Variables and data types

(*) Data comes from ***observations***.

(*) Each observation yields values for one or more ***variables***.

(*) ***Qualitative variables:*** The characteristic is *categorical.* E.g., gender, ethnicity, treatment group vs. control group.

(*) ***Quantitative variables:*** The characteristic is *numerical.* E.g., income level, age, blood pressure. Quantitative variables can be *discrete* or *continuous.*

- ***Discrete*** variables can take values that differ by fixed amounts, usually used to *count* things. E.g., number of children in a household.
- ***Continuous*** variables can take values that differ by arbitrarily small amounts. E.g., height or temperature. Continuous variables come in two flavors. *Interval* variables — the difference between two values has a reasonable interpretation, but the ratio does not. E.g., temperature. *Ratio* variables — the ratio between two values has a reasonable interpretation. E.g., height.

Example: 500 households are surveyed by a marketing research firm. The investigators collect data on: size of each household; monthly household income; occupation of head-of-household ; number of computers in house; type of internet connection.

(*) 500 observations, each producing data for five variables.

(*) Household size, monthly income and number of computers — these are quantitative variables.

- Income is (treated like) a continuous variable.
- Household size and number of computers are discrete variables.

(*) Occupation of head of household and type of internet connection are qualitative variables.

Tables – categorical data

(*) Categorical data can be ***summarized*** in tables by recording the ***frequency*** or ***relative frequency*** of the data in each category.

Example. The table below describes the results of the randomized, double-blind field test of the Salk polio vaccine.

Group	***size***	***infections/100,00***
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

(*) Observations: children.

(*) Variables: group to which child belongs (three categories) and infection status (two categories).

Distribution tables – quantitative data

(*) To summarize quantitative data in a table, the typical approach is to treat it like *categorical data*, using different ranges of values as the categories.

(*) The ***range*** (smallest to biggest) of the observed values is divided into ***class intervals***, also called ***bins***. The bins play the role of the categories.

(*) The ***frequency***, or ***relative frequency*** of each class interval is recorded in a ***distribution table***.

- The frequency of a bin is the ***number*** of the observations that fall into that bin.
- The relative frequency of a bin is the ***proportion*** of the observations that fall into that bin. Proportions are typically recorded as ***percentages***.

Comment: Data can be divided into class intervals in different ways. How this is done can affect the way that the data is perceived.

Example: US household incomes from 2015 — frequency distribution.

Income level	*Frequency*
\$0 - \$14,999	14,595,004
\$15,000 - \$24,999	13,210,995
\$25,000 - \$34,999	12,581,900
\$35,000 - \$49,999	15,979,013
\$50,000 - \$74,999	21,011,773
\$75,000 - \$99,999	15,224,099
\$100,000 - \$149,999	17,740,479
\$150,000 - \$199,999	7,800,778
\$200,000 and over	7,674,959

Example: US household incomes from 2015 — *relative* frequency distribution.

Income level	*Relative frequency*
\$0 - \$14,999	11.6%
\$15,000 - \$24,999	10.5%
\$25,000 - \$34,999	10%
\$35,000 - \$49,999	12.7%
\$50,000 - \$74,999	16.7%
\$75,000 - \$99,999	12.1%
\$100,000 - \$149,999	14.1%
\$150,000 - \$199,999	6.2%
\$200,000 and over	6.1%

Comment: A distribution table makes it much easier to absorb large amounts of data. The price we pay is the loss of information inherent in *summarizing*.

When determining the class intervals for the table, you have to decide how much of the fine detail you are willing to lose.

In the table below we lose (hide?) information about the very poor and the very rich, and the middle class is divided into two large groups, upper and lower.

Income level	*Relative frequency*
\$0 - \$24,999	22.1%
\$25,000 - \$49,999	22.7%
\$50,000 - \$99,999	28.8%
\$100,000 and over	26.4%

Cross-tabulation

(*) In studies with more than one category (or more than one quantitative variable), we can produce different distribution tables for different categories. The separate distribution tables can be combined into one table (with many columns).

(*) The result of this process is called a ***cross-tabulation***, and it helps to ***control for*** (observe the effect of) confounding variables.

Example. Oral contraceptives and blood pressure. The following table summarizes the results of the study on the effects of oral contraceptives on the blood pressure of women who use them done by the Kaiser clinic in Walnut Creek, CA.

(*) Qualitative variable: ***user/nonuser***

(*) Quantitative variable: ***blood pressure***.

(*) Variable controlled for: ***age***.

Table 2. Systolic blood pressure by age and pill use, for women in the Contraceptive Drug Study, excluding those who were pregnant or taking hormonal medication other than the pill. Class intervals include the left endpoint, but not the right. – means negligible. Table entries are in percent; columns may not add to 100 due to rounding.

Blood	*Age 17–24*		*Age 25–34*		*Age 35–44*		*Age 45–58*	
pressure	*Non-*		*Non-*		*Non-*		*Non-*	
(millimeters)	*users*	*Users*	*users*	*Users*	*users*	*Users*	*users*	*Users*
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
under 90	–	1	1	–	1	1	1	–
90–95	1	–	1	–	2	1	1	1
95–100	3	1	5	4	5	4	4	2
100–105	10	6	11	5	9	5	6	4
105–110	11	9	11	10	11	7	7	7
110–115	15	12	17	15	15	12	11	10
115–120	20	16	18	17	16	14	12	9
120–125	13	14	11	13	9	11	9	8
125–130	10	14	9	12	10	11	11	11
130–135	8	12	7	10	8	10	10	9
135–140	4	6	4	5	5	7	8	8
140–145	3	4	2	4	4	6	7	9
145–150	2	2	2	2	2	5	7	9
150–155	–	1	1	1	1	3	2	4
155–160	–	–	–	1	1	1	1	3
160 and over	–	–	–	–	1	2	2	5
Total percent	100	98	100	99	100	100	99	99
Total number	1,206	1,024	3,040	1,747	3,494	1,028	2,172	437

Histograms

A ***histogram*** is a graphical representation of a distribution table (for quantitative data).

- Histograms *for data* are usually drawn as bar-charts.
- The horizontal axis of the chart is divided into class intervals.
- The scale on the vertical axis of the chart is typically one of following three:

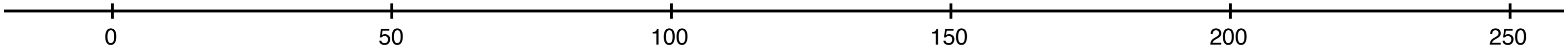
(*) The *frequency* – the number of all observations in a given bin.

(*) The *relative frequency* – the percentage of all observations in a given bin.

(*) The *density* – the relative frequency of the bin divided by its width.

- If one uses the ***density scale*** (as we will be in this class), then the ***area*** of the region drawn above a class interval represents the *relative frequency* of (pecentage of data in) that class interval.

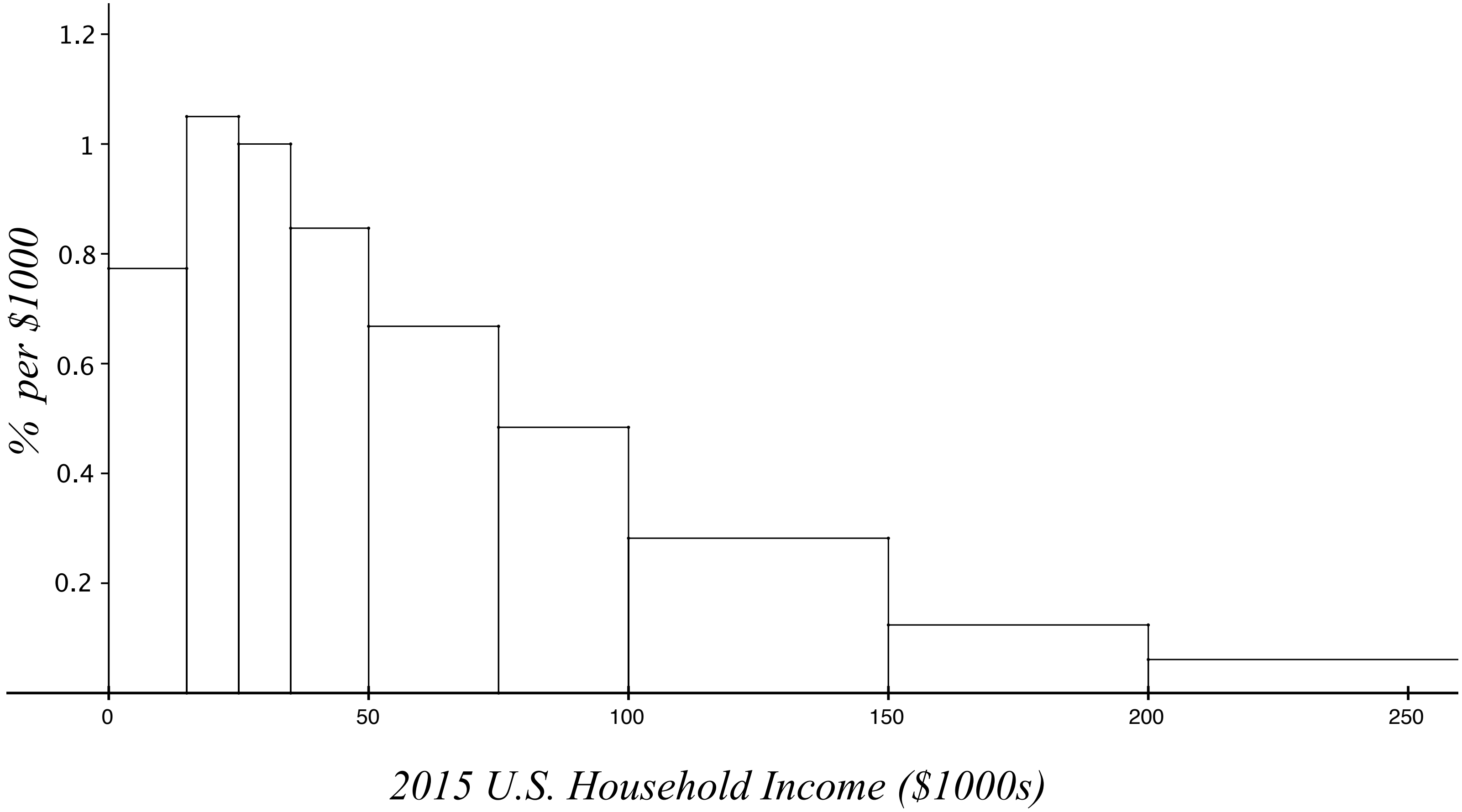
Example: Starting with the table of income distribution we saw earlier, we first draw the horizontal axis...



... Using a density scale, we draw rectangles over each class interval whose ***areas*** equal the percentages of the families in those intervals.

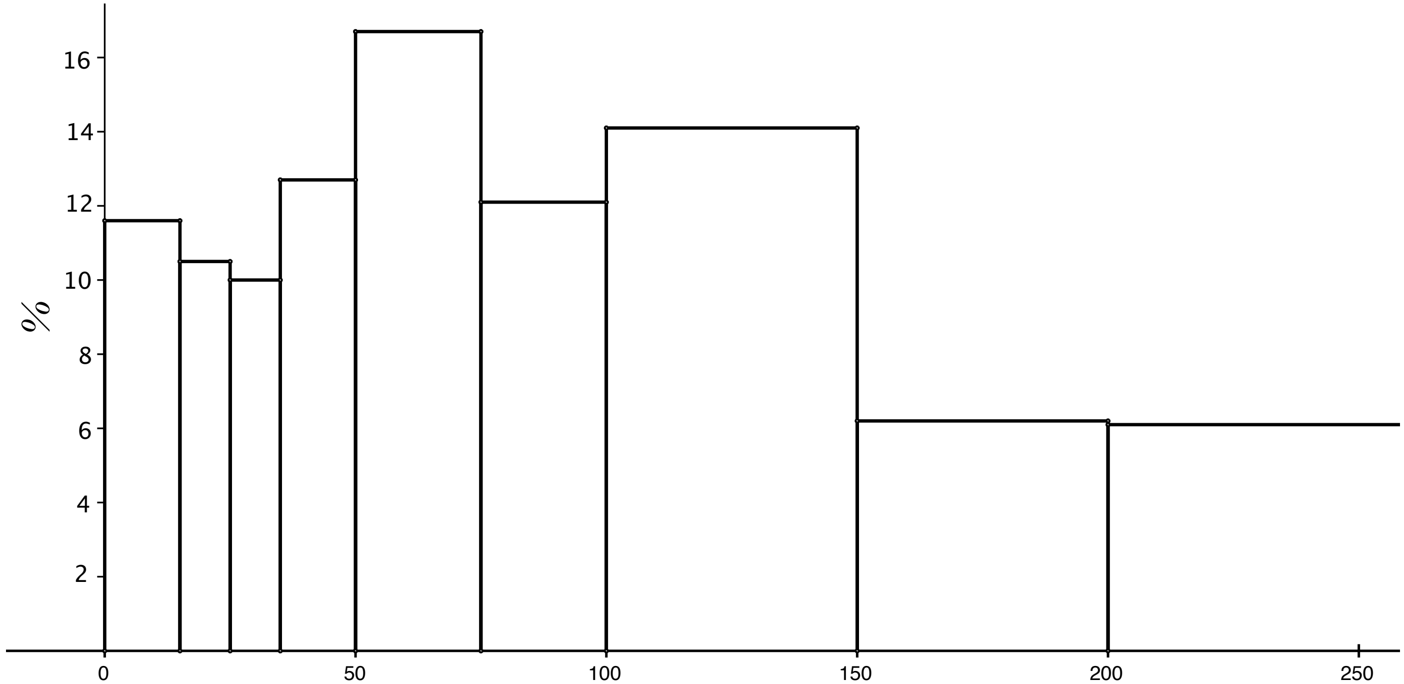
The height of each rectangle is equal to the percentage of the observations in the corresponding class interval divided by the length of the class interval (the width of the rectangle).

The end-result should look like this



The vertical scale here is *percent per $1000* – i.e., it is the relative frequency (percentage) divided by the width of the intervals (which in this case are measured in $1000s). It's always a good idea to label the axes.

If, for example, we use the *relative frequency* scale instead of the *density* scale, the histogram looks like this:



2015 U.S. Household Income ($1000s)

This histogram reports the information accurately, but it is misleading. The bins for the higher incomes seem to be much bigger than the bins for the lower incomes *because they are wider*.

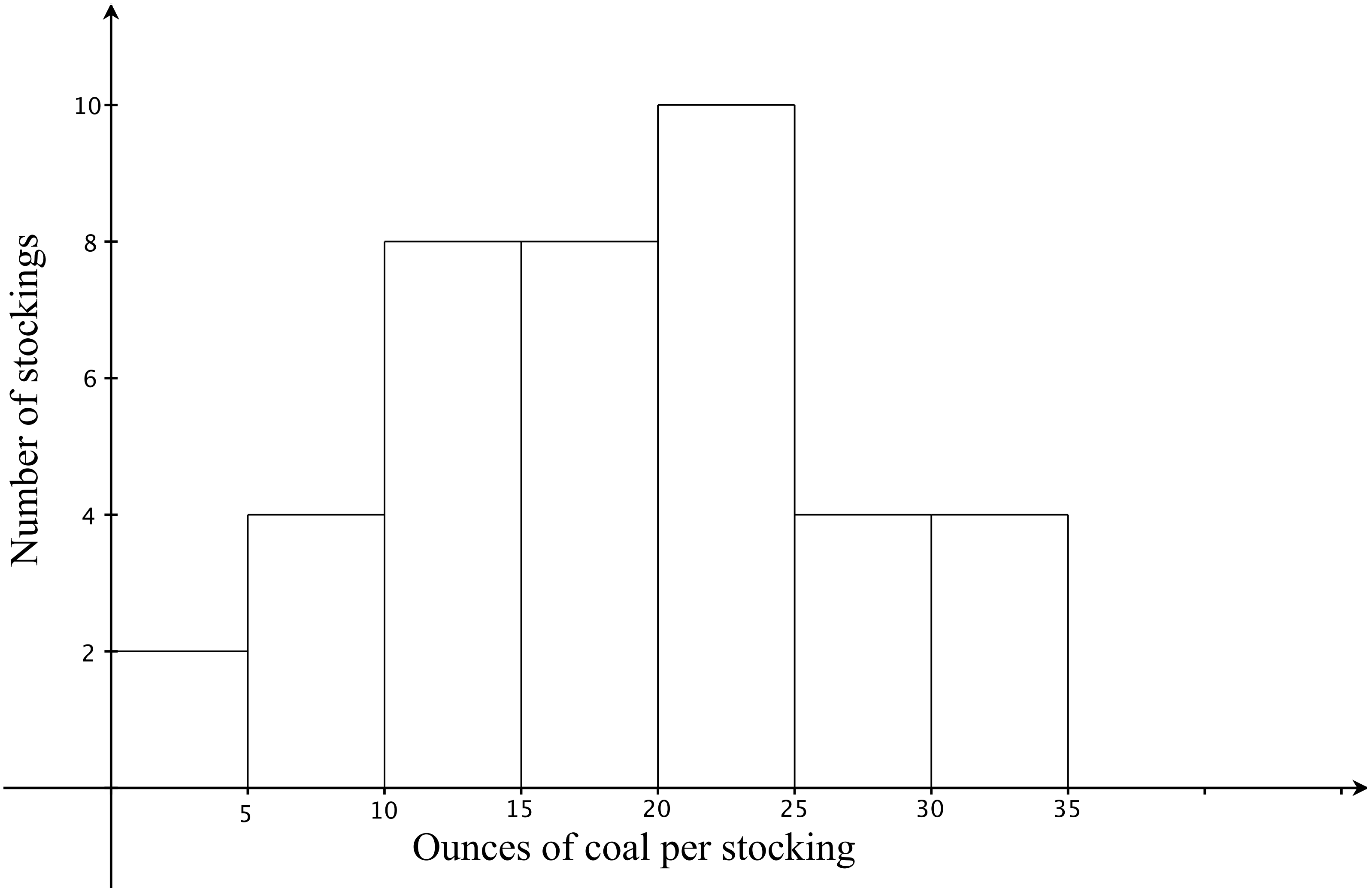
(*) ***If bins have different widths — use the density scale.***

Comment: If all the bins in the distribution table have the same width, then the appearance of the histogram will be the same for all three scales. Only the units (and numbers) on the vertical scale will change.

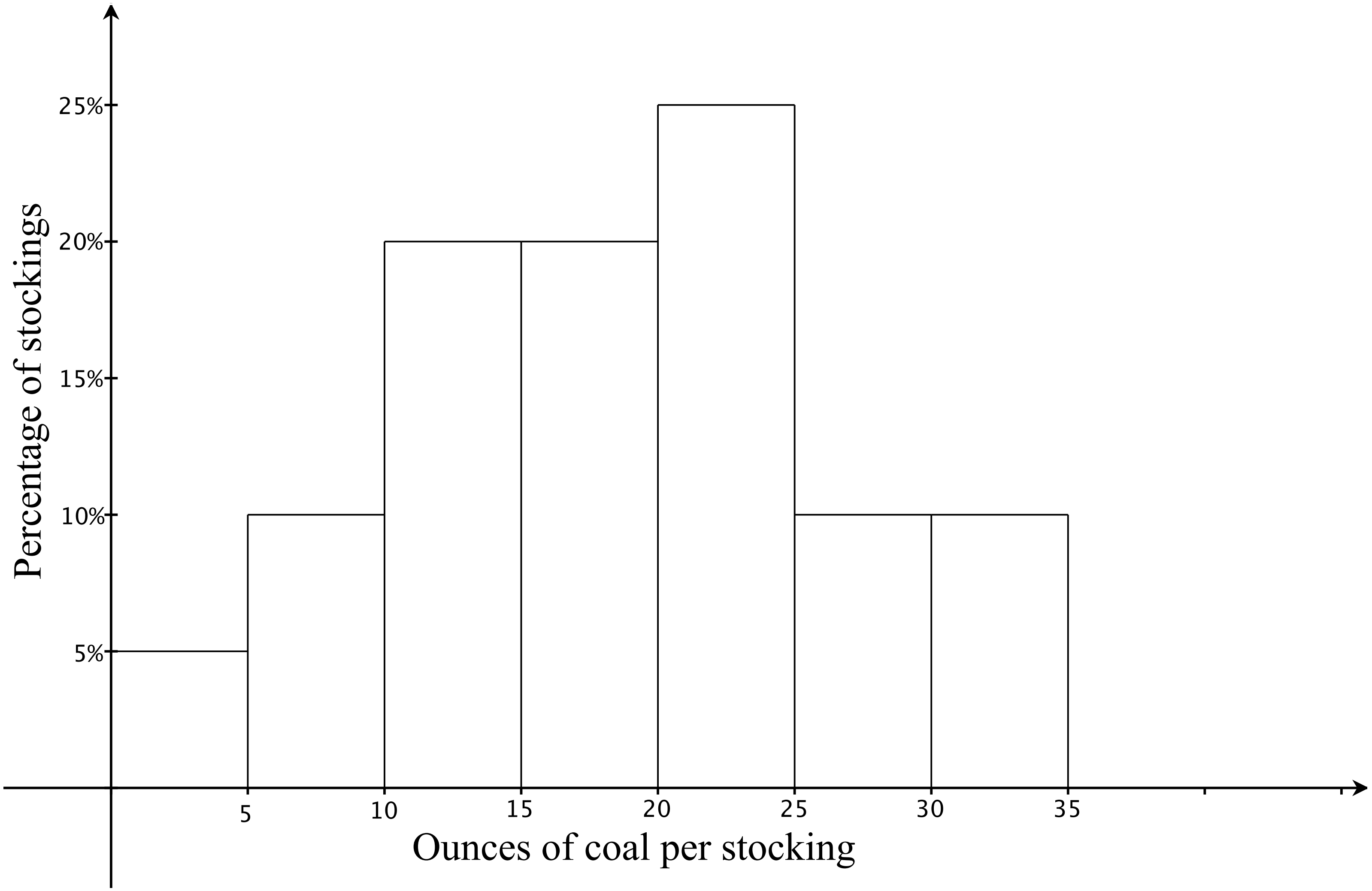
Example: Distribution of coal (by weight) in Christmas stockings of 40 children at Wool's orphanage.

ounces of coal	number of stockings
0 – 5	2
5 – 10	4
10 – 15	8
15 – 20	8
20 – 25	10
25 – 30	4
30 – 35	4

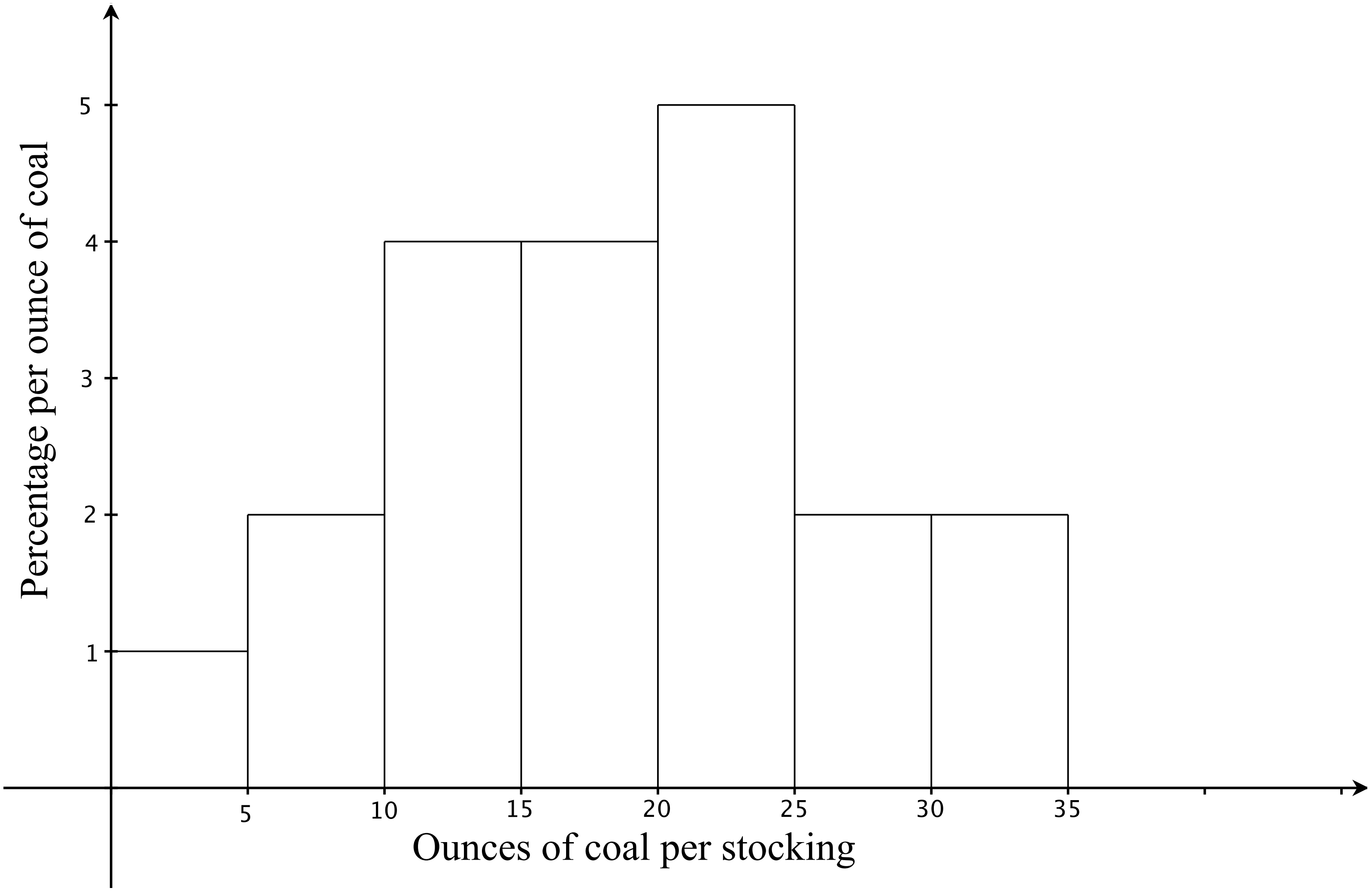
Histogram with frequency scale:



Histogram with relative frequency scale:

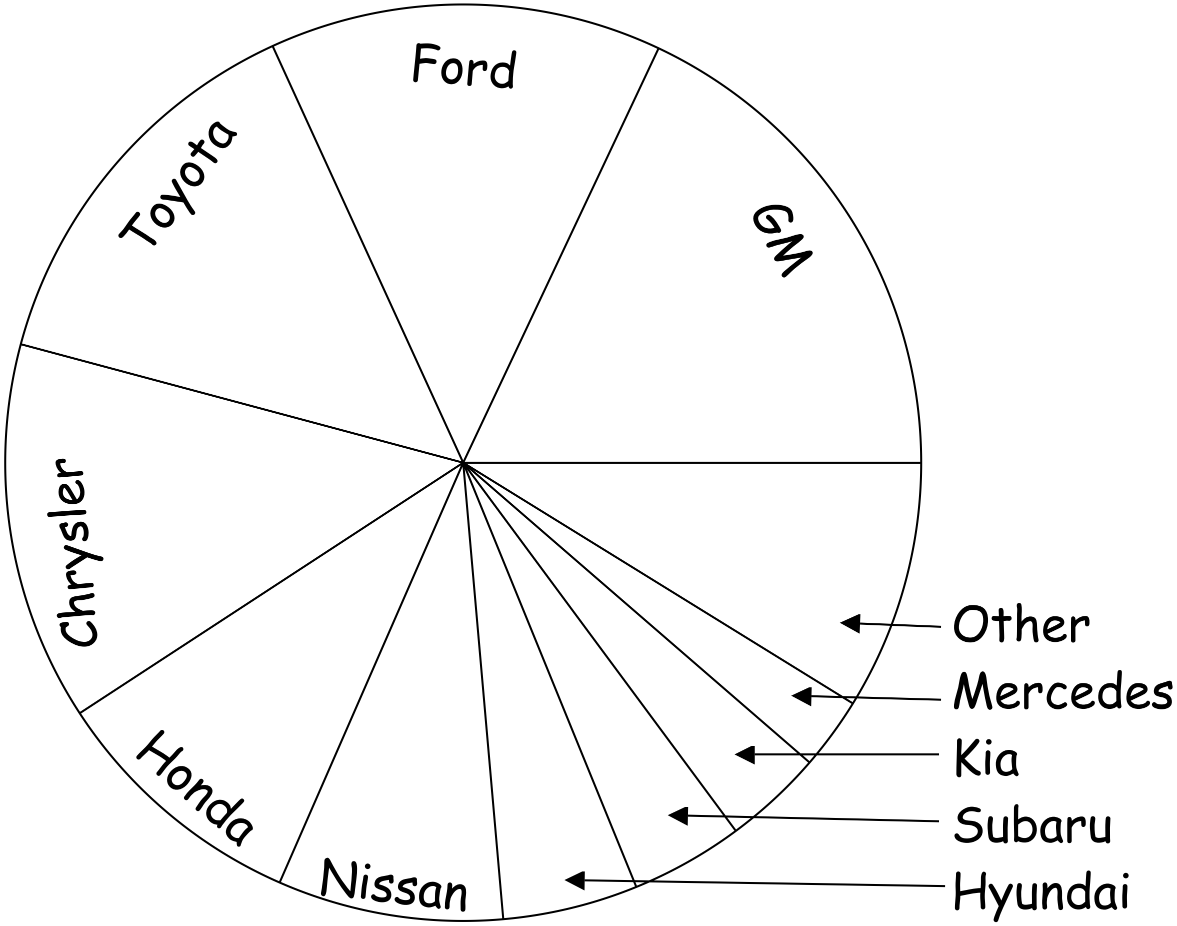


Histogram with density scale:



Distribution tables for *categorical data* are often represented as *Pie Charts*.

Example: Market share in the U.S. of automobile manufacturers.



Statistics and parameters

Tables, histograms and other charts are used to summarize large amounts of data. Often, an even more extreme summary is desirable.

- A number that summarizes ***population*** data is called a ***parameter***.
- A number that summarizes ***sample*** data is called a ***statistic***.

Observations:

- Population parameters are (more or less) ***constant.***
- Sample statistics ***vary with the sample***, i.e., their values depend on the particular sample chosen. A sample statistic can be thought of as a ***variable***.
- Sample statistics are *known* because we can compute them from the (available) sample data, while population parameters are often *unknown*, because data for the entire population is often unavailable.
- One of the most common uses of sample statistics is to ***estimate*** population parameters.

Measures of ***central tendency***

The most extreme way to summarize a list of numbers is with a single, *typical* value. The most common choices are the *mean* and *median*.

- The ***mean*** (*average*) of a set of numbers is the sum of all the values divided by the number of values in the set.
- The ***median*** of a set of number is the middle number, when the numbers are listed in increasing (or decreasing) order. The median splits the data into two equally sized sets—50% of the data lies below the median and 50% lies above.

  (If the number of numbers in the set is *even*, then the median is the average of the two middle values.)

The mean and median are different ways of describing the *center* of the data. Another statistic that is often used to describe the typical value is the ***mode***, which is the *most frequently occurring* value in the data.

Example. Find the mean, median and mode of the following set of numbers:

$$\{12, 5, 6, 8, 12, 17, 7, 6, 14, 6, 5, 16\}.$$

- The ***mean*** (average).

$$\frac{12+5+6+8+12+17+7+6+14+6+5+16}{12} = \frac{114}{12} = 9.5.$$

- The ***median***. Arrange the data in ascending order, and find the average of the middle two values in this case, since there are an even number of values:

$$5, 5, 6, 6, 6, 7, 8, 12, 12, 14, 16, 17 \ \longrightarrow\ \text{median} = \frac{7+8}{2} = 7.5.$$

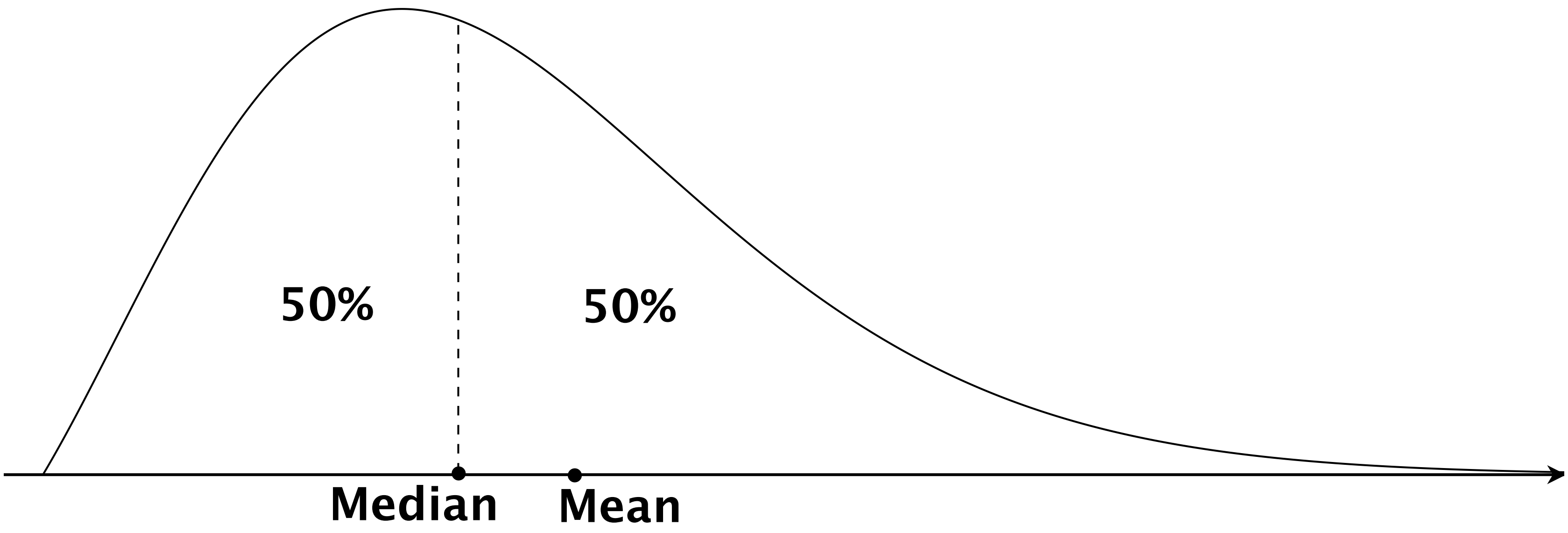
- The ***mode*** is 6, because 6 occurs most frequently (three times).

Comments:

- The mean is sensitive to ***outliers***—extreme values in the data (much bigger or much smaller than most of the data). Big outliers pull the mean up and small outliers pull the mean down.
- The median gives a better sense of 'middle' when the data is skewed in one direction or the other.
- The mean is easier to use in mathematical formulas.
- Both the median and the mean leave out a lot of information. E.g., each one separately tells us nothing about the *spread* of the data or where we might find 'peaks' (modes) in the distribution, etc.

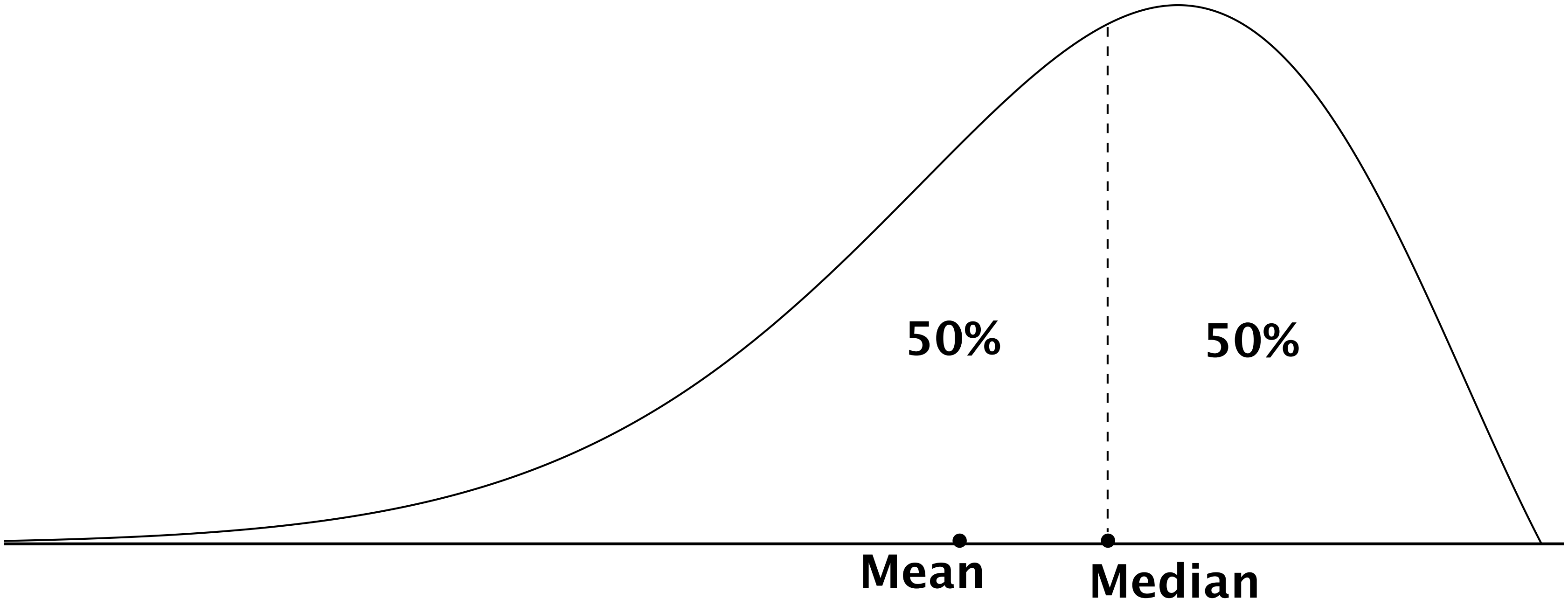
On the other hand, if we know both, then the *relative positions* of the mean and median provide some information about how the data is distributed...

In this histogram the mean is bigger than the median.



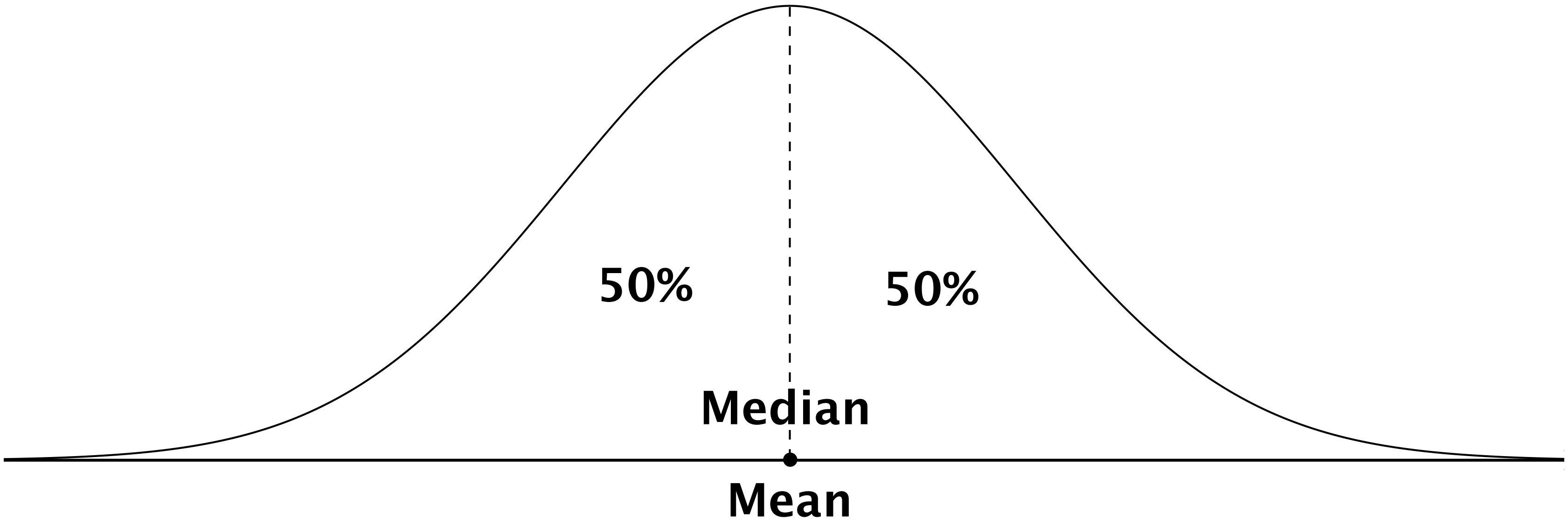
This is an indication that there are large outliers — the histogram has a longer tail on the right. We say that the data is *skewed to the right*.

In this histogram the mean is smaller than the median.

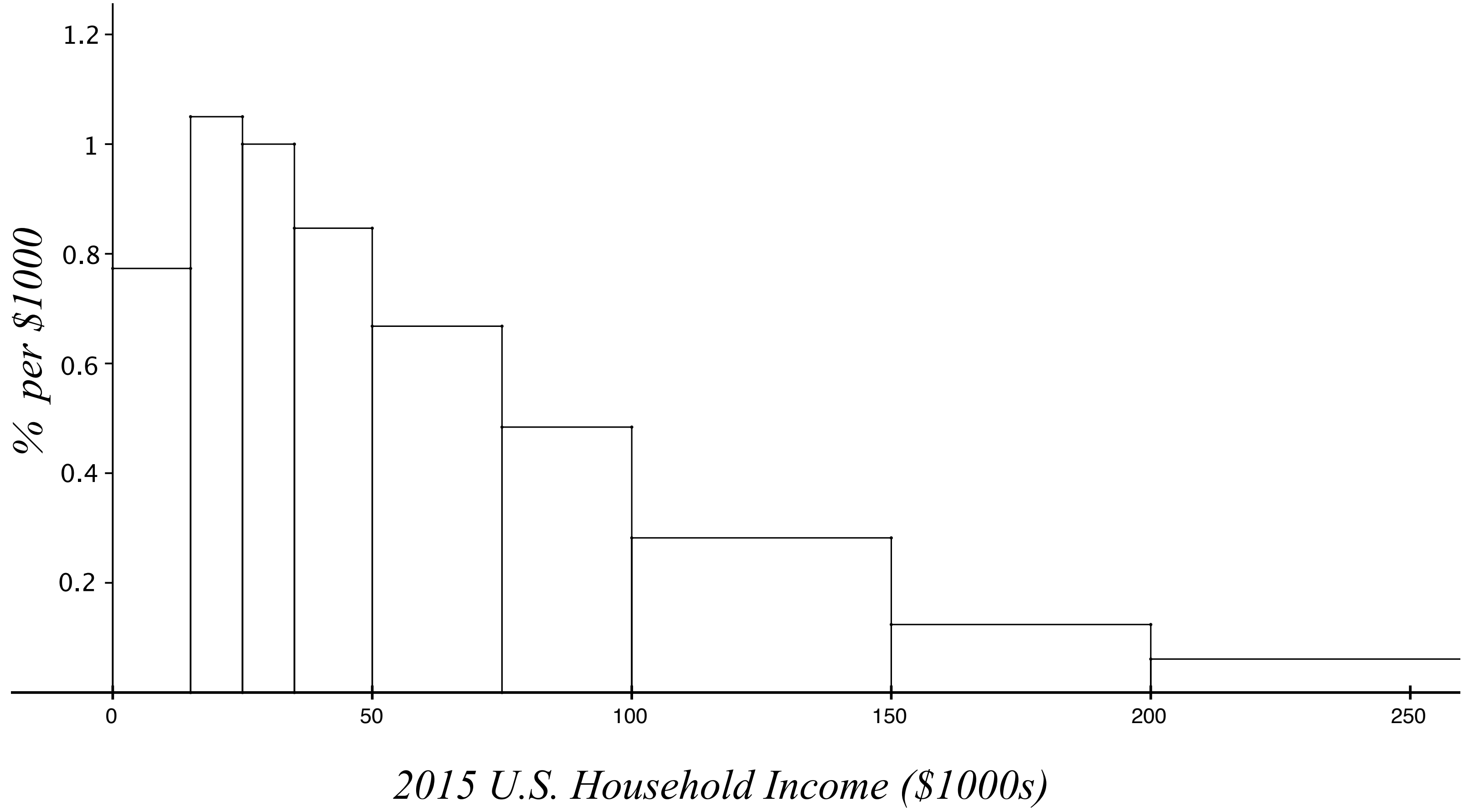


This is an indication that there are small outliers — the histogram has a longer tail on the left. We say that the data is *skewed to the left*.

If the mean and median are (more or less) equal, then the tails of the distribution are (more or less) the same, and the data has a (more or less) symmetric distribution around the mean/median, as depicted below.



Example: Here is the histogram that we constructed before:

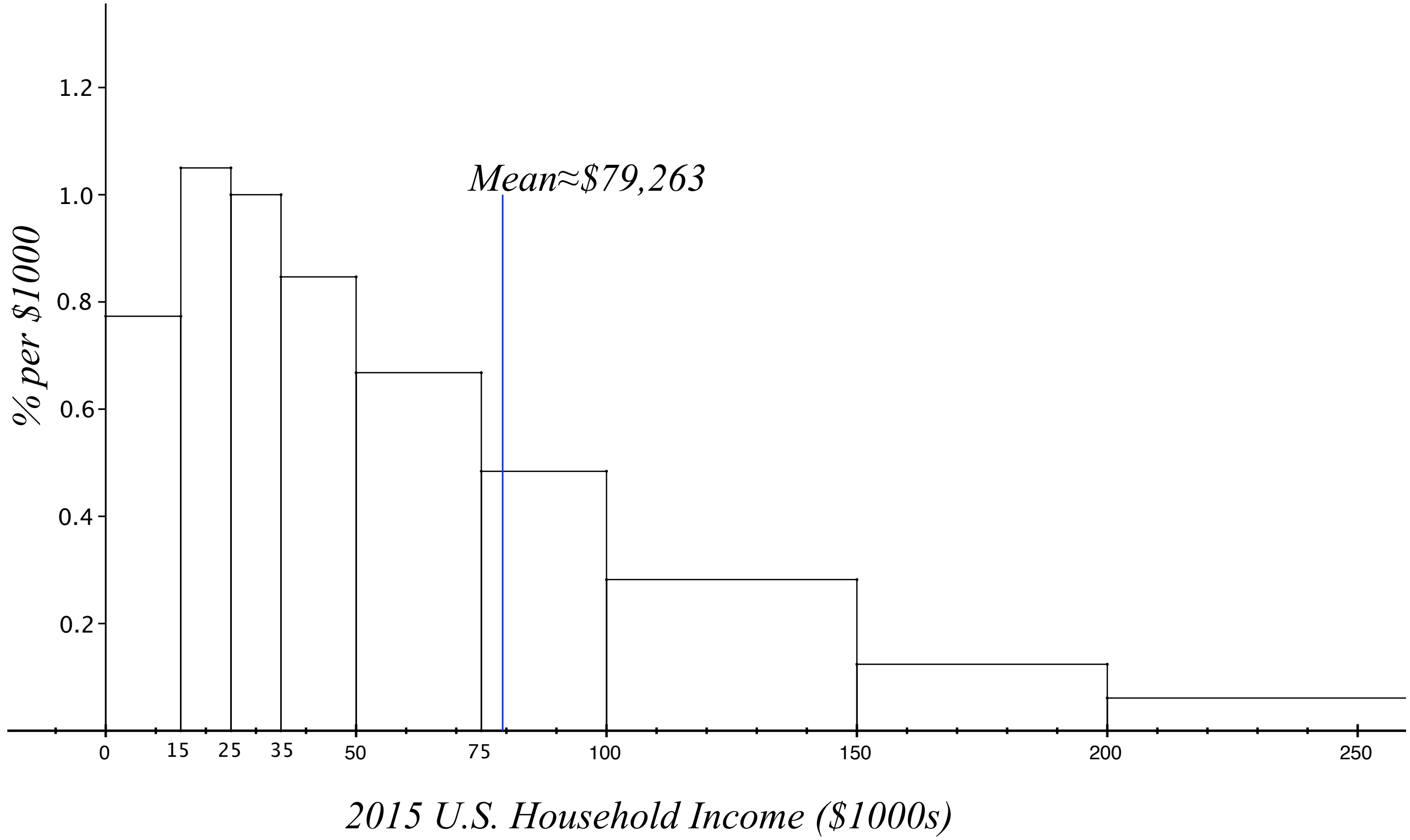


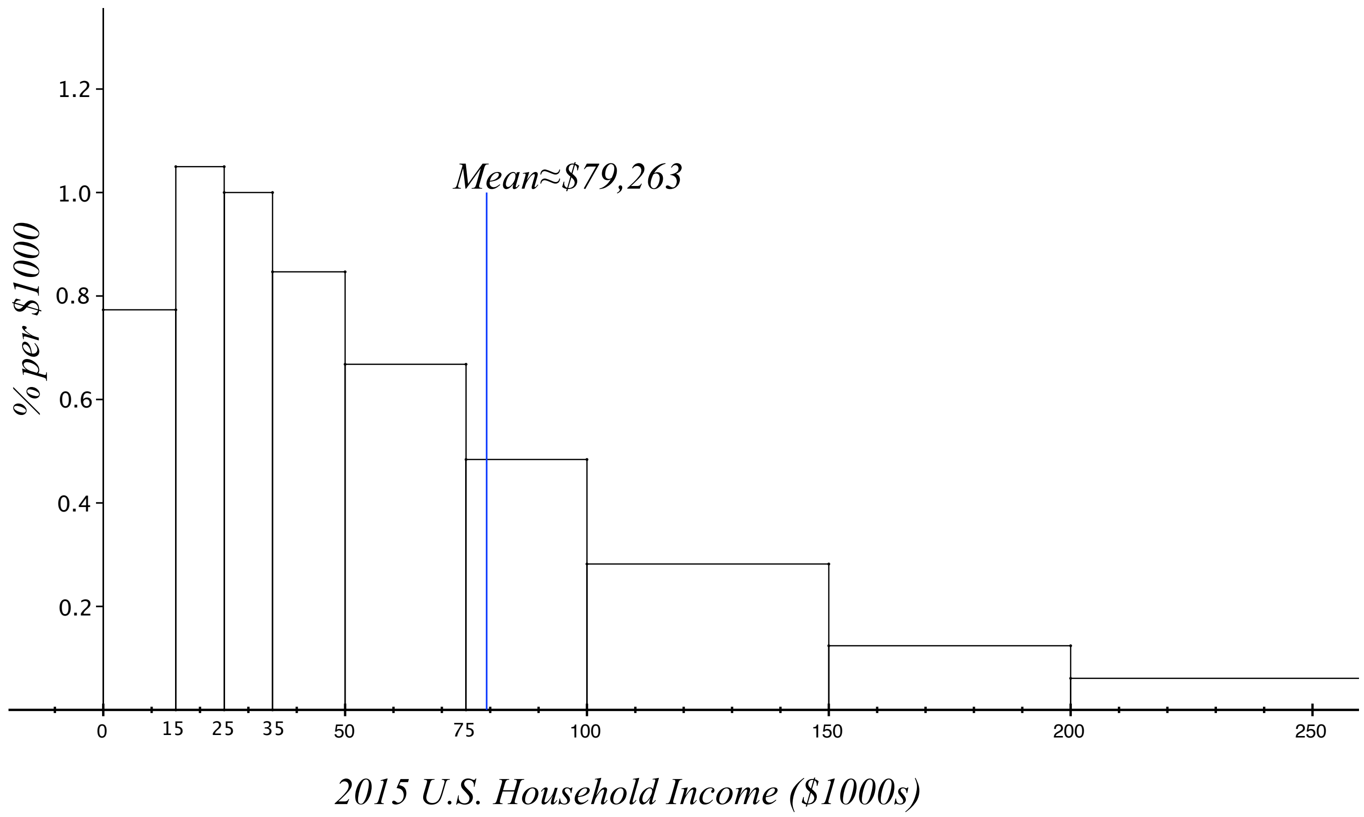
The histogram is skewed to the right, indicating that the mean will be larger than the median in this case.

(*) The mean income (estimated from the sample data) is about $79,263.

(*) We can find the (approximate) median by *reading* the histogram.

(*) Remember: the area of each bar represents the percentage of the population with income in the corresponding range. We find the areas of the bars, starting from the leftmost interval (0–15), and stop when we reach 50%.





0 to 15: $\approx 0.78 \frac{\%}{\$1000} \times \$15000 = 11.7\%$, 15 to 25: $\approx 1.05 \frac{\%}{\$1000} \times \$10000 = 10.5\%$

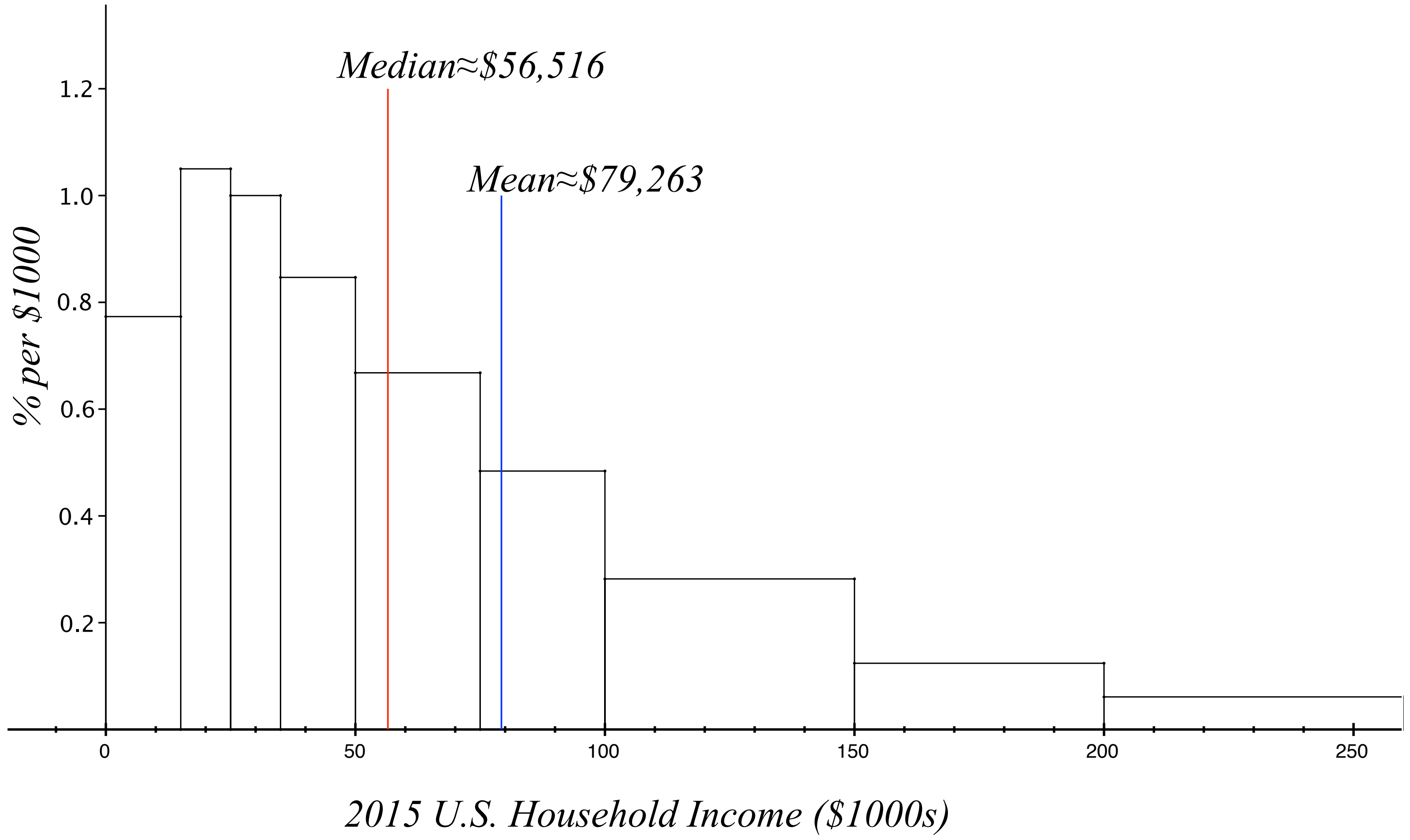
25 to 35: $\approx 1 \frac{\%}{\$1000} \times \$10000 = 10\%$, 35 to 50: $\approx 0.85 \frac{\%}{\$1000} \times \$15000 = 12.75\%$

0 to 50: area $\approx 11.7\% + 10.5\% + 10\% + 12.75\% = 44.95\%$... Need another 5%.

50 to 75: area $\approx 0.66 \frac{\%}{\$1000} \times \$25000 = 16.5\%$. Need to go a little less than one third the way from 50 to 75 to get another 5%...

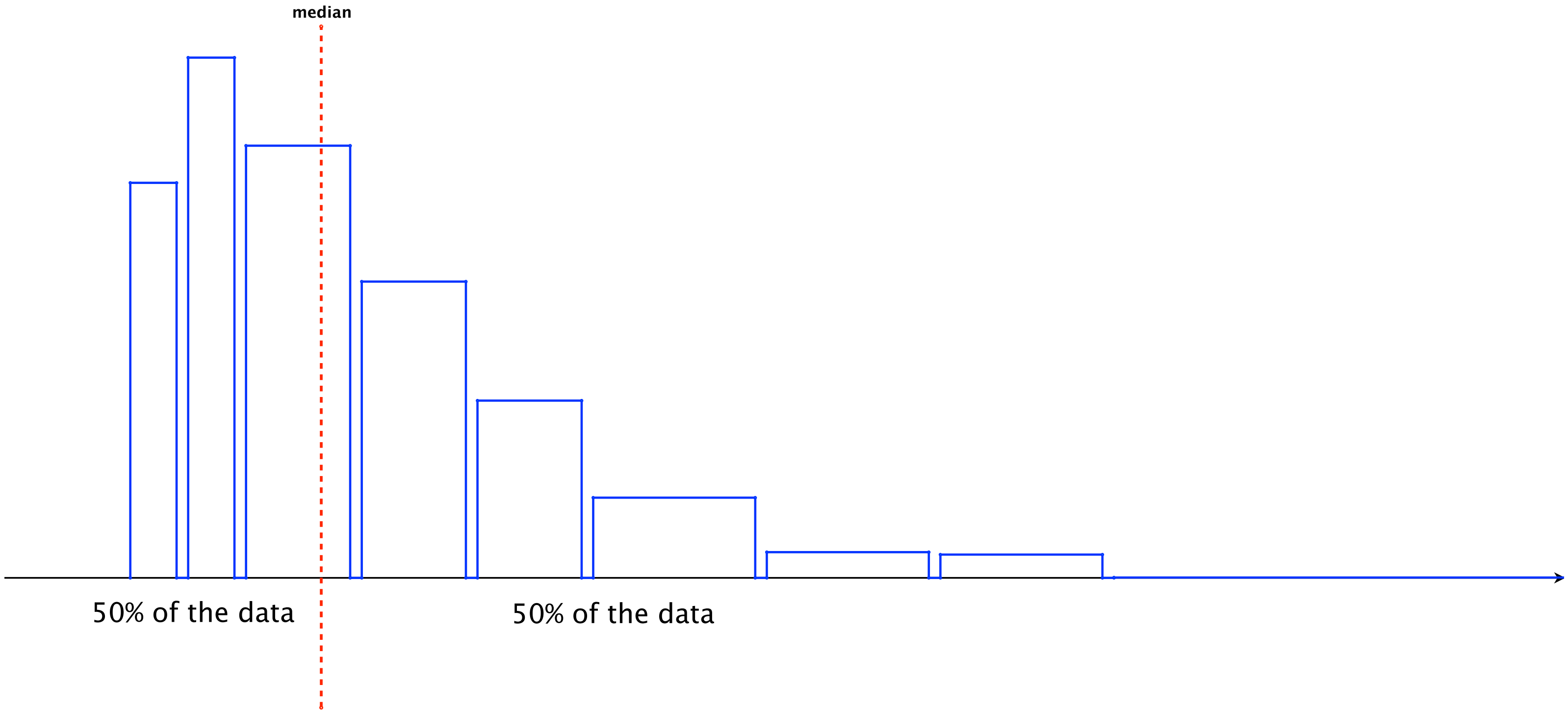
Median ≈ \$57, 500.

More precise estimate (using all of the survey data): Median ≈ \$56, 516

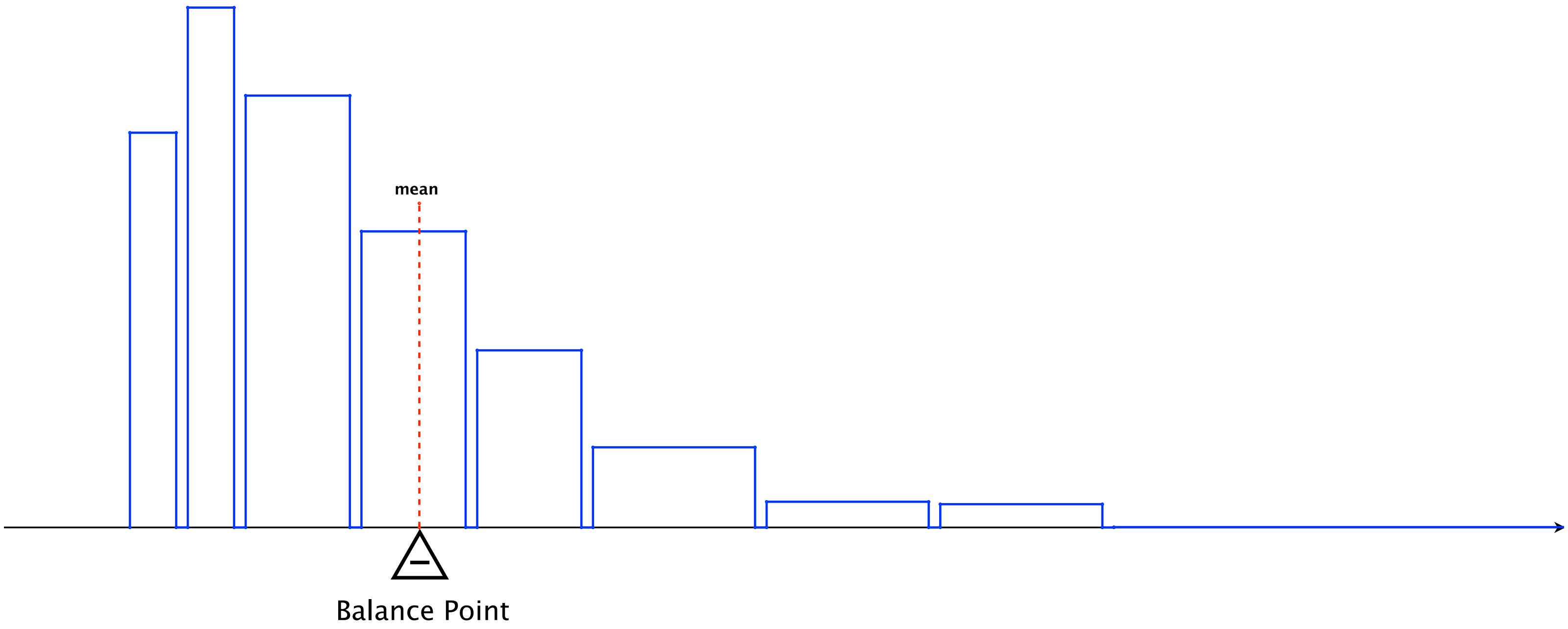


The mean and median describe the middle of the data in somewhat different ways:

- The median divides the histogram into two halves of equal area: it divides the data into two equal halves.



- The mean is the ‘balance point’ of the data:

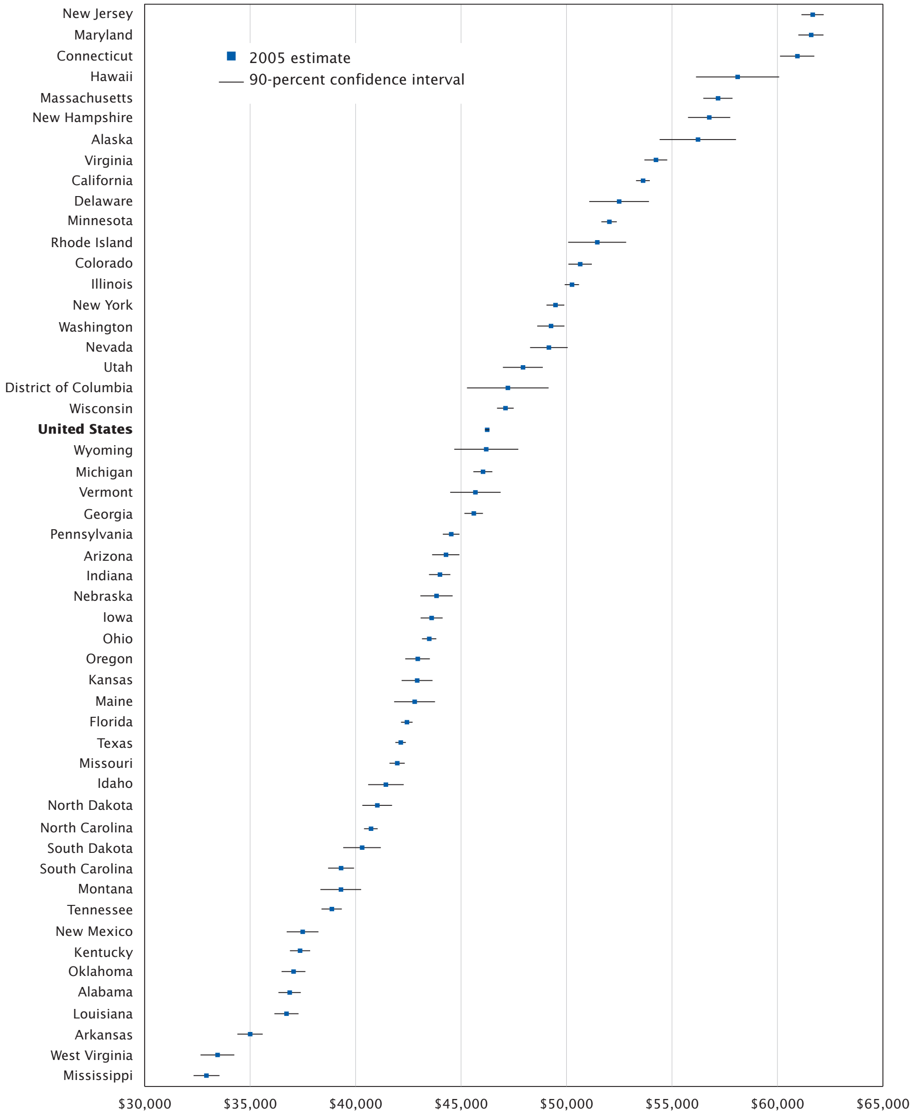


Averages and medians give a snapshot of a set of data. If the data comprises more than one variable, we can divide the data into categories with respect to one variable, and study the average/median of another variable in each category separately.

This allows researchers to discern relationships between different variables.

Example: The following graph comes from the 2005 American Community Survey of the US Census Bureau. It plots median household income by state.

Figure 1.
Median Household Income in the Past 12 Months With 90-Percent Confidence Intervals by State: 2005



Source: U.S. Census Bureau, 2005 American Community Survey.

Notational Interlude:

- The *population* mean (a parameter) is denoted by the Greek letter μ ('mu'). If there are several variables being studied, we put a subscript on the μ to tell us which variable it pertains to. For example, if we have data for population height (h) and population weight (w), the mean height would be denoted by μ_h and the mean weight by μ_w.
- The mean of a set of *sample* data (a statistic) is denoted by putting a bar over the variable. E.g., if $\{h_1, h_2, h_3, \ldots, h_n\}$ is a sample of heights, then the average of this sample would be denoted by $\overline{h}$.
- The median is usually denoted by m or M, and sometimes by Q_2 (more on this later).
- We can use ***summation notation*** to simplify the writing of (long) sums:

$$h_1 + h_2 + h_3 + \cdots + h_n = \sum_{j=1}^{n} h_j = \sum h_j.$$

For example we can write:

$$\begin{aligned}\overline{h} &= \frac{h_1 + h_2 + \cdots + h_n}{n} \\ &= \frac{1}{n}\left(h_1 + h_2 + \cdots + h_n\right) = \frac{1}{n}\sum h_j.\end{aligned}$$

Comment: The point of summation notation is to simplify expressions that involve sums with many terms, or in some cases, an unspecified number of terms. All the usual rules/properties of addition continue to hold. In particular

(i) $\sum (h_j \pm w_j) = \sum h_j \pm \sum w_j$

(ii) $\sum (a \cdot h_j) = a \cdot \left(\sum h_j\right)$

and

(iii) $\sum c = n \cdot c$ (here n is the number of constant terms).

Measuring the spread of the data

The mean and median describe the middle of the data. To get a better sense of how the data is *distributed*, statisticians also use '*measures of dispersion*'.

- The ***range*** is the distance between the smallest and largest values in the data.
- The ***interquartile range*** is the distance between the value separating the bottom 25% of the data from the rest and the value separating the top 25% of the data from the rest. In other words, it is the *range* of the middle 50% of the data.

Example: In the histogram describing household income distribution, about 25% of all households have incomes below $28,000 and about 25% of all households have incomes above $145,000, so the interquartile range is $\$145,000 - \$28,000 = \$117,000$.

The standard deviation: The standard deviation of a set of numbers is *something like* the average distance of the numbers from their mean. Technically, it is a little more complicated than that.

If $x_1, x_2, x_3, \ldots, x_n$ are numbers and $\overline{x}$ is their mean, then one candidate for measuring spread is the *average deviation* from the mean:

$$\frac{(x_1-\overline{x})+(x_2-\overline{x})+\cdots+(x_n-\overline{x})}{n}=\frac{1}{n}\sum(x_j-\overline{x}).$$

Potential problem: positive terms and negative terms in the sum can cancel each other out... How much cancellation?

$$\begin{aligned}\frac{1}{n}\sum(x_j-\overline{x}) &= \frac{1}{n}\sum x_j-\frac{1}{n}\sum\overline{x}\\ &= \overline{x}-\frac{\overbrace{\overline{x}+\overline{x}+\cdots+\overline{x}}^{n}}{n}\\ &= \overline{x}-\frac{\cancel{n}\cdot\overline{x}}{\cancel{n}}=\overline{x}-\overline{x}=0\end{aligned}$$

Complete cancellation!

Instead, statisticians use the ***standard deviation***, which is given by

$$SD_x = \sqrt{\frac{1}{n}\sum(x_j - \overline{x})^2}.$$

In words, the SD is the ***root of the mean of the squared deviations of the numbers from their mean***.

(*) Squaring the deviations fixes the cancellation problem...

(*) ... but exaggerates both very small deviations (making them smaller) and very large deviations (making them bigger)...

(*) ... and also *changes the scale* (e.g., from inches to squared inches).

(*) Taking the square root of the average squared deviation fixes both of these problems (to a certain extent).

(*) If a lot of the data is far from the mean, then many of the $(x_j - \overline{x})^2$ terms will be quite large, so the mean of these terms will be large and the SD of the data will be large.

(*) In particular, outliers can make the SD bigger. (Outliers have an even bigger effect on the *range* of the data.)

(*) On the other hand, if the data is all clustered close to the mean, then all of the $(x_j - \overline{x})^2$ terms will be fairly small, so their mean will be small and the SD will be small.

Example: Find the SD of the set $\{x_j\} = \{2, 4, 5, 8, 5, 11, 7\}$.

- Step 1. Find the mean: $\overline{x} = \dfrac{2+4+5+8+5+11+7}{7} = \dfrac{42}{7} = 6$.
- Step 2. Find the mean of the squared deviations of the numbers from their mean:

$$\frac{(2-6)^2 + (4-6)^2 + (5-6)^2 + \cdots + (7-6)^2}{7} = \frac{52}{7}.$$

- Step 3. $SD_x = \sqrt{52/7} \approx 2.726$.

(*) (Very) useful shortcut (for calculations done by hand):

$$\boxed{\frac{1}{n}\sum(x_j-\overline{x})^2=\left(\frac{1}{n}\sum x_j^2\right)-(\overline{x})^2}$$

so

$$SD_x=\sqrt{\frac{1}{n}\sum(x_j-\overline{x})^2}=\sqrt{\left(\frac{1}{n}\sum x_j^2\right)-(\overline{x})^2}$$

Check with example:

$\{x_j\}=\{2,4,5,8,5,11,7\}$ and $\overline{x}=6$:

$$\left(\frac{1}{7}\sum x_j^2\right)-\overline{x}^2=\frac{304}{7}-36=\frac{52}{7}\quad\checkmark$$

Very useful special case: *All the numbers in the data are* 0*s and* 1*s.*

- m 1s and $n - m$ 0s (n numbers in all).

$$\Rightarrow\ \overline{x} = \frac{\overbrace{1 + 1 + \cdots + 1}^{m} + \overbrace{0 + 0 + \cdots + 0}^{n-m}}{n} = \frac{m}{n}$$

(I.e., the average is equal to the ***proportion*** *of* 1*s in the data).*

$$\Rightarrow\ SD_x = \sqrt{\frac{\overbrace{1^2 + 1^2 + \cdots + 1^2}^{m} + \overbrace{0^2 + 0^2 + \cdots + 0^2}^{n-m}}{n} - \left(\frac{m}{n}\right)^2}$$

$$= \sqrt{\frac{m}{n} - \left(\frac{m}{n}\right)^2} = \sqrt{\frac{m}{n}\left(1 - \frac{m}{n}\right)}$$

$$= \sqrt{\frac{m}{n} \cdot \frac{n-m}{m}}$$

$$= \sqrt{(\text{proportion of 1s}) \cdot (\text{proportion of 0s})}$$