1. Bernoulli random variables take (only) the values 1 and 0.
A. True. Bernoulli distribution can arise only when there are two or binary outcomes.

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. Center Limit theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
B. Modelling bounced count data

4. Point out the correct statement.
D. All of them

5. ______ random variables are used to model rates.
C. Poisson Distribution

6. Usually replacing the standard error by its estimated value does change the CLT.
B. False

7. Which of the following testing is concerned with making decisions using data?
B. Hypothesis

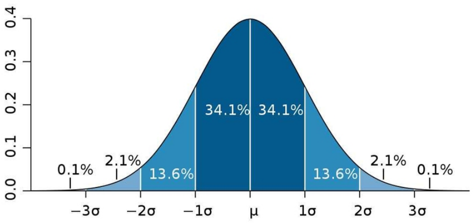
8.Normalized data are centred at______and have units equal to standard deviations of the original data.
A.0

9.Which of the following statement is incorrect with respect to outliers?
C. Outliers cannot conform to the regression relationship

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

Normal distribution is a bell-shaped curved distribution where the mean, median, and mode all are at same location in the curve. On the chart, left hand side is minus infinity and right-hand side is positive infinity.

Population mean (μ) and population standard deviation (σ) will decide the shape and location of the curve. If the standard deviation is higher, the curve will be flatter and if the standard deviation is less, the curve will be flatter.



The above curve shows the normal distribution. Also it represents the rule of 68-95-99.7% rule. On x axis we see the distance and on y axis we see probabilities.

In ideal conditions, when we go 1σ away from mean, we cover 68% of total area, meaning 68% of total datapoints lies in this area.
when we go 2σ away from mean, we cover 95% of total area, meaning 95% of total datapoints lies in this area.
And when we go 3σ away from mean, we cover 99.7% of total area, meaning 99.7% of total datapoints lies in this area.

Remaining 0.3% data lies on any left-out area of any side of the curve. The points lying far away from 3σ, are considered to be the outliers. They drag the bell curve on left or right side as unlike median, mean is highly affected by the outliers.

The parameter we use to locate the datapoint is called z, and to calculate z score we can use below formula

$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

By calculating z scores, we can find out the outliers and we can remove them from the dataset.

11. How do you handle missing data? What imputation techniques do you recommend?
There are situations when the dataset has missing values and we need to deal with them. When there are very fewer missing data, we can simply omit the complete row. But when there are much more data missing from the dataset, we need to use another approach to tackle such issue.

First of all, we need to ask to the client if they can help us with the missing data from their data sources. If they say that they can support us with the missing data then we should go with their support.

If client says that we cannot support on this, then we need to try with different methods. One of the replacements for missing data can be replacing them with mean of that particular variable. The reason that we are using mean as a replacement is that mean is considered to be the representative of that particular variable.

Another value that we can use as a replacement for the missing values is median. Median is widely used replacement for the missing data is because mean is easily affected by the outliers easily where is median do not get any impact with outliers or it has very less impact from the outliers.

12. What is A/B testing?
A/b testing is an experiment where there are two or more variables / products are used and observed which performs better under the controlled environment.

For example,
Company designed two products, one they say product A and another they say product B. So, company can ask the group of people to give their preference to both the products and based on this outcome, company can decide that which product they can go with.

13. Is mean imputation of missing data acceptable practice?
Mean is not much acceptable practice for imputation. There are some limitations due to which mean is not acceptable practice.

The first reason behind that is it do not preserve the relationship among the variables. And most of time, we are after the relationships among the variables.

Mean is also impacted by the outliers. Hence one outlier can make entire normal distribution into skewed distribution. So instead, we should use median or any other imputation technique like MICE.

A second reason  is any statistic that uses mean as the imputed data will have a standard error that's too low. The reason behind that is after imputing mean, when we get new standard deviation, it will be extremely small and we know very well that this standard deviation is not ideal to take into consideration.

14. What is linear regression in statistics?

The linear regression is understanding the relationship between dependent variable and independent variable.

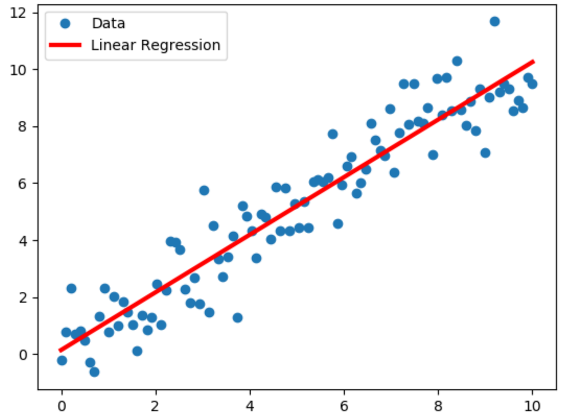
This relationship shows that if there is one unit change in independent variable(x), how much will be the change in dependent variable(y).

It is called as linear regression because it is based on the assumption that there is linear relationship between dependent variable and independent variable.

While we perform linear regression, it tries to draw a linear line that tries to best cover most points on the chart.

y = c + b*x

This is the formula for deciding the slope and intercept of the line where,
 C is intercept
 B is conefficient
X is independent variable



As we can see in the chart above, the red line passes through most of the datapoints. The line should be drawn in such a manner that most of the datapoints stays near the line and cover most of the datapoints.

The points which are near the line generated the error. By summing the square of all the errors, we get sum of squared errors. This error must be minimum, so that the model / line that we have fitted performs at the best. Lesser the error, better the performance of the model.

Simple Linear regression:  When there is one is one independent variable and one dependent variable then we have to use simple linear regression.

Multiple Linear regression: when there are more than one independent variable and one dependent variable then we have to use multiple linear regression.

While selecting the variable in multiple linear regression, we need to make sure that there should not be high collinearity among the independent variables and independent variables should have good correlation we dependent variable so that they can contribute to the prediction of y.

With the help of this model, we can also predict the value if y, when we give the value of x. Hence linear regression is the part of predictive statistics.

15. What are the various branches of statistics?

There are two branches of statistics

a. Descriptive statistics: descriptive statistics talks about describing the data on the basis of statistics like mean, median, mode, and standard deviation et cetera. Descriptive statistics gives us brief idea about the data and a general picture of what data is talking about.
   In descriptive statistics, we can analyse single variable which is also called as univariate analysis which consist of mean, median, mode, interquartile range, standard deviation (five-point summary) et cetera.
   When we analyse more than one variable, that is called multivariate analysis where we can check the relationships between variables. These relationships can be negative or positive which can help us to predict the effect of one variable on other

b. Predictive statistics: predictive statistics also known as inferential statistics. The reason behind this is that on the basis of available data we try to infer or predict something. With the help of statistical methods like regression modelling we try to find the relationships between the variables and on the basis of this relationships we try to predict the future. For example, if there is a change in dependent variable how much will be the change in dependent variable.