Unit-5

Estimation Theory

Estimation is a technique for calculating information about a bigger group from a smaller sample, and statistics are crucial to analyzing data. For instance, the average age of a city's population may be obtained by taking the age of a sample of 1,000 residents. While estimates aren't perfect, they are typically trustworthy enough to be of value.

What is Estimation?

Estimation in statistics involves using sample data to make educated guesses about a population's characteristics, such as mean, variance, or proportion. The population refers to the entire interest group, like all people in a country or all products made by a company.

Purpose of Estimation in Statistics

Statistical estimation is essential for making inferences about populations using sample data, helping to determine parameters like mean and variance without individual measurements.
- This evaluation is vital for decision-making in business and healthcare, informing strategies and treatment options.
- It is closely linked to hypothesis testing, contributing to scientific development, political decisions, public health, and economic choices.
- Risk assessment benefits from evaluation in managing probabilities and risk in finance and insurance.
- Quality control also relies on evaluation to ensure products and services meet standards by identifying and correcting deviations.

Types of Estimation

Estimation is of two types that includes:
- Point Estimation
- Interval Estimation

Point Estimation

In statistics, the sample mean is used to estimate a population mean, while the sample proportion is used to estimate a population percentage. These measurements help approximate unknown population parameters accurately.
- Identifying a single number to represent a large group is like a point estimate. For instance, measuring the heights of random people can be used to estimate the average height of the entire group.

If individuals measured were 5 feet, 6 feet, and 5 feet.
We could estimate the average height to be around 5 feet. This single number, called a score estimator, gives a rough idea of the group's characteristics.

- Population mean is estimated using the sample mean.
- Similar techniques can be applied to estimate other attributes like percentages of specific characteristics in a population.

Properties of Point Estimators

It is desirable for a point estimate to be the following:

- **Consistent** - We can say that the larger is the sample size, the more accurate is the estimate.
- **Unbiased** - The expectation of the observed values of various samples equals the corresponding population parameter. Let's take, for example, We can say that sample mean is an unbiased estimator for the population mean.
- **Most Efficient That is also Known as Best Unbiased** - of all the various consistent, unbiased estimates, the one possessing the smallest variance (a measure of the amount of dispersion away from the estimate). In simple words, we can say that the estimator varies least from sample to sample and this generally depends on the particular distribution of the population. For example, the mean is more efficient than the median (that is the middle value) for the normal distribution but not for more "skewed" (also known as asymmetrical) distributions.

Interval Estimation

Interval estimation involves constructing a range of values, called a confidence interval, that is likely to contain the unknown population parameter with a certain level of confidence. For example, a 95% confidence interval for the population mean would contain the true population mean in 95% of all possible samples.

The formula for a confidence interval for the population mean is:

$$\bar{x} \pm z\ \alpha/2 * \sigma / \sqrt{n}$$

where ,

- $\bar{x}$ is the sample mean
- $z\ \alpha/2$ is the z-score that corresponds to the desired level of confidence
- σ is the population standard deviation (if known) or the sample standard deviation (if unknown)
- and n is the sample size.

The formula for a confidence interval for the population proportion is:

$\hat{p} \pm z\alpha/2 * \sqrt{(\hat{p}(1 - \hat{p}))} / n$

- where **$\hat{p}$** is the sample proportion,
- **$z\alpha/2$** is the z-score that corresponds to the desired level of confidence,
- and **n** is the sample size.

Properties of Confidence intervals

- A good confidence interval should have two desirable properties: coverage probability and margin of error.
- Coverage probability means that the confidence interval will contain the true population parameter with a certain level of confidence. For example, a 95% confidence interval should contain the true population parameter in 95% of all possible samples.
- Margin of error is a measure of the precision of the confidence interval. The margin of error is defined as half the width of the confidence interval. A narrower confidence interval will have a smaller margin of error and be more precise.

Confidence Interval Estimation

- Interval estimation is another method of estimation that provides a range of plausible values for a population parameter. Unlike point estimation, it provides a range of values rather than a single value. The range is called the confidence interval, and it represents a level of certainty that the true population parameter falls within the interval.
- The confidence interval is computed by taking the point estimate and adding and subtracting a margin of error. The margin of error is based on the level of confidence, the sample size, and the standard error of the point estimate. A common level of confidence used in interval estimation is 95%.
- For example, suppose we want to estimate the mean weight of all dogs in a population. We take a sample of 100 dogs and compute the sample mean weight to be 30 pounds with a standard deviation of 5 pounds. We want to construct a 95% confidence interval for the population mean weight.
- Using the formula for a confidence interval for the population mean, we can compute the margin of error as follows:

Margin of error = $Z_{(\alpha/2)} * (s/\sqrt{n}) = 1.96 * (5/\sqrt{100}) = 0.98$

We then construct the confidence interval as follows:

Confidence interval = sample mean ± margin of error = 30 ± 0.98 = (29.02, 30.98)

This means we are 95% confident that the true population mean weight falls within the interval of 29.02 to 30.98 pounds.

Types of Interval Estimation

There are different types of interval estimation, depending on the parameter being estimated and the method used to compute the confidence interval. Some common types are:

- Confidence interval for the population mean: This is used to estimate the population mean when the population standard deviation is unknown. It uses the t-distribution to compute the critical value for the confidence interval.
- Confidence interval for the population proportion: This is used to estimate the population proportion, such as the proportion of voters who support a certain candidate. It uses the normal distribution to compute the critical value for the confidence interval.
- Confidence interval for the difference between two means: This is used to estimate the difference between two population means. It uses either the t-distribution or normal distribution, depending on the sample sizes and whether the variances are assumed to be equal or not.
- Confidence interval for the difference between two proportions: This is used to estimate the difference between two population proportions. It uses either the normal distribution or the chi-square distribution, depending on the sample sizes and whether the variances are assumed to be equal or not.

Limitations and Assumptions of Interval Estimation

Interval estimation, like point estimation, relies on certain assumptions and has some limitations. Some of the key assumptions and limitations are:

- The sample is representative of the population: Interval estimation assumes that the sample is randomly selected and represents the population of interest. If the sample is biased or non-random, the confidence interval may not be accurate.
- Normality assumption: Interval estimation assumes that the population is normally distributed or that the sample size is large enough for the central limit theorem to apply. If the data is not normally distributed and the sample size is small, the confidence interval may not be accurate.
- Independence assumption: Interval estimation assumes that the observations are independent of each other. If there is correlation or dependence between the observations, the confidence interval may not be accurate.

- Finite population correction: If the sample size is a significant portion of the population size, a finite population correction factor may need to be applied to adjust the confidence interval.