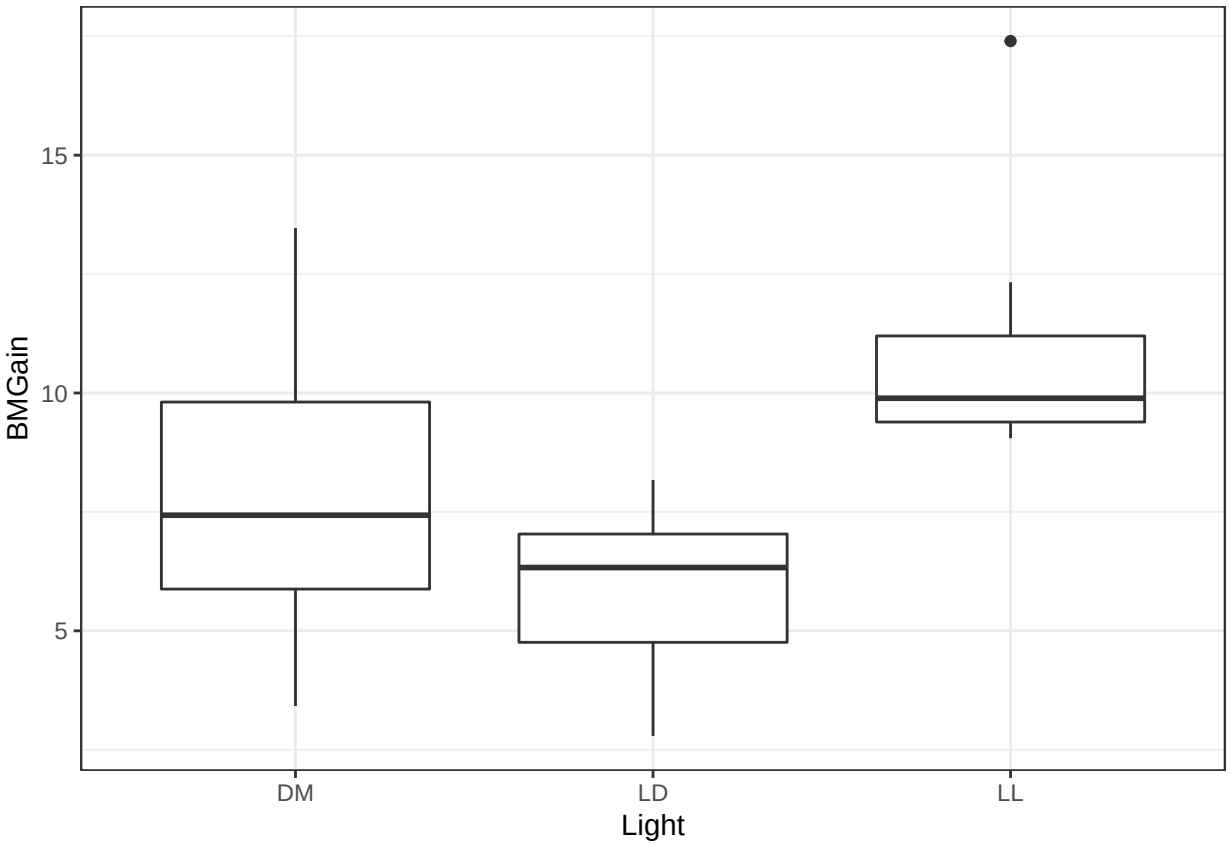
Homework 2

Ngoc Duong

10/13/2020

1. Plot the outcome by treatment group

We have outcome is body mass gain (from start to end of study), and three treatment groups – darkness (DM), dim light (LD), and bright light (LL).



Comments: Based on the boxplots, we can see some skewness in the LD (dim light) and LL (bright light) groups. The variability in outcome also seems to differ across groups. The median body mass gain in the LL group is the highest, while that of LD group is the lowest. In comparing DM (darkness) and LL treatment, we can see some noticeable difference in median body mass gain between the two groups.

2. We want to compare the mice exposed to darkness to the mice exposed to bright light. We subsetted data to only consider these two groups:

Light	BMGain	Corticosterone	DayPct	Consumption	GlucoseInt	GTT15	GTT120	Activity
DM	10.20	128.560	40.848	3.414	No	319.266	94.495	1409
DM	7.29	124.430	47.450	3.219	Yes	335.772	279.675	509
DM	7.57	98.517	56.429	3.613	Yes	343.590	412.821	2003

Light	BMGain	Corticosterone	DayPct	Consumption	GlucoseInt	GTT15	GTT120	Activity
DM	3.42	208.260	55.051	3.857	No	271.717	148.485	1084
DM	5.82	80.685	48.352	3.587	Yes	402.941	335.294	1848
DM	10.92	26.410	67.635	4.514	Yes	380.808	274.747	1841
DM	5.21	3.000	42.969	4.231	No	400.000	169.369	2716
DM	13.47	3.000	72.864	5.324	Yes	328.571	328.571	4622
DM	8.64	49.142	66.746	4.633	Yes	445.833	398.958	1744
DM	6.05	11.994	56.816	4.849	No	159.048	144.762	7253
LL	9.89	42.132	71.552	3.387	Yes	378.704	328.704	5752
LL	9.58	48.238	61.453	3.451	No	379.091	227.273	1256
LL	11.20	92.191	85.978	3.501	Yes	366.129	383.871	244
LL	9.05	51.999	64.827	4.240	No	392.373	250.000	931
LL	12.33	12.252	81.600	3.479	Yes	466.346	470.192	3582
LL	9.39	3.000	87.257	5.940	Yes	259.615	413.462	2657
LL	10.88	132.400	70.441	4.586	No	348.780	126.016	153
LL	9.37	8.615	84.415	4.873	Yes	335.652	286.957	4482
LL	17.40	66.679	81.636	7.177	Yes	435.644	405.941	6702

We end up with a subset of the data with 19 observations (9 in the Bright light group and 10 in the Darkness group).

3. Set up data with generic names

The quantities needed to evaluate the causal effect of light at night on weight gain are: mice body mass gain – BMGain, and Light (Darkness versus Bright light treatment). Here, we change the name of the outcome of interest – BMGain – to "Y_obs", treatment – Light – to "A", and the levels of Light such that "Bright Light" is 1, and "Darkness" is 0.

```
ll_dm_gen = ll_dm %>% rename(Y_obs = BMGain, A = Light) %>%
  mutate(A = ifelse(A == "LL", 1, 0)) #relevel the treatment to 1 and 0

#make object for the two quantities needed to evaluate causal effect
Y_obs = ll_dm_gen$Y_obs #outcome of interest (continuous variable)
A = ll_dm_gen$A #treatment assignment (1 or 0)
```

4. Calculate the statistic T_{obs} as the difference in means between two treatment groups:

$$T_{obs} = \frac{\Sigma_{i=1}^{n} A_i Y_{1i}}{N_1} - \frac{\Sigma_{i=1}^{n} A_i Y_{0i}}{N_0} = 3.151$$

5. There are 19 observations in the combinded dataset with two groups. We want to consider the number of possible ways to choose N_1 from the total N (choose 9 from 19, or equivalently, choose 10 from 19). This adds up to 92378 different possibilities.

Below are the first 10 possibilities (each column is a randomization scenario)

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1    1    1    1    1    1    1    1    1     1
##  [2,]    1    1    1    1    1    1    1    1    1     1
##  [3,]    1    1    1    1    1    1    1    1    1     1
##  [4,]    1    1    1    1    1    1    1    1    1     1
##  [5,]    1    1    1    1    1    1    1    1    1     1
##  [6,]    1    1    1    1    1    1    1    1    1     1
##  [7,]    1    1    1    1    1    1    1    1    1     1
##  [8,]    1    1    1    1    1    1    1    1    1     1
##  [9,]    1    0    0    0    0    0    0    0    0     0
```

```
## [10,]    0    1    0    0    0    0    0    0    0    0
## [11,]    0    0    1    0    0    0    0    0    0    0
## [12,]    0    0    0    1    0    0    0    0    0    0
## [13,]    0    0    0    0    1    0    0    0    0    0
## [14,]    0    0    0    0    0    1    0    0    0    0
## [15,]    0    0    0    0    0    0    1    0    0    0
## [16,]    0    0    0    0    0    0    0    1    0    0
## [17,]    0    0    0    0    0    0    0    0    1    0
## [18,]    0    0    0    0    0    0    0    0    0    1
## [19,]    0    0    0    0    0    0    0    0    0    0
```

6. The sharp null hypothesis of no individual difference is:

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

for all i

In other words, there is no treatment effect on the outcome for each individual observation.

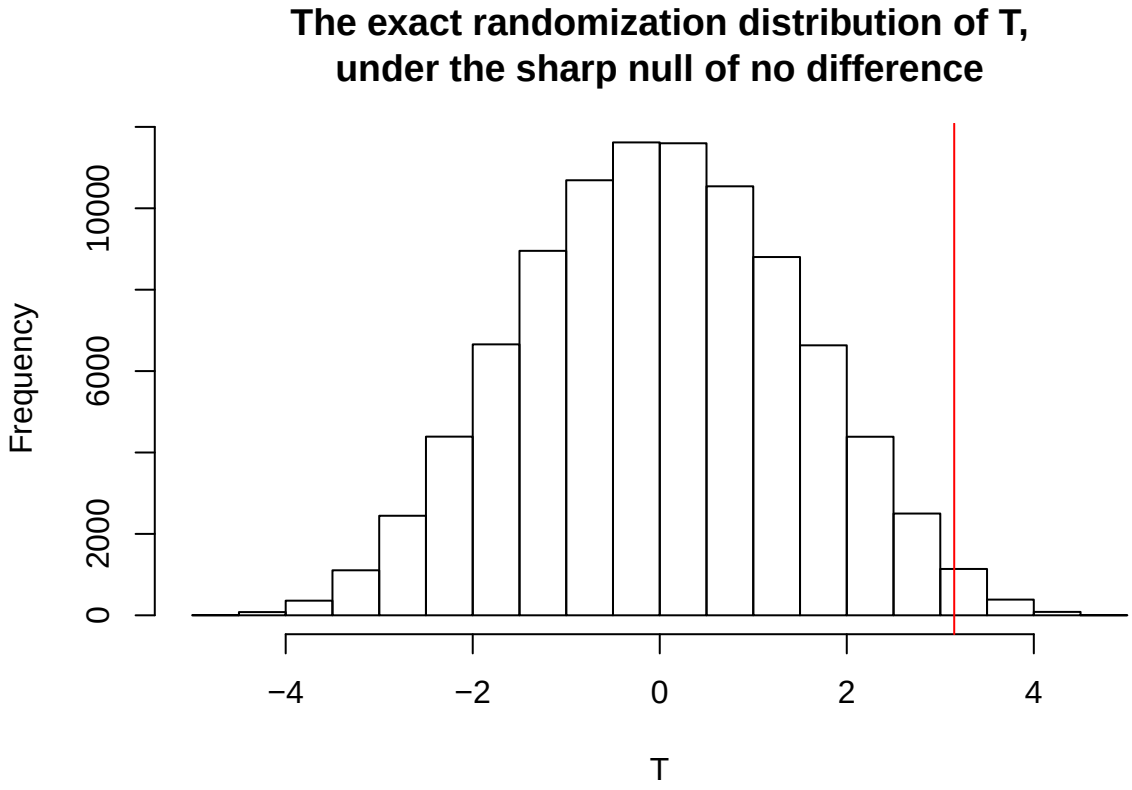
The test statistic under one of these probabilities for A (the first one), under the sharp null hypothesis is -2.1416

7. Generate the exact randomization distribution for T, under the sharp null hypothesis of no difference

```
##  [1] -2.14155556 -0.76722222 -1.06488889 -1.38788889 -2.27033333 -1.07333333
##  [7] -0.52655556 -1.39211111 -1.54622222  0.30311111 -2.09300000 -0.26055556
## [13] -0.55822222 -0.88122222 -1.76366667 -0.56666667 -0.01988889 -0.88544444
## [19] -1.03955556  0.80977778
```

Above are the first 20 values of T under the sharp null hypothesis.

8. Plot the distribution, and mark the observed test statistic



9.

The exact p-value based on this distribution is 0.0126.

10. Our observed statistic appears to be quite extreme in terms of the empirical exact randomization distribution. This is confirmed by the exact p-value for the observed test statistic based on this distribution being < 0.05, so we can reject the sharp null hypothesis and conclude that there is an individual effect of treatment type (Bright Light versus Darkness) on body mass gain in mice.

Appendix

```
#read in data
data = read_csv("light.csv")

#Q1 -- visualization
ggplot(data, aes(x = Light, y = BMGain)) + geom_boxplot() + theme_bw()

#Q2 -- subset treatment LL (bright) and DM (dark) from data
ll_dm = subset(data, Light %in% c("LL", "DM"))
ll_dm %>% arrange(Light) %>% knitr::kable()

#Q3 -- make generic names for variables of interest
ll_dm_gen = ll_dm %>% rename(Y_obs = BMGain, A = Light) %>%
mutate(A = ifelse(A == "LL", 1, 0)) #relevel the treatment to 1 (LL) and 0 (DM)

#make object for the two quantities needed to evaluate causal effect
Y_obs = ll_dm_gen$Y_obs #outcome of interest (continuous variable)
A = ll_dm_gen$A #treatment assignment (1 or 0)

#Q4 -- calculate observed statistic
t_obs = mean(Y_obs[A == 1]) - mean(Y_obs[A == 0])

#Q5 -- enumarate all these probabilies in a matrix Amat
Amat = chooseMatrix(19, 9)
#transpose so each column is a randomization scenario
Amat = t(Amat)
#show the first 10 randomization scenarios from all the possibilities
Amat[,1:10]

#Q6 -- calculate statistic under first possibility in Amat
A_tilde = Amat[,1] #choose the first randomization scenario
t_stat = mean(Y_obs[A_tilde == 1]) - mean(Y_obs[A_tilde == 0]) #obtain the test statistic for this scen
t_stat

#Q7 -- generate exact randomization distribution of statistic
#create a vector of empty spots/placeholders for each statistic under each randomization scenario
rdist <- rep(NA, times = ncol(Amat))

#run a for loop through each randomization scenario and calculate the corresponding test statistic, the
for (i in 1:ncol(Amat)) {
  A_tilde <- Amat[, i]
  rdist[i] <- mean(Y_obs[A_tilde == 1]) - mean(Y_obs[A_tilde == 0])
}
rdist[1:20]
```

```r
#Q8 -- visualize distribution
#p-value calculated as the proportion of statistics equal or more extreme than the statistic under
#all possible randomizations
pval <- mean(rdist >= t_obs)
quant <- quantile(rdist,probs = 1-pval) #get the quantile in the distribution of this pval
hist(rdist, xlab = "T", main = "The exact randomization distribution of T,\nunder the sharp null of no
abline(v = quant,col="red") #create red line to show where the t_stat is

#Q9 -- calculate exact p-value
#proportion of statistics equal or more extreme than the observed statistic under all possible randomiz
pval <- mean(rdist >= t_obs)
pval
```