The standard deviation

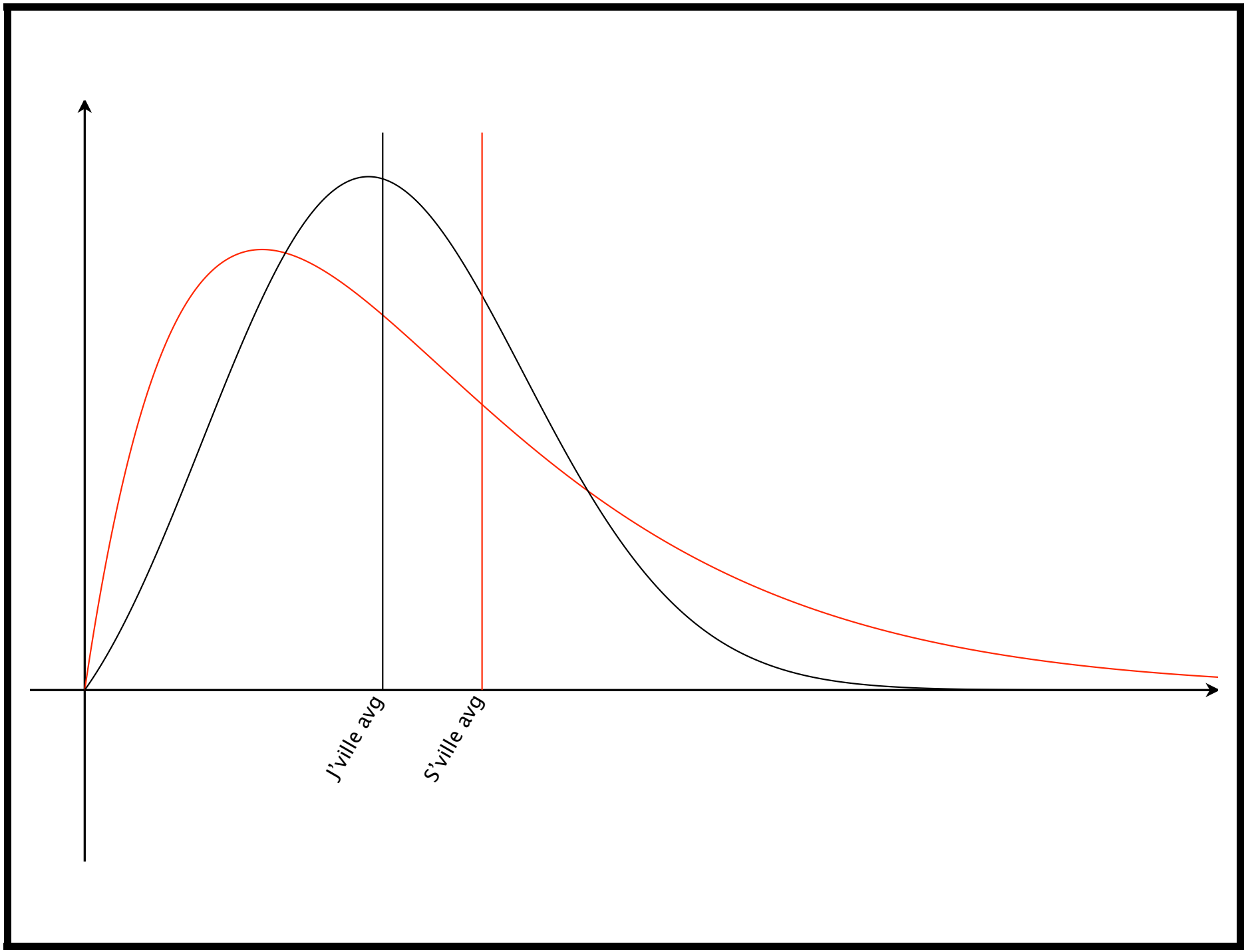
(*) $SD = \sqrt{\frac{1}{n}\sum(x_j - \overline{x})^2}$... Measures the spread of the data around the average.

Example 1. (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 disparity?

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

(*) The histogram for household income in Smithsville is more skewed to the right. The histogram for household income in Jonesville is likely to be more symmetric (but still skewed to the right).


J'ville avg
S'ville avg

Technical facts:

- Useful shortcut for calculations done by hand. First:

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

because arithmetic.

This means that

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

- Even shorter shortcut, in an important special case.

If $\{u_j\} = \{\overbrace{1, 1, \ldots, 1}^{m}, \overbrace{0, 0, \ldots, 0}^{n-m}\}$ (n numbers in all), then $\overline{u} = m/n$ (= the proportion of 1s), and

$$SD_u = \sqrt{\frac{m}{n} \cdot \frac{n-m}{n}} = \sqrt{(\text{proportion of 1s}) \cdot (\text{proportion of 0s})}$$

- The SD and the mean are both sensitive to scale:

If $\{x_1, x_2, \ldots, x_n\}$ is a set of numbers and $y_j = a \cdot x_j + b$ for each $1 \leq j \leq n$, then

$$\overline{y} = a \cdot \overline{x} + b \qquad \text{and} \qquad SD_y = a \cdot SD_x.$$

Examples:

(*) If the average weight of a group of students is 148 lbs, then the average weight, measured in kilograms, of the same students is $148/2.2 \approx 67.27$ kg.

(*) If the standard deviation of the weights is 15 lbs, then in kilograms, the standard deviation is $15/2.2 \approx 6.82$ kg.

(*) If the average temperature in January Podunk is 29° fahrenheit, with a standard deviation of 5° (fahrenheit), then in degrees celsius the average is $(29 - 32) \cdot \frac{5}{9} = -\frac{5}{3} \approx -1.67°$ and the standard deviation in degrees celsius is $5 \cdot \frac{5}{9} \approx 2.78°$.

(*) The SD and the mean together tell us where most of the data lies.

Chebyshev's Inequality: In ***any set of data***, most of the numbers lie within two or three SDs of the average. Specifically,

> *the proportion of the data that lies more than k SDs away from average is* ***always less than*** $\dfrac{100\%}{k^2}$.

- Less than $\dfrac{100\%}{4} = 25\%$ of the numbers in any data set is more than 2 SDs away from average.

  This means that *at least* 75% of the numbers *in any data set* can be found within 2 SDs of the average.

- Less than $\dfrac{100\%}{9} \approx 11.11\%$ of the numbers in any data set is more than 3 SDs away from average.

  This means that *at least* 88.88% of the numbers *in any data set* can be found within 3 SDs of the average.

Example 1. (continued)

(*) In Jonesville, at least 75% of the households have income between \$30,000 ($= \$70,000 - 2 \cdot \$20,000$) and \$110,000 ($= \$70,000 + 2 \cdot \$20,000$).

(*) In Smithsville, at least 75% of the households have income between \$0 (?) and \$210,000 ($= \$90,000 + 2 \cdot \$60,000$).