5-2024

AN INVESTIGATION OF THE EFFECTIVENESS OF STUDENT'S t-TEST UNDER HETEROGENEITY OF VARIANCE

Hayden Nelson
Western Kentucky University, haydenpnelson@gmail.com

AN INVESTIGATION OF THE EFFECTIVENESS OF STUDENT'S *t*-TEST UNDER HETEROGENEITY OF VARIANCE

A Thesis submitted in partial fulfillment of the requirements for the degree Master of Science in Industrial-Organizational Psychology

Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

By
Hayden Peter Nelson

May, 2024

An Investigation Of The Effectiveness Of Student's T-Test Under Heterogeneity Of Variance

Hayden Peter Nelson

Date Recommended 4/11/2024

DocuSigned by:
Reagan Brown
B66126D3D0A24BF...
Chair

DocuSigned by:
Katrina Burch
75BB04A0959F41C...
Committee Member

DocuSigned by:
Xiaowen Chen
EE91D1B056514FB...
Committee Member

Committee Member

DocuSigned by:
Jennifer Hammonds
EBE3858E068F42D...
Interim Director of the Graduate School

ABSTRACT

AN INVESTIGATION OF THE EFFECTIVENESS OF STUDENT'S *t*-TEST UNDER HETEROGENEITY OF VARIANCE

Within the field of psychology, few tests have been as thoroughly investigated as Student's *t*-test. One area of criticism is the use of the test when the assumption for heterogeneity of variance between two samples is violated, such as when sample sizes and observed sample variances are unequal. The current study proposes a Monte Carlo analysis to observe a broad range of conditions in efforts to identify the resulting fluctuations in the proportion obtained significant results for two conditions: no mean difference ($\mu_1 = \mu_2$) compared to the set level of alpha, and small-to-moderate mean differences ($\mu_1 \neq \mu_2$) compared to the expected power. For each condition, population standard deviations and sample sizes will be changed incrementally. Results indicate that outside of conditions with extreme differences in population standard deviations and relative sample sizes will produce results comparable to conditions with homogenous sampling conditions at roughly the same rate. As differences between population means are increased, researchers also need not be concerned with massive losses to statistical power. Future directions for researchers are discussed further.

Keywords: Monte Carlo, Pooled samples *t*-test, Heterogeneity, Standard deviation, Sample size

I dedicate this thesis to my parents, Christine and Kirk Nelson, my siblings, Eric, Jenna, Karlie, and Adam, and my grandparents.

ACKNOWLEDGEMENTS

I would like to acknowledge the teachers and mentors I have had that helped me define myself as a student. Though I may not have started education while taking myself seriously, a select few instructors have enabled me to achieve academic success. I would like to thank Quinten Lyon for inspiring my maturity in academia and for instructing me over the basics of writing and argument structure, and Jennifer Higgins for allowing my interest in mathematics to manifest. I would like to thank Dr. Chris Waples for introducing me to the fields of I-O psychology and statistics, and for being a figure in my life I could work for to impress. I would like to thank Dr. Julie Lanz for introducing me to research in psychology and connecting me to opportunities I would otherwise not have found on my own. Finally, I would like to thank Dr. Reagan Brown, Dr. Xiaowen Chen, and Dr. Katrina Burch for their patience, expertise, and instruction during one of the greatest periods of change in my life. If not for individuals like them and their support, I could have never made it this far in my career. I am eternally grateful.

TABLE OF CONTENTS

An Investigation of the Effectiveness of Student's *t*-Test Under Heterogeneity of Variance

Within the field of psychology, few concepts have been as deeply investigated as Student's *t*-test (Student, 1908). Student's *t*-test, also known as the pooled *t*-test, is one of the foundational significance tests within the field as it is the obvious test of choice for testing mean differences in a two independent samples experiment (e.g., treatment-control groups). In an industrial-organizational setting, Student's *t*-test can see use when determining the mean differences of two independent samples. For example: determining whether a group of employees introduced to a training program have improved job performance over those that were not (treatment-control design). In addition, the test has clear connections to one-way ANOVA. The popularity of the test may also be due to research indicating that it is one of the most uniformly powerful tests when conditions are met (Sawilowsky & Blair, 1992). As with all significance tests, research and professional guidance caution against the use of the pooled sample *t*-test when the assumptions for its use are not met. These cautions primarily concern the case when the assumption of homogeneity of variance (equal variance in the population) is violated (Bradley, 1978).

Student's *t*-test is computed as follows:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\,. \tag{1}$$

Where

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\,, \tag{2}$$

and

$$df = n_1 + n_2 - 2\,. \tag{3}$$

Even when the assumption of homogeneity of variance is true, the sample variances are likely to never be equal due to the vagaries of sampling error. Student's *t*-test deals with discrepancies in the sample variances by creating a weighted average of the sample variances to create a pooled estimate of the variance (i.e., S_p^2).

Samples from populations with unequal variance can reduce the effectiveness of the significance test leading to erroneous conclusions, a limitation known as the Behrens-Fisher problem (Kim & Cohen, 1998). When the null hypothesis is true (i.e., no difference between population means), a violation of the assumption of homogeneity of variance can result in an inflation of the Type I error rate (Bradley, 1978). When the null hypothesis is false (i.e., unequal population means), heterogeneous variance reduces the statistical power of the test (Sawilowsky & Blair, 1992). In these scenarios, there is much risk posed to organizations implementing the test. When there is no difference between groups, businesses risk wasting resources on practices with no potential return on investment. When there is a difference between groups, and the observable effect is not correctly identified, they potentially lose out on practices that improve internal functioning.

In response to the Behrens-Fisher problem, a few solutions have been identified (Scheffe, 1970). The most practical and well-regarded of these solutions is the Welch's approximate *t*-test (Welch, 1937), which separates each sample variance within the study. The main benefit to using Welch's solution is that when sample variances are equal, it acts as an equivalent test to the original pooled samples *t*-test (Derrick et al., 2016). The Welch approximate *t*-test is also a much simpler solution compared to other alternatives (Scheffe, 1970). Because of its improved robustness over the original, some researchers have argued for the replacement of the original with the Welch (e.g., Delacre et al., 2017). The Welch approximate *t*-test is computed as follows:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}. \quad (4)$$

Degrees of freedom are computed as:

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}. \quad (5)$$

Although the benefits of using Welch approximate *t*-test in place of Student's t-test have been described by several studies (Delacre et al., 2017; Lu & Yuan, 2010; Wang, 1971), there has yet to be a detailed description of the exact conditions under which the pooled samples *t*-test fails to function effectively. Investigation of the point (in terms of unequal sample sizes and/or population variances) at which the results of the pooled sample *t*-test has elevated Type I error (when the null is true) or decreased statistical power (when the null is false) allows researchers to determine which test is called for in a given situation. This information may be particularly valuable in applied settings (e.g., an evaluation of training program effectiveness in which organizational decisions and resources are contingent on the results).

Heterogeneity of Variance and Test Effectiveness

Although the pooled samples *t*-test has been found to have elevated Type I error rates when the population variances are heterogeneous, it is still a reasonably robust significance test . In the case where obtained sample variances are different, the pooled test is still able to achieve robust results as long as the sample sizes are equal (Sawilowsky & Blair, 1992). The conditions under which the pooled *t*-test becomes non-robust has been debated somewhat (Bradley, 1978), but it has generally been stated that when both the sample size and variance differ the effectiveness of the test is diminished (Boneau, 1960; Zimmerman & Zumbo, 2009).

The Role of Relative Sample Sizes

When both the sample size and variance are unequal, there is also the matter of which group has the larger portion of the variance. If the larger sample also has greater variance, the pooled shows some reduction from the expected rate of alpha (i.e., the null is rejected at a rate less than alpha). In the opposite case, in which the smaller sample has the larger variance, the Type I error rate is elevated (Derrick et al., 2016). In considering the results of the significance test, a researcher may consider sample sizes when there appears to be a substantial difference between sample variances as a potential explanation of error in the findings (Zimmerman & Zumbo, 2009). These studies extend the understanding of the Behrens-Fisher problem by expanding on the conditions that may lead to errors in interpretation (Kim & Cohen, 1998).

Welch's Approximate Solution

The solution offered by Welch (1937) takes a different approach to handling sample variances; instead of forming a pooled variance of the two samples, the Welch method includes the sample variances independently. In practical terms, this approach helps reduce the rate of Type I error inflation (Wang, 1971). The estimated level of alpha within the results of the Welch roughly equals the set level of alpha under a variety of sample size and variance conditions (Lu & Yuan, 2010). This property is particularly relevant when the conditions of the study reflect those examined by Derrick et al. (2016), in which two samples each of relatively small size exhibit large differences in variances. When presented with these conditions, the estimated Type I error for the Welch test reflects the rates observed when sample groups are normally distributed.

The Welch approximate *t*-test has a similar history of review to that of the pooled samples *t*-test and is often used to illustrate the shortcomings of the latter. The robustness of the

Welch approximate *t*-test has been investigated in terms of the expected loss of statistical power and Type I error rate. When samples are unequal in terms of their obtained sizes, the Welch approximate *t*-test exhibits estimated power comparable to normally distributed data (Sawilowsky & Blair, 1992). Furthermore, the Type I error rate for both equal samples and unequal samples (e.g., n_1 = 10 and n_2 = 20) reflected the set level of alpha (Wang, 1971). Being able to express the limits of the usability of the pooled *t*-test will help both researchers and practitioners make decisions regarding the use of a certain test.

Delacre et al. (2017) argued that researchers should never assume equality of variances and should always use the Welch approximate *t*-test. In their Monte Carlo analysis of situations where the null hypothesis is true and population variances are unequal, Delacre et al. observed that Student's *t*-test consistently underperformed (compared to expected Type I error rates as well as Welch's approximate *t*-test) when the sample sizes were unequal. Under the condition where the larger sample (n_1 = 60 vs n_2 = 40) was drawn from the population with the greater standard deviation, the frequency of rejecting the null was 2.8 percent. In the reverse condition where the smaller sample had a larger standard deviation, the frequency of rejecting the null increased to 8.3 percent. Neither of these results reflect the actual Type I error rate of .05. In addition, the Welch's approximate *t*-test displayed consistent results of approximately .05 across all conditions.

Although these findings support much of what has been discussed in the literature of the two significance tests, there is a lack of clear definition for when exactly Student's *t*-test loses the ability to perform at an acceptable level. In the aforementioned study, the samples generated had considerably large differences in sample sizes (n_1 was 50% greater than n_2), and the ratio of the larger the standard deviation to the smaller the standard deviation was 2.0. By focusing on

extreme cases of relative sample sizes heterogeneity of variance, we may fail to acknowledge the meaningful range where Student's *t*-test provides robust results.

There is value to identifying the limit for Student's *t*-test's tolerance to heterogeneity in the obtained samples, especially given its use in applied settings. Although Welch's approximate *t*-test appears to be a stronger test for determining mean group differences, Student's *t*-test offers simplicity in how the results of the significance test are obtained. In addition, the pooled *t*-test (two-tailed test) yields the same result as an ANOVA run on the same data ($(t_{obs})^2 = F_{obs}$ when $a = 2$), In the case where an industrial-organizational practitioner has to elaborate on the decision to use Welch's approximate *t*-test, they may run into difficulties conveying to a lay audience why a different calculation method is warranted.

Monte Carlo Simulation

The nature of researching topics such as the effectiveness of statistical formulas makes typical research methods impractical. Monte Carlo research designs allow researchers to study conditions that would otherwise be impractical to manipulate (Harrison, 2010). A Monte Carlo simulation is a research strategy that allows researchers to conduct a large number (e.g., millions) of replications of an experiment within a short period of time (Mooney, 1997). Additionally, the population parameters from which the studies are drawn from are set by the researcher to observe results under a variety of conditions and to allow the researcher to compare the results to these known population parameters.

The Current Study

A Monte Carlo simulation was used to determine the point at which the Student's *t*-test fails to perform as designed. Based on results from previous research (Delacre et al., 2017; Derrick et al., 2016; Wang, 1971), heterogeneity of variance should result in rejection rates that

depart from alpha when the null is true and from expected power when the null is false as the relative sample sizes depart from equality.

Method

Procedure

To simulate the results of the pooled samples *t*-test under various conditions, the open-source statistical software program R was used. The effectiveness of the pooled *t*-test was examined under two general conditions: equal population means (i.e., the null hypothesis is true) and unequal population means (i.e., the null hypothesis is false). For the true null hypothesis condition the populations means were set to zero for both groups (i.e., $\mu_1 = \mu_2 = 0$). For the false null hypothesis conditions, the population means were set to differing values reflecting varying levels of population effect sizes ($\mu_1 = .10$, $\mu_2 = 0.0$; $\mu_1 = .25$, $\mu_2 = 0.0$; $\mu_1 = .50$, $\mu_2 = 0.0$). Population standard deviations were ranged from equal ($\sigma_1 = \sigma_2 = 1.0$) to unequal (e.g., $\sigma_1 = 1.0$, σ_2 as low as .25).

The total sample size in each condition was held constant ($N = 40$). The respective sample sizes were investigated at values ranging from equal ($n_1 = n_2 = 20$) to extremely different ($n_1 = 12$, $n_2 = 28$). This sample size was selected because problems with the pooled *t*-test were more likely to be found at smaller sample sizes (Boneau, 1960).

Analyses

The outcome of each iteration of the study was recorded as whether the null hypothesis was rejected for a two-tailed test with alpha set to .05 (where a retained null hypothesis result equals 0 and a rejected null hypothesis result equals 1). Results were averaged across 100,000 iterations (i.e., independent samples). Finally, observed rejection rates were compared to the

expected rates (alpha for the true null and power for the false null) to identify conditions under which Student's *t*-test does not function properly.

Results

Equal Population Means Condition

To determine the effectiveness of the Student's *t*-test when the population means are equal (i.e., the null is true), the proportion of results that were significant was computed across an array of heterogeneous sampling conditions (see Appendix A). As the disparity between sample size and standard deviation increases, observed significance rates deviate from alpha. Similar to previous findings (Sawilowsky & Blair, 1992), the pooled *t*-test produces accurate results (i.e., significant results in approximately five percent, the chosen alpha level, of the samples) in situations for which either sample sizes or standard deviations are equal.

Furthermore, as long as the population standard deviations are not extremely different (i.e., the smaller standard deviation is at least 75% of the larger or 56% of the variance), even the most extreme sample size disparity ($n_1 = 12$, $n_2 = 28$, as well as the inverse) resulted in significance rates that deviated from alpha by less than three percentage points. Large departures from alpha were observed only when the smaller population standard deviation was less than or equal to 50% of the larger (25% in terms of variance) and sample sizes were not close (e.g., $n_1 = 16$, $n_2 = 24$).

Unequal Population Means Condition

To determine the conditions in which the pooled *t*-test fails to produce accurate results when the population means are not equal (i.e., the null is false), the frequency of significant results was compared to the theoretical estimate of power for each condition. In the condition where the population mean difference is small ($d = .10$ when population standard deviations are

equal), the errors (difference between observed significance and expected significance rates) for concluding significant results approximately reflected the trend of error when population means are equal (see Appendix B). When compared to the theoretical estimate of power that should be obtained from the given conditions, there is an increasingly larger departure from expected results as both disparities between sample size and standard deviation increase. This increased error seems to be exacerbated when the group with the smaller sample size has the greater population standard deviation ($n_1,\ n_2\ = 12, 28$; $\sigma_1,\ \sigma_2 = 1, .75$). Notably, these departures from expected rates of significant results are small (approximately than three percentage points); thus, the additional risk taken in interpreting the results of similar conditions is limited.

In the condition where there is a small to medium difference ($d = .25$ when population standard deviations are equal) in population means, results vary little from the previous two population mean conditions (see Appendix C). This finding implies a greater degree of resilience to heterogenous sampling conditions than what was originally expected (Wang, 1971) of the pooled samples *t*-test.

When the population means differed by a moderate amount ($d = .50$ when population standard deviations are equal), the trends in error rates are more pronounced (see Appendix D). Similar to the condition where there was a small-to-medium difference in sample means, the rate of significant result approximated expected rates as long as sample sizes or population standard deviations were equal. Furthermore, as long as population standard deviations were even somewhat close (i.e., the smaller standard deviation was at least 75% of the larger, 56% in terms of variance), the departure from expected rates of significance was less than two percentage points.

The condition that led to the greatest deviation from theoretical power was a fairly extreme condition where population standard deviations were not close in any sense ($\sigma_1,\ \sigma_2 = 1$, .25, a 4:1 standard deviation ratio and an 8:1 variance ratio) and sample size disparities were large ($n_1,\ n_2 = 12, 28$). The departure from expected significance rates was a decrease of approximately five percentage points in the proportion of obtained significant findings (.6320 vs. .6843). This specific combination was the only scenario of those investigated where a departure of this magnitude was observed. Furthermore, previous conditions involving smaller mean differences never approached deviations as pronounced as this. These results imply that errors of this magnitude are only present when there are substantial differences between the parameters of two populations coupled with a large sample size disparity.

Discussion

The trends in deviations of observed results from expected results were not entirely anticipated when beginning data collection. Although previous research has concluded that both Type I and Type II error rates can be inflated when the pooled samples *t*-test is used while in violation of the assumption of homogeneity of variance (Derrick et al., 2016), the manner in which this trend manifested across a spectrum of sampling conditions was unexpected; error rates even when population standard deviations differed by almost 50% were much smaller than expected.

This pattern of results observed can be expected to increase in strength as the difference in sample means increases. Despite this, the degree to which Student's *t*-test's departs from expected power remains fairly consistent until population mean differences are moderately sized. However, concerns with statistical power likely become less relevant when dealing with large

differences in sample means holding other factors constant. Therefore, these conditions should be more resilient to increasingly heterogenous sampling conditions.

Recommendations for Researchers

In deciding when it is acceptable to use the pooled samples *t*-test, results indicate that any differences between the groups results in some degree of additional error. However, these errors are small unless the population means are moderately different ($d = .50$). When population means are equal or only slightly different ($d < .25$), a researcher can expect to have accurate conclusions for results when either sample sizes or population standard deviations are equal. Furthermore, comparable results can be found when the larger sample sizes are up to 50% greater than the smaller (e.g., $n_1,\ n_2 = 16, 24$) as long as the smaller population standard deviation is at least 90% of the greater (81% in terms of variance). Within these ranges, the observed rate of significant results deviates less than one percentage point from the expected rate. These results provide a solid basis for drawing conclusions from the results of the pooled samples *t*-test.

For situations where there is a moderate difference between population means, one should take a more conservative approach to relying on the pooled samples *t*-test. Increases in the frequency of obtained significant results are much more substantial than those observed in the conditions where mean differences were smaller. In the small population mean difference condition (see Appendix B) when sample sizes are held constant and the difference in population standard deviations is made larger, there is a one percentage point increase in obtaining significant results. When comparing these same conditions to two populations with a slightly larger mean difference (see Appendix C), this deviation from expected rates of significance is approximately two percentage points.

However, a researcher may not have to worry about losing statistical power, as only in extreme conditions where the population standard deviations differences are very large (i.e., the larger standard deviation is four times greater than the smaller, eight times greater in terms of variance) and sample sizes are heavily weighted to favor the population with the smaller standard deviation ($n_1 = 12, n_2 = 28$) will the proportion of obtained significant results substantially depart from the expected estimate of power. Any conditions less extreme than these will result in a proportion of significant results well within five percentage points of the expected power.

In practice, researchers do not know population mean or standard deviation differences. The only relevant factor that is known are the sample sizes. Thus, recommendations for practice must crafted around available information. In the event where there are slight disparities between the sample sizes (e.g., $n_1 = 18, n_2 = 22$), the error presented in these conditions is minimal (i.e., less than a two percentage point difference from expected rates of significance) as long as the smaller population variability is remotely close to the larger (i.e., at least half the standard deviation or a quarter of the variance), a condition likely true in almost every scenario. Even larger sample size disparities (i.e., 16 vs. 24) will yield results that differ from expected values by less than 3.5 percentage points as long as the population variability is within the same limits.

Setting the recommendation to these ranges provides a simple standard for decision making when evaluating whether an individual should instead rely on Welch's approximate *t*-test (e.g., If the sample sizes are not close, is there any reason to believe that one group has a population variance four times greater than the other?). In deciding whether to run one test versus the other, referencing the ranges set above can simplify the analysis for research projects by eliminating the need to choose between the results of two tests.

The field of statistics in psychology can present a number of challenges to individuals conducting scientific research. Given the test's status and popularity of use, it is imperative to set an easily observable benchmark for researchers to reference. By further defining the conditions in which it is acceptable to rely on the results of Student's *t*-test, this study offers clarity to an otherwise arbitrary standard for decision-making. Although arguments have been made that it is simpler to default to Welch's *t*-test (Delacre et al., 2017), researchers need not discard the validity of results obtained via Student's *t*-test in less-than-ideal conditions.

References

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*(1), 49-64. https://psycnet.apa.org/doi/10.1037/h0041412

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student;s *t*-test. *International Review of Social Psychology, 30*(1), 92-101. https://doi.org/10.5334/irsp.82

Derrick, B., Toher, D., & White, R. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology, 12*(1), 30-38. https://doi.org/10.20982/tqmp.12.1.p030

Harrison, R. L. (2010). Introduction to Monte Carlo simulation. In: AIP Conference Proceedings, *1204*, 17–21. https://doi.org/10.1063/1.3295638

Kim, S., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics, 23*(4), 356-377. https://doi.org/10.3102/10769986023004356

Lee, A. F. S. (1992). Optimal sample sizes determined by two-sample Welch's *t*-test. *Communications in Statistics – Simulation and Computation, 21*(3), 689-696. https://doi.org/10.1080/03610919208813043

Lu, Z. L., & Yuan, K. (2010). Welch's *t*-test. In N. J. Salkind (Ed.) *Encyclopedia of Research Design* (1st ed., pp. 1620-1623). Thousand Oaks. https://doi.orgg/10.13140/rg.2.1.3057.9607

Mooney, C. Z. (1997). *Monte Carlo Simulation*. SAGE Publications, Inc.,

https://dx.doi.org/10.4135/9781412985116

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the *t*-test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360. https://psycnet.apa.org/doi/10.1037/0033-2909.111.2.352

Scheffe, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association, 65*(332), 1501-1508. https://doi.org/10.2307/2284332

Student (1908). The probable error of a mean. *Biometrika, 6*(1), 1-25. https://doi.org/10.1093/biomet/6.1.1

Wang, Y. Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem. *Journal of the American Statistical Association, 66*(335), 605-608. https://doi.org/10.2307/2283538

Welch, B. L. (1937). The Significance of the Difference Between Two Means When the Population Variances Are Unequal. *Biometrika, 29*(3/4), 350-362. https://doi.org/10.2307/2332010

Zimmerman, D. W., & Zimbo, B. D. (2009). Hazards in choosing between pooled and separate-variances *t*-tests. *Psicologica, 30*(2), 371-390. https://www.redalyc.org/comocitar.oa?id=16911991012

Appendix A

Proportion of statistically significant results: Equal population means ($\mu_1 = \mu_2 = 0$).

Group 2 standard deviation (σ_2)	Sample sizes: $n_1,\ n_2$								
	12, 28	14, 26	16, 24	18, 22	20, 20	22, 18	24, 16	26, 14	28, 12
1.0	.0502	.0503	.0507	.0505	.0501	.0494	.0498	.0499	.0500
.95	.0551	.0526	.0528	.0508	.0496	.0500	.0477	.0457	.0449
.90	.0592	.0580	.0542	.0537	.0503	.0476	.0447	.0432	.0412
.85	.0661	.0608	.0584	.0536	.0496	.0471	.0434	.0398	.0378
.80	.0733	.0674	.0618	.0550	.0498	.0456	.0413	.0374	.0330
.75	.0786	.0737	.0647	.0568	.0495	.0444	.0381	.0336	.0289
.50	.1270	.1036	.0854	.0671	.0523	.0392	.0302	.0205	.0137
.25	.1888	.1459	.1090	.0811	.0558	.0361	.0215	.0126	.0066

Note: $\alpha = .05$, $\sigma_1 = 1.0$, number of independent samples per condition = 100,000.

Appendix B

Proportion of statistically significant results: Unequal population means (μ_1 = 0.10, μ_2 = 0.00) and theoretical power.

Group 2 standard deviation (σ_2)	Sample sizes: n_1, n_2								
	12, 28	14, 26	16, 24	18, 22	20, 20	22, 18	24, 16	26, 14	28, 12
1.0	.0592	.0607	.0598	.0605	.0616	.0592	.0597	.0597	.0596
	(.0585)	(.0593)	(.0598)	(.0601)	(.0602)	(.0601)	(.0598)	(.0593)	(.0585)
.95	.0638	.0647	.0628	.0624	.0630	.0596	.0586	.0566	.0553
	(.0592)	(.0599)	(.0604)	(.0607)	(.0607)	(.0605)	(.0602)	(.0596)	(.0588)
.90	.0713	.0680	.0670	.0647	.0612	.0581	.0574	.0538	.0508
	(.0599)	(.0606)	(.0611)	(.0613)	(.0613)	(.0610)	(.0606)	(.0599)	(.0590)
.85	.0787	.0741	.0717	.0680	.0620	.0599	.0540	.0508	.0462
	(.0607)	(.0614)	(.0618)	(.0619)	(.0618)	(.0615)	(.0610)	(.0602)	(.0593)
.80	.0843	.0809	.0750	.0692	.0640	.0579	.0522	.0483	.0433
	(.0615)	(.0620)	(.0620)	(.0626)	(.0624)	(.0620)	(.0614)	(.0606)	(.0595)
.75	.0940	.0871	.0793	.0727	.0662	.0557	.0521	.0454	.0395
	(.0624)	(.0630)	(.0633)	(.0633)	(.0631)	(.0626)	(.0618)	(.0609)	(.0598)

Note: α = .05, σ_1 =1.0, number of iterations for each condition = 100,000, parenthetical values are estimates of power.

Appendix C

Proportion of statistically significant results for the unequal population means condition ($\mu_1 = 0.25$, $\mu_2 = 0.00$) and theoretical power.

Group 2 standard deviation (σ_2)	Sample sizes: $n_1,\ n_2$								
	12, 28	14, 26	16, 24	18, 22	20, 20	22, 18	24, 16	26, 14	28, 12
1.0	.1097	.1148	.1177	.1191	.1211	.1193	.1168	.1148	.1093
	(.1053)	(.1101)	(.1135)	(.1156)	(.1163)	(.1156)	(.1135)	(.1101)	(.1053)
.95	.1173	.1227	.1225	.1257	.1248	.1211	.1202	.1120	.1044
	(.1096)	(.1144)	(.1177)	(.1195)	(.1198)	(.1187)	(.1162)	(.1122)	(.1070)
.90	.1315	.1306	.1341	.1317	.1290	.1221	.1196	.1099	.1021
	(.1143)	(.1190)	(.1221)	(.1236)	(.1235)	(.1219)	(.1189)	(.1144)	(.1086)
.85	.1394	.1421	.1384	.1364	.1332	.1285	.1197	.1101	.1001
	(.1195)	(.1241)	(.1269)	(.1280)	(.1274)	(.1253)	(.1217)	(.1166)	(.1103)
.80	.1542	.1528	.1486	.1422	.1381	.1278	.1186	.1091	.0946
	(.1252)	(.1296)	(.1321)	(.1326)	(.1315)	(.1287)	(.1245)	(.1189)	(.1120)
.75	.1670	.1625	.1578	.1529	.1400	.1311	.1193	.1045	.0925
	(.1315)	(.1356)	(.1375)	(.1375)	(.1357)	(.1323)	(.1274)	(.1211)	(.1136)

Note: $\alpha = .05$, $\sigma_1 = 1.0$, number of iterations for each condition = 100,000. , parenthetical values are estimates of power.

Appendix D

Proportion of statistically significant results for the unequal population means condition ($\mu_1 =$ 0.50, $\mu_2 =$0.00) and theoretical power.

Group 2 standard deviation(σ_2)	Sample sizes: n_1, n_2								
	12, 28	14, 26	16, 24	18, 22	20, 20	22, 18	24, 16	26, 14	28, 12
1.0	.2918	.3120	.3278	.3372	.3375	.3348	.3270	.3136	.2909
	(.2849)	(.3049)	(.3192)	(.3277)	(.3305)	(.3277)	(.3192)	(.3049)	(.2849)
.95	.3146	.3314	.3429	.3502	.3532	.3474	.3351	.3178	.2960
	(.3028)	(.3227)	(.3362)	(.3436)	(.3449)	(.3404)	(.3301)	(.3139)	(.2919)
.90	.3344	.3553	.3625	.3693	.3672	.3597	.3446	.3268	.2976
	(.3223)	(.3418)	(.3543)	(.3602)	(.3599)	(.3535)	(.3412)	(.3229)	(.2989)
.85	.3640	.3738	.3850	.3863	.3797	.3744	.3560	.3303	.3026
	(.3435)	(.3622)	(.3734)	(.3777)	(.3754)	(.3669)	(.3524)	(.3320)	(.3059)
.80	.3858	.4005	.4049	.4055	.4021	.3861	.3675	.3397	.3076
	(.3665)	(.3840)	(.3936)	(.3959)	(.3914)	(.3806)	(.3638)	(.3411)	(.3128)
.75	.4088	.4224	.4263	.4211	.4142	.4004	.3743	.3472	.3086
	(.3913)	(.4072)	(.4147)	(.4147)	(.4078)	(.3945)	(.3753)	(.3502)	(.3197)
.50	.5343	.5370	.5287	.5181	.4972	.4715	.4293	.3812	.3264
	(.5381)	(.5376)	(.5286)	(.5126)	(.4905)	(.4628)	(.4301)	(.3927)	(.3510)
.25	.6320	.6264	.6142	.5921	.5667	.5241	.4792	.4127	.3411
	(.6843)	(.6587)	(.6288)	(.5950)	(.5573)	(.5161)	(.4714)	(.4237)	(.3732)

Note: $\alpha = .05$, $\sigma_1 =$1.0, number of iterations for each condition = 100,000. , parenthetical values are estimates of power.

Copyright Permission

Name: Nelson, Hayden Peter

Email (to receive future readership statistics): haydenpnelson@gmail.com

Type of document: ['Thesis']

Title: AN INVESTIGATION OF THE EFFECTIVENESS OF STUDENT'S t-TEST UNDER HETEROGENEITY OF VARIANCE

Keywords (3-5 keywords not included in the title that uniquely describe content): Monte Carlo, Pooled samples t-test, Heterogeneity, Standard deviation, Sample size

Committee Chair: Dr. Reagan Brown

Additional Committee Members: Dr. Katrina Burch Dr. Xiaowen Chen

Select 3-5 TopSCHOLAR® disciplines for indexing your research topic in TopSCHOLAR®: Social and Behavioral Sciences Psychology Social Statistics Industrial-Organizational Psychology Quantitative Psychology