Final Exam

STAT II (Spring 2025)

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

Guidelines

- You will have **24 hours** to complete the exam. Please upload your answers to Brightspace by **Wednesday, April 23rd, at 10:00 AM**.
- This final exam is open-book. You can use any relevant materials, but you always have to write the answers in your own words and not copy-paste from the sources.
- We can only answer clarifying questions. Please ask all questions by sending an email to Gosha Syunyaev (g.syunyaev@vanderbilt.edu) and Alex Dean (alexander.r.dean@vanderbilt.edu). You always need to send an email to all of us. The title of the email should be "[Question: Final Exam in PSCI8357]".
- To make sure we can answer questions precisely, we answer questions about the final via email between 10:00 AM and 7:00 PM on April 22. Please plan accordingly.
- You **cannot** discuss the final exam with your classmates or anyone. Asking questions to your classmates is not allowed. If you have clarifying questions, you should only ask them to the instructors as instructed above. Submitted derivations and answers must be your own work.
- No late submission will be accepted unless discussed and approved with the instructor beforehand.
- Grading: Please show every step of your derivations. We grade steps of derivations as well as your final answers. So, even if you cannot solve the problem entirely, we can give you partial points to your derivations. Even if your final answer is correct, you might not get full points if your derivations are incomplete.
- Stylistic Requirements: Please follow the same rules as those used in problem sets.
- Please ensure your answers are included in the "Main Answer" PDF or HTML file (preferably produced using LaTeX or Quarto Markdown). Also submit the source file containing all R code needed to reproduce your answers. If you have questions, please email us.

Part 1. True/False (40 points, 4 points per question)

Question 1.1

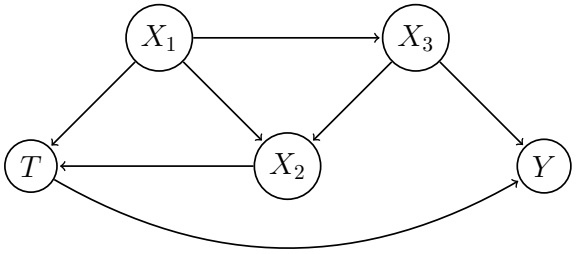
In the Fisherian approach, we are often interested in the sharp null hypothesis. **One of the key properties of the Fisherian approach is that we always obtain the same p-value regardless of the choice of test statistics.**

Question 1.2

Consider a randomized experiment in which a researcher randomly assigns political messages to experimental subjects. Suppose she finds that the estimated average treatment effect for Democrats is much larger than the one for non-Democrats. **She cannot argue that the difference in these treatment effects is caused by the partisanship.**

Question 1.3

We are interested in identifying the causal effect of T on Y. Consider the following causal DAG. **A set (X_2, X_3) satisfies the backdoor criteria for identifying the causal effect of T on Y.**



Question 1.4

Suppose we assume the conditional ignorability and positivity assumptions with pre-treatment covariates $\mathbf{X}_i$. To estimate the ATE (the average treatment effect), we can use the Hajek estimator, which can be computed by the weighted linear regression where we only include an intercept and one binary treatment variable T_i, while specifying a weight, w_i, for each unit.

The correct weights are:

Group	Weight formula
Treated ($T_i = 1$)	$w_i = 1/\{1 - \pi(\mathbf{X}_i)\}$
Control ($T_i = 0$)	$w_i = 1/\pi(\mathbf{X}_i)$

where $\pi(\mathbf{X}_i)$ is the propensity score.

Question 1.5

Suppose we have the binary instrumental variable, the binary treatment variable, and the continuous outcome variable. We found that the estimated causal effect of the instrumental variable on the treatment variable was positive among males and negative among females. **This is statistical evidence against the exclusion restriction assumption.**

Question 1.6

Suppose we have the binary instrumental variable, the binary treatment variable, and the continuous outcome variable. Suppose the estimated average effect of the instrument on the treatment variable is positive. **Under this setup, the estimate of the Local Average Treatment Effect (or the complier average treatment effect) is always less than or equal to the estimate of the Intent-to-Treat Effect (ITT).**

Question 1.7

Consider the basic DiD design where we use a panel data and there are two groups (treatment and control groups) and two time periods (before and after the treatment assignment). If the treatment assignment is completely randomized, **the parallel trend assumption, which is necessary for identifying the average treatment effect on treated (ATT), always holds.**

Question 1.8

Consider the staggered adoption design where different units can receive the treatment in different time periods, but once units receive the treatment, they remain exposed to the treatment. Suppose we assume the parallel trends assumption holds for all groups for all time periods. **When treatment effects are constant across units, the two-way fixed effects estimator is unbiased for the ATT.**

Question 1.9

Under the sharp regression discontinuity design (RDD), we require the identification assumption of the continuity of the average potential outcomes. **To perform reliable causal inference, we also require the positivity assumption, i.e.,** $0 < \Pr(T_i = 1 \mid X_i = x) < 1$ **for all** x **where** T_i **is the treatment variable and** X_i **is the forcing variable.**

Question 1.10

The key difference between the sharp and fuzzy RD designs is that the instrumental variable rather than the treatment variable is determined by a threshold under the fuzzy RDD. Therefore, to estimate the Local Average Treatment Effect for compliers at the threshold under the fuzzy RDD, **we just need to update the continuity assumption; we assume the continuity of the potential outcomes and the potential treatments under the fuzzy RDD, but we do not need other assumptions compared to the sharp RDD.**

Part 2. Analytical questions (30 points)

Question 2.1: Randomized Experiment (10 points)

In this question, we examine a randomized experiment. When you provide answers, use the following notations. If you need to introduce additional notations, explain them in detail. We also assume no interference throughout this question.

- Index experimental units with $i \in \{1, \ldots, N\}$, where N is the total number of experimental units.
- Define $T_i \in \{1, 0\}$ to be a binary treatment variable.
- Define $Y_i(t)$ to be the potential outcome when unit i receives $T_i = t$, where $t \in \{0, 1\}$.
- Define Y_i to be the observed outcome for unit i, and we assume the consistency of the potential outcomes, $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$.

We formally define the average treatment effect (ATE) as follows:

$$\tau = \frac{1}{N} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}.$$

Suppose a researcher employs a block randomized experiment. That is, the treatment assignment is completely randomized within blocks. More specifically, units are partitioned into two equally sized blocks, defined by a binary pre-treatment variable $X_i \in \{0, 1\}$. There are $N/2$ units with $X_i = 1$ and $N/2$ units with $X_i = 0$.

Within block 1 (units with $X_i = 1$), 50% of the units receive the treatment and the other 50% receive the control based on a completely randomized design. Use N_{11} (N_{01}) to denote the number of treated (control) units within block 1, where $N_{11} + N_{01} = N/2$. On the other hand, within block 0 (units with $X_i = 0$), 25% of the units receive the treatment and the other 75% receive the control based on a completely randomized design. Use N_{10} (N_{00}) to denote the number of treated (control) units within block 0, where $N_{10} + N_{00} = N/2$.

The design can be summarized as the following table, where each cell shows the number of units for each combination of X_i and T_i:

	$T_i = 1$	$T_i = 0$
$X_i = 1$	$N_{11} = N/4$	$N_{01} = N/4$
$X_i = 0$	$N_{10} = N/8$	$N_{00} = 3N/8$

To simplify the problem, we also assume that $Y_i(0) = 0$ for everyone in the experiment.

Question: The difference-in-means estimator, defined below, is not unbiased for the ATE under this design in general. Derive an exact expression of the bias. In some special cases, the bias can be zero. Discuss at least one concrete situation under which the bias is zero. The bias is formally defined as $\text{Bias} = \mathbb{E}[\widehat{\tau}_{\mathsf{DiM}} \mid \mathcal{O}_N] - \tau$, where $\mathcal{O}_N = \{Y_i(1), Y_i(0), X_i\}_{i=1}^N$.

$$\widehat{\tau}_{\mathsf{DiM}} = \frac{1}{N_{11} + N_{10}} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N_{01} + N_{00}} \sum_{i=1}^{N} (1 - T_i) Y_i.$$

Hint: *You can use the following equalities in your proof.*

- $\mathbb{E}[T_i X_i \mid \mathcal{O}_N] = \frac{X_i}{2}$
- $\mathbb{E}[(1-T_i)X_i \mid \mathcal{O}_N] = \frac{X_i}{2}$
- $\mathbb{E}[T_i(1-X_i) \mid \mathcal{O}_N] = \frac{(1-X_i)}{4}$
- $\mathbb{E}[(1-T_i)(1-X_i) \mid \mathcal{O}_N] = \frac{3(1-X_i)}{4}$

Question 2.2: Difference-in-Differences

In this question, we consider how to analyze the basic DiD design.

(a) (10 points) Consider the simple case with the following notation:

Symbol	Description
i	Index for units, $i \in \{1, \ldots, n\}$, where n is the total number of unique units
t	Time period, $t = 0$ (pre-treatment), $t = 1$ (post-treatment)
G_i	Binary variable: 1 if unit i belongs to Group 1, 0 if unit i belongs to Group 0
T_{it}	Binary treatment: 1 if unit i is treated at time t, 0 otherwise
$Y_{it}(t)$	Potential outcome for unit i when $T_{it} = d$
Y_{it}	Observed outcome: $Y_{it} = Y_{it}(T_{it})$

Assume that the following parallel trends assumption holds:

$$\mathbb{E}\{Y_{i1}(0) - Y_{i0}(0) \mid G_i = 1\} = \mathbb{E}\{Y_{i1}(0) - Y_{i0}(0) \mid G_i = 0\}.$$

Question: Under this parallel trends assumption, show that the following difference-in-differences (DiD) estimator is unbiased for the Average Treatment Effect for Treated (ATT):

$$\widehat{\tau}_{\mathsf{DiD}} = \left\{ \frac{1}{n_1} \sum_{i=1}^{n} G_i(Y_{i1} - Y_{i0}) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - G_i)(Y_{i1} - Y_{i0}) \right\},$$

where $n_g = \sum_{i=1}^{n} \mathbb{1}\{G_i = g\}$ is the number of units in each group ($g \in \{0, 1\}$).

(b) (10 points) Now, we consider cases where the parallel trends assumption is violated. However, we have another placebo state where the treatment was not implemented at all. We use this placebo state to implement the triple difference-in-differences. The notation is as follows:

Symbol	Description
i	Index for units, $i \in \{1, \ldots, n\}$, where n is the total number of unique units
t	Time period, $t = 0$ (pre-treatment), $t = 1$ (post-treatment)
S_i	Binary variable: 1 if unit i lives in State 1, 0 if in State 0
G_i	Binary variable: 1 if unit i belongs to Group 1, 0 if in Group 0
T_{it}	Binary treatment: defined by time, state, and group membership
$Y_{it}(d)$	Potential outcome for unit i when $T_{it} = d$
Y_{it}	Observed outcome: $Y_{it} = Y_{it}(T_{it})$

Treatment assignment:

- If $S_i = 1$: $T_{i0} = 0$, $T_{i1} = 1$ if $G_i = 1$; $T_{i0} = T_{i1} = 0$ if $G_i = 0$

- If $S_i = 0$: $T_{i0} = T_{i1} = 0$ for all units

Assume the following "difference-in-differences-in-differences" (triple difference) parallel trends assumption:

$$\begin{aligned}&\mathbb{E}\{Y_{i1}(0) - Y_{i0}(0) \mid G_i = 1, S_i = 1\} - \mathbb{E}\{Y_{i1}(0) - Y_{i0}(0) \mid G_i = 0, S_i = 1\} \\ &= \mathbb{E}\{Y_{i1}(0) - Y_{i0}(0) \mid G_i = 1, S_i = 0\} - \mathbb{E}\{Y_{i1}(0) - Y_{i0}(0) \mid G_i = 0, S_i = 0\}\end{aligned}$$

Question: Under this assumption, show that the following triple difference-in-differences estimator is unbiased for the Average Treatment Effect for Treated in State 1, i.e., $\mathbb{E}\{Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1, S_i = 1\}$:

$$\begin{aligned}\widehat{\tau}_{\mathsf{DDD}} = &\left\{\frac{1}{N_{11}}\sum_{i=1}^{n} S_i G_i (Y_{i1} - Y_{i0}) - \frac{1}{N_{10}}\sum_{i=1}^{n} S_i (1 - G_i)(Y_{i1} - Y_{i0})\right\} \\ &- \left\{\frac{1}{N_{01}}\sum_{i=1}^{n} (1 - S_i) G_i (Y_{i1} - Y_{i0}) - \frac{1}{N_{00}}\sum_{i=1}^{n} (1 - S_i)(1 - G_i)(Y_{i1} - Y_{i0})\right\}\end{aligned}$$

where $N_{sg} = \sum_{i=1}^{n} \mathbb{1}\{S_i = s\}\mathbb{1}\{G_i = g\}$ is the number of units in group g at State s ($g \in \{0, 1\}$, $s \in \{0, 1\}$).

Part 3. Data analysis (30 points)

This question is based on the paper by Lee (2008) we discussed in class. The study investigates the causal effect of incumbency in U.S. House elections on subsequent electoral outcomes by exploiting close electoral races. A simplified version of the dataset (`lee_final.rds`) that we analyze in this question is shared on Brightspace and GitHub and can be loaded using the `read_rds()` function from the `readr` package. The variables in the dataset are described below:

Variable	Description
`state`	Identifiers for states
`distid`	Identifiers for districts
`yearel`	Year of elections
`share_t`	The vote share of the Democratic party in the current election
`margin_t`	The margin of victory for the Democratic party in the current election
`margin_lag1`	The margin of victory for the Democratic party in the last election
`margin_lag2`	The margin of victory for the Democratic party in the election before last

We perform a sharp regression discontinuity analysis to estimate the incumbency advantage. We use the margin of victory for the Democratic party in the last election (`margin_lag1`) as the forcing variable and the vote share in the current election (`share_t`) as the outcome variable.

(a) (5 points) State the causal estimand under the sharp RDD and provide its substantive meaning in the context of Lee (2008).

(b) (5 points) State the key identification assumption and provide its substantive meaning in the context of Lee (2008).

(c) (10 points) Conduct two tests of RDD assumptions. First, using the margin of victory for the Democratic party in the election before the last election (`margin_lag2`) as the placebo outcome. Second, provide a plot of density `margin_lag1` on both sides of the cutoff and conduct formally the McCrary (2008) test. Please briefly interpret the results and explain which assumption(s) we can assess with these tests.

(d) (10 points) Estimate the causal effect using a local quadratic polynomial regression (including both linear and quadratic terms for the running variable). Use `rdrobust::rdrobust()` function to apply Epanechnikov kernel weights and select the bandwidth with `bwselect = "mserd"`. Report the point estimate, standard error, and 95% confidence interval. Do not forget to cluster standard errors by the `state` variable. Also, create an RDD plot showing the fitted quadratic trends on both sides of the cutoff, using the selected bandwidth (`bws` from the `rdrobust` fit object). Use `rdplot()` from the `rdrobust` package. Briefly interpret your results.

References

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in US House Elections." *Journal of Econometrics* 142 (2): 675–97.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142 (2): 698–714.