1) Bernoulli random variables take (only) the values 1 and0
   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. d) All of the mentioned

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

Ans. c) Poisson

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

Ans. b) False

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

Ans. b) Hypothesis

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

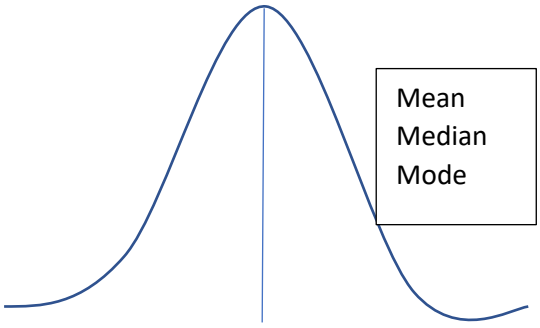
Ans. a) 0

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. c) Outliers cannot conform to the regression relationship

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

Ans. We call this Bell-shaped curve a Normal Distribution. Carl Friedrich Gauss discovered it so sometimes we also call it a Gaussian Distribution as well. We can simplify the Normal Distribution's Probability Density by using only two parameters: $\boldsymbol{\mu}$ Mean and $\boldsymbol{\sigma}2$. This curve is symmetric around the Mean. Also as you can see for this distribution, the Mean, Median, and Mode are all the same.



According to the Empirical Rule for Normal Distribution:

- 68.27% of data lies within 1 standard deviation of the mean
- 95.45% of data lies within 2 standard deviations of the mean
- 99.73% of data lies within 3 standard deviations of the mean

Thus, almost all the data lies within 3 standard deviations. This rule enables us to check for Outliers and is very helpful when determining the normality of any distribution.

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

Ans. Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical program will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea. Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values. And how would you choose that estimate? The following are some of the most prevalent methods:

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all the methods described below are superior to mean imputation.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilizing the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

12) What is A/B testing?

Ans. A/B Testing is a tried-and-true method commonly performed using a traditional statistical inference approach grounded in a hypothesis test (e.g., t-test, z-score, chi-squared test). In plain English, 2 tests are run in parallel:

1. Treatment Group (Group A) - This group is exposed to the new web page, popup form, etc.
2. Control Group (Group B) - This group experiences no change from the current setup.

The goal of the A/B is then to compare the conversion rates of the two groups using statistical inference.

For example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It’s important to note that all other variables need to be held constant when performing an A/B test.

In technical language, A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

13) Is mean imputation of missing data acceptable practice?

Ans. The process of replacing null values in a data collection with the data’s mean is known as mean imputation. Mean imputation is typically considered terrible 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 less accurate, and the confidence interval is narrower.

14) What is linear regression in statistics?

Ans. Linear regression is probably one of the most important and widely used regression techniques. It's among the simplest regression methods. One of its main advantages is the ease of interpreting results.

When implementing linear regression of some dependent variable y on the set of independent variables $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of predictors, you assume a linear relationship between y and $\mathbf{x}$: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. This equation is the regression equation. $\beta_0, \beta_1, \ldots, \beta_r$ are the regression coefficients, and ε is the random error.

Linear regression calculates the estimators of the regression coefficients or simply the predicted weights, denoted with $b_0, b_1, \ldots, b_r$. These estimators define the estimated regression function $f(\mathbf{x}) = b_0 + b_1 x_1 + \cdots + b_r x_r$. This function should capture the dependencies between the inputs and output sufficiently well.

The estimated or predicted response, $f(\mathbf{x}_i)$, for each observation $i = 1, \ldots, n$, should be as close as possible to the corresponding actual response y_i. The differences $y_i - f(\mathbf{x}_i)$ for all observations $i = 1, \ldots, n$, are called the residuals. Regression is about determining the best predicted weights—that is, the weights corresponding to the smallest residuals.

To get the best weights, you usually minimize the sum of squared residuals (SSR) for all observations $i = 1, \ldots, n$: $\text{SSR} = \Sigma_i (y_i - f(\mathbf{x}_i))^2$. This approach is called the method of ordinary least squares.

15) What are the various branches of statistics?

Ans. The various branches of statistics are:

1. Descriptive Statistics: Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc.
    a. Measure of Central Tendency: Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.
        i. Mean
        ii. Median
        iii. Mode
    b. Measure of Variabiliy: Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population.
        i. Range
        ii. Variance
        iii. Dispersion
2. Inferential Statistics: Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive stats. It is simply used to analyze, interpret result, and draw conclusion. Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis. Hypothesis testing is a type of inferential procedure that takes help of sample data to evaluate and assess credibility of a hypothesis about a population. Inferential

statistics are generally used to determine how strong relationship is within sample. But it is very difficult to obtain a population list and draw a random sample.

a. One sample test of difference/One sample hypothesis test
b. T-test or Anova
c. Contingency Tables and Chi-Square Statistic