entropy

Article

Indicators of Evidence for Bioequivalence

Stephan Morgenthaler [1,†] **and Robert Staudte** [2,*,†]

[1] Département de Mathématiques, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland; stephan.morgenthaler@epfl.ch
[2] Department of Mathematics and Statistics, La Trobe University, Melbourne 3086, Australia
[*] Correspondence: r.staudte@latrobe.edu.au; Tel.: +61-3-9455-3206; Fax: +61-3-9479-2466
[†] Both authors contributed equally to this work.

Academic Editors: Julio Stern, Adriano Polpo and Kevin H. Knuth
Received: 29 May 2016; Accepted: 2 August 2016; Published: 9 August 2016

Abstract: Some equivalence tests are based on two one-sided tests, where in many applications the test statistics are approximately normal. We define and find evidence for equivalence in Z-tests and then one- and two-sample binomial tests as well as for t-tests. Multivariate equivalence tests are typically based on statistics with non-central chi-squared or non-central F distributions in which the non-centrality parameter λ is a measure of heterogeneity of several groups. Classical tests of the null $\lambda \geq \lambda_0$ versus the equivalence alternative $\lambda < \lambda_0$ are available, but simple formulae for power functions are not. In these tests, the equivalence limit λ_0 is typically chosen by context. We provide extensions of classical variance stabilizing transformations for the non-central chi-squared and F distributions that are easy to implement and which lead to indicators of evidence for equivalence. Approximate power functions are also obtained via simple expressions for the expected evidence in these equivalence tests.

Keywords: evidence for alternatives; Kullback–Leibler divergence; sample size determination; two by two tables; two one-sided tests; variance stabilization

1. Introduction

Our purpose is to extend the concept of "evidence for the alternative hypothesis", already available in classical one-sided testing, to contexts where that alternative is "equivalence" of two or more distributions. We abbreviate the term "bioequivalence" to "equivalence" for simplicity and because these results are of much more general applicability.

1.1. Background and Summary

Why should we introduce another approach to equivalence testing? Because, even though some equivalence tests [1] are well established and embraced by the USA Food and Drug Administration (FDA) and the European Medicines Agency (EMA), there are substantial critiques [2–4], as well as novel, competing approaches to multivariate equivalence testing [4–8].

We endorse the proposal by [4] to define a hierarchy of bioequivalence models, "average bioequivalence" within "population bioequivalence" within "individual bioequivalence", in terms of the Kullback–Leibler symmetrized distance (KLD) between distributions arising in equivalence tests. Somewhat surprisingly, we advocate estimating these distances indirectly using variance stabilized test statistics (VSTs) rather than plug-in parameter estimates for parameters appearing in the KLD. The close tie between the mean of a VST and the KLD for non-central chi-squared and F distributions will be illustrated in Appendix A.

In the remainder of this introduction we describe the notion of "evidence for the alternative" in the context of Z-tests and show how it is connected to level and power of such tests; these ideas were first introduced in [9]. After these preliminaries, we extend the notion of evidence for equivalence to two one-sided Z-tests (TOSTs), and show how it is related to a one-sided test based on a non-central chi-squared statistic with one degree of freedom.

This notion of evidence is applicable to many situations because VSTs can carry many test statistics into normally distributed statistics Z. For exponential families, Reference [10] show that the expected evidence of the variance stabilized statistic T is approximately equal to the signed square root of the Kullback–Leibler symmetrized divergence. Examples not from exponential families are in [11–13]. These results gives support to calling T the "evidence for the alternative hypothesis".

Throughout the paper we employ standard notation and properties of the non-central t, χ^2 and F distributions which are introduced and studied in depth in [14,15]. Below we first discuss specific examples of two one-sided tests (TOST) in Section 2, namely for binomial, two by two tables and t-tests; all methods are illustrated using data examples from the literature. In the multivariate setting, chi-squared tests for equivalence are given in Section 3 and F-tests for equivalence of K normal populations in Section 4. New methods for choosing λ_0 and for variance stabilizing the test statistics are provided. Discussion follows in Section 5, and R scripts for implementing some procedures are found in the Appendix B.

1.2. Properties of Evidence in One-Sided Z-Tests

Univariate equivalence tests are often based on statistics having the non-central t distribution or normal approximations to the binomial distribution, so we begin our introduction to evidence contained in such tests with the approximating, and simpler, Z-tests.

In the prototypical model $X \sim N(\mu, 1)$ where μ is unknown, and one tests the null hypothesis $\mu \leq \mu_0$ against the alternative $\mu > \mu_0$ by rejecting the null if the p-value is sufficiently small. For an observed $X = x$, the p-value is $\Phi(\mu_0 - x)$, where Φ is the standard normal distribution function. The "evidence for the alternative" $\mu > \mu_0$ is defined to be $T \equiv X - \mu_0$. It is normally distributed with mean $\mu - \mu_0$, which is linearly increasing in μ. In addition, T estimates its mean with standard error 1, regardless of the value of μ. For reasons given in [9,16] values of T near 1.645, 3.3 and 5 are called "weak", "moderate" and "strong" evidence for the alternative. Note that for an observed $T = t = x - \mu_0$, the p-value can be recovered from $1 - \Phi(t)$.

In Table 1 are shown two sets of numbers with very different interpretations, resulting from different assumptions. They are based on one observation $T = t$, where $T \sim N(\mu, 1)$ and $\mu \geq 0$. If one assumes the boundary hypothesis $\mu = 0$, the second row of p-values gives the correct "degree of surprise" at having observed $T = t$; the smaller the p-value, the more surprised one is with the outcome. However, if one only assumes $\mu \geq 0$, the first row gives an estimate of the expected evidence $\mathrm{E}[T] = \mu$ for the alternative $\mu > 0$. This estimate has an additive standard normal error. When one interprets p-values, one must be careful not to interpret the smallness of their magnitudes as though they were evidence for the alternative on a linear scale. The first row in Table 1 gives a much more reasonable estimate of "evidence for the alternative" together with an easily understood standard error. The compatibility of this calibration scale for evidence with Bayesian calibration scales for p-values and Bayes factors is discussed in [10] (Section 4.3).

Table 1. p-values for testing $\mu = 0$ against $\mu > 0$ and evidence estimates for alternatives based on one observation $T = t$, where $T \sim N(\mu, 1)$. Keep in mind that the standard error of the observed value of t is equal to 1.

t	0	1.281	1.645	2.326	3.090	3.3	3.719	5
p-value	0.5	0.10	0.05	0.01	0.001	0.0005	0.0001	0.0000003

The generality of this definition of evidence for the alternative stems from the fact that in many situations the natural test statistics X can be transformed to T, which has approximately a normal distribution with unit variance. Also, it is a more basic concept of the test than level and power. For a normally distributed test statistic T with unit variance, the expected evidence for one-sided alternatives $\mu > \mu_0$ is related to the level α and power $1 - \beta(\mu)$ through the sum of the probits:

$$\mathrm{E}_\mu[T] = z_{1-\alpha} + z_{1-\beta(\mu)}\,. \tag{1}$$

In the same testing problem the sample size n required to detect an alternative μ_1 with power 0.8 at level 0.05 is the solution of $\sqrt{n}\,(\mu_1 - \mu_0) = z_{0.95} + z_{0.8} \approx 2.5$; the expected evidence in such an experiment therefore lies between weak and moderate.

For negative T, $-T$ is interpreted as evidence for the null $\mu \leq \mu_0$. Because of the symmetry of the problem, if we had begun with the null hypothesis $\mu \geq \mu_0$ against the alternative $\mu < \mu_0$ the evidence for the alternative would be defined as $\mu_0 - X$.

1.3. Properties of Evidence in Two One-Sided Z-Tests (TOST)

Consider the simplest example with one observation $X \sim N(\mu, 1)$ for testing $H_0 : |\mu| \geq \mu_0$, where $\mu_0 > 0$ defines the equivalence alternative $H_1 : |\mu| < \mu_0$. The null hypothesis consists of two possibilities: $\mu \leq -\mu_0$ and $\mu \geq \mu_0$. The left hand part is rejected at level α if $X + \mu_0 \geq z_{1-\alpha}$, and the evidence for its alternative is $T_- = X + \mu_0$, because $T_- \sim N(\mu + \mu_0, 1)$ and T_- has an expected value that increases in μ and is 0 at the boundary $\mu = -\mu_0$. The right hand part is rejected at level α if $X - \mu_0 \leq z_\alpha$, and the evidence for its alternative is $T_+ = \mu_0 - X$, because $T_+ \sim N(\mu_0 - \mu, 1)$, whose expected value is increasing with *decreasing* μ and is 0 at its null boundary. The two one-sided testing procedure (equivalence test) rejects in favor of equivalence only if both of the one-sided tests reject their respective null hypotheses, and this has level α, because only one null hypothesis can hold.

The *evidence for the alternative hypothesis of equivalence* is logically the minimum of the evidences for the two one-sided tests:

$$T = \min\{T_-, T_+\} = \mu_0 + T_0\,, \tag{2}$$

where $T_0 = \min\{X, -X\}$ and $X \sim N(\mu, 1)$. Now $-T_0 = |X|$ has a folded (to the right) normal distribution ([14] (p. 170), and [16]), with parameters $(\mu, 1)$, so T_0 has a folded (to the left) normal distribution with the same parameters. The density of T_0 is given in terms of the standard normal density φ for $t < 0$ by $f_{T_0}(t;\mu) = \varphi(t - \mu) + \varphi(t + \mu)$, so the density of T is for $t < 0$

$$f_T(t;\mu) = \varphi(t - \mu - \mu_0) + \varphi(t + \mu - \mu_0). \tag{3}$$

In Figure 1 are shown in black lines some examples of $f_T(t)$ for $\mu_0 = 4$ and several choices of μ. When $\mu = 0$, (exact equivalence), the density is negative half-normal with upper bound $\mu_0 = 4$, but as $|\mu|$ increases, the distribution rapidly approaches normality.

Also of interest are the mean and standard deviation of T as μ varies. The first two moments of T_0 are $\mathrm{E}_\mu[T_0] = \mu\{1 - 2\Phi(\mu)\} - 2\varphi(\mu)$ and $\mathrm{E}_\mu[T_0^2] = 1 + \mu^2$, so

$$\begin{aligned} \mathrm{E}_\mu[T] &= \mu_0 + \mathrm{E}_\mu[T_0] = \mu_0 + \mu\{1 - 2\Phi(\mu)\} - 2\varphi(\mu) & (4)\\ \mathrm{Var}_\mu[T] &= 1 + \mu^2 - \{\mu\,(1 - 2\Phi(\mu)) - 2\varphi(\mu)\}^2\,. & (5) \end{aligned}$$

The mean and standard deviation of T are shown in Figure 2 as black lines. The top left-hand plot reveals that even in the case of perfect equivalence $\mu = 0$, the equivalence test with $\mu_0 = 3$ will only yield, on average, weak evidence for it. With $\mu_0 = 4$, this average evidence for equivalence when $\mu = 0$ becomes moderate.

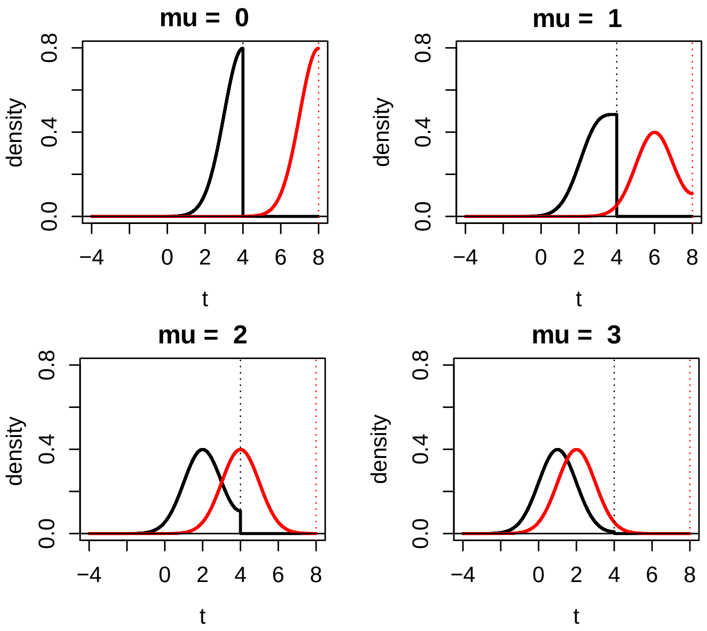

Figure 1. The black lines show the densities $f_T(t)$ of the evidence for equivalence (3), for $\mu_0 = 4$ and various μ. The red lines show the densities of evidence for equivalence (6) when $n = 4$.

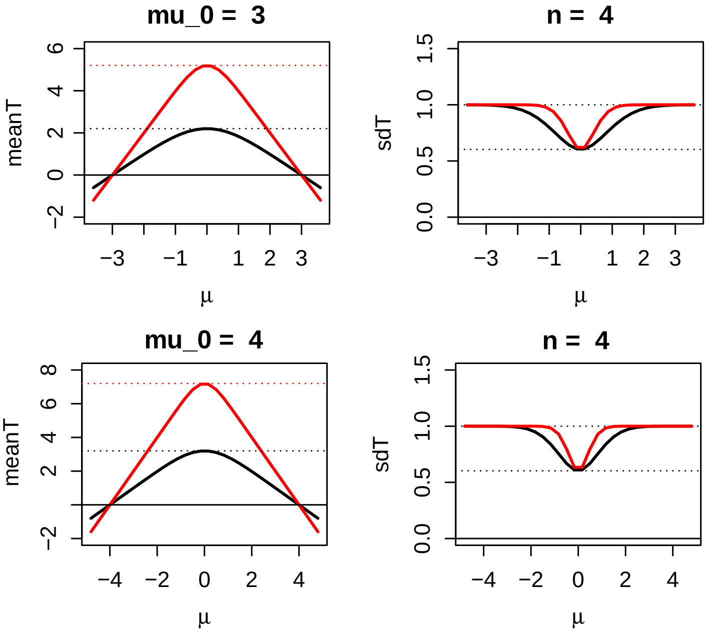

Figure 2. In the top row $\mu_0 = 3$; in the bottom row $\mu_0 = 4$. Exact values of the mean (left plot) and standard deviation (right plot) of TOST evidence T for equivalence are shown in black lines, while the red solid lines show the graphs based on $n = 4$ observations. The horizontal black dotted line is at height $\mu_0 - \sqrt{2/\pi}$, while the red dotted line is $2\,\mu_0 - \sqrt{2/\pi}$. The horizontal dotted lines in the right hand plot are at $\sqrt{1 - 2/\pi}$ and 1.

The plots in Figure 2 suggest that, except for μ near 0, the distribution of T defined by (2) is approximately normal with mean near $\min\{\mu + \mu_0, \mu_0 - \mu\}$ and standard deviation near one. For μ near 0, which is of interest in the case of equivalence, it is not normal, but this is perhaps compensated for by having a smaller standard error. Its distribution approaches a negative half-normal as

$|\mu| \to 0$, with mean and standard deviation from (4) and (5) converging to $\mu_0 - \sqrt{2/\pi}$ and $\sqrt{1-2/\pi}$, respectively.

1.3.1. How Evidence Grows with Sample Size in Two One-Sided Z-Tests

The evidence for a one-sided Z-test grows with the square root of the sample size $\sqrt{n}$ for if the sample mean $\bar{X}_n \sim N(\mu, 1/n)$ then the evidence for the alternative $\mu > -\mu_0$ to the null $\mu \leq -\mu_0$ is $T_{n,-} = \sqrt{n}\,(\bar{X} + \mu_0) \sim N(\sqrt{n}\,(\mu + \mu_0), 1)$, which is increasing in μ, has variance 1, and has expected value 0 at the boundary $\mu = -\mu_0$. Similarly, $T_{n,+} = \sqrt{n}\,(\mu_0 - \bar{X}) \sim N(\sqrt{n}\,(\mu_0 - \mu), 1)$ the evidence for the alternative $\mu < \mu_0$ to the null $\mu \geq \mu_0$. Thus the evidence for equivalence $|\mu| < \mu_0$ based on n observations is:

$$T_n = \min\{T_{n,-}, T_{n,+}\} = \sqrt{n}\,\mu_0 + T_{n,0}, \tag{6}$$

where $T_{n,0} = \sqrt{n}\,\min\{\bar{X}, -\bar{X}\}$ and $\sqrt{n}\,\bar{X} \sim N(\sqrt{n}\,\mu, 1)$. Now $T_{n,0}$ has a folded to the left normal distribution with parameters $\sqrt{n}\,\mu, 1$ and T_n is a shift by $\sqrt{n}\,\mu_0$ of $T_{n,0}$, so its mean and variance are:

$$\begin{aligned} \mathrm{E}_\mu[T_n] &= \sqrt{n}\,\mu_0 + \sqrt{n}\,\mu\{1 - 2\Phi(\sqrt{n}\,\mu)\} - 2\varphi(\sqrt{n}\,\mu) \\ \mathrm{Var}_\mu[T_n] &= 1 + n\,\mu^2 - \left\{\sqrt{n}\,\mu\,(1 - 2\Phi(\sqrt{n}\,\mu)) - 2\varphi(\sqrt{n}\,\mu)\right\}^2 . \end{aligned}$$

A plot of the densities of T_4 compared to T defined by (3) are also shown in Figure 1 as red lines, and similarly for the mean and standard deviation of T_4 in Figure 2.

1.3.2. Sample Size Determination

For a one-sided Z-test based on n observations one can obtain a given expected evidence 2.5, say, for an alternative distant μ_0 from the null, by taking n to satisfy $\sqrt{n}\,\mu_0 = 2.5$. So $n_{\text{1-sided}} = \lceil (2.5/\mu_0)^2 \rceil$, where $\lceil r \rceil$ is the smallest integer greater than or equal to r. For a TOST Z-test with equivalence alternative $|\mu| < \mu_0$, to obtain the same expected evidence when in fact $\mu = 0$ one needs by (7) to have $\sqrt{n}\,\mu_0 - \sqrt{2/\pi} = 2.5$, or $n_{\text{TOST}} = \lceil (2.5 + 0.8)/\mu_0)^2 \rceil$, which is 74% larger than $n_{\text{1-sided}}$.

If one had asked for only weak expected evidence 1.645 instead of 2.5 in the above paragraph, the ratio of sample sizes required by TOST Z-tests to a one-sided test is 2.2, so the equivalence test would require 120% more observations than the one-sided test.

1.4. Connection of Evidence in TOST with a One-Sided Test

The evidence $T = T_{\text{TOST}}$ defined by (2) is based on $X \sim N(\mu, 1)$ for the null hypothesis $H_0 : |\mu| \geq \mu_0 > 0$, composed of two disjoint sets, with the equivalence alternative $H_1 : |\mu| < \mu_0$. One could equally study the evidence in the equivalent experiment $S = X^2 \sim \chi_1^2(\lambda)$, where $\lambda = \mu^2$, for the hypotheses restated as $H_0 : \lambda \geq \lambda_0$ against the equivalence alternative $H_1 : \lambda < \lambda_0$, where $\lambda_0 = \mu_0^2$. The evidence for equivalence in an experiment with $S \sim \chi_1^2(\lambda)$ is a special case of (13) found in Section 3:

$$T = T_{1,\lambda_0}(S) = \begin{cases} \sqrt{\lambda_0} - S/\sqrt{2} & \text{for } S < 1\,; \\ \sqrt{\lambda_0} - \sqrt{S - 1/2} & \text{for } S \geq 1\,. \end{cases} \tag{7}$$

The top left plot of Figure 3 compares the graph of this T with that of T_{TOST} when $\mu_0 = 2$. Further, its expected evidence is by (12) approximately $\sqrt{\lambda_0 + 1/2} - \sqrt{\lambda + 1/2}$. The top right plot of Figure 3 shows the graph of this approximate expected evidence for equivalence as a function of μ for the case $\mu_0 = 2$ as a dashed line, to be compared with the previously obtained expected evidence in the two one-sided test experiment (4), whose graph is shown as a solid line. The bottom plots are for $\mu_0 = 4$.

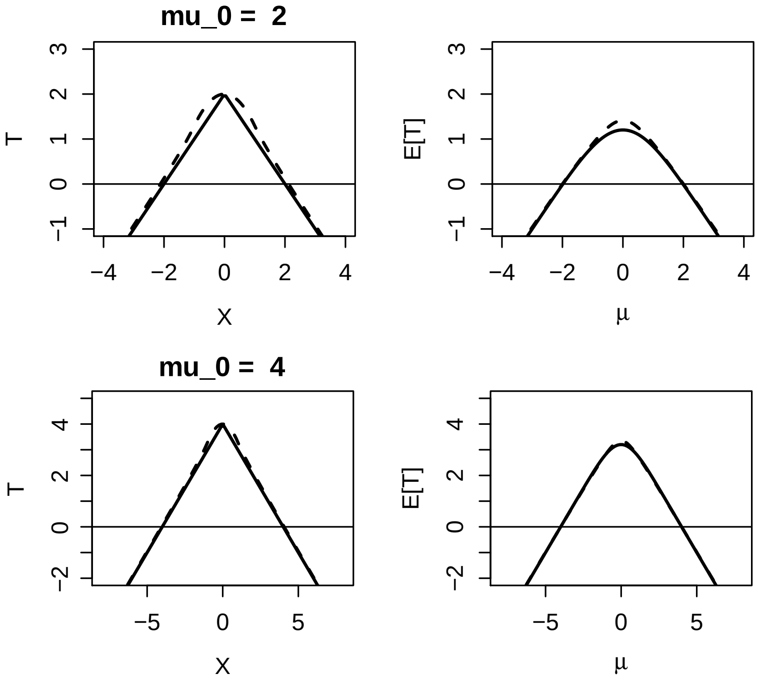

Figure 3. The top plots are for $\mu_0 = 2$; the bottom for $\mu_0 = 4$. On the left is a comparison of T_{μ_0} given by (2) as a function of data $X = x$ plotted as a solid line, to be compared with the one-sided evidence in the corresponding chi-squared test (7) as a dashed line. On the right are comparisons of two approximations for the expected TOST evidence for equivalence, a solid line depicting the exact value (4), and that given by the first order approximation $\sqrt{\lambda_0 + 1/2} - \sqrt{\mu^2 + 1/2}$ for the equivalent chi-squared test, shown as a dashed line.

2. More Examples of Two One-Sided Tests (TOSTS)

2.1. Evidence for Equivalence in Two One-Sided Binomial Tests

Given $X \sim \text{Binomial}(n, p)$, and let $\hat{p} = X/n$. Let $p_1 < p < p_2$ define the region of equivalence. (Often this region will also be of the form $|p - p_0| < \Delta_0$ for some p_0, Δ_0.) We want to test at level α the null $H_0 : \; p \le p_1$ or $p \ge p_2$ against the equivalence alternative $H_1 : p_1 < p < p_2$. The null hypothesis is two-sided and its right-hand part is rejected at level α if $\hat{p} \le p_2 - z_\alpha \sqrt{p_2(1-p_2)/n}$, whereas the left-hand part is rejected at level α if $\hat{p} \ge p_1 + z_{1-\alpha}\sqrt{p_1(1-p_1)/n}$. Only one of these tests can reject the null, so the level of the combined tests is α. We have assumed that n is large enough so that normal critical points give accurate levels.

The VST of $\hat{p}$ is the well-known arc-sine transformation $h(\hat{p}) = 2\sqrt{n}\,\arcsin(\sqrt{\hat{p}})$, which is asymptotically normal with variance 1 and asymptotic mean $2\sqrt{n}\,\arcsin(\sqrt{p})$. Large values of $h(\hat{p})$ indicate evidence for large p. The evidence in the test of $p \le p_1$ for an alternative $p > p_1$ is therefore $T_- = h(\hat{p}) - h(p_1)$, while the evidence in the test of $p \ge p_2$ for an alternative $p < p_2$ is $T_+ = h(p_2) - h(\hat{p})$. For the combined two one-sided tests, the evidence for equivalence is the minimum evidence in these two one-sided tests; that is, $T = \min\{T_-, T_+\}$.

Example 1. (Intervention success of new treatment.) In Example 4.2 of [17] (p. 59) a highly toxic drug used in chemotherapy treatment for a tumor led to a 73% two-year progression-free survival period. A new combined and much more tolerable treatment was administered to 361 patients and it was deemed equivalent to the previous treatment if the success rate fell in the interval [0.65, 0.75]. In the new treatment 191 patients survived a two-year progression-free period, and Wellek used two one-sided binomial tests to find that non-equivalence was rejected at the 0.05 level with estimated power 0.12. For these data $T_+ = 2.84$, $T_- = 0.793$ so the evidence for equivalence of the treatment effect is $T = 0.793$, which is "weak". The standard error of $T = 0.793$ is known (see the comments at the end of Section 2), to satisfy $0.60 \leq \text{SE}[T] \leq 1$, but is likely to near the smaller bound, because $\hat{p} = 191/273$ is close to the center of the equivalence interval. This result is consistent with the analysis of [17], and it is much simpler.

2.2. Evidence for Equivalence of Risks

It is often the case that one wants to compare risks associated with new and standard treatments, with data often displayed in 2 by 2 tables; an example is given below after we introduce notation and explain how to find evidence for equivalence in this context.

Let X_1 and X_2 be two independent binomial random variables with parameters (n_1, p_1) and (n_2, p_2), respectively. Letting $\hat{p}_i = (X_i + 0.5)/(n_i + 1)$ for $i = 1, 2$ the unknown risk difference $\Delta = p_1 - p_2$ is estimated by $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$. We want the evidence for equivalence hypothesis $\Delta_1 < \Delta < \Delta_2$, where Δ_1, Δ_2 are specified bounds, usually of the form $-\Delta_0, \Delta_0$. This can be achieved by combining the results of two one-sided tests: $\Delta \leq \Delta_1$ versus $\Delta > \Delta_1$ and $\Delta \geq \Delta_2$ versus $\Delta < \Delta_2$. To find the evidence for the alternative in the first test, we use the VST of $\hat{\Delta}$ derived in Kulinskaya et al. [11]. This is a family of VSTs indexed by a parameter $0 < A < 1$. For the choice $A = 1/2$ the nuisance parameter is $\psi = \bar{p} = (p_1 + p_2)/2$, and we also require $N = n_1 + n_2$, $v = (1 - 2\bar{p})(1/2 - n2/N)$ and $w = \sqrt{\bar{p}(1 - \bar{p}) + v^2}$. Then Equation 2.3 of [11], can be written:

$$T(\hat{\Delta}, \bar{p}, \Delta_1) = \sqrt{\frac{4n_1 n_2}{N}} \left(\text{arcsin}\left(\frac{\hat{\Delta}/2 + v}{w} \right) - \text{arcsin}\left(\frac{\Delta_1/2 + v}{w} \right) \right). \tag{8}$$

Reference [11] show that the statistic $T(\hat{\Delta}, \hat{\bar{p}}, \Delta_1)$ obtained by replacing $\bar{p}$, v and w by their plug-in estimates, is for large n_1, n_2 normally distributed with mean that is monotone increasing in Δ from 0 at the null $\Delta = \Delta_1$. Further, this statistic has variance 1 at the null, which allows them to derived large-sample confidence intervals for Δ; these intervals are shown to be quite competitive for even small to moderate sample sizes in [13]. Next define

$$\begin{aligned} T_- &= T(\hat{\Delta}, \hat{\bar{p}}, \Delta_1) \\ T_+ &= -T(\hat{\Delta}, \hat{\bar{p}}, \Delta_2) \\ T &= \min\{T_-, T_+\}. \end{aligned} \tag{9}$$

T_- gives the putative evidence for the alternative $\Delta > \Delta_1$ for the null $\Delta \leq \Delta_1$, while T_+ gives the putative evidence for the alternative $\Delta < \Delta_2$ for the null $\Delta \geq \Delta_2$. We say "putative" because, as [11] point out, the variances of these statistics can stray far from 1 if Δ is not near the null. *However, the evidence T for equivalence $\Delta_1 < \Delta < \Delta_2$ is better behaved, and has standard error similar to that for the two one-sided Z-test evidence discussed in Section 2.* R scripts for computing (9) are in Appendix B.

Example 2. (Comparing methods of patient care.) As described in [18], the objective of a randomized trial was to determine whether a standard method of care for patients by doctors was comparable to nurse-practitioner care. For the first group, there were $n_1 = 225$ patients and of these $X_1 = 148$ were found to have adequate care. For the second group, of $n_2 = 167$ patients, $X_2 = 115$ were found to have adequate care. Letting p_1, p_2 be the probability of adequate care for the first, second methods and $\Delta = p_1 - p_2$, it was desired to test for "equivalence of treatments" defined by $|\Delta| \leq \Delta_0 = 0.1$. For these data, $\hat{\Delta} = -0.03$, $T_- = 1.461$, $T_+ = 2.719$ and $T = 1.461$, which is close to 1.65 with a standard error less than 1. That is, the evidence for equivalence is positive but weak. By way of comparison, Reference [18] found the p-value for the equivalence alternative $|\Delta| \leq \Delta_0 = 0.1$ to be 0.005, but in a later corrected analysis in [19] calculated it to be 0.07.

In order to obtain expected moderate evidence 3.3 ± 1 for equivalence $|\Delta| \leq \Delta_0 = 0.1$ in this setting when in fact there is near equivalence, one would need sample sizes in each group near 1000.

2.3. Evidence for Equivalence in Two One-Sided t-Tests

For the t-test, the equivalence test of $H_0 : |\mu| \geq \mu_0$ against the alternative $H_1 : |\mu| < \mu_0$ is based on n measurements of the differential effect. The t-statistic is a function of the estimated mean $\bar{y}$ and standard deviation s obtained from the sample. The null hypothesis is two-sided and its right-hand part is rejected if $S_+ = \sqrt{n}\,(\bar{y} - \mu_0)/s < qt_{n-1,\alpha}$, whereas the left-hand part is rejected if $S_- = \sqrt{n}\,(\bar{y} + \mu_0)/s > qt_{n-1,1-\alpha}$. In both of these expressions, $qt_{n-1,p}$ denotes the p-quantile of the corresponding t-distribution. Both parts of the null hypothesis must be rejected in order to get significant evidence for equivalence. This holds if the confidence interval $[\bar{y} \pm qt_{n-1,1-\alpha}s/\sqrt{n}]$ is contained within $[\pm\mu_0]$.

The t-statistic S, if the true mean is μ, has a non-central t-distribution with $n-1$ degrees of freedom and noncentrality parameter $\lambda = \sqrt{n}(\mu - \mu_0)/\sigma$. We are interested in evidence in favor of small $|\mu|$. For the left-hand part, the non-centrality parameter is $\lambda = \sqrt{n}(\mu + \mu_0)/\sigma$ and we are interested in evidence in favor of large μ. The VST is derived in [9,20] and is defined by $h(S) = \sqrt{2n}\,\sinh^{-1}\left(S/\sqrt{2n}\right)$, where $\sinh^{-1}(x) = \ln(x + \sqrt{x^2+1})$. This is an increasing function and measures the evidence in favor of large μ. The evidence contained in the data in favor of equivalence for the right-hand part of the null hypothesis is thus, $-h(S_+)$, while for the left-hand part it is $h(S_-)$. Both need to be sufficiently large in order to conclude in favor of equivalence, that is, the empirical evidence

$$\hat{E} = \min\left\{-\sqrt{2n}\,\sinh^{-1}\left(\frac{\bar{y} - \mu_0}{\sqrt{2}s}\right), \sqrt{2n}\,\sinh^{-1}\left(\frac{\bar{y} + \mu_0}{\sqrt{2}s}\right)\right\}$$

must be at least 2 and better 3. Negative values of the empirical evidence can occur and they have to be interpreted as evidence in favor of non-equivalence.

Example 3. Figure 4 shows a plot of the empirical evidence as a function of the average of the measurements. The evidence from the t-statistic is nearly linear in $\bar{y}$ and largest if $\bar{y}$ is exactly halfway between the equivalence limits. The difference with the usual statistical tests is striking. There, the evidence will grow with the distance from the null hypothesis and can become arbitrarily large. Here, the maximal size of the evidence is limited by the equivalence limits.

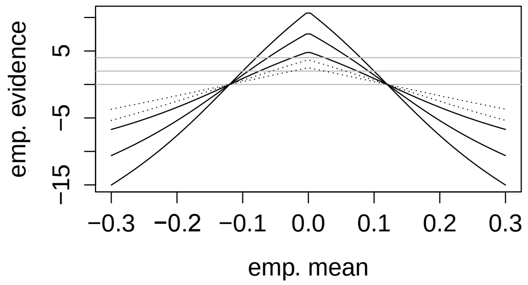

Figure 4. The solid curves show the evidence in favor of equivalence as a function of the average $\bar{y}$. It is assumed that the equivalence limits are $\mu_0 = \pm 0.12$ and the empirical standard deviation is $s = 0.1$. As the sample size grows from $n = 20$ to $n = 40$ and $n = 100$, the evidence grows. The dotted lines are for $n = 40$ and show the decreasing evidence if $s = 0.2$ and $s = 0.3$. The horizontal grey lines are at 0, 2 and 4.

The behavior of the evidence is as expected. If the sample size grows, so does the amount of evidence. If the standard deviation grows, there is less evidence if all other conditions are the same. The amount of evidence is bigger than a desired amount (2 or 4, for example), if $\bar{y}$ is within an interval centered at the halfway mark between the equivalence limits.

Approximate Normality of the Variance Stabilized *t*-Statistic

The VST is symmetric with regard to the origin, because $(x+\sqrt{x^2+1}) = 1/(-x+\sqrt{x^2+1})$, that is, $\ln(x+\sqrt{x^2+1}) = -\ln(-x+\sqrt{x^2+1})$. The expansion

$$h(S) = S - \frac{1}{6}\frac{S^3}{2n} + \frac{3}{40}\frac{S^5}{4n^2}O(n^{-3}),$$

shows that for values of S up to order $O(n^{1/3})$, the deviation from the identity is small. Only for values of the t-statistic S further into the tail does the VST pull them towards zero, that is, $h(S) < S$. For very large values of S the function $h(S)$ is logarithmic. The tail of the t-density evaluated at x is $O(x^{-n})$ as $x \to \infty$ and thus has a tail index of $n-1$. The VST transforms this to an infinite tail index.

3. Evidence in Multivariate Equivalence Tests

Multivariate equivalence tests are often based on a test statistic S having an exact or approximate non-central chi-squared distribution, denoted $S \sim \chi^2_\nu(\lambda)$, where ν is the known degrees of freedom (*df*), and the non-centrality parameter (*ncp*) $\lambda \geq 0$ is unknown. Others are based on the non-central F distribution, see Section 4. The null hypothesis postulates non-equivalence between the samples, $\lambda \geq \lambda_0$, whereas the alternatives postulate practical equivalence $\lambda < \lambda_0$. The limit λ_0 is a positive constant adapted to the context; examples are given in [17] and the following sections.

Wellek [8,20] looks at the case of possibly dependent measurements (K of them) that are done independently on n subjects. He then wants to test whether the K measurements have equal means. His first proposal is to pass to the $K-1$ differences between the K measurements and to use Hotelling's T^2 test for 0 means. He then remarks on the elliptical shape of the equivalence region, which might be criticized as being arbitrary. Reference [20] then discusses rectangular regions, as we do, and comments on the difficulties of this approach. The material in Section 3.2 below proposes a possible compromise solution. A fully Bayesian approach to multivariate equivalence testing is found in [5].

3.1. A VST for the Non-Central Chi-Squared Statistic

Once the "equivalence limit" λ_0 is chosen, one can carry out a Neyman–Pearson test which rejects non-equivalence $\lambda \geq \lambda_0$ at level α in favor of equivalence when the test statistic is sufficiently small, that is, S less than the α-quantile of the $\chi^2_\nu(\lambda_0)$ distribution ($c_\alpha = \chi^2_{\nu,\alpha}(\lambda_0)$). The power function of this test is the probability of deciding in favor of equivalence

$$1-\beta(\lambda) = P_\lambda(S \leq c_\alpha), \qquad 0 \leq \lambda < \lambda_0\,, \tag{10}$$

where $\beta(\lambda)$ is the probability of falsely reaching a conclusion of non-equivalence.

The testing approach underlying (10) is easier to understand when the test statistic is variance stabilized. In this context a VST is a monotone *decreasing* function $h(S)$ of the test statistic S, which for all values of λ is approximately normal with variance one. Rather than summarizing the evidence by an accept/reject decision, by a p-value or by a confidence interval, we propose to use the statistic $T = h(S) - h(\mathrm{E}_{\lambda_0}[S])$ because it provides a more informed and interpretable measure of the *evidence in favor of equivalence.* The larger its value, the more evidence resides in the data in favor of equivalence. Since its variance remains close to one for all λ, it is only the value of T that matters. The expected evidence $\mathrm{E}_\lambda[T]$ is

$$\mathcal{K}_{\lambda_0}(\lambda) = \mathrm{E}_\lambda[T] \approx h(\mathrm{E}_\lambda[S]) - h(\mathrm{E}_{\lambda_0}[S])\,. \tag{11}$$

This is a quantity that increases monotonically as λ *decreases* to 0. By construction, it is 0 at the equivalence limit λ_0 and has a maximal value of $\mathcal{K}_{\lambda_0}(0)$. The observed evidence for equivalence can be reported as $T \pm 1$, indicating that evidence T for equivalence has a standard normal error.

One can derive an approximation for $\mathcal{K}_{\lambda_0}(\lambda) = \mathrm{E}_\lambda[T(S)]$ for the evidence in $S \sim \chi^2_\nu(\lambda)$ for testing $\lambda \geq \lambda_0$ against $\lambda < \lambda_0$ as follows. The variance $\mathrm{Var}_\lambda[S] = 2\nu + 4\lambda$ is not constant (not stable) and in order to stabilize it asymptotically one can use the standard delta-method [9] (p. 242). Using the fact that $\mathrm{E}_\lambda[S] = \nu + \lambda \geq \nu$, one has $\mathrm{Var}_\lambda[S] = g(\mathrm{E}_\lambda[S])$, where $g(s) = 4s - 2\nu$ is defined for $s \geq \nu$. The transformation $h(s)$ that removes the dependence on λ is equal to an antiderivative of $-1/\sqrt{g(s)}$, which leads to $h(s) = -\sqrt{s - \nu/2}$, where the negative sign was chosen in order to obtain a decreasing function in s. This standard procedure fails to define a VST for all $s \geq 0$, because strictly speaking our function $h(s)$ should only be applied in the range $s \geq \nu$, and even if we tried to extend it towards $s = 0$, its value is undefined for $s < \nu/2$

This problem is created by the zero crossing of $g(s)$ at $s = \nu/2$ and various ideas for extending $h(s)$ to the entire positive real line could be tried; see (13) below. For this or any other such choice agreeing with $h(s)$ for $s > \nu$, the transformed statistic $h^*(S)$ has approximate expected value $\mathrm{E}_\lambda[h^*(S)] \doteq h^*(\mathrm{E}_\lambda[S]) = h^*(\nu + \lambda) \doteq h(\nu + \lambda) = -\sqrt{\lambda + \nu/2}$. After centering $h^*(S)$ at the limit λ_0, one obtains the observed evidence $T = h^*(S) + \sqrt{\lambda_0 + \nu/2}$. The expected evidence for equivalence, to first order, is thus

$$\mathcal{K}_{\lambda_0}(\lambda) = \mathrm{E}_\lambda[T] = \sqrt{\lambda_0 + \nu/2} - \sqrt{\lambda + \nu/2}\,. \tag{12}$$

The expected evidence (12) in the experiment has a maximum at $\lambda = 0$, namely $\mathcal{K}_{\lambda_0}(0) = \sqrt{\lambda_0 + \nu/2} - \sqrt{\nu/2}$. The dotted line in Figure 5 shows what the equivalence values λ_0 must be as a function of ν, namely $\lambda_0 = \mathcal{K}^2_{\lambda_0}(0) + \sqrt{2\nu}\,\mathcal{K}_{\lambda_0}(0)$, where the maximal expected evidence is of varying strengths. For example, when $\nu = 15$, for moderate maximal expected evidence, λ_0 must be at least 29. More frequently, the equivalence bound λ_0 is determined by context, and the degrees of freedom ν determined by $\nu \geq (\lambda_0 - \mathcal{K}^2_{\lambda_0}(0))^2/\{2\mathcal{K}^2_{\lambda_0}(0)\}$.

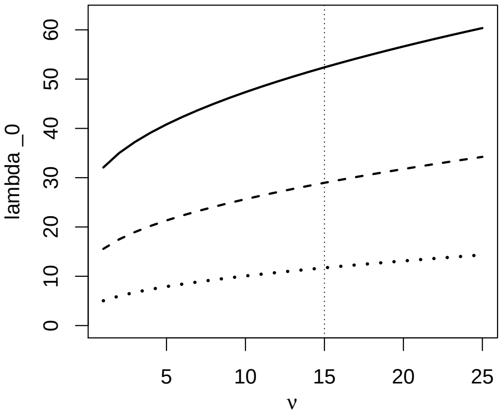

Figure 5. Plot of λ_0 against degrees of freedom ν based on (12) with $\lambda = 0$ for weak maximum expected evidence (dotted line), moderate maximum expected evidence (dashed line) and strong maximum expected evidence (solid line). The vertical line marks the df $\nu = 15$.

The evidence for $\lambda < \lambda_0$ in $S \sim \chi^2_\nu(\lambda)$ can be defined for certain $M > \mathcal{K}_{\lambda_0}(0)$:

$$T = T_{\nu,\lambda_0}(S) = \begin{cases} M - S/\sqrt{2\nu} & \text{for } S < \nu\,; \\ M - \sqrt{S - \nu/2} & \text{for } S \geq \nu\,. \end{cases} \tag{13}$$

This T has a negative continuous derivative, and for the choice $M = \sqrt{\lambda_0}$ has $T \approx N(0,1)$ at the null $\lambda = \lambda_0$. Further, at perfect equivalence $\lambda = 0$ it has $\mathrm{E}[T] \approx \mathcal{K}_{\lambda_0}(0)$ for $\mathcal{K}_{\lambda_0}$ given by (12). These claims are made based on simulation studies, using R scripts in Appendix B.

3.2. Evidence for Equal Means

Given independent $X_k \sim N(\mu_k, 1)$, $k = 1, \ldots, K$, we want to find evidence for equivalence in the sense that all means equal $\mu = (\sum_k \mu_k)/K$ simultaneously. A test statistic is $S = \sum_k (X_k - \bar{X})^2$, which has distribution $S \sim \chi^2_{K-1}(\lambda)$, where $\lambda = \sum_k (\mu_k - \mu)^2$. In practice, the μ_k are considered "equal" if $\max_k \{ |\mu_k - \mu| \} \leq \epsilon$ for a given $\epsilon > 0$. How can this last notion of equivalence be translated into a value for the limit λ_0? To solve this problem, we note that $\{(\mu_1 - \mu), \ldots, (\mu_K - \mu)\}$ defines a $K-1$ dimensional hyperplane of R^K, which when shifted to the origin is a $K-1$ dimensional subspace of R^K, and distances between points in the hyperplane are preserved under translation. Thus it suffices to solve the following problem for arbitrary $K \geq 2$: given $\mu_1, \ldots, \mu_K$ with $\mu = (\sum_k \mu_k)/K = 0$ and $\max_k \{ |\mu_k| \} \leq \epsilon$, find an appropriate choice of $\lambda_0 = \sum_k \mu_k^2 = r^2$, which is the square of the Euclidean distance r of the point $(\mu_1, \ldots, \mu_K)$ from the origin in R^K. After a solution for the equivalence boundary $\lambda_0 = \lambda_0(\epsilon, K)$ is found, it can be implemented in the case of unknown μ by replacing K by $K-1$.

The largest ball contained in the K-dimensional cube with edge 2ϵ, both having the same center, has radius ϵ, whereas the smallest ball containing the same cube has radius $\sqrt{K}\,\epsilon$. The latter choice would allow for one μ_k to be as large as $\sqrt{K}\,\epsilon$ (if all other $\mu_j = 0$) and overall equivalence would be claimed even though it was violated in one case. To reduce this violation, the radius ought to be in-between the extremes. A "reasonable" compromise requires the volume of the L_2-ball to equal $(2\epsilon)^K$, the volume of the cube with side 2ϵ. The volume of a ball in R^K with radius r is given by:

$$V_K(r) = \begin{cases} \frac{\pi^m \, r^{2m}}{\Gamma(m+1)}, & K = 2m; \\ \frac{\pi^m \, (2r)^{2m+1} \, \Gamma(m+1)}{\Gamma(2m+2)}, & K = 2m+1. \end{cases} \tag{14}$$

To ensure equal volumes for the hypersphere and the hypercube for all K, one needs the radius of the ball to be $r_0 = \sqrt{K b_K}\,\epsilon$, where

$$b_K = \begin{cases} \frac{4\,\{\Gamma(1+K/2)\}^{2/K}}{\pi K}, & K = 2m; \\ \frac{1}{\pi^{1-1/K}\,K} \left\{ \frac{\Gamma(K+1)}{\Gamma((K+1)/2)} \right\}^{2/K}, & K = 2m+1. \end{cases} \tag{15}$$

A good approximation to b_K is given by $c_K = 2/(\pi e) + (1.7 \cdot K)^{-5/6}$, see Table 2 and Figure 6. This makes it easier to choose the desired equivalence limit $\lambda_0 = r_0^2 = b_K K \epsilon^2 \sim 2K\epsilon^2/(\pi e)$ for moderate and large K.

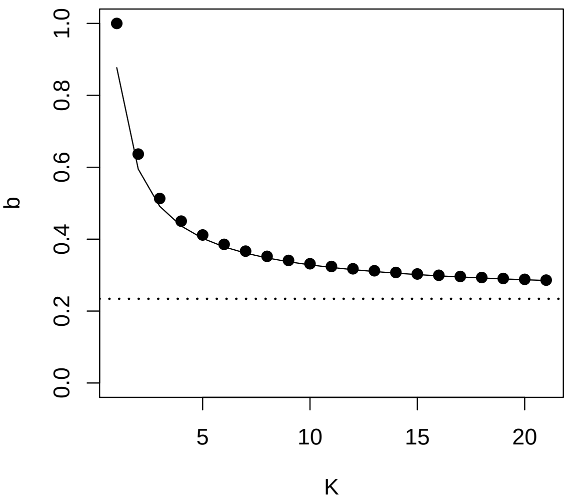

Figure 6. Plot of b_K (points) and approximation $c_K = 2/(\pi e) + (1.7\ K)^{-5/6}$ (continuous line) against K ranging from 0 to 21. The dotted horizontal line gives the asymptotic limit $2/(\pi e)$.

Table 2. For selected values of K are shown the exact b_K coefficients (15) to 3 decimal places so that the K-dimensional ball of radius $\sqrt{K\,b_K}\,\epsilon$, has the same volume as the K-dimensional cube of side 2ϵ. The approximate value is $c_K = 2/(\pi e) + (1.7 \cdot K)^{-5/6}$.

K	2	3	4	5	6	7	8	10	50	100	∞
b_K	$2/\pi$	0.513	0.450	0.412	0.386	0.367	0.352	0.332	0.259	0.248	0.23420...
c_K	0.595	0.492	0.437	0.402	0.379	0.361	0.348	0.329	0.259	0.248	$2/(\pi e)$

Example 4. To illustrate the use of this table, suppose $K = 4$, μ is unknown, and we want each of μ_k to satisfy $|\,\mu_k - \mu\,| \le \epsilon = 1/2$, say, in order to claim "equivalence of all means". As discussed above the corresponding problem with μ known to equal 0 in $K-1 = 3$ dimensions is to utilize $\lambda_0 = \lambda_0(\epsilon, K-1) = \lambda_0(1/2, 3) = r_0^2 = 3\,b_3\,\epsilon^2 = 3 \times 0.513 \times 0.25 = 0.3375$. From Equation (12) it then follows that for $K = 4$, $\nu = 3$ and this λ_0 the maximum expected evidence possible for equivalence of all 4 means is about 0.13, which is almost negligible.

If we had begun this section with each X_k replaced by $\bar{X}_k \sim N(\mu_k, 1/n)$, for some sample size $n \ge 2$, then the test statistic would be $S_n = \sum_k (\bar{X}_k - \bar{\bar{X}})^2$, which has distribution $S_n \sim \chi^2_\nu(\lambda)$, where $\nu = K-1$ and $\lambda = n\sum_k(\mu_k - \mu)^2$. By imposing the same condition $\max_k \{\,|\,\mu_k - \mu|\,\} \le \epsilon$, where μ is unknown, the "equal-volume" solution $\sum_k(\mu_k - \mu)^2 \le= \nu b_\nu\,\epsilon^2$, so the appropriate equivalence hypothesis is $\lambda \le \lambda_0 = n\,\nu\,b_\nu\,\epsilon^2$. The maximum expected evidence in $T_n = T(S_n)$ is attained when all means are equal, and this maximum is $\sqrt{\lambda_0 + \nu/2} - \sqrt{\nu/2}$, or $\sqrt{n\,\nu\,b_\nu\,\epsilon^2 + \nu/2} - \sqrt{\nu/2}$ which is growing at rate $\sqrt{n}$. Continuing Example 4, $n = 20$ will yield weak maximum expected evidence 1.65 for equivalence.

3.3. Application to between Group Sum of Squares

Given independent observations from K groups with different means $X_{ki} \sim N(\mu_k, 1)$, $i = 1, \ldots, n_k$, $k = 1, \ldots, K$, we denote the total sample size by $N = \sum_k n_k$, and the group sample proportions by $q_k = n_k/N$. Let the kth sample mean be $\bar{X}_k$ and the weighted group sample mean by $\bar{X} = \sum_k q_k \bar{X}_k$; it is an unbiased estimator of the weighted population mean $\mu = \sum_k q_k \mu_k$. Then the between group sum of squares $SSB_{\text{between}} = N\sum_k q_k(\bar{X}_k - \bar{X})^2 \sim \chi^2_\nu(\lambda)$, where $\nu = K-1$ and $\lambda = N\sum_k q_k(\mu_k - \mu)^2$; see, for example, [9] (p. 184). The evidence in $S \sim \chi^2_\nu(\lambda)$ for equivalence $\lambda < \lambda_0$ was derived in Section 3.1.The practical problem is to choose λ_0. Once λ_0 is chosen one can compute the maximum evidence in the experiment; it is by (12) equal to $\sqrt{\lambda_0 + (K-1)/2} - \sqrt{(K-1)/2}$. As explained in Section 3.1, this determines the maximum power for equivalence at any given level.

First we consider the argument of [17] (p. 164) for choosing λ_0. He introduces the parameter $\psi^2 = \sum_k \frac{n_k}{\bar{n}}(\mu_k - \mu)^2$, where $\bar{n} = N/K$, which he calls a *generalized squared Euclidean distance* between $(\mu_1, \ldots, \mu_K)$ and $(\mu, \ldots, \mu)$. He proposes to define homogeneity in terms of ψ^2 and the equivalence hypothesis by $\psi^2 \le \epsilon^2$, where ϵ is to be chosen. Note that our *ncp* $\lambda = \bar{n}\psi^2$, the same as his. For $K = 2$ condition $\psi^2 \le \epsilon^2$ reduces to $|\mu_2 - \mu_1| \le \sqrt{2}\,\epsilon$. This leads [17] (p. 164) using conditions for comparing 2 normal populations, to suggest that, in general, one take $\lambda_{0,Wellek} = \bar{n}\epsilon^2$, with ϵ ranging from 1/4 to 1/2 where 1/4 yields a "strict" equivalence limit $\bar{n}/16$ while $\epsilon = 1/2$ leads to what he calls a "liberal" limit of $\bar{n}/4$. However, this approach assumes that the requirement $\max_k \{\,|\,\mu_k - \mu|\,\} \le \epsilon$ for a pre-specified ϵ holds for all K. From our point of view the choice of λ_0 should grow with $\sqrt{K}$ and ϵ should be determined by context.

In the balanced case $n_k \equiv n$, we can make direct comparisons between Wellek's criterion and ours, derived near the end of Section 3.2, which yielded $\lambda_{0,MS} = n\,\nu\,b_\nu\,\epsilon^2$. Assuming the same choice of ϵ, the ratio of $\lambda_{0,MS}/\lambda_{0,Wellek} = \nu\,b_\nu$. Using Table 2, this ratio varies with $\nu = 1, 2, 3, \ldots$ considerably and equals, respectively, $0.64, 1.03, 1.35, 1.65, 1.93, \ldots$; further it grows with ν as $2\nu/(\pi\,e)$. Only for $\nu = 2$ are the two criteria the same.

4. Testing for Equivalence of K Groups

Given independent observations from K groups with common unknown variance $\sigma^2 > 0$ and different means $X_{ki} \sim N(\mu_k, \sigma^2)$, $i = 1, \ldots, n_k$, $k = 1, \ldots, K$. Let $N = \sum_k n_k$, $q_k = n_k/N$ for all k and the overall mean $\mu = \sum_k q_k \mu_k$. We define *equivalence of means* for a given $\epsilon > 0$ if none of the standardized μ_k differs more than ϵ from zero:

$$\max_k \left\{ \frac{|\,\mu_k - \mu|}{\sigma} \right\} \le \epsilon. \tag{16}$$

Just as in Section 3.3 where σ was assumed known (and without loss of generaliity set equal to 1), we can define $\lambda = N \sum_k q_k \{(\mu_k - \mu)/\sigma\}^2$. Given λ_0 we want evidence for the hypothesis of equivalence $\lambda < \lambda_0$. The arguments for choosing $\lambda_0 = \lambda_{0,MS} = n\,\nu_1\, b_{\nu_1}\, \epsilon^2$ where $\nu_1 = K - 1$ carry over from Sections 3.2 and 3.3, provided sampling is nearly balanced so that all $n_k \approx \bar{n} = n$.

The *within* group sum of squares is defined by

$$SS_{\text{within}} = \sum_{k-1}^{K} \sum_{i=1}^{n_i} (X_{ki} - \bar{X}_k)^2. \tag{17}$$

Standard theory [9] (p. 196) shows $S = (S_{\text{between}}/\nu_1)/\{SS_{\text{within}}/\nu_2\}$ has a non-central F distribution with *df* $\nu_1 = K - 1$, $\nu_2 = N - K$ and ncp λ. A VST for the statistic S has been derived by [21] and also [9] (p. 197). It assumes $\nu_2 > 4$. Let $a^2 = (\nu_2 - 4)/2$ and $c^2 = \nu_2^2(\nu_1 + \nu_2 - 2)/\{\nu_1^2(\nu_2 - 2)\}$. The VST is $h = h(S)$, defined by $h(s) = \;-a \cosh^{-1}\left((s + \nu_2/\nu_1)/c\right)$. where the inverse hyperbolic cosine function is defined by $\cosh^{-1}(x) = \ln(x + \sqrt{x^2 - 1}\,)$. Now $\cosh^{-1}(x)$ is only defined for $|x| \le 1$ so the VST $h(s)$ is only defined for $s > c - \nu_2/\mu_1$.

The evidence for $\lambda < \lambda_0$ is defined by $T = h(S) - h(\mathrm{E}_{\lambda_0}[S])$. The expected value of the non-central F-statistic S is $\mathrm{E}_\lambda[S] = (\nu_1 + \lambda)\nu_2/\{\nu_1(\nu_2 - 2)\}$, so the expected evidence, to first order, is given by:

$$\mathcal{K}_{\lambda_0}(\lambda) = \mathrm{E}_\lambda[h(S)] - h(\mathrm{E}_{\lambda_0}[S]). \tag{18}$$

Unfortunately the VST itself and hence the evidence $T = h(S) - h(\mathrm{E}_{\lambda_0}[S])$ for equivalence is undefined for informative small values of S. To make it useful, we extend the evidence function to small values via monotone linearization in (19). Strictly speaking, the F-statistic VST was only derived for $s > b = \nu_2/(\nu_2 - 2)$, as mentioned in [9] (p. 197), for the same reason the chi-squared VST derivation had a limited domain, see Section 3.1. And, for the same reasons given there, we can extend it to $0 \le s \le b = \nu_2/(\nu_2 - 2)$ without changing much the expectation (18). Evidence for equivalence $\lambda < \lambda_0$ based on $S \sim F_{\nu_1, \nu_2, \lambda}$ is defined for certain $M > \mathcal{K}_{\lambda_0}(0)$ by

$$T = T_{\nu_1, \nu_2, \lambda_0}(S) = \begin{cases} M - \;a\, S \cosh^{-1}\left((b + \nu_2/\nu_1)/c\right)/b & \text{for } S < b\,; \\ M - \;a\, \cosh^{-1}\left((S + \nu_2/\nu_1)/c\right) & \text{for } S \ge b\,. \end{cases} \tag{19}$$

This T is continuous and increasing as S moves to 0. For the choice $M = \sqrt{\lambda_0 - \nu_1/(\nu_1 + \nu_2)}$ it has $T \approx N(0,1)$ at the null $\lambda = \lambda_0$; and, at perfect equivalence $\lambda = 0$ it has $\mathrm{E}[T] \approx \mathcal{K}_{\lambda_0}(0)$ for the $\mathcal{K}_{\lambda_0}$ of (18).

Example 5. Example 7.1 of [17] (p. 165) considers four treatments for hypertension with measurements taken on diastolic blood pressure averaged over an interval. To test for equivalence of the treatments, the following summary data (sample size, mean, standard deviation) were recorded: $n_1 = 10$, $\bar{x}_1 = 99.8120$, $s_1 = 7.56391$; $n_2 = 12$, $\bar{x}_2 = 99.2903$, $s_2 = 5.9968$; $n_3 = 13$, $\bar{x}_3 = 100.0024$, $s_3 = 10.4809$; and $n_4 = 15$, $\bar{x}_4 = 98.6407$, $s_4 = 4.5309$. This yields $N = 50$, $\bar{n} = 12.5$ and $\bar{\bar{x}} = 99.3849$. Continuing, $SS_{between} = 15.196$ and $SS_{within} = 2516.135$ so the F-test statistic $S = (SS_{between}/$

$(K-1))/(SS_{within}/(N-K)=0.0926$. For $\epsilon=0.5$ and $\lambda_{0,MS}=\bar{n}\,\nu_1 b_{\nu_1}\epsilon^2=4.219$ we have by (19) the evidence for equivalence $T=1.934$, which is slightly more than weak.

By way of comparison, [17] uses traditional hypothesis testing with the smaller equivalence boundary $\lambda_{0,Wellek}=\bar{n}\epsilon^2=3.125$ and finds the non-central F-test significant at level 0.05 with an estimated power 0.18 for detecting perfect equivalence.

5. Summary and Discussion

For the test statistic $S \sim N(\mu,1)$ of a null hypothesis $\mu \le \mu_0$ and alternative $\mu > \mu_0$ Neyman–Pearson methods help one make a decision; while a p-value can provide a measure of surprise regarding the boundary hypothesis $\mu=\mu_0$. But what one often wants from S is a measure of evidence *for* the alternative $\mu>\mu_0$. In this simplest of statistical tests, $T=S-\mu_0$ is such a measure of evidence for the alternative, because T estimates the unknown expected evidence $\mathrm{E}[T]=\mu-\mu_0$ which is linearly increasing with μ and comes with an easily understood standard normal error. Values of T near 1.645, 3.3 and 5 are interpreted as weak, moderate and strong evidence for the alternative $\mu>\mu_0$. In addition the expected evidence can be written as the sum of the probits of level and power.

The vast majority of routine statistical tests can be transformed into the above setting through variance stabilization. And the mean of a variance stabilized test statistic $T=\mathrm{VST}(S)$, after centering at μ_0, is very close to the signed square root of the Kullback–Leibler symmetrized divergence between the null and alternative distributions. This result gives more theoretical support for calling T the *evidence for the alternative*; references are listed in Section 1.

What we have done here is to extend the above ideas to the more complicated hypotheses of the form $|\mu|\ge\mu_0$ versus the equivalence alternative $|\mu|<\mu_0$, with applications for TOST based on one- and two-sample binomial experiments and two one-sided t-tests. Then, in the multivariate setting, we have found modifications of the classical VSTs for non-central chi-squared and non-central F-tests which make it practicable to find evidence for the equivalence alternative $\lambda<\lambda_0$ to the null hypothesis $\lambda\ge\lambda_0$ of non-equivalence.

The practical choice of equivalence limit λ_0 is also an important ingredient, and we have provided a new approach to assist in its choice. In particular, when testing for equivalence of means in K arms of a study, we found the value of the radius required so that the K-ball has the same volume as the K-cube of edge 2ϵ. This leads to a proposal for converting the condition $\max_k|\mu_k-\mu|\le\epsilon$ into an approximate equivalence condition $\lambda<\lambda_0$.

The new extensions of classical VSTs for non-central chi-squared and F statistics require more work in the choice of M to center them properly at the null λ_0. Simulation studies show that adjusting M so that the mean of the VST is 0 when $\lambda=\lambda_0$ automatically ensures that the mean of the VST statistic when $\lambda=0$ is near its expected maximum value. The choice $M\approx\sqrt{\lambda_0}$ works adequately, but simple formulae for M that depend on the degree(s) of freedom as well as λ_0 would be useful.

Finally, we note that finding simple expressions for the expected evidence in a VST greatly assists one in finding minimal sample sizes when planning an experiment for determining equivalence; by knowing the maximum expected evidence for equivalence, one also learns of the power to detect perfect equivalence at any given level.

Another application is in goodness-of-fit tests, where instead of “backing into” a model by not rejecting it a liberal level such as 0.1, one could find the evidence for the model. Chapter 8 of [17] is a good starting point for solving this problem. Further research will also take into account recent results on bio-equivalence defined in terms of the Kullback–Leibler divergences, see [4,7].

Acknowledgments: The authors thank the Editors and referees for their many helpful comments and suggestions which have improved the breadth and clarity of the manuscript.

Author Contributions: The two authors collaborated on both the research for, and writing of this manuscript. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Quality of KLD Approximations

Let S have density f_λ belonging to a family of densities indexed by a real parameter λ. Our purpose in this section is to compare the expected evidence $\mathcal{K}_{\lambda_0}(\lambda)$ of the VST for testing $\lambda \geq \lambda_0$ against the equivalence alternative $\lambda < \lambda_0$ with the signed square root of the [22] symmetrized divergence (KLD) between the null f_{λ_0} and alternative f_λ distributions. That is, to examine the approximation

$$\mathcal{K}_{\lambda_0}(\lambda) \approx \operatorname{sgn}(\lambda_0 - \lambda)\,\sqrt{\mathcal{J}(\lambda_0, \lambda)}, \tag{A1}$$

where the KLD is defined by

$$\mathcal{J}(\lambda_0, \lambda) = \mathrm{E}_{\lambda_0}[\log\{f_{\lambda_0}(S)/f_\lambda(S)\}] + \mathrm{E}_\lambda[\log\{f_\lambda(S)/f_{\lambda_0}(S)\}]. \tag{A2}$$

References [10,12,13] discuss the theory behind such approximations, but here we consider numerical examples relevant to multivariate equivalence testing.

Appendix A.1. Non-Central Chi-Squared Distribution

Let f_λ be the noncentral χ^2 density with ν degrees of freedom, so that by (12), $\mathcal{K}_{\lambda_0}(\lambda) = \mathrm{E}_\lambda[T] = \sqrt{\lambda_0 + \nu/2} - \sqrt{\lambda + \nu/2}$. There is no simple analytic expression for $\mathcal{J}_{\chi^2}(\lambda_0; \lambda)$, so we use numerical approximation to compute it and its signed square root. The graph of the latter (for $\nu = 1$, $\lambda_0 = 6$) is shown in the top left plot of Figure A1 as a dashed line, and is to be compared with the graph of the $\mathcal{K}_6(\lambda)$ defined by (12) as a solid line. Note that they are very close over the range of λ of interest although both pass through 0 at $\lambda = \lambda_0$. Our main point is to emphasize the quality of the approximation (A1), which is clear from the plot to its right which shows the absolute *relative* error is less than 1 in 16 over this range of λ.

Figure A2 and others not shown demonstrate that quite generally the absolute relative error in the approximation (A1) is less than 1 in 20 over a wide range of equivalence experiments with non-central chi-squared distributed outcomes.

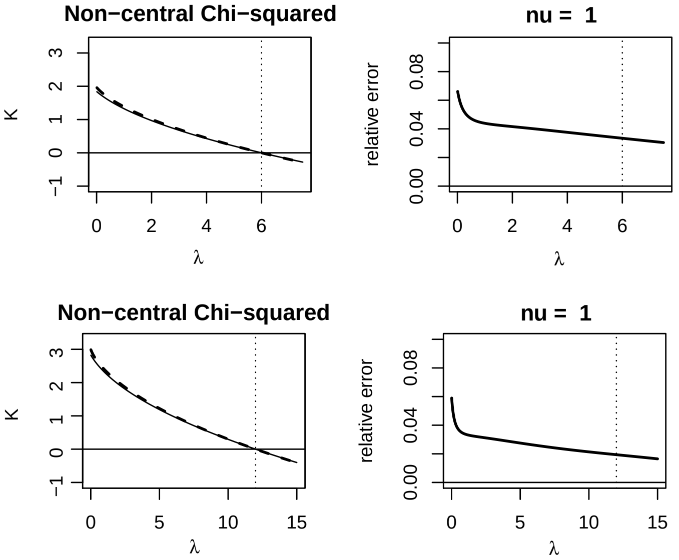

Figure A1. Non-central Chi-squared with parameters ν, λ: The upper left plot shows the graph of the expected evidence $\mathcal{K}_6(\lambda)$ when $\nu = 1$ as a solid line, to be compared with the signed square root of the KLD, shown as a dashed line. The absolute relative error in this approximation is shown on its right. The lower two plots are for $\lambda_0 = 12$ and $\nu = 1$.

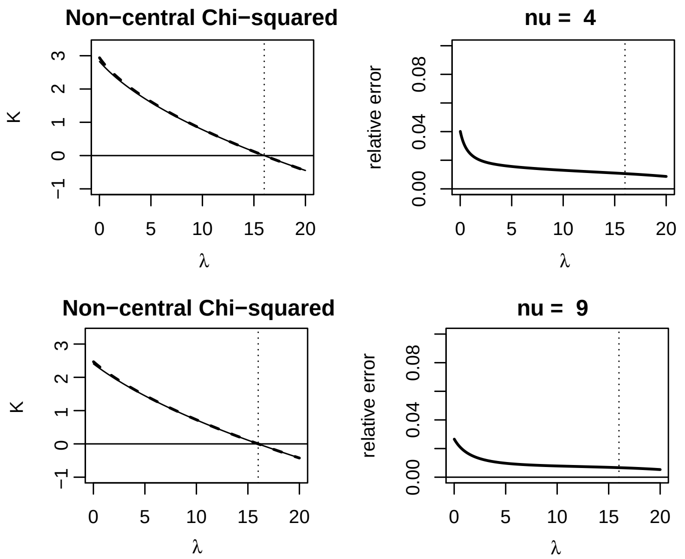

Figure A2. The same notation as in Figure A1, but for different parameters.

Appendix A.2. Non-Central F Distribution

If f_λ denotes the noncentral F density with ν_1, ν_2 degrees of freedom and $ncp\lambda$, then by (18), and using $\mathrm{E}_\lambda[S] = (\nu_1 + \lambda)\nu_2 / \{\nu_1(\nu_2 - 2)\}$,

$$\mathcal{K}_{\lambda_0}(\lambda) = \mathrm{E}_\lambda[T] = a \cosh^{-1}\left((\mathrm{E}_{\lambda_0}[S] + \nu_2/\nu_1)/c\right) - a \cosh^{-1}\left((\mathrm{E}_\lambda[S] + \nu_2/\nu_1)/c\right).$$

As for the chi-squared case, there is no simple analytic expression for $\mathcal{J}_F(\lambda_0; \lambda)$, so we use numerical approximation to compute it and its signed square root. Figures A3 and A4 show typical comparative results between the above expected evidence and the signed square root of the KLD between the null f_{λ_0} and alternative f_λ distributions. These and similar plots provide more numerical support for the approximation (A1).

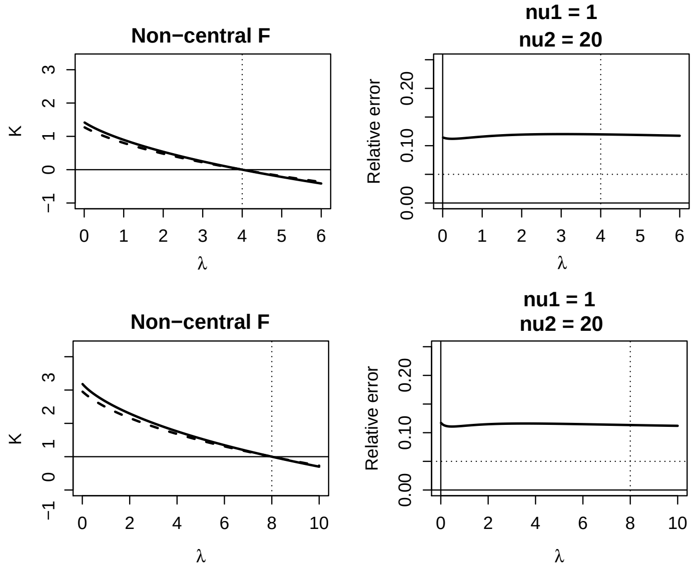

Figure A3. Non-central F with parameters ν_1, ν_2 and λ: The upper left plot shows the graph of the expected evidence $\mathcal{K}_{10}(\lambda)$ when $\nu_1 = 1, \nu_2 = 20$ as a solid line, to be compared with the signed square root of the KLD, shown as a dashed line. The absolute relative error in this approximation is shown on its right. The lower two plots are also for $\lambda_0 = 10$ but now $\nu_1 = 4$ and $\nu_2 = 20$.

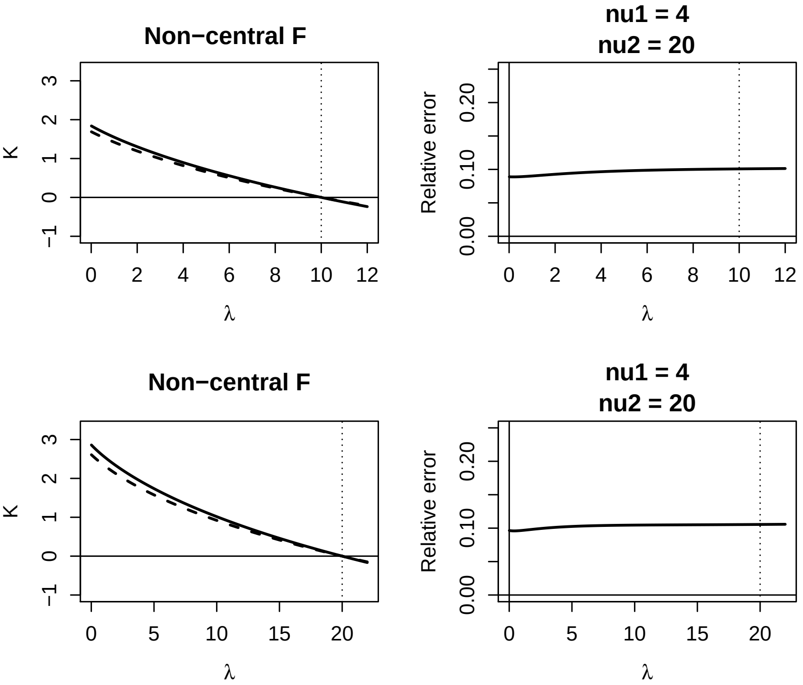

Figure A4. The same notation as in Figure A3, but for different parameters.

Appendix B. R Scripts for Computing VSTs

```
#############  Evidence for equivalence
#############  using TOST, and one-sample binomial data
vstbinom <- function(n,p)
{h <- 2*sqrt(n)*asin(sqrt(p))
return(h)}

############# Usually p1=p0-Delta0,p2=p0+Delta0
bioevid <- function(n,p1,p2,phat)
{Tminus <-  vstbinom(n,phat)-vstbinom(n,p1)
Tplus  <-  vstbinom(n,p2)-vstbinom(n,phat)
evidforequiv <- min(Tminus,Tplus)
out <- c(Tminus,Tplus,evidforequiv)
outrd <- round(out,digits=3)
return(outrd)}

#############  Example 1 of the text. (data from Example 4.2 of Weller, page 59.)
n <- 273
phat <- 199/273
p1 <- 0.65
p2 <- 0.75
bioevid(n,p1,p2,phat)

############################################################################
############# Evidence for equivalence for risk difference

vstRD <- function(n1,p1hat,n2,p2hat,Delta0)
{Deltahat <- p1hat-p2hat
pbarhat <- (p1hat+p2hat)/2
N <- n1+n2
vhat <- (1-2*pbarhat)*(1/2-n2/N)
what <- sqrt(pbarhat*(1-pbarhat)+vhat^2)
vst <- sqrt(4*n1*n2/N)*(asin((Deltahat/2 +vhat)/what)-asin((Delta0/2 +vhat)/what))
return(vst)}
```

```
############## Usually Delta1= -Delta0 and Delta2= +Delta0
bioevidRD <- function(x1,n1,x2,n2, Delta1,Delta2)
{p1hat <- (x1+0.5)/(n1+1)
p2hat <- (x2+0.5)/(n2+1)
Deltahat <- p1hat-p2hat
Tminus <-  vstRD(n1,p1hat,n2,p2hat,Delta1)
Tplus  <-  -vstRD(n1,p1hat,n2,p2hat,Delta2)
T <- min(Tminus,Tplus)
out <- c(Deltahat,Tminus,Tplus,T)
outrd <- round(out,digits=3)
return(outrd)}

####### Example 2 of the text. (data from Dunnett and Gent(1977) Biometrics)
x1 <- 148
n1 <- 225
x2 <- 115
n2 <- 167
Delta1 <- -0.1
Delta2 <- 0.1
bioevidRD(x1,n1,x2,n2, Delta1,Delta2)

###########################################################################
## Linear extension of vst for chisq(nu,lambda) (Equation (13) of the text.)

evidchisqlin <- function(s,nu,lambda0,M)
{smalls <- s[s <= nu]
Tsmall <- M-sqrt(s[s>nu]-nu/2)  ## usual vst
grad <- -sqrt(nu/2)/nu
Tbig <- M +grad*s[s<=nu]
T <-  c(Tbig,Tsmall)
return(T)}

Mfun <- function(nu,lambda0)    ## requires lambda0 > nu/2
{return(sqrt(lambda0))}

# Mfun2 <- function(nu,lambda0) ##  The user may want to supply their own M.
# {return(sqrt(lambda0+nu)-sqrt(nu/10))}

############### Plot evidence function (illustrative example)
nu = 3
lambda0 = 6
maxexev <- sqrt(lambda0+nu/2)-sqrt(nu/2)
maxexev
M <- Mfun(nu,lambda0)
M
s <- c(seq(0,lambda0+nu,.01))
T <-  evidchisqlin(s,nu,lambda0,M)
plot(s,T,type="l",lwd=2,main="Evidence for equivalence")
abline(h=maxexev,lty=3)
abline(h=0)

##################### To examine properties of evidence function.
lambda = lambda0   ## Try this and other lambda, where 0 <= lambda < lambda0.
s <- rchisq(10000,nu,lambda)
T <- evidchisqlin(s,nu,lambda0,M)
mean(T)
sd(T)
hist(T)

###########################################################################
## Linear extension of VST for F(nu1,nu2,lambda) (Equation (19) of the text.)
```

```
evidFlin <- function(s,nu1,nu2,lambda0,M)   ### nu2>4
{c <- (nu2/nu1)*sqrt((nu1+nu2-2)/(nu2-2))
b <- nu2/(nu2-2)
a <- sqrt((nu2-4)/2)
Tsmall <-  M-a*acosh((s[s>b]+nu2/nu1)/c)
grad <- -a*acosh((b+nu2/nu1)/c)/b
Tbig <- M+grad*s[s<=b]
T <-  c(Tbig,Tsmall)
return(T)}

##################    Example 5 of the text. (data from Wellek, Example 7.2, p.165.)

n1 <- 10
x1bar <- 99.8120
s1 <- 7.5639
n2 <- 12
x2bar <- 99.2903
s2 <- 5.9968
n3 <- 13
x3bar <- 100.0024
s3 <- 10.4809
n4 <- 15
x4bar <- 98.6407
s4 <- 4.5309
N <- n1+n2+n3+n4  ##
xbarbar <- (n1*x1bar+n2*x2bar+n3*x3bar+n4*x4bar)/N  ## 99.3849
K <- 4
nbar <- N/K
ssqW <- (n1-1)*s1^2+(n2-1)*s2^2+(n3-1)*s3^2+(n4-1)*s4^2
ssqB <- n1*(x1bar-xbarbar)^2+n2*(x2bar-xbarbar)^2+n3*(x3bar-xbarbar)^2
ssqB <- ssqB+n4*(x4bar-xbarbar)^2

epsilon <- 1/2                 ## Wellek's choice
nu1 <- K-1
nu2 <- N-K
lambda0 <- 12.5*0.25*1.35      ## nbar*eps^2*vu*bnu = 4.21875

nu1 <- K-1
nu2 <- N-K

epsilon <- 1/2                 ## this is arbitrary; want max_k |mu_k-mu|/sigma < epsilon
lambda0 <- 12.5*0.25*1.35      ## nbar*eps^2*vu1*bnu1 = 4.21875

MFfun <- function(nu1,nu2,lambda0)
{return(sqrt(lambda0+nu1/(nu1+nu2)))}

M <- MFfun(nu1,nu2,lambda0)
S <-   (ssqB/nu1)/(ssqW/nu2)     ## 0.0926 (Value of F-statoistic.)

evidFlin(S,nu1,nu2,lambda0,M)    ## T = 1.934

################  To compute maximum expected evidence, require:

exS <- function(lambda,nu1,nu2)   ##### this computes expected value of S
{return(nu2*(n1+lambda)/(nu1*(nu2-2)))}

c <- (nu2/nu1)*sqrt((nu1+nu2-2)/(nu2-2))
b <- nu2/(nu2-2)
a <- sqrt((nu2-4)/2)
M <- sqrt(lambda0-nu1/(nu1+nu2))
M
maxexev <- -a*acosh((exS(0,nu1,nu2)+nu2/nu1)/c)+
a*acosh((exS(lambda0,nu1,nu2)+nu2/nu1)/c)
```

```
maxexev

############### Plot evidence function (illustrative example)

s <- c(seq(0,exS(lambda0,nu1,nu2)+1,.01))
T <-  evidFlin(s,nu1,nu2,lambda0,M)
plot(s,T,type="l",lwd=2,ylim=c(-1,M),main="Evidence for equivalence")
abline(h=maxexev,lty=3)
abline(h=0)
abline(v=lambda0,lty=3)

################################# To examine properties of evidence function
lambda=0 ## Try this and other lambda, where 0 <= lambda < lambda0.lambda = lambda0

s <- rf(10000,nu1,nu2,lambda)
T <-  evidFlin(s,nu1,nu2,lambda0,M)
mean(T)
sd(T)
hist(T)
```

References

1. Schuirmann, D.J. A comparison of two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharm. Biopharm.* **1987**, *15*, 657–680.
2. Berger, R.; Hsu, J.C. Bioequivalence trials, intersection-union tests and equivalence confidence sets with discussion. *Stat. Sci.* **1996**, *11*, 283–302.
3. Senn, S. Statistical issues in equivalence testing. *Stat. Med.* **2001**, *20*, 2785–2799.
4. Dragalin, V.; Fedorov, V.; Patterson, S.; Jones, B. Kullback–leibler divergence for evaluating bioequivalence. *Stat. Med.* **2003**, *22*, 913–930.
5. Lauretto, M.; Pereira, C.A.B.; Stern, J.M.; Zacks, S. Full Bayesian Signicance Test Applied to Multivariate Normal Structure Models. *Braz. J. Probab. Stat.* **2003**, *17*, 147–168.
6. Chervoneva, I.; Hyslop, T.; Hauck, W.W. A multivariate test for population bioequivalence. *Stat. Med.* **2007**, *26*, 1208–1223.
7. Ocaña, J.; Sanchez, M.P.; Sanchez, A.; Carrasco, J.L. On equivalence and bioequivalence testing. *Sort* **2008**, *32*, 151–158.
8. Tsai, C.A.; Huang, C.Y.; Liu, J.P. An approximate approach to sample size determination in bioequivalence testing with multiple pharmacokinetic responses. *Stat. Med.* **2014**, *33*, 3300–3317.
9. Kulinskaya, E.; Morgenthaler, S.; Staudte, R.G. *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
10. Morgenthaler, S.; Staudte, R.G. Advantages of variance stabilization. *Scand. J. Stat.* **2012**, *39*, 714–728.
11. Kulinskaya, E.; Morgenthaler, S.; Staudte, R.G. Variance stabilizing the difference of two binomial proportions. *Am. Stat.* **2010**, *64*, 350–356.
12. Morgenthaler, S.; Staudte, R.G. Evidence for alternative hypotheses. In *Robustness and Complex Data Structures*; Becker, C., Fried, R., Kuhnt, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 315–329.
13. Prendergast, L.A.; Staudte, R.G. Better than you think: Interval estimators of the difference of binomial proportions. *J. Stat. Plan. Inference* **2014**, *148*, 38–48.
14. Johnson, N.L.; Kotz, S.; Balakrishnan, N. *Continuous Univariate Distributions*, 2nd ed.; Wiley: New York, NY, USA, 1994; Volume 1.
15. Johnson, N.L.; Kotz, S.; Balakrishnan, N. *Continuous Univariate Distributions*, 2nd ed.; Wiley: New York, NY, USA, 1995; Volume 2.
16. Leone, F.C.; Nelson, L.S.; Nottingham, R.B. The folded normal distribution. *Technometrics* **1961**, *3*, 543–550.
17. Wellek, S. *Testing Statistical Hypotheses of Equivalence*; CRC Press: Boca Raton, FL, USA, 2003.
18. Dunnett, C.W.; Gent, M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics* **1977**, *33*, 593–602.
19. Johnson, R.; Dunnett, C.W.; Gent, M. p-values in 2 × 2 tables. *Biometrics* **1988**, *44*, 907–910.

20. Wellek, S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2014.
21. Laubscher, N.F. Normalizing the noncentral t and f distributions. *Ann. Math. Stat.* **1960**, *31*, 1105–1112.
22. Kullback, S. *Information Theory and Statistics*; Dover: Mineola, NY, USA, 1968.