World Journal of Social Science Research
ISSN 2375-9747 (Print) ISSN 2332-5534 (Online)
Vol. 3, No. 1, 2016
www.scholink.org/ojs/index.php/wjssr

Sample Size Guideline for Correlation Analysis

Mohamad Adam Bujang[1,2] & Nurakmal Baharum[1]

[1] Biostatistics Unit, National Clinical Research Centre, Hospital Kuala Lumpur, Kuala Lumpur, Malaysia

[2] Faculty Computer and Mathematical Sciences, Universiti Teknologi Mara, Shah Alam, Selangor, Malaysia

[*] Mohamad Adam Bujang, E-mail: adam@crc.gov.my

Abstract

Correlation analysis is a common statistical analysis in various fields. The aim is usually to determine to what extent two numerical variables are correlated with each other. One of the issues that are important to be considered before conducting any correlation analysis is to plan for the sufficient sample size. This is to ensure, the results that to be derived from the analysis be able to reach a desired minimum correlation coefficient value with sufficient power and desired type I error or p-value. Sample size estimation for correlation analysis should be in line with the study objective. Researchers who are not statistician need simpler guideline to determine the sufficient sample size for correlation analysis. Therefore, this study aims to tabulate tables that show sample size calculation based on desired correlation coefficient, power and type 1 error (p-value) values. Moving towards that, simpler guidelines are proposed to estimate sufficient sample size requirements in different scenarios.

Keywords

coefficient, correlation, sample size

1. Introduction

Pearson correlation test is a univariate statistical test to measures the magnitude of correlation between two numerical variables. This statistical test is commonly used for research in various fields. The different between correlation and Simple Linear Regression (SLR) is with regards to prediction. In SLR, researchers are able to predict outcome based on a set of predictors. Usually in SLR, the condition of variables whether is a predictor or outcome is justified. Unlike correlation test, both variables may correlate with each other but it may not a relationship of predictor and outcome (Krzanowsk, 1988; Rodriguez, 1982).

Sample size guideline has to be guided with the determination of sizeable effect size that researchers can accept or tolerate. The lacking on this may contribute to publication bias in research (Kühberger et al., 2014). Meaning that, the desired effect size should also be prioritized besides setting the reasonable cut offs for p-value and power (example: p-value with 0.05 and power with 80%) (American Educational Research Association, 2006; Thompson, 2009; American Psychological Association,

2010). There are free software available to calculate sample for correlation test (StatsToDo, 2014; Sample Size Calculators, 2016). Besides that, there were also papers published with regards to sample size requirement for correlation test (Faul et al., 2009; Bonett & Wright, 2000). Although the free software and published papers are available, however researchers especially who majority are non statisticians may need simpler guideline regarding sample size requirement for correlation analysis. This is because sample size determination is very much depends on the study objective besides knowing the statistical test to be used to answer the objective. Researchers need to understand the scenario and later to determine an appropriate sample size for their study.

Therefore, this study was conducted with the aims are to produce tables of sample size requirement for correlation test and explain the guideline on how to use these tables. The scope of this paper is suitable for non statisticians for example clinicians who do research in clinical setting. They will know how many sample need to be collected and why they planned for the estimated sample size.

2. Methodology

Sample size for correlation was generated using Power Analysis and Sample Size Software (PASS) version 11.0. Sample size calculations were conducted for one correlation test and inequality tests for two correlations. The procedures of the two tests are listed under *Correlation* category. Once the procedure is selected, the *Data* tab would appear. In this *Data* tab contains parameters and options that helps to calculate the sample size. The option for "Solve For" provides selection of parameters to be solved. The parameters are *Power*, *Beta* or N. Parameter N was selected since our interest is to calculate the sample size. *Power* is the probability of rejecting a false null hypothesis. It is also equal to one minus *Beta*. While *Beta* is the probability of type II error. The value of both *Power* and *Beta* must between zero and one. *Power* of 0.80 (*Beta* = 0.20) or 0.90 (*Beta* = 0.10) are an acceptable value and commonly used. In this study, selection of *Power* are provide which are 0.80 and 0.90. The value of *Alpha* represents the probability of a type I error. Value of *Alpha* is between zero and one however, value of 0.05 is used in the study. R0 (Baseline Correlation) is the value of correlation sets at null hypothesis or the value of ρ_0. Note that the value of correlation ranges between -1 and +1. The value of R0 usually set as 0 however in this study we used diverse value of R0 which set from 0.0 to 0.8 as an option. R1 (Alternative Correlation) is the value of correlation at alternative hypothesis or the value ρ_1. In this study, value of R1 are considered between 0.1 to 0.9.

2.1 Hypothesis Testing

This part provides an option for the alternative hypothesis. It specifies the direction of the hypothesis. The null hypothesis is set as $H_0 : \rho_0 = \rho_1$.

The selections for the alternative hypothesis are:

$H_a : R_0 \neq R_1$.

This is for two-tailed test. It is use when the objective is to test whether the correlation values are different.

$H_a : R_0 < R_1$ or $H_a : R_0 > R_1$ are for one-tailed test.

All calculations are based on the algorithm described by Guenther (1977) for calculating the cumulative correlation coefficient distribution.

3. Results

Sample size tables for one correlation test are presented in Table 1 and Table 2. For R_0 is equal to zero, sample size of 782 is needed to be able to detect correlation coefficient of 0.1 with alpha of 0.05 and power of 80.0%. When correlation coefficients are increased to 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9, the sample sizes are reduced to 193, 84, 46, 29, 19, 13, 9 and 6 respectively. For R_0 is not equal to zero, the larger sample size is needed when the different between R_0 and R_1 is smaller. For example, to detect low difference of 0.1 unit different based on alpha of 0.05 and power of 80%, the estimated highest minimum sample size is between 751 (R_0 = 0.1 and R_1 = 0.2) and the estimated lowest minimum sample size is 59 (R_0 = 0.8 and R_1 = 0.9). Sample size table for inequality tests of two correlations tests is presented in Table 3 and Table 4. The larger sample size is needed when the different between the two correlations from different samples are smaller. For example, to detect low difference of 0.1 unit different based on alpha of 0.05 and power of 80%, the estimated highest minimum sample size is between 1563 (R_0 = 0.0 and R_1 = 0.1) and the estimated lowest minimum sample size is 116 (R_0 = 0.8 and R_1 = 0.9) as shown in Table 3. The larger sample size is needed when to detect power of 90.0% as shown in Table 4.

Table 1. Sample Size Requirement for One Correlation Test with *Power* = 80%, *Alpha* = 0.05

R_0	R_1	No. of subjects	R_0	R_1	No. of subjects
0.0	0.1	782	0.3	0.4	605
	0.2	193		0.5	139
	0.3	84		0.6	56
	0.4	46		0.7	28
	0.5	29		0.8	15
	0.6	19		0.9	9
	0.7	13	0.4	0.5	500
	0.8	9		0.6	111
	0.9	6		0.7	43
0.1	0.2	751		0.8	20
	0.3	182		0.9	10
	0.4	78	0.5	0.6	382
	0.5	41		0.7	80
	0.6	25		0.8	29

	0.7	16		0.9	12
	0.8	11	0.6	0.7	262
	0.9	7		0.8	51
0.2	0.3	691		0.9	16
	0.4	163	0.7	0.8	149
	0.5	68		0.9	24
	0.6	35	0.8	0.9	59
	0.7	20			
	0.8	12			
	0.9	8			

Table 2. Sample Size Requirement for One Correlation Test with *Power* = 90%, *Alpha* = 0.05

R_0	R_1	No. of subjects	R_0	R_1	No. of subjects
0	0.1	1046	0.3	0.4	809
	0.2	258		0.5	185
	0.3	112		0.6	74
	0.4	61		0.7	36
	0.5	37		0.8	20
	0.6	24		0.9	11
	0.7	17	0.4	0.5	668
	0.8	11		0.6	147
	0.9	8		0.7	56
0.1	0.2	1005		0.8	26
	0.3	243		0.9	12
	0.4	103	0.5	0.6	510
	0.5	55		0.7	107
	0.6	32		0.8	38
	0.7	20		0.9	15
	0.8	13	0.6	0.7	349
	0.9	8		0.8	67
0.2	0.3	924		0.9	20
	0.4	218	0.7	0.8	199
	0.5	90		0.9	32
	0.6	46	0.8	0.9	78
	0.7	26			
	0.8	16			

	0.9	9

Table 3. Sample Size Requirement for Inequality Tests for Two Correlations with *Power* = 80%, *Alpha* = 0.05

R_0	R_1	No. of subjects	R_0	R_1	No. of subjects
0	0.1	1563	0.3	0.4	1209
	0.2	385		0.5	277
	0.3	167		0.6	110
	0.4	91		0.7	54
	0.5	56		0.8	29
	0.6	36		0.9	15
	0.7	24	0.4	0.5	998
	0.8	17		0.6	220
	0.9	11		0.7	83
0.1	0.2	1501		0.8	38
	0.3	362		0.9	18
	0.4	154	0.5	0.6	762
	0.5	81		0.7	159
	0.6	48		0.8	56
	0.7	30		0.9	22
	0.8	19	0.6	0.7	521
	0.9	12		0.8	99
0.2	0.3	1380		0.9	29
	0.4	325	0.7	0.8	297
	0.5	134		0.9	46
	0.6	69	0.8	0.9	116
	0.7	39			
	0.8	23			
	0.9	13			

Table 4. Sample Size Requirement for Inequality Tests for Two Correlations with *Power* = 90%, *Alpha* = 0.05

R_0	R_1	No. of subjects	R_0	R_1	No. of subjects
0	0.1	2091	0.3	0.4	1617
	0.2	515		0.5	369
	0.3	223		0.6	146

	0.4	121		0.7	71
	0.5	73		0.8	37
	0.6	47		0.9	19
	0.7	31	0.4	0.5	1334
	0.8	21		0.6	293
	0.9	13		0.7	110
0.1	0.2	2008		0.8	50
	0.3	484		0.9	23
	0.4	205	0.5	0.6	1019
	0.5	108		0.7	211
	0.6	63		0.8	73
	0.7	39		0.9	28
	0.8	25	0.6	0.7	696
	0.9	15		0.8	131
0.2	0.3	1846		0.9	38
	0.4	434	0.7	0.8	396
	0.5	178		0.9	61
	0.6	91	0.8	0.9	154
	0.7	51			
	0.8	30			
	0.9	17			

4. Discussions

When to estimate sample size for correlation analysis? Sample size for correlation test has to be considered if one of the aims of a study is to determine the magnitude of correlation between two numerical variables. Usually in general, correlation analysis can be analyzed in three different scenarios. First is to determine the correlation when we assumed null hypothesis is equal to zero, second to determine the correlation when we assume null hypothesis is not equal to zero and third to determine the correlation for inequality tests for two correlations.

4.1 When Null Hypothesis is Equal to Zero

When null hypothesis is equal to zero, it means that the researchers assumed that the two variables have no association in the null hypothesis. In other words, researchers aim to determine the magnitude of association between the two variables. Usually in research, researchers will aim higher effect size in order to prove that the two variables have a correlation. Generally, correlation coefficient of 0.3 is considered sizeable (Cohen, 1992). Thus to measure the medium correlation coefficient with power of 80%, minimum sample size of 84 is needed to achieve correlation coefficient of at least 0.3. The lower

sample size is needed to detect the higher correlation coefficient such as sample size of minimum 29 and 13 to be able to detect correlation coefficient of 0.5 and 0.7 respectively.

As for example, researchers would like to determine the correlation between "variable A" and "variable B". From past literatures, the correlation between both variables was estimated within 0.42-0.55. What is the minimum sample size that the researcher needs to collect to get significant result? Since the lower effect size will yield the larger sample, therefore from Table 1, the minimum required sample size is 46. In Table 1, correlation coefficient for 0.42 was not calculated. However, the lower but closest correlation coefficient can be used such as 0.4. If the researchers have no issue in collecting more samples, then researcher may consider collecting minimum samples of 193 to be able to detect at least correlation coefficient of 0.2. In this case, although the correlation coefficient of 0.4 might not be achievable, at least researcher is able to detect the lower correlation coefficient with significant results.

The example of sample size statement can be written as follow: "The aim of the study is to determine the correlation between "variable A" and "variable B". The aim is to get significant result ($p < 0.05$) with sufficient power (80%) to detect at least correlation coefficient of 0.4. Therefore, the minimum required sample size for this study is 46. The formula for calculation is based on two-tailed test (Guenther, 1977)".

4.2 When Null Hypothesis is Not Equal to Zero

When null hypothesis is not equal to zero, it means that the researcher assumed that the two variables have a certain magnitude of correlation coefficient in the null hypothesis. In other words, researchers aim to determine the magnitude of association between the two variables is higher or lower than an estimated correlation coefficient in the null hypothesis. For example, suppose from past experience showed that the correlation between the exams score between mathematics and science subjects is 0.50 in a particular class. A teacher wants to test whether the correlation between the two subjects were more than 0.50 in a different class. How many students are there that need to be collected? Here, the correlation coefficient in the null hypothesis is = 0.50. In this scenario, the larger sample size is needed when the different between the correlation coefficient in the null and alternative hypothesis is smaller. Based on power of 80% for instance, a minimum sample size of 382 is needed to be able to detect at least 0.1 units different in correlation coefficient (e.g., 0.5 versus 0.6). However, only a minimum sample size of 80 samples is needed to be able to detect at least 0.2 units different in correlation coefficient (e.g., 0.5 versus 0.7). Then it is up to the researcher to decide to what extent the researcher can accept the different between the correlation coefficient in the null and alternative hypothesis is sizeable enough to indicate there is a different of correlation between the two classes.

The example of sample size statement can be written as follow: "The aim of the study is to determine the correlation between "variable A" and "variable B". "Variable A" and "variable B" from a previous study is said to have a correlation coefficient of 0.5. Researcher estimated that the correlation coefficient between "variable A" and "variable B" is able to achieve at least 0.7 from new set of data. Thus, the aims is to get significant result ($p < 0.05$) with sufficient power (80%) to detect at least

correlation coefficient of 0.7 when the correlation coefficient in the null hypothesis is 0.5. Therefore, the minimum required sample size for this study is 80. The formula for calculation is based on two-tailed test (Guenther, 1977)".

4.3 Inequality Tests for Two Correlations

In a scenario of inequality tests for two correlations, researcher aims to determine whether the magnitude of association between the two variables is different from two different groups of sample. For example, a teacher wants to test whether the correlation between the two subjects (mathematics and science) are different between males and females students. How many students are there that need to be collected? In this scenario, the larger sample size is needed when the different of correlation coefficient of the two subjects between males and females are smaller. Based on power of 80% for instance, a minimum sample size of 762 is needed to be able to detect at least 0.1 units different in correlation coefficient when correlation coefficient of males is estimated at 0.5 versus correlation coefficient of females is estimated at 0.6. However, only a minimum sample size of 56 samples is needed to be able to detect at least 0.3 units different in correlation coefficient (e.g., 0.5 versus 0.8).

The example of sample size statement can be written as follow: "The aim of the study is to determine whether the correlation coefficient between two exam scores (mathematics and science subjects) are different between gender. Researchers estimated that the correlation coefficient between the two exams scores among males student is 0.5. Researchers have estimated that the correlation coefficients between the two exams scores are higher among females' students which are about at least 0.3 units different. Thus, the aims is to get significant result ($p < 0.05$) with sufficient power (80%) to detect at least the different of correlation coefficient of 0.3 when the correlation coefficient for males and females student are estimated at 0.5 and 0.8 respectively. Therefore, the minimum required sample size for this study is 56 samples among male's students and 56 samples among female's students. The formula for calculation is based on two-tailed test (Guenther, 1977)".

4.4 Other Considerations

Say a researcher intends to compare the correlation coefficient between more than two groups. Authors suggested the researcher could use the approach from inequality tests for two correlations. For example, A teacher wants to test whether the correlation between the two subjects (mathematics and science) are different between three different schools (school A, school B and school C). The teacher needs to estimate the lowest correlation coefficient from either one of the schools. Then he/she needs to estimate the sizeable different of correlation coefficient to defeats the lowest correlation coefficient that was set earlier. For example, the teacher estimated correlation coefficient of the two subjects in school C is 0.5 and he/she satisfies if one of the schools could achieve the correlation coefficient of 0.7. Then the minimum required sample size is 159 in each school.

Taking example from "*When Null Hypothesis is Not Equal to Zero*", lower sample is needed to be able to detect high effect size and vice versa. Generally, it is rare the correlation of two variables is studied just to determine the correlation is weak or poor. If the aim is to determine whether the two variables

have a sizeable correlation, thus the minimum effect size or correlation coefficient usually to be set is 0.3 (Cohen, 1992). However, it is not encourage for researcher to plan for small sample although they target the finding will yield strong correlation. Result may yield otherwise after the analysis is conducted. Hence, we proposed minimum of 29 samples or subject to be recruited to determine the reasonable high correlation of two variables (to be able to detect effect size to detect correlation coefficient = 0.5 with alpha and power are set at 0.05 and 80% respectively). Results for sample size determination presented in this study actually as a guideline for minimum sample. In fact, researchers are encouraged to take more than the minimum sample. Therefore, taking large sample is also reasonable since it will increase the accuracy of the estimates. Some studies suggest that by taking sample of more than 300, the statistics that will be derived from the sample will be likely the same with the true value in the intended population (Bujang et al., 2009; Bujang et al., 2015). This guideline perhaps is suitable for survey that has relatively large population.

One of the limitations of this paper is the limited in terms of statistics for the purpose of sample size calculation. For example, this paper did not take into account the sample size requirement for correlation coefficient with two decimal points such 0.35 or 0.37. Generally, researchers are only estimating the sample size and therefore estimating the correlation coefficient with one decimal point is sufficient. For instance, researcher can estimate the correlation coefficient of 0.3 instead of 0.35 and 0.37 since it will yield larger sample size. Besides that, this paper also did not include non parametric test such as spearman rank correlation. However, when estimating for sample size, researcher will not know whether the assumption for Pearson correlation is met or not. Therefore, researchers can assume the assumptions for Pearson correlation is met when use these tables for estimating the sample size.

Acknowledgments

The authors would like to acknowledge the Director General of the Ministry of Health for his support in our effort in the registry. We would like to extend the appreciation to Mr John Hon Yoon Khee for his effort in proofreading this manuscript.

References

American Educational Research Association. (2006). Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*, *35*, 33-40.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC.

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

Bujang, M. A., Ghani, P. A., Zolkepali, N. A., Selvarajah, S., & Haniff, J. A. (2012). Comparison between convenience sampling versus systematic sampling in getting the true parameter in a population: Explore from a clinical database: The Audit Diabetes Control Management (ADCM)

registry. In *Proceedings of the International Conference Statistics Sciences Business Engineering*. http://dx.doi.org/10.1109/ICSSBE. 2012.6396615

Bujang, M. A., Sa'at, N., Joys, A., Reena, A., & Mariana, M. (2015). An audit of the statistics and the comparison with the parameter in the population. *AIP Conference Proceedings*, 1682. http://dx.doi.org/10.1063/1.4932510

Cohen, J. A. (1992). Power primer. *Psychological Bulletin*, *112*, 155-159.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *4*, 1149-1160.

Guenther, W. C. (1977). Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient. *The American Statistician*, *31*(1), 45-48.

Krzanowsk, W. J. (1988). *Principles of multivariate analysis: A user's perspective* (pp. 405-432). Oxford, England.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). *A Diagnosis Based on the Correlation between Effect Size and Sample Size*. http://dx.doi.org/10.1371/journal.pone.0105825

Power Analysis and Sample Size Software. (n. d.). NCSS, LLC.

Rodriguez RN. Correlation. (1982). In S. Kotz, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 193-204). New York: Wiley.

Sample Size Calculators. (2016). Retrieved from http://www.sample-size.net/

StatsToDo Home Page. (2014). Retrieved from https://www.statstodo.com/SSizCorr_Pgm.php

Thompson, B. A. (2009). Brief Primer on Effect Sizes. *Journal of Teaching in Physical Education*, *28*, 251-254.