Chapter 21

What is a Confidence Interval?

Class Poll:

We'll estimate the proportion of college students who engaged in binge drinking during the last year. Take the definition of binge drinking to be 5 or more drinks on one occasion for males or 4 or more drinks for one occasion for females.

Have you engaged in "binge drinking" in the past year?

A. **YES**
B. **NO**

Record the estimate of the proportion of college students who engaged in binge drinking at least once in the last year.

Recall the definition of parameter and statistic:

A **parameter** is a number used to describe a population.

A **statistic** is a number calculated from a sample and is used to estimate the parameter.

For the class poll on binge drinking, what is the statistic? What is the parameter?

We want to use our statistic to estimate the parameter – will our statistic be exactly the same as the population parameter?

Recall the notion of **sampling variability**: Different samples from the same population may yield different values of the sample statistic.

Clicker Question:
How can you reduce the variability of the statistic?

A. **Reduce the sample size to decrease variability**
B. **Increase the sample size to decrease variability**
C. **We can do nothing to control sampling variability**

DEFINITION
The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples from the same population.

CENTRAL LIMIT THEOREM (CLT)
The sampling distribution of a sum or percentage will become approximately normal as the sample size gets larger.

SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

Take a Simple Random Sample (SRS) of size n from a population that contains a certain true proportion of successes, p. Let $\hat{p}$ be the sample proportion of successes.

If the sample size is large enough, then

a. The sampling distribution of $\hat{p}$ is approximately normal
b. The mean of this normal distribution is p
c. The standard deviation of this distribution is $\sqrt{\frac{p(1-p)}{n}}$

Draw a picture:

Using the 68-95-99.7 Rule, approximately 95% of samples will produce $\hat{p}$ between

But, we do not know p! (If we did, we wouldn't need to make a confidence interval…) So use $\hat{p}$ in place of p. This is an approximate 95% confidence interval for p.