4 *Tests of Hypotheses: One Sample*

Chapter 3 answered the question ''what is the parameter?'' with an interval of plausible values. This chapter answers a different question: ''is a specific *claim* about the parameter supported by the data?'' A supplier claims the mean is 115. A vendor claims the failure rate is below 2%. A teammate claims the new process reduced the variance. Each claim is either rejected or not rejected, using the same sampling distributions that built the confidence intervals. The chapter is organized in the following order: a short recap of the interval toolkit; statistical hypotheses and the two types of decision error; test statistics and rejection regions; p-values, including how to compute or bound them from the printed tables; the seven-step testing procedure; the four one-sample tests (z for a mean, t for a mean, χ^2 for a variance, z for a proportion); the duality between tests and confidence intervals; type II error, power, and sample size; and a practice section that runs the full procedure on three problems side by side.

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

Q1. A test of $H_0 : \mu = 50$ rejects H_0 at $\alpha = 0.05$. Does this *prove* that $\mu \neq 50$?

Q2. An A/B test on a signup page reports ''p-value $= 0.03$.'' What does the 0.03 measure?

Q3. You tighten a test from $\alpha = 0.05$ to $\alpha = 0.01$ without changing the sample size. What happens to the probability β of a type II error?

Answers.

Q1. *No.* Rejection means the observed data would be unlikely (probability at most 0.05) if μ were really 50. There is still up to a 5% chance that H_0 is true and we rejected

it anyway—a type I error. Hypothesis tests weigh evidence; they never prove.

Q2. The probability of observing a test statistic at least as extreme as the one actually observed, *computed assuming H_0 is true*. It is not the probability that H_0 is true. Since $0.03 < 0.05$, the result is significant at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Q3. *β increases.* Demanding stronger evidence before rejecting makes false alarms rarer but missed detections more common. For fixed n, the two error probabilities move in opposite directions; only a larger sample reduces both.

Lecture 8 starts here — *From intervals to tests*

Lecture 8. This section is covered by Lecture 8.

Reference. Montgomery & Runger, 6th ed., Sections 8.1–8.7 and 9.1.

4.1 From Intervals to Tests: The Toolkit So Far

Chapter 3 built one tool for each parameter: an interval of plausible values, at a stated confidence level, computed from one sample. Table 4.1 collects the toolkit in one place, because every test in this chapter reuses one of these sampling distributions. The pattern to notice: each interval has the form *estimate* $\pm$ (*table value*) $\times$ (*standard error*), and the table value comes from the sampling distribution of the estimator—standard normal, t, or chi-square.

Parameter	Conditions	Interval	Equation
μ	σ known	$\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$	(3.1)
μ, one-sided bound	σ known	$\bar{x} + z_{\alpha}\,\sigma/\sqrt{n}$ or $\bar{x} - z_{\alpha}\,\sigma/\sqrt{n}$	(3.2)
error bound	σ known	$E = z_{\alpha/2}\,\sigma/\sqrt{n}$	(3.3)
n for target E	σ known	$n = (z_{\alpha/2}\,\sigma/E)^2$	(3.4)
μ	σ unknown	$\bar{x} \pm t_{\alpha/2,n-1}\,s/\sqrt{n}$	(3.5)
σ^2	normal population	$\dfrac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$	(3.6)
p	large n	$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$	(3.7)
n for proportion	large n	$n = (z_{\alpha/2}/E)^2\,p(1-p)$	(3.8)
one future observation	normal population	$\bar{x} \pm t_{\alpha/2,n-1}\,s\sqrt{1+1/n}$	(3.10)
proportion of population	normal population	$\bar{x} \pm k\,s$ (tolerance factor k)	(3.11)

Table 4.1. The interval toolkit of Chapter 3 at a glance. Every test in this chapter reuses one of these sampling distributions.

A confidence interval answers an *estimation* question: ''what values of μ are

plausible?'' A hypothesis test answers a *decision* question: ''is the specific value μ_0 plausible?'' The two questions use the same standard error, the same table values, and the same distributional assumptions. What changes is the direction of the logic. An interval starts from the data and reports a range. A test starts from a claimed value, computes how far the data fall from that claim, and decides whether the distance is too large to blame on sampling variability.

Here is the situation that motivates the whole chapter. A container terminal operator states in a service agreement that the mean container dwell time is 115 hours. Your logistics team measures $n = 64$ randomly selected containers and finds $\bar{x} = 120$ hours; long experience with this terminal justifies treating $\sigma = 16$ hours as known. The 95% confidence interval from (3.1) is

$$120 \pm 1.96 \times \frac{16}{\sqrt{64}} = 120 \pm 3.92 = (116.08,\ 123.92) \text{ hours.}$$

The claimed value 115 lies *outside* this interval. Every value inside the interval is compatible with the data; 115 is not. The claim looks false—but how do we turn ''looks false'' into a defensible decision with a controlled error rate? That is exactly what a hypothesis test does.

Startup lens. Every A/B test your startup will ever run is a hypothesis test: H_0 says the new variant changed nothing, and the data must argue against that default. Learning the mechanics here is learning the mechanics of product experimentation.

Reference. Montgomery & Runger, 6th ed., Section 9.1.1.

4.2 Statistical Hypotheses

A *statistical hypothesis* is a statement about a population parameter. It is a statement about μ, σ^2, or p—never about the sample. ''$\bar{x} = 120$'' is not a hypothesis; it is a fact we computed. ''$\mu = 115$'' is a hypothesis: it may be true or false, and we will never know for certain, because we never observe the whole population.

Every test involves a pair of competing hypotheses:

$$H_0\colon\ \theta = \theta_0 \qquad \text{versus} \qquad H_1\colon\ \theta \neq \theta_0 \quad (\text{or } \theta > \theta_0,\ \text{or } \theta < \theta_0), \tag{4.1}$$

where θ is the parameter of interest (μ, σ^2, or p) and θ_0 is a specific claimed value. The *null hypothesis* H_0 is the default or status quo: it is what we assume true at the start, and it always contains the equality sign. The *alternative hypothesis* H_1 is the research claim: it is what we are trying to establish, and it carries $\neq$, $>$, or $<$. The burden of proof is on H_1. The data must provide convincing evidence against H_0 before we abandon it.

Statistical hypothesis. A statement about the parameters of one or more populations, such as $\mu = 115$ or $p < 0.02$.

Null hypothesis H_0. The default claim, assumed true until the data argue otherwise. It always contains the equality.

Alternative hypothesis H_1. The claim we require evidence for. It contains $\neq$, $>$, or $<$.

Insight. Why does H_0 get the benefit of the doubt? Because the test is designed around controlling the probability of *falsely* abandoning the default. Think of a trial: the defendant is presumed innocent (H_0), the prosecution must present evidence, and the court convicts only when the evidence is strong enough. A weak case does not prove innocence—it only fails to prove guilt. In exactly the same way, a test never proves H_0; it either rejects H_0 or *fails to reject* it. Please pay attention here, because this is a common trap: we never say "accept H_0."

Two-sided test. A test with $H_1 : \theta \neq \theta_0$; departures in either direction count as evidence.

One-sided test. A test with $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$; only one direction counts as evidence.

The form of H_1 decides the *sidedness* of the test, and the sidedness must come from the question being asked, not from the data.

- **Two-sided (two-tailed):** $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Used when a departure in either direction matters: "is the dwell time *different* from 115 hours?"
- **Upper-tailed (right-tailed):** $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. Used when the claim to establish is "greater than": "does the mean exceed 200 km?"
- **Lower-tailed (left-tailed):** $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$. Used when the claim to establish is "less than": "is the pass rate below 95%?"

Engineering judgment. Sidedness is an engineering decision made before data collection. A pipeline pressure that is too high bursts pipes and one that is too low starves the pumps—two-sided. A safety limit that must not be exceeded is one-sided by design. Write the question down first; the inequality in H_1 should copy the inequality in the engineering requirement.

For a one-sided pair we will often write H_0 with an inequality ($\mu \leq \mu_0$), but the test is always carried out *at the boundary value* $\mu = \mu_0$, because that is the hardest case for H_1 to beat. If the data can reject $\mu = \mu_0$ in favor of $\mu > \mu_0$, they can reject every smaller value of μ as well.

Example 4.1 — Framing hypotheses in three settings

Write H_0 and H_1, state the parameter, and state the sidedness for each claim.

1. An AV software vendor claims its perception model's disengagement rate is below 2 events per 10,000 km, i.e., the proportion of 1-km segments with a disengagement is below $p_0 = 0.0002$.
2. A freight forwarder's contract specifies that the mean truck route completion time equals 150 minutes; both faster and slower schedules disrupt the loading docks.
3. Your startup redesigned its onboarding flow and wants to show that the

mean time users spend in the first session now exceeds 12.5 minutes.

Solution.

Claim 1 — AI safety.

The parameter is a proportion p. The vendor's claim is the thing that needs proof, so it goes in H_1:

$$H_0 \colon p \geq 0.0002 \qquad \text{versus} \qquad H_1 \colon p < 0.0002,$$

a lower-tailed test. If we instead put the vendor's claim in H_0, a small or noisy sample would "support" the vendor by default—the burden of proof would be backwards.

Claim 2 — supply chains.

The parameter is a mean μ, and departures in either direction matter:

$$H_0 \colon \mu = 150 \qquad \text{versus} \qquad H_1 \colon \mu \neq 150,$$

a two-sided test.

Claim 3 — your startup.

The parameter is a mean μ, and the claim to establish is "greater than":

$$H_0 \colon \mu \leq 12.5 \qquad \text{versus} \qquad H_1 \colon \mu > 12.5,$$

an upper-tailed test, carried out at the boundary $\mu_0 = 12.5$.

Safety lens. In AI safety testing the null hypothesis should be "the system is not safe enough" whenever a release decision hangs on the test. Then a type I error (shipping an unsafe system) is the error whose probability the test explicitly controls at α. Framing the hypotheses the other way around silently shifts the controlled error to the harmless one.

Reference. Montgomery & Runger, 6th ed., Sections 9.1.2–9.1.3.

4.3 *Decision Errors: Type I and Type II*

A test ends in one of two decisions—reject H_0 or fail to reject H_0—and reality is one of two states—H_0 true or H_0 false. That gives four outcomes, two correct and two wrong, shown in Table 4.2.

	Truth (unknown)	
Decision	H_0 is true	H_0 is false
Reject H_0	type I error (prob. α)	correct (prob. $1-\beta$, the power)
Fail to reject H_0	correct (prob. $1-\alpha$)	type II error (prob. β)

Type I error. Rejecting H_0 when H_0 is true — a false alarm. Its probability is α.

Type II error. Failing to reject H_0 when H_0 is false — a missed detection. Its probability is β.

Table 4.2. The four possible outcomes of a hypothesis test. The two error probabilities α and β are defined in (4.2).

The two error probabilities have standard names:

Significance level. The value α chosen before the test: the largest type-I-error probability the analyst is willing to tolerate.

Power. $1 - \beta$, the probability of correctly rejecting a false H_0 — the test's ability to detect a real effect.

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ true}), \qquad \beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H \tag{4.2}$$

A type I error is a *false alarm*: the process was fine and we raised a flag anyway. A type II error is a *missed detection*: something changed and we did not notice. The two errors usually have very different costs. Rejecting a good production batch wastes rework; accepting a bad batch causes recalls and reputation damage. Deciding which error is more expensive is an engineering and business judgment, not a statistical one.

The analyst chooses α *before* seeing the data. This chosen value is called the *significance level*. Common choices are $\alpha = 0.10$ (lenient), $\alpha = 0.05$ (the usual default), and $\alpha = 0.01$ (strict). You met these numbers in Chapter 3: a test at $\alpha = 0.05$ corresponds to a 95% confidence interval, a connection made precise in Section 4.11. The complement of β is the *power* $1 - \beta$: the probability that the test detects a departure from H_0 when one truly exists.

Figure 4.1 shows both errors on one picture for the dwell time problem. The green curve is the sampling distribution of $\bar{X}$ *if* H_0 *is true* ($\mu = 115$, standard error $\sigma/\sqrt{n} = 16/8 = 2$). The test rejects when $\bar{x}$ falls outside $115 \pm 1.96 \times 2 = (111.08,\ 118.92)$; the two maroon tails under the green curve have total area $\alpha = 0.05$. The maroon curve is the sampling distribution if the truth is actually $\mu = 119$. The gold region is the part of the maroon curve that still lands *inside* the fail-to-reject interval: its area is β, the probability of missing the shift. The two curves overlap, and that overlap is the whole difficulty of testing: with $n = 64$ the data cannot cleanly separate $\mu = 115$ from $\mu = 119$.

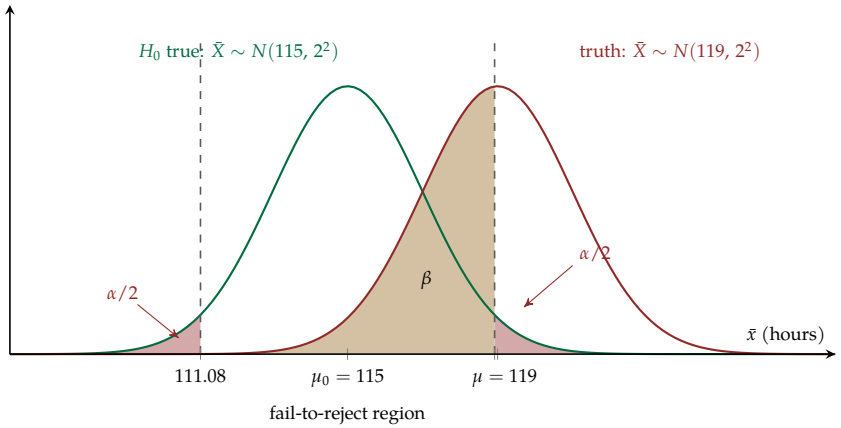

Figure 4.1. Type I and type II errors for the dwell time test ($n = 64$, $\sigma = 16$, $\alpha = 0.05$). The maroon tails under the H_0 curve have total area α; the gold area under the true-mean curve is β, computed in Example 4.7.

Three facts about α and β govern every test design. First, for a fixed sample size, they trade off: shrinking the rejection region to lower α necessarily enlarges the fail-to-reject region and raises β. Second, β is not a single number—it depends on the true parameter value. A large shift ($\mu = 130$, say) is easy to detect and has small β; a small shift ($\mu = 116$) hides inside the overlap and has large β. Third, the only way to reduce both errors at once is to collect more evidence: increasing n narrows both curves in Figure 4.1, shrinking the overlap. Section 4.12 turns these facts into computations.

Engineering judgment. Regulators and standards fix α because the false-alarm rate is what the test's user can control by design. Nobody fixes β in the same breath, because β depends on the unknown truth. Good engineering practice is to pick the smallest shift that matters operationally and size the sample so that this shift is detected with high power.

Reference. Montgomery & Runger, 6th ed., Sections 9.1.2 and 9.2.1.

4.4 *Test Statistics and the Rejection Region*

The decision needs a measuring stick. A *test statistic* converts the sample into a single number that measures the distance between what we observed and what H_0 predicts, expressed in standard-error units. For a mean with σ known, the test statistic is

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}, \tag{4.3}$$

which is exactly the CLT standardization (2.8) with the claimed value μ_0 substituted for μ. If H_0 is true, Z_0 follows the standard normal distribution, so values like 0.4 or -1.1 are ordinary and values like 3.2 are rare. Z_0 answers the question: *how many standard errors does $\bar{x}$ sit from the claimed mean?*

Test statistic. A single number, computed from the sample, that measures how far the data fall from what H_0 predicts, in standard-error units.

For the dwell time data,

$$z_0 = \frac{120 - 115}{16/\sqrt{64}} = \frac{5}{2} = 2.50.$$

Rejection region. The set of test-statistic values that lead to rejecting H_0; its probability under H_0 equals α. Also called the critical region.

Critical values. The boundaries of the rejection region, read from the distribution table at the chosen α.

The observed mean sits 2.5 standard errors above the claim. Is that far enough to reject? We need a rule decided in advance.

The *rejection region* (or critical region) is the set of test-statistic values considered too extreme to blame on chance, and it is constructed so that its total probability under H_0 equals α. Its boundaries are the *critical values*. Where the region sits depends on the sidedness of H_1, as Figure 4.2 shows:

- **Two-sided** ($H_1 : \mu \neq \mu_0$): reject H_0 if $|Z_0| > z_{\alpha/2}$. The area α is split evenly between the two tails.
- **Upper-tailed** ($H_1 : \mu > \mu_0$): reject H_0 if $Z_0 > z_\alpha$. All of α sits in the right tail.
- **Lower-tailed** ($H_1 : \mu < \mu_0$): reject H_0 if $Z_0 < -z_\alpha$. All of α sits in the left tail.

The reject direction always matches the direction of H_1. Please pay attention here, because this is a common trap: at the same $\alpha = 0.05$, the two-sided critical value is $z_{0.025} = 1.96$ but the one-sided critical value is $z_{0.05} = 1.645$. Mixing them up changes conclusions for any $|z_0|$ between 1.645 and 1.96.

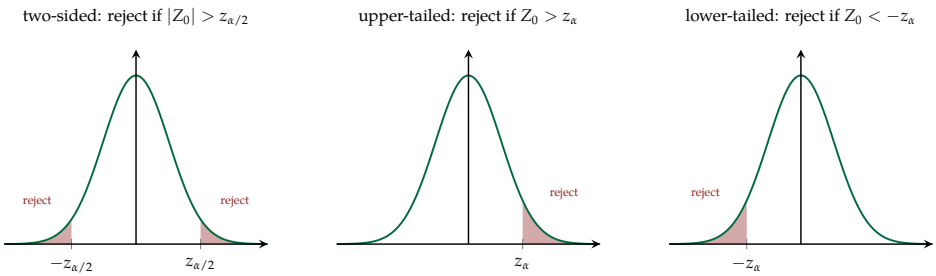

Figure 4.2. Rejection regions for the z test at significance level α. The shaded area is α in every panel; only its placement changes with the direction of H_1.

Example 4.2 — Testing the terminal operator's claim

A container terminal operator claims the mean container dwell time is 115 hours. Your team measures $n = 64$ containers, obtaining $\bar{x} = 120$ hours; take $\sigma = 16$ hours as known. Test the claim at $\alpha = 0.05$.

Solution.

Step 1 — Hypotheses.

Both shorter and longer dwell times matter for berth planning, so the test is two-sided:

$$H_0 : \mu = 115 \qquad \text{versus} \qquad H_1 : \mu \neq 115, \qquad \alpha = 0.05.$$

Step 2 — Test statistic and rejection region.

σ is known, so use Z_0 from (4.3). Reject H_0 if $|Z_0| > z_{0.025} = 1.96$.

Step 3 — Compute.

$$z_0 = \frac{120 - 115}{16/\sqrt{64}} = \frac{5}{2} = 2.50.$$

Step 4 — Decide and conclude.

Since $|2.50| = 2.50 > 1.96$, the statistic falls in the rejection region, so we $\boxed{\text{reject } H_0}$. There is sufficient evidence at the 5% significance level that the true mean dwell time is not 115 hours. The sample mean of 120 suggests the operator is understating dwell time, which matters for every downstream delivery promise built on it.

Notice how this agrees with the confidence interval computed in Section 4.1: the 95% CI $(116.08,\ 123.92)$ excluded 115, and the $\alpha = 0.05$ test rejected $\mu = 115$. This is not a coincidence; it is a theorem (Section 4.11).

Lecture 9 starts here — *Hypothesis testing*

Lecture 9. This section is covered by Lecture 9.

Reference. Montgomery & Runger, 6th ed., Section 9.1.4.

4.5 P-Values

P-value. The probability, computed assuming H_0 is true, of a test statistic at least as extreme as the one observed. Equivalently, the smallest α at which the data reject H_0.

The rejection-region approach reports a binary verdict: reject or not. It does not report *how strong* the evidence was. A statistic of $z_0 = 1.97$ and a statistic of $z_0 = 6.2$ both "reject at $\alpha = 0.05$," yet the second is overwhelming and the first is borderline. The p-value repairs this.

$$\text{p-value} \;=\; P\big(\text{test statistic at least as extreme as the observed value} \;\big|\; H_0 \text{ true}\big). \tag{4.4}$$

"At least as extreme" is read in the direction(s) of H_1. An equivalent reading, often more useful: the p-value is the *smallest significance level at which the observed data would reject* H_0. A small p-value means the data would be surprising if H_0 were true, which is evidence against H_0. The decision rule becomes: reject H_0 if and only if p-value $\le \alpha$. Both approaches—critical value and p-value—always reach the same conclusion at the same α; the p-value simply carries more information.

For the z test, "extreme" translates into tail areas of the standard normal curve. Writing Φ for the standard normal cdf (the z table),

$$\text{p-value} = \begin{cases} 2\big[1-\Phi(|z_0|)\big] & \text{two-sided } (H_1\colon \mu \neq \mu_0), \\ 1-\Phi(z_0) & \text{upper-tailed } (H_1\colon \mu > \mu_0), \\ \Phi(z_0) & \text{lower-tailed } (H_1\colon \mu < \mu_0). \end{cases} \tag{4.5}$$

The two-sided case doubles the one-tail area because extreme values on *either* side argue against H_0, as Figure 4.3 shows.

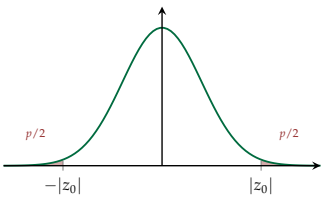

Figure 4.3. Two-sided p-value: the total shaded area $2[1-\Phi(|z_0|)]$, drawn for the dwell time statistic $z_0 = 2.50$.

z	0.00	0.01	0.02
2.4	0.9918	0.9920	0.9922
2.5	0.9938	0.9940	0.9941
2.6	0.9953	0.9955	0.9956

Table 4.3. Cumulative z-table excerpt: $\Phi(2.50) = 0.9938$, so the upper-tail area beyond 2.50 is 0.0062.

Let us recompute Example 4.2 with a p-value. The z table gives $\Phi(2.50) = 0.9938$ (Table 4.3), so the upper-tail area is $1 - 0.9938 = 0.0062$ and, doubling for the two-sided alternative,

$$\text{p-value} = 2(0.0062) = 0.0124.$$

Since $0.0124 < 0.05$, reject H_0—the same verdict as before, but now with a grade attached: the data are significant at $\alpha = 0.05$, and would remain significant down to any $\alpha \ge 0.0124$, but *not* at $\alpha = 0.01$.

Procedure 4.1 — Computing and interpreting a p-value.

1. Compute the test statistic from the sample.
2. Read the direction of "more extreme" from H_1: both tails for $\neq$, the right tail for $>$, the left tail for $<$.

3. Compute the tail probability under the null distribution (for the z test, use (4.5) and the cumulative table). Double the one-tail area for a two-sided test.
4. Decide: reject H_0 if p-value $\leq \alpha$; otherwise fail to reject.
5. Interpret: the p-value is the smallest α at which these data reject H_0. Report it alongside the decision, so the reader can apply their own α.

How small is small? Table 4.4 gives the conventional reading. Treat these bands as vocabulary, not as physics: there is nothing magical about 0.05, and a p-value of 0.049 is essentially the same evidence as 0.051.

p-value	Evidence against H_0
> 0.10	little or none
0.05 to 0.10	weak
0.01 to 0.05	moderate
0.001 to 0.01	strong
< 0.001	very strong

Table 4.4. Conventional interpretation of p-value magnitudes.

Insight. What a p-value is *not*. It is not the probability that H_0 is true—H_0 is a fixed statement, not a random event, and the p-value is computed *assuming* H_0 holds. It is not the probability that the result arose "by chance." And a large p-value does not confirm H_0; it only says the data cannot distinguish H_0 from nearby alternatives. Misreading the p-value as "the chance the claim is true" is the most common statistical error in published engineering reports.

4.5.1 *Reading p-values from the z table*

Software prints exact p-values, but on exams and in the field you will work from the printed cumulative table, which lists $\Phi(z)$ for z in steps of 0.01. The strategy: compute $|z_0|$, round to two decimals, look up $\Phi(|z_0|)$, take $1 - \Phi(|z_0|)$ for the tail, and double if two-sided. Because the z table is cumulative and finely spaced, it usually gives an *exact or near-exact* p-value—unlike the t and χ^2

tables of Section 4.8.1, which only bracket it. If $|z_0|$ carries more precision than two decimals, bracket it between consecutive table entries $z_a < |z_0| < z_b$; then $1 - \Phi(z_b) <$ tail area $< 1 - \Phi(z_a)$, and the p-value is bounded accordingly.

A handful of benchmark values, listed in Table 4.5, let you bound a z p-value without any table at all. If $|z_0| > 1.96$, then $p < 0.05$ two-sided; if $|z_0| > 2.576$, then $p < 0.01$ two-sided.

z_α	tail α	two-sided p
1.282	0.10	0.20
1.645	0.05	0.10
1.960	0.025	0.05
2.326	0.01	0.02
2.576	0.005	0.01
3.090	0.001	0.002

Table 4.5. Benchmark z values worth memorizing for quick p-value bounds.

Startup lens. Report the p-value, not just "significant." An investor update that says "conversion improved, $p = 0.048$, $n = 310$" invites the right follow-up questions. One that says "the improvement is statistically proven" invites regret.

Reference. Montgomery & Runger, 6th ed., Section 9.1.6.

4.6 *The Seven-Step Testing Procedure*

Every test in this course—and in the chapters on two-sample inference, regression, and designed experiments that follow—uses the same fixed sequence. Learn it once; only the test statistic and its table change from problem to problem.

Procedure 4.2 — The seven steps of a hypothesis test.

1. **Identify** the parameter of interest (μ, σ^2, or p) and the conditions: is σ known? is the population normal? is n large?
2. **State** the null hypothesis H_0.
3. **State** the alternative hypothesis H_1 (this fixes the sidedness) and choose the significance level α.
4. **Select** the appropriate test statistic (Z_0, T_0, or χ^2_0).
5. **Determine** the rejection region at level α (or plan to compute the p-value).
6. **Compute** the test statistic from the sample data (and the p-value, if using that route).
7. **Decide and conclude in context**: reject or fail to reject, then state what that means for the original engineering question.

In the lecture slides, steps 2 and 3 are sometimes compressed into a single line ("state H_0, H_1, and α"), giving a six-item list; the content is identical. Step 7 deserves emphasis. A bare "reject H_0" is not a conclusion. The conclusion is a

sentence about the problem: *''We reject the null hypothesis: there is sufficient evidence at the* 5% *significance level that the true mean dwell time is not* 115 *hours.''* And when the test fails to reject, say exactly that—*''there is not enough evidence that…''*—never ''we proved the claim is true.''

Reference. Montgomery & Runger, 6th ed., Section 9.2.

4.7 Tests on the Mean of a Normal Distribution, Variance Known

The z test applies when the parameter is a mean and the population standard deviation σ is known—from long production history, a calibrated measurement system, or a specification—or when n is large enough that the CLT (2.8) and the substitution $s \approx \sigma$ make the normal approximation accurate (as a working rule, $n \geq 40$). The test statistic is Z_0 from (4.3); the rejection regions are those of Figure 4.2; the p-values come from (4.5). Example 4.2 was a two-sided z test. The next example runs a one-sided test through all seven steps.

Example 4.3 — A vendor's disengagement claim

An autonomous-vehicle software vendor claims that its new perception stack drives *more than* 200 km between disengagements on average. Your validation team runs $n = 36$ standardized test routes and records $\bar{x} = 206$ km between disengagements; fleet history supports $\sigma = 18$ km. Test the vendor's claim at $\alpha = 0.05$ and report the p-value.

Solution.

Step 1 — Parameter and conditions.

The parameter is μ, the mean distance between disengagements. σ is known, so the z test applies.

Steps 2–3 — Hypotheses and α.

The vendor's claim (''more than 200'') must carry the burden of proof, so it is H_1:

$$H_0 : \mu \leq 200 \qquad \text{versus} \qquad H_1 : \mu > 200, \qquad \alpha = 0.05 \quad \text{(upper-tailed)}.$$

Steps 4–5 — Statistic and rejection region.

Use Z_0 from (4.3). Reject H_0 if $Z_0 > z_{0.05} = 1.645$.

Step 6 — Compute.

$$z_0 = \frac{206 - 200}{18/\sqrt{36}} = \frac{6}{3} = 2.00, \qquad \text{p-value} = 1 - \Phi(2.00) = 1 - 0.9772 = 0.0228.$$

Step 7 — Decide and conclude.

Since $z_0 = 2.00 > 1.645$ (equivalently, $0.0228 < 0.05$), we $\boxed{\text{reject } H_0}$. There is sufficient evidence at the 5% level that the mean distance between disengagements exceeds 200 km; the vendor's claim is supported. Note the p-value also tells us the support is moderate, not overwhelming: at $\alpha = 0.01$ the same data would fail to reject.

Safety lens. Example 4.3 accepts the vendor's framing: the vendor must prove improvement. If instead the question were "is this system safe to deploy without a driver?", the hypotheses should flip so that *safety* is what requires proof. Who carries the burden of proof is the most consequential line in a V&V test plan.

Reference. Montgomery & Runger, 6th ed., Section 9.3.

4.8 Tests on the Mean of a Normal Distribution, Variance Unknown

In most real problems σ is not known and must be estimated by the sample standard deviation s. Exactly as with confidence intervals (3.5), replacing σ by s replaces the standard normal reference distribution with the t distribution on $n - 1$ degrees of freedom. The test statistic is

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, \tag{4.6}$$

and under H_0 (with an approximately normal population) T_0 follows the t_{n-1} distribution. The decision rules mirror the z test with t critical values: reject H_0 if $|t_0| > t_{\alpha/2,n-1}$ (two-sided), if $t_0 > t_{\alpha,n-1}$ (upper-tailed), or if $t_0 < -t_{\alpha,n-1}$ (lower-tailed). For small n the normality of the population matters; check it with a normal probability plot or, at minimum, a dot plot free of strong skew and outliers.

Example 4.4 — Onboarding engagement time

Your startup redesigned its onboarding flow and claims that users now spend *more than* 12.5 minutes in their first session. A pilot with $n = 9$ new users gives $\bar{x} = 13.2$ minutes and $s = 1.5$ minutes; session times in past cohorts have looked approximately normal. Test the claim at $\alpha = 0.05$.

Solution.

Steps 1–3 — Set up.

Parameter: μ, mean first-session time; σ unknown, n small, population approximately normal, so use the t test with $n - 1 = 8$ degrees of freedom.

$$H_0 \colon \mu \le 12.5 \qquad \text{versus} \qquad H_1 \colon \mu > 12.5, \qquad \alpha = 0.05 \quad \text{(upper-tailed)}.$$

Steps 4–5 — Statistic and rejection region.

Use T_0 from (4.6). Reject H_0 if $t_0 > t_{0.05,8} = 1.860$ (Table 4.6).

Step 6 — Compute.

$$t_0 = \frac{13.2 - 12.5}{1.5/\sqrt{9}} = \frac{0.7}{0.5} = 1.40.$$

Step 7 — Decide and conclude.

Since $t_0 = 1.40 < 1.860$, we $\boxed{\text{fail to reject } H_0}$. There is not enough evidence at the 5% level that mean first-session time exceeds 12.5 minutes. This does *not* show the redesign failed—with $n = 9$ the test has little power to detect a real but modest improvement. The honest summary for the team: promising direction, sample too small to call.

Reference. Course handout: p-value bounds from statistical tables.

4.8.1 *Bounding the p-value from the t table*

Unlike the cumulative z table, the t table lists only *critical values* at selected tail areas ($\alpha = 0.25,\ 0.10,\ 0.05,\ 0.025,\ 0.01,\ 0.005$), one row per degrees of freedom. So an exact p-value cannot be read off; instead we *bound* it by bracketing the observed statistic between two critical values in its row.

Procedure 4.3 — Bounding a p-value from a critical-value table.

1. Go to the table row for your degrees of freedom.
2. Bracket the observed statistic between two adjacent critical values: find tail areas $\alpha_1 > \alpha_2$ with $t_{\alpha_1,\nu} < |t_0| < t_{\alpha_2,\nu}$. The *larger* critical value corresponds to the *smaller* tail area.

3. Read off the one-tail bound: $\alpha_2 <$ tail area $< \alpha_1$.
4. Convert to the p-value: one-sided tests use the bound directly, $\alpha_2 < p < \alpha_1$; two-sided tests double both ends, $2\alpha_2 < p < 2\alpha_1$.
5. For a *lower-tail* χ^2 test (Section 4.9), bracket χ^2_0 using the large-α columns and subtract each column header from 1 before applying step 4, because the header always denotes an *upper*-tail area.

Apply this to Example 4.4. In the $\nu = 8$ row of the t table (Table 4.6), the statistic $t_0 = 1.40$ falls between $t_{0.10,8} = 1.397$ and $t_{0.05,8} = 1.860$. The test is one-sided, so

$$0.05 < \text{p-value} < 0.10.$$

This bound agrees with the decision: $p > 0.05$, fail to reject at $\alpha = 0.05$—but the evidence is only just short of the threshold, which is far more informative than the bare verdict.

$\nu \backslash \alpha$	0.10	0.05	0.025	0.01
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
15	1.341	1.753	2.131	2.602

Table 4.6. t-table excerpt. For Example 4.4, $t_0 = 1.40$ falls between the two highlighted entries in the $\nu = 8$ row, so $0.05 < p < 0.10$ one-sided.

Please pay attention here, because this is the most common exam mistake with p-value bounds: forgetting to double the bounds for a two-sided test, or doubling them when the test is one-sided. A two-sided t test with $t_{0.025,\nu} < |t_0| < t_{0.01,\nu}$ has $0.02 < p < 0.05$, not $0.01 < p < 0.025$.

Lecture 10 starts here — *Practice: hypothesis testing*

Lecture 10. This section is covered by Lecture 10.

Reference. Montgomery & Runger, 6th ed., Section 9.4.

4.9 *Tests on the Variance of a Normal Distribution*

Sometimes the claim is about spread, not center: a contract caps the variability of delivery times; a re-slotted warehouse should have made picking times *more consistent*; a sensor's error variance must stay within specification. For a normal population, the test statistic for $H_0 : \sigma^2 = \sigma_0^2$ is

$$\chi^2_0 = \frac{(n-1)\,S^2}{\sigma_0^2}, \tag{4.7}$$

which under H_0 follows the chi-square distribution with $n-1$ degrees of freedom—the same sampling distribution that built the variance interval (3.6). The normality requirement is not cosmetic: the chi-square test for variance is sensitive to non-normality, and the CLT does not rescue it for large n.

The chi-square distribution is not symmetric, so the left and right critical values must be looked up separately, and the notation matters: $\chi^2_{\alpha,\nu}$ is the value with *upper*-tail area α, so $P(\chi^2_\nu > \chi^2_{\alpha,\nu}) = \alpha$. A left-tail cutoff with area 0.05 below it is therefore written $\chi^2_{0.95,\nu}$. Table 4.7 collects the decision rules.

Test	H_1	Reject H_0 if
two-sided	$\sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha/2,\,n-1}$ or $\chi_0^2 < \chi^2_{1-\alpha/2,\,n-1}$
upper-tailed	$\sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha,\,n-1}$
lower-tailed	$\sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi^2_{1-\alpha,\,n-1}$

Table 4.7. Decision rules for the chi-square test on a variance (4.7). The distribution is not symmetric, so left and right critical values differ.

Example 4.5 — Did re-slotting make picking times more consistent?

A warehouse re-slotted its fast-moving items to *reduce the variability* of order picking times below the historical value $\sigma_0^2 = 0.25\ \text{min}^2$. A sample of $n = 20$ picks after the change gives $s^2 = 0.14\ \text{min}^2$; picking times are approximately normal. Test at $\alpha = 0.05$ and bound the p-value.

Solution.

Steps 1–3 — Set up.

Parameter: σ^2, the variance of picking time. The improvement claim must be proven, so

$$H_0\colon \sigma^2 \geq 0.25 \qquad \text{versus} \qquad H_1\colon \sigma^2 < 0.25, \qquad \alpha = 0.05 \quad \text{(lower-tailed)}.$$

Steps 4–5 — Statistic and rejection region.

Use χ_0^2 from (4.7) with $n - 1 = 19$ degrees of freedom. From Table 4.7, reject H_0 if $\chi_0^2 < \chi^2_{0.95,\,19} = 10.117$ (Table 4.8 and Figure 4.4).

Step 6 — Compute.

$$\chi_0^2 = \frac{(20-1)(0.14)}{0.25} = \frac{2.66}{0.25} = 10.64.$$

For the p-value, bracket 10.64 in the $\nu = 19$ row using the large-α columns (Procedure 4.3, step 5): $\chi^2_{0.95,\,19} = 10.117 < 10.64 < 11.651 = \chi^2_{0.90,\,19}$, so the lower-tail area satisfies $1 - 0.95 < p < 1 - 0.90$, i.e., $0.05 < p < 0.10$.

Step 7 — Decide and conclude.

Since $\chi_0^2 = 10.64 > 10.117$, the statistic does *not* fall in the rejection region (equivalently, $p > 0.05$), so we fail to reject H_0. There is not enough evidence at the 5% level that re-slotting reduced the picking-time variance below 0.25 min^2. The point estimate $s^2 = 0.14$ looks like a real improvement, but with $n = 20$ the variance estimate is itself too variable to be sure. Note also the sensitivity to α: at $\alpha = 0.10$ the critical value becomes $\chi^2_{0.90,19} = 11.651$ and the same data *would* reject.

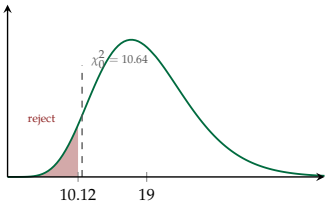

Figure 4.4. Lower-tailed chi-square test with $\nu = 19$: the shaded rejection region lies below $\chi^2_{0.95,19} = 10.117$; the observed $\chi_0^2 = 10.64$ falls just outside it.

$\nu \backslash \alpha$	0.95	0.90	0.10	0.05	0.025
19	10.117	11.651	27.204	30.144	32.852
20	10.851	12.443	28.412	31.410	34.170

Table 4.8. Chi-square table excerpt. Column headers are *upper-tail* areas: $\chi^2_{0.95,19} = 10.117$ leaves only 5% in the *left* tail. The highlighted pair brackets $\chi_0^2 = 10.64$ in Example 4.5.

Engineering judgment. A variance test often matters more than a mean test in operations. A warehouse whose mean pick time drops but whose variance grows will miss *more* shipping cutoffs, not fewer, because deadlines are hit by the tail of the distribution, not by its center.

Reference. Montgomery & Runger, 6th ed., Section 9.5.

4.10 Tests on a Population Proportion

Claims about rates—defect rates, pass rates, conversion rates—are claims about a proportion p. With a sample of n Bernoulli trials and X successes, $\hat{P} = X/n$, and for large n the normal approximation gives the test statistic for $H_0\colon p = p_0$:

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}. \tag{4.8}$$

Note carefully that the standard error in the denominator uses the *claimed* value p_0, not $\hat{p}$—under H_0, the variance of $\hat{P}$ is $p_0(1-p_0)/n$, so the null hypothesis supplies the yardstick. (The confidence interval (3.7), which assumes no null value, uses $\hat{p}$ instead.) The approximation is adequate when $np_0 \geq 5$ and $n(1-p_0) \geq 5$. Rejection regions and p-values are exactly those of the z test (Figure 4.2, equation (4.5)).

Example 4.6 — Perception model pass rate

A perception-model vendor claims that *at least* 95% of scenarios in your safety test suite are handled correctly. Your team runs a random sample of $n = 200$ scenarios and the model passes 183 of them ($\hat{p} = 0.915$). At $\alpha = 0.05$, is there evidence that the true pass rate is below 95%? Report the p-value.

Solution.

Steps 1–3 — Set up.

Parameter: p, the pass rate over the scenario population. Check the approx-

imation: $np_0 = 200(0.95) = 190 \geq 5$ and $n(1-p_0) = 200(0.05) = 10 \geq 5$.

$$H_0 \colon p \geq 0.95 \qquad \text{versus} \qquad H_1 \colon p < 0.95, \qquad \alpha = 0.05 \quad \text{(lower-tailed)}.$$

Steps 4–5 — Statistic and rejection region.
Use Z_0 from (4.8). Reject H_0 if $Z_0 < -z_{0.05} = -1.645$.

Step 6 — Compute.
The null standard error is $\sqrt{0.95 \times 0.05/200} = \sqrt{0.0002375} = 0.01541$, so

$$z_0 = \frac{0.915 - 0.95}{0.01541} = \frac{-0.035}{0.01541} = -2.27, \qquad \text{p-value} = \Phi(-2.27) = 1 - 0.9884 = 0.0116.$$

Step 7 — Decide and conclude.
Since $z_0 = -2.27 < -1.645$ (equivalently, $0.0116 < 0.05$), we $\boxed{\text{reject } H_0}$. There is significant evidence at the 5% level that the pass rate is below the claimed 95%. The vendor's number should not go into the safety case as stated.

Safety lens. Example 4.6 treats the 200 scenarios as a random sample from the scenario population the safety case covers. If the suite oversamples easy highway scenes, no p-value can repair the bias: inference is only as good as the sampling design behind it.

Reference. Montgomery & Runger, 6th ed., Section 9.2.4.

4.11 *The Duality Between Tests and Confidence Intervals*

You have now seen the same pair of facts twice: the dwell time CI $(116.08,\ 123.92)$ excluded 115, and the $\alpha = 0.05$ test rejected $\mu_0 = 115$. The general statement:

Insight. A two-sided test of $H_0 \colon \mu = \mu_0$ at significance level α rejects H_0 *exactly when* the $100(1-\alpha)\%$ two-sided confidence interval fails to contain μ_0. The two computations rearrange the same inequality: $|\bar{x} - \mu_0| > z_{\alpha/2}\,\sigma/\sqrt{n}$ says both "Z_0 is in the rejection region" and "μ_0 is outside $\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$." A confidence interval is the set of all null values that the data would *not* reject.

The same duality links one-sided tests to the one-sided confidence bounds (3.2), the chi-square variance test to the variance interval (3.6), and the t test to the t interval (3.5). In practice, report both: the test gives the decision and its error control; the interval shows the range of parameter values the data actually support, which guards against confusing *statistical* significance with *practical* im-

portance. With n large enough, a dwell time of 115.2 hours would also "reject $\mu_0 = 115$"—statistically detectable, operationally irrelevant.

Startup lens. When a stakeholder asks "did the experiment win?", answer with the interval, not only the verdict. "Conversion changed by $+0.8$ to $+3.1$ percentage points (95% CI)" supports a business decision; "$p < 0.05$" alone does not say whether the effect is worth shipping.

Reference. Montgomery & Runger, 6th ed., Sections 9.2.2–9.2.3.

4.12 *Type II Error, Power, and Sample Size*

The significance level α is chosen; the type II error probability β must be *computed*, and it can only be computed against a specific alternative. The question "what is β?" is incomplete; the complete question is "what is β *if the true mean is* μ_{true}?"

The computation has a simple structure. The two-sided z test fails to reject exactly when $\bar{X}$ lands in the fail-to-reject interval

$$\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;\le\; \bar{X} \;\le\; \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{4.9}$$

If the truth is $\mu = \mu_{\text{true}}$, then $\bar{X}$ is normal with mean μ_{true} and standard error $\sigma/\sqrt{n}$, and β is just the probability that this normal variable lands inside (4.9). Standardizing, with $\delta = \mu_{\text{true}} - \mu_0$ denoting the true shift,

$$\beta \;=\; \Phi\!\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \;-\; \Phi\!\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right), \qquad \delta = \mu_{\text{true}} - \mu_0. \tag{4.10}$$

For a one-sided test, drop the term that corresponds to the tail the test does not use and replace $z_{\alpha/2}$ by z_α. The power is $1 - \beta$.

Procedure 4.4 — Computing β for a z test.

1. Write the fail-to-reject region for $\bar{X}$ under H_0 at the chosen α, using (4.9).
2. Specify the alternative μ_{true} you want protection against (an engineering choice: the smallest shift that matters).
3. Compute $\beta = P(\bar{X} \in \text{fail-to-reject region} \mid \mu = \mu_{\text{true}})$ by standardizing both endpoints with mean μ_{true} and standard error $\sigma/\sqrt{n}$, or apply (4.10) directly.
4. Report the power $1 - \beta$ alongside α; if the power is too low, increase n

(Equation (4.11)).

Example 4.7 — How often would we miss a real shift in dwell time?

Continue the terminal-dwell-time setting of Example 4.2: $\mu_0 = 115$, $\sigma = 16$, $n = 64$, standard error $\sigma/\sqrt{n} = 2$. (a) Suppose the true mean is actually $\mu_{\text{true}} = 119$ hours. Compute β for the two-sided test at $\alpha = 0.05$ and at $\alpha = 0.01$, and interpret. (b) How many containers must be sampled so that a true shift of $\delta = 4$ hours is detected with power 0.90 at $\alpha = 0.05$?

Solution.

Part (a), $\alpha = 0.05$ — fail-to-reject region, then a normal probability.

By (4.9) the test fails to reject when

$$115 - 1.96(2) \le \bar{X} \le 115 + 1.96(2) \quad \Longleftrightarrow \quad 111.08 \le \bar{X} \le 118.92.$$

If $\mu = 119$, standardize both ends with mean 119 and standard error 2:

$$\beta = P\left(\frac{111.08 - 119}{2} \le Z \le \frac{118.92 - 119}{2}\right) = P(-3.96 \le Z \le -0.04) = \Phi(-0.04) - \Phi(-3.96) = 0.4840 - 0.0000,$$

so $\beta \approx 0.484$: the gold area in Figure 4.1 and the shaded area in Figure 4.5. Even though the terminal is understating dwell time by 4 hours, this test misses the shift almost half the time. Checking with (4.10): $\delta\sqrt{n}/\sigma = 4(8)/16 = 2$, and $\Phi(1.96 - 2) - \Phi(-1.96 - 2) = \Phi(-0.04) - \Phi(-3.96) = 0.4840$.

Part (a), $\alpha = 0.01$.

The fail-to-reject interval widens to $115 \pm 2.576(2) = (109.85,\ 120.15)$, and

$$\beta = P\left(\frac{109.85 - 119}{2} \le Z \le \frac{120.15 - 119}{2}\right) = P(-4.58 \le Z \le 0.58) = \Phi(0.58) - 0.0000 = 0.7190.$$

So $\beta \approx \boxed{0.72}$. Tightening α from 0.05 to 0.01 (fewer false alarms) raised the missed-detection probability from 0.484 to about 0.72—the α–β tradeoff in numbers.

Part (b) — sample size from power.

Power 0.90 means $\beta = 0.10$, so $z_\beta = z_{0.10} = 1.28$. Using (4.11) below,

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\,\sigma^2}{\delta^2} = \frac{(1.96 + 1.28)^2(16)^2}{4^2} = \frac{(3.24)^2 \times 256}{16} = 10.50 \times 16 = 168.0,$$

so sample $\boxed{n = 168}$ containers (always round up). Tripling the sample turns a coin-flip detector into a reliable one.

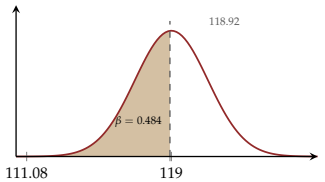

Figure 4.5. β for Example 4.7(a): the probability that $\bar{X}$, distributed around the true mean 119, still lands inside the fail-to-reject interval (111.08, 118.92).

Operating characteristic curve. A plot of β against the true parameter value (or the shift δ), one curve per sample size.

The formula used in part (b) comes from requiring the test to have type II error at most β at the shift δ. Setting the right-hand side of (4.10) equal to β and ignoring the far-tail term (which is negligible whenever $\delta \neq 0$) gives $z_\beta \approx \delta\sqrt{n}/\sigma - z_{\alpha/2}$, and solving for n:

$$n \;\approx\; \frac{(z_{\alpha/2} + z_\beta)^2\,\sigma^2}{\delta^2}, \qquad \delta = \mu_{\text{true}} - \mu_0 \quad \text{(two-sided; use } z_\alpha \text{ for one-sided)}. \tag{4.11}$$

Read the structure: sample size grows with the square of the noise-to-signal ratio σ/δ. Halving the detectable shift quadruples the required sample.

Plotting β from (4.10) against the true mean gives the *operating characteristic (OC) curve* of the test, Figure 4.6. Every curve equals $1 - \alpha$ at $\mu = \mu_0$ (no shift to detect) and falls toward zero as the true mean moves away; larger samples make the curve drop faster. Montgomery & Runger provide printed OC charts in their appendix; the figure here is computed directly from (4.10).

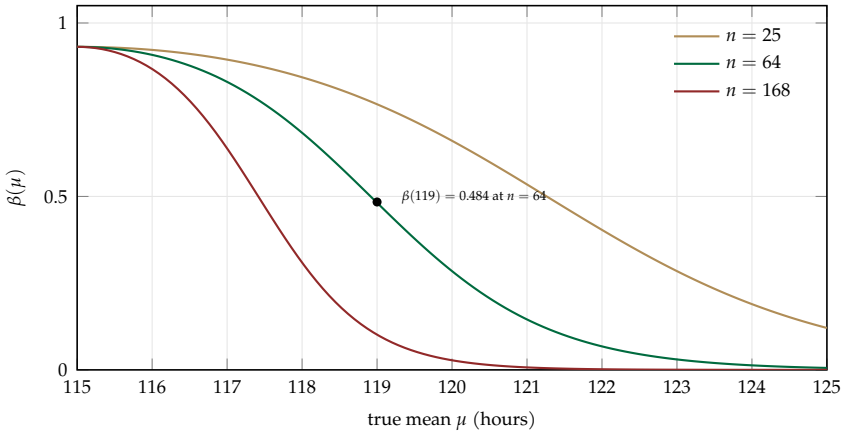

Figure 4.6. Operating characteristic curves for the two-sided dwell-time z test ($\mu_0 = 115$, $\sigma = 16$,

$\alpha = 0.05$), computed from (4.10). Larger samples detect the same shift with much smaller β; at $n = 168$ the shift to 119 is missed only about 10% of the time.

Engineering judgment. Choose δ from the process, not from the data. The question "what shift in mean dwell time would force us to renegotiate berth schedules?" has an operational answer—say 4 hours—and that answer, not the observed $\bar{x}$, belongs in (4.11).

Reference. Montgomery & Runger, 6th ed., Chapter 9 roadmap.

4.13 *Choosing the Right Test*

Four one-sample tests are now on the table. Choosing among them is a short decision sequence, identical in spirit to choosing the right interval (Procedure 3.5).

Procedure 4.5 — Choosing the one-sample test.

1. Identify the parameter the claim is about. A statement about an average $\rightarrow \mu$; about consistency, spread, or precision $\rightarrow \sigma^2$; about a rate or percentage of items $\rightarrow p$.
2. If the parameter is μ: use the z statistic (4.3) when σ is known (or $n \geq 40$, using s for σ); use the t statistic (4.6) when σ is unknown and the population is approximately normal.
3. If the parameter is σ^2: use the chi-square statistic (4.7); the population must be normal.
4. If the parameter is p: use the proportion z statistic (4.8); check $np_0 \geq 5$ and $n(1-p_0) \geq 5$.
5. Fix the sidedness from the claim—before looking at the data—and run Procedure 4.2.

Parameter	Condition	Test statistic	Null distribution	Equation
μ	σ known	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	standard normal	(4.3)
μ	σ unknown, normal pop.	$T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$	t_{n-1}	(4.6)
σ^2	normal population	$\chi_0^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$	χ^2_{n-1}	(4.7)
p	$np_0 \geq 5$, $n(1-p_0) \geq 5$	$Z_0 = \dfrac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$	standard normal	(4.8)

Table 4.9. The four one-sample tests. The sampling distributions are exactly those behind the corresponding confidence intervals in Table 4.1.

Reference. Lecture 10 practice sets A, B, and C.

4.14 *Practice: Three Problems, One Procedure*

The three worked problems below deliberately span the three test statistics and the three sidedness patterns. In each, watch how Procedure 4.2 carries the entire solution and how the p-value locates the evidence more finely than the verdict alone.

Example 4.8 — Practice A — two-sided z test: route completion times

A freight forwarder's dock schedule assumes the mean truck route completion time is 150 minutes. A telematics sample of $n = 49$ routes gives $\bar{x} = 154$ minutes; historical records support $\sigma = 14$ minutes. Test whether the mean differs from 150 at $\alpha = 0.05$; then state what happens at $\alpha = 0.01$.

Solution.

Steps 1–3.

Parameter μ, σ known. Both early and late arrivals disrupt the docks: $H_0\colon \mu = 150$ versus $H_1\colon \mu \neq 150$, $\alpha = 0.05$ (two-sided).

Steps 4–5.

z test (4.3); reject if $|Z_0| > z_{0.025} = 1.96$.

Step 6.

$$z_0 = \frac{154 - 150}{14/\sqrt{49}} = \frac{4}{2} = 2.00, \qquad \text{p-value} = 2\big[1 - \Phi(2.00)\big] = 2(0.0228) = 0.0455.$$

Step 7.

$|2.00| > 1.96$, so $\boxed{\text{reject } H_0}$: sufficient evidence at the 5% level that the mean route time differs from 150 minutes. At $\alpha = 0.01$ the same data give $0.0455 > 0.01$: fail to reject. Same sample, same statistic, different conclusion—the stricter α demands stronger evidence, and (Example 4.7) quietly raises β.

Example 4.9 — Practice B — upper-tailed t test: engagement sessions

Your startup claims that pilot users now average *more than* 10.0 feature sessions per week after the activation redesign. A sample of $n = 16$ pilot users gives $\bar{x} = 10.8$ sessions and $s = 1.2$ sessions; weekly session counts for pilot users are approximately normal. Test at $\alpha = 0.05$, bound the p-value, and check $\alpha = 0.01$.

Solution.

Steps 1–3.

Parameter μ, σ unknown, small n, approximately normal population: t test with $\nu = 15$. $H_0 : \mu \leq 10.0$ versus $H_1 : \mu > 10.0$, $\alpha = 0.05$ (upper-tailed).

Steps 4–5.

t statistic (4.6); reject if $t_0 > t_{0.05,\,15} = 1.753$.

Step 6.

$$t_0 = \frac{10.8 - 10.0}{1.2/\sqrt{16}} = \frac{0.8}{0.3} = 2.67.$$

P-value bound (Procedure 4.3), row $\nu = 15$: $t_{0.01,\,15} = 2.602 < 2.67 < 2.947 = t_{0.005,\,15}$, so, one-tailed, $0.005 < p < 0.01$.

Step 7.

$2.67 > 1.753$, so $\boxed{\text{reject } H_0}$: sufficient evidence at the 5% level that mean weekly sessions exceed 10.0. Because $p < 0.01$, the conclusion survives even at $\alpha = 0.01$ ($2.67 > t_{0.01,\,15} = 2.602$)—strong evidence, worth acting on.

Example 4.10 — Practice C — two-sided χ^2 test: sensor-fusion error variance

The specification for a sensor-fusion module states that the variance of its lateral position error is $\sigma_0^2 = 0.020\ \text{m}^2$. A validation sample of $n = 20$ runs gives $s^2 = 0.032\ \text{m}^2$; the errors are approximately normal. Test whether the variance *differs* from specification at $\alpha = 0.05$ and bound the p-value.

Solution.

Steps 1–3.

Parameter σ^2; normal population; deviation in either direction matters for the safety case (too large is dangerous, too small suggests a data problem): $H_0\colon \sigma^2 = 0.020$ versus $H_1\colon \sigma^2 \neq 0.020$, $\alpha = 0.05$ (two-sided), $\nu = 19$.

Steps 4–5.

Chi-square statistic (4.7); from Table 4.7, reject if $\chi_0^2 > \chi^2_{0.025,\,19} = 32.852$ or $\chi_0^2 < \chi^2_{0.975,\,19} = 8.907$.

Step 6.

$$\chi_0^2 = \frac{19 \times 0.032}{0.020} = 30.4 \quad \text{(upper side)}.$$

P-value bound: in the $\nu = 19$ row (Table 4.8), $\chi^2_{0.05,\,19} = 30.144 < 30.4 < 32.852 = \chi^2_{0.025,\,19}$, so the upper-tail area is between 0.025 and 0.05; doubling for the two-sided test, $0.05 < p < 0.10$.

Step 7.

30.4 is neither above 32.852 nor below 8.907, so $\boxed{\text{fail to reject } H_0}$ (equivalently $p > 0.05$). There is not enough evidence at the 5% level that the error variance differs from specification—but the p-value bound shows the data sit close to the line, and $s^2 = 0.032$ is 60% above spec. The right engineering response is not ''all clear'' but ''collect more runs before sign-off.''

	Practice A	Practice B	Practice C
Parameter	μ	μ	σ^2
σ known?	yes	no	—
Test	z (4.3)	t (4.6)	χ^2 (4.7)
Sidedness	two	upper	two
Statistic	2.00	2.67	30.4
P-value	0.0455	0.005–0.01	0.05–0.10
Decision at 0.05	reject	reject	fail to reject

Table 4.10. The three practice problems side by side. The seven-step procedure is identical; only the statistic and its table change.

Chapter Summary

A hypothesis test decides between a default claim H_0 and a research claim H_1 (4.1), with the burden of proof on H_1. The test can err in two ways—false alarm (α) and missed detection (β) (4.2)—and for fixed n the two trade off; only a larger sample reduces both. The decision runs through the seven steps of Procedure 4.2: identify the parameter, state H_0 and H_1 with α, pick the statistic from Procedure 4.5, set the rejection region, compute, decide, and conclude in context. Each test statistic—Z_0 (4.3), T_0 (4.6), χ^2_0 (4.7), and the proportion statistic (4.8)—measures the distance between data and claim in standard-error units, using the same sampling distributions as Chapter 3, which is why a two-sided test at level α and a $100(1-\alpha)\%$ interval always agree. The p-value (4.4) is the smallest α at which the data reject: computed exactly from the z table (4.5), bounded from the t and χ^2 tables (Procedure 4.3). Finally, β is a computation, not a hope: fix an alternative that matters, apply (4.10) (Procedure 4.4), and if the power is too low, size the sample with (4.11).

Quantity	Formula	Equation
Hypothesis pair	$H_0\colon \theta = \theta_0$ vs. $H_1\colon \theta \neq \theta_0$ (or $>$, $<$)	(4.1)
Error probabilities	$\alpha = P(\text{reject} \mid H_0)$, $\beta = P(\text{fail to reject} \mid H_1)$	(4.2)
z statistic (mean, σ known)	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	(4.3)
p-value	$P(\text{at least as extreme} \mid H_0)$	(4.4)
p-value, z test	$2[1-\Phi(\lvert z_0\rvert)]$; $1-\Phi(z_0)$; $\Phi(z_0)$	(4.5)
t statistic (mean, σ unknown)	$T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$, $\nu = n-1$	(4.6)
χ^2 statistic (variance)	$\chi^2_0 = \dfrac{(n-1)S^2}{\sigma_0^2}$, $\nu = n-1$	(4.7)
z statistic (proportion)	$Z_0 = \dfrac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$	(4.8)
Fail-to-reject region for $\bar{X}$	$\mu_0 \pm z_{\alpha/2}\,\sigma/\sqrt{n}$	(4.9)
β for the z test	$\Phi\big(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\big) - \Phi\big(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\big)$	(4.10)
Sample size from power	$n \approx \dfrac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}$	(4.11)

Table 4.11. Key formulas of Chapter 4 at a glance.

Exercises

Framing hypotheses

Exercise 4.1. Your startup's monthly churn rate has historically been 6%. A new onboarding flow is meant to *reduce* churn. Which hypothesis pair is correctly framed for deciding whether to ship the new flow?

(A) $H_0: p < 0.06$ versus $H_1: p = 0.06$

(B) $H_0: p \geq 0.06$ versus $H_1: p < 0.06$

(C) $H_0: p \leq 0.06$ versus $H_1: p > 0.06$

(D) $H_0: \hat{p} = 0.06$ versus $H_1: \hat{p} < 0.06$

Exercise 4.2. A guardrail filter for an LLM product is tested before release with H_0: ''the false-negative rate meets the 1% specification'' versus H_1: ''the rate exceeds 1%.'' Describe, in the words of this application, (a) the type I error, (b) the type II error, and (c) which error the choice of α controls, and why that framing is dangerous here.

z tests: mean with σ known

Exercise 4.3. A truck fleet's route-planning model assumes mean fuel use of 32.0 L/100 km with known $\sigma = 2.4$. After a driver-training program, a sample of $n = 36$ trips gives $\bar{x} = 31.1$ L/100 km. Test $H_0: \mu = 32.0$ versus $H_1: \mu \neq 32.0$ at $\alpha = 0.05$, and compute the p-value.

Exercise 4.4. Your landing-page team claims the mean session length now *exceeds* 4.0 minutes. Historical data give $\sigma = 1.5$ minutes; a sample of $n = 100$ sessions yields $\bar{x} = 4.3$ minutes. Run the seven steps at $\alpha = 0.05$, report the p-value, and state the conclusion at $\alpha = 0.01$ as well.

t tests: mean with σ unknown

Exercise 4.5. After a warehouse re-layout, management claims the mean order picking time is now *below* 3.5 minutes. A sample of $n = 25$ picks gives $\bar{x} = 3.32$ minutes and $s = 0.60$ minutes; picking times are approximately normal. Test at $\alpha = 0.05$ and bound the p-value from the t table ($t_{0.10,24} = 1.318$, $t_{0.05,24} = 1.711$).

Exercise 4.6. An LLM evaluation benchmark has a contract target score of 75. Twelve independent evaluation runs give $\bar{x} = 78.1$ and $s = 4.4$; run scores are approximately normal. Test $H_0 : \mu = 75$ versus $H_1 : \mu \neq 75$ at $\alpha = 0.05$, bound the p-value ($t_{0.025,11} = 2.201$, $t_{0.01,11} = 2.718$), and state the decision at $\alpha = 0.01$.

Variance and proportion

Exercise 4.7. A carrier contract caps the variance of daily demand-forecast errors at 25 (units^2). A sample of $n = 15$ days gives $s^2 = 41$; errors are approximately normal. Is there evidence at $\alpha = 0.05$ that the variance *exceeds* the cap? Bound the p-value ($\chi^2_{0.10,14} = 21.064$, $\chi^2_{0.05,14} = 23.685$).

Exercise 4.8. A red-team vendor claims its guardrail blocks *at least* 98% of a standard jailbreak prompt set. In a random sample of $n = 400$ prompts, 380 are blocked. Test at $\alpha = 0.05$ and compute the p-value.

Power and sample size

Exercise 4.9. A delivery service tests $H_0 : \mu = 30$ minutes (mean lateness) versus $H_1 : \mu \neq 30$ with $\sigma = 8$, $n = 64$, $\alpha = 0.05$. Compute β if the true mean lateness is actually 33 minutes, and state the power.

Exercise 4.10. For the lateness test of Exercise 4.9, how large must n be so that a true shift of $\delta = 3$ minutes is detected with power 0.95 at $\alpha = 0.05$ (two-sided)?

Exam-style problems

Exercise 4.11. **[20 pt]** A port authority's service agreement states a mean container dwell time of 60 hours. An audit samples $n = 49$ containers: $\bar{x} = 63.5$ hours, with $\sigma = 10.5$ hours known from terminal history.

1. **[4 pt]** State the parameter, the hypotheses, and the sidedness for testing the agreement.
2. **[5 pt]** At $\alpha = 0.05$, give the rejection region, compute the test statistic, and decide.

3. **[4 pt]** Compute the p-value and interpret it in one sentence.
4. **[3 pt]** Construct the 95% confidence interval for μ and show that it agrees with your decision in (b).
5. **[4 pt]** If the true mean dwell time is actually 64 hours, compute β for this test.

Exercise 4.12. **[15 pt]** A safety team evaluates an LLM assistant on a 16-run benchmark: $\bar{x} = 71.5$, $s = 3.0$; run scores are approximately normal.

1. **[6 pt]** The release bar requires mean score *above* 70. Test at $\alpha = 0.05$ and bound the p-value ($t_{0.05,15} = 1.753$, $t_{0.025,15} = 2.131$).
2. **[6 pt]** The release bar also caps the score variance at 4.0. With $s^2 = 9.0$, test $H_0 : \sigma^2 \leq 4.0$ versus $H_1 : \sigma^2 > 4.0$ at $\alpha = 0.05$ ($\chi^2_{0.05,15} = 24.996$, $\chi^2_{0.005,15} = 32.801$), and bound the p-value.
3. **[3 pt]** Combine (a) and (b) into a one-sentence release recommendation.

Assessing outside recommendations

Exercise 4.13. A consultant analyzes truck turnaround times at your cross-dock. Testing $H_0 : \mu = 25$ minutes versus $H_1 : \mu \neq 25$ with $n = 10$, $\bar{x} = 26.8$, $s = 4.1$, they compute $t_0 = 1.39$, compare it with $t_{0.025,9} = 2.262$, and write in the report: ''*We accept the null hypothesis. The analysis proves that the mean turnaround time equals 25 minutes, so no process change is needed.*'' Identify every error in this conclusion and rewrite it correctly. ($t_{0.10,9} = 1.383$, $t_{0.05,9} = 1.833$.)

Exercise 4.14. Your co-founder asks an AI assistant whether a pricing experiment moved mean revenue per user from its baseline of SAR 50. The assistant inspects the data, notices $\bar{x} = 52.4 > 50$, and reports: ''*Since the sample mean is above 50, I tested* $H_0 : \mu = 50$ *versus* $H_1 : \mu > 50$ *and obtained p-value* $= 0.03 < 0.05$. *There is a 97% probability the price change increased revenue; the result is significant.*'' The experiment was registered in advance as a two-sided question. Find the two distinct errors and give the corrected analysis and conclusion.

In-Class Activity 4.1: Whose Burden of Proof?

Week 4 — Lecture 8

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

Your team is the release gate for a new perception model in an autonomous shuttle. The current model is in production; the new model claims fewer failures. A test drive campaign will produce failure data, and your team must frame the hypothesis test that decides whether the new model ships.

Part 1 — Frame it twice

Write two possible hypothesis pairs: (i) with ''the new model is no better'' as H_0, and (ii) with ''the new model is unsafe to ship'' as H_0. For each, state in words what a type I and a type II error would mean (which model ends up on the road, wrongly?).

Part 2 — Which error is which price?

For each framing in Part 1, say which error is controlled at α and which is left as β. Then decide as a team: which framing is appropriate for a *safety* release gate, and why?

Part 3 — Choose α

Your manager proposes $\alpha = 0.10$ ''so we can move fast.'' Write two sentences to the manager
what $\alpha = 0.10$ means in plain operational terms for your chosen framing, and one stating what
(not ''luck'') to reduce both error probabilities at once.

Exit check

Complete in one sentence: ''Failing to reject H_0 is not the same as proving H_0 because...''

In-Class Activity 4.2: Run the Seven Steps

Week 5 — Lecture 9

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

A parcel carrier promises your fulfillment startup a mean line-haul time of at most 48 hours. Your tracking data: $n = 36$ shipments, $\bar{x} = 49.2$ hours, and $\sigma = 3.6$ hours (known from a full year of telemetry). You will test the promise at $\alpha = 0.05$. A z-table excerpt: $\Phi(2.00) = 0.9772$.

Part 1 — Steps 1 to 5

Identify the parameter and conditions, state H_0 and H_1 with the correct sidedness (who carries the burden of proof?), pick the test statistic, and write the rejection region at $\alpha = 0.05$.

Part 2 — Steps 6 and 7

Compute the test statistic and the p-value, make the decision, and write the conclusion as one sentence *about shipments*, not about symbols.

Part 3 — The carrier pushes back

The carrier's analyst replies: "At $\alpha = 0.01$ your own data fail to reject; therefore our promis
Which half of that sentence is correct arithmetic and which half is a reasoning error? Explain

Exit check

Your p-value in Part 2, and in one phrase what it measures.

In-Class Activity 4.3: Which Test, and How Sure Are We?

Week 5 — Lecture 10

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

Your startup claims the redesigned signup flow cut mean onboarding time below 8.0 minutes. Data: $n = 16$ new users, $\bar{x} = 7.4$ minutes, $s = 1.2$ minutes; onboarding times are approximately normal. A t-table excerpt for $\nu = 15$: $t_{0.10} = 1.341$, $t_{0.05} = 1.753$, $t_{0.025} = 2.131$.

Part 1 — Choose the test

Which of the four one-sample tests applies, and why not the other three? State H_0, H_1, and the rejection region at $\alpha = 0.05$.

Part 2 — Compute and bound

Compute the test statistic, decide, and bound the p-value from the table excerpt. Write the conclusion in the language of onboarding.

Part 3 — How could we be wrong?

Suppose the redesign truly cut the mean to 7.5 minutes. Which error type could your Par commit? Without computing β, list two concrete changes to the study that would reduce it, and would reduce α instead—and say what that would cost you.

Exit check

One sentence: when do you use the t statistic instead of the z statistic for a mean?