Sample size

The issue of sample size is a matter often concerns researches and all those who are related to formal research activity. **Sample size- optimum**

1. Efficiency
2. Reliability
3. Flexibility and Representativeness

Factors determine the sample size

Criteria	**Sample size**	
	Large	**Small**
Nature of population	Heterogeneous	Homogeneous
Nature of respondents	Non- cooperative	Easily accessible
Nature of study	One time study	Continuous
Complexity of tabulation	Larger the number of categories	Small number of categories
Problems related with collection of data	Time and money Problems are high	Lower the problems
Type of sampling	Multistage sampling	Stratified sampling
Degree of accuracy Law of statistical regularity Law of inertia of large numbers	Increases	Decreases
Desired level of confidence	S . E decreases	S . E Increases

A sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data and the results found in the sample do not represent the results that would be obtained from the entire population. Sampling is an analysis performed by selecting a number of observations from a larger population, and the selection can produce both sampling errors and non-sampling errors.

A sampling error is a deviation in sampled value versus the true population value due to the fact the sample is not representative of the population or biased in some way. Even randomized samples will have some sampling error since it is only an approximation of the population from which it is drawn. Sampling errors can be eliminated when the sample size is increased and also by ensuring that the sample adequately represents the entire population.

What is a Non-Sampling Error?

A non-sampling error is a statistical term that refers to an error that results during data collection, causing the data to differ from the true values. A non-sampling error differs from a sampling error. A sampling error is limited to any differences between sample values and universe values that arise because the sample size was limited. (The entire universe cannot be sampled in a survey or a census)

Two causes for incorrect inference resulting from the data.

1. **Non-sampling error (Systematic bias)**

Non-sampling errors are other errors which can impact the final survey estimates, caused by problems in data collection, processing, or sample design. It cannot be reduced or eliminated by increasing the size of the samples. They include:

1. **Unsuitable sample frame** or source list.
2. **Over coverage**: Inclusion of data from outside of the population.
3. **Under coverage**: Sampling frame does not include elements in the population.
4. **Measurement error**: e.g. when respondents misunderstand a question, or find it difficult to answer.
5. **Processing error**: Mistakes in data coding.
6. **Indeterminacy principle**: it means that the individuals act differently when kept under observation than when they do when kept in non-observed situation.
7. **Non-response**: Failure to obtain complete data from all selected individuals.

Sampling errors and biases

Sampling errors and biases are induced by the sample design. Sampling errors can be decreased by increasing the sample size. They include:

1. **Selection bias**: When the true selection probabilities differ from those assumed in calculating the results.

2. **Random sampling error**: Random variation in the results due to the elements in the sample being selected at random.

Types of sampling errors

1. **Biased errors:**

These are errors, which are occurring due to the faulty selection of the sampling method due to prejudice of the researcher. It is also known as cumulative errors or non – compensating errors. Such type of errors does not decrease when the sample size is increased.

Causes of bias

1. Faulty process selection
2. Faculty work during the collection of information.
3. Faulty method of analysis.

1 Faulty Process Selection

Faulty selection of sample may arise due to the following ways.

1. Deliberate selection of the sample by the researcher.
2. Possibility of conscious or unconscious bias in the selection of the random sample. The researcher or investigator claims that the sample is selected randomly even though it not selected randomly.
3. If an item chosen through random sampling is substituted by another item, it will be a source for bias.

4. If all items, which are selected randomly, are covered for study it will be a cause for bias.
5. If the respondents are not filling the entire questionnaire and return it the researcher, it will be a reason for creating bias.

2. Faulty work during the collection of information

The regular error in measurement will be a reason for bias for bias. If the error in measurement is occurred in samples, it will be create much bigger errors. Bias may arise due to wrong definition of population, improper formulation of decision, securing inadequate source list etc. Improper design of questionnaires, interview schedule, ill trained interviewer, etc will also make bias.

3. Faulty methods of analysis

If proper methods of analysis are not adopted, it will be the reasons for bias.

2. Unbiased errors

Unbiased errors are those which arise just on account of chance. They are not the results of any prejudice or bias. If figure are approximated to the nearest whole number, the error would be unbiased, as in some cases the approximated number would be less than the actual ones while in others they would be more than the true values.

Unbiased errors are generally compensating. One error compensates the other. The law of statistical regularity works here and since errors are both positive and negative they usually cancel each other. It is also known as random sampling errors or non-cumulative error.ie. Increase the size of the sample can decrease the error. Combining the two findings above, we have the following general rules related to sampling error of means.

1. As the size of the random sample increases, the amount of sampling error of means decreases.

2. As the variability in the population increases, the amount of sampling error of means increases.

Respondent errors

1. If the respondent fails to provide information, we call it non-response error.
2. If the respondent gives wrong information, we call it response-bias.

Sometimes, respondent do not exactly understand what specific information is wanted. In some cases the respondents may not have the required knowledge or they may not reveal the truth.

Administrative errors

The errors that have arisen due to improper administration of the research process are called administrative errors.

1. There may be errors in sample selection.

2. Eliciting information indirectly from persons other than the real respondents may provide incorrect data.

3. Due to poor structuring of the questions, the response may be biased towards some specific direction.

4. Errors may occur while editing, coding and entering the data into computer for further processing and analysis.

Control of error

It is not possible to eliminate completely the sources of errors. However, a researcher objectives and effort should be to minimize the sources of errors as much as possible. Some of the ways of reducing errors are as follows:

1. Designing and executing good questionnaire.

2. Selection of appropriate sampling method.

3. Adequate sample size

4. Employing trained investigation to collect the data.

5. Care in editing, coding and entering the data into the computer.

Sample Size Determination

When you survey a large population of respondents, you're interested in the entire group, but it's not realistically possible to get answers or results from absolutely everyone. So you take a random sample of individuals which represents the population as a whole.

The size of the sample is very important for getting accurate, statistically significant results and running your study successfully.

If your sample is too small, you may include a disproportionate number of individuals which are outliers and anomalies. These skew the results and you don't get a fair picture of the whole population.

If the sample is too big, the whole study becomes complex, expensive and time-consuming to run, and although the results are more accurate, the benefits don't outweigh the costs.

When you only survey a small sample of the population, uncertainty creeps in to your statistics. If you can only survey a certain percentage of the true population, you can never be 100% sure that your statistics are a complete and accurate representation of the population. This uncertainty is called sampling error and is usually measured by a confidence interval. For example, you might state that your results are at a 90% confidence level. That means if you were to repeat your survey over and over, 90% of the time your would get the same results. So you must learn how to determine sample size

To choose the correct **sample size**, **(A sample size is a part of the population chosen for a survey or experiment) Sample size determination is the act of choosing the number of observations or replicates to include in a statistical sample. The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample. In practice, the sample size used in a study is usually determined based on the cost, time, or convenience of collecting the data, and the need for it to offer sufficient statistical power.**

You need to consider a few different factors that affect your research, and gain a basic understanding of the statistics involved. You'll then be able to use a sample size formula to bring everything together and sample confidently, knowing that there is a high probability that your survey is statistically accurate.

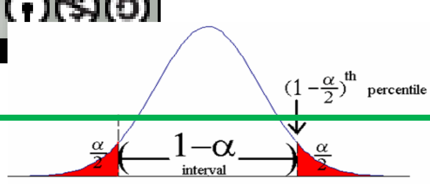
Stage 1: Consider your sample size variables

Before you can calculate a sample size, you need to determine a few things about the target population and the level of accuracy you need: In addition to the purpose of the study and population size, three criteria usually will need to be specified to determine the appropriate sample size: the level of precision, the level of confidence or risk, and the degree of variability in the attributes being measured

1. Population size

How many people are you talking about in total? To find this out, you need to be clear about who does and doesn't fit into your group.

2. Margin of error: Errors are inevitable – the question is how much error you'll allow. The margin of error is a statistic expressing the amount of random sampling error in the results of a survey. The margin of error is the amount of error that you can tolerate. Greater the desired precision (Lower margin of error), larger will be the sample size. The desired precision of the estimate (also



sometimes called the allowable or acceptable error in the estimate). For example if you would like the confidence interval width to be about 0.5 (5%) you would enter a precision of +/- 0.025.

3. Confidence interval.

In statistics, a confidence interval is a type of estimate computed from the statistics of the observed data. A confidence level refers to the percentage of all possible samples that can be expected to include the true population parameter. It deals with how confident you want to be that the actual mean falls within your margin of error. The most common confidence intervals are 90%, 98%, 95% confident, and 99% etc.

A 0% confidence level means you have no faith at all that if you repeated the survey that you would get the same results. A 100% confidence level means there is no doubt at all that if you repeated the survey you would get the same results. In reality, you would never publish the results from a survey where you had no confidence at all that your statistics were accurate (you would probably repeat the survey with better techniques). A 100% confidence level doesn't exist in statistics, unless you surveyed an entire population — and even then you probably couldn't be 100 percent sure that your survey wasn't open to some kind or error or bias. In general, the higher the coefficient, the more certain you are that your results are accurate.

Degree of variability:

In order to estimate the sample size, three issues need to be studied, i.e. the level of precisions, confidence or risk level and the variability. Regarding the last issue, which your questions is concentrated the degree of variability in the attributes being measured refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain a given level of precision. The less variable (more homogeneous) a population, the smaller the sample size. Note that a proportion of 50% indicates a greater level of variability. This is because a proportion of .5 indicates the maximum variability in a population, it is often used in determining a more conservative sample size, that is, the sample size may be larger than if the true variability of the population attribute were used.

Use the sample size formula

There are many different formulas you can use, depending on what you know (or don't know) about your population. You're ready to calculate the sample size you need. This can be done using an online sample size calculator or with paper and pencil.

Sample Size Calculator

This Sample Size Calculator is presented as a public service of Creative Research Systems survey software. You can use it to determine how many people you need to interview in order to get results that reflect the target population as precisely as needed. You can use any online system for the estimation of appropriate sample size.

Use a sample size calculator

1. Raosoft
2. Survey monkey
3. Survey system

you can use any online system

There are many different formulas you can use, depending on what you know (or don't know) about your population. **Cochran's formula** is considered especially appropriate in situations with **large populations**.

Cochran (1977) has given a modus operandi for sample size determination**. In order to decide upon the sample size, according to Cochran, the researcher has to be able to make out the boundaries of mistakes and errors in the items which have been considered crucial in the survey. Cochran holds that an approximate guess of the required sample size is made disjointedly for each item in the survey.** The researcher who is undertaking the task will then use the help of a wide range of sample sizes which includes smaller sample sizes for dichotomous categorical variables. Sampling decisions should be made by the researcher based on the data acquired. The researcher uses the largest sample size if the range of the sample size is close to the variable of interest.

When the researcher does not have direct influence over the variance, he must take in the variance estimates. This is a serious component in sample size determination. This is because the estimation or approximation of the difference in the important variables of interest under the study is an essential module for sample size determination.

To estimate the population for sample size determination, Cochran followed four steps. In the first step of estimating the population variances and differences for sample size determination, the researcher obtains the samples in two steps. He uses the results of the first step in order to settle on the desired number of extra responses to achieve an appropriate sample size based on the differences studied in the first step. Secondly, while determining the sample size, the researcher estimates the population variances for sample size determination by using the results of the pilot study. Next, the data from prior studies of the population is used by the researcher to determine the sample size. Finally, the researcher makes the required estimation for sample size determination by the formation of the population using the assistance of some logical mathematical results.

Cochran's formula large (infinite) populations

Where: e is the desired level of precision (i.e. the margin of error),

p is the (estimated) proportion of the population which has the attribute in question, (degree of variability) q is 1 – p.

The z-value is found in a Z table.

90% – Z Score = 1.645

95% – Z Score = 1.96

99% – Z Score = 2.576

If you chose a different confidence level, use this Z-score table to find your score.

Use the sample size formula

Plug in your Z-score, standard of deviation, and confidence interval into the sample size calculator or use this sample size formula to work it out yourself:

Necessary Sample Size = (Z-score)2 * StdDev*(1-StdDev) / (margin of error)2

Here's a worked example, assuming you chose a 95% confidence level, .5 standard deviation or degree of variability is maximum at 50%, and a margin of error (confidence interval) of +/- 5%.

((1.96)2 x .5(.5)) / (.05)2

$$n_0 = \frac{Z^2pq}{e^2}$$

(3.8416 x .25) / .0025

.9604 / .0025

384.16

385 respondents are needed

2. If the population we're studying is small (finite or known), we can modify the sample size we calculated in the above formula by using this equation: $\frac{n0}{1+(n0-1)/N}$

We know the population is 5000 what is the appropriate sample size. Use the above equation; here n0 represents the number of respondents needed when the population is infinite. Here 385 respondents so the required sample size is 359 respondents

$\frac{385}{1+(385-1)/5000}$ =359

Taro Yamane formula

Yamane (1967:886) provides a simplified formula to calculate sample sizes. This formula was used to calculate the sample size $n = \frac{N}{1+N(e)^2}$

N is the population size, and e is the level of precision.

n= sample size, N = population size, and e = Margin of error (MoE), e = 0.05 based on the research condition. In a finite population, when the original sample collected is more than 5% of the population size, the corrected sample size is determined by using the Yamane's formula.

$$n = \frac{5000}{1+5000(.05)^2} = 370$$

- The second method is to use the formula for the sample size for the mean. The formula of the sample size for the mean is similar to that of the proportion, except for the measure of variability. The formula for the mean employs σ^2 instead of (p x q), as shown in

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

σ^2 is the variance of an attribute in the population.

The disadvantage of the sample size based on the mean is that a "good" estimate of the population variance is necessary. Often, an estimate is not available. Furthermore, the sample size can vary widely from one attribute to another because each is likely to have a different variance. Because of these problems, the sample size for the proportion is frequently preferred.

Summary

You can use any formula for the determination of appropriate sample size for your study. But one thing kept in your mind don’t follow blindly these formula for estimating sample size. Use proper review of literature then you will get an idea about the adequate sample size needed for the study.

&&&&&&&&&&&&&&&