Lecture 5: Visualizing Numerical Data

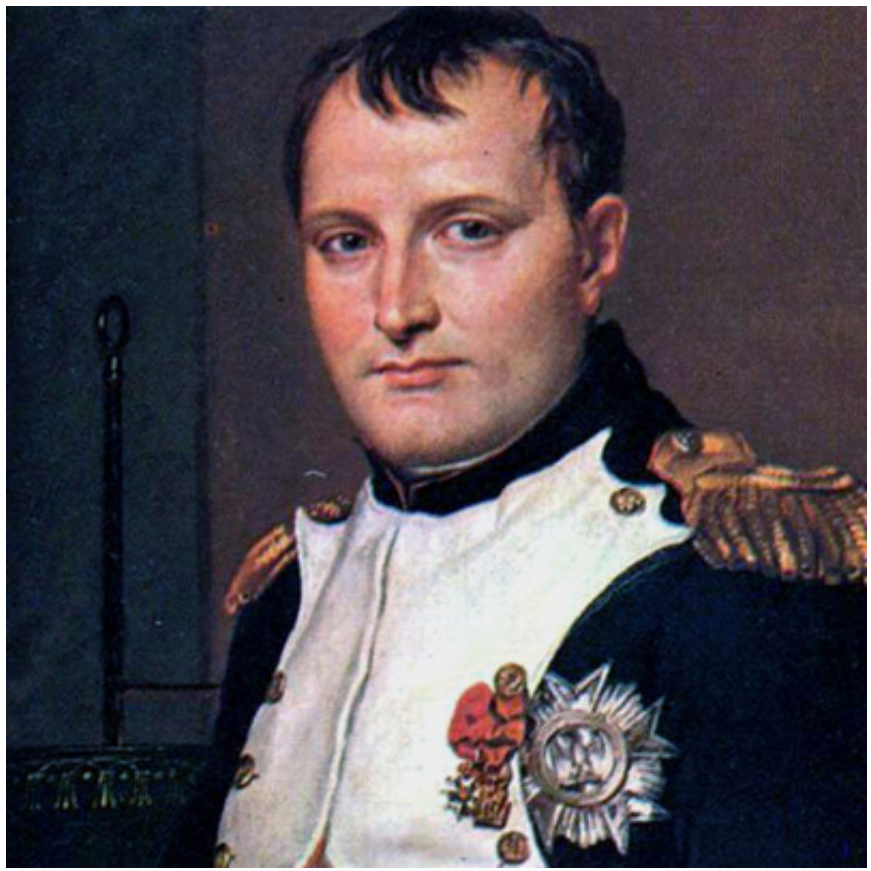
Chapter 1.6 + 1.7

Goals for Today

- Visualizing numerical data
- Histograms
- Measures of Central Tendency: Mean, Median, and Mode
- Measure of Spread: Sample variance and sample standard deviation

Famous Example 1: Napoleon's March on Russia in 1812

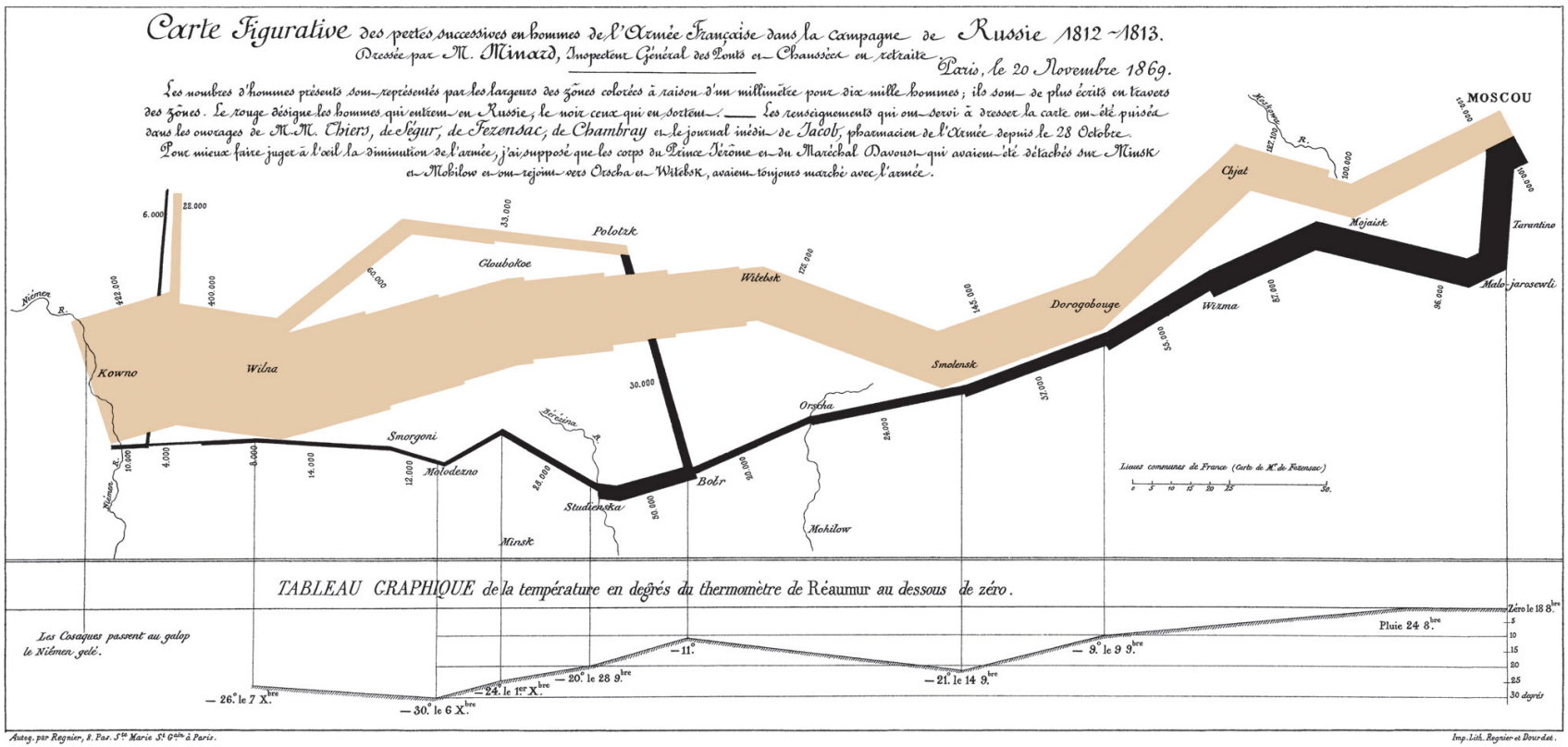
In 1812, Napoleon led a French invasion of Russia, at one point marching on Moscow.



Famous Example 1: Napoleon's March on Russia in 1812

The advance and retreat on Moscow was an unmitigated disaster:

Famous Example 1: Napoleon's March on Russia in 1812



Famous Example 2: 1854 Broad Street Cholera Outbreak

On August 31 1854, an epidemic of cholera began in the Soho neighborhood of London. Over the next three days 127 people near Broad Street had died.

Famous Example 2: 1854 Broad Street Cholera Outbreak

On August 31 1854, an epidemic of cholera began in the Soho neighborhood of London. Over the next three days 127 people near Broad Street had died.

(Wikipedia) Dr. John Snow was skeptical of the then-dominant miasma theory that diseases like cholera/plague were caused by "bad air."

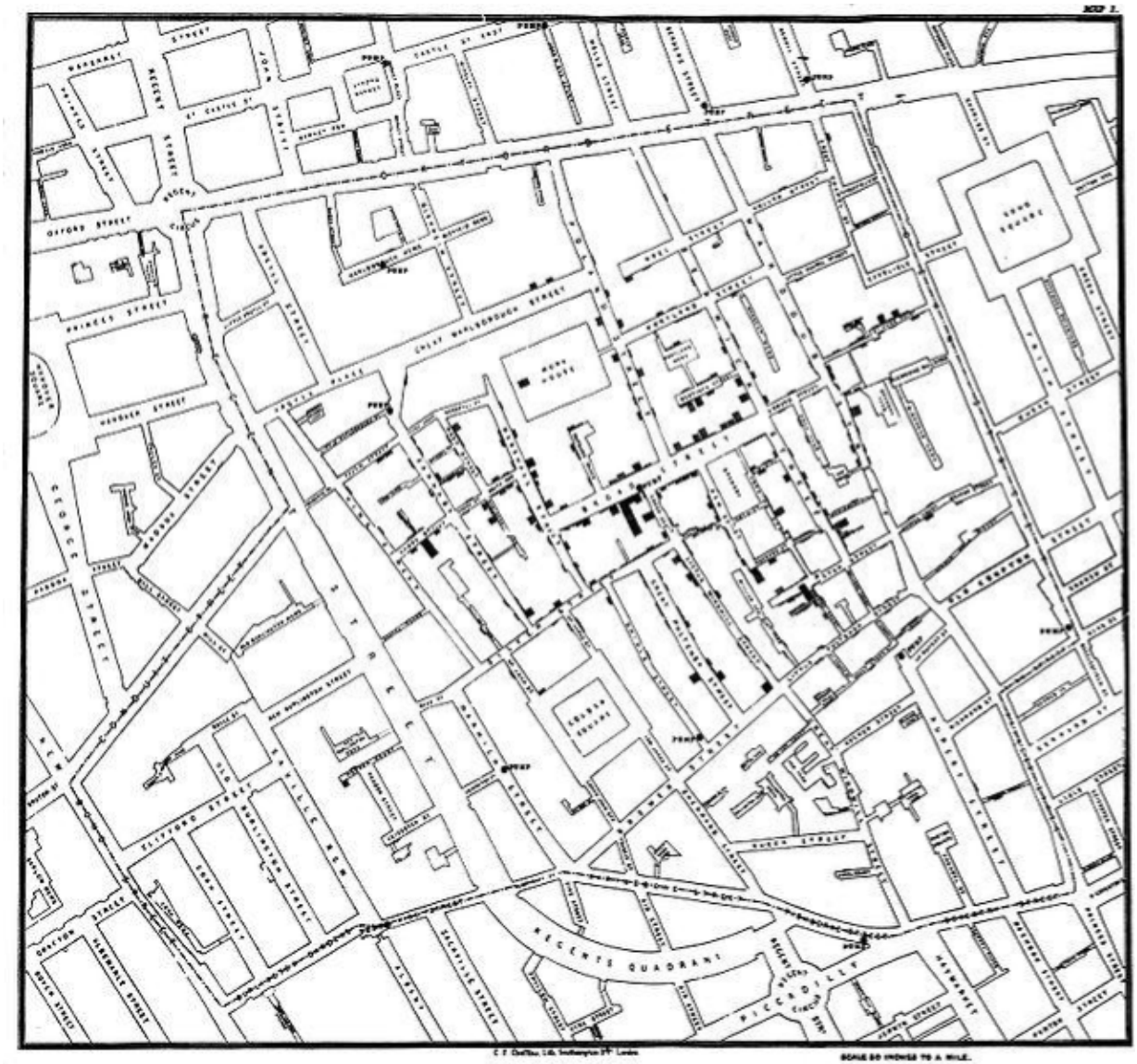
Famous Example 2: 1854 Broad Street Cholera Outbreak

On August 31 1854, an epidemic of cholera began in the Soho neighborhood of London. Over the next three days 127 people near Broad Street had died.

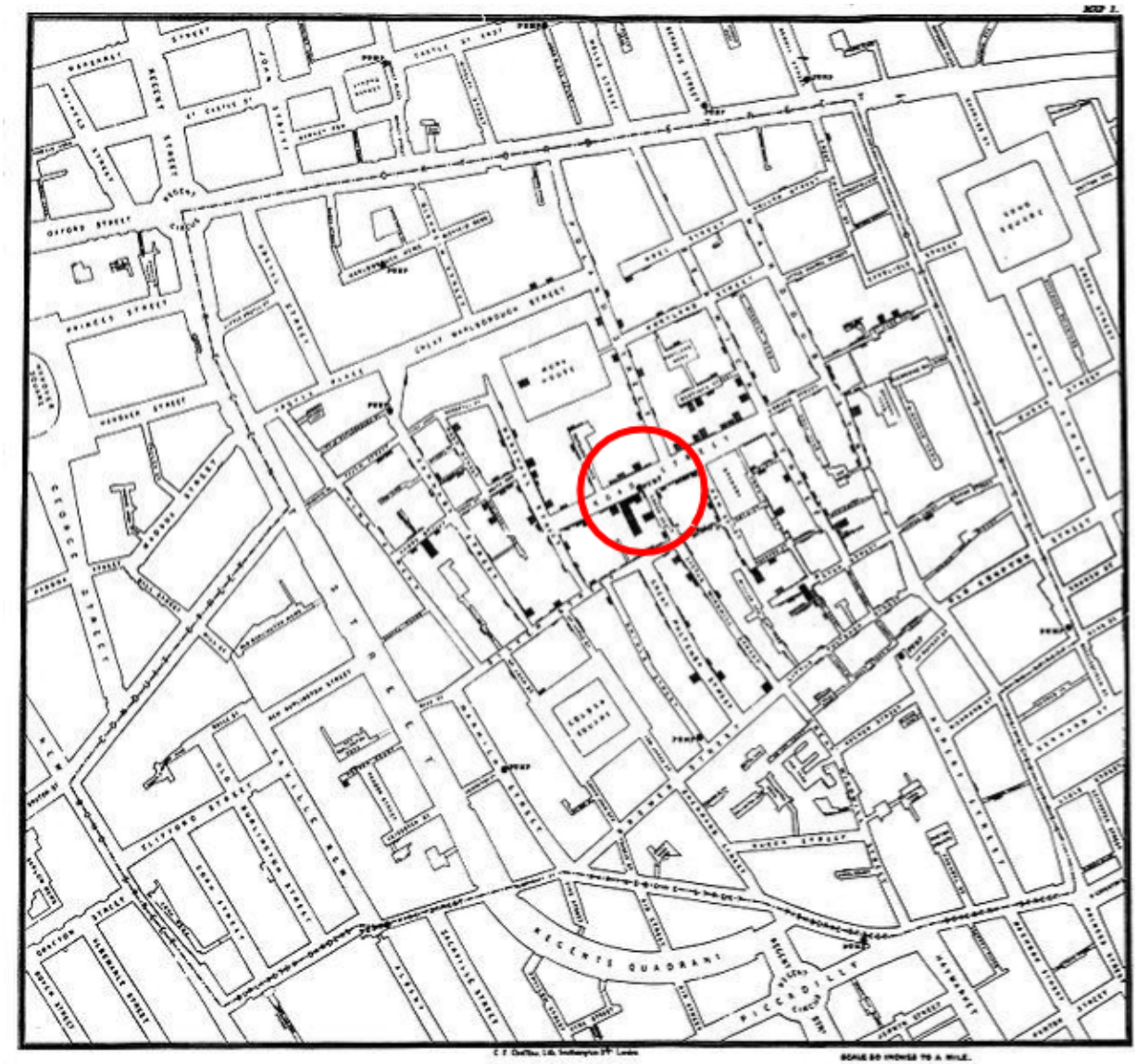
(Wikipedia) Dr. John Snow was skeptical of the then-dominant miasma theory that diseases like cholera/plague were caused by "bad air."

Snow created the following map to investigate:

Famous Example 2: 1854 Broad Street Cholera Outbreak



Famous Example 2: 1854 Broad Street Cholera Outbreak



Famous Example 2: 1854 Broad Street Cholera Outbreak

He identified the source of the outbreak as water from the Broad Street Pump, which was near a cesspit that began to leak.

Famous Example 2: 1854 Broad Street Cholera Outbreak

He identified the source of the outbreak as water from the Broad Street Pump, which was near a cesspit that began to leak.

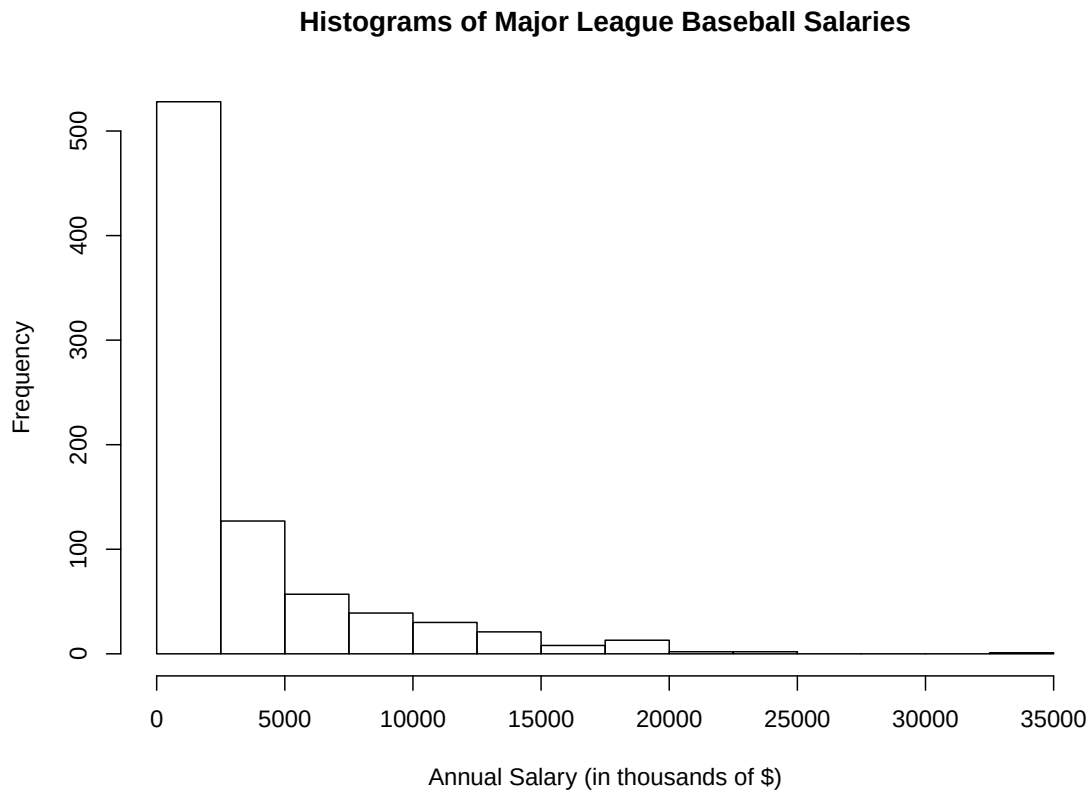
This led to the discovery that cholera was transmitted by food and water being contaminated by fecal matter and not via the air. This was a watershed moment in the emerging field of epidemiology.

Histograms

http://rpubs.com/moonstomper/histograms

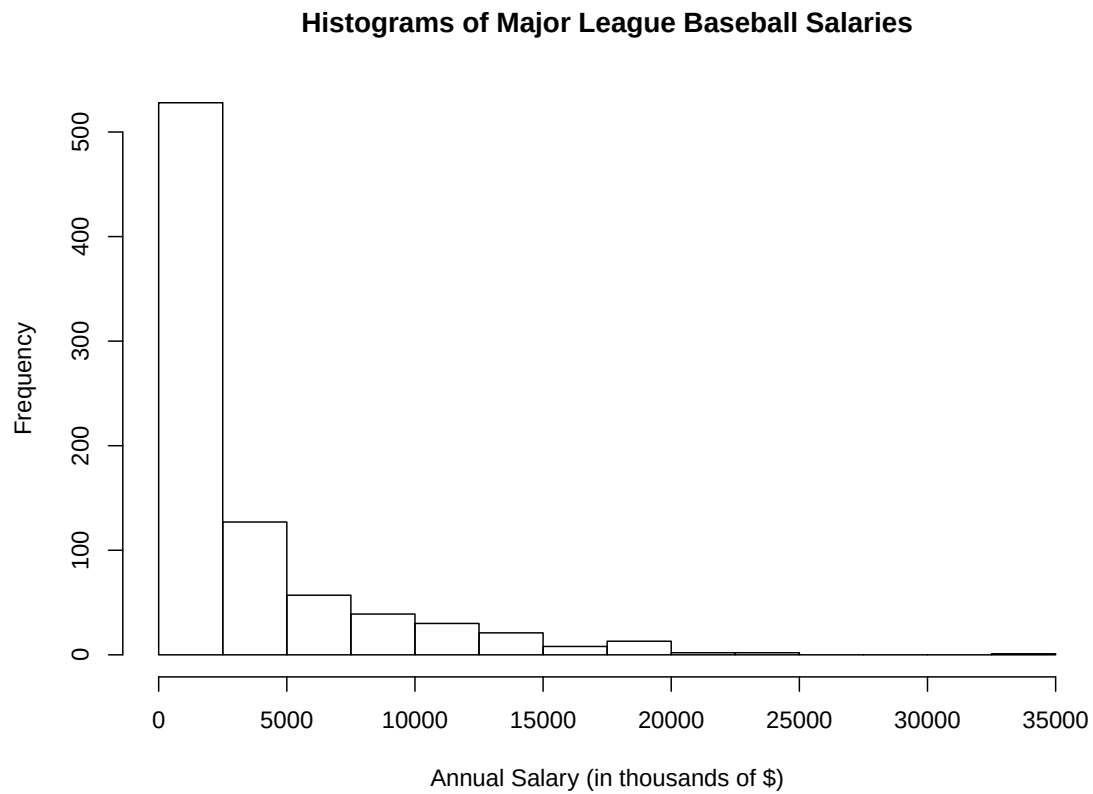
Skew and Long Tail

In the `openintro` package is MLB salary data in 2010. The histogram:



Skew and Long Tail

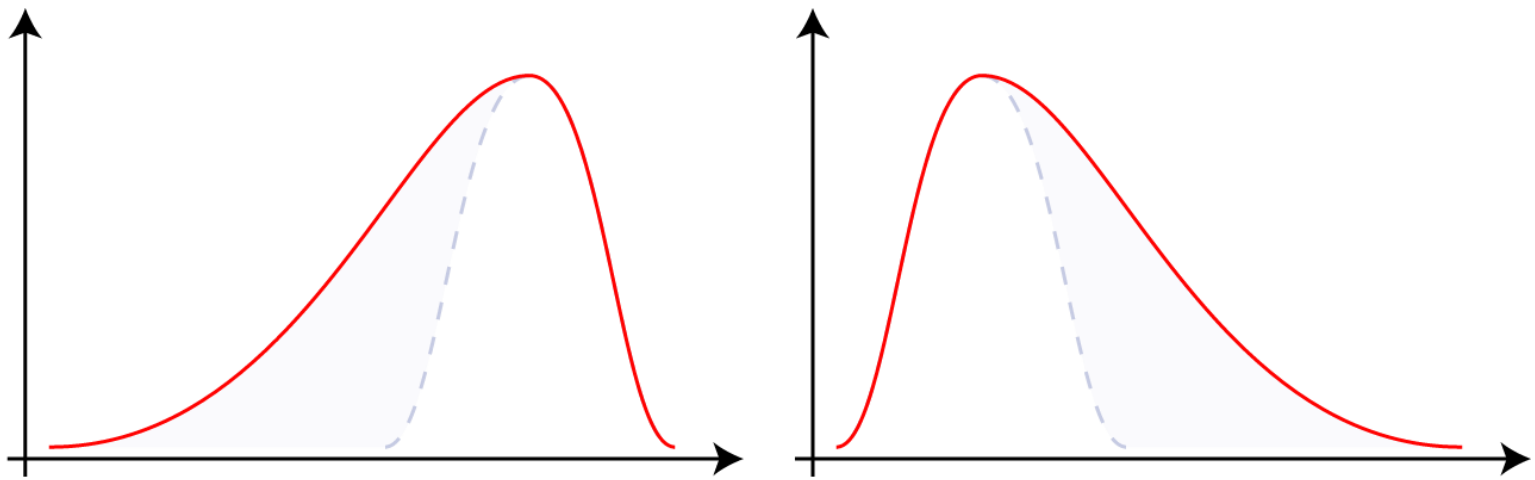
In the `openintro` package is MLB salary data in 2010. The histogram:



The data has a long tail to the right: data is right-skewed. i.e. a small number of players who make a VERY large amount of money.

Trick to Remembering Which Skew is Which

Trick to Remembering Which Skew is Which



Mean

Median

Mean vs Median: Imaginary Scenario

But why use the median at all?

- Say at company X, there 5 employees: the CEO and everyone else.

Mean vs Median: Imaginary Scenario

But why use the median at all?

- Say at company X, there 5 employees: the CEO and everyone else.
- The CEO earns \$1000 an hour, while the others earn \$20, \$21, \$30, and \$40 an hour.

Mean vs Median: Imaginary Scenario

But why use the median at all?

- Say at company X, there 5 employees: the CEO and everyone else.
- The CEO earns \$1000 an hour, while the others earn \$20, \$21, \$30, and \$40 an hour.
- The employees complain that they are paid too little.

Mean vs Median: Imaginary Scenario

But why use the median at all?

- Say at company X, there 5 employees: the CEO and everyone else.
- The CEO earns \$1000 an hour, while the others earn \$20, \$21, \$30, and \$40 an hour.
- The employees complain that they are paid too little.
- The CEO counters that the mean hourly salary is $\overline{x} = \frac{20+21+30+40+1000}{5} = 222.20$ an hour, which is really high.

Mean vs Median: Imaginary Scenario

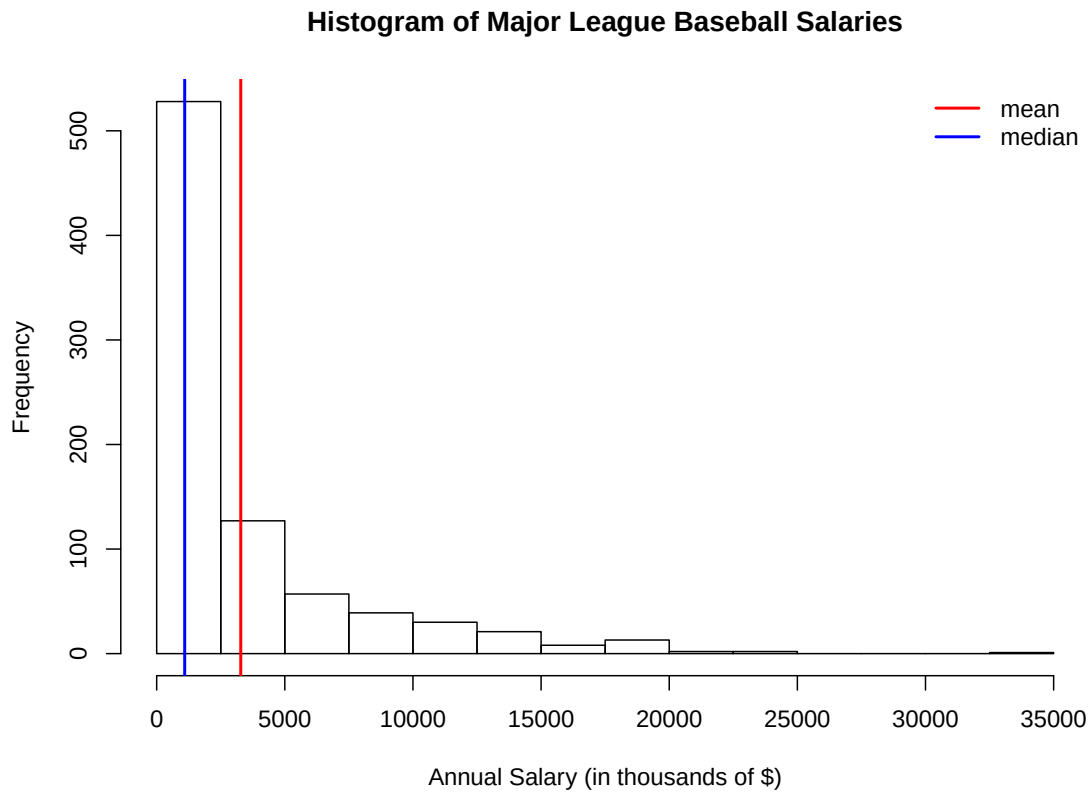
The CEO's extreme salary is inflating the mean. A more appropriate measure is the median hourly salary of 30.

Mean vs Median: Imaginary Scenario

The CEO's extreme salary is inflating the mean. A more appropriate measure is the median hourly salary of 30.

Ex: http://www.zillow.com/home-values/

Mean vs Median: Back to MLB Salary Data



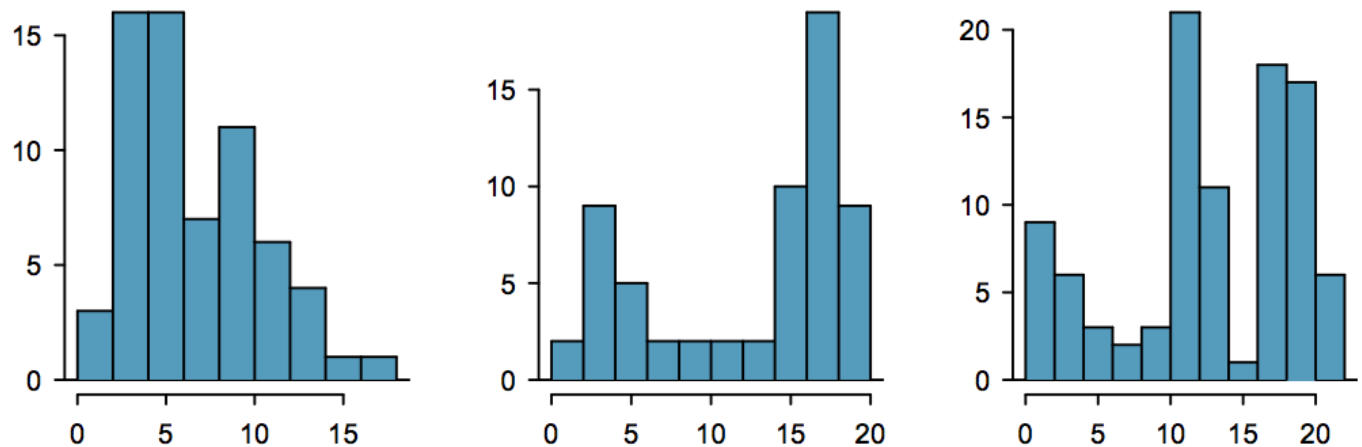
Mode

A mode is the value that appears the most often in a data set. So out of (1, 3, 3, 5, 6), the modal value is 3.

Mode

A mode is the value that appears the most often in a data set. So out of (1, 3, 3, 5, 6), the modal value is 3.

Modes also describe peaks, but this can get subjective. A distribution can be unimodal, bimodal, or multimodal:

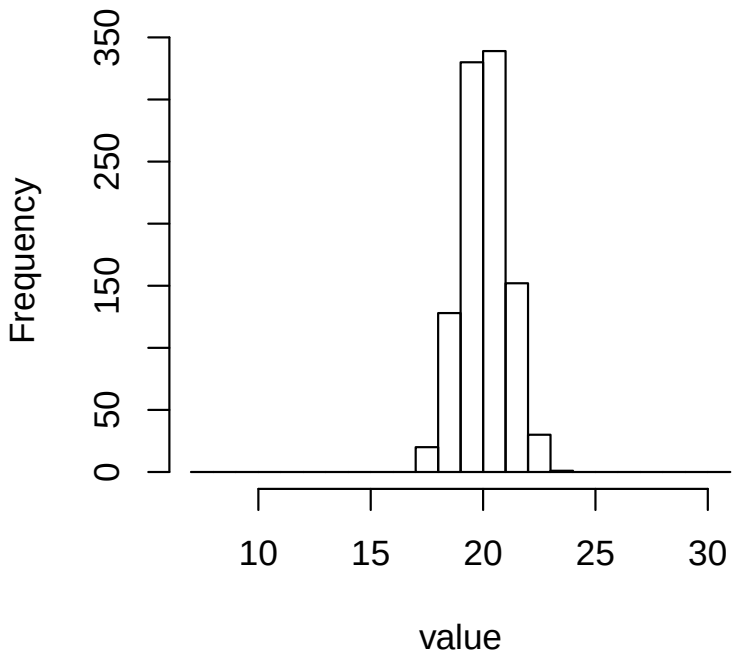


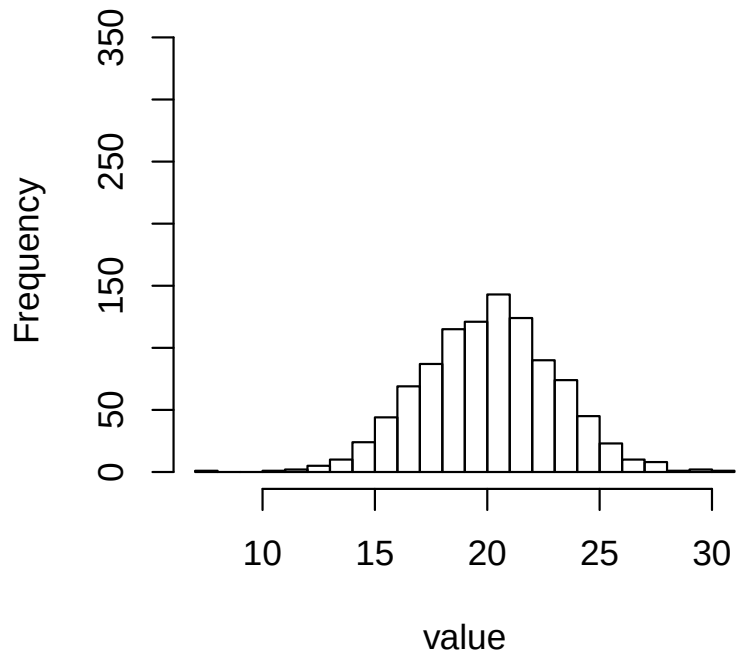
Measure of Spread

Next, consider the following two histograms: Both have mean of about 20. What is the difference between them?

Measure of Spread

Next, consider the following two histograms: Both have mean of about 20. What is the difference between them?

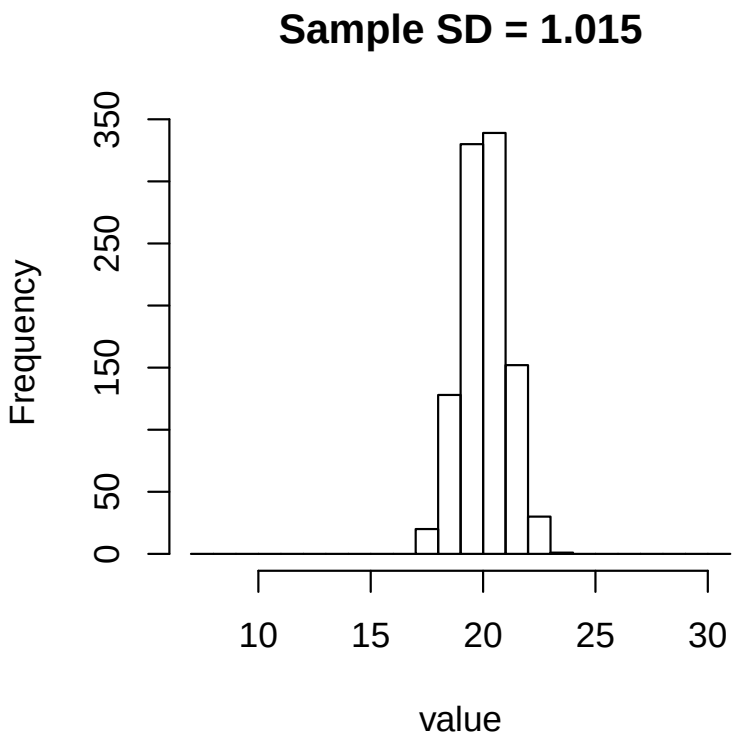


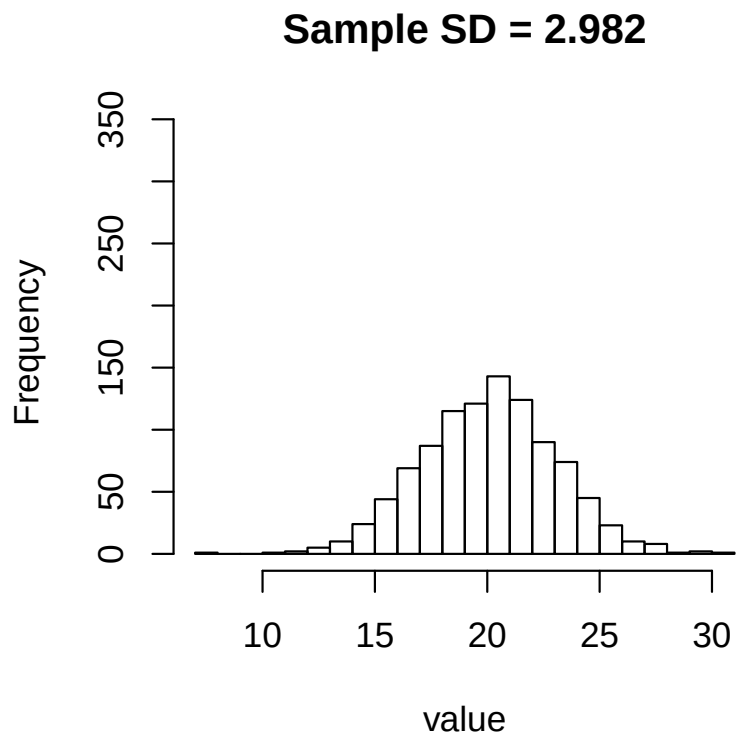


Measure of Spread

Measure of Spread

Back to example:





How to Compute the Sample Standard Deviation

Read section 1.6.4. The formula really doesn't make much intuitive sense, but is the way it is due to mathematical convenience. Fortunately there is an `R` command: `sd()`

Next Time

- Another simple data visualization tool: boxplots
- Examining/Visualizing Categorical Data