

Summary---Topic 1: Introduction to Statistics

Measure of Central Tendency

- Indicate where the data group around a central value
- Can be positive or negative or zero

Measurement	Mean	Median (Q2)	Mode
Definition	Population, μ : $\mu = \frac{\sum_{i=1}^N X_i}{N}$ Sample, \bar{X} : $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	Middle value in the ordered array <ul style="list-style-type: none">• If n or N is even, Median is the average of the 2 middle numbers• If n or N is odd, the median is the middle number	Value that occurs most often
Affected by extreme values	Yes	No	No
Single value	Yes	Yes	No

Measure of Variation

- Indicate how much the data spread out; the more the data spread out, the greater the value would be
- Must be nonnegative (i.e. positive or zero)

Measurement	Range	Interquartile range	Variance	Standard deviation
Definition	$X_{\text{largest}} - X_{\text{smallest}}$	$Q_3 - Q_1$	Population, σ^2 : $\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$ Sample, S^2 : $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$	Population, σ : $\sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$ Sample, S : $\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$
Same unit as data	Yes	Yes	No	Yes

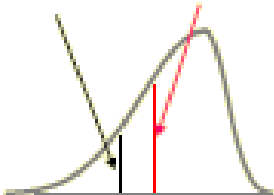
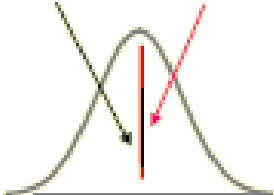
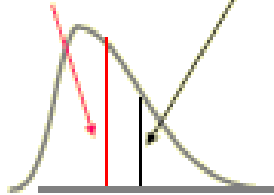
Q_1 position: $0.25 * (n + 1)$; Q_2 position: $0.5 * (n + 1)$; Q_3 position: $0.75 * (n + 1)$

When calculating the ranked position, use the following rules:

- If the result is a **whole number**, it is the ranked position to use
- If the result is a **fractional half** (e.g. 2.5, 8.5, ...), average the two corresponding data values
- If the result is **not a whole number or a fractional half**, round the result to the nearest integer to find the ranked position

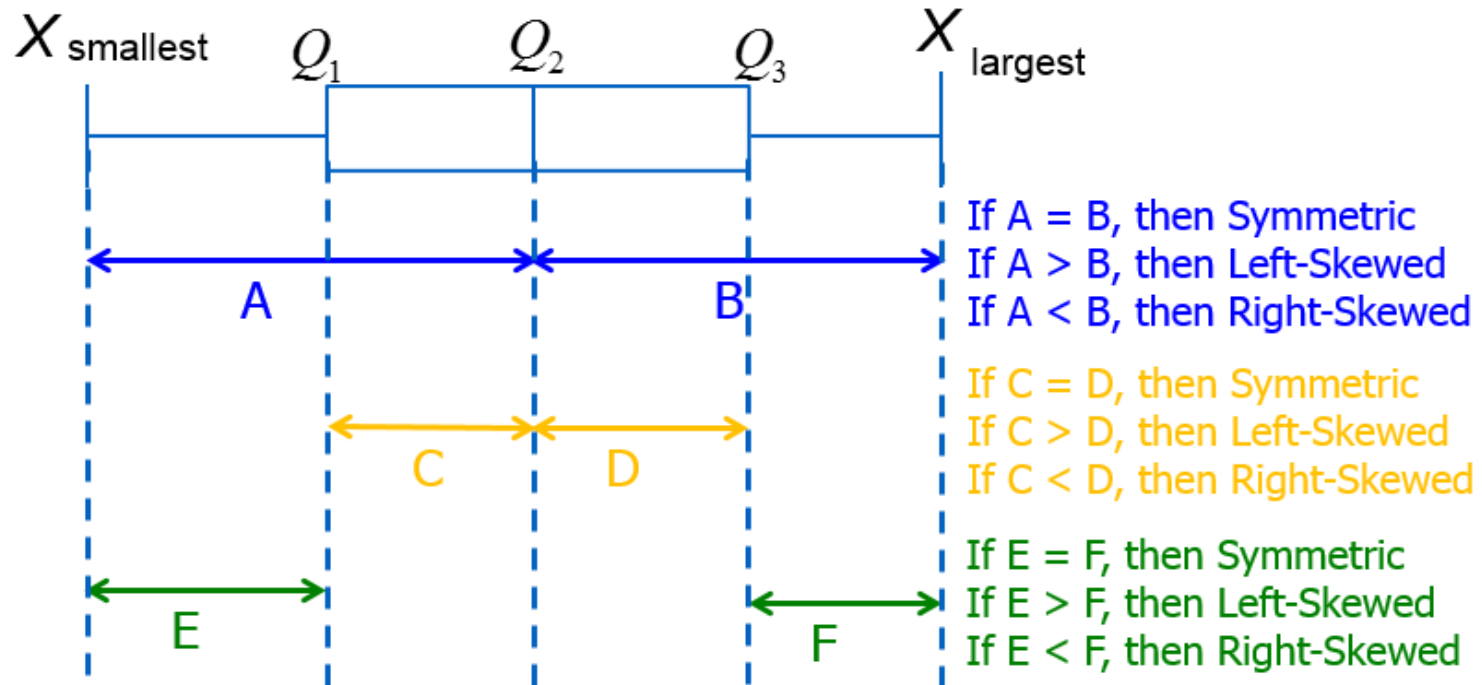
Shape of Distribution

- The shape is the pattern of the distribution of values from the lowest value to the highest value

Left-skewed	Symmetric	Right-skewed
<p>Mean < Median</p> 	<p>Mean = Median</p> 	<p>Median < Mean</p> 
<p>Tail on the left: Relatively few low values Skewness < 0</p>	<p>Evenly distributed around the mean Skewness = 0</p>	<p>Tail on the right: Relatively few high values Skewness > 0</p>

Distribution Shape and Boxplot

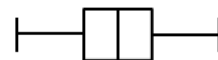
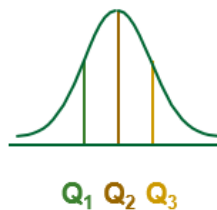
The five numbers that help describe the center, spread and shape of data are



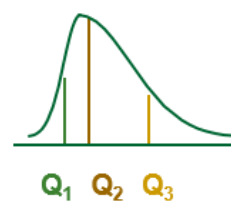
Left-Skewed



Symmetric



Right-Skewed



Exercises and Solutions

Q7. The following is a set of data for a population of size $N=10$:

7 5 11 8 3 5 2 1 10 8

- a) Compute the population mean.
- b) Compute the population standard deviation

Solution:

a) Population Mean: $\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{7+5+11+8+3+5+2+1+10+8}{10} = 6.$

b) Population variance:

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \\ &= \frac{(7-6)^2 + (5-6)^2 + (11-6)^2 + \dots + (1-6)^2 + (10-6)^2 + (8-6)^2}{10} = 10.2\end{aligned}$$

Population std: $\sigma = \sqrt{10.2} = 3.1937$

Q8. A food inspector, examining 10 bottles of a certain brand of honey, obtained the following percentages of impurities:

23.5 19.8 21.3 22.6 19.4 18.2 24.7 21.9 20.0 21.1

What are the mean and standard deviation of this sample?

Solution:

$$\text{Sample Mean: } \bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{23.5+19.8+\dots+20+21.1}{10} = 21.25.$$

Sample variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = \frac{(23.5 - 21.25)^2 + (19.8 - 21.25)^2 + \dots + (21.1 - 21.25)^2}{10 - 1} = 3.96$$

$$\text{Sample std: } S = \sqrt{3.958} = 1.9896$$

Q9. The data contain the price for two tickets with online service charges, large popcorn, and two medium soft drinks at a sample of six theater chains:

\$36.15 \$31.00 \$35.05 \$40.25 \$33.75 \$43.00

- a) Compute the mean and median
- b) Compute the variance, standard deviation and range
- c) Are the data skewed? If so, how?
- d) Based on the results of (a) through (c), describe the data.

Solution:

a) Sample Mean: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{36.15 + \dots + 33.75 + 43}{6} = 36.53$

Sorted Array: 31, 33.75, 35.05, 36.15, 40.25, 43

Find the position of median: $(n+1)*0.5 = (6+1)*0.5 = 3.5$

Compute the median: $(3^{\text{rd}} \text{ obs} + 4^{\text{th}} \text{ obs})/2 = (35.05 + 36.15)/2 = 35.6$

b) variance: $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{(36.15 - 36.53)^2 + \dots + (43 - 36.53)^2}{6-1} = 19.27$

standard deviation: $S = \sqrt{19.27} = 4.39$

range: $X_{largest}(\text{maximum}) - X_{smallest}(\text{minimum}) = 43 - 31 = 12$

- c) Since the mean is only slightly greater than the median, the data are slightly right-skewed.

Q9. The data contain the price for two tickets with online service charges, large popcorn, and two medium soft drinks at a sample of six theater chains:

\$36.15 \$31.00 \$35.05 \$40.25 \$33.75 \$43.00

- a) Compute the mean and median
- b) Compute the variance, standard deviation and range
- c) Are the data skewed? If so, how?
- d) Based on the results of (a) through (c), describe the data.

Solution:

- a) Sample Mean: $\bar{X} = 36.53$;
Sorted Array: 31, 33.75, 35.05, 36.15, 40.25, 43
median: = 35.6
- b) variance: $S^2 = 19.27$
standard deviation: $S = 4.39$
range = 12
- c) Since the mean is only slightly greater than the median, the data are slightly right-skewed
- d) There is a \$12 difference between the most expensive and the least expensive outlet. The prices vary around \$36.53 with half of the outlets being more expensive than \$35.6.

Q10. The data contains the total fat, in grams per serving, for a sample of 20 chicken sandwiches from fast-food chains. The data are as follows:

4	5	7	8	16	19	19	20	20	23
24	25	29	29	30	30	30	30	50	56

- Compute the first quartile (Q_1), the third quartile (Q_3), and the interquartile range.
- List the five-number summary.
- Construct a boxplot and describe the shape.

Q_1 position: $0.25 * (n + 1)$; Q_2 position: $0.5 * (n + 1)$; Q_3 position: $0.75 * (n + 1)$

When calculating the ranked position, use the following rules:

- If the result is a **whole number**, it is the ranked position to use
- If the result is a **fractional half** (e.g. 2.5, 8.5, ...), average the two corresponding data values
- If the result is **not a whole number or a fractional half**, round the result to the nearest integer to find the ranked position

Solution:

- Find the position of Q_1, Q_2, Q_3 :

- Q_1 position: $(20+1)*0.25 = 5.25 \rightarrow Q_1 = 5^{\text{th}} \text{ obs} = 16$
- Q_2 position: $(20+1)*0.5 = 10.5 \rightarrow Q_2 = (10^{\text{th}} \text{ obs} + 11^{\text{th}} \text{ obs})/2 = (23+24)/2 = 23.5$
- Q_3 position: $(20+1)*0.75 = 15.75 \rightarrow Q_3 = 16^{\text{th}} \text{ obs} = 30$
- interquartile range = $Q_3 - Q_1 = 30 - 16 = 14$

Q10. The data contains the total fat, in grams per serving, for a sample of 20 chicken sandwiches from fast-food chains. The data are as follows:

4	5	7	8	16	19	19	20	20	23
24	25	29	29	30	30	30	30	50	56

b) List the five-number summary.

Solution:

b) The five numbers that help describe the center, spread and shape of data are

minimum -- Q_1 -- Q_2 (Median) -- Q_3 -- maximum

4 ----- 16 ----- 23.5 ----- 30 ---- 56

Q10. The data contains the total fat, in grams per serving, for a sample of 20 chicken sandwiches from fast-food chains. The data are as follows:

4	5	7	8	16	19	19	20	20	23
24	25	29	29	30	30	30	30	50	56

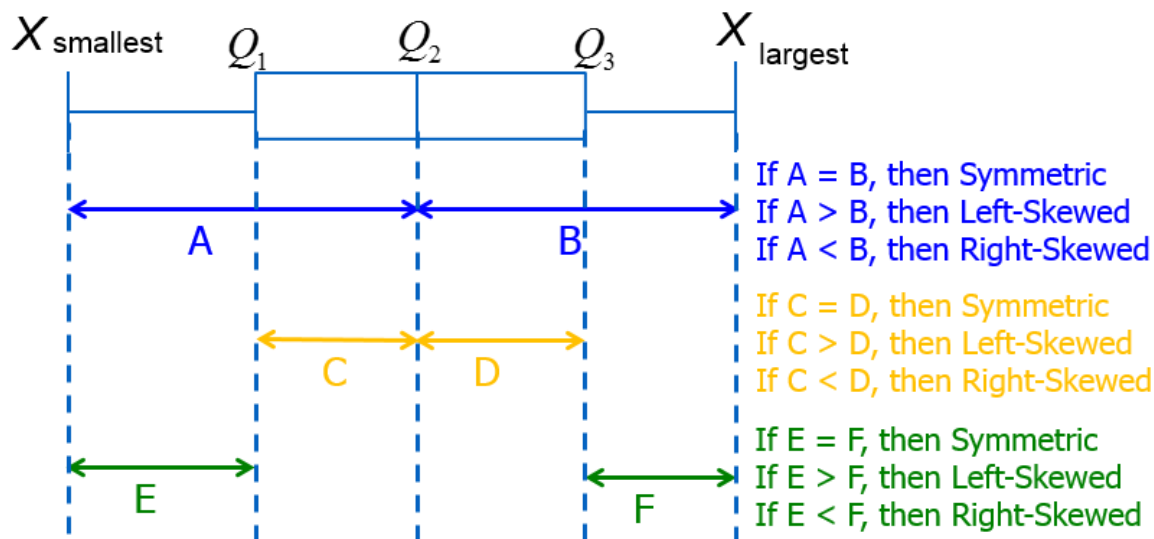
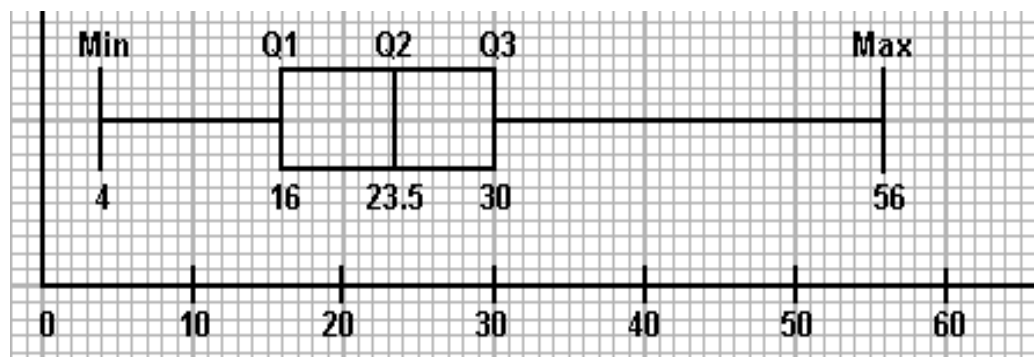
c) Construct a boxplot and describe the shape.

Solution:

c) Boxplot

- $A = Q_2 - \text{minimum} = 23.5 - 4 = 19.5$;
 $B = \text{maximum} - Q_2 = 56 - 23.5 = 32.5$;
 $A < B \rightarrow \text{right-skewed}$
- $C = Q_2 - Q_1 = 23.5 - 16 = 7.5$;
 $D = Q_3 - Q_2 = 30 - 23.5 = 6.5$;
 $C > D \rightarrow \text{left-skewed}$
- $E = Q_1 - \text{minimum} = 16 - 4 = 12$;
 $F = \text{maximum} - Q_3 = 56 - 30 = 26$;
 $E < F \rightarrow \text{right-skewed}$

Therefore, the distribution is right-skewed.



Q11. The following data is the number of vitamin supplements sold by a health food store in a sample of 11 days:

19 19 20 20 20 22 23 25 26 27 30

- What are the average and standard deviation of daily sale of vitamin supplements of the health food store?
- Work out a five-number summary of the data in the sample. Comment on the distribution of the sample data.

Solution:

a) Sample Mean/Average: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{19+\dots+30}{11} = 22.8182$

Std: $S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{(19-22.82)^2 + \dots + (30-22.82)^2}{11-1}} = 3.7099$

b) Min. value = 19

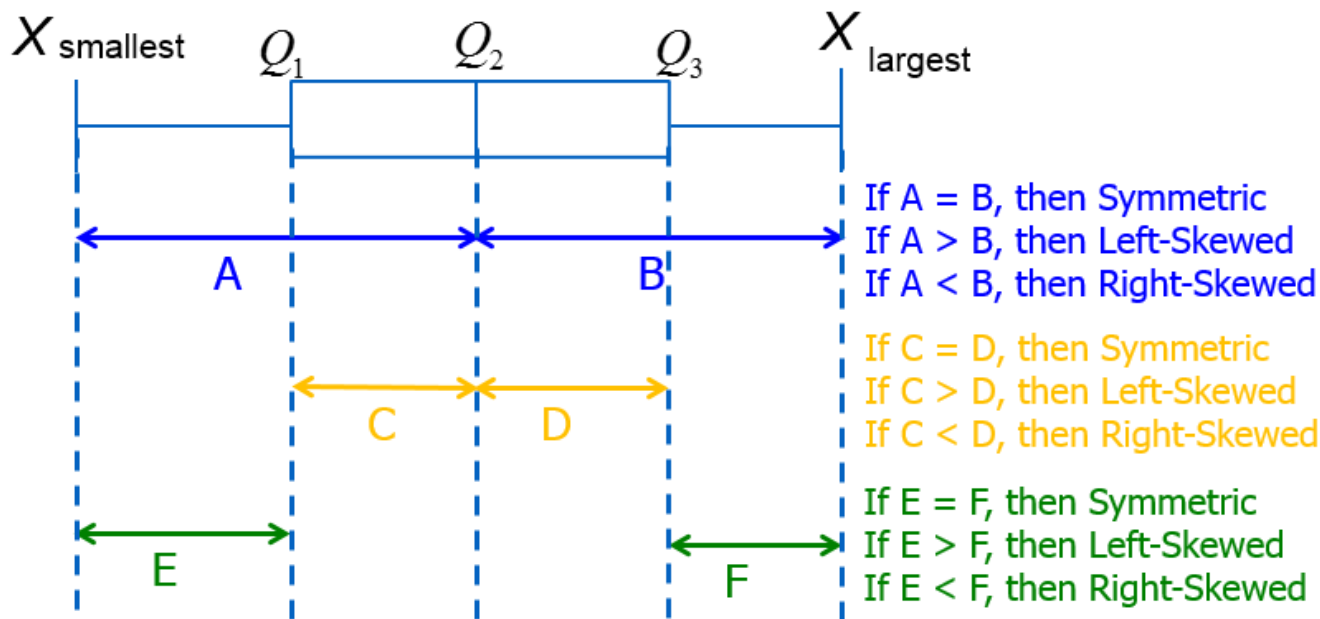
Q1 position: $(11+1)*0.25=3 \rightarrow Q1=3^{\text{th}} \text{ obs} = 20$

Q2 position: $(11+1)*0.5=6 \rightarrow Q2=6^{\text{th}} \text{ obs} = 22$

Q3 position: $(11+1)*0.75=9 \rightarrow Q3=9^{\text{th}} \text{ obs} = 26$

Max. value = 30

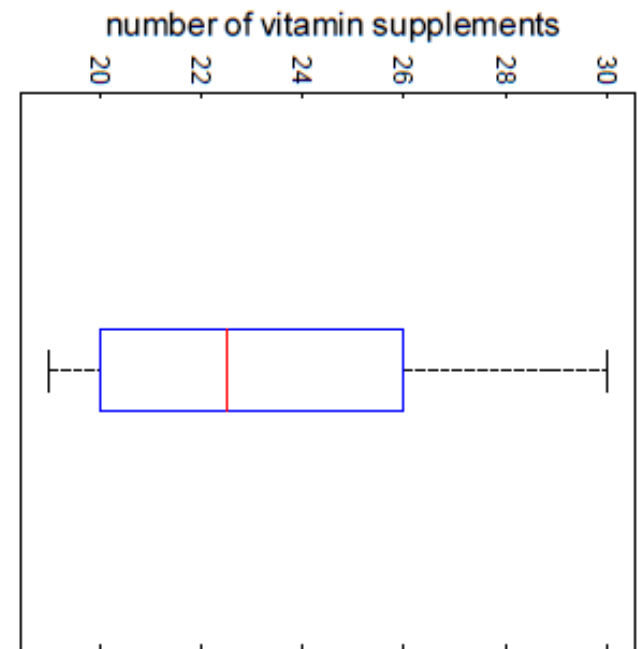
Five-number summary: 19, 20, 22, 26, 30



b) Min. value = 19; $Q_1 = 20$; $Q_2 = 22$; $Q_3 = 26$
 Max. value = 30

- $A = Q_2 - \text{minimum} = 22 - 19 = 3$; $B = \text{maximum} - Q_2 = 30 - 22 = 8$;
 $A < B \rightarrow$ right-skewed
- $C = Q_2 - Q_1 = 22 - 20 = 2$; $D = Q_3 - Q_2 = 26 - 22 = 4$;
 $C < D \rightarrow$ right-skewed
- $E = Q_1 - \text{minimum} = 20 - 19 = 1$; $F = \text{maximum} - Q_3 = 30 - 26 = 4$;
 $E < F \rightarrow$ right-skewed

Therefore, the distribution is right-skewed.



Q12. A bank branch located in a commercial district of a city has developed an improved process for serving customers during the 12:00 to 1 p.m. peak lunch period. The waiting time in minutes (operationally defined as the time the customer enters the line to the time he or she is served) of all customers during this hour is recorded over a period of a week. A random sample of 15 customers is selected, and the results are as follows:

0.38	2.34	3.02	3.2	3.54	4.21	4.5	4.77
5.12	5.13	5.55	6.1	6.19	6.46	3.79	

Another branch located in a residential area is most concerned with the Friday evening hours from 5 to 7 p.m. The waiting time in minutes (operationally defined as the time the customer enters the line to the time he or she is served) of all customers during these hours is recorded over a period of a week. A random sample of 15 customers is selected, and the results are as follows:

3.82	4.08	5.47	5.64	5.79	5.9	6.17	
6.68	8.01	8.02	8.35	8.73	9.66	9.91	10.5

a) For each bank branch, compute the mean, median and interquartile range.

Solution:

a) Bank branch in commercial district:

Sample Mean/Average: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = 4.287 \text{ mins}$

Bank branch in residential area:

Sample Mean/Average: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = 7.115 \text{ mins}$

Q12. Bank branch in commercial district: A random sample of 15 customers' waiting time:

0.38	2.34	3.02	3.2	3.54	3.79	4.21	
4.5	4.77	5.12	5.13	5.55	6.1	6.19	6.46

Bank branch in residential area: A random sample of 15 customers' waiting time:

3.82	4.08	5.47	5.64	5.79	5.9	6.17	
6.68	8.01	8.02	8.35	8.73	9.66	9.91	10.5

a) For each bank branch, compute the mean, median and interquartile range.

Solution:

a) Find the position of Q1,Q2,Q3:

- Q1 position: $(15+1)*0.25 = 4$
➔ Q1_commercial = 4th obs = 3.2 mins; Q1_residential = 4th obs = 5.64 mins;
- Q2 position: $(15+1)*0.5 = 8$
➔ Q2_commercial = 8th obs = 4.5 mins; Q2_residential = 8th obs = 6.68 mins;
- Q3 position: $(15+1)*0.75 = 12$
➔ Q3_commercial = 12th obs = 5.55 mins; Q3_residential = 12th obs = 8.73mins;

interquartile range_commercial = Q3_commercial - Q1_commercial = 2.35 mins;

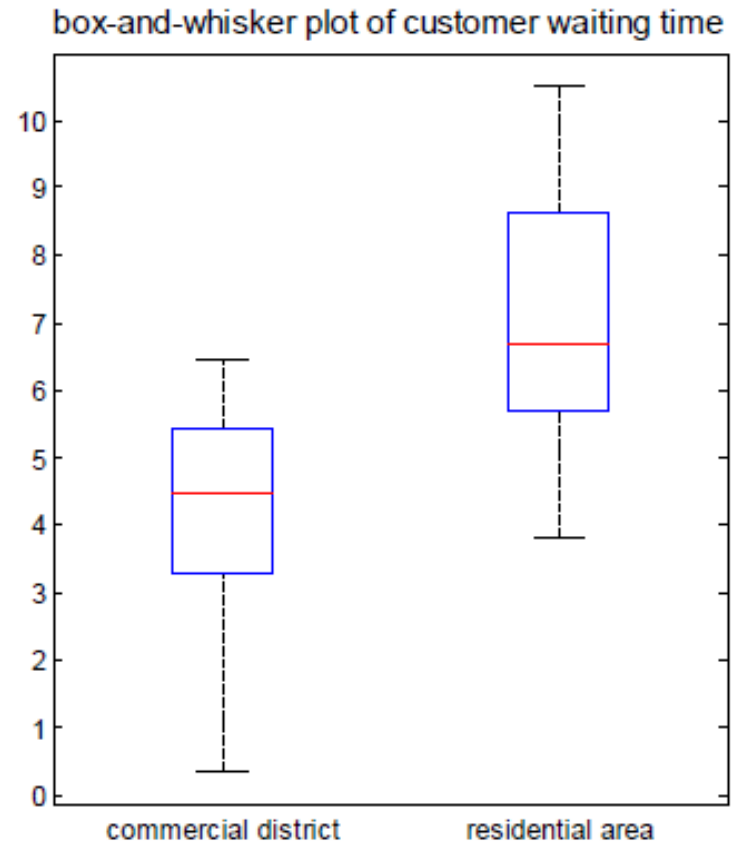
interquartile range _ residential = Q3_residential - Q1 _residential = 3.09 mins;

Q12. b) Form the box-and-whisker plot, and describe the shape of the distribution of waiting time at the two bank branches.

	min	Q1	Q2	Q3	max
commercial district	0.38	3.2	4.5	5.55	6.46
residential area	3.82	5.64	6.68	8.73	10.5

	A	B	C	D	E	F
commercial district	4.12	1.96	1.3	1.05	2.82	0.91
residential area	2.86	3.82	1.04	2.05	1.82	1.77

- The distribution of waiting time for the bank branch in commercial district is left-skewed;
- The distribution of waiting time for the bank branch in residential area is right-skewed;



c) Compare and contrast the distributions of the waiting time at the two bank branches.

Solution: The central tendency of the waiting time for the bank branch located in the commercial district is lower than that of the branch located in residential area.

Also, the normal waiting time for residential area is longer than that of commercial area.