Simplified Tools for Sample Size Determination for Correlation Coefficient Inference

Justine May[1], Jessica Ketchum[2], and Stephen Looney[1]
[1,3]Dept. of Biostatistics and Epidemiology, Augusta University, Augusta, GA 30912
[2]Craig Hospital, 3425 S. Clarkson St., Englewood, CO 80113

Abstract

Bivariate correlation analysis is one of the most commonly used statistical methods. Unfortunately, it is generally the case that little or no attention is given to sample size determination when planning a study in which correlation analysis will be used. For example, our review of clinical research journals indicated that none of the 111 articles published in 2014 that presented correlation results provided a justification for the sample size used in the correlation analysis. In this presentation, we discuss the issues associated with sample size determination for bivariate correlation analysis and provide simplified tools, including nomograms, for determining the required sample size. These tools make use of recent improvements in methods for sample size calculations for correlation analysis. Tools are provided that can be used for sample size determination for Pearson, Spearman, and Kendall coefficients.

Keywords: Effect size; Confidence interval; Pearson correlation; Spearman correlation; Kendall coefficient; Power

1. Introduction

Bivariate correlation analysis is one of the most commonly used statistical methods. Unfortunately, little or no attention is given to sample size determination when planning a study in which correlation will be the primary analysis. Our review of clinical research journals indicated that none of the 111 articles published in 2014 that used correlation as the primary analysis provided a sample size justification or power calculation.

2. Available Tools

There are many tools available for investigators who wish to perform a sample size or power calculation for correlation coefficient inference. There include tables (Cohen 1988, Looney 1996, Looney and Hagan 2015), formulas (Bonett and Wright 2000, Looney and Hagan (2015), software code (Looney and Hagan 2015), software packages (Table 1), and internet-based Java applets (Table 2). However, these resources vary widely in terms of capabilities and sometimes they cannot be relied upon to provide accurate results. For example, none of the widely used packages listed in Table 1 have the capability of performing sample size calculations for either a Spearman correlation coefficient (SCC) or a Kendall coefficient of concordance (KCC). None of the applets in Table 2 are capable of performing a sample size calculation based on the width of the confidence interval. One of the applets in Table 2 sometimes gives incorrect results, as illustrated in Section 6.1.

3. Issues in Sample Size Determination for Correlation Coefficients

Given the wide availability of tools for performing sample size and power calculation for correlation coefficients, one must ask: "why are sample size and power calculations not done?" There are several possible explanations, including (1) status quo, (2)

lack of availability of easy-to-use tools that provide correct results, and (3) correlation analysis is often considered to be an exploratory analysis, so sample size determination (or justification) may seem unnecessary.

However, these attitudes are hardly justifiable, especially among statistical professionals, as any practicing statistician is well aware of the importance of sample size determination in general. If the sample size is too small, it will be impossible for the statistical test of the correlation coefficient to detect a scientifically meaningful association, *even if one is present*. Furthermore, negative (i.e., non-significant) statistical results can give the mistaken impression that the two variables are not associated with each other if the sample size is too small. On the other hand, if the sample size is too large, *valuable resources will be wasted* since a scientifically meaningful association could have been detected with fewer subjects. In addition, associations that are not scientifically meaningful may be statistically significant if the sample size is too large.

4. Our Contribution

In this presentation, we provide the following new developments: (1) improved sample size formulas for the Spearman coefficient, (2) sample size formulas for the Kendall coefficient, and (3) sample size nomograms for the Spearman coefficient.

5. Example

We will illustrate many of our ideas using information from the study entitled "Development of Lithogenic Bile During Puberty in Pima Indians" (Bennion et al. 1979). The primary goal of this study was to examine the association between age and bile cholesterol saturation (BCS) in young Pima Indians. Bennion et al. found a Pearson correlation of $r = 0.40$ in their sample of 36 males. Suppose that we want to plan a study to examine the association between age and BCS in young Cherokee Indians.

6. Methods for Sample Size Determination for Correlation Coefficients

6.1 Usual Approach for Pearson Correlation

The most commonly used method for calculating the required sample size for inference based on a Pearson correlation coefficient (PCC) is to use pilot data, previously published research, published guidelines or recommendations, clinical judgment or expertise of the research team, or other relevant information to identify a "planning value" for the correlation coefficient of interest. For example, we could use the value of 0.40 obtained by Bennion et al. (1979) as our planning value.

Once an appropriate planning value has been identified, one simply uses Table 3.4.1 in Cohen (1988, p. 102) to determine the required sample size. This table is based on the "usual" test of the hypotheses

$$H_o: \rho = 0 \text{ vs. } H_a: \rho \neq 0, \tag{1}$$

that uses the test statistic

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{2}$$

which follows a t distribution with n - 2 degrees of freedom when H_0 in (1) is true. Based on Cohen's table, $n = 46$ yields 80% power for detecting departures from zero as small as $|\rho| = 0.40$ when $\alpha = 0.05$. With one exception, applying any of the tools mentioned in Tables 1 and 2 yields either $n = 46$ or $n = 47$. AI-Therapy Statistics https://www.ai-

therapy.com/psychology-statistics/sample-size-calculator yields a sample size of $n = 40$, not $n = 46$.

While the "usual" approach described above is straightforward to apply, with easily accessible tools, there are problems with basing the sample size calculation for the PCC on the hypotheses in (1). For one thing, H_o: $\rho = 0$ is not the appropriate null hypothesis to test in most situations. It is usually of little interest to determine if $\rho \neq 0$. (The lone exception would be that the primary null hypothesis is that X and Y are independent, and the assumption can be made that X and Y have a bivariate normal distribution.) Other authors agree with our assertion: Strike (1996, p. 170) argues that the test of $\rho = 0$ is "utterly redundant" and Shoukri (2011, p. 92) asserts that a test of H_o: $\rho = 0$ is "meaningless." One may then reasonably ask why the most commonly used approach for sample size determination for a correlation coefficient is based on the test of H_o: $\rho = 0$.

Another problem with basing the sample size calculation on the "usual" test of (1) is that sample sizes required to yield 80% power for the test of H_o: $\rho = 0$ using a significance level of 0.05, for example, are generally too small to yield a 95% confidence interval (C.I.) of usable width, even when r is rather large. For example, using any of the tools in Table 1 or 2, we find that $n = 6$ is sufficient to achieve 80% power against the alternative value $\rho_1 = 0.90$ when testing the hypotheses in (1). Now, for the sake of argument, suppose the value of the PCC in the sample of $n = 6$ turns out to be exactly $r = 0.9$. We obtain a statistically significant result, as expected, with a 2-tailed p-value of 0.015. However, the 95% C.I. based on these data is (0.33, 0.99). Thus, we have statistical significance, but a C.I. that is too wide to provide any useful information about the true magnitude of $|\rho|$.

A third problem with basing the sample size calculation for the PCC on the test of (1) is that the test statistic in (2) often rejects H_o for small values of $|r|$, even when n is relatively small. For example, suppose a sample of $n = 30$ yields $r = 0.361$. Then the 2-tailed p-value = 0.0495, with a 95% C.I. of (0.001, 0.638). Again, we have statistical significance, but a C.I. that is too wide to provide any useful information about the true magnitude of $|\rho|$.

6.2 Alternative Approach 1

An alternative approach for determining the sample size required for inference based on the PCC is to choose n based on the desired width of the resulting C.I., not the power of the test of H_o: $\rho = 0$. In general, a confidence interval for the true value of a Pearson coefficient can be derived using Fisher's z-transform of the sample coefficient r:

$$z = \tanh^{-1} r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right], \qquad (3)$$

which is asymptotically distributed as $N\left(\tanh^{-1}\rho,\ \sigma^2\right)$, where ρ denotes the true value of the PCC and σ^2 denotes the asymptotic variance of z. The same transformation can be applied to the sample value of a Spearman or Kendall coefficient, yielding an approximately normally distributed transformed coefficient. For the PCC, $\sigma^2 = 1/(n-3)$ (Fisher 1925); for the SCC, $\sigma^2 = (1+\rho_s^2/2)/(n-3)$, where ρ_s denotes the true value of the Spearman coefficient (Bonett and Wright 2000); and, for the KCC, $\sigma^2 = 0.437/(n-4)$ (Fieller et al. 1957).

Basing sample size determination on obtaining a confidence interval of desirable width is easily justified, given that the use of C.I.'s in correlation analysis (or any other type of statistical inference) should by now be standard practice, given that leading

statistical practitioners have emphasized the use of C.I.'s as an alternative to *p*-values for the last 30 years (e.g., Gardner and Altman 1986).

Bonett and Wright (2000) proposed a two-stage process for determining *n* based on the desired width of the C.I. for the true value of the coefficient. This method can be applied to Pearson, Spearman, or Kendall coefficients. The true value of the coefficient to be estimated will be denoted by ξ in what follows.

Stage 1: Compute $n_0 = 4c^2\left(1-\widehat{\xi}^2\right)\left(z_{\alpha/2}/w\right)^2 + b,$

Stage 2: Compute $n = (n_0 - b)\left(w_0/w\right)^2 + b$ if the desired confidence interval width has not been attained using n_0. In these formulas,

$\hat{\xi}$ = a planning value for the coefficient,
w = desired width of the C.I. for the true value ξ,
$z_{\alpha/2}$ = z-score corresponding to a two-sided 100(1-α)% C.I.,
b and c^2 are taken from Table 3,
n_o is the sample size estimate from Stage 1,
w_o is the width of the Stage 1 interval based on n_0.

A nomogram for finding the required sample size for estimating a Spearman coefficient using a 95% C.I. that is based on this two-stage procedure is provided in Figure 1. To use this nomogram, first locate the planning value along the *x*-axis. (In the example, this is 0.4.) Then, draw a vertical line that intersects with the curve corresponding to the desired width of the confidence interval in the study being planned. (In the example, this is 0.2.) Finally, draw a horizontal line from the curve to the *y*-axis. The point of intersection is the desired sample size. This method is illustrated in Figure 1 for the example described below.

Returning to the Example, suppose that we decide to use the SCC as the measure of association in the study we are planning because of the apparent non-normality of the BCS data in the study by Bennion et al. Let r_s denote the sample SCC and let ρ_s denote the true value. Using the nomogram in Figure 1, we find that a sample size of n = 300 would yield a 95% C.I. for ρ_s of width 0.20 based on a planning value of 0.4. Suppose for the sake of argument that the resulting sample of size 300 yields exactly r_s = 0.4. Then we obtain a 2-tailed p-value < 0.0001 and a 95% C.I. of (0.29, 0.50). The results certainly indicate statistically significance, but the C.I. is narrow enough to indicate that ρ_s is "weak" according to the cutoffs proposed by Morton, Hebel, and McCarter (1996), who recommended that a correlation between 0.2 and 0.5 be classified as weak.

6.3 Alternative Approach 2

Another alternative to basing the sample size calculation on the test of (1) is to specify another null value in the null hypothesis. We can use the test statistic based on the Fisher z-transform in (3) to test

$$H_o: \xi = \xi_0, \qquad (4)$$

where ξ_0 is the null value of interest of the desired measure of association (Pearson, Spearman, or Kendall). For a 1-tailed test of the null hypothesis in (4), Fisher's *z* yields the following sample size formula:

$$n = b + c^2\left[\frac{(z_\alpha + z_\beta)}{z(\xi_1) - z(\xi_0)}\right]^2, \qquad (5)$$

where z_γ =upper γ-percentage point of N(0,1),
$z(\xi)$ denotes the Fisher z-transform of ξ, and

b and c^2 are taken from Table 3.

For a 2-tailed test of (4), simply replace z_α in (5) by $z_{\alpha/2}$.

A nomogram for finding the required sample size for a lower-tail test of a Spearman coefficient using significance level 0.05 that will yield 80% power is provided in Figure 2. To use this nomogram, first locate the null value of the SCC along the *x*-axis. (In the example, this is 0.7). Then, draw a vertical line that intersects with the curve corresponding to the alternative value in the lower-tailed hypothesis test to be used in the study being planned. (In the example, this is 0.5.) Finally, draw a horizontal line from the curve to the *y*-axis. The point of intersection is the desired sample size. This method is illustrated in Figure 2 for the example described below.

Sometimes the primary purpose of a new study is to determine if the correlation is significantly different from some scientifically or clinically meaningful value. Suppose that we wish to use multiple regression based on ranks (Iman and Conover 1979) to model BCS as a function of various characteristics of the Cherokee Indians we are planning to study (e.g., age, gender, diabetic status). As part of our variable screening, we wish to examine the percent variability in the ranked bile cholesterol saturation levels that is explained by ranked age, and we assume that the scientifically meaningful cutoff is $R_s^2 = 0.50$, corresponding to $\rho_s = 0.7$. We will test H_o: $\rho_s \geq 0.7$ using the sample data from the study we are planning. If the data indicate that H_o should be rejected (and hence we conclude that $\rho_s < 0.7$), then we will eliminate age from further consideration in our rank-based multiple regression model for BCS. Using the nomogram in Figure 2, we see that we could detect alternative values of ρ_s as large as 0.5 ($R_s^2 = 0.25$) in our 1-tailed test with a sample of $n = 70$.

6. Discussion

As we pointed out in Section 3, H_o: $\rho = 0$ is usually *not* the appropriate null hypothesis to test, and using *n* obtained from sample size tables tailored to the usual *t*-test of this hypothesis can yield samples that provide very little useful information about the magnitude of ρ. As alternatives to basing the sample size determination on the test of H_o: $\rho = 0$, we propose that one consider either (1) testing a null value other than $\rho_0 = 0$ and choosing a value of *n* to achieve acceptable power for this test, or (2) choosing *n* so as to achieve a desired level of precision of the estimate of the correlation coefficient, as measured by the width of a confidence interval. However, we acknowledge that using either of these alternative methods can yield sample sizes that are much larger than those required to achieve good power for the usual test of H_o: $\rho = 0$. Depending on the accessible population and the available resources for collecting the data from these larger samples, these sample sizes may not be feasible in the context of the applied research problem. In this case, the investigators should consider re-formulating their research question(s), or proposing another statistical analysis that will require fewer subjects. The results of a correlation analysis should be interpreted in light of the *p*-value calculated for the appropriate hypothesis test of the true value of the correlation parameter, as well as the confidence interval for the true value. If the confidence interval resulting from the sample size used in the study is too wide to be of practical use, then one may legitimately question the validity and generalizability of the study results.

References

1. Bennion et al. (1979), "Development of Lithogenic Bile During Puberty in Pima Indians," *New England Journal of Medicine,* 300, 873-876.

2. Bonett D.G., and Wright T.A. (2000), “Sample Size Requirements for Estimating Pearson, Kendall, and Spearman Correlations,” *Psychometrika*, 65, 23-28.
3. Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates.
4. Gardner M.J., and Altman, D.G. (1986), "Confidence Intervals Rather than P Values: Estimation Rather than Hypothesis Testing," *British Medical Journal*, 292, 746-750.
5. Iman, R.L., and Conover, W.J. (1979), "The Use of the Rank Transform in Regression," *Technometrics*, 21, 499-509.
6. Looney, S.W. (1996), “Sample Size Determination for Correlation Coefficient Inference: Practical Problems and Practical Solutions,” *Proceedings of the ASA Statistical Computing Section, 1996 Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 240-245.
7. Looney, S.W., and Hagan, J.L. (2015), *Analysis of Biomarker Data: A Practical Guide*, New York: John Wiley & Sons.
8. Morton, R.F., Hebel, J.R., and McCarter, R.J. (1996), *A Study Guide to Epidemiology and Biostatistics*, Gaithersburg, MD: Aspen Publishers, p. 96.
9. Shoukri, M.M. (2011), *Measures of Interobserver Agreement and Reliability* (2nd ed.), Boca Raton, FL: CRC Press, p. 92.
10. Strike, P.W. (1996), “Assay Method Comparison Studies,” in *Measurement in Laboratory Medicine: A Primer on Control and Interpretation*, Oxford, UK: Butterworth-Heinemann, p. 170)

Table 1. Summary of Software Capabilities for Sample Size and Power Calculations for Correlation Coefficients

Method	PASS	nQuery	SAS Proc POWER	R package pwr
Test of H_o: $\rho = 0$	√	√	√	√
Test of H_o: $\rho = \rho_1$	√	√	√	---
C.I. Width	√	√	---	---
Spearman	---	---	---	---
Kendall	---	---	---	---

Table 2. Summary of Applet Capabilities for Sample Size and Power Calculations for Correlation Coefficients

Method	UCSF[1]	Stat Decision Tree[2]	SISA[3]	AI-Therapy Statistics[4]
Test of H_o: $\rho = 0$	√	√	√	√
Test of H_o: $\rho = \rho_1$	---	---	√	---
C.I. Width	---	---	---	---
Spearman	---	---	---	---
Kendall	---	---	---	---

[1] http://www.sample-size.net/correlation-sample-size/
[2] https://www.anzmtg.org/stats/PowerCalculator/PowerCorrelation
[3] http://www.quantitativeskills.com/sisa/statistics/correl.htm
[4] https://www.ai-therapy.com/psychology-statistics/sample-size-calculator

(All were accessed on July 19, 2016.)

Table 3. Constants Needed to Apply Fisher z-Transform to Measures of Association

Measure of Association	b	c^2	Source
Pearson	3	1	Fisher (1925)
Spearman	3	$1+\left(\hat{\rho}_s^2/2\right)$	Bonett and Wright (2000)
Kendall	4	0.437	Fieller et al. (1957)

Note: For purposes of sample size determination based on a confidence interval, $\widehat{\rho}_s$ is a "planning value." For purposes of sample size determination based on an hypothesis test, $\widehat{\rho}_s$ is the null value.

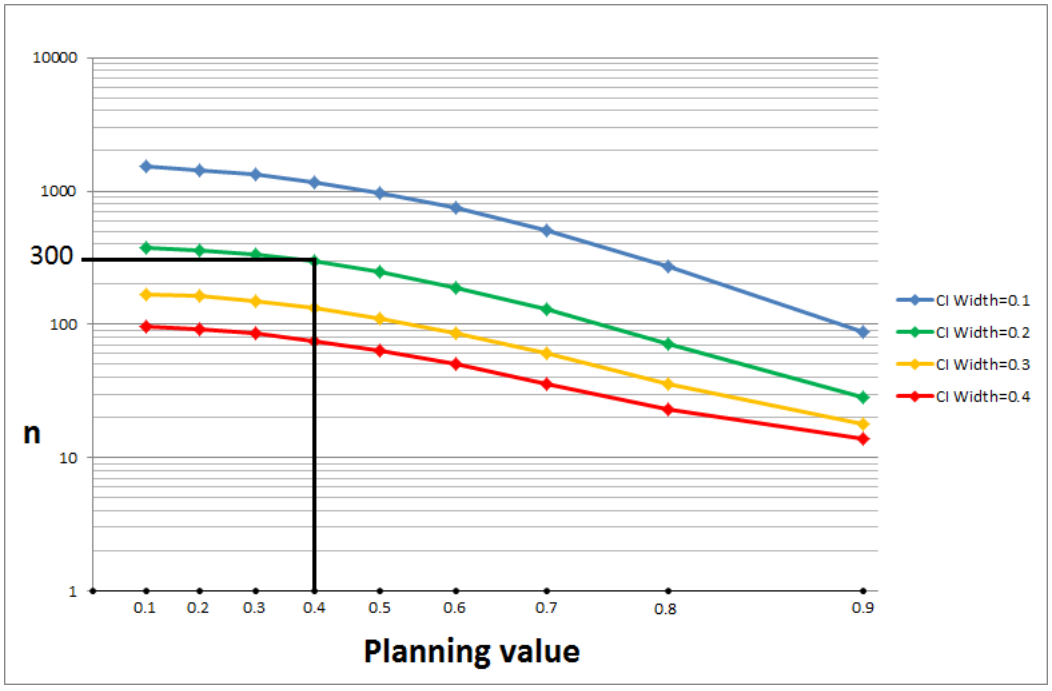

Figure 1. Sample Size Nomogram for 95% Confidence Interval Estimation of Spearman's Coefficient

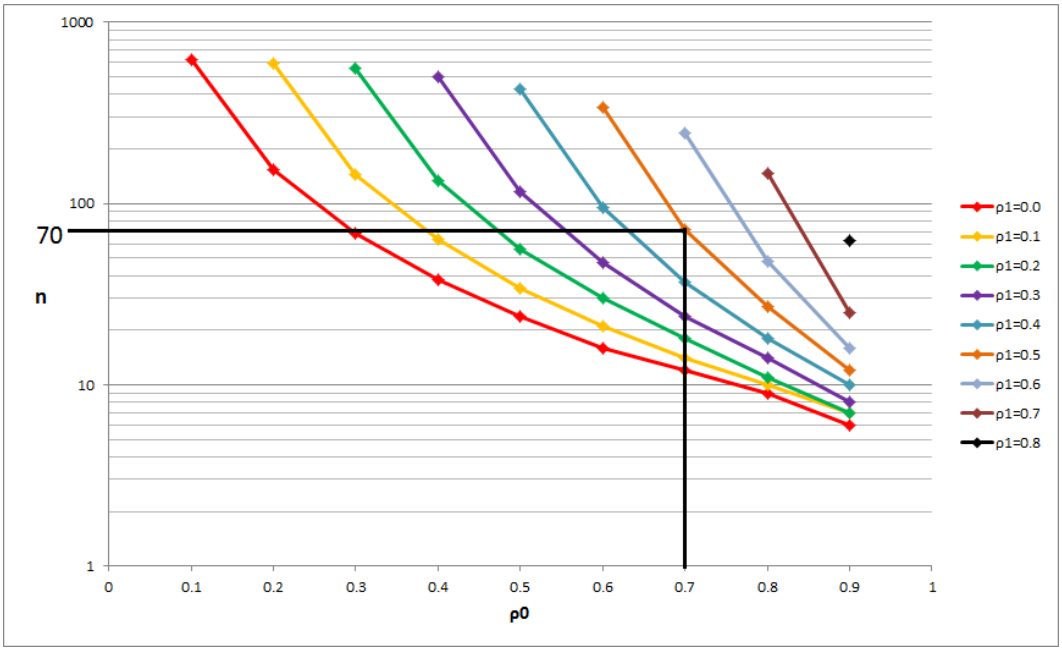

Figure 2. Sample Size Nomogram for Lower-tailed Test of Spearman's Coefficient, 80% Power, Significance Level 0.05