PYTHON – WORKSHEET 1

Anwers

1. C
2. D
3. C
4. A
5. D
6. C
7. A
8. C
9. A ,C
10. A,B

Q11 to Q15 Answers in Python Note Book

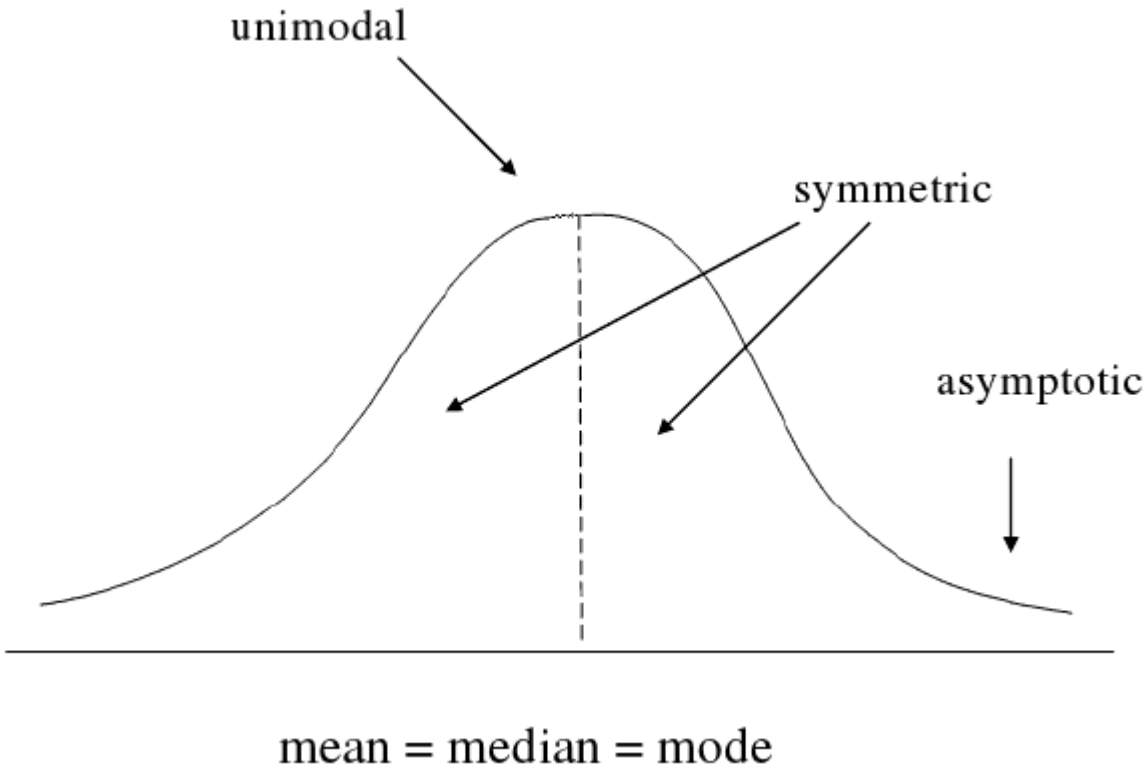
STATICS – WORKSHEET 1

Anwers

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C

Ans 10. A normal distribution is a bell-shaped frequency distribution curve. Most of the data values in a normal distribution tend to cluster around the mean. The further a data point is from the mean, the less likely it is to occur. There are many things, such as intelligence, height, and blood pressure, that naturally follow a normal distribution. For example, if you took the height of one hundred 22-year-old women and created a histogram by plotting height on the x-axis, and the frequency at which each of the heights occurred on the y-axis, you would get a normal distribution.

Characteristics of Normal Distribution



Here, we see the four characteristics of a normal distribution. Normal distributions are **symmetric**, **unimodal**, and **asymptotic**, and the **mean**, **median**, and **mode** are all equal.

A normal distribution is perfectly symmetrical around its center. That is, the right side of the center is a mirror image of the left side. There is also only one mode, or peak, in a normal distribution. Normal distributions are continuous and have tails that are asymptotic, which means that they approach but

never touch the x-axis. The **center of a normal distribution** is located at its peak, and 50% of the data lies above the mean, while 50% lies below. It follows that the mean, median, and mode are all equal in a normal distribution.

Ans 11. Missing data can occur due to many reasons. The data is collected from various sources and, while mining the data, there is a chance to lose the data. However, most of the time cause for missing data is item nonresponse**,** which means people are not willing to answer the questions in a survey, and some people unwillingness to react to sensitive questions like age, salary, gender.

Types of Missing data

Before dealing with the missing values, it is necessary to understand the category of missing values. There are 3 major categories of missing values.

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

Imputation Techniques

- Imputation with a constant value
- Imputation using the statistics (mean, median, mode)
- K-Nearest Neighbor Imputation.

Ans 12. *A/B testing,* also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

Ans 13. Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

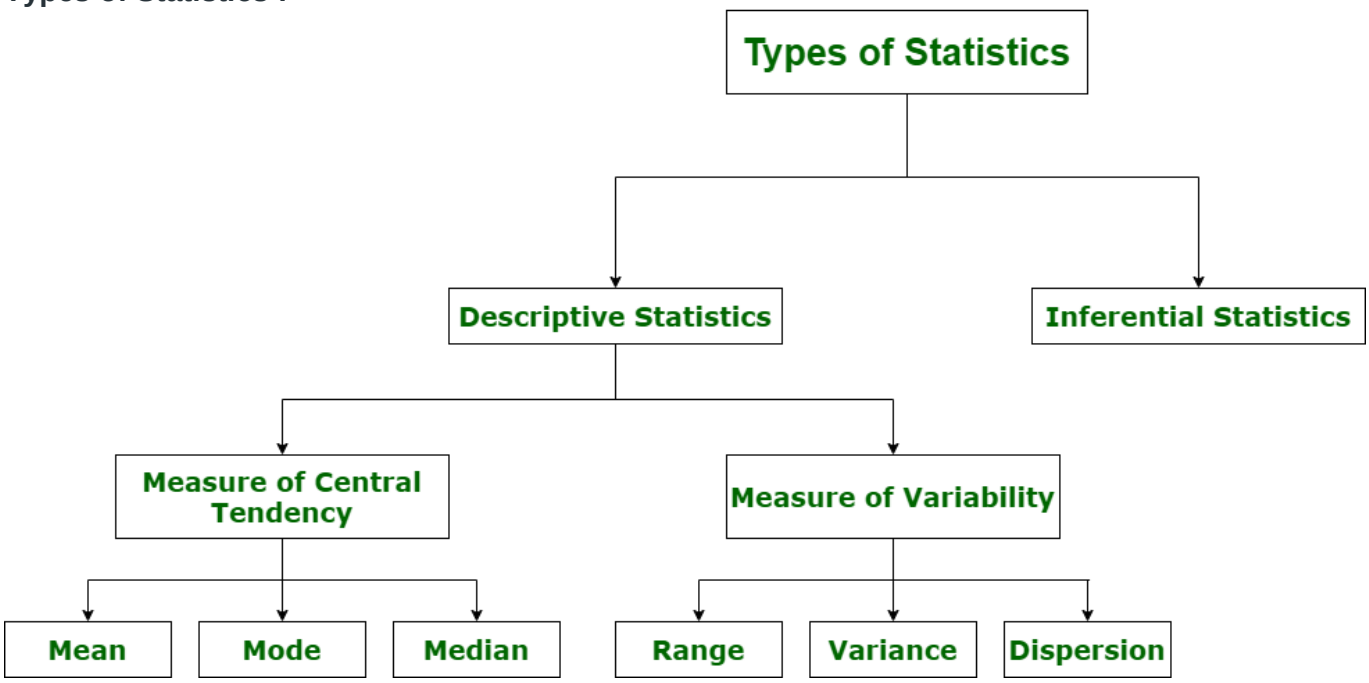
Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Ans 14. 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).

Ans 15. Statistics simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data. It is actually a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life.

Types of Statistics :



1. **Descriptive Statistics :**
   Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

(a). Measure of central tendency –
Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.
In statistics, there are three common measures of central tendency as shown below:

(i) Mean :
It is measure of average of all value in a sample set.
For example,

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5

$$\text{Mean (m)} = \frac{\text{Sum of all the terms}}{\text{Total no. of terms}}$$

$$m = \frac{21.3 + 20.8 + 19}{3}$$

$$= 20.366$$

(ii) Median :
It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.
For example,

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5
i 20	15	4

Ordering the set from lowest to highest = 15 19 20.8 21.3

$$\text{Median} = \frac{19 + 20.8}{2}$$

Median = 23.5

(iii) Mode :
It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.
For example,

2 3 4 2 4 6 4 7 7 4 2 4

Mode = 4

(b). Measure of Variability –
Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population. In statistics, there are three common measures of variability as shown below:

(i) Range :
It is given measure of how to spread apart values in sample set or data set.

Range = Maximum value - Minimum value

(ii) Variance :
It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.

$$\mathbf{S^2=} \sum^{n}_{i=1} [(x_i - \overline{x})^2 \div n]$$

In these formula, **n** represent total data points, $\overline{\mathbf{x}}$ represent mean of data points and $\mathbf{x_i}$ represent individual data points.

- **(iii) Dispersion :**

It is measure of dispersion of set of data from its mean.

$$\mathbf{\sigma=} \sqrt{} (1 \div n) \sum^{n}_{i=1} (x_i - \mu)^2$$

2. Inferential Statistics :

Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive stats. It is simply used to analyze, interpret result, and draw conclusion. Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis.
Hypothesis testing is a type of inferential procedure that takes help of sample data to evaluate and assess credibility of a hypothesis about a population. Inferential statistics are generally used to determine how strong relationship is within sample. But it is very difficult to obtain a population list and draw a random sample.

Types of inferential statistics –
Various types of inferential statistics are used widely nowadays and are very easy to interpret. These are given below:

- One sample test of difference/One sample hypothesis test
- Confidence Interval
- Contingency Tables and Chi-Square Statistic
- T-test or Anova
- Pearson Correlation
- Bi-variate Regression
- Multi-variate Regression

MACHINE LEARNING -WORKSHEET 1

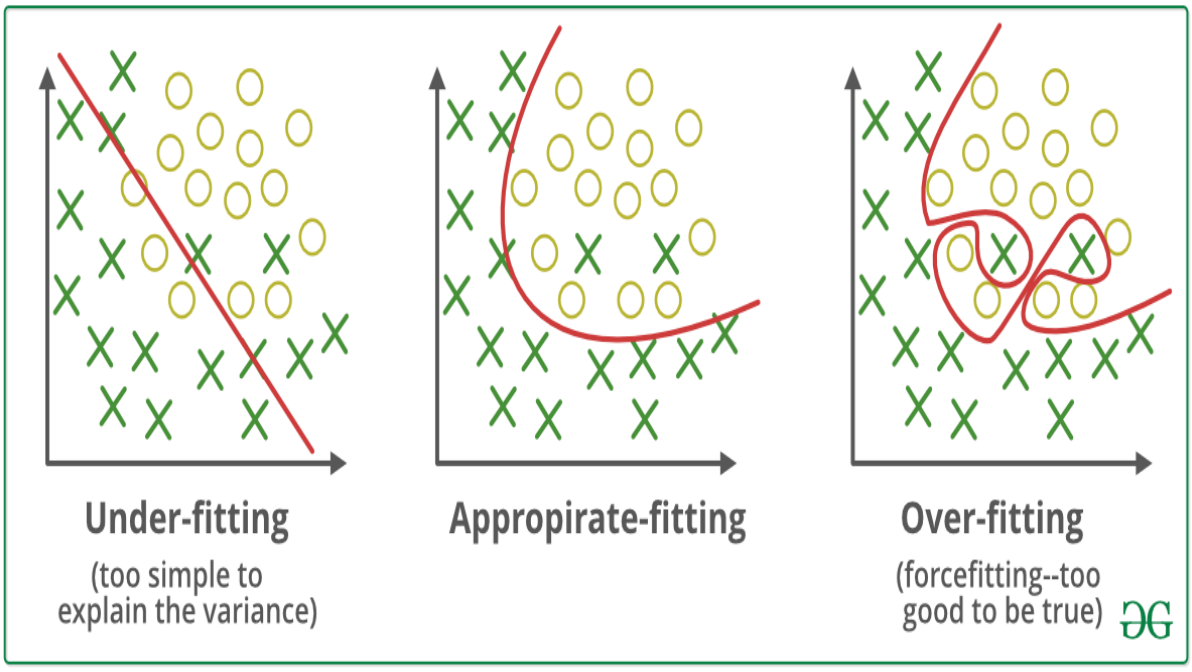
Answers

1. A
2. A
3. B
4. B
5. C
6. B
7. D
8. D
9. A
10. B
11. B
12. A,B

Ans 13. Regularization is a technique used in regression to reduce the complexity of the model and to shrink the coefficients of the independent features.

In simple words Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it. Sometimes the machine learning model performs well with the training data but does not perform well with the test data.

Overfitting is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data.



Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
The commonly used regularization techniques are :

1. L1 regularization
2. L2 regularization
3. Dropout regularization

A regression model which uses **L1 Regularization** technique is called **LASSO(Least Absolute Shrinkage and Selection Operator)** regression.

A regression model that uses **L2 regularization** technique is called **Ridge regression**.
Lasso Regression adds *"absolute value of magnitude"* of coefficient as penalty term to the loss function(L).

Ans 14. Regularization algorithms

- Ridge Regression
- LASSO (Least Absolute Shrinkage and Selection Operator) Regression
- Elastic-Net Regression

Ridge Regression

Ridge regression is a method for analyzing data that suffer from multi-collinearity. Ridge regression shrinks the coefficients as it helps to reduce the model complexity and **multi-collinearity.**

LASSO is a regression analysis method that performs both feature selection and regularization in order to enhance the prediction accuracy of the model.LASSO regression adds a penalty **(L1 penalty*)*** to the loss function that is equivalent to the magnitude of the coefficients.

In LASSO regression, the penalty has the effect of forcing some of the coefficient estimates to be **exactly equal to zero** when the regularization parameter λ is sufficiently large.*LASSO regression is also known as the* L1 Regularization (L1 penalty).

Elastic-Net Regression

Elastic-Net is a regularized regression method that linearly combines the L1 and L2 penalties of the LASSO and Ridge methods respectively.

A standard least-squares model tends to have some variance in it i.e. the model won't generalize well for a data set different than its training data. ***Regularization, significantly reduces the variance of the model, without a substantial increase in its bias***.

Ans15. The error term is the stuff that isn't explained by the model.

For a very simple example, suppose you are predicting the weight of adult human males based on their height. Well, height is certainly related to weight - taller people tend to be heavier - but the model won't be perfect because there is a range of weights at each height. The error is the **difference between the predicted value and the actual value.**

$$y=a+bx+e$$

The algorithm solves for the straight line function y=a+bx, where

y is the outcome

x is a predictor

b is the effect coefficient for x

a is the y-intercept (the predicted value of y when x=0).

e is the error

Given y and x, the algorithm solves for a and b, which are the slope and y-intercept of the straight line that produces the smallest average deviation from the points (x,y).