P8122 Homework 2

Due: 09/30/2022 at 5pm

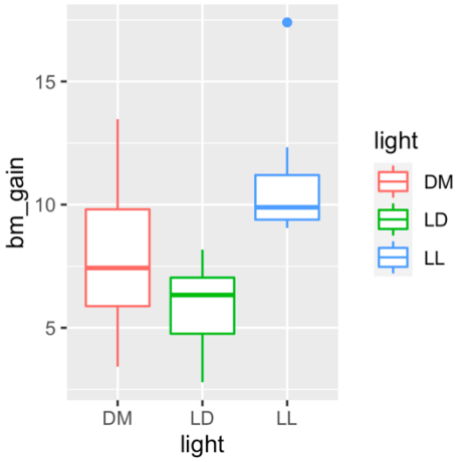
Roxy Zhang rz2570

Consider a study on the effect that light at night has on weight gain and other variables in mice. The goal of this study was to determine whether light at night may play a causal role in the obesity epidemic. The data comprises $n = 30$ mice randomized to three different treatment groups. All mice spent 16 hours in light, and the explanatory variable was the level of light during the remaining 8 hours. Some of the mice were randomized to darkness during those 8 hours (as is typical for regular mice), some were randomized to a dim light (equivalent to a TV on in the room for humans), and the remaining mice were exposed to bright light for all 24 hours. Mice are nocturnal, and typically most of their activity and eating happen at night. The hypothesis of this study was that having a light on at night may alter mouse eating habits and/or metabolism, and so increase body mass. The data are available on canvas. There are some missing values because one mouse died, and one mouse did not receive the full glucose injection for it's glucose tolerance test.

Table 1: Data Description

Variable Name	Variable Description
Light	Treatment group (dark, dim, bright)
BMGain	Change in body mass from the beginning of the experiment to the end (week 8)
Consumption	Average daily consumption (in grams)
Corticosterone	Blood corticosterene level (a measure of stress)
DayPct	Percentage of calories consumed during the day
GlucoseInt	Glucose intolerant at end of study? (Yes or No)
GTT15	Glucose level in the blood 15 minutes after a glucose injection
GTT120	Glucose level in the blood 120 minutes after a glucose injection
Activity	A measure of physical activity level

1. (5 points) We are interested in the causal effect of light at night on weight gain. Plot the outcome by treatment group.



2. (5 points) Here we will compare the mice exposed to darkness to the mice exposed to bright light overnight (once you have the code it is easy to rerun the analysis for the dim light group, if you are interested). Subset the data to only consider these two groups.

Consider LD as dark, DM as dim, LL as bright. Filter out DM rows.

```
df_sub = df %>%
  filter(light != "DM")
```

3. (15 points) Set up the data such that everything you will need has generic names (such as Y_{obs} or whatever you want to call them). Everything specific to the context of your data (variable names, sample sizes) should only be in your R Script here. Everything else should be generic so you can copy/paste it for later use. What quantities will you need to evaluate the causal effect of light at night on weight gain?

Rename light as "A", bm_gain as "Y". Re-code treatment "dark" as 0, "light" as 1.

```
df_setup = df_sub %>%
  select("light", "bm_gain") %>%
  rename(A = light,
         Y = bm_gain) %>%
  mutate(A = case_when(A == "LD" ~ 0,
                       A == "LL" ~ 1)) %>%
```

```
> head(df_setup)
# A tibble: 6 × 2
      A     Y
  <dbl> <dbl>
1     1  9.89
2     0  5.02
3     1  9.58
4     1 11.2
5     0  6.67
6     1  9.05
```

Quantities needed to evaluate causal effect:

- Average outcome in the dark group: $E(Y_{i0})$
- Average outcome in the bright group: $E(Y_{i1})$
- Population average causal effect (PACE): $E(Y_{i1} - Y_{i0})$
- Exact p-value

4. (10 points) Suppose we want the statistic to be the difference in means between the two treatment groups. Calculate T_{obs}.

```
> avg_y0
[1] 5.92625
> round(avg_y0, 2)
[1] 5.93
> round(avg_y1, 2)
[1] 11.01
> t_obs = round(avg_y1 - avg_y0, 2)
> t_obs
[1] 5.08
```

$T_{obs} = E(Y_{i1} - Y_{i0}) = 11.01 - 5.93 = 5.08$, where Y_{i1} stands for outcome in the bright group, Y_{i1} stands for outcome in the dark group.

5. (10 points) How many different possibilities are there for *A*? Enumerate all of these possibilities in a matrix. (Hint: it's probably easiest to first install the *ri* or *perm* package, have a look at the function *chooseMatrix* in R, it may come in handy.)

There are $C_9^{17} = 24310$ different possibilities for A.

Used *genperm()* function in *ri* to generate the matrix.

```
72  A = c(rep(1, 9), rep(0, 10))
73  Y = df_setup$Y
74
75  Abold = genperms(df_setup$A, maxiter = 30000)
76
77  Abold[, 1:15]
78
```

79:1 | Chunk 7 | R Markdown

Console | Terminal | Render | Jobs

R 4.1.2 · ~/Desktop/P8122/causal_inference/

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1
10	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

```
> ncol(Abold)
[1] 24310
```

6. (15 points) State the sharp null hypothesis of no difference. Calculate the test statistic under one of these possibilities for *A* (the first one), under the sharp null hypothesis.

The sharp null hypothesis: there is no treatment effect between two treatment groups.

$$H_0: \tau_i = Y_{0i} - Y_{1i} = 0$$

In this study, the sharp null hypothesis is that the weight gain for each mouse would be the same regardless it was assigned to light group or dark group.

The test statistic is -0.675.

```
> t_stat = mean(Y[A == 1]) - mean(Y[A == 0])
> t_stat
[1] -0.675
```

7. (10 points) Generate the exact randomization distribution for *T*, under the sharp null hypothesis of no difference.

```
rdist <- rep(NA, times = ncol(Abold))
for (i in 1:ncol(Abold)) {
  A_tilde <- Abold[, i]
  rdist[i] <- mean(Y[A_tilde == 1]) - mean(Y[A_tilde == 0])
}

rdist[1:50]
hist(rdist)
```

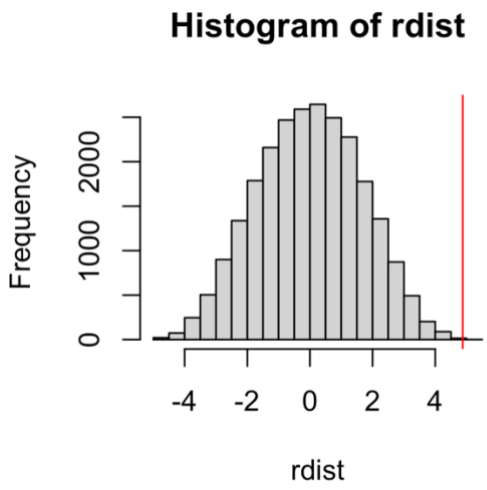
93:1 Chunk 8 R Markdown

Console Terminal Render Jobs

R 4.1.2 · ~/Desktop/P8122/causal_inference/

```
> rdist[1:50]
 [1] -0.67500000 -1.66666667 -1.36916667 -1.01736111 -2.08930556 -1.37388889
 [7] -2.09402778  0.52208333 -2.64888889  0.58583333  0.88333333  1.23513889
[13]  0.16319444  0.87861111  0.15847222  2.77458333 -0.39638889 -0.10833333
[19]  0.24347222 -0.82847222 -0.11305556 -0.83319444  1.78291667 -1.38805556
[25]  0.54097222 -0.53097222  0.18444444 -0.53569444  2.08041667 -1.09055556
[31] -0.17916667  0.53625000 -0.18388889  2.43222222 -0.73875000 -0.53569444
[37] -1.25583333  1.36027778 -1.81069444 -0.54041667  2.07569444 -1.09527778
[43]  1.35555556 -1.81541667  0.80069444 -0.68444444 -0.38694444 -0.03513889
[49] -1.10708333 -0.39166667
```

8. (10 points) Plot this distribution, and mark the observed test statistic.



9. (10 points) Calculate the exact p-value, based on this distribution.

$$(T(a,Y) \geq T(A,Y)|\tau = 0) = 4.113534e-05$$

```
103  pval <- mean(rdist >= t_obs)
104  pval
> pval
[1] 4.113534e-05
```

10. (10 points) What do you conclude?

The p-value is smaller than 0.05, we can conclude that under a 95% confidence level, we can reject the sharp null hypothesis that there is no difference in means of body mass change for mice in the dark group and light group. So light at night may play a causal role in the mouse's eating habits/metabolism, and therefore increase body mass.

Appendix: R code

```
```{r setup, include=FALSE}

# install ri from local file

#library(devtools)

#install_local("~/Desktop/P8122/causal_inference/hw2/ri_0.9.tar.gz")
```
```

```
library(ri)

library(tidyverse)
```

```{r}
# read and clean data
df = read_csv("~/Desktop/P8122/causal_inference/hw2/light.csv") %>%
  janitor::clean_names()
```

```{r}
# plot the outcome by treatment group
ggplot(data = df, aes(x = light, y = bm_gain, color = light)) +
  geom_boxplot()
```

```{r}
# only interested in mice exposed to darkness and bright
# subset the data into two groups
df_sub = df %>%
  filter(light != "DM")
```

```{r}
# set up the data such that everything needed has generic names
# re-code treatment: dark as 0 and bright as 1
df_setup = df_sub %>%
  select("light", "bm_gain") %>%
  rename(A = light,
         Y = bm_gain) %>%
  mutate(A = case_when(A == "LD" ~ 0,
                       A == "LL" ~ 1))
```

```{r}
# calculate the difference in means between the two treatment groups
y0 = data.frame(df_setup %>%
                  filter(A == 0) %>%
                  select(Y))

avg_y0 = mean(y0$Y)
round(avg_y0, 2)

y1 = data.frame(df_setup %>%
                  filter(A == 1) %>%
                  select(Y))
```

```
avg_y1 = mean(y1$Y)

round(avg_y1, 2)


t_obs = round(avg_y1 - avg_y0, 2)

t_obs
```


```{r}
# enumerate different possibilities for A in a matrix

A = c(rep(1, 9), rep(0, 8))

Y = df_setup$Y


Abold = genperms(df_setup$A, maxiter = 30000)


Abold[, 1:15]


ncol(Abold)
```



```{r}
# generate the exact randomization distribution for T

# under the sharp null hypothesis of no difference

rdist <- rep(NA, times = ncol(Abold))

for (i in 1:ncol(Abold)) {
```

```
  A_tilde <- Abold[, i]
  rdist[i] <- mean(Y[A_tilde == 1]) - mean(Y[A_tilde == 0])
}

rdist[1:50]
```

```{r}
# t statistics
t_stat = mean(Y[A == 1]) - mean(Y[A == 0])
t_stat
```

```{r echo=TRUE}
# p-value
pval <- mean(rdist >= t_obs)
pval

quant <- quantile(rdist, probs = 1 - pval)
```

```{r}
# plot this distribution, and mark the observed test statistic
hist(rdist)
abline(v = quant,col="red")
```

```
```