# Probability and Statistics

Slides set 3

# Measure of position (arrangement is necessary)

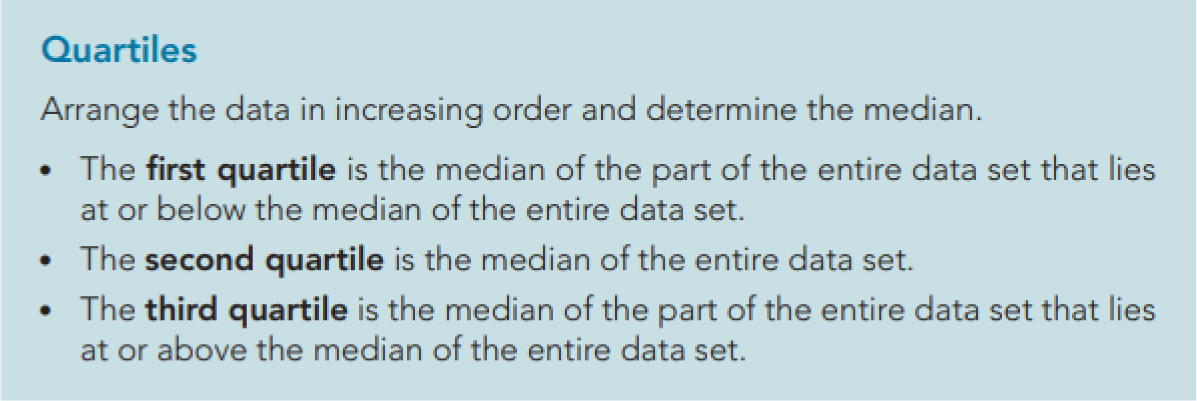
- Quartiles (divide the arranged data in four equal parts)
- Deciles (divide the arranged data in ten equal parts)
- Percentiles (divide the arranged data in hundred equal parts)

# Quartiles (divide the arranged data in four equal parts)

$$Q_1 = lower\ quartile$$
$$Q_2 = Median$$
$$Q_3 = upper\ quartile$$

$$data < Q_1\ are\ 25\%\ and\ data > Q_1\ are\ 75\%$$
$$data < Q_2\ are\ 50\%\ and\ data > Q_2\ are\ 50\%$$
$$data < Q_3\ are\ 75\%\ and\ data > Q_3\ are\ 25\%$$

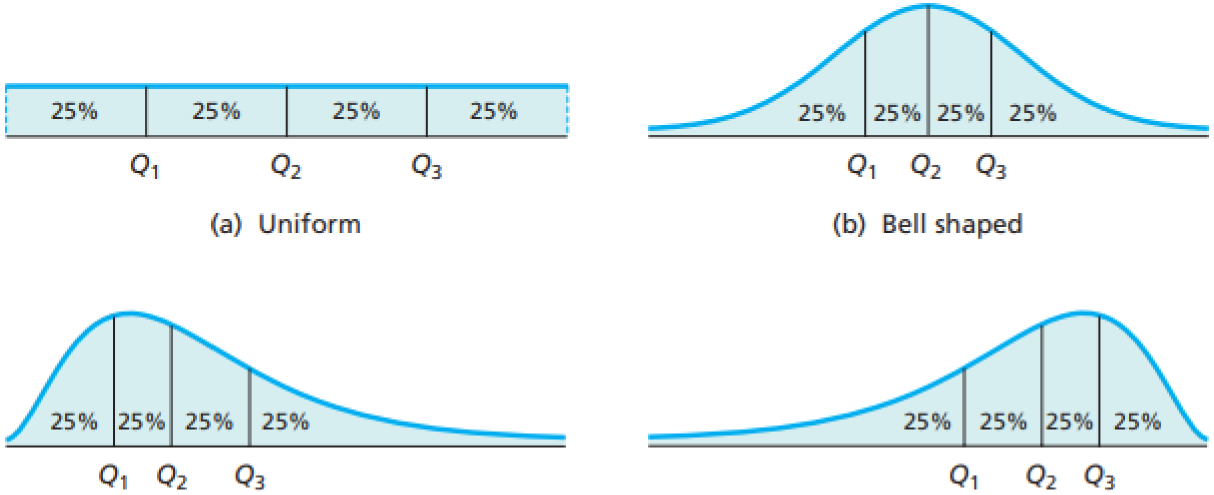
# Quartiles (divide the arranged data in four equal parts)



### Quartiles

Arrange the data in increasing order and determine the median.

- The **first quartile** is the median of the part of the entire data set that lies at or below the median of the entire data set.
- The **second quartile** is the median of the entire data set.
- The **third quartile** is the median of the part of the entire data set that lies at or above the median of the entire data set.



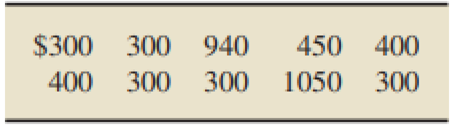
(a) Uniform

(b) Bell shaped

(c) Right skewed

(d) Left skewed

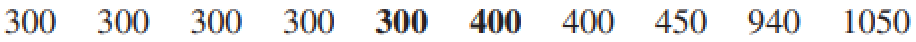
# Example (Quartiles of raw data)



| | | | | |
|---|---|---|---|---|
| $300 | 300 | 940 | 450 | 400 |
| 400 | 300 | 300 | 1050 | 300 |

Compute $Q_1$ *and* $Q_3$ *of above data*

Arrangement of above data is,



300 300 300 300 **300** **400** 400 450 940 1050

$$Q_i = \frac{i(n+1)}{4}^{th\,value}$$

$$Q_1 = \frac{(n+1)}{4}^{th}$$

$$Q_1 = \frac{11}{4} = 3.75^{th} = 3^{rd}\ value + 0.75\left(4^{th} - 3^{rd}\right) = 300 + 0.75(300 - 300) = 300$$

$$Q_3 = \frac{3(n+1)}{4}^{th}$$

$$Q_3 = \frac{33}{4} = 8.25^{th} = 8^{th}\ value + 0.25\left(9^{th} - 8^{th}\right) = 450 + 0.25(940 - 450) = 572.5$$

Example (Quartiles of single-value grouping type data)

| x | f |
|---|---|
| 0 | 11 |
| 1 | 23 |
| 2 | 3 |
| 3 | 12 |
| 4 | 10 |
| 5 | 11 |
| 6 | 21 |
|  | 91 |

$$\sum f$$

For finding ith quartile class ($Q_i$), compute $\frac{i\sum f}{4}$

For finding $Q_1$, compute $\frac{\sum f}{4} = \frac{91}{4} = 22.75^{th}\ value$

| x | f | c.f(<) |
|---|---|---|
| 0 | 11 | 11 |
| 1 | 23 | 34 |
| 2 | 3 | 37 |
| 3 | 12 | 49 |
| 4 | 10 | 59 |
| 5 | 11 | 70 |
| 6 | 21 | 91 |
|  | 91 |  |

$Q_1$ class

$Q_1$

# Example (Quartiles of grouped data)

| Class Interval | Class Boundaries | Frequency (f) |
|---|---|---|
| 30-39 | 29.5-39.5 | 3 |
| 40-49 | 39.5-49.5 | 1 |
| 50-59 | 49.5-59.5 | 8 |
| 60-69 | 59.5-69.5 | 10 |
| 70-79 | 69.5-79.5 | 7 |
| 80-89 | 79.5-89.5 | 7 |
| 90-99 | 89.5-99.5 | 4 |
| | | 40 |

For finding $Q_3$, first compute $\frac{3\sum f}{4} = \frac{120}{4} = 30^{th}\ \ value$

| Class Interval | Class Boundaries | Frequency (f) | c.f(<) |
|---|---|---|---|
| 30-39 | 29.5-39.5 | 3 | 3 |
| 40-49 | 39.5-49.5 | 1 | 4 |
| 50-59 | 49.5-59.5 | 8 | 12 |
| 60-69 | 59.5-69.5 | 10 | 22 |
| 70-79 | 69.5-79.5 | 7 | 29 |
| 80-89 | 79.5-89.5 | 7 | 36 |
| 90-99 | 89.5-99.5 | 4 | 40 |
| | | 40 | |

$Q_3$ class

Now, use formula for $Q_i = l + \frac{h}{f}\left(i\frac{\sum f}{4} - c.f(<)\right)$

Where

$$l = L.C.B\ of\ Q_i\ class$$
$$h = width\ of\ Q_i\ class$$
$$f = frequency\ of\ Q_i\ class$$
$$c, f(<) = c.f(<)\ of\ previous\ class\ to\ Q_i class$$

Formula : $\text{Q}_3 = l + \frac{h}{f}\left(3\frac{\sum f}{4} - c.f(<)\right)$

$$= 79.5 + \frac{10}{7}\left(\frac{120}{4} - 29\right)$$

$$= 80.928\ ans$$

- Deciles (divide the arranged data in ten equal parts)

For ith deciles $(d_i)$ , in all formulae, denominator of $Q_i$ is replaced by 10

- Percentiles (divide the arranged data in hundred equal parts)

For ith deciles $(P_i)$ , in all formulae, denominator of $Q_i$ is replaced by 100

## The Interquartile Range

### Interquartile Range

The **interquartile range,** or **IQR,** is the difference between the first and third quartiles; that is, $\text{IQR} = Q_3 - Q_1$.

## Outliers

In data analysis, the identification of **outliers**—observations that fall well outside the overall pattern of the data—is important. An outlier requires special attention. It may be the result of a measurement or recording error, an observation from a different population, or an unusual extreme observation. Note that an extreme observation need not be an outlier; it may instead be an indication of skewness.

As an example of an outlier, consider the data set consisting of the individual wealths (in dollars) of all U.S. residents. For this data set, the wealth of Bill Gates is an outlier—in this case, an unusual extreme observation.

Whenever you observe an outlier, try to determine its cause. If an outlier is caused by a measurement or recording error, or if for some other reason it clearly does not belong in the data set, the outlier can simply be removed. However, if no explanation for the outlier is apparent, the decision whether to retain it in the data set can be a difficult judgment call.

We can use quartiles and the IQR to identify potential outliers, that is, as a diagnostic tool for spotting observations that may be outliers. To do so, we first define the *lower limit* and the *upper limit* of a data set.

### Lower and Upper Limits

The **lower limit** and **upper limit** of a data set are

$$\text{Lower limit} = Q_1 - 1.5 \cdot \text{IQR};$$
$$\text{Upper limit} = Q_3 + 1.5 \cdot \text{IQR}.$$

Observations that lie below the lower limit or above the upper limit are **potential outliers.** To determine whether a potential outlier is truly an outlier, you should perform further data analyses by constructing a histogram, stem-and-leaf diagram, and other appropriate graphics that we present later.

## The Five-Number Summary

From the three quartiles, we can obtain a measure of center (the median, $Q_2$) and measures of variation of the two middle quarters of the data, $Q_2 - Q_1$ for the second quarter and $Q_3 - Q_2$ for the third quarter. But the three quartiles don't tell us anything about the variation of the first and fourth quarters.

To gain that information, we need only include the minimum and maximum observations as well. Then the variation of the first quarter can be measured as the difference between the minimum and the first quartile, $Q_1 - \text{Min}$, and the variation of the fourth quarter can be measured as the difference between the third quartile and the maximum, $\text{Max} - Q_3$.

Thus the minimum, maximum, and quartiles together provide, among other things, information on center and variation.

### Five-Number Summary

The **five-number summary** of a data set is Min, $Q_1$, $Q_2$, $Q_3$, Max.

In Example 3.17, we show how to obtain and interpret the five-number summary of a set of data.

## Boxplots

A **boxplot,** also called a **box-and-whisker diagram,** is based on the five-number summary and can be used to provide a graphical display of the center and variation of a data set. These diagrams, like stem-and-leaf diagrams, were invented by Professor John Tukey.[†]

To construct a boxplot, we also need the concept of *adjacent values.* The **adjacent values** of a data set are the most extreme observations that still lie within the lower and upper limits; they are the most extreme observations that are not potential outliers. Note that, if a data set has no potential outliers, the adjacent values are just the minimum and maximum observations.

### PROCEDURE 3.1 To Construct a Boxplot

**Step 1** **Determine the quartiles.**

**Step 2** **Determine potential outliers and the adjacent values.**

**Step 3** **Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.**

**Step 4** **Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.**

**Step 5** **Plot each potential outlier with an asterisk.**

**Note:**

- In a boxplot, the two lines emanating from the box are called **whiskers.**
- Boxplots are frequently drawn vertically instead of horizontally.
- Symbols other than an asterisk are often used to plot potential outliers.

## Boxplots

*Weekly TV-Viewing Times* The weekly TV-viewing times for a sample of 20 people are given in Table 3.12 on page 116. Construct a boxplot for these data.

**Solution** We apply Procedure 3.1. For easy reference, we repeat here the ordered list of the TV-viewing times.

5 15 16 20 21 25 26 27 30 30 31 32 32 34 35 38 38 41 43 66

**Step 1 Determine the quartiles.**

In Example 3.15, we found the quartiles for the TV-viewing times to be $Q_1 = 23$, $Q_2 = 30.5$, and $Q_3 = 36.5$.

**Step 2 Determine potential outliers and the adjacent values.**

As we found in Example 3.18(b), the TV-viewing times contain one potential outlier, 66. Therefore, from the ordered list of the data, we see that the adjacent values are 5 and 43.

**Step 3 Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.**

See Fig. 3.9(a).

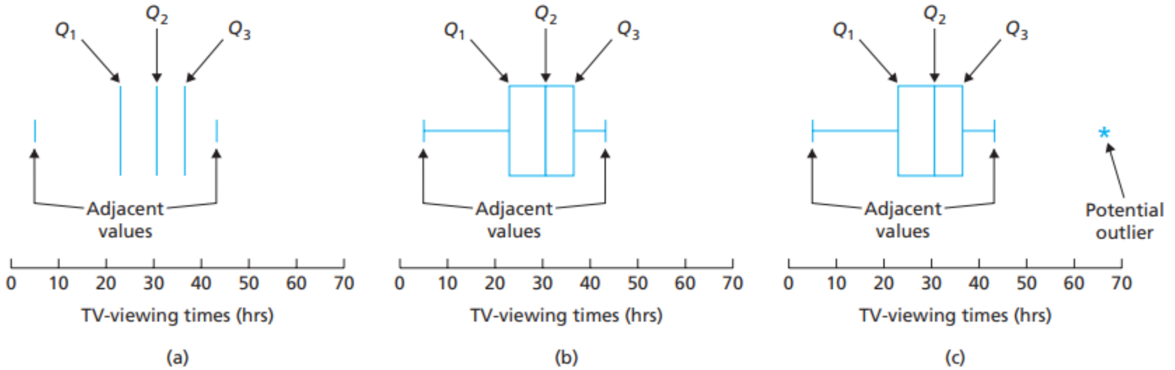
**Step 4 Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.**

See Fig. 3.9(b).

**Step 5 Plot each potential outlier with an asterisk.**

As we noted in Step 2, this data set contains one potential outlier—namely, 66. It is plotted with an asterisk in Fig. 3.9(c).

**FIGURE 3.9** Constructing a boxplot for the TV-viewing times

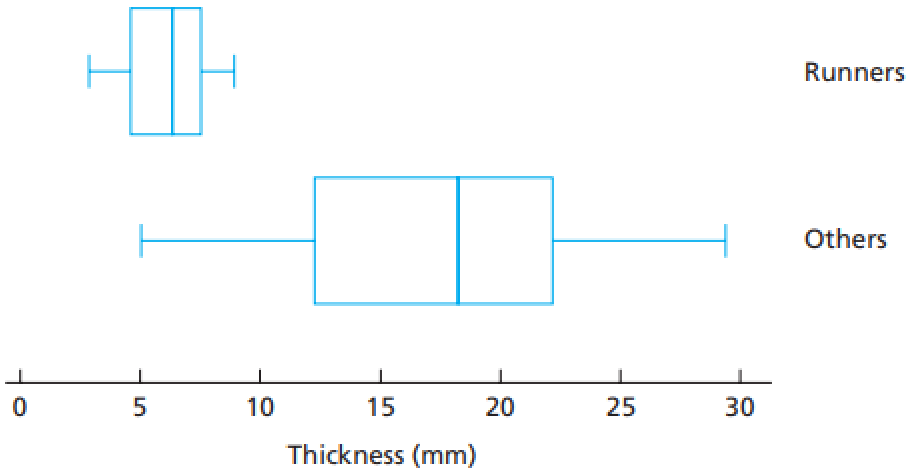


## Comparing Data Sets by Using Boxplots

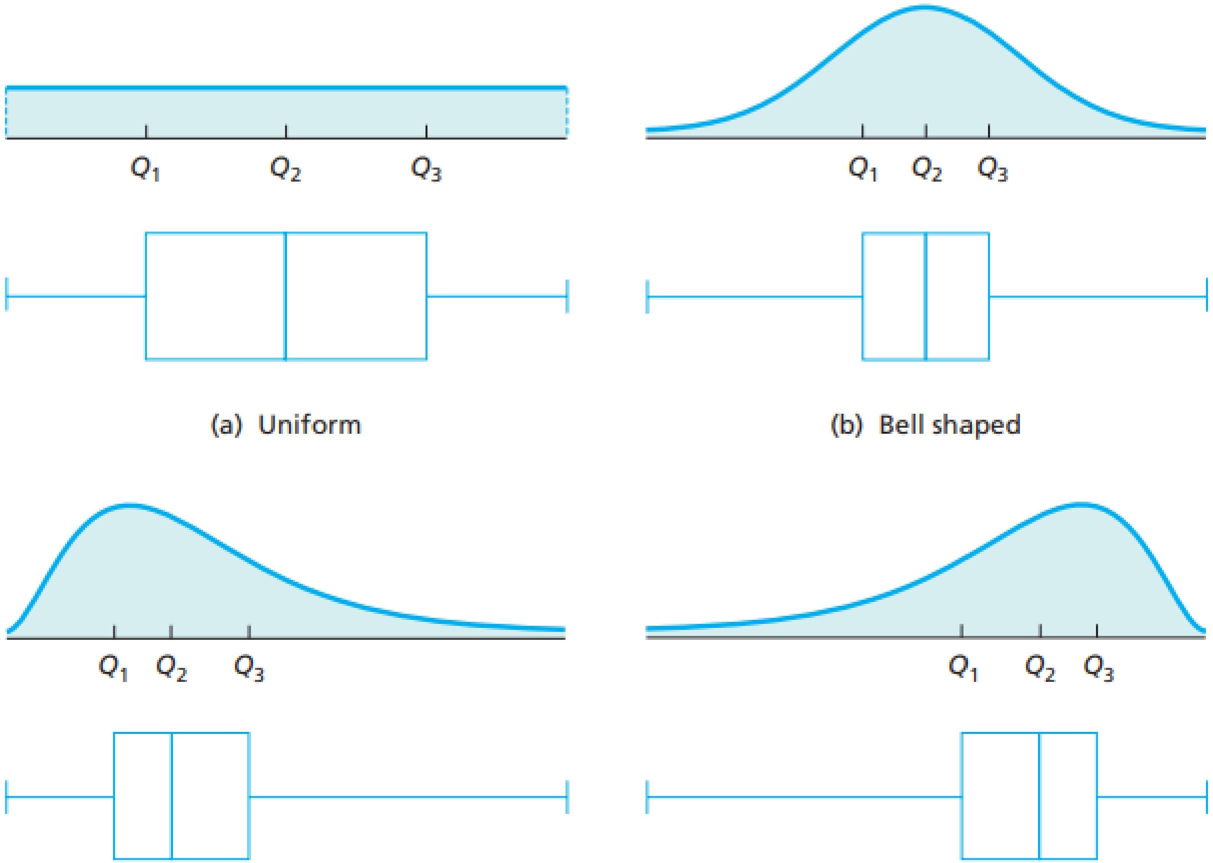
***Skinfold Thickness*** A study titled "Body Composition of Elite Class Distance Runners" was conducted by M. Pollock et al. to determine whether elite distance runners are actually thinner than other people. Their results were published in *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies* (P. Milvey (ed.), New York: New York Academy of Sciences, p. 366). The researchers measured skinfold thickness, an indirect indicator of body fat, of samples of runners and nonrunners in the same age group. The sample data, in millimeters (mm), presented in Table 3.13 are based on their results. Use boxplots to compare these two data sets, paying special attention to center and variation.

| Runners | | | Others | | | |
|---|---|---|---|---|---|---|
| 7.3 | 6.7 | 8.7 | 24.0 | 19.9 | 7.5 | 18.4 |
| 3.0 | 5.1 | 8.8 | 28.0 | 29.4 | 20.3 | 19.0 |
| 7.8 | 3.8 | 6.2 | 9.3 | 18.1 | 22.8 | 24.2 |
| 5.4 | 6.4 | 6.3 | 9.6 | 19.4 | 16.3 | 16.3 |
| 3.7 | 7.5 | 4.6 | 12.4 | 5.2 | 12.2 | 15.6 |

**Solution** Figure 3.10 displays boxplots for the two data sets, using the same scale.



From Fig. 3.10, it is apparent that, on average, the elite runners sampled have smaller skinfold thickness than the other people sampled. Furthermore, there is much less variation in skinfold thickness among the elite runners sampled than among the other people sampled. By the way, when you study inferential statistics, you will be able to decide whether these descriptive properties of the samples can be extended to the populations from which the samples were drawn.



(a) Uniform

(b) Bell shaped

(c) Right skewed

(d) Left skewed

Do questions from Ex # 3.3 (Neil.wises book)
Page number 124-126
Questions 3.113-3.135