Journal of Statistical Theory and Practice (2020) 14:49
https://doi.org/10.1007/s42519-020-00115-6

ORIGINAL ARTICLE

Relationships Between *p*-values and Pearson Correlation Coefficients, Type 1 Errors and Effect Size Errors, Under a True Null Hypothesis

Eugene Komaroff[1]

Published online: 26 June 2020
© Grace Scientific Publishing 2020

Abstract
The American Statistical Association (ASA) published a statement in 2016 in *The American Statistician* for "researchers, practitioners and science writers who are not primarily statisticians" on the proper use and interpretation of *p*-values. Three years later, the ASA acknowledged that telling researchers what not to do with *p*-values was insufficient. Consequently, 3 years later an open access, special issue appeared with 43 papers proposing various novel and sophisticated alternatives to classical *p*-values for use with scientific methods in the twenty-first century. In the opening remarks, the editors stated that "no *p*-value can reveal the plausibility, presence, truth, or importance of an association or effect" and banned statistical significance: "don't say it, don't use it." This paper questions both statements with a simulated data study. It is shown that *p*-values are strongly related to correlation coefficients under a true null hypothesis; hence, can reveal the "importance of an association or effect." Furthermore, it demonstrates why a cut point for statistical significance is still a viable, ancillary tool for assessing the substantive significance of statistical effects with small sample sizes ($n < 1000$).

Keywords *p*-values · Statistical significance · Substantive significance · Pearson correlation coefficient · Fisher *r* to *z* transformation · Effect size

1 Introduction

The American Statistical Association (ASA) published a statement in *The American Statistician* to address the confusion, misinterpretation and abuse of *p*-values, noting that "statisticians and others have been sounding the alarm about these matters for decades, to little avail" [1, p. 5]. The ASA acknowledged their paper underscored

✉ Eugene Komaroff
komaroffeugene@gmail.com

[1] Keiser University Graduate School, 1900 W. Commercial Blvd., Fort Lauderdale, FL 33309, USA

the problems but offered little in terms of solutions. Consequently, 3 years later an open access, special issue of *The American Statistician* was published with 43 papers proposing various sophisticated, alternatives to classical *p*-values. In their opening remarks, the editors proclaimed that a "*p*-value, or statistical significance, does not measure the size of an effect or the importance of a result." Although they acquiesced with the use of *p*-values on a continuous scale, they called for a ban on the concept of statistical significance: "don't say it and don't use it [2, p. 2]. This paper demonstrates that *p*-values are a measure of the importance of an effect and that statistical significance is still a viable concept. However, this paper does not dismiss the sound warnings against the indiscriminate use of *p*-values by researchers who ignore potential systematic errors and other confounding design issues [3].

There is no doubt that the twenty-first century of big data requires innovative alternatives to classical frequentist statistical techniques [4]. Nonetheless, as Fisher remarked: "it is with small samples, less than 100, that the practical research worker ordinarily wishes to use the correlation coefficient" [5, p. 195]. Furthermore, "it is not true…that valid conclusions cannot be drawn from small samples, if accurate methods are used in calculating the probability" [5, p. 198].

Fisher [6] developed the classical concepts of *p*-values and statistical significance at the beginning of the twentieth century as tools for use by applied researchers who worked with small sample sizes ($n < 1000$). These tools remain valuable in the twenty-first century. This paper demonstrates how these tools work with correlational analyses. Contrary to the ASA statement, this paper shows that *p*-values do reveal the size of an effect or importance of a result. Under a true null hypothesis, correlation coefficients and their corresponding *p*-value are strongly and inversely related. As correlations increase to unity ($|r| = 1.0$), their corresponding *p*-values decrease to zero. Furthermore, the ASA ban on statistical significance is unwise because the ban surely opens the correlational research literature to even more irreproducible results [7].

2 Methods

SAS Base Procedures [8] were used with the University Edition of SAS for WINDOWS [9] for simulating raw data and subsequent analyses. The simulations consisted of drawing independent, identically distributed normal random variables from a standard normal distribution ($\mu = 0$, $\sigma = 1$). The number of variables (k) could be viewed as items on a survey that presumably measured a common construct. The number of items increased from 5 to 30 by 5 ($k = 5$, 10, 15, 20, 25, or 30). Also, there were 13 sample size conditions starting with $n = 10$, increased by 5–50, and finally 100, 500 and 1000. Imagine the variables as items on surveys that varied in length from 5 to 30 items and administered to 13 groups comprised by different number of people.

Each k and n combination produced its own unique empirical sampling distribution of bivariate correlations. For example, with $n = 10$, there were 10 bivariate correlations (${}_5C_2$) with $k = 5$. Next, with $n = 15$, a new correlation matrix with 10 unique pairwise correlations. This process was followed until $n = 1000$. The theoretical center or expected value of each sampling distribution was zero ($\rho = 0$) because all correlations were computed with independent, identically distributed, normal random variables drawn from the standard normal distribution ($\mu = 0$, $\sigma = 1$).

The p-value for each correlation coefficient was computed with the z-test using the Fisher transformation of r to z_r to test the null hypothesis H_0: $\rho = \rho_0$, with $\rho_0 = 0$. In the notation, z_r represents the value of the correlation coefficient that results from applying the Fisher transformation, which is the hyperbolic tangent function. The major benefit of the transformation is the conversion of a potentially skewed r sampling distribution to a relatively normal z_r sampling distribution. As a result, the z-test with the z_r statistic is used to determine the significance of an observed correlation coefficient. The standard error, which is discussed next, plays a crucial role for detection of statistical significance.

Gosset ('Student') [10] focused on differences in means in his statistical research and recognized that estimates of the "standard errors of the mean ($\frac{\sigma}{\sqrt{n}}$) can be obtained by replacing the unknown population standard deviation (σ) with the known sample estimate (s). The use of n in the denominator was before Fisher developed the degrees of freedom concept for small sample sizes. Gosset proposed two ways for dealing with the uncertainty of the standard error: "An experiment may be repeated many times, until such a long series is obtained that the standard deviation is determined once and for all with sufficient accuracy" [10, p. 2]. Gosset recognized that often it was not easy to repeat the same experiment many times, so the sample standard deviation is used for the calculation of the standard error of the mean. However, the sample standard deviation as an estimate of the population standard deviation is subject to sampling error. Student's t-distributions take this variability into account by stretching the tails of the t-distribution beyond the z distribution for small sample sizes ($n < 1000$).

Fisher [6] explained that the standard error of the correlation coefficient was derived from large sample theory:

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{n - 1}} \tag{1}$$

and calculated as

$$s_r = \frac{1 - r^2}{\sqrt{n - 1}} \tag{2}$$

where "with small samples the value of r is often very different from the true value, ρ, and the factor $1 - r^2$, correspondingly in error; in addition the distribution of r is far from normal, so that tests of significance based on the large sample formula are often very deceptive" [6, p. 195]. As a result, Fisher recommended the standard error that was used to test for significance of an r with the Student's t-test

$$t = (n - 2)^{\frac{1}{2}} \left(\frac{r^2}{1 - r^2} \right)^{\frac{1}{2}} \tag{3}$$

where the probability is computed by treating t-statistic as coming from a t sampling distribution with $(n-2)$ degrees of freedom. SAS uses this formula for computing r and the corresponding p-value with PROC CORR [8, p. 20].

Nonetheless, Fisher [6] developed a simpler standard error (σ_z) for his r to z transformed correlation (z_r)

$$\sigma_z = \frac{1}{\sqrt{n-3}} \tag{4}$$

He refers to z_r as simply z in the following where he promotes the benefits of this standard error because it is approximately:

> independent of the value of the correlation in the population from which the sample was drawn. In the second place, the distribution of r is not normal in small samples, and even for large samples it remains far from normal for high correlations The distribution of z is not strictly normal, but it tends to normality rapidly as the sample is increased, whatever maybe the value of the correlation [6, pp. 200–201].

In short, although the sampling distributions of r could be skewed if $\rho \neq 0$, the distribution of z_r is approximately normal for all values of ρ. However, because the mean of the z_r sampling distribution (ζ) is slightly bigger than the mean r of the theoretical sampling distribution (ρ), especially when ρ is near the boundary, Fisher developed a bias adjustment $\frac{\rho_0}{2(n-1)}$ (see 5).

For this paper, all statistical tests of the null hypothesis were conducted by first applying Fisher's transformation to both an observed r and ρ_0 such that $z_r = \tanh^{-1}$ (r) and $\zeta_0 = \tanh^{-1}$ (ρ_0), where $\tanh^{-1}$ is the inverse hyperbolic tangent function. The standard error of Fisher's z appears in the denominator of the z-test

$$z = \frac{z_r - \zeta_0 - \frac{\rho_0}{2(n-1)}}{\frac{1}{\sqrt{n-3}}} \tag{5}$$

SAS [8] states that the bias adjustment is always used for computing p-values for the z-test with the Fisher option in PROC CORR. However, the default null hypothesis is H_0: $\rho = \rho_0$ where $\rho_0 = 0$, so the bias adjustment by default is zero. For other values of ρ_0, the bias adjustment becomes increasingly negligible as sample size increases. For example, for $\rho_0 = .30$ and $n = 10$ the bias adjustment is 0.01667 but with $n = 1000$ the adjustment is relatively trivial 0.00015.

With large sample sizes (i.e., $n \geq 1000$), provided assumptions are satisfied (linearity, bivariate normality and no outliers), any observed z_r is a precise estimate of ζ, the center of Fisher's z sampling distribution. Similarly, it follows that any observed r is also precise estimate of ρ, the center of Pearson's r sampling distribution. This can be deduced from the standard errors of each distribution. For example, suppose an $r = .30$ was found with $n = 1000$. Fisher's transformation produces $z_r = .31$ with standard error 0.032 (4). Because this z_r distribution is normal, 99.7% of the values are between $-.095$ and .095. This is a narrow range revealing that almost all z_r

are near a fixed, central parameter. An estimate of ζ was not needed for calculating Fisher's standard error, but the large sample value for standard error of r requires ρ (1). Just like Gosset's [10] sample standard deviation was inserted for the population standard deviation to calculate a standard error of the mean; the observed r serves as the estimate of the population parameter ρ (2). For instance, with $r = .30$, the standard error (σ_r) is 0.029 where the theoretical sampling distribution of r is assumed normal because of the large sample size. As a result, 99.7% of the correlations in this sampling distribution are between $-.39$ and .39, or very near $\rho = .30$. It is important to recognize the role of sample size in these calculations. For instance, with $n = 10$ and $\rho = .30$ the standard error (σ_r) is 0.303, whereas with $n = 1000$ it is 0.029. The increase in sample size from 10 to 1000 produced a 91% reduction in the standard error, hence, provided all assumptions are satisfied, and all observed correlations are excellent and precise estimates of ρ. The same cannot be said with small sample sizes as will be demonstrated next with simulated data.

To assess the nature of the relationship between statistical significance and substantive significance, the cut point for statistical significance was $\alpha = 0.05$. Because it was known that all simulated correlations were derived from the null sampling distribution that was centered at zero, all statistically significant correlations were type 1 errors. For determining substantive significance, Cohen's [11] criteria for correlation coefficients as effect sizes ($|r|$) was utilized where a small effect size was $r \geq .10$ to $< .30$, medium effect was $r \geq .30$ to $< .50$, and large effect was $r \geq .50$. Similarly, because all simulated correlations were random realizations of a null parameter, all substantively significant correlations were effect size errors (ES errors).

3 Results

3.1 Relationship Between *p*-values and Correlation Coefficients

Table 1 presents just the summary statistics for sampling distributions of 435 correlation coefficients over 13 sample size conditions. This table demonstrates the properties that also applied to all the other sets of correlation coefficients (data not shown). Table 1 reveals close agreement between theoretical and empirical standard errors as sample size increases. In calculating the large sample theoretical standard errors of r (2), the mean ($\bar{r}$) of each sampling distribution was the estimate of ρ because each individual r was subject to more sampling error (evident from the min and max values).

In contradistinction to the ASA statement [2] that p-values cannot reveal the importance of an association or effect, the scatterplots in Fig. 1 revealed strong monotonic relationships between p-values and their corresponding Fisher's z values.

Figure 1 displays consistent patterns. As the number of correlation coefficients increased, the shapes of the relationships between r and their corresponding Fisher's r to z p-values came into sharper focus. For example, with $k = 5$ there was 10 unique bivariate correlations for each sample size condition, but with $k = 30$ there were 435 correlations. Like pixels in a photograph, more dots (coordinate pairs of Fisher's

Table 1 Summary statistics for theoretical and empirical (simulated) sampling distributions of 435 correlation coefficients over 13 sample size conditions

k	n	Number of correlations	Variables	Theoretical			Empirical				
				Mean	SE	Variance	Mean	SD	Variance	Min	Max
30	10	435	Pearson's r	0.00	0.32	0.10	0.03	0.34	0.11	−0.84	0.87
			Fisher's z	0.00	0.38	0.14	0.04	0.38	0.15	−1.21	1.34
30	15	435	Pearson's r	0.00	0.26	0.07	0.01	0.34	0.12	−0.78	0.83
			Fisher's z	0.00	0.29	0.08	0.01	0.38	0.15	−1.03	1.18
30	20	435	Pearson's r	0.00	0.22	0.05	−0.02	0.26	0.07	−0.68	0.69
			Fisher's z	0.00	0.24	0.06	−0.02	0.28	0.08	−0.83	0.86
30	25	435	Pearson's r	0.00	0.20	0.04	0.00	0.23	0.05	−0.53	0.63
			Fisher's z	0.00	0.21	0.05	0.00	0.24	0.06	−0.59	0.74
30	30	435	Pearson's r	0.00	0.18	0.03	0.00	0.21	0.04	−0.65	0.63
			Fisher's z	0.00	0.19	0.04	0.00	0.21	0.05	−0.77	0.74
30	35	435	Pearson's r	0.00	0.17	0.03	−0.01	0.20	0.04	−0.49	0.50
			Fisher's z	0.00	0.18	0.03	−0.01	0.21	0.04	−0.54	0.55
30	40	435	Pearson's r	0.00	0.16	0.02	0.01	0.17	0.03	−0.44	0.49
			Fisher's z	0.00	0.16	0.03	0.01	0.17	0.03	−0.48	0.54
30	45	435	Pearson's r	0.00	0.15	0.02	0.01	0.16	0.03	−0.40	0.50
			Fisher's z	0.00	0.15	0.02	0.01	0.16	0.03	−0.43	0.54
30	50	435	Pearson's r	0.00	0.14	0.02	0.00	0.15	0.02	−0.39	0.48
			Fisher's z	0.00	0.15	0.02	0.00	0.16	0.03	−0.41	0.52
30	75	435	Pearson's r	0.00	0.12	0.01	0.00	0.14	0.02	−0.46	0.43
			Fisher's z	0.00	0.12	0.01	0.00	0.15	0.02	−0.50	0.46
30	100	435	Pearson's r	0.00	0.10	0.01	−0.01	0.11	0.01	−0.33	0.35
			Fisher's z	0.00	0.10	0.01	−0.01	0.12	0.01	−0.35	0.36
30	500	435	Pearson's r	0.00	0.04	0.00	0.00	0.04	0.00	−0.11	0.11
			Fisher's z	0.00	0.04	0.00	0.00	0.04	0.00	−0.11	0.11
30	1000	435	Pearson's r	0.00	0.03	0.00	0.00	0.03	0.00	−0.08	0.09
			Fisher's z	0.00	0.03	0.00	0.00	0.03	0.00	−0.08	0.09

z and p-value) produced higher resolution. Notice there is no "scatter" because r and p-values are in a monotonic one-to-one relationship. Each scatterplot includes a dashed reference line at $p = 0.05$ below which a correlation coefficient would be considered statistically significant.

The visual impression of a strong relationships was confirmed with 36 second order polynomial regression models where. Fisher's z values were the dependent variable and their corresponding p-values and their squares were two independent variables. There were 18 regression models run on the subset of only positive correlations and another 18 for only negative correlations. Although not perfect, the regression models confirmed the visual impression of strong relationships. R-square across all 36 models ranged from .93 to .99 and because a Fisher's z can be inversely transformed to a Pearson's r, strong monotonic (albeit nonlinear) relationships between r and p-values must also exist on the Pearson r scale.

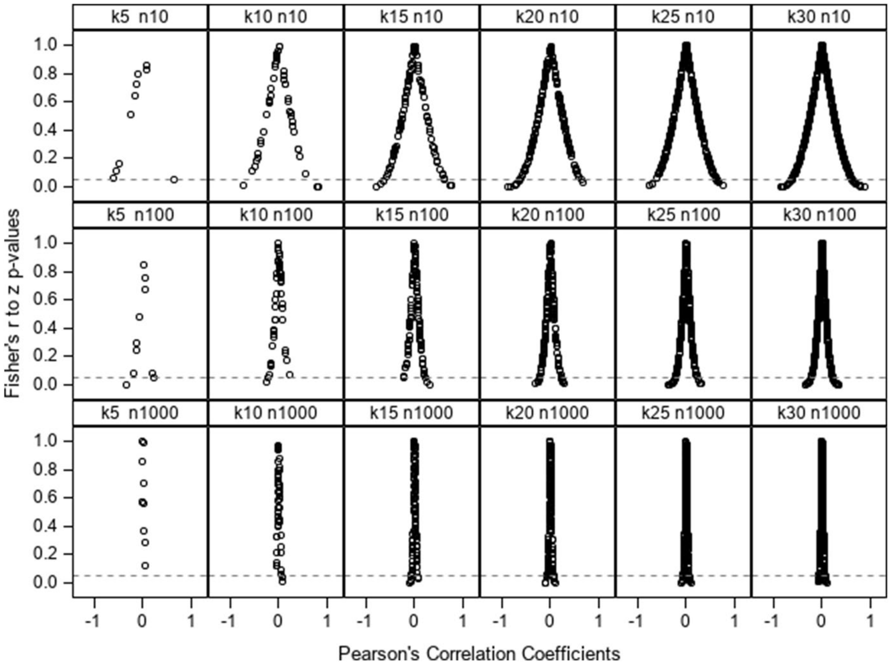

Fig. 1 Scatterplots of Pearson's correlation coefficients and corresponding Fisher's r to z p-values for select sample sizes and number of correlation coefficients

3.2 Statistical Significance of Pearson Correlation Coefficients

Fisher defined statistical significance as "the value for which $p = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not… small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty" [6, p. 44]. Fisher most likely meant small sample size by "insufficiently numerous." For example, to detect a small effect such as $\rho = .20$, the sampled r is an estimate ($\hat{\rho}$) of the small effect. With alpha = 0.05 and $n = 193$, there is 80% power to detect the difference $\hat{\rho} - \rho_0$ as significantly different from zero. However, with only $n = 10$, with all else being equal, there is only 8% power to detect the same small effect size; hence, this small effect will escape notice.

Yet, another way to understand Fisher's "insufficiently numerous" could be in terms of number of estimates. For instance, a sampling distribution of 10 correlations with $n = 10$ produced an empirical range from $-.61$ to .63, compared to $-.84$ to .87 that was found with $n = 10$ but now with 435 correlations. In short, even though the standard errors (4) were the same for both scenarios, the larger correlation matrix simply contained more correlations to evaluate. Of course, testing 435 correlations with the same null hypothesis introduces a serious alpha inflation problem. Fortunately, the problem can be mollified with statistical techniques like the Bonferroni correction among others [12]. For example, a Bonferroni correction with $\alpha = 0.05$ and 435 statistical tests of the same null hypothesis, a p-value < 0.00011 is required for statistical significance.

With $\alpha = 0.05$, statistical significance is defined as an observed correlation that deviates from a hypothesized ρ_0 by at least two (1.96) standard error units. It is important to recall that in this study, all r's were obtained from a sampling distribution that was centered at $\rho = 0$ because they were calculated with independent, identically distributed random variables. Therefore, any detection of a statistically significant difference between an observed r ($\hat{\rho}$) and any postulated parameter (ρ_0) was a type 1 error. A statistically significant r could still have come from the null sampling distribution centered at zero despite a very small probability of such an event [13]. Nonetheless, a significant correlation is interpreted as a correlation that was obtained from an alternative sampling distribution, or one that is not centered at the postulated null parameter. But what is the center of this alternative distribution? Unfortunately, this question cannot be answered with classical statistical tests. Confidence intervals for estimating parameters may help, but just like statistical significance, there is always some doubt, especially with small sample sizes. For instance, if there is a 95% chance the confidence interval covers the fixed alternative parameter, there remains a 5% chance that every value contained in the confidence interval is wrong. This error cannot be eliminated by increasing sample size and furthermore if alpha inflation is induced by multiplicity or p-hacking, that increases the probability of a wrong conclusion [14].

3.3 Null Parameters Other Than Zero

It is possible to test for other values of ρ_0 besides 0 with a z-test. PROC CORR permits the specification of any reasonable ρ_0 between but not including -1.0 and 1.0. Say a researcher observed an $r = .10$ and the null hypothesis stated $\rho_0 = .30$. Figure 2 demonstrates the probability of detecting statistically significant correlations from sets of 435 correlations with ρ_0 set equal to .30 under three sample size conditions. As sample size increased, many more significant correlations were detected. With $n = 1000$, all correlations were significant or were more than two standard errors away from .30. To understand this phenomenon, recognize that Fisher's z standard error is 0.032 with $n = 1000$. The difference between the observed r ($\hat{\rho}$) and the null parameter ρ_0 is approximately .2 on the Fisher's z scale. The z-test reveals that .2 divided by σ_z (4) corresponds to a z score of 6.61, which is significantly different from zero. Notice the x-axis for the $n = 100$ figure ranges from $-.1$ to .1 therefore any differences smaller than .10 or bigger than .10 (such as .2) must be statistically significant.

The dashed reference line at .30 indicates the hypothesized population rho that was tested for statistical significance. Nevertheless, there is potential for a confusing and misleading conclusion with classical statistical reasoning. After the null is rejected, all possible correlations, except for the one specified under the null (.30), became plausible estimates of the fixed but unknown alternative parameter. In this case, zero is also plausible and a researcher could mistake this finding as proof of

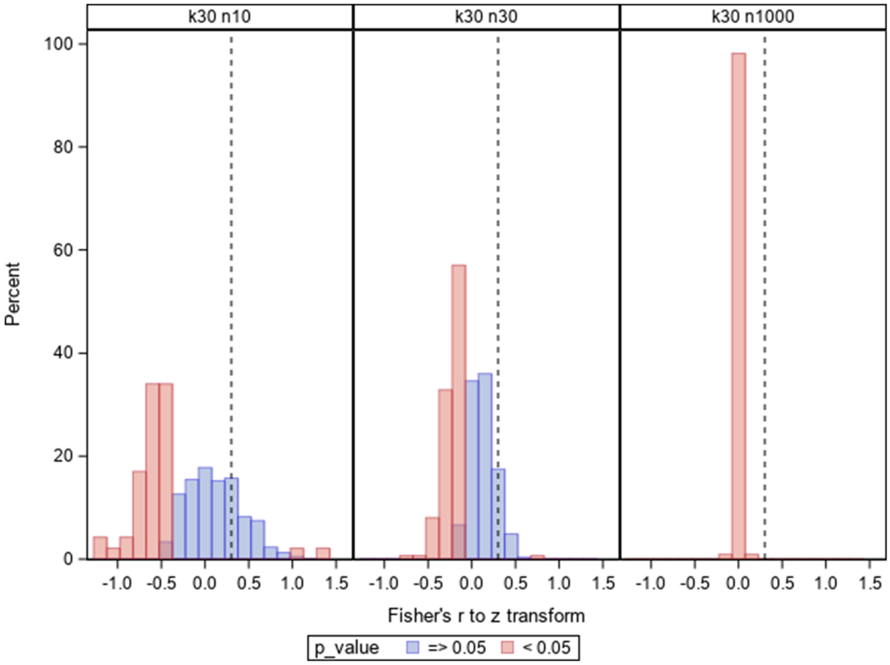

Fig. 2 Three empirical sampling distributions with 435 correlation coefficients and $n = 10$, 30 or 1000

no relationship. It is bad enough that applied researchers seem to mistake $p \geq 0.05$ as indicating the null is true or there is "no relationship." This misunderstanding is pervasive [1]. If the goal is to find evidence of no relationship, then an appropriate equivalence testing is required [15].

3.4 Relationship Between Statistical and Substantive Significance

Because all observed r came from a sampling distributions of correlations (population) that was known to be centered at zero, rejection of the null (acceptance of the alternative hypothesis, H_a: $\rho \neq \rho_0$) was a type 1 error. However, suppose a researcher decided to abide by the ASA's ban on statistical significance and evaluated the observed r using only Cohen's [11] criteria for correlation coefficients as effect sizes. Figure 3 demonstrates the most likely outcomes under different sample size conditions with 435 correlation coefficients.

The dashed line in Fig. 3 represents "ES errors" that resulted from flagging any correlations smaller than $-.10$ or bigger than .10 as substantive or meaningful based on Cohen's [11] criteria. In each sample size condition, the percentage of ES errors was the count that fit Cohen's criteria divided by the total number of correlations(435). Statistically significant r had corresponding p-values < 0.05 based on Fisher's r to z transformation. Figure 3 reveals a high percentage of ES errors for small sample sizes that progressively diminished and finally disappeared with

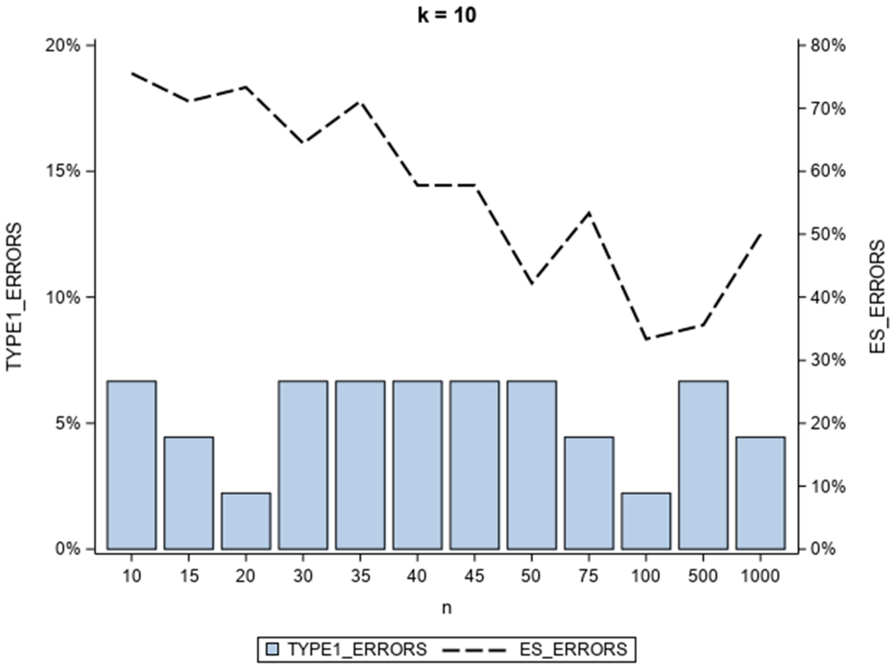

Fig. 3 Percentage of wrong rejections of null hypothesis (type 1 error), based on $p < 0.05$) and wrong acceptance of effect sizes (ES error) based on Cohen's criteria for correlations between |0.1| and |1.0|

$n = 1000$. On the other hand, statistical significance (p-values < 0.05) remained relatively stable around 5% for all sample size conditions.

Provided all assumptions are met, p-values follow a uniform distribution under a true null; therefore, there is always a 5% chance of getting a p-value < 0.05 under the null hypothesis. However, the percentage of significant correlations increases when the alternative distribution is truly not zero. This is the fundamental concept that underlies the power calculations for correlation coefficients [16]. For sample size calculations, a researcher needs to postulate a specific parameter value (center) for both the null sampling distribution as well as the alternative sampling distribution of correlations. The difference between these two centers produces the required sample size that would yield, typically, 80% significant p-values in the long run. The problem here is that the center of the null distribution is known to be "0," but the center of the alternative distribution cannot be a wild guess [17] nor driven by ancillary considerations such as the cost of recruiting participants.

4 Conclusion

Although there is a strong relationship between p-values and their corresponding correlation coefficients that says nothing about the important role that statistical significance plays for statistical inference. The ASA ban on statistical significance denies "researchers, practitioners and science writers who are not primarily statisticians," (the audience the ASA [1] was addressing with their p-value principles) a very simple, effective statistical tool for probabilistic decision-making. Regardless if $\alpha = 0.05$ or any other level, statistical significance excludes many misleading and meaningless correlations that will appear randomly with small sample sizes. For example, for $n = 10$ with 435 correlations, 81% (320/435) fit Cohen's criteria (ES $\geq .10$), yet only 6% (22/320) were statistically significant. Even if alpha were set at 0.50, that would mean about 50% of the correlations would have been significant under the true null; however, that would still be a sizeable reduction in the number of meaningless correlations or false effect sizes. It appears the ASA ban on statistical significance will end up stoking the reproducibility crisis with more unreliable nonsense.

Standard error is the fundamental concept that underlies statistical significance. Researchers should be discouraged from simply reporting statistically significant p-values. Say a researcher had $n = 10$ and discovered an almost perfect correlation between two continuous variables ($r = .998$, $z_r = 3.45$, $p < 0.001$), which is statistically significant. (Note, z_r is not the z-statistic, which is $z_{r/}\sigma_r = 9.13$). The researcher proceeds to speculate about this strong correlation without understanding that the standard error (σ_z) is 0.378. Recognizing the large standard error should make it difficult to crown $r = .998$ as the true population correlation when 99.7% of the sampling distribution of z_r under the null lies between $-.74$ to .74. Because of the small sample size ($n = 10$) in a single study the range of possible values precludes premature conclusions that any observed r with a small sample size is a good estimate of an alternative population correlation. This is like getting 8 heads in 10 tosses of a two-sided coin that was hypothesized to be fair. The binomial probability mass function returns a two-tailed $p = 0.0273$ of getting 8 or more heads in 10 tosses. Rejecting the null hypothesis that the coin is fair implies only that the coin maybe biased. This conclusion would need to be strengthened considerably by more tosses (replications/sample size).

On the other hand, with a large sample size such as $n = 1000$ the standard error is .032, so there is little variation in the sampling distribution of correlation coefficients. Suppose a researcher discovers a weak but statistically significant correlation [$r = 0.065$ (1000), $p = 0.0396$]. The researcher would still need to explain why such a tiny correlation is meaningful (notice does not even fit Cohen's small effect size criteria). The simulated data revealed that both statistical and substantive significance are important considerations for correlational analyses with small sample sizes ($n < 1000$). There are important issues besides standard errors that affect statistical significance, such as ignoring obvious violation assumptions, multiplicity induced alpha inflation also known as p-hacking [7], but these are ethical transgressions that are beyond the scope of this paper.

The simulated data produced results that are well understood in statistical theory. There is a mathematical link between t and r as seen in the t-test formula (3). Because t maps to a specific p-value indexed by degrees of freedom ($n - 2$), it is obvious that p-values and correlations coefficients are related. Surely this relationship between p-values and their corresponding correlation coefficients must be known by mathematical statisticians; therefore, it is puzzling to read the ASA statement that *p*-values do not measure the size of an effect. This paper demonstrated that *p*-values are related to correlation coefficients therefore must also measure the size of an effect. Furthermore, perhaps the ASA ban on statistical significance was their attempt to rescue researchers from the common misinterpretation that *p*-values are the probability of the null hypothesis given the data. However, that confusion should be corrected with education, not by eradication of statistical significance. The confusion between the correct interpretation (pr. D|NH) and the incorrect one (pr. NH|D) could stem from characterizing the null hypothesis as being either "true or false," as typically stated in textbooks. Applied researchers most likely understand true or false as a dichotomy. Therefore, they believe if *p*-value <0.05 indicates the null hypothesis is likely to be false; then, *p*-value ≥ 0.05 must indicate that the null hypothesis is likely to be true. This intuitive thinking is logical but does not apply to classical statistical reasoning.

4.1 Recommendations for Moving to a World Beyond $p<0.05$

Applied researchers do not need to know Kolmogorov's probability axioms to understand statistical significance. A *p*-value <0.05 indicates the observed statistic (such as a correlation coefficient), probably came from an alternative sampling distribution. Of course, that decision does not immediately mean that the observed r is the center of the alternative distribution. Estimating the center requires a big random sample ($n \geq 1000$) and a confidence interval. On the other hand, a *p*-value ≥ 0.05 indicates the statistic probably came from the null sampling distribution centered at zero as specified with the null hypothesis. However, if the researcher still believes the statistic came from an alternative sampling distribution, the study should be re-designed with new and more data. Fisher stated that "in relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" [18, p. 14]. We should also know how to conduct observational studies (such as surveys) that rarely fail to produce evidence of reliable correlations.

The editors of the New England Journal of Medicine took measured opposition to ASA's diminution of *p*-values and ban on statistical significance: "Despite the difficulties they pose, *p* values continue to have an important role in medical research, and we do not believe that *p* values and significance tests should be eliminated altogether" [19, p. 286]. Regarding $p<0.05$ as the level of significance, Fisher offered the following refreshing perspective: "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" [5, p. 45]. Wherever the cut point, statistical significance serves a valuable purpose for correlational analyses with small sample sizes. Provided all assumptions are satisfied, statistical significance reduces the number of meaningless and irreproducible correlations (effect size errors) that are inevitable with small sample sizes.

Funding Not applicable.

Availability of Data and Material Yes, SAS macros with fixed seed will reproduce data.

Compliance with Ethical Standards

Conflict of interest The author declares that he has no conflict of interest.

Code Availability Yes, SAS analysis program.

References

1. Wasserstein RL, Lazar NA (2016) The ASA statement on *p*-values: context, process, and purpose. Am Stat 70(2):129–133
2. Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond $p<0.05$. Am Stat 73(sup1):1–19
3. McShane BB, Gal D, Gelman A, Robert C, Tackett JL (2019) Abandon statistical significance. Am Stat 73(sup1):235–245
4. Efron B (1998) R. A. Fisher in the 21st century. Stat Sci 13(2):95–114
5. Fisher RA (1973) Statistical methods and scientific inference, 3rd edn. Hafner, New York. Reproduced in Statistical methods, experimental design and scientific inference (1995). Oxford University Press, New York
6. Fisher RA (1973) Statistical methods for research workers, 14th edn. Hafner Publishing, New York. Reproduced in Statistical methods, experimental design and scientific inference (1995). Oxford University Press, New York
7. Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8):e124
8. SAS Institute Inc (2013) Base SAS® 9.4 procedures guide: statistical procedures, 2nd edn. SAS Institute Inc, Cary, NC
9. SAS Institute Inc (2017) SAS® university edition: installation guide for windows. SAS Institute Inc, Cary, NC
10. Gosset WS ("Student," 1908) The probable error of a mean. Biometrika 6(1):1–25
11. Cohen J (1968) Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, Mahwah, NJ
12. Westfall PH, Tobias RD, Wolfinger RD (2011) Multiple comparisons and multiple tests using SAS®, 2nd edn. SAS Institute Inc., Cary
13. Hand DJ (2014) The improbability principle. Scientific American/Farrar, Straus and Giroux, New York
14. Meeks SL, D'Agostino RB (1983) A note on the use of confidence limits following rejection of a null hypothesis. Am Stat 37(2):134–136
15. Castelloe J, Watts D (2015) Equivalence and noninferiority testing using SAS/STAT® software. SAS Institute Inc. SAS users group paper: SAS1911-2015
16. 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. Behav Res Methods 39:175–191
17. Gelman A, Carlin J (2014) Beyond power calculations assessing type s (sign) and type m (magnitude) errors. Perspect Psychol Sci 9(6):641–651
18. Fisher RA (1971) Design of experiments, 8th edn. Hafner Publishing, New York. Reproduced in Statistical methods, experimental design and scientific inference (1995). Oxford University Press, New York
19. 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 2019(381):285–286

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.