THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Scientific Research Sample Size Determination

Bostley Muyembe Asenahabi
Lecturer, Department of Physics and Computer Science, Alupe University, Kenya
Peters Anselemo Ikoha
Lecturer, Department of Information Technology, Kibabii University, Kenya

Abstract:
Determining the appropriate sample size is an essential factor for an empirical study. A sample size that is too small will not yield valid results or adequately represent the reality of the population under study. On the other hand, despite a large sample size yielding a smaller margin of error and being more representative, it significantly increases the time and cost required to conduct the study. Specific sampling techniques are used for specific research problems since one technique may not be applicable to all problems. Similarly, if the sample size is inappropriate, it may lead to an erroneous conclusion. The complexity of a study population determines the sampling strategy to be used. A given study may require more than one sample size determination approach. This review paper discusses the various methods that can be used to determine an appropriate sample size.

Keywords: *Sampling, sample size, sampling formula, published tables, z-score and sampling strategy*

1. Introduction

A sample size is a portion of a population selected to represent the entire population in a study. The sample size is achieved by selecting a number of observations to be used in a study. Different factors have to be put into consideration while determining the appropriate sample size, including the purpose of the study, population size, and the risk involved in using a wrong sample size and the allowable sampling error (Israel, 2003).

To determine an appropriate sample size, the researcher has to consider the level of precision/sampling error, the confidence level and the degree of variability (Singh & Masuku, 2014). The level of precision is the degree of accuracy with which a parameter is estimated by a researcher. Sampling error occurs when a researcher does not select a sample that represents the entire population of data. In most cases, it is expressed as a percentage, for instance, ± 5 percent. If a researcher finds 70% of students in a certain university own smartphones with a precision level of ± 5%, then it can be concluded that 65% to 75% of the learners have smartphones.

The confidence level is based on the idea of Central Limit Theorem – when a population is repeatedly sampled, the mean value of the attribute obtained by these samples is equal to the true population value. Additionally, the values obtained by these samples are normally distributed about the mean value, with some samples having a higher value while others having a lower value than the mean. If a 95% confidence level is selected for a study, then 95 out of 100 samples will have the true population value within the precision range.

The degree of variability in the variables being measured refers to how spread out a group of data is with respect to the mean. If a population is heterogeneous, a study requires a large sample size to obtain a given level of precision. A proportion of 50% indicates a greater level of variability than either 20% or 80%. The recommended variance value is .50 (Krejcie & Morgan, 1970). This proportion results in the maximization of variance, which will also produce the maximum sample size.

2. Sample Size Determination

Different methods can be used to determine the sample size scientifically. A census is recommended for small populations. A researcher can adopt a sample size used in previous similar studies. Besides, a researcher may use a formula, a percentage or a published table to calculate a sample size. This section discusses the different strategies.

2.1. Census

Census calls for a researcher to collect data from the entire population. This is an approach adopted for small populations, presumably less than 200 or when the data for each entity within the population has to be factored in the study. Cost implications make it difficult for the census to be used for large populations. On the other hand, census eliminates sampling error.

2.2. Sample Size of a Similar Study

A researcher may adopt a sample size previously used in a similar study based on the review of literature in the discipline under study. A literature review can guide a researcher on the relevant sample sizes previously used in similar

studies. The main undoing of this approach is that if the procedure employed in arriving at the sample size is not reviewed, the researcher may end up repeating the errors that were made in determining the sample size (Israel, 1992).

2.3. Published Tables

Published tables can be used by a researcher to determine the sample size of a study. These tables are generated based on the desired precision, confidence level and variability (p). The sample sizes in published tables reflect the number of obtained responses to take part in the study.

The sample size for ± 5%, ± 7 % and ± 10 % precision levels where the confidence level is 95% and p = 0.5.

Population Size	**Sample Size (n) for Precision (e) of:**		
	± 5%	**± 7 %**	**± 10%**
100	81	67	51
200	134	101	67
300	172	121	76
400	201	135	81
500	222	145	83
600	240	152	86
700	255	158	88
800	267	163	89
900	277	166	90
1 000	286	169	91
2 000	333	185	95
3 000	353	191	97
4 000	364	194	98
5 000	370	196	98
6 000	375	197	98
7 000	378	198	99
8 000	381	199	99
9 000	383	200	99
10 000	385	200	99
15 000	390	201	99
20 000	392	204	100
25 000	394	204	100
50 000	397	204	100
100 000	398	204	100
> 100 000	400	204	100

Table 1: Sample Size for Different Precision Levels
Adopted from Yamane (1967)

If the population size is less than 500, the attributes being measured have to be normally distributed or nearly so. Otherwise, the entire population may be surveyed.

2.4. Using a Percentage of the Accessible Population

A percentage of the accessible population can be used by a researcher to determine a sample size with much ease. Different authors have provided their insights in relation to the use of a percentage to determine sample size.

Singh (2006) suggests that a researcher should select 10 – 20 percent of the accessible population for the sample.

According to Mugenda and Mugenda (2013), if the accessible population is less than 10,000, a researcher can select 10 – 30 percent to be the representative sample size.

Additionally, Kothari (2004) points out that a representative sample should have at least 10 percent of the population.

2.5. Using Formulae

Formulae can be used to determine sample size. They enable a researcher to calculate the required sample size using different combinations of levels of precision, confidence and variability. In most instances, the confidence level is converted into a z-score. Table 2 shows the z-score for the most common confidence levels:

Confidence level (α)	z-score
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58

Table 2: z-score Values for Different Confidence Levels

The margin of error for *α* confidence interval is *z-score* times the standard error. For instance, a 95% confidence interval is 1.96 times the standard error.

2.5.1. Sample Size Calculation, If Mean Is the Parameter of Research

The confidence interval is an estimate of the margin of error. It shows the accuracy of the estimate and is based on the variability of the estimate.

$$n = \frac{Z^2 \times \sigma^2}{e^2}$$ Equation 1

where:

- n = sample size
- z = abscissa of the normal curve that cuts off an area α at the tails (1 - α equals the desired confidence level, e.g., 95%). The value for Z is found in statistical Z-tables, which contain the area under the normal curve.
- e = margin of error
- α = mean

2.5.2. Cochran Formula

It is used when the proportion is the parameter of research. To get a sample size for a large population, Cochran (1963) developed Equation 1 to yield a representative sample for proportions. It enables a researcher to get an ideal sample size given a desired level of precision, desired level of confidence and the estimated proportion of the attribute present in the population. A sample of any given size provides more information about a smaller population than a larger one, so there is a variation through which the number given by Cochran formula can be reduced if the whole population is relatively small.

The Cochran formula is:

$$n_o = \frac{Z^2 pq}{e^2}$$ Equation 2

Where:

- n_o is the sample size
- *e* is the desired level of precision/ confidence interval/margin of error
- *p* is the estimated (proportion of the) population which has the attribute in question
- *q* is 1 - *p*
- *z* is the abscissa of the normal curve that cuts off an area α at the tails (1 - α equals the desired confidence level, e.g., 95%).

The value for Z is found in statistical Z-tables, which contain the area under the normal curve.

Cochran's Formula Example:

Consider a study being done on students learning computing in universities and the researcher wants to find out how many of these learners will end up being professional computer programmers, this being one of the skills learnt in computing courses. The researcher does not have much information on the subject, to begin with, and thus assumes that half of the learners have an interest in computer programming. This gives maximum variability, p = 0.5. The researcher wants 95% confidence and at least ± 5% precision. A 95% confidence level gives a Z-value of 1.96, per the normal tables. Thus, the sample size is:

$$n_o = \frac{Z^2 pq}{e^2} = ((1.96)^2. (0.5). (0.5)) / (0.05)^2 = 385$$

2.5.3. Yamane Formula

A different equation by Yamane (1967) can be used to determine the sample size. This formula requires the researcher to know the study population before calculating the sample size.

The Yamane Formula is:

$$n = \frac{N}{1+N(e)^2}$$ Equation 3

Where:

- n is the sample size
- *N* is the population size
- *e* is the level of precision

If we have a study population of 75 000 respondents, a 95% confidence level and p = 0.5, the appropriate sample size will be:

$n = \frac{N}{1+N(e)^2} = \frac{75\,000}{1+(75\,000\,(0.05^2))} = 398$

2.5.4. Lehr's Formula

Lehr's formula can be used for quick calculations of sample size for the comparison of two equal-sized groups and the research study is to be verified through t-tests and Chi-squared tests (Lehr, Fraga, Belen & Cekirge, 1984).

Lehr's formula is:

$$n = \frac{16}{Standardized\ difference^2} \quad \text{Equation 4}$$

For the power of 80%, the numerator is 16; for the power of 90%, it becomes 21.

The standardized effect is the ratio of effect size to SD.

The main undoing of this formula is that if the standardized difference is small, the sample size is overestimated.

2.5.5. Slovin's Formula

Slovin's formula is used to calculate an appropriate sample size from a population. It is used in a study when the researcher does not know how the attributes being measured are distributed within the study population.

Slovin's formula is:

$$n = \frac{N}{1+N(e)^2} \quad \text{Equation 5}$$

Where:

- n is the sample size
- *N* is the total population
- *e* is the level of precision

2.5.6. Andrew Fisher's Formula

Fisher's formula is used to calculate the sample size of a study. To use this formula, the researcher needs to:

- Determine the confidence interval
- Determine the confidence level
- Determine the standard deviation (a standard deviation of 0.5 is a safe choice if the figure is not known)
- Convert the confidence level into a z-score, as shown in table 2.

Fisher's Formula is:

$$\text{Sample size} = \frac{z\text{-}score^2 \times StdDev \times (1-StdDev)}{Confidence\ Interval^2} \quad \text{Equation 6}$$

3. Recommendation

Determining the specific sampling strategy is a science that requires critical thinking techniques. The complexity of a study population determines the sampling strategy to be used as a given study may require more than one sample size determination approach. For instance, to get a representative sample size from a study population spanning an entire country, a researcher can divide the country into regions and get representative regions using a percentage of the regions. A formula can then be applied to get the exact number of respondents from the representative regions.

4. Conclusion

Determining an appropriate sample size is essential for an empirical study, especially if the researcher intends to infer about the population from a sample. This review paper discusses the different methods:

- Census,
- Sample size used in previous related studies,
- Use of a percentage of the study population, and
- Use of formulae and tables through which a researcher can determine an appropriate sample size for a specific study

The complexity of a study population determines the sampling strategy to be used.

5. References

i. Cochran, W. G. (1963). *Sampling Techniques* (2nd ed.). New York: John Wiley and Sons, Inc.
ii. Israel, G. D. (1992). *Sampling the Evidence of Extension Program Impact. Program Evaluation and Organizational Development.* University of Florida, IFAS.
iii. Israel, G. D. (2003). *Determining Sample Size.* PEOD6, University of Florida, IFAS.
iv. Kothari, C. R. (2004). *Research Methodology: Methods and Techniques.* New Age International.
v. Krejcie, R., and Morgan, D. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*, 607–610.
vi. Lehr, W., Fraga, R. J., Belen, M. S., and Cekirge, H. M. (1984). A new technique to estimate initial spill size using a modified Fay-type spreading formula. *Marine Pollution Bulletin, 15*(9), 326–329.
vii. Mugenda, A., and Mugenda, O. (2013). *Research Methods: Quantitative and qualitative approaches.* Nairobi: ACTS press.

viii. Singh, A., and Masuku, M. (2014). Sampling Techniques and Determination of Sample Size in Applied Statistics Research: AN Overview. *International Journal of Economics, Commerce and Management, 2*(11), 1–22.
ix. Singh, Y. K. (2006). *Fundamental of Research Methodology and Statistics.* New Delhi: New Age International Publishers.
x. Yamane, T. (1967). *Statistics, An Introductory Analysis* (2nd ed.). New York: Harper and Row.