1. Exploring Data

1.1 Analyzing Categorical Data

Pie charts and bar graphs display the distribution of a categorical variable.

Bar Graphs can also compare any set of quantities measured in the same units.

1. Exploring Data

1.1 Analyzing Categorical Data

A two-way table of counts organizes data about two categorical variables measured for the same set of individuals.

There is an association between two variables if knowing the value of one variable helps predict the value of the other.

1. Exploring Data

1.2 Displaying Quantitative Data with Graphs

You can use a dotplot, stemplot, or histogram to show the distribution of a quantitative variable.

Remember: histograms are for quantitative data; bar graphs are for categorical data.

1. Exploring Data

1.2 Displaying Quantitative Data with Graphs

Shape, **center**, and **spread** describe the overall pattern of the distribution of a quantitative variable.

Outliers are observations that lie **outside** the overall pattern of a distribution.

1. Exploring Data

1.2 Displaying Quantitative Data with Graphs

Some distributions have simple shapes, such as **symmetric**, **skewed to the left**, or **skewed to the right**.

The other aspect are **the number of modes** (major peaks), **clusters**, and **gaps**.

1. Exploring Data

1.3 Describing Quantitative Data with Numbers

The **mean** is the **average** of the observations, and the **median** is the **midpoint** of the values.

The **median** is a **resistant** measure of center because it is relatively unaffected by extreme observations. The **mean** is **nonresistant**.

1. Exploring Data

1.3 Describing Quantitative Data with Numbers

The **five-number summary** consisting of the **median**, the **quartiles**(Q_1,Q_3), and the **maximum** and **minimum** values provides a quick overall description of a distribution.

Boxplots based on the **five-number summary** are useful for comparing distributions.

1. Exploring Data

1.3 Describing Quantitative Data with Numbers

The **interquartile range (IQR)** is **the range of the middle 50%** of the observations and is found by $IQR = Q_3 - Q_1$.

An extreme observation is an **outlier** if it is **smaller than** $Q1 - (1.5 \times IQR)$ or **larger than** $Q3 + (1.5 \times IQR)$.

1. Exploring Data

1.3 Describing Quantitative Data with Numbers

The **variance** s_x^2 and especially its square root, the **standard deviation** s_x, are common measures of **spread** about the mean.

The **standard deviation** s_x is **zero** when there is **no variability** and gets larger as the spread increases.

1. Exploring Data

1.3 Describing Quantitative Data with Numbers

The **mean** and **standard deviation** are good descriptions for **roughly symmetric distributions without outliers**.

The **median** and **IQR** are a better description for **skewed distributions**.