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.

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?

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?

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.

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.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

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

a) True

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

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

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

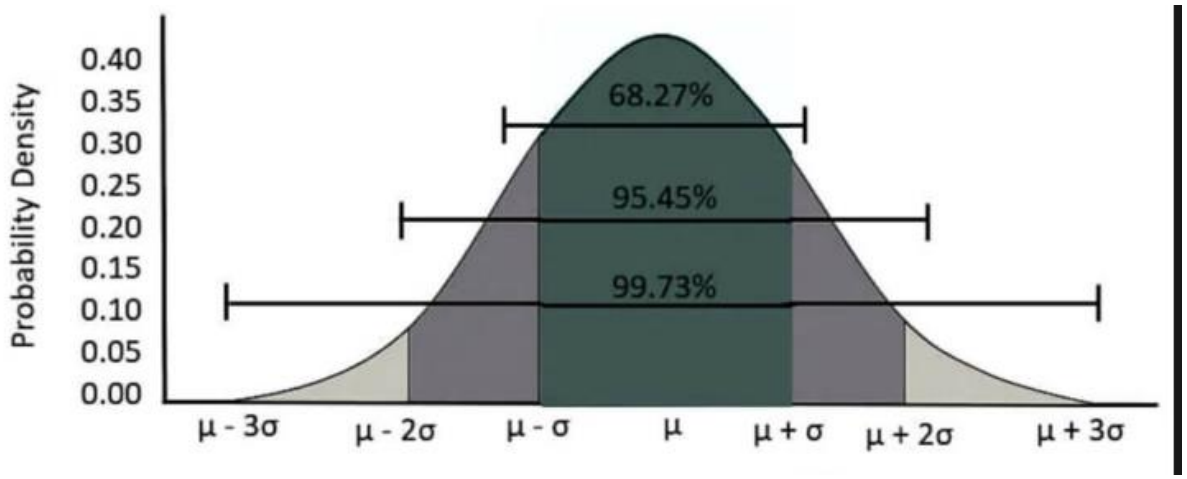
c) Outliers cannot conform to the regression relationship

d) None of the mentioned

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

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

Ans: A normal distribution is perfectly symmetrical around its center. That is, the right side of the center is a mirror image of the left side. Normal Distribution is also called Gaussian distribution and bell shaped curve.



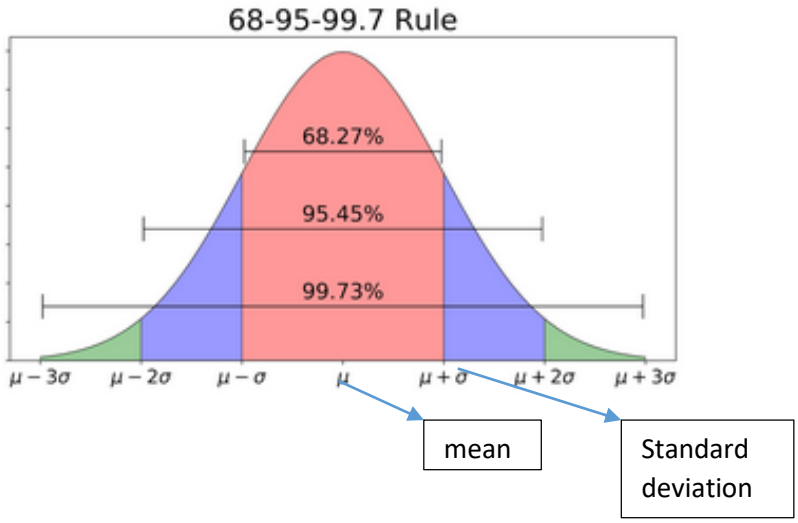
The normal distribution is a core concept in statistics, the backbone of data science.

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.

Empirical Rule for Normal Distribution



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

Ans: Handle missing values in categorical variables:

1. Delete the rows (losing information)
2. Replace with the most FREQUENT Values (Imbalance data)
3. Apply classifier algorithm to predict
4. Apply Unsupervised Machine Learning

Handle missing values in numerical variables:

1. List wise deletion (losing information)
2. Mean/Median/Mode
3. Deck Imputation
   (i) Cold
   (ii) Hot
4. Model based imputation
   (i) KNN(K-nearest neighbour)
   (ii) EM (expectation maximization)
   (iii) ML(maximum likelihood)
   (iv) Regression
5. Prior knowledge

12. What is A/B testing?

Ans: A/B testing (also known as bucket testing or split-run testing) is **a user experience research methodology**. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.

A/B testing Helps Identify Issues: A lot of marketing campaigns fail due to small errors. Best AB testing tools can recognize these errors so that a business can run smoothly. It can help identify a lot of problems such as poor UX design. This is important because a better design can increase conversion by up to 400 percent.

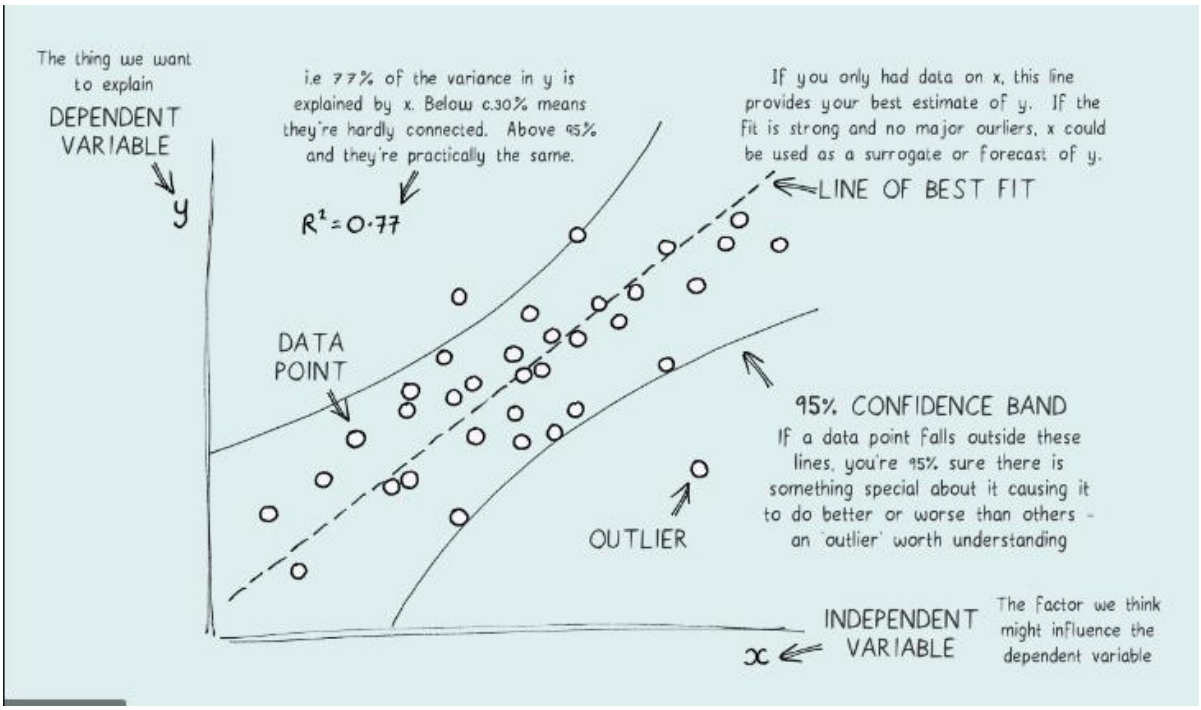
13. Is mean imputation of missing data acceptable practice?

Ans: Outliers data points will have a significant impact on the mean and hence, in such cases, **it is not recommended to use the mean for replacing the missing values**. Using mean values for replacing missing values may not create a great model and hence gets ruled out.

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

14. What is linear regression in statistics?

Ans: In statistics, linear regression is **a linear approach for modelling the relationship between a dependent and independent variables**.



Example of Linear Regression:
The weight of the person is linearly related to their height. So, this shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase.

15. What are the various branches of statistics?

Ans: Statistics simply means **numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data.**
There are three real branches of statistics:

1. **Data collection**
2. **Descriptive statistics**
3. **Inferential statistics**.