CAAP Statistics - Lec11

Jul 22, 2022

Review: Distributions of Random Variable

- Normal Distribution: $N(\mu, \sigma)$
  - To model most of the numerical variables
  - Symmetric, Unimodal, Bell-shaped curve
  - Normal probability using `pnorm`
  - Normal percentile using `qnorm`
- Bernoulli Distribution: $Ber(p)$
  - 2 possible outcomes: success OR failure
- Binomial Distribution: $Bin(n, p)$
  - To model the number of success out of n trials
  - Independent sum of Bernoulli trials
  - Normal Approximation when n is large enough
- Poisson distribution: $Poi(\lambda)$
  - To model the number of rare event in a give unit of time

Learning Objectives

- Point estimates and sampling variability
  - What is sampling distribution
  - Central Limit Theorem
- Confidence intervals for a proportion
  - Interpreting the confidence interval
- Hypothesis testing for a proportion
  - Null hypothesis vs. Alternative hypothesis
  - Decision Error(Type I error, Type II error)

Point Estimates and Sampling Variability

Parameter—a.k.a a quantity of Interest—estimation

- We are often interested in *population parameters*.
- Since complete populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the *margin of error* associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

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

Suppose we randomly sample 1,000 adults from each state in the US. 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.

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 support the expansion of solar energy?

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 support the expansion of solar energy?

More likely.

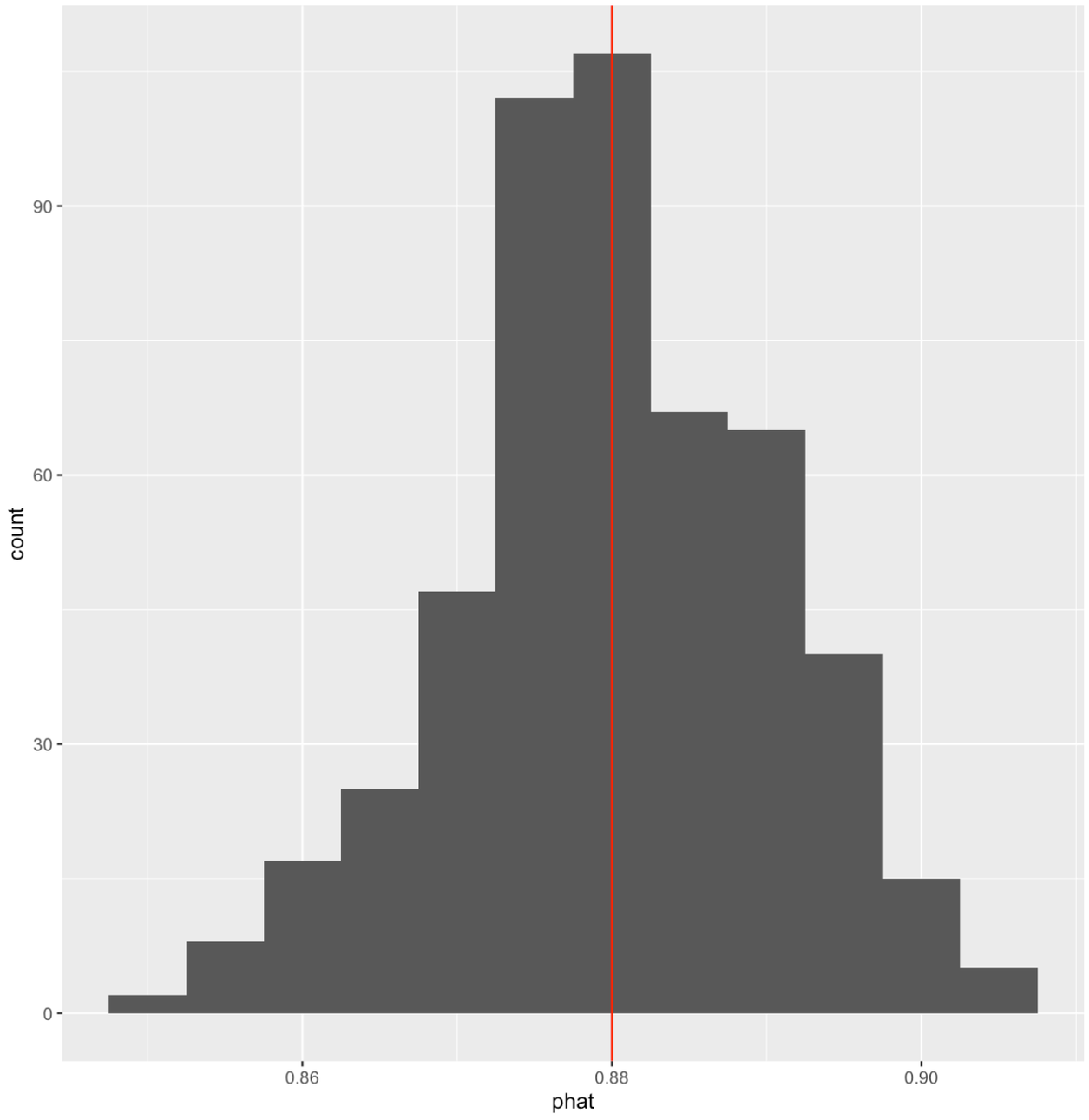
Suppose that you don’t have access to the population of all American adults, which is a quite likely scenario. In order to estimate the proportion of American adults who support solar power expansion, you might sample from the population and use your sample proportion as the best guess for the unknown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

- Sample, with replacement, 1000 American adults from the population, and record whether they support solar power or not expansion.
- Find the sample proportion.
- Plot the distribution of the sample proportions obtained by members of the lass.

```r
library(ggplot2)

# 1. Create a set of 250 million entries, where 88% of
# them are "support" and 12% are "not".

pop_size <- 250000000
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)

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

res = data.frame(phat = rep(0,100))
set.seed(2022)
for(i in 1:500){
  sampled_entries <- sample(possible_entries, size = 1000)
  res[i,] = sum(sampled_entries == "support") / 1000
}
```

Sampling distribution

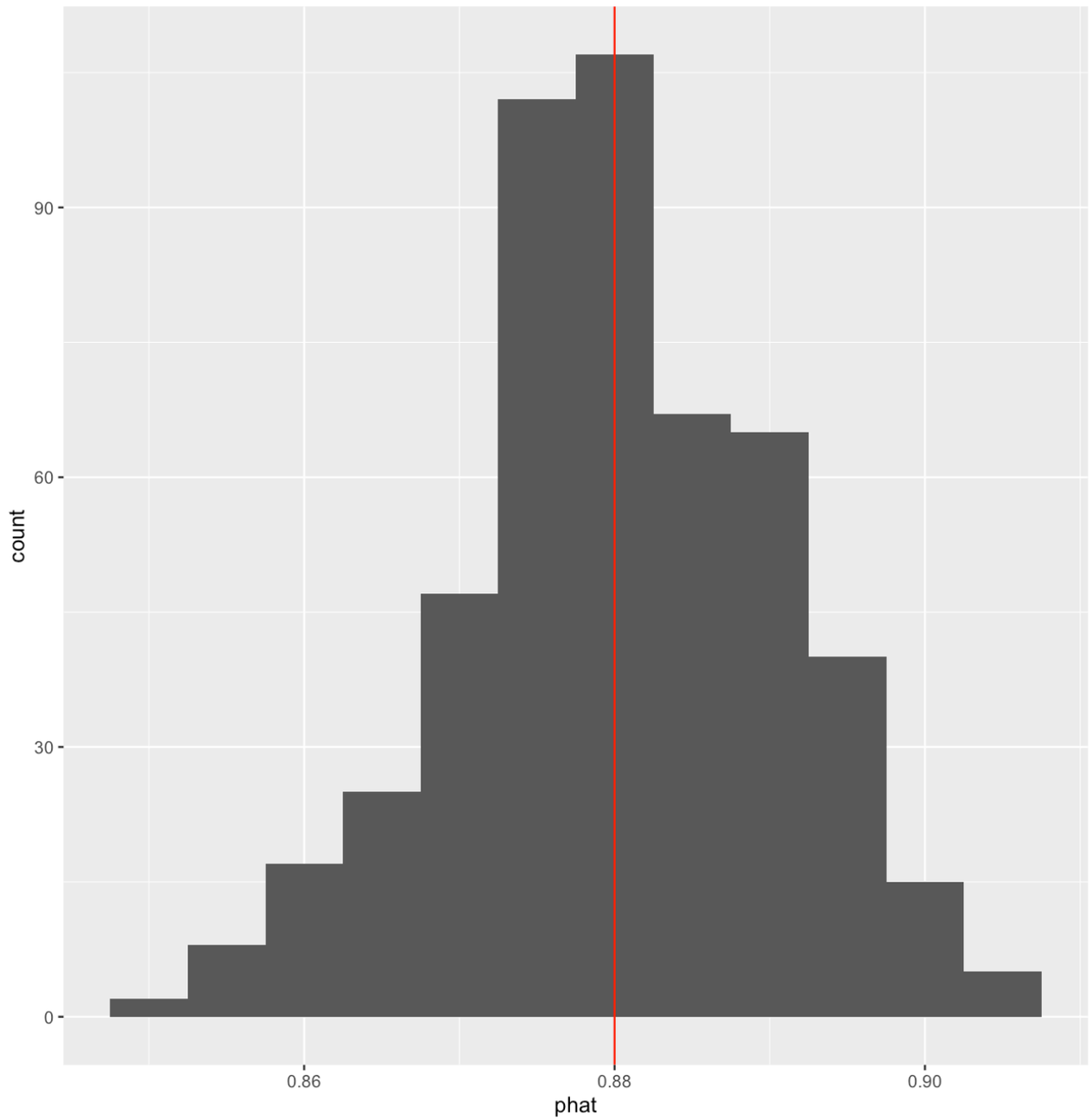
Suppose you were to repeat this process many times and plot the results.
What you just constructed is called a sampling distribution.



Sampling distribution

What is the shape and center of this distribution?

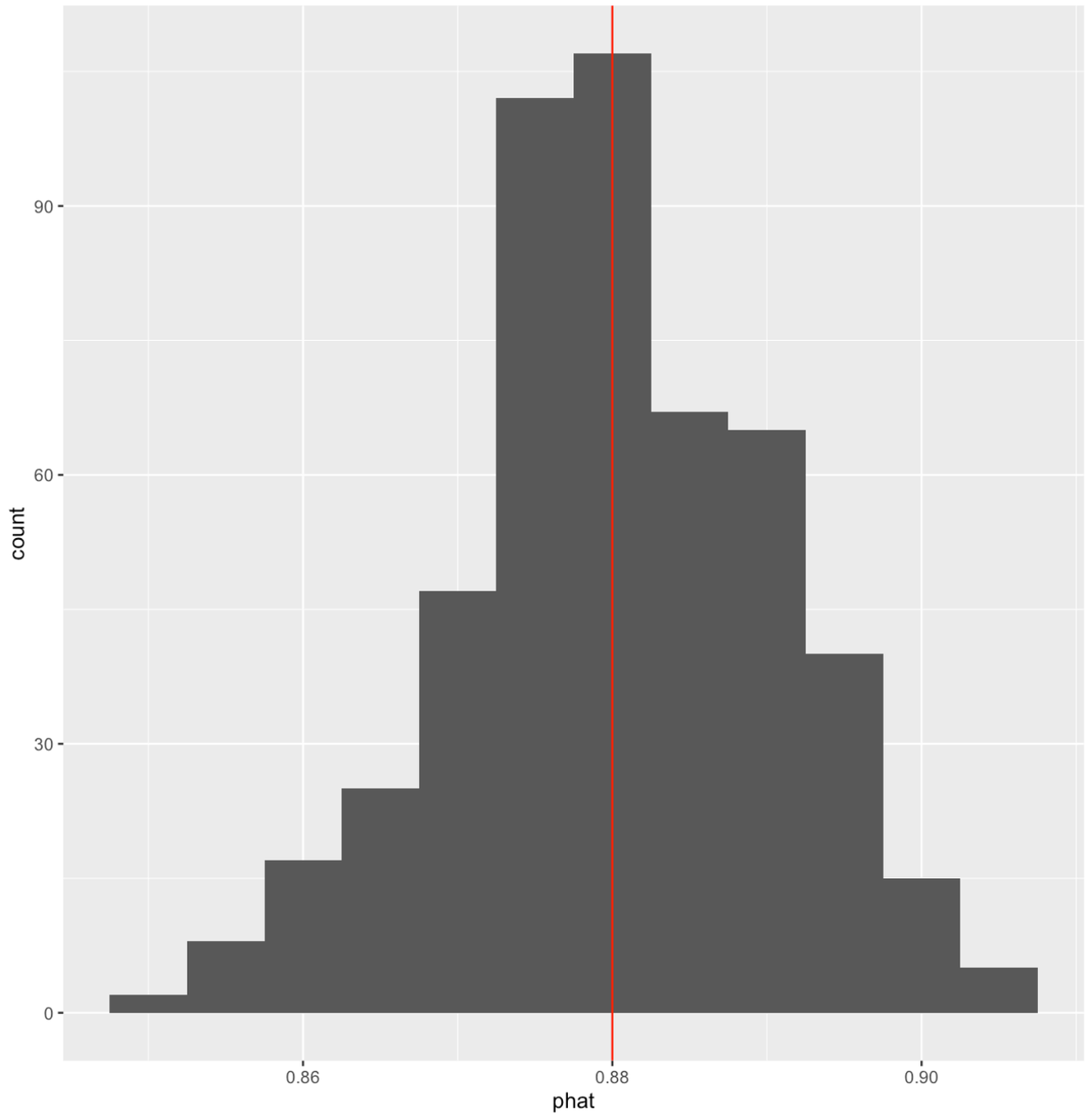
The distribution looks symmetric and somewhat bell-shaped.



Sampling distribution

Based on this distribution, what do you think is the true population proportion?

The center of the distribution: about 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 population.

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:

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.

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

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

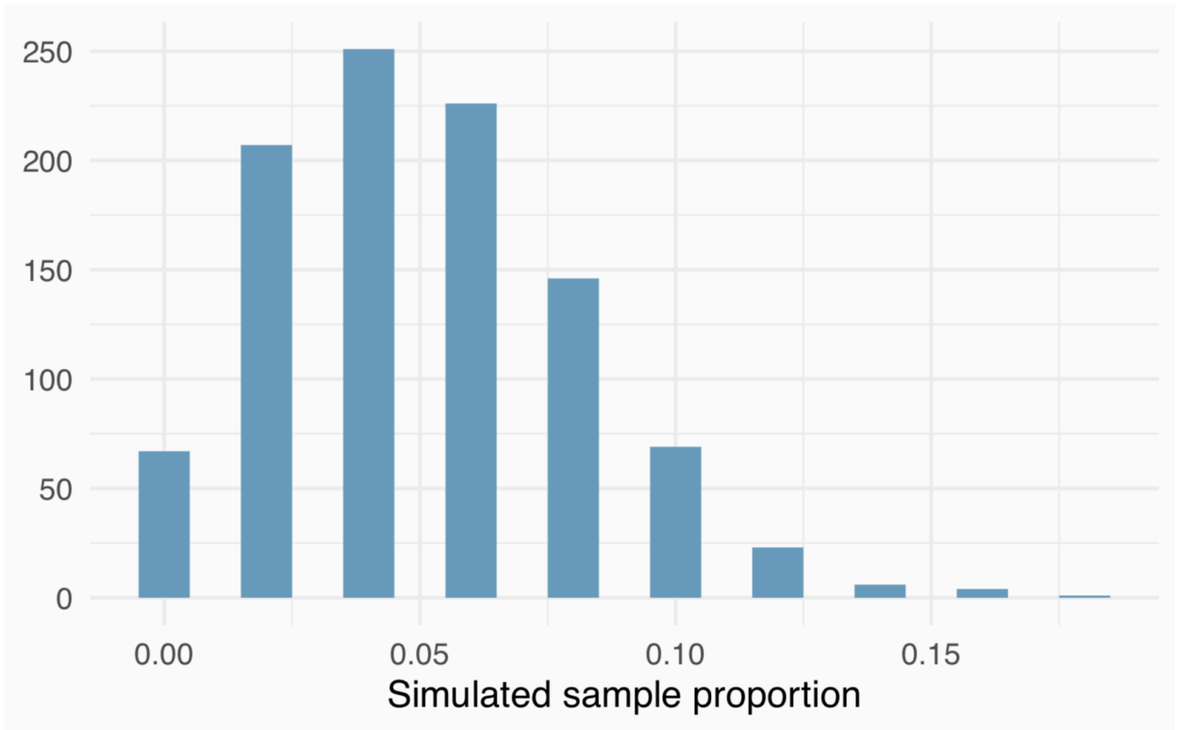
with the condition that *np* and *n(1 − p)* are at least 10.

However, we often don't know the value of *p*, the population proportion. In these cases we substitute $\hat{p}$ for *p*.

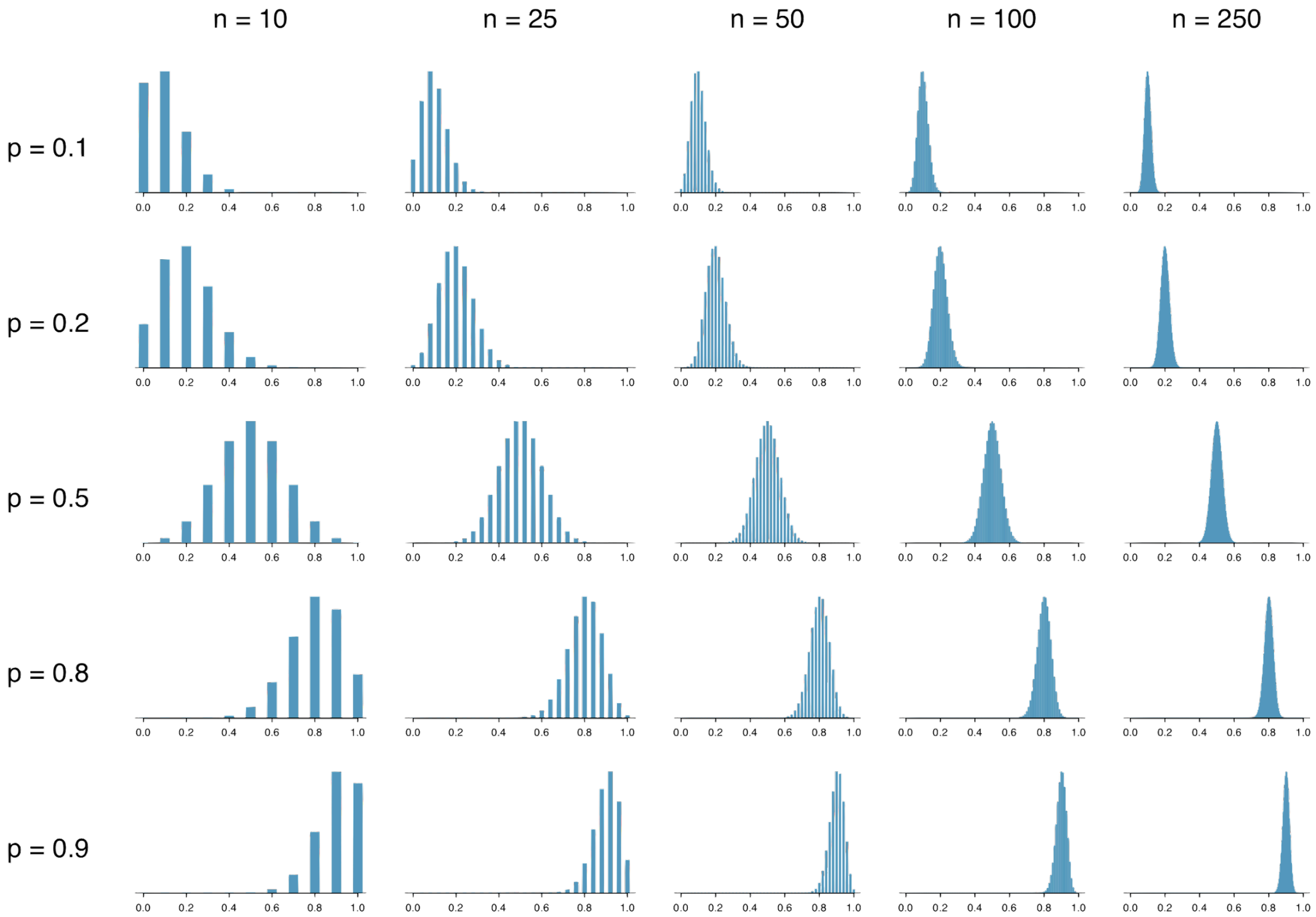
When *np* or *n(1 - p)* is small

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 x 0.05 = 2.5), so 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

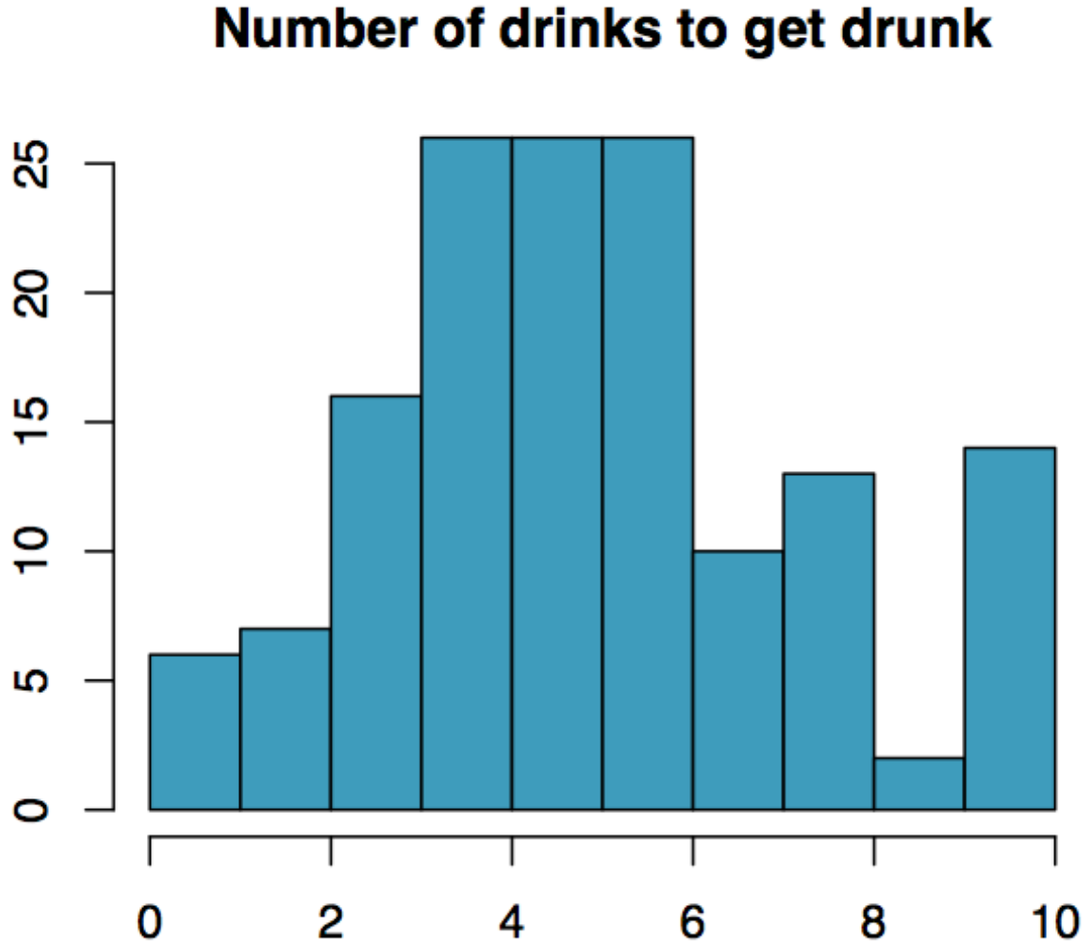
The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion.

- 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.

Practice

The following histogram shows the distribution of number of drinks it takes a group of college students to get drunk. We will assume that this is our population of interest. If we randomly select observations from this data set, which values are most likely to be selected, which are least likely?



Sampling distribution

What you just constructed is called a *sampling distribution*.

Sampling distribution

What you just constructed is called a *sampling distribution*.

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

Sampling distribution

What you just constructed is called a *sampling distribution*.

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

Approximately 5.39, the true population mean.

Central Limit Theorem

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right),$$

where *SE* is represents *standard error*, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use *s*.

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population mean.
- We won't go through a detailed proof of why $SE = \sigma / \sqrt{n}$, but note that as *n* increases *SE* decreases.
  - As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

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:

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.

Sample size / skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.

- the more skewed the population distribution, the larger sample size we need for the CLT to apply
- for moderately skewed distributions n > 30 is a widely used rule of thumb

This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.