# Chapter 2

# Stratified and Paired Randomized Experiments

## 2.1 Introduction

In this chapter we introduce a different class of randomized experiments, stratified and paired randomized experiments. The issues raised in this chapter, however, go beyond randomized experiments and are important in many observational study settings too. The starting point is a population that is partitioned into $G$ subpopulations, referred to as groups, and, in other settings, as clusters or strata. It is important that this is a partitioning of the population. Each unit in the population is a member of only one group or stratum. The characteristics or features of a unit which determine which subpopulation a unit is a member of may vary. In fact, group membership can be chosen by the researcher. What is important is that any uncertainty measures take as fixed the subpopulation membership.

Let $G_i \in \mathcal{G} = \{1, \ldots, G\}$ denote the group or stratum that unit $i$ is a member of, and let $G_{ig} = \mathbf{1}_{G_i=g}$ be a binary indicator for unit $i$ being a member of stratum $g$. The number of units in stratum $g$ is $N_g$, with the total population size $N = \sum_{g=1}^{G} N_g$.

## 2.2 Stratified Randomized Experiments

Within stratum $g$ we randomly assign $N_{\mathrm{c},g}$ and $N_{\mathrm{t},g}$ units to the control and treatment groups respectively. The repeatedly sampling perspective is the same as in the previous chapter: the repeated sampling variances are conditional on the potential outcomes and the number of control and treated units in each stratum.

The estimand remains the same, the sample average effect of the treatment. Defining stratum specific sample average effects as

$$\tau_{\mathrm{fs},g} = \frac{1}{N_g} \sum_{i:G_i=g} \Big( Y_i(1) - Y_i(0) \Big),$$

we can write the estimand as a weighted average of the within-stratum average effects:

$$\tau_{\mathrm{fs}} = \frac{1}{N} \sum_{i=1}^{N} \Big( Y_i(1) - Y_i(0) \Big) = \sum_{g=1}^{G} \frac{N_g}{N} \tau_{\mathrm{fs},g}.$$

The assignment mechanism in a stratified randomized experiment fixes the number of treated and control units in each stratum, $N_{\mathrm{t},g}$ and $N_{\mathrm{c},g}$ in stratum $g$ respectively, so that

$$\mathrm{pr}(\mathbf{W} = \mathbf{w}) = \prod_{g=1}^{G} \left( \begin{array}{c} N_g \\ N_{\mathrm{t},g} \end{array} \right)^{-1} \qquad \text{for all } \mathbf{w} \text{ s.t. for all } g,\ \sum_{i=1}^{N} G_{ig} W_i = N_{\mathrm{t},g}.$$

This has two implications. First, it limits the set of assignments that have positive probability. Second, it may lead to variation in the probabilities for feasible assignments. The hope is that the design eliminates values for the assignment vector that are uninformative about the questions of interest.

Here we discuss the estimation of $\tau_{\mathrm{fs}}$ using Neyman's approach. Testing sharp null hypotheses using Fisher's approach is a direct extension of the methods discussed in the previous chapter. First consider estimation of and inference for a stratum-specific average effect, $\tau_{\mathrm{fs},g}$. For this estimand the results in the previous chapter directly apply. The natural estimator is

$$\hat{\tau}_g = \overline{Y}^{\mathrm{obs}}_{\mathrm{t},g} - \overline{Y}^{\mathrm{obs}}_{\mathrm{c},g},$$

where

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

The exact variance of this estimator is

$$\mathbb{V}_g = \mathbb{V}\left(\hat{\tau}_g\right) = \frac{S^2_{\mathrm{c},g}}{N_{\mathrm{c},g}} + \frac{S^2_{\mathrm{t},g}}{N_{\mathrm{t},g}} - \frac{S^2_{\mathrm{ct},g}}{N_g}, \tag{2.1}$$

where

$$S^2_{\mathrm{c},g} = \frac{1}{N_g - 1} \sum_{i:G_{i,g}=1} \Big(Y_i(0) - \overline{Y}(0)\Big)^2, \qquad S^2_{\mathrm{t},g} = \frac{1}{N_g - 1} \sum_{i:G_{i,g}=1} \Big(Y_i(1) - \overline{Y}(1)\Big)^2,$$

and

$$S^2_{\mathrm{ct},g} = \frac{1}{N_g - 1} \sum_{i:G_{i,g}=1} \Big(Y_i(1) - Y_i(0) - \big(\overline{Y}(1) - \overline{Y}(0)\big)\Big)^2.$$

As in the previous chapter, there is no unbiased estimator for $S^2_{\mathrm{ct},g}$, so the standard approach is to ignore this term. Then the natural, but conservative, estimator for $\mathbb{V}(\hat{\tau}_g)$ is

$$\hat{\mathbb{V}}_g = \frac{s^2_{\mathrm{c},s}}{N_{\mathrm{c},g}} + \frac{s^2_{\mathrm{t},g}}{N_{\mathrm{t},g}},$$

where

$$s^2_{\mathrm{c},g} = \frac{1}{N_{\mathrm{c},g} - 1} \sum_{i:G_{i,g}=1,W_i=0} \Big(Y_i^{\mathrm{obs}} - \overline{Y}^{\mathrm{obs}}_{\mathrm{c},g}\Big)^2, \qquad \text{and } s^2_{\mathrm{t},g} = \frac{1}{N_{\mathrm{t},g} - 1} \sum_{i:G_{i,g}=1,W_i=1} \Big(Y_i^{\mathrm{obs}} - \overline{Y}^{\mathrm{obs}}_{\mathrm{t},g}\Big)^2,$$

are unbiased estimators for $S^2_{\mathrm{c},g}$ and $S^2_{\mathrm{t},g}$ respectively. Given these within-stratum estimates of the average effect, we estimate the overall average effect as a weighted average:

$$\hat{\tau}^{\mathrm{strat}} = \sum_{g=1}^{G} \frac{N_g}{N} \hat{\tau}_g.$$

Conditional on the stratum sizes, and the fractions of treated units within each stratum, and still conditional on the potential outcomes, the variance for this estimator is

$$\mathbb{V}\left(\hat{\tau}^{\mathrm{strat}}\right) = \sum_{g=1}^{G} \left(\frac{N_g}{N}\right)^2 \mathbb{V}_g,$$

for which a conservative estimator is

$$\hat{\mathbb{V}}^{\mathrm{strat}} = \sum_{g=1}^{G} \left(\frac{N_g}{N}\right)^2 \hat{\mathbb{V}}_g.$$

## 2.3 Paired Randomized Experiments

A paired randomized experiment is a special case of a stratified randomized experiment where the number of strata is half the number of units, $G = N/2$, the number of units in each stratum is $N_g = 2$, with a single treated and control unit, $N_{\mathrm{c},g} = N_{\mathrm{t},g} = 1$.

Although one can write the estimator for the finite sample average treatment effect in exactly the same way as in the general stratified case, it is useful to modify the notation to take account of the special features of this design. Let $t(g) \in \{1, \ldots, N\}$ be the index for the treated unit in stratum $g$, and let $c(g) \in \{1, \ldots, N\}$ be the index for the control unit in stratum $g$. Now we can write the estimator for the pair-specific estimator as

$$\hat{\tau}_g = Y_{t(g)}^{\mathrm{obs}} - Y_{c(g)}^{\mathrm{obs}},$$

with the estimator for the average treatment effect

$$\hat{\tau} = \frac{1}{G}\sum_{g=1}^{G}\Big(Y_{t(g)}^{\mathrm{obs}} - Y_{c(g)}^{\mathrm{obs}}\Big) = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}.$$

Its variance is the same as in the general stratified case. The main difference with the general stratified case arises with inference. Because there is only a single unit for each treatment in each stratum, it is not possible to use the within-treatment within-stratum variance estimator used before,

$$s_{\mathrm{c},g}^2 = \frac{1}{N_{\mathrm{c},g}-1}\sum_{i:G_i=g,W_i=0}\Big(Y_i^{\mathrm{obs}} - \overline{Y}_{\mathrm{c},g}^{\mathrm{obs}}\Big)^2, \qquad \text{and} \quad s_{\mathrm{t},g}^2 = \frac{1}{N_{\mathrm{t},g}-1}\sum_{i:G_i=g,W_i=1}\Big(Y_i^{\mathrm{obs}} - \overline{Y}_{\mathrm{t},g}^{\mathrm{obs}}\Big)^2.$$

In practice a common variance estimator is

$$\hat{\mathbb{V}}_{\mathcal{P}} = \frac{1}{G(G-1)}\sum_{g=1}^{G}\big(Y_{t(g)}^{\mathrm{obs}} - Y_{c(g)}^{\mathrm{obs}} - \hat{\tau}\big)^2.$$

This variance is conservative in two ways. First, as before, it assumes perfect correlation between the potential outcomes for the same unit. Second, it ignores the variation in treatment effects by stratum. This is important, because stratified designs where the variance is estimated first within strata, and then averaged over strata, do capture this variation. The implication is that we recommend, whenever possible, to stratify using small strata, with at least two treated and control units within each stratum, rather than pairwise randomization. Although pairwise randomization may lead to smaller variances, the complications arising from the estimation of the variance, reduce their appeal.

## 2.4 Adjusting by Design versus Adjusting by Analysis

In this section we investigate the benefits of stratification in randomized experiments, and compare it to ex post adjustment for pre-treatment variables. We show that in terms of expected-squared-error, stratification (with the same treatment probabilities in each stratum) cannot be worse than complete randomization, even in small samples, and even with little or no correlation between covariates and outcomes. *Ex ante*, committing to stratification can only improve precision, not lower it. There are two important qualifications to this result. First, *ex post*, given the joint distribution of the covariates in the sample, a particular stratification may be inferior to complete randomization. Second, the result requires that the sample can be viewed as a (stratified) random sample from an infinitely large population, with infinitely large strata, with the expectation in the expected-squared-error taken over this super-population. This requirement guarantees that outcomes within strata cannot be negatively correlated.

The lack of any finite sample cost to ex ante stratification in terms of expected-squared-error contrasts with the potential cost of ex post stratification, or covariance adjustment. Although ex post adjustment for covariates through regression has no cost in large samples, it does have a cost in finite samples, and may in fact increase the finite sample variance. It will do so on average for any sample size, even in the case where the covariates have no predictive power whatsoever.

Although there is no cost to stratification in terms of the variance of the estimator for the average treatment effect, there is a cost in terms of estimation of this variance. In expectation the estimated variance based on taking account of the stratification is less than or equal to the variance in a completely randomized experiment. However, this estimator for the variance typically has itself a larger variance, related to the degrees of freedom adjustment. In my view this should not be interpreted, however, as an argument against stratification. One can always use the variance that ignores the stratification: this is conservative if the stratification did in fact reduce the variance.

### 2.4.1 Set Up

There is an infinitely large (super-)population. For ease of exposition we focus on the simplest case where this population is partitioned into two strata, say f (females) and m (males). The share of females in the population is $q$. In the population the average of $Y_i(0)$ for females and males is $\mu_{\mathrm{c,f}}$ and $\mu_{\mathrm{c,m}}$ respectively, and the corresponding averages for $Y_i(1)$ are $\mu_{\mathrm{t,f}}$ and $\mu_{\mathrm{t,m}}$. The corresponding conditional population variances are $\sigma^2_{\mathrm{c,f}}$, $\sigma^2_{\mathrm{c,m}}$, $\sigma^2_{\mathrm{t,f}}$, and $\sigma^2_{\mathrm{t,m}}$. The population average treatment effect is $\tau^{\mathrm{sp}} = \mu_{\mathrm{t}} - \mu_{\mathrm{c}}$, where the two marginal means are $\mu_{\mathrm{t}} = q\mu_{\mathrm{t,f}} + (1-q)\mu_{\mathrm{t,m}}$, and $\mu_{\mathrm{c}} = q\mu_{\mathrm{c,f}} + (1-q)\mu_{\mathrm{c,m}}$. Define the marginal variance $\sigma^2_{\mathrm{c}} = \mathbb{V}(Y_i(0))$, which can be written, using the law of iterated expectations, in terms of the conditional variances and means as $q\sigma^2_{\mathrm{c,f}} + (1-q)\sigma^2_{\mathrm{c,m}} + q(1-q)(\mu_{\mathrm{c,f}} - \mu_{\mathrm{c,m}})^2$, and similarly define $\sigma^2_{\mathrm{t}} = \mathbb{V}(Y_i(1))$, which can be written as $q\sigma^2_{\mathrm{t,f}} + (1-q)\sigma^2_{\mathrm{t,m}} + q(1-q)(\mu_{\mathrm{t,f}} - \mu_{\mathrm{t,m}})^2$.

The sampling scheme is also stratified, in order to allow for exact finite sample calculations. Formally, we draw $N_{\mathrm{f}}$ individuals randomly from the female subpopulation, and $N_{\mathrm{m}}$ from the male subpopulation, with $N = N_{\mathrm{f}} + N_{\mathrm{m}}$ the total sample size. We fix these strata shares to exactly equal population shares, so $q = N_{\mathrm{f}}/N$.

We consider two experimental designs. In the first, the completely randomized design, denoted by $\mathcal{C}$, we randomly select $N_{\mathrm{t}}$ individuals from the sample to be exposed to the treatment, and assign the remaining $N_{\mathrm{c}} = N - N_{\mathrm{t}}$ individuals to the control group. Let $p = N_{\mathrm{t}}/N$ be the share of treated units. In the second, stratified design, denoted by $\mathcal{S}$, we randomly select $N_{\mathrm{t,f}}$ individuals from the female subsample and assign those to the treatment group, and randomly select $N_{\mathrm{t,m}} = N_{\mathrm{t}} - N_{\mathrm{t,f}}$ individuals from the male subsample and assign those to the treatment group, and assign the remaining individuals in the sample to the control treatment. To ease the comparison of the designs we enforce in the stratified design equal assignment probabilities within the strata, and equal marginal assignment probabilities in the two designs, so that $N_{\mathrm{t,f}}/N_{\mathrm{f}} = N_{\mathrm{t,m}}/N_{\mathrm{m}} = p$.

### 2.4.2 The Benefits from Stratification for Point Estimation in Finite Samples

In the completely randomized design the natural estimator is

$$\hat{\tau} = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}. \tag{2.2}$$

The variance of this estimator, relative to the population average treatment effect $\tau^{\rm sp}$, is, using the arguments from the previous chapter,

$$\mathbb{V}_{\mathcal{C}} = \mathbb{V}(\hat{\tau}|\mathcal{C}) = \frac{\sigma_{\rm t}^2}{pN} + \frac{\sigma_{\rm c}^2}{(1-p)N}.$$

Note that we do not have the term involving the variance of the treatment effect that arose in the earlier representations, e.g., (2.1)because we focus on the super-population average treatment effect rather than the finite sample average treatment effect. Substituting in for $\sigma_{\rm t}^2$ and $\sigma_{\rm c}^2$, this can be written as,

$$q(1-q)(\mu_{\rm c,f} - \mu_{\rm c,m})^2 + \frac{q\sigma_{\rm c,f}^2}{(1-p)N} + \frac{(1-q)\sigma_{\rm c,m}^2}{(1-p)N}$$

$$+q(1-q)(\mu_{\rm t,f} - \mu_{\rm t,m})^2 + \frac{q\sigma_{\rm t,f}^2}{pN} + \frac{(1-q)\sigma_{\rm t,m}^2}{pN},$$

Next, we calculate the variance under the stratified design. In that one natural estimator is the one that averages the within-stratum differences by treatment status:

$$\hat{\tau}^{\rm strat} = q\hat{\tau}_{\rm f} + (1-q)\hat{\tau}_{\rm m}, \tag{2.3}$$

where

$$\hat{\tau}_{\rm f} = \frac{1}{pqN} \sum_{i:G_i={\rm f},W_i=1} Y_i^{\rm obs} - \frac{1}{(1-p)qN} \sum_{i:G_i={\rm f},W_i=0} Y_i^{\rm obs},$$

and

$$\hat{\tau}_{\rm m} = \frac{1}{p(1-q)N} \sum_{i:G_i={\rm m},W_i=1} Y_i^{\rm obs} - \frac{1}{(1-p)(1-q)N} \sum_{i:G_i={\rm m},W_i=0} Y_i^{\rm obs}.$$

Because in this special case the assignment probabilities are the same in the two strata, it follows that the regression estimator has the same form as the simple difference by treatment status:

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

The variance in general satisfies

$$\mathbb{V}_{\mathcal{S}} = \mathbb{V}(\hat{\tau}|\mathcal{S}) = q^2\mathbb{V}(\hat{\tau}_{\rm f}) + (1-q)^2\mathbb{V}(\hat{\tau}_{\rm m}).$$

The two subpopulation variances are, using the arguments from the previous chapter,

$$\mathbb{V}(\hat{\tau}_{\rm f}) = \frac{\sigma^2_{\rm c,f}}{pqN} + \frac{\sigma^2_{\rm t,f}}{(1-p)qN}, \qquad \text{and} \quad \mathbb{V}(\hat{\tau}_{\rm m}) = \frac{\sigma^2_{\rm c,m}}{p(1-q)N} + \frac{\sigma^2_{\rm t,m}}{(1-p)(1-q)N}.$$

Substituting these variances in leads to

$$\mathbb{V}_{\mathcal{S}} = \frac{q}{N}\left(\frac{\sigma^2_{\rm t,f}}{p} + \frac{\sigma^2_{\rm c,f}}{1-p}\right) + \frac{1-q}{N}\left(\frac{\sigma^2_{\rm t,m}}{p} + \frac{\sigma^2_{\rm c,m}}{1-p}\right).$$

Comparing the variances under the two experimental designs we find

$$\mathbb{V}_{\mathcal{C}} - \mathbb{V}_{\mathcal{S}} = q(1-q)(\mu_{\rm c,f} - \mu_{\rm c,m})^2 + q(1-q)(\mu_{\rm t,f} - \mu_{\rm t,m})^2 \geq 0. \tag{2.4}$$

The stratified design leads to a weakly lower variance than the completely randomized design in this setting.

A couple of comments. Here we have equal treatment probabilities in all strata. That need not be optimal, but without it you cannot generally rank the designs. We focused on the case with two strata. The result generalizes to the case with multiple strata. We used stratified sampling. Without that we cannot get the exact result, because it need no longer be possible to get exactly equal treatment probabilities in all strata, but the result would hold approximately. A final comment concerns the use of a large population, with all the strata large. This is important. Suppose all the strata consist of a few units. Then it is possible that the potential outcomes are negatively correlated within a stratum. This can happen in practice, for example, if the strata are litters of puppies. This is not possible if there are a large number of units per stratum, which guarantees that the potential outcomes within a stratum have non-negative correlation.

For the second result of this section we return to the completely randomized design but change the estimator to the adjusted (stratified) estimator $\hat{\tau}^{\rm strat}$, and consider the variance

$$\mathbb{V}^{\rm strat}_{\mathcal{C}} = \mathbb{V}(\hat{\tau}^{\rm strat}|\mathcal{C}).$$

In that case the number of treated and control individuals in both strata, $N_{\rm t,f}$, $N_{\rm t,m}$, $N_{\rm c,f}$, and $N_{\rm c,m}$, and in particular the ratios of treated, $p_{\rm f} = N_{\rm t,f}/N_{\rm f}$ and $p_{\rm m} = N_{\rm t,m}/N_{\rm m}$ are

stochastic, with $p = qp_{\rm f} + (1-q)p_{\rm m}$. Conditional on their values the variance is

$$\mathbb{V}\left(\hat{\tau}^{\rm strat}\,|\mathcal{C}, p_{\rm f}, p_{\rm m}\right) = \frac{q}{N}\left(\frac{\sigma^2_{\rm t,f}}{p_{\rm f}} + \frac{\sigma^2_{\rm c,f}}{1-p_{\rm f}}\right) + \frac{1-q}{N}\left(\frac{\sigma^2_{\rm t,m}}{p_{\rm m}} + \frac{\sigma^2_{\rm c,m}}{1-p_{\rm m}}\right).$$

In large samples, this variance will be approximately the same as the variance of $\hat{\tau}$ (or, which is the same thing, $\hat{\tau}^{\rm strat}$) under the stratified design and exploit the advantages of stratification. In small samples, however, there is now a cost. Consider the special case with homoskedasticity, by stratum and treatments status, $\sigma^2_{\rm c,f} = \sigma^2_{\rm c,m} = \sigma^2_{\rm t,f} = \sigma^2_{\rm t,m}$. In addition suppose $\mu_{\rm c,f} = \mu_{\rm c,m}$ and $\mu_{\rm t,f} = \mu_{\rm t,m}$. The implication is that now all the variances are equal. Denoting this common variance by $\sigma^2$ we have $\sigma^2 = \sigma^2_{\rm c} = \sigma^2_{\rm t} = \sigma^2_{\rm c,f} = \sigma^2_{\rm c,m} = \sigma^2_{\rm t,f} = \sigma^2_{\rm t,m}$. Note that in this case the stratification is pointless: the distributions of the potential outcomes do not differ between the two strata. Nevertheless, the earlier result still holds that the stratification does not hurt the variance: in this case $\mathbb{V}_{\mathcal{S}} = \mathbb{V}_{\mathcal{C}}$. We can write the variance of the two estimators under a completely randomized design in this special case as

$$\mathbb{V}(\hat{\tau}|\mathcal{C}, p_{\rm f}, p_{\rm m}) = \frac{\sigma^2}{N}\frac{1}{p(1-p)},$$

and

$$\mathbb{V}(\hat{\tau}^{\rm strat}|\mathcal{C}, p_{\rm f}, p_{\rm m}) = \frac{\sigma^2}{N}\left(\frac{q}{p_{\rm f}(1-p_{\rm f})} + \frac{1-q}{p_{\rm m}(1-p_{\rm m})}\right).$$

(Note that we condition the variance of $\hat{\tau}$ on $p_{\rm f}$ and $p_{\rm m}$, to facilitate the comparison, although the variance does not actually vary with $p_{\rm f}$ and $p_{\rm m}$.) Because $p = qp_{\rm f} + (1-q)p_{\rm m}$ it follows that in this case

$$\mathbb{V}(\hat{\tau}^{\rm strat}|\mathcal{C}, p_{\rm f}, p_{\rm m}) \geq \mathbb{V}(\hat{\tau}|\mathcal{C}, p_{\rm f}, p_{\rm m}), \tag{2.5}$$

for all $p_{\rm f}$ and $p_{\rm m}$, with equality only if $p_{\rm f} = p_{\rm m} = p$. If $p_{\rm f} \neq p_{\rm m}$, then the variance for the stratification estimator is strictly worse than the variance for the simple difference in means. The take-away is that ex post stratification can increase the variance relative to no stratification, whereas ex ante stratification cannot increase the variance. In other words, design always trumps analysis.

### 2.4.3 The Cost of Stratification for Variance Estimation in Finite Samples

Although there is no cost to stratification in terms of the exact finite sample variance, there is a potential cost in terms of inference. First consider the natural estimator for the variance of $\hat{\tau}$ under the completely randomized design, using the results from the previous chapter:

$$\hat{\mathbb{V}}_{\mathcal{C}} = \frac{s_{\mathrm{c}}^2}{N_{\mathrm{c}}} + \frac{s_{\mathrm{t}}^2}{N_{\mathrm{t}}}.$$

For a stratified randomized experiment the natural variance estimator, taking into account the stratification, is:

$$\hat{\mathbb{V}}_{\mathcal{S}} = q^2 \left( \frac{s_{\mathrm{c,f}}^2}{N_{\mathrm{fc}}} + \frac{s_{\mathrm{t,f}}^2}{N_{\mathrm{ft}}} \right) + (1-q)^2 \left( \frac{s_{\mathrm{c,m}}^2}{N_{\mathrm{mc}}} + \frac{s_{\mathrm{t,m}}^2}{N_{\mathrm{mt}}} \right).$$

Under the stratified sampling scheme where $N_{\mathrm{fc}}/N_{\mathrm{c}} = N_{\mathrm{ft}}/N_{\mathrm{t}} = q$ and $N_{\mathrm{mc}}/N_{\mathrm{c}} = N_{\mathrm{mt}}/N_{\mathrm{t}} = 1-q$, we can write this as

$$\hat{\mathbb{V}}_{\mathcal{S}} = \frac{q s_{\mathrm{c,f}}^2 + (1-q) s_{\mathrm{c,m}}^2}{N_{\mathrm{c}}} + \frac{q s_{\mathrm{t,f}}^2 + (1-q) s_{\mathrm{t,m}}^2}{N_{\mathrm{t}}}.$$

Both estimators, $\hat{\mathbb{V}}_{\mathcal{C}}$ and $\hat{\mathbb{V}}_{\mathcal{S}}$ are unbiased for their respective true variances:

$$\mathbb{E}\left[\hat{\mathbb{V}}_{\mathcal{C}}\right] = \mathbb{V}_{\mathcal{C}} \qquad \text{and} \quad \mathbb{E}\left[\hat{\mathbb{V}}_{\mathcal{S}}\right] = \mathbb{V}_{\mathcal{S}},$$

so that by the results in the previous section $\mathbb{E}\left[\hat{\mathbb{V}}_{\mathcal{C}}\right] \geq \mathbb{E}\left[\hat{\mathbb{V}}_{\mathcal{S}}\right]$. However, there is in general no stochastic dominance, and it may be the case that

$$\mathbb{V}\left(\hat{\mathbb{V}}_{\mathcal{S}}\right) > \mathbb{V}\left(\hat{\mathbb{V}}_{\mathcal{C}}\right).$$

This is most easily seen in the case where the stratification is completely irrelevant and $\mu_{\mathrm{c,f}} = \mu_{\mathrm{c,m}}$, $\mu_{\mathrm{t,f}} = \mu_{\mathrm{t,m}}$, $\sigma_{\mathrm{c,f}}^2 = \sigma_{\mathrm{c,m}}^2$, and $\sigma_{\mathrm{t,f}}^2 = \sigma_{\mathrm{t,m}}^2$. In that case

$$\mathbb{E}\left[s_{\mathrm{c,f}}^2\right] = \mathbb{E}\left[s_{\mathrm{c,m}}^2\right] = \mathbb{E}\left[s_{\mathrm{c}}^2\right] = \sigma_{\mathrm{c}}^2,$$

but

$$\mathbb{V}\left(s_{\mathrm{c,f}}^2\right) > \mathbb{V}\left(s_{\mathrm{c}}^2\right), \qquad \text{and} \quad \mathbb{V}\left(s_{\mathrm{c,m}}^2\right) > \mathbb{V}\left(s_{\mathrm{c}}^2\right),$$

so that

$$\mathbb{V}\left(qs^2_{\rm c,f}+(1-q)s^2_{\rm c,m}\right) > \mathbb{V}(s^2_{\rm c}), \qquad \text{and} \quad \mathbb{V}\left(qs^2_{\rm t,f}+(1-q)s^2_{\rm t,m}\right) > \mathbb{V}(s^2_{\rm t})$$

and therefore,

$$\mathbb{V}\left(\hat{\mathbb{V}}_{\mathcal{S}}\right) > \mathbb{V}\left(\hat{\mathbb{V}}_{\mathcal{C}}\right).$$

The cost of the stratification is that the natural estimator for the variance under stratification is potentially noiser (and definitely so if the stratification is irrelevant), and that therefore by using it one would lose statistical power.

Of course one solution is to stratify (and therefore gain the variance benefits), and then use the variance based on complete randomization. That leads to conservative confidence intervals, but with a better coverage than the confidence intervals would have under complete randomization.

## 2.5 Re-randomization

In the previous sections we discussed systematic alternatives to a completely randomized design. Another alternative is to inspect balance in pre-treatment variables after the randomization. If the balance is adequate, by some measure, the researcher would then proceed and expose the units to the assigned treatment. If the balance is not adequate, the researcher would go back and re-randomize the assignment until the balance was adequate. In this section we discuss such re-randomization schemes.

### 2.5.1 Re-randomization: Some Basic Points

The first point we make is that implicitly some of the designs we have considered can be interpreted as re-randomization schemes. This perspective is useful because it allows us to see that at least in some simple cases whether and how inference needs to be adjusted to take into account the re-randomization.

Suppose we assign units to the treatment by tossing a fair coin. With $N$ units in the population (say with $N$ even), we expect to see $N/2$ units in the treatment group.

However, the actual number of treated units will often be different from $N/2$. The variance of the standard estimator $\hat{\tau} = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$ is difficult to calculate. Conditional on $N_{\mathrm{t}}$ and $N_{\mathrm{c}}$ it is straightforward, and equal to

$$\mathbb{V}(\hat{\tau}|N_{\mathrm{t}}, N_{\mathrm{c}}) = \frac{S_{\mathrm{t}}^2}{N_{\mathrm{t}}} + \frac{S_{\mathrm{c}}^2}{N_{\mathrm{c}}} - \frac{S_{\mathrm{c,t}}^2}{N}.$$

The unconditional variance

$$\mathbb{V}(\hat{\tau}) = \mathbb{E}\left[\frac{S_{\mathrm{t}}^2}{N_{\mathrm{t}}} + \frac{S_{\mathrm{c}}^2}{N_{\mathrm{c}}} - \frac{S_{\mathrm{c,t}}^2}{N}\right],$$

is more difficult to calculate, and requires defining the estimator in the case where there are only treated or only control units. Even if we condition on that contingency not occurring, the exact variance of this Bernoulli trial will be strictly larger than the variance for the experiment where we fix the number of treated and control units at $N_{\mathrm{t}} = N_{\mathrm{c}} = N/2$.

Now consider checking balance in the treatment groups and rejecting randomizations where the number of treated units differs from $N/2$. If we re-randomize until the number of treated units is equal to $N/2$, the variance of the estimator taking into account the re-randomization is

$$\mathbb{V}(\hat{\tau}|N_{\mathrm{t}} = N_{\mathrm{c}} = N/2) = \frac{S_{\mathrm{t}}^2}{N/2} + \frac{S_{\mathrm{c}}^2}{N/2} - \frac{S_{\mathrm{c,t}}^2}{N}.$$

This is different from the unconditional variance for the original experiment.

You can also see stratification as a special case of re-randomization. Suppose we have $N$ units in our sample, $N_{\mathrm{f}}$ women and $N_{\mathrm{m}}$ men. We can do a completely randomized design where we draw $M = pN$ units at random from this sample and assign those to the treatment group. Now suppose that after the randomization we consider the difference in the share of women and men in the treatment and control group,

$$\Delta_X = \overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}},$$

where

$$\overline{X}_{\mathrm{t}} = \frac{1}{N_{\mathrm{t}}} \sum_{i:W_i=1} \mathbf{1}_{X_i=\mathrm{f}}, \qquad \text{and} \quad \overline{X}_{\mathrm{c}} = \frac{1}{N_{\mathrm{c}}} \sum_{i:W_i=0} \mathbf{1}_{X_i=\mathrm{f}}.$$

Alternatively we can also look at the t-statistic

$$t_X = \frac{\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}}}{S_X^2},$$

where

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{1}_{X_i=\mathrm{f}} - \overline{X}\right)^2, \qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{X_i=\mathrm{f}},$$

but in the current setting that makes no difference.

Now suppose that we re-randomize if $|\Delta_X| > 0$ (or equivalently, $|t_X| > 0$). This presumes that $pN_{\mathrm{f}}$ and $pN_{\mathrm{m}}$ are integers, if that is not the case we can simply focus on checking whether $\Delta_X$ is as small as possible. We can keep re-randomizing until $\Delta_X$ is exactly equal to zero (or its minimum feasible value). In that case we end up with a randomization design that is exactly equivalent to a completely stratified design. This discussion shows that in simple cases re-randomization is simply an indirect way of getting to a well-known design. It also shows that re-randomization does affect the variance of the natural estimator (the difference in means) for the average treatment effect, and that in general the effects of the re-randomization needs to be taken into account.

These points hold more generally. The problem is that unless the exact algorithm for re-randomization is precisely defined, it is impossible to infer the adjustment, and even if it is precisely defined, it may be difficult to develop unbiased estimators for the average treatment effect. Nevertheless, because re-randomization generally improves the variance of the estimators, it is often preferable over simply proceeding with a randomized assignment that is characterized by imbalances in important pre-treatment variables.

### 2.5.2 The Benefits from Re-randomization

Now let us consider a particular way to implement re-randomization, based on Morgan and Rubin [2012]. There are $K$ covariates or features. Given an assignment vector $\mathbf{W}$, with marginal assignment probability $p$, there is a function $\psi\{0,1\}^N \times \mathbb{X}^N \mapsto \{0,1\}$ that measures the imbalance of the covariates. We focus here on

$$\psi(\mathbf{W},\mathbf{X}) = Np(1-p)\left(\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}}\right)^{\top}\Sigma_X^{-1}\left(\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}}\right).$$

Given the randomization this statistic has approximately a Chi-squared distribution with degrees of freedom equal to $K$.

Now we can do the re-randomization in a number of different ways. We can do randomize $L$ times, and choose the randomization vector that minimizes $\psi(\mathbf{W}_l, \mathbf{X})$ over the $L$ assignment vectors $\mathbf{W}_l$. Alternatively we can choose a threshold $a$ and re-randomize until $\psi(\mathbf{W}, \mathbf{X}) \leq a$. Morgan and Rubin [2012] consider the latter, and show that the covariance matrix of $\mathbf{X}_\mathrm{t} - \mathbf{X}_\mathrm{c}$ conditional on $\psi(\mathbf{W}, \mathbf{X}) \leq a$ is $r_a$ times the unconditional covariance matrix, with $r_a$ equal to the ratio of the probabilities of Chi-squared random variables being less than $a$:

$$r_a = \frac{\text{pr}\left(\mathcal{X}^2_{K+2} \leq a\right)}{\text{pr}\left(\mathcal{X}^2_K \leq a\right)}.$$

Suppose the potential outcomes follow a linear model, with

$$Y_i(w)|X_i \sim \mathcal{N}(\alpha + X_i^\top \beta, \sigma^2),$$

with $R^2$ equal to $\beta^\top \Sigma \beta / (\beta^\top \Sigma \beta + \sigma^2)$, then the variance for the difference-in-means estimator $\hat{\tau} = \overline{Y}_\mathrm{t} - \overline{Y}_\mathrm{c}$ after re-randomization is

$$\mathbb{V}(\hat{\tau}) = \frac{1}{Np(1-p)} r_a \beta^\top \Sigma \beta + \frac{\sigma^2(1-R^2)}{Np(1-p)}.$$

### 2.5.3 Concerns with Re-randomization

The main concern with re-randomization is that one needs to ensure that there remains sufficient randomization in the set of acceptable assignment vectors. Suppose that instead of accepting any assignment vector with $\psi(\mathbf{W}, \mathbf{X}) \leq a$, we simply look for the assignment vector that minimizes $\psi(\mathbf{W}, \mathbf{X})$. The structure of $\psi(\mathbf{W}, \mathbf{X})$ implies that for two vectors $\mathbf{W}$ and $\mathbf{W}'$, if $W'_i = 1 - W_i$ for all $i$, we have $\psi(\mathbf{W}, \mathbf{X}) = \psi(\mathbf{W}', \mathbf{X})$, so that there will always be at least two assignment vectors that minimize this objective function. However, if we limit the set of acceptable assignments to this pair, there is not much randomness left to base inference on. For example, if we have a single covariate, $X_i$, with $X_i \in [0, 1]$ for $i = 1, \ldots, N-1$, and $X_N = 1000$, re-randomizing to minimize $\psi(\mathbf{W}, \mathbf{X})$ would lead to a very particular assignment vector, with all the smallest $N/2 - 1$ units together with $X_N$ in the same treatment group.

## 2.6 Conclusion

## Notes

Bruhn and McKenzie [2009], Morgan and Rubin [2012, 2015]

Although it is well known that stratification on covariates is beneficial if based on pre-treatment variables that are sufficiently strongly correlated with the potential outcomes, there appears to be confusion concerning the benefits in small samples if this correlation is weak. Bruhn and McKenzie [2009] document this in a survey of researchers in development economics, but the confusion is also apparent in the statistics literature. For example, Snedecor and Cochran (1989, page 101) write:

> "If the criterion has no correlation with the response variable, a small loss in accuracy results from the pairing due to the adjustment for degrees of freedom. A substantial loss may even occur if the criterion is badly chosen so that member of a pair are negatively correlated."

Box, Hunter and Hunter (2005, page 93) also suggest that there is a tradeoff in terms of accuracy or variance in the decision to stratify, writing:

> "Thus you would gain from the paired design only if the reduction in variance from pairing outweighed the effect of the decrease in the number of degrees of freedom of the $t$ distribution."

This is somewhat counterintuitive: if one stratifies on a covariate that is independent of all other variables, then stratification is obviously equivalent to complete randomization. In the current chapter we argue that this intuition is correct and that in fact there is no tradeoff.

# Chapter 3

# Power Analyses

## 3.1 Introduction

Prior to conducting a randomized experiment a researcher has a number of important decisions to make regarding the design. One of these decisions involves the sample size. There is no point in doing a randomized experiment that has little or not chance of finding reasonably sized effects. This is particularly important because there is a tendency to (mis-)interpret findings of no statistically significant effects as findings of zero effects. Power Analyses involve the calculation of sample sizes that have a reasonable chance of finding effects of substantively interesting size. Such analyses involve the choice of a number of parameters, including the effects that are viewed as being of a reasonable size. Here we discuss power analyses in two leading cases.

## 3.2 Testing a Mean against Zero

Suppose we have a random sample $X_1, \ldots, X_N$ from a Normal distribution with mean $\mu$ and variance $\sigma^2$. We wish to test the null hypothesis

$$H_0: \quad \mu = 0, \qquad \text{against the alternative} \quad H_a: \quad \mu \neq 0.$$

We want the probability of a type I error to be less than $\alpha$ (in other words, we want the size of the test to be $\alpha$). Often, going back to Fisher [1937], we use $\alpha = 0.05$, or $\alpha = 0.10$. ("It is usual and convenient for experimenters to take 5 percent. as a standard level of

significance Fisher [1937], p. 13) We want the test to have power $\beta$ for an alternative where $\mu = \mu_0$, for some pre-specified values of $\mu_0 \neq 0$. In other words, the probability of rejecting the null hypothesis, when the null is false with $\mu = \mu_0$, should be $\beta$. A common choice for $\beta$ is 0.8.

We base the statistical test on the t-statistic

$$T = \frac{\overline{X}}{\sqrt{S_X^2/N}},$$

where $\overline{X}$ and $S_X^2$ are the sample average and sample variance respectively:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \text{and} \quad S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2.$$

If $N$ is very small, the chance of rejecting the null when $\mu = \mu_0$ is small, and if $N$ is very large this probability will be close to one.

The question is now, given $\alpha$, $\beta$, $\sigma^2$, and $\mu_0$, what is the minimum sample size such that the rejection probability is at least $\beta$? If the size of the test is $\alpha$, we will reject the null hypothesis if

$$|T| \geq \Phi^{-1}\left(1 - \alpha/2\right),$$

where $\Phi^{-1}(a)$ is the inverse of, $\Phi(x)$, the cumulative distribution function for the Normal distribution,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) dx.$$

So, with $\alpha = 0.05$, so that $1-\alpha/2 = 0.975$, we get the familiar critical value $\Phi^{-1}\left(1 - 0.05/2\right) = 1.96$. If we choose $\alpha = 0.10$, the critical value would be $\Phi^{-1}\left(1 - 0.10/2\right) = 1.645$

Now let us look at the probability of rejecting the null hypothesis if the null hypothesis is in fact false, and the true mean is $\mu_0 \neq 0$. For ease of exposition let us assume that $\mu_0 > 0$. In that case most of my rejections will occur when $T$ is larger than $\Phi^{-1}\left(1 - \alpha/2\right)$, rather than because $T$ is smaller than $-\Phi^{-1}\left(1 - \alpha/2\right)$. For the power calculations we can essentially ignore the probability of rejecting for a low value of $T$, so

$$\text{pr}\left(|T| > \Phi^{-1}\left(1 - \alpha/2\right)\right) = \text{pr}\left(T > \Phi^{-1}\left(1 - \alpha/2\right)\right) + \text{pr}\left(T < -\Phi^{-1}\left(1 - \alpha/2\right)\right)$$

$$
\begin{aligned}
&\approx \text{pr}\left(T > \Phi^{-1}\left(1-\alpha/2\right)\right) \\
&= \text{pr}\left(\frac{\overline{X}}{\sqrt{S_X^2/N}} > \Phi^{-1}\left(1-\alpha/2\right)\right) \\
&\approx \text{pr}\left(\frac{\overline{X}}{\sqrt{\sigma_X^2/N}} > \Phi^{-1}\left(1-\alpha/2\right)\right) \\
&= \text{pr}\left(\frac{\overline{X}-\mu_0}{\sqrt{\sigma_X^2/N}} > \Phi^{-1}\left(1-\alpha/2\right) - \frac{\mu_0}{\sqrt{\sigma_X^2/N}}\right) \\
&= \text{pr}\left(-\frac{\overline{X}-\mu_0}{\sqrt{\sigma_X^2/N}} < -\Phi^{-1}\left(1-\alpha/2\right) + \frac{\mu_0}{\sqrt{\sigma_X^2/N}}\right).
\end{aligned}
$$

Note that under the alternative hypothesis with $\mu = \mu_0$,

$$
-\frac{\overline{X}-\mu_0}{\sqrt{\sigma_X^2/N}} \sim \mathcal{N}(0,1),
$$

and so

$$
\begin{aligned}
&\text{pr}\left(|T| > \Phi^{-1}\left(1-\alpha/2\right)\right) \\
&\approx \Phi\left(-\Phi^{-1}\left(1-\alpha/2\right) + \frac{\mu_0}{\sqrt{\sigma_X^2/N}}\right).
\end{aligned}
$$

We want to find the sample size $N$ such that this probability is equal to, or exceeds, $\beta$. The probability is equal to $\beta$ if

$$
\beta = \Phi\left(-\Phi^{-1}\left(1-\alpha/2\right) + \frac{\mu_0}{\sqrt{\sigma_X^2/N}}\right),
$$

leading to

$$
\Phi^{-1}\left(\beta\right) = -\Phi^{-1}\left(1-\alpha/2\right) + \frac{\mu_0}{\sqrt{\sigma_X^2/N}},
$$

and thus

$$
N = \left(\frac{\Phi^{-1}\left(\beta\right) + \Phi^{-1}\left(1-\alpha/2\right)}{\mu_0/\sigma_X}\right)^2. \tag{3.1}
$$

For example, if $\alpha = 0.05$ (we test at the 5% level), $\mu_0/\sigma_X = 0.1$ (the population mean is 10% of a standard deviation), and we wish to detect such a difference from zero with

probability 0.8 (type II error should be less than $1 - \beta = 0.2$). Then we need a sample size of

$$N = \left( \frac{\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2)}{\mu_0/\sigma_X} \right)^2 = \left( \frac{\Phi^{-1}(0.8) + \Phi^{-1}(0.975)}{0.1} \right)^2 = 784.89,$$

so the minimum sample size is 785. Suppose we want the probability of a type I error to be less than $\alpha = 0.1$, then we need

$$N = \left( \frac{\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2)}{\mu_0/\sigma_X} \right)^2 = \left( \frac{\Phi^{-1}(0.8) + \Phi^{-1}(0.95)}{0.1} \right)^2 = 618.26,$$

so the minimum sample size is now 619.

## 3.3 Testing a Difference of Means with Unequal Sample Sizes and Unequal Variances

Now let us look at the case of a randomized experiment where we want to detect a causal effect. We have a random sample from a large population, $Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}}$, and a binary treatment indicator $W_1, \ldots, W_N$, with sample size $N$. Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes, so that the realized outcome is

$$Y_i^{\text{obs}} = Y_i(W_i).$$

We are interested in testing the hypothesis that the population average treatment effect $\tau^{\text{sp}} = \mathbb{E}[Y_i(1) - Y_i(0)]$ is zero:

$$H_0: \quad \mathbb{E}[Y_i(1) - Y_i(0)] = 0,$$

against the alternative that it differs from zero:

$$H_a: \quad \mathbb{E}[Y_i(1) - Y_i(0)] \neq 0.$$

The size of the test is again $\alpha$, and we want power $\beta$, against an alternative that the average treatment effect is $\tau^{\text{sp}} = \tau_0$ for some $\tau_0 \neq 0$. Let $\gamma = \sum_i W_i/N$ be the proportion of treated units. We look for the minimum sample size $N = N_{\text{c}} + N_{\text{t}} = N_{\text{c}}(1 + \gamma)$, as a function of $\alpha$, $\beta$, $\tau$, $\sigma^2$, and $\gamma$. We assume for simplicity that the variance is not affected by

the treatment. Although this is often not realistic, in practice allowing different variances does not affect the power calculations much, and assuming homoskedasticity reduces the number of parameters that need to be fixed in advance: it is rare that we would a priori have credible information about the magnitude of the heteroskedasticity.

Let $\overline{Y}_{\rm t}^{\rm obs}$ and $\overline{Y}_{\rm c}^{\rm obs}$ be the average outcomes in the two subsamples. We look at the T-statistic

$$T = \frac{\overline{Y}_{\rm t} - \overline{Y}_{\rm c}}{\sqrt{S^2/N_{\rm t} + S^2/N_{\rm c}}}.$$

Again, for the purposes of the power calculations we use the t-statistic based on homoskedasticity, although with the actual data we may wish to allow for heteroskedasticity. Note also that this is also the t-statistic for the test that the slope coefficient is zero in the linear regression

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

We reject the null hypothesis of no difference if $|T|$ exceeds $\Phi^{-1}\left(1-\alpha/2\right)$, and want the rejection probability to be at least $\beta$, given that the alternative hypothesis is true with $\tau^{\rm sp} = \tau_0$. Under that scenario

$$\frac{\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs} - \tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}} \sim \mathcal{N}(0,1),$$

so approximately

$$-\left(T - \frac{\tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}}\right) \sim \mathcal{N}(0,1).$$

Now, as before, focusing on the case where $\tau_0 > 0$,

$$\begin{aligned}
&{\rm pr}\left(|T| > \Phi^{-1}\left(1-\alpha/2\right)\right) \approx {\rm pr}\left(T > \Phi^{-1}\left(1-\alpha/2\right)\right) \\
&\quad = {\rm pr}\left(T - \frac{\tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}} > \Phi^{-1}\left(1-\alpha/2\right) - \frac{\tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}}\right) \\
&\quad = {\rm pr}\left(-\left(T - \frac{\tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}}\right) < -\Phi^{-1}\left(1-\alpha/2\right) + \frac{\tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}}\right) \\
&\quad = \Phi\left(-\Phi^{-1}\left(1-\alpha/2\right) + \frac{\tau_0}{\sqrt{\sigma^2/N_{\rm t} + \sigma^2/N_{\rm c}}}\right).
\end{aligned}$$

This rejection probability is equal to $\beta$ if

$$\begin{aligned}\Phi^{-1}(\beta) &= -\Phi^{-1}(1-\alpha/2) + \frac{\tau_0}{\sqrt{\sigma^2/N_{\mathrm{t}} + \sigma^2/N_{\mathrm{c}}}},\\ &= -\Phi^{-1}(1-\alpha/2) + \frac{\tau_0}{\sqrt{\sigma^2/(N\gamma) + \sigma^2/(N(1-\gamma)}},\\ &= -\Phi^{-1}(1-\alpha/2) + \frac{\tau_0\sqrt{N}\sqrt{(\gamma(1-\gamma)}}{\sigma},\end{aligned}$$

leading to an overall sample size

$$N = \frac{\left\{\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2)\right\}^2}{(\tau_0^2/\sigma^2)\gamma(1-\gamma)}. \tag{3.2}$$

For the two subsample sizes we have $N_{\mathrm{t}} = N\gamma$, and $N_{\mathrm{c}} = N(1-\gamma)$.

For example, suppose we chose $\gamma = 0.5$ (equal sample sizes), $\alpha = 0.05$ (test at 0.05 level), $\tau_0/\sigma = 0.1$ (looking for effect of 0.1 standard deviation), $\beta = 0.8$, power of 0.8. Then

$$\begin{aligned}N &= \frac{\left(\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2)\right)^2}{(\tau_0^2/\sigma^2)\gamma(1-\gamma)}\\ &= \frac{\left(\Phi^{-1}(0.8) + \Phi^{-1}(0.975)\right)^2}{0.1^2 0.5^2} = 3,139.6,\end{aligned}$$

so that the minimum sample size is 3,140, with 1,570 treated and 1,570 controls.

## Notes

Cohen [1977] on power analyses, other references.

Page 54-55 give values for $N_{\rm c} = N_{\rm t}$ for this case under $\gamma = 0.5$ (equal sample sizes), for different values of $\beta$ (power), $\alpha$ ($a_2$ in Cohen's notation), and different values for $\tau_0/\sigma$ ($d$ in the notation of Cohen's text).