5 *Inference for Two Samples*

Chapters 3 and 4 built confidence intervals and hypothesis tests for one population at a time. Most engineering questions, however, are comparisons: which of two suppliers is more consistent, does the new software version fail less often than the old one, did the process change actually improve throughput? This chapter extends every one-sample tool to two populations. It is organized in the following order: the two-sample framework; inference on the difference of two means when the variances are known; the pooled t procedure when the variances are unknown but equal; the Welch t procedure when they are unequal; the paired t procedure when the observations come in matched pairs; inference on two proportions; the F test for two variances; and a decision map that ties all the procedures together.

The chapter then does two more things. Because Major Exam 1 covers Chapters 1–5, the chapter closes with a marked review section that compresses everything so far into one master table and one problem-attack procedure, and a practice section with fully worked exam-style problems. Keep this chapter close during exam week: it is both new material and your revision map.

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

Q1. An engineer measures the fuel efficiency of the *same* eight trucks before and after an engine tune-up. Are these two independent samples?

Q2. A pooled two-sample t test uses samples of sizes $n_1 = 12$ and $n_2 = 15$. How many degrees of freedom does the test have?

Q3. A 95% confidence interval for $\mu_1 - \mu_2$ comes out as $(-1.2,\ 3.8)$. At $\alpha = 0.05$, what does a two-sided test of $H_0\colon \mu_1 = \mu_2$ conclude?

Answers.

Q1. *No.* Each "after" measurement is naturally linked to the "before" measurement on the same truck, so the two columns of numbers are dependent. The correct analysis works with the eight *differences*, one per truck: the paired t procedure, with $n - 1 = 7$ degrees of freedom.

Q2. $\nu = n_1 + n_2 - 2 = 25$. One degree of freedom is spent estimating each sample mean, and both samples contribute their remaining information to the single pooled variance estimate.

Q3. *Fail to reject* H_0. The interval contains 0, and by the CI–test duality a two-sided test at level α rejects exactly when the hypothesized value falls outside the $100(1-\alpha)\%$ interval.

Lecture 11. This section is covered by Lecture 11.

Reference. Montgomery & Runger, 6th ed., Section 10.1.

Two-sample problem. An inference problem with two populations, each described by its own parameters and each observed through its own random sample.

Lecture 11 starts here — *Two-sample inference, part 1*

5.1 From One Sample to Two

The one-sample chapters asked questions about a single population parameter: "is $\mu = 50$?" The two-sample problem asks a comparison question: "is $\mu_1 = \mu_2$?" There are now two populations. Population 1 has mean μ_1 and variance σ_1^2; population 2 has mean μ_2 and variance σ_2^2. From population 1 we draw a random sample of size n_1 and compute $\bar{X}_1$ and S_1^2; from population 2 we draw an *independent* random sample of size n_2 and compute $\bar{X}_2$ and S_2^2. Figure 5.1 shows the setup.

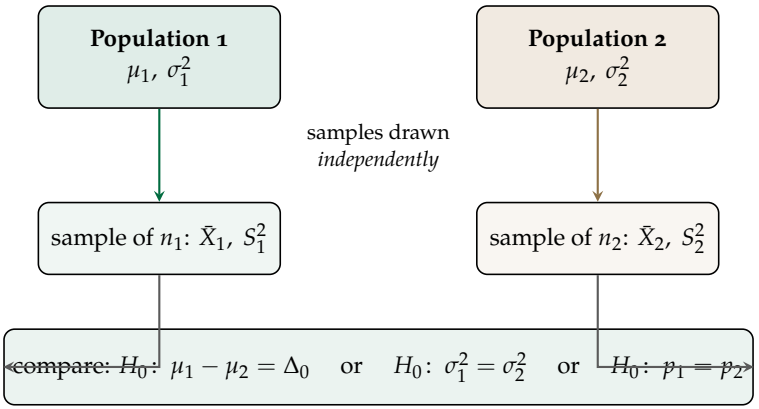


Figure 5.1. The two-sample framework. Two populations are observed through two independent random samples, and inference targets the *difference* or *ratio* of their parameters.

The natural point estimator of the difference $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2$, and Chapter 2 already derived its sampling distribution in (2.13): when both populations are normal, or both sample sizes are large,

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

Please pay attention to the plus sign in the variance, because this is a common trap: even though we are *subtracting* the means, the variances *add*. Both samples contribute noise to the comparison, and neither cancels the other. Everything in this chapter is a standardization of this one distribution (or of its t, F, and proportion analogues), so the seven-step testing procedure of Chapter 4 (Procedure 4.2) carries over unchanged. Only the test statistic and its reference distribution change from section to section.

Safety lens. Almost every model-evaluation report is a two-sample problem: version 2 of a perception model against version 1, evaluated on independent scenario sets. Naming the two populations precisely—which model, which scenario distribution—is the first step of a defensible comparison.

Reference. Montgomery & Runger, 6th ed., Section 10.1.

5.2 *Two Means, Variances Known: the Two-Sample z*

Suppose both population variances σ_1^2 and σ_2^2 are known, from long production history or instrument specifications. We test whether the difference in means equals a stated value Δ_0 (very often $\Delta_0 = 0$):

$$H_0:\ \mu_1 - \mu_2 = \Delta_0 \qquad \text{vs} \qquad H_1:\ \mu_1 - \mu_2 \neq \Delta_0.$$

Standardizing $\bar{X}_1 - \bar{X}_2$ under H_0 gives the test statistic

$$Z_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}, \tag{5.1}$$

which follows the standard normal distribution when H_0 is true. The structure is identical to the one-sample z statistic (4.3): an estimate, minus its hypothesized value, divided by its standard error. Only the standard error is new, and it is exactly the standard error of a difference from (2.14). Table 5.1 collects the decision rules and P-values for the three alternatives; they are the same rules as in Chapter 4, applied to the new statistic.

Alternative H_1	**Reject H_0 if**	**P-value**
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$	$2\,[1 - \Phi(\lvert z_0 \rvert)]$
$\mu_1 - \mu_2 > \Delta_0$	$Z_0 > z_\alpha$	$1 - \Phi(z_0)$
$\mu_1 - \mu_2 < \Delta_0$	$Z_0 < -z_\alpha$	$\Phi(z_0)$

Table 5.1. Decision rules for the two-sample z test. The pooled and Welch t tests of the next sections use the same three rows with t critical values in place of z.

Inverting the test in the usual way gives the confidence interval. A $100(1-\alpha)\%$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \;\pm\; z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{5.2}$$

Compare with the one-sample interval (3.1): the center is now a difference of sample means, and the standard error combines two sample sizes. If this interval contains 0, a two-sided test at the same α fails to reject $H_0\colon\ \mu_1 = \mu_2$; the CI–test duality from Chapter 4 applies to every interval in this chapter.

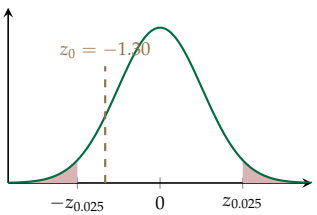


Figure 5.2. Two-sided rejection region at $\alpha = 0.05$ for Example 5.1. The observed $z_0 = -1.30$ does not reach either shaded tail.

Example 5.1 — Two container terminals, known variances

A logistics company routes import containers through two terminals and wants to know whether the mean gate-processing time differs between them. Years of gate-system data make the standard deviations effectively known. Terminal A: $n_1 = 15$ trucks sampled, $\bar{x}_1 = 48.5$ minutes, $\sigma_1 = 6.0$. Terminal B: $n_2 = 17$ trucks, $\bar{x}_2 = 51.6$ minutes, $\sigma_2 = 7.5$. Test at $\alpha = 0.05$ whether the

mean processing times differ, report the P-value, and give a 95% confidence interval for $\mu_1 - \mu_2$.

Solution.

Step 1 — Identify the setting.

The parameter is $\mu_1 - \mu_2$. Both variances are known and the samples are independent, so the two-sample z test (5.1) applies.

Step 2 — Hypotheses.

$H_0\colon\ \mu_1 - \mu_2 = 0$ vs $H_1\colon\ \mu_1 - \mu_2 \neq 0$, with $\alpha = 0.05$.

Step 3 — Compute the statistic.

The standard error is

$$\sqrt{\frac{6.0^2}{15} + \frac{7.5^2}{17}} = \sqrt{2.400 + 3.309} = \sqrt{5.709} = 2.389,$$

so

$$z_0 = \frac{(48.5 - 51.6) - 0}{2.389} = \frac{-3.1}{2.389} = -1.297 \approx -1.30.$$

Step 4 — Decide.

Two-sided at $\alpha = 0.05$: reject if $|z_0| > z_{0.025} = 1.96$. Here $|z_0| = 1.30 < 1.96$, so the statistic is not in the rejection region (Figure 5.2). The P-value is $2\,[1 - \Phi(1.30)] = 2\,(1 - 0.9032) = 0.1936$, well above 0.05.

Step 5 — Interval.

By (5.2),

$$-3.1 \pm 1.96\,(2.389) = -3.1 \pm 4.683 \quad\Longrightarrow\quad (-7.78,\ 1.58) \text{ minutes.}$$

Step 6 — Conclude in context.

We fail to reject H_0: at $\alpha = 0.05$ there is not sufficient evidence that the terminals differ in mean gate-processing time. The interval says the data are consistent with anything from Terminal A being 7.8 minutes faster to 1.6 minutes slower, and it contains 0, consistent with the test. The answer is $\boxed{z_0 = -1.30,\ \text{fail to reject } H_0,\ \text{CI } (-7.78,\ 1.58)}$.

Engineering judgment. "Known variance" is an engineering claim, not a mathematical convenience. It is defensible when a mature, instrumented process has years of stable history. For a new process or a new sensor, treat σ as unknown and use the t procedures that follow.

Reference. Montgomery & Runger, 6th ed., Section 10.2.

5.3 *Two Means, Variances Unknown but Equal: the Pooled t*

In practice σ_1^2 and σ_2^2 are almost never known; we estimate them with S_1^2 and S_2^2. Two cases arise, and they lead to two different procedures. In the first case we are willing to assume the two populations share a common variance, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. This is often reasonable when the two samples come from the same kind of process under two treatments: the treatment may shift the mean without changing the spread. In the second case the variances may differ, and the next section handles it.

Pooled variance. The weighted average of two sample variances, with weights equal to their degrees of freedom; it estimates the common variance σ^2 when $\sigma_1^2 = \sigma_2^2$.

If both samples estimate the *same* σ^2, throwing either estimate away would waste data. Instead we pool them:

$$S_p^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}. \tag{5.3}$$

The pooled variance is a weighted average of S_1^2 and S_2^2, with weights $(n_1 - 1)$ and $(n_2 - 1)$: the sample with more observations, and therefore the more reliable variance estimate, gets more weight. When $n_1 = n_2$ the weights are equal and S_p^2 is the simple average of the two sample variances. Note that S_p^2 is *not* the variance of the merged data set—merging would mix the two group means into the spread and inflate the estimate whenever $\mu_1 \neq \mu_2$.

Replacing σ by S_p in the two-sample z statistic gives the *pooled t statistic*:

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}, \tag{5.4}$$

which under H_0 follows the t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. The decision rules are the three rows of Table 5.1 with $t_{\alpha/2,\nu}$ and $t_{\alpha,\nu}$ in place of the z critical values. The matching $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \;\pm\; t_{\alpha/2,\, n_1+n_2-2}\; s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{5.5}$$

Insight. Why $n_1 + n_2 - 2$ degrees of freedom. Each sample variance is computed around its own sample mean, and estimating a mean costs one degree of freedom, exactly as in the one-sample case. Sample 1 therefore contributes $n_1 - 1$ independent

squared deviations and sample 2 contributes $n_2 - 1$. The pooled estimate stacks them: $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. More degrees of freedom means a smaller t critical value and a narrower interval—this is the reward for being able to assume a common variance. The assumption is not free, though: if it is wrong, the standard error in (5.4) is wrong, and the stated α is no longer the true type I error rate.

Startup lens. Early-stage experiments run on small pilots: eight users on the old flow, ten on the new one. With samples this small, the pooled t's extra degrees of freedom visibly tighten the interval—but only claim the equal-variance assumption if the two user groups genuinely face the same product surface.

Example 5.2 — Session length under two layouts

Your startup tests two dashboard layouts with separate pilot groups and records the mean session length in minutes. Layout 1: $n_1 = 12$ users, $\bar{x}_1 = 22.7$, $s_1 = 1.8$. Layout 2: $n_2 = 10$ users, $\bar{x}_2 = 21.3$, $s_2 = 2.1$. Assume the population variances are equal (the ratio $s_2/s_1 = 1.17$ raises no alarm; Section 5.7 gives the formal check). Test at $\alpha = 0.05$ whether the mean session lengths differ, bound the P-value from the t table, and give a 95% confidence interval.

Solution.

Step 1 — Identify the setting.

Parameter $\mu_1 - \mu_2$; variances unknown but assumed equal; independent samples. Use the pooled t (5.4).

Step 2 — Hypotheses.

$H_0\colon\ \mu_1 - \mu_2 = 0$ vs $H_1\colon\ \mu_1 - \mu_2 \neq 0$, $\alpha = 0.05$.

Step 3 — Pool the variances.

By (5.3),

$$s_p^2 = \frac{(12-1)(1.8)^2 + (10-1)(2.1)^2}{12+10-2} = \frac{11(3.24) + 9(4.41)}{20} = \frac{35.64 + 39.69}{20} = 3.767,$$

so $s_p = \sqrt{3.767} = 1.941$.

Step 4 — Compute the statistic.

$$t_0 = \frac{(22.7 - 21.3) - 0}{1.941\sqrt{\dfrac{1}{12} + \dfrac{1}{10}}} = \frac{1.4}{1.941 \times 0.4282} = \frac{1.4}{0.8311} = 1.685, \qquad \nu = 12 + 10 - 2 = 20.$$

Step 5 — Decide.

From Table 5.2, $t_{0.025,20} = 2.086$. Since $|t_0| = 1.685 < 2.086$, we fail to reject. For the P-value, the same row gives $t_{0.10,20} = 1.325$ and $t_{0.05,20} = 1.725$; since $1.325 < 1.685 < 1.725$, the one-tail area is between 0.05 and 0.10, so $0.10 < P < 0.20$ for the two-sided test.

Step 6 — Interval and conclusion.

By (5.5),

$$1.4 \pm 2.086\,(0.8311) = 1.4 \pm 1.734 \quad \Longrightarrow \quad (-0.33,\ 3.13) \text{ minutes},$$

which contains 0. At $\alpha = 0.05$ there is not sufficient evidence that the layouts differ in mean session length; a true difference of up to about 3 minutes in either direction is still compatible with these small pilots. The answer is $\boxed{t_0 = 1.685,\ \text{fail to reject } H_0,\ 0.10 < P < 0.20}$.

ν	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$
18	1.330	1.734	2.101	2.552
19	1.328	1.729	2.093	2.539
20	1.325	1.725	2.086	2.528
21	1.323	1.721	2.080	2.518

Table 5.2. Excerpt of the t table around $\nu = 20$. The highlighted cell is $t_{0.025,20} = 2.086$, used in Example 5.2; the neighboring columns bound its P-value.

Lecture 12. This section is covered by Lecture 12.

Reference. Montgomery & Runger, 6th ed., Section 10.2.2.

Lecture 12 starts here — *Two-sample inference, part 2*

5.4 Two Means, Unequal Variances: the Welch t

When the two population variances cannot be assumed equal, do *not* pool. Pooling forces one number to stand for two different spreads, and the resulting standard error is too small for one group and too large for the other. Instead, keep the two sample variances separate:

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}. \tag{5.6}$$

This statistic looks like the two-sample z (5.1) with S_1^2, S_2^2 replacing σ_1^2, σ_2^2, but its distribution is no longer exactly t with a simple degree-of-freedom count. It is

approximately t_ν, with ν given by the Welch–Satterthwaite formula:

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}. \tag{5.7}$$

Welch–Satterthwaite degrees of freedom. The effective degrees of freedom of the separate-variances t statistic, computed from (5.7) and always rounded *down* to an integer.

Please pay attention to the rounding rule, because it is the opposite of the sample-size rule: sample sizes round *up*, but the Welch degrees of freedom round *down*. Rounding down is the conservative direction: it gives a slightly larger critical value and a slightly wider interval. Three facts help you sanity-check a computed ν:

- If $s_1^2 \approx s_2^2$ and $n_1 \approx n_2$, then $\nu \approx n_1 + n_2 - 2$, the pooled value.
- If the variances are very different, ν can be much smaller than $n_1 + n_2 - 2$; it can never fall below $\min(n_1 - 1,\ n_2 - 1)$.
- A smaller ν means a larger critical value: unequal variances cost information (Figure 5.3).

The confidence interval keeps the same shape as always:

$$(\bar{x}_1 - \bar{x}_2) \ \pm \ t_{\alpha/2,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

with ν from (5.7).

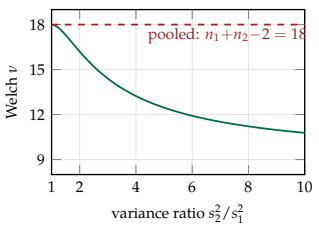


Figure 5.3. Welch degrees of freedom (5.7) for $n_1 = n_2 = 10$ as the variance ratio grows. At equal variances $\nu = 18$; strongly unequal variances push ν toward $n - 1 = 9$.

Example 5.3 — Two red-team scenario suites

An AI safety team evaluates a perception model's detection accuracy (in percent) on two scenario suites built by different red teams. Suite 1: $n_1 = 12$ scenarios, $\bar{x}_1 = 85.6$, $s_1 = 3.2$. Suite 2: $n_2 = 15$ scenarios, $\bar{x}_2 = 81.9$, $s_2 = 6.8$. Does the mean accuracy differ between suites at $\alpha = 0.05$?

Solution.

Step 1 — Check the variance ratio.

$s_2/s_1 = 6.8/3.2 = 2.13 > 2$, so the equal-variance assumption is not credible. Use Welch's t (5.6).

Step 2 — Hypotheses.

$H_0\colon\ \mu_1 - \mu_2 = 0$ vs $H_1\colon\ \mu_1 - \mu_2 \neq 0$, $\alpha = 0.05$.

Step 3 — Compute the statistic.

$$\frac{s_1^2}{n_1} = \frac{10.24}{12} = 0.8533, \qquad \frac{s_2^2}{n_2} = \frac{46.24}{15} = 3.0827,$$

$$t_0 = \frac{(85.6 - 81.9) - 0}{\sqrt{0.8533 + 3.0827}} = \frac{3.7}{\sqrt{3.9360}} = \frac{3.7}{1.984} = 1.865.$$

Step 4 — Degrees of freedom.

By (5.7),

$$\nu = \frac{(0.8533 + 3.0827)^2}{\dfrac{0.8533^2}{11} + \dfrac{3.0827^2}{14}} = \frac{15.492}{0.0662 + 0.6788} = \frac{15.492}{0.7450} = 20.80 \implies \nu = 20 \text{ (round down)}.$$

Step 5 — Decide and conclude.

$t_{0.025,\,20} = 2.086$, and $|t_0| = 1.865 < 2.086$: fail to reject. At $\alpha = 0.05$ there is not sufficient evidence that the model's mean accuracy differs across the two suites, even though the point difference is 3.7 percentage points; the noisier second suite makes the comparison imprecise. The answer is $\boxed{t_0 = 1.865,\ \nu = 20,\ \text{fail to reject } H_0}$.

Safety lens. Red-team suites often differ in difficulty spread by design: one probes a narrow attack family, the other samples broadly. Unequal variances are then the expected situation, not a nuisance—defaulting to Welch's t is standard practice in evaluation pipelines for exactly this reason.

Reference. Montgomery & Runger, 6th ed., Section 10.4.

5.5 *Paired Samples: the Paired t*

Paired design. A study in which each observation in sample 1 is naturally matched to one observation in sample 2—the same unit measured twice, or two treatments applied to the same specimen.

Sometimes the two samples are not independent at all. Each observation in the first sample is naturally linked to one observation in the second: before-and-after measurements on the same machine, two measurement methods applied to the same specimen, the same prompt scored under two model versions. The keywords are ''same,'' ''matched,'' ''before/after,'' and data printed in corresponding pairs.

The correct analysis does not treat the columns as two samples. It reduces the problem to *one* sample by computing the differences $D_j = X_{1j} - X_{2j}$ for each pair $j = 1, \ldots, n$, then applies the one-sample t machinery of Chapter 4 to the D_j.

With

$$\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j, \qquad S_D = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(D_j - \bar{D}\right)^2},$$

the hypotheses $H_0\colon\ \mu_D = \Delta_0$ vs $H_1\colon\ \mu_D \neq \Delta_0$ (often $\Delta_0 = 0$) are tested with

$$T_0 = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}, \tag{5.8}$$

which under H_0 follows t_{n-1}, where n is the number of *pairs*. The $100(1-\alpha)\%$ confidence interval for μ_D is

$$\bar{d} \;\pm\; t_{\alpha/2,\,n-1}\,\frac{s_D}{\sqrt{n}}. \tag{5.9}$$

Insight. Why pairing helps. Write the difference for pair j as $D_j = X_{1j} - X_{2j}$. If the two measurements on the same unit are positively correlated with correlation ρ—a strong unit is strong both times—then $\mathrm{Var}(D_j) = \sigma_1^2 + \sigma_2^2 - 2\rho\,\sigma_1\sigma_2$. The subtraction cancels the unit-to-unit variability that both measurements share, so the differences are far less noisy than the raw columns whenever ρ is large. An independent-samples analysis has no access to this cancellation: its standard error carries the full $\sigma_1^2 + \sigma_2^2$. Pairing trades degrees of freedom ($n-1$ instead of $2n-2$) for a much smaller standard error, and when units differ a lot, that trade is heavily favorable.

Example 5.4 — Fuel efficiency before and after a tune-up

Eight fleet vehicles are tested for fuel efficiency (km/L) before and after an engine tune-up. The data and differences $d_j = \text{after} - \text{before}$ appear in Table 5.3 and Figure 5.4. Does the tune-up *improve* fuel efficiency at $\alpha = 0.05$? Give a 95% lower confidence bound for the mean improvement. Then repeat the test the *wrong* way—as two independent samples—and compare.

Vehicle	1	2	3	4	5	6	7	8
Before (x_{1j})	12.1	14.3	11.8	13.5	15.2	12.7	14.0	13.1
After (x_{2j})	12.9	14.8	12.5	14.1	15.6	13.4	14.9	13.6
$d_j = x_{2j} - x_{1j}$	0.8	0.5	0.7	0.6	0.4	0.7	0.9	0.5

Table 5.3. Fuel efficiency (km/L) of the same eight vehicles before and after the tune-up, with per-vehicle differences.

Solution.

Step 1 — Recognize the design.

The same eight vehicles are measured twice, so the data are paired. Work with the differences; use the paired t (5.8) with $n = 8$ pairs.

Step 2 — Hypotheses.

With $d_j = \text{after} - \text{before}$, improvement means positive differences: $H_0 : \mu_D = 0$ vs $H_1 : \mu_D > 0$, $\alpha = 0.05$.

Step 3 — Summarize the differences.

$$\bar{d} = \frac{0.8 + 0.5 + 0.7 + 0.6 + 0.4 + 0.7 + 0.9 + 0.5}{8} = \frac{5.1}{8} = 0.6375,$$

$$\sum d_j^2 = 3.45, \qquad s_D^2 = \frac{\sum d_j^2 - n\bar{d}^2}{n-1} = \frac{3.45 - 8(0.6375)^2}{7} = \frac{0.19875}{7} = 0.02839,$$

so $s_D = 0.1685$ and the standard error is $s_D/\sqrt{8} = 0.1685/2.828 = 0.05957$.

Step 4 — Test.

$$t_0 = \frac{0.6375 - 0}{0.05957} = 10.70, \qquad \nu = 8 - 1 = 7, \qquad t_{0.05,7} = 1.895.$$

Since $10.70 \gg 1.895$, reject H_0 decisively; the P-value is far below 0.0005. A 95% lower bound for μ_D (one-sided version of (5.9)) is $0.6375 - 1.895\,(0.05957) = 0.525$ km/L.

Step 5 — The wrong analysis, for contrast.

Treating the columns as independent samples gives $\bar{x}_1 = 13.34$, $\bar{x}_2 = 13.98$, $s_1 = 1.148$, $s_2 = 1.069$, and a pooled analysis by (5.3) and (5.4) yields $s_p = 1.109$,

$$t_0 = \frac{0.6375}{1.109\sqrt{\frac{1}{8} + \frac{1}{8}}} = \frac{0.6375}{0.5544} = 1.15, \qquad \nu = 14, \qquad t_{0.05,14} = 1.761,$$

which *fails* to reject. The vehicles differ from each other by ± 1.1 km/L, and that vehicle-to-vehicle spread buries the consistent 0.6 km/L improvement. Pairing removes it: every line in Figure 5.4 tilts upward.

Step 6 — Conclude.

The tune-up improves mean fuel efficiency; we are 95% confident the improvement is at least 0.525 km/L. The answer is $\boxed{t_0 = 10.70,\ \text{reject } H_0,\ \mu_D \geq 0.525 \text{ km/L}}$.

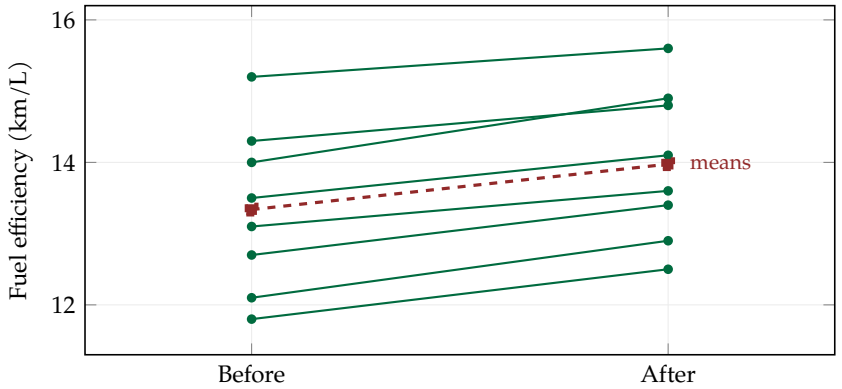


Figure 5.4. The paired fuel-efficiency data of Example 5.4. Each line is one vehicle. The vehicles are spread over almost 4 km/L, yet every line tilts up by a consistent 0.4–0.9 km/L: pairing isolates exactly that tilt, while an independent-samples analysis must fight the full vertical spread.

Engineering judgment. Pairing is a design decision, made before data collection. If you can run both treatments on the same unit—the same vehicle, the same specimen, the same prompt set—you buy a large variance reduction at the price of half as many degrees of freedom. Plan for it; you cannot pair data after the fact if the units were never matched.

Table 5.4 summarizes when each design applies.

	Two independent samples	**Paired samples**
Design	Two separate groups	Same units, two conditions
Sample sizes	May differ ($n_1 \neq n_2$)	Always equal (n pairs)
Parameter	$\mu_1 - \mu_2$	μ_D
Statistic	(5.4) or (5.6)	(5.8)
Degrees of freedom	$n_1 + n_2 - 2$ or (5.7)	$n - 1$
Advantage	Simpler logistics	Removes unit-to-unit variability

Table 5.4. Independent-samples versus paired designs. If the problem says ''same,'' ''matched,'' or ''before/after,'' or prints the data in corresponding pairs, use the paired t.

Reference. Montgomery & Runger, 6th ed., Section 10.6.

5.6 *Two Proportions*

Pooled proportion. The combined success rate $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$, used to estimate the common p when the null hypothesis says $p_1 = p_2$.

Comparisons of rates follow the same pattern. Sample 1 has n_1 trials with X_1 successes, so $\hat{P}_1 = X_1/n_1$; independently, sample 2 has n_2 trials with X_2 successes and $\hat{P}_2 = X_2/n_2$. We test $H_0 : \; p_1 = p_2$ against a one- or two-sided alternative, and both samples should be large: $n_i\hat{p}_i \geq 5$ and $n_i(1 - \hat{p}_i) \geq 5$ for $i = 1, 2$.

Under H_0 the two populations share a common proportion p, and the best estimate of that common value pools the raw counts:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Please pay attention here, because this is a common trap: pool the *counts*, not the proportions. Averaging $\hat{p}_1$ and $\hat{p}_2$ happens to agree only when $n_1 = n_2$. The test statistic standardizes $\hat{P}_1 - \hat{P}_2$ using the pooled estimate in the standard error:

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}\,(1 - \hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \tag{5.10}$$

approximately standard normal under H_0, with the decision rules of Table 5.1. The confidence interval, in contrast, does not assume $p_1 = p_2$, so it uses the *individual* sample proportions in the standard error:

$$(\hat{p}_1 - \hat{p}_2) \;\pm\; z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}. \tag{5.11}$$

Startup lens. A/B testing is exactly this machinery: design A converts x_1 of n_1 visitors, design B converts x_2 of n_2. The test in (5.10) answers "is the lift real?"; the interval in (5.11) answers the more useful business question, "how big could the lift plausibly be?" Report both.

The division of labor mirrors the one-sample case ((4.8) versus (3.7)): the test's standard error is computed under the null hypothesis, the interval's from the data alone.

Example 5.5 — Guardrail false positives, two model versions

A safety team compares the false-positive rates of two guardrail versions on benign prompts. Version A flagged $x_1 = 18$ of $n_1 = 200$ benign prompts ($\hat{p}_1 = 0.090$); version B flagged $x_2 = 32$ of $n_2 = 250$ ($\hat{p}_2 = 0.128$). Do the false-positive rates differ at $\alpha = 0.05$? Give the P-value and a 95% confidence interval for $p_1 - p_2$.

Solution.

Step 1 — Check conditions.

All four counts $(18, 182, 32, 218)$ exceed 5, so the normal approximation is appropriate.

Step 2 — Hypotheses.

$H_0\colon\ p_1 = p_2$ vs $H_1\colon\ p_1 \neq p_2$, $\alpha = 0.05$.

Step 3 — Pool and compute.

$$\hat{p} = \frac{18 + 32}{200 + 250} = \frac{50}{450} = 0.1111,$$

$$z_0 = \frac{0.090 - 0.128}{\sqrt{0.1111(0.8889)\left(\dfrac{1}{200} + \dfrac{1}{250}\right)}} = \frac{-0.038}{\sqrt{0.0008888}} = \frac{-0.038}{0.02981} = -1.27.$$

Step 4 — Decide.

$|z_0| = 1.27 < z_{0.025} = 1.96$: fail to reject. P-value $= 2\,[1 - \Phi(1.27)] = 2\,(1 - 0.8980) = 0.2040$.

Step 5 — Interval.

By (5.11), with individual proportions in the standard error,

$$\sqrt{\frac{0.090(0.910)}{200} + \frac{0.128(0.872)}{250}} = \sqrt{0.0004095 + 0.0004465} = 0.02926,$$

$$-0.038 \pm 1.96\,(0.02926) = -0.038 \pm 0.0573 \quad \Longrightarrow \quad (-0.095,\ 0.019).$$

Step 6 — Conclude.

The interval contains 0, consistent with the test: at $\alpha = 0.05$ there is not sufficient evidence that the two guardrail versions differ in false-positive rate, though a difference as large as 9.5 percentage points in version A's favor is still compatible with the data. The answer is $\boxed{z_0 = -1.27,\ P = 0.204,\ \text{fail to reject } H_0}$.

Reference. Montgomery & Runger, 6th ed., Section 10.5.

5.7 *Two Variances: the F Test*

Sometimes the comparison of interest is variability itself: which supplier is more consistent, which process holds tolerance better. And even when the means are the target, Section 5.3 needs a way to check the assumption $\sigma_1^2 = \sigma_2^2$. Both jobs belong to the F distribution.

F distribution. The distribution of a ratio of two independent sample variances, each divided by its population variance. It has two degree-of-freedom parameters: numerator u and denominator v.

For independent samples from two *normal* populations,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,\, n_2-1},$$

the F distribution with numerator degrees of freedom $u = n_1 - 1$ and denominator degrees of freedom $v = n_2 - 1$. The distribution is skewed to the right and concentrated near 1 (Figure 5.5). Its critical values $f_{\alpha,u,v}$ satisfy $P(F > f_{\alpha,u,v}) = \alpha$ and are tabulated for upper-tail areas only; lower-tail values come from the reciprocal identity

$$f_{1-\alpha,\,u,\,v} = \frac{1}{f_{\alpha,\,v,\,u}}, \tag{5.12}$$

in which the two degrees of freedom *swap places*. Please pay attention to the swap, because it is the most common table-reading error in this section.

To test $H_0\colon\ \sigma_1^2 = \sigma_2^2$, set the ratio under the null:

$$F_0 = \frac{S_1^2}{S_2^2}, \tag{5.13}$$

which under H_0 follows $F_{n_1-1,\,n_2-1}$.

Procedure 5.1 — The two-variance F test.

1. Verify that both populations are plausibly normal (normal probability plots). The F test is *not* robust to non-normality.
2. State $H_0\colon\ \sigma_1^2 = \sigma_2^2$ and the alternative; choose α. For a two-sided test, label the sample with the larger variance as sample 1, so $F_0 \geq 1$ and only the upper critical value is needed.
3. Compute $F_0 = s_1^2/s_2^2$ (5.13), with $u = n_1 - 1$ (numerator) and $v = n_2 - 1$ (denominator).
4. Reject H_0 when: $H_1:\ \sigma_1^2 \neq \sigma_2^2$ and $F_0 > f_{\alpha/2,u,v}$ or $F_0 < f_{1-\alpha/2,u,v}$; $H_1:\ \sigma_1^2 > \sigma_2^2$ and $F_0 > f_{\alpha,u,v}$; $H_1\colon\ \sigma_1^2 < \sigma_2^2$ and $F_0 < f_{1-\alpha,u,v}$. Use (5.12) for any lower-tail value.

5. Conclude in the units of the problem: consistency, tolerance, repeatability.

Inverting the ratio distribution gives a $100(1-\alpha)\%$ confidence interval for the variance ratio:

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{f_{\alpha/2,\,n_1-1,\,n_2-1}} \;\le\; \frac{\sigma_1^2}{\sigma_2^2} \;\le\; \frac{s_1^2}{s_2^2}\cdot f_{\alpha/2,\,n_2-1,\,n_1-1}. \tag{5.14}$$

If this interval contains 1, the data are consistent with equal variances at level α.

Insight. Why the two F values in the interval have swapped degrees of freedom. The interval starts from $P\big(f_{1-\alpha/2,u,v} \le F \le f_{\alpha/2,u,v}\big) = 1-\alpha$ with $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ and solves the two inequalities for σ_1^2/σ_2^2. The lower table value $f_{1-\alpha/2,u,v}$ is rewritten through the reciprocal identity (5.12) as $1/f_{\alpha/2,v,u}$, and after the algebra that reciprocal lands on the *upper* end of the interval with its degrees of freedom reversed. The swap is not a typographical quirk; it is the identity (5.12) doing its work.

Example 5.6 — Consistency of two roller suppliers

A warehouse-automation firm buys conveyor rollers from two suppliers and measures the diameter of incoming lots. Supplier 1: $n_1 = 16$ rollers, $s_1^2 = 0.042$ mm^2. Supplier 2: $n_2 = 21$ rollers, $s_2^2 = 0.018$ mm^2. Diameters are approximately normal for both. Do the suppliers differ in variability at $\alpha = 0.05$? Give a 95% confidence interval for σ_1^2/σ_2^2.

Solution.

Step 1 — Set up.

Parameter: σ_1^2/σ_2^2. Both populations plausibly normal, samples independent: use Procedure 5.1. $H_0\colon\ \sigma_1^2 = \sigma_2^2$ vs $H_1\colon\ \sigma_1^2 \neq \sigma_2^2$, $\alpha = 0.05$. The larger variance is already in sample 1.

Step 2 — Compute.

By (5.13),

$$f_0 = \frac{0.042}{0.018} = 2.333, \qquad u = 15, \quad v = 20.$$

Step 3 — Critical value.

From the F table excerpt (Table 5.5), $f_{0.025,15,20} = 2.57$. Since $f_0 = 2.333 < 2.57$ (Figure 5.5), we fail to reject.

Step 4 — Interval.

By (5.14), with the swapped value $f_{0.025,20,15} = 2.76$ from the same table,

$$2.333 \times \frac{1}{2.57} \;\le\; \frac{\sigma_1^2}{\sigma_2^2} \;\le\; 2.333 \times 2.76 \quad\Longrightarrow\quad (0.91,\ 6.44).$$

Step 5 — Conclude.

The interval contains 1, consistent with the test: at $\alpha = 0.05$ the data do not establish a difference in variability—but note how wide the interval is. Supplier 1 could plausibly be anywhere from equally consistent to 6 times more variable. Variance comparisons need large samples. The answer is $f_0 = 2.333$, fail to reject H_0, CI $(0.91,\ 6.44)$.

	numerator ν_1			
ν_2	10	12	15	20
15	3.06	2.96	2.86	2.76
20	2.77	2.68	2.57	2.46

Table 5.5. Excerpt of the F table, upper 0.025 critical values $f_{0.025,\nu_1,\nu_2}$. Green: $f_{0.025,15,20} = 2.57$ for the test in Example 5.6; gold: the swapped-df value $f_{0.025,20,15} = 2.76$ needed by the interval (5.14).

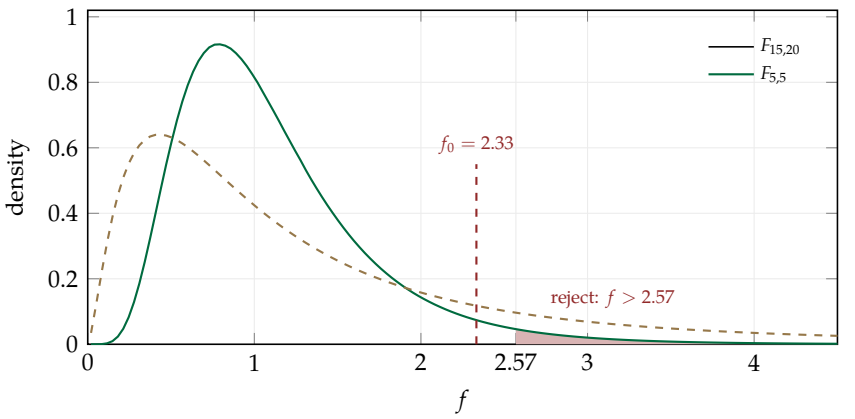


Figure 5.5. The $F_{15,20}$ density with the upper $\alpha/2 = 0.025$ rejection region shaded, for Example 5.6; the observed $f_0 = 2.33$ falls just short of the critical value. The dashed $F_{5,5}$ curve shows how strongly the shape depends on the degrees of freedom.

Engineering judgment. The F test inherits the fragility of the one-sample χ^2 procedures: it is derived exactly from normality and is badly miscalibrated for skewed or heavy-tailed populations. Before comparing variances, look at normal probability plots of both samples. If normality is doubtful, say so in the report rather than quoting a precise P-value.

Reference. Montgomery & Runger, 6th ed., Sections 10.1–10.6.

5.8 *Which Procedure? A Decision Map*

Five procedures for comparing means have now accumulated, and choosing among them is itself an exam skill. The choice depends on three questions, asked in order: are the data paired, are the variances known, and can the unknown variances be assumed equal? Figure 5.6 draws the map and Procedure 5.2 states it in words.

Procedure 5.2 — Choosing the two-sample procedure for means.

1. **Paired?** If each observation in sample 1 is naturally linked to one in sample 2 (same unit, before/after, matched specimens), compute differences and use the paired t (5.8) with $\nu = n - 1$. Stop.
2. **Variances known?** If σ_1^2 and σ_2^2 are known from history or specification, use the two-sample z (5.1). Stop.
3. **Equal variances defensible?** Compare the sample standard deviations: if $s_{\max}/s_{\min} < 2$ (or a two-sided F test, Procedure 5.1, does not reject), the pooled t (5.4) with $\nu = n_1 + n_2 - 2$ is appropriate.
4. **Otherwise**, use the Welch t (5.6) with the degrees of freedom from (5.7), rounded down. When in doubt, Welch is the safe default: it loses little when the variances are equal and protects you when they are not.

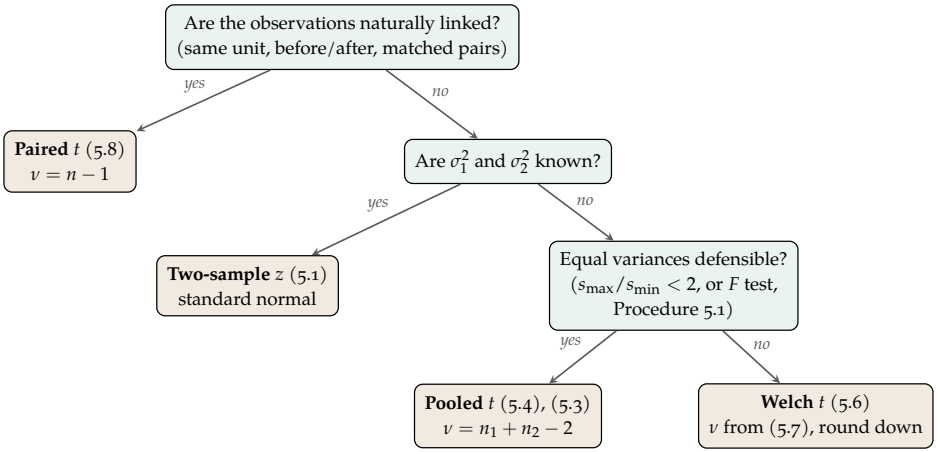


Figure 5.6. Decision map for comparing two means. Ask the questions in order: pairing first, then knowledge of the variances, then their equality. Proportions and variances have their own procedures ((5.10) and (5.13)).

Two reminders complete the map. First, pairing beats everything: if the data are paired, the independence assumed by all the other procedures is violated, and no choice of variance treatment repairs that. Second, the rule of thumb $s_{\max}/s_{\min} < 2$ is a screen, not a proof; when the consequences of a wrong assumption are serious, run the F test of Section 5.7 or simply use Welch.

Startup lens. Consultants and analytics tools default to different choices here—some pool automatically, some always use Welch. When a vendor hands you a two-sample result, the first diligence question is: which branch of Figure 5.6 did they take, and do the data justify it?

Lecture 13. This section is covered by Lecture 13.

Reference. Montgomery & Runger, 6th ed., Chapters 7–10.

Lecture 13 starts here — *Major Exam 1 review*

5.9 *Major Exam 1 Review: Chapters 1–5 in One Map*

Major Exam 1 covers everything so far: the estimation framework of Chapters 1 and 2, the one-sample intervals of Chapter 3, the one-sample tests of Chapter 4, and the two-sample procedures of this chapter. This section compresses all of it into one master table, a short list of relationships worth memorizing, and one procedure for attacking any exam problem. Nothing here is new; everything here is examinable.

Table 5.6 is the map. Every interval and test in the course so far appears as one row: the parameter it targets, the conditions it needs, the statistic, the reference distribution, and the equation numbers where the formulas live. In the exam, your first job on every problem is to find its row.

Parameter	Conditions	Statistic	Distribution	Test	CI
μ	σ known; normal or large n	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$	(4.3)	(3.1)
μ	σ unknown; approx. normal	$T_0 = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$	t_{n-1}	(4.6)	(3.5)
σ^2	normal population (critical)	$\chi^2_0 = \dfrac{(n-1)S^2}{\sigma_0^2}$	χ^2_{n-1}	(4.7)	(3.6)
p	large sample	$Z_0 = \dfrac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$	$N(0,1)$	(4.8)	(3.7)
$\mu_1 - \mu_2$	σ_1, σ_2 known	two-sample Z_0	$N(0,1)$	(5.1)	(5.2)
$\mu_1 - \mu_2$	σ^2's unknown, equal	pooled T_0, S_p^2 from (5.3)	$t_{n_1+n_2-2}$	(5.4)	(5.5)
$\mu_1 - \mu_2$	σ^2's unknown, unequal	Welch T_0	t_ν, ν from (5.7)	(5.6)	—
μ_D	paired data	$T_0 = \dfrac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$	t_{n-1}	(5.8)	(5.9)
σ_1^2/σ_2^2	both populations normal	$F_0 = S_1^2/S_2^2$	$F_{n_1-1,\,n_2-1}$	(5.13)	(5.14)
$p_1 - p_2$	both samples large	Z_0 with pooled $\hat{p}$	$N(0,1)$	(5.10)	(5.11)

Table 5.6. Master table for Major Exam 1: every interval and test from Chapters 3–5. One-sample procedures above the rule, two-sample procedures below it. The Welch CI uses the same standard error as its test with $t_{\alpha/2,\nu}$.

Beyond the table, a short list of relationships shows up in some form on almost every exam:

- **CI–test duality.** A two-sided test at level α rejects exactly when the hypothesized value lies outside the $100(1-\alpha)\%$ two-sided interval. This works for every row of Table 5.6 and lets you answer test questions from an interval and vice versa.
- **Standard error versus standard deviation.** Averages vary less than observations: the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$ (2.5), never σ.
- **Precision is bought at $\sqrt{n}$.** Halving a margin of error requires four times the sample size (3.4).
- **The α–β tradeoff.** For fixed n, shrinking α inflates β; only a larger n shrinks both (4.2).
- **Rounding rules.** Sample sizes round *up*; the Welch degrees of freedom (5.7) round *down*.
- **Which $\hat{p}$ where.** One-sample: p_0 in the test's standard error, $\hat{p}$ in the CI. Two-sample: pooled $\hat{p}$ in the test (5.10), individual $\hat{p}_1, \hat{p}_2$ in the CI (5.11).

Safety lens. Regulators and audit teams read statistical claims the way this review reads exam problems: identify the parameter, check the conditions, verify the arithmetic. Writing your exam solutions in that order is also practice for writing V&V reports that survive review.

- **Normality sensitivity.** The t procedures for means are robust for large n (CLT); the χ^2 and F procedures for variances are not robust at any n.

Finally, the attack procedure. Exam problems rarely announce which row of the master table they belong to; extracting that is the graded skill.

Procedure 5.3 — How to attack an exam problem.

1. **Name the parameter(s).** Mean, variance, or proportion? One population or two? Write the symbol (μ, σ^2, p, $\mu_1 - \mu_2$, μ_D, σ_1^2/σ_2^2, $p_1 - p_2$).
2. **Classify the data.** Paired or independent? Is σ known or estimated? Is the population normal, or is n large? For two-sample means, walk Figure 5.6.
3. **Find the row** in Table 5.6: statistic, distribution, degrees of freedom.
4. **Fix the direction before computing.** Translate the words (''improves,'' ''exceeds,'' ''differs'') into H_1 and decide one- versus two-sided *now*, not after seeing the number.
5. **Compute cleanly.** Standard error first, then the statistic; keep four significant figures in intermediate values.
6. **Decide two ways when possible.** Critical value and P-value (or CI duality) must agree; if they do not, hunt the arithmetic error.
7. **Conclude in context**, with units, in one sentence naming the population and the parameter—not just ''reject H_0.''

Lecture 14 starts here — *Major Exam 1 practice*

Lecture 14. This section is covered by Lecture 14.

5.10 Major Exam 1 Practice: Worked Problems

This section works four exam-style problems end to end, one from each major block of the course: a paired comparison, a one-sample z test with power and

sample size, a proportion interval with sample-size planning, and a pooled two-sample t. Attempt each problem yourself before reading the solution; the exam allows about 20 minutes per problem of this size.

Example 5.7 — Practice A: re-slotting a warehouse (paired t)

A fulfillment operations manager tests whether a re-slotting of storage locations improves the hourly pick-completion yield (%) of picking stations. Eight stations are measured before and after the re-slotting:

Station	1	2	3	4	5	6	7	8
Before	85	78	92	88	76	95	82	90
After	88	82	90	94	80	93	87	94

(a) Compute the differences $d_j = \text{after} - \text{before}$, then $\bar{d}$ and s_D. (b) Test at $\alpha = 0.05$ whether the re-slotting *improved* mean yield. (c) Bound the P-value from the t table. (d) Construct a 95% two-sided CI for μ_D and check consistency with (b).

Solution.

Part (a) — Differences.

d_j: 3, 4, −2, 6, 4, −2, 5, 4, so $\sum d_j = 22$ and $\sum d_j^2 = 126$. Then

$$\bar{d} = \frac{22}{8} = 2.750, \qquad s_D^2 = \frac{126 - 8(2.750)^2}{7} = \frac{65.50}{7} = 9.357, \qquad s_D = 3.059.$$

Part (b) — Test.

Same stations twice: paired (Procedure 5.3, steps 1–2). $H_0: \ \mu_D = 0$ vs $H_1: \ \mu_D > 0$, $\alpha = 0.05$. By (5.8),

$$t_0 = \frac{2.750}{3.059/\sqrt{8}} = \frac{2.750}{1.082} = 2.543, \qquad \nu = 7, \qquad t_{0.05,7} = 1.895.$$

Since $2.543 > 1.895$, reject H_0: the re-slotting improved mean yield.

Part (c) — P-value bounds.

With $\nu = 7$: $t_{0.025,7} = 2.365$ and $t_{0.01,7} = 2.998$. Since $2.365 < 2.543 < 2.998$, the one-tail P-value satisfies $\boxed{0.01 < P < 0.025}$.

Part (d) — Interval.

By (5.9),

$$2.750 \pm 2.365\,(1.082) = 2.750 \pm 2.558 \quad \Longrightarrow \quad (0.19,\ 5.31) \text{ percentage points.}$$

The interval excludes 0 on the positive side, consistent with rejecting in (b). The answer is $\boxed{t_0 = 2.543,\ \text{reject } H_0}$.

Example 5.8 — Practice B: your startup's inspection drones (z test, β, n)

Your startup builds battery-powered inspection drones. The current design has mean flight endurance $\mu_0 = 100$ minutes, and endurance is normal with known $\sigma^2 = 144$ min^2. A new thermal-management system is meant to *increase* endurance. You test $n = 36$ retrofitted drones and observe $\bar{x} = 104$ minutes. (a) Test the claim at $\alpha = 0.05$ and interpret. (b) If the true mean is actually $\mu = 106$, find the probability of a type II error. (c) What sample size detects the shift to 106 with power 0.90 at $\alpha = 0.05$?

Solution.

Part (a) — Test.

One population, σ known: one-sample z (4.3). $H_0\colon\ \mu = 100$ vs $H_1\colon\ \mu > 100$, $\alpha = 0.05$.

$$z_0 = \frac{104 - 100}{12/\sqrt{36}} = \frac{4}{2} = 2.00, \qquad z_{0.05} = 1.645.$$

Since $2.00 > 1.645$, reject H_0; the P-value is $1 - \Phi(2.00) = 0.0228$. There is evidence at the 5% level that the thermal system increases mean endurance beyond 100 minutes—the pitch deck claim survives, at this significance level.

Part (b) — Type II error at $\mu = 106$.

The test fails to reject when $\bar{X} \le \bar{x}_c$, where

$$\bar{x}_c = \mu_0 + z_{0.05}\frac{\sigma}{\sqrt{n}} = 100 + 1.645\,(2) = 103.29.$$

Under the true mean 106 (definition (4.2)),

$$\beta = P(\bar{X} \le 103.29 \mid \mu = 106) = P\left(Z \le \frac{103.29 - 106}{2}\right) = \Phi(-1.36) = 0.0877 \approx 0.088.$$

The test misses a genuine 6-minute improvement about 9% of the time.

Part (c) — Sample size for power 0.90**.**

Here $z_\alpha = 1.645$, $z_\beta = z_{0.10} = 1.282$, and $\delta = 106 - 100 = 6$; the one-sided version of the power-based sample-size formula (compare (4.11)) gives

$$n = \left\lceil \frac{(z_\alpha + z_\beta)^2\,\sigma^2}{\delta^2} \right\rceil = \left\lceil \frac{(1.645 + 1.282)^2 (144)}{36} \right\rceil = \lceil 34.27 \rceil = 35.$$

Sample sizes round up, so test $\boxed{n = 35}$ drones.

Example 5.9 — Practice C: perception-model false positives (p, CI, n)

A validation lab reviews $n = 250$ alerts raised by a perception model on held-out driving scenes; independent human raters judge $x = 40$ of them to be false positives. (a) Check the large-sample conditions and give a 95% CI for the true false-positive proportion p. (b) A follow-up study must achieve a margin of error of at most $E = 0.04$ at 95% confidence; find the required n using part (a) as a prior estimate. (c) Find n with no prior estimate. (d) Explain why (c) exceeds (b).

Solution.

Part (a) — Interval.

$\hat{p} = 40/250 = 0.160$. Conditions: $n\hat{p} = 40 \ge 5$ and $n(1 - \hat{p}) = 210 \ge 5$. By (3.7),

$$0.160 \pm 1.96\sqrt{\frac{0.160(0.840)}{250}} = 0.160 \pm 1.96\,(0.02319) = 0.160 \pm 0.0454,$$

so the interval is $\boxed{(0.115,\ 0.205)}$.

Part (b) — Sample size with a prior.

By (3.8),

$$n = \left\lceil \frac{z_{0.025}^2\, \hat{p}(1-\hat{p})}{E^2} \right\rceil = \left\lceil \frac{(1.96)^2(0.160)(0.840)}{(0.04)^2} \right\rceil = \lceil 322.7 \rceil = 323.$$

Part (c) — Conservative sample size.

With no prior, use $\hat{p} = 0.5$, which maximizes $\hat{p}(1-\hat{p})$ at 0.25:

$$n = \left\lceil \frac{(1.96)^2(0.25)}{(0.04)^2} \right\rceil = \lceil 600.3 \rceil = 601.$$

Part (d) — Why larger.

The conservative formula plans for the worst-case variability $p = 0.5$. Since $0.160 \times 0.840 = 0.134 < 0.25$, the prior-informed plan needs barely more than half the scenes. The answers are $\boxed{n = 323}$ with the prior and $\boxed{n = 601}$ without.

Example 5.10 — Practice D: two onboarding flows (pooled t)

Your startup pilots two onboarding flows with independent user groups and scores each user's first-week engagement (points). Flow A: $n_1 = 10$, $\bar{x}_1 = 85$, $s_1 = 4$. Flow B: $n_2 = 10$, $\bar{x}_2 = 80$, $s_2 = 6$. Assume equal population variances ($s_2/s_1 = 1.5 < 2$). (a) Test at $\alpha = 0.05$ whether Flow A yields *higher* mean engagement. (b) Bound the P-value. (c) Give a 95% two-sided CI for $\mu_1 - \mu_2$ and check consistency.

Solution.

Part (a) — Test.

Independent samples, variances unknown but assumed equal: pooled t. $H_0\colon\ \mu_1 - \mu_2 = 0$ vs $H_1\colon\ \mu_1 - \mu_2 > 0$, $\alpha = 0.05$. By (5.3),

$$s_p^2 = \frac{9(16) + 9(36)}{18} = \frac{468}{18} = 26, \qquad s_p = 5.099,$$

and by (5.4),

$$t_0 = \frac{85 - 80}{5.099\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{5.0}{2.280} = 2.193, \qquad \nu = 18, \qquad t_{0.05,\,18} = 1.734.$$

Since $2.193 > 1.734$, reject H_0: Flow A produces higher mean engagement at the 5% level.

Part (b) — P-value bounds.

With $\nu = 18$: $t_{0.025,\,18} = 2.101$ and $t_{0.01,\,18} = 2.552$; since $2.101 < 2.193 < 2.552$, the one-tail P-value satisfies $\boxed{0.01 < P < 0.025}$.

Part (c) — Interval.

By (5.5),

$$5.0 \pm 2.101\,(2.280) = 5.0 \pm 4.791 \quad \Longrightarrow \quad (0.21,\ 9.79) \text{ points.}$$

The interval excludes 0, consistent with part (a); but it is wide—the advantage could be anywhere from negligible to nearly 10 points. Before rebuilding the product around Flow A, run a larger test. The answer is $\boxed{t_0 = 2.193,\ \text{reject } H_0,\ \text{CI } (0.21,\ 9.79)}$.

Chapter Summary

This chapter extended one-sample inference to comparisons. For two means there are four procedures, chosen by Procedure 5.2: the two-sample z (5.1) when variances are known, the pooled t (5.4) when they are unknown but equal (pooling by (5.3)), the Welch t (5.6) with degrees of freedom (5.7) when they are unequal, and the paired t (5.8) when the observations are matched. Pairing removes unit-to-unit variability and often multiplies the effective precision. Two proportions are compared with the pooled-proportion z (5.10) and the individual-proportion interval (5.11); two variances with the F statistic (5.13) and interval (5.14), both requiring normal populations. The CI–test duality holds for every procedure. The chapter closed with the Major Exam 1 master table (Table 5.6) and attack procedure (Procedure 5.3). Table 5.7 collects the formulas.

Quantity	Formula	Eq.
Two-sample z statistic	$Z_0 = \dfrac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$	(5.1)
CI, variances known	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$	(5.2)
Pooled variance	$S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$	(5.3)
Pooled t statistic	$T_0 = \dfrac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{S_p\sqrt{1/n_1 + 1/n_2}}$, $\nu = n_1+n_2-2$	(5.4)
CI, pooled	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\nu}\, s_p\sqrt{1/n_1 + 1/n_2}$	(5.5)
Welch t statistic	$T_0 = \dfrac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	(5.6)
Welch df (round down)	$\nu = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$	(5.7)
Paired t statistic	$T_0 = \dfrac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$, $\nu = n-1$	(5.8)
Paired CI	$\bar{d} \pm t_{\alpha/2,n-1}\, s_D/\sqrt{n}$	(5.9)
Two-proportion z	$Z_0 = \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$, $\hat{p} = \dfrac{x_1+x_2}{n_1+n_2}$	(5.10)
CI for $p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	(5.11)
F statistic	$F_0 = S_1^2/S_2^2 \sim F_{n_1-1,\,n_2-1}$	(5.13)
CI for σ_1^2/σ_2^2	$\left[\dfrac{s_1^2}{s_2^2}\dfrac{1}{f_{\alpha/2,n_1-1,n_2-1}},\ \dfrac{s_1^2}{s_2^2} f_{\alpha/2,n_2-1,n_1-1}\right]$	(5.14)

Table 5.7. Key formulas of Chapter 5 at a glance.

Exercises

Two independent means

Exercise 5.1. A port authority compares mean container dwell time at two yards, using gate-system records whose long-run standard deviations are known. Yard A: $n_1 = 36$ containers, $\bar{x}_1 = 58.2$ hours, $\sigma_1 = 9.0$. Yard B: $n_2 = 49$ containers, $\bar{x}_2 = 63.0$ hours, $\sigma_2 = 10.5$. (a) Test $H_0\colon\ \mu_1 = \mu_2$ against a two-sided alternative

at $\alpha = 0.05$ and give the P-value. (b) Construct the matching 95% CI for $\mu_1 - \mu_2$.

Exercise 5.2. Your startup measures the time (minutes) for pilot customers to complete a setup task under two support plans. Plan A: $n_1 = 8$, $\bar{x}_1 = 42.0$, $s_1 = 6.0$. Plan B: $n_2 = 10$, $\bar{x}_2 = 36.0$, $s_2 = 5.0$. Assume equal population variances. Test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ at $\alpha = 0.05$, bound the P-value, and give the 95% CI for $\mu_1 - \mu_2$.

Exercise 5.3. An evaluation team scores an LLM on two benchmark suites (mean score per item, 0–1 scale). Suite 1: $n_1 = 10$, $\bar{x}_1 = 0.842$, $s_1 = 0.031$. Suite 2: $n_2 = 16$, $\bar{x}_2 = 0.815$, $s_2 = 0.078$. (a) Explain why the pooled t is a poor choice here. (b) Carry out the appropriate test of $H_0 : \mu_1 = \mu_2$ at $\alpha = 0.05$, showing the degrees-of-freedom computation.

Paired samples

Exercise 5.4. Six pilot customers complete the same workflow before and after a UI redesign (task time, minutes): before 14.2, 11.8, 16.0, 12.4, 13.5, 15.1; after 12.8, 11.0, 14.6, 12.0, 12.4, 14.3. Test at $\alpha = 0.05$ whether the redesign *reduces* mean task time, and give a 95% two-sided CI for the mean reduction.

Exercise 5.5. A safety team wants to compare two model versions on evaluation prompts. Design I: score the *same* 12 prompts under both versions. Design II: score 12 prompts under version 1 and a different, independent set of 12 prompts under version 2. (a) Name the analysis and its degrees of freedom for each design. (b) Which design typically detects a given quality difference with higher power, and why? (c) Give one situation where Design II is the right choice anyway.

Two proportions

Exercise 5.6. An A/B test of a signup page shows: design A, $x_1 = 60$ conversions from $n_1 = 500$ visitors; design B, $x_2 = 36$ from $n_2 = 450$. Test $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$ at $\alpha = 0.05$ and give the P-value.

Exercise 5.7. A guardrail update is meant to lower the false-positive rate on benign prompts. Version 1 (old): $x_1 = 27$ false positives in $n_1 = 300$ prompts. Version 2 (new): $x_2 = 15$ in $n_2 = 300$. (a) Give a 95% CI for $p_1 - p_2$. (b) Using the CI–test duality, what does a two-sided test at $\alpha = 0.05$ conclude? What do you recommend to the team?

Two variances

Exercise 5.8. Two carriers' delivery lateness (minutes) is compared for consistency. Carrier 1: $n_1 = 13$ deliveries, $s_1^2 = 48$. Carrier 2: $n_2 = 16$, $s_2^2 = 20$. Lateness is approximately normal for both. Test $H_0 \colon \sigma_1^2 = \sigma_2^2$ vs $H_1 \colon \sigma_1^2 \neq \sigma_2^2$ at $\alpha = 0.05$, given $f_{0.025,\,12,\,15} = 2.96$.

Exercise 5.9. Two filling machines are sampled: machine 1, $n_1 = 16$, $s_1^2 = 10.8$ g^2; machine 2, $n_2 = 16$, $s_2^2 = 4.5$ g^2; both fill weights approximately normal. (a) Construct a 95% CI for σ_1^2/σ_2^2, given $f_{0.025,\,15,\,15} = 2.86$. (b) An engineer wants to pool these variances in a two-sample t test. Based on your interval, comment.

Exam-style problems

Exercise 5.10. Two loading docks process identical trailer types, and management wants to compare mean unload cycle times. Dock 1: $n_1 = 12$ cycles, $\bar{x}_1 = 18.5$ min, $s_1 = 2.4$. Dock 2: $n_2 = 10$ cycles, $\bar{x}_2 = 20.3$ min, $s_2 = 2.6$. Cycle times are approximately normal.

1. **[4 pt]** Test $H_0 \colon \sigma_1^2 = \sigma_2^2$ at $\alpha = 0.05$, given $f_{0.025,\,9,\,11} = 3.59$, and state which mean-comparison procedure your result licenses.
2. **[6 pt]** Carry out the two-sided test of $H_0 \colon \mu_1 = \mu_2$ at $\alpha = 0.05$ and bound the P-value.
3. **[3 pt]** Give the 95% CI for $\mu_1 - \mu_2$.
4. **[2 pt]** State the practical conclusion in one sentence.

Exercise 5.11. A red team attacks two releases of a language model with independent attack sets. Release 1: $x_1 = 64$ successful jailbreaks in $n_1 = 400$ attempts. Release 2 (hardened): $x_2 = 35$ in $n_2 = 350$.

1. **[6 pt]** Test at $\alpha = 0.05$ whether the hardened release has a *lower* breach rate, and give the P-value.
2. **[5 pt]** Construct a 95% two-sided CI for $p_1 - p_2$.
3. **[3 pt]** The team lead says: ''the one-sided test rejected, so the CI must exclude zero.'' Comment carefully.

Exercise 5.12. Five pilot customers' weekly active usage (hours) is recorded before and after a new feature ships: before 6.2, 8.1, 5.4, 7.3, 6.8; after 7.0, 8.9, 5.9, 8.2, 7.4.

1. **[3 pt]** Explain why a two-sample t test is wrong here and set up the correct hypotheses for ''the feature increases usage.''
2. **[6 pt]** Carry out the test at $\alpha = 0.025$.
3. **[4 pt]** Give a 95% two-sided CI for the mean usage gain and interpret it for the founders.

Assessing outside recommendations

Exercise 5.13. A fleet analytics vendor evaluates an aerodynamic kit fitted to 10 trucks, with fuel economy (km/L) measured on the same trucks before and after installation. Their report: ''Group statistics: before $\bar{x}_1 = 8.90$, $s_1 = 1.20$; after $\bar{x}_2 = 9.32$, $s_2 = 1.25$ ($n = 10$ each). Pooled two-sample t: $s_p = 1.225$, $t_0 = 0.42/\left(1.225\sqrt{1/10 + 1/10}\right) = 0.77$ with 18 df; not significant. The kit has no measurable effect.'' The per-truck differences have $\bar{d} = 0.42$ and $s_D = 0.35$. Find the error, redo the analysis correctly at $\alpha = 0.05$, and state what changed.

Exercise 5.14. A vendor pitching a ''more consistent'' inference server benchmarks response-time variability against your current stack, with $n_1 = n_2 = 10$ runs each: current stack $s_1^2 = 5.1$ ms^2, vendor stack $s_2^2 = 1.6$ ms^2, both approximately normal. Their slide: ''two-sided F test of $H_1 : \sigma_1^2 \neq \sigma_2^2$: $F_0 = 5.1/1.6 = 3.19 > f_{0.05,9,9} = 3.18$, significant at $\alpha = 0.05$.'' Given $f_{0.025,9,9} = 4.03$, assess the claim and state what the vendor could legitimately have claimed.

In-Class Activity 5.1: Two Gates, One Bottleneck

Week 6 — Lecture 11

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

A container terminal runs an automated gate (A) and a manual gate (B), and your team must advise whether the automated gate is faster on average. Historical gate-system data give known standard deviations. Gate A: $n_1 = 36$ trucks, $\bar{x}_1 = 6.2$ min, $\sigma_1 = 1.8$. Gate B: $n_2 = 45$ trucks, $\bar{x}_2 = 7.1$ min, $\sigma_2 = 2.4$. Useful value: $\Phi(1.93) = 0.9732$.

Part 1 — Pick the procedure

Which test from this chapter applies, and why (two reasons: what is known, and how the samples relate)? Write the hypotheses for a *two-sided* comparison.

Part 2 — Run it

Compute the standard error and the test statistic, and state the decision at $\alpha = 0.05$ with the P-value.

one-sided argument

s: ''We built the automated gate to be faster; we should have tested $H_1: \mu_A < \mu_B$ from the at test reject at $\alpha = 0.05$? Write one sentence on why the sidedness must be chosen *before* seeing

le standard error, why do the two variance terms add even though the means are subtracted?

In-Class Activity 5.2: Paired or Not?

Week 6 — Lecture 12

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

Your AI validation team must pick the right procedure before any computation.

Part 1 — Classify three studies

For each study, name the correct procedure (paired t / pooled t / Welch t / two-proportion z) and its degrees of freedom or reference distribution: (i) the same 20 prompts scored under model v1 and v2; (ii) crash rates of 500 simulation runs on v1 vs 500 independent runs on v2; (iii) latency of v1 on 15 scenes ($s_1 = 4.1$ ms) vs v2 on 12 different scenes ($s_2 = 9.7$ ms).

Part 2 — A quick paired test

Five scenario suites are scored under both model versions; the per-suite gains (v2 − v1, points) are 2, 3, 1, 4, 2. Test $H_0\colon\ \mu_D = 0$ vs $H_1\colon\ \mu_D \neq 0$ at $\alpha = 0.05$ ($t_{0.025,\,4} = 2.776$). Show $\bar{d}$, s_D, and t_0.

nd the choice

of Part 1, a colleague insists on pooling "because it gives more degrees of freedom." In two in what pooling assumes and what the s_2/s_1 ratio says about that assumption here.

stion do you ask *first* when handed any two-column data set?

In-Class Activity 5.3: The Test-Picker Tournament

Week 7 — Lecture 13

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

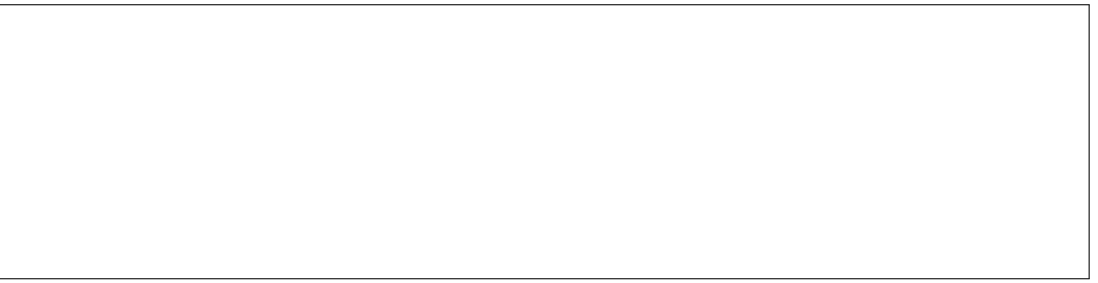
Mixed review for Major Exam 1. Race your team through Table 5.6: for each scenario, write only the parameter, the statistic, the reference distribution, and the degrees of freedom. No arithmetic.

Part 1 — Six scenarios, four lines each

(i) Is the mean picking time under 90 s? ($n = 14$, s from data.) (ii) Is a batch's fill-volume variance within spec σ_0^2? ($n = 20$, normal.) (iii) Do two independent ad campaigns differ in click-through rate? ($n_1 = 900$, $n_2 = 1100$.) (iv) Does a coating change mean corrosion on the same 9 coupons measured before and after? (v) Do two suppliers differ in tensile-strength variability? ($n_1 = 12$, $n_2 = 15$, normal.) (vi) Do two machine shifts differ in mean output, σ known for both?

Part 2 — Spot the errors

A classmate's solution reads: ''Paired study, $n = 8$ pairs, one-sided H_1: $\mu_D > 0$: $t_0 = 2.10$ vs $t_{0.025,8} = 2.306$, fail to reject.'' Find *two* distinct errors and state the correct comparison ($t_{0.05,7} = 1.895$). Does the conclusion flip?



e master table is your team slowest to identify? Write it down—that is tonight's revision target.

In-Class Activity 5.4: Mini Major

Week 7 — Lecture 14

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

One exam-style problem, under exam conditions: 15 minutes, tables only, reasoning shown. Your startup A/B-tests a new signup page. Old page: $x_1 = 96$ signups from $n_1 = 800$ visitors. New page: $x_2 = 68$ from $n_2 = 800$. Useful values: $\Phi(2.31) = 0.9896$, $z_{0.025} = 1.96$.

Part (a) — Set up

Define p_1, p_2, state the hypotheses for ''the pages differ,'' and verify the large-sample conditions.

Part (b) — Test

Compute the pooled proportion, the test statistic (5.10), the decision at $\alpha = 0.05$, and the P-value.

erval and recommendation

% CI for $p_1 - p_2$ using (5.11) and write one decision sentence for the founders in the form ''We ecause Y.''

pooled $\hat{p}$ but the CI did not. One sentence: why?