CHAPTER 6

Neyman's Repeated Sampling Approach to Completely Randomized Experiments

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 to fill in the values for all missing potential outcomes in the experiment. This was the basis for deriving the randomization distributions of various statistics, that is, the distributions induced by the random assignment of the treatments given fixed potential outcomes under that sharp null hypothesis. During the same period in which Fisher was developing this method, Neyman (1923, 1990) was focused on methods for the estimation of, and inference for, average treatment effects, also using the distribution induced by randomization, sometimes in combination with repeated sampling of the units in the experiment from a larger population of units. At a general level, he was interested in the long-run operating characteristics of statistical procedures under both repeated sampling from the population and randomized assignment of treatments to the units in the sample. Specifically, 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, his focus on average effects was different 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.

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? Most importantly, what is the difference between these averages, the average treatment effect $\tau_{\rm fs} = \overline{Y}(1) - \overline{Y}(0) = \sum_{i=1}^{N} (Y_i(1) - Y_i(0))/N$? (Here we use the subscript fs to be explicit about the fact that the estimand is the finite-sample average treatment effect. Later we use the notation $\tau_{\rm sp}$ to denote the super-population average treatment effect.) Neyman's approach was to develop an estimator of the average treatment effect and derive its mean and variance under repeated sampling. By repeated sampling we refer to the sampling generated by drawing from both the population of units, and from the randomization distribution (the assignment vector $\mathbf{W}$), although 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 $\mathbf{Y}^{\rm obs}$) 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, without such an assumption we can often still obtain good estimators of aspects of this distribution, for example, first and second moments. Neyman's primary concern was whether an estimator was unbiased for the average treatment effect τ_{fs}. 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 probability, over repeated draws, at least equal to some fixed value, the confidence coefficient.

The remainder of this chapter is organized as follows. In Section 6.2 we begin by describing the data that will be used to illustrate the concepts discussed in this chapter. These data are from a randomized experiment conducted by Duflo, Hanna, and Ryan (2012) to assess the effect of a teacher-incentive program on teacher performance. Next, in Section 6.3, 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, in Section 6.4, the sampling variance of this estimator and propose an estimator of this variance in Section 6.5. There are several approaches one can take in this latter step, depending on whether one assumes a constant additive treatment effect. In Section 6.6 we discuss the construction of confidence intervals. 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 this population was selected; the entire analysis is conditional on the population itself. In Section 6.7 we relax this assumption and instead consider, as did Neyman (1923, 1990), a population of units so that we can view the sample of N units as a random sample 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 for the average effect. In Section 6.8 we discuss the role of covariates in Neyman's approach. In the current chapter we allow only for discrete covariates. With continuous covariates the analysis is more complicated, and we discuss various methods in Chapters 7 and 8. Next, in Section 6.9, we apply Neyman's approach to the data from the Duflo-Hanna-Ryan teacher-incentive experiment. Section 6.10 concludes. Throughout the chapter we maintain the stability assumption, SUTVA.

6.2 THE DUFLO-HANNA-RYAN TEACHER-INCENTIVE EXPERIMENT DATA

To illustrate the methods discussed in this chapter, we use data from a randomized experiment conducted in rural India by Duflo, Hanna, and Ryan (2012), designed to study the effect of financial incentives on teacher performance, measured both directly by teacher absences and indirectly by educational output measures, such as average class test scores. A sample of 113 single-teacher schools was selected, and in a randomly selected subset

Table 6.1. *Summary Statistics for Duflo-Hanna-Ryan Teacher-Incentive Observed Data*

	Variable	Control ($N_c = 54$)		Treated ($N_t = 53$)			
		Average	(S.D.)	Average	(S.D.)	Min	Max
Pre-treatment	pctprewritten	0.19	(0.19)	0.16	(0.17)	0.00	0.67
Post-treatment	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

of 57 schools, the salary structure was changed so that teachers were given a salary that was tied to their (i.e., the teachers') attendance over a month-long period, whereas in the remaining 56 schools, the salary structure was unchanged. In both 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, both at the beginning and at the 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 complicating issues of unintended missing data, and we drop six schools with missing data and use the $N = 107$ 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, $N_t = 53$ were in the treatment group with a salary schedule tied to teacher attendance, and $N_c = 54$ were in the control sample. In our analyses, we use four 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 value of the writing test score (`written`), averaged over all the students in each school who took the test. Even though not all students took the test, in each class at least some students took the writing test at the end of the study. The fourth outcome variable is the average writing test score with zeros imputed for the students who did not take the test (`written_all`). We use one covariate in the analysis, the percentage of students who completed the written test prior to the study (`pctprewritten`).

Table 6.1 presents summary statistics for the data set. For all five variables (the pretreatment variables `pctprewritten`, and the four outcome variables `open`, `pctpostwritten`, `written`, and `written_all`), we present averages and standard deviations by treatment status, and the minimum and maximum values 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 considered fixed. As a result, the only random component is the vector of treatment assignments, $\mathbf{W}$, with i^{th} element W_i, which by definition has a known distribution in a completely randomized experiment.

Neyman was interested in the population average treatment effect:

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big) = \overline{Y}(1) - \overline{Y}(0),$$

where $\overline{Y}(0)$ and $\overline{Y}(1)$ are the averages of the potential control and treated outcomes respectively:

$$\overline{Y}(0) = \frac{1}{N}\sum_{i=1}^{N} Y_i(0), \quad \text{and} \quad \overline{Y}(1) = \frac{1}{N}\sum_{i=1}^{N} Y_i(1).$$

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

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

where

$$\overline{Y}_{\text{c}}^{\text{obs}} = \frac{1}{N_{\text{c}}}\sum_{i:W_i=0} Y_i^{\text{obs}} \quad \text{and} \quad \overline{Y}_{\text{t}}^{\text{obs}} = \frac{1}{N_{\text{t}}}\sum_{i:W_i=1} Y_i^{\text{obs}}.$$

Theorem 6.1 *The estimator $\hat{\tau}^{\text{dif}}$ is unbiased for τ_{fs}.*

Proof of Theorem 6.1. Using the fact that $Y_i^{\text{obs}} = Y_i(1)$ if $W_i = 1$, and $Y_i^{\text{obs}} = Y_i(0)$ if $W_i = 0$, we can write the estimator $\hat{\tau}^{\text{dif}}$ as:

$$\hat{\tau}^{\text{dif}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i(1)}{N_{\text{t}}/N} - \frac{(1 - W_i)\cdot Y_i(0)}{N_{\text{c}}/N}\right).$$

Because we view the potential outcomes as fixed, the only component in this statistic that is random is the treatment assignment, W_i. Given the setup of a completely randomized experiment (N units, with N_t randomly assigned to the treatment), by Section 3.5, $\Pr_W(W_i = 1|\mathbf{Y}(0), \mathbf{Y}(1)) = \mathbb{E}_W[W_i|\mathbf{Y}(0), \mathbf{Y}(1)] = N_{\text{t}}/N$. (Here we index the probability and expectation, and later the variance, operators by W to stress 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.) Thus,

$\hat{\tau}^{\text{dif}}$ is unbiased for the average treatment effect τ_{fs}:

$$\mathbb{E}_W\left[\hat{\tau}^{\text{dif}}\middle|\mathbf{Y}(0),\mathbf{Y}(1)\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{E}_W[W_i]\cdot Y_i(1)}{N_{\text{t}}/N} - \frac{\mathbb{E}_W[1-W_i])\cdot Y_i(0)}{N_{\text{c}}/N}\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1)-Y_i(0)\big) = \tau_{\text{fs}}.$$

□

Note that the estimator is unbiased, irrespective of the share of treated and control units in the randomized experiment. This does not imply, however, that this share is irrelevant for inference; it can greatly affect the precision of the estimator, as we 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

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 0.80 - 0.58 = 0.22,$$

as can be seen from the numbers in Table 6.1.

6.4 THE SAMPLING VARIANCE OF THE NEYMAN ESTIMATOR

Neyman was also interested in constructing interval estimates for the average treatment effect, which he later (Neyman, 1934) termed 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 variance from step 2 to create a large-sample confidence interval for the average treatment effect τ_{fs}.

In this section we focus on the first step, deriving the sampling variance of the proposed estimator $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$. 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 i is assigned to the active treatment lowers the probability that unit i' will receive active treatment. To show how to derive the sampling variance, we start with a simple example of only two units with one unit assigned to each treatment group. We then expand our discussion to the general case with N units and N_{t} randomly assigned to active treatment.

6.4.1 The Sampling Variance of the Neyman Estimator with Two Units

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

$$\tau_{\text{fs}} = \frac{1}{2}\cdot\big[(Y_1(1)-Y_1(0)) + (Y_2(1)-Y_2(0))\big]. \tag{6.1}$$

In a completely randomized experiment, both units cannot receive the same treatment; it follows that $W_1 = 1 - W_2$. The estimator for the average treatment effect is therefore:

$$\hat{\tau}^{\text{dif}} = W_1 \cdot \big(Y_1^{\text{obs}} - Y_2^{\text{obs}}\big) + (1 - W_1) \cdot \big(Y_2^{\text{obs}} - Y_1^{\text{obs}}\big).$$

If unit 1 receives the treatment ($W_1 = 1$), our estimate of the average treatment effect will be $\hat{\tau}^{\text{dif}} = Y_1^{\text{obs}} - Y_2^{\text{obs}} = Y_1(1) - Y_2(0)$. If on the other hand, $W_1 = 0$, the estimate will be $\hat{\tau} = Y_2^{\text{obs}} - Y_1^{\text{obs}} = Y_2(1) - Y_1(0)$, so that we can also write:

$$\hat{\tau}^{\text{dif}} = W_1 \cdot \big(Y_1(1) - Y_2(0)\big) + (1 - W_1) \cdot \big(Y_2(1) - Y_1(0)\big).$$

To simplify the following calculations of the sampling variance of this estimator, define the binary variable $D = 2 \cdot W_1 - 1$, so that $D \in \{-1, 1\}$, $W_1 = (1 + D)/2$ and $W_2 = 1 - W_1 = (1 - D)/2$. 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 $\mathbb{V}_W(D) = \mathbb{E}_W[D^2] = D^2 = 1$. In terms of D and the potential outcomes, we can write the estimator $\hat{\tau}$ as:

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

which can be rewritten as:

$$\begin{aligned}\hat{\tau}^{\text{dif}} &= \frac{1}{2} \cdot \big[\big(Y_1(1) - Y_1(0)\big) + \big(Y_2(1) - Y_2(0)\big)\big] \\ &\quad + \frac{D}{2} \cdot \big[\big(Y_1(1) + Y_1(0)\big) - \big(Y_2(1) + Y_2(0)\big)\big] \\ &= \tau_{\text{fs}} + \frac{D}{2} \cdot \big[\big(Y_1(1) + Y_1(0)\big) - \big(Y_2(1) + Y_2(0)\big)\big].\end{aligned}$$

Because $\mathbb{E}_W[D] = 0$, we can see immediately that $\hat{\tau}^{\text{dif}}$ is unbiased for τ_{fs} (which we already established in Section 6.3 for the general case). However, the representation in terms of D also makes the calculation of its sampling variance straightforward:

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

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

$$\mathbb{V}_W(\hat{\tau}^{\text{dif}}) = \frac{1}{4} \cdot \big[\big(Y_1(1) + Y_1(0)\big) - \big(Y_2(1) + Y_2(0)\big)\big]^2. \tag{6.2}$$

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 the different potential outcomes for the same unit that are never jointly observed.

6.4.2 The Sampling Variance of the Neyman Estimator with N Units

Next, we look at the general case with $N > 2$ units, of which N_t are randomly assigned to treatment. To calculate the sampling variance of $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, we need the expectations of the second and cross moments of the treatment indicators W_i for $i = 1, \ldots, N$. Because $W_i \in \{0, 1\}$, $W_i^2 = W_i$, and thus

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

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 i' being treated – are not independent. Therefore $\mathbb{E}_W\left[W_i \cdot W_{i'}\right] \neq \mathbb{E}_W\left[W_i\right] \cdot \mathbb{E}_W\left[W_{i'}\right] = (N_t/N)^2$. Rather:

$$\mathbb{E}_W[W_i \cdot W_{i'}] = \Pr_W(W_i = 1) \cdot \Pr_W(W_{i'} = 1 | W_i = 1) = \frac{N_{\text{t}}}{N} \cdot \frac{N_{\text{t}} - 1}{N - 1}, \quad \text{for } i \neq j,$$

because conditional on $W_i = 1$ there are $N_{\text{t}} - 1$ treated units remaining, out of a total of $N - 1$ units remaining. Given the sampling moments derived, we can infer the sampling variance and covariance of W_i and $W_{i'}$.

Theorem 6.2 *The sampling variance of* $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$ *is*

$$\mathbb{V}_W\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right) = \frac{S_c^2}{N_{\text{c}}} + \frac{S_t^2}{N_{\text{t}}} - \frac{S_{tc}^2}{N}, \tag{6.3}$$

where S_c^2 *and* S_t^2 *are the variances of* $Y_i(0)$ *and* $Y_i(1)$ *in the sample, defined as:*

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

and S_{tc}^2 *is the sample variance of the unit-level treatment effects, defined as:*

$$\begin{aligned} S_{tc}^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0))\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \tau_{\text{fs}}\right)^2. \end{aligned}$$

Proof of Theorem 6.2. See Appendix A.

Let us consider the interpretation of the three components of this variance in turn. The first two are related to sample variances for averages of random samples. Recall that the finite-sample average treatment effect is the difference in average potential outcomes: $\tau_{\text{fs}} = \overline{Y}(1) - \overline{Y}(0)$. To estimate τ_{fs}, we first estimate $\overline{Y}(1)$, the population average potential outcome under treatment, by the average outcome for the N_t treated units, $\overline{Y}_{\text{t}}^{\text{obs}}$. 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 of a random sample of size N_{t} would be $(S_t^2/N_{\text{t}}) \cdot (1 - N_{\text{t}}/N)$,

where the last factor is the finite sample correction. The first term has this form, except for the finite sample correction. Similarly, the average outcome for the $N_{\rm c}$ units assigned to control, $\overline{Y}_{\rm c}^{\rm obs}$, is unbiased for the population average outcome under the control treatment, $\overline{Y}(0)$, and its sampling variance, ignoring the finite population correction, 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 $N_{\rm t}$ treated units provide a simple random sample of the N values of $Y_i(1)$, and the $N_{\rm c}$ control units provide a simple random sample of the N values of $Y_i(0)$.

The third component of this sampling variance, S_{tc}^2/N, is the sample 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 $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$, Equation (6.3), 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 $\hat{\tau}^{\rm dif}$ 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

$$\rho_{tc} = \frac{1}{(N-1)\cdot S_c \cdot S_t} \sum_{i=1}^{N} \Big(Y_i(1) - \overline{Y}(1)\Big) \cdot \Big(Y_i(0)) - \overline{Y}(0)\Big). \tag{6.4}$$

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 (6.3), the alternative expression for the sampling variance of $\hat{\tau}^{\rm dif}$ (alternative to (6.3)) is:

$$\mathbb{V}_W\left(\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\right) = \frac{N_{\rm t}}{N\cdot N_{\rm c}} \cdot S_c^2 + \frac{N_{\rm c}}{N\cdot N_{\rm t}} \cdot S_t^2 + \frac{2}{N} \cdot \rho_{tc} \cdot S_c \cdot S_t. \tag{6.5}$$

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

$$\mathbb{V}_W\left(\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\,\middle|\, \rho_{tc} = -1\right) = \frac{N_{\rm t}}{N\cdot N_{\rm c}} \cdot S_c^2 + \frac{N_{\rm c}}{N\cdot N_{\rm t}} \cdot S_t^2 - \frac{2}{N} \cdot S_c \cdot S_t,$$

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

$$\mathbb{V}_W\left(\overline{Y}_{\rm t}^{\rm obs}-\overline{Y}_{\rm c}^{\rm obs}\,\middle|\,\rho_{tc}=1\right)=\frac{N_{\rm t}}{N\cdot N_{\rm c}}\cdot S_c^2+\frac{N_{\rm c}}{N\cdot N_{\rm t}}\cdot S_t^2+\frac{2}{N}\cdot S_c\cdot S_t$$
$$=\frac{S_c^2}{N_{\rm c}}+\frac{S_t^2}{N_{\rm t}}-\frac{(S_c-S_t)^2}{N}. \tag{6.6}$$

The most notable special case of perfect correlation arises when the treatment effect is constant and additive, $Y_i(1)-Y_i(0)=\tau$ for all $i=1,\ldots,N$. In that case,

$$\mathbb{V}^{\rm const}=\mathbb{V}_W\left(\overline{Y}_{\rm t}^{\rm obs}-\overline{Y}_{\rm c}^{\rm obs}\,\middle|\,\rho_{tc}=1,S_c^2=S_t^2\right)=\frac{S_c^2}{N_{\rm c}}+\frac{S_t^2}{N_{\rm t}}. \tag{6.7}$$

The fact that the sampling variance of $\overline{Y}_{\rm t}^{\rm obs}-\overline{Y}_{\rm c}^{\rm obs}$ is largest when the treatment effect is constant (i.e., not varying) across units may appear somewhat counterintuitive. Let us therefore return to the two-unit case and consider the form of the sampling variance there in more detail. In the two-unit case, the sampling variance, presented in Equation (6.2), is a function of the sum of the two potential outcomes for each of the two units. Consider two numerical examples. In the first example, $Y_i(0)=Y_i(1)=10$, and $Y_2(0)=Y_2(1)=-10$, corresponding to a zero treatment effect for both units. To calculate the correlation between the two potential outcomes, we use expression (6.4) for ρ_{tc} and find the numerator of ρ_{tc} 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 $Y_1(1)=Y_2(0)=10$. 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 case 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 20^2), 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, $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, the next step is to develop an estimator for this sampling variance. To do this, we consider separately each of the three elements of the sampling variance given in Equation (6.3).

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

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

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

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

The third term, S_{tc}^2 (the population variance of the unit-level treatment effects), is generally impossible to estimate empirically 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 ($Y_i(1) - Y_i(0) = \tau_{\text{fs}}$ for all units), then this component of the sampling variance is equal to zero and the third term vanishes. Thus we have proved:

Theorem 6.3 *If the treatment effect $Y_i(1) - Y_i(0)$ is constant, then an unbiased estimator for the sampling variance is*

$$\hat{\mathbb{V}}^{\text{neyman}} = \frac{s_c^2}{N_{\text{c}}} + \frac{s_t^2}{N_{\text{t}}}. \tag{6.8}$$

This estimator for the sampling variance is widely used, even when the assumption of an additive treatment effect may be known to 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 $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, irrespective of the heterogeneity in the treatment effect, because the third term is non-negative. 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 rely on the central limit theorem to construct normal-distribution-based confidence intervals.) 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 unbiased only 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 generally overestimates the true sampling variance of $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$. As a result, Neyman's interval is generally statistically conservative in large samples. The second reason for using $\hat{\mathbb{V}}^{\rm neyman}$ as an estimator for the sampling variance of $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$ is that it is always unbiased for the sampling variance of $\hat{\tau}^{\rm dif}$ as an estimator of the infinite super-population average treatment effect; we discuss this population interpretation at greater length in Section 6.7.

In the remainder of this section, we consider two alternative estimators for the sampling variance of $\hat{\tau}^{\rm dif}$. The first explicitly allows for treatment effect heterogeneity. Under treatment effect heterogeneity, the estimator for the sampling variance in Equation (6.8), $\hat{\mathbb{V}}^{\rm neyman}$, 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 under coverage in large samples.

To see that there is indeed information to do so, recall our argument 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 effects. To use this information to create a better estimator for the sampling variance of $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$, let us turn to the representation of the sampling variance in Equation (6.5), which incorporates ρ_{tc}, the population correlation coefficient between the potential outcomes:

$$\mathbb{V}_W\left(\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\right) = S_c^2 \cdot \frac{N_{\rm t}}{N \cdot N_{\rm c}} + S_c^2 \cdot \frac{N_{\rm c}}{N \cdot N_{\rm t}} + \rho_{tc} \cdot S_c \cdot S_t \cdot \frac{2}{N}.$$

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

$$\hat{\mathbb{V}}^{\rho_{tc}} = s_c^2 \cdot \frac{N_{\rm t}}{N \cdot N_{\rm c}} + s_t^2 \cdot \frac{N_{\rm c}}{N \cdot N_{\rm t}} + \rho_{tc} \cdot s_c \cdot s_t \cdot \frac{2}{N}. \tag{6.9}$$

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

[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 large samples. Fisher (1934) in his discussion did not agree with the propriety of this definition.

this bound is

$$\begin{aligned}\hat{\mathbb{V}}^{\rho_{tc}=1} &= s_c^2 \cdot \frac{N_{\rm t}}{N \cdot N_{\rm c}} + s_1^2 \cdot \frac{N_{\rm c}}{N \cdot N_{\rm t}} + s_c \cdot s_t \cdot \frac{2}{N} \\ &= \frac{s_c^2}{N_{\rm c}} + \frac{s_t^2}{N_{\rm t}} - \frac{(s_t - s_c)^2}{N}. \end{aligned} \tag{6.10}$$

If s_c^2 and s_t^2 are unequal, then $\hat{\mathbb{V}}^{\rho_{tc}=1}$ will be smaller than $\hat{\mathbb{V}}^{\rm neyman}$. Using $\hat{\mathbb{V}}^{\rho_{tc}=1}$ to construct confidence intervals will result in tighter confidence intervals than using $\hat{\mathbb{V}}^{\rm neyman}$, without compromising their large-sample validity. The intervals based on $\hat{\mathbb{V}}^{\rho_{tc}=1}$ will still be conservative in large samples, because $\hat{\mathbb{V}}^{\rho_{tc}=1}$ is still upwardly biased when the true correlation is smaller than one, although less so than $\hat{\mathbb{V}}^{\rm neyman}$. 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 an alternative sampling variance estimator under the additional assumption that the treatment effect is constant, $Y_i(1) - Y_i(0) = \tau$ for all i. This alternative estimator exploits the fact that under the constant treatment assumption, the population variances of the two potential outcomes, S_c^2 and S_t^2, must be equal. We can therefore define $S^2 \equiv S_c^2 = S_t^2$ and pool the outcomes for the the treated and control units to estimate this common variance:

$$\begin{aligned} s^2 &= \frac{1}{N-2} \cdot \left(s_c^2 \cdot (N_{\rm c} - 1) + s_t^2 \cdot (N_{\rm t} - 1) \right) \\ &= \frac{1}{N-2} \cdot \left(\sum_{i:W_i=0} \left(Y_i^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs} \right)^2 + \sum_{i:W_i=1} \left(Y_i^{\rm obs} - \overline{Y}_{\rm t}^{\rm obs} \right)^2 \right). \end{aligned} \tag{6.11}$$

The larger sample size for this estimator (from $N_{\rm c}$ and $N_{\rm t}$ 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 $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$ if the treatment effect is constant, namely

$$\hat{\mathbb{V}}^{\rm const} = s^2 \cdot \left(\frac{1}{N_{\rm c}} + \frac{1}{N_{\rm t}} \right). \tag{6.12}$$

When the treatment effects are constant this estimator is preferable to either $\hat{\mathbb{V}}^{\rm neyman}$ or $\hat{\mathbb{V}}^{\rho_{tc}=1}$, but if not, it need not be valid. Both $\hat{\mathbb{V}}^{\rm neyman}$ 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 probability 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

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

The two sample variances 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

$$\hat{\mathbb{V}}^{\text{neyman}} = \frac{s_c^2}{N_{\text{c}}} + \frac{s_t^2}{N_{\text{t}}} = 0.0311^2.$$

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

$$\hat{\mathbb{V}}^{\rho_{tc}=1} = s_c^2 \cdot \frac{N_{\text{t}}}{N \cdot N_{\text{c}}} + s_t^2 \cdot \frac{N_{\text{c}}}{N \cdot N_{\text{t}}} + s_c \cdot s_t \cdot \frac{2}{N} = 0.0305^2.$$

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 of 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 example. The third and final estimate of the sampling variance, which relies on a constant treatment effect for its validity, is

$$\hat{\mathbb{V}}^{\text{const}} = s^2 \cdot \left(\frac{1}{N_{\text{c}}} + \frac{1}{N_{\text{t}}} \right) = 0.0312^2,$$

slightly larger than the other estimates, but essentially the same for practical purposes.

6.6 CONFIDENCE INTERVALS AND TESTING

In the introduction to this chapter, 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 $1 - \alpha$, here we mean a pair of functions $C_L(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ and $C_U(\mathbf{Y}^{\text{obs}}, \mathbf{W})$, defining 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 $\mathbf{W}$. The distribution of the confidence limits 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 considered 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 $\hat{\tau}^{\text{dif}}$ over its randomization distribution (in practice we recommend using $\hat{\mathbb{V}}^{\text{neyman}}$). Suppose we wish to construct a 90% confidence interval. We base the interval on a normal approximation to the randomization distribution of $\hat{\tau}^{\text{dif}}$. This approximation is somewhat intellectually inconsistent with our stress on finite-sample properties of the estimator for τ and its sampling variance, but it is driven by the common lack of empirical *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 $\hat{\tau}^{\text{dif}}/\sqrt{\hat{\mathbb{V}}^{\text{neyman}}}$ 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 a nominal central 90% confidence interval as:

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

More generally, if we wish to construct a central confidence interval with nominal 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 $z_{\alpha/2}$, and construct the confidence interval:

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

This approximation applies when using any estimate of the sampling variance, and, in large samples, the resulting intervals are valid confidence intervals 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 three following 90% confidence intervals. First, based on $\hat{\mathbb{V}}^{\text{neyman}} = 0.0311^2$, we find

$$\begin{aligned}\text{CI}^{0.90}_{\text{neyman}}(\tau_{\text{fs}}) &= \left(\hat{\tau}^{\text{dif}} + z_{0.10/2} \cdot \sqrt{\hat{\mathbb{V}}^{\text{neyman}}}, \hat{\tau}^{\text{dif}} + z_{1-0.10/2} \cdot \sqrt{\hat{\mathbb{V}}^{\text{neyman}}}\right) \\ &= (0.2154 - 1.645 \cdot 0.0311, 0.2154 + 1.645 \cdot 0.0311) = (0.1642, 0.2667).\end{aligned}$$

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

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

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

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

With the estimates for the sampling variances so similar, the three 90% large-sample 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 hypothesis that the average effect differs from zero:

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

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

A natural test statistic to use for Neyman's "average null" is the ratio of the point estimate to the estimated standard error. For the teacher-incentive data, the point estimate is $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 0.2154$. 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, based on the normal approximation to the distribution of the t-statistic, 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 approaches. First, and most important, they assess different null hypotheses, for example, a zero average effect for Neyman versus a zero effect for all units for Fisher (although Fisher's null hypothesis implies Neyman's). Second, the Neyman test relies on a large-sample normal approximation for its validity, whereas the p-values based on the FEP approach are 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

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

and the (implicit) alternative hypothesis is

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

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 heterogeneous 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 equal to the average difference in outcomes by treatment status, has no power against this alternative. However, the FEP approach, with a different statistic, based on the average difference in outcomes after 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 samples.

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 *actual* variance ($\mathbb{V}_W(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}})$). Second, we use a normal *approximation* for the repeated sampling distribution of the difference in averages $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$. Both approximations are justified in large samples. If the sample is reasonably large, and if there are few or no outliers, as in the application in this chapter, these approximations will likely be accurate.

6.7 INFERENCE FOR 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 perspective, in which the N observed units are viewed as a random sample from an essentially infinite population. The second argument in favor of using the sampling variance estimator $\hat{\mathbb{V}}^{\text{neyman}}$ in Equation (6.8) 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 $\hat{\tau}^{\text{dif}}$ for the super-population, as opposed to the finite sample, 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, which, for simplicity, we assume is infinite. This is a slight departure from Neyman's explicit focus on the average treatment effect for a finite population. In many cases, however, 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 in the year of the experiment, in most social and medical science settings, experiments are, explicitly or implicitly, conducted with a view to inform policies for a larger population of units, often assumed to have generated the N units in our sample by random sampling. However, without additional information, we cannot hope to obtain more precise estimates for the treatment effects in the super-population than for the treatment effects in the sample. In fact, the estimates for the population estimands are typically strictly less precise. Ironically 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 random sample of the target super-population, rather than viewing them as the population itself, induces a distribution on the two potential outcomes for each unit. The pair of potential outcome values for an observed unit i is simply one draw from the distribution in the 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 effects and on the average of the unit-level treatment effects within the drawn sample. To be clear about this super-population perspective, we use the subscript fs to denote the finite-sample average treatment effect and sp to denote the super-population average treatment effect:

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

Analogously, the subscript sp on the expectations operator indicates that the expectation is taken over the distribution generated by random sampling from the super-population 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 super-population or, equivalently, the average treatment effect in the super-population. Because of the random sampling, τ_{sp} is also equal to the expected value of the finite-sample average treatment effect,

$$\mathbb{E}_{\text{sp}}\left[\tau_{\text{fs}}\right] = \mathbb{E}_{\text{sp}}\left[\overline{Y}(1) - \overline{Y}(0)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_i(0)\right] = \tau_{\text{sp}}. \tag{6.13}$$

See Appendix B for details on the super-population perspective. Let σ^2_{tc} be the variance of the unit-level treatment effect in 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:

$$\sigma^2_c = \mathbb{V}_{\text{sp}}(Y_i(0)) = \mathbb{E}_{\text{sp}}\left[(Y_i(0) - \mathbb{E}_{\text{sp}}[Y_i(0))^2\right],$$

and

$$\sigma^2_t = \mathbb{V}_{\text{sp}}(Y_i(1)) = \mathbb{E}_{\text{sp}}\left[(Y_i(1) - \mathbb{E}_{\text{sp}}[Y_i(1))^2\right].$$

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

$$\mathbb{V}_{\text{sp}}(\tau_{\text{fs}}) = \mathbb{V}_{\text{sp}}\left(\overline{Y}(1) - \overline{Y}(0)\right) = \sigma^2_{tc}/N. \tag{6.14}$$

Now let us consider the sampling variance of the standard estimator for the average treatment effect, $\hat{\tau}^{\text{dif}} = \overline{Y}^{\text{obs}}_{\text{t}} - \overline{Y}^{\text{obs}}_{\text{c}}$, given this sampling from the super-population. The expectation and variance operators without subscripts denote expectations and variances taken over both the randomization distribution and the random sampling from the super-population.

We have

$$\begin{aligned}\mathbb{V}\left(\hat{\tau}^{\text{dif}}\right) &= \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \mathbb{E}\left[\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \mathbb{E}_{\text{sp}}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)^2\right],\end{aligned}$$

where the second equality holds because $\mathbb{E}\left[\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right] = \mathbb{E}_{\text{sp}}[\overline{Y}(1) - \overline{Y}(0)] = \tau_{\text{sp}}$, 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:

$$\begin{aligned}&\mathbb{V}\left(\hat{\tau}^{\text{dif}}\right) \\ &= \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \left(\overline{Y}(1) - \overline{Y}(0)\right) + \left(\overline{Y}(1) - \overline{Y}(0)\right) - \mathbb{E}_{\text{sp}}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \left(\overline{Y}(1) - \overline{Y}(0)\right)\right)^2\right] \\ &\quad + \mathbb{E}_{\text{sp}}\left[\left(\left(\overline{Y}(1) - \overline{Y}(0)\right) - \mathbb{E}_{\text{sp}}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)^2\right] \\ &\quad + 2 \cdot \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \left(\overline{Y}(1) - \overline{Y}(0)\right)\right) \cdot \left(\left(\overline{Y}(1) - \overline{Y}(0)\right) - \mathbb{E}_{\text{sp}}\left[\overline{Y}(1) - \overline{Y}(0)\right]\right)\right].\end{aligned}$$

The third term of this last expression, the covariance term, is equal to zero because the expectation of the first factor, $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - (\overline{Y}(1) - \overline{Y}(0))$, conditional on the N-vectors $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ (taking the expectation just over the randomization distribution), is zero. Hence the sampling variance reduces to:

$$\begin{aligned}\mathbb{V}\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right) &= \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2\right] \\ &\quad + \mathbb{E}_{\text{sp}}\left[\left(\overline{Y}(1) - \overline{Y}(0) - \mathbb{E}_{\text{sp}}\left[Y(1) - Y(0)\right]\right)^2\right]. \qquad (6.15)\end{aligned}$$

Earlier we showed that $\mathbb{E}_W\left[\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \middle| \mathbf{Y}(0), \mathbf{Y}(1)\right] = \tau_{\text{fs}} = \overline{Y}(1) - \overline{Y}(0)$; hence by iterated expectations, the first term on the right side is equal to the expectation of the conditional (randomization-based) variance of $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$ (conditional on the N-vector of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$). This conditional variance is equal to

$$\mathbb{E}_W\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2 \middle| \mathbf{Y}(0), \mathbf{Y}(1)\right] = \frac{S_c^2}{N_{\text{c}}} + \frac{S_t^2}{N_{\text{t}}} - \frac{S_{tc}^2}{N}, \qquad (6.16)$$

as in Equation (6.3). Recall that these earlier calculations were made when assuming that the sample N was the population of interest and thus were conditional on $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$. The expectation of (6.16) over the distribution of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ generated by sampling

from the super-population is

$$\begin{aligned}
&\mathbb{E}\left[\left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2\right] \\
&\quad = \mathbb{E}_{\mathrm{sp}}\left[\mathbb{E}_W\left[\left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} - \overline{Y}(1) - \overline{Y}(0)\right)^2 \,\middle|\, \mathbf{Y}(0), \mathbf{Y}(1)\right]\right] \\
&\quad = \mathbb{E}_{\mathrm{sp}}\left[\frac{S_c^2}{N_{\mathrm{c}}} + \frac{S_t^2}{N_{\mathrm{t}}} - \frac{S_{tc}^2}{N}\right] = \frac{\sigma_c^2}{N_{\mathrm{c}}} + \frac{\sigma_t^2}{N_{\mathrm{t}}} - \frac{\sigma_{tc}^2}{N}.
\end{aligned}$$

The expectation of the second term on the right side of Equation (6.15) is equal to σ_{tc}^2/N, as we saw in Equation (6.14). Thus the sampling variance of $\hat{\tau}^{\mathrm{dif}}$ over sampling from the super-population equals:

$$\mathbb{V}_{\mathrm{sp}} = \mathbb{V}_{\mathrm{sp}}\left(\hat{\tau}^{\mathrm{dif}}\right) = \frac{\sigma_c^2}{N_{\mathrm{c}}} + \frac{\sigma_t^2}{N_{\mathrm{t}}}, \tag{6.17}$$

which we can estimate without bias by substituting s_c^2 and s_t^2 for σ_c^2 and σ_t^2, respectively:

$$\hat{\mathbb{V}}^{\mathrm{sp}} = \frac{s_c^2}{N_{\mathrm{c}}} + \frac{s_t^2}{N_{\mathrm{t}}}.$$

The estimator $\hat{\mathbb{V}}^{\mathrm{sp}}$ is identical to the previously introduced conservative estimator of the sampling variance for the finite population average treatment effect estimator, $\hat{\mathbb{V}}^{\mathrm{neyman}}$, presented in Equation 6.8. Under simple random sampling from the super-population, the expected value of the estimator $\hat{\mathbb{V}}^{\mathrm{neyman}}$ equals $\mathbb{V}_{\mathrm{sp}}$. Hence, considering the N observed units as a simple random sample from an infinite super-population, the estimator in (6.8) is an unbiased estimate of the sampling variance of the estimator of the super-population average treatment effect. Neither of the alternative estimators – $\hat{\mathbb{V}}^{\mathrm{const}}$ in Equation (6.12), which exploits the assumption of a constant treatment effect, nor $\hat{\mathbb{V}}^{\rho_{tc}=1}$ in Equation (6.10), derived through bounds on the correlation coefficient – has this attractive quality. Thus, despite the fact that $\hat{\mathbb{V}}^{\mathrm{const}}$ may be a better estimator of the sampling variance in the finite population when the treatment effect is constant, and $\hat{\mathbb{V}}^{\rho_{tc}=1}$ may be a better estimator of $\mathbb{V}_{\mathrm{fs}}$, $\hat{\mathbb{V}}^{\mathrm{neyman}}$ is used almost uniformly in practice in our experience, although the logic for it appears to be rarely explicitly discussed.

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 sample into subsamples defined by the values of the covariate and then conduct the analysis separately within these subsamples. The resulting within-subsample estimators would be unbiased for the within-subsample average treatment effect. Taking an average of these estimates, weighted by subsample sizes, gives an unbiased estimate of the overall average treatment effect. As we see in Chapter 9, we consider this method in the discussion on stratified random experiments.

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

6.9 RESULTS FOR THE DUFLO-HANNA-RYAN 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 four outcomes in turn, plus one "pseudo-outcome." For illustrative purposes, we report here a number of point, sampling variance, and interval estimates. The first variable we analyze, as if it were 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, -0.03, with a 95% confidence interval that comfortably includes zero, $(-0.10, 0.04)$.

Now we turn to the four "real" outcomes. In Table 6.2 we report estimates of the components of the variance, and in Table 6.3 we present estimates of and confidence intervals for the average treatment effects. First we focus on the causal effect of the attendance-related salary incentives on the proportion of days that the school was open during the 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 appeared to 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 statistically significant at the 5% level, but it is at the 10% 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 $[0.00, 0.28]$. As with the Fisher exact p-value approach, the interpretation of nominal levels for tests and interval estimates formally holds for only one such interval. In the final analysis, we look at estimates separately for two subsamples, 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. Again, these analyses are for illustrative purposes only, and we do not take account of the fact that we do multiple tests. The first subpopulation (`pctprewritten`$=0$) comprises 40 schools (37%) and the second (`pctprewritten`>0) 67 schools (63%). We analyze

Table 6.2. ***Estimates of Components of Variance of Estimator for the Effect of Teacher Incentives on the Proportion of Days that the School is Open;*** $N_c = 54$, $N_t = 53$, ***Duflo-Hanna-Ryan Data***

Estimated means	$\overline{Y}_c^{obs}$	0.58
	$\overline{Y}_t^{obs}$	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
Sampling variance estimates	$\hat{\mathbb{V}}^{neyman} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$	0.03^2
	$\hat{\mathbb{V}}^{const} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t}\right)$	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***

$\widehat{ATE}$	$\widehat{(\text{s.e.})}$	95% C.I.
0.22	(0.03)	(0.15,0.28)
0.05	(0.04)	(−0.03,0.13)
0.17	(0.08)	(0.00,0.34)
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***

Variable	`pctpre` = 0 ($N = 40$)			`pctprewritten` > 0 ($N = 67$)			Difference		
	$\hat{\tau}$	$\widehat{(\text{s.e.})}$	95% C.I.	$\hat{\tau}$	$\widehat{(\text{s.e.})}$	95% C.I.	EST	$\widehat{(\text{s.e.})}$	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)
`pctpost written`	−0.004	(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)

separately the effect of assignment to attendance-based teacher incentives on all four outcomes. The descriptive results are reported in Table 6.4. The main substantive finding is that the effect of the incentive scheme on writing skills (`written`) appears lower for schools where many students entered with insufficient writing skills to take the initial test. The 95% confidence interval comfortably includes zero (−0.41, 0.05), and 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 and random assignment of treatments. 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 address more complicated questions, albeit under additional assumptions or approximations.

NOTES

There was disagreement between Fisher and Neyman regarding the importance of the null hypothesis of a zero average effect versus zero effects for all units. In the reading of Neyman's 1935 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* (emphasis 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."

Given the tone of Fisher's remarks, it is all the more suprising how gracious Neyman is in later discussions, for example, the quotations in Chapter 5.

Much of the material in this chapter draws on Neyman (1923), translated as Neyman (1990). Also see Neyman (1934, 1935), with discussions, as well as the comments in Rubin (1990b) on Neyman's work in this area.