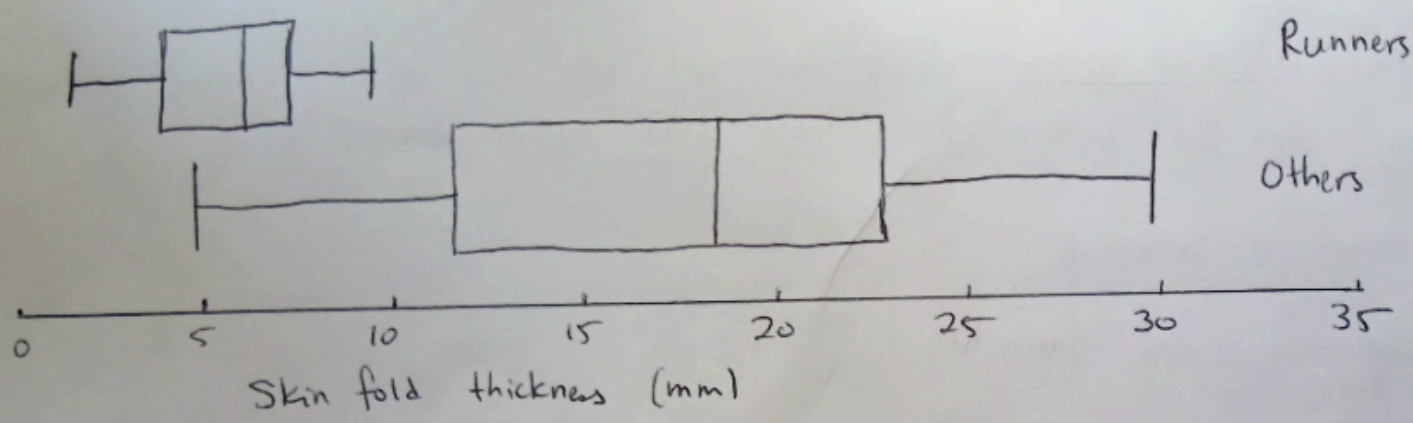
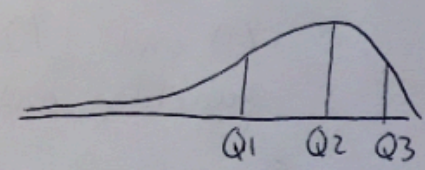
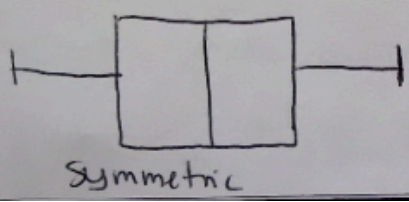
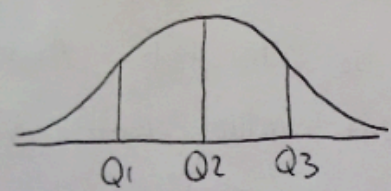
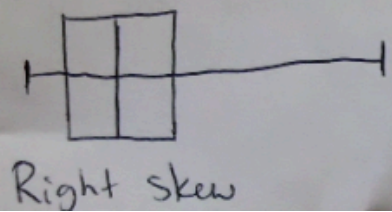
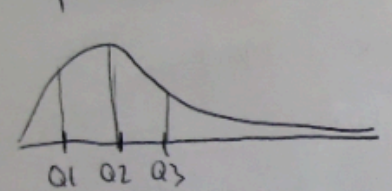


We can use boxplots to compare datasets.



Box plots to examine skew



Quartiles

For each P between 0 and 100, the p^{th} percentile is the number that divides the bottom $p\%$ of the data from the top $(100-p)\%$.

For example, P_{50} the 50th percentile is the median - 50% of the data is below and 50% is above.

The most commonly used percentiles are the quartiles, which divide the data into four parts: the 25th, 50th, and 75th percentiles. We denote these as Q_1 , Q_2 , and Q_3 respectively.

Determining Quartiles:

- 1) Arrange the data in increasing order.
- 2) Find the median of the entire dataset. This is Q_2 .
- 3) Divide the ordered dataset into two halves - top and bottom - if the number of observations is odd, include the median in both halves.
- 4) Find the median of the lower/bottom half of the data. This is Q_1 .
- 5) Find the median of the upper/top half of the data. This is Q_3 .

Ex Weekly TV-Viewing Times

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

Weekly TV-viewing times (hours)

$$\text{Median} = \frac{30+31}{2} = 30.5$$

$$Q_2 = 30.5$$

$$Q_1 = \frac{21+25}{2} = 23$$

Find the quartiles.

i) Arrange in increasing order.

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

$$n = 20 \rightarrow \frac{n+1}{2} = 10.5$$

$$n_{\text{bottom half}} = 10 \rightarrow \frac{n_{bh}+1}{2} = 5.5$$

$$n_{\text{top half}} = 10 \rightarrow \frac{n_{th}+1}{2} = 5.5$$

1) Arrange in increasing order

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

2) Find the median.

$$n = 20 \rightarrow \frac{n+1}{2} = 10.5$$

$$\frac{30+31}{2} = 30.5 \quad Q2$$

3) Bottom half: 5 15 16 20 21 | 25 26 27 30 30

$$\text{Median? } \frac{10+1}{2} = 5.5 \rightarrow \frac{21+25}{2} = 23 \quad Q1$$

4) Top half: 31 32 32 34 35 | 38 38 41 43 66

$$\frac{10+1}{2} = 5.5 \rightarrow \frac{35+38}{2} = 36.5 \quad Q3$$

So the quartiles are $Q1 = 23$ $Q2 = 30.5$ and $Q3 = 36.5$

So 25% of viewing times are less than $Q1 = 23$ hours,
25% are between $Q1 = 23$ hours and $Q2 = 30.5$ hours,
25% are between $Q2 = 30.5$ hours and $Q3 = 36.5$ hours,
and 25% are above $Q3 = 36.5$ hours.

The interquartile range or IQR is

$$IQR = Q3 - Q1$$

This is another measure of variability, used when the median is the preferred measure of center.

For the TV-viewing data,

$$IQR = 36.5 - 23 = 13.5 \text{ hours.}$$

So the middle 50% of TV-viewing times are spread over a 13.5 hour interval.

The five-number summary of a data set is
Min, Q1, Q2, Q3, Max

For the TV viewing times, the ordered data was

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

We found $Q1 = 23$, $Q2 = 30.5$, and $Q3 = 36.5$

The minimum is 5 and the maximum is 66.

So the 5-number summary is

5, 23, 30.5, 36.5, 66

We estimate first quarter variability as $Q1 - \text{Min} = 18$

second quarter as $Q2 - Q1 = 7.5$

etc.

$Q3 - Q2 = 6$

and $\text{Max} - Q3 = 29.5$

Outliers are values that fall far from the overall pattern of the data. We should consider outliers carefully for

- measurement error
- rounding error
- population or sampling problems

If there is a problem with the data, we may wish to remove the outliers. However, if there are no measurement errors, outliers can be an indicator of skew - we do NOT want to remove these data!

We can use the quartiles to look for outliers:

Lower Limit = $Q1 - 1.5 \text{ IQR}$ also "lower fence"

Upper Limit = $Q3 + 1.5 \text{ IQR}$ also "upper fence"

Observations below the lower limit or above the upper limit are potential outliers.

Ex For the TV-viewing times, we found $Q1 = 23$ and $Q3 = 36.5$
 $IQR = 36.5 - 23 = 13.5$

$$\text{Lower Limit} = Q1 - 1.5 IQR = 23 - 1.5(13.5) = 2.75 \text{ hours.}$$

$$\text{Upper Limit} = Q3 + 1.5 IQR = 36.5 + 1.5(13.5) = 56.75 \text{ hrs}$$

$\text{Min} = 5$, so there are no potential outliers below the lower limit.

$\text{Max} = 66$, which is above the upper limit. This is a potential outlier.

Boxplots AKA Box-and-Whisker diagrams

Def The adjacent values of a data set are the most extreme observations that still lie within the lower and upper limits.

Constructing a boxplot:

- 1) Determine the quartiles.
- 2) Determine potential outliers and the adjacent values.
- 3) Draw a horizontal axis on which the numbers obtained in (1) and (2) can be located. Above this axis, mark the quartiles and adjacent values with vertical lines.
- 4) Connect the quartiles to make a box, then connect the box to the adjacent value using horizontal lines.
- 5) Plot each potential outlier with an asterisk.

Notes:

- The lines emanating from the box are called whiskers.
- Boxplots may be drawn vertically instead of horizontally.
- Other symbols may be used to plot outliers.

Ex For the TV-viewing data, we found five-number summary
5, 23, 30.5, 36.5, 66
and identified 66 as a potential outlier.
The adjacent values, then, are 5 and 43.

