Statistics Worksheet-1

1) a
2) a
3) b
4) d
5) c
6) b
7) b
8) a
9) c

Question 10) What do you understand by the term Normal Distribution?

Answer. Normal distribution is also known as Gaussian distribution where it is bell shaped curve. 68% of data points lie between standard deviation +/- 1 standard deviation, 95% of data points lie between standard deviation +/-2 standard deviation and 99.7% are within +/- 3 standard deviation from the mean. It has zero skewness and kurtosis of 3.

Question 11) How do you handle missing data? What imputation techniques do you recommend?

Answer. Usually missing data is a common occurrence in a large dataset. Missing values are usually represented by NaN values in the dataset. A code isnull().sum() can help us identify missing values in a dataset and there after we can treat them in various ways which are listed below.
We can use .dropna function to delete the entire row or use pairwise deletion to delete the NaN values. If the data is continuous or discrete we can use .fillna with the mean, median or mode approach. If the data is a string then we can use .fillna and replace it with the most occurring string in that column.

Question 12) What is A/B testing?

Answer.

Question 13) Is mean imputation of missing data acceptable practice?

Answer. Yes, it is but it depends on the data to be imputed. If it’s a continuous data and the data points are fairly in range of each other with no outliers then it can be be used. Having said that, if we can somehow drop the outliers based on our understanding and if it makes sense to do so, we can use mean imputation. However, the non-negotiable part is the data has to be continuous or discrete in-order to use mean imputation to replace missing data.

Question 14) What is linear regression in statistics?

Answer) It is a kind of descriptive statistics where in the target variable “y” is a numeric value i.e. continuous or discrete. It uses an equation of line y=mx+c where m is the slope, c is the intercept, x is an independent variable(predictor) and y is a target variable or predicted variable. The errors or the metrics between the predicted “y” and test “y” can be found out using least squared method or root mean squared method or mean absolute error.

Question 15) What are the various branches of statistics?

Answer) There are two branches of statistics namely: a) Descriptive and b) Inferential statistics

a) Descriptive statistics: It describes the data using mean, median, mode, standard deviation or variance. It is subdivided into linear and logistic regression. It is represented by graphs and charts. They tell us more about the data which is already available.

b) Inferential statistics: It describes the data taken as a sample from population. It is represented by probability. It describes the data by using z-score hypothesis testing, anova testing or chi-squared testing, regression or classification. They use predictive analysis to predict the outcome.