Chapter 15

Sampling Distributions

Intended Learning Outcomes

- Distinguish parameters from statistics
- State the Law of Large Numbers
- Define and interpret a sampling distribution
- Calculate the standard error of $\bar{X}_n$
- State and apply the Central Limit Theorem
- Compute probabilities involving $\bar{X}_n$

Guiding Questions

Our goal in statistics is to learn about *populations* using *samples*.

But samples vary, so how can we make reliable inferences?

How do we know if an observed result reflects
real evidence or just random chance?

PART 1

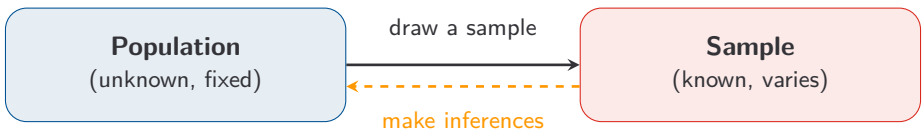
Parameters and Statistics

Distinguishing what we know from what we estimate

Parameters and Statistics

Parameter vs. Statistic

- A **parameter** is a number that describes the entire *population*. Parameters are typically **unknown** and **fixed**.
- A **statistic** is a number computed from *sample* data. Statistics are **known** but **vary** from sample to sample.



Identify Parameters and Statistics

Example 15.1

For each scenario, identify whether the value is a **parameter** or a **statistic**:

1. The average height of all 30,000 Western students is 170 cm.

2. You measure the height of 50 randomly selected students and find their average height is 172 cm.

3. The true proportion of left-handed people worldwide is 10%.

4. In a poll of 500 voters, 52% support a policy.

Notation: Parameters vs. Statistics

Some examples of common parameters and statistics of interest are:

Measure	Population (Parameter)	Sample (Statistic)
Mean		
Standard Deviation		
Proportion		

Focus: the sample mean as an estimatoer of the population mean

We often use the sample mean $\bar{X}_n$ to estimate the population mean μ.

But is this a good idea?

PART 2

Does $\bar{X}_n$ Aim at the Right Target?

The Law of Large Numbers

Building the Running Mean

Example 15.2: Days 1–5

Context: A café claims the average wait is $\mu = 4$ min. You time your wait each morning.

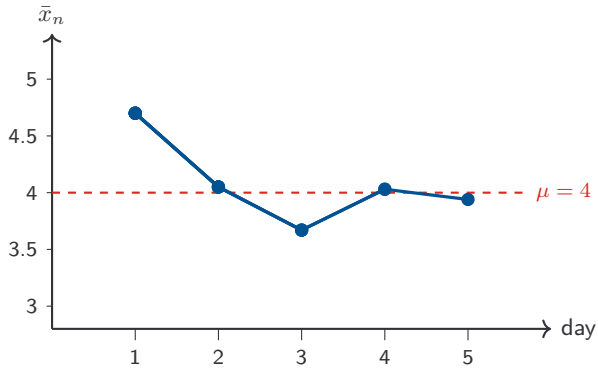
Day 1: waited 4.7 min
$\bar{x}_1 = 4.7/1 = 4.70$

Day 2: waited 3.4 min
$\bar{x}_2 = 8.1/2 = 4.05$

Day 3: waited 2.9 min
$\bar{x}_3 = 11.0/3 = 3.67$

Day 4: waited 5.1 min
$\bar{x}_4 = 16.1/4 = 4.03$

Day 5: waited 3.6 min
$\bar{x}_5 = 19.7/5 = 3.94$



The Law of Large Numbers in Action

Example 15.2: Convergence

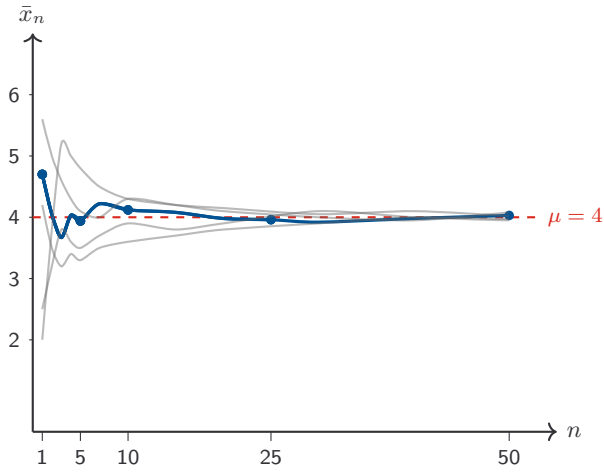
Context: After 5 mornings, $\bar{x}_5 = 3.94$. Keep going...

Day 10: $\bar{x}_{10} = 4.12$ (closer)

Day 25: $\bar{x}_{25} = 3.96$ (very close)

Day 50: $\bar{x}_{50} = 4.03$ (essentially at μ)

This happens **every time** you repeat the experiment.



The Law of Large Numbers

Law of Large Numbers (LLN)

If $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ, then as the sample size n increases, the sample mean $\bar{X}_n$ converges to μ with high probability.

In plain language:

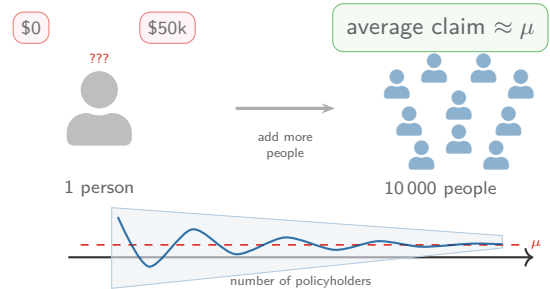
- Larger samples give **better estimates**
- $\bar{X}_n$ gets closer to μ as n grows
- This justifies using $\bar{X}_n$ to estimate μ

Intuition: The LLN tells us *where* $\bar{X}_n$ converges (to μ), but not *how much* it varies along the way. To understand variability, we need its **sampling distribution**.

LLN in Practice: Insurance & Risk Pooling

The core idea:

- Cannot predict whether *you* will file a claim
- With thousands of policyholders, the *average* claim cost converges reliably to μ



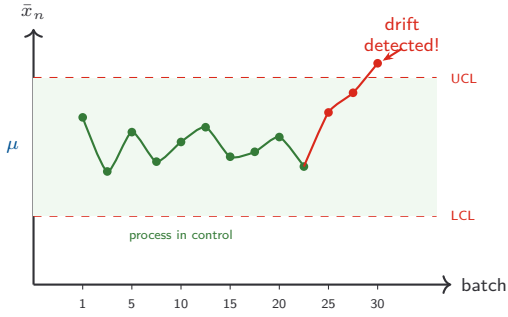
Intuition: This is why insurance *works*:
the company doesn't need to predict your outcome - only the average across the pool.

LLN in Practice: Quality Control

The core idea:

- A factory cannot test every unit
- Measures the average defect rate across each production run
- As the run grows, the sample average tracks the true defect rate, catching process drift early

Control Chart (Schematic)



Generally, control limits are set at $\mu \pm 3 \cdot \mathrm{SE}$.

Intuition: If the running average suddenly deviates from the known μ, something has *changed* in the production process.
This is the basis of **control charts** in statistical process control.

What the LLN Does *Not* Say

Common Misconception: The Gambler's Fallacy

"I've flipped 10 tails in a row. Heads must be due next."

What the LLN actually says:

As $n \to \infty$, $\bar{X}_n \to \mu$ because early deviations get *diluted*, not corrected.
Each flip is independent: past outcomes carry no information about future ones.

Real-world stakes:

- Casinos profit because gamblers misread the LLN as a short-run correction mechanism.
- Investors who "average down" after repeated losses may be committing the same error.
- The LLN is a *long-run* guarantee, not a short-run balancing force.

PART 3

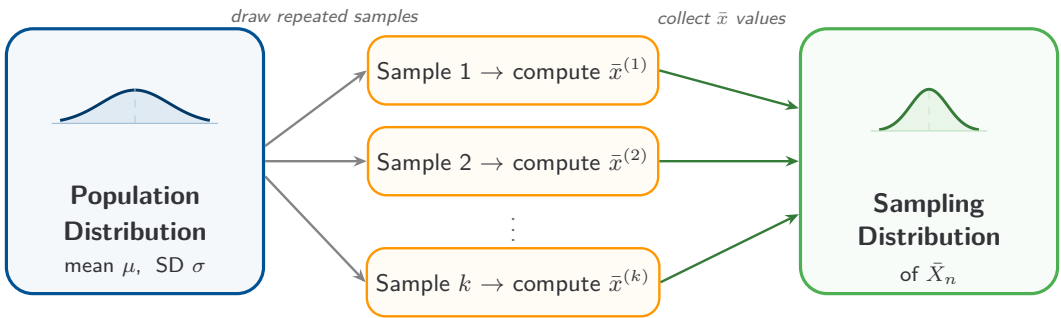
How Much Does $\bar{X}_n$ Vary?

Sampling distributions and the standard error

What Is a Sampling Distribution?

Sampling Distribution

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in *all possible samples* of the same size from the same population.



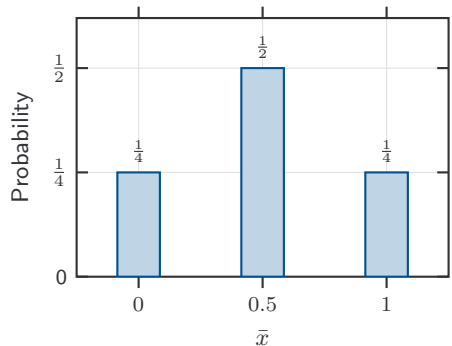
Sampling Distribution by Hand ($n = 2$)

Context: A fair coin is tossed. Let $X = 0$ (tails) or $X = 1$ (heads), each with probability $1/2$.

All $2^2 = 4$ possible samples of size 2:

Sample	$\bar{x}$
(0, 0)	0
(0, 1)	0.5
(1, 0)	0.5
(1, 1)	1

Sampling distribution of $\bar{X}_2$:



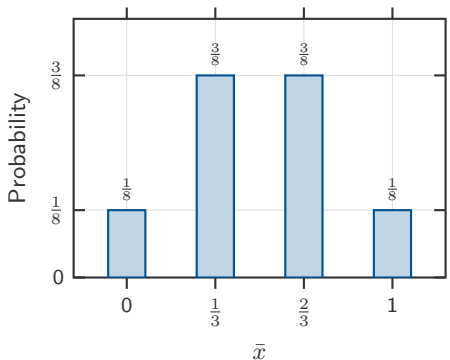
Sampling Distribution by Hand ($n = 3$)

Context: Same coin flip: $X = 0$ or 1, each with probability $1/2$.

All $2^3 = 8$ possible samples of size 3:

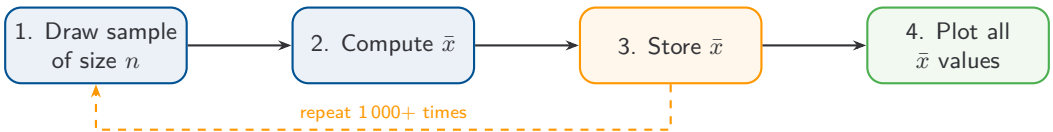
Sample	$\bar{x}$
(0, 0, 0)	0
(0, 0, 1)	$1/3$
(0, 1, 0)	$1/3$
(1, 0, 0)	$1/3$
(0, 1, 1)	$2/3$
(1, 0, 1)	$2/3$
(1, 1, 0)	$2/3$
(1, 1, 1)	1

Sampling distribution of $\bar{X}_3$:



From Enumeration to Simulation

With only a few outcomes, we can list every possible sample. For larger or continuous populations, we **simulate** instead:



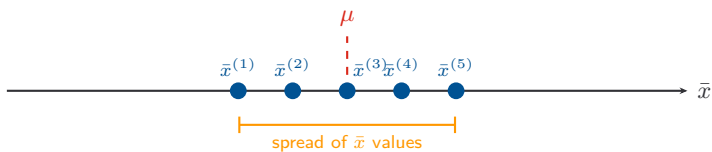
> 💬 **Note:** The result is technically an **empirical** sampling distribution.
> With enough repetitions, it closely approximates the true (theoretical) sampling distribution.

Building a Sampling Distribution

Why Sampling Distributions Matter

The LLN tells us $\bar{X}_n$ heads toward μ as n grows, but it doesn't tell us how *close* we can expect $\bar{X}_n$ to be for a given sample size.

Imagine drawing five different samples of the same size from the same population. Each gives a slightly different $\bar{x}$:

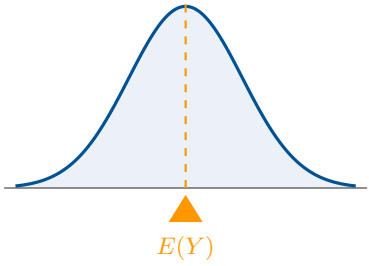


Two natural questions:

- How spread out are these sample means?
- What is the relationship between the spread and the sample size n?

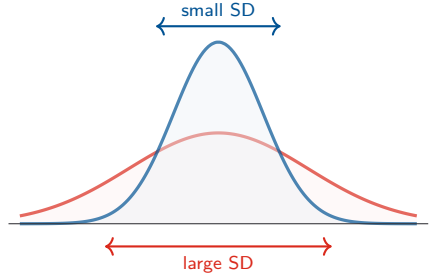
Recap: Centre and Spread

Centre: $E(Y)$



The **expected value** $E(Y)$ is the balance point of the distribution: the long-run average if you could repeat the random process forever.

Spread: $\mathrm{SD}(Y)$



The **standard deviation** $\mathrm{SD}(Y)$ measures how far values typically fall from the centre.

The **variance** is the same idea in squared units: $\mathrm{Var}(Y) = [\mathrm{SD}(Y)]^2$.

Standard Error vs. Standard Deviation

Standard Error (SE)

The **standard error** of a statistic is the standard deviation of its *sampling distribution*. For the sample mean:

$$\mathrm{SE}(\bar{X}_n) \;=\; \mathrm{SD}(\bar{X}_n) \;=\; \frac{\sigma}{\sqrt{n}}$$

Why not just say "standard deviation"?

- $\sigma = \mathrm{SD}(X)$ describes variability of **individual observations** in the population.
- $\mathrm{SE} = \sigma/\sqrt{n}$ describes variability of **sample means** across repeated samples.

Intuition: Both are standard deviations, but of *different things*: σ measures spread of raw data; SE measures precision of a statistic.

Mean and Standard Error of $\bar{X}_n$

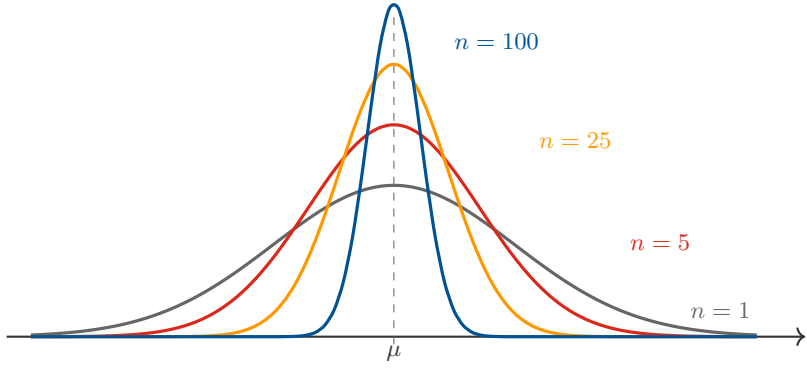
Mean and Standard Error of the Sample Mean

If X is a population with mean μ and standard deviation σ, then the sample mean $\bar{X}_n$ has:

- **Mean:**
- **Standard Error (SE):**

Principle: The sample mean is **unbiased** (on average equals μ), and the SE **shrinks** as n grows: larger samples give more precise estimates.

How Sample Size Affects the Sampling Distribution



Intuition: As n increases, the sampling distribution gets **narrower** (smaller SE), but the centre stays at μ.

Finding SE's

Example 15.3

Context: Adult heights have $\mu = 170$ cm and $\sigma = 10$ cm.

Calculate: The standard error for each sample size.

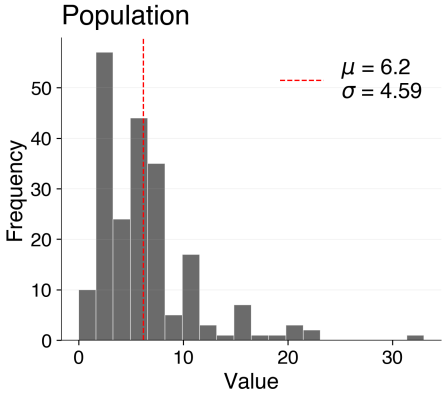
Sample size n	Calculation	SE
4		
9		
25		
100		
400		

Countries Visited

Example 15.4

Context: The population distribution of countries visited by DS 1000 students ($\mu = 6.2$, $\sigma = 4.59$).

Find: The sampling distribution of $\bar{X}_n$ for several sample sizes.



(a) What is the parameter of interest?

(b) What happens to mean and SD as n increases?

Using the SE to Find a Minimum Sample Size

The relationship $\mathrm{SE} = \sigma/\sqrt{n}$ works in both directions: given a desired standard error, we can solve for the required n.

Principle: Quadrupling the sample size *halves* the SE: precision is gained slowly, so plan your sample size in advance.

Computing Mean and Standard Error

Example 15.5

Context: The average number of countries visited by DS 1000 students has $\mu = 6.2$ and $\sigma = 4.59$.

Find: The required sample sizes for target precision levels.

(a) What sample size is needed to halve the SE?

(b) What sample size n is needed so that the SE is at most 0.5 countries?

PART 4

How is the Sample Mean Distributed?

From Normal populations to the Central Limit Theorem

Recap: What We Know So Far

So far, we answered two questions:

✔ **Does $\bar{X}_n$ aim at the right target?** Yes: the LLN says $\bar{X}_n \to \mu$ as n grows.

✔ **How much does $\bar{X}_n$ vary?** SE $= \sigma/\sqrt{n}$. Larger $n \to$ less variability.

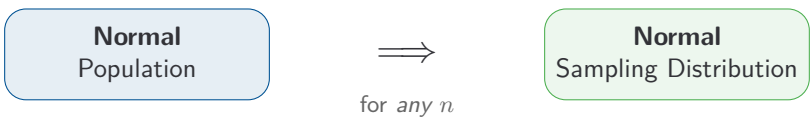
? **What distribution does the variability follow?** ← next

Sampling from a Normal Population

Sampling Distribution of Normal Populations

If X is Normally distributed with mean μ and standard deviation σ, then the sampling distribution of $\bar{X}_n$ for samples of size n is:

This result holds for *any* sample size $n \geq 1$.



Electric Vehicle Range

Example 15.6: Setup

Context: Real-world range for a popular EV model is Normally distributed with $\mu = 263$ miles and $\sigma = 25$ miles (Consumer Reports highway tests, 2023). A reviewer tests $n = 16$ vehicles.

(a) What is the sampling distribution of $\bar{X}_{16}$?

(b) What is the probability that $\bar{X}_{16} \geq 270$?

Electric Vehicle Range

Example 15.6: Calculation

(c) What is $P(250 < \bar{X}_{16} < 276)$?

(d) The reviewer finds $\bar{x} = 280$ miles. What is the probability of observing a sample mean this high or higher under the stated specs?

EV Range: What Sample Size?

Example 15.7

Context: EV range has $\mu = 263$ miles and $\sigma = 25$ miles (same as the previous example).

Find: The minimum sample size and the SE for a given n.

(a) A reviewer wants the SE to be at most 5 miles. What is the minimum sample size?

(b) If the reviewer tests $n = 100$ vehicles, what is the SE?

PART 5

The Central Limit Theorem

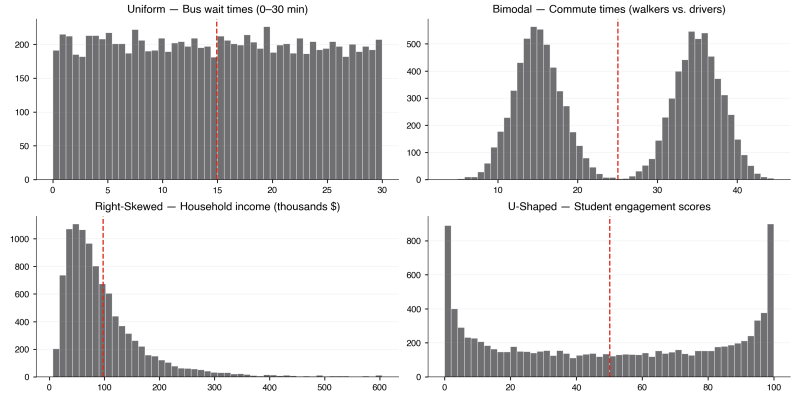
Why sample means become approximately Normal regardless of the population's distribution

Four Strange Populations

Example 15.8

None of these population distributions are normal. What will the sampling distribution of $\bar{X}_n$ look like for large n?

Let's investigate.



The Central Limit Theorem

Central Limit Theorem (CLT)

For a random sample of size n from *any* population with mean μ and standard deviation σ, as n increases, the sampling distribution of $\bar{X}_n$ approaches a Normal distribution:

$$\bar{X}_n \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

where $\dot{\sim}$ means "is approximately distributed as." The approximation improves as n increases.

Any
Population

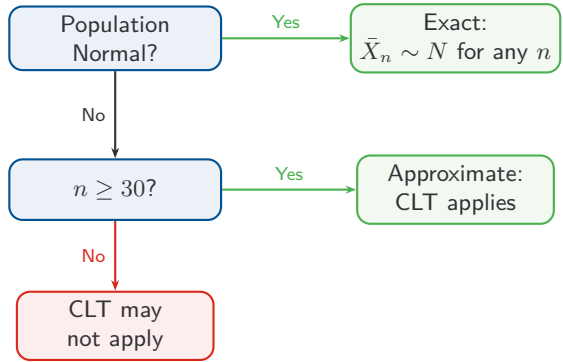
$\Longrightarrow$
for large n

Approximately Normal
Sampling Distribution

When Does the CLT Apply?

Conditions:

1. The sample is **random**
2. The sample size is **large enough**:
   - $n \geq 30$ (rule of thumb), *or*
   - The population is approximately Normal



Principle: The CLT lets us use Normal probability calculations for *any* population, as long as n is large enough. If the population is already Normal, the result is exact for any n.

Potato Crate Quality Control

Example 15.9: Setup

Context: A farm ships crates of potatoes. Potato weight has $\mu = 300\,\text{g}$ and $\sigma = 50\,\text{g}$. Each crate contains $n = 100$ potatoes. An inspector samples crates and calculates the average weight per potato $\bar{X}_{100}$.

(a) What are the mean and standard deviation of $\bar{X}_{100}$?

(b) Approximate the probability that the average weight per potato in a crate is between 295 g and 305 g.

Potato Crate Quality Control

Example 15.9: Continued

Context: Potato weight: $\mu = 300\,\text{g}$, $\sigma = 50\,\text{g}$, $n = 100$; $\bar{X}_{100} \dot{\sim} N(300,\ 5)$.

(c) Find the probability that the *total* weight of potatoes in a crate exceeds 31 kg.

CLT and Discrete distributions

The CLT applies to *any* distribution

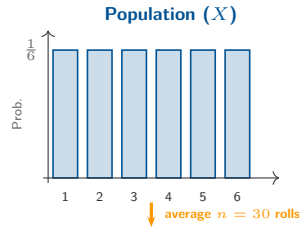
The CLT doesn't require a continuous population. Even if individual observations can only take on a **few discrete values**, the *average* of many such observations is approximately Normal.

Examples:

- Die rolls ($X \in \{1, \ldots, 6\}$)
- Coin flips ($X \in \{0, 1\}$)

As long as n is large enough, $\bar{X}_n$ is approximately Normal regardless of whether the population is discrete or continuous.

Die rolls: $\mu = 3.5, \ \sigma \approx 1.71$



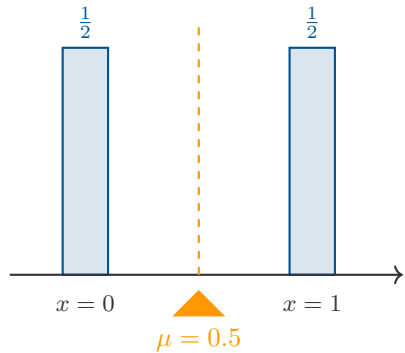
Digression: Mean of a Coin Toss

Context: Let $X = 1$ (heads) or $X = 0$ (tails), each with probability $\frac{1}{2}$.

Outcome	x	Prob.
Tails	0	$\frac{1}{2}$
Heads	1	$\frac{1}{2}$

Half the time you get 0, half the time you get 1.

The long-run average is the balance point:



Standard Deviation of a Coin Toss

Recall: $\mu = 0.5$. How far does each flip land from the mean?

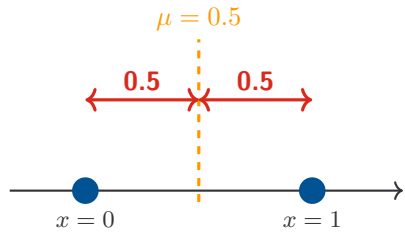
Every flip lands exactly 0.5 away from μ:

- Tails ($x = 0$):

  distance $= |0 - 0.5| = 0.5$

- Heads ($x = 1$):

  distance $= |1 - 0.5| = 0.5$



The typical distance from the mean:

Principle: In general, for a binary random variable having outcomes $\{0, 1\}$ where $P(X = 1) = p$, the mean is $E(X) = \mu = p$ and the SD is $\sigma = \sqrt{p(1-p)}$.

Coin Tosses and the CLT

Example 15.10: Setup

Context: A fair coin has $\mu = p = 0.5$ and $\sigma = 0.5$. We toss it $n = 100$ times and observe 60 heads ($\hat{p} = \bar{X}_{100} = 0.60$).

(a) What are the mean and SE of the sampling distribution of $\bar{X}_{100}$?

(b) What is the sampling distribution of $\bar{X}_{100}$?

Coin Tosses and the CLT

Example 15.10: Calculation

Context: Fair coin, $n = 100$; $\bar{X}_{100} \dot{\sim} N(0.5,\ 0.05)$. Observed $\hat{p} = 0.60$.

(c) What is $P(\bar{X}_{100} \geq 0.60)$?

Putting It All Together

Example 15.11

Task: For each population and sample size, write the sampling distribution of $\bar{X}_n$ and compute the SE.

Population	μ	σ	n	SE	Sampling dist.
Heights (Normal)	170	10	64		
EV range (Normal)	263	25	49		
Countries (skewed)	6.2	4.59	100		
Income (skewed)	95.1	74.73	50		

Principle: Normal population $\Rightarrow$ exact ($\sim$).
Non-Normal population with $n \geq 30 \Rightarrow$ approximately normal.

CHAPTER 15

Summary

Key ideas and formulas to carry forward

Key Takeaways

- **Parameters vs. Statistics:** A parameter (μ, σ, p) describes a population and is fixed but unknown. A statistic ($\bar{x}$, s, $\hat{p}$) is computed from a sample and varies from sample to sample.
- **Law of Large Numbers:** As the sample size grows, $\bar{X}_n$ converges to μ. Larger samples produce more reliable estimates.
- **Standard Error:** The SE $= \sigma/\sqrt{n}$ measures how much $\bar{X}_n$ varies across samples. Quadrupling n halves the SE.
- **Sampling distribution of $\bar{X}_n$:**
  If the population is Normal, then $\bar{X}_n \sim N(\mu,\ \sigma/\sqrt{n})$ exactly for any n.
  If the population is not Normal, the CLT gives $\bar{X}_n \dot{\sim} N(\mu,\ \sigma/\sqrt{n})$ for $n \geq 30$.

Chapter 15: Formula Card

Concept	Formula
Mean of $\bar{X}_n$	$E(\bar{X}_n) = \mu$
Standard Error	$\mathrm{SE} = \dfrac{\sigma}{\sqrt{n}}$
Normal population	$\bar{X}_n \sim N\left(\mu,\ \dfrac{\sigma}{\sqrt{n}}\right)$ for any n
CLT (any population)	$\bar{X}_n \dot{\sim} N\left(\mu,\ \dfrac{\sigma}{\sqrt{n}}\right)$ for $n \geq 30$
Minimum sample size	$n \geq \left(\dfrac{\sigma}{d}\right)^2$ where $d =$ target SE

PRACTICE

Practice Questions

Practice: Study Hours

Example 15.12

Context: University students study an average of $\mu = 15$ hours per week with $\sigma = 6$ hours. The distribution is right-skewed. You survey $n = 36$ randomly selected students and find $\bar{x} = 16.5$.

(a) Is $\mu = 15$ a parameter or a statistic? Is $\bar{x} = 16.5$?

(b) What is the SE of $\bar{X}_{36}$?

(c) Write the sampling distribution of $\bar{X}_{36}$. Justify your answer.

Practice: Delivery Times

Example 15.13: Setup

Context: A courier company's delivery times have $\mu = 45$ minutes and $\sigma = 12$ minutes. The distribution is right-skewed (most deliveries are quick, but some take much longer). A manager randomly selects $n = 36$ deliveries.

(a) Does the CLT apply to $\bar{X}_{36}$? Justify your answer.

(b) Write the sampling distribution of $\bar{X}_{36}$.

Practice: Delivery Times

Example 15.13: Calculation

Recall: $\bar{X}_{36} \dot{\sim} N(45,\ 2)$.

(c) Find the probability that the average delivery time in the sample exceeds 49 minutes.

(d) Would the CLT apply if the manager sampled only $n = 10$ deliveries?

Practice: Marathon Finish Times

Example 15.14: Setup

Context: Marathon finish times at a large city marathon have $\mu = 280$ minutes and $\sigma = 40$ minutes. Assume the population is approximately Normal.

(a) A race organizer wants the SE of the sample mean to be at most 5 minutes. What is the minimum sample size?

(b) Using $n = 64$, write the sampling distribution of $\bar{X}_{64}$. Do you need the CLT?

Practice: Marathon Finish Times

Example 15.14: Calculation

Recall: $\bar{X}_{64} \sim N(280,\ 5)$.

(c) Find the probability that the sample mean is below 275 minutes.

(d) Find $P(270 < \bar{X}_{64} < 290)$.

Ride-Share Tip Threshold

Example 15.15: Setup

Context: Tip amounts on a ride-share platform are right-skewed: $\mu = \$2.50$ and $\sigma = \$1.50$ per trip. A driver completes $n = 100$ trips per shift. The platform awards a *Top Earner* badge to drivers whose total shift tips rank in the top 2.3% under normal conditions.

(a) What are the mean and SE of $\bar{X}_{100}$?

(b) Let $S = \sum_{i=1}^{100} X_i$ be the total tips for a shift. What are $E(S)$ and $\mathrm{SD}(S)$?

Ride-Share Tip Threshold

Example 15.15: Calculation

Context: Tip amounts: $\mu = \$2.50$, $\sigma = \$1.50$, $n = 100$; $S \dot{\sim} N(250,\ 15)$.

(c) Find the total tip threshold t^* such that $P(S > t^*) \approx 0.023$. Drivers above this earn the badge.