PYTHON – WORKSHEET 1

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question

1. Which of the following operators is used to calculate remainder in a division?

A) # B) & **C) %** D) $

2. In python 2//3 is equal to?

A) 0.666 **B) 0** C) 1 D) 0.67

3. In python, 6<<2 is equal to?

A) 36 B) 10 **C) 24** D) 45

4. In python, 6&2 will give which of the following as output?

A) 2 B) True C) False D) 0

5. In python, 6|2 will give which of the following as output?

A) 2 B) 4 C) 0 **D) 6**

6. What does the finally keyword denotes in python?

A) It is used to mark the end of the code
B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block.
C) the finally block will be executed no matter if the try block raises an error or not.
D) None of the above

7. What does raise keyword is used for in python?

A) It is used to raise an exception. B) It is used to define lambda function
C) it's not a keyword in python. D) None of the above

8. Which of the following is a common use case of yield keyword in python?

A) in defining an iterator B) while defining a lambda function

C) in defining a generator D) in for loop

Q9 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

9. Which of the following are the valid variable names?

A) _abc B) 1abc **C) abc2** D) None of the above

10. Which of the following are the keywords in python?

A) yield **B) raise** C) look-in D) all of the above

STATISTICS

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. In hypothesis testing, type II error is represented by β and the power of the test is 1−β then β is:

A. The probability of rejecting H0 when H1 is true
B. The probability of failing to reject H0 when H1 is true
C. The probability of failing to reject H1 when H0 is true
D. The probability of rejecting H0 when H1 is true

2. In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the

A. correct hypothesis **B. null hypothesis**

C. alternative hypothesis D. level of significance

3. When the null hypothesis has been true, but the sample information has resulted in the rejection of the null, a _________ has been made

A. level of significance B. Type II error C. critical value **D. Type I error**

4.For finding the p-value when the population standard deviation is unknown, if it is reasonable to assume that the population is normal, we use

A. the z distribution
B. the t distribution with n - 1 degrees of freedom
C. the t distribution with n + 1 degrees of freedom
D. none of the above

5. A Type II error is the error of

A. accepting Ho when it is false B. accepting Ho when it is true

C. rejecting Ho when it is false **D. rejecting Ho when it is true**

6. A hypothesis test in which rejection of the null hypothesis occurs for values of the point estimator in either tail of the sampling distribution is called

A. the null hypothesis B. the alternative hypothesis

C. a one-tailed test **D. a two-tailed test**

7. In hypothesis testing, the level of significance is

A. the probability of committing a Type II error
B. the probability of committing a Type I error
C. the probability of either a Type I or Type II, depending on the hypothesis to be tested
D. none of the above

8. In hypothesis testing, b is

A. the probability of committing a Type II error

B. the probability of committing a Type I error

C. the probability of either a Type I or Type II, depending on the hypothesis to be test

D. none of the above

9. When testing the following hypotheses at an α level of significance H0: p = 0.7 H1: p > 0.7 The null hypothesis will be rejected if the test statistic Z is

A. z > zα B. z < zα C. z < -z D. none of the above

10. Which of the following does not need to be known in order to compute the P-value?

A. knowledge of whether the test is one-tailed or two-tail
B. the value of the test statistic
C. the level of significance
D. All of the above are needed

11. The maximum probability of a Type I error that the decision maker will tolerate is called the

A. level of significance B. critical value

C. decision value D. probability value

12. For t distribution, increasing the sample size, the effect will be on

A. Degrees of Freedom B. The t-ratio

C. Standard Error of the Means **D. All of the Above**

Q13 to Q15 are subjective answers type questions. Answers them in their own words briefly.

13. What is Anova in SPSS?

ANOVA IN SPSS:

ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations.

ANOVA in SPSS must have a dependent variable which should be metric (measured using an interval or ratio scale). ANOVA in SPSS must also have one or more independent variables, which should be categorical in nature. In ANOVA in SPSS, categorical independent variables are called factors. A particular combination of factor levels, or categories, is called a treatment.

In ANOVA in SPSS, there is one way ANOVA which involves only one categorical variable, or a single factor. For example, if a researcher wants to examine whether heavy, medium, light and nonusers of cereals differed in their preference for Total cereal, then the differences can be examined by the one way ANOVA in SPSS. In one way ANOVA in SPSS, a treatment is the same as the factor level.

If two or more factors are involved in ANOVA in SPSS, then it is termed as n way ANOVA. For example, if the researcher also wants to examine the preference for Total cereal by the customers who are loyal to it and those who are not, then we can use n way ANOVA in SPSS.
In ANOVA in SPSS, from the menu we choose:

“Analyze” then go to “Compare Means” and click on the “One-Way ANOVA.”

Now, let us discuss in detail how the software operates ANOVA:

The first step is to identify the dependent and independent variables. The dependent variable is generally denoted by Y and the independent variable is denoted by X. X is a categorical variable having c categories. The sample size in each category of X is generally denoted as n, and the total sample size N=nXc.

The next step in ANOVA in SPSS is to examine the differences among means. This involves decomposition of the total variation observed in the dependent variable. This variation in ANOVA in SPSS is measured by the sums of the squares of the mean.

The total variation in Y in ANOVA in SPSS is denoted by SSy, which can be decomposed into two components:

SSy=SSbetween+SSwithin

where the subscripts between and within refers to the categories of X in ANOVA in SPSS. SSbetween is the portion of the sum of squares in Y related to the independent variable or factor X. Thus it is generally referred to as the sum of squares of X. SSwithin is the variation in Y related to the variation within each category of X. It is generally referred to as the sum of squares for errors in ANOVA in SPSS.

The logic behind decomposing SSy is to examine the differences in group means. The next task in ANOVA in SPSS is to measure the effects of X on Y, which is generally done by the sum of squares of X, because it is related to the variation in the means of the categories of X. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the differences among the means of Y in categories of X increases. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the variation in Y within the categories of X decreases.

The strength of the effects of X on Y is measured with the help of η2 in ANOVA in SPSS .The value of η2 varies between 0 and 1. It assumes a value 0 in ANOVA in SPSS when all the category means are equal, indicating that X has no effect on Y. The value of η2 becomes 1, when there is no variability within each category of X but there is still some variability between the categories.

The final step in ANOVA in SPSS is to calculate the mean square which is obtained by dividing the sum of squares by the corresponding degrees of freedom. The null hypothesis of equal means, which is done by an F statistic, is the ratio between the mean square related to the independent variable and the mean square related to the error.

N way ANOVA in ANOVA in SPSS involves simultaneous examination of two or more categorical independent variables, which is also computed in a similar manner.

A major advantage of ANOVA in SPSS is that the interactions between the independent variables can be examined.

14. What are the assumptions of Anova?

There are three primary assumptions in ANOVA:

1. The responses for each factor level have a **normal population distribution**.
2. These distributions have the **same variance**.
3. The data are **independent**.
4. Within each sample, the observations are sampled randomly and independently of each other
5. Factor effects are additive

15. What is the difference between one way Anova and two way Anova?

	One-Way ANOVA	**Two-Way ANOVA**
Definition	A test that allows one to make comparisons between the means of three or more groups of data.	A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.
Number of Independent Variables	One.	Two.
What is Being Compared?	The means of three or more groups of an independent variable on a dependent variable.	The effect of multiple groups of two independent variables on a dependent variable and on each other.
Number of Groups of Samples	Three or more.	Each variable should have multiple samples.

MACHINE LEARNING ASSIGNMENT

In Q1 to Q7, only one option is correct, Choose the correct option:

1. What is the advantage of hierarchical clustering over K-means clustering?

A) Hierarchical clustering is computationally less expensive
B) In hierarchical clustering you don’t need to assign number of clusters in beginning
C) Both are equally proficient
D) None of these

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

A) max_depth B) n_estimators C) min_samples_leaf D) min_samples_splits

3. Which of the following is the least preferable resampling method in handling imbalance datasets?

A) SMOTE B) RandomOverSampler **C) RandomUnderSampler** D) ADASYN

4. Which of the following statements is/are true about “Type-1” and “Type-2” errors?

1. Type1 is known as false positive and Type2 is known as false negative.
2. Type1 is known as false negative and Type2 is known as false positive.
3. Type1 error occurs when we reject a null hypothesis when it is actually true.

A) 1 and 2 B) 1 only **C) 1 and 3** D) 2 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur:

1. Randomly selecting the cluster centroids
2. Updating the cluster centroids iteratively
3. Assigning the cluster points to their nearest center

A) 3-1-2 B) 2-1-3 C) 3-2-1 **D) 1-3-2**

6. Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?

A) Decision Trees **B) Support Vector Machines**
C) K-Nearest Neighbors D) Logistic Regression

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

A) CART is used for classification, and CHAID is used for regression.

B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of two children for a node).
C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)
D) None of the above

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

A) Ridge will lead to some of the coefficients to be very close to 0
B) Lasso will lead to some of the coefficients to be very close to 0
C) Ridge will cause some of the coefficients to become 0
D) Lasso will cause some of the coefficients to become 0.

9. Which of the following methods can be used to treat two multi-collinear features?

A) remove both features from the dataset
B) remove only one of the features
C) Use ridge regularization
D) use Lasso regularization

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

A) Overfitting B) Multicollinearity **C) Underfitting** D) Outliers

Q10 to Q15 are subjective answer type questions, Answer them briefly.

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Encoding Categorical Data

One-Hot Encoding

For categorical variables where no ordinal relationship exists, the integer encoding may not be enough, at best, or misleading to the model at worst.

Forcing an ordinal relationship via an ordinal encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a one-hot encoding can be applied to the ordinal representation. This is where the integer encoded variable is removed and one new binary variable is added for each unique integer value in the variable.

If you know all of the labels to be expected in the data, they can be specified via the "*categories*" argument as a list.
The encoder is fit on the training dataset, which likely contains at least one example of all expected labels for each categorical variable if you do not specify the list of labels. If new data contains categories not seen in the training dataset, the "*handle_unknown*" argument can be set to "*ignore*" to not raise an error, which will result in a zero value for each label. The one-hot encoding creates one binary variable for each category.

The problem is that this representation includes redundancy. For example, if we know that [1, 0, 0] represents "*blue*" and [0, 1, 0] represents "*green*" we don't need another binary variable to represent "*red*", instead we could use 0 values for both "*blue*" and "*green*" alone, e.g. [0, 0].
This is called a dummy variable encoding, and always represents C categories with C-1 binary variables.

In addition to being slightly less redundant, a dummy variable representation is required for some models.

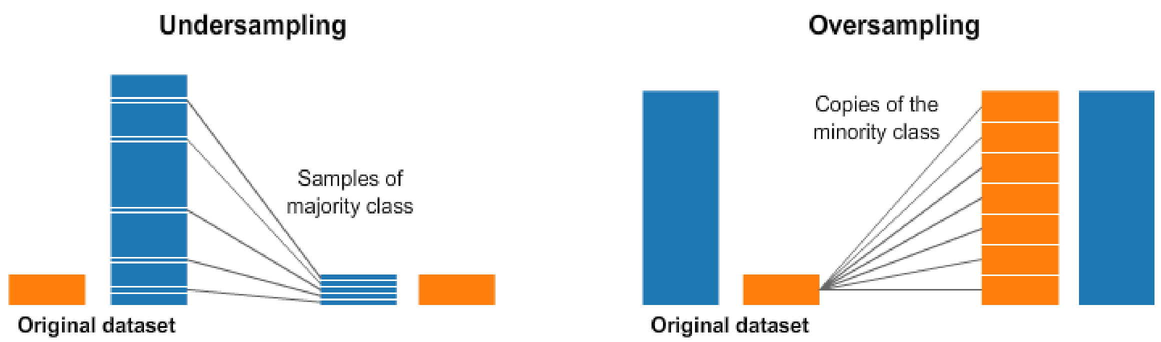
For example, in the case of a linear regression model (and other regression models that have a bias term), a one hot encoding will case the matrix of input data to become singular, meaning it cannot be inverted and the linear regression coefficients cannot be calculated using linear algebra. For these types of models a dummy variable encoding or label encoding must be used instead.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Imbalanced datasets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed by two classes: The majority (negative) class and the minority (positive) class

the various techniques you can use to handle imbalanced datasets.

1. Random Undersampling and Oversampling



A widely adopted and perhaps the most straightforward method for dealing with highly imbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling).

2. Undersampling and Oversampling using imbalanced-learn

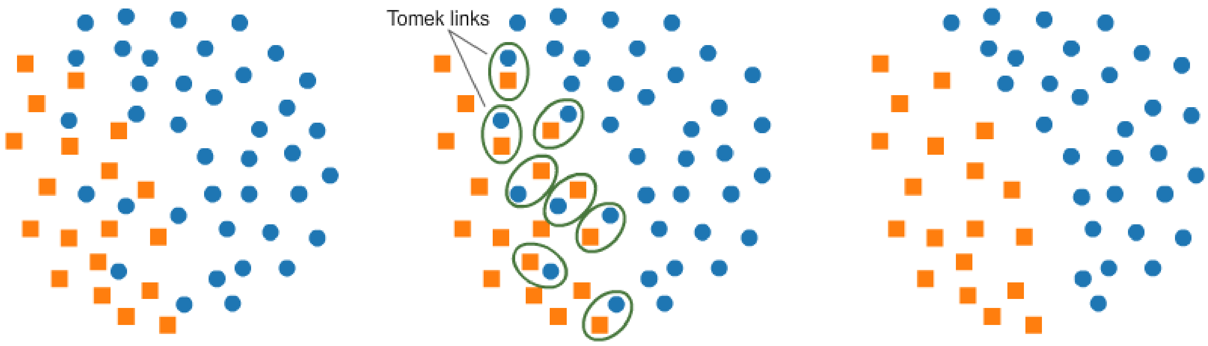
imbalanced-learn(imblearn) is a Python Package to tackle the curse of imbalanced datasets.

It provides a variety of methods to undersample and oversample.

a. Undersampling using Tomek Links:

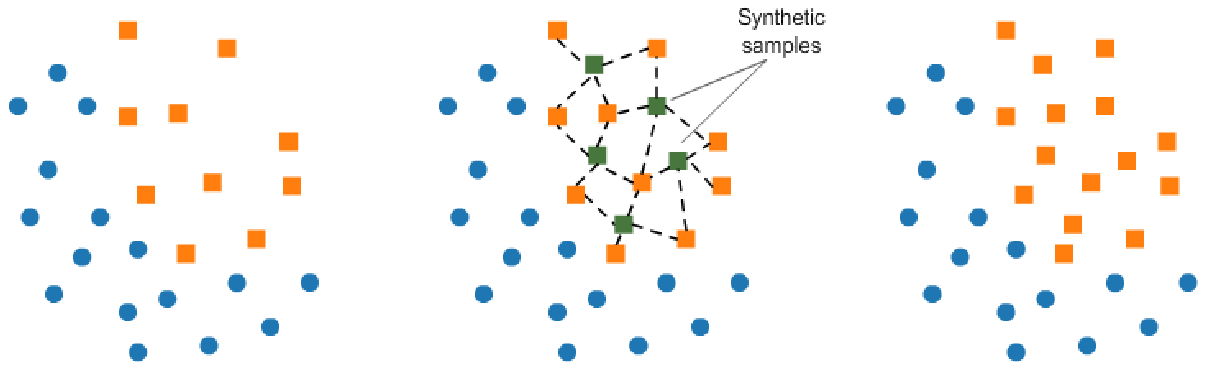
One of such methods it provides is called Tomek Links. Tomek links are pairs of examples of opposite classes in close vicinity.

In this algorithm, we end up removing the majority element from the Tomek link, which provides a better decision boundary for a classifier.



b. Oversampling using SMOTE:

In SMOTE (Synthetic Minority Oversampling Technique) we synthesize elements for the minority class, in the vicinity of already existing elements.



There are a variety of other methods in the imblearn package for both undersampling(Cluster Centroids, NearMiss, etc.) and oversampling(ADASYN and bSMOTE)

Miscellaneous

Various other methods might work depending on your use case and the problem you are trying to solve:

a) Collect more Data

This is a definite thing you should try if you can. Getting more data with more positive examples is going to help your models get a more varied perspective of both the majority and minority classes.

b) Treat the problem as anomaly detection

You might want to treat your classification problem as an anomaly detection problem.

Anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data

You can use Isolation forests or autoencoders for anomaly detection.

c) Model-based

Some models are particularly suited for imbalanced datasets.

For example, in boosting models, we give more weights to the cases that get misclassified in each tree iteration.

13. What is the difference between SMOTE and ADASYN sampling techniques?

SMOTE:
What smote does is simple. First it finds the n-nearest neighbors in the minority class for each of the samples in the class . Then it draws a line between the the neighbors an generates random points on the lines.
ADASYN:
Its a improved version of Smote. What it does is same as SMOTE just with a minor improvement. After creating those sample it adds a random small values to the points thus making it more realistic. In other words instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e they are bit scattered.

The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions.

ADASYN is similar to SMOTE, and derived from it, featuring just one important difference. it will bias the sample space (that is, the likelihood that any particular point will be chosen for duping) towards points which are located not in homogenous neighborhoods.

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

GridSearchCV is a technique for **finding the optimal parameter values from a given set of parameters in a grid**. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.

GridSearchCV is a method used to find the best hyperparameters for a given model in order to maximize its performance. It is preferable to use in case of large datasets as it is more efficient in finding the best hyperparameters compared to manual tuning. GridSearchCV automates the process of tuning hyperparameters, which would take a lot of time and effort if done manually.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are:

- Mean Squared Error (MSE).
- Root Mean Squared Error (RMSE).
- Mean Absolute Error (MAE)

Mean Squared Error

Mean Squared Error, or MSE for short, is a popular error metric for regression problems.

It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here "*least squares*" refers to minimizing the mean squared error between predictions and expected values.

The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

- MSE = 1 / N * sum for i to N (y_i – yhat_i)^2

Where *y_i* is the i'th expected value in the dataset and *yhat_i* is the i'th predicted value. The difference between these two values is squared, which has the effect of removing the sign, resulting in a positive error value.

The squaring also has the effect of inflating or magnifying large errors. That is, the larger the difference between the predicted and expected values, the larger the resulting squared positive error. This has the effect of "*punishing*" models more for larger errors when MSE is used as a loss function. It also has the effect of "*punishing*" models by inflating the average error score when used as a metric.

We can create a plot to get a feeling for how the change in prediction error impacts the squared error.

The units of the MSE are squared units.

Root Mean Squared Error

The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.

Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.

For example, if your target variable has the units "*dollars*," then the RMSE error score will also have the unit "*dollars*" and not "*squared dollars*" like the MSE.

As such, it may be common to use MSE loss to train a regression predictive model, and to use RMSE to evaluate and report its performance.

The RMSE can be calculated as follows:

- RMSE = sqrt(1 / N * sum for i to N (y_i – yhat_i)^2)

Where *y_i* is the i'th expected value in the dataset, *yhat_i* is the i'th predicted value, and *sqrt()* is the square root function.

We can restate the RMSE in terms of the MSE as:

- RMSE = sqrt(MSE)

Note that the RMSE cannot be calculated as the average of the square root of the mean squared error values

The root mean squared error between your expected and predicted values can be calculated using the mean_squared_error() function from the scikit-learn library.

Mean Absolute Error

Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.

Unlike the RMSE, the changes in MAE are linear and therefore intuitive.

That is, MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error.

As its name suggests, the MAE score is calculated as the average of the absolute error values. Absolute or *abs()* is a mathematical function that simply makes a number positive. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE.

The MAE can be calculated as follows:

- MAE = 1 / N * sum for i to N abs(y_i – yhat_i)

Where *y_i* is the i'th expected value in the dataset, *yhat_i* is the i'th predicted value and *abs()* is the absolute function.

A perfect mean absolute error value is 0.0, which means that all predictions matched the expected values exactly.

WORKSHEET 2 PYTHON

Q1 to Q7 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is not a core datatype in python?

A) list **B) struct** C) tuple C) set

2. Which of the following is an invalid variable name in python?

A) _init_ B) no_1 **C) 1_no** D) _1

3. Which one of the following is a keyword in python?

A) in B) _init_ C) on D) foo

4. In which of the following manner are the operators of the same precedence executed in python?

A) Left to Right B) BODMAS C) Right to Left D) None of these

5. Arrange the following in decreasing order of the precedence when they appear in an expression in python?

i) Multiplication ii) Division iii) Exponential iv) Parentheses

A) iii – iv – ii – i B) iii – iv – i – ii **C) iv – iii – ii – i** D) iii – ii – i – iv

6. (28//6)**3/3%3 = ?

A) 7.1111… B) 0 **C) 0.3333…** D) 1

7. a = input("Enter an integer"). What will be the data type of a?

A) int **B) str** C) float D) double

Q8 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

8. Which of the following statements are correct?

A) Division and multiplication have same precedence in python
B) Python's operators' precedence is based on PEDMAS
C) Python's operators' precedence is based on VBODMAS
D) In case of operators' having the same precedence, the one on the left side is executed first.

9. Which of the following is(are) valid statement(s) in python?

A) abc = 1,000,000 B) a b c = 1000 2000 3000

C) a,b,c = 1000, 2000, 3000 D) a_b_c = 1,000,000 ASSIGNMENT

10. Which of the following is not equal to x16 in python?

A) x**4**4 B) x**16 **C) x^16** D) (x**4)**4

Q11 to Q13 are subjective questions, answer them briefly

11. Differentiate between a list, tuple, set and dictionary.

List	Tuple	Set	Dictionary
List can be written as comma separated values(items) between squared bracket.	Tuple can be written as comma separated values(items) enclosed in parentheses	Set is collection of data with no duplicate value, it represented in curly brackets.	Dictionaries are enclosed in curly brackets in the form of [key:value] pairs.
List is an ordered collection of data.	Tuple is an unordered collection of data.	Set is an unordered collection of data.	Dictionary is an unordered collection of data.
List are mutable i.e. it can changed the data after creation.	Tuples are unmutable.	Sets are unmutable.	Dictionary are mutable.
Creating an empty list l=[ ]	Creating an empty Tuple t=( )	Creating an empty set a=set( )	Creating an empty dictionary d={ }
List allows duplicate elements	Tuple allows duplicate elements	Set will not allow duplicate elements	Dictionary doesn't allow duplicate keys.
List can be represented by [ ]	Tuple can be represented by ( )	Set can be represented by { }	Dictionary can be represented by { }

12. Are strings mutable in python? Suppose you have a string "I+Love+Python", write a small code to replace '+' with space in python.

Python strings are immutable i. e. it can not be changed.

If we have a string "I+Love+Python", we have to replace '+' with space in python, we use following code

```
In [2]: string ="I+Love+Python"
        string

Out[2]: 'I+Love+Python'

In [3]: string.replace("+"," ")

Out[3]: 'I Love Python'
```

13. What does the function ord() do in python? Explain with an example. Also, write down the function for getting the data type of a variable in python.

ord() : The ord() function returns the number representing the unicode code of a specified character.

This function accepts a string of unit length as an argument and returns the Unicode equivalence of the passed argument. In other words, given a string of length 1, the ord() function returns an integer representing the Unicode code point of the character when an argument is a Unicode object, or the value of the byte when the argument is an 8-bit string.

Syntax: ord(ch)

ch – A unicode character

example:

In [3]:
```
# inbuilt function return an
# integer representing the Unicode code
value = ord("A")

# writing in ' ' gives the same result
value1 = ord('A')

# prints the unicode value
print (value, value1)
```

```
65 65
```

Data type: data type of variable in python Python has a number of built-in data types, including integers, floats, strings, booleans, lists, tuples, dictionaries, and more.

To get the type of a variable in Python, you can use the built-in type() function.

The basic syntax looks like this:
type(variableName)

STATISTICS

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. The owner of a travel agency would like to determine whether or not the mean age of the agency's customers is over 24. If so, he plans to alter the destination of their special cruises and tours. If he concludes the mean age is over 24 when it is not, he makes a ________ error. If he concludes the mean age is not over 24 when it is, he makes a _______error.

A. Type II; Type II B. Type I; Type I **C. Type I; Type II** D. Type II; Type I

2. Suppose we wish to test H0: μ =53 vs H1: μ > 53. What will result if we conclude that the mean is greater than 53 when its true value is really 55?

a. We have made a Type I error **b. We have made a correct decision**

c. We have made a Type II error d. None of the above are correct

3. The value that separates a rejection region from an acceptance region is called a _____________.

A. parameter B. critical value C. confidence coefficient **D. significance level**

4. A hypothesis test is used to prevent a machine from under filling or overfilling quart bottles of beer. On the basis of sample, the machine is shut down for inspection. A thorough examination reveals there is nothing wrong with the filling machine. From a statistical point of view:

A. Both Type I and Type II errors were made. **B. A Type I error was made.**

C. A Type II error was made. D. A correct decision was made.

5. Suppose we wish to test H0 : μ =21 vs H1 : μ > 21. Which of the following possible sample results gives the most evidence to support H1 (i.e., reject H0)? Hint: Compute Z-score.

A. x = 23 s , = 3 B. x = 19 s , = 4 **C. x = 17 s , = 7** D. x = 18 s , = 6

6. Given H0: μ = 25, H1: $\mu \neq$ 25, and P-value = 0.041. Do you reject or fail to reject H0 at the 0.01 level of significance?

A. fail to reject H0 B. not sufficient information to decide

C. reject H0

7. A bottling company needs to produce bottles that will hold 12 ounces of liquid. Periodically, the company gets complaints that their bottles are not holding enough liquid. To test this claim, the bottling company randomly samples 36 bottles. Suppose the p-value of this test turned out to be 0.0455. State the proper conclusion.

A. At α = 0.085, fail to reject the null hypothesis.
B. At α = 0.035, accept the null hypothesis.
C. At α = 0.05, reject the null hypothesis.
D. At α = 0.025, reject the null hypothesis.

8. If a hypothesis test were conducted using α = 0.05, for which of the following p-values would the null hypothesis be rejected?

A. 0.100 **B. 0.041** C. 0.055 D. 0.060

9 . For H1: $\mu > \mu 0$ p-value is 0.042. What will be the p-value for Ha: $\mu < \mu o$?

A. 0.084 B. 0.021 **C. 0.958** D. 0.042

10. The test statistic is t = 2.63 and the p-value is 0.9849. What type of test is this?

A. Right tail B. Two tail **C. Left tail** D. Can't tell

11. The test statistic is z =2.75, the critical value is z = 2.326. The p- value is …

A. Less than the significance level B. Equal to the significance level

C. Large than the significance level

12. The area to the left of the test statistic is 0.375. What is the probability value if this is a left tail test?

A. 0.750 B. 0.375 C. 0.1885 **D. 0.625**

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13.What is T distribution and Z distribution?

T distribution :T-distribution is a probability distribution used to estimate population parameters when the sample size is small or the population variance is unknown. It is a type of continuous probability distribution that is similar to the standard normal distribution, but has heavier tails, meaning that it is more prone to producing values that are further away from the mean.

Z-distribution : Z-distribution is a type of continuous probability distribution that is based on the standard normal distribution. It is used to calculate the probability of an observed statistic being greater than, less than, or equal to a specified value. Z-distributions are used to determine the probability of an event occurring, or the probability of an event not occurring.

14.Is the T distribution normal?

No, the T distribution is not normal. It is a type of symmetric, bell-shaped distribution that is similar to the normal distribution, but has heavier tails.

15.What does the T distribution tell us?

The T distribution is a type of probability distribution that is used to calculate the probability of a statistic in a given population. It is most commonly used to calculate probabilities in small sample sizes when the population standard deviation is unknown. It is a symmetrical, bell-shaped distribution that is similar to the normal distribution, but has heavier tails, meaning that it is more likely to have extreme values at the end of the distribution. The T distribution is used to calculate confidence intervals, hypothesis testing, and other statistical tests.

MACHINE LEARNING ASSIGNMENT

1. Which of the following are disadvantages of using Hard Margin SVM classifier?

A) They allow misclassifications, that's why they are not optimal.
B) They cannot be used when the data is not completely linearly separable while allowing no errors.
C) They are not optimal to use in case of outliers.
D) None of the above.

2. Which of the following statements are true regarding maximal margin classifier?

A) It is the most optimal classifier in a completely linearly separable data.
B) It's the classifier for which the margin length or the distance between the closest data-point on either side of the classifier and the classifier is maximized.
C) Any possible classifier which can linearly separate the data of two classes is called maximal margin classifier.
D) All of the above.

3. Which of the following statements are true regarding soft margin SVM classifier?

A) They are less sensitive to outliers and can be used even in their presence.
B) They make sure that there is no data point present in the margin area.
C) They allow some degree of errors or misclassification.
D) They can be used in case data is not completely linearly separable.

4. Which of the following statements are true regarding SVMs?

A) They take the data from lower dimensional space to some higher dimensional space in case the data is not likely to be linearly separable.
B) They use the kernel tricks to escape the complex computations required to transform the data.
C) If the data is not linearly separable SVM technique cannot be used.
D) All of the above.

5. Which of the following Statements are true regarding the Kernel functions used in SVM?

A) These functions gives value of the dot product of pairs of data-points in the desired higher. dimensional space without even explicitly converting the whole data in to higher dimensional space.
B) We have to first convert the whole data in to the higher dimensional space before applying the kernel function.
C) The data product values given by the kernel functions are used to find the classifier in the higher dimensional space.
D) None of the above

6. How can SVM be classified?

A) It is a model trained using unsupervised learning. It can be used for classification and regression.
B) It is a model trained using unsupervised learning. It can be used for classification but not for regression
C) It is a model trained using supervised learning. It can be used for classification and regression.
D) It is a model trained using supervised learning. It can be used for classification not for regression.

7. The quality of an SVM model depends upon:

A) Selection of Kernel
B) Kernel Parameters
C) Soft Margin Parameter C
D) All of the above

8. The SVM's are less effective when:

A) The data is linearly separable.
B) The data is clean and ready to use.
C) The data is noisy and contains overlapping points.
D) None of these.

9. What would happen when you use very small C (C~0)?

A) Misclassification would happen.
B) Data will be correctly classified.
C) Can't say
D) None of these.

10. What do you mean by generalization error in terms of the SVM?

A) How far the hyperplane is from the support vectors.
B) How accurately the SVM can predict outcomes for unseen data.

C) The threshold amount of error in an SVM.
D) None of these

WORKSHEET 3 PYTHON

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following will raise a value error in python?

A) int(32) B) int(3.2) C) int(-3.2) D) **int(‘32’)**

2. What will be the output of round(3.567)?

A) 3.5 B) 3.0 **C) 4** D) 3

3. How is the function pow(a,b,c) evaluated in python?

A) a**b**c **B) (a**b)%c** C) (a**b)*c D) (a**b)**c

4. What will be the output of print(type(type(int))) in python 3?

A) < class ‘type’> B) < type ‘type’>

C) < class ‘int’> D) < type ‘type’>

5. What will be the output of ord(chr(65))?

A) ‘A’ B) ‘a’ **C) 65** D) TypeError

6. What is called when a function is defined inside a class?

A) Module B) Function C) _init_ function **D) Method**

7. What will be the output of all([1, 0, 5 ,7])?

A) 0 **B) False** C) True D) error

8. Is the output of the function abs() the same as that of the function math.fabs()?

A) Always B) Sometimes C) Never D) None of these

Q9 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

9. Select all correct float numbers in python?

A) -68.7e100 B) 42e3 **C) 4.2038** **D) 3.0**

10. Which of the following is(are) correct statement(s) in python?

A) You can pass positional arguments in any order.

B) You can pass keyword arguments in any order.

C) You can call a function with positional and keyword arguments.

D) Positional arguments must be before keyword arguments in a function call

STATISTICS

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Rejection of the null hypothesis is a conclusive proof that the alternative hypothesis is

A. True B. False C. Neither

2. Parametric test, unlike the non-parametric tests, make certain assumptions about

A. The population size B. The underlying distribution C. The sample size

3. The level of significance can be viewed as the amount of risk that an analyst will accept when making a decision

A. True **B. False**

4. By taking a level of significance of 5% it is the same as saying

A. We are 5% confident the results have not occurred by chance

B. We are 95% confident that the results have not occurred by chance

C. We are 95% confident that the results have occurred by chance

5. One or two tail test will determine

A. If the two extreme values (min or max) of the sample need to be rejected

B. If the hypothesis has one or possible two conclusions

C. If the region of rejection is located in one or two tails of the distribution

6. Two types of errors associated with hypothesis testing are Type I and Type II. Type II error is committed when

A. We reject the null hypothesis whilst the alternative hypothesis is true

B. We reject a null hypothesis when it is true

C. We accept a null hypothesis when it is not true

7. A randomly selected sample of 1,000 college students was asked whether they had ever used the drug Ecstasy. Sixteen percent (16% or 0.16) of the 1,000 students surveyed said they had. Which one of the following statements about the number 0.16 is correct?

A. It is a sample proportion.

B. It is a population proportion.

C. It is a margin of error.

D. It is a randomly chosen number.

8. In a random sample of 1000 students, pˆ = 0.80 (or 80%) were in favour of longer hours at the school library. The standard error of pˆ (the sample proportion) is

A. .013 B. .160 C. .640 D. .800

9. For a random sample of 9 women, the average resting pulse rate is x = 76 beats per minute, and the sample standard deviation is s = 5. The standard error of the sample mean is

A. 0.557 B. 0.745 **C. 1.667** D. 2.778

10. Assume the cholesterol levels in a certain population have mean μ= 200 and standard deviation σ = 24. The cholesterol levels for a random sample of n = 9 individuals are measured and the sample mean x is determined. What is the z-score for a sample mean x = 180?

A. –3.75 **B. –2.50** C. −0.83 D. 2.50

11. In a past General Social Survey, a random sample of men and women answered the question “Are you a member of any sports clubs?” Based on the sample data, 95% confidence intervals for the population proportion who would

answer “yes” are .13 to .19 for women and .247 to .33 for men. Based on these results, you can reasonably conclude that

A. At least 25% of American men and American women belong to sports clubs.

B. At least 16% of American women belong to sports clubs.

C. There is a difference between the proportions of American men and American women who belong to sports clubs.

D. There is no conclusive evidence of a gender difference in the proportion belonging to sports clubs.

12. Suppose a 95% confidence interval for the proportion of Americans who exercise regularly is 0.29 to 0.37. Which one of the following statements is FALSE?

A. It is reasonable to say that more than 25% of Americans exercise regularly.

B. It is reasonable to say that more than 40% of Americans exercise regularly.

C. The hypothesis that 33% of Americans exercise regularly cannot be rejected.

D. It is reasonable to say that fewer than 40% of Americans exercise regularly.

Q13 to Q15 are subjective answers type questions. Answers them in their own words briefly.

13. How do you find the test statistic for two samples?

Two-Sample Independent t-Test

A two-sample independent *t*-test can be run on sample data from a normally distributed numerical outcome variable to determine if its mean differs across two independent groups. For example, we could see if the mean GPA differs between freshman and senior college students by collecting a sample of each group of students and recording their GPAs.

Hypotheses:
H_0: The population mean of one group equals the population mean of the other group, or $\mu_1 = \mu_2$
H_A: The population mean of one group does not equal the population mean of the other group, or $\mu_1 \neq \mu_2$

This test can also be conducted with a directional alternate hypothesis:
H_0: The population mean of one group equals the population mean of the other group, or

$\mu_1 = \mu_2$
H_a: The population mean of one group is greater than the population mean of the other group, or $\mu_1 > \mu_2$

The test statistic for a two-sample independent *t*-test is calculated by taking the difference in the two sample means and dividing by either the pooled or unpooled estimated standard error. The estimated standard error is an aggregate measure of the amount of variation in both groups.

Relevant Equations:

Degrees of freedom: Varies by conditions, but the basic rule of thumb for hand calculations is the smaller of $n_1 - 1$ and $n_2 - 1$, where n is the sample size for each group.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

Assumptions:

- Random samples
- Independent observations
- The population of **each group** is normally distributed.
- The population variances are equal.

If the third assumption is not met, the alternative test is the Mann-Whitney *U*-Test, which can be run to see if there is a difference between two groups for a variable with any type of distribution.

If the fourth assumption is met, then the pooled estimated standard error is used in the calculation of the test statistic. If the fourth assumption fails, then the more conservative unpooled estimated standard error is used (and the test is referred to as "Welch's Test").

14. How do you find the sample mean difference?

Difference Between Means: Theory

Suppose we have two populations with means equal to μ_1 and μ_2. Suppose further that we take all possible samples of size n_1 and n_2. And finally, suppose that the following assumptions are valid.

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2. (In this context, populations are considered to be large if they are at least 20 times bigger than their sample.)
- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.
- The set of differences between sample means is normally distributed. This will be true if each population is normal or if the sample sizes are large. (Based on the central limit theorem, sample sizes of 40 would probably be large enough).

Given these assumptions, we know the following.

- The expected value of the difference between all possible sample means is equal to the difference between population means. Thus,

  $E(x_1 - x_2) = \mu_d = \mu_1 - \mu_2$.

- The standard deviation of the difference between sample means (σ_d) is approximately equal to:

  $\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$

It is straightforward to derive the last bullet point, based on material covered in previous lessons. The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

$\sigma^2{}_d = \sigma^2{}_{(x1 - x2)} = \sigma^2{}_{x1} + \sigma^2{}_{x2}$

If the populations N_1 and N_2 are both large relative to n_1 and n_2, respectively, then

$\sigma^2{}_{x1} = \sigma^2{}_1 / n_1$

$\sigma^2{}_{x2} = \sigma^2{}_2 / n_2$

$\sigma_d^2 = \sigma_1^2 / n_1 + \sigma_2^2 / n_2$

$\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$

Problem 1

For boys, the average number of absences in the first grade is 15 with a standard deviation of 7; for girls, the average number of absences is 10 with a standard deviation of 6.

In a nationwide survey, suppose 100 boys and 50 girls are sampled. What is the probability that the male sample will have *at most* three more days of absences than the female sample?

(A)0.025
(B) 0.035
(C) 0.045
(D) 0.055
(E) None of the above

Solution

The correct answer is B. The solution involves three or four steps, depending on whether you work directly with raw scores or z-scores. The "raw score" solution appears below:

- Find the mean difference (male absences minus female absences) in the population.

  $\mu_d = \mu_1 - \mu_2 = 15 - 10 = 5$

- Find the standard deviation of the difference.

  $\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$

  $\sigma_d = \text{sqrt}(7^2/100 + 6^2/50)$

  $\sigma_d = \text{sqrt}(49/100 + 36/50)$

  $\sigma_d = \text{sqrt}(0.49 + .72) = \text{sqrt}(1.21) = 1.1$

- Find the probability. This problem requires us to find the probability that the average number of absences in the boy sample minus the average number of absences in the girl sample is less than 3. To find this probability, we use Stat Trek's Normal Distribution Calculator. Specifically, we enter the following inputs: 3, for the normal random variable; 5, for the mean; and 1.1, for the standard deviation.

We find that the probability of the mean difference (male absences minus female absences) being 3 or less is about 0.035.

Alternatively, we could have worked with z-scores (which have a mean of 0 and a standard deviation of 1). Here's the z-score solution:

- Find the mean difference (male absences minus female absences) in the population.

  $\mu_d = \mu_1 - \mu_2 = 15 - 10 = 5$

- Find the standard deviation of the difference.

  $\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$

  $\sigma_d = \text{sqrt}(7^2/100 + 6^2/50) = \text{sqrt}(49/100 + 36/50)$

  $\sigma_d = \text{sqrt}(0.49 + .72) = \text{sqrt}(1.21) = 1.1$

- Find the z-score that is produced when boys have three more days of absences than girls. When boys have three more days of absences, the number of male absences minus female absences is three. And the associated z-score is

  $z = (x - \mu)/\sigma = (3 - 5)/1.1 = -2/1.1 = -1.818$

- Find the probability. To find this probability, we use Stat Trek's Normal Distribution Calculator. Specifically, we enter the following inputs: -1.818, for the normal random variable; 0, for the mean; and 1, for the standard deviation.

We find that the probability of probability of a z-score being -1.818 or less is about 0.035. Of course, the result is the same, whether you work with raw scores or with z-scores.

15. What is a two sample t test example?

Two Sample t-test: Example

Suppose we want to know whether or not the mean weight between two different species of turtles is equal. To test this, will perform a two sample t-test at significance level $\alpha = 0.05$ using the following steps:

Step 1: Gather the sample data.

Suppose we collect a random sample of turtles from each population with the following information:

Sample 1:

- Sample size $n_1 = 40$
- Sample mean weight $x_1 = 300$

- Sample standard deviation $s_1 = 18.5$

Sample 2:

- Sample size $n_2 = 38$
- Sample mean weight $x_2 = 305$
- Sample standard deviation $s_2 = 16.7$

Step 2: Define the hypotheses.

We will perform the two sample t-test with the following hypotheses:

- **H_0:** $\mu_1 = \mu_2$ (the two population means are equal)
- **H_1:** $\mu_1 \neq \mu_2$ (the two population means are not equal)

Step 3: Calculate the test statistic *t*.

First, we will calculate the pooled standard deviation s_p:

$$s_p = \sqrt{(n_1-1)s_1^2 + (n_2-1)s_2^2 / (n_1+n_2-2)} = \sqrt{(40-1)18.5^2 + (38-1)16.7^2 / (40+38-2)} = \mathbf{17.647}$$

Next, we will calculate the test statistic *t*:

$$t = (x_1 - x_2) / s_p(\sqrt{1/n_1 + 1/n_2}) = (300-305) / 17.647(\sqrt{1/40 + 1/38}) = \mathbf{-1.2508}$$

Step 4: Calculate the p-value of the test statistic *t*.

According to the T Score to P Value Calculator, the p-value associated with t = -1.2508 and degrees of freedom = n_1+n_2-2 = 40+38-2 = 76 is **0.21484**.

Step 5: Draw a conclusion.

Since this p-value is not less than our significance level α = 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the mean weight of turtles between these two populations is different.

MACHINE LEARNING ASSIGNMENT

In Q1 to Q8, only one option is correct, Choose the correct option:

1. In the linear regression equation y = θ0 + θ1x, θ0 is the:

A) Slope of the line B) Independent variable

C) y intercept D) Coefficient of determination

2. True or False: Linear Regression is a supervised learning algorithm.

A) True B) False

3. In regression analysis, the variable that is being predicted is:

A) the independent variable B) the dependent variable

C) usually denoted by x D) usually denoted by r

4. Generally, which of the following method(s) is used for predicting continuous dependent variables?

A) Logistic Regression B) Linear Regression

C) Both D) None of the above

5. The coefficient of determination is:

A) the square root of the correlation coefficient B) usually less than zero

C) the correlation coefficient squared D) equal to zero

6. If the slope of the regression equation is positive, then:

A) y decreases as x increases B) y increases as x increases

C) y decreases as x decreases D) None of these

7. Linear Regression works best for:

A) linear data B) non-linear data

C) both linear and non-linear data D) None of the above

8. The coefficient of determination can be in the range of:

A) 0 to 1 B) -1 to 1 C) -1 to 0 D) 0 to infinity

In Q9 to Q13, more than one options are correct, Choose all the correct options:

9. Which of the following evaluation metrics can be used for linear regression?

A) Classification Report **B) RMSE** C) ROC curve **D) MAE**

10. Which of the following is true for linear regression?

A) Linear regression is a supervised learning algorithm.

B) Linear regression supports multi-collinearity.

C) Shape of linear regression’s cost function is convex.

D) Linear regression is used to predict discrete dependent variable.

11. Which of the following regularizations can be applied to linear regression?

A) Ridge B) Lasso C) Pruning D) Elastic Net

12. Linear regression performs better for:

A) Large amount of training samples with small number of features.

B) Same number of features and training samples

C) Large number of features

D) The variables which are drawn independently, identically distributed

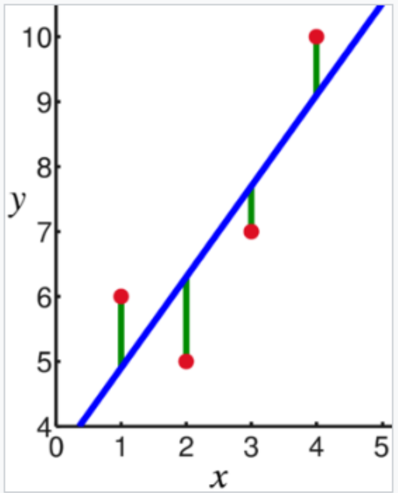
13. Which of the following assumptions are true for linear regression?

A) Linearity **B) Homoscedasticity** C) Non-Independent **D) Normality**

Q14 and Q15 are subjective answer type questions, Answer them briefly.

14. Explain Linear Regression?

Linear Regression: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).



In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a dependent variable (*y*) and an independent variable (*x*).

Important of linear regression

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

we’ll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.
Key assumptions of effective linear regression

Assumptions to be considered for success with linear-regression analysis:

- For each variable: Consider the number of valid cases, mean and standard deviation.
- For each model: Consider regression coefficients, correlation matrix, part and partial correlations, multiple R, R2, adjusted R2, change in R2, standard error of the estimate, analysis-of-variance table, predicted values and residuals. Also, consider 95-percent-confidence intervals for each regression coefficient, variance-covariance matrix, variance inflation factor, tolerance, Durbin-Watson test, distance measures (Mahalanobis, Cook and leverage values), DfBeta, DfFit, prediction intervals and case-wise diagnostic information.
- Plots: Consider scatterplots, partial plots, histograms and normal probability plots.

- Data: Dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.
- Other assumptions: For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear and all observations should be independent.

Here's how you can check for these assumptions:

1. The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.
2. Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
3. The observations should be independent of each other (that is, there should be no dependency).
4. Your data should have no significant outliers.
5. Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.
6. The residuals (errors) of the best-fit regression line follow normal distribution.

Examples of linear-regression success

- Evaluating trends and sales estimates

You can also use linear-regression analysis to try to predict a salesperson's total yearly sales (the dependent variable) from independent variables such as age, education and years of experience.

- Analyze pricing elasticity

Changes in pricing often impact consumer behavior — and linear regression can help you analyze how. For instance, if the price of a particular product keeps changing, you can use regression analysis to see whether consumption drops as the price increases. What if consumption does not drop significantly as the price increases? At what price point do buyers stop purchasing the product? This information would be very helpful for leaders in a retail business.

- Assess risk in an insurance company

Linear regression techniques can be used to analyze risk. For example, an insurance company might have limited resources with which to investigate homeowners' insurance claims; with linear regression, the company's team can build a model for estimating claims costs. The analysis could help company leaders make important business decisions about what risks to take.

- Sports analysis

Linear regression isn't always about business. It's also important in sports. For instance, you might wonder if the number of games won by a basketball team in a season is related to the average number of points the team scores per game. A scatterplot indicates that these variables are linearly related. The number of games won and the average number of points scored by the opponent are also linearly related. These variables have a negative relationship. As the number of games won increases, the average number of points scored by the opponent decreases. With linear regression, you can model the relationship of these variables. A good model can be used to predict how many games teams will win.

15. What is difference between simple linear and multiple linear regression?

Linear Regression	**multiple linear regression**
Simple linear regression has only one x and one y variable.	Multiple linear regression has one y and two or more x variables.
For instance, when we predict rent based on square feet alone that is simple linear regression.	When we predict rent based on square feet and age of the building that is an example of multiple linear regression.
A simple linear regression model usually takes the form of: $y = \beta_0 + \beta_1 x$	With multiple regression models, on the other hand, the equation looks more like this: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
Example: Assume you own chocolate business. A simple linear regression will entail you determining a connection between revenue and product texture, with revenue as the dependent variable and product texture as the independent variable.	Example: A researcher decides to study students' performance from a school over a period of time. He observed that as the lectures proceed to operate online, the performance of students started to decline as well. The parameters for the dependent variable "decrease in performance" are various independent variables like "lack of attention, more internet addiction, neglecting studies" and much more.