Application Development Laboratory (CS 33002)

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

School of Computer Engineering

2 Credit

Analytics Application Development using R

Lab Contents

Sr #	Major and Detailed Coverage Area	Lab#
1	Descriptive Analytics	6

Descriptive Statistics

All the data which is gathered for any analysis is useful when it is properly represented so that it is easily understandable by everyone and helps in proper decision making. After carrying out data analysis, we describe its summary so as to understand it in a much better way. This is known as **summarizing** the data.

We can summarize the data in several ways either by text manner or by pictorial representation. We can summarize our data in R as follows:

1. **Descriptive Statistics** – With the help of descriptive statistics (also called summary statistics), we can represent the information about our datasets. They also form the platform for carrying out complex computations as well as analysis. Therefore, even though they are developed with simple methods, they play a crucial role in the process of analysis.
2. **Tabulation** – Representing the data analyzed in tabular form for easy understanding.
3. **Graphical** – It is a way to represent data graphically.

Why Descriptive Statistics?

To answer:

1. How widely dispersed is the data?
2. Are there lot of different values?
3. Are the many of the values same?
4. What value is in the middle of the data?
5. Where does a particular data value stand with respect with other values in the data set?

Descriptive Statistics cont…

Choosing which descriptive statistics are appropriate depend on the type of data being examined. Different statistics should be used for qualitative and quantitative data.

Quantitative and Qualitative Data

1. Quantitative data are measures of values or counts and are expressed as numbers. Such data are data about **numeric** variables (e.g. how many; how much; or how often).

2. Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code. Such data are data about **categorical** variables (e.g. what type).

Quantitative = Quantity

Qualitative = Quality

Quantitative and Qualitative Data

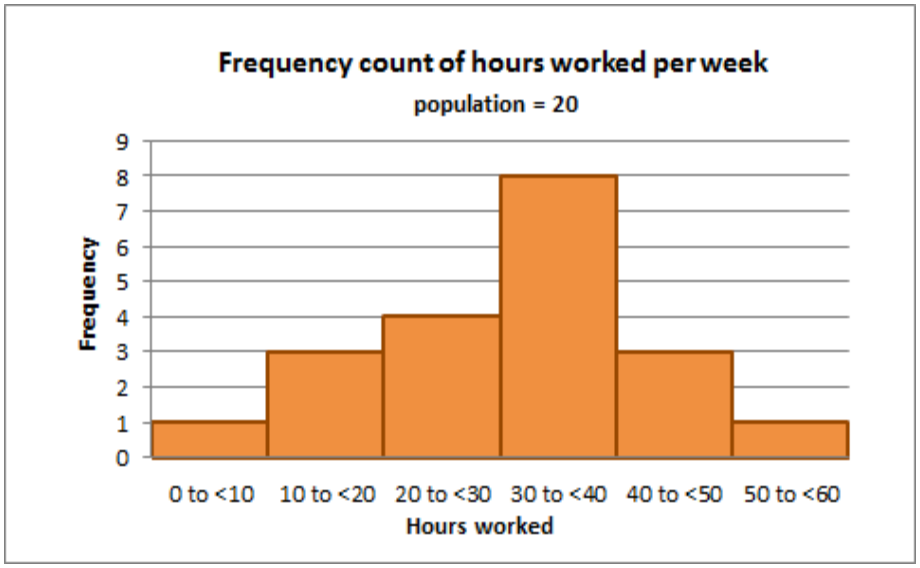
Data collected about a **numeric variable** will always be **quantitative** and data collected about a **categorical variable** will always be **qualitative**. Therefore, you can identify the type of data, prior to collection, based on whether the variable is numeric or categorical.

Example

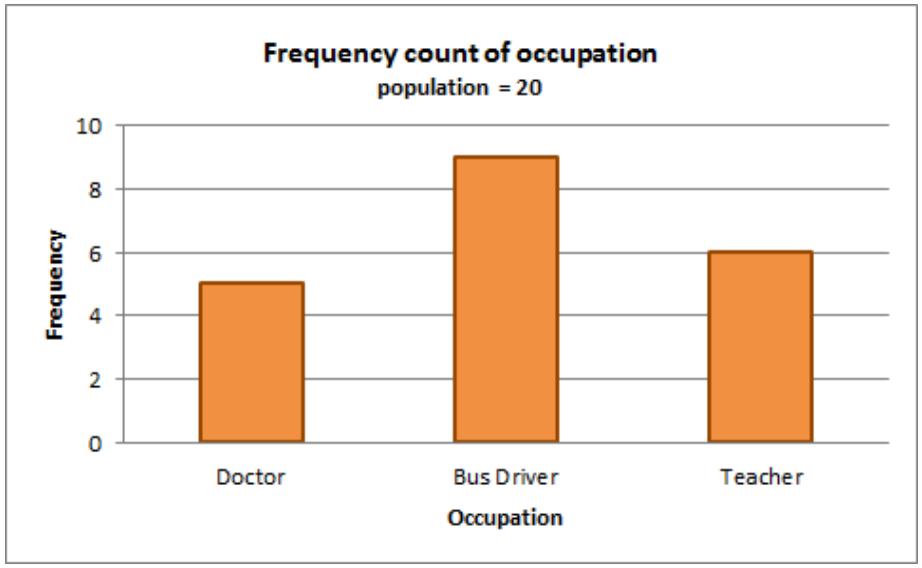
Data unit	Numeric variable	= Quantitative data	Categorical variable	= Qualitative data
A person	"*How many* children do you have?"	**4** children	"In *which country* were your children born?"	**Australia**
	"*How much* do you earn?"	**$60,000** p.a.	"*What* is your occupation?"	**Photographer**
	"How many hours do you work?"	**38** hours per week	"Do you work *full-time or part-time*?"	**Full-time**
A house	"*How many* square metres is the house?"	**200** square metres	"In *which city or town* is the house located?"	**Brisbane**
A business	"*How many* workers are currently employed?"	**264** employees	"*What* is the *industry* of the business?"	**Retail**
A farm	"*How many* milk cows are located on the farm?	**36** cows	"*What* is the main *activity* of the farm?"	**Dairy**

How can you use quantitative and qualitative data?

Quantitative Data



Qualitative Data



Sessional Exercise

Create a data frame with the following info:

id, name, salary, start.date, dept

1,Rick,623.3,2012-01-01,IT

2,Dan,515.2,2013-09-23,Operations

3,Michelle,611,2014-11-15,IT

4,Ryan,729,2014-05-11,HR

5,Gary,843.25,2015-03-27,Finance

6,Nina,578,2013-05-21,IT

7,Simon,632.8,2013-07-30,Operations

8,Guru,722.5,2014-06-17,Finance

Exercise

1. Examine the quantitative data and draw the frequency count.
2. Examine the qualitative data and draw the frequency count.

Descriptive Statistics cont…

Statistics that describe or summaries can be produced for quantitative data and to a lesser extent for qualitative data.

- As quantitative data are always numeric they can be ordered, added together, and the frequency of an observation can be counted. Therefore, all descriptive statistics can be calculated using quantitative data.

- As qualitative data represent individual (mutually exclusive) categories, the descriptive statistics that can be calculated are limited, as many of these techniques require numeric values which can be logically ordered from lowest to highest and which express a count.

Types of Descriptive Statistics

There are four major types of descriptive statistics:

- **Measures of Frequency**
  - *Count, Percent, Frequency*
  - Shows how often something occurs
  - Use this when you want to show how often a response is given.
- **Measures of Central Tendency**
  - *Mean, Median, and Mode*
  - Locates the distribution by various points
  - Use this when to show how an average or most commonly indicated response
- **Measures of Dispersion or Variation**
  - *Range, Variance, Standard Deviation*
  - Identifies the spread of scores by stating intervals
  - Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean
- **Measures of Position**
  - *Percentile Ranks, Quartile Ranks*
  - Describes how scores fall in relation to one another. Relies on standardized scores
  - Use this when you need to compare scores to a normalized score (e.g., a norm)

Measures of Frequency Exercise

CSE-1	
Grade	# of students
O	0
E	3
A	3
B	5
C	8
D	29
F	5

IT-4	
Grade	# of students
O	1
E	10
A	19
B	24
C	17
D	7
F	4

CSSE-1	
Grade	# of students
O	0
E	3
A	6
B	18
C	28
D	8
F	7

Lab Exercise

1. Calculate frequency of each grade for individual sections and at overall level.
2. Calculate percent of each grade for individual section and at overall level.
3. Calculate the total number of students for each section and at overall level.

Frequency = no. of students who received that score

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

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.

Mean

It is calculated by taking the sum of the values and dividing with the number of values in a data series. The function mean() is used to calculate this.

Example: The mean of 4, 1, and 7 is (4+1+7)/3=12/3=4

Syntax: mean(x, trim = 0, na.rm = FALSE, ...) where x is the input vector, trim is used to drop some observations from both end of the sorted vector, and na.rm is used to remove the missing values from the input vector.

Illustration 1:

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)

# Find Mean.
print(mean(x))
```

Illustration 2 (Applying Trim Option): When trim parameter is supplied, the values in the vector get sorted and then the required numbers of observations are dropped from calculating the mean. When trim = 0.3, 3 values from each end will be dropped from the calculations to find mean. In this case the sorted vector is (−21, −5, 2, 3, 4.2, 7, 8, 12, 18, 54) and the values removed from the vector for calculating mean are (−21,−5,2) from left and (12,18,54) from right.

print(mean(x, trim = 0.3))

Illustration 3 (Applying NA Option): If there are missing values, then the mean function returns NA. To drop the missing values from the calculation use na.rm = TRUE. which means remove the NA values. **x <- c(12,7,3,4.2,18,2,54,-21,8,-5,NA) print(mean(x,na.rm = TRUE))**

Median

The middle most value in a data series is called the median. The median() function is used to calculate this value.

Syntax:

median(x, na.rm = FALSE), where x is the input vector, and na.rm is used to remove the missing values from the input vector.

Illustration:

```
# Create the vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)

# Find the median.
median.result <- median(x)
print(median.result)
```

Sessional Exercise

Create a vector with starting value 1, end value 1000, with incremental step 2. Find the median.

Mode

The mode is the value that has highest number of occurrences in a set of data. Unlike mean and median, mode can have both numeric and character data. R does not have a standard in-built function to calculate mode. So we create a user function to calculate mode of a data set in R. This function takes the vector as input and gives the mode value as output.

Illustration:

```
# Create the function.
getMode <- function(v){
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
# Create the vector with numbers.
v <- c(2,1,2,3,1,2,3,4,1,5,5,3,2,3)
print(getMode (v))
```

```
# Create the vector with characters.
charv <- c("o","it","the","it","it")
print(getMode(charv))
```

Measures of Dispersion

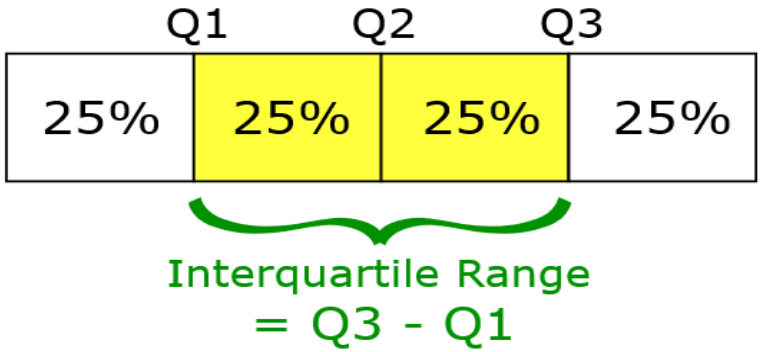
As the name suggests, the measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

There are four frequently used measures of variability:

- Range
- Interquartile range
- Variance
- Standard deviation

Range

- The range is the difference between the lowest and highest values. Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the range is 9 – 3 = 6. It is that simple!
- The range can sometimes be misleading when there are extremely high or low values. Example: In {8, 11, 5, 9, 7, 6, 3616}: the lowest value is 5, and the highest is 3616, So the range is 3616 – 5 = 3611. The single value of 3616 makes the range large, but most values are around 10.
- So we may be better off using **Interquartile Range**.

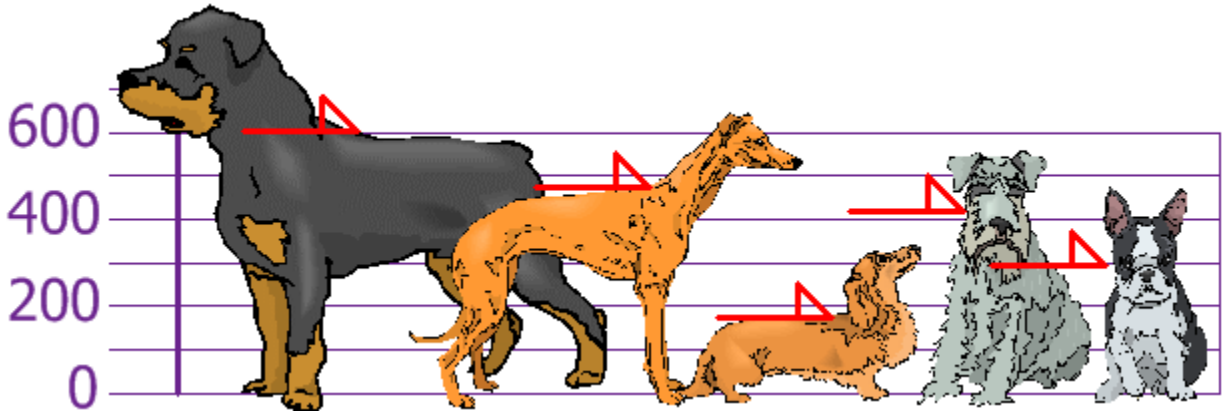


Variance

Variance measures how far a data set is spread out. It is mathematically defined as the average of the squared differences from the mean.

Illustration:

The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

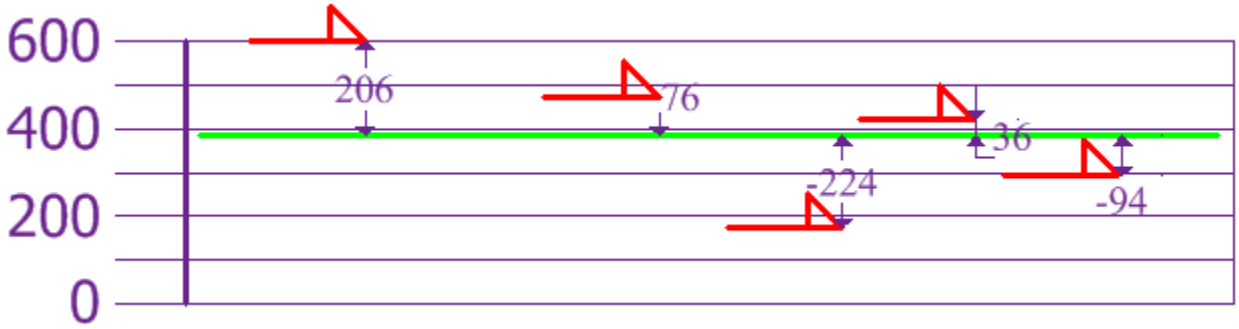


Mean = (600 + 470 + 170 + 430 + 300) / 5 = 394

so the mean (average) height is 394 mm.

Now we calculate each dog's difference from the Mean.

Variance cont…



To calculate the variance, take each difference, square it, and then average the result.

Variance = $\sigma^2 = (206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2) / 5 = 21704$

Use the **var()** function to calculate the variance. But which variance does it give? The one with N in the denominator or the one with N-1.

heights <- c(50, 47, 52, 46, 45)

It calculates the estimated variance (with N−1 in the denominator). To calculate that first variance with N in the denominator, one have to multiply this number by (N−1)/N. Using length() to calculate N, that's

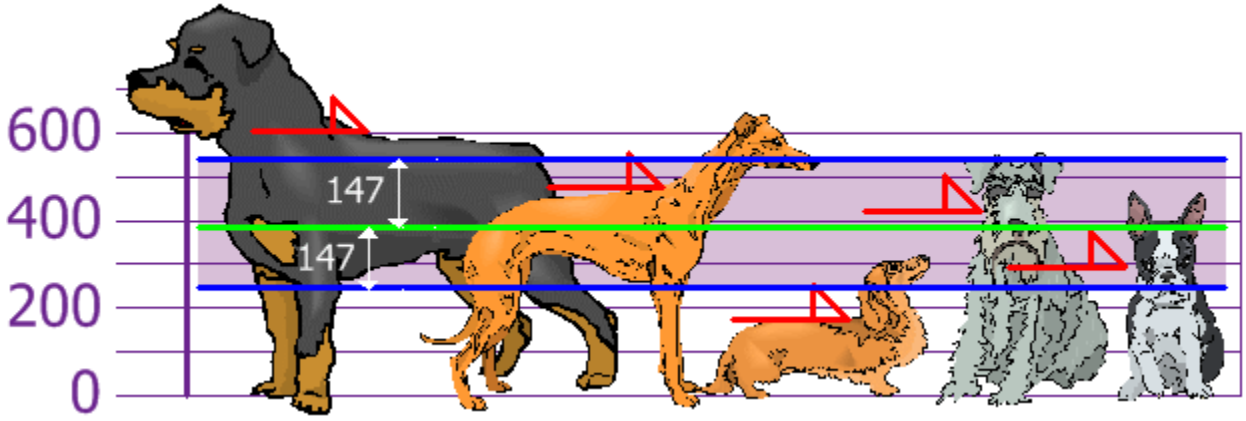
var(heights)*(length(heights)-1)/length(heights)

Standard Deviation

Standard Deviation is just the square root of Variance.

Standard Deviation $\sigma = \sqrt{21704} = 147.32... = 147$ (to the nearest mm)

The good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean.



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small. Rottweilers are tall dogs. Dachshunds are a bit short ...

```
heights <- c(50, 47, 52, 46, 45)
print(sd(heights))
```

Lab Exercise

Consider the following dataset

Student	Grade	Age	Sex	Height	Calories eaten	Attitude	Class rank
A	5	10	M	137	2000	5	4
B	5	11	M	140	1500	4	2
C	4	9	F	120	1200	3	5
D	4	10	F	140	1400	4	1
F	5	10	M	135	1600	4	8
G	4	9	M	130	1200	4	3
H	4	10	F	140	1800	3	7

Exercise

1. Create the data frame for the above dataset
2. Calculate mean, median , mode, variance and standard variation of quantitative data
3. Draw suitable visualization illustrating number of females and males in the dataset
4. Calculate max, min, range of quantitative data

Measures of Position

A measure of position is a method by which the position that a particular data value has within a given data set can be identified.

There are three frequently used measures of position:

- Standard Score
- Quartile
- The Five Number Summary

Standard Score

The standard score (often called the **z-score**) of a particular data value gives the position of that data value by determining its distance from the mean, in units of standard deviation.

The formula for the z-score = (data point – mean) / standard deviation

Illustration:

Suppose the score on AP lab exam averaged 78%, with a standard deviation of 5 percentage points. If Tom earned 69% on that exam, then his z-score:

$$\text{z-score} = (69 - 78) / 5 = -9 / 5 = -1.8$$

So Tom scored 1.8 standard deviations below the mean score.

Quartile / Percentile

Find the first and third quartiles of the set {4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11}.

Put them in order: 3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

Cut it into halves: 3, 4, 4 4, 7, 10 | 11, 12, 14, 16, 17, 18

In this case all the quartiles are between numbers:

Quartile 1 (Q1) = (4+4)/2 = 4

Quartile 2 (Q2) = (10+11)/2 = 10.5

Quartile 3 (Q3) = (14+16)/2 = 15

So,

25th Percentile (or Q1): 11.5

50th Percentile (or Q2) : 17

75th Percentile (or Q3): 20

The Five Number Summary

A five-number summary is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: **minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), maximum value**.

These values have been selected to give a summary of a data set because each value describes a specific part of a data set: the median identifies the centre of a data set; the upper and lower quartiles span the middle half of a data set; and the highest and lowest observations provide additional information about the actual dispersion of the data. This makes the five-number summary a useful measure of spread. A five-number summary can be represented in a diagram known as a box and whisker plot.

Exercise

Find five number summary of the set {1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57}.

Organizing data: observations and variables

In general, collected raw data is organized according to observations and variables. Variables represent a single measurement or characteristic for each observation.

Instructor	Date	Student	Rating
Bob Belcher	2015-01-01	a	4
Bob Belcher	2015-01-01	b	5
Bob Belcher	2015-01-01	c	4
Bob Belcher	2015-01-01	d	6
Bob Belcher	2015-02-05	e	6
Bob Belcher	2015-02-05	f	6
Bob Belcher	2015-02-05	g	10
Bob Belcher	2015-02-05	h	6
Linda Belcher	2015-01-01	a	8
Linda Belcher	2015-01-01	b	6
Linda Belcher	2015-01-01	c	8
Linda Belcher	2015-01-01	d	8
Linda Belcher	2015-02-05	e	8
Linda Belcher	2015-02-05	f	7
Linda Belcher	2015-02-05	g	10
Linda Belcher	2015-02-05	h	9

Each row represents an observation. So each observation contains the rating by a single student on a single date for an instructor. The variables are Instructor, Date, Student, and Rating.

Types of variables

The most common variables used in data analysis can be classified as one of four types of variables: nominal, ordinal, interval, and ratio. Understanding the differences in these types of variables is critical, since the variable type will determine which statistical analysis will be valid for that data. In addition, the way we summarize data with statistics and plots will be determined by the variable type.

- **Nominal :** A variable with a nominal scale characterised with categories that do not have a natural order or ranking. You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless. Example: Gender: Male, Female, Other. Hair Color: Brown, Black, Blonde, Red, Other.
- **Ordinal :** A variable with an ordinal scale is one where the order matters but not the difference between values. Example: High school class ranking: 1st, 9th, 87th… Socioeconomic status: poor, middle class, rich.
- **Interval scale:** A variable with an interval scale has values of equal intervals that mean something. Example: Celsius Temperature, Fahrenheit Temperature, SAT scores (200-800), credit score (300-850).
- **Ratio scale:** A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable. Example: a weight of zero doesn't exist; an age of zero doesn't exist. On the other hand, temperature is not a ratio scale, because zero exists (i.e. zero on the Celsius scale is just the freezing point; it doesn't mean that water ceases to exist).

Data and Distribution

DA Score

Score	Frequency
60	3
65	10
70	12
75	15
80	20
85	15
90	12
95	10
100	3

Frequency = no. of students who received that score

Plot Histogram

Data is normally distributed. Normal distribution curve is a bell-shaped frequency distribution curve. The bell curve gets its name quite simply because its shape resembles that of a bell. Most of the data values tend to cluster around the mean. Right and the left of the distribution are perfect mirror images of one another.

Data and Distribution cont...

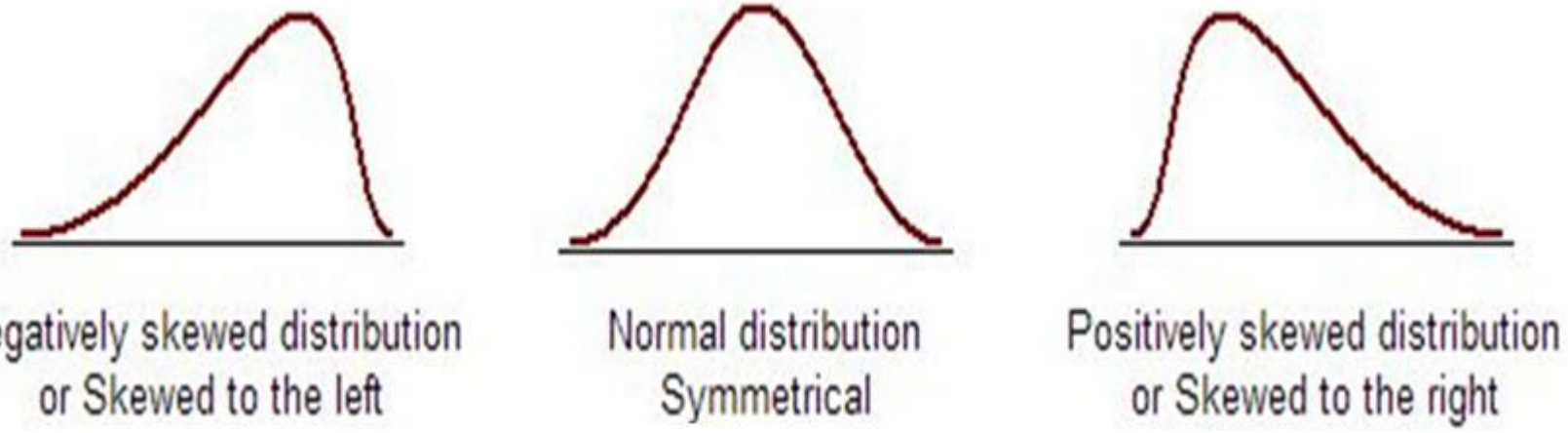
Size of Items	Frequency
2-4	5
4-6	12
6-8	27
8-10	10
10-12	8
12-14	3
14-16	1

Size of Items	Frequency
10-15	2
15-20	5
20-25	12
25-30	18
30-35	30
35-40	21
40-45	6

Plot Histogram

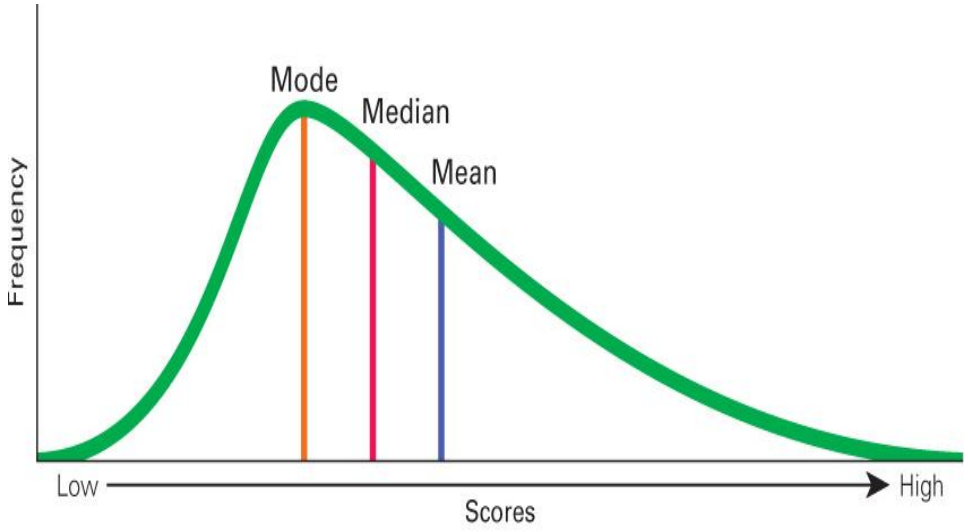
Skewness

- Not every distribution of data is symmetric. Sets of data that are not symmetric are said to be asymmetric. The measure of how asymmetric a distribution can be is called skewness.
- Skewness is a measure that refers to the extent of symmetry or asymmetry in a distribution.
- Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right.

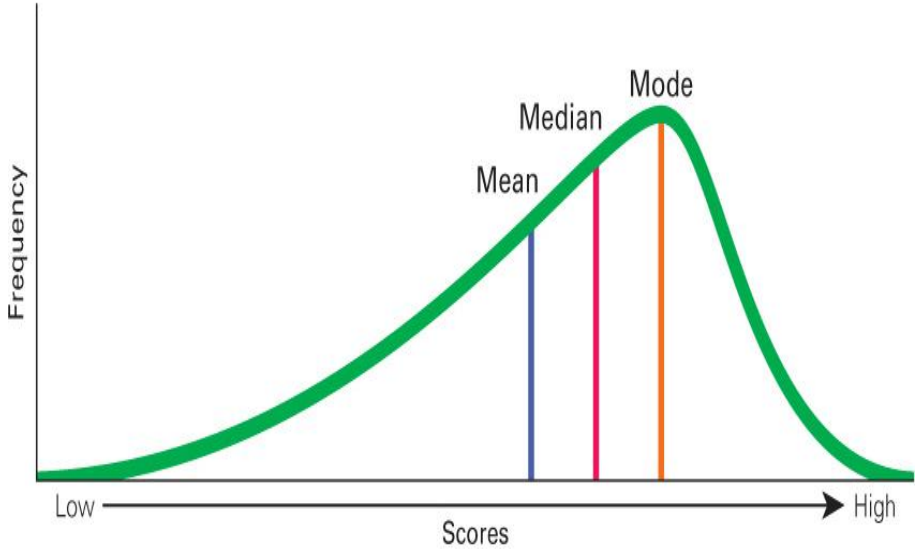


Skewness Types

- Skewed to the Right: Data that are skewed to the right have a long tail that extends to the right. In this situation, the mean and the median are both greater than the mode.
- Skewed to the Left: Data that are skewed to the left have a long tail that extends to the left. In this situation, the mean and the median are both less than the mode.



Positively Skew



Negatively Skew

Measures of Skewness

The measures of skewness can be both absolute and relative.

- Absolute Measure: It tells the extent of asymmetry and whether it is positive or negative.

  Symbolically: absolute skewness = **mean – mode** or **mean – median**

  If the value is > 0, skewness is positive else negative.

- Relative Measure: In order to make comparison between the skewness in two or more distributions, coefficient of skewness is computed for the given series or distribution. The formula used for measuring skewness is

$$\text{skewness} = SK_p = \frac{\textbf{mean – mode}}{\textbf{standard deviation}}$$

This is called **Karl Pearson's coefficient of skewness**

Illustration:

Section A: mean = 46.83, SD (standard deviation) = 14.8, and mode = 51.67

Section B: mean = 47.83, SD = 14.8, and mode = 47.07

Section A: SK_p = -0.327 and Section B: SK_p = 0.051

So, distribution of marks in section A is more skewed (relative). The skewness of Section A is negative , while that of B is positive.

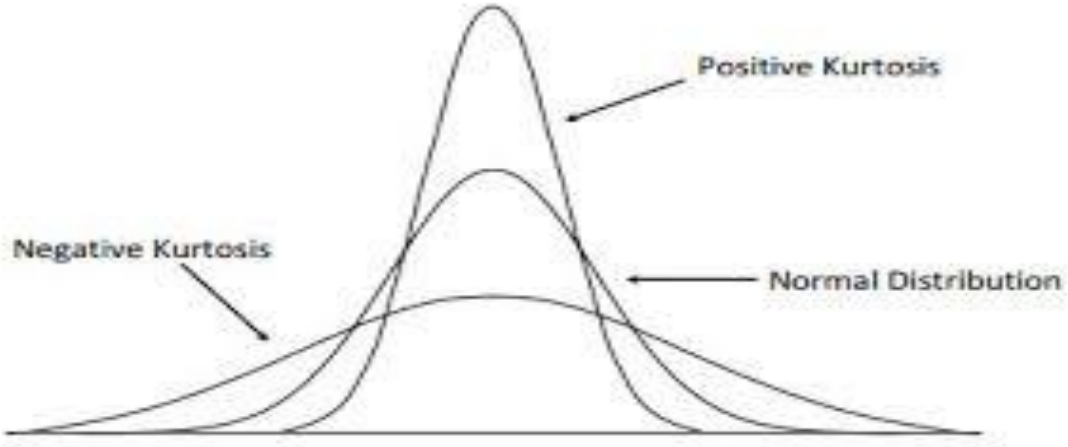
Measures of Skewness cont…

When is the skewness too much? The rule of thumb seems to be:

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed
- If the skewness is less than -1 or greater than 1, the data are highly skewed

Kurtosis

Along with skewness, kurtosis is an important descriptive statistic of data distribution. Kurtosis measures the degree to which the distribution has either fewer and less extreme outliers, or more and more extreme outliers than the normal distribution. In nutshell, skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails. In general, the higher the kurtosis, the sharper the peak and the longer the tails. This is called leptokurtic, and is indicated by positive kurtosis values. The opposite—platykurtosis—has negative kurtosis values.



Kurtosis cont...

Score	Frequency
75	3
80	3
85	3
90	4
95	27
100	32
105	27
110	4
115	3
120	3
125	3

Score	Frequency
75	19
80	20
85	27
90	26
95	27
100	27
105	27
110	26
115	27
120	22
125	21

Plot Histogram

Kurtosis cont…

- **High kurtosis** in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!
- **Low kurtosis** in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis(too good to be true), then also we need to investigate and trim the dataset of unwanted results.

Kurtosis Calculation

$$S_{kr} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{S^4}$$

Using the data from the example above (12 13 54 56 25), determine kurtosis.

$$X = \frac{(12 + 13 + \ldots + 25)}{5} = \frac{160}{5} = 32$$

$$S^2 = \frac{(12-32)^2 + \ldots + (25-32)^2}{4} = 467.5$$

$$\textit{Therefore,} \qquad S = 467.5^{\frac{1}{2}} = 21.62$$

$$S_{kr} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{S^4} = \frac{1}{5}\frac{-20^4 + (-19^4) + 22^4 + 24^4 + (-7^4)}{21.62^4} = 0.7861$$

Types of Kurtosis

The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values, as well as values close to zero.

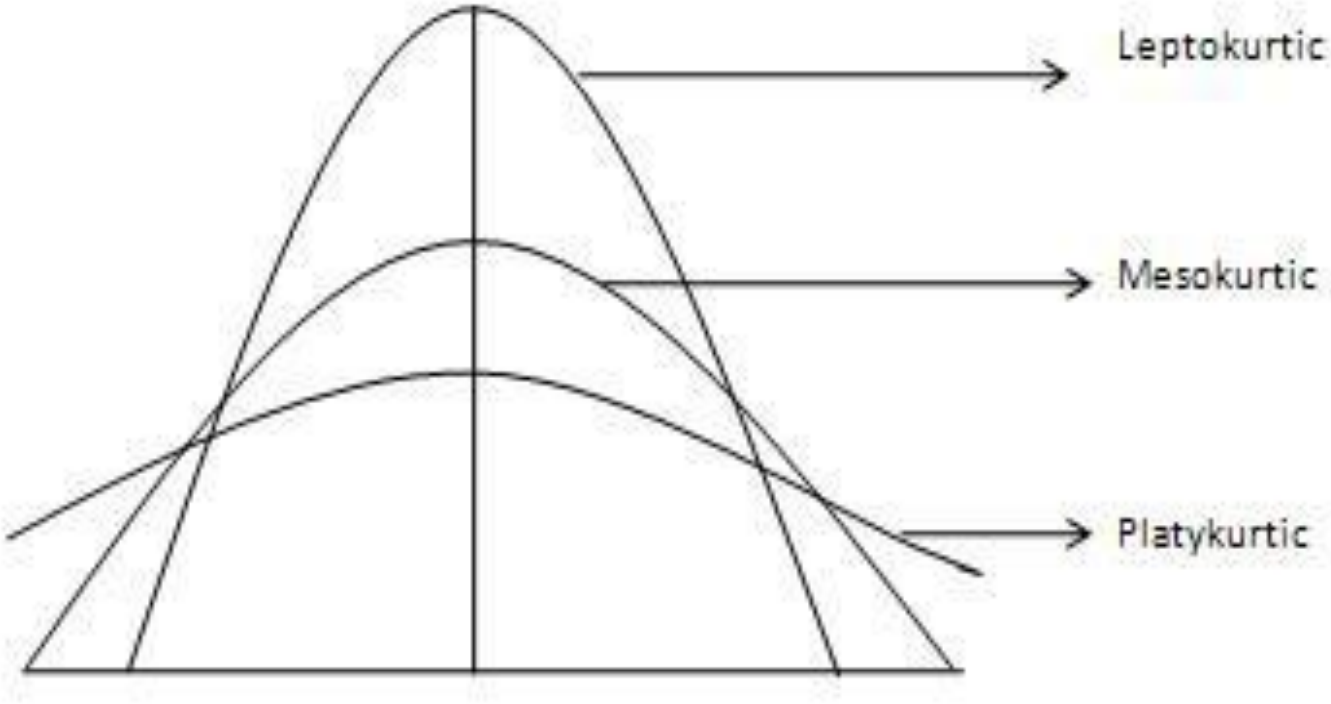
Excess Kurtosis = Kurtosis – 3

1. **Mesokurtic:** Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. It means that if the data follows a normal distribution, it follows a mesokurtic distribution.

2. **Leptokurtic**: Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating the large outliers. In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.

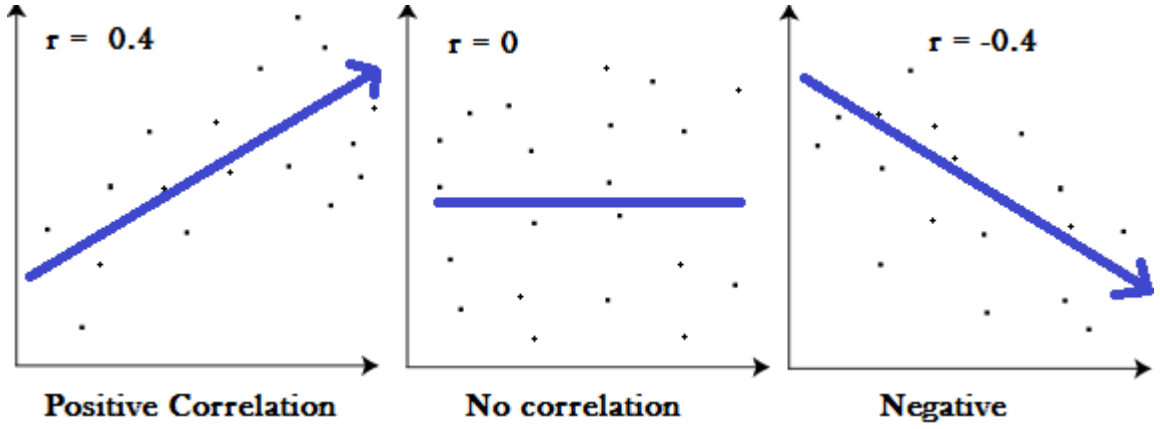
3. **Platykurtic:** A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution. In the finance context, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.

Types of Kurtosis cont...



Correlation

Correlation is a technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight. Nonetheless, the average weight of people 5'5'' is less than the average weight of people 5'6'', and their average weight is less than that of people 5'7'', etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights. The main result of a correlation is called the **correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.



Correlation cont…

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

Example

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days.

Ice Cream Sales vs. Temperature												
Temp	14.2	16.4	11.9	15.2	18.5	22.1	19.4	25.1	23.4	18.1	22.6	17.2
Sales	215	325	185	332	406	522	412	614	544	421	445	408

Draw a scatter plot

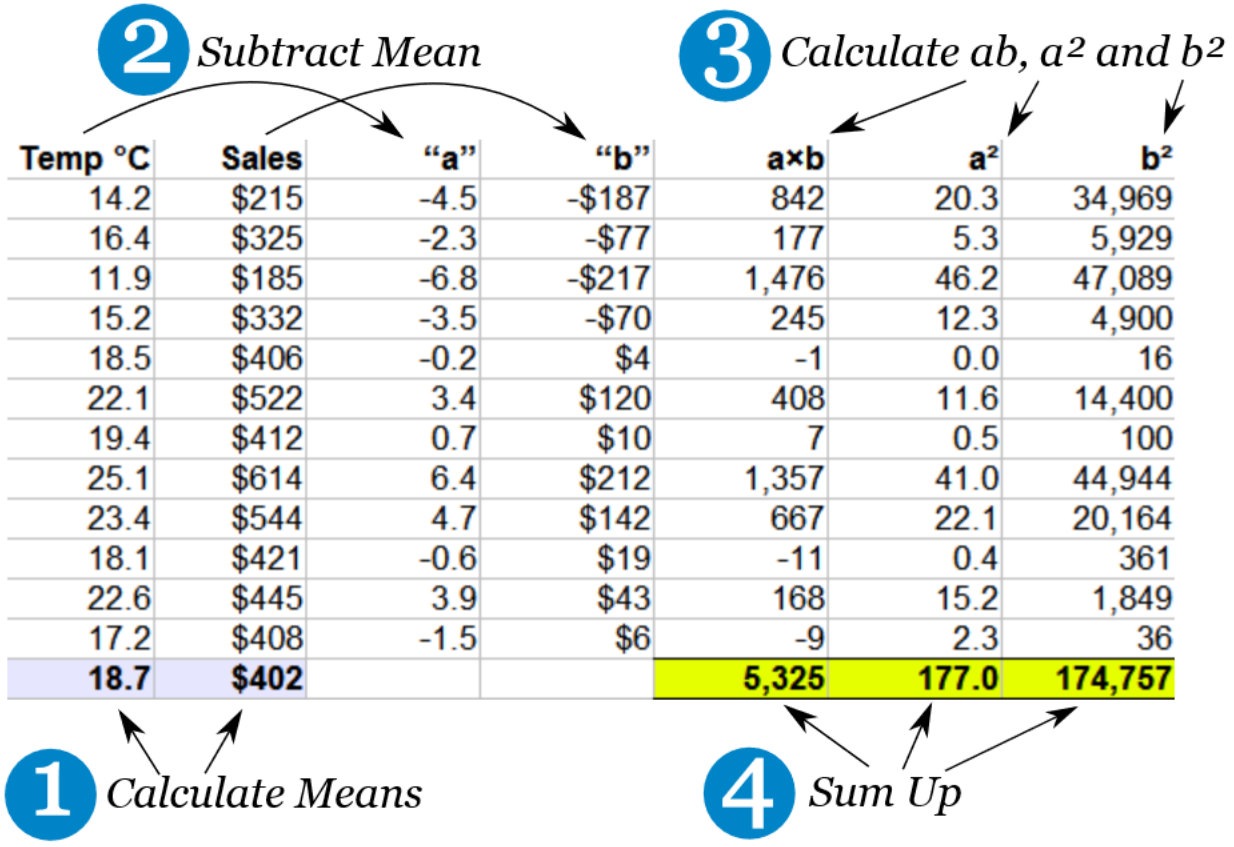
How to calculate Correlation Coefficient?

Let us call the two sets of data "x" and "y" (in our case Temperature is x and Ice Cream Sales is y)

1. **Step 1:** Find the mean of x, and the mean of y
2. **Step 2:** Subtract the mean of x from every x value (call them "a"), do the same for y (call them "b")
3. **Step 3:** Calculate: a*b, a^2 and b^2 for every value
4. **Step 4:** Sum up a*b, sum up a^2 and sum up b^2
5. **Step 5:** Divide the sum of a*b by the square root of [(sum of a^2) × (sum of b^2)]

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlation Coefficient Calculation



Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

5 $$\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$$

Covariance

It is a measure of the relationship between two variables.

Example: John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock of ABC Corp to his portfolio, he wants to assess the directional relationship between the stock of ABC Corp and the S&P 500.

John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction.

John can calculate the covariance between the stock of ABC Corp. and S&P 500. Unlike the correlation coefficient, covariance is measured in units and it can be positive or negative values. The values are interpreted as follows:

- **Positive covariance:** Indicates that two variables tend to move in the same direction.
- **Negative covariance:** Reveals that two variables tend to move in inverse directions.

Covariance vs. Correlation

Using covariance, we can only gauge the direction of the relationship (whether the variables tend to move in tandem or show an inverse relationship). However, it does not indicate the strength of the relationship, nor the dependency between the variables.

On the other hand, correlation measures the strength of the relationship between variables. Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

The relationship between the two concepts can be expressed using the formula:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Where

$\rho(X,Y)$ – the correlation between the variables X and Y

$Cov(X,Y)$ – the covariance between the variables X and Y

σX – the standard deviation of the X-variable

σY – the standard deviation of the Y-variable

Covariance cont…

The formula for calculating covariance of sample data is shown below.

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Where:

X_i – the values of the X-variable

Y_j – the values of the Y-variable

$\bar{X}$ – the mean (average) of the X-variable

$\bar{Y}$ – the mean (average) of the Y-variable

n – the number of data points

The prices of ABC Corp. and the S&P 500 are as follows. Find the covariance.

Year	S&P 500	ABC Corp
2013	1692	68
2014	1978	102
2015	1884	110
2016	2151	112
2017	2519	154

Thank You
End of Lab 6

Lab Experiments

1. You've been given a list of twenty blood types for emergency surgery patients: A, O, A, B, B, AB, B, B, O, A, O, O, O, AB, B, AB, AB, A, O, A. WAP to draw a frequency distribution table consisting of distinct items, frequency, proportion, percent and cumulative frequency.
2. WAP to calculate the coefficient of skeweness based on mean and median for the following distribution:

Class Interval	Frequency
0-10	6
10-20	12
20-30	22
30-40	48
40-50	56
50-60	32
60-70	18
70-80	6

Lab Experiments cont…

3. WAP to comment on the nature of distribution:
   - 14, 14, 14, 14, 14
   - 11, 12, 14, 16, 17
   - 1, 3, 6, 8, 42
4. The following facts were gathered from a firm before and after an industrial update. By making use of the above data, compare the position of the firm before and after the dispute as fully as possible.

	Before Dispute	After Dispute
Mean wages	850	900
Median wages	820	800
Number employed	600	550
Standard distribution	30	110
Quartiles	750 & 920	750 & 950
Modal wages	760	600

Lab Experiments cont…

5. WAP to comment on the nature of skewness:
   - Size of items: 10-12 12-14 14-16 16-18 18-20
     Frequency 27 20 12 6 3
   - Size of items: 10-12 12-14 14-16 16-18 18-20
     Frequency 3 6 12 20 27
6. For the following marks of 36 students in an examination, WAP to exhibit:
   - Measures of Frequency
   - Measures of Central Tendency
   - Measures of Dispersion or Variation
   - Measures of Position

55	75	65	30	90	55	40	50	60	80	80	76	95
75	55	45	65	80	30	50	75	85	80	90	75	75
90	65	78	72	82	52	62	67	66	65	88	45	70

Lab Experiments cont…

7. WAP to determine the types of kurtosis for the data values 0, 3, 4, 1, 2, 3, 0, 2, 1, 3, 2, 0, 2, 2, 3, 2, 5, 2, 3, 999.
8. A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams. WAP to calculate correlation coefficient and determine whether there is an association between the two variables?

Infant ID	1	2	3	4	5	6	7	8	9	10	11
Gestational Age (weeks)	34.7	36	29.3	40.1	35.7	42.4	40.3	37.3	40.9	38.3	38.5
Birth weight(gm)	1895	2030	1440	2835	3090	3827	3260	2690	3285	2920	3430

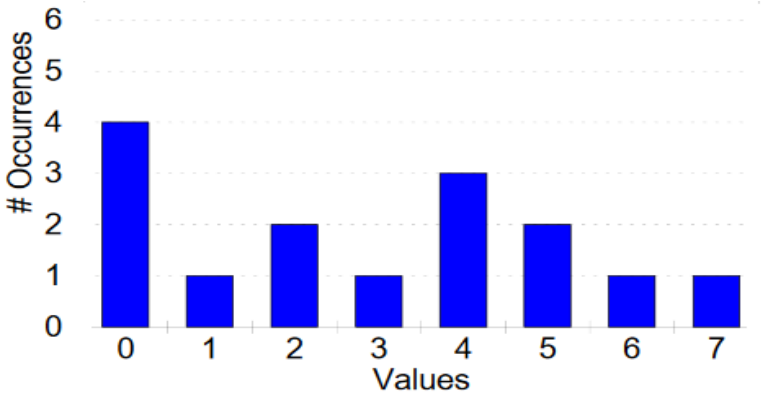
Infant ID	12	13	14	15	16	17
Gestational Age (weeks)	41.4	39.7	39.7	41.1	38.0	38.7
Birth weight(gm)	3657	3685	3345	3260	2680	2005

Lab Experiments cont…

9. Nine students held their breath, once after breathing normally and relaxing for one minute, and once after hyperventilating for one minute. The table indicates how long (in sec) they were able to hold their breath. Is there an association between the two variables?

Subject	A	B	C	D	E	F	G	H	I
Normal	56	56	65	65	50	25	87	44	35
Hypervent	87	91	85	91	75	28	122	66	58

10. Find the Standard Score, Quartile, and The Five Number Summary of the data in below figure.



Lab Experiments cont...

11. Given the following return information, what is the covariance and correlation coefficient between the return of Stock A and the return of the market index?

Month	Return of stock A	Return of Market Index
1	2.3	1.3
2	2.5	5.0
3	1.9	0.8
4	2.4	1.9
5	2.1	1.1

12. Find the covariance and correlation coefficient of eruption duration and waiting time in the data set **faithful**. Observe if there is any linear relationship between the two variables.