9 *Two-Factor and Factorial Experiments*

Chapter 8 analyzed experiments with a single factor: several treatments, one question. Most engineering systems, however, are driven by several factors at once, and the factors often work together. The effect of a warehouse layout can depend on the shift that uses it; the failure rate of a perception model can depend on the scenario it is tested in; the value of a new pricing page can depend on the onboarding flow behind it. This chapter develops the factorial experiment, the design that varies all factors together and can therefore measure not only the effect of each factor but also their *interaction*. The chapter is organized in the following order: why factorial designs beat changing one factor at a time; main effects and interactions, and how to read an interaction plot; the two-factor effects model and its ANOVA, with a complete hand computation; what to do when the interaction is significant; residual diagnostics for factorial models; the special cases of one observation per cell and of three or more factors; the two-level 2^k designs that carry most industrial experimentation, including single replicates, center points, blocking, and fractional replication; and a closing section on how factorial thinking appears in real engineering studies.

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

Q1. A red team compares two guardrail versions for a language model, but runs the comparison only on English prompts. Version B wins, and it is deployed for all languages. What is the team implicitly assuming?

Q2. A plot shows mean container dwell time against crane type, with one line for each of two ship sizes. The two lines rise and fall together and stay the same vertical distance apart everywhere. What does this say about the two factors?

Q3. A startup tests $a = 3$ landing pages against $b = 2$ signup flows with $n = 3$ replicate traffic batches per combination. How many degrees of freedom does

the interaction have, and how many are left for error?

Answers.

Q1. That the guardrail effect is the same at every level of the untested factor (language); that is, that guardrail and language do *not* interact. If version B is better on English prompts but worse on Arabic prompts, the one-condition test cannot detect it. Testing all combinations of guardrail and language is a factorial experiment, and it can.

Q2. The parallel lines say there is no interaction: the effect of crane type is the same for both ship sizes. Each factor may still matter on its own (the lines are not flat, and they are separated), but the two factors act additively.

Q3. The interaction has $(a-1)(b-1) = 2 \times 1 = 2$ degrees of freedom. The error has $ab(n-1) = 6 \times 2 = 12$. Together with $df_A = 2$ and $df_B = 1$, they add to $abn - 1 = 17$, the total.

Lecture 25 starts here — *Two-factor factorial experiments*

Lecture 25. This section is covered by Lecture 25.

Reference. Montgomery & Runger, 6th ed., Sections 14-1 and 14-2.

One-factor-at-a-time (OFAT). An experimental strategy that holds all factors fixed except one, finds the best level of that factor, then moves to the next factor.

Factorial experiment. A design that runs every combination of the levels of all factors, so that main effects and interactions can all be estimated.

9.1 Why Factorial Experiments?

Suppose an engineer must choose operating settings for two factors, A and B, to maximize a response. The strategy most people try first is *one-factor-at-a-time* (OFAT): fix B at some starting value, vary A and pick the best level; then hold A at that winner and vary B. The strategy feels systematic, but it has two defects, one of efficiency and one of substance.

The substantive defect is that OFAT cannot detect an *interaction*: a situation in which the effect of factor A depends on the level of factor B. OFAT explores the factor space along two perpendicular lines, so it only ever sees how the response behaves along those lines. If the best region lies off the lines, OFAT never visits it, and it can confidently recommend a setting that is far from the true optimum. Figure 9.1 shows the classic picture. The response surface has its peak at the upper right. The OFAT experiment first varies A with B held low, finds the best A along that slice, then varies B at that A. It settles near a response of about 67, and it has no way of knowing that combinations it never ran reach about 90. The

factorial experiment on the right runs a grid of all combinations. The grid covers the corner region the OFAT lines miss, and the pattern of the nine responses points directly toward the peak.

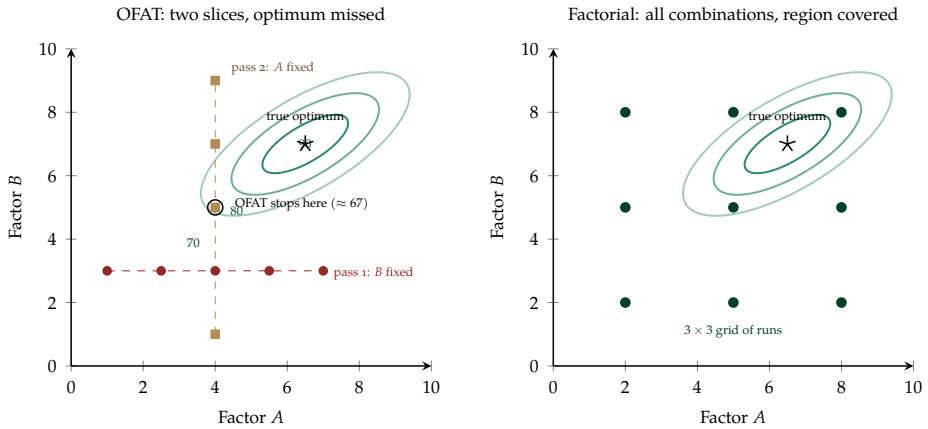

Figure 9.1. One-factor-at-a-time versus a factorial design on the same response surface (contours at 70, 80, 90). OFAT explores two perpendicular slices, settles at a response near 67, and never sees the peak. The factorial grid covers the whole region and reveals that the response keeps rising toward the upper right.

Startup lens. Running one A/B test after another — first the pricing page, then the onboarding flow — is OFAT with marketing vocabulary. If the two page elements interact, the sequential winners combined can convert worse than a combination the sequence never tested.

A *factorial experiment* removes both defects. With factor A at a levels and factor B at b levels, the design runs all ab combinations, each replicated n times, for $N = abn$ total runs. Its two advantages are:

- It detects interactions, because every level of A is observed at every level of B. OFAT structurally cannot do this.
- It is more efficient. Every one of the N observations contributes to the estimate of the A effect *and* to the estimate of the B effect at the same time. In an OFAT program, each run informs only the factor being varied. This reuse of the same data for every effect is sometimes called hidden replication.

Cell. One combination of factor levels in a factorial design. A two-factor design has ab cells, each holding n replicate observations.

Please pay attention to the vocabulary here, because the rest of the chapter uses it constantly. Each combination of one level of A with one level of B is a *cell* (also called a treatment combination). Each of the n repeated observations inside a cell is a *replicate*. Replication is what allows the experiment to estimate the random error σ^2 separately from the systematic effects; Section 9.9 shows what breaks when $n = 1$.

Reference. Montgomery & Runger, 6th ed., Section 14-3.

Main effect. The average change in the response when a factor moves between its levels, averaged over the levels of all other factors.

Interaction. Two factors interact when the effect of one factor on the response is different at different levels of the other factor.

9.2 *Main Effects and Interaction*

The *main effect* of factor A is the average change in the response when A changes level, where the average is taken over all levels of B. The main effect of B is defined the same way with the roles reversed. Main effects answer the question ''does this factor matter on average?''

The *interaction* between A and B, written AB, measures how far the cells depart from that averaged story. A and B interact when the effect of A is not the same at every level of B. A small numerical case makes this concrete. Suppose a 2×2 experiment gives cell means

	B_1	B_2
A_1	20	30
A_2	40	52

versus

	B_1	B_2
A_1	20	30
A_2	40	12

In the left table, moving from A_1 to A_2 raises the mean by 20 at B_1 and by 22 at B_2: nearly the same change at both levels, so there is essentially no interaction, and the sentence ''A_2 beats A_1 by about 21'' is a fair summary. In the right table, the same move raises the mean by 20 at B_1 but *lowers* it by 18 at B_2. No single sentence about A alone is honest there; the effect of A depends entirely on B.

The standard picture for this comparison is the *interaction plot*: plot the cell means against the levels of one factor, with one line per level of the other factor. Figure 9.2 shows the three patterns you must be able to recognize.

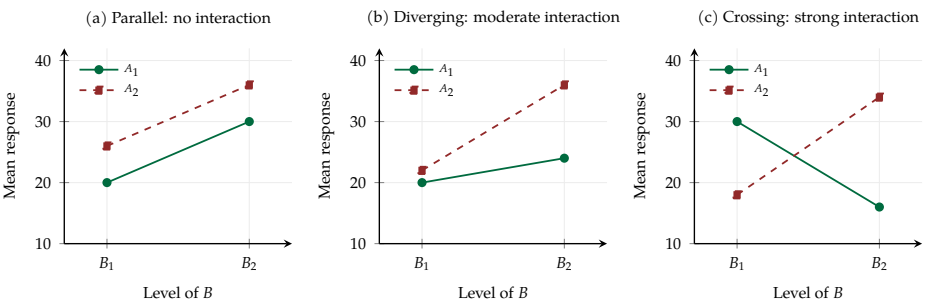

Figure 9.2. The three interaction-plot patterns. (a) Parallel lines: the factors act additively; main effects tell the whole story. (b) Lines that diverge without crossing: the direction of the A effect is stable but its size depends on B; a moderate interaction, common in real data. (c) Crossing lines: which level of A is best reverses with B; a strong interaction, and main effects alone are misleading.

Reading the plot follows three questions, always in this order. First, are the lines parallel? If yes, there is no visible interaction and each factor can be summarized

on its own. Second, if the lines are not parallel, do they cross? Diverging lines mean the size of the effect changes with the other factor; crossing lines mean even its *direction* changes. Third, only after the interaction question is settled, look at the average slope (the B main effect) and the average separation between lines (the A main effect). Remember that the plot shows sample means, which contain noise: lines from a real experiment are never perfectly parallel, and the ANOVA of Section 9.5 is the formal test of whether the departure from parallel is larger than chance.

Example 9.1 — Reading an interaction plot: fleet energy cost

A logistics company compares two truck types, diesel (A_1) and electric (A_2), on three route terrains: flat, hilly, and mountain. The response is energy cost in SAR per km, and the cell means are plotted in Figure 9.3. Interpret the plot.

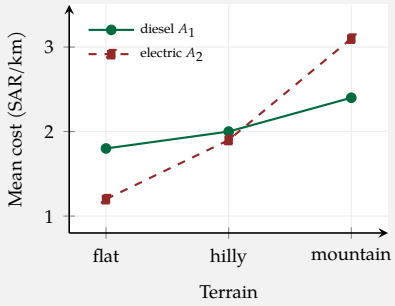

Figure 9.3. Cell-mean plot for truck type × terrain. The lines cross between hilly and mountain terrain.

Solution.

Step 1 — Parallel?

No. The diesel line rises gently ($1.8 \to 2.4$) while the electric line rises steeply ($1.2 \to 3.1$). The vertical gap between the lines changes from -0.6 on flat terrain to $+0.7$ on mountain terrain, so the effect of truck type depends on terrain: an interaction.

Step 2 — Crossing?

Yes, between hilly and mountain. Electric is cheaper on flat terrain (1.2 vs. 1.8 SAR/km) and roughly equal on hilly terrain, but diesel is cheaper on mountain routes (2.4 vs. 3.1 SAR/km). The *ranking* of truck types reverses.

Step 3 — What may be reported.

No statement of the form "electric trucks are cheaper to run" is defensible, even though the electric row mean, $(1.2 + 1.9 + 3.1)/3 = 2.067$, equals the diesel row mean, $(1.8 + 2.0 + 2.4)/3 = 2.067$. The honest report is terrain-by-terrain: assign electric to flat routes, diesel to mountain routes, with either type acceptable on hilly routes pending the formal test.

Safety lens. An AI evaluation grid — model versions down the rows, test scenario families across the columns — is exactly a two-factor factorial. A leaderboard that averages over scenarios reports only the main effect of model version and hides every version-by-scenario interaction, which is often where the safety-relevant behavior lives.

Reference. Montgomery & Runger, 6th ed., Section 14-3.1.

9.3 The Two-Factor Effects Model

To test main effects and interactions formally, we need a model. Let Y_{ijk} be the k-th replicate observation in the cell where A is at level i and B is at level j. The *two-factor effects model* is

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}, \qquad \begin{cases} i = 1, 2, \ldots, a, \\ j = 1, 2, \ldots, b, \\ k = 1, 2, \ldots, n, \end{cases} \tag{9.1}$$

where μ is the overall mean, τ_i is the main effect of the i-th level of A, β_j is the main effect of the j-th level of B, $(\tau\beta)_{ij}$ is the interaction effect of cell (i, j), and the errors ε_{ijk} are independent $N(0, \sigma^2)$ random variables. The symbol $(\tau\beta)_{ij}$ is a single parameter, not a product; the parentheses are part of the name. The effects are deviations from the overall mean, so they satisfy the constraints $\sum_i \tau_i = 0$, $\sum_j \beta_j = 0$, and $\sum_i (\tau\beta)_{ij} = \sum_j (\tau\beta)_{ij} = 0$.

Insight. The mean of cell (i, j) under the model is $\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij}$. The first three terms are the additive prediction: overall level, plus the row adjustment, plus the column adjustment. The interaction term $(\tau\beta)_{ij}$ is whatever is left: the amount by which that particular combination beats or misses its additive prediction. Parallel lines in the interaction plot are exactly the statement that every $(\tau\beta)_{ij}$ is zero.

The experiment tests three null hypotheses at once, one per set of effects:

$$H_0^{AB}: (\tau\beta)_{ij} = 0 \text{ for all } i, j, \qquad H_0^{A}: \tau_i = 0 \text{ for all } i, \qquad H_0^{B}: \beta_j = 0 \text{ for all } j. \tag{9.2}$$

Each is tested against the alternative that at least one of its effects is nonzero. The interaction hypothesis is listed first in (9.2) deliberately: Section 9.5 explains why it must also be *tested* first.

The notation for totals extends the dot convention of Chapter 8. A dot in a subscript position means "summed over that index": $y_{i\cdot\cdot}$ is the total of all bn observations at level i of A (a row total), $y_{\cdot j\cdot}$ is the total at level j of B (a column total), $y_{ij\cdot}$ is the total of the n replicates in cell (i, j), and $y_{\cdots}$ is the grand total of all $N = abn$ observations. Bars denote the corresponding means, so $\bar{y}_{ij\cdot} = y_{ij\cdot}/n$ is a cell mean and $\bar{y}_{\cdots}$ is the grand mean.

Reference. Montgomery & Runger, 6th ed., Section 14-3.1.

9.4 *Partitioning the Variability*

Exactly as in one-way ANOVA, the analysis rests on splitting the total sum of squares into recognizable pieces. For the two-factor factorial the split has *four* components:

$$SS_T \;=\; SS_A + SS_B + SS_{AB} + SS_E. \tag{9.3}$$

SS_A measures variability between the row averages, SS_B between the column averages, SS_{AB} the departure of the cell means from additivity, and SS_E the scatter of replicates inside their own cells. The identity is an algebraic fact, obtained by writing each observation as

$$y_{ijk} - \bar{y}_{\cdots} = (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}) + (y_{ijk} - \bar{y}_{ij\cdot}),$$

squaring, and summing; all cross-product terms cancel. The four terms on the right are the sample versions of τ_i, β_j, $(\tau\beta)_{ij}$, and ε_{ijk} from (9.1), so each sum of squares grows when its set of effects is real and stays small when they are zero.

For hand computation, the practical formulas use totals rather than means. With $N = abn$ and the correction term $y_{\cdots}^2/abn$,

$$SS_T \;=\; \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 \;-\; \frac{y_{\cdots}^2}{abn}, \tag{9.4}$$

$$SS_A \;=\; \frac{\sum_{i=1}^{a} y_{i\cdot\cdot}^2}{bn} \;-\; \frac{y_{\cdots}^2}{abn}, \tag{9.5}$$

$$SS_B \;=\; \frac{\sum_{j=1}^{b} y_{\cdot j\cdot}^2}{an} \;-\; \frac{y_{\cdots}^2}{abn}, \tag{9.6}$$

$$SS_{AB} \;=\; \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij\cdot}^{2} \;-\; \frac{y_{\cdots}^{2}}{abn} \;-\; SS_A \;-\; SS_B, \tag{9.7}$$

$$SS_E \;=\; SS_T - SS_A - SS_B - SS_{AB}. \tag{9.8}$$

Note the pattern shared by (9.5)–(9.7): square the totals, divide by the number of observations inside each total (bn per row, an per column, n per cell), and subtract the correction term. In (9.7), the first two terms give the variability between the ab cell totals; subtracting SS_A and SS_B then removes the part of that cell-to-cell variability already explained by the main effects, so what remains is pure interaction.

The degrees of freedom split the same way:

$$\underbrace{(a-1)}_{A} + \underbrace{(b-1)}_{B} + \underbrace{(a-1)(b-1)}_{AB} + \underbrace{ab(n-1)}_{\text{Error}} \;=\; abn - 1. \tag{9.9}$$

Each row average consumes $a-1$ free deviations, each column average $b-1$; the ab cell means, having already spent $1 + (a-1) + (b-1)$ degrees of freedom on the mean and main effects, leave $(a-1)(b-1)$ for interaction; and each cell contributes $n-1$ within-cell deviations to error. Please check this identity every time you build an ANOVA table; a degrees-of-freedom column that does not add to $abn - 1$ means an arithmetic slip somewhere upstream.

Reference. Montgomery & Runger, 6th ed., Section 14-3.1.

9.5 *The Two-Factor ANOVA Table and Its F Tests*

Dividing each sum of squares by its degrees of freedom gives the mean squares MS_A, MS_B, MS_{AB}, and $MS_E = SS_E/[ab(n-1)]$. Under the model, MS_E is always an unbiased estimator of σ^2, while each of the other mean squares estimates σ^2 *plus* a positive term contributed by its own effects. So each test statistic is a ratio with MS_E in the denominator, large when the corresponding effects are real. The interaction statistic is

$$F_0^{AB} \;=\; \frac{MS_{AB}}{MS_E} \;=\; \frac{SS_{AB}/[(a-1)(b-1)]}{SS_E/[ab(n-1)]}, \tag{9.10}$$

which under H_0^{AB} follows the F distribution with $(a-1)(b-1)$ numerator and $ab(n-1)$ denominator degrees of freedom. The main-effect statistics are

$$F_0^{A} \;=\; \frac{MS_A}{MS_E} \quad \text{with } (a-1) \text{ and } ab(n-1) \text{ df}, \qquad F_0^{B} \;=\; \frac{MS_B}{MS_E} \quad \text{with } (b-1) \text{ and } ab(n-1) \text{ df} \tag{9.11}$$

Each null hypothesis in (9.2) is rejected at level α when its statistic exceeds the critical value f_{α,ν_1,ν_2} with the matching degrees of freedom. Three tests, one shared error term. Table 9.1 collects the whole computation in the standard layout.

Source	Sum of squares	df	Mean square	F_0
A (rows)	SS_A	$a-1$	$MS_A = SS_A/(a-1)$	MS_A/MS_E
B (columns)	SS_B	$b-1$	$MS_B = SS_B/(b-1)$	MS_B/MS_E
AB (interaction)	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	MS_{AB}/MS_E
Error	SS_E	$ab(n-1)$	$MS_E = \frac{SS_E}{ab(n-1)}$	
Total	SS_T	$abn-1$		

Table 9.1. The ANOVA table for a two-factor factorial with n replicates per cell. All three F ratios share the denominator MS_E.

Insight. Why test the interaction first. The main effect of A is an average of the A effect over the levels of B. If the interaction is significant, that average may be averaging things that should not be averaged — a large gain at one level of B against a large loss at another — and can even come out near zero while A matters enormously level-by-level. So the interaction test decides how the other two tests may be read: if H_0^{AB} is rejected, interpret main effects cautiously and work with cell means (Section 9.7); if it is not rejected, the additive story holds and the main effects can be reported directly.

Procedure 9.1 — Building the two-factor ANOVA table.

1. Organize the data in an $a \times b$ table of cells. Compute the cell totals $y_{ij\cdot}$, the row totals $y_{i\cdot\cdot}$, the column totals $y_{\cdot j\cdot}$, and the grand total $y_{\cdots}$. Check: rows, columns, and cells must all add to the grand total.
2. Compute the correction term $C = y_{\cdots}^2/(abn)$ and the sum of all squared observations $\sum y_{ijk}^2$; then $SS_T = \sum y_{ijk}^2 - C$ by (9.4).
3. Compute SS_A from the row totals by (9.5) and SS_B from the column totals by (9.6).

4. Compute the cell-total sum $\frac{1}{n}\sum_{i,j} y_{ij\cdot}^2 - C$ and subtract SS_A and SS_B to get SS_{AB} by (9.7).
5. Obtain SS_E by subtraction, (9.8). It must be nonnegative; a negative value signals an arithmetic error.
6. Fill in the degrees of freedom $a-1$, $b-1$, $(a-1)(b-1)$, $ab(n-1)$, and verify they add to $abn-1$ as in (9.9). Divide to get the mean squares.
7. Compute F_0^{AB} by (9.10) and test it against $f_{\alpha,(a-1)(b-1),ab(n-1)}$ *first*.
8. Compute F_0^A and F_0^B by (9.11) and test each against its critical value. State every conclusion in the language of the problem, respecting the interaction result from step 7.

Example 9.2 — Counting degrees of freedom before running

A startup plans a factorial test of $a = 4$ landing-page designs against $b = 3$ traffic sources (search ads, social, email), with $n = 5$ replicate traffic batches per combination; the response is an engagement score. Before any data arrive, find the degrees of freedom for every line of the ANOVA table.

Solution.

Step 1 — Sizes.
$N = abn = 4 \times 3 \times 5 = 60$ runs, so $df_T = 59$.

Step 2 — Effects.
$df_A = a - 1 = 3$; $df_B = b - 1 = 2$; $df_{AB} = (a-1)(b-1) = 3 \times 2 = 6$.

Step 3 — Error and check.
$df_E = ab(n-1) = 12 \times 4 = 48$. Check with (9.9): $3 + 2 + 6 + 48 = 59 = abn - 1$. The design leaves a healthy $\boxed{df_E = 48}$ degrees of freedom for estimating σ^2. A common error is to confuse df_E with $df_T = 59$, or to use $ab = 12$ instead of $(a-1)(b-1) = 6$ for the interaction.

Reference. Montgomery & Runger, 6th ed., Example 14-1 (structure).

9.6 A Complete Hand Computation

The next example carries Procedure 9.1 through a full data set, keeping every intermediate number visible. Work through it with a calculator; this computation is a standard exam task.

Example 9.3 — Perception-model failures: a 2×3 factorial

An AI safety team validates two versions of a perception model (A: version V1 and version V2) across three test-scenario types (B: clear weather, rain, night). For each cell, the team runs $n = 2$ independent test batches of 1,000 frames and records the number of missed detections per batch. The data, with all totals, are in Table 9.2. Test for interaction and main effects at $\alpha = 0.05$.

Model	**Clear**	**Rain**	**Night**	$y_{i\cdot\cdot}$
V1	8, 10	14, 16	21, 23	92
V2	7, 9	9, 11	20, 22	78
$y_{\cdot j\cdot}$	34	50	86	$y_{\cdots} = 170$

Table 9.2. Missed detections per 1,000-frame batch: model version $\times$ scenario type, $n = 2$ batches per cell. Cell totals are $y_{11\cdot} = 18$, $y_{12\cdot} = 30$, $y_{13\cdot} = 44$, $y_{21\cdot} = 16$, $y_{22\cdot} = 20$, $y_{23\cdot} = 42$.

Solution.

Step 1 — Totals.

Here $a = 2$, $b = 3$, $n = 2$, so $N = abn = 12$. The totals are shown in Table 9.2; both row totals ($92 + 78$) and column totals ($34 + 50 + 86$) add to $y_{\cdots} = 170$, as required.

Step 2 — Correction term and SS_T.

The correction term is $C = 170^2/12 = 28900/12 = 2408.33$. The squared observations sum to

$$\sum y_{ijk}^2 = 8^2 + 10^2 + 14^2 + \cdots + 20^2 + 22^2 = 2802,$$

so by (9.4), $SS_T = 2802 - 2408.33 = 393.67$.

Step 3 — Main-effect sums of squares.

Each row total contains $bn = 6$ observations and each column total $an = 4$.

By (9.5) and (9.6),

$$SS_A = \frac{92^2 + 78^2}{6} - C = \frac{14548}{6} - 2408.33 = 2424.67 - 2408.33 = 16.33,$$

$$SS_B = \frac{34^2 + 50^2 + 86^2}{4} - C = \frac{11052}{4} - 2408.33 = 2763.00 - 2408.33 = 354.67.$$

Step 4 — Interaction.

The six cell totals give, by (9.7),

$$\frac{1}{n}\sum_{i,j} y_{ij\cdot}^2 - C = \frac{18^2 + 30^2 + 44^2 + 16^2 + 20^2 + 42^2}{2} - 2408.33 = \frac{5580}{2} - 2408.33 = 381.67,$$

so $SS_{AB} = 381.67 - 16.33 - 354.67 = 10.67$.

Step 5 — Error.

By (9.8), $SS_E = 393.67 - 16.33 - 354.67 - 10.67 = 12.00$.

Step 6 — Degrees of freedom and mean squares.

$df_A = 1$, $df_B = 2$, $df_{AB} = 2$, $df_E = ab(n-1) = 6$; check $1 + 2 + 2 + 6 = 11 = N - 1$. The mean squares are $MS_A = 16.33$, $MS_B = 354.67/2 = 177.33$, $MS_{AB} = 10.67/2 = 5.333$, and $MS_E = 12.00/6 = 2.000$.

Steps 7–8 — F tests, interaction first.

	ν_1 (numerator df)				
ν_2	1	2	3	4	5
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97

Table 9.3. Excerpt of the $f_{0.05,\nu_1,\nu_2}$ table. The tests in Example 9.3 use the $\nu_2 = 6$ row: 5.99 for one numerator df and 5.14 for two.

By (9.10), $F_0^{AB} = 5.333/2.000 = 2.67$, compared with $f_{0.05,2,6} = 5.14$ (Table 9.3). Since $2.67 < 5.14$, we do not reject H_0^{AB}: no significant interaction. The main effects can therefore be read directly. By (9.11), $F_0^A = 16.33/2.000 = 8.17 > f_{0.05,1,6} = 5.99$ and $F_0^B = 177.33/2.000 = 88.67 > 5.14$, so both main effects are significant. Table 9.4 assembles the results.

Source	SS	df	MS	F_0	$f_{0.05}$
Model version (A)	16.33	1	16.33	8.17	5.99
Scenario type (B)	354.67	2	177.33	88.67	5.14
Interaction (AB)	10.67	2	5.333	2.67	5.14
Error	12.00	6	2.000		
Total	393.67	11			

Table 9.4. ANOVA for the perception-model study of Example 9.3.

Conclusion in problem language.

Scenario type dominates: night scenes produce far more missed detections than clear scenes for both versions ($\bar{y}_{\cdot 1\cdot} = 8.5$ vs. $\bar{y}_{\cdot 3\cdot} = 21.5$ misses per batch). Averaged over scenarios, version V2 misses significantly fewer detections than V1 ($\bar{y}_{2\cdot\cdot} = 13.0$ vs. $\bar{y}_{1\cdot\cdot} = 15.33$), and because the interaction is not significant, that advantage can be quoted as a single number. The answer is $\boxed{F_0^{AB} = 2.67 \text{ (n.s.)},\ F_0^{A} = 8.17,\ F_0^{B} = 88.67 \text{ (both significant)}}$.

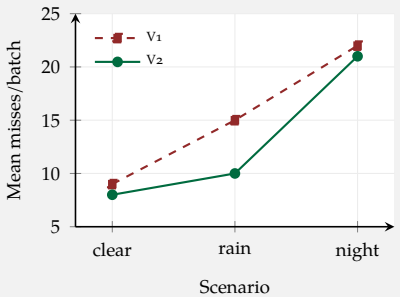

Figure 9.4. Cell means for Example 9.3. The lines are roughly parallel (the rain gap is within noise: $F_0^{AB} = 2.67$, not significant), so main effects summarize the study fairly.

Reference. Montgomery & Runger, 6th ed., Section 14-3 and Example 14-1.

9.7 When the Interaction Is Significant

In Example 9.3 the interaction was not significant, and the analysis could end with two clean main-effect sentences. The next example shows the other outcome, which changes how everything downstream must be reported.

Example 9.4 — Warehouse picking: layout × shift, 3×3 with $n = 4$

A fulfillment operations engineer studies the time to pick one order line (in seconds) under three warehouse layouts (A: layouts 1, 2, 3) and three shifts (B: morning, evening, night; order-volume patterns differ sharply by shift). Each cell is replicated $n = 4$ times, giving $N = 36$ timed picks. The full data are in Table 9.5. Analyze at $\alpha = 0.05$ and recommend a layout.

Layout	Morning	Evening	Night	$y_{i\cdot\cdot}$
1	130, 155, 74, 180	34, 40, 80, 75	20, 70, 82, 58	998
	$y_{11\cdot} = 539$	$y_{12\cdot} = 229$	$y_{13\cdot} = 230$	
2	150, 188, 159, 126	136, 122, 106, 115	25, 70, 58, 45	1300
	$y_{21\cdot} = 623$	$y_{22\cdot} = 479$	$y_{23\cdot} = 198$	
3	138, 110, 168, 160	174, 120, 150, 139	96, 104, 82, 60	1501
	$y_{31\cdot} = 576$	$y_{32\cdot} = 583$	$y_{33\cdot} = 342$	
$y_{\cdot j\cdot}$	1738	1291	770	$y_{\cdots} = 3799$

Table 9.5. Picking time (seconds per order line), 3×3 factorial with $n = 4$ replicates per cell, with cell, row, and column totals.

Solution.

Steps 1–2 — Totals, correction term, SS_T.

The correction term is $C = 3799^2/36 = 400900.03$. The 36 squared observations sum to $\sum y_{ijk}^2 = 478547$, so $SS_T = 478547 - 400900.03 = 77646.97$.

Step 3 — Main effects.

Rows contain $bn = 12$ observations each, columns $an = 12$:

$$SS_A = \frac{998^2 + 1300^2 + 1501^2}{12} - C = \frac{4939005}{12} - 400900.03 = 10683.72,$$

$$SS_B = \frac{1738^2 + 1291^2 + 770^2}{12} - C = \frac{5280225}{12} - 400900.03 = 39118.72.$$

Step 4 — Interaction.

The nine cell totals give

$$\frac{1}{4}\sum_{i,j} y_{ij\cdot}^2 - C = \frac{1841265}{4} - 400900.03 = 59416.22,$$

so $SS_{AB} = 59416.22 - 10683.72 - 39118.72 = 9613.78$.

Steps 5–6 — Error, df, mean squares.

$SS_E = 77646.97 - 59416.22 = 18230.75$, with $df_E = ab(n-1) = 9 \times 3 = 27$. The completed table is Table 9.6.

Source	SS	df	MS	F_0	$f_{0.05}$	Decision
Layout (A)	10683.72	2	5341.86	7.91	3.35	reject H_0^A
Shift (B)	39118.72	2	19559.36	28.97	3.35	reject H_0^B
Interaction (AB)	9613.78	4	2403.44	3.56	2.73	reject H_0^{AB}
Error	18230.75	27	675.21			
Total	77646.97	35				

Table 9.6. ANOVA for the picking study. All three effects are significant at $\alpha = 0.05$; the interaction test is read first.

Steps 7–8 — Tests.

The interaction comes first: $F_0^{AB} = 2403.44/675.21 = 3.56 > f_{0.05,4,27} = 2.73$, so the interaction is significant. Then $F_0^A = 7.91$ and $F_0^B = 28.97$, both above $f_{0.05,2,27} = 3.35$. All three hypotheses in (9.2) are rejected: layout matters, shift matters strongly, and *which layout is fastest depends on the shift*.

Follow-up — read the cell means, not the row means.

Because H_0^{AB} is rejected, the recommendation must come from the cell means $\bar{y}_{ij\cdot} = y_{ij\cdot}/4$ in Table 9.7, plotted in Figure 9.5.

Layout	Morning	Evening	Night
1	134.75	57.25	57.50
2	155.75	119.75	49.50
3	144.00	145.75	85.50

Table 9.7. Cell means (seconds per order line) for the picking study.

Reading the table column by column: in the morning shift the three layouts are roughly comparable (all congested at peak volume); in the evening, layout 1 is much faster (57.25 s) while layout 3 stays slow (145.75 s); at night, layout 2 is fastest (49.50 s) but layout 1 is close (57.50 s). No layout wins every shift — that is exactly what the significant interaction means. If a single layout must be chosen, the sensible choice is the robust one: $\boxed{\text{layout 1}}$, which is

fastest or within about 8 s of fastest in every shift, and has the lowest average time (83.17 s vs. 108.33 and 125.08). A recommendation built on row means alone would hide the fact that layout 2 beats it at night.

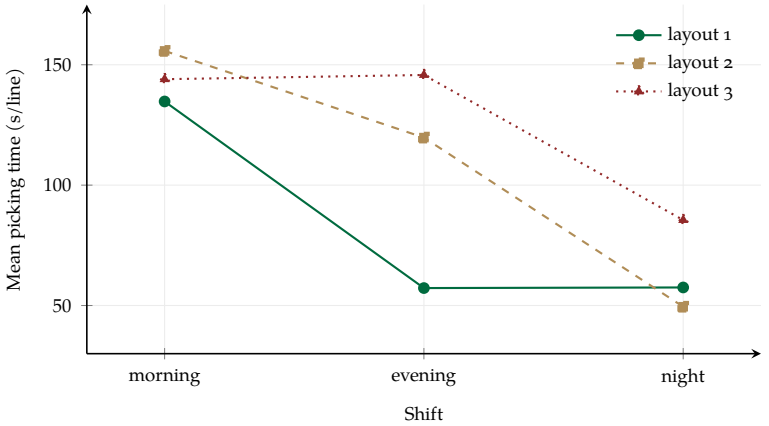

Figure 9.5. Interaction plot for the picking study (Example 9.4). The lines are far from parallel and the layout ranking changes across shifts: layout 1 leads in the evening, layout 2 at night. This is the picture behind $F_0^{AB} = 3.56$.

Engineering judgment. "Choose the robust cell" is an engineering decision rule, not a statistical one. If night-shift throughput is the operational bottleneck, the engineer may rightly pick layout 2 despite its poor evening numbers. The statistics establishes that the trade-off is real; the engineering context decides how to make it.

Two practical remarks complete the picture. First, when the interaction is *not* significant, follow-up comparisons among the levels of a significant main effect proceed exactly as in Chapter 8: Fisher LSD or Tukey comparisons applied to the row (or column) means, using MS_E from the factorial table with $ab(n-1)$ degrees of freedom. Second, when the interaction *is* significant, run those comparisons within a fixed level of the other factor — compare layouts within the night shift, for example — because averaged comparisons no longer describe any real operating condition.

Lecture 26 starts here — *Factorial diagnostics and applications*

Lecture 26. This section is covered by Lecture 26.

Reference. Montgomery & Runger, 6th ed., Section 14-3.2.

9.8 Residual Analysis and Model Adequacy

Every conclusion in this chapter rests on the model (9.1): independent normal errors with constant variance σ^2 in every cell. Before trusting the F tests, check

the residuals. The fitted value for any observation in a factorial model is simply its cell mean, so the residual is

$$e_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}. \tag{9.12}$$

Residual. The difference between an observation and its fitted value; in a factorial model, the deviation from its own cell mean.

Procedure 9.2 — Diagnostics for a factorial experiment.

1. Compute the fitted value of every observation (its cell mean $\bar{y}_{ij\cdot}$) and the residual e_{ijk} by (9.12). The residuals must sum to zero within every cell; use this as an arithmetic check.
2. Make a normal probability plot of all abn residuals. Points close to a straight line support the normality assumption; strong curvature or heavy tails (extreme points peeling away at both ends) do not.
3. Plot residuals against fitted values. The vertical spread should be roughly constant across the range. A funnel that widens with the fitted value indicates non-constant variance; consider transforming the response ($\log y$ or $\sqrt{y}$) and re-analyzing.
4. Plot residuals against the levels of each factor. The spread should be similar at every level; one level with visibly larger scatter means that factor also affects the *variance*, not just the mean.
5. Investigate any isolated large residual (a potential outlier): check the data record before deleting anything.
6. Read the interaction plot of cell means alongside the ANOVA: parallel lines with a significant F_0^{AB}, or wildly non-parallel lines with an insignificant one, usually means the error variance is large and the experiment is underpowered — interpret with care.

Example 9.5 — Residual check for the picking study

Check the model adequacy for Example 9.4.

Solution.

Step 1 — Residuals.

Fitted values are the cell means of Table 9.7. For the first cell (layout 1, morning), $\bar{y}_{11\cdot} = 134.75$, so the four residuals are $130 - 134.75 = -4.75$, $155 - 134.75 = 20.25$, $74 - 134.75 = -60.75$, and $180 - 134.75 = 45.25$; they sum to zero as required. Repeating for all nine cells gives 36 residuals ranging from -60.75 to 45.25.

Steps 2–4 — The four panels.

Figure 9.6 shows the diagnostic panels. The normal probability plot is acceptably straight, with the two largest residuals (-60.75 and 45.25, both from the layout-1 morning cell) sitting at the ends but not far off the line. The residual-versus-fitted plot shows no funnel. The plots against layout and against shift show mildly larger spread for layout 1 and for the morning shift — both driven by the same noisy cell — but nothing severe enough to invalidate the analysis.

Conclusion.

The assumptions are adequately met; the ANOVA conclusions of Table 9.6 stand. The layout-1 morning cell would be the first place to look if the study were repeated: its within-cell range (74 to 180 s) suggests uneven congestion during the morning peak. Model adequate; no transformation needed.

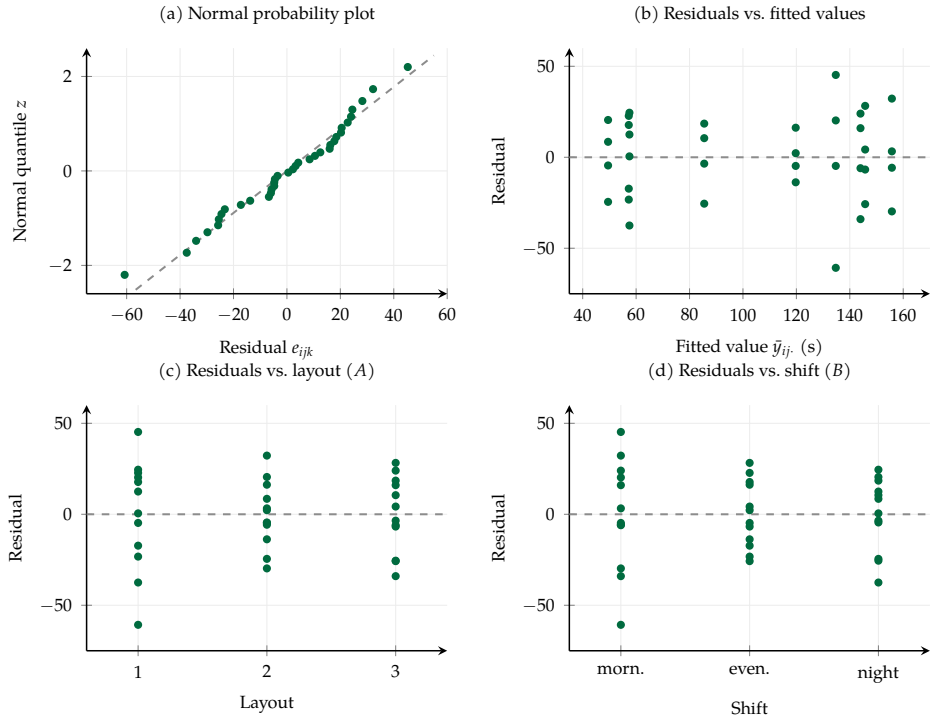

Figure 9.6. Residual diagnostics for the picking study, following Procedure 9.2. (a) The normal plot is acceptably straight. (b) No funnel against fitted values. (c), (d) Spread is broadly similar across factor levels, with the layout-1 morning cell contributing the largest residuals.

Reference. Montgomery & Runger, 6th ed., Section 14-3.3.

9.9 *One Observation per Cell*

The error degrees of freedom, $ab(n-1)$, are supplied entirely by replication. If the experiment runs only $n = 1$ observation per cell, then $df_E = 0$: there are no within-cell repeats, so the data provide no estimate of σ^2 that is free of the model effects. The cell ''mean'' is the observation itself, every residual in (9.12) is zero, and the interaction sum of squares and the error sum of squares collapse into one quantity that cannot be separated. No F test of the interaction is possible.

The standard workaround is to *assume* the interaction away. If $(\tau\beta)_{ij} = 0$ for all i, j, the model becomes additive,

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

which is exactly the structure of the randomized complete block design of Chapter 8, and the former interaction line, with its $(a-1)(b-1)$ degrees of freedom, becomes the error line. Both main effects can then be tested. The assumption is not free: if a real interaction exists, it is silently folded into the error term, inflating MS_E and making the main-effect tests conservative at best and misleading at worst. A diagnostic exists for this situation — Tukey's single-degree-of-freedom test of nonadditivity, which isolates one degree of freedom from the residual to test for a multiplicative form of interaction. If that test rejects, do not trust the additive analysis; transform the response or replicate the experiment.

Insight. Why replication is cheap insurance. Going from $n = 1$ to $n = 2$ doubles the runs, but it buys three things at once: an honest estimate of σ^2, a valid test of the interaction, and residuals that can actually be plotted. In most engineering studies the cost of being wrong about an interaction — shipping the wrong model version, committing to the wrong layout — is far larger than the cost of the extra runs. Whenever feasible, run $n \geq 2$.

Reference. Montgomery & Runger, 6th ed., Section 14-4.

9.10 *General Factorial Experiments*

Nothing in the factorial idea stops at two factors. With three factors A, B, C at a, b, c levels and n replicates, the design has $N = abcn$ runs, and the ANOVA partitions the variability into *seven* effect sources plus error: three main effects (A, B, C), three two-factor interactions (AB, AC, BC), and one three-factor interaction (ABC). The degrees of freedom follow the same construction rules as before, shown in Table 9.8. Every F statistic again uses MS_E in the denominator, and interactions are examined before the main effects they involve.

Source	df	Source	df
A	$a-1$	AB	$(a-1)(b-1)$
B	$b-1$	AC	$(a-1)(c-1)$
C	$c-1$	BC	$(b-1)(c-1)$
ABC	$(a-1)(b-1)(c-1)$	Error	$abc(n-1)$
Total			$abcn-1$

Table 9.8. Degrees of freedom for a three-factor factorial with n replicates. Each interaction df is the product of the main-effect df of the factors it involves.

As a check, take $a = 3$, $b = 2$, $c = 4$, $n = 2$: then $df_{ABC} = 2 \times 1 \times 3 = 6$ and $df_E = abc(n-1) = 24 \times 1 = 24$, and the total is $abcn - 1 = 47$. Notice the practical problem, however: the run count grows *multiplicatively*. Five factors at three levels each already demand $3^5 = 243$ cells before any replication. This cost is what drives industrial experimentation to the special case of the next section, in which every factor is restricted to exactly two levels.

Reference. Montgomery & Runger, 6th ed., Section 14-5.

9.11 *Two-Level Designs: The 2^k Workhorse*

2^k design. A factorial design with k factors, each at exactly two levels, written $-$ (low) and $+$ (high); it has 2^k treatment combinations.

A 2^k design studies k factors, each at only two levels, coded low $(-)$ and high $(+)$. Two levels per factor is the minimum that still detects which factors move the response and which factors interact, and it keeps the run count manageable: $2^5 = 32$ combinations instead of $3^5 = 243$. The price is that two levels cannot detect curvature between them; center points, discussed in Section 9.12, repair that blind spot cheaply. Because of this balance of information against cost, 2^k designs are the workhorse of engineering experimentation — screening process variables, tuning simulation parameters, structuring red-team test campaigns, and powering multivariate product tests.

The 2^2 design

Start with $k = 2$: factors A and B, each at two levels, four cells, n replicates each, $N = 4n$ runs. Each treatment combination has a standard label, and the label also names the *total* of the n replicates observed there (Table 9.9).

Label	A	B	**Meaning**
(1)	$-$	$-$	both factors low
a	$+$	$-$	A high, B low
b	$-$	$+$	A low, B high
ab	$+$	$+$	both factors high

Table 9.9. Treatment-combination notation for the 2^2 design. A lowercase letter appears when that factor is high; (1) means no factor is high. Each label also denotes the total of the n replicates at that combination.

The effect of A is the average response when A is high minus the average when A is low. Reading Table 9.9, A is high in a and ab (a total of $2n$ observations) and

Contrast. A signed sum of treatment totals; each effect in a 2^k design is its contrast divided by $n\,2^{k-1}$.

low in (1) and b, so

$$\text{Effect}_A = \frac{a + ab - (1) - b}{2n}, \quad \text{Effect}_B = \frac{b + ab - (1) - a}{2n}, \quad \text{Effect}_{AB} = \frac{(1) + ab - a - b}{2n}. \tag{9.13}$$

The numerator of each expression — a signed sum of the four treatment totals — is called the *contrast* of that effect. The interaction contrast $(1) + ab - a - b$ compares the A effect at high B with the A effect at low B, which is exactly the ''non-parallelism'' the interaction plot displays. Each effect carries one degree of freedom, and its sum of squares comes from one formula used three times:

$$SS_{\text{effect}} = \frac{(\text{Contrast})^2}{4n}, \qquad df = 1. \tag{9.14}$$

This is nothing new: (9.14) is the two-factor machinery of Section 9.4 simplified by the ± 1 coding. SS_T is still computed from (9.4), and $SS_E = SS_T - SS_A - SS_B - SS_{AB}$ with $df_E = N - 4 = 4(n-1)$.

Example 9.6 — A 2^2 growth experiment: pricing page $\times$ onboarding flow

A startup tests two versions of its pricing page (A: current $-$, redesigned $+$) and two onboarding flows (B: current $-$, guided $+$) on signup conversion. Each of the four combinations is shown to $n = 2$ independent weekly cohorts of 200 visitors, and the response is the conversion rate in percent. The treatment totals (each the sum of two weekly rates) are

$$(1) = 60, \qquad a = 76, \qquad b = 48, \qquad ab = 72,$$

and the total sum of squares is $SS_T = 248$. Analyze at $\alpha = 0.05$.

Solution.

Step 1 — Contrasts and effects.

By (9.13) with $2n = 4$:

$$\text{Contrast}_A = a + ab - (1) - b = 76 + 72 - 60 - 48 = 40, \qquad \text{Effect}_A = 40/4 = 10.0,$$
$$\text{Contrast}_B = b + ab - (1) - a = 48 + 72 - 60 - 76 = -16, \qquad \text{Effect}_B = -16/4 = -4.0,$$
$$\text{Contrast}_{AB} = (1) + ab - a - b = 60 + 72 - 76 - 48 = 8, \qquad \text{Effect}_{AB} = 8/4 = 2.0.$$

The redesigned pricing page raises conversion by about 10 percentage points

on average; the guided onboarding *lowers* it by about 4; the interaction estimate is small.

Step 2 — Sums of squares.

By (9.14) with $4n = 8$:

$$SS_A = \frac{40^2}{8} = 200, \qquad SS_B = \frac{(-16)^2}{8} = 32, \qquad SS_{AB} = \frac{8^2}{8} = 8.$$

Step 3 — Error.

$SS_E = 248 - 200 - 32 - 8 = 8$, with $df_E = N - 2^k = 8 - 4 = 4$, so $MS_E = 8/4 = 2$.

Step 4 — F tests.

Each effect has one degree of freedom, so $MS = SS$ for each. Against $f_{0.05,1,4} = 7.71$:

$$F_0^{AB} = 8/2 = 4.0 < 7.71 \text{ (n.s.)}, \qquad F_0^A = 200/2 = 100 > 7.71, \qquad F_0^B = 32/2 = 16 > 7.71.$$

The interaction is not significant, so the two page elements act additively; both main effects are significant. The recommendation is the redesigned pricing page with the *current* onboarding flow, worth about $10 + 4 = 14$ percentage points of conversion over the worst combination: deploy *a*: redesigned pricing, current onboarding.

Startup lens. Example 9.6 used 8 cohort-weeks to answer three questions at once: the pricing effect, the onboarding effect, and whether they interfere. Two sequential A/B tests of the same precision would have used the same traffic and left the interaction question unanswered — and it was only luck that the interaction here was negligible.

The general 2^k design and the sign table

For k factors the same logic scales up mechanically. There are $N = n\,2^k$ runs and $2^k - 1$ effects (for $k = 3$: three main effects, three two-factor interactions, one three-factor interaction). Two formulas replace all the special cases:

$$\text{Effect} = \frac{\text{Contrast}}{n\,2^{k-1}}, \tag{9.15}$$

$$SS_{\text{effect}} = \frac{(\text{Contrast})^2}{n\,2^k}, \qquad df = 1. \tag{9.16}$$

For $k = 2$ the denominators reduce to $2n$ and $4n$, matching (9.13) and (9.14).

The contrasts themselves come from the *sign table* (Table 9.10 shows the 2^3 case). The main-effect columns A, B, C record whether each factor is high or low

Sign table. The table of $\pm$ signs, one row per treatment combination and one column per effect, from which every contrast in a 2^k design is read off.

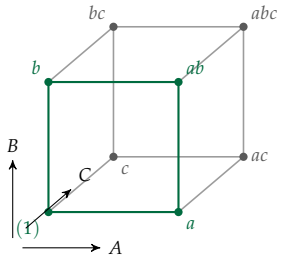

Figure 9.7. The 2^3 design as the corners of a cube: each axis is one factor, and each treatment label marks which factors are high.

in that run. Each interaction column is the entrywise product of its parents: the AB column is $A \times B$, the ABC column is $A \times B \times C$. To build any contrast, read down the effect's column, attach each sign to the matching treatment total, and add.

Run	A	B	C	AB	AC	BC	ABC
(1)	$-$	$-$	$-$	$+$	$+$	$+$	$-$
a	$+$	$-$	$-$	$-$	$-$	$+$	$+$
b	$-$	$+$	$-$	$-$	$+$	$-$	$+$
ab	$+$	$+$	$-$	$+$	$-$	$-$	$-$
c	$-$	$-$	$+$	$+$	$-$	$-$	$+$
ac	$+$	$-$	$+$	$-$	$+$	$-$	$-$
bc	$-$	$+$	$+$	$-$	$-$	$+$	$-$
abc	$+$	$+$	$+$	$+$	$+$	$+$	$+$

Table 9.10. The sign table for the 2^3 design. Main-effect columns code the factor levels; each interaction column is the entrywise product of its parent columns. Reading down the AC column, for example, gives $\text{Contrast}_{AC} = (1) - a + b - ab - c + ac - bc + abc$.

Procedure 9.3 — Analyzing a 2^k factorial.

1. Write the sign table for the 2^k treatment combinations (Table 9.10 for $k = 3$), and record the treatment totals.
2. For each of the $2^k - 1$ effects, multiply the column signs by the treatment totals and sum to obtain the contrast.
3. Compute each effect estimate from (9.15) and each sum of squares from (9.16). The denominator of the SS is $n\,2^k$ for every effect — a common exam error is to keep using $4n$ when $k > 2$.
4. Compute SS_T from the individual observations by (9.4), then $SS_E = SS_T - \sum SS_{\text{effect}}$ with $df_E = N - 2^k$.
5. Build the ANOVA table: every effect has $df = 1$, so $MS_{\text{effect}} = SS_{\text{effect}}$, and each $F_0 = MS_{\text{effect}}/MS_E$ is tested against $f_{\alpha,1,N-2^k}$. Read interaction

results before main-effect results, as always.

Example 9.7 — A 2^3 study of truck fuel savings

A fleet engineer studies the percentage fuel saving of delivery trucks under three two-level factors: A, aerodynamic fairing (absent, fitted); B, low-rolling-resistance tires (standard, fitted); and C, speed governor (off, on). Each of the 8 configurations is run on $n = 2$ matched routes, $N = 16$ runs. The treatment totals (sums of two replicate savings, in percent) are

$$(1) = 3, \quad a = 17, \quad b = 9, \quad ab = 27, \quad c = 7, \quad ac = 21, \quad bc = 13, \quad abc = 31,$$

and $SS_T = 356$. Analyze at $\alpha = 0.05$.

Solution.

Steps 1–2 — Contrasts from the sign table.

Reading down the A column of Table 9.10:

$$\text{Contrast}_A = -(1) + a - b + ab - c + ac - bc + abc = -3 + 17 - 9 + 27 - 7 + 21 - 13 + 31 = 64.$$

Repeating for the other six columns gives the contrasts in the table below.

Step 3 — Effects and sums of squares.

With $n\,2^{k-1} = 8$ and $n\,2^k = 16$, (9.15) and (9.16) give:

Effect	A	B	C	AB	AC	BC	ABC
Contrast	64	32	16	8	0	0	0
Effect estimate	8.0	4.0	2.0	1.0	0	0	0
$SS = \text{Contrast}^2/16$	256	64	16	4	0	0	0

The fairing is worth about 8 percentage points of saving on average, the tires about 4, the governor about 2.

Step 4 — Error.

$SS_E = 356 - (256 + 64 + 16 + 4 + 0 + 0 + 0) = 16$, with $df_E = N - 2^k = 16 - 8 = 8$ and $MS_E = 2$.

Step 5 — F tests.
Each effect has $df = 1$, so $F_0 = SS/2$, tested against $f_{0.05,1,8} = 5.32$:

$$F_0^A = 128, \quad F_0^B = 32, \quad F_0^C = 8 \quad (\text{all} > 5.32); \qquad F_0^{AB} = 2, \quad F_0^{AC} = F_0^{BC} = F_0^{ABC} = 0.$$

All three main effects are significant and no interaction is. The three measures act independently and their savings add: fitting all three is expected to save about $8 + 4 + 2 = 14\%$ of fuel. The answer is all three main effects significant; no interactions

Engineering judgment. The additivity found in Example 9.7 is itself engineering information: fairing, tires, and governor attack different loss mechanisms (drag, rolling resistance, speed profile), so their savings stack. Had AB been large, it would have pointed to a shared mechanism — physics you can go look for.

Reference. Montgomery & Runger, 6th ed., Sections 14-5 to 14-7.

Sparsity of effects. The empirical principle that in most real systems only a few main effects and low-order interactions are large; most high-order interactions are negligible.

9.12 *Single Replicates, Center Points, Blocking, and Fractions*

Four practical extensions make 2^k designs usable at industrial scale. Each solves a resource problem, and each pays for the saving with a stated assumption. Knowing what is paid is the engineering skill.

Single replicates and the normal plot of effects

For $k = 5$, one replicate is already 32 runs; $n = 2$ doubles that. Experimenters therefore often run $n = 1$ — and then $df_E = N - 2^k = 0$, so no F test is possible, the same wall as in Section 9.9. The escape uses the *sparsity-of-effects principle*: in real systems, only a few effects, mostly main effects and two-factor interactions, are genuinely large; the rest behave like noise. If the negligible effects are noise, they are approximately normal with mean zero, so plot all $2^k - 1$ effect estimates on a normal probability plot: the inactive effects line up on a straight line through zero, and the active effects fall visibly off the line (Figure 9.8). The effects on the line are then pooled into an error estimate, and the analysis proceeds.

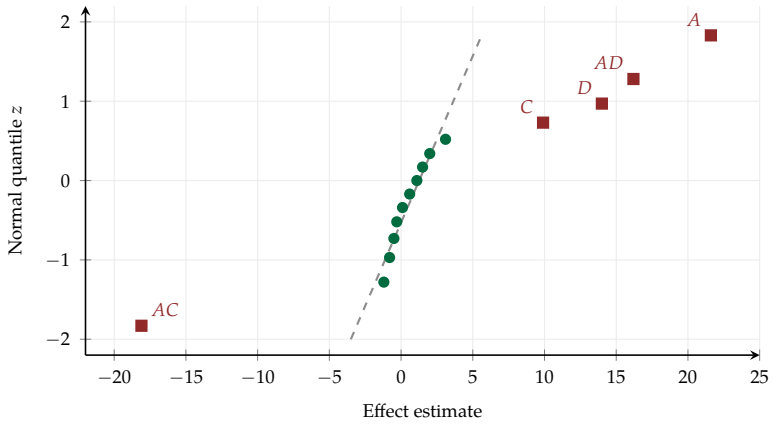

Figure 9.8. Normal probability plot of the 15 effect estimates from a single-replicate 2^4 experiment. The ten effects on the dashed line (green circles) behave like noise and are pooled as error; the five that depart from it (A, D, AD, C, AC; maroon squares) are the active effects. Note that the active set contains interactions, which no OFAT program would have found.

Center point. A run with every factor set at the midpoint of its low and high levels, added to a 2^k design to check for curvature and to estimate pure error.

Center points: a curvature check

A 2^k design observes only the corners of a k-dimensional box, so a straight-line (plus interactions) model is all it can fit; it is blind to curvature between the corners. The inexpensive fix is to add n_C runs (typically 3 to 5) at the *center point*, where every factor sits at the midpoint of its two levels. The center runs are genuine replicates of one condition, so they provide a pure-error estimate of σ^2 without disturbing the balance of the factorial runs, and they enable a curvature check: compare the average corner response $\bar{y}_F$ with the average center response $\bar{y}_C$. If $|\bar{y}_F - \bar{y}_C|$ is small, the linear picture is adequate over the region; if it is large, the surface bends, and a follow-up design with more than two levels per factor is needed. Please do not confuse center points with replicates of the corners — they sit at the midpoint, a condition the factorial itself never visits.

Confounding. Deliberately arranging a design so that a chosen effect cannot be distinguished from a nuisance (block) effect.

Blocking and confounding

Sometimes the 2^k runs cannot all be performed under uniform conditions: one shift, one batch of material, one day of stable traffic. Splitting the runs into blocks controls that nuisance variation — the same motive as the randomized block design of Chapter 8 — but in a 2^k design the split has a price: one effect becomes *confounded* with the blocks, meaning its contrast and the block contrast are the

same signed sum, so the two cannot be told apart. The standard choice is to sacrifice the highest-order interaction, the effect least likely to be real. For a 2^3 in two blocks of four, confound ABC: block 1 takes the runs with a $-$ sign in the ABC column of Table 9.10, namely $(1), ab, ac, bc$, and block 2 takes a, b, c, abc. All three main effects and all three two-factor interactions remain cleanly estimable; only ABC is lost to the block difference.

Aliasing. In a fractional factorial, the sharing of one contrast by two or more effects, which the data cannot then separate.

Fractional replication: 2^{k-p} designs

When k is large, even a single replicate is expensive: $2^7 = 128$ runs. By sparsity of effects, most of those runs are buying estimates of high-order interactions that are almost surely negligible. A *fractional* design runs only a chosen fraction of the combinations. The half-fraction 2^{k-1} keeps the runs on which a chosen *generator* column is $+$; for example, a 2^{5-1} with defining relation $I = ABCDE$ runs the 16 combinations (out of 32) with a plus sign in the $ABCDE$ column. The cost is *aliasing*: in the fraction, every effect shares its column of signs with a partner (multiply the effect by the defining relation to find it), so $A = BCDE$, $AB = CDE$, and so on. The data estimate the alias pair's sum, and only engineering judgment — usually sparsity — says which member to credit. Smaller fractions 2^{k-p} use p generators to cut the run count by 2^p, at the price of denser aliasing. The standard industrial workflow is: screen many factors with a small fraction, identify the two or three active ones, then follow up with a full factorial (or a design with center points) on those alone.

Example 9.8 — Choosing a screening design under a run budget

A red team must screen $k = 6$ binary configuration factors of an LLM deployment (safety fine-tune on/off, system-prompt hardening, output filter, retrieval guard, temperature cap, tool-use gate) for their effect on the jailbreak success rate. The evaluation pipeline can afford 16 runs. Which design fits, and what does it give up?

Solution.

Step 1 — Rule out the full designs.

A full 2^6 needs 64 runs; a half-fraction 2^{6-1} needs 32. Both exceed the budget.

Step 2 — Size the fraction.

A quarter-fraction 2^{6-2} has $2^4 = 16$ runs — exactly the budget. It uses two generators, so each effect is aliased with three partners.

Step 3 — State the price.

With standard generators (a resolution IV choice), main effects are aliased only with three-factor and higher interactions, so under sparsity the six main effects are interpretable; two-factor interactions, however, are aliased with each other in groups and cannot be separated by these 16 runs alone.

Conclusion.

Run the 2^{6-2}, identify the few active factors from a normal plot of effects, then spend the next testing budget on a full factorial in those factors only. The answer is $\boxed{2^{6-2} \text{ quarter-fraction, 16 runs, resolution IV}}$. Note what would be wrong with the alternatives: OFAT (12 runs, six factors two levels each) estimates no interactions at all, and a one-way comparison of six ''configurations'' ignores the factorial structure entirely.

Safety lens. Guardrail configurations multiply fast: six binary switches already define 64 deployment variants, and no evaluation budget tests them all with replication. Fractional factorials are how a finite red-team budget still yields defensible statements about which switches matter — with the aliasing assumptions stated in the report, where a reviewer can challenge them.

Reference. Montgomery & Runger, 6th ed., Sections 14-1 and 14-5 to 14-7.

9.13 *Factorial Thinking in Engineering Practice*

The machinery of this chapter appears in practice less often as a single grand experiment and more often as a way of *structuring* studies that would otherwise be run one factor at a time. Three patterns from the course's recurring settings show what to look for.

AI safety testing. An evaluation campaign is a factorial the moment it crosses model versions with test conditions: version × scenario family (Example 9.3), guardrail × language, fine-tune × prompt-attack class. The interaction line of the ANOVA is where regression risk lives — a new version that improves the average while getting worse on one scenario family shows up as a version-by-scenario interaction, not as a main effect. Evaluation reports that quote only per-version averages are reporting main effects; ask to see the cell means. When the number of configuration switches grows, fractional designs (Example 9.8) keep the test matrix affordable, and the aliasing structure belongs in the test plan.

Supply chains and transportation. Operational pilots almost always span heterogeneous conditions: layouts across shifts (Example 9.4), routing policies across

terminals, loading procedures across cargo types. Treating the condition as a second factor, rather than averaging over it, is what separates a defensible pilot from an anecdote. When conditions are nuisance rather than of interest — day of week, vessel arrival pattern — blocking enters, and in two-level pilots the block is confounded with a chosen high-order interaction. The interaction between policy and site is often the finding that matters most, because it tells the planner where a policy will and will not transfer.

Building your own startup. Growth experimentation lives on this chapter. Sequential A/B testing is OFAT (Section 9.1); multivariate testing is a factorial, and Example 9.6 is its smallest honest unit. The practical constraint is traffic: a 2^3 multivariate test splits visitors eight ways, so effects must be large or cohorts long. Fractions and single replicates with a normal plot of effects are the standard compromises, and center points have an analog whenever a factor is really continuous (price, trial length) — test the midpoint before believing a straight-line story between two endpoints.

Finally, two cautions that apply in every setting. First, a statistically significant effect is not automatically worth acting on: with enough data, an effect of 0.05 units will clear the F test even when the engineering tolerance is ± 2 units. Compare effect estimates — (9.15) gives them in the response's own units — against practical thresholds, not only against critical values. Second, the most common factorial mistakes on exams and in practice are the same four: using $4n$ in the SS denominator when $k > 2$ (it is $n\,2^k$); mis-signing a contrast (recheck the sign-table column); forgetting that $n = 1$ leaves $df_E = 0$; and reading center points as corner replicates. Slowing down at exactly these four points prevents most wrong answers.

Chapter Summary

A factorial experiment runs all combinations of factor levels, which lets it estimate each main effect from every observation and — unlike one-factor-at-a-time experimentation — detect interactions, where the effect of one factor depends on the level of another. The two-factor effects model (9.1) splits each observation into an overall mean, two main effects, an interaction, and error, and the total variability splits accordingly into four sums of squares (9.3), computed from row, column, and cell totals by (9.4)–(9.8). Three F statistics share the denominator MS_E; the interaction (9.10) is tested before the main effects (9.11), because a significant

interaction changes what the main effects mean: conclusions must then be drawn from cell means, level by level. Model adequacy is checked through the residuals $e_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}$ (Procedure 9.2). Replication supplies the error degrees of freedom $ab(n-1)$; with $n = 1$ the interaction cannot be tested without assuming additivity. With many factors the 2^k design is the workhorse: contrasts from the sign table give every effect and sum of squares through (9.15) and (9.16), single replicates are analyzed with a normal plot of effects under the sparsity-of-effects principle, center points check curvature, blocking confounds a chosen high-order interaction, and fractional designs 2^{k-p} trade aliasing for run count.

Quantity	Formula	Equation
Two-factor model	$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$	(9.1)
Decomposition	$SS_T = SS_A + SS_B + SS_{AB} + SS_E$	(9.3)
Total SS	$SS_T = \sum y_{ijk}^2 - y_{\cdots}^2/abn$	(9.4)
Factor A SS	$SS_A = \sum_i y_{i\cdot\cdot}^2/bn - y_{\cdots}^2/abn$	(9.5)
Factor B SS	$SS_B = \sum_j y_{\cdot j\cdot}^2/an - y_{\cdots}^2/abn$	(9.6)
Interaction SS	$SS_{AB} = \frac{1}{n}\sum_{i,j} y_{ij\cdot}^2 - \frac{y_{\cdots}^2}{abn} - SS_A - SS_B$	(9.7)
Error SS	$SS_E = SS_T - SS_A - SS_B - SS_{AB}$	(9.8)
Interaction F	$F_0^{AB} = MS_{AB}/MS_E$, df $(a-1)(b-1), ab(n-1)$	(9.10)
Main-effect F	$F_0^A = MS_A/MS_E,\ F_0^B = MS_B/MS_E$	(9.11)
Residual	$e_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}$	(9.12)
2^k effect	$\text{Effect} = \text{Contrast}/(n\,2^{k-1})$	(9.15)
2^k effect SS	$SS = (\text{Contrast})^2/(n\,2^k),\ df_E = N - 2^k$	(9.16)

Table 9.11. Key formulas of Chapter 9 at a glance.

Exercises

Reading interaction plots

Exercise 9.1. A startup measures weekly feature-engagement scores in a 2×3 factorial: notification style (A_1 digest, A_2 real-time) crossed with plan tier (free, pro, team). The cell means are:

	Free	Pro	Team
A_1 digest	42	55	63
A_2 real-time	40	44	47

Sketch or imagine the interaction plot (engagement vs. tier, one line per style). Which description fits best?

(A) Parallel lines; the factors act additively.

(B) Diverging lines that do not cross: engagement rises with tier for both styles, but much faster for the digest style — a moderate interaction.

(C) Crossing lines: the best style reverses across tiers.

(D) Flat lines; neither factor matters.

Exercise 9.2. An AI safety team compares guardrail versions G1 and G2 on the false-positive rate (benign prompts wrongly blocked, in percent) for two prompt languages. The cell means are: English — G1: 6, G2: 2; Arabic — G1: 3, G2: 7. The ANOVA finds the guardrail main effect not significant but the interaction highly significant. A teammate concludes: ''Guardrail version does not matter; the main effect is near zero.'' Explain, using the cell means, why the conclusion is wrong and what the correct statement is.

Two-factor computation

Exercise 9.3. A warehouse compares two picker teams (A) across four shifts (B) on the number of mispicks per 1,000 order lines, with $n = 3$ replicate days per cell. The sums of squares are $SS_A = 40$, $SS_B = 30$, $SS_{AB} = 22.5$, and $SS_E = 80$. Complete the ANOVA table (df, MS, F_0) and state each conclusion at $\alpha = 0.05$, given $f_{0.05,1,16} = 4.49$ and $f_{0.05,3,16} = 3.24$.

Exercise 9.4. An LLM evaluation crosses two prompt templates (A: T1, T2) with two model sizes (B: small, large) and scores each run out of 20, with $n = 3$ replicate evaluation seeds per cell:

	Small	**Large**	$y_{i\cdot\cdot}$
T1	9, 10, 11	13, 14, 15	72
T2	11, 12, 13	11, 12, 13	72
$y_{\cdot j\cdot}$	66	78	$y_{\cdots} = 144$

Given $\sum y_{ijk}^2 = 1760$ and $f_{0.05,1,8} = 5.32$, carry out the full ANOVA at $\alpha = 0.05$ following Procedure 9.1, and explain the seemingly strange outcome for factor A.

Exercise 9.5. A port authority plans a three-factor study of container dwell time: crane type ($a = 3$), yard policy ($b = 2$), and vessel size class ($c = 4$), with $n = 2$ replicate vessel calls per combination. The degrees of freedom for the ABC interaction and for error are:

(A) $df_{ABC} = 9$ and $df_E = 24$
(B) $df_{ABC} = 6$ and $df_E = 24$
(C) $df_{ABC} = 9$ and $df_E = 47$
(D) $df_{ABC} = 24$ and $df_E = 24$

Inference and diagnostics

Exercise 9.6. For the evaluation data of Exercise 9.4: (a) write the fitted value for every observation in the (T1, large) cell and compute its three residuals; (b) verify they sum to zero; (c) name the two plots from Procedure 9.2 you would draw first for the full data set and state what each checks.

Exercise 9.7. A startup runs a 2^2 test with A = checkout design (current $-$, new $+$) and B = free-trial length (14 days $-$, 30 days $+$), with $n = 2$ replicate cohorts of 200 visitors each; the response is conversions per cohort. The treatment totals are $(1) = 52$, $a = 68$, $b = 60$, $ab = 84$, and $SS_T = 292$. Using $f_{0.05,1,4} = 7.71$: (a) compute the three contrasts and effect estimates; (b) compute the sums of squares, SS_E, and MS_E; (c) test all three effects at $\alpha = 0.05$ and state the recommendation.

Exercise 9.8. A red team can afford a half-fraction 2^{5-1} (16 runs) to screen five binary safety-configuration factors A, B, C, D, E, using the defining relation $I = ABCDE$. (a) Find the alias of the main effect A and of the interaction AB. (b) In

one or two sentences, explain why this aliasing is usually acceptable in early-stage screening.

Exam-style problems

Exercise 9.9. A regional carrier studies delivery lateness (minutes) for three depots (A: D1, D2, D3) under two delivery time windows (B: W1 morning, W2 afternoon), with $n = 2$ replicate route-days per cell:

Depot	**W1**	**W2**	$y_{i\cdot\cdot}$
D1	6, 8	10, 12	36
D2	5, 7	15, 17	44
D3	9, 11	13, 15	48
$y_{\cdot j\cdot}$	46	82	$y_{\cdots} = 128$

Given $\sum y_{ijk}^2 = 1528$, $f_{0.05,2,6} = 5.14$, and $f_{0.05,1,6} = 5.99$:

1. **[4 pt]** Compute SS_T, SS_A, SS_B, SS_{AB}, and SS_E, showing the formula used for each.
2. **[3 pt]** Build the complete ANOVA table and carry out the three F tests at $\alpha = 0.05$, in the correct order.
3. **[3 pt]** Compute the six cell means and state, in plain language, what the significant effects mean for depot operations. Which depot should receive the afternoon congestion-mitigation pilot, and why?

Exercise 9.10. An AI safety group runs a 2^3 factorial with $n = 2$ replicate evaluation suites on the number of jailbreak prompts blocked (out of 20): A = safety fine-tune, B = system-prompt hardening, C = output filter (each: absent $-$, present $+$). Treatment totals (sums of two suites):

$$(1) = 8, \quad a = 18, \quad b = 12, \quad ab = 26, \quad c = 10, \quad ac = 22, \quad bc = 14, \quad abc = 30,$$

with $SS_T = 227$ and $f_{0.05,1,8} = 5.32$.

1. **[3 pt]** Using the sign table (Table 9.10), compute the contrasts for A, B, C, and AB.

2. **[3 pt]** Compute all seven effect estimates and sums of squares. (The remaining contrasts are $\text{Contrast}_{AC} = 4$, $\text{Contrast}_{BC} = 0$, $\text{Contrast}_{ABC} = 0$.)
3. **[2 pt]** Find SS_E, df_E, and MS_E.
4. **[2 pt]** Test all effects at $\alpha = 0.05$ and state the conclusion in the language of the deployment decision.

Assessing outside recommendations

Exercise 9.11. A startup's AI analytics assistant summarizes a 2×2 experiment (pricing page P1/P2 $\times$ onboarding flow O1/O2, $n = 4$ cohorts per cell, conversion in percent) as follows: ''Main effect of pricing page: $+3.1$ points for P2 ($p = 0.002$). Main effect of onboarding: $+0.2$ points for O2 ($p = 0.71$). Recommendation: switch everyone to P2; onboarding flow does not matter.'' The full ANOVA output, which the assistant also displays, includes the line AB: $F_0 = 18.4$, $p = 0.001$, and the cell means are

	O1	**O2**
P1	8.0	12.4
P2	15.3	11.5

Identify what the assistant did wrong, explain the consequence using the cell means, and write the corrected recommendation.

Exercise 9.12. A logistics consultant studies three dock-scheduling policies across three terminal types in a 3×3 design with $n = 1$ observation per cell, and delivers a report containing a full two-factor ANOVA table with an interaction line ($df_{AB} = 4$), an error line (''$df_E = 4$''), and an interaction F test that ''confirms policy-by-terminal interaction ($p = 0.03$).'' Explain what is impossible in this report, what the consultant's software most likely did, and what you would ask for or do instead.

In-Class Activity 9.1: Whose Model Is Better? It Depends

Week 14 — Lecture 25

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

Your team is the safety-review board for an autonomous-shuttle pilot. Two perception-model versions (V1, V2) were each tested on three scenario types with $n = 2$ batches per cell. The cell means (missed detections per 1,000 frames) are:

	Clear	**Rain**	**Night**	row mean
V1	9	15	22	15.33
V2	8	10	21	13.00

Part 1 — Sketch and read (5 min)

Sketch the interaction plot (misses vs. scenario, one line per version). Are the lines roughly parallel, diverging, or crossing? Where is the largest gap between versions, and what would you want to know before trusting that gap?

Part 2 — Which sentence may marketing use? (5 min)

The ANOVA gives $F_0^{AB} = 2.67$ against a critical value of 5.14, and both main effects are significant. Decide, as a board, which of these sentences is statistically defensible, and fix the indefensible one: (i) ''V2 misses fewer detections than V1, on average across scenarios.'' (ii) ''V2's advantage is concentrated in rain.''

gn the next test (5 min)

to add a third factor — sensor cleaning schedule (2 levels) — and keep $n = 2$. How many runs torial need? List the degrees of freedom for every line of the new ANOVA table and check they

hy must the interaction be tested before the main effects?

In-Class Activity 9.2: Sixteen Cohorts to a Better Signup Page

Week 14 — Lecture 26

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

Your startup can show experimental page variants to 16 cohorts of visitors this month. You are testing $A =$ headline (current $-$, new $+$) and $B =$ signup-form length (long $-$, short $+$), with $n = 4$ cohorts per combination. Conversions per cohort (totals of the 4 replicates): $(1) = 88$, $a = 112$, $b = 120$, $ab = 152$.

Part 1 — Contrasts by hand (5 min)

Compute the three contrasts and the three effect estimates (divide by $2n = 8$). Which change moves conversion more: the headline or the shorter form?

Part 2 — Is there synergy? (5 min)

The interaction effect estimate from Part 1 is small but positive. Explain in one or two sentences what a *large positive* interaction would have meant here, in plain product language, and what evidence would let you call the one you found ''real'' rather than noise.

ng up on a budget (5 min)

u want to screen five binary page factors, but you can afford 16 cohorts again. Name the design as 2^{k-p}), state its defining-relation idea in one sentence, and give one concrete thing you lose ll 2^5.

hat does the sparsity-of-effects principle let a single-replicate experimenter do?