WORKSHEET

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

Ans : Option 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
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

Ans : Option 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
c) Modeling contingency tables
d) All of the mentioned

Ans : Option b) Modelling 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

Ans : Option d) All of the mentioned.

5. _____ random variables are used to model rates.

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans : Option c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Ans: Option b)False.

7. 1. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans : Option b) Hypothesis.

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

Ans : Option 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

Ans : Option c) Outliers cannot conform to the regression relationship.

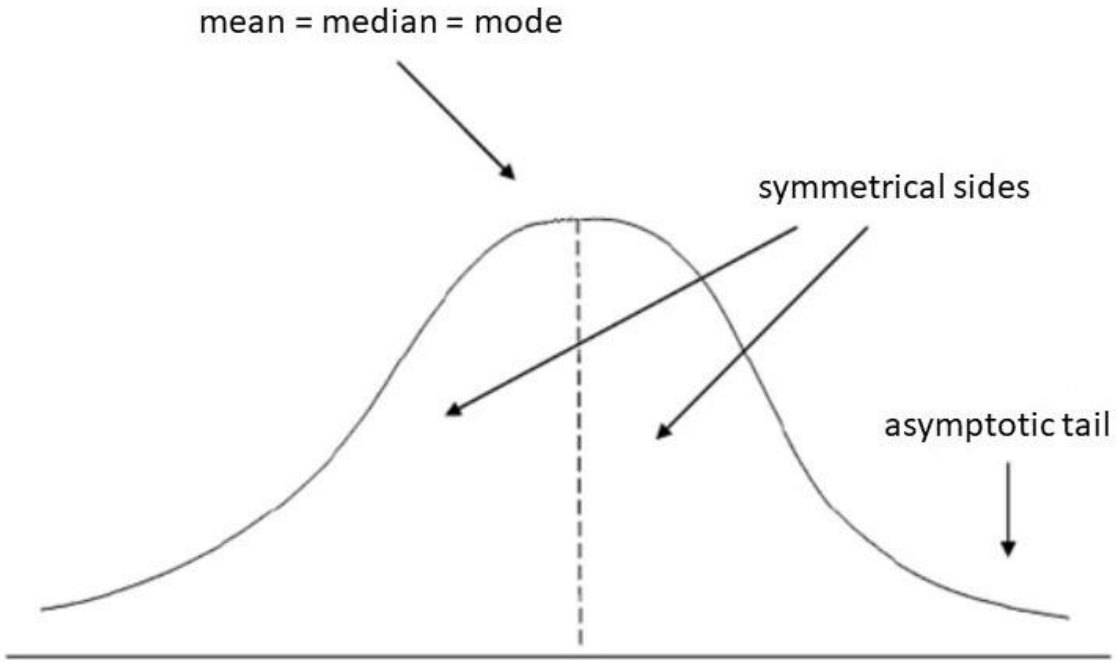
WORKSHEET

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

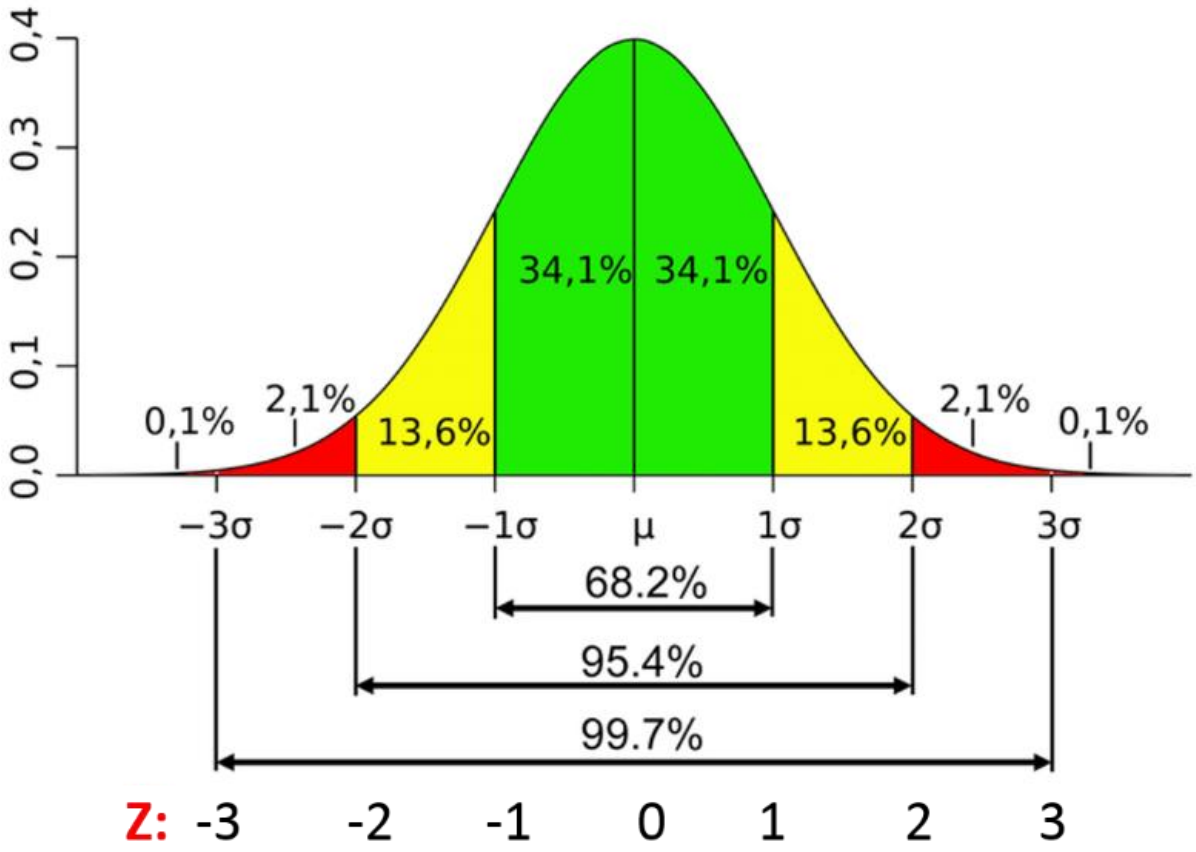
10. What do you understand by the term Normal Distribution?

Ans : The Normal Distribution also called Gaussian distribution where mean, median & mode are line up at the centre of distribution which is mean, because of this exactly half of the results fall to either side of the mean. It is identified by bell shape and sometime referred as bell curve. It is generally show distribution for continuous data and not for categorical data.

Further, when we receive data, it is expected to have all the values in a normal distribution. Which means, that the data should all revolve around some central value or the average/mean of the values should represent the entire data value range? Sometimes, there may be values which are slightly off the normal values expected, either on the lower side or the higher side. Such values which lie outside the normal values are called OUTLIERS. These outlier values can affect your analysis and hence the treatment of these outlier values depends on the company's decision to handle or ignore them.



<u>Characterstics of Normal Distribution</u> :



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

Ans : Missing data Can be Handled in three ways :

1. If there are missing values present in some rows say 5-6 rows we can drop the rows:

df.isnull().sum()

df = pd.dropna(axis=0)Or

` df = pd.dropna(axis=0,inplace=True)

Where Axis 0 : Representing row/Observation in a data set

2. Remove the entire attribute or Column if the NaN value > 40%age:

df = df.drop(['var1'], axis=1) OR

df = df.drop(['var1'], axis=1,inplace=True)

3.Set missing value to some value like zero, mean,median or mode.

a) Replacing NAN with zero for whole dataframe:

newdf=df.replace(np.nan, 0) OR

newdf=df.replace(np.nan, 0, Inplace=True)

b) Replacing NAN with mean if the data is continuous and normally distributed for whole data frame.

newdf=df.replace(np.nan,df.mean()) Or

newdf=df.replace(np.nan,df.mean(), Inplace=true)

c) Replacing NaN with Mean for continuous features if the feature is normally distributed.

df['column']=df['column'].replace(np.NAN,df['column'].mean(), inplace=True)

d) Replacing NaN with Mode for a Categorical feature.

df['feature1'].fillna(df['feature1'].value_counts().idxmax(),inplace=True) OR

df['feature1']= df['feature1'].fillna(df['feature1'].mode(), inplace=True)

e) Replacing NaN with Median if the data is skewed:

1) For whole dataframe : df.fillna(df.median(), inplace=True)

2) For one feature :

df['feature1']= df['feature1'].fillna(df['feature1'].median(), inplace=True)

Imputation Technique:

1.Simple imputer (Univariate) it only takes a single feature into account when performing imputation.: Most Common Technique for Simple imputer are Mean, Median & Mode.

a) Treating Categorical Variables by frequency or mode:

from sklearn.impute import SimpleImputer

Imp = SimpleImputer(strategy="most_frequent")

df['Feature1']=imp.fit_transform(df['Feature1'].values.reshape(-1,1))

b) Treating Conitnuous Variables by mean:

```
Imp = Simple Imputer(missing_values=np.nan, strategy='mean')
df['feature1']=imp.fit_transform(df['feature1'].values.reshape(-1,1))
df['feature2']=imp.fit_transform(df['feature2'].values.reshape(-1,1))
```

2.Iterative imputer(Multivariate Imputation) it treats the column with missing values as a target variable while the remaining columns are used are predictor variables to predict the target variable.

We have to instantiate [simple_imp = IterativeImputer()] which is applied to target variable

3. KNN(Multivariate Imputation) KNN imputer scans a dataset for k nearest rows to the row with missing values. It then proceeds to fill those missing values with the average of those nearest rows.

So Concluding, multivariate imputation results in better model predictions due to its lower mean cross-validation score.

12. What is A/B testing? :

Ans : A/B Testing is a traditional Statistical Analytics Technique where two test run in parallel grounded in a hypothesis test (e.g. t-test, z-score, chi-squared test). In plain English, 2 tests are run in parallel:

Treatment Group (Group A) - This group is exposed to the new web page, popup form, etc.

Control Group (Group B) - This group experiences no change from the current setup.

The goal of the A/B is then to compare the conversion rates of the two groups using statistical inference.

With the rise of digital marketing led by tools including Google Analytics, Google Adwords, and Facebook Ads, a key competitive advantage for businesses is using A/B testing to determine effects of digital marketing efforts. Why? In short, small changes can have big effects.

This is why A/B testing is a huge benefit. A/B Testing enables us to determine whether changes in landing pages, popup forms, article titles, and other digital marketing decisions improve conversion rates and ultimately customer purchasing behavior. A successful A/B Testing strategy can lead to massive gains - more satisfied users, more engagement, and more sales - Win-Win-Win.

13. Is mean imputation of missing data acceptable practice? :

Ans : Well it depends on how the data is distributed, if the data is skewed then there will be large number of data points that act as outliers. Outlier‘s data points will have a significant impact on the mean and hence it is not recommended to use mean for replacing the missing values and will not create a good model and hence it may ruled out.

If the data is symmetrical distributed then one may use mean value for imputing missing values, however it may induces some low standard error which may lead to underestimate of standard errors.

14. What is linear regression in statistics?

Ans : Linear regression is a type of predictive analysis .

Which are used to explain the relationship between one dependent
Variable and one or more independent variables?
Find out which independent variables is highly significant &
Influencing the dependent variables.
Which predictor is doing well in predicting the outcome (target variable).

Linear Regression equation is defined as $Y = \beta0 + \beta1X1 + \beta2X2….+Error$ Where .

-Y is the dependent variable which is influenced by other independent variables.

-β0 is called the Intercept or Constant

-β1, β2…. βn are the Regression Co-efficient (which are calculated –

Co-eff of Regression).

-X1, X2….Xn are independent /predictor variables affecting Y (these are actual values of independent variables)

-Error is the Residual (E) $(Y - \hat{Y})$ which is the difference between the Y- Actual Value & Y(hat) predicted target values.

15. What are the various branches of statistics? :

Ans : There are Two Branch of Statistics :

1.Descriptive Statistics: Descriptive statistics provide statistical insight of a sample of data taken from large a population which further helps to know as how much they vary or deviate from the central values, ways to handle negative values, categorization

of subsets from populations, simple comparisons, representation of these values in graphs, shapes of graphs/plots etc.

Descriptive statistics are broadly classified into:

1. Measure of Center : Mean, Median &Mode

2. Measure of Dispersion: Range, Percentile, Quartile, IQR, Variance, Standard deviation, Skewness, Kurtosis.

The overall objective of descriptive statistics is to give you a detailed description of sample data you have on hand. As we have only limited data or sample data on hand, we are mostly required to estimate the population parameters from the sample data. The parameter calculated from the sample data is not 100% accurate and might result in small errors while estimating the parameters for the population.

2.INFERENTIAL Statistics: That is, if a researcher gathers data from a sample and uses the statistics generated, to reach conclusions about a population from which the sample was taken, then those statistics are INFERENTIAL Statistics.

Population - is the collection of objects/people which form the larger group of the analysis.

Sample - is also the collection of objects/people which are usually a subset of the larger population or sometimes is the only size available for analysis.

Types of Inferential Statistics Commonly Used :

1.One Sample hypothesis Test.

2.Confidence Interval.

3.Chi Square Test.

4.T-Test or Annova.

5.Pearson Correlation

6.Bi-Variate Regression

7.Multi variate Regression