2014 ■ vol. 10 ■ no. 2

DOI: 10.20982/tqmp.10.2.p124

Sample size estimation for correlations with pre-specified confidence interval

Murray Moinester [a], Ruth Gottfried [b]

[a] School of Physics and Astronomy, Tel Aviv University, 69978, Tel-Aviv, Israel

[b] Faculty of Social Welfare & Health Sciences, University of Haifa, Mount Carmel, 3478601, Haifa, Israel

Abstract ■ A common measure of association between two variables x and y is the bivariate Pearson correlation coefficient ρ(x,y) that characterizes the strength and direction of any linear relationship between x and y. This article describes how to determine the optimal sample size for bivariate correlations, reviews available methods, and discusses their different ranges of applicability. A convenient equation is derived to help plan sample size for correlations by confidence interval analysis. In addition, a useful table for planning correlation studies is provided that gives sample sizes needed to achieve 95% confidence intervals (CI) for correlation values ranging from 0.05 to 0.95 and for CI widths ranging from 0.1 to 0.9. Sample size requirements are considered for planning correlation studies.

Keywords ■ sample size estimation, correlation, confidence interval

murraym@tauphy.tau.ac.il

Introduction

This article describes how to determine by confidence interval analysis the optimum sample size for studies that measure the strength of bivariate correlations between characteristics (variables) x and y. In cross-sectional correlational research for example, the x variable may measure exposure to some experience while the y variable may measure some subsequent behaviour or outcome. The Pearson correlation coefficient ρ(x,y) describes the strength and direction of an assumed linear relationship between x and y (Corty, 2007, Field, 2009). For a given correlation value, sample size determines the width of the confidence interval (CI), and conversely the width determines the sample size. Estimating sample size before conducting a study, or at the early stage of a study, is scientifically important in order to maximize the probability to detect any existing significant correlations (Beaulieu-Prévost, 2006, Corty, 2007, Field, 2009, Kelley, 2008). This article reviews existing methods of sample size estimation for measuring the strength of a correlation, and discusses their different ranges of applicability. A convenient equation is derived and presented to plan sample size to achieve a desired (narrow) CI width for correlations. In addition, a useful table for planning correlation studies is provided that gives sample sizes needed to achieve a 95% confidence interval (CI) for correlation values ranging from 0.05 to 0.95 and for CI widths ranging from 0.1 to 0.9. Sample size requirements are considered for planning correlation studies.

Alternative sample size estimations based on statistical power analyses have been described by Descoteaux (2007), Lachin (1981) and Lenth (2001). A power analysis allows defining for example a 95% power (probability) of rejecting a null hypothesis H_0 of no correlation in the sample and accepting an alternative hypothesis H_1 that a correlation exists. However, Beaulieu-Prévost (2006) and Cumming (2014) pointed out serious problems with the null hypothesis power analysis, and recommended instead that estimations should be based on effect sizes and confidence intervals. This article follows their recommendation.

Statistical Concepts and their Connection to Sample Size

The Pearson correlation coefficient is a numerical index that measures the strength and the direction of a linear relationship between two variables, x and y. If one variable increases (or decreases) as the other increases (or decreases), then the coefficient is positive (or negative). The strength of a relationship is indicated by the numeric value of the coefficient, which can take a range of values from +1 to -1. Formal requirements are: the selection of x,y pairs is random and independent, the joint distribution is multivariate normal, the linear regression line is straight for the relationship between variables x,y; and the variables are measured on a numerical interval scale (Corty,

2007, Field, 2009). A correlation coefficient (CC) that characterizes the entire population is denoted by ρ(x,y), while a CC evaluated for a particular sample of size N is denoted by r(x,y). When variables are correlated, knowledge of one allows estimating (predicting) the other. Medium to strong correlations are useful for establishing a predictive relationship between the variables. A CC value of zero means that there is no linear relationship between the two variables.

Consider the sampling distribution of the CCs, the probability distribution of all the CCs obtained from a large number of random data sampled from a large (parent) "population", each sample having size N. The sampling distribution for large N is expected to be approximately normal, with a single central peak at the mean value of ρ and with standard deviation equal to σ_ρ (Corty, 2007, Field, 2009). The σ_ρ value may be estimated using an infinite series given by Hotelling (1953), for which the first two terms are:

$$\sigma_\rho^2 = (1-\rho^2)^2[\frac{1}{N-1} + \frac{11\rho^2}{2(N-1)^2}]. \qquad (1)$$

For planning purposes, since the population correlation ρ is usually not known, a measured sample statistic r may be used to approximate ρ, or it may be estimated from previous research. Hotelling (1953) shows explicitly that the distribution shape is approximately symmetric and normal for N ≥ 55 and |ρ| ≤ 0.7. For these conditions, the first term of Eq. 1 provides better than 5% precision for the evaluation of σ_r:

$$\sigma_r = \frac{(1-r^2)}{\sqrt{N-1}} \qquad (2)$$

The American Psychological Association (APA, 2010) recommends that researchers provide estimates of the strength of a measured characteristic (effect size) of a population by means of a confidence interval. This procedure is usually referred to as accuracy in parameter estimation (AIPE). Beaulieu-Prévost (2006) and Cumming (2014) described this method in detail and emphasized its importance for presenting research results and for estimating sample size. The effect size of interest here is the smallest value of Pearson's ρ that the researcher decides would be scientifically meaningful to measure. Accuracy for a given sample size measures how close a measured r is to the true population size ρ. Although it may improve as sample size increases, it depends strongly on controlling systematic errors that may lead to various forms of bias. By contrast, precision as measured by the standard deviation σ_r improves as the sample size increases, approximately following Eq. 2. In the context of AIPE, "accuracy" is defined as the square root of the mean square error, which includes both precision and bias errors (Kelley, 2008). The "confidence interval analysis" discussed below deals only with sample size precision errors, not with bias errors.

Sample Size Estimation Associated with Confidence Interval Analysis

A two-sided confidence interval (CI) for the Pearson correlation coefficient ρ is an observed range of values that consists of a lower limit (*LO*) and an upper limit (*UP*), within which the true value of ρ is found with a specified probability (Corty, 2007; Field, 2009). The CC and its standard deviation σ_r for N measurements of x,y data pairs may be computed using standard statistics programs. Consider that a CC is determined as CC = r ± σ_r for a certain sample N. The CI provides an estimate of the unknown ρ value, and also indicates the reliability of the estimate. A 95% CI would capture the true value of ρ with 95% level of confidence, within lower (*LO*) and upper (*UP*) limits:

$$LO = r - 1.96\sigma_r, \qquad (3a)$$

$$UP = r + 1.96\sigma_r. \qquad (3b)$$

The total CI width will be here denoted by CI2w (CI2w = *UP* - *LO*), and the CI half-width by w.

The z-score multiplier 1.96 is used to define the 95% CI of a normal distribution (Corty, 2007; Field, 2009). This is so since 95% of the area under the standard normal distribution curve falls within the z-score interval [-1.96, 1.96]; or equivalently because the area under the standard normal curve for z < 1.96 equals 0.975. The z-score measures the deviation from the mean expressed in units of standard deviations. The values $z = 1.645, 1.96, 2.576$ define 90%, 95%, 99% CIs respectively. A 95% CI is associated with an $\alpha = 0.05$ level of significance (0.95 probability), via the relationship *confidence level = 1 - α*.

For example, consider that $\rho = 0.316$, and that the 95% CI measurement gives $LO = 0.204$ and $UP = 0.428$ for a given sample size. If the measurement were

Table 1 ■ Sample Size Requirements for Desired 95% Confidence Interval Half-Widths w for different Pearson Correlation Values | *r* |, based on CIxcorr.

	w								
| *r* |	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.05	1530	383	171	97	62	43	32	25	20
0.10	1507	378	168	95	61	43	32	25	20
0.15	1469	368	164	93	60	42	31	24	19
0.20	1418	355	159	90	58	41	30	23	19
0.25	1352	339	151	86	55	39	29	22	18
0.30	1274	320	143	81	53	37	28	21	17
0.35	1185	298	133	76	49	35	26	20	16
0.40	1086	273	123	70	46	32	24	19	15
0.45	980	247	111	64	42	30	22	18	14
0.50	867	219	99	57	37	27	20	16	13
0.55	751	190	86	50	33	24	18	15	12
0.60	633	161	74	43	29	21	16	13	11
0.65	517	132	61	36	25	18	14	12	_
0.70	404	105	49	30	20	15	12	_	_
0.75	299	79	38	23	17	13	_	_	_
0.80	205	56	28	18	13	_	_	_	_
0.85	125	36	19	13	_	_	_	_	_
0.90	62	20	12	_	_	_	_	_	_
0.95	22	_	_	_	_	_	_	_	_

repeated many times, 95% of the random sample intervals from *LO* to *UP* would "cover" (i.e. include) the population value $\rho = 0.316$. A narrow CI means that ρ is estimated with high precision. The research goal is to choose a sample size N that achieves a sufficiently narrow confidence interval for measuring the smallest CC of potential interest.

Methods for Confidence Interval Analysis

Method 1. Bonett's open source R function CIcorr.R (Bonett, 2014, R Foundation, 2011) and the StatsToDo internet calculator allow calculating CI widths for given values of r, sample size N, and significance α. A slightly modified version of Bonett's program, CIxcorr.R shown in Appendix A, can be used to iteratively find N. The iteration is carried out to find the highest value of N for which the output CI width CI2w is closest to but does not exceed a pre-specified CI width. This iteration method was previously described by Bonett and Wright (2000). The CIxcorr program output is shown in Appendix A for $r = 0.9$, where $N = 62$ is the highest value of N for which the output CI2w is closest to but does not exceed a pre-specified CI width of 0.1. Based on such iterations, Table 1 gives sample sizes needed to achieve 95% CIs for r values ranging from 0.05 to 0.95 and for CI half-widths w ranging from 0.05 to 0.45. Bonett and Wright calculated sample sizes by iteration for CI half-widths $w = 0.05$, 0.10, 0.15, and their results for these w values agree exactly with those of Table 1. Bonett's open source R function sizeCIcorr.R (Bonett, 2014, R Foundation, 2011), shown in Appendix B, may alternatively be used to compute a sample size N required to estimate a chosen Pearson correlation coefficient r with a given significance α and total width CI2w. The sizeCIcorr program and the CIxcorr iteration procedure give identical results (within 1 count) for all the values shown in Table 1.

Method 2. Bonett and Wright (2000) presented a two-stage approximation (based on a set of six equations given in their Eqs. 2, 3, 5) for precisely estimating sample size for a correlation with desired CI. Their two-stage approximation results agree very well with method 1 above.

Method 3. A sample size equation by Corty and Corty (2011) is available to estimate N for a given choice of r, w and α. Their equation is derived using Fisher's r-to-*z* transformation (Corty, 2007, Field, 2009, Fisher, 1915) to obtain a normal distribution in Fisher's z-variable: $z(r) = 0.5\ln[(1+r)/(1-r)]$. The two z-values [z(r) ± $1.96/\sqrt{(N\text{-}3)}$] define the 95% CI for the associated *z*-distribution, considering that the variance of the *z*-distribution is given by $V(z) = 1/(N\text{-}3)$. The inverse z-to-r transformation is then used to construct a CI for *r*. Beaulieu-Prévost (2006) previously outlined and

Table 2 ■ Sample Size Requirements for 95% Confidence Interval Width CI2w=0.1 for different Pearson Correlation Values | *r*|, based on CIxcorr, Eqs. 4, 7, 8.

$\|r\|$	CI2w	CIxcorr	Eq. 4	Eq. 7	Eq. 8
0.1	0.100	1507	1507	1508	1508
0.3	0.100	1274	1273	1274	1274
0.4	0.100	1086	1084	1086	1087
0.5	0.100	867	864	866	867
0.6	0.100	633	628	631	633
0.7	0.100	404	397	401	404
0.8	0.100	205	195	201	204
0.9	0.100	62	50	57	62

described this method. Corty and Corty's resulting sample size equation, where $r = |r|$, is:

$$N = 15.37/(\ln(B))^2 + 3, \tag{4}$$

where $B = \frac{\sqrt{(1+r+\mathrm{w})(1-r+\mathrm{w})}}{\sqrt{(1-r-\mathrm{w})(1+r-\mathrm{w})}}$.

Corty and Corty provided a table similar to Table 1, based on Eq. 4. A comparison of Eq. 4 values with Table 1 values (based on CIxcorr) shows that Eq. 4 gives values up to one count too low in the range r ≤ 0.3; and as many as 12 counts too low in the range r > 0.3. For planning purposes therefore, the more precise Table 1 is preferred.

Method 4. An alternative equation to plan sample size to achieve a desired (narrow) CI width for correlations is now derived. Referring to Eqs. 3, for 95% CI,

$$\mathrm{UP} - \mathrm{LO} = \mathrm{CI2w} = 2\mathrm{w} = 3.92\sigma_r \tag{5}$$

For a planned sample size N, one may approximate σ_r by the sampling distribution's standard deviation σ_r, given in Eq. 2. Combining Eq. 5 with Eq. 2 gives:

$$\mathrm{w} = 1.96\sigma_r = 1.96\frac{(1-r^2)}{\sqrt{N-1}}, \tag{6}$$

and therefore:

$$N = \frac{3.84(1-r^2)^2}{\mathrm{w}^2} + 1. \tag{7}$$

Eqs. 6-7 are convenient because of their particularly simple r and w dependences. Comparing Eq. 7 and Table 1 N-values (based on CIxcorr) shows that Eq. 7 gives values up to one count too low in the range **|r|** ≤ 0.5; and as many as 5 counts too low in the range **|r|** > 0.5. The excellent agreement for small r follows considering that the Eq. 2 approximation is most precise for small r. Eq. 7 based on Eq. 2 has a greater range of applicability than Eq. 4 based on Fisher r-to-*z* transformation. Eq. 7 provides the basis for deriving the useful Eqs. 8, 9 which follow. The Bonett and Wright (2000) first-stage equation (their Eq. 3) discussed previously differs from Eq. 7 by an additive constant (+3 replaces +1 in Eq. 7) that arises as a result of its derivation based on the variance of Fisher's z-distribution.

Method 5. Eq. 8 is based on a simple and small addition ($6r^2$) to Eq. 7; the addition is expressed mathematically as $\Delta N = 6r^2$. The resulting equation is:

$$N = \frac{3.84(1-r^2)^2}{\mathrm{w}^2} + 1 + 6r^2. \tag{8}$$

The corresponding equation for w is:

$$\mathrm{w} = 1.96\frac{(1-r^2)}{\sqrt{(N-1-6r^2)}}. \tag{9}$$

Eqs. 8, 9 provide an accurate alternative to method 2 of Bonett and Wright (2000). Eq. 8 may be conveniently used to a precision of 1 count for all r and w values in the range of Table 1.

Comparison of methods. Table 2 gives sample size estimates for $|r| = 0.1$ - 0.95 for all the methods discussed above for the particular choice CI width = CI2w = 0.1, for ease in comparing the different methods. All sample size values shown are rounded up to the next higher integer; for example 49.4 is rounded up to 50. The CIxcorr value shown is the highest value of N for which the output CI is closest to but does not exceed the input value CI = 0.100. The contents of Tables 1 and 2 were verified using Monte Carlo simulations. A very large number of correlations (50,000) were generated to obtain the lower and upper 2.5% percentiles. The difference between the two percentiles corresponds to the range CI2w for a 95% CI. Simulation results agreed very well with the CIxcorr and Eq. 8 methods.

Comparing results for $r = 0.9$, $\alpha = 0.05$, CI2w = 0.10, Eqs. 4, 7, 8, and `CIxcorr` give $N = 50$, $N = 57$, $N = 62$, $N = 62$, respectively. More generally, the `CIxcorr`, `sizeCIcorr` and Eq. 8 methods give

the most precise sample size values. For $|r|$ and w values not shown in Table 1, CIxcorr.R or `sizeCIcorr`.R or a simple interpolation based on Table 1 may be used. Finally, regarding the equations of methods 2-5, Eq. 8 is recommended; it is more accurate than Eqs. 4 and 7, and easier to use than the method 2 equations.

Sample Size Planning

For illustration, consider sample size requirements for planning a research project dealing with a binary correlation between characteristics x and y. Table 1 shows that if the correlation is $\rho = 0.1$, it can be measured with CI [0.05, 0.15] via a sample size $N = 1507$; and with CI [0.0,0.20] via a sample size $N = 378$. For another example, Table 1 shows that a sample size $N = 1086$ allows estimating $\rho = 0.4$ within CI [0.35, 0.45]; while $N = 273$ would yield CI [0.30, 0.50]. If one aims to measure $\rho = 0.2$ within CI [0.09, 0.31], meaning that $w \sim 0.11$, Table 1 by interpolation or Eq. 9 shows that this can be achieved with $N \sim 300$. These examples show that the sample size depends on the choices made of the minimum effect size (ρ) and CI width to be measured. Based on these choices, Eqs. 8 or Table 1 can conveniently help researchers select the optimal sample size for their planned projects.

Conclusions

The importance of estimating sample size before conducting quantitative research studies has been stressed. This article reviewed statistical concepts needed for estimating the sample size N to determine correlation coefficients (CCs) between two characteristics, reviewed available methods, and discussed their different ranges of applicability. A convenient equation was derived to help plan sample size for correlations by confidence interval analysis. In addition, a table for planning correlation studies was provided that gives sample sizes needed to achieve a 95% confidence interval (CI) for correlation values ranging from 0.05 to 0.95 and for CI widths ranging from 0.1 to 0.9. Sample size requirements were considered for planning correlation studies.

References

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.), Washington, DC: APA.

Beaulieu-Prévost, D. (2006). Confidence Intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorials in Quantitative Methods for Psychology*, 2, 11-19.

Bonett, D. G., Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, 65, 23-28.

Bonett, D. G. (2014). CIcorr.R and sizeCIcorr.R http://people.ucsc.edu/~dgbonett/ psyc181.html

Corty, E. W. (2007). Using and interpreting statistics: A practical text for the health, behavioral, and social sciences. St. Louis: Mosby Elsevier.

Corty, E. W., Corty, R. W. (2011). Setting sample size to ensure narrow confidence intervals for precise estimation of population values. *Nursing Research*, 60, 148-153.

Cumming, G. (2014). The New Statistics: Why and How, Psychological Science, 25, 7-29.

Descoteaux, J. (2007). Statistical power: An historical introduction. *Tutorials in Quantitative Methods for Psychology*, 3, 28-34.

Field, A. (2009). Discovering statistics using SPSS. London, Great Britain: Sage Publications Limited.

Fisher, R.A. (1915). "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population". *Biometrika, 10*, 507–521.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms, *Journal of the Royal Statistical Society*, 15, 193-232.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research*, 43, 524-555.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2, 93-113.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55, 187-193.

R Foundation for Statistical Computing. (2011). R: a language and environment for statistical computing. [WWW page]: http://cran. r-project. org

StatsToDo. (2014). confidence interval estimation. http://www.statstodo.com/SSizCorr_Pgm.php#

Appendix A

R-function CIxcorr is shown here based on CIcorr of Bonett (2011), with his permission. It calculates a confidence interval for a Pearson correlation (lower and upper levels LL and UL, width CI2w = UL-LL) for chosen correlation coefficient corr, significance alpha, and sample size N.

```
CIxcorr <- function(alpha, corr, N) {
  # Computes a confidence interval for a Pearson correlation
  # Args:
  #   alpha: alpha level for (1-alpha) confidence
  #   corr:  value of correlation
  #   N:     sample size
  # Returns:
  #  confidence interval
  Z   <- qnorm(1 - alpha/2)
  # Z = 1.9599640 for 95% CI
  se  <- sqrt(1/((N - 3)))
  zr  <- log((1 + corr)/(1 - corr))/2
  LL0 <- zr - Z*se
  UL0 <- zr + Z*se
  LL  <- (exp(2*LL0) - 1)/(exp(2*LL0) + 1)
  UL  <- (exp(2*UL0) - 1)/(exp(2*UL0) + 1)
  CI2w <- UL - LL
  #  CI width CI2w output added to Bonett's CIcorr R-function
  CI  <- c(LL, UL, CI2w)
  return(CI)
 }
```

For example, with *alpha* = 0.05 level of significance, 95% CI, and |*r*| = 0.9, input function arguments

```
CIxcorr(0.05, 0.9, 62)
```

to get function output

```
c(LL,UL,CI2w)= 0.83878, 0.93875, 0.09996;
```

and input function arguments

```
CIxcorr(0.05, 0.9, 61)
```

to get function output

```
c(LL,UL,2CIw)= 0.83813, 0.93901,0.10088
```

Appendix B

R-function sizeCIcorr by Bonett (2011) is shown here with his permission, with a minor change in a variable name. It computes a sample size N required to estimate a chosen Pearson correlation coefficient corr with a given significance alpha and CI width CI2w.

```
sizeCIcorr <- function(alpha, corr, CI2w) {
  # Computes sample size required to estimate a correlation with desired precision
  # Args:
  #   alpha: alpha level for 1-alpha confidence
  #   corr:  planning value of correlation
  #   CI2w:     desired confidence interval width
  # Returns:
  #   required sample size
  z   <- qnorm(1 - alpha/2)
  n1  <- ceiling(4*(1 - corr^2)^2*(z/CI2w)^2 + 3)
  zr  <- log((1 + corr)/(1 - corr))/2
  se  <- sqrt(1/(n1 - 3))
  LL0 <- zr - z*se
  UL0 <- zr + z*se
  LL  <- (exp(2*LL0) - 1)/(exp(2*LL0) + 1)
```

```
  UL  <- (exp(2*UL0) - 1)/(exp(2*UL0) + 1)
  N   <- ceiling((n1 - 3)*((UL - LL)/CI2w)^2 + 3)
  return(N)
}
```

For example, with alpha = 0.05 level of significance, 95% CI, and |r| = 0.85, and CI width CI2w = 0.1, input function arguments

```
sizeCIcorr(0.05, 0.85, 0.1)
```

to get function output

```
N = 125.
```

Citation

Moinester, M., & Gottfried, R. (2014). Sample size estimation for correlations with pre-specified confidence interval. *The Quantitative Methods for Psychology*, *10* (2), 124-130.

Received: 15/12/13 ~ Accepted: 10/03/14