The Sampling Distribution

Prof. Wells

Math 209, 3/1/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Review sampling activity from last week
- Discuss the framework for random sampling
- Investigate properties of the sampling distribution

Section 1

The Sampling Distribution

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
  - Ex: We may want to know average point value μ for a customized deck of playing cards

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
    - Ex: We may want to know average point value μ for a customized deck of playing cards
    - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
  - Ex: We may want to know average point value μ for a customized deck of playing cards
  - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.
  - The mean μ is a parameter, while the sample mean $\bar{x}$ is a statistic.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
  - Ex: We may want to know average point value μ for a customized deck of playing cards
  - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.
  - The mean μ is a parameter, while the sample mean $\bar{x}$ is a statistic.
- The sample statistics form a data set. The distribution of these sample statistics is called the **sampling distribution**.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
    - Ex: We may want to know average point value μ for a customized deck of playing cards
    - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.
    - The mean μ is a parameter, while the sample mean $\bar{x}$ is a statistic.
- The sample statistics form a data set. The distribution of these sample statistics is called the **sampling distribution**.
- In many circumstances, the mean of the sampling distribution is the *population parameter*.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
  - Ex: We may want to know average point value μ for a customized deck of playing cards
  - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.
  - The mean μ is a parameter, while the sample mean $\bar{x}$ is a statistic.
- The sample statistics form a data set. The distribution of these sample statistics is called the **sampling distribution**.
- In many circumstances, the mean of the sampling distribution is the *population parameter*.
  - This mean tells use the value of a typical statistic.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
  - Ex: We may want to know average point value μ for a customized deck of playing cards
  - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.
  - The mean μ is a parameter, while the sample mean $\bar{x}$ is a statistic.
- The sample statistics form a data set. The distribution of these sample statistics is called the **sampling distribution**.
- In many circumstances, the mean of the sampling distribution is the *population parameter*.
  - This mean tells use the value of a typical statistic.
- The standard deviation of the sampling distribution is called the **standard error**.

Sampling Distribution

- Researchers are interested in the value of a **parameter** in a population and use a **statistic** from a sample as a point estimate for the parameter.
  - Ex: We may want to know average point value μ for a customized deck of playing cards
  - We might not be able to take a census of the entire deck of cards (maybe it's kept hidden by a casino). So instead, we estimate the value of μ using the sample mean $\bar{x}$ in sample hand of 20 cards.
  - The mean μ is a parameter, while the sample mean $\bar{x}$ is a statistic.
- The sample statistics form a data set. The distribution of these sample statistics is called the **sampling distribution**.
- In many circumstances, the mean of the sampling distribution is the *population parameter*.
  - This mean tells use the value of a typical statistic.
- The standard deviation of the sampling distribution is called the **standard error**.
  - The standard deviation tells us how much the statistic varies from sample to sample.

Sampling Distribution vs. Population Distribution

- For most sample statistics and sufficiently large sample sizes ($n \geq 30$), the sampling distribution will be approximately bell-shaped (even if the population is not)

Sampling Distribution vs. Population Distribution

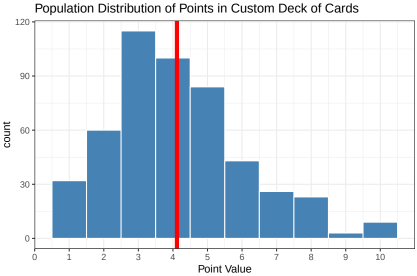
- For most sample statistics and sufficiently large sample sizes ($n \geq 30$), the sampling distribution will be approximately bell-shaped (even if the population is not)
- Additionally, the sampling distribution will have lower variability than the population distribution.

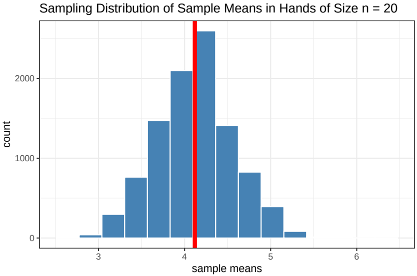
Sampling Distribution vs. Population Distribution

- For most sample statistics and sufficiently large sample sizes ($n \geq 30$), the sampling distribution will be approximately bell-shaped (even if the population is not)
- Additionally, the sampling distribution will have lower variability than the population distribution.
- Both distributions will have the same center.

Sampling Distribution vs. Population Distribution

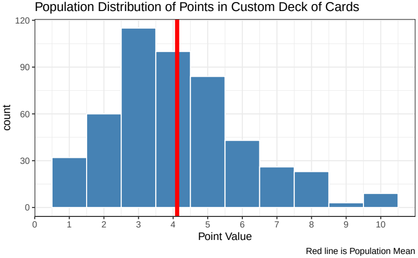
- For most sample statistics and sufficiently large sample sizes ($n \geq 30$), the sampling distribution will be approximately bell-shaped (even if the population is not)
- Additionally, the sampling distribution will have lower variability than the population distribution.
- Both distributions will have the same center.





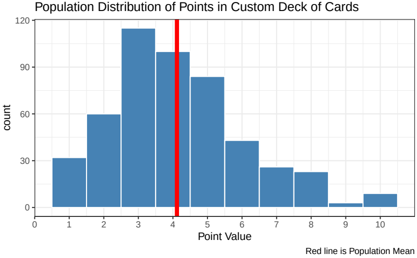
The Distributions Three

What we want to know:

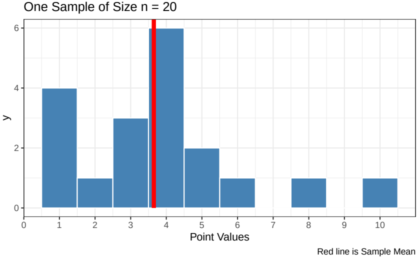


The Distributions Three

What we want to know:

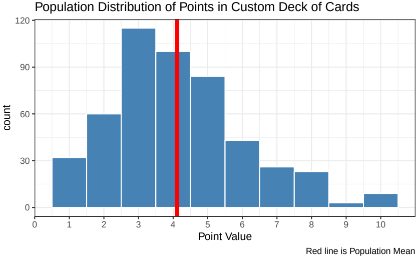


What we have:

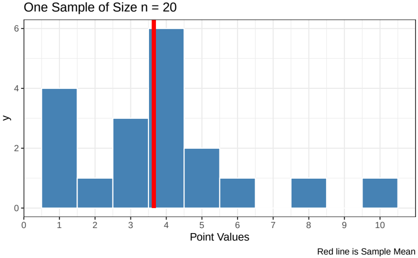


The Distributions Three

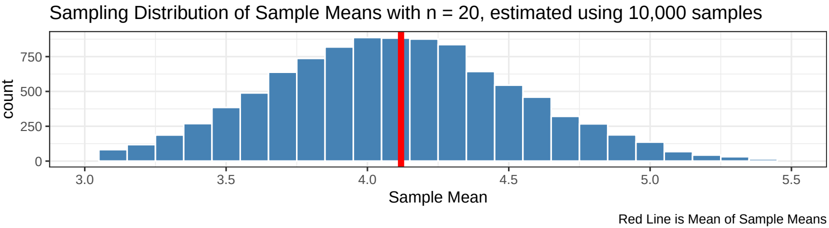
What we want to know:



What we have:



What we know about what we have:



Variability in Samples

- The standard error of a statistic (denoted SE), is the standard deviation of the sampling distribution.

Variability in Samples

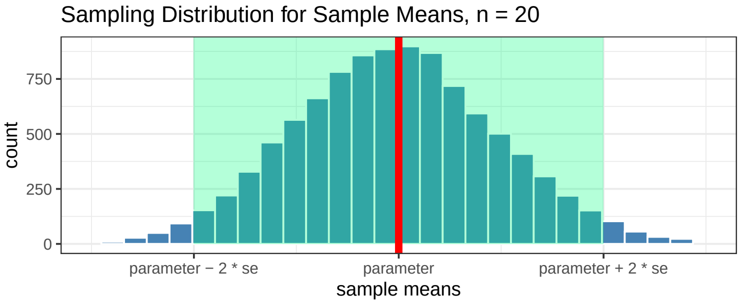
- The standard error of a statistic (denoted SE), is the standard deviation of the sampling distribution.
- For data with an approximately bell-shaped distribution, about 95% of observations fall within two standard deviations of the data's mean μ.

Variability in Samples

- The standard error of a statistic (denoted SE), is the standard deviation of the sampling distribution.
- For data with an approximately bell-shaped distribution, about 95% of observations fall within two standard deviations of the data's mean μ.
- The sampling distribution is approximately bell-shaped and centered at the population parameter, in most cases.
  - So about 95% of all sample statistics fall within 2 SE of the population parameter.

Variability in Samples

- The standard error of a statistic (denoted SE), is the standard deviation of the sampling distribution.
- For data with an approximately bell-shaped distribution, about 95% of observations fall within two standard deviations of the data's mean μ.
- The sampling distribution is approximately bell-shaped and centered at the population parameter, in most cases.
  - So about 95% of all sample statistics fall within 2 SE of the population parameter.

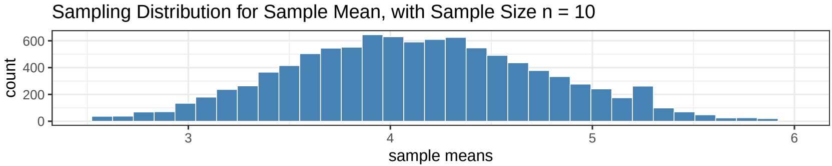


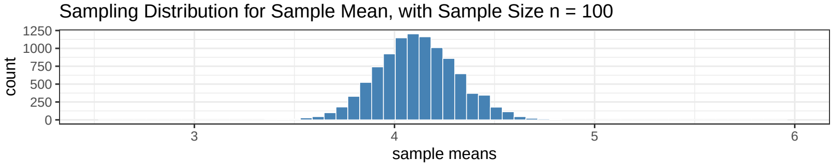
Standard Error and Sample Size

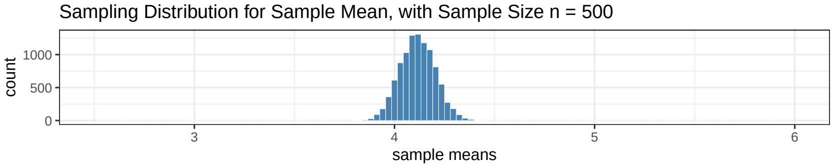
- How does the variability of the sampling distribution change as sample size changes?

Standard Error and Sample Size

- How does the variability of the sampling distribution change as sample size changes?







Variability and Sample Size II

- The sampling distributions for each of $n = 10$, $n = 100$, and $n = 1000$ are all approximately bell-shaped, and so 95% of sample means are within 2 standard errors of the sampling distribution mean.

Variability and Sample Size II

- The sampling distributions for each of $n = 10$, $n = 100$, and $n = 1000$ are all approximately bell-shaped, and so 95% of sample means are within 2 standard errors of the sampling distribution mean.
- We can compute the mean and standard deviation of each sampling distribution:

Variability and Sample Size II

- The sampling distributions for each of $n = 10$, $n = 100$, and $n = 1000$ are all approximately bell-shaped, and so 95% of sample means are within 2 standard errors of the sampling distribution mean.
- We can compute the mean and standard deviation of each sampling distribution:

Sample_Size	Mean	Standard_Deviation
10	4.12	0.63
100	4.12	0.20
500	4.12	0.09

Variability and Sample Size II

- The sampling distributions for each of $n = 10$, $n = 100$, and $n = 1000$ are all approximately bell-shaped, and so 95% of sample means are within 2 standard errors of the sampling distribution mean.
- We can compute the mean and standard deviation of each sampling distribution:

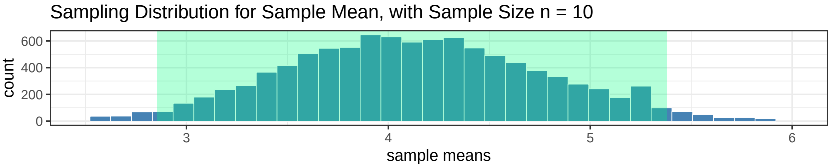
Sample_Size	Mean	Standard_Deviation
10	4.12	0.63
100	4.12	0.20
500	4.12	0.09

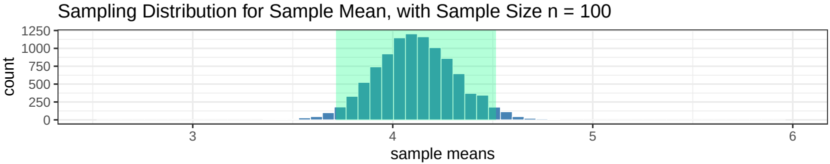
- Using $\text{mean} \pm 2 \cdot SE$, we can construct intervals for each distribution which contain 95% of all sample statistics:

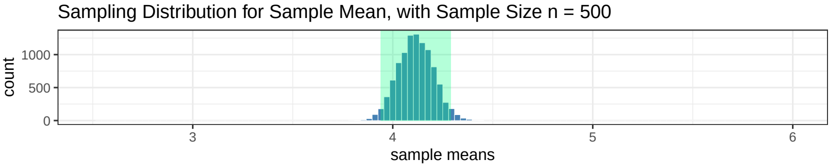
Sample_Size	Lower_Bound	Upper_Bound
10	2.86	5.38
100	3.72	4.52
500	3.94	4.29

Variability and Sample Size III

- Highlighted in green are the intervals containing 95% of all sample means:





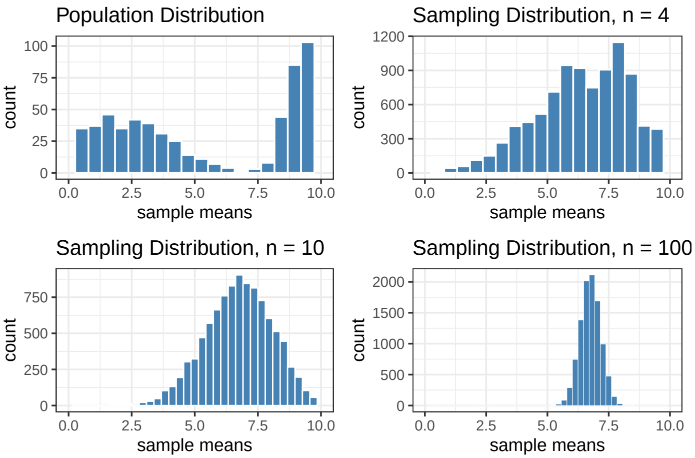


The Shape of the Sampling Distribution

- How does the shape of the sampling distribution change as sample size increases?

The Shape of the Sampling Distribution

- How does the shape of the sampling distribution change as sample size increases?



Section 2

Sampling Example

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking
  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.
- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9$ million)

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.
- **Sampling Method**: SRS(?) of size $n = 1020$ obtained using phone-numbers

Presidential Polling

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking
  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.
- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.
- **Sampling Method**: SRS(?) of size $n = 1020$ obtained using phone-numbers
- **Point Estimate/Sample Statistic**: The sample proportion $\hat{p}$ of Americans who plan to vote for Trump/Pence. In this case, $\hat{p} = 0.46$.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
  - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
  - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
  - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
  - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
  - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
  - But if we just want an estimate that is likely close true proportion, then we should be very confidant.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
  - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
  - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
  - But if we just want an estimate that is likely close true proportion, then we should be very confidant.
- The sampling distribution tells us how much variability to expect from sample to sample.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
    - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
    - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
    - But if we just want an estimate that is likely close true proportion, then we should be very confidant.
- The sampling distribution tells us how much variability to expect from sample to sample.
    - Using probability theory, we can show that the standard error for the sampling distribution of the proportion with sample size n is at most $\frac{1}{2\sqrt{n}}$

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
  - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
  - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
  - But if we just want an estimate that is likely close true proportion, then we should be very confidant.
- The sampling distribution tells us how much variability to expect from sample to sample.
  - Using probability theory, we can show that the standard error for the sampling distribution of the proportion with sample size n is at most $\frac{1}{2\sqrt{n}}$
  - For a sample of size $n = 1020$, the standard error is at most $\frac{1}{2\sqrt{1020}} = 0.016$.

Sampling Variability

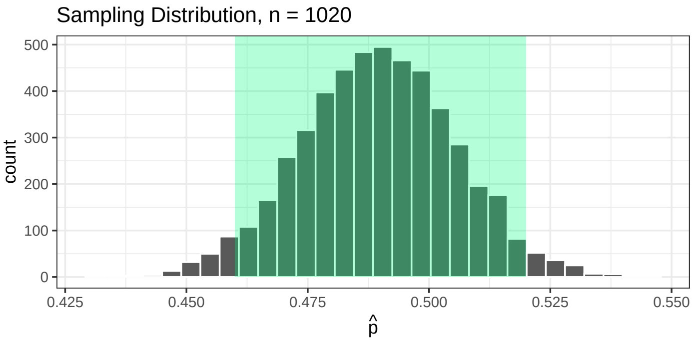
- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$

Sampling Variability

- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.

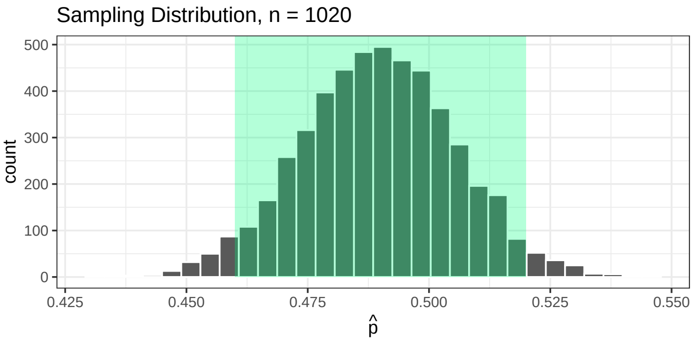
Sampling Variability

- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.



Sampling Variability

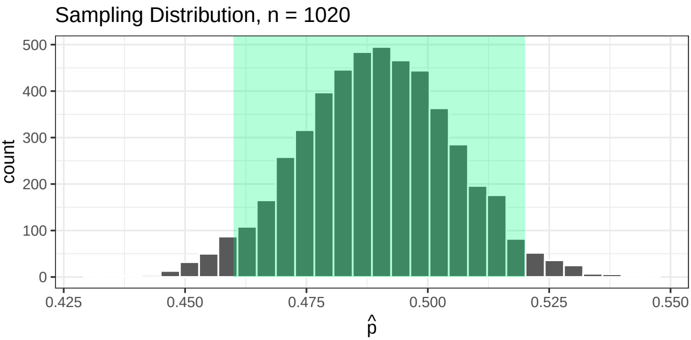
- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.



- Of these, only 6% differed from the true value $p = .49$ by more than .03

Sampling Variability

- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.



- Of these, only 6% differed from the true value $p = .49$ by more than .03
- But this also means that for 94% of samples, the true proportion p is within 0.03 of the sample proportion $\hat{p}$.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.
  - But if we can collect enough samples to form the sampling distribution, we probably can just take a census of the population.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.
  - But if we can collect enough samples to form the sampling distribution, we probably can just take a census of the population.
- The fix?

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.
  - But if we can collect enough samples to form the sampling distribution, we probably can just take a census of the population.
- The fix?
  - Discussed in class on Friday!