CHAPTER 4

Correlation and Regression

Inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists.

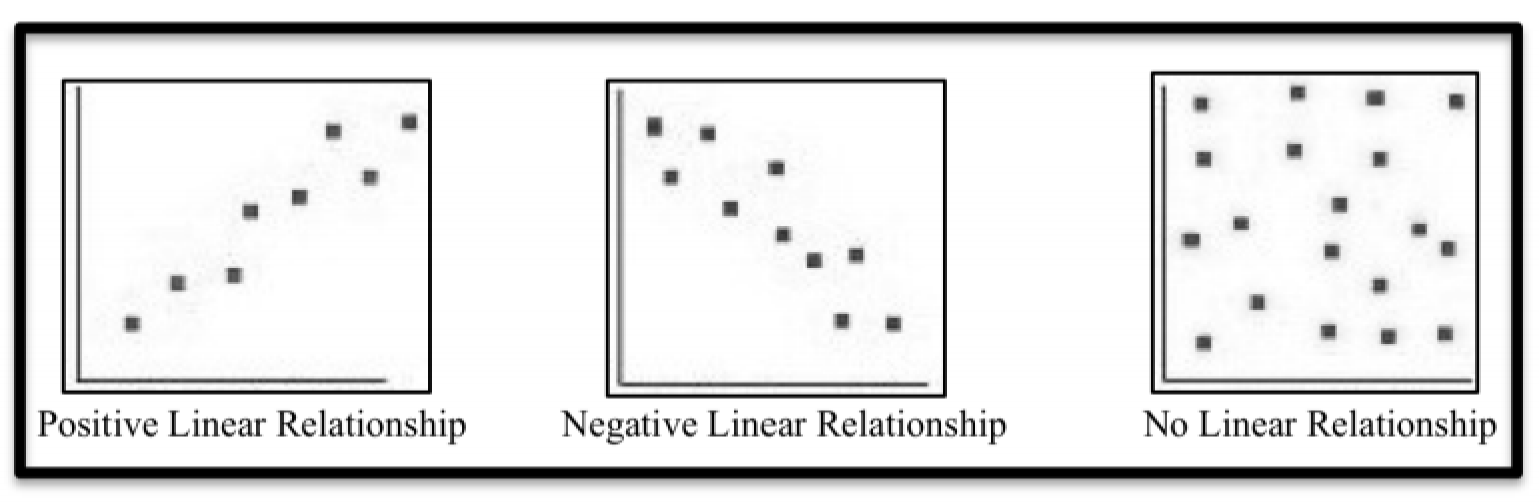
- *Correlation is a statistical method used to determine whether a relationship between variables exists.*

- *Regression is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.*

In a simple relationship, there are only two types of variables under study; an independent variable, and a dependent variable.

- *Simple relationship can be positive or negative:*
  - *A positive relationship exists when both variables increase or decrease at the same time.*
  - *A negative relationship exists when one variable increases and the other variable decreases.*

- *A scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables.*



Correlation Coefficient

- *The correlation coefficient is measure of how variables are related, it measures the strength and direction of a linear relationship between two variables.*
- *The symbol for the population correlation coefficient is* ρ *(rho).*
- *The symbol for the sample correlation coefficient is r.*
- *The range of the correlation coefficient is from -1 to +1*

- If there is a strong positive linear relationship between the variables, the value of r will be close to +1.

- If there is a strong negative linear relationship between the variables, the value of r will be close to -1.

- When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.

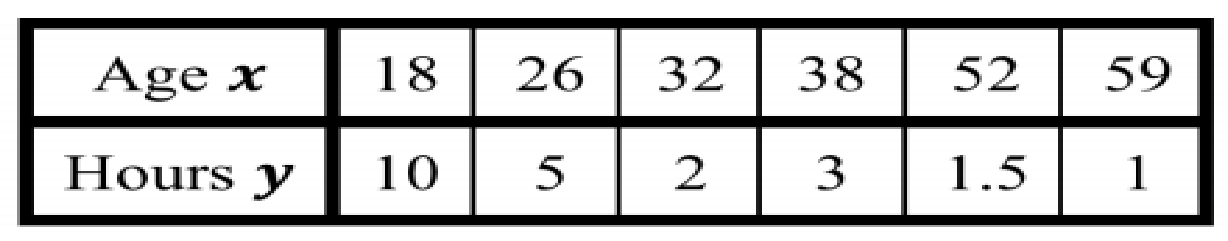
Pearson linear correlation coefficient

- *Formula for the Pearson linear correlation coefficient (r)*

$$\boldsymbol{r_P} = \frac{\boldsymbol{n}(\sum \boldsymbol{xy}) - (\sum \boldsymbol{x})(\sum \boldsymbol{y})}{\sqrt{[\boldsymbol{n}(\sum \boldsymbol{x}^2) - (\sum \boldsymbol{x})^2][\boldsymbol{n}(\sum \boldsymbol{y}^2) - (\sum \boldsymbol{y})^2]}}$$

where n is the number of data pairs (sample size).

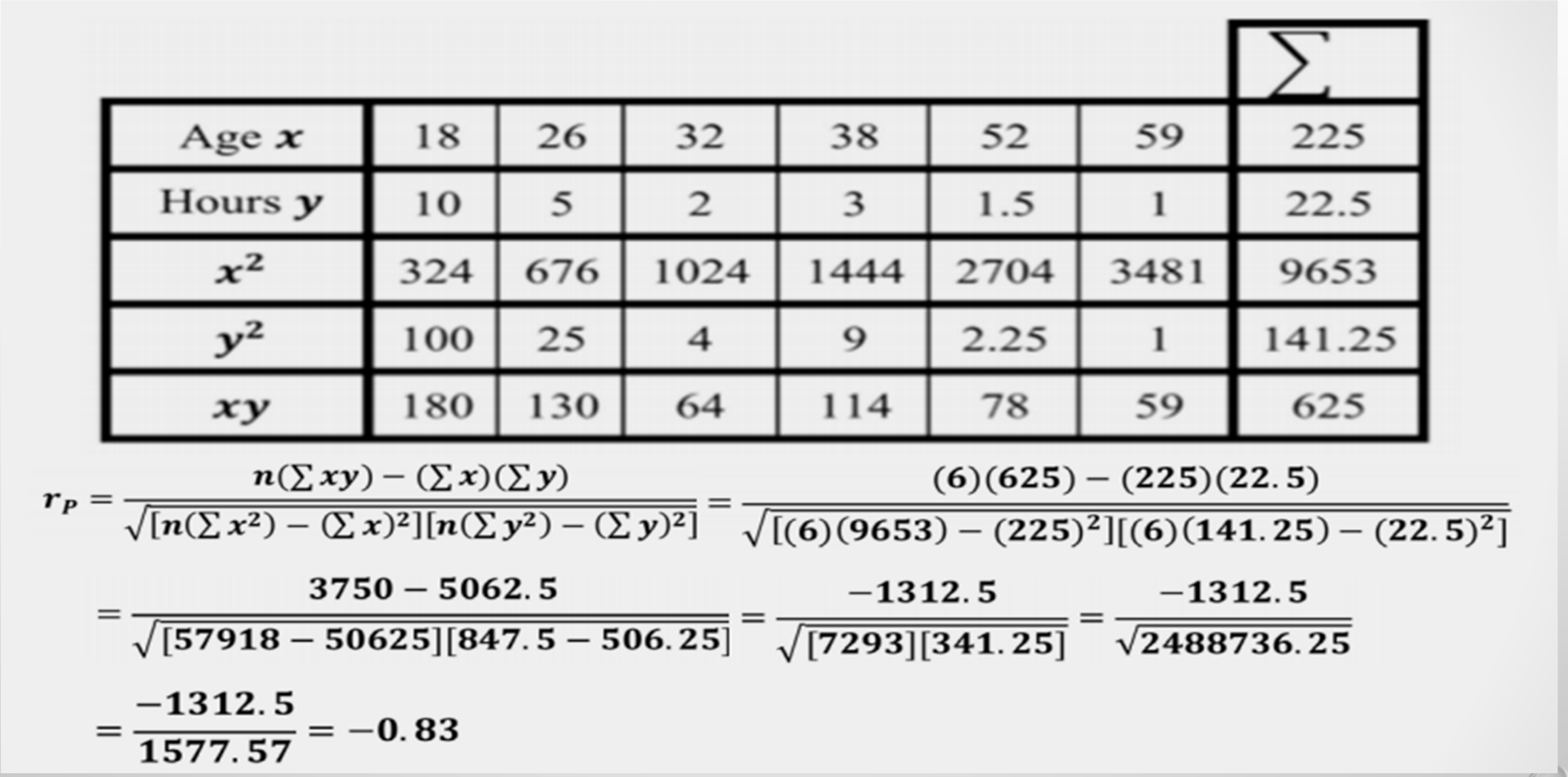
Ex: A researcher wishes to determine if a person's age is related to the number of hours he or she exercises per week. The data for the sample are shown below.



Age x	18	26	32	38	52	59
Hours y	10	5	2	3	1.5	1

Compute the value of the correlation coefficient.

Solution:



							$\sum$
Age x	18	26	32	38	52	59	225
Hours y	10	5	2	3	1.5	1	22.5
x^2	324	676	1024	1444	2704	3481	9653
y^2	100	25	4	9	2.25	1	141.25
xy	180	130	64	114	78	59	625

$$r_P = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} = \frac{(6)(625) - (225)(22.5)}{\sqrt{[(6)(9653) - (225)^2][(6)(141.25) - (22.5)^2]}}$$

$$= \frac{3750 - 5062.5}{\sqrt{[57918 - 50625][847.5 - 506.25]}} = \frac{-1312.5}{\sqrt{[7293][341.25]}} = \frac{-1312.5}{\sqrt{2488736.25}}$$

$$= \frac{-1312.5}{1577.57} = -0.83$$

There is a negative linear relationship, which means that older people tend to exercise less on average.

Spearman rank correlation coefficient

- *Formula for the Spearman rank correlation coefficient (rS)*

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

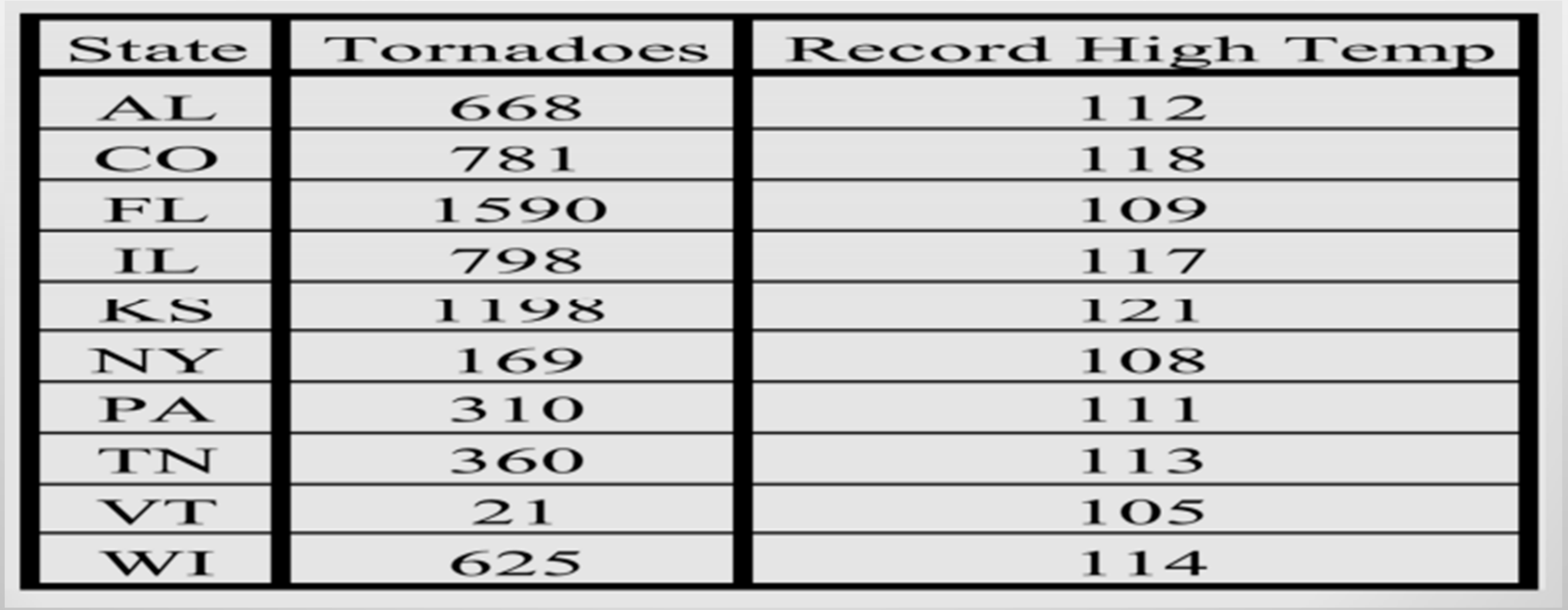
where

d = difference in the ranks and

n = number of data pairs

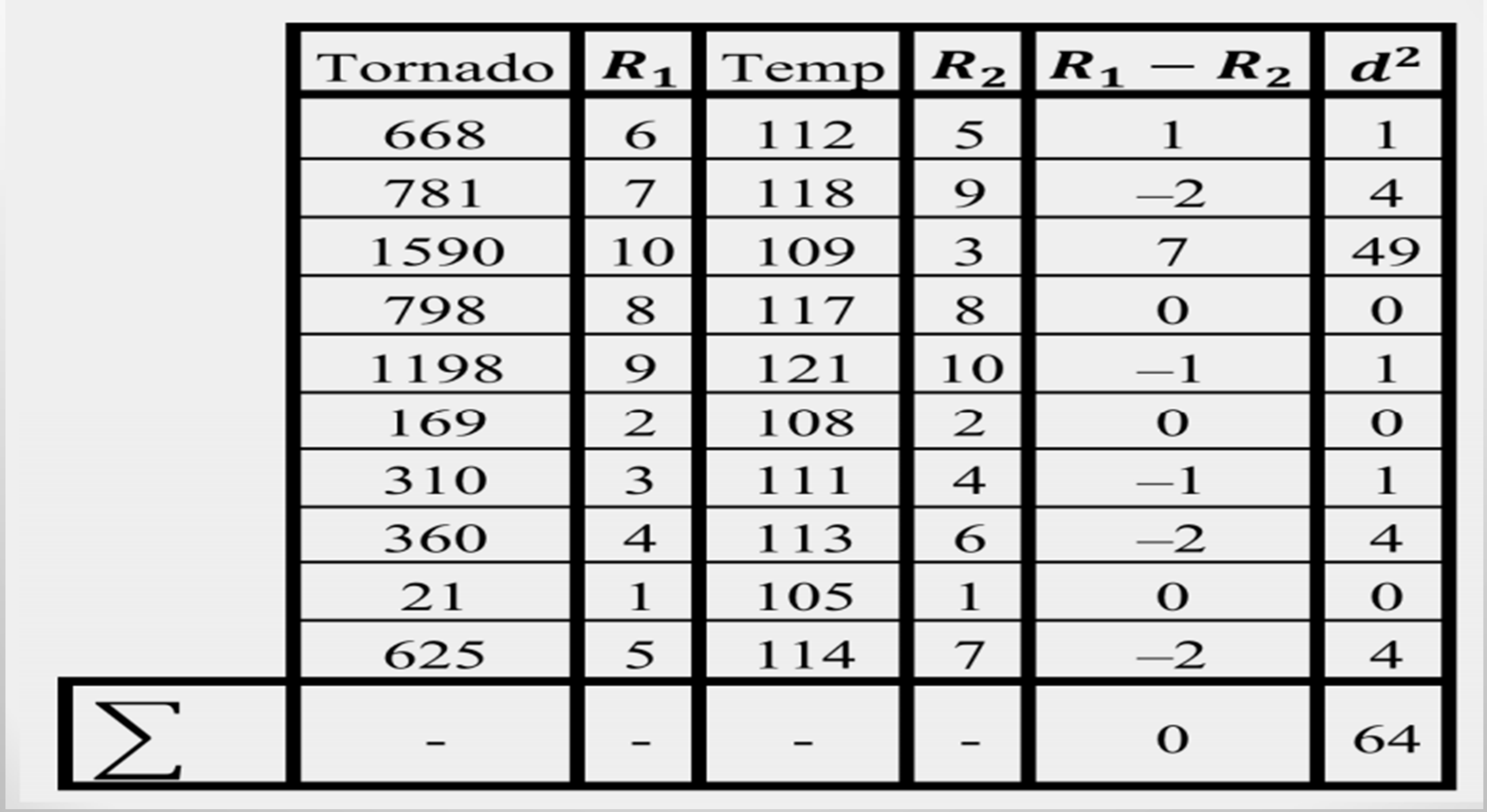
Ex: The table below shows the total number of tornadoes that occurred at some states from 1962 to 1991 and the record high temperatures for the same states.

Is there a relationship between the number of tornadoes and the record high temperatures?

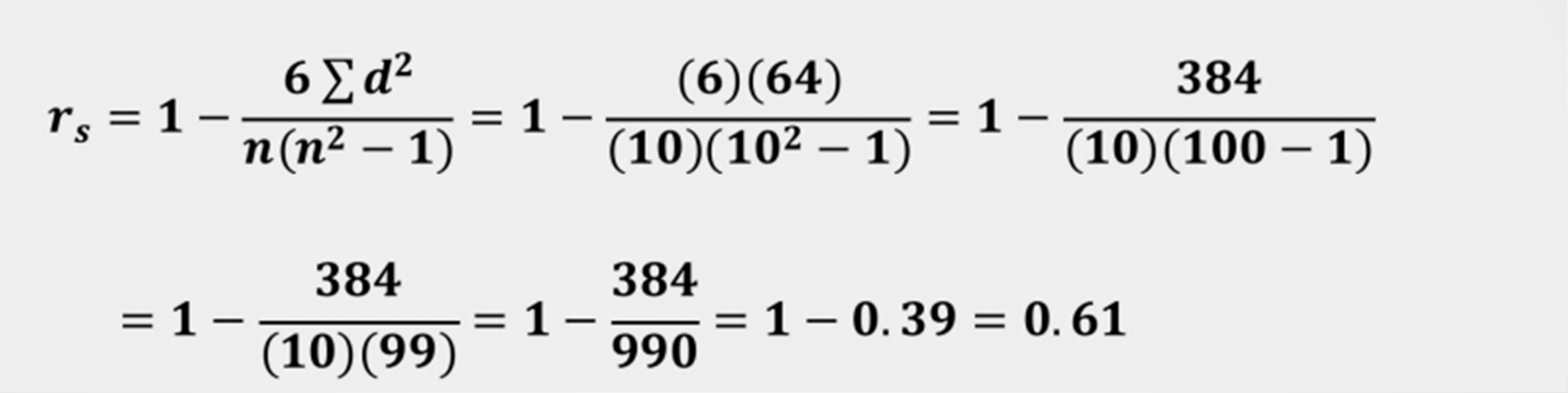


State	Tornadoes	Record High Temp
AL	668	112
CO	781	118
FL	1590	109
IL	798	117
KS	1198	121
NY	169	108
PA	310	111
TN	360	113
VT	21	105
WI	625	114

Solution:



	Tornado	R_1	Temp	R_2	$R_1 - R_2$	d^2
	668	6	112	5	1	1
	781	7	118	9	–2	4
	1590	10	109	3	7	49
	798	8	117	8	0	0
	1198	9	121	10	–1	1
	169	2	108	2	0	0
	310	3	111	4	–1	1
	360	4	113	6	–2	4
	21	1	105	1	0	0
	625	5	114	7	–2	4
$\sum$	-	-	-	-	0	64



$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{(6)(64)}{(10)(10^2-1)} = 1 - \frac{384}{(10)(100-1)}$$

$$= 1 - \frac{384}{(10)(99)} = 1 - \frac{384}{990} = 1 - 0.39 = 0.61$$

There is a positive linear relationship between the number of tornados and the record high temperatures.

Regression Line

- *If the value of the correlation coefficient is significant (will not be discussed here), the next step is to determine the equation of the regression line, which is the data's line of best fit.*

- *Best fit means that the sum of the squares of the vertical distance from each point to the line is at a minimum.*

*** The Weighted Mean*

The type of mean that considers an additional factor is called the weighted mean, and it is used when the values are not all equally represented.

$$\overline{X} = \frac{w_1X_1 + w_2X_2 + \cdot \cdot \cdot + w_nX_n}{w_1 + w_2 + \cdot \cdot \cdot + w_n} = \frac{\Sigma wX}{\Sigma w}$$

Find the weighted mean of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights. where w1, w2, . . . , wn are the weights and X1, X2, . . . , Xn are the values

A student received an A in English Composition I (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology I (4 credits), and a D in Physical Education (2 credits). Assuming A = 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

Solution

Course	*Credits (w)*	*Grade (X)*
English I	*3*	*A (4 points)*
Psychology	*3*	*C (2 points)*
Biology I	*4*	*B (3 points)*
Physical	*2*	*D (1 point)*

$$\overline{X} = \frac{\Sigma wX}{\Sigma w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Properties and Uses of Central Tendency

The Mean

1. The mean is found by using all the values of the data.

4. The mean for the data set is unique and not necessarily one of the data values.

6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The Median

1. The median is used to find the center or middle value of a data set.

2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.

4. The median is affected less than the mean by extremely high or extremely low values.

The Mode

1. The mode is used when the most typical case is desired.

2. The mode is the easiest average to compute.

3. The mode can be used when the data are nominal or categorical, such as religious preference, gender, or political affiliation.

4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

The Midrange

1. The midrange is easy to compute.

2. The midrange gives the midpoint.

3. The midrange is affected by extremely high or low values in a data set.

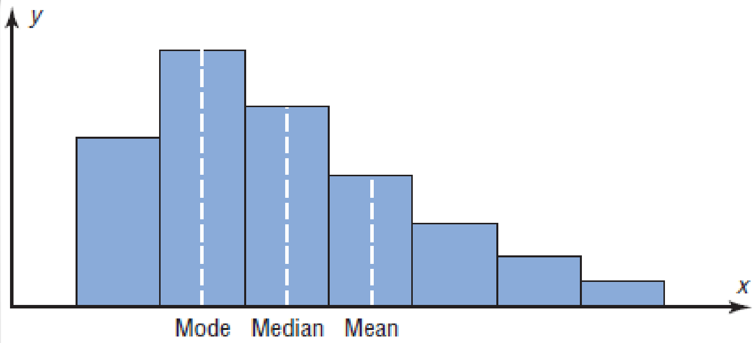
Distribution Shapes

Frequency distributions can assume many shapes:

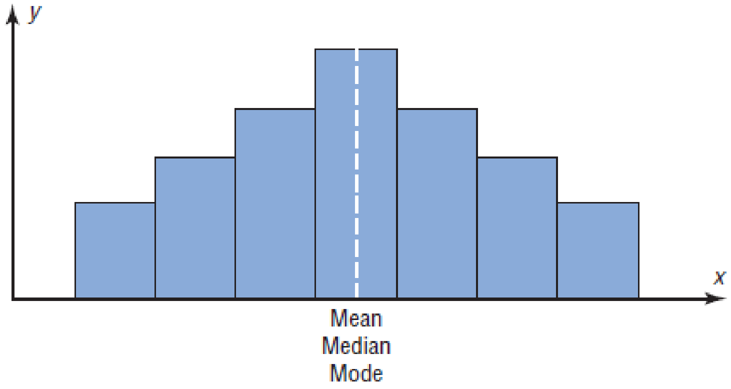
1) In a ***positively skewed or right-skewed distribution****, the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution. Also, the mean is to the right of the median, and the mode is to the left of the median.*

2) In a ***symmetric distribution****, the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution.*

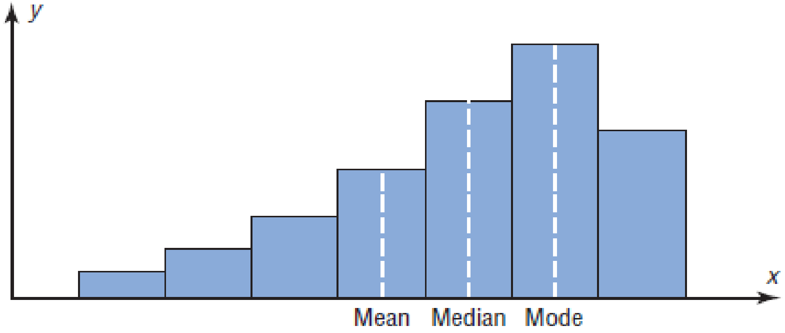
3) In a ***negatively skewed or left-skewed****, the data values fall to the right of the mean., the mean is to the left of the median, and the mode is to the right of the median.*



(a) Positively skewed or right-skewed



(b) Symmetric



(c) Negatively skewed or left-skewed

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency

Measures of Variation:

For the spread or variability of a data set, three measures are commonly used: range, variance, and standard deviation. Each measure will be discussed in this section.

Range

The range is the highest value minus the lowest value. The symbol R is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

Find the ranges for the paints:

(a) Brand A

10 20 30 35 40 50 60

(b) Brand B

25 30 35 40 45

Solution

For brand A, the range is

R = 60 - 10 = 50 months

For brand B, the range is

R = 45 - 25 = 20 months

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A’s range.

Population Variance and Standard Deviation

(Just the symbol & Formula)

To have a more meaningful statistic to measure the variability, statisticians use measures called the variance and standard deviation.

Population Variance:

The variance is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 *(s is the Greek lowercase letter sigma). The formula for the population variance is*

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

where

X : individual value

m : population mean

N : population size

standard deviation :

The standard deviation is the square root of the variance. The symbol for the population standard deviation is σ*. The corresponding formula for the population standard deviation is:*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Sample Variance and Standard Deviation

Sample Variance:

The formula for the sample variance, denoted by S^2*, is*

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$$

where

n: sample size

Standard Deviation :

The standard deviation is the square root of the variance. The symbol for the population standard deviation is S*. The corresponding formula for the population standard deviation is:*

$$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$$

Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Solution

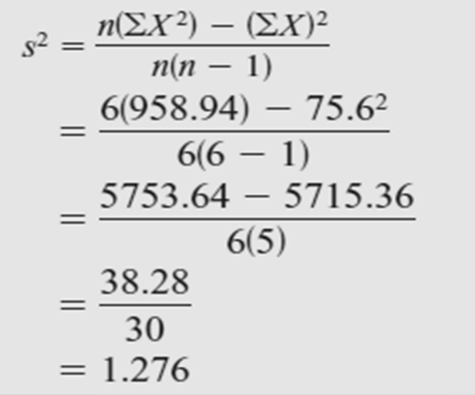
Step 1 Find the sum of the values.

$\sum X$ *= 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 + 75.6*

Step 2 Square each value and find the sum.

$\sum X^2$ *= 11.22 + 11.92 + 12.02 + 12.82 + 13.42 + 14.32 + 958.94*

Step 3 Substitute in the formulas and solve.



$$
\begin{aligned}
s^2 &= \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)} \\
&= \frac{6(958.94) - 75.6^2}{6(6-1)} \\
&= \frac{5753.64 - 5715.36}{6(5)} \\
&= \frac{38.28}{30} \\
&= 1.276
\end{aligned}
$$

The variance is 1.28 rounded.

***** Hence, the sample standard deviation is***

S = $\sqrt{1.28}$ = *1.13.*

Coefficient of Variation:

Whenever two samples have the same units of measure, the variance and standard deviation for each can be compared directly. But the statistic using coefficient of variation to compare the standard deviations when the units are different.

The ***coefficient of variation****, denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.*

For samples, the

$$CVar = \frac{S}{m} \times 100\%$$

Ex: The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution:

The coefficients of variation are

$$CVar = \frac{773}{5225} \times 100\% = 14.8\% \qquad \text{commissions}$$

$$CVar = \frac{5}{87} \times 100\% = 5.7\% \text{ sales}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Measures of Position

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set.

Standard Scores:

A z score or standard score for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is z. The formula for samples is:

$$z = \frac{value\ - mean}{\text{standard deviation}} = \frac{X\ -\ \bar{X}}{S}$$

The z score represents the number of standard deviations that a data value falls above or below the mean.

EX: A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution:

First, find the z scores. For calculus the z score is

$$z = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1$$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

When all data for a variable are transformed into z scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A z score, then, is actually the number of standard deviations each value is from the mean for a specific distribution. In the last example the calculus score of 65 was actually 1.5 standard deviations above the mean of 50.

Percentiles:

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group.

*** Percentiles** *divide the data set into 100 equal groups.*

Percentile Formula:

(1) *The percentile corresponding to a given value X is computed by using the following formula:*

$$\text{Percentile} = \frac{\text{number of values below } X \ + 0.5}{\text{total number of values}} * 100$$

(2) *The steps for finding a value corresponding to a given percentile:*

Step 1 *Arrange the data in order from lowest to highest.*

Step 2 *Substitute into the formula*

$$C = \frac{n \times p}{100} \quad \text{where: n: total number of values, p: percentile}$$

Step 3

(A) *If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.*

(B) *If c is a whole number, use the value halfway between the cth and (c 1)st values when counting up from the lowest value.*

EX: A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12 and find the value corresponding to the 25th percentile

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution:

** Arrange the data in order from lowest to highest.*

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula.

$$\text{Percentile} = \frac{6+0.5}{10} * 100 = 65\text{th}$$

*** Substitute in the formula* $C = \frac{10 \times 25}{100} = 2.5$

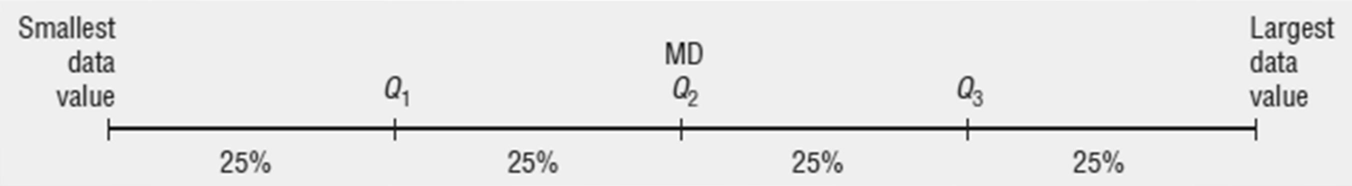
If c is not a whole number, round it up to the next whole number; in this case, c = 3. Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds

to the 25th percentile.

Quartiles:

Quartiles *divide the distribution into four groups, separated by Q1, Q2, Q3.*

Note that Q1 is the same as the 25th percentile; Q2 is the same as the 50th percentile, or the median; Q3 corresponds to the 75th percentile, as shown:



However, an easier method for finding quartiles is found in this Procedure Table:

1. Arrange the data in order from lowest to highest.

2. Find the median of the data values (Q2).

3. Find the median of the data values that fall bellow Q2 (Q1).

4. Find the median of the data values that fall above Q2 (Q3).

EX: Find Q1, Q2, and Q3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

Solution

Step 1 Arrange the data in order.

5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find the median (Q2).

5, 6, 12, 13, 15, 18, 22, 50

↑

MD = (13 +15)/2 = 14

Step 3 Find the median of the data values less than 14.

5, 6, 12, 13

↑

Q1 = (6+12)/2 = 9

Step 4 Find the median of the data values greater than 14.

15, 18, 22, 50

↑

Q3 = (18+22)/2 =20

Outlier :

An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.

- *Outliers can be identified as follows:*

1. Arrange the data in order and find Q1 and Q3.

2. Find the interquartile range: IQR=Q3-Q1.

3. The values that are smaller than Q1-(1.5)(IQR) or larger than Q3+(1.5)(IQR) are called outliers.

EX: Check the following data set for outliers. (using the last example)

Solution

Step 1 Arrange the data in order.

5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find Q1 and Q3.

Q1 = 9 and Q3 = 20.

2. Find the interquartile range: IQR=Q3-Q1 = 20 – 9 = 11

3. The values that are smaller than 9-(1.5)(11) = -7.5 or larger than 20+(1.5)(11) = 36.5 are called outliers. The value 50 is outside this interval; hence, it can be considered an outlier.

Exploratory Data Analysis:

In exploratory data analysis (EDA), the data are represented graphically using a boxplot (sometimes called a box-and-whisker plot). The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data such as the center and the spread

**** The Five-Number Summary and Boxplots***

A boxplot can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)

2. Q1

3. The median

4. Q3

5. The highest value of the data set (i.e., maximum)

These values are called a five-number summary of the data set.

Information obtained from a boxplot

* *Using the box:*

- If the median is near the center of the box, the distribution is approximately symmetric.

- If the median is to the left of the box, the distribution is positively skewed.

- If the median is to the right of the box, the distribution is negatively skewed.

* *Using the lines:*

- If the lines are about the same length, the distribution is approximately symmetric.

- If the right line is taller, the distribution is positively skewed.

- If the left line is taller, the distribution is negatively skewed.

Information obtained from a boxplot

** Using the box:*

- If the median is near the center of the box, the distribution is approximately symmetric.

- If the median is to the left of the box, the distribution is positively skewed.

- If the median is to the right of the box, the distribution is negatively skewed.

** Using the lines:*

- If the lines are about the same length, the distribution is approximately symmetric.

- If the right line is taller, the distribution is positively skewed.

- If the left line is taller, the distribution is negatively skewed.