9. SAMPLE SIZE

9.1. THE DETERMINATION OF THE APPROPRIATE SIZE OF A SAMPLE OBJECTIVES:

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 determined based on the expense of data collection, and the need to have sufficient statistical power.

In complicated studies there may be several different sample sizes involved in the study: for example, in a survey sampling involving stratified sampling there would be different sample sizes for each population. In a census, data are collected on the entire population, hence the sample size is equal to the population size.
In experimental design, where a study may be divided into different treatment groups, there may be different sample sizes for each group.

Sample sizes may be chosen in several different ways:

- expedience - For example, include those items readily available or convenient to collect. A choice of small sample sizes, though sometimes necessary, can result in wide confidence intervals or risks of errors in statistical hypothesis testing.
- using a target variance for an estimate to be derived from the sample eventually obtained
- using a target for the power of a statistical test to be applied once the sample is collected.

Larger sample sizes generally lead to increased precision when estimating unknown parameters. For example, if we wish to know the proportion of a certain species of fish that is infected with a pathogen, we would generally have a more accurate estimate of this proportion if we sampled and examined 200 rather than 100 fish. Several fundamental facts of mathematical statistics describe this phenomenon, including the law of large numbers and the central limit theorem.

In some situations, the increase in accuracy for larger sample sizes is minimal, or even non-existent. This can result from the presence of systematic errors or strong dependence in the data, or if the data follow a heavy-tailed distribution.

Sample sizes are judged based on the quality of the resulting estimates. For example, if a proportion is being estimated, one may wish to have the 95% confidence interval be less than 0.06 units wide. Alternatively, sample size may be assessed based on the power of a hypothesis test. For example, if we are comparing the support for a certain political candidate among women with the support for that candidate among men, we may wish to have 80% power to detect a difference in the support levels of 0.04 units.

Required sample sizes for hypothesis tests

A common problem faced by statisticians is calculating the sample size required to yield a certain power for a test, given a predetermined Type I error rate α. As follows, this can be estimated by pre-determined tables for certain values, by Mead's resource equation, or, more generally, by the cumulative distribution function:

By tables

Tables can be used in a two-sample t-test to estimate the sample sizes of an experimental group and a control group that are of equal size, that is, the total number of individuals in the trial is twice that of the number given, and the desired significance level is 0.05. The parameters used are:

- The desired statistical power of the trial, shown in column to the left.
- Cohen's d (=effect size), which is the expected difference between the means of the target values between the experimental group and the control group, divided by the expected standard deviation.

Mead's resource equation

Mead's resource equation is often used for estimating sample sizes of laboratory animals, as well as in many other laboratory experiments. It may not be as accurate as using other methods in estimating sample size, but gives a hint of what is the appropriate sample size where parameters such as expected standard deviations or expected differences in values between groups are unknown or very hard to estimate.[5]

All the parameters in the equation are in fact the degrees of freedom of the number of their concepts, and hence, their numbers are subtracted by 1 before insertion into the equation.

9.2 PARAMETER ESTIMATION: Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured/empirical data that has a random component. The parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements.

For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the parameter sought; the estimate is based on a small random sample of voters.

Or, for example, in radar the goal is to estimate the range of objects (airplanes, boats, etc.) by analyzing the two-way transit timing of received echoes of transmitted pulses. Since the reflected pulses are unavoidably embedded in electrical noise, their measured values are randomly distributed, so that the transit time must be estimated.

In estimation theory, two approaches are generally considered.

- The probabilistic approach (described in this article) assumes that the measured data is random with probability distribution dependent on the parameters of interest

- The set-membership approach assumes that the measured data vector belongs to a set which depends on the parameter vector.

For example, in electrical communication theory, the measurements which contain information regarding the parameters of interest are often associated with a noisy signal. Without randomness, or noise, the problem would be deterministic and estimation would not be needed.

9.3 ESTIMATE OF A PROPORTION: The proportion of something is the number of observations that meet a certain criterion, divided by the total number of observations. For example, the proportion of males in the population of Americans is the number of American males divided by the number of Americans. The population proportion is this for the entire population. This can rarely be calculated exactly, so it must be estimated.

Instructions

- 1
  Get a random sample of the population. If your sample is not random, estimates of the proportion (and other quantities) may be biased. For example, if you want to estimate the proportion of boys in an elementary school, you could assign a number to each student, then randomly pick a sample by choosing random numbers. The bigger your sample, the more accurate your estimate will be.
- 2
  Find the number of observations that meet the criterion in your sample. In our example, we would find how many of the children in our sample were boys.
- 3
  Divide this number by the total number of observations in the sample. This is the estimated proportion.
- 4
  To see how good this estimate is, the standard formula for a 95 percent confidence interval is p +- 1.96(pq/n) ^ .5, where p is the proportion found in step 3, q = 1 - p, and n is the number of observations.

9.4 ESTIMATION OF AN AVERAGE: In statistics, **estimation** refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Point Estimate vs. Interval Estimate

Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

- **Point estimate**. A point estimate of a population parameter is a single value of a statistic. For example, the sample mean x is a point estimate of the population mean μ. Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.
- **Interval estimate**. An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an

interval estimate of the population mean μ. It indicates that the population mean is greater than *a* but less than *b*.

Confidence Intervals

Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.
- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* ± *margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic* ± *margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level

The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

9.5 COMPARISON OF TWO PROPORTIONS: For statistical purposes, you can compare two populations or groups when the variable is categorical (for example, smoker/nonsmoker, Democrat/Republican, support/oppose an opinion, and so on) and you're interested in the proportion of individuals with a certain characteristic — for example, the proportion of smokers.

In order to make this comparison, two independent (separate) random samples need to be selected, one from each population. The null hypothesis H_0 is that the two population proportions are the same; in other words, that their difference is equal to 0. The notation for the null hypothesis is H_0: $p_1 = p_2$, where p_1 is the proportion from the first population, and p_2 is the proportion from the second population