Page No.
Date

Assignment No. 1

Q.1 a) What is descriptive statistics ? list out the types of descriptive statistics.

→ * Descriptive statistics :→

Descriptive statistic describes the characteristics or properties of the data. It helps to summarize the data in a meaningful data in a meaningful way. It allows important patterns to emerge from the data.

Data summarization techniques are used to identify the properties of data. It is helpful in understanding the distribution of data. They do not involve in generalizing beyond the data.

* Types of descriptive statistics :

There are two types of descriptive statistics. They are :

a) measures of central Tendency :

This includes the mean, median and mode.

b) measures of data spread or dispersion :

It includes range, quartiles, variance and standard deviation.

Page No.
Date

Q.1 b) Write in brief :

① mean :→

Mean is the most commonly used measure of central tendency. It actually represent the average of the given collection data. It applied for both continuous and discrete data.

It is equal to the sum of all the values in the collection of data, divided by the total no. of values

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

OR

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

③ median :→

Median represents the mid-value of the given set of data when arranged in a particular order.

The median for grouped data can be calculated using the formula,

$$\text{median} = l + \left(\frac{n/2 - c.f}{f} \right) \times h$$

③ mode :→

The mode is the most frequent score in our data set. The mode is used for categorical data where we want to know which is the most common category occuring in the population.

$$\text{mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] \times h$$

Q.2 a) What is inferential statistics ? Explain main two areas of inferential statistics.

→ * Inferential statistics :→

In inferential statistics, predictions are made by taking any group of data in which you are interested. It can be defined as random sample of data taken from a population to describe and make inferences about population.

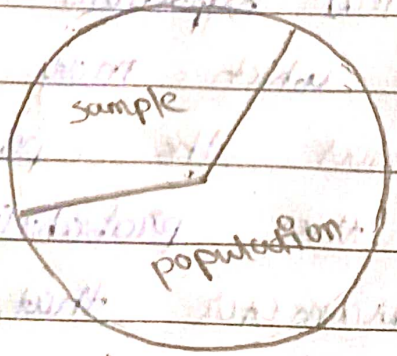

Fig: Inferential statistics

Date

* There are two main areas of inferential statistics :

① Estimating parameters : →

This means taking a statistic from the sample data and using it to infer about a population parameter. A good statistical estimator should have the unbiased, consistent and accuracy characteristics.

② Hypothesis tests : →

This is where sample data can be used to answer research questions. For ex. we might be interested in knowing if a new cancer drug is effective or if breakfast helps children perform better in school.

Q.2 b) Explain in detail about the statistical hypothesis.

→ ① A statistical hypothesis is defined as a statement, which may or may not be true about the population parameters or about the probability distribution of the parameter that we wish to

validate on the basis of sample information.
(2) most times, experiments are performed with random samples instead of the entire population. (3) But before drawing inferences about the population it should be always kept in mind that the observed results might have come due to chance factor.
(4) In order to have an accuracy or more precise inference, the chance factor should be ruled out.

Q.3a) Explain in detail data manipulation.
→ (1) It involves 'manipulation' or 'manipulating' data using available set of variables.
(2) This is done to enhance accuracy and association with data.
(3) Actually, the data collection process can have many loopholes.
(4) There are various uncontrollable factor which lead to inaccuracy in data such as mental situation of respondents, personal biases, difference / error in reading of machines etc.

Page No.
Date

(5) To lessen these inaccuracies, data manipulation is done to increase the possible accuracy in data.

(6) This stage is also known as data wrangling or data cleaning.

Q.3 b) Explain different ways to manipulate data.

→ * Different ways to manipulate / treat data are as followed: →

(1) manipulating data using inbuilt base R function: →

This is the first step, but is often repetitive and time consuming. Hence, it is less efficient way to solve the problem.

(2) use of packages for data manipulation:

CRAN has more than 8000 packages available today. These packages are a collection of pre written commonly used pieces of codes. They helps to perform the repetitive task fast, reduce errors in coding.

(3) use of ml algo. for data manipulation: →
ml algo. like tree based boosting algorithms to take care of missing data and outliers. These algorithms are less time consuming.

Q.4 a) Explain the following packages:
i) dplyr package: →
This package is created and maintained by Hadley Wickham. This package has everything to accelerate data manipulation efforts. It is known best for data exploration and transformation. It include 5 main manipulation commands:
(1) filter - It filters the data based on a condition.

(2) select - It is used to select col of interest from a data set.

(3) arrange - It is used to arrange data set values on ascending or descending order.

Page No.
Date

(4) mutate - It is used to create new variables from existing variables

(5) Summarise - It is used to perform analysis by commonly used operations such as min, max & etc.

iii) data.table package :→

This package allows to perform faster manipulation in datasets. A data table has 3 parts namely DT [i, j, by]. We can tell R to subset the rows using 'i', to calculate 'j' which is grouped by 'by'. Most of the time, 'by' relates to categorical variables.

Syntax:

DT [where, select/update/do, by]

Page No.
Date

Q.4b) Explain the following with limitation and advantages :

i) Scatter plot :→

A scatter plot is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

Limitation:

① With scatter diagram we cannot get the exact extent of correlation.

② Quantitative measure of the relationship betn the variable cannot be viewed. Only show quantitative expression.

③ The relationship can only show for two variables.

Advantages :

① Relationship betn two variable can be viewed.

② plotting the diagram is very simple.

③ observation and reading is easy to understand.

④ For non-linear pattern, this is the best method.

ii) Histogram: →

A histogram represent the frequency distribution of continuous variables. It presents numerical data and as drawn in such way that there is no gap betn bars.

Limitations:

A histogram can present data that is misleading as it has many bars.

only two sets of data are used, but to analyse certain types of statistical data, more than two sets of data are necessary.

Advantages:

It helps to identify different data, the frequency of the data occuring in the dataset and categories which are difficult to interpret in a tabular form.

It helps to visualize the distribution of data.

Page No.
Date

Q.5 a) What is data type? List out the types with example.

→ A datatype describes the characteristic of a variable. It defines what type of data we are going to store in variable.

There are six standard data types:

① Number: →

In numbers, there are mainly 3 types which includes int, float and complex

Example:

a = 5 → integer number

b = 2.5 → float number

c = 6+2j → complex number

② String: →

A string is an ordered sequence of character. We can use single or double quotes to represent string.

Example:

str = "Welcome"

Page No.
Date

③ List :→

A list can contain a series of values. list variables are declared by using brackets [ ]. A list is mutable, which means we can modify the list.

Example:

list = [2,3,4,5.5,'Hi']

④ Tuple :→

A tuple is a sequence of python objects separated by commas. Tuples are immutable, which means tuples once created cannot be modified. They are defined using parentheses ( ).

Example:

Tuple = (50,25.6, "python")

⑤ Set :→

A set is an unordered collection of items. set is defined by values separated by a comma inside braces { }

Example:

set = {5,1,2.6,"python"}

⑥ Dictionary:

Dictionaries items are stored and fetched by using key. They are used to store huge amount of data. Dictionaries are defined within braces { }

Syntax :→ key : value

Example:

Dict = { 1 : 'Hi' , 2 : 7.5 , 3 : 'class' }

Q.5 b) Enumerate the list and its method with example.

→ ① List is a collection which is ordered and changeable. It also allows duplicate members.

② It is defined by using brackets [ ].

③ Methods:

a) append() :→

used for appending and adding elements to the end of the list

example:

```
months = ['Jan', 'Feb', 'Mar']
months.append('Apr')
print(months)
```

b) sort():→

Sort a list in ascending, descending or user-defined order.

ex.

```
prices = [23.8, 287.81, 238.91]
prices.sort()
print(prices)
```

c) Extend():

Adds each element of iterable to the end of list.

Example:

```
x = [1,2,3]
x.extend([4,5])
print(x)
```

d) Index():

Returns the lowest index where the element appears.

example:

```
months = ['Jan', 'Feb', 'Mar']
months.index('Feb')
```

e) max():

The max() function will return the highest value of list.

Example:

```
prices = [159.34, 37.13, 209.89]
price_max = max(prices)
print(price_max)
```

f) min():

Calculates the minimum of all the element of list.

```
prices = [159.34, 37.13, 209.89]
price_min = min(prices)
print(price_min).
```

g) len():

len() function shows the no. of element in a list. Example: stocks = [75.14, 89.16, 90.1]

```
print(len(stocks)).
```

h) pop():

This method removes the item at specified index.

Ex. prime_num = [2,3,5,7]

```
prime_num.pop(2)
print(prime_num)
```

i) clear():

This method removes all the element from a list

Ex. list.clear().

Page No.
Date

Q.6 a) Elucidate the string and its methods with ex.

→ ① String represents a sequence of characters.
Ex. str = "Data"

② String can be created using single or double quotes.

③ In python, string can be accessed by using the method of indexing
ex. str1 = "Sahir"
print (str1[0])

④ To access a range of characters in string, the method of slicing is used. It is done by using colon.
ex. str1 = "Sahir"
print (str1[1:5]

⑤ String in python can be formatted with the use of format() method which is a very versatile & powerful tool for formatting.
Ex.
str1 = "{}{}{}{}".format("Cricks", "Urban", "bad")
print (str1)

Page No.
Date

Q.6b) What is dictionary ? Explain methods available in dictionary.

→ (1) Dictionaries are used to store data values in key-value pairs.

(2) A dictionary is a collection which is ordered, changeable and do not allow duplicates.

(3) Creating dictionary : →

With many key-value pairs surrounded in curly brackets and a colon separating each key.

Syntax : Dict = {"Name" : "Sahil" , "Roll No" : 140 }

(4) Accessing the dictionary values : →

The key of dict can be used to obtain the values because they are unique from one another.

Ex. Employee = {"Name" : "Dev" , "Age" : 20 }

print ("Name : %s" %. Employee ["Name"])

print ("Age : %d" %. Employee ["Age"])

(5) Deleting element using pop() method is one of the ways to get rid of element from dict.

ex. Dict = {1: "Hello" , 2: "Hey" , 3 : "Hii" }

pop_key = Dict .pop(2)

print (Dict)

(6) Iterating dictionary can be iterated using for loop. Example :

Employee = {"Name" : "Sahil" "Age" : 21 }

for x in Employee :

print (x)

(7) Delete keyword can be used to in-place delete the key that is present in the dictionary in python.

dict = {"Arushi": 22, "mani": 21}

del dict["mani"]