Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Dr. Saeed's Website

Statistics
for
Business & Economics

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Describing Data: Numerical Measures

Chapter 3

McGraw-Hill/Irwin

Learning Objectives

LO1 Differentiate between a statistic and a parameter.

LO2 Compute and interpret the measures of location.

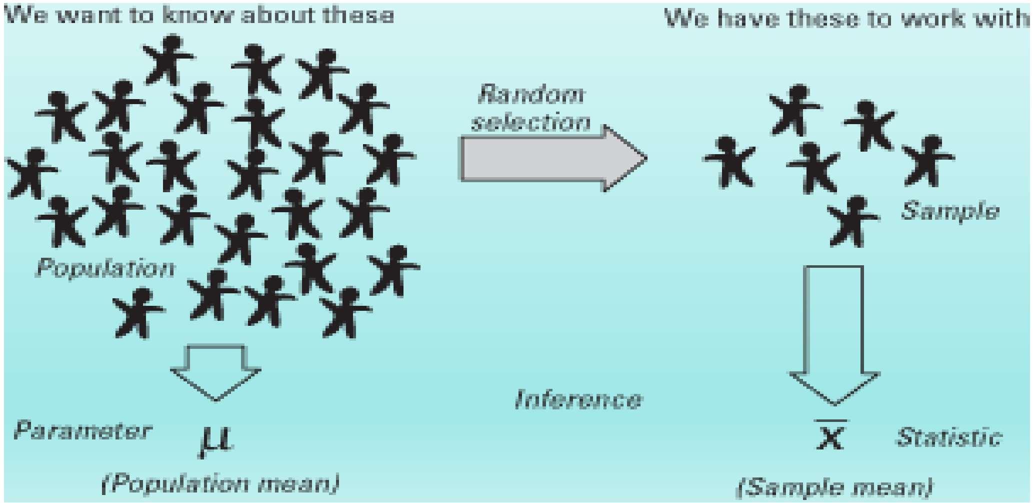
LO3 Compute and interpret the measures of variability.

LO4 Compute and interpret the measures of shape

LO5 Identify the techniques of exploratory data analysis.

LO6 Compute and interpret the measures of association.

Parameter Versus Statistics



PARAMETER is a measurable characteristic of a *population*. It is calculated using all the data values of the population under study, e.g., the population mean.

STATISTIC is a measurable characteristic of a *sample*. It is calculated using the data values of a sample from the population under study, e.g., the sample mean.

Mean

MEAN is the sum of the values divided by the number of values.

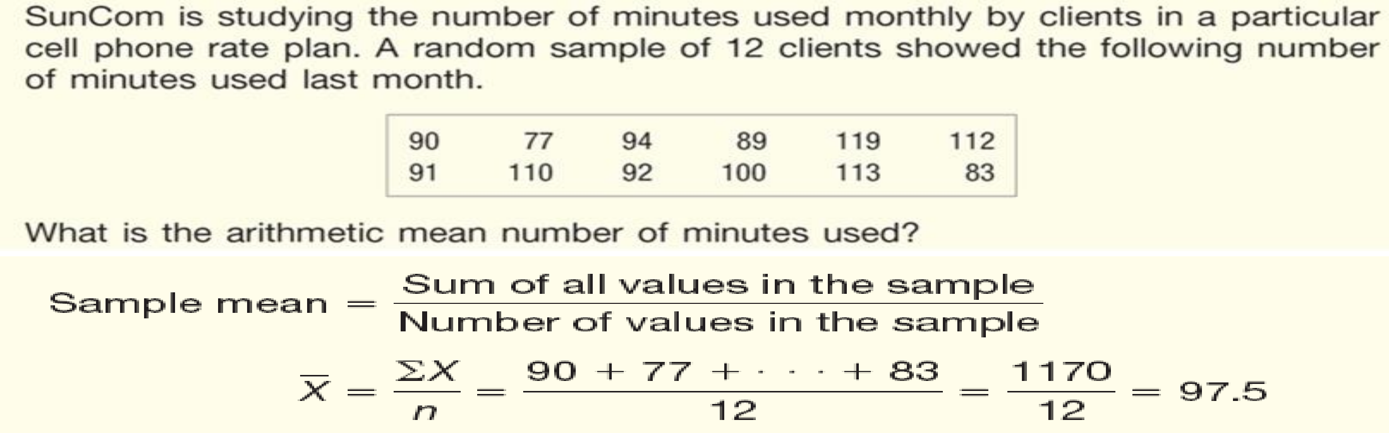
- The Greek letter μ (mu) is used to represent the population mean.

$$\mu = \frac{\sum X}{N} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

- The symbol $\bar{X}$(“x-bar”) is used to represents the sample mean.

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

EXAMPLE:



SunCom is studying the number of minutes used monthly by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

90	77	94	89	119	112
91	110	92	100	113	83

What is the arithmetic mean number of minutes used?

$$\text{Sample mean} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{90 + 77 + \cdots + 83}{12} = \frac{1170}{12} = 97.5$$

Properties of The Mean

- The most important measures of location for a variable.
- The *mean* is computed by using all the values of a data set.
- The *mean* varies less than the other measures of central tendency.
- The *mean* for a data set is unique, and not necessarily one of the data values.
- The sum of the deviatic $\sum(X - \bar{X}) = 0$ om the mean is zero. Thus,
- The *mean* is affected by extremely high or low values and may not be the appropriate average.

Weighted Mean

- The ***weighted mean*** is used when the values in a data set are not all equally represented. Thus, there are several observations of the same value.
- For example, suppose a grocery store sold small, medium and large-sized drinks for 1, 1.5 and 2 SAR respectively. Of the last 10 drinks sold, 5 were small, 2 were medium and 3 were large. To find the mean price of the last 10 sold, we could use the mean formula as

$$\bar{x} = \frac{\sum x}{n} = \frac{1+1+1+1+1+1.5+1.5+2+2+2}{10}$$

or use the mean formula as

$$\bar{x} = \frac{5\times1+2\times1.5+3\times2}{5+2+3}$$

which is the weighted mean formula.

The **WEIGHTED MEAN of a variable** x is found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights

$$\bar{X}_w = \frac{w_1x_1+w_2x_2+\ldots+w_nx_n}{w_1+w_2+\ldots+w_n} = \frac{\sum wx}{\sum w}$$

Median

MEDIAN is the midpoint of the values after they have been ordered from the smallest to the largest, or the largest to the smallest.

PROPERTIES OF THE MEDIAN

1. There is a unique median for each data set.
2. It is not affected by extremely large or small values and is therefore a valuable measure of central tendency when such values occur.
3. It can be computed for ratio-level, interval-level, and ordinal-level data.
4. It can be computed for an open-ended frequency distribution if the median does not lie in an open-ended class.

EXAMPLES:

The ages for a sample of five college students are: **21, 25, 19, 20, 22**

Arranging the data in ascending order gives: **19, 20, 21, 22, 25**

Thus the median is 21.

The heights of four basketball players, in inches, are: **76, 73, 80, 75**

Arranging the data in ascending order gives: **73, 75, 76, 80**

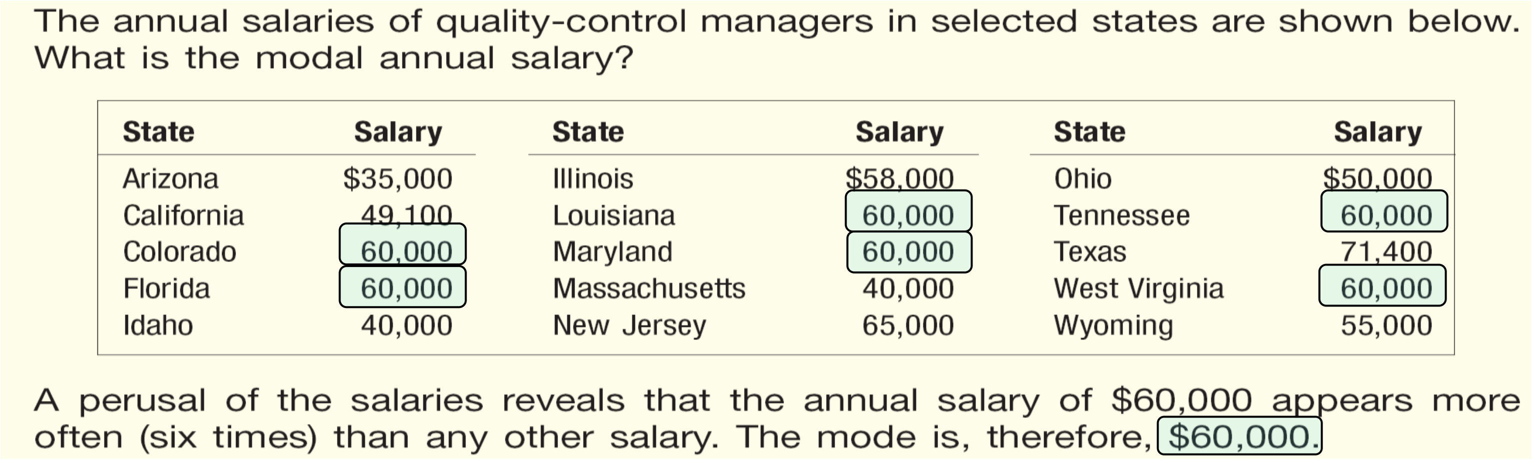
Thus the median is 75.5

Mode

MODE is the value of the observation that appears most frequently.

- The *mode* value is the value that occurs most often in a data set.
- A data set with one value that occurs with greatest frequency is said to be *unimodal* ,*e.g., (3,2,1,2,4,5,6)* .
- A data set with two values that occur with greatest frequency is said to be *bimodal* , *e.g., (3,2,1,2,4,5,1).*
- A data set with more than two values that occur with greatest frequency is said to be *multimodal* , *e.g., (4,5,3,3,2,1,2,6,1)*
- When all the values in a data set occur with the same frequency is said to have *no mode* , *e.g., (3,2,3,2,1,5,5,1)*

EXAMPLE:



The annual salaries of quality-control managers in selected states are shown below. What is the modal annual salary?

State	Salary	State	Salary	State	Salary
Arizona	$35,000	Illinois	$58,000	Ohio	$50,000
California	49,100	Louisiana	60,000	Tennessee	60,000
Colorado	60,000	Maryland	60,000	Texas	71,400
Florida	60,000	Massachusetts	40,000	West Virginia	60,000
Idaho	40,000	New Jersey	65,000	Wyoming	55,000

A perusal of the salaries reveals that the annual salary of $60,000 appears more often (six times) than any other salary. The mode is, therefore, $60,000.

Percentiles

The *pth percentiles* is a value such that at least *p percent* of the observations are less than or equal to this value.

- *A percentile* provides information about how the data are spread over the interval from the smallest value to the largest value.
- Finding a data value corresponding to a given percentile

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

In Excel			
Select Data	HOME	Sort & Filter	Sort Smallest to Largest

2- Substitute into the formula

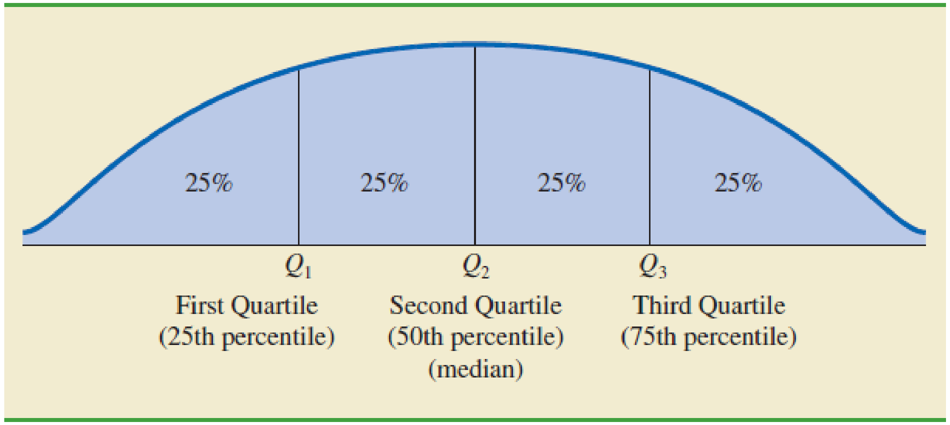
Formula	*In Excel*
$c = \frac{n * p}{100}$	$= \boldsymbol{n * p/100}$

3- If *c* is not an integer, round up. The next integer *greater* than *c* denotes the position of the *p*th percentile.

4- if *c* is an integer, use the mid-value between the *cth* and *(c+1)st* values when counting up from lowest value.

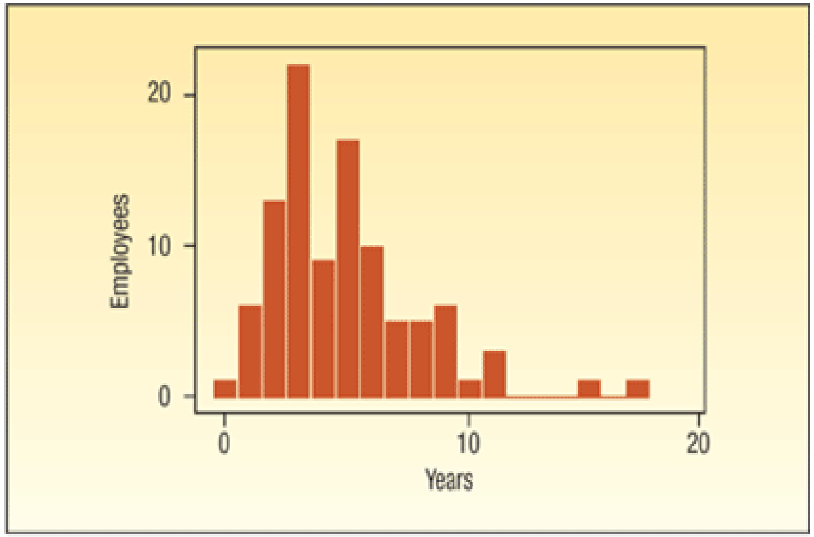
Quartiles

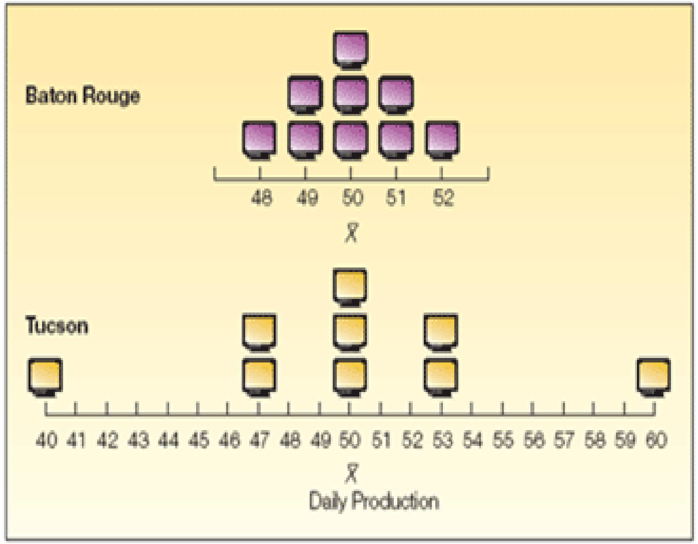
- *Quartiles* divide the distribution into four groups, denoted by Q_1, Q_2, Q_3. Note that Q_1 is the same as the 25th percentile; Q_2 is the same as the 50th percentile or the median; and Q_3 corresponds to the 75th percentile.
- An *outlier* is an extremely high or an extremely low data value when compared with the rest of the data values.
- *Outliers* can be identified using the *interquartile range (IQR)* which is a *measure of variation* that can be used when the data contains outlier values. IQR = Q_3 - Q_1
- Outliers can be the result of measurement or observational error.



Measures of Variability

A small value of variation of dispersion indicates that the data are clustered closely around the mean. The mean is therefore considered representative of the data. Conversely, a large variation indicates that the mean is not reliable.





The simplest measure of variability is the *range* which is a measure of the largest variability in a data set.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

Variance

VARIANCE is the arithmetic mean of the squared deviations from the mean.

POPULATION VARIANCE

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

σ² is the population variance (σ is the lowercase Greek letter sigma). It is read as "sigma squared."

X is the value of an observation in the population.

μ is the arithmetic mean of the population.

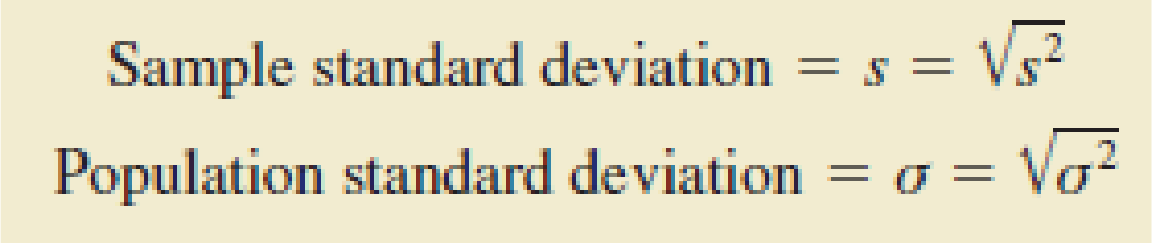
N is the number of observations in the population.

SAMPLE VARIANCE

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

Standard Deviation

STANDARD DEVIATION is the positive square root of the variance.



$$\text{Sample standard deviation} = s = \sqrt{s^2}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

- The variance and standard deviations are nonnegative and are zero only if all observations are the same.
- For populations whose values are *near the mean*, the variance and standard deviation will be small.
- The variance overcomes the weakness of the range by using all the values in the population

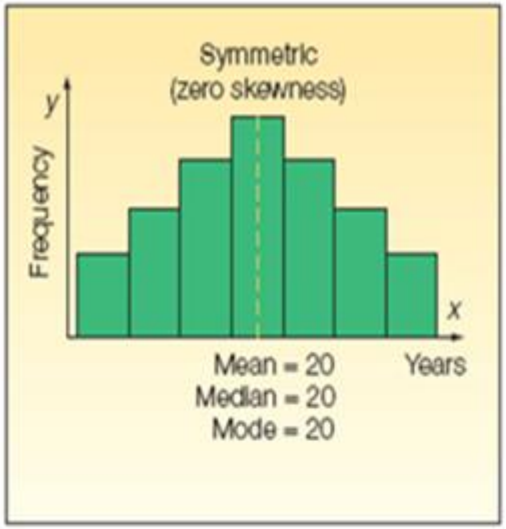
Coefficient of Variation

- □ The *coefficient of variation* is a measure of the dispersion of data values around the mean value of the data. It is calculated by dividing the standard deviation by the mean expressed as a percentage.

$$CV = \frac{\sigma}{\mu} * 100$$

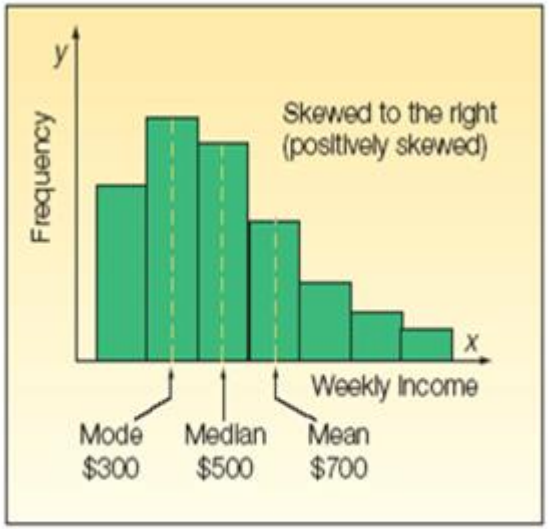
- □ The *coefficient of variation* is mostly used to compare standard deviations of two variables or more when the units or the values of the means are different.
- □ Large coefficient of variation means large variability.

Relative Positions of the Mean, Median and Mode



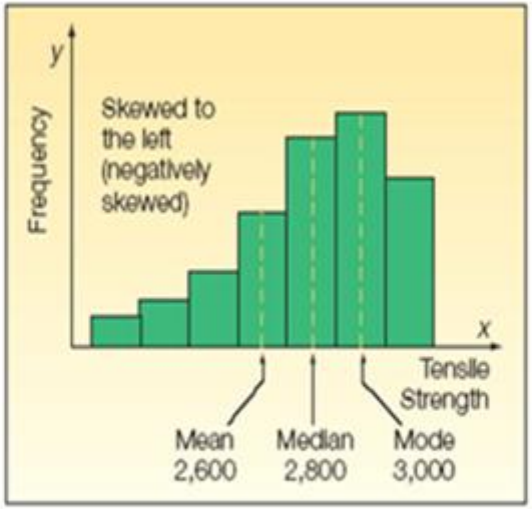
zero skewness

mode = median = mean



positive skewness

mode < median < mean



negative skewness

mode > median > mean

An important numerical measure of the shape of a distribution is called *skewness*

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- The distribution is skewed to the left if the skewness value is negative.
- The distribution is skewed to the right if the skewness value is positive.
- The distribution is symmetric if the skewness value is zero

Z-Score

- A *standard score* or *z score* is used when direct comparison of raw scores is impossible.
- The z score represents the number of standard deviations a data value falls above or below the mean.

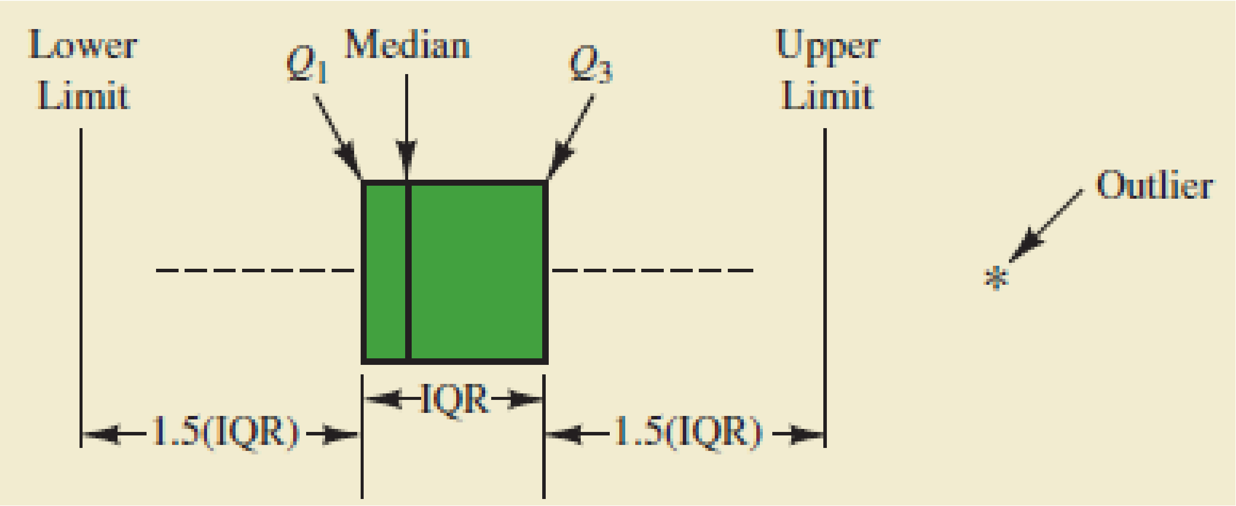
Population z-score	*Sample z-score*	In Excel
$z = \frac{x - \mu}{\sigma}$	$z = \frac{x - \bar{x}}{s}$	$= (x - \bar{x})/s$

- Positive z value means that the value is above the mean and negative z value mean that the value is below the mean

Five-number Summary and Boxplots

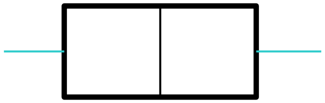
Exploratory data analysis includes the *box plot* and the *five-number summary*.

- □ *Boxplots* are graphical representations of a *five-number summary* of a data set.
- □ The five specific values that make up a *five-number summary* are minimum, Q_1, Q_2, Q_3 and maximum.

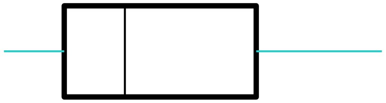


Skewness and Boxplots

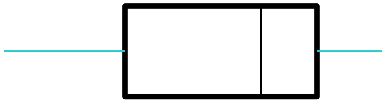
- If the median is near the center of the box, the distribution is approximately symmetric.



- 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.



Correlation

- Often a manager or decision maker is interested in the *relationship between two or more variables*
- Inferential statistics involves determining whether a relationship between two or more numerical variables exists.
- *Correlation* is a statistical method used to determine whether a relationship between two variables, that is not affected by the units of measurements, exists.

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT: SAMPLE DATA

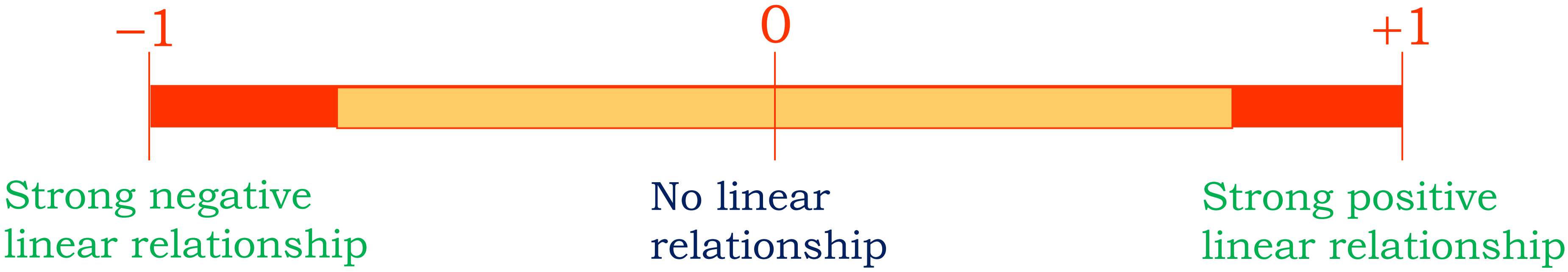
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

SAMPLE COVARIANCE

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Correlation

- The *correlation coefficient* is a 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.


–1
0
+1
Strong negative linear relationship
No linear relationship
Strong positive linear relationship

Applying the concepts

- The following data represent salaries from a school district

  10,000 11,000 11,000 12,500 14,300 17,500 18,200 14,700
  18,000 16,600 19,200 21,100 15,400 50,000 15,700 15,200

- If you work for the school board and do not wish to increase salaries. Compute the measures of central tendency and decide which one would best support your position.
- If you work for the teachers' union and want a raise for the teachers. Use the best measure of central tendency to support your position.
- Explain how outliers can be used to support one or the other position.
- If the salaries represented every teacher in the school district, would the averages be parameters or statistics?
- Which measure of central tendency can be misleading when a data set contains outliers?
- When you are comparing the measures of central tendency, does the distribution display any skewness? Explain.

Applying the concepts

- The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

- If you work for the school board and do not wish to show a large variation in salaries. Compute the measures of variation and decide which one would best support your position.
- If you work for the teachers’ union and want to show a large variation in salaries. Use the best measure of variation to support your position.
- Which measure of variation can be misleading when a data set contains outliers?
- From the coefficient of skewness, does the distribution display any skewness? Explain.
- Find the z score for a teacher’s salary of 14000 and for a teacher’s salary 18000. Explain.
- What value corresponds to the 30th percentile?
- What value corresponds to the 50th percentile?

Applying the concepts

- The following data represent salaries from a school district

10,000	11,000	11,000	12,500	14,300	17,500	18,200	14,700
18,000	16,600	19,200	21,100	15,400	50,000	15,700	15,200

- Calculate the values of Q_1, Q_2 and Q_3 and decide in which quartile a teacher's salary of 17000 falls.
- How many observations falls between the minimum and the median?
- Is the data containing any outlier values?
- From the boxplot, comment on the skewness of the distribution.
- If you take the first eight salaries as one group and the rest as another group, which group is more variable?
- Taking the first eight salaries as group one and the second eight salaries as group two, calculate the correlation coefficient and interpret the result.

Summary

Sample statistic A numerical value used as a summary measure for a sample.

Population parameter A numerical value used as a summary measure for a population.

Point estimator The sample statistic when used to estimate the corresponding population parameter.

Mean A measure of central location computed by summing the data values and dividing by the number of observations.

Weighted mean The mean obtained by assigning each observation a weight that reflects its importance.

Median A measure of central location provided by the value in the middle when the data are arranged in ascending order.

Mode A measure of location, defined as the value that occurs with greatest frequency.

Percentile A value such that at least p percent of the observations are less than or equal to this value and at least (100p) percent of the observations are greater than or equal to this value. The 50th percentile is the median.

Summary

Quartiles The 25th, 50th, and 75th percentiles, referred to as the first quartile, the second quartile (median), and third quartile, respectively. The quartiles can be used to divide a data set into four parts, with each part containing approximately 25% of the data.

Interquartile range (IQR) A measure of variability, defined to be the difference between the third and first quartiles.

Outlier An unusually small or unusually large data value.

Range A measure of variability, defined to be the largest value minus the smallest value.

Variance A measure of variability based on the squared deviations of the data values about the mean.

Standard deviation A measure of variability computed by taking the positive square root of the variance.

Coefficient of variation A measure of relative variability computed by dividing the standard deviation by the mean and multiplying by 100.

Skewness A measure of the shape of a data distribution. Data skewed to the left result in negative skewness; a symmetric data distribution results in zero skewness; and data skewed to the right result in positive skewness.

Summary

z-score A value computed by dividing the deviation about the mean (xi) by the standard deviation s. A z-score is referred to as a standardized value and denotes the number of standard deviations xi is from the mean.

Five-number summary An exploratory data analysis technique that uses five numbers to summarize the data: smallest value, first quartile, median, third quartile, and largest value.

Box plot A graphical summary of data based on a five-number summary.

Covariance A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.

Correlation coefficient A measure of linear association between two variables that takes on values between 1 and 1. Values near 1 indicate a strong positive linear relationship; values near 1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.