The sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum x.$$

The mean is sensitive to extreme values.

Notation for population mean: μ.

The median

The **median** is the midpoint of a distribution.

To find the median of a data set:
Arrange all the n observations from smallest to largest.
If n is odd, the median is the center observation. If n is even, the median is the average of the two center observations.

Find the median of this data set: 1, 19,12, 3, 21, 25, 11

Ordered values are:

1, 3, 11, 12, 19, 21, 25

The median is 12.

Find the median of this data set: 1, 19,12, 3, 21, 25, 11, 32

Ordered values are:

1, 3, 11, 12, 19, 21, 25, 32

The median is (12+19)/2=15.5

Compare the mean and the median

- for roughly symmetric distribution: mean close to median
- right skewed: mean $>$ median
- left skewed: mean $<$ median

The median is not sensitive to extreme data values.

The mode

The **mode** is the value (values) that appears most frequently.
The mode can be any value of a data set.
Finding the mode:
0, 1, 1, 1, 1, 2, 2, 3, 3, 6.
The mode is 1.

The spread of data: Range and variance

Range = largest value - smallest value

The **variance** s^2 measures how far the values are from the mean, on average.
$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots(x_n-\bar{x})^2}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$.

The **standard deviation** s is the square root of the variance s^2:
$s = \sqrt{s^2}$.

Notation for population variance and standard deviation: σ^2, σ.

Find variance

Exercise: find the variance of the data set: 3,4,6,5,1,5.

$n = 6, \bar{x} = 4$ ($n = 6$ as the data has 6 values, $\bar{x}$ is the sample mean).

$s^2 = \frac{(3-4)^2+(4-4)^2+(6-4)^2+(5-4)^2+(1-4)^2+(5-4)^2}{6-1} = \frac{16}{5} = 3.2.$

$s = \sqrt{3.2} = 1.789.$

Empirical rule

If a population has a bell-shaped histogram,

- About 68% of the data are between $(\mu - \sigma, \mu + \sigma)$.
- About 95% of the data are between $(\mu - 2\sigma, \mu + 2\sigma)$.
- Almost all the data are between $(\mu - 3\sigma, \mu + 3\sigma)$.

This rule says what percentage of data are concentrated around the mean for a bell-shaped distribution.
we will examine this rule more in the next chapter.

Percentile

pth **percentile**: p% of the data are below it.
e.g., 23rth percentile of a data set is a value such that 23% of the data values are below it.

Special percentiles:
First quartile Q_1: 25th percentile: A quarter of the data are below it.
Third quartile Q_3: 75th percentile: Three quarters of the data are below it.
Second quartile Q_2: 50th percentile, median: half of the data values are below it

Find quartiles

Find : Q_1, Q_3.

1. Find the median.
2. Q_1 is the median of the observations to the left of the median (excluding median).

Q_3 is the median of the observations to the right of the median (excluding median).

Find Q_1, Q_2, Q_3 for this data set: 1, 19, 12, 3, 21, 25, 11, 32, 38

```
Ordered values are:

1, 3, 11, 12,        19,          21, 25, 32, 38
    Q1=7          median=19          Q3=28.5
```

Note Q_1 is the median of the data set (1, 3, 11, 12)
and Q_3 is the median of the data set (21,25,32,38).

Exercise

The amounts of money (cents) in coins carried by 10 students in a class:
50,35,5,97,75,0,5,87,23,65.
Find the median.
Find Q_1, Q_3.

```
0    5    5    23   35              50   65    75    87   97
          --                                   --
          Q1=5          Q2=42.5          Q3=75
```

$IQR = 75 - 5 = 70$
$1.5IQR = 1.5 * 70 = 105$
$Q_1 - 105 = -100,$
$Q_3 + 105 = 180.$

Five number summary

The **five-number summary** of a distribution consists of:
Minimum, Q1, Median, Q3, Maximum

Use quartiles to find outliers:
Inter quartile range: IQR $= Q_3 - Q_1$
Lower outlier boundary: $Q_1 - (1.5 \times \text{IQR})$
Upper outlier boundary: $Q_3 + (1.5 \times \text{IQR})$.

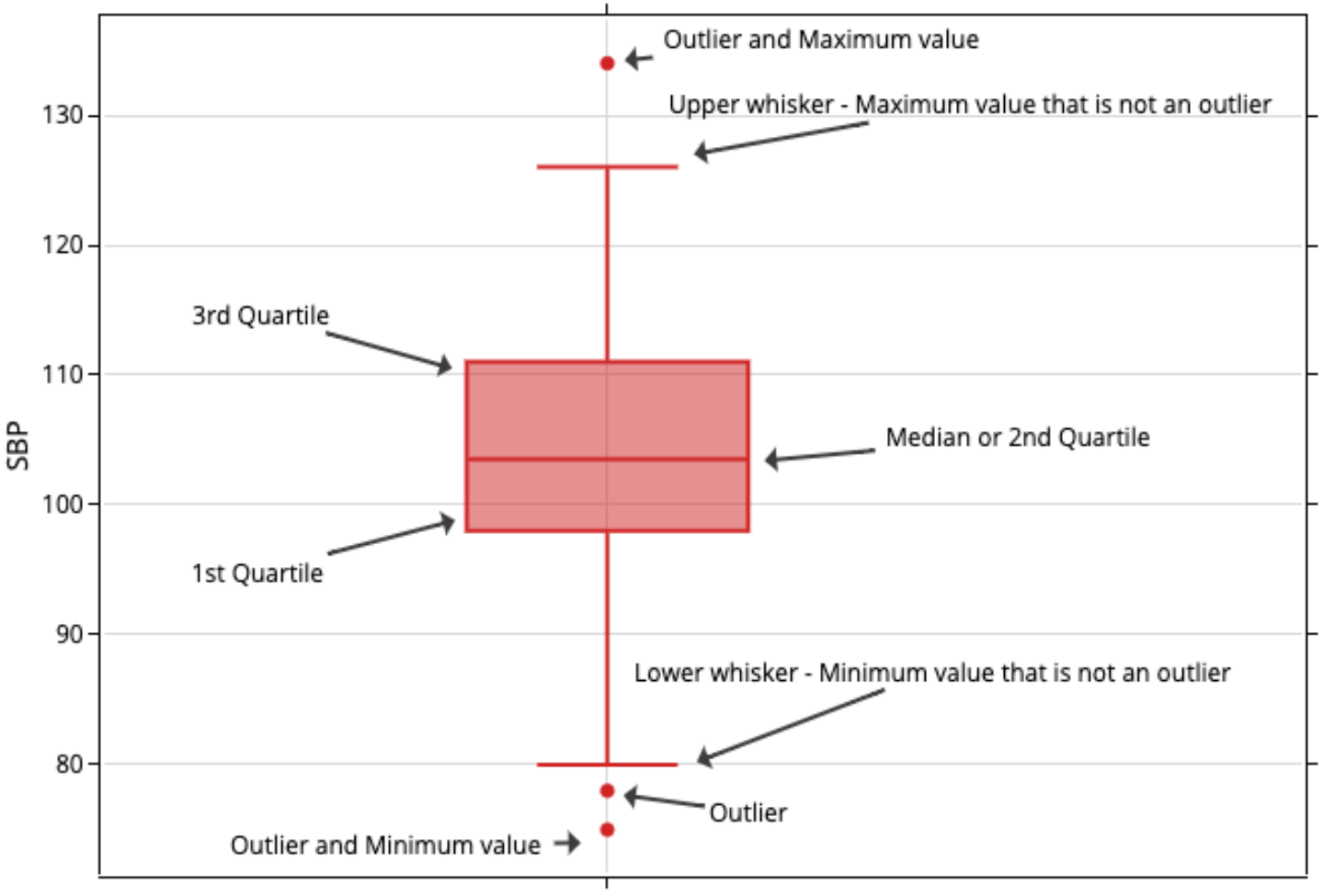
Are their any outliers in the previous data set (the coin money data)?
No. There are no values outside the outlier boundaries (computed on page 16 of the slides)

Box plot

A **boxplot** is a graph of the five-number summary.

- A central box spans Q_1 and Q_3.
- A line in the box marks the median.
- Lines extend from the box out to the smallest and largest observations that are within the outlier boundaries.
- Outliers are displayed separately.

Components of a boxplot



Make a boxplot of this data set

Number of absent students in a high school:

Table 3.11 Number of Absences

Date	Number Absent	Date	Number Absent	Date	Number Absent
Jan. 2	65	Jan. 14	59	Jan. 23	42
Jan. 3	67	Jan. 15	49	Jan. 24	45
Jan. 4	71	Jan. 16	42	Jan. 25	46
Jan. 7	57	Jan. 17	56	Jan. 28	100
Jan. 8	51	Jan. 18	45	Jan. 29	59
Jan. 9	49	Jan. 21	77	Jan. 30	53
Jan. 10	44	Jan. 22	44	Jan. 31	51
Jan. 11	41				

Make a boxplot

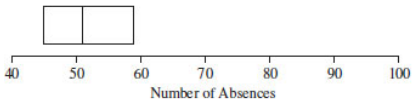
EXAMPLE 3.31 **Constructing a boxplot**

Construct a boxplot for the absence data in Table 3.11.

Solution

Step 1: In Example 3.27, we computed the median to be 51 and the first and third quartiles to be $Q_1 = 45$ and $Q_3 = 59$.

Step 2: We draw vertical lines at 45, 51, and 59, then horizontal lines to complete the box, as follows:

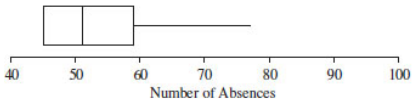


Step 3: We compute the outlier boundaries as shown in Example 3.30:

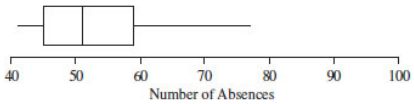
Lower outlier boundary = 45 − 1.5(14) = 24

Upper outlier boundary = 59 + 1.5(14) = 80

Step 4: The largest data value that is less than the upper boundary is 77. We draw a horizontal line from 59 up to 77, as follows:



Step 5: The smallest data value that is greater than the lower boundary is 41. We draw a horizontal line from 45 down to 41, as follows:



Step 6: We determine, as shown in Example 3.30, that the value 100 is the only outlier. We plot this point separately, to produce the boxplot shown in Figure 3.12.

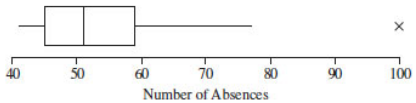


Figure 3.12 Boxplot for the absence data in Table 3.11

Side-by-side boxplots

Boxplots are best used for side-by-side comparison. We can compare multiple data sets in one plot.