Research Article

Statistical and Substantive Significance of Pearson Bivariate Correlation Coefficients Under True or False Null Hypotheses

Eugene Komaroff[1]

1. Keiser University, Fort Lauderdale, United States

Modern, massive digital data requires computer-intensive algorithms (data science) for analysis. However, small data sets continue to be analyzed with classical, inferential statistical methods. Regrettably, these methods have been tainted by the abuse, misuse, and misunderstanding of statistical significance. Understanding statistical significance requires an appreciation of theoretical sampling distributions of summary statistics under a true null hypothesis. This paper demonstrates a method to teach statistical and substantive significance with empirical computer-simulated sampling distributions of Pearson's correlation coefficients. Sampling distributions of Pearson's correlation coefficients and p-values reveal that statistical significance with small sample sizes filters out effect size errors that would otherwise be considered substantively significant under a true null hypothesis.

Corresponding author: Eugene Komaroff, komaroffeugene@gmail.com

About 45 years ago, Cox[1] stated that criticism of statistical significance fills volumes. The p-value debates continue. For a recent review of the controversy, see https://en.wikipedia.org/wiki/Statistical_hypothesis_test. This paper does not attempt to unify the three classical statistical theories[2], promote modern data science, nor call for a paradigm shift to qualitative research methods. This paper has two purposes: (1) to contribute to the p-value debate with evidence that statistical significance is still a viable tool for a binary decision when working with small sample sizes, and (2) to teach simple, intuitive, foolproof, and proper understanding of statistical and substantive significance with empirical sampling distributions.

Greenland et al.[3] attempted to teach the proper interpretation of statistical significance by cataloging misinterpretation, misuse, and abuse. Others also attempted education[4] but ultimately decided it was futile and, among others, banned statistical significance from the scientific literature[5][6][7]. This banishment is regrettable because statistical significance is a viable tool for filtering out false effect sizes when working with

 doi.org/10.32388/PS72PK.3

small sample sizes. However, appreciating this fact requires understanding theoretical sampling distributions. Students can interpret histograms of data from samples. However, they most likely do not fully grasp the concept of sampling distributions of summary statistics because these are derived from complex mathematical theorems (Central Limit Theorem, Law of Large Numbers). That is a problem. Understanding sampling distributions of summary statistics is foundational for appreciating both statistical and substantive significance. A related problem is the common misinterpretation that a null hypothesis can be accepted as true when the p-value is not statistically significant[6]. This misunderstanding is obviated here as data are simulated under a known, rather than merely assumed, true null hypothesis. In frequentist theory, the population parameter, as stated with a null hypothesis, is a specific, exact, and fixed value. There is no probability (uncertainty) associated with the population parameter. The population correlation rho (ρ) is declared with null hypothesis, and here rho is equal to zero (H_0: $\rho = \rho_0 = 0$).

Theoretical Framework

Student[8] described the theory of small-sample sampling distributions: “Any experiment may be regarded as forming an individual of a ‘population’ of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population” (p. 1). Fisher[9] echoed the concept: “The entire result of an extensive experiment may be regarded as but one of a possible population of such experiments” (p. 2). Moore et al.[10] present histograms of sampling distributions comprised of many summary statistics (means and proportions) repeatedly and randomly sampled with replacement from a human population. However, the human perspective obscures the important fact that sampling distributions are not physiological, physical, psychological, sociological, or economic phenomena. They are probability distributions of summary statistics that exist only in dense mathematical theory. This paper does not require derivations with advanced mathematics, relying instead on graphs and simple counting numbers.

Methodology

Because correlation sampling distributions with small sample sizes are skewed[9], Pearson’s correlation coefficients r were converted to z_r with the Fisher r to z transformation. These were subtracted from the population correlation specified under the null hypothesis ($z_r - \rho_0$) and the difference was divided by the standard error, $1/\sqrt{n-3}$, which is “practically independent of the value of the correlation in the population from which the sample is drawn”[9]. The resulting ratio is the z-test $\left(\frac{z_r - \rho_0}{1/\sqrt{n-3}}\right)$ which produces a z-score for a sample correlation and the corresponding p-value. These calculations were performed 1,000 times for each of six sample size conditions, with the “Fisher” option in PROC CORR[11] and the data was output for graphing

and tabulation. With a 5% alpha level of statistical significance ($\alpha = .05$), an indicator variable was coded as "1" when p-value < α, otherwise it was "0". The count of "1's" out of 4095 p-values was an empirical estimate of the theoretical Type 1 error rate. Note because the null hypothesis was true, and all statistically significant p-values were Type 1 errors (rejections of the true null hypothesis).

Substantive significance was determined by Cohen's[12] classification of Pearson's continuous correlation coefficients (r) as categorical effect sizes: $|r| < .10$ is trivial, $|r| \geq 0.10$ is small, $|r| \geq 0.30$ is medium, and $|r| \geq 0.50$ is a large effect size. The absolute value ($|r|$) conveys that Pearson's correlations span both negative and positive values, which are bound between -1.0 and +1.0. Each correlation was coded into one of the four mutually exclusive effect size categories (0 = trivial, 1 = small, 2 = medium, 3 = large) with a 4-level indicator variable. Although Fisher's z was tested for statistical significance, the back-transformed r was evaluated of substantive significance. Note that the population Cohen's D = 0.0 (i.e., $\rho = 0.0$); therefore, all substantively significant correlations ($|r| \geq .10$) were false, spurious, or effect size errors. For additional details about the methodology see Komaroff[13].

Data Source

SAS onDemand for Academics[14], freely available on the internet from SAS Institute Inc., was used to generate bivariate Pearson correlation coefficients. These were computed with PROC CORR[11] using 100 independent, identically distributed (i.i.d.) random variables drawn from the standard normal population ($\mu = 0$, $\sigma = 1$). This population distribution guaranteed that $\rho = 0.0$ (for mathematical proof, see [15]). Six separate data sets were created, each producing 4,950 observed correlations, but with different sample sizes: n = 5, 15, 30, 100, 1000, and 2000. The six data sets were combined (appended) into one data set for graphing and tabulating by sample size.

Results

The empirical sampling distribution with n = 5 is far from normally distributed (see Figure 1).

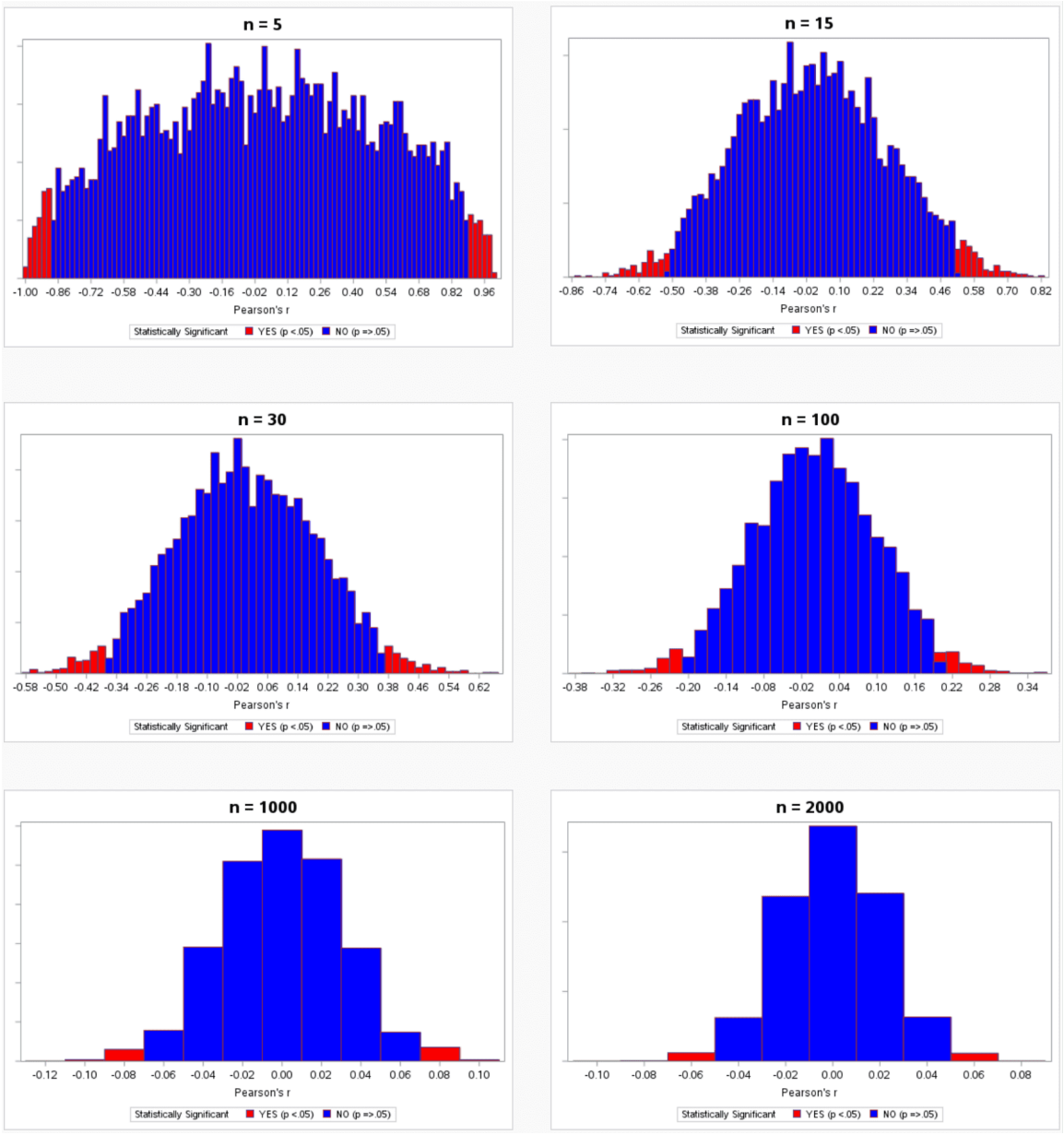

Figure 1. *Six Empirical Sampling Distributions with 4950 Correlation Coefficients by Sample Size.*

There is a discrepancy between the grand mean and the median of the sampling distributions of correlations (see Table 1).

Sample Size	Variable	Mean	Median	Std Dev	Minimum	Maximum
n = 5	Pearson's r	0.00199	0.0067	0.4976	-0.99947	0.99727
	Effect Sizes				0	3
	Fisher's r to z	0.00025	0.0067	0.68438	-4.11405	3.29728
	P-values	0.53263	0.54856	0.28778	0	0.99985
n = 15	Pearson's r	0.0019	0.00048	0.26527	-0.84938	0.81396
	Effect Sizes				0	3
	Fisher's r to z	0.00251	0.00048	0.28577	-1.25393	1.13865
	P-values	0.50587	0.50261	0.28771	0.00001	0.99998
n = 30	Pearson's r	0.00052	-0.00425	0.18654	-0.57243	0.6575
	Effect Sizes				0	3
	Fisher's r to z	0.00063	-0.00425	0.19326	-0.65112	0.7884
	P-values	0.4964	0.49416	0.28684	0.00004	0.99993
n = 100	Pearson's r	0.00119	0.00087	0.10057	-0.36525	0.36478
	Effect Sizes				0	2
	Fisher's r to z	0.0012	0.00087	0.1016	-0.38293	0.38239
	P-values	0.50028	0.50522	0.28974	0.00016	0.99998
n = 1000	Pearson's r	0.00026	-0.00005	0.0317	-0.10179	0.10221
	Effect Sizes				0	1
	Fisher's r to z	0.00026	-0.00005	0.03173	-0.10214	0.10257
	P-values	0.49537	0.49427	0.28634	0.0012	0.99953
n = 2000	Pearson's r	-0.00001	0.0005	0.02228	-0.08551	0.07342
	Effect Sizes				0	0
	Fisher's r to z	-0.00001	0.0005	0.0223	-0.08572	0.07355
	P-values	0.49688	0.50065	0.2846	0.00013	0.99949

Table 1. *Descriptive Summary Statistics of 4950 Correlation Coefficients by Sample Size*

The standard deviation, which is an empirical estimate of the standard error, is 0.4976, indicating the dispersion of the correlations around the grand center correlation of 0.00199. This is a large spread, as evident from the range (minimum and maximum), which is very close to the boundary of Pearson's correlation coefficients (-1.0 to +1.0). These data reveal that many correlations computed with small sample sizes seriously under- or over-estimate ρ. As sample sizes increase, the empirical sampling distributions approximate bell-shaped (normal) curves, and the standard deviations (empirical standard errors) become smaller. For example, with n = 2000, the range (-0.09 to +0.07) indicates that the entire sampling distribution is comprised of only trivial, or ignorable, effect sizes ($|r| < .10$).

In contrast, the empirical sampling distributions of p-values are relatively uniform regardless of sample size, with small p-values in the fifth percentile (highlighted in red) being statistically significant (see Figure 2).

doi.org/10.32388/PS72PK.3

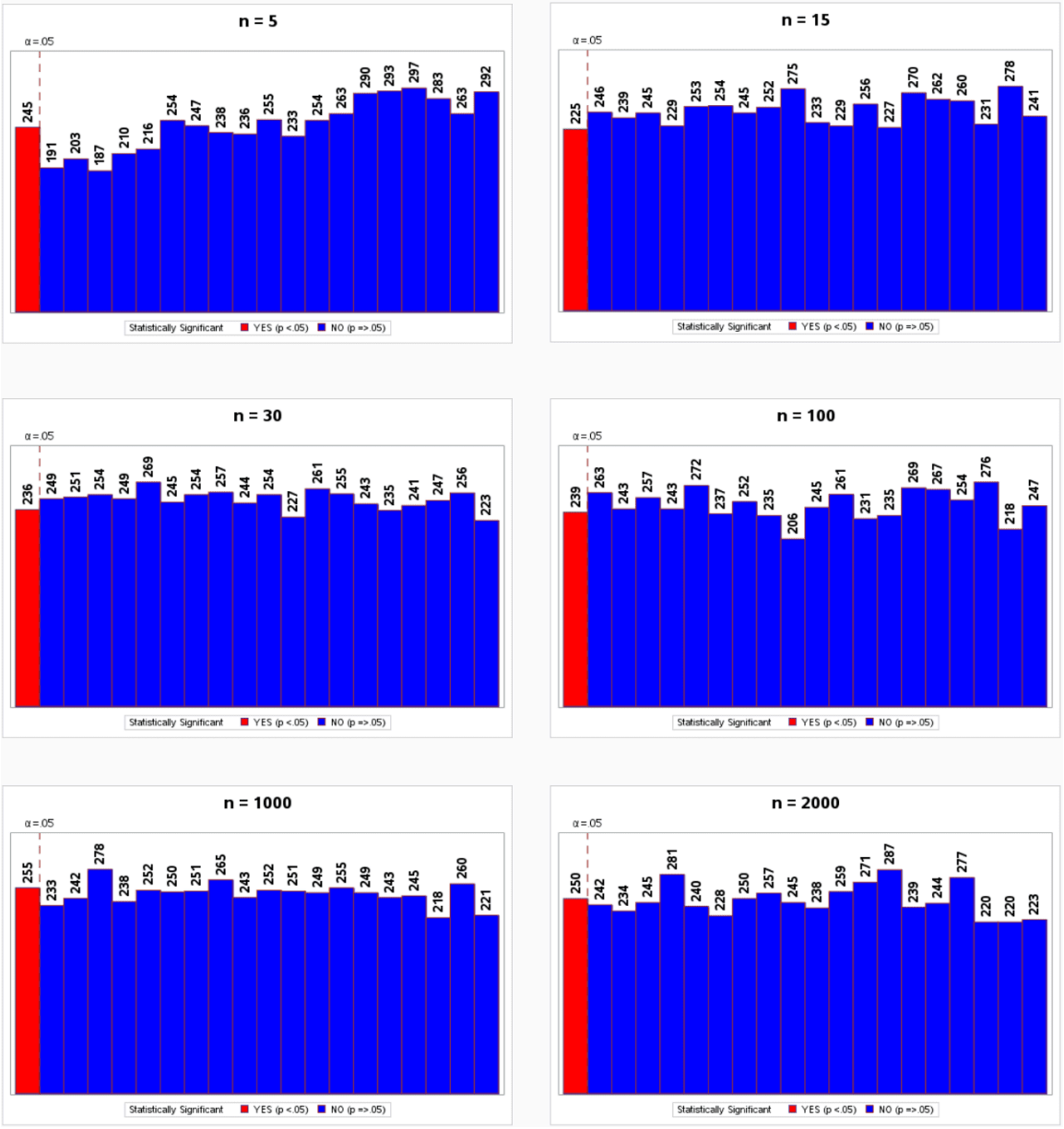

Figure 2. *Empirical Sampling Distribution of 4950 P-values by Sample Size.*

Unlike the sampling distributions of correlation coefficients, p-values do not converge on the grand central P-value (.50), which is evident from the relatively consistent standard deviations and ranges of p-values by sample size in Table 1.

The vertical bar charts in Figure 3 were created using the same correlations as in Figure 1 but now displayed as Cohen's effect sizes (small $|r| \geq 0.10$, medium $|r| \geq 0.30$, large $|r| \geq 0.50$).

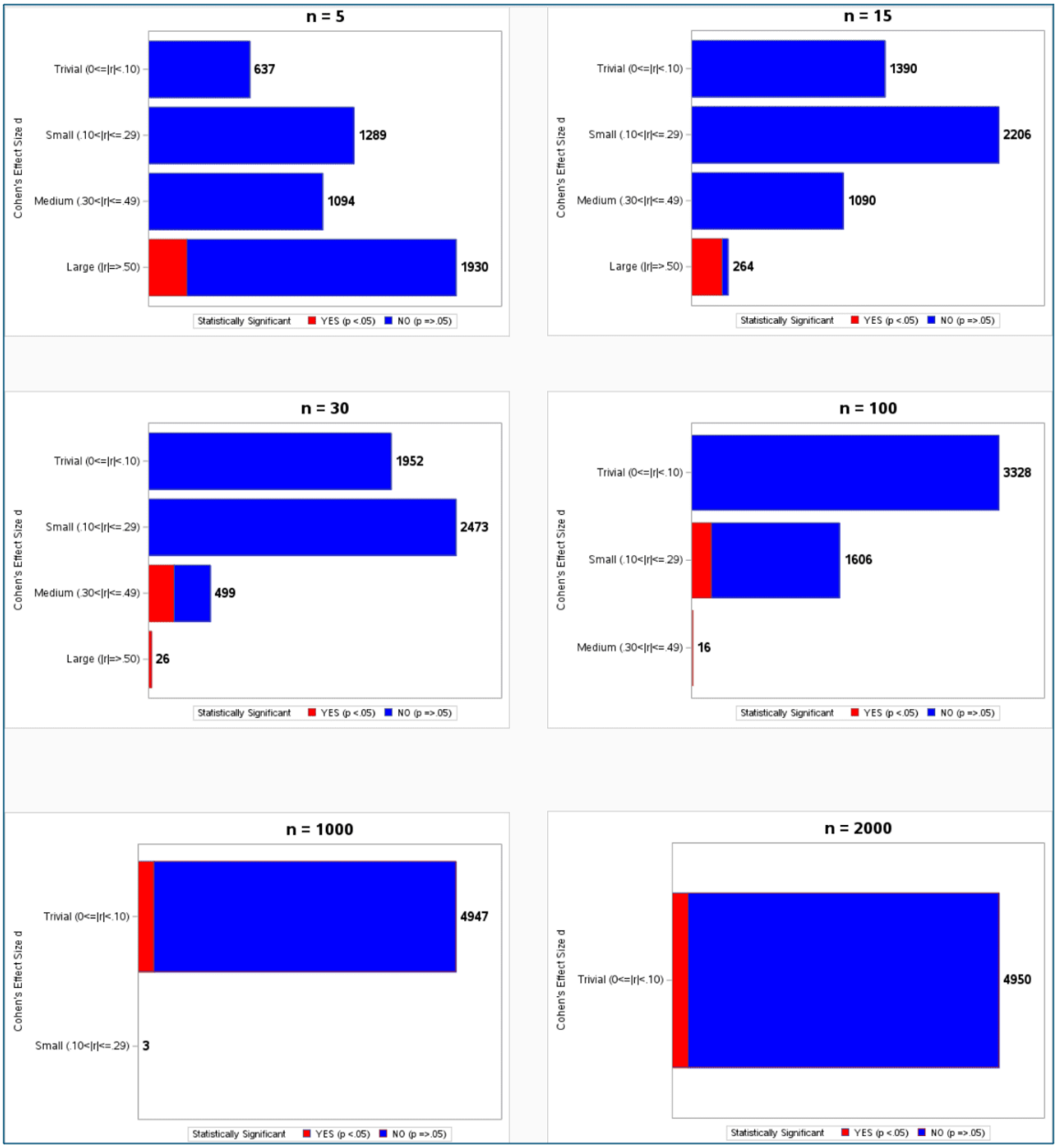

Figure 3. *Empirical Sampling Distribution of 4950 Effect Sizes by Sample Size.*

With small sample sizes, only the largest effect sizes were statistically significant (see also Tables 2 and 3).

Sample Size	Statistically Significant p-values	Cohen's Effect Sizes d	Number and Percentage of Correlation Coefficients	
n = 5	YES (p <.05)	Large (|r|=>.50)	245	4.9%
	NO (p ≥.05)	Trivial (0<=|r|<.10)	637	12.9%
	NO (p ≥.05)	Small (.10<|r|<=.29)	1289	26.0%
	NO (p ≥.05)	Medium (.30<|r|<=.49)	1094	22.1%
	NO (p ≥.05)	Large (|r|=>.50)	1685	34.0%
n = 15	YES (p <.05)	Large (|r|=>.50)	225	4.5%
	NO (p ≥.05)	Trivial (0<=|r|<.10)	1390	28.1%
	NO (p ≥.05)	Small (.10<|r|<=.29)	2206	44.6%
	NO (p ≥.05)	Medium (.30<|r|<=.49)	1090	22.0%
	NO (p ≥.05)	Large (|r|=>.50)	39	0.8%
n = 30	YES (p <.05)	Medium (.30<|r|<=.49)	210	4.2%
	YES (p <.05)	Large (|r|=>.50)	26	0.5%
	NO (p ≥.05)	Trivial (0<=|r|<.10)	1952	39.4%
	NO (p ≥.05)	Small (.10<|r|<=.29)	2473	50.0%
	NO (p ≥.05)	Medium (.30<|r|<=.49)	289	5.8%
n = 100	YES (p <.05)	Small (.10<|r|<=.29)	223	4.5%
	YES (p <.05)	Medium (.30<|r|<=.49)	16	0.3%
	NO (p ≥.05)	Trivial (0<=|r|<.10)	3328	67.2%
	NO (p ≥.05)	Small (.10<|r|<=.29)	1383	27.9%
n = 1000	YES (p <.05)	Trivial (0<=|r|<.10)	252	5.1%
	YES (p <.05)	Small (.10<|r|<=.29)	3	0.1%
	NO (p ≥.05)	Trivial (0<=|r|<.10)	4695	94.8%
n = 2000	YES (p <.05)	Trivial (0<=|r|<.10)	250	5.1%
	NO (p ≥.05)	Trivial (0<=|r|<.10)	4700	94.9%

Table 2. *Intersections of Type 1 Errors and Effect Size Errors by Sample Size*

Sample Size	Number of Statistically Significant P-values	Percentage of Statistically Significant P-values	Number of Substatantively Significant Effect Sizes	Percentage of Substantively Significant Effect Sizes	Percentage of Statistically Significant, Substantive Effect Sizes
n = 5	245	4.9%	4313	87.1%	5.7%
n = 15	225	4.5%	3560	71.9%	6.3%
n = 30	236	4.8%	2998	60.6%	7.9%
n = 100	239	4.8%	1622	32.8%	14.7%
n = 1000	255	5.2%	3	0.1%	100.0%
n = 2000	250	5.1%	0	0.0%	0.0%

Table 3. *Summary of Intersection of Statistical and Substantive Significance from Table 2.*

As sample size increased, smaller effect sizes became statistically significant. Finally, only trivial or ignorable effect sizes materialized with n = 2000, where 5.1% were statistically significant.

Table 3 reveals the benefit of screening for statistical significance before interpreting substantive significance. For example, with n = 5, there were 87.1% (4313/4950) substantive effect sizes; however, in reality, all were false or effect size errors. Statistical significance reduced the number to only 5.7% (245/4313), with the remaining 94.3% (4068/4313) excluded from consideration. Recall that the null hypothesis is true (H_0: $\rho = \rho_0 = 0.0$); therefore, Cohen's D = 0.0 is also true, so any sampled effect size other than zero is an effect size error. In any case, it would be unwise to draw a firm conclusion about a correlation based on such a small sample size. Therefore, consider a more realistic study with n = 30. More than half (60.6%) were non-trivial effect sizes, but this number is reduced to 7.9% when statistical significance is considered. An interesting and instructive phenomenon occurred with n = 1000. Here, only a tenth of a percent would be considered substantively significant, but there were 5.2% (255) statistically significant p-values. Of these, three were small effect sizes, and the remaining were trivial or ignorable. Small-sample statistical significance theory is not useful in this case. Finally, with n = 2000, it became useless because 250 (5.1%) statistically significant p-values were not substantively significant effect sizes.

Conclusion and Discussion

There are assumptions underlying the statistical test of a population correlation as specified with the null hypothesis: bivariate normality, linearity, and the absence of outliers. If these assumptions are satisfied, the sampling distribution of p-values is uniform under a true null hypothesis[16]. Therefore, any p-value in the open interval from 0.0 to 1.0 materializes by chance regardless of sample size under a true null hypothesis. This reveals an underappreciated fact about power calculations. When the null hypothesis is true, increasing the sample size does not increase power (produce more statistically significant p-values). Increasing sample size increases power only when the null hypothesis is false.

Greenland et al.[3] stated: "Every method of statistical inference depends on a complex web of assumptions about how data were collected and analyzed, and how the analysis results were selected for presentation" (p. 338) Wasserstein and Lazar[4] warned against a naïve and single-minded obsession with a statistically significant p-value: "Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" (p. 9). Indeed, ignoring such considerations distorts or skews the shape of the p-value sampling distribution under a true null hypothesis, thereby invalidating α that was declared a priori. Incidentally, α does not have to be 5% (2-sigma), for example, a 5-sigma is used in theoretical physics and engineering.

Student[8] and Fisher[9] developed a small-sample statistical theory that relies on statistical significance for decision-making. In his later writings, Fisher[17] viewed statistical significance as a guide or milestone and not the end of the research project: “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). Nonetheless, statistical significance has been criticized for contributing to the replication crisis[18]. Ironically, the solution is not a ban on statistical significance, but replication of statistical significance with fresh, new data[19]. With small sample sizes, large effect sizes can occur by chance under a true null hypothesis of $\rho = 0$, thus warranting replication even if statistically significant. With large sample sizes, a substantively significant correlation is unusual under a true null hypothesis with $\rho = 0$; therefore, the study also merits replication regardless of statistical significance. For a similar conclusion involving differences in means or Cohen’s effect size d under a true null hypothesis, see Komaroff[20].

Statements and Declarations

Data Availability

The author guarantees that the results in this paper are reproducible and replicable. Reproducing the results in this manuscript is possible by downloading the data sets and analysis files used in writing this paper from the public repository at https://figshare.com or by contacting the author with a reasonable request at komaroffeugene@gmail.com. Replicating the results can be easily achieved because the computer clock time initiates the random data sampling.

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. ^*Efron B (1998). "R. A. Fisher in the 21st Century (Invited Paper Presented at the 1996 R. A. Fisher Lecture)." Statis t Sci. **13**(2):95–122. doi:10.1214/ss/1028905930.*
3. ^a, ^b*Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016). "Statistical Tests, P Val ues, Confidence Intervals, and Power: A Guide to Misinterpretations." Eur J Epidemiol. **31**:337–350.*

4. ^a, ^b*Wasserstein RL, Lazar NA (2016). "The ASA's Statement on P-Values: Context, Process, and Purpose." Am Stat. doi:10.1080/00031305.2016.1154108.*
5. ^*Amrhein V, Greenland S (2018). "Remove, Rather Than Redefine, Statistical Significance." Nat Hum Behav. 2:4. doi:10.1038/s41562-017-0224-0.*
6. ^a, ^b*Trafimow D, Marks M (2015). "Editorial." Basic Appl Soc Psych. **37**(1):1–2. doi:10.1080/01973533.2015.1012991.*
7. ^*Wasserstein RL, Schirm AL, Lazar NA (2019). "Moving to a World Beyond P < 0.05." Am Stat. **73**(sup1):1–19.*
8. ^a, ^b*Student (1908). "Probable Error of the Mean." Biometrika. **6**(1):1–25.*
9. ^a, ^b, ^c, ^d*Fisher RA (1970). Statistical Methods For Research Workers. 14th ed. Oxford: Oxford University Press.*
10. ^*Moore DS, Notz WI, Fligner MA (2021). The Basic Practice of Statistics. 9th ed. New York: W. H. Freeman.*
11. ^a, ^b*SAS Institute Inc. (2019). SAS/STAT 9.4 User's Guide. Cary: SAS Institute Inc.*
12. ^*Cohen J (1968). Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates.*
13. ^*Komaroff E (2020). "Relationships Between P-Values and Pearson Correlation Coefficients, Type 1 Errors and Effect Size Errors, Under a True Null Hypothesis." J Stat Theory Pract. **14**:49. doi:10.1007/s42519-020-00115-6.*
14. ^*SAS Institute Inc. (2014). SAS OnDemand For Academics: User's Guide. Cary: SAS Institute Inc.*
15. ^*Hogg RV, McKean JW, Craig AT (2013). Introduction to Mathematical Statistics. Boston: Pearson Education, Inc.*
16. ^*Westfall PH, Tobias RD, Wolfinger RD (2011). Multiple Comparisons and Multiple Tests Using SAS. 2nd ed. Cary: SAS Institute, Inc. Press. doi:10.17226/25303.*
17. ^*Fisher RA (1973). Statistical Methods and Scientific Inference. 3rd ed. New York: Hafner.*
18. ^*Ioannidis JP (2005). "Why Most Published Research Findings Are False." PLoS Med. **2**(8):e124.*
19. ^*National Academies of Sciences, Engineering, and Medicine (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. doi:10.17226/25303.*
20. ^*Komaroff E (2025). "A Redemption Song for Statistical Significance." Qeios. https://www.qeios.com/read/3QQSNC.5.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.