Point Estimates and Sampling Variability

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

Plan for Today

- Quick Review
- Point Estimates
- Sampling Variability

Review

Work with whoever you are sitting near to do the following:

- Researchers conducted a randomized controlled study where patients were randomly assigned to one of two groups: treatment or control. The patients in the treatment group received acupuncture that was specially designed to treat migraines. The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. What is the research question for this study? What are the explanatory and response variables in this study?
- A data set that contains sepal length and width, and petal length and width from three species of iris flowers (setosa, versicolor and virginica). There are 50 flowers from each species in the data set.
  - How many observations are in the data?
  - How many numerical variables are in the data? Indicate what they are, and if they are continuous or discrete.
  - How many categorical variables are in the data, and what are they? List the corresponding levels (categories).
- Consider the distribution: $N(\mu = 0, \sigma = 1)$. Sketch graphs to show what percentage of the normal distribution is found in the following regions
  - $Z < -1.35$
  - $|Z| > 2$
  - $-0.4 < Z < 1.5$

Parameter estimation

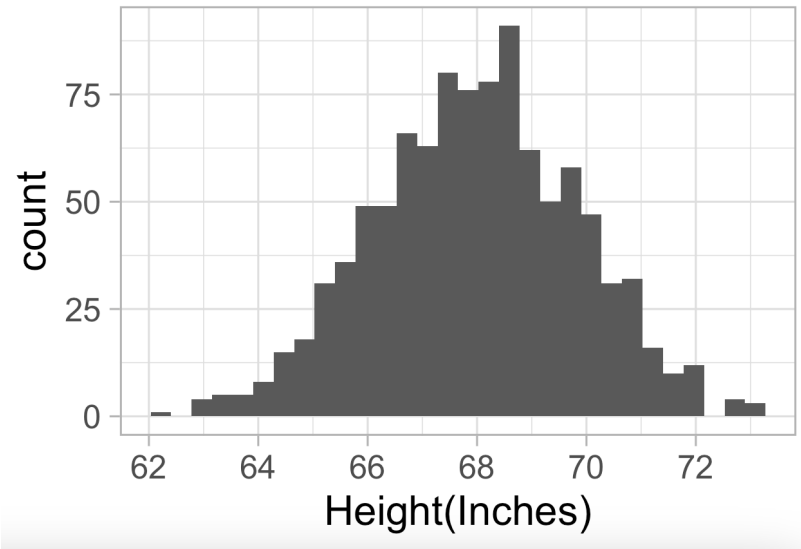
Suppose we randomly sample 1,000 adults from the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Parameter estimation

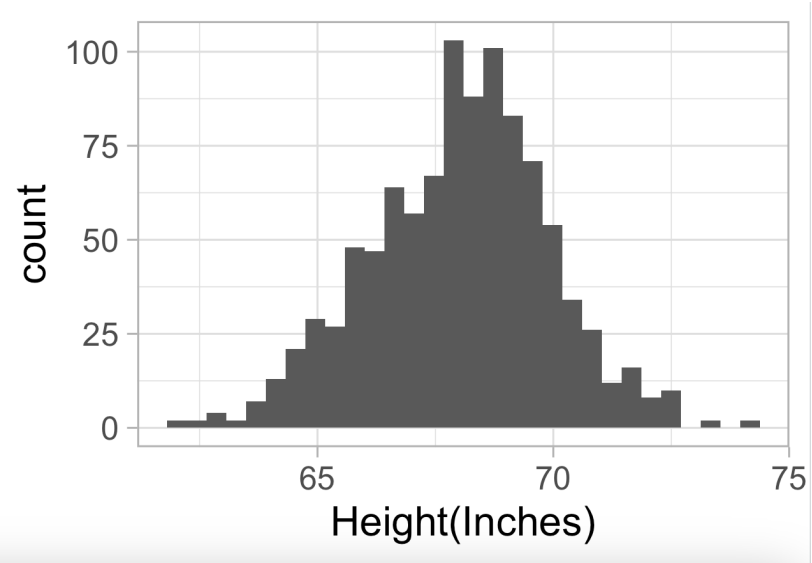
Suppose we randomly sample 1,000 adults from the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

$\bar{x} = 67.97$



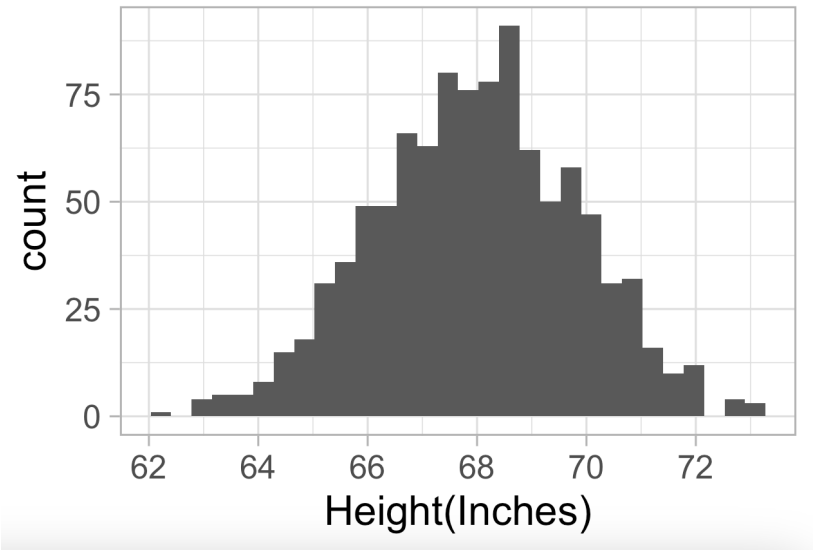
$\bar{x} = 68.02$



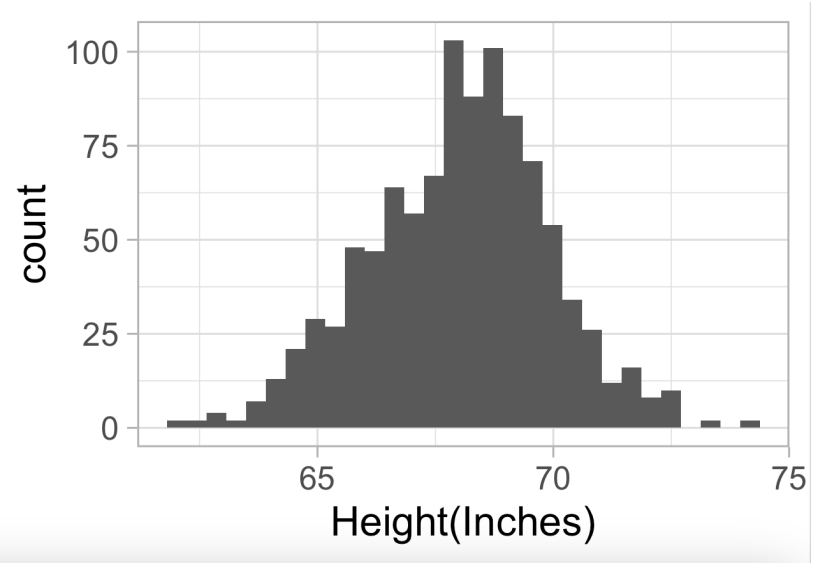
Parameter estimation

Suppose we randomly sample 1,000 adults from the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?
What about standard deviation?

$\bar{x} = 67.97$



$\bar{x} = 68.02$



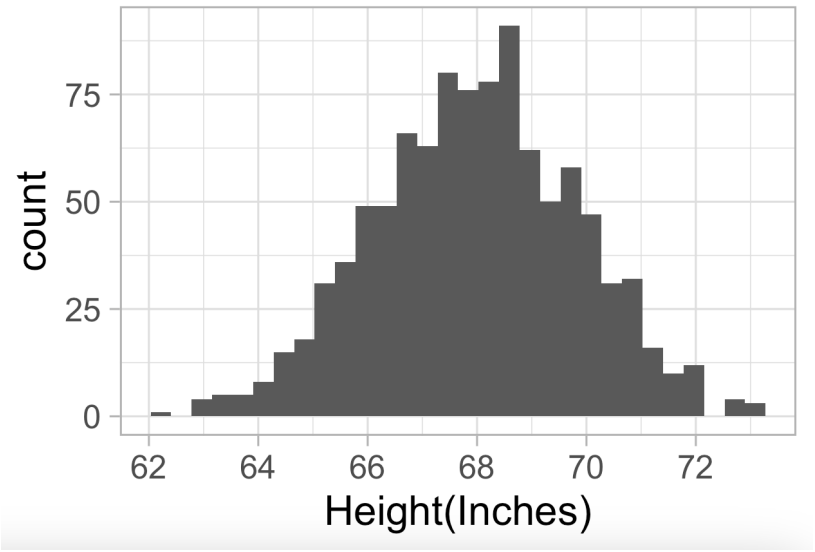
Parameter estimation

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

What about standard deviation? Not the same, but only somewhat different.

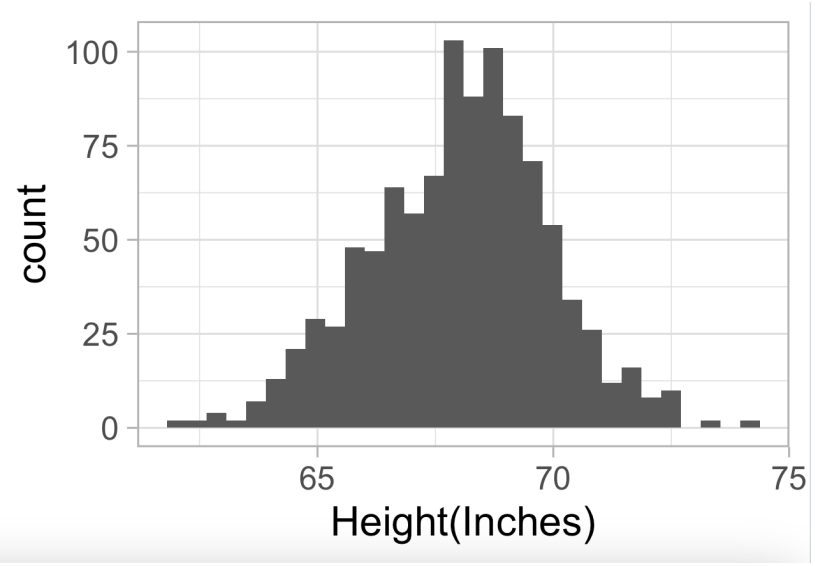
$\bar{x} = 67.97$

s = 1.86



$\bar{x} = 68.02$

s = 1.90



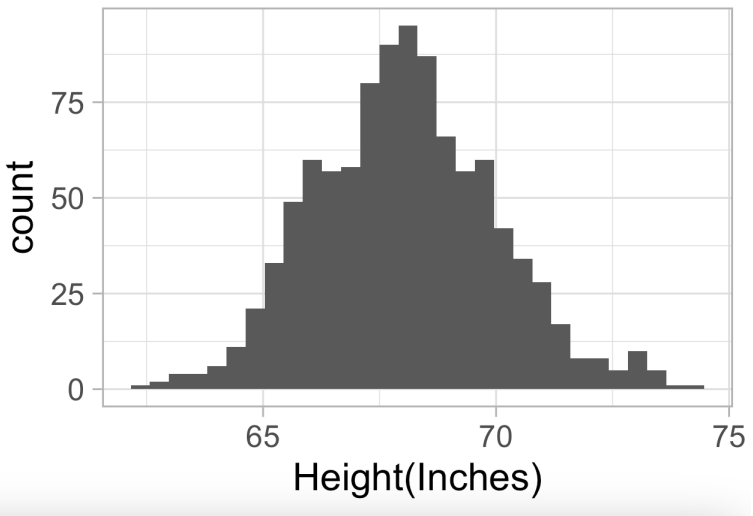
Parameter estimation

Suppose we randomly sample 1,000, 5,000, and 10,000 adults from the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different? What about standard deviation?

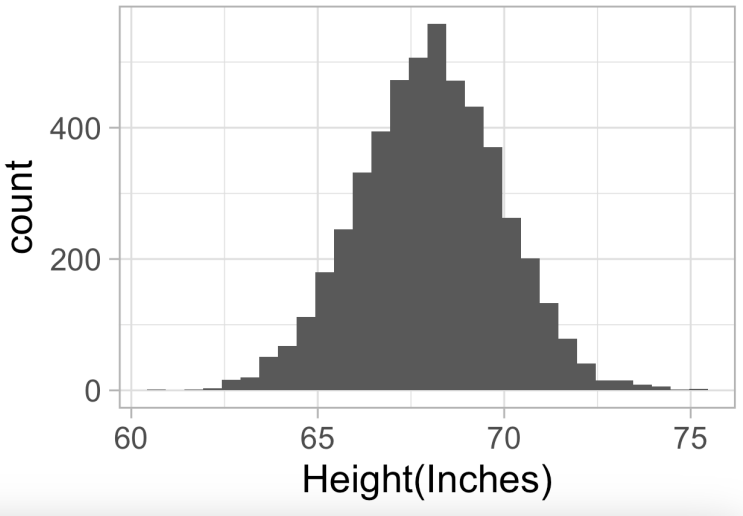
Parameter estimation

Suppose we randomly sample 1,000, 5,000, and 10,000 adults from the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different? What about standard deviation?

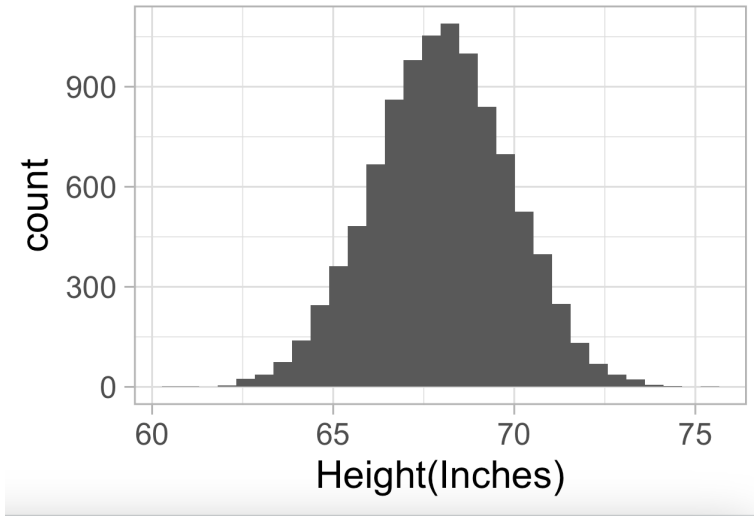
$n = 1000$
$\bar{x} = 68.05$
s = 1.94



$n = 5000$
$\bar{x} = 68.00$
s = 1.91



$n = 10000$
$\bar{x} = 67.98$
s = 1.89



Parameter 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.

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

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

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

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.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

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

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.

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. (41% ± 2.9%)

We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills. (49% ± 4.4%)

Margin of error

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Parameter estimation

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.

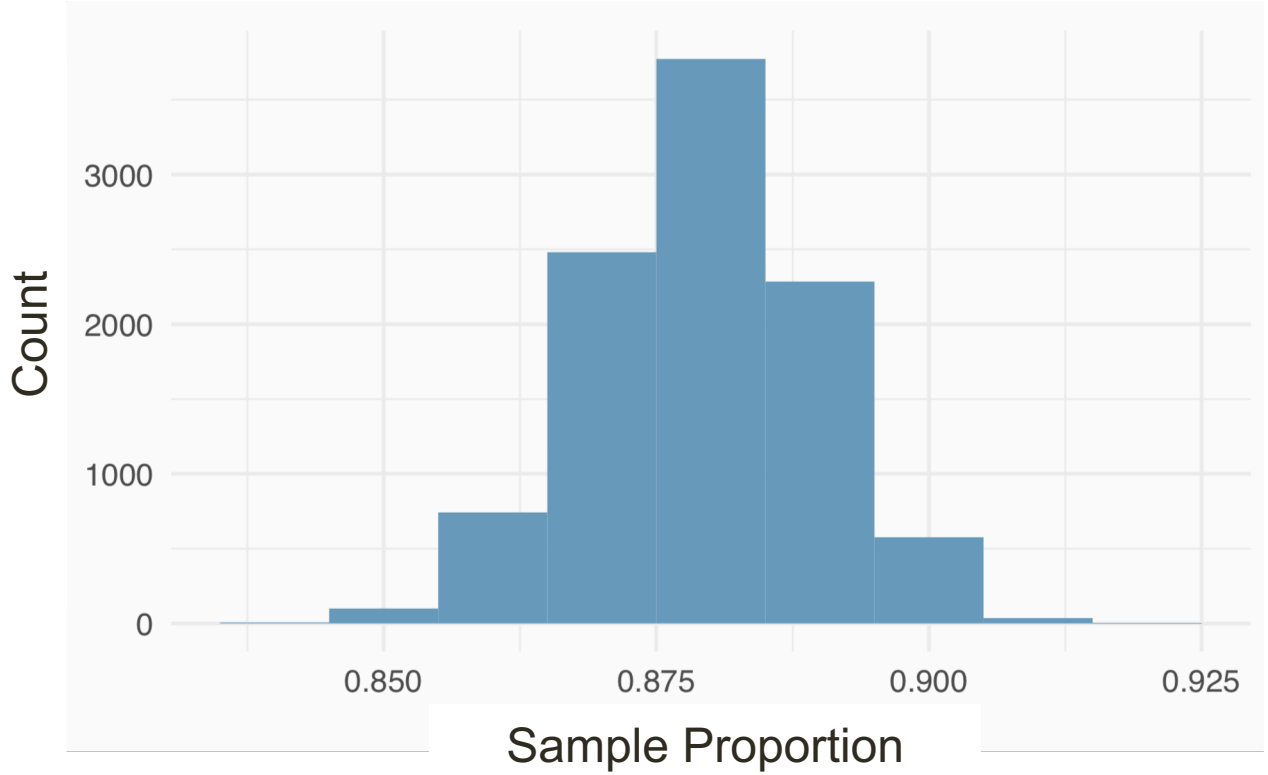
Parameter estimation

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.

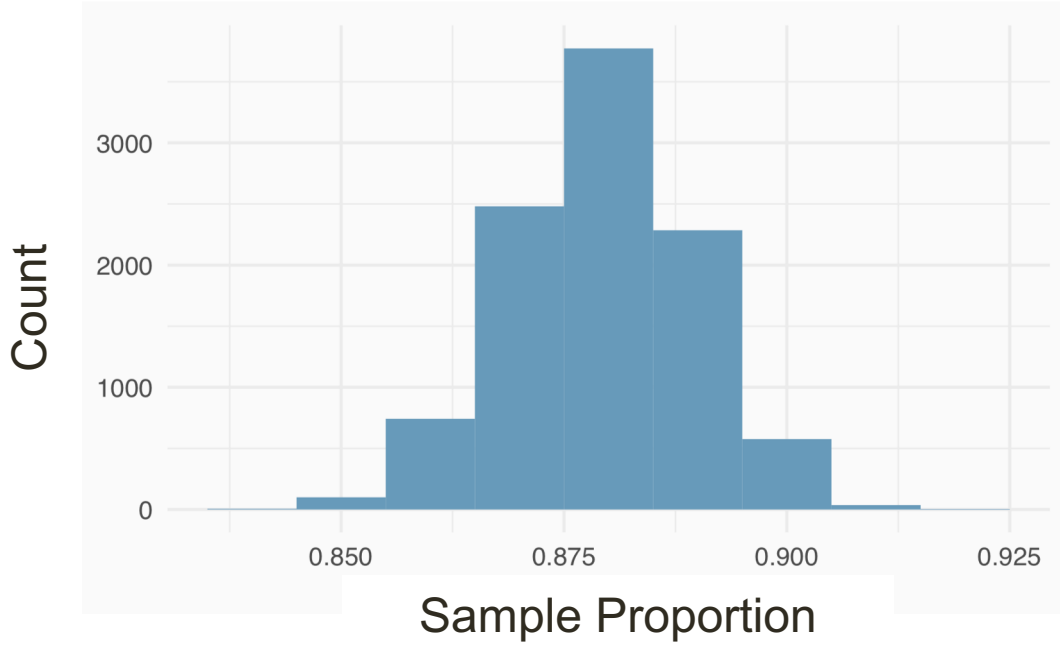
Sampling distribution

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



Sampling distribution

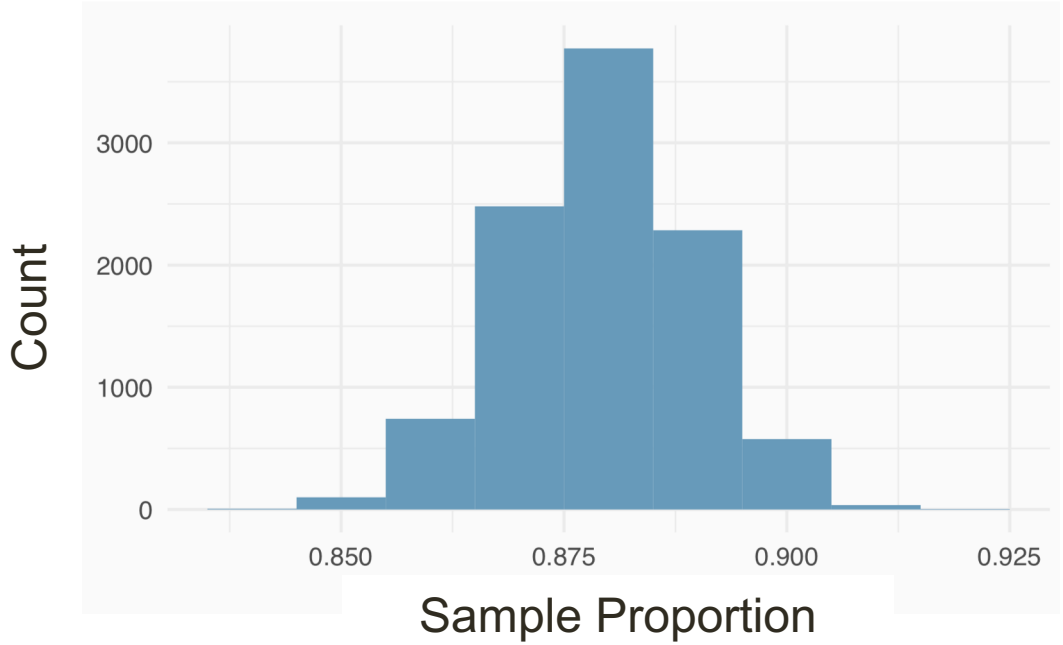
What is the shape and center of this distribution?



Sampling distribution

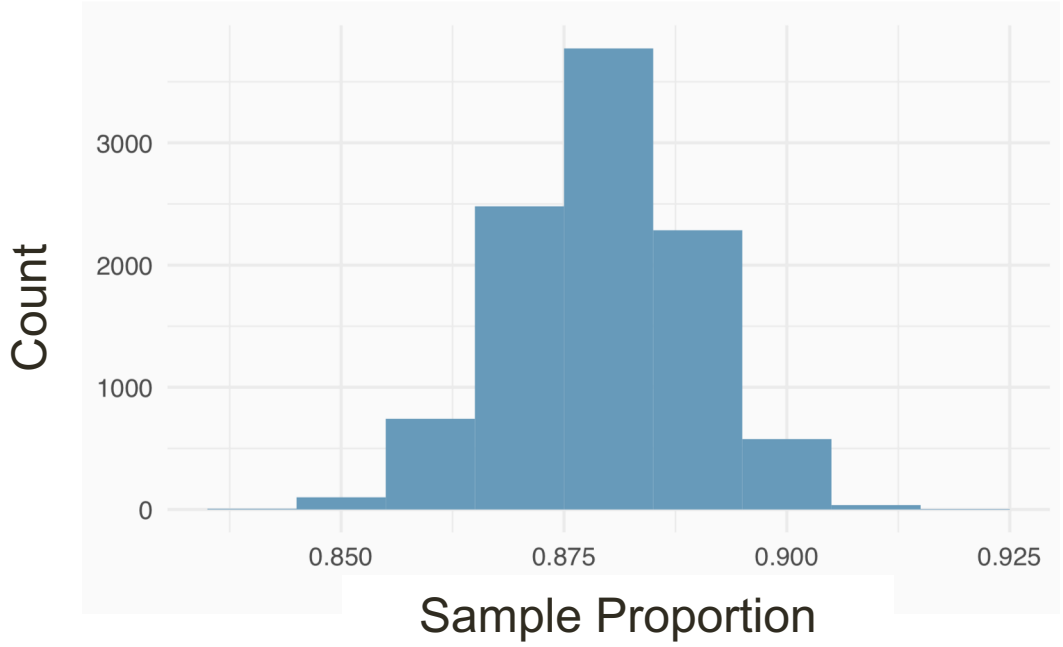
What is the shape and center of this distribution?

The distribution looks symmetric and unimodal.



Sampling distribution

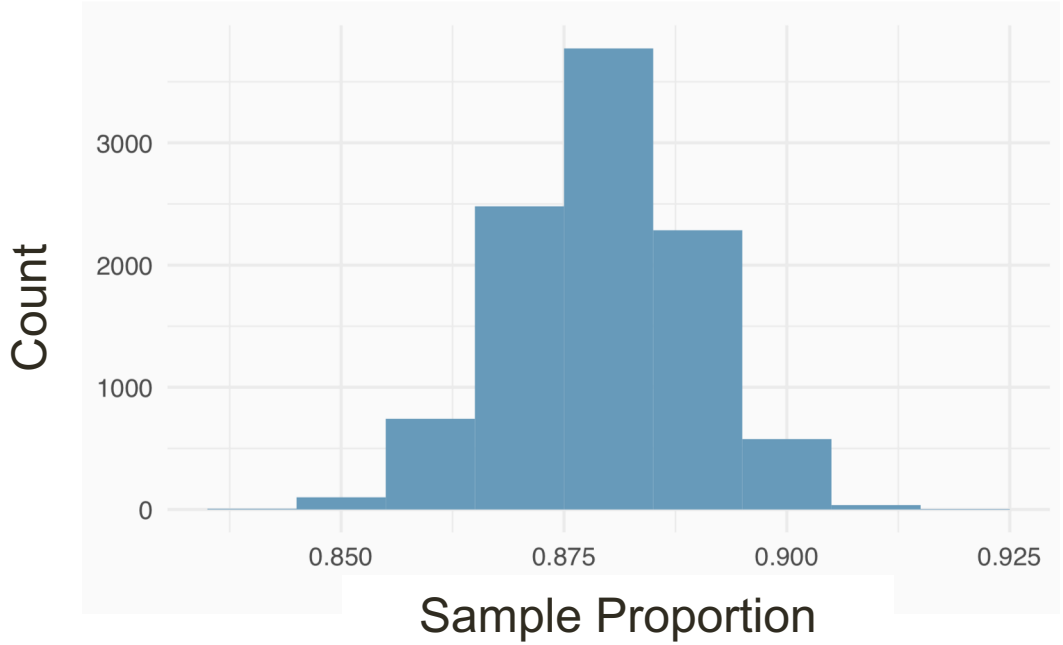
Based on this distribution, what do you think is the true population proportion?



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*.

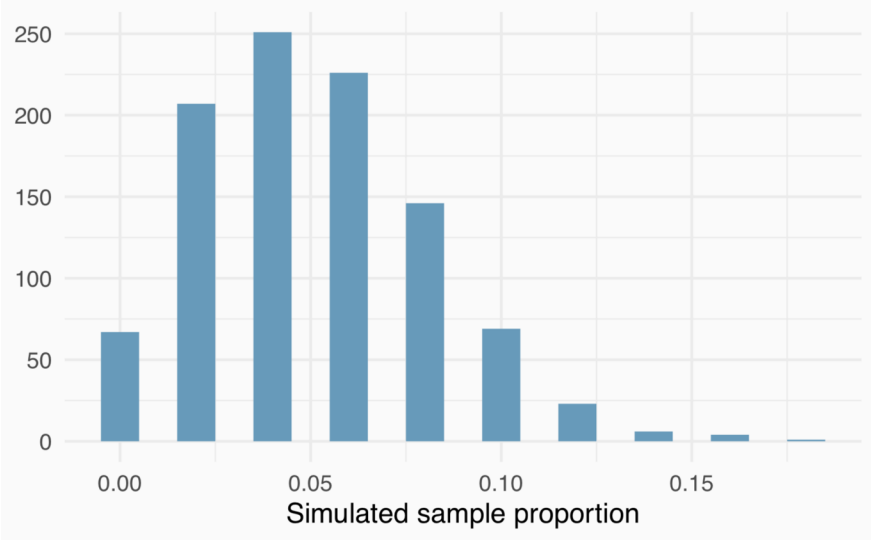
When np or $n(1 - p)$ is small

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?

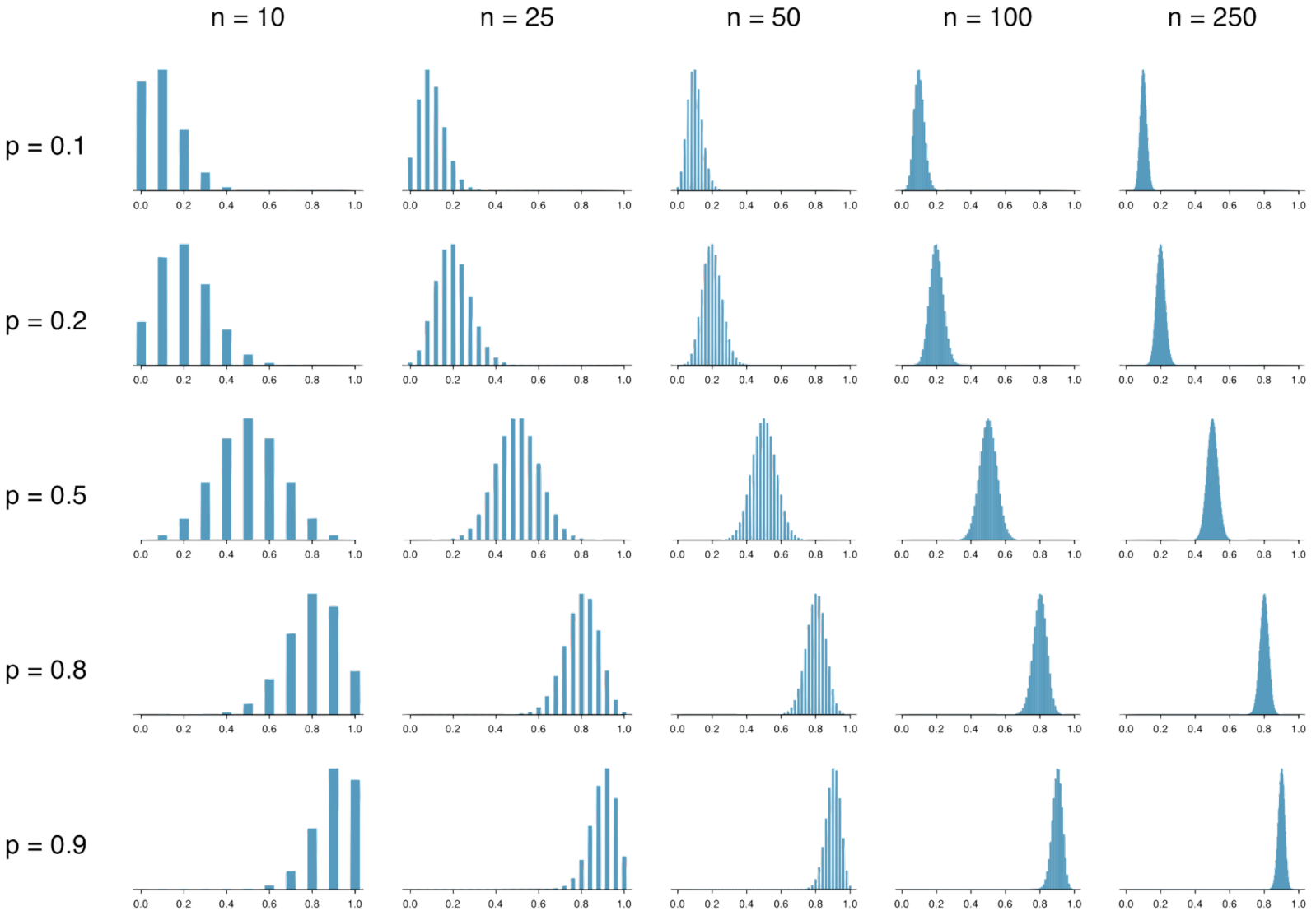
When *np* or *n(1 - p)* is small

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 CLT applies to other parameters as well, even if the details change a little.