Lecture 5: Visualizing Numerical Data

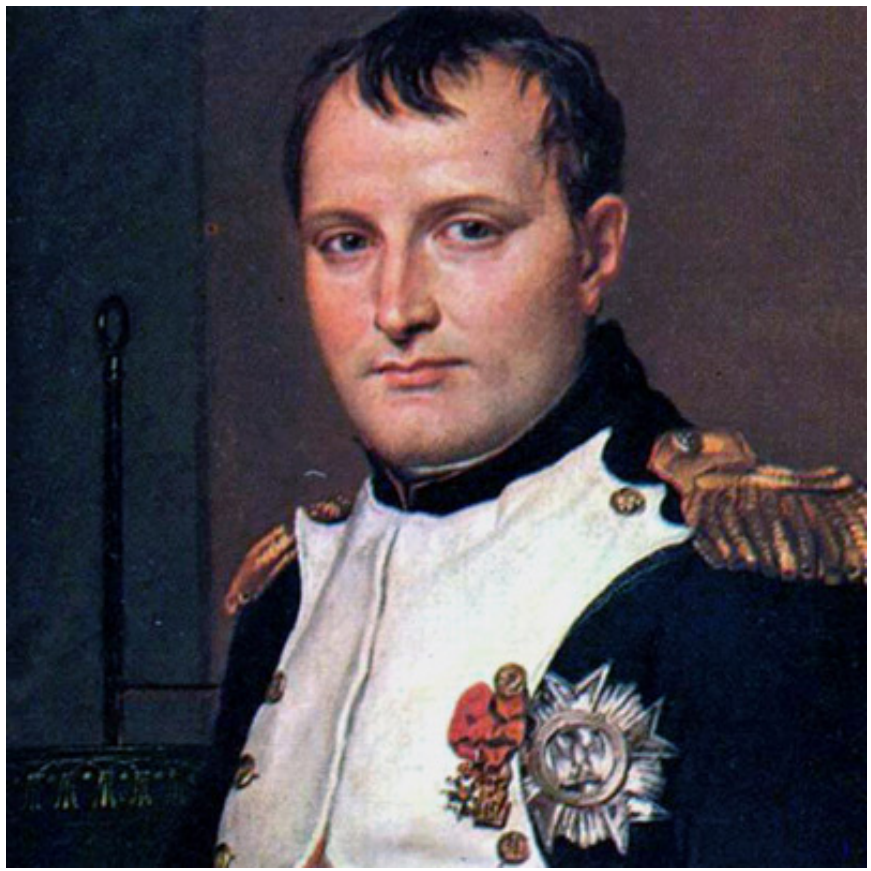
Chapter 1.6 + 1.7

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

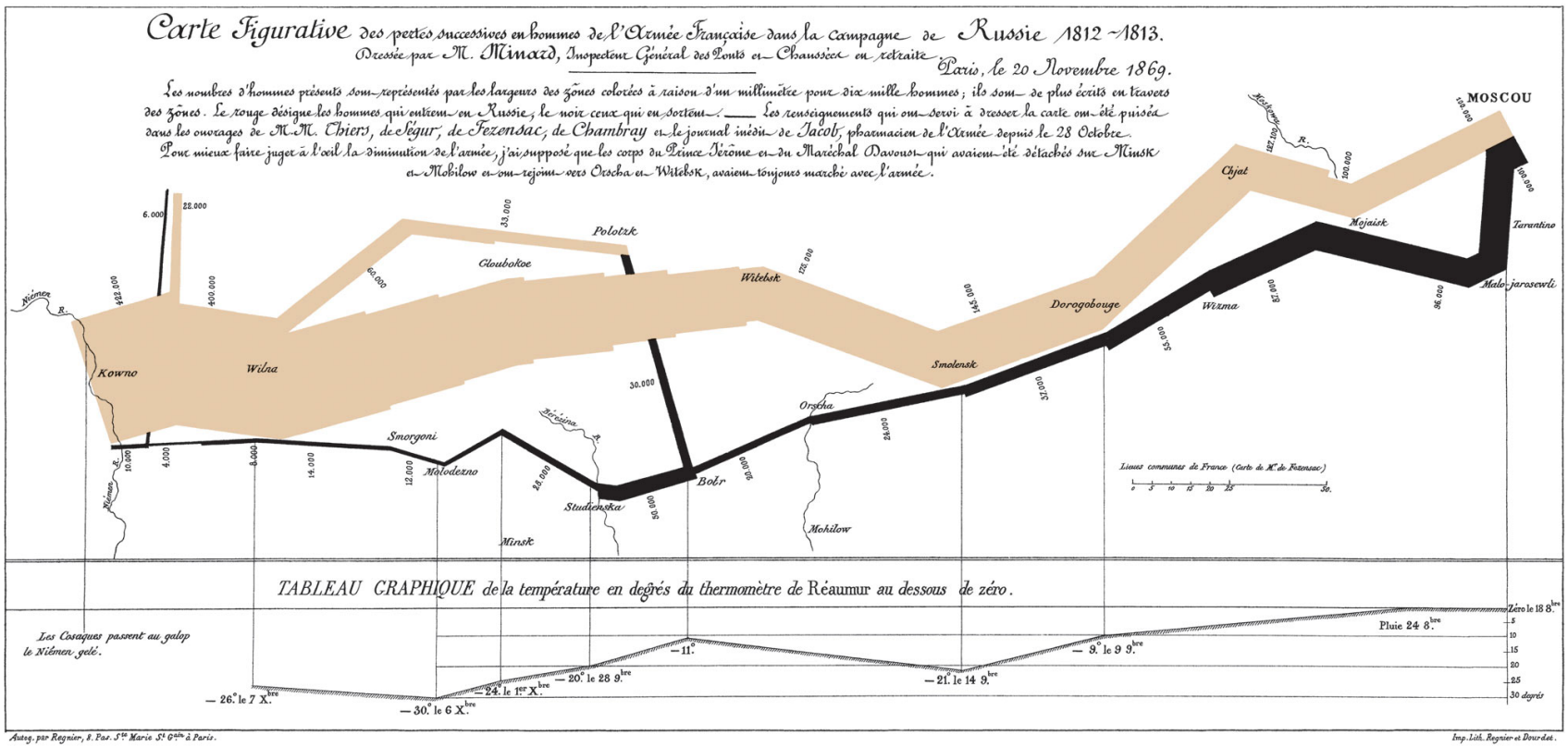
In 1812, Napoleon led a French invasion of Russia, at one point marching on Moscow.



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 1: Napolean's March on Russia in 1812

Why is this visualization big deal? On a two-dimensional page, it displays 6 variables (i.e. dimensions of information) at once:

1. Size of the army (width of bars)
2. Latitude
3. Longitude
4. Direction of the army: advance (brown) or retreat (black)
5. Date
6. Temperature (on the bottom)

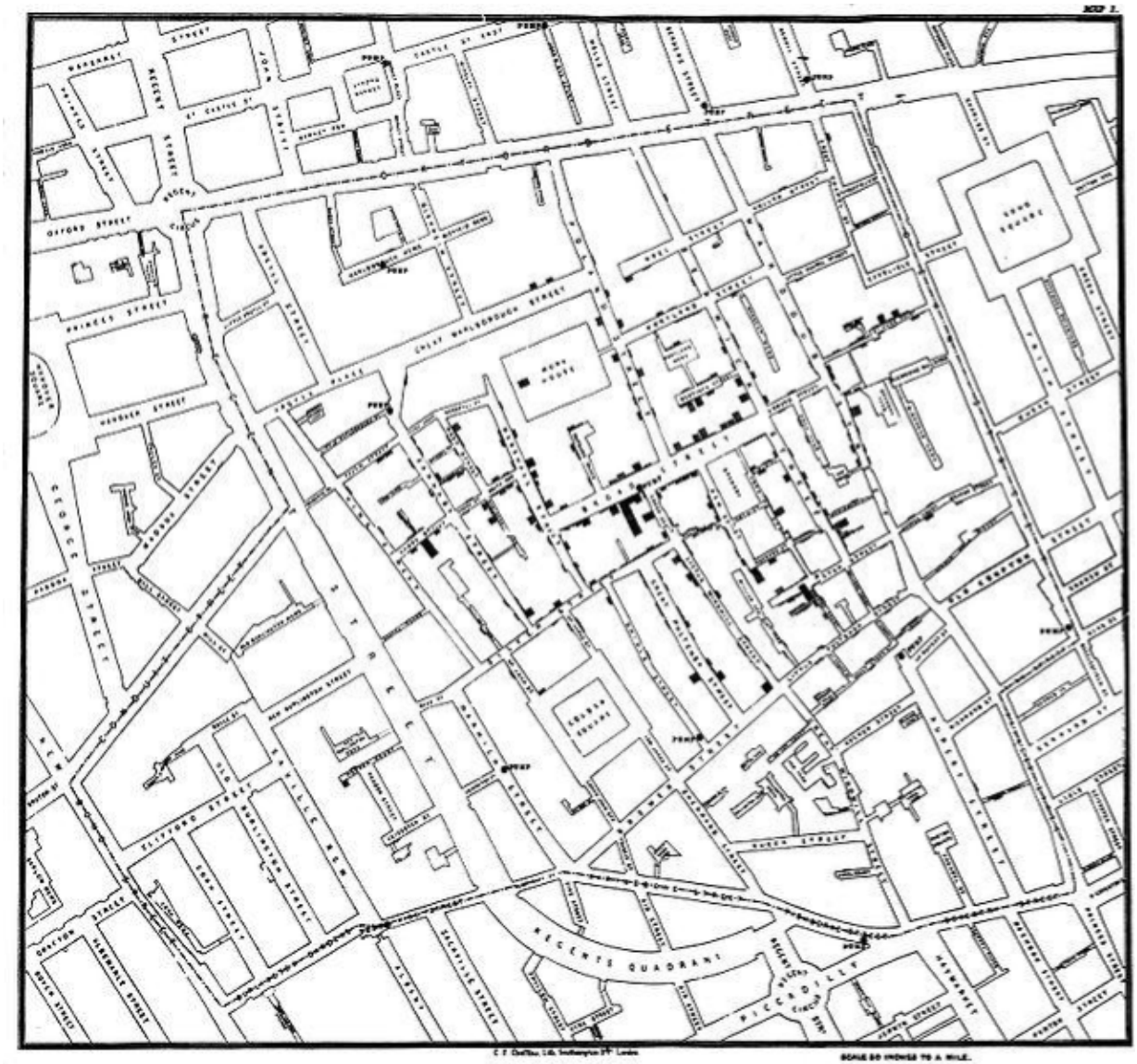
Famous Example 2: 1854 Broad Street Cholera Outbreak

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.

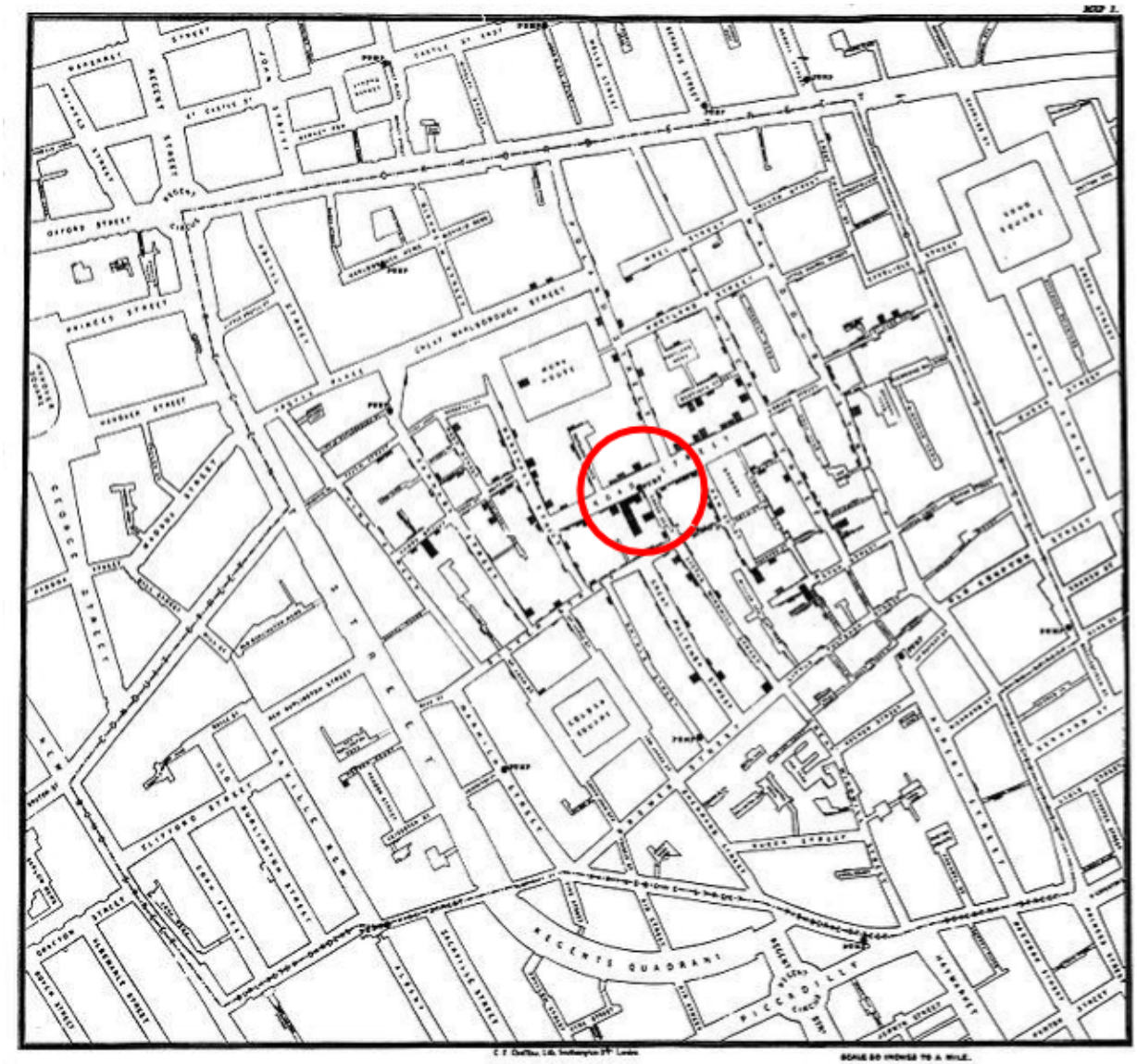
(Wikipedia) Dr. John Snow, a physician, was a skeptic of the then-dominant miasma theory that diseases such as cholera/plague were caused by pollution or a noxious form of "bad air."

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.

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

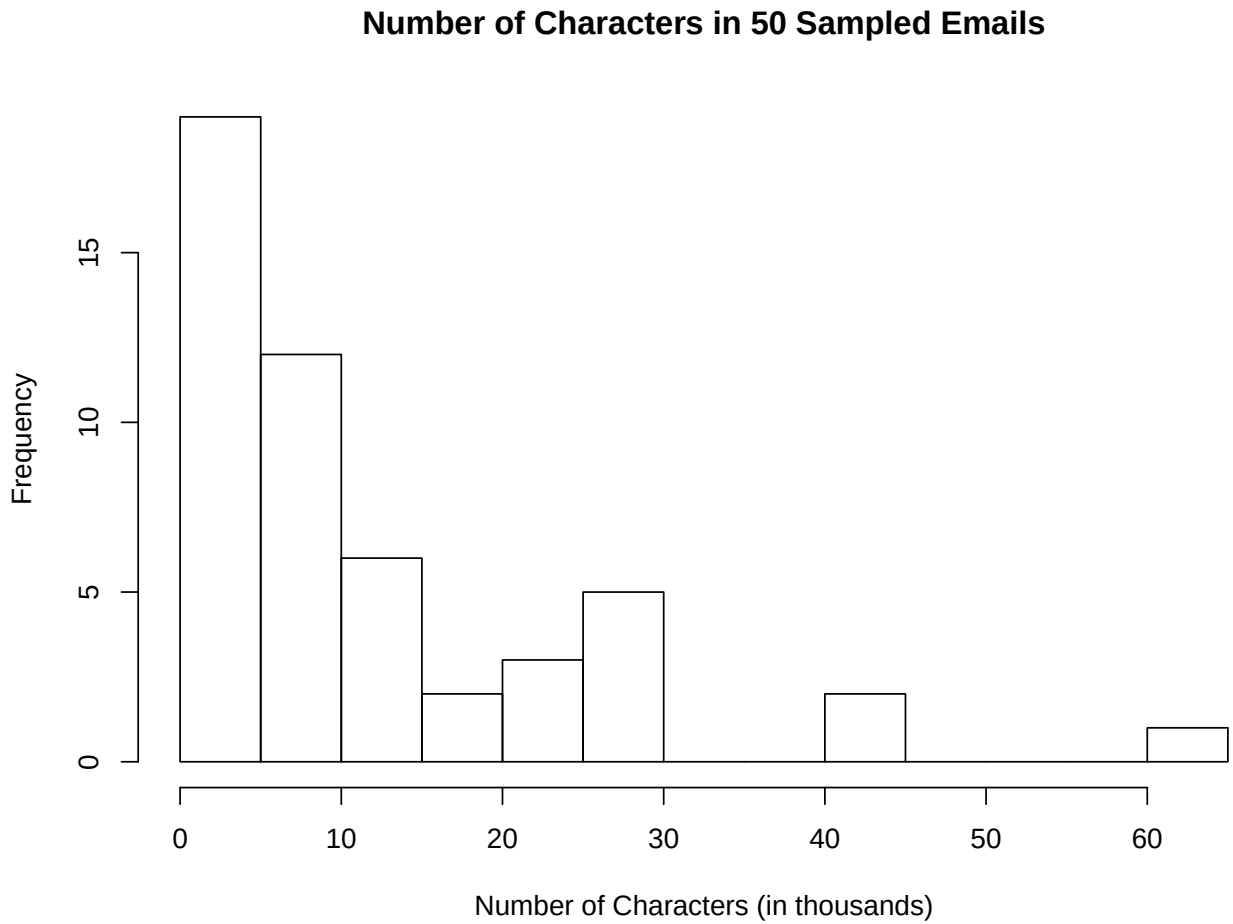
In the `openintro` package, the `email50` dataset contains a random sample of 50 emails, in which researchers try to identify emails as spam. One variable is the # of characters:

Characters (in 1000's)	0-4.999	5-9.999	10-14.999	...	60-64.999
Count	19	12	6	...	1

So each of the intervals 0-5, 5-10, 10-15, etc. are buckets/bins and we count the number of emails in each bucket/bin.

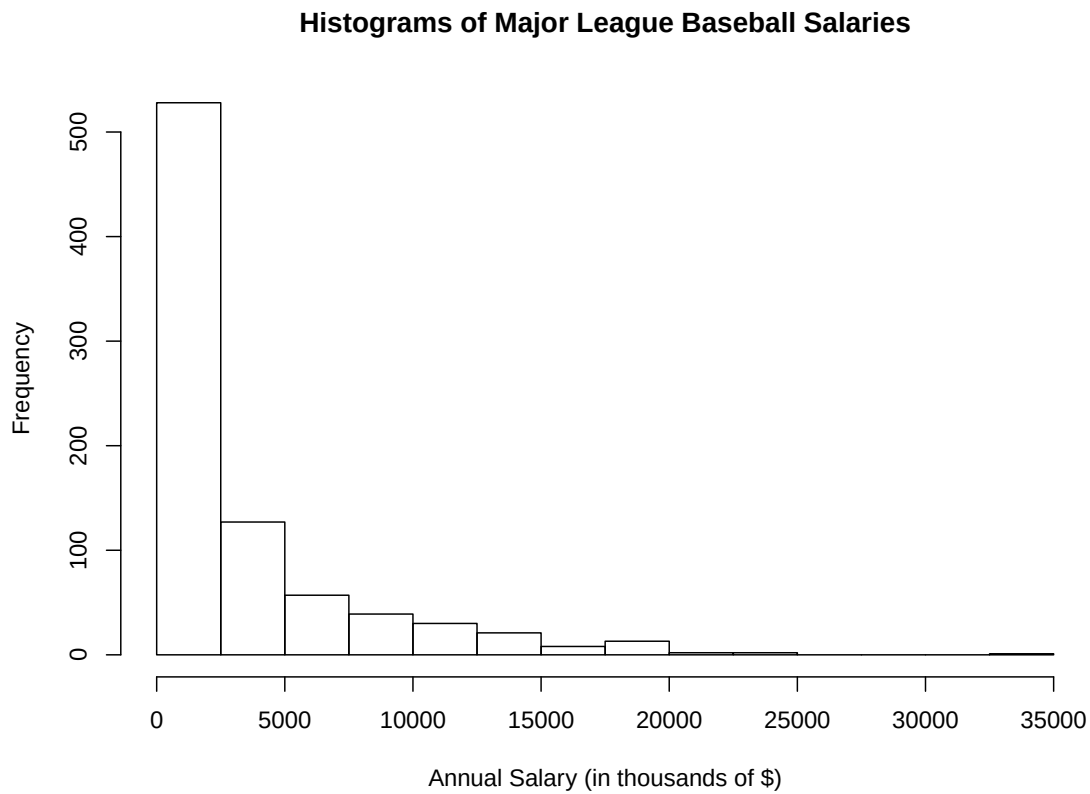
Histograms

Histograms provide a description of the shape of the distribution of data.



Skew and Long Tail

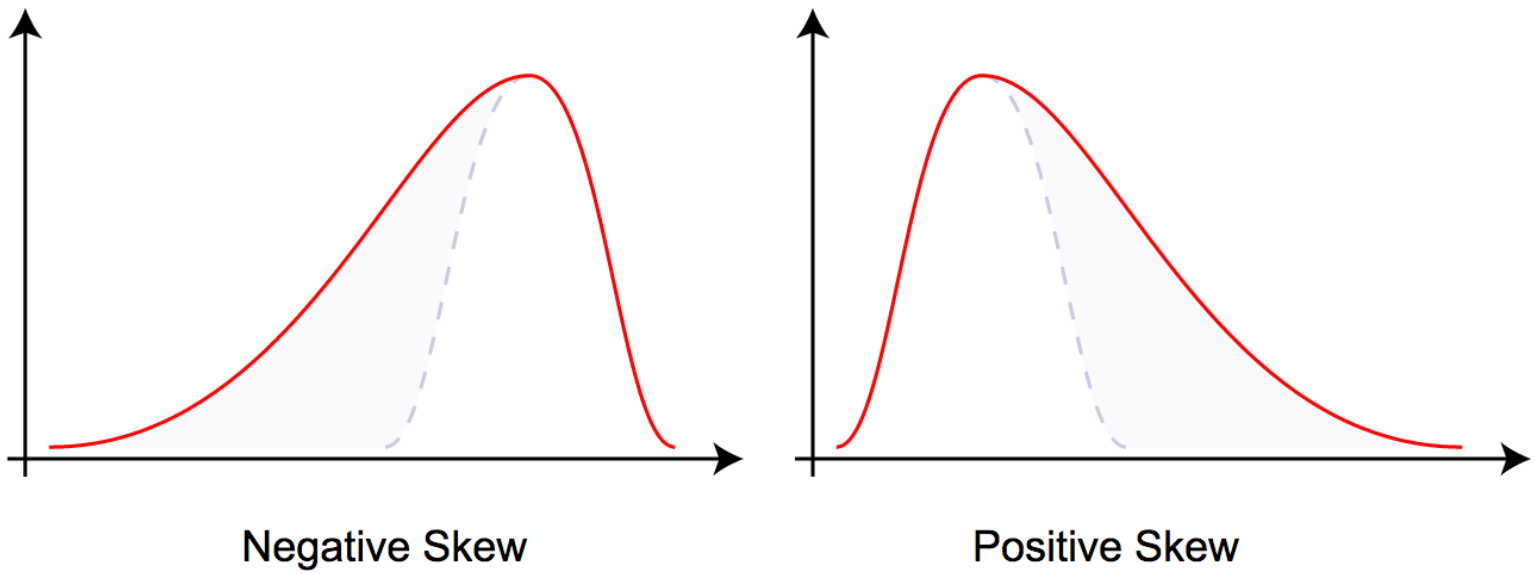
Also in the `openintro` package is MLB salary data in 2010. If we plot a histogram:



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

- Long tail to the right: data is right-skewed AKA positively-skewed
- Long tail to the left: data is left-skewed AKA negatively-skewed



Reed's 2013 US-Originating Entering Class

What can we do about skewed data?

http://rpubs.com/rudeboybert/reed2013

Mean

The mean, AKA average, is a common way to measure the center of the data. So for example, the mean of 1, 2, 5, 3, and 7 is

$$\frac{1+2+5+3+7}{5} = 3.6$$

We label the sample mean $\overline{x}$ (pronounced "x bar"):

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ are the n observed/sampled values.

Median

The median, however, is the middle number.

Two cases:

- Odd number of values: the median of (1, 3, 5, 8, 10) is 5.
- Even number of values: the median of (1, 3, 5, 8) is the average of the middle two values: $\frac{3+5}{2} = 4$

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.

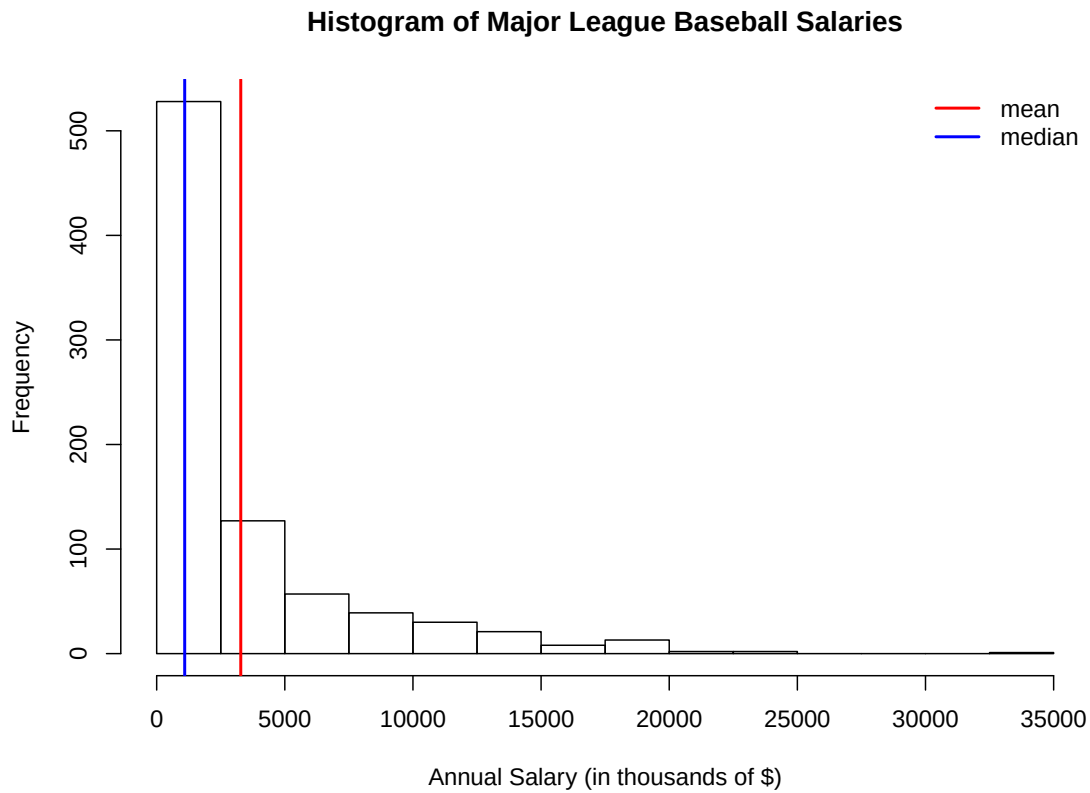
Mean vs Median: Imaginary Scenario

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

Medians are less sensitive to (i.e. more robust to) outliers than the mean.

Ex: the "median home price" is typically used, because it isn't as sensitive as the mean to the few very expensive houses.

Mean vs Median: Back to MLB Salary Data

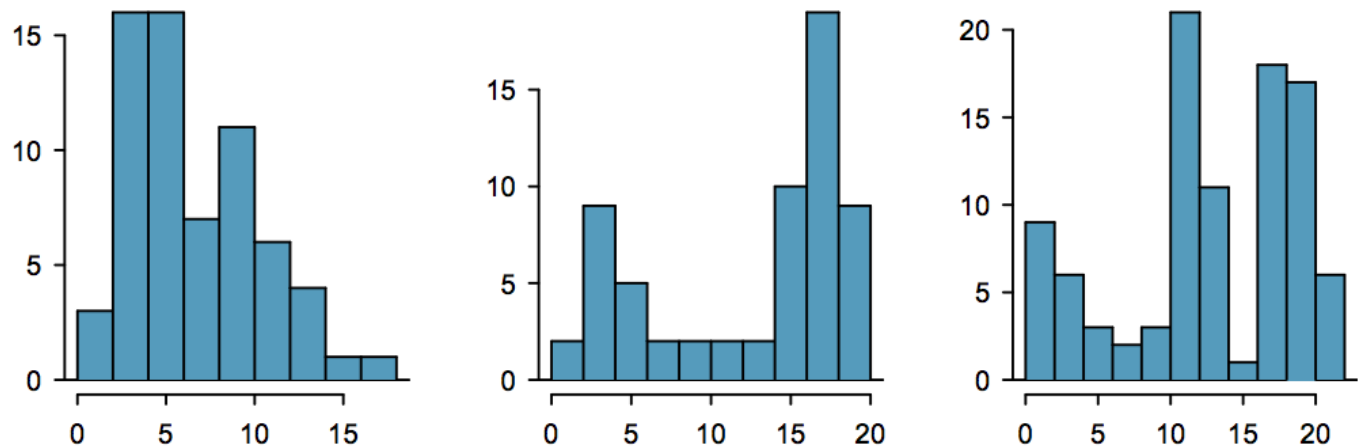


Mode

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.

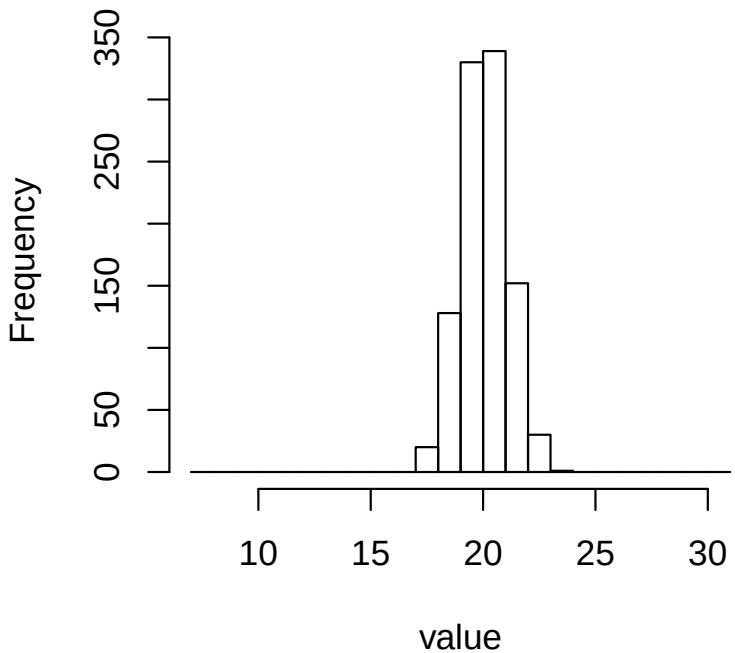
Modes also describe peaks, but this can get subjective.

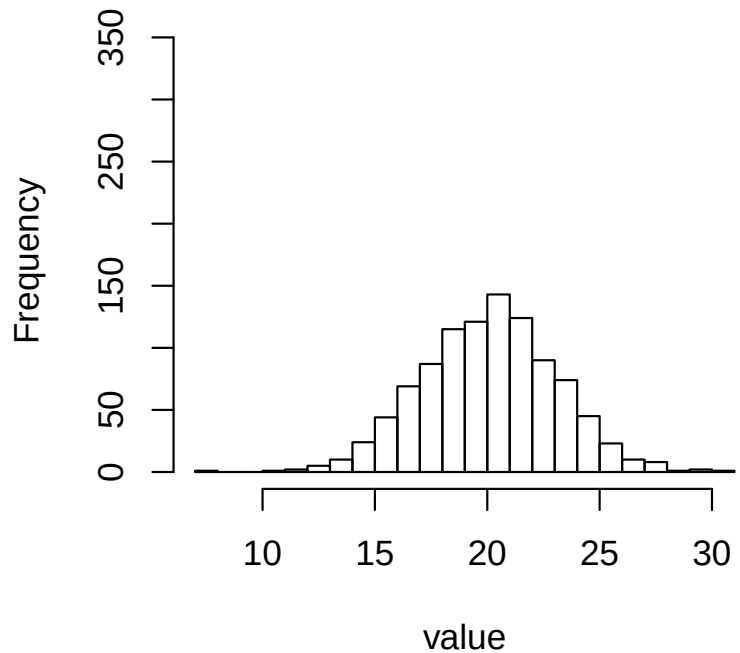
A distribution can be unimodal, bimodal, or multimodal:



Measure of Spread

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





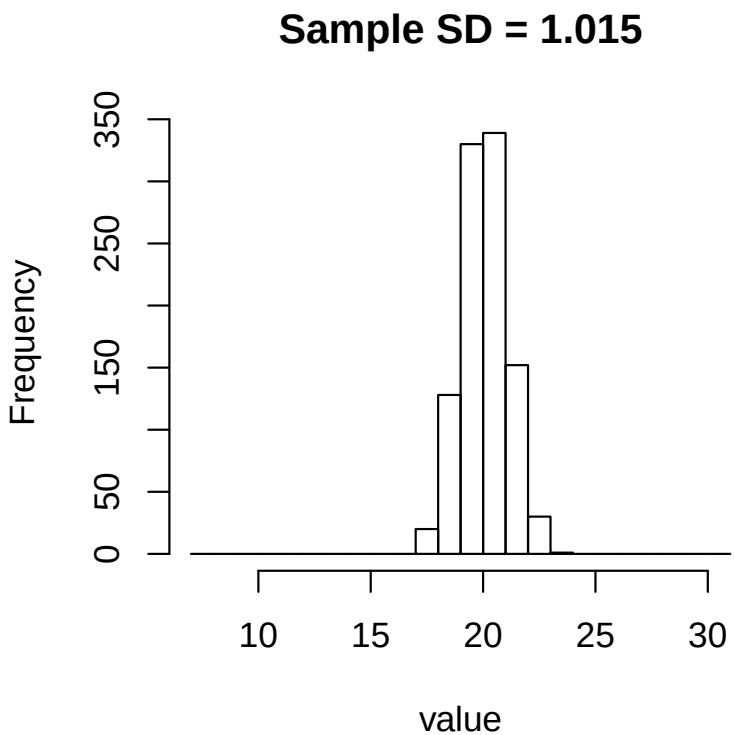
Measure of Spread

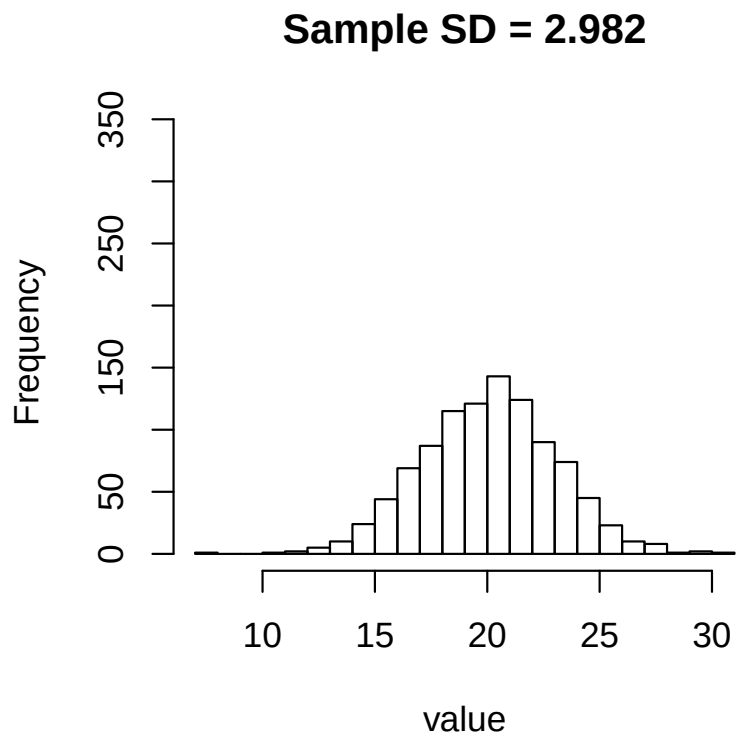
We need a measure of spread/variability. The sample variance s^2 is roughly the average squared distance from the mean.

The sample standard deviation s is the square root of the sample variance. The sample standard deviation is useful when considering how close the data are to the mean.

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 that computes it for you: `sd()`

Next Time

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