v2: 23 June 2025

Research Article

A Redemption Song for Statistical Significance

Preprinted: 11 June 2024
Peer-approved: 27 May 2025

Qeios, Vol. 7 (2025)
ISSN: 2632-3834

Eugene Komaroff[1]

1. Keiser University, Fort Lauderdale, United States

Disagreement is a common occurrence in the field of statistics. In the last century, Ronald Fisher focused on the data-generating probability model known as the null hypothesis. Jerzy Neyman and Egon Pearson elaborated on Fisher's null model, incorporating alternative data-generating probability models. Bayesians, such as Harold Jeffreys, mathematically combined subjective probabilities with the objective ones derived from the data. In the current century, these classical methodologies have been surpassed by modern, computer-intensive machine learning algorithms utilizing massive datasets, requiring implementation with advanced calculus and interpretation informed by domain-specific knowledge. This paper does not attempt to unify statistical theories, predict the future of statistical science, claim superiority for any methodology, or advocate for a radical methodological paradigm shift to qualitative research. This paper focuses on Fisher's statistical significance and the null hypothesis model. Computer-simulated data sets with different sample sizes were used to test a true null hypothesis of zero difference between two independent population means with independent samples t-tests. Statistical significance was determined with a 5% cut score for p-values, and substantive significance was evaluated with Cohen's "effect size index d." The results demonstrate that statistical significance is a viable tool for filtering out false effect sizes (effect size errors) that would otherwise be misinterpreted as substantively significant.

Corresponding author: Eugene Komaroff, komaroffeugene@gmail.com

Cox[1] opined that criticism of statistical significance fills volumes. (see https://en.wikipedia.org/wiki/Statistical_hypothesis_test for an overview). There has been misuse and abuse of statistical significance as catalogued by Greenland et al.[2]. Consequently, some have banished the use of statistical significance[3][4]. That is a mistake. Resolving the problem requires proper education. There is no attempt here to unify the three classical statistical theories[5], promote modern machine learning methods, or call for a radical methodological paradigm shift to qualitative research. This paper illustrates, using graphs and a few numerical examples, that Fisher's[6] statistical significance needs to be understood as a viable screening tool for filtering out misleading substantively significant effect sizes when working with relatively small sample sizes.

Student[7] identified three foundational assumptions essential for ensuring valid t-test results: (1) independent observations, (2) homogeneity of variances, and (3) normally distributed dependent variables. Independent observations are dictated by research design, while the latter two can be assessed using statistical methods. For instance, normality can be evaluated with the Shapiro-Wilks test, and homogeneity of variance can be tested through Levene's test or an F-Max ratio. If any assumption is violated, nonparametric tests, such as the Wilcoxon Rank-Sum test, may serve as alternatives. Typically, textbooks address statistical significance by directing the reader to "assume the null hypothesis is true." However, unlike the assumptions of the t-test, researchers should aim to reject the assumption that "the null hypothesis is true." This might seem counterintuitive to those unfamiliar with sampling distribution theory. This paper aims to elucidate this theory through graphs and a few numerical examples, avoiding reliance on mathematical proofs that require advanced calculus.

Student[7] stated: "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. Now, any series of experiments is only of value insofar as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases, the question finally turns on the value of a mean, either directly or as the mean difference between the two quantities." (pp 1-2). Fisher[6] 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). The "populations" for Student and Fisher were not social or physical phenomena; they were the bell-shaped normal distribution of means that exist only in statistical/probability theory.

A histogram of a human trait, such as height, is relatively easy to understand. Counts (or percentages) of people are categorized into intervals based on a graduated measurement scale, such as feet and inches. However, histograms of sampling distributions of means were derived using complex mathematical theorems: the Central Limit Theorem and the Law of Large Numbers[8]. The mathematical proofs can be

doi.org/10.32388/3QQSNC.5

found in textbooks, such as Hogg, et al.[9] and online (e.g., https://online.stat.psu.edu/stat414/lesson/24/24.2). Here, histograms of the sampling distribution were simulated, which do not require calculus for understanding.

There are two key parameters in the sampling distributions of means. One is the central value, the “mean of means” or grand mean, and the other is the standard error of the means. The grand mean is relatively straightforward to understand as the central value of a sampling distribution of means. However, the dispersion of these distributions is a complex calculation that depends on the sample standard deviation and the square root of the sample size. Fisher[6] stated: “The fundamental proposition upon which the statistical treatment of mean values is based is that – If a quantity be normally distributed with variance σ^2, then the mean of a random sample of n such quantities is normally distributed with variance σ^2/n (p. 114). However, because the population standard deviation (σ) is usually unknown in practice, Student[7] derived that a sample variance (s^2) is an unbiased estimate of the population variance (σ^2). Now, the variance of a theoretical sampling distribution of means can be estimated with a random sample of data (s^2/n). As a result, Student’s t-test replaced the z-test, which requires a known population variance. In summary, for a t-test, “assume the null hypothesis is true” requires a researcher to state a central value or grand mean of a theoretical sampling distribution of means. Collect data with a fixed sample size and assume that the standard deviation of a random sample of data is an accurate (unbiased) estimate of the population standard deviation. Finally, one must assume that the theoretical sampling distribution of means follows a standard normal distribution (μ = 0, σ = 1).

Imagine a novel teaching method designed to help sixth-grade elementary school students develop a grade-level academic skill, such as reading, writing, or arithmetic. The researcher hypothesized that a novel teaching method would improve the skill as predicted by a theory of cognitive development. However, the researcher did not know what to expect in terms of improvement. A statistician recommended designing a small (proof of concept), randomized, controlled experiment with two independent groups: experimental (E) and control (C). However, the researcher was bewildered when the statistician hypothesized a zero or no difference in means. The researcher did not fully understand the logic of a null hypothesis test. Two groups of students took the same test at the end of the interventions. In the numerator of the independent samples t-test, there was the difference in sample means subtracted from the difference in their respective population means: $(\bar{y}_E - \bar{y}_C) - (\mu_E - \mu_C)$. The null hypothesis was the puzzling zero difference in the population means $(\bar{y}_E - \bar{y}_C) - (\mu_E - \mu_C)$.. There was no speculation about a possible difference in sample means. The results revealed that students taught with the experimental pedagogy achieved higher test scores ($\bar{y}_E = 73$) than those taught with the traditional method ($\bar{y}_C = 43$); however, the p-value for the mean difference ($\bar{y}_E - \bar{y}_C = 40$) was not statistically significant (p =.2114) with α =.05. The researcher was annoyed with the statistician’s conclusion that the results were not conclusive. The researcher believed the sizeable 40-point difference in favor of the experimental intervention was important. Furthermore, this difference divided by the pooled (average) standard deviation was a very large effect size (Cohen’s d = 1.57). The researcher decided to ignore the lack of statistical significance. In the write-up, the researcher belittled the mysterious 5% p-value theory under a “true null hypothesis” and enthusiastically speculated about the effectiveness of the novel pedagogy. If you would do the same, please continue reading, as you may also be chasing an illusion down a dark rabbit hole.

Methodology

Sampling distributions of means were simulated using the free online statistical software called “SAS OnDemand for Academics”[10]. Two variables (y_E and Y_C) were randomly sampled 1,000 times from the same normal distribution: N (μ = 50, σ = 25). The process was replicated eight times with the following sample sizes: n = 5, 15, 30, 64, 100, 250, 500, 1000 per group. The difference in population means was tested for statistical significance with independent samples t-tests[11]. The SAS output provided descriptive summary statistics, including sample sizes, means, and standard deviations, as well as inferential statistics, such as t-values, degrees of freedom, and p-values. The null hypothesis of zero difference in population means was true because $\mu_E = \mu_C = 50$, and the variances were equal: $\sigma_E = \sigma_C = 25$.

Statistical Significance

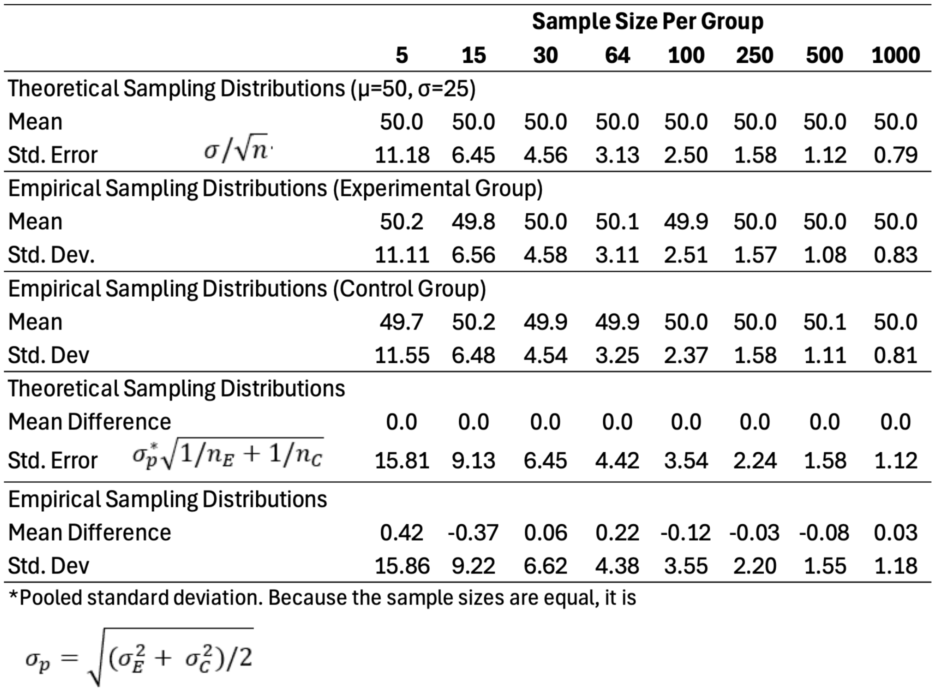
The level of statistical significance was 5% (α =.05). An indicator variable was used to count the number of statistically significant p-values (p <.05) in a sampling distribution of p-values. The indicator was coded as “1” if p < α; otherwise, it was “0.” The count (percentage) of statistically significant p-values in 1,000 p-values was an empirical estimate of the conventional type 1 error rate of 5% under a true null hypothesis. Notice the null hypothesis was not merely assumed but was known to be true. As a result, all statistically significant p-values were type 1 errors (false rejections of a true null hypothesis).

Substantive Significance

Cohen's effect size d was computed by dividing the difference in two independent means by the pooled (average) standard deviation. Cohen[12] recommended the following categories for substantive significance: |d| < 0.20 was considered trivial, |d| ≥ 0.20 to 0.49 was classified as small, |d| ≥ 0.50 to 0.79 was categorized as medium, and |d| ≥ 0.80 was deemed a large effect size. The two vertical lines around d indicate that both negative and positive values were computed, as the t-tests were two-sided. The percentages of substantively significant effect sizes were captured with an indicator variable coded "1" if Cohen's |d| ≥ 0.20 (either small, medium, or large); otherwise, it was "0." Recognize that all the substantively significant (non-trivial) d's were "effect size errors" under the true null hypothesis because the population Cohen's D was zero [(50 − 50) / 25 = 0.00].

Results

Table 1 has descriptive summary data. The parameters of the data-generating probability distribution were μ = 50 and σ = 25. The standard errors (St. Error) of the theoretical sampling distributions differed only by sample size ($\sqrt{n}$). The empirical standard deviations (St. Dev.) were estimates of the theoretical standard errors. The empirical grand means differ slightly from the theoretical grand means because of random sampling. The close agreement between theoretical and empirical data validates the coding and analyses.



		Sample Size Per Group							
		5	15	30	64	100	250	500	1000
Theoretical Sampling Distributions (μ=50, σ=25)									
Mean		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Std. Error	$\sigma/\sqrt{n}$	11.18	6.45	4.56	3.13	2.50	1.58	1.12	0.79
Empirical Sampling Distributions (Experimental Group)									
Mean		50.2	49.8	50.0	50.1	49.9	50.0	50.0	50.0
Std. Dev.		11.11	6.56	4.58	3.11	2.51	1.57	1.08	0.83
Empirical Sampling Distributions (Control Group)									
Mean		49.7	50.2	49.9	49.9	50.0	50.0	50.1	50.0
Std. Dev		11.55	6.48	4.54	3.25	2.37	1.58	1.11	0.81
Theoretical Sampling Distributions									
Mean Difference		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Std. Error	$\sigma_p^*\sqrt{1/n_E + 1/n_C}$	15.81	9.13	6.45	4.42	3.54	2.24	1.58	1.12
Empirical Sampling Distributions									
Mean Difference		0.42	-0.37	0.06	0.22	-0.12	-0.03	-0.08	0.03
Std. Dev		15.86	9.22	6.62	4.38	3.55	2.20	1.55	1.18

*Pooled standard deviation. Because the sample sizes are equal, it is

$$\sigma_p = \sqrt{(\sigma_E^2 + \sigma_C^2)/2}$$

Table 1. Theoretical and Empirical Sampling Distributions of Means

Figure 1 shows the histograms of sampling distributions of observed mean differences by sample sizes. The grand means of the sampling distributions closely approximate the zero difference in population means. The standard deviations (standard errors) of the empirical sampling distributions decrease as the sample size increases. The shrinkage in dispersion can also be discerned from the range (maximum minus minimum).

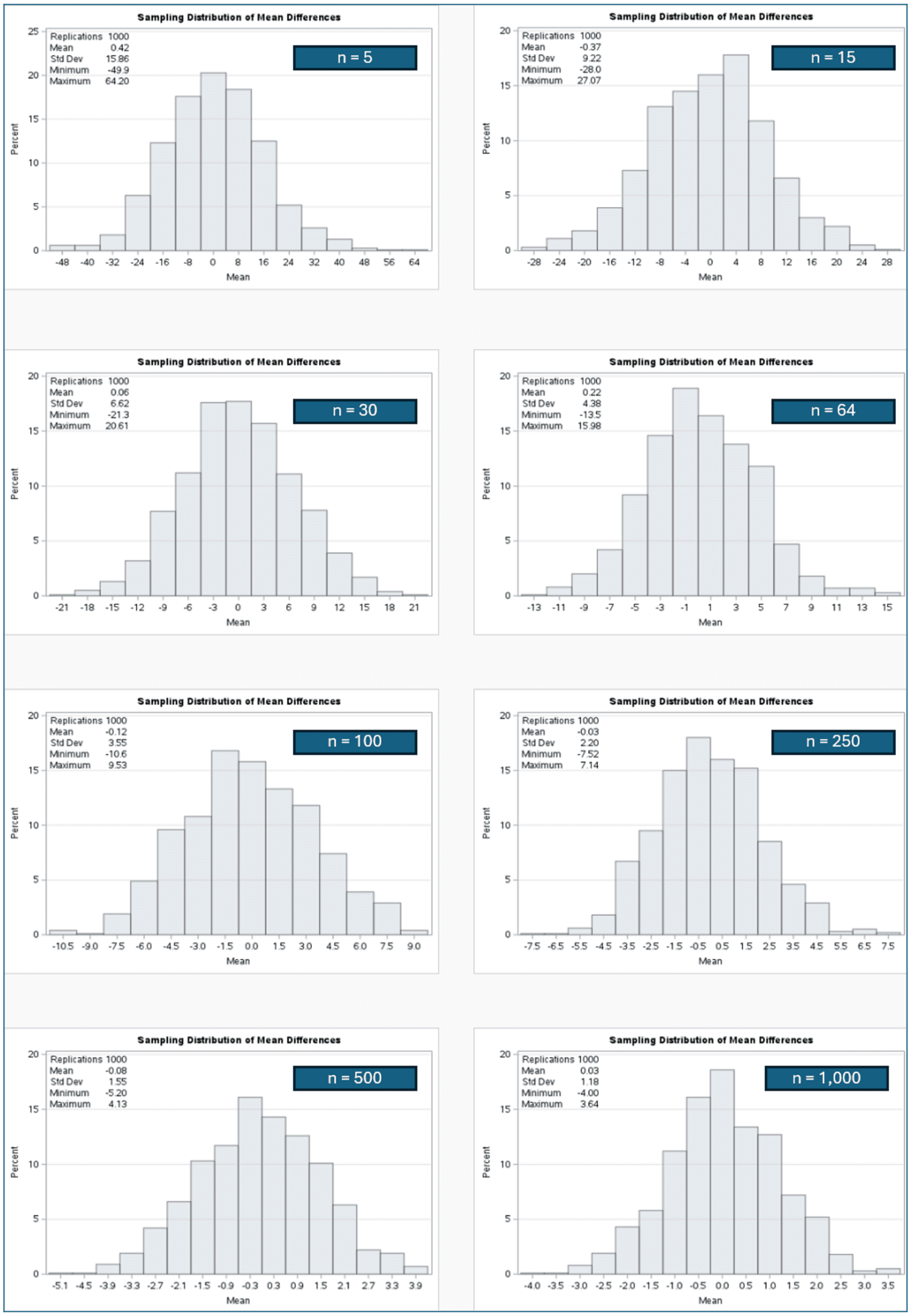


Figure 1. Sampling Distribution of Mean Differences under a True Null Hypothesis.

Figure 2 provides empirical support for Fisher's null hypothesis paradigm: Every p-value has an equal chance of occurring under a true null hypothesis[13]. The empirical histograms are approximately uniform (rectangular) but would be perfectly uniform with infinitely countable replications.

doi.org/10.32388/3QQSNC.5

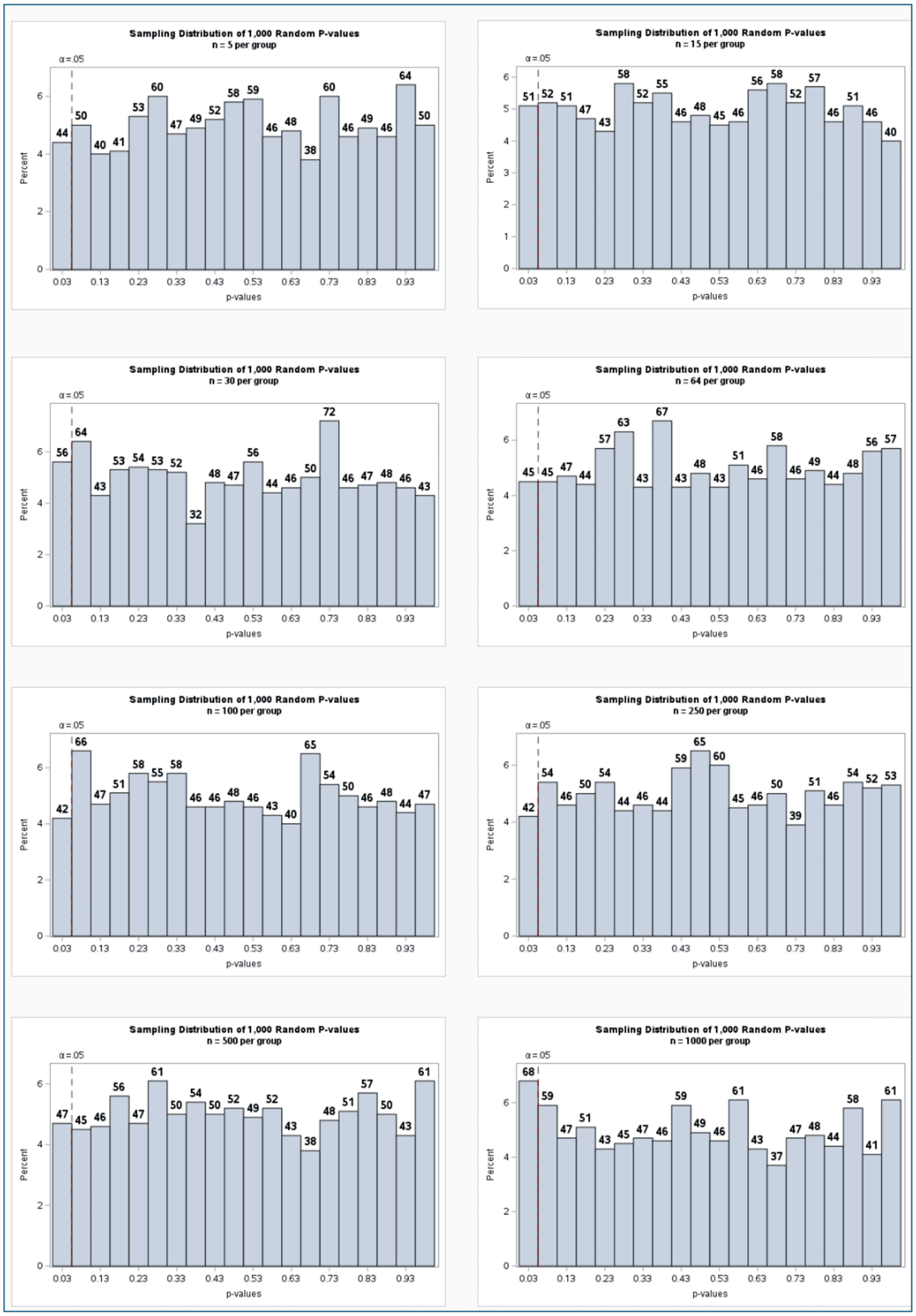


Figure 2. Histograms of p-values from independent samples t-tests of zero difference in population means.

Figure 3 shows distributions of Cohen's d as continuous effect sizes, highlighting the statistically significant ones in the tails. As sample sizes increased, the d's converged (became better estimates) of the population D = 0.00. As a result, the two reference lines denoting the substantively significant effect sizes (|d| ≥ 0.20) appear to be moving farther apart.

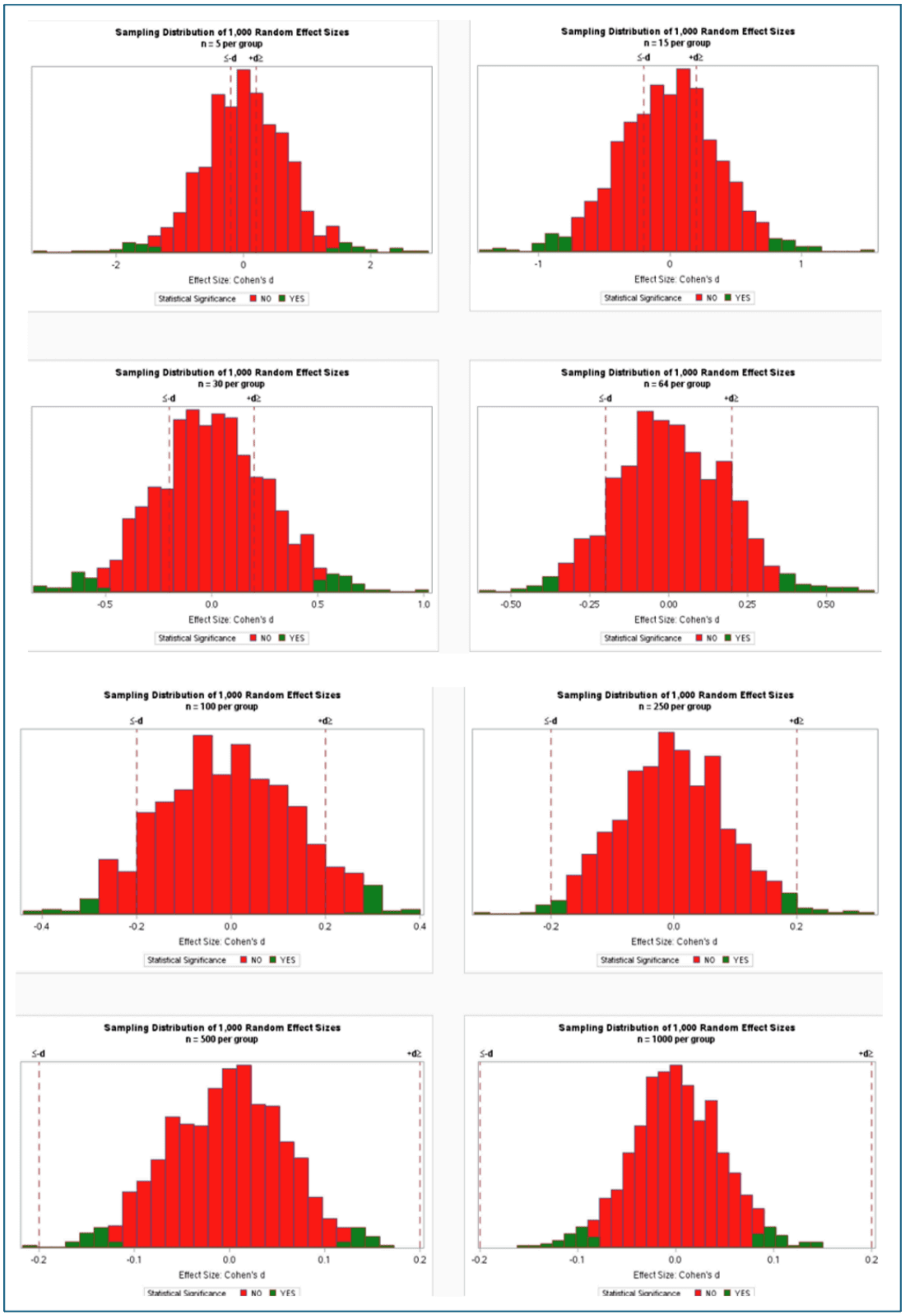


Figure 3. Sampling Distributions of Continuous Cohen's d under a True Null Hypothesis

The bar graphs in Figure 4 were created by grouping the continuous d's according to Cohen's[12] criteria: |d|'s < 0.20 are trivial, |d|'s ≥ 0.20 and < 0.49 are small, |d|'s ≥ 0.50 and < 0.79 are medium, and |d|'s ≥ 0.80 are large effect size. With an increasing sample size, the percentage (count) of statistically significant p-values remains relatively constant at 5%, but the effect sizes decrease. Finally, with n = 1,000 per group, all effect sizes are statistically significant, but none are substantively significant.

doi.org/10.32388/3QQSNC.5

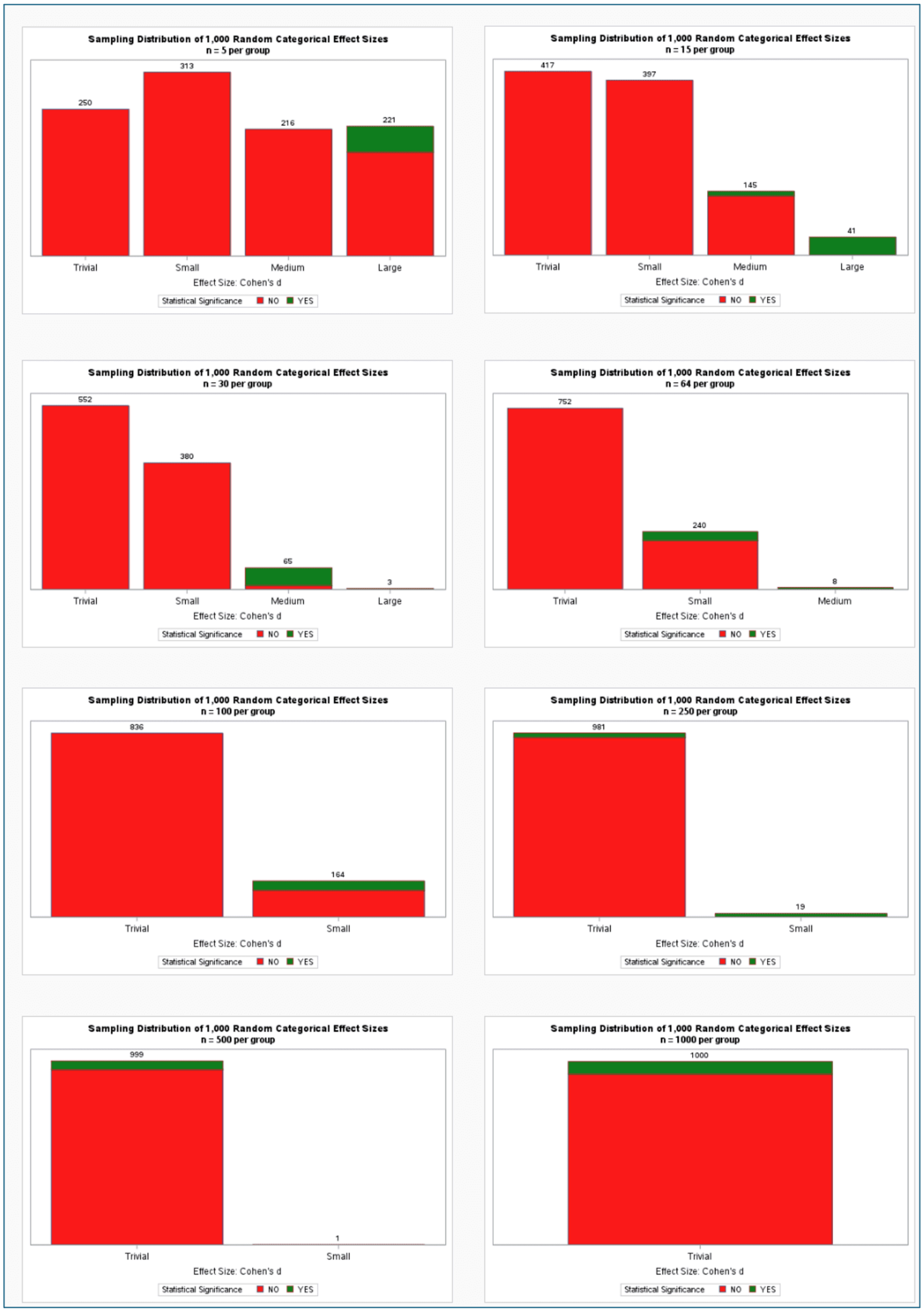
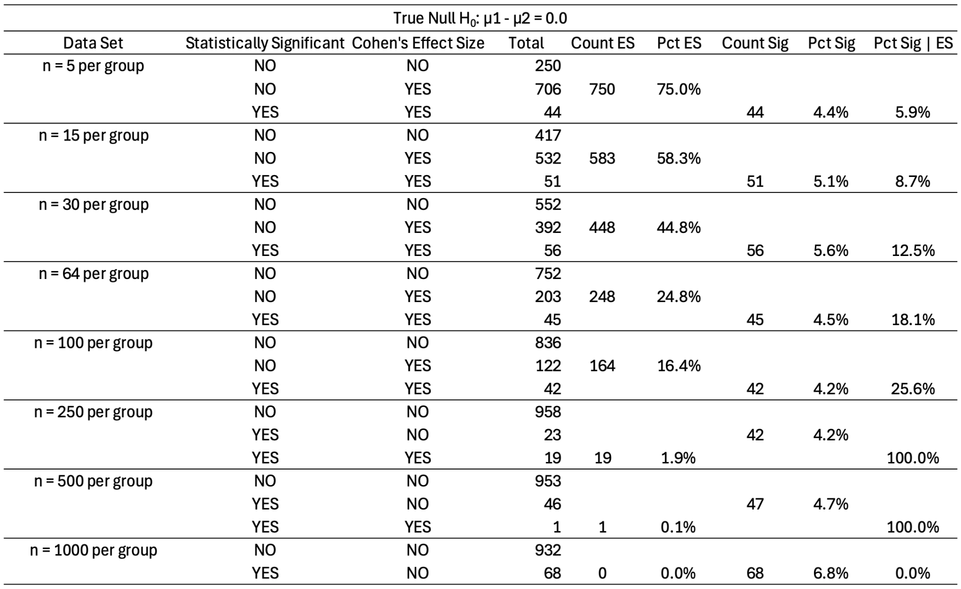


Figure 4. Distributions of Effect Sizes according to Cohen's Criteria and Statistical Significance

Table 2 provides the counts and percentages of statistically significant p-values and substantively significant effect sizes.



True Null H_0: μ1 - μ2 = 0.0								
Data Set	Statistically Significant	Cohen's Effect Size	Total	Count ES	Pct ES	Count Sig	Pct Sig	Pct Sig | ES
n = 5 per group	NO	NO	250					
	NO	YES	706	750	75.0%			
	YES	YES	44			44	4.4%	5.9%
n = 15 per group	NO	NO	417					
	NO	YES	532	583	58.3%			
	YES	YES	51			51	5.1%	8.7%
n = 30 per group	NO	NO	552					
	NO	YES	392	448	44.8%			
	YES	YES	56			56	5.6%	12.5%
n = 64 per group	NO	NO	752					
	NO	YES	203	248	24.8%			
	YES	YES	45			45	4.5%	18.1%
n = 100 per group	NO	NO	836					
	NO	YES	122	164	16.4%			
	YES	YES	42			42	4.2%	25.6%
n = 250 per group	NO	NO	958					
	YES	NO	23			42	4.2%	
	YES	YES	19	19	1.9%			100.0%
n = 500 per group	NO	NO	953					
	YES	NO	46			47	4.7%	
	YES	YES	1	1	0.1%			100.0%
n = 1000 per group	NO	NO	932					
	YES	NO	68	0	0.0%	68	6.8%	0.0%

Table 2. Count and Percentages of Statistically Significant P-values and Substantively Significant Effect Sizes (Cohen's d) under a True Null Hypothesis by Sample Size.

With n = 5 per group, 44 (4%) of the 1,000 p-values were statistically significant (p <.05). However, all of these were type 1 errors because the null hypothesis was true. A type 1 error is the probability of "rejecting a true null hypothesis." Similarly, 750 Cohen's d were substantively significant (small, medium, or large effect sizes), but these were all "effect size errors" because Cohen's D = 0.00. The column labeled "Pct Sig | ES" indicates the percentage of statistically significant p-values given that the effect size was substantively significant. Statistical significance filtered out 94% of the effect size errors, leaving only a small portion (6%) for consideration. As the sample size increased, the percentage of substantively significant effect sizes decreased; however, all were statistically significant until n = 250 per group. At this point, only 19 were both statistically and substantively significant, whereas the remaining 42 statistically significant effect sizes were trivial. This phenomenon became more pronounced with n = 500 per group, where only one substantive effect size was statistically significant, whereas the remaining 49 were statistically significant but trivial effect sizes. Finally, statistical significance was useless with n = 1,000 per group (total n = 2,000) because all 68 statistically significant effect sizes were trivial.

Statistical Significance and Cohen's d Under a False Null Hypothesis

The false null hypothesis was created by adding 20 points to the original experimental variable (Y_E). This is similar to sampling Y_E from a normal population where μ_E = 70 and σ_E = 25. The Y_C (control) variable was again randomly sampled from μ_C = 50 and σ_C = 25. The null hypothesis $(H_0 : \mu_{\bar{y}E} - \mu_{\bar{y}C} = 0)$ was now false, and the alternative hypothesis was true $(H_a : \mu_{\bar{y}E} - \mu_{\bar{y}C} = 20)$. However, it was the null hypothesis that was tested for statistical significance, not the alternative hypothesis. Additionally, the population Cohen's d = (20/25) = 0.80 was now a large effect size.

Figure 5 reveals right-skewed p-value distributions that are expected under a false null hypothesis. With α =.05, as the sample size increased, more than 5% of the p-values were statistically significant, and none were type 1 errors. The histograms with n > 64 per group are not shown because all 1,000 p-values were clustered in the 5th percentile (see Table 2).

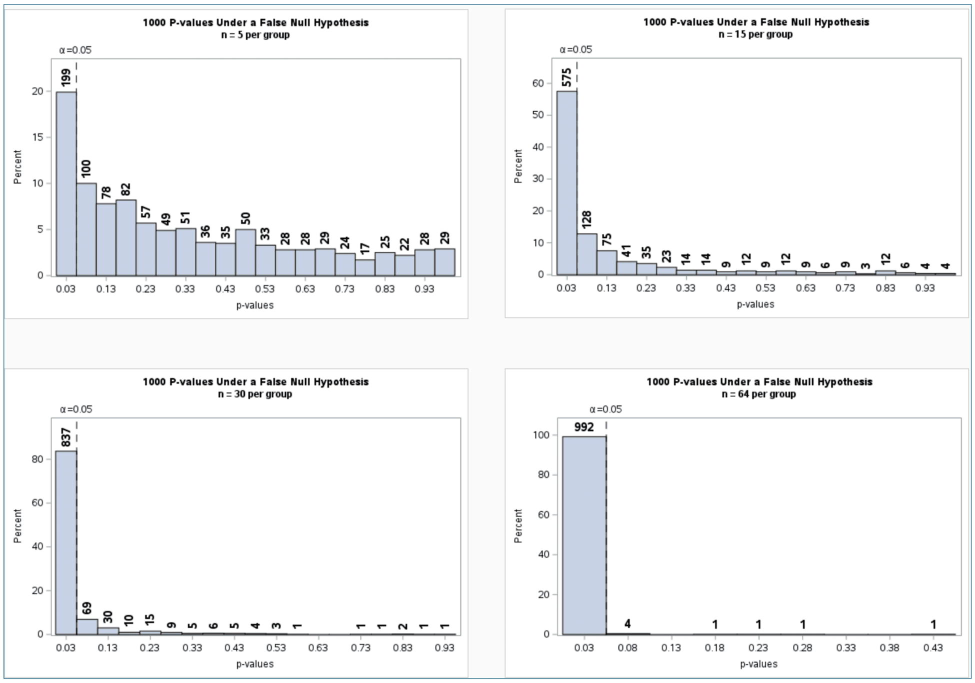
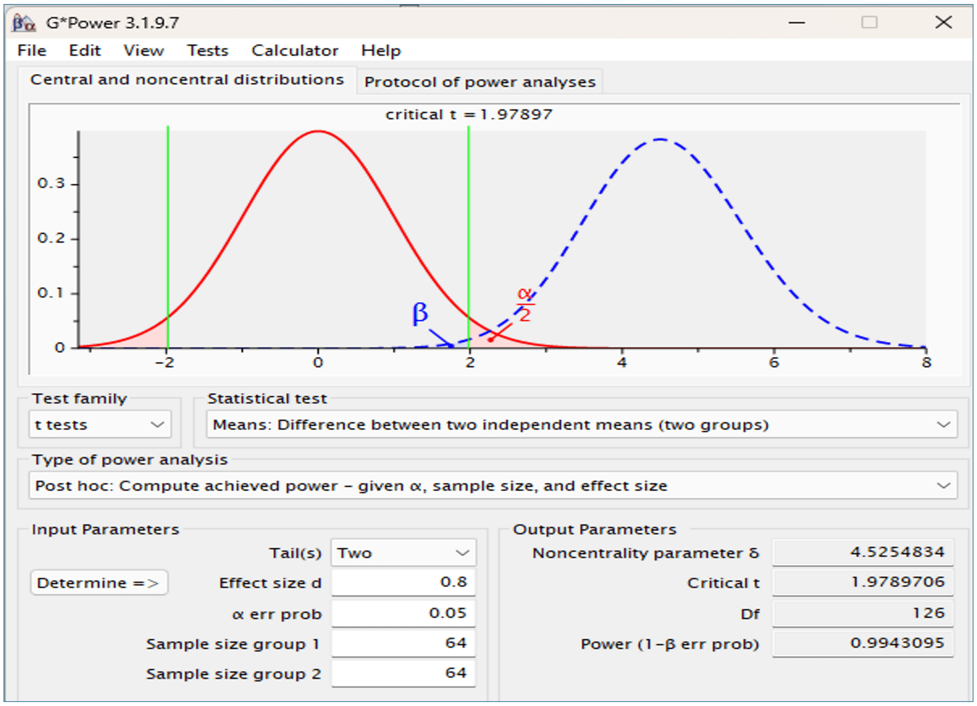


Figure 5. P-value Sampling Distributions Under a False Null Hypothesis

The histogram with n = 64 per group (total n = 120) reveals that 99% of the p-values were statistically significant. This indicates that with n = 64 per group and α =.05, there is 99% power to reject the assumption: “null hypothesis is true.” G*Power[14] confirms the empirical analysis (Fig. 6).



Test family: t tests

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

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters

Tail(s)	Two
Effect size d	0.8
α err prob	0.05
Sample size group 1	64
Sample size group 2	64

Output Parameters

Noncentrality parameter δ	4.5254834
Critical t	1.9789706
Df	126
Power (1-β err prob)	0.9943095

Figure 6. Power to Reject A Null Hypothesis with an Independent Samples T-Test for an Effect Size D = 0.80 and Alpha =.05.

Conclusion

Moore et al. (2022) stated: “The smaller the P-value, the more substantial the evidence against H_0 provided by the data” (p 387). However, “more substantial” is confusing in two ways. P-values should not be used to interpret substantive significance, and adding adjectives, like “more” to statistically significant, fuels misinterpretation. Once the parameter under the null hypothesis has been rejected ($p < \alpha$), there is nothing more to say about statistical significance. The null parameter is a fixed constant, not a random variable as in the Bayesian paradigm. After determining statistical significance, attention must shift to an alternative sampling distribution with a similar dispersion (standard error) but a different grand mean. Researchers typically estimate the alternative grand mean by placing a confidence interval around the observed (sample) mean, where the confidence level is typically 90%, 95%, or 99%. However, the confidence interval indicates precision. There is no question in statistical theory that the mean is an unbiased estimator of the population mean[15]; however, that does not guarantee that an observed sample mean is an accurate estimate of the population (or grand) mean, particularly with small sample sizes (as demonstrated here).

Without statistical significance, many effect size errors are liable to be misinterpreted as substantively significant. Consequently, a ban on statistical significance implies that the scientific research literature will be inundated with even more irreplicable results than have been attributed to the misuse and abuse of statistical significance (Ioannidis, 2005). Ironically, replication will help resolve the replication crisis[16]; however, replication should not be confused with mere reproduction. The report from the National Academies of Sciences, Engineering, and Medicine (2019) explained: “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 make 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). Fisher[17] also called for replication because statistical significance was a signpost and not the final destination: “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” (p. 103). Furthermore, Fisher believed that a level of statistical significance is required, but “no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.”[17].

In conclusion, I hope the results in this paper have convinced the reader that, regardless of a chosen α-level, statistical significance remains a viable tool when running independent samples t-tests with small sample sizes. If the null hypothesis is true, many false effect sizes are excluded from further consideration. For independent samples t-tests with large sample sizes ($n \geq 2{,}000$), regardless of statistical significance, substantive effect sizes are unusual under a true null hypothesis and thus merit scrutiny for scientific plausibility. In the scenario presented at the beginning of this paper, the researcher should have been pleased with the statistical analysis. The next step would have been to design and conduct a new, appropriately powered, randomized, controlled experiment.

Statements and Declarations

Reproducibility

The author guarantees that the results in this paper are reproducible and replicable. Reproducible because the data sets were saved to a hard drive, and replicable because the SAS program uses computer clock time to initiate the random data streams. Please try to simulate your own sampling distributions of mean differences (effect sizes) and their corresponding p-values from independent samples t-tests under a true null hypothesis. You may also agree with Mayo and Hand[18]: “Recommendations to replace, abandon, or retire statistical significance undermine a central function of statistics in science: to test whether observed patterns in the data are genuine or due to background variability” (p. 219).

Conflicts of interest

The author has no conflicts of interest to disclose.

Data Availability

The SAS program code used to perform the simulations and data analysis described in this paper is available from the corresponding author, Eugene Komaroff (komaroffeugene@gmail.com), upon reasonable request. The author states that the SAS program uses computer clock time to initiate the random data streams, therefore enabling statistical replication of the findings. The primary datasets generated during the study and all analysis files to reproduce the published figures and tables in this manuscript are also available from the corresponding author upon reasonable request. You can also download the data sets and SAS programs (code) to reproduce and replicate the results at https://figshare.com

doi.org/10.32388/3QQSNC.5

Author Contributions

EK: Conceptualization, Methodology, Software, Validation, Programming and Statistical Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing, Tabulating & Graphing.

References

1. ^*Cox DR (1982). "Statistical significance tests." Br. J. clin. Pharmac.* ***14****:325–331.*
2. ^*Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016). "Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations." European Journal of Epidemiology.* ***31****:337–350.*
3. ^*Trafimow D, Marks M (2015). "Editorial." Basic and Applied Social Psychology.* ***37****(1):1–2. doi:10.1080/01973533.2015.1012991.*
4. ^*Wasserstein RL, Schirm AL, Lazar NA (2019). "Moving to a world beyond pThe American Statistician.* ***73****(sup1):1–19.*
5. ^*Efron B (1998). "R. A. Fisher in the 21st century (Invited paper presented at the 1996 R. A. Fisher Lecture)." Statistical Science.* ***13****(2):95–122. doi:10.1214/ss/1028905930.*
6. a, b, c*Fisher RA (1970). Statistical Methods for Research Workers. 14th ed. Oxford University Press.*
7. a, b, c*Student (1908). "The probable error of a mean." Biometrika.* ***6****(1):1–25.*
8. ^*Moore DS, Notz WI, Fligner M (2021). Basic Practice of Statistics. 9th ed. Macmillan Learning.*
9. ^*Hogg RV, McKean JW, Craig AT (2005). Introduction to Mathematical Statistics. 5th ed. Prentice Hall. http://www.pearsonhighered.com/educator/product/Introduction-to-Mathematical-Statistics/9780130085078.page.*
10. ^*SAS Institute Inc. (2014). SAS® OnDemand for Academics: User's Guide. SAS Institute Inc.*
11. ^*SAS Institute Inc. (2019). SAS/STAT® 9.4 User's Guide. Cary NC: SAS Institute Inc.*
12. a, b*Cohen J (1968). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates. https://books.google.com/books?id=Tl0N2lRAO9oC&printsec=frontcover&dq.*
13. ^*Westfall PH, Tobias RD, Wolfinger RD (2011). Multiple Comparisons and Multiple Tests Using SAS. 2nd ed. SAS Institute Inc.*
14. ^*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.*
15. ^*Fisher RA (1922). "On the mathematical foundations of theoretical statistics." Phil. Trans. R. Soc. Lond. A Containing Papers of a Mathematical or Physical Character.* ***222****:309–368. doi:10.1098/rsta.1922.0009.*
16. ^*Vidgen B, Yasseri T (2016). "P-values: misunderstood and misused." Frontiers in Physics.* ***4****:6.*
17. a, b*Fisher RA (1973). Statistical Methods and Scientific Inference. Hafner Press.*
18. ^*Mayo DG, Hand D (2022). "Statistical significance and its critics: Practicing damaging science, or damaging scientific practice?" Synthese.* ***200****:1–33. doi:10.1007/s11229-022-03692-0.*

Declarations

Funding: No specific funding was received for this work.
Potential competing interests: No potential competing interests to declare.