1. Bernoulli random variables take (only) the values 1 and 0.

a) 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

3. Which of the following is incorrect with respect to use of Poisson distribution?

b) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

5. \_\_\_\_\_\_ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

8. 4. Normalized data are centered 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, also known as the Gaussian distribution, is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a [bell curve](https://www.investopedia.com/terms/b/bell-curve.asp).

* A normal distribution is the proper term for a probability bell curve.
* In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
* Normal distributions are symmetrical, but not all symmetrical distributions are normal.
* In reality, most pricing distributions are not perfectly normal.

The normal distribution model is motivated by the [Central Limit Theorem.](https://www.investopedia.com/terms/c/central_limit_theorem.asp) This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance).

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

Missing data can skew anything for data scientists, from economic analysis to clinical trials. After all, any analysis is only as good as the data. A data scientist doesn’t want to produce biased estimates that lead to invalid results.

When dealing with missing data, [data scientists](http://www.mastersindatascience.org/careers/data-scientist) can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Before deciding which approach to employ, data scientists must understand why the data is missing.

Data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results.

Mean, Median and Mode:

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can [calculate the mean or median of the existing observations](https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4).

12. What is A/B testing?

[A/B testing](https://vwo.com/testing/ab-testing/), also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

Essentially, A/B testing eliminates all the guesswork out of [website optimization](https://vwo.com/website-optimization/) and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

To run an A/B test, you need to create two different versions of one piece of content, with changes to a single **variable**. Then, you'll show these two versions to two similarly sized audiences and analyze which one performed better over a specific period of time (long enough to make accurate conclusions about your results).

13.Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data’s mean is known as mean imputation.imputing the mean preserves the mean of the observed data.  So if the data are [missing completely at random](https://www.theanalysisfactor.com/causes-of-missing-data/), the estimate of the mean remains unbiased. That’s a good thing.Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That’s good too.This is the original logic involved in mean imputation. If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

Mean imputation is typically considered terrible practice since it ignores feature correlation. 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?

Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data, such as in cancer diagnoses or in stock prices.

Linear regression shows a relationship between an independent variable and a dependent variable being studied.

There are a number of ways to calculate linear regression. One of the most common is the ordinary least-squares method, which estimates unknown variables in the data, which visually turns into the sum of the vertical distances between the data points and the trend line.

The calculations to perform linear regressions can be quite complex. Fortunately, linear regression models are included in most major calculations packages, such as Excel, R, MATLAB and Mathematica.

15. What are the various branches of statistics?

The two main branches of statistics are [descriptive statistics](https://explorable.com/descriptive-statistics) and [inferential statistics](https://explorable.com/inferential-statistics). Both of these are employed in scientific analysis of data.

#### Descriptive Statistics:

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

#### Inferential Statistics:

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.