STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

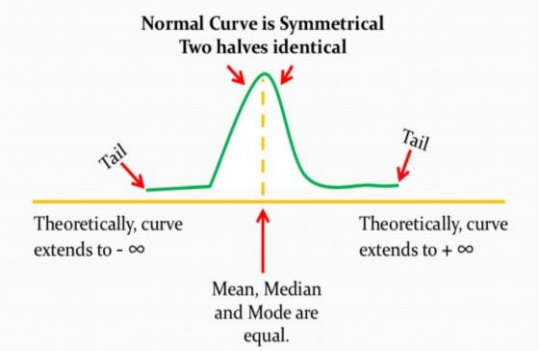
1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central 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?

What is Normal Distribution?

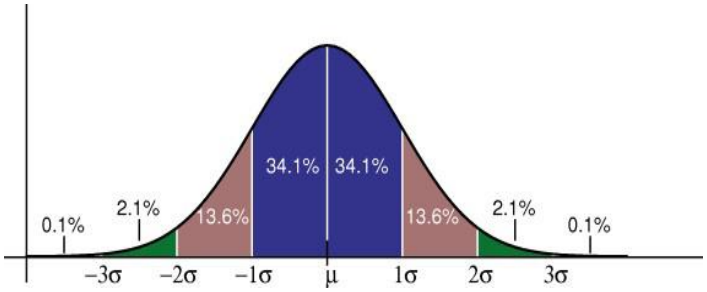
Normal distribution, also known as the **Gaussian distribution**, after the German mathematician Carl Gauss who first described it, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



The normal distribution is often called the **bell curve** because the graph of its probability density looks like a bell.
For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve
A normal distribution is determined by two parameters the mean and the variance.
A normal distribution with **a mean of 0 and a standard deviation of 1** is called a **standard normal distribution.**

The **empirical rule** tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- **68%** of the data falls within **one standard deviation** of the mean
- **95%** of the data falls within **two standard deviations** of the mean.
- **99.7%** of the data falls within **three standard deviations** of the mean.



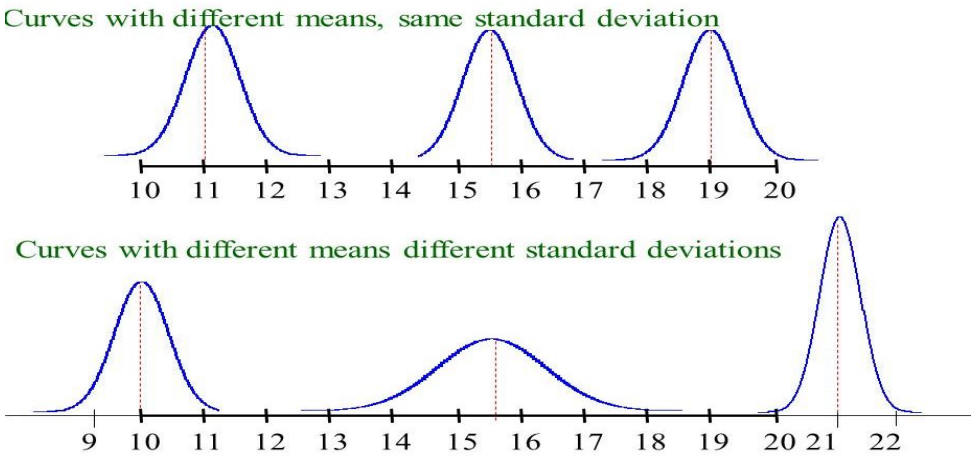
Properties of a normal distribution:

1. The mean, mode and median are all equal.
2. The curve is symmetric at the center (i.e. around the mean, μ).
3. Exactly half of the values are to the left of center and exactly half the values are to the right.
4. The total area under the curve is 1.
5. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
6. Normal distributions are symmetric, unimodal, and the mean, median, and mode are all equal.
7. The normal distribution is asymptotic –i.e. the curve gets closer and closer to the X-axis but never

actually touches it.

8. Theoretically the curve extends to infinity.

   The points at which the curvature changes are called inflection points.



Note: Normal distributions are symmetrical, but not all symmetrical distributions are normal

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

Missing values are common occurrences in data. Unfortunately, most predictive modeling techniques cannot handle any missing values. Therefore, this problem must be addressed prior to modeling. The handling of missing data is very important during the preprocessing of the dataset .The cause of missing values can be data corruption or failure to record data.

Ways to handle missing values in the dataset:

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods like simple imputer, KNN imputer and iterative imputer
5. Multiple Imputation
6. Using algorithms that support missing values.

Delete Rows with Missing Values:

The simplest approach for dealing with missing values is to remove entire predictor(s) and/or sample(s) that contain missing values.

If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped

A model trained with the removal of all missing values creates a robust model but may cause loss of a lot of information.

Example:

```
#dropna function in Pandas removes all the rows with missing values
data.dropna(inplace=True)


#Putting axis=1 removes the columns with missing values
data.dropna(inplace=True, axis=1)
```

This method works poorly if the percentage of missing values is excessive in comparison to the complete dataset.

Impute missing values with Mean/Median:

Imputation helps substitute the missing data by some statistical methods. Imputation is useful in the sense that it preserves all cases by replacing missing data with an estimated value based on other available information

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the two approximations (mean, median) is a statistical approach to handle the missing values.

Example:

```
1  data["Age"] = data["Age"].replace(np.NaN, data["Age"].mean())
2  data["Age"] = data["Age"].replace(np.NaN, data["Age"].median())
```

```
[10] data["Age"][:20]

0     22.0
1     38.0
2     26.0
3     35.0
4     35.0
5      NaN
6     54.0
7      2.0
8     27.0
9     14.0
10     4.0
11    58.0
12    20.0
13    39.0
14    14.0
15    55.0
16     2.0
17     NaN
18    31.0
19     NaN
Name: Age, dtype: float64
```

```
data["Age"] = data["Age"].replace(np.NaN, data["Age"].mean())
print(data["Age"][:20])

0     22.000000
1     38.000000
2     26.000000
3     35.000000
4     35.000000
5     29.699118
6     54.000000
7      2.000000
8     27.000000
9     14.000000
10     4.000000
11    58.000000
12    20.000000
13    39.000000
14    14.000000
15    55.000000
16     2.000000
17    29.699118
18    31.000000
19    29.699118
Name: Age, dtype: float64
```

This method prevents data loss which results in deletion of rows or columns and works well with a small dataset and is easy to implement.
Mean/median imputation works only with numerical continuous variables and can cause data leakage.

Imputation method for categorical columns:

When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category. This method Negates the loss of data by adding a unique category but addition of new features may result in poor performance.

Example:

```
# categorical columns
>>> df.fillna(df.select_dtypes(include='object').mode().iloc[0], inplace=True)
```

Simple Imputer:

The scikit-learn library provides the SimpleImputer pre-processing class that can be used to replace missing values.
It is a flexible class that allows you to specify the value to replace (it can be something other than NaN) and the technique used to replace it (such as mean, median, or mode). The SimpleImputer class operates directly on the NumPy array instead of the DataFrame.

Example:

```
>>> import numpy as np
>>> from sklearn.impute import SimpleImputer
>>> imp_mean = SimpleImputer(missing_values=np.nan, strategy='mean')
>>> imp_mean.fit([[7, 2, 3], [4, np.nan, 6], [10, 5, 9]])
SimpleImputer()
>>> X = [[np.nan, 2, 3], [4, np.nan, 6], [10, np.nan, 9]]
>>> print(imp_mean.transform(X))
[[ 7.   2.   3. ]
 [ 4.   3.5  6. ]
 [10.   3.5  9. ]]
```

Multiple Imputation:

Multiple imputations is an iterative method in which multiple values are estimated for the missing data points using the distribution of the observed data. The advantage of this method is that it reflects the uncertainty around the true value and returns unbiased estimates.
MI involves the following three basic steps:
Imputation: The missing data are filled in with estimated values and a complete data set is created. This process of imputation is repeated m times and m datasets are created.
Analysis: Each of the m complete data sets is then analyzed using a statistical method of interest (e.g. Linear regression).
Pooling: The parameter estimates (e.g. coefficients and standard errors) obtained from each analyzed data set is then averaged to get a single point estimate.

Python's Scikit-learn has methods — impute. SimpleImputer for univariate (single variable) imputations and impute.IterativeImputer for multivariate imputations.

Several versions of the same data set are created, which are then combined to make the best values.
One of the most used method for imputation is known as MICE (Multivariate Imputation by Chained Equations). The MICE package in R supports the multiple imputation functionality. Python does not directly support multiple imputations but Iterative Imputer can be used for multiple imputations by applying it repeatedly to the same dataset with different random seeds when sample_posterior=True.

Example:

```
from fancyimpute import IterativeImputer as MICE
data_fit = pd.DataFrame(MICE().fit_transform(data))
```

This is the most preferred method for imputation since it is easy to use and produces no biases as long as the imputation model is correct.

KNN Imputer:

KNN Imputer helps to impute missing values present in the observations by finding the nearest neighbors with the Euclidean distance matrix. The idea in is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.

Example:

```
[23] from sklearn.impute import KNNImputer

     X = [[3, np.nan, 5], [1, 0, 0], [3, 3, 3]]

     imputer = KNNImputer(n_neighbors=1)
     imputer.fit_transform(X)
```

```
array([[3., 3., 5.],
       [1., 0., 0.],
       [3., 3., 3.]])
```

```
[24] imputer = KNNImputer(n_neighbors=2)
     imputer.fit_transform(X)
```

```
array([[3. , 1.5, 5. ],
       [1. , 0. , 0. ],
       [3. , 3. , 3. ]])
```

KNN imputer for categorical values:

```
# map the Gender values to 0 and 1
df['Gendermap'] = df.Gender.map({'female':1,'male':0})

# print data frame
df[['Age', 'Gendermap']]
```

	Age	Gendermap
0	10	0.0
1	15	1.0
2	14	NaN
3	12	1.0
4	9	0.0

```
from sklearn.impute import KNNImputer

# knn based imputation for categorical variables
imputer = KNNImputer(n_neighbors=2)
df_filled = imputer.fit_transform(df[['Age', 'Gendermap']])

# print the completed dataframe
df_filled
```

```
array([[10.,  0.],
       [15.,  1.],
       [14.,  1.],
       [12.,  1.],
       [ 9.,  0.]])
```

Algorithms that Support Missing Values:

Not all algorithms fail when there is missing data. There are algorithms that can be made robust to missing data, such as k-Nearest Neighbors that can ignore a column from a distance measure when a value is missing. Naive Bayes can also support missing values when making a prediction.
One of the really nice things about Naive Bayes is that missing values are no problem at all.

There are also algorithms that can use the missing value as a unique and different value when building the predictive model, such as classification and regression trees.

Another algorithm that can be used here is Random Forest that works well on non-linear and the categorical data. It adapts to the data structure taking into consideration the high variance or the bias, producing better results on large datasets.

Note: Sadly, the scikit-learn implementations of naive Bayes, decision trees and k-Nearest Neighbors in python are not robust to missing values. Although it is being considered.
Nevertheless, this remains as an option if you consider using another algorithm implementation (such as xgboost) or developing your own implementation

What imputation techniques do you recommend?

We cannot always stick to any particular approach until and unless you try out things and have some idea about the model performance.

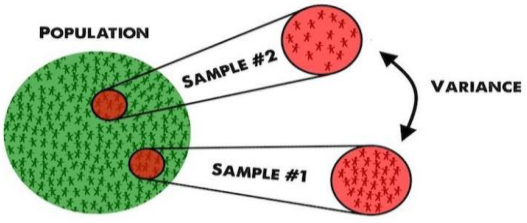
I would keep the following in mind before going for any one:

1. If the missing data is more than 60% of the observations and only if that variable is insignificant I would go with dropping the column (observation). But in most cases since we lose a lot of potential information dropping data is not a good approach.
2. Though the most preferred solution depends on the computational resources, as well as tolerance to errors in approximating missing values and several other factors .The advanced imputation methods address the problem of missing data by exploiting the relationships between variables and they impute multiple values rather than a single value hence I prefer using the iterative imputers, KNN imputer or simple imputer instead of mean/mean/mode imputation.
3. I do not prefer imputation by Mean/Mode/Median because this method reduces the variance of the imputed variables and also reduces the correlation between the imputed variables and other variables .The imputed values are just estimates and will not be related to other values inherently.
4. As handling the missing data is crucial for a data science project. One should keep in mind that the data distribution should not be changed while handling missing data. Any missing data treatment method should satisfy the following rules:
    - Estimation without bias — any missing data treatment method should not change the data distribution.
    - The relationship among the attributes should be retained.

Hence I would prefer the imputers like simple imputer, multiple imputers and KNN imputer to fill the missing values.

12. What is A/B testing?

A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.



In statistics, a **population** is the pool of individuals from which a statistical sample is drawn for a study. Thus, any selection of individuals grouped together by a common feature can be said to be a population.

A sample refers to a smaller, manageable version of a larger group. It is a subset containing the characteristics of a larger population. Samples are used in statistical testing when population sizes are too large for the test to include all possible members or observations.

An AB 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 A/B testing process can be simplified as follows:

1. You start the A/B testing process by making a claim (hypothesis).
2. You launch your test to gather statistical evidence to accept or reject a claim (hypothesis).
3. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

In statistics your hypothesis breaks down into:

1. Null hypothesis
2. Alternative hypothesis

1. Null hypothesis or H0:

A statement in which no difference or effect is expected. If the null hypothesis is not rejected, no changes will be made. The null hypothesis is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant groups.

2. Alternative Hypothesis or H0:

A statement that some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions. It is the opposite of the null hypothesis and is basically a hypothesis that the researcher believes to be true.

Types of errors in hypothesis testing:

1. Type I error
2. Type II error

Type I error: Type 1 error occurs when you incorrectly reject the null hypothesis and conclude that there is actually a difference between the original and the variation when there really isn't. In other words, you obtain **false positive** test results. Like the name indicates, a false positive is when you think one of your test challengers is a winner while in reality it is not.

Type II error: this type of 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 while in reality one of them actually did.

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Type I and type II errors cannot happen at the same time:

1. Type I error happens only when the null hypothesis is true
2. Type II error happens only when hypothesis is false

To summarize:

- Type I error occurs when we incorrectly reject the null hypothesis.
- Type II error occurs when the null hypothesis is false, but we incorrectly fail to reject it.

Example of A/B testing:

Let's say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format.

Our objective here is to check which newsletter brings higher traffic on the website i.e. the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

In our example, the hypothesis can be "By making changes in the language of the newsletter, we can get more traffic on the website".

Null hypothesis H0 is "there is no difference in the conversion rate in customers receiving newsletter A and B".

Alternate hypothesis Ha is "the conversion rate of newsletter B is higher than those who receive newsletter A".

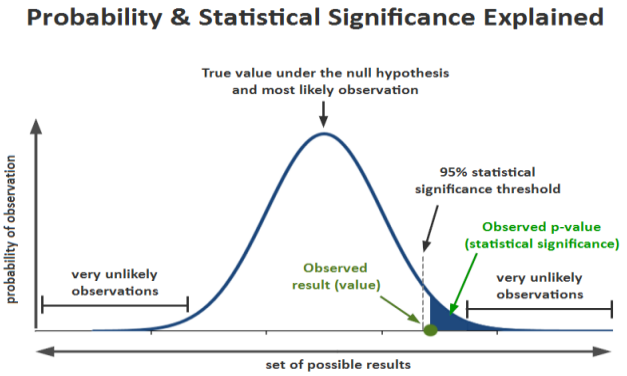
How to decide on will hypothesis is to be accepted:

Let us understand few terms:

Significance level (alpha): The significance level, also denoted as alpha or α, is the threshold probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05. This is also known as the type I error rate.

P-Value: the p-value is the smallest level of significance at which a null hypothesis can be rejected.

Confidence interval: As the name suggests a level of confidence: how confident are we in taking out decisions. Generally, we take a 95% confidence interval



If p value < 0.05: Reject Null Hypothesis and accept alternate hypothesis

If p value>0.05: Accept the Null Hypothesis and reject Ha

Thus whenever we want to make claims about the distribution of data or whether one set of results are different from another set of results in applied machine learning, we must rely on statistical hypothesis tests.

13. Is mean imputation of missing data acceptable practice?

In Mean Imputation, the mean of all values within the same attribute is calculated and then imputed in the missing data cells. The method works only if the attribute examined is not nominal

This method reduces the variance of the imputed variables and also reduces the correlation between the imputed variables and other variables .The imputed values are just estimates and will not be related to other values inherently. Mean imputation does not preserve the relationships among variables.

Also the mean is not robust to outliers and is affected by magnitude of extreme values.

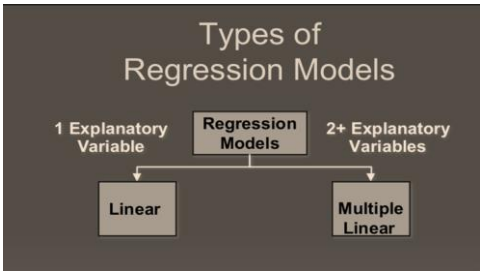
For an example:

List 1 =[1,1,1,2,2,3,3,3,200] the mean will be 24 whereas almost all the other numbers are very small as compared to the mean value this will decrease the dataset's robustness.
Hence we can say the mean values are sensitive to any change in values unlike the mode/median.

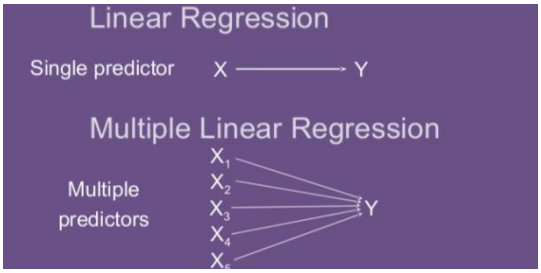
Therefore though the Mean imputation is simple it does not make it a good solution

14. What is linear regression in statistics?

Linear regression is the simplest and extensively used statistical technique for predictive modeling analysis
It is a way to explain the relationship between one dependent variable (target variable) and one or more independent variables. (predictors/explanatory variables) using a straight line.



Based on the number of independent variables the linear regression is classified into simple linear and multiple linear regression.



A simple linear regression uses one independent variable to explain or predict the outcome of dependent variable

Equation:

Y= a+bX+e

Where,

Y=dependent variable (variable that you are trying to predict)
X=Independent variable
a=the intercept
b= slope
e=the residual error

Multiple Linear Regression: it uses two or more independent variables to predict the output.

Equation:

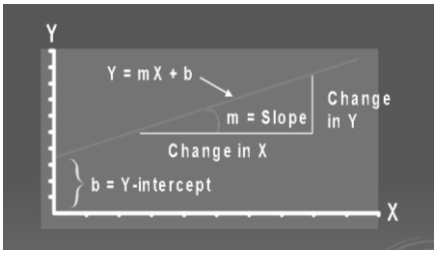
Y= a+b1X1+b2X2+b3X3+b4X4+…… +bnXn+e

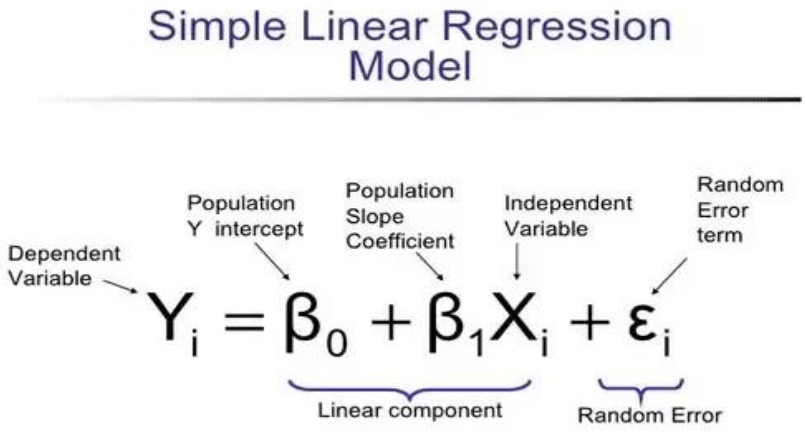
Where,

Y=dependent variable (variable that you are trying to predict)
X1 to Xn =Independent variables
a=the intercept
b= slope
e=the residual error

Examples of linear regression:

- Income and educational qualifications
- Home sales and interest rates
- Weight and height

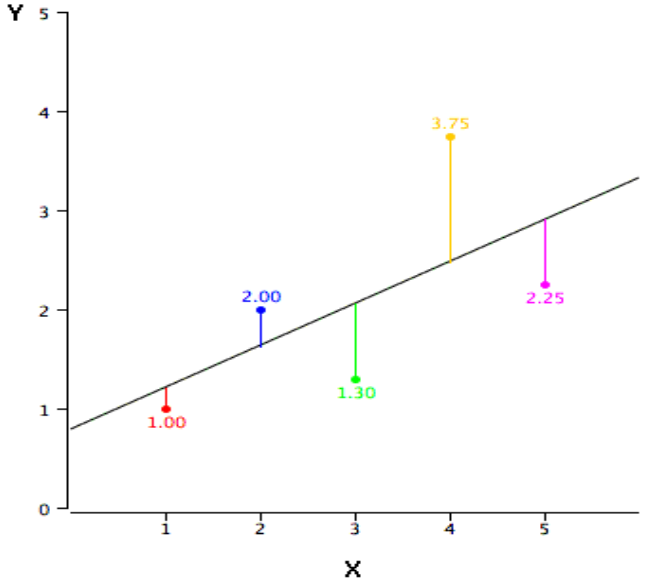




The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line.

Best Fit Line:

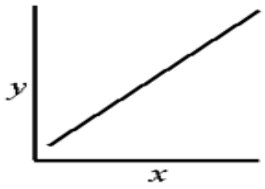
Linear regression consists of finding the best-fitting straight line through the points. **The best-fitting line is called a regression line.**
The black diagonal line in Figure below is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.



The correlation coefficient (r) r is a measure of the linear (straight-line) association between two variables and has the following characteristics**.**

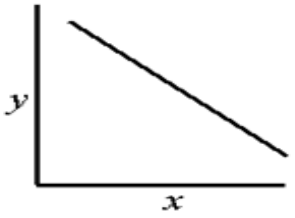
1. The range of r is between -1 and 1, inclusive.

2. If r = 1, the observations fall on a straight line with positive slope.



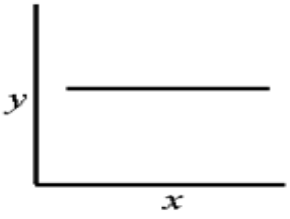
Positive slope

3. If r = -1, the observations fall on a straight line with negative slope.



Negative slope

4. If r = 0, there is no linear relationship between the two variables.



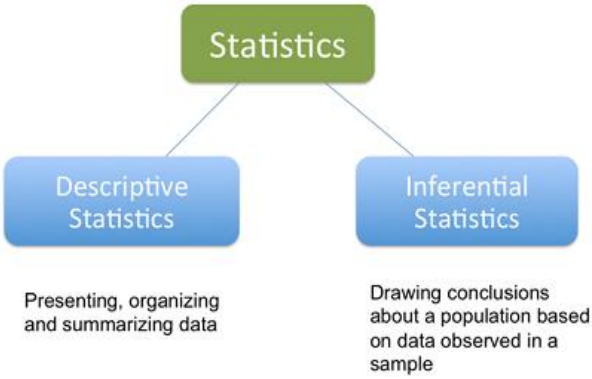
Zero slope

Thus the Linear regression is used to predict the value of one variable (dependent variable) on the basis of other variables (independent variables)

15. What are the various branches of statistics?

The two main branches of statistics are

1. Descriptive statistics
2. Inferential statistics.



Both of these are employed in scientific analysis of data and both are equally important branches of statistics.

Descriptive Statistics

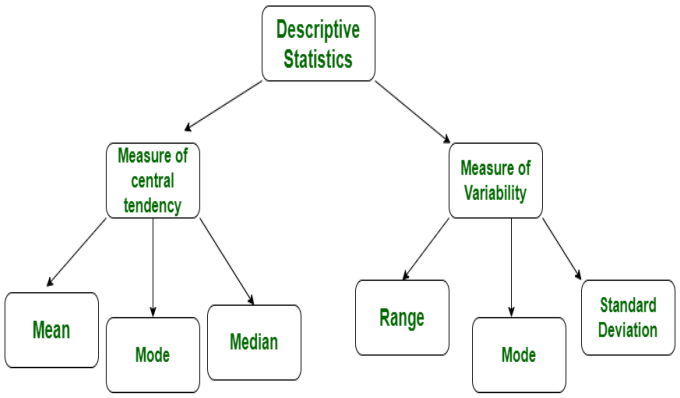
Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a

statistical analysis. It focuses on collecting, summarizing and presenting set of data

Example: A house hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full. Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this particular branch “descriptive-statistics".

Descriptive statistics have two parts

1. Central tendency measures
2. Variability measures



1. **Measures of Central Tendency:**

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

Mean

Mean is a conventional method used to describe the central tendency. Typically, to calculate the average of values, count all values, and then divide them with the number of available values.

Median

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

Mode

The mode is the frequently occurring value in the given data set.

2. **Measures of Variability**

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistic
It analyses sample data to draw conclusion about population

The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Inference statistics often speak in terms of probability by using descriptive statistics. Besides, these techniques are used primarily by a statistician for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.
Most predictions of the future and generalization on a population study of a smaller specimen are in the scope of the inference statistics

Example:

We want to have an idea about percentage of illiterates in a country. We take a sample from a population and the proportion of illiterates in the sample. That sample with the help of probability enables us to find the proportion to the original population. We start with a hypothesis, and see whether the data is consistent with that hypothesis.

Different types of inferential statistics include:

1. Regression analysis
2. Analysis of variance (ANOVA)
3. Analysis of covariance (ANCOVA)
4. Statistical significance (t-test)
5. Correlation analysis

Sr. No	**Descriptive statistics**	**Inferential Statistics**
1.	It gives information about raw data which describes the data in some manner.	It makes inference about population using data drawn from the population.
2.	It helps in organizing, analyzing and to present data in a meaningful manner.	It allows us to compare data, make hypothesis and predictions.
3.	It is used to describe a situation.	It is used to explain the chance of occurrence of an event.
4.	It explain already known data and limited to a sample or population having small size.	It attempts to reach the conclusion about the population.
5.	It can be achieved with the help of charts, graphs, tables etc.	It can be achieved by probability.
6.	Measure of tendency, variations etc. is used	Uses hypothesis testing,ANOVA test etc.
7.	Works with smaller dataset	Works with a large dataset

Thank-you.