CHAPTER 10

Pairwise Randomized Experiments

10.1 INTRODUCTION

In the previous chapter we analyzed stratified randomized experiments, where a sample of size N was partitioned into J strata, and within each stratum a completely randomized experiment was conducted. In this chapter we consider a special case of the stratified randomized experiment. Each stratum contains exactly two units, with one randomly selected to be assigned to the treatment group, and the other one assigned to the control group. Such a design is known as a *pairwise randomized experiment* or *paired comparison*. Although this can be viewed simply as a special case of a stratified randomized experiment, there are two features of this design that warrant special attention. First, the fact that there is only a single unit in each treatment group in each stratum (or pair in this case) implies that the Neyman sampling variance estimator that we discussed in the chapters on completely randomized experiments (Chapter 6) and stratified randomized experiments (Chapter 9) cannot be used; that estimator requires the presence of at least two units assigned to each treatment in each stratum. Second, each stratum has the same proportion of treated units, which allows us to analyze the within-stratum estimates symmetrically; the natural estimator for the average treatment effect weights each stratum equally.

As in the case of stratified randomized experiments, the motivation for eliminating some of the possible assignments in pairwise randomized experiments is that *a priori* those values of the assignment vectors that are eliminated are expected to lead to less informative inferences. This argument relies on the within-pair variation in potential outcomes being small relative to the between-pair variation. Often the assignment to pairs is based on covariates. Units are matched to other units based on their similarity in these covariates, with the expectation that this similarity corresponds to similarity in the potential outcomes under each treatment. Suppose, for example, that the treatment is an expensive surgical procedure for a relatively common medical condition. It may not be financially feasible to apply the treatment to many individuals. To increase the precision of an experiment, it may, in such cases, be sensible to use the following steps. First randomly draw J individuals from the target population of individuals who have the condition for which the surgery may be beneficial. Then, for each of these J individuals, find a matching individual in the same population, as similar as possible to the original

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

unit in terms of the characteristics that may be correlated with potential outcomes and efficacy of the treatment. If the population is relatively large, it may be possible to get very close matches with respect to a large number of characteristics, thereby reducing the variation in treatment-control differences in potential outcomes. Given these J matched pairs, one can then conduct a pairwise randomized experiment by randomly selecting one member of each pair to be assigned to the active treatment.

In this chapter we discuss analyses for such pairwise randomized experiments. In particular we discuss the calculation of Fisher exact p-values and Neyman's repeated sampling perspective, as well as regression and model-based inference. We focus primarily on conceptual issues that are special to this design.

Section 10.2 describes the data set we use to illustrate the concepts discussed in this chapter, which comes from a randomized experiment conducted around 1970 to evaluate the effect of an educational children's television program on reading ability as measured through test scores. Section 10.3 discusses the structure of paired randomized experiments and introduces some additional notation. In 10.4 we discuss the application of Fisher's exact p-value calculations in the setting of paired randomized experiments. Next, in Section 10.5 we discuss the implications of pairwise randomization for the methods discussed in Chapter 6 based on Neyman's repeated sampling perspective. In Sections 10.6 and 10.7 we analyze regression and model-based imputation methods. Section 10.8 concludes.

10.2 THE CHILDREN'S TELEVISION WORKSHOP EXPERIMENT DATA

The Children's Television Workshop experiment was designed by Ball, Bogatz, Rubin, and Beaton (1973) to evaluate *The Electric Company*, an educational television program aimed at improving reading skills for young children, somewhat similar to *Sesame Street*. The experiment was conducted in two locations, Youngstown, Ohio, and Fresno, California, where *The Electric Company* was not broadcast on local stations. In each location a number of schools was selected. Within each school, a pair of two classes was selected. Within each pair, one class was randomly assigned to be shown *The Electric Company* show during the standard reading-class period, and the other class continued with the regular reading curriculum.

Here we focus on the data from Youngstown, where two first-grade classes from each of eight schools participated in the experiment. The data for the sixteen classes for the Youngstown location from this experiment are displayed in Table 10.1, which presents values of a pre-test score, the post-test score (the primary outcome), an indicator for the pair or school to which the unit belongs, and an indicator for the treatment (one for classes that viewed *The Electric Company* program, and zero for classes in the control group).

10.3 PAIRWISE RANDOMIZED EXPERIMENTS

A pairwise randomized experiment is a special case of a stratified randomized experiment where the number of units, N, is even, the number of strata is $J = N/2$, with one

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

Table 10.1. *Data from Youngstown Children's Television Workshop Experiment*

Pair G_i	Treatment W_i	Pre-Test Score X_i	Post-Test Score Y_i^{obs}	Normalized Rank Post-Test Score R_i
1	0	12.9	54.6	−7.5
1	1	12.0	60.6	2.5
2	0	15.1	56.5	−4.5
2	1	12.3	55.5	5.5
3	0	16.8	75.2	0.5
3	1	17.2	84.8	4.5
4	0	15.8	75.6	1.5
4	1	18.9	101.9	7.5
5	0	13.9	55.3	−6.5
5	1	15.3	70.6	−1.5
6	0	14.5	59.3	−3.5
6	1	16.6	78.4	2.5
7	0	17.0	87.0	5.5
7	1	16.0	84.2	3.5
8	0	15.8	73.7	−0.5
8	1	20.1	108.6	7.5

treated unit and one control unit in each stratum ($N_t(j) = N_c(j) = 1$ and $N(j) = 2$ for all $j = 1, \ldots, J$), so that each stratum is a pair. Let G_i be the variable indicating the pair, with $G_i \in \{1, \ldots, N/2\}$. The pair indicator can be thought of as a function of covariates. Of course this indicator is a pre-treatment variable in the sense that it is not affected by the treatment. Within each pair there are $\binom{N(j)}{N_t(j)} = \binom{2}{1} = 2$ possible assignments, so that the probability for any assignment vector $\mathbf{W}$ is

$$p(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{N/2} \binom{N(j)}{N_t(j)}^{-1} = \prod_{j=1}^{N/2} \frac{1}{2} = 2^{-N/2}, \quad \text{for } \mathbf{W} \in \mathbb{W}^+,$$

where

$$\mathbb{W}^+ = \left\{ \mathbf{W} \,\middle|\, \sum_{i:G_i=j} W_i = 1 \text{ for } j = 1, \ldots, N/2 \right\}.$$

Because the assignment mechanism fits into the stratified randomized experiments discussed in Chapter 9, we can directly use many of the methods discussed in that chapter. However, there is one important difference. Because all strata have the property that they contain exactly one treated and one control unit, methods that rely on the presence of multiple control or multiple treated units cannot be applied.

To facilitate the discussion of pairwise randomized experiments, it is useful to introduce some additional notation. We arbitrarily label the two units within a pair as units A and B. Then, for all pairs $j = 1, \ldots, N/2$, let $(Y_{j,A}(0), Y_{j,A}(1))$ and $(Y_{j,B}(0), Y_{j,B}(1))$ be the potential outcomes for units A and B, respectively, in pair j, and let $W_{j,A}$ and $W_{j,B}$ be

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

Table 10.2. ***Potential Outcomes and Covariates from Children's Television Workshop Experiment, from Table 10.1***

Pair	Unit A					Unit B				
	$Y_{i,A}(0)$	$Y_{i,A}(1)$	$W_{i,A}$	$Y^{\text{obs}}_{i,A}$	$X_{i,A}$	$Y_{i,B}(0)$	$Y_{i,B}(1)$	$W_{i,B}$	$Y^{\text{obs}}_{i,B}$	$X_{i,B}$
1	54.6	?	0	54.6	12.9	?	60.6	1	60.6	12.0
2	56.5	?	0	56.5	15.1	?	55.5	1	55.5	13.9
3	75.2	?	0	75.2	16.8	?	84.8	1	84.8	17.2
4	76.6	?	0	75.6	15.8	?	101.9	1	101.9	18.9
5	55.3	?	0	55.3	13.9	?	70.6	1	70.6	15.3
6	59.3	?	0	59.3	14.5	?	78.4	1	78.4	16.6
7	87.0	?	0	87.0	17.0	?	84.2	1	84.2	16.0
8	73.7	?	0	73.7	15.8	?	108.6	1	108.6	20.1

the treatment indicators for these units. In a pairwise randomized experiment, one unit in each pair is randomly assigned to the active treatment, and the other unit is assigned to the control treatment, thus $W_{j,A} = 1 - W_{j,B}$, with $\Pr(W_{j,A} = 1|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = 1/2$. Define also

$$Y^{\text{obs}}_{j,A} = \begin{cases} Y_{j,A}(0) & \text{if } W_{j,A} = 0, \\ Y_{j,A}(1) & \text{if } W_{j,A} = 1, \end{cases} \quad \text{and} \quad Y^{\text{obs}}_{j,B} = \begin{cases} Y_{j,B}(0) & \text{if } W_{j,A} = 1, \\ Y_{j,B}(1) & \text{if } W_{ji,A} = 0. \end{cases}$$

The average treatment effect within pair j is $\tau_{\text{pair}}(j)$,

$$\tau^{\text{pair}}(j) = \frac{1}{2} \sum_{i:G_i=j} \big(Y_i(1) - Y_i(0)\big) = \frac{1}{2}\Big(\big(Y_{j,A}(1) - Y_{j,A}(0)\big) + \big(Y_{j,B}(1) - Y_{j,B}(0)\big)\Big).$$

The finite-sample average treatment effect is

$$\tau_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i(1) - Y_i(0)\big) = \frac{2}{N} \sum_{j=1}^{N/2} \tau^{\text{pair}}(j).$$

Also define the pair of observed variables, one treated and one control from each pair:

$$Y^{\text{obs}}_{j,c} = \begin{cases} Y^{\text{obs}}_{j,A} & \text{if } W_{i,A} = 0, \\ Y^{\text{obs}}_{j,B} & \text{if } W_{i,A} = 1, \end{cases} \quad \text{and} \quad Y^{\text{obs}}_{j,t} = \begin{cases} Y^{\text{obs}}_{j,B} & \text{if } W_{i,A} = 0, \\ Y^{\text{obs}}_{j,A} & \text{if } W_{i,A} = 1. \end{cases}$$

Table 10.2 displays some of these variables for the 16 classes in the Children's Television Workshop Experiment.

10.4 FISHER'S EXACT P-VALUES IN PAIRWISE RANDOMIZED EXPERIMENTS

The same way stratified randomization did not pose any conceptual difficulties for the calculation of Fisher Exact P-values (FEPs), pairwise randomization does not introduce

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

any new issues. Let us focus in this discussion on the usual Fisher null hypothesis of absolutely no treatment effects for any units,

$$H_0: \quad Y_i(0) = Y_i(1), \quad \text{for all } i = 1, \ldots, N.$$

With the assignment mechanism fully known, we can, under H_0, for any fixed statistic, derive the randomization distribution and thus calculate the corresponding p-value. An obvious statistic is the average, over the $J = N/2$ pairs, of the difference between the treated and control outcomes within each pair:

$$\begin{aligned} T^{\text{dif}} &= \left| \frac{1}{J} \sum_{j=1}^{J} \left(Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} \right) \right| \\ &= \left| \frac{1}{J} \sum_{j=1}^{J} \left(W_{i,A} \cdot \left(Y_{j,A}^{\text{obs}} - Y_{j,B}^{\text{obs}} \right) + (1 - W_{i,A}) \cdot \left(Y_{j,B}^{\text{obs}} - Y_{j,A}^{\text{obs}} \right) \right) \right|. \end{aligned}$$

Because each pair has a single treated and a single control unit, this also equals the difference between average outcomes for treated and control units, $T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right|$, the statistic that was the starting point of the discussion of the FEP approach in Chapter 5. However, the p-value for this statistic will be different than that calculated under the randomization distribution considered in Chapter 5 because here the randomization distribution is based on the assignment mechanism corresponding to a pairwise randomized experiment, not the assignment mechanism corresponding to a completely randomized experiment, leading to fewer elements in $\mathbb{W}^+$.

Alternative statistics include the average of within-pair differences in logarithms or other transformations of the basic outcomes, such as ranks. To calculate the rank statistic, let R_i be the rank of Y_i^{obs} among the N values $Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}}$, normalized to have mean zero, and let $R_{j,A}$ and $R_{j,B}$ be the rank of the A and B units in pair j, among all N units. For the Children's Television Workshop data, the ranks for the sixteen classes are displayed in the last column in Table 10.1. Then the rank statistic is

$$T^{\text{rank}} = \left| \overline{R}_{\text{t}} - \overline{R}_{\text{c}} \right| = \left| \frac{1}{J} \sum_{j=1}^{J} \left(W_{j,A} \cdot \left(R_{j,A} - R_{j,B} \right) + (1 - W_{j,A}) \cdot \left(R_{j,B} - R_{j,A} \right) \right) \right|.$$

Using ranks in pairwise randomized experiments has the same advantages as using ranks in completely randomized experiments, namely reducing the sensitivity to outliers. Another statistic that is specific to pairwise randomized experiments is based on the average within-pair rank of the observed outcomes. That is, for each pair we calculate an indicator for whether the observed outcome for the treated unit is larger than the observed outcome for the control unit, and an indicator whether the observed outcome for the control unit is larger than the observed outcome for the treated unit. (Using the two indicators, rather than one of the indicators alone, allows for a simpler way of

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

dealing with within-pair ties.) We then average the difference between these indicators,

$$T^{\text{rank,pair}} = \left| \frac{2}{N} \sum_{j=1}^{N/2} \left(\mathbf{1}_{Y_{j,1}^{\text{obs}} > Y_{j,0}^{\text{obs}}} - \mathbf{1}_{Y_{j,1}^{\text{obs}} < Y_{j,0}^{\text{obs}}} \right) \right|,$$

similar to the statistic $T^{\text{rank,stratum}}$ in Chapter 9. Like the rank-based statistic, T^{rank}, this statistic is particularly insensitive to the presence of outliers in the observed potential outcomes, and when there is substantial variation in the level of the outcomes between the pairs, it has more power than the statistic T^{rank} against alternatives under which the treatment effect is constant.

We apply these Fisher exact p-value calculations to the Children's Television Workshop data, using the null hypothesis of no effect whatsoever. Although the p-value is valid only for a single statistic, for illustrative purposes we do the analysis for all three statistics. For the statistic based on the absolute value of the difference in average outcomes by treatment status, we find

$$T^{\text{dif}} = 13.4, \quad \text{p-value} = 0.031.$$

Using the rank statistic, we find

$$T^{\text{rank}} = 3.8, \quad \text{p-value} = 0.031.$$

The last statistic, based on the indicator for whether within the pair the treated outcome was larger or smaller than the control outcome, leads to

$$T^{\text{rank,pair}} = 0.5, \quad \text{p-value} = 0.145.$$

The mechanical reason that the p-value for the within-pair rank statistic is less significant than for the other statistics is that for the two pairs where the outcome for the treated unit is less than the outcome for the control unit in the pair, the difference in outcomes is small. These small differences do not affect the average difference much, but they do affect the within-pair rank statistic. The other two p-values suggest that the television program did affect reading ability at conventional significance levels.

10.5 THE ANALYSIS OF PAIRWISE RANDOMIZED EXPERIMENTS FROM NEYMAN'S REPEATED SAMPLING PERSPECTIVE

Consider first the analysis of the average treatment effect in a single pair. The obvious estimator for the average treatment effect in pair j, $\tau^{\text{pair}}(j)$, is

$$\hat{\tau}^{\text{pair}}(j) = Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} = \sum_{i:G_i=j} (2 \cdot W_i - 1) \cdot Y_i^{\text{obs}}.$$

The values of $\hat{\tau}^{\text{pair}}(j)$ for the eight pairs in the Children's Television Workshop data are displayed in Table 10.3.

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

Table 10.3. ***Observed Outcome Data from Children's Television Workshop Experiment by Pair***

Pair	Outcome for Control Unit	Outcome for Treated Unit	Difference
1	54.6	60.6	6.0
2	56.5	55.5	−1.0
3	75.2	84.8	9.6
4	75.6	101.9	26.3
5	55.3	70.6	15.3
6	59.3	78.4	19.1
7	87.0	84.2	−2.8
8	73.7	108.6	34.9
Mean	67.2	80.6	13.4
(S.D.)	(12.2)	(18.6)	(13.1)

Next, let us consider inference, first for the within-pair average treatment effect $\tau^{\text{pair}}(j)$. For each pair we have a completely randomized experiment with two units of which one unit is assigned to active treatment. From the results in Chapter 6 on Neyman's repeated sampling approach, it follows that the estimator $\hat{\tau}^{\text{pair}}(j)$ is unbiased for the average treatment effect $\tau^{\text{pair}}(j)$ within this pair and that its sampling variance, based on the randomization distribution, is equal to

$$\mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = \frac{S_{\text{c}}(j)^2}{N_{\text{c}}(j)} + \frac{S_{\text{t}}^2(j)}{N_{\text{t}}(j)} - \frac{S_{\text{ct}}(j)^2}{N(j)}.$$

With $N(j) = 2$ and $N_{\text{c}}(j) = N_{\text{t}}(j) = 1$, this expression simplifies to

$$\mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = S_{\text{c}}(j)^2 + S_{\text{t}}^2(j) - \frac{S_{\text{ct}}(j)^2}{2}.$$

The within-pair variances can be written as

$$S_{\text{c}}^2(j) = \sum_{i:G_i=j} \left(Y_i(0) - \overline{Y}_j(0)\right)^2 = \frac{1}{2} \cdot \left(Y_{j,A}(0) - Y_{j,B}(0)\right)^2,$$

$$S_{\text{t}}^2(j) = \sum_{i:P_i=j} \left(Y_i(1) - \overline{Y}_j(1)\right)^2 = \frac{1}{2} \cdot \left(Y_{j,A}(1) - Y_{j,B}(1)\right)^2,$$

and

$$S_{\text{ct}}^2(j) = \frac{1}{2} \cdot \left(\left(Y_{j,A}(1) - Y_{j,A}(0)\right) - \left(Y_{j,B}(1) - Y_{j,B}(0)\right)\right)^2,$$

where

$$\overline{Y}_j(0) = \frac{1}{2} \cdot \left(Y_{j,A}(0) + Y_{j,B}(0)\right) \quad \text{and} \quad \overline{Y}_j(1) = \frac{1}{2} \cdot \left(Y_{j,A}(1) + Y_{j,B}(1)\right).$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

If the primary interest is in the finite-sample average treatment effect, $\tau_{\rm fs}$, that is, the within-pair average treatment effect averaged over the $N/2$ pairs,

$$\tau_{\rm fs} = \frac{1}{N/2}\sum_{j=1}^{N/2} \tau^{\rm pair}(j),$$

the natural estimator is

$$\hat{\tau}^{\rm dif} = \frac{1}{N/2}\sum_{j=1}^{N/2} \hat{\tau}^{\rm pair}(j) = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}. \tag{10.1}$$

By unbiasedness of the within-pair estimators, $\hat{\tau}$ is unbiased for the sample average treatment effect, $\tau_{\rm S}$. Its sampling variance over the randomization distribution is

$$\mathbb{V}_W\left(\hat{\tau}^{\rm dif}\right) = \frac{1}{(N/2)^2}\sum_{j=1}^{N/2}\left(S_{\rm c}^2(j) + S_{\rm t}^2(j) - \frac{S_{\rm ct}^2(j)}{2}\right).$$

So far the discussion is exactly analogous to the discussion for stratified randomized experiments in the previous chapter. However, one of the special features of pairwise randomized experiments, alluded to in the introduction to this chapter, creates a complication for the estimation of the sampling variance. In a completely randomized experiment (and similarly, within a stratum in the stratified randomized experiment), the standard estimator for the sampling variance for the observed difference in treatment and control averages is

$$\hat{\mathbb{V}}^{\rm neyman}\left(\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\right) = \frac{s_{\rm c}^2}{N_{\rm c}} + \frac{s_{\rm t}^2}{N_{\rm t}},$$

with

$$s_{\rm c}^2 = \frac{1}{N_{\rm c}-1}\sum_{i:W_i=0}\left(Y_i(0) - \overline{Y}_{\rm c}^{\rm obs}\right)^2 = \frac{1}{N_{\rm c}-1}\sum_{i:W_i=0}\left(Y_i^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\right)^2,$$

and analogously

$$s_{\rm t}^2 = \frac{1}{N_{\rm t}-1}\sum_{i:W_i=1}\left(Y_i^{\rm obs} - \overline{Y}_{\rm t}^{\rm obs}\right)^2.$$

Because within each stratum (or pair in this case) the numbers of control and treated units are $N_{\rm c} = N_{\rm t} = 1$, these estimators, $s_{\rm c}^2$ and $s_{\rm t}^2$, cannot be used, and the standard estimator for the sampling variance of the estimated overall average effect is not feasible.

One solution to this problem is to assume that the treatment effect is constant and additive, not only within pairs but also across pairs. Because of the assumption of a constant treatment effect within pairs, it follows that the within-pair sampling variance is

$$\mathbb{V}_W(\hat{\tau}^{\rm pair}(j)) = 2\cdot S^2(j), \quad \text{where } S^2(j) = S_{\rm c}^2(j) = S_{\rm t}^2(j).$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

Moreover, if the treatment effect is constant across pairs, $\tau^{\text{pair}}(j) = \tau_S$ for all j, the within-pair variances are constant, $S^2(j) = S^2$ for all j, and

$$\mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{(N/2)^2}\sum_{j=1}^{N/2}\left(S_c^2(j) + S_t^2(j) - \frac{S_{ct}^2(j)}{2}\right) = \frac{4}{N}\cdot S^2,$$

which can be estimated by calculating the sample variance of the pair-level treatment effect estimates:

$$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{4}{N\cdot(N-2)}\cdot\sum_{j=1}^{N/2}\left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2.$$

If there is heterogeneity in the treatment effects, then this sampling variance estimator is upwardly biased, and the corresponding confidence intervals will be conservative in the usual statistical sense.

Theorem 10.1 *Suppose we have J pairs of units, and randomly assign one unit from each pair to the active treatment and the other unit to the control treatment. Then (i) $\hat{\tau}^{\text{dif}}$ is unbiased for τ_{fs}, (ii) the sampling variance of $\hat{\tau}^{\text{dif}}$ is*

$$\mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{N^2}\sum_{j=1}^{N/2}(Y_{j,A}(0) + Y_{j,A}(1) - \left(Y_{j,B}(0) + Y_{j,B}(1)\right))^2,$$

and (iii) the estimator for the sampling variance

$$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{4}{N\cdot(N-2)}\cdot\sum_{j=1}^{N/2}\left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2,$$

satisfies

$$\mathbb{E}\left[\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)\right] = \mathbb{V}_W(\hat{\tau}^{\text{dif}}) + \frac{4}{N\cdot(N-2)}\cdot\sum_{j=1}^{N/2}\left(\tau^{\text{pair}}(j) - \tau\right)^2,$$

with the expected value equal to $\mathbb{V}_W(\hat{\tau}^{\text{dif}})$ if the treatment effect is constant across and within pairs.

Proof of Theorem 10.1: See Appendix.

Let us return to the data from the Children's Television Workshop experiment. The within-pair differences $\hat{\tau}^{\text{pair}}(j)$ are displayed in Table 10.3. Their average is

$$\hat{\tau}^{\text{dif}} = \frac{1}{8}\cdot\sum_{j=1}^{8}\hat{\tau}^{\text{pair}}(j) = 13.4,$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

and its estimated sampling variance is

$$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{8 \cdot (8-1)} \cdot \sum_{j=1}^{8} \left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2 = 4.6^2.$$

The standard, Gaussian-distribution-based asymptotic 95% confidence interval is

$$\text{CI}^{0.95}(\tau_{\text{fs}}) = \left(\hat{\tau} - 1.96 \times \sqrt{\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)}, \hat{\tau} + 1.96 \times \sqrt{\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)}\right) = (4.3, 22.5). \tag{10.2}$$

Because we have only eight pairs of classes, one may wish to use a confidence interval based on the t-distribution with degrees of freedom equal to $N/2 - 1 = 7$, with 0.975 quantile equal to 2.365, leading to a slightly wider confidence interval

$$\begin{aligned}\text{CI}^{0.95}_{\text{t}(7)}(\tau_{\text{fs}}) &= \left(\hat{\tau} - 2.365 \times \sqrt{\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)}, \hat{\tau} + 2.365 \times \sqrt{\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)}\right) \\ &= (2.5, 24.3).\end{aligned} \tag{10.3}$$

Let us now illustrate the benefits of doing a pairwise randomized experiment instead of a completely randomized experiment. Suppose we had done a completely randomized experiment and had the same assignment vector. In that case we would have the same point estimate for the average treatment effect, namely $\hat{\tau}^{\text{dif}} = \overline{Y}^{\text{obs}}_{\text{t}} - \overline{Y}^{\text{obs}}_{\text{c}} = 13.4$. However, we would have a different estimate of the sampling variance. Using the standard Neyman estimated sampling variance discussed in Chapter 6, we would have estimated the sampling variance of the two potential outcomes as

$$s^2_{\text{c}} = \frac{1}{N_{\text{c}} - 1} \sum_{i:W_i=0} \left(Y^{\text{obs}}_i - \overline{Y}^{\text{obs}}_{\text{c}}\right)^2 = 18.5^2, \quad \text{and} \quad s^2_{\text{t}} = 12.2^2,$$

leading to an estimate for the sampling variance of the estimated average effect of

$$\hat{\mathbb{V}}^{\text{neyman}} = \frac{s^2_{\text{c}}}{8} + \frac{s^2_{\text{t}}}{8} = 7.8^2.$$

This sampling variance estimate is substantially larger than the estimate based on the pairwise randomization, $\hat{\mathbb{V}}^{\text{pair}} = 4.6^2$, because the observed variance of potential outcome within pairs is substantially smaller than it would be if units were randomly assigned to pairs. In other words, in this application, the assignment to pairs is effective, in the sense that it is based on factors that make the within-pair units substantially more similar than randomly selected units, probably leading to substantially more precise estimates.

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

10.6 REGRESSION-BASED ANALYSIS OF PAIRWISE RANDOMIZED EXPERIMENTS

In this section the second special feature of pairwise randomized experiments, alluded to in the introduction of this chapter, motivates an analysis that is different from that discussed for stratified randomized experiments. In the discussions of regression-based analyses in completely and stratified randomized experiments, the basic outcome in the analysis was Y_i^{obs}, the observed outcome for unit i. Here, instead, we use as the primary outcome in the regression analysis the within-pair difference in observed outcomes of the treated and the control unit in the pair,

$$\hat{\tau}^{\text{pair}}(j) = Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}},$$

with the pair serving as the unit of analysis. We take a super-population perspective, where the pairs of units are drawn randomly from a large population, and one member of each pair is randomly assigned to the treatment group, and the other to the control group. The population average treatment effect is $\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[\tau^{\text{pair}}(j)]$, with the expectation taken over the random sampling of the pairs.

The standard estimator for the average treatment effect in a pairwise randomized experiment is the simple average of the within-pair differences,

$$\hat{\tau}^{\text{dif}} = \frac{2}{N} \sum_{j=1}^{N/2} \hat{\tau}^{\text{pair}}(j).$$

This estimator can also be interpreted as a regression estimator, where the regression function is specified simply as a constant:

$$\hat{\tau}^{\text{pair}}(j) = \tau_{\text{sp}} + \varepsilon_j.$$

The more interesting question is how to include additional covariates, beyond the implicit use of the pair indicators, into the regression function. As before, because of the randomization, we do not *need* to include additional covariates in order to remove bias, because the estimator $\hat{\tau}$ is unbiased over the randomization distribution without including covariates. The goal when including additional covariates is to improve the precision of the estimator in cases where the covariates are strongly correlated with the treatment-control differences in potential outcomes. Before discussing particular specifications, we first define $X_{j,A}$ and $X_{j,B}$ to be the covariate values for units A and B respectively within pair j. Then we define the within-pair observed difference in covariates between the treated and control units,

$$\Delta_{X,j} = \left(W_{j,A} \cdot \left(X_{j,A} - X_{j,B}\right) + (1 - W_{j,A}) \cdot \left(X_{j,B} - X_{j,A}\right)\right),$$

and the average covariate value within the pair,

$$\overline{X}_j = \left(X_{j,A} + X_{j,B}\right)/2.$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

There are two leading approaches to including the covariates in the regression analysis. First, we can include them in the form of the within-pair difference $\Delta_{X,j}$. This is an attractive option if one thinks the conditional expectation given covariates of the pairwise difference of potential outcomes is additive and linear in the treatment minus control difference in covariates. In other words, the inclusion of $\Delta_{X,j}$ in the regression function makes sense if the covariate X_i is associated with both potential outcomes $Y_i(0)$ and $Y_i(1)$ to approximately equal degrees. Second, we can include the average value of the covariates $\overline{X}_j$. This is a natural specification if one thinks the treatment effect, the difference in potential treated and control outcomes, rather than the level of the potential outcomes, is linear in the covariates. The most general version of the regression function we consider includes the covariates both as within-pair differences and pair averages, where the latter is in deviations from the overall covariate mean $\overline{X}$:

$$\hat{\tau}^{\text{pair}}(j) = \tau + \beta \cdot \Delta_{X,j} + \gamma \cdot (\overline{X}_j - \overline{X}) + \varepsilon_j.$$

Let $(\tau^*, \beta^*, \gamma^*)$ be the population values, defined analogously to the way they were defined in Chapter 7:

$$\left(\tau^*, \beta^*, \gamma^*\right) = \arg\min_{\tau,\beta,\gamma} \mathbb{E}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right],$$

where $\mu_X = \mathbb{E}_{\text{sp}}(X)$ is the super-population mean of X_i. Here we use again the convention that the expectation operator without subscript is both over the randomization distribution and over the distribution induced by the random sampling from the super-population. Also let $(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}})$ be the least squares estimators,

$$(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}) = \arg\min_{\tau,\beta,\gamma} \sum_{i=1}^{N} \left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \overline{X})\right)^2.$$

Theorem 10.2 *Suppose we conduct a pairwise randomized experiment in a sample of pairs drawn at random from the super-population. Then,* (*i*),

$$\tau^* = \tau_{\text{sp}},$$

and (*ii*),

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_{\text{sp}}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^* - \beta^* \cdot \Delta_{X,j} - \gamma^* \cdot (\overline{X}_j - \mu_X)\right)^2\right]\right).$$

Proof of Theorem 10.2 See Appendix.

Now let us estimate the average treatment effect using four different specifications for the regression function. First, for the regression model with only a constant, the least squares estimator for τ is

$$\hat{\tau}^{\text{ols}} = \frac{2}{N} \sum_{j=1}^{N/2} \hat{\tau}^{\text{pair}}(j) = \hat{\tau}^{\text{dif}},$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

equal to the estimator in Equation (10.1). Note that we do not directly include the treatment indicator, because the unit of the least squares analysis here is the pair, not the individual unit. Applying this to the Children's Television experiment data leads to

$$\hat{\tau}^{\text{ols}} = 13.4 \quad (\widehat{\text{s.e.}}\ 4.3)$$

(standard errors in brackets). The next specification for the regression function includes the within-pair difference $\Delta_{X,j}$:

$$\hat{\tau}^{\text{pair}}(j) = \tau + \beta \cdot \Delta_{X,j} + \varepsilon_j.$$

With the Children's Television Workshop data, this specification leads to

$$\hat{\tau}^{\text{pair}}(j) = \underset{(1.5)}{9.0} + \underset{(0.6)}{5.4} \times \Delta_{X,j},$$

with a substantially smaller standard error for $\hat{\tau}^{\text{ols}}$, 1.5 instead of 4.3, because the covariate $\Delta_{X,j}$ is a strong predictor of the observed within-pair difference in outcomes. The next specification includes $\overline{X}_j$ as an additional regressor.

$$\hat{\tau}^{\text{pair}}(j) = \tau + \gamma \cdot \overline{X}_j + \varepsilon_j.$$

This leads to

$$\hat{\tau}^{\text{pair}}(j) = \underset{(3.5)}{13.4} + \underset{(1.7)}{3.9} \times \overline{X}_j.$$

Whereas including $\Delta_{X,j}$ in the regression reduced the standard error of the estimator of the average treatment effect from 4.3 to 1.5, including $\overline{X}_j$ instead of $\Delta_{X,j}$ gives a standard error of 3.5. The final specification includes both $\Delta_{X,j}$ and $\overline{X}_j$, leading to

$$\hat{\tau}^{\text{pair}}(j) = \underset{(1.5)}{8.5} + \underset{(0.8)}{5.9} \times \Delta_{X,j} \underset{(0.7)}{-1.0} \times \overline{X}_j,$$

with again a substantial reduction of the standard error, to 1.5, relative to that using the specification without covariates, but basically the same as the specification that includes only $\Delta_{X,j}$ but not $\overline{X}_j$.

10.7 MODEL-BASED ANALYSIS OF PAIRWISE RANDOMIZED EXPERIMENTS

In principle the model-based imputation approach to the analysis of pairwise randomized experiments is little different from that for the case of stratified randomized experiments. In both cases we can carry out the analysis using the covariate that indicates pair or stratum membership, G_i. In practice, the fact that each pair contains only two units implies that we cannot be as flexible regarding the specification of the joint distribution of the potential outcomes within pairs as would be possible within strata in the stratified

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

case where we have a larger number of units in each stratum. More appropriate is an analysis with some structure on the variance within pairs, such as a hierarchical structure.

The starting point is, as in the chapter on the model-based approach to completely randomized experiments, a model for the joint distribution of the potential outcomes given the covariates, including the pair indicators, in terms of an unknown vector parameter θ:

$$f(\mathbf{Y}(0), \mathbf{Y}(1)|\mathbf{X}, \mathbf{G}, \theta),$$

in combination with a prior distribution on θ, $p(\theta)$. These two components, in combination with the known assignment mechanism, allow us to obtain the joint distribution of the missing potential outcomes $\mathbf{Y}^{\text{mis}}$ given the observed data $(\mathbf{X}, \mathbf{G}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$, and thus allow us to obtain the posterior distribution of the estimand of interest (e.g., the average effect of the treatment).

First we assume that, conditional on $(\mathbf{X}, \mathbf{G}, \mathbf{W})$ and the parameter θ, the potential outcomes are independent by the usual appeal to de Finetti's theorem:

$$f(\mathbf{Y}(0), \mathbf{Y}(1)|\mathbf{X}, \mathbf{G}, \mathbf{W}, \theta) = \prod_{i=1}^{N} f(Y_i(0), Y_i(1)|X_i, G_i, \theta),$$

where we implicitly assume that the parameters governing the marginal distribution of (X_i, G_i) are distinct from θ. The specific model we consider has a hierarchical structure, with pair-specific mean parameters μ_j, for $j = 1, \ldots, J$. Conditional on pair indicators, covariates, and parameters,

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| \; G_i = j, X_i = x, \mu(1), \ldots, \mu(N/2), \gamma, \beta, \sigma_{\text{c}}^2, \sigma_{\text{t}}^2$$

$$\sim \mathcal{N}\left(\left(\begin{array}{c} \mu(j) + x \cdot \beta \\ \mu(j) + \gamma + x \cdot \beta \end{array} \right), \left(\begin{array}{cc} \sigma_{\text{c}}^2 & 0 \\ 0 & \sigma_{\text{t}}^2 \end{array} \right) \right).$$

Conditional on pair-specific mean parameters μ_j, and common parameters γ and β, we assume that the mean of the two potential outcomes is linear in x. We assume the variances are constant across pairs but allow them to differ between potential outcomes. This model is similar in spirit to the regression model where the difference in within-pair observed outcomes was modeled as linear in the difference in within-pair covariate values. Note that given this model, the parameter γ corresponds to the super-population average treatment effect, τ_{sp}. However, in this discussion we focus on inference for the finite-sample average treatment effect, τ_{fs}, by multiply imputing the missing potential outcomes. For that reason, the interpretation of the parameters in the statistical model is incidental.

Next, we specify a model for the pair-specific means μ_j:

$$\left(\begin{array}{c} \mu(1) \\ \vdots \\ \mu(N/2) \end{array} \right) \Bigg| \mathbf{G}, \mathbf{X}, \mathbf{W}, \gamma, \beta, \sigma_{\text{c}}^2, \sigma_{\text{t}}^2, \mu \sim \mathcal{N}\left(\left(\begin{array}{c} \mu \\ \vdots \\ \mu \end{array} \right), \left(\begin{array}{ccc} \sigma_\mu^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_\mu^2 \end{array} \right) \right).$$

Just as in the previous chapter, using simulation methods is generally essential here for the purpose of doing inference. Even in simple cases, there are no analytic expressions

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

Table 10.4. ***Posterior Moments and Quantiles for Youngstown Children's Television Workshop Experiment Data from Table 10.1***

Parameter	Mean	(S.D.)	Quantiles	
			0.025	0.975
γ	8.6	(1.6)	5.1	11.7
β	5.9	(0.6)	4.8	7.0
$\ln(\sigma_c)$	1.1	(0.5)	−0.3	1.9
$\ln(\sigma_t)$	0.5	(0.7)	−0.8	1.7
μ	−9.2	(2.2)	−13.6	−4.7
$\ln(\sigma_\mu)$	1.5	(0.4)	0.4	2.2

for the posterior distributions for estimands of interest in such hierarchical models. However, as we discussed in Chapter 8, this is of no intrinsic importance. Modern Bayesian simulation methods offer efficient algorithms for drawing from the posterior distribution of the estimands given the data. We provide some details in the Appendix for this specific case.

We now implement this model on the Children's Television Workshop data. The single covariate X_i is the pre-test score. We specify independent prior distributions for μ, σ_μ^2, σ_c^2, σ_t^2, γ, and β. For the mean parameters (μ, γ, β), we use normal prior distributions centered at zero, with variance 100^2. For the three variance parameters $(\sigma_\mu^2, \sigma_c^2, \sigma_t^2)$, we use, as we did in Chapter 8, inverse Chi-squared distributions, here with parameters 1 and 1. Based on the Children's Television Workshop data, the posterior mean and variance for the average treatment effect are

$$\mathbb{E}[\tau_{fs}|\mathbf{Y}^{obs}, \mathbf{W}, \mathbf{X}, \mathbf{G}] = 8.4, \qquad \mathbb{V}(\tau_{fs}|\mathbf{Y}^{obs}, \mathbf{W}, \mathbf{X}, \mathbf{G}) = 1.7^2.$$

These estimates are quite similar to those for the regression model with the covariate equal to differences in pre-treatment variables, where we estimated the average effect to be 9.0 with a standard error of 1.5. In Table 10.4 we report posterior means, standard deviations, as well as upper and lower limits for 95% posterior intervals for all parameters.

10.8 CONCLUSION

In this chapter we analyze a special case of stratified randomized experiments: paired randomized experiments. In this special case, each of the strata, now called pairs, contains two units, one assigned to the treatment group and one assigned to the control group. This simplifies some analyses and complicates others. The Fisher exact p-value approach is conceptually not affected by the restrictions on the set of assignments. The Neyman and model-based analyses are modified to take account of the special features of this design. Within each pair there is a natural estimator for the treatment effect, namely the difference in observed outcomes for the treated unit in the pair and the control unit in the same pair. Estimation of the sampling variance for estimators is more complicated in the pairwise randomized experiment because we cannot estimate the sampling variance within each pair separately the way we could estimate the sampling

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

variance within each stratum in the previous chapter on randomized block designs. In the Neyman analysis, we therefore focus on a statistically conservative estimator for the overall sampling variance, based on the sample variance of the within-pair differences. In the regression analyses, the differences between the stratified randomized experiment case and the pairwise randomized experiment case are reflected by the focus on the within-pair difference in outcomes as the dependent variable and the pair as the unit of analysis. Finally, just like in the randomized block design, in the model-based analyses the difference between a completely randomized and a pairwise randomized experiment is reflected by the utility of a hierarchical structure for the latter case.

NOTES

The Children's Television Workshop experiment is discussed in detail in Ball, Bogatz, Rubin, and Beaton (1973). See also Gelman and Hill (2006).

The analysis of pairwise randomized experiments is discussed in detail in standard references on classical experimental design: Hinkelmann and Kempthorne (2005, 2008), Cox and Reid (2000), Cox (1958), and Snedecor and Cochran (1967, 1989). To address the issue of the variance estimation, Lynn and McCulloch (1992) suggest estimating the variance assuming homoskedasticity, ignoring the paired design. See also Donner (1987), Diehr, Martin, Koepsell, and Cheadle (1995). Shipley, Smith, and Dramaix (1989) discuss power calculations for pairwise randomized experiments. Rosenbaum (1989b) analyzes optimal matching strategies to construct matched samples that can then be analyzed using the methods for pairwise randomized experiments discussed in this chapter.

Imai (2008) obtains the same expression for the statistically conservative estimator of the sampling variance as we do in Theorem 10.1.

APPENDIX: PROOFS

Proof of Theorem 10.1

Within each pair we have a completely randomized experiment. Therefore we can use the results on the sampling variance from Chapter 6. This directly implies unbiasedness of $\hat{\tau}^{\text{pair}}(j)$ for $\tau^{\text{pair}}(j)$, and thus unbiasedness of $\hat{\tau}$ for τ_{fs}. This proves part (*i*) of the theorem.

Next consider part (*ii*). The sampling variance expression from Chapter 6 implies

$$\mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = \frac{S_{\text{c}}(j)^2}{N_{\text{c}}(j)} + \frac{S_{\text{t}}^2(j)}{N_{\text{t}}(j)} - \frac{S_{\text{ct}}^2(j)}{N(j)}.$$

With $N(j) = 2$ and $N_{\text{c}}(j) = N_{\text{t}}(j) = 1$, this expression simplifies to

$$\mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = S_{\text{c}}(j)^2 + S_{\text{t}}^2(j) - \frac{S_{\text{ct}}^2(j)}{2}.$$

The within-pair variances can be written as

$$S_{\text{c}}^2(j) = \sum_{i:G_i=j} \left(Y_i(0) - \overline{Y}_j(0)\right)^2,$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

$$S_t^2(j) = \sum_{i:G_i=j} \left(Y_i(1) - \overline{Y}_j(1)\right)^2,$$

and

$$S_{ct}^2(j) = \sum_{i:G_i=j} \left(Y_i(1) - Y_i(0) - \tau^{\text{pair}}(j)\right)^2,$$

where

$$\overline{Y}_j(0) = \frac{1}{2} \cdot \sum_{i:G_i=j} Y_i(0) = \frac{1}{2} \cdot \left(Y_{j,A}(0) + Y_{j,B}(0)\right),$$

and

$$\overline{Y}_j(1) = \frac{1}{2} \cdot \sum_{i:G_i=j} Y_i(1) = \frac{1}{2} \cdot \left(Y_{j,A}(1) + Y_{j,B}(1)\right).$$

Because pair j comprises two units, indexed by A and B, we can rewrite these expressions as

$$S_c^2(j) = \frac{1}{2} \cdot \left(Y_{j,A}(0) - Y_{j,B}(0)\right)^2, \qquad S_t^2(j) = \frac{1}{2} \cdot \left(Y_{j,A}(1) - Y_{j,B}(1)\right)^2,$$

and

$$S_{ct}^2(j) = \frac{1}{2} \cdot \left(\left(Y_{j,A}(1) - Y_{j,A}(0)\right) - \left(Y_{j,B}(1) - Y_{j,B}(0)\right)\right)^2.$$

Hence the sampling variance of $\hat{\tau}^{\text{dif}} = (2/N)\sum_{j=1}^{N/2} \hat{\tau}^{\text{pair}}(j)$ is

$$\mathbb{V}_W(\hat{\tau}^{\text{dif}}) = \frac{4}{N^2} \sum_{j=1}^{N/2} \mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = \frac{4}{N^2} \sum_{j=1}^{N/2} \left(S_c(j)^2 + S_t^2(j) - \frac{S_{ct}^2(j)}{2}\right).$$

Substituting for $S_c^2(j)$, $S_t^2(j)$, and $S_{ct}^2(j)$ leads to

$$\mathbb{V}_W(\hat{\tau}^{\text{dif}}) = \frac{1}{N^2} \sum_{j=1}^{N/2} \Big(2 \cdot \left(Y_{j,A}(0) - Y_{j,B}(0)\right)^2 + 2 \cdot \left(Y_{j,A}(1) - Y_{j,B}(1)\right)^2$$
$$- \left(\left(Y_{j,A}(1) - Y_{j,A}(0)\right) - \left(Y_{j,B}(1) - Y_{j,B}(0)\right)\right)^2\Big),$$

which simplifies to

$$\mathbb{V}_W(\hat{\tau}^{\text{dif}}) = \frac{1}{N^2} \sum_{j=1}^{N/2} \left(Y_{j,A}(0) + Y_{j,A}(1) - \left(Y_{j,B}(0) + Y_{j,B}(1)\right)\right)^2.$$

Finally, consider part (*iii*). If the treatment effect is constant, then $Y_{j,A}(1) = Y_{j,A}(0) + \tau$ and $Y_{j,B}(1) = Y_{j,B}(0) + \tau$ for all j. Hence the expression for the sampling variance

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

simplifies to

$$\begin{aligned}\mathbb{V}_W(\hat{\tau}) &= \frac{1}{N^2}\sum_{j=1}^{N/2}\left(Y_{j,A}(0)+Y_{j,A}(1)-\left(Y_{j,B}(0)+Y_{j,B}(1)\right)\right)^2 \\ &= \frac{1}{N^2}\sum_{j=1}^{N/2}\left(2\cdot Y_{j,A}(0)+\tau-\left(2\cdot Y_{j,B}(0)+\tau\right)\right)^2 \\ &= \frac{4}{N^2}\sum_{j=1}^{N/2}\left(Y_{j,A}(0)-Y_{j,B}(0)\right)^2.\end{aligned}$$

Now consider the variance estimator $\hat{\mathbb{V}}^{\text{pair}}$,

$$\hat{\mathbb{V}}^{\text{pair}} = \frac{4}{N\cdot(N-2)}\cdot\sum_{j=1}^{N/2}\left(\hat{\tau}^{\text{pair}}(j)-\hat{\tau}\right)^2.$$

We calculate the expectation of $\hat{\mathbb{V}}^{\text{pair}}$. Note that

$$\mathbb{E}_W\left[\hat{\tau}^{\text{pair}}(j)\right] = \tau^{\text{pair}}(j),$$

and

$$\mathbb{E}_W\left[\hat{\tau}^{\text{pair}}(j)\cdot\hat{\tau}^{\text{pair}}(k)\right] =$$

$$\begin{cases}\tau^{\text{pair}}(j)\cdot\tau^{\text{pair}}(k) & \text{if } j\neq k, \\ \tau^{\text{pair}}(j)^2+\frac{1}{4}\cdot\left(Y_{j,A}(0)+Y_{j,A}(1)-\left(Y_{j,B}(0)+Y_{j,B}(1)\right)\right)^2 & \text{if } j=k.\end{cases}$$

Then:

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{N/2}\left(\hat{\tau}^{\text{pair}}(j)-\hat{\tau}^{\text{dif}}\right)^2\right] &= \mathbb{E}\left[\sum_{j=1}^{N/2}\left(\hat{\tau}^{\text{pair}}(j)-\frac{2}{N}\cdot\sum_{k=1}^{N/2}\hat{\tau}^{\text{pair}}(k)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j)^2-\frac{4}{N}\cdot\sum_{j=1}^{N/2}\sum_{k=1}^{N/2}\hat{\tau}^{\text{pair}}(j)\cdot\hat{\tau}^{\text{pair}}(k)+\frac{2}{N}\left(\sum_{k=1}^{N/2}\hat{\tau}^{\text{pair}}(k)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j)^2-\frac{4}{N}\cdot\sum_{j=1}^{N/2}\sum_{k=1}^{N/2}\hat{\tau}^{\text{pair}}(j)\cdot\hat{\tau}^{\text{pair}}(k)+\frac{2}{N}\sum_{j=1}^{N/2}\sum_{k=1}^{N/2}\hat{\tau}^{\text{pair}}(j)\cdot\hat{\tau}^{\text{pair}}(k)\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j)^2-\frac{4}{N}\cdot\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j)^2-\frac{4}{N}\cdot\sum_{j=1}^{N/2}\sum_{k\neq j}\hat{\tau}^{\text{pair}}(j)\cdot\hat{\tau}^{\text{pair}}(k)\right. \\ &\quad \left.+\frac{2}{N}\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j)^2+\frac{2}{N}\sum_{j=1}^{N/2}\sum_{k\neq j}\hat{\tau}^{\text{pair}}(j)\cdot\hat{\tau}^{\text{pair}}(k)\right]\end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

$$= \frac{N-2}{N} \cdot \mathbb{E}\left[\sum_{j=1}^{N/2} \hat{\tau}^{\text{pair}}(j)^2\right] - \frac{2}{N} \cdot \mathbb{E}\left[\sum_{j=1}^{N/2} \sum_{k \neq j} \hat{\tau}^{\text{pair}}(j) \cdot \hat{\tau}^{\text{pair}}(k)\right]$$

$$= \frac{N-2}{N} \cdot \sum_{j=1}^{N/2} \tau^{\text{pair}}(j)^2 - \frac{2}{N} \cdot \sum_{j=1}^{N/2} \sum_{k \neq j} \tau^{\text{pair}}(j) \cdot \tau^{\text{pair}}(k)$$

$$+ \frac{N-2}{4 \cdot N} \cdot \sum_{j=1}^{N/2} \left(Y_{j,A}(0) + Y_{j,A}(1) - \left(Y_{j,B}(0) + Y_{j,B}(1)\right)\right)^2$$

$$= \sum_{j=1}^{N/2} \left(\tau^{\text{pair}}(j) - \tau_{\text{S}}\right)^2 + \frac{N \cdot (N-2)}{4} \cdot \mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right).$$

Thus,

$$\mathbb{E}\left[\hat{\mathbb{V}}^{\text{pair}}\right] = \mathbb{E}\left[\frac{4}{N \cdot (N-2)} \cdot \sum_{j=1}^{N/2} \left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2\right]$$

$$= \frac{4}{N \cdot (N-2)} \cdot \left(\sum_{j=1}^{N/2} \left(\tau^{\text{pair}}(j) - \tau_{\text{fs}}\right)^2 + \frac{N \cdot (N-2)}{4} \cdot \mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right)\right)$$

$$= \mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) + \frac{4}{N \cdot (N-2)} \cdot \sum_{j=1}^{N/2} \left(\tau^{\text{pair}}(j) - \tau_{\text{fs}}\right)^2.$$

□

Proof of Theorem 10.2

First let us expand the expectation:

$$\mathbb{E}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right] \tag{A.1}$$

$$= \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right]\right]$$

$$= \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\tau^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right]\right] \tag{A.2}$$

$$+ \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right)^2\right]\right]$$

$$+ 2 \cdot \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right) \cdot \left(\tau^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)\right]\right].$$

Consider the three terms separately.

The first term equals

$$\mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\tau^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right]\right]$$

$$= \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\tau^{\text{pair}}(j) - \tau - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right]\right] + \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\beta \cdot \Delta_{X,j}\right)^2\right]\right]$$

$$- 2 \cdot \beta \cdot \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\Delta_{X,j}\left(\tau^{\text{pair}}(j) - \tau - \gamma \cdot (\overline{X}_j - \mu_X)\right)\right]\right]$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

$$
\begin{aligned}
&= \mathbb{E}_{\text{sp}}\left[\left(\tau^{\text{pair}}(j) - \tau - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right] + \beta^2 \cdot \mathbb{E}\left[\left(\Delta_{X,j}\right)^2\right] \\
&= \mathbb{E}_{\text{sp}}\left[\left(\tau^{\text{pair}}(j) - \tau\right)^2\right] + \mathbb{E}_{\text{sp}}\left[\left(\gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right] \\
&\quad - 2 \cdot \mathbb{E}_{\text{sp}}\left[\left(\tau^{\text{pair}}(j) - \tau\right) \cdot \left(\gamma \cdot (\overline{X}_j - \mu_X)\right)\right] + \beta^2 \cdot \mathbb{E}\left[\left(\Delta_{X,j}\right)^2\right] \\
&= \mathbb{E}_{\text{sp}}\left[\left(\tau^{\text{pair}}(j) - \tau\right)^2\right] + \gamma^2 \cdot \mathbb{E}_{\text{sp}}\left[\left(\overline{X}_j - \mu_X\right)^2\right] \\
&\quad - 2 \cdot \mathbb{E}_{\text{sp}}\left[\tau^{\text{pair}}(j) \cdot \left(\gamma \cdot (\overline{X}_j - \mu_X)\right)\right] + \beta^2 \cdot \mathbb{E}\left[\left(\Delta_{X,j}\right)^2\right].
\end{aligned}
$$

The second term equals

$$
\begin{aligned}
&\mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right)^2\right]\right] \\
&\quad = \mathbb{E}_{\text{sp}}\left[\frac{1}{4} \cdot \left(Y_{j,A}(0) + Y_{j,A}(1) - \left(Y_{j,B}(0) + Y_{j,B}(1)\right)\right)^2\right],
\end{aligned}
$$

which does not depend on the parameters (τ, β, γ), and therefore can be ignored for the purpose of determining the minimand of the objective function (A.1).

The third term equals

$$
\begin{aligned}
&2 \cdot \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right) \cdot \left(\tau^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)\right]\right] \\
&\quad = -2 \cdot \beta \cdot \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right) \cdot \Delta_{X,j}\right]\right] \\
&\quad = -2 \cdot \beta \cdot \mathbb{E}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right) \cdot \Delta_{X,j}\right].
\end{aligned}
$$

Collecting the terms that depend on (τ, β, γ) leads to

$$
\begin{aligned}
&= \mathbb{E}_{\text{sp}}\left[\left(\tau^{\text{pair}}(j) - \tau\right)^2\right] + \gamma^2 \cdot \mathbb{E}_{\text{sp}}\left[\left(\overline{X}_j - \mu_X\right)^2\right] \\
&\quad - 2 \cdot \gamma \cdot \mathbb{E}_{\text{sp}}\left[\tau^{\text{pair}}(j) \cdot \left((\overline{X}_j - \mu_X)\right)\right] + \beta^2 \cdot \mathbb{E}\left[\left(\Delta_{X,j}\right)^2\right] \\
&\quad - 2 \cdot \beta \cdot \mathbb{E}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right) \cdot \Delta_{X,j}\right].
\end{aligned}
$$

Minimizing this over (τ, β, γ) leads to

$$
\tau^* = \mathbb{E}_{\text{sp}}\left[\tau^{\text{pair}}(j)\right] = \tau_{\text{sp}},
$$

$$
\gamma^* = \frac{\mathbb{E}_{\text{sp}}\left[\tau^{\text{pair}}(j) \cdot \left((\overline{X}_j - \mu_X)\right)\right]}{\mathbb{E}_{\text{sp}}\left[\left(\overline{X}_j - \mu_X\right)^2\right]}, \quad \text{and} \quad \beta^* = \frac{\mathbb{E}_{\text{sp}}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^{\text{pair}}(j)\right) \cdot \Delta_{X,j}\right]}{\mathbb{E}_{\text{sp}}\left[\left(\Delta_{X,j}\right)^2\right]}.
$$

Next, consider part (*ii*) of the theorem:

$$
(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}) = \arg\min_{\tau,\beta,\gamma} \sum_{i=1}^{N} \left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \overline{X})\right)^2. \tag{A.3}
$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press

Define $\Delta_{Y,j} = \hat{\tau}^{\text{pair}}(j)$ and $\hat{\mu} = \overline{X}$. The first-order conditions for the estimators $(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}, \hat{\mu}^{\text{ols}})$ in the minimization problem (A.3) are

$$\sum_{j=1}^{N/2} \psi(\Delta_{Y,j}, \Delta_{X,j}, \overline{X}_j, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}, \hat{\mu}^{\text{ols}}) = 0,$$

where

$$\psi(\Delta_y, \Delta_x, \overline{x}, \tau, \beta, \gamma, \mu) = \begin{pmatrix} \Delta_y - \tau - \beta \cdot \Delta_x - \gamma \cdot (\overline{x} - \mu) \\ \Delta_x \cdot \left(\Delta_y - \tau - \beta \cdot \Delta_x - \gamma \cdot (\overline{x} - \mu)\right) \\ (\overline{x} - \mu) \cdot \left(\Delta_y - \tau - \beta \cdot \Delta_x - \gamma \cdot (\overline{x} - \mu)\right) \\ \overline{x} - \mu \end{pmatrix}.$$

By the same arguments as used in the proofs in Chapter 7,

$$\sqrt{N} \cdot \begin{pmatrix} \hat{\tau}^{\text{ols}} - \tau_{\text{sp}} \\ \hat{\beta}^{\text{ols}} - \beta^* \\ \hat{\gamma}^{\text{ols}} - \gamma^* \\ \hat{\mu}^{\text{ols}} - \mu_X \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Gamma^{-1} \Delta (\Gamma')^{-1} \right),$$

where the two components of the covariance matrix are

$$\Gamma = \mathbb{E}\left[\frac{\partial}{\partial(\tau, \beta, \gamma, \mu)} \psi\left(\Delta_{Y,j}, \Delta_{X,j}, \overline{X}_j, \tau, \beta, \gamma, \mu\right)\right]\Bigg|_{(\tau_{\text{sp}}, \beta^*, \gamma^*, \mu_X)}$$

and

$$\Delta = \mathbb{E}\left[\psi\left(\Delta_{Y,j}, \Delta_{X,j}, \overline{X}_j, \tau_{\text{sp}}, \beta^*, \gamma^*, \mu_X\right) \cdot \psi\left(\Delta_{Y,j}, \Delta_{X,j}, \overline{X}_j, \tau_{\text{sp}}, \beta^*, \gamma^*, \mu_X\right)' \right].$$

$$\Gamma = \begin{pmatrix} -1 & -\mathbb{E}\left[\Delta_{X,j}\right] & -\mathbb{E}\left[\overline{X} - \mu_X\right] & \gamma^* \\ -\mathbb{E}\left[\Delta_{X,j}\right] & -\mathbb{E}\left[\Delta_{X,j}^2\right] & -\mathbb{E}\left[\Delta_{X,j} \cdot (\overline{X} - \mu_X)\right] & \gamma^* \cdot \mathbb{E}\left[\Delta_{X,j}\right] \\ -\mathbb{E}\left[\overline{X} - \mu_X\right] & -\mathbb{E}\left[\Delta_{X,j} \cdot (\overline{X} - \mu_X)\right] & -\mathbb{E}\left[(\overline{X} - \mu_X)^2\right] & 2 \cdot \gamma^* \cdot \mathbb{E}\left[\overline{X}_j - \mu_X\right] \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} -1 & 0 & 0 & \gamma^* \\ 0 & -\mathbb{E}\left[\Delta_{X,j}^2\right] & 0 & 0 \\ 0 & 0 & -\mathbb{E}\left[(\overline{X}_j - \mu_X)^2\right] & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Thus $\mathbb{V}(\hat{\tau}^{\text{ols}})$, the $(1, 1)$ element of $\Gamma^{-1}\Delta(\Gamma')^{-1}$, is equal to $\Delta_{11} - \gamma^* \cdot \Delta_{14}$, where Δ_{km} is the (k, m) element of Δ. Because

$$\Delta_{14} = \mathbb{E}\left[\left(\Delta_{Y,j} - \tau_{\text{sp}} - \beta^* \cdot \Delta_{X,j} - \gamma^* \cdot (\overline{X}_j - \mu_X)\right) \cdot \left(\overline{X}_j - \mu_X\right)\right] = 0,$$

it follows that the $(1, 1)$ element of $\Gamma^{-1}\Delta(\Gamma')^{-1}$ is equal to

$$\mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{ols}}) = \Delta_{11} = \mathbb{E}\left[\left(\Delta_{Y,j} - \tau_{\text{sp}} - \beta^* \cdot \Delta_{X,j} - \gamma^* \cdot (\overline{x} - \mu_X)\right)^2\right].$$

https://doi.org/10.1017/CBO9781139025751.011 Published online by Cambridge University Press