Chapter 6

Neyman's Repeated Sampling Approach

6.1 Introduction

In the last chapter we introduced the Fisher Exact P-value (FEP) approach for assessing sharp null hypotheses. As we saw, a sharp null hypothesis allowed us fill in the values for all missing potential outcomes in the experiment. This was the basis for deriving the randomization distributions of various statistics, the distributions induced by the random assignment of the treatments given fixed potential outcomes. During the same period in which Fisher was developing this method, Jerzey Neyman was instead focusing on methods for the estimation of, and inference for, average treatment effects, also using the distribution induced by randomization and repeated sampling from a population. In particular, he was interested in the long-run operating characteristics of statistical procedures under repeated sampling and randomizations. Thus, he attempted to find point estimators that were unbiased, and also interval estimators that had the specified nominal coverage in large samples. As noted before, focusing on average effects is different from from the focus of Fisher; the average effect across a population may be equal to zero even when some or even all unit-level treatment effects differ from zero. It is interesting to note that Neyman's analysis shows Fisher's analysis to be conservative in a certain asymptotic sense. Neyman's own proposal for estimating the sampling variance of the difference in average outcomes for treated and control units generally over-estimates the exact sampling variance, unless the treatment effect is constant and additive accross all units, which is automatically satisfied under Fisher's null hypothesis of no effects whatsoever.

Neyman's basic questions were the following. What would the average outcome be if all units were exposed to the active treatment, $\overline{Y}(1)$ in our notation? How did that compare to the average outcome if all units were exposed to the control treatment, $\overline{Y}(0)$ in our notation? Neyman's approach was to develop an estimator of the difference between these two averages and derive its mean and variance under repeated sampling. By repeated sampling we refer to the sampling generated by drawing from both the superpopulation (the potential outcomes), and from the randomization distribution (the assignment vector $\mathbf{W}$), although originally Neyman never described his analysis this way. His approach is similar to Fisher's, in that both consider the distribution of statistics (functions of the observed $\mathbf{W}$ and $\mathbf{Y}^{\text{obs}}$) under the randomization distribution, with all potential outcomes regarded as fixed. The similarity ends there. In Neyman's analysis, we do not start with an assumption that allows us to fill in all values of the missing potential outcomes, and so we cannot derive the exact randomization distribution of statistics of interest. However, we can often obtain good estimators of aspects of this distribution, for example, the first and second moments. In addition, we may be able to generalize to units not in the experiment. Neyman's primary concern was whether an estimator was unbiased for the average treatment effect. A secondary goal was to construct an interval estimator for the causal estimand, which he hoped to base on an unbiased estimator for the sampling variance of the average treatment effect estimator. Confidence intervals, as they were called later by Neyman (1934) , are stochastic intervals that are constructed in such a way that they include the true value of the estimand with proportion, over repeated draws, at least equal to some fixed probability, the confidence coefficient.

The remainder of this chapter is organized as follows. We begin by describing the data that will be used to illustrate the concepts discussed in this chapter. These data are from a study by Duflo and Hanna (2006) to asses the effect of teacher incentive program on teacher performance. Next, we introduce Neyman's estimator for the average treatment effect and show that it is unbiased for the average treatment effect, given a completely randomized experiment. We then calculate the sampling variance of this estimator and propose an estimator of this variance. There are several approaches one can take in this latter step, depending on whether one assumes a constant treatment effect.

Throughout the first part of this discussion, we assume that our interest is in a finite population of size N. Because we do not attempt to infer anything about units outside this population, it does not matter how it was selected; the entire analysis is conditional

on the population itself. In the second part of this discussion, we relax this assumption and instead consider, as did Neyman (1923), a super-population so that we can view the sample of N units as randomly drawn from this population. Given this shift in perspective, we reinterpret the original results, especially with respect to the choice of estimator for the sampling variance, and the associated large sample confidence interval. Throughout the chapter we maintain SUTVA.

We then apply this approach to the data from the teacher incentive experiment. We conclude by discussing how to apply Neyman's approach in the presence of discrete covariates (maintaining the setting of a completely randomized experiment). With discrete covariates, this approach is straightforward. One simply partitions the population into subpopulations based on the values of the covariate, conducts the analyses on each subpopulation, and uses the weighted average of these within-subpopulation treatment effects as the estimator of the average treatment effect within the population as a whole, with the weights proportional to the subpopulation sizes. With continuous covariates, this approach is infeasible. In this setting a more appealing alternative is to construct a model for the potential outcomes under each treatment level, derive an estimator of the average treatment effect under such a model, estimate the sampling variance of this estimator, and apply the same logic as in this chapter; such models are the topic of our next chapter.

6.2 The Duflo-Hanna-Ryan Teacher Incentive Experiment

To illustrate the methods discussed in this chapter, we use data from a randomized experiment conducted in rural India by Esther Duflo, Rema Hanna, and Steve Ryan (2008), designed to study the effect of financial incentives on teacher performance, both measured directly by teacher absences as well as by output measures, such as average class test scores. For the purpose of the study, a population of $N = 113$ single-teacher schools was selected. In a randomly selected subset of $N_t = 57$ schools, teachers were given a salary that was tied to attendance over a month long period, whereas in the remaining $N_c = 56$ schools the salary structure unchanged. In both the treatment and control schools, the teachers were given cameras with time stamps and asked to have students take pictures of the class with the teacher at the beginning and end of every school day. In addition, there were random

unannounced visits to the schools by program officials to see whether the school was open or not.

In the current chapter, to focus on Neyman's approach, we avoid considering issues of unintended missing data, and we use the 107 schools with recorded values for all five key variables, in addition to the treatment indicator: four outcomes and one covariate. Out of these 107 schools/teachers, 53 were chosen randomly to be given a salary schedule tied to teacher attendance, and 54 were not, and assigned to the control sample. In our analysis we use three outcome variables. The first is the proportion of times the school was open during a random visit (`open`). The second outcome is the percentage of students who completed a writing test (`pctpostwritten`). The third is the school average of the value of the writing test score (`written`), averaged over all the students who took the test. Even though not all students took the test, in all classes at least some students took the writing test at the end of the study. The fourth outcome variable is the average of the writing test score with zeros imputed for the students who did not take the test. We use one covariate in the analysis, the percentage of the students who took the written test prior to the study (`pctprewritten`).

Table 6.1 presents summary statistics for the data set. For all four variables (the pretreatment variables `pctprewritten`, and the three outcome variables `open`, `pctpostwritten`, and `written`), we present averages and standard deviations by treatment status, and the minimum and maximum over the full sample.

6.3 Unbiased Estimation of the Average Treatment Effect

Suppose we have a population consisting of N units. As before, for each unit there exist two potential outcomes, $Y_i(0)$ and $Y_i(1)$, corresponding to the outcome under control and treatment respectively. As with the Fisher Exact P-value (FEP) approach discussed in the previous chapter, the potential outcomes are assumed fixed. As before, the only random component is the treatment assignment, $\mathbf{W}$, which has a known distribution. The randomization distribution of the treatment assignment defines which potential outcome is observed for each unit.

Neyman was interested in the population average treatment effect:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big) = \overline{Y}(1) - \overline{Y}(0).$$

Suppose that we observe data from a completely randomized experiment in which $N_t = \sum_{i=1}^{N} W_i$ units are assigned to treatment and $N_c = \sum_{i=1}^{N}(1 - W_i)$ are assigned to control. Because of the randomization, a natural estimator for the average treatment effect is the difference in the average outcomes for those assigned to treatment versus those assigned to control:

$$\hat{\tau} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{\rm obs} - \frac{1}{N_c}\sum_{i:W_i=0} Y_i^{\rm obs} = \overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs},$$

where $\overline{Y}_c^{\rm obs} = \sum_{i:W_i=0} Y_i^{\rm obs}/N_c$ and $\overline{Y}_c^{\rm obs} = \sum_{i:W_i=1} Y_i^{\rm obs}/N_t$. First let us show that this estimator is unbiased for τ. Using the fact that $Y_i^{\rm obs} = Y_i(1)$ if $W_i = 1$, and $Y_i^{\rm obs} = Y_i(0)$ if $W_i = 0$, we can rewrite the statistic $\hat{\tau}$ as:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1 - W_i)\cdot Y_i(0)}{N_c/N}\right).$$

Because we view the potential outcomes as fixed, the only component in this statistic that is random is the treatment assignment, W_i. Given our completely randomized experiment (N units, with N_t randomly assigned to the treatment), by Section 3.5, $\Pr_W(W_i = 1) = \mathbb{E}_W[W_i] = N_t/N$. (Here we index the probability and expectation (and later the variance) operators by W to denote that the probability, expectation or variance is taken solely over the randomization distribution, keeping fixed the potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, and keeping fixed the population of size N.) Thus, $\hat{\tau}$ is unbiased for the average treatment effect τ:

$$\mathbb{E}_W\left[\hat{\tau}\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{E}_W[W_i]\cdot Y_i(1)}{N_t/N} - \frac{\mathbb{E}_W[1 - W_i])\cdot Y_i(0)}{N_c/N}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right) = \tau. \tag{6.1}$$

Note that, in terms of bias, the share of treated and control units in the randomized experiment is immaterial. This statement does not imply, however, that this share is irrelevant for inference; it can greatly affect the precision of the estimator, as we shall see in the next section.

For the teacher-incentive experiment, taking the proportion of days that the school was open (`open`) as the outcome of interest, this estimator for the average effect is

$$\hat{\tau} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} = 0.80 - 0.58 = 0.22,$$

as can be seen from the numbers in Table 6.1.

6.4 The Sampling Variance of $\hat{\tau}$

Neyman was also interested in constructing interval estimates for the average treatment effect, which he later (Neyman, 1934), called confidence intervals. This construction involves three steps. First, derive the sampling variance of the estimator for the average treatment effect. Second, develop estimators for this sampling variance. Third, appeal to a central limit argument for the large sample normality of $\hat{\tau}$ over its randomization distribution, and use its estimated sampling variances to create a large sample confidence interval for the average treatment effect τ.

In this section we focus on the first step, calculating the sampling variance of our estimator $\hat{\tau} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, under a completed randomized experiment. This derivation is relatively cumbersome because the assignments for different units are not independent in a completely randomized experiment. With the number of treated units fixed at N_t, the fact that unit 1 is assigned to the active treatment lowers the probability that unit 2 will receive the same active treatment. To show how to derive the sampling variance, we start with a simple example of only two units. We then expand our discussion to the more general case with N units and N_t randomly assigned to active treatment.

6.4.1 The Sampling Variance of the Unbiased Estimator With Two Units

Consider the simple case with one treated and one control unit. The average treatment effect in this case is

$$\tau = \frac{1}{2} \cdot \Big[(Y_1(1) - Y_1(0)) + (Y_2(1) - Y_2(0)) \Big]. \tag{6.2}$$

Assuming a completely randomized experiment—hence both units cannot receive the same treatment—it follows that $W_1 = 1 - W_2$. The estimator for the average treatment effect is

therefore:

$$\hat{\tau} = W_1 \cdot \Big(Y_1(1) - Y_2(0)\Big) + (1 - W_1) \cdot \Big(Y_2(1) - Y_1(0)\Big).$$

If unit 1 receives the treatment ($W_1 = 1$) our estimate of the average treatment effect will be $\hat{\tau} = Y_1^{\rm obs} - Y_2^{\rm obs} = Y_1(1) - Y_2(0)$. If on the other hand, $W_1 = 0$, the estimate will be $\hat{\tau} = Y_2^{\rm obs} - Y_1^{\rm obs} = Y_2(1) - Y_1(0)$.

To simplify the following calculations of the sampling variance of this estimator, define $D = 2 \cdot W_1 - 1$, so that $W_1 = (D + 1)/2$. It follows that $D \in \{-1, 1\}$, and $D^2 = 1$. Furthermore, because the expected value of the random variable W_1 is equal to 1/2, the expected value of D, over the randomization distribution, is $\mathbb{E}_W[D] = 0$ and the variance is $\mathbb{V}_W(D) = \mathbb{E}_W[D^2] = 1$. In terms of D, we can write our estimator $\hat{\tau}$ as:

$$\hat{\tau} = \frac{D+1}{2} \cdot \Big(Y_1(1) - Y_2(0)\Big) + \frac{1-D}{2} \cdot \Big(Y_2(1) - Y_1(0)\Big),$$

which can be rearranged as:

$$\begin{aligned}\hat{\tau} &= \frac{1}{2} \cdot \Big[\Big(Y_1(1) - Y_1(0)\Big) + \Big(Y_2(1) - Y_2(0)\Big)\Big] + \frac{D}{2} \cdot \Big[\Big(Y_1(1) + Y_1(0)\Big) - \Big(Y_2(1) + Y_2(0)\Big)\Big] \\ &= \tau + \frac{D}{2} \cdot \Big[\Big(Y_1(1) + Y_1(0)\Big) - \Big(Y_2(1) + Y_2(0)\Big)\Big].\end{aligned}$$

Because $\mathbb{E}_W[D] = 0$, we can see that $\hat{\tau}$ is unbiased for τ, (which we already established). However, this representation also makes the calculation of its sampling variance straightforward:

$$\begin{aligned}\mathbb{V}_W(\hat{\tau}) &= \mathbb{V}_W\left(\tau + \frac{D}{2} \cdot \Big[\Big(Y_1(1) + Y_1(0)\Big) - \Big(Y_2(1) + Y_2(0)\Big)\Big]\right) \\ &= \frac{1}{4} \cdot \mathbb{V}_W(D) \cdot \Big[\Big(Y_1(1) + Y_1(0)\Big) - \Big(Y_2(1) + Y_2(0)\Big)\Big]^2,\end{aligned}$$

because τ and the potential outcomes are fixed. Given that $\mathbb{V}_W(D) = 1$, it follows that the sampling variance of our estimator $\hat{\tau}$ is equal to:

$$\mathbb{V}_W(\hat{\tau}) = \frac{1}{4} \cdot \Big[\Big(Y_1(1) + Y_1(0)\Big) - \Big(Y_2(1) + Y_2(0)\Big)\Big]^2. \tag{6.3}$$

This representation of the sampling variance shows that this will be an awkward object to estimate, because it depends on all four potential outcomes, including products of potential outcomes for the same unit that are never jointly observed.

6.4.2 The Sampling Variance of $\hat{\tau}$ with N Units

To interpret expression (6.3), and to see how it can be estimated, we look at the general case with N units, of which N_t are randomly assigned to treatment. To calculate the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, we need the expectations of the second and cross moments of the treatment indicators W_i for $i = 1, \ldots, N$. Because $W_i \in \{0, 1\}$ is binary, $W_i^2 = W_i$, and thus

$$\mathbb{E}_W\left[W_i^2\right] = \mathbb{E}_W\left[W_i\right] = \frac{N_t}{N}, \qquad \text{and} \quad \mathbb{V}_W(W_i) = \frac{N_t}{N} \cdot \left(1 - \frac{N_t}{N}\right).$$

To calculate the cross moment in a completely randomized experiment, recall that with the number of treated units fixed at N_t, the two events—unit i being treated and unit j being treated—are not independent. Therefore $\mathbb{E}_W\left[W_i \cdot W_j\right] \neq \mathbb{E}_W\left[W_i\right] \cdot \mathbb{E}_W\left[W_j\right] = (N_t/N)^2$. Rather:

$$\mathbb{E}_W[W_i \cdot W_j] = \Pr_W(W_i = 1) \cdot \Pr_W(W_j = 1 | W_i = 1) = \frac{N_t}{N} \cdot \frac{N_t - 1}{N - 1}, \qquad \text{for} \quad i \neq j,$$

because conditional on $W_i = 1$ there are $N_t - 1$ treated units remaining, out of $N - 1$ units remaining.

Given the sampling moments derived above, we know the sampling variance and covariance of W_i and W_j. A straightforward but surprisingly long and tedious calculation (given in detail in Appendix B to this chapter) shows that the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ is equal to:

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}, \tag{6.4}$$

where S_c^2 and S_t^2 are the variances of $Y_i(0)$ and $Y_i(0)$ in the population, defined as:

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(0) - \overline{Y}(0)\Big)^2, \qquad \text{and} \quad S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - \overline{Y}(1)\Big)^2,$$

(and equal to zero if $N = 1$), and S_{tc}^2 is the population variance of the unit-level treatment effects, defined as:

$$S_{tc}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0))\Big)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0) - \tau\Big)^2.$$

Let us consider the interpretation of the three components of this variance in turn. The first two are fairly intuitive. Recall that the population average treatment effect is the

difference in average potential outcomes: $\tau = \overline{Y}(1) - \overline{Y}(0)$. To estimate τ, we first estimate $\overline{Y}(1)$, the population average potential outcome under treatment, by the average outcome for the N_t treated units, $\overline{Y}_t^{\rm obs}$. This estimator is unbiased for $\overline{Y}(1)$. The population variance of $Y_i(1)$ is $S_t^2 = \sum_i (Y_i(1) - \overline{Y}_1)^2/(N-1)$. Given this population variance for $Y_i(1)$, the sampling variance for an average from a sample of size N_t is $S_t^2/N_t = \sum_i (Y_i(1) - \overline{Y}(1))^2/(N_t(N-1))$. Similarly the average outcome for the N_c units assigned to control, $\overline{Y}_c^{\rm obs}$, is unbiased for the population average outcome under the control, $\overline{Y}(0)$, and its sampling variance is S_c^2/N_c. These results follow by direct calculation, or by using standard results from the analysis of simple random samples. Given a completely randomized experiment, the N_t treated units provide a simple random sample of the N values $Y_i(1)$ and the N_c control units provide a simple random sample of the N values $Y_i(0)$.

The third component of this sampling variance, S_{tc}^2/N, is the population variance of the unit-level treatment effects, $Y_i(1) - Y_i(0)$. If the treatment effect is constant in the population, this third term is equal to zero. If the treatment effect is not constant, S_{tc}^2 is positive. Because it is subtracted from the sum of the first two elements in the expression for the sampling variance of $\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}$, equation (6.4), the positive value for S_{tc}^2 reduces the sampling variance of this estimator for the average treatment effect.

There is an alternative representation of the sampling variance of $\hat{\tau}$ that is useful. First we write the variance of the unit-level treatment effect as a function of ρ_{tc}, the population correlation coefficient between the potential outcomes $Y_i(1)$ and $Y_i(0)$:

$$S_{tc}^2 = S_c^2 + S_t^2 - 2 \cdot \rho_{tc} \cdot S_c \cdot S_t,$$

where

$$\rho_{tc} = \frac{1}{(N-1) \cdot S_c \cdot S_t} \sum_{i=1}^{N} \big(Y_i(1) - \overline{Y}(1)\big) \cdot \big(Y_i(0)) - \overline{Y}(0)\big) \,. \tag{6.5}$$

By definition, ρ_{tc} is a correlation coefficient, and so lies in the interval $[-1, 1]$. Substituting this representation of S_{tc}^2 into equation (6.4), the alternative expression for the sampling variance of $\hat{\tau}$ (alternative to (6.4)) is:

$$\mathbb{V}_W \left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} \right) = \frac{N_t}{N \cdot N_c} \cdot S_c^2 + \frac{N_c}{N \cdot N_t} \cdot S_t^2 + \frac{2}{N} \cdot \rho_{tc} \cdot S_c \cdot S_t. \tag{6.6}$$

The sampling variance of our estimator is smallest when the potential outcomes are perfectly

negatively correlated ($\rho_{tc} = -1$), so that

$$S_{tc}^2 = S_c^2 + S_t^2 + 2 \cdot S_c \cdot S_t,$$

and

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \middle| \rho_{tc} = -1\right) = \frac{N_t}{N \cdot N_c} \cdot S_c^2 + \frac{N_c}{N \cdot N_t} \cdot S_t^2 - \frac{2}{N} \cdot S_c \cdot S_t,$$

and largest when the two potential outcomes are perfectly positively correlated ($\rho_{tc} = +1$), so that

$$S_{tc}^2 = S_c^2 + S_t^2 - 2 \cdot S_c \cdot S_t,$$

and

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \middle| \rho_{tc} = 1\right) = \frac{N_t}{N \cdot N_c} \cdot S_c^2 + \frac{N_c}{N \cdot N_t} \cdot S_t^2 + \frac{2}{N} \cdot S_c \cdot S_t$$

$$= \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{(S_c - S_t)^2}{N}. \tag{6.7}$$

The most notable special case of perfect correlation arises when the treatment effect is constant and additive. In that case,

$$\mathbb{V}_{\text{const}} = \mathbb{V}_W\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \middle| \rho_{tc} = 1, S_c^2 = S_t^2\right) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t}. \tag{6.8}$$

The fact that the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ is largest when the treatment effect is constant (*i.e.*, not varying) across units may be somewhat counterintuitive. Let us therefore return to the two-unit case and consider the form of the sampling variance in more detail. In this case, the sampling variance, presented in equation (6.3), is a function of the sum of the two potential outcomes for each of the two units. Consider two examples. In the first, $Y_1(0) = Y_1(1) = 10$, and $Y_2(0) = Y_2(1) = -10$. To calculate the correlation between the two potential outcomes we use expression (6.5) for ρ_{tc} and find the numerator equals

$$\frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - \overline{Y}(1)\Big) \cdot \Big(Y_i(0) - \overline{Y}(0)\Big)$$

$$= \Big((Y_1(1) - 0) \cdot (Y_1(0) - 0) + (Y_2(1) - 0) \cdot (Y_2(0) - 0)\Big) = 200,$$

and the two components of the denominator of ρ_{tc} equal

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(0) - \overline{Y}(0)\Big)^2 = \Big((10-0)^2 + (-10-0)^2\Big) = 200,$$

and

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - \overline{Y}(1)\Big)^2 = \Big((10-0)^2 + (-10-0)^2\Big) = 200,$$

so that the correlation between the two potential outcomes is 1. In the second example, suppose that $Y_1(0) = Y_2(1) = 10$, and $Y_1(0) = Y_2(1) = -10$. A similar calculation shows that the correlation between the two potential outcomes is now -1. In both examples, the average treatment effect is zero, but in the first, the treatment effect is constant and thus equal to 0 for each unit, whereas in the second case the treatment effect for unit 1 is equal to 20, and for unit 2 the treatment effect is equal to -20. As a result, when estimating the average treatment effect, in the first case the two possible values of the estimator are $Y_1^{\rm obs} - Y_2^{\rm obs} = 20$ (if $W_1 = 1$ and $W_2 = 0$) and $Y_2^{\rm obs} - Y_1^{\rm obs}$=-20 (if $W_1 = 0$ and $W_2 = 1$). In contrast, in the second case, the two values of the estimator are both equal to 0. Hence the sampling variance of the estimator in the first case, with $\rho_{tc} = +1$, is positive (in fact, equal to 400), whereas in the second case, with $\rho_{tc} = -1$, the sampling variance is 0.

6.5 Estimating the Sampling Variance

Now that we have derived the sampling variance of our estimator, $\hat{\tau} = \overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}$, given a completely randomized experiment, the next step is to develop an estimator of this sampling variance. To do this, we consider separately each of the three elements of the sampling variance, shown in equation (6.4).

The numerator of the first term, the population variance of the potential control outcome vector, $\mathbf{Y}(0)$, is equal to S_c^2. As shown in Appendix B, or from standard results on simple random samples, an unbiased estimator for S_c^2 is

$$s_c^2 = \frac{1}{N_c - 1}\sum_{i:W_i=0}\Big(Y_i(0) - \overline{Y}_c^{\rm obs}\Big)^2 = \frac{1}{N_c - 1}\sum_{i:W_i=0}\Big(Y_i^{\rm obs} - \overline{Y}_c^{\rm obs}\Big)^2.$$

Analogously, we can estimate S_t^2, the population variance of $Y_i(1)$, by

$$s_t^2 = \frac{1}{N_t - 1}\sum_{i:W_i=1}\Big(Y_i(1) - \overline{Y}_t^{\rm obs}\Big)^2 = \frac{1}{N_t - 1}\sum_{i:W_i=1}\Big(Y_i^{\rm obs} - \overline{Y}_t^{\rm obs}\Big)^2.$$

The third term, S_{tc}^2 (the population variance of the unit-level treatment effects) is more challenging to estimate because we never observe both $Y_i(1)$ and $Y_i(0)$ for the same unit. We therefore have no direct observations on the variation in the treatment effects across the population and therefore cannot directly estimate S_{tc}^2. As noted previously, if the treatment effects are constant and additive ($Y_i(1) - Y_i(0) = c$ for all i), then this component of the sampling variance is equal to zero and the third term vanishes. Under this circumstance we can obtain an unbiased estimator for the sampling variance as:

$$\hat{\mathbb{V}}_{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}. \tag{6.9}$$

This estimator for the sampling variance is widely used, even when the assumption of an additive treatment effect may be inaccurate. There are two main reasons for the popularity of this estimator for the sampling variance. First, by implicitly setting the third element of the estimated sampling variance equal to zero, the expected value of $\hat{\mathbb{V}}_{\text{neyman}}$ is at least as large as the true sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, irrespective of the heterogeneity in the treatment effect. Hence, in large samples, confidence intervals generated using this estimator of the sampling variance will have coverage at least as large, but not necessarily equal to, their nominal coverage.[1] (Note that this statement still needs to be qualified by the clause "in large samples," because we will rely on the central limit theorem to construct confidence intervals.) As mentioned in the introduction to this chapter, it is interesting to return to the discussion between Fisher and Neyman regarding the general interest in average treatment effects and sharp null hypotheses. Neyman's proposed estimator for the sampling variance is only unbiased in the case of a constant additive treatment effect, which is satisfied under the sharp null hypothesis of no treatment effects whatsoever, which was the case considered by Fisher. In other cases the proposed estimator of the sampling variance overestimates the true sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$. As a result, Neyman's interval is generally conservative. The second reason for using this estimator for the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ is that it is always unbiased for the sampling variance of $\hat{\tau}$ as an estimator of the super-population average treatment effect; we shall discuss this super-population interpretation, at greater length in Section 6.7.

[1]This potential difference between actual and nominal coverage of confidence intervals in randomized experiments concerned Neyman, and probably with this in mind, he formally defined confidence intervals in 1934 to allow for the possibility that the actual coverage could be greater than the nominal coverage. Thus the proposed conservative intervals are still valid.

In the remainder of this section, we consider two alternative estimators for the sampling variance of $\hat{\tau}$. The first explicitly allows for treatment effect heterogeneity. Under treatment effect heterogeneity, the estimator for the sampling variance in equation (6.9), $\hat{\mathbb{V}}_{\text{neyman}}$, provides an upwardly biased estimate: the third term, which vanishes if the treatment effect is constant, is now negative. The question arises whether we can improve upon the Neyman variance estimator without risking undercoverage.

To see that there is indeed information to do so, recall the argument above that an implication of constant treatment effects is that the variances S_c^2 and S_t^2 are equal. A difference between these variances, which would in large samples lead to a difference in the corresponding estimates s_c^2 and s_t^2, indicates variation in the treatment effect.

To use this information to create a better estimator for the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, let us turn to the representation of the sampling variance in equation (6.6), which incorporates ρ_{tc}, the population correlation coefficient between the potential outcomes:

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right) = S_c^2 \cdot \frac{N_t}{N \cdot N_c} + S_c^2 \cdot \frac{N_c}{N \cdot N_t} + \rho_{tc} \cdot S_c \cdot S_t \cdot \frac{2}{N}.$$

Conditional on a value for the correlation coefficient, ρ_{tc}, we can estimate this sampling variance as

$$\hat{\mathbb{V}}_{\rho_{tc}} = s_c^2 \cdot \frac{N_t}{N \cdot N_c} + s_t^2 \cdot \frac{N_c}{N \cdot N_t} + \rho_{tc} \cdot s_c \cdot s_t \cdot \frac{2}{N}. \tag{6.10}$$

This variance is again largest if the two potential outcomes are perfectly correlated, that is, $\rho_{01} = 1$. An alternative conservative estimator of the sampling variance that exploits this bound is:

$$\begin{aligned}\hat{\mathbb{V}}_{\rho_{tc}=1} &= s_c^2 \cdot \frac{N_t}{N \cdot N_c} + s_1^2 \cdot \frac{N_c}{N \cdot N_t} + s_c \cdot s_t \cdot \frac{2}{N} \\ &= \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} - \frac{(s_t - s_c)^2}{N}. \end{aligned} \tag{6.11}$$

If s_c^2 and s_t^2 are unequal, then $\hat{\mathbb{V}}_{\rho_{tc}=1}$ will be smaller than $\hat{\mathbb{V}}_{\text{neyman}}$. Because $\hat{\mathbb{V}}_{\rho_{tc}=1}$ is still conservative, in large samples it will lead to tighter confidence intervals for τ than using the sampling variance estimate $\hat{\mathbb{V}}_{\text{neyman}}$. Using $\hat{\mathbb{V}}_{\rho_{tc}=1}$ to construct confidence intervals will result is tighter confidence intervals than using $\hat{\mathbb{V}}_{\text{neyman}}$, without compromising their validity. The intervals based on $\hat{\mathbb{V}}_{\rho_{tc}=1}$ will still be conservative because $\hat{\mathbb{V}}_{\rho_{tc}=1}$ is still upwardly biased when the true correlation is smaller than one, although less so than $\hat{\mathbb{V}}_{\text{neyman}}$. Note, however,

that with no information beyond the fact that $S_c^2 \neq S_t^2$, all choices for ρ_{tc} smaller than unity raise the possibility that we will underestimate the sampling variance and construct invalid confidence intervals.

Next consider sampling variance estimation under the additional assumption that the treatment effect is constant, $Y_i(1) - Y_i(0) = \tau$ for all i. valid only if the treatment effect is in fact constant. In that case the third term in (6.11) is zero, and we can simply estimate the variance as

$$\frac{s_c^2}{N_c} + \frac{s_t^2}{N_t},$$

which is equal to $\hat{\mathbb{V}}_{\text{neyman}}$. However, an alternative is to exploit the fact that under the constant treatment assumption, the population variance of the two potential outcomes, S_t^2 and S_t^2, must be equal. We can therefore define $S^2 \equiv S_c^2 = S_t^2$ and pool the treated and control vectors to estimate this common variance:

$$\begin{aligned} s^2 &= \frac{1}{N-2} \cdot \Big(s_c^2 \cdot (N_c - 1) + s_t^2 \cdot (N_t - 1) \Big) \\ &= \frac{1}{N-2} \cdot \left(\sum_{i:W_i=0} \left(Y_i^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right)^2 + \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2 \right). \end{aligned} \tag{6.12}$$

The larger sample size for this estimator (from N_c and N_t for s_c^2 and s_t^2 respectively, to N for s^2), leads to a more precise estimator for the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$:

$$\hat{\mathbb{V}}_{\text{const}} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t} \right). \tag{6.13}$$

If the treatment effect is constant, this estimator is preferable to either $\hat{\mathbb{V}}_{\text{neyman}}$ or $\hat{\mathbb{V}}_{\rho_{tc}=1}$, but if not, it need not be valid. Both $\hat{\mathbb{V}}_{\text{neyman}}$ and $\hat{\mathbb{V}}_{\rho_{tc}=1}$ are valid generally, and therefore may be preferred.

Let us return to the Duflo-Hanna-Ryan teacher incentive data. The estimate for the average effect of assignment to the incentives-based salary rather than the conventional salary structure, on the likelihood that the school is open, is, as discussed in the previous section, equal to 0.22. Now let us consider estimators for the sampling variance. First we estimate the sample variances S_c^2, S_t^2, and the combined variance S^2. The estimates are

$$s_c^2 = 0.19^2, \qquad s_t^2 = 0.13^2, \qquad \text{and} \quad s^2 = 0.16^2.$$

The two sample variance s_c^2 and s_t^2 are quite different, with their ratio being larger than two.

Next we use the sample variances of the potential outcomes to estimate the sampling variance for the average treatment effect estimator. The first estimate for the sampling variance, which is in general conservative but allows for unrestricted treatment effect heterogeneity, is

$$\hat{\mathbb{V}}_{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} = 0.0311^2.$$

(We report four digits behind the decimal point to make explicit the small differences between the various estimators for the sampling variance, although in practice one would only report two or three digits.) The second estimate, still conservative, but exploiting differences in the variance of the outcome by treatment group, and again allowing for unrestricted treatment effect heterogeneity, is

$$\hat{\mathbb{V}}_{\rho_{tc}=1} = s_c^2 \cdot \frac{N_t}{N \cdot N_c} + s_t^2 \cdot \frac{N_c}{N \cdot N_t} + s_c \cdot s_t \cdot \frac{2}{N} = 0.0305^2.$$

By construction this estimator is smaller than $\hat{\mathbb{V}}_{\text{neyman}}$. However, even though the variances s_c^2 and s_t^2 differ by more than a factor two, the difference in the estimated sampling variances $\hat{\mathbb{V}}_{\rho_{tc}=1}$ and $\hat{\mathbb{V}}_{\text{neyman}}$ is very small in this example, less than 1%. In general, the standard variance $\hat{\mathbb{V}}_{\text{neyman}}$ is unlikely to be substantially larger than $\hat{\mathbb{V}}_{\rho_{tc}=1}$, as suggested by this artificial example. The third and final estimate of the sampling variance, which relies on a constant treatment effect for its validity, is

$$\hat{\mathbb{V}}_{\text{const}} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t} \right) = 0.0312^2.$$

6.6 Confidence Intervals and Testing

In the introduction we noted that Neyman's interest in estimating the precision of the estimator for the average treatment effect was largely driven by an interest in constructing confidence intervals. By a confidence interval with confidence coefficient $1-\alpha$, we mean an interval $[C_L(\mathbf{Y}^{\text{obs}}, \mathbf{W}), C_U(\mathbf{Y}^{\text{obs}}, \mathbf{W})]$, such that

$$\Pr_W(C_L(\mathbf{Y}^{\text{obs}}, \mathbf{W}) \leq \tau \leq C_U(\mathbf{Y}^{\text{obs}}, \mathbf{W})) \geq 1-\alpha.$$

The only reason the lower and upper bounds in this interval are random is through their dependence on $\mathbf{W}$. Its distribution is therefore generated by the randomization. Note that,

in this expression, the probability of including the true value τ may exceed $1-\alpha$, in which case the interval is valid but conservative. Here we discuss a number of ways to construct such confidence intervals and to conduct tests for hypotheses concerning the average treatment effect. We will use the Duflo-Hanna-Ryan data to illustrate the steps of Neyman's approach.

6.6.1 Confidence Intervals

Let $\hat{\mathbb{V}}$ be an estimate of the sampling variance of $\hat{\tau}$ under the randomization distribution (we recommend using $\hat{\mathbb{V}}_{\text{neyman}}$). Suppose we wish to construct a 90% confidence interval. We use a normal approximation to the randomization distribution of $\hat{\tau}$. This approximation is somewhat inconsistent with our stress on finite sample properties of the estimator for τ and its sampling variance, but it is driven by the lack of *a priori* information about the joint distribution of the potential outcomes. As we will see, normality is often a good approximation to the randomization distribution of standard estimates, even in fairly small samples. To further improve on this approximation, we could approximate the distribution of $\hat{\mathbb{V}}_{\text{neyman}}$by a chi-squared distribution, and then use that to approximate the distribution of $\hat{\tau}/\sqrt{\hat{\mathbb{V}}_{\text{neyman}}}$ by a t-distribution. For simplicity here, we use the 5th and 95th percentile of the standard normal distribution, -1.645 and 1.645, to calculate the central 90% confidence interval as:

$$\text{CI}^{\tau}_{0.90} = \left(\hat{\tau} - 1.645 \cdot \sqrt{\hat{\mathbb{V}}}, \hat{\tau} + 1.645 \cdot \sqrt{\hat{\mathbb{V}}}\right).$$

More generally, if we wish to construct a central confidence interval with confidence level $(1-\alpha)\times 100\%$, as usual we look up the $\alpha/2$ and $1-\alpha/2$ quantiles of the standard normal distribution, denoted by $c_{\alpha/2}$, and construct the confidence interval:

$$\text{CI}^{\tau}_{1-\alpha} = \left(\hat{\tau} + c_{\alpha/2} \cdot \sqrt{\hat{\mathbb{V}}}, \hat{\tau} + c_{1-\alpha/2} \cdot \sqrt{\hat{\mathbb{V}}}\right).$$

This approximation applies when using any estimate of the sampling variance, and, in large samples, the resulting confidence intervals are valid under the same assumptions that make the corresponding estimator for the sampling variance an unbiased or upwardly biased estimator of the true sampling variance.

Based on the three sampling variance estimates reported in the previous section for the outcome that the school is open, we obtain the following three 90% confidence intervals.

First, based on $\hat{\mathbb{V}}_{\text{neyman}} = 0.0311^2$, we get

$$\text{CI}^{\tau}_{0.90,\text{neyman}} = \left(\hat{\tau} + c_{0.10/2} \cdot \sqrt{\hat{\mathbb{V}}_{\text{neyman}}}, \hat{\tau} + c_{1-0.10/2} \cdot \sqrt{\hat{\mathbb{V}}_{\text{neyman}}}\right)$$

$$= (0.2154 - 1.645 \cdot 0.0311, 0.2154 + 1.645 \cdot 0.0311) = (0.1642, 0.2667).$$

Second, based on the sampling variance estimator assuming a constant treatment effect, $\hat{\mathbb{V}}_{\text{const}} = 0.0312^2$, we get a very similar interval,

$$\text{CI}^{\tau}_{0.90,\text{const}} = (0.1640, 0.2668).$$

Finally, based on the third sampling variance estimator, $\hat{\mathbb{V}}_{\rho_{tc}=1} = 0.0305^2$, we get again a fairly similar interval,

$$\text{CI}^{\tau}_{0.90,\rho_{tc}=1} = (0.1652, 0.2657).$$

With the estimates for the sampling variances so similar, the three confidence intervals are also very similar.

6.6.2 Testing

We can also use the sampling variance estimates to carry out tests of hypotheses concerning the average treatment effect. Suppose we wish to test the null hypothesis that the average treatment effect is zero against the alternative that it differs from zero:

$$H_0^{\text{neyman}}: \qquad \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = 0$$

$$H_a^{\text{neyman}}: \qquad \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) \neq 0$$

A natural test statistic to use for Neyman's "average null" is the the ratio of the point estimate to the estimated standard error. For the teacher incentive data, the point estimate is $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} = 0.2154$. The estimated standard error is, using the conservative estimator for the sampling variance $\hat{\mathbb{V}}_{\text{neyman}}$, equal to 0.0311. The resulting t-statistic is therefore

$$t = \frac{\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}}{\sqrt{\hat{\mathbb{V}}_{\text{neyman}}}} = -\frac{0.2154}{0.0311} = 6.9.$$

The associated p-value for a two-sided test is $2 \cdot (1 - \Phi(6.9)) < 0.001$. At conventional significance levels, we clearly reject the (Neyman) null hypothesis that the average treatment effect is zero.

It is interesting to compare this test based on Neyman's approach to the FEP approach. There are two important differences between the two tests. First, they assess different null hypotheses, for example a zero average effect for Neyman versus a zero effect for all units for Fisher. Second, the Neyman test relies on a large sample normal approximation for its validity, whereas the FEP approach is exact.

Let us discuss both differences in more detail. First consider the difference in hypotheses. The Neyman test assesses whether the average treatment effect is zero, whereas the FEP assesses whether the treatment effect is zero for all units in the experiment. Formally, in the Fisher approach the null hypothesis is

$$H_0^{\text{fisher}}: \qquad Y_i(1) - Y_i(0) = 0 \qquad \text{for all } i = 1, \ldots, N,$$

and the (implicit) alternative hypothesis is

$$H_a^{\text{fisher}}: \qquad Y_i(1) - Y_i(0) \neq 0 \qquad \text{for some } i = 1, \ldots, N.$$

Depending on the implementation of the FEP approach, this difference in null hypotheses may be unimportant. If we choose to use a test statistic proportional to the average difference, we end up with a test that has virtually no power against alternatives with heterogenous treatment effects that average out to zero. We would have power against at least some of those alternatives if we choose a different statistic. Consider as an example, a population where for all units $Y_i(0) = 2$. For 1/3 of the units the treatment effect is 2. For 2/3 of the units the treatment effect is -1. In this case the Neyman null hypothesis of a zero average effect is true. The Fisher null hypothesis of no effect whatsoever is not true. Whether we can detect this violation depends on the choice of statistic. The FEP approach with the statistic based on the average difference in outcomes by treatment status has no power against this alternative. However, the FEP approach with a different statistic, based on transforming the outcomes by taking logarithms, does have power in this setting. In this artificial example, the expected difference in logarithms by treatment status is -0.23. The FEP based on the difference in average logarithms will detect this difference in large enough

samples. The Neyman approach does not gives us a choice of test statistic, unless we wish to change the null hypothesis.

The second difference between the two procedures is in the approximate nature of the Neyman test, compared to the exact results for the FEP approach. We use two approximations in the Neyman approach. First we use the estimated variance (e.g., $\hat{\mathbb{V}}_{\text{neyman}}$) instead of the exact variance ($\mathbb{V}_W(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}})$). Second, we use a normal approximation for the repeated sampling distribution of the difference in averages $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$. Both approximations have a justification in large samples. If the sample is reasonably large, and if there are few or not outliers, as in the application in this chapter, these approximations will likely be accurate.

6.7 Inference for Super-population Average Treatment Effects

In the introduction to this chapter, we commented on the distinction between a finite population interpretation, in which the sample of size N is considered the population of interest, and a super-population interpretation, in which the N observed units are considered a random sample from a larger population. The second argument in favor of using the sampling variance estimator $\hat{\mathbb{V}}_{\text{neyman}}$ in equation (6.9) is that, regardless of the level of heterogeneity in the unit-level treatment effect, $\hat{\mathbb{V}}_{\text{neyman}}$ is unbiased for the sampling variance of the estimator $\hat{\tau}$ for the super-population average treatment effect. Here we further explore this argument, address how it affects our interpretation of the estimator of the average treatment effect, and discuss the various choices of estimators for its sampling variance.

Suppose that the population of N subjects taking part in the completely randomized experiment is itself a simple random sample from a larger population. For simplicity, we assume that this "super-population" is infinitely large. This is a slight departure from Neyman's explicit focus on the average treatment effect for a finite population. In many cases this change of focus is immaterial. Although in some agricultural experiments farmers may be genuinely interested in which fertilizer was best for their specific fields the year of the experiment, in most social and medical science settings, experiments are, explicitly or implicitly, conducted with a view to informing policies for a larger population of units. However, without additional information, we cannot hope to obtain more precise estimates

for the effects in the super-population than for the effects in the sample. In fact, the estimates for the super-population estimands are typically strictly less precise. Ironically it it is exactly this loss in precision that enables us to obtain unbiased estimates of the sampling variance of the traditional estimator for the average treatment effect in the super-population.

Viewing our N units as a sample of the target population, rather than viewing them as the population itself, induces a distribution on the two potential outcomes. The pair of potential outcome values of an observed unit i is simply one draw from the distribution within the full population and is, therefore, itself stochastic. The distribution of the pair of two potential outcomes in turn induces a distribution on the unit-level treatment effect and on the average of the unit-level treatment effect within the drawn sample. To be clear about this superpopulation perspective, we use the subscript SP to denote the super-population average treatment effect and FS to denote the finite sample average treatment effect:

$$\tau_{\text{SP}} = \mathbb{E}_{\text{SP}}\left[Y_i(1) - Y_i(0)\right] \qquad \text{and} \quad \tau_{\text{FS}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right).$$

Analogously, the subscript SP to the expectations operator indicates that the expectation is taken over the distribution generated by random sampling from the superpopulation, and not solely over the randomization distribution. Thus $\tau_{\text{SP}} = \mathbb{E}_{\text{SP}}[Y_i(1) - Y_i(0)]$ is the expected value of the unit-level treatment effect, under the distribution induced by sampling from the superpopulation, or, equivalently, the average treatment effect in the super-population. See Appendix A for details on the superpopulation perspective. Let σ^2_{tc} be the variance of the unit-level treatment effect within this super-population, $\sigma^2_{tc} = \mathbb{V}_{\text{SP}}(Y_i(1) - Y_i(0)) = \mathbb{E}_{\text{SP}}[(Y_i(1) - Y_i(0) - \tau_{\text{SP}})^2]$, and let σ^2_c and σ^2_t denote the population variances of the two potential outcomes, or the super-population expectations of S^2_c and S^2_t: $\sigma^2_c = \mathbb{V}_{\text{SP}}(Y_i(0)) = \mathbb{E}\left[(Y_i(0) - \mathbb{E}_{\text{SP}}[Y_i(0))^2\right]$ and $\sigma^2_t = \mathbb{V}_{\text{SP}}(Y_i(1)) = \mathbb{E}\left[(Y_i(1) - \mathbb{E}_{\text{SP}}[Y_i(1))^2\right]$.

As stated above, the sample of size N, previously the target population, is now assumed to be a simple random sample from this target super-population. This implies that the average treatment effect within the sample, that is, the sample average treatment effect, $\tau_{\text{FS}} = \overline{Y}(1) - \overline{Y}(0)$, can be viewed as a random variable with expectation equal to the population average treatment effect in the super-population, τ_{SP}:

$$\mathbb{E}_{\text{SP}}\left[\tau_{\text{FS}}\right] = \mathbb{E}_{\text{SP}}\left[\overline{Y}(1) - \overline{Y}(0)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\text{SP}}\left[Y_i(1) - Y_i(0)\right] = \tau_{\text{SP}}. \tag{6.14}$$

Thus, (6.14) states that, assuming that the N units in our sample arose from a simple random sample, the average treatment effect in the sample is unbiased for the average treatment effect in the super-population.

The definition of the variance of the unit-level treatment effect within the super-population, σ^2_{tc}, implies that the variance of $\tau_{\rm FS}$ across repeated random samples is equal to

$$\mathbb{V}_{\rm SP}(\tau_{\rm SP}) = \mathbb{V}_{\rm SP}\left(\overline{Y}(1) - \overline{Y}(0)\right) = \sigma^2_{tc}/N. \tag{6.15}$$

Now let us consider the sampling variance of the standard estimator for the average treatment effect, $\hat{\tau} = \overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}$, given this sampling from a super-population. The expectation and variance operator without subscript denote expectations and variances taken over both the randomization distribution and the random sampling. We have

$$\begin{aligned}\mathbb{V}\left(\hat{\tau}\right) &= \mathbb{V}\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}\right) = \mathbb{E}\left[\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - \mathbb{E}\left[\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - \mathbb{E}_{\rm SP}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)^2\right],\end{aligned}$$

where the second equality holds because $\mathbb{E}\left[\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}\right]$ and $\mathbb{E}_{\rm SP}[\overline{Y}(1) - \overline{Y}(0)]$ are both equal to $\tau_{\rm SP}$, as shown above. Adding and subtracting $\overline{Y}(1) - \overline{Y}(0)$ within the expectation, this sampling variance, over both randomization and random sampling, is equal to:

$$\begin{aligned}\mathbb{V}\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}\right) &= \mathbb{E}\left[\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - \left(\overline{Y}(1) - \overline{Y}(0)\right) + \left(\overline{Y}(1) - \overline{Y}(0)\right) - \mathbb{E}_{\rm SP}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - (\overline{Y}(1) - \overline{Y}(0))\right)^2\right] \\ &\quad + \mathbb{E}_{\rm SP}\left[\left(\left(\overline{Y}(1) - \overline{Y}(0)\right) - \mathbb{E}_{\rm SP}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)^2\right] \\ &\quad + 2\cdot\mathbb{E}\left[\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - \left(\overline{Y}(1) - \overline{Y}(0)\right)\right)\cdot\left(\left(\overline{Y}(1) - \overline{Y}(0)\right) - \mathbb{E}_{\rm SP}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)\right].\end{aligned}$$

The third term of this last equation is equal to zero because the expectation of the first factor, $\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - (\overline{Y}(1) - \overline{Y}(0))$, conditional on the N–vectors $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, is zero. Hence the sampling variance reduces to:

$$\begin{aligned}&\mathbb{V}\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}\right) = \\ &\qquad \mathbb{E}\left[\left(\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs} - \overline{Y}(1) - \overline{Y}(0)\right)^2\right] + \mathbb{E}_{\rm SP}\left[\left(\overline{Y}(1) - \overline{Y}(0) - \mathbb{E}_{\rm SP}\left[Y(1) - Y(0)\right]\right)^2\right].\end{aligned}$$

(6.16)

In equation (6.1) we showed that $\mathbb{E}_W\left[\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \middle| \mathbf{Y}(0), \mathbf{Y}(1)\right] = \overline{Y}(1) - \overline{Y}(0)$, hence by iterated expectations, the first term on the right side is equal to the expectation of the conditional variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ (conditional on the N-vector of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$). This is equal to

$$\mathbb{E}_W\left[\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2 \middle| \mathbf{Y}(0), \mathbf{Y}(1)\right] = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}, \tag{6.17}$$

as in equation (6.4). Recall that these earlier calculations were made when assuming that the sample N was the target population, and thus were conditional on $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$. The expectation of (6.17) over the distribution of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ generated by sampling from the superpopulation is

$$\begin{aligned}
&\mathbb{E}\left[\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2\right] \\
&\quad = \mathbb{E}_{\text{SP}}\left[\mathbb{E}_W\left[\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2 \middle| \mathbf{Y}(0), \mathbf{Y}(1)\right]\right] \\
&\quad = \mathbb{E}_{\text{SP}}\left[\frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}\right] = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t} - \frac{\sigma_{tc}^2}{N}.
\end{aligned}$$

The expectation of the second term on the right side of equation (6.16), is equal to σ_{tc}^2/N, as we saw in equation (6.15). Thus the sampling variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ over sampling from the superpopulation equals:

$$\mathbb{V}_{\text{SP}} = \mathbb{V}\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right) = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t}, \tag{6.18}$$

which we can estimate by substituting s_c^2 and s_t^2 for σ_c^2 and σ_t^2, respectively:

$$\hat{\mathbb{V}}_{\text{SP}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}.$$

Notice that $\hat{\mathbb{V}}_{\text{SP}}$ is identical to the previously introduced conservative estimator of the sampling variance for the finite population average treatment effect estimator, $\hat{\mathbb{V}}_{\text{neyman}}$, presented in Equation 6.9. Under simple random sampling from the super-population, the expected value of the estimator $\hat{\mathbb{V}}_{\text{neyman}}$ equals $\mathbb{V}_{\text{SP}}$. Hence, considering the N observed units as a simple random sample from an infinite super-population, the estimator in (6.9) is an unbiased estimate of the sampling variance of the estimator of the super-population

average treatment effect. Neither of the alternative estimators—$\hat{\mathbb{V}}_{\text{const}}$ in equation (6.13), which exploits the assumption of a constant treatment effect, nor $\hat{\mathbb{V}}_{\rho_{tc}=1}$ in equation (6.11), derived through bounds on the correlation coefficient—have this attractive quality. Thus, despite the fact that $\hat{\mathbb{V}}_{\text{const}}$ may be a better estimator of the sampling variance when the treatment effect is constant, and $\hat{\mathbb{V}}_{\rho_{tc}=1}$ may be a better estimator of $\mathbb{V}_{\text{FS}}$, $\hat{\mathbb{V}}_{\text{neyman}}$ is used far more frequently in practice in our experience, although the logic for it appears to be rarely articulated.

6.8 Neyman's Approach With Covariates

One can easily extend Neyman's approach for estimating average treatment effects to settings with discrete covariates. In this case, one would partition the full samples into subsamples defined by the values of the covariate and then conduct the analysis separately within these subpopulations. The resulting within-subpopulation estimators would be unbiased for the within-subpopulation average treatment effect. Taking an average of these estimates, weighted by subpopulation sizes, gives an unbiased estimate of the overall average treatment effect. As we see in Chapter 9, we in fact consider this method in the discussion on stratified random experiments.

It is impossible, however, to derive estimators that are exactly unbiased under the randomization distribution, conditional on the covariates, if there are covariate values for which we have only treated or only control units. In settings with covariates that take on many values, this is likely to happen with great frequency. In such settings building a model for the potential outcomes, and using this model to create an estimator of the average treatment effect, becomes a more appealing option. We turn to this topic in the next two chapters.

6.9 Results for the Duflo-Hanna Teacher Incentive Data

Now let us return to the teacher incentive data and systematically look at the results based on the methods discussed in the current chapter. We analyze five "outcomes" in turn. For illustrative purposes we report here a number of point, sampling variance and interval estimates. It should be kept in mind that the confidence intervals will only have their nominal coverage if only one is evaluated. The first variable we analyze as an outcome is a

pre-treatment variable, and so we know *a priori* that the causal effect of the treatment on this variable is zero, both at the unit level and on average. In general it can be useful to carry out such analyses as a check on the success of the randomization: that is, we know here that the Fisher null hypothesis of no effect whatsoever is true. The pre-treatment variable is `pctprewritten`, the percentage of students in a school that took the pre-program writing test. For this variable, we estimate, as anticipated, the average effect to be small, with a 95% confidence interval that comfortably includes zero.

Next we turn to the four "real" outcomes. First, the causal effect of the attendance-related salary incentives on the proportion of days that the school was open during the of days it was subject to a random check. The estimated effect is 0.22, with a 95% confidence interval of $[0.15, 0.28]$. It is clear that the attendance-related salary incentives lead to a higher proportion of days with the school open. We also look at the effect on the percentage of students in the school who took the written test, `pctpostwritten`. Here the estimated treatment effect is 0.05, with a 95% confidence interval of $[-0.03, 0.13]$. The effect is not distinguishable from zero at the 95% significance level, but it is at the 90% significance level. Next, we look at the average score on the writing test, which leads to a point estimate of 0.17, with a 95% confidence interval of $[0.00, 0.34]$. Finally, we examine the average test score, assigning zeros to students not taking the test. Now we estimate an average effect of 0.14, with a 95% confidence interval of $[0.00, 0.28]$.

In the final analysis, we look at estimates separately for two subpopulations, defined by whether the proportion of students taking the initial writing test was zero or positive, to illustrate the application of the methods developed in this chapter to subpopulations defined by covariates. These analyses are for illustrative purposes only, and we do not take account of the fact that we do multiple tests here. The first subpopulation (`pctprewritten`$= 0$) comprised 40 schools (37%) and the second (`pctprewritten`> 0) 67 schools (63%). We analyze separately the effect of assignment to attendance-based teacher incentives on all four outcomes. The results are reported in Table 6.4. The main substantive finding is that the effect of the incentive scheme on writing skills appears higher for schools where many students entered with some writing skills (sufficient to at least take the initial test). Although the 95% confidence interval includes zero (-0.41,0.05), the difference from the effect for the group with no students taking the initial test is almost signicant at the 10% level. The 90% confidence interval is (-0.37,0.01).

6.10 Conclusion

In this chapter we discussed Neyman's approach to estimation and inference in completely randomized experiments. He was interested in assessing the operating characteristics of statistical procedures under repeated sampling. Neyman focused on the average effect of the treatment. He proposed an estimator for the average treatment effect in the finite sample, and showed that it was unbiased under repeated sampling. He also derived the sampling variance for this estimator. Finding an estimator for this sampling variance that itself is unbiased turned out to be impossible in general. Instead Neyman showed that the standard estimator for the sampling variance of this estimator is positively biased, unless the treatment effects are constant and additive, in which case it is unbiased. Like Fisher's approach, Neyman's methods have great appeal in the settings where they apply. However, again like Fisher's methods, there are many situations where we are interested in questions beyond those answered by their approaches. For example, we may want to estimate average treatment effects adjusting for differences in covariates in settings where some covariate values appear only in treatment or control groups. In the next two chapters we discuss methods that do not have the exact (finite sample) statistical properties that make the Neyman and Fisher approaches so elegant in their simplicity, but that do answer additional questions, albeit under additional assumptions.

Appendix A: Random Sampling from a Superpopulation

In this chapter we introduced the superpopulation perspective. In this Appendix we provide more details of this approach, and the differences from the finite population perspective. Let $N_{\rm SP}$ be the size of the superpopulation, with $N_{\rm SP}$ large, but finite. Each unit in this population is characterized by the triple $(Y_i(0), Y_i(1))$, for $i = 1, \ldots, N_{\rm SP}$. Let $\mathbf{Y}_{\rm SP}(0)$ and $\mathbf{Y}_{\rm SP}(1)$ denote the $N_{\rm SP}$-component vectors with ith element equal to $Y_i(0)$ and $Y_i(1)$ respectively. We continue to view these potential outcomes as fixed. Our finite sample is a Simple Random Sample (SRS) of size N from this large superpopulation. We take N as fixed. Let R_i denote the sampling indicator, so that $R_i = 1$ if unit i is sampled, and $R_i = 0$ if unit i is not sampled, with $\sum_{i=1}^{N_{\rm SP}} R_i = N$. The sampling indicator is a binomial random variable with mean $N/N_{\rm SP}$ and variance $(N/N_{\rm SP}) \cdot (1 - N/N_{\rm SP})$. The covariance between R_i and R_j, for $i \neq j$, is $-(N/N_{\rm SP})^2$. Within the finite sample of size N, we carry out a completely randomized experiment, with N_t units randomly selected to receive the active treatment, and the remaining $N_c = N - N_t$ units assigned to the control treatment. For the units in the finite sample $W_i = 1$ for units assigned to the treatment group, and $W_i = 0$ for units assigned to the control group. To simplify the exposition, let us assign $W_i = 0$ to all units not sampled (with $R_i = 0$).
The superpopulation average treatment effect is

$$\tau_{\rm SP} = \frac{1}{N_{\rm SP}} \sum_{i=1}^{N_{\rm SP}} \left(Y_i(1) - Y_i(0) \right),$$

and the variance of the treatment effect in the superpopulation is

$$\sigma^2_{tc} = \frac{1}{N_{\rm SP}} \sum_{i=1}^{N_{\rm SP}} \left(Y_i(1) - Y_i(0) - \tau_{\rm SP} \right)^2.$$

Now consider the finite population average treatment effect:

$$\tau_{\rm FS} = \frac{1}{N} \sum_{i=1}^{N_{\rm SP}} R_i \cdot \left(Y_i(1) - Y_i(0) \right).$$

Viewing R_i as random, but keeping $(Y_i(0), Y_i(1))$, for $i = 1, \ldots, N_{\rm SP}$ fixed, we can take the expectation of $\tau_{\rm FS}$ over the distribution generated by the random sampling. Indexing the expectations operator by subscript W to be explicit about the fact that the expectation is taken over the distribution generated by the random sampling, and thus over R_i, $i = 1, \ldots, N$, we have

$$\mathbb{E}_{\rm SP}\left[\tau_{\rm FS} \middle| \mathbf{Y}_{\rm SP}(0), \mathbf{Y}_{\rm SP}(1) \right] = \frac{1}{N} \sum_{i=1}^{N_{\rm SP}} \mathbb{E}_{\rm SP}\left[R_i \right] \cdot \left(Y_i(1) - Y_i(0) \right) = \frac{1}{N} \sum_{i=1}^{N_{\rm SP}} \frac{N}{N_{\rm SP}} \cdot \left(Y_i(1) - Y_i(0) \right) = \tau_{\rm SP}.$$

The variance of the finite sample average treatment effect is

$$\begin{aligned}
\mathbb{V}_{\rm SP}\left(\tau_{\rm FS} \middle| \mathbf{Y}_{\rm SP}(0), \mathbf{Y}_{\rm SP}(1) \right) &= \mathbb{E}_{\rm SP}\left[\left(\frac{1}{N} \sum_{i=1}^{N_{\rm SP}} R_i \cdot \left(Y_i(1) - Y_i(0) \right) - \tau_{\rm SP} \right)^2 \middle| \mathbf{Y}_{\rm SP}(0), \mathbf{Y}_{\rm SP}(1) \right] \\
&= \mathbb{E}_{\rm SP}\left[\left(\frac{1}{N} \sum_{i=1}^{N_{\rm SP}} \left(R_i - \frac{N}{N_{\rm SP}} \right) \cdot \left(Y_i(1) - Y_i(0) - \tau_{\rm SP} \right) \right)^2 \middle| \mathbf{Y}_{\rm SP}(0), \mathbf{Y}_{\rm SP}(1) \right] \\
&= \frac{1}{N^2} \sum_{i=1}^{N_{\rm SP}} \sum_{i=1}^{N_{\rm SP}} \mathbb{E}_{\rm SP}\left[\left(R_i - \frac{N}{N_{\rm SP}} \right) \cdot \left(R_j - \frac{N}{N_{\rm SP}} \right) \cdot \left(Y_i(1) - Y_i(0) - \tau_{\rm SP} \right) \cdot \left(Y_j(1) - Y_j(0) - \tau_{\rm SP} \right) \middle| \mathbf{Y}_{\rm SP}(0), \mathbf{Y}_{\rm SP}(1) \right] \\
&= \frac{1 - N/N_{\rm SP}}{N \cdot N_{\rm SP}} \sum_{i=1}^{N_{\rm SP}} \left(Y_i(1) - Y_i(0) - \tau_{\rm SP} \right)^2 - \frac{1}{N_{\rm SP}^2} \sum_{i=1}^{N_{\rm SP}} \sum_{j \neq i} \left(Y_i(1) - Y_i(0) - \tau_{\rm SP} \right) \cdot \left(Y_j(1) - Y_j(0) - \tau_{\rm SP} \right)
\end{aligned}$$

$$= \frac{\sigma^2_{tc}}{N} - \frac{\sigma^2_{tc}}{N_{\rm SP}} - \frac{1}{N^2_{\rm SP}} \sum_{i=1}^{N_{\rm SP}} \sum_{j\neq i} \left(Y_i(1) - Y_i(0) - \tau_{\rm SP}\right) \cdot \left(Y_j(1) - Y_j(0) - \tau_{\rm SP}\right).$$

If $N_{\rm SP}$ is large relative to N, the last two terms are small relative to the first one, and the variance of $\tau_{\rm FS}$ over the superpopulation is approximately equal to

$$\mathbb{V}_{\rm SP}\left(\tau_{\rm FS} \middle| \mathbf{Y}_{\rm SP}(0), \mathbf{Y}_{\rm SP}(1)\right) \approx \frac{\sigma^2_{\rm SP}}{N}.$$

Now let us consider the estimator $\hat{\tau} = \overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}$. We can write this in terms of the superpopulation as

$$\hat{\tau} = \frac{1}{N_t} \sum_{i=1}^{N_{\rm SP}} R_i \cdot W_i \cdot Y_i^{\rm obs} - \frac{1}{N_c} \sum_{i=1}^{N_{\rm SP}} R_i \cdot (1-W_i) \cdot Y_i^{\rm obs}.$$

We can take the expectation of this estimator, first conditional on $\mathbf{R}$ (and always conditional on $\mathbf{Y}_{\rm SP}(1)$ and $\mathbf{Y}_{\rm SP}(0)$), so the expectation is over the randomization distribution:

$$\mathbb{E}_W\left[\hat{\tau} \middle| \mathbf{R}, \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] = \frac{1}{N_t} \sum_{i=1}^{N_{\rm SP}} R_i \cdot \mathbb{E}_W[W_i] \cdot Y_i^{\rm obs} - \frac{1}{N_c} \sum_{i=1}^{N_{\rm SP}} R_i \cdot \mathbb{E}_W[1-W_i] \cdot Y_i^{\rm obs}$$

$$= \frac{1}{N} \sum_{i=1}^{N_{\rm SP}} R_i \cdot \left(Y_i(1) - Y_i(0)\right) = \tau_{\rm FS}.$$

Thus, the unconditional expectation of $\hat{\tau}$, taken both over the randomization distribution and the sampling distribution, is

$$\mathbb{E}\left[\hat{\tau} \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] = \mathbb{E}_{\rm SP}\left[\mathbb{E}_W\left[\hat{\tau} \middle| \mathbf{R}, \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right]$$

$$= \mathbb{E}_{\rm SP}\left[\tau_{\rm FS} \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] = \tau_{\rm SP}.$$

Next we calculate the sampling variance, both over the randomization distribution and over the sampling distribution. By iterated expectations,

$$\mathbb{V}_{\rm SP} = \mathbb{V}\left(\hat{\tau} \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right)$$

$$= \mathbb{E}_{\rm SP}\left[\mathbb{V}_W\left(\hat{\tau} \middle| \mathbf{R}, \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right) \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] + \mathbb{V}_{\rm SP}\left(\mathbb{E}_W\left[\hat{\tau} \middle| \mathbf{R}, \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right)$$

$$= \mathbb{E}_{\rm SP}\left[\left.\frac{S^2_c}{N_c} + \frac{S^2_t}{N_t} - \frac{S^2_{tc}}{N}\right| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right] + \mathbb{V}_{\rm SP}\left(\tau_{\rm FS} \middle| \mathbf{Y}_{\rm SP}(1), \mathbf{Y}_{\rm SP}(0)\right)$$

$$= \frac{\sigma^2_c}{N_c} + \frac{\sigma^2_t}{N_t} - \frac{\sigma^2_{tc}}{N} + \frac{\sigma^2_{tc}}{N} - \frac{\sigma^2_{tc}}{N_{\rm SP}} - \frac{1}{N^2_{\rm SP}} \sum_{i=1}^{N_{\rm SP}} \sum_{j\neq i} \left(Y_i(1) - Y_i(0) - \tau_{\rm SP}\right) \cdot \left(Y_j(1) - Y_j(0) - \tau_{\rm SP}\right)$$

$$\approx \frac{\sigma^2_c}{N_c} + \frac{\sigma^2_t}{N_t},$$

when $N_{\rm SP}$ is large relative to N.

Appendix B: Sampling Variance Calculations

First we calculate the sampling variance of the estimator $\hat{\tau} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$. As before, we have N units, N_t receiving the treatment and N_c receiving the control. The average treatment effect is:

$$\overline{Y}(1) - \overline{Y}(0) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) = \tau.$$

The standard estimator of τ is:

$$\begin{aligned}\hat{\tau} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} &= \frac{1}{N_t}\sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_c}\sum_{i=1}^{N}(1-W_i)\cdot Y_i^{\text{obs}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{N}{N_t}\cdot W_i \cdot Y_i(1) - \frac{N}{N_c}\cdot(1-W_i)\cdot Y_i(0)\right).\end{aligned}$$

For the variance calculations is useful to work with a centered treatment indicator D_i, defined as

$$D_i = W_i - \frac{N_t}{N} = \begin{cases} \frac{N_c}{N} & \text{if } W_i = 1 \\ -\frac{N_t}{N} & \text{if } W_i = 0. \end{cases}$$

The expectation of D_i is zero, and its variance is $\mathbb{V}(D_i) = \mathbb{E}[D_i^2] = N_c N_t / N^2$. Later we also need its cross moment, $\mathbb{E}[D_i \cdot D_j]$. For $i \neq j$ the distribution of this cross product is

$$\Pr_W\left(D_i \cdot D_j = d\right) = \begin{cases} \frac{N_t\cdot(N_t-1)}{N\cdot(N-1)} & \text{if } d = N_c^2/N^2 \\ 2\cdot\frac{N_t\cdot N_c}{N\cdot(N-1)} & \text{if } d = -N_t N_c/N^2 \\ \frac{N_c\cdot(N_c-1)}{N\cdot(N-1)} & \text{if } d = N_t^2/N^2 \\ 0 & \text{otherwise,} \end{cases}$$

leading to

$$\mathbb{E}_W\left[D_i \cdot D_j\right] = \begin{cases} \frac{N_c\cdot N_t}{N^2} & \text{if } i = j \\ -\frac{N_t\cdot N_c}{N^2\cdot(N-1)} & \text{if } i \neq j \end{cases},$$

if $i \neq j$.
In terms of D_i our estimate of the average treatment effect is:

$$\begin{aligned}\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{N}{N_t}\cdot\left(D_i + \frac{N_t}{N}\right)\cdot Y_i(1) - \frac{N}{N_c}\cdot\left(\frac{N_c}{N} - D_i\right)\cdot Y_i(0)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) + \frac{1}{N}\sum_{i=1}^{N} D_i\cdot\left(\frac{N}{N_t}\cdot Y_i(1) + \frac{N}{N_c}\cdot Y_i(0)\right). \\ &= \tau + \frac{1}{N}\sum_{i=1}^{N} D_i\cdot\left(\frac{N}{N_t}\cdot Y_i(1) + \frac{N}{N_c}\cdot Y_i(0)\right). \end{aligned} \tag{B.1}$$

Because $\mathbb{E}_W[D_i] = 0$ and all potential outcome are fixed, the estimator $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ is unbiased for the average treatment effect, $\tau = \overline{Y}(1) - \overline{Y}(0)$.

Next, because the only random element in equation (B.1) is D_i, the variance of $\hat{\tau} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ is equal to the variance of the second term in equation (B.1). Using Y_i^+ as shorthand for $(N/N_t)Y_i(1) + (N/N_c)Y_i(0)$, this is equal to:

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right) = \frac{1}{N^2}\cdot\mathbb{E}_W\left[\left(\sum_{i=1}^{N} D_i\cdot Y_i^+\right)^2\right]. \tag{B.2}$$

Expanding equation (B.2) we get:

$$
\begin{aligned}
\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}}\right) &= \mathbb{E}_W\left[\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N D_iD_jY_i^+Y_j^+\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N\left(Y_i^+\right)^2\cdot\mathbb{E}_W\left[D_i^2\right]+\frac{1}{N^2}\sum_{i=1}^N\sum_{j\neq i}\mathbb{E}_W\left[D_i\cdot D_j\right]\cdot Y_i^+\cdot Y_j^+ \\
&= \frac{N_c\cdot N_t}{N^4}\sum_{i=1}^N\left(Y_i^+\right)^2-\frac{N_c\cdot N_t}{N^4\cdot(N-1)}\sum_{i=1}^N\sum_{j\neq i}Y_i^+\cdot Y_j^+ \\
&= \frac{N_c\cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^N\left(Y_i^+\right)^2-\frac{N_c\cdot N_t}{N^4\cdot(N-1)}\sum_{i=1}^N\sum_{j=1}^N Y_i^+\cdot Y_j^+ \\
&= \frac{N_c\cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^N\left(Y_i^+-\overline{Y^+}\right)^2 \\
&= \frac{N_c\cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^N\left(\frac{N}{N_t}\cdot Y_i(1)+\frac{N}{N_c}\cdot Y_i(0)-\left(\frac{N}{N_t}\cdot\overline{Y}(1)+\frac{N}{N_c}\cdot\overline{Y}(0)\right)\right)^2 \\
&= \frac{N_c\cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^N\left(\frac{N}{N_t}\cdot Y_i(1)-\frac{N}{N_t}\cdot\overline{Y}(1)\right)^2+\frac{N_c\cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^N\left(\frac{N}{N_c}\cdot Y_i(0)-\frac{N}{N_c}\cdot\overline{Y}(0)\right)^2 \\
&\quad+\frac{2\cdot N_c\cdot N_t}{N^3\cdot(N-1)}\sum_{i=1}^N\left(\frac{N}{N_t}\cdot Y_i(1)-\frac{N}{N_t}\cdot\overline{Y}(1)\right)\cdot\left(\frac{N}{N_c}\cdot Y_i(0)-\frac{N}{N_c}\cdot\overline{Y}(0)\right) \\
&= \frac{N_c}{N\cdot N_t\cdot(N-1)}\sum_{i=1}^N\left(Y_i(1)-\overline{Y}(1)\right)^2+\frac{N_t}{N\cdot N_c\cdot(N-1)}\sum_{i=1}^N\left(Y_i(0)-\overline{Y}(0)\right)^2 \\
&\quad+\frac{2}{N\cdot(N-1)}\sum_{i=1}^N\left(Y_i(1)-\overline{Y}(1)\right)\cdot\left(Y_i(0)-\overline{Y}(0)\right). \qquad \text{(B.3)}
\end{aligned}
$$

Recall the definition of S_{tc}^2, which implies that

$$
\begin{aligned}
S_{tc}^2 &= \frac{1}{N-1}\sum_{i=1}^N\left(Y_i(1)-\overline{Y}(1)-\left(Y_i(0)-\overline{Y}(0)\right)\right)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^N\left(Y_i(1)-\overline{Y}(1)\right)^2+\frac{1}{N-1}\sum_{i=1}^N\left(Y_i(0)-\overline{Y}(0)\right)^2 \\
&\quad-\frac{2}{N-1}\sum_{i=1}^N\left(Y_i(1)-\overline{Y}(1)\right)\cdot\left(Y_i(0)-\overline{Y}(0)\right) \\
&= S_t^2+S_c^2-\frac{2}{N-1}\sum_{i=1}^N\left(Y_i(1)-\overline{Y}(1)\right)\cdot\left(Y_i(0)-\overline{Y}(0)\right).
\end{aligned}
$$

Hence the expression in (B.3) is equal to

$$
\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}}\right)=\frac{N_c}{N\cdot N_t}\cdot S_t^2+\frac{N_t}{N\cdot N_c}\cdot S_c^2+\frac{1}{N}\cdot\left(S_t^2+S_c^2-S_{tc}^2\right)=\frac{S_t^2}{N_t}+\frac{S_c^2}{N_c}-\frac{S_{tc}^2}{N}.
$$

Now we investigate the bias of the Neyman estimator for the sampling variance, $\mathbb{V}_{\text{neyman}}$, under the assumption of a constant treatment effect. Assuming a constant treatment effect, S_{tc}^2 is equal to zero, so we need only find unbiased estimators for S_c^2 and S_t^2 to provide an unbiased estimator of the variance of $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$.[2] Consider the estimator

$$s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2.$$

The goal is to show that the expectation of s_t^2 is equal to

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(1) - \overline{Y}(1) \right)^2 = \frac{N}{N-1} \left(\overline{Y^2}(1) - \left(\overline{Y}(1) \right)^2 \right).$$

First,

$$\begin{aligned} s_t^2 &= \frac{1}{N_t - 1} \sum_{i=1}^{N} 1\{W_i = 1\} \cdot \left(Y_i^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2 = \frac{1}{N_t - 1} \sum_{i=1}^{N} 1\{W_i = 1\} \cdot \left(Y_i(1) - \overline{Y}_t^{\text{obs}} \right)^2 \\ &= \frac{1}{N_t - 1} \sum_{i=1}^{N} 1\{W_i = 1\} \cdot Y_i^2(1) - \frac{N_t}{N_t - 1} \left(\overline{Y}_t^{\text{obs}} \right)^2. \end{aligned} \tag{B.4}$$

Consider the expectation of the two terms in (B.4) in turn. Using again $D_i = 1\{W_i = 1\} - N_t/N$, with $\mathbb{E}[D_i] = 0$, we have

$$\begin{aligned} \mathbb{E}\left[\frac{1}{N_t - 1} \sum_{i=1}^{N} 1\{W_i = 1\} \cdot Y_i^2(1) \right] &= \frac{1}{N_t - 1} \sum_{i=1}^{N} \mathbb{E}\left[\left(D_i + \frac{N_t}{N} \right) \cdot Y_i^2(1) \right] \\ &= \frac{N_t}{N_t - 1} \cdot \overline{Y^2}(1). \end{aligned}$$

Next, the expectation of the second factor in the second term in (B.4):

$$\begin{aligned} \mathbb{E}_W \left[\left(\overline{Y}_t^{\text{obs}} \right)^2 \right] &= \mathbb{E}_W \left[\frac{1}{N_t^2} \sum_{i=1}^{N} \sum_{j=1}^{N} W_i \cdot W_j \cdot Y_i^{\text{obs}} \cdot Y_j^{\text{obs}} \right] \\ &= \mathbb{E}_W \left[\frac{1}{N_t^2} \sum_{i=1}^{N} \sum_{j=1}^{N} W_i \cdot W_j \cdot Y_i(1) \cdot Y_j(1) \right] \\ &= \frac{1}{N_t^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_W \left[\left(D_i + \frac{N_t}{N} \right) \cdot \left(D_j + \frac{N_t}{N} \right) \cdot Y_i(1) \cdot Y_j(1) \right] \\ &= \frac{1}{N_t^2} \sum_{i=1}^{N} \sum_{j=1}^{N} Y_i(1) \cdot Y_j(1) \cdot \left(\mathbb{E}\left[D_i \cdot D_j \right] + \frac{N_t^2}{N^2} \right) \\ &= \frac{1}{N_t^2} \sum_{i=1}^{N} Y_i^2(1) \cdot \left(\mathbb{E}_W\left[D_i^2 \right] + \frac{N_t^2}{N^2} \right) + \frac{1}{N_t^2} \sum_{i=1}^{N} \sum_{j \neq i} Y_i(1) \cdot Y_j(1) \cdot \left(\mathbb{E}_W\left[D_i \cdot D_j \right] + \frac{N_t^2}{N^2} \right) \end{aligned}$$

[2] As shown in Section 6.7, if we consider our sample N as a simple random sample from an infinite super-population, the variance of our estimator, $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, will equal $S_c^2/N_c + S_t^2/N_t$, so again we need only find unbiased estimators of S_c^2 and S_t^2.

$$= \frac{1}{N_t^2}\sum_{i=1}^{N} Y_i^2(1)\cdot\left(\frac{N_c\cdot N_t}{N^2}+\frac{N_t^2}{N^2}\right)+\frac{1}{N_t^2}\sum_{i=1}^{N}\sum_{j\neq i} Y_i(1)\cdot Y_j(1)\cdot\left(-\frac{N_c\cdot N_t}{N^2\cdot(N-1)}+\frac{N_t^2}{N^2}\right)$$

$$= \frac{1}{N_t}\cdot\overline{Y^2}(1)+\frac{N_t-1}{N\cdot(N-1)\cdot N_t}\sum_{i=1}^{N}\sum_{j\neq i} Y_i(1)\cdot Y_j(1)$$

$$= \frac{1}{N_t}\cdot\overline{Y^2}(1)-\frac{N_t-1}{N\cdot(N-1)\cdot N_t}\sum_{i=1}^{N} Y_i^2(1)+\frac{N_t-1}{N\cdot(N-1)\cdot N_t}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i(1)\cdot Y_j(1)$$

$$= \frac{1}{N_t}\cdot\overline{Y^2}(1)-\frac{N_t-1}{(N-1)\cdot N_t}\cdot\overline{Y^2}(1)+\frac{(N_t-1)\cdot N}{(N-1)\cdot N_t}\left(\overline{Y}(1)\right)^2$$

$$= \frac{N_c}{N_t\cdot(N-1)}\cdot\overline{Y^2}(1)+\frac{(N_t-1)\cdot N}{(N-1)\cdot N_t}\left(\overline{Y}(1)\right)^2$$

Hence the expectation of the second term in (B.4) equals

$$-\frac{N_c}{(N_t-1)\cdot(N-1)}\cdot\overline{Y^2}(1)+\frac{N}{(N-1)}\cdot\left(\overline{Y}(1)\right)^2,$$

and adding up the expectations of both terms in in (B.4) leads to

$$\mathbb{E}_W\left[s_t^2\right] = \frac{N_t}{N_t-1}\cdot\overline{Y^2}(1)-\frac{N_c}{(N_t-1)\cdot(N-1)}\cdot\overline{Y^2}(1)-\frac{N}{(N-1)}\cdot\left(\overline{Y}(1)\right)^2$$

$$= \frac{N}{N-1}\cdot\overline{Y^2}(1)-\frac{N}{(N-1)}\cdot\left(\overline{Y}(1)\right)^2 = S_t^2.$$

Following the same argument,

$$\mathbb{E}_W\left[s_c^2\right] = \frac{1}{N_c-1}\cdot\mathbb{E}_W\left[\sum_{i=1}^{N}(1-W_i)\cdot\left(Y_i^{\rm obs}-\overline{Y}_c^{\rm obs}\right)^2\right] = S_c^2.$$

Hence the estimators s_c^2 and s_t^2 are unbiased for S_c^2 and S_t^2, and can be used to create an unbiased estimator for the variance of $\overline{Y}_t^{\rm obs}-\overline{Y}_c^{\rm obs}$, our estimator of the average treatment effect.

Notes

There was a heated discussion between Fisher and Neyman regarding the importance of the null hypothesis of a zero average effect versus a zero effect for all units. In the reading of Neyman's 1923 paper in the Journal of the Royal Statistical Society on the interpretations of data from a set of agricultural experiments, the discussion became very heated:
(Neyman) "So long as the *average* (italics in original) yields of any treatments are identical, the question as to whether these treatments affect *separate* yields on *single* plots seems to be uninteresting and academic ... "
(Fisher) "... It may be foolish, but that is what the z [FEP] test was designed for, and the only purpose for which it has been used. ..."
(Neyman) "... I believe Professor Fisher himself described the problem of agricultural experimentation formerly not in the same manner as he does now. ..."
(Fisher) "... Dr. Neyman thinks another test would be more important. I am not going to argue that point. It may be that the question which Dr. Neyman thinks should be answered is more important than the one I have proposed and attempted to answer. I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose. Failing that it is surely quite unusual to claim to understand the purpose of previous work better than its author."
Much of the material in this chapter draws on Neyman (1923) translated as Neyman (). See also the comments in Rubin () on Neyman's work in this area.
The experiment from which the data used in this chapter are drawn is described in more detail in Duflo and Hanna (2006), and Duflo, Hanna, and Ryan (2007).

Table 6.1: Summary Statistics for Duflo-Hanna-Ryan Teacher-Incentive Data

	Variable	Control ($N_c = 54$) avg	(s.d.)	Treated ($N_t = 53$) avg	(s.d.)	min	max
pretreatment	`pctprewritten`	0.19	0.19	0.16	0.17	0.00	0.67
posttreatment	`open`	0.58	0.19	0.80	0.13	0.00	1.00
	`pctpostwritten`	0.47	0.19	0.52	0.23	0.05	0.92
	`written`	0.92	0.45	1.09	0.42	0.07	2.22
	`written_all`	0.46	0.32	0.60	0.39	0.04	1.43

Table 6.2: Estimates for Effect of Teacher Incentives on Proportion of Days that School is Open

Estimated Means	$\overline{Y}_c^{\text{obs}}$	0.58
	$\overline{Y}_t^{\text{obs}}$	0.80
	$\hat{\tau}$	0.22
Estimated Variance Components	s_c^2	0.19^2
	s_t^2	0.13^2
	s^2	0.16^2
	N_c	54
	N_t	53
Sampling Variance Estimates	$\hat{\mathbb{V}}_{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$	0.03^2
	$\hat{\mathbb{V}}_{\text{const}} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t}\right)$	0.03^2
	$\hat{\mathbb{V}}_{\rho_{tc}=1} = s_c^2 \cdot \frac{N_t}{N \cdot N_c} + s_t^2 \cdot \frac{N_c}{N \cdot N_t} + s_c \cdot s_t \cdot \frac{2}{N}$	0.03^2

Table 6.3: Estimates of, and Confidence Intervals for, Average Treatment Effects for Duflo-Hanna-Ryan Teacher-Incentive Data

		ate	(s.e.)	95% c.i.
pretreatment	`pctprewritten`	-0.03	(0.04)	[-0.10,0.04]
posttreatment	`open`	0.22	(0.03)	[0.15,0.28]
	`pctpostwritten`	0.05	(0.04)	[-0.03,0.13]
	`written`	0.17	(0.08)	[0.00,0.34]
	`written_all`	0.14	(0.07)	[0.00,0.28]

Table 6.4: Estimates of, and Confidence Intervals for, Average Treatment Effects for Duflo-Hanna-Ryan Teacher-Incentive Data

	`pctprewritten` $= 0$ ($N = 40$)			`pctprewritten` > 0 ($N = 67$)			Difference		
variable	$\hat{\tau}$	(s.e.)	95% c.i.	$\hat{\tau}$	(s.e.)	95% c.i.	est	(s.e.)	95% c.i.
`open`	0.23	(0.05)	[0.14,0.32]	0.21	(0.04)	[0.13,0.29]	0.02	(0.06)	[-0.10,0.14]
`pctpostwritten`	-.004	0.06	[-0.16,0.07]	0.11	(0.05)	[0.01,0.21]	-0.15	(0.08)	[-0.31,0.00]
`written`	0.20	(0.10)	[0.00,0.40]	0.18	(0.10)	[-0.03,0.38]	0.03	(0.15)	[-0.26,0.31]
`written_all`	0.04	(0.07)	[-0.10,0.19]	0.22	(0.09)	[0.04,0.40]	-0.18	(0.12)	[-0.41,0.05]