v1: 27 May 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 typical in the discipline 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 generalized Fisher's null model with alternative data-generating probability models. Bayesians, such as Harold Jeffries, mathematically conjugated subjective probabilities with objective ones derived from the data. In the current century, these classical methodologies have been supplemented with modern computer-intensive machine learning algorithms with massive data sets that require implementation with advanced calculus and interpretation with domain-specific knowledge. This paper does not try to unify the three classical statistical theories, forecast the future of statistical science, claim superiority for any methodology, or call for a radical paradigm shift to textual qualitative research methodology. This paper is focused on Fisher's pervasive statistical significance of a null hypothesis model. Computer-simulated data was used to test a zero difference between independent means under a true null hypothesis. Statistical significance was informed by p-values, and substantive significance was evaluated with Cohen's "effect size index d." The results demonstrate that statistical significance remains a viable tool for filtering out false effect sizes (effect size errors) that might 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 of the controversy. Some have taken an extreme position and called for a ban on statistical significance[2][3]. This author acknowledges the misuse and abuse catalogued by Greenland et al.[4]; However, this is not an attempt to unify the three classical statistical theories[5] nor call for a radical paradigm shift to textual, qualitative research methods. The purpose here is to demonstrate with graphs and a few numbers that Fisher's[6] statistical significance needs to be understood as a viable screening tool for substantive significance when working with relatively small sample sizes (e.g., n < 2,000).

Student[7] stated three underlying assumptions to ensure valid t-test results: (1) independent observations, (2) homogeneity of variances, and (3) normally distributed dependent variables. Independent observations are determined by research design, but the other two can be evaluated with statistical methods. For example, normality can be tested with the Shapiro-Wilks test, and homogeneity of variance can be evaluated with Levene's test or an F-Max ratio. If an

 doi.org/10.32388/3QQSNC.4

assumption is violated, nonparametric tests, such as the Wilcoxon Rank Sums test, can be used instead. Typically, textbooks discuss statistical significance by telling the reader, "assume the null hypothesis is true." However, unlike t-test assumptions, violating the "true null hypothesis " assumption is desirable. This may be counterintuitive for those who do not understand Fisher's sampling distribution theory.

Fisher's paradigm was motivated by Student[7]: "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, but were bell-shaped, or normal distribution of means that existed only in statistical/probability theory.

A histogram of a human trait, such as height, is relatively easy to understand. Counts (percentages) of people are classified into intervals of heights according to a graduated measurement scale, such as feet and inches, or meters and centimeters. However, histograms of sample means were derived with complex mathematical theorems: The Central Limit Theorem and the Law of Large Numbers[8]. The mathematical proofs can be found in textbooks, e.g., Schaeffer[9], and online (e.g., https://online.stat.psu.edu/stat414/lesson/24/24.2). Here, empirical sampling distribution histograms were created and do not require calculus to be appreciated.

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. The grand mean is relatively straightforward compared to the complex standard error. 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). 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). Remarkably, the variance of a theoretical sampling distribution was estimated with a random sample of data (s^2/n). Student's t-test replaced the z-test, which required the population variance. For a t-test, "assume the null hypothesis is true" requires a researcher to assume a central value or grand mean of a theoretical sampling distribution of means with a fixed sample size (i.e., degrees of freedom). Assume that the standard deviation of a random sample of data is an accurate and unbiased estimate of the population standard deviation. Finally, assume the theoretical sampling distribution of means is normally distributed.

Imagine a novel teaching method created to help sixth-grade elementary school students achieve a grade-level academic skill (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 how much improvement to expect. A statistician recommended

designing a small (proof of concept), randomized experiment with two independent groups: an 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 understand the logic of null hypothesis testing, but complied with the statistician. Two groups of students took the same test at the end of the intervention sessions. 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: ($\overline{x}_E -$ \({\overline{x}}_{C})\ - \\)($\mu_E - \mu_C$). The null hypothesis was the puzzling zero difference in the population means (H_0: $\mu_E - \mu_C$ = 0). Nothing was speculated about a difference in the sample means. The results revealed that students taught with the experimental pedagogy achieved higher test scores ($\overline{x}_E$ = 73) than those taught with the traditional method ($\overline{x}_C$ = 43); however, the p-value for the mean difference ($\overline{x}_E - \overline{x}_C$ = 40) was not statistically significant (p = .2114) with α = .05. The researcher became angry when the statistician concluded the results were promising/suggestive but not convincing. The researcher observed a noteworthy difference of 40 points in favor of the experimental intervention. The difference divided by the pooled (average) standard deviation was a huge effect size (Cohen's d = 1.57). The researcher decided to ignore the lack of statistical significance. The researcher did not understand or care about the mysterious p-value under a "true null hypothesis" and enthusiastically speculated about the effectiveness of the novel pedagogy in the write-up. If you agree with the researcher, please keep reading because you may also be chasing illusions 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 (x_E and x_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 data: sample sizes, means, and standard deviations, as well as the inferential statistics: 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 statistically significant p-values (p < .05) in the 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 from 1,000 p-values for each sample size was an empirical estimate of the conventional type 1 error rate of 5% under a true null hypothesis. Notice that 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 a trivial, |d| ≥ 0.20 to 0.49 was a small, |d| ≥ 0.50 to .0.79 was a medium, and |d| ≥ 0.80 was a large effect size. The two vertical lines surrounding d indicate that negative and positive values were computed because 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 substantively significant (non-trivial) d's were "effect size errors" 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 errors. The close agreement between theoretical and empirical data validates the SAS programming 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	50.1	49.8	50.0	50.1	49.9	50.0	50.0	50.0
Std. Dev.	11.10	6.60	4.60	3.10	2.50	1.60	1.10	0.80
Empirical Sampling Distributions (Control Group)								
Mean	49.7	50.2	49.9	49.9	50	50	50.1	50
Std. Dev	11.60	6.50	4.50	3.20	2.40	1.60	1.10	0.80
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	0.4	-0.4	0.1	0.2	-0.1	0.0	-0.1	0.0
Std. Dev	15.86	9.22	6.62	4.38	3.55	2.20	1.55	1.18

Footnote: *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 empirical sampling distributions' standard deviations (standard errors) decrease with

doi.org/10.32388/3QQSNC.4

increased sample size. The shrinkage in dispersion can also be discerned from the range (maximum—minimum).

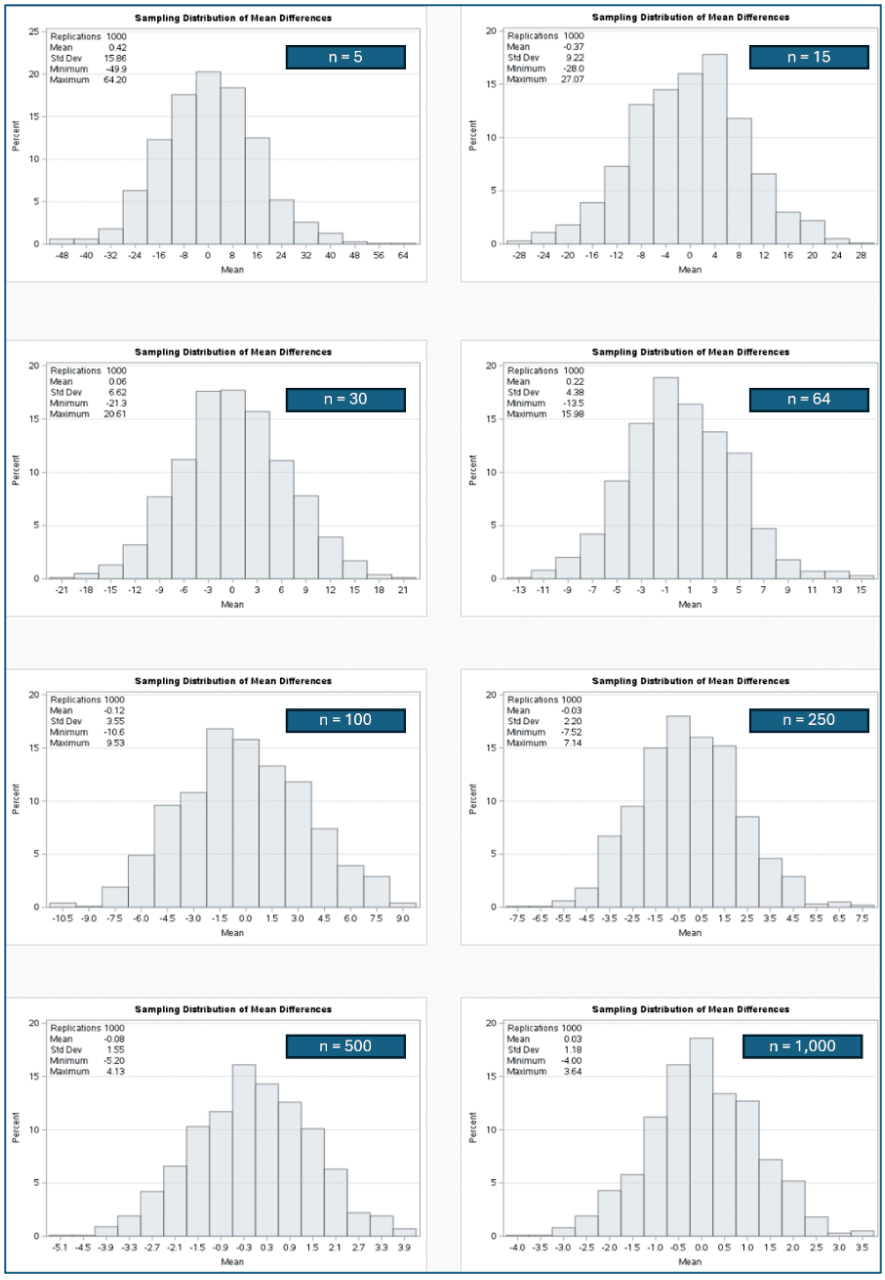

Figure 1. Sampling Distribution of Mean Differences under a True Null Hypothesis.

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 replication.

doi.org/10.32388/3QQSNC.4

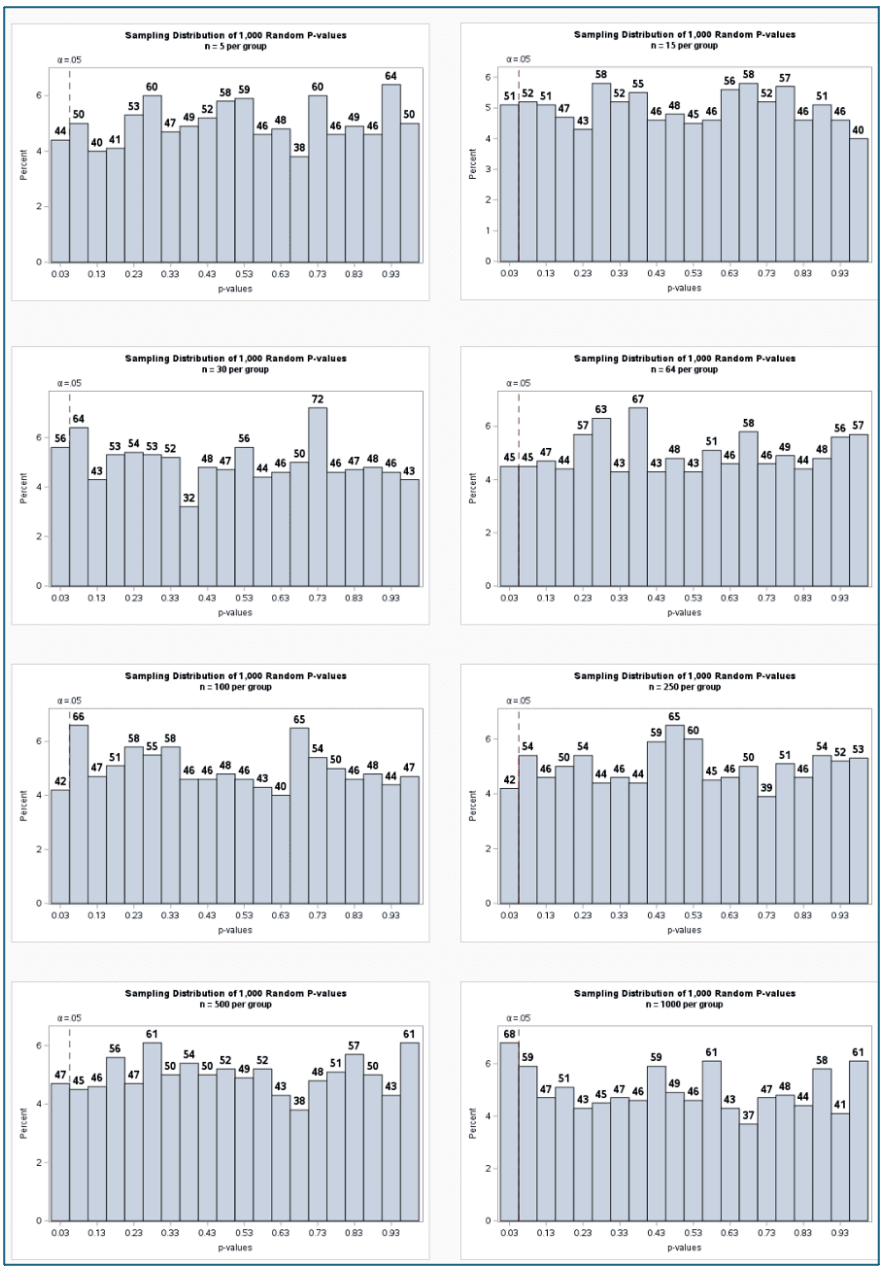

Figure 2. Histograms of p-values from independent samples t-tests of zero difference in population means.

Figure 3 shows distributions of Cohen's d as continuous effect sizes where the statistically significant d's are in the tails. As sample sizes increase, the ds converge (are better estimates) of the population D = 0.00. As a result, the two reference lines denoting the substantively significant effect sizes ($|d| \geq 0.20$) appear to be moving farther apart.

doi.org/10.32388/3QQSNC.4

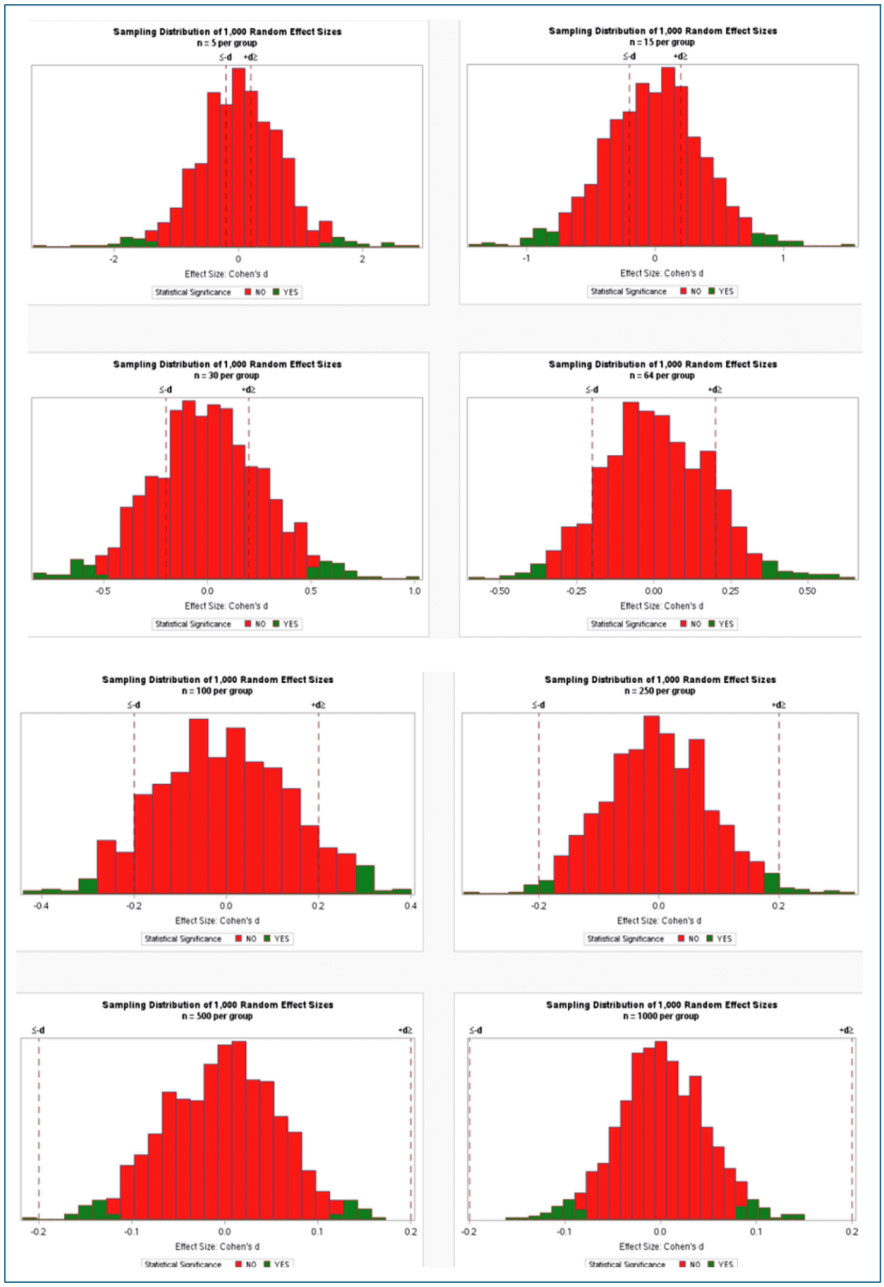

Figure 3. Sampling Distributions of Continuous Cohen's d under a True Null Hypothesis

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 increasing sample size, the percentage (count) of statistically significant p-values remains relatively constant at 5%, but the effect sizes become smaller. Finally, with n = 1,000 per group, all effect sizes are statistically significant, but none are substantively significant.

doi.org/10.32388/3QQSNC.4

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 (as depicted in Figure 4).

doi.org/10.32388/3QQSNC.4

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, these were all 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" states 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 few (6%) for contemplation. As sample size increased, the percentage of substantively significant effect sizes decreased, but all were statistically significant until n = 250 per group. Here, only 19 were statistically and substantively significant effect sizes, but 42 statistically significant effect sizes were not substantively significant. This phenomenon became more pronounced with n = 500 per group, where only one substantive effect size was statistically significant, with the remaining 46 not substantively significant. Finally, statistical significance was no help with n = 1,000 per group (total n = 2,000) because all 68 effect sizes were statistically but not substantively significant.

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

The false null hypothesis was created by adding 20 points to the original x_E (experimental) variable. This is similar to sampling x_E from a normal population where μ_E = 70 and σ_E = 25. The x_C (control) variable was randomly sampled from the previous population (μ_C = 50 and σ_C = 25. The null hypothesis (H_0: $\mu_{\bar{x}_E} - \mu_{\bar{x}_C} = 0$) was now false, and the alternative hypothesis was true (H_a: $\mu_{\bar{x}_E} - \mu_{\bar{x}_C} = 20$). However, the null hypothesis was tested for statistical significance, not the alternative hypothesis. Also, the population Cohen's D = (20/25) = 0.80 was now a large population effect size.

Figure 5 reveals right-skewed p-value distributions that are expected under a false null hypothesis. As the sample size increased, although α = .05, more than

doi.org/10.32388/3QQSNC.4

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 clustered in the 5th percentile (see Table 2).

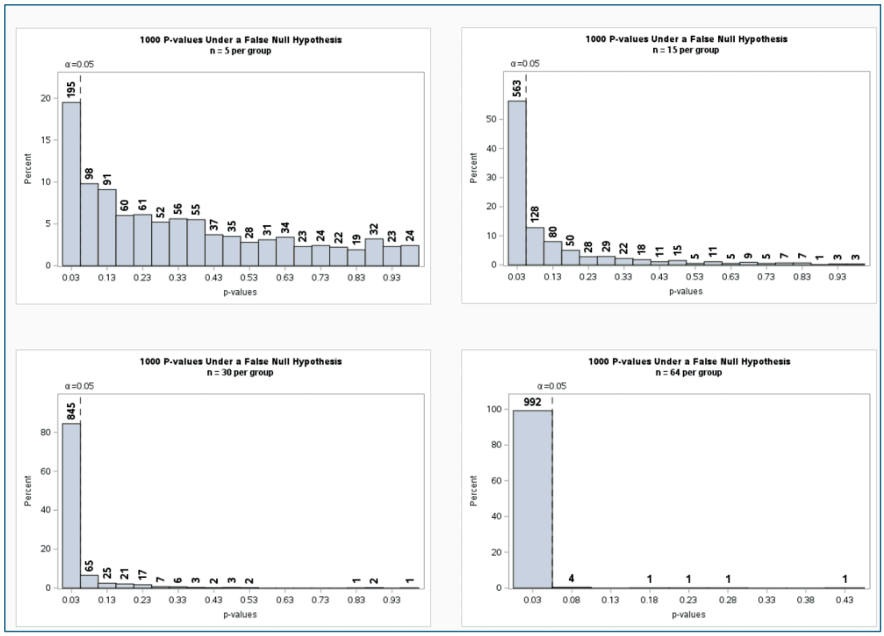

Figure 5. P-value Sampling Distributions Under a False Null Hypothesis

The histogram of the n=64 per group (total n = 120) reveals that 99% (992/1000) of the p-values were statistically significant. This reveals that with n = 64 per group and α = .05, there is 99% power to reject the "null hypothesis is true" assumption. Figure 6 confirms the empirical analysis with a formula from G*Power[14].

doi.org/10.32388/3QQSNC.4

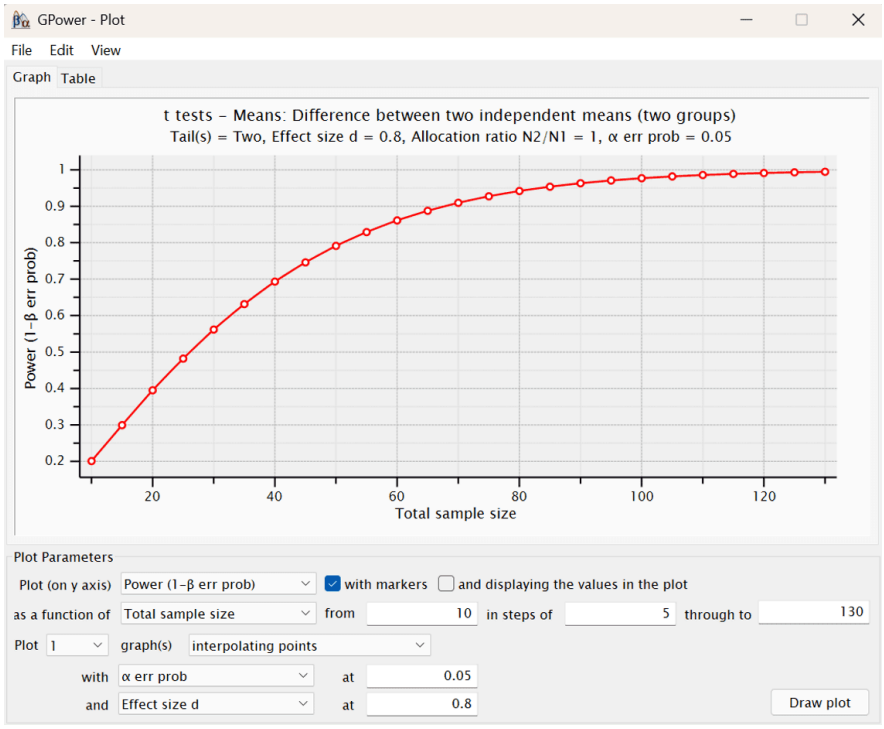

Figure 6. Power to Reject A Null Hypothesis with an Independent Samples T-Test for an Effect Size D = 0.80 and Alpha = .05.

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 presumably estimate the alternative grand mean by putting an X% confidence interval around the observed (sample) mean, where X is typically 90, 95, or 99. However, that indicates precision but does not guarantee the sample mean is an accurate estimate of the grand mean, particularly with small sample sizes.

The data demonstrated that, without statistical significance, many effect size errors are liable to be misinterpreted as substantively significant. Consequently, the scientific research literature will be inundated with even more irreplicable results than attributed to the misuse and abuse of statistical significance (Ioannidis, 2005). Ironically, replication solves the replication crisis[15] but should not be confused with merely 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[16] 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."[16].

In conclusion, I hope the results in this paper have convinced the reader that, regardless of a chosen α-level (bright line), statistical significance is a viable screening tool when working with a small sample size (total $n < 2{,}000$). If the null hypothesis is true, many false effect sizes are excluded from further consideration. With a large sample size ($n \geq 2{,}000$), substantive effect sizes are unusual under a true null hypothesis, regardless of statistical significance, and merit replication and scrutiny for scientific plausibility. The researcher in the scenario at the start of this paper should have been happy with the statistical analysis. The next step would be appropriately powering another experiment with a strong chance of rejecting the null hypothesis.

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 under a true null hypothesis. You may also agree with Mayo and Hand[17]: "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).

Statements and Declarations

Data availability

If you would like a copy of the SAS program that produced the results in this paper, please request a copy from komaroffeugene@gmail.com

References

1. ^*Cox DR (1982). "Statistical significance tests." Br J Clin Pharmac. 14:325–331.*
2. ^*Trafimow D, Marks M (2015). "Editorial." Basic and Applied Social Psychology. 3 7(1):1–2. doi:10.1080/01973533.2015.1012991.*
3. ^*Wasserstein RL, Schirm AL, Lazar NA (2019). "Moving to a world beyond p < 0.0 5." The American Statistician. 73(sup1):1–19.*
4. ^*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 misi nterpretations." European Journal of Epidemiology. 31:337–350.*
5. ^*Efron B (1998). "R. A. Fisher in the 21st century (Invited paper presented at the 19 96 R. A. Fisher Lecture)." Statistical Science. 13(2):95–122. doi:10.1214/ss/10289059 30.*

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. ^*Scheaffer RL (1995). Introduction to probability and its applications. 2nd ed. Duxbury Press.*
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=%22jacob%2Bcohen%22&hl=en&ei=GfE4TNSZHMK6cai36foO&sa=X&oi=book_result&ct=result&resnum=1&ved=0CCgQ6AEwAA%23v%3Donepage&q&f=false.*
13. ^*Westfall PH, Tobias RD, Wolfinger RD (2011). Multiple Comparisons and Multiple Tests Using SAS. 2nd ed. Cary: 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. ^*Vidgen B, Yasseri T (2016). "P-values: misunderstood and misused." Frontiers in Physics. 4:6.*
16. ^a, ^b*Fisher RA (1973). Statistical Methods and Scientific Inference. Hafner Press.*
17. ^*Mayo D, 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.