Sampling & Confidence

How do we learn about populations from samples?

Announcements

📝 Pset 6 is due tonight at 10 pm

✅ Checkoffs begin tomorrow and are due next Tuesday at 9 pm

📝 Pset 7 is due Friday at 10 pm

No checkoff — the multiple choice questions on the pset page will be graded as your checkoff score

🏋️ All finger exercises are due Friday at 11:59 pm

No exceptions

📊 Course evaluations are open until Monday, Dec 15 at 9 am

https://eduapps.mit.edu/subjeval/studenthome.htm

🚫 Andrew won't have instructor office hours this Thursday

01 Distributions & Random Variables

02 Populations vs. Samples

03 Variance & Standard Deviation

04 The Central Limit Theorem

05 Confidence Intervals

It is 8:30 AM. The Exam starts at 9:00 AM.

The Bus

Average arrival: **8:50 AM**

The Train

Average arrival: **8:55 AM**

Which would you take? 🖐 Bus? Train? Not enough info?

Reveal σ

It is 8:30 AM. The Exam starts at 9:00 AM.

🚌

The Bus

Average arrival: **8:50 AM**

Std dev: **σ = 10 min**

🚆

The Train

Average arrival: **8:55 AM**

Std dev: **σ = 3 min**

Now who wants the bus? Who wants the train?

Back

Show Distributions

It is 8:30 AM. The Exam starts at 9:00 AM.

🚌

The Bus

Average arrival: **8:50 AM**

Std dev: **σ = 10 min**

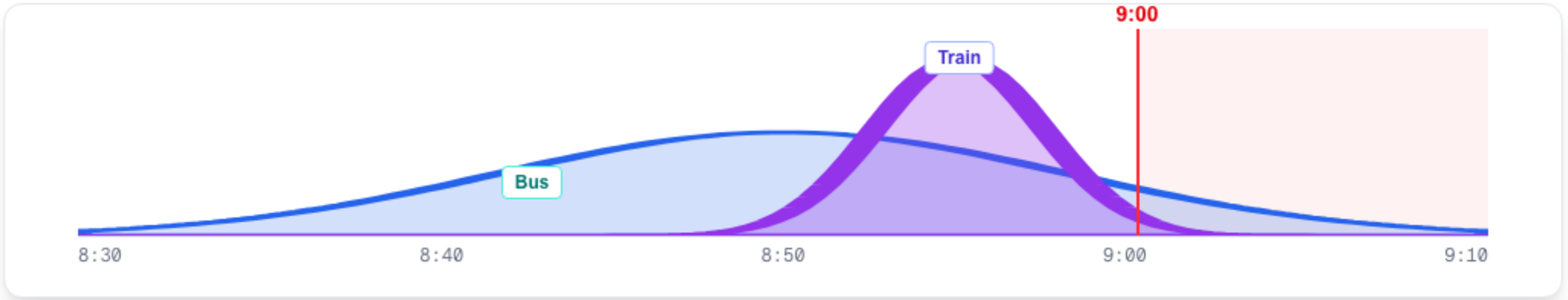
🚆

The Train

Average arrival: **8:55 AM**

Std dev: **σ = 3 min**

Let's see what the distributions look like...



Back

Show Probabilities

It is 8:30 AM. The Exam starts at 9:00 AM.

🚌

The Bus

Average arrival: **8:50 AM**

Std dev: **σ = 10 min**

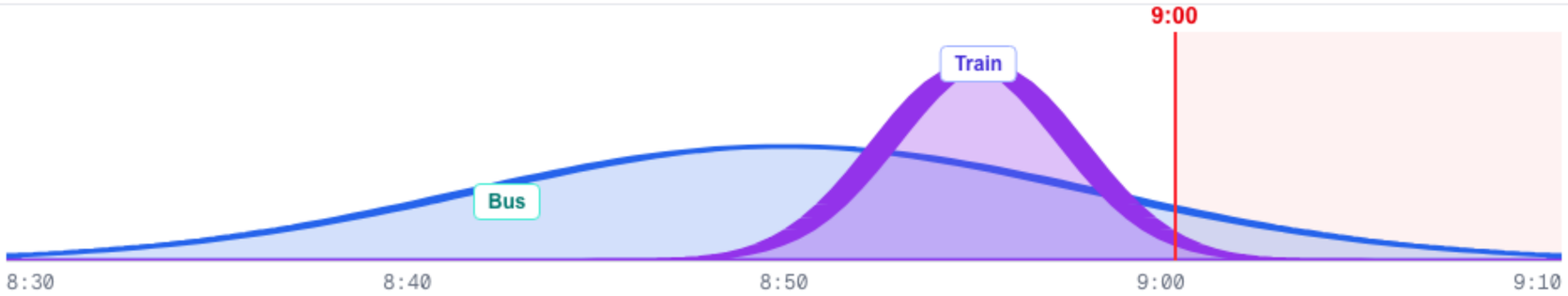
🚆

The Train

Average arrival: **8:55 AM**

Std dev: **σ = 3 min**

Let's see what the distributions look like...



P(Bus arrives after 9:00)

≈ 16%

P(Train arrives after 9:00)

< 1%

Back

Random Variable X

A quantity whose value is **uncertain** until we observe it.

X = "What time will the bus arrive?"

Realization x

The **actual value** we observe.

x = 8:47 AM

Before: X is uncertain. After: x is known.

BUS ARRIVAL TIME

?:??

X = uncertain

Simulate Bus Arrival

A **random variable** can be **Discrete** or **Continuous**

Discrete | Continuous

Discrete: X takes values you can *list*. We compute $P(X = k)$.

DISCRETE DISTRIBUTIONS

- **Discrete Uniform** — Rolling a fair die
- **Bernoulli** — Coin flip (Heads=1, Tails=0)
- **Binomial** — # of heads in 10 coin flips
- **Poisson** — # of emails per hour

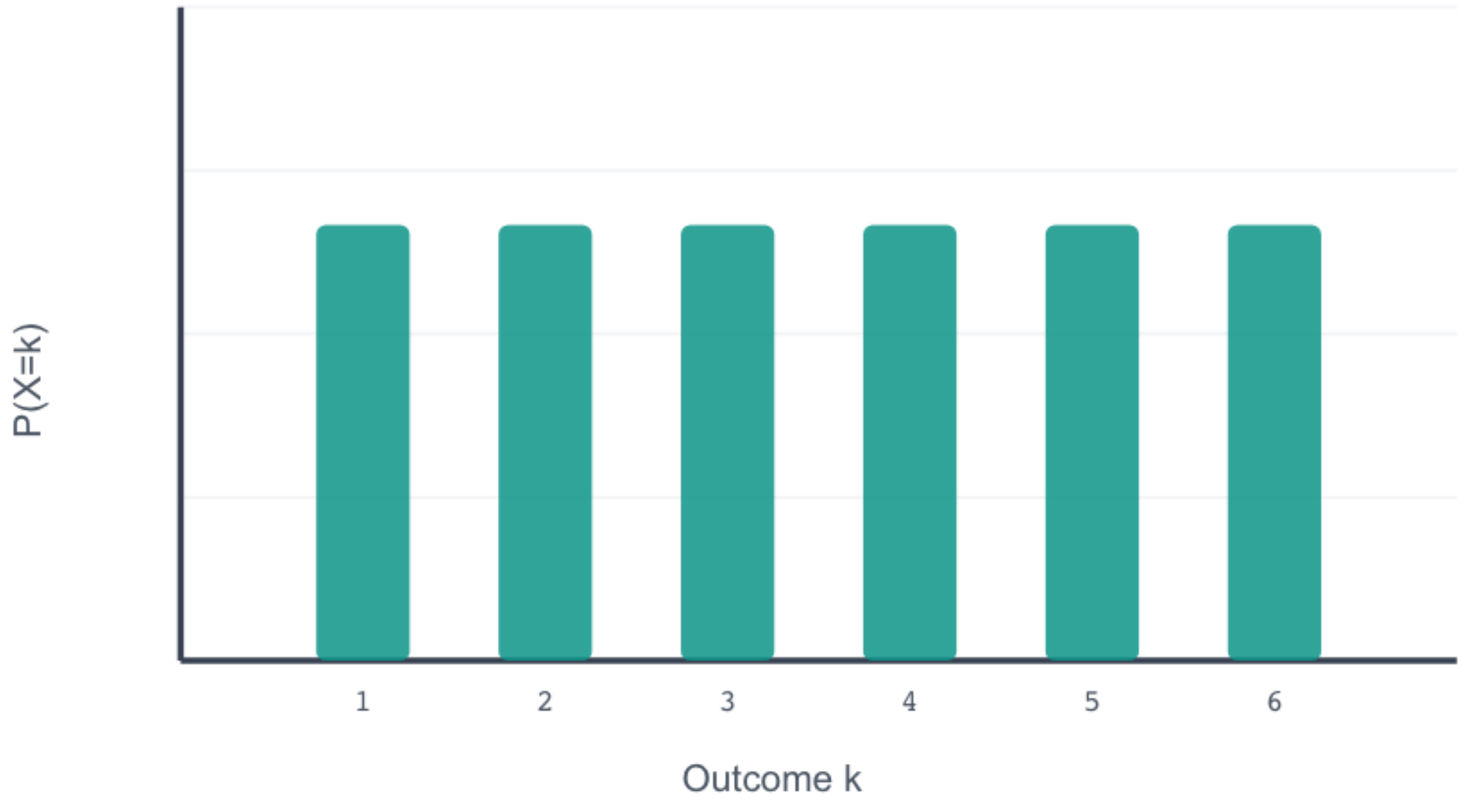
Discrete Uniform

DISCRETE | Real-World Examples

Each of the n outcomes has equal probability 1/n.

PMF

$$P(X = k) = \frac{1}{n}$$



Bar heights = exact probabilities. Sum of all bars = 1.

A **random variable** can be **Discrete** or **Continuous**

Discrete | Continuous

Discrete: X takes values you can *list*. We compute $P(X = k)$.

DISCRETE DISTRIBUTIONS

- **Discrete Uniform** — Rolling a fair die
- **Bernoulli** — Coin flip (Heads=1, Tails=0)
- **Binomial** — # of heads in 10 coin flips
- **Poisson** — # of emails per hour

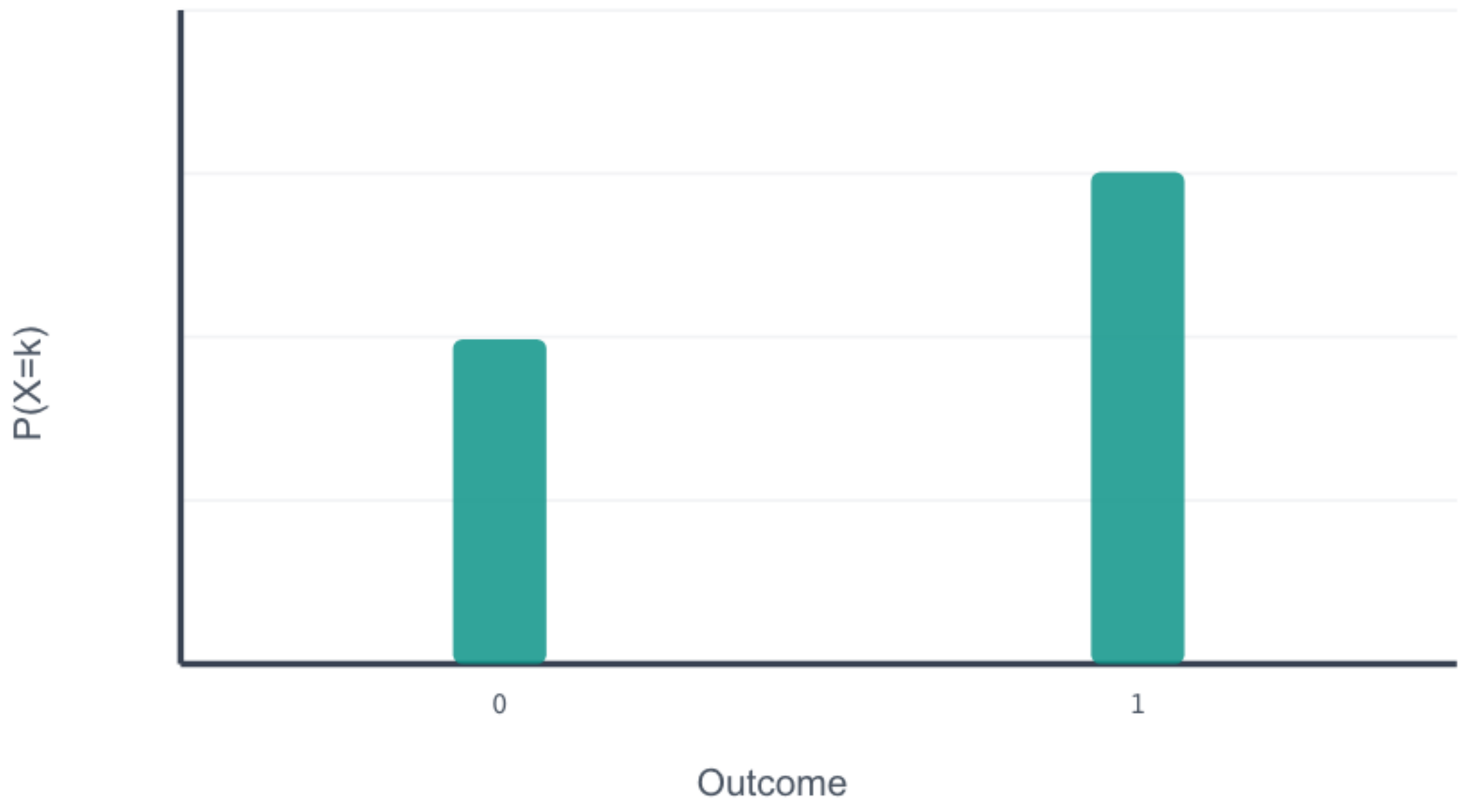
Bernoulli

DISCRETE | Real-World Examples

Binary outcome: success (1) with probability p, failure (0) with probability 1-p.

PMF

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$



Bar heights = exact probabilities. Sum of all bars = 1.

A **random variable** can be **Discrete** or **Continuous**

Discrete | Continuous

Discrete: X takes values you can *list*. We compute $P(X = k)$.

DISCRETE DISTRIBUTIONS

- **Discrete Uniform** — Rolling a fair die
- **Bernoulli** — Coin flip (Heads=1, Tails=0)
- **Binomial** — # of heads in 10 coin flips
- **Poisson** — # of emails per hour

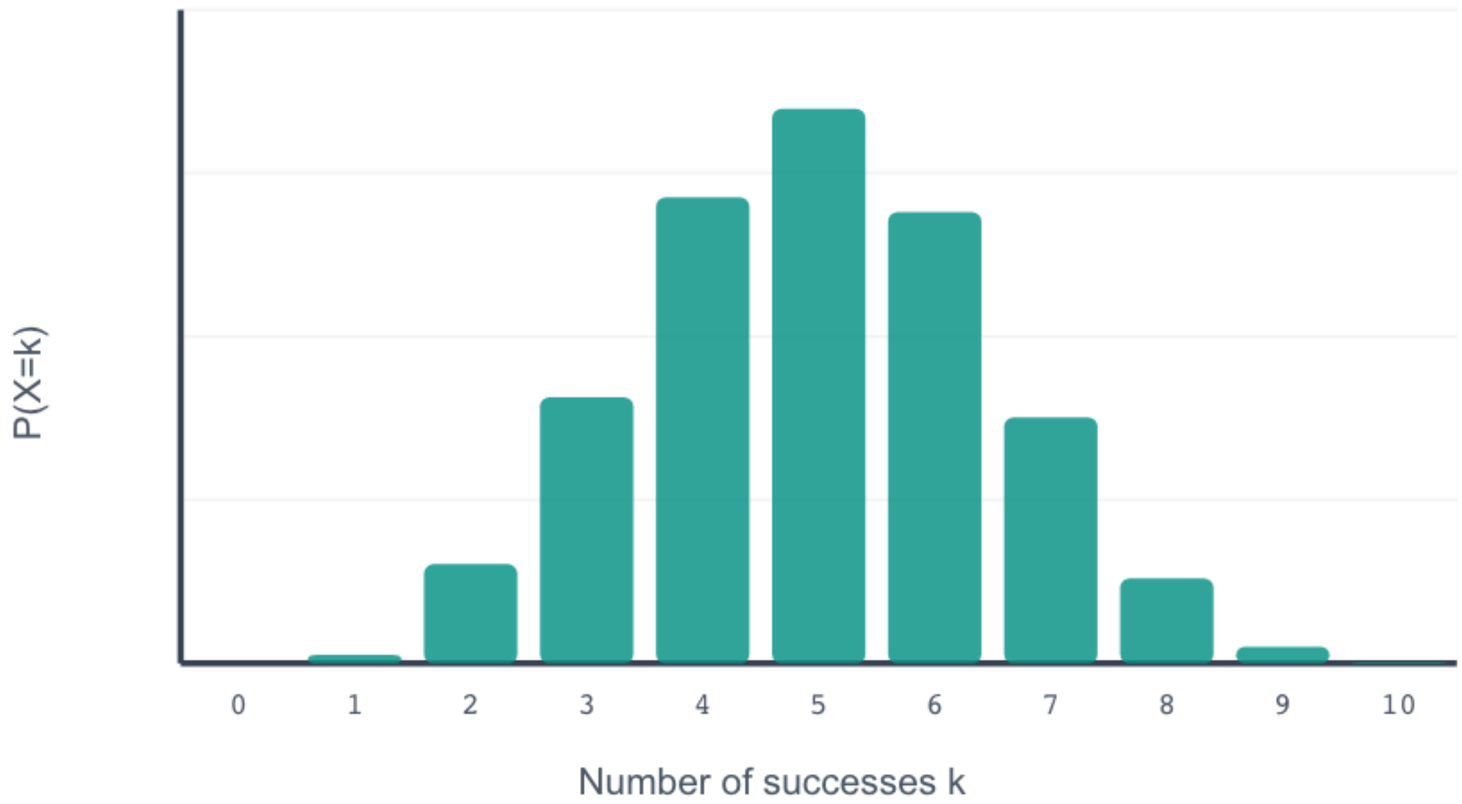
Binomial

DISCRETE | Real-World Examples

Number of successes in n independent Bernoulli trials.

PMF

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Bar heights = exact probabilities. Sum of all bars = 1.

A **random variable** can be **Discrete** or **Continuous**

Discrete | Continuous

Discrete: X takes values you can *list*. We compute $P(X = k)$.

DISCRETE DISTRIBUTIONS

Discrete Uniform
Rolling a fair die

Bernoulli
Coin flip (Heads=1, Tails=0)

Binomial
# of heads in 10 coin flips

Poisson
# of emails per hour

Poisson

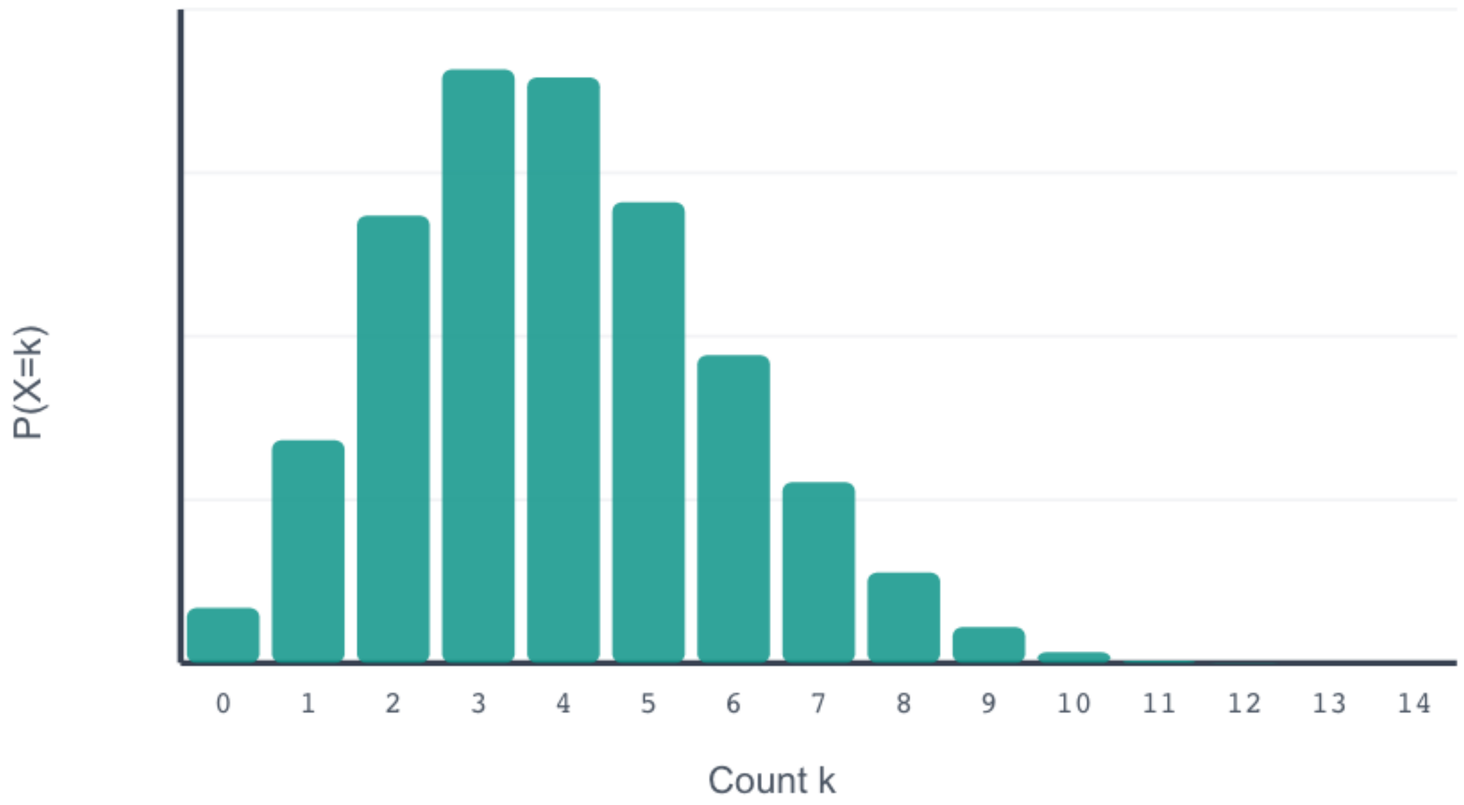
DISCRETE

Real-World Examples

Count of events in a fixed interval when events occur at constant average rate λ.

PMF

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Bar heights = exact probabilities. Sum of all bars = 1.

A **random variable** can be **Discrete** or **Continuous**

Discrete | **Continuous**

Continuous: X takes *any* value in a range. $P(X = x) = 0$ always!

CONTINUOUS DISTRIBUTIONS

📈 **Normal (Gaussian)**
Human heights, measurement errors

⏱️ **Exponential**
Wait time until next bus

➖ **Continuous Uniform**
Random point on a line segment

📈 Normal (Gaussian)

CONTINUOUS | Real-World Examples

The bell curve. Most values cluster around the mean μ, with spread controlled by σ.

PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

📈 **Area under curve** = probability. $P(a < X < b) = \int_a^b f(x)\,dx$. Total area = 1.

A **random variable** can be **Discrete** or **Continuous**

Discrete | Continuous

Continuous: X takes *any* value in a range. $P(X = x) = 0$ always!

CONTINUOUS DISTRIBUTIONS

Normal (Gaussian)
Human heights, measurement errors

Exponential
Wait time until next bus

Continuous Uniform
Random point on a line segment

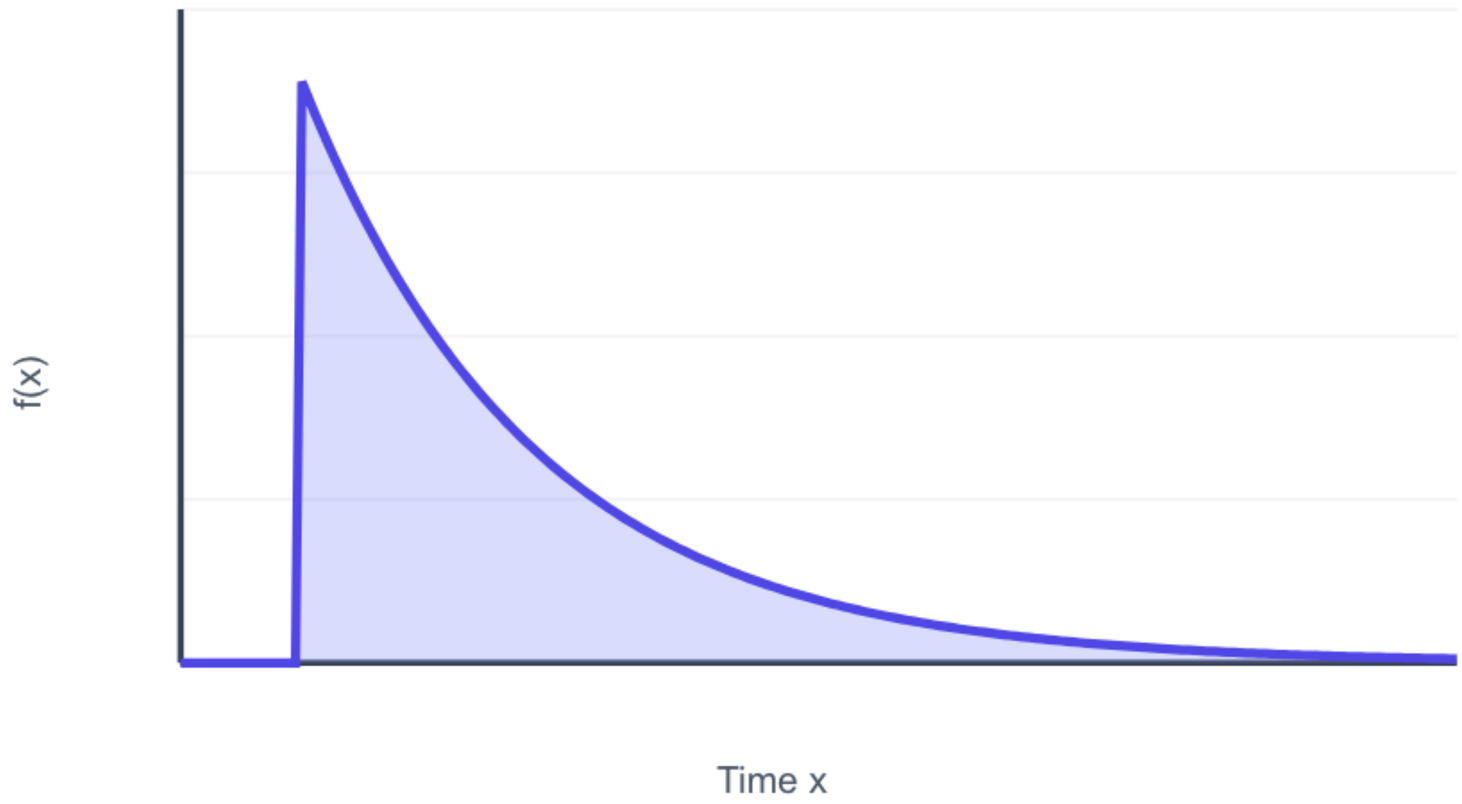
Exponential

CONTINUOUS | Real-World Examples

Time until next event in a Poisson process. Memoryless property.

PDF

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$



Area under curve = probability. $P(a < X < b) = \int_a^b f(x)\,dx$. Total area = 1.

A **random variable** can be **Discrete** or **Continuous**

Discrete | Continuous

Continuous: X takes *any* value in a range. $P(X = x) = 0$ always!

CONTINUOUS DISTRIBUTIONS

📈 **Normal (Gaussian)**
Human heights, measurement errors

⏱️ **Exponential**
Wait time until next bus

▬ **Continuous Uniform**
Random point on a line segment

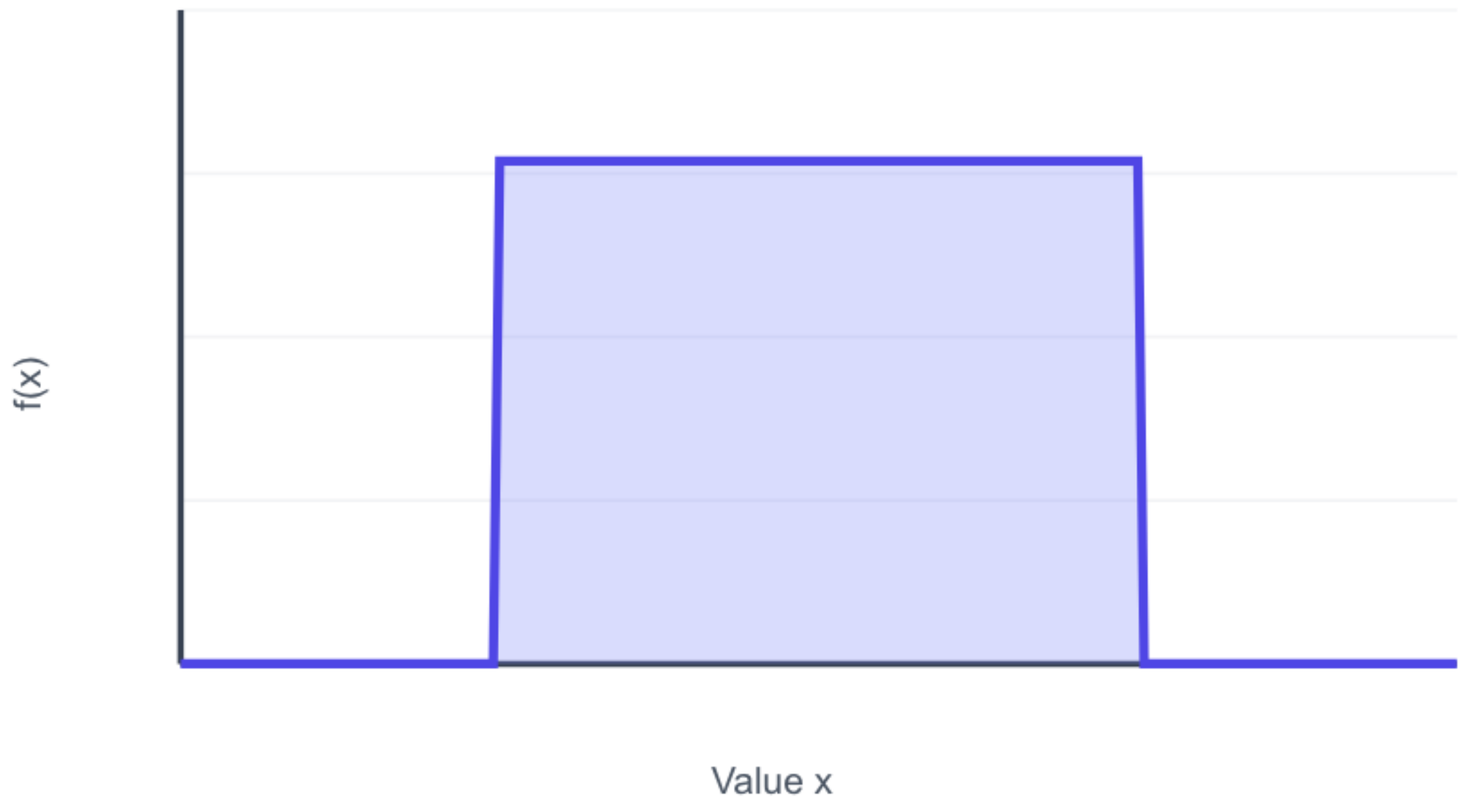
▬ Continuous Uniform

CONTINUOUS | Real-World Examples

Every value in the interval [a, b] is equally likely. Flat density.

PDF

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b$$



📈 **Area under curve** = probability. $P(a < X < b) = \int_a^b f(x)\,dx$. Total area = 1.

The Population

$N = 119,626$ stars in HYG catalog

True mean: $\mu = 15.7$

POPULATION DISTRIBUTION

SAMPLE DISTRIBUTION (N=30)

SAMPLE MEAN ($N = 30$)

—

Measure 30 Stars

SAMPLING ERROR

Take a sample to see the error.

HISTORY

No samples yet

Sampling: With vs Without Replacement

🔄 With Replacement

After selecting an item, **put it back** before the next draw.

EXAMPLES

- Rolling a die multiple times
- Slot machine spins
- Flipping a coin repeatedly

Key property:
Each draw is **independent**

❌ Without Replacement

After selecting an item, **remove it** from the pool.

EXAMPLES

- Drawing cards from a deck
- Lottery number drawings
- Picking team members

Key property:
Draws are **dependent**

In Practice

When the population is **much larger** than the sample ($n \ll N$), sampling without replacement behaves like sampling with replacement. Our star catalog has $N = 119{,}626$ — sampling 50 stars is only 0.04% of the population!

Variance (σ^2) measures spread. **Standard Deviation** (σ) is its square root.

PARAMETERS

Mean (μ) 0

Std Dev (σ) 30

Resample Data

STD DEV DEFINITION

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

NORMAL PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

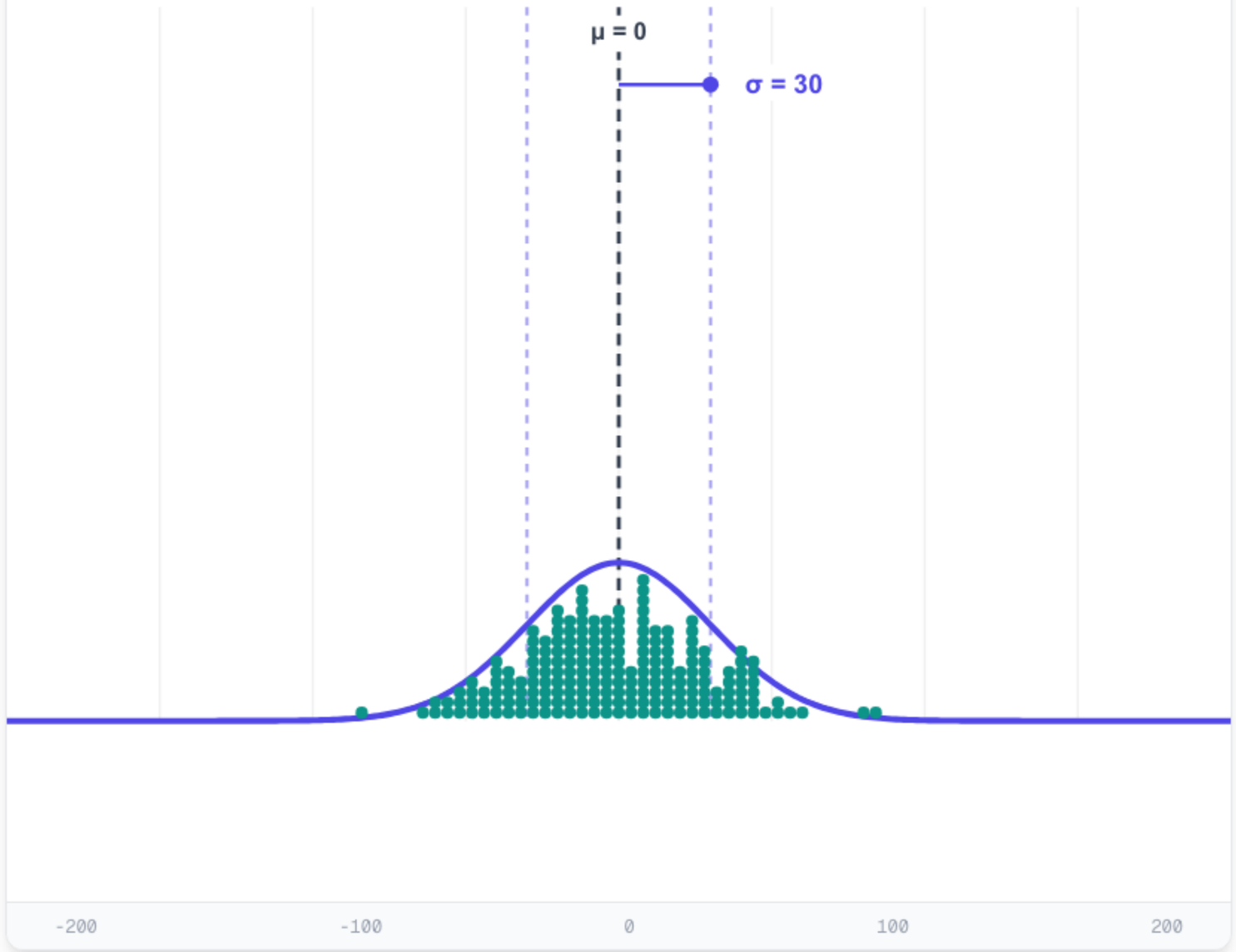
EMPIRICAL RULE

68% within ±1σ

95% within ±2σ

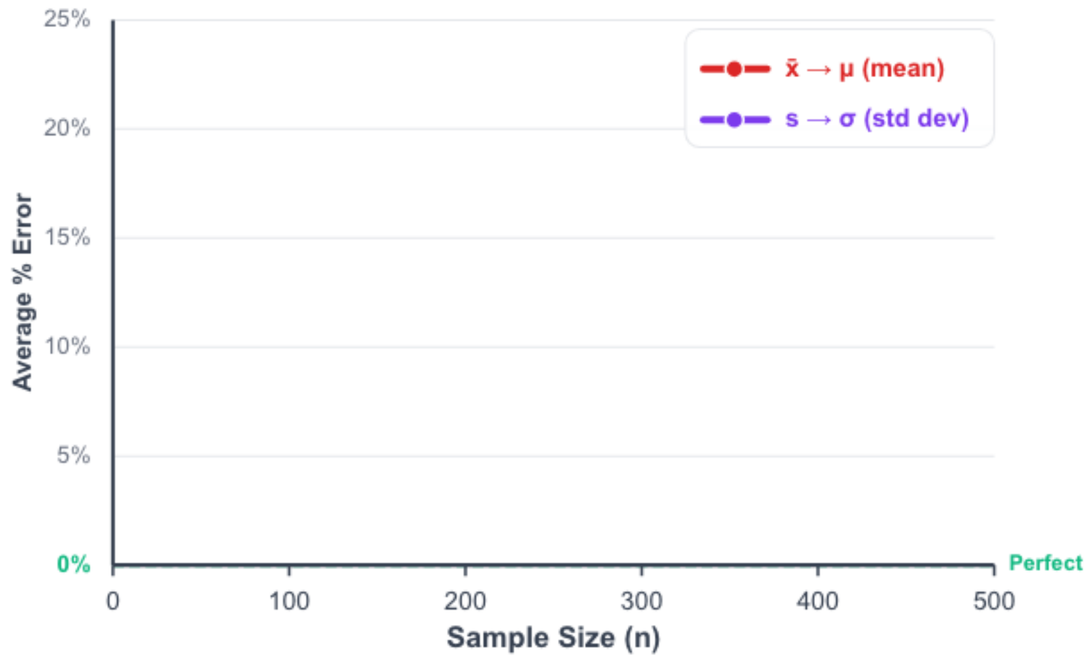
99.7% within ±3σ

Show theoretical curve



Estimators Converge to Truth

As sample size n increases, both **sample mean (x̄)** and **sample std (s)** converge to the true population values



▶ Run Simulation

Law of Large Numbers

Both x̄ and s are **consistent estimators**.

As $n \to \infty$:

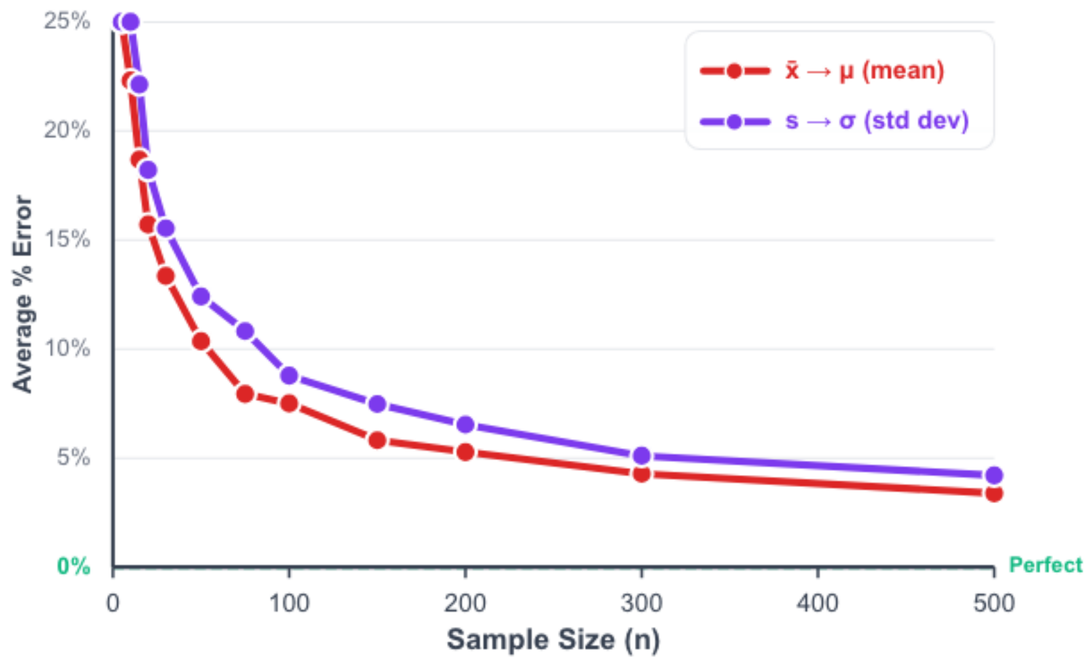
$$\bar{X} \xrightarrow{P} \mu$$
$$s \xrightarrow{P} \sigma$$

Sample statistics **converge in probability** to population parameters.

Stars: N = 119,626 | μ = 15.7 | σ = 14.3

Estimators Converge to Truth

As sample size n increases, both **sample mean (x̄)** and **sample std (s)** converge to the true population values



▶ Run Simulation

📐 Law of Large Numbers

Both **x̄** and **s** are **consistent estimators**.

As $n \to \infty$:

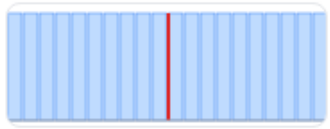
$$\bar{X} \xrightarrow{P} \mu$$

$$s \xrightarrow{P} \sigma$$

Sample statistics **converge in probability** to population parameters.

Stars: N = 119,626 | μ = 15.7 | σ = 14.3

Does Distribution Shape Matter?



Uniform

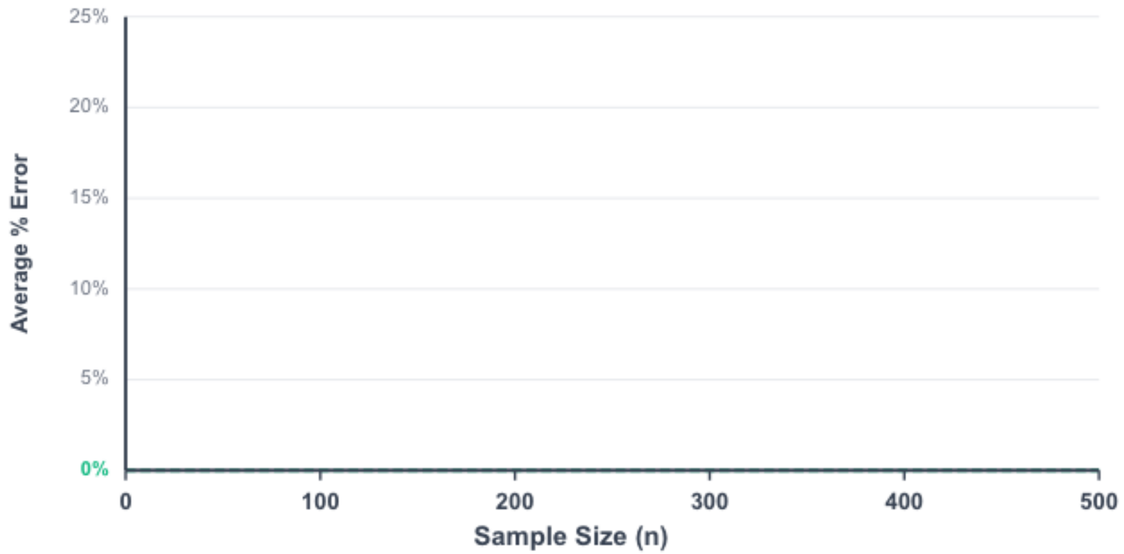
Skewness: **Symmetric**

Star Brightness

Skewness: **Right Skewed**

% Error in Sample Standard Deviation

For each n: draw n values → compute sample std (s) → compare to true σ



▶ Run Simulation

Real Data vs Theory

- **Uniform:** Textbook distribution, perfectly symmetric
- **Stars:** Real astronomical data with natural skew

The gap shows the cost of skewness

Skewed data needs more samples to estimate σ accurately

Star Dataset

N = 119,626 stars

μ = 18.5, σ = 11.2

Does Distribution Shape Matter?

Uniform

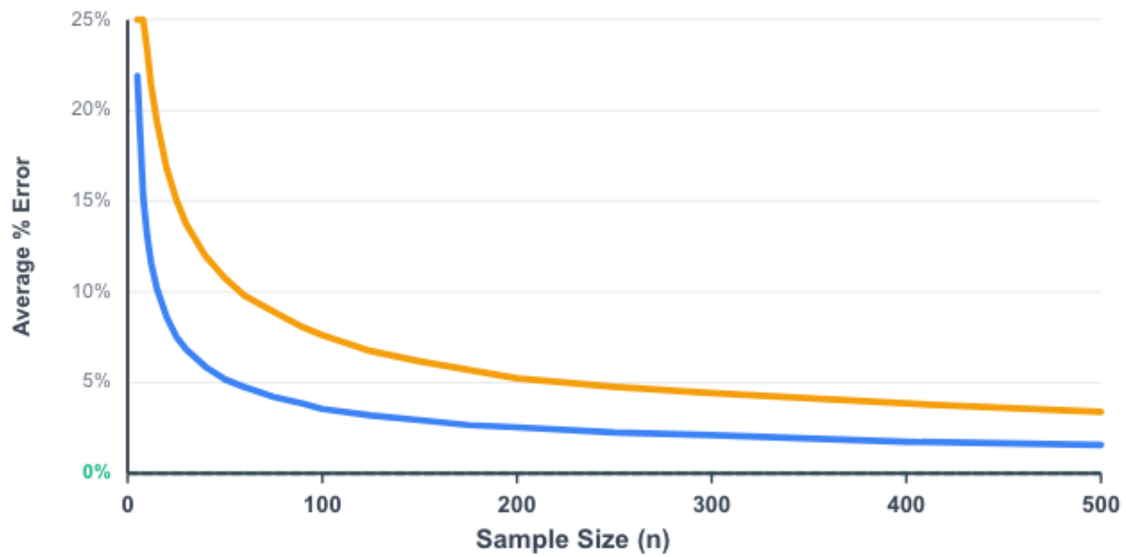
Skewness: **Symmetric**

Star Brightness

Skewness: **Right Skewed**

% Error in Sample Standard Deviation

For each n: draw n values → compute sample std (s) → compare to true σ



▶ Run Simulation

Real Data vs Theory

- **Uniform:** Textbook distribution, perfectly symmetric
- **Stars:** Real astronomical data with natural skew

The gap shows the cost of skewness

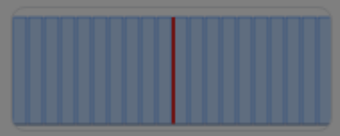
Skewed data needs more samples to estimate σ accurately

Star Dataset

N = 119,626 stars

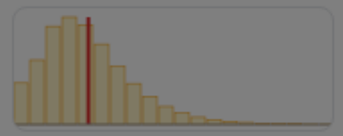
μ = 18.5, σ = 11.2

Does Distribution Shape Matter?



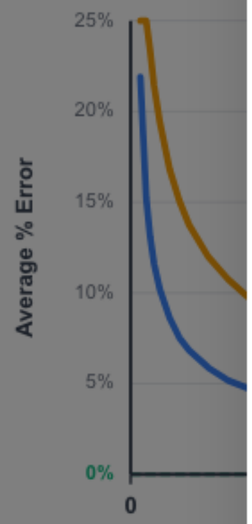
Uniform

Skewness: `Symmetric`



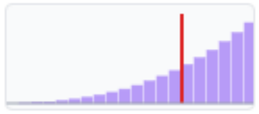
Star Brightness

Skewness: `Right Skewed`



What is Skewness?

Skewness measures the *asymmetry* of a distribution. It tells us whether the data is spread more to one side of the mean.

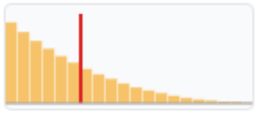


Left Skewed

Skewness: `-0.8`

No Skew

Skewness: `0.0`



Right Skewed

Skewness: `+0.8`

Left (Negative)

Long tail stretches left. Most values are high, but some extreme low values pull the mean down.

Symmetric (Zero)

Data is balanced around the mean. Normal and uniform distributions have zero skewness.

Right (Positive)

Long tail stretches right. Most values are low, but some extreme high values pull the mean up.

Star brightness has positive skewness: many dim stars, few very bright ones.

▶ Run Simulation

Real Data vs Theory

- **Uniform:** Textbook distribution, perfectly symmetric
- **Stars:** Real astronomical data with natural skew

The gap shows the cost of skewness

Skewed data needs more samples to estimate σ accurately

Star Dataset

N = 119,626 stars
μ = 18.5, σ = 11.2

The Central Limit Theorem

Sample n values, compute $\bar{x}$, repeat. **The distribution of $\bar{x}$ becomes Normal.**

POPULATION

Discrete | Continuous

- Normal
- Exponential
- Uniform [0,1]

SAMPLE SIZE

n = 5

Larger n → narrower spread

▶ Run | Reset

Samples: 0

What is CLT?

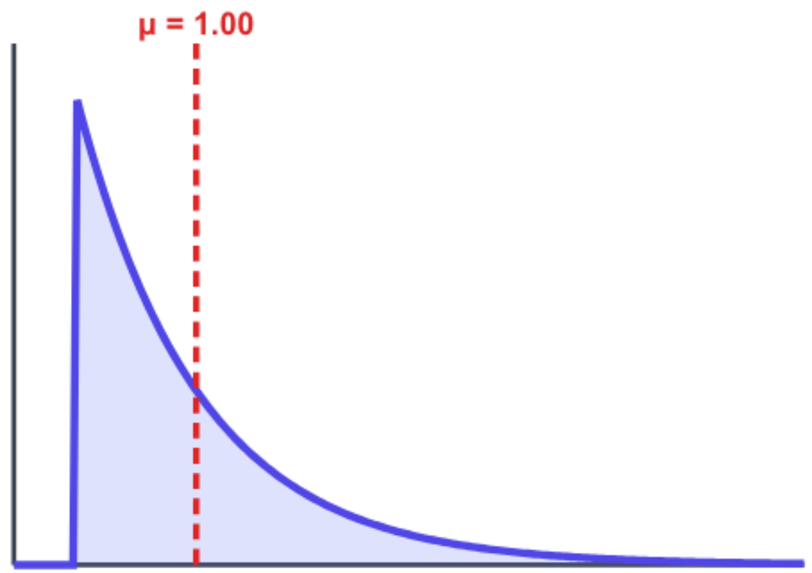
Show 95% CI

⚠ Common Misconception

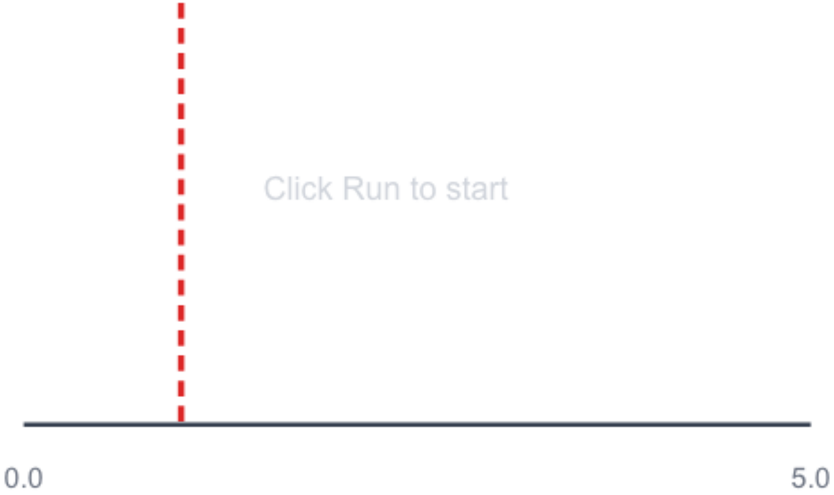
When CLT Fails

Speed:

Population: Exponential



Distribution of Sample Means ($\bar{x}$)



The Central Limit Theorem

Sample n values, compute $\bar{x}$, repeat. **The distribution of $\bar{x}$ becomes Normal.**

POPULATION

Discrete | Continuous

- Normal
- Exponential
- Uniform [0,1]

SAMPLE SIZE

n = 5

Larger n → narrower spread

Pause | Reset

Samples: **80**

What is CLT?

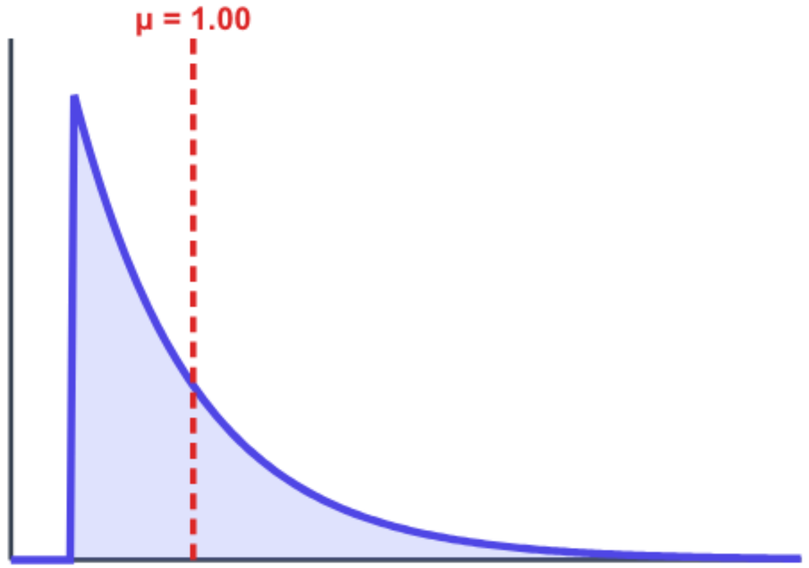
Show 95% CI

Common Misconception

When CLT Fails

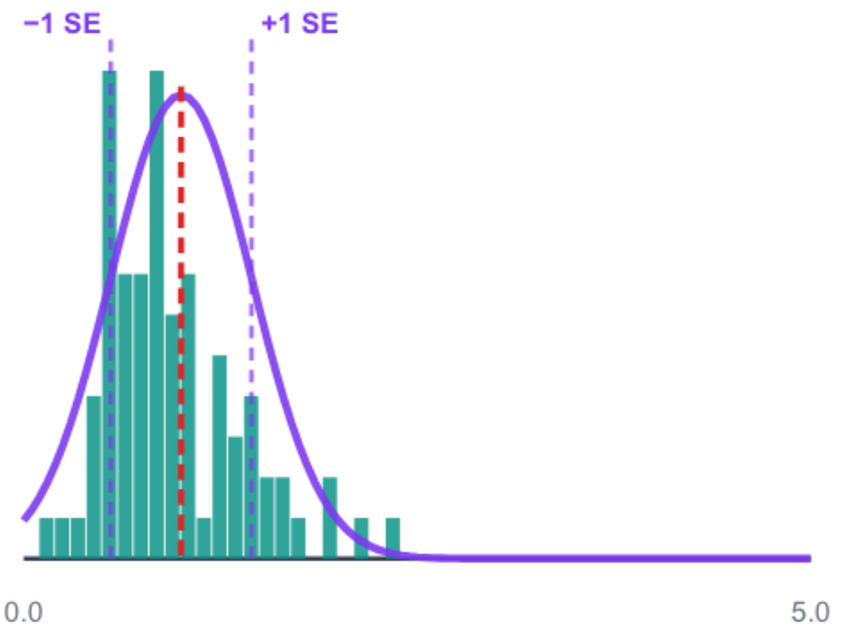
Speed:

Population: Exponential



Distribution of Sample Means ($\bar{x}$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



What is Standard Error (SE)?

The **standard deviation of sample means**. It measures how much x̄ typically varies from sample to sample. Larger n → smaller SE → more precise estimates.

$$SE = \sigma / \sqrt{n} \approx 0.447$$

The Central Limit Theorem

Sample n values, compute $\bar{x}$, repeat. **The distribution of $\bar{x}$ becomes Normal.**

POPULATION

Discrete | Continuous

- Normal
- Exponential
- Uniform [0,1]

SAMPLE SIZE

n = 5

Larger n → narrower spread

⏸ Pause | Reset

Samples: **90**

📐 What is CLT?

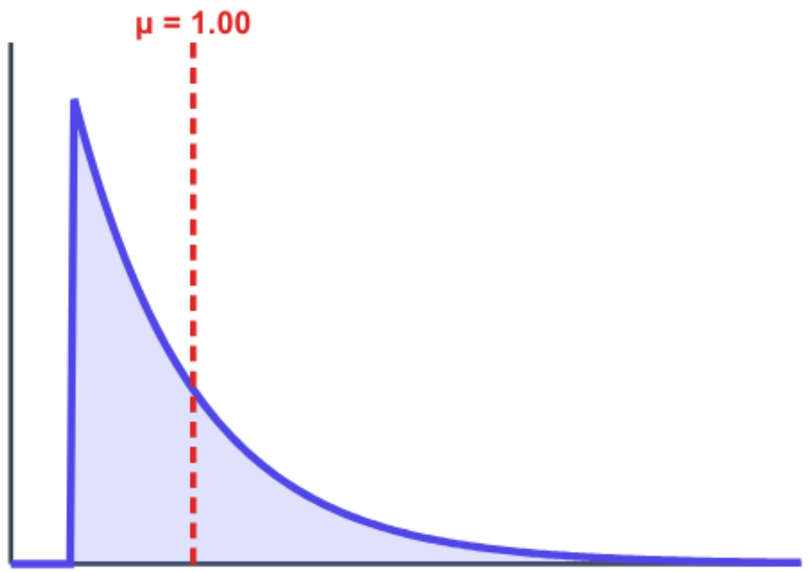
✓ 95% CI Shown

⚠️ Common Misconception

⚡ When CLT Fails

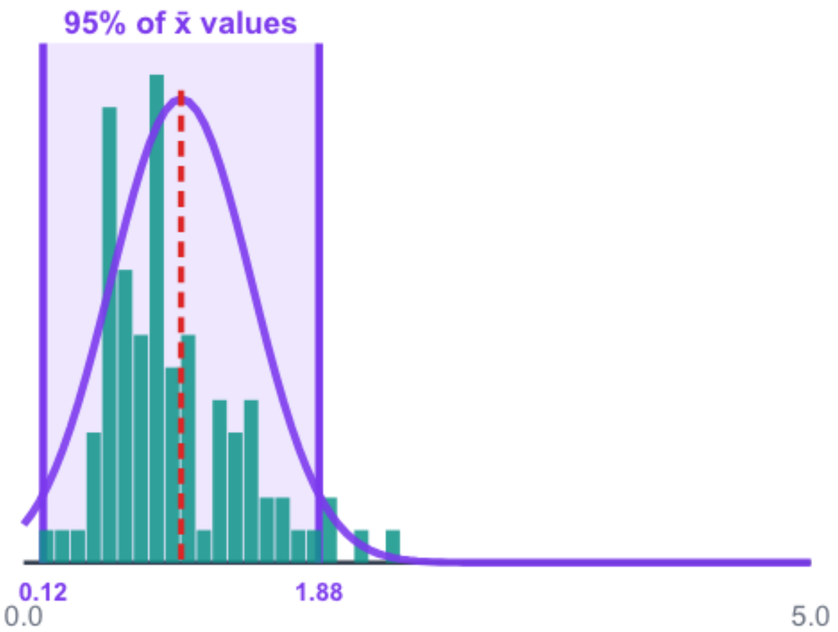
Speed:

Population: Exponential



Distribution of Sample Means ($\bar{x}$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



What is a 95% Confidence Interval?

If we repeated this sampling process many times, **95% of the intervals** we construct would contain the true population mean μ.

μ ± 1.96 × SE

[0.12, 1.88]

POPULATION

Discrete | Continuous

Normal

Exponential

Uniform [0,1]

SAMPLE SIZE

n = 5

Larger n → narrower spread

Pause | Reset

Samples: 100

✓ CLT Definition

✓ 95% CI Shown

⚠ Common Misconception

⚡ When CLT Fails

Speed:

of Sample Means ($\bar{x}$)

$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

8

5.0

val?

ny times, **95% of**

n the true

μ ± 1.96 × SE

[0.12, 1.88]

population mean μ.

The Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are **independent** random variables from **any distribution** with mean μ and variance σ^2, then for large n:

$$\bar{X} \approx \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The sample mean is approximately Normal with mean μ and standard deviation σ/√n

In Plain English

The average of many samples becomes **normally distributed**, regardless of the original distribution's shape.

Key Parameters

- Mean of $\bar{X}$: same as μ
- Std of $\bar{X}$: $s/\sqrt{n}$ (SE)

The Key Implication

We can estimate SE using just **one sample**! Replace σ with s (sample std):

$$\text{SE} \approx \frac{s}{\sqrt{n}}$$

Building Confidence Intervals

With just one sample, we can build a CI for the unknown μ:

$$\bar{x} \pm 1.96 \times \frac{s}{\sqrt{n}}$$

No need to sample repeatedly — one sample gives us everything!

Got it!

Sample n values, compute $\bar{x}$, repeat. **The distribution of $\bar{x}$ becomes Normal.**

POPULATION

Discrete | Continuous

Normal

Exponential

Uniform [0,1]

SAMPLE SIZE

n = 5

Larger n → narrower spread

Pause | Reset

Samples: 117

What is CLT?

✓ 95% CI Shown

✗ Hide Misconception

When CLT Fails

Speed:

Population: Exponential

Distribution of Sample Means ($\bar{x}$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

lues

0.12

1.88

0.0

5.0

Common CI Misconception

✗ WRONG: "There's a 95% probability that μ is in this interval"

✓ RIGHT: "95% of intervals constructed this way contain μ"

The population mean μ is a **fixed value**, not random. Once you calculate an interval, μ is either inside it or it isn't—there's no probability about it. The 95% refers to the **procedure's long-run success rate**, not the probability for any single interval.

Got it!

What is a 95% Confidence Interval?

If we repeated this sampling process many times, **95% of the intervals** we construct would contain the true population mean μ.

$\mu \pm 1.96 \times SE$

[0.12, 1.88]

Sample n values, compute $\bar{x}$, repeat. **The distribution of $\bar{x}$ becomes Normal.**

POPULATION

Discrete | Continuous

Normal

Exponential

Uniform [0,1]

SAMPLE SIZE

n = 5

Larger n → narrower spread

Pause | Reset

Samples: **134**

What is CLT?

✓ 95% CI Shown

⚠ Common Misconception

✗ Hide Caveats

Speed:

Population: Exponential

Distribution of Sample Means ($\bar{x}$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

lues

1.88

5.0

When CLT Can Fail

Small Sample Size (n < 30)

The approximation may be poor for small n, especially with skewed populations. Use n ≥ 30 as a rule of thumb, or larger for very skewed data.

Heavy-Tailed Distributions

Distributions with extreme outliers (like Cauchy or Pareto) may need very large n. Some have undefined variance, so CLT doesn't apply at all!

Non-Independent Samples

CLT assumes observations are independent. Time series, clustered data, or sampling without replacement from small populations violate this.

Understood!

What is a 95% Confidence Interval?

If we repeated this sampling process many times, **95% of the intervals** we construct would contain the true population mean μ.

μ ± 1.96 × SE

[0.12, 1.88]

How Many Patients Do We Need?

💊 A drug claims to lower blood pressure by **5 mmHg**. Each patient costs **$50k**. How many to prove it works?

95% Confidence Interval: —

Fewer patients **Drag or type number of patients** More patients

9 Calculate CI

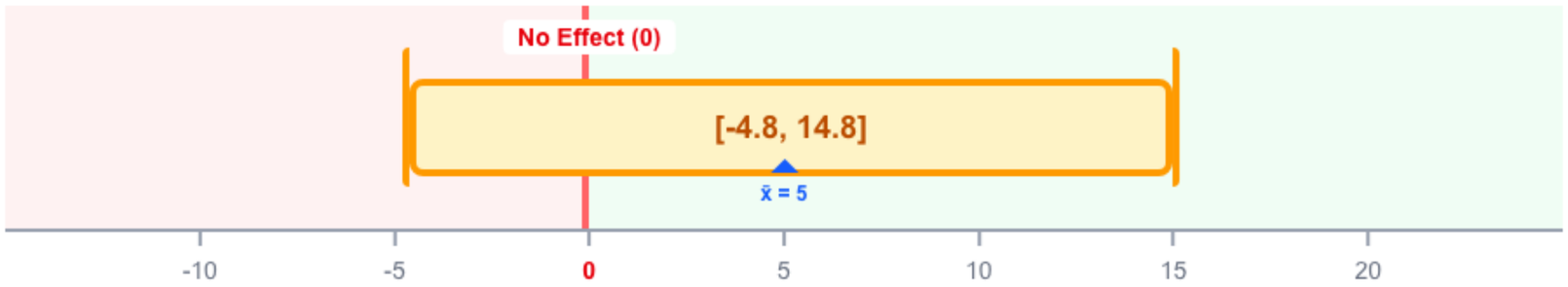
PATIENTS	BUDGET	CI WIDTH
9	$450k	±9.8

$SE = \sigma/\sqrt{n}$ — to halve the CI width, you need 4× the patients

How Many Patients Do We Need?

💊 A drug claims to lower blood pressure by **5 mmHg**. Each patient costs **$50k**. How many to prove it works?

95% Confidence Interval: **? Inconclusive**



Fewer patients **Drag or type number of patients** More patients

9 ✓ CI Shown

PATIENTS	BUDGET	CI WIDTH
9	$450k	±9.8

$SE = \sigma/\sqrt{n}$ — to halve the CI width, you need 4× the patients

How Many Patients Do We Need?

💊 A drug claims to lower blood pressure by **5 mmHg**. Each patient costs **$50k**. How many to prove it works?

95% Confidence Interval: ✓ **Drug Works!**

Fewer patients **Drag or type number of patients** More patients

50

✓ CI Shown

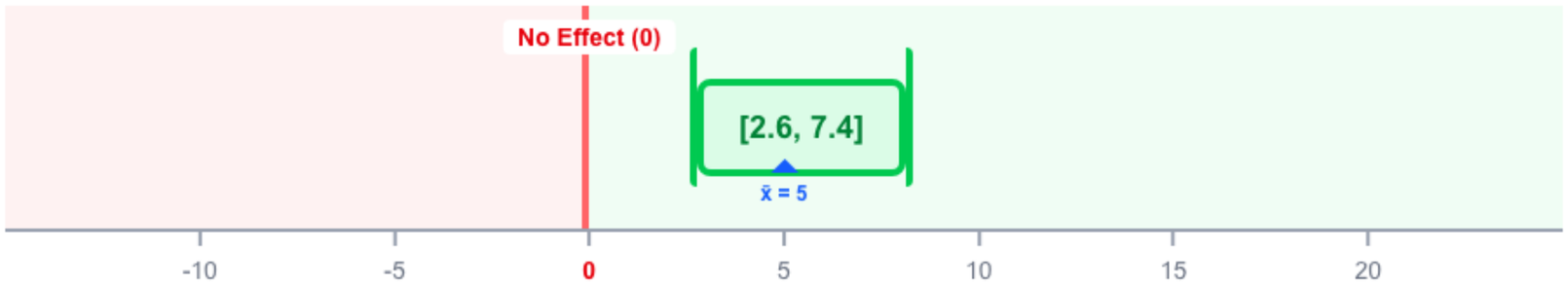
PATIENTS	BUDGET	CI WIDTH
50	$2.5M	±4.2

$SE = \sigma/\sqrt{n}$ — to halve the CI width, you need 4× the patients

How Many Patients Do We Need?

💊 A drug claims to lower blood pressure by **5 mmHg**. Each patient costs **$50k**. How many to prove it works?

95% Confidence Interval: ✓ **Drug Works!**



Fewer patients **Drag or type number of patients** More patients

150

✓ CI Shown

PATIENTS	BUDGET	CI WIDTH
150	$7.5M	±2.4

$SE = \sigma/\sqrt{n}$ — to halve the CI width, you need 4× the patients

A **95% Confidence Interval** means: if we repeated this study many times, **95% of the intervals** would contain the true population mean.

PARAMETERS

True Mean μ	100
Population σ	20
Sample Size n	30

Adjust sample size

95% CONFIDENCE INTERVAL FORMULA

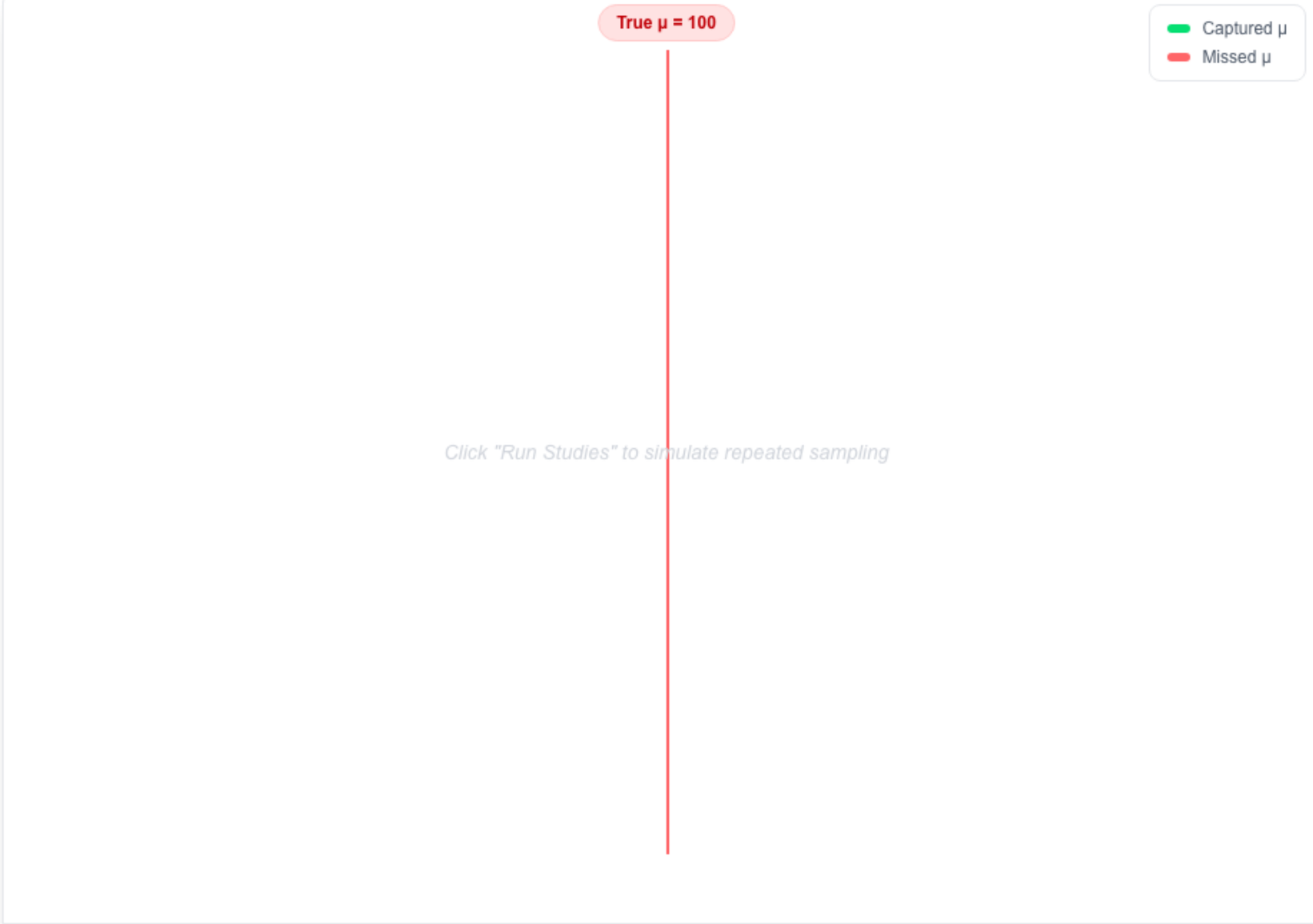
$$\bar{x} \pm 2 \times \frac{\sigma}{\sqrt{n}}$$

Standard Error (SE)	3.65
Margin of Error	±7.30

▶ Run Studies

Reset

STUDIES RUN	CAPTURE RATE
0	0.0%



A **95% Confidence Interval** means: if we repeated this study many times, **95% of the intervals** would contain the true population mean.

PARAMETERS

True Mean μ	100
Population σ	20
Sample Size n	30

Adjust sample size

95% CONFIDENCE INTERVAL FORMULA

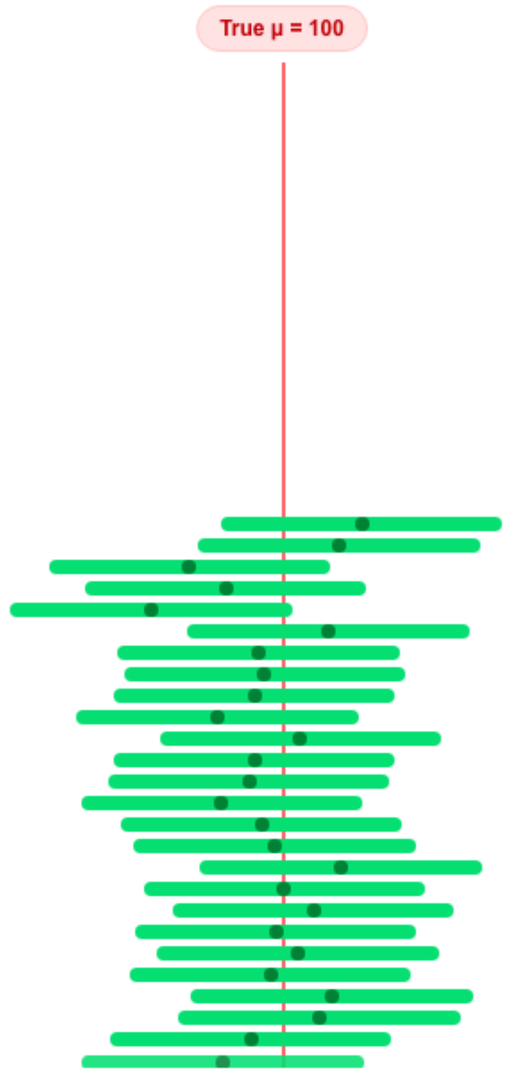
$$\bar{x} \pm 2 \times \frac{\sigma}{\sqrt{n}}$$

Standard Error (SE)	3.65
Margin of Error	±7.30

Pause

Reset

STUDIES RUN	CAPTURE RATE
26	100.0%



Captured μ (26) Missed μ (0) True Population Mean

Conclusion: Sampling & Confidence

1. **Samples Estimate Populations**
   We use samples (n) to learn about populations (N) we can't fully measure.

2. **Standard Error**
   SE = s / √n tells us how much sample means typically vary.

3. **Central Limit Theorem**
   Sample means are approximately Normal, regardless of the population shape.

4. **The √n Trade-off**
   To halve your error, you need 4× the sample size.

5. **95% Confidence Intervals**
   CI = $\bar{x} \pm 2 \times SE$. About 95% of intervals will contain the true mean.

Lab: Apply to real astronomical data (120k stars)