STATISTICS WORKSHEET-1

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

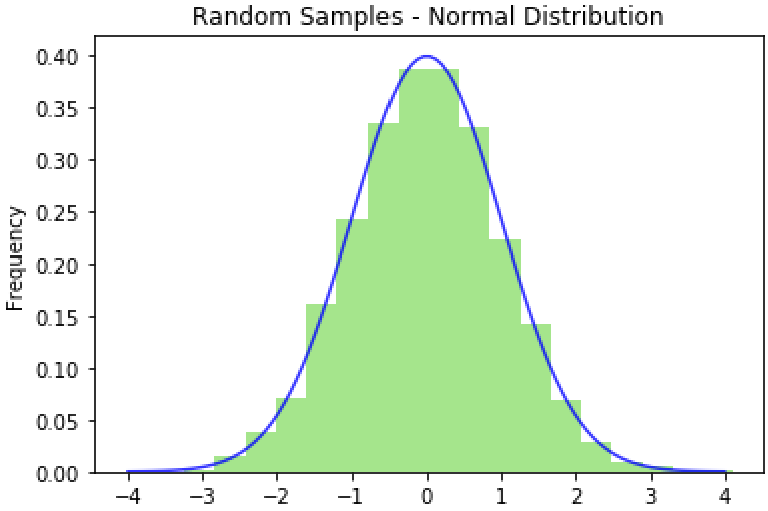
1. Bernoulli random variables take (only) the values 1 and 0. Answer: - A
   a) True
   b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   Answer: - A
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution? Answer: - B
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement. Answer: - C
   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
5. _______ random variables are used to model rates. Answer: - C
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT. Answer: - F
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data? Answer: - B
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_______and have units equal to standard deviations of the original data. Answer: - A
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers? Answer: - C
   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

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

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

Answer:

Normal Distribution is symmetric about the mean. Data near the mean are more frequent in occurrence comparatively data far from the mean. In graphical representation Normal Distribution looks like bell curve as shown in below image.



In a normal distribution the mean is zero, the standard deviation is 1 and it has zero skewness. Normal distributions are symmetrical, but not all symmetrical distributions are normal.

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

Answer:

Missing data reduces the statistical power of the analysis of the data, which can reduce the accuracy of result or prediction. When dealing with missing data, I use two primary methods to process the data (Removing of data or Imputation of data).

If missing data is low, I prefer Imputation of data which develops the reasonable guesses from missing data. If missing data is very high, the results lack natural variation that could results in ineffective model. So, it's better to use removing of data method if missing data is very high.

12. What is A/B testing?

Answer:

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

Answers:

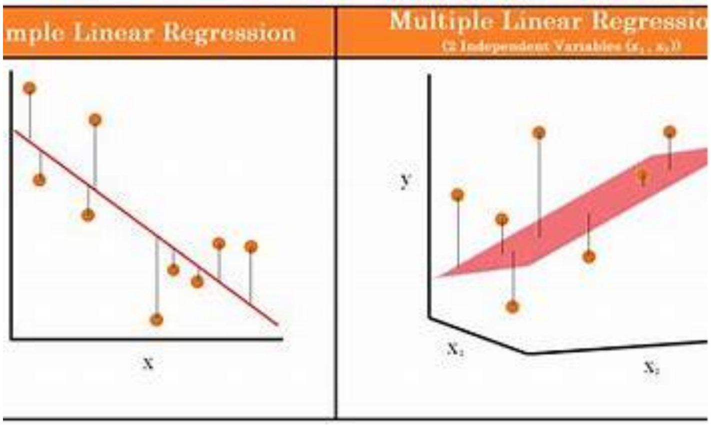
Mean imputation of missing data is not acceptable and it is a bad practice. Mean imputation preserves the mean of the observed data. Which can lead to underestimate of the standard deviation and also destroy the correlation between variables.

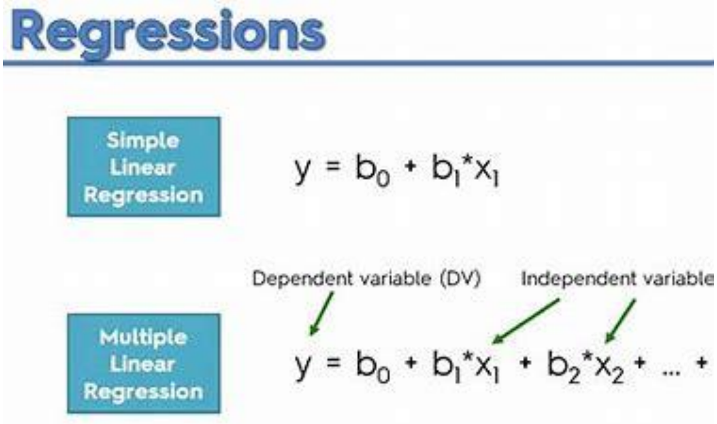
For example, imagine we have a table showing age and fitness score and imagine that an 80 year old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the 80 year old will appear to have a much higher fitness score that they actually should.

14. **What is linear regression in statistics?**

Answers:

Linear regression is a model between two variables by fitting a linear equation to observed data. One variable is considered to be an independent variable, and the other is considered to be a dependent variable. If there is only one independent variable in model it is called simple linear regression, in case of multiple independent variable it is called multiple linear regression.





Regressions

Simple Linear Regression

$y = b_0 + b_1*x_1$

Dependent variable (DV)

Independent variable

Multiple Linear Regression

$y = b_0 + b_1*x_1 + b_2*x_2 + \dots +$

15. **What are the various branches of statistics?**

Answer:

There are two main branches of statistics
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.