Descriptive Statistics

Objectives :

This chapter will provide different methods to summarize the given data.

Central Tendency: Central tendency is the measure of average and includes mean, median, mode, and midrange. For example, the average American man is five feet, nine inches tall; the average women is five feet, 3.6 inches.

Measure of Variation: The central tendency gives average of data. However, knowing only average of data is not enough to describe the entire data. Even though, the average man's show size is 10, the shoe store owner cannot make her business viable if she orderes only size 10 shoes. For this one most know the spread of data. This include variance, standard deviation, and range.

Measure of Position: To know where the specific data lies within data set or its relative position with respect to other data we need measure of position. This include decile, quartile, and percentile.

The codes are written in Python programming language. You can refer to "https://Getting Started with Python" to get understanding of the language.

Central Tendency

Mean

Also known as arithmetic average, the mean is the ratio of the sum of all the given data to the total number of the data.

For example: The number of records broken in the first six days of the Olympics games are 3, 5, 7, 2, 9, and 1. Calculate the mean of the broken records on each day.

Data (X)= 3, 5, 7, 2, 9, 1

Total number of data (N) = 6

Sum of the data (ΣX)= 3 + 5 + 7 + 2 + 9 + 1 = 27

sample mean $(\bar{x}) = \frac{\Sigma X}{N}$ = $\frac{27}{6}$ = 4.5

Let's calculate mean using Python.

To run the code below click on play button or press "Shift + Enter" on keyboard.

```
# Python Code for to calculate the above question
data1 = [3, 5, 7, 2, 9, 12, 15, 1] # Given data
total_sum = sum(data1) # Calculating sum of the given number
number = len(data1) # Counting total number of the data
mean = total_sum/number # Formula of the data
print ("The mean of record broken per day is ", mean) # Displaying the mean
```

```
The mean of record broken per day is  4.5
```

In the above code, we used Python in-built function "sum" and "len" to calculate the sum of the given numbers and to count the total numbers repectively in the "data" variable. The formula to calculate the mean we used formula as given in the definition, ratio of total sum to the total number.

Python ignores the characters written after "#". It is used to add comment in the code.

Using Python built in funtion to calculate mean.

```
1 import statistics # Importing module
2 data2 = [3, 5, 7, 2, 9, 12, 25, 1]
3 arithmetic_mean = statistics.mean(data2) # Using in-built function of the module
4 print ("The mean of record broken per day is", arithmetic_mean) # Displaying the result
```

```
The mean of record broken per day is 8
```

In the above code, we imported 'statitics' from Python's standard library. It has built-in function to calculate basics of statistics. In the 2nd line of code, we called built-in 'mean' funtion in the statistics module. After that, 'data' is used as an argument of the 'mean' function. The calculation is stored in the 'arithmetic_mean' variable.

Using the examples above, calculate mean of the following questions.

Question 1

The number of calls that a local police department responded to for a sample of 9 months is shown. Find the mean. (The data were hypothetical). 475, 447, 440, 761, 993, 1051, 783, 676, and 620.

```
# write your code here
# Import required module from Python's Standard Library
# Call the funtion and store in a variable
# Display the result
# Correct answer: 694
```

Question 2

The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the mean. 110, 76, 29, 38, 105, and 31

```
# write your code here
# Import required module from Python's Standard Library
# Call the funtion and store in a variable
# Display the result
# Correct answer: 64
```

Mean of Groups

The students in a class are carrying money which is divided into following classes.

Calculate the mean money in the class per students.

Class Interval	Frequency
0 - 10	3
10 - 20	5
20 - 30	3
30 - 40	9
40 - 50	6
50 - 60	3
60 - 70	5
70 - 80	3
80 - 90	4
90 - 100	2

To calculate mean, first we need to calculate midpoint of all the money group. The formula to calculate midpoint is:

$$mid\ point = \frac{lower\ point + higher\ point}{2}$$

The fomula to calculte total money is:

$$total\ sum = \Sigma(\ midpoint\ *\ number\ of\ students)$$

$$total\ students = \Sigma\ frequency$$

Finally,

$$\bar{x} = \frac{total\ sum}{total\ students}$$

```
# importing libraries.
import pandas as pd

# Using dictionary
data3 = { 'lower_limit': [0, 10, 20, 30, 40, 50, 60, 70, 80, 90],
        'upper_limit': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100],
         'frequency': [3, 5, 3, 9, 6, 3, 5, 3, 4, 2]
}
data_table = pd.DataFrame(data3) # converting out data in table

print(data_table) # Displaying the table
```

```
   lower_limit  upper_limit  frequency
0            0           10          3
1           10           20          5
2           20           30          3
3           30           40          9
4           40           50          6
5           50           60          3
6           60           70          5
7           70           80          3
8           80           90          4
9           90          100          2
```

The above code uses 'pandas' library. I used the pandas to built table for easy viualization.

```
# calculating midpoint
data_table['midpoint'] = (data_table['lower_limit'] +  data_table['upper_limit'])/2
print (data_table)
```

```
   lower_limit  upper_limit  frequency  midpoint
0            0           10          3       5.0
1           10           20          5      15.0
2           20           30          3      25.0
3           30           40          9      35.0
4           40           50          6      45.0
5           50           60          3      55.0
6           60           70          5      65.0
7           70           80          3      75.0
8           80           90          4      85.0
9           90          100          2      95.0
```

```
# calculating mean
total_summed = (data_table['midpoint'] * data_table['frequency']).sum() # sum is the bui
total_number = data_table['frequency'].sum()
money_mean = total_summed/total_number # mean formula
print("The students have mean money of", round(money_mean, 2)) # round function to displ
```

```
The students have mean money of 46.4
```

Question 3

The data represent the number of miles run during one week for a sample of 20 runners.

Class Interval	Frequency
5.5 - 10.5	1
10.5 - 15.5	2
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2
Total	20

```
# Write your code here
# create dictionary, unique dictionary name
# create dataframe table
# display the table, to check if the code working
```

```
# Write your code here
# create new column 'midpoint' and use formula to fill the data in it
# display the table, to check if the code working
```

```
# Write your code here
# use formula to calculate mean
# display your result
```

Median

Median is the midpoint of the given data which divides the data into two equal parts.

For example: from the above first example, the data is: 3, 5, 7, 2, 9, and 1.

Arranging the data in ascending order: 1, 2, 3, 5, 7, and 9

If the total number of data is odd, the median is the middle data value.

IF the total number of data is even, the median is the midpoint of the middle two data values.

In the given data it is even (6) so the median is the midpoint of the two middle data values: 3 and 5.

Thus, $median = \frac{3+5}{2} = 4$

Using Python

Let's use Python to calculate median of the data.

```
# importing module
import statistics
data4 = [3,45, 5, 7, 2, 9, 1] # creating list for the given data
median_value = statistics.median(data4) # calculating median
print("The median value of the data is", median_value) # displaying the result
```

```
The median value of the data is 5
```

In the above example, Python automatically arrange the data into ascending order. After arranging, it calculates the median value. Regardless of the number of data being even or odd, it will calculate median value in just 4 line of codes.

Question 4

The number of calls that a local police department responded to for a sample of 9 months is shown. Find the median. (The data were hypothetical). 475, 447, 440, 761, 993, 1051, 783, 676, and 620.

```
# import module
# create list for the given data, assign a unique variable
# call the built-in median function, from the module
# display the result
# Hint answer: 676
```

Question 5

The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the median. 110, 76, 29, 38, 105, and 31

```
# import module
# create list for the given data, assign a unique variable
# call the built-in median function, from the module
# display the result
# Hint answer: 57
```

Mode

Mode is the value in the data set which occurs the most. Mode can be of different types:

unimodal: If only one data has highest occurance.

bimodal: If two data has same number of occurance.

multimodal: If more than two data has same number of occurance.

no mode: If all the data set occurs only once. It cannot be called as 'mode is zero', because temperature has '0' as value.

Example:

```
# importing module
import statistics
data5 = [2, 3, 9, 4, 5, 6] #data set
mode_value = statistics.multimode(data5) #Use the funtion 'multimode' if there are more
print('The mode of the given data set is', mode_value)
```

```
The mode of the given data set is [2, 3, 9, 4, 5, 6]
```

Question 6

Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dolloars are: 18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

```
# import module
# create list for the given data
# call the built-in mode function, from the module
# display the result
# Hint answer: 10
```

Midrange

Midrange is the midpoint of the lowest and the highest value in the dataset and is denoted by MR. It gives a rough estimation of the average.

$$MR = \frac{lowest\ value + highest\ value}{2}$$

For example: The number of bank failures for a recent five-year period is shown. Find the midrange.

3, 30, 148, 157, 71

```
# Calculating midrange is simple and straight forward.
# There is no built-in midrange function
data6 = [3, 30, 148, 157, 71] # creating list with dataset
highest = max(data6) # assigns maximum value from the list to the 'highest' variable
lowest = min(data6) # assigns minimum value from the list to the 'lowest' variable
MR = (lowest + highest)/2
print('Midrange is', MR)
```

```
Midrange is 80.0
```

Question 7

Find the midrange of data for the NFL signing bonuses. The bonuses in millions of dollars are: 18.0, 14.0,34.5, 10.0, 11.3, 10.0, 12.4, 10.00

```
# Write your code here
# create list for the given data set
# find the lowest value
# find the highest value
# calculate the midrange
# print the result
```

Double-click (or enter) to edit