STATISTICS WORKSHEET

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? a) Modeling event/time data .

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 d) All of the mentioned

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 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?

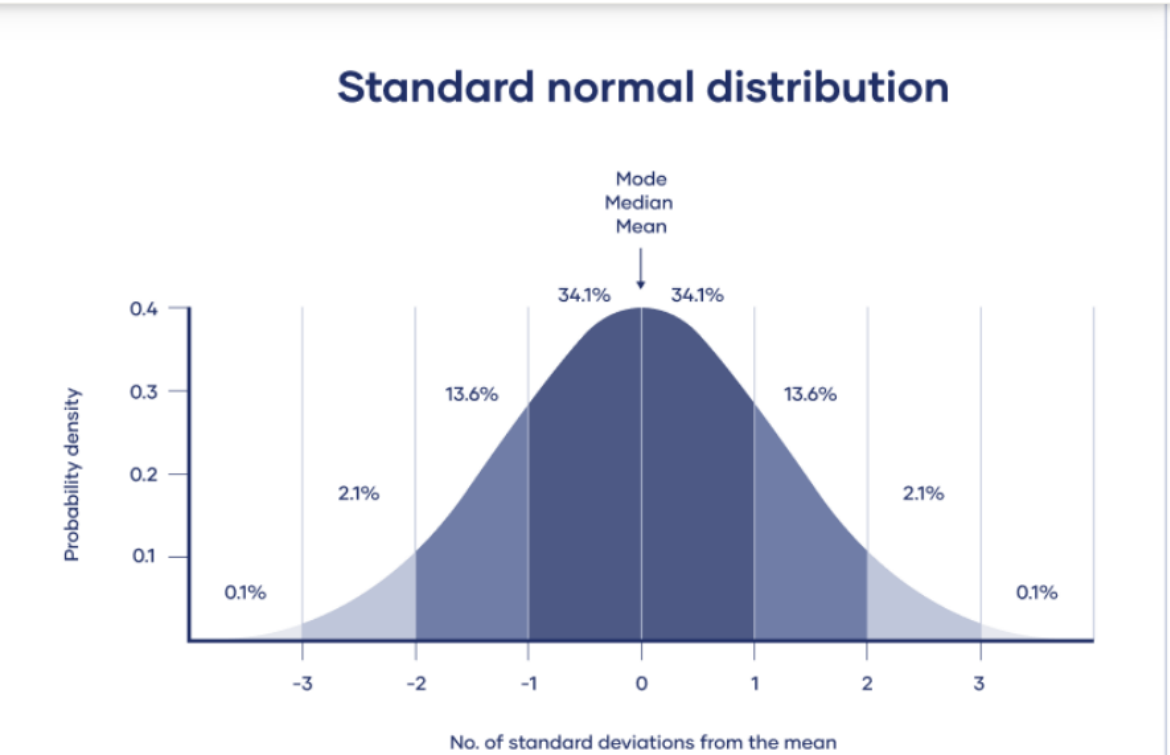
A **Normal distribution** is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

A z-score, or z-statistic, is a number representing how many standard deviations above or below the mean population the score derived from a z-test is. Essentially, it is a numerical measurement that describes a value's relationship to the mean of a group of values.

$$z = (X - \mu) / \sigma$$

where X is a normal random variable
μ is the mean of X
σ is the standard deviation of X.



The normal distribution is the most commonly known and used of all distributions. Because the normal distribution relates to many natural phenomena so well, it has become a standard of reference for many probability problems.

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

When dealing with missing data, we 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.

Mean, Median and Mode imputation technic is the commonly recommended method

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. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. his method does not use time-series characteristics or depend on the relationship between the variables.

12. What is A/B testing?

An **A/B test** is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

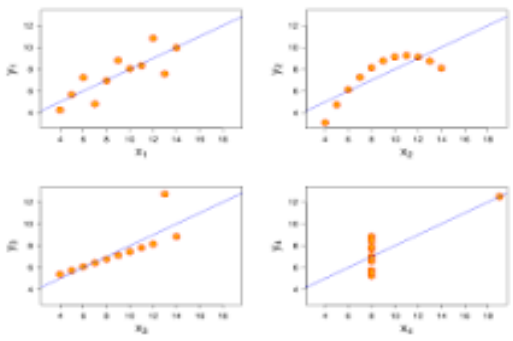
This method compare two versions of a webpage or app against each other to determine which one performs better. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.

13. Is mean imputation of missing data acceptable practice?

No, Mean imputation is not a good practise, because . It does improve power, but your results will be so biased, the improved power won't help much. Mean imputation reduces the variance of the imputed variables. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval. Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).



There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis.
It is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.

15. What are the various branches of statistics?

Statistical methods were classified into four categories: **descriptive methods, parametric inferential methods, nonparametric inferential methods, and predictive methods**.

DESCRIPTIVE METHOD :

Descriptive statistics are used to describe or summarize data in ways that are meaningful and useful. For example, it would not be useful to know that all of the participants in our example wore blue shoes. However, it would be useful to know how spread out their anxiety ratings were.

PARAMETRIC INFERENTIAL METHOD:

Parametric statistical procedures rely on assumptions about the shape of the distribution (i.e., assume a normal distribution) in the underlying population and about the form or parameters (i.e., means and standard deviations) of the assumed distribution.

NON- PARAMETRIC INFERENTIAL METHOD:

Nonparametric inference refers to statistical techniques that use data to infer unknown quantities of interest while making as few assumptions as possible. Typically, this involves working with large and flexible infinite-dimensional statistical model.

PREDICTIVE METHOD

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities.