CHAPTER 5: STATISTICS

Objectives:
a. Identify and differentiate Patterns in Nature.
b. Understand the Fibonacci Sequence.
c. Appreciate the beauty of Mathematics in terms of Patterns and Number in Nature and in the World.

Lesson 1: Measures of Central Tendency

Many problems in statistics are concerned with averages. The most common measures that attempt to locate the center of a set of data are: *average or mean, median* and *mode.*

The mean is the part of the distribution around which the values balance.

Symbol for mean: $\overline{X}$, read as "x bar"

The median provides the necessary information about the value of the middle position in the distribution.

Symbol for median: $\tilde{X}$ read as "x tilde"

The mode is the score with the highest frequency.

Symbol for mode: $\hat{X}$ read as "x hat"

Definition:

The **mean** (commonly called the average) of a set of **n** numbers is the sum of all numbers divided by **n.**

The **median** is the middle number in a set of data when arranged in decreasing order. When there are even numbers of elements, the median is the mean of two middle number.

The **mode** is the number that occurs most often in a set of data. A set of data can have more than one mode. If all the numbers appear the same number of times, there is no mode for that data set.

A. The Mean

The most widely used average, the *arithmetic mean*, is defined as the sum of the observations divided by the number of observations.

Mean (Ungrouped data)

$$\bar{X} = \frac{\sum X_i}{N}$$

Example:

A motorist records the time in it takes him to travel to work by car during the peak hour traffic for a 10-day period. The times (to the nearest minute) are as follows:

$$36, 33, 28, 28, 32, 29, 33, 34, 32, 33$$

Find the mean time it took him to get to work during the two weeks (ten working days).

Solution:

$$\bar{X} = \frac{\sum X_i}{N}$$

$$= \frac{36+33+28+28+32+29+33+34+32+33}{10}$$

$$= \frac{318}{10}$$

$$= 31.8 \text{ minutes}$$

Weighted Mean

If a set of data is in the form of a frequency distribution in which an observation x occurs with frequency *f*, we may use the formula for the arithmetic mean as:

$$\bar{X} = \frac{\sum fX}{N}$$

Where: $\bar{X}$ = mean, *f*= frequency, X= score

$\sum fX$ = sum of the products of frequency and score

N= total frequency

Example:

Suppose a particular math curse is graded in the following manner:

Assignments	10%
Project	20%
Midterm Exams	30%
Final Exams	40%
	100 %

Jason obtained marks of 85%in assignments, 74% in project, 76% in midterm exams and 87% in his final examples. Find his weighted mean mark for math course.

Solution:

$$\bar{X} = \frac{\sum fX}{N}$$

$$= \frac{\sum 85(10) + 74(20) + 76(30) + 87(40)}{100}$$

$$= 80.9\%$$

Mean (Grouped data)

$$\bar{X} = \frac{\sum fX_m}{N}$$

Where: $\bar{X}$ = mean

f = frequency

X_m = class mark (average of lower interval and upper interval)

$\sum fX_m$ = sum of the product of frequencies and class mark

N= total frequency

Example:

Shown below in Table are the sources of 50 junior students in an achievement test. Calculate the mean score.

Class Interval	Frequency (f)
95-99	1
90-94	9
85-89	8
80-84	14
75-79	11
70-74	5
65-69	2

Solution:

Add two columns for the class mark (X_m) and fX_m in the given table. Find the summations of f and fX_m.

Class Interval	Frequency (f)	X_m	fX_m
95-99	1	97	97
90-94	9	92	828
85-89	8	87	696
80-84	14	82	1148
75-79	11	77	847
70-74	5	72	360

65-69	2	67	134
	$\sum f$ = 50		$\sum fX_m$ = 4110

Solve for the mean.

$$\bar{X} = \frac{\sum fX_m}{N} = \frac{4110}{50} = 82.2$$

B. The Median

The median is the value of the middle observation if the data are arranged in the form of an array. Thus, the median is the value of an array which divides it so that there an equal number of observations on either side of it. The median is often used when describing an educational and sociological data, such as ages, income, family size, etc.

If there is an *odd* number (n) of observations, then the median is the value of $\left(\frac{n+1}{2}\right)th$ observation. If n is an *even* number, the median is usually defined as the mean of the $\left(\frac{n}{2}\right)th$ and $\left(\frac{n+1}{2}\right)th$ observation.

Example 1:

In a certain clinic, the waiting time (in minutes) for 11 randomly chosen out patients on a particular day are:

$$12, 36, 34, 15, 17, 14, 8, 40, 16, 36, 17$$

Find the median waiting time.

Solution:

Arrange the data in order of magnitude.

$$8, 12, 14, 15, 16, 17, 17, 26, 34, 36, 40$$

Since $n = 11$, there are observations. Therefore:

$$Median \text{ waiting time} = \left(\frac{11+1}{2}\right)th \text{ observations}$$

$= 6^{th}$ observations
= 17 minutes

Example 2:

A sample of 50 students was given an inventory test (based on a possible score of 0-5). The result sore as follows:

Score	0	1	2	3	4	5
Frequency	5	9	12	16	6	2

Find the median score.

Solution:

Construct a cumulative frequency distribution.

Score	Frequency`	Cum. Frequency
0	5	5
1	9	14
2	12	26
3	16	42
4	6	48
5	2	50
	$\sum f = 50$	

There are $\sum f = 50$ observations. The median is the mean of the 25^{th} and 26^{th} observations. Since the data is already in an array form, it can easily be seen that both 25^{th} and 26^{th} observations are 2. Hence**, the median score is 2.**

The first step in the computation of the median of a grouped data is to determine the class interval which contains the $\left(\frac{n}{2}\right)th$ score. This can be located under the column $<cf$ of the cumulative frequency distribution. The class interval that contains the $\left(\frac{n}{2}\right)th$ score is called the *median class* of the distribution. To calculate the median, we use the formula:

Median (Grouped Data)

$$\tilde{X} = X_{LB} + \left[\frac{\frac{N}{2} - cf_b}{f_m}\right]i$$

Where:

$\tilde{X}$ = median
X_{LB} = the lower boundary or true lower limit of the median class
N = total frequency
cf = cumulative frequency before the median class
f_m = frequency of the median class
i = size of the class interval

Example:

Calculate the median score of 50 junior students in an achievement test in Math given in the table below.

Solution:

Achievement test Results in Math of 50 Junior Students

Class Interval	Frequency (*f*)
95-99	1
90-94	9
85-89	8
80-84	14
75-79	11
70-74	5
65-69	2

Add the entitles in the column for $<cf$.

Achievement test Results in Math of 50 Junior Students

Class Interval	Frequency (*f*)	$<cf$
95-99	1	50
90-94	9	49
85-89	8	40
80-84	14	32
75-79	11	18
70-74	5	7
65-69	2	2
N=50		

$\frac{N}{2}th$ score= $\left(\frac{50}{2}\right)th$ score

= 25th score

The class interval that contains the 25th score is 80-84.

X_{LB} = 79.5
cf_b = 18

$f_m = 14$

$i = 5$

$$\tilde{X} = X_{LB} + \left[\frac{\frac{N}{2} - cf_b}{f_m} \right] i$$

$$\tilde{X} = 79.5 + \left[\frac{\frac{50}{2} - 18}{14} \right] 5$$

$$\tilde{X} = 79.5 + 2.5$$

$$\tilde{X} = 82.2$$

This means that 50 percent of the students got scores below 82.

C. The Mode

The *mode* is defined as the observation which occurs the most often in a set of data

This is the observation which has the largest frequency. It is frequently used to determine those products which are in greatest demand.

Example:

Find the mode of the following scores:

$14, 17, 17, 17, 18, 18, 19, 20, 21, 21, 23$

Solution:

By inspection, the mode is 17 since it occurs 3 times in the distribution.

Note: A distribution can have one or more modes.

Example:

An ice cream parlor sells 6 flavors of ice cream The numbers for each type sold on a particular day are shown below.

Flavors of Ice Cream	Frequency of sale
Cheese	16
Chocolate	12
Vanilla	22

Macapuno	26
Strawberry	23
Fruit Salad	18

Determine the most popular flavor of ice cream for that day.

Solution:

The most popular flavor is that which is most frequently sold. The highest frequency is 26. Hence, the modal (most popular) flavor for that day is *macapuno.*

In the computation of the mode given a frequency distribution, the first step is to get the modal class. The modal class is that class interval with the highest frequency. To compute for the mode, we use the formula:

Mode (Grouped data)

$$\hat{X} = X_{LB} + \left(\frac{d_1}{d_1 + d_2} \right) i$$

Where:

X_{LB} = lower boundary of the modal class

d_1 = difference of the frequency of the modal class and the frequency preceding it.

d_2 = difference of the frequency of the modal class and the frequency succeeding it.

i = size of the class interval

Example:

Find the mode for the following grouped frequency distribution.

Achievement test Results in Math of 50 Junior Students

Class Interval	Frequency (*f*)
95-99	1
90-94	9
85-89	8
80-84	14
75-79	11
70-74	5
65-69	2

Solution:

The modals class is the class interval 80-84 since it has the highest frequency. Therefore,

X_{LB} = 79.5 $\qquad$ $d_2 = 14 - 8 = 6$

$$d_1 = 14 - 11 = 3 \qquad i = 5$$

$$\hat{X} = X_{LB} + \left(\frac{d_1}{d_1 + d_2}\right) i$$

$$\hat{X} = 79.5 + \left(\frac{3}{3+6}\right) 5$$

$$\hat{X} = 79.5 + 1.67$$

$$\hat{X} = 81.17$$

For more knowledge about Measures of Central Tendency, please check the link provided;
http://onlinestatbook.com/2/summarizing_distributions/measures.html
https://study.com/academy/lesson/central-tendency-measures-definition-examples.html

REMEMBER

- **Mean** is the most widely used average, the *arithmetic mean*, is defined as the sum of the observations divided by the number of observations.
- **Median** is the value of the middle observation if the data are arranged in the form of an array.
- **Mode** is defined as the observation which occurs the most often in a set of data

ACTIVITY:
Choose 10 of your classmates and ask them if how much money left in their pocket. In your collected data, compute for the mean, median and mode.

Lesson 2: Measures of Dispersion

The measure of central tendency is not in itself sufficient to adequately describe a set of data. In addition, a *measure of dispersion* (or spread) of data is also required. This measure describes the extent to which individual observations vary above and below the average. The need for a measure of dispersion is just as important as the average. A measure of dispersion gives an indication of the reliability of the average value.

The most commonly measure of dispersion are: the range, the quartile deviation, the mean deviation, the variance and the standard deviation.

A. The Range

The easiest and the simplest way to determine measure of dispersion is the range. The *range* is simply defined as the difference of the highest score (H.S) and the lowest score (L.S). It shows the extreme scores of a set of data.

When we talk of grouped data, the range can be calculated data by subtracting the lower boundary (L.B) of the lowest class interval from the upper boundary (U.B) of the highest class interval. That is,

$$R = H.S - LS = U.B - LB$$

Example 1:

a. The range of the set of scores in $12, 14, 14, 16, 16$ is $16 - 12$ or 4.
b. The range of the set of scores in $10, 14, 14, 18, 25$ is $25 - 10$ or 15.

Example 2:

Find the range of the frequency distribution below.

Class interval	Frequency
38-39	1
36-37	3
34-35	3
32-33	3
30-31	6
28-29	6
26-27	8
24-25	6

22-23	10
20-21	14

Solution:

$$Range = U.B - L.B$$
$$Range = 39.5 - 19.5$$
$$Range = 20$$

B. The Quartile Deviation

An extension of the median is the concept of *quartiles* which divide the data into four equal parts. Quartiles are often used with scores on aptitude test, examinations, and other testing stations. When the data is divided into four equal parts, the points of separation are:

- 1st quartile (Q_1). There are 25% of the observations *below* Q_1 and 75% of the observation *above* Q_1.
- 2nd quartile (Q_2). There are 50% of the observations *below* Q_2 and 50% of the observations *above* Q_2. The second quartile is also the median.
- 3rd quartile (Q_3). There are 75% of the observations *below* Q_3 and 25% of the observations *above* Q_3.

If there are n observation in a set of data, then Q_1 can be identified as the $\left(\frac{n+1}{4}\right)^{th}$ observation, and Q_3 as the $\left[\frac{3(n+1)}{4}\right]^{th}$ observation.

Example:

Mr. Basanez is interested in the amount of time it takes his bank tellers to service customers. One particular morning, her records the service times for 15 customers. The times (to the nearest minute) are given below.

$$6, 9, 7, 5, 16, 11, 9, 7, 4, 9, 7, 11, 10, 8, 6$$

a. Find the median time
b. Find Q_1 and Q_3 of the service times.

Solution:

The number of observations is n=15. Arrange the data in array

$$\begin{array}{cccc} 4,5,6,6 & 7,7,7,8 & 9,9,9,10 & 11,11,16 \\ \uparrow & \uparrow & \uparrow & \\ Q_1 & Q_2 & Q_3 : 12th & \end{array}$$

a. The median or $Q_2 = \left(\frac{n+1}{2}\right)^{th}$ *observations*

$= 8th$ observation

$= 8$ minutes

b. The first quartile: $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ *observation*

$= 4^{th}$*observation*

$= 6$ minutes

c. The third quartile: $Q_3 = \frac{3(n+1)}{4}^{th}$ *observation*

$= 12^{th}$*observation*

$= 10$ minutes

Q_1 and Q_3 Grouped data

$$Q_1 = X_{LB} + \left(\frac{\frac{N}{4} - cf_b}{f_{q1}}\right) i$$

Where: X_{LB} = lower boundary of the Q_1 class

N = total frequency

cf_b = cumulative frequency before the Q_1 class

f_{q1} = frequency of the Q_1 class

i = size of the class interval

Q_1 and Q_3 Grouped data

$$Q_3 = X_{LB} + \left(\frac{\frac{3N}{4} - cf_b}{f_{q3}}\right) i$$

Where: X_{LB} = lower boundary of the Q_3 class

N = total frequency

cf_b = cumulative frequency before the Q_3 class

f_{q3} = frequency of the Q_3 class

i = size of the class interval

Example: From the given frequency distribution table, compute for Q_1, Q_2 and Q_3.

Class interval	f
28-32	3
23-27	8
18-22	15
13-17	12
8-12	5
3-7	2

Solution:

a. The first step is to add the entitles in the column for $< cf$.

Class interval	f	$< cf$
28-32	3	45
23-27	8	42
18-22	15	34
13-17	12	19
8-12	5	7
3-7	2	2

b. Calculate for Q_1:

$Q_1 class = \frac{N}{4} = \frac{45}{4} = 11.25$ Class interval 13-17

X_{LB} = 12.5

cf_b = 7

f_{q1} = 12

i =5

$$Q_1 = X_{LB} + \left(\frac{\frac{N}{4} - cf_b}{f_{q1}} \right) i$$

$$Q_1 = 12.5 + \left(\frac{11.25 - 7}{12} \right) 5$$

$= 12.5+1.77$

$= 14.27$

C. Calculate for Q_2:

$Q_2 class = \frac{2N}{4} = \frac{90}{4} = 22.5$ Class interval 18-22

X_{LB} = 17.5

cf_b = 19

f_{q1} = 15

i =5

$$Q_2 = 17.5 + \left(\frac{22.5 - 19}{15}\right)5$$
$= 17.5+1.17$

$= 18.67$

d. Calculate for Q_3:

$$Q_3 class = \frac{3N}{4} = \frac{135}{4} = 33.75 \text{ Class interval 18-22}$$

X_{LB} = 17.5

cf_b = 19

f_{q1} = 15

i =5

$$Q_3 = 17.5 + \left(\frac{33.75 - 19}{15}\right)5$$
$= 17.5+4.92$

$= 22.42$

Like the range, the *quartile deviation* is a measure of dispersion which is determined by the distance between two particular observations. The first step is to calculate the *interquartile range.*

The **interquartile range or (I.R)** is a more reliable measures of variability. It is the difference of the 75th percentile or Q_3 and the 25th percentile or Q_1, hence we can conclude that 50 percent of the distribution will be falling within the interquartile range, 25 percent will be falling below Q_1 and 25 percent will be above Q_3.

Interquartile Deviation

$$I.R = Q_3 - Q_1$$

The formula for finding the interquartile range shows the distance between Q_3 and Q_1. The value obtained half of this distance is called the *quartile deviation* or (Q.D) and the formula is given by:

Quartile Deviation

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Example:

A farmer has his corn crop spread over 15 fields each of equal size. The output (in cubic meters) for each of the fields is

$$226, 174, 185, 203, 193, 216, 164, 228, 244, 208, 235, 200, 216, 196, 188$$

a. Find the range of the output.
b. Find the quartile deviation of the outputs.

Solution:

Arrange the data in an array:

164,174,185,188,193,196, 200, 202, 208, 216, , 216, 226, 228, 235, 244

a. The range of the output $= 244 - 164 = 80$ cubic meters

b. The first quartile =4th observations
=188 cubic meters

The third quartile =12th observations
=226 cubic meters

The interquartile range $= 226 - 1888$
=38 cubic meters

The quartile deviation $= \frac{\text{interquartile range}}{2} = \frac{226 - 188}{2}$

$= \frac{38}{2}$

$= 19$ cubic meters

C. **The Mean Deviation**

A measure of dispersion which takes into account each observation is the *mean deviation*. This is more reliable than the range and the quartile deviation because each makes use of only two values in the distribution, namely: the two most extreme values in the range; and Q_3 and Q_1 in the quartile deviation. The formulas for the computation of the mean deviation will be shown.

Ungrouped Distribution Mean Deviation (M.D)

$$M.D = \frac{\sum\left|X - \bar{X}\right|}{N}$$

Where:

X = represents the scores of the distribution

$\bar{X}$ = is the mean

N = is the number of observations

The formula tells us that we have to follow the following steps:

1. Calculate the mean of the data.
2. Add a column for $\left|X - \bar{X}\right|$.
3. Subtract the mean of each score and record the differences.
4. Write down the absolute values of each of the differences.
5. Get the total of the score under the heading $\left|X - \bar{X}\right|$.
6. Divide the sum obtained in Step 3 by N.

Example: Find the mean deviation of the following ungrouped distribution.

x	5	8	11

Solution:

a. Calculate the mean.

$$\bar{X} = \frac{\sum X}{N} = \frac{24}{8} = 8$$

b. Add the column for $|X - \bar{X}|$.

c.

X	$\lvert X - \bar{X}\rvert$
5	3
8	0
11	3

d. $\sum |X - \bar{X}| = 6$

e. $M.D = \frac{6}{3} = 2$

Grouped Frequency Distribution: Mean Deviation (M.D)

$$M.D = \frac{\sum f|X - \bar{X}|}{N} \text{ or } M.D = \frac{\sum f|X_m - \bar{X}|}{N}$$

Example: Find the mean deviation of the following distribution.

X	**f**
20	5
18	3
16	7
14	15
12	12

10	8
	N= 50

Solution:

a. Calculate the mean by using the formula $\bar{X} = \frac{\sum fX}{N}$. This means we are going to add the entitles in the column for fX.

X	f	fX
20	5	100
18	3	54
16	7	112
14	15	210
12	12	144
10	8	80
	N= 50	$\sum fX$ = 700

$$\bar{X} = \frac{\sum fX}{N}$$
$$= \frac{700}{50}$$
$$= 14$$

b. Add two columns for $\left|X - \bar{X}\right|$ and $f\left|X - \bar{X}\right|$.

X	f	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$
20	5	6	30
18	3	4	12
16	7	2	14
14	15	0	0
12	12	2	24
10	8	4	32
	N= 50		$\sum f\lvert X-\bar{X}\rvert=112$

c. Divide $\sum f\lvert X-\bar{X}\rvert$ by N.

$$M.D=\frac{112}{50}=2.24$$

D. The Variance and Standard Deviation

The variance and the standard deviation are the most commonly used measures of variation of a set of data. The standard deviation allows us to immediately compare the spread of different sets of score and enables us also to interpret the scores of a given set of data. Like the mean deviation, the variance and standard deviation are based on the deviation of the individual observations about the arithmetic mean. Also, the more widely scattered the observations are about the mean, the larger the value of the standard deviation.

The ***variance*** is defined as the quotient of the sum of the squared deviations from the mean divided by N-1 while the ***standard deviation*** is the square root of the variance. The formulas are given below

Variance $\left(S^2\right)$ **and Standard Deviation** (S)

$$S^2=\frac{\sum\left(X-\bar{X}\right)^2}{N-1} \quad \text{and} \quad S=\sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{N-1}}$$

These are formulas use the mean deviation method and tell us to follow the following steps:

1. Calculate the mean.
2. Get the difference of each score and the mean, then get the square of this difference.
3. Get the sum of the squared deviations in Step 2.
4. Substitute in the formulas: $S^2 = \frac{\sum\left(X - \bar{X}\right)^2}{N-1}$ and $S = \sqrt{\frac{\sum\left(X - \bar{X}\right)^2}{N-1}}$

Example: Find the variance and standard deviation of the following distribution:

X	5	8	11

Solution:

a. Calculate the mean: $\bar{X} = 8$

b. Add the column for $\left(X - \bar{X}\right)^2$ and sum up the scores.

X	$\left(X - \bar{X}\right)^2$
5	9
8	0
11	9
	$\sum\left(X - \bar{X}\right)^2 = 18$

c. Divide $\sum\left(X - \bar{X}\right)^2$ by 2 since $N - 1 = 3 - 1 = 2.\backslash$

$$s^2 = \frac{18}{2} = 9$$

$$s = \sqrt{9} = 3$$

The following are important points to remember regarding the calculations of the standard deviation.

a. The standard deviation cannot be negative.
b. The standard deviation of a s et of data is zero if and only if the observations are of equal value.
c. As a rough guide, the standard deviation should have a value which is equal to approximately one-third of the range.
d. The standard deviation cannot be more than the range of the data.
e. If a constant k is added to each observation in a set of data, the standard deviation of the new det of data has the same value as the standard deviation of the original set of data.

This method of computing the variance and standard deviation is called the *raw score method*. The formulas are given below.

Variance	**Standard Deviation**

For Ungrouped Data

$$S^2 = \frac{N\sum X^2 - \left(\sum X\right)^2}{N(N-1)} \qquad S = \sqrt{\frac{N\sum X^2 - \left(\sum X\right)^2}{N(N-1)}}$$

For Ungrouped Frequency Distribution

$$S^2 = \frac{N\sum fX^2 - \left(\sum fX\right)^2}{N(N-1)} \qquad S = \sqrt{\frac{N\sum fX^2 - \left(\sum fX\right)^2}{N(N-1)}}$$

For Grouped Frequency Distribution

$$S^2 = \frac{N\sum fX_m{}^2 - \left(\sum fX_m\right)^2}{N(N-1)} \qquad S = \sqrt{\frac{N\sum fX_m{}^2 - \left(\sum fX_m\right)^2}{N(N-1)}}$$

Example: Find the variance and the standard deviation of the following distribution.

X	5	8	11

Solution:

a. Get $\sum X$.

X
5
8
11
$\sum X$ =24

b. Add a column for X^2, square all the scores and get their sum.

X	X^2
5	25
8	64
11	121
$\sum X$ =24	$\sum X^2 = 210$

c. Substitute $\sum X = 24$=24 and $\sum X^2 = 210$ in the formula.

$$S^2 = \frac{N\sum X^2 - \left(\sum X\right)^2}{N(N-1)}$$

$$S^2 = \frac{3(210) - (24)^2}{3(3-1)}$$

$$S^2 = \frac{630 - 576}{6}$$

$$S^2 = \frac{54}{6}$$

$$S^2 = 9$$

$$S = \sqrt{9}$$

$$S = 3$$

For more knowledge about Measures of Dispersion, please check the link provided;
https://study.com/academy/lesson/measures-of-dispersion-definition-equations-examples.html
https://www.slideshare.net/BirinderSinghGulati/measures-of-dispersion-

REMEMBER

- The ***range*** is simply defined as the difference of the highest score (H.S) and the lowest score (L.S).
- A measure of dispersion which takes into account each observation is the ***mean deviation*.**
- The **standard deviation** allows us to immediately compare the spread of different sets of score and enables us also to interpret the scores of a given set of data.

ACTIVITY:

Calculate the:
a. range
b. quartile deviation
c. mean deviation
d. standard deviation
e. variance

for these data. 1,6,3,5,5,3,4,1,2,7,3,2,4

Lesson 3: Measures of Relative Position

A measure of position is a method by which the position that a particular data value has within a given data set can be identified. As with other types of measures, there is more than one approach to defining such a measure.

- **Standard Score (z-score)**

The **standard score** (often called the **z-score**) for a given data value *x* is the number of standard deviations that *x* is above or below the mean of the data. The following formulas show how to calculate the z-score for a data value *x* in a population and sample.

population data $z = \frac{x-\mu}{\sigma}$ and sample data $z = \frac{x-\bar{x}}{s}$

To compute a standard score, only the mean and standard deviation are required. However, since both of those quantities do depend on every value in the data set, a small change in one data value will change every z-score.

Example:

Scores on a history test have an average of 80 with a standard deviation of 6. What is the *z*-score for a student who earned a 75 on the test?

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

$$z = \frac{75-80}{6}$$
$$z = -0.833$$

Example:

The weight of chocolate bars from a particular chocolate factory has a mean of 8 ounces with a standard deviation of .1 ounce. What is the *z*-score corresponding to a weight of 8.17 ounces?

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

$$z = \frac{8.17-8}{.1}$$
$$z = 1.7$$

- **Percentiles**

A value *x* is called the pth percentile of a data set provided p% of the data values are less than *x*.

Example:

In a recent year, the median annual salary for a physical therapist was Php 74,480. If the 90th percentile for the annual salary of a physical therapist was Php 105,900; find the percent of physical therapists whose annual salary was

a. More than Php 74,480
b. Less than Php 105,900
c. Between Php 74,480 and Php 105,900

Solution:

a. By definition, the median is the 50th percentile. Therefore, 50% of the physical therapists earned more than Php 74,480 per year.
b. Because Php 105,900 is the 90th percentile, 90% of all physical therapists earned less than Php 105,900.
c. From parts a and b,

$$90\%-50\% = 40\%$$

40% of the physical therapists earned between Php 74,480 and Php 105,900.

Percentile for a Given Data Value

Given a set of data and a data value *x*.

$$\text{Percentile of score x} = \frac{\text{number of data values less than x}}{\text{total number of data values}} \cdot 100$$

Example:

On a reading examination given to 900 students. Benedict's score of 602 was higher than the scores of 576 of the students who took the examination. What is the percentile for Benedict's score?

Solution:

$$\text{Percentile} = \frac{\text{number of data values less than 602}}{\text{total number of data values}} \cdot 100$$

$$\text{Percentile} = \frac{576}{900} \cdot 100$$

$$= 64$$

Therefore, Benedict's score of 602 places him at the 64th percentile.

- **Quartile**

The three numbers $Q_{1,}$ $Q_{2,}$ and Q_3 that partitions a ranked data into four equal groups are called *quartiles.*

- 1st quartile (Q_1). There are 25% of the observations *below* Q_1 and 75% of the observation *above* Q_1.
- 2nd quartile (Q_2). There are 50% of the observations *below* Q_2 and 50% of the observations *above* Q_2. The second quartile is also the median.
- 3rd quartile (Q_3). There are 75% of the observations *below* Q_3 and 25% of the observations *above* Q_3.

The Median procedure for Finding Quartiles

1. Rank the data.
2. Find the median of the data. This is the second quartile Q_2.
3. The first quartile Q_1 is the median of the data values less than Q_2. The third quartile Q_3 is the median of the data values greater than Q_2

Example:

The following table lists the calories per 100 milliliters of 25 popular sodas. Find the quartiles for the data.

Calories per 100 milliliters of Selected sodas

43	37	42	40	53	62	36	32	50	49
26	53	73	48	45	39	45	48	40	56
41	36	58	42	39					

Solution:

Step 1: Rank the data as shown as the following table.

1) 26	2) 32	3) 36	4) 36	5) 37	6) 39	7) 39	8) 40	9) 40
10) 41	11) 42	12) 42	13) 43	14) 45	15) 45	16) 48	17) 48	18) 49
19) 50	20) 53	21) 53	22) 56	23) 58	24) 62	25) 73		

Step 2: The median of these 25 data values has a rank of 13. Thus, the median is 43. The second quartile Q_2 is the median of the data, so $Q_2 = 43$.

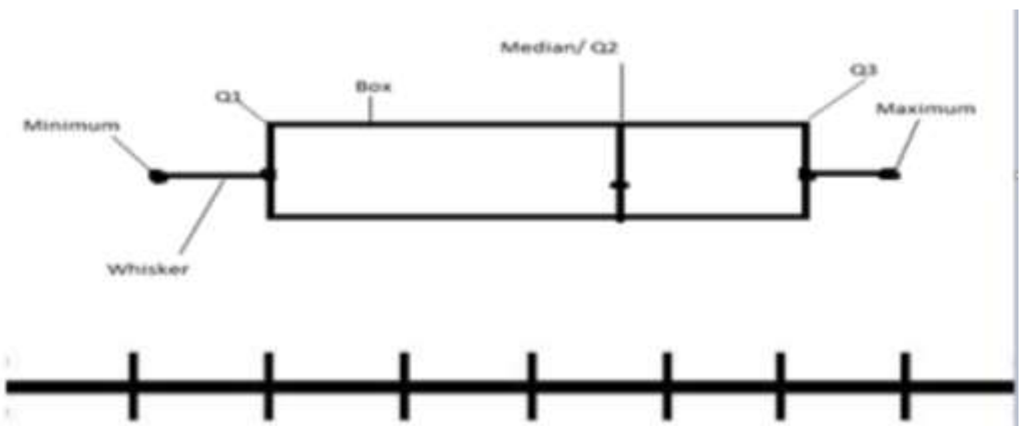
Step 3: There are 12 data values less than the median and 12 data values greater than the median. The first quartile is the median of the data values less than the median. Thus, Q_1 is the mean of the data values with rank of 6 and 7.

$$Q_1 = \frac{39+39}{2} = 39$$

The third quartile is the median of the data values greater than the median. Thus, Q_3 is the mean of the data values with ranks of 19 and 20.

$$Q_3 = \frac{50+53}{2} = 51.5$$

- **Box-and-Whisker Plot** It is sometimes called as box-plot. It is often used to provide a visual summary of a set of data. A box-and-whisker plot shows the median, the first and the third quartiles, and the minimum and maximum values of a data set.



Construction of a Box-and-Whisker Plot

1. Draw a horizontal scale that extends from minimum data value the maximum data value.
2. Above the scale, draw a rectangle (a box) with left side at Q_1 and its right at Q_3.
3. Draw a vertical line segment across the rectangle at the median, Q_2.
4. Draw a horizontal line segment, called a whisker, that extends from Q_1 to the minimum and another whisker that extends from Q_3 to the maximum.

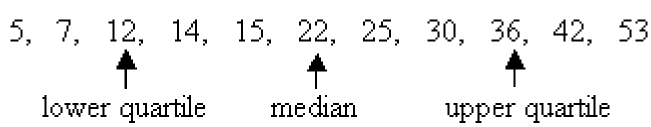
Example:

Construct a box plot for the following data:

$$12,\ 5,\ 22,\ 30,\ 7,\ 36,\ 14,\ 42,\ 15,\ 53,\ 25$$

Solution:

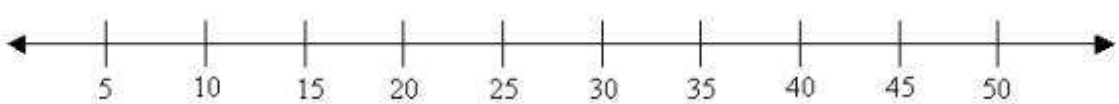
Step 1: Arrange the data in ascending order.
Step 2: Find the median, lower quartile and upper quartile



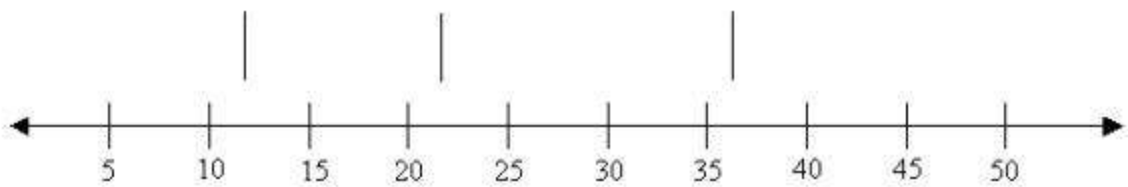
Median (middle value) = 22
Lower quartile (middle value of the lower half) = 12
Upper quartile (middle value of the upper half) = 36

(If there is an even number of data items, then we need to get the average of the middle numbers.)

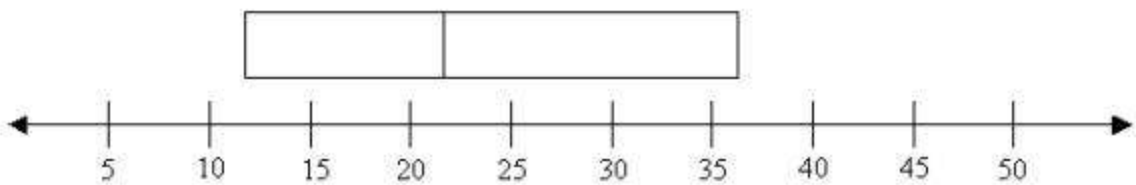
Step 3: Draw a number line that will include the smallest and the largest data.



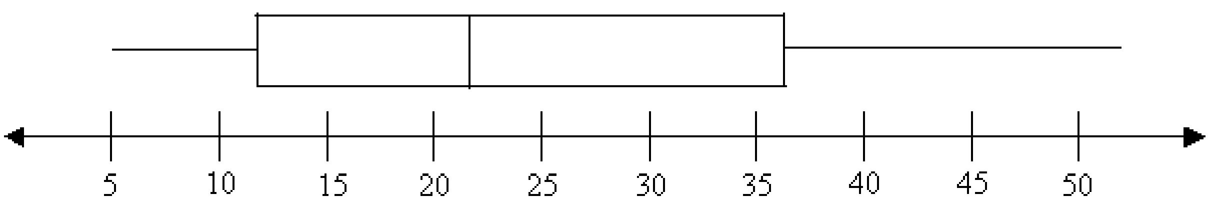
Step 4: Draw three vertical lines at the lower quartile (12), median (22) and the upper quartile (36), just above the number line.



Step 5: Join the lines for the lower quartile and the upper quartile to form a box.



Step 6: Draw a line from the smallest value (5) to the left side of the box and draw a line from the right side of the box to the biggest value (53).



For more knowledge about Measures of Relative Position, please check the link provided;
https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/02%3A_Descriptive_Statistics/2.04%3A_Relative_Position_of_Data
https://www.slideshare.net/AbdulAleem95/measures-of-relative-position

REMEMBER

- The percentile rank and *z*-score of a measurement indicate its relative position with regard to the other measurements in a data set.
- The three quartiles divide a data set into fourths.
- The five-number summary and its associated box plot summarize the location and distribution of the data.

ACTIVITY:

a. Consider the data set

69 93 70 53 92 75 85 70 68 76 88 70 77 82 85 82 80 100 96 85

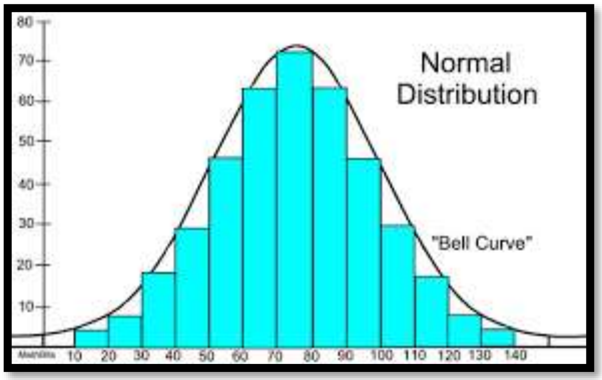
1. Find the percentile rank of 82.
2. Find the percentile rank of 68.

b. Find the z-score of each measurement in the following sample data set.

$$-5 \quad 6 \quad 2 \quad -1 \quad 0$$

Lesson 4: Normal Distribution

The **normal distribution** forms a bell-shaped curve that is symmetric about a vertical line through the mean of the data. A graph of a normal distribution is shown below:



Properties of Normal Distribution

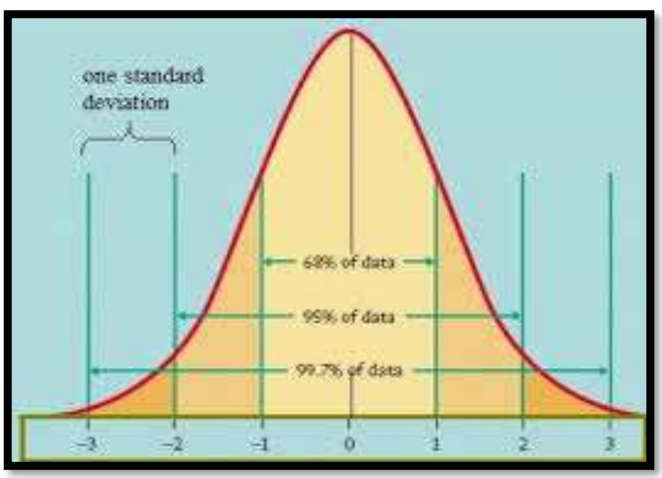
Every normal distribution has the following properties.

- The graph is symmetric about a vertical line through the mean of the distribution.
- The mean, median and mode are equal.
- The y-value of each point on the curve I the *percent* (expressed as a decimal) of the data at the corresponding *x-value.*
- Areas under the curve that are symmetric about the mean are equal.
- The total area under the curve is 1.

Empirical Rule for a Normal Distribution

In a normal distribution, approximately

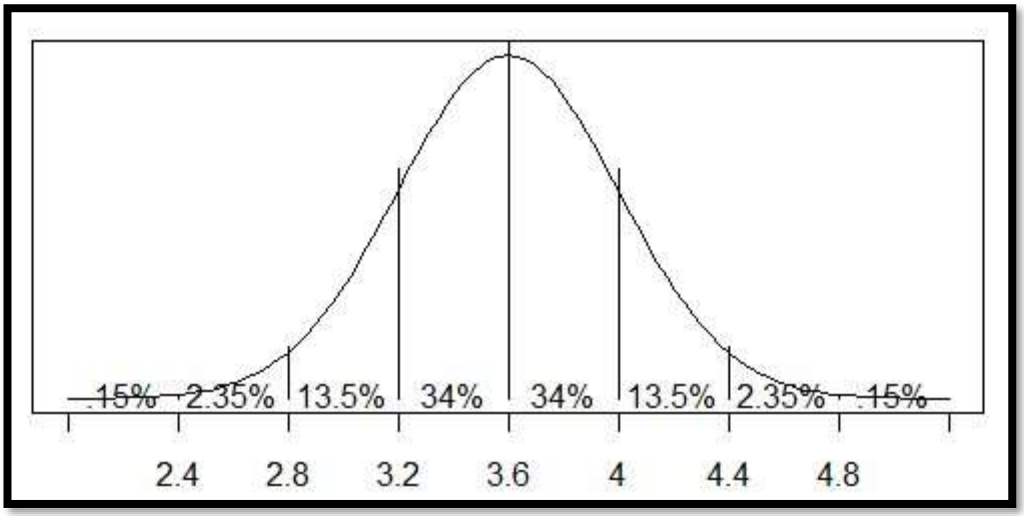
- 68% of the data lie within 1 standard deviation of the mean.
- 95% of the data lie within 2 standard deviations of the mean.
- 99.7% of the data is within 3 standard deviations of the mean.



Example:

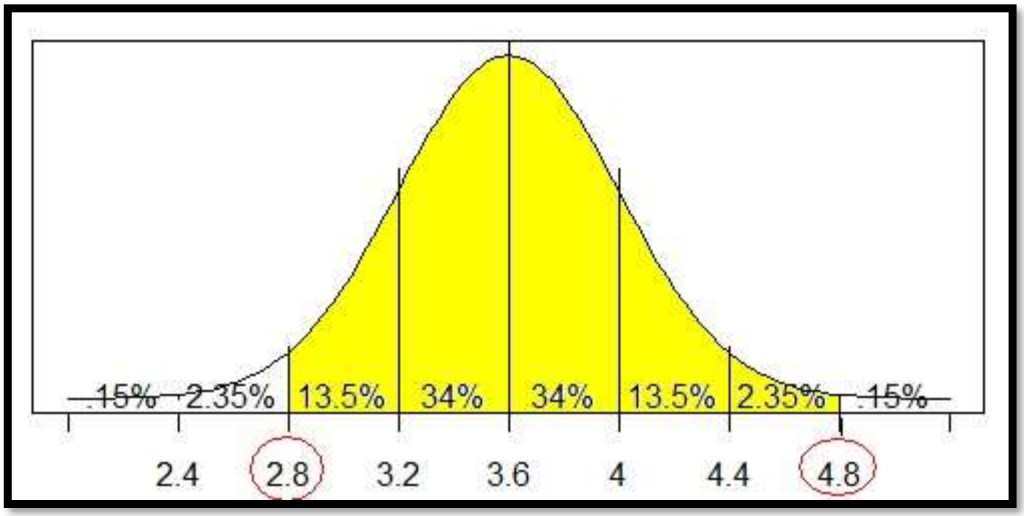
The weights of adorable, fluffy kittens are normally distributed with a mean of 3.6 pounds and a standard deviation of 0.4 pounds.

First, draw your Empirical curve with the 4 percentages! (Steps 1-3 are completed below.)



What percent of adorable, fluffy kittens weigh between 2.8 and 4.8 pounds?

Step 4: We need to shade the region they are asking for.

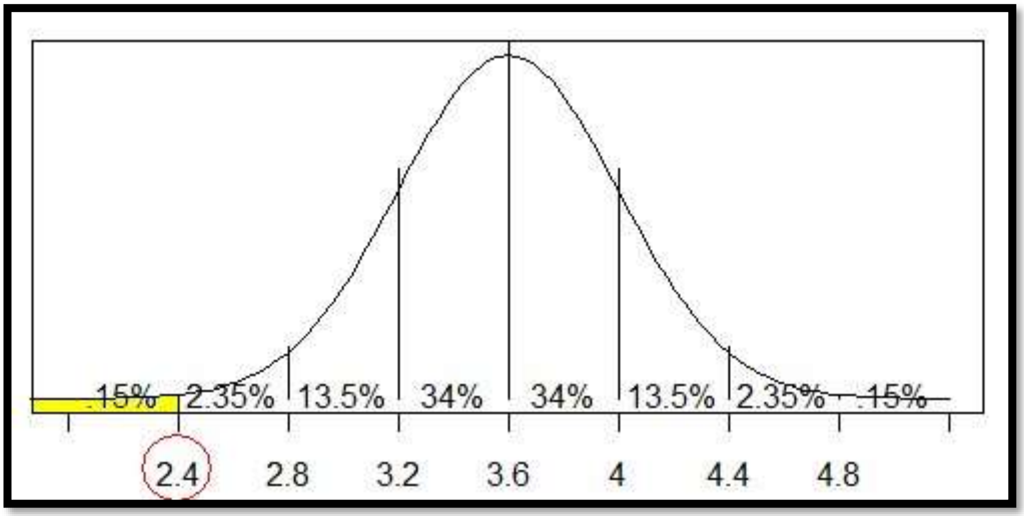


Step 5: We need to add the percent in the shaded areas.

13.5% + 34% + 34% + 13.5% + 2.35% = 97.35%

What percent of adorable, fluffy kittens weigh less than 2.4 pounds?

Step 4: We need to shade the region they are asking for.

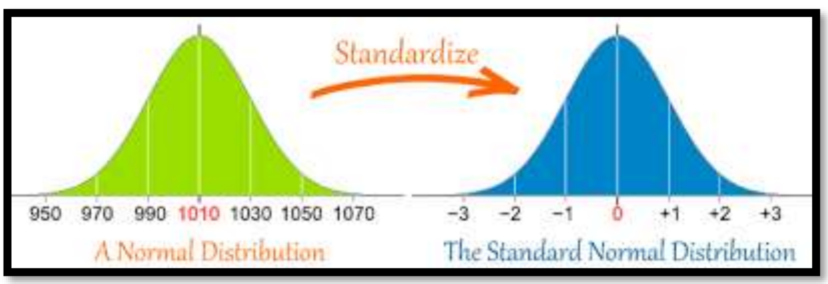


Step 5: We need to add the percents in the shaded areas.

0.15%

- **Standard Normal Distribution**

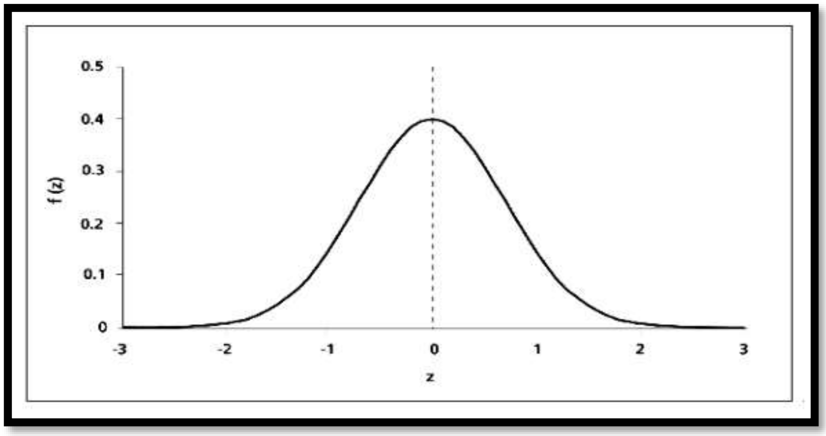
A normal distribution with a mean of 0 (u=0) and a standard deviation of 1 (o= 1).



The standard normal distribution (graph below) is a mathematical-or theoretical distribution that is frequently used by researchers to assess whether the distributions of the variables they are studying approximately follow a normal curve.

Every score in a normally distributed data set has an equivalent score in the standard normal distribution. This means that the standard normal distribution can be used to calculate the exact percentage of scores between any two points on the normal curve.

Statisticians have worked out tables for the standard normal curve that give the percentage of scores between any two points. In order to be able to use this table, scores need to be converted into Z scores.



For finding the area under the curve, the table is shown below:

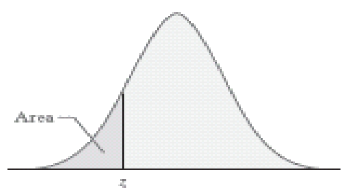


Table IV

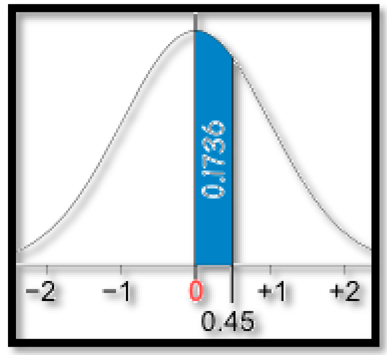
Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Example:

Percent of Population Between 0 and 0.45

Solution:



Start at the row for 0.4, and read along until 0.45: there is the value 0.1736

And 0.1736 is **17.36%**

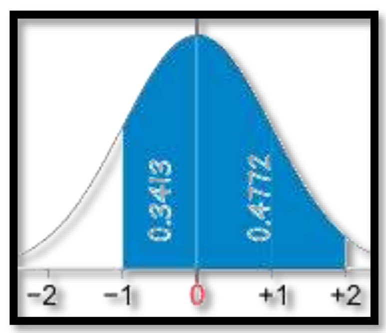
So 17.36% of the population are between 0 and 0.45 Standard Deviations from the Mean.

Because the curve is symmetrical, the same table can be used for values going either direction, so a negative 0.45 also has an area of 0.1736

Example:

Percent of Population Z Between -1 and 2

Solution:



From **−1 to 0** is the same as from **0 to +1**:

At the row for 1.0, first column 1.00, there is the value **0.3413**

From **0 to +2** is:

At the row for 2.0, first column 2.00, there is the value **0.4772**

Add the two to get the total between -1 and 2:

0.3413 + 0.4772 = **0.8185**

And **0.8185** is **81.85%**

So 81.85% of the population are between -1 and +2 Standard Deviations from the Mean.

Linear Regression

Linear regression finds the line that best fits the data points. There are actually a number of different definitions of "best fit," and therefore a number of different methods of linear regression that fit somewhat different lines. By far the most common is "ordinary least-squares regression"; when someone just says "least-squares regression" or "linear regression" or "regression," they mean ordinary least-squares regression.

Naming the Variables.

There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are:
(1) determining the strength of predictors
(2) forecasting an effect, and
(3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.

Third, regression analysis predicts trends and future values.

If the plot of n pairs of data (x , y) for an experiment appear to indicate a "linear relationship" between y and x, then the method of least squares may be used to write a linear relationship between x and y.
The least squares regression line is the line that minimizes the sum of the squares $(d1 + d2 + d3 + d4)$ of the vertical deviation from each data point to the line (see figure below as an example of 4 points).

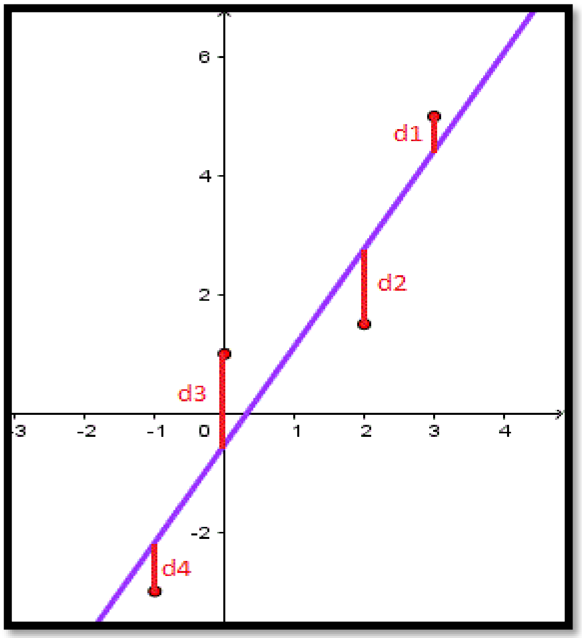


Figure 1. Linear regression where the sum of vertical distances d1 + d2 + d3 + d4 between observed and predicted (line and its equation) values is minimized.

The least square regression line for the set of n data points is given by the equation of a line in slope intercept form:

$$y = a x + b$$

where a and b are given by

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

Figure 2. Formulas for the constants a and b included in the linear regression

Example:

Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}
a) Find the least square regression line for the given data points.
b) Plot the given points and the regression line in the same rectangular system of axes.

Solution:

a) Let us organize the data in a table.

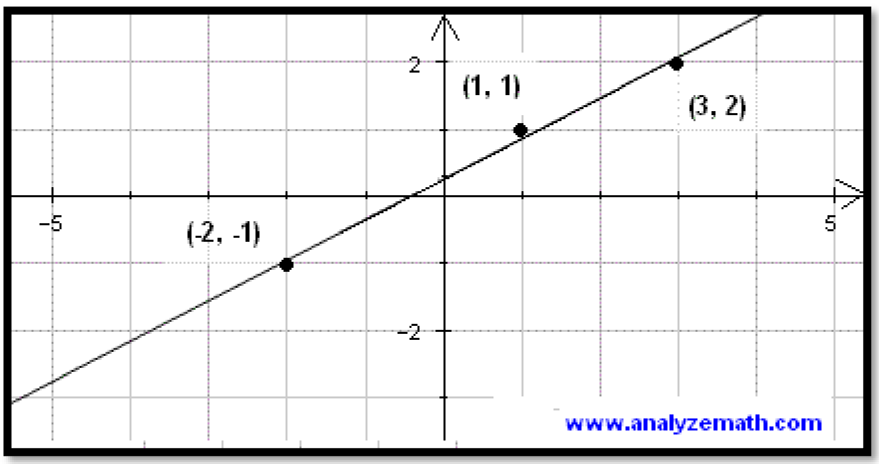
x	y	xy	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
$\Sigma x = 2$	$\Sigma y = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

We now use the above formula to calculate a and b as follows

$$a = \frac{(n\Sigma x\,y - \Sigma x \Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \;/\; = \frac{((3)(9) - (2)(2))}{((3)(14) - (2^2))} = \frac{23}{38}$$

$$b = \left(\frac{1}{n}\right)(\Sigma y - a\,\Sigma x) = \left(\frac{1}{3}\right)\left(2 - \left(\frac{23}{38}\right)(2)\right) = \frac{5}{19}$$

b) We now graph the regression line given by y = a x + b and the given points.

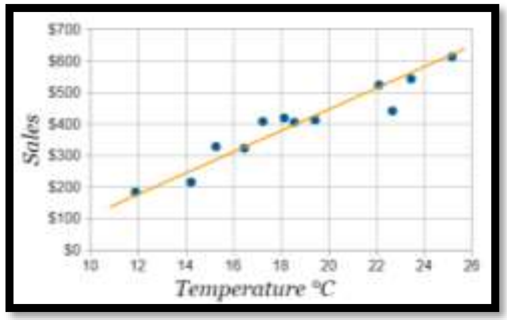


Graph of linear regression in problem 1.

The Least-Squares Regression Line

The **Least-Square Regression Line** for a set of bivariate data is the line that minimizes the sum of the squares of the vertical deviations from each data point to the line.

Imagine you have some points, and want to have a **line** that best fits them like this:



We can place the line "by eye": try to have the line as close as possible to all points, and a similar number of points above and below the line.

But for better accuracy let's see how to calculate the line using **Least Squares Regression**.

The Line

Our aim is to calculate the values **m** (slope) and **b** (y-intercept) in the equation of a line :

$$y = mx + b$$

Where:

- y = how far up
- x = how far along
- m = Slope or Gradient (how steep the line is)
- b = the Y Intercept (where the line crosses the Y axis)

Steps

To find the line of best fit for **N** points:

Step 1: For each (x, y) point calculate x^2 and xy

Step 2: Sum all x, y, x^2 and xy, which gives us $\Sigma x, \Sigma y, \Sigma x^2$ and Σxy

Step 3: Calculate Slope m:

$$\mathbf{m} = N\,\Sigma(xy) - \Sigma x\,\Sigma y N\,\Sigma(x^2) - (\Sigma x)^2$$

(N is the number of points.)

Step 4: Calculate Intercept $\mathbf{b}$:

$$\mathbf{b} = \Sigma y - m\, \Sigma x N$$

Step 5: Assemble the equation of a line

$$y = mx + b$$

Example:

Sam found how many hours of sunshine vs how many ice creams were sold at the shop from Monday to Friday:

"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

$$y = mx + b$$

Step 1: For each (x, y), calculate x^2 *and* xy

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum all $x, y, x^2 \, and \, xy$, which gives us $\Sigma x, \Sigma y, \Sigma x^2 \, and \, \Sigma xy$

x	y	x^2	Xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	**Σy: 41**	**Σx^2: 168**	**Σxy: 263**

Also **N** (number of data values) = **5**

Step 3: Calculate Slope **m**:

$$\mathbf{m} = N\ \Sigma(xy)\ -\ \Sigma x\ \Sigma y N\ \Sigma\left(x^2\right)\ -\ (\Sigma x)^2$$

$$\mathbf{m} = 5\ x\ 263\ -\ 26\ x\ 415\ x\ 168\ -\ 26^2$$

$$\mathbf{m} = 1315\ -\ 1066840\ -\ 676$$

$$\mathbf{m} = 249164$$

$$\mathbf{m} =\ 1.5183...$$

Step 4: Calculate Intercept **b**:

$$\mathbf{b} = \Sigma y\ -\ m\ \Sigma x N$$

$$\mathbf{b} = 41\ -\ 1.5183\ x\ 265$$

$$\mathbf{b} =\ 0.3049...$$

Step 5: Assemble the equation of a line:

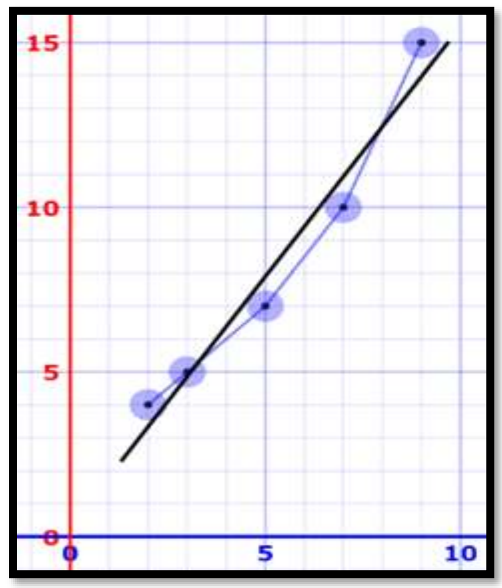
$$y\ =\ mx\ +\ b$$

$$y\ =\ 1.518x\ +\ 0.305$$

Let's see how it works out:

x	y	y = 1.518x + 0.305	error
2	4	3.34	−0.66
3	5	4.86	−0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	−1.03

Here are the (x,y) points and the line $\mathbf{y} = \mathbf{1.518x} + \mathbf{0.305}$ on a graph:



Sam hears the weather forecast which says "we expect 8 hours of sun tomorrow", so he uses the above equation to estimate that he will sell

$$y = 1.518\, x\; 8 + 0.305 = 12.45\ \textit{Ice Creams}$$

Sam makes fresh waffle cone mixture for 14 ice creams just in case. Yum.

Linear Correlation

To determine the strength of a linear relationship between two variables, statisticians use a statistics called the *linear correlation coefficient,* which is denoted by the variable *r* and is defined as follow:

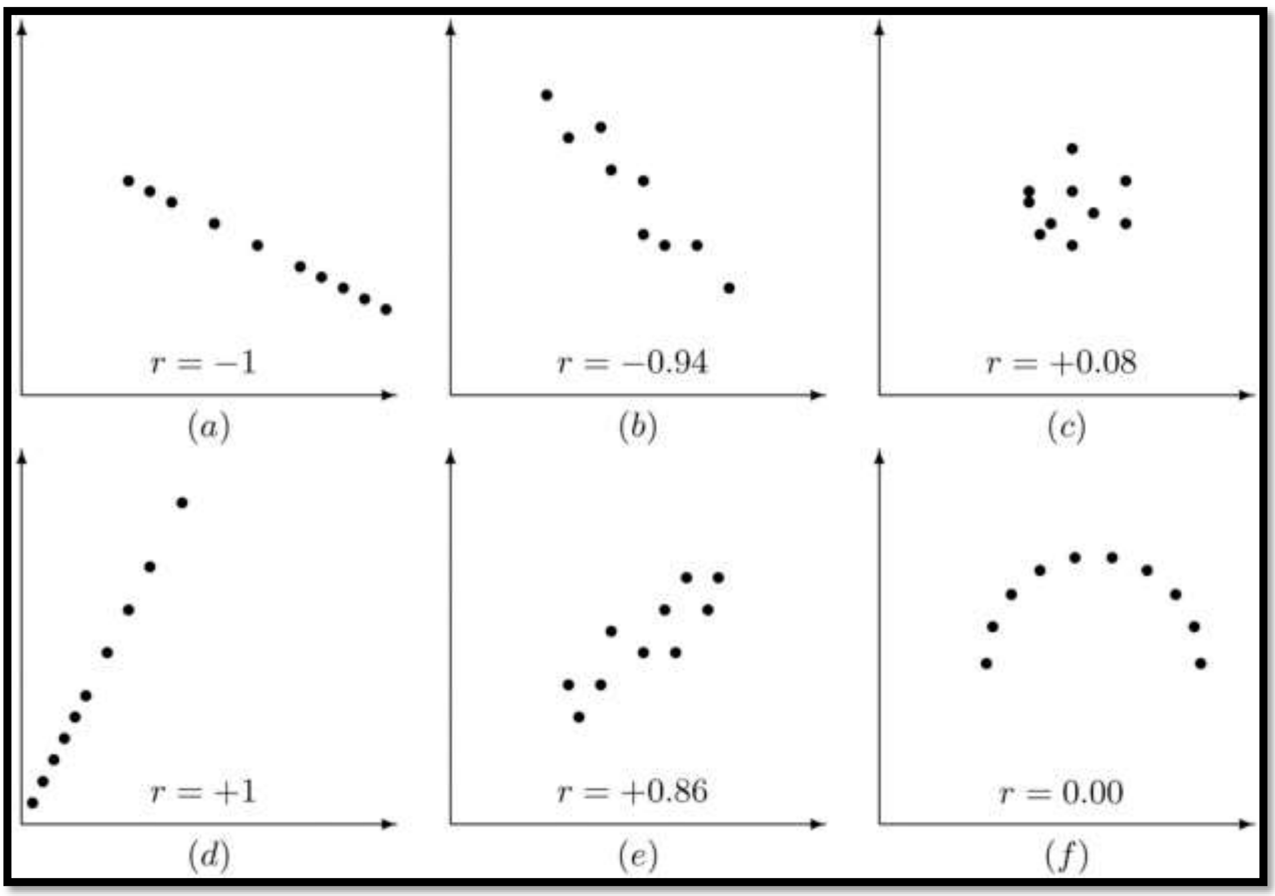
Linear Correlation Coefficient

For the *n* ordered pairs $(x_{1,}y_1),(x_{2,}y_2),(x_{3,}y_3),...,(x_{n,}y_n)$, the **linear correlation coefficient** *r* is given by

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

Properties of Linear Correlation Coefficient

1. The value of *r* lies between −1 and 1, inclusive.
2. The sign of *r* indicates the direction of the linear relationship between *x* and *y*:
    1. If r $r < 0$ then *y* tends to decrease as *x* is increased.
    2. If $r > 0$ then *y* tends to increase as *x* is increased.
3. The size of |*r*| indicates the strength of the linear relationship between *x* and *y*:
    1. If |*r*| is near 1 (that is, if *r* is near either 1 or −1) then the linear relationship between *x* and *y* is strong.
    2. If |*r*| is near 0 (that is, if *r* is near 0 and of either sign) then the linear relationship between *x* and *y* is weak.



Example:

Calculate the linear correlation coefficient for the following data.

$X = 4, 8, 12, 16$ and $Y = 5, 10, 15, 20.$

Solution:

Given variables are,

$X = 4, 8, 12, 16$ and $Y = 5, 10, 15, 20.$

For finding the linear coefficient of these data, we need to first construct a table for the required values.

X	y	x^2	y^2	XY
4	5	16	25	20
8	10	64	100	80
12	15	144	225	180
16	20	256	400	320
Σ x = 40	Σ y =50	480	750	600

According to the formula of linear correlation we have,

$$r(xy) = \frac{400}{\sqrt{320} \cdot \sqrt{500}}$$

$$r(xy) = \frac{400}{17.89 \cdot 22.36}$$

$$r(xy) = \frac{400}{400}$$

$$r(xy) = 1$$

Therefore, $r(xy) = 1$

For more knowledge about Normal Distribution please check the link provided;
https://statisticsbyjim.com/basics/normal-distribution/
https://study.com/academy/lesson/standard-normal-distribution-definition-example.html

REMEMBER

- The linear correlation coefficient measures the strength and direction of the linear relationship between two variables *x* and *y*.
- The sign of the linear correlation coefficient indicates the direction of the linear relationship between *x* and *y*.
- When *r* is near 1 or −1 the linear relationship is strong; when it is near 0 the linear relationship is weak.

ACTIVITY:
Solve the following problem.

1. X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find

   a) $P(x < 40)$

   b) $P(x > 21)$

   c) $P(30 < x < 35)$

https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+measures+of+central+tendency
https://www.toppr.com/guides/business-mathematics-and-statistics/measures-of-central-tendency-and-dispersion/measure-of-dispersion/
https://stattrek.com/descriptive-statistics/measures-of-position.aspx
https://statisticsbyjim.com/basics/normal-distribution/