Restor Dent Endod. 2024 May;49(2):e21
https://doi.org/10.5395/rde.2024.49.e21
eISSN 2234-7666

Restorative Dentistry & Endodontics RDE

Statistical Research Article

Check for updates

An elaboration on sample size determination for correlations based on effect sizes and confidence interval width: a guide for researchers

Mohamad Adam Bujang

Clinical Research Centre, Sarawak General Hospital, Ministry of Health Malaysia, Kuching, Sarawak, Malaysia

OPEN ACCESS

Received: Oct 11, 2023
Accepted: Jan 4, 2024
Published online: May 2, 2024

Citation
Bujang MA. An elaboration on sample size determination for correlations based on effect sizes and confidence interval width: a guide for researchers. Restor Dent Endod 2024;49(2):e21.

***Correspondence to**
Mohamad Adam Bujang, PhD
Clinical Research Centre, Sarawak General Hospital, Jalan Tun Ahmad Zaidi Adruce, 93586 Kuching, Sarawak, Malaysia.
Email: adam@crc.gov.my

Copyright © 2024. The Korean Academy of Conservative Dentistry
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Conflict of Interest
No potential conflict of interest relevant to this article was reported.

Author Contributions
Conceptualization: Bujang MA. Data curation: Bujang MA. Formal analysis: Bujang MA. Funding acquisition: Bujang MA. Investigation: Bujang MA. Methodology: Bujang MA. Project

ABSTRACT

Objectives: This paper aims to serve as a useful guide for sample size determination for various correlation analyses that are based on effect sizes and confidence interval width.
Materials and Methods: Sample size determinations are calculated for Pearson's correlation, Spearman's rank correlation, and Kendall's Tau-b correlation. Examples of sample size statements and their justification are also included.
Results: Using the same effect sizes, there are differences between the sample size determination of the 3 statistical tests. Based on an empirical calculation, a minimum sample size of 149 is usually adequate for performing both parametric and non-parametric correlation analysis to determine at least a moderate to an excellent degree of correlation with acceptable confidence interval width.
Conclusions: Determining data assumption(s) is one of the challenges to offering a valid technique to estimate the required sample size for correlation analyses. Sample size tables are provided and these will help researchers to estimate a minimum sample size requirement based on correlation analyses.

Keywords: Confidence interval; Correlation; Parametric and non-parametric; Sample size

INTRODUCTION

Correlation analysis is a common statistical technique in most types of research. It requires 2 variables that are either expressed in a numerical form or measured on at least an ordinal scale. In other words, the correlation test aims to measure whether or not there is a correlation between the 2 variables, as well as the direction and magnitude (or strength) of the correlation. The correlation is denoted by the symbol *r* and its coefficient shall hold a value that can range from -1 (perfect negative correlation) to +1 (perfect positive correlation) where a correlation of 0 indicates the absence of a relationship between the 2 variables being studied [1].

Generated by Xmlinkpress

administration: Bujang MA. Resources: Bujang MA. Software: Bujang MA. Supervision: Bujang MA. Validation: Bujang MA. Visualization: Bujang MA. Writing - original draft: Bujang MA. Writing - review & editing: Bujang MA.

ORCID iDs
Mohamad Adam Bujang
https://orcid.org/0000-0001-6231-4062

There are 3 popular measures of correlation such as Pearson's correlation, Kendall-Tau b correlation, and Spearman's rank correlation. The important assumption for performing these correlation analyses is that the 2 values should be a related pair of values for each observation. The measurement scale of the data for conducting Pearson's correlation and Spearman's rank correlation analyses are the same, which is preferably in an interval or a ratio scale. Pearson's correlation is a parametric test and therefore its analysis will require the fulfillment of the parametric assumptions. If any of these parametric assumptions have been violated, then it is recommended to use the non-parametric alternative to Pearson's correlation, *i.e.*, Spearman's rank correlation instead [2].

Kendall-Tau b is also a non-parametric type of analysis that commands other special characteristics in the data pair whereby it is a type of correlation test that can only accept 2 variables that are being measured at least in an ordinal scale and also following a monotonicity pattern [3]. This means that there is a consistent direction of the relationship between the 2 variables in that when 1 variable goes up, the other goes up. In this case, a plot of the 2 variables would move consistently in the upward direction. The correlation can also be monotonic if when 1 variable goes up, the other goes down. Thus, the plot of the 2 variables would move consistently in the downward direction.

One of the common problems encountered by a researcher in designing a study involving correlation analysis is to determine its minimum required sample size. In the case of correlation analyses, the calculation allows the researcher to set any value for the effect size, which can range from extremely small (*i.e.* $r = 0.1$) to extremely large (*i.e.* $r = 0.9$). It is necessary to establish appropriate guidelines for sample size calculation and justification for these analyses to assist the researcher in making a prudent decision for sample size planning involving a correlation analysis. Therefore, this study aims to provide a useful guideline for sample size determination for conducting correlation analyses based on an estimation by the confidence interval. Although the sample size determination for correlations has always been facilitated by the use of appropriate statistical software and published sample size tables which are readily available in the existing literature, a list of useful guidelines does serve to assist the researchers in determining the minimum required sample size for their correlation studies [4-6].

MATERIALS AND METHODS

Sample size formula

Sample size calculation was conducted for Pearson's correlation, Spearman's rank correlation, and Kendall's Tau-b correlation. The calculation for these statistical analyses was based on a formula introduced in a previous study [4]. For all these calculations, the confidence interval is set at 95.0%. The target correlation coefficients are set at 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Correspondingly, the precision of the 95% confidence interval of the correlation coefficient will be set by pre-specifying the widths of the confidence intervals to be 0.1, 0.2, 0.3, 0.4, and 0.5. A careful perusal followed by a thorough discussion of the above results shall form the basis for guiding a researcher in determining the minimum required sample size for their studies.

Software

The minimum sample size requirement was estimated by using PASS 2022 (NCSS LLC; Kaysville, UT, USA), Power Analysis and Sample Size Software. The PASS software is a commercial software that provides a list of tools for determining sample size requirements for thousands of statistical tests and confidence interval scenarios. Each sample size technique in this software has been carefully validated and substantiated by scientifically sound published articles [7].

RESULTS

A careful observation of the results has found that the sample size requirement increases for a smaller target correlation coefficient and also for a smaller or narrower width of the 95% confidence interval, both of which are set for the sample size calculation. Using the same effect sizes, the sample size calculation obtained from Spearman's rank correlation shall yield the largest value in comparison to those values obtained from the other 2 statistical analyses. There are also several highly significant differences between the 3 different statistical analyses in their sample size determination (**Table 1**).

By fixing the width of confidence interval at 0.3 and level of the confidence at 95.0%, the recommended sample size is between 12 (for $r_p = 0.9$) to 143 (for $r_p = 0.3$) for Pearson's correlation, between 8 (for $r_k = 0.9$) to 65 (for $r_k = 0.3$) for Kendall's Tau-b's correlation, and between 16 (for $r_s = 0.9$) to 149 (for $r_s = 0.3$) for Spearman's rank correlation. For a study that aims to achieve a target correlation with a high level of precision such as fixing its 95% confidence level width at 0.1, the recommended sample size requirement is between 62 (for $r = 0.9$) to 1274 (for $r = 0.3$) for Pearson's correlation, and between 30 (for $r_k = 0.9$) to 560 (for $r_k = 0.3$) for Kendall's Tau-b's correlation. On the other hand, Spearman's rank correlation is commanding the highest sample size requirement by necessitating a sample size of between 86 (for $r = 0.9$) to 1331 (for $r = 0.3$) (**Table 1**).

The results have indicated that some of the calculations have reported a lower bound for a 95% level of confidence such as at zero or even negative values, especially for those with very low target correlation coefficient values (*i.e.* 0.1 and 0.2). Therefore, it is not necessary to set a bigger width of 95% confidence interval for very low values of target correlation coefficient such as 0.1 and 0.2. This means that only a very narrow width of 0.1 will be sufficient for a low target correlation coefficient value of 0.1, and a width of 0.1, 0.2, and 0.3 will also be sufficient for a slightly higher target correlation coefficient value of 0.2.

On the other hand, it was found that some upper bound of 95% confidence intervals yielded a confidence level of more than 1.0, especially for a high target correlation coefficient value such as 0.8 and 0.9. Again, it is also not necessary to set a big width of 95% confidence interval for a high target correlation coefficient value of 0.8 and 0.9. Hence, only a narrow width of 0.1, 0.2, 0.3, and 0.4 will be required for a high target correlation coefficient value of 0.8. Likewise, an even narrower width of 0.1 and 0.2 will be required for a higher target correlation coefficient value of 0.9. Hence, it is advisable to take such measures when adopting such guidelines for the determination of sample size requirements for correlation tests to avoid erroneously deciding upon an extremely small sample size for our study (**Table 1**).

Table 1. Sample size calculation for Pearson's, Kendall-Tau's, and Spearman's rank correlations based on 95% confidence interval (CI) width

Effect size	Width	Width for both sides	95% CI	n_p	n_k	n_s	diff.*
0.1	0.1	± 0.05	0.050, 0.150	1,507	662	1,515	8
	0.2†	± 0.10†	0.000, 0.200†	378†	168†	379†	1†
	0.3†	± 0.15†	−0.050, 0.250†	168†	77†	169†	1†
	0.4†	± 0.20†	−0.100, 0.300†	95†	45†	96†	1†
	0.5†	± 0.25†	−0.150, 0.350†	61†	30†	62†	1†
0.2	0.1	± 0.05	0.150, 0.250	1,417	622	1,446	29
	0.2	± 0.10	0.100, 0.300	355	158	362	7
	0.3	± 0.15	0.050, 0.350	159	72	162	3
	0.4†	± 0.20†	0.000, 0.400†	90†	42†	91†	1†
	0.5†	± 0.25†	−0.050, 0.450†	58†	28†	59†	1†
0.3	0.1	± 0.05	0.250, 0.350	1,274	560	1,331	57
	0.2	± 0.10	0.200, 0.400	320	143	334	14
	0.3	± 0.15	0.150, 0.450	143	65	149	6
	0.4	± 0.20	0.100, 0.500	81	38	85	4
	0.5	± 0.25	0.050, 0.550	53	26	55	2
0.4	0.1	± 0.05	0.350, 0.450	1,086	478	1,173	87
	0.2	± 0.10	0.300, 0.500	273	122	295	22
	0.3	± 0.15	0.250, 0.550	123	57	132	9
	0.4	± 0.20	0.200, 0.600	70	34	75	5
	0.5	± 0.25	0.150, 0.650	46	23	49	3
0.5	0.1	± 0.05	0.450, 0.550	867	382	975	108
	0.2	± 0.10	0.400, 0.600	219	99	246	27
	0.3	± 0.15	0.350, 0.650	99	46	111	12
	0.4	± 0.20	0.300, 0.700	57	28	64	7
	0.5	± 0.25	0.250, 0.750	37	19	42	5
0.6	0.1	± 0.05	0.550, 0.650	633	280	746	113
	0.2	± 0.10	0.500, 0.700	161	73	189	28
	0.3	± 0.15	0.450, 0.750	74	35	86	12
	0.4	± 0.20	0.400, 0.800	43	22	50	7
	0.5	± 0.25	0.350, 0.850	29	16	33	4
0.7	0.1	± 0.05	0.650, 0.750	404	180	503	99
	0.2	± 0.10	0.600, 0.800	105	49	129	24
	0.3	± 0.15	0.550, 0.850	49	24	60	11
	0.4	± 0.20	0.500, 0.900	30	16	36	6
	0.5	± 0.25	0.450, 0.750	20	12	25	5
0.8	0.1	± 0.05	0.750, 0.850	205	93	269	64
	0.2	± 0.10	0.700, 0.900	56	27	72	16
	0.3	± 0.15	0.650, 0.950	28	15	35	7
	0.4	± 0.20	0.600, 1.000	18	11	22	4
	0.5‡	± 0.25‡	0.550, 1.000‡	13‡	9‡	16‡	3‡
0.9	0.1	± 0.05	0.850, 0.950	62	30	86	24
	0.2	± 0.10	0.800, 1.000	20	12	27	7
	0.3‡	± 0.15‡	0.750, 1.000‡	12‡	8‡	16‡	4‡
	0.4‡	± 0.20‡	0.700, 1.000‡	9‡	7‡	12‡	3‡
	0.5‡	± 0.25‡	0.650, 1.000‡	8‡	6‡	10‡	2‡

n_p, sample size for Pearson's correlation; n_k, sample size for Kendall Tau-b's correlation; n_s, sample size for Spearman's correlation.
*diff. represents the difference between n_s - n_p; †Lower bound with either zero or negative value; ‡Upper bound with more than 1.000.

DISCUSSION

This paper discusses the concept and guidelines for sample size determination for correlation analyses specifically based on a specific method. Although a theory regarding the sample size calculation technique that was based on confidence interval width for correlation analyses has already been published, this paper emphasizes further elaboration in the practices and step-by-step guidelines on how to conduct sample size planning for correlation analyses. These discussions are necessary to strengthen the understanding regarding sample size among the researchers who are the majority without adequate knowledge of statistics. Literature has proven that such discussion papers are very useful for researchers [8,9]. The discussion is divided into several sections to elaborate on the various aspects of the sample size determination for performing correlation analyses, as well as to raise any issues of concern in this area.

Advantage of performing a sample size calculation based on precision (via a confidence interval)

There is a fundamental difference between sample size calculation based on hypothesis testing and the same calculation based on a 95% confidence interval. Based on the calculation done by a previous study involving hypothesis testing, the researcher will not be able to determine the level of accuracy of the correlation coefficients based on the sample size alone. Instead, if the sample size calculation is derived from an estimation obtained from the confidence interval of the target correlation coefficient value, then it will enable the researcher to impose clear expectations on its target value [10]. For example, if a study attempts to target a correlation coefficient of 0.8 by conducting the Pearson's correlation, then the researcher will know that to prove the truth of such a high level of correlation with high accuracy such as between 0.75 to 0.85, a minimum sample size of 205 subjects will be required. If the researcher manages to recruit only 20 subjects, he/she will know that the correlation coefficients in the target population will probably range from 0.45 to 0.75. In other words, as the accuracy for the derived value of the correlation coefficient will probably range from low to moderate only, therefore he/she will not be able to confidently determine exactly what is the magnitude of the correlation.

Besides that, the calculation of the correlation coefficient based on a 95% confidence interval will also provide a hint for setting an appropriate width of the confidence interval for sample size calculation. For example, setting the confidence interval's width of 0.3 for a very low target correlation coefficient value of 0.1 is unnecessary since this will yield a lower bound which indicates a too low level of confidence, which will probably range from 0 to negative values. Likewise, it will also not be necessary to set a bigger width of the confidence interval to reduce the sample size requirement if the resulting calculation produces too low a level of confidence with 0 or even negative values for its lower bound of the confidence interval, or produces too high a level of confidence with more than 1.0 for its upper bound of the confidence interval.

A comparison between sample size calculation based on parametric versus non-parametric

This study has prepared 3 sample size tables for use in guiding the conduct of both parametric and non-parametric statistical analyses for correlation. Among the 3 statistical analyses, namely: Pearson's correlation, Spearman's rank correlation, and Kendall's Tau-b correlation, it was found that the sample size requirement for performing Kendall's Tau-b correlation is the smallest. Results of this study have also shown that based on the same effect size, Spearman's rank correlation will usually yield a larger sample size requirement than Pearson's correlation. Bearing in mind that this difference can often be considerably large, it is therefore necessary to first check the validity of these assumptions in the data involved such as from the literature, before proceeding to the sample size estimation. Without giving due consideration to these assumptions, the sample size calculation may not be valid since it is always possible for a study to be rendered underpowered subsequently, and hence valid conclusions cannot be drawn from its findings.

Guideline for assessing the sample size requirement based on the predetermined effect size and width of 95% confidence interval

Using sample size tables is one of the smartest and easiest ways to estimate the minimum sample size requirement [8,9]. Such sample size tables shall facilitate a researcher in deciding the minimum required sample size without the need to identify the correct formula for sample size calculation which is based on a specific statistical analysis. Nevertheless, it is still recommended to formulate a list of very useful guidelines to assist a researcher in making the correct decision for sample size determination when planning a correlation study. In the case of correlation, the calculation shall allow the researchers to pre-specify any values of its effect size which can range from an extremely small value (*i.e.* $r = 0.1$) to an extremely large value (*i.e.* $r = 0.9$). In fulfillment of this aim, this study thereby proposes and recommends a simplified yet useful guideline for which a small, moderate, or high value of the effect size has been predetermined in a study involving the correlation coefficient.

It is very rare for a researcher to purposefully select a small effect size (*i.e.* effect size = 0.1 or 0.2) for a study. One of the reasons for setting a smaller effect size is to merely prove that the 2 variables have either very low or almost 0 correlation. In the multiple regression model, one of the assumptions is to ensure all the predictors have no strong correlation with each other [11,12]. One of the ways to confirm this is by measuring the level of correlation between the 2 variables. By convention, it can be assumed that 2 variables have a weak correlation when the correlation between them is less than 0.3 since less than 0.3 indicates a low correlation [13]. On the other hand, other studies can also opt to use a much bigger cut-off such as a correlation coefficient of less than 0.8 to indicate no strong correlation [12,14]. Another possible scenario is when the researcher aims to determine which predictors (or variables) should be regarded as not important or less important. In some cases, proving that a variable does not correlate with another variable can also be considered an important finding [15,16].

Conventionally, most research studies aim to prove that there is a correlation between 2 variables. In this case, the ideal effect size is 0.3 since a cut-off value of 0.3 and above shall indicate that the correlation is sizeable or of a significant magnitude [13]. Before conducting an analysis, a researcher usually will not know the exact value of the effect size which a study can realistically achieve since its true value can be anything from 0.3 and above. However, for sample size calculation, it is often recommended to set an effect size of 0.3 because by doing so, the sample size calculation can yield a sufficiently large sample size requirement which can detect at least a moderate to high effect size.

Another rare condition is when the researcher is required to set an excellent effect size for a study, which can often be particularly high (*i.e.* effect size = 0.8 or 0.9). However, the researcher shall need to exercise additional caution since an unusually high effect size will also yield a very low sample size requirement [4-6]. Normally, a smaller sample size requirement is often preferable by many researchers since it will save cost and time but it is also possible for a researcher to be exposed to a risk of obtaining insignificant results of the study if it is severely underpowered. Anyway, we must always bear in mind there are always highly relevant and/or valid reasons for setting a very high effect size when conducting correlation studies. One of the plausible reasons is to establish an exceptionally high level of dependence between the 2 variables, namely the predictor and outcome variables. In this circumstance, the aim is not to check for the fulfillment of the assumption of independence for the regression model but instead to determine a high-accuracy model to enable it to predict an outcome [17].

Sample size statement

This study proposes the formulation of sample size statements based on the recommendation from a previous study [18]. Bujang in his paper recommends 5 steps (Step 1: To understand the objective of a study, Step 2: To choose the appropriate statistical analysis, Step 3: To estimate or calculate the sample size, Step 4: To make allowances to cater for the possibility of non-response, Step 5: To write down a sample size statement for its sample size determination. However, reporting type I error and power of the study is not necessary in this case since sample size calculation is based on the 95% confidence interval width and assumed correlation coefficient.

The examples are as follows:

Sample size statement based on Pearson's correlation

This study aims to determine a correlation between variable A and variable B. Both variables are measured at least on an interval scale and parametric assumptions are assumed. Therefore, Pearson's correlation test will be used for analysis. By setting the correlation coefficient at 0.7 with a 95% confidence interval of width 0.3 (0.55, 0.85), therefore the minimum required sample size is 49. To incorporate a non-response rate of 20.0%, a minimum sample size of 62 respondents will be required.

Sample size statement based on Spearman's rank correlation

This study aims to determine a correlation between variable C and variable D. Both variables are measured at least on an interval scale. Based on the findings of a previous study, the data did not fulfill the assumptions necessary for performing a parametric test (state the reference). Therefore, Spearman's rank correlation test will be used for its analysis. By setting the correlation coefficient at 0.7 with a 95% confidence interval of width 0.3 (0.55, 0.85), therefore the minimum required sample size is 60. To incorporate a non-response rate of 20.0%, a minimum sample size of 75 respondents will be required.

Sample size statement based on Kendall's Tau correlation

This study aims to determine a correlation between variable E and variable F. Both variables are measured on an ordinal scale. Based on the findings of a previous study, the pattern correlation of the 2 variables had shown a monotonic pattern (state the reference). Therefore, Kendall's Tau-b correlation test will be used for analysis. By setting the correlation coefficient at 0.7 with a 95% confidence interval of width 0.3 (0.55, 0.85), therefore the minimum required sample size is 24. To incorporate a non-response rate of 20.0%, a minimum sample size of 30 respondents will be required.

CONCLUSIONS

This paper proposes and recommends a simplified yet highly useful guide to determine the sample size requirements for correlation analyses based on the estimation by a 95% confidence interval. Different target effect sizes can be used depending on a specific scenario. In general, a minimum sample size of 149 is usually deemed adequate for performing both parametric and non-parametric correlation analyses to establish at least

https://doi.org/10.5395/rde.2024.49.e21

a moderate to excellent level of correlation. It must be noted that each statistical analysis will command a different sample size requirement even for the same effect size. Therefore, it is always recommended to check for the validity of the necessary assumptions in the data involved by ensuring they have fulfilled all the necessary assumptions to perform a particular statistical analysis, to determine which types of statistical analysis shall provide the scientifically valid approach for estimating the minimum sample size required for the study.

ACKNOWLEDGEMENTS

We thank the Director General Ministry of Health for permitting us to publish the paper.

REFERENCES

1. Pearson K. Notes on the history of correlation. Biometrika 1920;13:25-45. CROSSREF
2. Rodriguez RN. Correlation. In Kotz S, Johnson NL, editors. Encyclopedia of statistical sciences. New York, NY: Wiley; 1982. p193-204.
3. Kendall M. A new measure of rank correlation. Biometrika 1938;30:81-93. CROSSREF
4. Bonett DG, Wright TA. Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. Psychometrika 2000;65:23-28. CROSSREF
5. Moinesterm M, Gottfried R. Sample size estimation for correlations with pre-specified confidence interval. TQMP 2014;10:124-130. CROSSREF
6. Bujang MA, Nurakmal B. Sample size guideline for correlation analysis. World J Soc Sci Res 2016;3:37-46. CROSSREF
7. NCSS, LLC. PASS 2022 Power Analysis and Sample Size Software. Kaysville, UT: NCSS, LLC.; 2022.
8. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. J Clin Diagn Res 2016;10:YE01-YE06. PUBMED | CROSSREF
9. Bujang MA, Baharum N. A simplified guide to the determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. Arch Orofac Sci 2017;12:1-11.
10. Bujang MA. An elaboration on sample size planning for performing a one-sample sensitivity and specificity analysis by basing on calculations on a specified 95% confidence interval width. Diagnostics (Basel) 2023;13:1390. PUBMED | CROSSREF
11. Young DS. Handbook of regression methods. Boca Raton, FL: CRC Press; 2017.
12. Shrestha N. Detecting multicollinearity in regression analysis. Am J Appl Math Stat 2020;8:39-42. CROSSREF
13. Cohen J. A power primer. Psychol Bull 1992;112:155-159. PUBMED | CROSSREF
14. Belinda B, Peat J. Medical statistics: a guide to SPSS, data analysis, and critical appraisal. 2nd ed. Oxford: Wiley; 2014.
15. El Hangouche AJ, Jniene A, Aboudrar S, Errguig L, Rkain H, Cherti M, et al. Relationship between poor quality sleep, excessive daytime sleepiness and low academic performance in medical students. Adv Med Educ Pract 2018;9:631-638. PUBMED | CROSSREF
16. Ottaviani FM, Marco AD. Multiple linear regression model for improved project cost forecasting. Procedia Comput Sci 2022;196:808-815. CROSSREF
17. Chu M, Nguyen T, Pandey V, Zhou Y, Pham HN, Bar-Yoseph R, et al. Respiration rate and volume measurements using wearable strain sensors. NPJ Digit Med 2019;2:8. PUBMED | CROSSREF
18. Bujang MA. A step-by-step process on sample size determination for medical research. Malays J Med Sci 2021;28:15-27. PUBMED | CROSSREF