ISIMS - UNIVERSITY OF SFAX
LSI ADBD - BACHELOR 2
2023/2024

***

DESCRIPTIVE AND INFERENTIAL STATISTICS

***

LECTURE II

CENTRAL TENDENCY, VARIATION, AND SHAPE

Lecture Content

1. Central Tendency
   - The Mean
   - The Median
   - The Mode
2. Variation
   - The Range
   - The Variance and the Standard Deviation
   - The Coefficient of Variation
3. The Shape
   - The Skewness
   - The Kurtosis
4. Quartile Analysis and Box Plot
5. Population Mean and Variance
6. Covariance and Coefficient of Correlation

CENTRAL TENDENCY

Central Tendency - The Mean

Central Tendency is the extent to which the values of a numerical variable group around a typical, or central, value.

The **arithmetic mean** (in everyday usage, the **mean**) is the most common measure of central tendency. It is a typical or central value and serves as a "balance point" in a set of data. Using a sample of data, we will refer to the arithmetic mean as the **sample mean**.
Definition: If the n observations in a sample are denoted by $X_1, X_2, \ldots, X_n$, the sample mean is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Example 1: The following data give the time to get ready in the morning from when you get out of bed to when you leave your home, collected for 10 consecutive workdays.

Day:	1	2	3	4	5	6	7	8	9	10
Time (minutes):	39	29	43	52	39	44	40	31	44	35

Central Tendency - The Median

The **median** is the middle value in an array of data that has been ranked from the smallest to the largest. Half the values are smaller than or equal to the median, and half the values are larger than or equal to the median.

You compute the median by following one of the two rules:

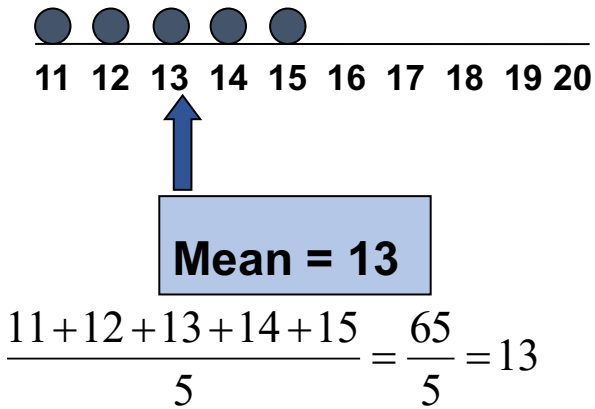
- **Rule 1:** If the data set contains an *odd* number of values, the median is the measurement associated with the middle-ranked value.
- **Rule 2:** If the data set contains an *even* number of values, the median is the measurement associated with the average of the two middle-ranked values.

Example 2: The median time to get ready the morning is 39.5.

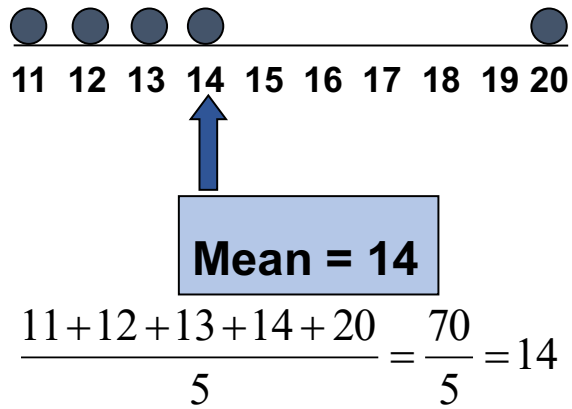
Ranked values:	1	2	3	4	5	6	7	8	9	10
Time (minutes):	29	31	35	39	39	40	43	44	44	52

Central Tendency - Mean vs Median

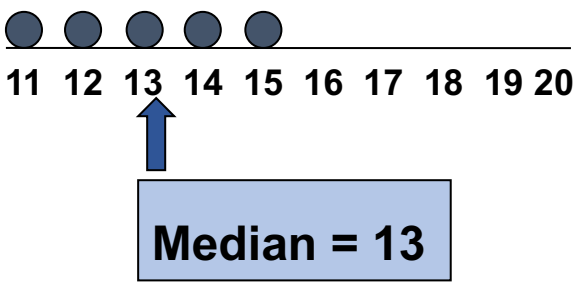
The median is less sensitive than the mean to the extreme values (outliers).

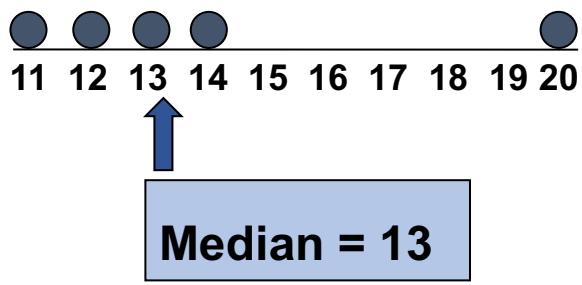


$$\frac{11+12+13+14+15}{5} = \frac{65}{5} = 13$$



$$\frac{11+12+13+14+20}{5} = \frac{70}{5} = 14$$





Central Tendency - The Mode

The **Mode** is the value that appears most frequently. For a particular variable, there can be several modes or no mode at all.

Example 3: Rank daily times to get ready the morning. There are two modes 39 and 44.

Exercise 1: The female students in an undergraduate business core course self-reported their heights to the nearest inch. Data are in Excel Sheet **DataEx1**.

1. Construct a stem-and-leaf diagram for the height and comment on any important features that you notice.
2. Calculate the sample mean, median and mode.

Central Tendency - The Geometric Mean

Definition: The **geometric mean** is the nth root of the product of n values:

$$\overline{X}_G = (X_1 \times X_2 \times ... \times X_n)^{1/n}$$

The geometric mean measures the rate of change of a variable over time.

Particularly, the **geometric mean rate of return** measures the mean percentage return of an investment per period of time.

Definition: If R_i is the rate of return un time period i, the geometric mean rate of return is:

$$\overline{R}_G = [(1 + R_1) \times (1 + R_2) \times ... \times (1 + R_n)]^{1/n} - 1$$

Central Tendency - Why the geometric mean ?

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$$X_1 = \$100{,}000 \quad X_2 = \$50{,}000 \quad X_3 = \$100{,}000$$

50% decrease 100% increase

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

- The arithmetic mean gives a **Misleading result**:

$$\overline{R} = \frac{(-0.5) + 1}{2} = 0.25 = 25\%$$

- The geometric mean is **more adequate** in **finane**:

$$\overline{R}_G = [(1 + (-0.5)) \times (1 + 1)]^{1/2} - 1 = [0.5 \times 2]^{1/2} - 1 = 0$$

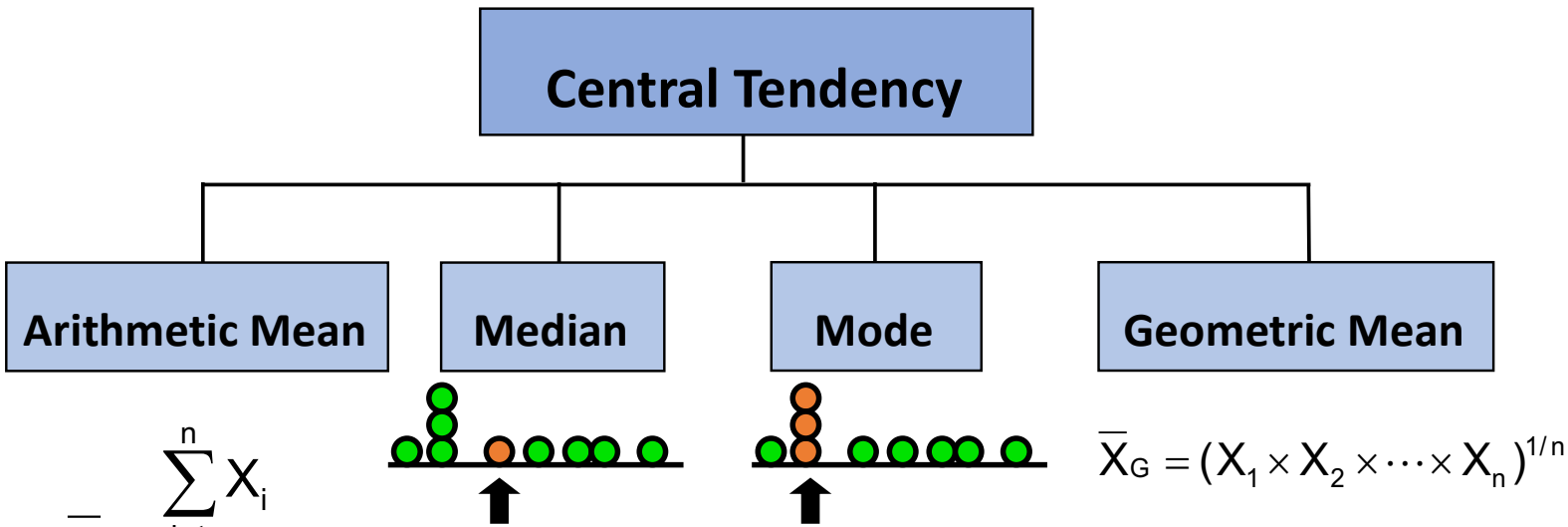
Central Tendency - The Geometric Mean

Exercise 2: Table below gives the total rate of return percentage for the Dow Jones Industrial Average (DJIA), the Standard Poor's 500 (SP 500) and the technology-heavy NASDAQ composite (NASDAQ) from 2013 through 2016.

Year	DJIA	S&P 500	NASDAQ
2013	26.5	29.6	28.3
2014	7.5	11.4	13.4
2015	-2.2	-0.7	5.7
2016	13.4	9.5	7.5

1. Compute the geometric mean rate of return per year for the DJIA, SP 500, and NASDAQ from 2013 through 2016.
2. What conclusions can you reach concerning the geometric mean rate of return of these three market indices ?

Measures of Central Tendency - Summary



VARIATION AND SHAPE

Measure of Variation - The Range

Variation measure the **spread**, or **dispersion**, of the values.

The **Range** is the difference between the largest and the smallest value and is the simplest descriptive measure of variation for a numerical variable.

Example 4: In the time-to-get-ready example, the range is $52 - 29 = 23$.

The range measures the total spread. It does not give any idea about how the values are distributed between the smallest and largest values. It does not indicate whether the values are evenly distributed, clustered near the middle, or clustered near one or both extremes.

Measure of Variation - Variance and Standard Deviation

The **variance** and **standard deviation** are the commonly used measures that account for how values fluctuate above and below the sample mean.
In the case of a sample of observations on a numerical variable, we talk about the **sample variance** S^2, and the **sample standard deviation** S.

Definition: If we denote by n the sample size, $\overline{X}$ the sample mean, and X_i the ith value of the variable X, the sample variance is given by,

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Measure of Variation - Variance and Standard Deviation

Definition: The sample standard deviation is:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Notice that:

- Because the sum of squares can never be a negative value, the variance and standard deviation will always be non-negative.
- The variance and standard deviation will be zero, meaning no variation, only for the special case in which every value in a sample is the same value.

Example 5: Compute the sample standard deviation of the time-to-get-ready data.

39 29 43 52 39 44 40 31 44 35

Measure of Variation - The Coefficient of Variation

The **coefficient of variation** measures the scatter in the data relative to the mean. It is a relative measure of variation that is always expressed as a percentage.

Definition: The coefficient of variation is equal to the standard deviation divided by the mean, multiplied by 100%.

$$CV = (\frac{S}{\overline{X}})100\%$$

Example 5: For the sample of 10 get ready times, $CV = 17.10\%$. In these data the standard deviation is 17.1% of the size of the mean.

The coefficient of variation can be used to compare the variability of two or more sets of data measured in different units.

Measure of Variation - Comparing Coefficients of Variations

- **Stock A :** Mean price last year = \$50 ; Standard deviation = \$5

$$CV_A = \frac{S}{\overline{X}}.100\% = \frac{5\$}{50\$}.100\% = 10\%$$

- **Stock B :** Mean price last year = \$100 ; Standard deviation = \$5

$$CV_B = \frac{S}{\overline{X}}.100\% = \frac{5\$}{100\$}.100\% = 5\%$$

- **Stock C :** Mean price last year = \$8 ; Standard deviation = \$2

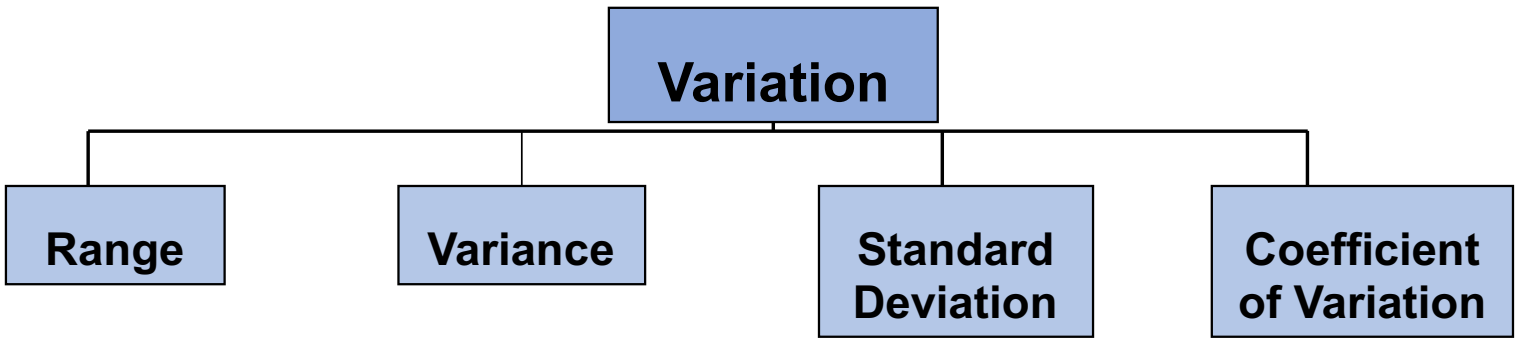
$$CV_C = \frac{S}{\overline{X}}.100\% = \frac{2\$}{8\$}.100\% = 25\%$$

***

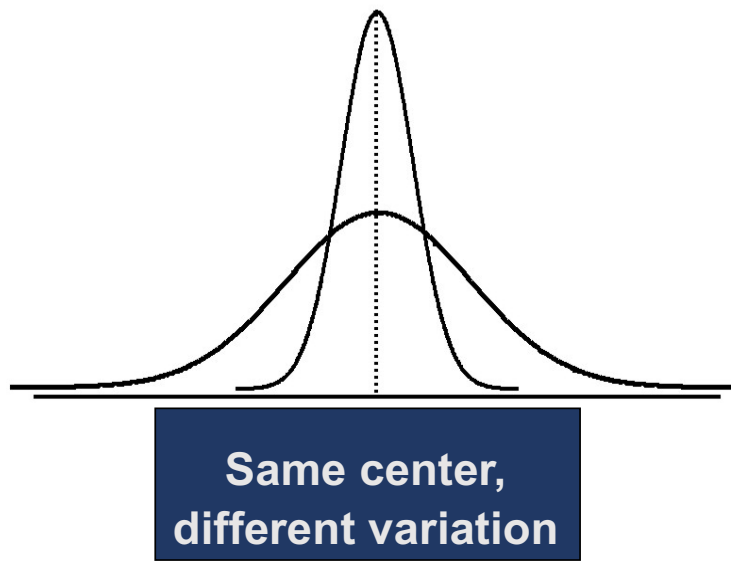
Stocks A and B have the same standard deviation, but stock B is less variable relative to its mean price than stock A.

Stock C has a much smaller standard deviation than A but a much higher coefficient of variation.

Measures of Variation - Summary



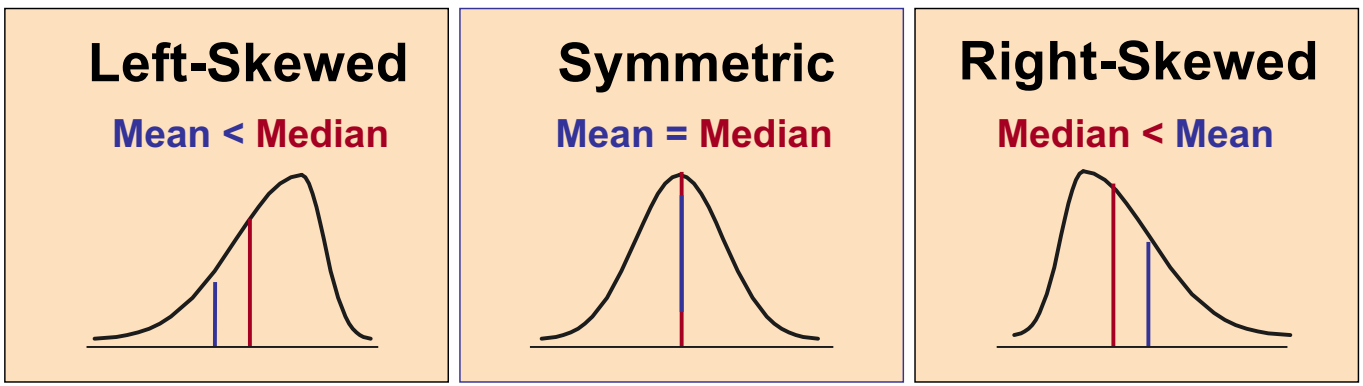
Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.



Measure of Shape - The Skewness

Skewness measures the extent to which the data values are not symmetrical around the mean. The three possibilities are:

- Mean < Median : **left-skewed distribution**; Skewness < 0.
- Mean = Median : **Symmetrical distribution**; skewness = 0.
- Mean > Median : **right-skewed distribution**; skewness > 0.



In symmetrical distributions, the values below the mean are distributed exactly the same way the values above the mean. In skewed distributions there is an imbalance of data values below and above the mean.

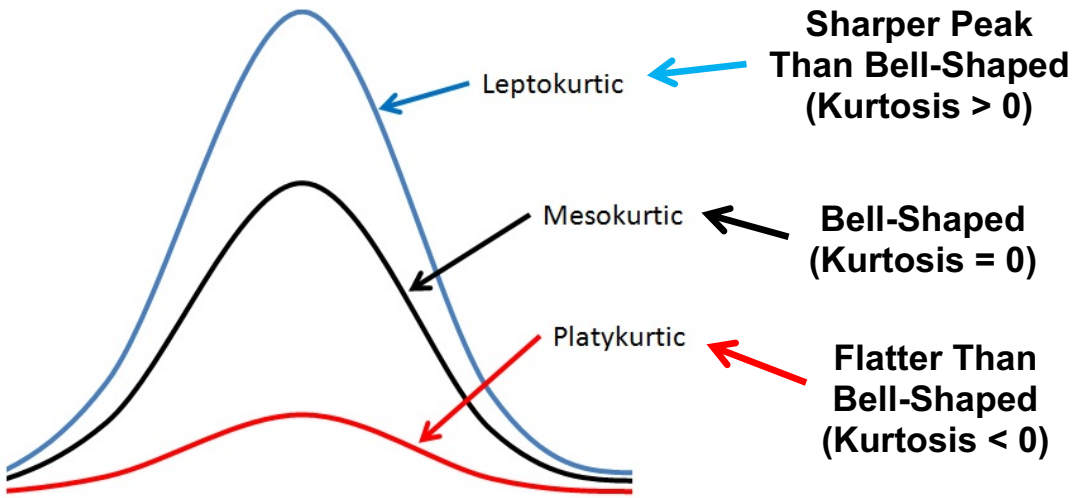
Measure of Shape - The Kurtosis

Kurtosis measures the peakedness of the curve of the distribution. It gives how the curve rises approaching the center of the distribution.

Kurtosis compares the shape of the peak to the shape of the peak of a bell-shaped distribution, which by definition has a kurtosis of zero.

- Kurtosis<0 : **platykurtic**; The center peak is slower-rising (or flatter) than the normal distribution peak. A platykurtic has a lower concentration of values near the mean compared to the normal distribution.
- Kurtosis>0 : **leptokurtic**; The center peak is sharper-rising than the normal distribution peak. A leptokurtic distribution has a higher concentration of values near the mean.

Measure of Shape - The Kurtosis



Notice that: In affecting the shape of the central peak, the relative concentration of values near the mean also affects the ends, of ***tails*** of the curve of a distribution. A leptokurtic distribution has fatter tails, i.e. many more values in the tails, than a normal distribution has.

Central Tendency, Variation, and Shape - Exercises

Exercise 3: Excel sheet **DataEx3** contains the overall download and upload speeds in mbps for nine carriers in the United States. For the download speed and upload speed separately:

1. Compute the mean and the median
2. Compute the variance, standard deviation, range and coefficient of variation.
3. Are the data skewed? If so, how ?
4. Based on the results of (1) through (3), what conclusions can you reach concerning the download and upload speed of various carriers?

Central Tendency, Variation, and Shape - Exercises

Exercise 4: The following data give the average room price (in US$) paid by various nationalities while traveling abroad in 2016

124 101 115 126 114 112 138 85 138 96 130 116 132.

1. Compute the mean, median, and mode.
2. Compute the range, variance, and standard deviation.
3. What conclusions can you reach concerning the room price paid by international travelers while traveling to various countries in 2016 ?
4. Suppose that the last value was 175 instead of 132. Repeat (1) through (3), using this value. Comment on the difference in the results.

QUARTILES AND BOX PLOTS

Quartiles of Numerical Variables

Quartiles split the values into four parts:

- The **first quartile** Q_1 divides the smallest 25% of the values from the other 75% that are larger.

$$Q_1 = \frac{n+1}{4} \quad \text{ranked values}$$

- The **second quartile** Q_2 is the median. 50% of the values are smaller than or equal to the median, and 50% are larger than or equal to the median

$$Q_2 = \frac{n+1}{2} \quad \text{ranked values}$$

- The **third quartile** Q_3 divides the smallest 75% of the values from the other 25% that are larger.

$$Q_3 = \frac{3(n+1)}{4} \quad \text{ranked values}$$

Quartiles of Numerical Variables - Rules of Calculation

- **Rule 1** If the ranked value is a whole number, the quartile is equal to the measurement that corresponds to that ranked value. Example n=7.
- **Rule 2** If the ranked value is a fractional half (2.5, 4.5, etc.) the quartile is equal to the measurement that corresponds to the average of the measurements corresponding to the two ranked values involved. Example: n=9.
- **Rule 3** If the ranked value is neither a whole number nor a fractional half, round the result to the nearest integer, and select the measurement corresponding to the ranked value. Example n=10.

Example 5: Excel sheet **DataMob** contains the mobile commerce penetration values, the percentage of the country population that bought something online via a mobile phone in the past month, for twenty-eight of the world's economies. Compute the quartiles of these data.

Quartiles of Numerical Variables - Other Statistics

1. **The Interquartile Range IQR:** Also called the **midspread**. It measures the difference in the center if a distribution between the third and first quartile.

$$\text{Interquartile range} = Q_3 - Q_1$$

   The IQR is called midspread because it covers the middle 50% of the data. It measures variability that is not influenced by outliers or extreme values.

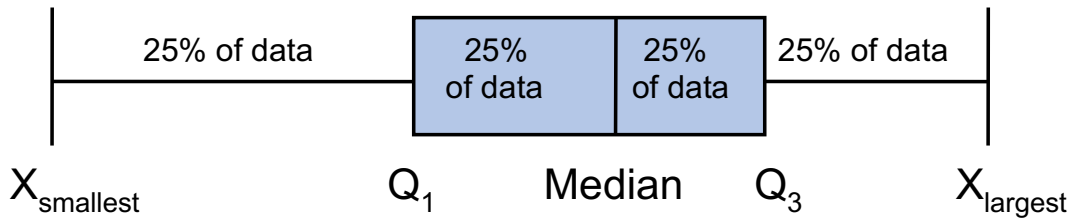
2. **The Five-Number Summary:** The five-number summary are:

$$X_{smallest} \qquad Q_1 \qquad Median \qquad Q_3 \qquad X_{largest}$$

3. **The Percentiles:** They are Related to quartiles. **Percentiles** split the variable into 100 equal parts. By this definition the first quartile is equivalent to the 25th percentile, the second quartile to the 50th percentile, and the third quartile to the 75th percentile.

The Boxplot

The **boxplot**, sometimes called *box-and-whisker plots*, is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry and identification of unusual observations or outliers.

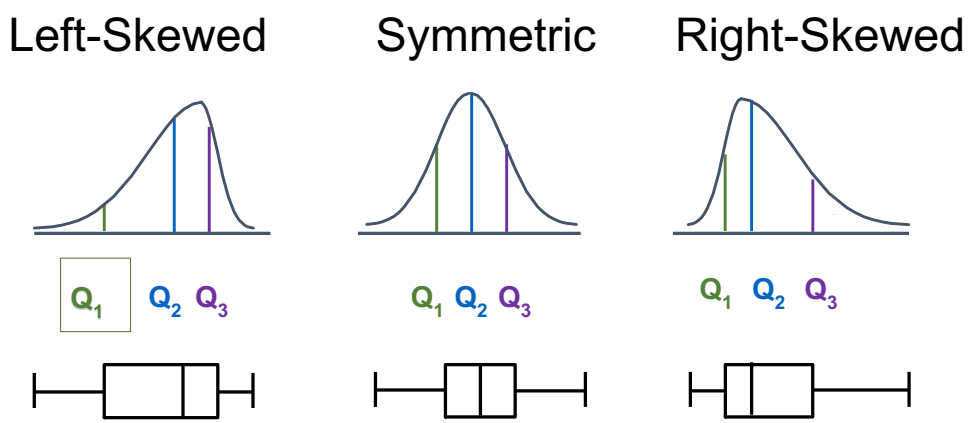


- It displays the three quartiles, the minimum, and the maximum of the data on a rectangular box.
- The box encloses the interquartile range with the left (lower) edge at the first quartile Q_1, and the right (upper) edge at the third quartile Q_3.
- A line is drawn through the box at the second quartile.
- A lower **whisker** line extends from Q_1 to $X_{smallest}$. And an upper whisker extends from the Q_3 to $X_{largest}$.

The Boxplot

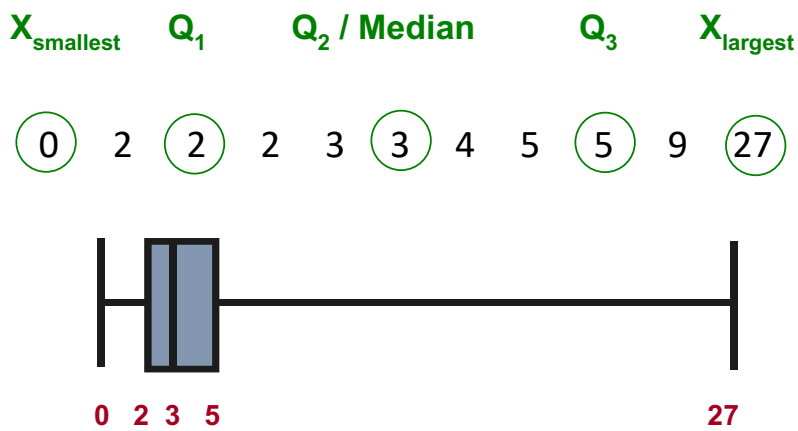
The figure below demonstrates the relationship between the box plot and the density curve for three different types of distributions.

- In the symmetric distribution the mean and the median are equal. the length of the right tail is equal to the length of the left tail.
- In the left-skewed distribution there is a heavy clustering of values at the high end of the scale. There is a long left tail that contains the smallest 25% of the values.
- In the right-skewed distribution the concentration of values is on the low end of the scale. There is a long right tail that contains the largest 25% of the values.



The Boxplot

Example 7: Below is a Boxplot for the following data. The data are right skewed, as the plot depicts.



Example 8: Consider the compressive strength data of 80 aluminum specimens given in data sheet $\text{Data}_{Example8.Describethesedatausingaboxplot}$.

Example 9: Consider the **DataMob** dataset of example 5. List the five-number summary, construct a boxplot, and describe its shape.

The Boxplot - Exercises

Exercise 5: The following data give the average room price (in US$) paid by various nationalities while traveling abroad in 2016

124 101 115 126 114 112 138 85 138 96 130 116.

1. Compute the first quartile Q_1, and the third quartile Q_3, and the interquartile range.
2. List the five-number summary.
3. Construct a boxplot and describe its shape.

NUMERICAL DESCRIPTIVE VALUES OF A POPULATION

The Population Mean

The **population mean** is the main measure of central tendency in a population. The Greek lowercase mu, μ, represents, this parameter which is defined by the following equation.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

where,

μ = population mean

X_i = ith value of the variable X

$\sum_{i=1}^{N} X_i$ = summation of all X_i values in the population

N = number of values in the population

The Population variance and Standard Deviation

The **population variance** and the **population standard deviation** parameters measure variation in the population.

Population Variance: Represented by the Greek letter σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

where,

$\sum_{i=1}^{N}(X_i - \mu)^2$ = sum of all the squared differences between X_i and μ.

Population Standard Deviation: Represented by σ

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

The empirical Rule

In Symmetrical datasets, where the median and mean are the same, the values often tend to cluster around the median and the mean, often producing a bell-shaped normal distribution.

The **empirical rule** states that for population data from a symmetrical mound-shaped distribution such as the normal distribution, the following are true:

- Approximately 68% of the values are within ± 1 standard deviation from the mean.
- Approximately 95% of the values are within ± 2 standard deviation from the mean.
- Approximately 99.7% of the values are within ± 3 standard deviation from the mean.

The empirical Rule

The empirical rule helps you examine variability in a population as well as identify outliers. As a general rule, you can consider values not found in the interval $\mu \pm 2\sigma$ as potential outliers. The rule also implies that only about 3 in 1,000 will be beyond standard deviations from the mean.

Example 10: A population of 2-liter bottles of cola is known to have a mean fill-weight of 2.06 liters and a standard deviation of 0.02 liter. The population is known to be bell-shaped. Describe the distribution of fill-weights. Is it very likely that a bottle will contain less than 2 liters of cola?

RELATIONSHIP BETWEEN TWO NUMERICAL VARIABLES

The Covariance

The **covariance** measures the strength of the linear relationship between two numerical variables X and Y.

The Sample Covariance:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- $cov(X,Y) > 0$: X and Y move in the ***same*** direction.
- $cov(X,Y) < 0$: X and Y move in the ***opposite*** direction.
- $cov(X,Y) = 0$: X and Y are ***independent***.

The covariance gives only the sign of the relationship between X and Y. It is not possible to determine the relative strength of this relationship from the size of the covariance.

The Coefficient of Correlation

The **coefficient of correlation** measures the strength of the linear relationship between two numerical variables X and Y.

Sample Coefficient of Correlation:

$$r = \frac{cov(X,Y)}{S_X S_Y}$$

where,

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

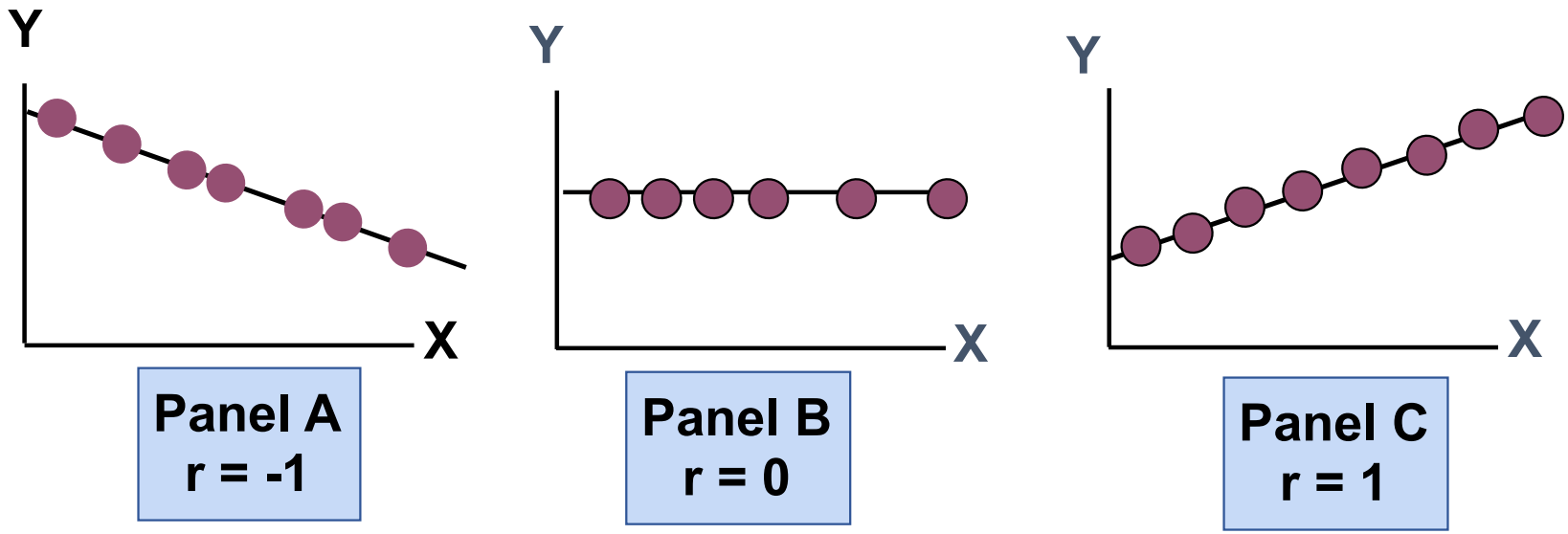
$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}, \quad \text{and} \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

The Coefficient of Correlation - Main Features

- The population coefficient of correlation is referred as ρ.
- The sample coefficient of correlation is referred to as r.
- Either ρ or r have the following features:
  - Unit free.
  - Range between –1 for a perfect negative correlation and 1 for a perfect positive correlation.
  - The closer to –1, the stronger the negative linear relationship.
  - The closer to 1, the stronger the positive linear relationship.
  - The closer to 0, the weaker the linear relationship.

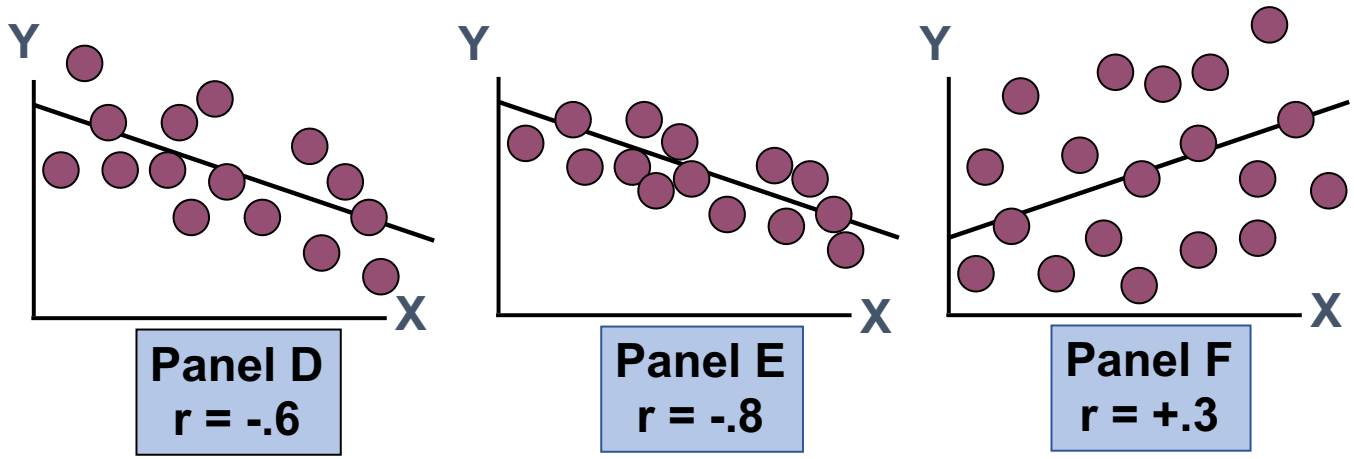
Perfect correlation means that if the points were plotted on a scatter plot, all the points could be connected with a straight line.

The Coefficient of Correlation - Scatter Plot



- In Panel A there is a perfect negative linear relationship between X and Y. When X increases, Y decreases in a perfectly predictable manner.
- Panel B shows a situation in which there is no relationship between X and Y. As X increases, there is no tendency for Y to increase or decrease.
- Panel C illustrates a perfect positive relationship. In this case, Y increases in a perfectly predictable manner when X increases.

The Coefficient of Correlation - Scatter Plot



- In Panel D, the small values of X tend to be paired with large values of Y. The linear relationship between X and Y in Panel D is not as strong as in Panel E.
- In Panel E, you can see that for small values of X, there is a very strong tendency for Y to be large, and inversely
- Panel F depicts data set that has a weak positive coefficient of correlation because small values of X tend to be paired with large values of Y. However dots are not very close to the line of perfect correlation.

The Coefficient of Correlation - To sum-up

In summary, the coefficient of correlation indicates the linear relationship, or association between two numerical variables. When the coefficient of correlation gets closer to $+1$ or -1, the linear relationship between the two variables is stronger. When the coefficient of correlation is near 0, little or no relationship exists.

The sign of the coefficient of correlation indicates whether the data are positively correlated (i.e., the larger values of X are typically paired with the larger values of Y) or negatively correlated (i.e., the larger values of X are typically paired with the smaller values of Y). The existence of a strong correlation however *does not imply a causation effect*. It only indicates the tendency present in the data.

Problems Set

Exercise 6: The following is a set of data from a sample 11 items:

X	8	6	9	4	7	11	13	5	10	16	19
Y	22	15	25	10	19	31	37	13	27	45	54

1. Calculate the covariance
2. Calculate the coefficient of correlation
3. Describe the relationship between the two variables X and Y.

Exercise 7: Consider the mobile speed data set in **DataEx3**.

1. Calculate the covariance between upload speed and download speeds.
2. Calculate the coefficient of correlation.
3. Based on 1. and 2., what conclusions can you reach about the relationship between download speed and upload speed?