Section 5.1
Point estimates and sampling variability

Sampling Variability

- Recall, we are often interested in *population parameters*.
- Complete populations are difficult to collect data on, so we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- *Error* in the estimate = difference between population parameter and sample statistic
- *Bias* is systematic tendency to over- or under-estimate the true population parameter.
- *Sampling error* describes how much an estimate will tend to vary from one sample to the next.

Two commonly studied population parameters

Population Mean

- symbol we use: μ
- μ represents the (unknown) mean of a numerical variable describing a population
- Example: $\mu =$ the mean wingspan of monarch butterflies

Population Proportion

- symbol we use: p
- p represents the (unknown) proportion of some population having a particular feature.
- Example: $p =$ the proportion of adults living in the U.S. who view climate change as an existential threat.

According to the Linfield Fact book for 20-21, updated in April 2021, 32% of the 1,911 students are First Generation Students.

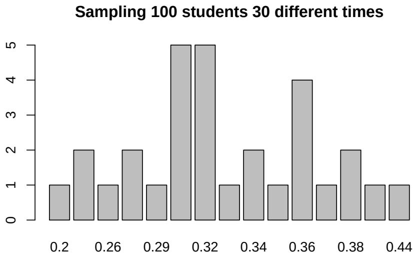
https://inside.linfield.edu/institutional-research/factbook.html

- $p = .32$ is a *population proportion*. (The population here is all Linfield students, and p is the proportion of all 2021 Linfield students who were First Gen.)
- Suppose each of us gathers an independent sample of 100 Linfield students, and determines $\hat{p}$, the proportion of First Gen students in our sample.
- The sample proportion $\hat{p}$ is a *point estimate* for the population proportion p, and $n = 100$ is the *sample size*.
- We would expect our different samples to yield slightly different values of $\hat{p}$.
- How much sampling variability is there likely to be? We can repeat the sampling to get a sense.

```
pop_size <- 1911
samp_size<- 100
linfield <- c(rep("first gen", round(0.32 * pop_size,0)),
                      rep("not", round(0.68 * pop_size,0)))

# 2. Sample 100 students.

sampled_entries <- sample(linfield, size = samp_size)

# 3. Compute p-hat: count the number that are "first gen",
# then divide by  the sample size.

sum(sampled_entries == "first gen") / samp_size
```

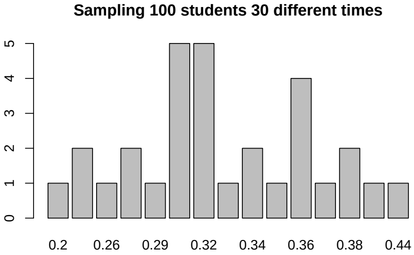
Sampling distribution

- Now, if each of us gathers our own sample, what would the resulting distribution of sample proportions look like?



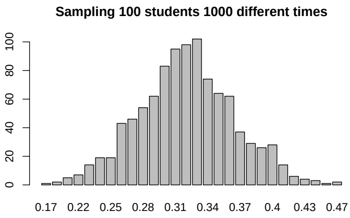
Sampling distribution

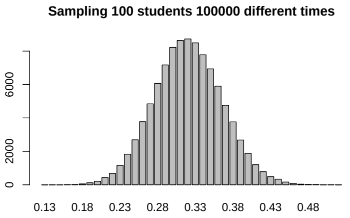
- Now, if each of us gathers our own sample, what would the resulting distribution of sample proportions look like?



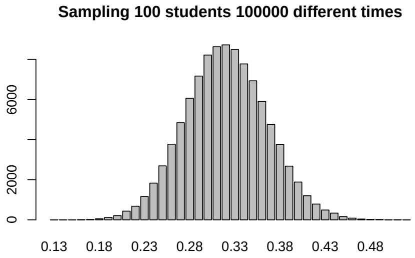
- Suppose you were to repeat this process many, many times and obtain many, many $\hat{p}$s. The distribution of values for $\hat{p}$ is called a *sampling distribution*.

Sampling distribution

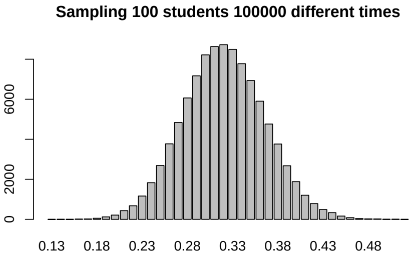




Q: *What is the shape and center of this distribution? Based on this distribution, what do you think is the true population proportion?*



Q: *What is the shape and center of this distribution? Based on this distribution, what do you think is the true population proportion?*



The distribution is unimodal and symmetric. A reasonable guess for the true population proportion is the center of this distribution, approximately 0.32.

Sampling distributions are never observed

- In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution.
- Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

Central Limit Theorem for proportions

Central limit theorem for proportions

If certain conditions are met, sample proportions will be nearly normally distributed with mean equal to the pop'n proportion, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

$$\hat{p} \sim N\left(p, \sqrt{\frac{p\ (1-p)}{n}}\right)$$

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population proportion.
- Note that as n increases SE decreases. Does this make sense?

Central Limit Theorem for proportions

Central limit theorem for proportions

If certain conditions are met, sample proportions will be nearly normally distributed with mean equal to the pop'n proportion, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

$$\hat{p} \sim N\left(p, \sqrt{\frac{p\ (1-p)}{n}}\right)$$

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population proportion.
- Note that as n increases SE decreases. Does this make sense?
  - As n increases samples will yield more consistent $\hat{p}$s, i.e. variability among $\hat{p}$s will be lower.

CLT - Conditions

Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent. This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - if sampling without replacement, $n < 10\%$ of the population.

CLT - Conditions

Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent. This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - if sampling without replacement, $n < 10\%$ of the population.
2. *Sample size:* There should be at least 10 expected successes and 10 expected failures in the observed sample.
   This is difficult to verify if you don't know the population proportion (or can't assume a value for it). In those cases we look for the number of observed successes and failures to be at least 10.

The CLT and the Linfield simulation

The true proportion of Linfield students who are First Gen: $p = .32$. In a sample of $n = 100$ the distribution for $\hat{p}$ will be approximately normal with

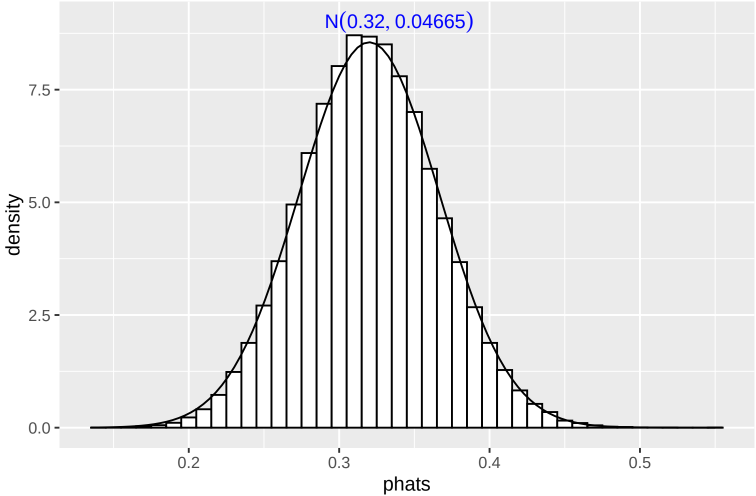
$$\text{mean} = .32$$

and

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.32 \cdot .68}{100}} \approx 0.04665.$$

Let's compare this bell curve with the histogram of 100,000 simulations of finding $\hat{p}$

The CLT and the Linfield simulation



When p is unknown

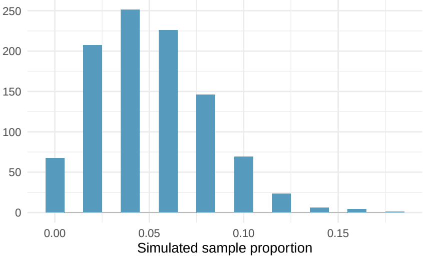
- The CLT states $SE = \sqrt{\frac{p\ (1-p)}{n}}$, with the condition that np and $n(1-p)$ are at least 10, however we often don't know the value of p, the population proportion
- In these cases we substitute $\hat{p}$ for p

When p is low

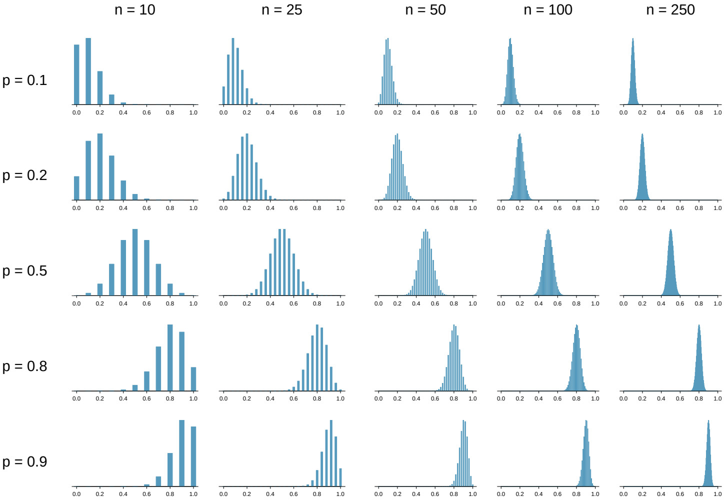
Q: *Suppose we have a population where the true population proportion is $p = 0.05$, and we take random samples of size $n = 50$ from this population. We calculate the sample proportion in each sample and plot these proportions. Would you expect this distribution to be nearly normal? Why, or why not?*

When p is low

Q: *Suppose we have a population where the true population proportion is $p = 0.05$, and we take random samples of size $n = 50$ from this population. We calculate the sample proportion in each sample and plot these proportions. Would you expect this distribution to be nearly normal? Why, or why not?*
No, the success-failure condition is not met ($50 * 0.05 = 2.5$), hence we would not expect the sampling distribution to be nearly normal.



Q: *What happens when* np *and/or* $n(1-p) < 10$*?*



When the conditions are not met...

- When either *np* or $n(1-p)$ is small, the distribution is more discrete.
- When *np* or $n(1-p) < 10$, the distribution is more skewed.
- The larger both *np* and $n(1-p)$, the more normal the distribution.
- When *np* and $n(1-p)$ are both very large, the discreteness of the distribution is hardly evident, and the distribution looks much more like a normal distribution.

Extending the framework for other statistics

- The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion.
  - Take a random sample of students at a college and ask them how many extracurricular activities they are involved in to estimate the average number of extra curricular activities all students in this college are interested in.
- The principles and general ideas are from this chapter apply to other parameters as well, even if the details change a little.

Using CLT to approximate probabilities

Linfield Fact Book 2020-21

According to the Linfield Fact book for 20-21, updated in April 2021, 22% of undergraduates came to Linfield as transfer students.
In a simple random sample of 100 Linfield students, use the CLT to approximate the probability that fewer than 10 of them are transfer students?

Using CLT to approximate probabilities

Linfield Fact Book 2020-21

According to the Linfield Fact book for 20-21, updated in April 2021, 22% of undergraduates came to Linfield as transfer students.
In a simple random sample of 100 Linfield students, use the CLT to approximate the probability that fewer than 10 of them are transfer students?

First check: are conditions of CLT being met?

- *Independence*: Yep! SRS, and $n = 100 < 10\%$ of the pop'n
- *Sample size*: Yep!, $p = .22$, so $np = 22$ and $n(1-p) = 88$, both are at least 10.

Transfer Students in a SRS of Linfield Students

Recall, $n = 100$, $p = .22$

- $\hat{p}$ - sample proportion of transfer students. By the CLT, $\hat{p} \sim N(0.22, 0.04142)$
- We estimate $P(\hat{p} < 0.1)$ by first converting to z-scores:

$$\begin{aligned} P(\hat{p} < 0.1) &= P(Z < (0.1 - 0.22)/0.04142) \\ &= P(Z < -2.90) \\ &\approx 0.0019. \end{aligned}$$

- So there is about a 0.2% chance that a SRS of 100 Linfield students would include fewer than 10 transfer students.
- In other words, there is a about a 1 in 500 chance that an SRS of 100 students would have fewer than 10 transfer students.