Descriptive Statistics vs Inferential Statistics

Descriptive Statistics

Descriptive Statistics helps in understanding characteristics of data without making any inferences or drawing conclusions beyond the data itself. The goal of descriptive statistics is to organize, summarize, and present data in a meaningful way, making it easier to comprehend and analyze.

- **Measures of Central Tendency (mean, median, mode):** These indicate the center or average value of a dataset.
- **Measures of Dispersion (range, variance, standard deviation):** These quantify the spread or variability of the data.
- **Frequency Distributions and Histograms:** These show how often different values occur in a dataset.
- **Graphs, charts, and plots**: These visually represent the data.

Inferential statistics

Inferential statistics are used to make predictions or inferences about a population based on a sample of data.

- **Hypothesis Testing**: This involves making a statement about a population parameter and then using sample data to either support or refute that statement.
- **Confidence Intervals**: These provide a range of values within which a population parameter is likely to fall.
- **Regression Analysis**: This helps to understand the relationship between variables and make predictions based on that relationship.

Central limit theorem

In simpler terms, if you take a large number of random samples from a population, calculate the mean of each sample, and then plot those means, the resulting distribution will be approximately normal, even if the original population is not normally distributed.

Alpha (α) vs P-value

Alpha (α):

- Alpha, often denoted as α, is the significance level of a hypothesis test. It represents the probability of rejecting a true null hypothesis.
- Commonly used alpha levels are 0.05 (5%) and 0.01 (1%), but other values can be chosen depending on the specific requirements of the study.
- If you set α = 0.05, for example, it means that you're willing to accept a 5% chance of incorrectly rejecting a true null hypothesis.

P-value:

- The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed sample statistic, assuming that the null hypothesis is true.
- It provides a way to quantify the strength of evidence against the null hypothesis. A small p-value (usually less than the chosen α level) indicates strong evidence against the null hypothesis, leading to its rejection.
- A large p-value suggests that the observed data is consistent with the null hypothesis, and there's not enough evidence to reject it.

- ❖ **If the p-value is less than or equal to α, you reject the null hypothesis.**
- ❖ **If the p-value is greater than α, you fail to reject the null hypothesis.**

Type I Error (False Positive) vs Type II Error (False Negative):

Type I Error (False Positive):

- A Type I error occurs when you reject a null hypothesis that is actually true. In other words, you conclude that there is a significant effect or difference when there isn't.
- The probability of committing a Type I error is denoted by the symbol α (alpha), which is the chosen significance level. For example, if you use a significance level of 0.05, you're willing to accept a 5% chance of making a Type I error.

Type II Error (False Negative):

- A Type II error occurs when you fail to reject a null hypothesis that is actually false. In other words, you miss detecting a real effect or difference.
- The probability of committing a Type II error is denoted by the symbol β (beta). The complement of β, which is 1 - β, is known as the power of the test. Power is the probability of correctly rejecting a false null hypothesis (i.e., finding a true effect).

One-Tail Test vs Two-Tail Test:

One-Tail Test:

- A one-tail test (also known as a one-sided test) is used when you want to test a specific direction of the effect or relationship between variables. It is suitable when you have a clear hypothesis about the direction of the effect (e.g., you expect a change to be either positive or negative, but not both).
- The critical region (the region where you would reject the null hypothesis) is only on one side of the distribution (either the right tail or the left tail).
- For example, if you're testing whether a new treatment is more effective than an existing one, you would use a one-tail test because you're only interested in the direction of improvement.
- The null hypothesis for a one-tail test typically contains a "less than" or "greater than" sign, indicating the direction of interest.

Example:

- **Null Hypothesis (H0):** The new drug is no more effective than the existing drug (μ_new ≤ μ_existing).
- **Alternative Hypothesis (H1):** The new drug is more effective than the existing drug (μ_new > μ_existing).

Two-Tail Test:

- A two-tail test (also known as a two-sided test) is used when you want to test for the possibility of an effect in both directions. It is suitable when you do not have a specific hypothesis about the direction of the effect, and you want to see if there is a significant difference in either direction.
- The critical region is split between both tails of the distribution.
- For example, if you're testing whether there is a difference in IQ scores between two groups, but you don't have a specific hypothesis about which group will score higher, you would use a two-tail test.
- The null hypothesis for a two-tail test typically contains an "equals" sign, indicating that you are testing for any difference, not a specific direction.

Example:

- **Null Hypothesis (H0):** There is no difference in mean IQ scores between the two groups (μ_group1 = μ_group2).
- **Alternative Hypothesis (H1):** There is a difference in mean IQ scores between the two groups (μ_group1 ≠ μ_group2).

Hypothesis Testing

Z-test : A z-test is a statistical hypothesis test used to compare a sample mean to a known population mean when the sample size is large (typically n ≥ 30) and the population standard deviation is known.

t-test : A t-test is a statistical hypothesis test used to determine if there is a significant difference between the means of two groups. It's particularly useful when you have a small sample size (typically less than 30) and you want to make inferences about a population. There are two main types of t-tests: the one-sample t-test and the two-sample t-test.

1. One-sample t-test : It's used to determine if the sample mean is significantly different from a known or hypothesized population mean.
Example: Suppose you want to know if the average height of a sample of 30 students is significantly different from the national average height, which is known to be 65 inches. You collect the sample heights and perform a one-sample t-test.

- **Null hypothesis (H0):** The sample mean is equal to the population mean ($\mu = 65$).
- **Alternative hypothesis (H1):** The sample mean is significantly different from the population mean ($\mu \neq 65$).

2. Two-sample t-test : It's used to determine if there is a significant difference between the means of two independent groups.
Example: Suppose you want to know if there's a significant difference in the exam scores between two different teaching methods. You have two groups: Group A taught with Method 1 and Group B taught with Method 2.

- **Null hypothesis (H0):** The means of the two groups are equal ($\mu 1 = \mu 2$).
- **Alternative hypothesis (H1):** The means of the two groups are not equal ($\mu 1 \neq \mu 2$).

What is Degree of Freedom? Why we need them?

- In a t-test, degrees of freedom are used to select the appropriate t-distribution for making inferences about the population(s) based on sample data. The choice of distribution affects the critical values that determine whether the test statistic falls in the rejection region.
- The concept of degrees of freedom is especially relevant when working with t-tests.

Here's **why degrees of freedom are important** in different types of t-tests:
One-Sample t-test:

- In a one-sample t-test, degrees of freedom ($df=n-1$) represent the number of independent pieces of information in the sample data.
- They are used to determine the critical value from the t-distribution, which helps decide whether to reject the null hypothesis.

Two-Sample t-test:

- The degrees of freedom in a two-sample t-test are calculated using a more complex formula that depends on the sample sizes and variances of the two groups.
- They help determine the shape of the t-distribution, which affects the critical values for the test statistic.

Chi-square (χ^2): Chi-square (χ^2) is a statistical test used to determine if there is a significant association between two categorical variables. It's particularly useful for analyzing data in situations where you have categorical data.
Chi-square test for independence (or association):

- This test is used to determine if there is a significant association between two categorical variables.
- For example, you might want to know if there is a relationship between gender (male or female) and smoking status (smoker or non-smoker) in a population.

Chi-square goodness-of-fit test:

- This test is used to determine if there is a significant difference between the observed and expected frequencies of a categorical variable.
- For example, if you expect a certain distribution of colors of candies in a bag (e.g., 30% red, 30% blue, 40% green) and you actually observe a different distribution, you can use a chi-square goodness-of-fit test to determine if the difference is significant.

ANOVA: ANOVA , which stands for Analysis of Variance, is a statistical technique used to analyze the differences among group means in a sample. It helps determine whether there are any statistically significant differences between the means of two or more groups. It's important to note that ANOVA is applicable when you are comparing means, not individual data points.

Correlation test: A correlation test in statistics is used to measure and quantify the relationship between two or more numerical variables. It assesses the strength and direction of the association between the variables. The correlation coefficient, which ranges from -1 to +1, indicates the degree to which the variables tend to move together.

- **Pearson's Correlation Coefficient (r):** This measures the linear relationship between two continuous variables. It assumes a linear relationship between the variables and is sensitive to outliers.
- **Spearman's Rank Correlation Coefficient (ρ or rho**): This assesses the strength and direction of association between two variables, but without assuming a linear relationship. It is based on the rank order of the data points and is more robust to outliers.

Hypothesis Test	**Purpose**	**When to Use**	**Example**
Z-test	The Z-test is used when you want to compare a sample mean to a known population mean, assuming that you know the population standard deviation.	When the sample size is large (usually n > 30). When the population standard deviation is known.	Suppose you want to test whether the average height of a sample of 100 adults is significantly different from the known population mean height of 65 inches, with a known population standard deviation of 5 inches.
t-test	The t-test is used when you want to compare the means of two groups or when you want to test the difference between a sample mean and a known population mean, but you don't know the population standard deviation.	When comparing two groups or samples. When the population standard deviation is unknown.	You want to test if there is a significant difference in exam scores between two groups of students (Group A and Group B). You collect sample data and perform a t-test to determine if there is a statistically significant difference between the groups.
Chi-square test	The Chi-square test is used to assess the association between categorical variables. It tests whether there is a significant relationship or association between the variables.	When dealing with categorical data (e.g., counts or proportions in different categories). When you want to test for independence or association between two categorical variables.	You want to investigate if there is a relationship between smoking status (Smoker or Non-smoker) and the occurrence of a specific health condition (e.g., Lung Cancer). You collect data on a sample of individuals and perform a Chi-square test.
ANOVA test	ANOVA is used to compare means across three or more groups. It tests whether there are any significant differences between the means of multiple groups.	When you have more than two groups to compare.	You are studying the effect of different teaching methods (Method A, Method B, Method C) on student test scores. You collect data from three groups of students and want to determine if there is a significant difference in mean scores among the three groups.