3 *Confidence Intervals: Inference for a Single Sample*

Chapter 2 taught us to compute a point estimate and to describe how much that estimate varies from sample to sample. This chapter takes the natural next step: instead of reporting a single number, we report an *interval* of plausible values together with a stated level of confidence. The chapter is organized in the following order: why a point estimate alone is not enough; the confidence interval for a mean when the population standard deviation is known, derived directly from the sampling distribution; the correct interpretation of the word ''confidence''; one-sided bounds; the choice of sample size; the t distribution and the interval for a mean when the standard deviation is unknown; the chi-square interval for a variance; the large-sample interval for a proportion; prediction intervals for a single future observation; tolerance intervals for a stated fraction of the population; and a decision procedure for choosing the right interval. The chapter closes with a full practice section that mirrors the in-class practice lecture, because reading a problem and selecting the correct interval is a skill of its own.

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

Q1. A logistics analyst reports a 95% confidence interval of $(57.8,\ 64.8)$ hours for the *mean* container dwell time at a port. A colleague reads this as ''95% of containers dwell between 57.8 and 64.8 hours.'' Is the colleague right?

Q2. A startup runs an experiment with $n = 25$ users and builds a 95% confidence interval for mean session length. If the experiment is repeated with $n = 100$ users, what happens to the width of the interval?

Q3. An AI safety team has $n = 12$ benchmark scores for a model and does not know the population standard deviation. Should they build the interval for the

mean score with $z_{\alpha/2}$ or with something else?

Answers.

Q1. *No.* The interval describes the plausible location of the population *mean* μ, not the spread of individual containers. Individual dwell times vary far more than the mean does. An interval that covers a stated fraction of individual observations is a *tolerance interval*, which is a different (and much wider) object introduced at the end of this chapter.

Q2. It is cut in half. The half-width is $z_{\alpha/2}\,\sigma/\sqrt{n}$, so multiplying n by 4 divides the width by $\sqrt{4} = 2$. Extra precision must be bought with the *square* of the desired improvement in width.

Q3. With the t critical value $t_{\alpha/2,\,n-1} = t_{0.025,\,11}$. When σ is replaced by the sample standard deviation S, the standardized statistic no longer follows $N(0,1)$; it follows the t distribution with $n-1 = 11$ degrees of freedom, which has heavier tails and therefore a larger critical value. Using z would make the interval too narrow.

Lecture 5 starts here — *Interval estimation*

Lecture 5. This section is covered by Lecture 5.

Reference. Montgomery & Runger, 6th ed., Section 8.1.

3.1 From Point Estimate to Interval Estimate

A point estimate answers the question ''what is μ?'' with a single number, for example $\hat{\mu} = \bar{x} = 85.3$. The number is useful, but it carries no measure of its own reliability. If a second sample would have given 79.1 and a third 91.6, then reporting 85.3 alone hides everything we learned in Chapter 2 about sampling variability. The reader of the report cannot tell whether the estimate is precise to within one unit or to within twenty.

Interval estimate. A range $(L,\ U)$, computed from the sample, that is likely to contain the unknown parameter.

The remedy is to report an *interval estimate*: a range $(L,\ U)$, computed from the sample, that is likely to contain the true parameter. The interval carries its own honesty. A wide interval says ''the data are thin; do not lean on this number,'' and a narrow interval says ''the data pin the parameter down tightly.''

Confidence interval. An interval estimate with a stated confidence level $100(1-\alpha)\%$; the endpoints are computed from the sample.

Confidence level. The long-run fraction $1-\alpha$ of intervals, produced by the same method over repeated samples, that contain the true parameter.

A *confidence interval* (CI) is an interval estimate with a stated *confidence level*, written $100(1-\alpha)\%$. The most common levels are 90% ($\alpha = 0.10$), 95% ($\alpha = 0.05$), and 99% ($\alpha = 0.01$). Every confidence interval in this chapter has the same

anatomy:

$$\underbrace{\text{point estimate}}_{\text{center}} \pm \underbrace{\text{critical value} \times \text{standard error}}_{\text{margin of error } E}.$$

Four ingredients appear again and again, so learn their names now. The *lower confidence limit* L and *upper confidence limit* U are the two endpoints. The *margin of error* E is the half-width, so a two-sided interval is $\bar{x} \pm E = (\bar{x} - E,\ \bar{x} + E)$. The *critical value* is the multiplier taken from the reference distribution; for the standard normal it is $z_{\alpha/2}$, the value that cuts off area $\alpha/2$ in the upper tail. The *standard error* is the standard deviation of the sampling distribution of the estimator; for $\bar{X}$ it is $\sigma/\sqrt{n}$, exactly as in (2.5). The margin of error multiplies the last two: $E = z_{\alpha/2} \times \sigma/\sqrt{n}$.

Intervals also come in one-sided versions. A *two-sided* interval bounds the parameter from both directions, $L \le \mu \le U$. A *one-sided lower bound* states only "$\mu \ge L$" (useful for quantities that must be large enough, such as strength or accuracy), and a *one-sided upper bound* states only "$\mu \le U$" (useful for quantities that must be small enough, such as contamination or error rates). Section 3.4 develops them.

Margin of error. The half-width E of a two-sided interval; the interval is (point estimate $\pm E$).

Critical value. The percentile of the sampling distribution that cuts off the required tail area, e.g. $z_{\alpha/2}$ for the standard normal.

Startup lens. An investor update that says "average activation time is 18.2 minutes" invites the question "how sure are you?" An update that says "18.2 minutes, 95% CI (16.6, 19.8)" answers it before it is asked. Intervals are how a data-driven team communicates what it does not know.

Reference. Montgomery & Runger, 6th ed., Section 8-1.1.

3.2 *Confidence Interval on the Mean, σ Known*

We begin with the cleanest case: the sample $X_1, X_2, \ldots, X_n$ comes from a normal population (or n is large enough for the CLT), the mean μ is unknown, and the standard deviation σ is *known*, typically from long production history or a validated measurement system. The entire derivation is three steps, and each step is something we already know.

Step 1 — Start from the sampling distribution.

From Chapter 2, the sample mean satisfies $\bar{X} \sim N(\mu,\ \sigma^2/n)$, so the standardized statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,\ 1)$$

exactly as in the CLT statement (2.8).

Step 2 — Trap Z between two percentiles.

By the definition of $z_{\alpha/2}$, the standard normal variable lands between $-z_{\alpha/2}$

and $z_{\alpha/2}$ with probability $1-\alpha$ (Figure 3.1):

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha.$$

Step 3 — Rearrange to isolate μ.

Multiply through by $\sigma/\sqrt{n}$, subtract $\bar{X}$, and flip the signs (which reverses the inequalities):

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Replacing the random variable $\bar{X}$ by its observed value $\bar{x}$ gives the confidence interval.

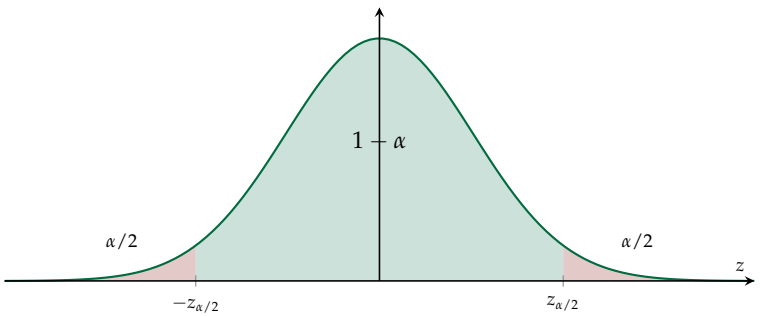

Figure 3.1. The standard normal density with the central area $1-\alpha$ (green) between $-z_{\alpha/2}$ and $z_{\alpha/2}$, and area $\alpha/2$ in each tail (maroon). For a 95% interval, $\alpha = 0.05$ and $z_{0.025} = 1.96$.

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;\le\; \mu \;\le\; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{(two-sided } 100(1-\alpha)\%\text{ CI, } \sigma \text{ known)} \tag{3.1}$$

Equation (3.1) is often abbreviated as $\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$. Please pay attention to what is random here, because this point decides how the interval may be interpreted. Before the data are collected, the endpoints are random (they contain $\bar{X}$) and μ is a fixed unknown constant. The probability statement in Step 3 is about the random endpoints capturing the fixed μ, not about μ moving around inside a fixed interval. Section 3.3 returns to this.

Confidence level	α	Two-sided		One-sided	
		$\Phi(z_{\alpha/2})$	$z_{\alpha/2}$	$\Phi(z_\alpha)$	z_α
90%	0.10	0.950	1.645	0.90	1.282
95%	0.05	0.975	1.960	0.95	1.645
99%	0.01	0.995	2.576	0.99	2.326

Table 3.1. Common critical values of the standard normal distribution. Two-sided intervals split α between the tails and use $z_{\alpha/2}$; one-sided bounds put all of α in one tail and use z_α.

Table 3.1 lists the critical values used most often. Memorize the 95% two-sided value $z_{0.025} = 1.96 \approx 2$; the others can always be recovered from a z table, and Section 3.6 shows exactly how.

Procedure 3.1 — Constructing a z confidence interval for μ.

1. Identify what is given: n, $\bar{x}$, the known σ, the confidence level $100(1-\alpha)\%$, and whether the interval is two-sided or one-sided.
2. Find the critical value: $z_{\alpha/2}$ for a two-sided interval, z_α for a one-sided bound (Table 3.1).
3. Compute the standard error $\mathrm{SE} = \sigma/\sqrt{n}$.
4. Compute the margin of error $E = z \times \mathrm{SE}$.
5. Assemble the interval: $\bar{x} \pm E$ (two-sided), $\mu \geq \bar{x} - E$ (lower bound), or $\mu \leq \bar{x} + E$ (upper bound).
6. Report the interval with its units and one interpretive sentence.

3.3 *What ''95% Confident'' Actually Means*

The confidence level describes the *procedure*, not any single computed interval. ''We are 95% confident'' means: if we repeated the whole process, drawing a fresh sample and recomputing the interval each time, then about 95 out of every 100 intervals produced this way would contain the true μ. Figure 3.2 shows twenty

Engineering judgment. Knowing σ is an engineering claim, not a mathematical one. It means the measurement system and the process have been stable long enough that historical variation can be trusted. If the process was retuned last month, the historical σ may no longer describe it, and the t interval of Section 3.7 is the honest choice.

Reference. Montgomery & Runger, 6th ed., Section 8-1.1.

intervals built from twenty independent samples of the same population; the method did its job for eighteen of them, and missed for two.

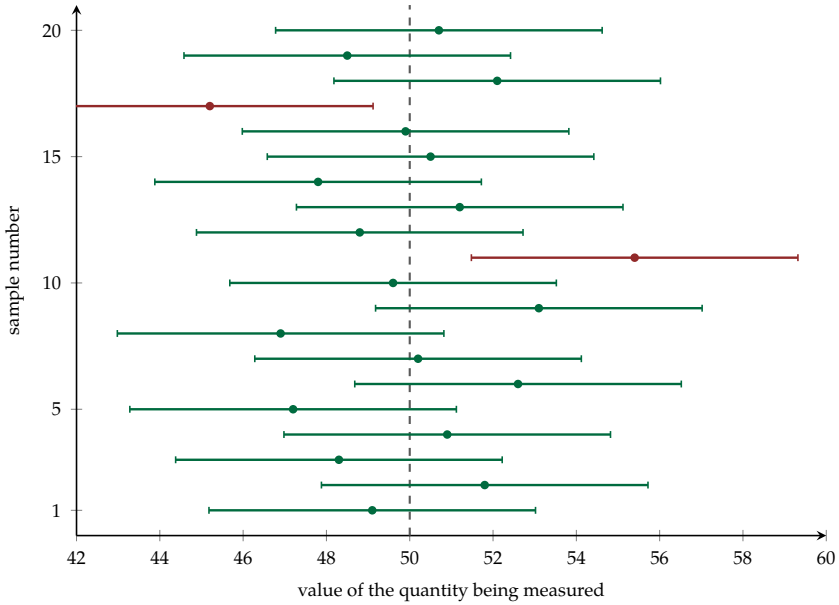

Figure 3.2. Twenty 95% confidence intervals, each built from an independent sample of $n = 25$ from a population with $\mu = 50$ and $\sigma = 10$. Eighteen intervals (green) contain μ; samples 11 and 17 (maroon) happened to produce unusually high and unusually low means, and their intervals miss. In the long run about one interval in twenty misses.

Please pay attention here, because this is a common trap. The statement ''there is a 95% probability that μ is inside $(57.8,\ 64.8)$'' is *wrong*, and it is wrong for a precise reason: once the interval is computed, nothing in it is random anymore. The true μ is a fixed constant — it is either inside $(57.8,\ 64.8)$ or it is not, and no probability attaches to a fact that is already settled, merely unknown to us. What was random was the *interval*: before sampling, the endpoints were random variables, and the method that generates them succeeds 95% of the time.

Two phrasings are correct and safe to use in reports:

- ''We are 95% confident that μ lies between L and U.''
- ''The interval $(L,\ U)$ was constructed by a method that captures μ in 95% of repeated samples.''

Insight. Why the confidence attaches to the method and not to the interval. The derivation in Section 3.2 made one probability statement, in Step 2, and it was a statement about the random variable Z — equivalently, about the random endpoints $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$. When we substitute the observed $\bar{x}$, we cash in that guarantee for one realized interval, the way a factory's 95% pass rate applies to the production line and not to one particular unit sitting on the desk. The unit either passed or it did not; the 95% describes the line.

3.4 One-Sided Confidence Bounds

Many engineering questions only care about one direction. A red-team lead wants to guarantee a model's mean accuracy is *at least* some value; a port authority wants mean emissions *at most* some value. In these cases a two-sided interval wastes half of the error budget on a direction nobody asked about. The fix is to put the entire tail area α on one side, which replaces $z_{\alpha/2}$ by the smaller value z_α (Figure 3.3):

$$\mu \;\geq\; \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \mu \;\leq\; \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \qquad \text{(one-sided } 100(1-\alpha)\% \text{ bounds)} \tag{3.2}$$

Please pay attention here, because this is a common trap: at the same confidence level, the one-sided critical value is *smaller* than the two-sided one (1.645 versus 1.96 at 95%), because the whole error budget sits in one tail. Students who reflexively write 1.96 in a one-sided problem lose the benefit that motivated the one-sided bound in the first place.

Safety lens. Regulators reviewing an autonomous-vehicle safety case read confidence statements literally. Writing "the failure rate is in (0.8%, 1.6%) with probability 0.95" in a safety report is a methodological error a reviewer will flag; writing "a 95% CI for the failure rate is (0.8%, 1.6%)" is the defensible claim.

Reference. Montgomery & Runger, 6th ed., Section 8-1.3.

One-sided confidence bound. A bound of the form $\mu \geq L$ or $\mu \leq U$ that spends the entire error probability α in a single tail.

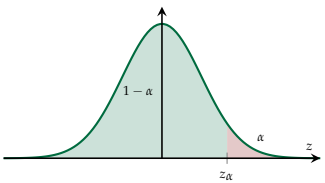

Figure 3.3. A one-sided bound puts all of α in a single tail, so the critical value is z_α, not $z_{\alpha/2}$. At 95%: $z_{0.05} = 1.645 < 1.96 = z_{0.025}$.

Example 3.1 — Container dwell time at a port

A port operations team measures the dwell time of $n = 36$ randomly selected import containers and finds $\bar{x} = 61.3$ hours. Terminal records over several stable years give $\sigma = 10.8$ hours.

(a) Construct a 95% two-sided confidence interval for the true mean dwell time μ.

(b) The terminal's service-level agreement only requires a guarantee that mean dwell is not too long. Construct a 95% one-sided *upper* confidence bound for μ.

Solution.

Step 1 — Set up.

Given: $n = 36$, $\bar{x} = 61.3$, $\sigma = 10.8$ (known), confidence 95%. Since σ is known, Procedure 3.1 applies with critical values from Table 3.1.

Step 2 — Standard error.

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{10.8}{\sqrt{36}} = \frac{10.8}{6} = 1.800 \text{ hours}.$$

Step 3 — Part (a): two-sided interval.

With $\alpha = 0.05$, the two-sided critical value is $z_{0.025} = 1.96$, so $E = 1.96 \times 1.800 = 3.528$, and by (3.1)

$$61.3 \pm 3.528 \quad \Longrightarrow \quad \boxed{(57.77,\ 64.83) \text{ hours}}.$$

We are 95% confident that the mean dwell time of all import containers is between 57.77 and 64.83 hours.

Step 4 — Part (b): one-sided upper bound.

Now all of $\alpha = 0.05$ goes into one tail, so the critical value is $z_{0.05} = 1.645$, and by (3.2)

$$\mu \ \leq \ 61.3 + 1.645 \times 1.800 = 61.3 + 2.961 = \boxed{64.26 \text{ hours}}.$$

We are 95% confident that mean dwell is *at most* 64.26 hours. Notice the one-sided bound (64.26) is tighter than the upper endpoint of the two-sided interval (64.83): when only one direction matters, the one-sided bound gives a stronger guarantee from the same data.

Reference. Montgomery & Runger, 6th ed., Section 8-1.2.

3.5 *Precision and the Choice of Sample Size*

The half-width of the two-sided interval (3.1) is the margin of error, also called the *error bound*:

$$E = z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}. \tag{3.3}$$

Equation (3.3) makes explicit the three levers that control precision. The interval narrows when the sample size n grows, when the population is less variable (smaller σ), or when we accept a lower confidence level (smaller $z_{\alpha/2}$). Only the first lever is usually under our control, and it is priced by a square root: Figure 3.4 shows that each halving of E costs a quadrupling of n.

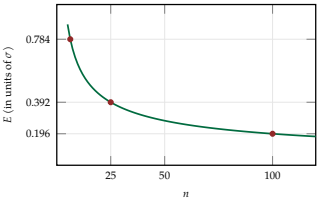

Figure 3.4. The margin of error $E = 1.96\sigma/\sqrt{n}$ at 95% confidence. Halving E (from 0.784σ to 0.392σ to 0.196σ) requires quadrupling n (from about 6 to 25 to 100): precision is bought at a square-root rate.

In planning a study the question runs in reverse: the engineer states the precision E that the decision requires, and asks how large the sample must be. Solving (3.3) for n gives

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2 \qquad \text{(round } \textit{up} \text{ to the next integer)}. \tag{3.4}$$

Always round up. Rounding down, even from 61.47 to 61, delivers a margin of error slightly larger than the one promised, which defeats the purpose of the computation.

Example 3.2 — Planning an onboarding study

Your startup wants to estimate the mean duration of customer onboarding calls. Product telemetry from the past two quarters gives a stable $\sigma = 6$ minutes. The operations lead wants the mean estimated to within $E = 1.5$ minutes.

(a) How many calls must be sampled for 95% confidence?

(b) How many for 99% confidence?

Solution.

Step 1 — Set up.

Given: $\sigma = 6$, $E = 1.5$. This is a direct application of (3.4).

Step 2 — Part (a): 95% confidence.

With $z_{0.025} = 1.96$:

$$n = \left(\frac{1.96 \times 6}{1.5}\right)^2 = \left(\frac{11.76}{1.5}\right)^2 = (7.840)^2 = 61.47 \quad \Longrightarrow \quad \boxed{n = 62 \text{ calls}}.$$

Step 3 — Part (b): 99% ***confidence.***

With $z_{0.005} = 2.576$:

$$n = \left(\frac{2.576 \times 6}{1.5}\right)^2 = \left(\frac{15.46}{1.5}\right)^2 = (10.30)^2 = 106.2 \quad \Longrightarrow \quad \boxed{n = 107 \text{ calls}}.$$

Raising the confidence level from 95% to 99% at the same precision costs about 73% more data. Confidence is not free; it is paid for in sample size.

Startup lens. Sample-size formulas turn statistics into budgeting. If each sampled onboarding call costs 40 SAR of an analyst's time to review, part (a) is a 2,480 SAR study and part (b) a 4,280 SAR study. Deciding between 95% and 99% is a business decision, and (3.4) prices it.

Lecture 6. This section is covered by Lecture 6.

Reference. Montgomery & Runger, 6th ed., Appendix Table III.

Lecture 6 starts here — *Interval estimation, part 2*

3.6 Reading Critical Values from the z Table

Table 3.1 covers the standard confidence levels, but exams and real problems use other levels too, so you must be able to extract any critical value from the cumulative z table. The table stores $\Phi(z) = P(Z \le z)$: rows give the first two digits of z (e.g. 1.9) and columns give the second decimal (e.g. 0.06).

Finding a critical value is the *inverse* lookup: given a probability, search for it *inside* the table and read the row and column headers. For example, to find $z_{0.025}$ we need the z with upper-tail area 0.025, i.e. $\Phi(z) = 1 - 0.025 = 0.975$. Searching the body of the table for 0.9750 (Table 3.2) locates it in row 1.9, column 0.06, so $z_{0.025} = 1.96$.

z	0.00	0.05	0.06	0.07
1.8	0.9641	0.9678	0.9686	0.9693
1.9	0.9713	0.9744	0.9750	0.9756
2.0	0.9772	0.9798	0.9803	0.9808

Table 3.2. Excerpt of the cumulative z table. Searching the body for 0.9750 gives row 1.9 + column 0.06, so $z_{0.025} = 1.96$.

The t and chi-square tables used later in this chapter are laid out the other way around: their columns are indexed directly by the upper-tail area α, so critical values are read off without any inverse search. What those tables add is a second index, the degrees of freedom, one row per value.

Reference. Montgomery & Runger, 6th ed., Section 8-2.

3.7 The t Distribution and the CI on the Mean, σ Unknown

In most real problems nobody hands us σ. The obvious repair is to estimate it with the sample standard deviation S and reuse the z machinery. The obvious repair is not quite right, and the reason deserves a careful look.

When σ is known, the only randomness in $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ comes from the numerator. When we replace σ by S, the denominator becomes random too: S varies from sample to sample, and in small samples it varies a lot. The statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

therefore carries *two* sources of uncertainty, and it does not follow $N(0,1)$. For a normal population it follows the *t distribution* with $\nu = n - 1$ degrees of freedom.

t distribution. A symmetric, bell-shaped distribution with heavier tails than the normal; indexed by its degrees of freedom ν, and converging to $N(0,1)$ as $\nu \to \infty$.

Degrees of freedom. The number of independent pieces of information available for estimating variability; for a single sample estimating μ, $\nu = n - 1$.

The degrees of freedom count the independent deviations available to compute S: the n deviations $x_i - \bar{x}$ always sum to zero, so only $n - 1$ of them are free. The t distribution is symmetric and bell-shaped like the normal, but it has *heavier tails*: extreme values of T are more likely, because an unluckily small S can inflate the ratio. Figure 3.5 shows the effect. As ν grows, S becomes a reliable stand-in for σ and the t density merges into the standard normal; the last row of every t table ($\nu = \infty$) is exactly the z row.

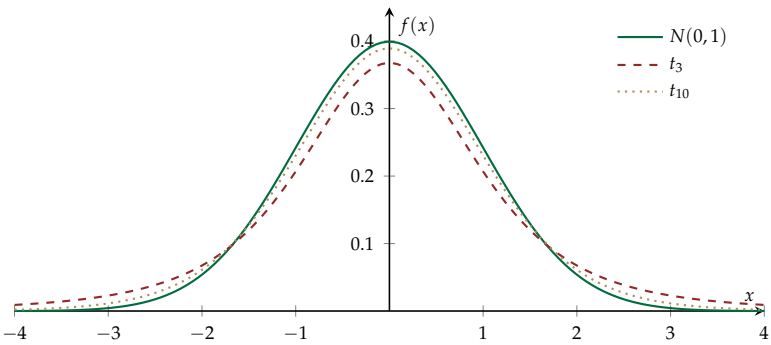

Figure 3.5. The t density for $\nu = 3$ and $\nu = 10$ degrees of freedom against the standard normal. Smaller ν means heavier tails, larger critical values, and wider confidence intervals; as $\nu \to \infty$ the t curve becomes the normal curve.

The critical value $t_{\alpha/2,\,n-1}$ cuts off area $\alpha/2$ in the upper tail of the t distribution with $n - 1$ degrees of freedom, and the confidence interval keeps the familiar anatomy:

$$\bar{x} - t_{\alpha/2,\,n-1}\frac{s}{\sqrt{n}} \;\le\; \mu \;\le\; \bar{x} + t_{\alpha/2,\,n-1}\frac{s}{\sqrt{n}} \qquad (\sigma \text{ unknown}) \qquad (3.5)$$

One-sided t bounds follow the same pattern as (3.2): use $t_{\alpha,\,n-1}$ and keep only the needed side, $\mu \ge \bar{x} - t_{\alpha,\,n-1}\,s/\sqrt{n}$ or $\mu \le \bar{x} + t_{\alpha,\,n-1}\,s/\sqrt{n}$.

Because $t_{\alpha/2,\,n-1} > z_{\alpha/2}$ for every finite n, the t interval is always wider than the z interval computed from the same numbers. That extra width is not a defect; it is the honest price of not knowing σ. For $n \ge 30$ the two critical values are

close (at $\nu = 30$, $t_{0.025} = 2.042$ versus $z_{0.025} = 1.960$), but the rule in this course is simple: *if σ is unknown, use t*, whatever the sample size.

Procedure 3.2 — Constructing a t confidence interval for μ.

1. Confirm that σ is unknown and that the data are plausibly from a normal population (for small n; for larger n the interval is robust to moderate non-normality).
2. Compute (or read off) $\bar{x}$ and s; set $\nu = n - 1$.
3. Find $t_{\alpha/2,\nu}$ (two-sided) or $t_{\alpha,\nu}$ (one-sided) from the t table.
4. Compute $\text{SE} = s/\sqrt{n}$ and $E = t \times \text{SE}$.
5. Assemble $\bar{x} \pm E$, or the one-sided bound, and interpret with units.

Example 3.3 — Mean safety-benchmark score of a fine-tuned model

An AI safety team evaluates a fine-tuned language model on $n = 12$ independent benchmark suites. The scores (out of 100) give $\bar{x} = 78.4$ and $s = 6.2$. No historical σ exists for this model. Construct a 95% confidence interval for the model's true mean benchmark score μ.

Solution.

Step 1 — Set up.

σ is unknown and estimated by $s = 6.2$, so Procedure 3.2 applies with $\nu = n - 1 = 11$.

Step 2 — Critical value.

From the t table (Table 3.3): $t_{0.025,11} = 2.201$.

	upper-tail area α			
ν	0.10	0.05	0.025	0.01
5	1.476	2.015	2.571	3.365
10	1.372	1.812	2.228	2.764
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681
15	1.341	1.753	2.131	2.602
24	1.318	1.711	2.064	2.492
∞	1.282	1.645	1.960	2.326

Table 3.3. Excerpt of the t table, $t_{\alpha,\nu}$ with $P(T > t_{\alpha,\nu}) = \alpha$. For a 95% two-sided CI with $n = 12$: row $\nu = 11$, column 0.025, giving $t_{0.025,11} = 2.201$.

Step 3 — Standard error and margin of error.

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{6.2}{\sqrt{12}} = \frac{6.2}{3.464} = 1.790, \qquad E = 2.201 \times 1.790 = 3.939.$$

Step 4 — Assemble.

By (3.5),

$$78.4 \pm 3.939 \quad \Longrightarrow \quad \boxed{(74.46,\ 82.34) \text{ points}}.$$

We are 95% confident that the model's true mean benchmark score is between 74.46 and 82.34. Had we wrongly used $z_{0.025} = 1.96$, the margin would have been $1.96 \times 1.790 = 3.508$ and the interval $(74.89,\ 81.91)$ — narrower than the data justify. With only 12 suites, the extra width of the t interval is exactly the uncertainty about σ made visible.

Safety lens. Model evaluations are chronically small-sample: benchmark suites are expensive to build, so $n = 10$–20 is typical. This is t-interval territory. An evaluation report that quotes z-based error bars on a dozen suites is systematically overconfident about every model it ranks.

Reference. Montgomery & Runger, 6th ed., Section 8-3.

3.8 *Confidence Interval on the Variance*

Sometimes the parameter of interest is not the center but the *spread*. A process with the right mean but erratic variation still fails its users: delivery times that average 30 minutes but swing between 5 and 90 are a bad product. For a normal population, inference about σ^2 is built on the distribution of the sample variance: the scaled quantity

Chi-square distribution. A right-skewed distribution on the positive axis, indexed by degrees of freedom ν; the distribution of $(n-1)S^2/\sigma^2$ for normal samples.

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

follows the *chi-square distribution* with $\nu = n-1$ degrees of freedom.

The chi-square distribution takes only positive values and is skewed to the right, so it is *not* symmetric (Figure 3.6). This has a practical consequence: unlike the z and t cases, we need *two* different critical values, one for each tail. The value $\chi^2_{\alpha/2,\,n-1}$ cuts off area $\alpha/2$ in the upper tail, and $\chi^2_{1-\alpha/2,\,n-1}$ cuts off area $\alpha/2$ in the lower tail (its upper-tail area is $1-\alpha/2$).

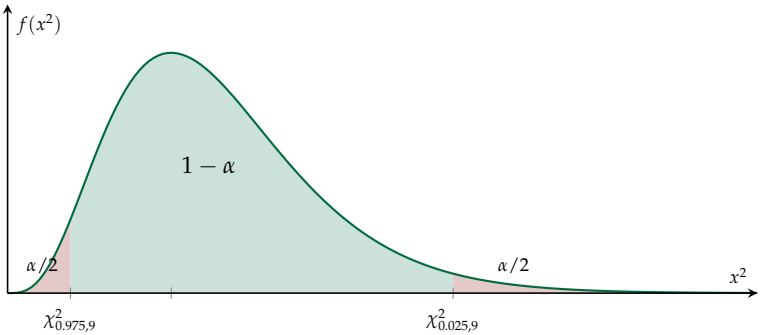

Figure 3.6. The chi-square density with $\nu = 9$ degrees of freedom. The distribution is right-skewed, so the two critical values ($\chi^2_{0.975,9} = 2.700$ and $\chi^2_{0.025,9} = 19.023$ for $\alpha = 0.05$) are not symmetric around the center, and the resulting CI for σ^2 is not symmetric around s^2.

Trapping X^2 between the two critical values and solving for σ^2 (the algebra flips the fractions, which swaps which critical value lands in which endpoint) gives

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,\,n-1}} \;\le\; \sigma^2 \;\le\; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,\,n-1}} \qquad \text{(two-sided } 100(1-\alpha)\%\text{ CI for } \sigma^2\text{)} \quad (3.6)$$

Please pay attention here, because this is a common trap: the *lower* CI endpoint uses the *larger* chi-square value (it sits in the denominator), and the upper endpoint uses the smaller one. If your computed interval comes out with $L > U$, you have swapped them. A CI for the standard deviation σ is obtained by taking square roots of both endpoints. Finally, the chi-square interval leans hard on the normality of the population — much harder than the t interval does — so check normality (for example with the plots of Chapter 1) before trusting it.

Procedure 3.3 — Constructing a chi-square confidence interval for σ^2.

1. Verify the population is plausibly normal; this interval is sensitive to non-normality.
2. Compute s^2 and set $\nu = n - 1$.
3. Read both critical values from the chi-square table: $\chi^2_{\alpha/2,\,\nu}$ (large) and $\chi^2_{1-\alpha/2,\,\nu}$ (small).
4. Compute $(n-1)s^2$ once, then divide it by the large value for the lower endpoint and by the small value for the upper endpoint, as in (3.6).
5. For a CI on σ, take square roots of the endpoints.

Example 3.4 — Consistency of LLM response times

An AI product team monitors the response time of a deployed language model. Users tolerate a slow-but-steady service better than an erratic one, so the team tracks the *variance*. A sample of $n = 15$ requests (response times approximately normal) gives $s = 180$ ms, i.e. $s^2 = 32{,}400$ ms^2. Construct a 95% confidence interval for σ^2 and for σ.

Solution.

Step 1 — Set up.

$\nu = n - 1 = 14$ and $\alpha = 0.05$, so we need $\chi^2_{0.025,\,14}$ and $\chi^2_{0.975,\,14}$.

Step 2 — Critical values.

From the chi-square table (Table 3.4): $\chi^2_{0.025,\,14} = 26.119$ and $\chi^2_{0.975,\,14} = 5.629$.

ν	upper-tail area α 0.975	0.95	0.05	0.025
9	2.700	3.325	16.919	19.023
10	3.247	3.940	18.307	20.483
14	5.629	6.571	23.685	26.119
15	6.262	7.261	24.996	27.488
19	8.907	10.117	30.144	32.852
24	12.401	13.848	36.415	39.364

Table 3.4. Excerpt of the chi-square table, $\chi^2_{\alpha,\nu}$ with $P(X^2 > \chi^2_{\alpha,\nu}) = \alpha$. A 95% CI for σ^2 with $n = 15$ uses *both* highlighted values in row $\nu = 14$.

Step 3 — Compute $(n-1)s^2$ ***and assemble.***

$(n-1)s^2 = 14 \times 32{,}400 = 453{,}600$. By (3.6),

$$\frac{453{,}600}{26.119} \le \sigma^2 \le \frac{453{,}600}{5.629} \quad \Longrightarrow \quad \boxed{17{,}367 \le \sigma^2 \le 80{,}583 \text{ ms}^2}.$$

Step 4 — Convert to standard deviation.

Taking square roots: $\sqrt{17{,}367} = 131.8$ and $\sqrt{80{,}583} = 283.9$, so we are 95% confident that $131.8 \le \sigma \le 283.9$ ms. Notice the interval is far from symmetric around $s = 180$ ms: the right-skew of the chi-square distribution passes through to the interval. With only 15 observations, the variance is known only to within roughly a factor of five, which is normal — variances are much harder to estimate than means.

Engineering judgment. Variance is often the engineering specification. A robot arm whose mean placement is perfect but whose spread exceeds the part tolerance produces scrap; a delivery service whose spread is huge cannot promise windows. When the contract is written on consistency, interval (3.6) is the one to report.

Reference. Montgomery & Runger, 6th ed., Section 8-4.

3.9 *Large-Sample Confidence Interval on a Proportion*

Many quantities of interest are proportions: the fraction of prompts a guardrail wrongly blocks, the fraction of shipments that arrive late, the fraction of trial users who convert to paid. If X counts the successes in n independent trials, the sample proportion is $\hat{p} = X/n$, and Chapter 2 gave its standard error in (2.11). For large samples $\hat{p}$ is approximately normal, and substituting $\hat{p}$ for p in the standard error gives the large-sample interval

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;\le\; p \;\le\; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \text{(large-sample CI for } p\text{)} \tag{3.7}$$

Large-sample condition. The normal approximation behind (3.7) is adequate when $n\hat{p} \ge 10$ and $n(1-\hat{p}) \ge 10$: at least ten successes and ten failures.

The word ''large-sample'' is a working condition, not decoration. Before using (3.7), check

$$n\hat{p} \ge 10 \qquad \text{and} \qquad n(1-\hat{p}) \ge 10,$$

that is, at least ten successes *and* ten failures in the data. When the condition fails — four failures observed in thirty tests, for instance — the normal approximation distorts the tails badly, and exact binomial methods (outside the scope of this chapter) should be used instead. One-sided bounds for p follow the usual pattern with z_α.

Procedure 3.4 — Constructing a large-sample confidence interval for p.

1. Compute $\hat{p} = x/n$ from the count of successes x.
2. Check the conditions $n\hat{p} \ge 10$ and $n(1-\hat{p}) \ge 10$. If either fails, stop: this interval does not apply.
3. Find $z_{\alpha/2}$ (two-sided) or z_α (one-sided) from Table 3.1.
4. Compute $\text{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$ and $E = z \times \text{SE}$.
5. Assemble $\hat{p} \pm E$ and report, usually in percent.

Sample size for a proportion

Setting the margin of error of (3.7) equal to a target E and solving for n gives the planning formula

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}) \qquad \text{(round up; use a planning guess for } \hat{p}\text{)}, \tag{3.8}$$

where $\hat{p}$ is a planning value: a pilot-study estimate, or last quarter's rate. When no guess at all is available, note that the product $p(1-p)$ is largest at $p = 0.5$,

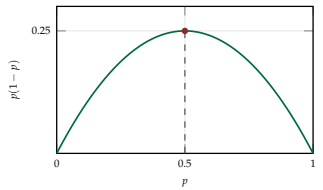

Figure 3.7. The variance factor $p(1-p)$ peaks at $p = 0.5$ with value 0.25. Planning with $\hat{p} = 0.5$ in (3.8) is therefore the worst case: it can never under-deliver on precision.

where it equals 0.25 (Figure 3.7). Substituting this worst case gives the *conservative* version

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \times 0.25, \tag{3.9}$$

which guarantees the target precision whatever the true p turns out to be, at the cost of possibly sampling more than necessary.

Example 3.5 — Signup conversion for a pricing page

Your startup redesigns its pricing page. Out of $n = 320$ visitors randomly routed to the new page, $x = 96$ signed up.

(a) Construct a 95% confidence interval for the true conversion rate p.

(b) The next experiment must estimate conversion to within $E = 0.02$. Using part (a)'s estimate as the planning value, how many visitors are needed?

(c) How many visitors are needed with *no* planning value at all?

Solution.

Step 1 — Estimate and check conditions.

$\hat{p} = 96/320 = 0.30$. Checks: $n\hat{p} = 96 \geq 10$ and $n(1-\hat{p}) = 224 \geq 10$, so Procedure 3.4 applies.

Step 2 — Part (a): the interval.

$$\text{SE} = \sqrt{\frac{0.30 \times 0.70}{320}} = \sqrt{0.0006563} = 0.02562,$$

$$E = 1.96 \times 0.02562 = 0.05021.$$

By (3.7), 0.30 ± 0.0502, so $\boxed{(0.2498,\ 0.3502)}$: we are 95% confident the true conversion rate is between about 25.0% and 35.0%.

Step 3 — Part (b): sample size with a planning value.

By (3.8) with $\hat{p} = 0.30$:

$$n = \left(\frac{1.96}{0.02}\right)^2 \times 0.30 \times 0.70 = 9604 \times 0.21 = 2016.8,$$

so $\boxed{n = 2017 \text{ visitors}}$.

Step 4 — Part (c): conservative sample size.

By (3.9):

$$n = 9604 \times 0.25 = 2401 \quad \Longrightarrow \quad \boxed{n = 2401 \text{ visitors}}.$$

The pilot estimate saves $2401 - 2017 = 384$ visitors of traffic. A rough planning value is worth real money; a wrong one, however, under-delivers precision, which is why the conservative figure is the safe default.

Startup lens. A/B testing platforms report exactly interval (3.7) on each variant. When the two variants' intervals overlap heavily, the experiment has not yet decided anything — resist the urge to ship the variant that is ahead today. Chapter 5 treats the two-variant comparison properly.

Reference. Montgomery & Runger, 6th ed., Section 8-6.

Prediction interval. An interval that will contain *one future observation* X_{n+1} with stated confidence; wider than the CI for μ because a single observation carries its own variability.

3.10 *Prediction Interval for a Single Future Observation*

A confidence interval locates the population *mean*. A different, very practical question is: where will the *next single observation* fall? The next container, the next truck, the next request. This is a *prediction interval* (PI), and it must be wider than the CI, for a reason worth internalizing: the error of predicting X_{n+1} by $\bar{X}$ has two independent parts, the sampling error of $\bar{X}$ (variance σ^2/n) and the inherent variability of the new observation itself (variance σ^2). The variances add:

$$\operatorname{Var}(X_{n+1} - \bar{X}) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right).$$

For a normal population with σ estimated by s, standardizing by this larger spread gives a t statistic with $n-1$ degrees of freedom, and the $100(1-\alpha)\%$ prediction interval is

$$\bar{x} - t_{\alpha/2,\,n-1}\, s\sqrt{1 + \tfrac{1}{n}} \;\le\; X_{n+1} \;\le\; \bar{x} + t_{\alpha/2,\,n-1}\, s\sqrt{1 + \tfrac{1}{n}}. \tag{3.10}$$

Compare the widths: the CI half-width shrinks like $s/\sqrt{n}$ and goes to zero as n grows, but the PI half-width tends to $t_{\alpha/2,\,n-1}\, s \approx z_{\alpha/2}\,\sigma$, which never goes to zero. No amount of data removes the randomness of a single future observation; more data only removes our uncertainty about where its distribution sits.

Insight. Why the "1" under the square root matters more than the "$1/n$". In (3.10) the term $1/n$ is the part of the width we can buy away with data, and the term 1 is the part we cannot. Already at $n = 25$ the factor is $\sqrt{1.04} \approx 1.02$: the PI is dominated by the population's own spread. This is the quantitative version of a distinction every engineer needs: knowing a process well is not the same as the process being repeatable.

Reference. Montgomery & Runger, 6th ed., Section 8-7.

Tolerance interval. An interval computed so that, with confidence $100(1-\alpha)\%$, it covers at least a stated proportion γ of the whole population.

3.11 *Tolerance Intervals*

The third and last member of the family answers a coverage question: give an interval that contains at least a stated proportion γ (say 95%) of the *entire population*, with stated confidence. If μ and σ were known exactly, the interval $\mu \pm 1.96\,\sigma$ would cover exactly 95% of a normal population. Since we only have $\bar{x}$ and s, the multiplier must be inflated to account for estimation error, giving the *tolerance interval*

$$\bar{x} - k\,s \;\le\; (\text{at least a fraction } \gamma \text{ of the population}) \;\le\; \bar{x} + k\,s, \tag{3.11}$$

where the factor $k = k(n,\ \gamma,\ 1-\alpha)$ is tabulated (Table 3.5). As $n \to \infty$, k decreases toward the known-parameter value $z_{(1-\gamma)/2}$ (e.g. 1.96 for $\gamma = 0.95$); for small n, k is much larger, because $\bar{x}$ and s are themselves uncertain.

	coverage γ (at 95% confidence)		
n	90%	95%	99%
10	2.839	3.379	4.433
15	2.480	2.954	3.878
20	2.310	2.752	3.615
25	2.208	2.631	3.457
30	2.140	2.549	3.350
50	1.996	2.379	3.126
∞	1.645	1.960	2.576

Table 3.5. Two-sided tolerance-interval factors k for a normal population at 95% confidence. The highlighted cell ($n = 25$, coverage 95%) is used in Example 3.6.

Keep the three interval types firmly separated, because they answer three different questions from the same $\bar{x}$ and s: the *confidence interval* locates the mean;

the *prediction interval* brackets one future observation; the *tolerance interval* brackets a stated fraction of all observations. Their widths ordinarily satisfy CI $<$ PI $<$ TI.

Example 3.6 — Truck turnaround at a distribution center

A distribution center measures the dock turnaround time of $n = 25$ trucks: $\bar{x} = 52.0$ minutes, $s = 6.0$ minutes, and the data pass a normality check.

(a) Construct a 95% prediction interval for the turnaround time of the *next* truck.

(b) Construct a tolerance interval that covers 95% of *all* truck turnarounds with 95% confidence.

(c) Compare both with the 95% CI for the mean, $52.0 \pm 2.064 \times 6.0/\sqrt{25} = (49.52,\ 54.48)$.

Solution.

Step 1 — Part (a): prediction interval.

With $\nu = 24$, $t_{0.025,\,24} = 2.064$ (Table 3.3). By (3.10),

$$s\sqrt{1 + \tfrac{1}{25}} = 6.0\sqrt{1.04} = 6.0 \times 1.020 = 6.119, \qquad E = 2.064 \times 6.119 = 12.63,$$

so the PI is $52.0 \pm 12.63 = \boxed{(39.37,\ 64.63) \text{ minutes}}$.

Step 2 — Part (b): tolerance interval.

From Table 3.5, $k = 2.631$ for $n = 25$, coverage 95%, confidence 95%. By (3.11),

$$52.0 \pm 2.631 \times 6.0 = 52.0 \pm 15.79 = \boxed{(36.21,\ 67.79) \text{ minutes}}.$$

Step 3 — Part (c): compare.

Same data, three intervals: CI $(49.52,\ 54.48)$, half-width 2.5 minutes; PI $(39.37,\ 64.63)$, half-width 12.6; TI $(36.21,\ 67.79)$, half-width 15.8. The CI is a statement about the average truck; the PI is a promise about the next truck; the TI is a promise about 95% of all trucks. Quoting the CI when a customer asks ''when will *my* truck be done?'' understates the risk by a factor of five.

Engineering judgment. Dock scheduling, buffer sizing, and staffing all consume the PI and TI, not the CI. A planner who books dock slots every 54 minutes because "the CI says the mean is below 54.5" has confused the average truck with every truck; the tolerance interval says slots must absorb turnarounds up to about 68 minutes.

Reference. Montgomery & Runger, 6th ed., Sections 8-1 to 8-7.

3.12 *Choosing the Right Interval*

Exam problems and real problems alike begin the same way: read the scenario and identify the parameter, the available information, and the question being asked. The computation is the easy part; the selection is where marks and decisions are lost.

Procedure 3.5 — Choosing the interval.

1. Identify the parameter the question is about: a mean μ, a variance σ^2 (or standard deviation σ), a proportion p, a single future observation, or a fraction of the population.
2. For a mean: if σ is known (stated, or from long stable history), use the z interval (3.1); if σ is unknown and estimated by s, use the t interval (3.5) with $\nu = n - 1$.
3. For a variance or standard deviation: use the chi-square interval (3.6); verify normality first.
4. For a proportion: check $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$, then use (3.7).
5. For one future observation: use the prediction interval (3.10). For a stated fraction of the whole population: use the tolerance interval (3.11).
6. Decide two-sided versus one-sided from the wording ("between" versus "at least" / "at most"), and split α accordingly ($z_{\alpha/2}$ or $t_{\alpha/2,\nu}$ versus z_α or $t_{\alpha,\nu}$).

Question is about	Information	Distribution	Interval
mean μ	σ known	z	(3.1)
mean μ	σ unknown (s)	t, $\nu = n-1$	(3.5)
variance σ^2, sd σ	normal data	χ^2, $\nu = n-1$	(3.6)
proportion p	$n\hat{p}$, $n(1-\hat{p}) \geq 10$	z	(3.7)
next observation X_{n+1}	normal data	t, $\nu = n-1$	(3.10)
fraction γ of population	normal data	factor k table	(3.11)

Table 3.6. Interval selection at a glance. The wording of the question picks the row; the available information picks the critical value.

Lecture 7 starts here — *Practice: interval estimation*

Lecture 7. This section is covered by Lecture 7.

Reference. Montgomery & Runger, 6th ed., Chapter 8 exercises.

3.13 *Worked Practice*

This section replays the full toolkit on fresh problems, in the same style as the in-class practice session. For each problem, run Procedure 3.5 first — name the parameter, name the distribution — and only then compute. The three examples below cover all six problem types of the practice lecture: a z interval and a sample-size plan; a t interval, a one-sided t bound, and a variance interval; a proportion interval and a proportion sample-size plan.

Example 3.7 — Delivery route times: z interval and sample size

A regional carrier samples $n = 64$ runs of a delivery route and finds $\bar{x} = 120$ minutes. Route-time variability is stable across seasons, with known $\sigma = 16$ minutes.

(a) Construct a 95% confidence interval for the true mean route time.

(b) The planning team wants the mean pinned down to within $E = 2$ minutes at 95% confidence. How many runs are needed?

Solution.

Step 1 — Select the interval.

The parameter is a mean and σ is known: z interval, Procedure 3.1.

Step 2 — Part (a).

SE $= 16/\sqrt{64} = 16/8 = 2.000$; $E = 1.96 \times 2.000 = 3.920$;

$$120 \pm 3.92 \quad \Longrightarrow \quad \boxed{(116.08,\ 123.92) \text{ minutes}}.$$

Step 3 — Part (b).

By (3.4),

$$n = \left(\frac{1.96 \times 16}{2}\right)^2 = (15.68)^2 = 245.9 \quad \Longrightarrow \quad \boxed{n = 246 \text{ runs}}.$$

Nearly four times the current sample — consistent with Figure 3.4, since the target margin (2) is about half the current one (3.92).

Example 3.8 — Sensor-fusion localization error: t, one-sided t, and variance

An AV team measures the localization error of a sensor-fusion stack on $n = 16$ calibration runs: $\bar{x} = 85$ mm, $s = 20$ mm, data approximately normal. No historical σ exists for this stack version.

(a) Construct a 95% confidence interval for the mean error μ.

(b) The safety case needs only a guarantee that the mean error is not too large: give a 95% upper confidence bound for μ.

(c) Construct a 95% confidence interval for σ^2 and for σ.

Solution.

Step 1 — Select the interval.

Mean with σ unknown: t interval with $\nu = 15$ (Procedure 3.2). Variance: chi-square with $\nu = 15$ (Procedure 3.3).

Step 2 — Part (a).

SE $= 20/\sqrt{16} = 5.000$; $t_{0.025,\,15} = 2.131$ (Table 3.3); $E = 2.131 \times 5.000 =$

10.66;

$$85 \pm 10.66 \quad \Longrightarrow \quad \boxed{(74.35,\ 95.66)\ \text{mm}}.$$

Step 3 — Part (b).

One-sided at 95% uses $t_{0.05,\,15} = 1.753$, not $t_{0.025,\,15}$:

$$\mu \ \leq\ 85 + 1.753 \times 5.000 = 85 + 8.765 = \boxed{93.77\ \text{mm}}.$$

Step 4 — Part (c).

$s^2 = 400$, $(n-1)s^2 = 15 \times 400 = 6000$. From Table 3.4, row $\nu = 15$: $\chi^2_{0.025,\,15} = 27.488$ and $\chi^2_{0.975,\,15} = 6.262$. By (3.6),

$$\frac{6000}{27.488} \leq \sigma^2 \leq \frac{6000}{6.262} \quad \Longrightarrow \quad \boxed{(218.3,\ 958.2)\ \text{mm}^2},$$

and taking square roots, $14.77 \leq \sigma \leq 30.95$ mm. The safety case should quote part (b) for the mean and the upper end of part (c) for the spread: both directions of ''how bad could it be'' are answered by one sample.

Example 3.9 — Trial-to-paid conversion: proportion interval and sample size

Of $n = 400$ users who started a free trial of your product last month, $x = 60$ converted to a paid plan.

(a) Construct a 95% confidence interval for the true conversion rate p.

(b) For next quarter's pricing experiment the team wants $E = 0.02$ at 95% confidence. Find the required sample size using $\hat{p}$ from part (a), and with no planning value.

Solution.

Step 1 — Select the interval and check conditions.

$\hat{p} = 60/400 = 0.15$; $n\hat{p} = 60 \geq 10$ and $n(1-\hat{p}) = 340 \geq 10$, so Procedure 3.4 applies.

Step 2 — Part (a).

$$\text{SE} = \sqrt{\frac{0.15 \times 0.85}{400}} = \sqrt{0.0003188} = 0.01785,$$
$$E = 1.96 \times 0.01785 = 0.03499.$$

By (3.7), 0.15 ± 0.035, so $\boxed{(0.115,\ 0.185)}$ — between 11.5% and 18.5%.

Step 3 — Part (b).

With the planning value $\hat{p} = 0.15$, by (3.8):

$$n = \left(\frac{1.96}{0.02}\right)^2 \times 0.15 \times 0.85 = 9604 \times 0.1275 = 1224.5,$$

so $\boxed{n = 1225 \text{ trial users}}$; with no planning value, by (3.9): $n = 9604 \times 0.25 = 2401$. If last month's rate is trustworthy, the pilot value nearly halves the required traffic.

Common mistakes to avoid

The recurring errors in quizzes and exams are few and predictable. Check your own work against this list before submitting anything.

- Using z when σ is unknown. If the problem gives s, the distribution is t with $\nu = n - 1$.
- Using z or t for a variance. Variance intervals use chi-square, with *two* critical values.
- Forgetting $\nu = n - 1$ for both t and chi-square.
- Swapping the chi-square bounds. The lower CI endpoint takes the *larger* chi-square value in its denominator.
- Skipping the checks $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ for a proportion.
- Rounding a sample size down. Always up.
- Confusing α with $\alpha/2$: two-sided intervals split α; one-sided bounds do not.
- Interpreting a CI as a probability statement about μ, or as a range for individual observations.

Safety lens. Verification work inherits every one of these failure modes at scale: an evaluation pipeline that hard-codes 1.96 silently commits the first mistake on every small-sample benchmark it processes. Auditing the statistics inside tools is part of auditing the model.

Chapter Summary

A confidence interval converts a point estimate and its standard error into a range with a stated long-run success rate. Every interval in this chapter has the same anatomy — estimate ± critical value × standard error — and differs only in which sampling distribution supplies the critical value: z when σ is known and for large-sample proportions, t with $n-1$ degrees of freedom when σ is estimated, chi-square for variances. Confidence describes the procedure, never a single computed interval; the parameter is fixed, the interval is what varies from sample to sample. One-sided bounds spend the whole error budget in one tail (z_α, $t_{\alpha,\nu}$). Sample-size formulas invert the margin of error and are always rounded up. Prediction intervals bracket one future observation and tolerance intervals bracket a stated fraction of the population; both are wider than the CI, and neither shrinks to zero with more data. Table 3.7 collects the formulas.

Target	Interval	Equation
mean, σ known	$\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$	(3.1)
one-sided z bounds	$\mu \geq \bar{x} - z_\alpha \sigma/\sqrt{n}$; $\quad \mu \leq \bar{x} + z_\alpha \sigma/\sqrt{n}$	(3.2)
margin of error	$E = z_{\alpha/2}\,\sigma/\sqrt{n}$	(3.3)
sample size, mean	$n = (z_{\alpha/2}\,\sigma/E)^2$, round up	(3.4)
mean, σ unknown	$\bar{x} \pm t_{\alpha/2,n-1}\,s/\sqrt{n}$	(3.5)
variance	$(n-1)s^2/\chi^2_{\alpha/2,n-1} \leq \sigma^2 \leq (n-1)s^2/\chi^2_{1-\alpha/2,n-1}$	(3.6)
proportion	$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$	(3.7)
sample size, proportion	$n = (z_{\alpha/2}/E)^2\,\hat{p}(1-\hat{p})$; worst case $\hat{p} = 0.5$	(3.8), (3.9)
next observation	$\bar{x} \pm t_{\alpha/2,n-1}\,s\sqrt{1+1/n}$	(3.10)
population fraction	$\bar{x} \pm k\,s$, k from Table 3.5	(3.11)

Table 3.7. Key formulas of Chapter 3 at a glance. All two-sided forms have one-sided versions with z_α or $t_{\alpha,n-1}$.

Exercises

z intervals

Exercise 3.1. A rail operator weighs $n = 49$ randomly selected loaded freight cars on a certified scale and finds $\bar{x} = 88.4$ tonnes. Loading variability is well established at $\sigma = 5.6$ tonnes. Construct a 90% confidence interval for the true mean car load.

Exercise 3.2. An AV fleet logs the distance driven between disengagements on $n = 25$ randomly chosen shifts, giving $\bar{x} = 310$ km. Fleet history supports $\sigma = 45$ km. The safety report must state a distance the mean is *at least*, with 99% confidence. Compute the bound and write the reporting sentence.

Exercise 3.3. A team computes a 95% confidence interval $(41.2,\ 46.8)$ for a mean μ. Which one of the following statements is a correct reading of it?

(A) There is a 95% probability that μ lies in $(41.2,\ 46.8)$.

(B) 95% of all observations in the population lie in $(41.2,\ 46.8)$.

(C) If the sampling were repeated many times, about 95% of the intervals built this way would contain μ.

(D) 95% of future sample means will lie in $(41.2,\ 46.8)$.

t intervals

Exercise 3.4. Your startup times how long pilot customers take to respond to a feature proposal: $n = 9$ responses give $\bar{x} = 4.8$ hours and $s = 1.2$ hours (approximately normal). Construct a 95% confidence interval for the mean response time.

Exercise 3.5. A warehouse samples $n = 25$ forklift shifts and finds mean idle time $\bar{x} = 2.4$ hours with $s = 0.5$ hours. Management wants a 95% *upper* confidence bound on mean idle time. Compute it and interpret.

Variance and proportion intervals

Exercise 3.6. An AI safety team measures the response-time consistency of a model serving endpoint: $n = 20$ queries give $s^2 = 144$ ms^2 (times approximately normal). Construct a 95% confidence interval for σ^2, and from it one for σ.

Exercise 3.7. Of $n = 300$ customers who joined your subscription product in January, $x = 45$ canceled within their first month. Construct a 90% confidence interval for the true first-month churn rate.

Exercise 3.8. A red team runs $n = 30$ adversarial test cases against a guardrail and observes $x = 4$ bypasses. A teammate plugs the numbers into (3.7) and reports a 95% CI of $(0.012,\ 0.255)$ for the bypass rate. Is this interval trustworthy? Explain what to check, what fails, and what to do instead.

Sample size

Exercise 3.9. Crane cycle times at a container berth have $\sigma = 0.9$ minutes from long operating history. How many cycles must be timed so that a 95% confidence interval for the mean cycle time has margin of error at most 0.2 minutes?

Exercise 3.10. Your team must estimate the signup conversion rate of a new landing page to within $E = 0.04$ at 95% confidence. (a) How many visitors are needed if nothing is known about the rate? (b) How many if a soft launch suggests $p \approx 0.2$?

Exam-style problems

Exercise 3.11. A red team scores the severity of $n = 10$ discovered defects in a perception model on a 0–10 scale: $\bar{x} = 6.9$, $s = 1.5$, data approximately normal.

1. **[3 pt]** Which distribution must be used for a CI on the mean severity, and why? State the degrees of freedom.
2. **[5 pt]** Construct the 95% confidence interval for the mean severity μ.
3. **[4 pt]** Construct the 95% confidence interval for σ^2.
4. **[3 pt]** The team lead writes: ''there is a 95% chance the true mean severity is in the interval from (b).'' Rewrite this sentence correctly.

Exercise 3.12. Delivery lateness on a parcel network is approximately normal with known $\sigma = 4.2$ minutes. A sample of $n = 49$ deliveries gives $\bar{x} = 12.6$ minutes late.

1. **[4 pt]** Construct a 95% two-sided confidence interval for the mean lateness.
2. **[4 pt]** Construct a 95% one-sided upper confidence bound for the mean lateness.

3. **[4 pt]** How many deliveries are needed to estimate the mean lateness to within 0.8 minutes at 95% confidence?

4. **[3 pt]** A manager reads part (a) and announces: ''95% of parcels are between 11.4 and 13.8 minutes late.'' Explain the error and name the interval that would support such a claim.

Assessing outside recommendations

Exercise 3.13. Your team asks an AI assistant for error bars on a model's mean inference latency from $n = 12$ measurements with $\bar{x} = 142$ ms and $s = 8.0$ ms. The assistant answers: ''95% CI $= 142 \pm 1.96 \times 8.0/\sqrt{12} = 142 \pm 4.53 = (137.47,\ 146.53)$ ms.'' Find the error, explain why it matters at this sample size, and give the corrected interval.

Exercise 3.14. A consultant's report on port operations states: ''The 95% confidence interval for mean container dwell is $(57.8,\ 64.8)$ hours. Therefore there is a 95% probability that the true mean lies in this interval, and 95% of containers dwell between 57.8 and 64.8 hours.'' Identify *both* errors, explain each in one or two sentences, and rewrite the paragraph correctly.

In-Class Activity 3.1: How Wide Is the Truth?

Week 3 — Lecture 5

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

Your startup's activation funnel is under review before a board meeting. Activation time (signup to first successful use) has a stable known $\sigma = 4$ minutes from telemetry. A random sample of $n = 25$ new users gives $\bar{x} = 18.2$ minutes.

Part 1 — Build it

Construct the 95% confidence interval for the true mean activation time. Show the standard error, the critical value, and the margin of error as separate steps.

Part 2 — Say it right

Your co-founder drafts three sentences for the board deck. Mark each one CORRECT or WRONG and fix the wrong ones: (i) ''We are 95% confident mean activation is between these bounds.'' (ii) ''There is a 95% probability that mean activation is in this range.'' (iii) ''95% of our users activate within this range.''

Part 3 — Buy precision

The board wants the mean pinned to within ± 0.8 minutes at the same confidence. How many
next sample contain? What does this say about the cost of halving a margin of error?

Exit check

In one sentence: what exactly is the ''95%'' a property of — the interval you computed, or som

In-Class Activity 3.2: Choose Your Distribution

Week 3 — Lecture 6

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

Your AI safety team is preparing an evaluation report and must pick the right interval for each measurement.

Part 1 — Match the tool

For each scenario, name the distribution (z, t, or χ^2) and the degrees of freedom if any: (i) mean latency, $n = 9$ measurements, only s available; (ii) mean accuracy, $n = 200$ prompts, σ known from a fixed benchmark harness; (iii) variance of latency, $n = 9$, data normal; (iv) fraction of prompts wrongly refused, $n = 160$.

Part 2 — Small-sample mean

The $n = 9$ latency measurements give $\bar{x} = 210$ ms and $s = 30$ ms. Construct the 95% confidence interval for the mean latency ($t_{0.025,\,8} = 2.306$). Then state what critical value a careless analyst would have used instead, and whether their interval would be too wide or too narrow.

Part 3 — The refusal rate

Out of $n = 160$ benign prompts, 24 were wrongly refused by the guardrail. Check the large-sa
then construct the 95% confidence interval for the true false-refusal rate.

Exit check

One sentence: why are t critical values larger than z critical values at the same confidence leve

In-Class Activity 3.3: The Dashboard Audit

Week 4 — Lecture 7

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

A port authority's new analytics dashboard shows uncertainty ranges next to every metric. Your team is hired to audit three of its claims.

Part 1 — Name the interval

For each dashboard claim, state which interval from Procedure 3.5 it needs (interval for a mean with z or t, variance, proportion, prediction, or tolerance): (i) ''mean crane cycle time is 2.05 ± 0.11 min'' (based on $n = 10$ cycles, no historical σ); (ii) ''the fraction of late gate appointments is $8\% \pm 2\%$'' (from $n = 450$ appointments); (iii) ''the next vessel's unloading time will be between 14.1 and 19.3 hours.''

Part 2 — Recompute one

For claim (i), the $n = 10$ cycles give $s = 0.6$ min. Construct the 95% confidence interval for the *standard deviation* σ of cycle times ($\chi^2_{0.025,\,9} = 19.023$, $\chi^2_{0.975,\,9} = 2.700$). Work through σ^2 first.

Part 3 — Challenge the vendor

The dashboard vendor says: ''To make every interval narrower, we switched all confidence lev 99%.'' Explain in two sentences what this change actually does to the intervals, and what the c to narrow them are.

Exit check

One sentence: of the confidence interval, prediction interval, and tolerance interval computed sample, which is widest and why?