Research Article

A Redemption Song for Statistical Significance

Eugene Komaroff[1]

1. Keiser University, Fort Lauderdale, United States

Controversy is not new in Statistics. Since the start of the 20th Century, proponents of three theories have claimed superiority. Bayesian theorists mathematically mix subjective theoretical probabilities with the probability of the data. R.A. Fisher reenvisioned Bayes' theory by eliminating subjective probability and inventing a data-generating probability model called the null hypothesis. With this approach, only the probability of the data can be computed. Subsequently, Neyman-Pearson supplemented Fisher's null model with alternative data-generating probability models. In this century, massive "omics" data are analyzed with a complex amalgam of computer science, advanced mathematics, statistics, and domain-specific knowledge. This paper does not attempt to predict the future of statistics, unify the three classical statistical theories, argue the superiority of one of the others, propose a new theory, or call for a radical shift to a new paradigm (e.g., qualitative or mixed methods research). The statistical analyses in this paper are grounded in Fisher's paradigm. Independent samples t-tests were run with simulated data under a true and a false null hypothesis. Statistical significance was evaluated with p-values and substantive significance was determined using Cohen's "effect size index d." It is shown with graphs and a few numbers that statistical significance is a viable tool for filtering out effect size errors that would otherwise be misinterpreted as substantively significant. Finally, it is shown that increasing sample size does not improve power under a true null hypothesis – that happens only under a false null hypothesis.

Corresponding author: Eugene Komaroff, komaroffeugene@gmail.com

Trafimow and Marks[1], editors of Basic and Applied Social Psychology (BASP) stated, "analogous to how null hypothesis testing fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the

population parameter of interest is likely to be within the stated interval." Bayesian procedures were neither required nor banned from BASP, but "strong descriptive statistics, including effect sizes" were required. This paper demonstrates that evaluating effect sizes for substantive significance with small sample sizes but ignoring statistical significance produces many spurious effect sizes (effect size errors) under a true null hypothesis. Fricker et al.[2] 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)

Cox[3] stated that criticism of significance testing fills volumes. An overview of the controversies is at https://en.wikipedia.org/wiki/Statistical_hypothesis_test. Some authors believe that misunderstanding and abuse of statistical significance are inherent to the method, so they recommend a paradigm shift: Benjamin & Berger[4]; Benjamin et al.[5]; Goodman[6]; McShane et al.[7]; Trafimow & Marks[1]; Wellek[8]; Westover et al.[9]. Others believe that continuous p-values are fine but want to retire the dichotomization into statistical significance: Andrade[10]; Amrhein & Greenland[11]; Amrhein et al.[12]; Blakeley et al.[13]; Greenland et al.[14]; Greenland et al.[15]; Gigerenzer[16]; Haller & Krauss[17]; Imbens[18]; Utts[19]; Wasserstein et al.[20]. Finally, others acknowledged problems but argue that better education, and not a ban, is the solution: Begg[21]; Benjamini et al.[22]; Chen et al.[23]; Haller & Krauss[17]; Lane-Getazis[24]; Harrington et al.[25]; Komaroff[26]; Laken's[27]; Lytsy et al.[28]; Mayo & Hand[29]; Nickerson[30]; Spence & Stanely[31]; Wasserstein & Lazar[32]; Vidgen & Yasseri[33]. This paper does not attempt to unify the three classical statistical theories (Bayes, Fisher, or Neyman-Pearson), argue for supremacy, propose a new theory, or encourage a radical paradigm shift. This author agrees with the last group that proper education is the solution. With computer-simulated data, the results from many independent sample t-tests are graphed and tabulated, demonstrating that statistical significance should not be banned or retired: it is still a viable tool for decision-making when working with small sample sizes.

The independent samples t-test has three underlying assumptions that must be satisfied to ensure the results are valid: independent observations, homogeneity of variances, and normally distributed dependent variable. The assumptions can be evaluated with statistical tests. For example, normality can be tested with the Shapiro-Wilk or Kolmogorov-Smirnov tests, and homogeneity or equality of variance can be tested with Levene's test or an F-Max ratio. If one or more assumptions are suspected, nonparametric tests, such as the Wilcoxon Rank Sums test, Mann-Whitney U test, or Sign test, can be used instead. Typically, textbooks and tutorial papers start a discussion of statistical significance by

"assuming the null hypothesis is true." However, unlike the violation of t-test assumptions, a researcher should rejoice if this assumption is untenable. This is counterintuitive and may be a reason for the misunderstanding of statistical significance. In this paper, the t-test assumptions were satisfied, but there was no need to assume the "null hypothesis is true." The null hypothesis was true, the foundation upon which R.A. Fisher[34] built his significance testing paradigm.

Under the pseudonym Student[35], W.S. Gosset described the fundamental concept of a sampling distribution that undergirds the t-test he invented: "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 in so far 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). R.A. Fisher[34] 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). It is essential to recognize that the "population" in these quotes is not a social or physical phenomenon in nature. This population is a mathematical construct about a sampling distribution of means that exists only in statistical theory.

A histogram of a human trait such as height is relatively easy to appreciate. The histogram represents counts (percentages) of people with specific values on a graduated scale of measurements, such as feet and inches or meters and centimeters. Histograms of many sample means are difficult to understand because they depend on mathematical theorems: The Central Limit Theorem and the Law of Large Numbers[36]. This paper does not present mathematical proofs found in textbooks (e.g., Schaeffer[37]) and online courses (e.g., https://online.stat.psu.edu/stat414/lesson/24/24.2). This author created the complex mathematical/statistical phenomena of sampling distributions of means by simulating data with known parameters.

Two parameters of a sampling distribution of means determine p-values and statistical significance. One parameter is the central value of the sampling distribution, the mean of means or the grand mean, and the other is the standard error. Fisher[34] 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). Usually, the population standard deviation is in the notation $\sigma/\sqrt{n}$ and is called a standard error. The

population standard deviation (σ) is rarely known in practice, but Gosset[35] showed that a sample standard deviation (s) can be used to estimate the population standard deviation. In summary, when running a t-test, "assume the null hypothesis is true" is a statement (1) that the hypothesized central value (grand mean) of a theoretical sampling distribution of means is correct. (2) The standard deviation of the sample is an accurate estimate of the population standard deviation, and (3) the sampling distribution of means is normal. (4) Differences in sample means are attributed to sampling error.

Imagine a novel teaching method created to help sixth-grade elementary school students achieve grade-level fundamental academic skills (reading and arithmetic). The researcher hypothesized that the novel experimental intervention would be effective as predicted by a theory of cognitive development. The effect of the intervention was to be evaluated with a standardized test. The researcher hypothesizes that the novel or experimental teaching method would produce higher test scores on average than traditional pedagogy with the same content. However, unable to state how much higher, a small, proof of concept, randomized, controlled experiment was designed. The researcher stated a null hypothesis of zero difference between means but was confused. The researcher wanted evidence that supported the research hypothesis that the intervention worked but was required to postulate a contrary hypothesis that there was no difference between the experimental intervention (E) and the control intervention (C). This is bewildering unless the foundational sampling distribution concept is understood. The experiment was run after school hours, and both groups of students took a test at the end of the sessions. An independent samples t-test was used to determine the statistical significance of the mean difference.

The numerator of the t-test was the difference between the sample means $(\overline{x}_E - \overline{x}_C)$ subtracted from the difference between the population means $(\mu_E - \mu_C)$, that is $(\overline{x}_E - \overline{x}_C) - (\mu_E - \mu_C)$. The null hypothesis of zero difference in population means $(H_0 : \mu_{E-C} = 0)$ was "assumed to be true." Furthermore, the sample means were assumed to be accurate estimates of the population means. The researcher reported the following results in a paper: Five students taught with the experimental method had significantly higher test scores (M_E = 72.9, SD_E = 12.51) than those taught with the traditional method (M_C = 43.3, SD_C = 23.56). A two-sided independent samples t-test revealed a mean difference of about 30 points (M_{E-C} = 29.5, SD_{E-C} = 18.87) with a 95% confidence interval, 2.02 to 57.06. The p-value was statistically significant [t (8) = 2.48, p = .038] because α = 0.05. Finally, the standardized mean difference, Cohen's d = 1.57, was a huge effect size, indicating that the experimental pedagogy effectively improved the test scores. The researcher concluded there was only a 5% probability that the null hypothesis was true but a hefty 95% probability that the alternative hypothesis was true and enthusiastically espoused

reasons for the significant effect of the experimental intervention. Finally, although the wide 95% confidence interval indicated poor precision, the researcher argued that this could be easily corrected by replicating the experiment with a larger sample size.

“Significant effect “is ambiguous because the researcher may have been referring to the statistically significant p-value (p = .038), the fact that the null hypothesis parameter was not contained in the confidence interval, the raw mean difference (30 points), or the standardized mean difference (d = 1.57). However, there are other issues with the researcher’s conclusions. If you do not recognize them, please keep reading; these will be discussed after the simulation results are presented.

Since the start of the 20th Century, three competing theories have dominated statistical methodology[38]. Bayesian theorists require an a priori subjective probability of a hypothesis and evaluate the probability of an interaction between the subjective hypotheses and observed data. R.A. Fisher reenvisioned Bayes’ theory by eliminating subjective probabilities and inventing a data-generating probability model called the null hypothesis. In Fisher’s theory, only the probability of the data can be computed, not the probability of a hypothesis. Subsequently, Neyman-Pearson’s theory supplemented Fisher’s null hypothesis model with alternative hypotheses. This paper is based on Fisher’s significance testing paradigm:

> In order to be used as a null hypothesis, a hypothesis must specify the frequencies with which the different results of our experiment shall occur, and that the interpretation of the experiment consisted in dividing these results into two classes, one of which is to be judged as opposed to, and the other as conformable with the null hypothesis. If these classes of results are chosen, such that the first will occur when the null hypothesis is true with a known degree of rarity in, for example, 5 percent or 1 percent of trials, then we have a test by which to judge, at a known level of significance, whether or not the data contradict the hypothesis to be tested”[39].

Textbooks have cobbled Fisher’s null hypothesis paradigm with Neyman-Pearson’s alternative hypotheses paradigm[16]. Fisher[40] disapproved of the Neyman-Pearson paradigm as an “acceptance procedure” and argued that one can never accept but only reject or fail to reject a hypothesis: “The simple rejection of *a hypothesis (emphasis added)*, at an assigned level of significance…is often all that is needed, and all that is proper, for the consideration of a hypothesis about the body of the experimental data available” (p. 40). In textbooks like Moore et al.[36], the null hypothesis for means is presented as H_0: μ =

0, and a two-tailed alternative hypothesis as H_a: $\mu \neq 0$. In Fisher's paradigm, there is no mention of the alternative parameter. It can be any value on the number line except for the one specified under the null hypothesis. Fisher did not contemplate an alternative parameter in his significance testing paradigm. The phrase: "dividing these results into two classes" refers to statistical significance where one either rejects or fails to reject the parameter specified by the null hypothesis. For power/sample size calculations[41], a specific alternative parameter can be postulated, for example, with an "effect size index "d"[42] (p. 20). However, the null parameter is tested for statistical significance, not the hypothesized d.

If all assumptions are satisfied, the Central Limit Theorem and the Law of Large Numbers guarantee that the central value of a sampling distribution of means equals the population mean ($\mu_{\bar{x}} = \mu$). There is only one population parameter (central value of the sampling distribution) under the null hypothesis, but the sampling distribution's standard deviation (variance) is complicated. There are three standard deviations in statistics: one for the population distribution, another for the sample distribution, and the third for the sampling distribution of a summary statistic like the mean. According to Fisher[34]: "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). Textbooks present the standard deviation of the sampling distribution of means as $\sigma / \sqrt{n}$ and give it a new name, the standard error. For the independent samples t-test, Gosset (Student)[35] proved that the sample standard deviation (s) can be used as an estimate of the population standard deviation (i.e., $s / \sqrt{n}$). For testing a difference between two independent means, the standard error expands to $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. The standard error formulas reveal that sample sizes (n) reduce the magnitude of the standard error.

According to Moore et al.[36]: "A test statistic calculated from the sample data measures how far the data diverge from the expected value *if the null hypothesis* H_0 *were true* (emphasis added). Unusually large values of the statistic show that the data are not consistent with H_0. The probability computed *assuming that* H_0 *is true*, (emphasis added) that the test statistic would take a value as extreme as or more extreme than that actually observed is called the P-value of the test. The smaller the P-value, the stronger the evidence against H_0 provided by the data" (p 387). Cohen's[42] "effect size index d" a standardized mean difference can be used to evaluate substantive significance. However, d is closely linked to the t-test statistic: $t = Cohen^{'}d \sqrt{\frac{n_E n_C}{n_{E+} n_C}}$ (see Cohen[42], p. 67). With equal sample sizes, the sample size multiplier

simplifies to $\sqrt{\frac{n}{2}}$. Because every t-value maps to a specific p-value, the formula reveals that the corresponding d-value maps to the same p-value.,

There was no need for the assumptions “null hypothesis is true” or the “standard error is true” in this paper. The probability data generating parameters for the simulated sampling distributions were created with several sample sizes (n) from known parameters (μ, σ) of the standard normal curve. The fifth percentile of sampling distributions of p-values (e.g., Bland[43]; Murdoch[44]; Hung et al.[45]; Wang, et al.[46]; Verykouki & Nakas[47]) was used to determine statistical significance. Substantive significance was determined by evaluating the standardizing mean differences (d) according to Cohen's[42] criteria as trivial, small, medium, or large effect size.

Methodology

Sampling distributions were simulated with “do loops” on the free online statistical software called “SAS OnDemand for Academics”[48] (SAS, 2014, see APPENDIX). Under a true null hypothesis of a zero difference in population means, two variables (y_E and y_C) were randomly sampled from the same normal distribution [N (50, 25)]. The two variables were randomly replicated 1,000 times under eight sample size conditions: n = 5, 15, 30, 64, 100, 250, 500, and 1000. The difference in population means was tested for statistical significance with independent samples t-tests[49] (SAS, 2019). The SAS output provided the descriptive statistics: sample sizes, means, and standard deviations; and the inferential statistics: t-values, degrees of freedom, and p-values under the “equal variance assumption.” These summary data were concatenated into one analysis data set by sample size. The null parameter of zero was true: $H_0 : \mu_{E-C} = 0.0$, and the “equal variance assumption” was unquestionable because the data for both treatment groups were randomly sampled from one normally distributed parent population: Y_E and Y_C ~ N (μ = 50, σ = 25).

Statistical Significance

The a priori 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 percentage (count) of statistically significant p-values was an empirical estimate of the theoretical type 1 error rate of 5% as predicted by Fisher's frequentist theory (19under a true null hypothesis.

Substantive Significance

Cohen's d was computed by dividing the difference between two sample means by the pooled standard deviation. Cohen's[42] effect size categories determined substantive significance, where |d| < 0.20 was trivial, |d| ≥ 0.20 to 0.49 was small, |d| ≥ 0.50 to .0.79 was medium, and |d| ≥ 0.80 was large. Note that the |d| in the notation indicates negative and positive d values were relevant because two-sided t-tests were run. To compute the percentages of substantively significant effect sizes, an indicator variable was coded "1" if Cohen's |d| ≥ 0.20 (either small, medium, or large); otherwise, it was "0." Under the true null hypothesis, all statistically significant p-values were type 1 errors. All substantively significant (non-trivial) d's were "effect size errors" because the population Cohen's D was zero, i.e. (50 − 50) / 25 = 0.00.

Results

Figure 1 displays the sample means from the "experimental" group. The tick marks on the X-axis were kept constant to produce a visual impression of sample means converging on the "mean of means" or the same grand mean ($\mu_{\bar{x}}$.) with increasing sample size. The standard deviation of the sampling distribution of means became smaller because dispersion around the grand mean decreased as the sample size increased. The sampling distributions for the control group are not shown because they had essentially the same convergence patterns.

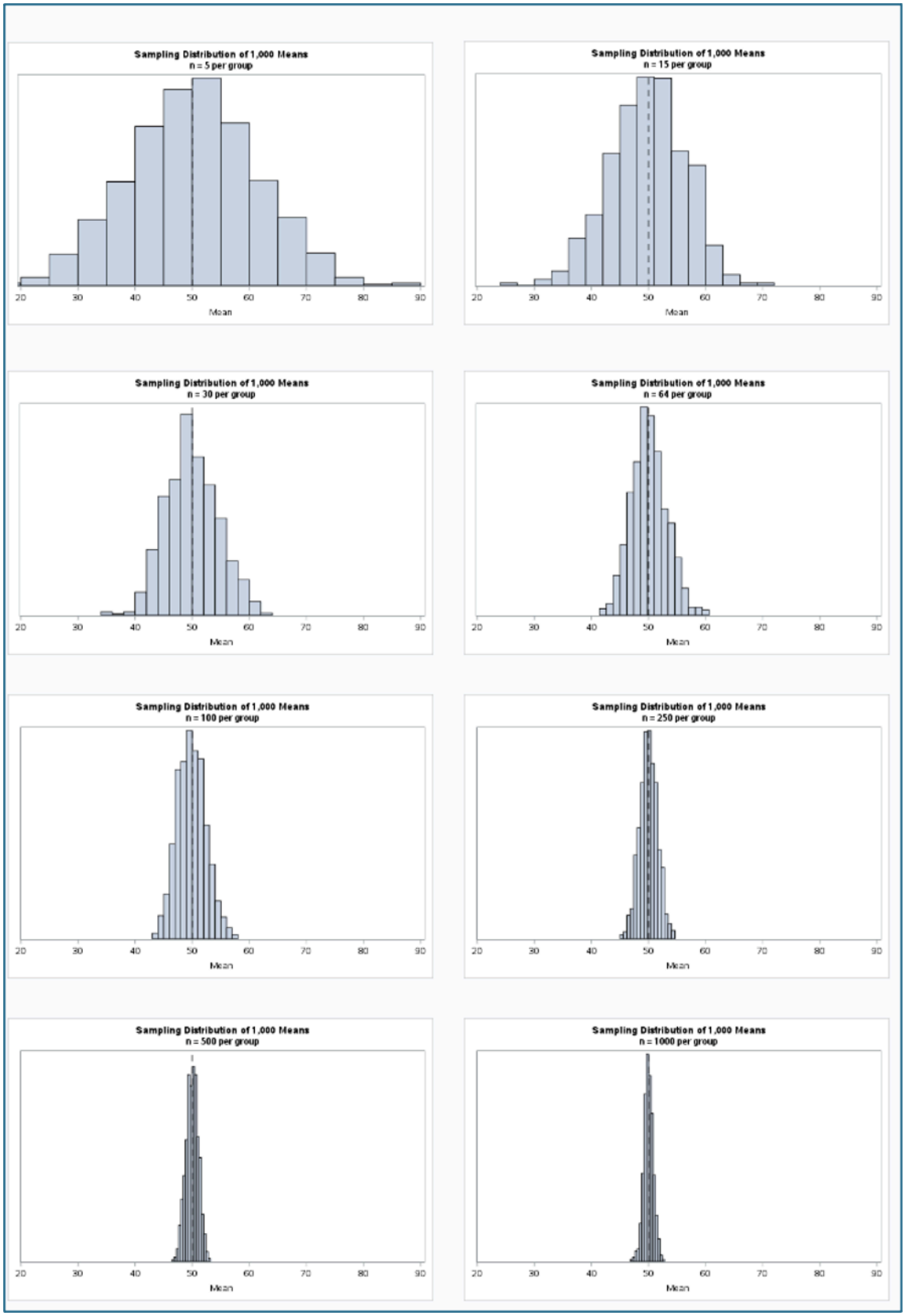


Figure 1. Sampling Distribution of Means of the Experimental Group under a True Null Hypothesis.

doi.org/10.32388/3QQSNC.3

Table 1 has the theoretical and empirical parameters under a true null hypothesis by sample size. The top of Table 1 shows the parameters from the data-generating probability distribution. Below the parameters are the theoretical standard errors (St. Error) calculated as $\sigma / \sqrt{n}$. The treatment group means are the "grand means" ($\mu_{\bar{x}}$)of the sampling distributions. The standard deviations (Std. Dev.) of the sampling distributions of means are empirical estimates of the theoretical standard errors $s / \sqrt{n}$.There is a close agreement between μ and each $\mu_{\bar{x}}$ as well as between the theoretical standard errors and empirical standard deviations. The consistency among the results validates the SAS simulation algorithm (see APPENDIX).

Normal Parent Population (μ = 50, σ = 25)								
n per group	5	15	30	64	100	250	500	1000
Theoretical Sampling Distributions								
Mean	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
St. Error	11.2	6.5	4.6	3.1	2.5	1.6	1.1	0.8
Empirical Experimental Group Sampling Distributions								
Mean	50.1	49.8	50	50.1	49.9	50	50	50
Std. Dev.	11.1	6.6	4.6	3.1	2.5	1.6	1.1	0.8
Empirical Control Group Sampling Distributions								
Mean	49.7	50.2	49.9	49.9	50	50	50.1	50
Std. Dev.	11.6	6.5	4.5	3.2	2.4	1.6	1.1	0.8
Total n	10	30	60	128	200	500	1000	2000
Theoretical Mean Difference Sampling Distributions								
Mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
St. Error	15.8	9.1	6.5	4.4	3.5	2.2	1.6	1.1
Empirical Mean Difference Sampling Distributions								
Mean	0.4	-0.4	0.1	0.2	-0.1	0.0	-0.1	0.0
Std. Dev.	15.9	9.2	6.6	4.4	3.5	2.2	1.5	1.2

Table 1. Theoretical and Empirical Parameters of Sampling Distributions of Means

Table 2 has the summary statistics for Figure 1, which confirm the convergence (Std. Dev., Min and Max) of the sample estimates on the central values of the sampling distributions with increasing sample size.

Treatment Group	Data Set	Grand Mean	Std. Dev.	Min	Max
Experimental	n = 5 per group	50.1	11.1	17.3	87.9
	n = 15 per group	49.8	6.6	25.5	69.7
	n = 30 per group	50.0	4.6	35.0	63.1
	n = 64 per group	50.1	3.1	41.9	60.3
	n = 100 per group	49.9	2.5	43.2	57.6
	n = 250 per group	50.0	1.6	45.4	54.4
	n = 500 per group	50.0	1.1	46.7	53.1
	n = 1000 per group	50.0	0.8	47.0	52.5
Control	n = 5 per group	49.7	11.6	11.6	82.7
	n = 15 per group	50.2	6.5	29.3	72.6
	n = 30 per group	49.9	4.5	33.8	67.0
	n = 64 per group	49.9	3.2	39.8	59.9
	n = 100 per group	50.0	2.4	43.1	56.9
	n = 250 per group	50.0	1.6	43.9	55.2
	n = 500 per group	50.1	1.1	46.4	53.9
	n = 1000 per group	50.0	0.8	47.7	53.2
Mean Differences (E - C)	n = 5 per group	0.4	15.9	-49.9	64.2
	n = 15 per group	-0.4	9.2	-28.0	27.1
	n = 30 per group	0.1	6.6	-21.3	20.6
	n = 64 per group	0.2	4.4	-13.5	16.0
	n = 100 per group	-0.1	3.5	-10.6	9.5
	n = 250 per group	0.0	2.2	-7.5	7.1
	n = 500 per group	-0.1	1.5	-5.2	4.1
	n = 1000 per group	0.0	1.2	-4.0	3.6

Table 2. Means, standard deviations of sampling distributions by increasing sample size

Figure 2 displays Cohen's d's as a continuous variable on the X-axis and the corresponding p-values from the t-tests on the Y-axis. Two perfectly monotonic relationships (Spearman |r| = 1.0) are evident between the d's and p's on the left and again on the right of population Cohens' D = 0.00. These correlations confirm the mathematical link between the t and d: $t = Cohen^{'}d\sqrt{\frac{n_E n_C}{n_{E+}n_C}}$ [42]. Consequently, the following from Wasserstein et al.[20] is puzzling: "No p-value can reveal the plausibility, presence, truth, or importance of an association or effect"(p. 2). Because p-values are related to Cohen's d, they can be used to "reveal the strength of an association," but they are restricted to an exclusive 0.0 to 1.0 scale. Cohen's d is more informative because it can be any value on the real number line.

In Figure 2, the line graph for n = 5 per group (top left) has 44 (4%) statistically significant d's. Of these, 23 ranged between -3.26 and -1.46 on the far left, and 21 were between 1.46 and 2.83 on the far right of Cohen's D = 0.00. Although statistically significant (below the α = 0.05 reference line), these are grandiose

overestimates of Cohen's D. More important is that fact that "not statistically significant" filtered out 706 effect size errors (all d's $\leq$ - 0.20 and all d's $\geq$ 0.20) that otherwise would be misinterpreted as substantively significant.

doi.org/10.32388/3QQSNC.3

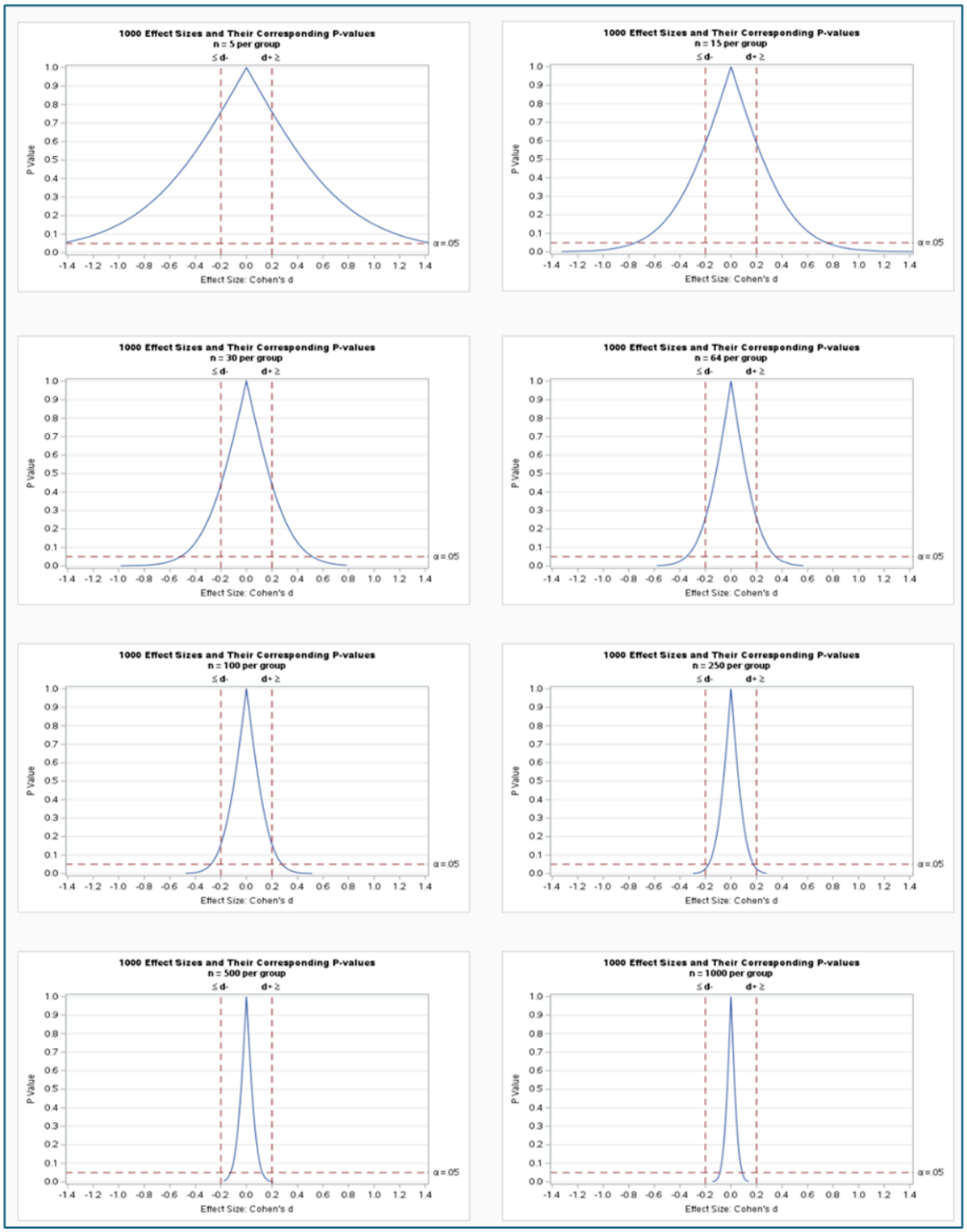


Figure 2. Line Graphs of P-Values from Independent Samples T-Tests of Differences in Means represented as Cohen's d Effect Sizes.

doi.org/10.32388/3QQSNC.3

Figure 3 shows that the empirical histograms of p-values are almost uniform (rectangular). Uniform distributions are predicted by statistical theory under a true null hypothesis if all statistical and research design assumptions are satisfied[50]. In other words, in the open interval (0.00 to 1.00), every p-value has the same chance of materializing under the true null hypothesis. Notice that the 5^{th} percentile (demarcated by the α reference line) contains all the statistically significant p-values for every sample size.

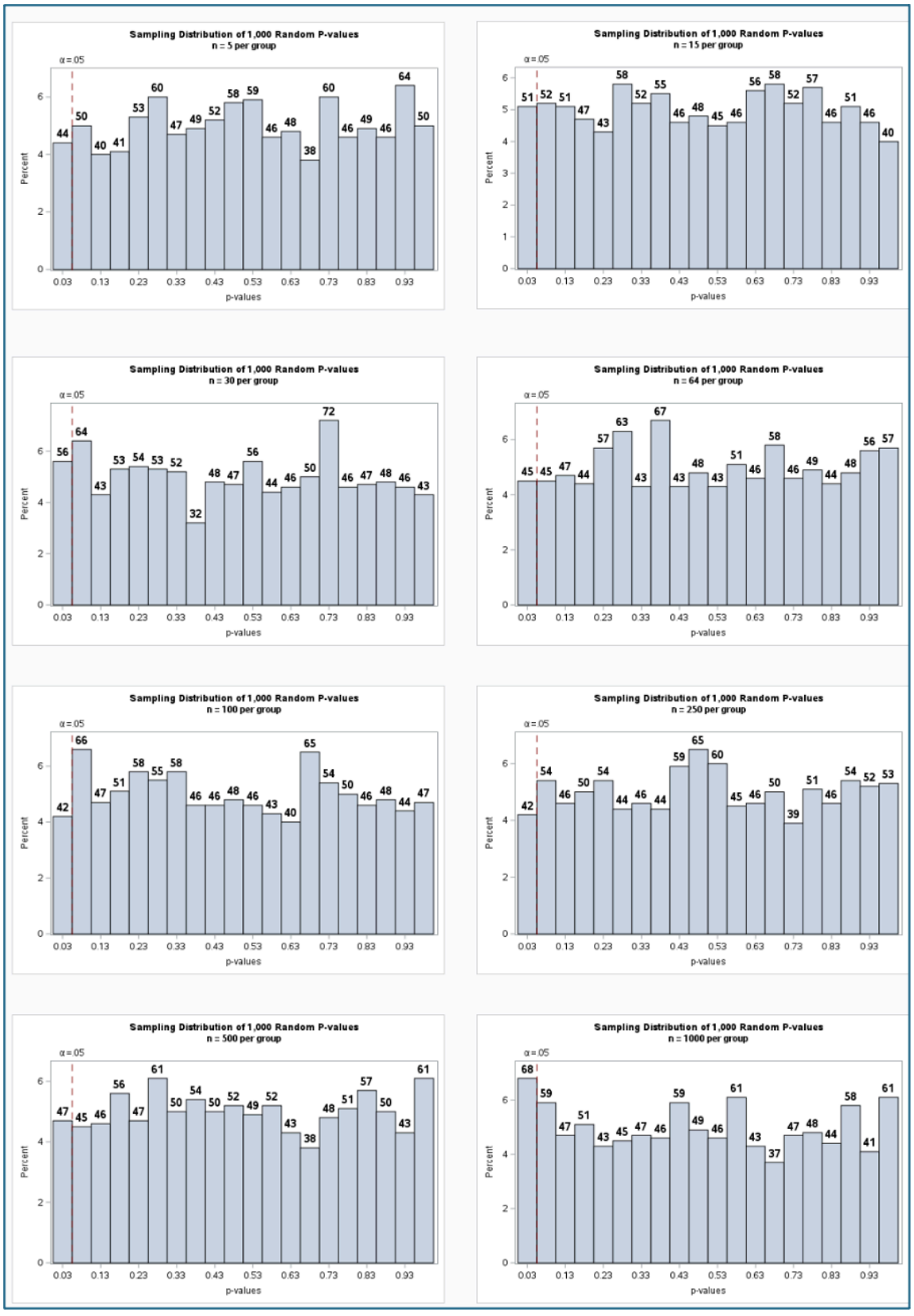


Figure 3. Empirical Histograms of P-values Under a True Null Hypothesis.

doi.org/10.32388/3QQSNC.3

Figure 4 shows distributions of Cohen's d as continuous effect sizes. As the sample size increased, the ds converged on the population D = 0.00. The same phenomenon occurred with the raw mean differences converging on μ_{E-C} = 0.00. The software (SAS) determined the X-axis tick marks for the graphs. As a result, the two reference lines denoting substantively significant effect sizes ($|d| \geq 0.20$) appear to move farther apart with increasing sample size. This, again, is the convergence phenomenon towards the population D = 0.00 as was evident in Figure 2. Notice that statistically significant d's are only in the tails of the distributions.

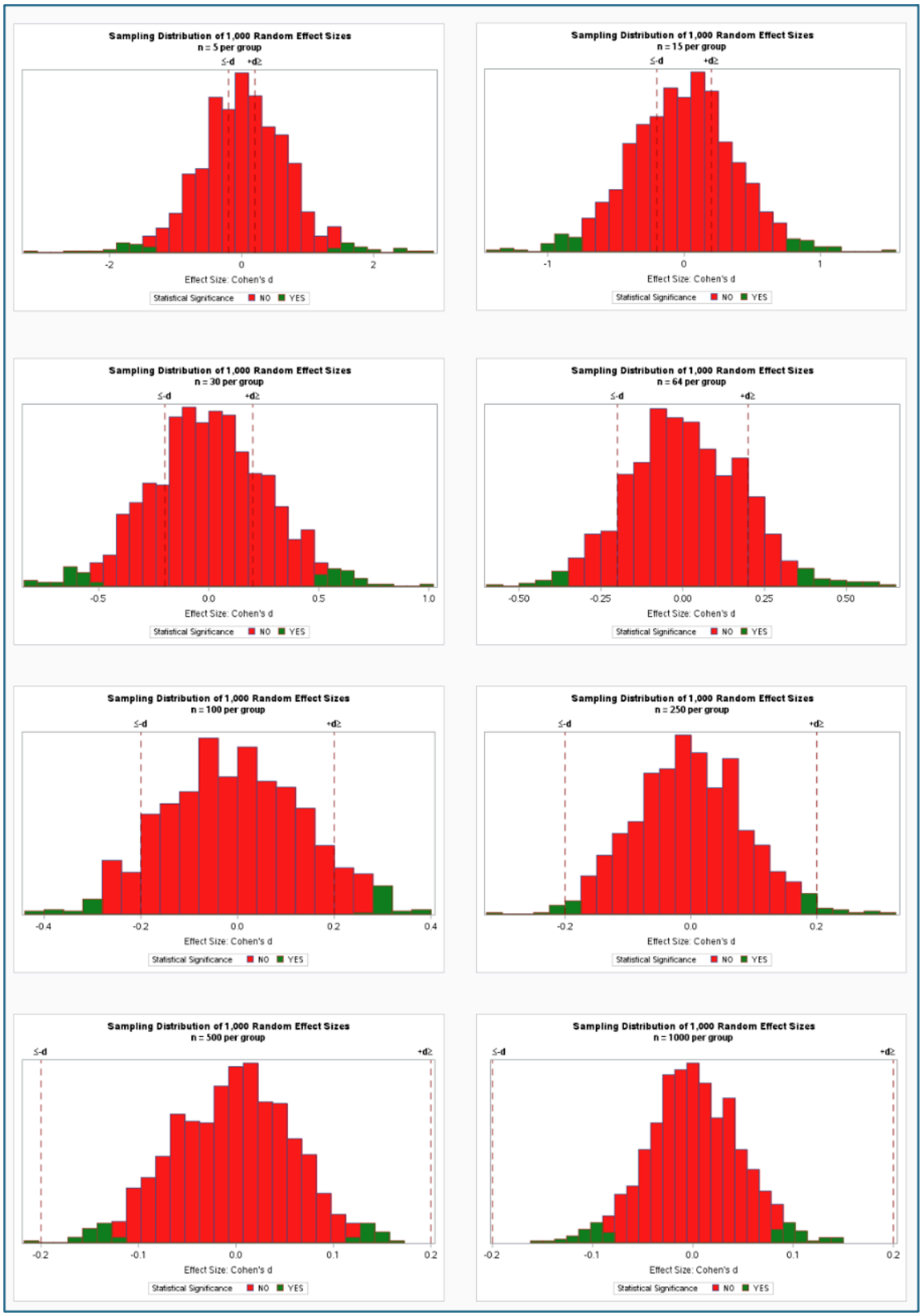


Figure 4. Sampling Distributions of Continuous Cohen's d under a True Null Hypothesis

doi.org/10.32388/3QQSNC.3

The bar graphs in Figure 5 were created by grouping the continuous d's according to Cohen's[42] 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 remained relatively constant at 5% and smaller effect sizes became statistically significant. Finally, with n = 1,000 per group, only trivial effect sizes materialized, and all were statistically significant.

Figure 5. Distributions of Effect Size Index d according to Cohen's Criteria

doi.org/10.32388/3QQSNC.3

Table 3 provides the counts and percentages of statistically significant p-values, and substantively significant effect sizes corresponding to Figures 2 to 5.

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

For instance, with n = 5 per group, 44 (4 %) of the 1,000 p-values were statistically significant ($p < .05$), but these were type 1 errors because the rejected null hypothesis was true. Similarly, 750 d's were substantively significant effect sizes (small, medium, or large) but were effect size errors because D = 0.00. Nevertheless, statistical significance filtered out 94% of the effect size errors, leaving only 6% (44/750) for consideration as substantively significant effect sizes. As the sample size increased, the percentage of effect size errors decreased, but all were statistically and substantively significant until n = 250. Only 19 statistically significant effect size errors materialized, but there were 42 statistically significant p-values. The statistical significance filter now also caught trivial effect sizes. Statistical significance lost much utility as a screening tool for effect size errors. This is more evident with n = 500 per group, where only one effect size error was detected as statistically significant, and the remaining 46 were trivial effect sizes. Finally, with n = 1,000 per group, all 68 statistically significant p-values corresponded to trivial effect sizes.

Confidence Intervals under a True Null Hypothesis

Figure 5 has subsets of 20 confidence intervals for the raw (unstandardized) mean differences under a true null hypothesis because the graph of 1,000 confidence intervals was a big, incomprehensible smear. Although confidence intervals for d's can be calculated[51], they were unavailable from SAS. Nevertheless, the raw mean difference confidence intervals display the same convergence phenomena as those for the standardized mean differences (Cohen's d). The precision increased (width decreased) as the sample size increased and the intervals converged on the population parameter. Moore et al.[36] stated that 95% confidence is a statement about the method's success rate in the long run and is not the probability of the population parameter contained in any given interval. For instance, with n = 1,000 per group, two intervals do not include the population parameter.

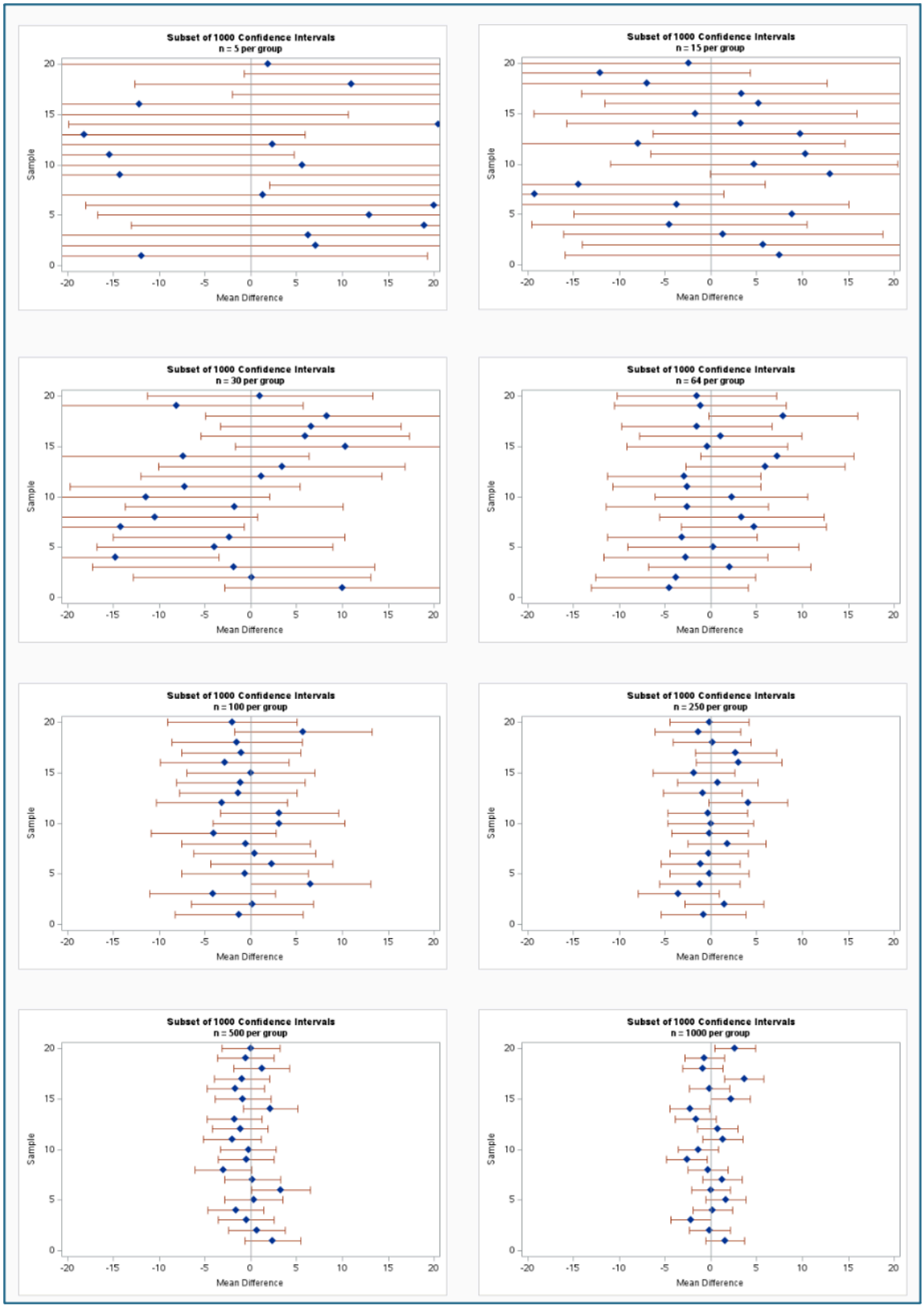


Figure 5. Subsets of Confidence Intervals Around Sample Mean Estimates of the Difference in Population

doi.org/10.32388/3QQSNC.3

Means.

P-values, Statistical Significance, and Cohen's d Under a False Null Hypothesis

The false null hypothesis was created by adding 20 points to the original y_E (experimental) variables where $\mu = 50$ and $\sigma = 25$ (see APPENDIX). This is similar to sampling y_E from a normal population distribution where $\mu_E = 70$. The y_C (control) variables were again randomly sampled from the same population distribution ($\mu_C = 50$, $\sigma_C = 25$) as under the true null hypothesis. The null hypothesis H_0: $\mu_{(E-C)} = 0$ was now false because $\mu_{E-C} = 20$; however, the null hypothesis of zero population mean difference was tested for statistical significance. Cohen's D = (20/25) = 0.80 is now a large population effect size under the false null hypothesis.

Figure 6 reveals skewed p-value distributions under a false null hypothesis. As the sample size increased, more than 5% of p-values became statistically significant. The p-value histograms for n > 64 per group are not shown because only one small interval (bar) contained all 1,000 p-values < 0.05 (see Table 4).

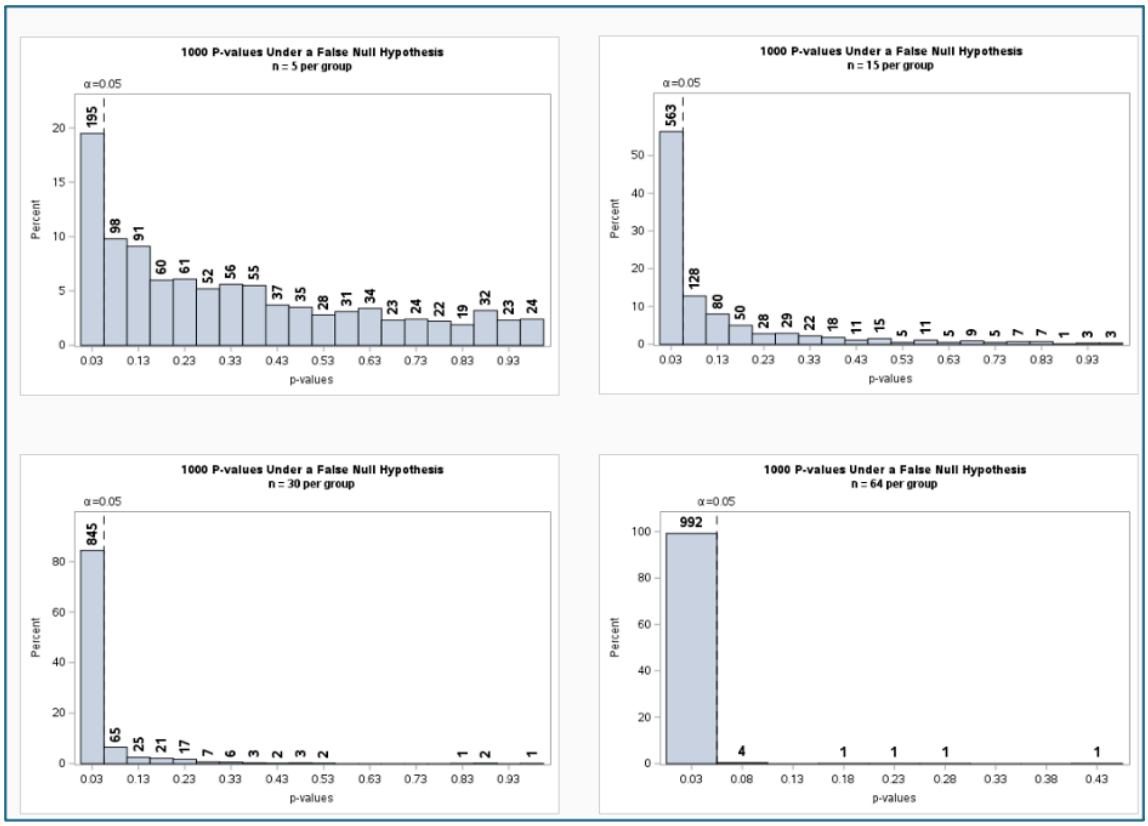
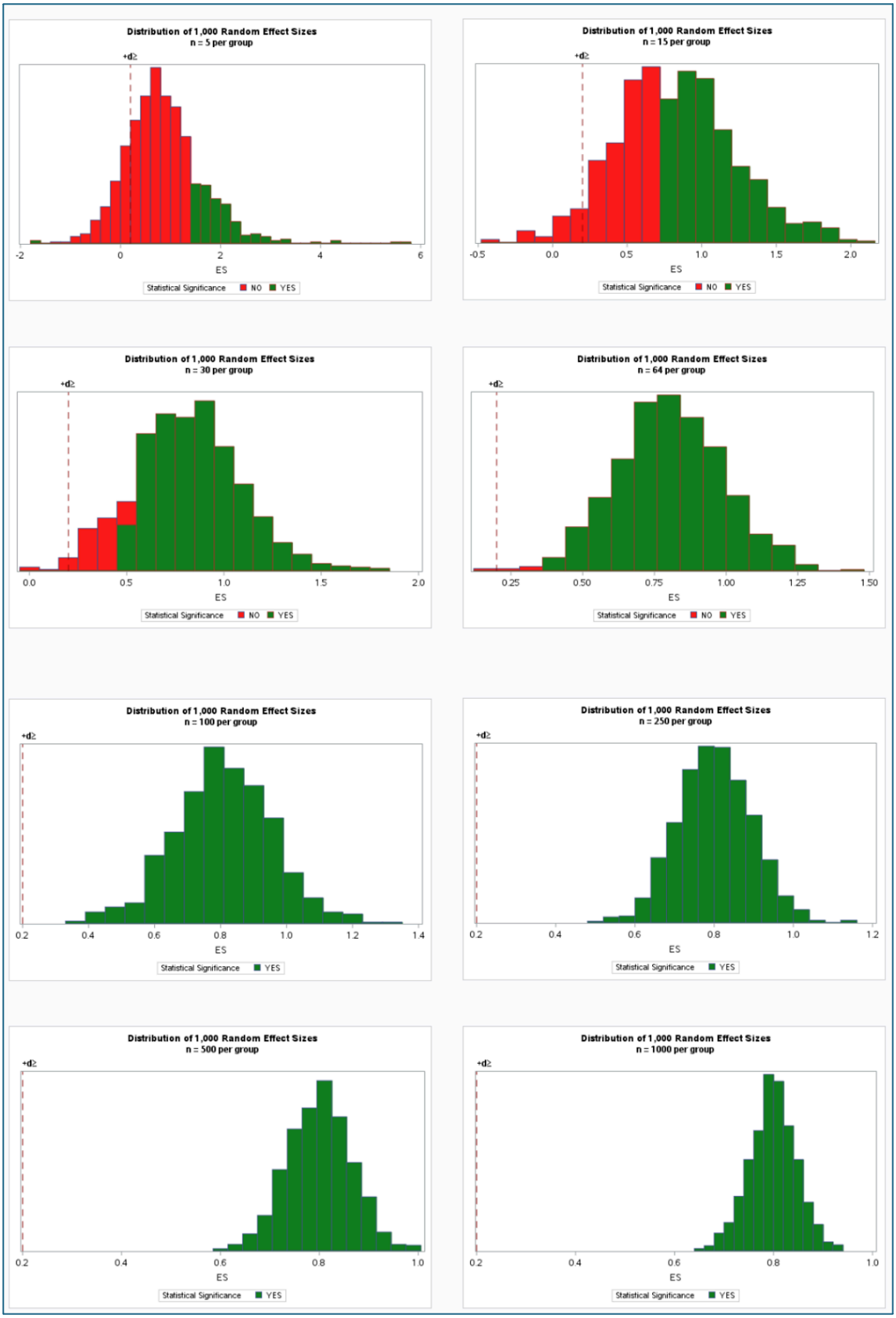


Figure 6. P-value Sampling Distributions Under a False Null Hypothesis

doi.org/10.32388/3QQSNC.3

Figure 7 shows that the sample d's are normally distributed but are now centered at the population D= 0.80. With increasing sample size, much higher percentages (counts) of substantive effect sizes and statistically significant p-values materialized under the false null hypothesis compared to the true null hypothesis.



doi.org/10.32388/3QQSNC.3

Figure 7. Sampling Distribution of Continuous Effect Sizes Overlayed with Statistically Significant P-values Under a False Null Hypothesis.

Table 4 contains the counts and percentages of statistically significant p-values and substantively significant effect sizes ($|d| \geq 0.20$) under the false null hypothesis. Starting with n = 100 per group, 100% of the effect sizes were substantively significant, and 100% of the p-values were statistically significant.

False Null Hypothesis H_0: $\mu_E - \mu_C = 0$								
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	118					
	NO	YES	687	882	88.2%			22.1%
	YES	YES	195			195	19.5%	
n = 15 per group	NO	NO	42					
	NO	YES	395	958	95.8%			58.8%
	YES	YES	563			563	56.3%	
n = 30 per group	NO	NO	10					
	NO	YES	145	990	99.0%			85.4%
	YES	YES	845			845	84.5%	
n = 64 per group	NO	NO	2					
	NO	YES	6	998	99.8%			99.4%
	YES	YES	992			992	99.2%	
n = 100 per group	YES	YES	1000	1000	100%	1000	100%	100%
n = 250 per group	YES	YES	1000	1000	100%	1000	100%	100%
n = 500 per group	YES	YES	1000	1000	100%	1000	100%	100%
n = 1000 per group	YES	YES	1000	1000	100%	1000	100%	100%

Table 4. Statistically Significant P-Values and Substantively Significant Effect Sizes ($|d| \geq 0.20$) Under a False Null Hypothesis

As predicted by statistical theory, Figure 8 shows that the confidence intervals converged on the parameter $\mu_{E-C} = 20$ as the sample size increased.

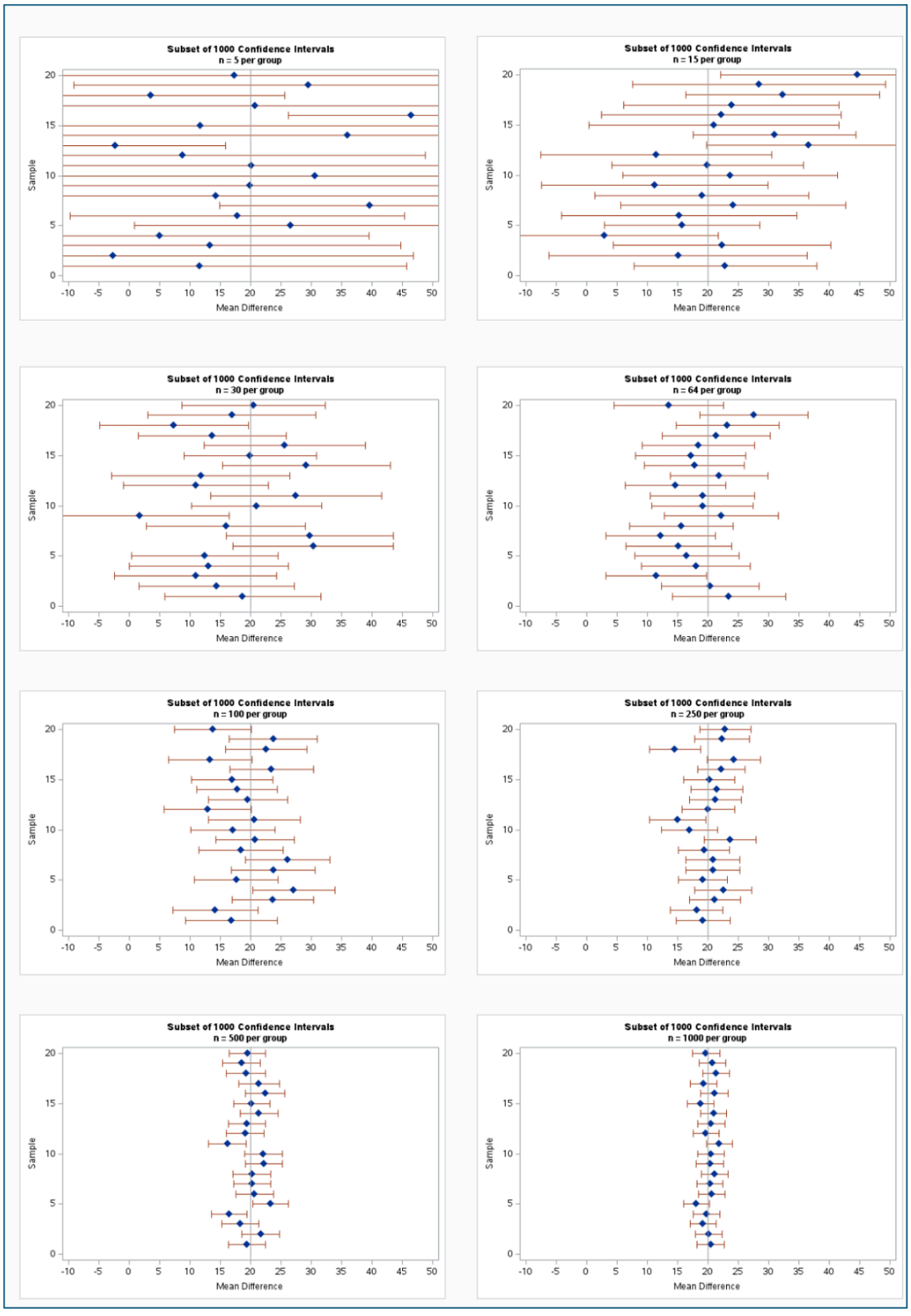


Figure 8. Subsets of Confidence Intervals for the Difference in Population Means Under a False Null

doi.org/10.32388/3QQSNC.3

Hypothesis.

In Figure 6, the p-value distributions are increasingly skewed right with increasing sample size. Table 3 has the minimum and maximum p-values by sample size.

Data Set	Minimum	Maximum
n = 5 per group	0.00002	0.9992
n = 15 per group	0.000004	0.9873
n = 30 per group	0.000000003	0.9987
n = 64 per group	0.0000000000003	0.4004
n = 100 per group	0.00000000000000001	0.0120
n = 250 per group	0.000000000000000000000000000002	0.00000002
n = 500 per group	0.0000000000000000000000000000000000000000000000001	0.00000000000000000002
n = 1000 per group	0.00000000000000000000000000000000000000000000000000000000000000000000000000000008	0.00000000000000000000000000000000000000000000000000000000000000000000000000000005

Table 3. Minimum and Maximum P-values with Increasing Sample Size Under a False Null Hypothesis.

The parameter tested under the false null hypothesis was a zero difference in population means. However, the true population difference was 20, and the sample means were now converging to 20 with increasing sample size. As a result, more sample means were farther away (deviations) from the null parameter of zero. Thus, t-statistics became bigger, and many smaller p-values materialized. Table 4 reveals 100% power at n = 100 per group under the false null hypothesis because the maximum p-value = .0120. Moore et al.[36] stated: The smaller the P-value, the more substantial the evidence against H_0 provided by the data" (p 387). A similar idea is conveyed by researchers who add the adjective "very" to statistical significance. This extra information about p-values produces misunderstanding of statistical significance. Once the null hypothesis has been "knocked down" ($p < \alpha$), there is nothing more to be gained by kicking and spitting on it. For instance, with $\alpha = .05$, the maximum p = .0120 with n = 100 is statistically significant as p < .000001 (p = 8.094055E-88 in scientific notation) with n = 1,000. There is no need for "more substantial evidence against the null hypothesis" a $p < \alpha$ is necessary and sufficient. However, the smallness of a statistically significant p-value is a crucial consideration when controlling for alpha inflation[52].

Conclusion

Westover et al.[9] reported the results of a short survey administered to 246 physicians at three major US teaching hospitals. The physicians were asked: "Consider a typical medical research study designed to test the efficacy of a drug in which a null hypothesis H0 ('no effect') is tested against an alternative

hypothesis H1 (‘some effect’). Suppose that the study results pass a statistical significance test (that is, P-value <0.05) in favor of H1. What has been shown?” The physicians had to choose one of the following seven response options: (1) H0 is false. (2) H0 is probably false. (3) H1 is true. (4) H1 is probably true. (5) Both 1 and 3. (6) Both 2 and 4. (7) None of the above. Only 7% got the correct answer: "None of the above.” However, instead of explaining why the other answers were wrong, Westover et al. encouraged readers to study Bayesian statistical theory, where the probability of hypotheses can be computed.

The simulations in this paper reveal that the probability of a true or false null hypothesis is irrelevant in Fisher’s frequentist paradigm. The parameter under the null hypothesis is not a random variable but is a fixed, specific value that the researcher determines, e.g., H_0: $\mu_{E-C} = 0$. The t-statistic (or Cohen’s d) probability is computed as the area under the curve of a t-sampling distribution. The null parameter is also the hypothesized central value of the sampling distribution of the t-statistics, and the distance (deviation) of the sample t-statistic from the central value corresponds to a specific p-value. The bigger the distance, the smaller the p-value. Incidentally, the null parameter does not need to be zero. The default parameter for the t-test in SAS is zero, but that can easily be changed to some other reasonable value.

The simulated data demonstrated that a ban on statistical significance increases the risk that many effect size errors will be interpreted as substantively significant. Although the risk is reduced under a false null hypothesis, this is not reassuring because the alternative parameter is unknown in the Fisher frequentist paradigm. Researchers have reported “no difference,” indicating acceptance of the null hypothesis[12]. If the goal was to determine no difference, an independent samples t-test was the wrong statistical method. The proper analysis requires positing a margin of equivalence and running two, one-sided t-tests[27][53][54]. In short, failing to reject a null hypothesis does not justify a “no difference” conclusion; it merely warrants further scientific investigation.

Proof of concept or pilot experiments are typically done with small sample sizes. The hypothetical study (at the start of this paper) came from an independent samples t-test run under a true null hypothesis, H_0: $\mu_{E-C} = 0.0$, with n = 5 per group and produced a statistically significant p-value. There cannot be much confidence in the results of a study with such a small sample size. As shown with sampling distributions of Cohen’s d as a continuous measure, large effect sizes materialized by chance, and sample means varied widely around the grand mean of the sampling distribution. Confidence intervals become narrower with increasing sample size, but the correct parameter is not guaranteed to be contained by either the observed or any future confidence interval. Finally, it is naïve for anyone working with

Fisher's[39] frequentist paradigm to interpret a p-value as a probability of either a null or alternative hypothesis (1.00 – p-value). The null parameter is fixed by design, and the alternative parameter is s unknown.

The improved precision of the confidence intervals with increasing sample size is understandable because the calculation includes a standard error. However, sample size is not the only driver of precision. The magnitude of the standard deviation also affects the standard error. As the sample size increases, a complex mathematical relationship exists between the margins of error (standard error multiplied by a critical t value) added and subtracted from the sample mean as the confidence bounds. It is easier to understand this phenomenon with simulated data where the standard deviation of the sampling distribution of means (standard error) became smaller as more sample means became better estimates of the grand mean of the sampling distribution

Imagine a researcher abides by BAPS's ban on statistical significance, has a small sample size, and interprets only the effect size. The results in this paper suggest that the scientific research literature will be inundated with even more irreplicable effect sizes than have already been blamed on the misuse and abuse of statistical significance[55]. Ironically, a solution for the replication crisis is reproduction and replication[33]. The report from the National Academies of Sciences, Engineering, and Medicine[56] defined the two concepts: "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 sampling error may prevent replication of exploratory research findings with small sample sizes, as seen here with unrealistically large effect sizes.

Fisher[40] also called for replication because statistical significance was a guide, not a final adjudication: "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 (α cut point) 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."[40].

Wasserstein et al.[20] ban on statistical significance, but not p-values, is reminiscent of Karl Pearson's objection to Fisher's and Neyman-Pearson's statistical significance testing with the chi-square goodness

of fit test: “It is very unwise, in my opinion, to form tables which provide only the values of P = 0.01 and P = 0.05, and consider 'hypotheses' which give a value of P < 0.01 as 'false,' and those with a value between 0.01 and 0.05 as 'doubtful,' and for the rest of the scale of P have no descriptive category for you must not say that such values prove hypotheses to be true”[57].

Nonetheless, the results in this paper support the need for a cut score that separates p-values into two classes: one class is statistically significant, and the other is not. However, the cut score or level of statistical significance does not have to be 0.05. It can be of any value in the open interval (0.0 to 1.0) consistent with the purpose of a scientific investigation. For instance, $\alpha \leq 0.20$ is recommended for selecting covariates for a multiple logistic regression model[58]. Similarly, $p > 0.05$ is desirable when testing assumptions such as normality or homogeneity of variance. In Structural Equation Modeling (SEM) theory, a statistically non-significant ($p \geq 0.05$) chi-square p-value indicates an acceptable fit between the data and the theoretical model[59]. Similarly, a chi-square goodness of fit tests the null hypothesis that a categorical variable follows a specific (theoretical) probability distribution. If the null hypothesis states a probability distribution, the researcher believes it is true (for example, 7 is the most frequent sum of two rolled fair dice). Rejecting the null hypothesis ($p < \alpha$) indicates the null hypothesis is false or the dice are loaded (biased). However, one does not accept the null if $p \geq \alpha$. According to Fisher[34], If p is between .1 and .9 there is certainly no reason to suspect the hypothesis tested, but that does mean the hypothesis has been proven. The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P the more satisfactorily is the hypothesis verified. Values over .999 have sometimes been reported, which, if the hypothesis were true, would only occur once in a thousand trials. The high p-value represents a rare event, just like p = .001, but with such a high p-value the null hypothesis raises suspicion (see Mendelian paradox at https://en.wikipedia.org/wiki/Gregor_Mendel)

Benjamini & Hochberg[52] argued that multiple testing of the same hypothesis but selectively reporting only small p-values is alpha inflation: “Conducting the analysis for many subgroups and highlighting or reaching decisions about the selected few that come out to be statistically significant raises a danger that the conclusions from the study will not be a result of a natural phenomenon but merely reflect the selection of the extremes among the extensively tested noise“ (p. 60). Theoretical physicists recognized the alpha inflation problem when running many experiments to test the same phenomenon, so they used a 5-sigma level of statistical significance[60]. Compare this to the relatively lax 2-sigma level ubiquitous in the social sciences.

I hope the results in this paper have convinced readers that statistical significance is a viable tool with small sample sizes because it filters or screens out false effect sizes from further consideration under a true null hypothesis. Nonetheless, statistical significance was useless with n = 1,000 per group. With large sample sizes, a substantively significant Cohen's d is unusual under a true null hypothesis parameter = 0.0; therefore, it merits scrutiny regardless of statistical significance. Similarly, with small sample sizes, huge effect sizes materialize by chance, which also merits scrutiny regardless of statistical significance. In conclusion, 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 computer clock time initiated the random data streams. The SAS code (in the APPENDIX) can be run on the free Internet version of SAS OnDemand for Academics[48]. Please simulate your sampling distributions of p-values and effect sizes under a true and a false null hypothesis. You may also agree, as I do, with Mayo and Hand[29]: "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).

Appendix

SAS code based on Wicklin's[61] method simulating data under a true null hypothesis.

```
libname N'/home/ekomaroff0/TRUENULL';

options VALIDVARNAME=ANY;

ODS EXCLUDE ALL;

/* True Null simulation of many random samples */

%let N = 1000;  /* sample size */

%let Mu0 = 50;

%let Sigma=25; /* true value of parameter */

%let run = 1000; /* number of random samples */

data Simt;

call streaminit(0);

do Sample = 1 to &run;

do i = 1 to &N;

do group = 1 to 2 by 1;

y = rand("Normal", &Mu0, &Sigma);

output;

end;

end;

end;

run;

ods exclude all;

proc ttest data=Simt alpha=0.05 H0=0 Sides=2;

by sample;

class group;

var y;

ods output ttests = Ttests statistics=stats;

run;

data ttests2;

set ttests;
```


doi.org/10.32388/3QQSNC.3

```
If method="Satterthwaite" then delete;

run;

data stats2;

set stats;

drop n Variable LowerCLStdDev UpperCLStdDev UMPULowerCLStdDev UMPUUpperCLStdDev StdErr;

If method="Satterthwaite" then delete;

If method="Pooled" then do;

es = Mean / StdDev;

end;

run;

data N.T&n;

merge ttests2 stats2;

by sample;

n = &n;

run;
```

SAS code simulates data under a false null hypothesis.

```
/* False Null simulation of many random samples */

libname N '/home/ekomaroff0/FALSENULL';

options VALIDVARNAME=ANY;

%let N = 26;  /* sample size */

%let MU0 = 50;

%let Sigma=25; /* true value of parameter */

%let run = 1000; /* number of random samples */

data SimF;

call streaminit(0);

do Sample = 1 to &run;

do i = 1 to &N;

do group = 1 to 2 by 1;

y = rand("Normal", &MU0, &sigma);

output;

end;

end;

end;

run;

Data simf2;

set simf;

if group = 1 then y = y + 20;

/* Population Mean for Group 1 changed to 70 */

run;

ods exclude all;

proc ttest data=SimF2 alpha=0.05 H0=0 Sides=2;

by sample;

class group;

var y;
```


doi.org/10.32388/3QQSNC.3

```
ods output ttests = Ttests statistics=stats;

run;

data ttests2;

set ttests;

If method="Satterthwaite" then delete;

run;

data stats2;

set stats;

drop n Variable LowerCLStdDev UpperCLStdDev UMPULowerCLStdDev UMPUUpperCLStdDev StdErr;

If method="Satterthwaite" then delete;

If method="Pooled" then do;

ES = Mean / StdDev;

end;

run;

data F&n;

merge ttests2 stats2;

by sample;

n = &n;

run;

/*Go back to the top, change the sample size, and re-run program  */
```

Statements and Declarations

The author has no conflicts of interest to disclose.

References

1. a, b *Trafimow D, Marks M. (2015). Editorial. Basic and Applied Social Psychology. 37(1): 1-2. doi:10.1080/01973533.2015.1012991.*

2. ^Fricker Jr RD, Burke K, Han X, William H. Woodall (2019). Assessing the statistical analyses used in basic an d applied social psychology after their p-value ban. The American Statistician. 73:sup1, 374-384. doi:10.108 0/00031305.2018.1537892
3. ^Cox DR (1982). Statistical significance tests. Br. J. clin. Pharmac. 14: 325-331.
4. ^Benjamin DJ, Berger JO (2019). Three recommendations for improving the use of p-values. The American S tatistician. 73:sup1, 186-191. doi:10.1080/00031305.2018.1543135
5. ^Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, ... Johnson VE (2018). Redefine statistical significance. Nature Human Behaviour. 2(1): 6-10.
6. ^Goodman S (2008). A Dirty Dozen: Twelve P-Value Misconceptions. Seminars in Hematology. 45: 135-140.
7. ^McShane BB, Gal D, Gelman A, Robert C, Tackett JL. (2019). Abandon statistical significance. The American Statistician. 73(sup1): 235-245.
8. ^Wellek S. (2017). A critical evaluation of the current "p-value controversy". Biometrical Journal. 59(5): 854-872.
9. a, bWestover MB, Westover KD, Bianchi MT. (2011). Significance testing as perverse probabilistic reasoning. BMC medicine. 9: 1-20.
10. ^Andrade C (2019). The P value and statistical significance: Misunderstandings, explanations, challenges, a nd alternatives. Indian J Psychol Med; 41: 210-215.
11. ^Amrhein V, Greenland S (2017). Remove, rather than redefine, statistical significance. Correspondence publ ished online: doi:10.1038/s41562-017-0224-0
12. a, bAmrhein V, Greenland S, McShane B (2019). Comment: Retire statistical significance. Nature. 567(7748): 3 05-307.
13. ^Blakeley B, McShane, David Gal, Andrew Gelman, Christian Robert, Jennifer L. Tackett (2019). Abandon St atistical Significance. The American Statistician. 73:sup1, 235-245. doi:10.1080/00031305.2018.1527253
14. ^Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016). Statistical tests, P va lues, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology. 31: 337-350.
15. ^Greenland S (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. The American Statistician. 73:sup1, 106-114. doi:10.1080/00031305.2018.1 529625
16. a, bGigerenzer G (2004). Mindless statistics. The Journal of Socio-Economics. 33: 587–606.

17. ^ a, b Haller H, Krauss S (2002). Misinterpretations of significance: A problem students share with their teacher s. Methods of Psychological Research. 7(1): 1–20.

18. ^ Imbens GW. (2021). Statistical significance, p-Values, and the reporting of uncertainty. The Journal of Econ omic Perspectives. 35(3): 157–174.

19. ^ Utts J. (2018). Understanding p-values and the controversy surrounding them. Accessed 10.14.2024 from htt ps://ics.uci.edu/~jutts/UnderstandingP-Values

20. ^ a, b, c Wasserstein RL, Schirm AL, Lazar NA. (2019). Moving to a world beyond p < 0.05. The American Statisti cian. 73(sup1): 1–19.

21. ^ Begg CB (2020). In defense of p-values. JNCI Cancer Spectrum. 4(2): 1–4. doi:10.1093/jncics/pkaa012

22. ^ Benjamini Y, De Veaux RD, Efron B, Evans S, Glickman M, Graubard BI, He X, Meng X, Reid N, Stigler SM, Va rdeman SB, Wikle CK, Wright T, Young LJ, Kafadar K (2021). The ASA president's task force statement on stat istical significance and replicability. Ann. Appl. Stat. 15(3): 1084–1085. doi:10.1214/21-AOAS1501

23. ^ Chen OY, Bodelet JS, Saraiva RG, Phan H, Di J, Nagels G, Schwantje T, Cao H, Gou J, Reinen JM, Xiong B (202 3). The roles, challenges, and merits of the p value. Patterns. 4(12).

24. ^ Lane-Getazis SJ. (2017). The p-value really dead? Assessing inference learning outcomes for social science students in an introductory statistics course. Statistics Education Research Journal. 16(1): 357–399.

25. ^ Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, Drazen JM, Hamel BM (2019). New guidelines for statistical reporting in the Journal. N Engl J Med. 381: 285–286.

26. ^ Komaroff E. (2020). Relationships between p-values and Pearson correlation coefficients, Type 1 errors an d effect size errors, under a true null hypothesis. Journal of Statistical Theory and Practice. 14(3): 49. doi:10.1 007/s42519-020-00115-6.

27. ^ a, b Lakens D. (2021). The practical alternative to the p value is the correctly used p value. Perspectives on Ps ychological Science. 16(3): 639–648.

28. ^ Lytsy P, Hartman M, Pingel R. (2022). Misinterpretations of P-values and statistical tests persist among res earchers and professionals working with statistics and epidemiology. Upsala Journal of Medical Sciences. 12 7.

29. ^ a, b Mayo D, Hand D. (2022). Statistical significance and its critics: Practicing damaging science, or damagin g scientific practice? Synthese. 200(3): 1–33. doi:10.1007/s11229-022-03692-0.

30. ^ Nickerson RS. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. Psychological methods. 5(2): 241.

31. ^Spence JR, Stanley DJ. (2018). Concise, simple, and not wrong: In search of a short-hand interpretation of st atistical significance. Frontiers in Psychology. 9: 1-5. doi:10.3389/fpsyg.2018.02185.

32. ^Wasserstein RL, Lazar NA. (2016). The ASA statement on p-values: context, process, and purpose. The Ame rican Statistician. 70(2): 129-133.

33. a, bVidgen B, Yasseri T. (2016). P-values: misunderstood and misused. Frontiers in Physics. 4: 6.

34. a, b, c, d, eFisher RA (1970). Statistical Methods for Research Workers (14th ed.). Reprinted in 1993 as Statistic al Methods, Experimental Designs and Scientific Inference by Oxford University Press.

35. a, b, cStudent. (1908). The probable error of a mean. Biometrika. 6(1): 1–25.

36. a, b, c, d, eMoore DS, Notz WI, Fligner M. (2021). Basic Practice of Statistics (9th ed.). Macmillan Learning.

37. ^Scheaffer RL. (1995). Introduction to probability and its applications (2nd ed.). Doxbury Press.

38. ^Efron B (1998). R. A. Fisher in the 21st century (Invited paper presented at the 1996 R. A. Fisher Lecture). St atistical Science. 13(2): 95–122. doi:10.1214/ss/1028905930

39. a, bFisher RA (1966). Design of Experiments (8th Ed.) New York: Hafner Publishing. Reprinted in 1993 as Stat istical Methods, Experimental Designs and Scientific Inference by Oxford University Press.

40. a, b, cFisher RA (1973). Statistical Methods and Scientific Inference. Hafner Press. Reprinted in 1993 as Statist ical Methods, Experimental Designs and Scientific Inference by Oxford University Press.

41. ^Faul F, Erdfelder E, Lang AG, Buchner A (2007). G*Power 3: A flexible statistical power analysis program fo r the social, behavioral, and biomedical sciences. Behavior Research Methods. 39: 175-191.

42. a, b, c, d, e, f, gCohen J (1968). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Assoc iates.

43. ^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

44. ^Murdoch DJ, Tsai YL, Adcock J. (2008). P-values are random variables. The American Statistician. 62(3): 24 2-245.

45. ^Hung JHM, O'Neill RT, Bauer P, Köhne K (1997). The behavior of the p-value when the alternative hypothes is is true. Biometrics. 53(1): 11-22.

46. ^Wang B, Zhou Z, Wang H, Tu XM, Feng C. (2019). The p-value and model specification in statistics. Gen Psy chiatr. Jul 9; 32(3): e100081. doi:10.1136/gpsych-2019-100081. PMID: 31360911; PMCID: PMC6629378.

47. ^Verykouki E, Nakas CT. (2023). Adaptations on the Use of p-Values for Statistical Inference: An Interpretati on of Messages from Recent Public Discussions. Stats. 6(2): 539-551.

48. ^a, ^b*SAS Institute Inc. (2014). SAS® OnDemand for Academics: User's Guide. SAS Institute Inc.*

49. ^*SAS Institute Inc. (2019). SAS/STAT® 9.4 User's Guide. Cary NC: SAS Institute Inc.*

50. ^*Westfall PH, Tobias RD, Wolfinger RD (2011). Multiple Comparisons and Multiple Tests Using SAS (2nd ed.). SAS Institute Inc.*

51. ^*Howell DC (n.d.). Confidence Intervals on Effect Size. Accessed July 17, 2024 at https://www.uvm.edu/~statdhtx/methods8/Supplements/MISC/Confidence%20Intervals%20on%20Effect%20Size.pdf*

52. ^a, ^b*Benjamini Y, Hochberg Y (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics. 25(1): 60–83.*

53. ^*Schuirmann DJ. (1987). A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. Journal of Pharmacokinetics and Biopharmaceutics. 15: 657–680.*

54. ^*Wellek S. (2010). Testing Statistical Hypotheses of equivalence and noninferiority. Second Edition. CRC Press.*

55. ^*Ioannidis JPA. (2005). Why most published research findings are false. PLoS Med. 2(8): e124.*

56. ^*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.*

57. ^*Inman HF. (1994). Karl Pearson and RA Fisher on statistical tests: a 1935 exchange from Nature. The American Statistician. 48(1): 2–11.*

58. ^*Hosmer DW, Lemeshow S (2000). Applied Logistic Regression. 2nd Edition, Wiley, New York. doi:10.1002/0471722146*

59. ^*Hayduk LA (2014). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. BMC Med Res Methodol. 14: 124. doi:10.1186/1471-2288-14-124*

60. ^*Louis L. (2013). Discovering the Significance of 5 sigma. (Accessed 10.11.2024 from https://doi.org/10.48550/arXiv.1310.1284)*

61. ^*Wicklin R (2013). Simulating Data with SAS. SAS Institute Inc.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.