Chapter 2

Determination of Appropriate Sample Size for a Research

2 Determination of Appropriate Sample Size for a Research

Felix Kutsanedzie[1]; Sylvester Achio[1]; Edmund Ameko[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

In research studies, it is often difficult if not impossible to use the whole population for a study. This is however different when the study covers population that is relatively small that the whole population can be considered for the study. Otherwise it is expensive, sometimes the destructive nature of the study requires that only small fraction of the population referred to as a sample which is used to make inferences about the population understudy. Most often would–be researchers fall short of not considering the appropriate sample size for a study and thus the inferences made about the population are misconstrued. This paper addresses the challenge of giving a comprehensive understanding of how an appropriate size of sample can be taken from a population for a study. It explains the various calculations involved in sizing a sample of a given population size.

Keywords

Sample, Population, Representative, Size, Research

2.1 Introduction

Sample is defined as a fraction of a population taken that is considered for a study in order to make inferences about the population. For appropriate inferences to be made of the population using a selected sample requires that the sample in question should be representative in terms of its composition and size. In this regard the sampling technique to be used for selecting the sample as well as the way the sample size is to be selected from the population are of equal importance in constituting the representative sample. Sample size determination is the act of choosing the number of observations or replicates that should be included in a statistical sample. The sample size to use for a study depends on the data to be collected and the statistics that is required to be derived from it.

Sample sizes to be taken from a population depend on the expediency and availability of data. A sample size can result in wide confidence interval or risks of error in order to increase precision of the data to be collected but in other cases the accuracy for using larger sample sizes cannot be guaranteed because of systematic errors. All these notwithstanding, the law of large numbers and central limit theory underpin the use of large size samples for increasing statistical power of a selected sample.

In selecting the right sample size one needs to understand the concepts of confidence level, confidence interval as well as margin of error. Confidence level is expressed in percentage and it informs the percentage of confidence ascribed to obtaining the true mean of a selected sample being considered for a study. For instance confidence level of 95% or 0.95 means that there is a surety that 95% of the true mean of the sample would be obtained when the study is repeated for 100 times within an interval called the confidence interval. The

confidence interval is the range within which the true mean is expected to be located at the chosen confidence level.

2.2 Determination of Sample Size

In order to determine the sample size, parameters such as the margin of error and the confidence interval need to be considered. Usually when data is collected and the sample mean is calculated, it tends to differ from the population mean, and this difference between the two is termed the margin of error. The margin of error is simply the maximum difference between the observed sample mean ($\bar{x}$) and the true value of the population mean (μ). The margin of error is mathematically expressed as:

$$E = \frac{Z_{\frac{\alpha}{2}}}{(2\sqrt{n})}$$

To determine the margin of error, a confidence level must be chosen and this is often a value less than 100%, however mostly 99%, 95% and 90% is used. However, 99% is mostly used for medical experiments that require higher precision whereas 95% for other situations. Once the confidence level is chosen, say 95%, the α is derived by subtracting the chosen confidence level from 100% and expressing it in units:

$$\propto = 100\% - 95\% = 5\% = 1 - 0.95 = 0.05$$

From the normal distribution diagram, the total area under the curve is equal to 1; that under symmetric half of the curve is equal to 0.5; and the area of $Z_{\frac{\alpha}{2}}$ both to the left or right coloured red is $Z_{\frac{5}{2}} = 0.025$. The region of $Z_{\frac{\alpha}{2}}$ to the left and $Z_{\frac{\alpha}{2}}$ to the right of $Z = 0$ is therefore equal to:

$$0.5 - 0.025 = 0.475$$

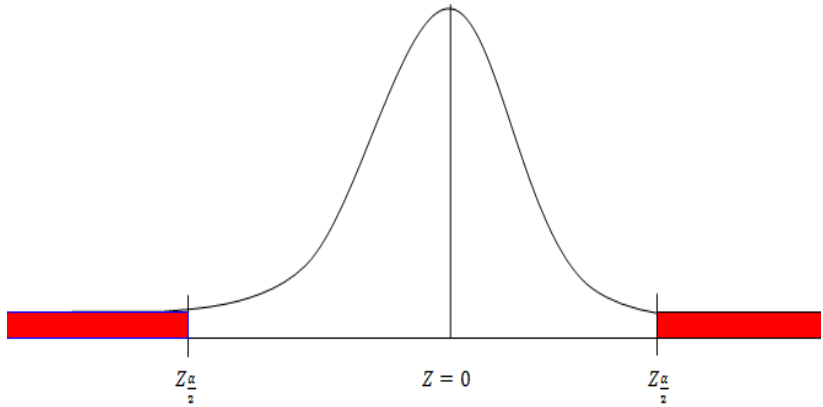

Figure 2.1 *Illustration* $\propto$ *on a normal curve.*

The critical Z value corresponding the area 0.475 under the normal distribution table is 1.9 located vertically plus 0.06 located horizontally (1.9 + 0.06 = 1.96) as show in the table below:

Table 2.1 *Areas for a Standard Normal Distribution.*

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857

Thus taking a sample size of 100, this can be substituted into the equation to obtain the margin of error as below:

$$E = \frac{1.96}{(2\sqrt{100})}$$

$$E = \frac{1.96}{(2)(10)}$$

$$E = \frac{1.96}{20}$$

$$E = 0.098 = 0.1 \; or \; 10\%$$

$$E = \pm 0.1 \; or \; \pm 10$$

$$Z_{\frac{\alpha}{2}} = 1.96, n = 100$$

However, if the margin of error is estimated at the beginning of the study as 5%, this same formula can be used to calculate the appropriate sample size to work with by making the n the subject of the equation as below:

$$n = (\frac{Z_{\frac{\propto}{2}}}{2E})^2 = (\frac{1.96}{2(0.05)})^2 = (\frac{1.96}{0.1})^2 = (19.6)^2 = 384.16 \cong 385$$

The formula $n = (\frac{Z_{\frac{\propto}{2}} \; x \; \partial}{2E})^2$, can be used when the population standard deviation is known. This might be obtained from other studies or a pilot test conducted or else might not be available at all.

For an identically and independently distributed data with a population variance of σ^2, the formula below is used in estimating the sample size required:

$$n = \frac{16\sigma^2}{W^2}$$

n=sample size, σ^2 =population variance, W=width of confidence interval (Wald's unit)

For estimating a required sample size for survey data collection, the formula below can be used:

$$n = \frac{4}{W^2} = \frac{1}{B^2}$$

n=sample size, W= width of confidence interval, B=error bound on estimate (usually given as $\pm B$ and in precentage)

Also the Mead's resource equation is used for estimating the required size of laboratory animals to be used in an experiment as well as other types of resources with regards to experimental work. Its used for estimating the sample size that might be as accurate as the other methods but is very helpful where the standard deviations or differences between values of groups being considered for an experiment are difficult or hard to estimate. Mead's resource equation is given as:

$$E = N - B - T$$

E=degree of freedom of the error component which should range between 10 and 20, N=the total number of individuals or units in the study minus 1 or df of sample size, T=the number of treatment groups including the control or the number of questions asked minus 1 or df of treatment, B=the blocking component minus 1 or df of block, however when there is no stratification B=0.

Let assume an experiment needs to be conducted where forty (40) animals of eight (8) animals each put in five (5) treatment groups. This can be worked out as follows:

$$E = (40 - 1) - 0 - (5 - 1)$$

$$E = 39 - 0 - 4$$

$$E = 35$$

The value 35 exceeds the expected range of E which is between 10 and 20. It presupposes that eight (8) is not the right sample size. Assuming 2 animals per group of 20 is considered, and then E becomes:

$$E = (40 - 1) - 0 - (20 - 1)$$

$$E = 39 - 19$$

$$E = 10$$

Therefore the required sample size can be 20.

2.3 Determination of Sample Size with Levels of Significance

Another way of determining the appropriate size of sample to be selected from a known population size for a research study is to use the levels of significance as per the formula:

$$n = \frac{N}{1 + N(level\ of\ sgf.)^2}$$

where n=sample size, N=population size, level of sgf.=level of significance in units (5%=0.05, 1%=0.01)

Thus using the formula above, and considering a population size of 200 and a level of significance of 5%, the sample size to be used can be calculated as follows:

$$n = \frac{20}{1 + 20(0.05)^2}$$

$$n = \frac{20}{1 + 20(0.0025)}$$

$$n = \frac{20}{1 + 0.5} = \frac{20}{1.5} = 19.04 \cong 19$$

When above formula is used a table below can be obtained for the various populations sizes and their respective sample sizes

Table 2.2 *Population Sizes and their calculated Sample Sizes.*

Population Size	Sample Size
20	19
40	36
60	52
80	66
100	80
150	108
200	132

Bibliography

[1] Ary, D., Jacobs, L. C., Razavieh, A. (1996). *Introduction to research in education*. Fort Worth, TX: Harcourt Brace College Publishers.

[2] Browner, W. S., Newman, T. B. (1998). Sample size and power based on the population attributable fraction. Am J Public Health, 79(9): 1289-94.

[3] Castelloe, J. (2000), "Sample Size Computations and Power Analysis with the SAS System," Paper 265-25 in *Proceedings of the Twenty-Fifth Annual SAS User's Group International Conference*, Cary, NC: SAS Institute, Inc.

[4] Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press, New York.

[5] Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

[6] Desu, M. M., and Raghavarao, D. (1990). *Sample Size Methodology*, Boston: Academic Press.

[7] Donald, M. N. (1967). Implications of non-response for the interpretation of mail questionnaire data. *Public Opinion Quarterly, 24*(1), 99-114.

[8] Fink, A. (1995). *The survey handbook*. Thousand Oaks, CA: Sage Publications.

[9] Freiman, J. A., Chalmers, T. C., Smith, H. Jr., Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and

interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med., 299(13): 690-4.

[10] Hagbert, E. C. (1968). Validity of questionnaire data: Reported and observed attendance in an adult education program. *Public Opinion Quarterly*, *25*: 453-456.

[11] Hair, J., Anderson, R., Tatham, R., Black, W. (1995). *Multivariate data analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

[12] Halinski, R. S. & Feldt, L. S. (1970). The selection of variables in multiple regression analyses. *Journal of Educational Measurement*, *7*(3), 151-158.

[13] Holton, E. H., Burnett, M. B. (1997). Qualitative research methods. In R. A. Swanson, & E. F.

[14] Houston, W. J. (1983). The analysis of errors in orthodontic measurements. Am J Orthod., 83(5): 382-90.