Statistics: Assignment no.3

1) Difference between Descriptive and Inferential statistics

	Descriptive Statistics	Inferential Statistics
1	It gives information about raw data which describes the data in some manner.	It makes inferences about the population using data drawn from the population.
2	It helps in organizing, analyzing, and to present data in a meaningful manner.	It allows us to compare data, and make hypotheses and predictions.
3	It is used to describe a situation.	It is used to explain the chance of occurrence of an event.
4	It explains already known data and is limited to a sample or population having a small size.	It attempts to reach the conclusion about the population.
5	It can be achieved with the help of charts, graphs, tables, etc.	It can be achieved by probability.

2) Difference between Population and Sample

	Population	Sample
1.	Universe of elements to be studied.	Selection of a part of the population.
2.	It can be classified according to the number of individuals that make it up.	It is part of the population: it should comprise between 5% and 10% to be most effective.
3.	It has statistical variables.	Variable could be random.

Submitted by Subhamita Kanungo

January 27, 2023

4.	To analyze the data collected regarding the common characteristics shared by the elements for various purposes.	To study the behaviour, characteristics, tastes, or properties of a representative part of the population.
5.	The people who inhabit a country. The number of cars in a city. Students of a university.	Study the performance of students from five universities in a city in a specific subject. 500 students are randomly taken as a sample (100 from each institution) studying at the same level so that the sample is representative.

3) What is hypothesis? Differentiate between Null and Alternative hypothesis.

Hypothesis are statement about the given problem. Hypothesis testing is a statistical method that is used in making a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Example:
You say an average student in the class is 30 or a boy is taller than girls. All those are an example in which we assume or need some statistic way to prove those. We need some mathematical conclusion whatever we are assuming is true.

Need for Hypothesis Testing
Hypothesis testing is an important procedure in statistics. Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data. When we say that the findings are statistically significant.

Statistics: Assignment no.3

	Null Hypothesis	Alternative Hypothesis
1.	In the null hypothesis, there is no relationship between the two variables.	In the alternative hypothesis, there is some relationship between the two variables i.e. They are dependent upon each other.
2.	Generally, researchers and scientists try to reject or disprove the null hypothesis.	Generally, researchers and scientists try to accept or approve the null hypothesis
3.	If the null hypothesis is accepted researchers have to make changes in their opinions and statements.	If the alternative hypothesis gets accepted researchers do not have to make changes in their opinions and statements.
4.	Here no effect can be observed i.e., it does not affect output.	Here effect can be observed i.e., it affects the output.
5.	Here the testing process is implicit and indirect.	Here the testing process is explicit and direct.
6.	This hypothesis is denoted by H0.	This hypothesis is denoted by Ha or H1.
7.	It is generally used when we reject the null hypothesis.	It gets accepted if we fail to reject the null hypothesis.
8.	In this hypothesis, the p-value is smaller than the significance level.	In this hypothesis, the p-value is greater than the significance level.

Submitted by Subhamita Kanungo

January 27, 2023

Statistics: Assignment no.3

4) What is Central Limit Theorem?

Central Limit Theorem:

The Central Limit Theorem states that as the sample size grows higher, the sample size of the sampling values approaches a normal distribution, regardless of the form of the data distribution. The mean of sample means will be the population mean, according to the Central Limit Theorem.
Likewise, if you average all the degrees of separation in your sample, you'll get the population's true standard deviation.

- The sample mean is equal to the population mean.
- The sample standard deviation is equal to the population standard deviation divided by the square root of the sample size.

5) Difference between Type-I and Type-II Error?

Type -I Error

- It is also known as a false-positive.
- It occurs if the researcher rejects a correct null hypothesis in the population. i.e., incorrect rejection of the null hypothesis.
- Measured by alpha (significance level).
- If the significance level is fixed at 5%,
- It means there are about five chances of type – 1 error out of 100.

Type -II Error

- It is also known as a false negative.
- It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.
- Measured by beta (the power of test).
- The probability of committing a type -2 error is calculated by 1 – beta (the power of test).

Submitted by Subhamita Kanungo

January 27, 2023

Statistics: Assignment no.3

6)What is Linear Regression?

Linear regression is a statistical technique that models the magnitude and direction of an impact on the dependent variable explained by the independent variables. Linear regression is commonly used in predictive analysis.

7) What are the assumptions required for Linear Regression?

Five main assumptions underlying multiple regression models must be satisfied:

(1) linearity

(2) homoskedasticity

(3) independence of errors

(4) normality

(5) independence of independent variables. Diagnostic plots can help detect

whether these assumptions are satisfied.

8) How is the statistical significance of an insight assessed?

- Statistical significance can be accessed using hypothesis testing:
- Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
- Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
- Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
- We calculate the observed test statistics from the data and check whether it lies in the critical region
  Common tests:
  – One sample Z test
  – Two-sample Z test
  – One sample t-test
  – paired t-test
  – Two sample pooled equal variances t-test
  – Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)
  – Chi-squared test for variances

Submitted by Subhamita Kanungo

January 27, 2023

- Chi-squared test for goodness of fit
- Anova (for instance: are the two regression models equals? F-test)
- Regression F-test (i.e: is at least one of the predictors useful in predicting the response

9) What is Mean?

The mean in math and statistics summarizes an entire dataset with a single number representing the data’s center point or typical value. It is also known as the arithmetic mean, and it is the most common measure of central tendency. It is frequently called the “average.”

10) What is the meaning of standard deviation?

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a “typical” deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore, a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

11) What is correlation?

Correlation refers to the statistical relationship between two entities. In other words, it's how two variables move in relation to one another. Correlation can be used for various data sets, as well. In some cases, you might have predicted how things will correlate, while in others, the relationship will be a surprise to you. It's important to understand that correlation does not mean the relationship is causal.

Positive correlation: A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.

Negative correlation: A negative correlation is -1. This means the two variables moved in opposite directions.

Zero or no correlation: A correlation of zero means there is no relationship between the two variables. In other words, as one variable moves one way, the other moved in another unrelated direction

Statistics: Assignment no.3

12) What is meaning of covariance?

Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

The formula is:
Cov(X,Y) = Σ E((X – μ) E(Y – ν)) / n-1 where:

X is a random variable

E(X) = μ is the expected value (the mean) of the random variable X and

E(Y) = ν is the expected value (the mean) of the random variable Y

n = the number of items in the data set.

Σ summation notation.

13) Where is inferential statistics used?

Inferential statistics are often used to compare the differences between the treatment groups. Inferential statistics use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.

14) What is one sample t-test?

- The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.
- We can use the test for continuous data. Your data should be a random sample from a normal population.

15) What is the relationship between standard deviation and standard variance?

Variance and Standard Deviation are the two important measurements in statistics. Variance is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data. The basic difference between both is standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units.

Submitted by Subhamita Kanungo

January 27, 2023

Statistics: Assignment no.3

16) What is one-way ANOVA test?

- The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. This guide will provide a brief introduction to the one-way ANOVA, including the assumptions of the test and when you should use this test. If you are familiar with the one-way ANOVA, but would like to carry out a one-way ANOVA analysis.
- The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

  $$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

  where μ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (HA), which is that there are at least two group means that are statistically significantly different from each other.

Submitted by Subhamita Kanungo

January 27, 2023