Chapter 5 Summary
Describing Distributions Numerically

What did you learn?
Distributions of quantitative variables can be summarized numerically.
- The 5-number summary displays the two quartiles, the median, and the two extremes for a variable.
- Measures of center include mean and median.
- Measures of spread include range, IQR, and standard deviation.
- When the distribution is skewed, we report the median and the IQR. When the distribution is symmetric, we report the mean and the standard deviation (and possibly the median and the IQR as well).

Data distributions can be displayed using boxplots.
- A boxplot reveals some features of a distribution not easily seen in a histogram – the center, the middle 50%, and outliers. Histograms are better at showing the shape of the distribution.
- Boxplots are effective for comparing groups graphically. When discussing group comparisons, we discuss their shape, center, spread, and unusual features.

n always refers to the number of data values.

Typical values for a data set are usually the center value

Center	We summarize the center of a distribution with the mean or the median.
Mean	The mean is found by summing all the data values and dividing by the count.
Midrange	The mean of the minimum and maximum values of a set of data.
Median	The median is the middle value of an organized data set, with half of the data above and half below it.

Spread	We summarize the spread of a distribution with the standard deviation, interquartile range, and range.
Range	The difference between the lowest and highest values in a data set (maximum - minimum).

Quartiles	The median and the quartiles divide data into four equal parts.
Lower Quartile	The lower quartile (Q1) is the value with a quarter of the data below it (the median of the lower half of the data set).
Upper Quartile	The upper quartile (Q3) is the value with the a quarter of the data above it (the median of the upper half of the data set).
Interquartile range (IQR)	The difference between the first and third quartile (IQR = Q3 - Q1).
Percentiles	The ith percentile is the number that falls above i% of the data.

5-number summary	A 5-number summary for a variable consists of the minimum, the lower quartile, the median, the upper quartile, and the maximum.
Boxplot	A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Boxplots are particularly effective for comparing groups of different sizes.

Deviations are distances between an individual data point and the 'center'

y represent individual data of n data values and M represents the middle, so

$\sum (y - M)$ represents the difference between each datum and the center

We want the deviations to be 0, so

$\sum (y - M) = 0$ becomes $\sum y = \sum M$ which becomes

$\sum y = M + M + M + \cdots + M = nM$ and $M = \dfrac{\sum y}{n}$, or the mean

The mean for statistics is referred to as $\bar{y}$ and pronounced "y=bar"

Mean is the balancing point of a histogram

For skewed data, it is better to report the median than the mean as the measure of center

Standard deviation measures the distance each value is from the center and should only be used with symmetric data

Standard deviation	The standard deviation is the square root of the variance.
Variance	The variance is the sum of squared deviations from the mean, divided by the count minus one.

In order for the distances from center to not cancel each other out due to positive and negative differences, we square the difference

Variance is $s^2 = \dfrac{\sum \left(y - \bar{y} \right)^2}{n-1}$ Standard deviation is $s = \sqrt{\dfrac{\sum \left(y - \bar{y} \right)^2}{n-1}}$

What can go wrong?

- Don't forget to do a reality check.
- Don't forget to sort the values before finding the median or percentiles.
- Don't compute numerical summaries of a categorical variable.
- Watch out for multiple modes.
- Be aware of slightly different methods.
- Beware of outliers.
- Make a picture.
- Be careful when comparing groups that have different spreads.