Introductory Statistics

Instructor: Guenther Walther

We will look at six main topics:

- Descriptive statistics for exploring data, especially visualization
- Some elementary probability
- Sampling distributions and the central limit theorem
- Regression
- Confidence intervals and tests of significance
- Multiple comparisions, reproducibility

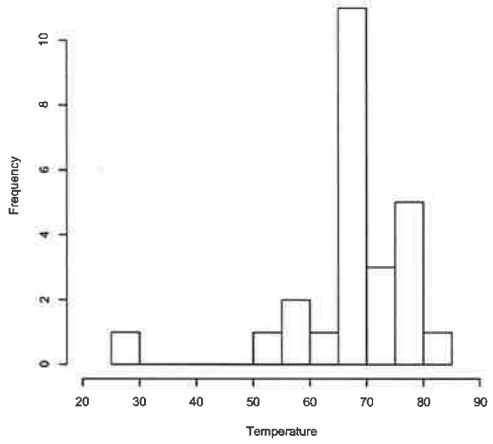
There won't be many formulas, rather we will look at the important statistical ideas behind these methods. Once you understand these ideas, then it's not difficult to look up detailed formulas and more advanced methodology.

Good introductory textbooks are *Statistics* by Freedman, Pisani, and Purves , or *Introduction to Probability & Statistics* by Mendenhall and Beaver (which is a more formal exposition).

Descriptive statistics - why is it important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29° F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F):
66,70,69,80,68,67,72,70,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58,29



The two most important functions of descriptive statistics are:

- Communicate information
- Support reasoning about data

When exploring data of large size, it becomes essential to use summaries.

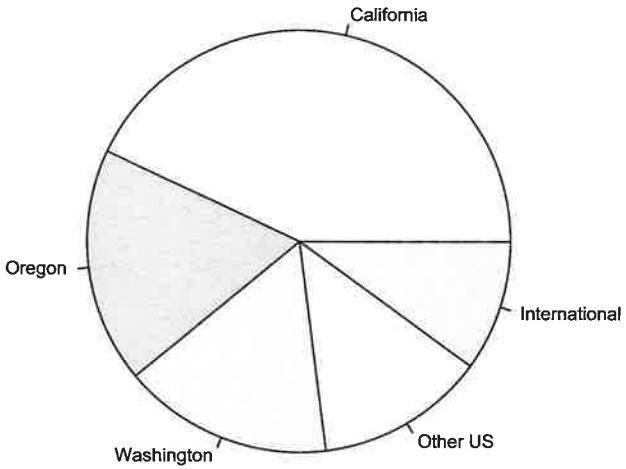
Graphical summaries of data

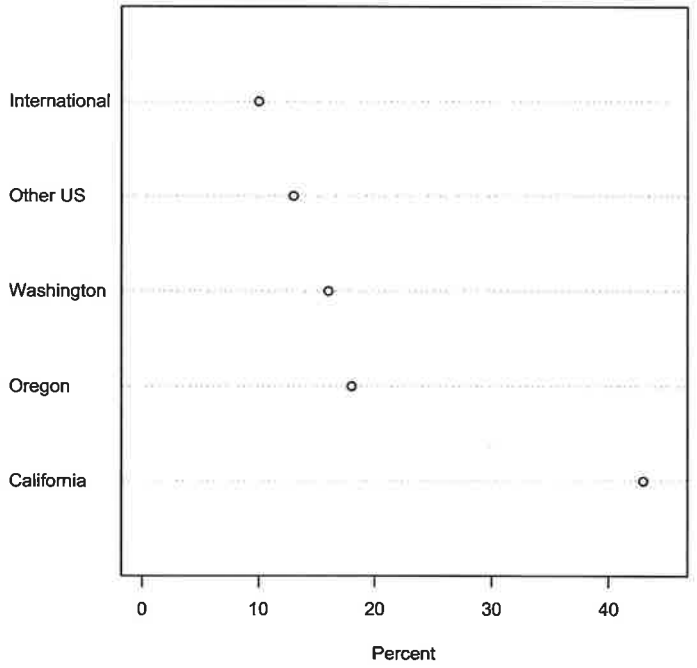
It is best to use a graphical summary to communicate information, because people prefer to look at pictures rather than at numbers.

There are many ways to visualize data. The nature of the data and the goal of the visualization determine which method to choose.

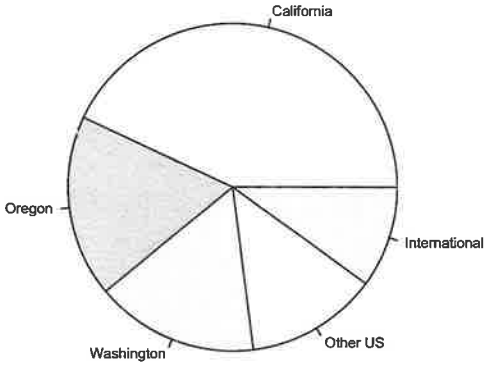
Pie chart and dot plot

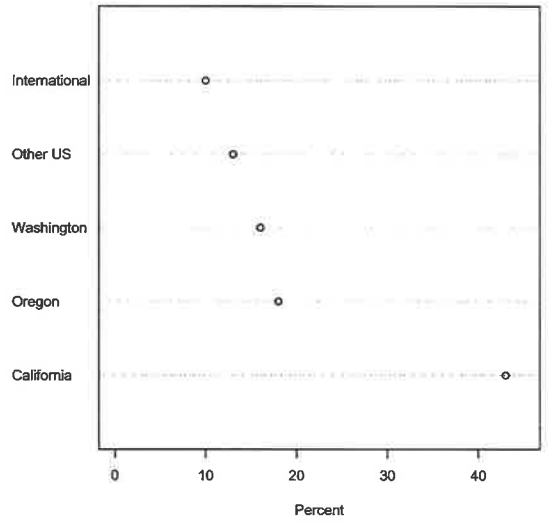
For data that a *qualitative* (e.g. colors, car types,...), use a **pie chart** or a **dot plot**.





Pie chart and dot plot



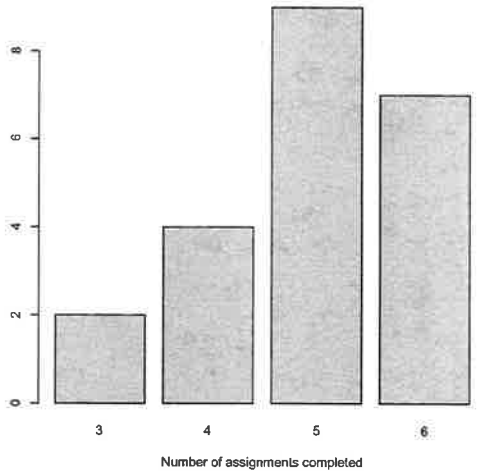


The dot plot makes it easier to compare frequencies of various categories, while the pie chart allows more easily to eyeball what fraction of the total a category corresponds to.

Bar graph

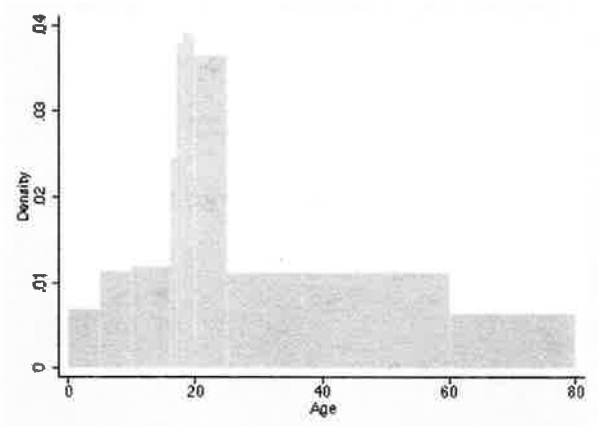
When the data are *quantitative* (i.e. numbers), then they should be put on a number line. This is because the ordering and the distance between the numbers convey important information.

The **bar graph** is essentially a dot plot put on its side.



The histogram

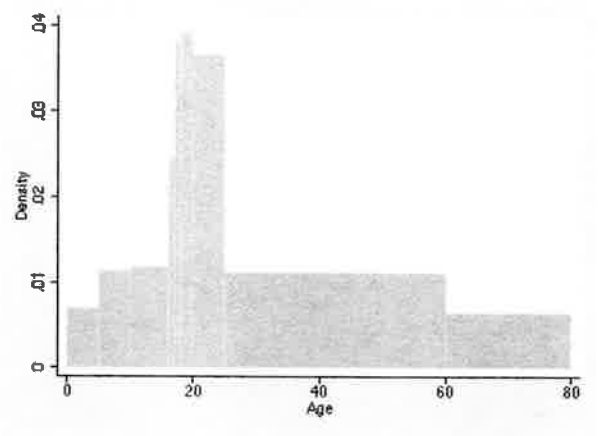
The **histogram** allows to use blocks with different widths.



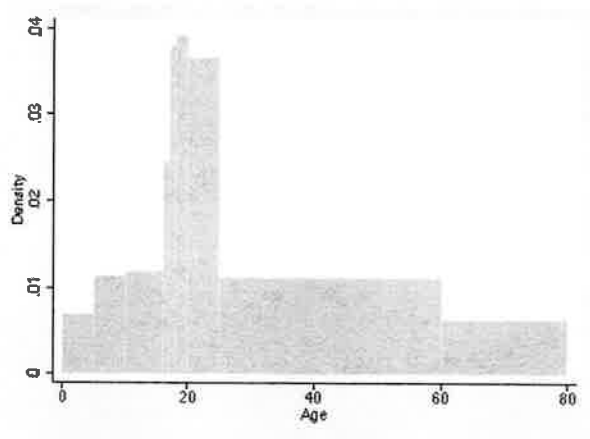
Key point: The areas of the blocks are proportional to frequency.

So the percentage falling into a block can be figured without a vertical scale since the total area equals 100%.
But it's helpful to have a vertical scale (*density scale*). Its unit is '% per unit', so in the above example the vertical unit is '% per year'.

The histogram gives two kinds of information about the data:



1. **Density** (crowding): The height of the bar tells how many subjects there are for one unit on the horizontal scale. For example, the highest density is around age 19 as $.04 = 4\%$ of all subjects are age 19. In contrast, only about 0.7% of subjects fall into each one year range for ages 60–80.



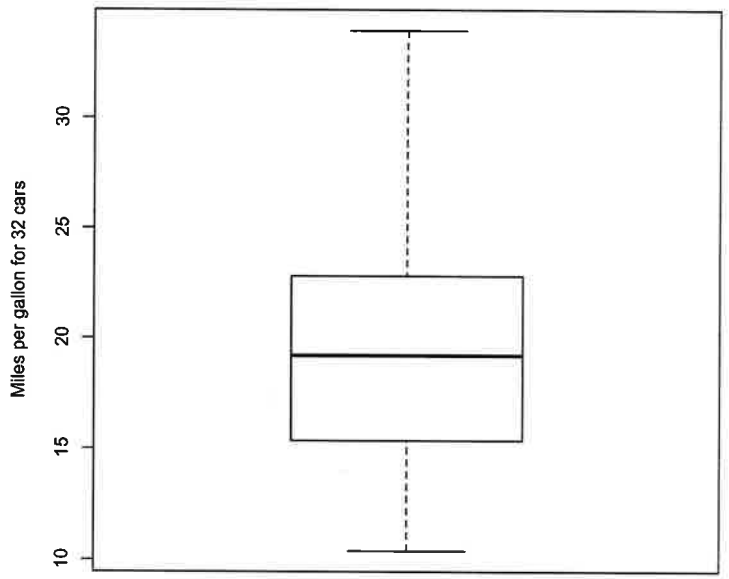
2. **Percentages** (relative frequences): Those are given by

$$\text{area} = \text{height} \times \text{width}.$$

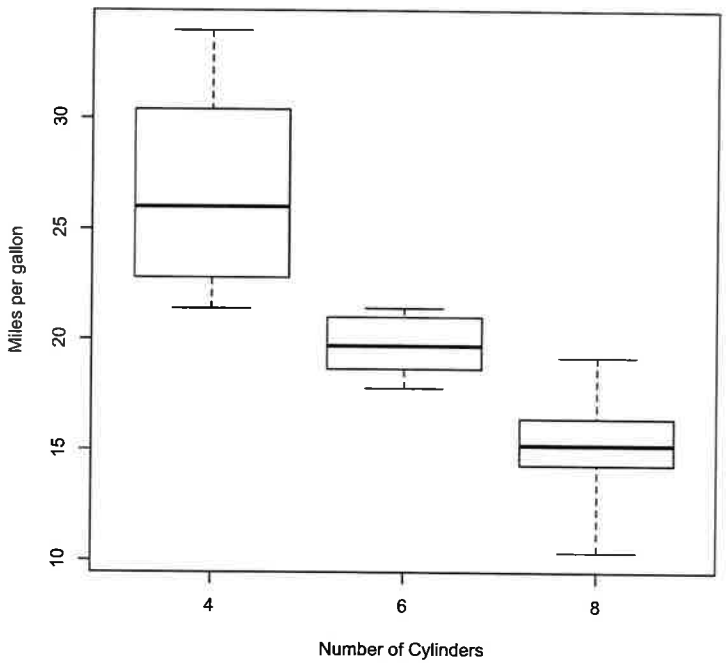
For example, about 14% of all subjects fall into the age range 60–80, because the corresponding area is (20 years) x (0.7 % per year)=14 %. Alternatively, you can find this answer by eyeballing that this area makes up roughly 1/7 of the total area of the histogram, so roughly 1/7=14% of all subjects fall in that range.

The boxplot (box-and-whisker plot)

The **boxplot** depicts five key numbers of the data:

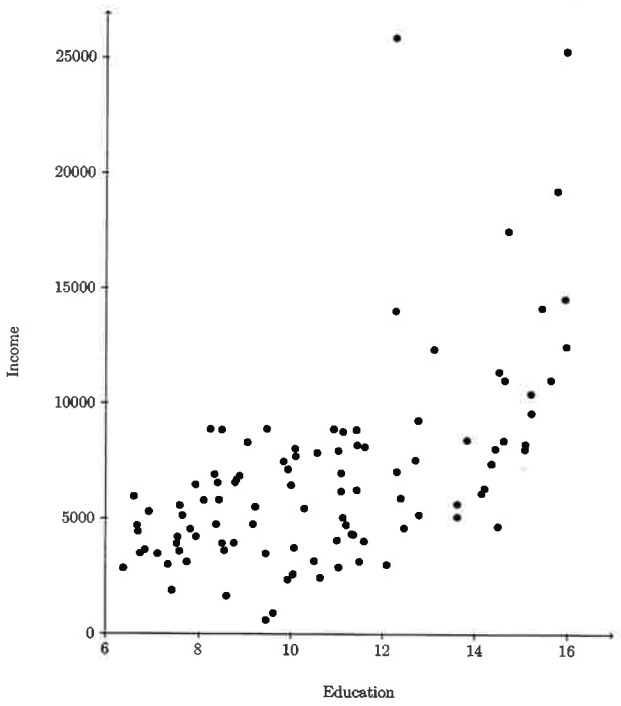


The **boxplot** conveys less information than a histogram, but it takes up less space and so is well suited to compare several datasets:



The scatterplot

The **scatterplot** is used to depict data that come as *pairs*.



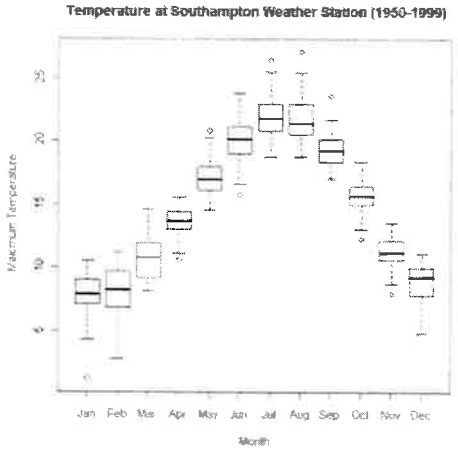
The scatterplot visualizes the relationship between the two variables.

Providing context is important

Statistical analyses typically compare the observed data to a reference. Therefore context is essential for graphical integrity.

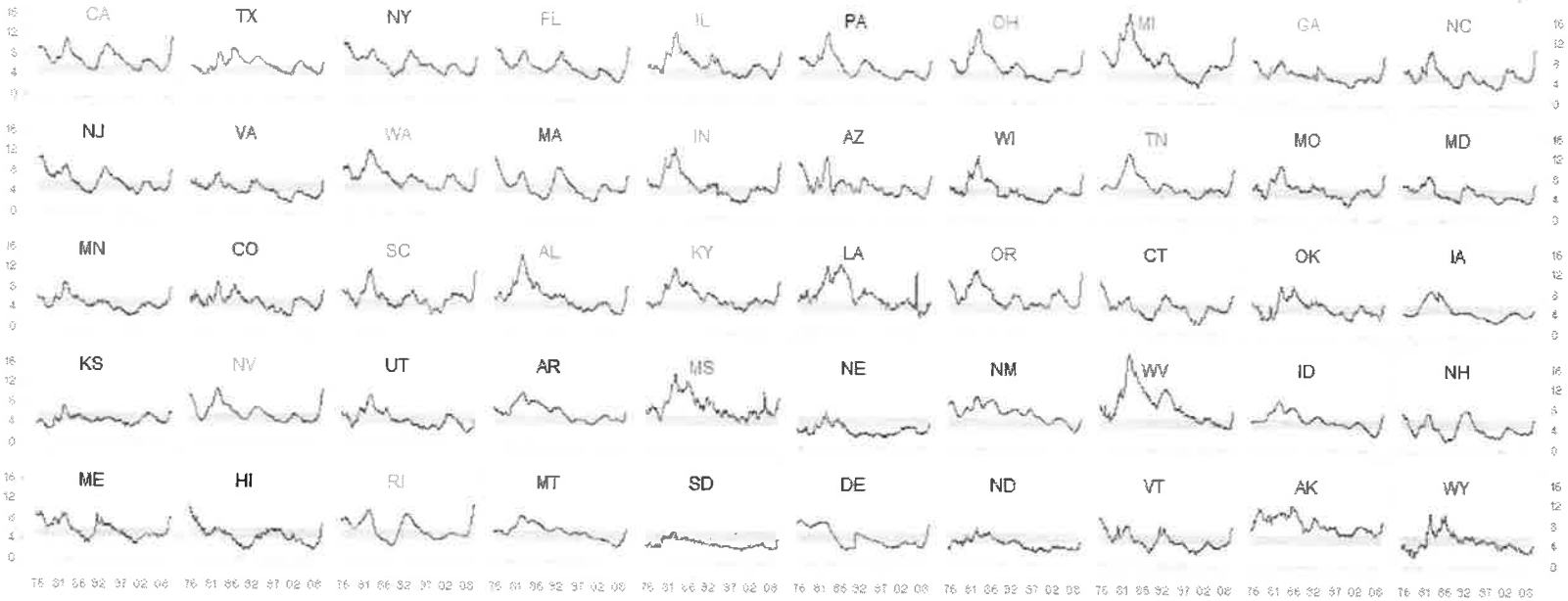
- 'The Visual Display of Quantitative Information' by Edward Tufte (p.74)

One way to provide context is by using *small multiples*. The compact design of the boxplot makes it well suited for this task:



Providing context with small multiples

Monthly Unemployment Rates by State, Jan 1976 - Apr 2009

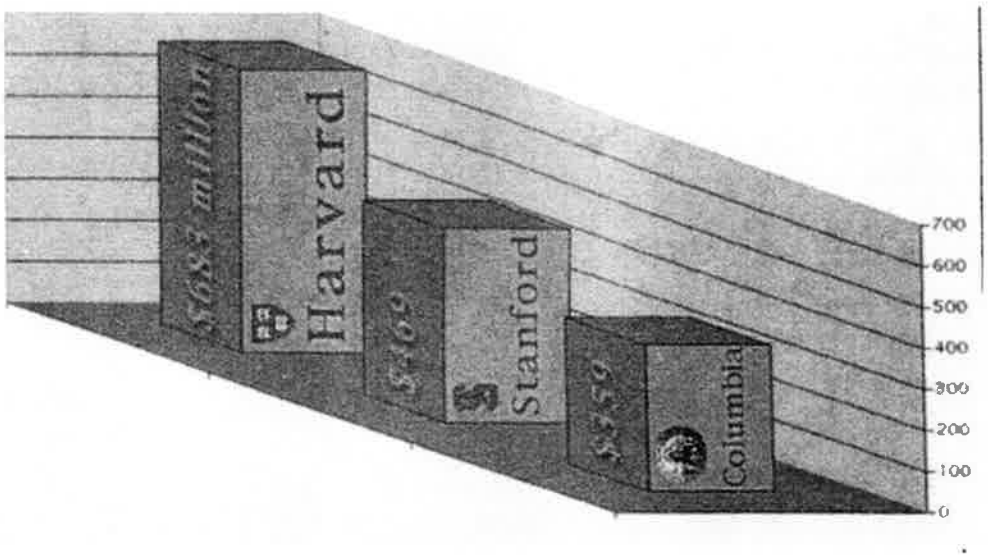


Source: Bureau of Labor Statistics

Notes: The orange band denotes a "normal" unemployment rate (4%-6%);
State code in red: unemployment rate in April 2009 is higher than the US average

Pitfalls when visualizing data

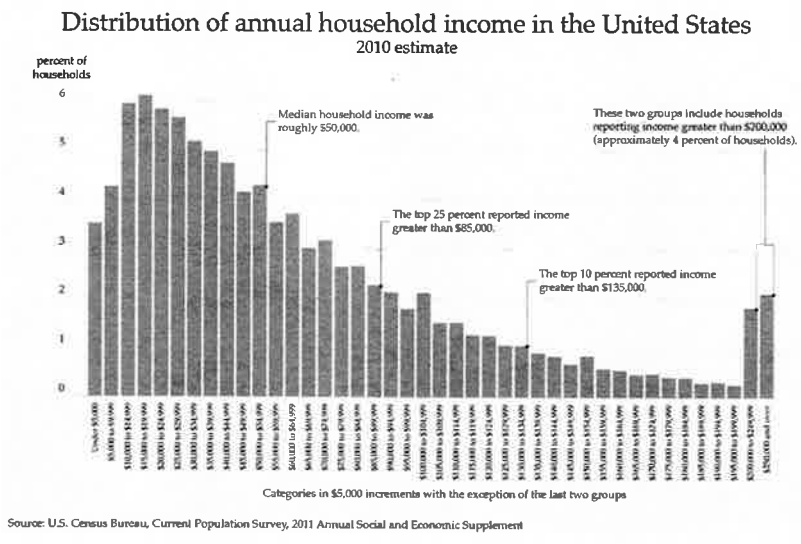
Sophisticated software makes it tempting to produce showy but poor visualizations:



- The 'Ghettysburg Powerpoint Presentation' by Peter Norvig

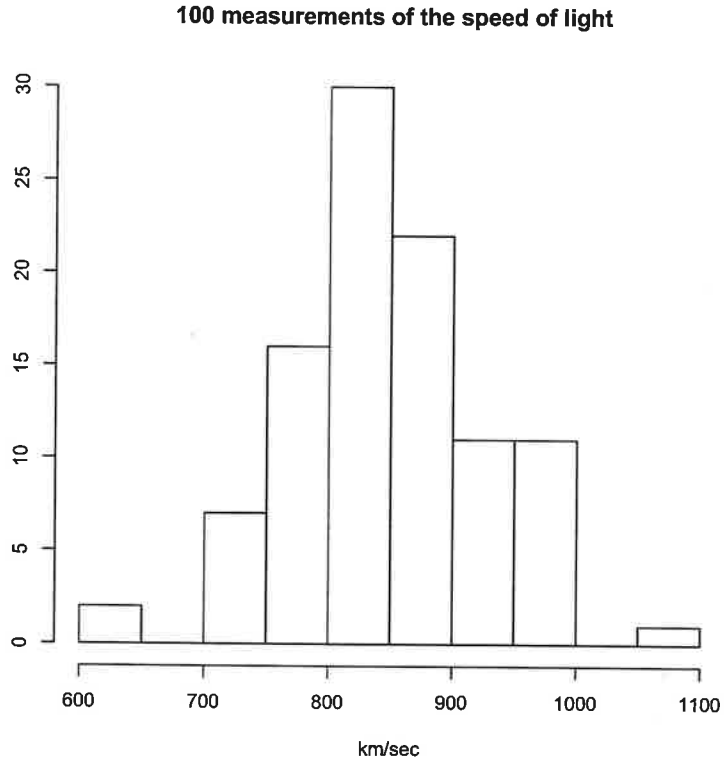
Numerical summary measures

For summerizing data with one number, use the **mean** (=average) or the **median**. The median is the number that is larger than half the data and smaller than the other half.



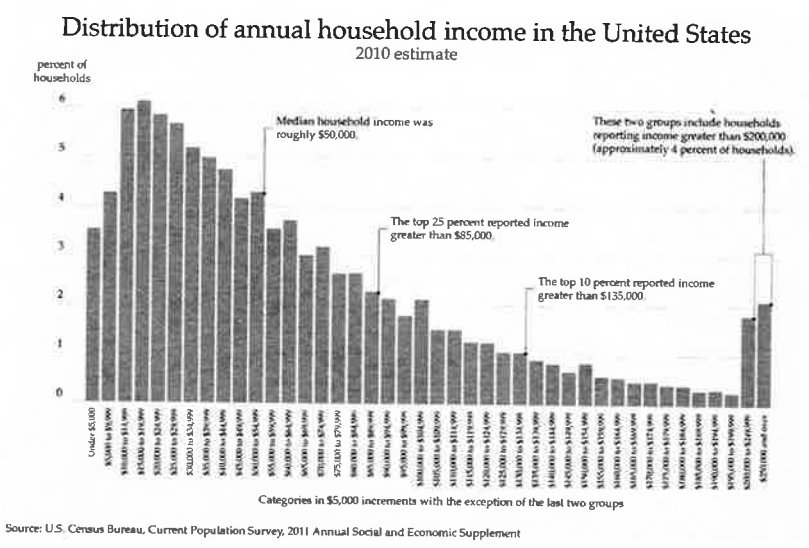
Mean vs. median

Mean and median are the same when the histogram is symmetric.



Mean vs. median

When the histogram is *skewed to the right*, then the mean can be much larger than the median.



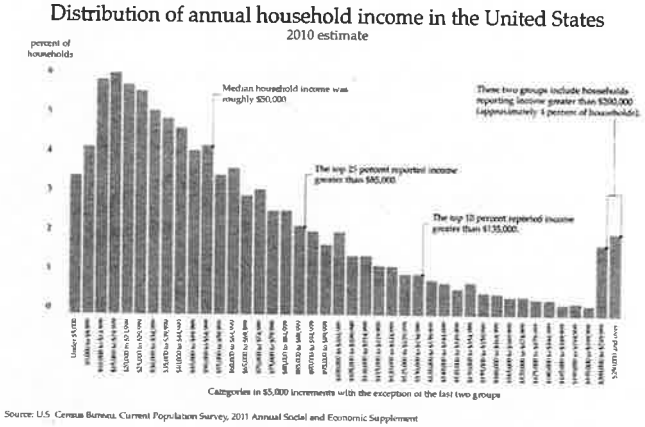
So if the histogram is very skewed, then use the median.

Mean vs. median

If the median sales price of 10 homes is $ 1 million, then we know that 5 homes sold for $ 1 million or more.

If we are told that the average sale price is $ 1 million, then we can't draw such a conclusion:

Percentiles



The 90th percentile of incomes is \$ 135,000: 90% of households report an income of \$ 135,000 or less, 10% report more.

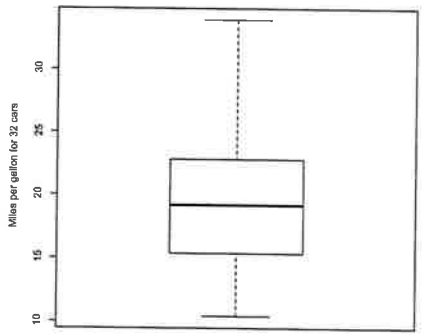
The 75th percentile is called **3rd quartile**: \$ 85,000

The 50th percentile is the **median**: \$ 50,000

The 25th percentile is called **1st quartile**.

Five-number summary

Recall that the boxplot gives a **five-number summary** of the data:
the smallest number, 1st quartile, median, 3rd quartile, largest number.



The **interquartile range** = 3rd quartile − 1st quartile.
It measures how spread out the data are.

The standard deviation

A more commonly used measure of spread is the **standard deviation**.

$\bar{x}$ stands for the average of the numbers $x_1, \ldots, x_n$.

The standard deviation of these numbers is

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{or} \quad \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The two numbers $\bar{x}$ and s are often used to summarize data. Both are sensitive to a few large or small data.

If that is a concern, use the median and the interquartile range.

Mini quiz

- For each of the following two data sets state an appropriate way for visualizing the data:
  a) A list of the eye colors of 120 people.
  b) A list that gives both the size (measured in square feet) and the number of bedrooms for 1513 houses.
- The average sales price for houses in a certain county during the last year was $ 342,000. Do the houses that sold for more than $ 342,000 constitute more/equal/less than 50% of all sales?