

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. It is useful (and productive) to think of sample statistics as a *variables*.
- 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:

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

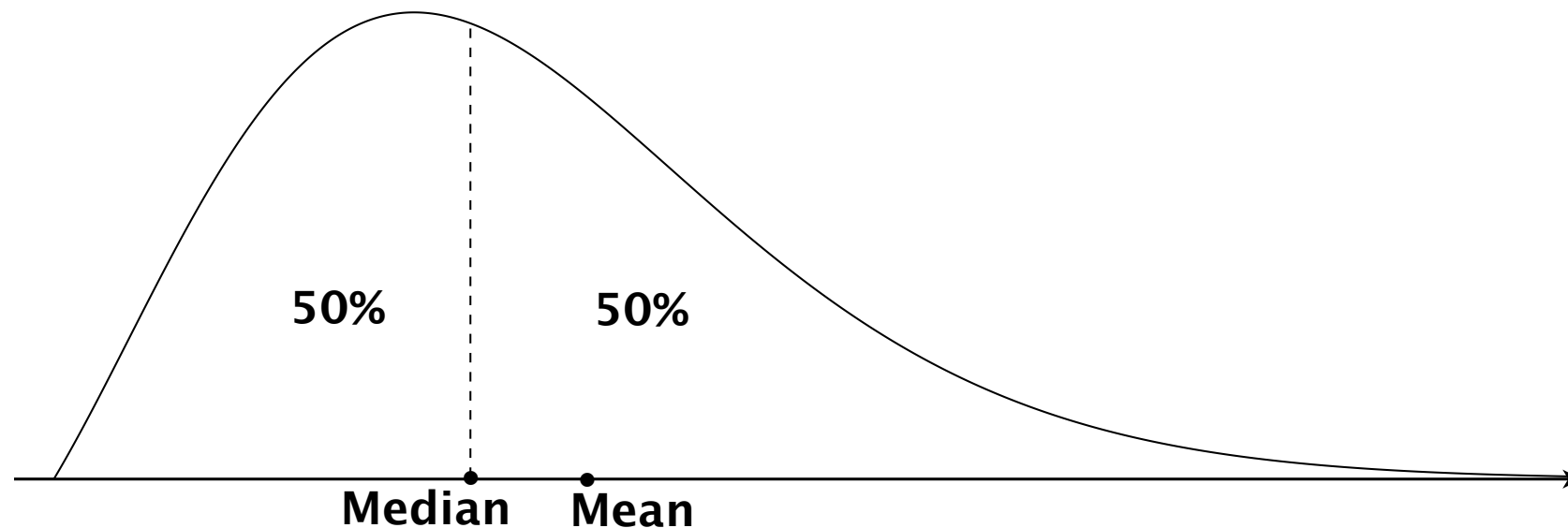
- The *mode* is 6, because 6 occurs most frequently (three times).

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. Most importantly, neither one says anything about how the data is *spread* out.

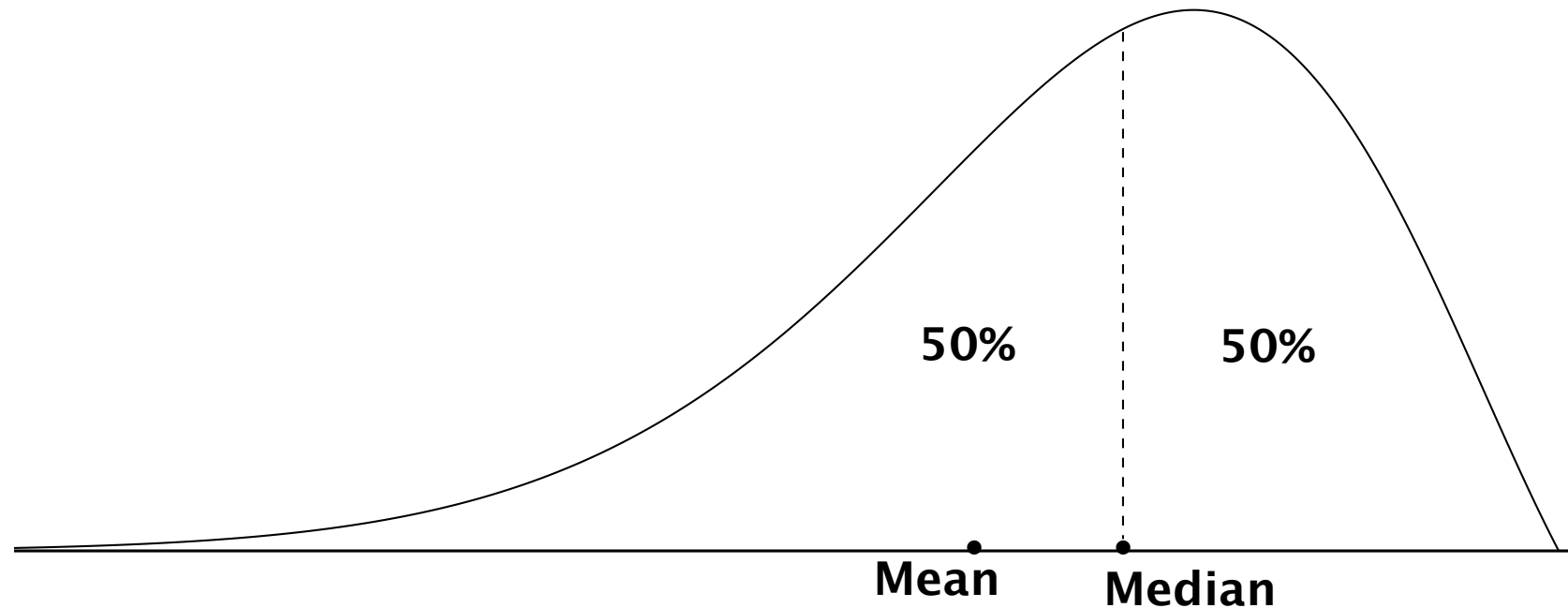
On the other hand, if we know both, then the *relative positions* of the mean and median can provide some information about how the data is distributed...

In this histogram the mean is bigger than the median.



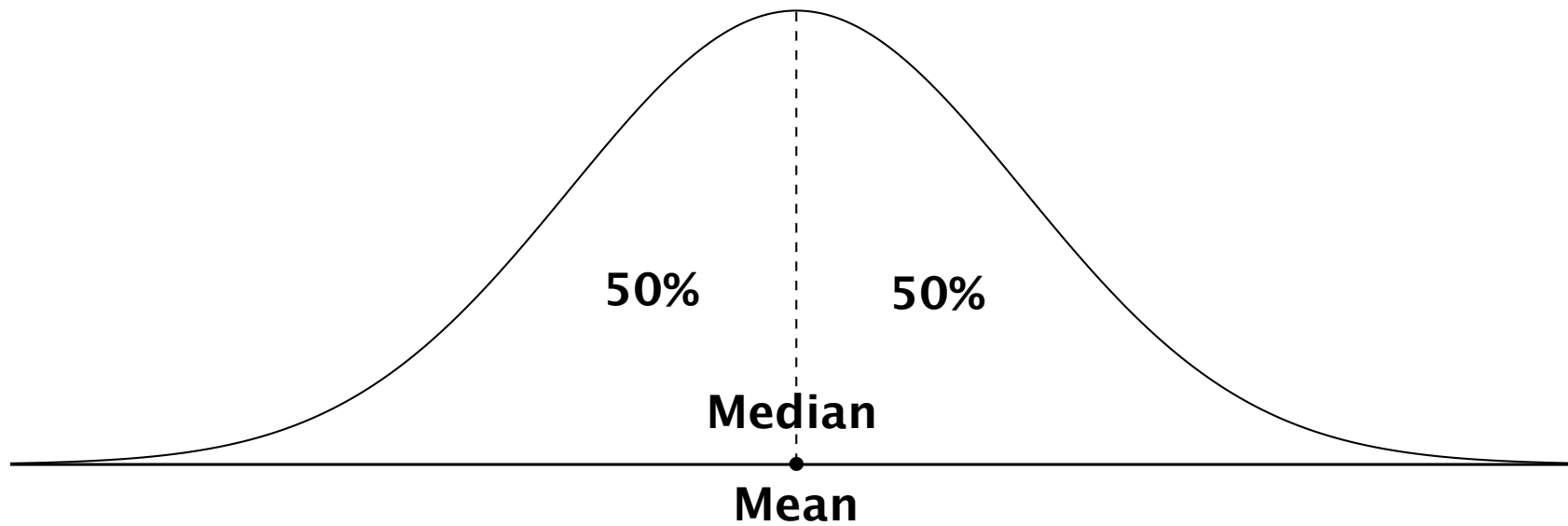
This is an indication that there are large outliers — the histogram has a longer tail on the right. We say that the data is *skewed to the right*.

In this histogram the mean is smaller than the median.

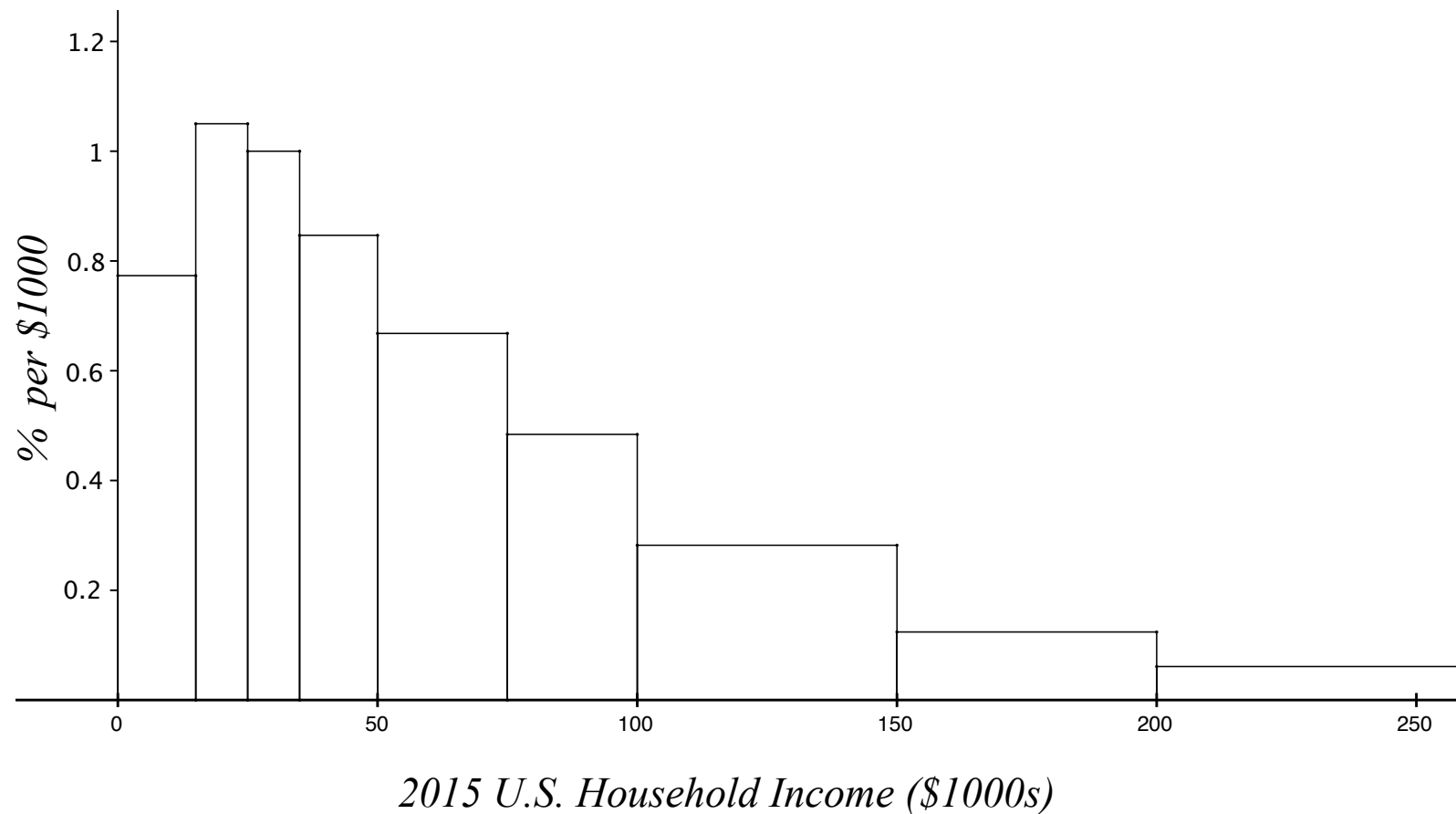


This is an indication that there are small outliers — the histogram has a longer tail on the left. We say that the data is *skewed to the left*.

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.

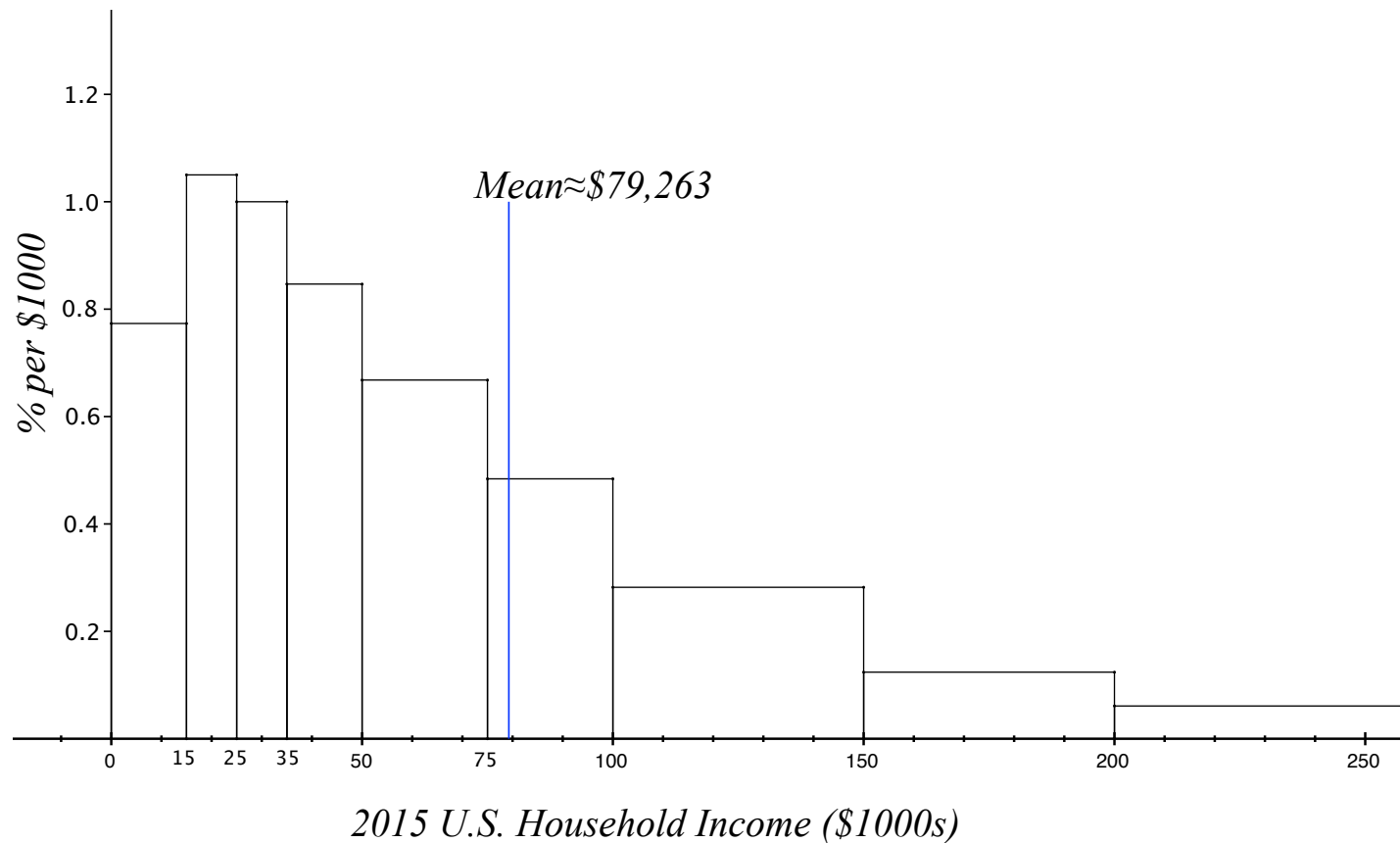


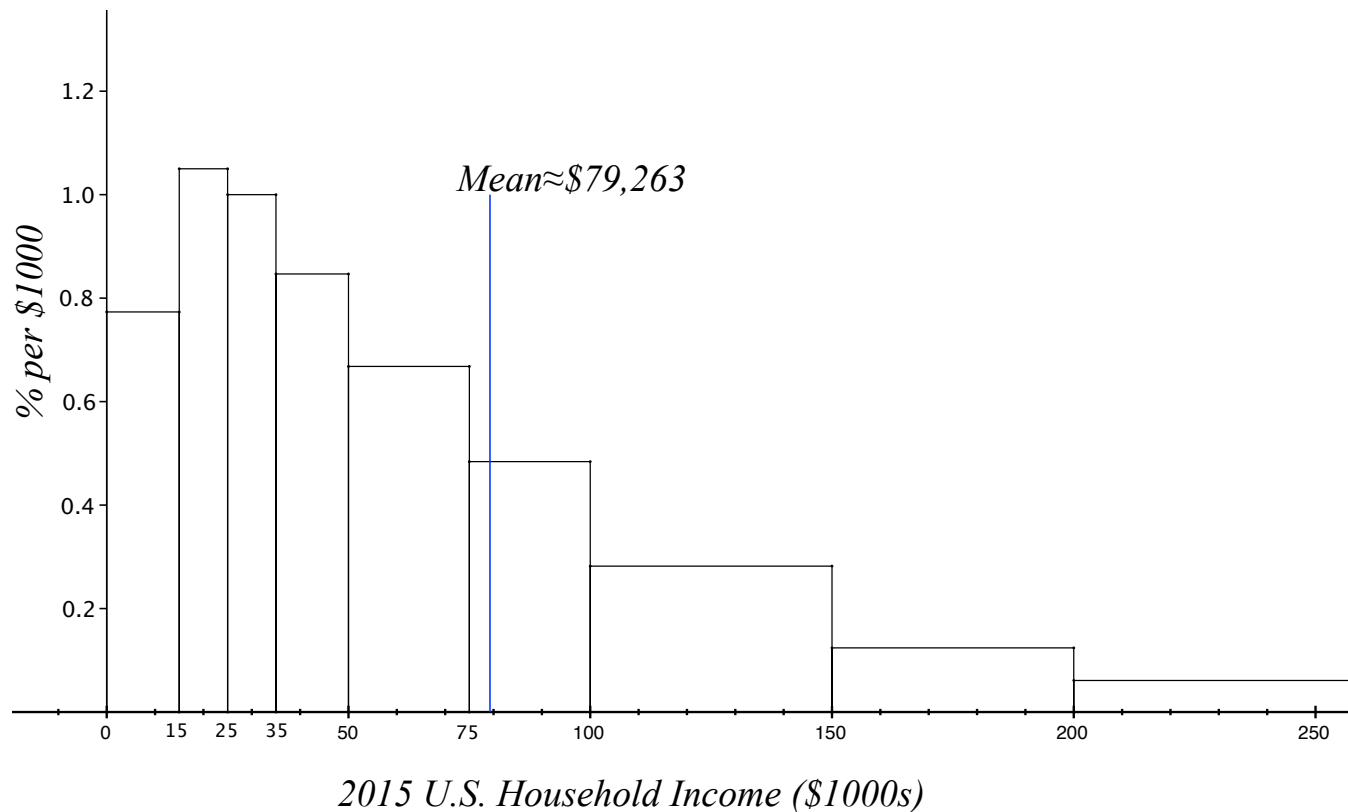
Example: Here is the histogram that we constructed before:



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: $\approx 1.05 \frac{\%}{\$1000} \times \$10000 = 10.5\%$

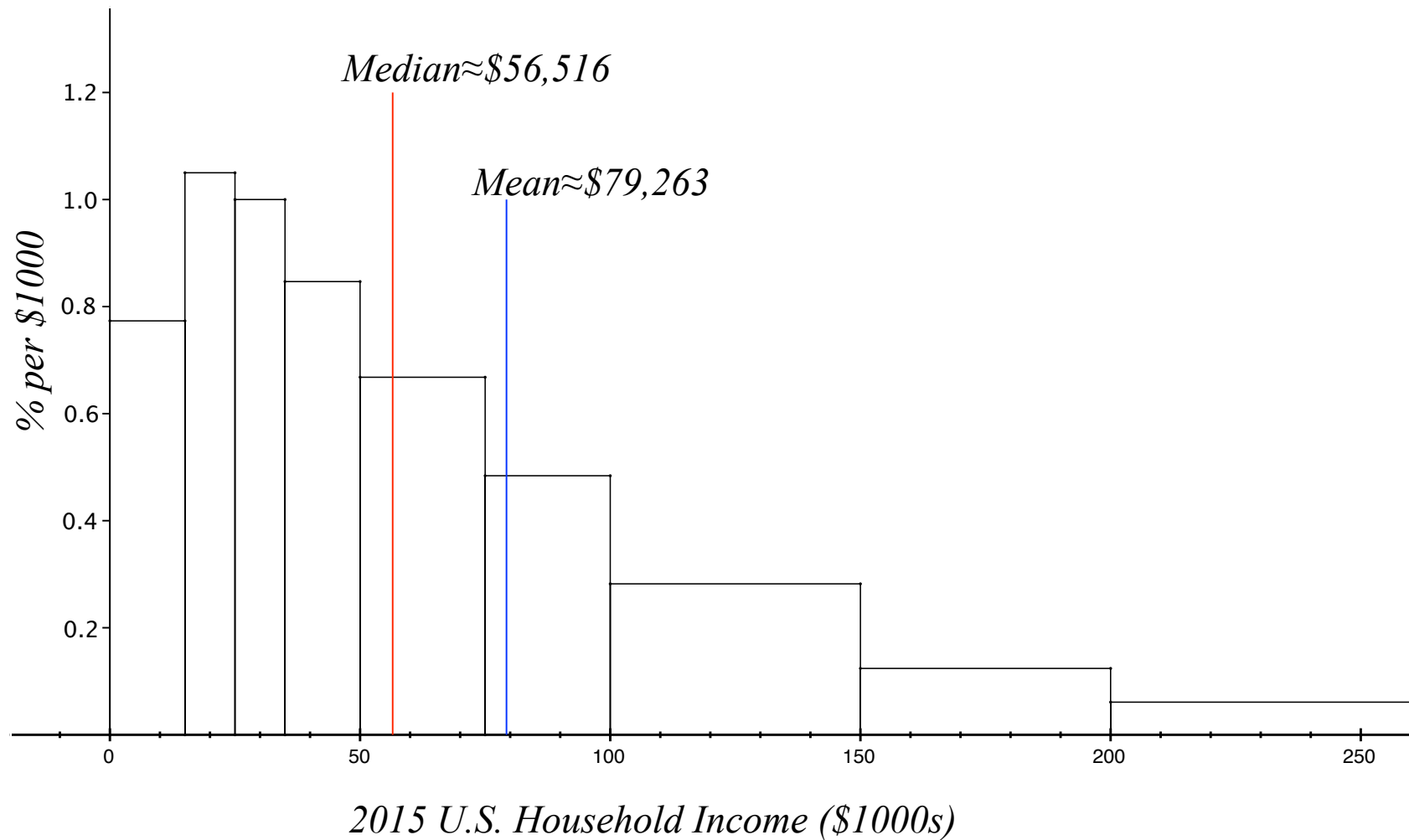
25 to 35: $\approx 1 \frac{\%}{\$1000} \times \$10000 = 10\%$, 35 to 50: $\approx 0.85 \frac{\%}{\$1000} \times \$15000 = 12.75\%$

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%...

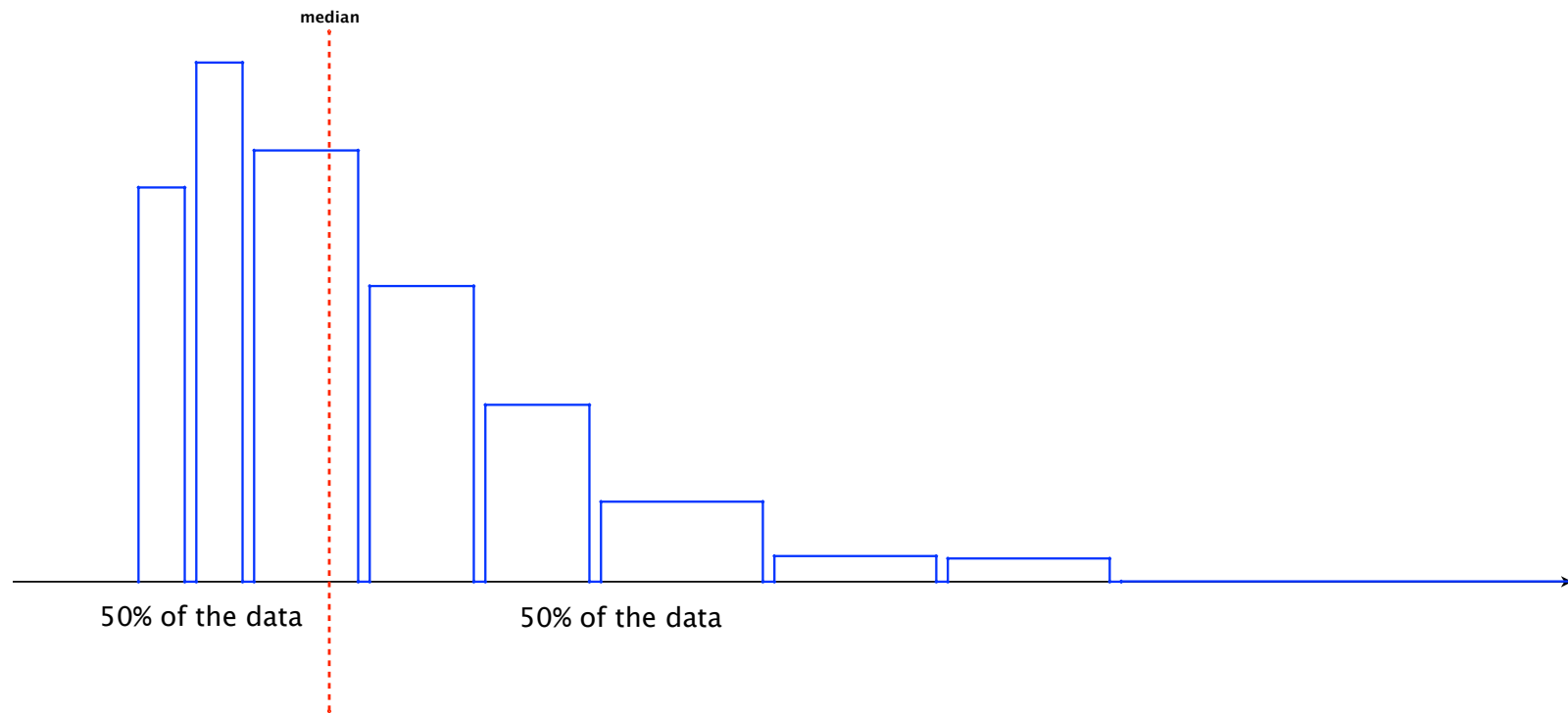
Median $\approx \$57,500$.

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

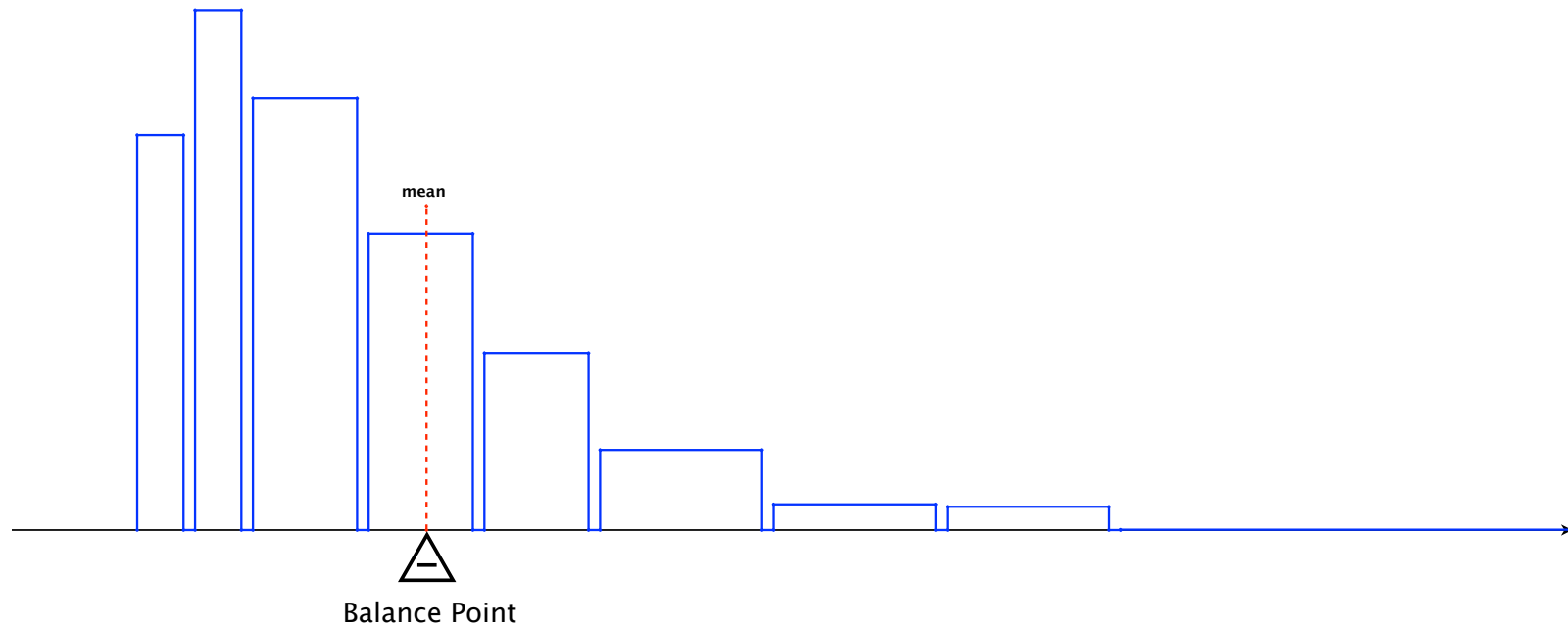


The mean and median describe the middle of the data in somewhat different ways:

- 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:



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 $\{h_1, h_2, h_3, \dots, h_n\}$ is a sample of heights, then the average of this sample would be denoted by \bar{h} .
- The median is usually denoted by m or M , and sometimes by Q_2 .
- 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:

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

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

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

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

and

$$(iii) \quad \sum c = n \cdot c \quad \text{in this formula, all the terms are the same (equal to the constant } c) \text{ and } n \text{ is the number of terms.}$$

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 (the first *quartile*, Q_1) and the value separating the top 25% of the data from the rest (the third *quartile*, Q_3). In other words, $Q_3 - Q_1$ is the *range* (or spread) of the middle 50% of the data. (The second quartile, Q_2 , is the median).

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

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

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

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

Complete cancellation!

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

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

In words, the standard deviation 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).

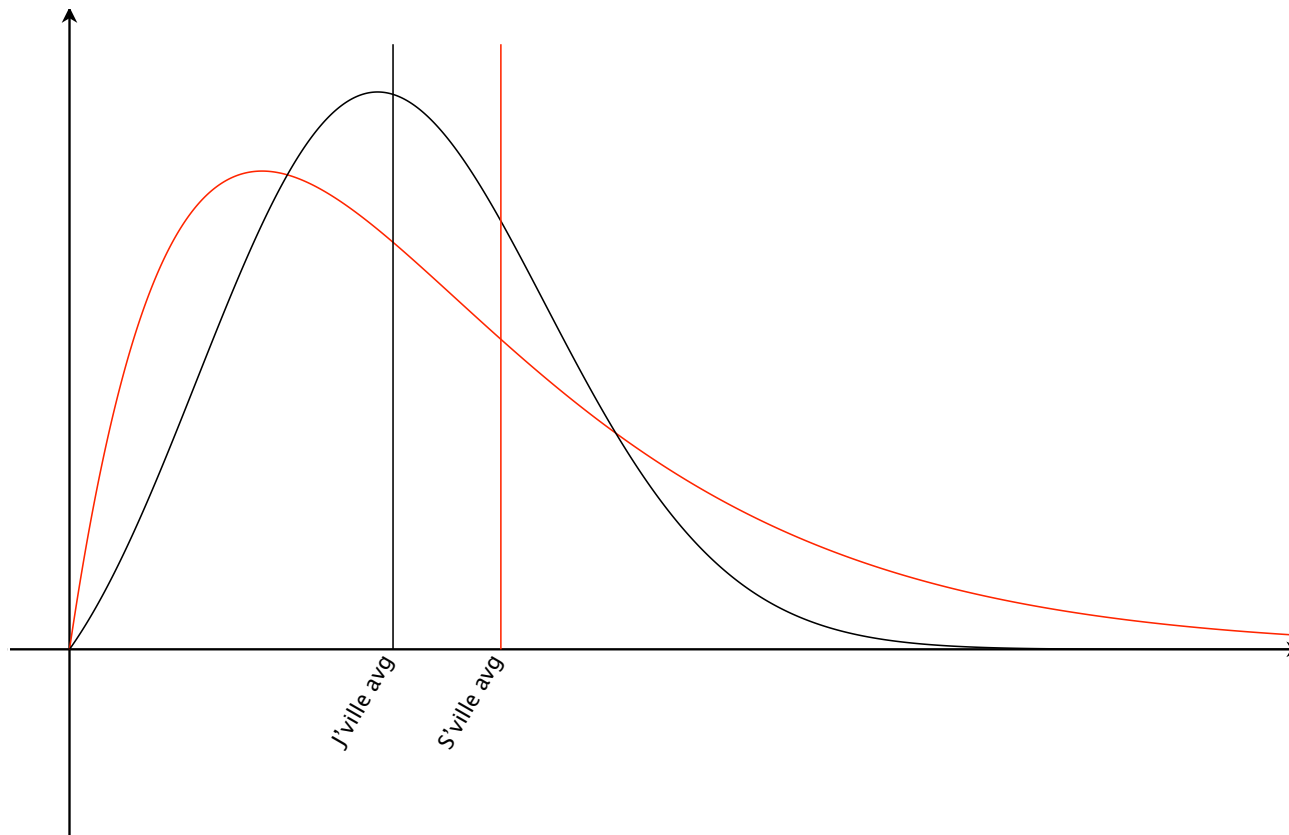
(*) Taking the square root of the average squared deviation fixes both of these problems (to a certain extent).

(*) **Notation:** The standard deviation of a set of population data is traditionally denoted by σ (lower case sigma). The standard deviation of a set of sample data is traditionally denoted by s . Our book uses SD for both (but sometimes SD^+ for sample standard deviation).

Example. (Hypothetical) In Jonesville (population 100,000), the average annual household income is \$70,000 with a standard deviation of \$20,000. In Smithsville (population 200,000) the average annual household income is \$90,000 with a standard deviation of \$60,000.

Question: Where is there more income inequality?

Answer: Smithsville. There is greater spread around the average income, indicating a bigger range of incomes.



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

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

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

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

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

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

so

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

Check with example:

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

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

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

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

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

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

$$\begin{aligned}\Rightarrow SD_x &= \sqrt{\frac{\overbrace{1^2 + 1^2 + \dots + 1^2}^m + \overbrace{0^2 + 0^2 + \dots + 0^2}^{n-m}}{n} - \left(\frac{m}{n}\right)^2} \\ &= \sqrt{\frac{m}{n} - \left(\frac{m}{n}\right)^2} = \sqrt{\frac{m}{n} \left(1 - \frac{m}{n}\right)} \\ &= \sqrt{\frac{m}{n} \cdot \frac{n-m}{n}} \\ &= \sqrt{(\text{proportion of 1s}) \cdot (\text{proportion of 0s})}\end{aligned}$$