2 *Point Estimation and Sampling Distributions*

Chapter 1 ended with the central problem of this course: we want to know something about a population, but we can only afford to measure a sample. This chapter builds the machinery that connects the two. It is organized in the following order: the statistical inference framework; point estimators and point estimates; the sampling distribution of a statistic; how to judge an estimator using bias, standard error, and mean squared error; the Central Limit Theorem, which tells us the shape of the sampling distribution of the sample mean; and probability computations for the sample mean, the sample proportion, and differences between two populations.

Every later chapter stands on this one. A confidence interval (Chapter 3) is a statement about a sampling distribution. A hypothesis test (Chapter 4) is a probability computation on a sampling distribution. If you can answer the question ''how does $\bar{X}$ behave from sample to sample?'' quickly and correctly, the rest of the course becomes a sequence of variations on that answer.

Pre-class quiz. *Try these before lecture; the answers follow at the end of the quiz.*

Q1. A port operations manager computes the average dwell time of 40 sampled containers and gets 61.3 hours. Is 61.3 a parameter or a statistic?

Q2. A startup measures signup conversion with a sample of $n = 25$ visitors and computes the standard error of the sample mean. If the sample size is increased to $n = 100$, what happens to the standard error?

Q3. Truck fuel-consumption readings are strongly right-skewed. If you average $n = 50$ readings, is the distribution of that average also strongly right-skewed?

Answers.

Q1. A *statistic*. It is computed from sample data, and a second sample of 40 containers would give a different value. The parameter is the mean dwell time μ of *all* containers passing through the port, and it remains unknown.

Q2. It is cut in half. The standard error of $\bar{X}$ is $\sigma/\sqrt{n}$, so multiplying n by 4 divides the standard error by $\sqrt{4} = 2$. Precision grows with the *square root* of the sample size, not with the sample size itself.

Q3. *No.* By the Central Limit Theorem, the sampling distribution of $\bar{X}$ is approximately normal for large n, regardless of the shape of the population distribution. With $n = 50$ the approximation is usually very good.

Lecture 2 starts here — *Review of estimation*

Lecture 2. This section is covered by Lecture 2.

Reference. Montgomery & Runger, 6th ed., Section 7.1.

Population. The entire collection of units or measurements we want to draw conclusions about.

Parameter. A fixed, unknown numerical property of the population, such as μ, σ^2, or p.

Sample. The subset of the population that is actually observed and measured.

2.1 The Statistical Inference Framework

Statistical inference always involves two worlds. The first is the *population*: the complete collection of measurements we care about. The population is described by *parameters* such as the mean μ, the variance σ^2, or a proportion p. Parameters are fixed numbers, but they are unknown, because we never observe the whole population. The second world is the *sample*: the part of the population we actually measure. From the sample we compute *statistics* such as the sample mean $\bar{x}$, the sample variance s^2, or the sample proportion $\hat{p}$. Statistics are known numbers, but they change from sample to sample.

Probability and statistics move between these two worlds in opposite directions. In a probability course (ISE 205), the population distribution is given and we compute what a sample will look like: known $\rightarrow$ observed. In this course we do the reverse: we start from an observed sample and reason back to the unknown population: observed $\rightarrow$ unknown. This reverse direction is called *statistical inference*, and it splits into two tasks. The first task is *parameter estimation*: ''what is μ?'' The second task is *hypothesis testing*: ''is $\mu = 50$?'' This chapter and the next develop estimation; Chapter 4 develops testing. Figure 2.1 shows the framework.

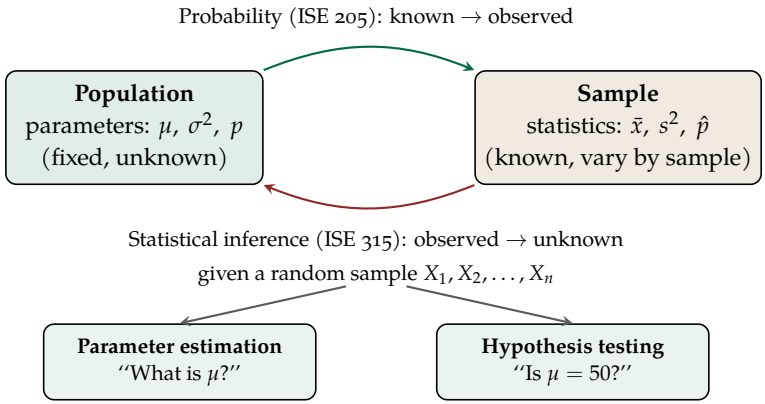

Figure 2.1. The statistical inference framework. Probability reasons from a known population to a sample; inference reasons from an observed sample back to the unknown population.

The bridge between the two worlds is the *random sample*. The random variables $X_1, X_2, \ldots, X_n$ form a random sample of size n if two conditions hold: the X_i are *independent* of one another, and every X_i has the *same probability distribution* as the population. Before the data are collected, each X_i is a random variable; after collection, we observe fixed numbers $x_1, x_2, \ldots, x_n$. Please pay attention to the capitalization convention, because the whole chapter depends on it: capital letters (X_i, $\bar{X}$) denote random variables before sampling; lowercase letters (x_i, $\bar{x}$) denote the numbers actually observed.

Random sample. Random variables $X_1, \ldots, X_n$ that are independent and all follow the same population distribution.

Safety lens. Independence is a modeling claim, not a formality. If a red team probes an AI model with 50 prompts that are small variations of one attack, the effective sample is much smaller than 50, and every formula in this chapter will overstate the precision of the results.

Example 2.1 — Reading a testing problem through the framework

An AI safety team validates the perception module of an autonomous shuttle fleet. The team wants the mean driving distance between disengagements (a human safety driver taking over) across all driving the fleet will ever do under the current software version. The team runs $n = 12$ standardized test drives and records the distance between disengagements on each; the average of the 12 recorded distances is 158.4 km. Identify the population, the parameter, the sample, the statistic, and the estimate.

Solution.

Population.

All driving the fleet performs under the current software version, described by the distribution of distance between disengagements.

Parameter.

The mean distance between disengagements over that whole population, μ. It is a fixed number, and it is unknown.

Sample.

The 12 test drives, giving observations $x_1, x_2, \ldots, x_{12}$. Before the drives, the distances $X_1, \ldots, X_{12}$ are random variables; we treat them as a random sample if the drives are independent and representative of fleet driving.

Statistic and estimate.

The statistic is the sample mean $\bar{X}$, a rule for combining the observations. The estimate is its observed value, $\boxed{\bar{x} = 158.4 \text{ km}}$. A second set of 12 drives would give a different estimate; μ itself would not change.

Reference. Montgomery & Runger, 6th ed., Section 7.1.

Statistic. Any function of the observations in a random sample. A statistic is itself a random variable.

Point estimator. A statistic $\hat{\Theta}$ chosen to estimate an unknown parameter θ.

Point estimate. The observed numerical value $\hat{\theta}$ of a point estimator, computed from one sample.

2.2 *Point Estimators and Point Estimates*

A *statistic* is any function of the sample observations: the sample mean, the sample variance, the largest observation, the mode. Because the observations are random variables, every statistic is a random variable too. This single sentence carries most of the chapter, so we repeat it plainly: a statistic has a probability distribution of its own, because different samples produce different values of it.

When a statistic is used to estimate an unknown parameter θ, we call it a *point estimator* and write it $\hat{\Theta}$ (“theta hat”). The number it produces from one particular sample is the *point estimate* $\hat{\theta}$. In general, a point estimator is a rule

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n), \qquad \hat{\theta} = h(x_1, x_2, \ldots, x_n), \tag{2.1}$$

where h is the formula that turns the sample into a single number. The estimator is the rule; the estimate is the output of the rule on one data set. For example, $\hat{\mu} = \bar{X}$ uses the sample mean to estimate the population mean, and $\hat{\sigma}^2 = S^2$ uses the sample variance to estimate the population variance.

Table 2.1 lists the parameters this course cares about and the standard estimator for each. All of them are functions of a random sample, so all of them are covered by the notation in (2.1).

Startup lens. Every number on a startup dashboard—average session length, conversion rate, churn—is a point estimate, not a parameter. When the dashboard moves from 4.1% to 4.4% overnight, the first question is not “what did we change?” but “is this movement larger than the estimator’s own sample-to-sample variation?” This chapter gives the tools to answer that.

Parameter	Meaning	Standard point estimator
μ	population mean	$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
σ^2	population variance	$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
p	population proportion	$\hat{P} = \frac{X}{n}$ (X = number of successes)
$\mu_1 - \mu_2$	difference of two means	$\bar{X}_1 - \bar{X}_2$
$p_1 - p_2$	difference of two proportions	$\hat{P}_1 - \hat{P}_2$

Table 2.1. The parameters studied in this course and their standard point estimators. Each estimator is a statistic, so each has a sampling distribution.

Reference. Montgomery & Runger, 6th ed., Section 7.2.

2.3 *Sampling Distributions*

Sampling distribution. The probability distribution of a statistic over all possible samples of size n.

Suppose the whole class draws samples of size $n = 10$ from the same population and each student computes $\bar{x}$. The values will not agree. This variation is not error in the arithmetic; it is the natural behavior of a statistic. The probability distribution that describes this sample-to-sample variation is called the *sampling distribution* of the statistic.

Please pay attention here, because this is the most common confusion in the whole chapter: the sampling distribution is the distribution of the *statistic*, not the distribution of the *population*. The population distribution is fixed; it describes single observations X_i. The sampling distribution describes $\bar{X}$ (or S^2, or $\hat{P}$), and its shape depends on the statistic's formula and on the sample size n.

Insight. Why does a statistic have a distribution at all? Because it is built from random inputs. The formula $\bar{X} = (X_1 + \cdots + X_n)/n$ is fixed, but its inputs are random variables, so its output is a random variable. A fixed function of random quantities is random. Everything in this chapter—bias, standard error, the Central Limit Theorem—is a statement about this induced randomness.

A concrete case makes the distinction visible. Let the population be $X_i \sim \text{Uniform}(0, 1)$. From ISE 205, this population has mean $\mu = (0 + 1)/2 = 0.5$ and variance $\sigma^2 = (1 - 0)^2/12 = 1/12$. The population density is flat, and it never

changes. Now take samples of size n and compute $\bar{X}$ each time. Figure 2.2 shows what happens: the sampling distribution of $\bar{X}$ is centered at the same value 0.5, but it is not flat—it is bell-shaped, and it becomes narrower as n grows. Same center, different shape, shrinking spread. The next two sections quantify each of these three observations.

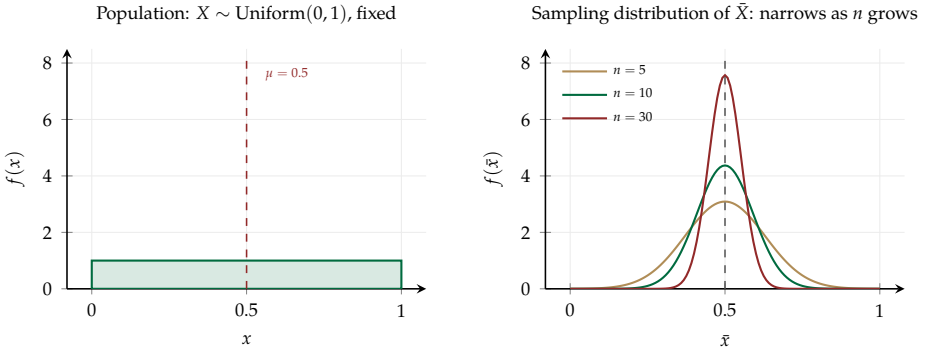

Figure 2.2. Population distribution versus sampling distribution for $X \sim \text{Uniform}(0,1)$. The population density (left) is flat and never changes. The sampling distribution of $\bar{X}$ (right) is centered at the same $\mu = 0.5$ but is bell-shaped, with spread σ^2/n that shrinks as n increases.

Engineering judgment. The population variance σ^2 is often supplied by engineering knowledge, not by the current data: instrument calibration records, process capability studies, historical operations logs. When a problem says "σ is known," this is where the number comes from.

Reference. Montgomery & Runger, 6th ed., Sections 7.3.1 and 7.3.3.

2.4 *Unbiasedness and the Standard Error of $\bar{X}$*

We now quantify the two most important properties of the sampling distribution of $\bar{X}$: where it is centered, and how spread out it is. Both results follow from the rules for expectation and variance of sums of independent random variables, which you saw in ISE 205 and which we will reuse throughout the course.

2.4.1 *The center: $\bar{X}$ is unbiased*

Take a random sample $X_1, \ldots, X_n$ from any population with mean μ and variance σ^2. The expected value of the sample mean is

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\mu) = \mu. \tag{2.2}$$

Bias. The systematic offset of an estimator: $\text{Bias}(\hat{\Theta}) = E[\hat{\Theta}] - \theta$.

Unbiased estimator. An estimator whose bias is zero: $E[\hat{\Theta}] = \theta$.

The first step pulls the constant $1/n$ out of the expectation; the second uses linearity of expectation, which requires no independence at all; the third uses

the fact that every X_i has the same mean μ. So the sampling distribution of $\bar{X}$ is centered exactly on the parameter it estimates, for every sample size.

This property has a name. The *bias* of an estimator $\hat{\Theta}$ for a parameter θ is

$$\text{Bias}(\hat{\Theta}) = E[\hat{\Theta}] - \theta, \tag{2.3}$$

and an estimator is *unbiased* when its bias equals zero. Equation (2.2) says that $\bar{X}$ is an unbiased estimator of μ: it has no systematic tendency to overshoot or undershoot. Unbiasedness does not say any single estimate equals μ; it says the estimator is correct *on average* across all possible samples.

Safety lens. Bias direction matters in safety work. An estimator that systematically understates a perception model's failure rate is more dangerous than one that overstates it, even at the same magnitude of bias, because it leads to shipping a system that looks safer than it is. Always ask not only "how large is the bias?" but "which way does it point?"

2.4.2 *The spread: variance and standard error*

For the spread we do need independence. For independent random variables, variances add, and constants come out squared:

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}. \tag{2.4}$$

The variance of the sample mean is the population variance divided by n. Averaging cancels part of the randomness: individual observations that come out high are offset by others that come out low, and the more observations we average, the more complete the cancellation.

Standard error. The standard deviation of an estimator's sampling distribution. For $\bar{X}$ it is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

The standard deviation of a sampling distribution is used so often that it has its own name: the *standard error*. For the sample mean,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}, \qquad \text{estimated by} \qquad \widehat{\text{se}}(\bar{X}) = \frac{s}{\sqrt{n}} \tag{2.5}$$

when σ is unknown and must be replaced by the sample standard deviation s. The standard error is the single most used number in the rest of this course: confidence-interval widths and test statistics are all built from it.

Two consequences of (2.5) deserve emphasis. First, the standard error shrinks like $1/\sqrt{n}$, not like $1/n$ (Figure 2.3). To cut the standard error in half you must collect *four* times as much data; to cut it to one tenth you need one hundred times as much. Data buys precision at a square-root exchange rate. Second, the standard error depends on the population spread σ: a noisier population needs more data to reach the same precision.

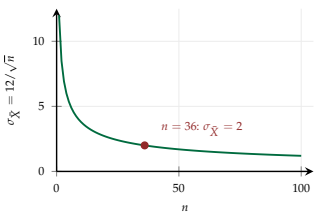

Figure 2.3. The standard error of $\bar{X}$ for $\sigma = 12$. Doubling precision requires quadrupling the sample size.

For the uniform population of Section 2.3 with $n = 10$, these formulas give concrete numbers: $\mu_{\bar{X}} = \mu = 0.5$ by (2.2); $\mathrm{Var}(\bar{X}) = (1/12)/10 = 1/120 \approx 0.008333$ by (2.4); and $\sigma_{\bar{X}} = \sqrt{1/120} \approx 0.09129$ by (2.5). These are exactly the center and spread of the curves in Figure 2.2. What the formulas do *not* yet give is the bell shape of those curves; that is the job of the Central Limit Theorem in Section 2.6.

Reference. Montgomery & Runger, 6th ed., Section 7.3.4.

2.5 *Mean Squared Error: Combining Bias and Variance*

Unbiasedness and low variance are both desirable, but they can pull in opposite directions: one estimator may be perfectly centered but widely spread, while another is slightly off-center but tightly concentrated. To compare such estimators we need one number that accounts for both effects. That number is the *mean squared error*.

Mean squared error. The expected squared distance between an estimator and the true parameter: $\mathrm{MSE}(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2]$.

The mean squared error of an estimator $\hat{\Theta}$ for θ is the expected squared distance between the two, and it decomposes exactly into a variance part and a bias part:

$$\mathrm{MSE}(\hat{\Theta}) = E\Big[\big(\hat{\Theta} - \theta\big)^2\Big] = \mathrm{Var}(\hat{\Theta}) + \big[\mathrm{Bias}(\hat{\Theta})\big]^2. \tag{2.6}$$

The decomposition is worth deriving once, because it explains why bias enters *squared*. Write $m = E[\hat{\Theta}]$ for the estimator's own mean, and add and subtract m inside the square:

$$\begin{aligned} E\Big[\big(\hat{\Theta} - \theta\big)^2\Big] &= E\Big[\big((\hat{\Theta} - m) + (m - \theta)\big)^2\Big] \\ &= E\Big[(\hat{\Theta} - m)^2\Big] + 2\,(m - \theta)\,E\big[\hat{\Theta} - m\big] + (m - \theta)^2. \end{aligned}$$

The first term is $\mathrm{Var}(\hat{\Theta})$ by definition. The middle term vanishes because $E[\hat{\Theta} - m] = m - m = 0$. The last term is the squared bias. Nothing is lost and nothing is approximated: MSE is *exactly* variance plus bias squared.

> **Insight.** The decomposition (2.6) says an estimator can afford a small bias if it buys a large reduction in variance. Shrinking an estimator toward zero, as in $X^* = 0.7\bar{X}$, introduces bias -0.3μ but multiplies the variance by 0.49. Whether the trade is worth it depends on how large μ really is—which is exactly the thing we do not know.

This tension between bias and variance reappears throughout statistics and machine learning.

Figure 2.4 shows the trade-off for two estimators of the same parameter $\theta = 10$. Estimator A is unbiased with variance 1, so $\mathrm{MSE}(A) = 1$. Estimator B is centered at 9.6, so its bias is -0.4, but its variance is only 0.25; its MSE is $0.25 + (-0.4)^2 = 0.41$. Even though B is systematically off-center, a value drawn from B's sampling distribution is, on average, closer to θ than a value drawn from A's.

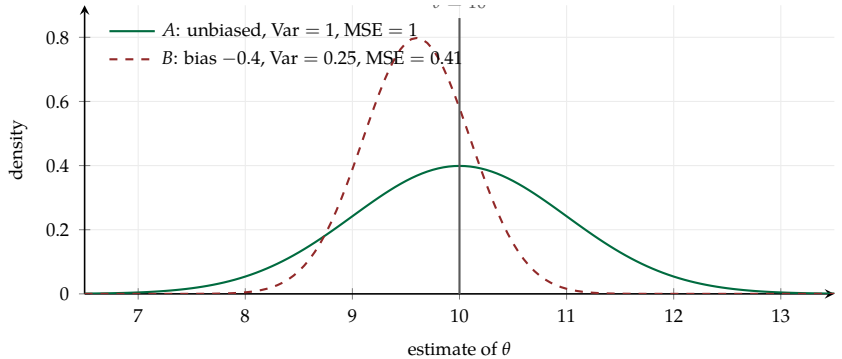

Figure 2.4. Two estimators of the same parameter $\theta = 10$. Estimator A (solid) is unbiased but spread out. Estimator B (dashed) is biased low by 0.4 but tightly concentrated, and its mean squared error is less than half of A's.

Judging an estimator is a routine we will use many times, so we record it as a procedure.

Procedure 2.1 — Judging the quality of an estimator.

1. **Compute the mean.** Find $E[\hat{\Theta}]$ using linearity of expectation, and compare it with the target θ. The difference is the bias (2.3). State whether the estimator is unbiased and, if not, which direction the bias points.
2. **Compute the variance.** Find $\mathrm{Var}(\hat{\Theta})$ using the rules for variances of independent sums; take the square root if the standard error is requested. Check how the variance behaves as n grows—a good estimator's variance shrinks toward zero.

3. **Combine into MSE.** Compute $\text{MSE} = \text{Var} + \text{Bias}^2$ by (2.6).
4. **Compare.** Between competing estimators, prefer the one with the smaller MSE. If the MSEs depend on the unknown θ, find the range of θ over which each estimator wins, and say so explicitly.

Example 2.2 — Four estimators for one startup metric

Your startup tracks daily engagement: the number of minutes a user spends in the product per day. Engagement minutes have unknown population mean μ and known standard deviation $\sigma = 10$ minutes. From a random sample of $n = 25$ users, four candidate estimators of μ are proposed:

- the sample mean $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$;
- the half-range $\hat{X} = \left(\max_i X_i + \min_i X_i\right)/2$;
- the first observation $X^{\dagger} = X_1$;
- the shrinkage estimator $X^{*} = 0.7\,\bar{X}$.

Which are unbiased? Compute the variance and MSE of $\bar{X}$, $X^{\dagger}$, and X^{*}, and decide which estimator you would use.

Solution.

Step 1 — Bias.

By (2.2), $E[\bar{X}] = \mu$, so $\bar{X}$ is unbiased. The single observation satisfies $E[X^{\dagger}] = E[X_1] = \mu$, so it is also unbiased. For the shrinkage estimator, $E[X^{*}] = 0.7\,E[\bar{X}] = 0.7\mu$, so its bias is $0.7\mu - \mu = -0.3\mu$: it systematically undershoots whenever $\mu > 0$. The half-range is unbiased when the population distribution is symmetric, but for skewed engagement data (a few very heavy users) it is biased; its exact behavior depends on the population shape.

Step 2 — Variance.

By (2.4), $\text{Var}(\bar{X}) = \sigma^2/n = 100/25 = 4.000$. The first observation ignores 24 of the 25 data points: $\text{Var}(X^{\dagger}) = \sigma^2 = 100.0$, and this never shrinks as n grows. For the shrinkage estimator, $\text{Var}(X^{*}) = 0.7^2\,\text{Var}(\bar{X}) = 0.49 \times 4.000 = 1.960$.

***Step 3 — MSE by* (2.6).**
$\mathrm{MSE}(\bar{X}) = 4 + 0 = 4.000$; $\mathrm{MSE}(X^{\dagger}) = 100 + 0 = 100.0$; $\mathrm{MSE}(X^{*}) = 1.960 + (0.3\mu)^2 = 1.960 + 0.09\mu^2$.

***Step 4 — Compare* (*Procedure 2.1*).**
$X^{\dagger}$ is unbiased but hopeless: its MSE is 25 times that of $\bar{X}$. The shrinkage estimator beats $\bar{X}$ only when $1.960 + 0.09\mu^2 < 4$, that is when $\mu^2 < 22.67$, or $|\mu| < 4.761$ minutes. Any realistic engagement product has mean session time well above 5 minutes, so the shrinkage estimator's bias term dominates: at $\mu = 18$, its MSE is $1.960 + 0.09(324) = 31.12$, nearly eight times worse than $\bar{X}$. The answer is $\boxed{\text{use } \bar{X}}$: unbiased, and by far the lowest MSE in the realistic range of μ.

Figure 2.5 draws the sampling distributions of three of these estimators for the case $\mu = 10$. Study it against the numbers in Example 2.2: the first-observation estimator is centered correctly but far too wide, and the shrinkage estimator is narrow but centered at 7 instead of 10.

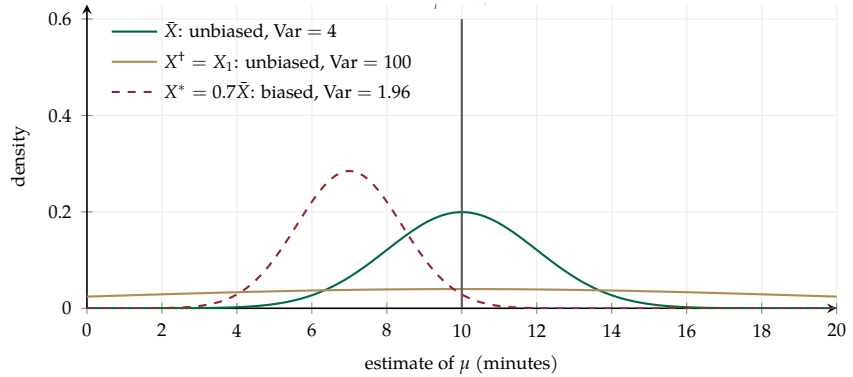

Figure 2.5. Sampling distributions of three estimators of the mean engagement time when $\mu = 10$, $\sigma = 10$, $n = 25$ (Example 2.2). The sample mean is both centered on μ and concentrated; the single observation is centered but wide; the shrinkage estimator is concentrated but centered at $0.7\mu = 7$.

Reference. Montgomery & Runger, 6th ed., Section 7.2.

2.6 *The Central Limit Theorem*

Equations (2.2) and (2.4) give the center and spread of the sampling distribution of $\bar{X}$, but not its shape. The shape is supplied by two results, one exact and one approximate.

2.6.1 *Exact result for normal populations*

If the population itself is normal, then $\bar{X}$ is a sum of independent normal random variables divided by a constant, and such a combination is exactly normal. Therefore, for a random sample of size n from $N(\mu, \sigma^2)$,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{exactly, for every } n \geq 1. \tag{2.7}$$

Please pay attention to the second argument, because this is a common trap: throughout this book, $N(\mu, \sigma^2)$ lists the *variance* in the second slot, following Montgomery and Runger. The standard deviation of the distribution in (2.7) is $\sigma/\sqrt{n}$, not σ^2/n. Always check which one a formula needs before substituting.

Central Limit Theorem. For large n, the sampling distribution of $\bar{X}$ is approximately normal, whatever the shape of the population.

2.6.2 *Approximate result for all populations*

The remarkable fact is that near-normality does not require a normal population. The *Central Limit Theorem* (CLT) states: if $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population with mean μ and finite variance σ^2, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad \text{is the standard normal distribution } N(0,1) \text{ as } n \to \infty. \tag{2.8}$$

In practice, for $n \geq 30$ the approximation $\bar{X} \sim N(\mu, \sigma^2/n)$ is good for almost any population shape, and for mildly non-normal populations it is already usable at much smaller n. If the population is normal, no approximation is needed at all, by (2.7).

Figure 2.6 shows the theorem acting on a strongly skewed population: an exponential distribution with mean $\mu = 1$. For $n = 1$ the sampling distribution of $\bar{X}$ *is* the population, with its sharp peak at zero and long right tail. At $n = 5$ the distribution of $\bar{X}$ has already pulled into a lopsided bell. At $n = 30$ it is nearly indistinguishable from the normal curve with the same mean μ and variance σ^2/n. The skewness dies out as the averaging cancels extreme values against ordinary ones.

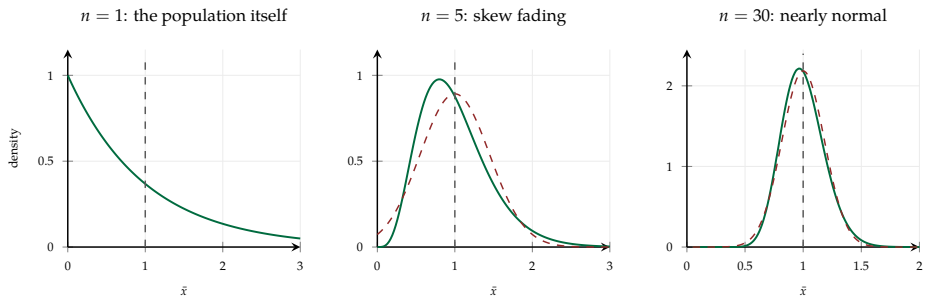

Figure 2.6. The Central Limit Theorem acting on a strongly right-skewed population (exponential with $\mu = 1$; dashed gray line at μ). Solid green: the exact sampling distribution of $\bar{X}$ for $n = 1, 5, 30$. Dashed maroon: the normal approximation $N(\mu, \sigma^2/n)$ from the CLT. By $n = 30$ the two curves nearly coincide.

2.6.3 *Probability computations for $\bar{X}$*

The CLT converts questions about $\bar{X}$ into standard normal table lookups. The routine below is the workhorse of this chapter; every problem in Lecture 3 and Lecture 4 is an instance of it, with only the statistic changing.

Procedure 2.2 — Probability computation for $\bar{X}$ via the CLT.

1. **Find the population parameters.** Identify μ and σ^2 of the population distribution, either given directly or computed from a known distribution family (e.g., Uniform(a, b) has $\mu = (a+b)/2$ and $\sigma^2 = (b-a)^2/12$).
2. **Write the sampling distribution.** By the CLT (2.8)—or exactly, by (2.7), if the population is normal—the sample mean satisfies $\bar{X} \mathrel{\dot\sim} N(\mu, \sigma^2/n)$ with standard error $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ from (2.5).
3. **Standardize.** Convert the event about $\bar{X}$ into an event about $Z = (\bar{X} - \mu)/\sigma_{\bar{X}}$. Every boundary value c maps to $z = (c - \mu)/\sigma_{\bar{X}}$.
4. **Read the table.** Use the standard normal table to find the probability, remembering that the table gives $\Phi(z) = P(Z \le z)$; use $P(Z > z) = 1 - \Phi(z)$ and $P(z_1 \le Z \le z_2) = \Phi(z_2) - \Phi(z_1)$.

Startup lens. A/B testing platforms lean on the CLT. Revenue per user is wildly skewed—most users pay nothing, a few pay a lot—yet the *average* revenue across thousands of users behaves like a normal random variable, which is what makes the platform's statistics valid.

Safety lens. The CLT is a statement about the *average*, not about the tails of individual observations. Averaging $n = 50$ sensor-fusion errors tells you about the mean error; it does not make the worst-case error normal, and worst cases are often what safety work cares about. Use the CLT for inference on means, not as a claim that the underlying data are normal.

5. **Answer in context.** State the probability and what it says about the engineering question, with units.

Example 2.3 — Mean of n LLM evaluation scores

An evaluation harness scores an LLM's responses on a safety rubric scaled to $[0, 1]$. Suppose individual scores behave like $X_i \sim \text{Uniform}(0, 1)$—maximally spread out, and certainly not normal. An audit run averages $n = 10$ independent scores. Find the approximate probability that the run's average score exceeds 0.6.

Solution.

Step 1 — Population parameters.

For Uniform$(0, 1)$: $\mu = 0.5$ and $\sigma^2 = 1/12 \approx 0.08333$.

Step 2 — Sampling distribution.

With $n = 10$, $\mu_{\bar{X}} = 0.5$ and, by (2.4), $\sigma^2_{\bar{X}} = (1/12)/10 = 1/120 \approx 0.008333$, so $\sigma_{\bar{X}} = 0.09129$. The uniform distribution is symmetric with short tails, so the CLT approximation $\bar{X} \dot{\sim} N(0.5,\ 1/120)$ is already good at $n = 10$.

Step 3 — Standardize.

$$P(\bar{X} > 0.6) = P\left(Z > \frac{0.6 - 0.5}{0.09129}\right) = P(Z > 1.10).$$

Step 4 — Read the table.

Table 2.2 gives $\Phi(1.10) = 0.8643$, so $P(Z > 1.10) = 1 - 0.8643 = 0.1357$.

z	0.00	0.01	0.02
1.0	0.8413	0.8438	0.8461
1.1	0.8643	0.8665	0.8686
1.2	0.8849	0.8869	0.8888

Table 2.2. Standard normal table excerpt: $\Phi(1.10) = 0.8643$.

Step 5 — Answer in context.

Even though a *single* score is above 0.6 with probability 0.40, the *average* of ten scores exceeds 0.6 with probability only $\boxed{P(\bar{X} > 0.6) \approx 0.1357}$. Averag-

ing concentrates the distribution around $\mu = 0.5$, so values far from the mean become much less likely for $\bar{X}$ than for any one X_i.

Three questions in this style will be asked again and again in this course. Given a population, what is the sampling distribution of $\bar{X}$ for sample size n? Given a population, what is the probability that $\bar{X}$ falls above, below, or between stated values? And given a target standard error, what is the smallest sample size n that achieves it? Procedure 2.2 answers the first two; the third inverts equation (2.5), and Example 2.6 will show it worked out. The same recipe also covers statistics other than $\bar{X}$: a shifted mean $\bar{X} - c$, a scaled mean $a(\bar{X} - c)$ (the Z-statistic itself is one), the sample proportion $\hat{P}$ (Section 2.8), and differences $\bar{X}_1 - \bar{X}_2$ and $\hat{P}_1 - \hat{P}_2$ (Sections 2.9 and 2.10). In every case the path is the same: population parameters → mean and standard error of the statistic → normal approximation → standardize → table.

Lecture 3 starts here — *Practice with estimators*

Lecture 3. This section is covered by Lecture 3.

Reference. Montgomery & Runger, 6th ed., Sections 7.2–7.3 and supplemental exercises.

2.7 *Practice with Estimators*

This section does what a practice lecture does: it works through the standard problem types one at a time, slowly, using Procedures 2.1 and 2.2. Each example is a type you should be able to solve on an exam without notes: a probability for $\bar{X}$ from a normal population; a bias–variance–MSE comparison of two estimators; and a sample-size determination. Work each one yourself before reading the solution.

Example 2.4 — Container dwell time: probability for $\bar{X}$

Container dwell times at a port terminal are normally distributed with mean $\mu = 50$ hours and standard deviation $\sigma = 12$ hours. A random sample of $n = 36$ containers is drawn from the terminal's records. Find the probability that the sample mean dwell time falls in the interval $47 \leq \bar{X} \leq 53$ hours. Is the assumption of normality important here? Why?

Solution.

Step 1 — Population parameters.
Given directly: $\mu = 50$, $\sigma = 12$, and the population is normal.

Step 2 — Sampling distribution.
By (2.7), $\bar{X}$ is *exactly* normal with

$$\mu_{\bar{X}} = \mu = 50, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = 2.000 \text{ hours}.$$

Step 3 — Standardize.

$$P(47 \le \bar{X} \le 53) = P\left(\frac{47-50}{2} \le Z \le \frac{53-50}{2}\right) = P(-1.50 \le Z \le 1.50).$$

Step 4 — Read the table.
Table 2.3 gives $\Phi(1.50) = 0.9332$; by symmetry $\Phi(-1.50) = 1 - 0.9332 = 0.0668$.

z	0.00	0.01	0.02
1.4	0.9192	0.9207	0.9222
1.5	0.9332	0.9345	0.9357
1.6	0.9452	0.9463	0.9474

Table 2.3. Standard normal table excerpt: $\Phi(1.50) = 0.9332$.

$$P(-1.50 \le Z \le 1.50) = 0.9332 - 0.0668 = 0.8664.$$

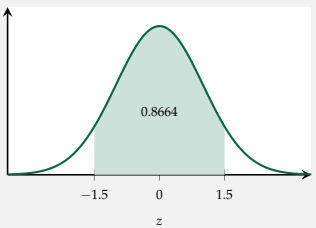

Figure 2.7. The shaded area $P(-1.50 \le Z \le 1.50) = 0.8664$ for Example 2.4.

Step 5 — Answer and discussion.
The probability is $\boxed{P(47 \le \bar{X} \le 53) = 0.8664}$. The normality assumption is *not* essential: even if dwell times were not normal, $n = 36 \ge 30$ means the

CLT (2.8) would make $\bar{X}$ approximately $N(50, 4)$ anyway, and the computation would be unchanged. Normality of the population makes the answer exact rather than approximate.

Example 2.5 — Shrinkage estimator for sensor-fusion error

An AV team estimates the mean lateral position error μ (in centimeters) of its sensor-fusion stack. Errors are normal with $\sigma^2 = 9$, and the test budget allows $n = 9$ runs. Two estimators are considered: $\hat{\Theta}_1 = \bar{X}$ and $\hat{\Theta}_2 = \frac{n}{n+1}\bar{X} = \frac{9}{10}\bar{X}$. Compute the bias, variance, and MSE of each. If $\mu = 1$ cm, which is better? If $\mu = 10$ cm? Find the exact cutoff.

Solution.

Step 1 — The unbiased estimator $\hat{\Theta}_1 = \bar{X}$.

Bias: $E[\bar{X}] - \mu = 0$ by (2.2). Variance: by (2.4), $\mathrm{Var}(\bar{X}) = \sigma^2/n = 9/9 = 1.000$. MSE: $\mathrm{MSE}(\hat{\Theta}_1) = 1.000 + 0^2 = 1.000$.

Step 2 — The shrinkage estimator $\hat{\Theta}_2 = \frac{9}{10}\bar{X}$.

Mean: $E[\hat{\Theta}_2] = \frac{9}{10}E[\bar{X}] = \frac{9}{10}\mu$, so by (2.3), $\mathrm{Bias}(\hat{\Theta}_2) = \frac{9}{10}\mu - \mu = -\mu/10$ (biased low). Variance: $\mathrm{Var}(\hat{\Theta}_2) = \left(\frac{9}{10}\right)^2 \mathrm{Var}(\bar{X}) = 0.81 \times 1.000 = 0.8100$. MSE by (2.6):

$$\mathrm{MSE}(\hat{\Theta}_2) = 0.8100 + \left(-\frac{\mu}{10}\right)^2 = 0.8100 + \frac{\mu^2}{100}.$$

Step 3 — Compare at the two stated values.

At $\mu = 1$: $\mathrm{MSE}(\hat{\Theta}_2) = 0.8100 + 0.0100 = 0.8200 < 1.000$, so the shrinkage estimator wins. At $\mu = 10$: $\mathrm{MSE}(\hat{\Theta}_2) = 0.8100 + 1.000 = 1.810 > 1.000$, so $\bar{X}$ wins.

Step 4 — Find the cutoff.

$$\begin{aligned} 0.8100 + \frac{\mu^2}{100} < 1.000 &\iff \frac{\mu^2}{100} < 0.1900 \\ &\iff \mu^2 < 19.00 \iff |\mu| < 4.359. \end{aligned}$$

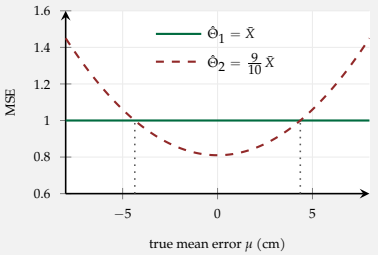

Figure 2.8. MSE of the two estimators in Example 2.5 as a function of the unknown μ. The curves cross at $|\mu| = 4.359$ cm.

The shrinkage estimator is better precisely when $\boxed{|\mu| < 4.359 \text{ cm}}$ (Figure 2.8).

Step 5 — Engineering reading.

The comparison depends on the very parameter we are estimating. If the team is confident from design analysis that the true error is small ($|\mu|$ under about 4 cm), shrinking toward zero pays off. If the error could be large, the biased estimator would understate it—exactly the failure mode a safety validation must avoid, so the conservative choice is $\bar{X}$.

Example 2.6 — Sample size for a pilot-response study

Your startup measures the time for pilot customers to respond to an onboarding email. From an earlier cohort, $\sigma = 8$ hours. How many pilot customers must be sampled so that the sample mean response time falls within 2 hours of the true mean μ with probability 0.95?

Solution.

Step 1 — Translate the requirement.

"Within 2 hours with probability 0.95" means $P\big(|\bar{X} - \mu| \le 2\big) = 0.95$.

Step 2 — Standardize.

Dividing through by $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ from (2.5) and applying the CLT (2.8),

$$P\left(|Z| \le \frac{2}{8/\sqrt{n}}\right) = 0.95.$$

From the standard normal table, $P(|Z| \le z) = 0.95$ requires $z = 1.96$ (because

$\Phi(1.96) = 0.9750$, leaving 0.025 in each tail).

Step 3 — Solve for n.

Setting the standardized half-width equal to 1.96 and solving,

$$\frac{E}{\sigma/\sqrt{n}} = z \quad \Longrightarrow \quad n = \left(\frac{z\,\sigma}{E}\right)^2, \tag{2.9}$$

where E is the required margin. Here $n = \left(1.96 \times 8/2\right)^2 = (7.840)^2 = 61.47$.

Step 4 — Round up.

Sample size must be an integer, and rounding *down* would miss the target precision, so take $\boxed{n = 62}$ pilot customers. Chapter 3 returns to formula (2.9) as part of confidence-interval design.

Reference. Montgomery & Runger, 6th ed., Sections 7.2 and 7.3.

Sample proportion. $\hat{P} = X/n$, the fraction of successes in a sample of n binary trials.

2.8 The Sample Proportion

Many engineering measurements are binary: a container arrives late or on time, a signup converts or does not, a guardrail flags a prompt or lets it pass. The parameter of interest is the population proportion p, and the natural estimator is the *sample proportion*

$$\hat{P} \;=\; \frac{X_1 + X_2 + \cdots + X_n}{n} \;=\; \frac{X}{n}, \tag{2.10}$$

where each $X_i \sim \text{Bernoulli}(p)$ equals 1 for a success and 0 for a failure, and $X = \sum_i X_i$ counts the successes. Look carefully at (2.10): $\hat{P}$ *is* a sample mean. It is the average of n binary random variables. Everything we established for $\bar{X}$ therefore applies at once; we only need the Bernoulli mean and variance, $E[X_i] = p$ and $\text{Var}(X_i) = p(1-p)$, from ISE 205.

Applying (2.2) and (2.4) with $\mu = p$ and $\sigma^2 = p(1-p)$:

$$E\big[\hat{P}\big] = p \qquad \text{(unbiased)}, \qquad \text{Var}\big(\hat{P}\big) = \frac{p(1-p)}{n},$$

so the standard error of the sample proportion is

$$\sigma_{\hat{P}} \;=\; \sqrt{\frac{p(1-p)}{n}}, \qquad \text{estimated by} \qquad \widehat{\text{se}}\big(\hat{P}\big) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.11}$$

when p is unknown. By the CLT (2.8), for large n,

$$\hat{P} \;\dot{\sim}\; N\!\left(p,\; \frac{p(1-p)}{n}\right). \tag{2.12}$$

The same result can be reached through the binomial distribution: since $X \sim \text{Binomial}(n, p)$ with $E[X] = np$ and $\text{Var}(X) = np(1-p)$, the scaled statistic $\hat{P} = X/n$ has $E[\hat{P}] = np/n = p$ and $\text{Var}(\hat{P}) = np(1-p)/n^2 = p(1-p)/n$—the same answers by a different route. A standard check before using the normal approximation (2.12) is $np \geq 10$ and $n(1-p) \geq 10$; if either count of expected successes or failures is smaller, the binomial distribution should be used directly.

Insight. The variance $p(1-p)$ is largest at $p = 0.5$ and shrinks toward zero as p approaches 0 or 1. A proportion near a boundary is easier to pin down—but the normal approximation also breaks down there, because np or $n(1-p)$ falls below 10. Rare-event proportions (defect rates, safety failure rates) are exactly the ones that need care.

Example 2.7 — Signup conversion rate

Your startup's landing page historically converts $p = 0.20$ of visitors into signups. You sample $n = 100$ independent visitors and compute the sample conversion rate $\hat{P}$. State the approximate sampling distribution of $\hat{P}$, then find the probability that $\hat{P}$ falls within 0.05 of the true rate.

Solution.

Step 1 — Population distribution.

Each visitor is a Bernoulli trial: $X_i \sim \text{Bernoulli}(0.20)$ with $E[X_i] = 0.20$ and $\text{Var}(X_i) = 0.20 \times 0.80 = 0.1600$.

Step 2 — Sampling distribution.

Check the approximation conditions: $np = 20 \geq 10$ and $n(1-p) = 80 \geq 10$. By (2.11),

$$\sigma_{\hat{p}} = \sqrt{\frac{0.20 \times 0.80}{100}} = \sqrt{0.001600} = 0.04000,$$

so by (2.12), $\hat{P} \dot{\sim} N(0.20,\ 0.0016)$.

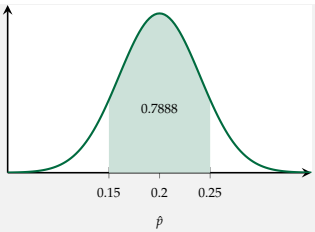

Figure 2.9. Sampling distribution of $\hat{P}$ in Example 2.7, with the event $|\hat{P} - p| \le 0.05$ shaded.

Step 3 — Standardize.

$$P\big(|\hat{P} - 0.20| \le 0.05\big) = P\left(|Z| \le \frac{0.05}{0.04000}\right) = P\big(|Z| \le 1.25\big).$$

Step 4 — Read the table.

$\Phi(1.25) = 0.8944$, so $P(|Z| \le 1.25) = 0.8944 - (1 - 0.8944) = 0.8944 - 0.1056 = 0.7888$.

Step 5 — Answer in context.

With 100 visitors, $\boxed{P\big(|\hat{P} - p| \le 0.05\big) \approx 0.7888}$: there is roughly a 21% chance the measured conversion rate misses the true rate by more than five percentage points. A dashboard based on 100 visitors is a rough instrument; Example 2.6's method shows how to size the sample for tighter precision.

Startup lens. Before celebrating that a new landing page "lifted conversion from 20% to 23%" on 100 visitors, note that Example 2.7 puts a one-standard-error band of ± 4 percentage points around the estimate. The observed lift is smaller than one standard error—it is indistinguishable from sampling noise.

Lecture 4. This section is covered by Lecture 4.

Reference. Montgomery & Runger, 6th ed., Section 7.2.

Lecture 4 starts here — *Estimators for multiple populations*

2.9 Estimators for Differences of Two Means

Engineering questions are often comparative: which supplier is faster, which model version is safer, which pricing page converts better. The parameter is then a *difference*, such as $\Delta\mu = \mu_1 - \mu_2$, and the natural estimator is the difference of the sample statistics, $\bar{X}_1 - \bar{X}_2$. Nothing new is needed to analyze it—only the expectation and variance rules from ISE 205, which we restate because every formula in this section follows from them. For *independent* random variables X and Y and constants a:

$$E[X \pm Y] = E[X] \pm E[Y], \qquad E[aX] = a\,E[X], \qquad E[X \pm a] = E[X] \pm a,$$

$$\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y),$$
$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X), \qquad \operatorname{Var}(X \pm a) = \operatorname{Var}(X).$$

Please pay attention to the first variance rule, because it is a common trap: for independent quantities, variances *add* even when the random variables are *subtracted*. Uncertainty accumulates under both addition and subtraction; it never cancels.

Insight. Why do variances add under subtraction? Because $\operatorname{Var}(-Y) = (-1)^2 \operatorname{Var}(Y) = \operatorname{Var}(Y)$: flipping the sign of a random variable flips its values around zero but leaves its spread unchanged. So $\operatorname{Var}(X - Y) = \operatorname{Var}(X) + \operatorname{Var}(-Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$. Each sample contributes its own noise to the comparison, and noise from one sample cannot erase noise from the other.

Now set up the two-population problem. Population 1 has mean μ_1 and variance σ_1^2, sampled with n_1 observations giving $\bar{X}_1$; population 2 has mean μ_2 and variance σ_2^2, sampled independently with n_2 observations giving $\bar{X}_2$. Applying the rules above together with (2.2) and (2.4):

$$E\left[\bar{X}_1 - \bar{X}_2\right] = \mu_1 - \mu_2 \qquad \text{(unbiased)}, \qquad \operatorname{Var}\left(\bar{X}_1 - \bar{X}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Since each sample mean is (approximately) normal by the CLT, and independent normals combine into a normal, the sampling distribution of the difference is

$$\bar{X}_1 - \bar{X}_2 \;\dot{\sim}\; N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right). \tag{2.13}$$

The pattern generalizes beyond means. For *any* two independent estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$, the standard error of their difference is

$$\operatorname{se}\left(\hat{\Theta}_1 - \hat{\Theta}_2\right) = \sqrt{\operatorname{se}\left(\hat{\Theta}_1\right)^2 + \operatorname{se}\left(\hat{\Theta}_2\right)^2}, \tag{2.14}$$

the square root of the *sum* of squared standard errors—never the difference. Equation (2.14) will be used for means here, for proportions in Section 2.10, and repeatedly in Chapter 5.

Procedure 2.3 — Probability computation for a difference of two independent estimators.

1. **Parameters of each population.** List μ_1, σ_1^2, n_1 and μ_2, σ_2^2, n_2 (or p_1, n_1 and p_2, n_2 for proportions). Confirm the two samples are independent.
2. **Mean of the difference.** The estimator $\hat{\Theta}_1 - \hat{\Theta}_2$ is unbiased for the parameter difference: its mean is $\mu_1 - \mu_2$ (or $p_1 - p_2$).
3. **Standard error of the difference.** Compute each estimator's standard error, then combine with (2.14): square, add, take the square root.
4. **Normal approximation and standardize.** Write the sampling distribution as in (2.13), convert the event to a Z event, and read the standard normal table.

Example 2.8 — Fuel use on two delivery route plans

A logistics company compares two routing plans for its truck fleet. Plan A is observed on $n_1 = 50$ delivery runs, with fuel-use standard deviation $\sigma_1 = 4$ L/100 km; Plan B is observed on $n_2 = 45$ independent runs, with $\sigma_2 = 5$ L/100 km. The true mean fuel uses μ_1 and μ_2 are unknown. Find the probability that the observed difference $\bar{X}_1 - \bar{X}_2$ falls within 1.5 L/100 km of the true difference $\mu_1 - \mu_2$.

Solution.

Step 1 — Parameters.

$\sigma_1 = 4$, $n_1 = 50$; $\sigma_2 = 5$, $n_2 = 45$. The means are unknown, but the question only involves the *deviation* of the estimator from them, so they will cancel.

Step 2 — Sampling distribution of the difference.

By (2.13), the mean of $\bar{X}_1 - \bar{X}_2$ is $\mu_1 - \mu_2$, and

$$\operatorname{Var}\left(\bar{X}_1 - \bar{X}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{16}{50} + \frac{25}{45} = 0.3200 + 0.5556 = 0.8756,$$

so by (2.14) the standard error is $\sqrt{0.8756} = 0.9357$ L/100 km.

Step 3 — Standardize.

$$P\Big(\big|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\big| < 1.5\Big) = P\left(|Z| < \frac{1.5}{0.9357}\right) = P\big(|Z| < 1.60\big).$$

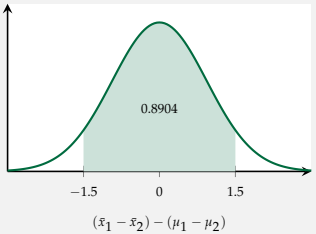

Figure 2.10. The deviation of $\bar{X}_1 - \bar{X}_2$ from the true difference in Example 2.8, with the event $|\text{deviation}| < 1.5$ shaded.

Step 4 — Read the table.

$\Phi(1.60) = 0.9452$, so

$$P\big(|Z| < 1.60\big) = 0.9452 - 0.0548 = 0.8904.$$

Step 5 — Answer in context.

With these sample sizes, $\boxed{P \approx 0.8904}$: there is about an 89% chance the measured fuel-use gap between the two plans is within 1.5 L/100 km of the true gap. If procurement needs the gap pinned down more tightly than that, more runs are required on both plans.

Engineering judgment. In two-population comparisons, engineering discipline is in the sampling, not the algebra: the runs must be independent *between* plans (different trucks and days, not the same truck driven twice) and measured in identical units and conditions. If Plan B's runs were all in summer traffic, no standard-error formula repairs the comparison.

Reference. Montgomery & Runger, 6th ed., Section 7.2.

2.10 *Differences of Two Proportions*

The same construction applies when both populations are binary. Let $\hat{P}_1$ be the sample proportion from n_1 trials with true proportion p_1, and $\hat{P}_2$ from an independent sample of n_2 trials with true proportion p_2. Each proportion has the standard error given by (2.11); combining them with (2.14) gives, for large samples,

$$\hat{P}_1 - \hat{P}_2 \;\dot{\sim}\; N\left(p_1 - p_2,\; \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right), \tag{2.15}$$

valid when all four counts $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, $n_2(1-p_2)$ are at least 10. Table 2.4 places the two difference estimators side by side; notice that both rows follow one pattern, which is the pattern of (2.14): unbiased for the parameter difference, and variances that add.

	Mean difference $\bar{X}_1 - \bar{X}_2$	Proportion difference $\hat{P}_1 - \hat{P}_2$
Mean of estimator	$\mu_1 - \mu_2$	$p_1 - p_2$
Variance of estimator	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
Reference	(2.13)	(2.15)

Table 2.4. Sampling distributions of the two multi-population estimators. Both are unbiased, and in both the variances of the two samples add.

Example 2.9 — Comparing two guardrail versions

An AI safety team runs two versions of a content guardrail on separate, independent streams of benign production prompts. Version 1 is evaluated on $n_1 = 200$ prompts and has true false-positive rate $p_1 = 0.08$; Version 2 is evaluated on $n_2 = 250$ prompts with true false-positive rate $p_2 = 0.05$. Find the probability that the difference in sample proportions $\hat{P}_1 - \hat{P}_2$ falls within 0.03 of the true difference $p_1 - p_2$.

Solution.

Step 1 — Parameters and checks.

$p_1 = 0.08$, $n_1 = 200$; $p_2 = 0.05$, $n_2 = 250$; independent streams. Approximation checks: $n_1 p_1 = 16$, $n_1(1-p_1) = 184$, $n_2 p_2 = 12.5$, $n_2(1-p_2) = 237.5$, all at least 10.

Step 2 — Standard error of the difference.

By (2.15),

$$\begin{aligned}\operatorname{Var}(\hat{P}_1 - \hat{P}_2) &= \frac{0.08 \times 0.92}{200} + \frac{0.05 \times 0.95}{250} \\ &= 0.0003680 + 0.0001900 = 0.0005580,\end{aligned}$$

so the standard error is $\sqrt{0.0005580} = 0.02362$.

Step 3 — Standardize.

$$P\Big(\big|(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)\big| \le 0.03\Big) = P\left(|Z| \le \frac{0.03}{0.02362}\right) = P\big(|Z| \le 1.27\big).$$

Step 4 — Read the table.

Table 2.5 gives $\Phi(1.27) = 0.8980$, so

$$P\big(|Z| \le 1.27\big) = 0.8980 - 0.1020 = 0.7960.$$

z	0.05	0.06	0.07
1.1	0.8749	0.8770	0.8790
1.2	0.8944	0.8962	0.8980
1.3	0.9115	0.9131	0.9147

Table 2.5. Standard normal table excerpt: $\Phi(1.27) = 0.8980$.

Step 5 — Answer in context.

$\boxed{P \approx 0.7960}$. Note what this number warns about: the true gap between the two guardrails is $p_1 - p_2 = 0.03$, and there is a 20% chance the measured gap misses the true gap by *more* than 0.03—enough to reverse the apparent ranking of the two versions. Comparing rare-event rates reliably needs large samples on both sides.

Safety lens. Example 2.9 is the statistical core of every "model $v2$ beats model $v1$" claim. Before accepting such a claim, ask for n_1, n_2, and the standard error of the difference from (2.14). If the observed improvement is smaller than about two standard errors, the ranking could plausibly flip on a re-run.

Chapter Summary

This chapter built the vocabulary and the machinery of estimation. A *parameter* is a fixed, unknown property of a population; a *statistic* is a quantity computed from a random sample, and it is a random variable with its own *sampling distribution*. A statistic used to estimate a parameter is a *point estimator* (2.1). An estimator is judged by its *bias* (2.3), its *standard error*, and their combination, the *mean squared error* (2.6)—exactly variance plus bias squared (Procedure 2.1). The sample mean is unbiased with variance σ^2/n (2.4), and by the Central Limit Theorem (2.8) its sampling distribution is approximately normal for large n, whatever the population shape—exactly normal for normal populations (2.7).

This turns questions about $\bar{X}$, $\hat{P}$, and differences of estimators into standard normal table lookups (Procedures 2.2 and 2.3). For independent estimators, standard errors combine by squaring, adding, and taking the square root (2.14): variances add, even for differences.

Quantity	Formula	Equation
Point estimator	$\hat{\Theta} = h(X_1, \dots, X_n)$	(2.1)
Mean of $\bar{X}$	$E[\bar{X}] = \mu$	(2.2)
Bias	$\text{Bias}(\hat{\Theta}) = E[\hat{\Theta}] - \theta$	(2.3)
Variance of $\bar{X}$	$\text{Var}(\bar{X}) = \sigma^2/n$	(2.4)
Standard error of $\bar{X}$	$\sigma_{\bar{X}} = \sigma/\sqrt{n}$, est. $s/\sqrt{n}$	(2.5)
Mean squared error	$\text{MSE} = \text{Var} + \text{Bias}^2$	(2.6)
Central Limit Theorem	$Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$	(2.8)
$\bar{X}$, normal population	$\bar{X} \sim N(\mu, \sigma^2/n)$ exactly	(2.7)
Sample size for margin E	$n = (z\sigma/E)^2$	(2.9)
SE of $\hat{P}$	$\sigma_{\hat{P}} = \sqrt{p(1-p)/n}$	(2.11)
Difference of means	$\bar{X}_1 - \bar{X}_2 \mathrel{\dot\sim} N\big(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\big)$	(2.13)
SE of a difference	$\text{se} = \sqrt{\text{se}_1^2 + \text{se}_2^2}$	(2.14)
Difference of proportions	$\hat{P}_1 - \hat{P}_2 \mathrel{\dot\sim} N\big(p_1 - p_2,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\big)$	(2.15)

Table 2.6. Key formulas of Chapter 2 at a glance.

Exercises

Concept checks

Exercise 2.1. A rail operator reports: ''the average load of the 60 freight cars we weighed this week is 88.4 tonnes.'' Which statement is correct?

(A) 88.4 is a parameter because it describes the fleet.

(B) 88.4 is a statistic, and it estimates the unknown parameter μ.

(C) 88.4 is a parameter because it was computed exactly, with no error.

(D) 88.4 is a statistic, so the fleet mean must also equal 88.4.

Exercise 2.2. An estimator $\hat{\Theta}$ of θ is called unbiased when:

(A) every estimate it produces equals θ.

(B) its variance shrinks to zero as n grows.

(C) $E[\hat{\Theta}] = \theta$.

(D) its MSE is smaller than that of every other estimator.

Also state, in one sentence, what unbiasedness does *not* guarantee.

Sampling distribution of the sample mean

Exercise 2.3. Warehouse picking times have standard deviation $\sigma = 45$ seconds. An analyst samples $n = 25$ picks and computes $\bar{X}$. (a) Find the standard error of $\bar{X}$. (b) The analyst wants the standard error cut in half. How many picks are needed, and what general rule does this illustrate?

Exercise 2.4. Fuel use for a truck model is normally distributed with mean $\mu = 32$ L/100 km and standard deviation $\sigma = 3$ L/100 km. A fleet audit samples $n = 9$ trucks. Find $P(\bar{X} > 33)$, and state why no approximation is involved.

Exercise 2.5. An evaluation team scores an LLM on a benchmark whose individual-item scores have $\sigma = 0.15$. How many items must be scored so that the standard error of the mean score is at most 0.02?

Estimator quality

Exercise 2.6. Delivery lateness (minutes) has unknown mean μ and variance $\sigma^2 = 16$. From $n = 16$ deliveries, an analyst proposes the estimator $\hat{\Theta} = 0.9\,\bar{X}$ instead of $\bar{X}$. (a) Find the bias, variance, and MSE of both estimators. (b) For which values of μ does $\hat{\Theta}$ have the smaller MSE?

Exercise 2.7. Two vendors offer defect-rate estimators for the same red-team pipeline. Vendor A's estimator is unbiased with variance 9. Vendor B's estimator has bias -2 and variance 4 (same units). Which has the smaller MSE? Would your recommendation change for a safety-critical deployment, and why?

Proportions and two populations

Exercise 2.8. A guardrail's true false-positive rate on benign prompts is $p = 0.10$. It is evaluated on $n = 400$ independent prompts. (a) Verify the normal approximation is appropriate and state the approximate sampling distribution of $\hat{P}$. (b) Find $P(\hat{P} > 0.13)$.

Exercise 2.9. Two warehouses are compared on picking time. Warehouse 1: $n_1 = 36$ picks, $\sigma_1 = 6$ s. Warehouse 2: $n_2 = 64$ independent picks, $\sigma_2 = 8$ s. Find the probability that $\bar{X}_1 - \bar{X}_2$ falls within 2 seconds of the true difference $\mu_1 - \mu_2$.

Exercise 2.10. An A/B test compares signup conversion on two page designs, with $n_1 = 300$ visitors on design 1 (true rate $p_1 = 0.12$) and an independent $n_2 = 300$ visitors on design 2 (true rate $p_2 = 0.10$). Find the probability that the test shows design 1 ahead, that is, $P(\hat{P}_1 - \hat{P}_2 > 0)$. Comment on what the result means for running small A/B tests.

Exam-style problems

Exercise 2.11. An autonomous-vehicle program logs the distance driven between disengagements. Across the fleet, the distance has mean $\mu = 180$ km and standard deviation $\sigma = 60$ km; the distribution is right-skewed, not normal. A regulator audits a random sample of $n = 36$ drives.

1. **[2 pt]** Identify the parameter, the estimator, and the estimate in this audit setting.
2. **[4 pt]** State the approximate sampling distribution of $\bar{X}$, justifying the approximation despite the skewed population.
3. **[4 pt]** Find the probability that the audit average falls below 160 km.
4. **[5 pt]** The regulator wants the standard error of the audit mean to be at most 5 km. How many drives must be audited?

Exercise 2.12. A shipper is choosing between two carriers using delivery lateness data. Carrier A: $n_1 = 40$ deliveries, $\sigma_1 = 9$ minutes. Carrier B: $n_2 = 50$ independent deliveries, $\sigma_2 = 10$ minutes.

1. **[4 pt]** State the approximate sampling distribution of $\bar{X}_A - \bar{X}_B$ in terms of $\mu_A - \mu_B$.
2. **[5 pt]** Find the probability that the observed difference is within 4 minutes of the true difference.
3. **[3 pt]** If both sample sizes are doubled, what happens to the standard error of the difference? Give the exact factor.

Assessing outside recommendations

Exercise 2.13. A consultant analyzes your startup's support desk. Ticket resolution times have mean $\mu = 40$ minutes and standard deviation $\sigma = 10$ minutes (approximately normal). For a weekly report averaging $n = 25$ tickets, the consultant writes: ''$P(\bar{X} > 44) = P\big(Z > (44 - 40)/10\big) = P(Z > 0.40) = 0.3446$, so about a third of weekly reports will show an average above 44 minutes.'' Find the error, explain it, and give the correct probability.

Exercise 2.14. An AI assistant is asked to compare mean benchmark scores of two model versions evaluated on independent test sets. Version 1: $n_1 = 64$ items, $\sigma_1 = 0.12$. Version 2: $n_2 = 81$ items, $\sigma_2 = 0.09$. The assistant writes: ''$\mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \sigma_1^2/n_1 - \sigma_2^2/n_2 = 0.000225 - 0.000100 = 0.000125$, so se $= 0.01118$, and the observed difference falls within 0.03 of the true difference with probability $P(|Z| < 2.68) = 0.9926$.'' Find the error, explain why it matters, and recompute the probability correctly.

In-Class Activity 2.1: One Fleet, Many Averages

Week 1 — Lecture 2

Work in teams of 3–4. Write your reasoning, not just the final number. One submission per team, all names on it.

Your team audits an autonomous shuttle fleet. The quantity of interest is the mean driving distance between disengagements across all fleet driving under the current software version.

Part 1 — Name the pieces

For this audit, identify in one line each: the *population*, the *parameter*, the *sample*, and the *statistic*. Say which of the four are random and which are fixed.

Part 2 — Two teams, two answers

Team Alpha runs 10 test drives and reports a mean of 142 km between disengagements. Team Beta independently runs 10 drives and reports 187 km. Neither team made any measurement or arithmetic error. Explain how both reports can be correct, and state which single concept from today's lecture describes this situation.

Part 3 — Buying precision

Fleet history suggests $\sigma \approx 60$ km. If a follow-up audit uses $n = 25$ drives, compute the stan
Then state how many drives would be needed to cut that standard error in half, and justify w
rather than trial and error.

Exit check

One sentence: what is the difference between the population distribution and the sampling di

In-Class Activity 2.2: Which Estimator Would You Ship?

Week 2 — Lecture 3

Work in teams of 3–4. Write your reasoning, not just the final number. One submission per team, all names on it.

Your startup needs one number for mean checkout time μ (minutes) to put in an investor update. Checkout times have $\sigma^2 = 16$, and you have a random sample of $n = 16$ recent checkouts. Three teammates propose estimators: (i) $\bar{X}$; (ii) X_1, the first checkout in the sample (''it's quicker to look up''); (iii) $0.8\,\bar{X}$ (''our times are inflated by a few outliers, shrink them'').

Part 1 — Bias first

For each of the three estimators, state its expected value in terms of μ and say whether it is unbiased. No numerical computation needed yet.

Part 2 — MSE showdown

Compute the variance and MSE of $\bar{X}$ and of $0.8\bar{X}$ (leave the MSE of $0.8\bar{X}$ as a function of μ). Then evaluate both MSEs at $\mu = 2$ and at $\mu = 10$, and state which estimator wins in each case.

Part 3 — Your shipping decision

Typical checkout times in your product are 6–12 minutes. Write one decision sentence in the fo
because Y,'' where Y refers to bias, variance, or MSE.

Exit check

In one sentence: when can a biased estimator be the better choice?

In-Class Activity 2.3: Two Carriers, One Contract

Week 2 — Lecture 4

Work in teams of 3–4. Write your reasoning, not just the final number. One submission per team, all names on it.

Procurement must choose between two trucking carriers using delivery-time data. Carrier A: $n_1 = 36$ deliveries observed, $\sigma_1 = 6$ minutes. Carrier B: $n_2 = 64$ independent deliveries, $\sigma_2 = 8$ minutes. The true mean delivery times μ_A and μ_B are unknown.

Part 1 — Write the distribution

Write the approximate sampling distribution of $\bar{X}_A - \bar{X}_B$: give its mean (in symbols) and compute its variance and standard error numerically. State the rule you used to combine the two samples' variances.

Part 2 — How trustworthy is the observed gap?

Using $\Phi(2.12) = 0.9830$, find the probability that the observed difference $\bar{X}_A - \bar{X}_B$ falls within 3 minutes of the true difference $\mu_A - \mu_B$. Show the standardization step.

Part 3 — Challenge the shortcut

A colleague says: "Skip the statistics—whichever carrier has the smaller sample mean is faster. S
Using your Part 2 result, write two sentences explaining when this shortcut is safe and when i
the wrong carrier.

Exit check

True or false, with one reason: for independent samples, the variance of a *difference* of samp
difference of their variances.