8 *Single-Factor Experiments: Analysis of Variance*

Every method so far compared at most two populations. Real engineering studies rarely stop at two. A red team probes an AI model with three attack strategies, a logistics group trials four routing policies, a startup tests three onboarding flows. This chapter develops the tool for comparing several means at once: the analysis of variance (ANOVA). The chapter is organized in the following order: why repeated pairwise t tests fail when there are more than two groups; the vocabulary of designed experiments and the completely randomized design; the effects model; the partition of total variability into a between-treatments part and a within-treatments part; the F test and the ANOVA table; confidence intervals for treatment means and their differences; the multiple-comparison procedures of Fisher and Tukey, which identify *which* means differ after the F test says *some* do; residual checks of the model assumptions; sample-size planning; the random-effects model, where the treatment levels are themselves a random sample; and the randomized complete block design, which removes a known nuisance source of variability before the treatments are compared.

The computations in this chapter are longer than in earlier chapters, but they are assembled from pieces you already own: sample means, sums of squared deviations, and one table lookup. If you can build the ANOVA table by hand once, every later design in this course (and the factorial designs of Chapter 9) is the same construction with more rows.

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

Q1. An AI safety team compares four guardrail configurations by running all $\binom{4}{2} = 6$ pairwise t tests, each at $\alpha = 0.05$. Is the chance of at least one false alarm

across the whole comparison still 5%?

Q2. A one-way ANOVA on delivery times for three routing policies reports $F_0 = MS_{\text{Treatments}} / MS_E = 9$. In one sentence, what two quantities does this ratio compare?

Q3. A startup wants to compare three onboarding flows, but weekly signup cohorts differ a lot: some weeks bring motivated users, some do not. What design principle handles the week-to-week differences, and how?

Answers.

Q1. *No.* Each test risks a type I error separately, and the risks accumulate. If the six tests were independent, the chance of at least one false rejection would be $1 - (0.95)^6 \approx 0.26$, about five times the intended level. This inflation is the reason ANOVA exists: one F test examines all four means simultaneously at a genuine 5% level.

Q2. It compares the variability *between* the policy averages with the variability *within* each policy. A ratio of 9 says the policy averages spread out nine times more than chance variation alone would explain, which is evidence that the policies genuinely differ.

Q3. *Blocking.* Treat each week as a block and expose every flow to every week's cohort. Week-to-week differences then become a recognized source of variability that the analysis subtracts out, instead of noise that hides the flow effect.

Lecture 22 starts here — *ANOVA: comparing several means*

Lecture 22. This section is covered by Lecture 22.

Reference. Montgomery & Runger, 6th ed., Sections 13-1 and 13-2.

8.1 Why Not Repeated Pairwise Tests?

Suppose an engineer must compare the means of $a = 3$ groups. Chapter 5 provided the two-sample t test, so a natural idea is to run it three times: group 1 against group 2, group 1 against group 3, and group 2 against group 3. Each test uses $\alpha = 0.05$. The idea looks harmless. It is not, and the reason is the first important computation of this chapter.

Family-wise error rate. The probability of making at least one type I error anywhere in a whole family of tests, not just in one test.

Each individual test has probability 0.05 of a false rejection. But the engineer will act on the whole *family* of conclusions, so the relevant error rate is the probability of at least one false rejection anywhere among the m tests. If the tests were

independent and every null hypothesis were true, each test survives (correctly fails to reject) with probability $1-\alpha$, so all m survive with probability $(1-\alpha)^m$, and

$$P(\text{at least one false rejection}) = 1-(1-\alpha)^m. \tag{8.1}$$

For $a=3$ groups there are $m=\binom{3}{2}=3$ pairwise tests, and

$$1-(1-0.05)^3 = 1-0.8574 = 0.1426.$$

The engineer believes the analysis runs at a 5% error rate, but the actual rate is above 14%. With $a=6$ groups there are $m=\binom{6}{2}=15$ pairs and the family-wise rate climbs to $1-(0.95)^{15}=0.5367$: a coin flip's chance of declaring at least one difference that does not exist. Figure 8.1 shows how fast the intended level is lost. In practice the pairwise tests share data and are not independent, so (8.1) is an approximation, but the conclusion stands: the more pairs you test, the more false discoveries you make.

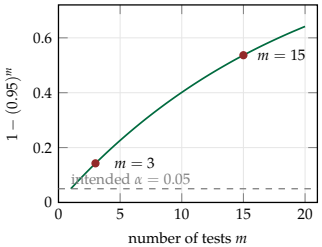

Figure 8.1. The family-wise type I error rate of (8.1) at $\alpha = 0.05$ per test. The intended 5% level is abandoned almost immediately.

Safety lens. Model-evaluation leaderboards invite exactly this trap. Comparing every new checkpoint against every baseline with separate tests at $\alpha = 0.05$ guarantees spurious "wins" as the number of comparisons grows. A single simultaneous test, followed by a controlled multiple-comparison procedure, is the defensible workflow.

Example 8.1 — How bad does it get?

An AI red team evaluates $a=6$ attack strategies against a language model and plans to compare every pair of strategies with a two-sample t test at $\alpha=0.05$. Assuming independence and that no strategy truly differs from another, find the probability that the team reports at least one spurious difference. Then find how small the per-test level α^* would need to be to keep the family-wise rate at 0.05.

Solution.

Step 1 — Count the tests.

There are $m=\binom{6}{2}=15$ pairwise comparisons.

Step 2 — Apply (8.1).

$$1-(1-0.05)^{15} = 1-0.9500^{15} = 1-0.4633 = 0.5367.$$

The team has a $\boxed{53.7\%}$ chance of at least one false discovery.

Step 3 — Reverse the computation.

Require $1-(1-\alpha^*)^{15}=0.05$, so $(1-\alpha^*)^{15}=0.95$ and $1-\alpha^*=0.95^{1/15}=0.99659$, giving $\alpha^* \approx \boxed{0.0034}$. Each individual test would have to run at about a third of one percent, which destroys its power to detect real differences.

Shrinking α per test is not a good fix; testing all means at once is. That simultaneous test is the subject of this chapter.

Analysis of variance (ANOVA). A procedure that tests the equality of several population means by partitioning the total variability in the data into between-group and within-group components.

The solution is a single test of the single hypothesis "all a means are equal," constructed so that its type I error rate is exactly α no matter how many groups are involved. The test works by comparing two estimates of variability, which is why it is called the *analysis of variance* even though it is a test about *means*.

8.2 Designed Experiments and the Completely Randomized Design

Reference. Montgomery & Runger, 6th ed., Section 13-1.

ANOVA is the analysis half of a subject called design of experiments. Before the formulas, four words must be exact, because every symbol in this chapter is indexed by them.

Factor. The controllable variable whose effect on the response is under study, such as attack strategy, routing policy, or onboarding flow.

Level (treatment). One specific setting of the factor. With a levels there are a treatments to compare.

Replicate. One independent observation of the response at a given treatment. With n replicates per treatment there are $N = an$ observations in total.

- The *factor* is the variable the engineer deliberately changes. In this chapter there is a single factor; Chapter 9 studies two or more.
- The *levels* of the factor, also called *treatments*, are the specific settings compared: three attack strategies, four routing policies, three onboarding flows. We write a for the number of levels.
- A *replicate* is one independent run of the experiment at one level. We write n for the number of replicates per level, so the total sample size is $N = an$.
- The *response* is the measured outcome: defects discovered, delivery time, activation rate.

Completely randomized design. A design in which treatments are assigned to experimental units, and run in an order, that is completely random.

Three principles make an experiment trustworthy. *Replication* means running each treatment more than once; without it there is no estimate of the experimental error variance σ^2, and no test is possible. *Randomization* means assigning treatments to experimental units, and choosing the run order, at random; it protects the comparison from unknown influences that drift over time. *Blocking*, the third principle, is deferred to Section 8.11: it removes the influence of a known nuisance variable. An experiment that uses replication and randomization but no blocking is called a *completely randomized design* (CRD), and it is the design analyzed in Sections 8.3 through 8.9.

Engineering judgment. Randomization is physical, not rhetorical. If a fleet trial runs policy A every morning and policy B every afternoon, traffic conditions ride along with the policy and no analysis can separate them afterward. Decide the run order with a random-number generator before the first run, and keep the record.

Throughout the chapter, y_{ij} denotes the j-th observation under treatment i. Totals and averages are written with the dot notation of Montgomery and Runger:

a dot in place of a subscript means that subscript has been summed over. Table 8.1 collects the notation; it is worth a minute of study now, because the formulas that follow are compact precisely because this notation does the bookkeeping.

Symbol	Meaning	Formula
y_{ij}	j-th observation under treatment i	—
$y_{i\cdot}$	total of treatment i	$\sum_{j=1}^{n} y_{ij}$
$\bar{y}_{i\cdot}$	sample mean of treatment i	$y_{i\cdot}/n$
$y_{\cdot\cdot}$	grand total	$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$
$\bar{y}_{\cdot\cdot}$	grand mean	$y_{\cdot\cdot}/N$
a	number of treatments (levels)	—
n	replicates per treatment	—
N	total observations	$N = an$

Table 8.1. Dot notation for the single-factor experiment. A dot replaces a subscript that has been summed over; a bar denotes an average.

Reference. Montgomery & Runger, 6th ed., Section 13-2.1.

8.3 *The Effects Model*

Treatment effect. The amount τ_i by which treatment i shifts the response above or below the overall mean μ.

Each observation is modeled as an overall level, plus a shift specific to its treatment, plus random error:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, \dots, a, \quad j = 1, \dots, n, \tag{8.2}$$

where μ is the overall (grand) mean, τ_i is the *effect* of treatment i, and the errors ε_{ij} are independent $N(0, \sigma^2)$ random variables with the same variance in every group. Equation (8.2) is called the *effects model*. The mean of treatment i is $\mu_i = \mu + \tau_i$, and the effects are defined as deviations from the overall mean, so they satisfy the constraint $\sum_{i=1}^{a} \tau_i = 0$.

The hypothesis ''all treatment means are equal'' becomes a statement about the effects:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \qquad \text{vs.} \qquad H_1 : \tau_i \neq 0 \text{ for at least one } i,$$

which is the same as $H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ against the alternative that at least one pair of means differs. Please pay attention to what H_1 does and does not say:

rejecting H_0 establishes that *some* means differ, not *which* ones, and not that *all* of them differ. Identifying the specific pairs is the job of the multiple-comparison procedures in Section 8.7.

The natural estimates replace population quantities by their sample counterparts:

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \qquad \hat{\tau}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \qquad i = 1, \dots, a. \tag{8.3}$$

Three assumptions carry the whole analysis, and they are the same three as in regression: the errors are *normally distributed*, they have *equal variance* σ^2 in every treatment, and they are *independent*. Normality and equal variance are checked with residual plots (Section 8.8); independence is earned, not checked, by randomizing the experiment.

Startup lens. In an A/B/n test the "treatments" are product variants and the "experimental units" are users. Random assignment of users to variants is what makes the independence assumption believable; letting users self-select their variant breaks it before any data arrive.

Reference. Montgomery & Runger, 6th ed., Section 13-2.2.

8.4 Partitioning Variability

The central idea of ANOVA is an identity about deviations. Take any observation and subtract the grand mean. That total deviation splits exactly into two pieces:

$$\underbrace{y_{ij} - \bar{y}_{\cdot\cdot}}_{\text{total deviation}} = \underbrace{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}_{\text{treatment deviation}} + \underbrace{(y_{ij} - \bar{y}_{i\cdot})}_{\text{error deviation}}. \tag{8.4}$$

The first piece asks: how far is this observation's *group average* from the overall average? That is the part of the deviation explained by the treatment. The second piece asks: how far is the observation from its *own group's* average? That is pure experimental noise, because observations inside one group all received the same treatment.

Square both sides of (8.4) and sum over all N observations. The cross-product term sums to zero (within each group, the deviations $y_{ij} - \bar{y}_{i\cdot}$ sum to zero while the group's treatment deviation is a constant), and what remains is the fundamental decomposition

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2}_{SS_T} = \underbrace{n\sum_{i=1}^{a}\left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2}_{SS_{\text{Treatments}}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2}_{SS_E}, \tag{8.5}$$

read as $SS_T = SS_{\text{Treatments}} + SS_E$: the total sum of squares equals the between-treatments sum of squares plus the within-treatments (error) sum of squares. Figure 8.2 shows the two kinds of variation on real data.

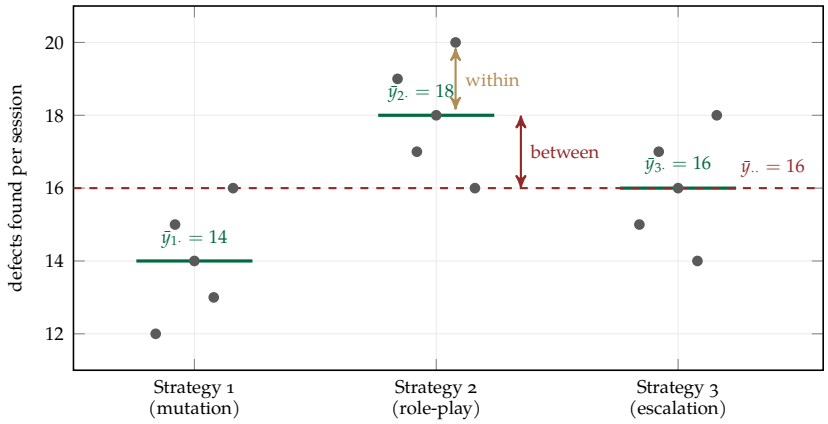

Figure 8.2. Between- versus within-treatment variation for the red-team data of Example 8.2. Solid green segments are the treatment means; the dashed line is the grand mean. $SS_{\text{Treatments}}$ measures the green segments' spread around the dashed line; SS_E measures the points' spread around their own green segment.

For hand computation, expanding the squares gives shortcut forms that use only totals. They are the forms on the course formula sheet, and they avoid computing N individual deviations:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N}, \tag{8.6}$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{a} y_{i\cdot}^2 - \frac{y_{\cdot\cdot}^2}{N}, \qquad SS_E = SS_T - SS_{\text{Treatments}}. \tag{8.7}$$

If the group sizes are unequal (n_i observations in group i, $N = \sum n_i$), replace n by n_i inside the sum in (8.7); everything else in the chapter goes through with $\sum_i (n_i - 1) = N - a$ error degrees of freedom.

A sum of squares only becomes comparable to another after dividing by its degrees of freedom. $SS_{\text{Treatments}}$ is built from a deviations of group means around the grand mean, which sum to zero, so it has $a - 1$ degrees of freedom. SS_E is built from $n - 1$ free deviations inside each of the a groups, so it has $a(n-1) = N - a$ degrees of freedom. Dividing gives the two *mean squares*:

$$MS_{\text{Treatments}} = \frac{SS_{\text{Treatments}}}{a - 1}, \tag{8.8}$$

$$MS_E = \frac{SS_E}{N - a}. \tag{8.9}$$

Mean square. A sum of squares divided by its degrees of freedom; an average squared deviation that estimates a variance.

MS_E deserves a second look. Inside group i, the quantity $\sum_j (y_{ij} - \bar{y}_{i\cdot})^2/(n-1)$ is exactly the sample variance s_i^2 of that group. MS_E is the pooled average of the a group variances, the direct generalization of the pooled variance s_p^2 from the two-sample t test. It estimates σ^2 whether or not H_0 is true.

Insight. Why a test about means works by comparing variances. Take expectations of the two mean squares under model (8.2):

$$E[MS_E] = \sigma^2, \qquad E[MS_{\text{Treatments}}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}.$$

MS_E estimates σ^2 always. $MS_{\text{Treatments}}$ estimates σ^2 *only* when every $\tau_i = 0$; any real treatment effect inflates it by a nonnegative amount. So under H_0 the ratio $MS_{\text{Treatments}}/MS_E$ hovers around 1, and real effects push it above 1. Large ratios are evidence against H_0 — and "how large is large" is answered by the F distribution.

The two expectations in the insight box are worth recording as a numbered result, because the random-effects model of Section 8.10 changes exactly one of them:

$$E[MS_E] = \sigma^2, \qquad E[MS_{\text{Treatments}}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}. \tag{8.10}$$

Reference. Montgomery & Runger, 6th ed., Section 13-2.2.

8.5 The F Test and the ANOVA Table

Under H_0, both mean squares estimate the same σ^2, and their ratio follows an F distribution — the same distribution used to compare two variances in Chapter 5. The test statistic is

$$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}, \qquad F_0 \sim F_{a-1,\,N-a} \text{ under } H_0, \tag{8.11}$$

and H_0 is rejected at level α when $F_0 > f_{\alpha,\,a-1,\,N-a}$. The test is one-tailed to the right even though the alternative is "means differ in any direction," because treatment effects of *any* sign inflate $MS_{\text{Treatments}}$ by (8.10); only the upper tail signals real effects.

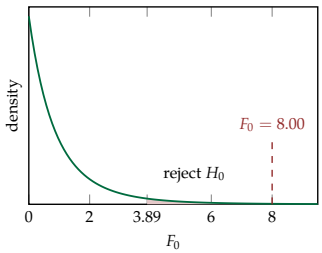

Figure 8.3. The $F_{2,12}$ density for Example 8.2. The shaded region above $f_{0.05,2,12} = 3.89$ has area 0.05; the observed $F_0 = 8.00$ falls deep inside it.

The whole computation is organized in a standard display called the ANOVA table, shown in Table 8.2. Software prints this table; in this course you must also be able to build it by hand, and Procedure 8.1 is the recipe.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}} = \dfrac{SS_{\text{Treatments}}}{a-1}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$N-a$	$MS_E = \dfrac{SS_E}{N-a}$	
Total	SS_T	$N-1$		

Table 8.2. The one-way ANOVA table. Both the SS column and the df column add down to the Total row — always verify both additions before reading off F_0.

Procedure 8.1 — Building the one-way ANOVA table.

1. Organize the data by treatment. Compute the treatment totals $y_{i\cdot}$, treatment means $\bar{y}_{i\cdot}$, grand total $y_{\cdot\cdot}$, and grand mean $\bar{y}_{\cdot\cdot}$. Record a, n, and $N = an$.
2. Compute SS_T from (8.6) (or directly as the sum of squared deviations from $\bar{y}_{\cdot\cdot}$).
3. Compute $SS_{\text{Treatments}}$ from (8.7) (or as $n\sum_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$), then $SS_E = SS_T - SS_{\text{Treatments}}$.
4. Fill in the degrees of freedom $a-1$, $N-a$, $N-1$ and check that the first two add to the third.
5. Compute the mean squares from (8.8) and (8.9), then F_0 from (8.11).
6. Look up $f_{\alpha,a-1,N-a}$. Reject H_0 if $F_0 > f_{\alpha,a-1,N-a}$ (equivalently, if the p-value is below α), and state the conclusion in the language of the problem.

Example 8.2 — Three red-team strategies

An AI safety group compares three red-team strategies for probing a deployed language model: prompt *mutation*, *role-play* jailbreaks, and multi-turn *escalation*. Each strategy is run in $n = 5$ independent four-hour sessions, in randomized order by the same analysts, and the response is the number of distinct safety defects discovered per session. The data are in Table 8.3. At

$\alpha = 0.05$, do the strategies differ in mean defect discovery?

Strategy	Defects per session					$y_{i\cdot}$	$\bar{y}_{i\cdot}$
1. Mutation	12	15	14	13	16	70	14.0
2. Role-play	19	17	18	20	16	90	18.0
3. Escalation	15	17	16	14	18	80	16.0
					Grand total $y_{\cdot\cdot}$	240	
					Grand mean $\bar{y}_{\cdot\cdot}$		16.0

Table 8.3. Red-team defect-discovery data: $a = 3$ strategies, $n = 5$ sessions each, $N = 15$.

Solution.

Step 1 — Hypotheses and setup.

H_0: $\mu_1 = \mu_2 = \mu_3$ (equivalently $\tau_1 = \tau_2 = \tau_3 = 0$) against H_1: at least one pair differs. Here $a = 3$, $n = 5$, $N = 15$, and the totals are in Table 8.3.

Step 2 — Total sum of squares.

The sum of all squared observations is

$$\sum_{i}\sum_{j} y_{ij}^2 = (12^2 + 15^2 + \cdots + 18^2) = 990 + 1630 + 1290 = 3910,$$

and $y_{\cdot\cdot}^2/N = 240^2/15 = 57600/15 = 3840$. By (8.6), $SS_T = 3910 - 3840 = 70.00$.

Step 3 — Treatment and error sums of squares.

By (8.7),

$$SS_{\text{Treatments}} = \frac{70^2 + 90^2 + 80^2}{5} - 3840 = \frac{19400}{5} - 3840 = 3880 - 3840 = 40.00.$$

The definition form agrees: $5[(14-16)^2 + (18-16)^2 + (16-16)^2] = 5(4 + 4 + 0) = 40$. Then $SS_E = 70 - 40 = 30.00$.

Step 4 — Mean squares and the statistic.

Degrees of freedom: $a - 1 = 2$, $N - a = 12$, $N - 1 = 14$ (and $2 + 12 = 14$, as required). By (8.8) and (8.9),

$$MS_{\text{Treatments}} = \frac{40}{2} = 20.00, \qquad MS_E = \frac{30}{12} = 2.500,$$

so by (8.11), $F_0 = 20.00/2.500 = 8.00$.

$\nu_2 \backslash \nu_1$	1	2	3
11	4.84	3.98	3.59
12	4.75	3.89	3.49
13	4.67	3.81	3.41

Table 8.4. Excerpt of the F table, upper 5% points $f_{0.05,\nu_1,\nu_2}$: $f_{0.05,2,12} = 3.89$.

Step 5 — Decision from the F table.

From Table 8.4, $f_{0.05,2,12} = 3.89$. Since $F_0 = 8.00 > 3.89$, reject H_0 (the p-value is about 0.006). The completed table is Table 8.5.

Source	SS	df	MS	F_0
Strategies	40.00	2	20.00	8.00
Error	30.00	12	2.500	
Total	70.00	14		

Table 8.5. Completed ANOVA table for the red-team study. Compare with Figure 8.3.

Step 6 — Conclusion in context.

The answer is $\boxed{F_0 = 8.00 > 3.89 \;\Rightarrow\; \text{reject } H_0}$: mean defect discovery is not the same across the three strategies. Role-play sessions found the most defects on average (18.0 per session), but the F test alone does not certify which specific pairs differ — that requires Section 8.7.

Safety lens. The response here — distinct defects per fixed-length session — is a discovery *rate*. Treating the sessions as independent replicates is justified only if sessions do not share seed prompts and analysts do not carry findings across sessions; otherwise the effective sample size is smaller than 15 and MS_E understates σ^2.

Reference. Montgomery & Runger, 6th ed., Section 13-2.2.

8.6 Estimating Treatment Effects and Confidence Intervals

Rejecting H_0 is rarely the end of the analysis; the engineer wants numbers. The point estimates come from (8.3), and the key fact for interval estimates is that MS_E, with its $N - a$ degrees of freedom, plays exactly the role s^2 played in Chapter 3. A $100(1-\alpha)\%$ confidence interval for a single treatment mean $\mu_i = \mu + \tau_i$ is

$$\bar{y}_{i\cdot} - t_{\alpha/2,\,N-a}\sqrt{\frac{MS_E}{n}} \;\le\; \mu_i \;\le\; \bar{y}_{i\cdot} + t_{\alpha/2,\,N-a}\sqrt{\frac{MS_E}{n}}, \tag{8.12}$$

and for the difference between two treatment means $\mu_i - \mu_k$,

$$(\bar{y}_{i\cdot} - \bar{y}_{k\cdot}) \pm t_{\alpha/2,\,N-a}\sqrt{\frac{2\,MS_E}{n}}. \tag{8.13}$$

Both intervals borrow strength from the whole experiment: even the interval for one group's mean uses the pooled MS_E from all a groups, which is why its degrees of freedom are $N - a$ rather than $n - 1$.

Example 8.3 — Effects and intervals for the red-team study

For the data of Example 8.2: estimate the model parameters, give a 95% confidence interval for the role-play mean μ_2, and a 95% confidence interval for the role-play advantage over mutation, $\mu_2 - \mu_1$.

Solution.

Step 1 — Point estimates.

By (8.3), $\hat{\mu} = 16.0$ and

$$\hat{\tau}_1 = 14 - 16 = -2.0, \qquad \hat{\tau}_2 = 18 - 16 = +2.0, \qquad \hat{\tau}_3 = 16 - 16 = 0.0,$$

which sum to zero as the model requires. The error variance estimate is $\hat{\sigma}^2 = MS_E = 2.500$.

Step 2 — Interval for one mean.

With $t_{0.025,12} = 2.179$ and $\sqrt{MS_E/n} = \sqrt{2.5/5} = \sqrt{0.5000} = 0.7071$, equation (8.12) gives

$$18.0 \pm 2.179(0.7071) = 18.0 \pm 1.541,$$

so $\boxed{16.46 \le \mu_2 \le 19.54}$ defects per session.

Step 3 — Interval for a difference.

With $\sqrt{2\,MS_E/n} = \sqrt{2(2.5)/5} = 1.000$, equation (8.13) gives

$$(18 - 14) \pm 2.179(1.000) = 4.0 \pm 2.179,$$

so $\boxed{1.82 \le \mu_2 - \mu_1 \le 6.18}$. The interval excludes zero: role-play discovers between about 2 and 6 more defects per session than mutation, with 95% confidence.

Lecture 23 starts here — *Multiple comparisons: LSD and HSD*

8.7 Multiple Comparisons After the F Test

A significant F test answers only the coarse question "are the means all equal?" The follow-up question — *which* means differ from which — is answered by multiple-comparison procedures. This section presents the two used in this course: Fisher's least significant difference (LSD), which is simple and powerful but does not control the family-wise error rate, and Tukey's honestly significant difference (HSD), which does. Both are applied to the same running example, a startup's onboarding experiment, so their behavior can be compared directly.

Engineering judgment. Statistical significance is the beginning of the engineering answer, not the end. The interval (1.82, 6.18) says the role-play advantage could plausibly be as small as two defects per session; whether that justifies its higher analyst cost is a judgment the interval informs but cannot make.

Lecture 23. This section is covered by Lecture 23.

Reference. Montgomery & Runger, 6th ed., Section 13-2.3.

Multiple comparisons. Procedures that identify which specific pairs of treatment means differ, applied after a significant overall F test.

The running example. A startup tests three onboarding flows for its product: flow A (the current design), flow B (a checklist added to the first screen), and flow C (a guided interactive tour). New signups are randomly assigned to one flow. The response is the activation rate: the percentage of a weekly cohort that completes a key action within seven days. Each flow accumulates $n = 6$ weekly cohorts (Table 8.6).

Flow	Activation rate (%) by cohort						$y_{i\cdot}$	$\bar{y}_{i\cdot}$
A. Current	48	55	51	56	50	52	312	52.0
B. Checklist	53	59	54	58	52	57	333	55.5
C. Tour	57	63	59	62	58	61	360	60.0
						Grand total $y_{\cdot\cdot}$	1005	
						Grand mean $\bar{y}_{\cdot\cdot}$		55.83

Table 8.6. Onboarding activation-rate data: $a = 3$ flows, $n = 6$ cohorts each, $N = 18$.

Following Procedure 8.1: $\sum\sum y_{ij}^2 = 56421$, $y_{\cdot\cdot}^2/N = 1005^2/18 = 56112.5$, so $SS_T = 308.5$; and $SS_{\text{Treatments}} = (312^2 + 333^2 + 360^2)/6 - 56112.5 = 56305.5 - 56112.5 = 193.0$, leaving $SS_E = 115.5$. The ANOVA table is Table 8.7. Since $F_0 = 12.53 > f_{0.05,2,15} = 3.68$, the flows differ ($p \approx 0.0006$), and multiple comparisons are warranted.

Source	SS	df	MS	F_0
Flows	193.0	2	96.50	12.53
Error	115.5	15	7.700	
Total	308.5	17		

Table 8.7. ANOVA table for the onboarding experiment. $F_0 = 12.53$ exceeds $f_{0.05,2,15} = 3.68$, so the pairwise analysis proceeds.

Least significant difference (LSD). The smallest gap between two sample means that a pairwise t test at level α would declare significant.

8.7.1 *Fisher's Least Significant Difference*

Fisher's idea is to run each pairwise comparison as a two-sample t test, but with one improvement over Chapter 5: instead of pooling only the two groups being compared, use MS_E from the full ANOVA as the variance estimate, with all $N - a$ error degrees of freedom. The pair (i, k) is declared different when

$$|t_0| = \frac{\left|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}\right|}{\sqrt{2\,MS_E/n}} > t_{\alpha/2,\,N-a},$$

and multiplying through by the denominator turns the rule into a single yardstick applied to every pair:

$$\mathrm{LSD} = t_{\alpha/2,\,N-a}\sqrt{\frac{2\,MS_E}{n}}, \tag{8.14}$$

declare $\bar{y}_{i\cdot}$ and $\bar{y}_{k\cdot}$ significantly different when $\left|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}\right| > \mathrm{LSD}$. Note that the LSD is exactly the half-width of the confidence interval (8.13): a pair is LSD-significant precisely when the confidence interval for its difference excludes zero. For unequal group sizes, replace $2/n$ by $1/n_i + 1/n_k$. Since $N - a = a(n-1)$ for equal group sizes, tables and formula sheets sometimes write the degrees of freedom as $a(n-1)$; the two are the same number.

Procedure 8.2 — Fisher LSD comparisons.

1. Confirm that the overall F test rejected H_0. If it did not, stop: LSD comparisons without a significant F test generate false discoveries.
2. Read MS_E and its degrees of freedom $N - a$ from the ANOVA table, and

look up $t_{\alpha/2,\,N-a}$.

3. Compute the yardstick LSD from (8.14).
4. List all $\binom{a}{2}$ pairs, compute each $|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|$, and flag the pairs that exceed the LSD.
5. Report the pattern of differences in the language of the problem, noting that the family-wise error rate is *not* controlled.

$\nu \backslash \alpha$	0.05	0.025	0.01
14	1.761	2.145	2.624
15	1.753	2.131	2.602
16	1.746	2.120	2.583

Table 8.8. Excerpt of the t table, upper-tail points $t_{\alpha,\nu}$: $t_{0.025,15} = 2.131$.

Example 8.4 — LSD for the onboarding flows

Apply Fisher's LSD at $\alpha = 0.05$ to the onboarding experiment of Table 8.7.

Solution.

Step 1 — Prerequisite.

The F test rejected H_0 ($F_0 = 12.53 > 3.68$), so LSD comparisons are justified (Procedure 8.2).

Step 2 — The yardstick.

$MS_E = 7.700$ with $N - a = 15$ degrees of freedom, and $t_{0.025,15} = 2.131$ from Table 8.8. By (8.14),

$$\text{LSD} = 2.131\sqrt{\frac{2(7.700)}{6}} = 2.131\sqrt{2.567} = 2.131(1.602) = 3.414.$$

Step 3 — Compare every pair.

$$|\bar{y}_A - \bar{y}_B| = |52.0 - 55.5| = 3.5 > 3.414,$$

$$|\bar{y}_A - \bar{y}_C| = |52.0 - 60.0| = 8.0 > 3.414, \qquad |\bar{y}_B - \bar{y}_C| = |55.5 - 60.0| = 4.5 > 3.414.$$

Step 4 — Conclusion.

By the LSD criterion $\boxed{\text{all three flows differ}}$: the tour beats the checklist, which in turn beats the current flow. But notice how close the A–B comparison is (3.5 against a yardstick of 3.414); the next example shows that a procedure with an honest family-wise error rate is not convinced by it.

The weakness of the LSD is the arithmetic of Section 8.1 all over again: each pair is tested at $\alpha = 0.05$, so the chance of at least one false flag among the three comparisons is roughly 14%, and it grows with a. Running LSD only after a significant F test (as Procedure 8.2 insists) limits the damage but does not repair it.

Studentized range. The random variable $q = (\bar{y}_{\max} - \bar{y}_{\min})/\sqrt{MS_E/n}$: the spread of the largest and smallest of a group means, in standard-error units.

8.7.2 *Tukey's Honestly Significant Difference*

Tukey's procedure fixes the family-wise error rate by asking a sharper question. Under H_0, how far apart should the *largest* and *smallest* of a sample means fall just by chance? The distribution of

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

is called the Studentized range distribution, with upper-α point $q_\alpha(a, f)$, where f is the error degrees of freedom. If the yardstick is calibrated to the most extreme pair, then *every* pair is simultaneously tested with family-wise error exactly α. The Tukey yardstick is

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}, \qquad f = N - a, \tag{8.15}$$

declare $\bar{y}_{i\cdot}$ and $\bar{y}_{k\cdot}$ significantly different when $\left|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}\right| > T_\alpha$.

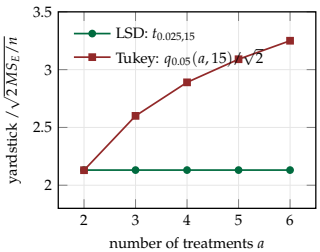

Figure 8.4. The two yardsticks on a common scale ($f = 15$ error df, $\alpha = 0.05$). They coincide at $a = 2$; for $a \geq 3$ Tukey's grows with the number of treatments, paying for family-wise error control.

Two properties are worth memorizing. First, for $a = 2$ the Studentized range satisfies $q_\alpha(2, f) = \sqrt{2}\, t_{\alpha/2,f}$, so Tukey and LSD agree exactly: with one comparison there is nothing to correct. Second, for $a \geq 3$, $q_\alpha(a, f) > \sqrt{2}\, t_{\alpha/2,f}$, so $T_\alpha >$ LSD always (Figure 8.4): Tukey demands a larger observed gap before it will call a pair different. Tukey rejects fewer pairs, and in exchange the probability of *any* false flag among all $\binom{a}{2}$ comparisons is held at α.

Procedure 8.3 — Tukey HSD comparisons.

1. Read MS_E and $f = N - a$ from the ANOVA table.
2. Look up the Studentized range point $q_\alpha(a, f)$ — note that its first argument is the number of treatments a, not the number of pairs.

3. Compute the yardstick T_α from (8.15).
4. Compare every $|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|$ against T_α and flag the pairs that exceed it.
5. Summarize with an underline diagram: sort the means, and underline every set of means whose members are *not* significantly different from one another. Report the groups in the language of the problem.

Example 8.5 — Tukey HSD, and the verdict it changes

Apply Tukey's HSD at $\alpha = 0.05$ to the onboarding experiment, and compare the conclusions with the LSD conclusions of Example 8.4.

$f \backslash a$	2	3	4
14	3.03	3.70	4.11
15	3.01	3.67	4.08
16	3.00	3.65	4.05

Table 8.9. Excerpt of the Studentized range table, upper 5% points $q_{0.05}(a, f)$: $q_{0.05}(3, 15) = 3.67$.

Solution.

Step 1 — The yardstick.

With $a = 3$, $f = 15$, Table 8.9 gives $q_{0.05}(3, 15) = 3.67$. By (8.15),

$$T_{0.05} = 3.67\sqrt{\frac{7.700}{6}} = 3.67\sqrt{1.283} = 3.67(1.133) = 4.157.$$

Step 2 — Compare every pair.

Table 8.10 runs both yardsticks side by side.

Pair	$\lvert\bar{y}_{i\cdot} - \bar{y}_{k\cdot}\rvert$	$>$ LSD $= 3.414$?	$> T_{0.05} = 4.157$?	Verdict
A vs. B	3.5	Yes	**No**	methods disagree
A vs. C	8.0	Yes	Yes	differ
B vs. C	4.5	Yes	Yes	differ

Table 8.10. LSD and Tukey HSD applied to the same onboarding data. The borderline A–B gap survives the per-pair yardstick but not the family-wise one.

Step 3 — Underline diagram.

Sort the means and underline the sets Tukey cannot separate (Figure 8.5): A and B share a line; C stands alone.

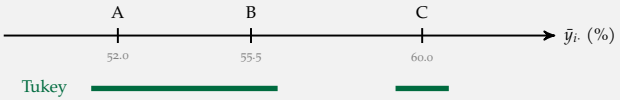

Figure 8.5. Underline diagram for the Tukey analysis. Means joined by a common line are not significantly different at family-wise $\alpha = 0.05$.

Step 4 — Conclusion.

Tukey certifies $\boxed{\text{C} > \text{B and C} > \text{A}}$ but declines to separate A from B. The honest summary for the product team: the guided tour is genuinely better than both alternatives; the evidence that the checklist beats the current flow is suggestive (3.5 points) but not conclusive at family-wise $\alpha = 0.05$. Fisher's LSD, which had claimed all three differ, was spending more type I error than it advertised. When the two procedures disagree, report Tukey; use LSD as an exploratory screen.

Startup lens. "Which variant won?" is a family-wise question, because the team will ship the pattern of winners, not one isolated comparison. Dashboards that flag every pairwise gap at $\alpha = 0.05$ will regularly flag ghosts; a Tukey correction is one honest line of code away.

Reference. Montgomery & Runger, 6th ed., Section 13-2.4.

Residual. The difference e_{ij} between an observation and its fitted value; in one-way ANOVA, the deviation from the treatment mean.

8.8 Checking the Model: Residual Analysis

The F test, the intervals, and both multiple-comparison procedures all lean on the model assumptions: normal errors, equal variances, independence. The raw material for checking them is the set of residuals. In the one-way model the fitted value of any observation is its treatment mean, $\hat{y}_{ij} = \bar{y}_{i\cdot}$, so

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}. \tag{8.16}$$

Three plots, the same three as in regression, each interrogate one assumption:

- a *normal probability plot* of the residuals checks normality — healthy residuals fall along a straight line; heavy tails or an S-shape are warnings;
- *residuals against fitted values* $\bar{y}_{i\cdot}$ checks equal variance — healthy plots show even vertical spread across all groups; a funnel that widens with the fitted value means σ^2 grows with the mean;
- *residuals against time or run order* checks independence — any drift or slow oscillation suggests the randomization failed to break a time effect.

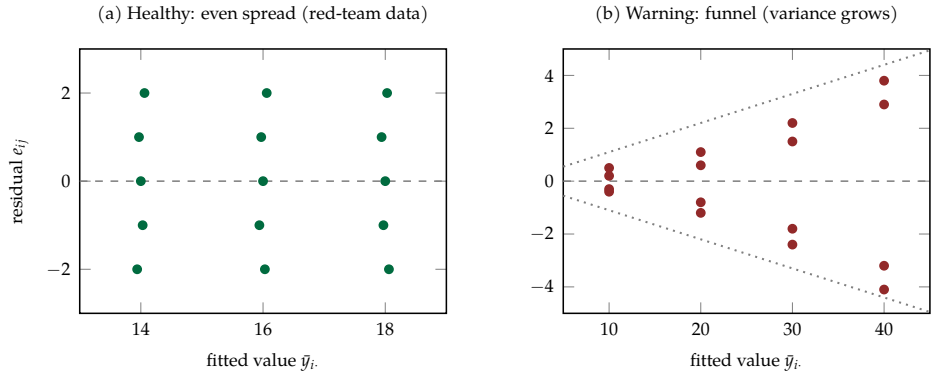

Figure 8.6. Residuals against fitted values. Panel (a): the red-team residuals from (8.16) show even spread around zero in all three groups — the equal-variance assumption is credible. Panel (b): a funnel pattern from a different experiment — the spread grows with the mean, and a transformation of the response is needed before ANOVA is trusted.

For the red-team data, the residuals in group 1 are $(-2, 1, 0, -1, 2)$, in group 2 $(1, -1, 0, 2, -2)$, and in group 3 $(-1, 1, 0, -2, 2)$; each set sums to zero by construction. Panel (a) of Figure 8.6 plots them against the fitted values: the spread is the same in all three groups and there are no outliers, so the analysis stands.

When a check fails, the standard remedies are: a variance-stabilizing transformation such as $\sqrt{y}$ or $\log y$ when the spread grows with the mean (common for counts and for times); a nonparametric alternative (the Kruskal–Wallis test) when non-normality is severe; and better randomization when residuals drift with run order. ANOVA is robust to mild non-normality, especially with equal group sizes; unequal variances, particularly with unequal group sizes, are the more serious violation. Please make the residual check a habit, because it costs minutes and it is the only part of the analysis that can tell you the rest of the analysis is built on sand.

8.9 *Choosing the Sample Size*

Reference. Montgomery & Runger, 6th ed., Section 13-2.5.

Operating characteristic curve. A plot of the probability of failing to reject H_0 against a non-centrality parameter, used to read off the power of a planned test.

How many replicates n per treatment does an experiment need? As in Chapter 4, the answer depends on the smallest effect worth detecting and on the desired power $1 - \beta$. For the ANOVA F test, the difficulty of the detection problem is summarized by the non-centrality parameter

$$\Phi^2 = \frac{n \sum_{i=1}^{a} \tau_i^2}{a\,\sigma^2}, \tag{8.17}$$

which is large when effects are big, replication is generous, and noise is small. The exact τ_i are unknown at planning time, so the standard shortcut supposes only that *some* two treatment means are at least D apart. The worst placement of the effects then gives the bound

$$\Phi^2 \;\geq\; \frac{nD^2}{2a\,\sigma^2},$$

and power is read from operating characteristic (OC) curves (Montgomery & Runger, Appendix Chart VII) as a function of Φ, $a-1$, and the error degrees of freedom. The planning inputs are therefore: a, α, the target power, the smallest meaningful difference D, and a pre-experiment estimate of σ (from a pilot study or historical data). Because D enters squared, halving the difference you must detect quadruples the required Φ^2, and roughly quadruples n; the same is true of doubling σ.

Example 8.6 — Planning a guardrail comparison

An AI safety team will compare $a = 4$ guardrail configurations by their false-positive rates on benign traffic, measured in percentage points per test batch. Pilot data suggest $\sigma \approx 2$ percentage points. The team wants to detect any gap of $D = 3$ points between two configurations with power about 0.90 at $\alpha = 0.05$. How many batches per configuration are needed?

Solution.

Step 1 — The non-centrality bound.

By the worst-case form of (8.17),

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} = \frac{n(3)^2}{2(4)(2)^2} = \frac{9n}{32} = 0.2813\,n.$$

Step 2 — Try candidate values of n.

For each n, compute Φ, the error degrees of freedom $a(n-1)$, and read the power from the OC chart (Table 8.11).

n	$\Phi^2 = 9n/32$	Φ	$\mathrm{df}_E = a(n-1)$	Power (OC chart)
4	1.13	1.06	12	≈ 0.55
5	1.41	1.19	16	≈ 0.70
6	1.69	1.30	20	≈ 0.82
7	1.97	1.40	24	≈ 0.90

Table 8.11. Sample-size search for the guardrail study. Power values are read from the operating characteristic curves for $a - 1 = 3$ numerator degrees of freedom at $\alpha = 0.05$.

Step 3 — Decide.

The first row reaching the target is $n = 7$. The answer is $n = 7$ batches per configuration (28 runs total) for approximately 90% power. Note the shape of the table: going from $n = 4$ to $n = 7$ nearly doubles the power. Replication is the cheapest insurance an experiment can buy, and it must be bought *before* the experiment runs.

Insight. Plan the sample size before collecting data, not after. A non-significant F from an underpowered experiment is not evidence that the treatments are equal — it is evidence that the experiment could not tell. The OC-curve computation, done in advance, is what separates ''we found no difference'' from ''we could not have found one.''

Lecture 24 starts here — *Random effects and blocking*

8.10 The Random-Effects Model

Everything so far treated the factor as *fixed*: the three onboarding flows were the exact flows the startup cares about, and the conclusions apply to those three and no others. Often the levels are instead a random sample from a larger population. A logistics network might select four warehouses at random from its thirty sites; a manufacturer might sample five operators from all qualified operators; an AI evaluation might sample six annotators from a labeling pool. The interesting question is then not ''do these particular four warehouses differ?'' but ''how much does the warehouse factor, across the whole network, contribute to variability?''

Lecture 24. This section is covered by Lecture 24.

Reference. Montgomery & Runger, 6th ed., Section 13-3.

Fixed factor. A factor whose levels are the specific settings of interest; conclusions apply to those levels only.

Random factor. A factor whose levels are a random sample from a larger population of levels; conclusions apply to the whole population.

The model equation looks identical to (8.2),

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

but now the treatment effects are themselves random: $\tau_i \sim N(0, \sigma_\tau^2)$, independent of the errors $\varepsilon_{ij} \sim N(0, \sigma^2)$. Because the two random terms are independent, the variance of any observation splits into two parts,

$$V(Y_{ij}) = \sigma_\tau^2 + \sigma^2, \tag{8.18}$$

Variance components. The pieces σ_τ^2 (between levels) and σ^2 (within levels) into which the total variance splits under the random-effects model.

called the *variance components*: the between-treatment component σ_τ^2 and the within-treatment component σ^2.

The hypotheses change accordingly. Individual τ_i are no longer parameters to compare; the factor matters if and only if its variance component is positive:

$$H_0 : \sigma_\tau^2 = 0 \qquad \text{vs.} \qquad H_1 : \sigma_\tau^2 > 0.$$

The pleasant surprise is that the arithmetic does not change at all. The sums of squares, the degrees of freedom, the mean squares, and the statistic $F_0 = MS_{\text{Treatments}}/MS_E$ are computed exactly as in Procedure 8.1, and H_0 is rejected when $F_0 > f_{\alpha, a-1, N-a}$. What changes is the interpretation, through the expected mean squares. Compare with (8.10): under the random-effects model,

$$E[MS_E] = \sigma^2, \qquad E[MS_{\text{Treatments}}] = \sigma^2 + n\,\sigma_\tau^2. \tag{8.19}$$

Solving the two equations for the two unknowns gives the method-of-moments estimators of the variance components:

$$\hat{\sigma}^2 = MS_E, \qquad \hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n}. \tag{8.20}$$

One caveat: if $MS_{\text{Treatments}} < MS_E$, formula (8.20) produces a negative estimate of a variance, which is impossible. The convention is to set $\hat{\sigma}_\tau^2 = 0$ and conclude that the factor contributes no detectable variability. A useful summary of the components is the proportion of total variance attributable to the factor, $\hat{\sigma}_\tau^2/(\hat{\sigma}_\tau^2 + \hat{\sigma}^2)$, sometimes called the intraclass correlation.

Procedure 8.4 — Random-effects analysis and variance components.

1. Confirm the factor is random: the levels were sampled from a larger population, and the conclusion is meant to apply to that population.
2. Build the ANOVA table exactly as in Procedure 8.1; test $H_0\colon \sigma_\tau^2 = 0$ by comparing F_0 with $f_{\alpha,\,a-1,\,N-a}$.
3. Estimate the components from (8.20); if the between estimate is negative, report $\hat{\sigma}_\tau^2 = 0$.
4. Report the proportion of total variability due to the factor, and state the conclusion about the *population* of levels, not about the particular levels sampled.

Example 8.7 — Warehouse-to-warehouse variability

A logistics network wants to know how much of the variability in order picking time is caused by differences between warehouses. Four warehouses are selected *at random* from the network's thirty sites, and $n = 5$ randomly chosen orders are timed at each (minutes). The ANOVA computation, following Procedure 8.1, yields Table 8.12. Test $H_0\colon \sigma_\tau^2 = 0$ at $\alpha = 0.05$ and estimate the variance components.

Source	SS	df	MS	F_0
Warehouses	126.0	3	42.00	7.00
Error	96.0	16	6.000	
Total	222.0	19		

Table 8.12. ANOVA table for the warehouse picking-time study ($a = 4$ randomly sampled warehouses, $n = 5$ orders each).

Solution.

Step 1 — The test.

This is a random factor: the four sites stand in for the whole network,

so the hypotheses are $H_0: \sigma_\tau^2 = 0$ against $H_1: \sigma_\tau^2 > 0$. The critical value is $f_{0.05,3,16} = 3.24$. Since $F_0 = 7.00 > 3.24$, reject H_0 ($p \approx 0.003$): warehouse-to-warehouse differences are a real source of variability across the network.

Step 2 — Variance components.

By (8.20),

$$\hat{\sigma}^2 = MS_E = 6.000, \qquad \hat{\sigma}_\tau^2 = \frac{42.00 - 6.000}{5} = \frac{36.00}{5} = 7.200.$$

Step 3 — Interpret the split.

The estimated total variance of a single picking time is (8.18): $7.200 + 6.000 = 13.20$ min^2, of which

$$\frac{7.200}{13.20} = 0.5455 \approx 55\%$$

is between warehouses and 45% is within a warehouse (Figure 8.7). The answer is $\boxed{\hat{\sigma}_\tau^2 = 7.20,\ \hat{\sigma}^2 = 6.00}$: more than half the picking-time variability is a site effect. Standardizing processes *across* warehouses attacks the larger component; training pickers *within* a warehouse attacks the smaller one.

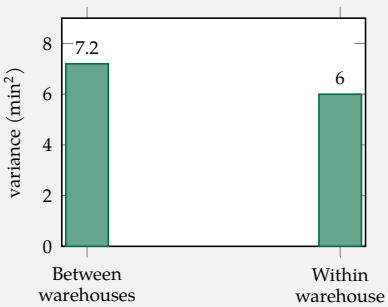

Figure 8.7. Estimated variance components for the warehouse study: 55% of picking-time variance is a between-site effect.

Safety lens. Fleet-level questions about AI systems are variance-component questions. "How much do disengagement rates vary from vehicle to vehicle, beyond within-vehicle noise?" samples vehicles as random levels; a large $\hat{\sigma}_\tau^2$ points to hardware or calibration drift rather than the shared software stack, and it changes where the engineering effort should go.

Engineering judgment. Fixed or random is decided by how the levels were chosen and what the conclusion must cover — before the data are collected. If you picked the levels on purpose, they are fixed. If you sampled them and want to speak about the population they came from, they are random. The ANOVA table cannot tell you which; the design can.

Reference. Montgomery & Runger, 6th ed., Section 13-4.

Nuisance factor. A variable that affects the response but is not itself of interest, such as day of week, operator, or raw-material batch.

Block. A group of experimental units that are alike with respect to a nuisance factor; every treatment appears once in each block.

8.11 *The Randomized Complete Block Design*

Return to fixed effects, and consider a complication every real experiment faces: a variable that influences the response but is not the question. A logistics team comparing routing policies knows that traffic varies by day of week. If the policies

are simply run on whatever days come up, the day-to-day variability lands in SS_E, inflates MS_E, and can bury a real policy difference. The nuisance variable cannot be eliminated — deliveries must happen every day — but it can be *blocked*: run every policy on every day, and let the analysis subtract the day effect out. The design principle is: *block what you can control; randomize what you cannot.*

Randomized complete block design (RCBD). A design with a treatments and b blocks in which every treatment appears exactly once in every block, in random order within each block.

In the randomized complete block design there are a treatments and b blocks, and each treatment is applied exactly once in each block, so $N = ab$. ''Complete'' means every block contains every treatment; ''randomized'' means the order (or assignment to units) *within* each block is random. The model adds one term to (8.2):

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \qquad i = 1, \dots, a, \quad j = 1, \dots, b, \tag{8.21}$$

where τ_i is the treatment effect ($\sum_i \tau_i = 0$), β_j is the block effect ($\sum_j \beta_j = 0$), and $\varepsilon_{ij} \sim N(0, \sigma^2)$ independently. The model is *additive*: it assumes a treatment's effect is the same in every block. The hypotheses of interest concern the treatments only: $H_0 \colon \tau_1 = \cdots = \tau_a = 0$ against $H_1 \colon$ at least one $\tau_i \neq 0$. Blocks are not on trial; they are there to be subtracted.

The total variability now partitions into *three* pieces,

$$SS_T = SS_{\text{Treatments}} + SS_{\text{Blocks}} + SS_E, \tag{8.22}$$

with computing forms (matching the course formula sheet)

$$\begin{aligned} SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{ab}, & SS_{\text{Treatments}} &= \frac{1}{b}\sum_{i=1}^{a} y_{i\cdot}^2 - \frac{y_{\cdot\cdot}^2}{ab}, \\ SS_{\text{Blocks}} &= \frac{1}{a}\sum_{j=1}^{b} y_{\cdot j}^2 - \frac{y_{\cdot\cdot}^2}{ab}, & SS_E &= SS_T - SS_{\text{Treatments}} - SS_{\text{Blocks}}, \end{aligned} \tag{8.23}$$

where $y_{\cdot j}$ is the total of block j. The degrees of freedom are $a - 1$ for treatments, $b - 1$ for blocks, and what remains, $(ab - 1) - (a - 1) - (b - 1) = (a - 1)(b - 1)$, for error. The treatment test statistic is

$$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}, \qquad F_0 \sim F_{a-1,\,(a-1)(b-1)} \text{ under } H_0, \tag{8.24}$$

rejecting when $F_0 > f_{\alpha,\,a-1,\,(a-1)(b-1)}$. Table 8.13 shows the layout.

Source	Sum of squares	df	Mean square	F_0
Treatments	$b\sum_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	$a-1$	$MS_{\text{Treatments}}$	$MS_{\text{Treatments}}/MS_E$
Blocks	$a\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	$b-1$	MS_{Blocks}	
Error	by subtraction	$(a-1)(b-1)$	MS_E	
Total	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot})^2$	$ab-1$		

Table 8.13. The RCBD ANOVA table. The block row absorbs nuisance variability that would otherwise inflate the error row.

Insight. Why blocking increases power. SS_{Blocks} is variability that, in a completely randomized analysis, would have been part of SS_E. Pulling it out shrinks MS_E, and since $F_0 = MS_{\text{Treatments}}/MS_E$, a smaller denominator makes real treatment effects easier to see. The price is modest: $b-1$ degrees of freedom move from the error row to the block row. Blocking pays whenever the nuisance variability it removes outweighs the degrees of freedom it costs — which is almost always, when the blocking variable genuinely affects the response.

After a significant treatment test, multiple comparisons work exactly as in Section 8.7, with two mechanical substitutions: each treatment mean now averages b observations (one per block), and the error degrees of freedom are $(a-1)(b-1)$. The yardsticks become

$$\text{LSD} = t_{\alpha/2,\,(a-1)(b-1)}\sqrt{\frac{2\,MS_E}{b}}, \tag{8.25}$$

$$T_\alpha = q_\alpha\big(a,\ (a-1)(b-1)\big)\sqrt{\frac{MS_E}{b}}. \tag{8.26}$$

Residuals also acquire one extra term. The fitted value under the additive model (8.21) combines the treatment mean and the block mean, $\hat{y}_{ij} = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$, so

$$e_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}. \tag{8.27}$$

The usual plots apply, plus one specific to the RCBD: look for a systematic pattern of residuals *within* blocks (for instance, one treatment always high in early blocks and always low in late ones). Such a pattern signals a treatment-by-block *interaction*, which violates the additivity assumption; the factorial designs of Chapter 9 are built to model it.

Procedure 8.5 — RCBD analysis.

1. Lay out the data as an $a \times b$ table (treatments as rows, blocks as columns). Compute row totals $y_{i\cdot}$, column totals $y_{\cdot j}$, the grand total $y_{\cdot\cdot}$, and the corresponding means.
2. Compute SS_T, $SS_{\text{Treatments}}$, and SS_{Blocks} from (8.23), then SS_E by subtraction.
3. Fill in the degrees of freedom $a-1$, $b-1$, $(a-1)(b-1)$, $ab-1$, and verify they add up.
4. Compute $MS_{\text{Treatments}}$, MS_{Blocks}, MS_E, and F_0 from (8.24); reject H_0 if $F_0 > f_{\alpha, a-1, (a-1)(b-1)}$.
5. If H_0 is rejected, run multiple comparisons with (8.25) or (8.26) (note b in place of n).
6. Check residuals (8.27): normal plot, residuals versus fitted values, and residuals inspected within blocks for interaction patterns.

Example 8.8 — Four routing policies, blocked by day of week

A last-mile logistics team compares four routing policies for its delivery vans: P1 (current commercial software), P2 (a new optimization engine), P3 (driver-chosen routes), and P4 (the optimization engine with live-traffic updates disabled). The response is the mean delivery time per package (minutes). Traffic varies strongly by day of week, so the team blocks on day: each policy is run once on each working day (Sunday through Thursday), with the within-day order randomized, giving $a = 4$, $b = 5$, $N = 20$ (Table 8.14). Test at $\alpha = 0.05$ whether the policies differ.

Policy	Sun	Mon	Tue	Wed	Thu	$y_{i\cdot}$	$\bar{y}_{i\cdot}$
P1. Current	41	40	42	43	44	210	42.0
P2. Optimizer	37	39	41	41	42	200	40.0
P3. Driver-chosen	43	45	45	45	47	225	45.0
P4. Optimizer, no live	39	40	40	43	43	205	41.0
$y_{\cdot j}$	160	164	168	172	176	840	
$\bar{y}_{\cdot j}$	40.0	41.0	42.0	43.0	44.0		42.0

Table 8.14. Delivery times (minutes per package) for four routing policies blocked by day of week. The last cell is the grand mean $\bar{y}_{\cdot\cdot} = 42.0$.

Solution.

Step 1 — Totals.

Row and column totals are in Table 8.14: $y_{\cdot\cdot} = 840$, $\bar{y}_{\cdot\cdot} = 42.0$. The block means climb steadily from 40 on Sunday to 44 on Thursday — the day effect the team blocked for is plainly there (Figure 8.8).

Step 2 — Sums of squares.

The sum of all squared observations is $\sum\sum y_{ij}^2 = 8830 + 8016 + 10133 + 8419 = 35398$, and $y_{\cdot\cdot}^2/ab = 840^2/20 = 35280$. By (8.23),

$$SS_T = 35398 - 35280 = 118.0,$$

$$SS_{\text{Treatments}} = \frac{210^2 + 200^2 + 225^2 + 205^2}{5} - 35280 = \frac{176450}{5} - 35280 = 70.00,$$

$$SS_{\text{Blocks}} = \frac{160^2 + 164^2 + 168^2 + 172^2 + 176^2}{4} - 35280 = \frac{141280}{4} - 35280 = 40.00,$$

$$SS_E = 118 - 70 - 40 = 8.000.$$

(The definition forms agree: $SS_{\text{Treatments}} = 5[(0)^2 + (-2)^2 + (3)^2 + (-1)^2] = 5(14) = 70$ and $SS_{\text{Blocks}} = 4[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2] = 4(10) = 40$.)

Step 3 — The table and the test.

Degrees of freedom: $a - 1 = 3$, $b - 1 = 4$, $(a-1)(b-1) = 12$, total 19; they add up. The completed table is Table 8.15. With $F_0 = 23.33/0.6667 = 35.00$ and $f_{0.05,3,12} = 3.49$, reject H_0 decisively ($p < 0.001$): the routing policies differ.

Source	SS	df	MS	F_0
Policies (treatments)	70.00	3	23.33	35.00
Days (blocks)	40.00	4	10.00	
Error	8.000	12	0.6667	
Total	118.0	19		

Table 8.15. RCBD ANOVA table for the routing study. Compare the error mean square with the unblocked value computed in Step 4.

Step 4 — What blocking bought.

Suppose the team had ignored days and analyzed the same numbers as a completely randomized design. The day variability would fold into the error row: $SS'_E = 40 + 8 = 48$ with 16 degrees of freedom, so $MS'_E = 3.000$ — four and a half times larger than the blocked $MS_E = 0.6667$ — and F_0 would drop from 35.00 to $23.33/3.000 = 7.78$. Still significant here, but a smaller policy effect would have been lost entirely. The answer is $\boxed{F_0 = 35.00 > 3.49 \;\Rightarrow\; \text{reject } H_0}$.

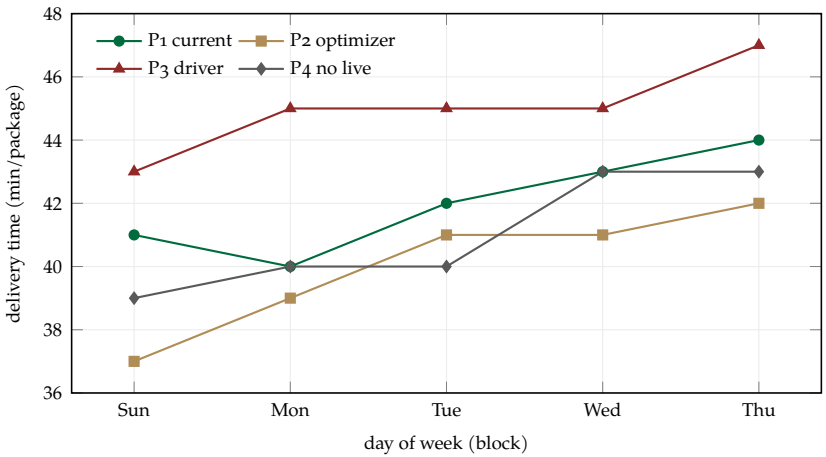

Figure 8.8. The routing data of Example 8.8 by block. All four policies drift upward through the week — the common day effect. The RCBD subtracts that shared drift, so the policies are compared by their vertical separation, not blurred by the slope.

Example 8.9 — After the RCBD: which policies differ, and does the model hold?

For the routing study of Example 8.8: (a) apply Tukey's HSD at family-wise $\alpha = 0.05$; (b) compute the residuals and check the additive model.

Solution.

Step (a) — Tukey with b in place of n.

Here $a = 4$, error df $= 12$, and $q_{0.05}(4, 12) = 4.20$. By (8.26),

$$T_{0.05} = 4.20\sqrt{\frac{0.6667}{5}} = 4.20\sqrt{0.1333} = 4.20(0.3651) = 1.534.$$

Ordering the means — P2 (40.0), P4 (41.0), P1 (42.0), P3 (45.0) — and comparing all six pairs: P2–P4 (1.0) and P4–P1 (1.0) do not exceed 1.534; P2–P1 (2.0), P2–P3 (5.0), P4–P3 (4.0), and P1–P3 (3.0) all do. The underline diagram (Figure 8.9) has two overlapping groups at the fast end and P3 isolated at the slow end. Engineering summary: the optimizer (P2) is certified faster than the current software (P1) and much faster than driver-chosen routes (P3); whether live-traffic updates matter (P2 vs. P4) is not settled by this experiment.

Figure 8.9. Tukey underline diagram for the routing policies. P2–P4 and P4–P1 share lines (not separated); P3 stands alone as the slowest.

Step (b) — Residuals.

By (8.27), for example $e_{11} = 41 - 42.0 - 40.0 + 42.0 = +1$. The full residual matrix is in Table 8.16. Every row and every column sums to zero (a built-in check of the arithmetic), all residuals are ± 1 or 0, and no policy is systematically high early in the week and low late. There is no sign of a policy-by-day interaction, so the additive model (8.21) is adequate, and $SS_E = 8$ is confirmed as the sum of the squared entries (8 entries of ± 1). The answer to (a) is P2, P4, P1 not fully separated; P3 slower than all.

Residual e_{ij}	Sun	Mon	Tue	Wed	Thu	Row sum
P1	$+1$	-1	0	0	0	0
P2	-1	0	$+1$	0	0	0
P3	0	$+1$	0	-1	0	0
P4	0	0	-1	$+1$	0	0
Column sum	0	0	0	0	0	0

Table 8.16. RCBD residuals (8.27) for the routing study. Zero margins are automatic; patterns within rows or columns would signal a treatment-by-block interaction.

Startup lens. Cohorts are blocks. Signups arriving in a product-launch week behave differently from signups in a quiet week. Testing onboarding flows *within* each weekly cohort, RCBD-style, keeps a lucky week from crowning the wrong variant.

Table 8.17 closes the loop on the chapter's three analyses: the choice among them is made by two design questions, not by the data.

Situation	Design and analysis	Inference target
Specific levels chosen on purpose	Fixed-effects one-way ANOVA (Procedure 8.1)	Compare those means; follow with Procedure 8.2 or 8.3
Levels sampled from a population	Random-effects model (Procedure 8.4)	Variance components σ_τ^2, σ^2
Known nuisance factor present	Randomized complete block design (Procedure 8.5)	Compare treatment means, nuisance removed

Table 8.17. Choosing the right single-factor design. The decision is made when the experiment is planned, before any data exist.

Chapter Summary

This chapter replaced the two-sample comparison with the general a-sample comparison. Repeated pairwise t tests inflate the family-wise type I error to $1-(1-\alpha)^m$ (8.1), so all means are tested at once. Under the effects model (8.2), total variability splits into a between-treatments piece and a within-treatments piece (8.5); dividing by degrees of freedom gives the mean squares (8.8)–(8.9), whose ratio F_0 (8.11) is compared with $f_{\alpha,a-1,N-a}$ (Procedure 8.1). MS_E then powers confidence intervals for means (8.12) and differences (8.13). After a significant F, Fisher's LSD (8.14) screens pairs quickly but lets the family-wise error grow, while Tukey's HSD (8.15) holds it at α using the Studentized range

(Procedures 8.2 and 8.3). Residuals (8.16) check the assumptions, and the non-centrality parameter (8.17) sets the sample size before the experiment runs. When the levels are themselves a random sample, the same ANOVA table tests $H_0 : \sigma_\tau^2 = 0$ and yields the variance-component estimates (8.20). When a known nuisance factor threatens the comparison, the randomized complete block design removes it: the decomposition gains a block term (8.22), the treatment test becomes (8.24) with $(a-1)(b-1)$ error degrees of freedom (Procedure 8.5), and the multiple-comparison yardsticks swap n for b (8.25)–(8.26).

Quantity	Formula	Equation
Family-wise error	$1-(1-\alpha)^m$	(8.1)
Effects model	$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$	(8.2)
Decomposition	$SS_T = SS_{\text{Treatments}} + SS_E$	(8.5)
SS_T shortcut	$\sum\sum y_{ij}^2 - y_{\cdot\cdot}^2/N$	(8.6)
$SS_{\text{Treatments}}$ shortcut	$\frac{1}{n}\sum_i y_{i\cdot}^2 - y_{\cdot\cdot}^2/N$	(8.7)
Mean squares	$MS_{\text{Trt}} = \frac{SS_{\text{Trt}}}{a-1}, \quad MS_E = \frac{SS_E}{N-a}$	(8.8), (8.9)
ANOVA F statistic	$F_0 = MS_{\text{Treatments}}/MS_E \sim F_{a-1,N-a}$	(8.11)
CI for μ_i	$\bar{y}_{i\cdot} \pm t_{\alpha/2,N-a}\sqrt{MS_E/n}$	(8.12)
CI for $\mu_i - \mu_k$	$(\bar{y}_{i\cdot} - \bar{y}_{k\cdot}) \pm t_{\alpha/2,N-a}\sqrt{2MS_E/n}$	(8.13)
Fisher LSD	$t_{\alpha/2,N-a}\sqrt{2\,MS_E/n}$	(8.14)
Tukey HSD	$T_\alpha = q_\alpha(a,f)\sqrt{MS_E/n}$	(8.15)
Non-centrality	$\Phi^2 = n\sum\tau_i^2/(a\sigma^2) \geq nD^2/(2a\sigma^2)$	(8.17)
Variance components	$\hat{\sigma}^2 = MS_E, \ \hat{\sigma}_\tau^2 = \frac{MS_{\text{Trt}} - MS_E}{n}$	(8.20)
RCBD model	$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$	(8.21)
RCBD decomposition	$SS_T = SS_{\text{Trt}} + SS_{\text{Blocks}} + SS_E$	(8.22)
RCBD F statistic	$F_0 = MS_{\text{Trt}}/MS_E \sim F_{a-1,(a-1)(b-1)}$	(8.24)
RCBD LSD	$t_{\alpha/2,(a-1)(b-1)}\sqrt{2\,MS_E/b}$	(8.25)
RCBD residual	$e_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}$	(8.27)

Table 8.18. Key formulas of Chapter 8 at a glance.

Exercises

One-way ANOVA computation

Exercise 8.1. An autonomous-vehicle team compares three sensor-fusion algorithms by their object-localization error (cm) on a fixed test track. Each algorithm is run on $n = 4$ randomized trials:

Algorithm	Error (cm)				$\bar{y}_{i\cdot}$
A	7	9	8	8	8.0
B	11	12	10	11	11.0
C	8	10	9	9	9.0

Build the complete ANOVA table and test, at $\alpha = 0.05$, whether the algorithms differ in mean error. Use $f_{0.05,2,9} = 4.26$.

Exercise 8.2. In a one-way ANOVA with $a = 5$ treatments and $n = 4$ replicates each, an engineer computes $F_0 = 1.85$; the critical value is $f_{0.05,4,15} = 3.06$. Which statement is most accurate?

(A) Since $F_0 > 1$, at least one mean differs from the others.

(B) Fail to reject H_0; the data do not provide evidence that the means differ.

(C) The test is inconclusive, so pairwise comparisons should be performed anyway.

(D) Reduce α to 0.10 and re-test, since the result is borderline.

Exercise 8.3. A freight broker compares $a = 4$ carriers with $n = 6$ observed deliveries each. The analyst reports $SS_T = 280$ and $SS_{\text{Treatments}} = 120$ but spills coffee on the rest of the output. Reconstruct the full ANOVA table and complete the test at $\alpha = 0.05$, given $f_{0.05,3,20} = 3.10$.

Multiple comparisons

Exercise 8.4. For the sensor-fusion study of Exercise 8.1 ($MS_E = 0.6667$, $n = 4$, error df $= 9$, means 8.0, 11.0, 9.0), compute Fisher's LSD at $\alpha = 0.05$ using $t_{0.025,9} = 2.262$, and state which algorithm pairs differ.

Exercise 8.5. Repeat the analysis of Exercise 8.4 with Tukey's HSD, using $q_{0.05}(3, 9) = 3.95$. Do the conclusions change? Sketch the underline diagram.

Exercise 8.6. An engineer compares $a = 8$ surface treatments; the overall F test is significant. To declare which pairs differ while controlling the family-wise type I error rate at $\alpha = 0.05$, the engineer should:

(A) use Fisher's LSD, because its critical value is smaller.

(B) use either procedure; both control the family-wise rate identically.

(C) use Tukey's HSD, because Fisher's LSD inflates the family-wise error rate when a is large.

(D) use neither; with $a = 8$ only the overall F test is meaningful.

Random effects and variance components

Exercise 8.7. An autonomous-vehicle operator randomly selects $a = 5$ vehicles from its fleet and measures a perception-latency score on $n = 6$ randomized runs per vehicle. The ANOVA gives $MS_{\text{Treatments}} = 12.0$ and $MS_E = 4.0$. (a) Test $H_0 : \sigma_\tau^2 = 0$ at $\alpha = 0.05$, given $f_{0.05,4,25} = 2.76$. (b) Estimate both variance components and the proportion of total variance due to vehicle-to-vehicle differences.

Exercise 8.8. An AI evaluation team randomly samples $a = 4$ annotators from its labeling pool; each scores $n = 5$ model outputs. The ANOVA gives $MS_{\text{Treatments}} = 3.2$ and $MS_E = 4.0$. Applying the variance-component formula, the team computes $\hat{\sigma}_\tau^2 = (3.2 - 4.0)/5 = -0.16$ and asks how a variance can be negative. Explain what happened and state the correct report.

Randomized complete block designs

Exercise 8.9. A startup tests three pricing-page variants (P, Q, R). Because visitors from different acquisition channels behave differently, each variant is shown to visitors within each of $b = 4$ channels (paid ads, organic search, referral, email), and the response is conversion rate (%):

Variant	Ads	Organic	Referral	Email	$\bar{y}_{i\cdot}$
P	16	16	19	21	18.0
Q	16	20	21	23	20.0
R	19	21	23	25	22.0
$\bar{y}_{\cdot j}$	17.0	19.0	21.0	23.0	20.0

Given $\sum\sum y_{ij}^2 = 4896$, build the RCBD ANOVA table and test at $\alpha = 0.05$ whether the variants differ, using $f_{0.05,2,6} = 5.14$.

Exercise 8.10. An engineer runs an RCBD with $a = 4$ treatments and $b = 5$ blocks. The error degrees of freedom for the treatment F test are:

(A) $N - 1 = 19$

(B) $N - a = 16$

(C) $(a-1)(b-1) = 12$

(D) $a(n-1) = 16$

Exam-style problems

Exercise 8.11. An AI safety team compares $a = 4$ prompt-defense configurations (A, B, C, D) for a customer-facing chatbot. Each configuration is evaluated on $n = 5$ independent randomized batches of adversarial prompts; the response is the number of blocked attacks per batch. The sample means are $\bar{y}_{A\cdot} = 10$, $\bar{y}_{B\cdot} = 14$, $\bar{y}_{C\cdot} = 12$, $\bar{y}_{D\cdot} = 12$, and the within-group sum of squares is $SS_E = 32$. Critical values: $f_{0.05,3,16} = 3.24$, $q_{0.05}(4, 16) = 4.05$.

1. **[2 pt]** State H_0 and H_1 in terms of the effects model (8.2).

2. **[6 pt]** Compute $SS_{\text{Treatments}}$ and build the complete ANOVA table.

3. **[3 pt]** Carry out the F test at $\alpha = 0.05$ and state the conclusion in context.

4. **[6 pt]** Apply Tukey's HSD at family-wise $\alpha = 0.05$ and report which configurations differ.

5. **[3 pt]** The team wants to ship the single best configuration. Write a one-sentence recommendation consistent with parts (c) and (d).

Exercise 8.12. A container terminal compares three truck-loading procedures (L1, L2, L3) on loading time (minutes). Crews differ in experience, so the study is run as an RCBD with $b = 4$ crews; each crew uses every procedure once, in random order:

Procedure	Crew 1	Crew 2	Crew 3	Crew 4	$\bar{y}_{i\cdot}$
L1	22	22	25	27	24.0
L2	23	27	28	30	27.0
L3	27	29	31	33	30.0
$\bar{y}_{\cdot j}$	24.0	26.0	28.0	30.0	27.0

Given $\sum\sum y_{ij}^2 = 8884$. Critical values: $f_{0.05,2,6} = 5.14$, $f_{0.05,2,9} = 4.26$, $t_{0.025,6} = 2.447$.

1. **[2 pt]** Why is blocking on crew appropriate here? Answer in one sentence.
2. **[6 pt]** Compute SS_T, $SS_{\text{Treatments}}$, SS_{Blocks}, and SS_E, and assemble the ANOVA table.
3. **[3 pt]** Test at $\alpha = 0.05$ whether the procedures differ.
4. **[4 pt]** Apply Fisher's LSD (8.25) at $\alpha = 0.05$ and rank the procedures.
5. **[3 pt]** Recompute F_0 as if the crews had been ignored (one-way analysis) and comment on what blocking accomplished.
6. **[2 pt]** One residual check is specific to the RCBD. Name it and say what pattern would be a warning.

Assessing outside recommendations

Exercise 8.13. A consultant compares four routing policies for your fleet by running all six pairwise two-sample t tests, each at $\alpha = 0.05$, and reports: ''Two comparisons came out significant, so we have solid evidence for two real differences at the 95% confidence level.'' Identify what is wrong with this claim, quantify it, and describe the analysis the consultant should have run.

Exercise 8.14. Your startup tested three onboarding flows across $b = 6$ weekly signup cohorts in a randomized complete block design (every flow, every week). An AI assistant analyzed the data as a one-way ANOVA, producing $SS_{\text{Treatments}} = 30$ with 2 df and a residual sum of squares of 100 with 15 df, hence $F_0 = 15.0/6.667 = 2.25 < f_{0.05,2,15} = 3.68$, and concluded: ''No significant flow effect; keep the current flow.'' A teammate notes that the week-to-week block sum of squares alone is 80. Find the error, redo the analysis correctly ($f_{0.05,2,10} = 4.10$), and state the correct conclusion.

In-Class Activity 8.1: Three Red Teams, One Verdict

Week 12 — Lecture 22

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

Your team manages safety evaluation for a chatbot release. Three red-team strategies were each run in 3 randomized sessions; the response is distinct defects found per session:

Strategy	Defects			$\bar{y}_{i\cdot}$
Mutation	4	6	5	5.0
Role-play	7	9	8	8.0
Escalation	4	5	6	5.0

Useful values: $f_{0.05,2,6} = 5.14$, grand mean $\bar{y}_{\cdot\cdot} = 6.0$.

Part 1 — Why not three t tests?

A teammate proposes three pairwise t tests at $\alpha = 0.05$ each. Compute the approximate family-wise error rate of that plan, and write one sentence to the teammate explaining why the team will run ANOVA instead.

Part 2 — Build the table

Compute $SS_{\text{Treatments}}$, SS_E, and the degrees of freedom, then assemble the ANOVA table and the statistic F_0. (Work from the treatment means and deviations; the numbers are small on purpose.)

de and say what it means

n at $\alpha = 0.05$ and write one sentence for the release notes: what did the experiment establish, *not* establish about which strategies differ?

^{2} whether or not H_0 is true. What does $MS_{\text{Treatments}}$ estimate when H_0 is false?

In-Class Activity 8.2: Which Flow Do We Ship?

Week 12 — Lecture 23

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

Your startup's onboarding experiment produced the ANOVA summary: $a = 3$ flows, $n = 5$ cohorts each, $\bar{y}_{A\cdot} = 62.0$, $\bar{y}_{B\cdot} = 65.5$, $\bar{y}_{C\cdot} = 67.0$ (activation %), $MS_E = 5.0$ with 12 error df, and a significant overall F test. Useful values: $t_{0.025,12} = 2.179$, $q_{0.05}(3, 12) = 3.77$.

Part 1 — Two yardsticks

Compute Fisher's LSD and Tukey's $T_{0.05}$ for these data. Before comparing any pair, predict which yardstick is larger and say why.

Part 2 — Apply both

For each pair (A–B, A–C, B–C), record the gap and whether it exceeds each yardstick. Draw the Tukey underline diagram.

recommendation

rison comes out differently under the two procedures. Write the team's shipping recommenda-
o sentences, stating which procedure you trust for the final report and why.

what exactly does ''family-wise $\alpha = 0.05$'' promise?

In-Class Activity 8.3: Block the Weekday

Week 13 — Lecture 24

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

A dispatch team compared three delivery carriers, running each carrier once on each of four days (an RCBD blocked on day). Summary: treatment means 31, 33, 35 minutes; block (day) means 30, 31, 35, 36; grand mean 33; total sum of squares $SS_T = 116$. Useful values: $f_{0.05,2,6} = 5.14$, $f_{0.05,2,9} = 4.26$.

Part 1 — Decompose

Compute $SS_{\text{Treatments}}$ and SS_{Blocks} from the means, then find SS_E by subtraction and list all degrees of freedom.

Part 2 — Test twice

Compute the treatment F_0 for the correct RCBD analysis. Then compute what F_0 would have been had the days been ignored (fold SS_{Blocks} back into the error). Compare both against their critical values.

sentence to the manager

s reach different verdicts. Explain to a manager, without formulas, what the blocked analysis nblocked one could not.

th one reason: blocking increases the treatment sum of squares.