Midterm 2

S350

Upload your typed answers as one PDF/Word/HTML document through the Assignments tab on Canvas by 11:59 pm, Thursday 16th April. Also upload your R code as one .R or .Rmd file.

Instructions, warnings, and veiled threats

- You may freely refer to books, notebooks, and the web. However, **you must not discuss this exam with anyone other than the instructor and the AI until the due date has passed**, except on Piazza to the extent discussed below.
- **Include all R code** as an appendix or as a .R or Markdown file. (Do not *just* upload a code file.)
- Include all graphs you're asked to draw in the body of your document.
- Give your answers in sentences, e.g. "The degrees of freedom for the test was 123." Just giving R output is not sufficient for full credit.
- Round answers sensibly. Confidence intervals stated to six significant figures will not get full credit.

What may I post about on Piazza?

- If you think there is an error in the exam, let us know as soon as possible.
- You may for clarification on the questions.
- You may ask general questions about course material.
- You can ask for help handling the data. However, it may be easier to talk about these issues during office hours.

What may I ask at the lecturer's/TAs' office hours?

- General questions about course material.
- Help entering and manipulating the data.
- Questions about Steph Curry, radishes, and eating people.

What can I ask other students?

Nothing.

1 Curry's mouth

The basketball player Steph Curry sometimes shoots free throws with his mouthguard in his mouth, and sometimes shoots free throws with his mouthguard outside of his mouth. His free throw statistics for one season were:

- Free throws with mouthguard in: 110 completed, 13 missed (89.4%)
- Free throws with mouthguard out: 198 completed, 16 missed (92.5%)

His observed free throw completion rate was slightly higher when his mouthguard was outside his mouth. However, we should check whether the difference could be plausibly explained as luck.

(a) Find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard *in*. Give a numerical answer.

(b) Find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard *out*.

(c) Suppose we wish to test the null hypothesis that Curry's probability of completing a free throw is the same with his mouthguard in as it is with his mouthguard out. The P-value for such a test is 0.33. What does this P-value tells you? Explain.

2 Kidneys and hearts

In part of a study reported by Perotta and Finch (1972), the blood films of 16 patients with severe renal (kidney) anemia and 10 patients with functional heart disease were measured for red blood cell counts. The percentage changes in red blood cell counts appear below:

- Renal anemia: 2.20, 1.52, 1.54, 0.77, 0.34, 0.45, 0.39, 0.29, 0.18, 0.16, 0.23, 0.24, 0.17, 0.08, 0.02, 0.02.
- Heart disease: 1.84, 0.44, 0.30, 0.06, 0.20, 0.14, 0.10, 0.09, 0.06, 0.04.

Because of non-normality, we'll do our analysis on the *logs* of the data. Suppose we wish to show that the mean log percentage change is *greater* for renal patients than for heart patients.

(a) Write down mathematical null and alternative hypotheses for such a test. Carefully define the parameters you use.

(b) Calculate the observed test statistic and the degrees of freedom for an appropriate test.

(c) Find a P-value for your test. What does it tell you about severe renal anemia, functional heart disease, and red blood cell counts?

(d) Find a a 95% confidence interval for the difference in the means of the log percentage changes. Transform the confidence interval back to the original scale. What does your interval tell you quantitatively about the difference in the distributions of red blood cell count changes between the two types of patients?

3 Radish roots

Rosene (1950) studied how quickly hairs on radish roots absorbed water when they were immersed. For each of eleven radishes, she measured the rate of influx of water for a young root hair and an old root hair on that radish. The data is given below.

Radish	Old	Young
A	0.89	2.13
B	0.49	1.16
C	0.91	2.60
D	0.80	1.58
E	0.56	1.53
F	0.79	1.70
G	0.47	2.67
H	0.50	2.64
I	1.08	2.19
J	1.65	2.54
K	1.94	4.46

Table 1: Radish root hair absorption data. Rates are in cubic microns per square micron per minute.

For each pair, the "Young" number is bigger than the "Old" number, so even without a test it's clear that young roots take in water more quickly. But how much more quickly?

(a) Explain why we should study this data using techniques for one sample rather than techniques for two independent samples.

(b) Draw a normal probability (qqnorm) plot of the differences (old minus young.) Explain why:

   (a) We should be hesitant do a t-test on these differences (old minus young);

   (b) We should not take the logs of these differences (old minus young.)

(c) Instead of using the differences, we can use the *ratios*: ~~young divided by old~~ old divided by young. The ratios come from a distribution that's much closer to normal, so we can use one-sample t inference. Enter the data into R and find a 95% confidence interval for the average value of this ratio. Interpret what this confidence interval tells you.

4 Handwashing and moral judgments

A widely publicized psychology experiment (Schnall, Benton & Harvey (2008)) suggested that handwashing reduced the severity of moral judgments. However, since the experiment was performed on a small sample, researchers at Michigan State wished to replicate the experiment with more participants.

132 participants (students at Michigan State) were individually shown a distasteful scene involving heroin addicts from the movie *Trainspotting*. 63 of the students, randomly selected, were instructed to wash their hands afterward; the remaining 69 students were not. All 132 students then had six different scenarios described to them. For instance, the "plane" scenario was as follows:

> "Your plane has crashed in the Himalayas. The only survivors are yourself, another man, and a young boy. The three of you travel for days, battling extreme cold and wind. Your only chance at survival is to find your way to small a village on the other side of the mountain, several days away. The boy has a broken leg and cannot move very quickly. His chances of surviving the journey are essentially zero. Without food, you and the other man will probably die as well. The other man suggests that you sacrifice the boy and eat his remains over the next few days. How wrong is it to kill this boy so that you and the other man may survive your journey to safety?"

For each scenario, they were asked to make a moral judgment on a scale of 1 to 7, where 1 meant "Nothing wrong at all" and 7 meant "Extremely wrong."

The purpose of the experiment was to determine whether handwashing caused the students' moral judgments to be *less* severe (i.e. smaller numbers.)

The file `handwashing.txt` in the Data folder of the Files section on Canvas contains seven variables:

- `Condition` is a binary variable, where 1 indicates the student was in the handwashing group, and 0 indicates they were in the control group.
- `Dog`, `Trolley`, `Wallet`, `Plane`, `Resume`, and `Kitten` give the students' responses to the six moral judgment questions. (If you wish, you can read all the questions at `https://osf.io/r7qbd/` — but this is not necessary to complete this exam.) For all of these, a higher number means a strong moral judgment.

  Note: Since all questions are on this 1 to 7 scale, the distributions are not normal. However, the sample sizes are large enough that this shouldn't be too much of a problem.

(a) What are the experimental units in this study? What measurements are taken on each unit? How many independent samples are there?

(b) Create a variable called `Total` that gives the *total* score on the six moral judgment questions for each individual. (Remember to give R code.) Draw one or two well-labeled, informative graphs that let you compare the values of `Total` for the handwashing and control groups.

Parts (c) and (d) of this question are on the next page.

(c) Perform a one-tailed significance test to study whether handwashing affected the average `Total` score for the morality questions. Carefully define hypotheses, state the name of the test you are using, state the test statistic and degrees of freedom (if any), calculate a P-value, and write a substantive conclusion that someone who has not taken this statistics course can understand.

(d) Find 95% confidence intervals for:

(i) the expected value (population mean) of `Total` for handwashing,

(ii) the expected value of `Total` for control,

(iii) the difference in the expected value of `Total` between the two groups.

5 Trump names

The 2016 Illinois Republican primary had an unusual system. In addition to voting directly for a Presidential *candidate*, each Republican primary voter could vote for three *delegates*. Each Presidential candidate (e.g. Donald Trump) nominates three delegates of their choosing (e.g. Trump nominated Doug Hartmann, James Kammer, and Raja Sadiq in Illinois' Thirteenth District); these delegates have the name of Presidential candidate they pledge to support next to them. Voters will not generally recognize the names of the delegates — the sole purpose of the delegates is to vote for the Presidential candidate they are pledged to support.

At the website FiveThirtyEight[1], Dave Wasserman note that in Illinois' Thirteenth District, a Trump delegate named "Doug Hartman" received 31,937 votes, while a Trump delegate named "Raja Singh" received only 24,103 votes. Since all voters would generally know about the delegates was their name and which candidate they supported, Wasserman reasoned that if two delegates for the same candidate in the same district received very different numbers of votes, the only explanation for the difference was the delegate's name. Wasserman's hypothesis was that Trump supporters were more likely to vote for delegates with "white" names than delegates with "nonwhite" names. We will test this hypothesis as objectively and rigorously as we can, given the available data.

Data

The file `trump.txt`, posted on Canvas, contains 51 observations. The individuals (experimental units) in this study are delegates who were pledged to Trump. There are six variables:

- `last.name`: Last name of delegate.
- `first.name`: First name of delegate.
- `name.type`: Looking at data from the 2000 Census: Were most of the people with that last name "white" or "nonwhite"? (This variable was an idea of Princeton student Evan Soltas. Note that it does *not* say whether the delegate was white or nonwhite — but that's not a problem here, as all we are interested in is whether the *name* belongs primarily to whites or nonwhites.)
- `district`: The Illinois Congressional District of the candidate (1 to 18; District 9 is missing for some reason.)
- `trump.vote`: The percentage of all votes in the delegate's district that went to that delegate's candidate. Since all the delegates in this data set were pledged to Trump, this is the percentage of all votes in the district that went to Trump delegates.
- `delegate.vote`: The percentage of all votes in the delegate's district that went to that individual delegate.
- `vote.diff`: If the null hypothesis is true, you would expect each delegate to get one-third of all the votes for that candidate. This variable calculates the difference between the delegate's actual vote percentage, and their expected percentage under this null hypothesis. That is, it's just

$$\texttt{vote.diff} = \texttt{delegate vote} - \frac{\texttt{trump.vote}}{3}$$

[1] `https://fivethirtyeight.com/features/trump-voters-aversion-to-foreign-sounding-names-cost-him-delegates/`

Read the .txt file into R, and split the data into Trump delegates with "white names" and Trump delegates with "nonwhite names." e.g.:

```
trump.data <- read.table("trump.txt", header = TRUE)
trump.white <- subset(trump.data, name.type == "white")
trump.nonwhite <- subset(trump.data, name.type == "nonwhite")
```

Let μ_W be the expected value (i.e. population mean) of the vote differential for a hypothetical population of Trump delegates with white names. Let μ_{NW} be the expected value of the vote differential for a hypothetical population of Trump delegates with nonwhite names.

(a) One way to approach this problem is to treat the observations of `vote.diff` in the `trump.nonwhite` data frame as a random sample from some population. Perform a one-sample t-test of the hypothesis μ_{NW} is zero. Give a P-value, and carefully state a conclusion.

(b) Another way to approach the problem is to treat the "white names" and "nonwhite names" sets of data as two independent samples from two populations. Perform a two-sample t-test of the hypothesis μ_W is the same μ_{NW}. Give a P-value, and carefully state a conclusion.

(c) Out of your one-sample t-test and the two-sample t-test, which is closer to having its assumptions met? Explain. (Hint: Independence is more important than normality here.)

SAMPLE REPUBLICAN PRIMARY BALLOT

Kathy Michael

Kathy Michael, County Clerk

MARCH 15, 2016
001-ALLIN 01
MCLEAN COUNTY, ILLINOIS

Judge's Initials _____

To vote, darken the oval to the LEFT of your choice, like this ⬬. To cast a write-in vote, darken the oval to the LEFT of the blank space provided and write the candidate's name in that space. For specific information, refer to the card of instruction posted in the voting booth. If you tear, spoil, deface or erroneously mark this ballot, return it to the election judge and obtain another.

FEDERAL

FOR PRESIDENT OF THE UNITED STATES
(Vote for one)

- ◯ **JEB BUSH**
- ◯ **CHRIS CHRISTIE**
- ◯ **DONALD J. TRUMP**
- ◯ **TED CRUZ**
- ◯ **MARCO RUBIO**
- ◯ **RAND PAUL**
- ◯ **CARLY FIORINA**
- ◯ **MIKE HUCKABEE**
- ◯ **RICK SANTORUM**
- ◯ **JOHN R. KASICH**
- ◯ **BEN CARSON**
- ◯ _____ Write-in

FOR UNITED STATES SENATOR
(Vote for one)

- ◯ **JAMES T. MARTER**
- ◯ **MARK STEVEN KIRK**

STATE

FOR COMPTROLLER
(For an unexpired two year term)
(Vote for one)

- ◯ **LESLIE GEISSLER MUNGER**

CONGRESSIONAL

FOR REPRESENTATIVE IN CONGRESS
EIGHTEENTH CONGRESSIONAL DISTRICT
(Vote for one)

- ◯ **DARIN LaHOOD**

CONGRESSIONAL

FOR DELEGATE TO THE
NATIONAL NOMINATING CONVENTION
EIGHTEENTH CONGRESSIONAL DISTRICT
(PLEASE NOTE: Next to the name of each candidate for delegate appears in parentheses the candidate's preference for President of the United States or the word "uncommitted".)
(Vote for not more than three)

- ◯ **ROBERT BROWNING** (CHRISTIE)
- ◯ **MARY K. BROOKHART** (CHRISTIE)
- ◯ **DONNA K. THOMPSON** (CHRISTIE)
- ◯ **JIM EDGAR** (BUSH)
- ◯ **BILL BRADY** (BUSH)
- ◯ **RAYMOND POE** (BUSH)
- ◯ **KENT GRAY** (TRUMP)
- ◯ **SANDRA YEH** (TRUMP)
- ◯ **WILLIAM GRAFF** (TRUMP)
- ◯ **H. LEE NEWCOM** (CRUZ)
- ◯ **MICHAEL FLYNN** (CRUZ)
- ◯ **CHRISTIAN H. GRAMM** (CRUZ)
- ◯ **KRISTINA RASMUSSEN** (FIORINA)
- ◯ **PHIL CHILES** (FIORINA)
- ◯ **CHUCK WEAVER** (FIORINA)
- ◯ **JIL TRACY** (KASICH)
- ◯ **RANDY E. FRESE** (KASICH)
- ◯ **ERIK M. WOEHRMANN** (KASICH)
- ◯ **JUDITH A. HANKS** (CARSON)
- ◯ **MATTHEW HOPPOCK** (CARSON)
- ◯ **STEVEN A. WAILAND** (CARSON)
- ◯ **ELIZABETH BLANKENSHIP** (PAUL)
- ◯ **MIKE BROOKS** (PAUL)
- ◯ **TIMOTHY TARVIN** (PAUL)
- ◯ **DARIN LaHOOD** (RUBIO)
- ◯ **JASON BARICKMAN** (RUBIO)
- ◯ **MICHAEL D. UNES** (RUBIO)