Lecture 7
Stratified randomized experiments

Outline

- Stratified randomized experiment
  - Fisher's exact p-value
  - Neyman's repeated sampling approach
  - Regression analysis
- Post stratification
- Suggested reading: Imbens and Rubin Chapter 9.1-9.6, Peng's book Chapter 5

Project STAR: why study class size?

STAR: Student–Teacher Achievement Ratio, Tennessee (1985–1989).

Would smaller classes improve reading and mathematics in the early grades?

Smaller classes require more teachers and classrooms. Policymakers needed credible evidence of their benefits.

A four-year experiment followed children from kindergarten through third grade. Reported cost: about $12 million.

Main finding: children in small classes performed better in reading and mathematics.

Mosteller (1997)

Project STAR: three treatment arms

STAR compared smaller classes with an additional adult in a regular class.

Assigned class type	Students per class	Staffing
Small	13–17	One teacher
Regular	22–25	One teacher
Regular with aide	22–25	Teacher and full-time aide

An aide adds adult support while leaving the class size roughly unchanged.

Does a lower pupil/adult ratio necessarily mean a smaller class?

Mosteller (1997); Imbens & Rubin, Chapter 9

Why randomize both teachers and students?

Both teachers and students were randomly assigned within each school.

Teacher assignment avoids systematically placing stronger or more experienced teachers in small classes.

Student assignment avoids systematic differences in student preparation across class types.

We analyze teacher/class assignments and class-average outcomes. Student randomization supports a class-size interpretation.

What could go wrong if only teachers, or only students, were randomized?

Why randomize within schools?

The 79 schools could differ substantially in students, teachers, and resources.

Each participating school needed enough students for all three class types. First decide how many classes to place in each arm.

Then randomize teachers and students within that school.

Each school is a stratum. Compare small and regular classes from the same school.

Schools contribute different numbers of classes. How should their estimated effects contribute to one overall effect?

The STAR comparison and its units

Analysis sample: 16 schools, 68 classes. Small versus regular classes.

Unit i: a teacher/class. Outcome: the class-average mathematics score.

$W_i = 1$: small class. $W_i = 0$: regular class. Stratum B_i: school.

Students in one class share a teacher and peers. We analyze class averages.

If each class receives equal weight, are we estimating an average effect per class or per student?

The analysis assumes no interference across teacher/class units.

Table 9.1. *Class Average Mathematics Scores from Project Star*

School/ Stratum	No. of 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
(S.D.)		(0.56)	(0.61)

- We focus on two arms (regular classes v.s. small classes) and 16 schools that have at least two classes per arm

STAR over time: practical complications

The intervention continued from kindergarten through third grade.

Complication in STAR	**Statistical issue**
Students switch class types after assignment.	Noncompliance: Treatment received differs from assignment.
Enrollment changes alter actual class sizes.	Treatment variation: The amount of treatment varies.
Students or schools leave, and their outcomes are missing.	Attrition: Missing outcomes can bias comparisons.

Later lectures: noncompliance and instrumental variables (IV).

Methods for attrition and varying treatment intensity are beyond this course.

The stratified assignment mechanism

Pre-treatment strata $B_i \in 1, \ldots, J$. Fix $N_{t,j}$ and $N_{c,j}$ within each stratum j.

$$N_j = N_{t,j} + N_{c,j}, \qquad q_j = \frac{N_j}{N}, \qquad e_j = \frac{N_{t,j}}{N_j}$$

Randomize independently across strata, complete randomization within each stratum.

$$\Pr(\mathrm{W} = \mathrm{w} \mid \mathrm{B}) = \prod_{j=1}^{J} \binom{N_j}{N_{t,j}}^{-1} \quad \text{if } \sum_{i:B_i=j}^{N} w_i = N_{t,j} \text{ for } j = 1, \cdots, J$$

All other assignments have probability zero.

Conditional unconfoundedness within strata

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid B_i, \qquad 0 < e_i = P(W_i = 1 \mid X) < 1$$

Within stratum j, consistency and randomization identify both means:

$$\mathrm{E}[Y_i^{obs} \mid W_i = w, B_i = j] = \mathrm{E}[Y_i(w) \mid B_i = j]$$

$$\tau_{sp,j} = \mathrm{E}[Y_i^{obs} \mid W_i = 1, B_i = j] - \mathrm{E}[Y_i^{obs} \mid W_i = 0, B_i = j]$$

Why do we compare treated and control outcomes within the same stratum?

The target determines the stratum weights

For the fixed study units, let τ_j be the average effect in stratum j.

$$\tau_j = \frac{1}{N_j} \sum_{i:\, B_i = j} [Y_i(1) - Y_i(0)]$$

$$\tau_{fs} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i:\, B_i = j} [Y_i(1) - Y_i(0)] = \sum_{j=1}^{J} \frac{N_j}{N} \tau_j$$

Estimate each stratum's effect, then use the same target weights:

$$\hat{\tau}_j = \overline{Y}_{t,j}^{obs} - \overline{Y}_{c,j}^{obs}, \qquad \hat{\tau}_{strat} = \sum_{j=1}^{J} q_j \hat{\tau}_j, \qquad q_j = \frac{N_j}{N}$$

A weighted estimate across two strata

Illustration: 100 units, with different treatment fractions in the two strata.

Stratum	N_j	$N_{t,j}$	$N_{c,j}$	**Treated mean**	**Control mean**	**Difference**
1	40	10	30	80	70	10
2	60	48	12	60	55	5

$$\hat{\tau}_{strat} = \frac{40}{100} \times 10 + \frac{60}{100} \times 5 = 7$$

Why do the weights use 40 and 60, rather than 10 and 48?
What would the equal-stratum average 7.5 estimate?

Simpson's paradox: a reversal from pooling

$$\hat{\tau}_{pool} = \sum_{j=1}^{J} \frac{N_{t,j}}{N_t} \overline{Y}_{t,j}^{obs} - \sum_{j=1}^{J} \frac{N_{c,j}}{N_c} \overline{Y}_{c,j}^{obs}$$

The treated and control means generally use different stratum weights.

$$\hat{\tau}_{pool} = \frac{10\times80 + 48\times60}{58} - \frac{30\times70 + 12\times55}{42} \approx -2.27$$

The same data gave a weighted within-stratum estimate of +7.

$$e_j = e \text{ for every j} \;\Rightarrow\; \frac{N_{t,j}}{N_t} = \frac{N_{c,j}}{N_c} = q_j \;\Rightarrow\; \hat{\tau}_{pool} = \hat{\tau}_{strat}$$

Simpson’s paradox: the kidney-stone table

Stone size	Treatment A successes	Treatment B successes
Small	81 / 87 = 93.1%	234 / 270 = 86.7%
Large	192 / 263 = 73.0%	55 / 80 = 68.8%
Combined	273 / 350 = 78.0%	289 / 350 = 82.6%

Why does pooling favor B when A does better in both strata?

$$\frac{357}{700}\left(\frac{81}{87}-\frac{234}{270}\right)+\frac{343}{700}\left(\frac{192}{263}-\frac{55}{80}\right)\approx 0.0537$$

Standardizing to the same stone-size mix gives +5.37 percentage points.
A causal interpretation still requires within-stratum comparability.

Unbiasedness and alternative targets

$$\mathrm{E}_W(\hat{\tau}_j) = \tau_j \;\Rightarrow\; \mathrm{E}_W(\hat{\tau}_{strat}) = \sum_{j=1}^{J} q_j \tau_j = \tau_{fs}$$

Target average	**Weight on stratum j**
Each study unit equally	$q_\mathrm{j} = N_\mathrm{j}/N$
Each stratum equally	$1/J$
A specified population mixture	π_j = target population share

If effects differ across strata, can changing the weights leave the estimand unchanged?

Fisher's test: the reference assignments

$$H_0\colon\ Y_i(1) = Y_i(0) \text{ for every i}, \qquad T_q = \left| \sum_{j=1}^{J} q_j \hat{\tau}_j \right|$$

Keep outcomes fixed under H_0. Shuffle treatment labels within each stratum.

$$|\Omega_B| = \prod_{j=1}^{J} \binom{N_j}{N_{t,j}}, \qquad p = \frac{\#\{w \in \Omega_B\colon\ T(w) \geq T_{obs}\}}{|\Omega_B|}$$

Monte Carlo: if b of M simulated statistics are at least as extreme,

$$\hat{p}_{MC} = \frac{b+1}{M+1}$$

Weights for a Fisher test statistic

- Generally, the test statistics has the form:

$$T_\lambda = |\sum_{j=1}^{J} \lambda_j \hat{\tau}_j|, \qquad \sum_{j=1}^{J} \lambda_j = 1$$

- Under the sharp null, with a common outcome variance S^2 across strata:

$$\mathrm{Var}_W(\hat{\tau}_j) = S^2(\frac{1}{N_{t,j}} + \frac{1}{N_{c,j}}) = \frac{S^2}{N_j e_j (1 - e_j)}$$

- **"inverse-variance-weighting":** weights $\lambda_j = \frac{N_j e_j (1-e_j)}{\sum_{k=1}^{J} N_k e_k (1-e_k)}$ to increase precision

- Precision weights can be useful for testing. SATE estimation still uses $q_j = N_j/N$.

Estimation weights and Fisher test statistics

For SATE estimation, the target determines the weights.

$$\hat{\tau}_{strat} = \sum_{j=1}^{J} q_j \hat{\tau}_j \,, \qquad q_j = \frac{N_j}{N}$$

For Fisher testing, the assignment mechanism determines validity.

Pooled and stratified statistics both give valid tests with within-stratum permutations.
Their relative power depends on the alternative and outcome variation.

Connection between pooling and precision weighting:

- Under the sharp null, after centering the pooled statistic:

$$\hat{\tau}_{pool} - \mathrm{E}_0(\hat{\tau}_{pool}) = \frac{N}{N_t N_c} \sum_{j=1}^{J} N_j e_j (1 - e_j) \hat{\tau}_j$$

- The weights are proportional to inverse variance weights

Can a biased estimator still be a valid test statistics?

Fisher tests in Project STAR

Reported two-sided p-values under within-school randomization:

Statistic	**Reported p-value**
Stratum differences weighted by N_j/N	0.034
Stratum differences with precision weights	0.023
Pooled difference in means	0.025
Difference in within-school ranks	0.15

- Difference in within-school ranks
  - Get R_i^{strat} as the within-strata rank of each individual i (definition page 196 of Imbens and Rubin's book)
  - Average difference of within-strata ranks between treatment and control

$$|\bar{R}_t^{\text{strat}} - \bar{R}_c^{\text{strat}}|$$

Neyman analysis within one stratum

Lecture 5 applies separately within each stratum:

$$\hat{\tau}_j = \overline{Y}_{t,j}^{obs} - \overline{Y}_{c,j}^{obs}, \qquad \mathrm{E}_W(\hat{\tau}_j) = \tau_j$$

$$\mathrm{Var}_W(\hat{\tau}_j) = \frac{S_{t,j}^2}{N_{t,j}} + \frac{S_{c,j}^2}{N_{c,j}} - \frac{S_{ct,j}^2}{N_j}$$

$$\hat{V}_j = \frac{s_{t,j}^2}{N_{t,j}} + \frac{s_{c,j}^2}{N_{c,j}}, \qquad \mathrm{E}_W(\hat{V}_j) - \mathrm{Var}_W(\hat{\tau}_j) = \frac{S_{ct,j}^2}{N_j} \geq 0$$

S^2: potential-outcome variances within stratum j. s^2: observed within-arm sample variances.

Combining the stratum variances

Assignments are independent across strata. Potential outcomes are fixed.

$$\mathrm{Var}_W(\hat{\tau}_{strat}) = \mathrm{Var}_W\left(\sum_{j=1}^{J} q_j \hat{\tau}_j\right) = \sum_{j=1}^{J} {q_j}^2 \mathrm{Var}_W(\hat{\tau}_j)$$

$$\hat{V}_{strat} = \sum_{j=1}^{J} {q_j}^2 \hat{V}_j$$

$$\mathrm{E}_W(\hat{V}_{strat}) - \mathrm{Var}_W(\hat{\tau}_{strat}) = \sum_{j=1}^{J} {q_j}^2 \frac{S_{ct,j}^2}{N_j} \geq 0$$

Worked example: weights and uncertainty

Stratum	q_j	Effect estimate	Variance estimate	SE
1	0.4	10	4	2
2	0.6	5	1	1

$$\hat{\tau}_{strat} = 0.4\times10 + 0.6\times5 = 7$$

$$\hat{V}_{strat} = 0.4^2\times4 + 0.6^2\times1 = 1, \qquad \mathrm{SE} = \sqrt{1} = 1$$

Pair check: why is 0.4 × 2 + 0.6 × 1 = 1.4 not the standard error of this estimator?

Confidence intervals and average-effect tests

$$\text{CI: } \hat{\tau}_{strat} \pm z_{1-\alpha/2}\sqrt{\hat{V}_{strat}}$$

$$H_0\colon \tau_{fs} = \tau_0, \qquad Z = \frac{\hat{\tau}_{strat} - \tau_0}{\sqrt{\hat{V}_{strat}}}, \qquad p \approx 2\{1 - \Phi(|Z|)\}$$

Example: 7 ± 1.96 × 1 = [5.04, 8.96].

When is normal approximation good?
A few large strata, or many independent strata with no stratum dominating the variance.

Project STAR: effect estimate and precision

Analysis	Estimate	Reported SE
Weighted within-school differences	0.241	0.092
Pooled comparison, ignoring strata	0.224	0.141

$$\text{Stratified } 95\% \text{ CI: } \; 0.241 \pm 1.96 \times 0.092 \approx [0.061, 0.421]$$

School adjustment changes both the point estimate and its uncertainty.

How chance imbalance affects the estimate

Two strata: 50 students would score 40 without treatment, and 50 would score 60.
Treatment adds 5 points to every student's score.

One possible assignment under complete randomization (CR):

Treatment group: 50 students

20 lower-score students score 45.
30 higher-score students score 65.

$$\text{Mean} = \frac{20\times45 + 30\times65}{50} = 57$$

Control group: 50 students

30 lower-score students score 40.
20 higher-score students score 60.

$$\text{Mean} = \frac{30\times40 + 20\times60}{50} = 48$$

Estimated effect = 57 − 48 = 9. True effect = 5.

More higher-score students in treatment add 4 points to the difference.

Stratified randomization (STR): put 25 students from each stratum in each group.
Treatment mean = (45 + 65)/2 = 55. Control mean = (40 + 60)/2 = 50.
The estimated effect is 5 for every allowed assignment in this example.

Precision gain: comparing assignment designs

$e = N_t/N$ is the treated fraction. STR fixes $\frac{N_{t,j}}{N_j} = e$ in every stratum.

$q_j = N_j/N$ is the proportion of study units in stratum j.

$$\text{Stratum mean difference: } \Delta_{wj} = \overline{Y}_j(w) - \overline{Y}(w)$$

$$\text{Var}_{CR}(\hat{\tau}_{pool}) - \text{Var}_{STR}(\hat{\tau}_{strat})$$

$$\approx \frac{1}{N}\sum_j q_j\{\sqrt{\frac{1-e}{e}}\Delta_{1j} + \sqrt{\frac{e}{1-e}}\Delta_{0j}\}^2 \geq 0$$

Slide 26: $e = 1/2, q_1 = q_2 = 1/2$, and $\Delta_{1j} = \Delta_{0j} = \pm 10$.

$$\text{Gain} \approx \frac{4}{N}\sum_j q_j\Delta_{0j}^2 = \frac{4}{100}(0.5\times 100 + 0.5\times 100) = 4$$

Equal stratum sizes can still give a gain when stratum means differ.

Regression with stratum-specific effects

$D_{\mathrm{ij}} = 1\{B_{\mathrm{i}} = j\}$. Randomization links the two conditional means:

$$\mathrm{E}[Y_i(w) \mid B_i = j] = \alpha_j + \tau_{sp,j} w$$

$$Y_i^{obs} = \sum_{j=1}^{J} D_{ij}(\alpha_j + \beta_j W_i) + \varepsilon_i$$

$$\hat{\beta}_j = \overline{Y}_{t,j}^{obs} - \overline{Y}_{c,j}^{obs} = \hat{\tau}_j, \qquad \hat{\tau}_{strat} = \sum_{j=1}^{J} q_j \hat{\beta}_j$$

Use the HC2 variance of this weighted coefficient contrast.
Without extra covariates, it equals the stratified Neyman variance.

An additive regression has implicit weights

$$Y_i^{obs} = \alpha_{B_i} + \beta W_i + \varepsilon_i$$

The solution can be written as

$$\hat{\beta} = \frac{\sum_{j=1}^{J} N_j e_j (1 - e_j) \hat{\tau}_j}{\sum_{j=1}^{J} N_j e_j (1 - e_j)}$$

These equal q_j weights when the treatment fraction is common across strata.

When fractions differ, does merely adding stratum indicators guarantee the sample average treatment effect?

Post-stratification: conditioning on counts

After complete randomization, form strata using baseline covariates.

Condition on C, the vector of realized treated counts across strata.

$$\Pr(\mathrm{W} = \mathrm{w} \mid C) = \frac{\binom{N}{N_t}^{-1}}{\prod_{j=1}^{J} \binom{N_j}{N_{t,j}} \Big/ \binom{N}{N_t}} = \prod_{j=1}^{J} \binom{N_j}{N_{t,j}}^{-1}$$

The conditional assignment law is the same as in a stratified experiment.

Use the same weighted estimate, variance, and within-stratum permutations.

Baseline strata, fixed independently of outcomes. Avoid empty or very small arm-by-stratum cells.

Post-stratification by age: a numerical example

Outcome: survival. Treatment: tolbutamide assignment. Control: placebo.

Age stratum	Treated survivors	Control survivors	Effect estimate	SE
Under 55	98 / 106	115 / 120	−0.034	0.031
55 or older	76 / 98	69 / 85	−0.036	0.060

$$\hat{\tau}_{strat} = \frac{226}{409} \times (-0.034) + \frac{183}{409} \times (-0.036) \approx -0.035$$

$$\text{SE} \approx \sqrt{\frac{226^2}{409} \times 0.031^2 + \frac{183^2}{409} \times 0.060^2} \approx 0.032$$

Pooled estimate: −0.045 (SE 0.033). Displayed inputs are rounded. Ding, §5.4.1

Computing the weighted estimate in R

```
a <- t(sapply(split(d, d$B), function(g) {
  yt <- g$Y[g$W == 1]; yc <- g$Y[g$W == 0]
  stopifnot(length(yt) >= 2, length(yc) >= 2)
  c(n = nrow(g), effect = mean(yt) - mean(yc),
    v = var(yt)/length(yt) + var(yc)/length(yc))
}))
q <- a[, "n"] / sum(a[, "n"])
est <- sum(q * a[, "effect"])
se <- sqrt(sum(q^2 * a[, "v"]))
est + c(-1, 1) * 1.96 * se
```

Which line changes if the target gives equal weight to every stratum?

What does stratification add?

The assignment mechanism preserves treated counts within each stratum.

$$\hat{\tau}_{strat} = \sum_{j=1}^{J} q_j \hat{\tau}_j \,, \qquad \hat{V}_{strat} = \sum_{j=1}^{J} {q_j}^2 \hat{V}_j$$

Exit check:

1. Why does SATE use N_j/N weights?
2. When does a pooled mean difference equal the stratified estimate?
3. Why can a biased SATE estimator still be a valid Fisher statistic?
4. What must we condition on for post-stratification?