

EVALUATING THE ASSUMPTION OF HOMOGENEITY OF VARIANCE
VIA EQUIVALENCE TESTING

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Philippe R. Gaillard

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Michael Rodriguez, Advisor

June 2009

© Philippe R. Gaillard 2009

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to the following individuals for helping me throughout my graduate studies:

To Pr. Michael Rodriguez, my advisor, thank you for your patience and for your support during my doctoral program, and for your insights on this thesis.

To Pr. John Connett, Pr. Shawn Curley, Pr. Robert delMas, and Pr. Kristen McMaster, thank you for serving on my thesis committee, and thank you for your careful reading and helpful feedback.

To Cynthia Davey, thank you for your comments and suggestions.

Finally, to my family, thank you for your constant support and encouragement.

ABSTRACT

Many research hypotheses involve a difference between levels of the independent variable(s), in terms of a dependent variable of interest. A research hypothesis of an existing difference is matched with a statistical alternative hypothesis of two parameters (e.g., population means) being different, and a null hypothesis of no difference. For the purposes of this thesis, the type of statistical testing used in such cases is referred to as “difference testing”. By contrast, statistical equivalence testing is a suitable approach in the less frequent situation where the research hypothesis involves the equivalence of several levels of the independent variable(s) in terms of a dependent variable. Once equivalence limits have been set, an equivalence test can be conducted, so as to test the null hypothesis of non-equivalence. The burden is on the researcher to accumulate enough empirical evidence to be able to reject this null hypothesis and conclude in favor of the alternative hypothesis of equivalence. After placing equivalence testing in context by reviewing its history, a formal definition and a numerical example of equivalence testing are given. Equivalence testing is then applied to the evaluation of the homogeneity of variance assumption, via a simulation study. An illustrative research example concerning classroom settings is also provided, in order to show how equivalence testing can be applied to research questions within Educational Psychology.

TABLE OF CONTENTS

ACKNOWLEDGMENTS i

ABSTRACT ii

LIST OF TABLES v

LIST OF FIGURES vi

CHAPTER 1 INTRODUCTION 1

 Statement of the Problem..... 1

 Historical context..... 3

 Research questions..... 4

 Limitations of the study 5

CHAPTER 2 REVIEW OF LITERATURE 7

CHAPTER 3 METHOD 16

 Formal definition of equivalence testing 16

 Numerical example of an equivalence test 22

 Methodological application: The assumption of homoskedasticity..... 26

 Computer simulations 30

 Pseudo code for the first set of simulations: PVT vs. SAT 34

 Pseudo code for the second set of simulations..... 36

 Substantive application: In class vs. online forms of the same course 39

CHAPTER 4 RESULTS AND ANALYSIS 41

 First set of simulations 41

 Second set of simulations..... 48

| | |
|---|----|
| Example of a substantive application: Comparison of students' scores between the in-class and online forms of the UMN course Public Health 6414: Biostatistical Methods I | 63 |
| CHAPTER 5 DISCUSSION..... | 71 |
| Summary of Findings..... | 71 |
| Contribution | 73 |
| Future research..... | 75 |
| REFERENCES | 77 |
| APPENDIX A Plots representing the equivalence test applied to the data set presented on page 23 | 83 |
| APPENDIX B PVT_SAT_alpha.sas program..... | 84 |
| APPENDIX C BF_ET_alpha_n50.sas program | 89 |
| APPENDIX D Pub6414 data set..... | 98 |
| APPENDIX E Plots representing the equivalence test of homoskedasticity (PubH5414 data)..... | 99 |

LIST OF TABLES

| | |
|---|----|
| Table 1. <i>Truth Table for Bioequivalence Decisions</i> | 10 |
| Table 2. <i>Combinations of sample ratios and variance ratios</i> | 32 |

LIST OF FIGURES

| | |
|--|----|
| <i>Figure 1.</i> Possible combinations of difference and equivalence results..... | 12 |
| <i>Figure 2.</i> Comparison of null hypotheses (nondirectional testing). | 15 |
| <i>Figure 3.</i> Mathcad worksheet for equivalence test p-value..... | 23 |
| <i>Figure 4.</i> Mathcad worksheet for equivalence test power..... | 24 |
| <i>Figure 5.</i> EquivTest output. | 25 |
| <i>Figure 6.</i> Simulation results for the control of alpha (PVT and SAT). | 42 |
| <i>Figure 7.</i> Actual alpha for the PVT. | 43 |
| <i>Figure 8.</i> Actual alpha for the SAT. | 44 |
| <i>Figure 9.</i> Simulation results for power (PVT and SAT). | 45 |
| <i>Figure 10.</i> Power for the PVT. | 46 |
| <i>Figure 11.</i> Power for the SAT. | 47 |
| <i>Figure 12.</i> Flow chart for the assessment of the BF and ET rules for switching between the PVT and the SAT as a function of the observed group variances..... | 51 |
| <i>Figure 13.</i> Agreement table between the BF and ET rules for H0 true and n = 50..... | 52 |
| <i>Figure 14.</i> Agreement table between the BF and ET rules for H0 true and n = 200..... | 53 |
| <i>Figure 15.</i> Simulation results for actual alpha (BF and ET) when n = 50..... | 54 |
| <i>Figure 16.</i> Simulation results for actual alpha (BF and ET) when n = 100..... | 55 |
| <i>Figure 17.</i> Simulation results for actual alpha (BF and ET) when n = 200..... | 56 |
| <i>Figure 18.</i> Agreement table between the BF and ET rules for H0 false and n = 50. | 57 |
| <i>Figure 19.</i> Agreement table between the BF and ET rules for H0 false and n = 200. | 58 |
| <i>Figure 20.</i> Simulation results for power (BF and ET) when n = 50..... | 59 |
| <i>Figure 21.</i> Simulation results for power (BF and ET) when n = 100..... | 60 |

| | |
|---|----|
| <i>Figure 22.</i> Simulation results for power (BF and ET) when total $n = 200$ | 61 |
| <i>Figure 23.</i> Probability density function for the central F distribution, and relevant characteristics of the data set. | 64 |
| <i>Figure 24.</i> Test of homoskedasticity. | 65 |
| <i>Figure 25.</i> Test of means. | 67 |
| <i>Figure 26.</i> SAS output for the equivalence test of group means (PubH5414 data)..... | 68 |
| <i>Figure 27.</i> SAS output for the conventional (i.e., difference) test of group means (PubH5414 data). | 69 |
| <i>Figure 28.</i> Decision chart for Equivalence Testing vs. Difference Testing. | 74 |

CHAPTER 1

INTRODUCTION

Statement of the Problem

Statistical hypothesis testing traditionally involves a null hypothesis of no difference. This is a result of the research hypothesis, a.k.a. substantive hypothesis, often being a hypothesis pertaining to an effect of some magnitude greater than zero. The statistical test is set up so as to match the substantive hypothesis with the alternative hypothesis of the test, H_A . The null hypothesis, H_0 , can then naturally represent the opposite possibility, that there is no effect. Being the default hypothesis, the null hypothesis is not rejected unless sufficient empirical evidence is accumulated against it. In this sense, the burden of the case rests on the researcher: Without enough evidence to support it, the claim of the existence of an effect is not accepted. As Oakes puts it, “The investigator seeks therefore to deny H_0 , or to ‘nullify’ it, and it is for this reason that it is known as the *null hypothesis*.” (1990, p. 5). It is important to note that the null hypothesis is essentially the hypothesis that the researcher attempts to reject, and only incidentally the hypothesis of no effect.

If the researcher’s claim pertains to an effect being just as large as another, then the alternative hypothesis should match this claim and entail the identity, or near identity, of the two effects, and by consequence the opposite hypothesis, i.e., the null hypothesis, should be one of a difference between the effects. In this way, the burden of the case rests on the researcher once again: The claim of similarity between the effects will need accumulated evidence to be supported, and without this evidence the null hypothesis of difference cannot be rejected.

Equivalence testing is the methodology for testing statistical hypotheses in exactly this situation: When similarity, rather than difference, is the research hypothesis, the appropriate definition of the alternative statistical hypothesis is one of equivalence, not of difference. In statistical analysis in general, and in Educational Research in particular, such instances occasionally arise, but have so far been handled mostly with difference testing, which is an ill-suited approach for this kind of research hypothesis. For instance, Terry (2007) has studied in-class and online modes of instruction, and asked whether the two were equivalent, but used traditional difference testing to answer this question. In pharmaceutical research however, and to some degree in medical research, equivalence testing has gained acceptance, and a sufficient body of literature on the subject currently exists to warrant an investigation as to how this approach could be applied to research in other disciplines, such as Educational Psychology.

Historical context

The need for a statistical method of hypothesis testing appropriate to demonstrate equivalence has long been recognized: “In many applications, the purpose of the experiment is to establish the equivalence of two or more treatments” (Dunnett & Gent, 1977, p. 593). Yet the methodical development of equivalence testing only began in response to evolving Food and Drug Administration (FDA) requirements. In the United States, the Food, Drug, and Cosmetic Act of 1938 gave the FDA authority to ensure that safety and efficacy data be generated to support claims for active ingredients in a drug prior to approval for sale. For a new formulation, clinical safety and efficacy trials are required, but for a re-formulation it is often sufficient to show therapeutic interchangeability (a.k.a. bioequivalence) with the reference formulation. According to Stegner (1996), “Over time a drug equivalence methodology has evolved as a result of an ongoing dialogue between the pharmaceutical industry and the Food and Drug Administration” (p. 194). In 1984, the Drug Price Competition and Patent Term Restoration Act further removed some barriers to generic drug development and allowed marketing of any off-patent drug, provided Abbreviated New Drug Application (ANDA) criteria are met. This more expedient process has fostered extensive development and marketing of generic drugs, and over half of the prescription drug units now sold in the U.S. are generic drugs that have been approved after demonstration of bioequivalence. Due to the large number of currently available medications, new medications (including but not limited to generic formulations) are increasingly being compared to active controls (i.e., existing formulations) as it would be unethical to continue using passive

controls (i.e., placebos) (Cleophas et al., 2006, p. 63). The importance of bioequivalence has resulted in well-defined statistical criteria for equivalence testing. At this time, the methodology of equivalence testing is well established in pharmaceutical research. While it is gaining acceptance more generally in medical research and biostatistics, the vast majority of textbooks in these disciplines do not yet include comprehensive coverage of the topic, with a few exceptions (such as Blair & Taylor, 2007).

Research questions

The purpose of this study is to assess whether equivalence testing could be useful in Educational Psychology. The application of this type of statistical testing can pertain to research methodology or to substantive research areas.

The main application examined in this study is a methodological one: When evaluating the assumption of homoskedasticity and whether it is tenable with a given set of data, the researcher can use a difference test, such as Levene's test, or an equivalence test. The question to be addressed in this study is whether the latter test would provide a sounder evaluation of this assumption.

A second application illustrated in this study is a substantive one. When a course is offered in two different formats, in class and online, it is believed, at least implicitly, that the coverage of the material and the grading of the students are similar, which justifies granting the same credit hours for the same course number, and allowing either form of the course to satisfy graduation requirements. To test this hypothesis, an equivalence test is used, in order to place the burden on the researcher to show that the two forms are indeed equivalent.

Limitations of the study

The first application of equivalence testing, with respect to the homoskedasticity assumption, has several limitations. As it is conducted via a computer Monte Carlo simulation, it is subject to the limitations generally attributed to simulations. The simulation involves the generation of random variables, which are intended to be samples from a theoretical distribution of infinite size (Stevenson, 2008, p. 782). This is done in order for the results to be generalizable to any such samples. One limitation is that each run of the same simulation will result in slightly different results. If the results are vastly different from one run to another, it is difficult to interpret the results unequivocally. By building into the simulation a large enough number of iterations per case scenario explored, it is possible to address this limitation and reduce the variability observed from one run to the next. Probably the greatest limitation of simulations is that they are completely dependent on the choice and settings of their parameters. If a simulation is set up in a way that does not match the real situations it is intended to represent, the simulation results may not be useful as they may not be applicable. This limitation can be addressed by setting up the simulation carefully and realistically, and by varying the parameter values sufficiently so that relevant case scenarios are not omitted. Another limitation of this simulation study is that it pertains to the formal testing of the assumption of variance homogeneity, which can be performed in order to decide which form of the t-test to use. Providing a good decision rule for this purpose is helpful only for researchers who do follow this decision-making path. For researchers whose policy to address possible doubts about homoskedasticity is, for instance, to always use a

nonparametric equivalent of the t-test (e.g., Mann-Whitney U-test), this decision rule, improved or not, is irrelevant.

The illustrative application of equivalence testing to the comparison of the in-class and online forms of a course also has limitations. While it is appropriate to use an equivalence test for this purpose, the data at hand includes only one dependent variable, the final scores given to the students for the course. This is of course an important variable, but it does not capture all the dimensions in which the course forms could be thought to be equivalent. Furthermore, the students are not randomly assigned to one form of the course or the other, so it cannot be ruled out that dimensions other than course form have a systematic effect on students' scores. Even if possible confounders were identified, the currently documented methodology of equivalence testing only allows for one independent variable at a time, a consequence of the development of this methodology within the context of clinical trials, where the ability to randomly assign the subjects to treatment levels eliminates the necessity (though possibly not the benefit) of including confounders as covariates in the model.

CHAPTER 2

REVIEW OF LITERATURE

Prior to the adoption of a special methodology for equivalence testing in the late 1980s, the methodology used to demonstrate bioequivalence relied on conventional difference testing. With this approach, failing to reject the null hypothesis of no difference is taken as evidence of equivalence. It soon became apparent that some studies were leading to conclusions of equivalence for the wrong reason: insufficient power prevents rejecting the null hypothesis, even when the point estimates for the treatment levels are quite far apart. To address the problem, this difference testing methodology was augmented with prospective power studies. This extension of the difference testing approach has been called the “power approach”. By requiring a minimum amount of power, this approach addresses the problem stemming from possibly low power.

The power approach to demonstrating equivalence is to say that if a difference of d exists and we have n observations and a significance level of α , then we have $1-\beta$ chance of finding the hypothesized difference. If, after the analysis is complete, we do not find a significant difference then we can say the evidence favors a conclusion of equivalence. (Stegner, 1996, p. 194)

However, the opposite problem remained: With a large sample size, it is quite possible that a very small difference will be found to be statistically significant, which would lead to a conclusion of non-equivalence, when in fact the difference is inconsequential. According to Ormsby (in Jackson, 1994), “It must be recognized that two formulations can never be identical (...) Therefore, it is the purpose of the bioequivalence study to demonstrate that the profiles produced by the formulations under

study do not differ significantly” (p. 2). The issue of what constitutes close enough to be considered bioequivalent (even if statistically different with a powerful test) has been addressed and resolved with a guideline of 20%, which should be superseded when the particulars of the study offer more meaningful equivalence limits.

In the pharmaceutical industry a standard for equivalence has evolved. That standard is that on agreed upon variables, the population mean of the test compound (μ_T) must be within 20% of the mean of reference compound (μ_R). (Stegner, 1996, p. 194)

Thus, the power approach involves the following three steps:

1. Set limits of equivalence, either 20% or a more appropriate value specific to the particular characteristics of the study. The proper setting of these limits is a substantive issue, not a statistical one.
2. Conduct a prospective power study, so as to ensure that the study would have enough power (80%) to detect an effect outside the limits of equivalence.
3. Conduct a statistical difference test, with a null hypothesis of no difference, and conclude in favor of equivalence if the null hypothesis is not rejected or if the point estimate of the difference and its confidence interval are included within the equivalence limits.

In 1987, Schuirmann published an article showing that even with the modifications outlined above, the difference testing approach is an awkward and ultimately ill-suited way of demonstrating equivalence. Among other problems, it does not offer the familiar correspondence between a low p-value and the desired conclusion (equivalence), and the rejection regions of the statistical tests do not have satisfactory statistical properties. Hauck and Anderson (1986) state that equivalence testing “provides

a way of quantifying (with p-values) what was actually determined from the study instead of saying what the study may or may not have accomplished with some degree of certainty (power).” (p. 203). Olmi (in Jackson, 1994) notes that “ this procedure [power approach] also has very poor operating characteristics in that studies with good precision may be rejected more often than studies with the same estimates but worse precision. This is opposite to what a rational procedure should do” (p. 10).

Schuirmann proposed instead that true equivalence tests be devised and conducted, where the null hypothesis is that of non-equivalence, and the alternative hypothesis is that of equivalence. Schuirmann’s method allows the researcher, when the desired conclusion of equivalence is reached, to obtain a p-value pertaining to the rejection of the null hypothesis of non-equivalence, which makes positive results interpretable much in the same way as they are when a traditional difference testing results in the rejection of null hypotheses of no difference. An additional attractive aspect of Schuirmann’s approach is that it gives results that are fully consistent with the confidence interval approach advocated by Westlake in 1976 (Stegner, 1996, p. 195).

The Two One-Sided Test (TOST) method proposed by Schuirmann has since been accepted as the standard method for equivalence testing, at least for experimental designs involving the random assignment of subjects to two treatment levels and the comparison of means or proportions. According to Dixon (in Newman & Strojan, 1998), “one strength of the two one-sided tests method is that it can be generalized to many circumstances” (p. 281).

Since the introduction of Schuirmann’s TOST method, the reliance on traditional difference testing as a means to provide evidence of equivalence has been discouraged in

medical and pharmaceutical research. In 1998, the International Conference on Harmonisation issued the following recommendation (ICH 09, cited by Lewis, 1999).

Concluding equivalence or non-inferiority based on observing a non-significant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is inappropriate (p. 1920).

In the context of pharmaceutical research, the decisions made based on equivalence testing can be summarized in Table 1 below (Metzler, in Welling et al., p. 56):

Table 1. *Truth Table for Bioequivalence Decisions*

| Truth table for bioequivalence decisions (H0: Two formulations not bioequivalent) | | TRUTH | |
|--|---|---|---|
| | | Bioequivalent | Not bioequivalent |
| DECISION | Bioequivalent (reject H0) | Right decision: Everyone gains (Power) | Wrong decision: Consumer loses (Type I error) |
| | Not bioequivalent (do not reject H0) | Wrong decision: Sponsor loses (Type II error) | Right decision: Consumer gains (Confidence) |

It is important to note that switching from the so-called “power approach” to the TOST method of equivalence testing does not eliminate the need for sufficient power. Indeed, the TOST method provides the researcher not only with a p-value relevant to a correctly stated set of null and alternative hypotheses, but with an estimate of the power

associated with the rejection of the null hypothesis. Like difference tests, equivalence tests need to have enough power, in order to keep the rate of Type II error sufficiently low. Here, insufficient power could lead to failing to conclude in favor of the research hypothesis (equivalence), even if the null hypothesis of non-equivalence were in fact false. This aspect of equivalence testing constitutes its main advantage, as it prevents an undesirable feature of a test (low power) to result in a desired outcome (concluding in favor of the research hypothesis of equivalence).

Difference testing and equivalence testing are not mutually exclusive (Allen et al, 2006, p. 77), and the conclusion reached via a difference test and the conclusion reached via an equivalence test are not redundant. Parkhust (2001, p. 1053) recommends using them concurrently.

These tests [of equivalence] can also be used (especially in basic science) in a second way, for which I use the term reverse tests. In this form of use, they are applied after a no-effect hypothesis test has failed to yield a “significant” result. Then, the reverse test can help in determining whether the original result could reasonably be interpreted as demonstrating lack of an important response, or whether it should more accurately be taken as evidence for lack of certainty caused by some combination of high variability and inadequate sample size.

Using Westlake's confidence interval approach, we can see how, depending on the power of the tests, four distinct case scenarios can occur, as depicted in Figure 1 below:

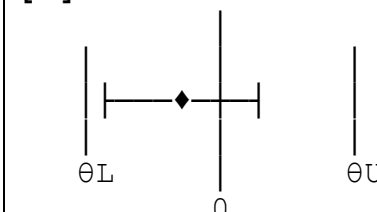
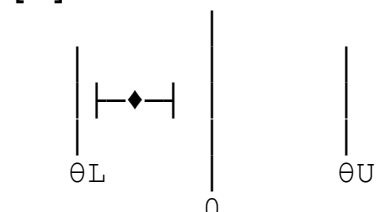
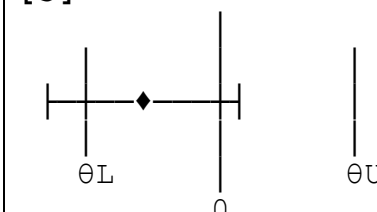
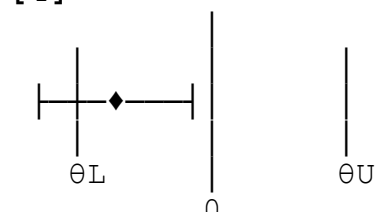
| | | Different | |
|------------|-----|---|---|
| | | No | Yes |
| Equivalent | Yes | [1]  | [2]  |
| | No | [3]  | [4]  |

Figure 1. Possible combinations of difference and equivalence results.

If the confidence interval is located strictly between the equivalence limits, then we can reject the null hypothesis of non-equivalence and conclude in favor of equivalence. The case-scenarios illustrated above in Figure 1 do not pertain to any particular actual data set, but are an illustration of different possible cases where the parameter of interest (e.g., sample mean) and its confidence interval may vary, with the related consequences in terms of difference and equivalence.

Improperly relying on statistical significance to draw conclusions about practical significance, we may mistakenly consider case [3] to be evidence of equivalence, or case [2] to be evidence of nonequivalence. Indeed, case [2], where groups are found to be

significantly different yet also significantly equivalent, serves to remind us of the need to explicitly define what constitutes a practical or consequential difference. Omitting this important step, and relying on statistical significance as a proxy for practical significance leads to difficulties in interpreting an apparently paradoxical situation of the type in case [2]. Misplaced concerns about “too much power” (McBride, 1999, p. 20) underscore the need to keep in mind the distinction between practical significance and statistical significance, and to regard high power as the desirable characteristic that it really is. The problems inherent in the confusion between statistical significance and practical significance have been well researched and presented by Ziliak and McCloskey (2007). The fact that equivalence testing requires that equivalence limits be set prior to any statistical considerations is a positive aspect of this approach as it helps avoid this confusion: Setting equivalence limits constitutes an operational definition of what “equivalent” means in a particular research context, based on substantive considerations, without dependence on sample size and statistical significance.

It should be noted that the very practice of null-hypothesis statistical testing proposed by Ronald Fisher (1925) has been thoroughly criticized over the years. Morrison and Henkel (1970) have documented a wide range of objections to this practice. Gigerenzer (1989) has described how the modern application of statistical testing is an awkward hybrid of the Fisher tradition and the Neyman-Pearson tradition, one that the original authors would probably not sanction. Nevertheless, null-hypothesis statistical testing continues to be the dominant paradigm in the social sciences, and has occasionally received additional theoretical support from its proponents, such as Chow (1996). Equivalence testing is a testing approach that is fully within the null-hypothesis tradition.

It is not an alternative to this tradition, but it does involve, as we have seen above, a null hypothesis that is a range hypothesis, even in the case of nondirectional testing. In that, it addresses one of the criticisms of null hypothesis testing, namely that with sufficient measurement precision and sample size, a point null hypothesis will surely be rejected, so its testing is not informative (Meehl, 1967).

This shortcoming of the point null hypothesis was recognized by Meehl, and later addressed by Serlin and Lapsley (1985), who recommended that a “good-enough belt” be placed around the null hypothesis, in order to make it less of a strawman hypothesis. The “good-enough” approach is similar to equivalence testing in that its null hypothesis, even in the case of nondirectional testing, is a range hypothesis (as opposed to a point hypothesis), but it is different from equivalence testing, in that its null hypothesis includes the no-difference point. Figure 2 below provides a visual description of these similarities and differences.

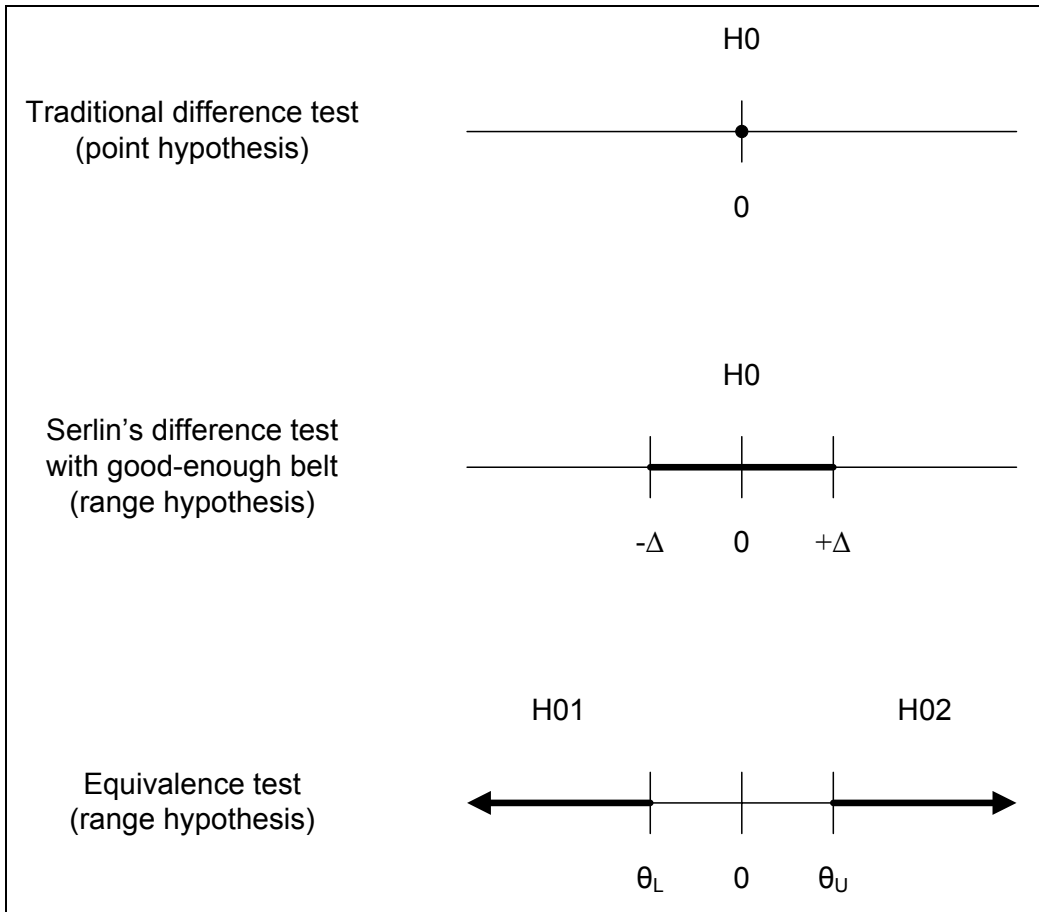


Figure 2. Comparison of null hypotheses (nondirectional testing).

CHAPTER 3

METHOD

Formal definition of equivalence testing

Equivalence testing is not about proving the null hypothesis (Streiner, 2003, p. 756) or even about accepting the null hypothesis (Frick, 1995, p. 132), it is rather a restatement of the null hypothesis so that the alternative hypothesis of equivalence matches the research hypothesis that two (or more) population parameters are close enough to be called equivalent.

In the context of medical research, Lewis (1999, p. 1940) provides the following definition of an equivalence trial.

A trial with the primary objective of showing that the response to two or more treatments differs by an amount which is clinically unimportant. This is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence margin of clinically acceptable differences.

When the research hypothesis is that one group (e.g., new generic product) is no worse than another group (e.g., established branded product), it is possible to omit one side of the equivalence test, and conduct only the single-sided alternative hypothesis of non-inferiority. Non-inferiority tests are simpler and provide more power than equivalence tests, and for that reason they are recommended for studies with such research hypotheses. This thesis focuses on the general case of equivalence testing, of which non-inferiority testing is a special case.

For the nondirectional (2-sided) comparison of two means, the conventional difference testing approach involves the following hypotheses:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

where μ_1 is the population mean for the first group (e.g., test group), and μ_2 is the population mean for the second group (e.g., reference group). Failing to reject this difference-test null hypothesis would lead to concluding that the groups are equivalent, with the problems mentioned above.

By contrast, the equivalence testing approach, proposed in 1987 by Schuirmann as the Two One-Sided Tests (TOST) method, involves a redefinition of the statistical hypotheses, in order for the alternative hypothesis to match the research hypothesis of equivalence:

$$H_0: \mu_1 - \mu_2 \leq \theta_L \text{ or } \mu_1 - \mu_2 \geq \theta_U$$

i.e., the null hypothesis of non-equivalence is retained if $\mu_1 - \mu_2$ is less than the lower equivalence limit, or if $\mu_1 - \mu_2$ is more than the upper equivalence limit.

$$H_A: \theta_L < \mu_1 - \mu_2 < \theta_U$$

i.e., the null hypothesis of non-equivalence is rejected if we conclude that $\mu_1 - \mu_2$ is both more than θ_L and less than θ_U .

Rejecting this equivalence-test null hypothesis leads to the conclusion that the groups are equivalent, with respect to the equivalence limits θ_L and θ_U . Prior to conducting the tests of significance, the equivalence testing approach, like the power approach, requires that the equivalence limits, θ_L and θ_U , be set. According to Rogers et al. (1993, p. 553), "Any difference small enough to fall within that equivalence interval

would be considered clinically and/or practically unimportant”. The basis for setting these limits is substantive, not statistical. They should be defined so as to be meaningful in the research area of interest, and could certainly depart from the 20% range sometimes suggested in pharmaceutical research. This point was emphasized by Schuirmann (1987): The specification of θ_L and θ_U “is made by the experts in the fields” (p. 659).

Once the equivalence interval (θ_L, θ_U) is set, the equivalence testing approach diverges from the power approach, in that the null and alternative hypotheses are split into two separate tests, in accordance to Schuirmann’s TOST method.

The first test is the “right side” test, in the sense that it is a right-side directional (one tail) test, centered on the lower equivalence limit. Rejecting H_{01} on the right side leads to the conclusion that the difference between the population means is more than the lower equivalence limit.

Test 1 (right side test involving the lower equivalence limit):

$$H_{01}: \mu_1 - \mu_2 \leq \theta_L$$

$$H_{A1}: \mu_1 - \mu_2 > \theta_L$$

The second test is the “left side” test, in the sense that it is a left-side directional test, centered on the upper equivalence limit. Rejecting H_{02} on the left side leads to the conclusion that the difference between the population means is less than the upper equivalence limit.

Test 2 (left side test involving the upper equivalence limit):

$$H_{02}: \mu_1 - \mu_2 \geq \theta_U$$

$$H_{A2}: \mu_1 - \mu_2 < \theta_U$$

The population value of $\mu_1 - \mu_2 = \delta$ is estimated in a sample by d :

$$d = \bar{X}_1 - \bar{X}_2$$

Using the example of a two-sample pooled-variance t test, the two tests are:

$$t_1 = \frac{d - \theta_L}{SEdbm}$$

and

$$t_2 = \frac{d - \theta_U}{SEdbm}$$

where the standard error of the difference between means (SEdmb) is:

$$SEdbm = \sqrt{S_{pooled}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

and the pooled variance is:

$$S_{pooled}^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

The critical value (95% confidence level) for the first one-sided test (right-side) is the value cv_1 for which:

$$\int_{cv_1}^{\infty} pdfct(x, df) dx = .05$$

where pdfct is the probability density function of the central t distribution with df degrees of freedom (n_1+n_2-2) and x is the integration variable.

The p-value for the first test is a function of t_1 and df :

$$p_1(t_1, df) = \int_{t_1}^{\infty} pdfct(x, df) dx$$

where t_1 is the value of the test statistic (t test) for the first test.

Power for the first test is a function of cv_1 , df , and t_1 :

$$power_1(cv_1, df, t_1) = \int_{cv_1}^{\infty} pdfnct(x, df, t_1) dx$$

where $pdfnct$ is the probability density function of the noncentral t distribution with df degrees of freedom and a noncentrality parameter value of t_1 .

The critical value for the second one-sided test (left side) is the value cv_2 for which:

$$\int_{-\infty}^{cv_2} pdfct(x, df) dx = .05$$

P-value and power for the second test are the following functions:

$$p_2(t_2, df) = \int_{-\infty}^{t_2} pdfct(x, df) dx$$

$$power_2(cv_2, df, t_2) = \int_{-\infty}^{cv_2} pdfnct(x, df, t_2) dx$$

Upon the completion of these two one-sided t tests, we need to compute the p-value for the overall test of equivalence. The p-value for the overall test of equivalence is the probability that the overall test results in a Type-1 error ($T=1$). This happens only when both tests result in a Type-1 error. If we define T1 as the test with the smallest p-value, then the overall p-value is:

$$p = P(T = 1) = P(T1 = 1 \cap T2 = 1) = P(T2 = 1) \cdot P(T1 = 1 | T2 = 1)$$

If T2 results in a Type-1 error ($T2=1$), then *a fortiori* T1 results in a Type-1 error ($T1=1$) since its p-value is smaller, so we know that $P(T1=1|T2=1) = 1$ and the expression simplifies to:

$$p = P(T2 = 1) \cdot P(T1 = 1 | T2 = 1) = P(T2 = 1) \cdot 1 = P(T2)$$

So the overall p-value p is $P(T2)$, the larger of the two p-values obtained. The rationale for this is presented by Rogers et al. (1993, p. 554).

In an equivalency test, both test statistics must be significant to lead an experimenter to reject the null hypothesis. Because the first test statistic and the second test statistic have to be significant by chance for a Type I error to be made, the experimenter in this case would multiply the Type I error probabilities of each test statistic to calculate the overall probability of making a Type I error. However (...) the probability of the larger (absolute value) of the two test statistics being significant, given that the smaller of the two is significant, equals 1. That is, the overall Type I error probability is the Type I error probability of the smaller test statistic multiplied by the conditional Type I error probability of the larger statistic, which will always be 1.

Similarly, the smallest of $power_1$ and $power_2$ is the overall power for the TOST:

$$p = \max(p_1, p_2)$$

$$power = \min(power_1, power_2)$$

When the overall p-value is less than alpha, the null hypothesis of non-equivalence is rejected, and the researcher concludes that the two population parameters of interest are equivalent, that is, the difference between them is likely to lie, in the population, within the equivalence interval.

Numerical example of an equivalence test

Using a small data set from Mendenhall et al. (2007, p. 409), we can illustrate equivalence testing with an application of the TOST version of a two-sample t test. The equivalence limits are arbitrarily set at -5 and 5. Mathcad worksheets can be used to present the necessary calculations to obtain p-value and power estimates.

| | |
|---|--|
| $\text{data} := \begin{pmatrix} 32 & 35 \\ 37 & 31 \\ 35 & 29 \\ 28 & 25 \\ 41 & 34 \\ 44 & 40 \\ 35 & 27 \\ 31 & 32 \\ 34 & 31 \end{pmatrix}$ | $\theta L := -5 \quad \theta U := 5 \quad \alpha := .05$ |
| | $X1 := \text{data}^{(1)} \quad X2 := \text{data}^{(2)}$ |
| | $\text{mean_X1} := \text{mean}(X1) = 35.222 \quad n1 := \text{rows}(X1) = 9$ |
| | $\text{mean_X2} := \text{mean}(X2) = 31.556 \quad n2 := \text{rows}(X2) = 9$ |
| | $d := \text{mean_X1} - \text{mean_X2} = 3.667 \quad n := n1 + n2 = 18$ |
| | $df := n1 + n2 - 2 = 16$ |
| $\text{sqdevX1} := (X1 - \text{mean_X1})^2$ | $SS1 := \sum \text{sqdevX1} = 195.56$ |
| $\text{sqdevX2} := (X2 - \text{mean_X2})^2$ | $SS2 := \sum \text{sqdevX2} = 160.22$ |
| $\text{PoolVar} := \frac{SS1 + SS2}{df} = 22.236 \quad \text{SEdbm} := \sqrt{\text{PoolVar} \cdot \left(\frac{1}{n1} + \frac{1}{n2} \right)} = 2.2229$ | |
| $t1 := \frac{d - \theta L}{\text{SEdbm}} = 3.89878$ | $t2 := \frac{d - \theta U}{\text{SEdbm}} = -0.59981$ |
| $cv1 := \text{qt}(1 - \alpha, df) = 1.74588$ | $cv2 := -\text{qt}(1 - \alpha, df) = -1.74588$ |
| <p>Central t probability density function:</p> | $\text{pdf_ct}(x, df) := \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{\pi \cdot (df)} \cdot \Gamma\left(\frac{df}{2}\right)} \cdot \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}}$ |
| $p1 := \int_{t1}^{\infty} \text{pdf_ct}(x, df) dx = 0.0006385$ | $p2 := \int_{-\infty}^{t2} \text{pdf_ct}(x, df) dx = 0.2785155$ |

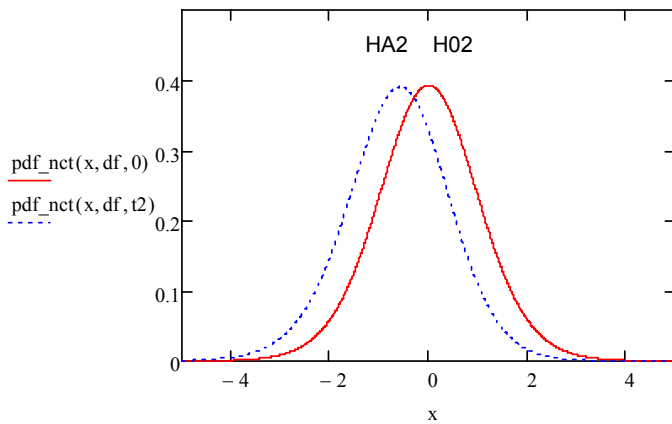
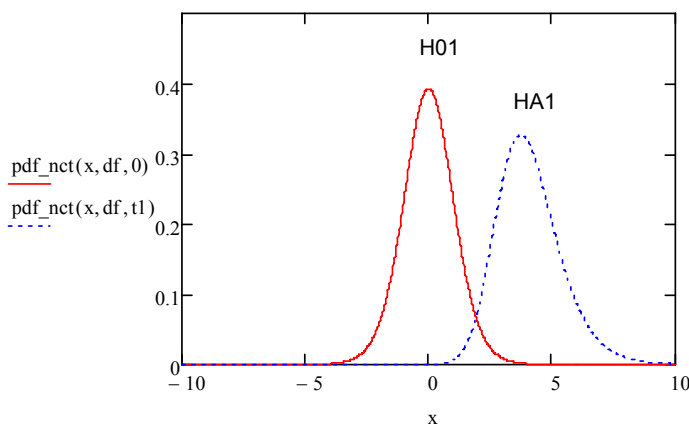
Figure 3. Mathcad worksheet for equivalence test p-value.

Noncentral t probability density function:

$$\text{pdf_nct}(x, \text{df}, t) := \frac{\frac{\text{df}}{\text{df}^2} \cdot \frac{-\text{df} \cdot t^2}{e^{2 \cdot (x^2 + \text{df})}}}{\sqrt{\pi} \cdot \Gamma\left(\frac{\text{df}}{2}\right) \cdot 2^{\frac{\text{df}-1}{2}} \cdot (x^2 + \text{df})^{\frac{\text{df}+1}{2}}} \int_0^{\infty} y^{\text{df}} \cdot e^{-\frac{1}{2} \cdot \left(y - \frac{t \cdot x}{\sqrt{x^2 + \text{df}}}\right)^2} dy$$

$$\text{power1} := \int_{\text{cv1}}^{\infty} \text{pdf_nct}(x, \text{df}, t1) dx = 0.981324$$

$$\text{power2} := \int_{-\infty}^{\text{cv2}} \text{pdf_nct}(x, \text{df}, t2) dx = 0.14228$$



$$p := \max(p1, p2) = 0.27852$$

$$\text{power} := \min(\text{power1}, \text{power2}) = 0.14228$$

Figure 4. Mathcad worksheet for equivalence test power.

The hypotheses for this equivalence test are the following:

$$H_0: \mu_1 - \mu_2 \leq -5 \text{ or } \mu_1 - \mu_2 \geq 5$$

$$H_A: -5 < \mu_1 - \mu_2 < 5$$

Since this p-value (.2785) is greater than alpha (.05), the null hypothesis of non-equivalence is not rejected: it is not sufficiently unlikely that d, in the population, is outside one of the equivalence limits (in this case 5, the upper equivalence limit). In other words, H02, the null hypothesis of non-equivalence for the second one-sided test (left side) cannot be rejected, so the overall null hypothesis of non-equivalence is retained and we must conclude that we do not have sufficient evidence supporting the hypothesis of equivalence.

These results are confirmed with the use of the software package EquivTest 2.0. In the output (presented in Figure 5 below), group 1 is labeled “T” (test group) and group 2 is labeled “R” (reference group).

| Schirmann OST/TOST: | | | | |
|--|------------------|-----------------|--|-----------------|
| Null Hypothesis L: Mean T- Mean R <= Lower Bound = -5.00 | | | | |
| Null Hypothesis U: Mean T- Mean R >= Upper Bound = 5.00 | | | | |
| | t-Value | | One-sided p-value to reject non-equivalence | |
| | Specified | Observed | Specified | Observed |
| Null Hypothesis L t-statistic | 1.7458 | 3.8987 | 0.0500 | 0.0006 |
| Null Hypothesis U t-statistic | -1.7458 | -0.5998 | 0.0500 | 0.2785 |

At least one of the one-sided tests does not reject the null hypothesis, so equivalence to within the specified equivalence bound cannot be claimed.

Figure 5. EquivTest output.

Plots representing the two one-sided tests are shown in Appendix A.

Methodological application: The assumption of homoskedasticity

One of the most common assumptions involved in statistical hypothesis testing is that the within-group variances are equal (homogeneity of variance, a.k.a. homoskedasticity). That is, when groups of subjects are formed based on their values on the independent variable(s), and the variance of the dependent variable is computed for each group, it should be the case that this variance takes on the same value from one group to another, in the population. Therefore, if the assumption is true, group variances in a random sample should not differ from one another by a greater magnitude than one would expect due to sampling error alone. Among the statistical methods relying on this assumption, we find the most widely used methods in Educational Psychology: the t and z tests, the analysis of variance/covariance, simple/multiple linear regression, and structural equation modeling.

The usual way to evaluate this assumption is to conduct a difference test of hypothesis: the null hypothesis is that of no difference between the group variances, and the alternative hypothesis is that of a difference. Such a test of assumption is sometimes referred to as a preliminary test, as it precedes the significance test of main interest. Over the years, several difference tests have been used for this purpose, including Hartley's F-max test, the Bartlett test, and the Levene test. The most commonly recommended test currently is a modified version of the Levene test (Howell, 2002). In his 1960 paper, Levene proposed, in the context of the analysis of variance (ANOVA), to check the assumption of homoskedasticity by computing, for each group, the absolute deviations from the group mean, and then conducting an ANOVA on these deviations. The rationale is that if we cannot reject the null hypothesis of equal deviations among the groups, then

the groups must have similar variances and we can consider that the homogeneity assumption is met. Brown and Forsythe (1974) introduced a modified version of the Levene test, using not deviations from the group means but deviations from the group medians, to improve the robustness of the test, in particular with respect to non-normal data.

Tests of the assumption of homoskedasticity are consequential to the extent that they are used to decide when to switch from the conventional pooled-variance t test to an alternative test that does not assume equal group variance. An alternative test frequently chosen is Satterthwaite's approximate test (1946), which uses separate variance estimates that need not be equal, and relies on the Student's t distribution with adjusted degrees of freedom (Moser et al., 1989). Satterthwaite's approximate test (SAT) is more robust to heteroskedasticity than the pooled-variance t test (PVT) but it has less power in certain conditions. Therefore, applying a sound decision rule to switch from the PVT to the SAT is important as it lets hypothesis tests have the best possible characteristics, in terms of the control of both Type-I and Type-II error rates. The PVT has been found to be quite robust (i.e., providing good control of the Type-I error rate) to heteroskedasticity when the two samples are of equal or nearly equal sizes. When these sample sizes are very different, however, heteroskedasticity can lead to a substantial inflation of the Type-I error rate, therefore in such cases it is useful to have a sound switching rule (Glass, 1966) so that data exhibiting heteroskedasticity and unequal sample sizes can be analyzed with an appropriately robust method.

The Levene test is used for the purpose of testing the homoskedasticity assumption, but being a difference test, it is not well-suited for the task. The main

problem is that it is, as all difference tests are, too conservative (i.e., too unlikely to lead to the rejection of H_0) when the sample size is small. Until sufficient evidence is available, the conclusion is that the variances are equal. It is reasonable to want a test to be conservative, and to retain the null hypothesis in the absence of sufficient evidence against it, when the course of action resulting from retaining the null hypothesis may be construed as the safest (e.g., not adopting a new unproven drug). In this case, however, the safest course of action would involve controlling alpha sufficiently well, and would result from considering the variances to be unequal and using the SAT instead of the PVT. With a small sample, however, sizeable differences in variances may not be found significant if one relies on a difference test. This is the opposite of what a sound decision rule should do, since using the PVT when the homoskedasticity assumption is not met has worse consequences with a small sample than with a large sample. As Hays (1994, p. 303) puts it: “In circumstances where they are needed most (small samples), the tests for homogeneity are poorest”. Conversely, when the sample size is larger, a difference test will be more likely to result in the rejection of the null hypothesis of equal variances, even if the difference is of a relatively small magnitude, thereby guiding the researcher toward switching away from the pooled-variance t-test, when in fact the larger sample size would make this switch less necessary. Box (1953, p. 333) describes this latter case scenario with a vivid analogy: “To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!”.

The hypothesis reversal involved in equivalence testing is helpful in this case, as it leads the researcher to the safest course of action when the sample is small: Slightly

less power but good protection against inflation of the Type-I error rate. When a larger sample size is involved, it is more likely that the equivalence test will lead to the rejection of the null hypothesis of unequal group variances, and provide the researcher with no reason to switch away from the pooled-variance t-test.

Over the last one hundred years, the Student's t-test (Student, a.k.a., William S. Gossett, 1908) has been the most widely used statistical test in behavioral sciences (Zabell, 2008), and it is also the simplest parametric statistical test. As such, it is a good choice for an illustration of the relative merits of difference testing vs. equivalence testing, as it is used extensively in applied research and its simplicity means that there is no modeling complication (e.g., variable selection in multiple regressions) that could introduce unnecessary ambiguity in the interpretation of the results. It can also be construed as a special case of the analysis of variance, therefore the Levene test, designed for the ANOVA, is readily applicable to it.

When conducting a preliminary test of homoskedasticity, prior to a two-sample t-test of substantive interest, setting up this preliminary test as an equivalence test can overcome the problem mentioned above. By reversing the null and alternative hypotheses, the equivalence test will lead the investigator toward the test alternative (SAT) that is safer, in terms of Type-I error rate, when the sample size is small, but toward the more powerful alternative (PVT) when the sample size is larger and there is no drastic departure from the homoskedasticity assumption. In addition, the equivalence testing approach requires setting up equivalence limits. These limits can be set so as to reflect the threshold of variance ratios beyond which the PVT is likely to result in excessive inflation of the Type-I error rate. Therefore, we can ask whether conducting the

Brown-Forsythe modified Levene test in the usual difference-testing approach (H_0 : equal variances) constitutes a worse decision rule than conducting it as an equivalence test (H_0 : non-equivalent variances) for this purpose, in terms of power and control of the Type-I error rate

Computer simulations

This question of which decision rule should be preferred can be answered with two sets of computer simulations. The first set is intended to compare the relative performances of the PVT and the SAT when the two group variances are equal and when they are not, under varying conditions of sample size and sample allocation. This helps identify appropriate equivalence limits for the equivalence-testing approach to the preliminary testing of homoskedasticity. The second set is intended to compare the respective performances of the Brown-Forsythe modification of the Levene test (BF) vs. the equivalence-testing form of the conventional F-ratio test (ET) as a switching rule between the PVT and the SAT. All simulations are conducted using the Statistical Analysis System (SAS) version 9.1 for Windows XP.

The first set includes two simulations. The first simulation estimates the respective abilities of the PVT and of the SAT to control alpha, under varying combinations of sample allocation and sample variances. The second simulation estimates power for these tests under the same varying conditions.

The second set also includes two simulations. These simulations are intended to compare the relative performances of the BF (Brown-Forsythe in its usual difference-test form) switching rule and the ET (equivalence-test version of the Brown-Forsythe test) switching rule. The better switching rule is the rule which most often selects the test best

suited for the data at hand, so that applying it results in relatively high levels of control of alpha and of power. The first simulation estimates the ability of these switching rules to control alpha, and the second simulation does the same with respect to power. For both sets of simulations, the varying combinations of sample allocations and sample variances are the following (Table 2):

Table 2. *Combinations of values for sample sizes and variances*

| combination | n1 | n2 | var1 | var2 |
|-------------|----|----|------|------|
| 1 | 50 | 50 | 20 | 20 |
| 2 | 50 | 50 | 25 | 15 |
| 3 | 50 | 50 | 30 | 10 |
| 4 | 50 | 50 | 35 | 5 |
| 5 | 50 | 50 | 37 | 3 |
| 6 | 50 | 50 | 38 | 2 |
| 7 | 40 | 60 | 20 | 20 |
| 8 | 40 | 60 | 25 | 15 |
| 9 | 40 | 60 | 30 | 10 |
| 10 | 40 | 60 | 35 | 5 |
| 11 | 40 | 60 | 37 | 3 |
| 12 | 40 | 60 | 38 | 2 |
| 13 | 30 | 70 | 20 | 20 |
| 14 | 30 | 70 | 25 | 15 |
| 15 | 30 | 70 | 30 | 10 |
| 16 | 30 | 70 | 35 | 5 |
| 17 | 30 | 70 | 37 | 3 |
| 18 | 30 | 70 | 38 | 2 |
| 19 | 20 | 80 | 20 | 20 |
| 20 | 20 | 80 | 25 | 15 |
| 21 | 20 | 80 | 30 | 10 |
| 22 | 20 | 80 | 35 | 5 |
| 23 | 20 | 80 | 37 | 3 |
| 24 | 20 | 80 | 38 | 2 |
| 25 | 10 | 90 | 20 | 20 |
| 26 | 10 | 90 | 25 | 15 |
| 27 | 10 | 90 | 30 | 10 |
| 28 | 10 | 90 | 35 | 5 |
| 29 | 10 | 90 | 37 | 3 |
| 30 | 10 | 90 | 38 | 2 |

For the second set, these simulations are conducted once with the “medium” sample size presented above ($n = 100$) and two additional times, with a total sample size set to be small (total $n = 50$) and large (total $n = 200$). The purpose is to discover the extent to which the relative performances are stable over different values of total sample size, under the same varying conditions of sample allocation.

For both sets of simulations, the control of alpha is estimated by calculating the proportion of rejection of the null hypothesis of equality between the group means when the group means are equal in the population from which the samples are drawn, and power is estimated by calculating this same proportion when the group means are different in the population.

Pseudo code for the first set of simulations: PVT vs. SAT

Simulation 1.1. (H0 true): Actual alpha for PVT and for SAT

- Set number of iterations for the simulation (10,000).
- Set fixed simulation parameters: alpha level (.05), total sample size (100), and group means (Mean1=Mean2=50).
- List combinations of values for varying parameters in a SAS data set (n1, n2, var1, var2).
- Begin macro (10,000 iterations per combination).
 - o Generate samples based on set parameters and on varying parameters
 - o Run PVT and count H0 rejections.
 - o Run SAT and count H0 rejections.
- End macro.
- Append output data set with each new iteration and compute the actual alpha as the proportion of tests for which H0 is rejected (given that H0 is true) separately for PVT and for SAT.
- Tabulate and export results

Simulation 1.2. (H0 false): Power for PVT and for SAT

- Set number of iterations for the simulation (10,000).
- Set fixed simulation parameters: alpha level (.05), total sample size (100), and group means (Mean1=50≠Mean2=53).
- List the 30 combinations of values for varying parameters in a SAS data set (n1, n2, var1, var2).
- Begin macro (10,000 iterations per combination).
 - o Generate samples based on set parameters and on varying parameters
 - o Run PVT and count H0 rejections.
 - o Run SAT and count H0 rejections.
- End macro.
- Append output data set with each new iteration and compute power as the proportion of tests for which H0 is rejected (given that H0 is false) separately for PVT and for SAT.
- Tabulate and export results.

Simulations 1.1. and 1.2. are implemented with, respectively, the SAS programs PVT_SAT_alpha.sas and PVT_SAT_power.sas (see Appendix B).

Pseudo code for the second set of simulations

We are comparing the traditional form of the Brown-Forsythe modification of Levene's test with the equivalence testing form, to evaluate homoskedasticity.

Simulation 2.1. (H0 true): Actual alpha (i.e., Type-I error rate) for the BF rule and the ET rule (for switching between PVT and SAT)

- Set number of iterations for the simulation (10,000).
- Set fixed simulation parameters: alpha level (.05), total sample size (50, 100, 200), and group means (Mean1=Mean2=50).
- List combinations of values for varying parameters in a SAS data set (n1, n2, var1, var2).
- Begin macro (10,000 iterations per combination).
 - o Generate samples based on set parameters and on varying parameters
 - o Run PVT and SAT
 - o Run the Brow-Forsythe test of homoskedasticity, if pass ($p \geq .05$) count H0 rejection using the PVT, if fail ($p < .05$) count H0 rejection using the SAT. These rejections accumulate in the BF counter.
 - o Run the Equivalence Testing test of homoskedasticity, if pass ($p < .05$) count H0 rejection using the PVT, if fail ($p \geq .05$) count H0 rejection using the SAT. These rejections accumulate in the ET counter.
- End macro.
- Append output data set with each new iteration and compute the actual alpha as the proportion of tests for which H0 is rejected (given that H0 is true) separately

for BF and for ET. Note when BF and ET rules agree on the recommendation for using PVT or SAT.

- Tabulate and export results.

Simulation 2.2. (H0 false): Power (i.e., $1 - \text{Type-II error rate}$) for the BF rule and the ET rule (for switching between PVT and SAT)

- Set number of iterations for the simulation (10,000).
- Set fixed simulation parameters: alpha level (.05), total sample size (50, 100, 200), and group means (Mean1=50≠Mean2=53).
- List combinations of values for varying parameters in a SAS data set (n1, n2, var1, var2).
- Begin macro (10,000 iterations per combination).
 - o Generate samples based on set parameters and on varying parameters
 - o Run PVT and SAT
 - o Run the Brow-Forsythe test of homoskedasticity, if pass ($p \geq .05$) count H0 rejection using the PVT, if fail ($p < .05$) count H0 rejection using the SAT. These rejections accumulate in the BF counter.
 - o Run the Equivalence Testing test of homoskedasticity, if pass ($p < .05$) count H0 rejection using the PVT, if fail ($p \geq .05$) count H0 rejection using the SAT. These rejections accumulate in the ET counter.
- End macro.
- Append output data set with each new iteration and compute power as the proportion of tests for which H0 is rejected (given that H0 is false) separately for

BF and for ET. Note when BF and ET rules agree on the recommendation for using PVT or SAT.

- Tabulate and export results.

Substantive application: In class vs. online forms of the same course

Many research questions in Educational Psychology can best be investigated with difference testing, where the alternative hypothesis is that of a new “treatment” (in the broadest sense) being superior to an established treatment (active control) or to no treatment at all. In certain instances, however, the research hypothesis may be one of equivalence. The reason equivalence testing is appropriate in the context of the introduction of a new generic drug is that the generic drug is already known to be superior in an important dimension (affordability), therefore it is sufficient to demonstrate that its effectiveness (the dependent variable) is equivalent to that of a more costly branded drug. Similarly, there are instances in Educational Psychology where one condition, known to be superior in some way, needs only to attain equivalence on the dependent variable of interest to be considered an attractive option. Some of the research areas where equivalence testing is beginning to be applied tend to be closer to Educational Research than the clinical trials where equivalence testing originated, for instance in that they do not allow random assignment. Examples of such research can be found in epidemiology, such as the study conducted by Barker et al. (2002), where equivalence testing is used to assess the extent to which public health policy has been effective in eliminating differences in immunization coverage that occur by gender, race/ethnicity, income, and geographic location. This suggests a need to extend equivalence testing beyond its original context of randomized trials.

Instruction, particularly at the university level, is increasingly provided in settings other than the conventional classroom. Online courses are currently gaining popularity, as

they offer easier access and more flexible schedules to students who may not otherwise be able to meet in a classroom setting. Given that these alternative settings have a clear advantage in terms of providing convenience and easier access, they may be shown to be a sensible choice, if they are found to be reasonably equivalent to the classroom setting in terms of their educational content and delivery. Several research studies have already investigated this issue (Lee et al., 2005; Terry, 2007), but without the use of equivalence testing. Equivalence tests are well-suited to the task of providing evidence that these alternate settings are, if not superior, at least comparable to the classroom settings, with respect to certain dependent variables such as grades or student evaluations.

Studies currently under way for which equivalence testing is applicable can generate data suitable for both equivalence testing and difference testing. The results can be interpreted in terms of the four possible difference/equivalence outcomes outlined in Figure 1 above (p. 12). For example, the same course offered at the University of Minnesota in a classroom version versus an online version (e.g., Public Health 6414: Biostatistical Methods I) can be examined in this manner.

The research question is whether the online version can be shown to be equivalent, in terms of grades, to the classroom version. Equivalence tests are conducted to answer this question, and they serve to further illustrate the use of equivalence testing in Educational Psychology.

CHAPTER 4

RESULTS AND ANALYSIS

First set of simulations

We are comparing the relative performances of the PVT and the SAT, for a two-group t-test on means.

Simulation 1.1.: Comparison of the PVT and the SAT with respect to the control of alpha

Figures 5, 6, and 7 below show how the PVT is robust to violations of the homogeneity assumption when sample sizes are equal or nearly equal: With a variance ratio of 19 to 1, the actual alpha is still very close to the observed alpha (.0505 vs. .05). It is also true that even with an n ratio as high as 9 to 1, the PVT still returns a barely inflated actual alpha (.0549). But when both the n ratio and the variance ratio are large, the actual alpha exhibits a high degree of inflation: For instance, with an n ratio of 9:1 and a variance ratio of 19:1, the PVT returns an actual alpha of .4502.

By contrast, we see that the SAT fares much better, as its control of alpha is uniformly excellent over the entire set of combinations of n ratios and variance ratios: In no instance does its actual alpha exceed .06.

This confirms that the SAT is indeed a good choice, with respect to minimizing the Type-1 error rate, when the homoskedasticity assumption is not met and when the n ratio differs from 1:1. These results suggest that an optimal switching rule between the PVT and the SAT should be based not only on the variance ratio but also on the n ratio. However, for the purpose of this comparison between the BF rule and an ET rule, only the variance ratio will be used, since the BF test is a test of variances and does not take into account n ratios. Therefore, for the comparison between the two rules to be fair and

meaningful, the ET rule should also be based on variance ratios alone. Inspection of these results leads to the conclusion that a reasonable variance ratio threshold could well be 2:1, since the PVT manages decent control of alpha up to this ratio, even when the n ratio departs greatly from 1:1, but performs much less well beyond that point. Figures 6, 7, and 8 below show the actual alpha values, for the pooled-variance t-test and for the Satterthwaite t-test.

| HO true: Mean1 = 50, Mean2 = 50 | | | | | | | | | | |
|---------------------------------|----|----|---------|--------|--------|--------|--------|--------|--------|---------|
| Pooled-variance t-test | | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | AvVar |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.0549 | 0.1110 | 0.2077 | 0.3310 | 0.4050 | 0.4502 | |
| 100 | 20 | 80 | 4.00 | 0.0508 | 0.0868 | 0.1493 | 0.2279 | 0.2750 | 0.2871 | |
| 100 | 30 | 70 | 2.33 | 0.0511 | 0.0786 | 0.1125 | 0.1590 | 0.1692 | 0.1829 | |
| 100 | 40 | 60 | 1.50 | 0.0468 | 0.0639 | 0.0741 | 0.0941 | 0.1014 | 0.1097 | |
| 100 | 50 | 50 | 1.00 | 0.0506 | 0.0534 | 0.0498 | 0.0499 | 0.0508 | 0.0505 | |
| Satterthwaite t-test | | | | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.0560 | 0.0537 | 0.0541 | 0.0486 | 0.0508 | 0.0485 | |
| 100 | 20 | 80 | 4.00 | 0.0512 | 0.0470 | 0.0489 | 0.0482 | 0.0497 | 0.0519 | |
| 100 | 30 | 70 | 2.33 | 0.0504 | 0.0510 | 0.0493 | 0.0502 | 0.0488 | 0.0477 | |
| 100 | 40 | 60 | 1.50 | 0.0469 | 0.0501 | 0.0497 | 0.0496 | 0.0501 | 0.0508 | |
| 100 | 50 | 50 | 1.00 | 0.0505 | 0.0532 | 0.0495 | 0.0488 | 0.0489 | 0.0483 | |

Figure 6. Simulation results for the control of alpha (PVT and SAT).

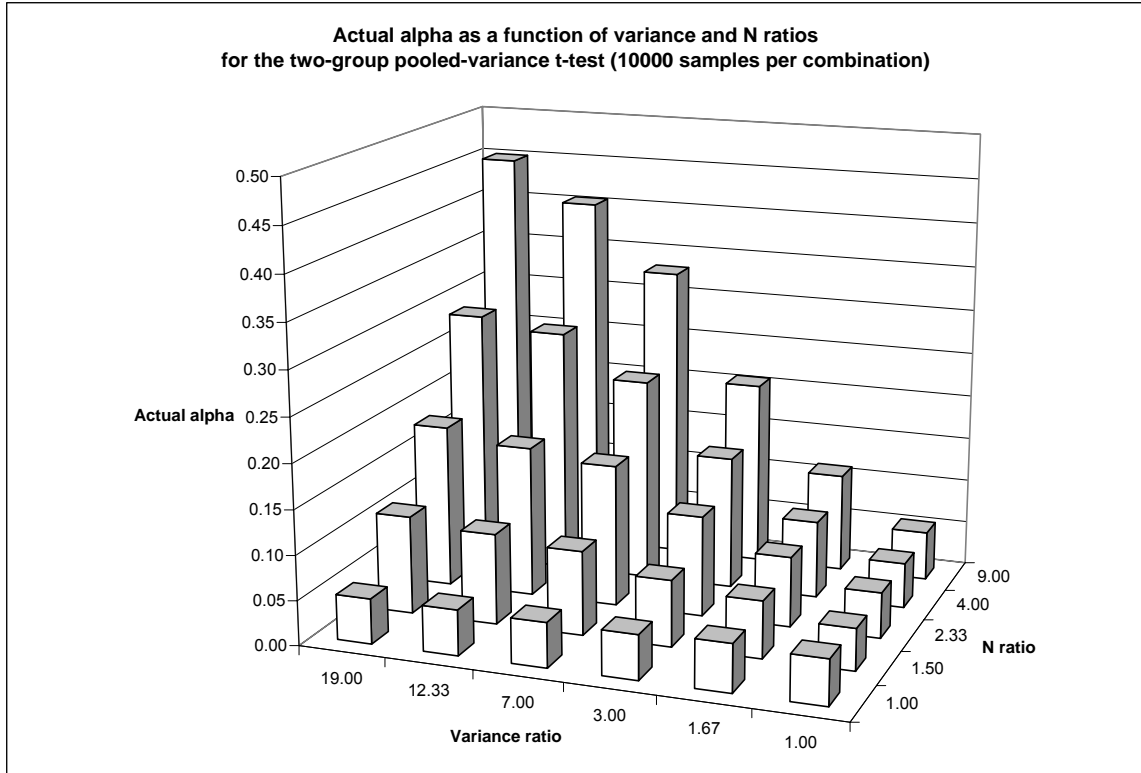


Figure 7. Actual alpha for the PVT.

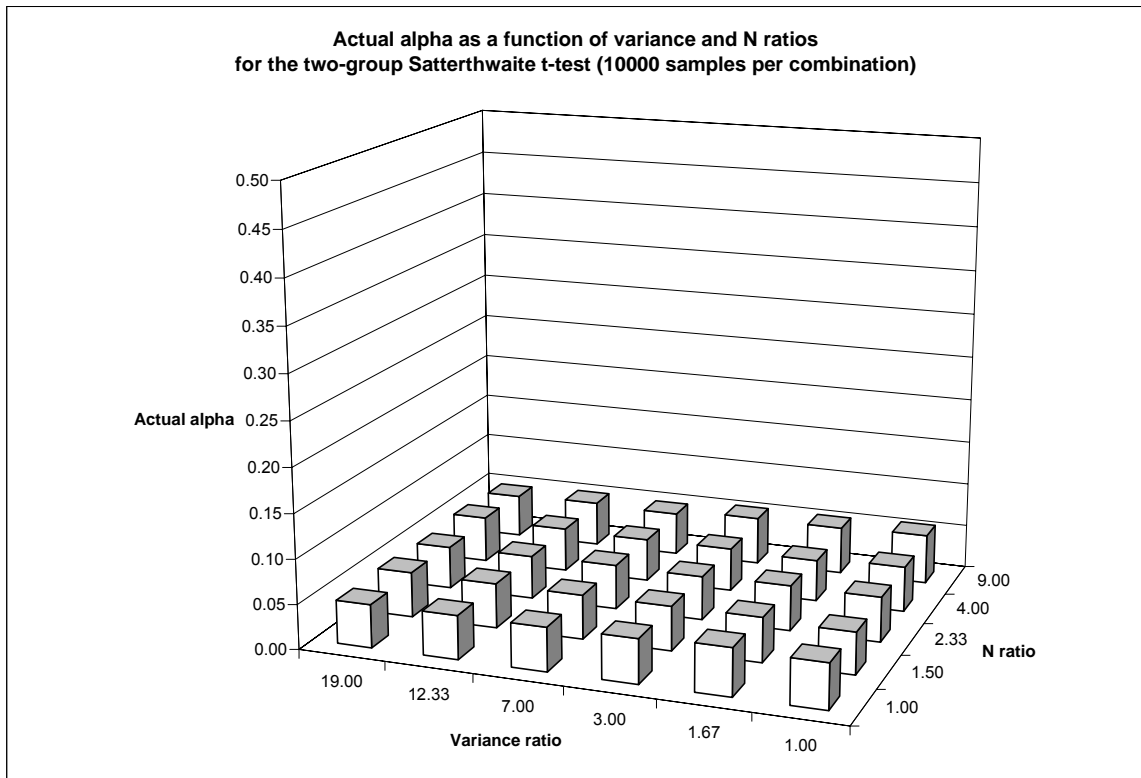


Figure 8. Actual alpha for the SAT.

Simulation 1.2. Comparison of the PVT and the SAT with respect to power

Figures 8, 9 and 10 below show how the PVT generally has more power than the SAT. When the variance ratio is above 2:1, this power advantage would not be consequential, if the rule outlined above is applied: Due to its poor control of alpha when the variance ratio exceeds 2:1, the PVT would not be used, and the researcher would use the SAT instead. When the variance ratio is below 2:1, and thus when the PVT might be used, there is some power advantage in doing so, as the PVT proves more powerful than the SAT, even if this advantage is slight. For instance, when the variance ratio is 5:3 (1.67) and the n ratio is 1:9, the power of the PVT is 58.57% while the power of the SAT is 37.82%. At this variance ratio, the power advantage of the PVT exists at each sample

allocation, although its magnitude is smaller when the sample allocation is more balanced. Figures 9, 10, and 11 below show the power values, for the pooled-variance t-test and for the Satterthwaite t-test.

| HO false: Mean1 = 50, Mean2 = 53 | | | | | | | | | | |
|----------------------------------|----|----|---------|--------|--------|--------|--------|--------|--------|---------|
| Pooled-variance t-test | | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.5092 | 0.5831 | 0.6566 | 0.7338 | 0.7706 | 0.7843 | |
| 100 | 20 | 80 | 4.00 | 0.7556 | 0.7861 | 0.8127 | 0.8452 | 0.8596 | 0.8628 | |
| 100 | 30 | 70 | 2.33 | 0.8617 | 0.8721 | 0.8833 | 0.8933 | 0.9005 | 0.8937 | |
| 100 | 40 | 60 | 1.50 | 0.9092 | 0.9062 | 0.9114 | 0.9100 | 0.9099 | 0.9119 | |
| 100 | 50 | 50 | 1.00 | 0.9161 | 0.9124 | 0.9129 | 0.9158 | 0.9130 | 0.9163 | |
| Satterthwaite t-test | | | | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.4459 | 0.3782 | 0.3340 | 0.3018 | 0.2834 | 0.2826 | |
| 100 | 20 | 80 | 4.00 | 0.7361 | 0.6727 | 0.6074 | 0.5650 | 0.5446 | 0.5396 | |
| 100 | 30 | 70 | 2.33 | 0.8573 | 0.8201 | 0.7814 | 0.7489 | 0.7306 | 0.7177 | |
| 100 | 40 | 60 | 1.50 | 0.9065 | 0.8861 | 0.8726 | 0.8501 | 0.8437 | 0.8423 | |
| 100 | 50 | 50 | 1.00 | 0.9159 | 0.9122 | 0.9118 | 0.9128 | 0.9102 | 0.9133 | |

Figure 9. Simulation results for power (PVT and SAT).

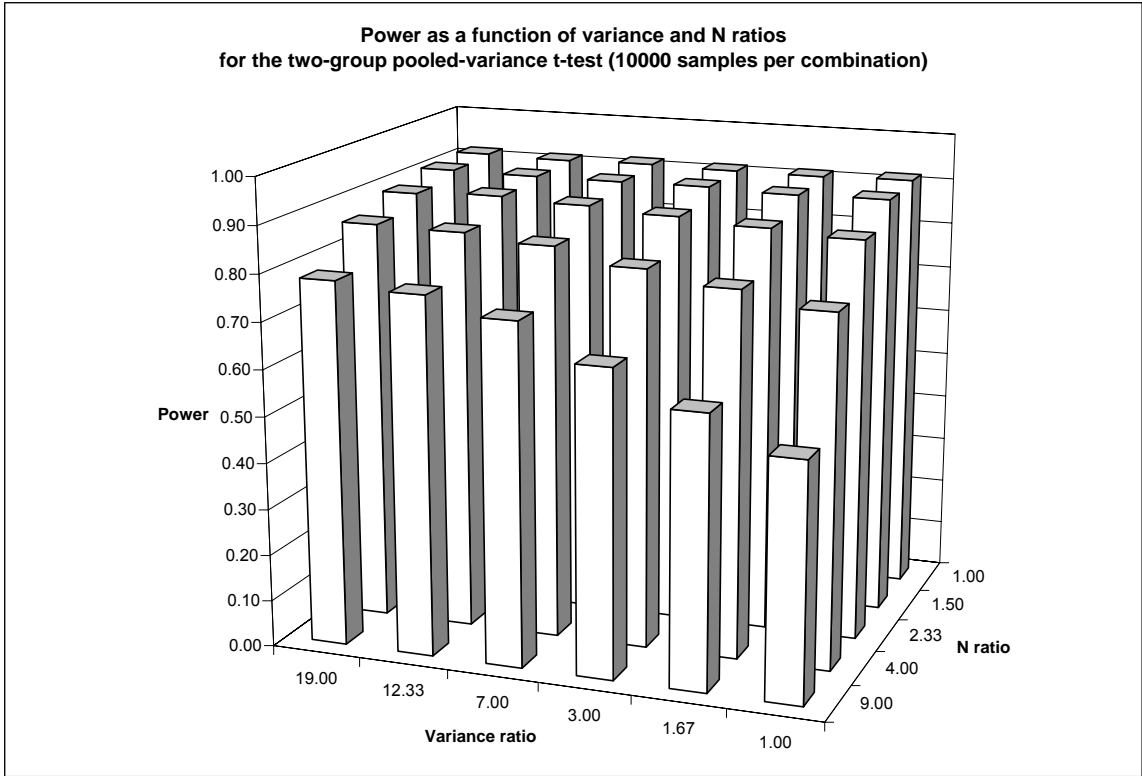


Figure 10. Power for the PVT.

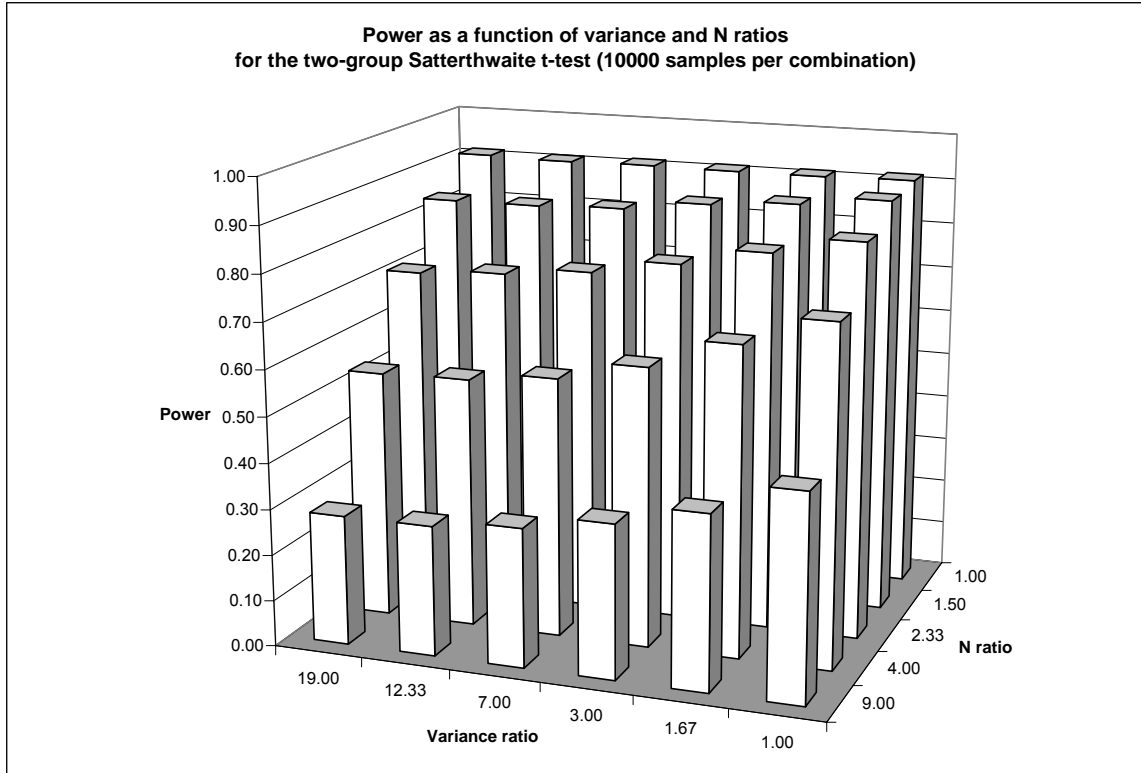


Figure 11. Power for the SAT.

Each estimate (of actual alpha or power) is computed as a proportion: For each simulation, it is the count of H_0 rejections divided by the total number of t-tests performed (number of iterations for that simulation). The large number of iterations results in a narrow confidence interval around these proportion estimates, which means that the results are minimally affected by sampling error.

For instance, the first estimate of actual alpha in Figure 5 is $549/10000 = .0549$. The limits of the 95% confidence interval around this point estimate, using the uncorrected normal approximation, are:

$$LowerLimit = p - z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} = .0549 - 1.96 \cdot \sqrt{\frac{0.051886}{10000}} = 0.050435$$

$$UpperLimit = p + z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} = .0549 + 1.96 \cdot \sqrt{\frac{0.051886}{10000}} = 0.059365$$

The results of the first set of simulations confirm that there is indeed a trade-off between the much greater control of alpha provided by the SAT and the slightly greater power achieved by the PVT. When the homoskedasticity assumption is met, it is preferable to use the PVT, as it provides more power, but when this assumption is not met, it is preferable to use the SAT, as it is more effective in preventing inflation of the actual alpha. Therefore it is sensible to use a data-driven switching rule to choose one or the other form of the t-test, such as using the 2:1 variance ratio as a threshold.

Second set of simulations

We are comparing the relative performances of the BF and the ET rules, with respect to evaluating the homoskedasticity assumption for the purpose of choosing between PVT and SAT

With the BF rule, the null hypothesis is that the variances are equal (assumption met) and therefore that the PVT can be used. If this null hypothesis is rejected, the conclusion is that the variances are not equal (assumption not met) and therefore the SAT should be used. By contrast, with the ET rule the null hypothesis is that the variances are not equal (assumption not met) and therefore the SAT should be used. If this null

hypothesis is rejected, the conclusion is that the variances are equivalent and therefore the PVT can be used. The ET rule requires equivalence limits to be set, and the first set of simulations provided guidance in this respect: a 2:1 ratio appears to be a reasonable threshold, so that the variance ratio lower limit can be set at $\frac{1}{2}$ and the upper limit at 2. In the same way as presented above for the equivalence test for two means, the ET procedure in this second set of simulations involves two one-sided tests. For the first test (right side) the null hypothesis is that the ratio of the group variances ($\text{var1}/\text{var2}$) is, in the population, less than or equal to $\frac{1}{2}$, and for the second test (left side) the null hypothesis is that this ratio is greater than or equal to 2. If both these null hypotheses are rejected, then we conclude that the ratio is, in the population, greater than $\frac{1}{2}$ and smaller than 2, and the overall null hypothesis of non-equivalence is rejected.

The relative performances of the BF and ET rules for switching between the PVT and the SAT are measured, much as in the first set of simulations, but performance is now defined in an indirect way: The control of alpha and power attained using these two switching rules are not those of the variance tests performed. They are, instead, those of the t-test on means, performed using the PVT or SAT forms of the t-test, depending on the recommendation given by the BF and ET rules (derived from the BF and ET tests on variances). In other words, the performance of a switching rule depends on the performance that is observed in the t-test that it has recommended. This is so because the researcher is not primarily concerned with the variance tests *per se*, rather the researcher is concerned with the performance of the t-test, and this performance is in part determined by whether the form of this t-test (PVT or SAT) is appropriate for the data at hand. The best switching rule is the rule that guides the researcher to select the form of

the t-test that is expected to result in the lowest alpha and the highest power for the t-test of substantive interest, given the variance characteristics of the data. Figure 12 below shows how the switching rules are assessed, in terms of their relative performance.

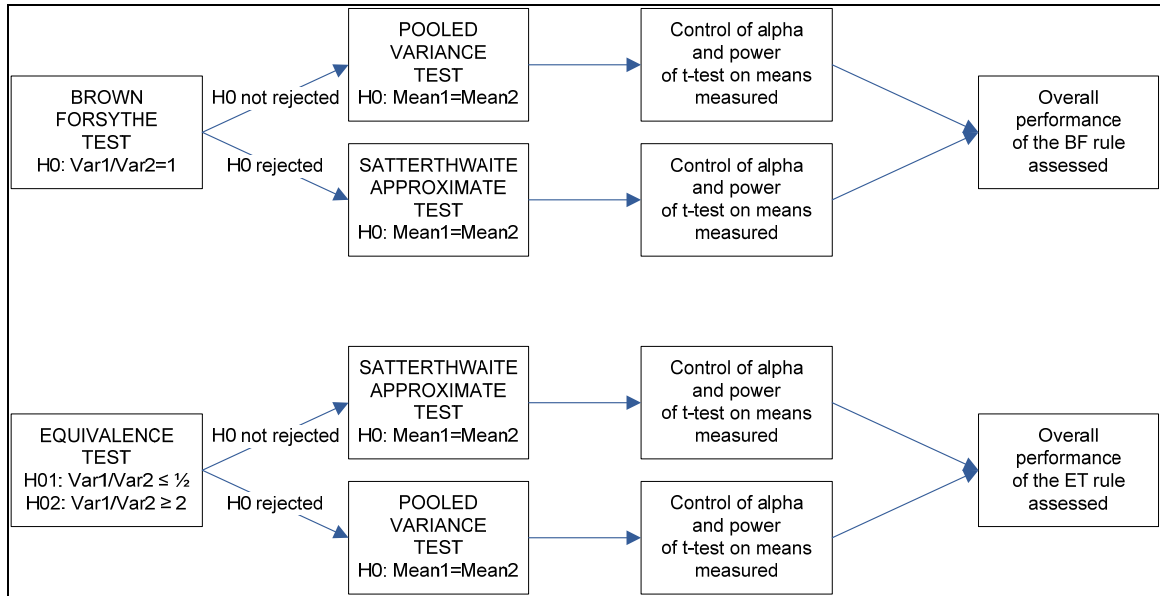


Figure 12. Flow chart for the assessment of the BF and ET rules for switching between the PVT and the SAT as a function of the observed group variances.

The BF and ET switching rules exhibit different performance levels only to the extent that their recommendations to select the PVT or the SAT forms of the t-test are not always in agreement. Figure 13 below shows the amount of agreement and disagreement between the two switching rules to estimate the control of alpha (H0 true), with a total sample size of 50. This crosstabulation includes four cells: the two cells on the downward diagonal represent the instances of agreement, and the two cells on the upward diagonal represent the instances of disagreement. Within each cell, the first number is the frequency count for this combination of ET and PVT recommendation, and the second number is the percent of the total represented by the frequency count. For instance, in the first cell (BF and ET recommend PVT), the frequency count is 215, which represents .07% of the total number of instances (300,000). Figure 13 shows that the two rules agree about 58% of the time (.07+58.42). There is no instance of the ET rule choosing the PVT

when the BF rule chooses the SAT, but the reverse occurs about 42% of the time (41.51). This is a consequence of the fact that with small sample sizes the null hypothesis is less likely to be rejected, so there are many instances, as the variance ratio varies in the low-to-middle range (which goes from 1 to 19), when both rules will retain H0, in which case the BF recommendation will be the PVT and the ET recommendation will be the SAT.

| BF_ET_alpha with 10000 iterations per combination (n=50) H0 true: mean1=50, mean2=50 | | | |
|---|-------------------------------|------------------------------------|------------------|
| BF | ET | | Total |
| | PVT | SAT | |
| Frequency | | | |
| Percent | | | |
| Row Pct | | | |
| Col Pct | | | |
| PVT | 215 0.07 0.17 100.00 | 124515 41.51 99.83 41.53 | 124730 41.58 |
| SAT | 0 0.00 0.00 0.00 | 175270 58.42 100.00 58.47 | 175270 58.42 |
| Total | 215 0.07 | 299785 99.93 | 300000 100.00 |

Figure 13. Agreement table between the BF and ET rules for H0 true and n = 50.

This result can be contrasted with the following crosstabulation (Figure 14), which shows the same measure of agreement as above when the total sample size is set at 200.

| BF_ET_alpha with 10000 iterations per combination (n=200) | | | |
|---|--------|--------|--------|
| H0 true: mean1=50, mean2=50 | | | |
| BF | ET | | |
| Frequency | | | |
| Percent | | | |
| Row Pct | | | |
| Col Pct | PVT | SAT | Total |
| PVT | 45860 | 27896 | 73756 |
| | 15.29 | 9.30 | 24.59 |
| | 62.18 | 37.82 | |
| | 100.00 | 10.98 | |
| SAT | 0 | 226244 | 226244 |
| | 0.00 | 75.41 | 75.41 |
| | 0.00 | 100.00 | |
| | 0.00 | 89.02 | |
| Total | 45860 | 254140 | 300000 |
| | 15.29 | 84.71 | 100.00 |

Figure 14. Agreement table between the BF and ET rules for H0 true and n = 200.

The greater sample size of 200 leads to a reduction in the number of disagreements (down to about 9%) as a consequence of the greater ease with which the BF rule can reject its null hypothesis of equal variances when n is larger.

Simulation 2.1.: Comparison of the BF and ET rules with respect to the control of alpha

Figures 15, 16 and 17 below permit the comparison of the respective performances of the BF and the ET rules in terms of control of alpha for the simulation runs when the total sample size is small (50), medium (100), and large (200).

| HO true: Mean1 = 50, Mean2 = 50 | | | | | | | | | | | |
|---------------------------------|----|----|---------|---------|--------|--------|--------|--------|--------|---------|-------|
| Brown-Forsythe rule | | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio | |
| 50 | 5 | 45 | 9.00 | 0.0606 | 0.1088 | 0.1629 | 0.1435 | 0.1109 | 0.0800 | | |
| 50 | 10 | 40 | 4.00 | 0.0526 | 0.0865 | 0.0995 | 0.0706 | 0.0526 | 0.0530 | | |
| 50 | 15 | 35 | 2.33 | 0.0505 | 0.0700 | 0.0755 | 0.0557 | 0.0490 | 0.0469 | | |
| 50 | 20 | 30 | 1.50 | 0.0479 | 0.0611 | 0.0582 | 0.0477 | 0.0500 | 0.0460 | | |
| 50 | 25 | 25 | 1.00 | 0.0482 | 0.0509 | 0.0526 | 0.0483 | 0.0491 | 0.0535 | | |
| | | | | | | | | | | | |
| Equivalence Testing rule | | | | 20 | 20 | 20 | 20 | 20 | 20 | AvVar | |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio | |
| 50 | 5 | 45 | 9.00 | 0.0623 | 0.0590 | 0.0576 | 0.0521 | 0.0554 | 0.0507 | | |
| 50 | 10 | 40 | 4.00 | 0.0521 | 0.0533 | 0.0507 | 0.0516 | 0.0486 | 0.0519 | | |
| 50 | 15 | 35 | 2.33 | 0.0514 | 0.0530 | 0.0540 | 0.0500 | 0.0488 | 0.0468 | | |
| 50 | 20 | 30 | 1.50 | 0.0479 | 0.0527 | 0.0501 | 0.0465 | 0.0500 | 0.0460 | | |
| 50 | 25 | 25 | 1.00 | 0.0481 | 0.0507 | 0.0525 | 0.0483 | 0.0491 | 0.0535 | | |
| | | | | | | | | | | | |
| BF - ET | | | | 20 | 20 | 20 | 20 | 20 | 20 | AvVar | |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio | |
| 50 | 5 | 45 | 9.00 | -0.0017 | 0.0498 | 0.1053 | 0.0914 | 0.0555 | 0.0293 | | |
| 50 | 10 | 40 | 4.00 | 0.0005 | 0.0332 | 0.0488 | 0.0190 | 0.0040 | 0.0011 | | |
| 50 | 15 | 35 | 2.33 | -0.0009 | 0.0170 | 0.0215 | 0.0057 | 0.0002 | 0.0001 | | |
| 50 | 20 | 30 | 1.50 | 0.0000 | 0.0084 | 0.0081 | 0.0012 | 0.0000 | 0.0000 | | |
| 50 | 25 | 25 | 1.00 | 0.0001 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | | |

Figure 15. Simulation results for actual alpha (BF and ET) when n = 50.

We see that the greatest discrepancies in actual alpha levels occur at the mid-level values of the variance ratio when the n ratio is large. These are instances where the SAT should be chosen, as the PVT does not control alpha well. The ET rule is better suited for the small sample size, in terms of preventing the inflation of alpha, as its null hypothesis corresponds to the SAT. Overall, the alpha inflation is smaller with the ET rule than with the BF rule in 21 out of the 30 combinations of variance ratio and n ratio.

The performance levels of the two rules are more similar when the total sample size is 100, due to the greater agreement. When $n=100$, the ET rule is slightly more conservative, but the BF rule does not exhibit a great deal of alpha inflation: only in one case does the actual alpha exceed .1 (n ratio = 9 and variance ratio = 3).

| HO true: Mean1 = 50, Mean2 = 50 | | | | | | | | | | |
|---------------------------------|----|----|---------|---------|--------|--------|--------|--------|----------|----------|
| Brown-Forsythe rule | | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 AvVar |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 V1 | |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.0528 | 0.0960 | 0.1099 | 0.0615 | 0.0590 | 0.0522 | |
| 100 | 20 | 80 | 4.00 | 0.0496 | 0.0715 | 0.0664 | 0.0566 | 0.0527 | 0.0535 | |
| 100 | 30 | 70 | 2.33 | 0.0489 | 0.0622 | 0.0557 | 0.0487 | 0.0476 | 0.0519 | |
| 100 | 40 | 60 | 1.50 | 0.0488 | 0.0577 | 0.0522 | 0.0487 | 0.0546 | 0.0523 | |
| 100 | 50 | 50 | 1.00 | 0.0494 | 0.0507 | 0.0495 | 0.0470 | 0.0475 | 0.0510 | |
| | | | | | | | | | | |
| Equivalence Testing rule | | | | 20 | 20 | 20 | 20 | 20 | 20 AvVar | |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 V1 | |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.0538 | 0.0517 | 0.0512 | 0.0471 | 0.0555 | 0.0518 | |
| 100 | 20 | 80 | 4.00 | 0.0519 | 0.0470 | 0.0502 | 0.0556 | 0.0527 | 0.0535 | |
| 100 | 30 | 70 | 2.33 | 0.0495 | 0.0503 | 0.0498 | 0.0487 | 0.0476 | 0.0519 | |
| 100 | 40 | 60 | 1.50 | 0.0477 | 0.0515 | 0.0509 | 0.0487 | 0.0546 | 0.0523 | |
| 100 | 50 | 50 | 1.00 | 0.0494 | 0.0507 | 0.0495 | 0.0470 | 0.0475 | 0.0510 | |
| | | | | | | | | | | |
| BF - ET | | | | 20 | 20 | 20 | 20 | 20 | 20 AvVar | |
| | | | | 20 | 25 | 30 | 35 | 37 | 38 V1 | |
| | | | | 20 | 15 | 10 | 5 | 3 | 2 V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | -0.0010 | 0.0443 | 0.0587 | 0.0144 | 0.0035 | 0.0004 | |
| 100 | 20 | 80 | 4.00 | -0.0023 | 0.0245 | 0.0162 | 0.0010 | 0.0000 | 0.0000 | |
| 100 | 30 | 70 | 2.33 | -0.0006 | 0.0119 | 0.0059 | 0.0000 | 0.0000 | 0.0000 | |
| 100 | 40 | 60 | 1.50 | 0.0011 | 0.0062 | 0.0013 | 0.0000 | 0.0000 | 0.0000 | |
| 100 | 50 | 50 | 1.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

Figure 16. Simulation results for actual alpha (BF and ET) when $n = 100$.

With the larger sample size ($n = 200$), the Brown-Forsythe rule leads to the rejection of the null hypothesis of equal variances, and therefore to the use of the SAT, about as often as the Equivalence Testing rule, when H_0 is true ($\text{mean1} = \text{mean2}$). As a result, the BF rule leads to very little difference from the ET rule, and there is practically no alpha inflation issue (we see that no actual alpha value exceeds .1).

| HO true: Mean1 = 50, Mean2 = 50 | | | | | | | | | | |
|---------------------------------|-----|-----|---------|--------|--------|--------|--------|--------|--------|---------|
| Brown-Forsythe rule | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 200 | 20 | 180 | 9.00 | 0.0541 | 0.0847 | 0.0665 | 0.0515 | 0.0516 | 0.0501 | |
| 200 | 40 | 160 | 4.00 | 0.0507 | 0.0711 | 0.0479 | 0.0499 | 0.0516 | 0.0481 | |
| 200 | 60 | 140 | 2.33 | 0.0495 | 0.0561 | 0.0548 | 0.0518 | 0.0516 | 0.0518 | |
| 200 | 80 | 120 | 1.50 | 0.0502 | 0.0578 | 0.0495 | 0.0509 | 0.0489 | 0.0474 | |
| 200 | 100 | 100 | 1.00 | 0.0493 | 0.0496 | 0.0485 | 0.0486 | 0.0457 | 0.0480 | |
| | | | | | | | | | | |
| Equivalence Testing rule | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 200 | 20 | 180 | 9.00 | 0.0532 | 0.0530 | 0.0503 | 0.0509 | 0.0516 | 0.0501 | |
| 200 | 40 | 160 | 4.00 | 0.0506 | 0.0548 | 0.0465 | 0.0499 | 0.0516 | 0.0481 | |
| 200 | 60 | 140 | 2.33 | 0.0494 | 0.0508 | 0.0547 | 0.0518 | 0.0516 | 0.0518 | |
| 200 | 80 | 120 | 1.50 | 0.0502 | 0.0546 | 0.0495 | 0.0509 | 0.0489 | 0.0474 | |
| 200 | 100 | 100 | 1.00 | 0.0493 | 0.0496 | 0.0485 | 0.0486 | 0.0457 | 0.0480 | |
| | | | | | | | | | | |
| BF - ET | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 200 | 20 | 180 | 9.00 | 0.0009 | 0.0317 | 0.0162 | 0.0006 | 0.0000 | 0.0000 | |
| 200 | 40 | 160 | 4.00 | 0.0001 | 0.0163 | 0.0014 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 60 | 140 | 2.33 | 0.0001 | 0.0053 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 80 | 120 | 1.50 | 0.0000 | 0.0032 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 100 | 100 | 1.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

Figure 17. Simulation results for actual alpha (BF and ET) when $n = 200$.

Simulation 1.1.: Comparison of the PVT and the SAT with respect to power

With respect to power, the agreement tables for total sample sizes of 50 and 200 are as follows.

| BF_ET_power with 10000 iterations per combination (n=50) H0 false: mean1=50, mean2=53 | | | |
|--|--------|--------|--------|
| BF | ET | | |
| Frequency | | | |
| Percent | | | |
| Row Pct | | | |
| Col Pct | PVT | SAT | Total |
| PVT | 211 | 124262 | 124473 |
| | 0.07 | 41.42 | 41.49 |
| | 0.17 | 99.83 | |
| | 100.00 | 41.45 | |
| SAT | 0 | 175527 | 175527 |
| | 0.00 | 58.51 | 58.51 |
| | 0.00 | 100.00 | |
| | 0.00 | 58.55 | |
| Total | 211 | 299789 | 300000 |
| | 0.07 | 99.93 | 100.00 |

Figure 18. Agreement table between the BF and ET rules for H0 false and n = 50.

BF_ET_power with 10000 iterations per combination (n=200)
H0 false: mean1=50, mean2=53

| BF | ET | | Total |
|-----------|-----------------------------------|------------------------------------|------------------|
| | PVT | SAT | |
| Frequency | | | |
| Percent | | | |
| Row Pct | | | |
| Col Pct | | | |
| PVT | 45800 15.27 62.16 100.00 | 27875 9.29 37.84 10.97 | 73675 24.56 |
| SAT | 0 0.00 0.00 0.00 | 226325 75.44 100.00 89.03 | 226325 75.44 |
| Total | 45800 15.27 | 254200 84.73 | 300000 100.00 |

Figure 19. Agreement table between the BF and ET rules for H0 false and n = 200.

As with the corresponding tables for actual alpha, we see that the greatest number of disagreements between the two rules occurs with the low sample size (n=50) where the ET rule is much more likely to recommend the SAT (99.93%) than the BF rule (58.51%). Figures 20, 21 and 22 below show the performances of the two rules in terms of power for total sample sizes of 50, 100 and 200, respectively.

| HO false: Mean1 = 50, Mean2 = 53 | | | | | | | | | | |
|----------------------------------|----|----|---------|--------|--------|--------|--------|--------|--------|---------|
| Brown-Forsythe rule | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 50 | 5 | 45 | 9.00 | 0.2957 | 0.3579 | 0.3695 | 0.2634 | 0.1891 | 0.1608 | |
| 50 | 10 | 40 | 4.00 | 0.4655 | 0.4796 | 0.4267 | 0.3141 | 0.2797 | 0.2842 | |
| 50 | 15 | 35 | 2.33 | 0.5705 | 0.5712 | 0.5139 | 0.4417 | 0.4138 | 0.4111 | |
| 50 | 20 | 30 | 1.50 | 0.6199 | 0.6314 | 0.5924 | 0.5461 | 0.5392 | 0.5334 | |
| 50 | 25 | 25 | 1.00 | 0.6357 | 0.6508 | 0.6360 | 0.6364 | 0.6245 | 0.6307 | |
| Equivalence Testing rule | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 50 | 5 | 45 | 9.00 | 0.2285 | 0.1843 | 0.1702 | 0.1488 | 0.1359 | 0.1363 | |
| 50 | 10 | 40 | 4.00 | 0.4339 | 0.3677 | 0.3323 | 0.2923 | 0.2760 | 0.2829 | |
| 50 | 15 | 35 | 2.33 | 0.5554 | 0.5034 | 0.4647 | 0.4378 | 0.4128 | 0.4111 | |
| 50 | 20 | 30 | 1.50 | 0.6156 | 0.6023 | 0.5718 | 0.5445 | 0.5391 | 0.5333 | |
| 50 | 25 | 25 | 1.00 | 0.6347 | 0.6495 | 0.6353 | 0.6362 | 0.6245 | 0.6307 | |
| BF - ET | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | V1 | |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | V2 | |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 50 | 5 | 45 | 9.00 | 0.0672 | 0.1736 | 0.1993 | 0.1146 | 0.0532 | 0.0245 | |
| 50 | 10 | 40 | 4.00 | 0.0316 | 0.1119 | 0.0944 | 0.0218 | 0.0037 | 0.0013 | |
| 50 | 15 | 35 | 2.33 | 0.0151 | 0.0678 | 0.0492 | 0.0039 | 0.0010 | 0.0000 | |
| 50 | 20 | 30 | 1.50 | 0.0043 | 0.0291 | 0.0206 | 0.0016 | 0.0001 | 0.0001 | |
| 50 | 25 | 25 | 1.00 | 0.0010 | 0.0013 | 0.0007 | 0.0002 | 0.0000 | 0.0000 | |

Figure 20. Simulation results for power (BF and ET) when n = 50.

| HO false: Mean1 = 50, Mean2 = 53 | | | | | | | | | | |
|----------------------------------|----|----|---------|--------|--------|--------|--------|--------|--------|---------|
| Brown-Forsythe rule | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | 38 | V1 |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.5195 | 0.5218 | 0.4264 | 0.3090 | 0.2877 | 0.2769 | |
| 100 | 20 | 80 | 4.00 | 0.7522 | 0.7359 | 0.6336 | 0.5722 | 0.5485 | 0.5371 | |
| 100 | 30 | 70 | 2.33 | 0.8600 | 0.8422 | 0.7837 | 0.7451 | 0.7362 | 0.7190 | |
| 100 | 40 | 60 | 1.50 | 0.9027 | 0.8972 | 0.8716 | 0.8492 | 0.8425 | 0.8360 | |
| 100 | 50 | 50 | 1.00 | 0.9163 | 0.9080 | 0.9115 | 0.9111 | 0.9096 | 0.9059 | |
| Equivalence Testing rule | | | | | | | | | | |
| | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | 38 | V1 |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.4566 | 0.3816 | 0.3312 | 0.2964 | 0.2865 | 0.2768 | 0 |
| 100 | 20 | 80 | 4.00 | 0.7345 | 0.6773 | 0.6106 | 0.5720 | 0.5485 | 0.5371 | 0 |
| 100 | 30 | 70 | 2.33 | 0.8537 | 0.8182 | 0.7787 | 0.7451 | 0.7362 | 0.7190 | 0 |
| 100 | 40 | 60 | 1.50 | 0.9014 | 0.8882 | 0.8704 | 0.8492 | 0.8425 | 0.8360 | 0 |
| 100 | 50 | 50 | 1.00 | 0.9162 | 0.9080 | 0.9115 | 0.9111 | 0.9096 | 0.9059 | 0 |
| BF - ET | | | | | | | | | | |
| | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | 38 | V1 |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 100 | 10 | 90 | 9.00 | 0.0629 | 0.1402 | 0.0952 | 0.0126 | 0.0012 | 0.0001 | |
| 100 | 20 | 80 | 4.00 | 0.0177 | 0.0586 | 0.0230 | 0.0002 | 0.0000 | 0.0000 | |
| 100 | 30 | 70 | 2.33 | 0.0063 | 0.0240 | 0.0050 | 0.0000 | 0.0000 | 0.0000 | |
| 100 | 40 | 60 | 1.50 | 0.0013 | 0.0090 | 0.0012 | 0.0000 | 0.0000 | 0.0000 | |
| 100 | 50 | 50 | 1.00 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

Figure 21. Simulation results for power (BF and ET) when n = 100.

| HO false: Mean1 = 50, Mean2 = 53 | | | | | | | | | | |
|----------------------------------|-----|-----|---------|---------|--------|--------|--------|--------|--------|---------|
| Brown-Forsythe rule | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | 38 | V1 |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 200 | 20 | 180 | 9.00 | 0.8068 | 0.7714 | 0.6453 | 0.5707 | 0.5510 | 0.5317 | |
| 200 | 40 | 160 | 4.00 | 0.9625 | 0.9459 | 0.9024 | 0.8677 | 0.8519 | 0.8457 | |
| 200 | 60 | 140 | 2.33 | 0.9904 | 0.9868 | 0.9756 | 0.9656 | 0.9609 | 0.9583 | |
| 200 | 80 | 120 | 1.50 | 0.9954 | 0.9938 | 0.9934 | 0.9893 | 0.9883 | 0.9886 | |
| 200 | 100 | 100 | 1.00 | 0.9975 | 0.9973 | 0.9985 | 0.9967 | 0.9965 | 0.9968 | |
| Equivalence Testing rule | | | | | | | | | | |
| | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | 38 | V1 |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 200 | 20 | 180 | 9.00 | 0.7819 | 0.7025 | 0.6283 | 0.5707 | 0.5510 | 0.5317 | |
| 200 | 40 | 160 | 4.00 | 0.9615 | 0.9370 | 0.9020 | 0.8677 | 0.8519 | 0.8457 | |
| 200 | 60 | 140 | 2.33 | 0.9905 | 0.9852 | 0.9756 | 0.9656 | 0.9609 | 0.9583 | |
| 200 | 80 | 120 | 1.50 | 0.9952 | 0.9936 | 0.9934 | 0.9893 | 0.9883 | 0.9886 | |
| 200 | 100 | 100 | 1.00 | 0.9975 | 0.9973 | 0.9985 | 0.9967 | 0.9965 | 0.9968 | |
| BF - ET | | | | | | | | | | |
| | | | 20 | 20 | 20 | 20 | 20 | 20 | 20 | AvVar |
| | | | 20 | 25 | 30 | 35 | 37 | 38 | 38 | V1 |
| | | | 20 | 15 | 10 | 5 | 3 | 2 | 2 | V2 |
| N | N1 | N2 | N ratio | 1.00 | 1.67 | 3.00 | 7.00 | 12.33 | 19.00 | V ratio |
| 200 | 20 | 180 | 9.00 | 0.0249 | 0.0689 | 0.0170 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 40 | 160 | 4.00 | 0.0010 | 0.0089 | 0.0004 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 60 | 140 | 2.33 | -0.0001 | 0.0016 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 80 | 120 | 1.50 | 0.0002 | 0.0002 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| 200 | 100 | 100 | 1.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

Figure 22. Simulation results for power (BF and ET) when total n = 200.

While the ET rule outperforms the BF rule in terms of control of alpha, particularly when n is small, the opposite is true for power: The BF rule results in slightly more power at the mid-levels of the variance ratio when the n ratio is large. This BF-rule power advantage is clear with n=50, becomes modest when n=100, and all but disappears when n=200.

Therefore, applying the difference-testing or the equivalence-testing form of the Brown-Forsythe test of homogeneity of variance results in different choices with respect to variance pooling, particularly when the sample size is small. Applying the ET rule leads the researcher to avoid using the pooled-variance t-test when there is insufficient evidence that the variances are indeed equal. The gain is a better control of alpha, at the

cost of some loss of power. It is important to note that the amount of trade-off can be fine-tuned, if the researcher so desires, by adjusting the width of the equivalence interval in the desired direction: A wider equivalence interval will make it easier to reject the H0 of non-equivalence, and therefore a higher probability of using the PVT, which will tend to increase power but also type-I error.

The SAS code for the BF-ET actual alpha simulation with total sample size of 50 can be found in Appendix C.

Example of a substantive application: Comparison of students' scores between the in-class and online forms of the UMN course Public Health 6414:

Biostatistical Methods I

The data set for this application of equivalence testing consists of the overall scores obtained by the students enrolled in PubH6414 at the University of Minnesota in the Spring of 2007 (in class) and in the Fall of 2007 (online). This data set can be found in Appendix D.

The research hypothesis is that the group difference between the two forms of the course is within a certain range of percentage points. This range is arbitrarily set at 5: The two forms are considered equivalent if the absolute value of this difference is no greater than 2.5.

Two equivalence tests are conducted: One test to evaluate the homoskedasticity assumption (with the same equivalence limits of $\frac{1}{2}$ and 2 as above), and one test to assess whether the two forms are indeed equivalent.

For the first equivalence test (variances), the hypotheses are:

$$H_0: \text{var1/var2} \leq \frac{1}{2} \text{ or } \text{var1/var2} \geq 2$$

$$H_A: \frac{1}{2} < \text{var1/var2} < 2$$

These hypotheses can also be expressed as follows:

$$H_{01}: \text{var1/var2} \leq \frac{1}{2}$$

$$H_{02}: \text{var1/var2} \geq 2$$

$$H_{A1}: \text{var1/var2} > \frac{1}{2}$$

$$H_{A2}: \text{var1/var2} < 2$$

The computations and results for the first equivalence test are shown in the Mathcad worksheet below (Figures 23 and 24):

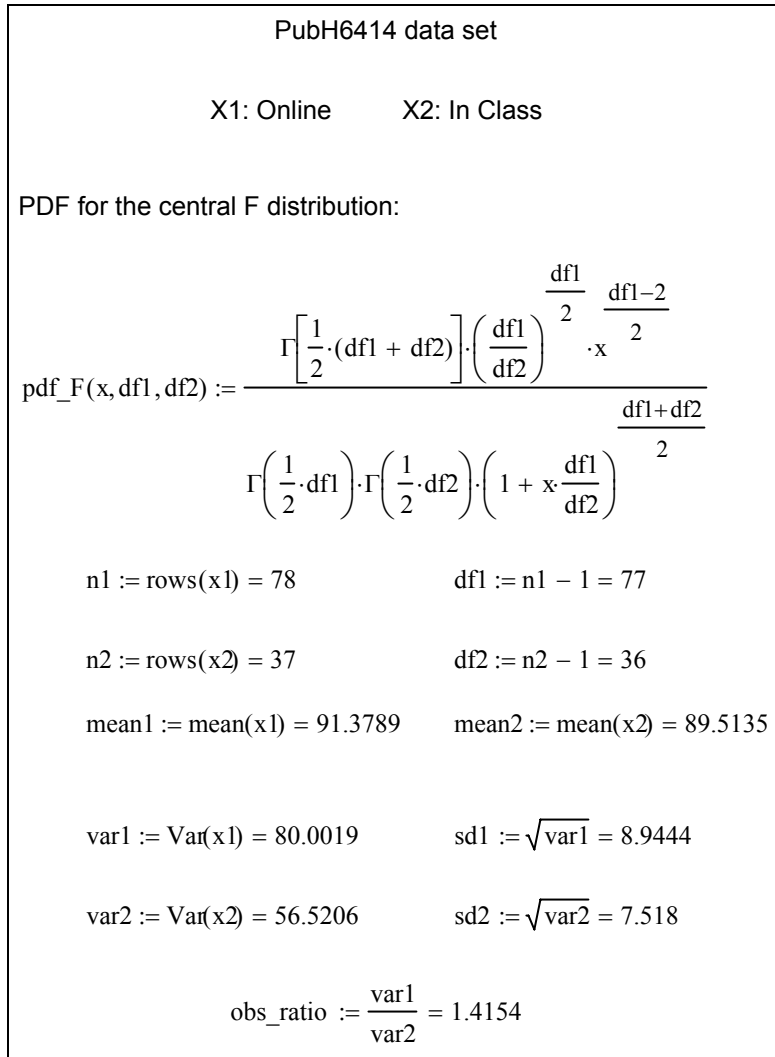


Figure 23. Probability density function for the central F distribution, and relevant characteristics of the data set.

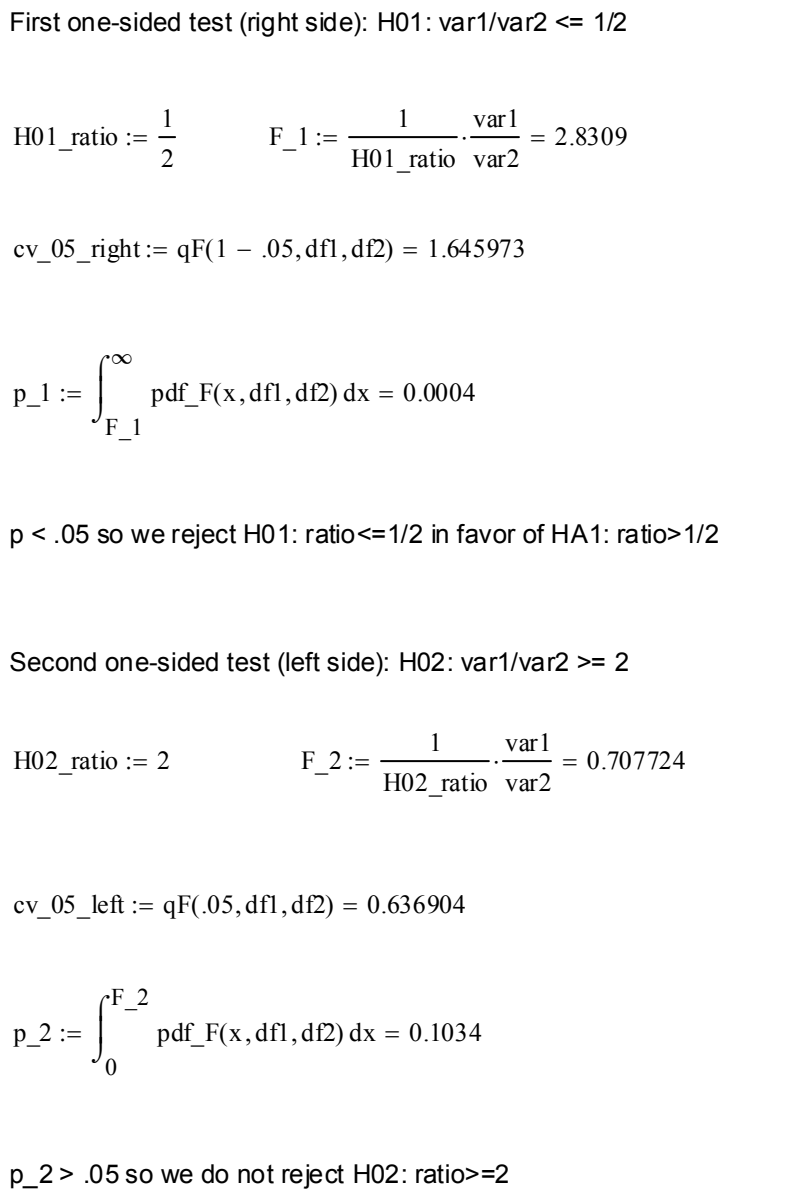


Figure 24. Test of homoskedasticity.

Since H_{01} and H_{02} are not both rejected, we retain the overall null hypothesis of heteroskedasticity, so the SAT is chosen for the test on the score means. Appendix E provides a visual representation of this equivalence test.

For the second equivalence test (means), the hypotheses are:

$H_0: \text{Mean1}-\text{Mean2} \leq -2.5 \text{ or } \text{Mean1}-\text{Mean2} \geq 2.5$

$H_A: -2.5 < \text{Mean1}-\text{Mean2} < 2.5$

These hypotheses can also be expressed as follows:

$H_{01}: \text{Mean1}-\text{Mean2} \leq -2.5$

$H_{02}: \text{Mean1}-\text{Mean2} \geq 2.5$

$H_{A1}: \text{Mean1}-\text{Mean2} > -2.5$

$H_{A2}: \text{Mean1}-\text{Mean2} < 2.5$

The computations and results for the second equivalence test are shown in Figure 25 below.

Probability density function for the central t distribution:

$$\text{pdf}_t(x, \text{df}) := \frac{\Gamma\left(\frac{\text{df} + 1}{2}\right)}{\sqrt{\pi \cdot (\text{df})} \cdot \Gamma\left(\frac{\text{df}}{2}\right)} \cdot \left(1 + \frac{x^2}{\text{df}}\right)^{-\frac{\text{df} + 1}{2}}$$

Degrees of freedom for Satterthwaite Approximate Test:

$$\text{df}_s := \frac{\left(\frac{\text{var1}}{n1} + \frac{\text{var2}}{n2}\right)^2}{\frac{\text{var1}^2}{n1^2 \cdot (n1 - 1)} + \frac{\text{var2}^2}{n2^2 \cdot (n2 - 1)}} = 83.065$$

Equivalence limits: $\theta_L := -2.5$ $\theta_U := 2.5$

$\text{mean}(x1) = 91.3789$ $\text{mean}(x2) = 89.5135$ $d := \text{mean}(x1) - \text{mean}(x2) = 1.8653$

First 1-sided test (right-side w/r to lower equivalence limit):

H01: Mean1 - Mean2 \leq Theta_L

HA1: Mean1 - Mean2 $>$ Theta_L

$$t_{s_right} := \frac{d - \theta_L}{\sqrt{\frac{\text{var1}}{n1} + \frac{\text{var2}}{n2}}} = 2.7319 \quad p_{t_s_right} := \int_{t_{s_right}}^{\infty} \text{pdf}_t(x, \text{df}_s) dx = 0.0038$$

For the right-side test, $p < .05$ so H01 is rejected: the observed difference of 1.87 is significantly greater than the lower equivalence limit of -2.5

Second 1-sided test (left-side w/r to upper equivalence limit):

H01: Mean1 - Mean2 \geq Theta_U

HA1: Mean1 - Mean2 $<$ Theta_U

$$t_{s_left} := \frac{d - \theta_U}{\sqrt{\frac{\text{var1}}{n1} + \frac{\text{var2}}{n2}}} = -0.3972 \quad p_{t_s_left} := \int_{-\infty}^{t_{s_left}} \text{pdf}_t(x, \text{df}_s) dx = 0.3461$$

For the left side test, $p > .05$ so H02 is not rejected: the observed difference of 1.87 is not significantly smaller than the upper equivalence limit of 2.5

The overall equivalence p-value is the largest of the two 1-sided p-values:

$$p_{t_s_eq} := \max(p_{t_s_right}, p_{t_s_left}) = 0.3461$$

Figure 25. Test of means.

Since H01 and H02 are not both rejected, we do not conclude in favor of equivalence. We have not found sufficient empirical evidence to reject the hypothesis that in the population represented by this sample, the average online score is higher than the average in-class score.

The equivalence t-test calculations above can be confirmed with SAS 9.2, the first version of SAS to include some capability to conduct equivalence tests within a SAS procedure.

```

The SAS System
The TTEST Procedure
Variable: Score

```

| Group | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|-------------|----|---------|---------|---------|---------|---------|
| 1: Online | 78 | 91.3788 | 8.9444 | 1.0128 | 39.6125 | 100.0 |
| 2: In Class | 37 | 89.5135 | 7.5180 | 1.2360 | 71.7500 | 100.0 |
| Diff (1-2) | | 1.8653 | 8.5159 | 1.6999 | | |

```

TOST Level 0.05 Equivalence Analysis

```

| Method | Variances | Test | Null | DF | t Value | P-Value |
|----------------------|----------------|----------------|------|--------|---------|---------------|
| Pooled | Equal | Upper | -2.5 | 113 | 2.57 | 0.0058 |
| Pooled | Equal | Lower | 2.5 | 113 | -0.37 | 0.3548 |
| Pooled | Equal | Overall | | | | 0.3548 |
| Satterthwaite | Unequal | Upper | -2.5 | 83.065 | 2.73 | 0.0038 |
| Satterthwaite | Unequal | Lower | 2.5 | 83.065 | -0.40 | 0.3461 |
| Satterthwaite | Unequal | Overall | | | | 0.3461 |

Figure 26. SAS output for the equivalence test of group means (PubH5414 data).

It is interesting to compare these equivalence-testing results with the results of the difference-testing approach, where the null hypothesis is that of equality between the group means.

| The SAS System | | | | | | | | | |
|---------------------|----------------|----|------------------|------------------|------------------|------------------|---------------------|---------------------|---------|
| The TTEST Procedure | | | | | | | | | |
| Statistics | | | | | | | | | |
| Variable | Group | N | Lower CL Mean | Upper CL Mean | Lower CL Mean | Upper CL Mean | Lower CL Std Dev | Upper CL Std Dev | Std Err |
| Score | 1: Online | 78 | 89.362 | 91.379 | 93.395 | 7.7276 | 8.9444 | 10.62 | 1.0128 |
| Score | 2: In Class | 37 | 87.007 | 89.514 | 92.02 | 6.1137 | 7.518 | 9.7656 | 1.236 |
| Score | Diff (1-2) | | -1.503 | 1.8653 | 5.2332 | 7.5357 | 8.5159 | 9.7917 | 1.6999 |

| T-Tests | | | | | |
|--------------|---------------|--------------|------------|-------------|---------------|
| Variable | Method | Variances | DF | t Value | Pr > t |
| Score | Pooled | Equal | 113 | 1.10 | 0.2748 |
| Score | Satterthwaite | Unequal | 83.1 | 1.17 | 0.2464 |

| Equality of Variances | | | | | |
|-----------------------|----------|--------|--------|---------|--------|
| Variable | Method | Num DF | Den DF | F Value | Pr > F |
| Score | Folded F | 77 | 36 | 1.42 | 0.2505 |

Figure 27. SAS output for the conventional (i.e., difference) test of group means (PubH5414 data).

With a p-value of .2748, we do not reject this null hypothesis of no mean difference, so the conclusion is inconsistent with that of the equivalence test. The sample size (37 in the in-class group) is small relative to the mean difference, and neither the difference test nor the equivalence test leads to the rejection of the null hypothesis, so the conclusions are opposite. If the purpose is to demonstrate that the two class formats are equivalent in terms of average grade, then one would be safer in concluding that this

small data set provides insufficient evidence. These results represent an example of cell [3] in Figure 1 above (p. 12): The group means are neither significantly different nor significantly equivalent.

CHAPTER 5

DISCUSSION

Summary of Findings

This thesis introduces equivalence testing as an alternative to traditional difference testing. Equivalence testing is appropriate when the research hypothesis is one of absence of effect (e.g., similarity between two group means). Using equivalence testing, the investigator sets up the null hypothesis to represent an effect, and produces data which are hoped to lead to the rejection of this null hypothesis.

The historical context of the development of equivalence testing is provided. As new drugs are being tested prior to FDA approval, the requirement often is to show equivalence – not necessarily superiority – to existing drugs, particularly when a new generic drug is compared to a branded drug. Thus, the methodology of equivalence testing has risen from a need to conduct hypothesis testing with the intent of showing not a difference but an equivalence between several conditions.

The primary research question of this thesis is whether equivalence testing could be useful when evaluating the assumption of homoskedasticity prior to conducting a Student t-test. The reason this assumption is evaluated is to decide whether the original form of the t-test can be used (pooled-variance estimate if the assumption is met) or another form should be used (separate-variance estimate, such as Satterthwaite's, if it is not met). Since the main risk is alpha inflation (actual alpha exceeding nominal alpha), it is arguably preferable to define the null hypothesis not as homoskedasticity, as it is done with traditional difference tests, but as heteroskedasticity, as it is done with equivalence testing. The superior rule to guide the decision to switch between the two forms of the t-

test is the rule that appropriately recommends either form, in a way that results in less alpha inflation without sacrificing too much power.

In order to set up an equivalence test to compare two group variances, equivalence limits need to be set, within which lies the alternative hypothesis of homoskedasticity. These limits should be set so that they reflect the risks involved in an erroneous conclusion with respect to the homoskedasticity assumption. The first set of simulations serves to help decide where to locate the equivalence limits. The results indicate that when the ratio of variances exceeds 2:1, there is a much greater risk of alpha inflation, especially when the sample sizes differ substantially (larger n ratio).

This finding is used to set the equivalence limits needed in the second set of simulations, where the two switching rules are compared. With the first rule (BF), a difference test is used, leading to the choice of the pooled-variance t-test when the null hypothesis of homoskedasticity is not rejected, and of the separate-variance t-test (Satterthwaite's test) when it is rejected. With the second rule (ET), an equivalence test is used, leading to the choice of Satterthwaite's test when the null hypothesis of heteroskedasticity is not rejected, and of the pooled-variance t-test when it is rejected.

The results show that using the equivalence test rule leads to less alpha inflation, particularly when the sample size is small ($n=50$), at the cost of a small loss of power in some cases. The superior control of alpha afforded by the ET switching rule makes the use of equivalence testing in this context an attractive option, as effectively controlling alpha is one of the main goals of statistical inference in Fisher's tradition.

The secondary research question of this thesis, presented as an illustrative example of a substantive application of equivalence testing in Educational Psychology, is

whether an in-class version and an online version of the same course can be considered equivalent, with respect to the average scores obtained by the students. With the limitations outlined on page 6 above, the results show that the two versions cannot be considered equivalent, based on the data available. If a difference test is used instead, the null hypothesis of zero difference cannot be rejected, which makes this analysis an example of how difference testing is inappropriate as a way to assess equivalence: the failure to reject the difference-testing null hypothesis is a result of low power, not of close similarity between the group means, at least not when similarity is defined in terms of the equivalence limits chosen here.

Contribution

The principal contribution of this thesis is to introduce equivalence testing to educational researchers, as an appropriate way to investigate research hypotheses involving the equivalence between several conditions. In Educational Psychology, and indeed in all social sciences, these research hypotheses have so far been investigated with difference testing. As noted above, the main problem with using difference testing for the purpose of demonstrating equivalence is that the desired conclusion can be reached as a result of low power. This can be avoided by using equivalence testing for these admittedly less common investigations when the researcher's claim is one of equivalence. The issue of substantive (a.k.a. practical) significance is often raised in the context of difference testing, after the statistical test is conducted and statistical significance is reached. A careful researcher knows not to confuse the significance of the effect with its magnitude, and addresses the issue of the practical significance of the statistically significant effect, in addition to having conducted a difference test of significance. One of the benefits of

using equivalence testing is that this consideration of practical significance is built in the test itself. Setting the equivalence limits, a necessary part of an equivalence test, constitutes a *de facto* operational definition of what is considered practical significance: A small effect can be significantly different from a null effect, yet not exceed a defined range of equivalence (see cell [2] in Figure 1 above).

Figure 28 below can help an investigator decide whether equivalence testing is appropriate, and if so, how to set up the null (H0) and alternative (HA) hypotheses. This decision process applies to a nondirectional (a.k.a. two-tail) statistical test.

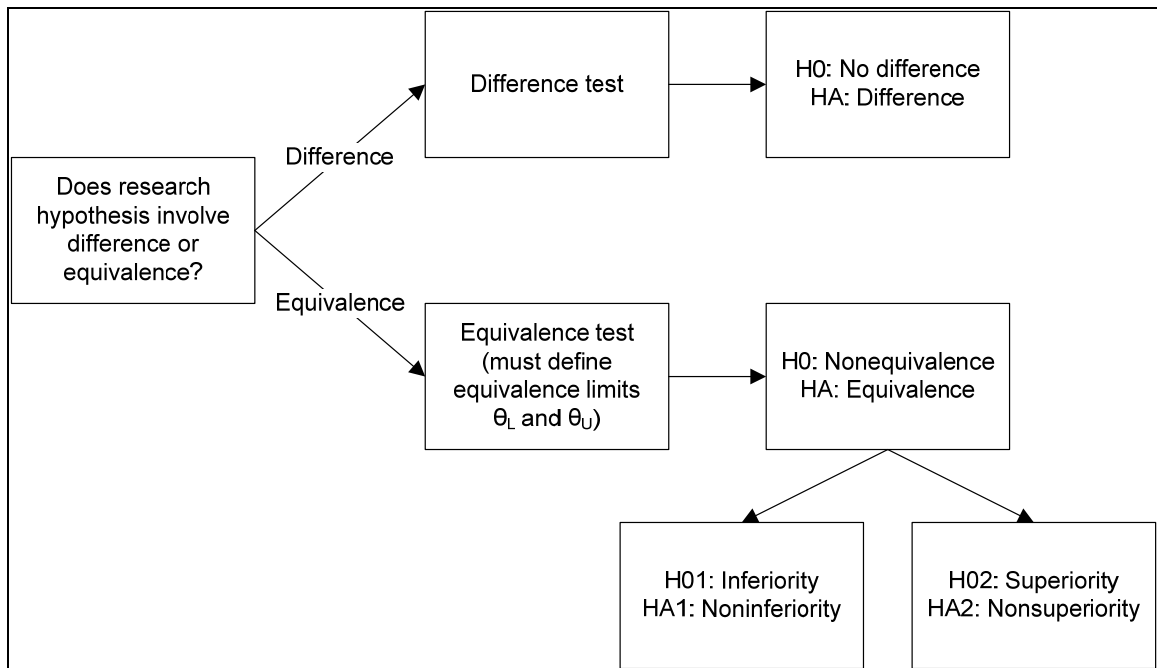


Figure 28. Decision chart for Equivalence Testing vs. Difference Testing.

Beyond introducing equivalence testing and emphasizing its benefits, the specific contributions of this thesis are to show that equivalence testing can be profitably used for testing the assumption of homoskedasticity, and that it is an appropriate way of

investigating applied research hypotheses of equivalence such as those relating to different versions of a course.

Future research

Other aspects of Educational Psychology can possibly benefit from the implementation of equivalence testing.

Potential future applications could include extending equivalence testing to statistical tests other than two group comparisons. Currently, the documented usage of equivalence testing in the statistical and medical literature is limited to the comparison of two groups on a continuous dependent variable, without the use of covariates. The capability of statistical software to conduct equivalence tests is nascent and currently reflects these limitations. SAS 9.2, EquivTest 2.0, and NCSS 2007 offer some equivalence testing, but only for the comparison of two means, two proportions, or two odd ratios. Naturally, statistical software including a programming language, such as SAS, allow for writing the necessary code for other tests, which is how equivalence tests comparing two variances were conducted in this thesis, but the inclusion of these other tests as documented procedures is still not in existence within the currently available statistical software. Therefore, extending equivalence testing to other forms of statistical testing, both in terms of theoretical development and software implementation, would constitute an attractive line of future research. These methods could include the analysis of variance, the analysis of covariance, and simple and multiple regression. As these methods allow for the inclusion of covariates, this would be a particularly useful extension for the field of Education Psychology, given that much of its empirical research occurs in an observational context, where experimental control is not possible, and where

the statistical control provided by the inclusion of covariates in the statistical model is crucial.

Whether the application is methodological or substantive, equivalence testing is helpful in testing statistical hypotheses where the burden is on the investigator to seek evidence in favor of equality or similarity. This allows the application of one method at a given confidence level to data sets of any size, with the understanding that the smaller the sample sizes, the less likely the alternative hypothesis of equivalence will be supported, which is appropriate for instances where this alternative statistical hypothesis corresponds to a research hypothesis of similarity.

REFERENCES

- Allen, I. E., & Christopher, A. S. (2006). Different, equivalent, both. *Quality Progress*, 39(7), 77.
- Altman, D. G., & Bland, J.M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, 311, 485.
- Aras, G. (2001). Superiority, noninferiority, equivalence and bioequivalence - revisited. *Drug Information Journal*, 35, 1157-1164.
- Barker, L. E., Luman, E. T., McCauley, M. M., & Chu, S. Y. (2002). Assessing equivalence: an alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, 156(11), 1056-1061(1056).
- Blair, R. C., & Taylor, R. A. (2007). *Biostatistics for the health sciences*. Upper Saddle River, NJ: Pearson Prentice Hall.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychology Bulletin*, 57, 49-64.
- Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40, 318-335.
- Bratley, P., Fox, B. & Schrage, L. (1987). *A Guide to Simulation*, 2nd ed. New York: Springer-Verlag.
- Brown, M. B., & Forsythe, Alan B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364-367.
- Chow, S. C. (1999). Individual bioequivalence: a review of the FDA guidance. *Drug Information Journal*, 33, 435-444.

- Chow, S. L. (1996). *Statistical Significance: Rationale, Validity and Utility*. London: Sage.
- Chuang-Stein, C. (1999). Clinical equivalence-a clarification. *Drug Information Journal*, 33, 1189-1194.
- Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variance, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4), 351-361.
- Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, 60(1), 1-10.
- Denis, D. J. (2003). Alternatives to null hypothesis significance testing. *Theory & Science [online]*, 4(1).
- Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics*, 33, 593-602.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Garrett, K. A. (1997). Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. *Phytopathology*, 87, 372-374.
- Glass, G. V. (1966). Testing homogeneity of variances. *American Educational Research Journal*, 3(3), 187-190.
- Gigerenzer, G., & al. (1989). *The empire of chance*. Cambridge University Press.
- Hall, I. J. (1972). Some comparisons of tests for equality of variances. *Journal of Statistical Computation and Simulation*, 1(2), 133-194.

- Hartley, H. O. (1950). The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, 37(3/4), 308-312.
- Hatch, J. P. (1996). Using statistical equivalence testing in clinical biofeedback research. *Biofeedback and Self-Regulation*, 21, 105-119.
- Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, 12(1), 83-91.
- Hauck, W. W., & Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine*, 5(3), 203-209.
- Hauck, W. W., & S., A. (1999). Some issues in the design and analysis of equivalence trials. *Drug Information Journal*, 33, 109-118.
- Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.
- Hwang, I. K., & Morikawa, T. (1999). Design issues in noninferiority/equivalence trials. *Drug Information Journal*, 33, 1205-1218.
- Jackson, A. J. (1994). *Generics and bioequivalence*. Boca Raton: CRC.
- Koyama, T., & Westfall, P. H. (2005). Decision-theoretic views on simultaneous testing of superiority and noninferiority. *Journal of Biopharmaceutical Statistics*, 15, 943-955.
- Le Henaff, A., Giraudeau, B., Baron, G., & Ravaud, P. (2006). Quality of reporting of noninferiority and equivalence randomized trials. *Journal of the American Medical Association*, 295.

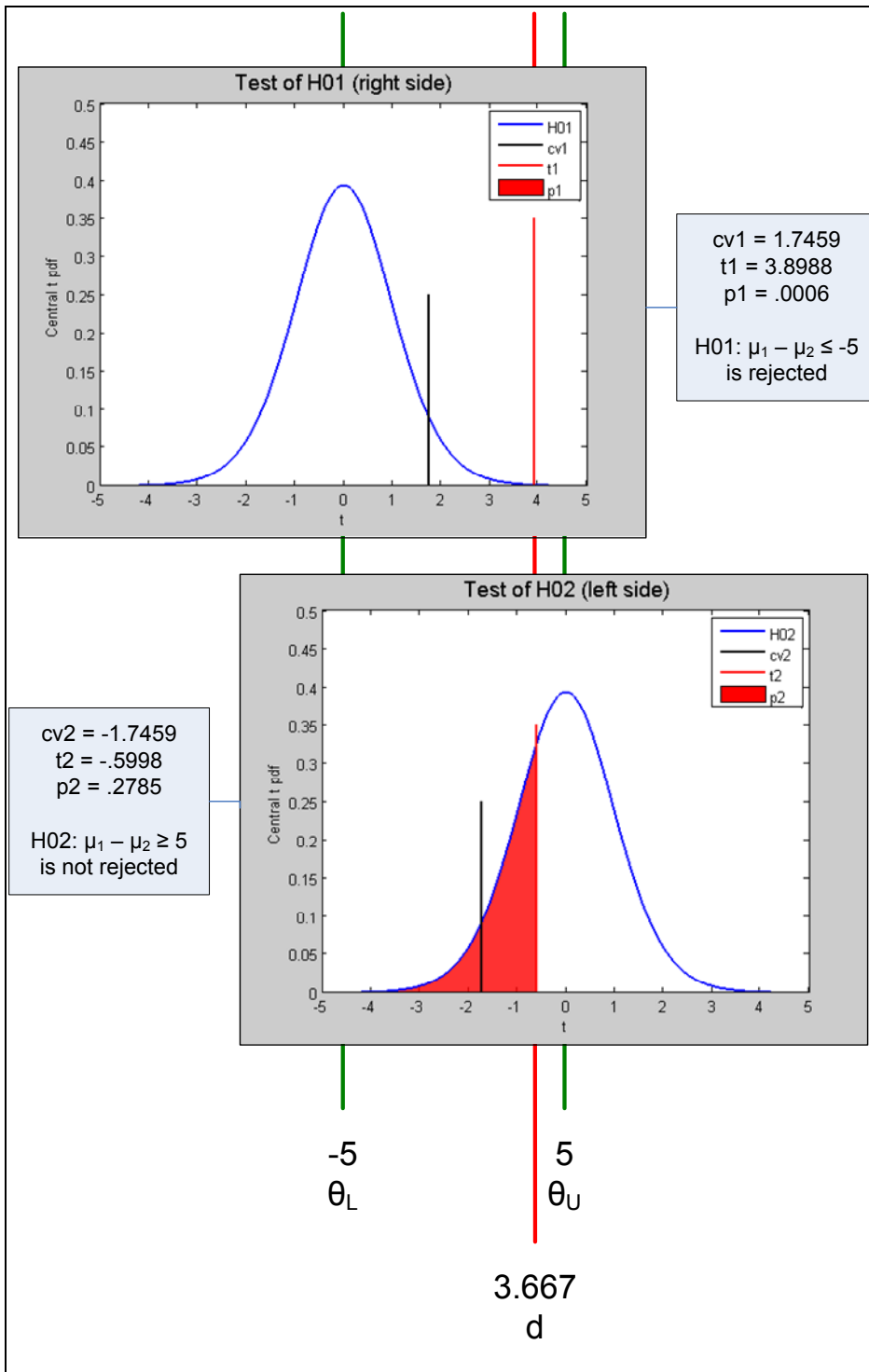
- Levene, H. (1960). Robust tests for equality of variances. In I. O. e. al. (Ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292): Stanford University Press.
- Lewis, S. E., & Lewis, J. E. (2005). The same or not the same: Equivalence as an issue in educational research. *Journal of Chemical Education*, 82, 1408-1412.
- McBride, G. B. (1999). Equivalence tests can enhance environmental science and management. *Australian and New Zealand Journal of Statistics*, 41, 19-29.
- Mendenhall, W. e. a. (2007). *Mathematical statistics with applications*. Boston: Duxbury Press.
- Meyners, M. (2007). Least equivalent allowable differences in equivalence testing. *Food quality and preference*, 18(3), 541-547.
- Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. Chicago, IL: Aldine Publishing Company.
- Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t test versus Satterthwaite's approximate F test. *Communications in Statistics - Theory and Methods*, 18(11), 3963 - 3975.
- Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician*, 46(1), 19-21.
- Newman, M. C., & Strojan, C. L. (1998). *Risk assessment : Logic and measurement*. Chelsea, MI: Ann Arbor Press.
- Oakes, M. (1990). *Statistical inference*. Newton Lower Falls, MA: Epidemiology Resources.

- Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *BioScience*, *51*, 1051-1057.
- Patterson, S. D., & Jones, B. (2006). *Bioequivalence and statistics in clinical pharmacology*. Boca Raton, FL: Chapman & Hall/CRC.
- Riffenburgh, R. H. (2006). *Statistics in medicine* (2nd ed.). Amsterdam ; Boston: Elsevier Academic Press.
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553-565.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110-114.
- Schuirman, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657-680.
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research. *American Psychologist*, *40*, 73-83.
- Stevenson, W. J. (2008). *Operations management, 10th ed.* Boston, MA: McGraw-Hill.
- Student (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.
- Streiner, D. L. (2003). Unicorns do exist: A tutorial on "proving" the null hypothesis. *Canadian Journal of Psychiatry*, *48*, 756-761.
- Terry, N. (2007). Assessing instruction modes for master of business administration (MBA) Courses. *Journal of Education for Business*, *82*(4), 220-226.

- Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. Boca Raton, Fla.: Chapman & Hall/CRC.
- Welling, P. G., Tse, F. L. S., & Dighe, S. V. (1991). *Pharmaceutical bioequivalence*. New York: Dekker.
- Woolsey, M. L., Harrison, T. J., & Gardner, R., III. (2004). A Preliminary examination of instructional arrangements, teaching behaviors, levels of academic responding of deaf middle school students in three different educational settings. *Education and Treatment of Children*, 27(3), 263-279.
- Zabell, S.L. (2008). On Student's 1908 Article "The Probable Error of a Mean". *Journal of the American Statistical Association*, 103(481), 1-7.
- Ziliak, S. T., & McCloskey, D. N. (2007). *The cult of statistical significance*. Ann Arbor, MI: The University of Michigan Press.
- Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *Journal of General Psychology*, 123.

APPENDIX A

Plots representing the equivalence test applied to the data set presented on page 23



APPENDIX B

PVT_SAT_alpha.sas program

```
/* PVT_SAT_alpha.sas */
/* Last revised on 11/28/2008 */
/* Purpose: count the prop of H0 rejections */
/* This prop is actual alpha, since H0 is true (mean1=mean2) */

data _null_;
  call symput('td_mv',trim(left(put(today()),date9.)));
run;

options nonotes nosource;
*options notes source;

ods noresults;
*ods results;

ods listing close;
*ods listing;

/* enter constants as macro variables */
/* negative seed: different data every time */
%let iternum=10000;
%let seed=-1;
%let mean1=50;
%let mean2=50;
%let n=100;
%let alpha=.05;

/* enter sample allocations and variances as a data set */
data Combinations;
  input combo n n1 n2 var1 var2;
  datalines;
1 100 50 50 20 20
2 100 50 50 25 15
3 100 50 50 30 10
4 100 50 50 35 5
5 100 50 50 37 3
6 100 50 50 38 2
7 100 40 60 20 20
8 100 40 60 25 15
9 100 40 60 30 10
10 100 40 60 35 5
11 100 40 60 37 3
12 100 40 60 38 2
13 100 30 70 20 20
14 100 30 70 25 15
15 100 30 70 30 10
16 100 30 70 35 5
17 100 30 70 37 3
```

```

18 100 30 70 38 2
19 100 20 80 20 20
20 100 20 80 25 15
21 100 20 80 30 10
22 100 20 80 35 5
23 100 20 80 37 3
24 100 20 80 38 2
25 100 10 90 20 20
26 100 10 90 25 15
27 100 10 90 30 10
28 100 10 90 35 5
29 100 10 90 37 3
30 100 10 90 38 2
;
run;

/* create base dataset in which generated values will accumulate */
/* this will grow into a tall dataset: c*iter=30*10000=300,000 rows */
data ttests;
  input Combo Iter n1 n2 var1 var2 Method $ tValue DF CV RejectH0;
  datalines;
  . . . . .
;
run;
data ttests;
  format Method $13.;
  set ttests;
  where Combo~=. ;
run;

/* begin macro */
%macro sampling;

/* outer loop: rows of the Combinations dataset */
%do c=1 %to 30;

/* inner loop: number of iterations per combination */
%do iter=1 %to &iternum;

/* capture n and var values pertaining to current combo & iter */
data _null_;
  set combinations;
  if combo=&c then do;
    call symput('n1',put(n1,8.));
    call symput('n2',put(n2,8.));
    call symput('var1',put(var1,8.));
    call symput('var2',put(var2,8.));
  end;
run;

```

```

/* create sample with current n and var values */
data combo&c._iter&iter;
  do i=1 to &n;
    select;
      when (1<=i<=&n1) do; group=1; mean=&mean1; var=&var1; end;
      when (&n1<i<=&n) do; group=2; mean=&mean2; var=&var2; end;
    end;
    output;
  end;
run;

/* add random variable to sample */
data combo&c._iter&iter;
  set combo&c._iter&iter;
  X=mean+sqrt(var)*rannor(&seed);
  format X 8.4;
run;

/* run t-test and capture test values */
proc ttest data=combo&c._iter&iter;
  class group;
  var x;
  ods output ttests=tt_combo&c._iter&iter;
run;

/* count H0 rejections */
/* the rejection proportion is the actual alpha */
data tt_combo&c._iter&iter;
  retain combo &c iter &iter n1 &n1 n2 &n2 var1 &var1 var2 &var2;
  set tt_combo&c._iter&iter;
  keep combo iter method n1 n2 var1 var2 cv tvalue df rejectH0;
  cv=TINV(1-(&alpha)/2,df);
  select;
    when (abs(tvalue)>cv) RejectH0=1;
    when (abs(tvalue)<=cv) RejectH0=0;
  end;
run;

/* accumulate values into base dataset */
proc append base=ttests data=tt_combo&c._iter&iter;
run;

proc datasets nolist;
  delete combo&c._iter&iter tt_combo&c._iter&iter;
quit;
run;

%end;

%end;

%mend;

/* end macro */

```

```

%sampling;

/* create summary dataset */
proc sort data=ttests out=ttests;
  by method combo iter;
proc means noprint data=ttests;
  output out=ttests_means
  mean(RejectH0)=ActualAlpha;
  by method combo;
  id n1 n2 var1 var2;
run;

proc datasets nolist;
  delete ttests;
quit;
run;

/* create PVT table */
data ttests_means_pool;
  set ttests_means;
  where method='Pooled';
  keep n2 var2 ActualAlpha;
run;
proc transpose prefix=var2_ data=ttests_means_pool out=PVT_alpha;
  by n2;
  var ActualAlpha;
  id var2;
run;
proc sort data=PVT_alpha;
  by descending n2;
run;

/* create SAT table */
data ttests_means_satt;
  set ttests_means;
  where method='Satterthwaite';
  keep n2 var2 ActualAlpha;
run;
proc transpose prefix=var2_ data=ttests_means_satt out=SAT_alpha;
  by n2;
  var ActualAlpha;
  id var2;
run;
proc sort data=SAT_alpha;
  by descending n2;
run;

/* print summary dataset */
ods listing;
options nodate pageno=1 pagesize=70 linesize=79;
proc printto file=
"C:\MyDocuments\dissertation
\Simulations\OutputFiles
\PVT_SAT_alpha_out_&td_mv..txt" new;
run;

```



```

proc print data=ttests_means;
  title1 "PVT_SAT_alpha with &iternum iterations per combination";
  title2 "H0 true: mean1=&mean1, mean2=&mean2";
  var method combo n1 n2 var1 var2 actualalpha;
run;
proc print noobs data=PVT_alpha;
  title3 'PVT table: Actual alpha';
run;
proc print noobs data=SAT_alpha;
  title3 'SAT table: Actual alpha';
run;

proc printto file=print;
run;

/* export PVT and SAT tables */
proc export
  data=PVT_alpha
  outfile=
"C:\MyDocuments\Dissertation
\Simulations\OutputFiles
\PVT_SAT.xls"
dbms=Excel2000 replace;
sheet="Alpha_PVT";
run;
proc export
  data=SAT_alpha
  outfile=
"C:\MyDocuments\Dissertation
\Simulations\OutputFiles
\PVT_SAT.xls"
dbms=Excel2000 replace;
sheet="Alpha_SAT";
run;

*options nonotes nosource;
options notes source;

*ods noresults;
ods results;

*ods listing close;
ods listing;

```

APPENDIX C

BF_ET_alpha_n50.sas program

```
/* SAS_programs\HOV\BF_ET_alpha_n50.sas */
/* Purpose: implement both rules and count the prop of H0 rejections */
/* last revised 12/05/2008 */

data _null_;
  call symput('td_mv',trim(left(put(today()),date9.)));
run;

options pageno=1 nodate;
options nonotes nosource;
*options notes source;

ods noresults;
*ods results;
ods listing close;
*ods listing;

/* enter constants as macro variables */
/* negative seed: different data every time */
%let iternum=1000;
%let n=50;
%let combonum=30;
%let seed=-1;
%let mean1=50;
%let mean2=50;
%let alpha=.05;

/* enter sample allocations and variances as a data set */
data Combinations;
  input combo n1 n2 var1 var2;
  datalines;
1 25 25 20 20
2 25 25 25 15
3 25 25 30 10
4 25 25 35 5
5 25 25 37 3
6 25 25 38 2
7 20 30 20 20
8 20 30 25 15
9 20 30 30 10
10 20 30 35 5
11 20 30 37 3
12 20 30 38 2
13 15 35 20 20
14 15 35 25 15
15 15 35 30 10
16 15 35 35 5
```

```

17 15 35 37 3
18 15 35 38 2
19 10 40 20 20
20 10 40 25 15
21 10 40 30 10
22 10 40 35 5
23 10 40 37 3
24 10 40 38 2
25 5 45 20 20
26 5 45 25 15
27 5 45 30 10
28 5 45 35 5
29 5 45 37 3
30 5 45 38 2
;
run;

/* create base dataset in which generated values will accumulate */
data ttests;
  input Combo Iter Rule $ Method $ n1 n2 var1 var2 tValue DF probt CV
  RejectH0;
  datalines;
  . . . . .
;
run;
data ttests;
  format Rule $2. Method $13.;
  set ttests;
  where Combo~=. ;
run;

/* begin macro */
%macro macro1;

/* outer loop: rows of the Combinations dataset */
%do c=1 %to &combonum;

/* inner loop: number of iterations per combination */
%do iter=1 %to &iternum;

/* capture n and var values pertaining to current combo & iter */
data _null_;
  set combinations;
  if combo=&c then do;
    call symput('n1',put(n1,8.));
    call symput('n2',put(n2,8.));
    call symput('var1',put(var1,8.));
    call symput('var2',put(var2,8.));
  end;
run;

```

```

/* create sample with current n and var values */
data combo&c._iter&iter;
  do i=1 to &n;
    select;
      when (1<=i<=&n1) do; group=1; mean=&mean1; var=&var1; end;
      when (&n1<i<=&n) do; group=2; mean=&mean2; var=&var2; end;
    end;
    output;
  end;
run;

/* add random variable to sample */
data combo&c._iter&iter;
  set combo&c._iter&iter;
  X=mean+sqrt(var)*rannor(&seed);
  format X 8.4;
run;

/* run ttest (with both PVT and SAT results) */
proc ttest data=combo&c._iter&iter;
  class group;
  var x;
  ods output ttests=tt_combo&c._iter&iter;
run;

/* run BF HOV test and record BF recommendation */
proc glm data=combo&c._iter&iter;
  class group;
  model x=group;
  means group / hovtest=bf;
  ods output hovFtest=bf1;
quit;
run;
data bf2;
  set bf1;
  where source='group';
  combo=&c;
  iter=&iter;
  select;
    when (probf<.05) bf_rec='Satterthwaite';
    when (probf>=.05) bf_rec='Pooled';
  end;
  keep combo iter probf bf_rec;
  call symput('bf_rec_mv',put(bf_rec,$13.));
run;

/* count H0 rejections for the BF rule */
/* the rejection proportion is the actual alpha */
data bf_tt_combo&c._iter&iter;
  retain rule 'BF' combo &c iter &iter n1 &n1 n2 &n2 var1 &var1 var2
&var2;
  set tt_combo&c._iter&iter;
  where method="&bf_rec_mv";
  keep rule combo iter method n1 n2 var1 var2 cv tvalue df probt
rejectH0;

```

```

cv=TINV(1-(&alpha)/2,df);
select;
  when (abs(tvalue)>cv) RejectH0=1;
  when (abs(tvalue)<=cv) RejectH0=0;
end;
run;

/* run ET HOV test and record ET recommendation */

/* compare variances using the F-ratio test */
/* ET test 1 (right side): H0_ratio=1/2 */
proc means noprint data=combo&c._iter&iter;
  output out=et_right_1
    var(x)=Var_X;
  by group;
run;
proc transpose prefix=var data=et_right_1 out=et_right_2;
  var var_x;
  id group;
run;
proc transpose prefix=n data=et_right_1 out=et_right_3;
  var _freq_;
  id group;
run;
data et_right_4;
  merge et_right_2 et_right_3;
  drop _name_;
  combo=&c;
  iter=&iter;
  df1=n1-1;
  df2=n2-1;
  H0_ratio=.5;
  obs_ratio=var1/var2;
  F_value=(1/H0_ratio)*(var1/var2);
  cv_right=FINV(1-.05,df1,df2);
  p_right=1-CDF('F',F_value,df1,df2);
select;
  when (p_right<.05) et_right='rejectH01';
  when (p_right>=.05) et_right='retainH01';
end;
run;

/* compare variances using the F-ratio test */
/* ET test 2 (left side): H0_ratio=2 */
proc means noprint data=combo&c._iter&iter;
  output out=et_left_1
    var(x)=Var_X;
  by group;
run;
proc transpose prefix=var data=et_left_1 out=et_left_2;
  var var_x;
  id group;
run;

```

```

proc transpose prefix=n data=et_left_1 out=et_left_3;
  var _freq_;
  id group;
run;
data et_left_4;
  merge et_left_2 et_left_3;
  drop _name_;
  combo=&c;
  iter=&iter;
  df1=n1-1;
  df2=n2-1;
  H0_ratio=2;
  obs_ratio=var1/var2;
  F_value=(1/H0_ratio)*(var1/var2);
  cv_left=FINV(.05,df1,df2);
  p_left=CDF('F',F_value,df1,df2);
  select;
    when (p_left<.05) et_left='rejectH02';
    when (p_left>=.05) et_left='retainH02';
  end;
run;

/* combine ET right and ET left to create ET HOV decision */
data et_right_5;
  set et_right_4;
  keep combo iter et_right;
run;
data et_left_5;
  set et_left_4;
  keep combo iter et_left;
run;
data et1;
  merge et_right_5 et_left_5;
  by combo iter;
  select;
    when (et_right='rejectH01' and et_left='retainH02')
et_rec='Satterthwaite';
    when (et_right='retainH01' and et_left='rejectH02')
et_rec='Satterthwaite';
    when (et_right='retainH01' and et_left='retainH02')
et_rec='Satterthwaite';
    when (et_right='rejectH01' and et_left='rejectH02')
et_rec='Pooled';
  end;
  call symput('et_rec_mv',put(et_rec,$13.));
run;

```

```

/* count H0 rejections for the ET rule */
/* the rejection proportion is the actual alpha */
data et_tt_combo&c._iter&iter;
  retain rule 'ET' combo &c iter &iter n1 &n1 n2 &n2 var1 &var1 var2
&var2;
  set tt_combo&c._iter&iter;
  where method="&et_rec_mv";
  keep rule combo iter method n1 n2 var1 var2 cv tvalue df probt
rejectH0;
  cv=TINV(1-(&alpha)/2,df);
  select;
    when (abs(tvalue)>cv) RejectH0=1;
    when (abs(tvalue)<=cv) RejectH0=0;
  end;
run;

data tt_combo&c._iter&iter;
  retain Combo Iter Rule Method n1 n2 var1 var2 tValue DF probt CV
RejectH0;
  set bf_tt_combo&c._iter&iter et_tt_combo&c._iter&iter;
run;

/* accumulate values into base dataset */
proc append base=ttests data=tt_combo&c._iter&iter;
run;
proc datasets nolist;
  delete combo&c._iter&iter tt_combo&c._iter&iter
    bf_tt_combo&c._iter&iter et_tt_combo&c._iter&iter
    bf1 bf2 et1 et_left_1 et_left_2 et_left_3 et_left_4 et_left_5
    et_right_1 et_right_2 et_right_3 et_right_4 et_right_5;
quit;
run;

%end;

%end;

%mend;

/* end macro */

%macro1;

/* create summary dataset */
proc sort data=ttests out=ttests;
  by rule combo iter;
proc means noprint data=ttests;
  output out=ttests_means
  mean(RejectH0)=ActualAlpha;
  by rule combo;
  id n1 n2 var1 var2;
run;
/*
proc datasets nolist;
  delete ttests;

```

```

quit;
*/
/* create BF table */
data ttests_means_bf;
  set ttests_means;
  where rule='BF';
  keep n2 var2 ActualAlpha;
run;
proc transpose prefix=var2_ data=ttests_means_bf out=BF_alpha;
  by n2;
  var ActualAlpha;
  id var2;
run;
proc sort data=BF_alpha;
  by descending n2;
run;

/* create ET table */
data ttests_means_et;
  set ttests_means;
  where rule='ET';
  keep n2 var2 ActualAlpha;
run;
proc transpose prefix=var2_ data=ttests_means_et out=ET_alpha;
  by n2;
  var ActualAlpha;
  id var2;
run;
proc sort data=ET_alpha;
  by descending n2;
run;

/* create dataset BF-ET agreement table */
data _null_;
  total_iter=&iternum*&combonum;
  call symput('total_iter',trim(left(put(total_iter,10.))));
run;
data agree1;
  set ttests;
  keep rule method;
  if method='Pooled' then method='PVT';
  if method='Satterthwaite' then method='SAT';
run;
proc sort data=agree1;
  by rule method;
run;
proc transpose data=agree1 out=agree2;
  by rule;
  var method;
run;
data agree2;
  set agree2;
  drop _name_;
run;

```



```

proc transpose data=agree2 out=agree3;
  var coll-col&total_iter;
  id rule;
run;
data agree3;
  set agree3;
  drop _name_;
run;
proc datasets nolist;
  delete agree1 agree2;
quit;
run;

/* print summary dataset and related tables */
ods listing;
options nodate pageno=1 pagesize=70 linesize=79;

proc printto file=
"C:\MyDocuments\Dissertation
\Simulations\OutputFiles
\BF_ET_alpha_n&n._&td_mv..txt" new;
run;
proc print data=ttests_means;
  title1 "BF_ET_alpha with &iternum iterations per combination (n=&n)";
  title2 "H0 true: mean1=&mean1, mean2=&mean2";
  var rule combo n1 n2 var1 var2 actualalpha;
run;
proc print noobs data=BF_alpha;
  title3 'BF table: Actual alpha';
run;
proc print noobs data=ET_alpha;
  title3 'ET table: Actual alpha';
run;
proc freq data=agree3;
  title3 'BF-ET agreement table';
  table bf*et;
run;
proc printto file=print;
run;

/* export BF, ET */
proc export
  data=BF_alpha
  outfile=
"C:\MyDocuments\Dissertation
\Simulations\OutputFiles
\BF_ET_&td_mv..xls"
dbms=Excel2000 replace;
sheet="Alpha_BF_n&n";
run;

```

```
proc export
  data=ET_alpha
  outfile=
  "C:\MyDocuments\Dissertation
  \Simulations\OutputFiles
  \BF_ET_&td_mv..xls"
  dbms=Excel2000 replace;
  sheet="Alpha_ET_n&n";
run;
proc printto file=print;
run;
title1 ' ';
title2 ' ';
title3 ' ';
*options nonotes nosource;
options notes source;
*ods noresults;
ods results;
```

APPENDIX D

Pub6414 data set

| Fall_2007_Online | | Spring_2007_InClass |
|------------------|--------|---------------------|
| 93.96 | 97.43 | 92.75 |
| 96.08 | 100.00 | 90.75 |
| 80.74 | 93.35 | 94.75 |
| 97.20 | 93.83 | 90.50 |
| 80.89 | 92.08 | 88.25 |
| 94.48 | 80.04 | 94.00 |
| 66.72 | 96.90 | 91.50 |
| 96.68 | 63.28 | 90.75 |
| 80.66 | 97.75 | 95.75 |
| 98.08 | 95.78 | 86.00 |
| 88.73 | 88.33 | 88.50 |
| 95.60 | 95.58 | 100.00 |
| 95.99 | 96.31 | 95.75 |
| 97.12 | 39.61 | 89.00 |
| 86.24 | 94.39 | 87.25 |
| 97.50 | 99.20 | 78.75 |
| 98.33 | 85.73 | 79.25 |
| 86.85 | 90.34 | 95.00 |
| 96.15 | 92.13 | 93.00 |
| 95.21 | 79.74 | 97.00 |
| 93.70 | 90.33 | 87.75 |
| 91.13 | 90.50 | 96.25 |
| 95.53 | 96.07 | 97.25 |
| 90.18 | 96.14 | 74.25 |
| 78.74 | 95.79 | 92.25 |
| 99.50 | 92.91 | 87.25 |
| 94.08 | 90.90 | 71.75 |
| 86.61 | 91.86 | 89.50 |
| 98.31 | 90.21 | 73.50 |
| 87.58 | 95.09 | 99.75 |
| 99.50 | 89.20 | 81.50 |
| 90.13 | 87.11 | 87.75 |
| 93.18 | 95.73 | 93.25 |
| 96.80 | 96.67 | 100.00 |
| 92.80 | 95.75 | 96.50 |
| 92.51 | 94.61 | 76.50 |
| 92.30 | 96.98 | 88.50 |
| | 89.84 | |
| | 98.46 | |
| | 92.63 | |
| | 93.29 | |

APPENDIX E

Plots representing the equivalence test of homoskedasticity (PubH5414 data)

