Framework of Random Sampling

Nate Wells

Math 141, 3/1/22

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Review Monday's group sampling activity
- Discuss the framework for random sampling
- Investigate properties of the sampling distribution

Section 1

Sampling Activity

Sampling Activity Discussion

- What is the theoretical mean value for the data set of card values?
- How does the distribution of sample means compare to the distribution of card values?
- What is the relationship between the centers of the two distributions?
- Which distribution appears to have more variability?
- How do the shapes of the two distributions compare?
- What does the variability of sample means suggest about the means in repeated samples?

Section 2

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 paramter.

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 paramter.
- Ex: We may want to know the proportion p of Reed community members infected with COVID-19 on March 2nd, 2022.
  - We cannot easily take a census of all Reed community members (at least on every day of the year), so we estimate p by using the proportion $\hat{p}$ in a sample of 100 individuals.
  - The proportion p is a parameter, while the proportion $\hat{p}$ 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 paramter.
- Ex: We may want to know the proportion p of Reed community members infected with COVID-19 on March 2nd, 2022.
  - We cannot easily take a census of all Reed community members (at least on every day of the year), so we estimate p by using the proportion $\hat{p}$ in a sample of 100 individuals.
  - The proportion p is a parameter, while the proportion $\hat{p}$ is a statistic.
- The sample statistics form a data set, so have their own mean, standard deviation (called the **standard error**), and distribution (called the **sampling distribution**)
  - Using theoretical tools, we can show that if the true proportion is $p = 0.002$, then the sampling distribution for $\hat{p}$ has mean 0.002 and standard error

$$SE = \sqrt{\frac{0.002(1 - .002)}{100}} \approx 0.004$$

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 paramter.
- Ex: We may want to know the proportion p of Reed community members infected with COVID-19 on March 2nd, 2022.
  - We cannot easily take a census of all Reed community members (at least on every day of the year), so we estimate p by using the proportion $\hat{p}$ in a sample of 100 individuals.
  - The proportion p is a parameter, while the proportion $\hat{p}$ is a statistic.
- The sample statistics form a data set, so have their own mean, standard deviation (called the **standard error**), and distribution (called the **sampling distribution**)
  - Using theoretical tools, we can show that if the true proportion is $p = 0.002$, then the sampling distribution for $\hat{p}$ has mean 0.002 and standard error

$$SE = \sqrt{\frac{0.002(1 - .002)}{100}} \approx 0.004$$

  - That is, typically, a sample will have proportion $\hat{p} = 0.002$, and most samples will have proportion between 0.00 and 0.01.

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

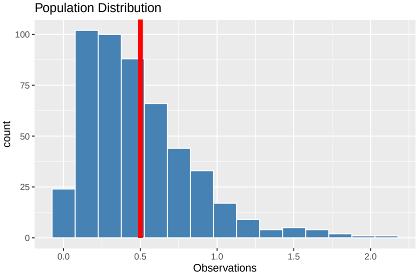
- For most sample statistics and sufficiently large sample sizes ($n \geq 30$), the sampling distribution will be approximately bell-shaped (even if the population is not)
- Additionally, the sampling distribution will have lower variability than the population distribution.

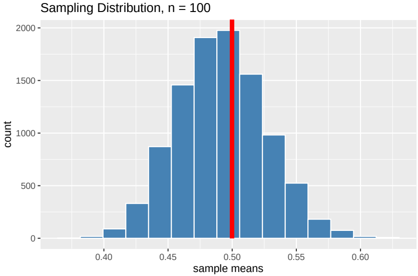
Sampling Distribution vs. Population Distribution

- 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

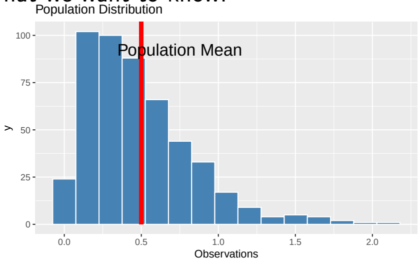
- For most sample statistics and sufficiently large sample sizes ($n \geq 30$), the sampling distribution will be approximately bell-shaped (even if the population is not)
- Additionally, the sampling distribution will have lower variability than the population distribution.
- Both distributions will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





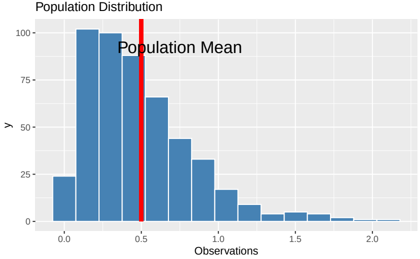
Why Use Sampling Distributions?

What we want to know:

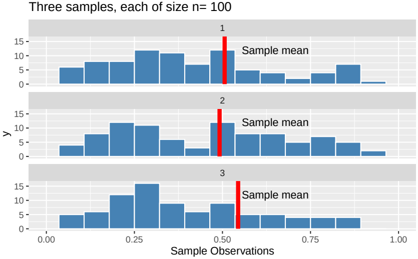


Why Use Sampling Distributions?

What we want to know:

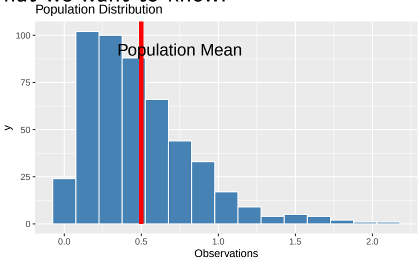


What we have:

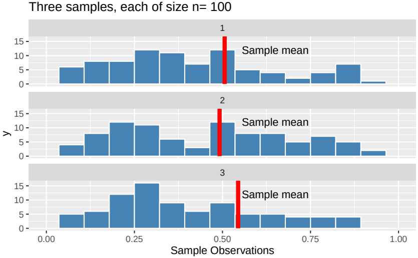


Why Use Sampling Distributions?

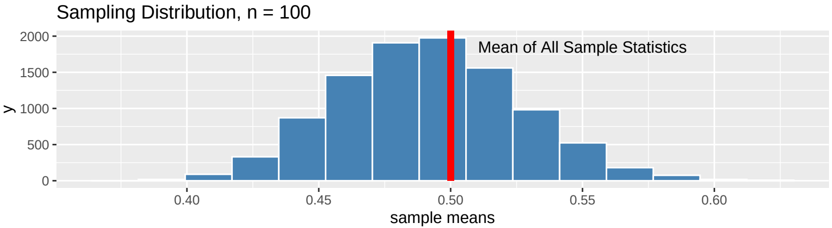
What we want to know:



What we have:



What we know about what we have:



Variability in Samples

- The standard error of the sample statistic measures variability between different samples.

Variability in Samples

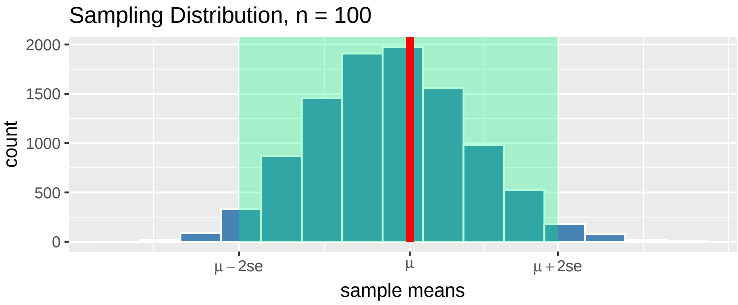
- The standard error of the sample statistic measures variability between different samples.
- For approximately bell-shaped distributions, about 95% of observations fall within two standard deviations of the population's mean μ.

Variability in Samples

- The standard error of the sample statistic measures variability between different samples.
- For approximately bell-shaped distributions, about 95% of observations fall within two standard deviations of the population's mean μ.
- Since the sampling distribution is approximately bell-shaped for most sample statistics, and is centered at the population mean, then 95% of all sample statistics fall within 2 standard error units of the population mean μ.

Variability in Samples

- The standard error of the sample statistic measures variability between different samples.
- For approximately bell-shaped distributions, about 95% of observations fall within two standard deviations of the population's mean μ.
- Since the sampling distribution is approximately bell-shaped for most sample statistics, and is centered at the population mean, then 95% of all sample statistics fall within 2 standard error units of the population mean μ.

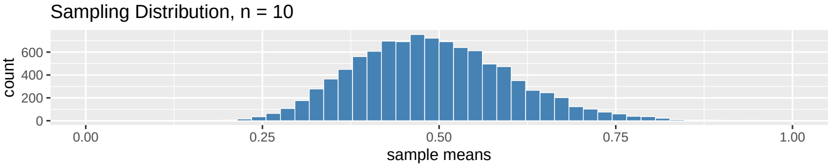


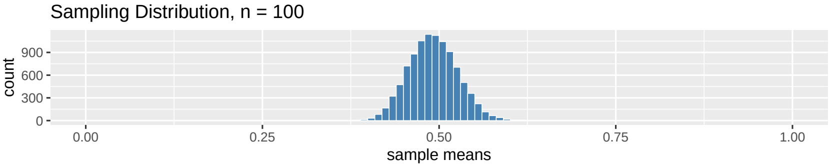
Standard Error and Sample Size

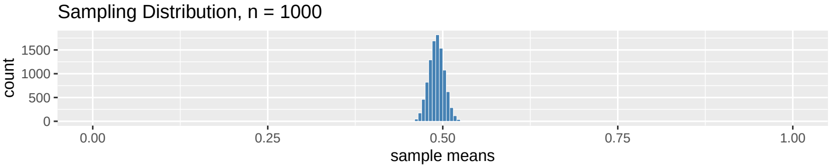
- How does the variability of the sampling distribution change as sample size changes?

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 $n = 10, 100, 1000$ are all approximately Normal, and so 95% of sample means are within 2 standard error units of the sampling distribution mean.

Variability and Sample Size II

- The sampling distributions for $n = 10, 100, 1000$ are all approximately Normal, and so 95% of sample means are within 2 standard error units of the sampling distribution mean.
- We can approximate the mean and standard error of each sampling distribution, and construct intervals which contain 95% of all sample means:

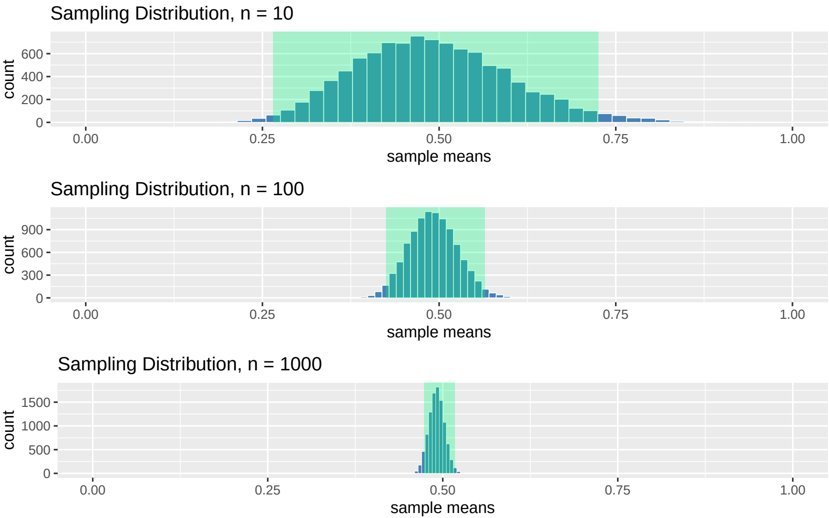
Variability and Sample Size II

- The sampling distributions for $n = 10, 100, 1000$ are all approximately Normal, and so 95% of sample means are within 2 standard error units of the sampling distribution mean.
- We can approximate the mean and standard error of each sampling distribution, and construct intervals which contain 95% of all sample means:

n	mean	standard error	lower	upper
10	0.5	0.11	0.28	.72
100	0.5	0.035	0.43	0.57
1000	0.5	0.011	0.48	0.52

Variability and Sample Size III

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

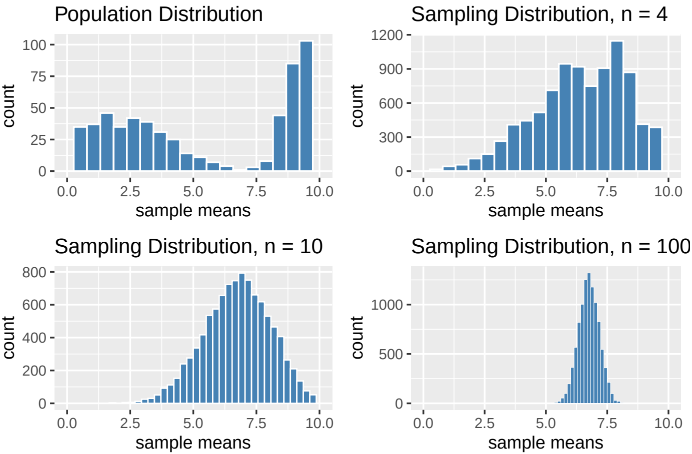


The Shape of the Sampling Distribution

- 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?



Polling Example

- 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.

Polling Example

- 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.

Polling Example

- 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.
  - The survey lists a margin of error of $\pm 3.8\%$, with 95% confidence (we'll discuss this next week)

Polling Example

- 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.
  - The survey lists a margin of error of $\pm 3.8\%$, with 95% confidence (we'll discuss this next week)
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris
- **Census Result**: We could compute the exact value of p by meticulously asking every registered voter in the population whether they plan to vote for Biden/Harris

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris
- **Census Result**: We could compute the exact value of p by meticulously asking every registered voter in the population whether they plan to vote for Biden/Harris
- **Sampling Method**: SRS(?) of size $n = 1020$ obtained using phone-numbers.

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris
- **Census Result**: We could compute the exact value of p by meticulously asking every registered voter in the population whether they plan to vote for Biden/Harris
- **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 Biden/Harris. In this case, $\hat{p} = 0.5$.

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris
- **Census Result**: We could compute the exact value of p by meticulously asking every registered voter in the population whether they plan to vote for Biden/Harris
- **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 Biden/Harris. In this case, $\hat{p} = 0.5$.
- Is the sampling procedure **representative**? Perhaps. Five-Thirty-Eight gives this pollster an A+ rating for its use of statistical weighting procedures to account for deviations in sample from known population characteristics.

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9$ million)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris
- **Census Result**: We could compute the exact value of p by meticulously asking every registered voter in the population whether they plan to vote for Biden/Harris
- **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 Biden/Harris. In this case, $\hat{p} = 0.5$.
- Is the sampling procedure **representative**? Perhaps. Five-Thirty-Eight gives this pollster an A+ rating for its use of statistical weighting procedures to account for deviations in sample from known population characteristics.
- Are the results **generalizable**? Yes, provided the sample was obtained randomly from the population.

Polling using Sampling Framework

- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Biden/Harris
- **Census Result**: We could compute the exact value of p by meticulously asking every registered voter in the population whether they plan to vote for Biden/Harris
- **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 Biden/Harris. In this case, $\hat{p} = 0.5$.
- Is the sampling procedure **representative**? Perhaps. Five-Thirty-Eight gives this pollster an A+ rating for its use of statistical weighting procedures to account for deviations in sample from known population characteristics.
- Are the results **generalizable**? Yes, provided the sample was obtained randomly from the population.
- Is it **biased**? Yes. Although hopefully bias was reduced through use of survey weighting.