STATISTICS WORKSHEET-1

1. a)
2. a)
3. b)
4. d)
5. c)
6. b)
7. b)
8. a)
9. c)
10. Normal distribution is an arrangement of a data set in which major values group in the middle of the range and the remaining taper off symmetrically towards either extremes. The distribution of data is normal when the data is symmetric around the mean, when 68% of the data lies within one standard deviation of the mean, 95% of the data lies within two standard deviation and 99.8% of the data lies within the three standard deviations of the mean. All the remaining data are outliers. It is a probability function that describes how the values of a variable are distributed.
11. Missing data can be managed in many ways. Mostly people ignore it. But depending on how much data is missing, ignorance may or may not be a good idea. Another common strategy for those who pay attention is imputation. Imputation is the procedure of replacing an estimation for misplaced values and analysing the entire data set as if the imputed values were the true

observed values. Some of the imputation techniques are-

- Mean imputation- Firstly calculate the mean of the observed values for that variable for all non-missing people. It has the benefit of upholding the same mean and sample size, but it also has a few of disadvantages. Almost all of the methods described below are superior to mean imputation.
- Substitution- Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.
- Hot Deck Imputation- A value picked at chance from a sample member who has similar values on other variables. To put it another way, select all the sample members who are comparable on other factors, then choose one of their lost variable values at random.

  One benefit is that you are limited to just possible values. In other words, if age is only permissible to be between 5 and 10 in your research, you will always obtain a value between 5 and 10. Another factor is the random element, which introduces some variation. For exact standard errors, this is vital.
- Regression Imputation- The result of regressing the absent variable on other factors to get a projected value. As a result, instead of utilising the mean, you're trusting on the anticipated value, which is

influenced by other factors. This keeps the links between the variables in the imputation model, but not the variability around the anticipated values.

12. A/B testing denotes the experimentations where two or more variations of the same webpage are related against each other by showing them to real-time visitors to determine which one performs better for a given goal. A/B testing is not limited by web pages only, you can A/B test your emails, popups, sign up forms, apps and more. A/B testing, at its most basic, is a way to check comparability between two versions of something to figure out which executes better. While it's most often associated with websites and apps, Fung says the method is almost 100 years old. You start an A/B test by deciding what it is you want to test.

13. The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is classically considered awful practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

    Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is not accurate and the confidence interval is narrower.

14. Linear regression is an elementary and frequently used type of analytical analysis. The overall idea of regression is to scrutinize two things: (1) does a set of analyst variables do a good job in predicting an outcome of dependent variable? (2) Which variables in particular are important predictors of the outcome variable, and in what way do they indicated by the scale and sign of the beta estimates impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regress and. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) predicting an effect, and (3) trend estimating.

15. There are two real branches of statistics: **descriptive statistics and inferential statistics.**

- Descriptive statistics- This branch of statistics focuses on collecting, summarizing, and presenting a set of data. The first aspect of statistics is descriptive statistics, which deals with the presentation and collection of data. It is not as simple as it appears. The statistician must know how to design and experiment, select the appropriate focus group, and prevent biases that are too easy to introduce into the experiment. Generally, descriptive statistics can be categorized into

  Measures of central tendency

  Measures of variability

  To understand both measures of tendency and variability, easily use graphs, tables, and general discussions.
- Inferential statistics- The branch of statistics that analyses sample data to draw conclusions about a population is inferential statistics. When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer.