Chapter 9 Stratified Randomized Experiments

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

August 19, 2021

Introduction

In stratified randomized experiments (SRE), units are stratified (or grouped or blocked) according to the values of (a function of) the covariates.

Within the strata, independent completely randomized experiments (CRE) are conducted, but possibly with different relative sizes of treatment and control groups.

Part of the motivation is interest in such experiments *per sé*.

Another, reason is that observational studies can be viewed in some way as analyzing the data as if they arose from hypothetical SRE.

Understanding these methods in the context of randomized experiments will aid their interpretation and implementation in observational studies.

Introduction

First we describe the data that will be used to illustrate the concepts. These data are from a randomized experiment designed to evaluate the effect of class size on academic achievement, known as Project Star.

Then we discuss the general structure of SRE.

We will discuss inferences for the four approaches; the FEP approach, the Neyman approach, the regression approach and the model-based imputation approach.

Then we discuss design issues, and specifically the benefits of SRE over CRE.

Finally computer assistant designs, e.g. rerandomization, will be briefly discussed

The Tennesee Project STAR Data

The data stems from a randomized evaluation of the effect of class size on test scores conducted in 1985-1986 in Tennesee called the Student/Teacher Achievement Ratio experiment, or Project STAR for short.

Mosteller (1995) calls it "one of the most important educational investigations ever carried out."

Here we use the kindergarten data where students and teachers were randomly assigned to small classes (13-17 students per teacher), regular classes (22-25 students per teacher), or regular classes with a teacher's aide.

To be eligible for Project Star, a school had to have a sufficient number of students to allow the formation of at least one class of each of the three types.

The Tennesee Project STAR Data

Once a school had been admitted to the program, a decision was made on the number of classes of each type (small, regular, regular with aide).

We take as fixed the number of classes of each type in each school. The unit of analysis for our analyses is the teacher or class, rather than the individual student, to help justify the no-interference part of SUTVA.

A school has a pool of at least 57 students, so they could support at least one small and two regular sized classes. Where these students come from, and how they differ from students in other schools is not important for the validity of our analysis (although it obviously affects the interpretation of the results and their generalizability).

The Tennesee Project STAR Data

Given the number of classes of each type, a school needs a certain number of teachers and teachers' aides. The availability of the teachers and aides may determine the number of classes, but this is again irrelevant for the validity of our analysis.

Two separate and independent randomizations took place.

One random assignment is that of teachers to classes of different types, small, regular, or regular with aide. The second randomization is of students to classes/teachers.

In our analysis, we mainly rely on the first randomization, of class-size and aides to teachers, using the teachers as the units of analysis.

Irrespective of the assignments of students to classes, the resulting inferences are valid for the effect on the teachers of being assigned to a particular type of class.

The Tennesee Project STAR Data

However, the second randomization is important for the interpretation of the results. Suppose we find that assignment to a small class leads on average to better outcomes for the teacher.

Without the randomization of students to classes, this could be due to systematic assignment of better students to the smaller classes. With the second randomization, this is ruled out, and systematic effects can be interpreted as the effects of class size.

This type of double randomization is somewhat similar to that in "split plot" designs (Cochran and Cox, 1957), although in split plot designs two different treatments are being applied by the double randomization.

The Tennesee Project STAR Data

Given the structure of the experiment, one could also focus on students as the unit of analysis, and investigate effects of class size on student-level outcomes.

The concern, however, is that the SUTVA is not plausible in that case.

Violations of SUTVA complicate the Neyman, regression and imputation approaches considerably, and we therefore primarily focus on class (*i.e.,* teacher-)level analyses in this chapter.

As we see, however, it remains straightforward to use the FEP approach to test the null hypothesis that assignment of students to different classes had no effect on test scores whatsoever, because SUTVA is automatically satisfied under Fisher's sharp null hypothesis of no effects of the treatment.

The Tennesee Project STAR Data

We focus on the comparison between regular (control) and small (treated) classes, and ignore the data for regular classes with teacher's aides.

We discard schools that do not have at least two classes of both the small size and the regular size.

Focusing on schools with at least two regular classes and two small classes leaves us with sixteen schools, which creates sixteen strata or blocks. Most have exactly two classes of each size, but one has two regular classes and four small classes, and a two other schools have three small classes and two regular-sized classes.

The total number of teachers and classes in this reduced data set is $N = 68$. Out of these $N_c = 32$ are assigned to regular-sized classes and $N_t = 36$ are assigned to small classes.

Table 9.1: Class Average Mathematics Scores from Project STAR

School	# classes	Regular Classes ($W_i = 0$)	Small Classes ($W_i = 1$)
1	4	-0.197, 0.236	0.165, 0.321
2	4	0.117, 1.190	0.918, -0.202
3	5	-0.496, 0.225	0.341, 0.561, -0.059
4	4	-1.104, -0.956	-0.024, -0.450
5	4	-0.126, 0.106	-0.258, -0.083
6	4	-0.597, -0.495	1.151, 0.707
7	4	0.685, 0.270	0.077, 0.371
8	6	-0.934, -0.633	-0.870, -0.496, -0.444, 0.392
9	4	-0.891, -0.856	-0.568, -1.189
10	4	-0.473, -0.807,	-0.727, -0.580
11	4	-0.383, 0.313	-0.533, 0.458
12	5	0.474, 0.140	1.001, 0.102, 0.484
13	4	0.205, 0.296	0.855, 0.509
14	4	0.742, 0.175	0.618, 0.978
15	4	-0.434, -0.293	-0.545, 0.234
16	4	0.355, -0.130	-0.240, -0.150
Average		-0.13	0.09
(std)		(0.56)	(0.61)

The Structure of Stratified Randomized Experiments

In SRE, units are grouped together according to some pre-treatment characteristics into strata.

Within each stratum, a CRE is conducted, and thus, within each stratum, the methods previously discussed are applicable.

However, the interest is not about hypotheses or treatment effects within a single stratum, but rather it is about hypotheses and treatment effects across all strata.

The Case with Two Strata

Divide the sample of N units into two subsamples, e.g., females and males, with subsample size N_f and N_m respectively, so that $N = N_f + N_m$.

It is useful to postulate for each unit a binary covariate, e.g., the unit's sex, with the membership in strata based on this covariate, denoted G_i for this particular covariate.

As with any other covariate, the value of G_i is not affected by the treatment. In this example G_i takes on the values f and m.

Define the finite sample average treatment effects in the two strata:

$$\tau_{\mathrm{FS}}(f) = \frac{1}{N_f} \sum_{i:G_i=f} \Big(Y_i(1) - Y_i(0) \Big), \qquad \text{and} \quad \tau_{\mathrm{FS}}(m) = \frac{1}{N_m} \sum_{i:G_i=m} \Big(Y_i(1) - Y_i(0) \Big).$$

The Case with Two Strata

Within each stratum, we conduct a CRE; with N_{tf} and N_{tm} 'treated' in the two subsamples respectively, and the remaining $N_{cf} = N_f - N_{tf}$ and $N_{cm} = N_m - N_{tm}$ 'controls'.

Let $N_t = N_{tf} + N_{tm}$ be the total number of 'treated' units and $N_c = N_{cf} + N_{cm}$ be the total number of 'control' units.

Let us consider the assignment mechanism.

The Case with Two Strata

Within the $G_i = f$ subpopulation, N_{tf} units out of N_f are randomly chosen to receive the treatment.

There are $\begin{pmatrix} N_f \\ N_{tf} \end{pmatrix}$ such allocations.

There are, furthermore $\begin{pmatrix} N_m \\ N_{tm} \end{pmatrix}$ allocations for the units with $G_i = m$

The assignment mechanism can thus be written as

$$\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{S}) = \begin{pmatrix} N_f \\ N_{tf} \end{pmatrix}^{-1} \cdot \begin{pmatrix} N_m \\ N_{tm} \end{pmatrix}^{-1} \quad \text{for} \quad \mathbf{W} \in \mathbb{W}^+,$$

$$\text{where } \mathbb{W}^+ = \left\{ \mathbf{W} \text{ such that } \sum_{i:G_i=f} W_i = N_{tf}, \sum_{i:G_i=m} W_i = N_{tm} \right\}.$$

The Case with Two Strata

Compare this to that for a CRE with $N_t = N_{tf} + N_{tm}$ assigned to treatment and $N_c = N_f - N_{tf} + N_m - N_{tm}$ assigned to control.

A large number of assignment vectors that would have positive probability with a CRE have probability zero with the SRE: all vectors with $\sum_{i=1}^{N} W_i = N_{tf} + N_{tm}$ but $\sum_{i:G_i=f} W_i \neq N_{tf}$ (or, equivalently, $\sum_{i:G_i=m} W_i \neq N_{tm}$).

If $N_{tf}/N_f \approx N_{tm}/N_m$ the stratification rules out substantial imbalances in the covariate distributions in the two treatment groups that could arise by chance in a CRE.

The possible disadvantage of the stratification is that a large number of possible assignment vectors are eliminated, just as a CRE eliminates assignment vectors that would be allowed under Bernoulli trials.

The Case with Two Strata

The advantage of a CRE over a Bernoulli trial is that of eliminating typically those assignment vectors with a severe imbalance between the number of controls and the number of treated.

Here the argument is similar, although not quite as obvious.

If we were to randomly partition the population into strata, the assignment vectors eliminated by the stratification are in expectation as helpful as the ones included, and the stratification will not produce a more informative experiment.

However, if the stratification is based on characteristics that are associated with the outcomes of interest, we shall see that SRE generally are more informative than CRE.

The Case with Two Strata

For example, in many drug trials, one may expect systematic differences in typical outcomes, both given the drug and without the drug, for men and women.

In that case, a SRE makes eminent sense. It can lead to more precise inferences, by eliminating the possibility of assignments with severe imbalances in sex distribution, for example, the extreme and uninformative assignment with all women exposed to the active treatment and all men exposed to the control treatment.

The Case with J Strata

Let J be the number of strata, and $N(j)$ and N_{tj} the total number of units and the number of treated units in strata j respectively, for $j = 1, \ldots, J$.

Let $G_i \in \{1, \ldots, J\}$ denote the stratum for unit i, and let $B_i(j) = \mathbf{1}_{G_i=j}$, be the indicator that is equal to one if unit i is in stratum j, and zero otherwise.

Within stratum j there are now $\begin{pmatrix} N(j) \\ N_{tj} \end{pmatrix}$ possible assignments, so that the assignment mechanism is

$$\Pr(\mathbf{W}|\mathbf{S}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{J} \begin{pmatrix} N(j) \\ N_{tj} \end{pmatrix}^{-1} \qquad \text{for} \quad \mathbf{W} \in \mathbb{W}^+,$$

where $\mathbf{W}^+ = \{\mathbf{W} | \sum_{i:G_i=j} W_i = N_{tj} \text{ for } j = 1, \ldots, J\}$.

Fisher's Exact P-values in SRE

Let $\overline{Y}_c^{\text{obs}}(j)$ and $\overline{Y}_t^{\text{obs}}(j)$ be the average observed outcome for units in stratum $G_i = j, j = 1, \ldots, J$ and let $N_c(j)$ and $N_t(j)$ be the number of units in stratum $G_i = j$ assigned to the control and treatment groups respectively:

$$\overline{Y}_c^{\text{obs}}(j) = \frac{1}{N_c(j)} \sum_{i:G_i=j} (1 - W_i) \cdot Y_i^{\text{obs}}, \qquad \overline{Y}_t^{\text{obs}}(j) = \frac{1}{N_t(j)} \sum_{i:G_i=j} W_i \cdot Y_i^{\text{obs}}$$

$$N_c(j) = \sum_{i=1}^{N} B_i(j) \cdot (1 - W_i), \qquad \text{and} \quad N_t(j) = \sum_{i=1}^{N} B_i(j) \cdot W_i.$$

For ease of exposition, we focus initially on the case with two strata, $G_i \in \{f, m\}$.

The Fisher's sharp null hypothesis is $H_0 : \ Y_i(0) = Y_i(1)$ for $i = 1, 2, \ldots, N$.

Fisher's Exact P-values in SRE

Obvious statistics are:

$$T^{\text{avg}}(f) = \left| \overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c(f)^{\text{obs}} \right| \qquad \text{and} \quad T^{\text{avg}}(m) = \left| \overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m) \right|.$$

Neither of the statistics is particularly attractive by itself: for either one an entire stratum is ignored, and thus the test would not be sensitive to violations of the null hypothesis in the stratum that is ignored.

A more appealing statistic is based on the combination of the two within-stratum statistics, $T^{\text{avg}}(f)$ and $T^{\text{avg}}(m)$, for example, the absolute value of a convex combination of the two difference in averages:

$$T^{\text{avg},\lambda} = \left| \lambda \cdot \left(\overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c^{\text{obs}}(f) \right) + (1-\lambda) \cdot \left(\overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m) \right) \right|,$$

for some $\lambda \in [0,1]$.

Fisher's Exact P-values in SRE

For any fixed value of λ, we can use the same FEP approach and find the distribution of the statistic under the null hypothesis, and thus calculate the corresponding p-value.

An obvious choice for λ is to weight $T^{\text{ave}}(f)$ and $T^{\text{ave}}(m)$ by the relative sample sizes (RSS) in the strata, and choose $\lambda = \lambda_{\text{RSS}} \equiv N_f/(N_f + N_m)$.

In that case, this choice for the weight parameter λ_{RSS} would lead to the natural statistic that is common in a CRE,

$$T^{\text{avg},\lambda_{\text{RSS}}} = \left| \frac{N_f}{N_f + N_m} \cdot \left(\overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c^{\text{obs}}(f) \right) + \frac{N_m}{N_f + N_m} \cdot \left(\overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m) \right) \right|.$$

Fisher's Exact P-values in SRE

If the relative proportions of treated and control units in each stratum, N_{tf}/N_f and N_{tm}/N_m respectively, are very different, however, this choice for λ does not necessarily lead to a very powerful test statistic.

Suppose, for example, that both strata contain 50 units, where in stratum f, only a single unit gets assigned to treatment, and the remaining 49 units get assigned to control, whereas in stratum m, the number of treated and control units is 25.

In that case, the test based on $T^{\text{avg}}(m)$ is likely to have substantially more power than the test based on $T^{\text{avg}}(f)$.

By using λ_{RSS} we are giving both stratum-specific average observed outcome differences $\hat{\tau}(f)$ and $\hat{\tau}(m)$ equal weights which would lead to that a test statistic with poor power properties.

Fisher's Exact P-values in SRE

An alternative choice for λ is motivated by considering against which alternative hypotheses we would like our test statistic to have power.

Often an important alternative hypothesis has a treatment effect that is constant both within, and between, strata.

Based on this perspective, it is useful to consider the sampling variances of $T^{\text{avg}}(f)$ and $T^{\text{avg}}(m)$, under Neyman's repeated sampling perspective.

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c^{\text{obs}}(f)\right) = \frac{S_t^2(f)}{N_{tf}} + \frac{S_c^2(f)}{N_{cf}} - \frac{S_{tc}(f)^2}{N_f},$$

and

$$\mathbb{V}_W\left(\overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m)\right) = \frac{S_t^2(m)}{N_{tm}} + \frac{S_c^2(m)}{N_{cm}} - \frac{S_{tc}^2(m)}{N_m}.$$

Fisher's Exact P-values in SRE

Suppose that the treatment effects are constant (i.e. $S^2_{tcf} = S^2_{tcm} = 0$).

Assume, in addition, that all four variances, S^2_{xw}, are equal to S^2.

Then the sampling variances of the two observed differences are

$$\mathbb{V}_W\left(\overline{Y}^{\text{obs}}_{tf} - \overline{Y}^{\text{obs}}_{cf}\right) = S^2 \cdot \left(\frac{1}{N_{tf}} + \frac{1}{N_{cf}}\right),$$

and

$$\mathbb{V}_W\left(\overline{Y}_{tm} - \overline{Y}_{cm}\right) = S^2 \cdot \left(\frac{1}{N_{tm}} + \frac{1}{N_{cm}}\right).$$

Fisher's Exact P-values in SRE

In that case, a sensible choice for λ would be the value that maximizes precision by weighting the two statistics by the inverse of their sampling variances, or

$$\lambda_{\text{opt}} = \frac{1}{\frac{1}{N_{tf}} + \frac{1}{N_{cf}}} \Bigg/ \left(\frac{1}{\frac{1}{N_{tm}} + \frac{1}{N_{cm}}} + \frac{1}{\frac{1}{N_{cm}} + \frac{1}{N_{tm}}} \right)$$

$$= \frac{N_f \cdot \frac{N_{tf}}{N_f} \cdot \frac{N_{cf}}{N_f}}{N_f \cdot \frac{N_{tf}}{N_f} \cdot \frac{N_{cf}}{N_f} + N_m \cdot \frac{N_{tm}}{N_m} \cdot \frac{N_{cm}}{N_m}},$$

with the weight for each stratum proportional to the product of the stratum size and the stratum proportions of treated and control units.

Fisher's Exact P-values in SRE

An alternative natural statistics is:

$$T^{\text{avg}} = \left| \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right|.$$

In the current setting of SRE, with two strata, this statistic can be written as:

$$T^{\text{avg}} = \left| \frac{1}{N_{tf} + N_{tm}} \sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_f - N_{tf} + N_m - N_{tm}} \sum_{i=1}^{N} (1 - W_i) \cdot Y_i^{\text{obs}} \right|.$$

Then we can write this statistic as

$$T^{\text{avg}} = \left| \frac{N_{tf}}{N_t} \cdot \overline{Y}_t^{\text{obs}}(f) - \frac{N_f - N_{tf}}{N_c} \cdot \overline{Y}_c^{\text{obs}}(f) + \frac{N_{tm}}{N_t} \cdot \overline{Y}_t^{\text{obs}}(m) - \frac{N_{cm}}{N_c} \cdot \overline{Y}_c^{\text{obs}}(m) \right|.$$

Fisher's Exact P-values in SRE

This statistic T^{avg} is a valid statistic for testing from the FEP perspective, but somewhat unnatural in the current context.

Because of Simpson's paradox one would not always expect small values for the statistic, even when the null hypothesis holds.

Suppose that the null hypothesis of zero treatment effects for all units holds, and that the potential outcomes are closely associated with the covariate that determines the strata

For example, assume $Y_i(0) = Y_i(1) = X_i$ for all units (e.g., $Y_i(0) = Y_i(1) = 1$ for units with $X_i = 1$ and $Y_i(0) = Y_i(1) = 2$ for units with $X_i = 2$).

Fisher's Exact P-values in SRE

In that case, the statistic T^{avg} is equal to

$$T^{\text{avg}} = \left| \frac{N_{tf}}{N_t} \cdot 1 - \frac{N_f - N_{tf}}{N_c} \cdot 1 + \frac{N_{tm}}{N_t} \cdot 2 - \frac{N_{cm}}{N_c} \cdot 2 \right|.$$

If $N_f = 10$, $N_{tf} = 5$, $N_m = 20$, and $N_{tm} = 5$, this is equal to

$$T^{\text{avg}} = \left| \frac{5}{10} \cdot 1 - \frac{5}{20} \cdot 1 + \frac{5}{10} \cdot 2 - \frac{15}{20} \cdot 2 \right| = \left| \frac{1}{2} + 1 - \frac{1}{4} - \frac{3}{2} \right| = \frac{1}{4}.$$

Under the sharp null hypothesis of no causal effects, the statistic $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ no longer has expectation equal to zero, whereas it did have expectation zero in the CRE.

Fisher's Exact P-values in SRE

Finally, let us consider rank-based statistics. In the setting with a CRE we focused on the difference in average ranks.

In that case we defined the normalized rank R_i (allowing for ties) as

$$R_i = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N+1}{2}.$$

Given the N ranks R_i, $i = 1, \ldots, N$, an obvious test statistic is the absolute value of the difference in average ranks for treated and control units:

$$T^{\text{rank}} = \left|\overline{R}_t - \overline{R}_c\right|, \qquad \text{where } \overline{R}_t = \frac{1}{N_t}\sum_{i:W_i=1} R_i, \quad \text{and } \overline{R}_1 = \frac{1}{N_c}\sum_{i:W_i=0} R_i.$$

where $\overline{R}_t$ and $\overline{R}_c$ are the average rank in the treatment and control groups respectively.

Fisher's Exact P-values in SRE

Although we can use this statistic for the FEP approach, this would not be attractive if there is substantial variation between strata.

We therefore propose modifying this statistic for the setting of a SRE.

Let R_i^{strat} be the normalized within-stratum rank of the observed outcome for unit i:

$$R_i^{\text{strat}} = \begin{cases} \sum_{j:G_i=f} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:G_i=f} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N_f+1}{2}, & \text{if } G_i = f, \\ \sum_{j:G_i=m} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:G_i=m} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N_m+1}{2}, & \text{if } G_i = m. \end{cases}$$

Fisher's Exact P-values in SRE

Then we can use the average value of the within-stratum ranks for treated and control units:

$$T^{\text{rank,stratum}} = \left| \overline{R}_t^{\text{strat}} - \overline{R}_c^{\text{strat}} \right|,$$

where

$$\overline{R}_t^{\text{strat}} = \frac{1}{N_t} \sum_{i:W_i=1} R_i^{\text{strat}}, \quad \text{and} \quad \overline{R}_c^{\text{strat}} = \frac{1}{N_c} \sum_{i:W_i=0} R_i^{\text{strat}}.$$

The FEP approach with J Strata

Most of the statistics discussed in the previous section extend naturally to the case with J strata.

Define for a general J-component vector λ the statistic

$$T^{\text{avg},\lambda} = \left| \sum_{j=1}^{J} \lambda_j \cdot \left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j) \right) \right|.$$

The first natural choice for λ has λ_j proportional to the stratum size,

$$\lambda_j = \frac{N(j)}{N}, \qquad \text{leading to} \quad T^{\text{avg},\lambda_{\text{RSS}}} = \left| \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j) \right) \right|.$$

The FEP approach with J Strata

The second choice for λ minimizes the sampling variance of the contrast between treated and control averages under homoskedasticity, leading to

$$\lambda_{\text{opt}} = \frac{N(j) \cdot \frac{N_t(j)}{N(j)} \cdot \frac{N_c(j)}{N(j)}}{\sum_{k=1}^{J} N(k) \cdot \frac{N_t(k)}{N(k)} \cdot \frac{N_c(k)}{N(k)}},$$

in turn leading to

$$T^{\text{ave},\lambda_{\text{opt}}} = \left| \frac{1}{\sum_{j=1}^{J} N(j) \cdot \frac{N_t(j)}{N(j)} \cdot \frac{N_c(j)}{N(j)}} \sum_{j=1}^{J} N(j) \cdot \frac{N_t(j)}{N(j)} \cdot \frac{N_c(j)}{N(j)} \cdot \left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j) \right) \right|.$$

The FEP approach with J Strata

For the modified rank statistic, we define R_i^{strat} to be the normalized within-stratum rank of the observed outcome for unit i, taking account of ties:

$$R_i^{\text{strat}} = \sum_{i':G_{i'}=G_i} \mathbf{1}_{Y_{i'}^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{i':G_{i'}=G_i} \mathbf{1}_{Y_{i'}^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N(G_i)+1}{2}.$$

Then we can use the average value of the within-stratum ranks for treated and control units:

$$T^{\text{rank,stratum}} = \left|\overline{R}_t^{\text{strat}} - \overline{R}_c^{\text{strat}}\right|,$$

where, as before, $\overline{R}_t^{\text{strat}}$ and $\overline{R}_c^{\text{strat}}$ are the averages of the ranks for treated and control units.

The FEP approach with Class-level Data from Project STAR

Let $B_i(j)$, $i = 1, \ldots, 68$, $j = 1, \ldots, 16$ be an indicator for unit (i.e., teacher) i being from stratum (school) j.

- For 13 schools with 2 classes of each type, the number of possible assignments are $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$.
- For 2 schools with 3 small classes and 2 regular classes, the number of possible assignments are $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10$
- For 1 school with 4 small and 2 regular classes, he number of possible assignments are $\begin{pmatrix} 6 \\ 2 \end{pmatrix} = 15$.

Hence, the total number of assignments of teachers to class type with positive probability is $(6^{13}) \times 10^2 \times 15 \approx 2 \times 10^{13}$. We therefore use numerical methods to approximate the p-values for the FEP approach.

The FEP approach with Class-level Data from Project STAR

We focus on the null hypothesis that there is no effect of class size on the average test score that a teacher would achieve for their students,

$$H_0: \quad Y_i(0) = Y_i(1), \quad \text{for all } i = 1, \ldots, 68 \textit{ classes}.$$

We consider the follow four test statistics.

(1) $T^{\text{avg}} = \left| \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right|$

(2) $T^{\text{avg},\lambda_{\text{RSS}}} = \left| \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j) \right) \right|.$

(3) $T^{\text{ave},\lambda_{\text{opt}}} = \left| \frac{1}{\sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_t(j)}{N(j)} \cdot \frac{N_c(j)}{N(j)}} \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_t(j)}{N(j)} \cdot \frac{N_c(j)}{N(j)} \cdot \left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j) \right) \right|.$

(4) $T^{\text{range}} = \frac{1}{N} \sum_{j=1}^{J} N(j) \cdot \Delta(j)$, where
$\Delta_c(j) = \max_{i:W_i=0,G_i=j} Y_i^{\text{obs}} - \min_{i:W_i=0,G_i=j} Y_i^{\text{obs}}$,
$\Delta_t(j) = \max_{i:W_i=1,G_i=j} Y_i^{\text{obs}} - \min_{i:W_i=1,G_i=j} Y_i^{\text{obs}}$ and $\Delta(j) = \Delta_t(j) - \Delta_c(j)$.

The FEP approach with Class-level Data from Project STAR

The realized value of

(1) is 0.224, with a corresponding p-value of $p = 0.034$.

(2) is 0.241, with a corresponding $p = 0.023$.

(3) is 0.238, with a corresponding $p = 0.025$

(4) is 0.226, with a corresponding $p = 0.109$

The first three test statistics suggesting that it is unlikely that the students of teachers assigned to the small classes had the same average test scores as the students of teachers assigned to large classes.

Wrt to the last, there is limited evidence against the null hypothesis that the variation in average scores differs between small and regular sized classes.

Recall that the p-value only has a valid interpretation if one statistic is specified *a priori*, and our exercise is for illustrative purposes only.

The FEP approach with Class-level Data from Project STAR

Note that (1) is not natural in this setting because one would not expect small values even when the null hypothesis is true (especially if there is substantial variation of the shares of treated units within the strata), although the results of the test are valid.

(3) is preferable when there is considerable variation in the proportion of treated and control units between strata, this statistic is more powerful against alternative hypotheses with constant additive treatment effects.

In the current application, these three test-statistics lead to very similar p-values. This is partly because most of the schools have two classes of each type. If there were more dispersion in the fraction of small classes by school and in the number of classes per school, the results could well differ more for the three statistics.

The FEP approach with Class-level Data from Project STAR

The value of the rank-based test $T^{\text{rank,stratum}}$ is 0.48, leading to a p-value of 0.15.

Because the outcomes themselves are averages (over students within the classes), there are few outliers, and in this case, the rank-based tests would not be expected to have an advantage over statistics based on simple averages.

The FEP approach with Class-level Data from Project STAR

Under the null hypothesis, the expected value of the average mathematics score in regular and small classes should be the same.

However, because in small classes the average is calculated over fewer students than in large classes, the small class averages should have a larger variance.

More precisely, if the individual test scores have a mean μ and variance σ^2, than the average in a class of size K should have mean μ and variance σ^2/K.

So, even if individual student scores are not affected by class size, the null hypothesis that at the teacher level, the average test score is not affected by the class size need not be true. This is the rational for using (4)

The FEP approach with Student-level Data from Project STAR

This analysis is specific to the FEP approach and the particular structure of the Project Star data, and not generally applicable to SRE. The purpose here, is to show the richness of the FEP approach.

The key issue is that under the null hypothesis of no effects whatsoever, the no interference assumption in the SUTVA holds automatically, but it need not hold under the alternative hypothesis.

Recall that the experiment assigned students and teachers randomly to the classes.

We index potential outcomes by the assignment vector that describes the class and teacher pair for each student.

The FEP approach with Student-level Data from Project STAR

First consider the data from a single school, j. This school has $N(j)$ students and $P(j)$ teachers and classes.

These students and teachers will be randomly assigned to $P(j)$ classes, with the class size for class s equal to $M_s(j)$.

The class sizes must add to the school size, or $\sum_{s=1}^{P(j)} M_s(j) = N(j)$. The total number of ways one can select the students, given class sizes, is

$$\prod_{s=1}^{P(j)-1} \left(\begin{array}{c} N(j) - \sum_{t<s} M_t(j) \\ M_s(j) \end{array} \right).$$

The FEP approach with Student-level Data from Project STAR

The $P(j)$ teachers can be assigned to the $P(j)$ classes in $P(j)!$ ways, so the total number of ways the students and teachers for school j can be assigned to classes is

$$\prod_{s=1}^{P(j)-1} \binom{N(j) - \sum_{t<s} M_t(j)}{M_s(j)} \cdot P(j)!.$$

For each student this is the total number of potential outcomes. The basis for the randomization distribution is this set of assignments, which are all equally likely. The total number of assignments is obtained by multiplying this for each school, across all schools:

$$\prod_{j=1}^{J} \prod_{s=1}^{S_j-1} \binom{N(j) - \sum_{t<s} M_t(j)}{M_s(j)} \cdot P(j)!.$$

The FEP approach with Student-level Data from Project STAR

The null hypothesis we consider is that of no effect whatsoever, against the alternative hypothesis that some potential outcomes differ.

The following stratum weighted test statistic is used

$$T^{\text{student}} = \left| \frac{1}{\sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_c(j)}{N(j)} \cdot \frac{N_t(j)}{N(j)}} \cdot \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \frac{N_c(j)}{N(j)} \cdot \frac{N_t(j)}{N(j)} \cdot \left(\overline{Y}_t(j)^{\text{obs}} - \overline{Y}_c(j)^{\text{obs}} \right) \right|.$$

The observed statistic is 0.242, with a p-value < 0.001. Thus, we get much stronger evidence against this null hypothesis than we did for the null hypothesis using class-level data.

If the no-interference assumption in SUTVA holds at the student level this is a valid test for an effect of class size.

The FEP approach with Student-level Data from Project STAR

In that case the student-level test will likely be more powerful than the teacher-level test.

However, here, the student-level stability assumption is strong and tenuous.

It is very plausible that there are interactions between children that would violate this assumption. Hence, even clear rejections of the null hypothesis of no differences by teacher assignment would not necessarily be credible evidence of systematic effects of class *size* — it may simply indicate the presence of effects of teachers or peers.

In contrast, the teacher-level assessment does not rely on within-class no-interference assumptions, and so clear evidence against the null hypothesis of no effect based on that assessment are more credible as evidence of class-size effects.

The Analysis of SRE from Neyman's Repeated Sampling Perspective

Initially we derive results with two strata (f and m), and then apply the framework to the Project Star data.

For the first stratum, the natural unbiased estimator for the average treatment effect $\tau_{\rm FS}(f)$ is

$$\hat{\tau}^{\rm dif}(f) = \overline{Y}_t^{\rm obs}(f) - \overline{Y}_c^{\rm obs}(f) = \frac{1}{N_{tf}} \sum_{i:G_i=f} W_i \cdot Y_i^{\rm obs} - \frac{1}{N_{cf}} \sum_{i:G_i=f} (1-W_i) \cdot Y_i^{\rm obs}.$$

The sampling variance of this estimator, under the randomization distribution, is

$$\mathbb{V}_W\left(\hat{\tau}^{\rm dif}(f)\right) = \frac{S_c^2(f)}{N_{cf}} + \frac{S_t^2(f)}{N_{tf}} - \frac{S_{ct}^2(f)}{N_f},$$

with analogous expressions for $\hat{\tau}^{\rm dif}(m)$

The Analysis of SRE from Neyman's Repeated Sampling Perspective

A natural estimand is the finite sample average treatment effect,

$$\tau_{\mathrm{S}} = \frac{N_f}{N_f + N_m} \cdot \tau_{\mathrm{FS}}(f) + \frac{N_m}{N_f + N_m} \cdot \tau_{\mathrm{FS}}(m) = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) \Big).$$

Because the stratum sizes are known, unbiasedness of the two within-stratum estimators implies unbiasedness of

$$\hat{\tau}^{\mathrm{strat}} = \frac{N_f}{N_f + N_m} \cdot \hat{\tau}^{\mathrm{dif}}(f) + \frac{N_m}{N_f + N_m} \cdot \hat{\tau}^{\mathrm{dif}}(m),$$

for the population average treatment effect τ_{S}.

The Analysis of SRE from Neyman's Repeated Sampling Perspective

Similarly, the assumption that the randomizations in the two strata are independent, implies

$$\begin{aligned}\mathbb{V}_W\left(\hat{\tau}^{\text{strat}}\right) &= \left(\frac{N_f}{N_f+N_m}\right)^2 \cdot \mathbb{V}_W(\hat{\tau}_f) + \left(\frac{N_m}{N_f+N_m}\right)^2 \cdot \mathbb{V}_W(\hat{\tau}_m) \\ &= \left(\frac{N_f}{N_f+N_m}\right)^2 \cdot \left(\frac{S_c(f)^2}{N_{cf}} + \frac{S_t^2(f)}{N_{tf}} - \frac{S_{ct}^2(f)}{N_f}\right) \\ &\quad + \left(\frac{N_m}{N_f+N_m}\right)^2 \cdot \left(\frac{S_c(m)^2}{N_{cm}} + \frac{S_t^2(m)}{N_{tm}} - \frac{S_{ct}^2(m)}{N_m}\right).\end{aligned}$$

The Analysis of SRE from Neyman's Repeated Sampling Perspective

As discussed in Chapter 6 there is no direct way to estimate $S^2_{ct}(f)$ and $S^2_{ct}(m)$, so typically those terms are ignored, giving the estimator:

$$\hat{\mathbb{V}}^{\text{neyman}} = \left(\frac{N_f}{N_f + N_m}\right)^2 \cdot \left(\frac{s^2_c(f)}{N_{cf}} + \frac{s^2_t(f)}{N_{tf}}\right) + \left(\frac{N_m}{N_f + N_m}\right)^2 \cdot \left(\frac{s^2_c(m)}{N_{cm}} + \frac{s^2_t(m)}{N_{tm}}\right).$$

This estimator of the sampling variance is unbiased if the within-stratum treatment effects are constant and additive, and overestimates the sampling variance in expectation otherwise.

Note that we do not need to make assumptions about the variation in treatment effects between strata.

The Analysis of SRE from Neyman's Repeated Sampling Perspective

In some cases we may be interested in a different weighted average than τ_{S} of the within-strata treatment effects.

For example, we may be interested in the average effect of the treatment on the outcome for the units who received the treatment.

Given the random assignment, and within the strata, this effect is equal to $\tau_{\mathrm{FS}}(f)$ and $\tau_{\mathrm{FS}}(m)$, respectively.

However, when the proportions of treated units differ between the strata, the weights have to be adjusted to obtain an unbiased estimate of the average effect of the treatment on the units who received treatment.

The Analysis of SRE from Neyman's Repeated Sampling Perspective

The appropriate weights are proportional to the fraction of treated units in each strata, leading to the estimand

$$\tau_{\mathrm{FS},t} = \frac{N_{tf}}{N_{tf}+N_{tm}} \cdot \tau_{\mathrm{FS}}(f) + \frac{N_{tm}}{N_{tf}+N_{tm}} \cdot \tau_{\mathrm{FS}}(m),$$

and thus to the natural unbiased estimator

$$\hat{\tau}_t = \frac{N_{tf}}{N_{tf}+N_{tm}} \cdot \hat{\tau}(f) + \frac{N_{tm}}{N_{tf}+N_{tm}} \cdot \hat{\tau}(m).$$

The sampling variance of $\hat{\tau}_t$ can be estimated in the same way as the sampling variance for the population average treatment effect, modifying the weights to reflect the new estimand:

$$\hat{\mathbb{V}}_t^{\mathrm{neyman}} = \left(\frac{N_{tf}}{N_{tf}+N_{tm}}\right)^2 \cdot \left(\frac{s_c(f)^2}{N_{cf}} + \frac{s_t^2(f)}{N_{tf}}\right) + \left(\frac{N_{tm}}{N_{tf}+N_{tm}}\right)^2 \cdot \left(\frac{s_c^2(m)}{N_{cm}} + \frac{s_t^2(m)}{N_{tm}}\right).$$

The Analysis of SRE from Neyman's Repeated Sampling Perspective

More generally we can look at other weighted averages (e.g. the average effect for the non-treated).

A natural unbiased estimator for the difference between $\tau_{\mathrm{FS}}(m)$ and $\tau_{\mathrm{FS},}(f)$ is

$$\hat{\tau}^{\mathrm{dif}}(m) - \hat{\tau}^{\mathrm{dif}}(f) = \overline{Y}_t^{\mathrm{obs}}(m) - \overline{Y}_c^{\mathrm{obs}}(m) - \left(\overline{Y}_t^{\mathrm{obs}}(f) - \overline{Y}_c^{\mathrm{obs}}(f)\right).$$

This estimator is unbiased with sampling variance

$$\mathbb{V}_W(\hat{\tau}^{\mathrm{dif}}(m) - \hat{\tau}^{\mathrm{dif}}(f)) = \frac{S_c^2(f)}{N_{cf}} + \frac{S_t^2(f)}{N_{tf}} - \frac{S_{ct}^2(f)}{N_f} + \frac{S_c^2(m)}{N_{cm}} + \frac{S_t^2(m)}{N_{tm}} - \frac{S_{ct}^2(m)}{N_m}.$$

We can use the upper bound estimator to create large-sample confidence intervals:

$$\hat{\mathbb{V}}^{\mathrm{neyman}}\left(\hat{\tau}^{\mathrm{dif}}(m) - \hat{\tau}^{\mathrm{dif}}(f)\right) = \frac{s_c^2(f)}{N_{cf}} + \frac{s_t^2(f)}{N_{tf}} + \frac{s_c^2(m)}{N_{cm}} + \frac{s_t^2(m)}{N_{tm}}.$$

The Neyman approach and Project STAR

For each school j, $j = 1, \ldots, 16$, the average effect of the treatment and the corresponding sampling variance were estimated as

$$\hat{\tau}^{\text{dif}}(j) = \overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j), \qquad \text{and} \quad \hat{\mathbb{V}}^{\text{neyman}}(j) = \frac{s_c(j)^2}{N_c(j)} + \frac{s_t(j)^2}{N_t(j)},$$

respectively.

The population average effect was estimated as

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \hat{\tau}(j) = 0.241,$$

and its sampling variance by

$$\hat{\mathbb{V}}^{\text{neyman}} = \sum_{j=1}^{J} \left(\frac{N(j)}{N} \right)^2 \cdot \hat{\mathbb{V}}^{\text{neyman}}(j) = 0.092^2.$$

Table 9.2: Within-School Estimates of Treatment Effect of Small Classes Relative to Regular Classes

School	Estimated Effect	Estimated Standard Error
1	0.223	(0.230)
2	-0.295	(0.776)
3	0.417	(0.404)
4	0.748	(0.215)
5	-0.077	(0.206)
6	1.655	(0.405)
7	-0.254	(0.255)
8	0.429	(0.306)
9	-0.006	(0.311)
10	-0.014	(0.182)
11	-0.003	(0.605)
12	0.222	(0.309)
13	0.432	(0.179)
14	0.340	(0.336)
15	0.207	(0.396)
16	-0.306	(0.245)
overall est	0.241	(0.092)

The Neyman approach and Project STAR

Hence the large sample 95% confidence interval for the average effect was

$$\mathrm{CI}^{0.95}(\tau_{\mathrm{FS}}) = \Big(0.061, 0.421\Big).$$

Treating the data as arising from a CRE, the point estimate of the average effect $\hat{\tau}^{\mathrm{dif}} = \overline{Y}_t^{\mathrm{obs}} - \overline{Y}_c^{\mathrm{obs}} = 0.224$, with an estimated standard error of 0.141, leading to a large sample 95% confidence interval of $(-0.053, 0.500)$.

This estimator of the sampling variance is biased if there is variation in the probability of treatment between the different strata, or if there is variation in the average potential outcomes by stratum.

The Neyman approach and Project STAR

We know the former is the case, with the probability of a small class equal to 0.5 in most schools, and equal to 0.60 and 0.67 in some schools.

Assessing the latter is more complicated, and we shall return to this in a couple of slides.

The fact that the point estimates differ under the assumptions of a CRE versus a SRE analyses, suggests that average potential outcomes also differ between strata.

The estimated standard error for the stratification-based analysis is smaller than that for the CRE, suggesting, again, that average potential outcomes differ between strata, which implies that there is a gain in precision from the stratification.

Regression Analysis of SRE

In order to interpret regression-based estimators, we take a super population perspective with a fixed number of strata, and an infinite number of units within each stratum.

Let $q(j) = N(j)/N$ and $e(j) = N_t(j)/N(j)$ be the proportion of each stratum in the sample from the infinite super-population, and the proportion of treated units in each stratum, or the propensity score, respectively.

We consider two specifications of the regression function

- the stratum indicators are included additively as additional regressors
- the stratum indicators and a full set of interactions of the stratum indicators with the treatment indicator.

and then investigate the large sample properties of the OLS estimator.

Regression Analysis of SRE

$$Y_i^{\text{obs}} = \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \varepsilon_i. \tag{1}$$

In this specification, a full set of stratum indicators $B_i(j)$, for $j = 1, \ldots, J$, is included why an intercept is *not* included.

We focus on the OLS estimator for τ,

$$(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\tau,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) \right)^2. \tag{2}$$

As before, define τ^* and β^* as the population counterparts to the OLS estimator,

$$(\tau^*, \beta^*) = \arg\min_{\tau,\beta} \mathbb{E} \left[\left(Y_i^{\text{obs}} - \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) \right)^2 \right]. \tag{3}$$

Regression Analysis of SRE

The first question concerns the population value τ^* corresponding to $\hat{\tau}^{\text{ols}}$.

In general $\hat{\tau}^{\text{ols}}$ is not consistent for the population average treatment effect τ_{SP}. Instead, it estimates a weighted average of the within-stratum average effects, with weights proportional to the product of the fraction of observations in the stratum and the probabilities of receiving and not receiving the treatment.

More specifically,

$$\omega(j) = q(j) \cdot e(j) \cdot (1 - e(j)), \qquad \text{and} \quad \tau_\omega = \sum_{j=1}^{J} \omega(j) \cdot \tau(j) \Big/ \left(\sum_{j=1}^{J} \omega(j) \right). \tag{4}$$

Then $\hat{\tau}^{\text{ols}}$ is consistent for τ_ω.

The following theorem formalizes this result.

Regression Analysis of SRE

Theorem

Suppose we conduct a SRE in a sample drawn at random from an infinite population. Then, for estimands τ^* and τ_ω defined in (3) and (4), the estimator $\hat{\tau}^{\text{ols}}$ satisfies, (i)

$$\tau^* = \tau_\omega,$$

and (ii),

$$\sqrt{N}\left(\hat{\tau}^{\text{ols}} - \tau_\omega\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\left(W_i - \sum_{j=1}^{J} q(j)B_i(j)\right)^2 \left(Y_i^{\text{obs}} - \tau^* W_i - \sum_{j=1}^{J} \beta_j^* B_i(j)\right)^2\right]}{\left(\sum_{j=1}^{J} q(j)\cdot e(j)\cdot(1-p(j))\right)^2}\right).$$

Regression Analysis of SRE

Suppose we estimate $\hat{\tau}^{\text{dif}}(j) = \overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)$.

The sampling variance of $\hat{\tau}^{\text{dif}}(j)$, under the assumption of a constant treatment effect, is $(S^2/N) \cdot (q(j) \cdot e(j) \cdot (1 - e(j)))^{-1}$.

Hence the weights ω_j are proportional to the precision of natural unbiased estimators of the within-stratum treatment effects, which leads to a relatively precisely estimated weighted average effect.

Regression Analysis of SRE

The second specification of the regression function is

$$Y_i^{\text{obs}} = \tau \cdot W_i \cdot \frac{B_i(j)}{N(j)/N} + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \sum_{j=1}^{J-1} \gamma(j) \cdot W_i \cdot \left(B_i(j) - B_i(J) \cdot \frac{N(j)}{N(J)} \right) + \varepsilon_i. \quad (5)$$

Note that in this specification we only include the first $J-1$ interactions to avoid perfect collinearity in the regression function.

In this case, the population value τ^*, corresponding to the large sample limit of the OLS estimator $\hat{\tau}^{\text{ols,inter}}$, is equal to the population average treatment effect τ_{SP}:

Regression Analysis of SRE

Theorem

Suppose we conduct a SRE in a sample drawn at random from an infinite population. Then, for $\hat{\tau}^{\text{ols,inter}}$ *defined as the OLS estimator corresponding to the regression function in (5), and* τ^* *defined as the population limit corresponding to that estimator,* (i)

$$\tau^* = \tau_{\text{SP}},$$

and (ii),

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols,inter}} - \tau_{\text{SP}} \right) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{J} q(j)^2 \cdot \left(\frac{\sigma^2_{j0}}{(1-e(j)) \cdot q(j)} + \frac{\sigma^2_t(j)}{e(j) \cdot q(j)} \right) \right).$$

Note: In general, the sampling variance of $\hat{\tau}^{\text{ols,inter}}$ is larger than that of $\hat{\tau}^{\text{ols}}$.

Regression Analysis of Project STAR

The point estimate and the estimated standard error from the first specification are are

$$\hat{\tau}^{\text{ols}} = 0.238 \quad (\widehat{\text{s.e.}}\ 0.103).$$

Recall from the previous discussion that this estimator is not necessarily consistent for the τ_{SP} if there is variation in the effect of the class-size by school.

The point estimate and the estimated standard error based on the second specification are

$$\hat{\tau}_{\text{ols,inter}} = 0.241 \quad (\widehat{\text{s.e.}}\ 0.095).$$

The two estimates for the average effect are fairly close, with similar standard errors.

Model-based Analysis of SRE

In a model-based analysis, it is conceptually straightforward to take account of the stratification.

As in the analysis of CRE, we combine the specification of the joint distribution of the potential outcomes with the known distribution of the vector of assignment indicators to derive the posterior distribution of the causal estimand.

There is one new issue that arises in this context: the link between the distributions of the potential outcomes in distinct strata.

One can choose to have distinct parameters for the distributions in different strata, *i.e.,* independent prior distributions.

Model-based Analysis of SRE

Alternatively the researcher may wish to link the parameters in the different strata either deterministically by imposing equality restrictions, or stochastically through a dependence structure in the prior distribution, that is, for example, through a hierarchical model.

In situations with few strata and many units per stratum, one may wish to pursue the first strategy and specify distinct distributions for the potential outcomes in each stratum, with independent prior distributions on the parameters of these distributions.

In contrast, in settings with a substantial number of strata, and a modest number of units per stratum, one may wish to link some of the parameters. One can do so by restricting them to be equal, or by incorporating dependence into the specification of the prior distribution.

Model-based Analysis of SRE

We make this more specific and illustrate the issues for the case with common and stratum-specific parameters.

Suppose we specify the joint distribution of the potential outcomes in stratum j as

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| B_i(j) = 1, \left(\mu_c(j), \mu_t(j), \sigma_c^2(j), \sigma_t^2(j)\right)_{j=1}^{J} \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c(j) \\ \mu_t(j) \end{array} \right), \left(\begin{array}{cc} \sigma_c^2(j) & 0 \\ 0 & \sigma_t^2(j) \end{array} \right) \right), \tag{6}$$

where the means $(\mu_c(j), \mu_t(j))$ and variances $(\sigma_c^2(j), \sigma_t^2(j))$ are specific to stratum j.

The full parameter vector is $\theta = (\mu_c(j), \mu_t(j), \sigma_c^2(j), \sigma_t^2(j),\ w = 0, 1,\ j = 1, \ldots, J)$.

Model-based Analysis of SRE

However, if there are many strata and the number of units per stratum is modest, we may wish to specify a hierarchical prior distribution for the means to obtain more precise estimates.

For example, we may wish to restrict the variances of the potential outcomes to be the same across strata, and to specify the means to have a joint normal prior distribution, independent of the variances σ_0^2 and σ_t^2:

$$\begin{pmatrix} \mu_c(1) \\ \mu_c(2) \\ \vdots \\ \mu_c(j) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma_c \\ \gamma_c \\ \vdots \\ \gamma_c \end{pmatrix}, \begin{pmatrix} \eta_c^2 & 0 & \dots & 0 \\ 0 & \eta_c^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & \eta_c^2 \end{pmatrix} \right),$$

and

Model-based Analysis of SRE

$$\begin{pmatrix} \mu_t(1) \\ \mu_t(2) \\ \vdots \\ \mu_t(j) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma_t \\ \gamma_t \\ \vdots \\ \gamma_t \end{pmatrix}, \begin{pmatrix} \eta_t^2 & 0 & \dots & 0 \\ 0 & \eta_t^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & \eta_t^2 \end{pmatrix} \right),$$

The full parameter vector is now $\theta = (\sigma_c^2, \sigma_t^2, \gamma_c, \gamma_t, \eta_c^2, \eta_t^2)$.

A Model-based Analysis of Project STAR

The model considered for the potential outcomes are

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| \mu_c(1), \mu_t(1), \ldots, \mu_c(j), \mu_t(j), \sigma^2, B_i(j) = 1 \sim \mathcal{N}\left(\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right),$$

in addition we assume

$$\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix} \Big| \sigma^2, \gamma_c, \gamma_t, \Sigma \sim \mathcal{N}\left(\begin{pmatrix} \gamma_c \\ \gamma_t \end{pmatrix}, \Sigma \right), \qquad \begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix} \perp\!\!\!\perp \begin{pmatrix} \mu_c(k) \\ \mu_t(k) \end{pmatrix} \Big| \sigma^2, \gamma_c, \gamma_t, \Sigma, j \neq k.$$

In this model, the two potential outcome means $(\mu_c(j), \mu_t(j))$ are specific to the stratum, and the variance σ^2 is common to all strata and both potential outcomes and $\theta = (\gamma_c, \gamma_t, \Sigma, \sigma^2)$.

A Model-based Analysis of Project STAR

For the prior distributions, we use conventional proper choices.

For the variance parameter σ^2 we use a standard inverse gamma prior distribution,

$$k_0 \cdot \nu_0^2 \cdot \sigma^{-2} \sim \mathcal{X}^2(k_0), \qquad \text{or} \quad \sigma^2 \sim \mathcal{X}^{-2}(k_0, \nu_0^2),$$

using the notation from Gelman, Carlin, Stern and Rubin (1995).

The choices for the parameters of the prior distribution are $k_0 = 2$ and $\nu_0^2 = 0.001$.

For γ_c and γ_t we use independent normal prior distributions,

$$\begin{pmatrix} \gamma_c \\ \gamma_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100^2 & 0 \\ 0 & 100^2 \end{pmatrix} \right).$$

A Model-based Analysis of Project STAR

The prior distribution for Σ is an inverse wishart distribution,

$$\Sigma \sim \mathcal{W}^{-1}(k_1, \Gamma_1^{-1}).$$

We consider two pairs of values for (k_1, Γ_1).

(1) $k_1 = 1000$, $\Gamma_1 = 1000 \cdot \mathcal{I}_2$.

(2) $k_1 = 3$ and $\Gamma_1^{-1} = 0.001 \cdot k_1 \cdot \mathcal{I}_2$.

(1) essentially corresponds to removing the link between the parameters in the different strata. We refer to this as the $\perp\!\!\!\perp$ prior, corresponding to independence between the stratum specific means.

(2) allows the hierarchical structure to influence answers. We refer to this as the hierarchical choice.

A Model-based Analysis of Project STAR

The posterior mean and standard deviation for the independent prior are

$$\mathbb{E}[\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\mathbf{B},\perp\!\!\!\perp] = 0.241, \qquad \mathbb{V}(\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\mathbf{B},\perp\!\!\!\perp) = 0.095^2.$$

Substantively it is difficult to see why one would wish to impose the *ex post* independence. Certainly, as we will see below, there is strong evidence in the data to suggest that the average potential outcomes within the schools are related.

The posterior mean and standard deviation for the hierarchical prior are

$$\mathbb{E}[\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\mathbf{B},\mathrm{hierarchical}] = 0.235, \qquad \mathbb{V}(\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\mathbf{B},\mathrm{hierarchical})^2 = 0.107^2.$$

A Model-based Analysis of Project STAR

To assess the evidence for variation in average potential outcomes and treatment effects by strata, we inspect the posterior distribution of Σ given the hierarchical prior distribution.

The logarithm of the square root of the two diagonal elements correspond to the logarithm of the standard deviation of $\mu_c(j)$ and $\mu_t(j)$ over the 16 schools.

The posterior means of logarithms of those two standard deviations are

$$\mathbb{E}\left[\ln(\sqrt{\Sigma}_{11})\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=-1.14,\ \mathbb{V}\left(\ln(\sqrt{\Sigma}_{11})\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.47^2,$$

and

$$\mathbb{E}\left[\ln(\sqrt{\Sigma}_{22})\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=-1.08,\ \mathbb{V}\left(\ln(\sqrt{\Sigma}_{22})\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.45^2.$$

A Model-based Analysis of Project STAR

There is clearly some evidence of heterogeneity in the stratum means.

However, the heterogeneity is highly correlated across potential outcomes, with the posterior mean for the Fisher Z transformation of the correlation between $\beta_c(j)$ and $\beta_t(j)$ (the (1,2) element of Σ divided by the square root of the product of the $(1,1)$ and $(2,2)$ elements) equal to

$$\mathbb{E}\left[\frac{1}{2}\ln\left(\frac{1+\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}{1-\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}\right)\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right]=2.63,$$

and the posterior variance equal to

$$\mathbb{V}\left(\frac{1}{2}\ln\left(\frac{1+\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}{1-\Sigma_{12}/(\sqrt{\Sigma_{11}\Sigma_{22}})}\right)\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.67^2.$$

A Model-based Analysis of Project STAR

The posterior mean of the correlation itself is 0.96.

The average treatment effect in school j is approximately $\tau(j) = \mu_{t,j} - \mu_{c,j}$. In terms of the parameters, the variance of the treatment effect across the 16 schools is $(-1\ 1)\Sigma(-1\ 1)' = \Sigma_{11} - \Sigma_{12} - \Sigma_{21} + \Sigma_{22}$.

We focus on the standard deviation of the treatment effect over the schools.

The posterior mean of the logarithm of the standard deviation of the treatment effect is

$$\mathbb{E}\left[\ln\left(\sqrt{\Sigma_{11} - \Sigma_{12} - \Sigma_{21} + \Sigma_{22}}\right)\middle|\, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{B}, \text{hierarchical}\right] = -2.33,$$

and the posterior variance is

A Model-based Analysis of Project STAR

$$\mathbb{V}\left(\ln\left(\sqrt{\Sigma_{11}-\Sigma_{12}-\Sigma_{21}+\Sigma_{22}}\right)\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{B},\text{hierarchical}\right)=0.59^2.$$

Comparing the posterior mean of the standard deviation of the stratum-specific treatment effect $\tau(j)$ over the 16 strata, (0.115), with the posterior mean of the standard deviation of the stratum specific level under the control treatment $\mu_{c,j}$ over the 16 strata, (0.349), suggests that, although there is considerable evidence that *levels* of the average test scores vary by school, there is little evidence that average class size *effects* vary much by school.

The former may be due to differences in teacher quality or differences in student populations. This type of conclusion highlights the advantage of a fully model-based analysis, which allows for the simultaneous investigation of multiple questions

Design Issues: SRE versus CRE

Here we study the implications of the choice between the different experimental designs for the expected sampling variance of the standard unbiased estimator for the average treatment effect.

There is a sense in which one is never worse off stratifying on a covariate. However, to make this point precise, we need to pose the question appropriately.

We analyze the problem in a super population setting. Each unit in this population has a binary characteristic G_i, $G_i \in \{f, m\}$.

Let q be the proportion of f types in the population.

Design Issues: SRE versus CRE

We consider two designs. We randomly draw

(1) N units from the population. We then randomly draw $M = q \cdot N$ units to receive the active treatment and $N_c = (1-q) \cdot N$ units to receive the control treatment

(2) $N(f) = q \cdot N$ units from the subpopulation of units who have $G_i = f$, and $N(m) = N \cdot (1-q)$ units from the population who have $G_i = m$.
- randomly select $N_t(f) = pqN$ and $N_t(m) = p(1-q)N$ units to be 'treated'
- let the remaining $N_c(f) = (1-p) \cdot q \cdot N$ and $N_c(m) = (1-p) \cdot (1-q) \cdot N$ units to be 'controls'.

Design Issues: SRE versus CRE

For design (1) we estimate the average treatment effect in the superpopulation as

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

with (super-population) sampling variance

$$\mathbb{V}_{\text{SP}}(\hat{\tau}^{\text{dif}}) = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t}.$$

For design (2) we estimate

$$\hat{\tau}^{\text{dif}}(f) = \overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c^{\text{obs}}(f), \qquad \text{and} \quad \hat{\tau}^{\text{dif}}(m) = \overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m),$$

Design Issues: SRE versus CRE

and the overall average effect as

$$\hat{\tau}^{\text{strat}} = \frac{N_f}{N} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N_m}{N} \cdot \hat{\tau}^{\text{dif}}(m) = q \cdot \hat{\tau}^{\text{dif}}(f) + (1-q) \cdot \hat{\tau}^{\text{dif}}(m).$$

The sampling variance for this estimator is

$$\mathbb{V}_{\text{SP}}(\hat{\tau}^{\text{strat}}) = \frac{q}{N} \cdot \left(\frac{\sigma_t^2(f)}{p} + \frac{\sigma_c^2(f)}{1-p} \right) + \frac{1-q}{N} \cdot \left(\frac{\sigma_t^2(m)}{p} + \frac{\sigma_c^2(m)}{1-p} \right).$$

The difference between the two sampling variances, normalized by the sample size N, is

$$N \cdot \left(\mathbb{V}_{\text{SP}}(\hat{\tau}^{\text{dif}}) - \mathbb{V}_{\text{SP}}(\hat{\tau}^{\text{strat}}) \right) = q(1-q) \cdot \left((\mu_c(f) - \mu_c(m))^2 + (\mu_t(f) - \mu_t(m))^2 \right) \geq 0.$$

Design Issues: SRE versus CRE

Although there is an unambiguous ranking of $\mathbb{V}_{\mathrm{SP}}(\hat{\tau}^{\mathrm{dif}})$ and $\mathbb{V}_{\mathrm{SP}}(\hat{\tau}^{\mathrm{strat}})$, the *estimated* sampling variance for the SRE may be larger than for the CRE.

The natural estimator for the sampling variance of the simple unbiased estimator in a SRE has a larger sampling variance than that for the natural estimator for the sampling variance in a CRE, because of the need to estimate the within-stratum potential outcome variances.

We can assess the benefits of having the stratification for an experiment with the size of project STAR.

Design Issues: SRE versus CRE, project STAR

Suppose we have S strata, each with N_t treated (small) and $N_c = N_t$ control (regular-sized) classes.

Suppose that the true within-stratum variance of the potential outcomes is $\sigma^2 = 0.43^2$, which is the posterior mean for the hierarchical model estimated on the STAR data.

Suppose also that the true variance of the within-stratum average potential outcomes over the strata is $\Sigma_{11} = 0.37^2$ for the control averages $\mu_{c,j}$ and $\Sigma_{22} = 0.37^2$ for the averages given the treatment $\mu_{t,j}$, again estimated on the STAR data.

Then the ratio of the variances under CRE versus a SRE would be $(0.43^2 + 0.37^2)/0.43^2) = 1.65$. Thus, the stratified design reduces the variance by 40%.

Computer assisted designs

An alternative, or complement, to stratification or blocking that has received attention lately is to utilize modern computational capabilities in finding allocations with balance in observed covariates.

Rerandomiziation: Morgan and Rubin (2012) formalized rerandomization by suggesting the experiment allocation to be one with a Mahalanobis distance (MD) in covariate means of treated and control units to be less than a given threshold. In Johansson and Schultzberg (2020) the experiment allocations are instead randomly chosen among the approximate 'best' subset of admissible allocations.

Algorithms: Bertsimas et al. (2015); Lauretto et al. (2017); Kallus (2018); Krieger et al. (2019); Kapelner et al. (2020) suggest using covariates in algorithms to find a set of (optimal) allocations from which the final experimental allocation should be chosen.

Computer assisted designs and inference.

As we saw, with stratification standard normal asymptotic inference is straightforward. This is no longer the case in these designs.

Morgan and Rubin (2012) suggested using the FEP based on $\hat{\tau}^{\text{dif}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$. The asymptotic distribution for this estimator was later derived in Li et al. (2018) and Li and Ding (2020) derived the asymptotics under regression adjustment.

Kallus (2018) suggested using bootstrap inference in his optimal design but gave no formal proof of its validity.

Computer assisted designs and inference.

Krieger et al. (2019); Kapelner et al. (2020); Johansson and Schultzberg (2020) all suggested FEP, while Bertsimas et al. (2015); Lauretto et al. (2017) did not discuss inference in their designs.

Zhang and Johansson (2019) suggested Bayesian inference and studied the small sample performance under the MD rerandomization design.

Rerandomization and "optimal designs"

The idea of rerandomization is to remove allocations in $\mathbb{W}$ with imbalance in observed covariates between treated and control units given a pre-specified imbalance criterion, thus $\mathbb{W} \rightarrow \mathbb{W}_a$. Different criteria give rise to different rerandomization designs.

Morgan and Rubin (2012) used the MD as the criterion for defining $\mathbb{W}_a$ with the aim to make inference about the τ_{FS}, estimated using $\hat{\tau}^{\text{dif}}$.

Johansson and Schultzberg (2020) suggested a rank based criterion and to find all allocations in $\mathbb{W}_a$ such that the Fisher exact test has a certain "resolution" (e.g. 200 possible assignments allows for a p-value of 0.01 and larger).

Kallus (2018) suggested finding the allocation that minimizes the estimated sampling variance of the $\hat{\tau}^{\text{dif}}$ estimator. Denote this set $\mathbb{W}_{Opt}$.

Mahalanobis-based rerandomization

Let $\mathbf{X}$ be the $N \times K$ matrix of fixed covariates with the finite population covariance matrix $\mathbf{S}_{xx}$.

With $\mathbf{x}_i, i = 1, ..., N$, the $K \times 1$ covariate vector, this covariance matrix is defined

$$\mathbf{S}_{xx} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \tag{7}$$

In a balanced experiment, i.e. $N_1 = N_0 = N/2$, there are $\binom{N}{N_1} = A$ possible treatment allocation (assignment) vectors, thus $\mathbf{W}^j$, $j = 1, ..., A$. The MD for each allocation j is defined:

$$M(\mathbf{W}^j, \mathbf{X}) = \frac{N}{4} (\widehat{\tau}_X^j \mathbf{S}_{xx}^{-1} \widehat{\tau}_X^j), \ j = 1, ..., A,$$

Mahalanobis-based rerandomization

where

$$\widehat{\tau}_x^j = \frac{1}{N_1}\sum_{i=1}^{N_1} W_i^j \mathbf{x}_i' - \frac{1}{N_0}\sum_{i=1}^{N_0}(1 - W_i^j)\mathbf{x}_i' = \overline{\mathbf{x}}_1^j - \overline{\mathbf{x}}_0^j.$$

Accept the treatment assignment vector $\mathbf{W}^j$ if

$$M(\mathbf{W}^j, \mathbf{X}) \leq a,$$

where a is a positive constant. Using the central limit theorem the means will be normally distributed why $M(\mathbf{W}^j, \mathbf{X}) \sim \chi^2_K$.

Mahalanobis-based rerandomization

Three properties of the MD criterion

1. χ^2_K distributed with moderate size N. This imply that a can be implicitly defined.
   1. Let a be a value such that $\mathrm{P}(M(\mathbf{W}^j, \mathbf{X}) \leq a) = P(\chi^2_K \leq a) = p_a$, then to randomize within the set of the 0.01% best balanced allocations implies setting $a : p_a = 0.0001$.
2. The percent reduction in variance of all of the (equally weighted) included covariates is equal to
$$100(1 - \nu_a)$$
where $\nu_a = \frac{\Pr(\chi^2_{K+2} \leq a)}{\Pr(\chi^2_K \leq a)}; 0 < \nu_a < 1$,
3. Assume (i) $\mathbf{Y}(0)$ conditionally on $\mathbf{X}$ to be normal and (ii) additive treatment effects then the percent reduction in variance on the treatment effect is equal to $R^2(1 - \nu_a)$,where R^2 is the coefficient of determination of a regression of $\mathbf{Y}(0)$ on $\mathbf{X}$.

Mahalanobis-based rerandomization and inference

Most often we do not have random sampling to the experiment. That is, interest in inferences to the sample only why the FRT can be used for inference.

If we however would like to make use of asymptotic inferences based on $\hat{\tau}^{\text{dif}}$ after rerandomization, $\widehat{\tau}^{rr}$, then we can no longer make use of the normal distribution.

Under rerandomization based on the MD criterion Li et al. (2018) shows that

$$\sqrt{N}(\widehat{\tau}^{rr} - \tau_{FS})|\widehat{\tau}_X \sim \sqrt{V_{\tau\tau}}Q,$$

Mahalanobis-based rerandomization and inference

where $Q = \sqrt{(1 - R^2)}\varepsilon_0 + \sqrt{R^2}L_{K,a}$.

Here ε_0 is a standard normal random variable orthogonal to the covariates while the other part

$$L_{K,a} \sim \chi_{K,a} S \sqrt{\beta_K},$$

projects into the space of covariates and is thus affected by the rerandomization. Here $\chi_{k,a} = \chi^2_K | \chi^2_K \leq a$, S a random sign taking ± 1 with probability $1/2$, and $\beta_K \sim$Beta$(1/2; (K-1)/2)$ a Beta random variable degenerating to a point mass at 1 when $K = 1$.

Rerandomization and inference, continued

Under Mahalanobis-based rerandomization, Li and Ding (2020) showed that asymptotically Lin (2013)'s approach is valid also under MD rerandomization. That is run the regression

$$Y_i = \beta_{00} + (\mathbf{x}_i - \bar{\mathbf{x}})'\beta_0 + W_i(\mathbf{x}_i - \bar{\mathbf{x}})'(\beta_1 - \beta_0) + \tau W_i + \eta_i$$

and for inference use the Eicker-Huber-White (EHW) covariance estimator (Eicker (1967); Huber (1967); White (1980). That is, use the $Q \times Q$ element in:

$$\widehat{V(\widehat{\tau})} = N\left(\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\sum_{i=1}^{N}\widehat{\eta}_i^2\mathbf{z}_i'\mathbf{z}_i\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\right)_{Q\times Q}, \tag{8}$$

where $\mathbf{z}_i = (1, \mathbf{x}_i, W_i(\mathbf{x}_i - \bar{\mathbf{x}})', W_i)'$ is a $Q \times 1$ vector and $\widehat{\eta}_i$ is the OLS residual from the regression.

Rerandomization and inference, continued

Zhang and Johansson (2019) proposed model-based Bayesian inference as yet another strategy, that is not restricted to the MD criterion

The strategy is based on the fact that, with rerandomization, treatment allocation only depends on the covariates of the experimental units and does not depend on the outcomes, so imputation of missing outcomes from the posterior distribution conditional on the covariates provides correct inference.

An advantage in comparison to Li and Ding (2020) is that it allows for efficient inference of discrete and censored outcome data after rerandomization.

Rerandomization and stratification

Johansson and Schultzberg (2020) suggested a sampling scheme for choosing the approximate 'best' subset of admissible allocations and also develop a rerandomization covariate balance measure that is easy to use when pre-experimental outcome data (possibly high frequency longitudinal) are available.

Based on this algorithm, Johansson and Schultzberg (2019) show that substantial computational and efficiency gains can be obtained by first stratifying and then finding the "optimal" allocations.

With discrete covariates, rearandomiziation with $a = 0$ can be seen to be equivalent to blocking in a within block balanced design

Rerandomization and stratification

This statement can illustrated using the stratification example above.

With $a = 0$, $M(\mathbf{W}^j, \mathbf{X}) = 0$. This imply $\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0 = 0$. This means these two possible allocations in the example $\mathbf{W}^j = (1, 0, 1, 0, 1, 0, 1, 0)'$ and $1 - \mathbf{W}^j$.

"Block or stratify what you can, rerandomize what you can't"

Rerandomization and "optimal designs", a discussion

The 'optimal' designs suggested by Kallus (2018) includes rerandomization based on the MD as a special case.

The underlying assumption of the proof in Li et al. (2018) is that of replicated randomization and that the sample size N go to infinite. The only thing stochastic is thus the allocation of treatments.

In the optimal design the cardinality of $\mathbb{W}_{Opt}$, $card(\mathbb{W}_{Opt})$, is reduced to a minimum. The implication, is that we cannot derive the asymptotic distribution $\hat{\tau}^{dif}$ as estimator of τ_{FS} (Johansson et al., 2019). For instance, in a deterministic design (i.e. $card(\mathbb{W}_{Opt}) = 1$) the resulting distribution have zero variance.

Note that with a limited set of allocations a FEP has no power.

Rerandomization and "optimal designs", a discussion

The assumption deriving the asymptotic distribution of $\hat{\tau}^{dif}$ as an estimator of τ_{SP} is that of random sampling to the experiment from a population of size N_p ($N_p > N$), or from a superpopulation.

When interest is on conduction inference to PATE there is no lower limit on the number of possible allocations as the asymptotic distribution is derived under the assumption of random sampling to the experiment only.

The consequence is that we, in theory, can have a deterministic design and then conduct inference to τ_{SP}, however no inference is possible regarding an effect in the experiment. This is an anomaly, but it is a consequence of the idea behind Neyman-Pearson inference.

Rerandomization and "optimal designs", a discussion

Schultzberg and Johansson (2020) show that when $card(\mathbb{W}_0) = 2$, in MD rerandomization, the asymptotic distribution of $\hat{\tau}^{dif}$ for inference to τ_{SP} is normally distributed with known variance.

Furthermore, the difference in efficiency compared to using the 'optimal' set, $card(\mathbb{W}_0) = 2$, is typically very small.

This means that, using a slightly larger 'near optimal' set, admits non-degenerate inference to both τ_{FS} and to τ_{SP} without substantially decreasing efficiency in inference to τ_{SP}.

Rerandomization and "optimal designs", a discussion

Lastly, the large sample asymptotic distribution of the $\hat{\tau}^{dif}$ estimator is well approximated by a normal distribution also when a larger 'near optimal' set is used.

The implication is important as the asymptotic inference after MD rerandomization is simplified in contrast to what is suggested in Li et al. (2018).

Bertsimas, D., Johnson, M., and Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 59–82. University California Press, Berkeley, CA.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 221–233. University California Press, Berkeley, CA.

Johansson, P. and Schultzberg, M. (2019). Re-randomization: A complement or substitute for stratification in randomized experiments? Working paper 2019:4, Department of Statistics, Uppsala University.

Johansson, P. and Schultzberg, M. (2020). Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. *Journal of Computational and Graphical Statistics*, 29(4):798–813.

Johansson, P., Schultzberg, M., and Rubin, D. B. (2019). On optimal re-randomization designs. Working paper 2019:3 Department of Statistics, Uppsala University. Forthcoming in JRSS (B).

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(1):85–112.

Kapelner, A., Krieger, A. M., Sklar, M., Shalit, U., and Azriel, D. (2020). Harmonizing optimized designs with classic randomization in experiments. *The American Statistician*, 0(0):1–12.

Krieger, A. M., Azriel, D., and Kapelner, A. (2019). Nearly random designs with greatly improved balance. *Biometrika*, 106(3):695–701.

Lauretto, M. S., Stern, R. B., Morgan, K. L., Clark, M. H., and Stern, J. M. (2017). Haphazard intentional allocation and rerandomization to improve covariate balance in experiments. *AIP Conference Proceedings*, 1853(June).

Li, X. and Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):241–268.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment Ccontrol experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115(37):9157–9162.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282.

Schultzberg, M. and Johansson, P. (2020). Asymptotic inference for optimal rerandomization designs. *Open Statistics*, 1(1):49–58.

White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Revi*, 21(1):149–170.

Zhang, J. and Johansson, P. (2019). A comparison of methods of inference in randomized experiments from a restricted set of allocations. Working paper 2019:5, Department of Statistics, Uppsala University.