SMGroup

SM Journal of Biometrics & Biostatistics

Research Article

Practical Issues in Sample Size Determination for Correlation Coefficient Inference

Stephen W Looney*

Department of Population Health Sciences, Augusta University, USA

Article Information

Received date: Mar 12, 2018
Accepted date: Mar 19, 2018
Published date: Mar 22, 2018

***Corresponding author**

Stephen W Looney, Department of Population Health Sciences, Augusta University, USA, Tel: (706) 721 4846; Email: slooney@augusta.edu

Keywords Confidence Interval; Fisher Z-Transform; Hypothesis Test; Interval Width; Kendall Coefficient; Pearson Correlation; Spearman Correlation; Statistical Power

Abstract

Determination of the appropriate sample size to use when performing inference for a single Pearson correlation coefficient ρ is usually based on achieving sufficient power for the test of H_0: $\rho = 0$. However, sample sizes found using this method can yield confidence intervals that are so wide that they provide very little useful information about the magnitude of the population correlation. Alternative approaches for determining the appropriate sample size are proposed and compared to the "usual" method.

Introduction

An important issue in planning any study that will require inference for a correlation coefficient is the determination of the appropriate sample size to use. The usual procedure of choosing n based on the power of the test of the hypothesis that the population correlation is zero often results in correlations of little practical importance being declared "significant," and confidence intervals that are too wide to be of any practical use. In this article, alternative methods for determining sample size are presented and compared to the "usual" procedure.

Example

Suppose that one is planning a study involving bivariate normal data in which statistical inference is to be performed for a Pearson Correlation Coefficient (PCC) and that the consensus of previous research in the area is that the population correlation is no smaller than 0.40. Reference to sample size tables for the "usual" t-test of the correlation coefficient [1] indicates that a sample of n= 46 will yield 80% power for detecting departures from zero as small as ρ = 0.40 when α = 0.05 (Table 1).

For the sake of argument, suppose that the value of the sample PCC (denoted here after by r) from a subsequent sample of 46 is exactly equal to 0.40. This yields a 2-tailed p-value of 0.006 and a 95% confidence interval of (0.12, 0.62). Although these results indicate statistical significance, their practical significance is unclear because the confidence interval is too wide to draw any reasonable conclusion about the true magnitude of ρ. For example, Hebel and McCarter [2] classify $0.0 \leq |\rho| \leq 0.2$ as "negligible," $0.2 < |\rho| < 0.5$ as "weak," $0.5 \leq |\rho| \leq 0.8$ as "moderate," and $0.8 < |\rho| \leq 1.0$ as "strong." Thus, using their classification scheme, all we can conclude from a confidence interval of (0.12, 0.62) is that ρ is somewhere between "negligible" and "moderate" (inclusive). If one prefers to interpret correlation coefficients in terms of effect size, Cohen [1] suggests that one classify $|\rho| = 0.1$ as a "small" effect size, $|\rho| = 0.3$ as "medium," and $|\rho| = 0.5$ as "large." Using this scheme, all that a confidence interval of (0.12, 0.62) tells us is that the effect size of $|\rho|$ is somewhere between "small" and "large" (inclusive).

One of the alternative approaches proposed in this article is to select n on the basis of the desired width of the resulting Confidence Interval (C.I.) for ρ rather than the power of the test of H_0: ρ= 0. For the aforementioned example, Table 2 indicates that a sample size of n = 273 is required to yield a 95% C.I. of width 0.20 using a "planning value" of r = 0.40. Assuming that a value of exactly r = 0.40 is obtained from a subsequent sample of 273, the resulting p-value is <0.001 and the 95% C.I. is (0.30, 0.50). While this result also indicates statistical significance, the C.I. is sufficiently narrow to indicate that the population correlation between the two variables is "weak" according to the classification scheme of Hebel and McCarter.

Background

The difficulty described in the previous section arises primarily from the fact that H_0: ρ= 0 is not the appropriate null hypothesis to test in most situations that require inference for a single

Table 1: Sample size required to achieve specified power (1 -β) for detecting $\rho = \rho 1$ when testing H_0: ρ=0 using α = 0.05 (2-tailed test).

ρ1	Sample Size		95% CI if r=ρ1		2-tailed ρ-value if r=ρ1	
	β=0.10	β=0.20	β=0.10	β=0.20	β=0.10	β=0.20
0.1	1047	783	(0.04,0.16)	(0.03,0.17)	0.001	0.005
0.2	259	194	(0.08,0.31)	(0.06,0.33)	0.001	0.005
0.3	113	85	(0.12,0.46)	(0.09,0.48)	0.001	0.005
0.4	62	46	(0.17,0.59)	(0.12,0.62)	0.001	0.006
0.5	37	28	(0.21,0.71)	(0.16,0.74)	0.002	0.007
0.6	24	18	(0.26,0.81)	(0.18,0.83)	0.002	0.009
0.7	16	12	(0.31,0.89)	(0.21,0.91)	0.003	0.011
0.8	11	9	(0.38,0.95)	(0.29,0.96)	0.003	0.01
0.9	7	6	(0.46,0.99)	(0.33,0.99)	0.006	0.015

The sample sizes in this table were obtained from Cohen [1].

correlation coefficient. It is usually of little interest to determine if there is sufficient evidence to conclude that $\rho = 0$. (An exception would be a study in which the primary null hypothesis is that variables X and Y are independent and X and Y can be assumed to have a bivariate normal distribution.) Most investigators would not proceed with a study unless they had sufficient reason to believe that the population correlation is non-zero, even though no formal statistical hypothesis test had ever been performed. Other authors agree with our assertion: Strike [4] argues that the test of ρ = 0 is "utterly redundant" and Shoukri [5] asserts that a test of H_0: $\rho = 0$ is "meaningless."

Another problem with testing H_0: $\rho = 0$ is that the usual t-test often rejects the null hypothesis for small values of the sample correlation, even when the sample size is small to moderate (Table 3). For example, when $n = 30$, a sample value of $r = 0.361$ ("weak," according to [2]) yields p = 0.05 (but note the extremely wide C.I. of (0.001, 0.638)). For a sample of $n = 100$, a correlation of only 0.197 yields $p = 0.05$. This correlation is "negligible" according to [2] and would be classified as a small effect size by Cohen [1]. Since the width of a C.I. for ρ varies inversely with both the sample size and the magnitude of r, the combination of small r and small to moderate n often yields C.I.'s that are too wide to be of any practical use, even though p < 0.05.

Table 2: Minimum Sample Size Required to Yield a 95% C.I.(ρ) of Specified Width*.

r	Width of 95% CI(r)			
	0.1	0.2	0.3	0.4
0.1	1507	378†	168†	95†
0.2	1417	355	159	90†
0.3	1274	320	143	81
0.4	1086	273	123	70
0.5	867	219	99	57
0.6	633	161	74	43
0.7	404	105	49	30
0.8	205	56	28	18
0.9	63	21	14	11

*The sample sizes in this table were obtained using the methods of Bonett and Wright [3].

†Confidence intervals based on these values of n and r are guaranteed to contain $\rho = 0$, resulting in a failure to reject H_0: $\rho = 0$.

A further concern is that the sample sizes required to yield 80% or 90% power for testing H_0: ρ = 0 are generally too small to yield C.I.'s of a usable width, even when the sample correlation is large (Table 1). For example, a sample of only $n = 6$ is required to achieve 80% power against the alternative value $\rho_1 = 0.90$. Assuming that the value of r from a subsequent sample of $n = 6$ is exactly equal to 0.9 yields $p = 0.015$ and a 95% C.I. of (0.33, 0.99) (Table 1), which merely indicates that ρ is somewhere between "weak" and "strong" (inclusive) according to [2]. If Cohen's Classification based on effect size is used, this interval only tells us that the effect size of ρ is at least "medium" in magnitude [1].

Alternative Approaches

One alternative to testing H_0: $\rho = 0$ is to specify another null value. Sometimes one is interested primarily in determining if the sample results are consistent with some relevant non-zero hypothesized value; for example, the smallest value of the correlation that would be considered to be clinically meaningful. Such a value may be determined from examining previously published research in the area, from published guidelines or recommendations, from the clinical judgment and expertise of the research team, etc. Applying the Fisher z-transform to r,

$$z(r) = \frac{1}{2}\log\frac{1+r}{1-r},$$

yields a new random variable that has an approximate normal distribution with mean 0 and variance 1/(n - 3). This result can be used to derive a test statistic for testing H_0: $\rho = \rho_0$:

$$z_0 = \frac{z(r) - z(\rho_0)}{\sqrt{n-3}}, \quad (1)$$

Table 3: Confidence intervals corresponding to minimum value of r required to yield $p \le 0.05$ when testing H_0: ρ= 0 (2-tailed test).

Sample Size	Minimum r Yielding p ≤ 0.05	95% C.I.(ρ)	Width of C.I./2
10	0.632	(0.004, 0.903)	0.45
20	0.444	(0.002, 0.741)	0.37
30	0.361	(0.001, 0.638)	0.319
40	0.312	(0.001, 0.568)	0.284
50	0.279	(0.001, 0.517)	0.259
60	0.255	(0.001, 0.478)	0.239
70	0.235	(0.000, 0.445)	0.223
80	0.22	(0.000, 0.419)	0.21
90	0.208	(0.001, 0.398)	0.199
100	0.197	(0.001, 0.379)	0.19
120	0.18	(0.001, 0.348)	0.174
150	0.161	(0.001, 0.313)	0.157
180	0.147	(0.001, 0.287)	0.144
200	0.139	(0.000, 0.272)	0.136

where n is the sample size, $z(r)$ is the Fisher z-transform applied to the sample value of the PCC, and z (ρ_0) is the Fisher z-transform applied to the hypothesized value of the PCC, which can be any ρ_0 such that $|\rho_0| < 1$. (Most commonly, $\rho_0 = 0$, in which case $z(\rho_0) = z(0) = 0$.) The value of z_0 in (1) is then evaluated against the standard normal distribution to obtain an approximate p-value.

In some instances, there may be no non-zero null value ρ_0 that is of primary interest. In this case, one could use the cutoffs advocated by Hebel and McCarter [2] (or other cutoffs that make sense in the context of the applied problem) to get a sense of the magnitude of the correlation in the population under study. For example, if it is known that ρ is positive, then one could test H_0: $\rho \leq 0.8$ to determine if the population correlation is "strong" or H_0: $\rho \leq 0.2$ to determine if the population correlation is "non-negligible" using the Hebel and McCarter criteria. Tables similar to Table 1 could then be constructed for these values of ρ_0. Alternatively, using the same notation as in Table 1, the following formula could be used for a 1-tailed test:

$$n = 3 + \left[\frac{z_\alpha + z_\beta}{z(\rho_1) - z(\rho_0)}\right]^2, \quad (2)$$

where z_r denotes the upper γ-percentage point of the standard normal and $z(\rho)$ denotes the Fisher z-transform of ρ.

Consider the example discussed previously in which one wishes to determine the appropriate sample size to use for a future study in which the PCC is of primary interest and there is reason to believe that ρ is positive and no smaller than 0.40. One approach would be to test H_0: $\rho \leq 0.2$; if this hypothesis is rejected, then one can conclude that the population correlation is non-negligible according to [2]. A calculation using Equation (2) indicates that samples of 130 and 179 would be required to achieve 80% and 90% power, respectively, to detect $\rho_1 = 0.40$ using an upper-tailed test.

Suppose that a subsequent sample of $n = 130$ yielded a sample value of exactly $r = 0.40$; the corresponding upper-tailed p-value for the test of H_0: $\rho \leq 0.2$ is 0.006 and the one-sided 95% C.I. is (0.27, 1.00). Thus, one can conclude that the population correlation is significantly greater than 0.2 ($p < 0.001$) and can be classified as "non-negligible" according to [2]. In the example in which $n = 30$ and $r = 0.361$, the p-value for the test of H_0: $\rho < 0.2$ is 0.181, insufficient evidence to conclude that the population correlation is "non-negligible," despite the fact that the test of H_0: $\rho = 0$ indicated that the result is "significant."

Another alternative is to focus one's attention on confidence interval estimation of the population correlation (derived using the Fisher z-transform of r) instead of the test of a particular hypothesized null value. This approach is consistent with the emphasis placed on confidence interval estimation over hypothesis testing by many authors [6-8] Table 2 can be used to determine the sample size required to obtain a 95% C.I. for ρ of a desired width. (This approach was illustrated in a previous section).

Discussion

The purpose of this article is to illustrate some of the practical problems encountered when attempting to determine the appropriate sample size to use when a proposed study requires inference for a single correlation coefficient. The argument is made that H_0 : $\rho = 0$ is usually not the appropriate null hypothesis to test and that using sample sizes that yield a desirable level of power (say, 80% or 90%) for this test can result in C.I.'s that are so wide that they provide very little useful information about the magnitude of the population correlation. Two alternative approaches were proposed: (1) testing null values other than $\rho_0 = 0$ and (2) determining the sample size so as to achieve a certain level of precision of the estimate of ρ, as measured by the width of the resulting C.I. Depending on the purpose of the statistical analysis, either or both of these approaches could be a useful alternative to the "usual" method of determining sample size. However, it must be noted that the sample sizes required for either of these approaches often will be much larger than those required to achieve acceptable power when testing H_0: $\rho = 0$. In the example considered in a previous section, the "usual" approach based on testing H_0: $\rho = 0$ yielded a sample of size $n = 46$, the approach based on testing H_0: $\rho \leq 0.2$ yielded $n = 163$ and the "C.I." approach yielded $n = 273$.

The alternative approaches described in this article could also be applied if one were performing inference for the Spearman Correlation Coefficient (SCC) or the Kendall Coefficient of Concordance (KCC). For example, using the same notation as in Equation (2) for a one-tailed test of the KCC, the "z-transform" developed by Fieller, Hartley, and Pearson [9] for the KCC yields the sample size formula.

$$n = 4 + 0.437\left[\frac{z_\alpha + z_\beta}{z(\tau_{b1}) - z(\tau_{b0})}\right]^2$$

Where τb_0 and τb_1 are the null and alternative values of the KCC, respectively ($\tau b1 > \tau b0$). For the SCC, the Fieller, Hartley and Pearson z-transform yields.

$$n = 3 + 1.06\left[\frac{z_\alpha + z_\beta}{z(\rho_{s1}) - z(\rho_{s0})}\right]^2, \quad (3)$$

where ρs_0 and ρs_1 are the null and alternative hypothesized values of the SCC, respectively ($\rho s_1 > \rho s_0$). Using the improved z-transform of the SCC proposed by Bonett and Wright [3], the formula in (3) becomes

$$n = 3 + \left[1 + \frac{\rho_{s1}^2}{2}\right]\left[\frac{z_\alpha + z_\beta}{z(\rho_{s1}) - z(\rho_{s0})}\right]^2. \quad (4)$$

Bonett and Wright recommend that (4) be used for $|\rho_{s1}| < 0.95$ and that (3) be used if $|\rho_{s1}| \geq 0.95$.

We are not necessarily advocating the use of the Hebel and McCarter criteria [2] for interpreting the magnitude of correlation coefficients. While we have found these to be useful in our own exploratory analyses of biomedical data, they may not be appropriate in other areas of investigation. However, we encourage the development and use of such guidelines because we feel that they can greatly enhance one's ability to interpret and communicate results to non-statisticians (e.g., see the guidelines proposed by Landis and Koch [10] for interpreting agreement coefficients. And the guidelines proposed by Fleiss et al. [11] for interpreting intra-class correlation coefficients).

Simply reporting that a correlation is "significant" just because the *p*-value for the test of $H_0 : \rho = 0$ is less than 0.05 is generally not sufficient.

Acknowledgement

Some of this work was completed while the first author was Acting Senior Research Fellow in Epidemiology, Department of Medicines Management and Visiting Professor, Department of Mathematics, Keele University, Staffordshire, England. This research was partially supported by Wellcome Research Travel Grant #0816.

References

1. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. 1988; 79-80,101.
2. Hebel JR, McCarter RJ. A Study Guide to Epidemiology and Biostatistics. Burlington, MA: Jones and Bartlett Learning. 2012; 90.
3. Bonett DG, Wright TA. Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. Psychometrika. 2000; 65: 23-28.
4. Strike PW. Assay method comparison studies, in Measurement in Laboratory Medicine: A Primer on Control and Interpretation, Oxford, UK: Butterworth-Heinemann. 1996; 170.
5. Shoukri MM. Measures of Interobserver Agreement and Reliability. Boca Raton, FL: CRC Press 2011; 92.
6. Gardner MJ, Altman DG. Confidence intervals rather than P values: Estimation rather than hypothesis testing. British Medical Journal 1986; 292: 746-750.
7. Hahn GJ, Meeker WQ. Statistical Intervals. New York: John Wiley & Sons. 1991; 39-40.
8. Schmidt F. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods. 1996; 1: 115-119.
9. Fieller EC, Hartley HO, Pearson ES. Tests for rank correlation coefficients. I. Biometrika 1957; 44: 470-481.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1975; 33: 159-174.
11. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions, Hoboken, NJ: John Wiley & Sons. 2003; 604.