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. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   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
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
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?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

10. Normal Distribution:
Normal Distribution is a probability distribution that is symmetric about the mean, showing the data near the mean and more frequent in occurrence the data far from the mean.

Key Point:
1. The Normal distribution is the proper term for a probability bell curve.
2. In a Normal distribution the mean is zero and standard deviation is 1.
3. Normal distributions are symmetrical, but Not all symmetrical distribution is Normal distribution.
4. In finance, most pricing distribution are not, however perfectly normal.

Understanding Normal Distribution:

Normal distribution is the most common type of distribution assumed in the technical stock market analysis, and other types of statical analysis. The standard Normal Distribution has two parameters: The Mean and the Standard Deviation.

The Normal Distribution model is important in statistic and is key to the Central Limit Theorem(CLT). This theory states that averages calculated from Independent, identically distributed random variables have approximately normal distributions, regardless of the type of distributions.

The Normal Distribution is one type of Symmetrical distribution. Symmetrical distribution occur when where a dividing line produces two mirror images.

Properties of Normal Distribution:
The Normal Distribution has several key features and properties that define it.

1. **Mean(Average) 2. Median(MidPoint) 3. Most(most frequent observation)**
   All those values are represent the peak or highest point of the distribution or the peak. All the values in the distribution then fall symmetrically around the mean. The width of the mean is defined by the standard variation.

The Formula for the Normal Distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Here X is value of variable;
f(x): is represents the probability density function
μ (mu): its mean
σ (sigma) is the standard deviation.

What is the Meant by the Normal Distribution:
The Normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the “Bell Curve”.

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

What is Missing Data:

Missing Data can be dealt with in a variety of ways. Real world data is messy and usually holds a lot of missing values. Missing Data is appear when No Values is available in One or More Variables of an Individual. Due to missing data the statistical power of the analysis can reduce, which can impact the validity of the results.

What are the reasons behind Missing Data?

Missing Data can occur due to many reasons. The Data collected from various sources and while mining the data, there is a chance to loose the data. However most of the time cause for missing data is item nonresponse, which means people are not willing to answer the question in a survey. Some people are don't want to share their contact no, age or salary.

Types of Missing Data:

1. Missing completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing Not at Random (MNAR)

Missing Completely at Random (MCAR):

A variable is missing completely at random (MCAR) if the missing values on a given variable (Y) don't have a relationship with other variables in a given data set or with the variable (Y). In other words, When data is MCAR, there is no relationship between the data missing and any values, and there is no particular reason for the missing values.

Missing at Random (MAR):

MAR occurs when the missingness is not random, but there is a systematic relationship between missing values and other observed data but not the missing data.

Missing Not at Random (MNAR):

The final and most difficult situation of missingness. MNAR occurs when the missingness is not random, and there is a systematic relationship between missing values, observed value, and missing itself.

Detecting Missing Data:

Detecting missing values numerically:

First, detect the percentage of missing values in every column of the dataset will give an idea about the distribution of missing values.

```
Import pandas as pd
Import numpy as np
Import matplotlib.pyplot as plt
Import seaborn as sns
Import warnings # Ignores any warning
Warning.filterwarnings("ignore")

Train = pd.read_csv("Train.csv")
```

Treating Missing Data:

Detection:

The Detection technique deletes the missing values from a dataset. Following are the types of missing Data.

Listwise Detection:
Listwise detection is preferred when there is missing completely at random case. In listwise deletion entire rows are detected, It is also known as complete case analysis as it removes all data that have one or more missing values.

Pairwise Deletion:
Pairwise deletion is used if missingness is missing completely at Random, MCAR
Pairwise deletion is preferred to the reduce the loss that happens in listwise deletion. It is also called an available case analysis as it removes only null observation, not the entire row.

Imputation Techniques:
The imputation techniques replaces missing values with substituted values. The Missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows.

- Imputation with constant value: It replaces the missing values with either zero or any constant value
- Imputation using Statistic: The syntax is the same as imputation with constant value. It can be 1 “Mean” or “Median” or “Most_Frequent”.
  “Mean” will replace missing values using the mean in each column. It is preferred data is numeric and not skewed.
  “Median” will replace missing values using the Median in each column, It is preferred data is numeric and skewed.
  “Most Frequent” will replacing missing values using the most_frequent in each column. It is preferred if data is a string or numeric.

Conclusion Of Missing Values:
There is no single method to handle missing values. Before applying any methods it is necessary to understand the types of missing values, then check the datatype and skewness of the missing column, and then decide which method is best for particular problem.

12. **What is A/B Testing:**
    A/B Testing is also called as Split testing. Refers to randomize experimentation process wherein two or more versions of a variable are shown to different segments of website visitors at the same time to determine which version levels the maximum impact and drives business metrics.

    Essentially, A/B Testing eliminates all the guesswork out of website optimization. And the enables experience optimizer to make data-blocked decisions. In A/B Testing, A refer to ‘Control’ or the original testing variable.

    How A/B testing Works:

    In A/B test, you take a webpage or app screen and modify it to create a second version of the same page. This change can be as simple as a single headline, button or be a complete redesign of the page. Then, half of your traffic is shown the original version of the page and half are shown the modified versions of the page.

    As visitors are served either the control or variation their engagement with each experience is measured and collected in a dashboard and analyzed through a statistical engine. You can determine whether changing had a positive, negative or neutral effect or visitor behavior.

Why you should A/B Test:
A/B testing allows individuals, teams and companies to make careful changes to their user experience while collecting data on the results. This allow them to construct hypotheses and to learn why certain elements of their experiences impact user behavior.

A/B testing Process:

- Collect Data
- Identify Goals
- Generate hypothesis
- Create variations
- Run Experiment
- Analyze Results.

A/B Test have several point for consider. Let's discuss about those and how its work:

1. Solve Visitor Pain Points
2. Get Better ROI from existing Traffic
3. Reduce Bounce Rate
4. Make Low Risk Modification
5. Achieve statistically significant improvements
6. Redesign website to increase future business gains

Solve Visitor Pain Points:
Visitors come to your website to achieve a special goal that they have in Mind. It may be to understand your products or services and buy a particular product, and learn more about a specific topic or simply browse. Whatever the visitors goal may be, they face some more common pain points with achieving their goal. Such as Google Analytics tool can help this problem or website survey to solve your visitor pain points.

Get Better ROI from Existing Traffic:
As most experience optimizers have come to realize the cost of acquiring traffic on your website huge, A/B testing lets you make the most out of your existing traffic and helps you increase conversions without having spend additional money on acquiring new traffic. A/B Testing give Most of the time High ROI and increase significant conversions on your website.

Reduce Bounce Rate:
We measure bounce rate on website as per website performance. It's a importance matrix to A/B testing. If your website bounce rate have high that means your website performance are not good at all. They may be many reasons behind your website high bounce rate, such as confusing navigation, not clear the products or services details too much time to load etc.

While A/B testing have multiple variation you can choose as per your sufficient for your website. They not only help friction and visitors pain points but help improve website visitors overall experience and making more time to spend on your website.

Make Low Risk Modification:
A/B testing lets you target your resources for maximum output and minimal modifications, result is increase ROI, You can perform an A/B test when you plan to remove or update your products descriptions. You'll be understand how do your visitors react those changes, By A/B test you can understand their reaction and ascertain which side the weighting scale may tilt.

Achieve Statistically significant improvements:
Since A/B testing is entirely data driven with no room for guesswork, gut feelings or instincts, you may quickly determine a "winner" or a "looser" based on statistically significant improvements on metrics like time spent on your page, demo request, home page messaging and call to action, click through rate etc.

Redesign website to increase better future business gains:
Redesign can range from a minor CTA text colour tweak to particular webpages to completely revamping the website. The decision to implement one version or the other should always be data driven when A/B test.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is the popular solution of missing data, despite its drawback, Its beat simple but so dangerous,

Here I can discuss problem about mean Imputation:

Problem: Mean imputation does not preserve the relationships among variables:
If the data are missing completely at random, the estimate of the mean remains unbiased. And imputing the mean, its able to keep sample size up to the full sample size, that's very good.

If all you are doing is estimating means and if the data are missing completely at random, mean imputation will not bias your parameter estimate.
Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

Linear Regression is a basic and commonly used type of predictive analysis.
These regression estimates are used to explain the relationship between one dependent variable and one or more independent variable.

Formula: Y = C+B*X
Y= Estimated dependent variable score
C= Constant
B= Regression Coefficient
X= Score on the independent variable

Three major uses for regression analysis are (1) determining the strength of predictors
(2) forecasting an effect (3)trend forecasting

Types of Linear Regression:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Ordinal Regression
5. Multinominal Regression
6. Discriminant analysis

15. What are the various branches of statistics?

The major two statistics are descriptive and inferential statistic.

Statistics is the idea we can learn about the properties of large sets of objective or events. Because many cases gathering comprehensive data about an entire population is too costly, difficult or flat out impossible, statistics start with a sample that can conveniently or affordably be observed

Descriptive Statistic: Descriptive statistics mostly focus on the central tendency, variability and distribution of sample data. Central tendency means estimate of the characteristics, a typical element of a sample population, and includes descriptive statistic such as mean, Median, and Mode. Variability refers to a set of statistic that show how much difference there is among the elements of a sample or population along with characteristic measured, and include metric such as Range, Variance, and standard deviation.

The distribution refers to the overall "Shape" of the data, which can be depicted on a chart such as histogram, or dot plot and include properties such as probability distribution function, skewness, and kurtosis.

Inferential Statistic:
Inferential statistic are tools that statisticians draw a conclusions about characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution statisticians can calculate the probability that statistic, which measure the central tendency, variability, distribution and relationship between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which sample is drawn.

Understanding Statistical Data:
The root of statistics is driven by variables. A Variable is a set of data set can be counted that makes a characteristic or attribute of an item.

There are two main types of variable: 1. Qualitive Variables 2. Quantitative Variables

Statistic Level of Measurement:

1. Nominal Level Measurement
2. Ordinal Level Measurement
3. Interval Level Measurement
4. Ratio Level Measurement

Nominal Level Measurement: There is no numerical or quantitative value, and qualities are not ranked. Instead, nominal level measurements are simply labels or categories assigned to other variables. It's easiest to think of nominal level measurement as non-numerical facts about a variable.

Ordinal Level Measurement: Outcomes can be arranged in an order, however all data values have the same value or weight.

Interval Level Measurement: Outcomes can be arranges in an order, however differences between data values may now have meaning. Two different data points are often used to compare the passing time of changing conditions within a data set.

Ratio Level Measurement: Outcomes can be arranged in order, and differences between data values now have meaning. However , there is now a starting point or "Zero Values" that can be used to further provide value to statistical value. The Ratio between data values now has meaning, including its distance away from zero.

Statistical Sampling Technique:

- Simple random sampling
- Systematic sampling
- Satisfied sampling
- Cluster sampling