

GVPT622 Lab Session

Hypothesis Testing I

Ryan Bookstein

October 23, 2026

University of Maryland

- ▶ Load the packages we'll use today and set the seed:

```
library(readxl)
library(tidyverse)
library(janitor)
library(ggdist)
library(MetBrewer)
library(DescTools)

set.seed(1234)
```

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?
- ▶ Today:

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?
- ▶ Today:
 - ▷ Populations, samples, and sampling error

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?
- ▶ Today:
 - ▷ Populations, samples, and sampling error
 - ▷ Sampling distributions and the Central Limit Theorem

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?
- ▶ Today:
 - ▷ Populations, samples, and sampling error
 - ▷ Sampling distributions and the Central Limit Theorem
 - ▷ Standard errors and confidence intervals

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?
- ▶ Today:
 - ▷ Populations, samples, and sampling error
 - ▷ Sampling distributions and the Central Limit Theorem
 - ▷ Standard errors and confidence intervals
 - ▷ Choosing a sample size

- ▶ Last lab: probability. If we know the true probabilities, what do repeated trials look like?
- ▶ Today we flip it around. We see one sample. What can it tell us about the population we can't see?
- ▶ Today:
 - ▷ Populations, samples, and sampling error
 - ▷ Sampling distributions and the Central Limit Theorem
 - ▷ Standard errors and confidence intervals
 - ▷ Choosing a sample size
 - ▷ A first hypothesis test

Population and Sample

- ▶ Say we want to know what share of likely voters will back a candidate. We can't ask every voter, so we ask a sample and generalize to the population

- ▶ Say we want to know what share of likely voters will back a candidate. We can't ask every voter, so we ask a sample and generalize to the population
- ▶ **Parameter:** the number we care about in the population. We *don't* know it

- ▶ Say we want to know what share of likely voters will back a candidate. We can't ask every voter, so we ask a sample and generalize to the population
- ▶ **Parameter:** the number we care about in the population. We *don't* know it
- ▶ **Statistic:** the same number computed in our sample. We *do* know it

- ▶ Say we want to know what share of likely voters will back a candidate. We can't ask every voter, so we ask a sample and generalize to the population
- ▶ **Parameter:** the number we care about in the population. We *don't* know it
- ▶ **Statistic:** the same number computed in our sample. We *do* know it
- ▶ Going from the statistic to the parameter is statistical inference

- ▶ Say we want to know what share of likely voters will back a candidate. We can't ask every voter, so we ask a sample and generalize to the population
- ▶ **Parameter:** the number we care about in the population. We *don't* know it
- ▶ **Statistic:** the same number computed in our sample. We *do* know it
- ▶ Going from the statistic to the parameter is statistical inference
- ▶ From lecture: a poll of 800 likely Maine voters found 50% for Jackson. That 50% is a statistic; the true share among all likely Maine voters is the parameter

What Makes a Good Sample?

- ▶ Random: in a pure random sample, every member of the population is equally likely to be drawn

What Makes a Good Sample?

- ▶ Random: in a pure random sample, every member of the population is equally likely to be drawn
 - ▷ Hard in practice! Phone polls miss people without landlines, people who screen unknown numbers, people at work during the day. . .

What Makes a Good Sample?

- ▶ Random: in a pure random sample, every member of the population is equally likely to be drawn
 - ▷ Hard in practice! Phone polls miss people without landlines, people who screen unknown numbers, people at work during the day. . .
- ▶ Large: the more observations, the closer we get to the truth (the law of large numbers)

What Makes a Good Sample?

- ▶ Random: in a pure random sample, every member of the population is equally likely to be drawn
 - ▷ Hard in practice! Phone polls miss people without landlines, people who screen unknown numbers, people at work during the day. . .
- ▶ Large: the more observations, the closer we get to the truth (the law of large numbers)
 - ▷ More is always better for inference, if not for your budget

What Makes a Good Sample?

- ▶ Random: in a pure random sample, every member of the population is equally likely to be drawn
 - ▷ Hard in practice! Phone polls miss people without landlines, people who screen unknown numbers, people at work during the day. . .
- ▶ Large: the more observations, the closer we get to the truth (the law of large numbers)
 - ▷ More is always better for inference, if not for your budget
- ▶ Even a large, random sample comes with some error. Let's explore why

Sampling Error

- ▶ Last lab's undecided member votes yea or nay with equal probability. We watch 100 votes:

```
possible_outcomes <- c("YEA", "NAY")  
  
votes_100 <- sample(possible_outcomes, 100, replace = T, prob = c(0.5,  
  0.5))  
table(votes_100)
```

```
votes_100  
NAY YEA  
  55  45
```

- ▶ Last lab's undecided member votes yea or nay with equal probability. We watch 100 votes:

```
possible_outcomes <- c("YEA", "NAY")  
  
votes_100 <- sample(possible_outcomes, 100, replace = T, prob = c(0.5,  
  0.5))  
table(votes_100)
```

```
votes_100  
NAY YEA  
  55  45
```

- ▶ The true probability is 0.5, so we “should” see 50 yeas. We got 45. That gap is sampling error

Many 100-Vote Trials

- ▶ Let's run that 100-vote trial 10,000 times and count the yeas each time:

```
whip_count <- function(possible_outcomes, n) {  
  
  outcomes <- sample(possible_outcomes, size = n, replace = T,  
                    prob = c(0.5, 0.5))  
  
  return(table(outcomes)["YEA"])  
  
}  
  
results <- tibble(trial = 1:1e4) |>  
  rowwise() |>  
  mutate(n_yea = whip_count(possible_outcomes, 100)) |>  
  ungroup()
```

Many 100-Vote Trials

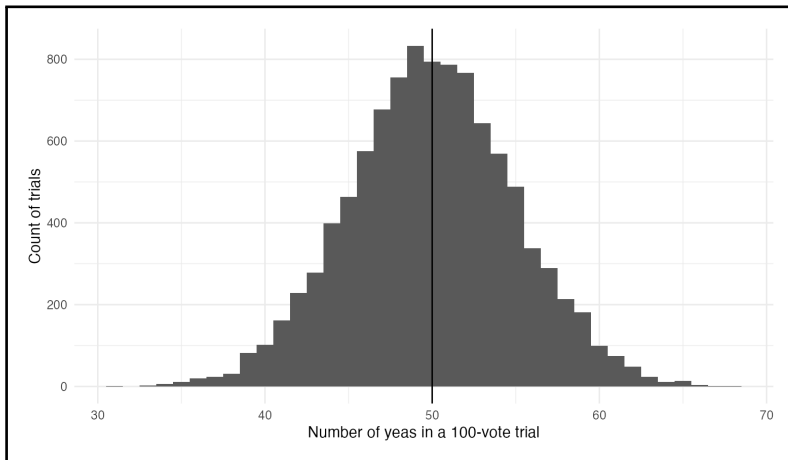
```
results
```

```
# A tibble: 10,000 x 2
  trial n_yea
  <int> <int>
1     1     52
2     2     52
3     3     48
4     4     60
5     5     61
# i 9,995 more rows
```

(Output trimmed to 5 rows.) Each row is one 100-vote trial. Same member, same 50/50 probability, different answers.

```
ggplot(results, aes(x = n_yea)) +  
  geom_histogram(binwidth = 1) +  
  geom_vline(xintercept = 50) +  
  theme_minimal() +  
  labs(x = "Number of yeas in a 100-vote trial",  
       y = "Count of trials")
```

The Distribution of Trials



Most trials land near 50 years, but some land far from it.

What the Trials Tell Us

```
mean(results$n_yea)  
range(results$n_yea)
```

```
[1] 49.9702  
[1] 31 68
```

- ▶ Some trials were way off: as few as 31 yeas, as many as 68

What the Trials Tell Us

```
mean(results$n_yea)  
range(results$n_yea)
```

```
[1] 49.9702  
[1] 31 68
```

- ▶ Some trials were way off: as few as 31 yeas, as many as 68
- ▶ But most cluster around 50, and the average is 49.97, almost exactly the truth

What the Trials Tell Us

```
mean(results$n_yea)  
range(results$n_yea)
```

```
[1] 49.9702  
[1] 31 68
```

- ▶ Some trials were way off: as few as 31 yeas, as many as 68
- ▶ But most cluster around 50, and the average is 49.97, almost exactly the truth
- ▶ That's what unbiased means: any one trial can miss, but the center of all the trials is the true value

Sampling Distributions

- ▶ What share of House members are Democrats?

Case Study: Democrats in the House

- ▶ What share of House members are Democrats?
- ▶ The CEL data cover every member in every Congress from the 93rd to the 118th (1973–2024): 11,606 member-terms

Case Study: Democrats in the House

- ▶ What share of House members are Democrats?
- ▶ The CEL data cover every member in every Congress from the 93rd to the 118th (1973–2024): 11,606 member-terms
- ▶ So for once we actually have the population, and we can compute the true parameter

Case Study: Democrats in the House

- ▶ What share of House members are Democrats?
- ▶ The CEL data cover every member in every Congress from the 93rd to the 118th (1973–2024): 11,606 member-terms
- ▶ So for once we actually have the population, and we can compute the true parameter
- ▶ Then we'll pretend we don't, draw samples, and see how close the samples get

Case Study: Democrats in the House

- ▶ What share of House members are Democrats?
- ▶ The CEL data cover every member in every Congress from the 93rd to the 118th (1973–2024): 11,606 member-terms
- ▶ So for once we actually have the population, and we can compute the true parameter
- ▶ Then we'll pretend we don't, draw samples, and see how close the samples get
- ▶ Note: a member who served five terms appears five times. Our population is member-terms, not people

Preparing the Data

```
cel <- read_excel("CELHouse93to118.xlsx", guess_max = 20000)

house <- cel |>
  rename(
    congress = "Congress number",
    name = "Legislator name, as given in THOMAS",
    dem = "1 = Democrat",
    female = "1 = female"
  ) |>
  select(congress, name, dem, female)
```

- ▶ dem and female are 0/1 indicators. We'll use female at the end of lab

```
head(house, 5)
```

```
# A tibble: 5 x 4
  congress name      dem female
  <dbl> <chr>      <dbl> <dbl>
1     93 Abdnor, James      0      0
2     93 Abzug, Bella       1      1
3     93 Adams, Brock       1      0
4     93 Addabbo, Joseph    1      0
5     93 Albert, Carl       1      0
```

```
tabyl(house, dem)
```

dem	n	percent
0	5317	0.4581251
1	6289	0.5418749

- ▶ 54.19% of all member-terms were Democrats. This is the parameter

```
tabyl(house, dem)
```

dem	n	percent
0	5317	0.4581251
1	6289	0.5418749

- ▶ 54.19% of all member-terms were Democrats. This is the parameter
- ▶ In research, you rarely get to see this

One Random Sample

- ▶ Draw a simple random sample of 1,000 member-terms. Every member-term is equally likely to be picked:

```
house_sample <- slice_sample(house, n = 1000)
tabyl(house_sample, dem)
```

dem	n	percent
0	446	0.446
1	554	0.554

One Random Sample

- ▶ Draw a simple random sample of 1,000 member-terms. Every member-term is equally likely to be picked:

```
house_sample <- slice_sample(house, n = 1000)
tabyl(house_sample, dem)
```

dem	n	percent
0	446	0.446
1	554	0.554

- ▶ 55.4%. Close to 54.19%, but not exactly

A Different Sample

```
house_sample_2 <- slice_sample(house, n = 1000)
tabyl(house_sample_2, dem)
```

dem	n	percent
0	457	0.457
1	543	0.543

- ▶ 54.3%. A different sample gives a different answer, just like our different 100-vote trials

A Different Sample

```
house_sample_2 <- slice_sample(house, n = 1000)
tabyl(house_sample_2, dem)
```

dem	n	percent
0	457	0.457
1	543	0.543

- ▶ 54.3%. A different sample gives a different answer, just like our different 100-vote trials
- ▶ This is sampling variability: a statistic changes from sample to sample because of random sampling error

1,000 Samples of 1,000

```
dem_survey <- function(df, n) {  
  
  slice_sample(df, n = n) |>  
    tabyl(dem) |>  
    filter(dem == 1) |>  
    pull(percent)  
  
}  
  
house_samples_1000 <- tibble(survey = 1:1000) |>  
  rowwise() |>  
  mutate(prop_dem = dem_survey(house, 1000)) |>  
  ungroup()
```

- ▶ `dem_survey()` draws one sample of size `n` and returns the share of Democrats in it

1,000 Samples of 1,000

```
house_samples_1000
```

```
# A tibble: 1,000 x 2
  survey prop_dem
  <int>   <dbl>
1     1     0.528
2     2     0.541
3     3     0.533
4     4     0.549
5     5     0.533
# i 995 more rows
```

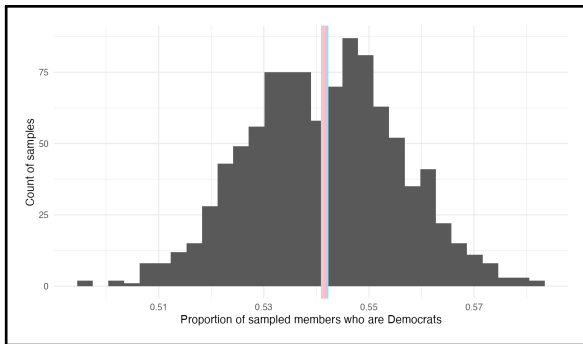
The Sampling Distribution

```
ggplot(house_samples_1000, aes(x = prop_dem)) +  
  geom_histogram() +  
  geom_vline(xintercept = tabyl(house, dem) |> filter(dem == 1) |>  
    pull(percent),  
             colour = "lightblue",  
             linewidth = 2) +  
  geom_vline(xintercept = mean(house_samples_1000$prop_dem),  
             colour = "pink",  
             linewidth = 2) +  
  theme_minimal() +  
  labs(x = "Proportion of sampled members who are Democrats",  
       y = "Count of samples")
```

The Sampling Distribution

```
mean(house_samples_1000$prop_dem)
```

```
[1] 0.541279
```



Pink: mean of the 1,000 samples (54.13%). **Blue:** the population parameter (54.19%).

- ▶ The average of our 1,000 sample proportions is 54.13%; the truth is 54.19%. With infinitely many samples, they'd match exactly

- ▶ The average of our 1,000 sample proportions is 54.13%; the truth is 54.19%. With infinitely many samples, they'd match exactly
- ▶ The sampling distribution is the distribution of a statistic across all possible samples of the same size

- ▶ The average of our 1,000 sample proportions is 54.13%; the truth is 54.19%. With infinitely many samples, they'd match exactly
- ▶ The sampling distribution is the distribution of a statistic across all possible samples of the same size
- ▶ The Central Limit Theorem: for large, random samples, the sampling distribution is approximately Normal, centered on the population parameter

The Central Limit Theorem

- ▶ The average of our 1,000 sample proportions is 54.13%; the truth is 54.19%. With infinitely many samples, they'd match exactly
- ▶ The sampling distribution is the distribution of a statistic across all possible samples of the same size
- ▶ The Central Limit Theorem: for large, random samples, the sampling distribution is approximately Normal, centered on the population parameter
- ▶ This holds for means, proportions, and counts, even though each member's dem is just a 0 or a 1

Inferring From a Single Sample

We Usually Get One Shot

- ▶ In real research we rarely get 1,000 samples. You can't rerun the 118th Congress

We Usually Get One Shot

- ▶ In real research we rarely get 1,000 samples. You can't rerun the 118th Congress
- ▶ And we usually don't know the population's size or shape

We Usually Get One Shot

- ▶ In real research we rarely get 1,000 samples. You can't rerun the 118th Congress
- ▶ And we usually don't know the population's size or shape
- ▶ Most of the time we have one statistic: one mean, one proportion, one count

We Usually Get One Shot

- ▶ In real research we rarely get 1,000 samples. You can't rerun the 118th Congress
- ▶ And we usually don't know the population's size or shape
- ▶ Most of the time we have one statistic: one mean, one proportion, one count
- ▶ What happens if we pick just one of our 1,000 samples?

Pick One Sample at Random

```
house_single <- slice_sample(house_samples_1000)
house_single
```

```
# A tibble: 1 x 2
  survey prop_dem
  <int>    <dbl>
1     628    0.523
```

- ▶ Sample 628 says 52.3%, about 2 points under the truth

Pick One Sample at Random

```
house_single <- slice_sample(house_samples_1000)
house_single
```

```
# A tibble: 1 x 2
  survey prop_dem
  <int>    <dbl>
1    628    0.523
```

- ▶ Sample 628 says 52.3%, about 2 points under the truth
- ▶ Not perfect, but close. Since most samples sit near the center, a random pick is more likely to be close than far

Pick One Sample at Random

```
house_single <- slice_sample(house_samples_1000)
house_single
```

```
# A tibble: 1 x 2
  survey prop_dem
  <int>     <dbl>
1     628     0.523
```

- ▶ Sample 628 says 52.3%, about 2 points under the truth
- ▶ Not perfect, but close. Since most samples sit near the center, a random pick is more likely to be close than far
- ▶ But we could get unlucky. How confident can we be in any one sample?

How Confident Can We Be?

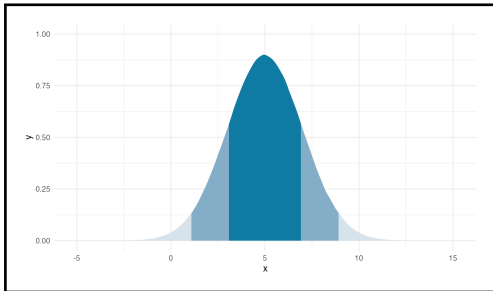
Properties of the Normal Distribution

```
norm_5_2 <- tibble(x = rnorm(n = 1e6, mean = 5, sd = 2))

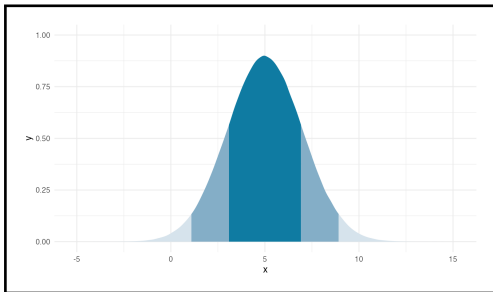
ggplot(norm_5_2, aes(x = x)) +
  stat_slab(
    aes(fill_ramp = after_stat(cut_cdf_qi(cdf))), fill =
      met.brewer("Egypt")[2]
  ) +
  scale_fill_ramp_discrete(range = c(1, 0.2), guide = "none") +
  theme_minimal()
```

- ▶ A Normal distribution with mean 5 and standard deviation 2, drawn 1,000,000 times

Properties of the Normal Distribution

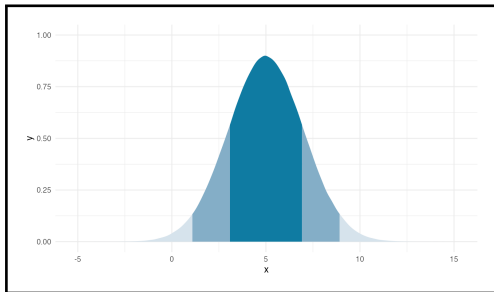


Properties of the Normal Distribution



The bands `cut_cdf_qi()` draws are the middle 66% (dark), the middle 95% (medium), and everything else (light). Close to the 68/95 rule, but not identical.

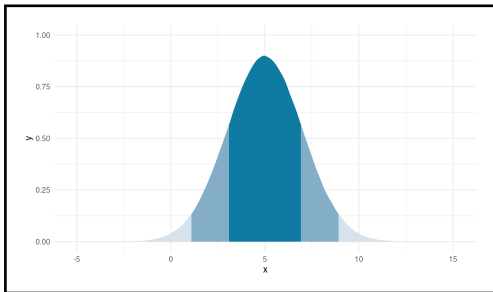
Properties of the Normal Distribution



- ▶ About 68% of the data fall within 1 SD of the mean

The bands `cut_cdf_qi()` draws are the middle 66% (dark), the middle 95% (medium), and everything else (light). Close to the 68/95 rule, but not identical.

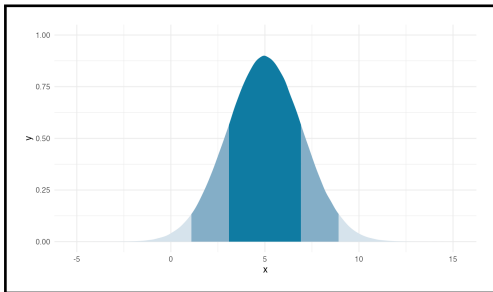
Properties of the Normal Distribution



- ▶ About 68% of the data fall within 1 SD of the mean
- ▶ About 95% fall within 1.96 (≈ 2) SDs

The bands `cut_cdf_qi()` draws are the middle 66% (dark), the middle 95% (medium), and everything else (light). Close to the 68/95 rule, but not identical.

Properties of the Normal Distribution



- ▶ About 68% of the data fall within 1 SD of the mean
- ▶ About 95% fall within 1.96 (≈ 2) SDs
- ▶ About 99.7% fall within 3 SDs

The bands `cut_cdf_qi()` draws are the middle 66% (dark), the middle 95% (medium), and everything else (light). Close to the 68/95 rule, but not identical.

From One Sample to a Sampling Distribution

- ▶ The trick: treat our one sample proportion as our best guess for the center of the sampling distribution

From One Sample to a Sampling Distribution

- ▶ The trick: treat our one sample proportion as our best guess for the center of the sampling distribution
- ▶ We know that distribution is Normal (CLT). To draw it, we just need its spread

From One Sample to a Sampling Distribution

- ▶ The trick: treat our one sample proportion as our best guess for the center of the sampling distribution
- ▶ We know that distribution is Normal (CLT). To draw it, we just need its spread
- ▶ The standard deviation of a sampling distribution is called the standard error. For a proportion:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

From One Sample to a Sampling Distribution

- ▶ The trick: treat our one sample proportion as our best guess for the center of the sampling distribution
- ▶ We know that distribution is Normal (CLT). To draw it, we just need its spread
- ▶ The standard deviation of a sampling distribution is called the standard error. For a proportion:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- ▶ $\hat{p}$ (“p-hat”) is the sample proportion. The hat marks a sample statistic

Our One Sample

```
my_sample <- slice_sample(house, n = 1000)

p_hat <- tabyl(my_sample, dem) |>
  filter(dem == 1) |>
  pull(percent)
p_hat

n <- 1000

se <- sqrt((p_hat * (1 - p_hat)) / n)
se
```

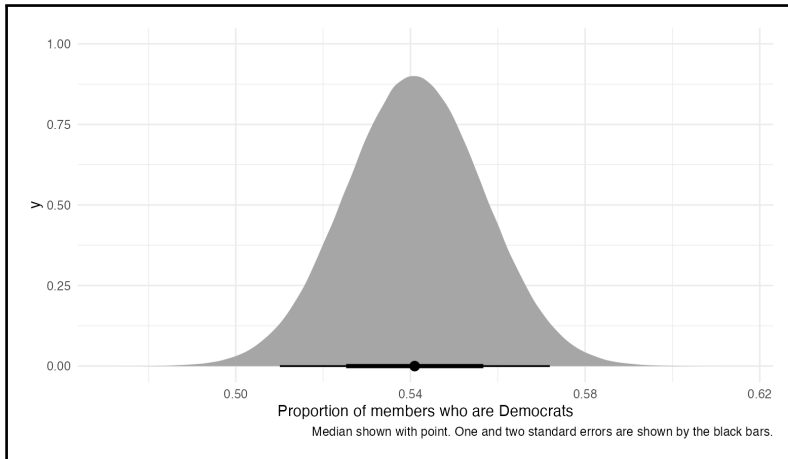
```
[1] 0.541
[1] 0.01575814
```

- A fresh sample: $\hat{p} = 54.1\%$, with a standard error of 1.6 points

Simulating Many Hypothetical Samples

```
ggplot(tibble(x = rnorm(1e6, mean = p_hat, sd = se)), aes(x = x)) +  
  stat_halfeye(.width = c(0.95, 0.68)) +  
  theme_minimal() +  
  labs(x = "Proportion of members who are Democrats",  
       caption = "Median shown with point. One and two standard errors  
are shown by the black bars.")
```

Simulating Many Hypothetical Samples



Our hypothetical sampling distribution, centered on $\hat{p} = 0.541$.

68% and 95% Intervals

68% and 95% Intervals

```
lower_68 <- p_hat - 1 * se  
lower_68
```

```
upper_68 <- p_hat + 1 * se  
upper_68
```

```
[1] 0.5252419
```

```
[1] 0.5567581
```

68% and 95% Intervals

```
lower_68 <- p_hat - 1 * se  
lower_68
```

```
upper_68 <- p_hat + 1 * se  
upper_68
```

```
[1] 0.5252419  
[1] 0.5567581
```

```
lower_95 <- p_hat - 1.96 * se  
lower_95
```

```
upper_95 <- p_hat + 1.96 * se  
upper_95
```

```
[1] 0.510114  
[1] 0.571886
```

68% and 95% Intervals

```
lower_68 <- p_hat - 1 * se  
lower_68
```

```
upper_68 <- p_hat + 1 * se  
upper_68
```

```
[1] 0.5252419  
[1] 0.5567581
```

```
lower_95 <- p_hat - 1.96 * se  
lower_95
```

```
upper_95 <- p_hat + 1.96 * se  
upper_95
```

```
[1] 0.510114  
[1] 0.571886
```

► 95% interval: 51.0% to 57.2%. It contains the true 54.19%

- ▶ **Point estimate:** our single best guess, $\hat{p} = 54.1\%$

- ▶ **Point estimate:** our single best guess, $\hat{p} = 54.1\%$
- ▶ **Interval estimate:** a range we're confident contains the parameter, 51.0% to 57.2%

- ▶ **Point estimate:** our single best guess, $\hat{p} = 54.1\%$
- ▶ **Interval estimate:** a range we're confident contains the parameter, 51.0% to 57.2%
- ▶ The margin of error is $1.96 \times se = 3.1$ points: “54.1%, give or take 3.1 points”

- ▶ **Point estimate:** our single best guess, $\hat{p} = 54.1\%$
- ▶ **Interval estimate:** a range we're confident contains the parameter, 51.0% to 57.2%
- ▶ The margin of error is $1.96 \times se = 3.1$ points: “54.1%, give or take 3.1 points”
- ▶ What “95% confident” means: if we repeated this method on many samples, about 95% of the intervals would contain the true parameter

- ▶ **Point estimate:** our single best guess, $\hat{p} = 54.1\%$
- ▶ **Interval estimate:** a range we're confident contains the parameter, 51.0% to 57.2%
- ▶ The margin of error is $1.96 \times se = 3.1$ points: “54.1%, give or take 3.1 points”
- ▶ What “95% confident” means: if we repeated this method on many samples, about 95% of the intervals would contain the true parameter
 - ▷ It's a statement about the method. Any one interval either contains the parameter, or it doesn't

- ▶ All of this leans hard on one sample. That's fine if the sample is large and random

- ▶ All of this leans hard on one sample. That's fine if the sample is large and random
- ▶ If it isn't random, the sampling distribution isn't centered on the parameter. That's bias, and a bigger sample won't fix it

- ▶ All of this leans hard on one sample. That's fine if the sample is large and random
- ▶ If it isn't random, the sampling distribution isn't centered on the parameter. That's bias, and a bigger sample won't fix it
- ▶ Know your data and your subject. The better you know them, the sooner you'll notice when a sample or a finding looks off

Choosing Your Sample Size

Standard Error and Sample Size

- ▶ n is in the denominator of $se(\hat{p})$: bigger samples, smaller standard errors, narrower intervals. But by how much?

```
se_from_sample <- function(n_samples) {  
  
  house |>  
    slice_sample(n = n_samples) |>  
    pull(dem) |>  
    MeanSE(na.rm = T)  
  
}
```

Standard Error and Sample Size

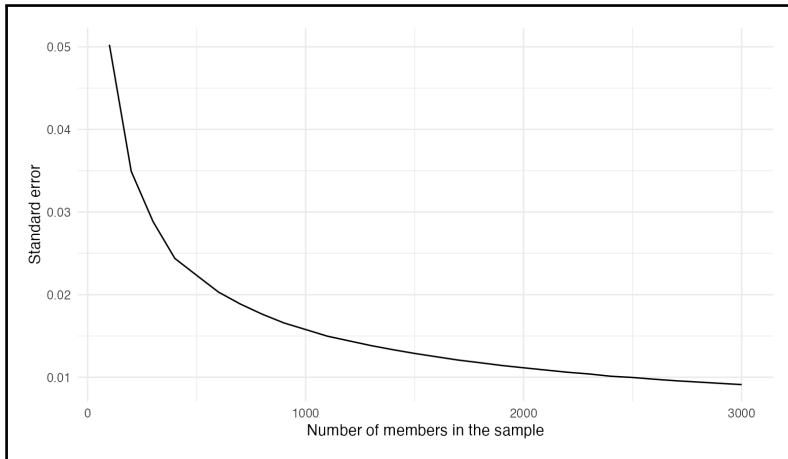
- ▶ n is in the denominator of $se(\hat{p})$: bigger samples, smaller standard errors, narrower intervals. But by how much?

```
se_from_sample <- function(n_samples) {  
  
  house |>  
    slice_sample(n = n_samples) |>  
    pull(dem) |>  
    MeanSE(na.rm = T)  
  
}
```

- ▶ MeanSE() (from DescTools) returns the standard error of a mean. The mean of a 0/1 variable is a proportion

```
tibble(sample_size = seq(from = 100, to = 3000, by = 100)) |>
  rowwise() |>
  mutate(se = se_from_sample(sample_size)) |>
  ggplot(aes(x = sample_size, y = se)) +
  geom_line() +
  theme_minimal() +
  labs(x = "Number of members in the sample",
       y = "Standard error")
```

Diminishing Returns



Steep at first, then flat: each extra member buys a smaller drop in the standard error.

- ▶ 100 $\rightarrow$ 1,000 members: the standard error drops from about 5 points to about 1.6

- ▶ 100 $\rightarrow$ 1,000 members: the standard error drops from about 5 points to about 1.6
- ▶ 1,000 $\rightarrow$ 1,900: it only drops to about 1.1

- ▶ 100 → 1,000 members: the standard error drops from about 5 points to about 1.6
- ▶ 1,000 → 1,900: it only drops to about 1.1
- ▶ Why? se shrinks with $\sqrt{n}$, so to halve your margin of error you need four times the sample

- ▶ 100 → 1,000 members: the standard error drops from about 5 points to about 1.6
- ▶ 1,000 → 1,900: it only drops to about 1.1
- ▶ Why? se shrinks with $\sqrt{n}$, so to halve your margin of error you need four times the sample
 - ▶ From lecture: the Maine poll's ± 3.5 at $n = 800$ only halves to ± 1.7 at $n = 3,200$

- ▶ 100 → 1,000 members: the standard error drops from about 5 points to about 1.6
- ▶ 1,000 → 1,900: it only drops to about 1.1
- ▶ Why? se shrinks with $\sqrt{n}$, so to halve your margin of error you need four times the sample
 - ▷ From lecture: the Maine poll's ± 3.5 at $n = 800$ only halves to ± 1.7 at $n = 3,200$
- ▶ More is always better, but every extra respondent buys you less, and survey costs keep adding up

Hypothesis Testing

- ▶ A hypothesis is a measurable, observable statement about a population

- ▶ A hypothesis is a measurable, observable statement about a population
 - ▷ Does trade dependence make war between states less likely?

- ▶ A hypothesis is a measurable, observable statement about a population
 - ▷ Does trade dependence make war between states less likely?
 - ▷ Does same-day registration increase turnout?

- ▶ A hypothesis is a measurable, observable statement about a population
 - ▷ Does trade dependence make war between states less likely?
 - ▷ Does same-day registration increase turnout?
- ▶ The frequentist approach: imagine a world in which you're wrong. That's the null hypothesis (H_0)

- ▶ A hypothesis is a measurable, observable statement about a population
 - ▷ Does trade dependence make war between states less likely?
 - ▷ Does same-day registration increase turnout?
- ▶ The frequentist approach: imagine a world in which you're wrong. That's the null hypothesis (H_0)
- ▶ Then ask: if H_0 were true, how likely would we be to see data like ours?

- ▶ Women in Congress: were most of them Democrats?

- ▶ Women in Congress: were most of them Democrats?
- ▶ **Alternative** (H_a): more than half of women member-terms were Democrats,
 $p > 0.5$

- ▶ Women in Congress: were most of them Democrats?
- ▶ **Alternative** (H_a): more than half of women member-terms were Democrats, $p > 0.5$
- ▶ **Null** (H_0): women were no more likely to be Democrats than Republicans, $p = 0.5$

- ▶ Women in Congress: were most of them Democrats?
- ▶ **Alternative** (H_a): more than half of women member-terms were Democrats, $p > 0.5$
- ▶ **Null** (H_0): women were no more likely to be Democrats than Republicans, $p = 0.5$
- ▶ The catch: we only get our one sample of 1,000 member-terms, `my_sample`

- ▶ Women in Congress: were most of them Democrats?
- ▶ **Alternative** (H_a): more than half of women member-terms were Democrats, $p > 0.5$
- ▶ **Null** (H_0): women were no more likely to be Democrats than Republicans, $p = 0.5$
- ▶ The catch: we only get our one sample of 1,000 member-terms, `my_sample`
- ▶ (Next week: comparing two groups, with difference-of-means and difference-of-proportions tests)

Women in Our Sample

```
my_women <- filter(my_sample, female == 1)
tabyl(my_women, dem)
```

dem	n	percent
0	37	0.2936508
1	89	0.7063492

- ▶ 126 women in our sample, and 70.6% of them are Democrats

Women in Our Sample

```
my_women <- filter(my_sample, female == 1)
tabyl(my_women, dem)
```

dem	n	percent
0	37	0.2936508
1	89	0.7063492

- ▶ 126 women in our sample, and 70.6% of them are Democrats
- ▶ That looks like support for H_a . But only 126 people: could this just be random noise?

The Standard Error

```
prop_women_dem <- mean(my_women$dem)
se_women <- sqrt((prop_women_dem * (1 - prop_women_dem)) /
  nrow(my_women))
se_women
```

```
[1] 0.04057327
```

- ▶ With only 126 observations, the standard error is 4.1 points, much wider than the 1.6 we had with 1,000

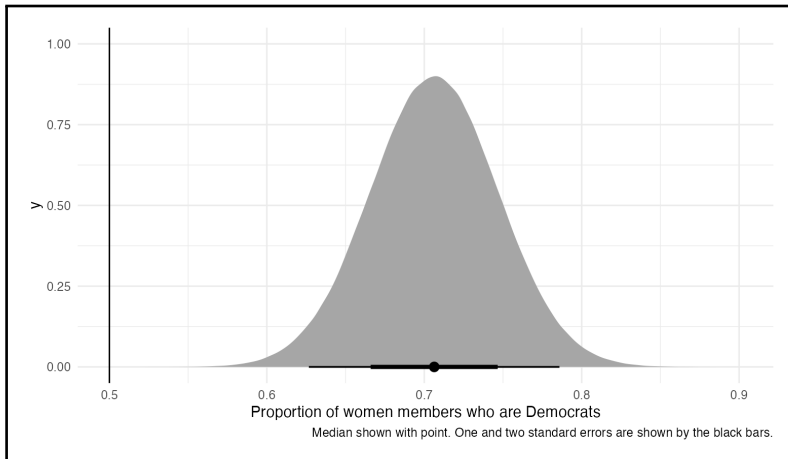
```
prop_women_dem <- mean(my_women$dem)
se_women <- sqrt((prop_women_dem * (1 - prop_women_dem)) /
  nrow(my_women))
se_women
```

```
[1] 0.04057327
```

- ▶ With only 126 observations, the standard error is 4.1 points, much wider than the 1.6 we had with 1,000
- ▶ 95% interval: $0.706 \pm 1.96 \times 0.041 = 62.7\%$ to 78.6%

```
ggplot(tibble(x = rnorm(1e6, mean = prop_women_dem, sd = se_women)),  
        aes(x = x)) +  
  stat_halfeye(.width = c(0.95, 0.68)) +  
  geom_vline(xintercept = 0.5) +  
  theme_minimal() +  
  labs(x = "Proportion of women members who are Democrats",  
        caption = "Median shown with point. One and two standard errors  
        are shown by the black bars.")
```

Adding the Null World



The null value (0.5, vertical line) is far outside the 95% interval.

- ▶ The vertical line at 0.5 is the null world

- ▶ The vertical line at 0.5 is the null world
- ▶ It sits far outside our 95% interval (62.7% to 78.6%). Almost none of our hypothetical samples come anywhere near it

- ▶ The vertical line at 0.5 is the null world
- ▶ It sits far outside our 95% interval (62.7% to 78.6%). Almost none of our hypothetical samples come anywhere near it
- ▶ If H_0 were true, it would be very unlikely to draw a random sample in which 70.6% of women are Democrats

- ▶ The vertical line at 0.5 is the null world
- ▶ It sits far outside our 95% interval (62.7% to 78.6%). Almost none of our hypothetical samples come anywhere near it
- ▶ If H_0 were true, it would be very unlikely to draw a random sample in which 70.6% of women are Democrats
- ▶ How unlikely? We can put a number on it

Test Statistic and p-value

```
se_null <- sqrt((0.5 * (1 - 0.5)) / nrow(my_women))  
z <- (prop_women_dem - 0.5) / se_null  
z  
  
2 * pnorm(-abs(z))
```

```
[1] 4.632528  
[1] 3.612271e-06
```

- ▶ Test statistic z : how many standard errors our estimate sits from the null value. Here, 4.6. (We use the SE under the null, with $p = 0.5$)

```
se_null <- sqrt((0.5 * (1 - 0.5)) / nrow(my_women))  
z <- (prop_women_dem - 0.5) / se_null  
z  
  
2 * pnorm(-abs(z))
```

```
[1] 4.632528  
[1] 3.612271e-06
```

- ▶ Test statistic z : how many standard errors our estimate sits from the null value. Here, 4.6. (We use the SE under the null, with $p = 0.5$)
- ▶ p-value: the probability of a result at least this extreme if H_0 were true.

- ▶ $p \approx 0.0000036$, far below the usual $\alpha = 0.05$

- ▶ $p \approx 0.0000036$, far below the usual $\alpha = 0.05$
- ▶ We reject the null: the share of women who are Democrats is statistically significantly different from 0.5

- ▶ $p \approx 0.0000036$, far below the usual $\alpha = 0.05$
- ▶ We reject the null: the share of women who are Democrats is statistically significantly different from 0.5
- ▶ This lines up with the interval: 0.5 is outside the 95% confidence interval, so we reject at the 0.05 level

- ▶ $p \approx 0.0000036$, far below the usual $\alpha = 0.05$
- ▶ We reject the null: the share of women who are Democrats is statistically significantly different from 0.5
- ▶ This lines up with the interval: 0.5 is outside the 95% confidence interval, so we reject at the 0.05 level
- ▶ A small p-value says the data would be surprising under H_0 . It doesn't tell us *why* women members skew Democratic

- Unlike in real research, we can look at the whole population:

```
house |>  
  filter(female == 1) |>  
  tabyl(dem)
```

dem	n	percent
0	424	0.2845638
1	1066	0.7154362

- ▶ Unlike in real research, we can look at the whole population:

```
house |>
  filter(female == 1) |>
  tabyl(dem)
```

dem	n	percent
0	424	0.2845638
1	1066	0.7154362

- ▶ The true share is 71.5%. Our sample of 126 said 70.6%, and the 95% interval (62.7% to 78.6%) caught it

- ▶ We use a sample statistic to learn about a population parameter

- ▶ We use a sample statistic to learn about a population parameter
- ▶ Every sample has sampling error, but with large, random samples the statistic is unbiased: its sampling distribution centers on the truth

- ▶ We use a sample statistic to learn about a population parameter
- ▶ Every sample has sampling error, but with large, random samples the statistic is unbiased: its sampling distribution centers on the truth
- ▶ The Central Limit Theorem: that sampling distribution is approximately Normal

- ▶ We use a sample statistic to learn about a population parameter
- ▶ Every sample has sampling error, but with large, random samples the statistic is unbiased: its sampling distribution centers on the truth
- ▶ The Central Limit Theorem: that sampling distribution is approximately Normal
- ▶ The standard error gives its spread. $\hat{p} \pm 1.96 \times se$ is a 95% confidence interval

- ▶ We use a sample statistic to learn about a population parameter
- ▶ Every sample has sampling error, but with large, random samples the statistic is unbiased: its sampling distribution centers on the truth
- ▶ The Central Limit Theorem: that sampling distribution is approximately Normal
- ▶ The standard error gives its spread. $\hat{p} \pm 1.96 \times se$ is a 95% confidence interval
- ▶ Standard errors shrink with $\sqrt{n}$: diminishing returns

- ▶ We use a sample statistic to learn about a population parameter
- ▶ Every sample has sampling error, but with large, random samples the statistic is unbiased: its sampling distribution centers on the truth
- ▶ The Central Limit Theorem: that sampling distribution is approximately Normal
- ▶ The standard error gives its spread. $\hat{p} \pm 1.96 \times se$ is a 95% confidence interval
- ▶ Standard errors shrink with $\sqrt{n}$: diminishing returns
- ▶ Hypothesis testing: how surprising would our data be in a world where we're wrong?

Center for Effective Lawmaking. 2025. "House Legislative Effectiveness Scores, 93rd–118th Congresses." <https://thelawmakers.org>.

Volden, Craig, and Alan E. Wiseman. 2014. *Legislative Effectiveness in the United States Congress: The Lawmakers*. New York: Cambridge University Press.

Questions?