CHAPTER 9

Stratified Randomized Experiments

9.1 INTRODUCTION

The focus in the previous chapters in Part II was on completely randomized experiments, where, in a fixed sample with N units, N_t are randomly chosen to receive the active treatment and the remaining $N_c = N - N_t$ are assigned to receive the control treatment. We considered four modes of inference: Fisher's exact p-values and associated intervals, Neyman's unbiased estimates and repeated sampling-based large-N confidence intervals, regression methods, and model-based imputation. In addition, we considered the benefits of observing covariates, that is, measurements on the units unaffected by the treatments, such as pre-treatment characteristics. In this chapter we consider the same issues for a different class of randomized experiments, stratified randomized experiments, also referred to as randomized blocks experiments to use the terminology of classical experimental design. In stratified randomized experiments, units are stratified (or grouped or blocked) according to the values of (a function of) the covariates. Within the strata, independent completely randomized experiments are conducted but possibly with different relative sizes of treatment and control groups.

Part of the motivation for considering alternative structures for randomized experiments is interest in such experiments per se. But there are other, arguably equally important reasons. In the discussion of observational studies in Parts III, IV, V, and VI of this text, we consider methods for (non-randomized) observational data that can be viewed in some way as analyzing the data as if they arose from hypothetical stratified randomized experiments. Understanding these methods in the context of randomized experiments will aid their interpretation and implementation in observational studies.

The main part of this chapter describes how the methods developed in the previous four chapters can be modified to apply in the context of stratified randomized experiments. In most cases these modifications are conceptually straightforward. We also discuss some design issues in relation to stratification. Specifically, we assess the benefits of stratification relative to complete randomization.

In the next section we describe the data used to illustrate the concepts discussed in this chapter. These data are from a randomized experiment designed to evaluate the effect of class size on academic achievement, known as Project Star. In Section 9.3 we discuss the general structure of stratified randomized experiments. In the next four sections

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

we discuss the four approaches we described previously for completely randomized experiments: in Section 9.4 the Fisher exact p-value approach; in Section 9.5 the Neyman approach; in Section 9.6 the regression approach; and in 9.7 the model-based imputation approach. Next, in Section 9.8, we discuss design issues and specifically the common benefits of stratified randomized experiments over completely randomized experiments. Section 9.9 concludes.

9.2 THE TENNESEE PROJECT STAR DATA

We illustrate the methods for randomized block experiments using data from a randomized evaluation of the effect of class size on test scores conducted in 1985–1986 in Tennessee called the Student/Teacher Achievement Ratio experiment, or Project Star for short. This was a very influential experiment; Mosteller (1995) calls it "one of the most important educational investigations ever carried out." In this chapter we use the kindergarten data from schools where students and teachers were randomly assigned to small classes (13–17 students per teacher), to regular classes (22–25 students per teacher), or to regular classes with a teacher's aide. To be eligible for Project Star, a school had to have a sufficient number of students to allow the formation of at least one class of each of the three types. Once a school had been admitted to the program, a decision was made on the number of classes of each type (small, regular, regular with aide). We take as fixed the number of classes of each type in each school. The unit of analysis is the teacher or class, rather than the individual student, to help justify the no-interference part of SUTVA.

The experiment is somewhat different from those we have discussed before, so we will be precise in its description. A school has a pool of at least 57 students, so they could support at least one small and two regular-sized classes. Two separate and independent randomizations took place. One random assignment is that of teachers to classes of different types, small, regular, or regular with aide. The second randomization is of students to classes/teachers. In our analysis, we mainly rely on the first randomization, of class-size and aides to teachers, using the teachers as the units of analysis. Irrespective of the assignments of students to classes, the resulting inferences are valid for the effect on the teachers of being assigned to a particular type of class. However, the second randomization is important for the interpretation of the results. Suppose we find that assignment to a small class leads on average to better outcomes for the teacher. Without the randomization of students to classes, this could be due to systematic assignment of better students to the smaller classes. With the second randomization, this is ruled out, and systematic effects can be interpreted as the effects of class size. This type of double randomization is somewhat similar to that in "split plot" designs (Cochran and Cox, 1957), although in split plot designs two different treatments are being applied by the double randomization.

Given the structure of the experiment, one could also focus on students as the unit of analysis, and investigate effects of class size on student-level outcomes. The concern, however, is that the Stable Unit Treatment Value Assumption (SUTVA) is not plausible in that case. Violations of SUTVA complicate the Neyman, regression, and imputation approaches considerably, and we therefore primarily focus on class-level

(i.e., teacher-level) analyses in this chapter. As we see in Section 9.4.4, however, it remains straightforward to use the FEP approach to test the null hypothesis that assignment of students to different classes had no effect on test scores whatsoever, because SUTVA is automatically satisfied under Fisher's sharp null hypothesis of no effects of the treatment.

In the analyses in this chapter, we focus on the comparison between regular (control) and small (treated) classes, and ignore the data for regular classes with teachers' aides. We discard schools that do not have at least two classes of both the small size and the regular size. Focusing on schools with at least two regular classes and two small classes leaves us with sixteen schools, which creates sixteen strata or blocks. Most have exactly two classes of each size, but one has two regular classes and four small classes, and two other schools have three small classes and two regular-sized classes. The total number of teachers and classes in this reduced data set is $N = 68$. Out of these 68 teachers, $N_c = 32$ are assigned to regular-sized classes, and $N_t = 36$ are assigned to small classes. Outcomes are defined at the class (i.e., teacher) level. The class-level outcomes we focus on are averages of test scores over all students for their teacher. One can, however, consider other outcomes, such as median test score of the students with a specific teacher or measures of within-teacher dispersion. The specific outcome we analyze here is the class average score on a mathematics test. The individual student scores were normalized to have mean equal to zero and standard deviation equal to one across all the students in the reduced data set. These individual scores then ranged from a minimum of -4.13 to a maximum of 2.94. The averages for each of the 68 classes in our analysis are reported in Table 9.1, organized by school. Overall, the average for the regular classes is -0.13 with a standard deviation of 0.56, and the average for the small classes is 0.09 with a standard deviation of 0.61. We return to these data after introducing methods for the analysis of such studies.

9.3 THE STRUCTURE OF STRATIFIED RANDOMIZED EXPERIMENTS

In stratified randomized experiments, units are grouped together according to some pre-treatment characteristics into strata. Within each stratum, a completely randomized experiment is conducted, and thus, within each stratum, the methods discussed in Chapters 5–8 are directly applicable. However, the interest is not about hypotheses or treatment effects within a single stratum, but rather it is about hypotheses and treatment effects across all strata. Moreover, the sample sizes are often such that we cannot obtain precise estimates of typical treatment effects within any one stratum. Here we discuss how the methods developed previously can be adapted to take account of the additional structure of the experiment.

9.3.1 The Case with Two Strata

As before, we are interested both in assessing null hypotheses concerning treatment effects and in estimating typical treatment effects (usually the average). First we focus

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

Table 9.1. *Class Average Mathematics Scores from Project Star*

School/ Stratum	No. of Classes	Regular Classes ($W_i = 0$)	Small Classes ($W_i = 1$)
1	4	−0.197, 0.236	0.165, 0.321
2	4	0.117, 1.190	0.918, −0.202
3	5	−0.496, 0.225	0.341, 0.561, −0.059
4	4	−1.104, −0.956	−0.024, −0.450
5	4	−0.126, 0.106	−0.258, −0.083
6	4	−0.597, −0.495	1.151, 0.707
7	4	0.685, 0.270	0.077, 0.371
8	6	−0.934, −0.633	−0.870, −0.496, −0.444, 0.392
9	4	−0.891, −0.856	−0.568, −1.189
10	4	−0.473, −0.807	−0.727, −0.580
11	4	−0.383, 0.313	−0.533, 0.458
12	5	0.474, 0.140	1.001, 0.102, 0.484
13	4	0.205, 0.296	0.855, 0.509
14	4	0.742, 0.175	0.618, 0.978
15	4	−0.434, −0.293	−0.545, 0.234
16	4	0.355, −0.130	−0.240, −0.150
Average		−0.13	0.09
(S.D.)		(0.56)	(0.61)

on the case with the sample of N units divided into two subsamples, for example, females (f) and males (m), with subsample size $N(f)$ and $N(m)$, respectively, so that $N = N(f) + N(m)$. To fit the division into two subsamples into the structure developed so far, it is useful to associate with each unit a binary covariate (e.g., the unit's sex) with the membership in strata based on this covariate. Although in general in this text we use the notation X_i for the covariate for unit i, here we use the notation G_i for this particular covariate that determines stratum or group membership, with $\mathbf{B}$ denoting the N-component vector with typical element B_i. As with any other covariate, the value of G_i is not affected by the treatment. In this example G_i takes on the values f and m. Define $\tau_{\rm fs}(f)$ and $\tau_{\rm fs}(m)$ to be the finite-sample average treatment effects in the two strata:

$$\tau_{\rm fs}(f) = \frac{1}{N(f)} \sum_{i:G_i=f} \big(Y_i(1) - Y_i(0)\big), \quad \text{and} \quad \tau_{\rm fs}(m) = \frac{1}{N(m)} \sum_{i:G_i=m} \big(Y_i(1) - Y_i(0)\big).$$

Within each stratum, we conduct a completely randomized experiment with $N_{\rm t}(f)$ and $N_{\rm t}(m)$ units assigned to the active treatment in the two subsamples respectively, and the remaining $N_{\rm c}(f) = N(f) - N_{\rm t}(f)$ and $N_{\rm c}(m) = N(m) - N_{\rm t}(m)$ units assigned to the control treatment. It need not be the case that the proportion of treated units, the propensity score, $e(f) = N_{\rm t}(f)/N(f)$ and $e(m) = N_{\rm t}(m)/N(m)$ for the female and male subpopulations, respectively, is the same in both subpopulations. Let $N_{\rm t} = N_{\rm t}(f) + N_{\rm t}(m)$ be the total number of units assigned to the treatment group, and $N_{\rm c} = N_{\rm c}(f) + N_{\rm c}(m)$ be the

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

total number of units assigned to the control group. Let us consider the assignment mechanism. Within the $G_i = f$ subpopulation, $N_t(f)$ units out of $N(f)$ are randomly chosen to receive the treatment. There are $\binom{N(f)}{N_t(f)}$ such allocations. For every allocation for the set of units with $G_i = m$, there are $\binom{N(m)}{N_t(m)}$ ways of choosing $N_t(m)$ units with $G_i = m$ to receive the treatment out of $N(m)$ units. All of these allocations are equally likely. Combining these two assignment vectors, the assignment mechanism for a stratified randomized experiment with two strata can be written as

$$\Pr(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{B}) = \binom{N(f)}{N_t(f)}^{-1} \cdot \binom{N(m)}{N_t(m)}^{-1} \quad \text{for } \mathbf{W} \in \mathbb{W}^+,$$

$$\text{where } \mathbb{W}^+ = \left\{ \mathbf{W} \text{ such that } \sum_{i:G_i=f} W_i = N_t(f), \sum_{i:G_i=m} W_i = N_t(m) \right\}.$$

Compare the assignment mechanism for a stratified randomized experiment to that for a completely randomized experiment with $N_t = N_t(f) + N_t(m)$ assigned to treatment and $N_c = N(f) - N_t(f) + N(m) - N_t(m)$ assigned to control. Many assignment vectors that would have positive probability with a completely randomized experiment have probability zero with the stratified randomized experiment: all vectors with $\sum_{i=1}^{N} W_i = N_t(f) + N_t(m)$ but $\sum_{i:G_i=f} W_i \neq N_t(f)$ (or, equivalently, $\sum_{i:G_i=m} W_i \neq N_t(m)$). If $N_t(f)/N(f) \approx N_t(m)/N(m)$, the stratification rules out substantial imbalances in the covariate distributions in the two treatment groups that could arise by chance in a completely randomized experiment. The possible disadvantage of the stratification is that a large number of possible assignment vectors are eliminated, just as a completely randomized experiment eliminates assignment vectors that would be allowed under Bernoulli trials (where assignment for each unit is determined independently of assignment for any other unit). The advantage of a completely randomized experiment over a Bernoulli trial for drawing causal inferences was argued to be the relative lack of information on treatment effects of the eliminated assignment vectors, typically those assignment vectors with a severe imbalance between the number of controls and the number of treated.

Here the argument is similar, although not quite as obvious. If we were to partition the population randomly into strata, the assignment vectors eliminated by the stratification are in expectation as helpful as the ones included, and the stratification will not produce a more informative experiment. However, if the stratification is based on characteristics that are associated with the outcomes of interest, we shall see that stratified randomized experiments generally are more informative than completely randomized experiments. For example, in many drug trials, one may expect systematic differences in typical outcomes, both given the drug and without the drug, for men and women. In that case, conducting the experiment by stratifying the population into males and females, rather than conducting a completely randomized experiment, makes eminent sense. It can lead to more precise inferences, by eliminating the possibility of assignments with severe imbalances in sex distribution – for example, the extreme and uninformative assignment with all women exposed to the active treatment and all men exposed to the control treatment.

9.3.2 The Case with J Strata

Here we generalize the notation to the situation with multiple strata. Let J be the number of strata, and $N(j)$, $N_c(j)$, and $N_t(j)$ the total number of units, and the number of control and treated units in strata j, respectively, for $j = 1, \ldots, J$. Let $G_i \in \{1, \ldots, J\}$ denote the stratum for unit i, and let $B_i(j) = \mathbf{1}_{G_i=j}$, be the indicator that is equal to one if unit i is in stratum j, and zero otherwise. Within stratum j there are now $\binom{N(j)}{N_t(j)}$ possible assignments, so that the assignment mechanism is

$$\Pr(\mathbf{W}|\mathbf{B}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \qquad \text{for } \mathbf{W} \in \mathbb{W}^+,$$

where $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} | \sum_{i=1}^{N} B_i(j) \cdot W_i = N_t(j) \text{ for } j = 1, \ldots, J\}$.

9.4 FISHER'S EXACT P-VALUES IN STRATIFIED RANDOMIZED EXPERIMENTS

In stratified randomized experiments, just as in completely randomized experiments, the assignment mechanism is completely known. Hence, given a sharp null hypothesis that specifies all unobserved potential outcomes given knowledge of the observed outcomes, we can directly apply Fisher's approach to calculate exact p-values as discussed in Chapter 5. Let us focus on Fisher's sharp null hypothesis that all treatment effects are zero: $H_0: \ Y_i(0) = Y_i(1)$ for $i = 1, 2, \ldots, N$. For ease of exposition, we focus initially on the case with two strata, $G_i \in \{f, m\}$.

9.4.1 The Choice of Statistics in the FEP Approach with Two Strata

Let $\overline{Y}_c^{\text{obs}}(j)$ and $\overline{Y}_t^{\text{obs}}(j)$ be the average observed outcome for units in stratum j (currently, in the two-stratum example for $j \in \{f, m\}$, later, in the general J-stratum case for $j = 1, \ldots, J$) in the control and treatment groups, and let $e(j)$ be the propensity score:

$$\overline{Y}_c^{\text{obs}}(j) = \frac{1}{N_c(j)} \sum_{i:G_i=j} (1 - W_i) \cdot Y_i^{\text{obs}}, \quad \overline{Y}_t^{\text{obs}}(j) = \frac{1}{N_t(j)} \sum_{i:G_i=j} W_i \cdot Y_i^{\text{obs}},$$

and

$$e(j) = N_t(j)/N(j).$$

Obvious statistics are the absolute value of the difference in the average observed outcome for treated and control units in the first and in the second stratum:

$$T^{\text{dif}}(f) = \left| \overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c(f)^{\text{obs}} \right| \quad \text{and} \quad T^{\text{dif}}(m) = \left| \overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m) \right|.$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

Neither of the statistics, $T^{\text{dif}}(f)$ or $T^{\text{dif}}(m)$, is particularly attractive by itself: for either one an entire stratum is ignored, and thus the test would not be sensitive to violations of the null hypothesis in the stratum that is ignored.

A more appealing statistic is based on the combination of the two within-stratum statistics, $T^{\text{dif}}(f)$ and $T^{\text{dif}}(m)$, for example, the absolute value of a convex combination of the two differences in averages,

$$T^{\text{dif},\lambda} = \left| \lambda \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) \right) + (1-\lambda) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right) \right|,$$

for some $\lambda \in [0,1]$. For any fixed value of λ, we can use the same FEP approach and find the randomized distribution of the statistic under the null hypothesis, and thus calculate the corresponding p-value. The question is what would be an attractive choice for λ? An obvious choice for λ is to weight the two differences $T^{\text{dif}}(f)$ and $T^{\text{dif}}(m)$ by the relative sample sizes (RSS) in the strata and choose $\lambda = \lambda_{\text{RSS}} \equiv N(f)/(N(f)+N(m))$. If the relative proportions of treated and control units in each stratum, $N_{\text{t}}(f)/N(f)$ and $N_{\text{t}}(m)/N(m)$ respectively, are similar, then the stratification from our stratified experiment is close to the stratification from a completely randomized experiment. In that case, this choice for the weight parameter λ_{RSS} would lead to the natural statistic that is common in a completely randomized experiment,

$$T^{\text{dif},\lambda_{\text{RSS}}} = \left| \frac{N(f)}{N(f)+N(m)} \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) \right) + \frac{N(m)}{N(f)+N(m)} \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right) \right|.$$

If the relative proportions of treated and control units are very different, however, this choice for λ does not necessarily lead to a very powerful test statistic. Suppose, for example, that both strata contain fifty units, where in stratum f, only a single unit gets assigned to treatment, and the remaining forty-nine units get assigned to control, whereas in stratum m, the number of treated and control units is twenty-five. In that case, the test based on $T^{\text{dif}}(m)$ is likely to have substantially more power than the test based on $T^{\text{dif}}(f)$. Combining $T^{\text{dif}}(f)$ and $T^{\text{dif}}(m)$ by the relative share of the two strata in the population, thereby giving both stratum-specific average observed outcome differences $\hat{\tau}(f)$ and $\hat{\tau}(m)$ equal weight, would lead to a test statistic with poor power properties because it gives equal weight to the f stratum that is characterized by a severe imbalance in the proportions of treated and control units.

An alternative choice for λ is motivated by considering against which alternative hypotheses we would like our test statistic to have power. Often an important alternative hypothesis has a treatment effect that is constant both within and between strata. To obtain a more attractive choice for λ based on this perspective, it is useful to consider the sampling variances of the two stratum-specific statistics, $T^{\text{dif}}(f)$ and $T^{\text{dif}}(m)$, under Neyman's repeated sampling perspective. Applying the results from Chapter 5, we find that under the randomization distribution, the sampling variance of the two within-stratum estimates of the average treatment effects are

$$\mathbb{V}_W \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) \right) = \frac{S_{\text{t}}^2(f)}{N_{\text{t}}(f)} + \frac{S_{\text{c}}^2(f)}{N_{\text{c}}(f)} - \frac{S_{\text{ct}}(f)^2}{N(f)},$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

and

$$\mathbb{V}_W\left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(m)-\overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(m)\right)=\frac{S_{\mathrm{t}}^2(m)}{N_{\mathrm{t}}(m)}+\frac{S_{\mathrm{c}}^2(m)}{N_{\mathrm{c}}(m)}-\frac{S_{\mathrm{ct}}^2(m)}{N(m)}.$$

Suppose that, within the strata, the treatment effects are constant. In that case, $S_{\mathrm{ct}}^2(f) = S_{\mathrm{ct}}^2(m) = 0$, and the last term drops from both expressions. Assume, in addition, that all four variances $S_{\mathrm{c}}^2(f)$, $S_{\mathrm{t}}^2(f)$, $S_{\mathrm{c}}^2(m)$, and $S_{\mathrm{t}}^2(m)$ are equal to S^2. Then the sampling variances of the two observed differences are

$$\mathbb{V}_W\left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(f)-\overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(f)\right)=S^2\cdot\left(\frac{1}{N_{\mathrm{t}}(f)}+\frac{1}{N_{\mathrm{c}}(f)}\right),$$

and

$$\mathbb{V}_W\left(\overline{Y}_{\mathrm{t}}(m)-\overline{Y}_{\mathrm{c}}(m)\right)=S^2\cdot\left(\frac{1}{N_{\mathrm{t}}(m)}+\frac{1}{N_{\mathrm{c}}(m)}\right).$$

In that case, a sensible choice for λ would be the value that maximizes precision by weighting the two statistics by the inverse of their sampling variances, or

$$\begin{aligned}\lambda_{\mathrm{opt}}&=\frac{1}{\frac{1}{N_{\mathrm{t}}(f)}+\frac{1}{N_{\mathrm{c}}(f)}}\Bigg/\left(\frac{1}{\frac{1}{N_{\mathrm{t}}(m)}+\frac{1}{N_{\mathrm{c}}(m)}}+\frac{1}{\frac{1}{N_{\mathrm{c}}(m)}+\frac{1}{N_{\mathrm{t}}(m)}}\right)\\&=\frac{N(f)\cdot\frac{N_{\mathrm{t}}(f)}{N(f)}\cdot\frac{N_{\mathrm{c}}(f)}{N(f)}}{N(f)\cdot\frac{N_{\mathrm{t}}(f)}{N(f)}\cdot\frac{N_{\mathrm{c}}(f)}{N(f)}+N(m)\cdot\frac{N_{\mathrm{t}}(m)}{N(m)}\cdot\frac{N_{\mathrm{c}}(m)}{N(m)}},\end{aligned}$$

with the weight for each stratum proportional to the product of the stratum size and the stratum proportions of treated and control units. The statistic $T^{\mathrm{dif},\lambda_{\mathrm{opt}}}$ often leads to a test statistic that is more powerful against alternatives with a constant treatment effect than $T^{\mathrm{dif},\lambda_{\mathrm{RSS}}}$, especially in settings with substantial variation in stratum-specific proportions of treated units.

We also could have used the exact same statistics we used in Chapter 5. For example, in the setting of a completely randomized experiment, a natural statistic was the difference between average observed treated and control outcomes:

$$T^{\mathrm{dif}}=\left|\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}-\overline{Y}_{\mathrm{c}}^{\mathrm{obs}}\right|.$$

In the current setting of stratified experiments, with two strata, this statistic can be written as

$$T^{\mathrm{dif}}=\left|\frac{1}{N_{\mathrm{t}}(f)+N_{\mathrm{t}}(m)}\sum_{i=1}^{N}W_i\cdot Y_i^{\mathrm{obs}}-\frac{1}{N_{\mathrm{c}}(f)+N_{\mathrm{c}}(m)}\sum_{i=1}^{N}(1-W_i)\cdot Y_i^{\mathrm{obs}}\right|.$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

Then we can write this statistic as

$$T^{\text{dif}} = \left| \frac{N_{\text{t}}(f)}{N_{\text{t}}} \cdot \overline{Y}_{\text{t}}^{\text{obs}}(f) - \frac{N(f) - N_{\text{t}}(f)}{N_{\text{c}}} \cdot \overline{Y}_{\text{c}}^{\text{obs}}(f) + \frac{N_{\text{t}}(m)}{N_{\text{t}}} \cdot \overline{Y}_{\text{t}}^{\text{obs}}(m) - \frac{N_{\text{c}}(m)}{N_{\text{c}}} \cdot \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|.$$

This statistic T^{dif} is a valid statistic for testing from the FEP perspective but somewhat unnatural in the current context. Because of Simpson's paradox, one would not always expect small values for the statistic, even when the null hypothesis holds. Suppose that the null hypothesis of zero treatment effects for all units holds and that the potential outcomes are closely associated with the covariate that determines the strata, for example, $Y_i(0) = Y_i(1) = X_i$ for all units ($Y_i(0) = Y_i(1) = 1$ for units with $X_i = 1$ and $Y_i(0) = Y_i(1) = 2$ for units with $X_i = 2$). In that case, the statistic T^{dif} is equal to

$$T^{\text{dif}} = \left| \frac{N_{\text{t}}(f)}{N_{\text{t}}} \cdot 1 - \frac{N(f) - N_{\text{t}}(f)}{N_{\text{c}}} \cdot 1 + \frac{N_{\text{t}}(m)}{N_{\text{t}}} \cdot 2 - \frac{N_{\text{c}}(m)}{N_{\text{c}}} \cdot 2 \right|.$$

If $N_f = 10$, $N_{\text{t}}(f) = 5$, $N(m) = 20$, and $N_{\text{t}}(m) = 5$, this is equal to

$$T^{\text{dif}} = \left| \frac{5}{10} \cdot 1 - \frac{5}{20} \cdot 1 + \frac{5}{10} \cdot 2 - \frac{15}{20} \cdot 2 \right| = \left| \frac{1}{2} + 1 - \frac{1}{4} - \frac{3}{2} \right| = \frac{1}{4}.$$

Under the sharp null hypothesis of no causal effects, the statistic $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$ no longer has expectation equal to zero, whereas it did have expectation zero in the completely randomized experiment. Nevertheless, T^{dif} is still a function of assignments, observed outcomes, and covariates, and as such its distribution under the null hypothesis can be tabulated, and p-values can be calculated.

Finally, let us consider rank-based statistics. In the setting with a completely randomized experiment we focused on the difference in average ranks. In that case we defined the normalized rank R_i (allowing for ties) as

$$R_i = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2} \left(1 + \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}} \right) - \frac{N+1}{2}.$$

Given the N ranks R_i, $i = 1, \ldots, N$, an obvious test statistic is the absolute value of the difference in average ranks for treated and control units:

$$T^{\text{rank}} = \left| \overline{R}_{\text{t}} - \overline{R}_{\text{c}} \right|, \quad \text{where} \quad \overline{R}_{\text{t}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} R_i, \quad \text{and} \quad \overline{R}_{\text{c}} = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} R_i,$$

where $\overline{R}_{\text{t}}$ and $\overline{R}_{\text{c}}$ are the average rank in the treatment and control groups respectively. Although we can use this statistic for the FEP approach, this would not be attractive if there is substantial variation between strata. We therefore propose modifying this statistic

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

for the setting of a stratified randomized experiment. Let R_i^{strat} be the normalized within-stratum rank of the observed outcome for unit i:

$$R_i^{\text{strat}} = \begin{cases} \sum_{j:G_i=f} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:G_i=f} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N(f)+1}{2}, & \text{if } G_i = f, \\ \sum_{j:G_i=m} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:G_i=m} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N(m)+1}{2}, & \text{if } G_i = m. \end{cases}$$

Then we can use the average value of the within-stratum ranks for treated and control units:

$$T^{\text{rank,stratum}} = \left| \overline{R}_{\text{t}}^{\text{strat}} - \overline{R}_{\text{c}}^{\text{strat}} \right|,$$

where

$$\overline{R}_{\text{t}}^{\text{strat}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} R_i^{\text{strat}}, \quad \text{and} \quad \overline{R}_{\text{c}}^{\text{strat}} = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} R_i^{\text{strat}}.$$

9.4.2 The FEP Approach with J Strata

Most of the statistics discussed in the previous section extend naturally to the case with J strata. Define for a general J-component vector λ the statistic

$$T^{\text{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|.$$

The first natural choice for λ has $\lambda(j)$ proportional to the stratum size,

$$\lambda(j) = \frac{N(j)}{N}, \quad \text{leading to } T^{\text{dif},\lambda_{\text{RSS}}} = \left| \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|.$$

The second choice for λ minimizes the sampling variance of the contrast between treated and control averages under homoskedasticity, leading to

$$\lambda^{\text{opt}}(j) = \frac{N(j) \cdot \frac{N_{\text{t}}(j)}{N(j)} \cdot \frac{N_{\text{c}}(j)}{N(j)}}{\sum_{k=1}^{J} N(k) \cdot \frac{N_{\text{t}}(k)}{N(k)} \cdot \frac{N_{\text{c}}(k)}{N(k)}},$$

in turn leading to

$$T^{\text{dif},\lambda_{\text{opt}}} = \left| \frac{1}{\sum_{j=1}^{J} N(j) \cdot \frac{N_{\text{t}}(j)}{N(j)} \cdot \frac{N_{\text{c}}(j)}{N(j)}} \sum_{j=1}^{J} N(j) \cdot \frac{N_{\text{t}}(j)}{N(j)} \cdot \frac{N_{\text{c}}(j)}{N(j)} \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|.$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

For the modified rank statistic, we define R_i^{strat} to be the normalized within-stratum rank of the observed outcome for unit i, taking account of ties:

$$R_i^{\text{strat}} = \sum_{i':G_{i'}=G_i} \mathbf{1}_{Y_{i'}^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{i':G_{i'}=G_i} \mathbf{1}_{Y_{i'}^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N(G_i)+1}{2}.$$

Then we can use the average value of the within-stratum ranks for treated and control units:

$$T^{\text{rank,stratum}} = \left|\overline{R}_{\text{t}}^{\text{strat}} - \overline{R}_{\text{c}}^{\text{strat}}\right|,$$

where, as before, $\overline{R}_{\text{t}}^{\text{strat}}$ and $\overline{R}_{\text{c}}^{\text{strat}}$ are the averages of the normalized within-stratum ranks for treated and control units.

9.4.3 The FEP Approach with Class-Level Data from Project Star

We now analyze the Project Star data using the FEP approach. Let $B_i(j)$, $i = 1, \ldots, 68$, $j = 1, \ldots, 13$ be an indicator for unit (i.e., teacher) i being from stratum (school) j. For the thirteen schools with two classes of each type, there are $\binom{4}{2} = 6$ different possible assignments. For the two schools with three small classes and two regular classes, there are $\binom{5}{2} = 10$ different possible assignments, and for the one school with four small and two regular classes, there are $\binom{6}{2} = 15$ different possible assignments. Hence, the total number of assignments of teachers to class type with positive probability is $(6^{13}) \times 10^2 \times 15 \approx 2 \times 10^{13}$. We therefore use numerical methods to approximate the p-values for the FEP approach.

We focus in this section on the null hypothesis that there is no effect of class size on the average test score that a teacher would achieve for their students,

$$H_0: \quad Y_i(0) = Y_i(1), \quad \text{for all } i = 1, \ldots, 68,$$

in any of the sixty-eight classes. We consider four test statistics based on the stratified class-level data. (Recall that the p-value has a valid interpretation only if one statistic is specified *a priori*, and our exercise is for illustrative purposes only.) The first test statistic is the absolute value of the difference in the average mathematics scores between small (treated) and regular-sized (control) classes:

$$T^{\text{dif}} = \left|\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right|.$$

As was discussed before, this statistic, which is natural in a completely randomized experiment, is not natural in this setting because one would not necessarily expect small values even when the null hypothesis is true (especially if there is substantial variation of the shares of treated units within the strata), although the results of the test are valid. The value of the statistic in the sample is 0.224. The p-value, here calculated as the probability under the randomization distribution of finding a value of the statistic at least as large as 0.224, is $p = 0.034$, thereby suggesting that it is unlikely that the students of

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

teachers assigned to the small classes had the same average test scores as the students of teachers assigned to large classes.

The second statistic is the average of the sixteen within-school average differences between small and regular class mathematics scores, weighted by the number of classes in the schools $N(j)$, divided by the total number of classes, $N = 68$:

$$T^{\text{dif},\lambda_{\text{RSS}}} = \left| \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|.$$

The realized value of the test statistic is 0.241. The p-value, now the probability under the randomization distribution of finding a value of the statistic at least as large as 0.241, is $p = 0.023$. This statistic also suggests that the teachers with smaller classes had different average test scores than teachers with regular-sized classes.

The third statistic also weights the within-school average differences, but now the weights are proportional to the product of the number of classes in each school and the proportions of treated and control classes within each school:

$$T^{\text{dif},\lambda_{\text{opt}}} = \left| \frac{1}{\sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_{\text{t}}(j)}{N(j)} \cdot \frac{N_{\text{c}}(j)}{N(j)}} \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_{\text{t}}(j)}{N(j)} \cdot \frac{N_{\text{c}}(j)}{N(j)} \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|.$$

Especially when there is considerable variation in the proportion of treated and control units between strata, this statistic is expected to be more powerful against alternative hypotheses with constant additive treatment effects. The realized value of the test statistic is 0.238, with a corresponding p-value of 0.025, leading to essentially the same substantive conclusion as that based on the previous two statistics.

In the current application, these three test-statistics lead to very similar p-values. This is partly because most of the schools have two classes of each type. If there were more dispersion in the fraction of small classes by school and in the number of classes per school, the results could well differ more for the three statistics. The value of the rank-based test $T^{\text{rank,stratum}}$ is 0.48, leading to a p-value of 0.15. Because the outcomes themselves are averages (over students within the classes), there are few outliers, and in this case, the rank-based tests would not be expected to have an advantage over statistics based on simple averages.

Another interesting test statistic here is based on the variation in average mathematics scores in small and regular classes. Suppose that at the individual-student level, it makes no difference to students whether they have many or few classmates, that is, whether they are in a regular or small class. In that case, the expected value of the average mathematics score in regular and small classes should be the same. However, because in small classes the average is calculated over fewer students than in large classes, the small class averages should have a larger variance. More precisely, if the individual test scores have a mean μ and variance σ^2, then the average in a class of size K should have mean μ and variance σ^2/K. So, even if individual student scores are not affected by class size, the null hypothesis that at the teacher level the average test score is not affected by the class size need not be true. We can investigate this phenomenon by choosing a new test statistic.

Now calculate for each school and class type the difference between the highest and the lowest average score:

$$\Delta_{\rm c}(j) = \max_{i:W_i=0,G_i=j} Y_i^{\rm obs} - \min_{i:W_i=0,G_i=j} Y_i^{\rm obs},$$

and

$$\Delta_{\rm t}(j) = \max_{i:W_i=1,G_i=j} Y_i^{\rm obs} - \min_{i:W_i=1,G_i=j} Y_i^{\rm obs}.$$

(For the schools with two small classes, this amounts to the absolute value of the difference between the two small classes.) We then take, for each school, the difference between this difference for small and regular classes:

$$\Delta(j) = \Delta_{\rm t}(j) - \Delta_{\rm c}(j).$$

We then average these differences over all 16 schools, weighted by the number of classes in each school:

$$T^{\rm range} = \frac{1}{N}\sum_{j=1}^{J} N(j)\cdot\Delta(j).$$

We find that the range does, indeed, on average appear to be larger in the small classes than in the regular classes, with the realized value of the test statistic equal to 0.226. The p-value based on the FEP calculations is 0.109. Thus there is only limited evidence against the null hypothesis that the variation in average scores differs between small and regular-sized classes.

9.4.4 The FEP Approach with Student-Level Data from Project Star

Here we consider an alternative analysis of the Project Star data, using the student-level data. This analysis is specific to the FEP approach and the particular structure of the Project Star data, and is not generally applicable to stratified randomized experiments. We present it here to show the richness of the FEP approach. This section can be bypassed without loss of continuity.

The key issue is that for this analysis, the no-interference part of the stability assumption, SUTVA, is automatically satisfied. More precisely, under the null hypothesis of no effects whatsoever, the no-interference assumption holds automatically, but it need not hold under the alternative hypothesis. Recall that the experiment assigned students and teachers randomly to the classes. Without the no-interference assumption, we index potential outcomes by the assignment vector that describes the class and teacher pair for each student. The discussion in this section is relatively informal. In Appendix A we present a more formal discussion of this example, which requires substantial new notation, which is not used in the rest of the text.

First consider the data from a single stratum, in this application a school, say school j. This school has $N(j)$ students and $P(j)$ teachers and classes. These students and teachers will be randomly assigned to $P(j)$ classes, with the class size for class s equal to $M_s(j)$.

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

The class sizes must add to the school size, or $\sum_{s=1}^{P(j)} M_s(j) = N(j)$. The total number of ways one can select the students, given class sizes, is

$$\prod_{s=1}^{P(j)-1} \binom{N(j) - \sum_{\mathrm{t}<s} M_{\mathrm{t}}(j)}{M_s(j)}.$$

The $P(j)$ teachers can be assigned to the $P(j)$ classes in $P(j)!$ ways, so the total number of ways the students and teachers for school j can be assigned to classes is

$$\prod_{s=1}^{P(j)-1} \binom{N(j) - \sum_{\mathrm{t}<s} M_{\mathrm{t}}(j)}{M_s(j)} \cdot P(j)!.$$

For each student this is the total number of potential outcomes. The basis for the randomization distribution is this set of assignments, which are all equally likely. The total number of assignments is obtained by multiplying this for each school, across all schools:

$$\prod_{j=1}^{J} \prod_{s=1}^{S(j)-1} \binom{N(j) - \sum_{\mathrm{t}<s} M_{\mathrm{t}}(j)}{M_s(j)} \cdot P(j)!.$$

The null hypothesis we consider is that of no effect whatsoever, against the alternative hypothesis that some potential outcomes differ. The test statistic we use is the average over the schools of the average student score for students in small classes minus the average student score for students in regular-sized classes.

$$T^{\text{student}} = \left| \frac{1}{\sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_{\mathrm{c}}(j)}{N(j)} \cdot \frac{N_{\mathrm{t}}(j)}{N(j)}} \cdot \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_{\mathrm{c}}(j)}{N(j)} \cdot \frac{N_{\mathrm{t}}(j)}{N(j)} \cdot \left(\overline{Y}_{\mathrm{t}}(j)^{\text{obs}} - \overline{Y}_{\mathrm{c}}(j)^{\text{obs}} \right) \right|,$$

with the stratum weight equal to

$$\frac{N(j)}{N} \cdot \frac{N_{\mathrm{c}}(j)}{N(j)} \cdot \frac{N_{\mathrm{t}}(j)}{N(j)}.$$

In the sample, the statistic is 0.242, with a p-value < 0.001. Thus we get much stronger evidence against this null hypothesis than we did for the null hypothesis using class-level data.

Now let us compare this analysis to that based on teacher-level data. If we were to maintain the no-interference assumption at the student level, the new null hypothesis requires only that changing student i's assignment from a regular to a small class does not change the outcome. In that case the student-level test score will tend to be more powerful than the class-level average test score, and the former would be preferable to the latter. However, in this application, the student-level stability assumption is a very strong and tenuous one to make. It is very plausible that there are interactions between children that would violate this assumption. Hence, even clear rejections of the null hypothesis of no differences by teacher assignment would not necessarily be credible evidence of systematic effects of class *size* – it may simply indicate the presence of

effects of teachers or peers. In contrast, the teacher-level assessment does not rely on within-class, no-interference assumptions, and so clear evidence against the null hypothesis of no effect based on that assessment is more credible evidence of class-size effects.

9.5 THE ANALYSIS OF STRATIFIED RANDOMIZED EXPERIMENTS FROM NEYMAN'S REPEATED SAMPLING PERSPECTIVE

The results in Chapter 6 for a completely randomized experiment can be used to analyze data within a stratum. Specifically, within each stratum those results can be used to obtain an estimate of the average treatment effect and to obtain a conservative estimator of the repeated sampling variance of this estimator.

9.5.1 The Two-Stratum Case

Initially we focus on the simple example with two strata and apply the framework to the Project Star data in Section 9.5.2. For the first stratum, the natural unbiased estimator for the average treatment effect $\tau_{\text{fs}}(f)$ is

$$\hat{\tau}^{\text{dif}}(f) = \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) = \frac{1}{N_{\text{t}}(f)} \sum_{i:G_i=f} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}(f)} \sum_{i:G_i=f} (1 - W_i) \cdot Y_i^{\text{obs}}.$$

The sampling variance of this estimator, under the randomization distribution, is

$$\mathbb{V}_W \left(\hat{\tau}^{\text{dif}}(f) \right) = \frac{S_{\text{c}}^2(f)}{N_{\text{c}}(f)} + \frac{S_{\text{t}}^2(f)}{N_{\text{t}}(f)} - \frac{S_{\text{ct}}^2(f)}{N(f)},$$

with analogous expressions for the estimator for the average treatment effect in the second stratum and its sampling variance. However, we are not necessarily interested in the two within-stratum average treatment effects. More commonly, we are interested in a weighted average of the two within-stratum average effects. A natural estimand is the finite-sample average treatment effect,

$$\tau_{\text{fs}} = \frac{N(f)}{N(f) + N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f) + N(m)} \cdot \tau_{\text{fs}}(m) = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i(1) - Y_i(0) \big).$$

With fixed stratum sizes, unbiasedness of the two within-stratum estimators implies unbiasedness of

$$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N(f) + N(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N(f) + N(m)} \cdot \hat{\tau}^{\text{dif}}(m),$$

for the population average treatment effect τ_{fs}. Similarly, the assumption that the randomizations in the two strata are independent, formalized in the assignment mechanism,

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

implies that the two estimators are uncorrelated, and thus

$$\begin{aligned}\mathbb{V}_W\left(\hat{\tau}^{\text{strat}}\right) &= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \mathbb{V}_W(\hat{\tau}_f) + \left(\frac{N(m)}{N_f+N(m)}\right)^2 \cdot \mathbb{V}_W(\hat{\tau}_m) \\ &= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{S_{\text{c}}(f)^2}{N_{\text{c}}(f)} + \frac{S_{\text{t}}^2(f)}{N_{\text{t}}(f)} - \frac{S_{\text{ct}}^2(f)}{N(f)}\right) \\ &\quad + \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{S_{\text{c}}(m)^2}{N_{\text{c}}(m)} + \frac{S_{\text{t}}^2(m)}{N_{\text{t}}(m)} - \frac{S_{\text{ct}}^2(m)}{N(m)}\right).\end{aligned}$$

The same issues that were discussed in Chapter 6 arise here in estimating this sampling variance. There is no direct way to estimate the components of this sampling variance involving the covariance of the unit-level potential outcomes, so typically those terms are ignored to obtain an estimated upper bound on the sampling variance by simply estimating the two within-stratum sampling variances:

$$\begin{aligned}\hat{\mathbb{V}}^{\text{neyman}} &= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_{\text{c}}^2(f)}{N_{\text{c}}(f)} + \frac{s_{\text{t}}^2(f)}{N_{\text{t}}(f)}\right) \\ &\quad + \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_{\text{c}}^2(m)}{N_{\text{c}}(m)} + \frac{s_{\text{t}}^2(m)}{N_{\text{t}}(m)}\right).\end{aligned}$$

This estimate of the sampling variance is unbiased if the within-stratum treatment effects are constant and additive, and overestimates the sampling variance in expectation otherwise. Note that we do not need to make assumptions about the variation in treatment effects between strata.

So far in this section, the discussion has focused on the estimation of the population average treatment effect, τ_{fs}. In some cases we may be interested in a different weighted average of the within-strata treatment effects. For example, we may be interested in the average effect of the treatment on the outcome for the units who received the treatment. Given the random assignment, and within the strata, this effect is equal to $\tau_{\text{fs}}(f)$ and $\tau_{\text{fs}}(m)$, respectively. Within each stratum this is, in expectation, the same as the average effect for the full stratum. However, when the proportions of treated units differ between the strata, the weights have to be adjusted to obtain an unbiased estimate of the average effect of the treatment on the units who received treatment. The appropriate weights are proportional to the fraction of treated units in each strata, leading to the estimand

$$\tau_{\text{fs},t} = \frac{N_{\text{t}}(f)}{N_{\text{t}}(f)+N_{\text{t}}(m)} \cdot \tau_{\text{fs}}(f) + \frac{N_{\text{t}}(m)}{N_{\text{t}}(f)+N_{\text{t}}(m)} \cdot \tau_{\text{fs}}(m),$$

and thus to the natural unbiased estimator

$$\hat{\tau}_{\text{t}}^{\text{strat}} = \frac{N_{\text{t}}(f)}{N_{\text{t}}(f)+N_{\text{t}}(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N_{\text{t}}(m)}{N_{\text{t}}(f)+N_{\text{t}}(m)} \cdot \hat{\tau}^{\text{dif}}(m).$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

The sampling variance of $\hat{\tau}_{\mathrm{t}}$ can be estimated in the same way as the sampling variance for the population average treatment effect, modifying the weights to reflect the new estimand:

$$\hat{\mathbb{V}}_{\mathrm{t}}^{\mathrm{neyman}} = \left(\frac{N_{\mathrm{t}}(f)}{N_{\mathrm{t}}(f)+N_{\mathrm{t}}(m)}\right)^2 \cdot \left(\frac{s_{\mathrm{c}}(f)^2}{N_{\mathrm{c}}(f)} + \frac{s_{\mathrm{t}}^2(f)}{N_{\mathrm{t}}(f)}\right) + \left(\frac{N_{\mathrm{t}}(m)}{N_{\mathrm{t}}(f)+N_{\mathrm{t}}(m)}\right)^2 \cdot \left(\frac{s_{\mathrm{c}}^2(m)}{N_{\mathrm{c}}(m)} + \frac{s_{\mathrm{t}}^2(m)}{N_{\mathrm{t}}(m)}\right).$$

More generally we can look at other weighted averages, such as the average effect for those who did not receive the treatment, but such averages are often more difficult to motivate as relevant.

Using Neyman's repeated sampling approach, we can also investigate other estimands, such as the differences between the stratum-specific average treatment effects. A natural unbiased estimator for the difference between $\tau_{\mathrm{fs}}(m)$ and $\tau_{\mathrm{fs}}(f)$ is

$$\hat{\tau}^{\mathrm{dif}}(m) - \hat{\tau}^{\mathrm{dif}}(f) = \left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(m) - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(m)\right) - \left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(f) - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(f)\right).$$

This estimator is unbiased for the difference in average treatment effects with sampling variance

$$\mathbb{V}_W\left(\hat{\tau}^{\mathrm{dif}}(m) - \hat{\tau}^{\mathrm{dif}}(f)\right) = \frac{S_{\mathrm{c}}^2(f)}{N_{\mathrm{c}}(f)} + \frac{S_{\mathrm{t}}^2(f)}{N_{\mathrm{t}}(f)} - \frac{S_{\mathrm{ct}}^2(f)}{N(f)} + \frac{S_{\mathrm{c}}^2(m)}{N_{\mathrm{c}}(m)} + \frac{S_{\mathrm{t}}^2(m)}{N_{\mathrm{t}}(m)} - \frac{S_{\mathrm{ct}}^2(m)}{N(m)}.$$

An estimator for the upper bound on this sampling variance is

$$\hat{\mathbb{V}}^{\mathrm{neyman}}\left(\hat{\tau}^{\mathrm{dif}}(m) - \hat{\tau}^{\mathrm{dif}}(f)\right) = \frac{s_{\mathrm{c}}^2(f)}{N_{\mathrm{c}}(f)} + \frac{s_{\mathrm{t}}^2(f)}{N_{\mathrm{t}}(f)} + \frac{s_{\mathrm{c}}^2(m)}{N_{\mathrm{c}}(m)} + \frac{s_{\mathrm{t}}^2(m)}{N_{\mathrm{t}}(m)}.$$

We can use any of the estimated sampling variances and the associated unbiased estimators to construct large-sample confidence intervals for the associated estimator.

9.5.2 The Neyman Approach and Project Star

Next, let us consider point estimates and confidence intervals for the average effect of the class size based on the stratified experiment. First we present estimates that account for the stratification. For each school j, for $j = 1, \ldots, 16$, the average effect of the treatment and its corresponding sampling variance are estimated as

$$\hat{\tau}^{\mathrm{dif}}(j) = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j), \quad \text{and} \quad \hat{\mathbb{V}}^{\mathrm{neyman}}(j) = \frac{s_{\mathrm{c}}(j)^2}{N_{\mathrm{c}}(j)} + \frac{s_{\mathrm{t}}(j)^2}{N_{\mathrm{t}}(j)},$$

respectively. For each school, the estimated average effect and the square root of the estimated sampling variance are reported in Table 9.2. The population average effect is estimated as

$$\hat{\tau}^{\mathrm{strat}} = \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \hat{\tau}(j) = 0.241,$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

Table 9.2. ***Within-School Estimates of Treatment Effect of Small Classes Relative to Regular Classes – Project Star***

School	Estimated Effect	$\widehat{(\text{s.e.})}$
1	0.223	(0.230)
2	−0.295	(0.776)
3	0.417	(0.404)
4	0.748	(0.215)
5	−0.077	(0.206)
6	1.655	(0.405)
7	−0.254	(0.255)
8	0.429	(0.306)
9	−0.006	(0.311)
10	−0.014	(0.182)
11	−0.003	(0.605)
12	0.222	(0.309)
13	0.432	(0.179)
14	0.340	(0.336)
15	0.207	(0.396)
16	−0.306	(0.245)
$\hat{\tau}^{\text{strat}}$	0.241	(0.092)

and its sampling variance by

$$\hat{\mathbb{V}}^{\text{neyman}} = \sum_{j=1}^{J} \left(\frac{N(j)}{N} \right)^2 \cdot \hat{\mathbb{V}}^{\text{neyman}}(j) = 0.092^2.$$

Hence the large sample 95% confidence interval for the average effect is

$$\text{CI}^{0.95}(\tau_{\text{fs}}) = \big(0.061, 0.421\big).$$

It is interesting to compare this point estimate and its associated standard error to that based on the analysis using the (incorrect) assumption that the data arose from a completely randomized experiment. The point estimate of the average effect is then $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 0.224$, with an estimated standard error of 0.141, leading to a large sample 95% confidence interval of $(-0.053, 0.500)$. This estimator of the sampling variance is biased if there is variation in the probability of treatment between the different strata, or if there is variation in the average potential outcomes by stratum. We know the former is the case, with the probability of a small class equal to 0.5 in most schools, and equal to 0.60 and 0.67 in some schools. Assessing the latter issue is more complicated, and we shall return to this in Section 9.7.2. The fact that the point estimates differ under the assumptions of a completely randomized experiment and a stratified randomized experiment suggests that average potential outcomes also differ between strata. The estimated standard error for the stratification-based analysis is smaller than that for the completely randomized experiment, suggesting, again, that average potential

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

outcomes differ between strata, which implies that there is a gain in precision from the stratification.

9.6 REGRESSION ANALYSIS OF STRATIFIED RANDOMIZED EXPERIMENTS

In order to interpret regression-based estimators, we take a super-population perspective with a fixed number of strata, and an infinite number of units within each stratum. Because there are few notational simplifications from considering the special case with only two strata, we look in this section immediately at the general situation with J strata.

9.6.1 The General Framework

Let $q(j) = N(j)/N$ and $e(j) = N_t(j)/N(j)$ be the proportion of each stratum in the sample from the infinite super-population, and the proportion of treated units in each stratum, or the propensity score, respectively. We consider two specifications of the regression function in this case. The first specification of the regression function treats the stratum indicators as additional regressors and includes them additively. The second specification includes a full set of interactions of the stratum indicators with the treatment indicator. We then investigate the large-sample properties of the least squares estimators of the coefficients on the treatment indicator.

Similar to the regression function specifications in Chapter 7, the first specification simply includes indicators for the strata additively in addition to the indicator for the treatment:

$$Y_i^{\text{obs}} = \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \varepsilon_i, \tag{9.1}$$

where $B_i(j)$ is an indicator for unit i belonging to stratum j. Because we include, in this specification, a full set of stratum indicators $B_i(j)$, for $j = 1, \ldots, J$, we do not include an intercept in the specification of the regression function. We focus on the least squares estimator for τ,

$$(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\tau,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) \right)^2. \tag{9.2}$$

As before, we define τ^* and β^* to be the population counterparts to these OLS estimators,

$$(\tau^*, \beta^*) = \arg\min_{\tau,\beta} \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) \right)^2 \right]. \tag{9.3}$$

The first question concerns the population value τ^* corresponding to $\hat{\tau}^{\text{ols}}$. In general $\hat{\tau}^{\text{ols}}$ is not consistent for the population average treatment effect τ_{sp}. Instead, it estimates a

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

weighted average of the within-stratum average effects, with weights proportional to the product of the fraction of observations in the stratum and the probabilities of receiving and not receiving the treatment. More specifically,

$$\omega(j) = q(j) \cdot e(j) \cdot (1 - e(j)), \quad \text{and} \quad \tau_\omega = \sum_{j=1}^{J} \omega(j) \cdot \tau_{\text{sp}}(j) \Bigg/ \left(\sum_{j=1}^{J} \omega(j) \right), \tag{9.4}$$

where $\tau_{\text{sp}}(j) = \mathbb{E}[Y_i(1) - Y_i(0) | B_i(j) = 1]$. Then $\hat{\tau}^{\text{ols}}$ is consistent for τ_ω. The following theorem formalizes this result.

Theorem 9.1 *Suppose we conduct a stratified randomized experiment in a sample drawn at random from an infinite population. Then, for estimands τ^* and τ_w defined in (9.3) and (9.4), the estimator $\hat{\tau}^{\text{ols}}$ satisfies,* (*i*)

$$\tau^* = \tau_\omega,$$

and (*ii*)*,*

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_\omega \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\left(W_i - \sum_{j=1}^{J} q(j) \cdot B_i(j) \right)^2 \cdot \left(Y_i^{\text{obs}} - \tau^* \cdot W_i - \sum_{j=1}^{J} \beta_j^* \cdot B_i(j) \right)^2 \right]}{\left(\sum_{j=1}^{J} q(j) \cdot e(j) \cdot (1 - e(j)) \right)^2} \right).$$

The proof appears in Appendix B.

The weights $\omega(j)$ have an interesting interpretation. Suppose we estimate the within-stratum average treatment effect $\tau^{\text{dif}}(j)$ as $\hat{\tau}^{\text{dif}}(j) = \overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j)$. The sampling variance of $\hat{\tau}^{\text{dif}}(j)$, under the assumption of a constant treatment effect, is $(S^2/N) \cdot (q(j) \cdot e(j) \cdot (1 - e(j)))^{-1}$. Hence the weights $\omega(j)$ are proportional to the precision of natural unbiased estimators of the within-stratum treatment effects, which leads to a relatively precisely estimated weighted average effect.

The second specification of the regression function includes a full set of interactions of the stratum indicators with the indicator for the treatment W_i. In order to be able to interpret the coefficient on the treatment indicator as an average causal effect, we include the interactions with the stratum indicators relative to their share in the sample and relative to the indicator for the last stratum:

$$Y_i^{\text{obs}} = \tau \cdot W_i \cdot \frac{B_i(j)}{N(j)/N} + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \sum_{j=1}^{J-1} \gamma(j) \cdot W_i \cdot \left(B_i(j) - B_i(J) \cdot \frac{N(j)}{N(J)} \right) + \varepsilon_i. \tag{9.5}$$

Note that in this specification we only include the first $J - 1$ interactions to avoid perfect collinearity in the regression function. In this case, the population value τ^*, corresponding to the large sample limit of the least squares estimator $\hat{\tau}^{\text{ols,inter}}$, is equal to the population average treatment effect τ_{sp}.

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

Theorem 9.2 *Suppose we conduct a stratified randomized experiment in a sample drawn at random from an infinite population. Then, for* $\hat{\tau}^{\text{ols,inter}}$ *defined as the least squares estimator corresponding to the regression function in (9.5), and* τ^* *defined as the population limit corresponding to that estimator,* (*i*)

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

and (*ii*),

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols,inter}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{J} q(j)^2 \cdot \left(\frac{\sigma_{\text{c}}^2(j)}{(1-e(j)) \cdot q(j)} + \frac{\sigma_{\text{t}}^2(j)}{e(j) \cdot q(j)}\right)\right).$$

It is interesting to compare the sampling variance of $\hat{\tau}^{\text{ols}}$ and $\hat{\tau}^{\text{ols,inter}}$. In general, the sampling variance of $\hat{\tau}^{\text{ols,inter}}$ is larger than that of $\hat{\tau}^{\text{ols}}$.

9.6.2 Regression Analysis of Project Star

The first specification of the regression function includes the treatment indicator and the indicators for the blocks:

$$Y_i^{\text{obs}} = \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \varepsilon_i.$$

The point estimate and standard error for τ_{fs} are

$$\hat{\tau}^{\text{ols}} = 0.238 \quad (\widehat{\text{s.e.}}\ 0.103).$$

Recall from the discussion in Section 9.6 that this estimator is not necessarily consistent for the average effect of the treatment in the population if there is variation in the effect of the class size by school.

The second specification of the regression function includes indicators for the strata, as well as interactions of the stratum indicators and the treatment indicator:

$$Y_i^{\text{obs}} = \tau \cdot W_i \cdot \frac{B_i(J)}{N(J)/N} + \sum_{j=1}^{J} \beta_j \cdot B_i(j) + \sum_{j=1}^{J-1} \tau(j) \cdot W_i \cdot \left(B_i(j) - B_i(J) \cdot \frac{N(j)}{N(J)}\right) + \varepsilon_i.$$

The point estimate and standard error for τ, based on this specification, are

$$\hat{\tau}^{\text{ols,inter}} = 0.241 \quad (\widehat{\text{s.e.}}\ 0.095).$$

The two estimates for the average effect are close, with similar standard errors, consistent with limited heterogeneity in the treatment effects.

9.7 MODEL-BASED ANALYSIS OF STRATIFIED RANDOMIZED EXPERIMENTS

In a model-based analysis, it is conceptually straightforward to take account of the stratification. As in the analysis of completely randomized experiments, we combine the specification of the joint distribution of the potential outcomes with the known distribution of the vector of assignment indicators to derive the posterior distribution of

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

the causal estimand. There is one new issue that arises in this context: the link between the distributions of the potential outcomes in distinct strata.

9.7.1 General Considerations

One can choose to have distinct parameters for the distributions in different strata, that is, independent prior distributions. Alternatively the researcher may wish to link the parameters in the different strata either deterministically by imposing equality restrictions or stochastically through a dependence structure in the prior distribution, that is, for example, through a hierarchical model. In situations with few strata and many units per stratum, one may wish to pursue the first strategy and specify distinct distributions for the potential outcomes in each stratum, with independent prior distributions on the parameters of these distributions. In contrast, in settings with a substantial number of strata, and a modest number of units per stratum, one may wish to link some of the parameters. One can do so by restricting them to be equal, or by incorporating dependence into the specification of the prior distribution.

We make this more specific and illustrate the issues for the case with common and stratum-specific parameters. Suppose we specify the joint distribution of the potential outcomes in stratum j as

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Bigg| B_i(j), \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix}, \begin{pmatrix} \sigma_c^2(j) & 0 \\ 0 & \sigma_t^2(j) \end{pmatrix} \right), \tag{9.6}$$

where the means $(\mu_c(j), \mu_t(j))$ and variances $(\sigma_c^2(j), \sigma_t^2(j))$ are specific to stratum j. The full parameter vector is $\theta = (\mu_c(j), \mu_t(j), \sigma_c^2(j), \sigma_t^2(j), w = 0, 1, j = 1, \ldots, J)$.

With few strata and a substantial number of units per stratum, we may wish to use a prior distribution that makes all elements of θ *a priori* independent, for example, using normal prior distributions for the $\mu_c(j)$ and $\mu_t(j)$ and inverse chi-squared prior distributions for the $\sigma_c^2(j)$ and $\sigma_t^2(j)$.

However, if there are many strata and the number of units per stratum is modest, we may wish to specify a hierarchical prior distribution for the means to obtain more precise estimates. For example, we may wish to restrict the variances of the potential outcomes to be the same across strata, σ_c^2 and σ_t^2 for all j, and to specify the means to have a joint normal prior distribution, independent of the variances σ_c^2 and σ_t^2:

$$\begin{pmatrix} \mu_c(1) \\ \mu_c(2) \\ \vdots \\ \mu_c(J) \\ \mu_t(1) \\ \mu_t(2) \\ \vdots \\ \mu_t(J) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma_c \\ \gamma_c \\ \vdots \\ \gamma_c \\ \gamma_t \\ \gamma_t \\ \vdots \\ \gamma \end{pmatrix}, \begin{pmatrix} \eta_c^2 & 0 & \ldots & 0 & \rho\sigma_c\sigma_t & 0 & \ldots & 0 \\ 0 & \eta_c^2 & & \vdots & 0 & \rho\sigma_c\sigma_t & & \vdots \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & \ldots & & \eta_c^2 & 0 & \ldots & & \rho\sigma_c\sigma_t \\ \rho\sigma_c\sigma_t & 0 & \ldots & 0 & \eta_t^2 & 0 & \ldots & 0 \\ 0 & \rho\sigma_c\sigma_t & & \vdots & 0 & \eta_t^2 & & \vdots \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & \ldots & & \rho\sigma_c\sigma_t & 0 & \ldots & & \eta_t^2 \end{pmatrix} \right).$$

The full parameter vector is now $\theta = (\sigma_c^2, \sigma_t^2, \gamma_c, \gamma_t, \eta_c^2, \eta_t^2, p)$.

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

9.7.2 A Model-Based Analysis of Project Star

We now conduct a model-based imputation analysis of the Project Star data. The model we consider for the potential outcomes is

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Bigg| B_i(j) = 1, \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right),$$

with a common variance σ^2. In addition we assume that the pairs of stratum-specific means $(\mu_c(j), \mu_t(j))$ are independent across strata given the hyperparameters,

$$\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix} \Bigg| \sigma^2, \gamma_c, \gamma_t, \Sigma \sim \mathcal{N}\left(\begin{pmatrix} \gamma_c \\ \gamma_t \end{pmatrix}, \Sigma \right), \quad \begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix} \perp\!\!\!\perp \begin{pmatrix} \mu_c(k) \\ \mu_t(k) \end{pmatrix} \Bigg| \sigma^2, \gamma_c, \gamma_t, \Sigma, j \neq k.$$

In this model, the two potential outcome means $(\mu_c(j), \mu_t(j))$ are specific to the stratum, and the variance σ^2 is common to all strata and both potential outcomes.

The full parameter vector is $\theta = (\gamma_c, \gamma_t, \Sigma, \sigma^2)$. For the prior distributions, we use conventional proper choices. For the variance parameter σ^2, we use a standard inverse Chi-squared prior distribution,

$$k_0 \cdot \nu_0^2 \cdot \sigma^{-2} \sim \mathcal{X}^2(k_0), \quad \text{or} \quad \sigma^2 \sim \mathcal{X}^{-2}(k_0, \nu_0^2),$$

using the notation from Gelman, Carlin, Stern, and Rubin (1995). Our choices for the parameters of the prior distribution are $k_0 = 2$ and $\nu_0^2 = 0.001$. For γ_c and γ_t, we use independent normal prior distributions,

$$\begin{pmatrix} \gamma_c \\ \gamma_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100^2 & 0 \\ 0 & 100^2 \end{pmatrix} \right).$$

The prior distribution for Σ is an inverse wishart distribution,

$$\Sigma \sim \mathcal{W}^{-1}(k_1, \Gamma_1^{-1}).$$

We consider two pairs of values for (k_1, Γ_1). The first is $k_1 = 1{,}000$, $\Gamma_1 = 1{,}000 \cdot \mathcal{I}_2$, where $\mathcal{I}_k$ is the $k \times k$ identity matrix. This essentially corresponds to removing the link between the parameters in the different strata. We refer to this as the "independent" prior, corresponding to independence between the stratum-specific means. The second choice for (k_1, Γ_1) is $k_1 = 3$ and $\Gamma_1^{-1} = 0.001 \cdot k_1 \cdot \mathcal{I}_2$, which allows the hierarchical structure to influence answers. We refer to this prior distribution as the hierarchical prior.

For the independent prior distribution, the posterior mean and standard deviation are

$$\mathbb{E}[\tau_{fs} | \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{B}, \text{independent}] = 0.241, \quad \mathbb{V}(\tau_{fs} | \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{B}, \text{independent}) = 0.095^2.$$

Substantively it is difficult to see why one would wish to impose the *ex post* independence. Certainly, as we will see, there is strong evidence in the data to suggest that the average potential outcomes within the schools are related.

For the hierarchical prior distribution, the posterior mean and standard deviation are

$$\mathbb{E}[\tau_{fs} | \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{B}, \text{hierarchical}] = 0.235, \quad \mathbb{V}(\tau_{fs} | \mathbf{Y}^{obs}, \mathbf{W}, \mathbf{B}, \text{hierarchical})^2 = 0.107^2.$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

It is also interesting to assess the evidence for variation in average potential outcomes and treatment effects by strata. In order to do so, we inspect the posterior distribution of Σ given the hierarchical prior distribution. The logarithm of the square root of the two diagonal elements corresponds to the logarithm of the standard deviation of $\mu_c(j)$ and $\mu_t(j)$ over the sixteen schools. The posterior means of logarithms of those two standard deviations are

$$\mathbb{E}\left[\ln\left(\sqrt{\Sigma_{11}}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=-1.14,$$
$$\mathbb{V}\left(\ln\left(\sqrt{\Sigma_{11}}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.47^2,$$

and

$$\mathbb{E}\left[\ln\left(\sqrt{\Sigma_{22}}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=-1.08,$$
$$\mathbb{V}\left(\ln\left(\sqrt{\Sigma_{22}}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.45^2.$$

There is clearly some evidence of heterogeneity in the stratum means. However, the heterogeneity is highly correlated across potential outcomes, with the posterior mean for the Fisher Z transformation of the correlation between $\beta_c(j)$ and $\beta_t(j)$ (the (1, 2) element of Σ divided by the square root of the product of the (1, 1) and (2, 2) elements) equal to

$$\mathbb{E}\left[\frac{1}{2}\ln\left(\frac{1+\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}{1-\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=2.63,$$

and the posterior variance equal to

$$\mathbb{V}\left(\frac{1}{2}\ln\left(\frac{1+\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}{1-\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.67^2.$$

The posterior mean of the correlation itself is 0.96. The average treatment effect in school j is approximately $\tau(j)=\mu_t(j)-\mu_c(j)$. In terms of the parameters, the variance of the treatment effect across the sixteen schools is $(-1\ 1)\Sigma(-1\ 1)'=\Sigma_{11}-\Sigma_{12}-\Sigma_{21}+\Sigma_{22}$. We focus on the square root of this, that is, the standard deviation of the treatment effect over the schools. The posterior mean of the logarithm of the standard deviation of the treatment effect is

$$\mathbb{E}\left[\ln\left(\sqrt{\Sigma_{11}-\Sigma_{12}-\Sigma_{21}+\Sigma_{22}}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=-2.33,$$

with posterior variance

$$\mathbb{V}\left(\ln\left(\sqrt{\Sigma_{11}-\Sigma_{12}-\Sigma_{21}+\Sigma_{22}}\right)\Big|\,\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.59^2.$$

Comparing the posterior mean of the standard deviation of the stratum-specific treatment effect $\tau(j)$ over the sixteen strata, (0.115), with the posterior mean of the standard deviation of the stratum-specific level under the control treatment $\mu_c(j)$ over the sixteen strata, (0.349), suggests that, although there is considerable evidence that *levels* of the average test scores vary by school, there is little evidence that average class size *effects* vary much

by school. The former may be due to differences in teacher quality or to differences in student populations. This type of conclusion highlights the advantage of a fully model-based analysis, which allows for the simultaneous investigation of multiple questions.

9.8 DESIGN ISSUES: STRATIFIED VERSUS COMPLETELY RANDOMIZED EXPERIMENTS

When designing an experimental evaluation, one may often have the choice between a completely randomized experiment and a stratified randomized experiment. Here we study the implications of the choice between the different experimental designs for the expected sampling variance of the standard unbiased estimator for the average treatment effect. There is a sense in which one is never worse off stratifying on a covariate. However, to make this point precise, we need to pose the question appropriately.

We analyze the problem in a super-population setting. Each unit in this population has a binary characteristic G_i, $G_i \in \{f, m\}$. The proportion of women ($G_i = f$ types) in the population is p. We consider the following two designs. In the first design we randomly draw N units from the population. Out of this sample of size N, we randomly draw $N_{\rm t} = q \cdot N$ units to receive the active treatment and $N_{\rm c} = (1-q) \cdot N$ units to receive the control treatment. Based on the randomized experiment, we estimate the average treatment effect in the super-population as

$$\hat{\tau}^{\rm dif} = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs},$$

with (super-population) sampling variance

$$\mathbb{V}_{\rm sp}(\hat{\tau}^{\rm dif}) = \frac{\sigma_{\rm c}^2}{N_{\rm c}} + \frac{\sigma_{\rm t}^2}{N_{\rm t}}.$$

In the second design, we randomly draw $N(f) = p \cdot N$ units from the subpopulation of units who have $G_i = f$, and $N(m) = (1-p) \cdot N$ units from the population who have $G_i = m$. In the first subsample, we randomly select $N_{\rm t}(f) = p \cdot q \cdot N$ units to receive the active treatment, and the remaining $N_{\rm c}(f) = p \cdot (1-q) \cdot N$ are assigned to receive the control treatment. In the second subsample $N_{\rm t}(m) = p \cdot q \cdot N$ units are randomly selected to receive the active treatment, and the remaining $N_{\rm c}(m) = (1-p) \cdot (1-q) \cdot N$ units to receive the control treatment. Note that we assign the same proportion of units in each subpopulation to the active treatment. In this experiment, we estimate the average treatment effect within the $G_i = f$ and $G_i = m$ subpopulations as

$$\hat{\tau}^{\rm dif}(f) = \overline{Y}_{\rm t}^{\rm obs}(f) - \overline{Y}_{\rm c}^{\rm obs}(f), \quad \text{and} \quad \hat{\tau}^{\rm dif}(m) = \overline{Y}_{\rm t}^{\rm obs}(m) - \overline{Y}_{\rm c}^{\rm obs}(m),$$

and the overall average effect as

$$\hat{\tau}^{\rm strat} = \frac{N(f)}{N} \cdot \hat{\tau}^{\rm dif}(f) + \frac{N(m)}{N} \cdot \hat{\tau}^{\rm dif}(m) = p \cdot \hat{\tau}^{\rm dif}(f) + (1-p) \cdot \hat{\tau}^{\rm dif}(m).$$

The super-population variance for this estimator is

$$\mathbb{V}_{\rm sp}(\hat{\tau}^{\rm strat}) = \frac{p}{N} \cdot \left(\frac{\sigma_{\rm t}^2(f)}{p} + \frac{\sigma_{\rm c}^2(f)}{1-p} \right) + \frac{1-p}{N} \cdot \left(\frac{\sigma_{\rm t}^2(m)}{p} + \frac{\sigma_{\rm c}^2(m)}{1-p} \right).$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

The difference between the two sampling variances, normalized by the sample size N, is

$$N \cdot \left(\mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{dif}}) - \mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{strat}}) \right) = p(1-p) \cdot \left((\mu_{\text{c}}(f) - \mu_{\text{c}}(m))^2 + (\mu_{\text{t}}(f) - \mu_{\text{t}}(m))^2 \right) \geq 0,$$

where $\mu_{\text{c}}(f)$ is the average of $Y_i(0)$ for women, and similarly for $\mu_{\text{c}}(f)$, $\mu_{\text{c}}(m)$, and $\mu_{\text{t}}(m)$.

Although under some conditions there is an unambiguous ranking of the population sampling variances, $\mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{dif}})$ and $\mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{strat}})$, the *estimated* sampling variance for the stratified experiment may be larger than for the completely randomized experiment. The natural estimator for the sampling variance of the simple unbiased estimator in a stratified randomized experiment can be larger than the natural estimators for the sampling variance in a completely randomized experiment, because of the need to estimate the within-stratum potential outcome variances.

We can assess the benefits of having the stratification for an experiment with the size of Project Star. Suppose we have J strata, each with N_{t} treated (small) and $N_{\text{c}} = N_{\text{t}}$ control (regular-sized) classes. Suppose that the true within-stratum variance of the potential outcomes is $\sigma^2 = 0.43^2$, which is the posterior mean for the hierarchical model estimated on the Project Star data. Suppose also that the true variance of the within-stratum average potential outcomes over the strata is $\Sigma_{11} = 0.37^2$ for the control averages $\mu_{\text{c}}(j)$ and $\Sigma_{22} = 0.37^2$ for the averages given the treatment $\mu_{\text{t}}(j)$, again estimated on the Project Star data. Then the ratio of the variances under a completely randomized experiment versus a stratified randomized experiment would be $(0.43^2+0.37^2)/0.43^2 = 1.65$. Using a stratified design reduces the variance by 40%. The stratification appears to be quite effective in Project Star.

9.9 CONCLUSION

In this chapter we discuss the analysis of stratified randomized experiments using the four approaches developed in the previous four chapters for completely randomized experiments. In general the stratification should not be ignored in design if treatment rates and potential outcomes vary systematically by stratum. All approaches can be adapted in a fairly straightforward manner to take account of the stratification. A key issue is that in the model-based analysis, a hierarchical model can be useful to take account of similarities in potential outcome distributions across strata. As we illustrate using data from the Project Star experiment on class size, stratification can increase precision of estimation when the strata are good predictors of the potential outcomes.

In the next chapter we extend these analyses to an extreme version of stratification in an experimental context, paired randomized experiments, where each stratum consists of only two units, one treated and one control.

NOTES

The Project Star data have been used by numerous researchers. For more recent research papers, see Krueger (1999), Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) and Graham (2008). Graham (2008) looks at implications of within-class interactions on variances, as discussed in Section 9.4.3.

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

To implement the Bayesian analysis discussed in Sections 9.7 and 9.7.2 it is useful to use modern numerical methods, in particular Markov-Chain-Monte-Carlo methods, which we discuss in some detail in Chapter 8.

In textbook discussions of the benefits of stratification, and its extreme version, pairing versus complete randomization, it is sometimes pointed out that there are costs associated with stratification and pairing in small population settings. For example, Snedecor and Cochran (1989, p. 101) write: "If the criterion has no correlation with the response variable, a small loss in accuracy results from the pairing due to the adjustment for degrees of freedom. A substantial loss may even occur if the criterion is badly chosen so that members of a pair are negatively correlated." The possibility of negative correlation arises only if in the populations in the strata are small. For example, as discussed in Snedecor and Cochran (1967, p. 294), if the strata correspond to litters of rats, then weights within strata may well be negatively correlated. On the other hand, if the within-strata samples are drawn from large strata, in expectation the stratification can only lead to non-negative correlations.

Box, Hunter, and Hunter (2005, p. 93) also suggest that there is a trade-off in terms of accuracy or variance in the decision to stratify, writing: "Thus you would gain from the paired design only if the reduction in variance from pairing outweighed the effect of the decrease in the number of degrees of freedom of the t distribution." These comments reflect on the implications for testing and interval estimation. In expectation, with large size strata, the sampling variance of the estimated average treatment effect can only decrease as a result of stratification or pairing, not increase.

Samii and Aronow (2012) discuss comparisons between regression approaches and Neyman repeated sampling variances in this setting.

APPENDIX A: STUDENT-LEVEL ANALYSES

Here we discuss the student-level significance tests in more detail. First consider the data from a single stratum, say school j. This school has $N(j)$ students with $P(j)$ classes/teachers. The class size for class s in school j is $M_s(j)$, with $\sum_{s=1}^{P(j)} M_s(j) = N(j)$. Note that we do not require the class sizes to be the same for all small or all regular-sized classes. Even if some classes are exactly the same size, we analyze them as distinct in the sense that having a particular group of twenty students and a teacher assigned to class 1, and a second group of ten students and another teacher assigned to class 2 is a different assignment from having the first group of students and their teacher assigned to class 2 and the others to class 1. This is not necessary, but interpreting those assignments as identical would require keeping track of classes that have identical sizes versus differ by small numbers. The $N(j)$ students and the $P(j)$ teachers are assigned randomly to the $P(j)$ classes. Start with the teachers. The $P(j)$ teachers can be assigned to the $P(j)$ classes in $P(j)!$ different ways. Selecting $M_1(j)$ students for the first class can be done in $\binom{N(j)}{M_1(j)}$ different ways. Selecting the students for the next class can be done in $\binom{N(j)-M_1(j)}{M_2(j)}$ different ways, and so on, implying that the students can be assigned in

$$\prod_{s=1}^{P(j)-1} \binom{N(j) - \sum_{t<s} M_t(j)}{M_s(j)}$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

different ways. Combining this with the teachers' assignments, the total number of ways the students and teachers for school j can be assigned is

$$\prod_{s=1}^{P(j)-1} \binom{N(j) - \sum_{t<s} M_t(j)}{M_s(j)} \cdot P(j)!.$$

For each student this is the total number of potential outcomes. Thus, let $\mathbf{W}(j)$ be the $N(j)$ vector of student assignments for school j, where the i^{th} element of $\mathbf{W}(j)$ takes on values in the set $\{1, \ldots, P(j)\}$, indicating which class student i is assigned to. In addition, $\mathbf{T}(j)$ is the $P(j)$-dimensional vector of teacher assignments in school j, again with each element of $\mathbf{T}(j)$ taking on values in the set $\{1, \ldots, P(j)\}$. Thus we can write the potential outcome for student i in school j as

$$Y_{ij}\left(\mathbf{W}(j), \mathbf{T}(j)\right).$$

The null hypothesis we consider is

$$H_0: \; Y_{ij}\Big(\mathbf{W}(j), \mathbf{T}(j)\Big) = Y_{ij}\left(\mathbf{W}'(j), \mathbf{T}'(j)\right) \quad \text{for all } \mathbf{W}(j), \mathbf{T}(j), \mathbf{W}'(j), \mathbf{T}'(j).$$

The basis for the randomization distribution is the full set of assignments, which are all equally likely. The total number of assignments is obtained by multiplying the number of assignments for each school:

$$\prod_{j=1}^{J} \prod_{s=1}^{P(j)-1} \binom{N(j) - \sum_{t<s} M_t(j)}{M_s(j)} \cdot P(j)!.$$

APPENDIX B: PROOFS OF THEOREMS 9.1 AND 9.2

It is convenient to reparametrize the model slightly. Instead of (τ, β), we parametrize the model as (τ, γ), where $\gamma(j) = \beta(j) - e(j) \cdot \tau$, which does not change the least squares estimate of τ. In terms of (τ, γ), the regression function is

$$Y_i^{\text{obs}} = \tau \cdot \left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right) + \sum_{j=1}^{J} \gamma(j) \cdot B_i(j) + \varepsilon_i.$$

The population values for the parameters are

$$(\tau^*, \gamma^*) = \arg\min_{\tau,\gamma} \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tau \cdot \left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right) - \sum_{j=1}^{J} \gamma(j) \cdot B_i(j)\right)^2\right].$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

We can write

$$Y_i^{\text{obs}} = \sum_{j=1}^{J} \alpha(j) \cdot B_i(j) + \sum_{j=1}^{J} \tau(j) \cdot W_i \cdot B_i(j) + \eta_i,$$

where $\alpha(j) = \mathbb{E}_{\text{sp}}[Y_i(0)|B_i(j) = 1]$ and $\tau_{\text{sp}}(j) = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|B_i(j) = 1]$, and where by definition $\mathbb{E}[\eta_i|B_i(1), \ldots, B_i(J), W_i] = 0$. Therefore,

$$\begin{aligned}
(\tau^*, \gamma^*) &= \arg\min_{\tau,\gamma} \mathbb{E}\left[\left(\sum_{j=1}^{J} \alpha(j) \cdot B_i(j) + \sum_{j=1}^{J} \tau(j) \cdot W_i \cdot B_i(j) - \tau \right.\right. \\
&\qquad \left.\left. \cdot \left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right) - \sum_{j=1}^{J} \gamma(j) \cdot B_i(j)\right)^2\right] \\
&= \arg\min_{\tau,\gamma} \mathbb{E}\left[\left(\sum_{j=1}^{J} B_i(j) \cdot (\alpha(j) - \gamma(j) + \tau(j) \cdot W_i) - \tau \right.\right. \\
&\qquad \left.\left. \cdot \left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right)\right)^2\right] \\
&= \arg\min_{\tau,\gamma} \left\{ \mathbb{E}\left[\left(\sum_{j=1}^{J} B_i(j) \cdot (\alpha(j) - \gamma(j) + \tau(j) \cdot W_i)\right)^2\right] \right. \\
&\qquad - 2 \cdot \tau \cdot \mathbb{E}\left[\sum_{j=1}^{J} B_i(j) \cdot (\alpha(j) - \gamma(j) + \tau(j) \cdot W_i) \cdot \left(W_i - \sum_{m=1}^{J} e(j) \cdot B_i(j)\right)\right] \\
&\qquad \left. + \tau^2 \cdot \mathbb{E}\left[\left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right)^2\right]\right\} \\
&= \arg\min_{\tau,\gamma} \left\{ \mathbb{E}\left[\left(\sum_{j=1}^{J} B_i(j) \cdot (\alpha(j) - \gamma(j) + \tau(j) \cdot W_i)\right)^2\right] \right. \\
&\qquad - 2 \cdot \tau \cdot \mathbb{E}\left[\sum_{j=1}^{J} B_i(j) \cdot \tau(j) \cdot W_i \cdot \left(W_i - \sum_{m=1}^{J} e(j) \cdot B_i(j)\right)\right] \\
&\qquad \left. + \tau^2 \cdot \mathbb{E}\left[\left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right)^2\right]\right\}
\end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

because $\mathbb{E}[W_i|B_i(1),\ldots,B_i(J)] = \sum_{j=1}^J e(j)\cdot B_i(j)$. Minimizing this over τ leads to

$$\tau^* = \frac{\mathbb{E}\left[\left(\sum_{j=1}^J B_i(j)\cdot\tau(j)\cdot W_i\cdot\left(W_i - \sum_{m=1}^J e(j)\cdot B_i(j)\right)\right)^2\right]}{\mathbb{E}\left[\left(W_i - \sum_{j=1}^J e(j)\cdot B_i(j)\right)^2\right]}.$$

Because $\Pr(W_i = 1) = \sum_{j=1}^J q(j)\cdot e(j)$, and $\Pr(B_i(j) = 1|W_i = 1) = q(j)\cdot e(j)/\sum_{m=1}^J q(j)\cdot e(j)$, it follows that the numerator is equal to $\sum_{j=1}^J e(j)\cdot(1-e(j))\cdot q(j)\cdot\tau(j)$, and that the denominator is equal to $\sum_{j=1}^J e(j)\cdot(1-e(j))\cdot q(j)$, which finishes the proof of the first part of Theorem 9.1.

The first-order conditions for the estimators $(\hat{\tau}^{\rm ols}, \hat{\gamma}^{\rm ols})$ are

$$\sum_{i=1}^N \psi(Y_i^{\rm obs}, W_i, B_i(1),\ldots,B_i(J),\hat{\tau}^{\rm ols},\hat{\gamma}^{\rm ols}) = 0,$$

where

$$\psi(y,w,b(1)\ldots,b(J),\tau,\gamma) = \begin{pmatrix} \left(w - \sum_{j=1}^J e(j)\cdot b(j)\right)\cdot\left(y - \tau\cdot\left(w - \sum_{j=1}^J e(j)\cdot b(j)\right) - \sum_{j=1}^J \gamma(j)\cdot b(j)\right) \\ b(j)\cdot\left(y - \tau\cdot\left(w - \sum_{j=1}^J e(j)\cdot b(j)\right) - \sum_{j=1}^J \gamma(j)\cdot b(j)\right) \end{pmatrix}.$$

Given the population values of the parameters, τ^* and γ^*, standard M-estimation (or generalized method of moments) results imply that, under standard regularity conditions, the estimator is consistent and asymptotically normally distributed:

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

where the two components of the covariance matrix are

$$\Gamma = \mathbb{E}\left[\frac{\partial}{\partial(\tau,\gamma')}\psi(Y_i^{\rm obs}, W_i, B_i(1),\ldots,B_i(J),\tau,\gamma)\right]\Bigg|_{(\tau^*,\gamma^*)}$$

$$= \mathbb{E}\left[\begin{pmatrix} \sum_{j=1}^J e(j)\cdot(1-e(j))\cdot q(j) & 0 & \ldots & 0 \\ 0 & B_i(1) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & B_i(J) \end{pmatrix}\right],$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

and

$$\begin{aligned}\Delta &= \mathbb{E}\left[\psi(Y_i^{\text{obs}}, W_i, B_i(1), \ldots, B_i(J), \tau^*, \gamma^*) \cdot \psi(Y_i^{\text{obs}}, W_i, B_i(1), \ldots, B_i(J), \tau^*, \gamma^*)'\right] \\ &= \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tau^* \cdot \left(W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j)\right) - \sum_{j=1}^{J} \gamma^*(j) \cdot B_i(j)\right)^2 \right. \\ &\quad \left. \cdot \begin{pmatrix} W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j) \\ B_i(j) \end{pmatrix} \begin{pmatrix} W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j) \\ B_i(j) \end{pmatrix}'\right] \\ &= \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tau^* \cdot W_i - \sum_{j=1}^{J} \beta^*(j) \cdot B_i(j)\right)^2 \right. \\ &\quad \left. \cdot \begin{pmatrix} W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j) \\ B_i(j) \end{pmatrix} \begin{pmatrix} W_i - \sum_{j=1}^{J} e(j) \cdot B_i(j) \\ B_i(j) \end{pmatrix}'\right].\end{aligned}$$

The sampling variance of $\hat{\tau}$ is the $(1,1)$ element of the covariance matrix. Because Γ is block diagonal, the $(1,1)$ element of $\Gamma^{-1}\Delta(\Gamma')^{-1}$ is equal to the $(1,1)$ element of Δ divided by the square of the $(1,1)$ element of Γ. Hence the sampling variance of $\hat{\tau}$, normalized by the sample size N, is equal to

$$\frac{\mathbb{E}\left[\left(W_i - \sum_{j=1}^{J} q(j) \cdot B_i(j)\right)^2 \cdot \left(Y_i^{\text{obs}} - \tau^* \cdot W_i - \sum_{j=1}^{J} \beta^*(j) \cdot B_i(j)\right)^2\right]}{\left(\sum_{j=1}^{J} q(j) \cdot e(j) \cdot (1 - e(j))\right)^2}.$$

□

Proof of Theorem 9.2

First write the regression function as

$$Y_i^{\text{obs}} = \sum_{j=1}^{J} \alpha(j) \cdot B_i(j) + \sum_{j=1}^{J} \tau(j) \cdot W_i \cdot B_i(j) + \varepsilon_i.$$

Estimating the parameters of this regression function by OLS leads to

$$\hat{\tau}^{\text{ols}}(j) = \overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j),$$

which is unbiased and consistent for $\tau(j)$. Then transform the parameter vector from $\tau(J)$ to $\tau = \sum_{j=1}^{J} q(j) \cdot \tau(j)$, with inverse transformation $\tau(J) = (\tau - \sum_{j'=1}^{J-1} q(j') \cdot \tau(j'))/q(J)$. In terms of the parameters $\alpha(1), \ldots, \alpha(J)$, $\tau(1), \ldots, \tau(J-1)$ and τ, the regression function is equal to

$$Y_i^{\text{obs}} = \tau \cdot W_i \cdot \frac{B_i(J)}{q(J)} + \sum_{j=1}^{J} \alpha(j) \cdot B_i(j) + \sum_{j=1}^{J-1} \tau(j) \cdot W_i \cdot \left(B_i(j) - B_i(J) \cdot \frac{q(j)}{q(J)}\right) + \varepsilon_i.$$

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press

Thus $\hat{\tau}^{\text{ols}}$ is identical to $\sum_{j=1}^{J} q(j) \cdot \hat{\tau}^{\text{ols}}(j)$, and therefore is consistent for $\sum_{j=1}^{J} q(j) \cdot \tau(j) = \tau_{\text{sp}}$.

Because the sampling variance of $\hat{\tau}^{\text{ols}}(j)$ is $(\sigma_{\text{c}}^2(j)/((1-e(j)) \cdot q(j)) + \sigma_{\text{t}}^2(j)/(e(j) \cdot q(j)))/N$, the sampling variance of $\sum_{j=1}^{J} q(j) \cdot \hat{\tau}^{\text{ols}}(j)$, normalized by N, is $N \cdot \sum_{j=1}^{J} q(j)^2 \cdot \mathbb{V}(\hat{\tau}^{\text{ols}}(j))$, equal to $\sum_{j=1}^{J} q(j)^2(\sigma_{\text{c}}^2(j)/((1-e(j)) \cdot q(j)) + \sigma_{\text{t}}^2(j)/(e(j) \cdot q(j)))$. □

https://doi.org/10.1017/CBO9781139025751.010 Published online by Cambridge University Press