Problem Set 2

STAT II (Spring 2025)

Disclaimer: *Please read the guidelines below carefully. Good luck!*

Guidelines

- Please upload your answers to Brightspace by **Thursday, March 6th, at 11:59 PM**.
- Direct all questions about the problem set to Brightspace under Contents > Discussions > Class Feed > Problem Set 2.
- For precise and expedited responses, we will address questions about this problem set on Brightspace until 6:00 PM on March 5rd.
- Collaboration is allowed, but I encourage you to attempt the problems on your own before seeking help from others. Regardless of collaboration, you must individually write up and submit your answers.
- Late submissions will not be accepted unless prior approval is obtained from the instructor at least one day before the due date.
- **The total points for the problem set are 100**. Individual bonus questions may appear. If your total points exceed 100, the excess points will carry over to subsequent problem sets (e.g., if you earn 120 points, 20 points will be added to future problem sets).
- **Grading:** Show every step of your derivations. We grade the steps of derivations as well as the final answers. If you are unable to solve the problem completely, partial credit can be given for your derivations. Conversely, a correct final answer without complete derivations may not receive full credit.
- **Stylistic Requirements:** Adhere to the guidelines for the format of your submitted answers. Ensure you follow these rules:
  1. Submit your answers as a PDF file compiled from LaTeX, R Markdown, or (ideally) a Quarto Markdown document.
  2. If using raw LaTeX(.tex) for your answers, submit an accompanying .R file for any computational tasks, with referenced line numbers corresponding to each specific task.
  3. If using R Markdown (.Rmd) or Quarto Markdown (.qmd), include your code as code chunks in the source file. Additionally, submit the source .Rmd or .qmd file along with the compiled PDF to allow us to run your code easily.
  4. To ensure reproducibility of your simulation results, use `set.seed()` at the beginning of your document.

Problem 1. Important concepts (10 points)

(a) What is a standard error? What is the difference between a standard error and a standard deviation?

Answer: The standard error is a measure of the statistical uncertainty surrounding a parameter estimate. The standard error is a measure of dispersion in a sampling distribution; the standard deviation is the measure of dispersion of any distribution but is most often used to describe the dispersion in an observed variable. The standard error is the standard deviation of the sampling distribution, or the set of estimates that could have arisen under all possible random assignments.

(b) How is randomization inference used to test the sharp null hypothesis of no effect for any subject?

Answer: The sharp null hypothesis of no effect is a case in which $Y_i(1) = Y_i(0)$; under this assumption, all potential outcomes are observed because treated and untreated potential outcomes are identical. In order to form the sampling distribution under the sharp null hypothesis of no effect, we simulate a random assignment and calculate the test statistic (for example, the difference-in-means between the assigned treatment and control groups). This simulation is repeated a large number of times in order to form the sampling distribution under the null hypothesis. The p-value of the test statistic that is observed in the actual experiment is calculated by finding its location in the sampling distribution under the null hypothesis. For example, if the observed test statistic is as large or larger than 9,000 of 10,000 simulated experiments, the one-tailed p-value is 0.10.

(c) What is a 95% confidence interval?

Answer: A confidence interval consists of two estimates, a lower number and an upper number, that are intended to bracket the true parameter of interest with a specified probability. An estimated confidence interval is a random variable that varies from one experiment to the next due to sampling variability. A 95% interval is designed to bracket the true parameter with a 0.95 probability. In other words, across hypothetical replications of a given experiment, 95% of the estimated 95% confidence intervals will bracket the true parameter.

(d) How does complete random assignment differ from block random assignment and cluster random assignment?

Answer: Under complete random assignment, each subject is assigned separately to treatment or control groups such that m of N subjects end up in the treatment condition. Under block random assignment, complete random assignment occurs within each block or subgroup. Under clustered assignment, groups of subjects are assigned jointly to treatment or control; the assignment procedure requires that if one member of the group is assigned to the treatment group, all others in the same group are also assigned to treatment.

(e) Experiments that assign the same number of subjects to the treatment group and control group are said to have a "balanced design." What are some desirable statistical properties of balanced designs?

Answer: One desirable property of a balanced design is that under certain conditions, it generates less sampling variability than unbalanced designs; this property of balanced designs holds when the variance of $Y_i(0)$ is approximately the same as the variance of $Y_i(1)$. Another attractive property is that estimated confidence intervals are, on average, conservative (they tend to overestimate the true amount of sampling variability) under balanced designs. (A final attractive property, which comes up in Chapter 4, is that regression is less prone to bias under balanced designs.)

Problem 2. Potential outcomes (10 points)

Consider the schedule of outcomes in the table below. If treatment A is administered, the potential outcome is $Y_i(A)$, and if treatment B is administered, the potential outcome is $Y_i(B)$. If no treatment is administered, the potential outcome is $Y_i(0)$. The treatment effects are defined as $Y_i(A) - Y_i(0)$ or $Y_i(B) - Y_i(0)$.

Subject	$Y_i(0)$	$Y_i(A)$	$Y_i(B)$
Miriam	1	2	3
Benjamin	2	3	3
Helen	3	4	3
Eva	4	5	3
Billie	5	6	3

Suppose a researcher plans to assign two observations to the control group and the remaining three observations to just one of the two treatment conditions. The researcher is unsure which treatment to use.

(a) Applying the equation for $\mathbb{V}[\widehat{\tau}_{DiM} \mid \mathcal{O}_N]$ from slide 18 on Randomized Experiments, determine which treatment, A or B, will generate a sampling distribution with a smaller standard error.

Answer:

First, notice that $Y_i(A) = Y_i(0) + 1$. Then using the results from lecture slides:

$$\sigma_B = \sqrt{\mathbb{V}[\widehat{\tau}_A \mid \mathcal{O}_N]} = \sqrt{\frac{2}{3} + \frac{2}{2} - 0} = 1.291$$

$$\sigma_A = \sqrt{\mathbb{V}[\widehat{\tau}_B \mid \mathcal{O}_N]} = \sqrt{\frac{0}{3} + \frac{2}{2} - \frac{2.5}{5}} = 0.707$$

The standard error for the B vs. control comparison is smaller than the standard error for the A vs. control comparison. Thus, administering treatment B gives rise to a narrower sampling distribution.

(b) What does the result in part (a) imply about the feasibility of studying interventions that attempt to close an existing "achievement gap"?

Answer:

When treatment B is administered, the achievement gap between the best and worst student narrows, leaving no variance in $Y_i(B)$. one of the three terms in the variance equation therefore become zero, and the variance of individual treatment becomes larger. The resulting standard error is much lower than it would be under treatment A, which has a constant effect across all subjects. The basic principle here is that it helps to study treatments that reduce the covariance between untreated and treated potential outcomes.

Problem 3. Randomization inference (20 points)

The file `Clingingsmith_subset.dta` contains data from a study by Clingingsmith, Khwaja, and Kremer (2009) that focuses on Pakistani Muslims who participated in a lottery to obtain visa for the pilgrimage to Mecca. The study is described as follows in chapter 3.5 of the Gerber and Green (2012):

"By comparing lottery winners to lottery losers, the authors are able to estimate the effects of the pilgrimage on the social, religious, and political views of the participants. Here, we consider the effect of winning the visa lottery on attitudes toward people from other countries. Winners and losers were asked to rate the Saudi, Indonesian, Turkish, African, European, and Chinese people on a five-point scale ranging from very negative (-2) to very positive ($+2$). Adding the responses to all six items creates an index ranging from -12 to $+12$."

(a) Use the data in `Clingingsmith subset.dta` to test the sharp null hypothesis that winning the visa lottery (variable `success` in dataset) for the pilgrimage to Mecca had no effect on the views of Pakistani Muslims toward people from other countries (variable `views` in data). Assume that the Pakistani authorities assigned visas using complete random assignment. Conduct $10,000$ simulated random assignments under the sharp null hypothesis.

- How many of the simulated random assignments generate an estimated ATE that is at least as large as the actual estimate of the ATE?
- What is the implied one-tailed (upper) p-value?
- How many of the simulated random assignments generate an estimated ATE that is at least as large in absolute value as the actual estimate of the ATE?
- What is the implied two-tailed p-value?

Answer:

```r
# package to load Stata data into R
pacman::p_load(
  tidyverse, # ggplot2, dplyr, tidyr, readr, purrr, and tibble.
  randomizr, # random assignments and conducting randomization-based inference
  haven,  # reading and writing data from SPSS, Stata, SAS
  pbapply) # progress bar support to vectorized R functions like lapply and sapply

# Change this path to your own directory that contains the files
data <- read_dta("../_data/Clingingsmith_subset.dta")

# Check how many people won the lottery
table(data$success, useNA = "always")
```

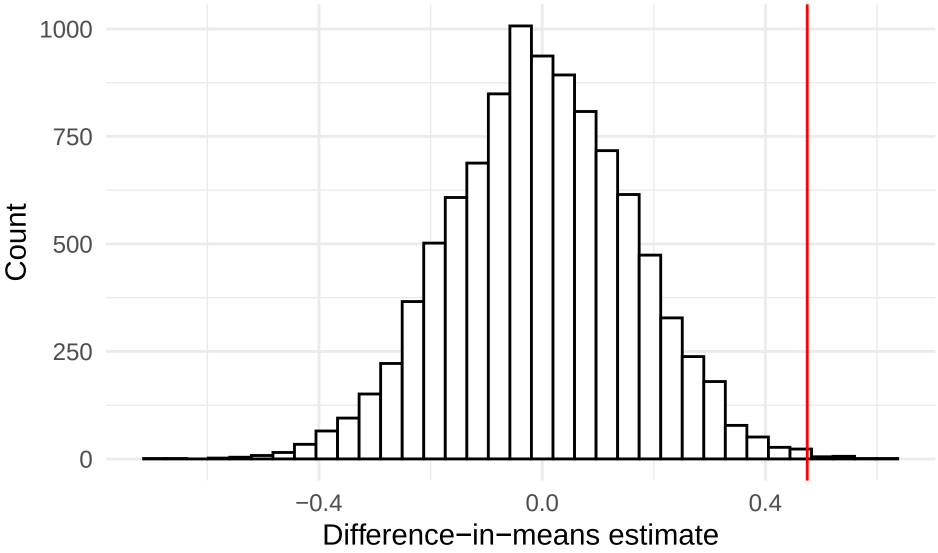
```
  0    1 <NA> 
448  510    0 
```

```r
# Estimate the observed ATE
hat_ate <- mean(data$views[data$success == 1]) -
  mean(data$views[data$success == 0])

# Generate the sampling distribution
# under the null hypothesis of no effect
# assuming complete random assignment of 510 units to treatment
set.seed(123)
Zs <- pbapply::pbreplicate(
   n = 10000,
   expr = complete_ra(N = nrow(data), m = 510), cl = 6)

estimates <- pbapply::pbapply(
   Zs, MARGIN = 2,
   function(x) {
      mean(data$views[x == 1]) - mean(data$views[x == 0])
   },
   cl = 6
   )

# Plot the sampling distribution and the observed estimate
data.frame(estimates = estimates) |>
  ggplot(aes(x = estimates)) +
  geom_histogram(bins = 35, fill = "white", color = "black") +
  theme_bw() +
  xlab("Difference-in-means estimate") +
  ylab("Count") +
  geom_vline(xintercept = hat_ate, color = "red") +
  theme_minimal()
```



```r
# Number of estimates at least as large as the observed one
sum(estimates >= hat_ate)
```

```
[1] 17
```

```r
# Implied one-tailed upper p-value
mean(estimates >= hat_ate)
```

```
[1] 0.0017
```

```r
# Number of estimates at least as large as the observed one in absolute value
sum(abs(estimates) >= abs(hat_ate))
```

```
[1] 37
```

```r
# Implied two-tailed p-value
mean(abs(estimates) >= abs(hat_ate))
```

```
[1] 0.0037
```

We reject the sharp null hypothesis that winning the visa lottery had no effect on the views toward people from other countries for any person in the sample (for conventional significance levels).

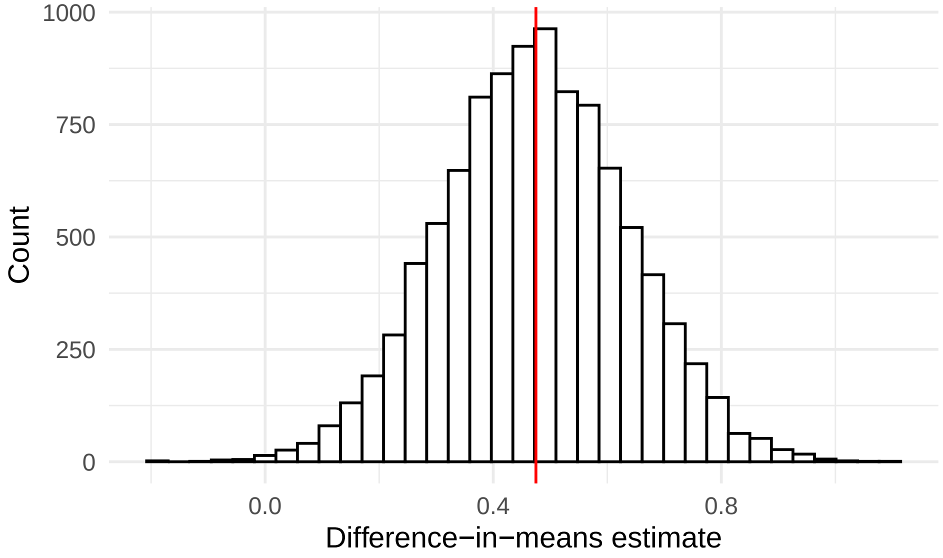
(b) Now, test the sharp null hypothesis that the effect of winning the visa lottery for the pilgrimage to Mecca on the views of Pakistani Muslims toward people from other countries equals your estimate of the *ATE*. What is the implied two-tailed p-value?

Answer:

```r
# Adjust the outcomes in line with sharp null hypothesis
data <-
  data |>
  mutate(
    views_treat = ifelse(success == 1, views, views + hat_ate),
    views_control = ifelse(success == 0, views, views - hat_ate)
  )

# Generate sampling distribution under null hypothesis
# Using the adjusted outcomes
estimates <- pbapply::pbapply(
  Zs,
  MARGIN = 2,
  FUN = function(x) {
   mean(data$views_treat[x == 1]) - mean(data$views_control[x == 0])
   }
)

# Plot the sampling distribution under the null hypothesis
# Together with the observed estimate
data.frame(estimates = estimates) %>%
  ggplot(aes(x = estimates)) +
  geom_histogram(bins = 35, fill = "white", color = "black") +
  theme_bw() +
  xlab("Difference-in-means estimate") +
  ylab("Count") +
  geom_vline(xintercept = hat_ate, color = "red") +
  theme_minimal()
```



```r
# Two-tailed p-value
mean(abs(estimates) >= abs(hat_ate))
```

```
[1] 0.4935
```

We cannot reject the sharp null hypothesis that the treatment effect is equal to `0.475` for every unit (for conventional significance levels).

Problem 4. Block random assignment (30 points)

Naturally occurring experiments sometimes involve what is, in effect, block random assignment. For example, Titiunik (2016) studies the effect of lotteries that determine whether state senators in Texas and Arkansas serve two-year or four-year terms in the aftermath of decennial redistricting. These lotteries are conducted within each state, and so there are effectively two distinct experiments on the effects of term length. An interesting outcome variable is the number of bills (legislative proposals) that each senator introduces during a legislative session.

The data set `Titiunik.dta` contains `term2year`–an indicator for whether a senator was assigned to treatment (a two-year rather than a four-year term); `bills_introduced`–the number of bills introduced by the senator (the outcome); `texas0_arkansas1`–an indicator for whether the senator is from Texas or Arkansas (the blocks).

(a) For each state, estimate the effect of having a two-year term on the number of bills introduced.

Answer:

```r
data <- read_dta("../_data/Titiunik.dta")

block_level_estimates <-
  data  |>
  group_by(texas0_arkansas1)  |>
  summarise(
    ate_hat = mean(bills_introduced[term2year == 1]) -
      mean(bills_introduced[term2year == 0]),
    sd_hat = sqrt(
      var(bills_introduced[term2year == 1]) /
        sum(term2year == 1) +
        var(bills_introduced[term2year == 0]) / sum(term2year == 0)
    ),
    n_resp = n(),
    weighted_ate_hat = ate_hat * n_resp / nrow(data),
    weighted_sd_hat_sq = sd_hat^2 * (n_resp / nrow(data))^2
  )

# Block-level ate estimates
block_level_estimates$ate_hat
```

```
[1] -16.74167 -10.09477
```

(b) For each state, estimate the standard error of the estimated ATE (using the estimator on the slides)

Answer:

```r
block_level_estimates$sd_hat
```

```
[1] 9.345871 3.395979
```

(c) Estimate the overall ATE for both states combined through a weighted average of the state-level ATE estimates (as discussed on the slides).

Answer:

```r
block_level_estimates  |>
  dplyr::summarize(hat_ate = sum(weighted_ate_hat)) |>
  unlist()
```

```
 hat_ate
-13.2168
```

(d) Explain why, in this study, simply pooling the data for the two states and comparing the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators leads to biased estimates of the overall ATE.

Answer: The probability of being assigned to treatment varries by block.

(e) Show that the weighted average of the state-level ATE estimates gives the same overall ATE estimate as a weighted regression that weights each observation by the inverse of its probability of being assigned to the condition that the observation was assigned to. For this task you can rely on the function `lm()` to run the regression and pass the inverse probability weights to the `weights` argument of this function.

Answer:

```r
# Calculating weights
data <- data |>
  group_by(texas0_arkansas1) |>
  mutate(p_treatment = mean(term2year == 1)) |>
  mutate(
    weight = ifelse(term2year == 1, 1 / p_treatment, 1 / (1 - p_treatment))
  )

# Running a weighted regression
(hat_ate <-
  estimatr::lm_robust(
    bills_introduced ~ term2year,
    weights = weight,
    data = data
  )$coefficients["term2year"])
```

```
term2year
 -13.2168
```

(f) Estimate the standard error for the overall ATE by combining estimates of the block-level standard errors as shown on the slides.

Answer:

```r
block_level_estimates  |>
  dplyr::summarize(hat_sd = sqrt(sum(weighted_sd_hat_sq))) |>
  unlist()
```

```
 hat_sd
4.74478
```

(g) Use randomization inference to test the sharp null hypothesis that the treatment effect is zero for all senators.

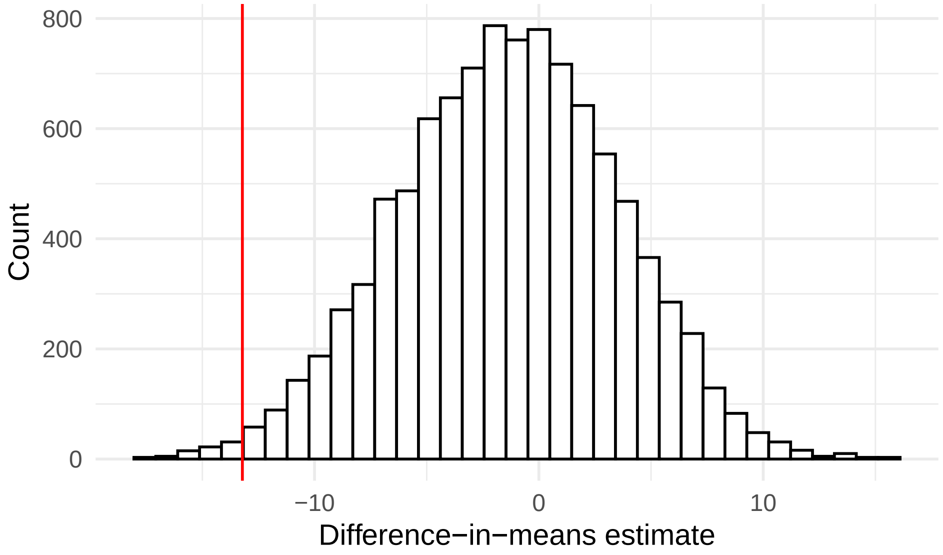
Answer:

```r
# Simulate 10,000 possible assignments
Zs <- pbapply::pbreplicate(
  n = 10000,
  expr = block_ra(blocks = data$texas0_arkansas1, block_m = c(15, 18)),
  cl = 6
)

# Function to get estimates for one assignment
getEstimate <- function(Z) {
  data$Z <- Z
  return(coef(lm(bills_introduced ~ Z, weights = weight, data = data))["Z"])
}

# Get estimates for all the simulated assignments
estimates <- pbapply::pbapply(
  Zs,
  MARGIN = 2,
  FUN = function(x) getEstimate(x),
  cl = 6
)


# Plot sampling distribution and observed estimate
data.frame(estimates = estimates) %>%
  ggplot(aes(x = estimates)) +
  geom_histogram(bins = 35, fill = "white", color = "black") +
  theme_bw() +
  xlab("Difference-in-means estimate") +
  ylab("Count") +
  geom_vline(xintercept = hat_ate, color = "red") +
  theme_minimal()
```



```r
# Two-tailed p-value
mean(abs(estimates) >= abs(hat_ate))
```

```
[1] 0.0089
```

We can reject the sharp null hypothesis that assignment to a two-year (instead of four-year) term had no effect on the number of bills introduced for any senator.

Problem 5. Cluster random assignment (30 points)

Consider the 2003 Kansas City voter mobilization experiment studied by Arceneaux (2005). In the experiment a group called ACORN targeted 28 low-income precincts in an effort to mobilize voters on behalf of a ballot measure designed to fund municipal bus service. ACORN wanted to work within selected precincts in order to make it easier to train and supervise its canvassers. Of the 28 precincts in the sample, 14 were randomly allocated to the treatment group, and ACORN made repeated attempts to canvass and call voters on its target list in those precincts. The 28 precincts contain a total of $9,712$ voters, and the number of targeted voters per precinct ranges from 31 to 655; the marked difference in cluster size leaves open the possibility for bias if precincts with large numbers of potential ACORN sympathizers have different potential outcomes from precincts with relatively few.

The study data in `Arceneaux_subset.dta` contains a wealth of covariates: the registrar recorded whether each voter participated in elections dating back to 1996.

(a) Assess the balance of the treatment and control groups by looking at whether past turnout predicts treatment assignment. To do so first regress treatment assignment on the entire set of past votes, and retrieve the observed F-statistic for the test of goodness of fit (e.g. using `estimatr::lm_robust()$fstatistic`). Use randomization inference to test the null hypothesis that none of the past turnout variables predict treatment assignment.[1] Judging from the p-value of this test, what does the F-statistic seem to suggest about whether subjects in the treatment and control groups have comparable background characteristics?

Answer:

```r
data <- haven::read_dta("../_data/Arceneaux_subset.dta")

Z <-  data$treatment
Y <- data$vote03
clust <- data$unit
covs <- as.matrix(data[,2:21])  # covariates are past voter turnout

Fstat <- summary(estimatr::lm_robust(Z~covs))$fstatistic[1]   # F-statistic from actual data

set.seed(123)
perms <- replicate(1000, randomizr::cluster_ra(clusters = clust, m = 14))  # clustered assignment
Fstatstore <- pbapply::pbapply(
   perms, MARGIN = 2, function(x) summary(estimatr::lm_robust(x~covs))$fstatistic[1],
   cl = 6)  # F-statistic from random assignment

p.value <- mean(Fstatstore >= Fstat)
```

Using randomization inference, we recover a p-value of `0.942`, meaning we cannot reject the null hypothesis of random assignment.

(b) Explain what intracluster correlation is and why it is important in the context of this experiment. Calculate the intracluster correlation for the 2003 Kansas City voter mobilization experiment using the turnout data from 2003. Interpret your results and calculate Moulton factor (design effect) using formula from slide 63 on Randomized Experiments.

Answer:

Intracluster correlation (ICC) is a measure of the similarity of observations within a cluster relative to other clusters. The ICC quantifies how much gives the proportion of the variance that arises from within-cluster variance compared to the variance between clusters. The ICC can be mathematically defined as:

$\rho = \frac{\sigma_B^2}{\sigma^2} = 1 - \frac{\sigma_W^2}{\sigma^2}$

The following code provides a calculation of the ICC:

[1] **Hint:** To simulate the distribution of the F-statistic, you must generate a large number of random cluster assignments and calculate the F-statistic for from regressing each simulated treatment assignment on the entire set of past votes.

```r
# Calculating variance across all clusters
overall_var <- var(data$vote03)

# Calculating the within cluster variance
within <- data |>
  group_by(unit) |>
  summarize(var = var(vote03), mn = mean(vote03), n = n())

# Calculating the within variance sigma squared
within_var <- sum(within$var * (within$n - 1)) / sum(within$n - 1)

# Calculating the ICC
( icc <- 1 - (within_var / overall_var) )
```

```
[1] 0.01653447
```

Based on the ICC, the Moulton factor (the design effect) shows how much cluster randomization inflates the sampling variance compared to complete randomization. The Moulton factor is given as:

$\frac{\mathbb{V}_{\hat{\tau}_{CL}}}{\mathbb{V}_{\hat{\tau}_{R}}} = 1 + (\bar{N} - 1)\rho$, where $\bar{N} = \frac{1}{G}\Sigma_1^G N_j$.

```r
n_bar <- mean(table(data$unit))

(moulton_factor <- 1 + (n_bar - 1)*icc)
```

```
[1] 6.718565
```

An ICC of `0.017` indicates that there is a substantial level of within-cluster similarity in the observed outcomes. As suggested by Moulton factor implied by this ICC, the sampling variance of estimator that correctly accounts for clustering is approximately `6.719` times larger than the sampling variance of difference-in-means estimator.

(c) Regress turnout in 2003 (after the treatment was administered) on the experimental assignment and the full set of covariates. Interpret the estimated *ATE*. Use randomization inference to test the sharp null hypothesis that experimental assignment had no effect on any subject's decision to vote.

Answer:

```r
# Estimate the ATE using a linear model
ate <- estimatr::lm_robust(Y ~ Z + covs)$coefficients["Z"]

# Generate the sampling distribution under the sharp null hypothesis
set.seed(123)
estimates <- pbapply::pbreplicate(10000, {
   treat_sim <- randomizr::cluster_ra(clusters = clust, m = 14)
   estimatr::lm_robust(Y ~ treat_sim + covs)$coefficients["treat_sim"]},
   cl = 6)

# Calculate the one-tailed p-value
p.value.onetailed <- mean(estimates >= ate)
```

The estimate of the treatment effect is 0.056, meaning that treatment increased turnout by 5.6 percentage points. This finding is statistically significant. Under the sharp null, estimates as large or larger only occur 0.19% of the time.

(d) Using the turnout data from 2003, compare the parametric variance estimates obtained with and without accounting for clustering (both can be done using `estimatr::lm_robust()` and specifying appropriate `se_type` argument). What do your results suggest about the importance of considering clustering in this experiment?

Answer:

```r
no_cluster_model <- estimatr::lm_robust(
  Y ~ Z + covs, data = data, se_type = "HC0"
  ) |> estimatr::tidy() |> as_tibble()

cluster_model <- estimatr::lm_robust(
  Y ~ Z + covs, data = data, clusters = clust, se_type = "CR2"
  ) |> estimatr::tidy()

bind_rows(
  no_cluster_model, cluster_model
) |>
  dplyr::filter(term == "Z") |>
  dplyr::mutate(model = c("No Cluster Std.", "Clustered Std.")) |>
  dplyr::select(model, term, estimate, std.error, p.value) |>
  knitr::kable(digits = 5, align = "lccc") |>
  kableExtra::kable_minimal()
```

model	term	estimate	std.error	p.value
No Cluster Std.	Z	0.05596	0.00775	0.00000
Clustered Std.	Z	0.05596	0.01720	0.00366

These results suggest that when ignoring the structure of the data (sampling units within clusters compared to complete random assignment) the standard errors are too small (are not conservative). While the results with clustered standard errors are still significant at traditional levels, this exercise highlights the importance of clustering your standard errors when your data generating process is similar to the data structure in Arceneaux (2005).

References

Arceneaux, Kevin. 2005. "Using Cluster Randomized Field Experiments to Study Voting Behavior." *The Annals of the American Academy of Political and Social Science* 601 (1): 169–79.

Clingingsmith, David, Asim Ijaz Khwaja, and Michael Kremer. 2009. "Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering." *The Quarterly Journal of Economics* 124 (3): 1133–70.

Gerber, Alan S, and Donald P Green. 2012. "Field Experiments: Design, Analysis, and Interpretation." *(No Title)*.

Titiunik, Rocio. 2016. "Drawing Your Senator from a Jar: Term Length and Legislative Behavior." *Political Science Research and Methods* 4 (2): 293–316.