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
   Answer: 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
   Answer: 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
   Answer: b) 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 mentioned

5. _______random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: c) Poisson

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

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?

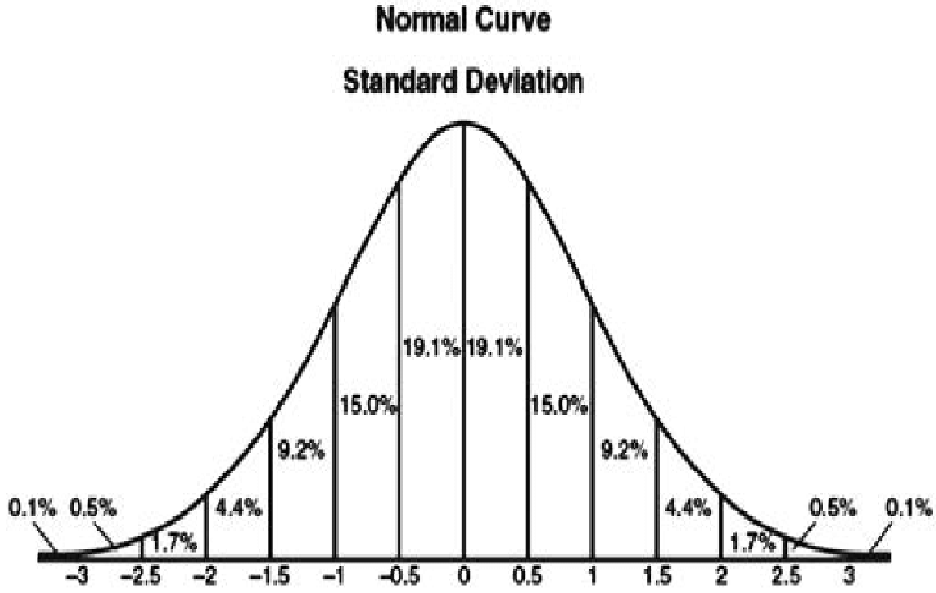
10) Normal Distribution:

- Normal Distribution is continuous probability distribution. It is described by the mean and the standard deviation. It is also called as Gaussian Distribution. It is a bell shape where mean is at the center position.

- The Normal Distribution is defined by the probability density function for a continuous random variable in a system.

- The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical.

- In a normal distribution the mean mode and median are all the same.



- In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

- Normal distribution, accuracy and precision are referred to as the *mean* and the *standard deviation*, respectively.

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu &= \text{Mean} \\ \sigma &= \text{Standard Deviation} \\ \pi &\approx 3.14159\cdots \\ e &\approx 2.71828\cdots \end{aligned}$$

- In a normal distribution, the **mean, mean and mode are equal**. (i.e., Mean = Median= Mode).
- The total area under the curve should be equal to 1.
- The normally distributed curve should be symmetric at the centre,
- There should be exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.
- The normal distribution should be defined by the mean and standard deviation.
- The normal distribution curve must have only one peak. (i.e., Unimodal)
- The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.
- When both sides of the distribution are **not distributed equally** then this is known as **Skewed Data**. It is not a symmetrical distribution.
- The normal distribution is the distribution of the probability without any skewness.
    - Positive Skewness.
    - Negative Skewness.

- The **standard normal distribution** is one of the forms of the normal distribution. It occurs when a normal random variable has a mean equal to zero and a standard deviation equal to one.
- The random variable of a standard normal distribution is known as the **standard score or a z-score**. It is possible to transform every normal random variable X into a z score using the following formula:
    - **z = (X – μ) / σ**

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

- Missing data can be dealt by calculating the mean of the observed values for that variable for all non-missing data.

- It has the advantage of maintaining the same mean and sample size.

- **Imputation** is the process of replacing missing values with substituted data. It is done as a preprocessing step.

- The missing values will be represented as NaN — Not a Number.

NORMAL IMPUTATION:

- If we have missing value in feature. We can replace the missing values with the below methods depending on the data type of feature.

  - Mean
  - Median
  - Mode

- If the data is numerical, we can use mean and median values to replace else if the data is categorical, we can use mode which is a frequently occurring value.

IMPUTATION BASED ON CLASS LABEL:

- Instead of taking the mean, median, or mode of all the values in the feature, we take based on class.

- The average of all the values in the feature that belongs to class 0 or 1 and replace the missing values. Same with median and mode.

MODEL-BASED IMPUTATION:

- We take feature as the class and all the remaining columns as features. Then we train our data with any model and predict the missing values.

CREATING MISSING VALUE FEATURE:

- Performing imputation on the features, we can create new corresponding features which will have binary values that say whether the data is missing in the features or not with 0 as not missing and 1 as missing.

12)What is A/B testing?

- An A/B test is an example of **statistical hypothesis testing**, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

- The **null hypothesis** is a baseline assumption that there is no relationship between two data sets. When a statistical hypothesis test is run, the results either disprove the null hypothesis or they fail to disprove the null hypothesis.

- **A/B testing** is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

- With the data we collected from the activity of users of our website, we can compare the efficacy of the two designs A and B. Simply comparing mean values wouldn't be very meaningful, as we would fail to assess the **statistical significance** of our observations. It is indeed fundamental to determine how likely it is that the observed discrepancy between the two samples originates from chance.

- A **two-tailed test** is preferable in our case, since we have no reason to know a priori whether the discrepancy between the results of A and B will be in favor of A or B. This means that we consider the alternative hypothesis **Ha** the hypothesis that A and B have different efficacy.

- Different kinds of metrics can be used to measure a website efficacy. With **discrete metrics**, also called **binomial metrics**, only the two values **0** and **1** are possible.

- With **continuous metrics**, also called **non-binomial metrics**, the metric may take continuous values that are not limited to a set two discrete states**.**

- A/B testing is perhaps one of the most used testing methods of today. After all, it is quite easy to learn, and it doesn't require a lot of technical know-how to implement.

- A/B testing is utilized by businesses of all sizes across various industries.

- Performing A/B tests on them can have a positive effect on overall performance.

  - Hypothesis testing (A/B testing) is a decision-making method. You can make the right decision or you can make a mistake.

In hypothesis testing there are three possible outcomes of the test:

- No error

- Type I error

- Type II error

- **Type I error** occurs when you incorrectly reject the null hypothesis and conclude that there is actually a difference between the original page and the variation
- **Type II error** occurs when you fail to reject the null hypothesis at the right moment, obtaining this time false negative test results. Type II error occurs when we conclude test with the assumption that none of the variations beat the original page while in reality one of them actually did.
- When we conduct an A/B test (or multivariate), we distribute visitors randomly amongst different variations.

Confidence Level

In different A/B testing software packages, you may see a column called:

- Confidence
- Statistical significance
- Significance

It usually shows you some percentage between 0 and 100% and determines how statistically significant the results are.

- Baseline conversion rate (thus, the conversion rate you have now)
- Desired increase (how much do you think the new design will beat the original)
- Confidence level: 90%, 95% or 99%
- The smaller the margin of error the more accurate result you get.
- Randomness is a part of your test and there are a number of statistical values that effect it.
- A/B testing is a decision-making method, but cannot give you a 100% accurate prediction

13) Is mean imputation of missing data acceptable practice?

- **Imputing the mean preserves the mean of the observed data**. So, if the data are missing completely at random, the estimate of the mean remains unbiased. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

- Mean imputation is generally bad practice because it doesn't take into account feature correlation.

- Mean imputation (or mean substitution) replaces missing values of a certain variable by the mean of non-missing cases of that variable.

- Missing values in more than one feature column, all missing values are first temporarily imputed with a basic imputation method, e.g. the mean value. Then the values for one column are set back to missing. The model is then trained and applied to fill in the missing values.

Advantage of Mean Imputation:

- Missing values in your data **do not reduce your sample size**.
- Mean imputation is very **simple to understand and to apply**
- The **sample mean of your variable is not biased**. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric.

Drawbacks of Mean Imputation:

- Mean substitution leads to bias in multivariate estimates such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
- **Standard errors and variance** of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable.
- Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.
- Mean imputation replaces missing values with the mean value of that feature/variable. Mean imputation is one of the most 'naive' imputation methods because unlike more complex methods like k-nearest neighbors' imputation, it does not use the information we have about an observation to estimate a value for it.
- The variable with missing data is used as the dependent variable. Cases with complete data for the predictor variables are used to generate the regression equation; the equation is then used to predict missing values for incomplete cases.

14)What is linear regression in statistics?

- Linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.
- The measure of the extent of the relationship between two variables is shown by the **correlation coefficient**. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables.
- A linear regression line equation is

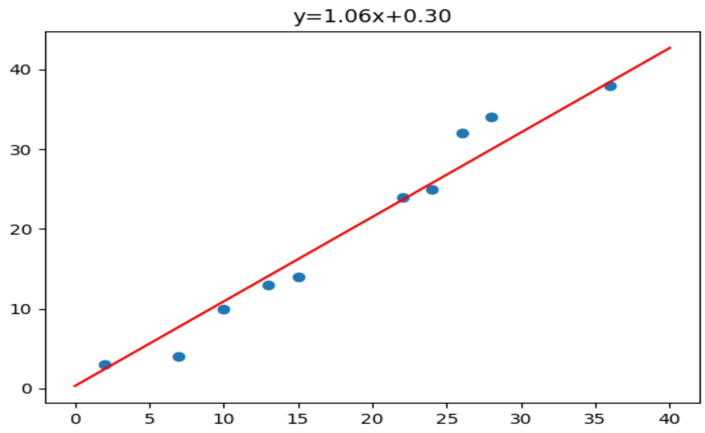
$$\mathbf{Y = a + bX}$$

- The most popular method to fit a regression line in the XY plot is the method of least-squares. This process determines the best-fitting line for the noted data by reducing the sum of the squares of the vertical deviations from each data point to the line.

Properties of Linear Regression:

For the regression line where the regression parameters b_0 and b_1 are defined, the properties are given as:

- The line reduces the sum of squared differences between observed values and predicted values.
- The regression line passes through the mean of X and Y variable values
- The regression constant (b_0) is equal to y-intercept the linear regression
- The regression coefficient (b_1) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X).



- **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables.
- It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

15)What are the various branches of statistics?

There are two Types of Statistics

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics:

- In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.
- In Descriptive statistics we are able to describe data such as Average mark of student, height, weight.
- Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position
- The frequency measurement displays the number of times a particular data occurs.
- Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data.
- Central tendencies are the mean, median and mode of the data.
- The measure of position describes the percentile and quartile ranks.

Inferential Statistics:

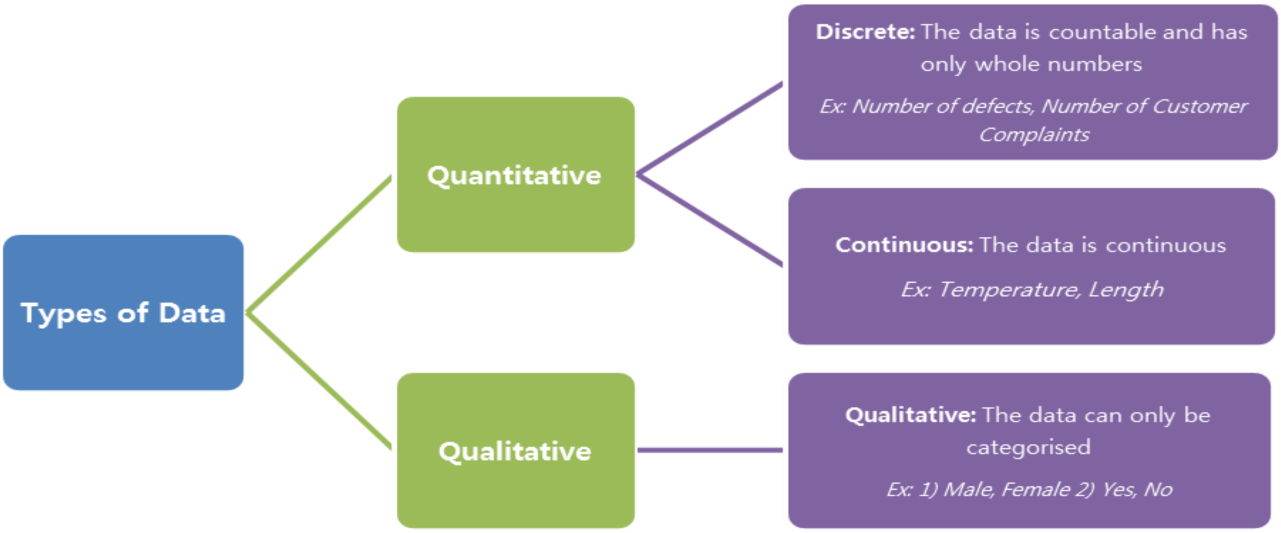
- In Inferential statistics the data is too large and it is difficult to explain.
- This type of statistics is used to interpret the meaning of Descriptive statistics.
- That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.
- Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population.
- It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Types of Inferential Statistics:

1. Regression Analysis
2. Analysis of Variance (ANOVA)
3. Analysis of Covariance (ANCOVA)
4. Statistical significance (t-test)
5. Correlation Analysis

Importance of Statistics:

- Statistics executes the work simply and gives a transparent picture of the work we do regularly.
- The statistical methods help us to examine different areas such as medicine, business, economic, social science and others.
- Statistics equips us with different kinds of organised data with the help of graphs, tables, diagrams and charts.
- Statistics helps to understand the variability of the data pattern in a quantitative way
- Statistics makes us understand the bulk of data in a simple way
- Statistics is the way to collecting accurate quantitative data.



Types of Data

Quantitative

Qualitative

Discrete: The data is countable and has only whole numbers

Ex: Number of defects, Number of Customer Complaints

Continuous: The data is continuous

Ex: Temperature, Length

Qualitative: The data can only be categorised

Ex: 1) Male, Female 2) Yes, No