Lecture 7
Stratified randomized experiments

Outline

- Stratified randomized experiment
  - Fisher’s exact p-value
  - Neyman’s repeated sampling approach
  - Regression analysis
- Post stratification
- HOMEFOOD case study

STAR (Student-Teacher Achievement Ratio) Project in Tennessee

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- What is STAR? (1985-1989)
  - A large-scale, four-year, longitudinal, experimental study of reduced class size
  - One the historically most important educational investigations
  - Cost of about $12 million

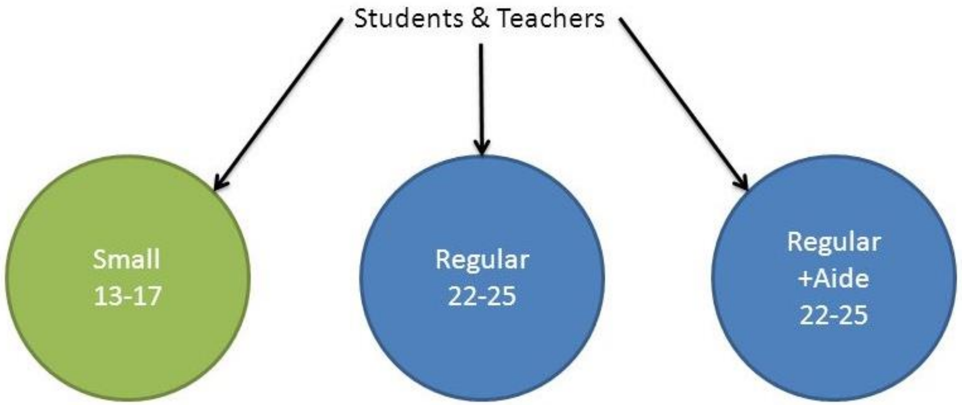
  - Conclusion: small classes have an advantage over larger classes in reading and math in the early primary grades

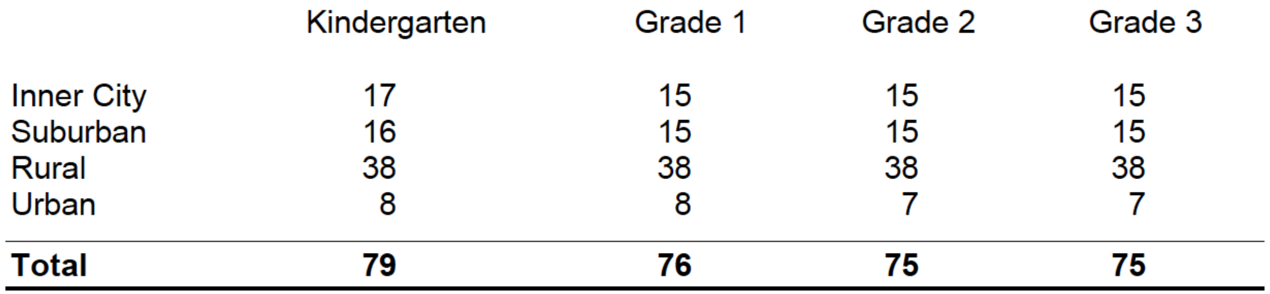
- Why was STAR needed?
  - Legislators and school administrators doubted the significance of smaller classes
  - Conducted at the elementary-school level as this is where the foundation is laid for children’s success in school.
  - The most credible study of class size

STAR (Student-Teacher Achievement Ratio) Project in Tennessee

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- How is the experiment designed?
  - Three levels of “treatment”: three types of classes
  - All schools are invited to participate
  - The study included 79 schools resulting in over 6,000 students per grade
    - A school need to have a minimum of 57 students in kindergarden (at least one for each type of class)
  - Once a school is admitted, a decision was made on the number of classes per arm
    - Difference between Class Size and Pupil/Teacher Ratio
    - The interventions were initiated as the students entered school in kindergarten and continued through third grade.





	Kindergarten	Grade 1	Grade 2	Grade 3
Inner City	17	15	15	15
Suburban	16	15	15	15
Rural	38	38	38	38
Urban	8	8	7	7
Total	**79**	**76**	**75**	**75**

The project STAR example

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- Stratified randomization procedure
  - potentially large differences in resources, teachers and students between schools
  - Randomization within each school
    - Students and teachers were randomly assigned to the one of the 3 arms
    - The unit is a teacher in a class, instead of a student to avoid violation of no interference assumption
- Practical issues faced in real experiment
  - Longitudinal experiment
    - Schools may drop out of the project
    - Classes may gain/lose students so that can become too small or too big
  - Selection bias in students’ involvement
    - Students’ parents were informed so may want their children to be in the smaller class

The project STAR example

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- Understanding the randomization procedure
  - Two randomizations happen in the experiment
    - Randomization of teachers
    - Randomization of students
  - Our causal analysis only relies on the randomization of teachers
    - The treatment effect on a particular teacher in a particular school is comparing the test score of being randomly assigned to a type of class and the test score of being randomly assigned to another type of class
  - The randomization of students helps interpretating our results
    - Treatment effect between two arms can be explained by the classroom size difference instead of the systematic differences of students

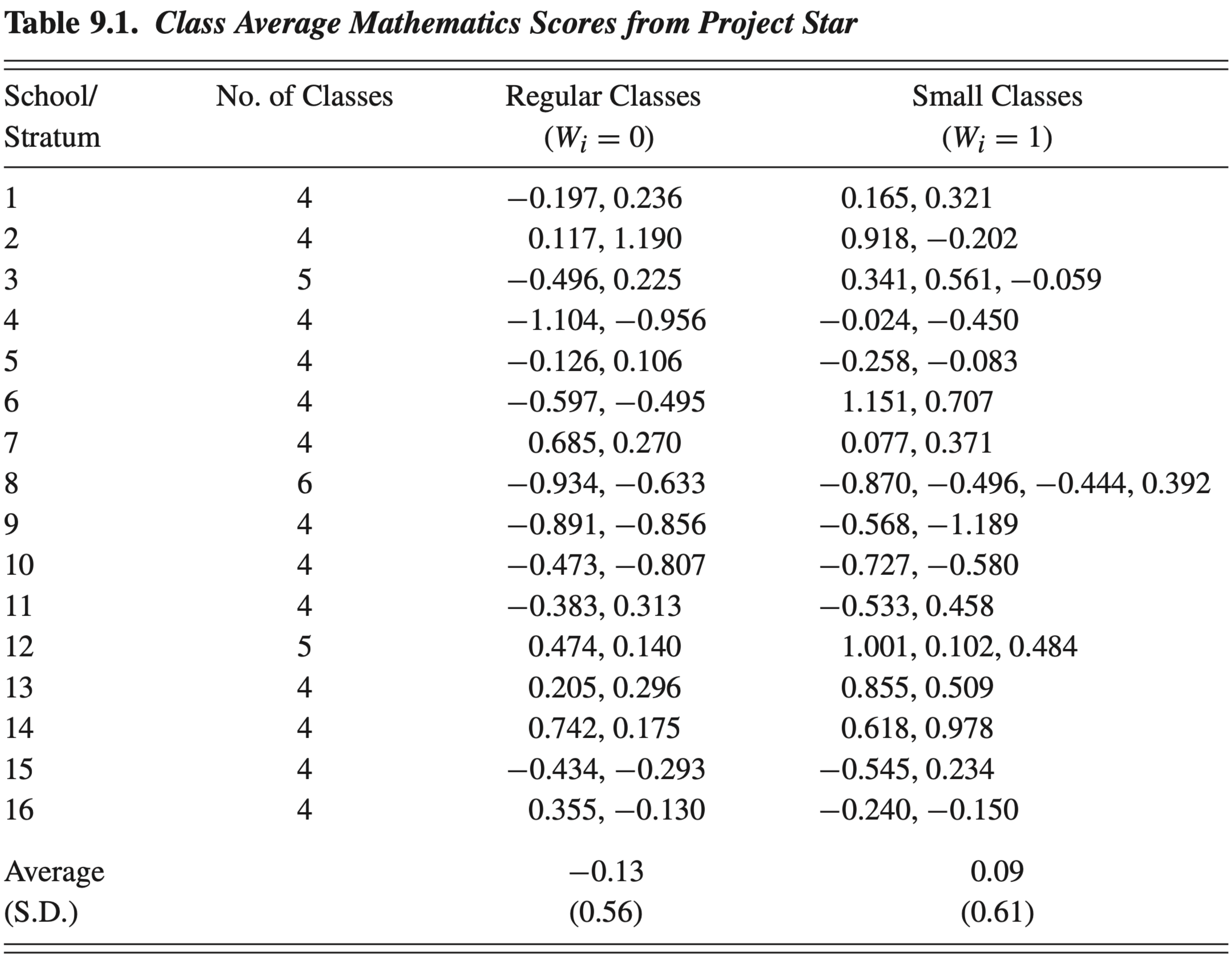


Table 9.1. *Class Average Mathematics Scores from Project Star*

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 (S.D.)		−0.13 (0.56)	0.09 (0.61)

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

Stratified randomized experiment

- Basic procedure:
  1. Blocking (Stratification): create groups of similar units based on pre-treatment covariates, let $B_i \in \{1, \cdots, J\}$ be the block indicator
  2. Block (Stratified) randomization: completely randomize treatment assignment within each group
- Blocking can improve the efficiency by minimizing the variance of the potential outcomes within each strata

"Block what you can and randomize what you cannot"

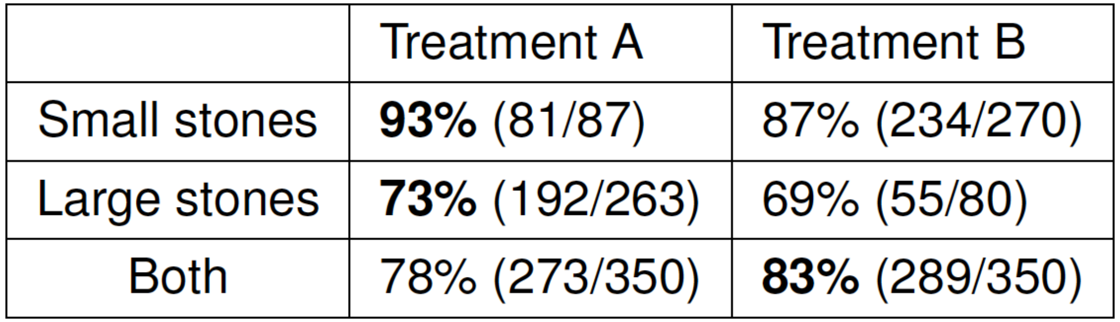
Box, et al. (2005). Statistics for Experimenters. 2nd eds. Wiley

- Assignment probability

$$P(\boldsymbol{W} = \boldsymbol{w} | \boldsymbol{X}) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \text{ if } \sum_{i:B_i=j}^{N} w_i = N_t(j) \text{ for } j = 1, \cdots, J \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

Compare treated v.s. control? Simpson’s paradox

- Compare the success rates of two treatment of kidney stores
- Treatment A: open surgery; treatment B: small puctures



	Treatment A	Treatment B
Small stones	**93%** (81/87)	87% (234/270)
Large stones	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

- Large difference in treatment assignment probability across strata
  - Small stone: assignment probability $\frac{87}{87+270} = 0.24$
  - Large stone: assignment probability is $\frac{263}{263+80} = 0.77$
- Compare within each strata and take a weighted average:
  - True average causal effect: $83.2\% - 78.2\% : (93\% - 87\%) \times 0.51\ - (73\% -\ 69\%) \times 0.49$

Fisher’s exact p-value

- We still focus on the **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, N$

- Choice of test statistics:
  Denote sample means for every strata / block

$$\overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j) = \frac{1}{N_{\mathrm{c}}(j)} \sum_{i:G_i=j} (1 - W_i) \cdot Y_i^{\mathrm{obs}}, \quad \overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) = \frac{1}{N_{\mathrm{t}}(j)} \sum_{i:G_i=j} W_i \cdot Y_i^{\mathrm{obs}}$$

  - Weighted combination of group mean differences across blocks

$$T^{\mathrm{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot \left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j) \right) \right|$$

    - Weights based on relative sample size $\lambda(j) = \frac{N(j)}{N}$
      sample difference is more accurate in larger strata
    - **“inverse-variance-weighting”:** assume that per-strata potential outcomes sample variances $S_c^2(j) \equiv S_t^2(j) \equiv S^2$ for all j, then under stratified randomization

$$\mathbb{V}_W\left[\bar{Y}_t^{\mathrm{obs}}(j) - \bar{Y}_c^{\mathrm{obs}}(j) | \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right] = S^2 \left(\frac{1}{N_c(j)} + \frac{1}{N_t(j)} \right)$$

Fisher’s exact p-value

- We still focus on the **Sharp null:** $H_0 : Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, N$

- Choice of test statistics:

  Denote sample means for every strata / block

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

  - Weighted combination of group mean differences across blocks

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

    - Weights based on relative sample size $\lambda(j) = \frac{N(j)}{N}$
      sample difference is more accurate in larger strata
    - **“inverse-variance-weighting”:** weights

$$\lambda(j) = \frac{1}{\left(\frac{1}{N_c(j)} + \frac{1}{N_t(j)}\right)} / \sum_{k=1}^{J} \frac{1}{\left(\frac{1}{N_c(k)} + \frac{1}{N_t(k)}\right)}$$

Fisher’s exact p-value

- Can we simply use the two-sample mean difference statistic $T = \left|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right|$?
  - This is still one test statistic and we will still get valid Fisher’s exact p-value if we follow the stratified randomization procedure to generate the reference distribution

  Simpson’s paradox:

  - We may not always get small value of T even wen the sharp null is true
    - Example:
      $Y_i(0) \equiv Y_i(1) = 1$ for strata 1 and $Y_i(0) \equiv Y_i(1) = 2$ for strata 2,
      $N_c(1) = N_t(1) = 5$, $N_c(2) = 15$ and $N_t(2) = 5$
      Then $\bar{Y}_t^{\text{obs}} = 1.5$ and $\bar{Y}_c^{\text{obs}} = 1.75$
  - Power of the Fisher’s test is affected

Fisher's exact p-value and the project STAR

- Choice of test statistics:
  - Rank-based statistics
    - 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

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

- Calculate the null distribution of test statistics
  - Randomly simulate treatment assignments following the same stratified randomization
- Project STAR results
  - P-values for the first 3 are similar as most schools have 4 classes
  - Large p-value for rank-based statistics as # classes too few in most schools

Test statistics	P-value
Weights $\lambda(j) = \frac{N(j)}{N}$	0.034
"inverse-variance-weighting"	0.023
$\left\|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right\|$	0.025
Rank-based statistics	0.15

Neyman's repeated sampling approach

- Target: PATE or SATE $\tau = \sum_j \frac{N(j)}{N} \tau(j)$ where $\tau(j)$ is the PATE or SATE for strata j

- Analysis procedure

1. Apply Neyman's analysis to each strata / block

$$\hat{\tau}^{\text{dif}}(j) = \overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j), \quad \text{and} \quad \hat{\mathbb{V}}^{\text{neyman}}(j) = \frac{s_{\text{c}}(j)^2}{N_{\text{c}}(j)} + \frac{s_{\text{t}}(j)^2}{N_{\text{t}}(j)}$$

   - Variance estimator is conservative within each strata as discussed before

2. Aggregate block-specific estimates and variances

$$\hat{\tau}^{\text{strat}} = \sum_j \frac{N(j)}{N} \hat{\tau}^{\text{dif}}(j), \qquad \hat{\mathbb{V}}(\hat{\tau}^{\text{strat}}) = \sum_j \left(\frac{N(j)}{N}\right)^2 \hat{\mathbb{V}}^{\text{neyman}}(j)$$

   - Both treatment assignments and potential outcomes are independent across strata

3. Statistical inference
   - Use normal approximation of the distribution of $\hat{\tau}^{\text{strat}}$
   - Normal approximation works as long as N is large enough
     - Either small strata size with many strata or large strata size with few strata

Power gain in Neyman's approach after stratification

- Variance decomposition

$$\underbrace{\mathbb{V}(X)}_{\text{total variance}} = \underbrace{\mathbb{E}\{\mathbb{V}(X \mid Y)\}}_{\text{within}-\text{block variance}} + \underbrace{\mathbb{V}\{\mathbb{E}(X \mid Y)\}}_{\text{across}-\text{block variance}}$$

- Assume that the treatment proportion $\frac{N(j)}{N}$ is the same across all strata
  - Then $\hat{\tau}^{\text{dif}} = \hat{\tau}^{\text{strat}}$

- $\mathbb{V}_{\text{complete}}(\hat{\tau}^{\text{dif}}) - \mathbb{V}_{\text{stratified}}(\hat{\tau}^{\text{strat}}) \geq 0$
  - Intuitively, we do not need to consider noise due to heterogeneity across blocks
  - For a rigorous proof, see Peng's book section 5.3.3

- Result in the project STAR
  - $\hat{\tau}^{\text{strat}} = 0.241$, $\widehat{\mathbb{V}}(\hat{\tau}^{\text{strat}}) = 0.092^2$
  - (In correct) if we analyze as if it is a completely randomized experiment
    - $\hat{\tau}^{\text{dif}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} = 0.224$ can be a biased estimate for τ
    - $\widehat{\mathbb{V}}(\hat{\tau}^{\text{dif}}) = 0.141^2$ larger standard deviation

Linear regression

- Run separate linear regressions within each strata
  - Does not work if each strata size is too small

- Denote $B_i(j)$ as the indicator variable of whether sample i belong to strata j
- If there are no covariates, equivalently, we can write separate linear regression models into a joint regression model

$$Y_i^{\text{obs}} = \alpha_j + \tau(j)W_i + \varepsilon_i$$

  - The underlying model for the potential outcomes

$$\mathbb{E}[Y_i(w)|\{B_i(j), j = 1, \cdots, J\}] = \alpha_j + \tau(j)w$$

  - Average causal effect for strata j is $\tau(j)$
  - The strata indicators $B_i(j)$ are treated as pre-treatment covariates
  - We need to adjust for the strata indicators as we only have conditional independence

$$\left(\boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) \perp \boldsymbol{W} \mid \boldsymbol{B}(j)$$

  - The homoscedastic error assumption for the joint model is assuming that

$$\mathbb{V}[Y_i(0)|\{B_i(j), j = 1, \cdots, J\}] = \mathbb{V}[Y_i(1)|\{B_i(j), j = 1, \cdots, J\}] = \sigma^2$$

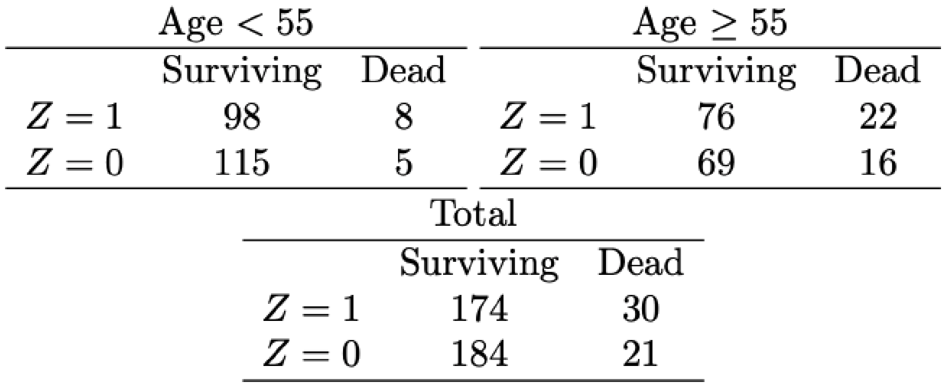
Post-stratification

- In a completely randomized experiment, each assignment vector has the sample probability ($P(\boldsymbol{W} = \boldsymbol{w})$) if $\sum_{i=1}^{N} w_i = N_t$
- If we focus on a subgroup S, conditional on $N_{t,S} = \sum_{i \in S} W_i$, the assignment vector for the individuals in the subgroup also has the same probability ($P(\boldsymbol{W}_S = \boldsymbol{w}_S)$) if $\sum_{i \in S} w_i = N_{t,S}$
- So conditional on $N_{t,S}$, we can treat the treatment assignment as from a completely randomized experiment also for the subgroup

- Post-stratification (Miratrix. et al. 1971. J. Royal Stat. Soc. B.)
  - Blocking after the experiment is conducted
  - Analyze the experiment as from a stratified randomized experiment by conditioning on $N_{t,S}$ for each strata S
  - By post-stratification, we can stratify individuals into relatively homogenous subpopulations
  - Post-stratification is nearly as efficient as pre-randomization blocking

Meinert et. al. (1970)'s example

- A completely randomized experiment.
- Treatment is tolbutamide ($Z = 1$) and control is a placebo ($Z = 0$)
- Causal effect: difference in the survival probability

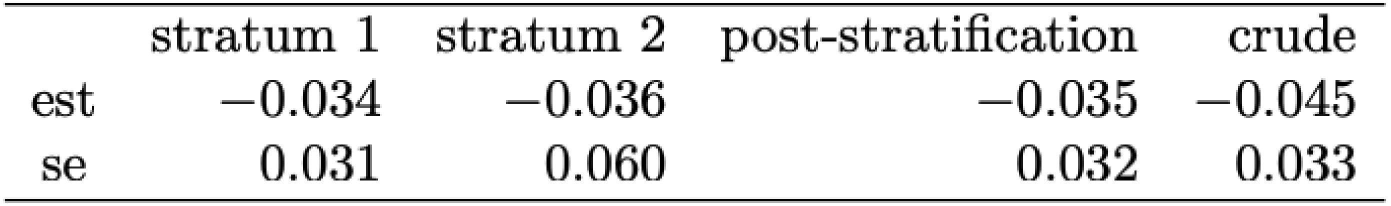


	Age < 55			Age ≥ 55	
	Surviving	Dead		Surviving	Dead
$Z = 1$	98	8	$Z = 1$	76	22
$Z = 0$	115	5	$Z = 0$	69	16

	Total	
	Surviving	Dead
$Z = 1$	174	30
$Z = 0$	184	21

Peng's book Section 5.4.1

- Subgroup and sample average estimates with post-stratification



	stratum 1	stratum 2	post-stratification	crude
est	-0.034	-0.036	-0.035	-0.045
se	0.031	0.060	0.032	0.033

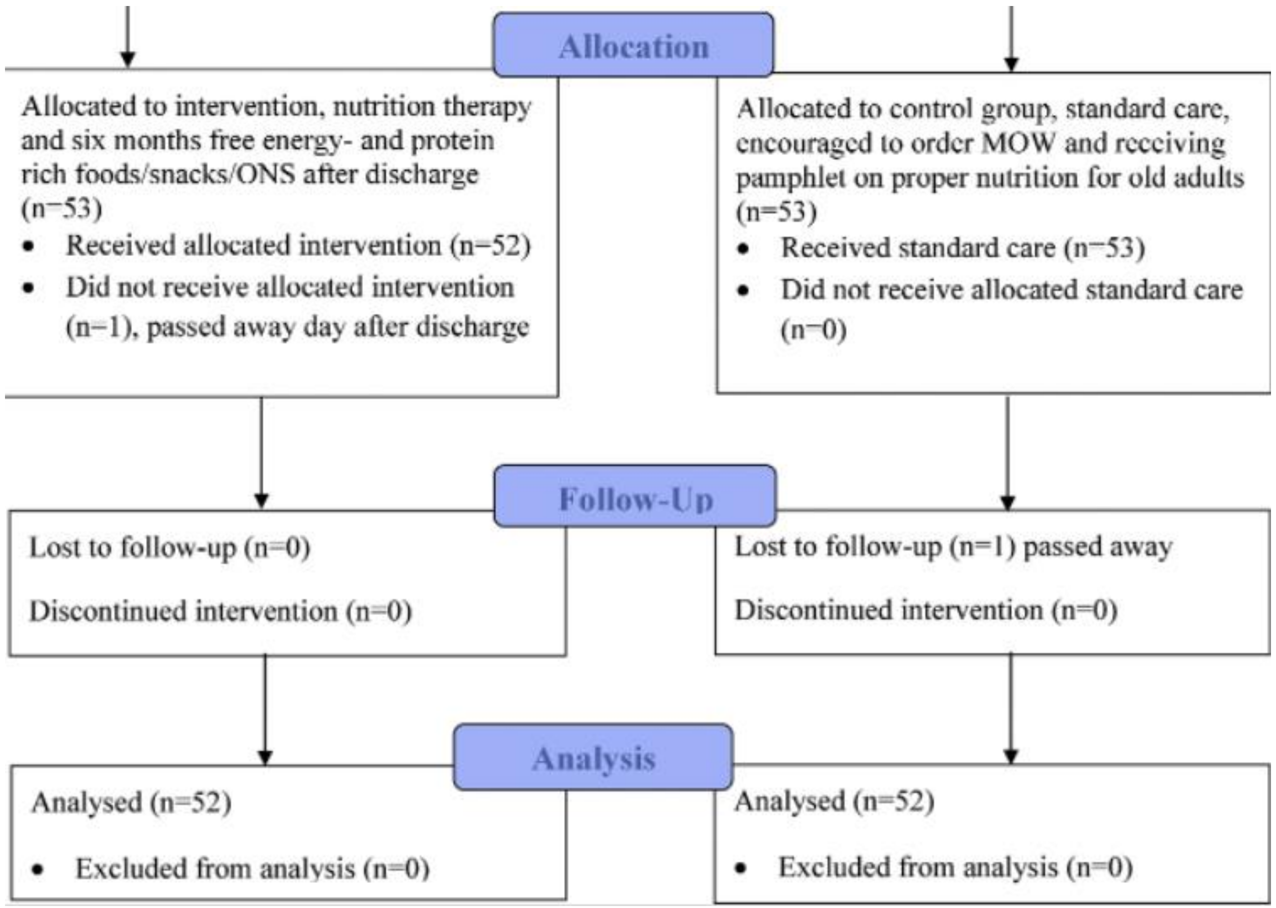
Case study: Analysis of HOMEFOOD randomized trial

[HOMEFOOD randomised trial–Six-month nutrition therapy improves quality of life, self-rated health, cognitive function, and depression in older adults after hospital discharge. *Clinical Nutrition ESPEN (2022)*.]

- Goal: investigate the effect of nutrition therapy on health-related quality of life
- Participants: Eligible participants were community dwelling patients discharging home from the hospital within 24 h, aged ≥65 years, and at risk for malnutrition
- Randomization: participants were randomly allocated (ratio = 1:1) to either the intervention or the control group by using a random number generated by the principal investigator
- Intervention: nutrition therapy from a clinical nutritionist consists of 5 home visits, 3 telephone calls, free supplemental energy- and protein-rich foods

Case study: Analysis of HOMEFOOD randomized trial

[HOMEFOOD randomised trial–Six-month nutrition therapy improves quality of life, self-rated health, cognitive function, and depression in older adults after hospital discharge. *Clinical Nutrition ESPEN (2022)*.]



Allocation

Allocated to intervention, nutrition therapy and six months free energy- and protein rich foods/snacks/ONS after discharge (n=53)

- Received allocated intervention (n=52)
- Did not receive allocated intervention (n=1), passed away day after discharge

Allocated to control group, standard care, encouraged to order MOW and receiving pamphlet on proper nutrition for old adults (n=53)

- Received standard care (n=53)
- Did not receive allocated standard care (n=0)

Follow-Up

Lost to follow-up (n=0)

Discontinued intervention (n=0)

Lost to follow-up (n=1) passed away

Discontinued intervention (n=0)

Analysis

Analysed (n=52)

- Excluded from analysis (n=0)

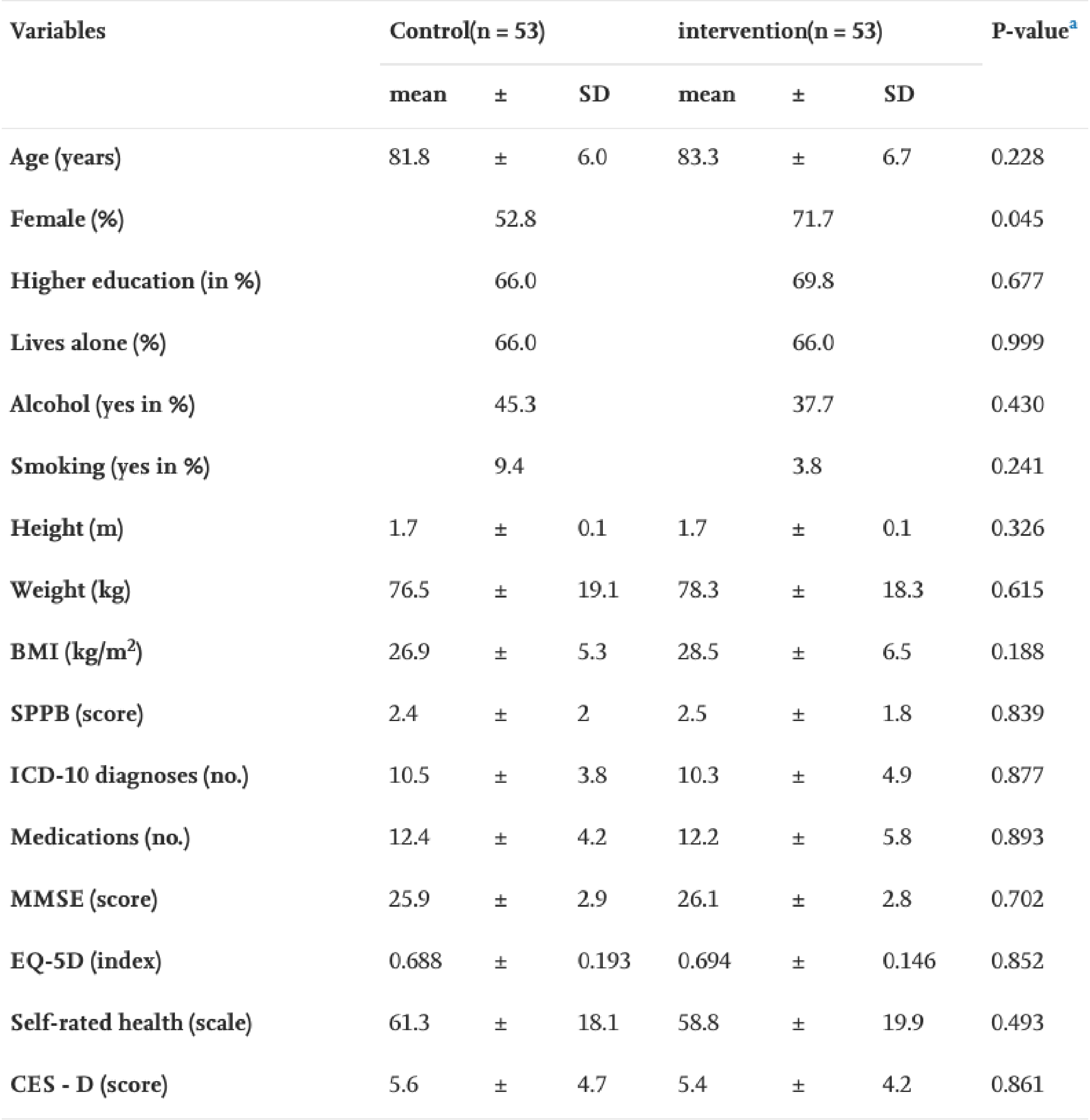
Analysed (n=52)

- Excluded from analysis (n=0)

- Non-compliance is a common issue in randomized experiments

- In this example, reasons that patients dropout are likely unrelated to the treatment

- Our analysis will be based on the N = 104 individuals



Variables	Control(n = 53)			intervention(n = 53)			P-value[a]
	mean	±	SD	mean	±	SD	
Age (years)	81.8	±	6.0	83.3	±	6.7	0.228
Female (%)		52.8			71.7		0.045
Higher education (in %)		66.0			69.8		0.677
Lives alone (%)		66.0			66.0		0.999
Alcohol (yes in %)		45.3			37.7		0.430
Smoking (yes in %)		9.4			3.8		0.241
Height (m)	1.7	±	0.1	1.7	±	0.1	0.326
Weight (kg)	76.5	±	19.1	78.3	±	18.3	0.615
BMI (kg/m^2)	26.9	±	5.3	28.5	±	6.5	0.188
SPPB (score)	2.4	±	2	2.5	±	1.8	0.839
ICD-10 diagnoses (no.)	10.5	±	3.8	10.3	±	4.9	0.877
Medications (no.)	12.4	±	4.2	12.2	±	5.8	0.893
MMSE (score)	25.9	±	2.9	26.1	±	2.8	0.702
EQ-5D (index)	0.688	±	0.193	0.694	±	0.146	0.852
Self-rated health (scale)	61.3	±	18.1	58.8	±	19.9	0.493
CES - D (score)	5.6	±	4.7	5.4	±	4.2	0.861

- We still want to check for covariates balancing even in randomized experiment

- If some covariates are not balanced, our analysis is still valid, but our conclusion can be very inaccurate

- Here sex is not balanced well, one solution is to use post-stratification and estimate causal effect on female and male groups separately

- Equivalently, we may also want to add sex as a covariate in linear regression

- Check R example 3 for data analysis