Chapter 5

Foundations for inference[1]

Department of Mathematics & Statistics
North Carolina A&T State University

[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Point estimates and sampling variability

Point estimates and error

- 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 $\rightarrow$ difference between population parameter and sample statistic.
- Generally, the **error** in the estimate consists of two aspects:
- **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.
- Much of statistics is focused on understanding and quantifying sampling error, and **sample size** is helpful for quantifying this error.

Practice

Suppose we randomly sample 1,000 adults from each state in the U.S. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Practice

Suppose we randomly sample 1,000 adults from each state in the U.S. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

Margin of error

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- $41\% \pm 2.9\%$: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- $49\% \pm 4.4\%$: We are 95% confident that 44.6% to 53.4% of 18-34 years old have taken a job they didn't want just to pay the bills.

Practice

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Is a randomly selected American adult more or less likely to supoort the expansion of solar energy?

Practice

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Is a randomly selected American adult more or less likely to supoort the expansion of solar energy?

More likely.

Suppose that you don't have access to the population of all American adults, which is a quite likely scenario. In order to estimate the proportion of American adults who support solar power expansion, you might sample from the population and use your sample proportion as the best guess for the unknown population proportion.

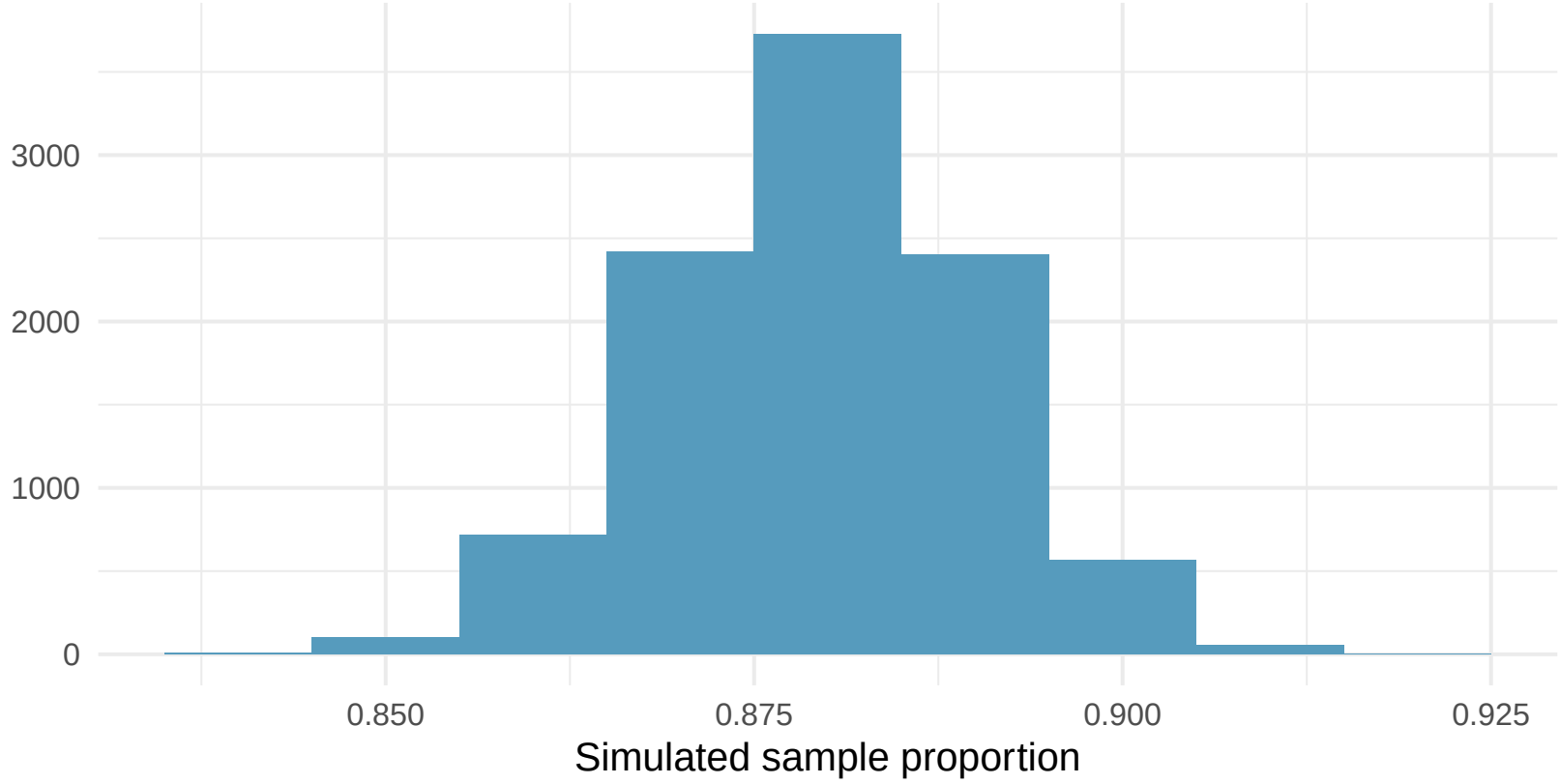
- Sample, without replacement, 1000 American adults from the population, and record whether they support or not solar power expansion.
- Find the sample proportion.
- Plot the distribution of the sample proportions obtained by members of the class.

```r
# 1. Create a set of 330 million entries, where 88%
# of them are "support" and 12% are "not".

pop_size <- 330000000
possible_entries <- c(rep("support", 0.88 * pop_size),
                      rep("not", 0.12 * pop_size))

# 2. Sample 1000 entries without replacement.

sampled_entries <- sample(possible_entries,
                          size = 1000, replace = F)

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

sum(sampled_entries == "support")/1000
```

```
## [1] 0.886
```

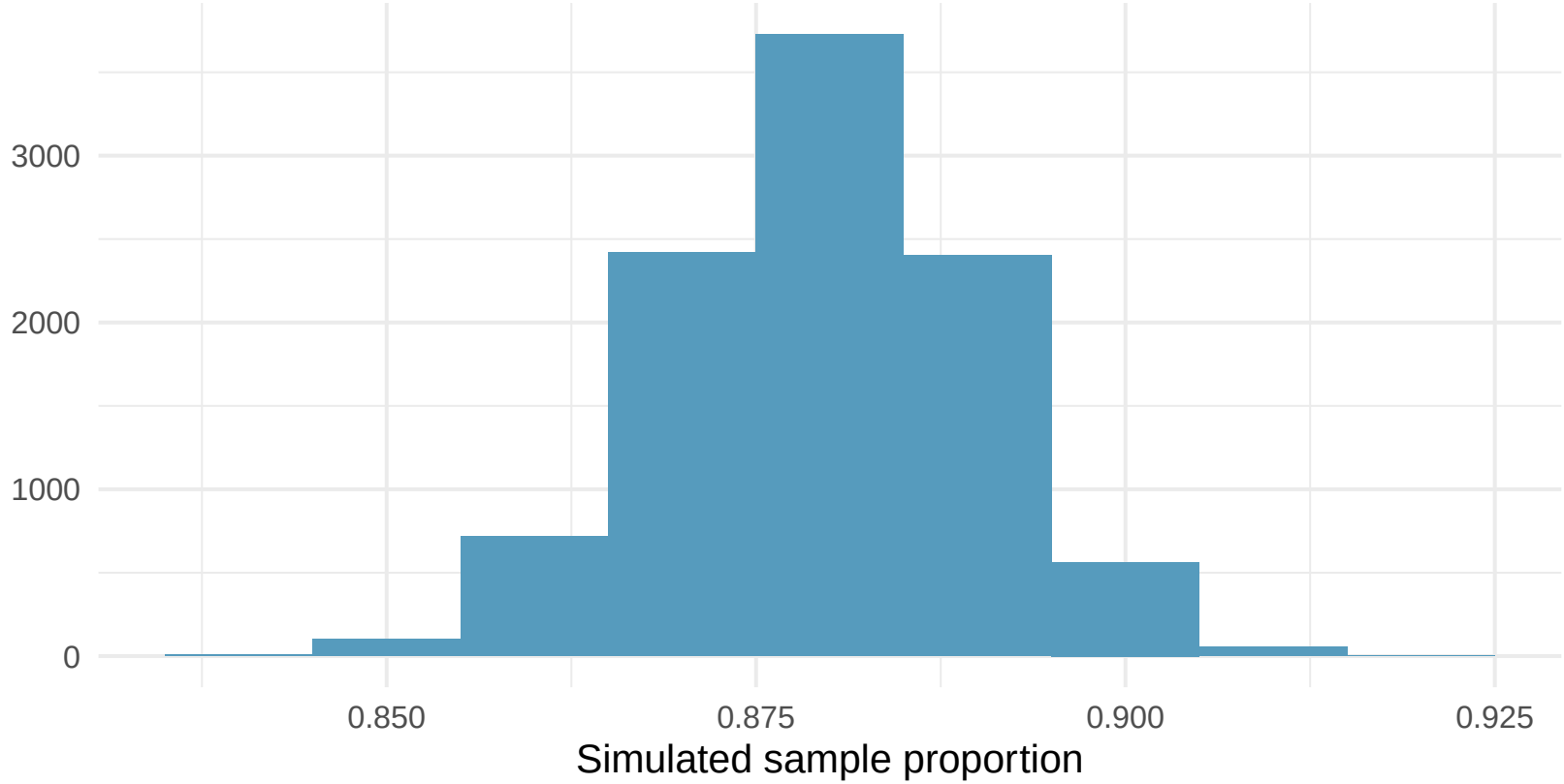
Sampling distribution

Suppose you were to repeat this process many times and obtain many $\hat{p}$s. This distribution is called a sampling distribution.



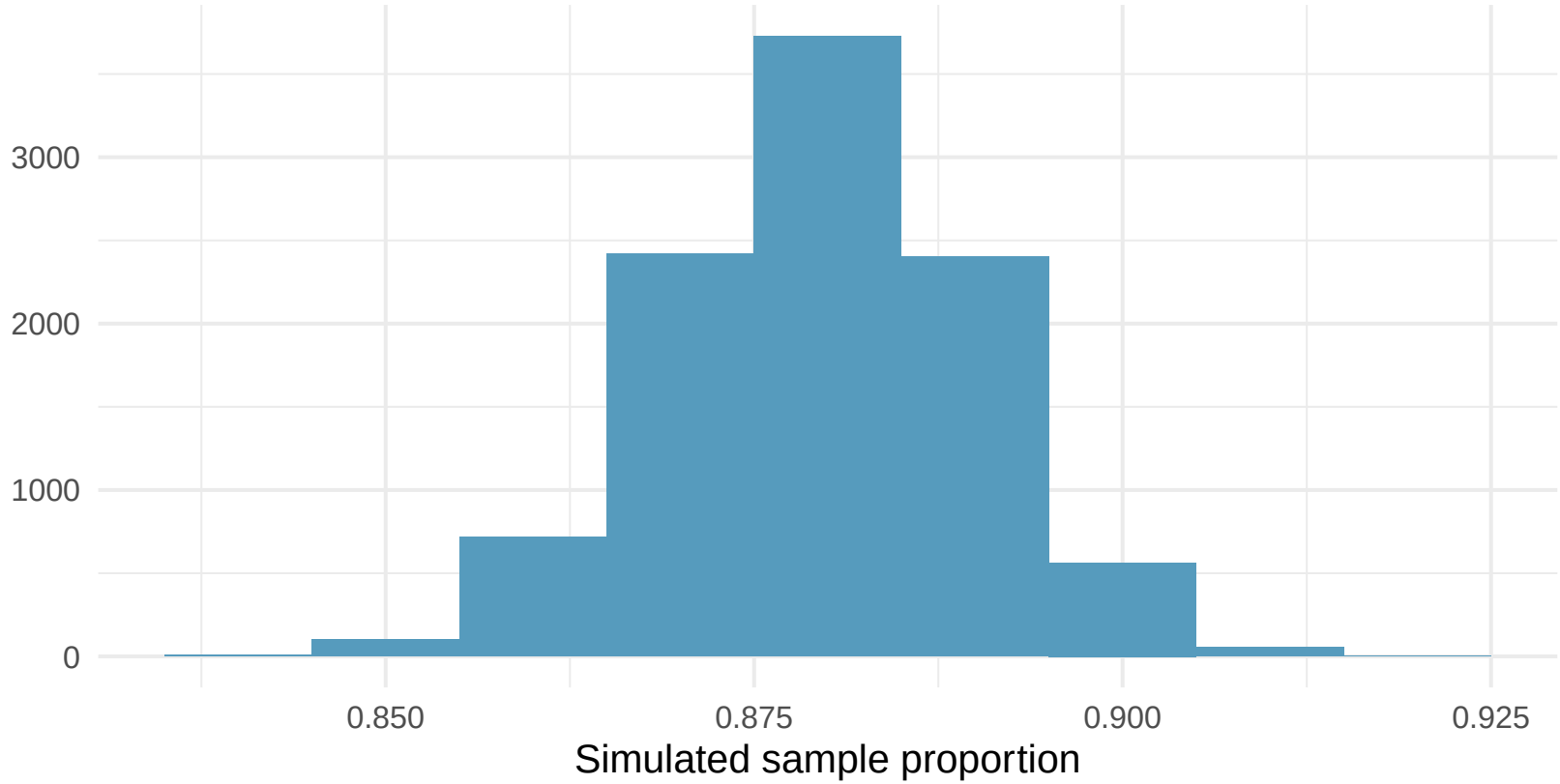
Practice

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



Practice

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 roughly symmetric. A reasonable guess for the true population proportion is the center of this distribution, approximately 0.88.

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

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p, and standard error equal to $\sqrt{\frac{p(1-p)}{n}}$.

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

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population proportion.
- We won't go through a detailed proof of why $SE = \sqrt{\frac{p(1-p)}{n}}$, but note that as n increases SE decreases.
  - 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.
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.

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.

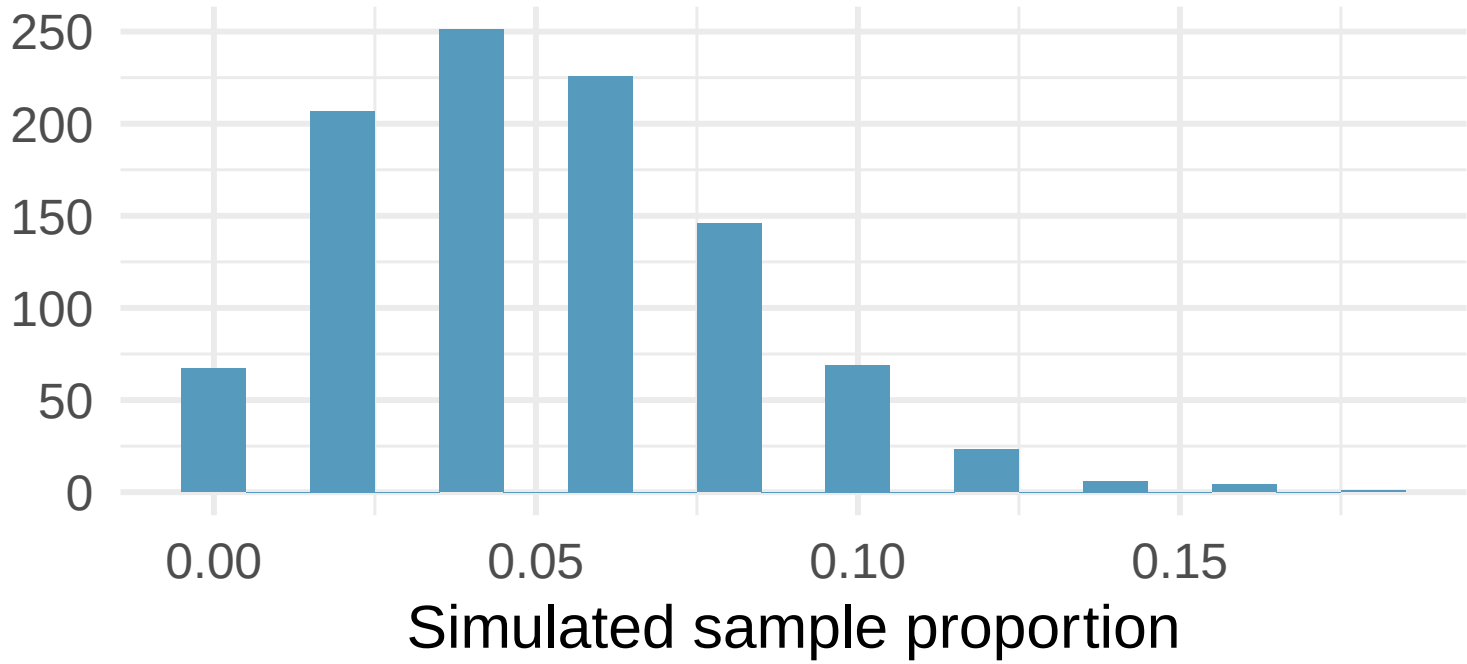
When p is low

Suppose we have a population where the true population proportion is $p = 0.05$, and we take random sample 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?

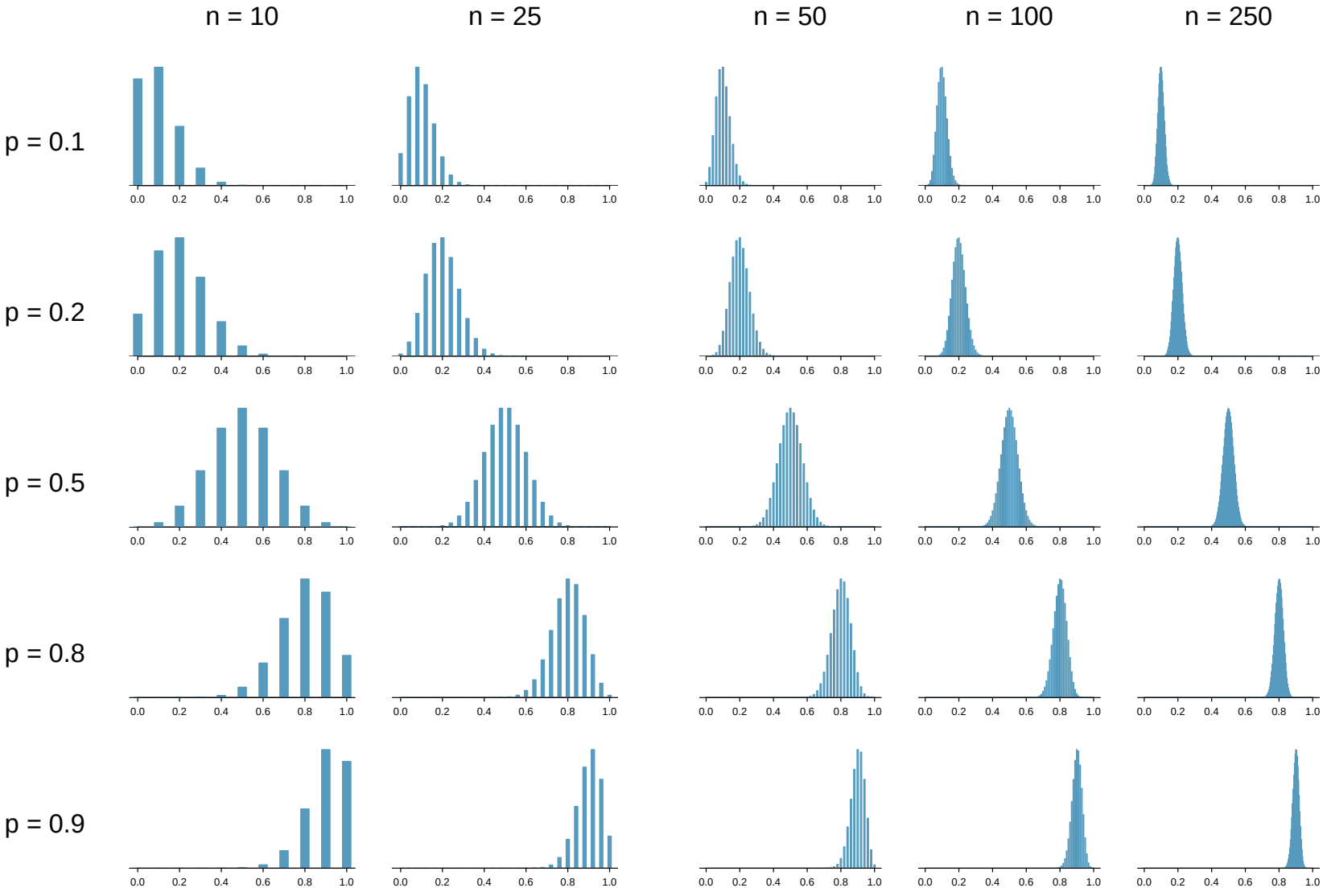
When p is low

Suppose we have a population where the true population proportion is $p = 0.05$, and we take random sample 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.



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 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.