STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False

   Ans : True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned

   Ans : Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned

   Ans : Modeling bounded count data

4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned

   Ans : All of the mentioned

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

   Ans : Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

   Ans : False

7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

   Ans : Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10

   Ans : 0

9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

   Ans : Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

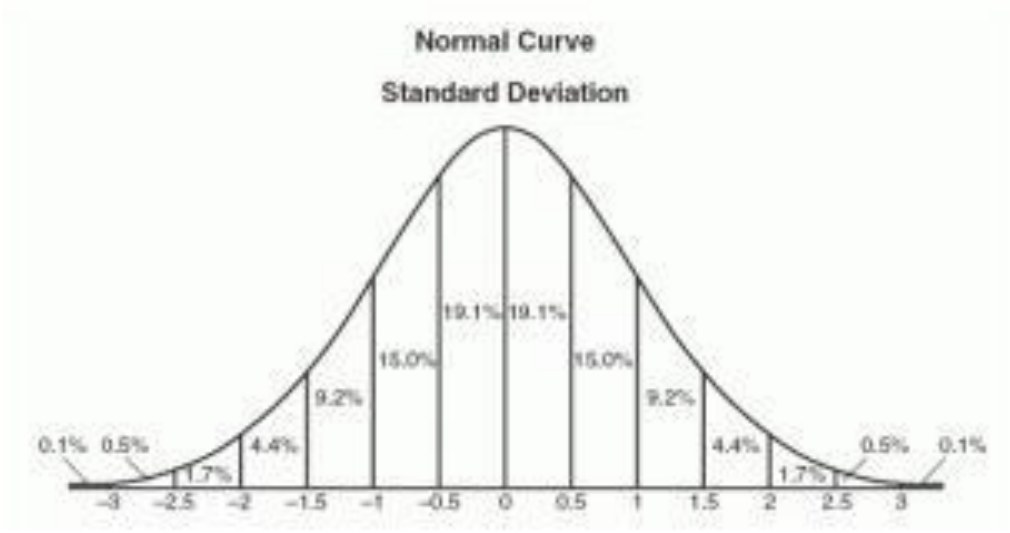
10. What do you understand by the term Normal Distribution?

Ans :

The data which is symmetrically distributed without skew is called normal distribution. When the data is plotted in graph it shows bell shaped curve. The most of the values are distributed towards center region.

The graph of the normal distribution is characterized by two parameters called mean, or average, which is the maximum of the graph and about which the graph is always symmetric and the second one is standard deviation, which determines the amount of dispersion away from the mean.

The normal distribution graph looks like below,



The random variables following the normal distribution are those whose values can find any unknown value in a given range. The normal distribution doesn't even bother about the range. The range can also extend to $-\infty$ to $+\infty$ and still we can find a smooth curve. These random variables are called Continuous Variables.

The normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out. If the standard deviation is smaller, the data are somewhat close to each other and the graph becomes narrower. If the standard deviation is larger, the data are dispersed more, and the graph becomes wider.

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

Ans:

It depends on what type of data is missing, how much is missing, and if it's randomly missing. We can do simple fixes for randomly missing data like taking mean, median or mode values. We can also create machine learning models based on other factors to predict missing values.

If missing data is not random, we will have problems and likely run into issues imputing. However, the fact that the value is missing can be informative, and flagging the data as missing can be useful, particularly for category data.

There are several machine learning protocalls which helps in high dimensional complex data imputation techniques, like KNN, RandomForest, Regression etc. each of them have their own pros and cons. So here the imputation is depended on the type of data we have and missing.

Skewness represents a distribution's degree of symmetry. Since the normal distribution is perfectly symmetric, it has a skewness of zero. In other distributions with a skewness less than or greater than zero, the left tail (left skewness) or the right tail (right skewness) will be longer, respectively.

12. What is A/B testing?

Ans :

A/B testing is a type of experiment in which we split our data or user base into two groups and show two different versions with the goal of comparing the results to find the more successful version. With A/B test one element is changed between the original and the test version to see if this modification has any impact on user behavior or conversion rates.

The null hypothesis, or H0, posits that there is no difference between two variables. In A/B testing, the null hypothesis would assume that changing one variable on data would have no impact on user behavior. The results of an A/B test are not due to rejecting the null hypothesis. This is calculated by measuring the p-value, or probability value. So, if the p-value is low, it is saying that it's unlikely the results of the A/B test were random. A rule of thumb tends to be that when the p-value is 5% or lower, the A/B test is much significant.

It is indeed fundamental to determine how likely it is that the observed discrepancy between the two samples originates from chance. In order to do that, we will use two sample hypothesis test. Our null hypothesis H0 **is that the two designs A and B have the same efficacy, i.e. that they produce an equivalent click-through rate, or average revenue per user, etc. The statistical significance is then measured by the** p-value**, i.e. the probability of observing a discrepancy between our samples at least as strong as the one that we actually observed.**

13. Is mean imputation of missing data acceptable practice?

Ans :

Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Imputing the mean preserves the mean of the observed data. So if the data are missing completely at random the estimate of the mean remains unbiased. By imputing the mean, we are able to keep your sample size up to the full sample size. If we are estimating means and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

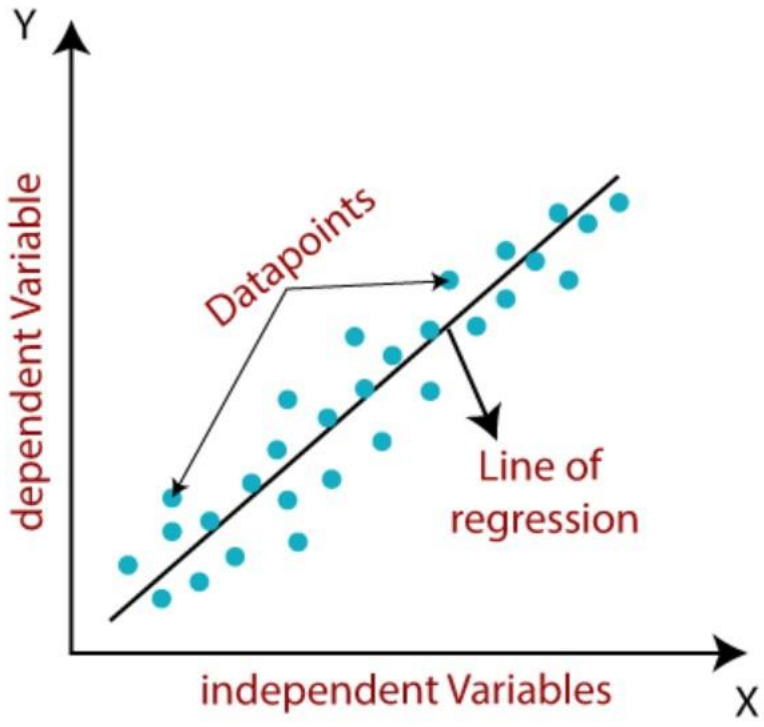
Mean imputation leads to an underestimate of standard errors. get the same mean from mean-imputed data that we would have gotten without the imputations. There are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them.

So according to me its all depends on the type of base data which we are handling. Also we have several best models for imputations depends on the situations we can switch to whichever is best.

14. What is linear regression in statistics?

Ans :

Linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. One variable, denoted *x*, is regarded as the predictor, explanatory, or independent variable. And the other variable, denoted *y*, is regarded as the response, outcome, or dependent variable.



This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line or Line of Regression for a set of paired data. then estimate the value of X (dependent variable) from Y (independent variable).

For a linear relationship, we can use a model of the form

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where y is the dependent or response variable and x is the independent or predictor variable. The random variable 1 is the error term in the model. In this context, error does not mean mistake but is a statistical term representing random fluctuations, measurement errors, or the effect of factors outside of our control.

Linear regression is of two different types such as the following:

1. **Simple linear regression: Simple linear regression is defined as linear regression with a single predictor variable. An example of a simple linear regression is $y = \beta_0 + \beta_1 x + \varepsilon$**

2. **Multiple linear regression: Multiple linear regression is defined as linear regression with more than one predictor variable along with its coefficients. An example of multiple linear regression is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots\ldots + \varepsilon$.**

Before you attempt to perform linear regression, you need to make sure that your data can be analysed using this procedure. Your data must pass through certain required assumptions.

Here's how you can check for these assumptions :

1. **The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.**
2. **Use a scatterplot to find out quickly if there is a linear relationship between those two variables.**
3. **The observations should be independent of each other (that is, there should be no dependency).**
4. **Your data should have no significant outliers.**
5. **statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.**
6. **The residuals (errors) of the best-fit regression line follow normal distribution.**

The linearity of the linear relationship can be determined by calculating the t- test statistic. The t-test statistic helps to determine how linear, or nonlinear.

15. What are the various branches of statistics?

Ans :

Statistics is a method of interpreting, analyzing and summarizing the data. Hence, the types of statistics are categorized based on these features: Descriptive and inferential statistics. Based on the representation of data such as using pie charts, bar graphs, or tables, we analyses and interpret it.

Statistics have majorly categorised into two types:

1. **Descriptive statistics**
2. **Inferential statistics**

Descriptive statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean standard deviation .

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Inferential statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.