Statistics Worksheet:

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

a) True

b) False

Answer : 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

Answer: 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

Answer: 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

Answer: d) All of the above.

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: c) Poisson

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

a) True

b) False

Answer: 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

Answer: 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

Answer: 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

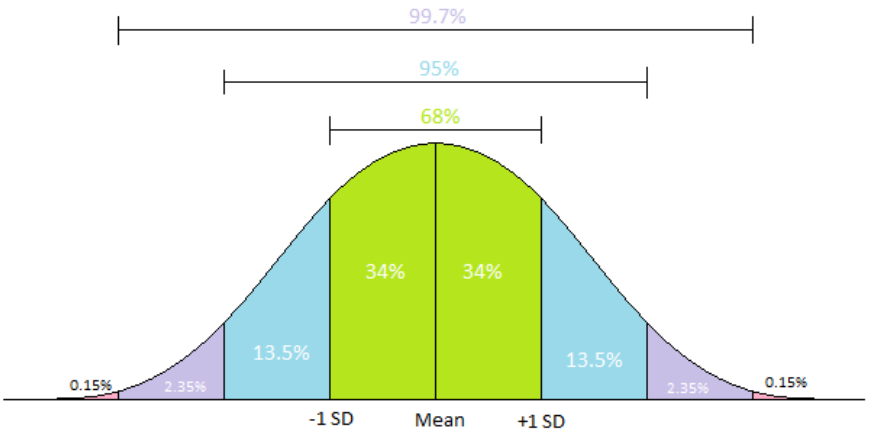
c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Answer: c) Outliers cannot conform to the regression relationship

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

Normal Distribution: A normal distribution is a bell-shaped frequency distribution curve consist of continuous random variable data. This normal distribution is also known as Gaussian Distribution where we have theoretically mean=median=mode. And data are symmetrically distributed around mean, median and mode. The form of normal distribution is determined by its mean and standard deviation where mean=0 and standard deviation = 1



In Normal distribution curve we have mean=0 and symmetrically both side of the mean(above and below the mean) in 1 standard deviation we have total 68% dispersion of data and similarly in 2 standard deviation we have total 95% dispersion of data and within 3 standard deviation we have total 99.7% dispersion of data. Below and above 3 standard deviation of the mean all the data are known as outliers.

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

There are different techniques to handle with missing data :

a. Dataframe['ColumnName'].fillna() method (pandas):
    i. df['Column_Name']=df['Column_Name'].fillna(df['Column_Name'].mean())
    ii. df['Column_Name']=df['Column_Name'].fillna(df['Column_Name'].median())
b. DataFrame.replace() Method (Numpy):
    i. df=df.replace(np.NaN, df['ColumnName'].mean())
    ii. df=df.replace(np.NaN, df['ColumnName'].median())
c. Imputation Method (sklearn.impute) :
    i. from sklearn.preprocessing import Imputer
    imp=Imputer(missing_values=np.NaN / pd.NA , strategy= 'mean' / 'median' / 'most_frequent', / 'constant')
    ii. from sklearn.impute import SimpleImputer
    imp=SimpleImputer(missing_values= np.nan / pd.NA, strategy= 'mean'/ 'median' / 'most_frequent' / 'constant')

Above all the techniques I will recommend SimpleImputer technique because with this one process we can perform in any dataset whether it is pandas.DataFrame or it is numeric dataset.

12. What is A/B testing?

A/B Testing: A/B testing is a type of experiment in which we split our web traffic or user base into two groups, and show two different versions of a web page, app, email, and so on, with the goal of comparing the results to find the more successful version. In the field of data science A/B testing is a form of statistical hypothesis testing or a significance test.

Null hypothesis (H_0): The null hypothesis, or H_0, posits that there is no difference between two variables in other word two variable are identical to each other. In A/B testing, the null hypothesis would assume that changing one variable on a web page (or marketing asset) would have no impact on user behavior.

Alternative hypothesis (H_A): On the other side, an alternative hypothesis suggests the opposite of the null hypothesis that is two variable are not identical to each other. i.e. changing an element will impact user behavior.

- if p-value > 0.05 then Null Hypothesis is accepted and Alternative Hypothesis is rejected.
- If p-value <0.05 then Alternative Hypothesis is accepted and Null Hypothesis is rejected.

Statistical significance: Statistical significance is meant to signify that the results of an A/B test are not due to chance (rejecting the null hypothesis). This is calculated by measuring the p-value, or probability value. So, if the p-value is low, it is saying that it's unlikely the results of the A/B test were random. A rule of thumb tends to be that when the p-value is 5% or lower, the A/B test is statistically significant.

Confidence level: Think of the confidence level as the inverse of the p-value. The confidence level is the indication of how likely it is that the results of your experiment are due to the changed variable (that is, these results are not random or a fluke occurrence). If a test is considered statistically significant when the p-value is at 5%, then the confidence level would be 95%.

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. First, 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 the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line.

15. What are the various branches of statistics?