SAMPLE SIZE DETERMINATION: A COMPARISON OF ATTRIBUTE AND CONTINUOUS VARIABLE METHODS

UDOFIA ITON UDOFIA
EDET E. NKEREUWEM
Library Department
University of Calabar
Calabar
Nigeria

Each of the two methods of sample size determination - the Attribute and the Continuous Variable Method has its use in the investigation of social science problems. The former allows the computation of sample size with reference to any parameters of the variable and, therefore, can substitute for the Continuous Variable Method, but, with and probable increase in sample size. The later is very useful when data are collected in ratio form. However, it demands estimates of dispersion from the mean which may be primary purpose of the research in the first place. The Attribute Method is highly recommended for library and information science since it can be substituted for Continuous Variable Method.

INTRODUCTION

Researchers intending to use sampling procedures for studying library and information science subject face many and diverse methodological problems. One of these problems is the decision about the sample size such as, the minimum amount of data to be collected or whether enough cases available to be statistically valid.

To determine appropriate sample size, one must have a thorough knowledge of the levels of measurement to be used, the hypotheses to be tested and the type of statistical tests most appropriate to the problem. Some knowledge of population parameters is necessary, for example, percentage of occurrence or the standard deviation as well as desired level of confidence and desired accuracy for the sample, expressed as tolerance. Hence, a thorough idea of statistical problems and training are required to decide which of the methods namely, the Attribute and Continuous Variable Method is more suitable for library and information science surveys.

One of the simple and informal method for determination of sample size is to use the sample size used by others studying a similar problem. For example, Roscoe [4] states that in behavioral research there are few occasions when samples smaller than 30 and larger than 500 in size can be justified. While the range from 30 to 500 may appear to be a large one, it does narrow the number to some extent. Unfortunately, no rationale is given for this recommendation, and this view is shared by Uko [5] and others.

The most frequently used statistical approaches are the Attribute and the Continuous Variable Methods. In this paper a basis for comparison with the two approaches is given for calculation of sample size for library and information science surveys. Convenience alone dictated the comparison of the two approaches.

STATISTICAL DETERMINATION OF SAMPLE SIZE

To generalize from a sample to a universe or population, that is, to hypothesize that the mean of the population variable falls within a certain range of values at a certain level of confidence, statistical techniques must be used.

In discussions of the Attribute Method, and the Continuous Variable Method, found in the documents on research methodology and statistics, two key factors are always mentioned [3]. First, the need to establish the level of confidence and the second, the need to establish the degree of accuracy or tolerance that is required.

The Attribute Method deals with the significance of proportions and requires an estimate of the percentage of occurrence of the key variable in the study. The Continuous Variable Method requires
an estimate of the dispersion within the key variable, usually, the standard deviation. In both cases, estimates based on prior knowledge are required. The attribute method does permit the selection of the largest possible sample size by estimating the occurrence of the key variable at 50%.

The confidence levels establish the degree of sampling error that will be permitted in the study. For example, when a confidence level of 95% is used, it is said that the probability is 95%. That means that, most of the time, the confidence interval will contain the true mean and in 5% of the time it will not. Carpenter and Vasu [1] stated that the most commonly used confidence levels were the 95% level and might be used as a standard. The 99% level increases the sample size and the 90% level is least used. To minimize total error that results from non-response, inaccuracy of recording responses or copying figures from one file to another and so on, the confidence level should be set at a conservatively high level. The degree of accuracy required of a sample is translated into statements such as, "The population mean falls with plus or minus 5 units of the sample mean". The researcher establishes this level of tolerance through inspection of the variable in question and a need for accuracy. But, as will be shown, sample size increases dramatically for higher level of tolerance.

THE ATTRIBUTE METHOD

The Attribute Method of sample size determination requires an estimate of the proportion of occurrence of a property or activity in the universe. Dougherty and Heinritz [2] used the example of "books that have been in circulation for fourteen days or more" to illustrate the concept of attribute. In this case, there are only two possible conditions - the case has the attribute or it does not. In effect, there are only two values. Other examples would be books borrowed for five days or more than five days; students who use libraries heavily or not heavily; documents that contain ten or more citations and those containing three or less.

A formula for the computation of Attribute sample size is given by Dougherty and Heinritz [2] and Uko [5].

\[ F = \frac{C^2}{t^2} p(1-p), \]

where \( F \) = the sample size
\( C^2 \) = the z-score squared representing the desired confidence level.
\( t^2 \) = the desired tolerance expressed as a fraction or decimal
\( p \) = the estimated percentage occurrence of the attribute being measured.

The z-score squared and tolerance squared have been used for ease of computation. For those who are more familiar with the z-score notation, the following gives the value of the z-score, and \( C^2 \), for six typical confidence levels:

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Z-score</th>
<th>( C^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>2.58</td>
<td>6.6564</td>
</tr>
<tr>
<td>98%</td>
<td>2.33</td>
<td>5.4289</td>
</tr>
<tr>
<td>97%</td>
<td>2.27</td>
<td>5.1529</td>
</tr>
<tr>
<td>96%</td>
<td>2.05</td>
<td>4.2025</td>
</tr>
<tr>
<td>95%</td>
<td>1.96</td>
<td>3.8416</td>
</tr>
<tr>
<td>90%</td>
<td>1.65</td>
<td>2.7225</td>
</tr>
</tbody>
</table>

The percentage occurrence portion of the formula, \( p(1-p) \), has the property of maximizing sample size at the 50% level of occurrence for a given tolerance. Hence, sample size is the same for both \( p = 40\% \) and \( p = 60\% \) or \( p = 30\% \) and \( p = 70\% \) [2, 5].

This, however, eases the problem of deciding on the criterion variable for sample size calculation in a multivariable study. For example, a hypothetical study might have three variables - age, distance and rate of visiting. Age could be defined as having the attribute of "greater than 18 years", distance having the attribute of "more than three kilometres" and rate of visiting as, "three or more visits per month". A problem in some people's mind would be to decide which of the three should be used in calculating sample size. In the Attribute Method, estimating the percentage of occurrence at 50% would maximize sample size for any variable. If it is necessary to decrease sample size, the choice should go to that variable with the percentage closest to 50%.

If in the above example, \( p = 65\% \) for age, \( p = 30\% \) for distance and \( p = 20\% \) for rate of visiting, sample size is highest for age because it is closest to 50%.
Table 1 illustrates the results of the computation at the 95%, 98% and 99% confidence levels with tolerance set at ±5 for all variables.

The rationale for choosing age as the criterion variable rather than distance or rate of visiting is simple. By choosing the greatest sample size, all the other variables will be generalizable to the population. If the maximum size is selected, it is not even necessary to estimate the percentage occurrence of any of the other variables. In some cases, the difference between the maximum size and the size computed for a variable may be great enough to choose the lower of the two. In Table 1, under the heading 95% confidence, there is a difference of 34 cases between the maximum size of 384 and the next highest age, 350. A difference of 34 cases might be a rationale for choosing the lower figure. This rationale could be used when the cost of collecting data for each case is high. If the cost of collecting data is low, choosing the maximum sample size would negate any errors in estimation of the percentage occurrence. This assumes, of course, the same level of tolerance for all variables, but, such an assumption may not be warranted. If the tolerance in Table 1 is changed to 3% for distance, and remains at 5% for both age and rate of visiting, the sample size would be more than double to 896 and would result in the highest sample size. It is important to bear in mind that this interrelationship between tolerance and percentage occurrence exists.

When the cost of data collection is low, the Attribute Method is very good to be used, but when the cost of data collection is high, estimates of the percentage occurrence on each variable should be made. Standard tables can be constructed for various levels of confidence and tolerance with p = 50%.

**THE CONTINUOUS VARIABLE METHOD**

The Continuous Variable Method is similar to the Attribute Method with the substitution of a measure of dispersion for the estimate of the percentage occurrence.

The formula used by Uko [5] is typical:

\[
\frac{z q^2 s^2}{t^2} = n
\]

where,

- \( n \) = sample size
- \( zq \) = z-score for the confidence level
- \( s \) = standard deviation
- \( t \) = tolerance or degree of accuracy.

The Continuous Variable Method is used when the variables are in the form of a ratio scale. In this case, an estimate is needed of the standard deviation. The standard deviation can be determined by using an electronic calculator and a small random sample of 10 or 20 cases from the intended sampling frame. A sampling frame consists of a list, directory, index, maps and other records, listing the population elements from which the sample
may be drawn. Some commonly used sampling frames are the telephone directory for telephone interviews, the staff nominal roll for sampling employees of institutions, and membership lists for associations and clubs.

To our example, we might estimate the standard deviations of distance as 0.5 kilometres, age as 5 years and rate of visiting as 0.65 visits per month.

Deciding on the tolerance level is also somewhat different than the procedure used in Attribute Method. A percentage of accuracy was decided upon in the previous method. In the Continuous Variable Method, an absolute value for the degree of accuracy is required such as, ±0.5 kilometres or ± 5 years of age. The desired degree of accuracy affects sample size considerably.

Table 2

Variation in sample size given different tolerance and standard deviations for three variables.

<table>
<thead>
<tr>
<th></th>
<th>Tolerance (±)</th>
<th>Confidence level</th>
<th>Standard deviation</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Years</td>
<td>95%</td>
<td>15.5</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>3 Years</td>
<td>95%</td>
<td>15.5</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>2 Years</td>
<td>95%</td>
<td>15.5</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>1 Year</td>
<td>95%</td>
<td>15.5</td>
<td>923</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50 Kilometres</td>
<td>95%</td>
<td>1.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>0.25 Kilometres</td>
<td>95%</td>
<td>1.5</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>0.20 Kilometres</td>
<td>95%</td>
<td>1.5</td>
<td>216</td>
<td></td>
</tr>
<tr>
<td>0.10 Kilometres</td>
<td>95%</td>
<td>1.5</td>
<td>864</td>
<td></td>
</tr>
<tr>
<td>Visits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50 visits</td>
<td>95%</td>
<td>1.05</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>0.25 visits</td>
<td>95%</td>
<td>1.05</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>0.20 visits</td>
<td>95%</td>
<td>1.05</td>
<td>106</td>
<td></td>
</tr>
<tr>
<td>0.10 visits</td>
<td>95%</td>
<td>1.05</td>
<td>435</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 illustrates the differences in sample size, given various hypothetical degrees of accuracy. With the Attribute Method, the sample size of the study should be based on the variable that yields the maximum sample size. In the example shown in Table 2, the tolerance of ±0.10 visits should be used as our criterion variable, as this would allow us to generalize to the population for distance as well as age.

GENERALIZATION

The formulae for calculating the sample size can be used to solve other aspects such as, tolerance, confidence level or proportion accounted for when \( n \), the sample size, is known. For example, the attribute formula for sample size is given as:

\[
F = \frac{C^2}{t^2} p(1-p). \tag{2}
\]

solving for \( t \)

\[
t = \sqrt{\frac{C^2}{F} p (1 - p)}
\]

The tolerance can be determined for a given sample size (\( F \)), confidence level (\( C^2 \)), and percentage of occurrence. Taking the sample size of
UDOFIA ITON UDOFIA and EDET E. NKEREUWEM

500, a 95% confidence level, and the 50% occurrence level, tolerance is

\[ t = \sqrt{\frac{3.8416}{500}} \cdot (0.5) \cdot (0.5) = \pm 4.4\% \]

Tolerances can be computed for various level of confidence, sample sizes and confidence rates, using this formula.

Continuous Variable Method can be used to calculate either confidence level or tolerance, given an estimate of standard deviation and a known sample size. The Continuous Variables sample size formula is given as:

\[ n = \frac{z^2 s^2}{t^2} \]

solving for \( t \)

\[ t = \frac{Zq}{{\sqrt{n}}(s)} \]

where \( Zq \) is the z-score, \( s \) is the standard deviation, and \( n \) is the sample size.

From the example in Table 2, where the standard deviation for age was 15.5 and with a 95% confidence level, and a sample size of 500, tolerance can be calculated thus:

\[ t = \frac{1.96 \cdot 15.5}{\sqrt{500}} = \pm 1.36 \text{ years} \]

It is possible to compute the confidence level given a tolerance, sample size and standard deviation.

In using the Continuous Variable Method with 95% confidence, ±2 units of tolerance, and a standard deviation of 15.5 yields a sample size of only 231. A calculation of sample size by the Attribute Method with 95% confidence, ±5% tolerance, and 50% occurrence would yield a sample size of 384 which is 153 cases larger.

CONCLUSION

The main intention of the study is to compare Attribute Method with Continuous Variable Method. The Attribute Method is highly recommended for library and information science surveys as it allows the computation of sample size with reference to any parameters of the variable and can be substituted for Continuous Variable Method.

The choice of the 95% level of confidence is arbitrary in many respects and the estimates of the desired tolerance or the standard deviation are often equally arbitrary depending on the knowledge the researcher has of the population parameters.

The ultimate criterion for choosing sample size is cost. If the costs of gathering each unit of information are high, the method that keeps sample size at a minimum can be used and when the costs per unit are low, methods that increase understanding, planning, and flexibility are to be considered. It is the author’s contention that statistically trained researchers will probably continue to use Attribute Method as previously recommended.

REFERENCES


