2

# NAVAL POSTGRADUATE SCHOOL
# Monterey, California

# THESIS

SAMPLE SIZE FOR CORRELATION ESTIMATES

by

Kemal SALAR

September 1989

Thesis Advisor Glenn F. LINDSAY

Approved for public release; distribution is unlimited.

Unclassified

security classification of this page

## REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a Report Security Classification Unclassified | 1b Restrictive Markings |
| 2a Security Classification Authority | 3 Distribution/Availability of Report<br>Approved for public release; distribution is unlimited. |
| 2b Declassification/Downgrading Schedule | |
| 4 Performing Organization Report Number(s) | 5 Monitoring Organization Report Number(s) |

| | | |
|---|---|---|
| 6a Name of Performing Organization<br>Naval Postgraduate School | 6b Office Symbol<br>*(if applicable)* 55 | 7a Name of Monitoring Organization<br>Naval Postgraduate School |
| 6c Address *(city, state, and ZIP code)*<br>Monterey, CA 93943-5000 | | 7b Address *(city, state, and ZIP code)*<br>Monterey, CA 93943-5000 |
| 8a Name of Funding/Sponsoring Organization | 8b Office Symbol<br>*(if applicable)* | 9 Procurement Instrument Identification Number |
| 8c Address *(city, state, and ZIP code)* | | 10 Source of Funding Numbers |

| | | | |
|---|---|---|---|
| Program Element No | Project No | Task No | Work Unit Accession No |

11 Title *(include security classification)* SAMPLE SIZE FOR CORRELATION ESTIMATES

12 Personal Author(s) Kemal SALAR

| | | | |
|---|---|---|---|
| 13a Type of Report<br>Master's Thesis | 13b Time Covered<br>From To | 14 Date of Report *(year, month, day)*<br>September 1989 | 15 Page Count<br>86 |

16 Supplementary Notation The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

| 17 Cosati Codes | | | 18 Subject Terms *(continue on reverse if necessary and identify by block number)* |
|---|---|---|---|
| Field | Group | Subgroup | Classical and nonparametric sample size determination, Pearson's R, Spearman's r and Kendall's tau |
| | | | |
| | | | |

19 Abstract *(continue on reverse if necessary and identify by block number)*

This thesis examines the classical measure of correlation (Pearson's R) and two nonparametric measures of correlation (Spearman's r and Kendall's $\tau$) with the goal of determining the number of samples needed to estimate a correlation coefficient with a 95% confidence level. For Pearson's R. tables, graphs, and computer programs are developed to find the sample number needed for a desired confidence interval size. Nonparametric measures of correlation (Spearman's r and Kendall's $\tau$) are also examined for appropriate sample numbers when a specific confidence interval size desired.

| | | |
|---|---|---|
| 20 Distribution/Availability of Abstract<br>☒ unclassified/unlimited □ same as report □ DTIC users | 21 Abstract Security Classification<br>Unclassified | |
| 22a Name of Responsible Individual<br>Glenn F. LINDSAY | 22b Telephone *(include Area code)*<br>(408) 373-6284 | 22c Office Symbol<br>55Ls |

DD FORM 1473,84 MAR

83 APR edition may be used until exhausted
All other editions are obsolete

security classification of this page

Unclassified

Approved for public release; distribution is unlimited.

# SAMPLE SIZE FOR CORRELATION ESTIMATES

by

Kemal SALAR
LTJG, Turkish Navy
B.S., Turkish Naval Academy, 1983

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1989

Author: ________________________________
Kemal SALAR

Approved by: Glenn F. Lindsay
Glenn F. LINDSAY, Thesis Advisor

William J Walsh
William J. WALSH, Second Reader

P. Purdue
Peter PURDUE, Chairman,
Department of Operations Research

# ABSTRACT

This thesis examines the classical measure of correlation (Pearson's R) and two nonparametric measures of correlation (Spearman's r and Kendall's tau τ) with the goal of determining the number of samples needed to estimate a correlation coefficient with a 95% confidence level. For Pearson's R, tables, graphs, and computer programs are developed to find the sample number needed for a desired confidence interval size. Nonparametric measures of correlation (Spearman's r and Kendall's τ) are also examined for appropriate sample numbers when a specific confidence interval size desired.

tau

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# I. INTRODUCTION

Everyone wants to know how big a sample is needed. In many forms of weapon system testing, there is always a decision about the sample size, and this decision is very important because an unnecessarily large sample takes extra time and increases costs. If the purpose of the testing is to estimate a value, then the test needs to give a good estimate (represented by a small confidence interval). At the same time it is desirea to use the smallest sample size required for the desired accuracy. The topic of this thesis is to develop a way to find sample sizes when the testing is to estimate a correlation coefficient.

There are many ways to find a sample size. In this thesis, the desired confidence interval size will be used as the basis for finding sample size. It is important to note that the size of the confidence interval depends upon the number of observations which are taken, and in general, if a bigger sample size, is used, then the confidence interval will be smaller.

The problem of finding the sample size for estimates of proportions, given a desired confidence interval size, has been studied for a variety of cases [Ref. 1], [Ref. 2] and [Ref. 3]. The work reported here looks at sampling done to estimate a correlation coefficient, and the sample size that is needed to produce a desired confidence interval for that correlation coefficient. This work investigates and gives some opinion about the necessary sample size that would be used when estimation involves Pearson's R, and also discusses the sample size problem when nonparametric statistical methods are employed. For each of these measures the relationship between sample size and confidence interval size will be analyzed, so that graphs and tables can be provided to assist a decision maker in finding the necessary sample size to obtain a desired confidence interval to estimate a correlation coefficient value.

In Chapter II a description of the classical sample measure of corre'ation (Pearson's R) and the confidence intervals that can be developed using the normal approximation method will be provided. The third chapter addresses sample size determination for estimating a correlation coefficient using confidence intervals. This chapter will discuss how computer programs wer developed and used, and graphs and tables were constructed to determine the required sample size to obtain a desired 95% confidence interval for a correlation coefficient. A comparison of methods is done to give easy to use results about sample sizes for estimating correlation coefficient values. Then, in Chapter IV, the use of Spearman and Kendall test statistics, and the problem of finding the sample size that is needed to produce a confidence interval of desired size will be described. Also, in this chapter a comparison will be done on the sample size results that are needed for a desired confidence interval size, using Pearson's R, Spearman's r and Kendall's *tau*.

The final chapter will summarize this research, and provide some suggestions for further research and study.

# II. CORRELATION AND THE PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT

In this chapter an explanation will be given on how to use the classical correlation coefficient method for a desired confidence interval. First, the F ,arson product-moment correlation coefficient will be studied. Then this information will be usғ d to show how estimates of the population correlation coefficient may be obtained. In the final part of this chapter, different procedures will be reviewed to find a confidence interval for population correlation coefficient by using the normal approximation method.

## A. THE PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT

Before determining any sample sizes, a brief introduction about the Pearson product-moment correlation coefficient will be provided. Gibbon states: "In general, if X and Y are two random variables with a bivariate probability distribution, their covariance, in a certain sense, reflects the direction and amount of correlation or correspondence between the variables. The covariance is large and positive if there is a high probability that large (small) values of X are associated with large (small) values of Y. On the other hand, if the correspondence is inverse so that large (small) values of X generally occur in conjunction with small (large) values of Y, their covariance is large and negative. This comparative type of correlation is referred to as concordance or agreement. The covariance parameter as a measure of correlation is difficult to interpret because its value depends on the orders of magnitude and units of the random variables concerned. A nonabsolute or relative measure of correlation circumvents this difficulty." [Ref. 4: p.206]

The Pearson product-moment correlation coefficient, defined as

$$\rho(X,Y) = \frac{cov(X.Y)}{\sqrt{(Var(X)Var(Y))}} \tag{2.1}$$

[Ref. 4: p.206] is variant under changes of scale and location in X and Y, and in classical statistics this parameter is usually employed as the measure of correlation in a bivariate distribution. The absolute value of the correlation coefficient does not exceed 1, and its sign is determined by the sign of the covariance. If X and Y are independent random variables, then their correlation should be zero, but the converse is not true in general. "If the main justification for the use of $\rho$ as a measure of association is that the bivariate normal is such an important distribution in classical statistics and zero correlation is equivalent to independence for that particular population, this reasoning has little significance in nonparametric statistics." [Ref. 4: p.206]

First of all, a measure of correlation between X and Y must satisfy the following requirements in order to be a good relative measure of association:

- The measure of correlation value should be between -1 and +1;
- If the larger values of X tend to be paired with the larger values of Y, and the smaller values of X tend to be paired with the smaller values of Y, then the measure of correlation should be positive, and if the tendency is strong then it is close to +1;
- If the larger values of X tend to be paired with the smaller values of Y, and vice versa, then the measure of correlation should be negative and if the tendency is strong then it is close to -1;
- If the values of X and the values of Y are randomly paired, then the measure of correlation should be fairly close to zero. It means that X and Y are independent.

## B. ESTIMATION OF THE POPULATION CORRELATION COEFFICIENT

Most of the time, the value of the population correlation coefficient ($\rho$) is unknown, but it must be estimated from our sample. The sample correlation coefficient is a random variable which is used in situations where the data consist of pairs of numbers. A bivariate random sample of size n is represented by $(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)$.

Suppose a random sample of n pairs $(X_1, Y_1),(X_2, Y_2), \ldots,(X_n, Y_n)$ is drawn from a bivariate population with Pearson product-moment correlation coefficient $\rho$. Then, in classical statistics, the estimate used for $\rho$ is the sample correlation coefficient R, defined as

$$R = \frac{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left(\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=0}^{n}(Y_i - \overline{Y})^2\right)^{\frac{1}{2}}} \quad , \tag{2.2}$$

[Ref. 5: p.244] where $\overline{X}$ and $\overline{Y}$ are the sample means

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.3}$$

and

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad . \tag{2.4}$$

If the numerator and denominator in Equation 2.2 are divided by n, then R becomes

$$R = \frac{\displaystyle\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left[\displaystyle\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{\frac{1}{2}}\left[\displaystyle\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{\frac{1}{2}}} \quad , \tag{2.5}$$

and it can be seen in Equation 2.5 that the numerator is the sample covariance and the denominator is the product of the two sample standard deviations (S). It means that this equation is similar in form to the population correlation coefficient defined in Equation 2.2.

This sample measure of correlation may be used on a set of data without any requirements, but it is difficult to interpret unless the scale of measurements is at least interval. The important point is that R is a random variable with a distribution function, and the distribution function of R depends on the bivariate distribution function of (X,Y).

## C. CONFIDENCE INTERVALS FOR THE CORRELATION COEFFICIENT.

If it is desired to determine confidence intervals for $\rho$ (population correlation coefficient), then the sampling distribution for the correlation coefficient R must be known. If (X,Y) is bivariate normal, then the expected value and variance of R are approximately

$$E(R) \cong \rho, \tag{2.6}$$

and

$$VAR(R) \cong \frac{(1-\rho^2)^2}{n} \quad \textit{provided n is not too small} \tag{2.7}$$

[Ref. 6: p.462]. There already exist confidence intervals for confidence coefficients of 95 percent. These were determined by F. N. David and are reproduced in Figure 5 on page 49 in Appendix A. In this figure, the abscissa is the estimated correlation coefficient from the sample data. For each given sample size and value of R there is a confidence interval for $\rho$, varying as R goes from -1.0 to +1.0. For example, for R = 0.60, n = 5 the 95 percent confidence interval is about -0.5 < $\rho$ < 0.91.

If a figure similar to that of Appendix A does not exist, or if we want to find the exact number for interval, the normal distribution can be used to obtain an approximation.

The statistic commonly used is

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) = \tanh^{-1}R, \tag{2.9}$$

which is distributed approximately normal with an expected value

$$E(Z) \cong \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \tag{2.10}$$

and variance

$$\sigma_z^2 \cong (n-3)^{-1} \qquad (2.11)$$

[Ref. 6: p.463]. Note here that Z is not the standard normal variable. Using this transformation, the confidence interval for $\rho$ can be calculated. Having calculated the estimate for $\rho$, namely R, we compute Z and the statistic

$$K_1 = \left[ Z - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\right]\sqrt{n-3} \cong \frac{Z - E(Z)}{\sigma(Z)} \qquad (2.12)$$

where $K_1$ approximately follows a standard normal distribution.

Using the normal approximation, there wi'l be 95% certainly that

$$-1.96 < \frac{Z - E(Z)}{\sigma(Z)} < 1.96 \qquad (2.13)$$

and the 95% confidence interval of E(Z) will be

$$\begin{aligned} \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) - 1.96\sigma(Z) < E(Z) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \\ < \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) + 1.96\sigma(Z). \end{aligned} \qquad (2.14)$$

From 2.10

$$\exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) - 1.96\sigma(Z)\right)\right\} < \left(\frac{1+\rho}{1-\rho}\right), \qquad (2.15a)$$

and

$$\left(\frac{1+\rho}{1-\rho}\right) < \exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) + 1.96\sigma(Z)\right)\right\}. \qquad (2.15b)$$

If the left side of 2.15a is $L_1$ and the right side of 2.15b is $U_1$, then

$$L_1 < \left( \frac{1+\rho}{1-\rho} \right) < U_1. \tag{2.16}$$

Values for $L_1$ and $U_1$ can be computed from sample results. From 2.16 the 95% confidence interval for $\rho$ will be

$$\left( \frac{L_1 - 1}{L_1 + 1} \right) < \rho < \left( \frac{U_1 - 1}{U_1 + 1} \right) \; . \tag{2.17}$$

For example, if the data has 10 observations and the sample correlation coefficient R = 0.60, the 95% confidence interval can be estimated. Using the confidence belts in Figure 5 on page 49 in Appendix A, the bounds are 0.05 and 0.89. These results are rough. Using Equations 2.9, 2.10, and 2.11, we have

$$Z = \frac{1}{2} \ln \left( \frac{1.6}{0.4} \right) = 0.6932$$

and

$$\sigma(Z) = \frac{1}{\sqrt{n-3}} = 0.378 \; .$$

The 95 percent confidence limits for E(Z) are then

$$0.6932 - 1.96 \times 0.378 < E(Z) < 0.6932 + 1.96 \times 0.378$$

which reduce to

$$-0.047768 < E(Z) < 1.4341 \; .$$

The inequalities can be written as

$$-\,0.04768 < \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right) < 1.4341 \; ,$$

and combining Equation 2.15a and 2.15b to obtain

$$\exp\{2 \times (-0.047768)\} < \left( \frac{1+\rho}{1-\rho} \right) < \exp\{2 \times (1.4341)\} \quad ,$$

results in $L_1 = 0.90905$, and $U_1 = 17.85$. Thus from Equation 2.17 the 95% confidence interval for $\rho$ is

$$\left( \frac{0.90905 - 1}{0.90905 + 1} \right) < \rho < \left( \frac{17.85 - 1}{17.85 + 1} \right)$$

which reduces to

$$-0.048 < \rho < 0.8925 \ .$$

Confidence interval size increases as the sample correlation coefficient R approaches zero, and the largest confidence interval that could result will occur when R = 0. Here

$$L_1 = \exp\left\{ -\frac{3.92}{\sqrt{n-3}} \right\}$$

and

$$U_1 = \exp\left\{ \frac{3.92}{\sqrt{n-3}} \right\}$$

so that the largest confidence interval size is

$$2A = \frac{\exp\left\{ \frac{3.92}{\sqrt{n-3}} \right\} - 1}{\exp\left\{ \frac{3.92}{\sqrt{n-3}} \right\} + 1} - \frac{\exp\left\{ -\frac{3.92}{\sqrt{n-3}} \right\} - 1}{\exp\left\{ -\frac{3.92}{\sqrt{n-3}} \right\} + 1} .$$

Results for this case are shown in Table 1. The table provides largest possible confidence interval sizes that could result for various sample sizes. For example, if a 95% confidence interval for $\rho$ is desired which is no greater than 0.2, then a minimum sample of size 367 would guarantee that result.

Table 1. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.00

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.00 | 5,085 | -0.025 | 0.025 | 0.05 |
| 0.00 | 3,638 | -0.03 | 0.03 | 0.06 |
| 0.00 | 2,738 | -0.035 | 0.035 | 0.07 |
| 0.00 | 2,128 | -0.04 | 0.04 | 0.08 |
| 0.00 | 1,704 | -0.045 | 0.045 | 0.09 |
| 0.00 | 1,395 | -0.05 | 0.05 | 0.10 |
| 0.00 | 1,163 | -0.055 | 0.055 | 0.11 |
| 0.00 | 984 | -0.06 | 0.06 | 0.12 |
| 0.00 | 844 | -0.065 | 0.065 | 0.13 |
| 0.00 | 732 | -0.07 | 0.07 | 0.14 |
| 0.00 | 641 | -0.075 | 0.075 | 0.15 |
| 0.00 | 565 | -0.08 | 0.08 | 0.16 |
| 0.00 | 503 | -0.085 | 0.085 | 0.17 |
| 0.00 | 450 | -0.09 | 0.09 | 0.18 |
| 0.00 | 367 | -0.10 | 0.10 | 0.20 |
| 0.00 | 333 | -0.105 | 0.105 | 0.21 |
| 0.00 | 279 | -0.115 | 0.115 | 0.23 |
| 0.00 | 237 | -0.125 | 0.125 | 0.25 |
| 0.00 | 220 | -0.13 | 0.13 | 0.26 |
| 0.00 | 190 | -0.14 | 0.14 | 0.28 |
| 0.00 | 166 | -0.15 | 0.15 | 0.30 |

This chapter explained how confidence intervals for the correlation coefficient may be obtained. The next chapter will present methods to determine the needed sample size for estimating a correlation coefficient by using confidence intervals.

# III. SAMPLE SIZE FOR ESTIMATING A CORRELATION COEFFICIENT USING CONFIDENCE INTERVALS

In Chapter II a discussion of the Pearson product-moment correlation coefficient, estimation of the population correlation coefficient, and confidence interval for population correlation coefficient was conducted. This chapter will study sample size determination for estimating the correlation coefficient, using confidence intervals and the normal approximation method that were explained in Chapter II. Then, we will discuss how we can develop and use computer programs, graphs and the tables to determine the required sample size to obtain a desired 95% confidence interval for a sample correlation coefficient value. The finai part of this chapter will show the required sample sizes for different sample correlation coefficient values.

## A. SAMPLE SIZE DETERMINATION USING THE NORMAL APPROXIMATION METHOD FOR THE ESTIMATED CORRELATION COEFFICIENT VALUE

Suppose a confidence intel.al of size 2A is desired for the correlation coefficient. Then, from Equation 2.17.

$$\begin{aligned} 2A &= \textit{Upper Confidence Limit} - \textit{Lower Confidence Limit} \\ &= \left(\frac{U_1 - 1}{U_1 + 1}\right) - \left(\frac{L_1 - 1}{L_1 + 1}\right) \end{aligned} \qquad (3.1)$$

where

$$U_1 = \exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) + 1.96\sigma(Z)\right)\right\}, \qquad (3.2)$$

and

$$L_1 = \exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) - 1.96\sigma(Z)\right)\right\}, \qquad (3.3)$$

and from Equation 2.11

$$\sigma(Z) = \frac{1}{\sqrt{n-3}}. \tag{3.4}$$

Thus, 95% confidence interval size (2A) will be equal to

$$2A = \frac{\exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) + \frac{1.96}{\sqrt{n-3}}\right)\right\} - 1}{\exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) + \frac{1.96}{\sqrt{n-3}}\right)\right\} + 1} - \frac{\exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) - \frac{1.96}{\sqrt{n-3}}\right)\right\} - 1}{\exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) - \frac{1.96}{\sqrt{n-3}}\right)\right\} + 1}. \tag{3.5}$$

If Equation 3.5 could be solved for n in terms of 2A, there would exist a closed expression by which the needed sample size could be computed. However, it is very hard or impossible to solve Equation 3.5 for n in terms of 2A, because of the complexity. Although a closed expression for n could not be obtained a table can still be constructed using n as the independent variable, and solving for 2A. Such a table could then be used to estimate the needed sample size. given a value for 2A.

However, a major difficulty still remains. From the form of $U_1$ and $L_1$ in Equation 3.2 and 3.3, it is seen that in subtracting the lower confidence limit from the upper confidence limit to obtain 2A, the sample result R does not vanish. Therefore, looking at Equation 3.5, to determine the required sample size, an estimate of the sample correlation coefficient value must be done.

It is a curious result to see that in order to determine the sample size needed to estimate a correlation coefficient $\rho$ by a value R, a first guess must be made at the result R of a sample not yet taken. However, in many cases some advance knowledge about R will be known, e.g., whether it is positive or negative, or whether it is greater or less than 0.5. It may not be likely that

R is to be very high, (say, $R < 0.8$). In any event, the tables will show that n is not extremely sensitive to the guessed value of R.

For example, suppose that R = 0.975 is estimated by the decision maker, and also suppose that the decision maker desires the confidence interval size to be 0.10. Then the decision maker can find n = 10 from Table 2. It is important to note that, when R = -0.975, the confidence interval size is the same with R = 0.975, but as can be seen from Table 2 and Table 3 on page 14, they have different upper and lower bounds because of the sign. For example, the values of the lower and the upper bounds will be 0.89 and 0.99 for R = 0.975 and -0.99 and -0.89 for R = -0.975. Because of this symmetry negative sample correlation coefficient values will not be discussed for the rest of the study. Also, since our purpose is finding sample size, there is no interest in the upper and the lower bound, but only the confidence interval size.

Table 2. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.975

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.975 | 20 | 0.94 | 0.99 | 0.05 |
| 0.975 | 16 | 0.93 | 0.99 | 0.06 |
| 0.975 | 14 | 0.92 | 0.99 | 0.07 |
| 0.975 | 12 | 0.91 | 0.99 | 0.08 |
| 0.975 | 11 | 0.90 | 0.99 | 0.09 |
| 0.975 | 10 | 0.89 | 0.99 | 0.10 |
| 0.975 | 9 | 0.88 | 0.99 | 0.11 |
| 0.975 | 8 | 0.861 | 0.99 | 0.13 |
| 0.975 | 7 | 0.841 | 1.0 | 0.16 |
| 0.975 | 6 | 0.781 | 1.0 | 0.21 |
| 0.975 | 5 | 0.661 | 1.0 | 0.34 |

Table 3. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = -0.975**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| -0.975 | 20 | -0.99 | -0.94 | 0.05 |
| -0.975 | 16 | -0.99 | -0.93 | 0.06 |
| -0.975 | 14 | -0.99 | -0.92 | 0.07 |
| -0.975 | 12 | -0.99 | -0.91 | 0.08 |
| -0.975 | 11 | -0.99 | -0.90 | 0.09 |
| -0.975 | 10 | -0.99 | -0.89 | 0.10 |
| -0.975 | 9 | -0.99 | -0.88 | 0.11 |
| -0.975 | 8 | -0.99 | -0.86 | 0.13 |
| -0.975 | 7 | -1.0 | -0.84 | 0.16 |
| -0.975 | 6 | -1.0 | -0.78 | 0.21 |
| -0.975 | 5 | -1.0 | -0.66 | 0.34 |

As a second example, suppose the estimate of the sample correlation coefficient is 0.90 and we calculate the 95% confidence interval using the normal approximation method. For a given sample size, this will yield the confidence limits. Table 4 on page 16 shows the number of samples required to obtain different 95% confidence interval sizes for various values of 2A when R = 0.90.

An APL program, named "Tez" was written to obtain the sample size, the upper and the lower confidence limits, and the confidence interval size (2A) after inputting any estimated value for sample correlation coefficient. For R = 0.90, Table 4 on page 16 was constructed by executing this APL program, and the program was used to create similar tables in this chapter and in Appendix B. Table 5 on page 17, Table 6 on page 18 and Table 7 on page 19 show the required sample size for 95% confidence intervals using R =

0.80, R = 0.75 and R = 0.10. Tables for the other values of R are in Appendix C.

It is important to note that when $R = \pm 1$, the Z statistic in Equation 2.9 goes to infinity. Because of this the sample size cannot be calculated for the desired confidence interval when $R = \pm 1$. However, an quess of $R = \pm 1$ will not used.

Table 4. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.90**

| Estimated Correiation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.90 | 195 | 0.87 | 0.92 | 0.05 |
| 0.90 | 160 | 0.87 | 0.93 | 0.06 |
| 0.90 | 120 | 0.86 | 0.93 | 0.07 |
| 0.90 | 95 | 0.85 | 0.93 | 0.08 |
| 0.90 | 75 | 0.85 | 0.94 | 0.09 |
| 0.90 | 60 | 0.84 | 0.94 | 0.10 |
| 0.90 | 50 | 0.83 | 0.94 | 0.11 |
| 0.90 | 45 | 0.82 | 0.94 | 0.12 |
| 0.90 | 40 | 0.82 | 0.95 | 0.13 |
| 0.90 | 35 | 0.81 | 0.95 | 0.14 |
| 0.90 | 30 | 0.80 | 0.95 | 0.15 |
| 0.90 | 27 | 0.79 | 0.96 | 0.16 |
| 0.90 | 25 | 0.78 | 0.96 | 0.17 |
| 0.90 | 23 | 0.78 | 0.96 | 0.18 |
| 0.90 | 21 | 0.77 | 0.96 | 0.19 |
| 0.90 | 20 | 0.76 | 0.96 | 0.20 |
| 0.90 | 19 | 0.75 | 0.96 | 0.21 |
| 0.90 | 17 | 0.74 | 0.96 | 0.22 |
| 0.90 | 16 | 0.73 | 0.97 | 0.24 |
| 0.90 | 15 | 0.72 | 0.97 | 0.25 |
| 0.90 | 14 | 0.71 | 0.97 | 0.26 |
| 0.90 | 13 | 0.69 | 0.97 | 0.28 |
| 0.90 | 12 | 0.67 | 0.97 | 0.30 |

Table 5. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.80

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.80 | 664 | 0.77 | 0.82 | 0.05 |
| 0.80 | 477 | 0.77 | 0.83 | 0.06 |
| 0.80 | 360 | 0.76 | 0.83 | 0.07 |
| 0.80 | 282 | 0.75 | 0.93 | 0.08 |
| 0.80 | 226 | 0.75 | 0.84 | 0.09 |
| 0.80 | 186 | 0.74 | 0.84 | 0.10 |
| 0.80 | 156 | 0.74 | 0.85 | 0.11 |
| 0.80 | 133 | 0.73 | 0.85 | 0.12 |
| 0.80 | 115 | 0.72 | 0.85 | 0.13 |
| 0.80 | 101 | 0.72 | 0.86 | 0.14 |
| 0.80 | 89 | 0.71 | 0.86 | 0.15 |
| 0.80 | 79 | 0.71 | 0.87 | 0.16 |
| 0.80 | 71 | 0.70 | 0.87 | 0.17 |
| 0.80 | 64 | 0.69 | 0.87 | 0.18 |
| 0.80 | 58 | 0.68 | 0.87 | 0.19 |
| 0.80 | 53 | 0.68 | 0.88 | 0.20 |
| 0.80 | 49 | 0.67 | 0.88 | 0.21 |
| 0.80 | 45 | 0.66 | 0.88 | 0.22 |
| 0.80 | 42 | 0.66 | 0.89 | 0.23 |
| 0.80 | 39 | 0.65 | 0.89 | 0.24 |
| 0.80 | 36 | 0.64 | 0.89 | 0.25 |
| 0.80 | 34 | 0.63 | 0.89 | 0.26 |
| 0.80 | 32 | 0.63 | 0.90 | 0.27 |
| 0.80 | 30 | 0.62 | 0.90 | 0.28 |
| 0.80 | 28 | 0.61 | 0.90 | 0.29 |
| 0.80 | 27 | 0.60 | 0.90 | 0.30 |

**Table 6. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.75**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.75 | 983 | 0.72 | 0.78 | 0.05 |
| 0.75 | 707 | 0.72 | 0.78 | 0.06 |
| 0.75 | 534 | 0.71 | 0.78 | 0.07 |
| 0.75 | 412 | 0.70 | 0.78 | 0.08 |
| 0.75 | 332 | 0.70 | 0.79 | 0.09 |
| 0.75 | 272 | 0.69 | 0.79 | 0.10 |
| 0.75 | 228 | 0.69 | 0.80 | 0.11 |
| 0.75 | 193 | 0.68 | 0.80 | 0.12 |
| 0.75 | 167 | 0.68 | 0.81 | 0.13 |
| 0.75 | 145 | 0.67 | 0.81 | 0.14 |
| 0.75 | 128 | 0.66 | 0.81 | 0.15 |
| 0.75 | 113 | 0.66 | 0.82 | 0.16 |
| 0.75 | 102 | 0.65 | 0.82 | 0.17 |
| 0.75 | 92 | 0.64 | 0.83 | 0.18 |
| 0.75 | 83 | 0.64 | o.83 | 0.19 |
| 0.75 | 75 | 0.63 | 0.33 | 0.20 |
| 0.75 | 69 | 0.62 | 0.83 | 0.21 |
| 0.75 | 63 | 0.62 | 0.84 | 0.22 |
| 0.75 | 58 | 0.61 | 0.84 | 0.23 |
| 0.75 | 54 | 0.60 | 0.84 | 0.24 |
| 0.75 | 50 | 0.60 | 0.85 | 0.25 |
| 0.75 | 47 | 0.59 | 0.85 | 0.26 |
| 0.75 | 44 | 0.58 | 0.86 | 0.27 |
| 0.75 | 41 | 0.58 | 0.86 | 0.28 |
| 0.75 | 39 | 0.57 | 0.86 | 0.29 |
| 0.75 | 37 | 0.56 | 0.86 | 0.30 |

Table 7. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.10

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.10 | 4998 | 0.07 | 0.13 | 0.05 |
| 0.10 | 3582 | 0.07 | 0.13 | 0.06 |
| 0.10 | 2682 | 0.06 | 0.14 | 0.07 |
| 0.10 | 2089 | 0.06 | 0.14 | 0.08 |
| 0.10 | 1677 | 0.05 | 0.14 | 0.09 |
| 0.10 | 1369 | 0.05 | 0.15 | 0.10 |
| 0.10 | 1140 | 0.04 | 0.15 | 0.11 |
| 0.10 | 965 | 0.04 | 0.16 | 0.12 |
| 0.10 | 828 | 0.03 | 0.17 | 0.13 |
| 0.10 | 717 | 0.03 | 0.17 | 0.14 |
| 0.10 | 629 | 0.03 | 0.18 | 0.15 |
| 0.10 | 555 | 0.02 | 0.18 | 0.16 |
| 0.10 | 495 | 0.01 | 0.18 | 0.17 |
| 0.10 | 443 | 0.01 | 0.19 | 0.18 |
| 0.10 | 398 | 0.00 | 0.19 | 0.19 |
| 0.10 | 360 | 0.00 | 0.20 | 0.20 |
| 0.10 | 327 | -0.01 | 0.21 | 0.21 |
| 0.10 | 299 | -0.01 | 0.21 | 0.22 |
| 0.10 | 274 | -0.02 | 0.21 | 0.23 |
| 0.10 | 253 | -0.02 | 0.22 | 0.24 |
| 0.10 | 234 | -0.03 | 0.22 | 0.25 |
| 0.10 | 215 | -0.03 | 0.23 | 0.26 |
| 0.10 | 200 | -0.04 | 0.23 | 0.27 |
| 0.10 | 186 | -0.04 | 0.24 | 0.28 |
| 0.10 | 174 | -0.05 | 0.24 | 0.29 |
| 0.10 | 163 | -0.05 | 0.25 | 0.30 |

Some graphs are provided which show the difference between sample sizes for different quesses of the sample correlation coefficient value. These graphs can also be used to determine the appropriate sample size for a desired confidence interval. Figure 1 on page 21 shows the sample size and confidence interval for R = 0.95 and R = 0.90. From this figure it is obvious that sample size increases as R decreases. Also there is a a high sensitivity in sample size to the guess of R. However, when our guess of correlation is smaller, (say, R less than 0.6) then n will not be as sensitive. In reality, we might assume R is not likely to be very high, (say, $R < 0.8$).

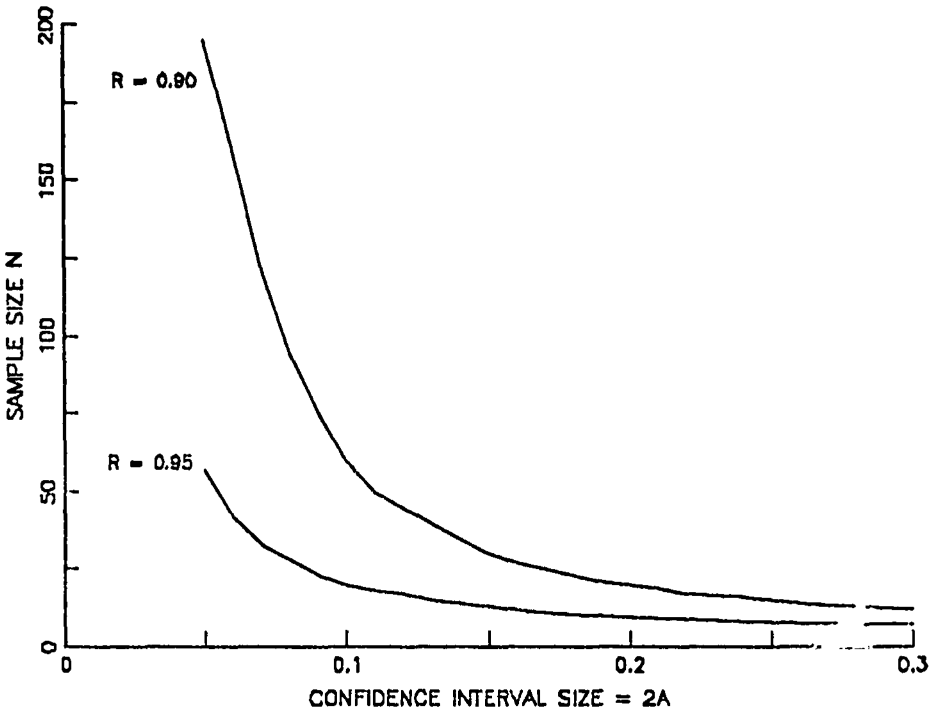


Figure 1. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.95 AND R = 0.90

Figure 2 on page 22 shows the sample size and the confidence interval for R = 0.65 and R = 0.45, and Figure 3 on page 23 shows the sample size and the confidence interval size for R = 0.55 and R = 0.35. From these two figures we can find the required sample sizes approximately, and we see that n is not too sensitive to the guessed value of R. The graphs of different sample correlation coefficient values are helpful in presenting the sensitivity differences in the sample size. Other graphs with different sample correlation coefficient values are in Appendix D.

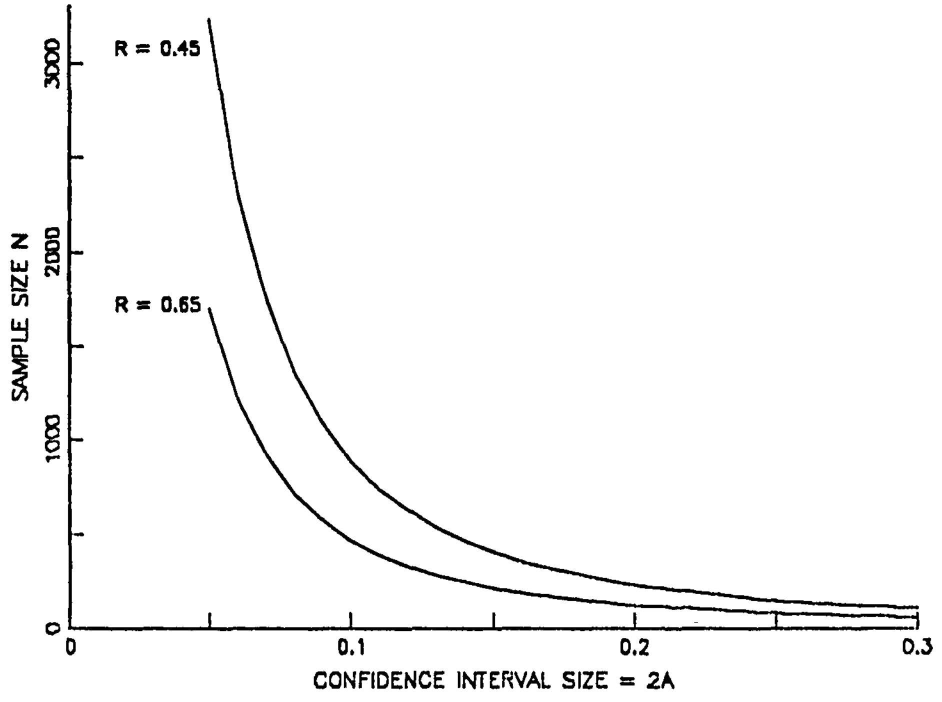


Figure 2. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.65 AND R = 0.45

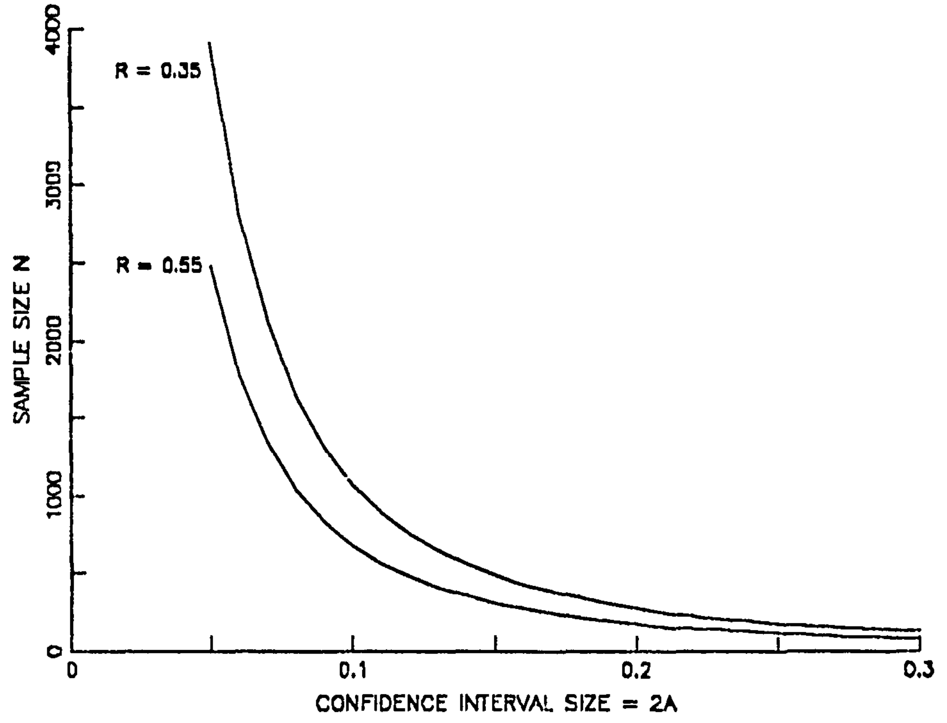


Figure 3. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.55 AND R = 0.35

## B. COMPARISON OF SAMPLE SIZES FOR DIFFERENT CORRELATION COEFFICIENT VALUES

A comparison of the results for different correlation coefficient values shows that as the correlation coefficient gets larger in absolute value, then the required sample size gets smaller for a desired confidence interval size. Table 8 on page 25 shows the results obtained from the computer program for different combinations of sample correlation coefficient estimates and confidence interval sizes. For example, if a confidence interval size 2A of 0.15 is desired then the required sample size is 20 for R = 0.925, 30 for R = 0.90, 171 for R = 0.70, 363 for R = 0.5 and 641 for R = 0.0. Further, a confidence interval of size 0.2 and a prior estimate of R = 0.6 could reduce the sample size to n = 153, which is less than half the n = 367 observations that would be required under total uncertainty about R (estimate R = 0.0).

How can this table be used when the sample correlation is not yet known?. The table can provide general guidance to relieve some of the mystery in choosing the size of a sample. For example, if a maximum confidence interval of size 0.2 is desired and the variables are assumed to be highly correlated, a sample of 50 should work; while if the correlation is assumed small, then several hundred observations will be needed.

Table 8. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL

| 95% Confidence Interval Size = 2A | Estimated sample correlation coefficient value | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.0 | 0.1 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.925 |
| 0 05 | 5,085 | 4,998 | 2,860 | 2,082 | 1,326 | 664 | 195 | 113 |
| 0.06 | 3,638 | 3,582 | 2,049 | 1,495 | 945 | 477 | 160 | 83 |
| 0.07 | 2,738 | 2,682 | 1,542 | 1,123 | 715 | 360 | 120 | 64 |
| 0.08 | 2,128 | 2,089 | 1,199 | 875 | 558 | 282 | 95 | 52 |
| 0.09 | 1,704 | 1,677 | 961 | 701 | 448 | 226 | 75 | 43 |
| 0.10 | 1,395 | 1,369 | 787 | 575 | 367 | 186 | 60 | 36 |
| 0.11 | 1,163 | 1,140 | 657 | 480 | 307 | 156 | 50 | 31 |
| 0.12 | 984 | 965 | 556 | 407 | 260 | 133 | 45 | 28 |
| 0.13 | 844 | 828 | 477 | 349 | 224 | 115 | 40 | 25 |
| 0.14 | 732 | 717 | 414 | 303 | 195 | 101 | 35 | 22 |
| 0.15 | 641 | 629 | 363 | 266 | 171 | 89 | 30 | 20 |
| 0.16 | 565 | 555 | 320 | 235 | 151 | 790 | 27 | 19 |
| 0.17 | 503 | 495 | 285 | 209 | 135 | 71 | 25 | 17 |
| 0.18 | 450 | 443 | 255 | 190 | 121 | 64 | 23 | 16 |
| 0.19 | 405 | 398 | 230 | 169 | 110 | 58 | 21 | 15 |
| 0.20 | 367 | 360 | 209 | 153 | 100 | 53 | 20 | 14 |
| 0.21 | 333 | 327 | 190 | 140 | 91 | 49 | 19 | 13 |
| 0.22 | 304 | 299 | 174 | 128 | 84 | 45 | 17 | 13 |
| 0.23 | 279 | 274 | 159 | 118 | 77 | 42 | 17 | 12 |
| 0.24 | 257 | 253 | 147 | 109 | 71 | 39 | 16 | 12 |
| 0.25 | 237 | 234 | 136 | 100 | 66 | 36 | 15 | 11 |
| 0.26 | 220 | 215 | 126 | 93 | 62 | 34 | 14 | 11 |
| 0.27 | 204 | 200 | 117 | 87 | 57 | 32 | 14 | 11 |
| 0.28 | 190 | 186 | 109 | 81 | 54 | 30 | 13 | 10 |
| 0.29 | 178 | 174 | 102 | 76 | 51 | 28 | 13 | 10 |
| 0.30 | 166 | 163 | 96 | 71 | 48 | 27 | 12 | 9 |

This chapter has discussed the problem of determining a sample size to estimate a correlation coefficient for a desired confidence interval, when Pearson's R is used. The next chapter will present two nonparametric measures of correlation (Spearman's and Kendall's test statistics) and explore the problem of finding a sample size for these cases.

## IV. NONPARAMETRIC MEASURES OF CORRELATION, AND SAMPLE SIZE

In the late 1930's a different approach to the problem of finding probabilities began to gather some momentum. This new package of statistical procedures became known as "nonparametric statistics," and the methods often involve less computational work, and therefore are often easier and quicker to apply than other statistical methods. [Ref. 5: p.3]

Included among methods described as "nonparametric" are procedures providing a measure of correlation when the bivariate data (X, Y) are on strict ordinal scales. An example of such data for sample size five could be

| X | Y |
|---|---|
| 2 | 4 |
| 3 | 2 |
| 5 | 3 |
| 1 | 1 |
| 4 | 5. |

where $X_1 < X_2$ would imply that $X_2$ possesses more of the property being measured than $X_1$, and $Y_1 < Y_2$ would imply that $Y_2$ possesses more of the property than $Y_1$. Ordinal data can arise directly from the measuring procedure in the experiment, or can be obtained from interval or ratio scaled data. An example of the latter would be bivariate data of the temperatures in centigrade in Istanbul and Izmir for five days:

| X | Y |
|---|---|
| 23 | 27 |
| 25 | 22 |
| 30 | 24 |
| 22 | 20 |
| 27 | 31. |

When reduced from an interval to an ordinal scale, this data would be as in the example shown above.

In the previous chapter we discussed sample size determination for estimating a correlation coefficient using confidence intervals and compared the sample sizes for different sample correlation coefficient values. In general, the sampling distributions of R depends upon the form of the bivariate population from which the sample of pairs is drawn. More importantly, Pearson's R as a correlation measure requires that data be on an interval or ratio scale.

Here we will discuss the Spearman and Kendall measures of correlation. First, some of the theory and examples of Spearman's measure of correlation will be provided. Then a discussion will be conducted in the use of the normal approximation method with Spearman's r, and how confidence intervals can be constructed. The next part of this chapter will summarize the theory and give examples of Kendall's measure of correlation. Likewise, the use of the normal approximation method to find confidence intervals with Kendall's $\tau_s$ will be presented. The final part of this chapter, will look at the results that can be obtained from Pearson's R, Spearman's r, and Kendall's $\tau_s$, and compare the sample sizes obtained from these three methods.

## A. SPEARMAN'S R

For this thesis, we let "r" be the notation for Spearman's coefficient of rank correlation. It is usually designated by $\rho$ but, the use of $\rho$ will cause some confusion between population correlation coefficient and this *rho*.

In general, the sampling distribution of R depends upon the bivariate population from which the sample of pairs is drawn. But, suppose that the X observations are ranked from smallest to largest using the integers 1,2,3,...,n, and the Y observations are ranked the same way. In other words, each observation is assigned a rank according to its magnitude relative to the others in its own group. Then, the data consists of n sets of paired ranks, and using these pairs, R as defined in Equation 2.5 can be calculated. The resulting statistic is called Spearman's coefficient of rank correlation (r). The difference between Pearson's R and Spearman's r is, Spearman's r measures the degree of correspondence between rankings, instead of actual variate values. However, it can still be considered a measure of correlation between X and Y in the continuous bivariate population. Let

$$R(X_i) = rank(X_i),$$

and

$$R(Y_i) = rank(Y_i).$$

Spearman's coefficient of rank correlation is

$$r = \frac{12\sum_{i=1}^{n}[R(X_i) - \overline{R(X)}][R(Y_i) - \overline{R(Y)}]}{n(n^2 - 1)} \qquad (4.1)$$

and if the data are replaced by their ranks, then $\overline{X}$ and $\overline{Y}$corresponds to $\overline{R(X)}$ and $\overline{R(Y)}$, and can be calculated as

$$\begin{aligned} \overline{R(X)} &= \frac{1}{n}\sum_{i=1}^{n}R(X_i) = \frac{1}{n}\sum_{i=1}^{n}i \\ &= \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}, \end{aligned} \qquad (4.2)$$

and in the same way

$$\overline{R(Y)} = \frac{n+1}{2}. \qquad (4.3)$$

Then, Spearman's coefficient of rank correlation becomes

$$r = \frac{12\sum_{i=1}^{n}\left[R(X_i) - \frac{n+1}{2}\right]\left[R(Y_i) - \frac{n+1}{2}\right]}{n(n^2 - 1)} \qquad (4.4)$$

[Ref. 5: p.246]. An equivalent but computationaly easier form is given by

$$r = \frac{1 - 6\sum_{i=1}^{n}[R(X_i) - R(Y_i)]^2}{n(n^2 - 1)}, \tag{4.5}$$

and if we take $T = \sum_{i=1}^{n}[R(X_i) - R(Y_i)]^2$ then, Equation 4.5 will be,

$$r = 1 - \frac{6T}{n(n^2 - 1)}. \tag{4.6}$$

It is important to note that Equation 4.5 and 4.6 are equivalent to 4.4 only if there are no ties.

If a small number of ties are present in the data, Equation 4.5 and 4.6 can be used because of the simplicity and there will be very little difference between the two coefficients obtained from 2.5 and 4.5. If there are many ties, then Pearson's R in Equation 2.5 should be used on the ranks as described below. In this manner. $\overline{X}$ corresponds to $\overline{R(X)}$ and $\overline{R(Y)}$ as explained before, and, $\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ corresponds to

$$\begin{aligned} \sum_{i=1}^{n}[R(X_i) - \overline{R(X)}]^2 &= \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 \\ &= \sum_{i=1}^{n}\left[i^2 - i(n+1) + \left(\frac{n+1}{2}\right)^2\right] \\ &= \frac{n(n+1)(1n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4} \\ &= \frac{n(n^2-1)}{12}. \end{aligned} \tag{4.7}$$

In the same way

$$\sum_{i=1}^{n}[R(Y_i) - R\overline{(Y)}]^2 = \frac{n(n^2-1)}{12}. \tag{4.8}$$

Thus Equation 2.5 becomes Equation 4.4, and this means that Pearson's R reduces to Spearman's r when the data are replaced by their ranks.

The following is an example to see the difference between Pearson's R and Spearman's r. Let's take 12 paired data like (86,88), (71,77), (77,76), (68,64), (91,96), (72,72), (77,65), (91,90), (70,65), (71,80), (88,81), (87,72). Suppose these are the math and the english scores of 12 students. The math scores of the students were ranked among themselves,

$$X_i = 68\ 70\ 71\ 71\ 72\ 77\ 77\ 86\ 87\ 88\ 91\ 91,$$

and the english scores of the students were ranked among themselves,

$$Y_i = 64\ 65\ 65\ 72\ 72\ 76\ 77\ 80\ 81\ 88\ 90\ 96.$$

There are 3 pairs of ties in X variables and 2 pairs of tie in Y variable. The pairs of ties will be given the average ranks for each pair. For example the first ties are when the X variable is 71; thus the rank will be $\frac{3+4}{2} = 3.5$. The other pairs of ties were similarly ranked and the general result is,

$$R(X_i) = 8,\ 3.5,\ 6.5,\ 1,\ 11.5,\ 5,\ 6.5,\ 11.5,\ 2,\ 3.5,\ 10,\ 9$$

and

$$R(Y_i) = 10,\ 7,\ 6,\ 1,\ 12,\ 4.5,\ 2.5,\ 11,\ 2.5,\ 8,\ 9,\ 4.5.$$

By using these values we can calculate

$$[R(x_i) - R(Y_i)]^2 = 4,\ 12.25,\ 0.25,\ 0,\ 0.25,\ 0.25,\ 16,\ 0.25,\ 0.25,\ 20.25,\ 1,\ 20.25$$

and then, calculate the statistic T in Equation 4.6 as

$$T = \sum_{i=1}^{12} [R(X_i) - R(Y_i)]^2 = 75.$$

Then r is obtained from Equation 4.6 as

$$r = 1 - \frac{6T}{n(n^2 - 1)} = 1 - \frac{6(75)}{12(143)} = 0.7378.$$

Using Equation 4.4 to calculate the r value, results in r = 0.729, and using Equations 2.13, 2.14, to calculate the Pearson's R on the ranks gives R = 0.7354. As can be seen, there is a very small differences between these values.

## B. CONFIDENCE INTERVALS FOR CORRELATION COEFFICIENT WHEN WE USE SPEARMAN'S R

If X and Y are independent and continuous then the population correlation coefficient will be equal to zero, and if this happens then the expected value of the sample correlation coefficient will essentially be zero too, because $E[R] \cong \rho$. The variance of the sample correlation coefficient will be equal to $\frac{1}{n}$, and from Equation 2.7 it is very clear that as a sample size gets bigger then variance of the sample correlation coefficient will approach zero.

To find a confidence interval for the population correlation coefficient by using Spearman's r, the statistic will be

$$Z = \frac{1}{2} \ln\left( \frac{1+r}{1-r} \right) = \tanh^{-1} r, \tag{4.9}$$

which is distributed approximately normally with expected value

$$E(Z) \cong \frac{1}{2} \ln\left( \frac{1+\rho}{1-\rho} \right), \tag{4.10}$$

and variance

$$\sigma_z^2 \cong (n-3)^{-1} \tag{4.11}$$

[Ref. 6: p.463].

Using this transformation, the confidence interval for $\rho$ can be found. Having calculated the estimate for $\rho$, namely r, we can compute Z and the statistic

$$K_2 = \left[ Z - \frac{1}{2} \ln\left( \frac{1+\rho}{1-\rho} \right) \right] \sqrt{n-3} \cong \frac{Z - E(Z)}{\sigma(Z)} \tag{4.12}$$

which is approximately normally distributed with expected value equal to 0.0 and variance equal to 1.

Using the normal approximation, there is 95% certainty that

$$-1.96 < \frac{Z - E(Z)}{\sigma} < 1.96, \tag{4.13}$$

and the 95% confidence interval of E(Z) will be

$$\begin{aligned} \frac{1}{2} \ln\left( \frac{1+r}{1-r} \right) - 1.96\sigma < E(Z) &= \frac{1}{2} \ln\left( \frac{1+\rho}{1-\rho} \right) \\ &< \frac{1}{2} \ln\left( \frac{1+r}{1-r} \right) + 1.96\sigma. \end{aligned} \tag{4.14}$$

Equation 4.10 may be used to obtain

$$\exp\left\{ 2\left( \frac{1}{2} \ln\left( \frac{1+r}{1-r} \right) - 1.96\sigma \right) \right\} < \left( \frac{1+\rho}{1-\rho} \right), \tag{4.15a}$$

and

$$\left( \frac{1+\rho}{1-\rho} \right) < \exp\left\{ 2\left( \frac{1}{2} \ln\left( \frac{1+r}{1-r} \right) + 1.96\sigma \right) \right\}. \tag{4.15b}$$

If the left side of 4.15a is $L_2$ and the right side of 4.15b is $U_2$ then

$$L_2 < \left( \frac{1+\rho}{1-\rho} \right) < U_2, \tag{4.16}$$

and from this equation the 95% confidence interval for $\rho$ will be

$$\left( \frac{L_2 - 1}{L_2 + 1} \right) < \rho < \left( \frac{U_2 - 1}{U_2 + 1} \right). \tag{4.17}$$

Spearman's r can be used to find a confidence interval for a population correlation coefficient, by using the normal approximation method. It is very important to note that when using this approximation the observations (X, Y) are independent. If these bivariate observations are independent then the measures of correlation values (Pearson's R and Spearman's r) will almost be equal. Thus, both of these methods can be used to find a confidence interval. If the observations are not independent then Spearman's r cannot be used in place of Pearson's R. Again, the largest sample size for a desired interval size that could occur will occur when r = 0, and we call this the worst case.

## C. KENDALL'S TAU

Another measure of correlation is Kendall's ($\tau_s$), which is usually considered more difficult to obtain than Spearman's r. The basic advantage of Kendall's $\tau_s$ is that its distribution approaches the normal distribution quite rapidly, so that the normal approximation is better for Kendall's $\tau_s$ than it is for Spearman's r. Another advantage of the Kendall test statistic is its direct and simple interpretation in terms of probabilities of observing concordant and discordant pairs. [Ref. 5: p.356]

For any two independent pairs of random variables ($X_i$, $Y_i$) and ($X_j$, $Y_j$), we denote by $p_c$ and $p_d$ the probabilities of concordance and discordance. Two observations, for example (2.3, 3.5) and (2.6, 1.7), are called concordant if both members of one observations are larger than their respective members of the other observation, and are called discordant otherwise. The probabilities $p_c$ and $p_d$ can be defined as

$$\begin{aligned} p_c &= P\{[(X_i < X_j) \cap (Y_i < Y_j)] \cup [(X_i > X_j) \cap (Y_i > Y_j)]\} \\ &= P[(X_j - X_i)(Y_j - Y_i) > 0] \\ &= P[(X_i < X_j) \cap (Y_i < Y_j)] + P[(X_i > X_j) \cap (Y_i > Y_j)], \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} p_d &= P[(X_j - X_i)(Y_j - Y_i) < 0] \\ &= P[(X_i < X_j) \cap (Y_i > Y_j)] + P[(X_i > X_j) \cap (Y_i < Y_j)] \end{aligned} \tag{4.19}$$

[Ref. 4: p.208].

If there is a perfect correlation between X and Y, then there is either perfect concordance or perfect discordance. The Kendall coefficient $\tau$ is defined as the difference

$$\tau = p_c - p_d. \tag{4.20}$$

If the marginal probability distributions of X and Y are continuous, so that the possibility of ties $X_i = X_j$ or $Y_i = Y_j$ within groups is eliminated, we have

$$\begin{aligned} p_c = &\{P(Y_i < Y_j) - P[(X_i > X_j) \cap (Y_i < Y_j)]\} \\ &+ \{P(Y_i > Y_j) - P[(X_i < X_j) \cap (Y_i > Y_j)]\} \end{aligned} \tag{4.21}$$

Thus,

$$\begin{aligned} p_c &= P(Y_i < Y_j) + P(Y_i > Y_j) - p_d \\ &= 1 - p_d. \end{aligned}$$

In this case, $\tau$ can be expressed as

$$\tau = 2p_c - 1 = 1 - 2p_d \tag{4.22}$$

[Ref. 4: p.208].

If X and Y are independent and continuous random variable then $p_c$ must be equal to $p_d$, and so we find $\tau = 0$. This means that for independent and continuous random variables $\tau$, will be equal to zero. In general, the converse is not true. [Ref 4: p.208]

All this explanation is about the population. However we are interested in the sample. If there are n observations then it means these n observations may be paired in $\binom{n}{2} = \frac{n(n-1)}{2}$ different ways. Suppose we compare all pairs and determine the number of concordant pairs and the number of discordant pairs. Let $c_i$ be the number of concordant pairs. Then, an unbiased estimate of $p_c$ will be

$$\hat{p}_c = \sum_{i=1}^{n} \frac{2c_i}{n(n-1)} . \qquad (4.23)$$

Now let $d_i$ be the number of discordant pairs and then

$$\hat{p}_d = \sum_{i=1}^{n} \frac{2d_i}{n(n-1)} \qquad (4.24)$$

will give an unbiased estimate of $p_d$. A measure of correlation of the sample will be

$$\tau_s = (\hat{p}_c - \hat{p}_d) \qquad (4.25)$$

[Ref. 4: p.210]. This is Kendall's sample *tau* coefficient $\tau_s$, which is an unbiased estimater of the parameter $\tau$ in any bivariate distribution. "It is important to note that the variance of $\tau_s$ approaches zero as the sample size approaches infinity." [Ref. 4: p.211]

Using the same data that we used in the Spearman example to calculate r, Kendall's $\tau_s$ will be calculated. Arrangement of the data $(X_i, Y_i)$ according to increasing values of X gives these pairs of observation: (68, 64), (70, 65), (71, 77), (71, 80), (72, 72), (77, 65), (77, 76), (86, 88), (87, 72),(88, 81), (91, 90), (91, 96). There are ties in scores 71, 77, and 91. We calculate

$$c_i = 11,\ 9,\ 4,\ 4,\ 5,\ 5,\ 4,\ 2,\ 3,\ 2,\ 0,\ 0$$

and

$$d_i = 0,\ 0,\ 4,\ 4,\ 1,\ 0,\ 1,\ 2,\ 0,\ 0,\ 0,\ 0,$$

and by using Equation 4.23 and 4.24 we find that $\hat{p}_c = \frac{2 \times 49}{12(11)}$ and $\hat{p}_d = \frac{2 \times 12}{12(11)}$. From Equation 4.25,

$$\tau_s = (\hat{p}_c - \hat{p}_d) = (0.7424 - 0.1818) = 0.5606$$

estimates a positive correlation between these variables. We already found r = 0.7378 with the same data. In general, the absolute value of Spearman's r will tend to be larger than the absolute value of Kendall's tau. As a test of significance there is no strong reason to prefer one over the other, because both usually give almost the same result. [Ref. 5: p.251]

## D. CONFIDENCE INTERVALS FOR CORRELATION COEFFICIENT WHEN WE USED KENDALL'S TAU

To find a confidence interval for the population correlation coefficient by using Kendall's $\tau_s$, the Z statistic will be

$$Z = \frac{1}{2} \ln\left( \frac{1 + \tau_s}{1 - \tau_s} \right) = \tanh^{-1}\tau_s, \tag{4.26}$$

which is approximately normally distributed with the expected value given in Equation 4.10 and variance given in Equation 4.11. [Ref. 6: p.463]

Again normalization on Z can be accomplished yielding

$$K_3 = \left[ Z - \frac{1}{2} \ln\left( \frac{1 + \rho}{1 - \rho} \right) \right] \sqrt{n - 3} \cong \frac{Z - E(Z)}{\sigma(Z)} \tag{4.27}$$

which is approximately normally distributed with expected value equal to 0 and variance equal to 1.

Using the normal approximation, there is 95% certainty that

$$-1.96 < \frac{Z - E(Z)}{\sigma(Z)} < 1.96, \tag{4.28}$$

and an 95% confidence interval of E(Z) will be

$$\begin{aligned}\frac{1}{2}\ln\left(\frac{1+\tau_s}{1-\tau_s}\right) - 1.96\sigma < E(Z) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\\ < \frac{1}{2}\ln\left(\frac{1+\tau_s}{1-\tau_s}\right) + 1.96\sigma,\end{aligned} \tag{4.29}$$

and

$$\exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+\tau_s}{1-\tau_s}\right) - 1.96\sigma\right)\right\} < \left(\frac{1+\rho}{1-\rho}\right), \tag{4.30a}$$

and

$$\left(\frac{1+\rho}{1-\rho}\right) < \exp\left\{2\left(\frac{1}{2}\ln\left(\frac{1+\tau_s}{1-\tau_s}\right) + 1.96\sigma\right)\right\}. \tag{4.30b}$$

If the left side of 4.30a is called $L_3$ and the right side of 4.30b is called $U_3$ then

$$L_3 < \left(\frac{1+\rho}{1-\rho}\right) < U_3, \tag{4.31}$$

and from this the 95% confidence interval for $\rho$ will be

$$\left(\frac{L_3 - 1}{L_3 + 1}\right) < \rho < \left(\frac{U_3 - 1}{U_3 + 1}\right). \tag{4.32}$$

Thus Kendall's $\tau_s$ in place of Pearson's R can be used to find a confidence interval for the population correlation coefficient by using the normal approximation method explained above. Again, if Kendall's $\tau_s$ is used in place of Pearson's R, the observations need to be independent. If they are not independent then the normal approximation method to find a sample size for desired confidence interval cannot be used.

## E. SAMPLE SIZE DETERMINATION FOR THE NONPARAMETRIC MEASURES

If the nonparametric measures of correlations are to be used to obtain a confidence interval, the bivariate observations must be independent. If they are not independent then the nonparametric measures of correlations cannot be used. If the variables are not independent, then the population correlation coefficient value will be different than zero. If the population correlation value is different than zero, then the standard normal approximation can not be used. The only knowledge is that if X and Y come from independent bivariate observations, then use of the normal approximation method to find a confidence interval for population correlation coefficient is valid. For this purpose, Pearson's R is used for determinating sample size when using the normal approximation method that was explained in Chapter II.

## F. THE RELATION BETWEEN PEARSON'S R SPEARMAN'S R AND KENDALL TAU

If the data are at least interval scaled with independent observations, then all three measures of correlation value can be used to find a maximum sample size for a desired confidence interval by using the normal approximation method. To see the difference between these three method, let's use the same 12 sample data pair we used in Chapter IV, Section A. Previous computations from the data resulted in r = 0.729, R = 0.7354 and $\tau_s = 0.5606$. If the confidence interval for population correlation coefficient is calculated by using R = 0.7374, the statistic will be

$$Z = \frac{1}{2} \ln\left( \frac{1.7374}{0.2626} \right) = 0.9448,$$

and standard deviation

$$\sigma(Z) = \frac{1}{\sqrt{n-3}} = 0.334.$$

The 95 percent confidence limits for E(Z) are

$$0.9448 - 1.96 \times 0.334 < E(Z) < 0.9448 + 1.96 \times 0.334,$$

which reduce to

$$0.2902 < E(Z) < 1.5995.$$

The inequalities can be written as

$$0.2902 < \frac{1}{2} \ln\left( \frac{1+\rho}{1-\rho} \right) < 1.5995,$$

and from Equation 2.15a and 2.15b

$$\exp\{2 \times (0.2902)\} < \left( \frac{1+\rho}{1-\rho} \right) < \exp\{2 \times (1.5995)\}.$$

Calculating $L_1 = 1.7868$. and $U_1 = 24.508$ and applying Equation 2.17, the confidence interval for $\rho$ will be

$$\left( \frac{1.7868 - 1}{1.7868 + 1} \right) < \rho < \left( \frac{24.508 - 1}{24.508 + 1} \right),$$

which reduces to

$$0.2823 < \rho < 0.9216.$$

So, the 95% confidence interval for $\rho$ by using Pearson's R is $0.2823 < \rho < 0.9216$, and confidence interval size (2A) is 0.6363.

If Spearman's r is used with r = 0.729, then

$$Z = \frac{1}{2} \ln\left( \frac{1.729}{0.271} \right) = 0.9266$$

and $\sigma(Z)$ is the same as with Pearson's R. The 95 percent confidence limits for E(Z) are

$$0.9266 - 1.96 \times 0.334 < E(Z) < 0.9266 + 1.96 \times 0.334$$

which reduce to

$$0.272 < E(Z) < 1.5813.$$

The inequalities can be written as

$$0.272 < \frac{1}{2} \ln\left( \frac{1+\rho}{1-\rho} \right) < 1.5813,$$

and from Equation 4.15a and 4.15b

$$\exp\{2 \times (0.272)\} < \left( \frac{1+\rho}{1-\rho} \right) < \exp\{2 \times (1.5813)\}.$$

Calculating $L_2 = 1.7229$ and $U_2 = 23.632$, the confidence interval for $\rho$ is

$$\left( \frac{1.7229 - 1}{1.7229 + 1} \right) < \rho < \left( \frac{23.632 - 1}{23.632 + 1} \right)$$

yielding

$$0.2655 < \rho < 0.9188.$$

So, the 95% confidence interval for $\rho$ by using Spearman's r is $0.2655 < \rho < 0.9188$, and confidence interval size (2A) is 0.6553.

Finally using Kendall's tau, $\tau_s = 0.5606$, gives the statistic

$$Z = \frac{1}{2} \ln\left( \frac{1.5606}{0.4394} \right) = 0.6337$$

and $\sigma(Z) = 0.334$. The 95 percent confidence limits for E(Z) are then

$$0.6337 - 1.96 \times 0.334 < E(Z) < 0.6337 + 1.96 \times 0.334$$

which reduces to

$$-0.021 < E(Z) < 1.2884.$$

The inequalities can be written as

$$-0.021 < \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) < 1.2884,$$

and from Equation 4.30a and 4.30b

$$\exp\{2 \times (-0.021)\} < \left(\frac{1+\rho}{1-\rho}\right) < \exp\{2 \times (1.2884)\}.$$

Again calculating $L_3 = 0.9589$ and $U_3 = 13.155$, the confidence interval for $\rho$ will be

$$\left(\frac{0.9589-1}{0.9589+1}\right) < \rho < \left(\frac{13.155-1}{13.155+1}\right)$$

or

$$-0.021 < \rho < 0.8587.$$

Thus the 95% confidence interval for $\rho$ using Kendall's tau is $-0.021 < \rho < 0.8587$ , and confidence interval size (2A) is 0.8797

As can be seen from these three results, Pearson's R and Spearman's r give approximately the same confidence interval size (2A). However, the confidence interval size that was obtained from Kendall's $\tau_s$ is noticeably different from the others. This seems to be a disadvantage for Kendall's tau, but Conover states that there is no strong reason to prefer one over another, because they will generally give roughly the same result. [Ref. 5: p.251]

Graphs are provided to show the difference among sample sizes from these three methods. At the same tim^ these graphs can be used to determine the appropriate sample size for a desired confidence interval. Figure 4 on page 43 shows the sample size and the confidence interval for these three methods.

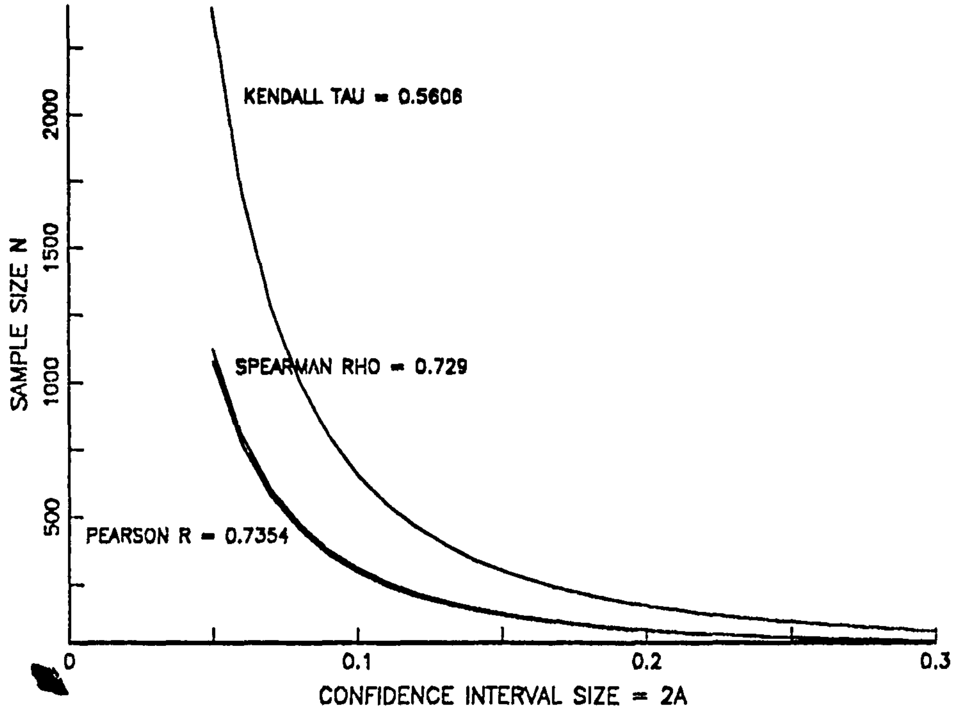


Figure 4. REQIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL SIZE BY USING DIFFERENT SAMPLE CORRELATION METHODS

A table can be developed to provide the exact value for different sample correlation coefficient methods. Table 9 shows these sample size values.

Table 9. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL SIZE BY USING DIFFERENT SAMPLE CORRELATION METHODS

| 95% Confidence Interval Size = 2A | Pearson's R = 0.7354 | Spearman's r = 0.729 | Kendall $\tau_s$ = 0.5606 |
|---|---|---|---|
| 0.05 | 1,076 | 1.120 | 2,392 |
| 0.06 | 772 | 804 | 1.714 |
| 0.07 | 581 | 605 | 1,287 |
| 0.08 | 454 | 472 | 1,002 |
| 0.09 | 364 | 379 | 804 |
| 0.10 | 299 | 311 | 659 |
| 0.11 | 250 | 260 | 550 |
| 0.12 | 212 | 221 | 466 |
| 0.13 | 183 | 190 | 400 |
| 0.14 | 159 | 165 | 347 |
| 0.15 | 140 | 145 | 304 |
| 0.16 | 124 | 129 | 269 |
| 0.17 | 111 | 115 | 239 |
| 0.18 | 100 | 104 | 214 |
| 0.19 | 90 | 94 | 193 |
| 0.20 | 82 | 85 | 175 |
| 0.21 | 75 | 78 | 160 |
| 0.22 | 69 | 72 | 146 |
| 0.23 | 64 | 66 | 134 |
| 0.24 | 59 | 61 | 124 |
| 0.25 | 55 | 57 | 114 |
| 0.26 | 51 | 53 | 106 |
| 0.27 | 48 | 50 | 99 |
| 0.28 | 45 | 46 | 92 |
| 0.29 | 42 | 44 | 86 |
| 0.30 | 40 | 41 | 81 |

This chapter explained Spearman's and Kendall's measures of correlation, and the problems of finding a sample size for a desired confidence interval size by using Spearman's r or Kendall's $\tau_{s.}$ The next chapter will summarize this study and give some suggestions for further research and study.

# V. SUMMARY AND SUGGESTION FOR FURTHER RESEARCH AND STUDY

In this chapter, a summary will be given of study of sample sizes for desired confidence intervals when the classical sample correlation coefficient method (Pearson's R), and the nonparametric statistical sample correlation coefficient methods (Spearman's r and Kendall's $\tau_s$) are used. Additionally, recommendations will be made for some additional study into the reduction of the number of observations needed to obtain a desired confidence interval for the correlation coefficient.

## A. SUMMARY

This study described the classical sample correlation coefficient (Pearson R) and the nonparametric statistical sample correlation coefficient methods (Spearman's r and Kendall's $\tau_s$) to obtain the number of samples needed to obtain a desired confidence interval size for a correlation coefficient.

First, a description was provided of the Pearson product-moment correlation coefficient, the estimated population correlation coefficient, and confidence intervals for the population correlation coefficient by the using the normal approximation method. In the next chapter, it was shown how the sample size for estimating a correlation coefficient using the confidence interval could be obtained, and a comparison was done of these results for different sample correlation coefficient values. The result had the limitation that one must guess at the sample result before taking the sample, but it was still possible to give general results about the magnitude of needed sample sizes. In Chapter IV, the Spearman and Kendall statistical sample correlation coefficient methods were described. Analysis concluded showing that there is no way to find a sample size by using the Spearman and Kendall statistical sample correlation coefficient method when *rho* is not equal to zero, due to the absence of any information about the cumulative distribution function when the population correlation coefficient is nonzero. Similarly, values for probabilities, expected values, and variances could not be determined.

However, most of the time the value of population correlation coefficient is unknown. If the observations are independent, then a sample size for a desired confidence interval using nonparametric measures of correlation can to found. If the observations are not independent then the normal approximation method can not be used for nonparametric statistics to find a needed sample size, and instead Pearson's R must be used to find a sample size.

To use the normal approximation method, the decision maker must estimate the measure of correlation value, and then determine the desired sample size for a confidence interval of size (2A). In order to find a sample for the desired confidence interval, the sample correlation coefficient must first be estimated without any data.

The results for different sample correlation coefficient values were compared. and it was observed that if the sample correlation coefficient value gets bigger in absolute value then the sample size gets smaller, and the largest sample size that could result will occur when R equals zero.

Computer programs were developed to calculate sample sizes for a desired confidence interval for different sample correlation coefficient values, and some tables and graphs giving the sample size needed for different R values were generated. These tables and graphs can be used by a decision maker to assist in determining the desired sample size.

## B. SUGGESTIONS FOR FURTHER STUDY

In this study, 95% confidence intervals were used. It would be useful if tables and graphs are developed for other confidence interval sizes, such as 90%, 97.5% and 99%.

The discussion about nonparametric statistics in this study centered on the: Spearman and Kendall test statistics. It was not concluded that these methods needed smaller sample sizes than the classical, Pearson's method. Additional research could be done searching for appropriate sample sizes for other nonparametric statistics.

It is sincerely hoped that the information about sample size needed to estimate correlation coefficients, and the tables, graphs and computer programs in this thesis be beneficial to decision makers in deciding the sample size for a desired confidence interval, when estimating a correlation coefficient.

# APPENDIX A. TABLE FOR CONFIDENCE BELTS FOR THE CORRELATION COEFFICIENT

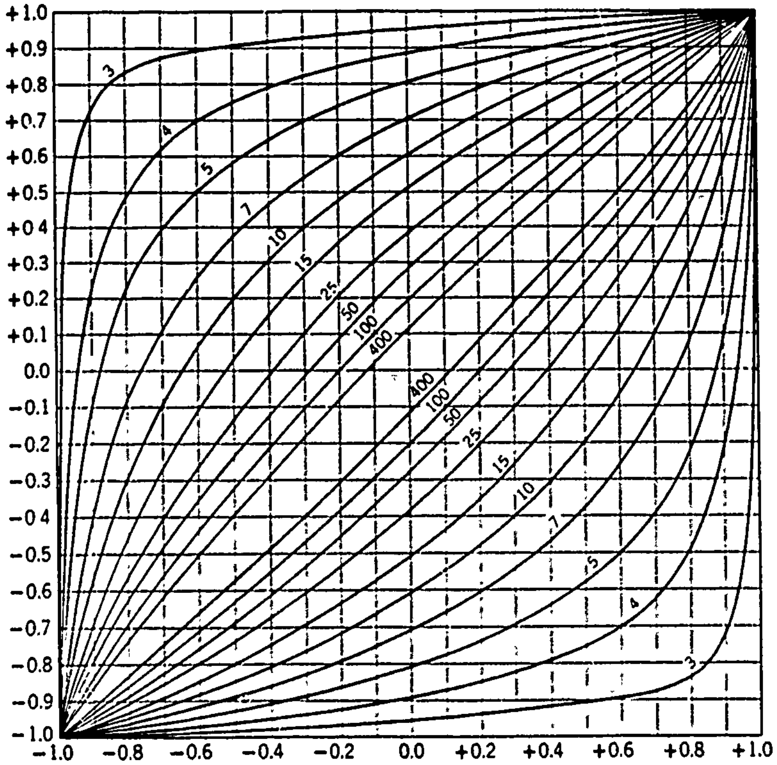


Figure 5. 95% CONFIDENCE BELTS FOR THE CORRELATION COEFFICIENT: The Vertical axis of this figure shows $\rho$, the Horizantal axis shows R.

[Ref. 6: p.545].

## APPENDIX B. THE APL PROGRAM "TEZ" USED TO COMPUTE CONFIDENCE INTERVAL FOR DESIRED SAMPLE CORRELATION COEFFICIENT VALUE

```
      ∇ TEZ R
[1]    ⍝ THIS PROGRAM COMPUTES THE CONFIDENCE INTERVAL WITH
[2]    ⍝ ESTIMATED SAMPLE CORRELATION COEFFICIENT VALUE FOR
[3]    ⍝ DIFFERENT SAMPLE SIZE. TO RUN THE PROGRAM, ENTER
[4]    ⍝ DESIRED CORRELATION COEFFICIENT. IT TERMINATES THE
[5]    ⍝ EXECUTION WHEN THE SAMPLE SIZE IS > 200. FOR BIGGER
[6]    ⍝ NUMBERS, THE VALUE OF N IN LINE 29 MUST BE INCREASED
[7]    ⍝ TO DESIRED SAMPLE SIZE. IF CONFIDENCE LEVEL DIFFERENT
[8]    ⍝ THAN 95 PERCENT, THEN THE STANDARD PROBABILITY VALUE
[9]    ⍝ MUST BE CHANGED IN LINE 15 AND 16. IT IS IMPORTANT
[10]   ⍝ TO NOTE THAT N CAN NOT BE LESS THAN 4.
[11]    Z←(1÷2)×(⍟((1+R)÷(1-R)))
[12]    N←4
[13]    'SAMPLE CORRELATION COEFFICIENTIS = ', 5 3 ⍕R
[14]    '-------------------------------------'
[15]    ''
[16]   L1:SIGMA←1÷((N-3)*(1÷2))
[17]   ⍝ 95 PERCENT C.I. FOR E(Z) ARE
[18]    Z1←Z-(1.96×SIGMA)
[19]    Z2←Z+(1.96×SIGMA)
[20]    A←Z1×2
[21]    A1  *A
[22]    LOW←(A1-1)÷(A1+1)
[23]    B←Z2×2
[24]    B1←*B
[25]    UPPER←(B1-1)÷(B1+1)
[26]    A2←UPPER-LOW
[27]    'FOR SAMPLE SIZE                = ', 5 0 ⍕N
[28]    'CONFIDENCE INTERVAL IS         = ', 4 2 ⍕LOW,UPPER
```

```
[29]  'CONFIDENCE INTERVAL SIZE = 2A IS = ', 4 2 ⍕A2
[30]  ''
[31]  N←N+1
[32]  →(N≤200)/L1
[33] ⍝
      ∇
```

# APPENDIX C. TABLES FOR DESIRED SAMPLE SIZE USING DIFFERENT ESTIMATED SAMPLE CORRELATION COEFFICIENT VALUES AND A 95% CONFIDENCE LEVEL

Table 10. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.95

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.95 | 57 | 0.92 | 0.97 | 0.05 |
| 0.95 | 42 | 0.91 | 0.97 | 0.06 |
| 0.95 | 33 | 0.90 | 0.98 | 0.07 |
| 0.95 | 28 | 0.89 | 0.98 | 0.08 |
| 0.95 | 23 | 0.88 | 0.98 | 0.09 |
| 0.95 | 20 | 0.88 | 0.98 | 0.10 |
| 0.95 | 18 | 0.87 | 0.98 | 0.11 |
| 0.95 | 17 | 0.86 | 0.98 | 0.12 |
| 0.95 | 15 | 0.85 | 0.98 | 0.13 |
| 0.95 | 14 | 0.85 | 0.98 | 0.14 |
| 0.95 | 13 | 0.84 | 0.99 | 0.15 |
| 0.95 | 12 | 0.83 | 0.99 | 0.16 |
| 0.95 | 11 | 0.81 | 0.99 | 0.17 |
| 0.95 | 10 | 0.80 | 0.99 | 0.19 |
| 0.95 | 9 | 0.77 | 0.99 | 0.22 |
| 0.95 | 8 | 0.74 | 0.99 | 0.25 |
| 0.95 | 7 | 0.69 | 0.99 | 0.30 |

Table 11. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.925

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.925 | 113 | 0.90 | 0.95 | 0.05 |
| 0.925 | 83 | 0.89 | 0.95 | 0.06 |
| 0.925 | 64 | 0.88 | 0.95 | 0.07 |
| 0.925 | 52 | 0.87 | 0.96 | 0.08 |
| 0.925 | 43 | 0.86 | 0.96 | 0.09 |
| 0.925 | 36 | 0.86 | 0.96 | 0.10 |
| 0.925 | 31 | 0.85 | 0.96 | 0.11 |
| 0.925 | 28 | 0.84 | 0.97 | 0.12 |
| 0.925 | 25 | 0.84 | 0.97 | 0.13 |
| 0.925 | 22 | 0.83 | 0.97 | 0.14 |
| 0.925 | 20 | 0.82 | 0.97 | 0.15 |
| 0.925 | 19 | 0.81 | 0.97 | 0.16 |
| 0.925 | 17 | 0.80 | 0.97 | 0.17 |
| 0.925 | 16 | 0.79 | 0.97 | 0.18 |
| 0.925 | 15 | 0.78 | o.98 | 0.19 |
| 0.925 | 14 | 0.77 | 0.98 | 0.20 |
| 0.925 | 13 | 0.76 | 0.98 | 0.21 |
| 0.925 | 12 | 0.75 | 0.98 | 0.23 |
| 0.925 | 11 | 0.73 | 0.98 | 0.25 |
| 0.925 | 10 | 0.71 | 0.98 | 0.28 |
| 0.925 | 9 | 0.68 | 0.98 | 0.30 |

Table 12. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.85**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.85 | 398 | 0.82 | 0.87 | 0.05 |
| 0.85 | 287 | 0.82 | 0.88 | 0.06 |
| 0.85 | 217 | 0.81 | 0.88 | 0.07 |
| 0.85 | 170 | 0.80 | 0.89 | 0.08 |
| 0.85 | 138 | 0.79 | 0.89 | 0.09 |
| 0.85 | 114 | 0.79 | 0.89 | 0.10 |
| 0.85 | 96 | 0.79 | 0.90 | 0.11 |
| 0.85 | 82 | 0.78 | 0.90 | 0.12 |
| 0.85 | 71 | 0.77 | 0.90 | 0.13 |
| 0.85 | 63 | 0.77 | 0.91 | 0.14 |
| 0.85 | 56 | 0.76 | 0.91 | 0.15 |
| 0.85 | 50 | 0.75 | 0.91 | 0.16 |
| 0.85 | 45 | 0.74 | 0.91 | 0.17 |
| 0.85 | 41 | 0.74 | 0.92 | 0.18 |
| 0.85 | 37 | 0.73 | o.92 | 0.19 |
| 0.85 | 34 | 0.72 | 0.92 | 0.20 |
| 0.85 | 32 | 0.71 | 0.92 | 0.21 |
| 0.85 | 30 | 0.71 | 0.93 | 0.22 |
| 0.85 | 28 | 0.70 | 0.93 | 0.23 |
| 0.85 | 26 | 0.69 | 0.93 | 0.24 |
| 0.85 | 24 | 0.68 | 0.93 | 0.25 |
| 0.85 | 23 | 0.67 | 0.93 | 0.26 |
| 0.85 | 22 | 0.67 | 0.94 | 0.27 |
| 0.85 | 21 | 0.66 | 0.94 | 0.28 |
| 0.85 | 20 | 0.94 | 0.73 | 0.29 |
| 0.85 | 19 | 0.64 | 0.94 | 0.30 |

Table 13. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.70

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.70 | 1326 | 0.67 | 0.72 | 0.05 |
| 0.70 | 745 | 0.67 | 0.73 | 0.06 |
| 0.70 | 715 | 0.66 | 0.73 | 0.07 |
| 0.70 | 558 | 0.66 | 0.74 | 0.08 |
| 0.70 | 448 | 0.65 | 0.74 | 0.09 |
| 0.70 | 367 | 0.65 | 0.75 | 0.10 |
| 0.70 | 307 | 0.64 | 0.75 | 0.11 |
| 0.70 | 260 | 0.64 | 0.76 | 0.12 |
| 0.70 | 224 | 0.63 | 0.76 | 0.13 |
| 0.70 | 195 | 0.62 | 0.77 | 0.14 |
| 0.70 | 171 | 0.62 | 0.77 | 0.15 |
| 0.70 | 151 | 0.61 | 0.77 | 0.16 |
| 0.70 | 135 | 0.60 | 0.77 | 0.17 |
| 0.70 | 121 | 0.60 | 0.78 | 0.18 |
| 0.70 | 110 | 0.59 | o.78 | 0.19 |
| 0.70 | 100 | 0.59 | 0.79 | 0.20 |
| 0.70 | 91 | 0.58 | 0.79 | 0.21 |
| 0.70 | 84 | 0.57 | 0.79 | 0.22 |
| 0.70 | 77 | 0.57 | 0.80 | 0.23 |
| 0.70 | 71 | 0.56 | 0.80 | 0.24 |
| 0.70 | 66 | 0.55 | 0.81 | 0.25 |
| 0.70 | 62 | 0.55 | 0.81 | 0.26 |
| 0.70 | 57 | 0.54 | 0.81 | 0.27 |
| 0.70 | 54 | 0.53 | 0.82 | 0.28 |
| 0.70 | 51 | 0.53 | 0.82 | 0.29 |
| 0.70 | 48 | 0.52 | 0.82 | 0.30 |

Table 14. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.65**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.65 | 1,698 | 0.62 | 0.67 | 0.05 |
| 0.65 | 1,217 | 0.62 | 0.68 | 0.06 |
| 0.65 | 915 | 0.61 | 0.68 | 0.07 |
| 0.65 | 712 | 0.61 | 0.69 | 0.08 |
| 0.65 | 572 | 0.60 | 0.69 | 0.09 |
| 0.65 | 469 | 0.60 | 0.70 | 0.10 |
| 0.65 | 392 | 0.59 | 0.70 | 0.11 |
| 0.65 | 332 | 0.59 | 0.71 | 0.12 |
| 0.65 | 285 | 0.58 | 0.71 | 0.13 |
| 0.65 | 248 | 0.57 | 0.71 | 0.14 |
| 0.65 | 217 | 0.57 | 0.72 | 0.15 |
| 0.65 | 192 | 0.56 | 0.72 | 0.16 |
| 0.65 | 172 | 0.56 | 0.73 | 0.17 |
| 0.65 | 154 | 0.55 | 0.73 | 0.18 |
| 0.65 | 139 | 0.54 | 0.73 | 0.19 |
| 0.65 | 126 | 0.54 | 0.74 | 0.20 |
| 0.65 | 115 | 0.53 | 0.74 | 0.21 |
| 0.65 | 105 | 0.53 | 0.75 | 0.22 |
| 0.65 | 97 | 0.52 | 0.75 | 0.23 |
| 0.65 | 90 | 0.51 | 0.75 | 0.24 |
| 0.65 | 83 | 0.83 | 0.51 | 0.76 |
| 0.65 | 77 | 0.50 | 0.76 | 0.26 |
| 0.65 | 72 | 0.49 | 0.76 | 0.27 |
| 0.65 | 67 | 0.49 | 0.77 | 0.28 |
| 0.65 | 63 | 0.48 | 0.77 | 0.29 |
| 0.65 | 59 | 0.47 | 0.77 | 0.30 |

Table 15. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.60

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.60 | 2,082 | 0.57 | 0.62 | 0.05 |
| 0.60 | 1,495 | 0.57 | 0.63 | 0.06 |
| 0.60 | 1,123 | 0.56 | 0.63 | 0.07 |
| 0.60 | 875 | 0.56 | 0.64 | 0.08 |
| 0.60 | 701 | 0.55 | 0.64 | 0.09 |
| 0.60 | 575 | 0.55 | 0.65 | 0.10 |
| 0.60 | 480 | 0.54 | 0.65 | 0.11 |
| 0.60 | 407 | 0.54 | 0.66 | 0.12 |
| 0.60 | 349 | 0.53 | 0.66 | 0.13 |
| 0.60 | 303 | 0.53 | 0.67 | 0.14 |
| 0.60 | 266 | 0.52 | 0.67 | 0.15 |
| 0.60 | 235 | 0.51 | 0.67 | 0.16 |
| 0.60 | 209 | 0.51 | 0.68 | 0.17 |
| 0.60 | 190 | 0.50 | 0.68 | 0.18 |
| 0.60 | 169 | 0.50 | o.69 | 0.19 |
| 0.60 | 153 | 0.49 | 0.69 | 0.20 |
| 0.60 | 140 | 0.48 | 0.69 | 0.21 |
| 0.60 | 128 | 0.48 | 0.70 | 0.22 |
| 0.60 | 118 | 0.47 | 0.70 | 0.23 |
| 0.60 | 109 | 0.47 | 0.71 | 0.24 |
| 0.60 | 100 | 0.46 | 0.71 | 0.25 |
| 0.60 | 93 | 0.45 | 0.71 | 0.26 |
| 0.60 | 87 | 0.45 | 0.72 | 0.27 |
| 0.60 | 81 | 0.44 | 0.72 | 0.28 |
| 0.60 | 76 | 0.44 | 0.73 | 0.29 |
| 0.60 | 71 | 0.43 | 0.73 | 0.30 |

Table 16. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.55

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.55 | 2,475 | 0.52 | 0.57 | 0.05 |
| 0.55 | 1,773 | 0.52 | 0.58 | 0.06 |
| 0.55 | 1,332 | 0.51 | 0.58 | 0.07 |
| 0.55 | 1,038 | 0.51 | 0.59 | 0.08 |
| 0.55 | 832 | 0.50 | 0.59 | 0.09 |
| 0.55 | 681 | 0.50 | 0.60 | 0.10 |
| 0.55 | 569 | 0.49 | 0.60 | 0.11 |
| 0.55 | 482 | 0.48 | 0.60 | 0.12 |
| 0.55 | 413 | 0.48 | 0.61 | 0.13 |
| 0.55 | 359 | 0.47 | 0.61 | 0.14 |
| 0.55 | 314 | 0.47 | 0.62 | 0.15 |
| 0.55 | 278 | 0.46 | 0.62 | 0.16 |
| 0.55 | 247 | 0.46 | 0.63 | 0.17 |
| 0.55 | 222 | 0.45 | 0.63 | 0.18 |
| 0.55 | 200 | 0.45 | 0.64 | 0.19 |
| 0.55 | 181 | 0.44 | 0.64 | 0.20 |
| 0.55 | 165 | 0.44 | 0.65 | 0.21 |
| 0.55 | 151 | 0.43 | 0.65 | 0.22 |
| 0.55 | 139 | 0.42 | 0.65 | 0.23 |
| 0.55 | 128 | 0.42 | 0.66 | 0.24 |
| 0.55 | 118 | 0.41 | 0.66 | 0.25 |
| 0.55 | 110 | 0.41 | 0.67 | 0.26 |
| 0.55 | 102 | 0.40 | 0.67 | 0.27 |
| 0.55 | 95 | 0.39 | 0.67 | 0.28 |
| 0.55 | 89 | 0.39 | 0.68 | 0.29 |
| 0.55 | 84 | 0.38 | 0.68 | 0.30 |

Table 17. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.50

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.50 | 2860 | 0.47 | 0.53 | 0.05 |
| 0.50 | 2049 | 0.47 | 0.53 | 0.06 |
| 0.50 | 1542 | 0.46 | 0.53 | 0.07 |
| 0.50 | 1199 | 0.46 | 0.54 | 0.08 |
| 0.50 | 961 | 0.46 | 0.55 | 0.09 |
| 0.50 | 787 | 0.45 | 0.55 | 0.10 |
| 0.50 | 657 | 0.44 | 0.56 | 0.11 |
| 0.50 | 556 | 0.44 | 0.56 | 0.12 |
| 0.50 | 477 | 0.43 | 0.56 | 0.13 |
| 0.50 | 414 | 0.42 | 0.56 | 0.14 |
| 0.50 | 363 | 0.42 | 0.57 | 0.15 |
| 0.50 | 320 | 0.41 | 0.58 | 0.16 |
| 0.50 | 285 | 0.41 | 0.58 | 0.17 |
| 0.50 | 255 | 0.40 | 0.58 | 0.18 |
| 0.50 | 230 | 0.40 | o.59 | 0.19 |
| 0.50 | 209 | 0.40 | 0.60 | 0.20 |
| 0.50 | 190 | 0.39 | 0.60 | 0.21 |
| 0.50 | 174 | 0.38 | 0.60 | 0.22 |
| 0.50 | 159 | 0.38 | 0.61 | 0.23 |
| 0.50 | 147 | 0.37 | 0.61 | 0.24 |
| 0.50 | 136 | 0.36 | 0.61 | 0.25 |
| 0.50 | 126 | 0.36 | 0.62 | 0.26 |
| 0.50 | 117 | 0.35 | 0.62 | 0.27 |
| 0.50 | 109 | 0.34 | 0.63 | 0.28 |
| 0.50 | 102 | 0.34 | 0.63 | 0.29 |
| 0.50 | 96 | 0.33 | 0.63 | 0.30 |

Table 18. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.45

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.45 | 3,237 | 0.42 | 0.47 | 0.05 |
| 0.45 | 2,316 | 0.42 | 0.48 | 0.06 |
| 0.45 | 1,738 | 0.41 | 0.48 | 0.07 |
| 0.45 | 1,356 | 0.41 | 0.49 | 0.08 |
| 0.45 | 1,086 | 0.40 | 0.49 | 0.09 |
| 0.45 | 889 | 0.40 | 0.50 | 0.10 |
| 0.45 | 744 | 0.39 | 0.50 | 0.11 |
| 0.45 | 628 | 0.39 | 0.51 | 0.12 |
| 0.45 | 539 | 0.38 | 0.51 | 0.13 |
| 0.45 | 467 | 0.37 | 0.51 | 0.14 |
| 0.45 | 409 | 0.37 | 0.52 | 0.15 |
| 0.45 | 361 | 0.36 | 0.42 | 0.16 |
| 0.45 | 322 | 0.36 | 0.53 | 0.17 |
| 0.45 | 288 | 0.35 | 0.53 | 0.18 |
| 0.45 | 260 | 0.35 | 0.54 | 0.19 |
| 0.45 | 235 | 0.34 | 0.54 | 0.20 |
| 0.45 | 214 | 0.34 | 0.55 | 0.21 |
| 0.45 | 196 | 0.33 | 0.55 | 0.22 |
| 0.45 | 179 | 0.33 | 0.56 | 0.23 |
| 0.45 | 165 | 0.32 | 0.56 | 0.24 |
| 0.45 | 153 | 0.31 | 0.56 | 0.25 |
| 0.45 | 142 | 0.31 | 0.57 | 0.26 |
| 0.45 | 132 | 0.30 | 0.57 | 0.27 |
| 0.45 | 123 | 0.30 | 0.58 | 0.28 |
| 0.45 | 115 | 0.29 | 0.58 | 0.29 |
| 0.45 | 108 | 0.29 | 0.59 | 0.30 |

Table 19. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.40**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.40 | 3.587 | 0.37 | 0.42 | 0.05 |
| 0.40 | 2,568 | 0.37 | 0.43 | 0.06 |
| 0.40 | 1,931 | 0.36 | 0.43 | 0.07 |
| 0.40 | 1,504 | 0.36 | 0.44 | 0.08 |
| 0.40 | 1,205 | 0.35 | 0.44 | 0.09 |
| 0.40 | 985 | 0.35 | 0.45 | 0.10 |
| 0.40 | 822 | 0.34 | 0.45 | 0.11 |
| 0.40 | 694 | 0.34 | 0.46 | 0.12 |
| 0.40 | 597 | 0.33 | 0.46 | 0.13 |
| 0.40 | 518 | 0.33 | 0.47 | 0.14 |
| 0.40 | 453 | 0.32 | 0.47 | 0.15 |
| 0.40 | 400 | 0.32 | 0.48 | 0.16 |
| 0.40 | 356 | 0.31 | 0.48 | 0.17 |
| 0.40 | 319 | 0.31 | 0.49 | 0.18 |
| 0.40 | 287 | 0.30 | o.49 | 0.19 |
| 0.40 | 260 | 0.30 | 0.50 | 0.20 |
| 0.40 | 237 | 0.29 | 0.50 | 0.21 |
| 0.40 | 216 | 0.28 | 0.50 | 0.22 |
| 0.40 | 198 | 0.28 | 0.51 | 0.23 |
| 0.40 | 183 | 0.27 | 0.51 | 0.24 |
| 0.40 | 169 | 0.27 | 0.52 | 0.25 |
| 0.40 | 157 | 0.26 | 0.52 | 0.26 |
| 0.40 | 146 | 0.26 | 0.53 | 0.27 |
| 0.40 | 136 | 0.25 | 0.53 | 0.28 |
| 0.40 | 127 | 0.25 | 0.54 | 0.29 |
| 0.40 | 119 | 0.24 | 0.54 | 0.30 |

Table 20. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.35

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.35 | 3,913 | 0.32 | 0.37 | 0.05 |
| 0.35 | 2,802 | 0.32 | 0.38 | 0.06 |
| 0.35 | 2,107 | 0.31 | 0.38 | 0.07 |
| 0.35 | 1,640 | 0.31 | 0.39 | 0.08 |
| 0.35 | 1,313 | 0.30 | 0.39 | 0.09 |
| 0.35 | 1,075 | 0.30 | 0.40 | 0.10 |
| 0.35 | 897 | 0.29 | 0.40 | 0.11 |
| 0.35 | 759 | 0.29 | 0.41 | 0.12 |
| 0.35 | 651 | 0.28 | 0.41 | 0.13 |
| 0.35 | 565 | 0.28 | 0.42 | 0.14 |
| 0.35 | 494 | 0.27 | 0.42 | 0.15 |
| 0.35 | 436 | 0.27 | 0.43 | 0.16 |
| 0.35 | 388 | 0.26 | 0.43 | 0.17 |
| 0.35 | 348 | 0.26 | 0.44 | 0.18 |
| 0.35 | 313 | 0.25 | 0.44 | 0.19 |
| 0.35 | 283 | 0.24 | 0.44 | 0.20 |
| 0.35 | 256 | 0.24 | 0.45 | 0.21 |
| 0.35 | 236 | 0.23 | 0.45 | 0.22 |
| 0.35 | 216 | 0.23 | 0.46 | 0.23 |
| 0.35 | 199 | 0.22 | 0.46 | 0.24 |
| 0.35 | 184 | 0.22 | 0.47 | 0.25 |
| 0.35 | 170 | 0.21 | 0.47 | 0.26 |
| 0.35 | 158 | 0.21 | 0.48 | 0.27 |
| 0.35 | 148 | 0.20 | 0.48 | 0.28 |
| 0.35 | 138 | 0.20 | 0.49 | 0.29 |
| 0.35 | 129 | 0.19 | 0.49 | 0.30 |

Table 21. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.30**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.30 | 4,203 | 0.27 | 0.32 | 0.05 |
| 0.30 | 3.015 | 0.27 | 0.33 | 0.06 |
| 0.30 | 2.263 | 0.26 | 0.33 | 0.07 |
| 0.30 | 1,746 | 0.26 | 0.34 | 0.08 |
| 0.30 | 1,413 | 0.25 | 0.34 | 0.09 |
| 0.30 | 1.156 | 0.25 | 0.35 | 0.10 |
| 0.30 | 965 | 0.24 | 0.35 | 0.11 |
| 0.30 | 816 | 0.24 | 0.36 | 0.12 |
| 0.30 | 700 | 0.23 | 0.36 | 0.13 |
| 0.30 | 607 | 0.23 | 0.37 | 0.14 |
| 0.30 | 531 | 0.22 | 0.37 | 0.15 |
| 0.30 | 469 | 0.22 | 0.38 | 0.16 |
| 0.30 | 417 | 0.21 | 0.38 | 0.17 |
| 0.30 | 373 | 0.21 | 0.39 | 0.18 |
| 0.30 | 336 | 0.20 | o.39 | 0.19 |
| 0.30 | 304 | 0.20 | 0.40 | 0.20 |
| 0.30 | 277 | 0.19 | 0.40 | 0.21 |
| 0.30 | 253 | 0.19 | 0.41 | 0.22 |
| 0.30 | 232 | 0.18 | 0.41 | 0.23 |
| 0.30 | 214 | 0.18 | 0.42 | 0.24 |
| 0.30 | 197 | 0.17 | 0.42 | 0.25 |
| 0.30 | 183 | 0.16 | 0.42 | 0.26 |
| 0.30 | 170 | 0.16 | 0.43 | 0.27 |
| 0.30 | 158 | 0.15 | 0.43 | 0.28 |
| 0.30 | 148 | 0.15 | 0.44 | 0.29 |
| 0.30 | 138 | 0.14 | 0.44 | 0.30 |

Table 22. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.25

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.25 | 4,466 | 0.22 | 0.27 | 0.05 |
| 0.25 | 3.198 | 0.22 | 0.28 | 0.06 |
| 0.25 | 2,402 | 0.21 | 0.29 | 0.07 |
| 0.25 | 1,871 | 0.21 | 0.29 | 0.08 |
| 0.25 | 1,498 | 0.20 | 0.29 | 0.09 |
| 0.25 | 1,226 | 0.20 | 0.30 | 0.10 |
| 0.25 | 1,023 | 0.19 | 0.30 | 0.11 |
| 0.25 | 866 | 0.19 | 0.31 | 0.12 |
| 0.25 | 742 | 0.18 | 0.31 | 0.13 |
| 0.25 | 644 | 0.18 | 0.32 | 0.14 |
| 0.25 | 564 | 0.17 | 0.32 | 0.15 |
| 0.25 | 497 | 0.17 | 0.33 | 0.16 |
| 0.25 | 442 | 0.16 | 0.33 | 0.17 |
| 0.25 | 396 | 0.16 | 0.34 | 0.18 |
| 0.25 | 357 | 0.15 | 0.34 | 0.19 |
| 0.25 | 323 | 0.15 | 0.35 | 0.20 |
| 0.25 | 294 | 0.14 | 0.35 | 0.21 |
| 0.25 | 268 | 0.14 | 0.36 | 0.22 |
| 0.25 | 246 | 0.13 | 0.36 | 0.23 |
| 0.25 | 226 | 0.13 | 0.37 | 0.24 |
| 0.25 | 209 | 0.12 | 0.37 | 0.25 |
| 0.25 | 194 | 0.12 | 0.38 | 0.26 |
| 0.25 | 181 | 0.11 | 0.38 | 0.27 |
| 0.25 | 168 | 0.11 | 0.39 | 0.28 |
| 0.25 | 157 | 0.10 | 0.39 | 0.29 |
| 0.25 | 147 | 0.09 | 0.39 | 0.30 |

Table 23. **REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.20**

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.20 | 4,784 | 0.17 | 0.22 | 0.05 |
| 0.20 | 3,615 | 0.17 | 0.23 | 0.06 |
| 0.20 | 2,522 | 0.16 | 0.23 | 0.07 |
| 0.20 | 1,963 | 0.16 | 0.24 | 0.08 |
| 0.20 | 1,574 | 0.15 | 0.24 | 0.09 |
| 0.20 | 1,288 | 0.15 | 0.25 | 0.10 |
| 0.20 | 1,076 | 0.14 | 0.25 | 0.11 |
| 0.20 | 908 | 0.14 | 0.26 | 0.12 |
| 0.20 | 778 | 0.13 | 0.26 | 0.13 |
| 0.20 | 675 | 0.13 | 0.27 | 0.14 |
| 0.20 | 591 | 0.12 | 0.27 | 0.15 |
| 0.20 | 521 | 0.12 | 0.28 | 0.16 |
| 0.20 | 464 | 0.11 | 0.28 | 0.17 |
| 0.20 | 415 | 0.11 | 0.29 | 0.18 |
| 0.20 | 374 | 0.10 | o.29 | 0.19 |
| 0.20 | 338 | 0.10 | 0.30 | 0.20 |
| 0.20 | 308 | 0.09 | 0.30 | 0.21 |
| 0.20 | 281 | 0.09 | 0.31 | 0.22 |
| 0.20 | 258 | 0.08 | 0.31 | 0.23 |
| 0.20 | 237 | 0.08 | 0.32 | 0.24 |
| 0.20 | 219 | 0.07 | 0.32 | 0.25 |
| 0.20 | 203 | 0.07 | 0.33 | 0.26 |
| 0.20 | 139 | 0.16 | 0.33 | 0.27 |
| 0.20 | 176 | 0.06 | 0.34 | 0.28 |
| 0.20 | 164 | 0.15 | 0.34 | 0.29 |
| 0.20 | 153 | 0.05 | 0.35 | 0.30 |

Table 24. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.15

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.15 | 4,855 | 0.12 | 0.17 | 0.05 |
| 0.15 | 3,476 | 0.12 | 0.18 | 0.06 |
| 0.15 | 2,610 | 0.11 | 0.18 | 0.07 |
| 0.15 | 2,032 | 0.11 | 0.19 | 0.08 |
| 0.15 | 1,631 | 0.10 | 0.19 | 0.09 |
| 0.15 | 1,333 | 0.10 | 0.20 | 0.10 |
| 0.15 | 1,111 | 0.09 | 0.20 | 0.11 |
| 0.15 | 946 | 0.09 | 0.21 | 0.12 |
| 0.15 | 807 | 0.08 | 0.21 | 0.13 |
| 0.15 | 699 | 0.08 | 0.22 | 0.14 |
| 0.15 | 612 | 0.07 | 0.22 | 0.15 |
| 0.15 | 540 | 0.07 | 0.23 | 0.16 |
| 0.15 | 481 | 0.06 | 0.23 | 0.17 |
| 0.15 | 430 | 0.06 | 0.24 | 0.18 |
| 0.15 | 387 | 0.05 | 0.24 | 0.19 |
| 0.15 | 350 | 0.05 | 0.25 | 0.20 |
| 0.15 | 319 | 0.04 | 0.25 | 0.21 |
| 0.15 | 291 | 0.04 | 0.26 | 0.22 |
| 0.15 | 267 | 0.03 | 0.26 | 0.23 |
| 0.15 | 246 | 0.03 | 0.27 | 0.24 |
| 0.15 | 227 | 0.02 | 0.27 | 0.25 |
| 0.15 | 210 | 0.02 | 0.28 | 0.26 |
| 0.15 | 195 | 0.01 | 0.28 | 0.27 |
| 0.15 | 182 | 0.01 | 0.29 | 0.28 |
| 0.15 | 170 | 0.00 | 0.29 | 0.29 |
| 0.15 | 159 | 0.00 | 0.30 | 0.30 |

Table 25. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.05

| Estimated Correlation Coefficient Value | Sample Size = n | Lower Confidence Limits | Upper Confidence Limits | 95% Confidence Interval Size = 2A |
|---|---|---|---|---|
| 0.05 | 5,072 | 0.02 | 0.07 | 0.05 |
| 0.05 | 3.627 | 0.02 | 0.08 | 0.06 |
| 0.05 | 2,719 | 0.01 | 0.08 | 0.07 |
| 0.05 | 2,122 | 0.01 | 0.09 | 0.08 |
| 0.05 | 1,696 | 0.00 | 0.09 | 0.09 |
| 0.05 | 1,388 | 0.00 | 0.10 | 0.10 |
| 0.05 | 1,157 | -0.01 | 0.10 | 0.11 |
| 0.05 | 979 | -0.01 | 0.11 | 0.12 |
| 0.05 | 840 | -0.02 | 0.11 | 0.13 |
| 0.05 | 728 | -0.02 | 0.12 | 0.14 |
| 0.05 | 638 | -0.03 | 0.12 | 0.15 |
| 0.05 | 563 | -0.03 | 0.13 | 0.16 |
| 0.05 | 500 | -0.04 | 0.13 | 0.17 |
| 0.05 | 448 | -0.04 | 0.14 | 0.18 |
| 0.05 | 403 | -0.05 | 0.14 | 0.19 |
| 0.05 | 365 | -0.05 | 0.15 | 0.20 |
| 0.05 | 332 | -0.06 | 0.15 | 0.21 |
| 0.05 | 303 | -0.06 | 0.16 | 0.22 |
| 0.05 | 278 | -0.07 | 0.16 | 0.23 |
| 0.05 | 256 | -0.07 | 0.17 | 0.24 |
| 0.05 | 236 | -0.08 | 0.17 | 0.25 |
| 0.05 | 219 | -0.08 | 0.18 | 0.26 |
| 0.05 | 203 | -0.09 | 0.18 | 0.27 |
| 0.05 | 189 | -0.09 | 0.19 | 0.28 |
| 0.05 | 177 | -0.10 | 0.19 | 0.29 |
| 0.05 | 165 | -0.10 | 0.20 | 0.30 |

# APPENDIX D. GRAPHS THAT CAN BE USED TO DETERMINE SAMPLE SIZES TO ESTIMATE CORRELATION COEFFICIENT VALUES

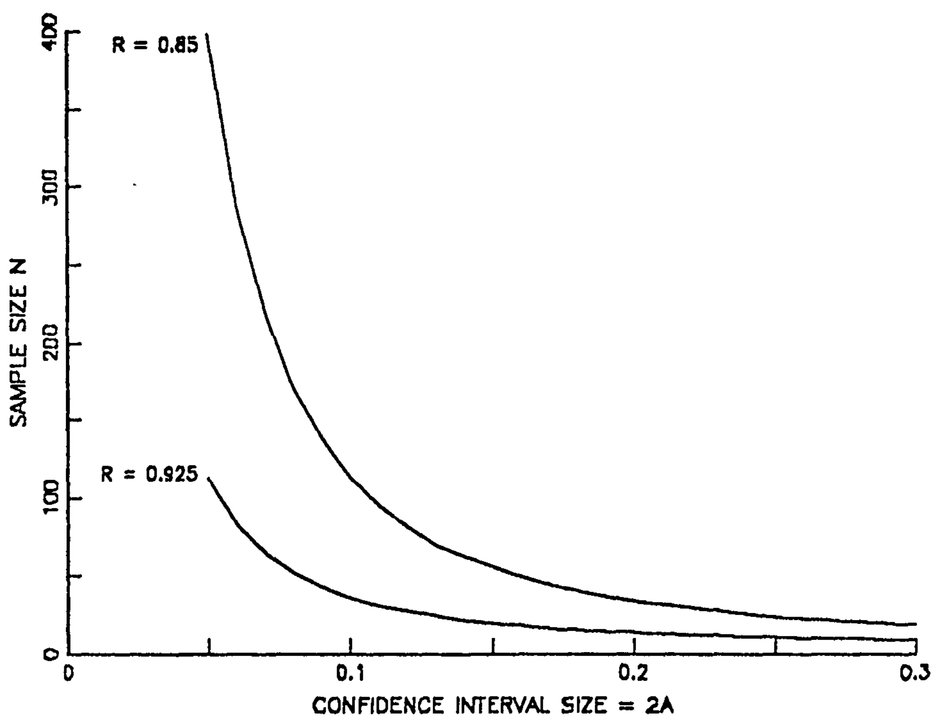


Figure 6. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.925 AND R = 0.85

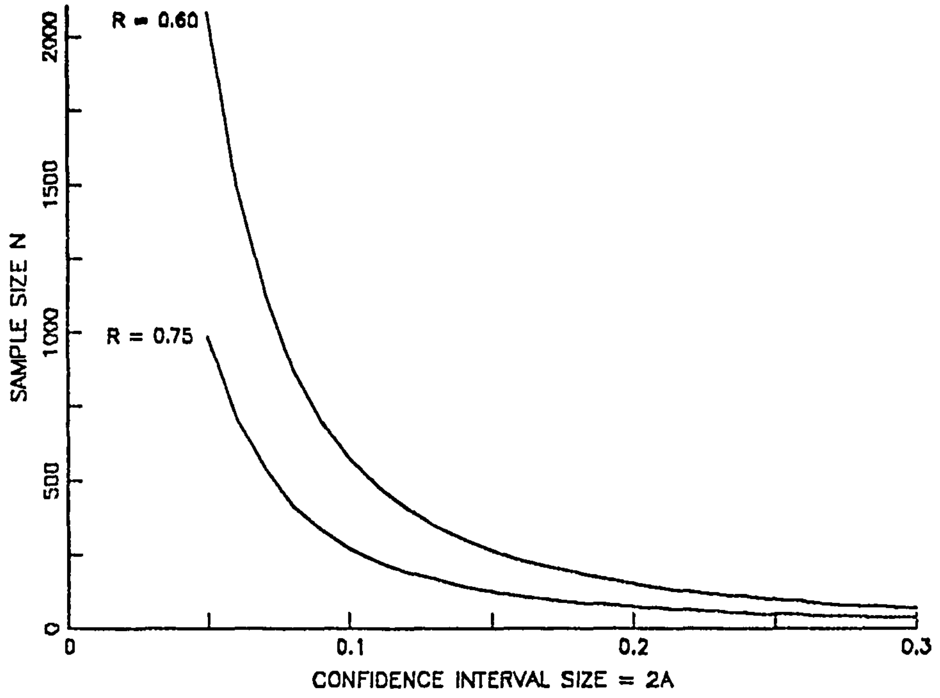


Figure 7. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.75 AND R = 0.60

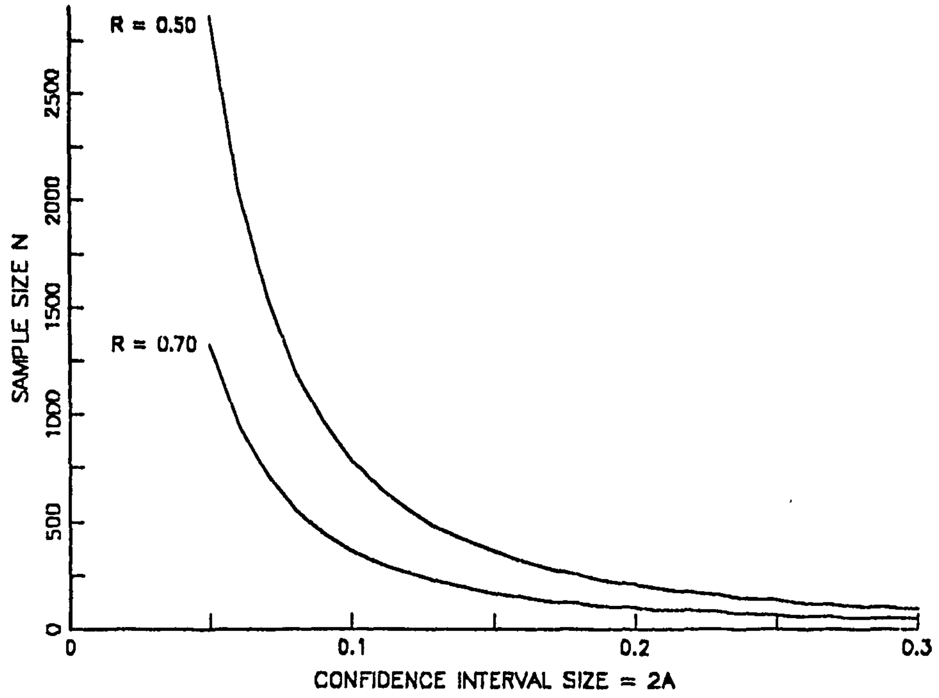


Figure 8. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.70 AND R = 0.50

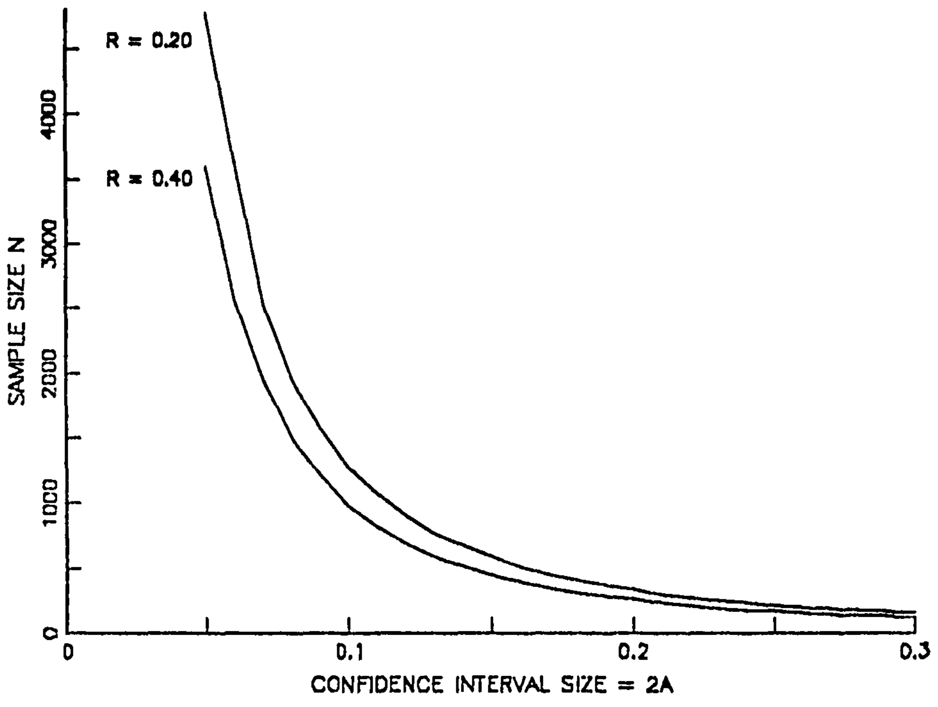


Figure 9. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.40 AND R = 0.20

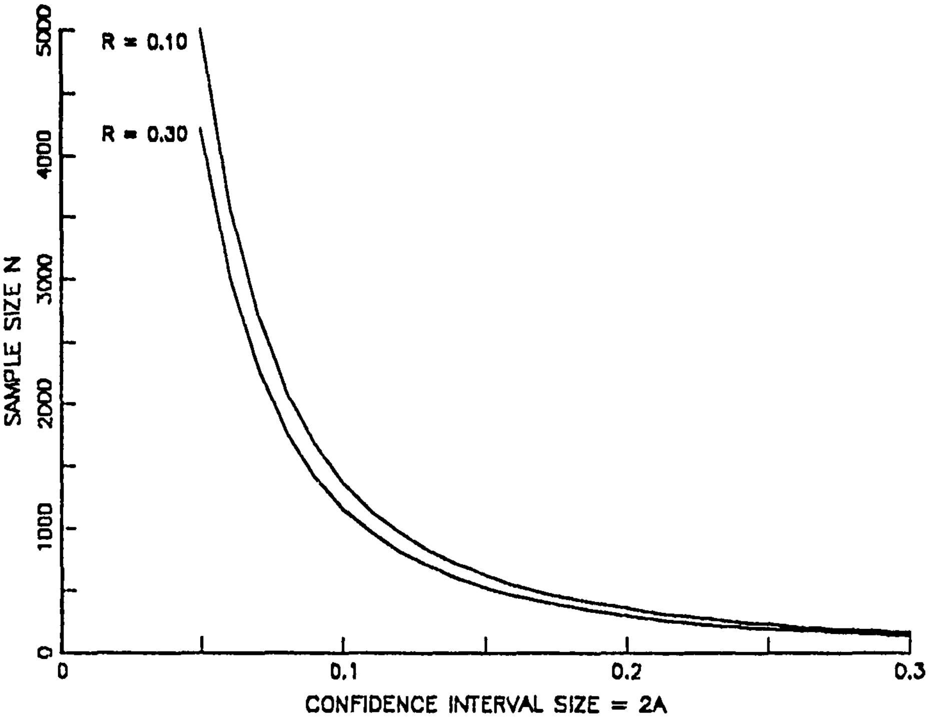


Figure 10. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.30 AND R = 0.10

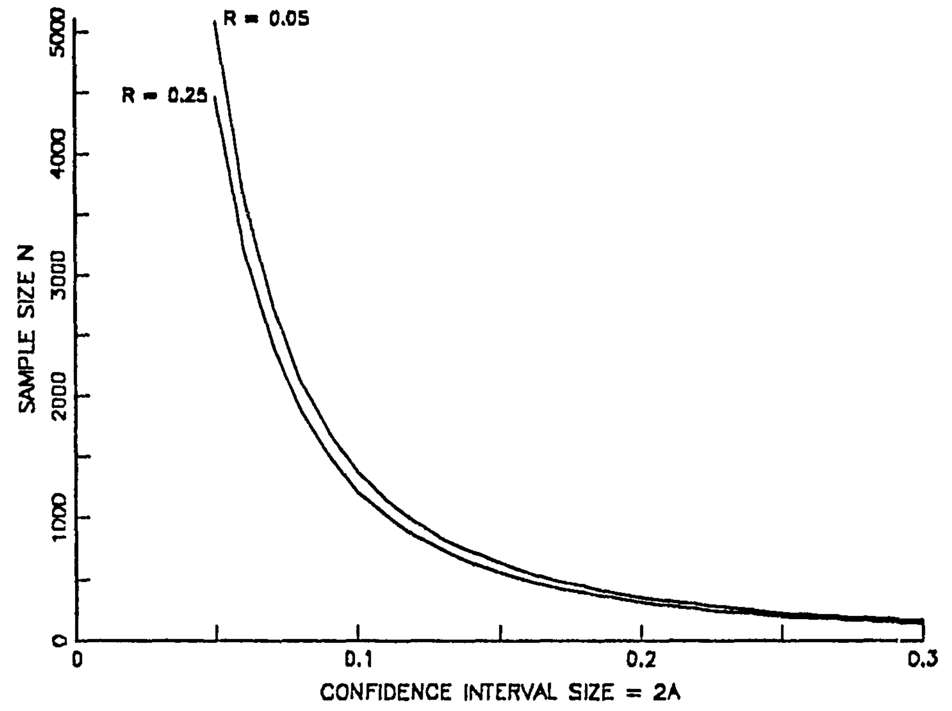
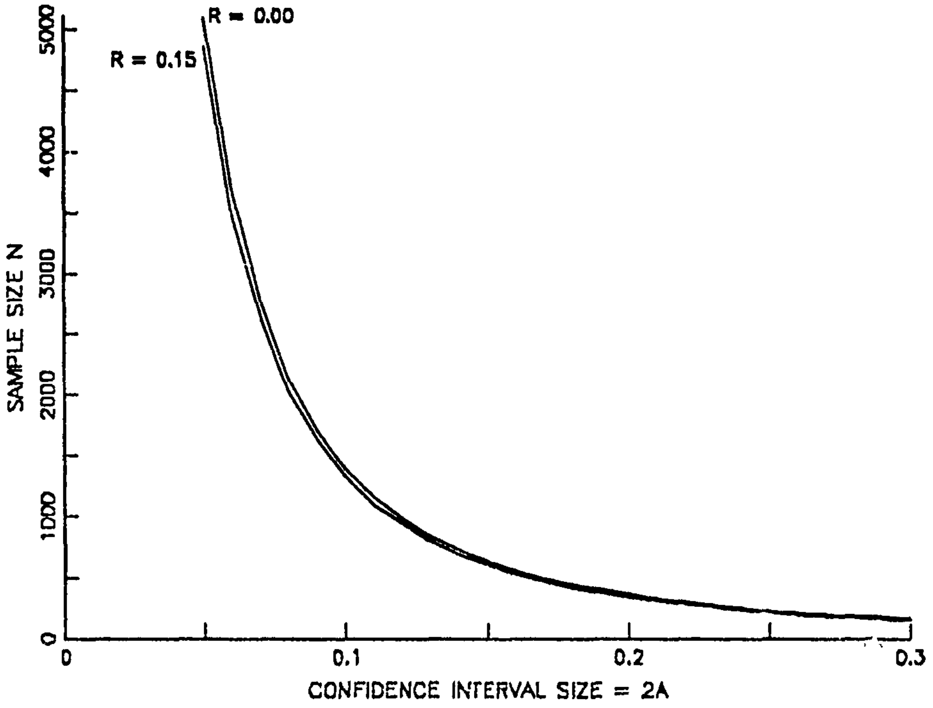


Figure 11. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.25 AND R = 0.05



Figu·e 12. REQUIRED SAMPLE SIZE FOR A 95% CONFIDENCE INTERVAL WHEN R = 0.15 AND R = 0.00

# LIST OF REFERENCES

1. Theodore, Floropoulus C., *A Bayesian Method to Improve Sampling in Weapons Testing,* Master's Thesis, Naval Postgraduate School, Monterey, California, December 1988.

2. Manion, Robert B., *Number of Samples Needed to Obtain Desired Bayesian Confidence Intervals for a Proportion,* Master's Thesis, Naval Postgraduate School, Monterey, California, March 1988.

3. Ipekkan, Ahmet Z., *Number of Test Samples Needed to Obtain a Desired Bayesian Confidence Interval for a Proportion,* Master's Thesis, Naval Postgraduate School, Monterey, California, March 1989.

4. Gibbons,Jean D., *Nonparametric Statistical Inference,* McGraw-Hill, Inc., New York, New York, 1971.

5. Conover, W. J., *Practical Nonparametric Statistics,* John Wiley & Sons, Inc., New York, New York, 1971.

6. Norman, L., Johnson & Fred, C., Leone, *Statistics and Experimental Design in Engineering and the Physical Sciences,* John Wiley & Sons, Inc., New York, 1977.

# INITIAL DISTRIBUTION LIST

No. Copies

1. Defense Technical Information Center 2
Cameron Station
Alexandria, VA 22304-6145

2. Library, Code 0142 2
Naval Postgraduate School
Monterey, CA 93943-5002

3. Deniz Kuvvetleri Komutanligi 1
Personel Daire Baskanligi
Bakanliklar-Ankara / TURKEY

4. Deniz Harp Okulu Komutanligi 1
Kutuphanesi
Tuzla - Istanbul / TURKEY

5. Hava Harp Okulu Komutanligi 1
Okul Kutuphanesi
Yesilyurt - Istanbul / TURKEY

6. Kara Harp Okulu Komutanligi 1
Okul Kutuphanesi
Bakanliklar - Ankara / TURKEY

7. Orta Dogu Teknik Universitesi 1
Okul Kutuphanesi
Ankara / TURKEY

8. Bogazici Universitesi 1
Okul Kutuphanesi
Bebek - Istanbul / TURKEY

9. Professor G. F. Lindsay, Code 55Ls 2
Operations Research Department
Naval Postgraduate School
Monterey, CA 93943

10. LCDR. Walsh, Code 55Wa 1
Operations Research Department
Naval Postgraduate School
Monterey, CA 93943

11. Kemal Salar 2
Inonu cad. Geyik Apt. No 672 Daire 10
Izmir / TURKEY