Descriptive Statistics

Contents

Measurement Scales

Data

Respondent	Gender	Region	Age	Satisfaction Level
1	M	1	23	3
2	M	2	45	4
3	M	2	33	3
4	F	2	25	4
5	F	3	37	2
6	M	1	35	1
7	M	2	41	5
8	F	3	27	2

Description

Region	1	Mumbai
	2	Delhi
	3	Kolkata
Satisfaction Level	1	Highly dissatisfied
	2	dissatisfied
	3	Neutral
	4	Satisfied
	5	Highly satisfied

Gender: Nominal
Region: Nominal
Age: Ratio
Satisfaction Level: Ordinal

Measures of Central Tendency

It is a single value Most commonly used measures of central tendency are :

Mean	Arithmetic Mean. Commonly known as Average.
	It is the sum of all values of the variable divided by the total number of values.
Median	Arrange the data in ascending order, Median is the middle value, if N is odd.
	If N is even, it is average of two middle values.
Mode	It is the most frequently occurring observation in a set of data.

Note : The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

Trimmed Mean

It is recommended to report 'Trimmed Mean' along with mean if outliers are present in the data.

Trimmed mean excludes extreme data points for the calculation of mean. Typically, 5% data points (5% at each end) are excluded.

Note that trimmed mean will give robust estimate if underlying distribution is symmetric.

Get an Edge!

Best Measure of Central Tendency

Type of Variable	Best Measure
Nominal	Mode
Ordinal	Median
Interval/Ratio (Symmetric)	Mean
Interval/Ratio (Not Symmetric)	Median

- Mean is appropriate when the distribution is symmetric. For symmetric distribution, the mean is at the centre.
- For a skewed (not symmetric) distribution, mean is generally not at the centre. Median is better measure of central tendency for a skewed distribution.

Measures of Variation

In addition to a measure of central tendency, it is desirable to have a measure of dispersion (variation) of data.

Measure of Dispersion :

- A measure of dispersion is an indication of the spread of measurements around the center of the distribution.

- Two data sets can have equal mean (measure of central tendency) but vastly different variability.
- Eg. Score of Batsman A = (78,62,73,54,76,77) & Score of Batsman B = (92,8,78,34,109,99)

  So Average scores of two batsmen in 6 innings is equal(=70) whereas Spread around mean is not identical.

Most commonly used measures of variation are :

- Range
- Inter-Quartile Range (IQR)

Coefficient of Variation (CV)

As variance has same units as that of the variable, it is inappropriate to use variance to compare two data sets having different units. Hence, there is a need of a quantity without unit like Coefficient of Variation (CV) for effective comparison.

CV is a relative measure of variation and is used to compare variability in two data sets.

The CV is defined as "Standard Deviation divided by Mean" and is generally expressed as a percentage.

Higher the value of CV, more is the variability.
CV is sometimes referred to as "Relative Standard Deviation".

Case Study - 1

Objective

- To compare the performance of two batsmen using the measures of central tendency and measure of variation

Available Information

- Runs scored by two batsman A and B in 6 matches

Runs Scored

Batsman A	Batsman B
78	92
62	8
73	78
54	34
76	109
77	99

Observation and Conclusion

Batsman A	Batsman B
78	92
62	8
73	78
54	34
76	109
77	99
MEAN = 70	MEAN = 70
CV = 13.97%	CV = 57.32%

- Average scores of two batsmen in 6 innings is equal(=70) but the spread around mean is not identical.

- We can see that variability in performance of Batsman B is more than that of Batsman A. Hence, we can infer that Batsman A is a more consistent performer than Batsman B.

Case Study - 2

To learn Descriptive Statistics in Python, we shall consider the below case as an example.

Background

Data of 100 retailers in platinum segment of FMCG companies.

Objective

To describe the variables present in the data

Sample Size

Sample size: 100
Variables: Retailer, Zone, Retailer_Age, Perindex, Growth, NPS_Category

Data Snapshot

Retail Data

Variables

Retailer	Zone	Retailer_Age	Perindex	Growth	NPS_Category
1	North	<=2	81.84	3.04	Promoter

Observations

Columns	Description	Type	Measurement	Possible values
Retailer	Retailer ID	numeric	-	-
Zone	Location of the retailer	character	East, West, North, South	4
Retailer_Age	Number of years doing business with the company	character	<=2, 2 to 5, >5	3
Perindex	Index of performance based on sales, buying frequency and buying recency	numeric	-	positive values
Growth	Annual sales growth	numeric	-	positive values
NPS_Category	Category indicating loyalty with the	character	Detractor, Passive,	3

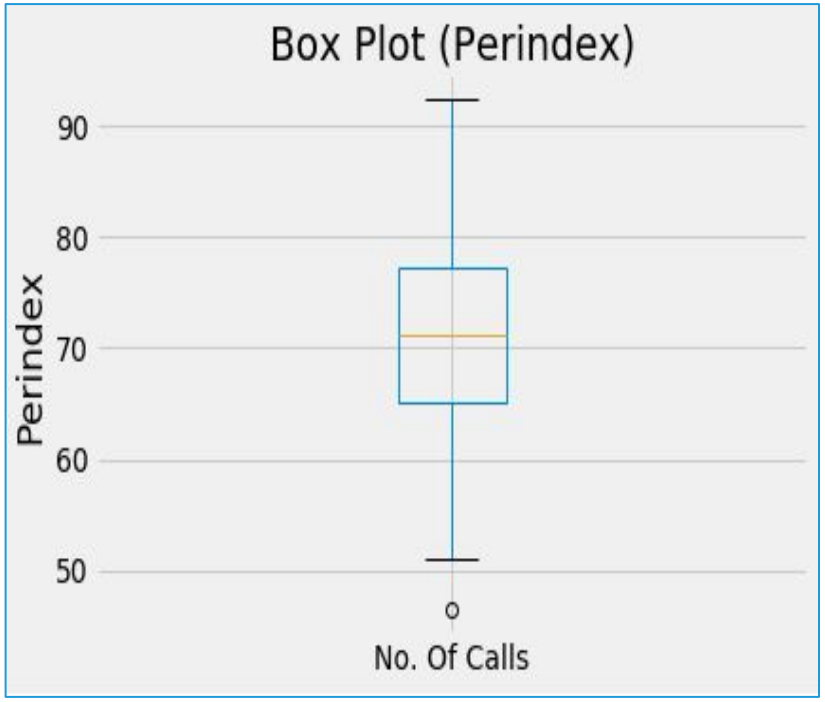
Describing Variables in Python

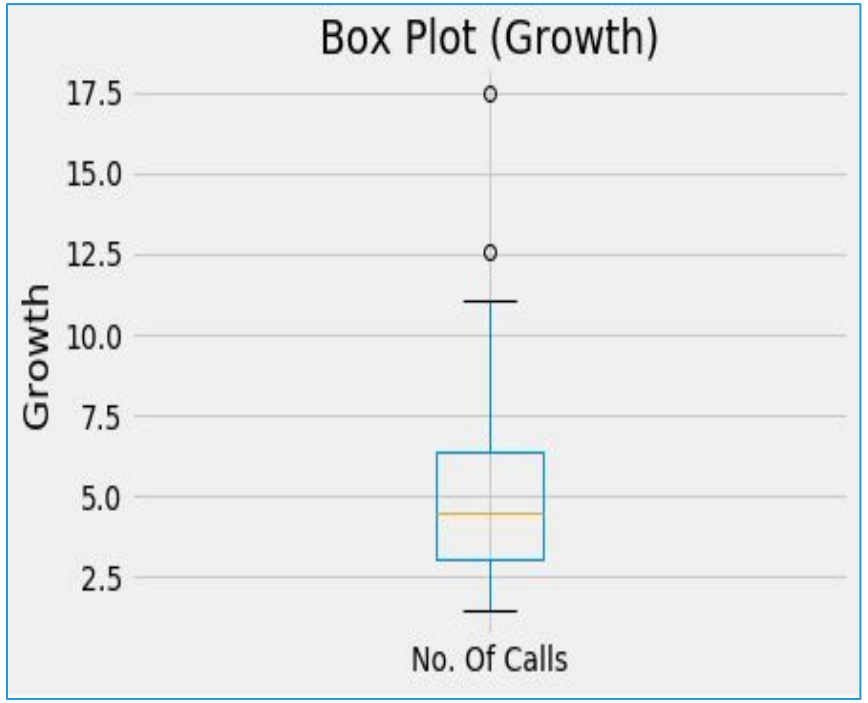
```
#Importing Data
import pandas as pd
retail_data =pd.read_csv('Retail_Data.csv')

#Checking the variable features using summary function
retail_data.describe(include = 'all')

# Output
```

	Retailer	Zone	Retailer_Age	Perindex	Growth	NPS_Category
min	1	NaN	NaN	46.53	1.47	NaN
25%	25.75	NaN	NaN	65.08	3.0575	NaN
std	29.01149	NaN	NaN	9.569232	2.620525	NaN
mean	50.5	NaN	NaN	70.49697	5.1528	NaN
50%	50.5	NaN	NaN	71.15	4.495	NaN
75%	75.25	NaN	NaN	77.175	6.34	NaN
count	100	100	100	99	100	100
max	100	NaN	NaN	92.49	17.5	NaN
unique	NaN	4	3	NaN	NaN	3
top	NaN	South	>5	NaN	NaN	Passive
freq	NaN	32	56	NaN	NaN	41

Understanding Data through Visualisation

```
from matplotlib import pyplot as plt
retail_data.Perindex.plot.box(label='No. Of Calls');plt.title('Box
Plot (Perindex)');plt.ylabel('Perindex')
retail_data.Growth.plot.box(label='No. Of Calls');plt.title('Box Plot
(Growth)');plt.ylabel('Growth')
```





Here we can see that Perindex variable is distributed symmetrically whereas Growth variable is Positively Skewed.

The concept of Skewness is explained in detail in next presentation
How to plot a box plot is explained in detail is Data Visualisation Module

Measures of Central Tendency in Python

```
# Mean for Perindex & Growth Variables
retail_data.Perindex.mean()
```

```
70.4969696969697
```

mean() in Python, gives mean of the variable. It excludes NAs by default

```
retail_data.Growth.mean()
```

```
5.152800000000002
```

```
# Median for Perindex & Growth Variables
retail_data.Perindex.median()
```

```
71.15
```

median() in Python, gives median of the variable.

```
retail_data.Growth.median()
```

```
4.495
```

So as we have seen, Perindex Variable is symmetric, hence it's mean value is appropriate whereas for Growth Variable which is Positively Skewed, Median would be a better measure.

Measures of Central Tendency in Python

```
# Import stats from scipy library
from scipy import stats
```

scipy is a python library used for advanced scientific operations.
stats includes statistical operations

```
# Trimmed Mean
trimmed_mean_PI = stats.trim_mean(retail_data['Perindex'], 0.1)
trimmed_mean_PI
```

```
70.76162500000001
```

Using 0.1 in the **trim_mean()**, excludes 10% observations from each side of the data from the mean

```
trimmed_mean_G = stats.trim_mean(retail_data['Growth'], 0.1)
trimmed_mean_G
```

```
4.825
```

```
# Mode
retail_data.Perindex.mode()
```

```
67.71
68.00
```

- In Python we can find mode directly by using **mode()** function.
- Here 67.71 and 68.00 has equal highest frequencies, hence 2 modes.

Measures of Central Tendency in Python

```
# Measure of Central Tendency for Categorical Variable
# Mode using Frequency Table

freq = retail_data['Zone'].value_counts()
freq

 # Output
```

South	32
West	28
North	25
East	15

value_counts() in Python, gives the frequency of counts of the variable mentioned.

Here Mode is 32 as the frequency is highest for South Zone.

Measures of Dispersion in Python

```
# Standard Deviation
retail_data['Perindex'].std()
```

```
9.56923188593669
```

std() in Python, gives standard deviation of the variable

```
# Variance
retail_data['Perindex'].var()
```

```
91.57019888682746
```

var() in Python, gives variance of the variable

```
# Coefficient of Variation
cv_PI = retail_data['Perindex'].std()/ retail_data.Perindex.mean()
cv_PI
```

```
0.135739620115161
```

We calculate CV manually by definition.

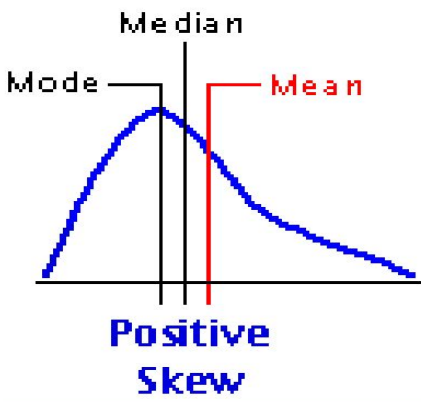
Skewness

Skewness gives us the Shape of the data. It is the 'Lack of Symmetry'

Positively Skewed

- Right Tail is longer
- Mass of the distribution is concentrated on the left

Mode<Median<Mean

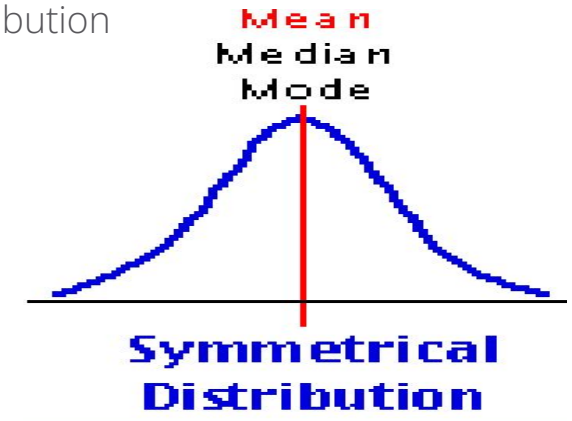


Symmetric

- Both tails are equal
- Mass of the distribution is equally distributed

Mean=Median=Mode

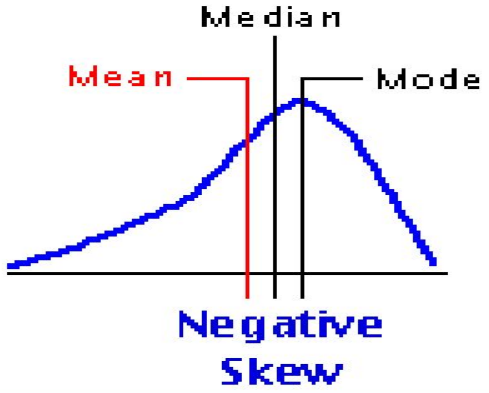
Normal Distribution is symmetric distribution



Negatively Skewed

- Left Tail is longer
- Mass of the distribution is concentrated on the right

Mean<Median<Mode



Kurtosis

Kurtosis is defined as a measure of 'peakedness'. It is generally measured relative to Normal distribution. (Which means 'excess of kurtosis' is measured)

Mesokurtic

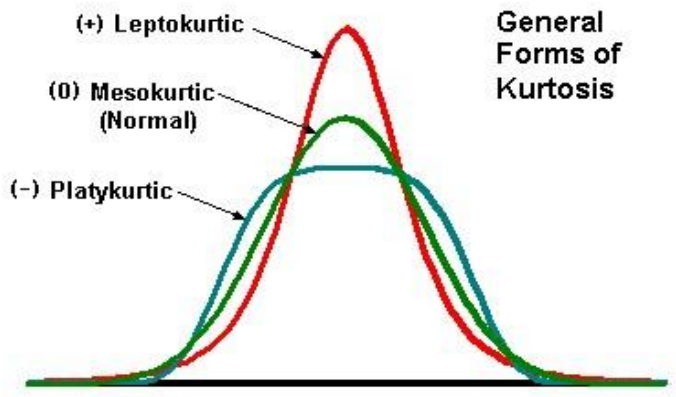
Normal distribution is termed as mesokurtic distribution

Leptokurtic

A leptokurtic distribution has a more acute peak. (positive kurtosis)

Platykurtic

A platykurtic distribution has a flatter peak. (negative kurtosis)



Skewness and Kurtosis in Python

```
#Importing Data
import pandas as pd
retail_data =pd.read_csv('Retail_Data.csv')
```

We have already seen that Growth variable is Positively Skewed, so we'll find out skewness & kurtosis value for the same

```
from scipy import stats
```

Using package **"scipy"** in Python is the easiest way to find skewness and kurtosis

```
# Skewness
retail_data['Growth'].skew()
```

```
1.5912357812381297
```

- **skew()** gives skewness of the variable.

```
# Kurtosis
retail_data['Growth'].kurtosis()
```

```
4.283885801046328
```

- **kurtosis()** gives kurtosis of the variable.

Quick Recap

In this session, we covered descriptive statistics using Python

Measures of Central Tendency/Variation	• Mean,Median,Mode • Standard Deviation, CV
Measures of Skewness and Kurtosis	• Skewness and Kurtosis
Working in Python	• Python codes for descriptive statistics