Chapter 5: Neyman's Repeated Sampling Perspective in Completely Randomized Experiments

Recall from the previous chapter that Fisher focused on testing the sharp null hypothesis that the treatment had no effect for any unit. Neyman (1923, 1935) was interested in a different question—he was interested in estimating the average effect of the treatment. The hypothesis that the average effect of the treatment is zero is clearly much weaker than the sharp Fisher hypothesis that the effect is zero for every single unit. The average effect may be zero even if for some units the treatment has a positive effect, as long as there are other units with negative, offsetting treatment effects.

Neyman's basic questions were the following: What would the average outcome be if all units were exposed to the treatment? How does that compare to the average outcome if all units are exposed to the control treatment? His approach was to consider an estimator of the difference of these two averages and derive its distribution under repeated sampling by drawing from the randomization distribution. Of concern to him was whether the estimator was unbiased for the average treatment effect. A second goal was to create confidence intervals for the causal estimand, and in order to achieve that goal he set about finding unbiased estimators for the variance of the estimator as well.

5.1 Unbiased Estimation of the Average Treatment Effect

Suppose we have a population consisting of N units. Let $\bar{Y}(1) - \bar{Y}(0)$ denote the population average treatment effect:

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

which is the estimand of interest. Suppose that a completely randomized experiment was conducted on these N units with M units assigned to treatment and $N - M$ assigned to control. The implication is that the indicator for assignment to treatment, W_i, has expectation equal to M/N. Then consider the statistic

$$T_i = \left(\frac{W_i \cdot Y_i^{\text{obs}}}{M/N} - \frac{(1-W_i) \cdot Y_i^{\text{obs}}}{(N-M)/N} \right) = \begin{cases} \frac{M}{N} Y_i^{\text{obs}} & \text{if } W_i = 1, \\ -\frac{N-M}{N} Y_i^{\text{obs}} & \text{if } W_i = 0, \end{cases}$$

Using the equality $Y_i^{\text{obs}} = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$ to substitute for the observed value Y_i^{obs} this can be written in terms of the potential outcomes as

$$T_i = \left(\frac{W_i \cdot Y_i(1)}{M/N} - \frac{(1-W_i) \cdot Y_i(0)}{(N-M)/N} \right).$$

In a completely randomized experiment with M treated and $N - M$ control units the expectation of W_i is equal to M/N. Hence the expectation of T_i is equal to the unit-level causal effect $Y_i(1) - Y_i(0)$:

$$E\Big[T_i\Big] = \left(\frac{E[W_i] \cdot Y_i(1)}{M/N} - \frac{(1 - E[W_i]) \cdot Y_i(0)}{(N-M)/N} \right) = Y_i(1) - Y_i(0).$$

Now consider the average of the unit-level statistic T_i over all N units:

$$\frac{1}{2N} \sum_{i=1}^{2N} T_i = \bar{y}_1 - \bar{y}_0,$$

where, as before, for $w = 0, 1$, $\bar{y}_w$ is the average outcome for units assigned treatment $W_i = w$:

$$\bar{y}_w = \sum_{i=1}^{N} Y_i^{\text{obs}} \cdot 1\{W_i = w\} \Big/ \sum_{i=1}^{N} 1\{W_i = w\}.$$

Since each T_i is unbiased for the unit-level causal effect, the expected value of $\bar{y}_1 - \bar{y}_0$ is the population average treatment effect:

$$E\Big[\bar{y}_1 - \bar{y}_0\Big] = \frac{1}{N} \sum_{i=1}^{N} E\Big[T_i\Big]$$

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

Just as in the discussion of the properties of the Fisher test this approach is non-parametric in the sense that we do not rely on assumptions about the distribution of the potential outcomes. The expectation of $\bar{y}_1 - \bar{y}_0$ is calculated purely on the basis of the randomization distribution and involves only the expectation of the assignment indicator W_i—treating the potential outcomes as fixed. Noe that, in terms of bias, the share of treated and control units in the randomized experiment is immaterial. This does not imply that this share is irrelevant, however. It can greatly affect the precision of the estimator, to which we turn next.

5.2 The Variance of Difference in Treatment and Control Averages

In addition to being interested in showing that the average treatment-control difference $\bar{y}_1 - \bar{y}_0$ is unbiased for the population average treatment effect $\bar{Y}(1) - \bar{Y}(0)$, Neyman was also interested in the precision of this estimator and how to estimate it.

First let us consider the simple case with one treated and one control unit. The average treatment effect is

$$\tau = \frac{1}{2}\left(Y_1(1) - Y_1(0) + Y_2(1) - Y_2(0)\right).$$

Let $W = W_1$, with $W_2 = 1 - W$. The estimator is then

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

To simplify the calculations, let $D = 2W - 1$, or $W = (D+1)/2$. Because $Pr(W = 1) = Pr(W = 0) = 1/2$, it follows that $D \in \{-1, 1\}$, and also $E[D] = 0$ and $D^2 = 1$. Then we can write:

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

Hence the variance of $\hat{\tau}$ is

$$V(\hat{\tau}) = \frac{1}{4}\left(Y_1(1) + Y_1(0) - Y_2(1) - Y_2(0)\right)^2.$$

To interpret this expression and see how in some cases it can be estimated, it is useful to look at the general case.

To calculate the variance of $\bar{y}_1 - \bar{y}_0$ for the general completely randomized experiment with M treated and $N - M$ controls, we require the second and cross moments of W_i. Because W_i is binary, the expectation of its square is equal to its expectation itself:

$$E[W_i^2] = E[W_i] = M/N.$$

To calculate the expectation of $W_i \cdot W_j$ in a completely randomized experiment, note that with the number of treated units fixed at M, the two events, unit i being treated and unit j being treated are not independent even if i differs from j, and thus $E[W_i \cdot W_j] \neq E[W_i] \cdot E[W_j] = M^2/N^2$. Rather:

$$E[W_i \cdot W_j] = Pr(W_i = 1, W_j = 1) = Pr(W_i = 1) \cdot Pr(W_j = 1|W_i = 1)$$

$$= \frac{M}{N} \cdot \frac{M-1}{N-1},$$

for $i \neq j$. Given the variance and covariance of W_i and W_j, a straightforward, but surprisingly long and tedious calculation, given in detail in the appendix to this chapter, shows that the variance of $\bar{y}_1 - \bar{y}_0$ is equal to

$$\text{Var}(\bar{y}_1 - \bar{y}_0) = \frac{S_0^2}{N-M} + \frac{S_1^2}{M} - \frac{S_{01}^2}{N}, \tag{1}$$

where S_0^2 is the variance of $Y_i(0)$ in the population, defined by:

$$S_0^2 = \frac{1}{N-1} \sum_{i=1}^{N} \Big(Y_i(0) - \bar{Y}(0) \Big)^2,$$

S_1^2 is the variance of $Y_i(1)$ in the population, defined by:

$$S_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} \Big(Y_i(1) - \bar{Y}(1) \Big)^2,$$

and S_{01}^2 is the population variance of the unit-level causal effect, defined by:

$$S_{01}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0))\Big)^2.$$

Let us consider the interpretation of the three components of this variance. The first and second components of this variance are easy to interpret. The average treatment effect is difference in average potential outcome: $\tau = \bar{Y}(1) - \bar{Y}(0)$. The first component of this difference, $\bar{Y}(1)$, the population average potential outcome under treatment, is estimated by the average outcome for the M treated units $\bar{y}_1$. This estimator is unbiased for the population average outcome under treatment, and its variance is S_1^2/M. Similarly the average outcome for the $N - M$ units assigned to control is unbiased for the population average outcome under the control treatment, and its variance is S_0^2/M.

If the treatment effect is constant, and $Y_i(1) - Y_i(0) = \bar{Y}(1) - \bar{Y}(0)$, the third component, S_{01}^2/N, is equal to zero. However, if there is variation in the treatment, the third term reduces the variance of the estimator for the average treatment effect. In general we can write

$$S_{01}^2 = S_0^2 + S_1^2 - 2 \cdot R_{01} \cdot S_1 \cdot S_0,$$

where R_{01} is the population correlation coefficient between the potential outcomes $Y(0)$ and $Y(1)$:

$$R_{01} = \frac{\sum_{i=1}^{N}\Big(Y_i(1) - \bar{Y}(1)\Big) \cdot \Big(Y_i(0) - \bar{Y}(0)\Big)/N}{\sqrt{\sum_{i=1}^{N}\Big(Y_i(1) - \bar{Y}(1)\Big)^2 \cdot \sum_{i=1}^{N}\Big(Y_i(0) - \bar{Y}(0)\Big)/N^2}},$$

which is restricted to the interval $[-1, 1]$. Substituting for S_{01}^2 and simplifying we can write:

$$\mathrm{Var}(\bar{y}_1 - \bar{y}_0) = S_0^2 \cdot \frac{M}{N \cdot (N - M)} + S_1^2 \cdot \frac{N - M}{N \cdot M} + 2 \cdot R_{01} \cdot S_0 \cdot S_1 \cdot \frac{1}{N}. \qquad (2)$$

The variance is highest when the correlation coefficient for $Y(0)$ and $Y(1)$ is equal to one, that is when the two potential outcomes are perfectly correlated. One important special case of this arises when the treatment effect is constant.

5.3 Estimating the Variance of $\bar{y}_1 - \bar{y}_0$

As discussed in Section 5.2, there are three components to the variance given in equation (1). The first two are easy to estimate. Consider the denominator of the first term,

$$S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(0) - \bar{Y}(0))^2.$$

An unbiased estimator for S_0^2

$$s_0^2 = \frac{1}{N-M-1}\sum_{i|W_i=0}(Y_i(0) - \bar{y}_0)^2 = \frac{1}{N-M-1}\sum_{i|W_i=0}(Y_i^{\text{obs}} - \bar{y}_0)^2.$$

(See the Appendix for a more formal derivation.) Similarly we can estimate S_1^2 by

$$s_1^2 = \frac{1}{M-1}\sum_{i|W_i=1}(Y_i(1) - \bar{y}_1)^2 = \frac{1}{M-1}\sum_{i|W_i=1}(Y_i^{\text{obs}} - \bar{y}_1)^2.$$

The third term is difficult to estimate because we have no direct observations on the variation in the treatment effect.

If the treatment effect is constant, the third term vanishes and we can obtain an unbiased estimator for the variance of the estimator for the average treatment effect as

$$V(\widehat{\bar{y}_1 - \bar{y}_0}) = \frac{s_0^2}{N-M} + \frac{s_1^2}{M}. \tag{3}$$

This estimator for the variance is widely used even when the assumption of constant treatment effects is inappropriate. There are two main reasons for this. One reason is that irrespective of the treatment effect heterogeneity, the expected value of this estimator of the variance is at least as large as the variance. Hence confidence intervals are in large samples going to be conservative, in line with the formal definition that requires coverage of confidence intervals to be at least equal to the pre-specified level. A second reason is that the variance is always unbiased for the variance of the estimator interpreted as an estimator for the super-population average treatment effect, as we shall discuss in more detail below.

An alternative estimator for the variance under constant treatment effects exploits the constant treatment effect assumption more fully. Under this assumption, it follows that the two potential outcome variances are equal, and thus we can define $S^2 \equiv S_0^2 = S_1^2$. Therefore we can combine the treated and control subpopulations for estimation of the common variance:

$$s^2 = \frac{1}{N-2}\Big(s_0^2 \cdot (N-M-1) + s_1^2 \cdot M\Big)$$

$$= \frac{1}{N-2}\Big(\sum_{i|W_i=0}(Y_i^{\text{obs}} - \bar{y}_0)^2 + \sum_{i|W_i=1}(Y_i^{\text{obs}} - \bar{y}_1)^2\Big),$$

to get a more efficient estimator for the variance of $\bar{y}_1 - \bar{y}_0$:

$$\tilde{V}(\bar{y}_1 - \bar{y}_0) = S_0^2 \cdot \left(\frac{1}{N-M} + \frac{1}{M}\right). \qquad (4)$$

Now let us consider estimation of the variance of $\bar{y}_1 - \bar{y}_0$ when the treatment effect is not assumed to be constant. It is clear that $\widehat{V}(\bar{y}_1 - \bar{y}_0)$, the variance estimator in equation (3), provides upwardly biased estimate of the variance in that case: the third term which vanishes if the treatment effect is constant, is now negative. The question arises whether we can improve upon this estimator for the variance. To see that there is indeed information to do so, note that the data do provide some information concerning the variation in the treatment effects, although they do not allow consistent estimation of its variance. As mentioned before, an implication of constant treatment effects is equality of the variances S_0^2 and S_1^2. Differences between these variances, measured by differences in the corresponding estimates s_0^2 and s_1^2, indicate variation in the treatment effect. To use this to find a better estimator for the variance of $\bar{y}_1 - \bar{y}_0$, let us turn to the representation of the variance in (2):

$$\text{Var}(\bar{y}_1 - \bar{y}_0) = S_0^2 \cdot \frac{M}{N \cdot (N-M)} + S_1^2 \cdot \frac{N-M}{N \cdot M} + 2 \cdot R_{01} \cdot S_0 \cdot S_1 \cdot \frac{1}{N}.$$

Conditional on a value for the correlation coefficient R_{01} we can estimate this variance as

$$\text{Var}(\bar{y}_1 - \bar{y}_0) = s_0^2 \cdot \frac{M}{N \cdot (N-M)} + s_1^2 \cdot \frac{N-M}{N \cdot M} + 2 \cdot R_{01} \cdot s_0 \cdot s_1 \cdot \frac{1}{N}. \tag{5}$$

The variance is highest if the two potential outcomes are perfectly correlated. In that case the estimator is

$$\bar{V}(\bar{y}_1 - \bar{y}_0) = s_0^2 \cdot \frac{M}{N \cdot (N-M)} + s_1^2 \cdot \frac{N-M}{N \cdot M} + 2 \cdot s_0 \cdot s_1 \cdot \frac{1}{N}. \tag{6}$$

In general this estimate of the variance is smaller than $\hat{V}(\bar{y}_1 - \bar{y}_0)$, although it is still upwardly biased if the correlation between the potential outcomes is less than unity. Alternatively we may choose another value for R_{01}, e.g., $R_{01} = 0$ and obtain a different estimate. In the example below we shall investigate the sensitivity of the results to this choice. Here it suffices to note that all other choices for R_{01} potentially lead to underestimates of the variance if the choice is in fact incorrect.

5.4 Inference for Superpopulation Average Treatment Effects

A second argument for the variance estimator $\hat{V}(\bar{y}_1 - \bar{y}_0)$, alluded to earlier, can be constructed through a re-direction of the focus on the average treatment effect. Suppose that the population subject to the randomized experiment is itself a random sample from a larger population. For simplicity we assume that this "super-population" is infinitely large. Furthermore, let us depart slightly from Neyman's focus on the finite population average treatment effect and focus on the superpopulation average treatment effect. In many cases this is not a major change of focus. Although in agricultural experiments it may be that a farmer is genuinely interested in which fertilizer is better for his specific plot of land, in most cases drug trials are conducted with a view to informing prescription policies for larger populations. It is clear, however, that this does not amount to much more than a slight of hand, as generally we cannot hope to learn anything more about the superpopulation than what we learn about the particular population in the study. In fact we typically learn strictly less about the superpopulation, but it is precisely this small amount of precision lost that enables us to obtain unbiased estimates of the variance of the estimator as an estimator of the superpopulation average treatment effect.

Sampling from the superpopulation induces a distribution on the two potential outcomes, and thus on the unit-level treatment effect, and finally on the average of the unit-level treatment effect in the experiment. Let $\tau_{SP} = E[Y(1) - Y(0)]$ be the expected value of the unit-level treatment effect under this distribution, or, equivalently, the average treatment effect in the superpopulation, and let σ_{01}^2 be the variance of the unit-level treatment effect.

The sample of size N, the population of interest in the discussion sofar, is assumed to be a simple random sample from this superpopulation. This implies that the average treatment effect in the sample, the finite population average treatment effect, $\tau_{FP} = \bar{Y}(1) - \bar{Y}(0)$, can be viewed as a random variable, with expectation and variance respectively equal to

$$E[\tau_{FP}] = E[\bar{Y}(1) - \bar{Y}(0)] = E\Big[Y(1) - Y(0)\Big] = \tau_{SP},$$

$$V(\tau_{FP}) = V(\bar{Y}(1) - \bar{Y}(0)) = \frac{\sigma_{01}^2}{N}.$$

In other words, the average treatment effect in the sample is unbiased for the superpopulation average treatment effect.

Next, let us consider the estimator $\bar{y}_1 - \bar{y}_0$ discussed in the previous subsections as an estimator for the superpopulation average treatment effect. The preceeding argument, combined with the law of iterated expectations, implies that this estimator is unbiased:

$$E[\bar{y}_1 - \bar{y}_0] = E\Big[E\Big[\bar{y}_1 - \bar{y}_0 \Big| Y_1(0), \ldots, Y_N(0), Y_1(1), \ldots, Y_N(1)\Big]\Big]$$

$$= E\Big[\bar{Y}(1) - \bar{Y}(0)\Big] = E[\tau_{FP}] = \tau_{SP}.$$

The variance of $\bar{y}_1 - \bar{y}_0$ can be written as

$$E\Big[\Big(\bar{y}_1 - \bar{y}_0 - E[Y(1) - Y(0)]\Big)^2\Big]$$

$$= E\Big[\Big(\bar{y}_1 - \bar{y}_0 - (\bar{Y}(1) - \bar{Y}(0)) + \bar{Y}(1) - \bar{Y}(0) - E[Y(1) - \bar{Y}(0)]\Big)^2\Big]$$

$$= E\Big[\Big(\bar{y}_1 - \bar{y}_0 - (\bar{Y}(1) - \bar{Y}(0))\Big)^2\Big] + E\Big[\Big(\bar{Y}(1) - \bar{Y}(0) - E[Y(1) - Y(0)]\Big)^2\Big]$$

$$+2 \cdot E\Big[\Big(\bar{y}_1 - \bar{y}_0 - (\bar{Y}(1) - \bar{Y}(0)) \cdot (\bar{Y}(1) - \bar{Y}(0) - E[Y(1) - Y(0)]\Big)\Big].$$

The third term in this variance is zero, because the expectation of the first factor in this term, conditional on $Y_1(0), \ldots, Y_N(0), Y_1(1), \ldots, Y_N(1)$ is zero. Hence the variance reduces to

$$V(\bar{y}_1 - \bar{y}_0)$$

$$= E\Big[\Big(\bar{y}_1 - \bar{y}_0 - \bar{Y}(1) - \bar{Y}(0)\Big)^2\Big] + E\Big[\Big(\bar{Y}(1) - \bar{Y}(0) - E[Y(1) - Y(0)]\Big)^2\Big]. \qquad (7)$$

The expectation of the first term, conditional on $Y_1(0), \ldots, Y_N(0), Y_1(1), \ldots, Y_N(1)$, is $S_0^2/(N-M) + S_1^2/M - S_{01}^2/N$, as in equation (1). The unconditional expectation of this term is $\sigma_0^2/(N-M) + \sigma_1^2/M - \sigma_{01}^2/N$. The expectation of the second term is σ_{01}^2/N. Hence the variance adds up to

$$V_{SP}(\bar{y}_1 - \bar{y}_0) = \frac{\sigma_0^2}{N-M} + \frac{\sigma_1^2}{M}.$$

We can estimate the super-population variance by substituting s_0^2 and s_1^2 for σ_0^2 and σ_1^2 respectively.

Let us compare this variance to the variance obtainedbefore for the finite population average treatment effect under the constant treatment effect assumption,

$$V_{FP}(\bar{y}_1 - \bar{y}_0) = \frac{S_0^2}{N-M} + \frac{S_1^2}{M}.$$

The expected value of this variance, under sampling from the super-population, is equal to V_{SP}.

Therefore, we can interpret the variance estimator in equation (3) in three different ways. First, it is an unbiased estimator for the variance of the average treatment effect in the finite population context under additive-constant treatment effects. Second, it is an upwardly biased estimator for the variance of this average treatment effect without assuming constant treatment effects. Finally, it is an unbiased estimator of the variance of the estimator of the superpopulation average treatment effect under random sampling from this superpopulation. The estimator for the variance that exploits the constant treatment effect assumption, given in (4), may be superior when that assumption holds, but it does not have the other two attractive properties, and therefore is rarely used. Similarly, the estimator in (6) may be better as a variance estimator for the finite sample average treatment effect, but it does not have desirable properties as an estimator for the variance of the superpopulation average treatment effect.

5.5 Confidence Intervals

In the introduction it was mentioned 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. Here we discuss a number of ways to construct confidence intervals. Let V be an estimate of the variance of $\hat{\tau}$. If we wish to construct a 90% confidence interval, we use the 5th and 95th percentile of the standard normal distribution, -1.645 and 1.645 respectively, with the implied 90% confidence interval equal to

$$\{\hat{\tau} - 1.645 \cdot \sqrt{V}, \hat{\tau} + 1.645 \cdot \sqrt{V}\}.$$

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

$$\{\hat{\tau} - c_{\alpha/2} \cdot \sqrt{V}, \hat{\tau} + c_{\alpha/2} \cdot \sqrt{V}\}.$$

This applies to all estimates of the variance, and the resulting confidence intervals are valid, in large samples, under the assumptions that make the corresponding variance estimates unbiased or upwardly biased estimates of the true variance.

In addition to the confidence intervals based on analytic variance estimates we consider two bootstrap based confidence intervals. First we consider the fully nonparametric bootstrap. Each bootstrap sample is constructed by drawing N units from the sample with replacement. Given bootstrap sample j we calculate $\hat{\tau}_j$ as the estimated average treatment effect for that sample. We repeat this B times. Given the B bootstrap estimates $\hat{\tau}_j$, $j = 1, \ldots, B$, we then calculate the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical distribution, $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively, and construct the $(1-\alpha) \times 100\%$ confidence interval as

$$\{q_{\alpha/2}, q_{1-\alpha/2}\}.$$

5.6 Comparison to Linear Regression

We can motivate the difference in treatment-control averages in an alternative way that is familiar in a different context, namely linear regression. In a linear regression one regress the observed outcome Y_i^{obs} on the indicator for the treatment, W_i and a constant:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i.$$

The least squares estimator for τ is based on minimizing, over α and τ,

$$\sum_{i=1}^{N} (Y_i^{\text{obs}} - \alpha - \tau \cdot W_i)^2,$$

with solution

$$\hat{\tau}_{ls} = \frac{\sum_{i=1}^{N} (W_i - \bar{W}) \cdot (Y_i^{\text{obs}} - \bar{Y}^{\text{obs}})}{\sum_{i=1}^{N} (W_i - \bar{W})^2},$$

and

$$\hat{\alpha}_{ls} = \bar{Y}^{\text{obs}} - \hat{\tau}\bar{W}.$$

The least squares estimates of τ would sometimes be interpreted as estimates of the causal effect of the treatment. More typically, textbooks would stress that these estimates measure "association" between the two variables, and that causal interpretations are not warranted in general. The assumptions typically used in this context are that the residuals ε_i are independent of, or at least uncorrelated with, the treatment indicator W_i. These assumptions are typically difficult to evaluate as it is rarely made explicit what the interpretation of these residuals is beyond a vague notion of capturing other factors affecting the outcomes of interest.

Here we shall look at the least squares estimator and its properties under the key assumption that assignment of treatment is random. In addition, we maintain the assumption made in Section 5.4 of random sampling from a superpopulation. Let α be the superpopulation average $E[Y_i(0)]$, and τ the population average treatment effect $E[Y_i(1) - Y_i(0)]$. Define the residual ε_i as

$$\varepsilon_i = Y_i^{\text{obs}} - \alpha - \tau \cdot W_i = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0) - \alpha - \tau \cdot W_i$$

$$= Y_i(0) - \alpha + W_i \cdot (Y_i(1) - Y_i(0) - \tau).$$

Random assignment implies that the assignments are independent of the potential outcomes:

$$W_i \perp (Y_i(0), Y_i(1)),$$

, which has two implications we shall use here. The first is that the assignment is independent of the outcomes under the null treatment:

$$W_i \perp Y_i(0),$$

and the second is that the assignment is independent of the causal effects:

$$W_i \perp Y_i(1) - Y_i(0).$$

The combination of random assignment and sampling from a super-population implies that

$$E[\varepsilon_i] = E\left[Y_i(0) - \alpha + W_i \cdot (Y_i(1) - Y_i(0) - \tau)\right] = 0,$$

and

$$E[W_i \cdot \varepsilon_i] = E\left[W_i \cdot (Y_i(0) - \alpha) + W_i \cdot (Y_i(1) - Y_i(0) - \tau)\right] = 0.$$

implying unbiasedness of the least squares estimator for $E[Y(1) - Y(0)]$.

Simple algebra shows that the least squares estimator is in fact equal to $\hat{\tau} = \bar{y}_1 - \bar{y}_0$, so these results could have been asserted directly using the arguments in the preceeding sections. However, the above derivation shows how the assumptions commonly made in least squares analyses in the randomized experiment context follow from the randomization, and thus have a scientific basis.

Assuming homoskedasticity, the variance of the residuals would be estimated as

$$\hat{\sigma}^2_\varepsilon = \frac{1}{N-2} \sum_{i=1}^N (Y_i^{\rm obs} - \hat{\alpha} - \bar{Y}(1) - \bar{Y}(0))^2$$

$$= \frac{1}{N-2} \Big(\sum_{i|W_i=0} (Y_i^{\rm obs} - \bar{y}_0)^2 + \sum_{i|W_i=1} (Y_i^{\rm obs} - \bar{y}_1)^2 \Big).$$

The corresponding estimator for the variance of $\hat{\tau} = \bar{y}_1 - \bar{y}_0$ is then

$$\hat{V}(\hat{\tau})_{homosk.} = \frac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^N (W_i - N/M)^2} = \hat{\sigma}^2_\varepsilon \cdot \Big(\frac{1}{N-M} + \frac{1}{M} \Big).$$

This is the same estimator for the variance justified under constant treatment effects. This is not surprising as homoskedasticity here implies that the variance of ε does not vary with W, and, by definition $\varepsilon = Y_i(0) - \alpha + W_i \cdot (Y_i(1) - Y_i(0) - \bar{Y}(1) - \bar{Y}(0))$, a constant variance for ε implies constant variances for the two potential outcomes.

Note however that the random assignment assumption implies zero correlation between the assignment and the residual, but not full independence. In general there is therefore no justification for the homoskedasticity assumption, and one may wish to use an estimator for the variance that allows for heteroskedasticity. One such estimator is:

$$heteroskedasticityconsistentvarianceestimator$$

In large samples this variance estimator is close to the variance estimator derived before, and given in equation (3).

5.7 An Example: The Fresno Educational Program

To illustrate the approach discussed in this chapter we take another look at the data introduced in the previous from a randomized experiment to evaluate the effect of an educational television program on reading skills. The unit is a class of students. The outcome of interest is the average reading score in the class. We take the $N = 23$ observations (classrooms) from Fresno, $M = 11$ of which are assigned to the reading program and $N - M = 12$ of which are assigned to the control group. The average testscore for the classrooms assigned to the program is 73.2 for the treated group and 73.0 for the control group. The two variance estimates are $s_0^2 = 180.1$ and $s_1^2 = 252.1$. The variance estimate under common variances is $s^2 = 217.8$. Table 2 presents estimates of the variance and confidence intervals for all seven estimators, with a description of all seven estimators presented in Table 1. Note that the confidence intervals are very sensitive to the assumption on the correlation between the two potential outcomes. If we, conservatively, assume that the correlation is equal to unity, the implied standard error is 6.1. If we assume the correlation is zero, this reduces to 4.3, and if we assume the correlation is negative one, the standard error is only 0.2.

To see how accurate these variance estimators are and assess the coverate rate of the corresponding confidence interval we carry out a small Monte Carlo experiment around this data set. We use four data generating processes. In each case we assign 110 units to the active treatment and 120 units to the control treatment. In the first data generating process the treatment effect is constant. The distbution of $Y_i(0)$ is normal with mean 71.3 and standrad deviation 15.4. The distribution of $Y_i(1)$ is normal with mean 73.5 and standard deviation 15.4. In the other three data generating processes we allow for treatment effect heterogeneity. In each case the marginal distribution of $Y_i(0)$ is normal with mean 71.3 and standard deviation is 10.7, and the marginal distribution of $Y_i(1)$ is normal with mean 73.5 and standard deviation 21.7. In the last three data generating processes the correlation between $Y_i(0)$ and $Y_i(1)$ is 1, 0, and -1 respectively.

In each case we draw 100,000 samples. For each sample we first calculate the finite sample population average treatment effect and its estimate. We then construct the seven 90% confidence intervals and check whether the finite sample and super population average treatment effect are inside the confidence intervals. We also calculate and report the average of the width of the confidence intervals. Note that for the boostrap confidence intervals we only carry out 1,000 replications, with for each sample 1,000 bootstrap samples.

Note that the first variance estimator leads to confidence intervals that are always conservative. If the treatment effect is constant, its coverage is very close to 90% both for the finite population and the super population average treatment effect. If the treatment effect varies, its coverage for the finite population average treatment effect is often much higher than 90%, but the coverage for the super population average treatment effect is always 90%. For the other four analytic variance estimates this is not the case. In some cases for the finite population average treatment effect their confidence intervals have coverage closer to 90%, and they are on average tighter, but this comes at the expense of losing some of the coverage of the super population average treatment effect. This is particularly important for the third variance estimator, based on the assumption of a correlation of unity between $Y_i(0)$ and $Y_i(1)$. As argued before, this estimator for the variance is conservative for the finite population average treatment effect, and at the same time is generally smaller than the constant treatment effect based variance. Nevertheless, it does not necessarily have the nominal coverage for the super population average treatment effect.

The fifth variance estimator based on a correlation of minus one performs well in terms of coverage for the finite population average treatment effect when this assumption is true, but does poorly in all other respects.

Overall, the simulations support the use of the constant-treatment effect based variance estimates.

Appendix: Variance Calculation

For simplicity we focus on the case with $2N$ units, N of which receive the treatment and N who receive control. The average treatment effect is

$$\bar{Y}(1) - \bar{Y}(0) = \sum_{i=1}^{2N} (Y_i(1) - Y_i(0))/(2N).$$

The estimate is

$$\bar{y}_1 - \bar{y}_0 = \sum_{i=1}^{2N} (w_i Y_i(1) - (1 - w_i) Y_i(0))/N,$$

where w_i is the treatment indicator. Define $d_i = 2w_i - 1$, so $d_i = 1$ for treated units and $d_i = -1$ for control units. The expectation of d_i is zero, both under the completely randomized and under the pairwise randomized design. In terms of d_i the estimate of the average treatment effect is

$$\begin{aligned} \bar{y}_1 - \bar{y}_0 &= \sum_{i=1}^{2N} \Big(\frac{d_i + 1}{2} Y_i(1) - \frac{1 - d_i}{2} Y_i(0) \Big)/N \\ &= \sum_{i=1}^{2N} (Y_i(1) - Y_i(0))/(2N) + d_i(Y_i(1) + Y_i(0))/(2N) \qquad (8) \end{aligned}$$

which has expectation equal to the first term which is the average treatment effect $\bar{Y}(1) - \bar{Y}(0)$.

The preceeding argument implies that the variance of $\bar{y}_1 - \bar{y}_0$ is equal to the variance of the second term in (8), which, using X_i as shorthand for $Y_i(1) + Y_i(0)$, is equal to the expectation of

$$\Big(\sum_{i=1}^{2N} d_i X_i \Big)^2 / (4N^2) \qquad (9)$$

As noted before, the expectation of d_i is equal to zero. Also note that $d_i^2 = 1$. Finally, the probability of $d_i = d_j$ for $j \neq i$ is equal to $Pr(w_i = w_j) = 2 \cdot Pr(w_i = w_j = 1) = 2 \cdot Pr(w_i =$

$1)\cdot Pr(w_j=1|w_i=1)=2\cdot(1/2)\cdot(N-1)/(2N-1)=(N-1)/(2N-1)$. The expectation of d_id_j is therefore $Pr(d_i=d_j)-Pr(d_i\neq d_j)=-1/(2N-1)$.

Writing out (9), we get

$$E\sum_{i=1}^{2N}\sum_{j=1}^{2N}d_id_jX_iX_j/(4N^2)$$

$$=\sum_{i=1}^{2N}X_i^2\Big(1+\frac{1}{2N-1}\Big)\Big/(4N^2)-\sum_{i=1}^{2N}\sum_{j=1}^{2N}\frac{1}{2N-1}X_iX_j\Big/(4N^2)$$

$$=\frac{1}{2N(2N-1)}\sum_{i=1}^{2N}X_i^2-\frac{1}{4N^2(2N-1)}\sum_{i=1}^{2N}\sum_{j=1}^{2N}X_iX_j$$

$$=\frac{1}{2N(2N-1)}\sum_{i=1}^{2N}(X_i-\overline{X})^2$$

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

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

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

$$+\frac{2}{2N(2N-1)}\sum_{i=1}^{2N}(Y_i(1)-\overline{Y(1)})(Y_i(0)-\overline{Y(0)}). \qquad (10)$$

Define

$$S_1^2=\frac{1}{2N-1}\sum_{i=1}^{2N}(Y_i(1)-\overline{Y(1)})^2,$$

$$S_0^2=\frac{1}{2N-1}\sum_{i=1}^{2N}(Y_i(0)-\overline{Y(0)})^2,$$

and

$$S_{01}^2=\frac{1}{2N-1}\sum_{i=1}^{2N}(Y_i(1)-Y_i(0)-\overline{Y(1)-Y(0)})^2.$$

Then, we have

$$S_{01}^2 = \frac{1}{2N-1}\sum_{i=1}^{2N}(Y_i(1) - \overline{Y(1)} - [Y_i(0) - \overline{Y(0)}])^2$$

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

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

Thus (10) equals

$$-\frac{S_{01}^2}{2N} + \frac{S_0^2}{2N} + \frac{S_1^2}{2N},$$

and the variance of $\bar{y}_1 - \bar{y}_0$ is equal to

$$-\frac{S_{01}^2}{2N} + \frac{S_0^2}{N} + \frac{S_1^2}{N}.$$

Under additivity S_{01}^2 is equal to zero, so we only need to find unbiased estimates for S_0^2 and S_1^2. Consider the estimator

$$S_1^2 = \frac{1}{N-1}\Big[\sum_{i=1}^{2N} w_i Y_i(1)^2 - N\Big(\frac{1}{N}\sum_{i=1}^{2N} w_i Y_i(1)\Big)^2\Big].$$

The expectation of the first term is

$$\frac{1}{2N-2}\sum_{i=1}^{2N} Y_i(1)^2. \qquad (11)$$

For the expectation of the second term, recall that the expectation of w_i^2 and the expectation of w_i are both equal to $1/2$, and that for $i \neq j$ the expectation of $w_i w_j$ is equal to $(N-1)/(4N-2)$. So, the second term in S_1^2,

$$-\frac{N}{N-1}\Big(\frac{1}{N}\sum_{i=1}^{2N} w_i Y_i(1)\Big)^2$$

has expectation

$$-E\frac{}{N(N-1)}\sum_{i=1}^{2N}\sum_{j=1}^{2N} w_i w_j Y_i(1)Y_j(1)$$

$$= -\frac{1}{N(N-1)}\sum_{i=1}^{2N}\Big(\frac{1}{2}-\frac{N-1}{4N-2}\Big)Y_i(1)^2 - \frac{1}{N(N-1)}\sum_{i=1}^{2N}\sum_{j=1}^{2N}\frac{N-1}{4N-2}Y_i(1)Y_j(1).$$

Adding in (11) and collecting terms we have

$$ES_1^2 = \sum_{i=1}^{2N} Y_i(1)^2\Big(\frac{1}{2N-2}-\frac{1}{N(N-1)2}+\frac{N-1}{N(N-1)(4N-2)}\Big)$$

$$-\sum_{i=1}^{2N}\sum_{j=1}^{2N} Y_i(1)Y_j(1)\frac{N-1}{N(N-1)(4N-2)}$$

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

$$= \frac{1}{2N-1}\sum_{i=1}^{2N}(Y_i(1)-\overline{Y(1)})^2 = S_1^2.$$

Similarly,

$$S_0^2 = ES_0^2 = E\frac{1}{N-1}\Big[\sum_{i=1}^{2N}(1-w_i)Y_i(0)^2 - 2N\Big(\frac{1}{N}\sum_{i=1}^{2N}(1-w_i)Y_i(0)\Big)^2\Big].$$

Table 1: VARIANCE ESTIMATORS

Var I	$s_0^2/(N-M)+s_1^2/M$
Var II	$s^2(1/(N-M)+1/M)$
Var III	$s_0^2 M/(N(N-M))+s_1^2(N-M)/(NM)+2s_0 s_1/N$
Var IV	$s_0^2 M/(N(N-M))+s_1^2(N-M)/(NM)$
Var V	$s_0^2 M/(N(N-M))+s_1^2(N-M)/(NM)-2s_0 s_1/N$
Var VI	Unconditional Bootstrap
Var VII	Bootstrap Conditional on N_c and N_t

Table 2: 95% Confidence Intervals for Average Treatment Effects: Fresno Data

Variance Estimator	Estimate of Standard Error	Lower Bound 95% CI	Upper Bound 95% CI
Var I	6.11	-9.90	10.21
Var II	6.16	-9.98	10.29
Var III	6.09	-9.87	10.18
Var IV	4.31	-6.94	7.25
Var V	0.25	-0.25	0.56
Var VI	5.96	-9.19	10.21
Var VII	5.92	-10.08	9.42

Table 3: 90% Confidence Interval Simulations: $N_t = 110$, $N_c = 120$, 100,000 replications. Coverage rates for finite population and super population average treatment effects

	Design I			Design II			Design III			Design IV		
$Y(0)$	$\mathcal{N}(71.3, 15.4^2)$			$\mathcal{N}(71.3, 10.7^2)$			$\mathcal{N}(71.3, 10.7^2)$			$\mathcal{N}(71.3, 10.7^2)$		
$Y(1)$	$\mathcal{N}(73.5, 15.4^2)$			$\mathcal{N}(73.5, 21.7^2)$			$\mathcal{N}(73.5, 21.7^2)$			$\mathcal{N}(73.5, 21.7^2)$		
R_{01}	1			1			0			-1		
	τ_{FP}	τ_{SP}	width	τ_{FP}	τ_{SP}	width	τ_{FP}	τ_{SP}	width	τ_{FP}	τ_{SP}	width
Var I	0.898	0.898	6.67	0.914	0.898	7.54	0.977	0.890	7.54	1.000	0.897	7.54
Var II	0.898	0.898	6.67	0.905	0.889	7.35	0.974	0.891	7.35	1.000	0.888	7.35
Var III	0.898	0.898	6.66	0.897	0.880	7.16	0.970	0.882	7.16	1.000	0.879	7.16
Var IV	0.754	0.755	4.72	0.783	0.760	5.41	0.900	0.762	5.41	0.999	0.760	5.41
Var V	0.070	0.070	0.35	0.457	0.438	2.66	0.580	0.438	2.66	0.890	0.436	2.66
Var VI	0.889	0.889	6.66	0.900	0.885	7.53	0.974	0.904	7.52	1.000	0.899	7.51
Var VII	0.896	0.896	6.63	0.914	0.897	7.50	0.976	0.900	7.48	1.000	0.910	7.52