Mar. 23 Statistic for the day:
Actors who have won an Oscar live an average of 3.9 years longer than actors who have not.

Assignment:
Read Chapter 20 and do Exercises 1, 2, 4, 6

What is the uncertainty in the mean?
We need a margin of error for the mean.

Suppose we take another sample of 237.
What will the mean be?
Will it be 152.5 again?

Probably not.
Consider what would happen if we took 1000 samples, each of size 237, and computed 1000 means.

Formula for estimating the standard deviation of the sample mean (don't need histogram)

Just like in the case of proportions, we would like to have a simple formula to find the standard deviation of the mean without having to resample a lot of times.

Suppose we have the standard deviation of the original sample. Then the standard deviation of the sample mean is:

\[
\text{standard deviation of the data} \div \sqrt{\text{sample size}}
\]

Sample means: measurement variables

Suppose we want to estimate the mean weight at PSU

Data from stat 100 survey, spring 2004. Sample size 237.
Mean value is 152.5 pounds.
Standard deviation is about \((240 - 100)/4 = 35\)

Hypothetical result, using a "population" that resembles our sample:

Extremely interesting: The histogram of means is bell-shaped, even though the original population was skewed!

Standard deviation is about
\[(157 - 148)/4 = 9/4 = 2.25\]

So in our example of weights:
The standard deviation of the sample is about 35.
Hence by our formula:
Standard deviation of the mean is 35 divided by the square root of 237:
\[35/15.4 = 2.3\]
(Recall we estimated it to be 2.25)
So the margin of error of the sample mean is
\[2 \times 2.3 = 4.6\]
Report 152.5 ± 4.6 (or 147.9 to 157.1)
Example: SAT math scores

Suppose nationally we know that the SAT math test has a mean of 100 points and a standard deviation of 100 points.

Draw by hand a picture of what you expect the distribution of sample means based on samples of size 100 to look like.

Sample means have a normal distribution
mean 500
standard deviation 100/10 = 10

So draw a bell shaped curve, centered at 500, with 95% of the bell between 500 – 20 = 480 and 500 + 20 = 520

Back to proportions:
Suppose the true proportion is known

When we know the true population proportion, then we can anticipate where a sample proportion will fall (give an interval of values).

It is known that about 12% of the population is left-handed. Take a sample of size 200.

We need the standard deviation of the sample proportion:

\[ SD = \sqrt{\frac{.12 \times (1-.12)}{200}} = .023 \]

True proportion known (cont’d)

If you play 100 games of craps, where will the proportion of games you win lie 95% of the time?

True proportion (mean of sample proportions): .493
Standard deviation of sample proportions:

\[ \sqrt{\frac{.493 \times (1-.493)}{100}} = .050 \]

Answer: Between .493 – 2(.050) and .493 + 2(.050), or between .393 and .593.
True proportion unknown

Next, suppose we do not know the true population proportion value.

How can we use information from the sample to estimate the true population proportion?

Suppose we have a sample of 100 stat100 students and find that 14 of them are left handed.

Our sample proportion is: .14

We can now estimate the standard deviation of the sample proportion based on a sample of size 100:

\[ SD \approx \sqrt{\frac{.14 \times (1-.14)}{100}} = .035 \]

Hence, 2 standard deviations = 2(.035) = .07

To estimate how we think the sample proportions are going to behave, we need a normal curve centered at .14 with 95% of the bell extending from .14−.07=.07 to .14+.07=.21.

95% confidence interval for the true population proportion:

An interval of values computed from the sample that is almost certain (95% certain in this case) to cover the true (but unknown) population proportion.

The plan:
1. Take a sample
2. Compute the sample proportion
3. Compute the estimate of the standard deviation of the sample proportion: \[ \sqrt{\frac{\text{sample proportion} \times (1 - \text{sample proportion})}{\text{sample size}}} \]
4. 95% confidence interval for the true population proportion: \[ \text{sample proportion} \pm 2(\text{SD}) \]

Other confidence coefficients

The confidence intervals we’ve been constructing (using ±2 std dev) are called 95% confidence intervals.

The confidence coefficient is 95%, or .95.

It means that we cover the middle 95% of the normal curve associated with sample proportions and this requires 2 standard deviations.

We can change the confidence coefficient by using the normal table (p. 157) to determine the appropriate number of standard deviations.
Other confidence coefficients: An example

Suppose we want a 90% confidence interval instead of 95%.

How many standard deviations span the middle 90% of the normal curve?

Back to the original example of sample size 100, 14% left-handed in sample.
We found the sample std dev to be .035.
Hence, to construct a 90% CI we have
\[ .14 - 1.64(.035) = .14 - .057 = .083 \]
\[ .14 + 1.64(.035) = .14 + .057 = .197 \]
90% confidence interval: .083 to .197
(95% confidence interval: .07 to .21)
So the 90% CI is shorter (a more precise estimate) but less accurate. It has a 10% chance of missing the true population proportion.

Law of diminishing returns

Again, consider the problem of estimating the proportion of left-handers at PSU. Suppose we have a sample of size 100 and a sample proportion of 14%.

\[ SD = \sqrt{\frac{.14 \times (1 - .14)}{100}} = .035 \]
This tells us how wide the normal curve is (since 95% of the bell is spanned by 4(.035) = .14
What if we take a sample size 4 times larger? (400 students)

Table of diminishing returns

<table>
<thead>
<tr>
<th>New SD / Old SD</th>
<th>Sample size increased by</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2</td>
<td>4 times</td>
</tr>
<tr>
<td>1/3</td>
<td>9 times</td>
</tr>
<tr>
<td>1/4</td>
<td>16 times</td>
</tr>
<tr>
<td>1/5</td>
<td>25 times</td>
</tr>
<tr>
<td>1/6</td>
<td>36 times</td>
</tr>
<tr>
<td>1/10</td>
<td>100 times</td>
</tr>
</tbody>
</table>
In our problem with sample proportion 14% and sample size 100, we had SD = .035.

If we wanted the SD to be .0035 (divided by 10), we would need a sample size of 100x100 = 10,000.

Hence, the law of diminishing returns: The effect of adding more people to the sample becomes smaller and smaller.

This also applies to the margin of error.