

Sampling Distribution

Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as the sample mean or sample proportion) computed from multiple independent samples of the same size from a population.

- Population : The entire data.
- Sample : Some part of the data.

Why Sampling Distribution is important?

Sampling distribution is important in statistics because it allows us to estimate the variability of a sample statistic, which is useful for making inferences about the population. By analysing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis tests, and make predictions about the population based on the sample data.

Central Limit Theorem

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution.

The conditions required for the CLT:

1. The sample size is large enough, typically greater than or equal to 30.
2. The random variables in the sample are independent and identically distributed.

Point Estimate

A point estimate is a single value, calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.

Confidence Interval

Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

Confidence Interval = Point Estimate \pm Margin of Error

Two Types of CI :

- Confidence Interval (Z Procedure)
- Confidence Interval (T Procedure)

1. Confidence Interval With Z Procedure:

Assumptions :

1. Random Sampling
2. Known population Standard Deviation
3. Normal Distribution , if not normal than apply central limit theorem to make distribution normal.

Formula : $CI = \bar{X} \pm Z_{\alpha/2} \frac{\delta}{\sqrt{n}}$

Confidence Interval With T Procedure:

Assumptions :

1. Random Sampling
2. Unknown population Standard Deviation
3. Normal Distribution , if not normal than apply central limit theorem to make distribution normal.

Formula : $CI = \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

Scientific Reason Why CI Work

Even with one sample, the CI works because:

1. **Sample is random is unbiased:**
Each random sample is equally likely to be above or below the population mean.
2. **CLT guarantees:**
The distribution of sample means is predictable (approximately normal).
3. **Probability interpretation:**
If you repeated the experiment many times, 95% of the confidence intervals calculated from different samples would contain the true population mean.

Hypothesis Testing:

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.

Example : You are a social media manager. You changed the thumbnail of your video to see if more people will click on it. You want to know, Does the new thumbnail increase the click rate compared to the old one?

- **Null Hypothesis(H_0):**

The null hypothesis is a statement that assumes there is no significant effect or relationship between the variables being studied. It serves as the starting point for hypothesis testing and represents the status quo or the assumption of no effect until proven otherwise. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

Example : The new thumbnail does not increase the click rate.

- **Alternate Hypothesis(H1):**

The alternative hypothesis, is a statement that contradicts the null hypothesis and claims there is a significant effect or relationship between the variables being studied. It represents the research hypothesis or the claim that the researcher wants to support through statistical analysis.

Example : The new thumbnail increases the click rate.

Steps involved in Hypothesis Testing

1. Formulate a Null and Alternate hypothesis
2. Select a significance level(This is the probability of rejecting the null hypothesis when it is actually true, usually set at 0.05 or 0.01)
3. Check assumptions (make distribution, check population std, data type working with, single or multiple columns)
4. Decide which test is appropriate(Z-test, T-test)
5. Conduct the test
6. Reject or not reject the Null Hypothesis.
7. Interpret the result

Example : Performing Test

Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.

