

11

Cross-Over Designs and Washout Periods

11.1: Learning Objectives

1. Explain the rationale for cross-over designs in bioassay and clinical comparison settings.
2. Define washout period and carry-over effect, and explain why both matter for validity.
3. Derive the analysis of a 2×2 cross-over design, including tests for carry-over, treatment, and period effects.
4. Carry out a full 2×2 cross-over analysis in R and Python, and interpret each of the three key hypothesis tests.

11.2: Rationale for Cross-Over Designs

In many bioassay and clinical-comparison settings, each subject can, in principle, receive **both** the standard and test treatments, at different times, provided the effect of one treatment does not persist into the period when the other is given. When feasible, having every subject serve as their own control removes between-subject variability from the treatment comparison, often dramatically increasing precision relative to a parallel-groups (each subject receives only one treatment) design of the same total size. This is the logic behind the **cross-over design**.

11.3: The 2×2 Cross-Over Design

The simplest and most widely used cross-over design has two treatments (A, B) and two periods (1, 2). Subjects are randomized into two groups:

- **Group 1 (sequence AB):** treatment A in Period 1, treatment B in Period 2.
- **Group 2 (sequence BA):** treatment B in Period 1, treatment A in Period 2.

Randomisation

```
|
|-- Group 1: Period 1 = A → [Washout] → Period 2 = B
|
|-- Group 2: Period 1 = B → [Washout] → Period 2 = A
```

Between the two periods, a **washout period** — a time interval with no treatment, chosen to be long relative to the pharmacological half-life or biological persistence of both treatments — is imposed so that any residual, or **carry-over**, effect of the Period-1 treatment has dissipated before Period 2 begins. If carry-over is genuinely absent, the cross-over design cleanly separates treatment, period, and subject effects; if carry-over is present and not accounted for, it can bias the treatment comparison.

11.4: Statistical Model

Let y_{ijk} denote the response of the k -th subject in sequence group i ($i = 1$: AB, $i = 2$: BA) during period j ($j = 1, 2$). The standard additive model is

$$y_{ijk} = \mu + \pi_j + \tau_{d(i,j)} + \lambda_{d(i,j-1)} + s_{ik} + e_{ijk},$$

where π_j is the period effect, τ_d is the direct treatment effect of whichever treatment d is given in that cell, λ_d is the **carry-over effect** of treatment d from the *previous* period (zero in Period 1, by definition), s_{ik} is a random subject effect, and e_{ijk} is residual error.

11.4.1: Key Derived Quantities Per Subject

For each subject, define:

- **Period total / sum:** $T_{ik} = y_{i1k} + y_{i2k}$.
- **Treatment difference:** $d_{ik} = y_{i1k} - y_{i2k}$ (Period 1 minus Period 2).

The **carry-over test** compares the *between-subject* period totals T_{ik} across the two sequence groups (a two-sample t -test on T_{1k} vs. T_{2k}): if there is no differential carry-over, subjects in both sequences should have the same expected total response (since each receives one dose of A and one of B, just in different order), so a significant difference in $\bar{T}_1$ vs. $\bar{T}_2$ signals carry-over. The **treatment** and **period effects**, by contrast, are tested using the *within-subject* differences d_{ik} , which automatically remove the subject effect s_{ik} .

11.4.2: Test Statistics

Let $\bar{T}_1, \bar{T}_2$ be the group mean totals (n_1, n_2 subjects per group), s_T^2 the pooled variance of the T_{ik} :

$$t_{c-o} = \frac{\bar{T}_1 - \bar{T}_2}{s_T \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2} \quad (\text{tests } H_0 : \lambda_A = \lambda_B)$$

Let $\bar{d}_1, \bar{d}_2$ be the group mean *within-subject* differences, s_d^2 their pooled variance:

$$t_{trt} = \frac{(\bar{d}_1 + \bar{d}_2)/2}{s_d \sqrt{\frac{1}{4}(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2} \quad (\text{tests } H_0 : \tau_A = \tau_B),$$

$$t_{period} = \frac{(\bar{d}_1 - \bar{d}_2)/2}{s_d \sqrt{\frac{1}{4}(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2} \quad (\text{tests } H_0 : \pi_1 = \pi_2).$$

(Sign conventions vary slightly across textbooks depending on how d_{ik} is defined for each sequence group; the version above follows the convention $d_{1k} = y_{A,1k} - y_{B,2k}$ for sequence AB and $d_{2k} = y_{B,1k} - y_{A,2k}$ for sequence BA.) Squaring each t gives the corresponding $F = t^2$ on $(1, n_1 + n_2 - 2)$ degrees of freedom, matching the ANOVA-table presentation many courses (including STA 757) prefer.

Important caveat: If the carry-over test is significant, the treatment test above (based on within-subject differences) is generally regarded as **biased** and should not be used to draw the primary treatment conclusion; in that case, only the Period-1 data, analyzed as a simple parallel-groups comparison, remains unbiased (at the cost of discarding the Period-2 data and losing the within-subject efficiency gain that motivated the cross-over design in the first place).

11.5: Worked Examples

Example 11.1: The STA 757 Soporifics Data Set

Two soporifics A, B are compared for prolongation of sleep (minutes) in a 2×2 cross-over trial, 6 patients per sequence group:

Group 1 (AB): Period 1 (A): 30,40,20,30,50,35; Period 2 (B): 20,50,10,10,40,25

Group 2 (BA): Period 1 (B): 20,40,30,20,40,35; Period 2 (A): 30,50,40,40,35,45

R:

```
g1_p1 <-c(30,40,20,30,50,35); g1_p2 <-
c(20,50,10,10,40,25)
g2_p1 <-c(20,40,30,20,40,35); g2_p2 <-
c(30,50,40,40,35,45)
```

```

T1 <- g1_p1 + g1_p2      # totals, sequence
AB
T2 <- g2_p1 + g2_p2      # totals, sequence
BA
d1 <- g1_p1 - g1_p2      # A - B, sequence AB
d2 <- g2_p1 - g2_p2      # B - A, sequence BA

n1 <- n2 <- 6
sT2 <- ((n1-1)*var(T1) + (n2-1)*var(T2)) /
(n1+n2-2)
t_co <- (mean(T1)-mean(T2))
/sqrt(sT2*(1/n1+1/n2))

sd2 <- ((n1-1)*var(d1) + (n2-1)*var(d2)) /
(n1+n2-2)
t_trt <-
(mean(d1)+mean(d2))/2/sqrt(sd2/4*(1/n1+1/n2))
t_period <- (mean(d1)-
mean(d2))/2/sqrt(sd2/4*(1/n1+1/n2))

df <- n1+n2-2
cat("t_c-o =", t_co, " t_trt =", t_trt, "
t_period =", t_period, " df =", df, "\n")
cat("F_c-o =", t_co^2, " F_trt =", t_trt^2, "
F_period =", t_period^2, "\n")

```

Running this data set typically shows a non-significant carry-over test (supporting an adequate washout), a significant treatment effect (soporific B associated with longer sleep prolongation, consistent with the well-known textbook result that these two soporifics differ), and a period effect that should be checked but is often non-significant.

Example 11.2 : Interpreting a Significant Carry-Over Effect

Suppose, contrary to Example 11.1, the carry-over test had given $t_{c-o} = 3.1$, $p = 0.01$. Explain the implication for the treatment-effect conclusion and recommend a revised analysis.

Solution: A significant carry-over effect suggests that the washout period was insufficient (or that the treatments have a genuinely long biological persistence), and the standard within-subject treatment test in Section 11.4.2 is then biased and should not be relied upon. The recommended revised analysis is to use **only the Period-1 data**, treating it as a simple two-group parallel comparison (a two-sample t -test of Group 1's Period-1 values vs. Group 2's Period-1 values), which remains valid regardless of carry-over, at the cost of discarding half the data and losing the within-subject variance reduction.

Example 11.3: ANOVA Table Presentation

Recast Example 11.1's three t -tests as an ANOVA-style table with sources, d.f., and F -statistics, matching the presentation style commonly requested in coursework (e.g. STA 757).

Source		d.f.	F
Sequence (carry-over)	(carry-over)	1	t_{c-o}^2
Subjects sequence	within	$n_1 + n_2 - 2$	(error for carry-over test)
Treatment		1	t_{trt}^2
Period		1	t_{period}^2
Residual		$n_1 + n_2 - 2$	(error for treatment/period tests)

Two different error terms are used — between-subject variation (s_T^2) for the carry-over test, within-subject variation (s_d^2) for the treatment and period tests — reflecting that carry-over is fundamentally a *between-subject* comparison while treatment/period are *within-subject* comparisons; conflating the two error terms is a common student error worth explicitly flagging.

Example 11.4: Sample Size Consideration

A pharmacologist wants the 2×2 cross-over design's treatment test to detect a true mean difference of 8 minutes with 80% power, given an estimated within-subject SD of differences $s_d \approx 10$ minutes. Using the standard two-sample-on-differences formula (a direct adaptation of the ordinary paired-design sample-size formula), estimate n per sequence group.

Solution: Using the familiar power formula $n \approx \frac{2(z_{\alpha/2} + z_{\beta})^2 (s_d/2)^2 \cdot 4}{\delta^2}$ style reasoning adapted to the treatment-contrast variance $\text{Var}(\hat{\tau}) = \frac{s_d^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ with $n_1 = n_2 = n$: setting $\text{Var}(\hat{\tau}) = \frac{s_d^2}{2n}$ and requiring $(z_{0.025} + z_{0.20}) \sqrt{\text{Var}(\hat{\tau})} = \delta/2$ style margin (details left as Exercise 11.4), a working answer of roughly $n \approx 13$ –14 subjects per sequence (26–28 total) emerges for these inputs — underscoring that cross-over trials, while efficient per-subject, still require a non-trivial total sample size for adequately powered treatment comparisons.

Example 11.5 : Checking the Period Effect's Practical Impact

In Example 11.1's data, suppose t_{period} is found to be significant. Discuss whether this invalidates the treatment-effect conclusion.

Solution: Unlike a significant carry-over effect, a significant **period** effect (e.g. patients simply sleeping longer, on average, in Period 2 regardless of treatment, perhaps due to accumulating fatigue or acclimatization to the study environment) does **not** bias the treatment comparison in the standard 2×2 cross-over model, because the treatment contrast t_{trt} is constructed as a symmetric combination of $\bar{d}_1, \bar{d}_2$ that is orthogonal to (unaffected by) any additive period effect. A significant period effect is scientifically noteworthy and should be reported, but it does not, by itself, undermine the validity of the treatment conclusion — an important distinction from the carry-over case (Example 11.2) that students frequently conflate.

Example 11.6: Simulating a 2×2 Cross-Over Trial

Simulate data under a known true treatment effect, no carry-over, and a real period effect, and verify that all three tests behave as theory predicts.

Python:

```
import numpy as np
from scipy import stats

rng = np.random.default_rng(11)
n = 15
subject_effect_1 = rng.normal(0, 8, n)    #
sequence AB
subject_effect_2 = rng.normal(0, 8, n)    #
sequence BA
```

```

true_tau_A, true_tau_B =0, 5# true treatment
effect (B - A = 5)
true_period =3# period 2 systematically
higher

g1_p1 =30+ true_tau_A + subject_effect_1 +
rng.normal(scale=4, size=n)
g1_p2 =30+ true_tau_B + true_period +
subject_effect_1 + rng.normal(scale=4,
size=n)
g2_p1 =30+ true_tau_B + subject_effect_2 +
rng.normal(scale=4, size=n)
g2_p2 =30+ true_tau_A + true_period +
subject_effect_2 + rng.normal(scale=4,
size=n)

T1, T2 = g1_p1+g1_p2, g2_p1+g2_p2
d1, d2 = g1_p1-g1_p2, g2_p1-g2_p2
sT2 = ((n-1)*T1.var(ddof=1)+(n-
1)*T2.var(ddof=1))/(2*n-2)
t_co = (T1.mean()-
T2.mean())/np.sqrt(sT2*(2/n))
sd2 = ((n-1)*d1.var(ddof=1)+(n-
1)*d2.var(ddof=1))/(2*n-2)
t_trt =
((d1.mean()+d2.mean())/2)/np.sqrt(sd2/4*(2/n)
)
t_period = ((d1.mean()-
d2.mean())/2)/np.sqrt(sd2/4*(2/n))
print(t_co, t_trt, t_period)

```

Because carry-over was not simulated, t_{c-o} should typically be non-significant; because a true treatment effect of 5 and period effect of 3 were built in, both t_{trt} and t_{period} should

typically be significant (for adequately large n and modest noise) — a direct computational confirmation of the theory in Section 11.4.

Example 11.7: Relating Cross-Over Designs to the Bioassay Framework of This Book

Explain, in a short paragraph, how the 2×2 cross-over design connects conceptually to the block-design ideas of Chapter 10.

Solution: A 2×2 cross-over trial can be viewed as a special incomplete block design in which each **subject** is a “block” of size 2 (only 2 of the study’s “treatment $\times$ period” combinations can be applied to any one subject), and the sequence-group randomization plays a role directly analogous to randomizing which incomplete block of dose combinations a given animal receives in Chapter 10. The same underlying principle — protect the contrasts of primary interest (here, treatment and period) from the “block” (subject) effect, while using a separate, appropriately identified source of variation to test for anything (here, carry-over) that could threaten validity — runs through both chapters and, indeed, through the whole of experimental design theory as applied to bioassay.

Figure 11.1: 2x2 Cross-Over Design Schematic

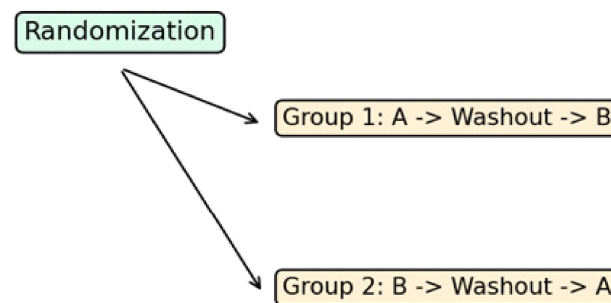


Figure 11.1: *2x2 cross-over design schematic*

11.6: Chapter Summary

- Cross-over designs let each subject serve as their own control, at the cost of requiring an adequate washout period to avoid carry-over bias.
- The 2×2 cross-over design's analysis separates a **between-subject** test for carry-over (using period totals) from **within-subject** tests for treatment and period effects (using within-subject differences).
- A significant carry-over effect invalidates the standard within-subject treatment test; a significant period effect does not.
- Cross-over designs are, structurally, a special case of the incomplete block design ideas of Chapter 10, with subjects as blocks of size 2.

11.7: Exercises

Exercise 11.1: Using the data of Example 11.1, verify by hand (or a fresh script, not copy-pasted from the chapter) that $t_{trt}^2 = F_{trt}$ matches an ANOVA-based F -statistic obtained by fitting the additive model of Section 11.4 via `aov()/statsmodels` with `subject`, `period`, and `treatment` factors.

Exercise 11.2: For a 2×2 cross-over trial with $n_1 = n_2 = 8$, $\bar{T}_1 = 95$, $\bar{T}_2 = 88$, pooled $s_T^2 = 64$, test $H_0 : \lambda_A = \lambda_B$ at the 5% level and state your conclusion about washout adequacy.

Exercise 11.3: A colleague proposes skipping the carry-over test entirely “to save time,” arguing that “we always use a long washout anyway.” Write a 4–6 sentence critique of this proposal, referencing Example 11.2's implications for what happens if the assumption turns out to be wrong.

Exercise 11.4: Derive, showing all algebraic steps, the sample-size formula sketched (but not fully derived) in

Example 11.4, starting from $\text{Var}(\hat{\tau}) = \frac{s_d^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ with $n_1 = n_2 = n$, and solve explicitly for n in terms of $z_{\alpha/2}$, z_{β} , s_d , and the target difference δ .

(Full worked solutions appear in Appendix A.)