Lecture 5: Examining/Visualizing Numerical Data

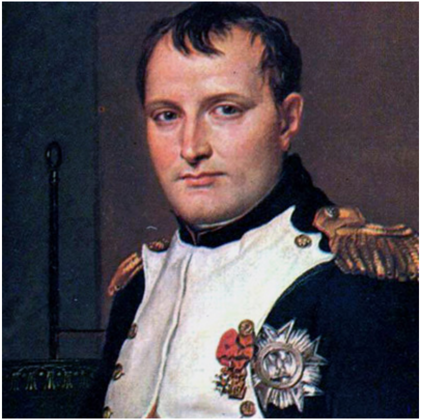
Chapter 1.6 + 1.7

Goals for Today

- Examples of Data Visualization
  - Two famous historical examples of data visualization
- 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

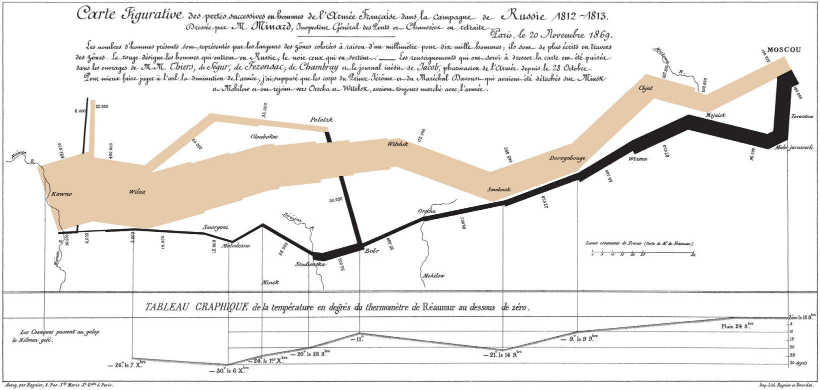
In 1812, Napoleon led a French invasion of Russia, at one point marching on Moscow.



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



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

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 (in others words, 6 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)

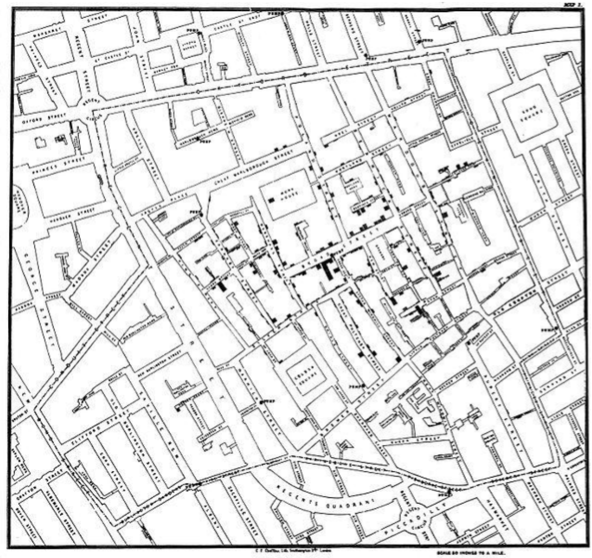
Famous Example 2: 1854 Broad Street Cholera Outbreak

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.

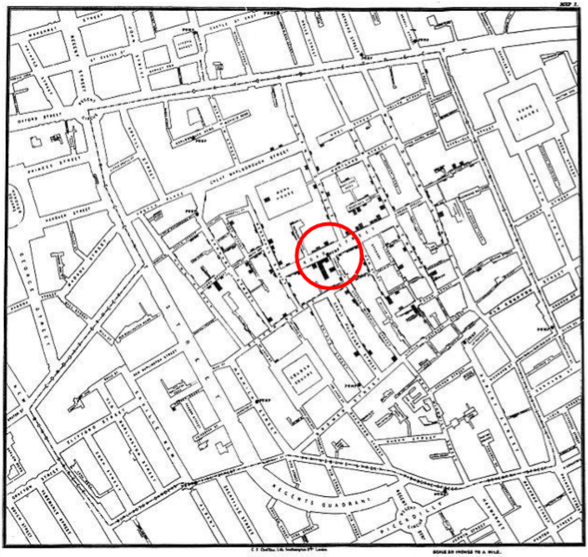
Dr. John Snow, a physician, was a student of the disease. (From Wikipedia) Snow was a skeptic of the then-dominant miasma theory that stated that diseases such as cholera or the Black Death were caused by pollution or a noxious form of "bad air."

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 the disease was transmitted by food and water being contaminated by fecal matter (chiefly due to poor sanitation) and not via the air. This was a watershed moment in the emerging field of epidemiology.

More Recent and Relevant to You: Where do Reedies go?

http://www.reed.edu/reed_magazine/september2013/articles/features/reed_degree.html

Histograms

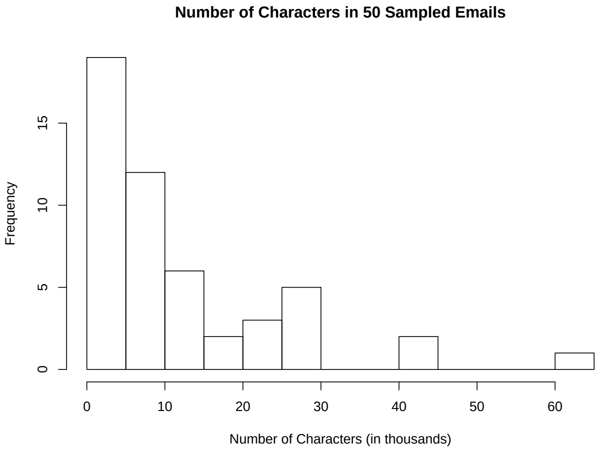
In the openintro package, there is the email50 dataset which is a random sample of 50 emails, in which researchers attempted to identify emails as spam or not. One variables is the # of characters:

```
Characters |  0-5  5-10  10-15  15-20  20-25  25-30  30-35  35-40  40-45  ...  60-65
(in 1000's)|
-------------------------------------------------------------------------------------------
Count      |   19    12      6      2      3      5      0      0      2  ...      1
```

So we can think of each of the intervals 0-5, 5-10, 10-15, etc. as buckets/bins and we want to count the number of emails in each bucket/bin. Let's show this visually...

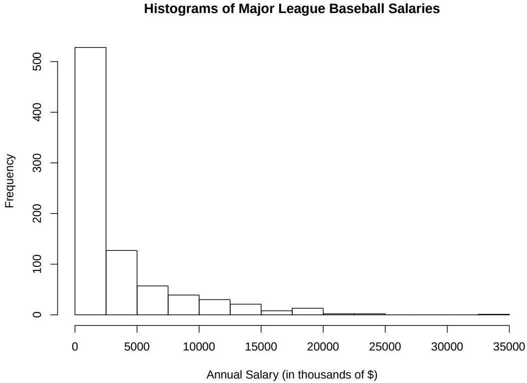
Histograms

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



Skew and Long Tail

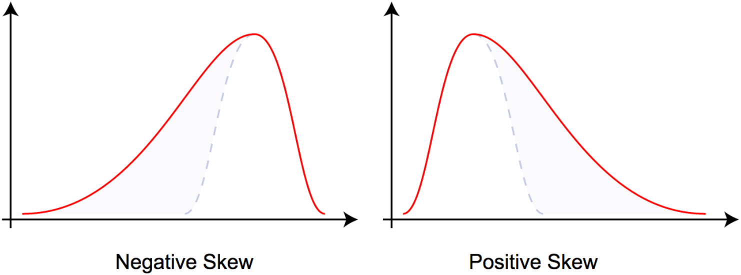
Also in the `openintro` package is MLB salary data. If we plot a histogram:



We see that the data has a long tail to the right. Then we say the data is right-skewed. i.e. we have 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



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 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: $\frac{3+5}{2} = 4$

But why use the median at all?

Mean vs Median: Imaginary Scenario

- Say at company X, there 5 employees: the CEO and everyone else.
- The CEO earns \$1000 an hour, while the other employees 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.

Why could you critique this argument?

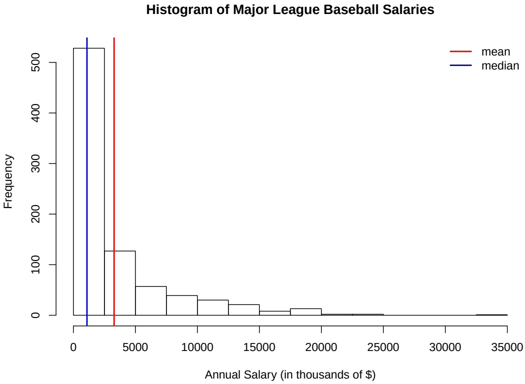
Mean vs Median: Imaginary Scenario

The CEO's extreme salary is inflating the mean! A more appropriate measure in this case would be the median hourly salary of 30.

Medians are much less sensitive to outliers than the mean.
i.e. values that are too extremely high or low values.

Hence, when reporting on house prices the "median home price" is typically used, because it isn't as sensitive as the mean to the few houses that are extremely expensive.

Mean vs Median: Back to MLB Salary Data

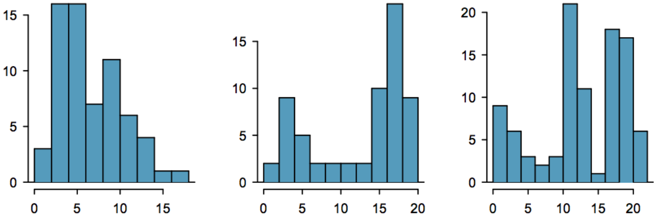


Mode

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

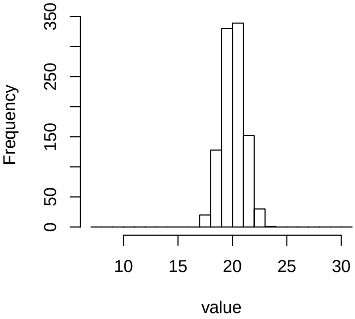
Modes also describe peaks, but this can get subjective.

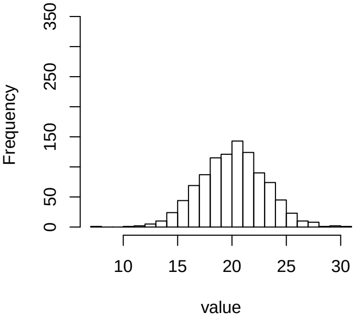
A distribution (of data) can be unimodal (single prominent peak), bimodal (two prominent peaks) and multimodal (two or more prominent peaks):



Measure of Spread

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





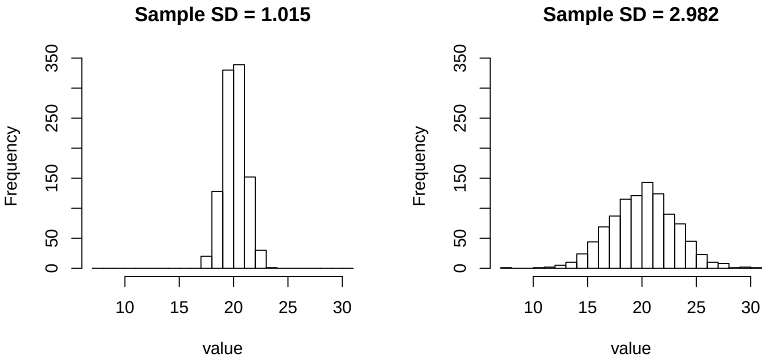
Measure of Spread

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: box plots
- Examining/Visualizing Categorical Data