MATH 348 Week 7: Operating Characteristics of Stratified and Matched RCTs

Prof. Kennedy-Shaffer

Feb. 26, 2024

1 Recap: Results for RCTs

In a standard RCT where we are estimating a difference in means (or proportions), we have:

$$\theta = E[Y \text{ if given treatment}] - E[Y \text{ if given control}]$$

$$\hat{\theta} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1,i} - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0,i}$$

$$Var(\hat{\theta}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

$$Power = 1 - \Phi\left(\Phi^{-1}(1-\alpha/2) - \frac{\theta}{\sqrt{Var(\hat{\theta})}}\right)$$

To improve power, we have a few options:

- Increase α (lower the bar to clear for significance)
- Increase θ (larger effect: requires changing outcome, intervention, and/or population)
- Reduce $Var(\hat{\theta})$:
  - Increase sample size
  - Reduce variance of each outcome: σ_1^2, σ_0^2

We could accomplish the last one by changing the outcome and/or population. We can also effectively change it by stratification, matching, or covariate adjustment. Details of the first two are described here; we'll talk about adjustment in class on Wednesday.

2 Stratification

Like in the survey sampling case, the principle of stratification for RCTs is:

(1) Find S strata (subsets of the population defined by covariates or groups of covariates) that are more similar to each other in their outcome values than the population as a whole.

(2) Within each of these strata, do the randomization, run the study as normal, and get a stratum-specific estimate $\hat{\theta}_s$ for each $s = 1, \ldots, S$. In RCTs, this means we will randomize to intervention and control *within* each stratum in the chosen randomization ratio. This ensures balance between the intervention and control arms in the covariates that define the strata.

(3) Combine these estimates together using a weighted average to get an overall estimate:

$$\hat{\theta}_{strat} = \sum_{s=1}^{S} w_s \hat{\theta}_s \text{ for weights } w_1, w_2, \ldots, w_S$$

The biggest difference in how we look at the operating characteristics between stratified RCTs and stratified surveys is that for stratified RCTs, we often assume that the effect of treatment is the same in each stratum. This is called a **homogeneous treatment effect** (if it's not the same, the effect is called **heterogeneous**). Note that the average outcomes themselves are still likely to be different in different strata (otherwise the variance won't go down), but we assume the differences are not. We'll come back to this assumption later on.

2.1 Usual Assumptions

For now, assume that every stratum has the same true treatment effect: $\theta_s = \theta$ for all s. Assume that within stratum s, we do a standard RCT and estimate the difference on those outcomes with our usual difference in means estimator.

2.2 Bias

We want an unbiased estimator, so we want $E[\hat{\theta}_{strat}] = \theta$:

$$\begin{aligned}
E[\hat{\theta}_{strat}] &= E\left[\sum_{s=1}^{S} w_s \hat{\theta}_s\right] = \sum_{s=1}^{S} w_s E[\hat{\theta}_s] \text{ as long as the weights are chosen prior to observing the data} \\
&= \sum_{s=1}^{S} w_s \theta \text{ as long as each within-stratum estimates is unbiased} \\
&= \theta \cdot \sum_{s=1}^{S} w_s
\end{aligned}$$

So the estimator is unbiased as long as the weights sum to one: $\sum_{s=1}^{S} w_s = 1$.

2.3 Variance

We need to add another index for the stratum. Let $Y_{k,s,i}$ be the outcome of individual i within stratum s in treatment arm k ($k = 0$ for control and $k = 1$ for intervention). Let $n_{k,s}$ be the number of people assigned to treatment arm k in stratum s.

$$\begin{aligned}
Var(\hat{\theta}_{strat}) &= Var\left(\sum_{s=1}^{S} w_s \hat{\theta}_s\right) = \sum_{s=1}^{S} w_s^2 Var(\hat{\theta}_s) \text{ since each stratum-specific result is independent} \\
&= \sum_{s=1}^{S} w_s^2 Var\left(\frac{1}{n_{1,s}} \sum_{i=1}^{n_{1,s}} Y_{1,s,i} - \frac{1}{n_{0,s}} \sum_{i=1}^{n_{0,s}} Y_{0,s,i}\right) \\
&= \sum_{s=1}^{S} w_s^2 \left(\frac{\sigma_{1,s}^2}{n_{1,s}} + \frac{\sigma_{0,s}^2}{n_{0,s}}\right)
\end{aligned}$$

Notice that if $S = 1$, then this simplifies to our result from a usual RCT.

These $\sigma_{0,s}^2$ and $\sigma_{1,s}^2$ parameters are *within-stratum* variance parameters. In other words, they describe the variability of outcomes from individuals in the same stratum. So the more predictive of the outcome the stratifying factors are, the lower these values will be. They are related to the overall variance through the Law of Total Variance.

As a general rule (from multivariable calculus), this variance will be minimized if:

$$w_s \propto \left(\frac{\sigma_{1,s}^2}{n_{1,s}} + \frac{\sigma_{0,s}^2}{n_{0,s}} \right)^{-1}$$

This is the **inverse-variance weighting** (which we also saw when combining multiple studies). Since we don't know the $\sigma_{k,s}^2$ values in advance, we often ignore those numerators and weight proportional to the sample size within the stratum: $w_s \propto n_{1,s} + n_{0,s}$. This is called a **self-weighting** stratified design. This is computationally convenient as well.

2.4 Wrong Assumptions

So far, we haven't said anything about how to select the sample size for each stratum $n_{1,s} + n_{0,s}$, only how to pick the weights once we know that sample size. A natural choice for the sample size is to make it proportional to that stratum's proportion in the population of interest. So if we are going to have a total sample size of the trial of 1,000 people, and stratum 1 represents 10% of the population of interest, let's make $n_{1,1} + n_{0,1} = 0.10 \cdot 1000 = 100$.

The advantage of this is it makes the overall estimator more interpretable if there are heterogeneous treatment effects. If the stratum-specific effects (the θ_s values) are not all equal, then this self-weighting stratified study will be unbiased for the population-weighted average of the treatment effects.

If we don't use weights that represent the population proportion (for example, if we have some variance estimates and try to improve our variance more), we run the risk of estimating an unclear weighted average of the treatment effects, which won't represent the effect of the treatment for any individual or for the population as a whole.

These issues are even more pronounced for ratio estimators, especially odds ratios, which are "not collapsible", and can even result in estimands that are outside of the range of stratum-specific treatment effects. This can cause a big generalizability issue.

2.5 Implementation and Sample Size Calculation

The usual approach would be:

(a) Decide on the stratification factors: we want a balance between having strata where individuals are similar to others in their same stratum and not having strata that are too small or hard to find people in.

(b) Estimate the proportion of the total population in each of the strata and use that to set the weights and to determine how much of the total sample size should be drawn from each stratum.

(c) Using a prior guess for the within-stratum variances $\sigma_{1,s}^2$ and $\sigma_{0,s}^2$ (often for continuous outcomes, we assume that these will be equal across strata and across treatment arms, so we get one variance parameter that is smaller than the variance in the unstratified design) and the other usual parameters for the sample size/power calculation, calculate the power or sample size you need given the other one.

2.6 Example

Suppose we are running an agricultural experiment with a new fertilizer to see its effect on plant height after one growing season. In the population as a whole, we estimate the average height under control would be 3 meters, with a variance of 0.5 m^2.

We have 12 plots available, will use a significance level of $\alpha = 0.05$, and we want the minimum detectable effect to be 0.75 m. If we put this into a power calculator, we get a power of 38.3%. This is very bad.

```
> power.t.test(n=6, delta=.75, sd=sqrt(.5), sig.level=0.05, power=NULL, type="two.sample", alternative="two.sided")

     Two-sample t test power calculation

              n = 6
          delta = 0.75
             sd = 0.7071068
      sig.level = 0.05
          power = 0.382661
    alternative = two.sided

NOTE: n is number in *each* group
```

We have 6 plots that are in the sum ($s = 1$) and 6 that are in the shade ($s = 2$), so decide to stratify on that since we expect it to be a big predictor of growth. We assume in advance that plants in the sun under control would grow to an average of 3.5m and those in the shade to 2.5m on average. We also assume that the variance in each group under control or treatment would be 0.25 m^2. (Note that this satisfies the law of total expectation and the law of total variance: marginally, $E[Y] = 3$ and $Var(Y) = 0.5$; conditionally, $Var(Y|S) = 0.25$, and $E[Y|S = 1] = 3.5$, $E[Y|S = 2] = 2.5$, $P[S = 1] = 0.5$, and $P[S = 2] = 0.5$).

We can go through the calculations above to get the power formula. Or we can find the weighted average within-stratum variance for a single outcome and plug that in as the variance into any of our power/sample size calculators. In this case, the within-stratum variance is always 0.25, so we put that in as our new variance parameter (or SD=0.5) and run the power calculation. Now we have a power of 65%.

```
> power.t.test(n=6, delta=.75, sd=sqrt(.25), sig.level=0.05, power=NULL, type="two.sample", alternative="two.sided")

     Two-sample t test power calculation

              n = 6
          delta = 0.75
             sd = 0.5
      sig.level = 0.05
          power = 0.6495744
    alternative = two.sided

NOTE: n is number in *each* group
```

One other note: Technically, we should adjust the t-distribution since we lost one degree of freedom. For large enough samples, this makes little difference. For small samples (and 12 probably does count), we probably have slightly lower power than we estimate. This is one reason we want to be careful about having lots of strata with a small sample size.

3 Matching

Matching is essentially an extreme form of stratification, where each stratum only has one (or very few) observations in each treatment arm. The idea is to pair up all of our study units with the one that we think will be closest to it in its outcome (usually, this means most similar across various covariates). Then we randomize one unit within each pair to the intervention and one to control (this can also be done by getting groups of 3 and randomizing two to intervention, one to control, or vice versa, etc.)

The difference in means between the two arms will then be equal to the mean difference within each pair. Let $j = 1, \ldots, n$ index the n pairs of observations, so Y_{1j} is the outcome of the treated unit in pair j and Y_{0j} is the outcome of the control unit in pair j. Then:

$$\begin{aligned}\hat{\theta} &= \frac{1}{n}\sum_{j=1}^{n} Y_{1j} - \frac{1}{n}\sum_{j=1}^{n} Y_{0j} \\ &= \frac{1}{n}\sum_{j=1}^{n}(Y_{1j} - Y_{0j}) = \frac{1}{n}\sum_{j=1}^{n} D_j,\end{aligned}$$

where $D_j = Y_{1j} - Y_{0j}$ is the difference in outcomes within pair j. Each D_j is unbiased for the treatment effect θ (assuming the treatment effect is homogeneous). And

$$\begin{aligned}Var(\hat{\theta}) &= Var\left(\frac{1}{n}\sum_{j=1}^{n} D_j\right) = \frac{1}{n^2}\sum_{j=1}^{n} Var(D_j) \\ &= \frac{\sigma_D^2}{n},\end{aligned}$$

where $\sigma_D^2 = Var(D_j)$ is the (assumed-constant) variance of each within-pair **difference**. The more similar the units within each pair are, the smaller σ_D^2 will be.

The test of the null hypothesis that $\theta = 0$ is then accomplished by testing whether the expected value of the D_j values is different from 0. So this is now a one-sample test, where the variance of each outcome is σ_D^2. Using the one-sample test options in our power/sample size calculator, then we can do this.

3.1 Example

In the same example as above, suppose we match the plots into pairs based on sunlight levels, natural water amount, soil quality, etc., and randomize within each pair. Suppose we estimate that the variance of plants within a pair will be $\sigma_D^2 = 0.2$. Then we calculate a power of 90% for our MDE of 0.75m.

```
> power.t.test(n=6, delta=.75, sd=sqrt(.2), sig.level=0.05, power=NULL, type="one.sample", alternative="two.sided")

     One-sample t test power calculation

              n = 6
          delta = 0.75
             sd = 0.4472136
      sig.level = 0.05
          power = 0.9021397
    alternative = two.sided
```

Note that using the one-sample test takes into account our reduced degrees of freedom, so this is the correct calculation.