are not just an expansion or shrinkage of scope of applicability, but they describe qualitatively different directions for looking at rejection of the null.

7. Discussion

We have described a framework where one DR may be interpreted under many different sets of assumptions or perspectives. We conclude by reemphasizing two major points highlighted by the MPDR framework:

- **Do not necessarily disregard results of a decision rule because it is obviously invalid from one perspective.** Perhaps it is valid or approximately valid from a different perspective. For example, consider a hypothesis test comparing two HIV vaccines, where the response is HIV viral load in the blood one year after vaccination. Even if both groups have median HIV viral loads of zero even under the alternative (which is very likely), invalidating the location shift perspective and all more restrictive perspectives than that (see Figure 1), that does not mean that a WMW DR cannot be applied under a different perspective (e.g., Perspective 3). As another example, suppose that a large clinical trial shows a significant difference in means by t-test but the test of normality determines the data are significantly non-normal. Then the t-test p-value can still be used under the general location shift perspective instead of under the normal shift perspective. Finally, consider the tests for interval censored data developed for continuous data which could be shown to be valid for discrete data as well (see Section 6.1).
- **Be careful of making conclusions by assumption.** In Section 6.2 we showed how the rejection of a genetic test of neutrality could be interpreted many different ways depending on the assumptions made. Every time a genetic test of neutrality is used, all the different perspectives (sets of assumptions) should be kept in mind, since focusing on only one perspective could lead to the totally wrong interpretation of the decision rule.

Therefore, the fact that a decision rule can have multiple perspectives can be either good or bad for clarification of a scientific theory; the MPDR may help support the theory by offering multiple statistical formulations consistent with

it, or the MPDR may highlight statistical formulations that may be consistent with alternative theories as well.

Appendix A: Nonparametric Behrens-Fisher decision rule of Brunner and Munzel

Let $R_1, \ldots, R_n$ be the mid-ranks of the Y_i values regardless of Z_i value, and let $W_1, \ldots, W_n$ be the within-group mid-ranks (e.g., if there are no ties and $W_i = w$ and $Z_i = 1$ then Y_i is the wth largest of the responses with $Z_i = 1$). Let $\bar{R}_1 = \frac{1}{n_1}\sum_{i=1}^n Z_i R_i$ and $\bar{R}_0 = \frac{1}{n_0}\sum_{i=1}^n (1-Z_i)R_i$, and

$$\begin{aligned}\hat{\tau}_1^2 &= \frac{1}{n_0^2(n_1-1)}\sum_{i=1}^n Z_i\left(R_i - W_i - \bar{R}_1 + \frac{n_1+1}{2}\right)^2 \\ \hat{\tau}_0^2 &= \frac{1}{n_1^2(n_0-1)}\sum_{i=1}^n (1-Z_i)\left(R_i - W_i - \bar{R}_0 + \frac{n_0+1}{2}\right)^2.\end{aligned}$$

More intuitively, we can write $\hat{\tau}_1^2$ and $\hat{\tau}_0^2$ as

$$\begin{aligned}\hat{\tau}_1^2 &= \frac{1}{n_1-1}\sum_{i=1}^n Z_i\left(\tilde{G}(Y_i) - \bar{G}\right)^2 \\ \hat{\tau}_0^2 &= \frac{1}{n_0-1}\sum_{i=1}^n (1-Z_i)\left(\tilde{F}(Y_i) - \bar{F}\right)^2\end{aligned}$$

where $\tilde{F}(t) = n_1^{-1}\sum_{i=1}^n Z_i\left\{I(Y_i < t) + \frac{1}{2}I(Y_i = t)\right\}$ and $\bar{F} = n_0^{-1}\sum_{i=1}^n (1 - Z_i)\tilde{F}(Y_i)$, and similarly for $\tilde{G}$ and $\bar{G}$. Then, let

$$V_J = \sqrt{\frac{\hat{\tau}_0^2}{n_0} + \frac{\hat{\tau}_1^2}{n_1}}$$

$$T_{NBF} = \frac{n^{-1}\left(\bar{R}_1 - \bar{R}_0\right)}{\sqrt{V_J}} = \frac{\phi(\hat{F}, \hat{G}) - \frac{1}{2}}{\sqrt{V_J}}.$$

When there is no overlap between the two groups, then $\tau_0 = \tau_1 = 0$ and $V_J = 0$; in this case Neubert and Brunner (2007), p. 5197, suggest setting V_J equal to the lowest possible non-zero value: $\frac{1}{\sqrt{2n_0^2 n_1^2}}$. We reject when $|T_{NBF}| > t_{d_B}^{-1}(1-\alpha/2)$ where d_B is given by d_W of equation 5.1 except that τ_0^2 and τ_1^2 replace σ_0^2 and σ_1^2.

Appendix B: Counterexample to uniform control of error rate for the t-test

Let $Y_1, \ldots, Y_n$ be iid Bernoulli random variables with parameter $p_n = \ln(2)/n$. Suppose that the two groups have equal numbers so that $n_0 = n_1 = n/2$. Let

$S_1 = \sum_{i=1}^n Z_i Y_i$ and $S_0 = \sum_{i=1}^n (1 - Z_i) Y_i$. Then

$$T_t(X) = \frac{S_1 - S_0}{\sqrt{\frac{n}{n-2}\left(S_1 - \frac{2S_1^2}{n} + S_0 - \frac{2S_0^2}{n}\right)}}.$$

When $S_1 \geq 1$ and $S_0 = 0$ then

$$T_t(X) = \frac{S_1}{\sqrt{\frac{n}{n-2}\frac{S_1(n-2S_1)}{n}}} \geq \frac{S_1}{\sqrt{\frac{n}{n-2}\frac{S_1(n-2)}{n}}} = \sqrt{S_1} \geq 1$$

Now suppose that $\alpha = .16$ one-sided. If $n \geq 100$ then $t_{n-2}^{-1}(.84) \geq .9995$ and we would reject whenever $S_1 \geq 1$ and $S_0 = 0$, which occurs with probability

$$Pr\left[S_1 \geq 1 \text{ and } S_0 = 0\right] = (1 - Pr[S_1 = 0])Pr[S_0 = 0] = (1 - q_{n_1})q_{n_0} \approx 0.25$$

where $q_n = \left(1 - \frac{\ln(2)}{n}\right)^n$ and the approximation uses $\lim_{n\to\infty} q_n = \exp(-\ln(2)) = \frac{1}{2}$; for $n = 100$ we get $q_{50}(1 - q_{50}) = 0.249994$ and for $n > 100$ the approximation gets closer to .25. Therefore, for this sequence of distributions, with $n > 100$ the type I error rate is 0.249 or more instead of the intended 0.16.

Appendix C: Sufficient conditions for uniform control for the t-test

The test δ_t is UAV if we impose certain conditions on the common distribution function F. For example, for $0 < B < \infty$ and $\epsilon > 0$, consider the class $\Psi_{B,\epsilon}$ of distribution functions such that $Var(Y) \geq \epsilon$ and $\mathrm{E}(Y^4) \leq B$. We will show that for every sequence of functions $F_n \in \Psi_{B,\epsilon}$, assuming $\lim_{n\to\infty} n_0/n = \lambda_0$ and $0 < \lambda_0 < 1$, the type 1 error rate has limit α. We do this in three steps: 1) we use the Berry-Esseen theorem to show that even though F_n may change with n, the distribution of the z-score associated with the sample mean converges uniformly to a normal distribution, 2) we show that the same is true of the z-score for the difference of two sample means, and 3) we show that the same is true for the t-score, which uses the sample variance instead of the population variance in the denominator.

Step 1

Consider a sequence of distribution functions $F_n \in \Psi_{B,\epsilon}$ with associated mean and variance equal to μ_n and σ_n^2. In this appendix Y_n denotes a random variable from F_n and the sample means for each group are $\hat{\mu}_{1n}$ and $\hat{\mu}_{0n}$. According to the Berry-Esseen theorem, for any sequence of distribution functions $F_n \in \Psi_{B,\epsilon}$,

$$\sup_z \left| \Pr\left\{ \frac{\hat{\mu}_{1n} - \mu_n}{\sigma_n/\sqrt{n_1}} \leq z \right\} - \Phi(z) \right| \leq \frac{33}{4} \frac{\mathrm{E}|Y_n - \mu_n|^3}{\sigma_n^3 \sqrt{n_1}},$$

for all n. Of course the same is true of the group with $Z_i = 0$ except with n_1 replaced by n_0, because they have the same distribution. Notice that $\mathrm{E}|Y_n - \mu_n|^3 \le \{\mathrm{E}|Y_n - \mu_n|^4\}^{3/4}$, and

$$\begin{aligned} |Y_n - \mu_n|^4 &\le (|Y_n| + |\mu_n|)^4 \le \{2\max(|Y_n|, |\mu_n|)\}^4 \\ &= 2^4\{\max(|Y_n|, |\mu_n|)\}^4 \le 2^4(|Y_n|^4 + |\mu_n|^4) \\ &= 16(Y_n^4 + |\mathrm{E}(Y_n)|^4) \le 16(Y_n^4 + \mathrm{E}(Y_n^4)) \\ &\le 16(Y_n^4 + B). \end{aligned}$$

It follows that $\mathrm{E}|Y_n - \mu_n|^4 \le 16(\mathrm{E}(Y_n^4) + B) \le 32B$, so $\mathrm{E}|Y_n - \mu_n|^3 \le (32B)^{3/4}$. Furthermore, because $\sigma_n^2 \ge \epsilon$, the Berry-Esseen bound is no greater than $A_{n_1} = (33/4)(32B)^{3/4}/\{n_1\epsilon^3\}^{1/2} \to 0$. The same result holds for the second group.

Step 2

We next show that even though the distribution function F_n may change with n, $(\hat{\mu}_{1n} - \hat{\mu}_{0n})/(\sigma_n^2(1/n_1 + 1/n_0))^{1/2}$ converges uniformly to a standard normal distribution. Specifically, we show that

$$\sup_z \left| \Pr\left\{ \frac{\hat{\mu}_{1n} - \mu_n}{\sigma_n b_n} - \frac{\bar{Y}_n - \mu_n}{\sigma_n b_n} \le z \right\} - \Phi(z) \right| \le A_{n_1} + A_{n_0},$$

where $b_n = (1/n_0 + 1/n_1)^{1/2}$.

Let $\lambda_{0n} = n_0/n$ and $\lambda_{1n} = 1 - \lambda_{0n} = n_1/n$, so that $\lambda_{0n}^{1/2} = b_n^{-1} n_1^{-1/2}$, and assume that $\lambda_{0n} \to \lambda_0$ and $0 < \lambda_0 < 1$. Let H_{n_1} and H_{n_0} denote the distribution functions for $Z_{n_1} = (\hat{\mu}_{1n} - \mu_n)/(\sigma_n/n_1^{1/2})$ and $Z'_{n_0} = (\hat{\mu}_{0n} - \mu_n)/(\sigma_n/n_0^{1/2})$, respectively. Then

$$\begin{aligned} \Pr\left(Z_n\sqrt{\lambda_{0n}} - Z'_n\sqrt{\lambda_{1n}} \le z\right) &= \int_{-\infty}^{\infty} H_{n_1}\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right) dH_{n_0}(u) \\ &= \int_{-\infty}^{\infty} \left\{ \Phi\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right) + H_n\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right) - \right. \\ &\quad \left. \Phi\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right) \right\} dH_{n_0}(u) \\ &= C_n + D_n, \end{aligned}$$

where

$$C_n = \int_{\infty}^{\infty} \Phi\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right) dH_{n_0}(u),$$

and

$$D_n = \int_{-\infty}^{\infty} \left\{ H_{n_1}\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right) - \Phi\left(\frac{z}{\sqrt{\lambda_{0n}}} + u\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}\right)\right\} dH_{n_0}(u).$$

Note that $C_n = \Pr(V < \frac{z}{\sqrt{\lambda_{0n}}} + Z'_{n_0}\sqrt{\frac{\lambda_{1n}}{\lambda_{0n}}}) = \Pr(Z'_{n_0} > V\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}})$, where V is independent of Z'_{n_0} and has a standard normal distribution. We can therefore rewrite C_n as

$$\begin{aligned}
C_n &= \int \left\{1 - H_{n_0}\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right\}\phi(v)dv \\
&= \int \left\{1 - \Phi\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right\}\phi(v)dv + \\
&\quad \int \left\{\Phi\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right) - H_{n_0}\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right\}\phi(v)dv \\
&= \Phi(z) - \int \left\{\Phi\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right) - H_{n_0}\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right\}\phi(v)dv
\end{aligned}$$

To show the last step, notice that if we let V and V' be independent standard normals, then $V\sqrt{\lambda_{0n}} - V'\sqrt{1-\lambda_{0n}}$ is standard normal, so that

$$\begin{aligned}
\int \left\{1 - \Phi\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right\}\phi(v)dv &= Pr\left[V' > V\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right] \\
&= Pr\left[V\sqrt{\lambda_{0n}} - V'\sqrt{\lambda_{1n}} \le z\right] = \Phi(z)
\end{aligned}$$

Thus,

$$\begin{aligned}
|C_n - \Phi(z)| &\le \int_{-\infty}^{\infty} \left|H_{n_0}\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right) - \Phi\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right|\phi(v)dv \\
&\le \int_{-\infty}^{\infty} A_{n_0}\phi(v)dv = A_{n_0},
\end{aligned}$$

where A_{n_0} is as specified at the end of step 1. Also,

$$\begin{aligned}
|D_n| &\le \int_{-\infty}^{\infty} \left|H_{n_1}\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right) - \Phi\left(v\sqrt{\frac{\lambda_{0n}}{\lambda_{1n}}} - \frac{z}{\sqrt{\lambda_{1n}}}\right)\right| dH_{n_0}(u) \\
&\le \int_{-\infty}^{\infty} A_{n_1} dH_{n_0}(u) = A_{n_1}.
\end{aligned}$$

In summary,

$$\begin{aligned} \Pr\left(Z_n\sqrt{\lambda_{0n}} - Z_n'\sqrt{\lambda_{1n}} \le z\right) &= \Phi(z) + C_n - \Phi(z) + D_n \\ \left|\Pr\left(Z_n\sqrt{\lambda_{0n}} - Z_n'\sqrt{\lambda_{1n}} \le z\right) - \Phi(z)\right| &\le |C_n - \Phi(z)| + |D_n| \\ &\le A_{n_0} + A_{n_1} \to 0. \end{aligned} \tag{C.1}$$

Step 3

The next step is to show that the t-statistic, T_t, also converges to a standard normal distribution. To do so, we will show that $\hat\sigma^2_{pn}/\sigma^2_n$ converges to 1 in probability, where $\hat\sigma^2_{pn}$ is the pooled variance. Because $\sigma^2_n \ge \epsilon$, this is equivalent to showing that $\hat\sigma^2_{np} - \sigma^2_n \to 0$ in probability. It is clearly sufficient to show that this is true of each of the two sample variances, say $\hat\sigma^2_{1n}$ and $\hat\sigma^2_{0n}$, since $\hat\sigma^2_{pn}$ is a weighted average of the two sample variances. In the following, we consider only the group with $Z_i = 1$ so all summations are over $\{i : Z_i = 1\}$, and further let $Y_{in} \sim F_n$ for all i.

$$\begin{aligned} (n_1-1)\hat\sigma^2_{n_1} &= \sum(Y_{in} - \mu_n + \mu_n - \hat\mu_{1n})^2 \\ &= \sum(Y_{in} - \mu_n)^2 + n_1(\mu_n - \hat\mu_{n1})^2 + 2(\mu_n - \hat\mu_{n1})\sum(Y_{in} - \mu_n) \\ &= \sum(Y_{in} - \mu_n)^2 + n_1(\hat\mu_{n1} - \mu_n)^2 - 2n_1(\hat\mu_{n1} - \mu_n)^2 \\ &= \sum(Y_{in} - \mu_n)^2 - n_1(\hat\mu_{n1} - \mu_n)^2. \end{aligned}$$

That is, $\hat\sigma^2_{1n} = (n_1-1)^{-1}\left\{\sum(Y_{in} - \mu_n)^2 - n_1(\hat\mu_{1n} - \mu_n)^2\right\}$. The proof is slightly easier if we replace $n_1 - 1$ with n_1. Therefore, we show that $\sum(Y_{in} - \mu_n)^2/n_1 - (\hat\mu_{1n} - \mu_n)^2 - \sigma^2_n \to 0$ in probability. Use Markov's inequality to conclude that $(\hat\mu_{1n} - \mu_n)^2$ converges to 0 in probability: $\Pr\{(\hat\mu_{1n} - \mu_n)^2 > \eta\} \le \mathrm{E}(\hat\mu_{1n} - \mu_n)^2/\eta = \mathrm{var}(\hat\mu_{1n})/\eta = \sigma^2_n/(n_1\eta)$, and $\sigma^2_n = \mathrm{E}(Y_n - \mu_n)^2 \le \sqrt{\mathrm{E(Y_n - \mu_n)^4}} \le \sqrt{32B}$ (from the calculations in step 1). Thus, $(\hat\mu_{1n} - \mu_n)^2 \to 0$ in probability. The only remaining task is to prove that $|\sum(Y_{in} - \mu_n)^2/n_1 - \sigma^2_n|$ converges in probability to 0. Note that

$$\begin{aligned} \Pr\left(\left|\frac{\sum(Y_{in} - \mu_n)^2}{n_1} - \sigma^2_n\right| > \eta\right) &\le \mathrm{var}\left(\sum(Y_{in} - \mu_n)^2/n_1\right)/\eta^2 \\ &= \frac{\mathrm{var}(Y_n - \mu_n)^2}{n_1\eta^2} \\ &= \frac{\mathrm{E}((Y_n - \mu)^4) - \{\mathrm{E}(Y_n - \mu_n)^2\}^2}{n_1\eta^2} \\ &\le \frac{\mathrm{E}(|Y_n - \mu|^4)}{n_1\eta^2} \le \frac{32B}{n_1\eta^2} \to 0. \end{aligned}$$

Appendix D: Justifications for Table 1

D.1. Validity

All the permutation-based DRs (e.g., δ_W, δ_{NBFp}, δ_{tp}, and δ_J) are valid whenever each permutation is equally likely under all $\theta \in \Theta_H$. These DRs can be expressed by either permuting the Z values or permuting the Y values; therefore, all perspectives with $F = G$ for all models in the null space, plus Perspective 9 (the randomization model) give valid tests.

Under H_{12} then δ_t is asymptotically equivalent to δ_{tp}, so when δ_t is not valid for less restrictive hypotheses (e.g., Perspectives 14 and 15, see below), δ_{tp} is also not valid. Since we have assumed $F, G \in \Psi_{fv}$ for all of Table 1, whenever δ_{tp} is valid then δ_t is at least PAV.

Since the denominator of T_{BF} can be written as $\hat{\sigma}_{BF}(1/n_1 + 1/n_0)^{1/2}$ with $\hat{\sigma}^2_{BF} = n^{-1}\left(n_0\hat{\sigma}^2_1 + n_1\hat{\sigma}^2_0\right)$, and we see that $\hat{\sigma}^2_{BF}$ is just a weighted average of the individual sample variances, then similar methods to Appendix C can be used to show that the other t-tests (δ_{tW} and δ_{tH}) are also UAV under Perspective 13.

The δ_t is valid under Perspective 11 and δ_{tH} is valid under Perspective 14 (see e.g., Dudewicz and Mishra, 1988, p. 502), so (δ_{tH}, A_{11}) is valid. Since δ_{tW} is equivalent to δ_{tH} except it has larger degrees of freedom (see e.g., Dudewicz and Mishra, 1988, p. 502), it is UAV under Perspectives 14 and 11. One can show that (δ_{tW}, A_{11}) is not valid for finite samples by simulation from standard normal with $n_1 = 2$ and $n_0 = 30$, which gives type I error of around 12%.

The rules δ_{NBF_a} and δ_{NBF_p} were shown to be PAV under H_{10} (see Brunner and Munzel, 2000 and Neubert and Brunner, 2007 respectively).

For any null where $F = G$ and $F \in \Psi_C$, any rank test that is PAV and only depends on the combined ranks is also UAV. To see that, note that such a DR is invariant to monotone transformations. Thus, if we let $Y^*_{in} = F_n(Y_{in})$ for all $i = 1, \ldots, n$, where F_n is the common distribution function regardless of Z_i, then the Y^*_{in}s are iid uniforms. The type I error rate of the rank test applied to the original data is the same as that applied to the uniforms, so the type I error rate is controlled uniformly over all continuous distributions.

Pratt (1964) showed that under the Perspective 14 (Behrens-Fisher), both the WMW and the t tests are not valid (not even PAV), so since $A_{14} \sqsubset A_{15}$ and and $A_{14} \sqsubset A_{10}$, both those DRs are not valid under Perspectives 15 and 10. However, note that when $n_1 = n_0$ then $T_t = T_{BF}$, and (δ_t, A_{14}) is UAV. We have previously shown in Section 5.2.2 that Perspective 15 is invalid for any DR which has power $> \alpha$ for some $\mathcal{P} \in K_{15}$. For Perspective 10 we can show the two simple discrete distributions with three values at $(-1, 0, 2)$ and probability density functions at those three points as $(.01, .98, .01)$ (associated with F) and $(.48, .02, .48)$ (associated with G) are in H_{10} but are not valid for δ_{tW}, δ_{tH} or δ_{BFp}.

D.2. Consistency

The consistency of δ_W was shown for Perspective 3 by Lehmann (1951) and Putter (1955) expanded this result showing consistency for discrete distributions. We can use the relationship among the assumptions (see Figure 1) to show consistency for most of Table 1. Brunner and Munzel (2000) showed the asymptotic normality of T_{NBF} whenever $\phi(F,G) = 1/2$ and $Var(F(Y_G)) > 0$ and $Var(G(Y_F)) > 0$, and it is straightforward to extend this to show the asymptotic normality of

$$T_{NBF}(\phi) = \frac{\phi(\hat{F},\hat{G}) - \phi(F,G)}{\sqrt{\frac{\hat{\tau}_0^2}{n_0} + \frac{\hat{\tau}_1^2}{n_1}}},$$

even when $\phi(F,G) \neq 1/2$. Thus, δ_{NBFa} is consistent for Perspective 10 and all more restrictive assumptions. Neubert and Brunner (2007) showed the asymptotic normality of the permutation version $T_{NBF}(\phi)$ and hence δ_{NBFp} is consistent for the same perspectives as δ_{NBFa}. For all three rank DRs (δ_W, δ_{NBFa} and δ_{NBFp}), the associated tests are not consistent when there are alternatives which include probability models with $\phi(F,G) = 1/2$.

The consistency of δ_t under K_{12} is shown in e.g., Lehmann and Romano (2005), p. 466, and it is straightforward to extend this to δ_{tW} and δ_{tH}. Thus, for all more restrictive assumptions those tests are also consistent. Similar methods analogously show that both δ_{tW} and δ_{tH} are consistent for K_{14}. For alternatives that contain probability models with $\mu_0 = \mu_1$, all the t-tests are not consistent.

Under K_{11} then δ_t is asymptotically equivalent to δ_{tp} (Lehmann and Romano, 2005, p. 683, problem 15.10); therefore δ_{tp} is consistent. Janssen (1997), Theorem 2.2, showed the consistency of δ_{tBFp} as long as $\mu_0 \neq \mu_1$ under a less restrictive assumptions than A_{14} which include A_6 and A_7. The DRs δ_{tp} and δ_{BFp} are consistent whenever respectively, δ_p and δ_W are consistent assuming finite variances (see van der Vaart, 1998, p. 188).

References

Andersen, P.K., Borgan, O, Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes* Springer-Verlag: New York. MR1198884

Berger, R.L. and Boos, D.D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89** 1012–1016. MR1294746

Blair, R. C. and Higgins, J.J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. *Journal of Educational Statistics* **5** 309–334.

Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery, second edition.* Wiley: New York. MR2140250

Brunner, E. and Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal* **42** 17–25. MR1744561

Cao, H. (2007). Moderate deviations for two sample t-statistics. *ESAIM: Probability and Statistics* **11** 264–271. MR2320820

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics* Chapman and Hall: London. MR0370837

Dowdy, S., Wearden, S., and Chilko, D. (2004). *Statistics for Research, third edition* Wiley: New York.

Dudewicz, E.J. and Mishra, S.N. (1988). *Modern Mathematical Statistics* Wiley: New York. MR0920168

Durrett, R. (2002). *Probability models for DNA sequence evolution.* Springer: New York. MR1903526

Edgington,E.S. (1995). *Randomization Tests, third edition.* Marcel Dekker, Inc.: New York.

Ewens, W.J. (2004). *Mathematical Population Genetics. I. Theoretical Introduction, second edition* Springer: New York. MR2026891

Fay, M.P. (1999). Comparing several score tests for interval censored data. *Statistics in Medicine* **18** 273–285 (Correction: 1999; 2681).

Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42** 845–854. MR0872963

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests* Academic Press: New York.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions* Wiley: New York. MR0829458

He, X., Simpson, D.G., and Portnoy, S.L. (1990). Breakdown robustness of tests. *Journal of the American Statistical Association* **85** 446–452. MR1141746

Hennekens, C.H., Eberlein, K.A., for the Physicians' Health Study Research Group. (1985). A randomized trial of aspirin and β-carotene among U.S. physicians. *Preventive Medicine* **14** 165–168.

Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks.* Krieger Publishing Company: Malabar, Florida. MR0758442

Hodges, J.L. and Lehmann, E.L. (1963). Estimates of Location Based on Rank Tests. *Annals of Mathematical Statistics* **34** 598–611. MR0152070

Huber, P.J. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics* **36** 1753–1758. MR0185747

Huber, P.J., and Ronchetti, E.M. (2009). *Robust Statistics, second edition* Wiley: New York. MR2488795

Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses on the generalized Behrens-Fisher problem. *Statistics and Probability Letters* **36** 9–21. MR1491070

Janssen, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference* **81** 71–93. MR1718393

Jureckova, J. and Sen, P.K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations* Wiley: New York. MR1387346

Kempthorne, O. and Doerfler, T.E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika* **56** 231–248. MR0254965

Lehmann, E.L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics* **22** 165–179. MR0040632

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks* Holden-Day, Inc.: Oakland, CA. MR0395032

Lehmann, E.L. (1997). Review of *Error and the Growth of Experimental Knowledge* by D.G. Mayo. *Journal of the American Statistical Association* **92** 789.

Lehmann, E.L. (1999). *Elements of Large-Sample Theory* Springer: New York. MR1663158

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses, third edition* Springer, New York. MR2135927

Ludbrook, J. and Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *American Statistician* **52** 127–132.

Mallows, C.L. (2000). Letter to the Editor in Response to Ludbrook and Dudley(1998. **54** 86–87.

Mann, H.B. and Whitney, D.R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other. *Annals of Mathematical Statistics* **18** 50–60. MR0022058

Mayo, D.G. (1996). *Error and the Growth of Experimental Knowledge* University of Chicago Press, Chicago.

Mayo, D.G. (2003). Comment on Could Fisher, Jefferys and Neyman Have Agreed on Testing. by J.O. Berger. *Statistical Science* **18** 19–24. MR1997064

Mayo, D.G. and Spanos, A. (2004). Methodology in Practice: Statistical Misspecification Testing. *Philosophy of Science* **71** 1007–1025. MR2133711

Mayo, D.G. and Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. **57** 323–357. MR2249183

McCullagh, P. (1980). Regression models for ordinal data. (with discussion) *Journal of the Royal Statistical Society, series B* **42** 109–142. MR0583347

McDermott, M.P. and Wang, Y. (1999). Comment on "The emperor's new tests" by Perlman and Wu *Statistical Science* **14** 374–377.

Mee, R.W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association.* **85** 793–800. MR1138359

Mehta, C.R., Patel, N.R., and Tsiatis, A.A. (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **40** 819–825. MR0775388

Moser, B.K., Stevens, G.R., and Watts, C.L. (1989). The two-sample t test versus Satterhwaite's approximate F test. *Communications in Statistics: Theory and Methods* **18** 3963–3975. MR1058922

Neubert, K. and Brunner, E. (2007). A Studentized permutation test for the non-parametric Behrens-Fisher problem. *Computational Statistics and Data Analysis* **51** 5192–5204. MR2370717

Perlman, M. and Wu, L. (1999). The emperor's new tests (with discussion). *Statistical Science* **14**, 355–381. MR1765215

Pratt, J.W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association* **59** 665–680. MR0166871

Putter, J. (1955). The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics* **26** 368–386. MR0070923

Proschan, M. and Follmann, D. (2008). Cluster without fluster: the effect of correlated outcomes on inference in randomized clinical trials. *Statistics in Medicine* DOI 10.1002/sim.2977. MR2420113

Sawilowsky, S.S. and Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin* **111 (2)** 352–360.

Sen, P.K. (1967). A note on asymptotically distribution-free confidence bounds for $Pr\{Y < X\}$ based on two independent samples. *Sankhya, Ser. A*, **29 (Pt. 1)** 95–102. MR0226772

Simonsen, K.L., Churchill, G.A., and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141** 413–429.

Sterring Committee of the Physicians' Health Study Research Group (1988). Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New England Journal of Medicine* **318** 262–264.

Sterring Committee of the Physicians' Health Study Research Group (1989). Final Report on the Aspirin Component of the Ongoing Physicians' Health Study. *New England Journal of Medicine* **321** 129–135.

Sun, J. (1996). A non-parametric test for interval-censored failure time data with applications to AIDS studies. *Statistics in Medicine*, **15**, 1387–1395.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123** 585–595.

van der Vaart, A.W. (1998). *Asymptotic Statistics* Cambridge University Press, Cambridge. MR1652247

Wilcoxon, F. (1945). Individual comparisons by Ranking Methods. *Biometrics Bulletin* **1** 80–83.

Winter, P.C., Hickey, G.I., and Fletcher, H.L. (2002). *Instant Notes: Genetics, second edition*. Bios Scientific Publishers: Oxford.

Whitt, W. (1988). Stochastic Ordering. in *Encyclopedia of Statistics* Vol. 8, S. Kotz and N.L. Johnson (editors). Wiley: New York.