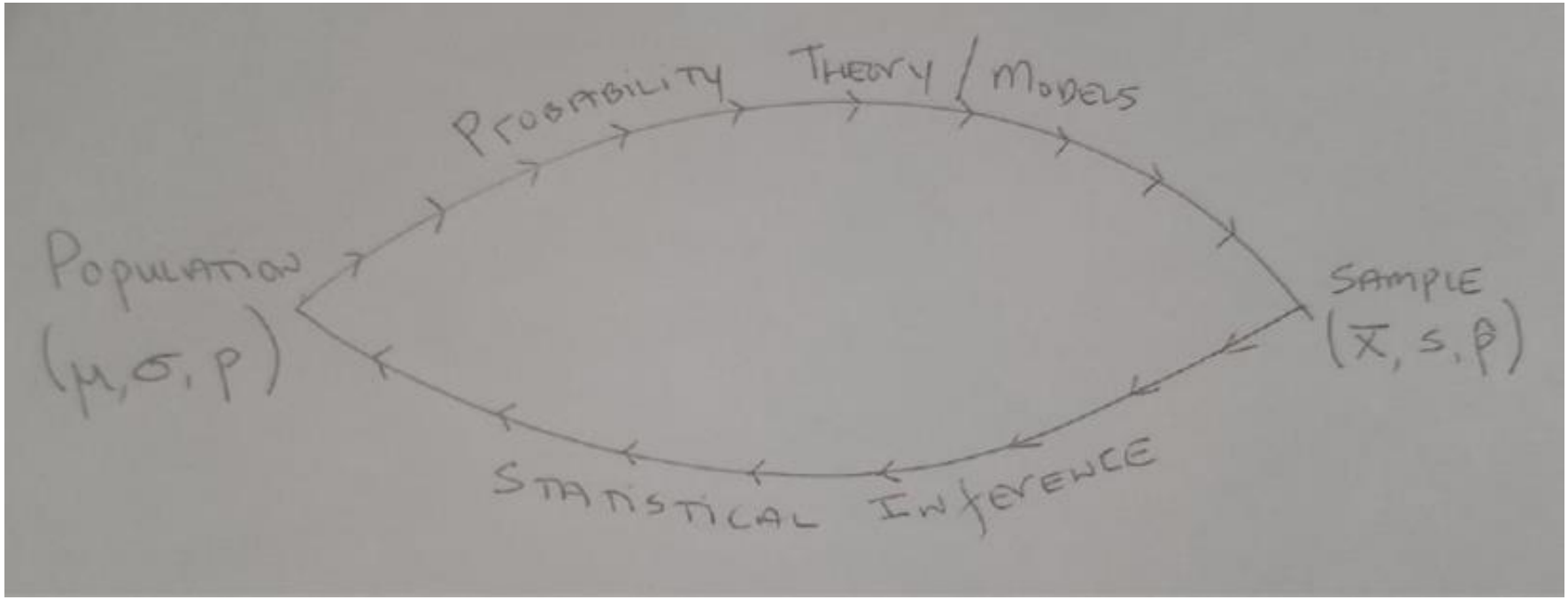
Probability and Mathematical Statistics in Data Science

Lecture 28: Section 8.1 – 9.4: Summary of Statistical Inference

Probability and Statistical Inference



Variation in Sample Statistics - Sample Proportions

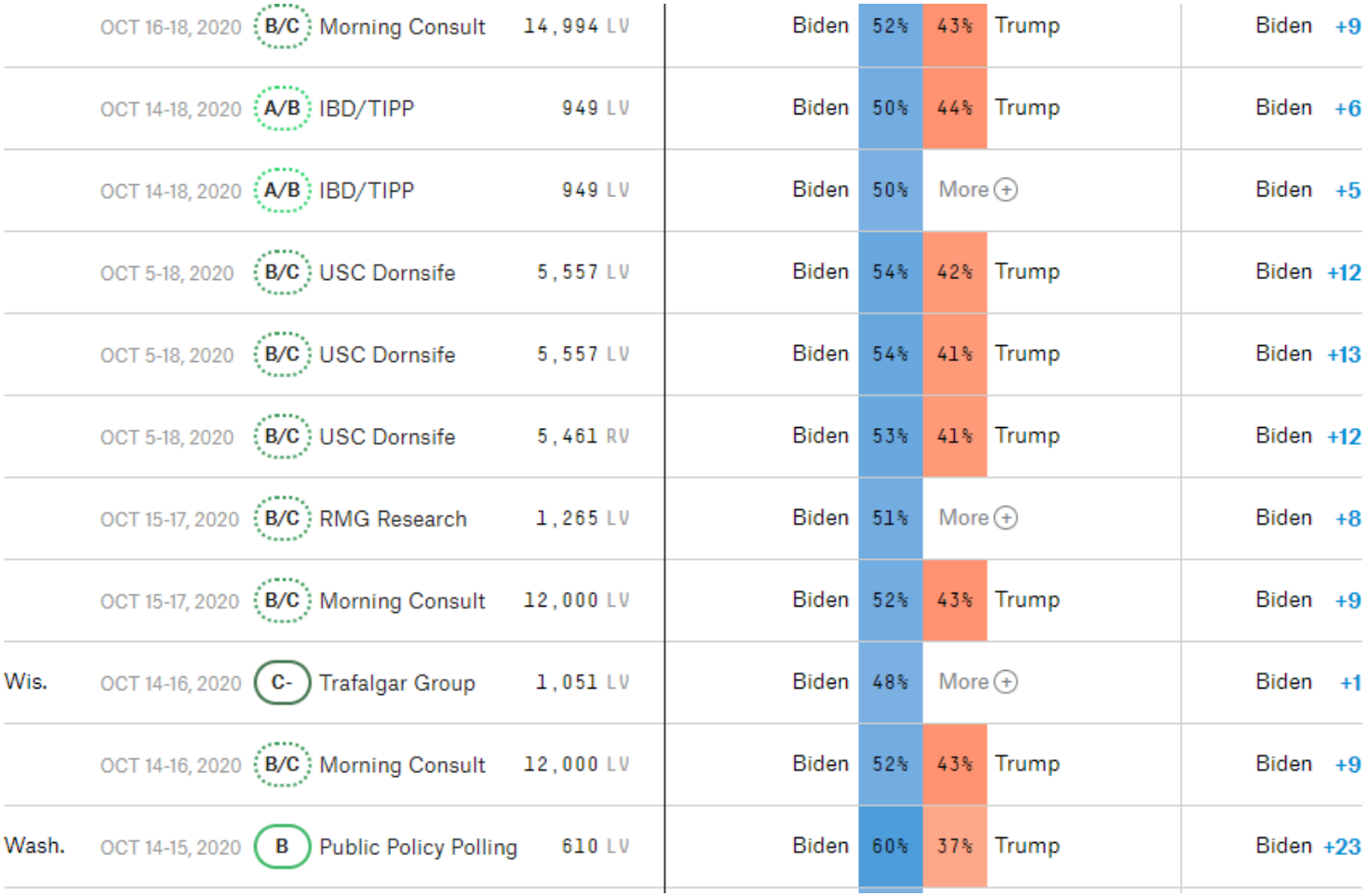
- The **sample proportion (percentage)** is an estimate of the **population proportion (percentage)**, the parameter of interest when looking at a single sample of categorical data.

- For example, we might be interested in estimating the proportion of the population who will vote for a particular candidate in a national election.

- The sample proportion is simply the number of people in sample who say they will vote for candidate over the sample size

- The standard error measures how far (on average) we expect a sample proportion to deviate from the population proportion.

Election Polling – FiveThirtyEight



	OCT 16-18, 2020	B/C	Morning Consult	14,994 LV	Biden	52%	43%	Trump	Biden +9
	OCT 14-18, 2020	A/B	IBD/TIPP	949 LV	Biden	50%	44%	Trump	Biden +6
	OCT 14-18, 2020	A/B	IBD/TIPP	949 LV	Biden	50%	More ⊕		Biden +5
	OCT 5-18, 2020	B/C	USC Dornsife	5,557 LV	Biden	54%	42%	Trump	Biden +12
	OCT 5-18, 2020	B/C	USC Dornsife	5,557 LV	Biden	54%	41%	Trump	Biden +13
	OCT 5-18, 2020	B/C	USC Dornsife	5,461 RV	Biden	53%	41%	Trump	Biden +12
	OCT 15-17, 2020	B/C	RMG Research	1,265 LV	Biden	51%	More ⊕		Biden +8
	OCT 15-17, 2020	B/C	Morning Consult	12,000 LV	Biden	52%	43%	Trump	Biden +9
Wis.	OCT 14-16, 2020	C-	Trafalgar Group	1,051 LV	Biden	48%	More ⊕		Biden +1
	OCT 14-16, 2020	B/C	Morning Consult	12,000 LV	Biden	52%	43%	Trump	Biden +9
Wash.	OCT 14-15, 2020	B	Public Policy Polling	610 LV	Biden	60%	37%	Trump	Biden +23

The results of polls for the same questions will differ due to **sampling variation**

https://projects.fivethirtyeight.com/polls/

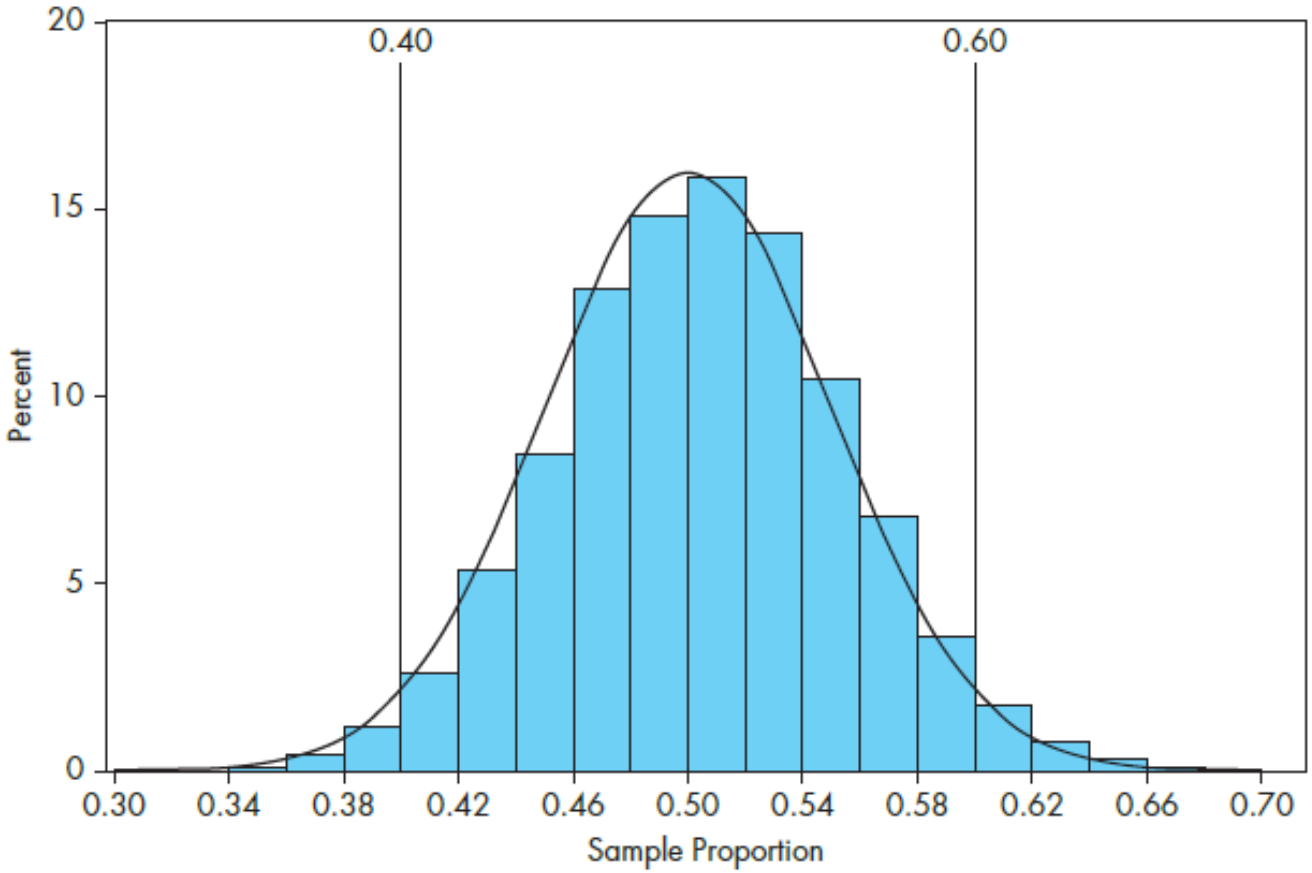
Variation in Sample Statistics - Sample Proportions

- The coin-toss experiment can be thought of as a question regarding a population proportion.
- If we toss a fair coin a very large number of times, we expect the proportion of heads to be 0.50.
- We can think of this value as the population proportion.
- Our sample proportion is the proportion of heads we obtain from a particular number of coin tosses

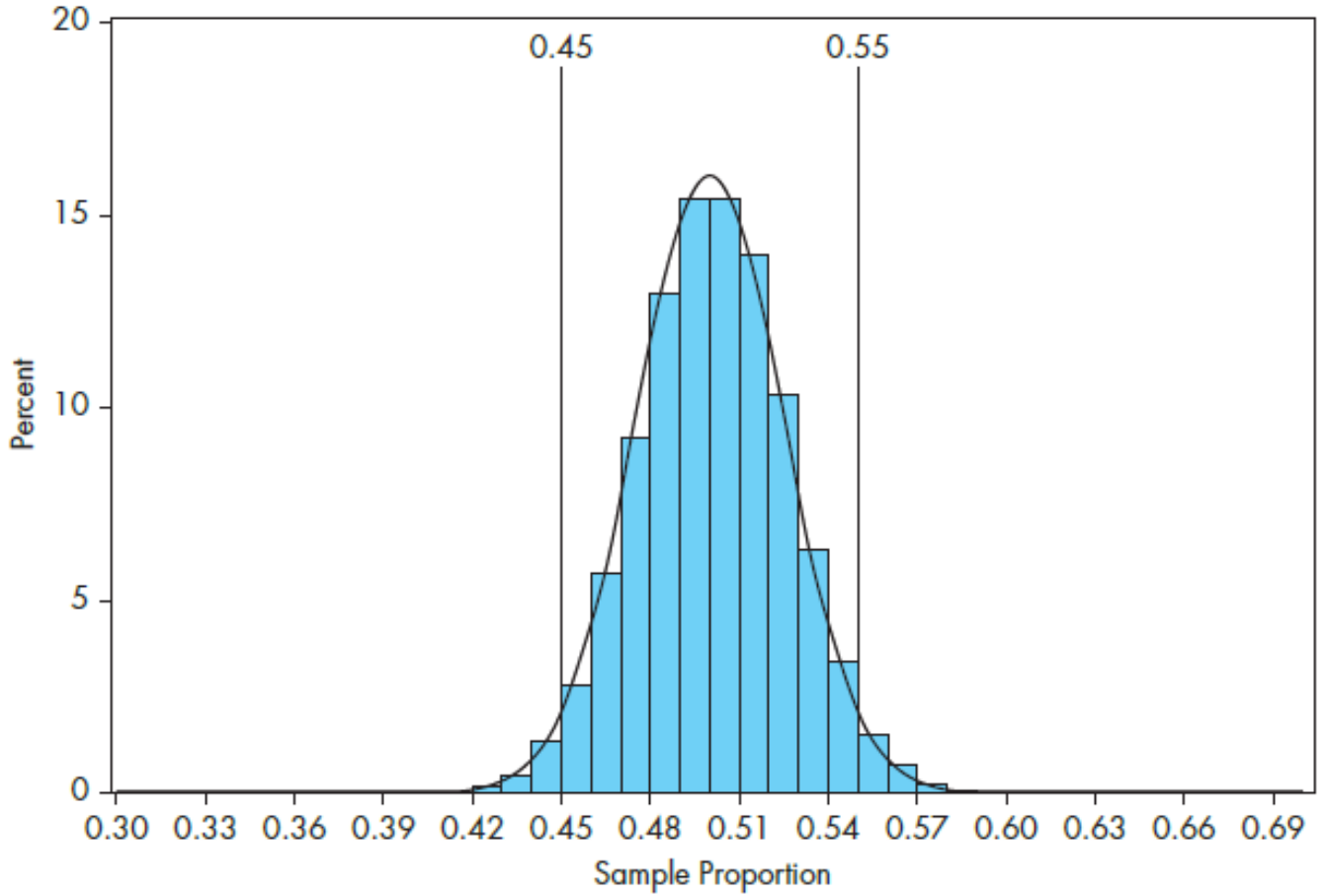
Variation in Sample Statistics - Sample Proportions

- If you toss a coin a 100 times, you expect to get 50 heads
- However, in reality the number of heads you get will vary around 50
- Example:
  - 1st 100 Tosses: 44 heads
  - 2nd 100 Tosses: 52 heads
  - 3rd 100 Tosses: 58 heads
- and so on… again the distribution of all possible outcomes will take on a familiar shape

Simulation of 10,000 Samples Proportions of Heads (Based on 100 Coin Tosses)



Simulation of 10,000 Samples Proportions of Heads (Based on 400 Coin Tosses)



Measuring Variation in Sample Proportions

- The standard error measures how far (on average) we expect a sample proportion to deviate from the population proportion (p).

standard error = square root of ((population proportion*(1 – population proportion))/sample size)

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

- When constructing **confidence intervals**, we will use our sample proportion (from our data) in this calculation.
- When conducting an **hypothesis test**, we use the null value for population proportion in this calculation.
- We will explain why when we cover these topics.

Measuring Variation in Sample Proportions

When the number of coin tosses is 100, the standard error is calculated as follows:

$$\text{standard error} = \text{square root of } ((0.50*(1 - 0.50))/100)$$
$$= \text{square root of } (0.025)$$
$$= 0.05$$

When we increase the number of coin tosses to 400, the standard error is calculated as follows:

$$= \text{square root of } ((0.50*(1 - 0.50))/400)$$
$$= \text{square root of } (0.000625)$$
$$= 0.025$$

Variation in Sample Statistics

source: https://mcu.edu/

Sample Means from 12 samples with a sample size equal to 16 men

68.3, 68.7, 69.2, 69.4, 69.6, 69.9, 70.1, 70.3, 70.5, 70.9, 71.1, 71.4.

Distribution of Men's Heights

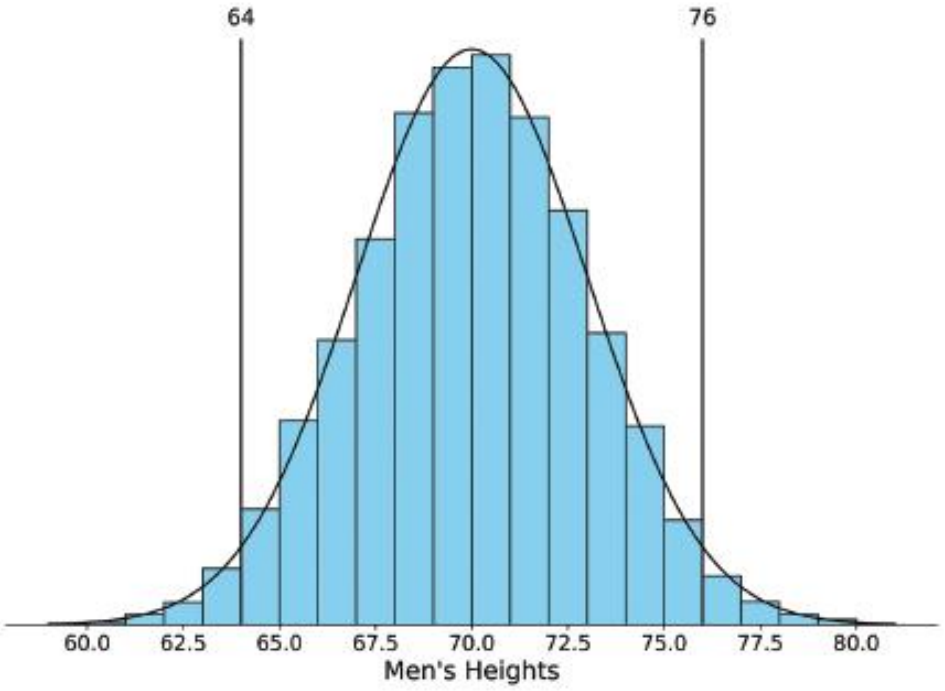
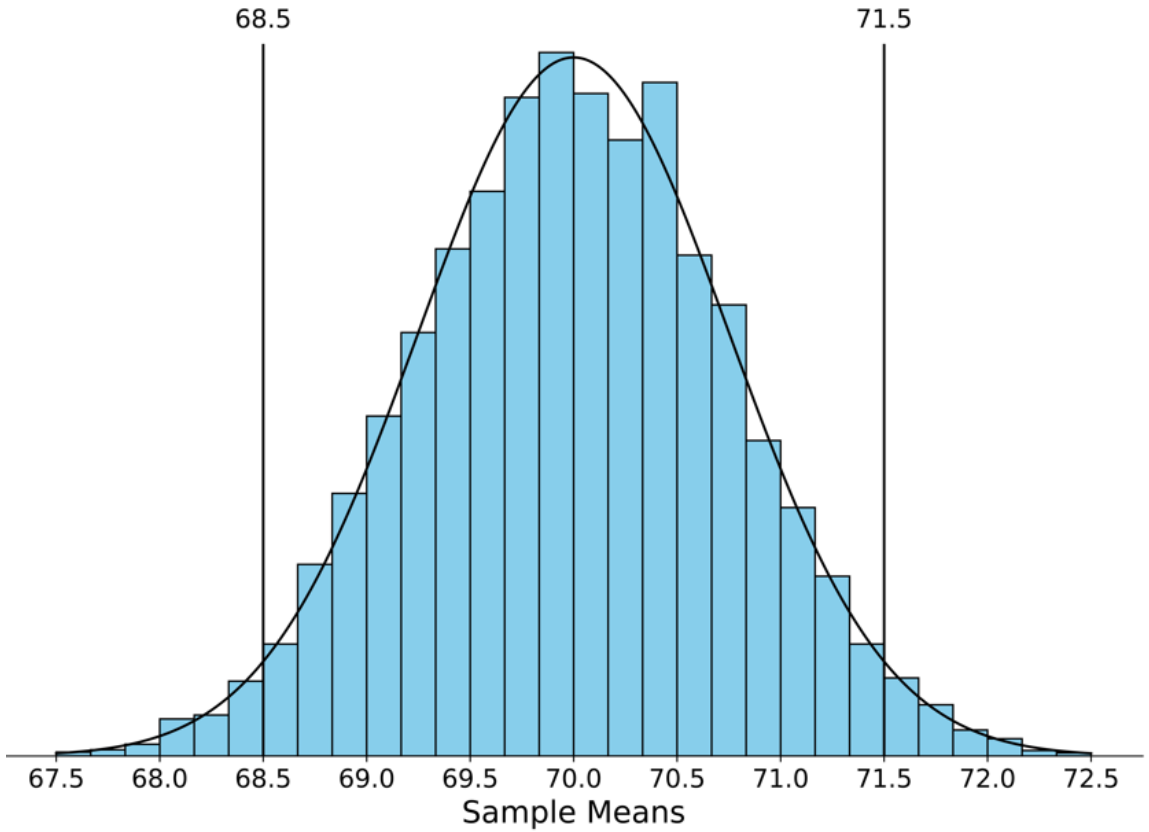


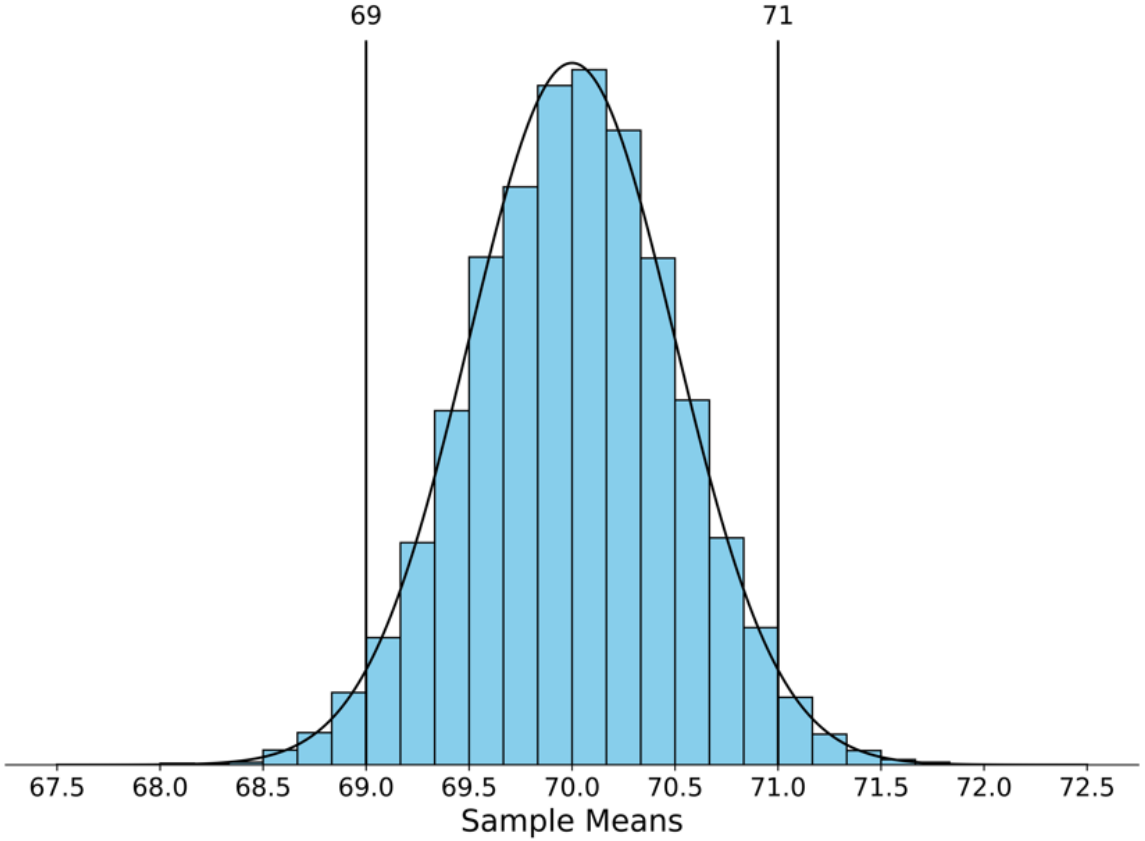
Figure 6.2: Population Distribution of Men's Heights in the United States

Distribution of Sample Mean Heights based on Sample Sizes of 16 Men



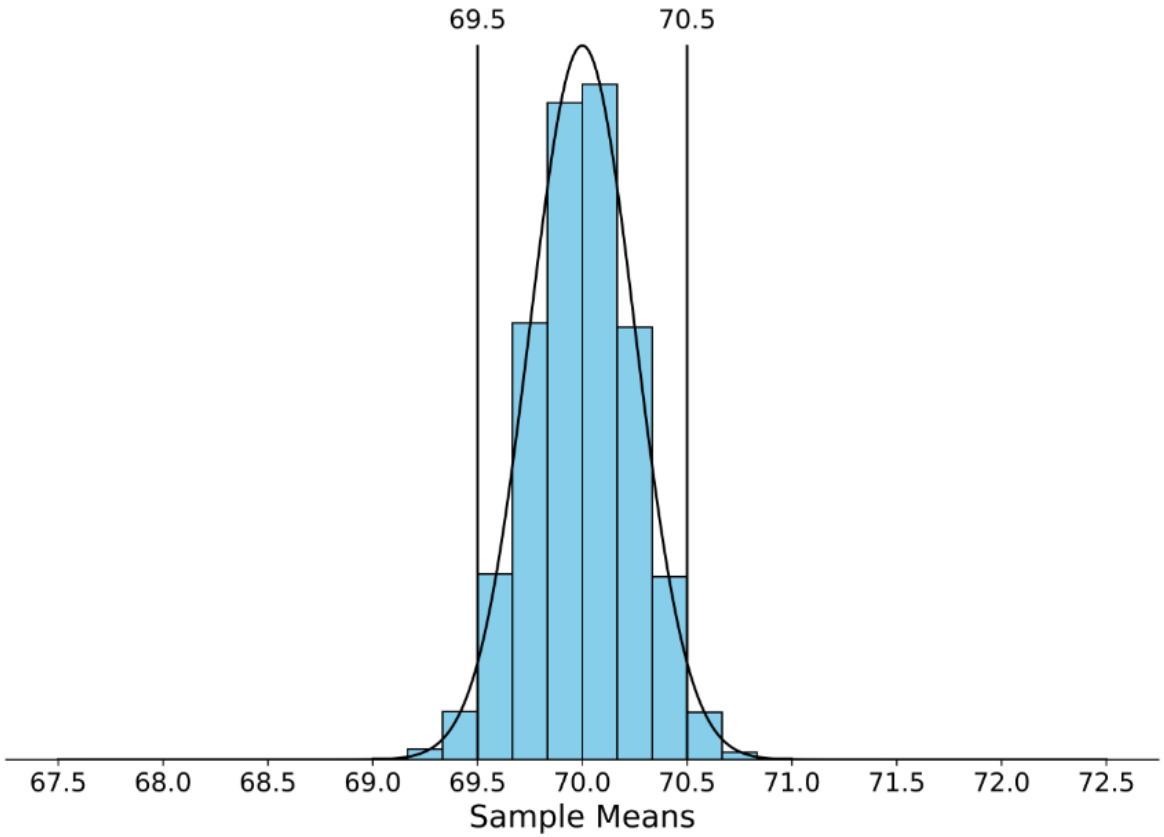
- For a sample of 16 men the standard error is 3/sqrt(16) = 0.75 inches
- 95% of possible sample mean heights between 68.5 inches and 71.5 inches

Distribution of Sample Mean Heights based on Sample Sizes of 36 Men



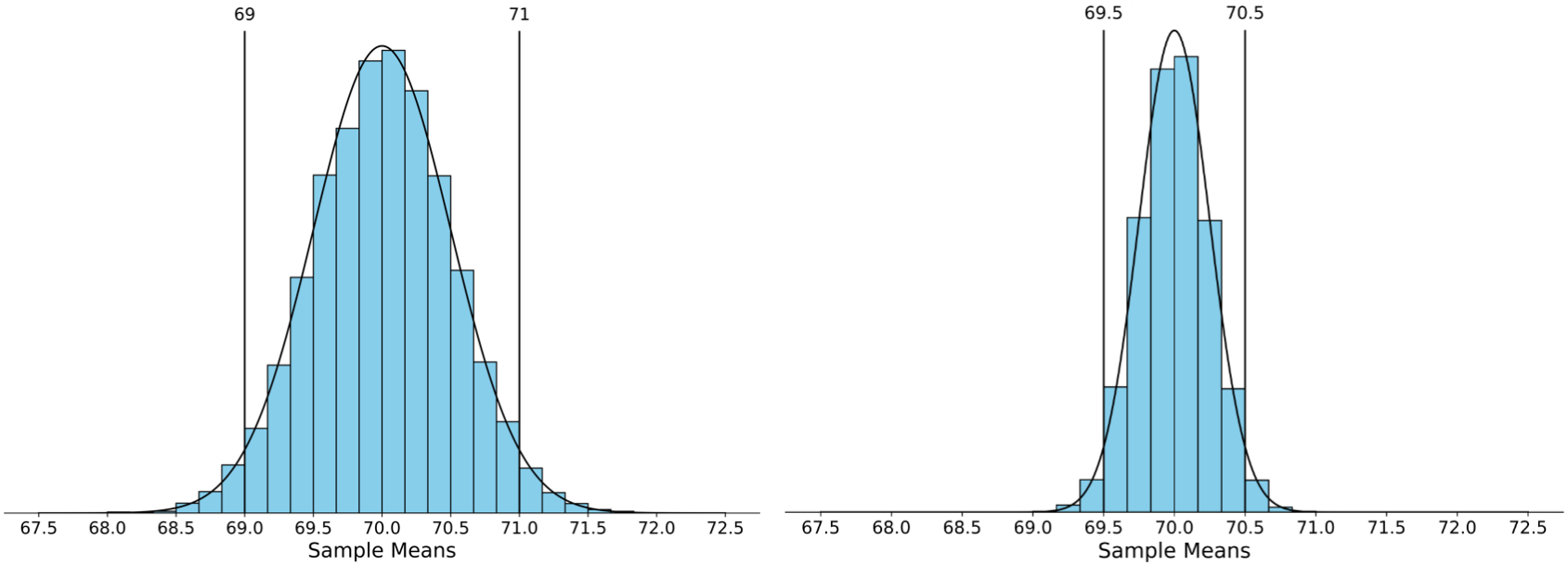
- For a sample of 36 men the standard error is 3/sqrt(36) = 0.5 inches
- 95% of possible sample mean heights between 69 inches and 71 inches

Distribution of Sample Mean Heights based on Sample Sizes of 144 Men



- For a sample of 144 men the standard error is 3/sqrt(144) = 0.25 inches
- 95% chance our sample mean will be between 69.5 inches and 70.5 inches

The Width of the Sampling Distribution



- The larger the sample size, the narrower the sampling distribution
- In reality, we don't know what the population mean height is
- Our sample mean height is one of a range of possible sample mean heights normally distributed around the unknown truth (or population mean)

Sampling Distribution of Men's Heights

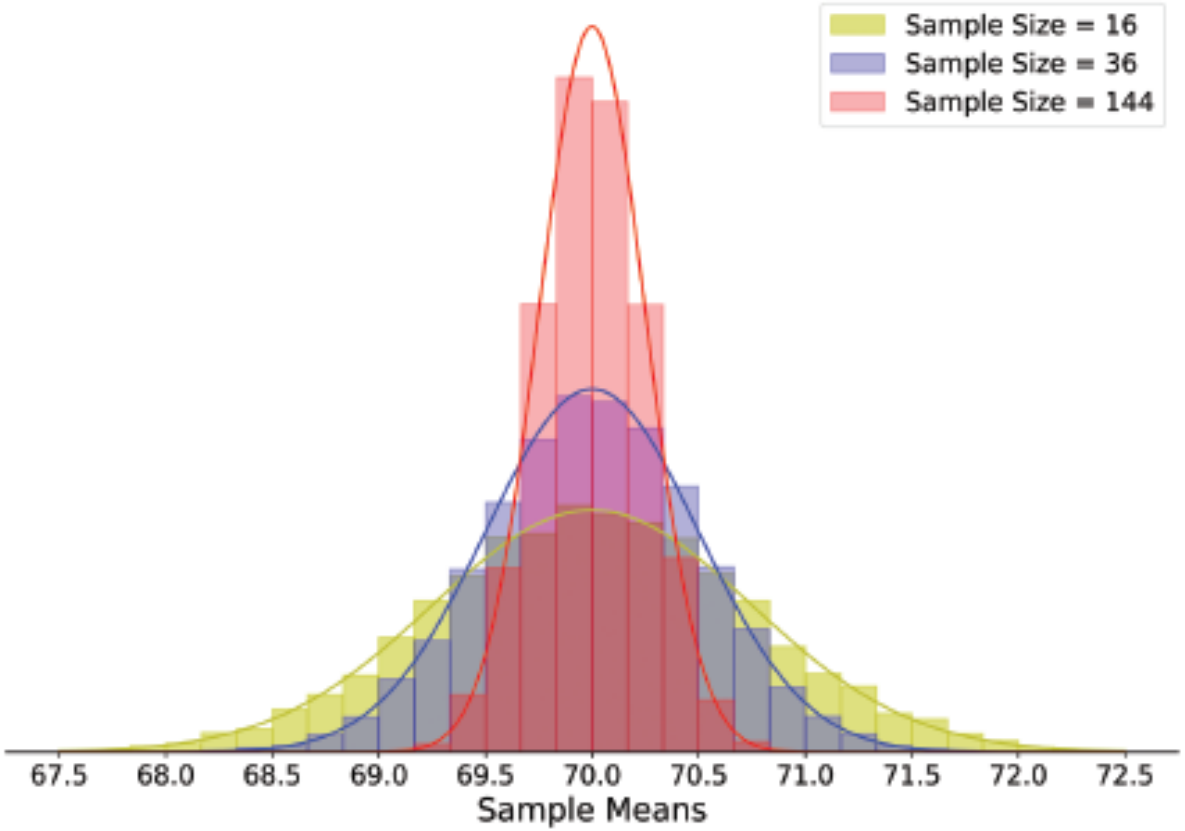
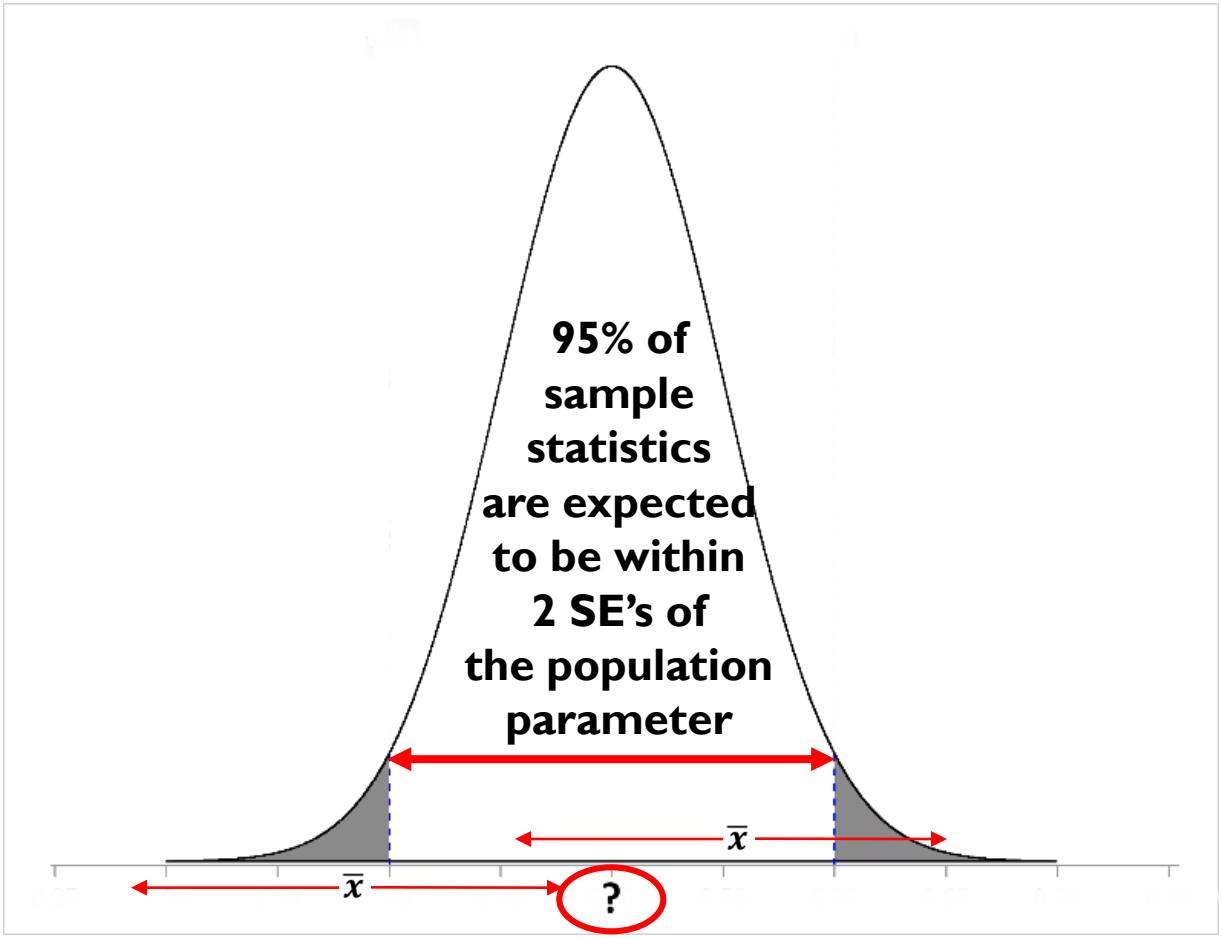


Figure 6.5: Simulation of 10,000 Sample Mean Heights—Sample Sizes Equal to 16, 36 and 144 Men

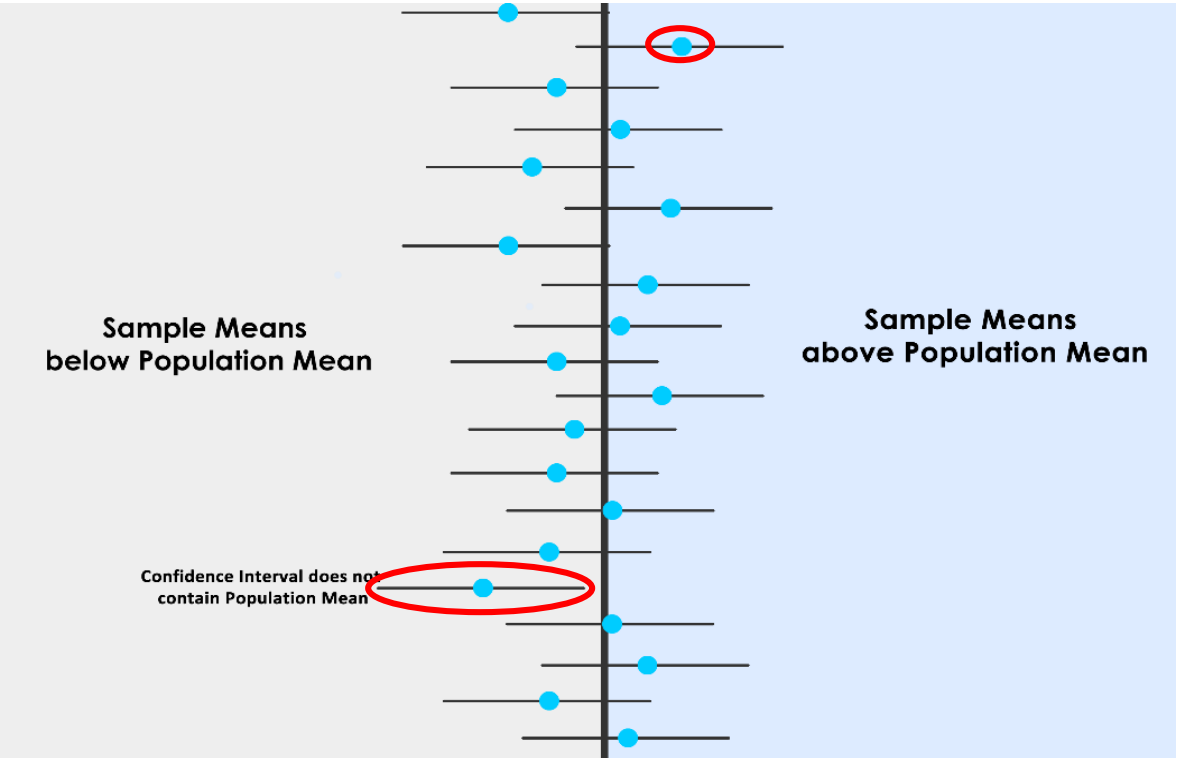
Confidence Intervals: Analogy

The Sampling Distribution



95% Confidence Interval: sample statistic ± 2 x (standard error)

Confidence Level: 95% Confidence Interval



- We expect19 out of every 20 "95% confidence intervals"
- ….to contain the population parameter

Understanding the Confidence Level

- For a confidence level of 95%, ***we expect that about 95% of all confidence intervals (based on a particular sample size) to actually include the population parameter of interest***. The remaining 5% will not.

- Our confidence is in the process over the long run.

  - 90% confidence level => multiplier = 1.645
  - 95% confidence level => multiplier = 1.96 (we round to 2)
  - 99% confidence level => multiplier = 2.576

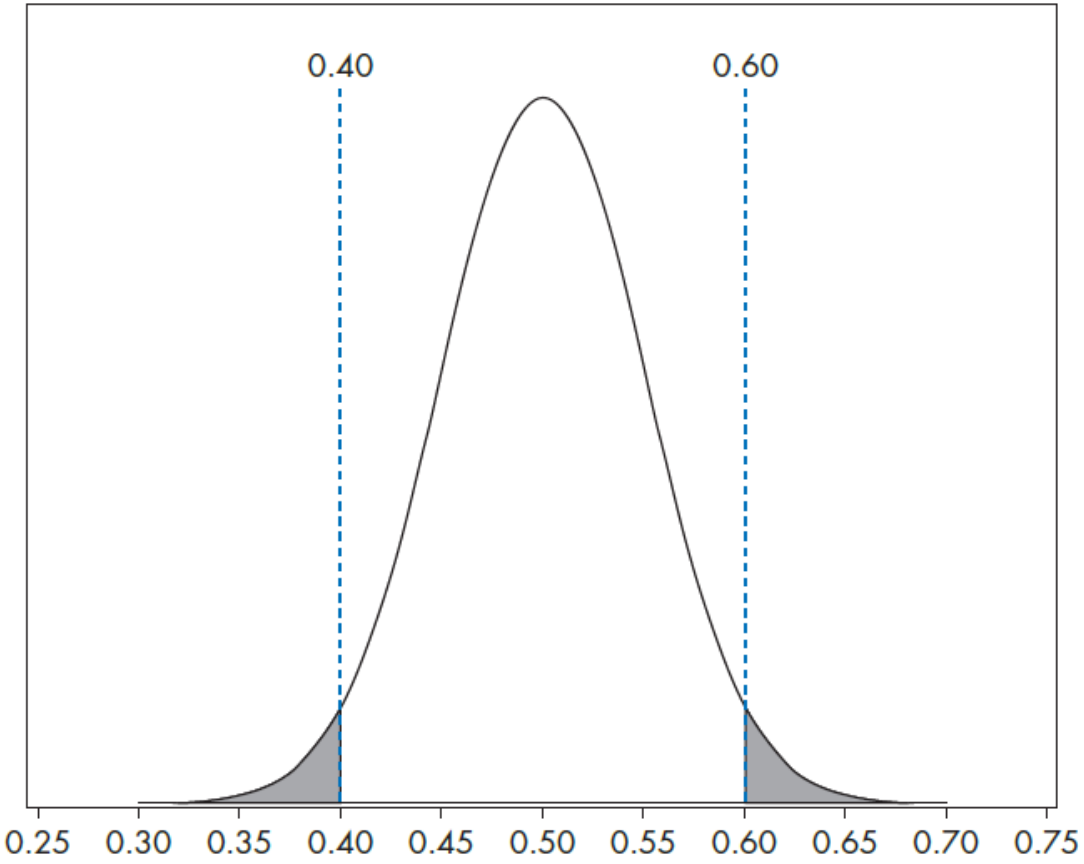
- More confidence implies an wider interval of plausible values for the population parameter of interest

Hypothesis Testing: Courtroom Jury Trial Analogy

- **Null Hypothesis**: *Defendant is innocent*
- **Alternative Hypothesis**: *Defendant is guilty*

source: www.defenselitigationinsider.com.com

Hypothesis Testing: Tossing Coin 100 Times



- **Null Hypothesis**: *Coin is fair*
- **Alternative Hypothesis**: *Coin is not fair*