Open Peer Review on Qeios

Intersections of Statistical and Substantive Significance Under a True and False Null Hypothesis

Eugene Komaroff[1]

1 Keiser University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

This paper has two objectives: (1) Provide intuitive insight into statistical and substantive significance intersections with histograms, bar graphs, and crosstabs with data from independent samples t-tests. (2) Convincingly demonstrate with graphs and a few numbers that statistically significant p-values from independent samples t-tests are valuable for screening out standardized mean differences, known as Cohen's d (effect size), that would otherwise be misinterpreted as substantively significant. The author hopes the empirical sampling distributions in this paper help students, applied researchers, and science writers to properly understand and appreciate the value of statistical significance for scientific inference and decision-making with small sample sizes (n < 1,000) in the face of uncertainty.

Eugene Komaroff[a,*]

Keiser University Graduate School

[a]ORCID iD: 0000-0003-2659-7821

[*]Correspondence can be addressed to Eugene Komaroff, Keiser University Graduate School, 1900 W. Commercial Blvd., Fort Lauderdale, FL 33309, but preferably by email to ekomaroff@keiseruniversity.edu

Keywords: True Null Hypothesis, False Null Hypothesis, T-test, Cohen's d, Effect Size, P-values, Statistical Significance.

Massive data sets in this 21st century may have thousands of variables with complex, multidimensional structures and millions of observations requiring computer-intensive data science for their statistical/mathematical analysis. Nonetheless, small data sets (n < 1,000) also exist, and their analysis requires more accessible inferential statistical techniques developed at the beginning of the 20th century (e.g., Student, 1908). These analyses are now done with statistical software on laptops that swiftly produces many p-values for evaluating statistical significance. Regrettably, the misunderstanding and abuse of p-values and statistical significance have diminished scientific credibility (Wasserstein & Lazar, 2016; Greenland et al., 2016; Ioannidis, 2005).

The editors of Basic and Applied Social Psychology (Trafimow & Marks, 2015) highlighted the common misinterpretation of a non-statistically significant p-value as the probability that the null hypothesis is true. As a result, they banned statistical significance, p-values, and confidence intervals from articles submitted for publication. A year later, the American Statistical Association (ASA) published six principles about p-values to guide their appropriate use and interpretation (Wasserstein & Lazar, 2016). Nevertheless, three years later, an editorial in a special issue of The American Statistician (TAS) called “Moving to a world beyond $p < 0.05$” stated that “no p-value can reveal the plausibility, presence, truth, or importance of an association or effect….regardless of whether it was ever useful, a declaration of ‘statistical significance’ has today become meaningless.” The editors called for a ban on statistical significance by proclaiming, “Don’t say it, Don’t use it” (Wasserstein et al., 2019, p. 2).

Coining a new expression for statistical significance makes sense because it has been conflated with substantive significance. However, no consensus exists for “don’t use it” (e.g., Begg, 2020; Benjamini, 2021; Harrington et al., 2019; Mayo & Hand, 2022; Komaroff, 2020; Komaroff, 2024). This paper is focused on sampling distributions of random p-values. Others have also considered p-value distributions (e.g., Bland, 2013; Verykouki & Nakas, 2023; Wang et al., 2019). This author agrees that the credibility of statistical significance is undermined by violated assumptions, poor execution of research designs, multiple testing, data dredging or fishing, p-hacking, and p-harking (hypothesis after the results are known). See Greenland et al. (2019) and Wasserstein and Lazar (2016) for a comprehensive discussion about misinterpretation and abuse of p-values. This paper focuses only on the effect of random sampling errors on sampling distributions of means, mean differences, and p-values related to statistical significance, as well as the substantive significance of effect sizes as measured with Cohen’s d (1968).

Theoretical sampling distributions are foundational for understanding statistical significance. However, this concept is taught with complex mathematical/statistical theorems that are too difficult for students and applied researchers who are not mathematicians. The concept is simplified by showing a sparse histogram depicting repeated summary statistics randomly sampled from a human population overlayed with a smooth, bell-shaped, standard normal curve (e.g., Moore et al., 2021). However, such simplification does not provide any insight into statistical significance. Sampling distributions are not about people but are theoretical distributions with infinitely many random, summary statistics that follow a predictable form.

Student (1908) stated that some experiments cannot be easily repeated; therefore, the certainty of a result must be judged with a tiny sample. Furthermore: ”Any experiment may be regarded as forming an individual of a ‘population’ of experiments which might be performed under the same conditions (Student, 1908, p. 1). Fisher (1970) echoed: “The entire result of an extensive experiment may be regarded as but one of a possible population of such experiments” (Fisher, 1970, p. 2). “Population” in these sentences exists only in mathematical, statistical theory. Students and applied researchers easily grasp the concept of sample and population frequency distributions because they exist in practice (reality). However, theoretical sampling distributions exist only in theory. The author hopes the empirical sampling distributions in this paper help students, applied researchers, and science writers to properly understand and appreciate the value of p-values and statistical significance for scientific inference and decision-making in the face of uncertainty.

This paper has two objectives: (1) Provide intuitive insight into the intersection of statistical and substantive significance with histograms, bar graphs, and crosstabs of data from independent samples t-tests. (2) Convincingly demonstrate with graphs and a few numbers that statistical significance is an essential scientific tool for screening out standardized mean differences known as Cohen's d (effect size) that would otherwise be misinterpreted as substantively significant.

Methodology

Imagine a small, early-phase, randomized controlled trial (RCT) designed to determine the efficacy of a novel or experimental treatment (T) to cause a desired outcome (endpoint) compared to a control treatment (C) with a placebo or standard of care. An independent samples t-test determined whether the difference in means was statistically significant. The numerator of the t-test is the difference between the sample means ($\bar{x}_T - \bar{x}_C$) subtracted from the difference between two population means ($\mu_T - \mu_C$). A typical null hypothesis is that the population means are equivalent, or stated another way, the difference in population means is zero ($H_0: \mu_T - \mu_C = 0$). If the sample means are accurate estimates of the population means, the difference in the sample means must also be zero. However, sample means differ in statistical theory because of random sampling errors. Nevertheless, when statistical significance is detected, the null hypothesis of zero difference in population means is rejected. That eliminates the need to consider random sampling errors as a cause of the difference in sample means. The conclusion is that the experimental treatment will be effective in the target population exposed to the experimental intervention.

Statistical Significance

Empirical sampling distributions were simulated with SAS OnDemand for Academics statistical software (SAS, 2014). The process started with a matrix of two random variables (X) sampled from the standard normal distribution [X ~ N (0,1)]. Summary descriptive statistics: sample size, means, standard deviations, and p-values from the "equal variance assumption" of the independent samples t-test (SAS, 2019) were saved into an analysis data set. The true null hypothesis stated that the difference in population means was zero $\left(H_0: \mu_T - \mu_C = 0\right)$. The process was replicated 1,000 times and repeated with four equal sample sizes per group (n = 15, 64, 500, 1000), creating four analysis data sets.

The level of statistical significance was 5% (α =.05). An indicator (binary) variable was coded as "1" if the p-value was statistically significant; otherwise, it was "0." The percentage of statistically significant p-values out of 1,000 was an empirical estimate of the theoretical type 1 error rate of 5% under a true null hypothesis. Please recognize that the null hypothesis of zero difference in the population means was not merely assumed to be true. It was known to be true because all the independent and identically distributed random variables were sampled from a standard normal population (mu = 0, sigma = 1).

An additional data set with a medium effect size of 0.50 was added to each observation in the experimental treatment condition to demonstrate testing a false null hypothesis. The results revealed the probability of a statistically significance p-value when the null hypothesis is false: $\mu_T - \mu_C = 0$ and the alternative hypothesis is true: $\mu_T - \mu_C = 0.50$. In statistical literature, this is called a power analysis, where the alternative hypothesis is assumed to be true.

Substantive Significance

Cohen's d (effect size) was computed as the difference between two sample means divided by a pooled standard deviation. Cohen's (1968) effect size categories were used to evaluate the substantive significance of a standardized difference in means: small $|d| \geq 0.20$ to 0.49, medium $|d| \geq 0.50$ to.0.79, large $|d| \geq 0.80$. To compute the percentages of substantively significant effect sizes (small + medium + large), an indicator (binary) variable was coded "1" if Cohen's $|d| \geq 0.20$. When Cohen's $|d| < 0.20$, the indicator variable was coded as "0" and labeled as "none-effect size." Just as all statistically significant p-values were false (type 1 errors) under the true null hypothesis, all substantive effect sizes were false (effect size errors) because the population effect size (Cohen's D) was known to be equal to 0.00 under the true null hypothesis.

Results

N = 15 Per Group

Figure 1 is the empirical sampling distribution of 1000 random means from the experimental treatment with n = 15 per group.

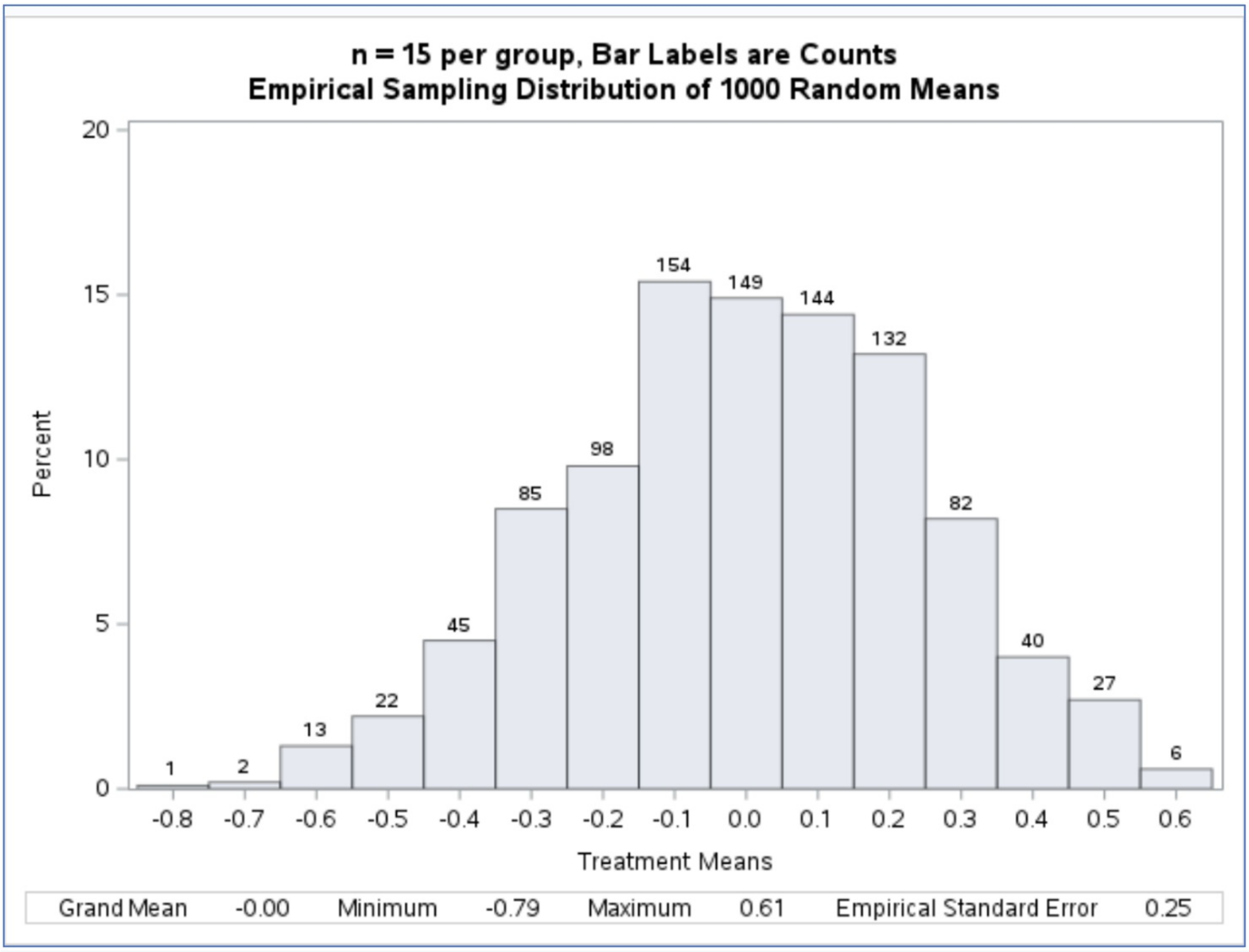


Figure 1.

The empirical distribution of treatment means is approximately symmetric as predicted by the Central Limit Theorem and centered at the population mean (mu) zero as predicted by the Law of Large Numbers.

Figure 2 shows the sample distribution of 1000 random means from the control treatment with n = 15 per group.

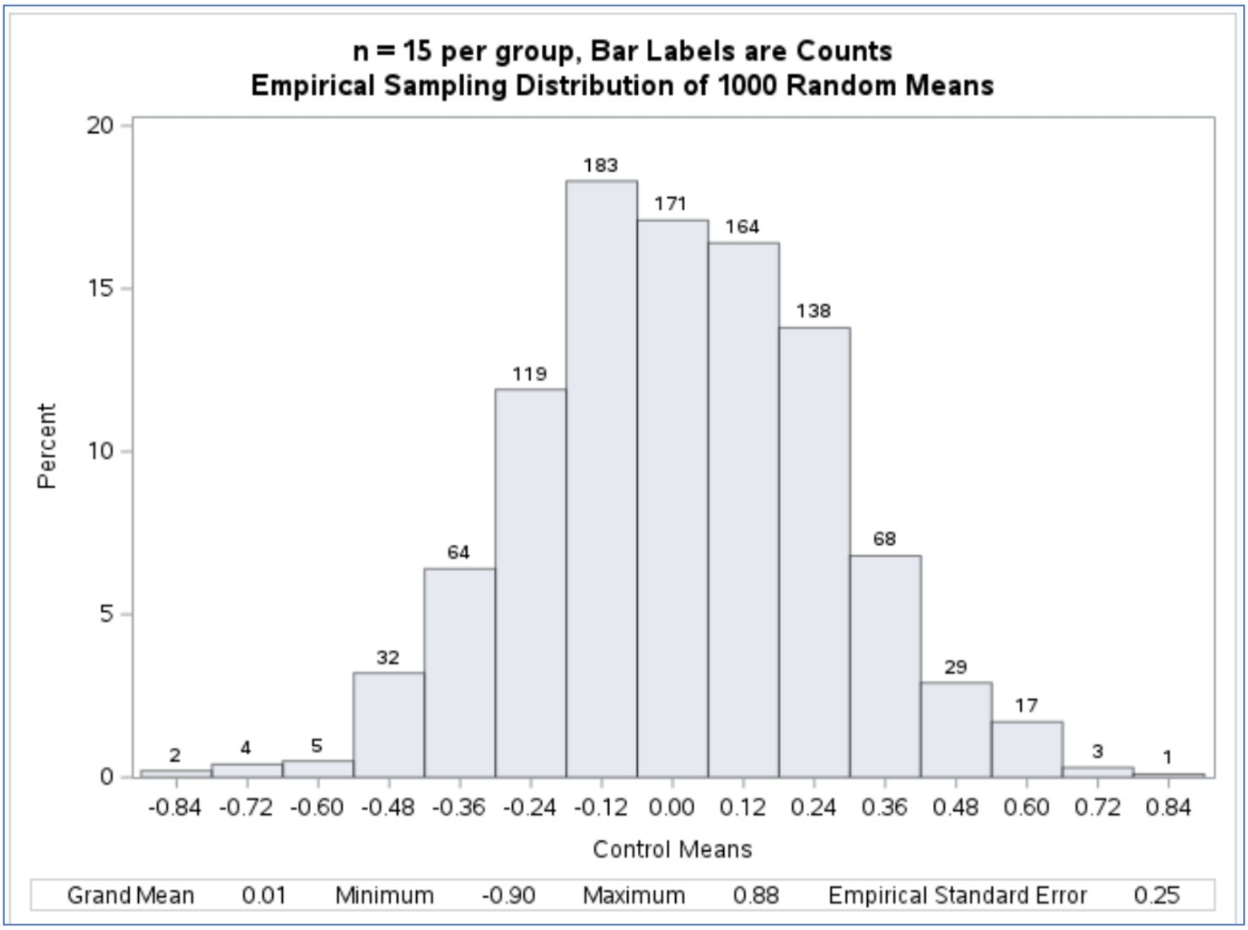


Figure 2.

The empirical sampling distribution of control means also appears symmetric and is centered at approximately the population mean (mu) = 0.00.

Figure 3 is the empirical sampling distribution of differences in the means. The T and C means were indexed by each simulation run from 1 to 1,000. The first T mean was subtracted from the first C mean, then the following T mean was subtracted from the next C mean, and so forth. This process created 1,000 mean differences.

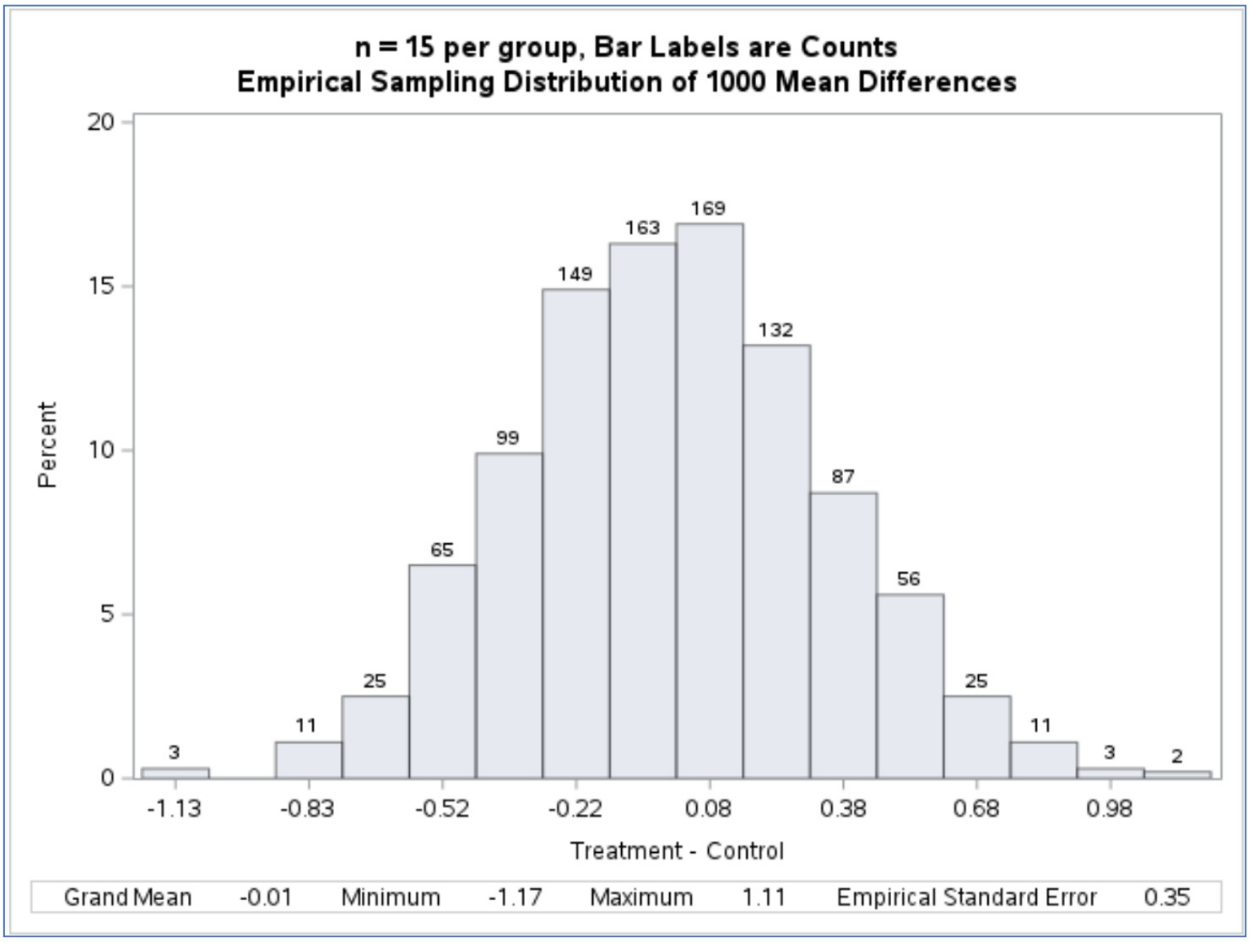


Figure 3.

The empirical sampling distribution of the mean differences is also approximately symmetric and is centered at zero.

Figure 4 shows the p-values from 1,000 independent samples t-tests.

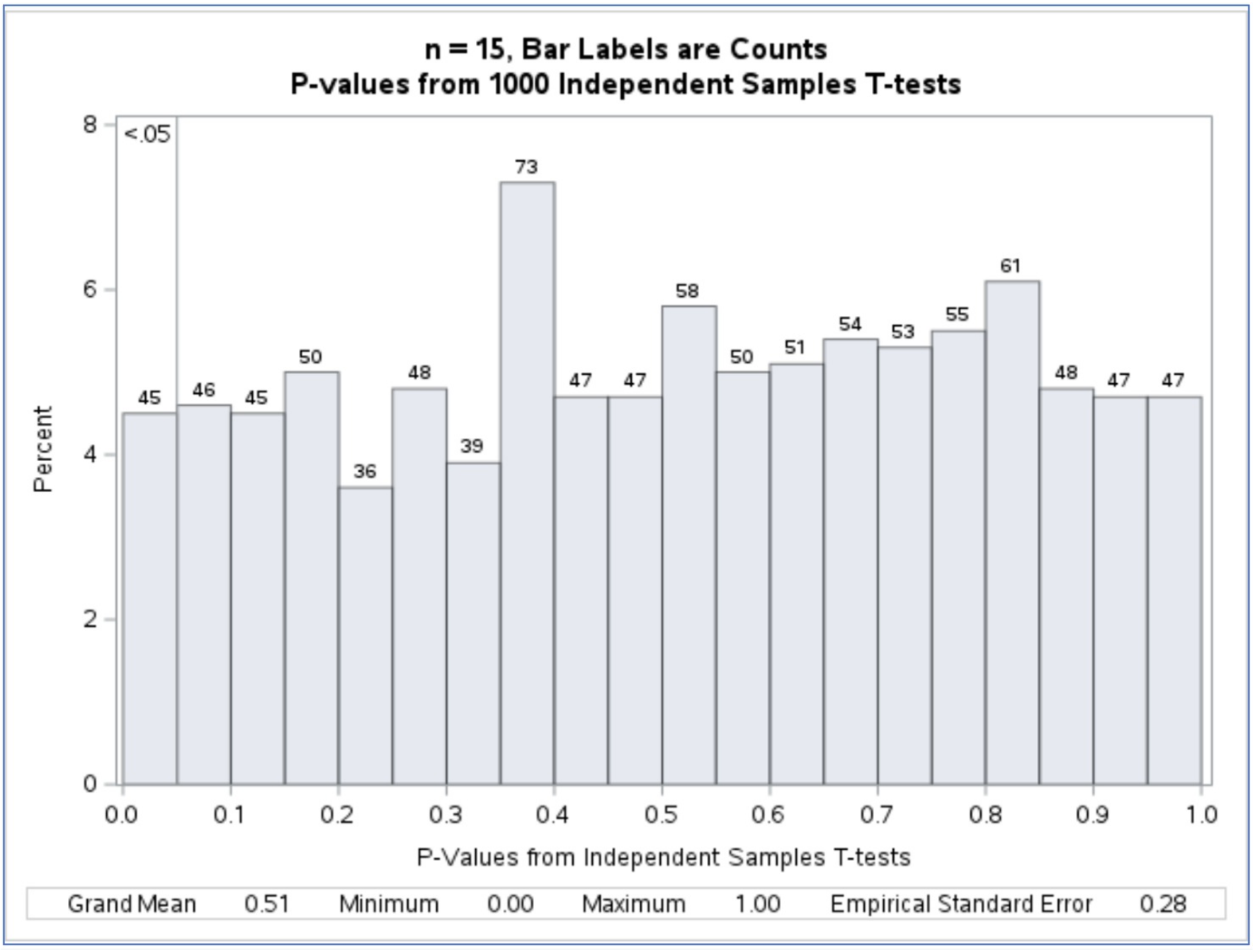


Figure 4.

The empirical distribution of p-values is close to the uniform distribution that is predicted by statistical theory when all assumptions are satisfied (Bland, 2103; Westfall et al., 2011; Wang et al., 2019). The 45 p-values to the left of the reference line at.05 are statistically significant.

Figure 5 is an empirical sampling distribution of Cohen’s d as a continuous variable.

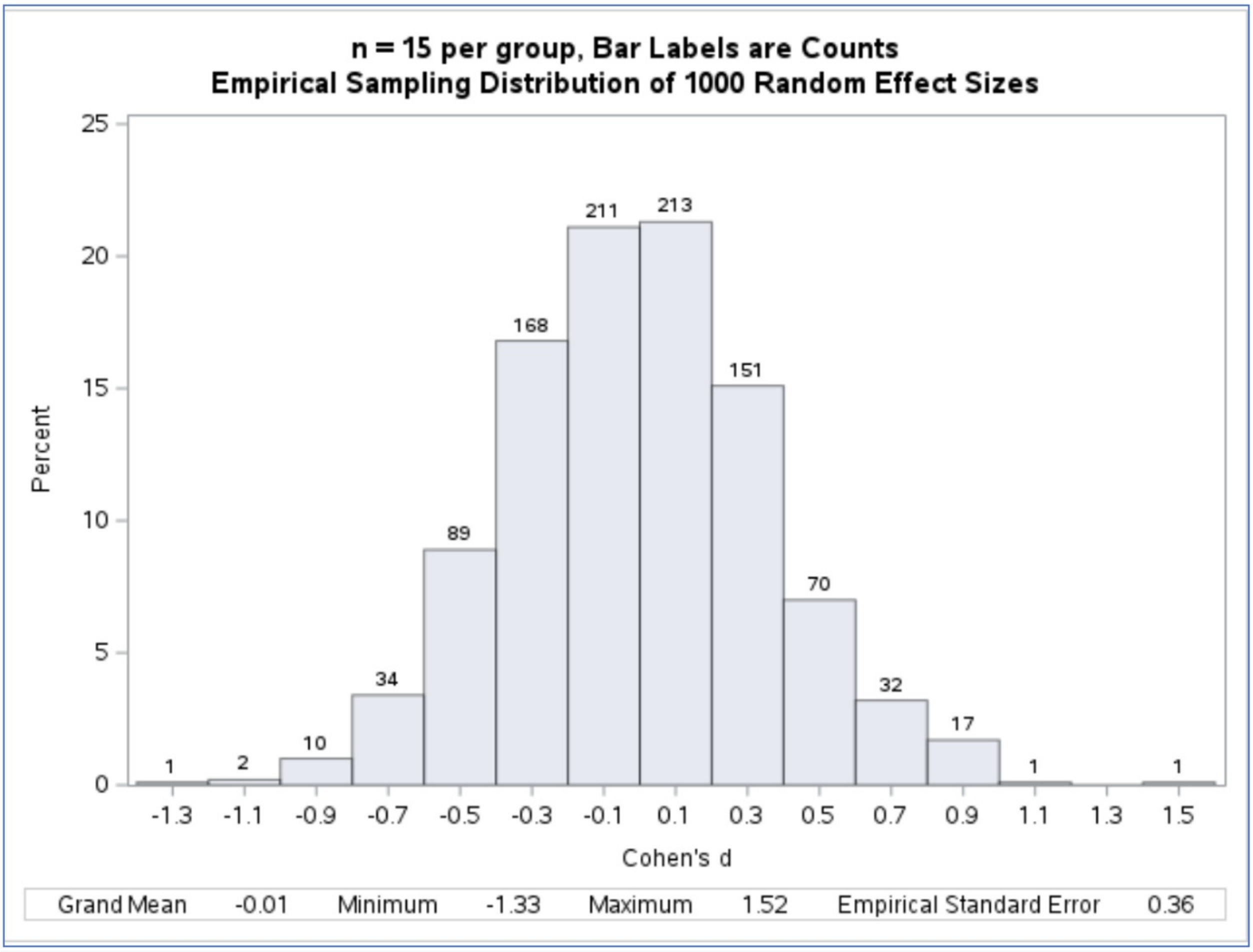


Figure 5.

This empirical sampling distribution of Cohen’s d as a continuous variable appears symmetric and is centered close to zero.

Figure 6 displays bar graphs of effect sizes according to Cohen’s d categories: small |d| (≥ 0.20 to 0.49), medium |d| ≥ 0.50 to.0.79, large |d| ≥ 0.80.

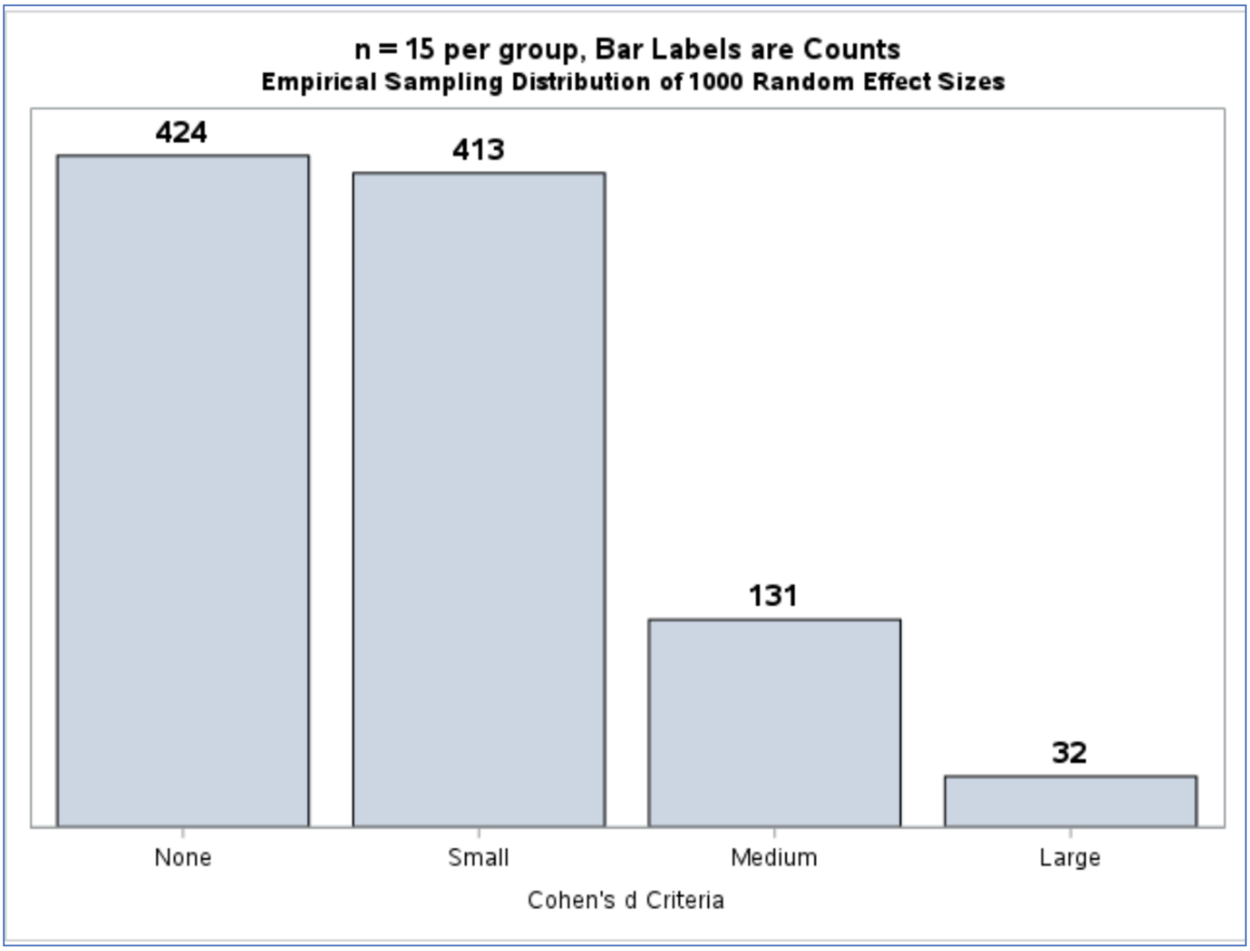


Figure 6.

Although almost half are in the no-effect size category, small, medium, and large effect sizes are evident even though the population effect size is zero (Cohen's D = 0.00).

Figure 7 shows the dichotomous distribution of effect sizes.

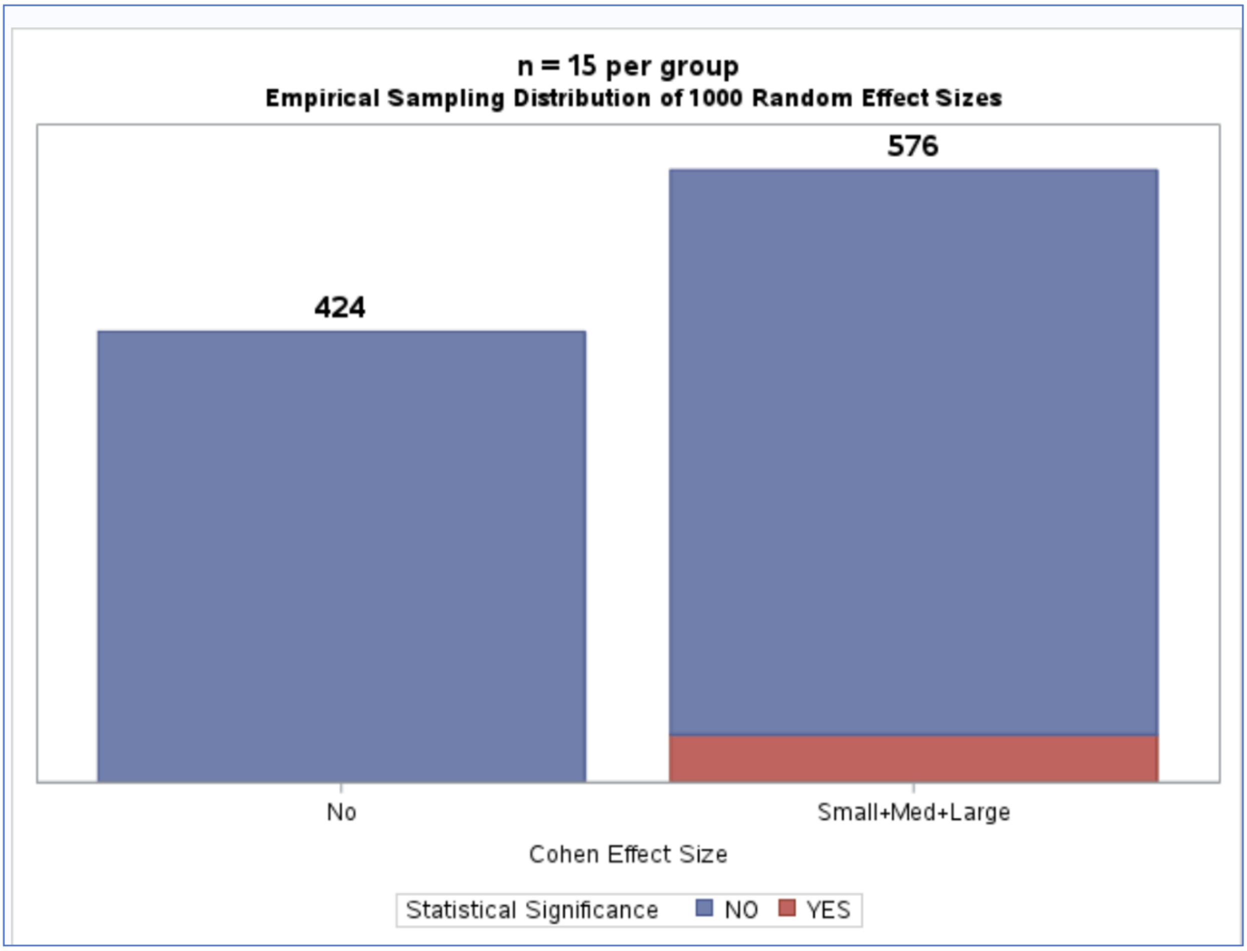
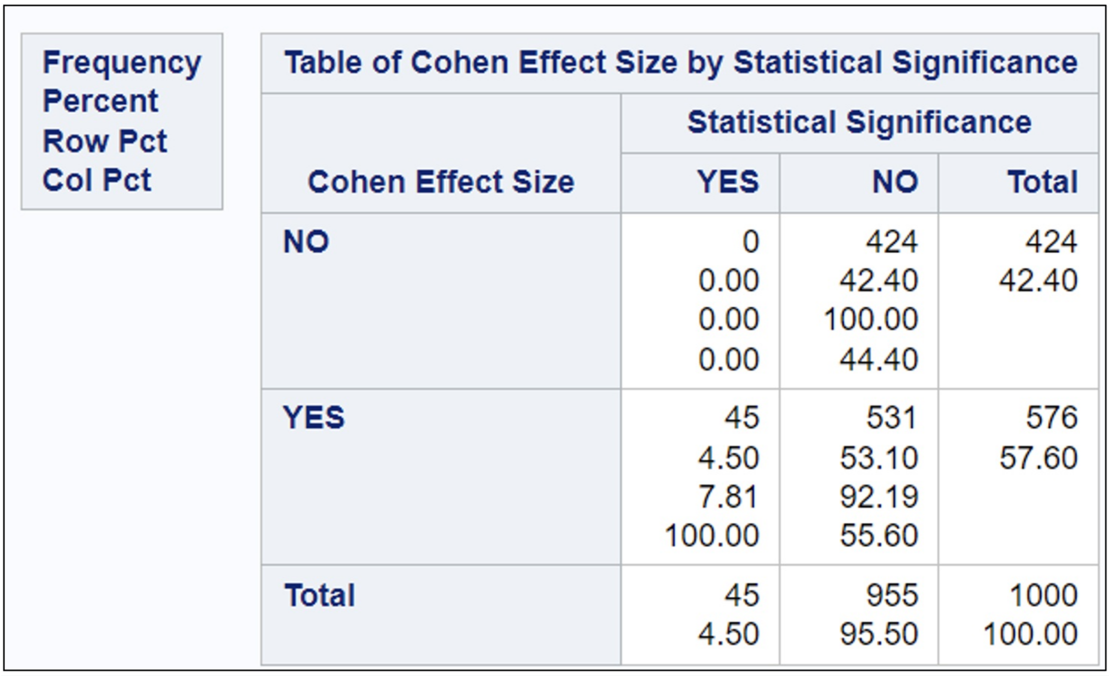


Figure 7.

Only a few effect sizes that fit Cohen's (1968) criteria were statistically significant.

Table 1 has the counts and percentages of statistical and substantive significance. The lower row total reveals that approximately 5% (45) were statistically significant. Of these 45, 100% corresponded to Cohen’s d substantively significant categories. Consequently, instead of a researcher contemplating/interpreting 576 (58%) effect sizes, statistical significance markedly reduced the count to 45 (8%), which would be misinterpreted as substantively significant. The population of Cohens D = 0.00. Therefore, all effect sizes > 0.00 were effect sizes caused by random sampling errors.



Frequency
Percent
Row Pct
Col Pct

Table of Cohen Effect Size by Statistical Significance

	Statistical Significance		
Cohen Effect Size	YES	NO	Total
NO	0 0.00 0.00 0.00	424 42.40 100.00 44.40	424 42.40
YES	45 4.50 7.81 100.00	531 53.10 92.19 55.60	576 57.60
Total	45 4.50	955 95.50	1000 100.00

Table 1.

Figure 8 overlays the statistically significant effect sizes (red) on the distribution of Cohen's d as a continuous variable.

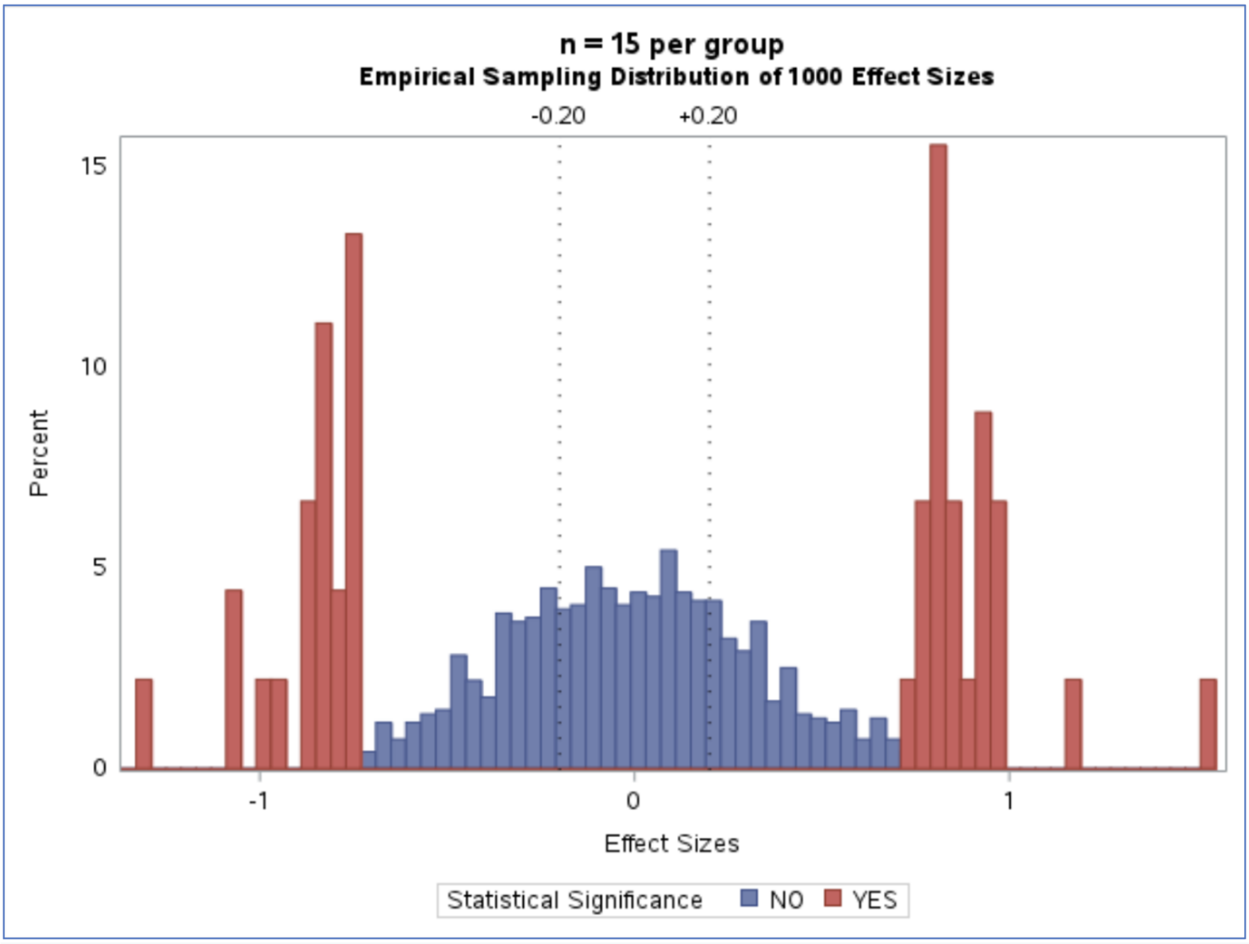


Figure 8.

The reference lines at -0.20 and +0.20 indicate the smallest effect sizes that would be considered substantively significant by Cohen's (1968) criteria. Nonetheless, it is important to note that statistical significance was detected for medium/large effect sizes (|d| > 0.70), which is reassuring. If a type 1 error was made, at least the effect size was worthy of consideration.

N = 64 Per Group

Figure 9 shows the empirical sampling distribution of continuous Cohen's d with n = 64 per group.

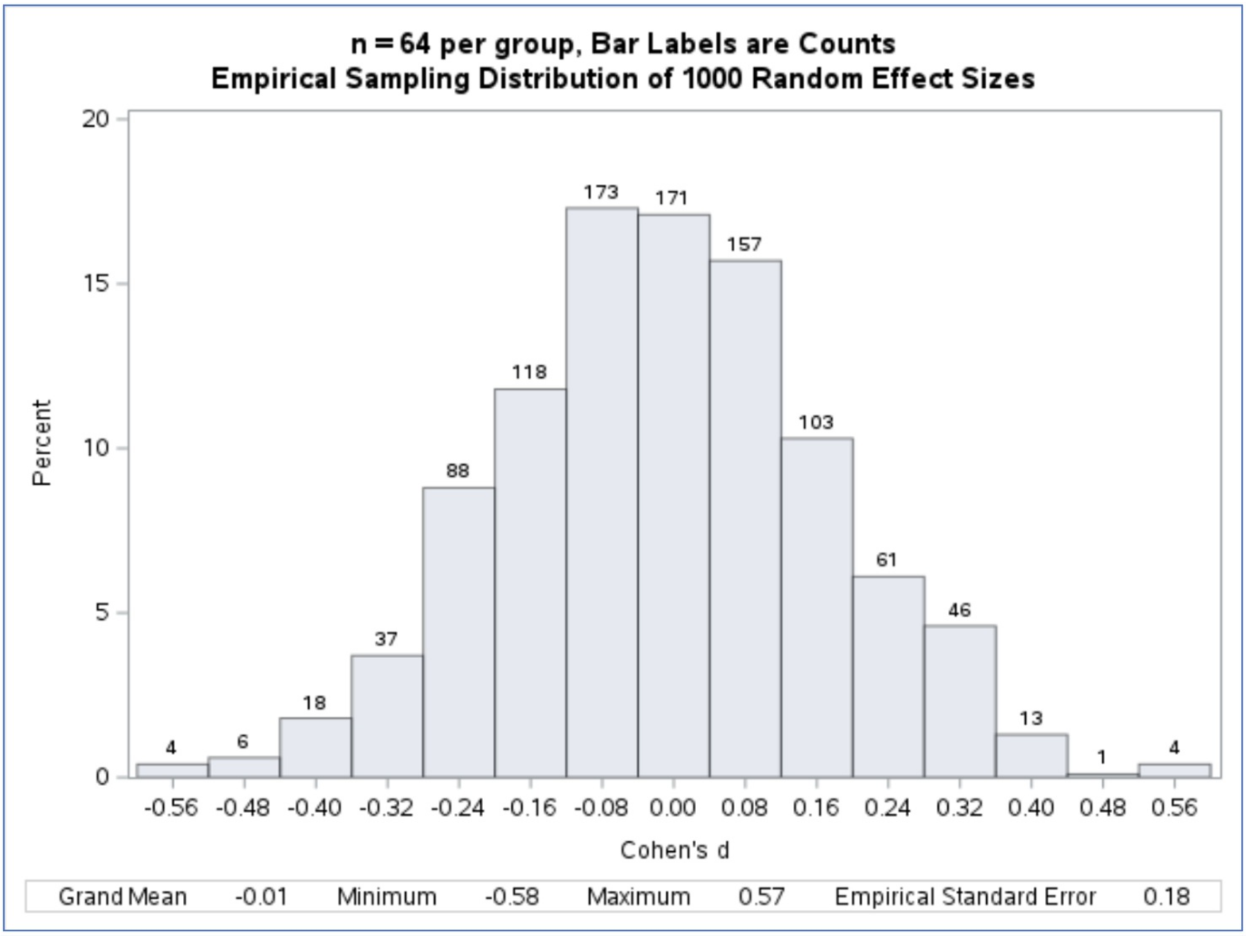


Figure 9.

The empirical sampling distribution of Cohen's d appears symmetric with a central value close to zero consistent with the known population Cohen's D = 0.00.

Figure 10 displays the empirical sampling distribution of p-values with n = 64 per group.

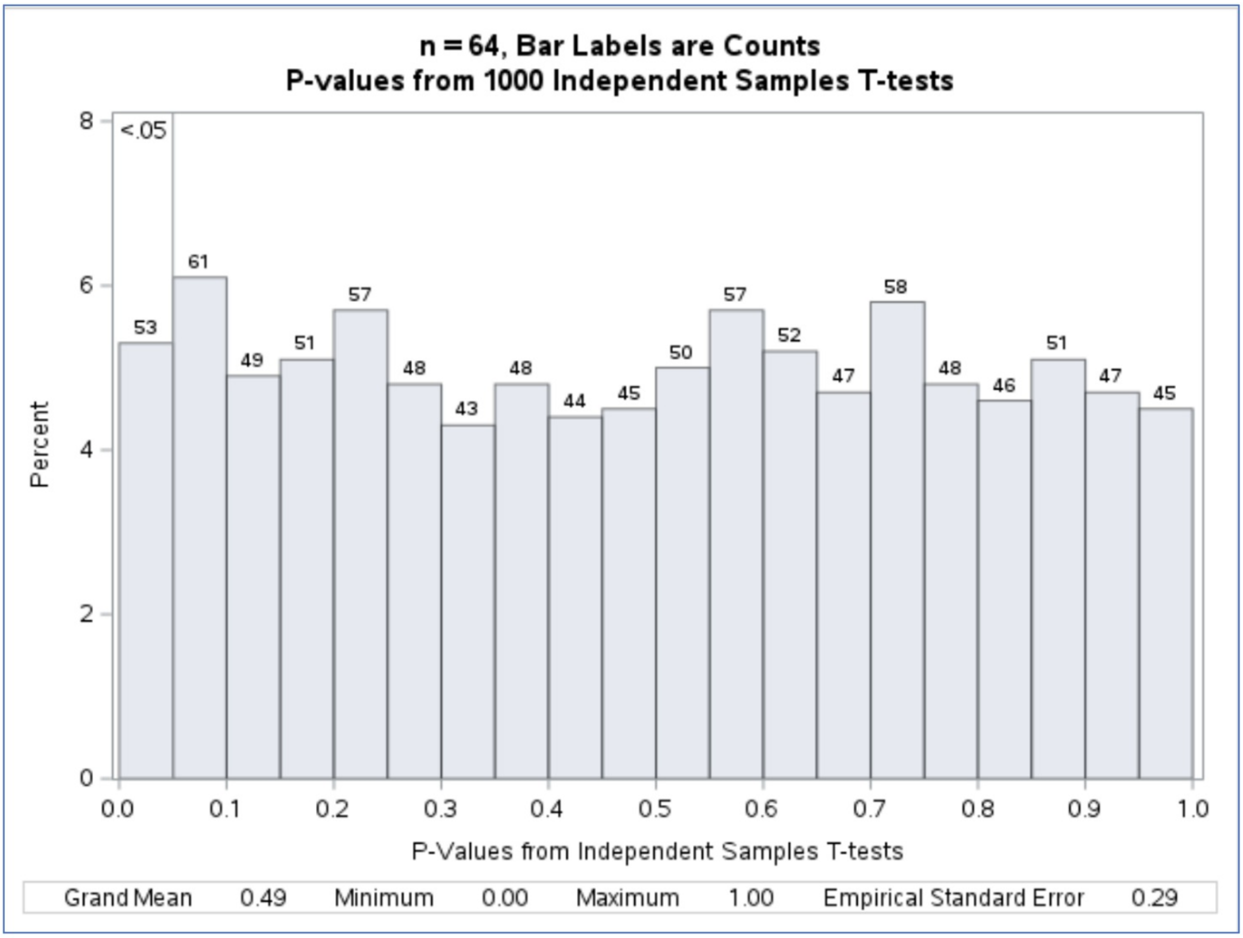


Figure 10.

Figure 10 is an approximately uniform distribution with 53 statistically significant p-values from 1000 t-tests.

Figure 11 indicates no large effect sizes under the true null hypothesis.

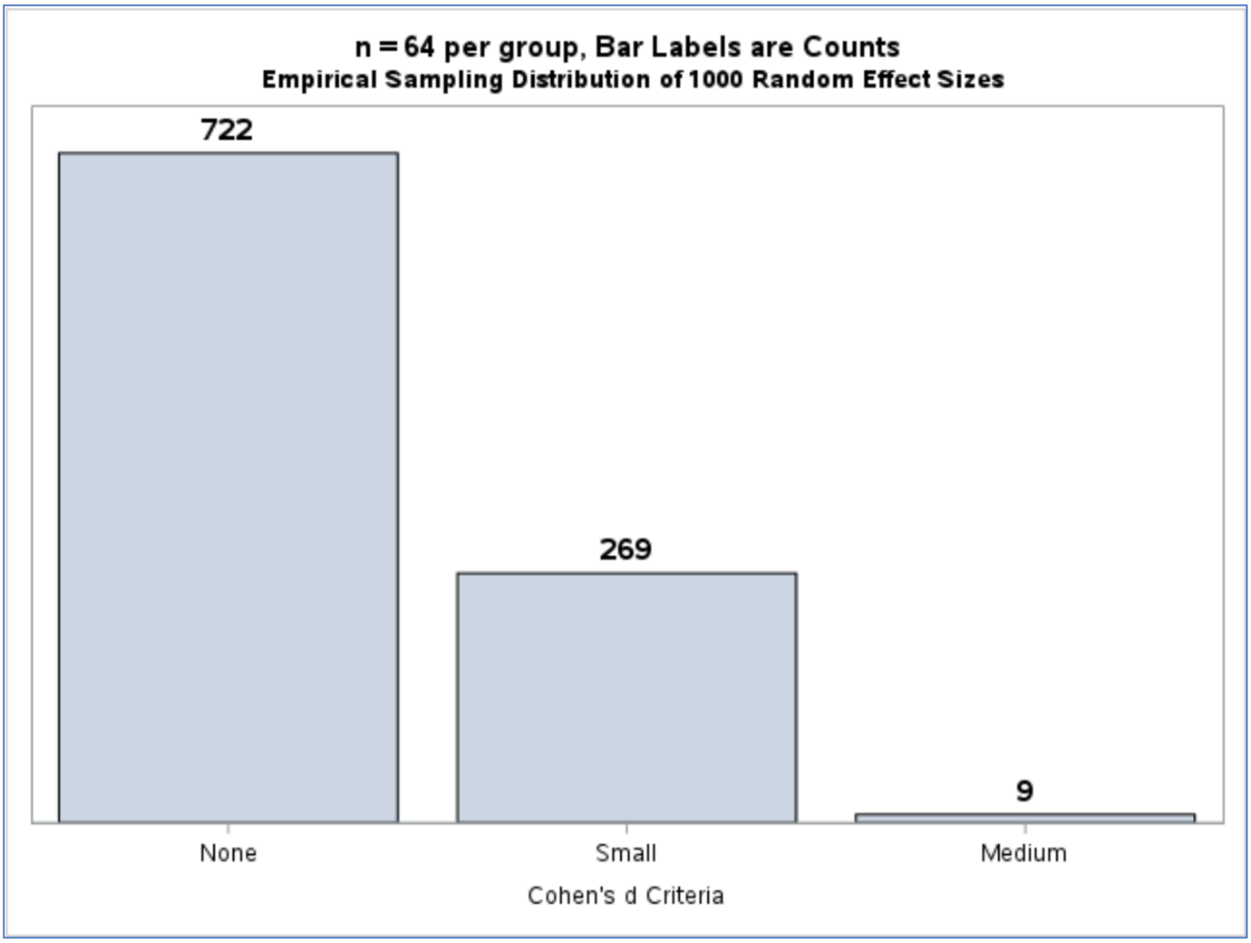


Figure 11.

Figure 12 reveals that statistically significant p-values were effect sizes that fit Cohen’s d substantive significance criteria.

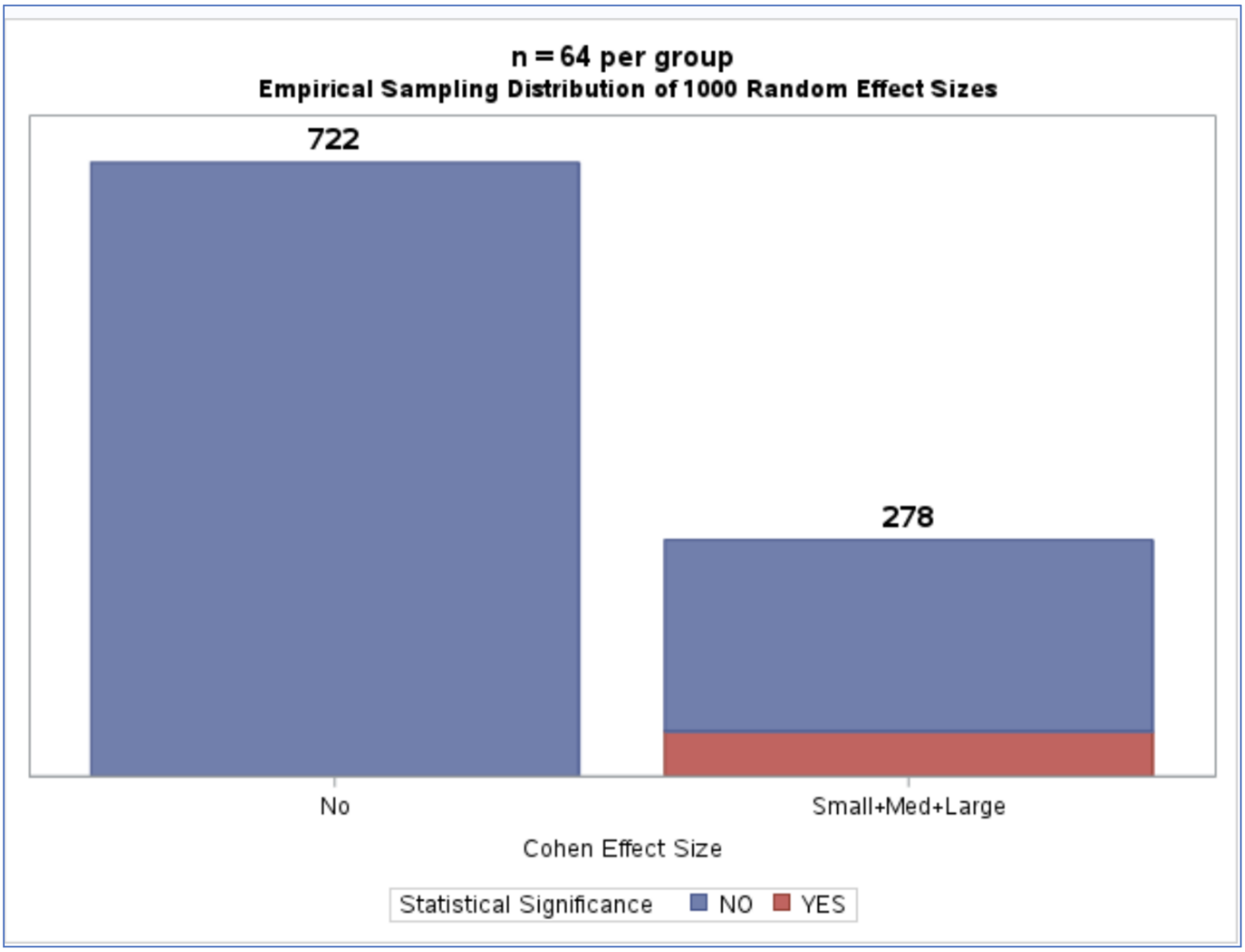
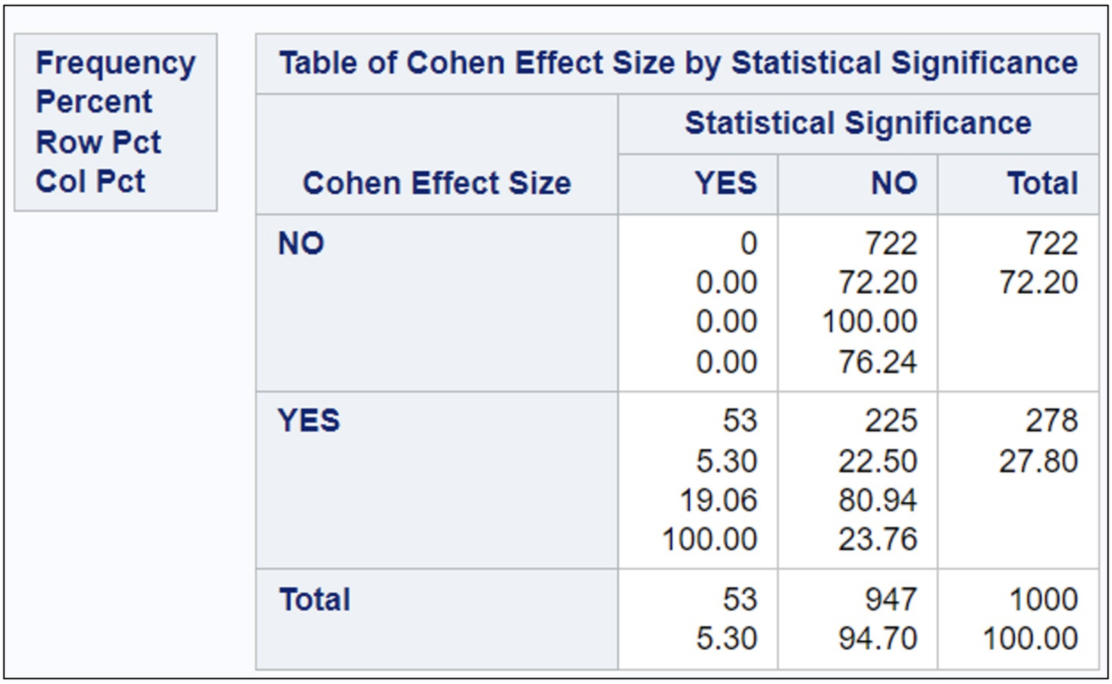


Figure 12.

Table 2 has the counts and percentages at the intersection of statistical significance and substantive significance.



Frequency
Percent
Row Pct
Col Pct

Table of Cohen Effect Size by Statistical Significance

Cohen Effect Size	Statistical Significance		
	YES	NO	Total
NO	0 0.00 0.00 0.00	722 72.20 100.00 76.24	722 72.20
YES	53 5.30 19.06 100.00	225 22.50 80.94 23.76	278 27.80
Total	53 5.30	947 94.70	1000 100.00

Table 2.

The marginal distributions in Table 2 reveal approximately 5% (53) statistically significant p-values and 28% (278) substantively significant effect sizes. Statistical significance screened out 81% of the effect size errors. It is again noteworthy that 100% (53) of the statistically significant p-values corresponded to Cohen's substantive effect sizes criteria of small, medium, or large.

N = 500 Per Group

Figure 13 shows the empirical sampling distribution of Cohen's d as a continuous variable with n = 500 per group. The range of Cohen's d is -0.18 (none effect size) to 0.21 (barely small effect size).

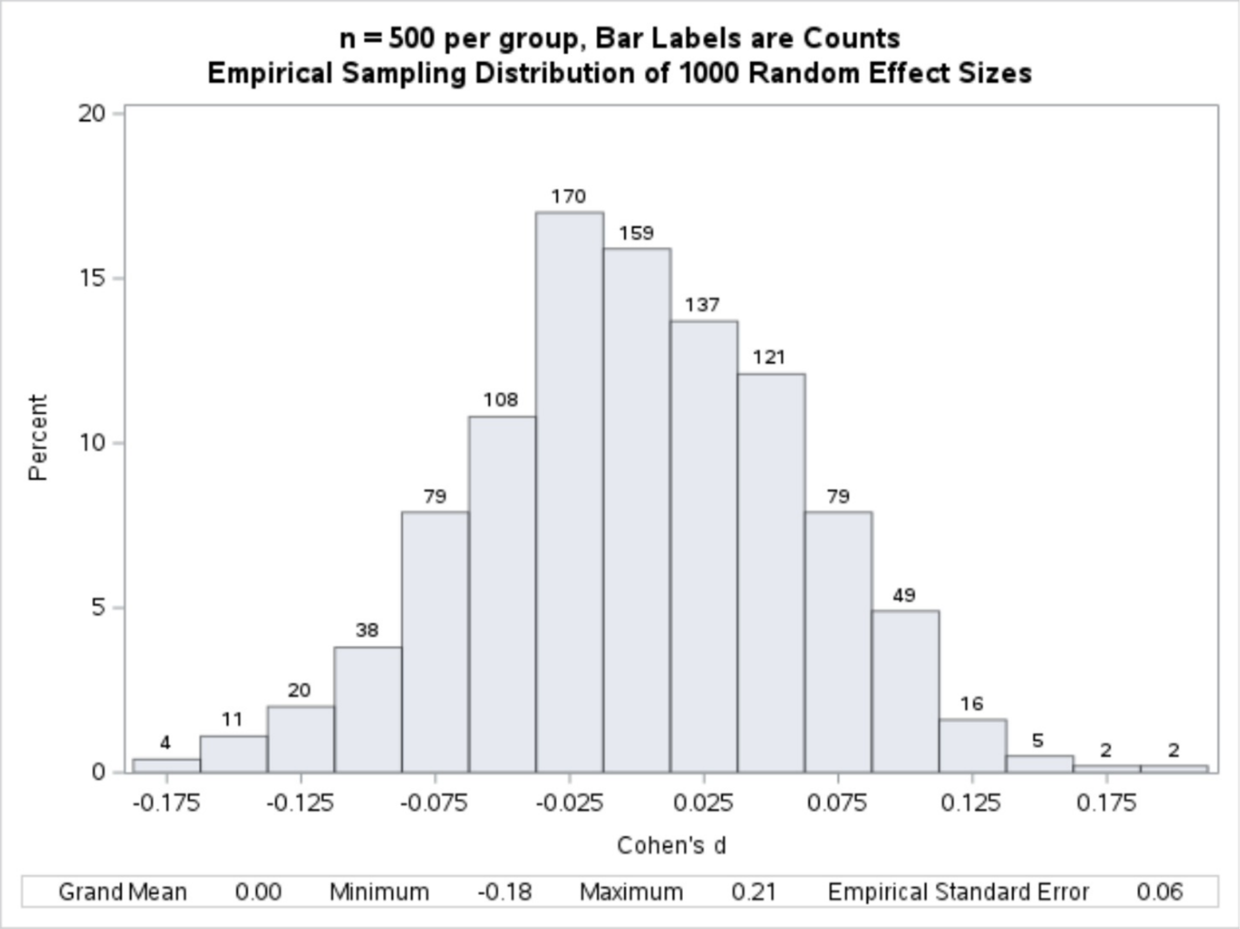


Figure 13.

Figure 14 reveals an approximately uniform distribution of p-values where 40 were statistically significant (p <.05).

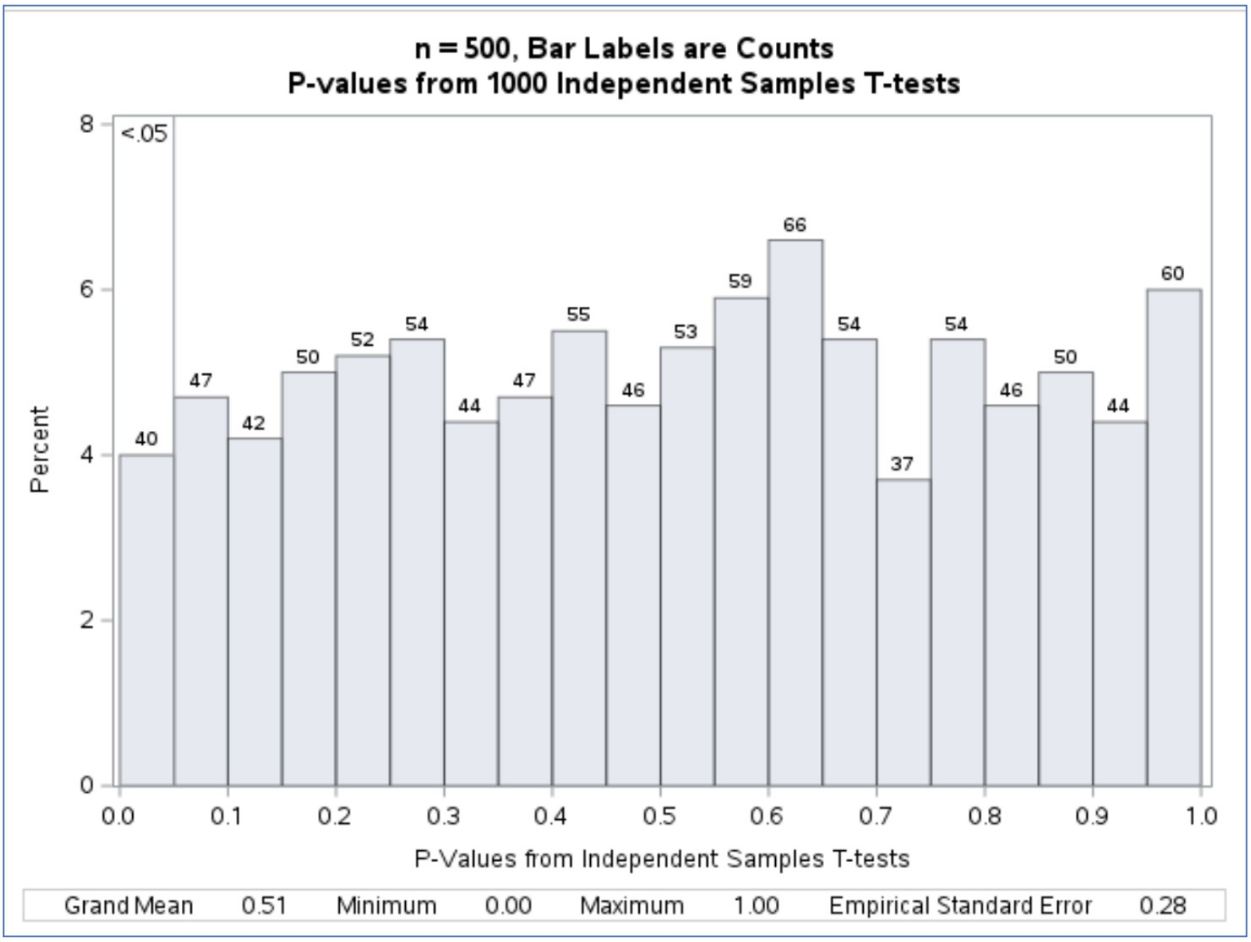


Figure 14.

Figure 15 revealed at least one statistically significant effect size, but with the total sample size = 1,000, a few non-effect sizes were detected as statistically significant.

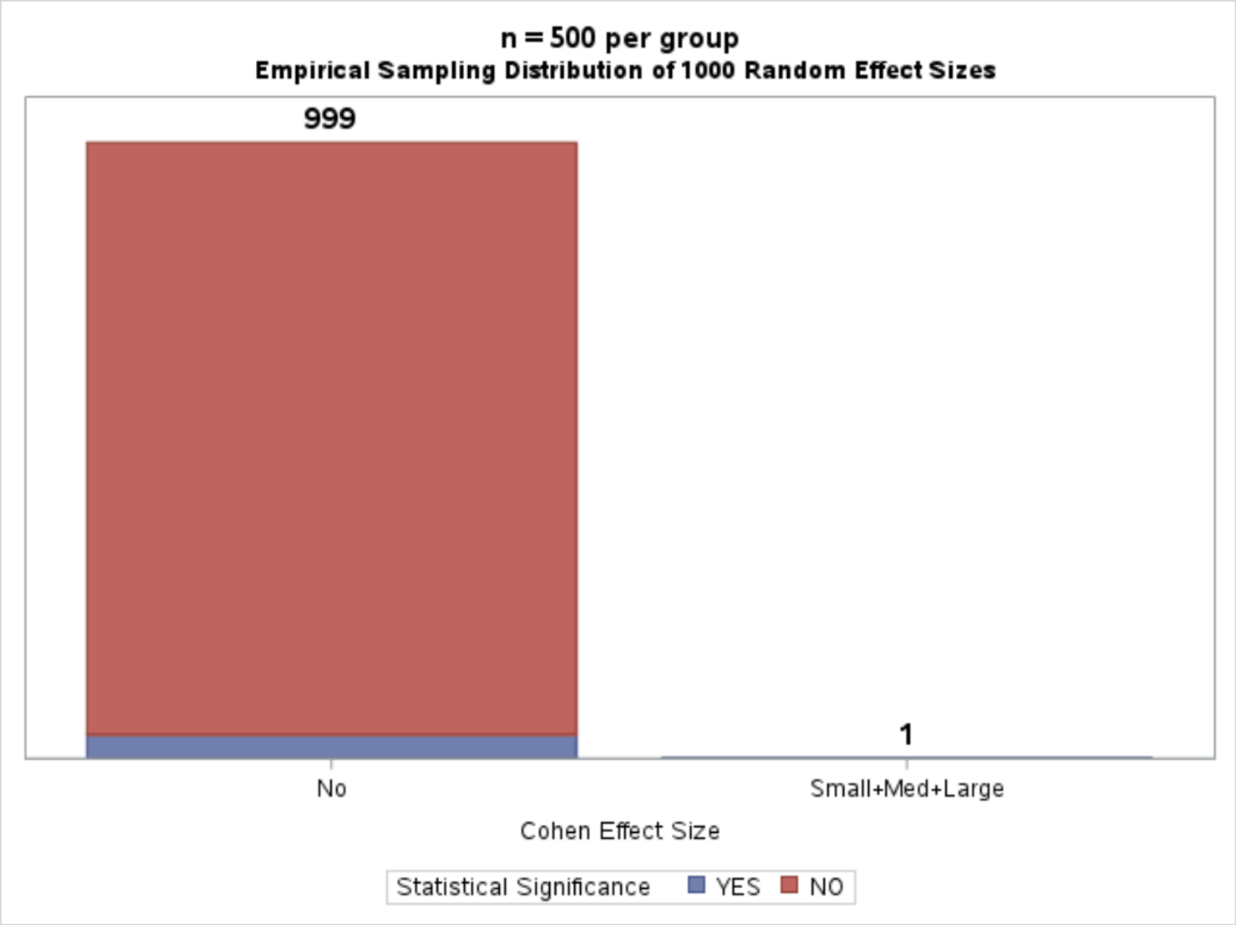
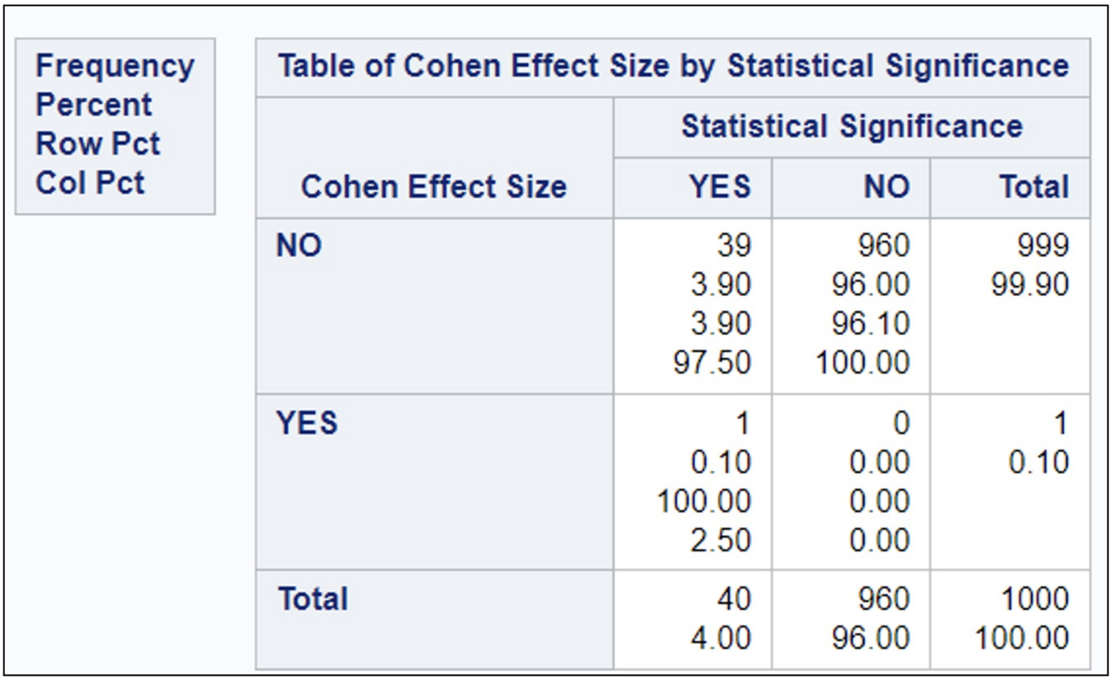


Figure 15.

Table 3 provides the counts and percentages of the intersection of statistical significance with substantive significance with n = 500 per group. The marginal distribution of statistical significance reveals that 4% (40) p-values were statistically significant. Approximately 2% (1) corresponded to an effect size fitting Cohen's criteria, whereas the remaining 98% (39) were statistically significant non-effect sizes.



Frequency
Percent
Row Pct
Col Pct

Table of Cohen Effect Size by Statistical Significance

	Statistical Significance		
Cohen Effect Size	YES	NO	Total
NO	39 3.90 3.90 97.50	960 96.00 96.10 100.00	999 99.90
YES	1 0.10 100.00 2.50	0 0.00 0.00 0.00	1 0.10
Total	40 4.00	960 96.00	1000 100.00

Statistical Significance Under a False Null Hypothesis with N = 64 Per Group

A medium effect size (0.50) was added to each of the 64 observations in the treatment condition, thereby the null hypothesis was false, H_0: $(\bar{x}_T - \bar{x}_C) - (\mu_T - \mu_C) = 0$

Figure 16 shows the empirical sampling distribution of the means from the Treatment condition.

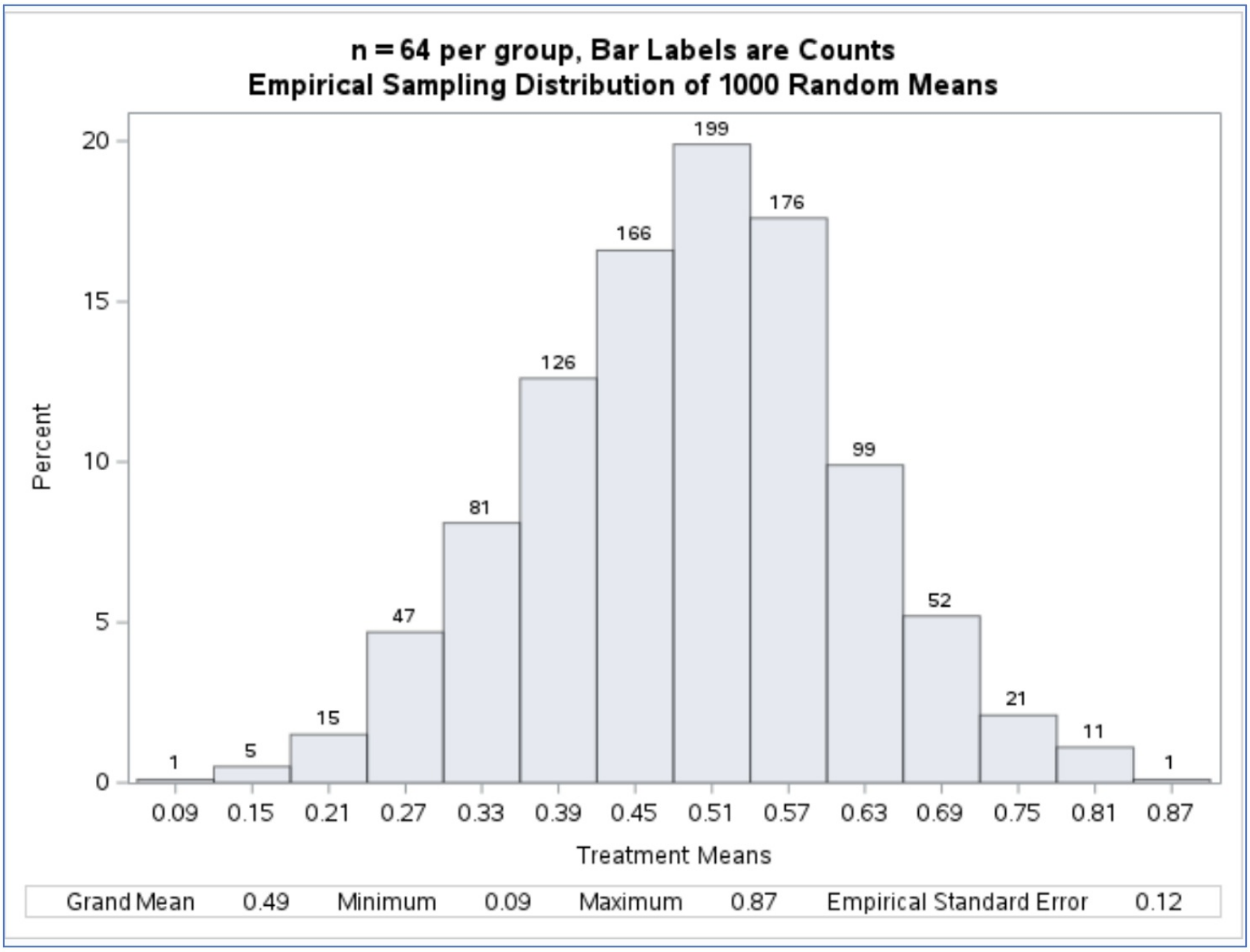


Figure 16.

The distribution of treatment means is approximately symmetric but is centered at 0.49, consistent with the induced effect size.

Figure 17 shows the empirical sampling distribution of the control means.

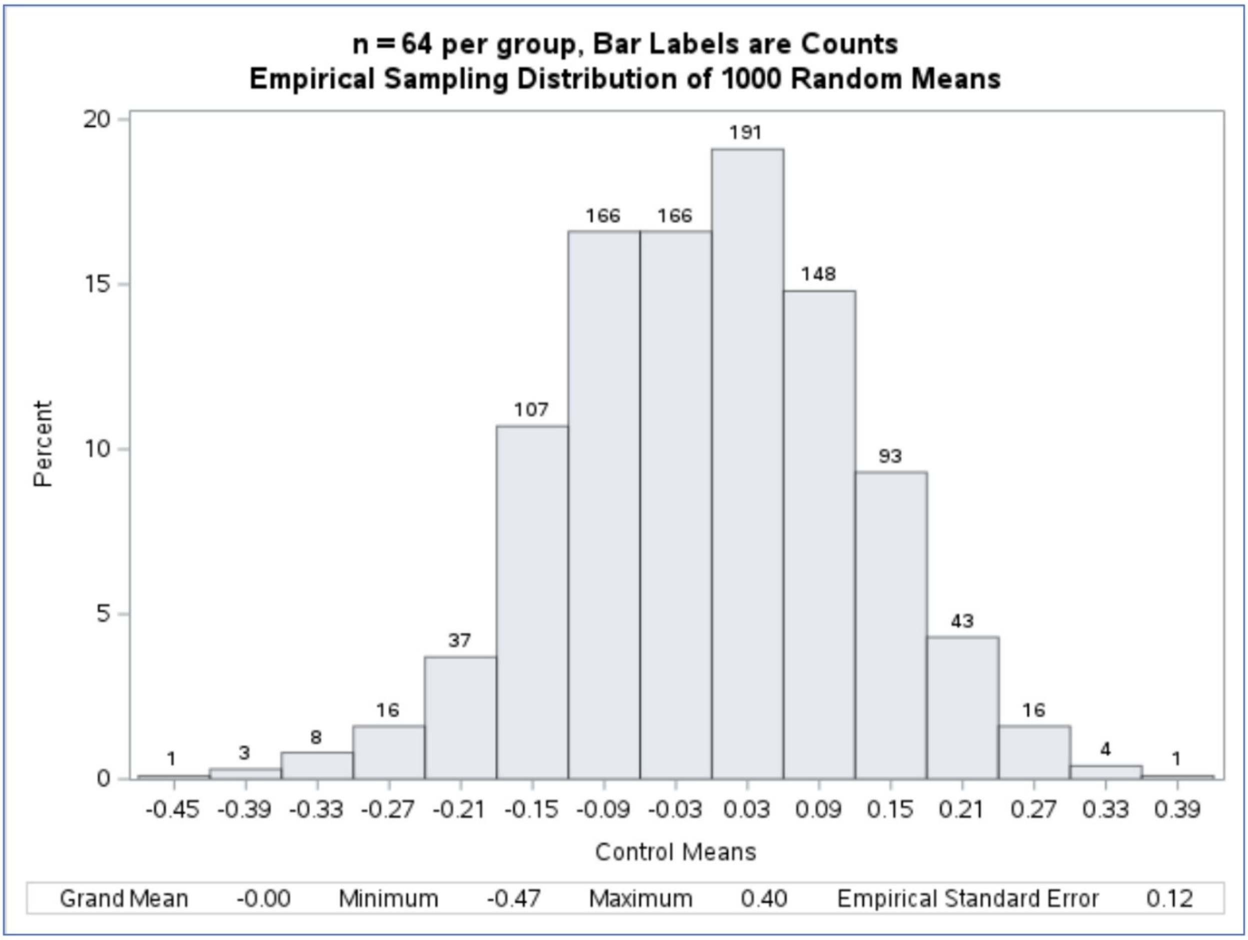


Figure 17.

The empirical sampling distribution of control means is approximately symmetric and is centered at zero.

Figure 18 displays the empirical sampling distribution of mean differences.

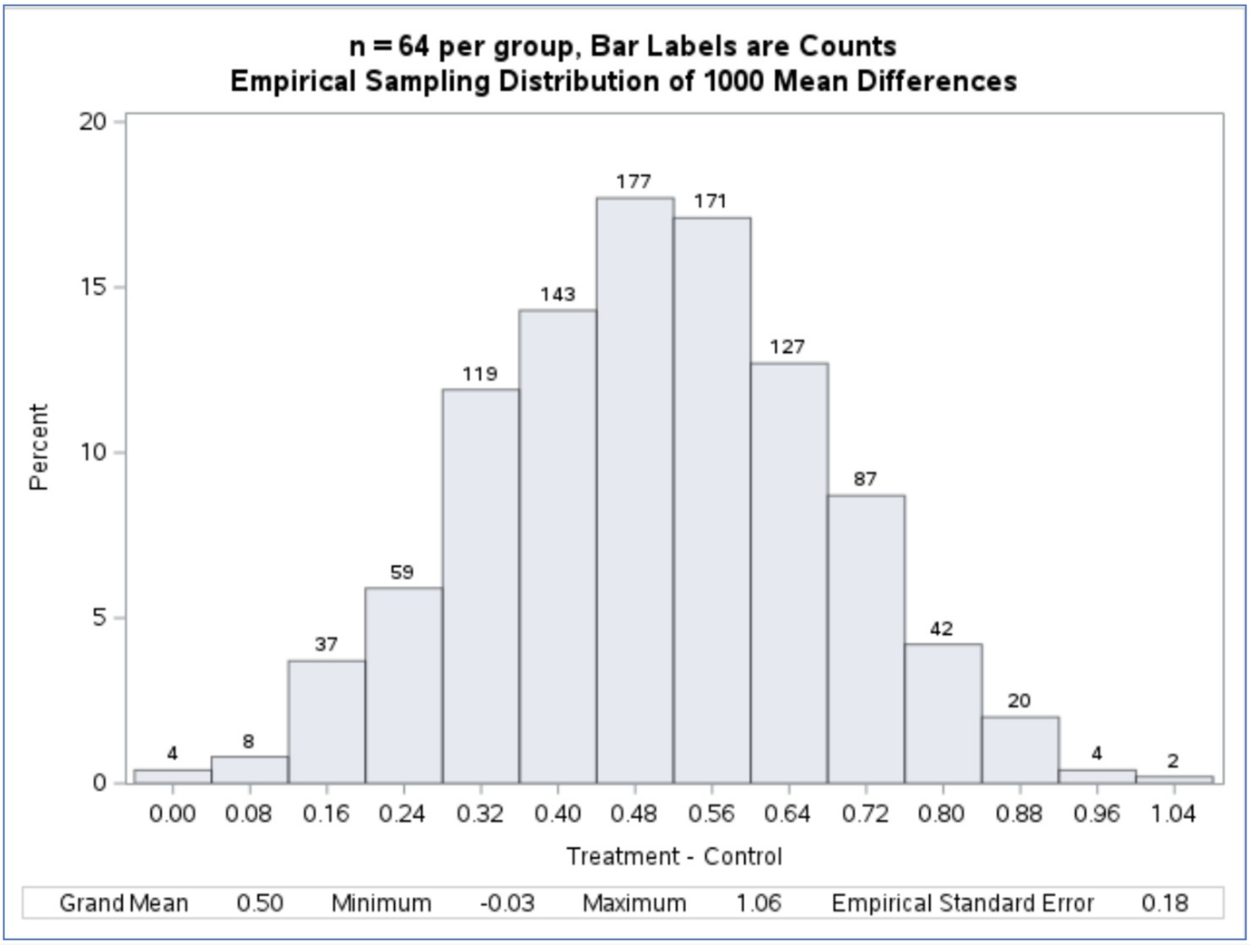


Figure 18.

The distribution of mean differences appears symmetric and centered at 0.50, the true difference in population parameters.

Figure 19 shows that the sampling distribution of p-values under the false null hypothesis is no longer uniform, as it was under the true null hypothesis.

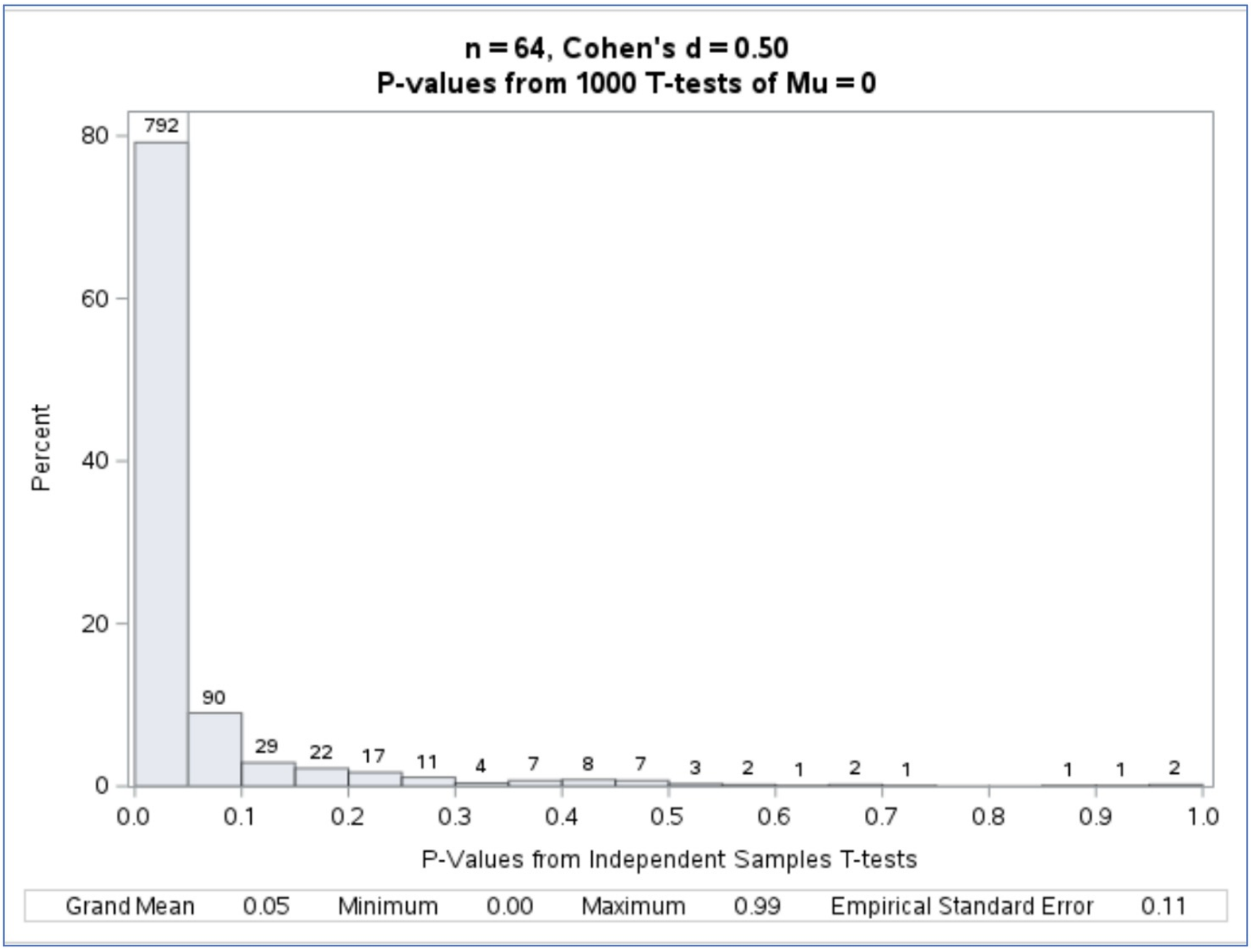


Figure 19.

There are more (792) statistically significant p-values than previously seen under the true null hypothesis with n = 64 per group (see Figure 10). These statistically significant p-values are not type 1 errors but are correct rejections of the tested false null hypothesis: μ_T - μ_C = 0, because the simulation ensured that the alternative hypothesis, μ_T - $\mu_C \neq 0$, was true

Figure 20 has the empirical sampling distribution of Cohen's d as a continuous variable.

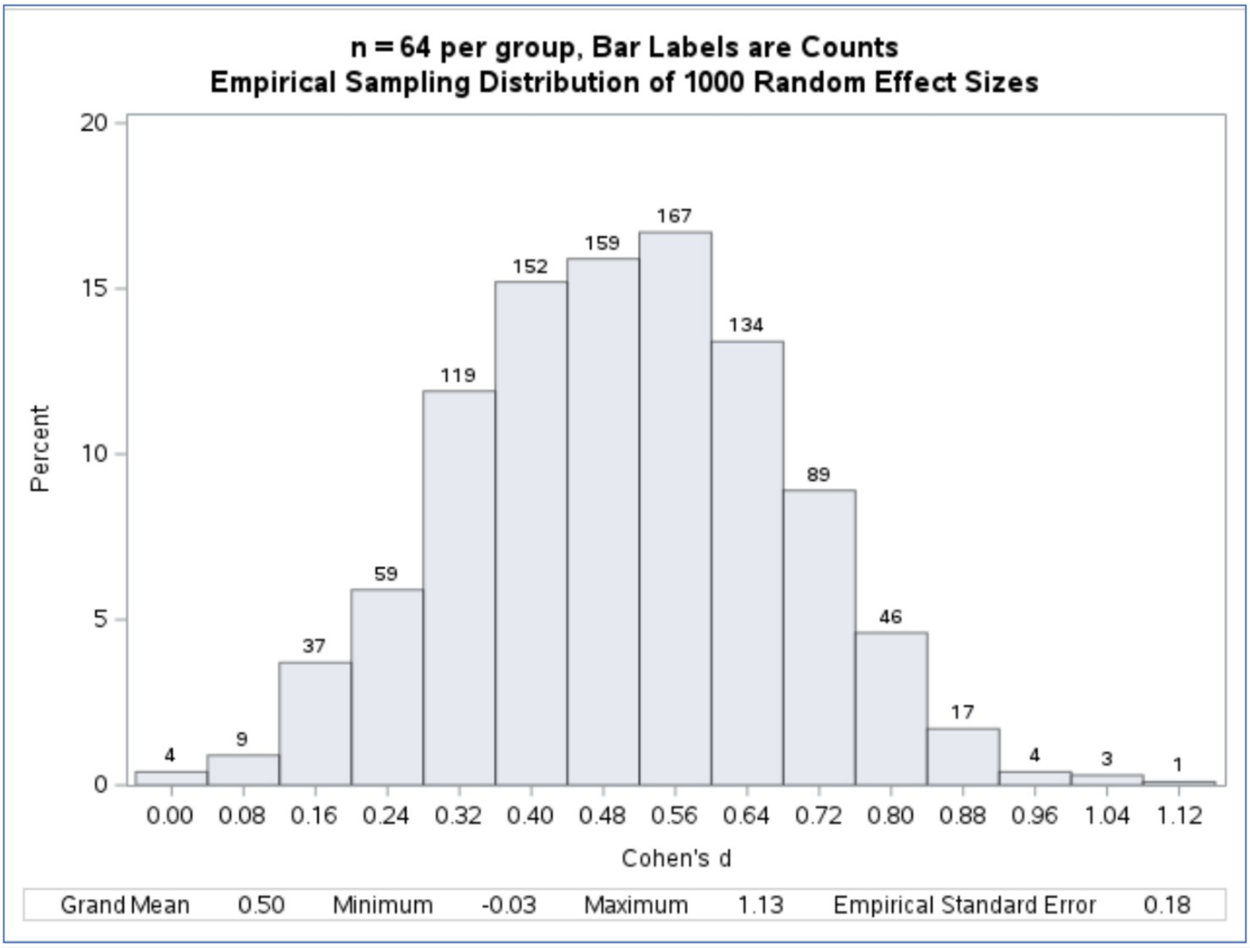


Figure 20.

The empirical sampling distribution of Cohen's d is approximately symmetric and centered at Cohen's d = 0.50, the simulated population effect size.

Figure 21 shows the binary split of Cohen's d into either no effect size or a small, medium, or large effect size.

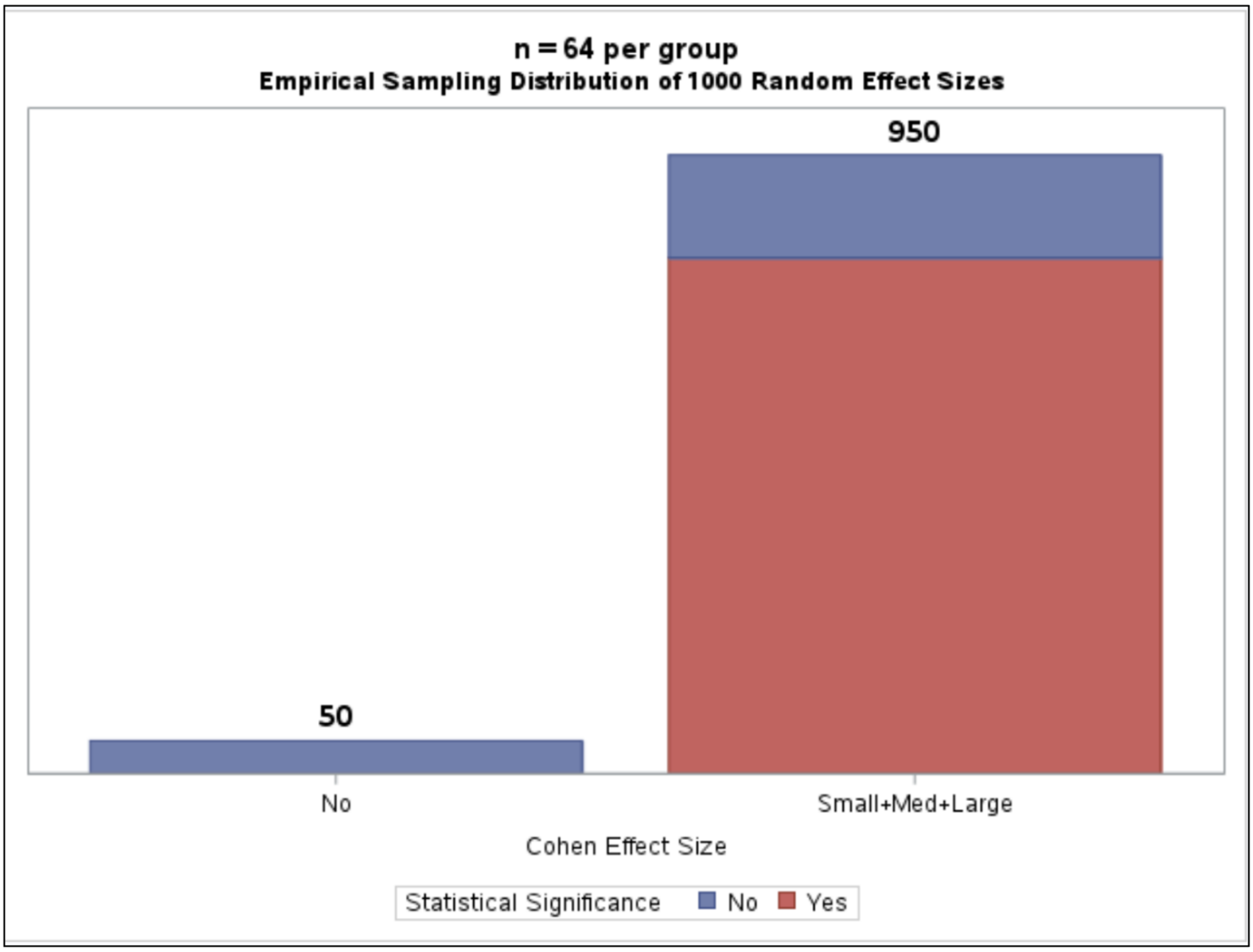
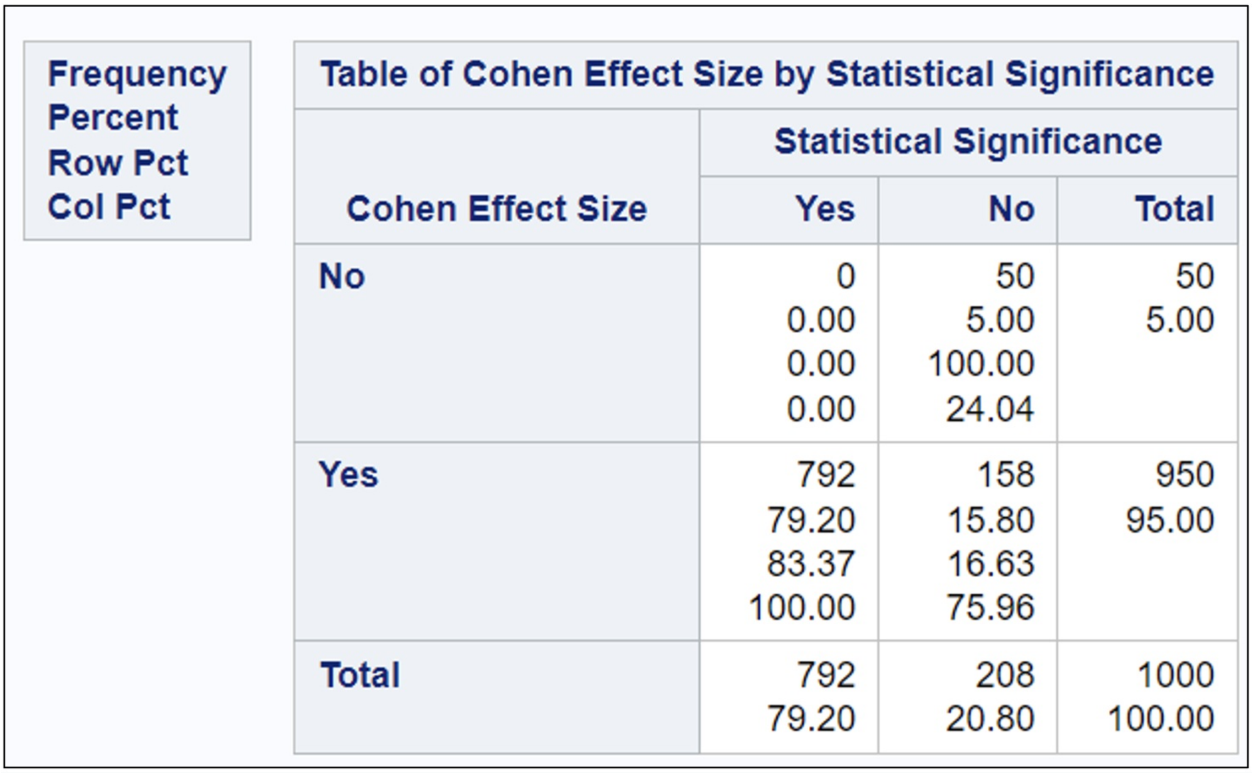


Figure 21.

Under the false null hypothesis, more effect sizes fit Cohen's categories compared to n = 64 per group under the true null hypothesis (see Figure 12).

Table 4 has the counts and percentages of statistical significance and substantive significance under the false null hypothesis.



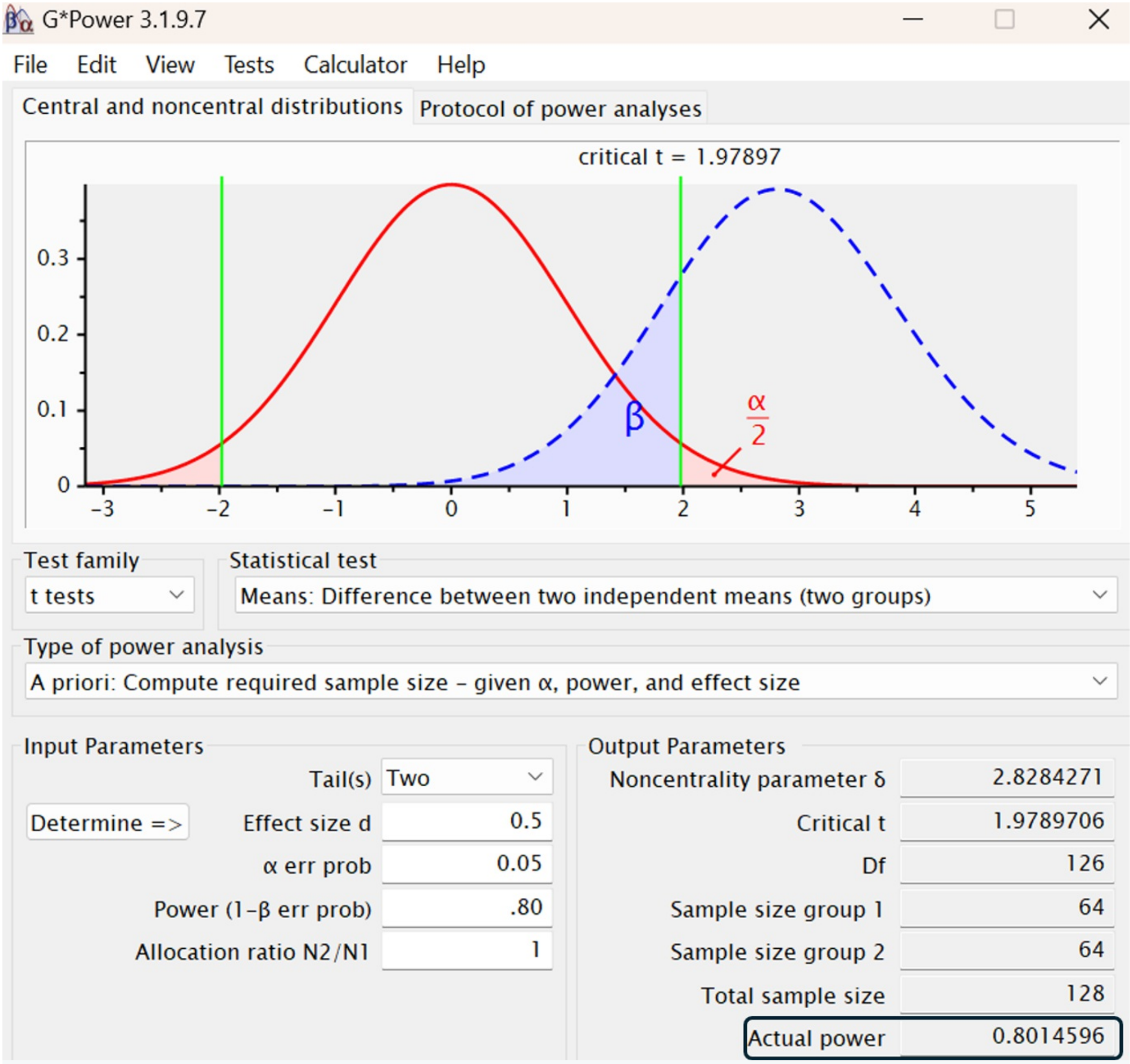
Frequency
Percent
Row Pct
Col Pct

Table of Cohen Effect Size by Statistical Significance

Cohen Effect Size	Statistical Significance		
	Yes	No	Total
No	0 0.00 0.00 0.00	50 5.00 100.00 24.04	50 5.00
Yes	792 79.20 83.37 100.00	158 15.80 16.63 75.96	950 95.00
Total	792 79.20	208 20.80	1000 100.00

Table 4.

The marginal distributions in Table 4 reveal 79% (792) statistically significant p-values with 95% (950) effect sizes that fit Cohen’s criteria for substantive significance. Of the 792 statistically significant p-values, 100% corresponded to substantive effect sizes. The simulation produced data consistent with power calculations using formulae. According to G*Power (Faul et al., 2007), with n = 64 per group, 80% power is obtained for a two-tailed independent samples t-test where Cohen’s d = 0.50, and alpha =.05 (see Figure 22).


G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.97897
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3
4
5
β
α
2
Determine =>

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	Two
Effect size d	0.5
α err prob	0.05
Power (1−β err prob)	.80
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.8284271
Critical t	1.9789706
Df	126
Sample size group 1	64
Sample size group 2	64
Total sample size	128
Actual power	0.8014596

Figure 22.

Conclusion

This paper was written to help students, applied researchers, and science writers better understand and appreciate the strengths and weaknesses of statistical significance. With graphs and a few numbers, this paper showed that statistical significance is a viable decision tool when working with small sample sizes (e.g., n < 1,000) and testing for differences in means with independent samples t-tests. Whatever alpha value is desired as the level of statistical significance, under a true null hypothesis, the probability of a statistically significant p-value does not increase with increasing sample size. Fisher (1973) noted: “Small effects will escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty” (p. 44). Notice that small effect sizes became statistically significant in only the n = 500 per group analyses. With the total sample size of n = 1,000, the utility of statistical significance as a screening tool was lost. Researchers can ignore statistical significance and decide whether trivial (non-effect sizes according to Cohen's criteria) are meaningful or substantively significant in their discipline. Similarly, statistical significance is irrelevant when a large effect size is observed with this large sample. The null hypothesis must be false because large effects are not predicted by the "Theory of Errors, one of the oldest and most fruitful lines of statistical investigation" (Fisher, 1970, p. 2), the foundation for the theory of sampling distributions. With a small sample size, if a p-value is statistically significant, the next step is to consider substantive significance. This could be with Cohen's d but more compelling if the researcher's measurement scale had a referent in the physical world.

Imagine a researcher abides by a ban on statistical significance. The results in this paper imply that the scientific research literature will be inundated with more irreplicable or unreliable results (Fricker et al., 2019). Ironically, the solution for the replication crisis may be reproduction and replication as defined in a report from the National Academies of Sciences, Engineering, and Medicine (2019): "Reproducibility includes the act of a second researcher recomputing the original results, and it can be satisfied with the availability of data, code, and methods that makes that re-computation possible. When a new study is conducted, and new data are collected, aimed at the same, or a similar scientific question as a previous one, we define it as a replication" (p. 45). Nevertheless, the report acknowledged that replication may not be possible with exploratory research designs.

Fisher conducted exploratory research with small data sets to determine the statistical significance of the effects of experimental interventions. However, he also explored large data sets created with observational research designs: "At the present time, very little can be claimed to be known as to the effects of weather upon farm crops" (Fisher, 1925, p. 1). His thoughts about developing a research hypothesis (not null or alternative) with small data sets from experiments and large data sets from observational research designs are clear. "A hypothesis is conceived and defined with all necessary exactitude; its logical consequences are ascertained by a deductive argument.; these consequences are compared with the available observations; if these are completely in accord with deductions, the hypothesis is justified until fresh and more stringent observations are available (Fisher, 1970, p. 8). The following reveals a similar thought: "An important difference is that decisions are final, while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation but of revision" (Fisher, 1973, p. 145). Regarding confirmatory analyses, "we may say that a phenomenon is experimentally demonstrated when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (Fisher, 1966, p. 14).

In conclusion, the author guarantees that the results in this paper are reproducible. Readers are encouraged to request a copy of the author's SAS program to reproduce all the results with the free Internet version of SAS OnDemand for Academics (2014). Readers can also write programs with their favorite statistical software to replicate the results under a true-and-false null hypothesis with different sample sizes.

The author has no conflicts of interest to disclose.

References

- Begg C.B. (2020). In defense of p-values. JNCI Cancer Spectrum 4(2), 1-4.https://doi.org/10.1093/jncics/pkaa012
- Benjamini Y., De Veaux, R.D., Efron B., Evans S., Glickman M., Graubard, B.I., He X., Meng, X. Reid N, Stigler, S.M., Vardeman, S.B., Wikle, C.K., Wright T., Young L.J., Kafadar, K. (2021). The ASA president's task force statement on statistical significance and replicability. Ann. Appl. Stat. 15(3): 1084-1085. DOI: 10.1214/21-AOAS1501
- Bland M (2013).Do baseline p-values follow a uniform distribution in randomized trials? PLoS ONE 8(10): e76010. doi:10.1371/journal.pone.0076010.
- Cohen, J. (1968). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.
- Fisher R. A. (1925). The influence of rainfall on the yield of wheat at Rothamsted. Phil. Trans. R. Soc. Lond. B213: 89–142. http://doi.org/10.1098/rstb.1925.0003
- Fisher R.A. (1970). Statistical Methods for Research Workers (14th ed.). Reprinted in 1993 as Statistical Methods, Experimental Designs and Scientific Inference by Oxford University Press.
- Fisher, R.A. (1973). Statistical Methods and Scientific Inference (. Hafner Press. Reprinted in 1993 as Statistical Methods, Experimental Designs and Scientific Inference by Oxford University Press.
- Fisher R.A. (1966). Design of Experiments (8th Ed.) New York: Hafner Publishing. Reprinted in 1993 as Statistical Methods, Experimental Designs and Scientific Inference by Oxford University Press.
- Fricker Jr., R.D., Burke, K., Han X. & William H. Woodall (2019). Assessing the statistical analyses used in basic and applied social psychology after their p-value ban, The American Statistician, 73:sup1, 374-384, DOI: 10.1080/00031305.2018.1537892
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31, 337-350.
- Harrington, D., D'Agostino, R.B., Gatsonis, C. Hogan, J.W., Hunter, D.J., Normand, S.T., Drazen, J.M., Hamel, B.M. (2019). New guidelines for statistical reporting in the Journal. N Engl J Med 2019; 381:285-286
- Ioannidis, J.P.A. (2005) Why most published research findings are false. PLoS Med 2(8): e124.
- Komaroff, E. (2024). Intersections of statistical significance and substantive significance: Pearson's correlation coefficients under a known true null hypothesis. QEIOS, doi:10.32388/PS72PK.
- Komaroff, E. (2020) Relationships between p-values and Pearson correlation coefficients, Type 1 errors and effect size errors, under a true null hypothesis. Journal of Statistical Theory and Practice, 14, 49. https://doi.org/10.1007/s42519-020-00115-6
- Mayo, D. & Hand, D. (2022). Statistical significance and its critics: Practicing damaging science, or damaging scientific practice? Synthese, 200, 1 - 33. https://doi.org/10.1007/s11229-022-03692-0
- National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

- SAS Institute Inc. (2014). SAS® OnDemand for Academics: User's Guide. SAS Institute Inc.
- SAS Institute Inc. (2019). SAS/STAT® 9.4 User’s Guide. Cary NC: SAS Institute Inc.
- Student (1908). The probable error of a mean. Biometrika. 6 (1), 1–25
- Trafimow D. & Marks, M. (2015). Editorial. Basic and Applied Social Psychology, 37:1, 1-2, DOI: 10.1080/01973533.2015.1012991
- Verykouki, E., & Nakas, C. T. (2023). Adaptations on the Use of p-Values for Statistical Inference: An Interpretation of Messages from Recent Public Discussions. Stats, 6(2), 539-551.
- Wang B, Zhou Z, Wang H, Tu XM, Feng C. (2019). The p-value and model specification in statistics. Gen Psychiatr. Jul 9;32(3):e100081. doi: 10.1136/gpsych-2019-100081. PMID: 31360911; PMCID: PMC6629378.
- Wasserstein, R.L, Lazar N.A. (2016). The ASA statement on p-values: context, process, and purpose. The American Statistician, 70:2, 129-133.
- Wasserstein, R.L, Schirm A.L. & Lazar N.A. (2019). Moving to a world beyond p .05. The American Statistician, 73:sup1, 1-19.
- Westfall, P. H., Tobias, R.D., Wolfinger, R.D. (2011). Multiple Comparisons and Multiple Tests Using SAS (2nd ed.). SAS Institute Inc.