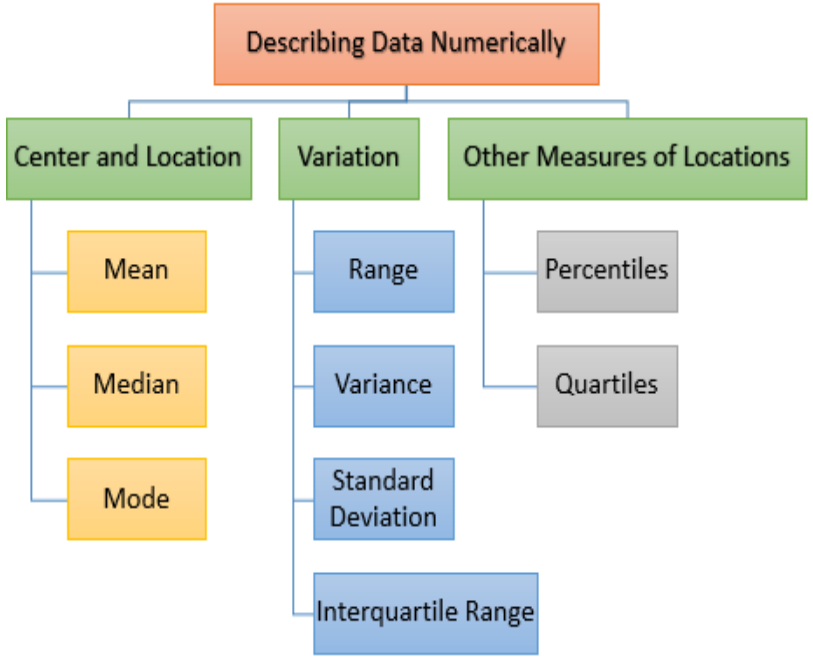
Introduction to Probability and Statistics

Topic (2): “Describing Data with Numerical Measures”



Dr. Heba Ayyoub

Philadelphia University

Topic (2): “Describing Data with Numerical Measures”

1 Describing a Set of Data with Numerical Measures

In this section, we will introduce the concept of summarization of the data by means of a single number called "a descriptive measure".

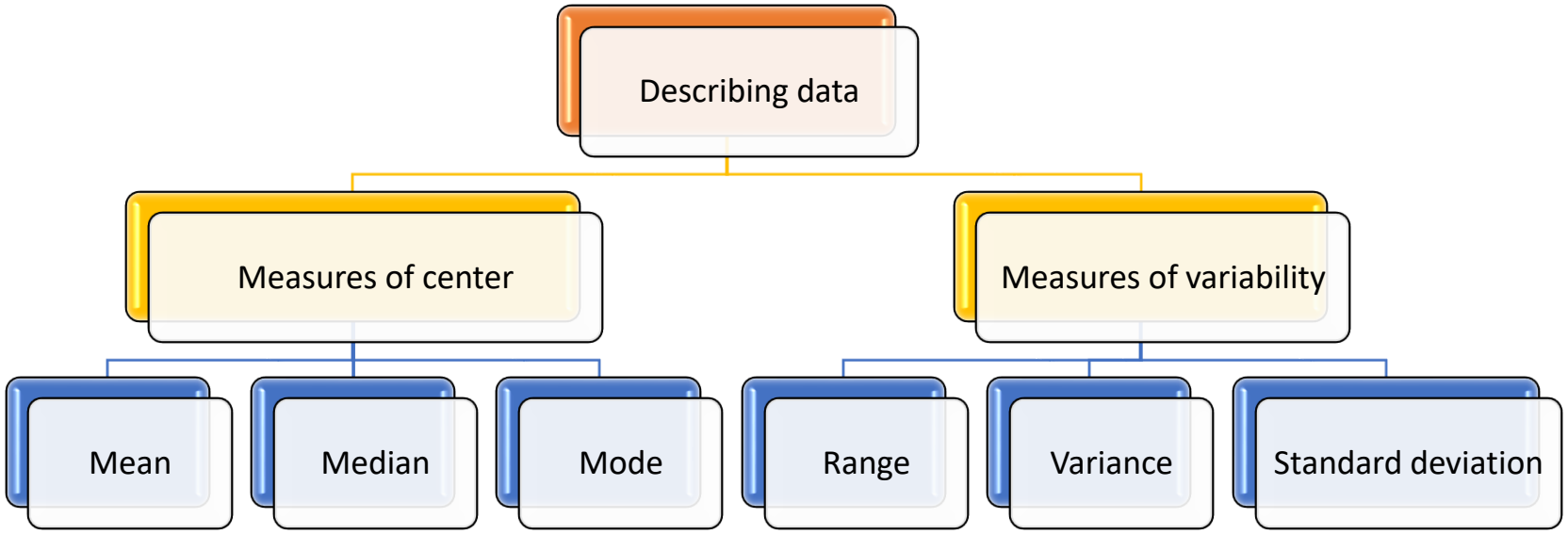
- A descriptive measure computed from the values of a population is called a "parameter".
- A descriptive measure computed from the values of a sample is called a "statistic".

(i) A **parameter** is a measure (or number) obtained from the population values: $X_1, X_2, X_3, \dots, X_N$

- Values of the parameters are unknown in general.
- We are interested to know true values of the parameters.

(ii) A **statistic** is a measure (or number) obtained from the sample values: $x_1, x_2, x_3, \dots, x_n$

- Values of statistics are known in general.
- Since parameters are unknown, statistics are used to approximate (estimate) parameters.



2 Measures of center

- **Mean:** The average value of the data set. It is calculated by summing all the data points and dividing by the number of data points.

  Let $x_1, x_2, \dots, x_n$ be a random sample then the sample mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  Let $X_1, X_2, \dots, X_N$ be observations in a population then the population mean is:

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

Example (1): Given the following data:

12, 2, −4, 3, 9, 5, 11, 12, 14, 0, −3. Calculate the sample mean?

Exercise (1): The mean of 10 numbers is 8. If an eleventh number is now added to the data, the mean becomes 9. What is the value of the new number?

Exercise (2): Given that $n = 10$, $\bar{x} = 12$ and $\sum_{i=1}^{9} x_i = 100$, then x_{10} is:

Exercise (3): A data set $5, c, 8, 2, 3, 7$, where c is unknown observation. Find the value of c if the sample mean is 5.

Exercise (4): The mean of n numbers is 5. If the number 12 is now removed from the n numbers, the mean is 4. Find the value of n.

Exercise (5): The mean of 4 numbers is 5, and the mean of 3 other numbers is 12. What is the mean of the 7 numbers together?

How to calculate the sample mean in case of **frequency table**?

Sample (x_i)	Frequency (f_i)
x_1	f_1
x_2	f_2
x_3	f_3
.	.
.	.
.	.
x_k	f_k

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$$

Example (2): Given the following table:

x_i	f_i	
3	2	
5	9	
8	15	
14	3	

Calculate the sample mean?

Exercise (6): For each of the following distributions, find the value of b for the given value of $\bar{x}$.

a) $\bar{x} = 1.05$

Value	Frequency
-2	5
0	2
1	**b**
3	9

b) $\bar{x} = 5.4$

Value	Frequency
2	5
b	2
8	2
10	1

How to calculate the sample mean in case of **group data**?

Interval	Frequency	x_i
$(A - B)$	f_1	$\frac{(A+B)}{2}$
$(C - D)$	f_2	$\frac{(C+D)}{2}$
$(E - F)$	f_3	$\frac{(E+F)}{2}$
. . .	. . .	. . .
$(K - L)$	f_k	$\frac{(K+L)}{2}$

→ Mid-point

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$$

Example (3): Given the following group table:

Interval	f_i		
(34.5 – 46.5)	3		
(47.5 – 59.5)	7		
(60.5 – 72.5)	10		
(73.5 – 85.5)	13		
(86.5 – 98.5)	7		

Calculate the sample mean?

Exercise (7): Find the mean of the following continuous distribution.

Class	Frequency
(1 – 5)	9
(6 – 10)	11
(11 – 15)	16
(16 – 20)	14

➢ Median (m):

Let $x_1, x_2, \dots, x_n$ be a random sample of n measurement ordered from smallest to the largest, then the median is:

1) If n is odd the median is the middle observation.
2) If n is even the median is the average of the two middle observations.

Note: Position of the median = $\frac{n+1}{2}$

Example (4): Given the following data:

13, 8, 2, 7, 9, 4, 10. Find the median and its position?

Example (5): Given the following data: 10, 15, 14, 2, 5, 8.

Find the median and its position?

Example (6): Given the following stem and leaf:

Stem	leaf								
2	1	2	2	3					
3	0	1	9	9					
4	2	3	4	5	7	7	8		
5	0	1	2	2	3	5	5	9	9

with Leaf unit = 0.1.

1) Describe the shape of the data?

2) Find the range?

3) Find the median?

4) Find the sample mean?

Example (7): In data (1) and (2) Calculate the mean and the median? What did you notice?

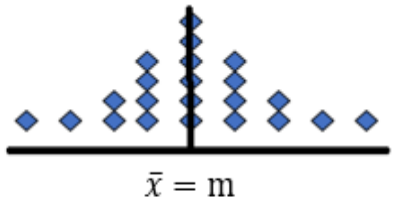
Data (1): 3, 4, 5, 9, 10

Data (2): 3, 4, 5, 9, 100

Example (8): Calculate the median for the following frequency table:

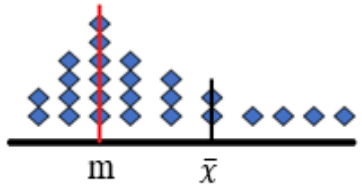
Value	Frequency
1	2
2	5
4	7
6	6

Exercise (8): If the median of the data 9, 1, x, 12 is 7, what is the value of x?



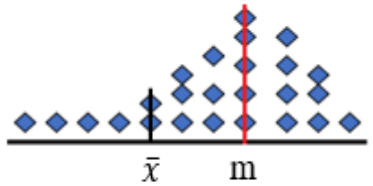
Symmetric: Mean = Median

$\bar{x} = m$



Skewed right: Mean > Median

$\bar{x} > m$



Skewed left: Mean < Median

$\bar{x} < m$

➢ **Mode:**

The value of x that occur most (or the category that occurs most frequently) .
A data set may have no mode, one mode, or multiple modes.

Example(9): for the following data sets. Find the mode(s).

1) 3, 5, 9, 5, 10, 11
2) 4, 3, 4 ,4, 4, 9, 3, 3, 3, 11, 12, 14, 12, 3, 4
3) 3, 4, 5, 9, 10

Example (10): Given the following table. Find the mode.

x	f
3	2
4	5
5	10
9	3
20	2

3 Measures of variability

1) Range: The difference between the highest and lowest values in the data set.

Range = max – min

Example (11): Given the following data: 2, 5, 10, 1, 4 and 3. Find the range.

2) Variance and Standard Deviation:

Measures the average squared deviation of each data point from the mean. It helps determine how spread out the data points are.

Deviations of sample values from the sample mean:

Let $x_1, x_2, x_3, \dots, x_n$ are the sample values, and $\bar{x}$ be the sample mean.

Note: The sum of deviation of observation from there mean equal zero. $[\sum(x_i - \bar{x}) = 0]$

Example (12): If the deviation of 5 observations from there mean are: 3, -2, 4, A, 3. Find the value of A?

The Population variance (σ^2): (Variance computed from the population)

Variance measures how far each number in the set from the mean.

Let $X_1, X_2, \dots, X_N$ be the population values. The Population variance (σ^2) is:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_N - \mu)^2}{N} \ (\text{unit})^2$$

where $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ is the population mean and N is the population size.

Note: σ^2 is a parameter because it is obtained from the population values (it is unknown in general).

$$\sigma^2 \geq 0$$

The Sample variance (S^2): (Variance computed from the sample)

Let $x_1, x_2, \dots, x_n$ be the sample values. The sample variance is:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} (\text{unit})^2$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean and n is the sample size.

Note: S^2 is a statistic because it is obtained from the sample values (it is known).

S^2 is used to approximate (estimate) σ^2

$$S^2 \geq 0$$

Note: If $S^2 = 0$, all the observation have the same value, there is no dispersion (no variation).

Sample variance (S^2):

Let $x_1, x_2, \dots, x_n$ be observations in a sample then the sample variance is:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \text{ where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Note that: $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$

Standard Deviation: The square root of the variance, providing a measure of the spread of data in the same units as the original data.

- The sample standard deviation (S) is $\sqrt{S^2}$.

Example (13): Given the following data: 1, 14, 15, 9, 4, 3

1) Calculate the sample variance and standard deviation.

2) Calculate the sample standard deviation using the second formula.

Example (14): If the deviation of five observation around their mean: $-2, 2, 4, 3, -7$. Calculate the sample variance.

Example (15): Given the following data.

9, 9, 9, 9, 9, 9, 9, 9, 9, 9.

Calculate the sample variance.

Note that if all the observation have the same value the sample variance (S^2) is always zero (= 0).

- The largest S^2 → the greater the variability.

Example (16): Given the following table:

Sample 1	Sample 2
$\bar{x} = 18$	$\bar{x} = 18$
$S^2 = 9$	$S^2 = 25$

Which sample is the best?

Example (17): Given the following information:

$\sum_{i=1}^{50} x_i = 300$ and the sample variance $S^2 = 3$. Calculate $\sum_{i=1}^{n} x_i^2$.

Exercise (9): You are given n = 10 measurements: 3, 5, 4, 6, 10, 5, 6, 9, 2, 8.

a. Calculate the sample mean .

b. Find median.

c. Find the mode.

Exercise (10): You are given n = 8 measurements: 4, 1, 3, 1, 3, 1, 2, 2.

a. Find the range.

b. Calculate $\bar{x}$.

c. Calculate S^2 and S.

Exercise (11): For a set of 10 numbers, $\sum_{i=1}^{10} x_i$ = 270 and $\sum_{i=1}^{10} x_i^2$ = 9054. Find the mean and the variance.

Exercise (12): For a set of 18 numbers, $\bar{x}$ = 23 and S = 12. Find $\sum_{i=1}^{18} x_i$ and $\sum_{i=1}^{18} x_i^2$.

Exercise (13): The numbers a, b, 6, 4, 7 have mean 5 and variance 4. Find the values of a, b.

Exercise (14): For a set of 20 numbers,$\sum_{i=1}^{20} x_i$ = 20 and $\sum_{i=1}^{20} x_i^2$= 96. For a second set of 30 numbers, $\sum_{i=1}^{30} x_i$ = 60 and $\sum_{i=1}^{30} x_i^2$ = 236. Find the mean and standard deviation for the combined set of 50 numbers.

The Variance of Discrete Frequency Distribution

$$S^2 = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})^2 . f_i\right)}{\sum_{i=1}^{n} f_i - 1} = \frac{\sum_{i=1}^{n}\left(x_i^2 . f_i\right) - \frac{\left(\sum_{i=1}^{n} x_i f_i\right)^2}{\sum_{i=1}^{n} f_i}}{\sum_{i=1}^{n} f_i - 1}$$

Example (18): : Find the variance of the following distribution.

Value (x_i)	Frequency (f_i)		
1	2		
2	9		
3	6		
4	8		

Exercise (15): Find the standard deviation of the following distribution.

Value (x_i)	Frequency (f_i)
3	4
4	5
5	8
6	3

4 On the Practical Significance of the Standard Deviation

❶ Tchebysheff's Theorem

Given the number *k* any real number greater than or equal 1 and a set of *n* measurement, at least $(1 - \frac{1}{k^2})$ of the measurement will lie within *k* standard deviation of their mean).

$$\bar{x} \pm kS$$

Example (19):
Find the proportion of data that will lie within 2 standard deviation of their mean.

Find the proportion of data that will lie within 3 standard deviation of their mean.

Find the proportion of data that will lie within 1 standard deviation of their mean.

Example (20): Find the proportion of the measurement that fall within 2.5 standard deviation of the mean.

Example (21): Given a set of 40 measurements that have mean 60 and variance 100.

1) Find the proportion of the measurement that lie in the interval [40, 80].

2) Find the number of measurements that lie in the interval [40, 80].

3) Find the proportion of the measurement that lie in the interval [30, 90].

Example (22): If the mean is 75 and the standard deviation is 10. Find the interval that at least $\frac{8}{9}$ of the measurement will lie in it.

❷ Empirical Rule

Given a distribution of measurement that is approximately mound shape.

1) The interval $(\mu \pm \sigma)$ contains approximately 68% of measurements.
2) The interval $(\mu \pm 2\sigma)$ contains approximately 95% of measurements.
3) The interval $(\mu \pm 3\sigma)$ contains approximately 99.7% of measurements.

Empirical	Tchebysheff's
Mound shape	Any distribution
The value of k is only 1, 2, 3	Use any value within k , $k \geq 1$

Example (23): Find the proportion of the measurement that fall within 2 standard deviation of the mean using:

- Empirical Rule:

- Tchebysheff's Theorem:

Example (24): Given that the distribution of a measurement is mound- shape with mean and variance 40 and 81, respectively.

1) What is the proportion of measurement that lie within one standard deviation of the mean.

2) What is the proportion of measurement that lie within the interval (22, 58).

Example (25): A distribution of measurements is relatively mound-shaped with mean 50 and standard deviation 10.

a. What proportion of the measurements will fall between 40 and 60?

b. What proportion of the measurements will fall between 30 and 70?

c. What proportion of the measurements will fall between 30 and 60?

d. If a measurement is chosen at random from this distribution, what is the probability that it will be greater than 60?

5 A Check on the Calculation of S

The approximate value of the standard deviation is

$$S = \frac{R}{4}$$

where R = max – min

Example (26): Given the following data: 4, 9, 15, 20, 13, 14, 8, 14, 17, 5.

1) Calculate the approximate value of the standard deviation.

2) Calculate the standard deviation.

3) Calculate the approximate value of the variance.

6 Measures of Relative Standing

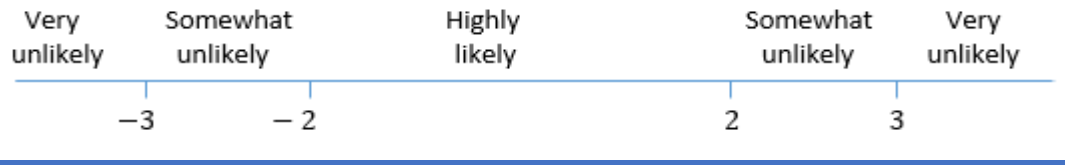
1) z-score: is a measure of how many standard deviations a data point is away from the mean of the data set. It helps in understanding the relative position of a particular data point within a distribution.

$$z\text{-score} = \frac{x-\bar{x}}{s}$$

- If $z\text{-score} = 0 \quad \Rightarrow \quad x = \bar{x}$
- If $z\text{-score} > 0 \quad \Rightarrow \quad x > \bar{x}$
- If $z\text{-score} < 0 \quad \Rightarrow \quad x < \bar{x}$

Example (27): In exam out of 50 points we have mean 25 and variance 16. Find the z-score of grade 30 and give a comment?

➢ The z-score is a valuable tool for determining whether a particular observation is highly likely to occur quite frequently (z-score between -2 and 2) or whether it is unlikely and might be considered an *outlier* (z-score exceed $|3|$).



Example (28): Let $\bar{x} = 25$ and $S^2 = 16$.

1. Find the z-score for $x = 30$.

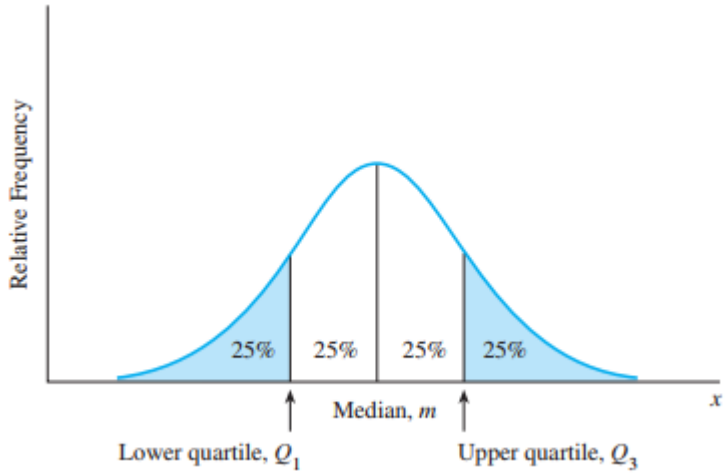
2. Find the z-score for $x = 5$.

2) Percentiles: Divide the data into 100 equal parts. For example, the 50th percentile is the median.

Definition: The p-th percentile of n ordered measurement is the value of x that is greater than $p\%$ of the measurement and less than the reaming $(100 - p)\%$ of the measurements.

Example (29): Suppose that the GPA of student is 77.8 in the summer semester which placed the student at the 60th percentile in the distribution of GPAs. Where does the GPA of 77.8 stand in relation to the GPAs of others who was in the same batch?

The 25-th and 75-th percentiles, called the **lower** and **upper quartiles,** along with the median (the 50-th percentile).



1) The first quartile = Q_1 = 25-th percentile. (Lower Quartile)
2) The median (m) = The second quartile = 50-th percentile.
3) The third quartile = Q_3 = 75-th percentile. (Upper Quartile)

How to calculate the p-th percentile?

1. Calculate the position of the p-th percentile (g).

$$g = \frac{p}{100}(n+1)$$

2. If g is an integer, the p-th percentile is $x_{(g)}$.
   If g is not an integer, the p-th percentile is

$$x_{(g)} = x_{(h)} + (g-h)(x_{(h+1)} - x_{(h)}).$$

where h is the nearest integer smaller than g and $h+1$ the nearest integer larger than g.

Example (30): Given the following data: 16, 25, 4, 18, 11, 13, 20, 8, 11 and 9. Calculate the following:

1) The first quartile.
2) The second quartile.
3) The third quartile.
4) The 60-th percentile.

❖ **Interquartile Range (IQR):** a very useful tool for understanding the spread of data and identifying extreme values in a dataset.

$$\text{IQR} = Q_3 - Q_1$$

Example (31): From the previous example find the IQR.

7 The Five-Number Summary and the Box Plot

❖ **Five number summary:**

$$\text{minimum}, Q_1, \text{median}, Q_3, \text{maximum}$$

Example (32): From the previous example find the five number summary.

Note: $\text{min} \le Q_1 \le \text{m} \le Q_3 \le \text{max}$

Example (33): Given the following data: 18, 15, 20, 3, 9, 2, 1.

1) Calculate the Interquartile Range.

2) Find the five number summary.

❖ **Box-plot:** Used to describe the distribution of the data and to detect outlier.

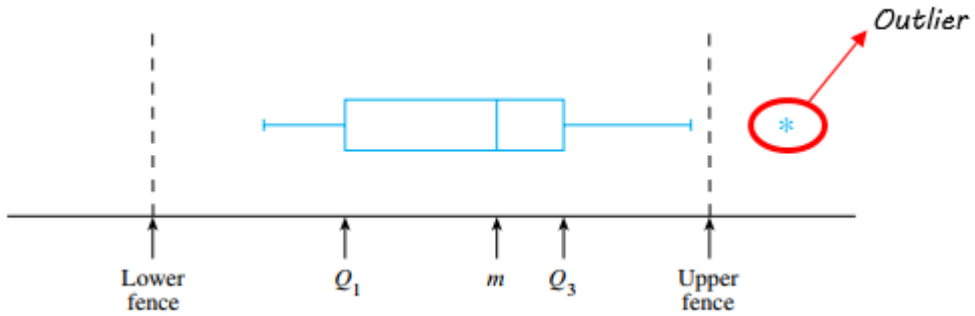
How to construct the box-plot?

Calculate the median, the upper and lower quartiles, and the IQR for the data set.

Calculate: Lower and upper fences.

Lower fence = $Q_1 - 1.5(IQR)$

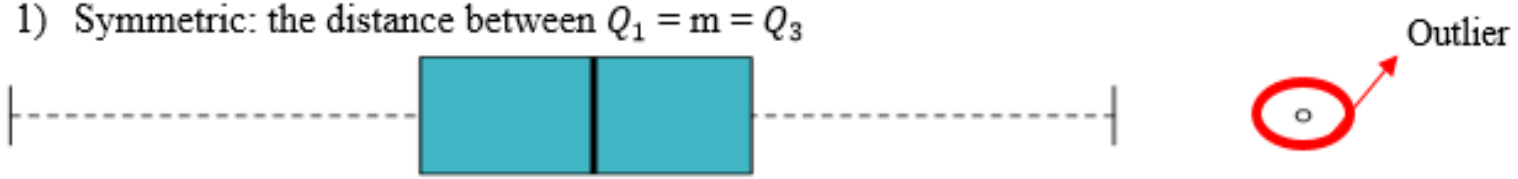
Upper fence = $Q_3 + 1.5(IQR)$



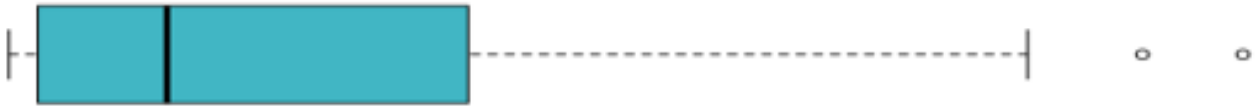
Outlier: One observation or more that are not compatible with rest of observations.

The shape of the distributions using the box-plot:

1) Symmetric: the distance between $Q_1 = \text{m} = Q_3$



2) Skewed to the right: $Q_3 - m > m - Q_1$



3) Skewed to the left: $m - Q_1 > Q_3 - m$

Example (34): Given the following data: 16, 25, 4, 18, 11, 13, 20, 8, 11 and 9.
where Q_1 = 8.75, m = 12, Q_3 = 18.5. Draw the box-plot.

Exercise (16): Given the following data: 340, 300, 520, 340, 320, 290, 260, 330.

1) Calculate the first quartile.
2) Calculate the median.
3) Calculate the third quartile.
4) Calculate the Interquartile Range.
5) Draw the box-plot.
6) Describe the shape of the distribution.
7) Determine if there is an outlier, and list the outlier if exist.

Exercise (17): True or False

1) The mean is sensitive to extreme values in a dataset.

2) The median is the middle value of a sorted dataset and is unaffected by outliers.

3) The mode is the value that occurs most frequently in a dataset.

4) The range is the difference between the largest and smallest values in a dataset.

5) The interquartile range (IQR) measures the spread of the middle 50% of the data.

6) A data set can have no mode, one mode, or multiple modes.

7) The variance and standard deviation both measure the spread or variability of the data.

8) The standard deviation is equal to the square of the variance.

9) A z-score tells how many standard deviations a data point is from the mean.

10) In a skewed distribution, the mean is always closer to the peak than the median. F

11) The five-number summary includes the minimum, maximum, mean, median, and mode.

12)A boxplot is used to show the spread and center of a dataset but does not display individual data points.

13) The empirical rule applies to all datasets, regardless of the shape of the distribution.

14) A negative z-score indicates a value below the mean.

15) The mode of a dataset can only be determined from a bar chart.