Research Article

Intersections of Statistical Significance and Substantive Significance: Pearson's Correlation Coefficients Under a Known True Null Hypothesis

Eugene Komaroff[1]

1. Keiser University, Fort Lauderdale, United States

The editors of a special issue of The American Statistician stated: "Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless." This resonates with the author's view, as "statistical significance" has been conflated with substantive significance. However, the author disagrees with the editors' call for "don't use it." With relatively simple graphs and tables, this author demonstrates that small sample sizes (n < 1000) require Pearson's correlation coefficients to be screened for statistical significance (p < .05) to reduce the number of effect size errors that would otherwise be considered substantively significant under a true null hypothesis. Note here that the null hypothesis is not merely assumed true but is indeed known to be true.

Corresponding author: Eugene Komaroff, ekomaroff@keiseruniversity.edu

The Board of Directors of the American Statistical Association (ASA) published a statement in "non-technical terms" for "researchers, practitioners, and science writers" who were not statisticians about the proper use and interpretation of statistical significance (Wasserstein & Lazar, 2016, p. 129). Nonetheless, the editors in a subsequent editorial abandoned teaching statistical significance and called for a ban with the slogan "statistically significant—don't say it and don't use it" (Wasserstein et al., 2019, p. 2). Greenland et al. (2016) noted that "misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof" (p. 1). The paper is dedicated to providing a simple, intuitive, and proper understanding of statistical significance for

students, applied researchers, and science writers who are not statisticians, aiming to reassure them about the clarity of the information.

A small sample theory of sampling distributions was explained by Student (1908): “Any experiment may be regarded as forming an individual of a ‘population’ of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population” (p. 1). Fisher (1970) echoed the idea: “The entire result of an extensive experiment may be regarded as but one of a possible population of such experiments” (p. 2). Moore et al. (2021) present a graph of a sampling distribution comprised of many summary statistics drawn repeatedly with replacement from a human population. The human perspective obscures that sampling distributions are not physiological, physical, psychological, sociological, or economic phenomena that exist in the real world. They are theoretical probability distributions of summary statistics like means and proportions. For example, consider a two-sided fair coin where, by definition, the probability of heads is .50. The probability of seeing two heads after two flips is .25, using the multiplication rule of independent events. This mathematical solution can be simulated with a sampling distribution where the two flips are independently simulated 5000 times. The resulting empirical sampling distribution will show approximately 1250 heads, from which the probability (p-value) of seeing two heads on two flips is 1250/5000 = .25.

Method

This paper aims to provide in “non-technical terms” for “researchers, practitioners, and science writers” who were not statisticians an intuitive understanding of statistical significance and effect sizes with graphs and a few numbers. Komaroff (2020) provides the theoretical details but used only 435 bivariate correlations computed with 30 iid random variables and 13 different sample sizes sampled from the standard normal distribution [N(0,1]. This paper extends the results in Komaroff (2020) with much bigger empirical sampling distributions comprised of 4950 bivariate correlations computed with 100 iid random variables with only five instructive sample sizes: n = 4, 30, 100, 1000, 2000. As in the previous paper, the null hypothesis H_0: $\rho_0 = 0$ was tested for statistical significance ($\alpha = .05$) with the “Fisher” option in PROC CORR (SAS, 2019). Type 1 errors (false rejection of the true null hypothesis) were counted when $p < \alpha$ because the population parameter (ρ), as specified with the null hypothesis (H_0), was known to equal zero (H_0: $\rho_0 = 0$) by mathematical theorem (Hogg et al., 2013).

Cohen (1968) proposed categories for observed Pearson’s r as effect sizes: $|r| = 0.10$ is small, $|r| \geq 0.30$ is medium, and $|r| \geq 0.50$ is a large effect size. Although Fisher’s r to z transformation (zr) was used to test

$\rho_0 = 0$ for statistical significance, zr was back-transformed to Pearson's r to evaluate effect sizes. Because $\rho_0 = 0$ all $|r| \geq .10$ were effect size errors.

Results

Fisher (1970) said: "The distribution of r is not normal in small samples, and even for large samples, it remains far from normal for high correlations" (pp. 200-201). Figure 1 shows the shape of the empirical sampling distribution of 4950 Pearson correlations with n = 4.

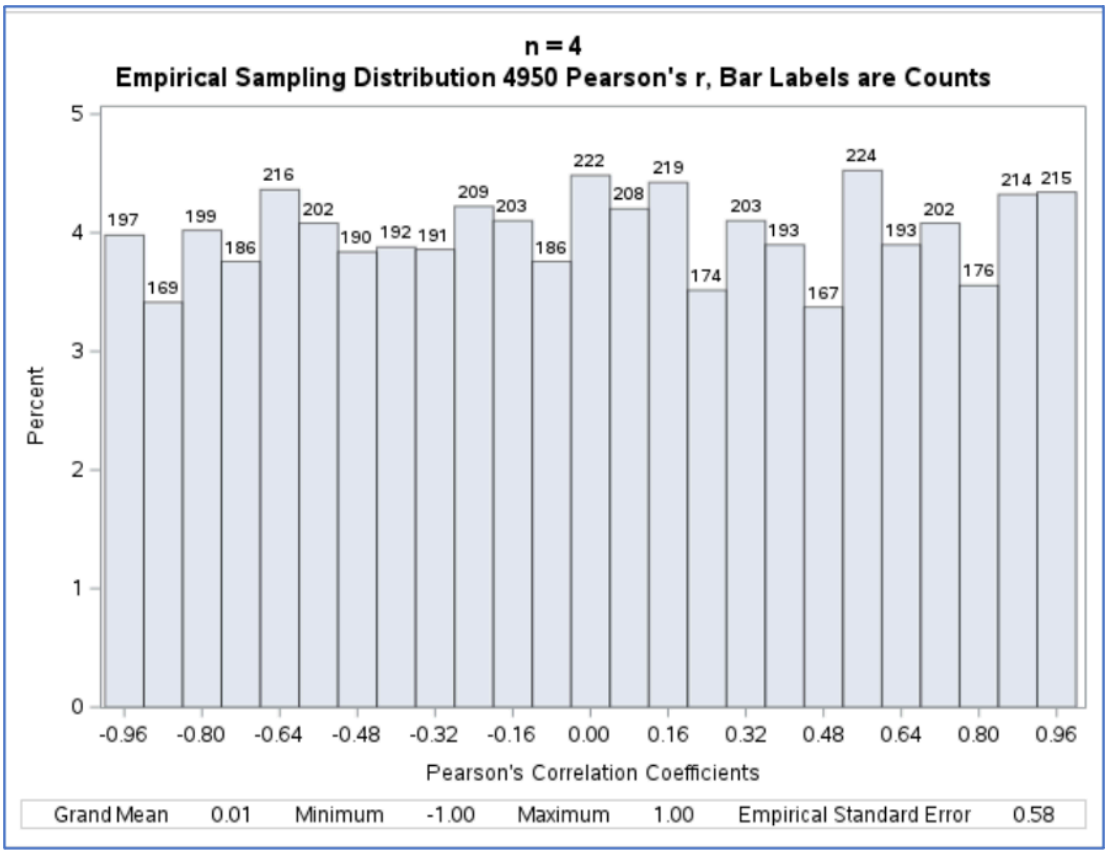

Figure 1.

This empirical sampling distribution is far from a symmetric, normal distribution. The empirical standard error (i.e., the standard deviation of this distribution) is 0.58, indicating a considerable dispersion around the central value (Grand Mean) of zero in the range from -1.00 to +1.00. It is evident that many observed correlations are wrong estimates of the actual population correlation coefficient; however, the overall mean correlation (Grand Mean) is very close to zero, which is consistent with the

Law of Large Numbers (Moore et al., 2021, p. 345). Figure 2 is the empirical sampling distribution of Fisher's r to z transformation (zr) of the correlations in Figure 1. Despite the small sample size, this empirical sampling distribution is approximately normal. This was Fisher's motivation for inventing zr because now the properties of the well-known standard normal distribution (zr ~ N(0,1) can be used to determine statistical significance (Fisher, 1970, p. 201).

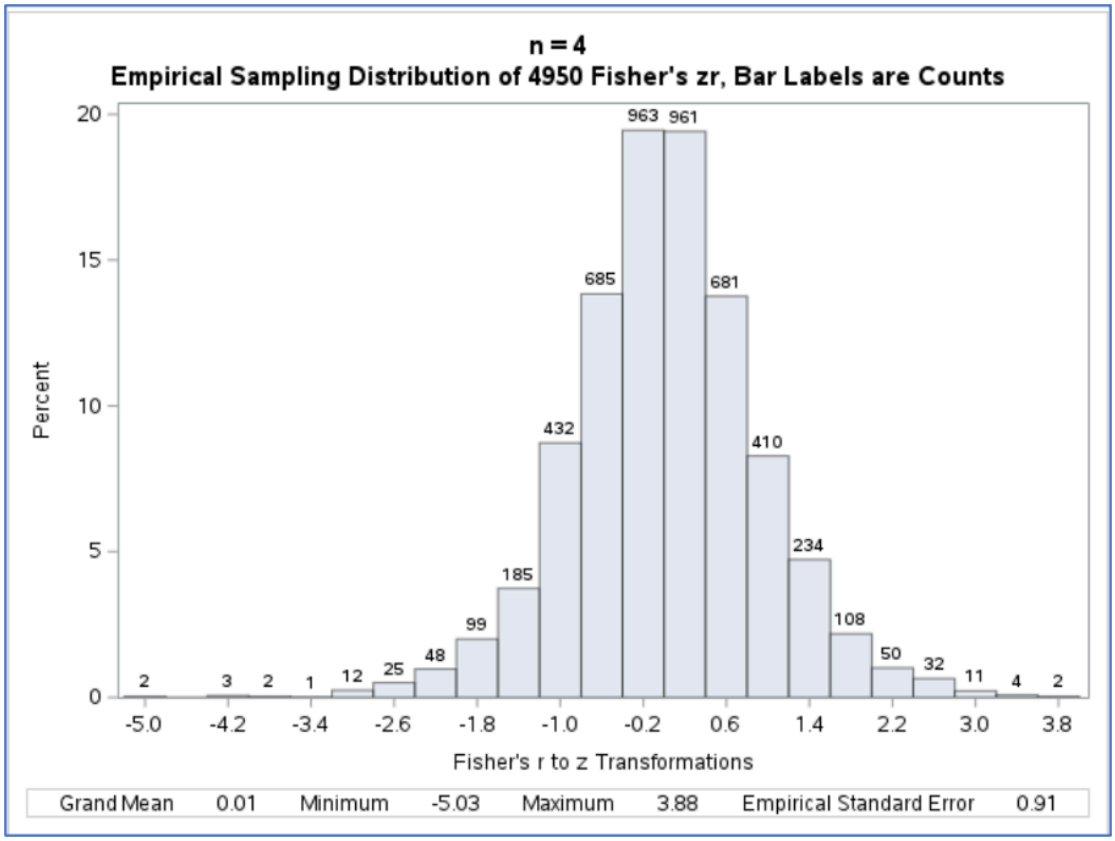

Figure 2.

A two-sided, 5% level of statistical significance corresponds to a |zr| > 1.96. These correspond to statistically significant p-values. Figure 3 reveals the count.

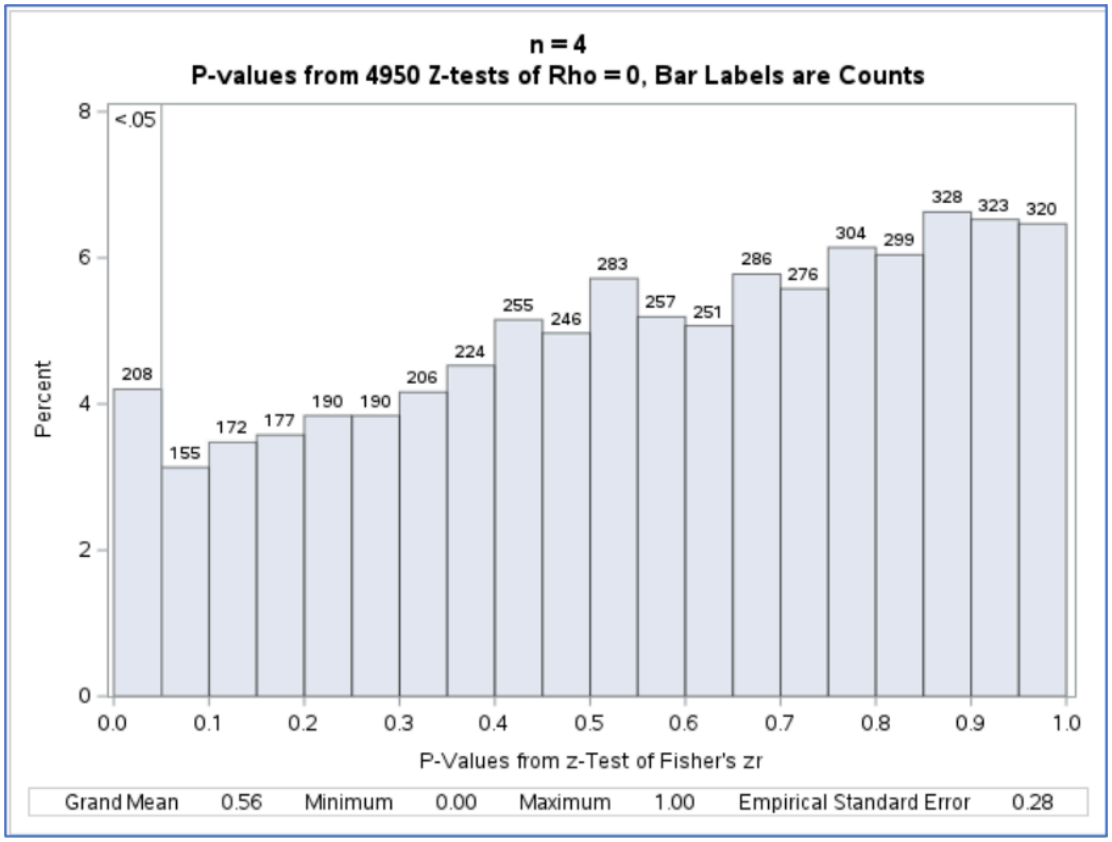

Figure 3.

In this empirical sampling distribution of zr under the null hypothesis of $\rho_0 = 0$, there are 208 (4%) statistically significant p-values (type 1 errors) because $p < \alpha$ and $\alpha = .05$ set a priori as the level of statistical significance. Figure 4 displays the empirical sampling distribution of Pearson's r as Cohen's effect sizes.

doi.org/10.32388/PS72PK

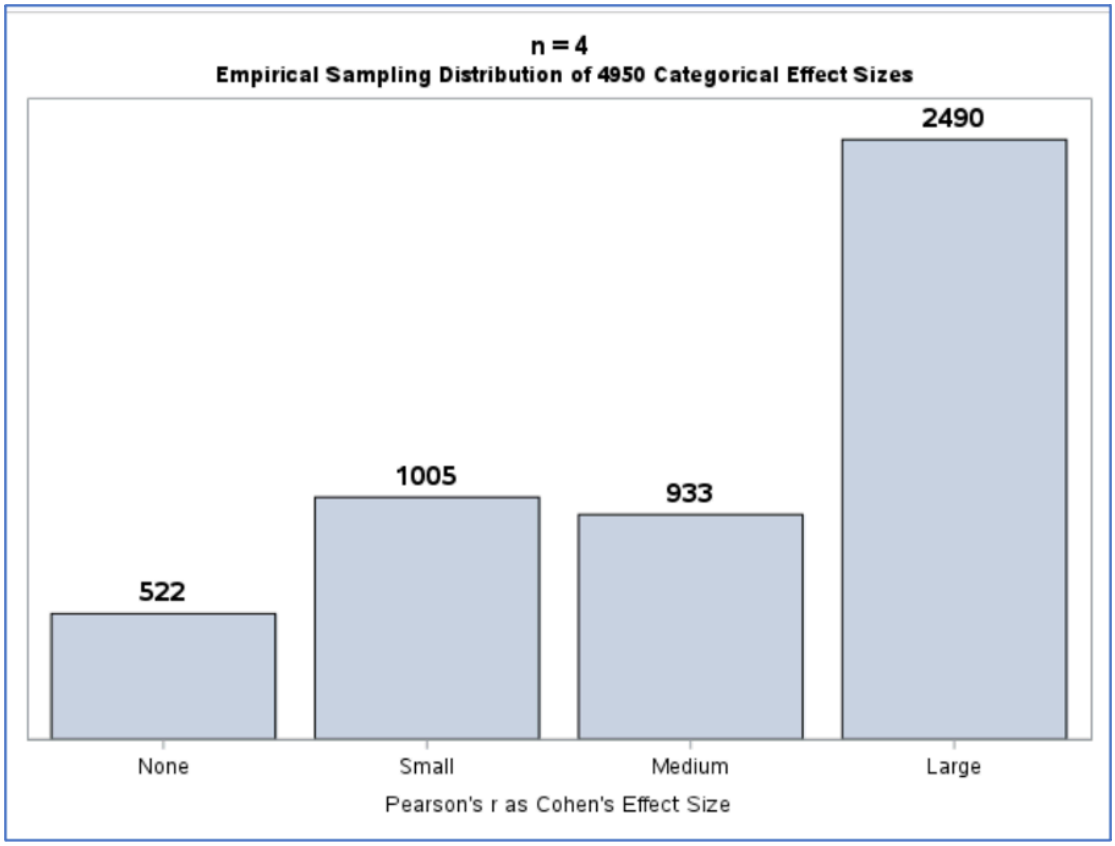

Figure 4.

Approximately 10% (522) are not effect sizes, leaving 90% to be misinterpreted as substantive or meaningful effect sizes when they are merely effect size errors under the true null hypothesis: $\rho 0 = 0$. Figure 5 shows that screening for statistical significance excludes many effect size errors from consideration.

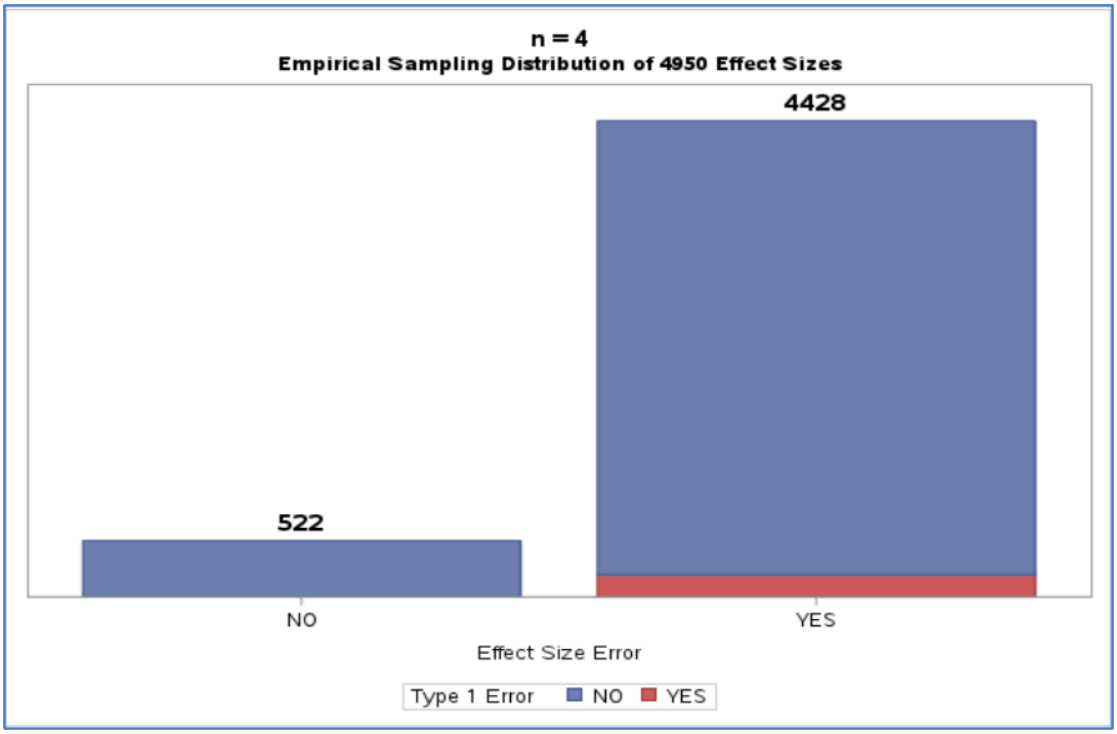

Figure 5.

Table 1 reveals that 95% (4220) are excluded because they are not statistically significant, leaving only 5% (208) effect size errors to be misinterpreted as substantive effect sizes. Classical Fisherian statistical theory predicts a 5% type 1 error under a true null hypothesis. It is important to recognize that no statistically significant effect size errors exist.

doi.org/10.32388/PS72PK

n = 4

Intersection of Effect Size Errors and Type 1 Errors

The FREQ Procedure

Frequency Row Pct	Table of Effect Size Error by Type 1 Error		
	Type 1 Error		
Effect Size Error	YES	NO	Total
NO	0 0.00	522 100.00	522
YES	208 4.70	4220 95.30	4428
Total	208	4742	4950

Table 1.

Table 2 shows that with n = 4, remarkedly high correlations appeared purely by chance but were nonetheless merely statistically significant effect size errors.

n = 4

Statistically Significant Effect Size Errors

The MEANS Procedure

Analysis Variable : Corr Sample Correlation		
N	Minimum	Maximum
208	-1.00	1.00

Table 2.

Figure 6 is the empirical sampling distribution of Pearson correlations with n = 30. This distribution is approximately normal with an empirical standard error of 0.19, indicating that the correlations' dispersion is closer to the true $\rho = 0$ than 0.28 with n = 4.

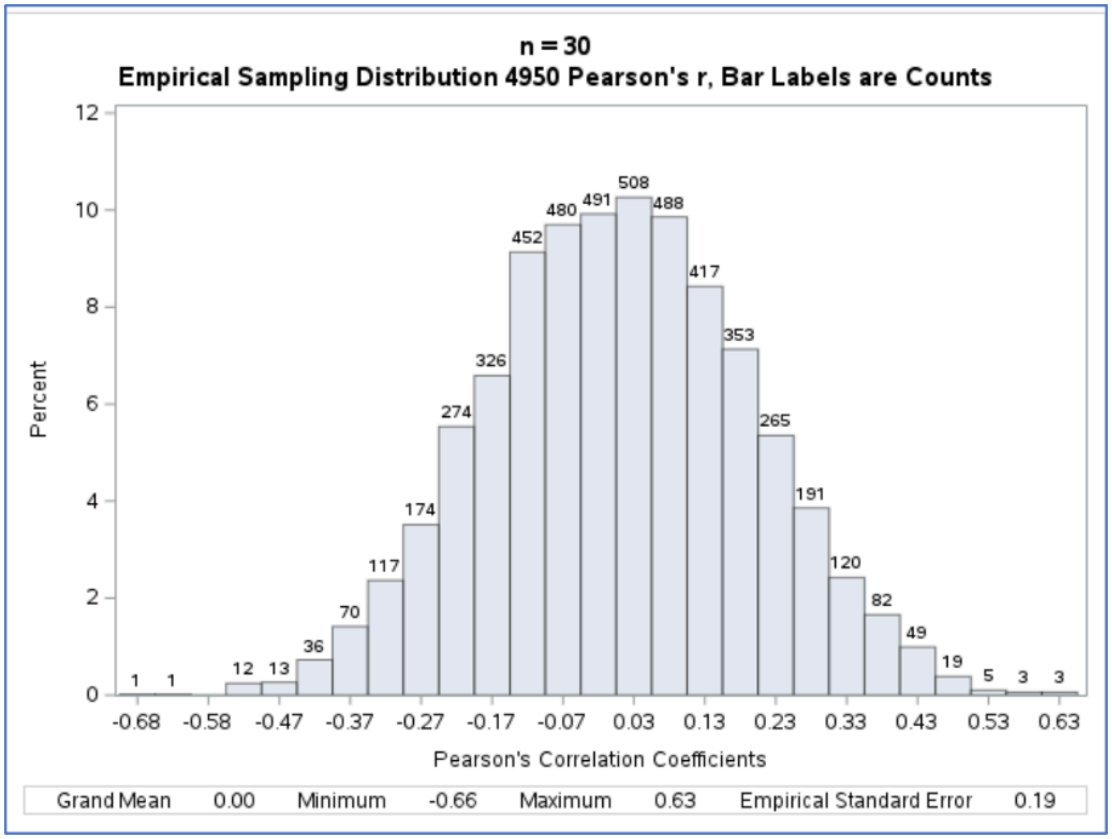

Figure 6.

Figure 7 is an empirical sampling distribution of zr values corresponding to the observed correlations in Figure 6.

doi.org/10.32388/PS72PK

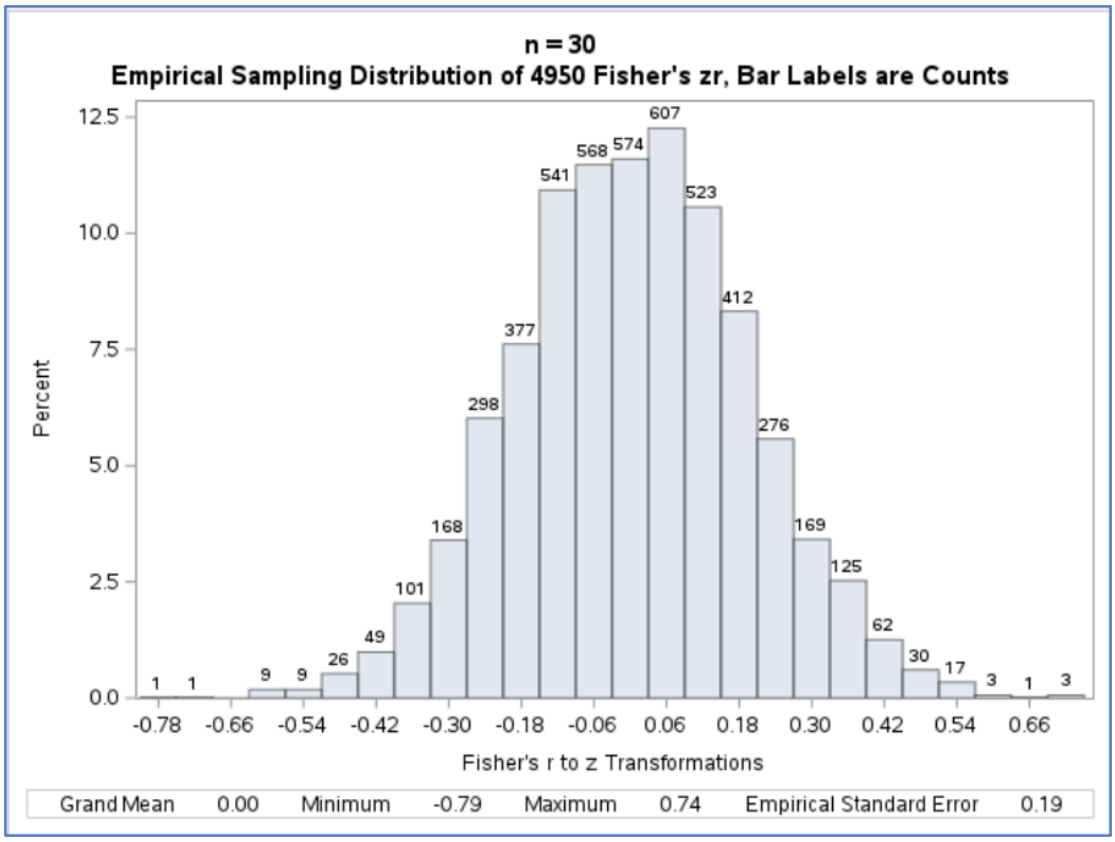

Figure 7.

This distribution appears normal with the same standard error = 0.19 as in Figure 6. However, the zr range is wider, -0.72 to 0.72, compared to the range -0.66 to 0.63 with n = 4 (Figure 6). Figure 8 shows the empirical sampling distribution of p-values corresponding to zr values.

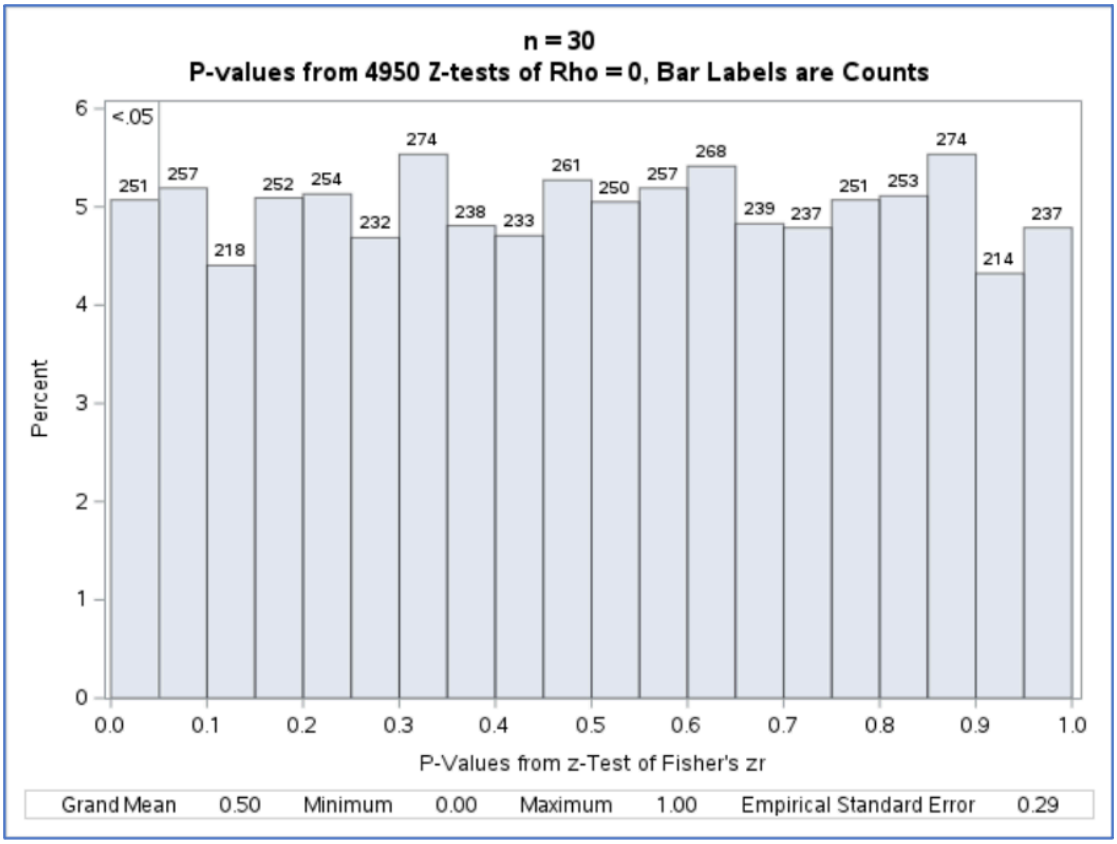

Figure 8.

As predicted by classical Fisherian statistical theory, approximately 5% (249/4950) are type 1 errors under a true null hypothesis. Figure 9 displays the 4950 Pearson correlations as Cohen's effect sizes.

doi.org/10.32388/PS72PK

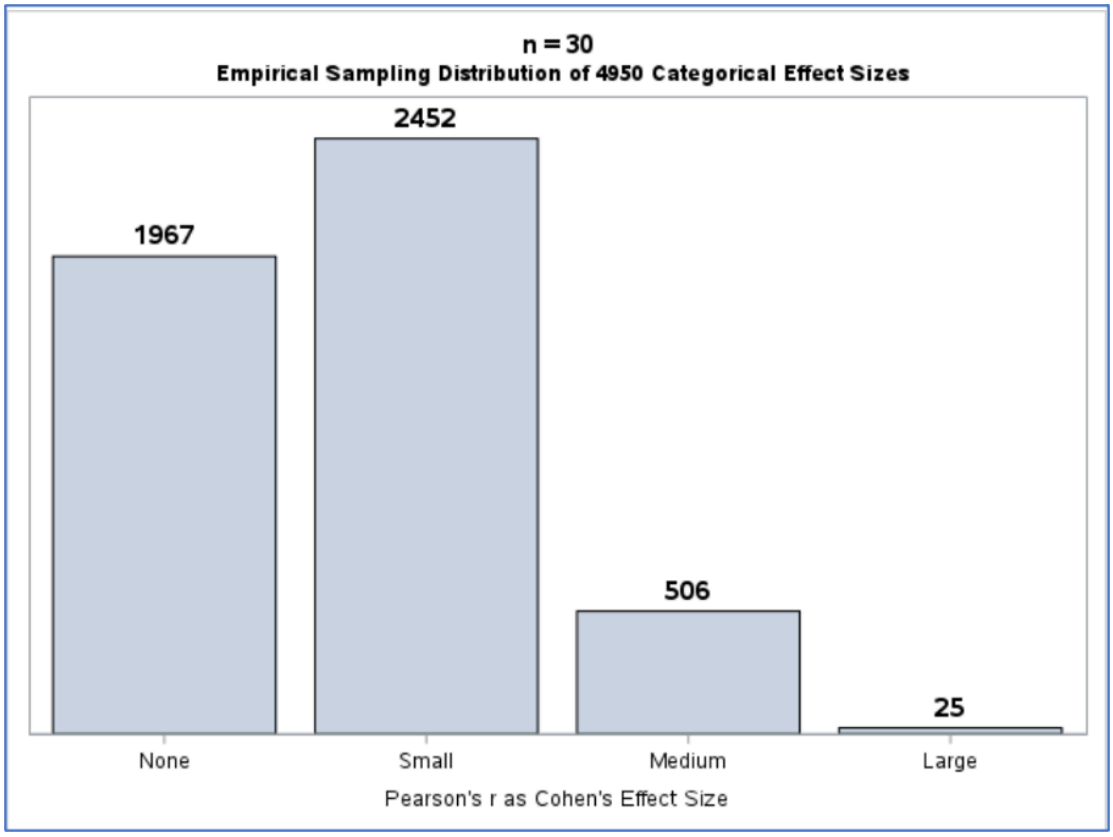

Figure 9.

Approximately 40% (1967) can be ignored, leaving about 60% effect size errors that could be easily misinterpreted as substantive or meaningful effect sizes if statistical significance is not considered. Figure 10 demonstrates that a relatively small percentage would be considered statistically significant. Table 3 shows the counts.

doi.org/10.32388/PS72PK

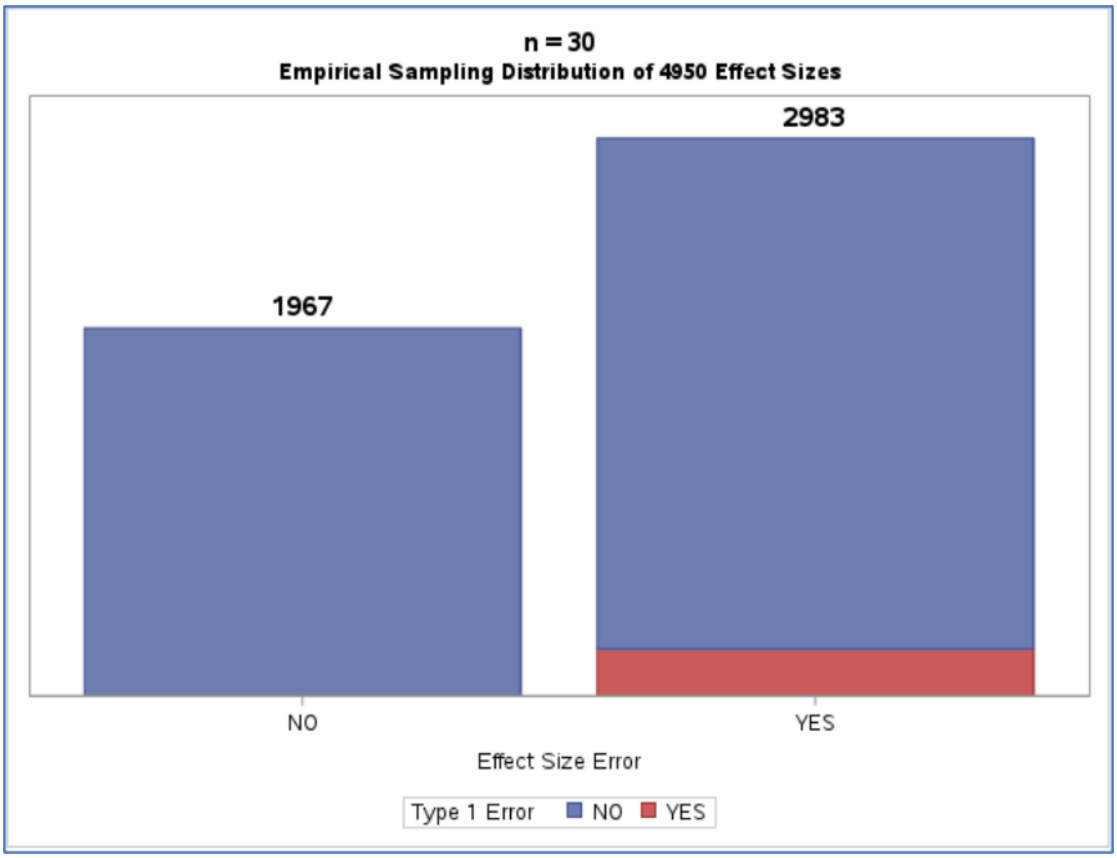

Figure 10.

Statistical significance excluded approximately 92% (2732) effect size errors from further consideration, leaving approximately 8% (251) to be misinterpreted as substantively significant.

doi.org/10.32388/PS72PK

n = 30
Intersection of Effect Size Errors and Type 1 Errors

The FREQ Procedure

Frequency
Row Pct

Table of Effect Size Error by Type 1 Error			
	Type 1 Error		
Effect Size Error	YES	NO	Total
NO	0 0.00	1967 100.00	1967
YES	251 8.41	2732 91.59	2983
Total	251	4699	4950

Table 3.

Again, it is noteworthy that statistical significance detected only substantive effect sizes (|r| > .10). Table 4 shows the range of the 251 statistically significant correlations. However, this range does not have the high correlations seen with n = 4.

n = 30
Statistically Significant Effect Size Errors

The MEANS Procedure

Analysis Variable : Corr Sample Correlation		
N	Minimum	Maximum
251	-0.66	0.63

Table 4.

Figure 11 shows the empirical sampling distribution of 4950 Pearson correlations with n = 100. This distribution is approximately normal, with an empirical standard error of 0.10. This indicates a smaller dispersion of observed correlations in this sampling distribution, centered at zero compared to n = 4 or n

= 30. In other words, fewer misleading estimates of $\rho = 0$ appeared in this empirical sampling distribution with the increase in sample size.

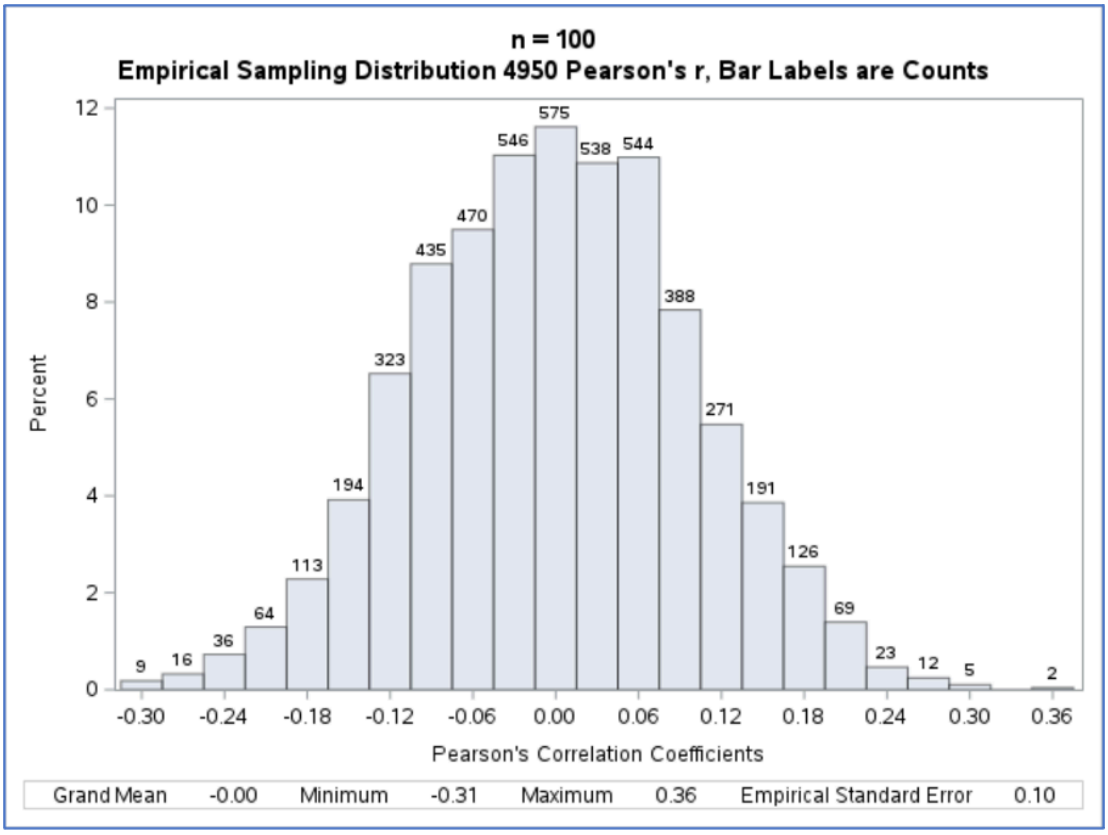

Figure 11.

Figure 12 shows the empirical sampling distribution of zr values corresponding to the observed correlations with n = 100. This distribution also appears normal with the same empirical standard error of 0.10 as Figure 11. Perhaps this is why Fisher's (1970 table of critical values of zr to determine statistical significance stops at n = 100 (p. 211).

doi.org/10.32388/PS72PK

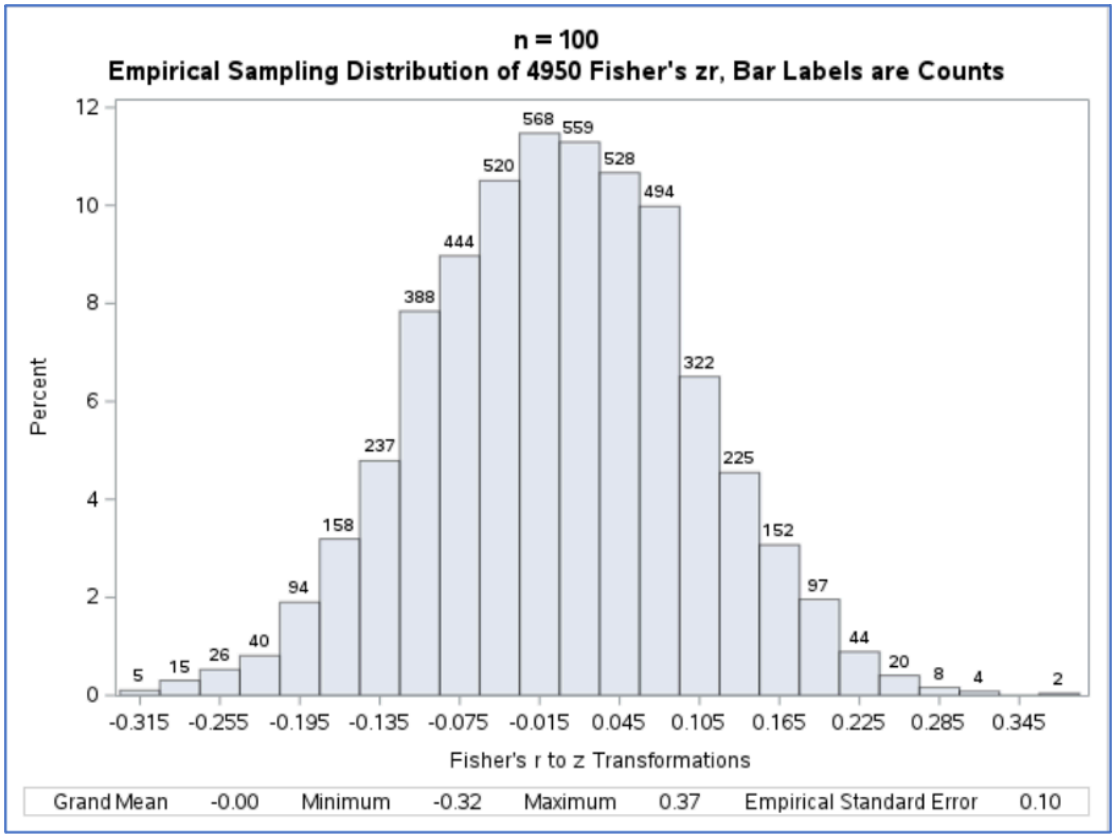

Figure 12.

Figure 13 displays the p-values from the significance test of the zr values. Approximately 5% (221) are statistically significant p-values.

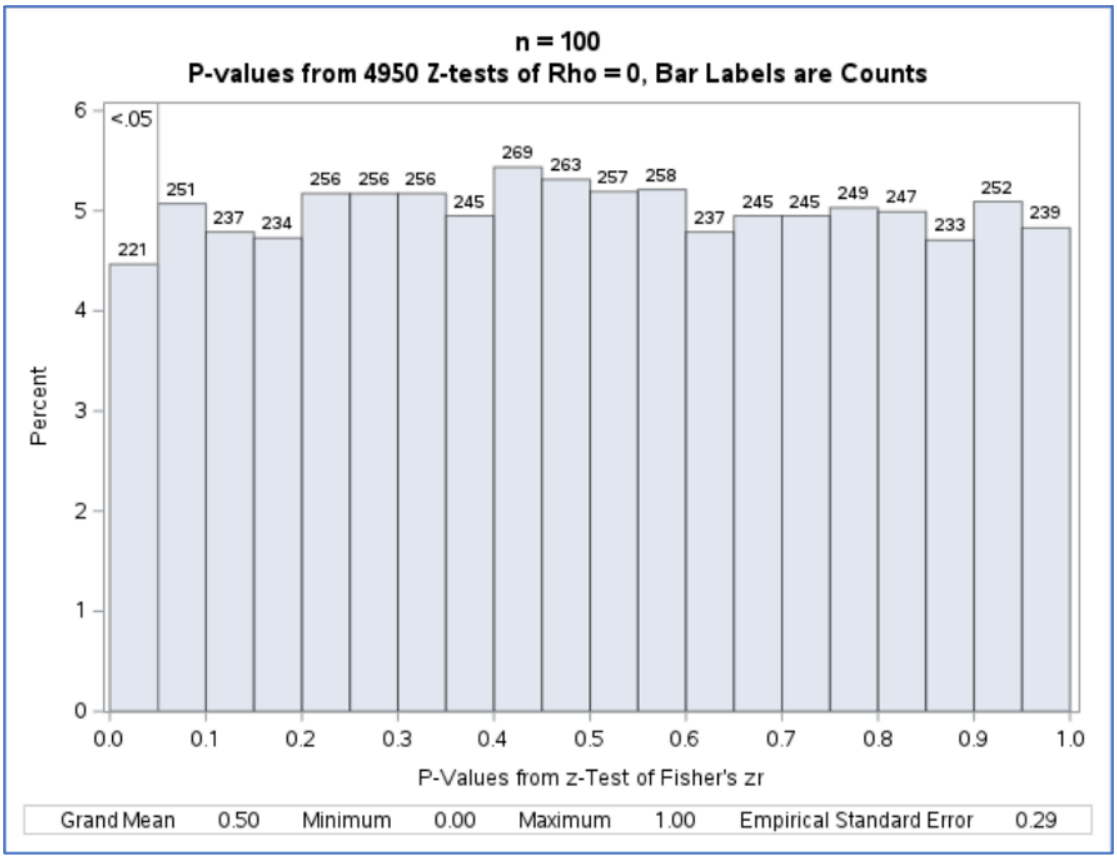

Figure 13.

Figure 14 displays 4950 Pearson's correlations categorized as Cohen's effect sizes. Approximately 69% (3394) can be ignored as non effet sizes.

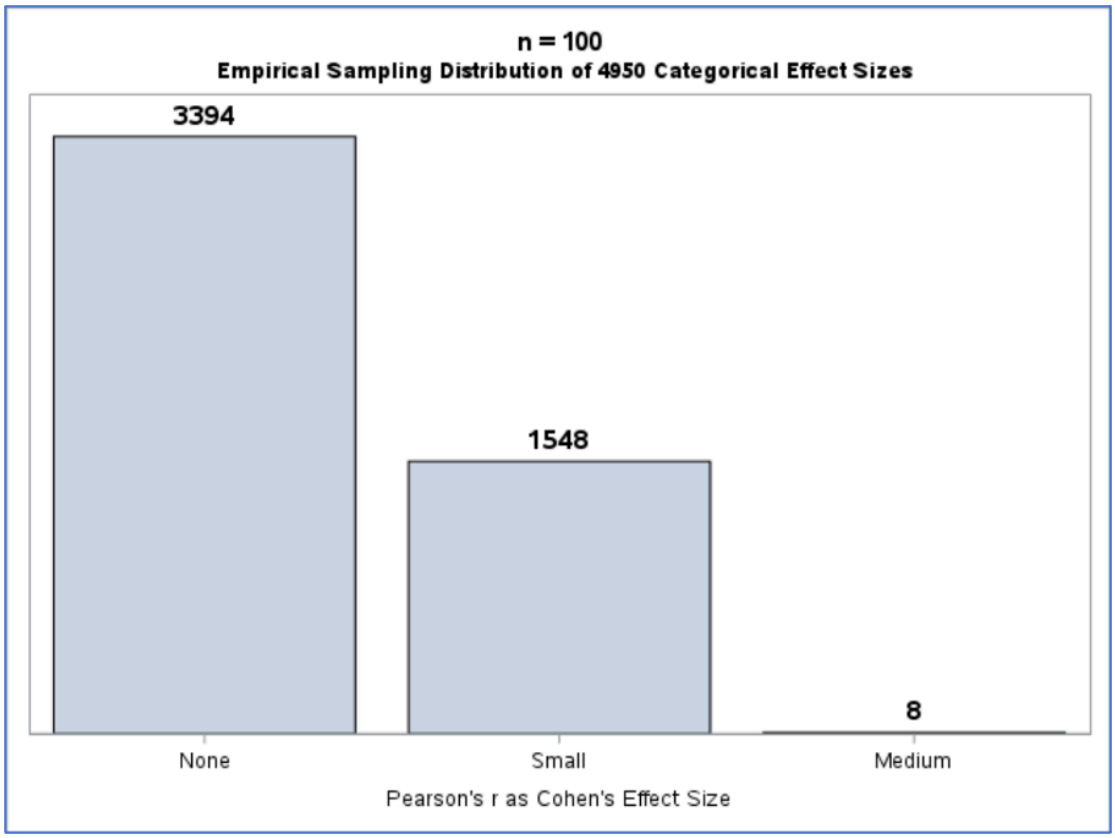

Figure 14.

Figure 15 reveals that relatively few are statistically significant.

doi.org/10.32388/PS72PK

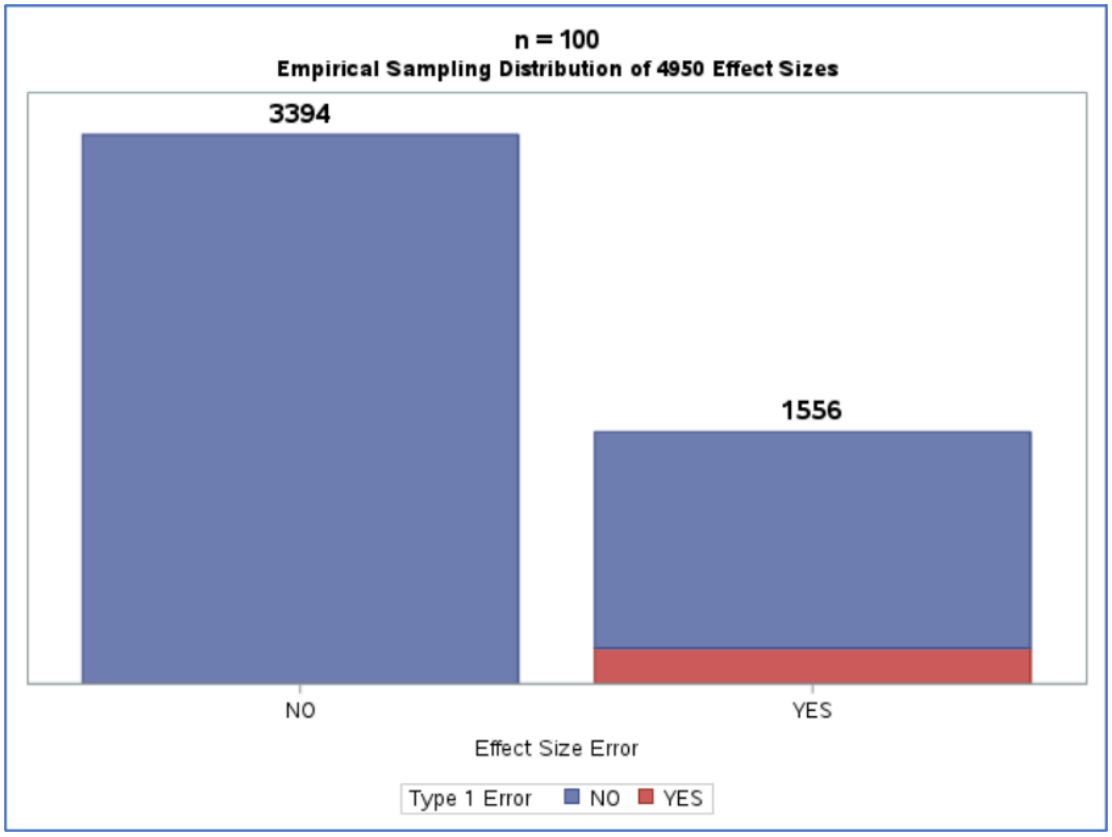

Figure 15.

Table 5 reveals that statistical significance would exclude approximately 86% (1335) effect size errors from further consideration, leaving 14% (263) to be misinterpreted as meaningful effect sizes. Again, it is noteworthy that the type 1 error occurred only with Cohen's effect sizes |r| > .10.

doi.org/10.32388/PS72PK

n = 100
Intersection of Effect Size Errors and Type 1 Errors

The FREQ Procedure

Frequency
Row Pct

Table of Effect Size Error by Type 1 Error			
	Type 1 Error		
Effect Size Error	YES	NO	Total
NO	0 0.00	3394 100.00	3394
YES	221 14.20	1335 85.80	1556
Total	221	4729	4950

Table 5.

Table 6 shows that the range of statistically significant correlations is smaller than previously seen with either n = 4 or 30.

n = 100
Statistically Significant Effect Size Errors

The MEANS Procedure

Analysis Variable : Corr Sample Correlation		
N	Minimum	Maximum
221	-0.31	0.36

Table 6.

Figure 16 shows the empirical sampling distribution of 4950 Pearson correlations with n = 1000. This distribution is approximately normal, with an empirical standard error of 0.03, indicating a much smaller dispersion around zero than the previous distributions with smaller sample sizes.

doi.org/10.32388/PS72PK

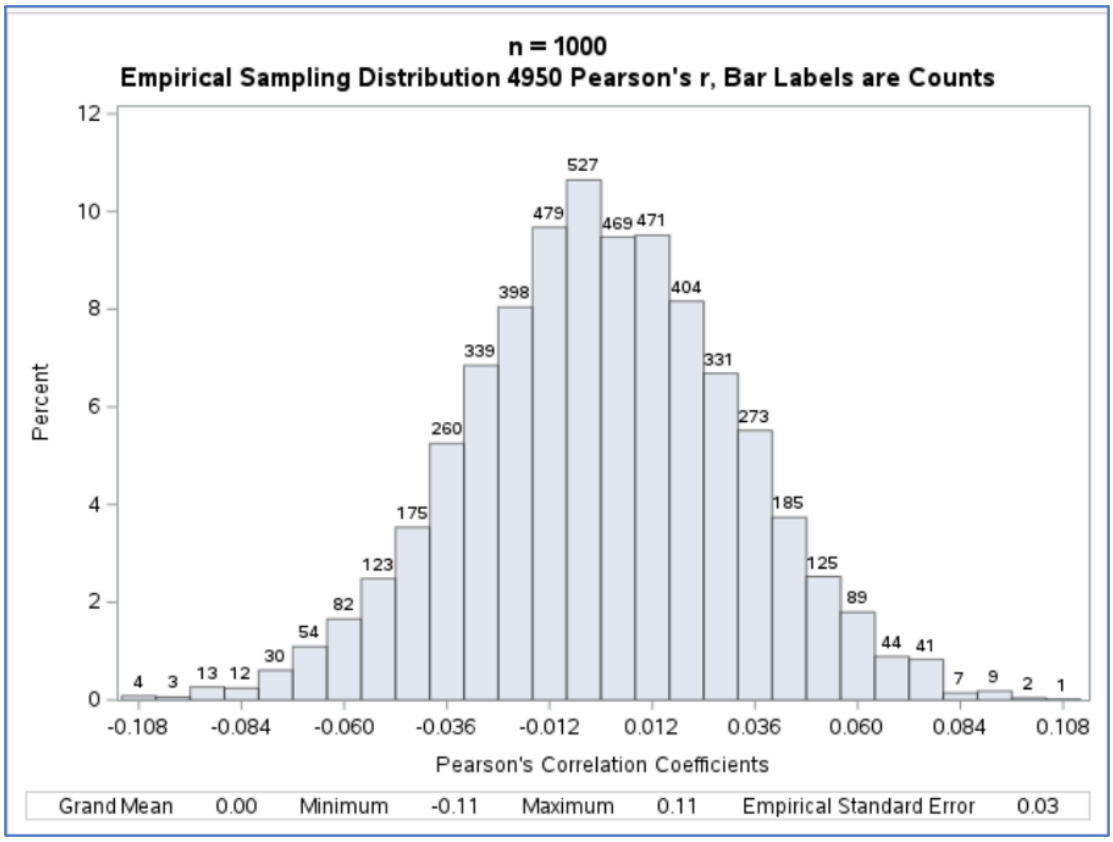

Figure 16.

Figure 17 shows the empirical sampling distribution of zr values corresponding to the observed correlations with n = 1000. This distribution appears normal, with the same empirical standard error of 0.10 and the same minimum and maximum values as in Figure 16. In effect, Fisher's r-to-z transform is unnecessary, which makes sense because the technique was created to detect the statistical significance of Pearson's correlation coefficients with small sample sizes.

doi.org/10.32388/PS72PK

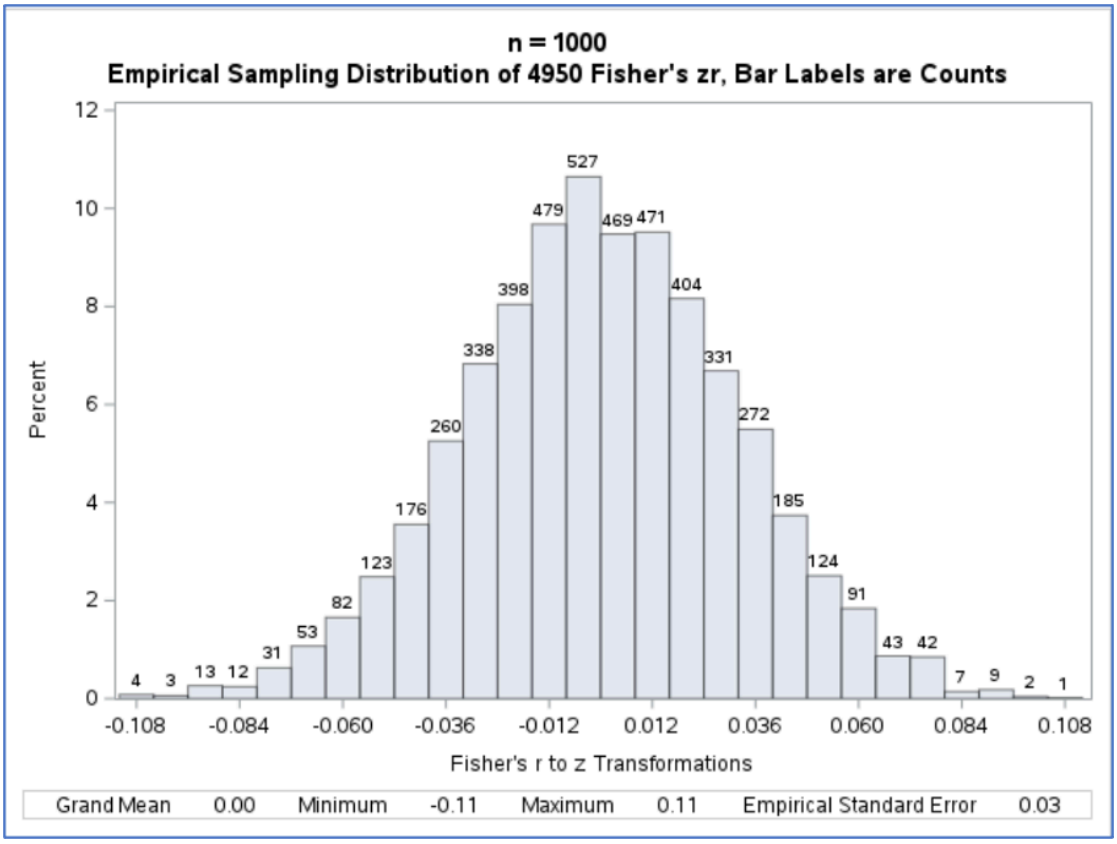

Figure 17.

Figure 18 displays the p-values from the significance test of the zr values. Approximately 5% (260) are statistically significant p-values (type 1 errors).

doi.org/10.32388/PS72PK

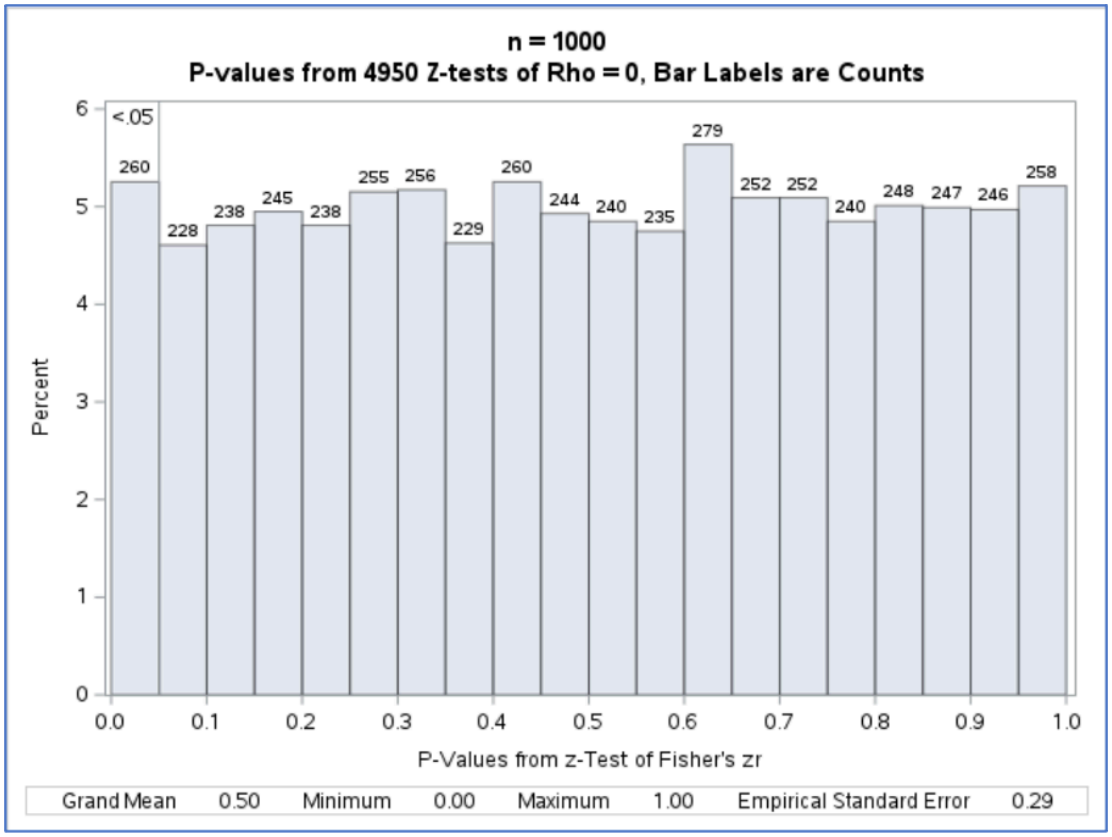

Figure 18.

Figure 19 displays 4950 Pearson's correlations categorized as Cohen's effect sizes. Approximately 4942 (99.8%) can be ignored leaving 8 (0.2%)for further consideration.

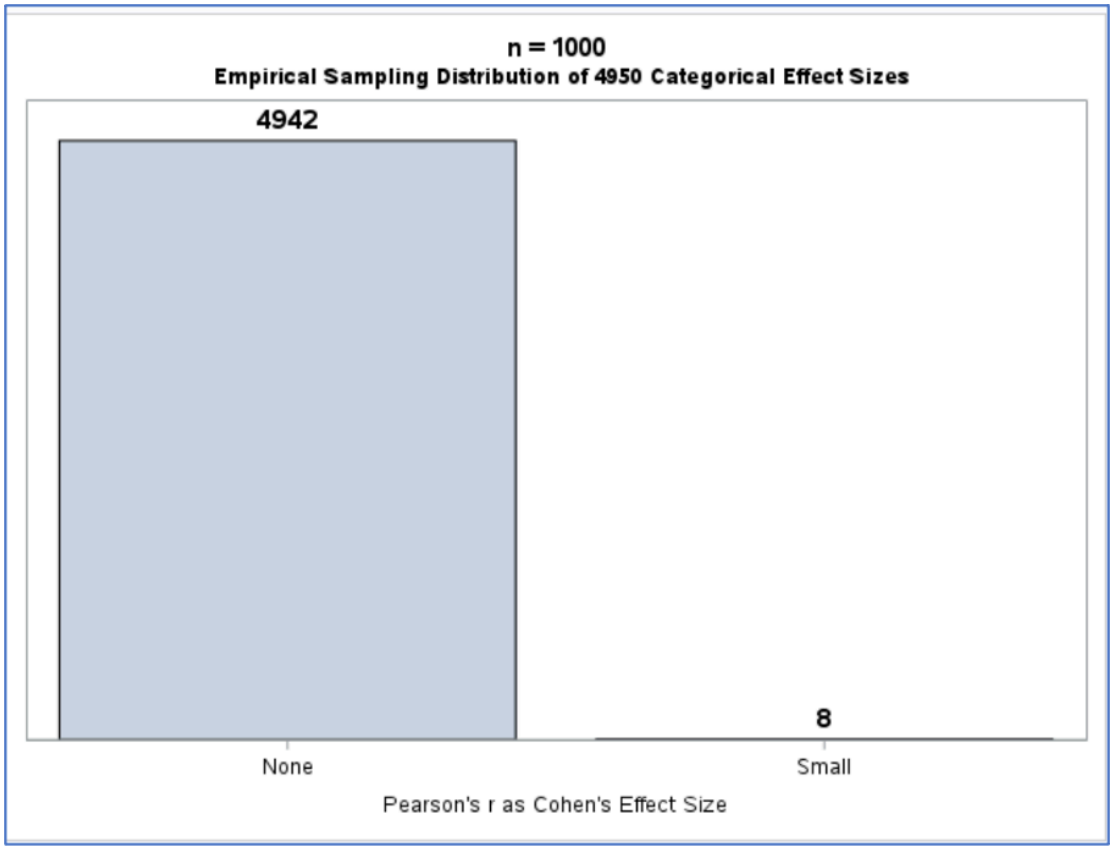

Figure 19.

Figure 20 reveals the relatively few statistically significant effect size errors.

doi.org/10.32388/PS72PK

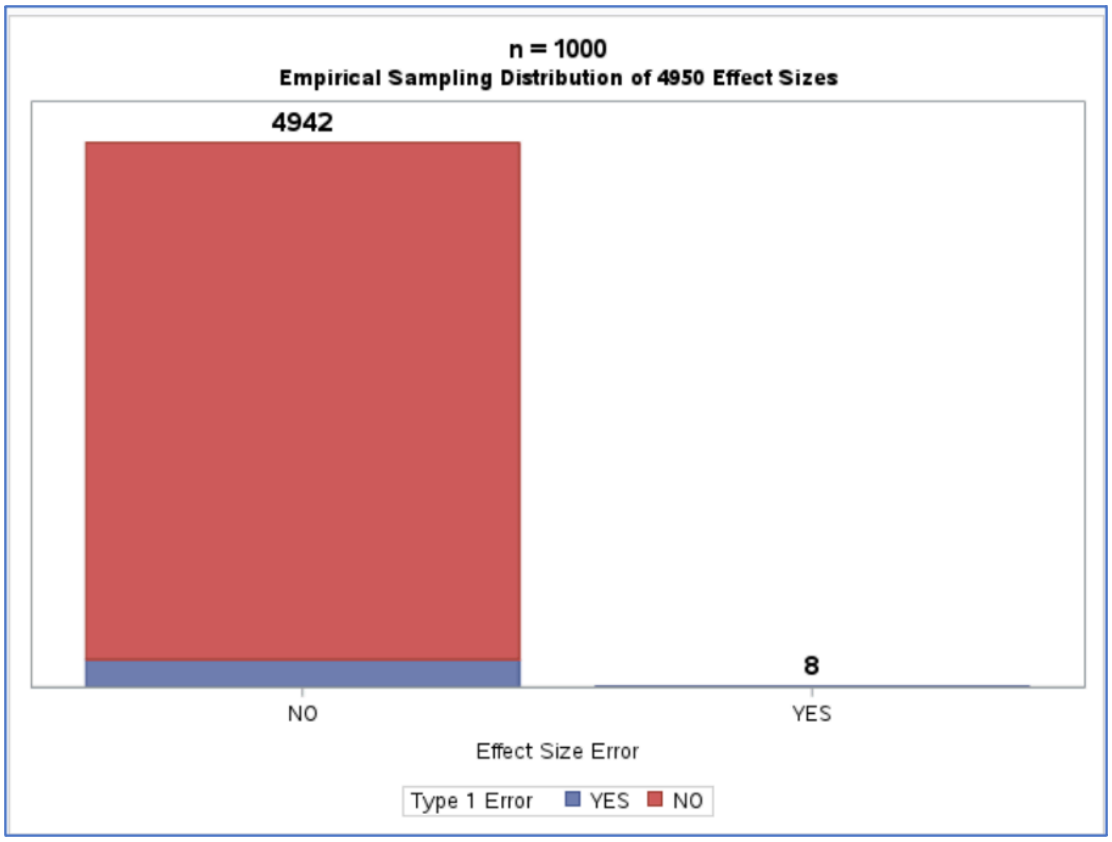

Figure 20.

Table 7 reveals only eight effect size errors, but now there are 252 none-effect sizes that are also statistically significant.

doi.org/10.32388/PS72PK

n = 1000
Intersection of Effect Size Errors and Type 1 Errors

The FREQ Procedure

Frequency Row Pct	Table of Effect Size Error by Type 1 Error		
	Type 1 Error		
Effect Size Error	YES	NO	Total
NO	252 5.10	4690 94.90	4942
YES	8 100.00	0 0.00	8
Total	260	4690	4950

Table 7.

Table 8 reveals that these statistically significant correlations were small effect sizes only.

n = 1000
Statistically Significant Effect Size Errors

The MEANS Procedure

Analysis Variable : Corr Sample Correlation		
N	Minimum	Maximum
260	-0.11	0.11

Table 8.

Previously, no non-effect sizes were detected as statistically significant. This reveals that with n=1000, statistical significance is no longer a useful tool under a true null hypothesis.

An increase in sample size to 2000 revealed only non-effect sizes materializing by chance under a true null hypothesis: $\rho_0 = 0$. Figure 21 shows the sampling distribution of Pearson correlations, with a range of -.08 to +.08, which is below Cohen's threshold of $|r| > .10$.

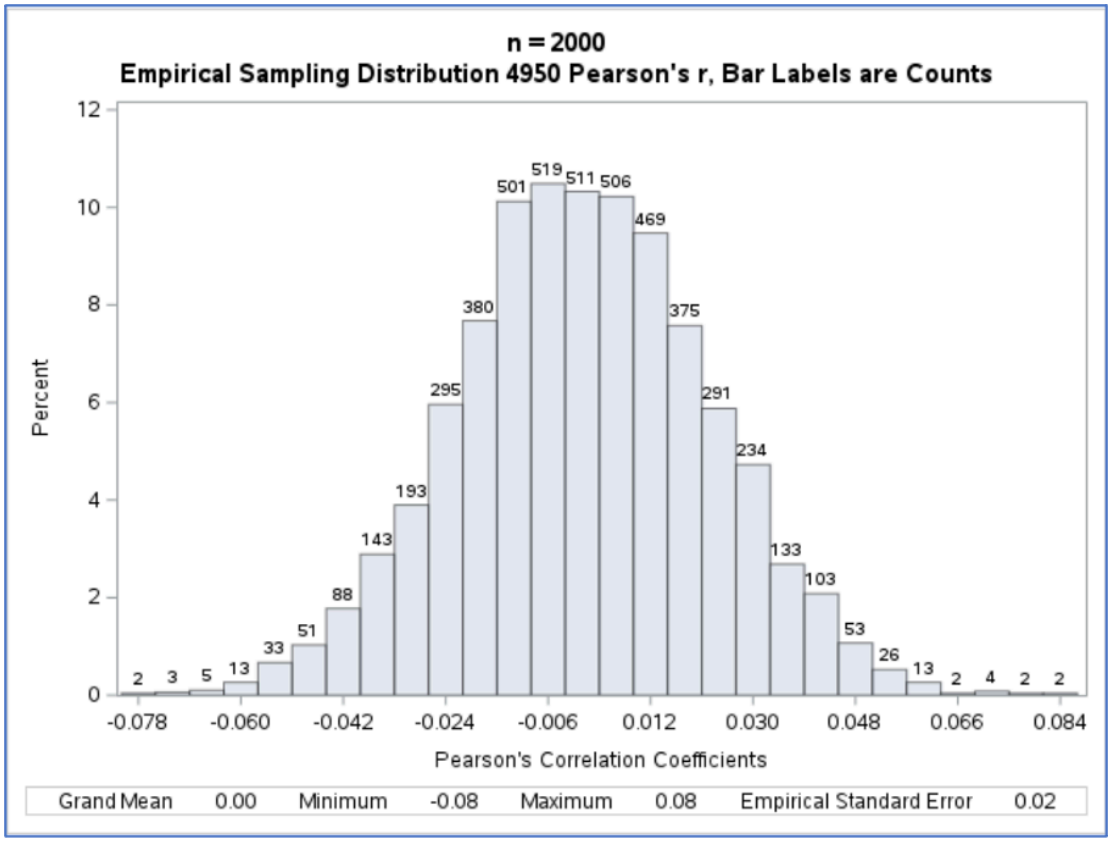

Figure 21.

Figure 22 shows the empirical sampling distribution of zr, which of course also has the same descriptive statistics as Figure 21

doi.org/10.32388/PS72PK

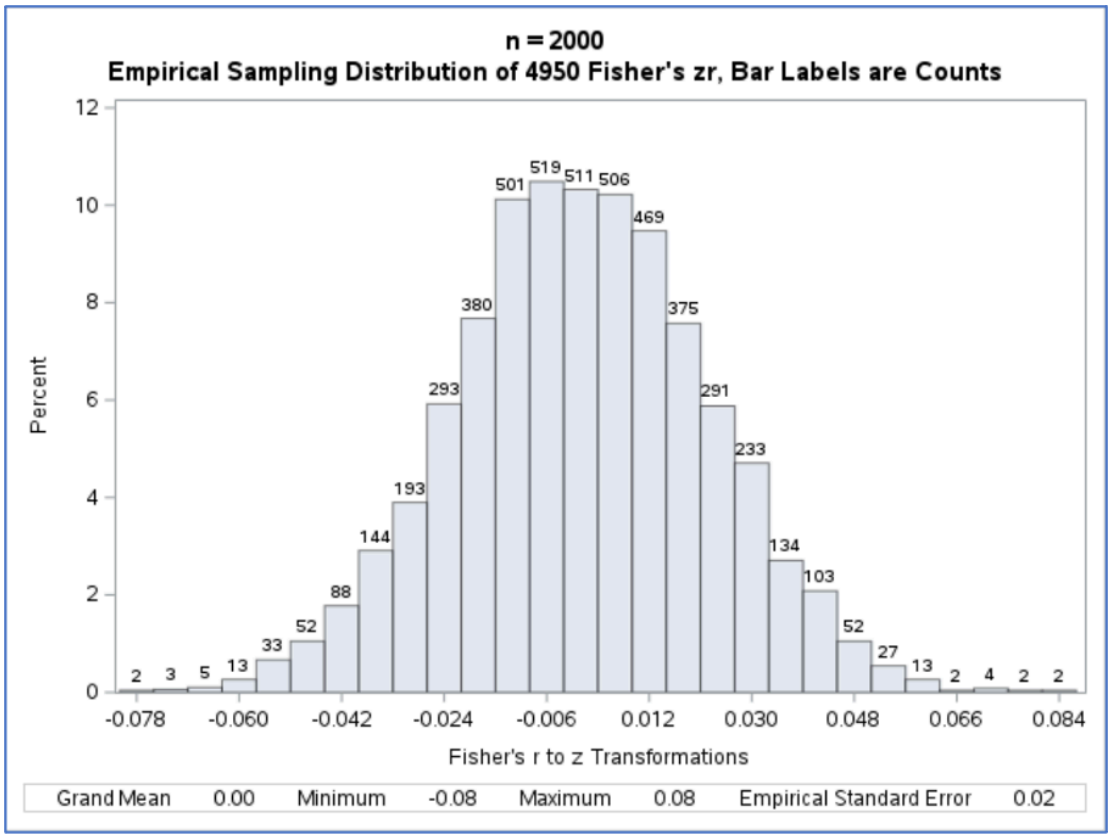

Figure 22.

Figure 23 has the empirical sampling distribution of p-values.

doi.org/10.32388/PS72PK

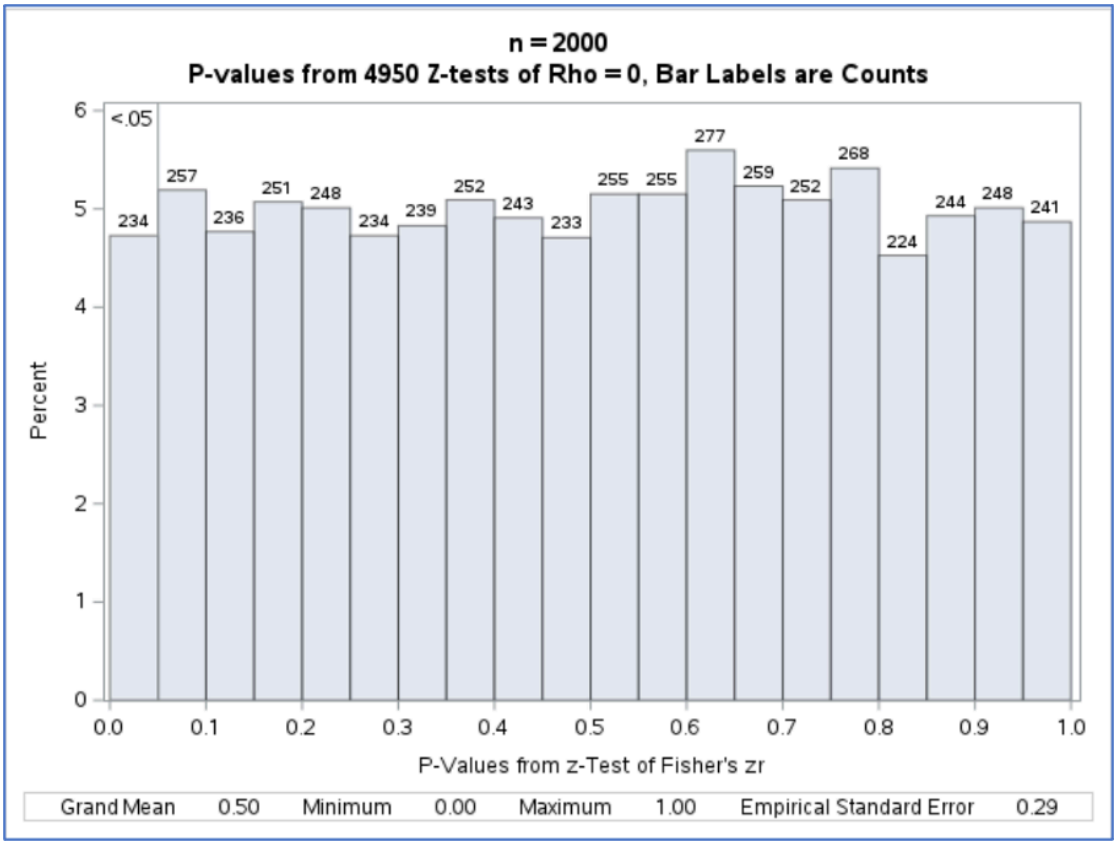

Figure 23.

Approximately 5% (234) of the p-values were statistically significant. Figure 24 conveys the same information as Figure 21, revealing that all correlations are non-effect sizes.

doi.org/10.32388/PS72PK

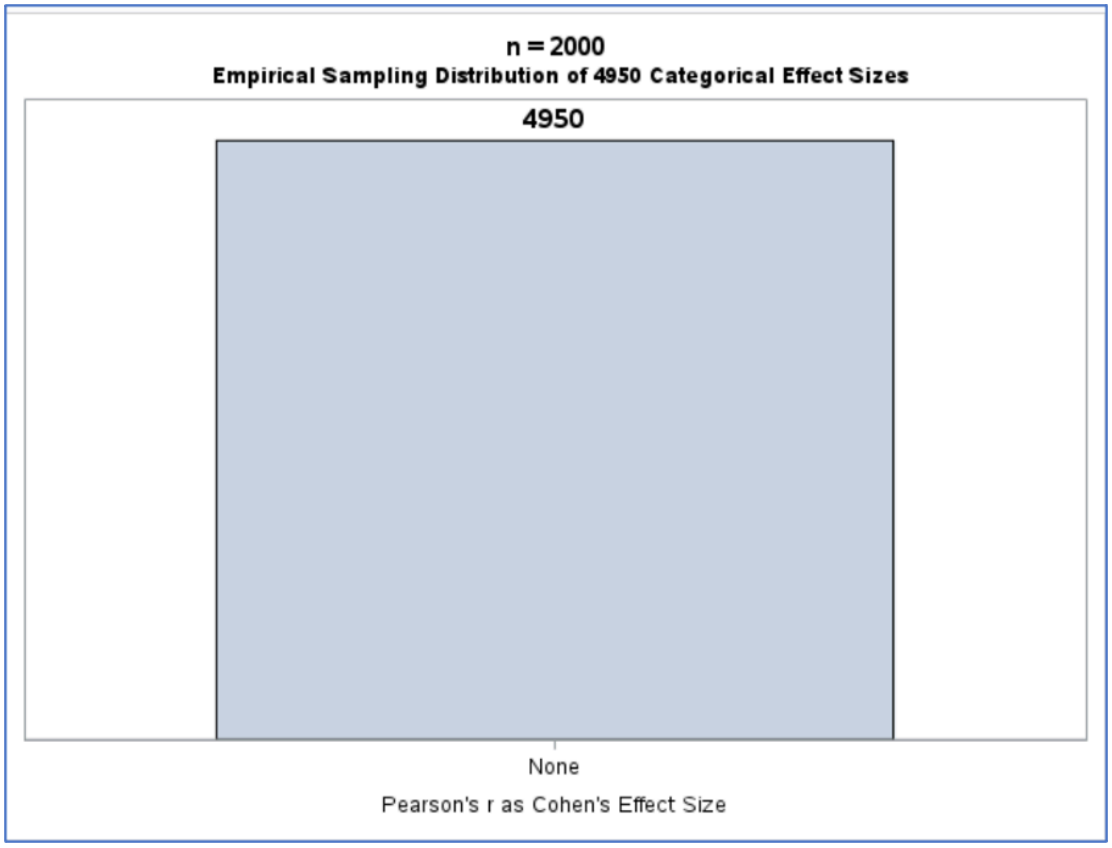

Figure 24.

Figure 25 indicates that relatively few were statistically significant among the non-effet sizes.

doi.org/10.32388/PS72PK

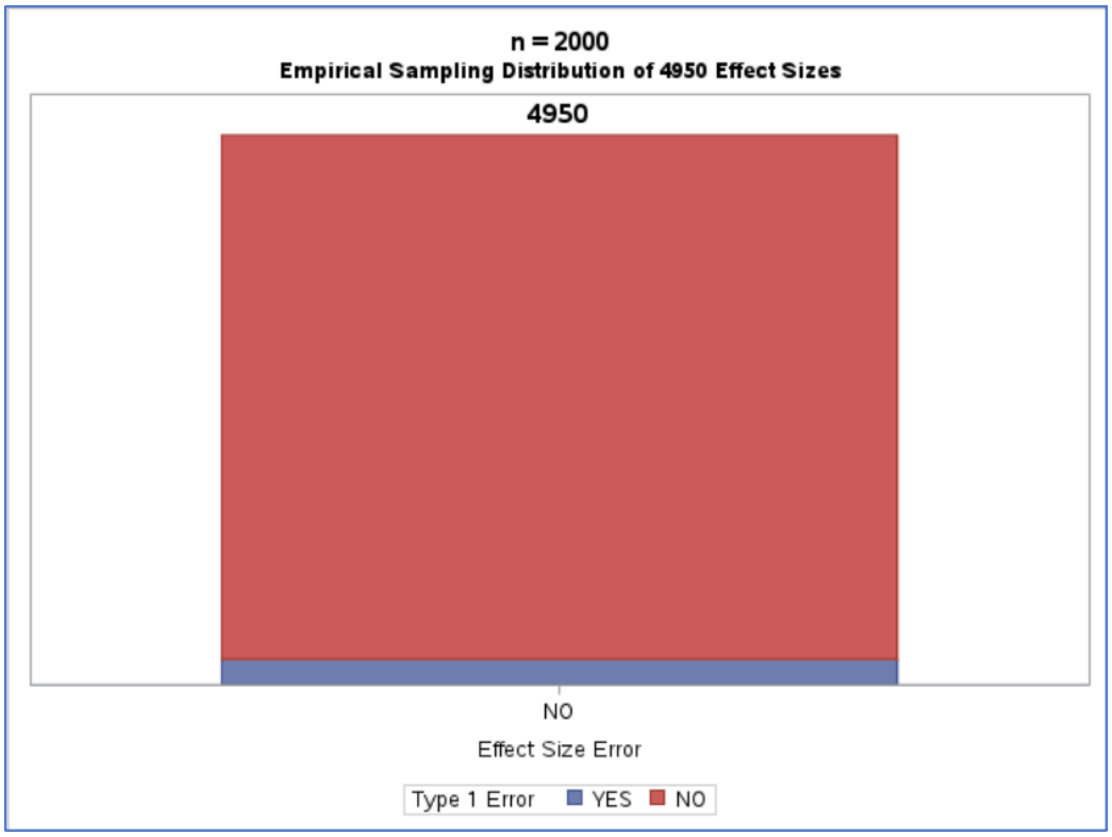

Figure 25.

Table 9 reveals that approximately 5% were statistically significant non-effect sizes.

n = 2000

Intersection of Effect Size Errors and Type 1 Errors

The FREQ Procedure

Frequency
Row Pct

Table of Effect Size Error by Type 1 Error			
	Type 1 Error		
Effect Size Error	YES	NO	Total
NO	234 4.73	4716 95.27	4950
Total	234	4716	4950

Table 9.

doi.org/10.32388/PS72PK

Table 10 confirms that the statistically significant correlation range is 0.08 to +0.08

n = 2000
Statistically Significant Effect Size Errors

The MEANS Procedure

Analysis Variable : Corr Sample Correlation		
N	Minimum	Maximum
234	-0.08	0.08

Table 10.

Conclusion

There are assumptions underlying the significance test of a population correlation, namely bivariate normality, linearity, and no overly influential coordinates. If the assumptions are satisfied, under a true null hypothesis, p-values follow a uniform sampling distribution (Westfall et al., 2011). If the population parameter declared with the null hypothesis is true, any p-value in the open interval from 0.0 to 1.0 can materialize regardless of sample size. More importantly, the percentage of type 1 errors under a true null hypothesis is the constant alpha (e.g., 5%) independent of sample size. In contrast, the percentage of effect size errors under a true null hypothesis is not constant because it decreases with sample size.

Discussion

Provided all assumptions are satisfied, alpha is the 5^{th} percentile value of a uniform sampling distribution of p-values under a true null hypothesis (Westfall et al., 2011). This phenomenon was demonstrated here with empirical sampling distributions. However, to this author's knowledge, no statistical theory predicts the percentage of effect size errors to expect under a true null hypothesis. Incidentally, the parameter specified with the null hypothesis does not have to be zero. Any reasonable value excluding 0.0 and |1.0| can be postulated for the null parameter. However, when the parameter is not zero, the statistical test requires a Fisher r to z transformation to get the propoer p-value because the sampling distribution of correlations is not a symmetric, bell-shaped, normal curve (Fisher, 1970, p. 202).

Imagine a researcher submitting an article to Basic and Applied Social Psychology, which banned statistical significance (Trafimow & Mark, 2015) and relied only on Cohen's effect size criteria to interpret the observed correlation coefficient. With a relatively small sample size and a true null hypothesis, there is a high probability that an effect size error would be misinterpreted as a substantively significant effect size. This scenario is realistic. Fricker et al. (2019) reviewed 31 quantitative research articles published by BASP after the ban on statistical significance and "found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered" (p. 374)

Greenland et al. (2016) stated: "Every method of statistical inference depends on a complex web of assumptions about how data were collected and analyzed, and how the analysis results were selected for presentation" (p. 338) Wasserstein and Lazare (2016) warned against a naïve and single-minded obsession with a statistically significant p-value." Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" (p. 9). Indeed, ignoring these considerations invalidates p-values and, thereby, statistical significance. Still, as stated by Fisher (1973): "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" (p. 103). Statistical significance has been blamed for the replication crisis (Ioannidis, 2005). However, ironically, the solution is not a ban on statistical significance but a replication of statistical significance with fresh new data before publication in a peer-reviewed, high-impact, leading journal.

The author has no conflict of interest with SAS Institute Inc. or any other statistical software company.

References

- Cohen, J. (1968). *Statistical Power Analysis for The Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Fisher, R.A. (1970). *Statistical Methods For Research Workers* (14th ed.). Reprinted in 1990 as *Statistical Methods, Experimental Designs, and Scientific Inference.* Oxford University Press.
- Fricker, Jr., R.D., Burke, K., Han X. & Woodall, W.H. (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.
- Hogg, R.V., McKean, J. W. & Craig, A.T. (2013). *Introduction to Mathematical Statistics*. Pearson Education, Inc.
- Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Med* 2(8): e124.
- 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
- Moore, D.S.,Notz,W.I. & Fligner, M.A.(2021). *The Basic Practice of Statistics* (9th ed.). W. H. Freeman
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. The National Academies Press. https://doi.org/10.17226/25303.
- SAS Institute Inc. (2011). *Base SAS® 9.3 Procedures Guide: Statistical Procedures*. Cary, NC: SAS Institute Inc.
- 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*. SAS Institute Inc.
- Student (1908). Probable error of the 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
- Wasserstein, R.L, & Lazar N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*. DOI: 10.1080/00031305.2016.1154108
- Wasserstein, R.L, Schirm A.L. & Lazar N.A. (2019). Moving to a world beyond $p < 0.05$. *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. Press. https://doi.org/10.17226/25303.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.