Inferential Statistics -Assignment(06-07-2024)

(Haritha P V)

Inferential statistics involves making predictions or inferences about a population based on a sample of data. Used to make generalizations about a population, test hypotheses, and determine relationships between variables.

Population and Sample:

- **Population**: The entire group of individuals or instances about whom we hope to learn.
- **Sample**: A subset of the population used to collect data and make inferences.

Parameter and Statistic:

- **Parameter**: A measurable characteristic of a population (e.g., mean, variance).
- **Statistic**: A measurable characteristic of a sample used to estimate a population parameter.

Sampling

Sampling is the process of selecting a sub-group of data points from the population based on a certain logic. This logic is provided by the type of technique used.

- **Simple Random Sampling**: Every individual has an equal chance of being selected.
- **Stratified Sampling**: The population is divided into subgroups, and random samples are taken from each subgroup.
- **Cluster Sampling**: The population is divided into clusters, and entire clusters are randomly selected.
- **Systematic Sampling**: Every nth individual is selected from a list of the population.

Central Limit Theorem

The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean approaches a normal distribution, regardless of the population's distribution, as the sample size becomes larger.

The standard deviation of the means of the samples is called “**Standard Error**”. It is denied by the formula,

Standard Error=$\sigma/\sqrt{n}$

where, σ= standard deviation of the population(use sample standard deviation “s” if population standard deviation is unknown), n= sample size

The difference between the stated value and the calculated sample value is called “**Sampling Error**”.

Sampling Error= population parameter-sample statistic

Estimation

1. **Point Estimation**

A single value estimate of a population parameter (e.g., sample mean as an estimate of population mean).We are not sure how accurate this point estimated parameter is, which is a drawback.

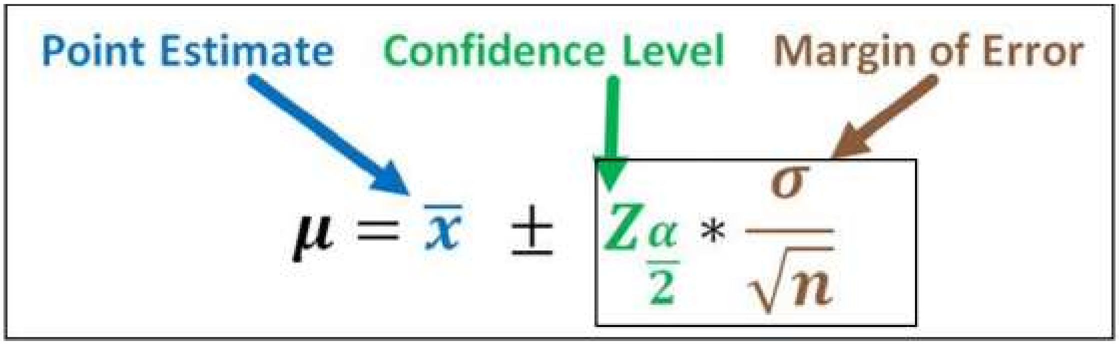
2. Interval Estimation

While deriving / summarizing our sample, if we define one estimated range instead of point estimation, then this is considered as an Interval Estimation.

Confidence Interval

A range of values used to estimate the population parameter. It provides a level of confidence that the parameter lies within the interval.

The range of the values from the point estimate on either side till the error magnitude is called **“ Margin of Error”**



μ is the mean of the population in the interval range given,

$\bar{x}$ is the sample mean,

$Z\alpha/2$ is the critical value corresponding to the desired confidence level

σ is the population standard deviation,

n is the sample size.

The term $Z\alpha/2$ refers to the critical value from the standard normal distribution that corresponds to the tail area of α/2. This value is crucial in constructing confidence intervals for population parameters when using the normal distribution,which is the "**level of significance**".

Confidence Interval is dependent on:

1. **Sample Size:** More the sample size, lesser is the margin of error and narrower will be the confidence interval
2. **Variability of Population:** More the Variability in population Data, more will be the standard deviation of population and more will be the Margin of Error
3. **Confidence Level:** Higher the confidence level, wider will be the confidence interval

1. Type I and Type II Errors:
   - Type I Error: Rejecting the null hypothesis when it is true (false positive).

    - Type II Error: Failing to reject the null hypothesis when it is false (false negative).
2. P-Value:
    - The probability of observing the data, or something more extreme, assuming the null hypothesis is true.
    - If the P-value is less than the significance level (usually 0.05), we reject the null hypothesis.
3. Test Statistics:
    - 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 categorical data to assess how likely it is that an observed distribution is due to chance.
    - ANOVA (Analysis of Variance): Used to compare the means of three or more samples.

Hypothesis and Hypothesis Testing

A hypothesis in statistics is a testable claim or assumption about a parameter of the population. It should be capable of being tested, either by experiment or observation.

Types of Hypotheses

a) Null Hypothesis (H0):

- States that there is no variation in the outcome or no real effect.
- Examples:
    - Special training on students does not affect their performance.
    - Different teaching methods do not affect students' performance.
    - The drug used for headaches does not affect the application.

b) Alternate Hypothesis (Ha):

- Contrasting statement to H0, suggesting there is a real effect.
- Examples:
    - Special training on students has a significant effect.
    - Different teaching methods have a significant effect on students' performance.
    - The drug used for headaches has a significant effect after application.

Hypothesis Testing Process

1. Define Hypotheses:
   - Formulate H0 and Ha.
2. Assign Confidence Level:
   - Typically, a 95% confidence level is used ($\alpha=0.05$), meaning there's a 5% chance of incorrectly rejecting H0.
3. Determine the Test:
   - Decide on a left-tailed, right-tailed, or two-tailed test based on the hypotheses.
4. Conduct the Test:
   - Use one of the following methods:
     - Critical value approach
     - p-value approach
     - Confidence interval approach
5. Make a Decision:
   - Compare the test statistic to the critical value or compare the p-value to α to accept or reject H0.

Test Types

Two-Tailed Test:

- Used when testing if the observed mean is equal to the hypothesized mean.
- H0 includes "=" (e.g., $\mu=\mu_0$).

One-Tailed Test:

- Used when testing if the observed mean is significantly greater or less than the hypothesized mean.
- Right-tailed test: Ha includes ">" (e.g., $\mu>\mu_0$).
- Left-tailed test: Ha includes "<" (e.g., $\mu<\mu_0$).

Methods for Hypothesis Testing

a) Critical Value Approach:

Steps:

1. Define the null and alternate hypotheses.
2. Choose the significance level (α).

3. Determine the critical value(s) from the appropriate distribution (Z or T).
4. Compute the test statistic from the sample data.
5. Compare the test statistic to the critical value to decide on H0.

Formulas:

- Left-tailed test: critical=scipy.stats.norm.ppf(α)   (Z-distribution)

  critical=scipy.stats.t.ppf(α,n−1)   (T-distribution)

- Right-tailed test: critical=scipy.stats.norm.isf(α)   (Z-distribution)

  critical=scipy.stats.t.isf(α,n−1)(T-distribution

- Two-tailed test:
  - Use the appropriate formula based on the sign of the test statistic.

b) p-value Approach:

Steps:

1. Define the null and alternate hypotheses.
2. Choose the significance level (α).
3. Compute the test statistic from the sample data.
4. Determine the p-value from the appropriate distribution.
5. Compare the p-value to α\alphaα to decide on H0.

Formulas:

- Left-tailed test

  p_value= scipy.stats.norm.cdf(test_stat) using Z-distribution for "σ "(known)

  p_value= scipy.stats.t.cdf(test_stat,n-1) using T-distribution for "σ "(unknown)

- Right-tailed test

  p_value= scipy.stats.norm.sf(test_stat) using Z-distribution for “σ “(known)

  p_value= scipy.stats.t.sf(test_stat,n-1) using T-distribution for “σ “(unknown)

- Two-tailed test:
  - Compute the p-value for the appropriate tail and multiply by 2.

c) Confidence Interval Approach:

Steps:

1. Define the null and alternate hypotheses.
2. Choose the confidence level (1−α).
3. Compute the confidence interval for the parameter.
4. Check if the hypothesized parameter lies within the confidence interval to decide on H0.