The standard deviation

- \( SD = \sqrt{\frac{1}{n} \sum (x_j - \bar{x})^2} \)
- The SD gives a measure of how the data are \textit{clustered around the mean}.
  - If the SD is larger, then the data are more spread out — we are more likely to find data that are relatively far from the mean.
  - If the SD is smaller, then the data is more tightly clustered around the mean — most of the data are fairly close to the average.
- But the SD and the Mean are both \textit{sensitive to scale}.
  If \( y_j = c \cdot x_j \), for \( j = 1, \ldots n \), then
  - \( \bar{y} = c \cdot \bar{x} \), and
  - \( SD_y = |c| \cdot SD_x \).
Chebyshev’s Inequality: For any data set, most of the data lies within several SDs of the mean. Specifically, \( \text{the proportion of the data that lies more than } k \text{ SDs away from the mean is always less than } \frac{100\%}{k^2} \).

E.g., for any data set:

- less than \( \frac{100\%}{4} \) = 25% of the data is more than 2 SDs away from average;
- less than \( \frac{100\%}{9} \) \approx 11.11% of the data is more than 3 SDs away from average;
- less than \( \frac{100\%}{16} \) = 6.25% of the data is more than 4 SDs away from average;
- less than \( \frac{100\%}{25} \) = 4% of the data is more than 5 SDs away from average; etc.
Example (from the book): $\bar{h} = 63.5$ inches and $SD_h \approx 3$ inches...

Figure 8. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within one SD of the average is shaded: 72% of the women differed from average by one SD (3 inches) or less.
Example (continued): $\bar{h} = 63.5$ inches and $SD_h \approx 3$ inches...

Figure 9. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within two SDs of the average is shaded: 97% of the women differed from average by two SDs (6 inches) or less.
Standard units

To compare the distributions of different data sets, we can put both on the same scale, by converting the data in each distribution to standard units—the number of standard deviations the data lies above or below the average for its distribution.

If \( x_j \) comes from a distribution with average \( \bar{x} \) and standard deviation \( SD_x \), we convert \( x_j \) to its standard units, \( z_j \), by setting

\[
    z_j = \frac{x_j - \bar{x}}{SD_x}.
\]

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.

\( z_j \) tells us how many SDs \( x_j \) is from \( \bar{x} \), and whether \( x_j \) is above (\( z_j > 0 \)) or below (\( z_j < 0 \)) average.
The normal approximation, I

- Different sets of data may be seen to have very similar distributions, *once they have been converted to standard units*.

- Converting to standard units moves the center of the histogram (the average of the data) to 0, and scales the data as a whole so that one SD is converted to 1 unit.

- In many cases, the histogram of the data in standard units takes on a somewhat *bell-shaped* form — the form of the *normal curve*.

The normal curve is the graph of the function

\[ y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \]

(where \( e = 2.7182818 \ldots \)). This graph appears on the next slide.
The normal curve is symmetric around the line $x = 0$, and the total area under the curve is equal to 1 (or 100%, if you prefer).
Example: The distribution of heights of women age 18 and over in HANES5 (Health and Nutrition Examination Study, ’03 - ’04) appears in the histogram below (from page 81 in chapter 5 of FPP).

The average height is 63.5 and the SD is about 3. The shaded region represents the heights that fall within one SD of average.
To see how well the distribution of the height data is approximated by the normal curve, we must convert the data to standard units and sketch the histogram for the ‘standardized’ (or normalized) data. To save a lot of drawing time, we observe that the conversion to standard units is just a *rescaling*. This means that instead of actually converting all of the heights to their standard units and then drawing a new histogram, we can simply *change the horizontal and vertical scales on the original histogram*. This is done in the figure (also from page 81 in FPP) on the next slide.
• If the (rescaled) histogram is similar to the normal curve, then area of regions under the histogram will be approximately equal to areas under the normal curve for the same range of standard units.

• I.e., the percentage of the data that lies within 1 SD of the average will be approximately equal to the area under the normal curve between -1 and 1; the percentage of the data lying within 2 SDs of the average will be approximately equal to the area under the normal curve between -2 and 2; and so forth.

• This is useful, because the distribution of the area under the normal curve is well understood.

• In particular:
  ○ The area under the normal curve between $-1$ and 1 is approximately $0.68 = 68\%$;
  ○ The area under the normal curve between $-2$ and 2 is approximately $0.95 = 95\%$;
  ○ The area under the normal curve between $-3$ and 3 is approximately $0.99 = 99\%$;

This gives the following ‘rule of thumb’...
Rule of thumb:

*If a set of data has an approximately normal distribution, then:*

- *About 68% of the data lies within one SD of average;*
- *About 95% of the data lies within two SDs of average;*
- *About 99% of the data lies within three SDs of average;*

To calculate areas under the normal curve for regions other than those above (−1 to 1, −2 to 2 and −3 to 3), we use a *normal table*, like the one found in the back of the textbook.

**Remember:** This rule *only applies to data that is (approximately) normally distributed!* Absent that condition, or any other assumption about how the data is distributed, we have to rely on weaker estimates, like those given by Chebyshev’s inequality.