EDU ENGINEERING

PIONEER OF ENGINEERING NOTES

TAMIL NADU'S BEST EDTECH PLATFORM FOR ENGINEERING

CONNECT WITH US

WEBSITE: www.eduengineering.net

TELEGRAM: @eduengineering

INSTAGRAM: @eduengineering

- **Regular Updates for all Semesters**
- **All Department** Notes AVAILABLE
- **Handwritten** Notes AVAILABLE
- **Past Year** Question Papers AVAILABLE
- **Subject wise** Question Banks AVAILABLE
- **Important Questions** for Semesters AVAILABLE
- Various **Author Books** AVAILABLE

AD3491 FUNDAMENTALS OF DATA SCIENCE

UNIT 2

Frequency Distribution and Data: Types, Tables, and Graphs

Frequency distribution in statistics provides the information of the number of occurrences (frequency) of distinct values distributed within a given period of time or interval, in a list, table, or graphical representation.

Types of Frequency Distribution:

There are two types of Frequency Distribution.

- Grouped
- Ungrouped

There are two types Data is a **collection of numbers** or **values**

Data: Any bit of information that is expressed in a **value or numerical number** is data. Data is basically a collection of information, measurements or observations.

For example

- The marks you scored in your Math exam is data
- The number of cars that pass through a bridge in a day.

Raw data :

Raw data is an initial collection of information. This information has not yet been organized. After the very first step of data collection, you will get raw data. For example,

A group of five friends their favourite colour. The answers are Blue, Green, Blue, Red, and Red. This collection of information is the raw data.

Discrete data :***Discrete data*** is that which is recorded in whole numbers, like the number of children in a school or number of tigers in a zoo. It cannot be in decimals or fractions.

Continuous data :Continuous data need not be in whole numbers, it can be in decimals. Examples are the temperature in a city for a week, your percentage of marks for the last exam etc.

Example of Data Handling:

- Pictographs

- Bar Graphs
- Histogram and Pie-Charts
- Chance and Probability
- Arithmetic Mean and Median and Mode

Frequency

The frequency of any value is the number of times that value appears in a data set. So from the above examples of colours, we can say two children like the colour blue, so its frequency is two. So to make meaning of the raw data, we must organize. And finding out the frequency of the data values is how this organisation is done.

Frequency Distribution

Many times it is not easy or feasible to find the frequency of data from a very large dataset. So to make sense of the data we make a frequency table and graphs. Let us take the example of the heights of ten students in cms.

Frequency Distribution Table

139, 145, 150, 145, 136, 150, 152, 144, 138, 138

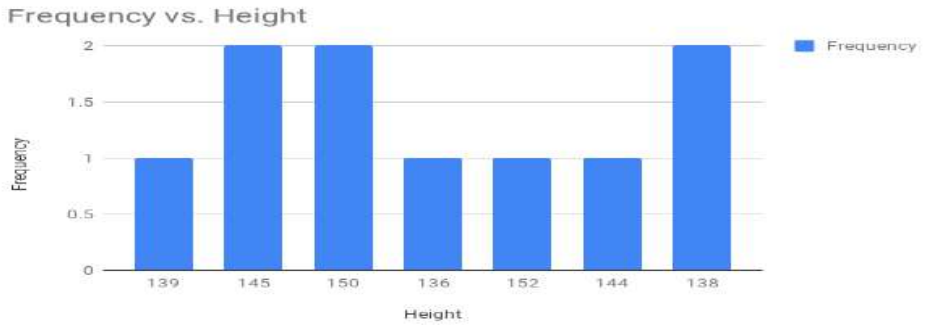
Height	Frequency
139	1
145	2
150	2
136	1
152	1
144	1
138	2

This frequency table will help us make better sense of the data given. Also when the data set is too big (say if we were dealing with 100 students) we use tally marks for counting. It makes the task more organized and easy. Below is an example of how we use tally marks.

1	I	6	𝍸 I
2	II	7	𝍸 II
3	III	8	𝍸 III
4	IIII	9	𝍸 IIII
5	𝍸	10	𝍸 𝍸

Frequency Distribution Graph

Using the same above example we can make the following graph:



Learn more about Bar Graphs and Histogram here.

Types of Frequency Distribution

- Grouped frequency distribution.
- Ungrouped frequency distribution.
- Cumulative frequency distribution.
- Relative frequency distribution.
- Relative cumulative frequency distribution.

Grouped Data

At certain times to ensure that we are making correct and relevant observations from the data set, we may need to group the data into class intervals. This ensures that the frequency distribution best represents the data. example :the height of students.

Class Interval	Frequency
130-140	4
140-150	3
150-160	3

From the above table, you can see that the value of 150 is put in the class interval of 150-160 and not 140-150. This is the convention we must follow.

- **The table gives the number of snacks ordered and the number of days as a tally. Find the frequency of snacks ordered. 2**

snacks	Tally
2-4	||||
4-6	|||
6-8	𝍸|||||
8-10	𝍸|||||
10-12	||||𝍸

Answer: From the frequency table the number of snacks ordered ranging between

- 2-4 is 4 days
- 4 to 6 is 3 days
- 6 to 8 is 9 days
- 8 to 10 is 9 days
- 10 to 12 is 7 days.

So the frequencies for all snacks ordered are 4, 3, 9, 9, 7

- **How to find frequency distribution? 2**

Answer: We can find frequency distribution by the following steps:

- First of all, calculate the range of the data set.
- Next, divide the range by the number of the group you want your data in and then round up.
- After that, use class width to create groups
- Finally, find the frequency for each group.

- **Define frequency distribution in statistics? 2**

Answer: In an overview, the frequency distribution of all distinct values in some variables and the number of times they occur. Meaning that it tells how frequencies are distributed overvalues in a frequency distribution. However, mostly we use frequency distributions to summarize categorical variables.

- **Why are frequency distributions important? 2**

Answer: It has great importance in statistics. Also, a well-structured frequency distribution makes possible a detailed analysis of the structure of the population with respect to given characteristics. Therefore, the groups into which the population break down can be determined.

- **State the components of frequency distribution? 2**

Answer: The various components of the frequency distribution are: Class interval, types of class interval, class boundaries, midpoint or class mark, width or size o class interval, class frequency,

frequency density = class frequency/ class width,

relative frequency = class frequency/ total frequency, etc.

Descriptive Statistics

A population is the group to be studied, and population data is a collection of all elements in the population. For example:

- All the fish in Long Lake.
- All the lakes in the Adirondack Park.
- All the grizzly bears in Yellowstone National Park.

A sample is a subset of data drawn from the population of interest. For example:

- 100 fish randomly sampled from Long Lake.
- 25 lakes randomly selected from the Adirondack Park.
- 60 grizzly bears with a home range in Yellowstone National Park.

Populations are characterized by descriptive measures called parameters. Inferences about parameters are based on sample statistics.

For example,

The population mean (μ) is estimated by the sample mean ($\bar{x}$). The population variance (σ2) is estimated by the sample variance (s2).

Variables are the characteristics we are interested in.

For example:

- The length of fish in Long Lake.
- The pH of lakes in the Adirondack Park.

- The weight of grizzly bears in Yellowstone National Park.

Variables are divided into two major groups: **Qualitative And Quantitative**.

1. **Qualitative variables**

- Qualitative variables have values that are attributes or categories.
- Mathematical operations cannot be applied to qualitative variables.
- Examples of qualitative variables are gender, race, and petal color.
- Quantitative variables have values that are typically numeric, such as measurements.
- Mathematical operations can be applied to these data. Examples of quantitative variables are age, height, and length.

2. **Quantitative variables**

    - Quantitative variables can be broken down further into two more categories: discrete and continuous variables.
    - **Discrete variables** have a finite or countable number of possible values. Think of discrete variables as "hens." Hens can lay 1 egg, or 2 eggs, or 13 eggs… There are a limited, definable number of values that the variable could take on.
    - **Continuous variables** have an infinite number of possible values. Think of continuous variables as "cows." Cows can give 4.6713245 gallons of milk, or 7.0918754 gallons of milk, or 13.272698 gallons of milk … There are an almost infinite number of values that a continuous variable could take on.

Examples

Is the variable qualitative or quantitative?

Species	Weight	Diameter	Zip Code
(qualitative	quantitative,	quantitative,	qualitative)

Graphs

Data can be described clearly and concisely with the aid of a well-constructed frequency distribution.

GRAPHS FOR QUANTITATIVE DATA

Histograms

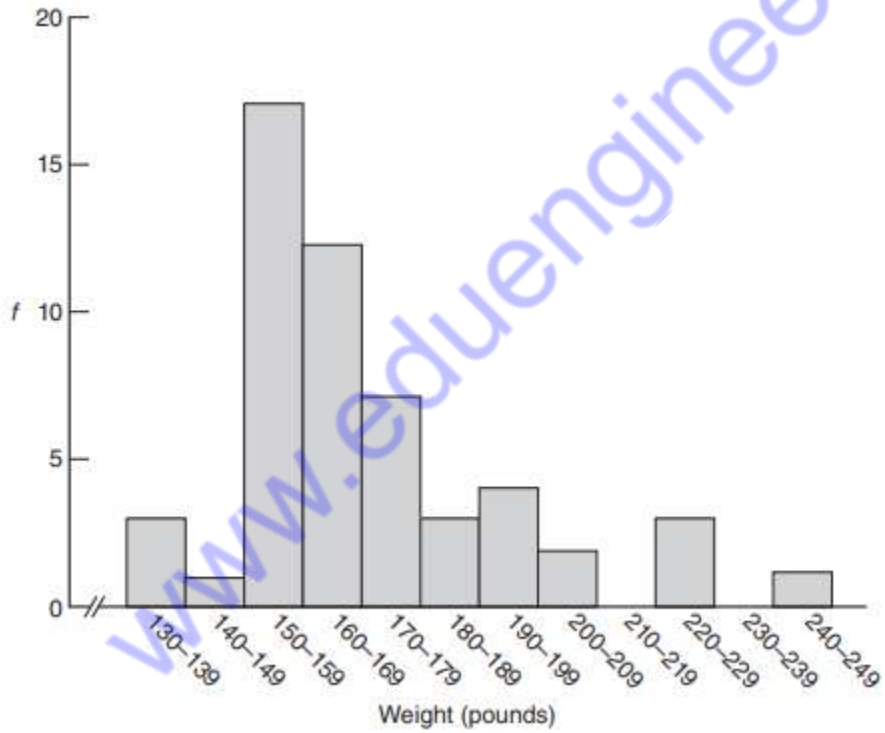
A bar-type graph for quantitative data. The common boundaries between adjacent bars emphasize the continuity of the data, as with continuous variables.

A histogram in Figure shows a casual glance at this histogram confirms previous conclusions: a dense concentration of weights among the 150s, 160s, and 170s, with a spread in the direction of the heavier weights. Let's pinpoint some of the more important features of histograms.

■ Equal units along the horizontal axis (the X axis, or abscissa) reflect the various class intervals of the frequency distribution.

■ Equal units along the vertical axis (the Y axis, or ordinate) reflect increases in frequency. (The units along the vertical axis do not have to be the same width as those along the horizontal axis.)

■ The intersection of the two axes defines the origin at which both numerical scales equal 0



Frequency Polygon

A line graph for quantitative data that also emphasizes the continuity of continuous variables

An important variation on a histogram is the frequency polygon, or line graph. Frequency polygons may be constructed directly from frequency distributions. However, we will follow the step-by-step transformation of a histogram into a frequency polygon, as described in panels A, B, C, and D of Figure 2.2. A. This panel shows the histogram for the weight distribution. B. Place dots at the midpoints of each bar top or, in the absence of bar tops, at midpoints for classes on the horizontal axis, and connect them with straight lines. [To find the midpoint of any class, such

as 160–169, simply add the two tabled boundaries (160 + 169 = 329) and divide this sum by 2 (329/2 = 164.5).] C. Anchor the frequency polygon to the horizontal axis. First, extend the upper tail to the midpoint of the first unoccupied class (250–259) on the upper flank of the histogram. Then extend the lower tail to the midpoint of the first unoccupied class (120–129) on the lower flank of the histogram. Now all of the area under the frequency polygon is enclosed completely. D. Finally, erase all of the histogram bars, leaving only the frequency polygon. Frequency polygons are particularly useful when two or more frequency distributions or relative frequency distributions are to be included in the same graph.

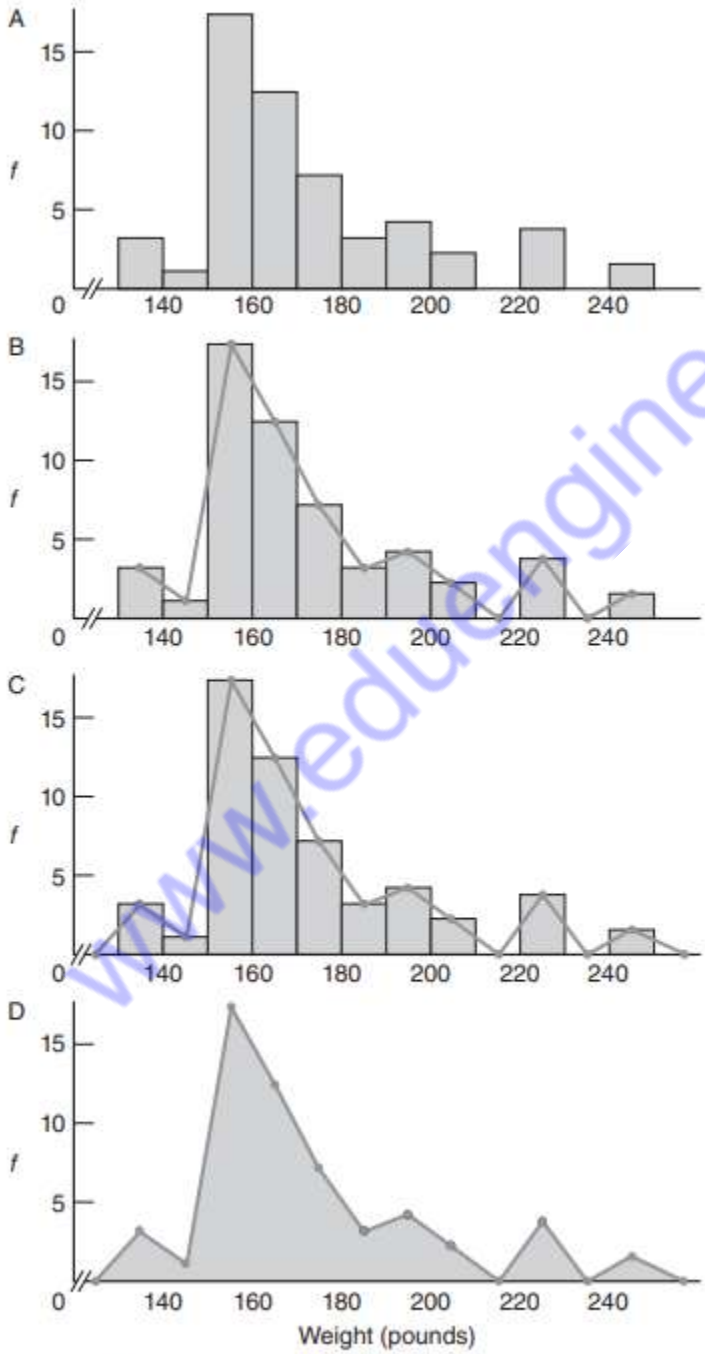


FIGURE 2.2

Transition from histogram to frequency polygon.

Stem and Leaf Displays:

A device for sorting quantitative data on the basis of leading and trailing digits.

Still another technique for summarizing quantitative data is a stem and leaf display. Stem and leaf displays are ideal for summarizing distributions, such as that for weight data, without destroying the identities of individual observations.

Constructing a Display

The stemplot (also called stem and leaf plot) is another graphical display ofthe distribution of quantitative variable.

To create a stemplot, the idea is to separate each data point into a stemand leaf, asfollows:

- The leaf is the right-most digit.
- The stem is everything except the right-most digit.
- So, if the data point is 34, then 3 is the stem and 4 is the leaf.
- If the data point is 3.41, then 3.4 is the stem and 1 is the leaf.
- Note: For this to work, ALL data points should be rounded to the same number of decimal places.

EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45
49 39 34 26 25 35 33

To make a stemplot:

- Separate each observation into a stem and a leaf.
- Write the stems in a vertical column with the smallest at the top, and draw avertical line at the right of this column.
- Go through the data points, and write each leaf in the row to the right of itsstem.
- Rearrange the leaves in an increasing order.

When some of the stems hold a large number of leaves, we can split each stem into two: one holding the leaves 0-4, and the other holding the leaves 5-9. A

```
Steps 1, 2 and 3          step 4

2|616965                  2|156669                  2|1
3|447513038359453         3|013333444555789         2|56669
4|21191359                4|11123599                3|013333444
5|                  ==>   5|                  ==>   3|555789
6|1                        6|1                       4|11123
7|4                       7|4                       4|599
8|0                       8|0                       5|
                                                    5|
                                                    6|1
                                                    6|
                                                    7|4
                                                    7|
                                                    8|0
```

statistical software package will often do the splitting for you, when appropriate.Note that when rotated 90 degrees counter-clockwise, the stemplot visuallyresembles a histogram:

```
      4
      4
      4
      3 9
    9 3 8 3
    6 3 7 2
    6 3 5 1 9
    6 1 5 1 9
1 5 0 5 1 5     1   4   0
-----------------------
2 2 3 3 4 4 5 5 6 6 7 7 8
```

The stemplot has additional unique features:

- preserves the original data.
- It sorts the data (which will become very useful in the next section).

Typical Shapes

Whether expressed as a histogram, a frequency polygon, or a stem and leaf display, an important characteristic of a frequency distribution is its shape. Figure 2.3 shows some of the more typical shapes for smoothed frequency polygons (which ignore the inevitable irregularities of real data).

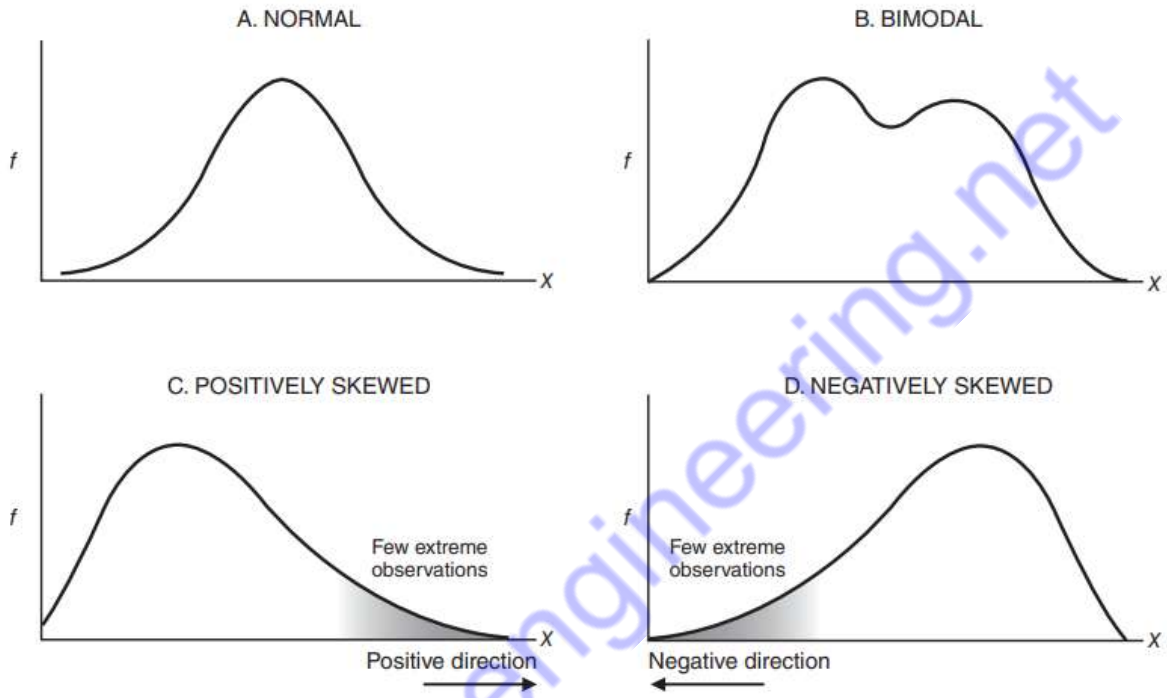


FIGURE 2.3
Typical shapes.

Normal

Any distribution that approximates the normal shape in panel A of Figure 2.3 can be analyzed, as we will see in Chapter 5, with the aid of the well-documented normal curve. The familiar bell-shaped silhouette of the normal curve can be superimposed on many frequency distributions, including those for uninterrupted gestation periods of human fetuses, scores on standardized tests, and even the popping times of individual kernels in a batch of popcorn.

Bimodal

Any distribution that approximates the bimodal shape in panel B of Figure 2.3 might, as suggested previously, reflect the coexistence of two different types of observations in the same distribution. For instance, the distribution of the ages of residents in a neighborhood consisting largely of either new parents or their infants has a bimodal shape.

Positively Skewed The two remaining shapes in Figure 2.3 are lopsided. A lopsided distribution caused by a few extreme observations in the positive direction (to the right of the majority of observations), as in panel C of Figure 2.3, is a positively skewed distribution.

The distribution of incomes among U.S. families has a pronounced positive skew, with most family incomes under $200,000 and relatively few family incomes spanning a wide range of values above $200,000. The distribution of weights in Figure 2.1 also is positively skewed.

Negatively Skewed A lopsided distribution caused by a few extreme observations in the negative direction (to the left of the majority of observations), as in panel D of Figure 2.3, is a negatively skewed distribution. The distribution of ages at retirement among U.S. job holders has a pronounced negative skew, with most retirement ages at 60 years or older and relatively few retirement ages spanning the wide range of ages younger than 60.

A GRAPH FOR QUALITATIVE (NOMINAL) DATA:

The distribution in Table 2.7, based on replies to the question "Do you have a Facebook profile?" appears as a bar graph in Figure 2.4. A glance at this graph confirms that Yes replies occur approximately twice as often as No replies. As with histograms, equal segments along the horizontal axis are allocated to the different words or classes that appear in the frequency distribution for qualitative data. Likewise, equal segments along the vertical axis reflect increases in frequency. The body of the bar graph consists of a series of bars whose heights reflect the frequencies for the various words or classes. A person's answer to the question "Do you have a Facebook profile?" is either Yes or No, not some impossible intermediate value, such as 40 percent Yes and 60 percent No. Gaps are placed between adjacent bars of bar graphs to emphasize the discontinuous nature of qualitative data. A bar graph also can be used with quantitative data to emphasize the discontinuous nature of a discrete variable, such as the number of children in a family.

FIGURE 2.4
Bar graph.

Misleading Graphs:

Graphs can be constructed in an unscrupulous manner to support a particular point of view. Indeed, this type of statistical fraud gives credibility to popular sayings, including "Numbers don't lie, but statisticians do" and "There are three kinds of lies—lies, damned lies, and statistics." For example, to imply that comparatively many students responded Yes to the Facebook profile question, an unscrupulous person might resort to the various tricks shown in Figure 2.5:

■ The width of the Yes bar is more than three times that of the No bar, thus violating the custom that bars be equal in width.
■ The lower end of the frequency scale is omitted, thus violating the custom that the entire scale be reproduced, beginning with zero. (Otherwise, a broken scale should be highlighted by crossover lines, as in Figures 2.1 and 2.2.)
■ The height of the vertical axis is several times the width of the horizontal axis, thus violating the custom, heretofore unmentioned, that the vertical axis be approximately as tall as the horizontal axis is wide. Beware of graphs in which, because the vertical axis is many times larger than the horizontal axis (as in Figure 2.5), frequency differences are exaggerated, or in which, because the vertical axis is many times smaller than the horizontal axis, frequency differences are suppressed.

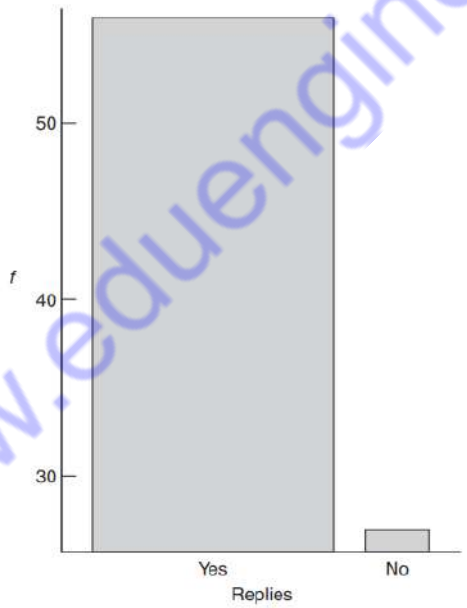


FIGURE 2.5
Distorted bar graph.

AVERAGES

A center of a data set is a way of describing a location. We can measure a center of a data in 3 different ways: the mean (average), the median and the mode.

The two main numerical measures for the center of a distribution are the mean and the median. Each one of these measures is based on a completely different idea of describing the center of a distribution. Let us first present each one of the measures, and then compare their properties.

MEAN

The mean is the average of a set of observations (i.e., the sum of the observations divided by the number of observations).

The mean is the average of a set of observations. If the n observations are written as their mean can be written mathematically as: their mean is:

$$x_1, x_2, \ldots, x_n$$

We read the symbol as "x-bar." The bar notation is commonly used to represent the samplemean, i.e. the mean of the sample.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example .

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45

49 39 34 26 25 35 33

The mean age of the 32 actresses is:

$$\bar{x} = \frac{34 + 34 + 26 + \ldots + 35 + 33}{32} = \frac{1233}{32} = 38.5$$

We add all of the ages to get 1233 and divide by the number of ages which was 32 to get 38.5. We denote this result as x-bar and called the sample mean.

EXAMPLE: World Cup Soccer

Often we have large sets of data and use a frequency table to display the data more efficiently. Data were collected from the last three World Cup soccer tournaments. A total of 192 games were played. The table below lists the number of goals scored per game (not including any goals scored in shootouts).

Total # Goals/Game	**Frequency**
0	17
1	45
2	51

3	37
4	25
5	11
6	3
7	2
8	1

To find the mean number of goals scored per game, we would need to find the sum of all 192 numbers, and then divide that sum by 192.

Rather than add 192 numbers, we use the fact that the same numbers appear many times. For example, the number 0 appears 17 times, the number 1 appears 45 times, the number2 appears 51 times, etc.

If we add up 17 zeros, we get 0. If we add up 45 ones, we get 45. If we add up 51 twos, we get 102. Repeated addition is multiplication.

Thus, the sum of the 192 numbers

= 0(17) + 1(45) + 2(51) + 3(37) + 4(25) + 5(11) + 6(3) + 7(2) + 8(1) = 453.

The sample mean is then 453 / 192 = 2.359.

Note that, in this example, the values of 1, 2, and 3 are the most common andour averagefalls in this range representing the bulk of the data.

MEDIAN

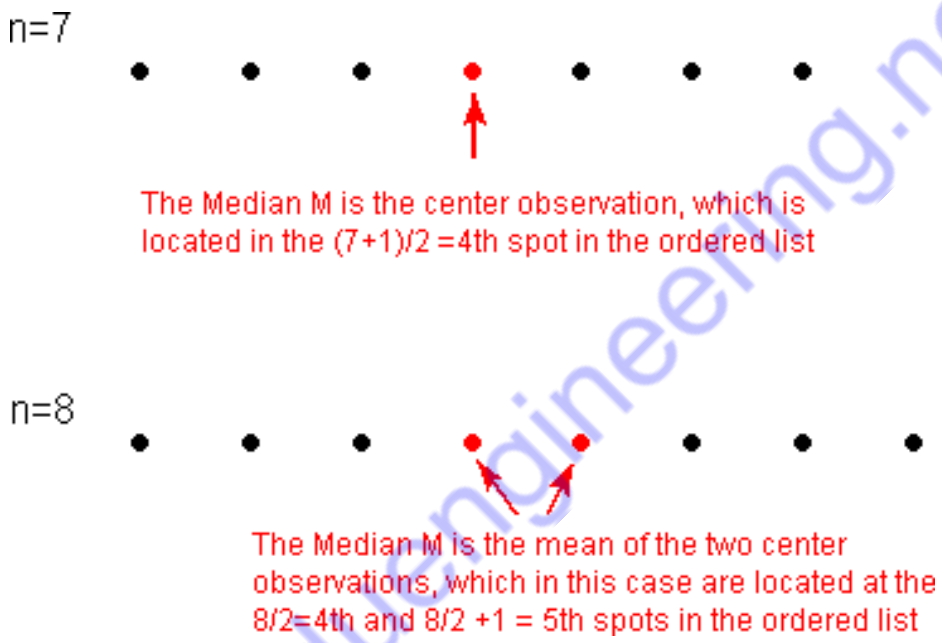
Define and calculate the sample median of a quantitative variable.
The median M is the midpoint of the distribution. It is the number suchthat half of the observations fall above, and half fall below.
To find the median:
Order the data from smallest to largest.
Consider whether n, the number of observations, is even or odd.
If n is odd, the median M is the center observation in the ordered list. Thisobservation is the one "sitting" in the (n + 1) / 2 spot in the ordered list.

If n is even, the median M is the mean of the two center observations in the ordered

list. These two observations are the ones “sitting” in the (n / 2) and(n / 2) + 1 spots in the ordered list.

EXAMPLE: Median (1)

For a simple visualization of the location of the median, consider the followingtwo simple cases of n = 7 and n = 8 ordered observations, with each observation represented by asolid circle:



Comments:

In the images above, the dots are equally spaced, this need not indicate the data values are actually equally spaced as we are only interested in listing them in order. In fact, in the above pictures, two subsequent dots could have exactly the same value. It is clear that the value of the median will be in the same position regardless of the distance between data values.

EXAMPLE: Median (2)

To find the median age of the Best Actress Oscar winners, we first need to order the data. It would be useful, then, to use the stemplot, a diagram in which the dataare already ordered.

Here n = 32 (an even number), so the median M, will be the mean of thetwo center observations.

These are located at the (n / 2) = 32 / 2 = 16th and(n / 2) +
1 = (32 / 2) + 1 = 17th

Counting from the top, we find that: the 16th ranked observation is 35the 17thranked observation also happens to be 35. Therefore, the median M = (35 + 35) / 2 = 35

```
2| 1
2| 56669
3| 013333444
3| 555789
4| 11123
4| 599
5|
5|
6| 1
6|
7| 4
7|
8| 0
```

Comparing the Mean and the Median

The mean and the median, the most common measures of center, each describe the centerof a distribution of values in a different way.

The mean describes the center as an average value, in which the actual values of the data points play an important role.

The median, on the other hand, locates the middle value as the center, and theorder of the data is the key.

To get a deeper understanding of the differences between these twomeasures of center,consider the following example. Here are two datasets:

Data set A → 64 65 66 68 70 71 73
Data set B → 64 65 66 68 70 71 730
For dataset A, the mean is 68.1, and the median is 68.

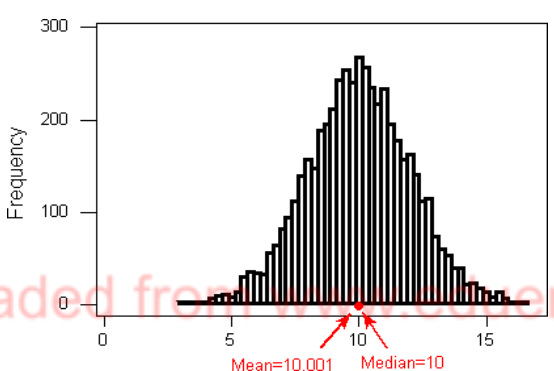
Looking at dataset B, notice that all of the observations except the last one are close together. The observation 730 is very large, and is certainly an outlier. In this case, the median is still 68, but the mean will be influenced by the high outlier, and shifted up to 162.
The message that we should take from this example is:
The mean is very sensitive to outliers (because it factors in their magnitude), while the median is resistant (or robust) to outliers.

MODE: 3rd Measure

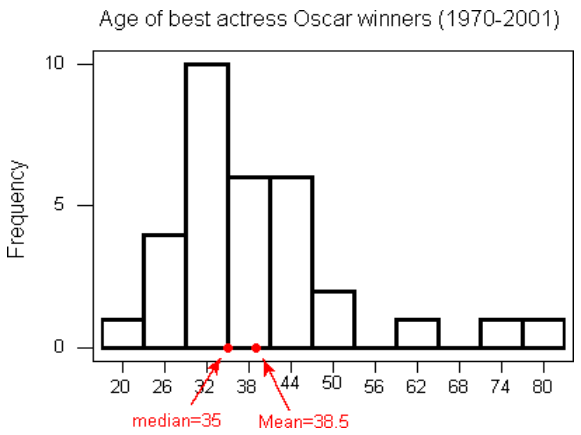
The mode of a data set is the number that occurs most frequently in the set.

- If no value appears more than once in the data set, the data set has no mode.
- If a there are two values that appear in the data set an equal number of times, theyboth will be modes etc.

For symmetric distributions with no outliers: the mean is approximately equaltothe median.

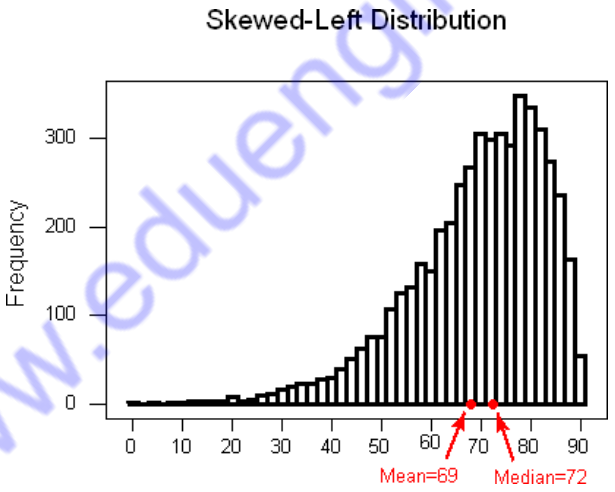


For skewed right distributions and/or datasets with high outliers: the mean is



greater than the median.

For skewed left distributions and/or datasets with low outliers: the mean is less than the median.



When to use which measures?

- Use the sample mean as a measure of center for symmetric distributions with no outliers. Otherwise, the median will be a more appropriate measure of the center of our data.

Let's Summarize

- The two main numerical measures for the center of a distribution are the

mean and the median. The mean is the average value, while the median is the middle value.

- The mean is very sensitive to outliers (as it factors in their magnitude), while the median is resistant to outliers.
- The mean is an appropriate measure of center for symmetric distributions with no outliers. In all other cases, the median is often a better measure of the center of the distribution.

Describing Variability

Intuitive Approach

- In Figure 4.1, each of the three frequency distributions consists of seven scores with the same mean (10) but with different variabilities. (Ignore the numbers in boxes; their significance will be explained later.) Before reading on, rank the three distributions from least to most variable. Your intuition was correct if you concluded that distribution A has the least variability, distribution B has intermediate variability, and distribution C has the most variability. If this conclusion is not obvious, look at each of the three distributions, one at a time, and note any differences among the values of individual scores. For distribution A with the least (zero) variability, all seven scores have the same value (10). For distribution B with intermediate variability, the values of scores vary slightly (one 9 and one 11), and for distribution C with most variability, they vary even more (one 7, two 9s, two 11s, and one 13). Importance of Variability Variability assumes a key role in an analysis of research results. For example, a researcher might ask: Does fitness training improve, on average, the scores of depressed patients on a mental-wellness test? To answer this question, depressed patients are randomly assigned to two groups, fitness training is given to one group, and wellness scores are obtained for both groups. Let's assume that the mean wellness score is larger for the group with fitness training. Is the observed mean difference between the two groups real or merely transitory? This decision depends not only on the size of the mean difference between the two groups but also on the inevitable variabilities of individual scores within each group. To illustrate the importance of variability, Figure 4.2 shows the outcomes for two fictitious experiments, each with the same mean difference of 2, but with the two groups in experiment B having less variability than the two groups in experiment C. Notice that groups B and C in Figure 4.2 are the same as their counterparts in Figure 4.1. Although the new group B* retains exactly the same (intermediate) variability

as group B, each of its seven scores and its mean have been shifted 2 units to the right. Likewise, although the new group C* retains exactly the same (most) variability as group C, each of its seven scores and its mean have been shifted 2 units to the right. Consequently, the crucial mean difference of 2 (from 12 − 10 = 2) is the same for both experiments. Before reading on, decide which mean difference of 2 in Figure 4.2 is more apparent. The mean difference for experiment B should seem more apparent because of the smaller variabilities within both groups B and B*. Just as it's easier to hear a phone message when static is reduced, it's easier to see a difference between group means when variabilities within groups are reduced.

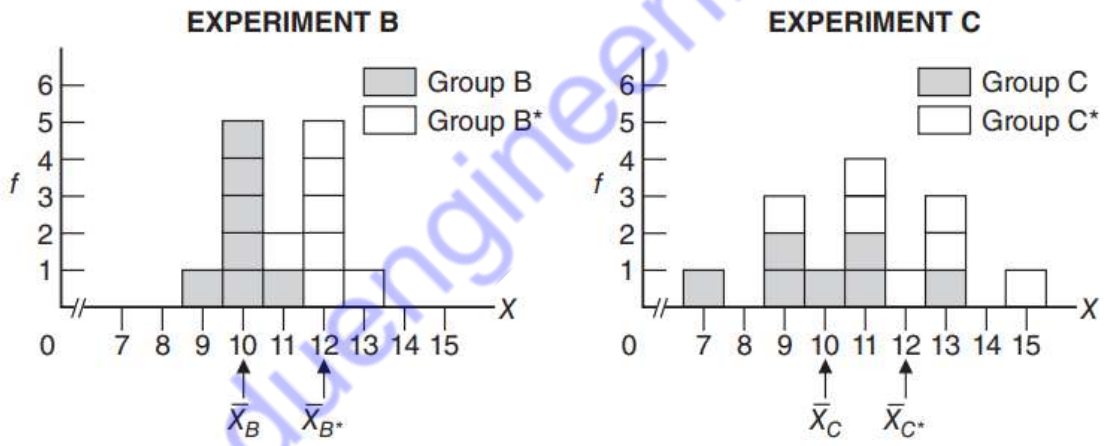


FIGURE 4.2

Two experiments with the same mean difference but dissimilar variabilities.

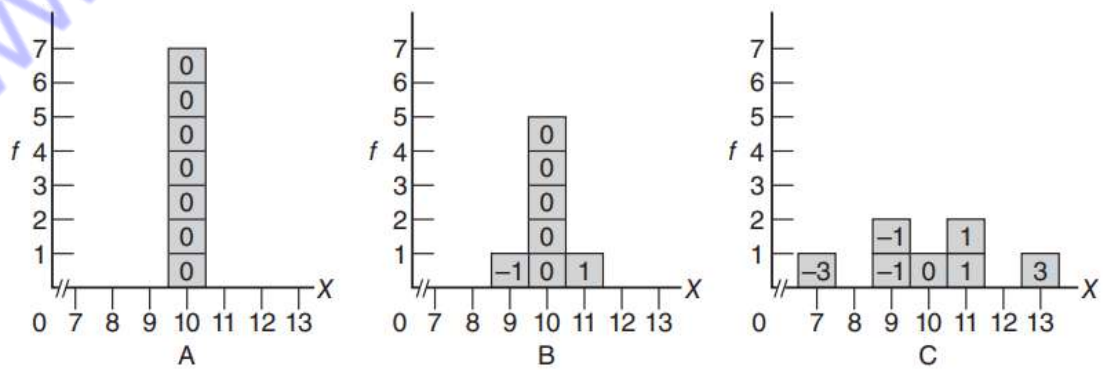


FIGURE 4.1

Three distributions with the same mean (10) but different amounts of variability. Numbers in the boxes indicate distances from the mean.

Range

A range measures the spread of a data inside the limits of a data set, it is calculated as a difference between the highest and lowest values in the data set. The larger the range, the greater the spread of the data.The range covered by the data is the most intuitive measure of variability. The range is exactly the distance between the smallest data point (min) and the largest one (Max).

Range = Max – min

Note: When we first looked at the histogram, and tried to get a first feel for thespread of the data, we were actually approximating the range, rather than calculating the exact range.

EXAMPLE: Best Actress Oscar Winners

Here we have the Best Actress Oscar winners' data
34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45
49 39 34 26 25 35 33
In this example:

min = 21 (Marlee Matlin for Children of a Lesser God, 1986) Max = 80 (Jessica Tandy for Driving Miss Daisy, 1989)

The range covered by all the data is 80 – 21 = 59 years.
Variance:
The mean of all squared deviation scores.
Although both the range and its most important spinoff, the interquartile range (discussed in Section 4.7), serve as valid measures of variability, neither is among the statistician's preferred measures of variability. Those roles are reserved for the variance and particularly for its square root, the standard deviation, because these measures serve as key components for other important statistical measures. Accordingly, the variance and standard deviation occupy the same exalted position among measures of variability as does the mean among measures of central tendency. Following the computational procedures described in later sections of this chapter, we could calculate the value of the variance for each of the three distributions in Figure 4.1. Its value equals 0.00 for the least variable distribution, A, 0.29 for the moderately

variable distribution, B, and 3.14 for the most variable distribution, C, in agreement with our intuitive judgments about the relative variability of these three distributions. Reconstructing the Variance To understand the variance better, let's reconstruct it step by step. Although a measure of variability, the variance also qualifies as a type of mean, that is, as the balance point for some distribution. To qualify as a type of mean, the values of all scores must be added and then divided by the total number of scores. In the case of the variance, each original score is re-expressed as a distance or deviation from the mean by subtracting the mean. For each of the three distributions in Figure 4.1, the face values of the seven original scores (shown as numbers along the X axis) have been re-expressed as deviation scores from their mean of 10 (shown as numbers in the boxes). For example, in distribution C, one score coincides with the mean of 10, four scores (two 9s and two 11s) deviate 1 unit from the mean, and two scores (one 7 and one 13) deviate 3 units from the mean, yielding a set of seven deviation scores: one 0, two –1s, two 1s, one –3, and one 3. (Deviation scores above the mean are assigned positive signs; those below the mean are assigned negative signs.) Mean of the Deviations Not a Useful Measure No useful measure of variability can be produced by calculating the mean of these seven deviations, since, as you will recall from Chapter 3, the sum of all deviations from their mean always equals zero. In effect, the sum of all negative deviations always counterbalances the sum of all positive deviations, regardless of the amount of variability in the group.

The standard deviation is to quantify the spread of a distribution by measuring how far the observations are from their mean. The standard deviation gives the average (or typicaldistance) between a data point and the mean.

Standard deviation is the measure of the overall spread (variability) of a data set valuesfrom the mean. The more spread out a data set is, the greater are thedistances from themean and the standard deviation.

There are many notations for the standard deviation: SD, s, Sd, StDev. Here, we'll use SDas an abbreviation for standard deviation, and use s as the symbol.Formula

The sample standard deviation formula is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Calculation

In order to get a better understanding of the standard deviation, it would beuseful tosee an example of how it is calculated.

EXAMPLE: Video Store Customers

The following are the number of customers who entered a video store in8 consecutivehours: 7, 9, 5, 13, 3, 11, 15, 9

To find the standard deviation of the number of hourly customers:

1. Find the mean, x-bar, of your data:

(7 + 9 + 5 + 13 + 3 + 11 + 15 + 9)/8 = 9

2. Find the deviations from the mean:

- The differences between each observation and the mean here are

  (7 – 9), (9 – 9), (5 – 9), (13 – 9), (3 – 9), (11 – 9), (15 – 9), (9 – 9)
  -2, 0, -4, 4, -6, 2, 6, 0

- Since the standard deviation attempts to measure the average (typical) distance between the data points and their mean, it would make sense to average the deviation we obtained.
- Note, however, that the sum of the deviations is zero.

3. To solve the previous problem, in our calculation, we square each of the deviations.

   $(-2)^2, (0)^2, (-4)^2, (4)^2, (-6)^2, (2)^2, (6)^2, (0)^2$

   4, 0, 16, 16, 36, 4, 36, 0

4. Sum the squared deviations and divide by n – 1:

   (4 + 0 + 16 + 16 + 36 + 4 + 36 + 0)/(8 – 1)

   (112)/(7) = 16

- This value, the sum of the squared deviations divided by n – 1, is called the variance. However, the variance is not used as a measure of spread directly as

the units are the square of the units of the original data.

5. The standard deviation of the data is the square root of the variance calculated in step.

In this case, we have the square root of 16 which is 4. We will use the lower case letter s represent the standard deviation. s = 4

- We take the square root to obtain a measure which is in the original units of the data. The units of the variance of 16 are in "squared customers" which is difficult to interpret.
- The units of the standard deviation are in "customers" which makes this measure ofvariation more useful in practice than the variance.

9. The interpretation of the standard deviation is that on average, the actual number of customers who enter the store each hour is 4 away from 9.

- The standard deviation is the square root of the variance (both population and sample).
- While the sample variance is the positive, unbiased estimator for the population variance, the units for the variance are squared.
- The standard deviation is a common method for numerically describing the distribution of a variable. The population standard deviation is σ (sigma) and sample standard deviation is *s*.

Population standard deviation Sample standard deviation

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2}$$

Example 7

Compute the standard deviation of the sample data: 3, 5, 7 with a sample mean of 5.

$$s = \sqrt{\frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1}} = \sqrt{4} = 2$$

DEGREES OF FREEDOM (d f)

Degrees of freedom (df) refers to the number of values that are free to vary, given one or more mathematical restrictions, in a sample being used to estimate a population characteristic.

The number of values free to vary, given one or more mathematical restrictions.

degrees of freedom, that is, df = n – 1.

Inter-Quartile Range (IQR)

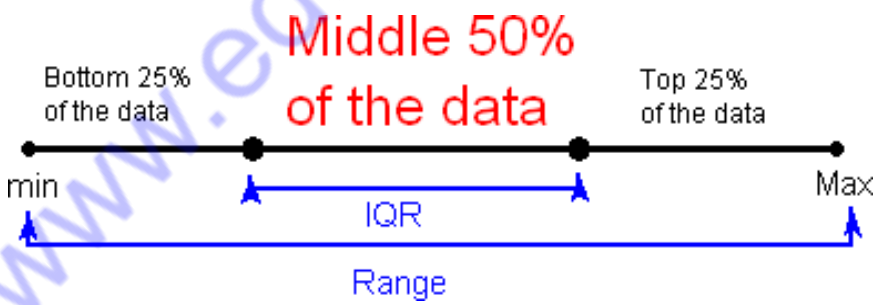
The Inter-Quartile Range or IQR measures the variability of a distribution by giving us the range covered by the MIDDLE 50% of the data.To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

IQR = Q3 – Q1

Q3 = 3rd Quartile = 75th PercentileQ1 = 1st

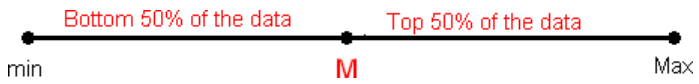
Quartile = 25th Percentile

The following picture illustrates this idea: (Think about the horizontal line as the data ranging from the min to the Max). IMPORTANT NOTE: The "lines" in the following illustrations are not to scale. The equal distances indicate equal amounts of data NOT equal distance between the numeric values.

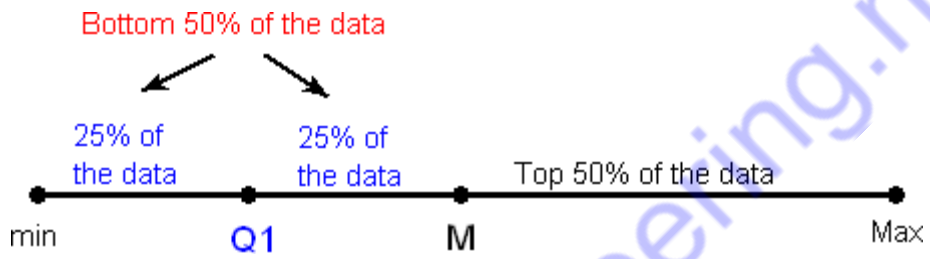


To calculate the IQR:

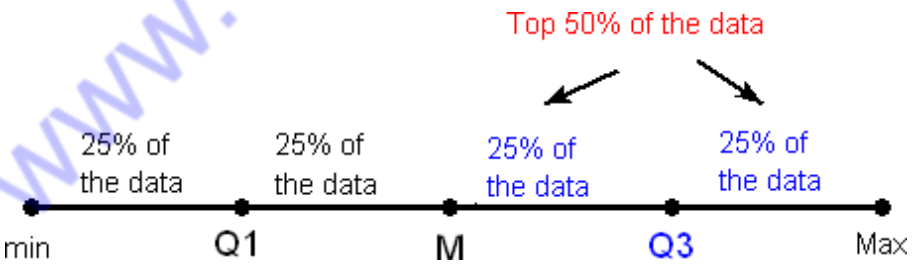
1. Arrange the data in increasing order, and find the median M. Recall that the median divides the data, so that 50% of the data points are below the median, and 50% of the data points are above the median.



2. Find the median of the lower 50% of the data. This is called the first quartile of the distribution, and the point is denoted by Q1. Note from the picture that Q1 divides the lower 50% of the data into two halves, containing 25% of the data points in eachhalf. Q1 is called the first quartile, since one quarter of the data points fall below it.

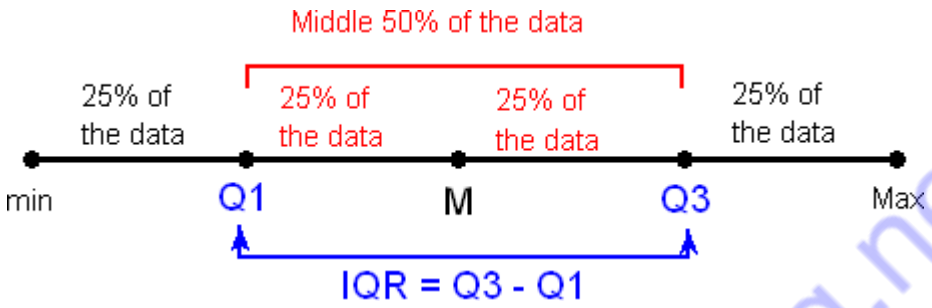


3. Repeat this again for the top 50% of the data. Find the median of the top 50% of the data. This point is called the third quartile of the distribution, and is denoted by Q3.Note from the picture that Q3 divides the top 50% of the data into two halves, with 25%of the data points in each.Q3 is called the third quartile,since three quarters of the data points fall below it.



4. The middle 50% of the data falls between Q1 and Q3, and therefore:

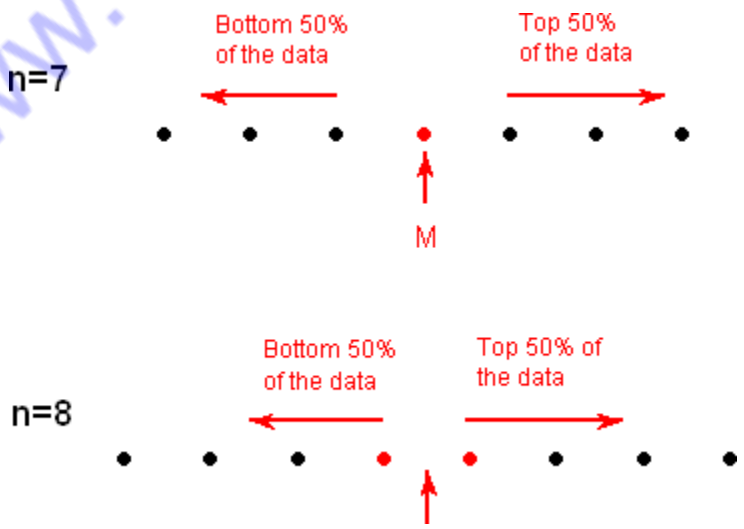
   IQR = Q3 – Q1.



Comments:

1. The last picture shows that Q1, M, and Q3 divide the data into four quarters with 25%of the data points in each, where the median is essentially the second quartile. The use of IQR = Q3 – Q1 as a measure of spread is therefore particularly appropriate when the median M is used as a measure ofcenter.

2. We can define a bit more precisely what is considered the bottom or top 50% of the data. The bottom (top) 50% of the data is all the observations whose position in the ordered list is to the left (right) of the location of the overall median M. The following picture will visually illustrate this for the simple cases of n = 7 and n = 8.



Note that when n is odd (as in n = 7 above), the median is not included in either the bottom or top half of the data; When n is even (as in n = 8 above), the data are naturally divided into two halves.

EXAMPLE: Best Actress Oscar Winners

To find the IQR of the Best Actress Oscar winners' distribution, it will be convenient touse the stemplot.

```
2| 1
2| 56669          Bottom Half
3| 013333444
3| 555789
4| 11123
4| 599
5|
5|
6| 1              Top half
6|
7| 4
7|
8| 0
```

Q1 is the median of the bottom half of the data. Since there are 16 observations in that half, Q1 is the mean of the 8th and 9th ranked observations in that half:

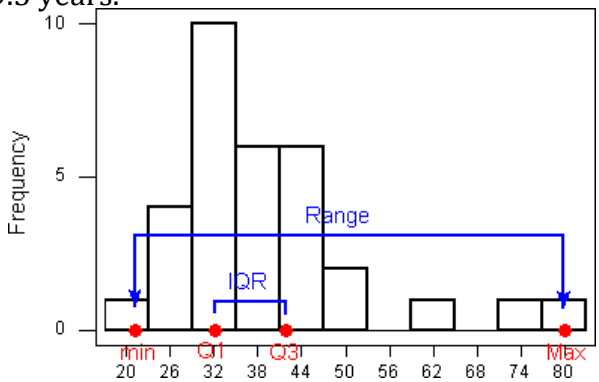
Q1 = (31 + 33) / 2 = 32

Similarly, Q3 is the median of the top half of the data, and since there are 16 observations in that half, Q3 is the mean of the 8th and 9th ranked observations in that half:

Q3 = (41 + 42) / 2 = 41.5

IQR = 41.5 – 32 = 9.5

Note that in this example, the range covered by all the ages is 59 years, while the range covered by the middle 50% of the ages is only 9.5 years. While the whole dataset is spread over a range of 59 years, the middle 50% of the datais packed into only 9.5 years.



The Normal Distribution

Many continuous random variables have a bell-shaped or somewhat symmetric distribution.

This is a normal distribution. In other words, the probability distribution of its relative frequency histogram follows a normal curve.

The curve is bell-shaped, symmetric about the mean, and defined by μ and σ (the mean and standard deviation).

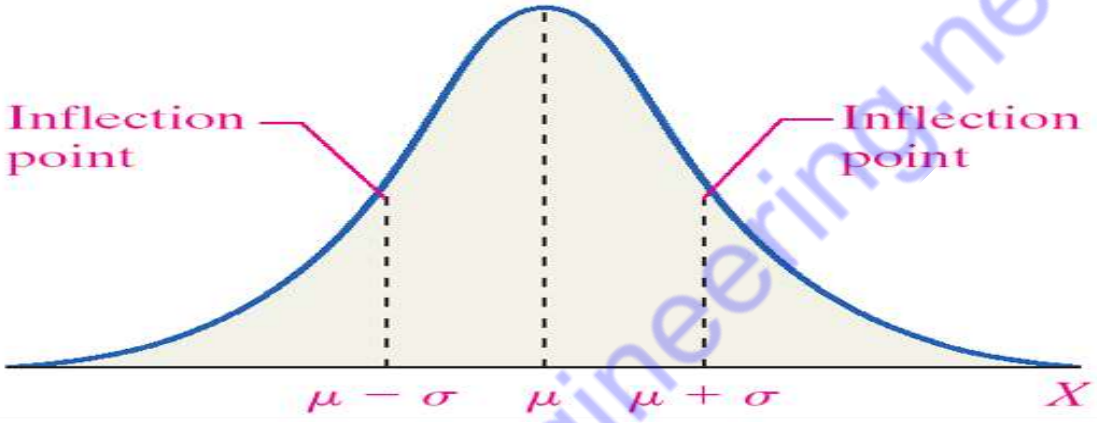


Figure 9. A normal distribution.

There are normal curves for every combination of μ and σ.

- The mean (μ) shifts the curve to the left or right.
- The standard deviation (σ) alters the spread of the curve.
- The first pair of curves have different means but the same standard deviation.
- The second pair of curves share the same mean (μ) but have different standard deviations.
- The pink curve has a smaller standard deviation. It is narrower and taller, and the probability is spread over a smaller range of values.

- The blue curve has a larger standard deviation. The curve is flatter and the tails are

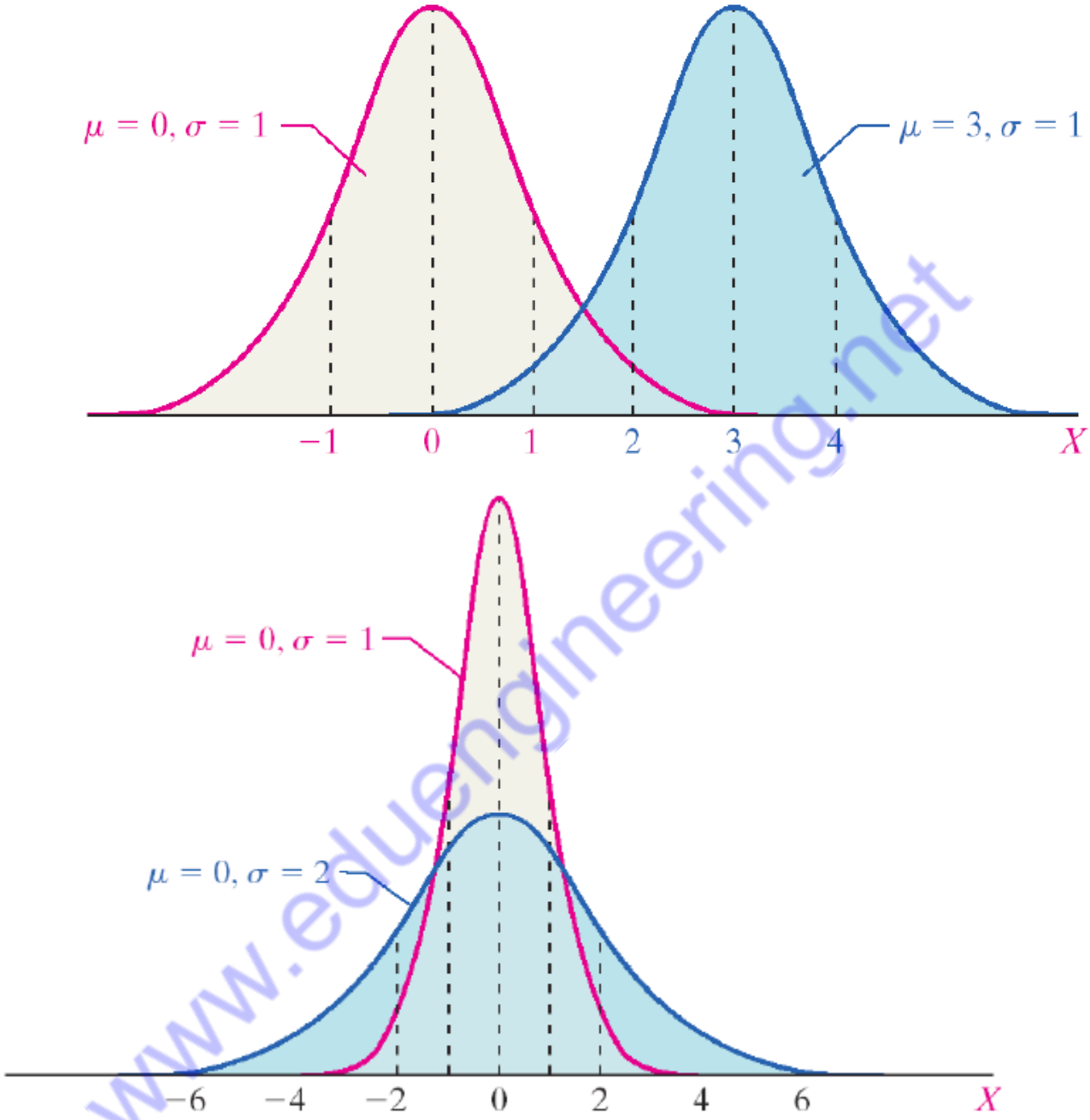


Figure 10. A comparison of normal curves.

Properties of the normal curve:

- The mean is the center of this distribution and the highest point.
- The curve is symmetric about the mean. (The area to the left of the mean equals the area to the right of the mean.)
- The total area under the curve is equal to one.
- As *x* increases and decreases, the curve goes to zero but never touches.

- The PDF of a normal curve is $y = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.
- A normal curve can be used to estimate probabilities.
- A normal curve can be used to estimate proportions of a population that have certain x-values.

The Standard Normal Distribution

There are millions of possible combinations of means and standard deviations for continuous random variables.

Finding probabilities associated with these variables would require us to integrate the PDF over the range of values we are interested in.

To avoid this, we can rely on the standard normal distribution. T

he standard normal distribution is a special normal distribution with a μ = 0 and **σ** = 1. We can use the Z-score to standardize any normal random variable, converting the x-values to Z-scores, thus allowing us to use probabilities from the standard normal table. So how do we find area under the curve associated with a Z-score?

Standard Normal Table

- The standard normal table gives probabilities associated with specific Z-scores.
- The table we use is cumulative from the left.
- The negative side is for all Z-scores less than zero (all values less than the mean).
- The positive side is for all Z-scores greater than zero (all values greater than the mean).
- Not all standard normal tables work the same way.

Example 10

What is the area associated with the Z-score 1.62?

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9595	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

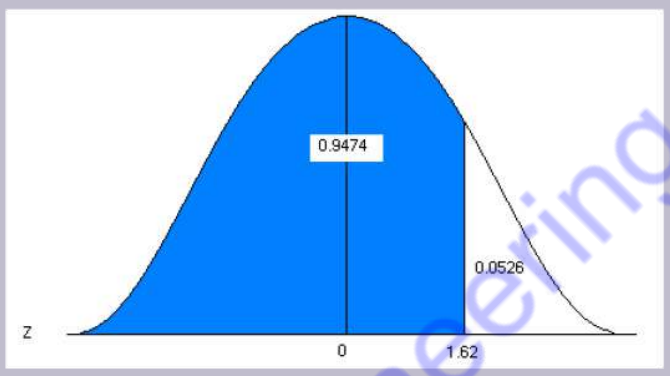


Figure 11. The standard normal table and associated area for z = 1.62.

Reading the Standard Normal Table

- Read down the Z-column to get the first part of the Z-score (1.6).
- Read across the top row to get the second decimal place in the Z-score (0.02).
- The intersection of this row and column gives the area under the curve to the left of the Z-score.

Finding Z-scores for a Given Area

- What if we have an area and we want to find the Z-score associated with that area?
- Instead of Z-score → area, we want area → Z-score.
- We can use the standard normal table to find the area in the body of values and read backwards to find the associated Z-score.
- Using the table, search the probabilities to find an area that is closest to the probability you are interested in.

Example 11

To find a Z-score for which the area to the right is 5%:

Since the table is cumulative from the left, you must use the complement of 5%.

1.000 – 0.05 = 0.9500

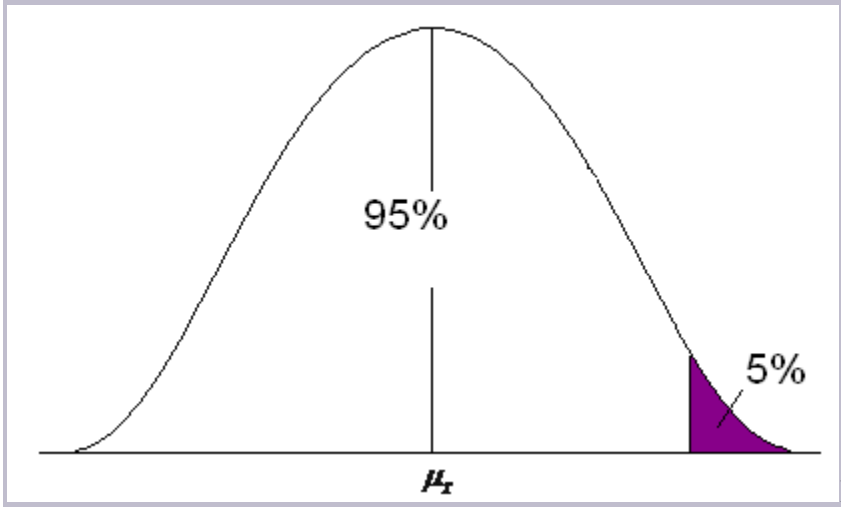


Figure 12. The upper 5% of the area under a normal curve.

- Find the Z-score for the area of 0.9500.
- Look at the probabilities and find a value as close to 0.9500 as possible.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9595	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

Figure 13. The standard normal table.

The Z-score for the 95th percentile is 1.64.**Area in between Two Z-scores**

Example 12

To find Z-scores that limit the middle 95%:

- The middle 95% has 2.5% on the right and 2.5% on the left.
- Use the symmetry of the curve.

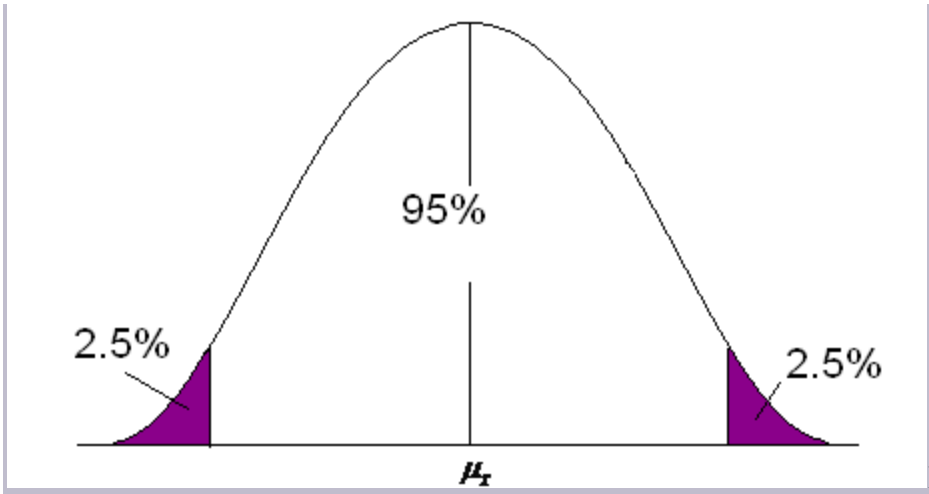


Figure 14. The middle 95% of the area under a normal curve.

- Look at your standard normal table. Since the table is cumulative from the left, it is easier to find the area to the left first.
- Find the area of 0.025 on the negative side of the table.
- The Z-score for the area to the left is -1.96.
- Since the curve is symmetric, the Z-score for the area to the right is 1.96.

Common Z-scores

There are many commonly used Z-scores:

- Z.05 = 1.645 and the area between -1.645 and 1.645 is 90%
- Z.025 = 1.96 and the area between -1.96 and 1.96 is 95%
- Z.005 = 2.575 and the area between -2.575 and 2.575 is 99%

Applications of the Normal Distribution

Typically, our normally distributed data do not have μ = 0 and **σ** = 1, but we can relate any normal distribution to the standard normal distributions using the Z-score. We can transform values of x to values of z.

$$z = \frac{x - \mu}{\sigma}$$

For example, if a normally distributed random variable has a μ = 6 and **σ** = 2, then a value of x = 7 corresponds to a Z-score of 0.5.

$$Z = \frac{7-6}{2} = 0.5$$

This tells you that 7 is one-half a standard deviation above its mean. We can use this relationship to find probabilities for any normal random variable.

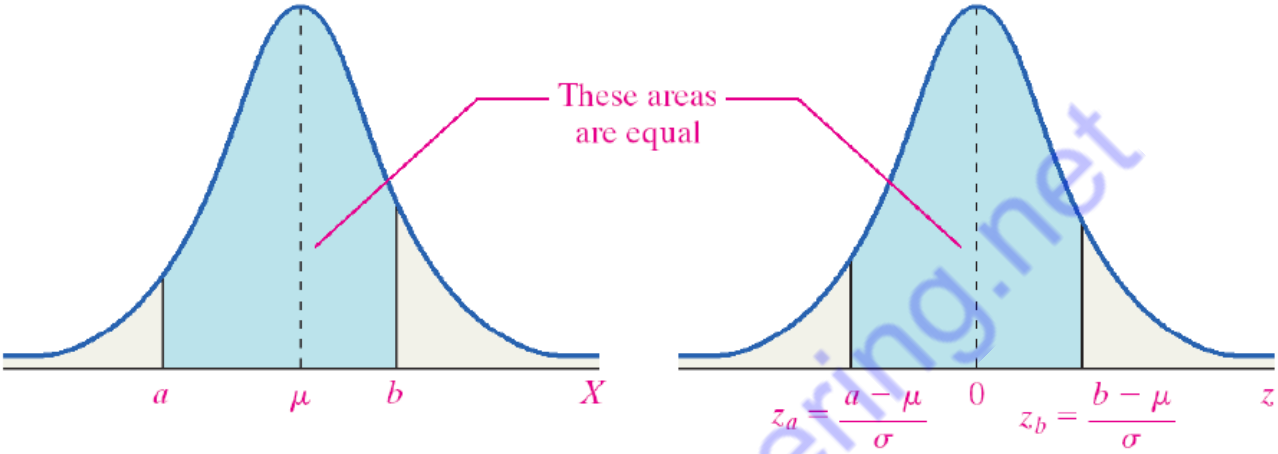


Figure 15. A normal and standard normal curve.

To find the area for values of X, a normal random variable, draw a picture of the area of interest, convert the x-values to Z-scores using the Z-score and then use the standard normal table to find areas to the left, to the right, or in between.

$$z = \frac{x-\mu}{\sigma}$$

Example 13

Adult deer population weights are normally distributed with μ = 110 lb. and **σ** = 29.7 lb. As a biologist you determine that a weight less than 82 lb. is unhealthy and you want to know what proportion of your population is unhealthy.

P(x<82)

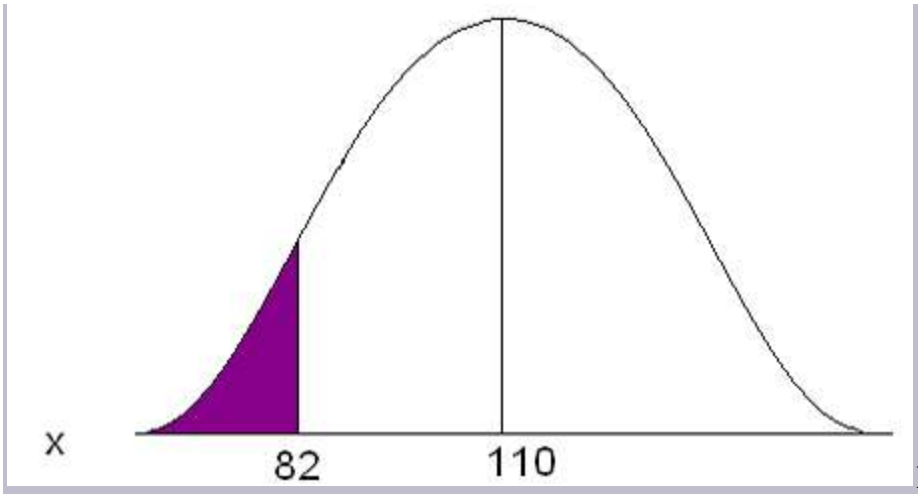


Figure 16. The area under a normal curve for P(x<82).

Convert 82 to a Z-score $$z = \frac{82 - 110}{29.7} = -0.94$$

The *x* value of 82 is 0.94 standard deviations below the mean.

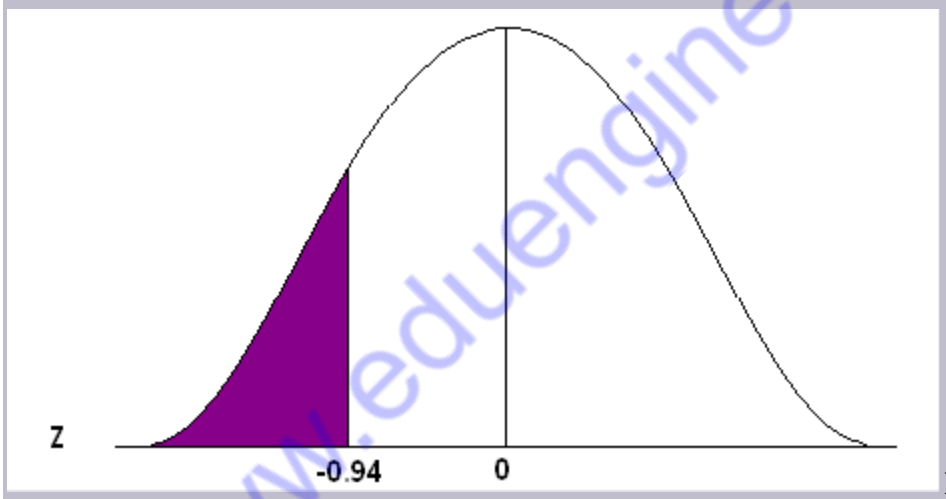


Figure 17. Area under a standard normal curve for P(z<-0.94).

Go to the standard normal table (negative side) and find the area associated with a Z-score of -0.94.

This is an "area to the left" problem so you can read directly from the table to get the probability.

P(x<82) = 0.1736

Approximately 17.36% of the population of adult deer is underweight, OR one deer chosen at random will have a 17.36% chance of weighing less than 82 lb.

Example 14

Statistics from the Midwest Regional Climate Center indicate that Jones City, which has a large wildlife refuge, gets an average of 36.7 in. of rain each year with a standard deviation of 5.1 in. The amount of rain is normally distributed. During what percent of the years does Jones City get more than 40 in. of rain?

P(x > 40)

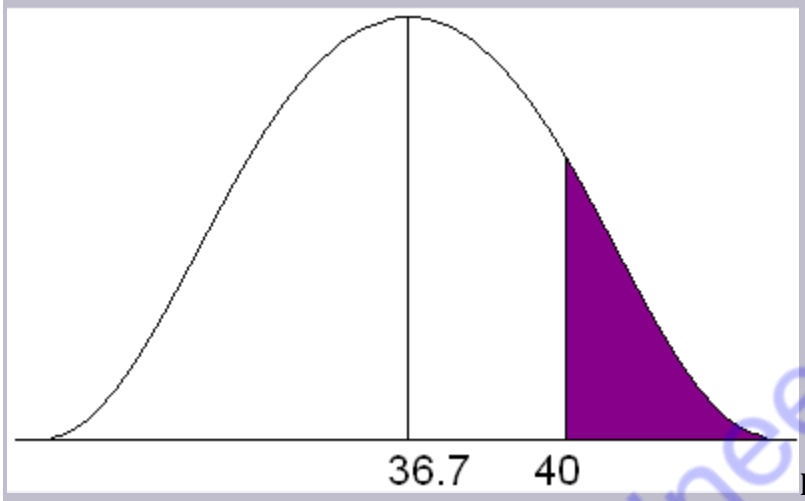


Figure 18. Area under a normal curve for P(x>40).

$$z = \frac{40 - 36.7}{5.1} = 0.65$$

P(x>40) = (1-0.7422) = 0.2578

For approximately 25.78% of the years, Jones City will get more than 40 in. of rain.

Assessing Normality

- If the distribution is unknown and the sample size is not greater than 30 (Central Limit Theorem), we have to assess the assumption of normality.
- Our primary method is the normal probability plot. This plot graphs the observed data, ranked in ascending order, against the "expected" Z-score of that rank.
- If the sample data were taken from a normally distributed random variable, then the plot would be approximately linear.
- Examine the following probability plot.
- The center line is the relationship we would expect to see if the data were drawn from a perfectly normal distribution.

- Notice how the observed data (red dots) loosely follow this linear relationship. Minitab also computes an Anderson-Darling test to assess normality.

- The null hypothesis for this test is that the sample data have been drawn from a normally distributed population. A p-value greater than 0.05 supports the assumption of normality.

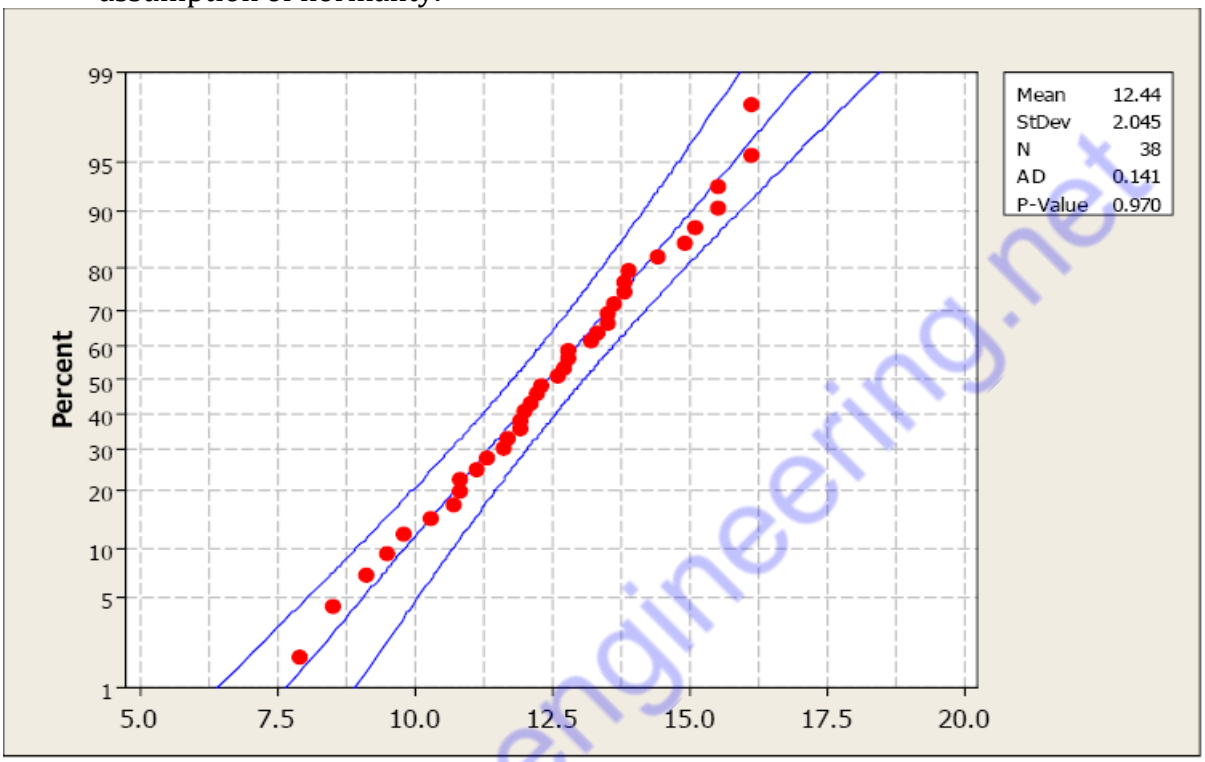


Figure 19. A normal probability plot generated using Minitab 16.

Compare the histogram and the normal probability plot in this next example. The histogram indicates a skewed right distribution.

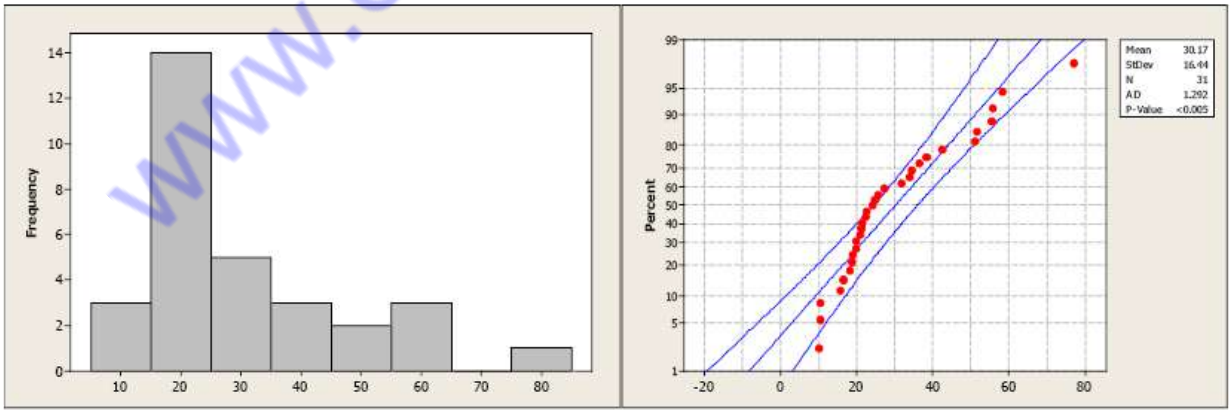


Figure 20. Histogram and normal probability plot for skewed right data.

The observed data do not follow a linear pattern and the p-value for the A-D test is less than 0.005 indicating a non-normal population distribution.

Normality cannot be assumed. You must always verify this assumption. Remember, the probabilities we are finding come from the standard NORMAL table. If our data are NOT normally distributed, then these probabilities DO NOT APPLY.

UNIT III DESCRIBING RELATIONSHIPS

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations – regression towards the mean

3.1 Scatterplot

- Define scatter plot(2M)
- Explain scatter plot with example(16M)
- How to interpret scatter plots(16M)

- The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.

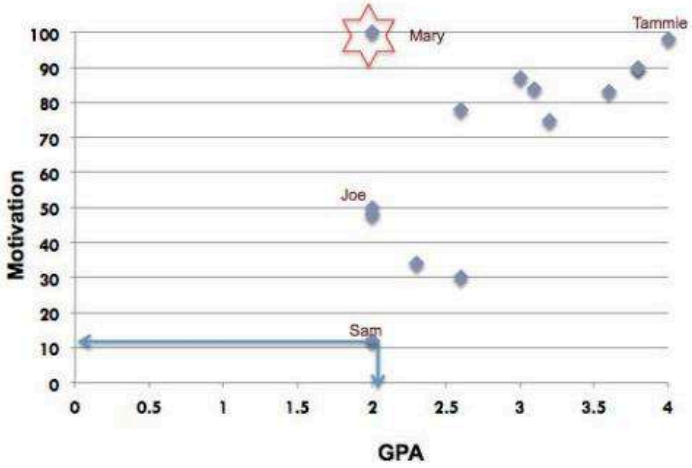
> A **scatterplot** shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the othervariable appear on the vertical axis. Each individual in the data appears as a point on the graph.

- Many research projects are **correlational studies** because they investigate the relationships that may exist between variables. Prior to investigating the relationship between two quantitative variables, it is always helpful to create a graphical representation that includes both of these variables. Such a graphical representation is called a **scatterplot**.

3.1.1 Scatterplot

What is the relationship between students' achievement motivation and GPA?

Student	Student GPA	Motivation
Joe	2.0	50
Lisa	2.0	48
Mary	2.0	100
Sam	2.0	12
Deana	2.3	34
Sarah	2.6	30
Jennifer	2.6	78
Gregory	3.0	87
Thomas	3.1	84
Cindy	3.2	75
Martha	3.6	83
Steve	3.8	90
Jamell	3.8	90
Tammie	4.0	98



- In this example, the relationship between students' achievement motivation and their GPA is being investigated.
- The table on the left includes a small group of individuals for whom GPA and scores on a motivation scale have been recorded. GPAs can range from 0 to 4 and motivation scores in this example range from 0 to 100. Individuals in this table were ordered based on their GPA.
- Simply looking at the table shows that, in general, as GPA increases, motivation scores also increase.
- However, with a real set of data, which may have hundreds or even thousands of individuals, a pattern cannot be detected by simply looking at the numbers. Therefore, a very useful strategy is to represent the two variables graphically to illustrate the relationship between them.
- A graphical representation of individual scores on two variables is called a **scatterplot**.
- The image on the right is an example of a scatterplot and displays the data from the table on the left. GPA scores are displayed on the horizontal axis and motivation scores are displayed on the vertical axis.
- Each dot on the scatterplot represents one individual from the data set. The location of each point on the graph depends on both the GPA and motivation scores. Individuals with higher GPAs are located further to the right and individuals with higher motivation scores are located higher up on the graph.
- Sam, for example, has a GPA of 2 so his point is located at 2 on the right. He also has a motivation score of 12, so his point is located at 12 going up.
- Scatterplots are not meant to be used in great detail because there are usually hundreds of individuals in a data set.
- The purpose of a scatterplot is to provide a general illustration of the relationship between the two variables.
- In this example, in general, as GPA increases so does an individual's motivation score.
- One of the students in this example does not seem to follow the general pattern: Mary. She is one of the students with the lowest GPA, but she has the maximum score on the motivation scale. This makes her an exception or an outlier.

3.1.2 Interpreting Scatterplots

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

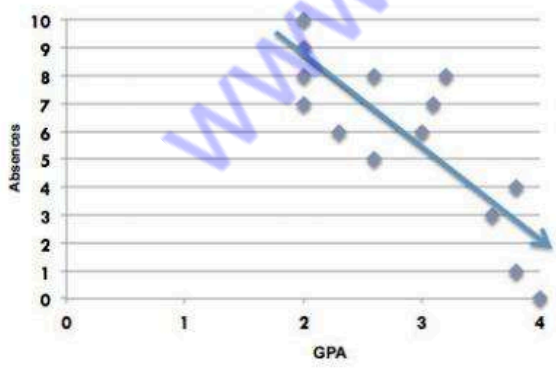
- The overall pattern of a scatterplot can be described by the **direction**, **form**, and **strength** of the relationship.
- An important kind of departure is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatterplots: Direction

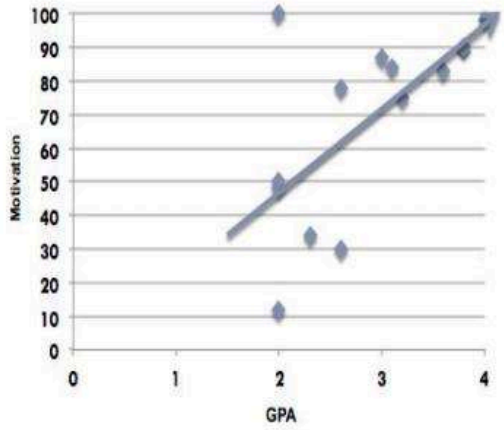
- One important component to a scatterplot is the **direction** of the relationshipbetween the two variables.

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.



This example comparesstudents' achievement
motivation and their GPA. These two variables have a **positive association** becauseas GPA increases, so does motivation.

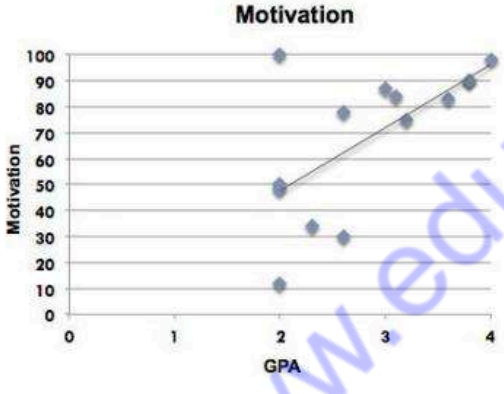


This example compares students' GPA and their numberof absences. These two variables have a **negative association** because, in general,as a student's number of absences decreases, their GPA increases

Interpreting Scatterplots: Form

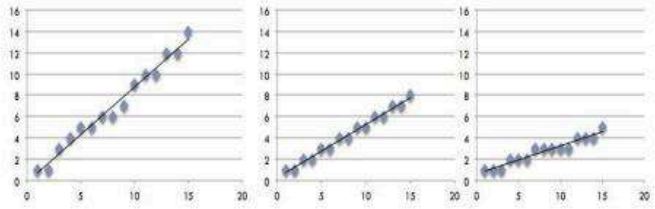
- Another important component to a scatterplot is the **form** of the relationshipbetween the two variables.

Linear relationship:

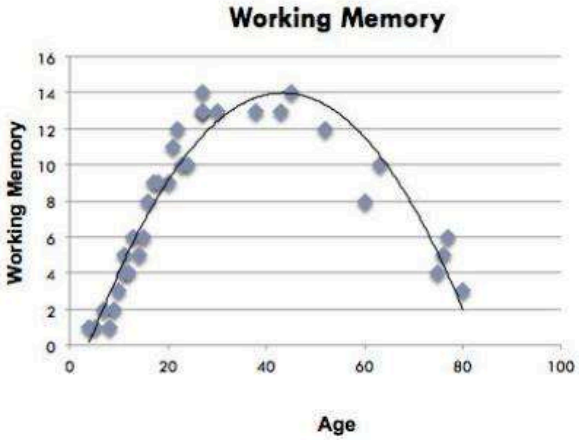


This example illustrates a linear relationship. This means that the points on the scatterplot closely resemble a straight line. A relationship is linear if one variable increases by approximately the same rate as theother variables changes by one unit.

Strong relationship: Moderate relationship: Weak relationship



Curvilinear relationship:



This example illustrates a relationship that has the form of acurve, rather than a straight line. This is due to the fact that one variable does not increase at a constant rate and may even start decreasing after a certain point.

This example describes a curvilinear relationship between the variable "age" and the variable "working memory." In this example, working memory increases throughout childhood, remains steady in adulthood, and begins decreasing around age 50.

Interpreting Scatterplots: Strength

- Another important component to a scatterplot is the **strength** of therelationship between the two variables.
- The **slope** provides information on the strength of the relationship.
- The strongest linear relationship occurs when the slope is 1. This means thatwhen one variable increases by one, the other variable also increases by the same amount. This line is at a 45 degree angle.
- The strength of the relationship between two variables is a crucial piece of information. Relying on the interpretation of a scatterplot is too subjective. More precise evidence is needed, and this evidence is obtained by computinga coefficient that measures the strength of the relationship under investigation.

Measuring Linear Association

- A scatterplot *displays* the strength, direction, and form of the relationshipbetween two quantitative variables.
- A correlation coefficient *measures* the strength of that relationship.

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

Pearson r:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- *r* is always a number between -1 and 1.
- *r* > 0 indicates a positive association.
- *r* < 0 indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1.
- The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship.

- Calculating a Pearson correlation coefficient requires the assumption that therelationship between the two variables is linear.
- There is a rule of thumb for interpreting the strength of a relationship basedon its r value (use the absolute value of the r value to make all values positive):

Absolute Value of r	Strength of Relationship
r < 0.3	None or very weak
0.3 < r <0.5	Weak
0.5 < r < 0.7	Moderate
r > 0.7	Strong

- The relationship between two variables is generally considered strong whentheir r value is larger than 0.7.

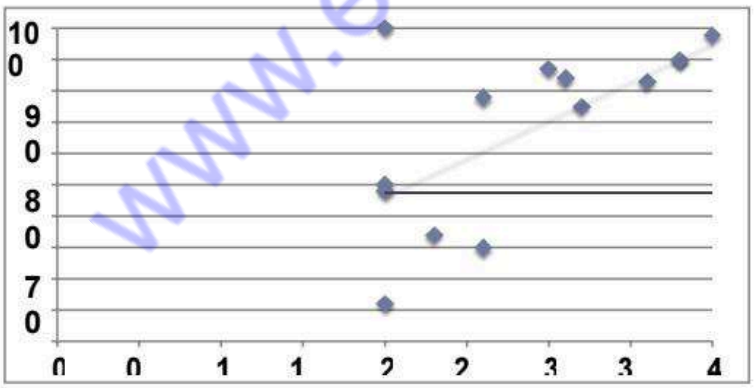
3.2 Correlations

- **What is correlations?(2M)**
- **Facts about correlations(2M)**

Example: There is a moderate, positive, linear relationship between GPA andachievement motivation.

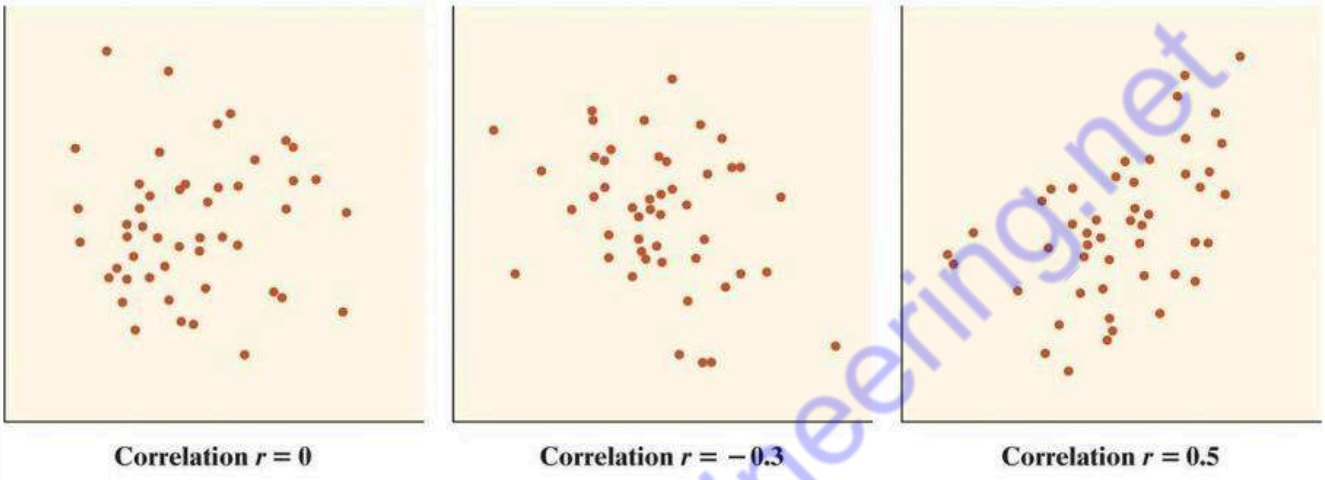
- Based on the criteria listed on the previous page, the value of r in this case (r = 0.62) indicates that there is a positive, linear relationship of **moderate** strength between achievement motivation and GPA.

r = 0.62

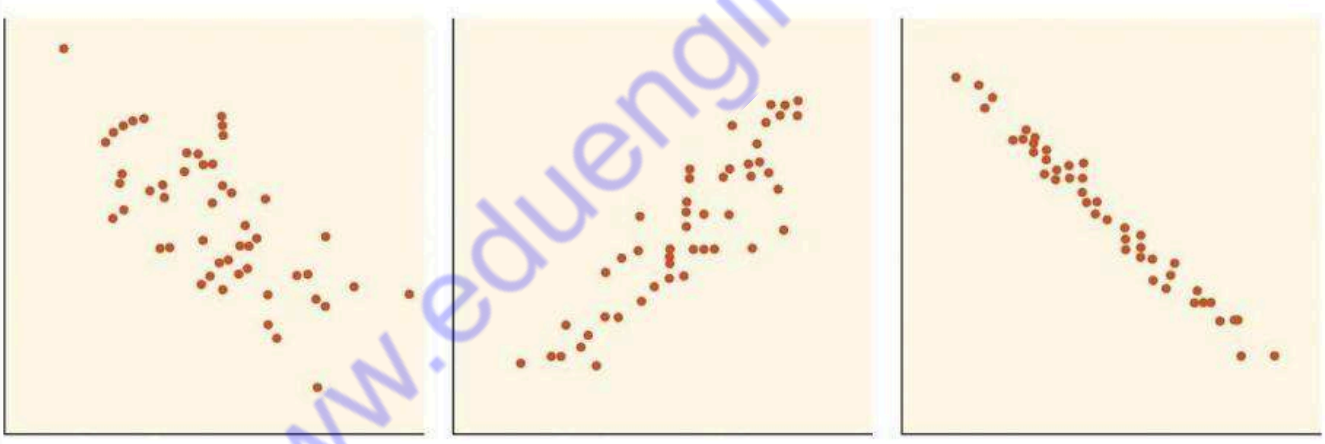


Correlation

- The images below illustrate what the relationships might look like atdifferent degrees of strength (for different values of r).



Correlation $r = 0$ | Correlation $r = -0.3$ | Correlation $r = 0.5$



Correlation $r = -0.7$ | Correlation $r = 0.9$ | Correlation $r = -0.99$

- For a correlation coefficient of zero, the points have no direction, the shape isalmost round, and a line does not fit to the points on the graph.
- As the correlation coefficient increases, the observations group closertogether in a linear shape.
- The line is difficult to detect when the relationship is weak (e.g., r = -0.3), butbecomes more clear as relationships become stronger (e.g., r = -0.99)

Facts About Correlation

1) The order of variables in a correlation is not important.

2) Correlations provide evidence of association, not causation.

3) *r* has no units and does not change when the units of measure of *x*, *y*, or bothare changed.

4) Positive *r* values indicate positive association between the variables, andnegative *r* values indicate negative associations.

5) The correlation *r* is always a number between -1 and 1.

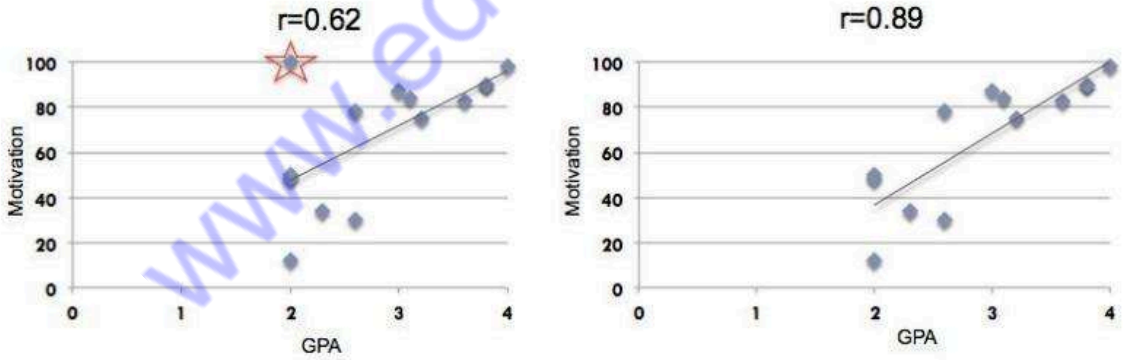
Pearson *r*: Assumptions

Assumptions:

- Correlation requires that both variables be quantitative.
- Correlation describes *linear* relationships. Correlation does not describe curve relationships between variables, no matter how strong the relationshipis.

Cautions:

- Correlation is not resistant. *r* is strongly affected by outliers.
- Correlation is not a complete summary of two-variable data.
- For example:



- The correlation coefficient is based on means and standard deviations, so it isnot robust to outliers; it is strongly affected by extreme observations. These individuals are sometimes referred to as *influential observations* because they have a strong impact on the correlation coefficient.
- For instance, in the above example the correlation coefficient is 0.62 on the left when the outlier is included in the analysis. However, when this outlier is removed, the correlation coefficient increases significantly to 0.89.
- This one case, when included in the analysis, reduces a strong relationship toa moderate relationship.
- This case makes such a big difference in this example because the data set contains a very small number of individuals. As a general rule, as the size ofthe sample increases, the influence of extreme observations decreases.
- When describing the relationship between two variables, correlations are just one piece of the puzzle. This information is necessary, but not sufficient.Other analyses should also be conducted to provide more information.

CORRELATION COEFFICIENT:

- What does a correlation coefficient tell you?(2M)
- Significance of correlation coefficient(2M)
- How to interpret correlation coefficient?(8M)
- Types of Correlation coefficient(16M)
- What are the assumptions our data has to meet for pearson'sr?(2M)
- Give Pearson's r formula with explaination(2M)
- Give spearmans's rho formula(2M)

A correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables. In other words, it reflects how similar the measurements of two or more variables are across a dataset.

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.

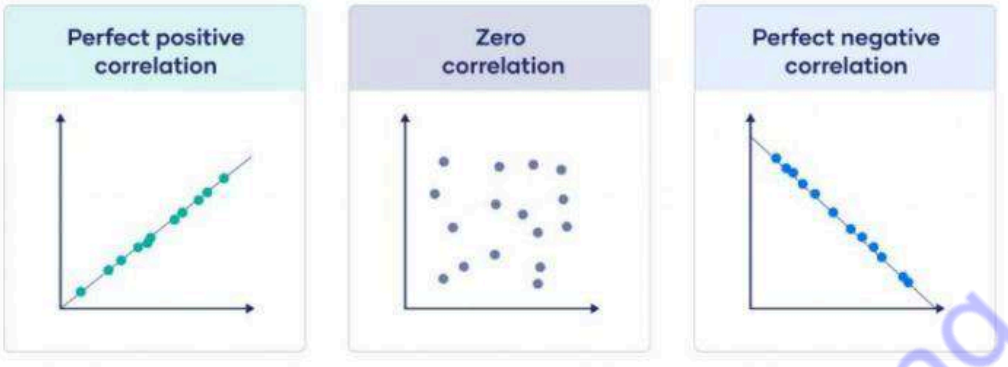


Figure : Co-relation

3.3 Correlation Coefficients

The Statistical Significance of Correlation Coefficients:

- Correlation coefficients have a probability (p-value), which shows **the probability that the relationship between the two variables is equal tozero** (null hypotheses; no relationship).
- **Strong** correlations have **low** p-values because the probability that they haveno relationship is very low.
- Correlations are typically considered statistically significant if the p-value islower than 0.05 in the social sciences, but the researcher has the liberty to decide the p-value for which he or she will consider the relationship to be significant.
- The value of p for which a correlation will be considered statisticallysignificant is called the **alpha level** and must be reported.
- SPSS notation for p values: Sig. (2 tailed)

In the previous example, r = 0.62 and p-value = 0.03. The p-value of 0.03 is less thanthe acceptable alpha level of 0.05, meaning the correlation is statistically significant.

Four things must be reported to describe a relationship:

1) The **strength** of the relationship given by the correlation coefficient.
2) The **direction** of the relationship, which can be positive or negative based onthe sign of the

correlation coefficient.

3) The **shape** of the relationship, which must always be linear to computer aPearson correlation coefficient.
4) Whether or not the relationship is **statistically significant**, which is basedon the p-value

What does a correlation coefficient tell you?

Correlation coefficients summarize data and help you compare results between studies.

Summarizing data

A correlation coefficient is a descriptive statistic. That means that it summarizes sample data without letting you infer anything about the population. A correlation coefficient is a bivariate statistic when it summarizes the relationship between two variables, and it's a multivariate statistic when you have more than two variables.

If your correlation coefficient is based on sample data, you'll need an inferential statistic if you want to generalize your results to the population. You can use an F test or a t test to calculate a test statistic that tells you the statistical significance of your finding.

Comparing studies

A correlation coefficient is also an effect size measure, which tells you the practical significance of a result. Correlation coefficients are unit-free, which makes it possible to directly compare coefficients between studies.

Using a correlation coefficient

In correlational research, you investigate whether changes in one variable are associated with changes in other variables.

Correlational research example

You investigate whether standardized scores from high school are related to academic grades in college. You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

After data collection, you can visualize your data with a scatterplot by plotting one variable on the x-axis and the other on the y-axis. It doesn't matter which variable you place on either axis.

Visually inspect your plot for a pattern and decide whether there is a linear or non-linear pattern between variables. A linear pattern means you can fit a straight line of best fit between the data points, while a non-linear or curvilinear pattern can take all sorts of different shapes, such as a U-shape or a line with a curve.

Visual inspection example

You gather a sample of 5,000 college graduates and survey them on their high school SAT scores and college GPAs. You visualize the data in a scatterplot to check for a linear

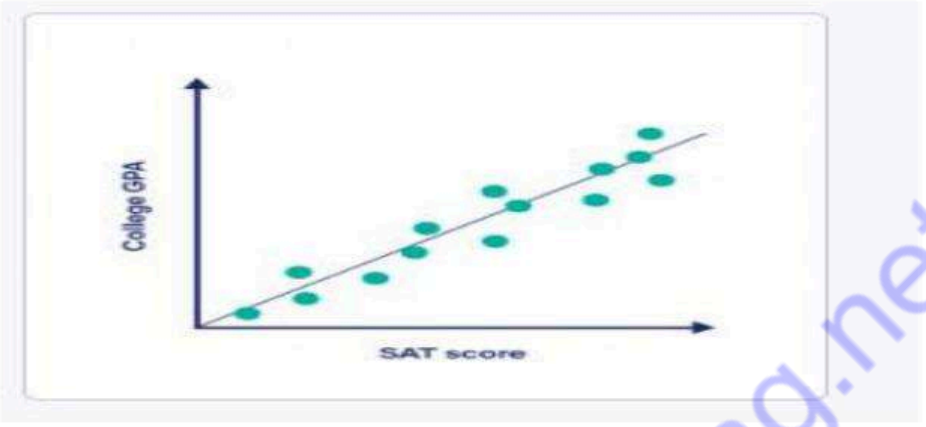


Figure :

There are many different correlation coefficients that you can calculate. After removing any outliers, select a correlation coefficient that's appropriate based on the general shape of the scatter plot pattern. Then you can perform a correlation analysis to find the correlation coefficient for your data.

You calculate a correlation coefficient to summarize the relationship between variables without drawing any conclusions about causation.

Correlation analysis example

You check whether the data meet all of the assumptions for the Pearson's r correlation test.
Both variables are quantitative and normally distributed with no outliers, so you calculate a Pearson's r correlation coefficient.

The correlation coefficient is strong at .58

Interpreting a correlation coefficient

The value of the correlation coefficient always ranges between 1 and -1, and you treat it as a general indicator of the strength of the relationship between variables.

The sign of the coefficient reflects whether the variables change in the same or opposite directions: a positive value means the variables change together in the same direction, while a negative value means they change together in opposite directions.

The absolute value of a number is equal to the number without its sign. The absolute value of a correlation coefficient tells you the magnitude of the correlation: the greater the absolute value, the stronger the correlation.

There are many different guidelines for interpreting the correlation coefficient because findings can vary a lot between study fields. You can use the table below as a general guideline for interpreting correlation strength from the value of the correlation coefficient.

While this guideline is helpful in a pinch, it's much more important to take your research context and purpose into account when forming conclusions. For example, if most studies in your field have correlation coefficients nearing .9, a correlation coefficient of .58 may be low in that context.

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

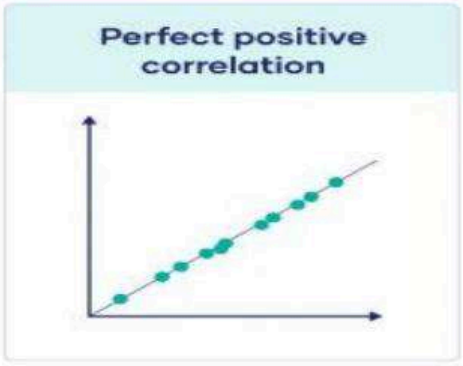
Table :

Visualizing linear correlations

The correlation coefficient tells you how closely your data fit on a line. If you have a linear relationship, you'll draw a straight line of best fit that takes all of your data points into account on a scatter plot.

The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

If all points are perfectly on this line, you have a perfect correlation.



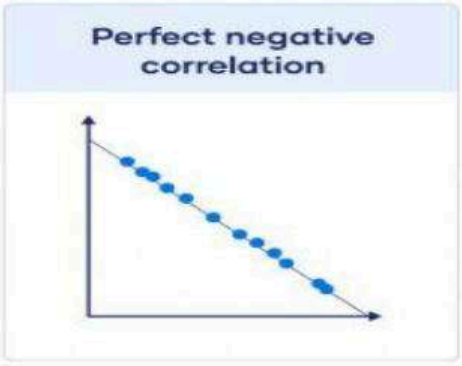
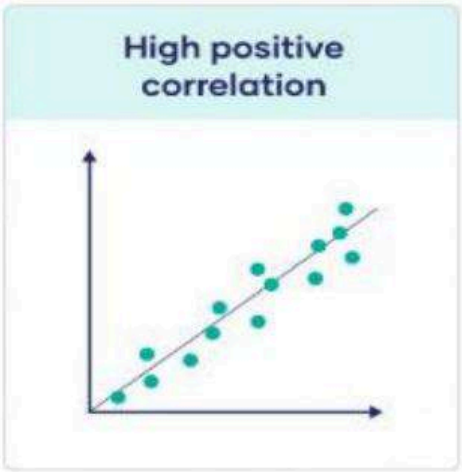


Figure :

If all points are close to this line, the absolute value of your correlation coefficient is high.



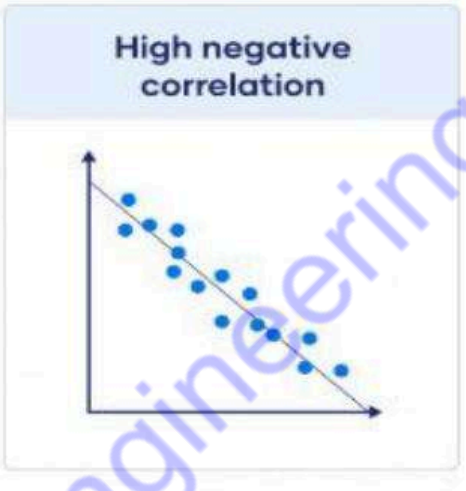
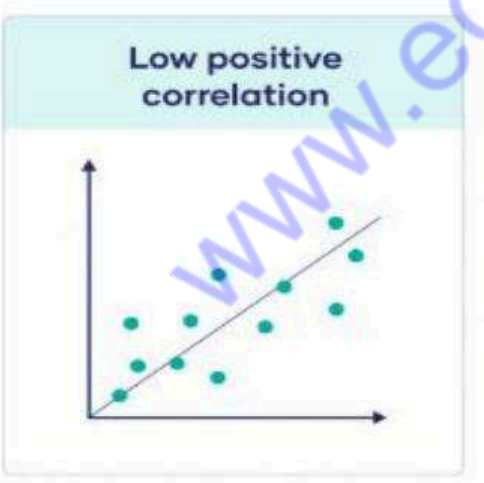


Figure:

If these points are spread far from this line, the absolute value of your correlation coefficient is low.



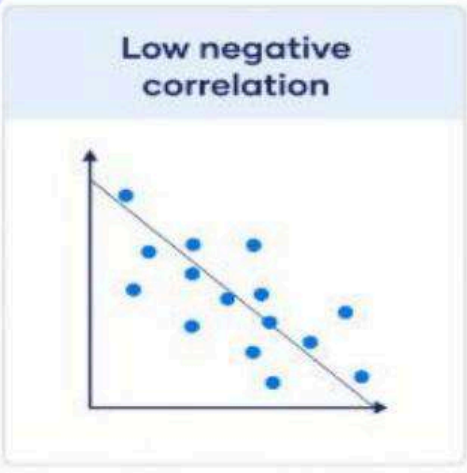
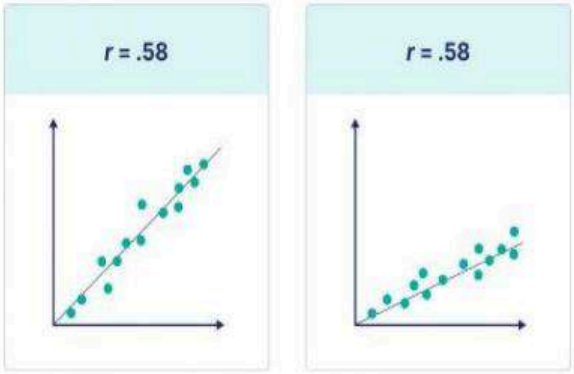


Figure:

Note that the steepness or slope of the line isn't related to the correlation coefficient value. The correlation coefficient doesn't help you predict how much one variable will change based on a given change in the other, because two datasets with the same correlation coefficient value can have lines
with very different slopes.



3.3.1. Types of correlation coefficients

You can choose from many different correlation coefficients based on the linearity of the relationship, the level of measurement of your variables, and the distribution of your data.

For high statistical power and accuracy, it's best to use the correlation coefficient that's most appropriate for your data.

The most commonly used correlation coefficient is Pearson's r because it allows for strong inferences. It's parametric and measures linear relationships. But if your data do not meet all assumptions for this test, you'll need to use a non-parametric test instead.

Non-parametric tests of rank correlation coefficients summarize non-linear relationships between variables. The Spearman's rho and Kendall's tau have the same conditions for use, but Kendall's tau is generally preferred for smaller samples whereas Spearman's rho is more widely used.

The table below is a selection of commonly used correlation coefficients, and we'll cover the two most widely used coefficients in detail in this article.

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Table

3.3.1.1 Pearson's r

The Pearson's product-moment correlation coefficient, also known as Pearson's r, describes the linear relationship between two quantitative variables.

These are the assumptions your data must meet if you want to use Pearson's r:

- Both variables are on an interval or ratio level of measurement
- Data from both variables follow normal distributions
- Your data have no outliers
- Your data is from a random or representative sample
- You expect a linear relationship between the two variables

The Pearson's r is a parametric test, so it has high power. But it's not a good measure of correlation if your variables have a nonlinear relationship, or if your data have outliers, skewed distributions, or come from categorical variables. If any of these assumptions are violated, you should consider a rank correlation measure.

The formula for the Pearson's r is complicated, but most computer programs can quickly churn out the correlation coefficient from your data. In a simpler form, the formula divides the covariance between the variables by the product of their standard deviations.

Formula	Explanation
$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$	• r_{xy} = strength of the correlation between variables x and y • n = sample size • $\sum$ = sum of what follows... • X = every x-variable value • Y = every y-variable value • XY = the product of each x-variable score and the corresponding y-variable score

3.3.1.1.1 Pearson sample vs population correlation coefficient formula

When using the Pearson correlation coefficient formula, you'll need to consider whether you're dealing with data from a sample or the whole population.

The sample and population formulas differ in their symbols and inputs. A sample correlation coefficient is called r, while a population correlation coefficient is called rho, the Greek letter ρ.

The sample correlation coefficient uses the sample covariance between variables and their sample standard deviations.

Sample correlation coefficient formula	Explanation
$r_{xy} = \frac{cov(x,y)}{s_x s_y}$	• r_{xy}= strength of the correlation between variables x and y • $cov(x,y)$ = covariance of x and y • s_x = sample standard deviation of x • s_y = sample standard deviation of y

The population correlation coefficient uses the population covariance between variables and their population standard deviations.

Population correlation coefficient formula	Explanation
$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	• ρ_{XY}= strength of the correlation between variables X and Y • $cov(X,Y)$ = covariance of X and Y • σ_X = population standard deviation of X • σ_Y = population standard deviation of Y

3.3.1.2 Spearman's rho

Spearman's rho, or Spearman's rank correlation coefficient, is the most common alternative to Pearson's r. It's a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

You should use Spearman's rho when your data fail to meet the assumptions of Pearson's r. This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.

While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

In a linear relationship, each variable changes in one direction at the same rate throughout the data range. In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate.

Positive monotonic: when one variable increases, the other also increases.
Negative monotonic: when one variable increases, the other decreases.
Monotonic relationships are less restrictive than linear relationships.

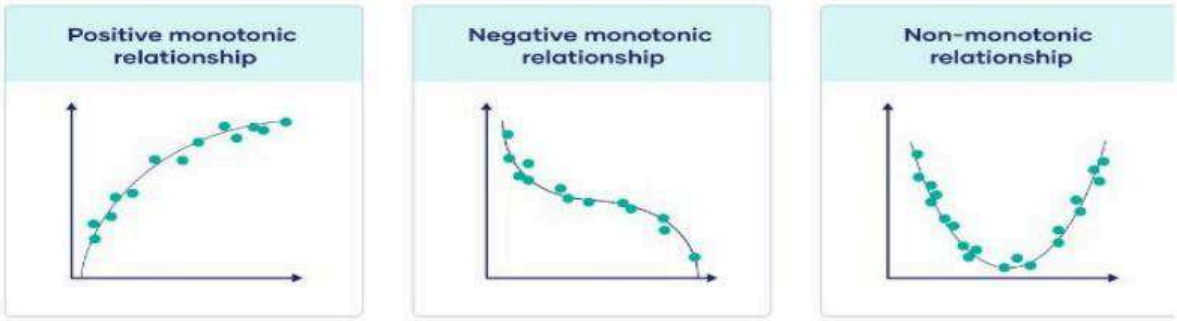


Figure :

3.3.1.2.1 Spearman's rank correlation coefficient formula

The symbols for Spearman's rho are ρ for the population coefficient and rs for the sample coefficient. The formula calculates the Pearson's r correlation coefficient between the rankings of the variable data.

To use this formula, you'll first rank the data from each variable separately from low to high: every datapoint gets a rank from first, second, or third, etc.

Then, you'll find the differences (di) between the ranks of your variables for each data pair and take that as the main input for the formula.

Spearman's rank correlation coefficient formula	Explanation
$r_s = 1 - \frac{6\sum d_i^2}{(n^3 - n)}$	• r_s= strength of the rank correlation between variables • d_i = the difference between the x-variable rank and the y-variable rank for each pair of data • $\sum d^2_i$ = sum of the squared differences between x- and y-variable ranks • n = sample size

If you have a correlation coefficient of 1, all of the rankings for each variable match up for every data pair. If you have a correlation coefficient of -1, the rankings for one variable are the exact opposite of the ranking of the other variable. A correlation coefficient near zero means that there's no monotonic relationship between the variable rankings.

The Least Squares Regression Line

Goodness of Fit of a Straight Line to Data

Once the scatter diagram of the data has been drawn and the model assumptions described in the previous sections at least visually verified (and perhaps the correlation coefficient r computed to quantitatively verify the linear trend), the next step in the analysis is to find the straight line that best fits the data. We will explain how to measure how well a straight line fits a collection of points by examining how well the line y=12x−1 fits the data set

x	2	2	6	8	10
y	0	1	2	3	3

(which will be used as a running example for the next three sections). We will write the equation of this line as $\hat{y}=12x-1$ with an accent on the y to indicate that the y-values computed using this equation are not from the data. We will do this with all lines approximating data sets. The line $\hat{y}=12x-1$ was selected as one that seems to fit the data reasonably well.

The idea for measuring the goodness of fit of a straight line to data is illustrated in Figure 10.6 "Plot of the Five-Point Data and the Line ", in which the graph of the line $\hat{y}=12x-1$ has been superimposed on the scatter plot for the sample data set.

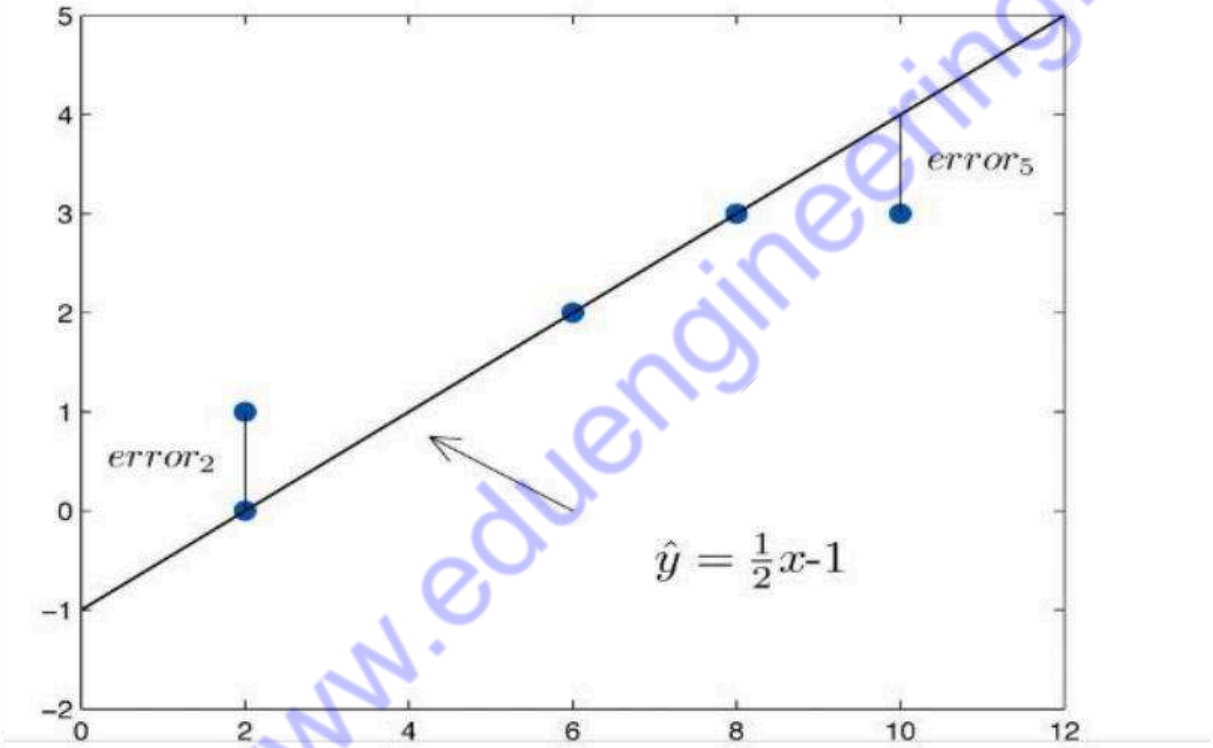


Figure Plot of the Five-Point Data and the Line $\hat{y}=12x-1$

To each point in the data set there is associated an “error,” the positive or negative vertical distance from the point to the line: positive if the point is above the line and negative if it is below the line. The error can be computed as the actual y-value of the point minus the y-value $\hat{y}$ that is “predicted” by inserting the x-value of the data point into the formula for the line:

$$\text{erroratdatapoint}(x,y)=(\text{truey})-(\text{predictedy})=y-\hat{y}$$

	x	y	$\hat{y} = \frac{1}{2}x - 1$	$y - \hat{y}$	$(y - \hat{y})^2$
	2	0	0	0	0
	2	1	0	1	1
	6	2	2	0	0
	8	3	3	0	0
	10	3	4	−1	1
Σ	-	-	-	0	2

Table The Errors in Fitting Data with a Straight Line

A first thought for a measure of the goodness of fit of the line to the data would be simply to add the errors at every point, but the example shows that this cannot work well in general. The line does not fit the data perfectly (no line can), yet because of cancellation of positive and negative errors the sum of the errors (the fourth column of numbers) is zero. Instead goodness of fit is measured by the sum of the squares of the errors. Squaring eliminates the minus signs, so no cancellation can occur. For the data and line in Figure 10.6 "Plot of the Five-Point Data and the Line " the sum of the squared errors (the last column of numbers) is 2. This number measures the goodness of fit of the line to the data.

Definition

The **goodness of fit** *of a line* $\hat{y}=mx+b$ *to a set of n pairs* (x,y) *of numbers in a sample is the sum of the squared errors*

$$\Sigma(y-\hat{y})^2$$

(n terms in the sum, one for each data pair).

The Least Squares Regression Line

Given any collection of pairs of numbers (except when all the x-values are the same) and the corresponding scatter diagram, there always exists exactly one straight line that fits the data better than any other, in the sense of minimizing the sum of the squared errors. It is called the least squares regression line. Moreover there are formulas for its slope and y-intercept.

Definition

Given a collection of pairs (x,y)(x,y) *of numbers (in which not all the x-values are the same), there is a line* y^=β^1x+β^0y^=β^1x+β^0 *that best fits the data in the sense of minimizing the sum of the squared errors. It is called the* least squares regression line. *Its slope* β^1β^1 *and y-intercept* β^0β^0 *are computed using the formulas*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2, \quad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)$$

x— — x- *is the mean of all the x-values,* y— — y- *is the mean of all the y-values, and n is the number of pairs in the data set.*

The equation y^=β^1x+β^0y^=β^1x+β^0 *specifying the least squares regression line is called the* least squares regression equation.

Remember from Section 10.3 "Modelling Linear Relationships with Randomness Present" that the line with the equation y=β1x+β0 is called the population regression line. The numbers β^1 and β^0 are statistics that estimate the population parameters β1 and β

EXAMPLE 1

Find the least squares regression line for the five-point data set and verify that it fits the data better than the line $\hat{y}=12x-1$ considered in Section 10.4.1 "Goodness of Fit of a Straight Line to Data".

x	2	2	6	8	10
y	0	1	2	3	3

Solution:

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.

	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
Σ	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 208 - \frac{1}{5}(28)^2 = 51.2$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) = 68 - \frac{1}{5}(28)(9) = 17.6$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{5} = 5.6$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{9}{5} = 1.8$$

so that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375 \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.8 - (0.34375)(5.6) = -0$$

The least squares regression line for these data is

$$\hat{y} = 0.34375x - 0.125$$

The computations for measuring how well it fits the sample data are given in Table 10.2 "The Errors in Fitting Data with the Least Squares Regression Line". The sum of the squared errors is the sum of the numbers in the last column, which is 0.75. It is less than 2, the sum of the squared errors for the fit of the line y^=12x−1 to this data set.

THE ERRORS IN FITTING DATA WITH THE LEAST SQUARES REGRESSION LINE

x	y	$\hat{y} = 0.34375x - 0.125$	$y - \hat{y}$	$(y - \hat{y})^2$
2	0	0.5625	−0.5625	0.31640625
2	1	0.5625	0.4375	0.19140625
6	2	1.9375	0.0625	0.00390625
8	3	2.6250	0.3750	0.14062500
10	3	3.3125	−0.3125	0.09765625

3.4 Regression

3.4.1 What is Regression?

- ❖ Define regression with example(2M,8M)
- ❖ Application of regression in real life(2M)

Regression allows researchers to predict or explain the variation in one variable based on another variable.

The variable that researchers are trying to explain or predict is called the response variable. It is also sometimes called the dependent variable because it depends on another variable.

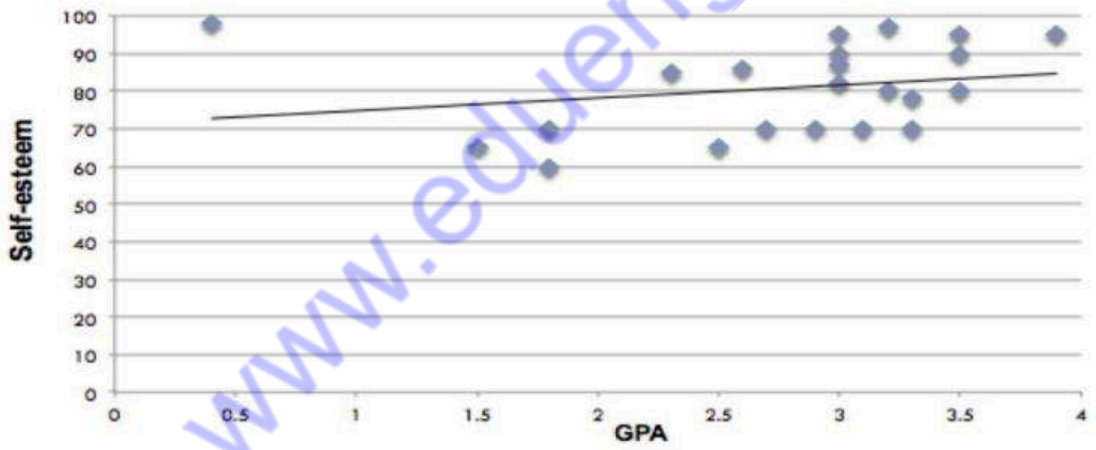
The variable that is used to explain or predict the response variable is called the explanatory variable. It is also sometimes called the independent variable because it is independent of the other variable.

In regression, the order of the variables is very important. The explanatory variable (or the independent variable) always belongs on the x-axis. The response variable (or the dependent variable) always belongs on the y-axis.

Example:

If it is already known that there is a significant correlation between students' GPA and their self-esteem, the next question researchers might ask is: Can students' scores on a self-esteem scale be predicted based on GPA? In other words, does GPA explain self-esteem? These are the types of questions that regression responds to.

**Note that these questions do not imply a causal relationship. In this example, GPA is the explanatory variable (or the independent variable) and self-esteem is the response variable (or the dependent variable). GPA belongs on the x-axis and self-esteem belongs on the y-axis.



Regression is essential for any machine learning problem that involves continuous numbers, which includes a vast array of real-life applications:

1. Financial forecasting, such as estimating housing or stock prices
2. Automobile testing
3. Weather analysis
4. Time series forecasting

3.4.2 Types of Regression

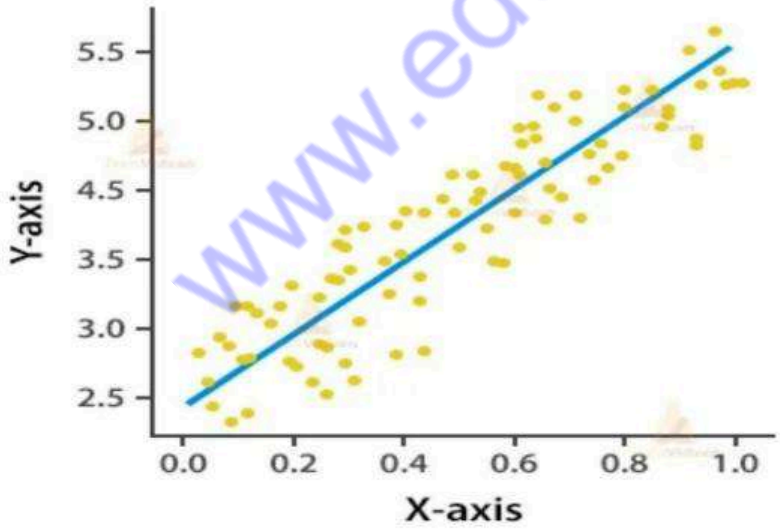
- ❖ **Types of regression(2M,16M)**
- ❖ **What are the three approaches in stepwise regression?(2M)**

- ➢ Linear Regression
- ➢ Logistic Regression
- ➢ Polynomial Regression
- ➢ Stepwise Regression
- ➢ Ridge Regression
- ➢ Lasso Regression
- ➢ Elastic Net Regression

3.4.2 .1 LINEAR REGRESSION:

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

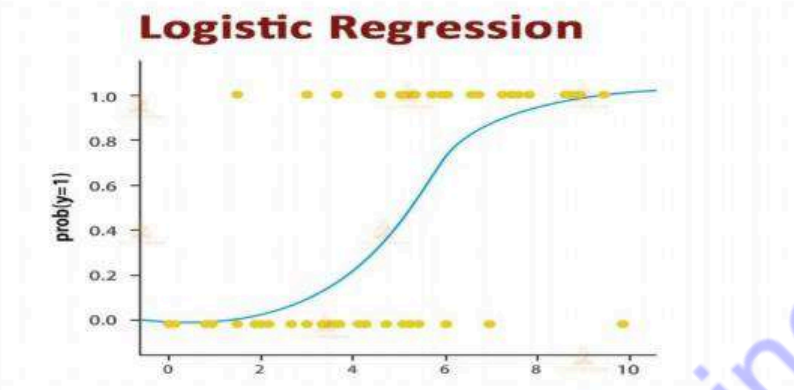


3.4.2.2 LOGISTIC REGRESSION:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college acceptance, the logistic function could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into one of two outcome categories.

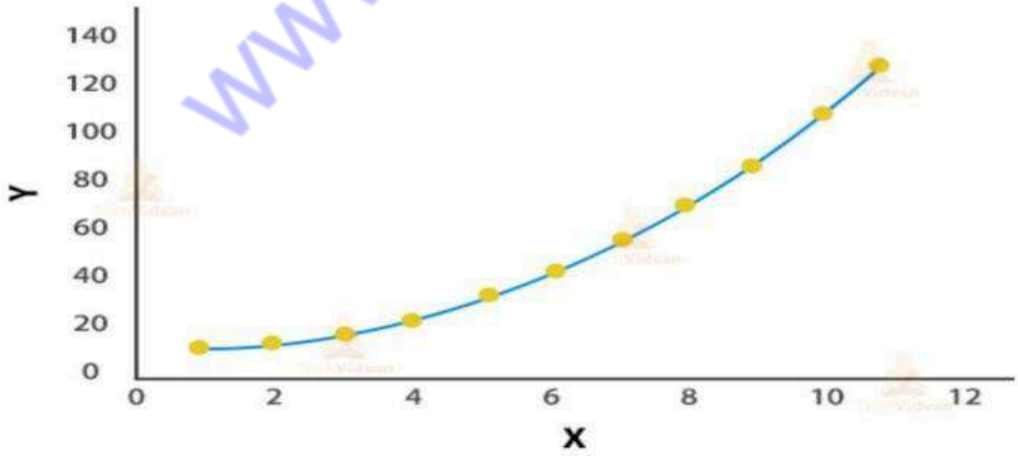


3.4.2.3 POLYNOMIAL REGRESSION:

In a polynomial regression, the power of the independent variable is more than 1. The equation below represents a polynomial equation:

$$y = a + bx2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



3.4.2.4 STEPWISE REGRESSION:

Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration.

The availability of statistical software packages makes stepwise regression possible, even in models with hundreds of variables.

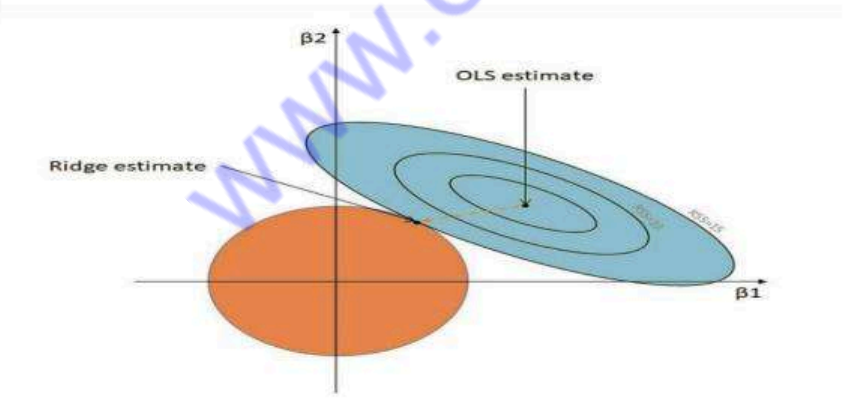
The underlying goal of stepwise regression is, through a series of tests (e.g. F-tests, t-tests) to find a set of independent variables that significantly influence the dependent variable.

there are three approaches to stepwise regression:

- **Forward selection** begins with no variables in the model, tests each variable as it is added to the model, then keeps those that are deemed most statistically significant—repeating the process until the results are optimal.
- **Backward elimination** starts with a set of independent variables, deleting one at a time, then testing to see if the removed variable is statistically significant.
- **Bidirectional elimination** is a combination of the first two methods that test which variables should be included or excluded.

3.4.2.5 RIDGE REGRESSION:

Ridge regression is a type of linear regression technique that is used in machine learning to reduce the overfitting of linear models. Recall that Linear regression is a method of modeling data that represents relationships between a response variable and one or more predictor variables. Ridge regression is used when there are multiple variables that are highly correlated. It helps to prevent overfitting by penalizing the coefficients of the variables. Ridge regression reduces the overfitting by adding a penalty term to the error function that shrinks the size of the coefficients. The penalty term is called the **L2 norm**.Ridge regression is similar to ordinary least squares regression, but the penalty term ensures that the coefficients do not become too large. This can be beneficial when there is a lot of noise in the data, as it prevents the model from being too sensitive to individual data points.



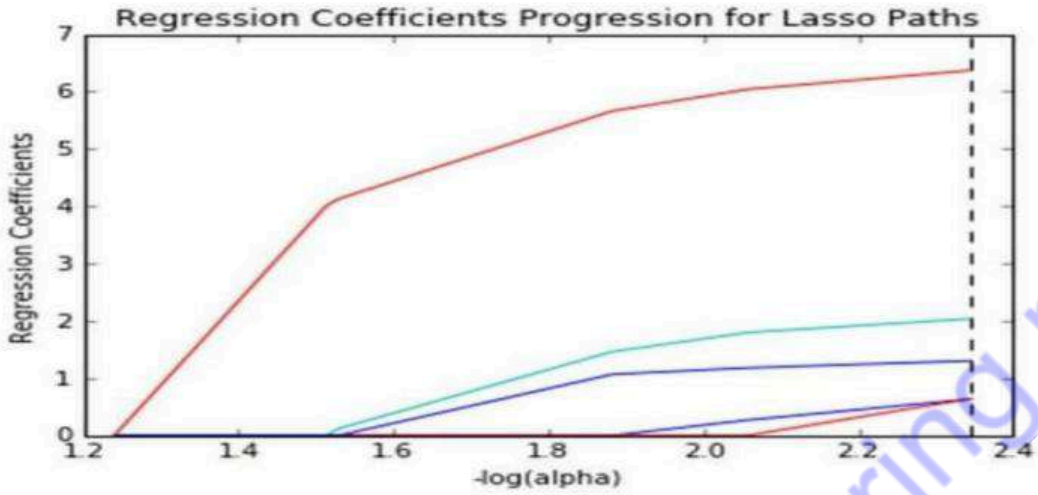
Below is the equation used to denote the Ridge Regression, λ (lambda) resolves the multicollinearity issue:

$$\beta = (X^{T}X + \lambda*I)^{-1}X^{T}y$$

3.4.2.6 LASSO REGRESSION:

The acronym "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.

In short, Lasso Regression is like Ridge Regression regarding its use. However, the only difference is that the data is being fed is not normal. In the case of Lasso Regression, only the required parameters are used, and the rest is made zero. This helps avoid the overfitting in the model. But if independent variables are highly collinear, then Lasso regression chooses only one variable and makes other variables reduce to zero.



3.4.2.7 Elastic Net Regression

Elastic Net regression is being utilized in the case of dominant independent variables being more than one amongst many correlated independent variables.

Also, seasonality & time value factors are made to work together to identify the type of regression.

Elastic Net Regression is a combination of Lasso Regression and Ridge Regression methods. It is prepared with L1 and L2 earlier as regularizer.

The equation represents as:

ElasticNet Regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

A clear advantage of trade-off among Lasso and Ridge is that it permits Elastic-Net to acquire a portion of Ridge's dependability under rotation.

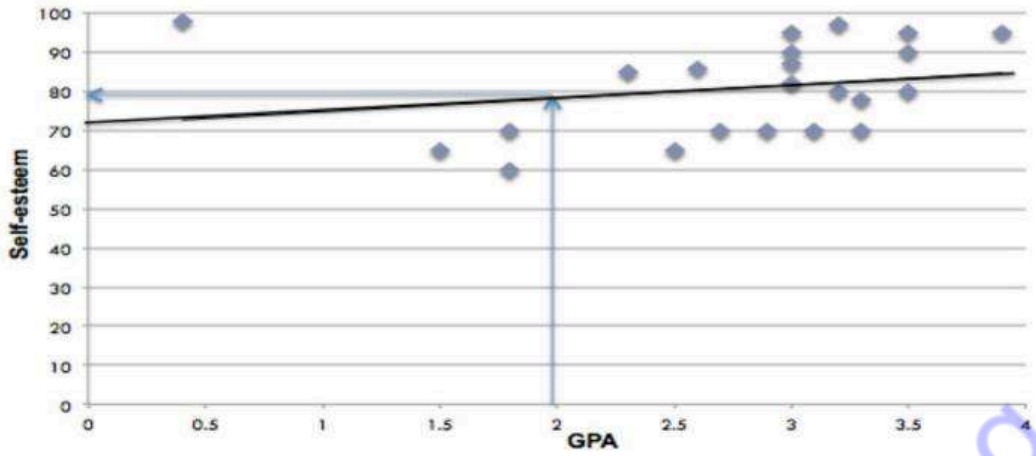
3.4.3 REGRESSION LINE:

- **Define regression lines or why regression lines are important?(2M)**
- **Explain regression line and give an example(13M)**

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. A regression line can be used to predict the value of y for a given value of x.

Regression analysis identifies a regression line. The regression line shows how much and in what direction the response variable changes when the explanatory variable changes. Most individuals in the sample are not located exactly on the line; the line closely approximates all the points. The way this line is computed will be described in more detail later

Example: Predict a student's self-esteem score based on her GPA



- The purpose of a regression line is to make predictions.

- In the above example, it is known how GPA is related to self-esteem. Therefore, if a student's self-esteem has not been measured, but her GPA is known, her self-esteem score can be predicted based on her GPA.
- As an example, if a student has a GPA of 2.0, this score matches up with a score of approximately 78 or 79 on the self-esteem scale. This score has been estimated by looking at the graph.

  ✓ Draw a straight line up from the point that represents a 2.0 GPA and find where this line intersects with the regression line.

  ✓ Then, draw a line straight from this point to the self-esteem axis to find the corresponding self-esteem score.

3.4.4 LEAST SQUARE METHOD:

- **Define least square method(2M)**
- **What is the formula to calculate Least Square Regression?(6M)**
- **Define Least Square Regression Line(2M)**
- **Explain Least Square Regression line With example(13M)**

The least squares method is a form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points. Each point of data represents the relationship between a known independent variable and an unknown dependent variable.

This method of regression analysis begins with a set of data points to be plotted on an x- and y-axis graph. An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

The least squares method is used in a wide variety of fields, including finance and investing. For financial analysts, the method can help to quantify the relationship between two or more variables—such as a stock's share price and its earnings per share (EPS). By performing this type of analysis investors often try to predict the future behavior of stock prices or other factors.

3.4.4.1 FORMULA TO CALCULATE LEAST SQUARE REGRESSION:

The regression line under the Least Squares method is calculated using the following formula –

$$\hat{y} = a + bx$$

Where,

$\hat{y}$ = dependent variable

x = independent variable

a = y-intercept

b = slope of the line

The slope of line b is calculated using the following formula –

$$b = \frac{\Sigma(x-\bar{x})\,(y-\overline{y})}{\Sigma(x-\overline{x})^2}$$

Or

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Y-intercept, 'a' is calculated using the following formula –

$$a = \frac{\sum y - (b \sum x)}{n}$$

Where,

$\hat{y}$ = dependent variable

x = independent variable

a = y-intercept

b = slope of the line

The slope of line b is calculated using the following formula –

$$b = \frac{\Sigma(x-\bar{x})\,(y-\overline{y})}{\Sigma(x-\overline{x})^2}$$

Or

$$b = \frac{\Sigma xy - \frac{\Sigma x\,\Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

Y-intercept, 'a' is calculated using the following formula –

$$a = \frac{\Sigma y - (b\,\Sigma x)}{n}$$

3.4.4.2 LEAST SQUARE REGRESSION LINE:

If the data shows a leaner relationship between two variables, the line that best fits this linear relationship is known as a least-squares regression line, which minimizes the vertical distance from the data points to the regression line. The term "least squares" is used because it is the smallest sum of squares of errors, which is also called the "variance."

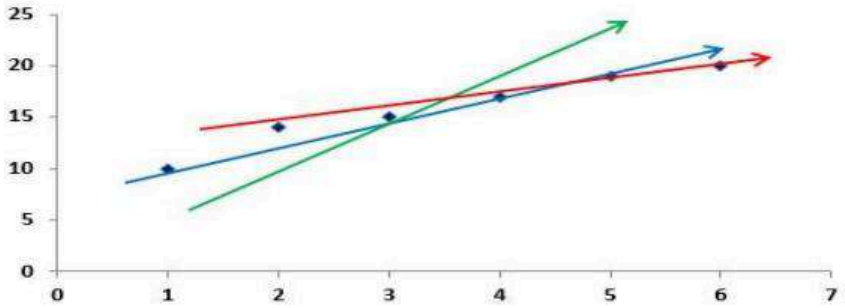
In regression analysis, dependent variables are illustrated on the vertical y-axis, while independent variables are illustrated on the horizontal x-axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

In contrast to a linear problem, a non-linear least-squares problem has no closed solution and is generally solved by iteration.

EXAMPLE:

The line of best fit is a straight line drawn through a scatter of data points that best represents the relationship between them.

Let us consider the following graph wherein a set of data is plotted along the x and y-axis. These data points are represented using the blue dots. Three lines are drawn through these points – a green, a red, and a blue line. The green line passes through a single point, and the red line passes through three data points. However, the blue line passes through four data points, and the distance between the residual points to the blue line is minimal as compared to the other two lines.



In the above graph, the blue line represents the line of best fit as it lies closest to all the values and the distance between the points outside the line to the line is minimal (i.e., the distance between the residuals to the line of best fit – also referred to as the sums of squares of residuals). In the other two lines, the orange and the green, the distance between the residuals to the lines is greater as compared to the blue line.

3.4.5 STANDARD ERROR OF ESTIMATE:

- ❖ Define Standard Error of Estimate(2M)
- ❖ How Standard Error of Estimate is calculated?(6M)

The **standard error of the estimate** is a way to measure the accuracy of the predictions made by a regression model.

Likewise, a standard deviation which measures the variation in the set of data from its mean, the standard error of estimate also measures the variation in the actual values of Y from the computed values of Y (predicted) on the regression line. It is computed as a standard deviation, and here the deviations are the vertical distance of every dot from the line of average relationship.

Often denoted σest, it is calculated as:

$\sigma est = \sqrt{\Sigma(y - \hat{y})2/n}$

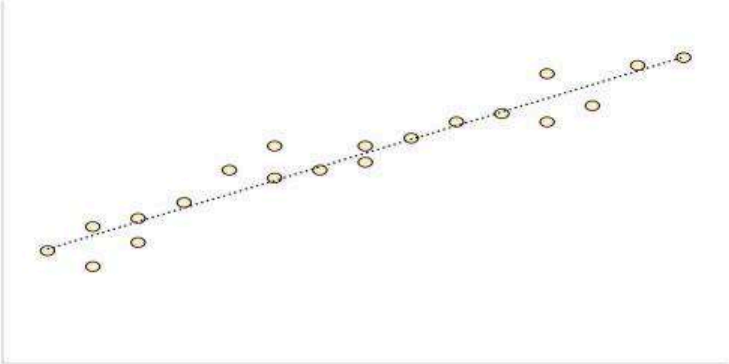
where:

- **y:** The observed value
- **ŷ:** The predicted value
- **n:** The total number of observations

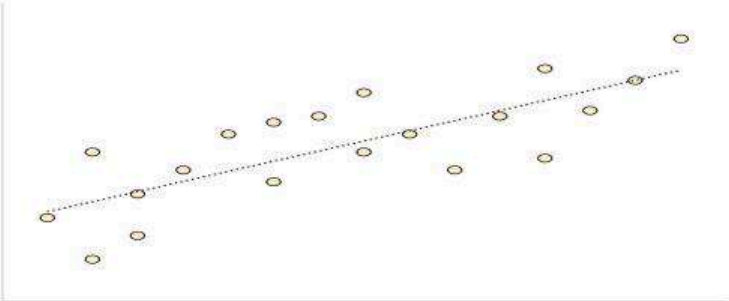
The standard error of the estimate gives us an idea of how well a regression model fits a data set. In particular:

- The smaller the value, the better the fit.
- The larger the value, the worse the fit.

For a regression model that has a small standard error of the estimate, the data points will be closely packed around the estimated regression line:



Conversely, for a regression model that has a large standard error of the estimate, the data points will be more loosely scattered around the regression line:



3.4.6 R-SQUARED:

- **Explain R-Squared(2M)**
- **What is the formula to calculate R-Squared?**
- **How to interpret R-Squared?(16M)**

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

After fitting a linear regression model, you need to determine how well the model fits the data.For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good.

R-squared is always between 0 and 100%:

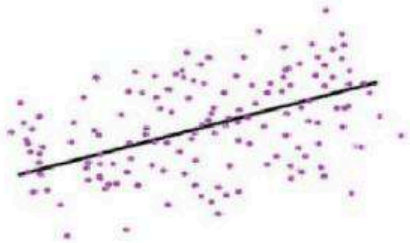
- 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- 100% represents a model that explains all the variation in the response variable around its mean.

Usually, the larger the R2, the better the regression model fits your observations.

3.4.6.1 INTERPRETATION OF R2:

Visual Representation of R-squared

You can have a visual demonstration of the plots of fitted values by observed values in a graphical manner. It illustrates how R-squared values represent the scatter around the regression line.



R-squared : 17%

R-squared : 83%

As observed in the pictures above, the value of R-squared for the regression model on the left side is 17%, and for the model on the right is 83%. In a regression model, when the variance accounts to be high, the data points tend to fall closer to the fitted regression line.

However, a regression model with an R2 of 100% is an ideal scenario which is actually not possible. In such a case, the predicted values equal the observed values and it causes all the data points to fall exactly on the regression line.

How to Interpret R squared

The simplest r squared interpretation is how well the regression model fits the observed data values. Let us take an example to understand this.

Consider a model where the R2 value is 70%. Here r squared meaning would be that the model explains 70% of the fitted data in the regression model. Usually, when the R2 value is high, it suggests a better fit for the model.

The correctness of the statistical measure does not only depend on R2 but can depend on other several factors like the nature of the variables, the units on which the variables are measured, etc. So, a high R-squared value is not always likely for the regression model and can indicate problems too.

A low R-squared value is a negative indicator for a model in general. However, if we consider the other factors, a low R2 value can also end up in a good predictive model.

Calculation of R-squared

R- squared can be evaluated using the following formula:

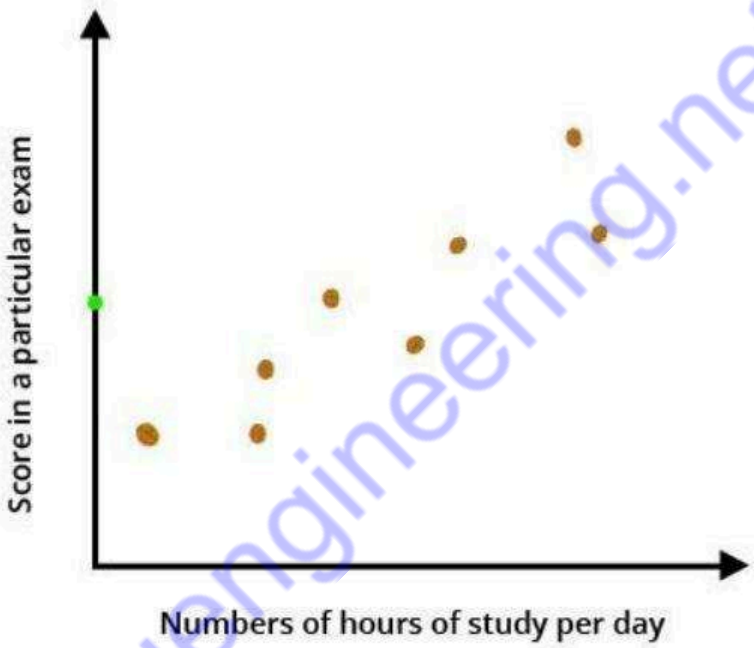
$$\text{R-squared} = \frac{SS_{regresion}}{SS_{total}}$$

Where:

- SSregression – Explained sum of squares due to the regression model.
- SStotal – The total sum of squares.

The sum of squares due to regression assesses how well the model represents the fitted data and the total sum of squares measures the variability in the data used in the regression model.

Now let us come back to the earlier situation where we have two factors: number of hours of study per day and the score in a particular exam to understand the calculation of R-squared more effectively. Here, the target variable is represented by the score and the independent variable by the number of hours of study per day.

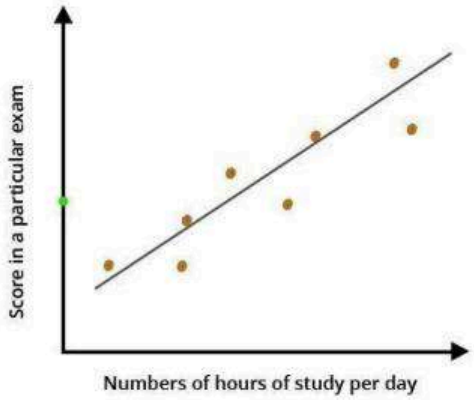


In this case, we will need a simple linear regression model and the equation of the model will be as follows:

$$\hat{y} = w1x1 + b$$

The parameters w1 and b can be calculated by reducing the squared error over all the data points. The following equation is called the least square function:

$$\text{minimize } \sum(yi - w1x1i - b)2$$



Now, to calculate the goodness-of-fit, we need to calculate the variance:

$$var(u) = 1/n\sum(ui - \bar{u})2$$

where, n represents the number of data points.

Now, R-squared calculates the amount of variance of the target variable explained by the model, i.e. function of the independent variable.

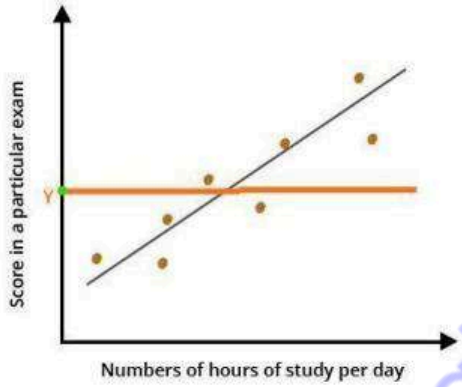
However, in order to achieve that, we need to calculate two things:

- Variance of the target variable:

$var(avg) = \sum(yi - \bar{y})2$

- Variance of the target variable around the best-fit line:

$var(model) = \sum(yi - \hat{y})2$



Finally, we can calculate the equation of R-squared as follows:

$R2 = 1 - [var(model)/var(avg)] = 1 - [\sum(yi - \hat{y})2/\sum(yi - \bar{y})2]$

3.4.7 MULTIPLE REGRESSION:

- ❖ Explain Multiple regression?(2M)
- ❖ Explain linear regression and multiple regression equation with example(13M)
- ❖ Assumptions of Multiple Regression Equations(2M)
- ❖ Benefits of Multiple Regression Equations(2M)

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

$Y=a+b1X1+b2X3+...+bnXn$

Here Y is the dependent variable, and X1,…,Xn are the *n* independent variables. In calculating the weights, a, b1,…,bn, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.

In the case of linear regression, although it is used commonly, it is limited to just one independent and one dependent variable. Apart from that, linear regression restricts the training data set and does not predict a non-linear regression.

For the same limitations and to cover them, we use multiple regression. It focuses on overcoming one particular limitation and that is allowing to analyze more than one independent variable.

3.4.7.1 Multiple regression equation

We will start the discussion by first taking a look at the linear regression equation:

y = bx + a

Where,

y is a dependent variable we need to find, x is an independent variable. The constants a and b drive the equation. But according to our definition, as the multiple regression takes several independent variables (x), so for the equation we will have multiple x values too:

y = b1x1 + b2x2 + … bnxn + a

Here, to calculate the value of the dependent variable y, we have multiple independent variables x1, x2, and so on. The number of independent variables can grow till n and the constant b with every variable denotes its numeric value. The purpose of the constant a is to denote the dependent variable's value in case when all the independent variable values turn to zero.

Example: A researcher decides to study students' performance at a school over a period of time. He observed that as the lectures proceed to operate online, the performance of students started to decline as well. The parameters for the dependent variable "decrease in performance" are various independent variables like "lack of attention, more internet addiction, neglecting studies" and much more.

So for the above example, the multiple regression equation would be:

y = b1 * attention + b2 * internet addiction + b3 * technology support + … bnxn + a

3.4.7.2 ASSUMPTIONS OF MULTIPLE REGRESSION ANALYSIS:

- ✧ The variables considered for the model should be relevant and the model should be reliable.
- ✧ The model should be linear and not non-linear.
- ✧ Variables must have a normal distribution
- ✧ The variance should be constant for all levels of the predicted variable.

3.4.7.3 BENEFITS OF MULTIPLE REGRESSION ANALYSIS:

- ✧ Multiple regression analysis helps us to better study the various predictor variables at hand.

- ✧ It increases reliability by avoiding dependency on just one variable and having more than one independent variable to support the event.
- ✧ Multiple regression analysis permits you to study more formulated hypotheses that are possible.

3.4.8 REGRESSION TOWARDS THE MEAN:

- ❖ Define regression towards mean(2M)
- ❖ Explain regression towards mean with example(13M)

In statistics, **regression toward the mean** (also called **reversion to the mean**, and **reversion to mediocrity**) is a concept that refers to the fact that if one sample of a random variable is extreme, the next sampling of the same random variable is likely to be closer to its mean. Furthermore, when many random variables are sampled and the most extreme results are intentionally picked out, it refers to the fact that (in many cases) a second sampling of these picked-out variables will result in "less extreme" results, closer to the initial mean of all of the variables.

Regression to the mean usually happens because of sampling error. A good sampling technique is to randomly sample from the population. If you don't (i.e. if you asymmetrically sample), then your results may be abnormally high or low for the average and therefore would regress back to the mean. Regression to the mean can also happen because you take a very small, unrepresentative sample (say, the highest 1 percent of the population or the lowest ten percent).

Formula for the Percent of Regression to the Mean:

You can use the following formula to find the percent for any set of data:

Percent of Regression to the Mean = 100(1-r)

where r is the correlation coefficient.

Why 1-r?
Note: In order to understand this discussion you should be very familiar with r, the correlation coefficient.

The percent of regression to the mean takes into account the correlation between the variables. Take two extremes:
If r=1 (i.e. perfect correlation), then 1-1 = 0 and the regression to the mean is zero. In other words, if your data has perfect correlation, it will never regress to the mean.
With an r of zero, there is 100 percent regression to the mean. In other words, data with an r of zero will *always* regress to the mean.

EXAMPLE:

If your favorite team won the championship last year, what does that mean for their chances for winning next season? This is an important question, often with money or pride on the line (The League, anyone?). To the extent this is due to skill (the team is in good condition, top coach etc.), their win signals that it's more likely they'll win next year. But the greater the extent this is due to luck (other teams embroiled in a drug scandal, favourable draw, draft picks turned out well etc.), the less likely it is they'll win next year. This is because of the statistical concept of regression to the mean.

Another example,
because of regression toward the mean, we would expect that students who made the top five scores on the first statistics exam would not make the top five scores on the second statistics exam. Although all five students might score above the mean on the second exam, some of their scores would regress back toward the mean. Most likely, the top five scores on the first exam reflect two components. One relatively permanent component reflects the fact that these students are superior because of good study habits, a strong aptitude for quantitative reasoning, and so forth. The other relatively transitory component reflects the fact that, on the day of the exam, at least some of these students were very lucky because all sorts of little chance factors, such as restful sleep, a pleas ant commute to campus, etc., worked in their favor. On the second test, even though the scores of these five students continue to reflect an above-average permanent component, some of their scores will suffer because of less good luck or even bad luck. The net effect is that the scores of at least some of the original five top students will drop below the top five scores—that is, regress *back* toward the mean—on the second exam. (When signifi cant regression toward the mean occurs after a spectacular performance by, for example, a rookie athlete or a first-time author, the term *sophomore jinx* often is invoked.)
There is good news for those students who made the five lowest scores on the first exam. Although all five students might score below the mean on the second exam, some of their scores probably will regress *up* toward the mean. On the second exam, some of them will not be as unlucky. The net effect is that the scores of at least some of the original five lowest scoring students will move above the bottom five scores—that is, regress up toward the mean—on the second

EDU ENGINEERING

PIONEER OF ENGINEERING NOTES

TAMIL NADU'S BEST EDTECH PLATFORM FOR ENGINEERING

CONNECT WITH US

WEBSITE: www.eduengineering.net

TELEGRAM: @eduengineering

INSTAGRAM: @eduengineering

- **Regular Updates for all Semesters**
- **All Department** Notes AVAILABLE
- **Handwritten** Notes AVAILABLE
- **Past Year** Question Papers AVAILABLE
- **Subject wise** Question Banks AVAILABLE
- **Important Questions** for Semesters AVAILABLE
- Various **Author Books** AVAILABLE