World Applied Sciences Journal 27 (2): 285-287, 2013
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2013.27.02.889

Heuristics for Sample Size Determination in Multivariate Statistical Techniques

Kamran Siddiqui

DHA Suffa University, Karachi, Pakistan

Abstract: This paper aims to present the guidelines given in the literature as to the appropriate sample size for the various statistical techniques (Factor Analysis, Regression Analysis, Conjoint Analysis, Canonical Correlation, Cluster Analysis and Structural Equation Modeling). Sample size estimation depends on the nature of research and statistical technique to be employed in research. Most of the statistical techniques are sample size sensitive. (a) The chi-square is sensitive to sample size; its significance becoming less reliable with sample sizes above 200 or less than 100 respondents. In large samples, differences of small size may be found to be significant, whereas in small samples even sizable differences may test as non-significant. (b) For factor analysis appropriate sample sizes depend upon the numbers of items available for factor analysis; for 10 items a sample size of 200 is required; for 25 250; for 90 items 400 and for 500 items a sample size of 1000 deemed necessary. (c) For multiple regression analyses the desired level is between 15 to 20 observations for each predictor variable. (d) Sample size for conjoint studies generally ranges from about 150 to 1,200 respondents; for non-comparative group a sample size of 300 respondents seems reasonable while for comparative groups 200 respondents for each group are required. (e) For SEM at least 15 cases per measured variable or indicator are needed (f) There is no rule of thumb for minimum sample size for Cluster Analysis.

Key words: Sample Size · Multivariate Analysis · Regression Analysis · Factor Analysis · Cluster Analysis · Conjoint Analysis · Chi Square · Canonical Correlation · Structural Equation Modeling

INTRODUCTION

Sample size estimation is the most difficult task of research study employing statistical techniques. Most of the methods to estimate the sample size are based on statistics and requires high level of computations. Experienced researchers developed heuristics for quickly determining sample sizes based on experience, rules-of-thumb and budget constraints. This paper offers reliable literature for successful and meaningful sample-size determination.

Literature Review: There are a number of different guidelines given in the literature as to the appropriate sample size needed for a typical type of research and required for the various statistical techniques i.e., Chi-Square, Factor Analysis, Multiple Regression Analysis, Conjoint Analysis, Cluster Analysis and Structural Equation Modeling.

Sample Size for Chi-square x^2: Despite numerous ambiguities associated with interpreting x^2, the value of the statistic itself holds the most promise for the development of an index of fit for which the sampling distribution is known [1]. It also forms the basis for nested model comparison and these values must be accompanied by the values of degrees of freedom and sample size. However, x^2 is sensitive to sample size; as the sample size increases (generally above 200); the x^2 test statistic has a tendency to indicate a significant probability level. In contrast, as the sample size decreases (generally below 100); the x^2 test statistic is prone to Indicate non-significant probability levels [2].

The chi-square is sensitive to sample size; its significance becoming less reliable with sample sizes outside this range. This makes it more likely to reject the model in structural equation modeling due to the larger the sample size. In large samples, differences of small size may be found to be significant, whereas in small samples even sizable differences may test as non-significant [3].

Sample Size for Factor Analysis: Different guidelines are available for judging the adequacy of the sample size for factor analysis. Comery and Lee describe it as more is better [4].

Correseonding Author: Dr. Kamran Siddiqui, DHA Suffa University, Karachi, Pakistan.

Table 1: Sample Size Classification

Sample Size	Quality
50	Very Poor
100	Poor
200	Fair
300	Good
500	Very Good
1,000	Excellent

Table 2: Appropriate Sample Size

Number of items	Sample Size
10	200
25	250
90	400
500	700-1000

Thompson, (2004) suggests that at least 200 respondents must be sampled in order to attain a stable solution through factor analysis [5]. Most promising work on the subject suggest that adequate sample size is not simple as listed above and it needs to be quite large (e.g., 400 or greater) to produce undistorted results. Meyers, Gamst, Guarino, (2006) suggest that appropriate sample sizes depend upon the numbers of items available for factor analysis as shown in Table 2 [7].

Sample Size for Regression Analysis: For multiple regression analyses the desired level is between 15 to 20 observations for each predictor variable [8]. However, if a stepwise procedure is employed, the recommended level must be increased to acquire a reasonable level of generalisability from the results. For example the predictor variables are 30, which require a sample size of 600, but raising this to a sample size of 1,000 will considered adequate as further analyses may require step-wise regression.

Sample Size for Conjoint Analysis: Sample size for conjoint studies generally ranges from about 150 to 1,200 respondents and it largely depends on the purpose of research [9]. For robust quantitative research where subgroup comparison is not the purposes, at least 300 respondents are required to make a meaningful insight. For investigational work and developing hypotheses about a market, between thirty and sixty respondents may do. If the purpose of research is to compare groups of respondents and detect significant differences, sample size to accommodate a minimum of about 200 per group would be required. Therefore, for a study planning to divide respondents into as many as four groups (i.e., through cluster analysis) it would be wise to include, at a minimum, $4 \times 200 = 800$ respondents.

Sample Size for Canonical Correlation: The sample size depends on reliability of variables. For variables with reliability around 0.8 about 10 cases are needed for every variable. For very high reliability higher a much lower ratio of cases to variables is acceptable [10].

Sample Size for Structural Equation Modeling: Sample size plays an important role in the estimation and interpretation of SEM results [11]. In general the literature suggests that sample sizes for structural equation models commonly run in the 200 to 400 range for models with 10 to 15 indicators. At least 100 cases are required for SEM and preferably 200 [14]. With more than 10 variables, sample sizes under 200 generally cause parameter estimates to be unstable and the tests of statistical significance to lack power. Furthermore, Loehlin explained a rule of thumb i.e., the sample size should be at least 50 more than eight times the number of variables in the model. Another rule of thumb is to have at least 15 cases per measured variable or dicator [13], while the minimum requirements has been defined as s five cases per parameter estimate [15]. For example a study that uses a maximum of 50 variables in any single model which requires a sample size of about 450 as per the guidelines provided above [16]. For smaller sample size or excessive kurtosis in SEM, the researcher should report the results in maximum likelihood (ML) estimation method as ML performs reasonably well under a variety of less-than-optimal analytic conditions [17].

Sample Size for Cluster Analysis: Unfortunately, there is no generally accepted rule of thumb regarding minimum sample sizes or the relationship between the objects and the number of clustering variables used [18]. This could be considered as one of the limitations of Cluster Anlysis as is it will always render a result no matter how many varaiables are cluterised with how many cases. However bigger sample size is needed to provide valid results.

CONCLUSIONS

Sample size estimation depends on the nature of research and statistical technique to be employed in research. Most of the statistical techniques are sample size sensitive. (a) The chi-square is sensitive to sample size; its significance becoming less reliable with sample sizes above 200 or less than 100 respondents. In large samples, differences of small size may be found to be significant, whereas in small samples even sizable differences may test as non-significant. (b) For factor

analysis appropriate sample sizes depend upon the numbers of items available for factor analysis; for 10 items a sample size of 200 is required; for 25 items 250; for 90 items 400 and for 500 items a sample size of 1000 deemed necessary. (c) For multiple regression analyses the desired level is between 15 to 20 observations for each predictor variable. (d) Sample size for conjoint studies generally ranges from about 150 to 1,200 respondents; for non-comparative group a sample size of 300 respondents seems reasonable while for comparative groups 200 respondents for each group are required. (e) For SEM at least 15 cases per measured variable or indicator are needed. In order to summarize we can say that for every additional variable requires an over-proportional increase in observations to ensure valid results. This is based on the observations from different analytical tools.

Practical Implications: Young social scientists and doctoral candidates are always inquisitive about the adequacy of sample size for their research where they might be applying sophisticated multivariate techniques. This research paper also highlights the limitation of various sample size determination formulae practiced in academia. As these formulae can only provide adequate sample size under normal circulastances.

Research Limitations: Exposure of author is limited to social sciences only therefore application of these findings may be more generalizeable for social sciences. Secondly this research paper only provides the guidelines for sample size estimation needed for statistical techniques rather than mathematical computation.

Originality/value: Although there are other studies providing a numerical solution to the problem; this paper answers to the demanding question from less-experienced researchers who would like to have some rule of thumb to decide upon. By contributing to the body of knowledge in this area, this research adds significant value.

REFERENCES

1. Hoyle, R.H., Ed, 1995. Structural Equation Modeling. Concepts, Issues and Applications, Thousand Oaks, CA: Sage Publications.
2. Schumacker, R.E, 1996. Editor's Note, Structural Equation Modeling, 3(1): 1-3.
3. Fan, X., B. Thompson and L. Wang, 1999. Effects of sample size, estimation method and model specification on structural equation modeling fit indexes. Structural Equation Modeling, 6: 56-83.
4. Comrey, A.L. and H.B. Lee, 1992. A first course in factor analysis, 2nd Ed.. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
5. Thompson, B. 2004. Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington, DC: American Psychological Association.
6. MacCallum, R.C., K.F. Widaman, K.J. Preacher and S. Hong, 2001. Sample Size in Factor Analysis: The Role of Model Error. Multivariate Behavioral Research, 36(4): 611-637.
7. Meyers, L.S., G. Gamst, A.J. Guarino, 2006. Applied Multivariate Research: Design and Interpretation, Sage Publications.
8. Hair, J.F., W.C. Black, B.J. Babin, R.E. anderson and R.L. Tatham, 2006. Multivariate Data Analysis, 6th Ed., Prentice Hall, New Jersey.
9. Orme, B., 2010. Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research. 2nd Ed, Madison, Wis.: Research Publishers LLC.
10. Tabachnick, B.G. and L.S. Fidell, (1996). Using Multivariate Statistics. 3rd Ed. New York: Harper-Collins.
11. Hair, J.F., W.C. Black, B.J. Babin, R.E. anderson, R.L. Tatham, 2006. Multivariate Data Analysis, 6th Ed., Prentice Hall, New Jersey.
13. Stevens, J.P., 2002. Applied Multivariate Statistics for Social Science, 4th Ed. Hillsdale, New Jersey.
14. Loehlin, J.C., 1992. Latent Variable Models: An introduction to factor, path and structural analysis, 2nd Ed, Hillsdale, New Jersey.
15. Bentler, P.M., C.P. Chu, 1987. Practical issues in structural modelling, Sociological Methods and Research, 16: 78-117.
16. Marsh, H.W., J.R. Balla and R.P. McDonald, 1988. Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103: 391-410.
17. Hoyle, R.H. and A.T. Panter, 1995. Writing About Structural Equation Models. In Hoyle, R.H. (Ed.) (1995). Structural Equation Modeling. Concepts, Issues and Applications, Thousand Oaks, CA: Sage Publications.
18. Dolnicar, S., 2002. A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation, ANZMAC Conference, Deakin University, Melbourne.