Today:

- Inter-Quartile Range,
- Outliers,
- Boxplots.

Reading for today: Start Chapter 4.

Quartiles and the Five Number Summary

- The five numbers are the Minimum (Q0), Lower Quartile (Q1), Median (Q2), Upper Quartile (Q3), and Maximum (Q4).

- Q1 means bigger than 1 Quarter of the data.
- Q3 means bigger than 3 Quarters of the data.

For the values {0, 1, 2, 4, 5, 5, 7, 10, 10, 12, 13, 17, 39},

the five number summary is: 0 → 3 → 7 → 12.5→ 39.

Inter-Quartile Range

- Even in the unimodal cases, neither the mean nor the median describes the data adequately.

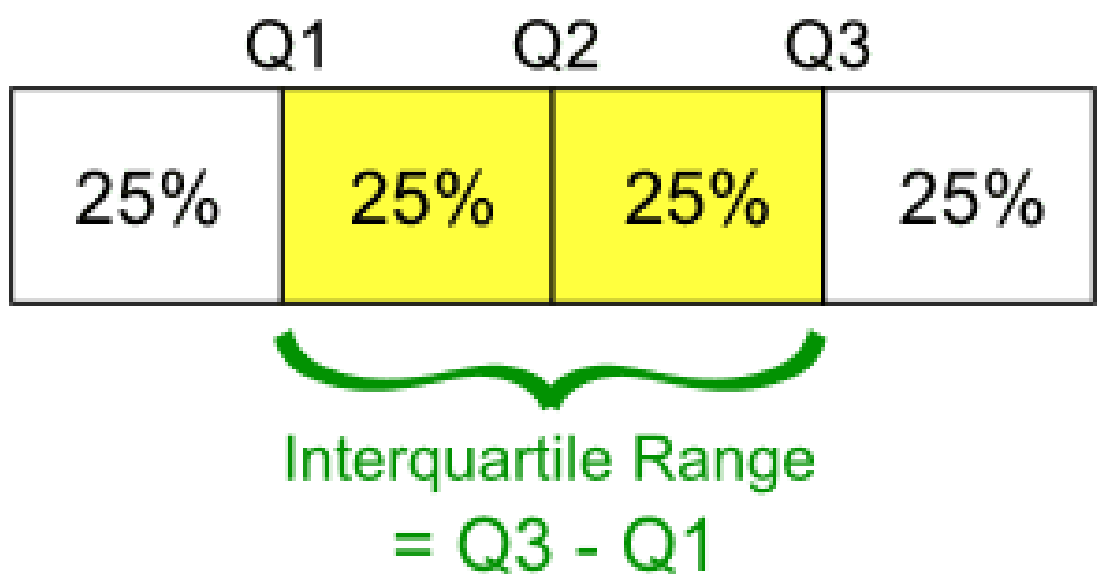
- The mean number of legs per Swede is 1.999, clearly there's something more we should know.

- The median of {30,31,32} is 31.

- The median of {-10000, 31, 10000} is also 31.

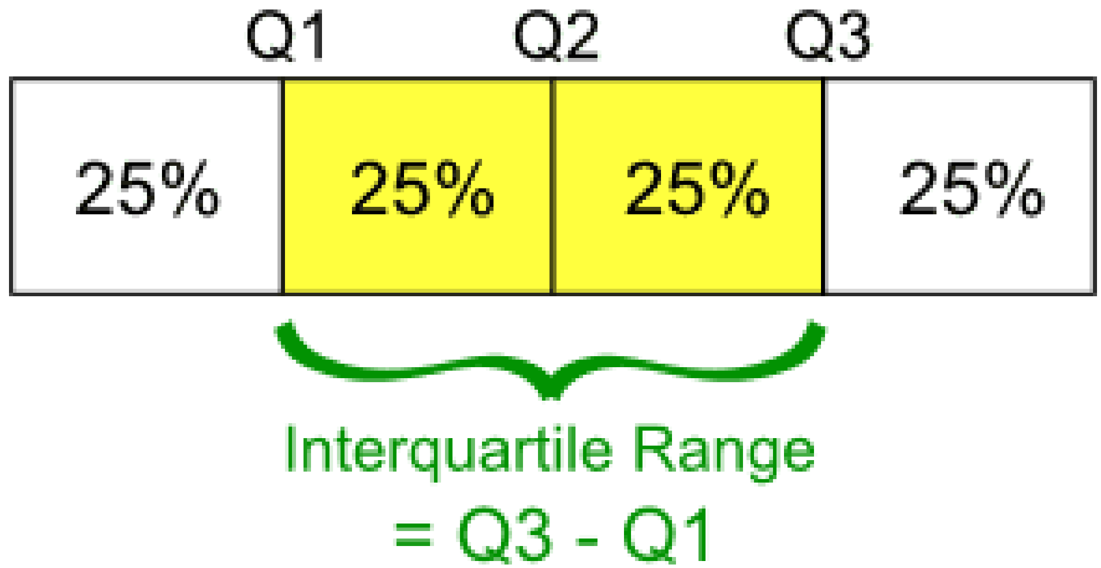
Inter-Quartile Range

- We also need measures of spread, like the Inter-Quartile Range. (Literally “range the between the quartiles”, called the IQR for short).



- The Inter-Quartile range is calculated:

$$IQR = Q3 - Q1$$



- The size of the IQR indicates how spread out the middle half of the data is.

Outliers (1.5 x IQR Rule)

- Now that we have a measure of spread, we can use it to identify values that are much farther from the center than usual.

- How? Spread measures like the IQR tell us how far a typical value could be from the average, so anything much more than the typical distance can be identified.

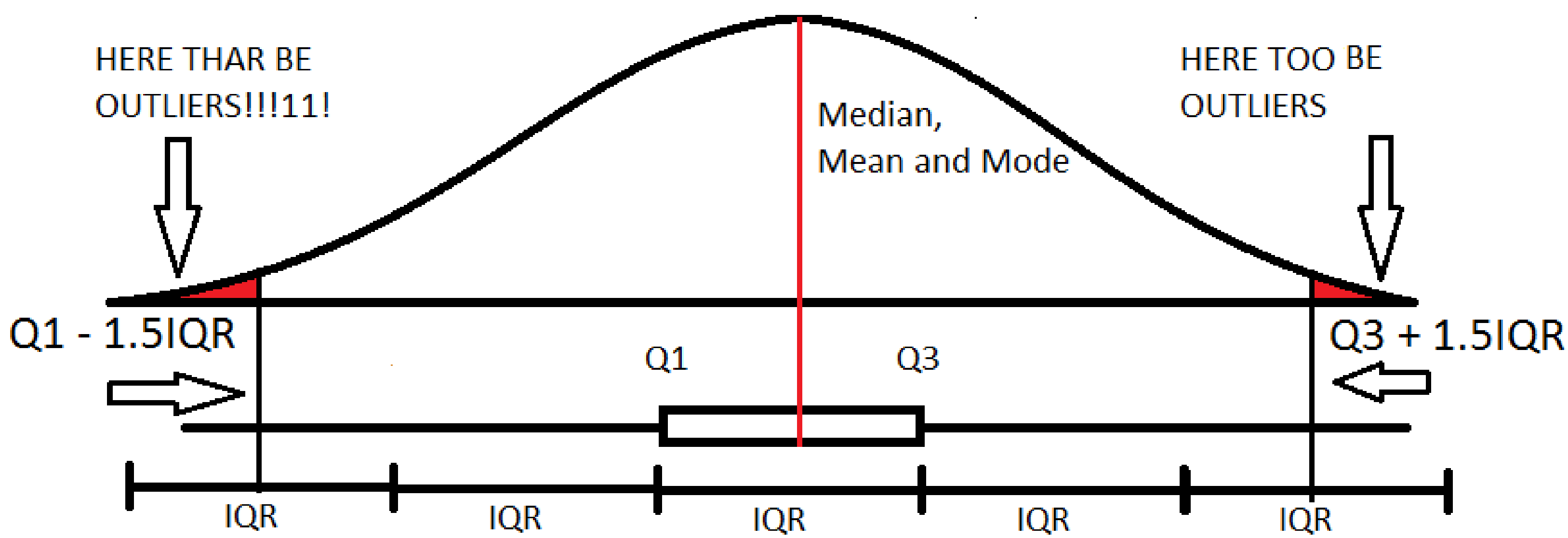
- We call these data points outliers.

  They (figuratively) lay outside the rest of the data.

- Because an outlier stands out from the rest of the data, it…
  - might not belong there, or
  - is worthy of extra attention.

- One way to define an outlier is
  - anything below Q1 – 1.5 IQR or…
  - above Q3 + 1.5 IQR.

This is called the 1.5 x IQR rule. (Important).



- Example: {0, 1, 2, 4, 5, 5, 7, 10, 10, 12, 13, 17, 39}

  Q1 = 3, Q3 = 12.5

  IQR = 12.5 - 3 = 9.5.

  Q1 – 1.5xIQR = 3 – 1.5(9.5)

  = 3 -14.25 = -11.25

  Anything less than -11.25 is an outlier.

  In this case there are no outliers on the low end.

- Example: {0, 1, 2, 4, 5, 5, 7, 10, 10, 12, 13, 17, 39}

  Q1 = 3, Q3 = 12.5

  IQR = 9.5

  Q3 + 1.5xIQR = 12.5 + 1.5*9.5

  = 12.5 + 14.25 = 26.75

  Anything more than 26.75 is an outlier.

  39 is the only outlier.

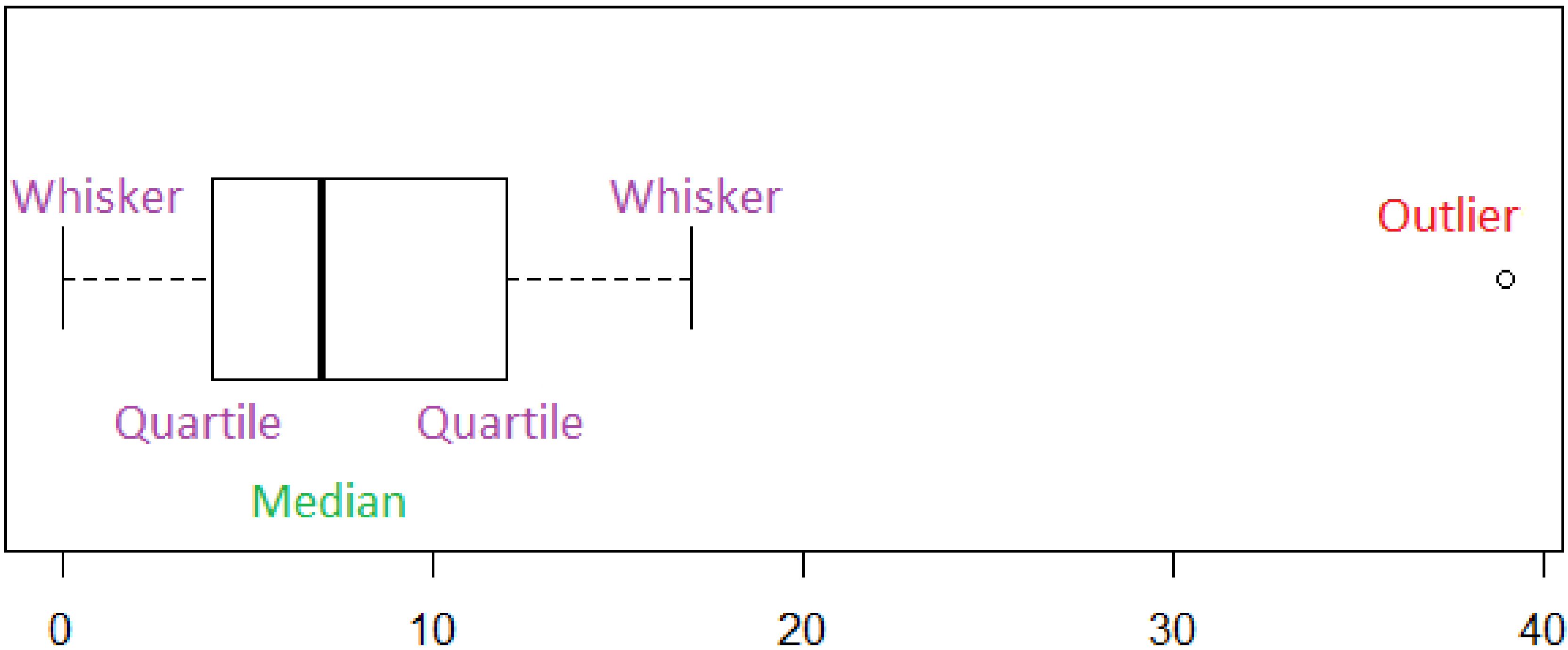
More on IQR and Outliers:

- There are other ways to define outliers, but 1.5xIQR is one of the most straightforward.

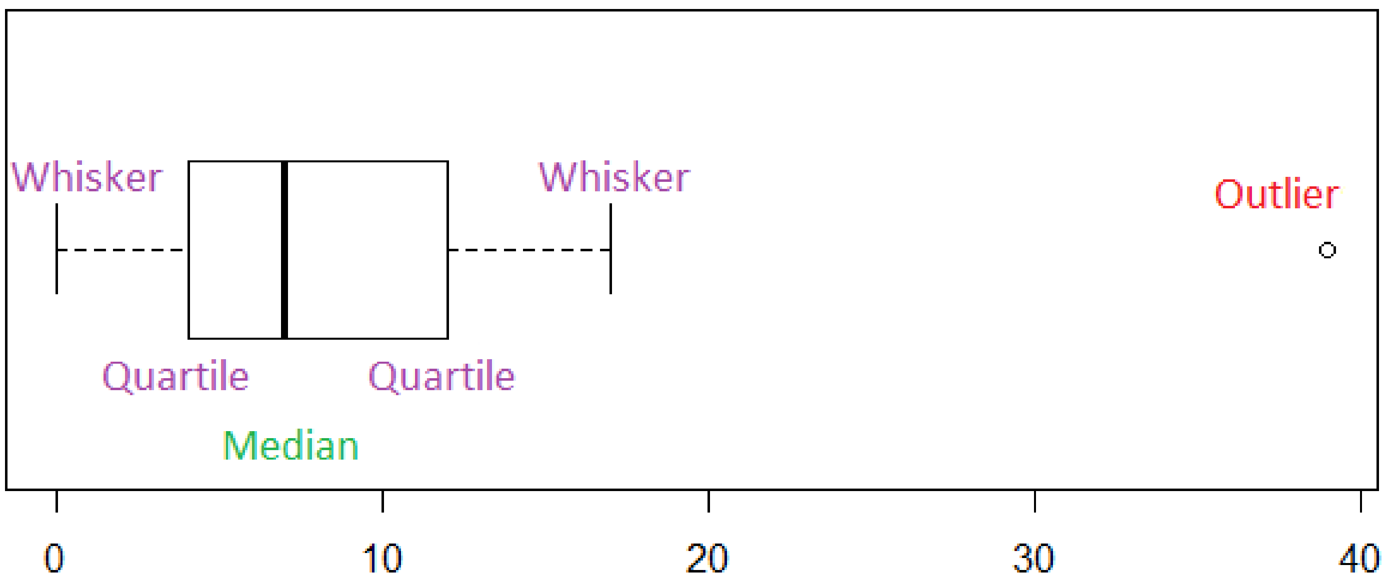
- If our range has a natural restriction, (like it can't possibly be negative), it's okay for an outlier limit to be beyond that restriction.

- If a value is more than Q3 + 3*IQR or less than Q1 – 3*IQR it is sometimes called an extreme outlier.

- The standard graph for showing the median, quartiles, and outliers of a data set is the boxplot, for {0, 1, 2, 4, 5, 5, 7, 10, 10, 12, 13, 17, 39} it looks like this:

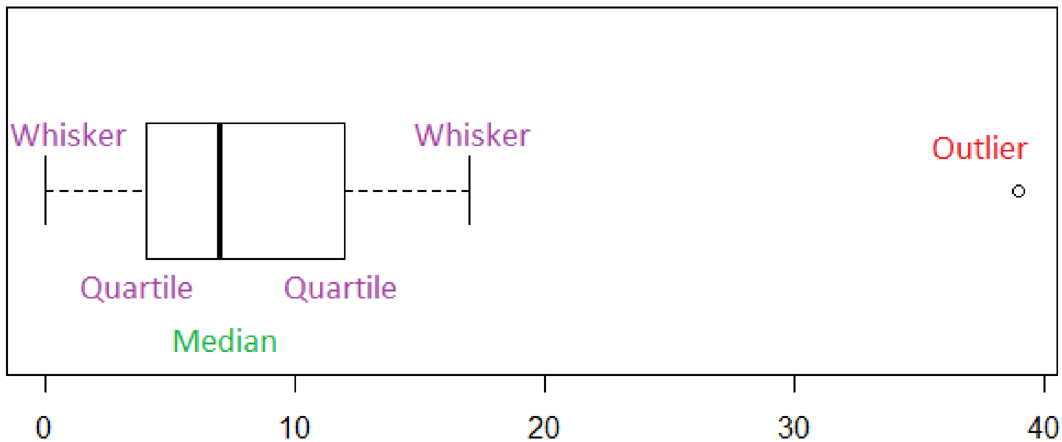


- The five-number summary is in the boxplot:
- The box from 3 to 12.5 is the region between Q1 and Q3.
- The line going through the middle of the box at 7 is the median.



-

- The lines going out the ends of the box are called the whiskers. They show the range of values that are not outliers.



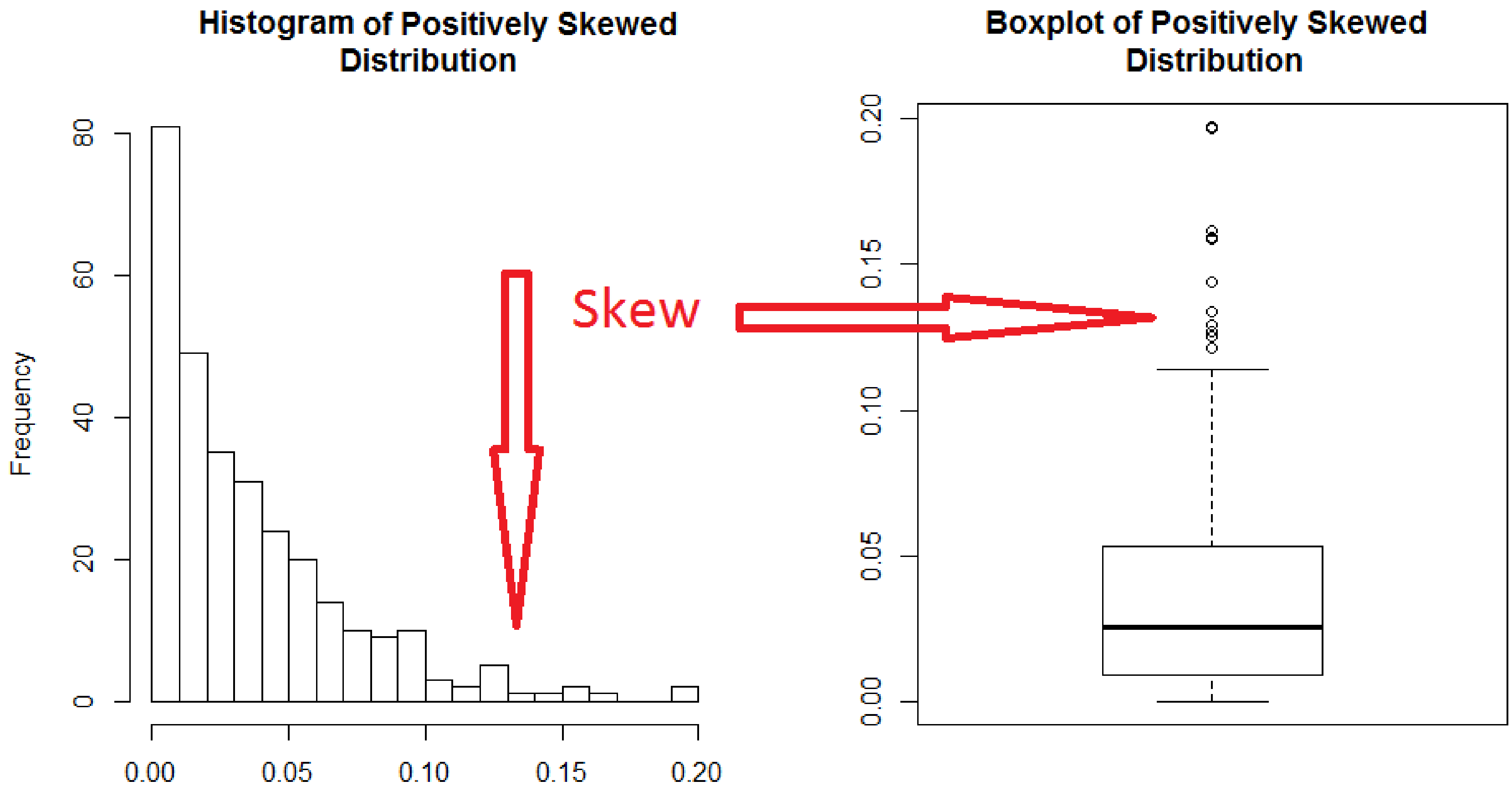
- The lower whisker goes to the lowest value, 1. The upper whisker goes to 17 because it’s the biggest value before the upper limit of 26.75 is hit.

- The individual dot at 39 shows an outlier.
- Outliers in SPSS are labelled with their row number so you can find them in data view.
- In SPSS extreme outliers are shown as stars.
- The farthest outliers on either side are the minimum and maximum.
- If there are no outliers on a side, the end of the whisker is that minimum or maximum.

Boxplots and Skew

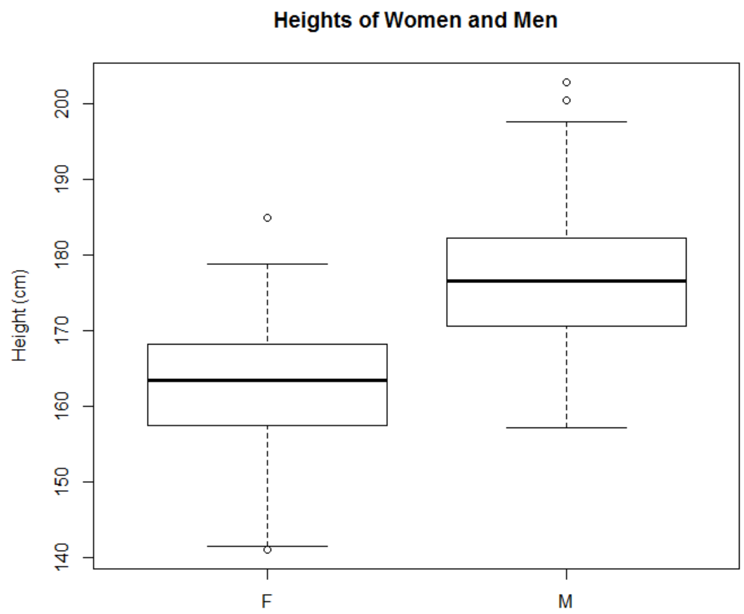
- Skewed distributions have more extreme values on one side, so a boxplot of a skewed distribution will have one whisker longer than the other.

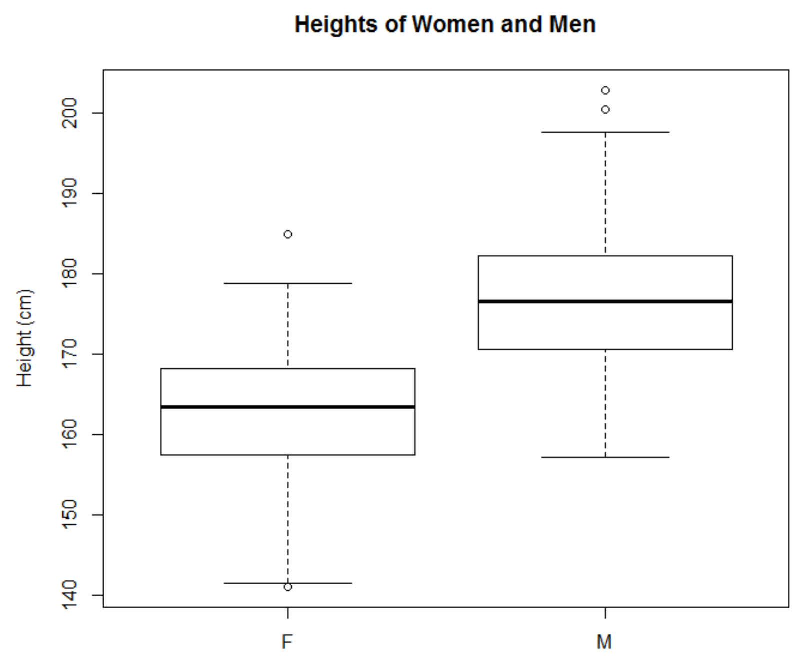
- There will also be more outliers on one side of the boxplot than the other.


Histogram of Positively Skewed Distribution
Frequency
0
20
40
60
80
0.00
0.05
0.10
0.15
0.20
Skew
Boxplot of Positively Skewed Distribution
0.00
0.05
0.10
0.15
0.20

Side-by-side Boxplots

- Boxplots can also be used to compare the distributions of two samples.
- Example: Heights of adult men and women.





- There is some overlap
- In general men are taller.
- The variance is about the same.
- Both distributions appear to be symmetric.

What exactly IS an outlier?

- It’s a value far from anything else that warrants special consideration aside from the rest of the data.

- Often it’s a mistake in data entry. If were recording a grade of 73%, mistyped, and recorded 3% or 730%, both of these values would be far from the rest of the data and would indicate that the data is not being represented properly.

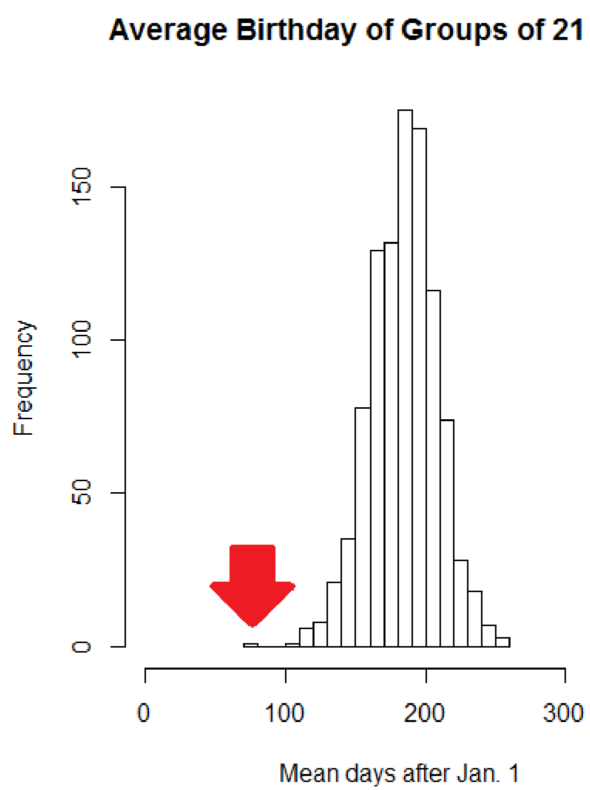
- If the times to finish a final exam had Q1 at 120 minutes and Q3 at 150 minutes, but someone finished in 62 minutes, that person could be a student with a stronger than recommended background for that course or someone who gave up during the exam.

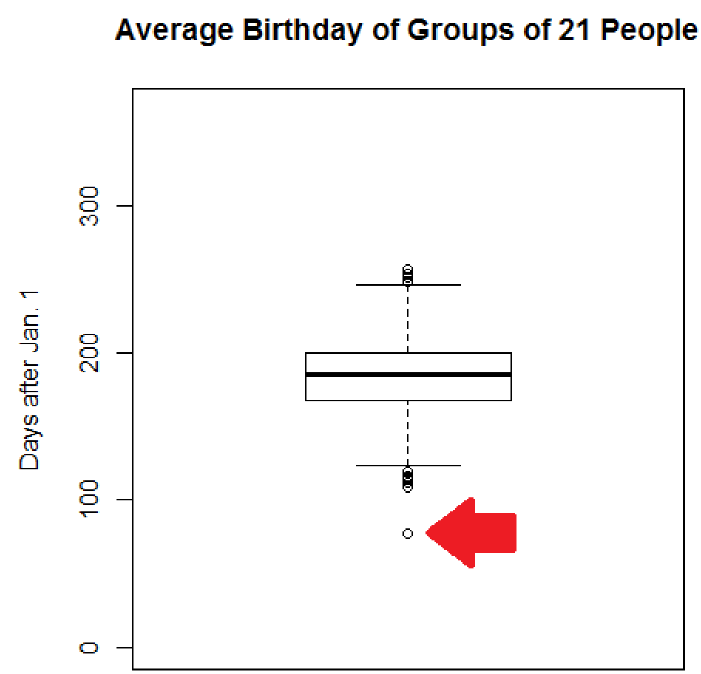
- In both cases, their exam wouldn't a good representation of the exams as whole.
- Sometimes outliers can tell your assumptions and expectations are wrong, like in minor hockey.

Minor Hockey and Outliers (See Malcolm Gladwell's Outliers) (this is for interest)

- If I took samples of 25 random people over and over and got their 'average' birthdays, I would get a bell curve around late June or early July. (Using days since January 1 as the value of people's birthdays) This would fit my assumption that the way I took my sample had nothing to do with when people were born, so the average birthday should be right near the middle of the year.

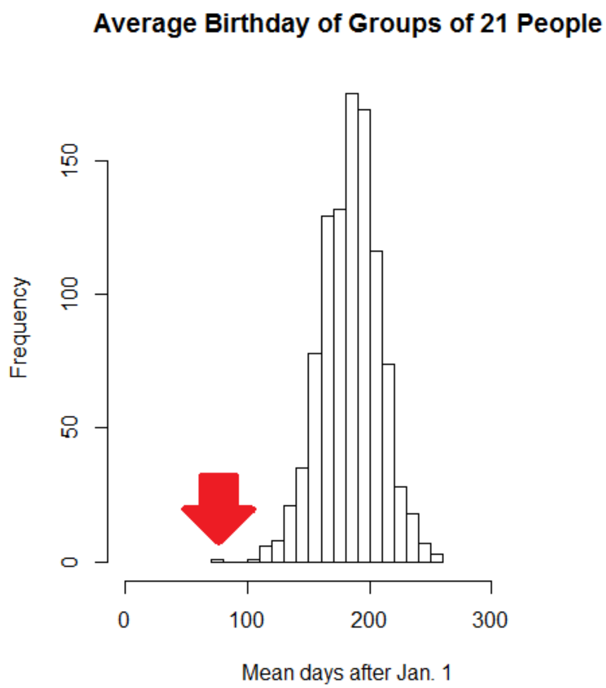
- If my sample is Team Canada in the World Junior Hockey Championship, however, the average is much closer to the beginning of the year.

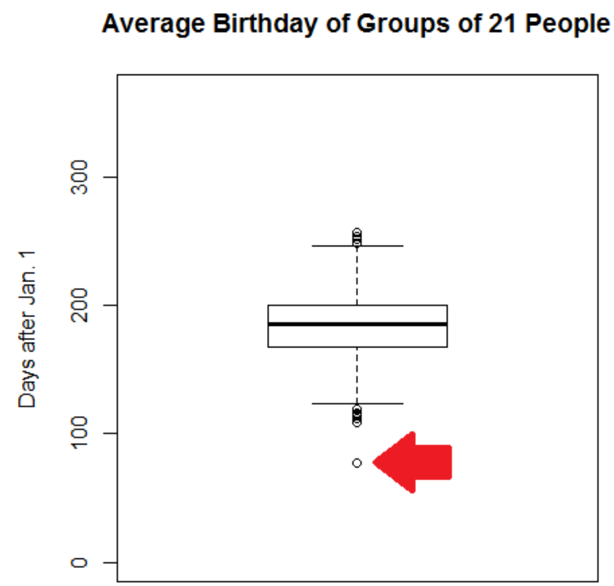




Indicates Junior Hockey Team Canada 77.09 Days (March 8) on Average

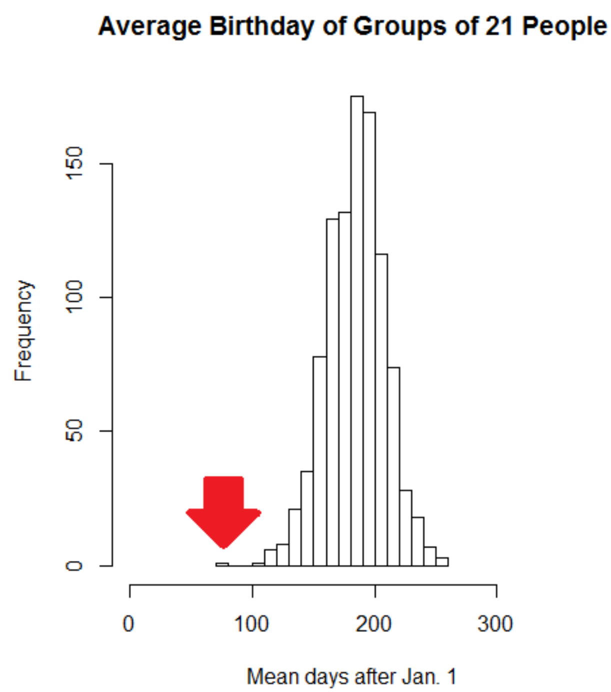
- The average birthday of champion hockey players is an outlier compared to the average of other groups of 25. The data was entered correctly, so the population must be different somehow.

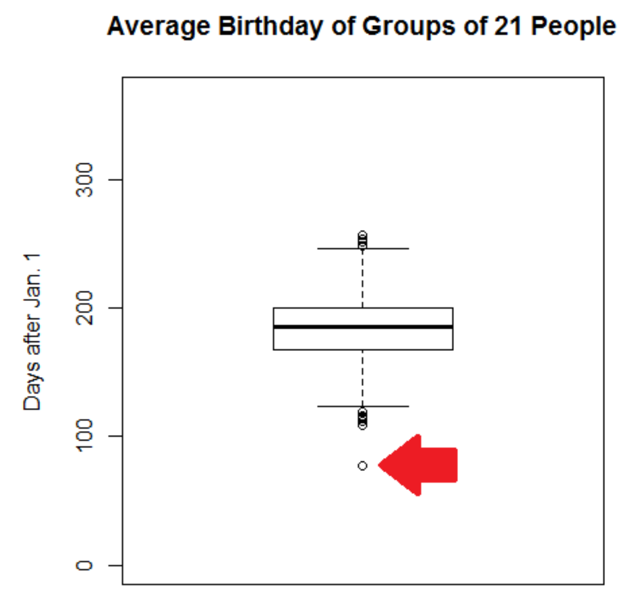




Indicates Junior Hockey Team Canada 77.09 Days (March 8) on Average

It’s possible this happened by chance, but unlikely in the context of the other samples. My assumption that birthdays and being a hockey champion are unrelated may be wrong.





Indicates Junior Hockey Team Canada 77.09 Days (March 8) on Average

Next lecture:

Standard Deviation and the Normal Curve
(read more of chapter 4)