Announcements

• Quiz 1 available at 1pm. If you do not receive an email telling you it is available, you need to contact me so I can add you to the list.
• Office hours on web were wrong. They were corrected on Mon evening.
• If you plan to use R Commander in the ICS labs, you need to get an account. See course webpage for information. In the meantime, you can use a temporary account:
  Username: ics-temp , Password: Anteat3r

Homework (due Friday, Oct 1):
Chapter 2: #81, 84, 99

Today:

• Finish material from last time (some of which I rushed through at end)
• Do Section 2.7
• Go over how to install and use R Commander. (See handouts on course webpage.)

Describing Spread (Variability):

• Range = high value – low value
• Interquartile Range (IQR) = upper quartile – lower quartile = $Q_3 - Q_1$ (to be defined)
• Standard Deviation

Example 2.13 Fastest Speeds Ever Driven

Five-Number Summary for 87 males

- Two extremes describe spread over 100% of data
  Range = 150 – 55 = 95 mph
- Two quartiles describe spread over middle 50% of data
  Interquartile Range = 120 – 95 = 25 mph

Notation and Finding the Quartiles

Split the ordered values into the half that is (at or) below the median and the half that is (at or) above the median.

$Q_1 = \text{lower quartile} = \text{median of data values that are (at or) below the median}$

$Q_3 = \text{upper quartile} = \text{median of data values that are (at or) above the median}$

Example 2.13 Fastest Speeds (cont)

Ordered Data (in rows of 10 values) for the 87 males:

- Median = (87+1)/2 = 44th value in the list = 110 mph
- $Q_1$ = median of the 43 values below the median = (43+1)/2 = 22nd value from the start of the list = 95 mph
- $Q_3$ = median of the 43 values above the median = (43+1)/2 = 22nd value from the end of the list = 120 mph
Percentiles

The $k^{th}$ percentile is a number that has $k\%$ of the data values at or below it and $(100 - k)\%$ of the data values at or above it.

- Lower quartile: 25th percentile
- Median: 50th percentile
- Upper quartile: 75th percentile

Describing Spread with Standard Deviation

Standard deviation measures variability by summarizing how far individual data values are from the mean.

Think of the standard deviation as *roughly the average distance values fall from the mean.*

Describing Spread with Standard Deviation: A very simple example

Both sets have same mean of 100.

Set 1: all values are equal to the mean so there is *no variability* at all.

Set 2: one value equals the mean and other four values are 10 points away from the mean, so the *average distance away from the mean is about 10.*

Calculating the Standard Deviation

Formula for the *(sample) standard deviation:*

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n - 1}}$$

The value of $s^2$ is called the *(sample) variance.* An equivalent formula, easier to compute, is:

$$s = \sqrt{\frac{\sum x_i^2 - n\overline{x}^2}{n - 1}}$$

Calculating the Standard Deviation

Example: 90, 90, 100, 110, 110

Step 1: Calculate $\overline{x}$, the sample mean. Ex: $\overline{x} = 100$

Step 2: For each observation, calculate the difference between the data value and the mean.
Ex: -10, -10, 0, 10, 10

Step 3: Square each difference in step 2.
Ex: 100, 100, 0, 100, 100

Step 4: Sum the squared differences in step 3, and then divide this sum by $n - 1$. Result = *variance $s^2$*
Ex: $400/(5 - 1) = 400/4 = 100$

Step 5: Take the square root of the value in step 4.
Ex: $s = \text{standard deviation} = \sqrt{100} = 10$

Population Standard Deviation

Data sets usually represent a sample from a larger population. If the data set includes measurements for an *entire population*, the notations for the mean and standard deviation are different, and the formula for the standard deviation is also slightly different. A *population mean* is represented by the Greek $\mu$ ("mu"), and the *population standard deviation* is represented by the Greek "sigma" (lower case)

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$
Bell-shaped distributions

- Measurements that have a bell-shape are so common in nature that they are said to have a normal distribution.
- Knowing the mean and standard deviation completely determines where all of the values fall for a normal distribution, assuming an infinite population!
- In practice we don’t have an infinite population (or sample) but if we have a large sample, we can get good approximations of where values fall.

Examples of bell-shaped data

- Women’s heights
  - mean = 64.5 inches, s = 2.5 inches
- Men’s heights
  - mean = 70 inches, s = 3 inches
- IQ scores
  - mean = 100, s = 15
- High school GPA for intro stat students
  - mean = 3.1, s = 0.5
- Verbal SAT scores for UCI incoming students
  - mean = 569, s = 75

Women’s heights from UCDavis data, n = 94
Note approximate bell-shape of histogram “Normal curve” with mean = 64, s = 2.5 superimposed over histogram

Interpreting the Standard Deviation for Bell-Shaped Curves: The Empirical Rule

For any bell-shaped curve, approximately
- 68% of the values fall within 1 standard deviation of the mean in either direction
- 95% of the values fall within 2 standard deviations of the mean in either direction
- 99.7% (almost all) of the values fall within 3 standard deviations of the mean in either direction

Ex: Hypothetical population of IQ scores

- 68% of IQ scores are between 85 and 115
- 95% of IQ scores are between 70 and 130
- 99.7% of IQ scores are between 55 and 145

Try Empirical Rule for these:

- Women’s heights
  - mean = 64.5 inches, s = 2.5 inches
- Men’s heights
  - mean = 70 inches, s = 3 inches
- High school GPA for intro stat students
  - mean = 3.1, s = 0.5
- Verbal SAT scores for UCI students
  - mean = 569, s = 75
**Example: Women’s Heights**

Mean height for the 94 UC Davis women was 64.5, and the standard deviation was 2.5 inches. Let’s compare actual with ranges from Empirical Rule:

<table>
<thead>
<tr>
<th>Range of Values:</th>
<th>Empirical Rule</th>
<th>Actual number</th>
<th>Actual percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ± 1 s.d.</td>
<td>68% in 62 to 67</td>
<td>70</td>
<td>70/94 = 74.5%</td>
</tr>
<tr>
<td>Mean ± 2 s.d.</td>
<td>95% in 59.5 to 69.5</td>
<td>89</td>
<td>89/94 = 94.7%</td>
</tr>
<tr>
<td>Mean ± 3 s.d.</td>
<td>99.7% in 57 to 72</td>
<td>94</td>
<td>94/94 = 100%</td>
</tr>
</tbody>
</table>

**The Empirical Rule, the Standard Deviation, and the Range**

- Empirical Rule tells us that the range from the minimum to the maximum data values equals about 4 to 6 standard deviations for data sets with an approximate bell shape.
- For a large data set, you can get a rough idea of the value of the standard deviation by dividing the range by 6.

\[
s \approx \frac{\text{Range}}{6}
\]

**Standardized z-Scores**

**Standardized score or z-score:**

\[
z = \frac{\text{Observed value} - \text{Mean}}{\text{Standard deviation}}
\]

**Example:** UCI Verbal SAT scores had mean = 569 and s = 75. Suppose someone had SAT = 670:

\[
z = \frac{674 - 569}{75} = 1.40
\]

Verbal SAT of 674 for UCI student is 1.40 standard deviations above the mean for UCI students.

**The Empirical Rule Restated for Standardize Scores (z-scores):**

- For bell-shaped data,
  - About 68% of the values have z-scores between −1 and +1.
  - About 95% of the values have z-scores between −2 and +2.
  - About 99.7% of the values have z-scores between −3 and +3.

**Installing and Using R Commander**

- “R” is a sophisticated and free statistical programming language.
- **R Commander** is an add-on, also free, that is menu-driven. It doesn’t do everything R does.
- You can use R Commander in the ICS Computer labs, or install it on your computer.
- See handouts on course webpage for installing R and R Commander, and for using R Commander for Chapters 2 and 5.
- Switch to laptop for R Commander demo.