Confidence Intervals

- **Example 1: How prevalent is sports gambling in America?**
- **2007 Gallup poll took a random sample of 1027 adult Americans.**
- **17% of the sampled adults had gambled on sports in the past year.**
- **We know this value 0.17 is an *estimate* of the true proportion of the *entire population* of American adults who have gambled on sports.**

Clicker Quiz 1

In the sports-gambling example, what is a reasonable interpretation of the poll results?

A. Exactly 17% of the sampled American adults have gambled on sports in the past year.

B. Somewhere around 17% of all American adults have gambled on sports in the past year.

C. Exactly 17% of all American adults have gambled on sports in the past year.

D. Both A and B are reasonable interpretations.

Statistics and Parameters

- **Recall that a *statistic* is a number that summarizes something about a *sample*.**
- **Recall that a *parameter* is a number that summarizes something about a *population*.**
- **The sample proportion (denoted $\hat{p}$), 0.17, *estimates* the population proportion (denoted p) in the gambling example.**
- **How can we make that phrase "Somewhere around 17% of all American adults" a bit more precise?**
- **What possible numbers would be reasonable values for the unknown p, given what our sample tells us?**

Sampling Distribution of $\hat{p}$

- **Note that if we had taken a *different* random sample of 1027 adults, our value of $\hat{p}$ would probably be slightly different.**
- **Imagine taking *many repeated* samples (each of size 1027) and calculating the sample proportion each time.**
- **The sampling distribution of $\hat{p}$ describes the pattern of those many sample proportion values.**
- **We know that when the sample size n is reasonably large, the sampling distribution of $\hat{p}$ is *approximately normal*.**

Sampling Distribution of $\hat{p}$ (Continued)

- **The sampling distribution of $\hat{p}$ has a mean of p, the true population proportion.**
- **The sampling distribution of $\hat{p}$ has a standard deviation of**

$$\sqrt{\frac{p(1-p)}{n}}$$

- **This assumes that the population is very large.**
- **Recall the 8th-grade marijuana use example from Chapter 18: We used computer simulation to look at the sampling distribution of $\hat{p}$.**

Clicker Quiz 2

In the Chapter 18 example, we said that the true proportion of all 8th-graders who had smoked marijuana was 0.10. Consider taking a random sample of 100 8th-graders and calculating the sample proportion $\hat{p}$. What is the *mean* of the sampling distribution of $\hat{p}$?

A. **0.10**

B. $\frac{0.10}{100}$ **= 0.001**

C. $\frac{0.10}{\sqrt{100}}$ **= 0.01**

D. **100**

Clicker Quiz 3

In the Chapter 18 example, we said that the true proportion of all 8th-graders who had smoked marijuana was 0.10. Consider taking a random sample of 100 8th-graders and calculating the sample proportion $\hat{p}$. What is the *standard deviation* of the sampling distribution of $\hat{p}$?

A. **0.10**

B. $\frac{0.10 \times 0.90}{100}$ **= 0.0009**

C. $\sqrt{\frac{0.10 \times 0.90}{100}}$ **= 0.03**

D. $\sqrt{100}$ **= 10**

Back to Example 1

- Recall our gambling example – here we don't know the true proportion of adults who gamble.
- This is a more realistic scenario when dealing with real-world data.
- The sample size was large (1027), so we can say that the sampling distribution of $\hat{p}$ is approximately normal.
- But we don't know the center or spread of the sampling distribution, because we don't know p.
- In fact, p (the proportion of gamblers in the adult population) is what we're trying to estimate precisely!

Using the Empirical (68-95-99.7) Rule

- **Since the sampling distribution of $\hat{p}$ is approximately normal, the empirical rule tells us about 95% of all possible samples will produce a value of $\hat{p}$ within 2 standard deviations of the true p (which is unknown).**

- **So in about 95% of samples, $\hat{p}$ will be between $\boldsymbol{p} - 2 \times (sd)$ and $\boldsymbol{p} + 2 \times (sd)$.**

- **Then logically, in about 95% of samples, the true p will be within 2 standard deviations of whatever $\hat{p}$ we got from that sample.**

- **In other words, in about 95% of samples, the unknown p will be between $\hat{\boldsymbol{p}} - 2 \times (sd)$ and $\hat{\boldsymbol{p}} + 2 \times (sd)$.**

Using the Empirical Rule (Continued)

- ***Important:*** **The population proportion p *does not change* from sample to sample.**
- **It is the sample proportion $\hat{p}$ that changes across different samples.**

A Confidence Interval for the Population Proportion

- **The interval** $\left(\hat{\boldsymbol{p}} - 2 \times (sd), \hat{\boldsymbol{p}} + 2 \times (sd)\right)$ **represents an approximate 95%** ***confidence interval*** **for** $\boldsymbol{p}$**.**
- **This gives us a set of reasonable values that** p **could take, given what our sample tells us.**
- **But . . . we still need to find the standard deviation to calculate this interval!**

Handling the Standard Deviation Part

- **Recall the standard deviation of this sampling distribution is**

$$\sqrt{\frac{p(1-p)}{n}}$$

- **We don't know p, so we will use $\hat{p}$ instead.**

- **This is not ideal, but if our sample size is large, we know $\hat{p}$ should be pretty close to p.**

- **The standard deviation of the sampling distribution will be very close to its true value.**

Confidence Interval Formula: Population Proportion

- **An approximate 95% confidence interval for a population proportion may be obtained using the formula:**

$$\left(\hat{\boldsymbol{p}} - 2 \times \sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}, \hat{\boldsymbol{p}} + 2 \times \sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}\right)$$

- **This interval is valid if the sample size is reasonably large.**
- **Usually it works pretty well if there are at least 30 observations in the sample.**

Confidence Interval: Gambling Example

- **Recall our gambling example: From our sample of n = 1027 adults, we found that $\hat{p}$ = 0.17 had gambled.**

- **The left endpoint of our 95% confidence interval would thus be** $0.17 - 2 \times \sqrt{\frac{0.17(0.83)}{1027}}$ **= 0.17 - 2 $\times$ 0.0117 = 0.1466.**

- **The right endpoint of our 95% confidence interval would be** $0.17 + 2 \times \sqrt{\frac{0.17(0.83)}{1027}}$ **= 0.17 + 2 $\times$ 0.0117 = 0.1934.**

- **So the 95% confidence interval for the true proportion of adults in the U.S. who gamble is (0.1466, 0.1934).**

Interpreting the Confidence Interval: Gambling Example

- **We are 95% confident that the true proportion of all adults in the U.S. who gamble is somewhere between 0.1466 and 0.1934.**
- **What exactly does this mean?**
- **It means this interval was obtained using a CI method that will “capture” the true parameter 95% of the time (i.e., in 95% of samples).**
- **So 95% of the time, we’ll get a “typical” sample and our method will “work.”**
- **But 5% of the time, we’ll get a “weird” sample and our method will NOT work (i.e., our interval *won’t* contain the true parameter value!)** . . . **see Figure 21.4.**

Interpreting the Confidence Interval (continued)

- **In our gambling example, was the sample we used one of the "lucky 95%," or one of the "unlucky 5%"?**
- **Unfortunately, we cannot know this – we just have to hope it was one of the lucky ones.**
- **Fortunately, the odds are with us . . . 95% of the time, our interval will be fine.**
- **What if we wanted to improve our chances of "getting lucky"?**
- **We could use, say, a 99% confidence interval – but there's a tradeoff!**

Clicker Quiz 4

Remember the formula $\hat{p} \pm 2 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ gave us an approximate 95% confidence interval. How could we change the formula to give us an approximate 99.7% interval?

A. $\hat{p} \pm 1 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

B. $\hat{p} \pm 2 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

C. $\hat{p} \pm 3 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

D. $\hat{p} \pm 0.5 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Different Confidence Levels

- **To change the level of confidence, we just adjust the number of standard deviations in the "margin of error" part of the formula.**
- **Actually for 99% confidence, we would use 2.58 rather than 2 or 3.**
- **So 99% confidence is better than 95% confidence, right?**
- **In some ways, it is: We have better odds that the interval we get will contain the true parameter value.**
- **But the interval will also be wider – less informative!**
- **The 99% interval for the true proportion of gamblers is (0.1398, 0.2002) . . . not as precise as the 95% interval.**

Different Confidence Levels (continued)

- OK, so let's go for a narrower interval, say 90%.
- For 90% confidence, we would use 1.64 rather than 2.
- The 90% interval for the true proportion of gamblers is (0.1508, 0.1892).
- This is more precise (more informative) than the 95% interval, but there's more of a chance that a 90% interval will miss the true parameter value, across repeated samples.
- Table 21.1 (page 497 of book) gives the appropriate numbers for a lot of different confidence levels.