Unit 6: Estimation

It is often of interest to learn about the characteristics of a large group of elements such as individuals, households, buildings, products, parts, customers, and so on. All the elements of interest in a particular study form the population. Because of time, cost, and other considerations, data often cannot be collected from every element of the population. In such cases, a subset of the population, called a sample, is used to provide the data. Data from the sample are then used to develop estimates of the characteristics of the larger population. The process of using a sample to make inferences about a population is called statistical inference.

Characteristics such as the population mean, the population variance, and the population proportion are called parameters of the population. Characteristics of the sample such as the sample mean, the sample variance, and the sample proportion are called sample statistics. There are two types of estimates: point and interval. A point estimate is a value of a sample statistic that is used as a single estimate of a population parameter. No statements are made about the quality or precision of a point estimate. Statisticians prefer interval estimates because interval estimates are accompanied by a statement concerning the degree of confidence that the interval contains the population parameter being estimated. Interval estimates of population parameters are called confidence intervals.

Sampling and sampling distributions

Although sample survey methods will be discussed in more detail below in the section Sample survey methods, it should be noted here that the methods of statistical inference, and estimation in particular, are based on the notion that a probability

sample has been taken. The key characteristic of a probability sample is that each element in the population has a known probability of being included in the sample. The most fundamental type is a simple random sample.

For a population of size *N*, a simple random sample is a sample selected such that each possible sample of size *n* has the same probability of being selected. Choosing the elements from the population one at a time so that each element has the same probability of being selected will provide a simple random sample. Tables of random numbers, or computer-generated random numbers, can be used to guarantee that each element has the same probability of being selected.

A sampling distribution is a probability distribution for a sample statistic. Knowledge of the sampling distribution is necessary for the construction of an interval estimate for a population parameter. This is why a probability sample is needed; without a probability sample, the sampling distribution cannot be determined and an interval estimate of a parameter cannot be constructed.

Estimation of a population mean

The most fundamental point and interval estimation process involves the estimation of a population mean. Suppose it is of interest to estimate the population mean, μ, for a quantitative variable. Data collected from a simple random sample can be used to compute the sample mean, $\bar{x}$, where the value of $\bar{x}$ provides a point estimate of μ.

When the sample mean is used as a point estimate of the population mean, some error can be expected owing to the fact that a sample, or subset of the population, is used to compute the point estimate. The absolute value of the difference between the sample mean, $\bar{x}$, and the population mean, μ, written $|\bar{x} - \mu|$, is called the sampling error. Interval estimation incorporates a probability statement about the magnitude of the sampling error. The sampling distribution of $\bar{x}$ provides the basis for such a statement.

Statisticians have shown that the mean of the sampling distribution of $\bar{x}$ is equal to the population mean, μ, and that the standard deviation is given by σ / Square root of $\sqrt{n}$, where σ is the population standard deviation. The standard deviation of a sampling distribution is called the standard error. For large sample sizes, the central limit theorem indicates that the sampling distribution of $\bar{x}$ can be approximated by a normal probability distribution. As a matter of practice, statisticians usually consider samples of size 30 or more to be large.

In the large-sample case, a 95% confidence interval estimate for the population mean is given by $\bar{x} \pm 1.96\sigma$ / Square root of $\sqrt{n}$. When the population standard deviation, σ, is unknown, the sample standard deviation is used to estimate σ in the confidence interval formula. The quantity 1.96σ / Square root of $\sqrt{n}$ is often called the margin of error for the estimate. The quantity σ / Square root of $\sqrt{n}$ is the standard error, and 1.96 is the number of standard errors from the mean necessary to include 95% of the values in a normal distribution. The interpretation of a 95% confidence interval is that 95% of the intervals constructed in this manner will contain the population mean. Thus, any interval computed in this manner has a 95% confidence of containing the population mean. By changing the constant from 1.96 to 1.645, a 90% confidence interval can be obtained. It should be noted from the formula for an interval estimate that a 90% confidence interval is narrower than a 95% confidence interval and as such has a slightly smaller confidence of including the population mean. Lower levels of confidence lead to even more narrow intervals. In practice, a 95% confidence interval is the most widely used.

Owing to the presence of the $n^{1/2}$ term in the formula for an interval estimate, the sample size affects the margin of error. Larger sample sizes lead to smaller margins of error. This observation forms the basis for procedures used to select the sample size. Sample sizes can be chosen such that the confidence interval satisfies any desired requirements about the size of the margin of error.

The procedure just described for developing interval estimates of a population mean is based on the use of a large sample. In the small-sample case—*i.e.*, where the sample size *n* is less than 30—the *t* distribution is used when specifying the margin of error and constructing a confidence interval estimate. For example, at a 95% level of confidence, a value from the *t* distribution, determined by the value of *n*, would replace the 1.96 value obtained from the normal distribution. The *t* values will always be larger, leading to wider confidence intervals, but, as the sample size becomes larger, the *t* values get closer to the corresponding values from a normal distribution. With a sample size of 25, the *t* value used would be 2.064, as compared with the normal probability distribution value of 1.96 in the large-sample case.

Estimation of other parameters

For qualitative variables, the population proportion is a parameter of interest. A point estimate of the population proportion is given by the sample proportion. With knowledge of the sampling distribution of the sample proportion, an interval estimate of a population proportion is obtained in much the same fashion as for a population mean. Point and interval estimation procedures such as these can be applied to other population parameters as well. For instance, interval estimation of a population variance, standard deviation, and total can be required in other applications.

Estimation procedures for two populations

The estimation procedures can be extended to two populations for comparative studies. For example, suppose a study is being conducted to determine differences between the salaries paid to a population of men and a population of women. Two independent simple random samples, one from the population of men and one from the population of women, would provide two sample means, $\bar{x}_1$ and $\bar{x}_2$. The difference between the two sample means, $\bar{x}_1 - \bar{x}_2$, would be used as a point estimate of the

difference between the two population means. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ would provide the basis for a confidence interval estimate of the difference between the two population means. For qualitative variables, point and interval estimates of the difference between population proportions can be constructed by considering the difference between sample proportions.

Comprehension Exercises

Choose a, b, c or d which best completes each item.

1) Characteristics such as the population mean, the population variance, and the population proportion are called _____ of the population.

a) parameters b) statistics c) quantities d) measures

2) The population proportion is a parameter of interest for _____ variables.

a) discrete b) continuous c) quantitative d) qualitative

3) A 90% confidence interval is _____ than a 95% confidence interval and as such has a slightly smaller confidence of including the population mean.

a) smaller b) larger c) narrower d) wider

4) The difference between sample means would be used as a _____ of the difference between the population means..

a) confidence interval b) point estimate c) approximation d) guess

Words to Learn

Find the Persian equivalents of the following terms and expressions.

estimate	process	one at a time	household
estimation	degree	absolute value	building
parameter	confidence	magnitude	product
sample statistic	confidence interval	narrow	part
point estimation	knowledge	closer	customer
interval estimation	replace	margin	fundamental
sample survey	basis	lower	guarantee
standard error	procedure	upper	conduct
small-sample	sample size	key	between
large-sample	total	fashion	practice