Journal of Data Science 23 (1), 171–187
January 2025
DOI: 10.6339/24-JDS1142
Statistical Data Science

A Joint Equivalence and Difference (JED) Test for Practical Use in Controlled Trials

Robert H. Riffenburgh[1,*] and Lingge Wang[1]

[1]*Mathematics and Statistics Department, San Diego State University, San Diego, CA, USA*

Abstract

A joint equivalence and difference (JED) test is needed because difference tests and equivalence (more exactly, similarity) tests each provide only a one-sided answer. The concept and underlying theory have appeared numerous times, noted and discussed here, but never in a form usable in workaday statistical applications. This work provides such a form as a straightforward simple test with a step-by-step guide and possible interpretations and formulas. For initial treatment, it restricts attention to a t test of two means. The guide is illustrated by a numerical example from the field of orthopedics. To assess the quality of the JED test, its sensitivity and specificity are examined for test outcomes depending on error risk α, total sample size, sub-sample size ratio, and variability ratio. These results are shown in tables. Interpretations are discussed. It is concluded that the test exhibits high power and effect size and that only quite small samples show any effect on the power or effect size of the JED test by commonly seen values of any of the parameters. Data for the example and computer codes for using the JED test are accessible through links to supplementary material. We recommend that this work be extended to other test forms and multivariate forms.

Keywords *decision-making; error rate estimation; means testing; medical decisions; statistical testing*

1 Introduction

1.1 Goal

The goal of this paper is twofold, first to provide users of statistical methods with a *joint equivalence and difference* (JED) *test* in a form that is easy to use, including use by non-statisticians, and second, to show that this form is statistically dependable and sensitive enough for common use. The JED test replaces the difference-versus-no-evidence and the equivalence-versus-no-evidence tests simultaneously. No power is lost by using such a test. We propose that it become the default method in common usage.

The concept and theory of a JED test as addressed in this paper are not new. The specific form of a JED test, the guidance for practical application, the plea for everyday pragmatic usage, and the sensitivity assessment are new.

*Corresponding author. Email: riffenbu@sdsu.edu.

Received February 5, 2024; Accepted June 8, 2024

1.2 The Need for a JED Test

In the current workaday methodology in which we want to compare the efficacy of two treatments, an early step an investigator must make is to choose between testing difference versus equivalence. The difference t test has been available since Gosset's historic paper in 1908 (Student, 1908) and equivalence testing since the 1980s.

Traditionally, if the user believes that the outcomes will be different, a difference test leads to one of two possible conclusions: either (a) reject the hypothesis that the two treatments are the same, concluding a difference, or (b) fail to reject it. The latter outcome effectively says that there are no grounds on which to make a decision. The user gets half the option.

If, on the other hand, the user believes that the outcomes will not be different, an equivalence test leads to either (c) reject the hypothesis that the two treatments are different, concluding similarity, or (d) fail to reject (no conclusion). Again the user gets half the option.

A joint equivalence and difference test provides the full option in a single test, allowing the user to know if the effects are different or are similar.

A note on the term "equivalence": A better term would be "similarity". The conclusion is not that the two treatments yield identical results but that the contrast between treatments is small enough for them to be considered similar rather than different. We will use the historical "equivalence" in this article because it represents the profession's established nomenclature.

An example of the need for a JED test might be the comparison of two new fertilizer formulations. The investigator has no idea if they provide equivalent benefits or if one is better than the other. To pose null and alternative hypotheses in which one hypothesis leads to indecision does not answer the need.

There are multiple possible applications that could benefit from a JED test, depending on the requirements of the goal and the nature of the data. Which of the several forms should a practical JED test take? The simplest and most commonly used need to be addressed is a comparison of two means in a randomized controlled trial (RCT). If this form should be accepted, other cases should be put into pragmatic formats for common use: tests on other parameters, observational data tests, multivariate tests, rank tests, etc.

1.3 Past Work Contributing to the Development of a JED Test

The concept of JED testing was first seen in Wald (1945) although it was not so named. It was embedded in the method of sequential analysis aimed at minimizing sample size. The sample increased unit by unit until a significance level was crossed. This simultaneous use of both types of error is based on the approach of decision theory that many say should supplant null hypothesis statistical testing (NHST) [see, for only the NHST issue, Betensky (2019) and Matthews (2019)].

A number of papers have given basics underlying a JED test or a form of the test itself as part of a different goal. However, none of them presented a JED test in a form usable in workaday statistical analysis.

Bofinger (1985, 1992) and Hsu et al. (1994) examined properties of confidence intervals (CIs) that later proved useful in JED testing. Da Silva et al. (2009), Christensen (2007), and Mascha (2010) looked at both equivalence testing and difference testing on the same data set, but did not pose using them simultaneously as a joint test. Christensen (2007) and Mascha (2010) provided graphical displays containing confidence intervals for forms of equivalence, difference, and indeterminacy.

Some papers gave theoretical treatments in which a JED test appeared as a special case but was not specified per se. This may be one cause as to why it has not been adopted into common

usage. In Bauer and Kieser (1996), the concept of JED testing appeared as an incidental case but was not the primary focus. They listed a family of equivalence and difference tests with decisions based on confidence intervals, bringing together results from prior papers. General hypotheses included a joint test as a subset case. The focus was on how confidence intervals relate, not on performing tests. The mechanism of a JED test was not given, nor were examples.

Procedures for a JED test have appeared in different formats. Rosenbaum and Silber (2009) used a rather general composite (complex) hypothesis theory. JED testing appeared as a subset case in a configuration limited to only observational (uncontrolled) studies that are subject to bias from covariates. Furthermore, while quite thorough and addressing sensitivity, the form of presentation was too abstract for common application.

Berger (1982) examined situations with multiple predictive variables, testing multiple parameters simultaneously. However, it did not focus on JED testing and provided no easily usable test form.

Tamhane and Logan (2004) looked at cases of multiple outcomes. They posed confidence intervals in two dimensions, one for equivalence and one for difference. They named their test *ui-iu* for the union-intersection test of Roy (1953) and the intersection-union test of Berger (1982). Starting with Hotelling's T^2 test, their test used the bivariate t distribution. They compared it in a simulation study to those of Bloch et al. (2001) and Perlman and Wu (2004). Bloch et al. (2001) did not control p-value, lacked monotonicity, and permited contradictory outcomes so was not recommended. The Perlman and Wu (2004) test was based on a one-sided likelihood ratio statistic from Perlman (1969), in turn based on the position of a multivariate difference vector respective to an orthant (a quadrant generalized to hyperspace). The simulation study showed that the discriminating power of *ui-iu* and Perlman and Wu (2004) are similar, but with Perlman and Wu (2004) slightly higher. This work is outstanding and is an up-to-2004 benchmark for JED test development and assessment. However, it is currently out of reach for practical use by non-statisticians (and many statisticians).

Observational methods treat measures over which there is no control and that are usually confounded by covariates. Cornfield et al. (1959) and Rosenbaum and Rubin (1983) introduced JED testing for such cases. Gastwirth (1992) and Rosenbaum and Rubin (1983) further examined the sensitivity of a covariate influencing a test outcome from observational data.

Waldhoer and Heinzl (2011) proposed a JED test creatively specialized to spatial measures, but these measures were not unique quantifications. As one issue, they tested if their areas differ from a reference value, but the use of areas did not allow for a conclusion of equivalence in the case of no significant difference. This approach even allowed an outcome that they intepreted as equivalence and superiority (which is non-equivalence) simultaneously. They relied completely on confidence intervals and did not address error risks or their size.

1.4 Cases in Which a JED Test Is Defined but Is Not Easily Usable

Cases that specifically define a JED test focus on purposes other than testing or employ methods that are incomplete or that contain errors. These cases constitute another cause as to why JED testing has not been adopted into common usage.

Öhrn and Jennison (2010), like Wald (1945), focus on stopping rules (i.e. minimizing sample size), specifically on adaptive designs. In addition, they do not allow for equivalence, treating only superiority v. non-superiority. Allen and Seaman (2006), Morikawa and Yoshida (1995), and Tryon (2001) all present a form of a JED test, but use only confidence intervals that avoid specific values useful in interpretation of results and that yield no indication of the strength of

the difference or equivalence through measures such as p-values and power. None are written in a form to guide a non-statistician statistics-user to conduct a JED test. Morikawa and Yoshida (1995) focused on confirmatory phase II trials. Tryon (2001) used two confidence intervals on the respective means rather than one confidence interval on the difference between means. In addition, the treatment included an error that was later corrected in Tryon and Lewis (2008).

Hirotsu (2007) gives a discussion that appears to be more readable than most by non-mathematical users. He provides a good logical breakdown of hypothesis concepts, but his presentation is involved mostly with the logic. His argument depends partially on Japanese standards. In addition, he breaks the application down into specific levels of weak versus strong non-inferiority, whereas the user should be able to choose these or other levels by selecting the appropriate α. His work contains no straightforward guide to application.

Some early but insufficient thinking on the approach appeared in Riffenburgh and Gillen (2020) (section 12.6, pp 307-309).

Goeman et al. (2010) provided a solid and comprehensive JED testing approach, first using a test statistic as in the common t test and then using confidence intervals. They stated that the test-statistic method is not compatible for simultaneous use with the confidence interval method but is more powerful and therefore should be the method of choice. However, the methodology advanced by Goeman et al. (2010) is again not easily followed by users not trained in mathematics. In addition, their method poses three test hypotheses (the mean difference is negative, is zero, or is positive) coupled with two error risks that lead to five possible outcomes (the mean difference is negative, is non-positive, is null, is non-negative, or is positive). Five outcomes occur because Goeman et al. (2010) use a total α risk of error for equivalence, leading to tail areas of $t_{\alpha/2}$ and $t_{1-\alpha/2}$ while the risk of error for difference tails are t_{α} and $t_{1-\alpha}$. These outcomes correspond respectively to the interpretations that one mean is inferior, non-superior, equivalent, non-inferior, or superior to the other. The non-superior and non-inferior options allow the acceptance of two hypotheses at once: inferior-plus-equivalent and equivalent-plus-superior, respectively.

If we confine our risk of error to α and use only tail regions t_{α} and $t_{1-\alpha}$, the test statistic will lie in one of three mutually exclusive and exhaustive sample space regions corresponding to the hypotheses: superior, equivalent, or inferior.

Because Goeman's JED test is comprehensive and appears to be the most powerful form, the approach we propose to follow is more closely aligned with theirs than with that of other authors, but we present our method in a much simpler and easily usable form.

1.5 Basis of a Practical JED Test in Randomized Controlled Trials

In our approach, we sought a form of the test that is simple to use, that will answer needs frequently met in scientific application, and that has adequate power and effect size. We address the question of whether sample evidence lies more strongly with a positive difference, a negative difference, or no difference between two means arising from normally distributed data. These three possible outcomes give rise to three hypotheses to be tested. The data arise from a RCT, so that bias from uncontrolled covariates need not be involved.

We propose that such a JED test be adopted by the community of statistical method users as a default form of comparing two means. We encourage statistical researchers to develop easily usable forms of other versions in the near future, versions based on other distribution characteristics (e.g. rank tests) or other parameters (e.g. variance tests).

In Section 2, we provide a step-by-step procedure to be followed by the user, including those

who are not statistically sophisticated. A numerical example is given in Section 3. Finally, some assessments of the quality of our JED test, including sensitivity, appear in Section 4.

2 A Practical Form of the JED Test

2.1 Terminology and Concepts

We have continuous observations (measurements) on some variable of interest. We take a sample of observations from each of two competing situations, each having a mean value. Our goal is to compare the means of the samples to decide if they are the same or different. The data are assumed to occur randomly and to be normally and independently distributed.

Definitions of concepts and symbols we use:

n_1, n_2: numbers of observations in the two samples.

μ_1, μ_2: population means for the two samples. The subscript "2" is assigned to the population having the larger sample mean.

m_1, m_2: sample means, estimating μ_1 and μ_2.

σ_1, σ_2: population standard deviations for the two samples.

s_1, s_2: sample standard deviations, estimating σ_1 and σ_2.

δ: the true but unknown difference μ_2–μ_1.

d: the observed difference m_2–m_1, estimating the true but unknown δ.

σ_d: population standard error of the difference.

s_d: sample standard error of the difference. If s_d is not given by statistical software and the user must calculate it, see Appendix A.

Δ: the minimum practically meaningful difference between the means of interest. (It may be thought of as the bound of the equivalence margin.)

α: the level of probability required to reject a hypothesis; often called significance level.

ν: degrees of freedom for evaluating probabilities. If ν is not given by statistical software and the user must calculate it, see Appendix A.

t_ν: critical value of the t distribution for ν degrees of freedom; for example, $t_{58} = -1.67$.

δ/σ_d: the difference (distance) between the two population means given as the number of standard errors.

d/s_d: the sample estimate of δ/σ_d. This value is distributed as t. It is used to allow the values to be measured on an axis in units of t. The degrees of freedom ν, contained in s_d, depends on assumptions and sample size; see Appendix A for forms.

Δ/s_d: the meaningful difference given in t units.

$\mathrm{F}_\nu(\cdot)$: the distribution function of the t distribution, where ν is degrees of freedom.

$\Phi(\cdot)$: the distribution function of the standard normal distribution.

There are three hypotheses:

H_+: $\delta = \Delta$. Acceptance of H_+ would imply a conclusion that $\delta \geqslant \Delta$.

H_0: $\delta = 0$. Acceptance of H_0 would imply a conclusion that $-\Delta < \delta < \Delta$.

H_-: $\delta = -\Delta$. Acceptance of H_- would imply a conclusion that $\delta \leqslant -\Delta$.

p: the probability of sufficient evidence to fail to reject the associated hypothesis. p has subscripts +, 0, or −, corresponding to the hypotheses respectively.

The three JED p-values are:

$p_+ = F_\nu(\text{test statistic for } \mathrm{H}_+)$.

$p_0 = 2F_\nu(-\,|\text{test statistic for } \mathrm{H}_0|)$.

$p_- = F_\nu(-\,\text{test statistic for } \mathrm{H}_-)$.

The outcome of the test posed here is the combination of acceptance or rejection of the three t-test hypotheses. If $p > \alpha$ for one hypothesis (not rejected) but $p < \alpha$ for the other two (rejected), the test outcome is acceptance of the not-rejected hypothesis. If two or three hypotheses are not rejected, the outcome is equivocal and no conclusion can be made. We believe that the majority of practical t-test applications will lead to the acceptance of one outcome. To address this issue, a table of outcomes for various values of d and s_d relative to Δ is provided in Section 4.

2.2 Steps in the Method

The mechanics of conducting the t-test form of the JED test are as follows.

- Verify the assumptions required for the test. Choose α.
- Evaluate the descriptive statistics (m_1, m_2, d, s_1, s_2, s_d, and ν).
- Calculate the test statistics as $(d - \Delta)/s_d$ when H_+ is posed; as $(d - 0)/s_d$ when H_0 is posed; and $(d + \Delta)/s_d$ when H_- is posed. (The test statistics are similar to those of a traditional t test.)
- Calculate the p-values corresponding to the posed hypotheses H_+, H_0, and H_-.
- If two hypotheses are rejected and one is not, accept that hypothesis as the outcome of the JED test. If less than two hypotheses are rejected, the outcome is equivocal; the test result is not conclusive.

2.3 The JED Test Compared to a Traditional t Test

The JED test embodies some new concepts different from the traditional t test. We will compare the JED test to a two-sample mean difference t test; an equivalence t test would follow the same reasoning.

For the traditional t test, a null hypothesis is posed, saying there is no difference between means. A significance level α is posed as the dividing point between the probability of evidence of a difference when there isn't one being small enough to accept, and therefore concluding that the difference is real, versus that probability being too large to accept, and therefore having no conclusion. This probability is named p. The difference in standard error units, distributed as t, is calculated. The estimated probability that a difference this large would occur by chance, viz. p, is calculated. If $p < \alpha$, we conclude a difference. If $p \geqslant \alpha$, we do not have enough evidence to make a conclusion.

For the JED test, we pose three hypotheses, $\mu_2 \geqslant \mu_1$ (superiority), $\mu_2 = \mu_1$ (equivalence), or $\mu_2 \leqslant \mu_1$ (inferiority). The label "null hypothesis" is not used. A significance level α is posed as the probability below which we lack adequate evidence to reject whichever hypothesis is being assessed. The difference in standard error units is calculated. We define p as the estimated probability that the data provide sufficient evidence to fail to reject the associated hypothesis. Rather than one p-value, we calculate three p-values, one for each hypothesis. The outcome of the test rests not on the assessment of one p but on the simultaneous assessment of all three p-values. If one p-value $> \alpha$ and the other two p-values $\leqslant \alpha$, we have evidence for a conclusion. The conclusion will be associated with the hypothesis for which its $p > \alpha$. If two or all three of the p-values $> \alpha$, the result of the test is inconclusive.

The user accustomed to the traditional t test may tend to think of each p-value assessment as a test and view the JED test as composed of three tests. However, a test is defined as a process leading to a decision. No decision is made after each p assessment. The three assessments of

p-values compose a single decision and is therefore a single test. We could define a single null-equivalent hypothesis composed of three sub-hypotheses, but that would be just a semantic difference, convoluted and confusing; the process would remain unchanged.

3 Comparing Treatments for a Broken Ankle

3.1 Setup

The following numerical example uses data from a randomized controlled trial designed and analyzed by the author (Riffenburgh, 2006) while at the Naval Medical Center San Diego. The data were obtained for process improvement, not for a research study, and are unpublished. They are available through a link given below under *Supplementary Material*. To treat a fractured ankle, the medical standard of care mandated pinning by device #1. (Pinning is fixation of shattered bones by screw- or nail-like pins during healing.) An investigator recorded outcomes using device #1, the standard in common use, compared with outcomes using device #2, a new, cheaper, and more easily installed pinning device, with 30 patients treated by each device. The measure of success used for comparison is the distance—measured in inches—covered in a triple hop on the injured leg 4 months after repair (the longer the hop, the better the healing). This study was chosen because it is straightforward and a distance in inches is easy to visualize.

The investigator judged that a difference in hop length must exceed 6 in., that is, half the length of a foot, to have clinical meaning; $\Delta = 6$. The investigator, having no idea whether the new device is better, worse, or no different than the current device, carried out a joint equivalence and difference (JED) test. The hop distance distributions appeared to be approximately normal so use of the t distribution was assumed to be appropriate. Sample means and standard deviations are $m_1 = 33.27$, $s_1 = 4.43$; and $m_2 = 35.13$, $s_2 = 6.12$. The difference between means is $d = 1.86$, its $v = 58$, and its standard error $s_d = 1.38$. α is chosen as 0.05.

3.2 Example Results

The regions associated with the three t-test hypotheses are $H_+ : \delta = 6$; $H_0 : 0$; or $H_- : \delta = -6$. The test statistics and their associated p-values are.

For H_+, $(d - \Delta)/s_d = (1.86 - 6)/1.38 = -3.00$. $p_+ = F_{58}(-3.00) = 0.002$. As $0.002 < 0.05$, we have evidence to reject H_+.

For H_0, $(d - 0)/s_d = 1.86/1.38 = 1.35$. $p_0 = 2F_{58}(-1.35) = 0.182$. As $0.182 > 0.05$, we do not have evidence to reject H_0.

For H_-, $(d + \Delta)/s_d = (1.86 + 6)/1.38 = 5.70$. $p_- = F_{58}(-5.70) < 0.001$. As $0.001 < 0.05$, we have evidence to reject H_-.

The test statistic is not statistically different from 0 but is statistically different from being $\geqslant \Delta$ or $\leqslant -\Delta$, we conclude that the post-operative hop distances for the two devices are not different.

As a further example, suppose d had been 9, larger by half again the hop distance considered to be clinically meaningful.

For H_+, $(d - \Delta)/s_d = (9 - 6)/1.38 = 3.00$. $p_+ = F_{58}(2.17) = 0.983$. As $0.983 > 0.05$, we have no evidence to reject H_+.

For H_0, $(d - 0)/s_d = 9/1.38 = 6.52$. $p_0 = 2F_{58}(-6.52) < 0.001$. As $0.001 < 0.05$, we have evidence to reject H_0.

For H_-, $(d + \Delta)/s_d = (9 + 6)/1.38 = 10.87$. $p_- = F_{58}(-10.87) < 0.001$. As $0.001 < 0.05$, we have evidence to reject H_-.

The hop distance for the new device would have been neither similar to nor smaller than that for the old device. We conclude that the new device would have given better post-operative recovery.

A note on the interpretation of our numerical example may be in order because Sections 2 and 3 are directed toward the user who is not always an experienced statistician. What we want to know is if the two devices provide the same or different health benefits from the surgery. The test does not tell us that. It speaks only about the similarity of hop distances, which gives us clues about relative health benefits only insofar as hop distances—and only at one point in time—represent such benefits. For example, hop distances might be different shortly after surgery but become similar over time. Hop distances might be the same but ankle motion might be different. Furthermore, it speaks only of average performance. The variabilities might be different. And it speaks only to one rather small sample, not the population of treated patients. As with statistical testing in general, the limitations of the conclusion from the JED test must be fully understood and the test used with appropriate caution.

4 Quality Assessments of the JED Test

4.1 Outcomes for Various Values of d and Variability Relative to Δ

Table 1 shows the behavior success of the JED test in the vicinity where it becomes capable of discerning an outcome.

Row quads (four-line coverage of the same Δ/s_d ratio) going down the table vertically represent Δ growing larger relative to the standard error or, alternatively, the variability grow-

Table 1: Outcomes of the JED test for various values of Δ and the standard error.

Δ	*p-values; outcome*	d							
		0.1 Δ	0.2 Δ	0.4 Δ	Δ	1.5 Δ	2 Δ	3 Δ	4 Δ
	p_+	0.184	0.212	0.274	0.500	0.691	0.841	0.997	0.999
1 std	p_0	0.920	0.841	0.689	0.317	0.134	0.046	0.003	<0.001
error	p_-	0.136	0.115	0.081	0.023	0.006	0.001	<0.001	<0.001
	outcome	none	none	none	none	none	superior	superior	superior
	p_+	0.036	0.055	0.115	0.500	0.841	0.977	>0.999	>0.999
2 std	p_0	0.841	0.689	0.424	0.046	0.003	<0.001	<0.001	<0.001
errors	p_-	0.014	0.008	0.003	<0.001	<0.001	<0.001	<0.001	<0.001
	outcome	equiv	equiv	none	superior	superior	superior	superior	superior
	p_+	0.003	0.008	0.036	0.500	0.933	0.999	>0.999	>0.999
3 std	p_0	0.764	0.549	0.230	0.003	<0.001	<0.001	<0.001	<0.001
errors	p_-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	outcome	equiv	equiv	equiv	superior	superior	superior	superior	superior
	p_+	<0.001	<0.001	0.008	0.5	0.997	>0.999	>0.999	>0.999
4 std	p_0	0.689	0.424	0.11	<0.001	<0.001	<0.001	<0.001	<0.001
errors	p_-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	outcome	equiv	equiv	equiv	superior	superior	superior	superior	superior

ing smaller relative to Δ. Columns going across the table horizontally represent values of the difference between sample means, d, relative to Δ.

The test is not very helpful when the variability is as large as Δ. However, that is a situation unlikely to be tested in real applications because the mean differences that suggest a need for a test would be obscured by variability. As the Δ/s_d ratio grows larger, the JED test begins to perceive differences and equivalences. When Δ reaches a little over two standard errors, the JED test can discern differences and equivalences for all cases other than when d is very close to Δ, that is, when the sample difference hovers about the value separating equivalence from difference. Users don't usually gather data and perform tests unless there is some indication of a difference or an equivalence from experience or prior data. In these cases, the JED test provides an answer. We note that, in our numerical example, the sample difference between means, $d = 1.86$, is less than 1/3 of the Δ of 6 inches.

4.2 Power for a JED Test

Another measure of quality is power. The p-values give an indication of decision quality based on the data; the true δ is unknown. Power gives an indication of decision quality when δ is known. For equivalence, we hypothesize that there is no difference between means, so the true $\delta = 0$. For difference, we hypothesize that $\delta > \Delta$ (or $\delta < -\Delta$), but the true δ is unknown, so we must use $\delta = \Delta$ (or $\delta = -\Delta$) as a lower (or upper) bound; we recognize that the power for a difference is at least this large.

For the JED test, we define power as the probability of selecting a hypothesis given it is true. Let us denote the JED power by W. (The letter "p" is so overused as to be confusing; W is the second consonant in the word "power".) A power can be calculated for each outcome. In Section 2.2, the hypotheses were designated H_+ H_0 and H_-. We can denote the corresponding powers as W_+, W_0 and W_-.

Analogous to NHST testing, *specificity* can be related to detecting an equivalence and *sensitivity* can be related to detecting a difference.

For evaluating the influence of various values of the parameters involved, we will assume that standard deviations σ_1 and σ_2 are known, leading to using the normal (z) distribution rather than the t distribution. We use the usual designation $\Phi(z_\alpha)$ to represent $P[x < z_\alpha]$, including the area under the standard normal distribution function up to z_α.

To find W_+, we hypothesize $\delta = 0$ and calculate the probability of rejecting the hypothesis given δ is truly Δ, i.e. we reject if $[(d - 0)/\sigma_d] > z_{1-\alpha}$. Subtracting and adding Δ to d, carrying the added Δ across the inequality, and taking the integral, we obtain W_+.

The power for accepting superiority when it is true (μ_2 is truly greater than μ_1) becomes

$$W_+ = 1 - \Phi[z_{1-\alpha} - (\Delta/\sigma_d)].$$

By symmetry, the power for accepting inferiority when it is true (μ_2 is truly less than μ_1) is

$$W_- = \Phi[z_\alpha + (\Delta/\sigma_d)].$$

The power for accepting equivalence when it is true (μ_2 is truly not different from μ_1), implying a difference (either superiority or inferiority; it could not be both) is false, is

$$W_0 = \Phi[(\Delta/\sigma_d) - z_{1-\alpha}].$$

These three expressions yield the same value, so we have a single power W for the JED test, whichever outcome we accept, as

$$W = \Phi[(\Delta/\sigma_d) - z_{1-\alpha}].$$

The numerical example's outcome was equivalence (or, said more exactly, similarity). Carrying out W with t-test calculations using the sample parameters yields

$$W = F_{58}[2.703] = 0.995.$$

The power of this sample to detect equivalence between μ_1 and μ_2 exceeds 99%.

Note that power is a non-negative value, that W_0 and W_+ are mutually exclusive and mutually exhaustive, and that joint (W_0 and W_+) $\geqslant W_+$ and joint (W_0 and W_+) $\geqslant W_0$. Therefore, joint (W_0 and W_+) is not less powerful than either W_0 or W_+ alone. The case for W_- follows by symmetry.

4.3 Some Assessment of Power for a JED Test

Several papers provided forays into various aspects of power. Rosenbaum and Silber (2009) looked at the effect of variability in mean differences. Tamhane and Logan (2004) examined variability in values of data parameters in a simulation study and compared their results with outcomes from the Bloch et al. (2001) and Perlman and Wu (2004) tests. Christensen (2007) looked at variability in the standardized mean differences for various value of α. Da Silva et al. (2009) examined the effect of various sample sizes on sensitivity. Öhrn and Jennison (2010) looked at the effect of using adaptive designs on sensitivity. Rosenbaum and Rubin (1983) showed the effect of binary and categorical covariates in binary observational studies. Gastwirth (1992) looked at sensitivity to missing data in observational studies.

In this section, we look at the power for our JED test. We assume the standard deviations σ are known and therefore use the normal distribution in calculating power.

We tabulate the effect of various design parameters on power (Table 2) and of sampling outcome values on power (Table 3). For both tables, three values of Δ, the size of a difference believed to be practicably meaningful, are used: $\Delta = 0.5$, the size of 0.5 σ_1 above 0; $\Delta = 1$, the size of σ_1 above 0; and $\Delta = 1.5$, the size of 1.5 σ_1 above 0.

In Table 2, sample size is assumed to be controllable by the experimenter. Sample variabilities ($\sigma_1 = \sigma_2 = 1$) are held constant so that the effects of sample size and subsample ratio can be seen. The power is examined for variability in the test design parameters: error risk α (= 0.10, 0.05, and 0.01), sample size $n(= n_1 + n_2)$ (= 10, 20, 30, and 60), and sample size disparity n_2/n_1 (= 1, 1.5, and 2).

In Table 3, sample size is assumed to be controlled by sample availability. The error risk α and sample size n are varied as before to facilitate comparison. The group sizes remain equal ($n_1 = n_2$). σ_2 is allowed to vary (= σ_1, 1.5 σ_1, and 2 σ_1) and with it the variability ratio σ_2/σ_1 (= 1, 1.5, and 2).

Let us examine the effect on the power of different parameters and how they interact with sample size for each of the two sample groups.

The effect of varying the *equivalence margin value* Δ: The closer the sample difference d is to the equivalence margin value, i.e. the minimum difference believed to be practicably meaningful, the greater is the sample size required to discern that difference. When the means are half a standard deviation apart, power is inadequate for all sample sizes posed. However, an

Table 2: JED-test power for various α, n, and n_2/n_1 for three values of Δ.

		Power of the JED test								
		Δ=0.5			Δ=1.0			Δ=1.5		
n	n_2/n_1	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
	1.0	0.312	0.196	0.062	0.618	0.475	0.228	0.862	0.766	0.518
10	1.5	0.306	0.192	0.060	0.606	0.462	0.219	0.851	0.751	0.499
	2.0	0.289	0.179	0.055	0.567	0.422	0.190	0.814	0.702	0.439
	1.0	0.435	0.299	0.113	0.830	0.723	0.464	0.981	0.956	0.848
20	1.5	0.426	0.291	0.109	0.818	0.707	0.446	0.978	0.950	0.831
	2.0	0.415	0.282	0.104	0.803	0.687	0.423	0.972	0.940	0.809
	1.0	0.535	0.391	0.169	0.927	0.863	0.660	0.998	0.993	0.963
30	1.5	0.524	0.381	0.162	0.920	0.850	0.639	0.997	0.991	0.955
	2.0	0.504	0.362	0.150	0.903	0.826	0.601	0.995	0.987	0.939
	1.0	0.744	0.615	0.348	0.995	0.987	0.939	>0.999	>0.999	>0.999
60	1.5	0.731	0.600	0.334	0.994	0.984	0.929	>0.999	>0.999	>0.999
	2.0	0.707	0.572	0.308	0.991	0.978	0.907	>0.999	>0.999	0.999

Table 3: JED-test specificity for true equivalence and sensitivity for three true superiority for various α, n, and σ_2/σ_1.

		Power of the JED test								
		$\Delta = 0.5$			$\Delta = 1.0$			$\Delta = 1.5$		
n	σ_2/σ_1	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
	1.0	0.312	0.196	0.062	0.618	0.475	0.228	0.862	0.766	0.518
10	1.5	0.254	0.153	0.044	0.484	0.343	0.139	0.719	0.585	0.321
	2.0	0.217	0.126	0.034	0.389	0.260	0.092	0.586	0.442	0.204
	1.0	0.435	0.299	0.113	0.830	0.723	0.464	0.981	0.956	0.848
20	1.5	0.343	0.221	0.074	0.682	0.544	0.284	0.911	0.838	0.620
	2.0	0.283	0.174	0.053	0.553	0.409	0.181	0.799	0.683	0.419
	1.0	0.535	0.391	0.169	0.927	0.863	0.660	0.998	0.993	0.963
30	1.5	0.418	0.284	0.105	0.807	0.693	0.429	0.974	0.943	0.815
	2.0	0.339	0.218	0.072	0.674	0.535	0.276	0.906	0.830	0.607
	1.0	0.744	0.615	0.348	0.995	0.987	0.939	>0.999	>0.999	>0.999
60	1.5	0.594	0.450	0.210	0.961	0.918	0.762	0.999	0.998	0.987
	2.0	0.477	0.337	0.135	0.879	0.789	0.549	0.992	0.979	0.911

experiment with variability so large is unlikely to be conducted. When the means are about a standard deviation apart, a group sample size of a little over 10 is required. When the distance apart increases to about 1.5 standard deviations, the required group sample size reduces to a little over 5.

In our numerical example, d is 2.7 standard errors from Δ and the power exceeds 0.99. Of course, in designing the experiment, the value of d will not be known until after data acquisition. We can say that sample size should be increased if the anticipated d will lie close to Δ.

The effect of varying the *risk of accepting an error* α: Power increases as α increases and vice versa. For a typical design in which α is chosen as 0.05 and d lies 1.5 standard deviations from Δ, adequate power is reached with a group sample size of just above 5. Again, the value of d will not be known until after data acquisition. We can say that sample size should be increased if a smaller than usual α is chosen.

The effect of varying the *relative size of sample groups* n_2/n_1: As the disparity between group sample sizes grows, the power decreases, but not greatly. When one group sample size is double the other, the power decreases about 10% for group sample size 5, lessening the decrease to 1% for group sample size 15. Increasing the sample size by 10% compensates for the loss in power due to unequal sample size. For example, if group sample sizes are 15, $n_2/n_1 = 1$, the power is 0.993. Changing n_2/n_1 to 2 ($n_1 = 10$; $n_2 = 20$) diminishes power to 0.983. Maintaining $n_2/n_1 = 2$ and changing the group sample sizes to 11 and 22 restores the power to 0.992.

The effect of varying the *relative size of sample group variability* σ_2/σ_1: For very small samples, power drops to about half if the variability ratio goes to 2, dropping a bit less for larger α and a bit more for smaller α. The loss of power lessens as sample size increases. By group sample sizes of 30, the drop is 20% for $\Delta = 1$ and 2% for $\Delta = 1.5$.

4.4 Some Assessment of Effect Size

Another measure of quality of the JED test is its effect size. The effect size for the JED t-test application is a measure of how meaningful is the difference between the means. The measure is usually taken as the difference between means divided by the pooled standard deviation, s_p, for the two independent samples. However, we are interested in the deviation of d from Δ, the point separating an inconsequential difference between means from a difference large enough to have practical implications. To show the effect of this deviation, we offset the d by Δ, so that our effect size is

$$D = (d - \Delta)/\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

The effect size, then, may be thought of as the number of pooled standard deviations apart is the difference between the two means and Δ. The most frequently used interpretive aid for t tests is due to Cohen (1998), who classes an effect size of 0.2 as a small effect, 0.5 as a medium effect, and 0.8 as a large effect. Effect sizes of 1.0 or greater are very large. Serdar et al. (2021) address further interpretation. When $d < \Delta$, $|D|$ indicates the strength of data to show no evidence that $\mu_2 > \mu_1$. When $d > \Delta$, D indicates the strength of data to show evidence that $\mu_2 > \mu_1$.

To simplify the effect size examination, we let the sample standard deviations and the sample size ratio equal one, leaving the pooled standard deviation of the means equal to one for all sample sizes. Our effect size for interpretation becomes

$$D = (d - \Delta) = (m_2 - m_1) - \Delta$$

and we interpret the effect size about Δ rather than about 0. We confine this effect size to $d \geqslant 0$. (If $d < 0$ and the $\Delta > 0$ is large enough for a practical test, $m_2 < m_1$ and it is most unlikely that $\mu_2 > \mu_1$; this region need not be pursued for useful applications.) The case for $\Delta < 0$ follows by symmetry.

When $d = 0$, $D = -\Delta$, the point at which μ_1 exceeds μ_2 by Δ, the amount judged as meaningful. When $d \subset (0, \Delta)$, $\mu_1 > \mu_2$. The JED test will conclude that m_2 is not larger than m_1 with effect size the number of standard deviations μ_1 is larger than μ_2. If $|D| \geqslant 0.8$, the effect size is judged to be large, etc. As $d \to \Delta$, the effect size $\to 0$. When $d = \Delta$, the effect size is 0 and the difference between means falls on the point dividing a meaningful difference from a non-meaningful difference. When d exceeds Δ, $m_2 > m_1$. When d becomes sufficiently large, the JED test concludes $\mu_2 > \mu_1$. The effect size of this conclusion is the number of pooled standard deviations by which m_2 exceeds m_1.

The numerical example's outcome was equivalence of the two means (or, said more exactly, their similarity). Carrying out D with t-test calculations using the sample parameters yields an effect size of $[(m_2 - m_1) - \Delta]/s_p = (1.86 - 6)/5.342 = -0.775$. The negative sign indicates that the outcome falls in the region of no difference. We take the effect size to be $|D| = 0.78$, just below the Cohen's interpretation as large. We find a large strength in our data to show no evidence that $\mu_2 > \mu_1$.

4.5 Summary of Interpretation

Experiments are seldom undertaken when data parameters are of a nature unlikely to give useful information. It is not unreasonable to assess a statistical method based on commonly used values of those parameters. We interpreted our assessment of the JED test in terms of parameter values from possibility regions likely to be used.

For sample size per sample group reaching at least 25, the power of the JED test is adequately high for any reasonably designed experiment. This agrees with the large sample assumption for historical t tests.

Looking at power for smaller samples, the parameter of greatest influence on JED test power is Δ, representing the true difference between means that has a practical implication. A useful experiment is likely to involve data in which the sample difference d is at least 1.5 standard deviations from Δ. As d moves closer to Δ, group sample sizes need to reach around 12 to 15 for adequate power. For the remainder of this section, we will assume $\Delta \geqslant 1.5$ standard deviations.

For the commonly assumed $\alpha = 0.05$, the JED test has adequate power for any group sample size larger than 6. If a smaller α is chosen, a larger group sample size should be used, at least 10.

If group sample sizes are very different, their sample sizes should be increased about 10%.

If variability in the two group samples is markedly different, vary small sample results are suspect. We recommend that group sample size be designed at 15 or greater.

Rules of thumb.

- If group sample sizes exceed 25, which satisfies the large sample requirement for ordinary t testing, the JED test has adequate power.
- If $|d - \Delta| \geqslant 1.5$ standard deviations, the JED test has adequate power and effect size. If $|d - \Delta| < 1.5$ standard deviations, design the experiment for group samples of at least 12 to 15.
- If $\alpha \geqslant 0.05$, the JED test has adequate power for group sample sizes of 6 or more. If $\alpha < 0.05$, group sample sizes should be increased up to 10 for $\alpha = 0.01$.

- If group sample sizes are very different, increase their sample sizes 10%.
- If variability in the two group samples is anticipated to be markedly different, design the experiment for at least 15 observations per group.

5 Conclusion

In this work, a joint equivalence and difference (JED) test is given to replace "one-sided" tests for either the equivalence or difference between two sample means. While not new, the test has not appeared in a form usable in workaday statistical applications. This article reviews the concept and underlying theory and presents the JED test for use in t-distribution applications in a straightforward, step-by-step guide, with possible interpretations and formulas for p-values (as slightly redefined for the JED test). The guide was illustrated by a numerical example from the medical field of orthopedics.

The quality of the JED test remained at question. It was noted that the joint test is at least as powerful as one-sided tests. Power calculation formulas were given. The sensitivity of the test was examined and shown in tables for common values of the parameters: the deviation of the mean difference (d) from the practically meaningful difference between means (Δ); the risk of accepting an error (α); representative group sample sizes (n_1 and n_2); the size ratio of the two samples (n_2/n_1); and the relative size of standard deviations of the two samples (s_2/s_1). Effect size was defined and discussed, giving methods of its calculation for applied use. Power and effect size were given and interpreted for the numerical example.

If group sample sizes exceed 24, as recommended for t tests in general, the JED test is adequately powerful and sized for unrestricted application. In most commonly seen designs, e.g. with frequently used values of α, roughly equal group sample sizes and variability, and a reasonable Δ, group sample sizes of 12 are adequate. For smaller samples than that, care should be taken in interpretation. We do not consider sample size limitations debilitating for JED test use so long as the user understands the risk of allowing lower test power for very small samples.

This work presents a reasonable justification for and guide for using a JED test when the assumptions underlying a t test are valid. To complete the JED portion of a statistical toolbox, the theory and practical guidance need to be extended to other tests, for example, ANOVA and nonparametric tests. It further needs to be extended to multivariate tests, for example Hotelling's T^2 test.

Supplementary Material

The dataset used in numerical example (Section 3) and R code for tables (Section 4) can be found at: https://github.com/wlingge/JED

A Appendix: Formulas for s_d and ν in the JED Test

The test statistic is

$$t = (m_2 - m_1)/s_d.$$

If it can be assumed that the population variances are equal and a difference between their estimates s_1^2 and s_2^2 is due to sampling variability, or if the sample sizes are large (say greater

than 50 or 100), then $\nu = n_1 + n_2 - 2$ and

$$s_d = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right]}.$$

If, however, the variances must be assumed unequal and the sample sizes are small, the Welch-Satterthwaite approximation may be used (Satterthwaite (1946); Welch (1947)):

$$s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and

$$approx(\nu) = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

The Welch-Satterthwaite approximation in the format used in this article is given in Riffenburgh and Gillen (2020) (section 11.3, p 248). Other approximations are available.

References

Allen IE, Seaman CA (2006). Different, equivalent or both? *Quality Progress*, 39(7): 77.

Bauer P, Kieser M (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*, 83(4): 934–937. https://doi.org/10.1093/biomet/83.4.934

Berger RL (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300. https://doi.org/10.2307/1267823

Betensky RA (2019). The p-value requires context, not a threshold. *American Statistician*, 73(sup1): 115–117. https://doi.org/10.1080/00031305.2018.1529624

Bloch DA, Lai TL, Tubert-Bitter P (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics*, 57(4): 1039–1047. https://doi.org/10.1111/j.0006-341X.2001.01039.x

Bofinger E (1985). Expanded confidence intervals. *Communications in Statistics - Theory and Methods*, 14(8): 1849–1864. https://doi.org/10.1080/03610928508829017

Bofinger E (1992). Expanded confidence intervals, one-sided tests, and equivalence testing. *Journal of Biopharmaceutical Statistics*, 2(2): 181–188. https://doi.org/10.1080/10543409208835038

Christensen E (2007). Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology*, 46(5): 947–954. https://doi.org/10.1016/j.jhep.2007.02.015

Cohen J (1998). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1): 173–203.

Da Silva GT, Logan BR, Klein JP (2009). Methods for equivalence and noninferiority testing. *Biology of Blood and Marrow Transplantation*, 15(1): 120–127. https://doi.org/10.1016/j.bbmt.2008.10.004

Gastwirth JL (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables. *Jurimetrics Journal*, 33: 19.

Goeman JJ, Solari A, Stijnen T (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine*, 29(20): 2117–2125. https://doi.org/10.1002/sim.4002

Hirotsu C (2007). A unifying approach to non-inferiority, equivalence and superiority tests via multiple decision processes. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 6(3): 193–203. https://doi.org/10.1002/pst.305

Hsu JC, Hwang JG, Liu HK, Ruberg SJ (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, 81(1): 103–114. https://doi.org/10.1093/biomet/81.1.103

Mascha EJ (2010). Equivalence and noninferiority testing in anesthesiology research. *The Journal of the American Society of Anesthesiologists*, 113(4): 779–781.

Matthews RA (2019). Moving towards the post p< 0.05 era via the analysis of credibility. *American Statistician*, 73(sup1): 202–212. https://doi.org/10.1080/00031305.2018.1543136

Morikawa T, Yoshida M (1995). A useful testing strategy in phase III trials: Combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics*, 5(3): 297–306. https://doi.org/10.1080/10543409508835115

Öhrn F, Jennison C (2010). Optimal group-sequential designs for simultaneous testing of superiority and non-inferiority. *Statistics in Medicine*, 29(7–8): 743–759. https://doi.org/10.1002/sim.3790

Perlman MD (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, 40(2): 549–567. https://doi.org/10.1214/aoms/1177697723

Perlman MD, Wu L (2004). A note on one-sided tests with multiple endpoints. *Biometrics*, 60(1): 276–280. https://doi.org/10.1111/j.0006-341X.2004.00159.x

Riffenburgh RH (2006). A Comparison of Two Fractured-ankle Pinning Devices. Unpublished process improvement data, Naval Medical Center San Diego. Personal data, collection of R. H. Riffenburgh.

Riffenburgh RH, Gillen DL (2020). *Statistics in Medicine*, 4th edition. Elsevier, Amsterdam.

Rosenbaum PR, Rubin DB (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B, Methodological*, 45(2): 212–218. https://doi.org/10.1111/j.2517-6161.1983.tb01242.x

Rosenbaum PR, Silber JH (2009). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *Journal of the American Statistical Association*, 104(486): 501–511. https://doi.org/10.1198/jasa.2009.0016

Roy SN (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2): 220–238. https://doi.org/10.1214/aoms/1177729029

Satterthwaite FE (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6): 110–114. https://doi.org/10.2307/3002019

Serdar CC, Cihan M, Yücel D, Serdar MA (2021). Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, 31(1): 27–53. https://doi.org/10.11613/BM.2021.010502

Student (1908). The probable error of a mean. *Biometrika*, 6(1): 1–25. https://doi.org/10.2307/2331554

Tamhane AC, Logan BR (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. *Biometrika*, 91(3): 715–727. https://doi.org/10.1093/biomet/91.3.715

Tryon WW (2001). Evaluating statistical difference, equivalence, and indeterminacy using in-

ferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6(4): 371. https://doi.org/10.1037/1082-989X.6.4.371

Tryon WW, Lewis C (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, 13(3): 272–277. https://doi.org/10.1037/a0013158

Wald A (1945). Sequential method of sampling for deciding between two courses of action. *Journal of the American Statistical Association*, 40(231): 277–306. https://doi.org/10.1080/01621459.1945.10500736

Waldhoer T, Heinzl H (2011). Combining difference and equivalence test results in spatial maps. *International Journal of Health Geographics*, 10: 1–10. https://doi.org/10.1186/1476-072X-10-1

Welch BL (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1–2): 28–35. https://doi.org/10.1093/biomet/34.1-2.28