COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 17: Central Limit Theorem

Nov 17, 2025

Reminders

- HW 6 due tonight
- Extra Credit (HW 5 Question 3) due tonight
- Final Project Proposal Due Wednesday, Nov 19
- HW 7 due next week Monday (skip Question 4 about the survey)

Data Science in this course

- Exploration: Discover patterns in data and articulate insights (visualizations)

- Inference: Make reliable conclusions about the world

  - Statistics is useful

- **Prediction: Informed guesses about unseen data**

Last Class: Normal Distributions

Standard Deviation (SD)

Standard deviation measures the variability around the mean

$$\sigma = \sqrt{\text{avg}\left((v - \mu)^2 \text{ for } v \in \vec{V}\right)}$$

- No matter the shape of the distribution, the bulk of the data is in the range “average plus or minus a few standard deviations”

Standard Units

- The quantity z (from "$\text{average} \pm z\ \text{SDs}$" in Chebychev's inequality) measures **standard units**

  - **Standard units** is the number of standard deviations away from the average

- To convert a value (v) to standard units, compare the deviation from the average (μ) with the standard deviation (SD):

$$z = \frac{v - \mu}{\text{SD}}$$

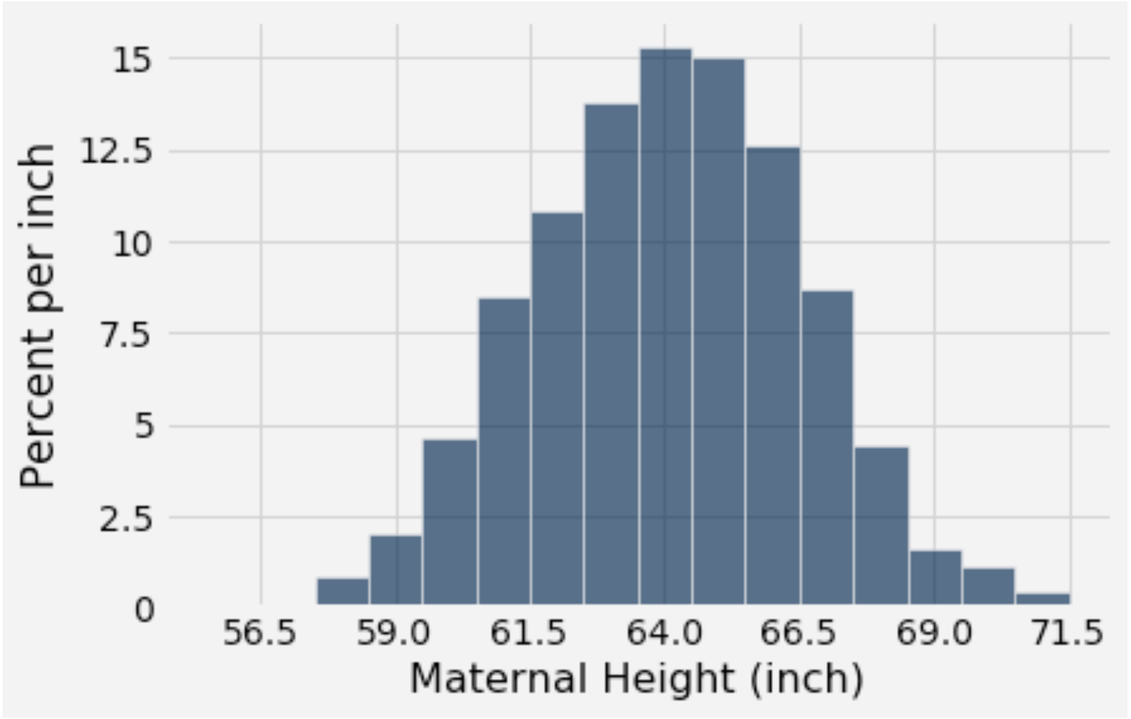
Bell Shaped Curves

- The normal curve / bell-curve is a very common distribution

- For bell-shaped (aka Gaussian distribution):

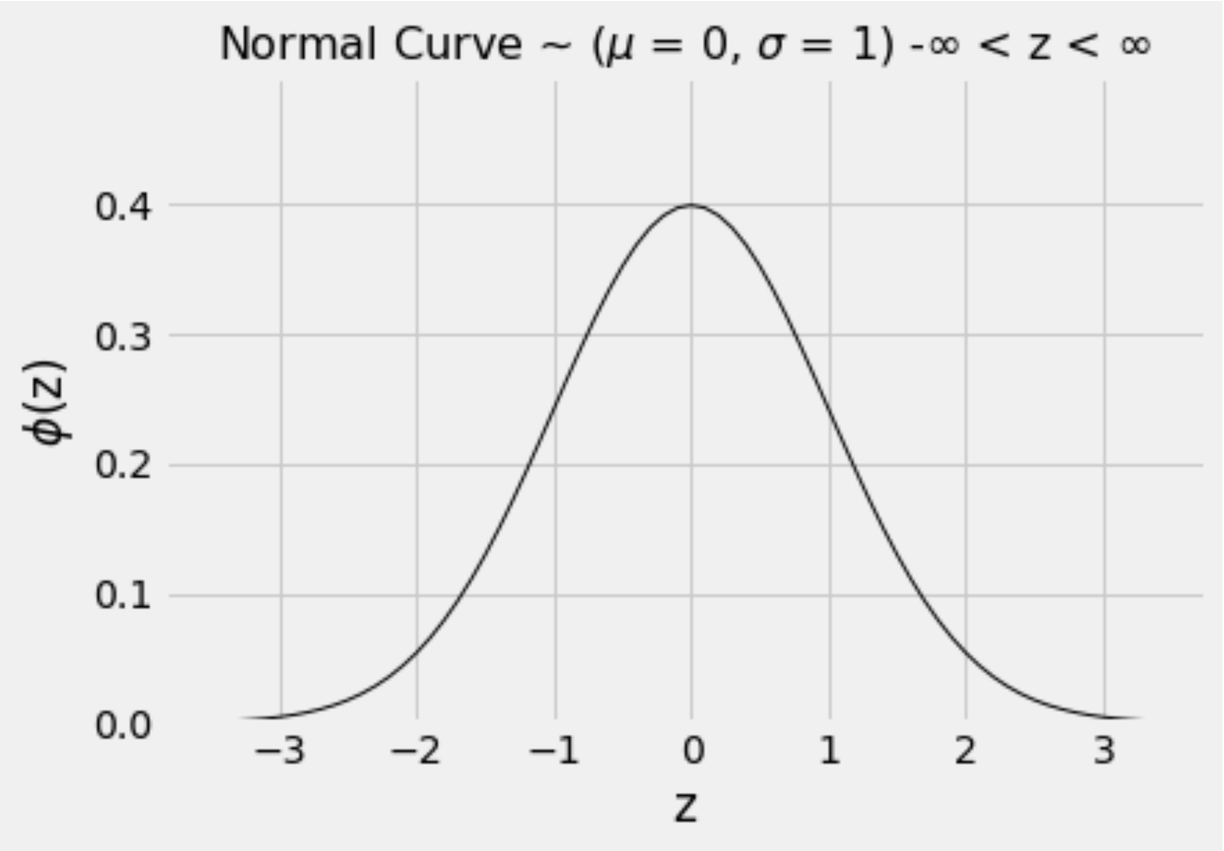
  - Average is at the center

  - SD is the distance between the average and the points of inflection on either side



Normal Distribution

- On a standard normal curve, x-axis units are standard units
- Total area of the curve is 1
- Curve is symmetric around 0 (mean and median are both 0)
- Points of inflection are -1 and 1
    - Standard deviation is 1

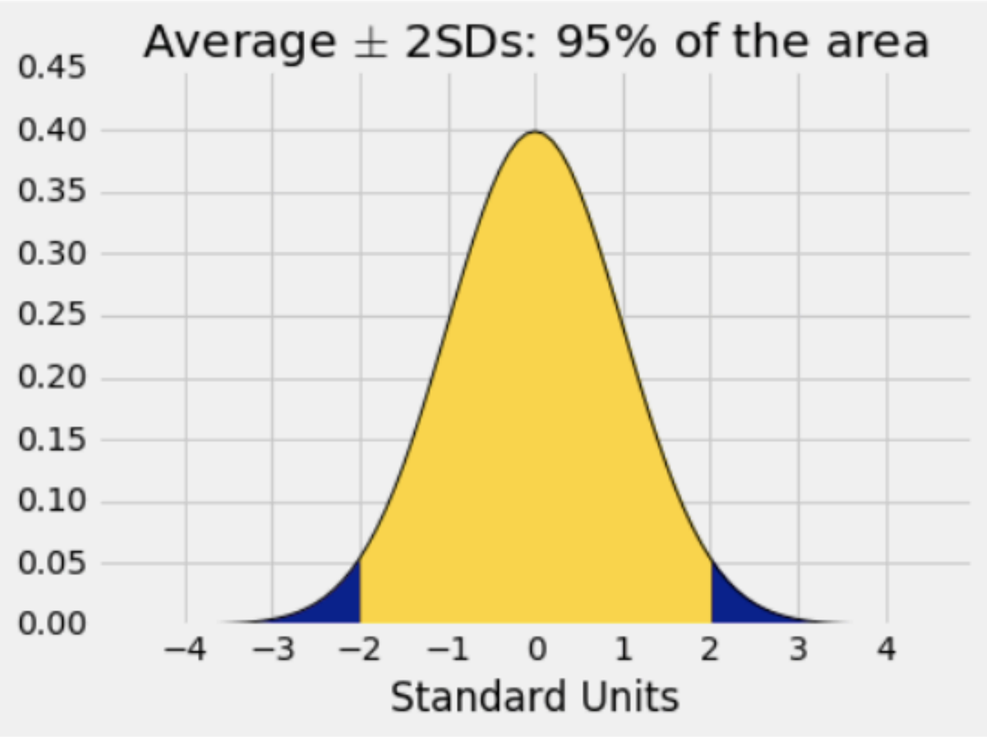


Application to Normal Distributions

- If a histogram is bell-shaped (normal), then 95% of the data is in the range
  $\text{average} \pm 2\ \text{SDs}$

- Note this is much higher than Chebychev's bound of 75%

  - 75% is a lower bound that applies to *all* distributions



Normal vs All Distributions

Range	All Distributions (Chebyshev's)	Normal Distribution
mean $\pm$ 1 SDs	At least 0%	At least 68%
mean $\pm$ 2 SDs	At least 75%	At least 95%
mean $\pm$ 3 SDs	At least 89%	At least 99%

Central Limit Theorem

- Describes how a normal distribution is connected to random sample averages (which helps us determine the population average)

- **Central Limit Theorem**: If a sample is large and drawn at random with replacement, then regardless of the distribution the **probability distribution of the sample average** is roughly normal

Central Limit Theorem

Central Limit Theorem

- Describes how a normal distribution is connected to random sample averages (the average of a sample we collect)

- We calculate sample averages because they can help us estimate population averages

Central Limit Theorem

Definition:

Is a sample is large and drawn at random with replacement

Then regardless of the distribution,

The probability distribution of the sample average is roughly normal

Distribution of Sample Averages

- If you have one random sample, you can take the average of values in that sample (one sample average)

- But that sample average could be different if you took different sample

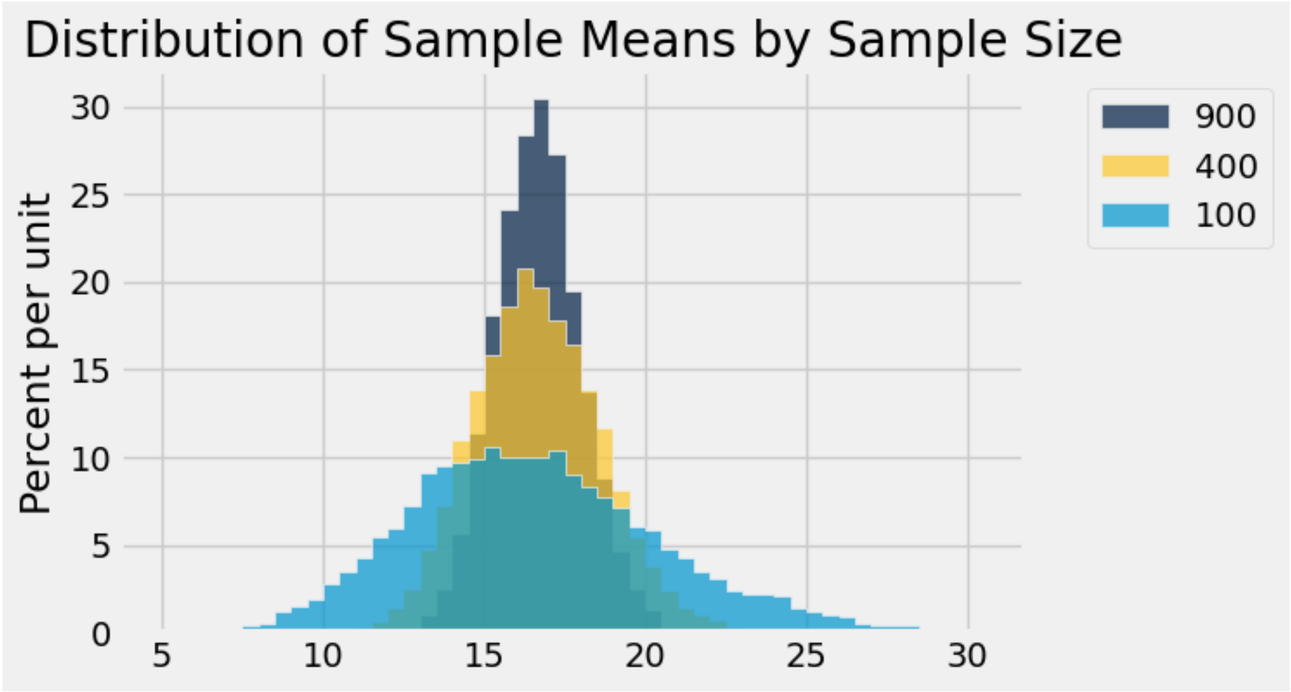
- Distribution of sample averages is the distribution of possible sample averages if you were to draw different samples

Notebook Demo: Central Limit Theorem

Distribution of Sample Means & Sample Size

When increasing sample size,

- Distributions get narrower (closer to the true mean)
- Also get taller (higher probability around true mean)



Notebook Demo: Standard Deviation + Sample Size

Central Limit Theorem for Sample Mean

Definition:

If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population,

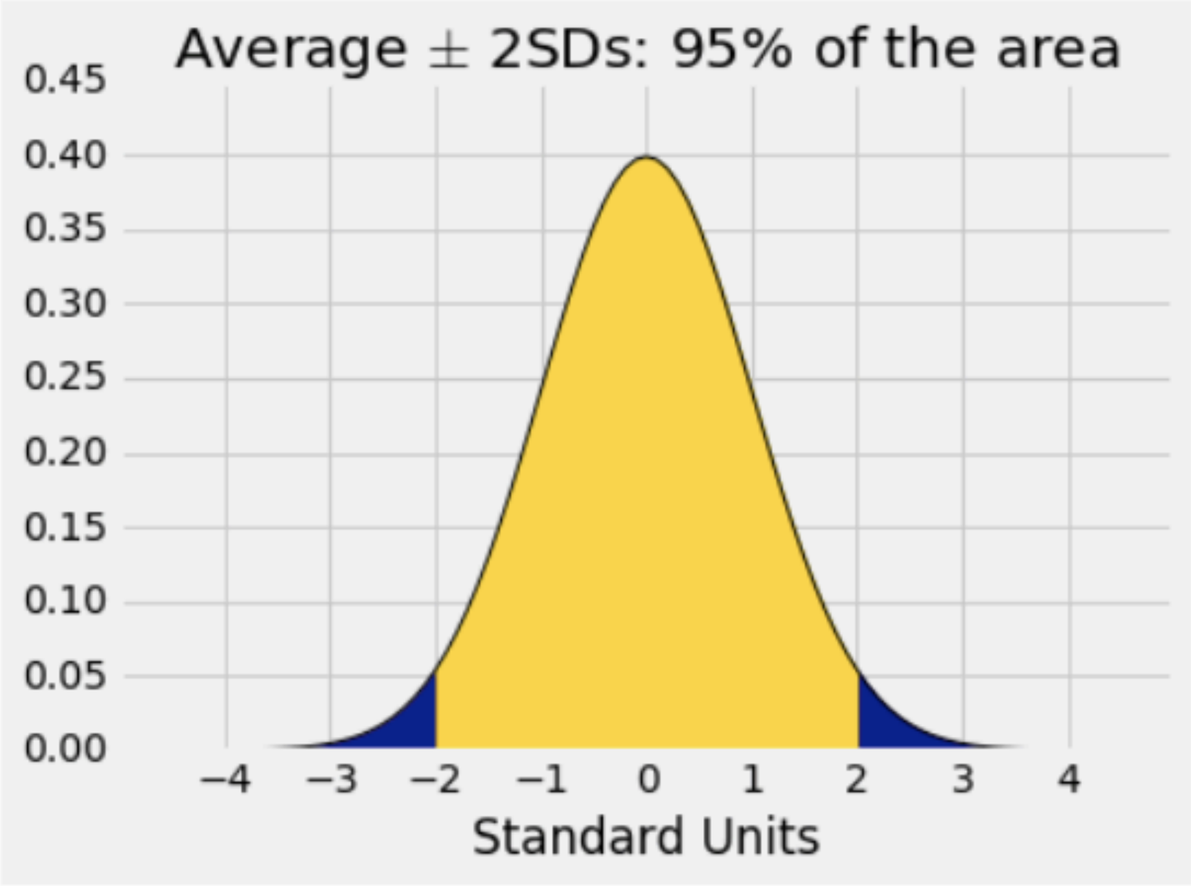
the probability distribution of the sample mean is roughly normal,

centered at the population mean, with

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

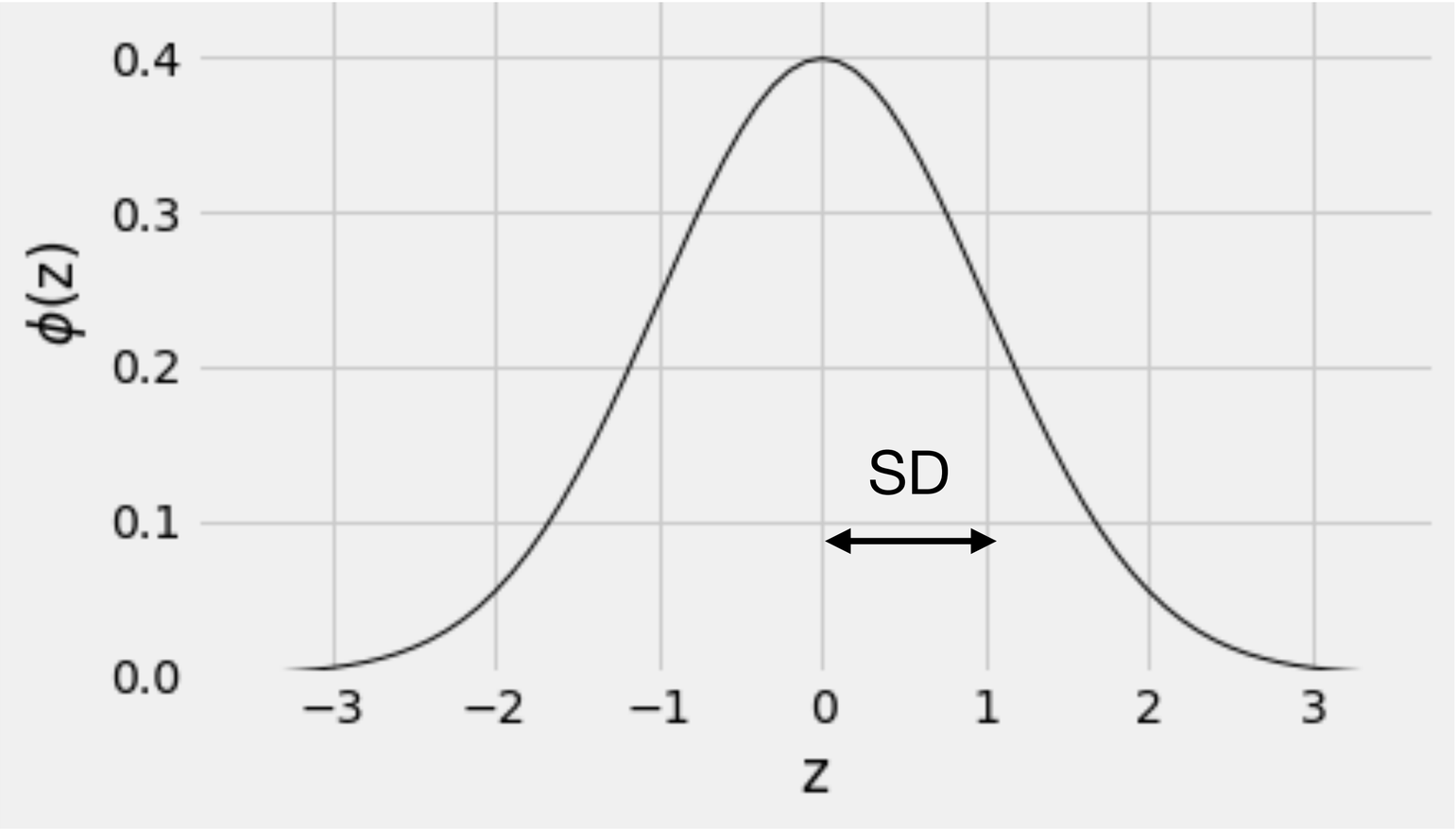
95% Confidence in Normal Distributions

- For a normal distribution, 95% of the data is in the range average + 2SDs



Connecting to confidence intervals

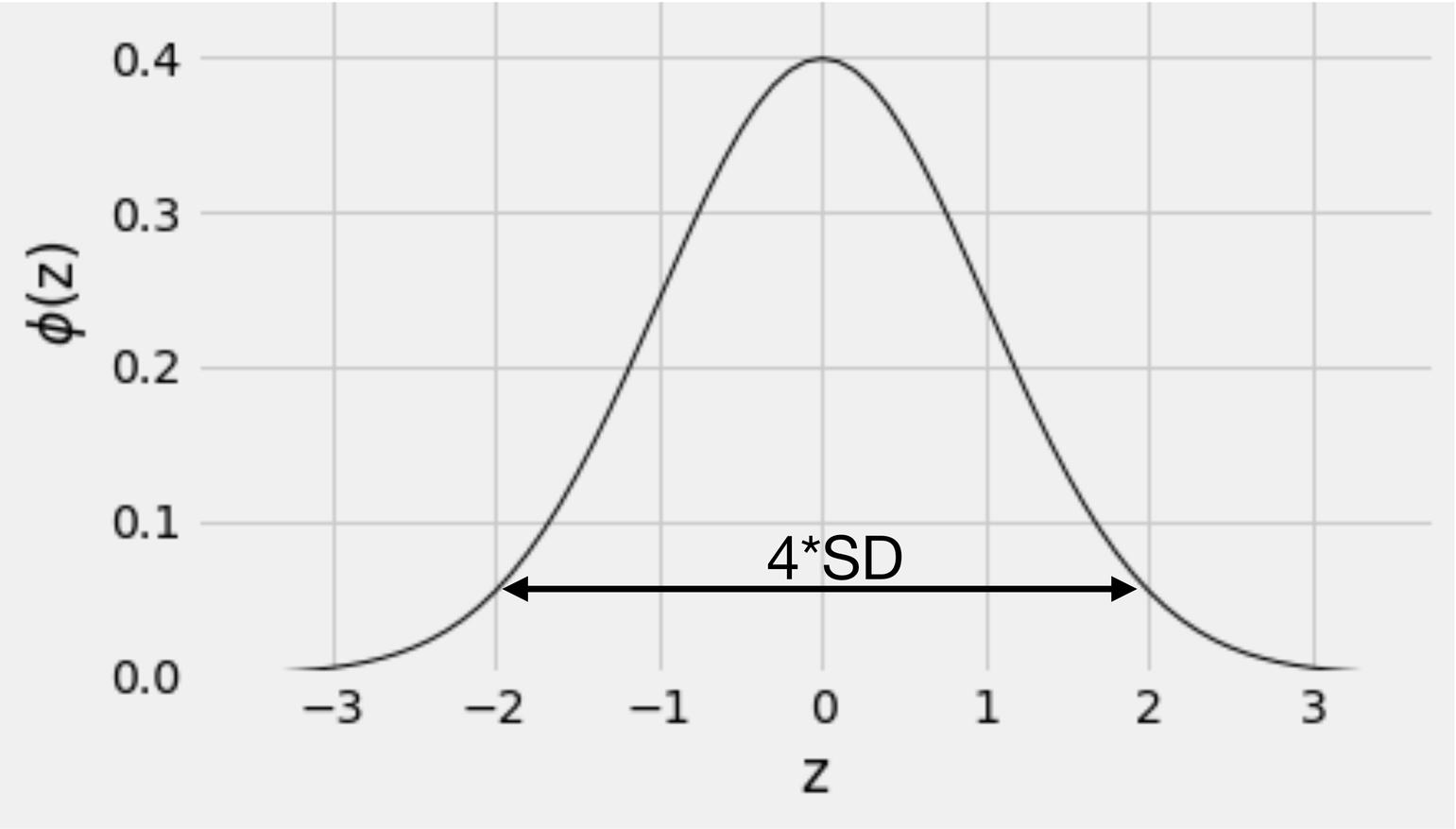
Distribution of Sample Averages (for Sample Size n)



$$\text{SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Connecting to confidence intervals

95% of the data is in the range average + 2SDs



width of the 95% confidence interval =

$$4 \times \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Example: Polling Sample Size

- Two candidates are up for election: Candidate A and Candidate B
- Candidate A wants to estimate with a 95% confidence interval what % of voters will vote for her
- How large of a sample should the candidate poll if they want to make this estimate with a desired accuracy of no wider than 1%?
  - Example: 95% confidence interval of (44%, 44.5%) is ok, but (44%, 46%) would be too inaccurate

How large of a sample do we need?

- Recall for normal distributions: Width of $95\,\%$ confidence interval $= 4 *$ SD
- Central Limit Theorem:

$$\text{SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- Putting it together:

$$0.01 \geq 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

How large of a sample do we need?

- Recall for normal distributions: Width of $95\,\%$ confidence interval $= 4 *$ SD
- Central Limit Theorem:

$$\text{SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- Putting it together:

$$0.01 \geq 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

How large of a sample do we need?

- Recall for normal distributions: Width of $95\,\%$ confidence interval $= 4 *$ SD
- Central Limit Theorem:

$$\text{SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

- Putting it together:

$$0.01 \geq 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

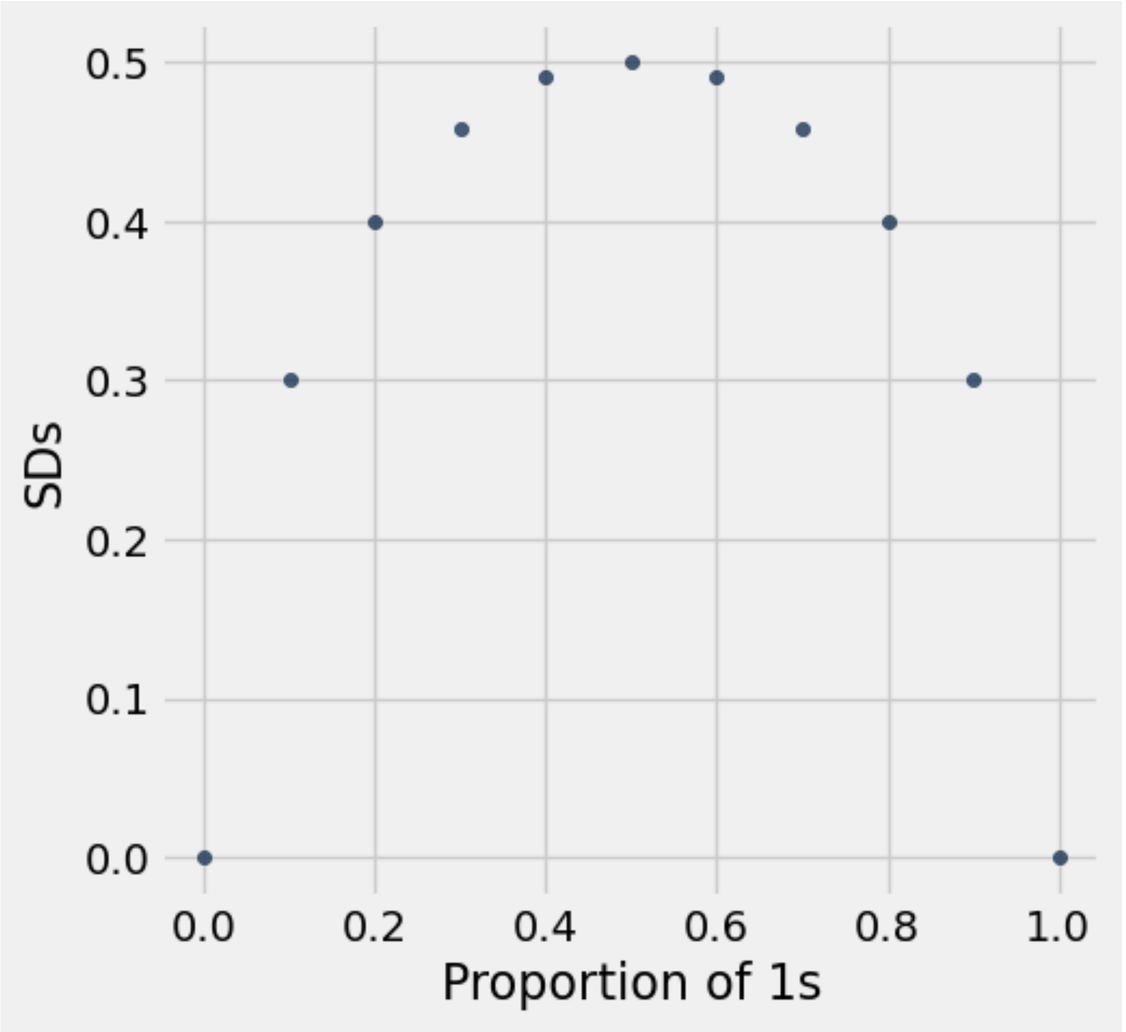
Notebook Demo: Population SD for Situations with Only 2 Outcomes

Textbook chapter on Choosing a Sample Size

Population SD for Situations with 2 Outcomes

- For situations with only 2 outcomes, the SD ranges from 0 to 0.5, with a max value of 0.5

- Thus, to estimate worst case scenario (most conservative sample size needed), you can use the maximum SD=0.5



Back to Polling Example

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

Back to Polling Example

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

Back to Polling Example

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * 50$$

Back to Polling Example

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * 50$$

$$\text{sample size} \geq 200^2$$

Back to Polling Example

$$\sqrt{\text{sample size}} \geq 4 * \frac{\text{Population SD}}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{0.01}$$

$$\sqrt{\text{sample size}} \geq 4 * 50$$

$$\text{sample size} \geq 200^2$$

$$\text{sample size} \geq 40{,}000$$

Next time

- Prediction
  - Correlation
  - Linear regression