Inferential Statistics

Introduction

Definition of Inferential Statistics

Inferential statistics is a branch of statistics that focuses on making inferences about populations based on data collected from samples. Unlike descriptive statistics, which simply describe the characteristics of a dataset, inferential statistics allows us to draw conclusions and make predictions about a larger group by analyzing a smaller, representative subset of that group.

Importance and Applications

Inferential statistics is crucial in various fields because it provides the tools needed to make evidence-based decisions and predictions. Some of the key applications include:

- **Medicine**: Determining the effectiveness of new treatments or drugs based on clinical trial data.
- **Economics**: Predicting economic trends and making policy decisions based on sample surveys.
- **Social Sciences**: Understanding social behaviors and trends through survey data.
- **Business**: Making informed business decisions, such as market research and quality control.

Difference Between Descriptive and Inferential Statistics

While both branches of statistics deal with data, they serve different purposes:

- **Descriptive Statistics**: Involves summarizing and organizing data to describe the characteristics of a dataset. This includes measures of central tendency (mean, median, mode), measures of variability (range, variance, standard deviation), and graphical representations (histograms, bar charts).
- **Inferential Statistics**: Involves using sample data to make inferences or generalizations about a population. This includes estimation (point estimates and confidence intervals) and hypothesis testing (determining if there is enough evidence to support a certain claim about the population).

Sampling Theory and Methods

Population and Sample

- **Population**: The complete set of all items or individuals of interest in a particular study. For example, if we are studying the heights of all students in a school, the population would be all the students in that school.
- **Sample**: A subset of the population selected for analysis. A sample is used to make inferences about the population. For example, if we measure the heights of 50 students from the school, this group represents a sample of the population.

Parameters and Statistics

- **Parameters**: Numerical characteristics of a population, such as the population mean (μ) or population standard deviation (σ). Parameters are often unknown and need to be estimated from sample data.
- **Statistics**: Numerical characteristics of a sample, such as the sample mean ($\bar{x}$) or sample standard deviation (s). Statistics are used to estimate the corresponding parameters of the population.

Sampling Methods

- **Random Sampling**: Each member of the population has an equal chance of being selected. This method helps ensure that the sample is representative of the population.
- **Stratified Sampling**: The population is divided into strata (groups) based on a specific characteristic, and random samples are taken from each stratum. This ensures that each subgroup is adequately represented in the sample.
- **Cluster Sampling**: The population is divided into clusters, usually based on geographical areas or natural groupings. A random sample of clusters is selected, and all individuals within the chosen clusters are included in the sample.
- **Systematic Sampling**: Every kth member of the population is selected, where k is a constant interval. For example, if k=10, every 10th member of the population is chosen. This method is easier to implement than simple random sampling but may introduce bias if there is a hidden pattern in the population.

Large Sample Theory

Large Sampling Theory in the study of the sequence of random variable when n tends to infinity. This helps us recognize the limiting behaviour of the sequence of random variables.

Mode of Convergence

Understanding how sequences of random variables converge is essential in inferential statistics. There are several modes of convergence:

- **Convergence in Distribution**: Let $\{F_n\}$ be a sequence of distribution functions corresponding to the sequences of random variables $\{X_n\}$. If there exist a distribution function F such that,

  $$F_n(x) \rightarrow F(x) \text{ as } n \rightarrow \infty$$

  at every point at which F is continuous. Then, we say F_n converges in law to F.

- **Convergence in Probability**: Let $\{X_n\}$ be a sequence of random variables defined on some probability space. We say X_n converges in probability to a random variable X, if for all $\varepsilon > 0$

  $$P(|X_n - X| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

- **Convergence Almost Surely**: Let $\{X_n\}$ be a sequence of random variable. We say that X_n converges almost surely to a random variable X iff

  $$P\{\lim_{n \rightarrow \infty} Xn(\omega) = X(\omega)\} = 1$$

  This is the strongest form of convergence.

Law of Large Numbers (LLN)

The Law of Large Numbers is a fundamental theorem that describes the result of performing the same experiment a large number of times. It guarantees that the average of the results obtained from a large number of trials will be close to the expected value and will tend to become closer as more trials are performed.

- **Weak Law of Large Numbers (WLLN)**: States that the sample average converges in probability towards the expected value as the sample size grows.

- **Strong Law of Large Numbers (SLLN)**: States that the sample average almost surely converges to the expected value as the sample size grows.

Central Limit Theorem (CLT)

The Central Limit Theorem is one of the most important results in statistics. It states that, for a large enough sample size, the sampling distribution of the sample mean will be approximately normally distributed, regardless of the original distribution of the population. This allows for the use of normal probability theory to make inferences about population parameters.

This theorem is foundational for many statistical procedures, including hypothesis testing and confidence interval estimation, and highlights the importance of normality in the realm of inferential statistics.

With an understanding of large sample theory, including modes of convergence, the law of large numbers, and the central limit theorem, we can proceed to the next section on sampling distributions.

Standard Error and Sampling Distributions

Definition and Importance

A sampling distribution is the probability distribution of a given statistic based on a random sample. It describes how the statistic varies from sample to sample and is fundamental in making inferences about population parameters. Understanding sampling distributions is essential for constructing confidence intervals and conducting hypothesis tests.

Standard Error

The standard error is the standard deviation of a sampling distribution. It measures the variability of a sample statistic (such as the sample mean) and decreases as the sample size increases.

Chi-Squared Distribution

The chi-squared distribution is a continuous probability distribution that arises in the context of estimating variances and conducting goodness-of-fit tests and tests of independence in contingency tables.

- **Degrees of Freedom** : The shape of the chi-squared distribution depends on the degrees of freedom, which is typically the number of independent pieces of information used to estimate a parameter.
- **Applications**:
    - **Goodness-of-Fit Test**: Tests whether a sample matches a distribution.
    - **Test of Independence**: Assesses the relationship between categorical variables.

Student's t-Distribution

The t-distribution is used instead of the normal distribution when the sample size is small, and the population standard deviation is unknown. It is similar in shape to the normal distribution but has heavier tails.

- **Degrees of Freedom** : The shape of the t-distribution depends on the degrees of freedom, calculated as n-1 where n is the sample size.
- **Applications**:
    - **Confidence Intervals**: For the mean when the population standard deviation is unknown.
    - **Hypothesis Testing**: Testing the mean when the population standard deviation is unknown.

F-Distribution

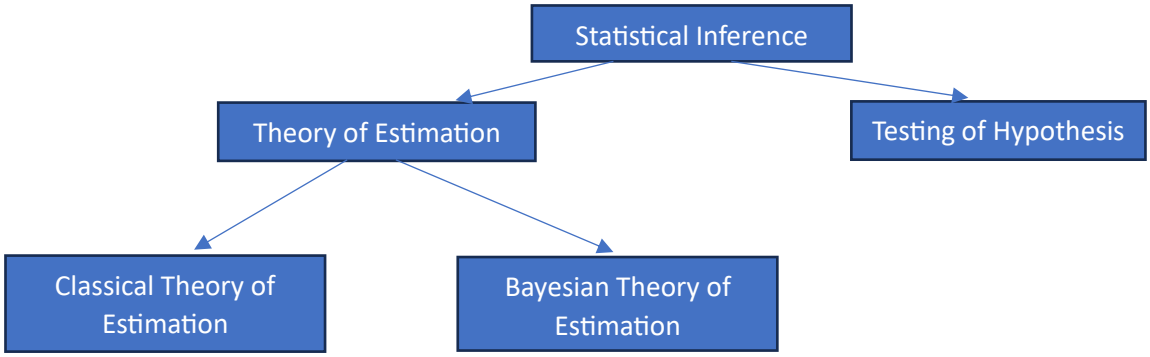
The F-distribution is used primarily in the analysis of variance (ANOVA) and in regression analysis to compare variances and test the overall fit of a model.

- **Degrees of Freedom (df)**: The F-distribution has two sets of degrees of freedom: one for the numerator and one for the denominator.
- **Applications**:
  - **ANOVA**: Testing the equality of three or more means.
  - **Regression Analysis**: Testing the significance of the overall regression model.

Estimation

Statistical Inference

Statistical inference is the process of drawing conclusions about on underlying population based on a sample or subset of data.



Parameters and Parameter Space

Parameters are functions of population observations. The parameter space is the space of possible parameter values and it is denoted by Θ.

Classical Theory of Estimator

In classical theory of estimation, we have density function $f(x, \theta)$, where x is the value of random variable and θ is the set of unknown parameters. Here, we estimate the value of unknown parameters. Method to find estimator in CTE:

- Method of Moments (MOM)
- Method of Maximum Likelihood Estimator
- Method of Minimum Variance
- Method of Least Square (Regression)
- Method of Minimum Chi-square (Discrete Population)

Bayesian Theory of Estimator

In Bayesian theory of estimation, we have density function $f(x,\theta)$, where x is the value of random variable and θ is also a random variable. Here, we estimate the distribution of unknown parameters.

Estimator

An Estimator is a statistic that estimates the population parameter. Properties of Estimator are:

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency
- Completeness

Point Estimation

Point estimation involves using sample data to calculate a single value, known as a point estimate, which serves as the best guess for an unknown population parameter. Common point estimates include:

- **Sample Mean ($\bar{X}$)**: Used to estimate the population mean (μ).
- **Sample Proportion ($\hat{p}$)**: Used to estimate the population proportion (p).
- **Sample Variance (s^2)**: Used to estimate the population variance (σ^2).

A good point estimator should be unbiased, meaning the expected value of the estimator equals the parameter being estimated, and have minimum variance among all unbiased estimators.

Interval Estimation

Interval estimation provides a range of values, known as a confidence interval, within which the population parameter is expected to lie with a certain level of confidence.

- **Confidence Interval for the Mean (when σ is known/unknown)**.
- **Confidence Interval for a Proportion**.

Confidence intervals provide a range that is likely to contain the population parameter, giving a measure of reliability to the point estimate.

Hypothesis Testing

Hypothesis testing is a method used to make decisions about population parameters based on sample data. It involves the following steps:

Null and Alternative Hypotheses

- **Null Hypothesis (H_0)**: A statement that there is no effect or no difference, and it serves as the default or baseline assumption.
- **Alternative Hypothesis (H_a or H_1)**: A statement that there is an effect or a difference. This is what the researcher aims to support.

Type I and Type II Errors

- **Type I Error (α)**: Rejecting the null hypothesis when it is actually true. The significance level (α) is the probability of making a Type I error.
- **Type II Error (β)**: Failing to reject the null hypothesis when it is actually false. Power of the test ($1 - \beta$) is the probability of correctly rejecting a false null hypothesis.

P-values and Significance Levels

- **P-value**: The probability of obtaining a test statistic as extreme as the one observed, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null hypothesis.
- **Significance Level (α)**: A threshold chosen by the researcher (commonly 0.05) to decide whether to reject the null hypothesis. If the p-value is less than α, the null hypothesis is rejected.

Steps in Hypothesis Testing

1. **State the hypotheses**: Formulate the null and alternative hypotheses.
2. **Choose the significance level (α)**: Decide the threshold for rejecting the null hypothesis.
3. **Collect and summarize the data**: Calculate the test statistic based on sample data.
4. **Calculate the p-value**: Determine the probability of observing the test statistic under the null hypothesis.
5. **Make a decision**: Compare the p-value with α\alphaα to decide whether to reject or fail to reject the null hypothesis.

Common Tests

- **Z-test**: Used when the population variance is known, and the sample size is large.
- **T-test**: Used when the population variance is unknown and the sample size is small.
- **Chi-square test**: Used for testing relationships between categorical variables or goodness of fit.
- **ANOVA (Analysis of Variance)**: Used for comparing means across three or more groups.