Homework 3

Xinyi Lin

9/27/2019

```
# Input data
light = read.csv("./light.csv")

library(ggplot2)
library(tidyverse)
library(perm)
library(ri)
```

Question 1

The null hypothesis is no treatment effect on average. Let 0 means getting treatment of 8 hours darkness and 1 means getting treatment of 8 hours bright light. One simulated randomization is as following.

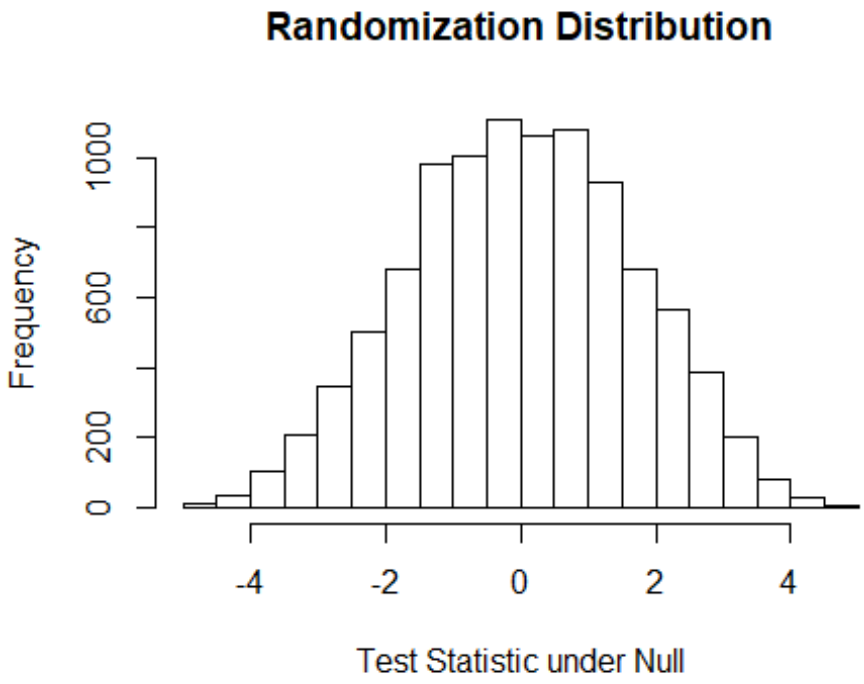
```
subset = light %>%
  filter(Light %in% c("LL", "LD"))
Y = subset$BMGain
A = rep(c(0,1),c(8,9))
set.seed(123)
A.sim = sample(A)
t_stat = mean(Y[A.sim==1])-mean(Y[A.sim==0])
round(t_stat, 2)

## [1] 1.18
```

The calculated T for this randomization assuming the null is around 1.18.

Quetion 2

The distribution simulated by randomizations is shown below.

```
set.seed(123)
n.sim = 10000 #number of simulations
T.sim = rep(NA, n.sim)
for (i in 1:n.sim) {
  A.sim = sample(A)
  T.sim[i] = mean(Y[A.sim==1], na.rm=TRUE) -  mean(Y[A.sim==0], na.rm=TRUE)
}
hist(T.sim, xlab = "Test Statistic under Null", main="Randomization
Distribution")
```



Question 3

```
T_obs = mean(Y[subset$Light=="LL"])-mean(Y[subset$Light=="LD"])
pval = mean(T.sim>=T_obs)
pval

## [1] 0
```

The approximate p-value is 0 and the exact p-value is 4.113533510^{-5}, so the approximate p-value is smaller than the exact p-value.

Question 4

```
subset_LL = subset %>%
  filter(Light == "LL")
subset_LD = subset %>%
  filter(Light == "LD")

set.seed(123)
outcome = ifelse(subset$Light=="LL", 1, 0)
perms = genperms(outcome)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 24310 to perform exact estimation.

probs = genprobexact(outcome)
low_ci = invert.ci(subset$BMGain,outcome,probs, perms,0.025)
high_ci = invert.ci(subset$BMGain,outcome,probs, perms,0.975)
```

The approximate 95% confidence interval of the sample average casual effects is (2.9, 7.51).

Interpretation: We have 95% confidence that the population average casual effects falls in (2.9, 7.51).

Question 5

$SACE = \overline{Y}_1 - \overline{Y}_0 = \frac{\sum_{i=1}^{N} Y_{1i}}{N} - \frac{\sum_{i=1}^{N} Y_{0i}}{N}$, where $\overline{Y}_1$ means the mean of weight gain among LL group and $\overline{Y}_0$ means the mean of weight gain among LD group. Y_{ji} menas the weight gain of each individual i.

$PACE = E(Y_{1i} - Y_{0i})$, where Y_{ji} means the weight gain of individaul i when it accept treatment j.

```
z = qnorm(0.975)
low_q2 = T_obs - z*sqrt(var(subset_LL$BMGain)/9+var(subset_LD$BMGain)/8)
high_q2 = T_obs + z*sqrt(var(subset_LL$BMGain)/9+var(subset_LD$BMGain)/8)
```

The point estimation of SACE is 5.08 and the approximate 95% confidence interval of the sample average casual effects is (2.92, 7.25).

The point estimation of PACE is 5.08 and the approximate 95% confidence interval of the sample average casual effects is (2.92, 7.25).

We can find that 95% confidence intervals of two methods are similar but the 95% confidence intervals get by Neyman's approach is narrower than that get by Fisher's approach.

Question 6

The two-way table is as following:

```
q6_data = light %>%
  filter(Light %in% c("LL", "LD")) %>%
  group_by(Light, GlucoseInt) %>%
  summarize(n = n()) %>%
  spread(key = GlucoseInt, value = n) %>%
  mutate(Yes = ifelse(is.na(Yes), 0, Yes)) %>%
  ungroup()
q6_data %>%
  knitr::kable()
```

Light	No	Yes
LD	8	0
LL	3	6

Null hypothesis: light at night and glucose intolerance in mice are independent.

If we use chi-square contingency table tests directly, results are as following:

Test statistic: $\sum_{all\ cells} \frac{(Observed-Expected)^2}{Expected}$, $Expected = \frac{row\ total \times column\ total}{overall\ total}$

```
q6_data %>%
  select(No, Yes) %>%
  chisq.test()

## Warning in chisq.test(.): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 5.5817, df = 1, p-value = 0.01815

#chisq.test(as.matrix(q6_data))
```

So the test statistic is 5.58 and p-value is 0.018.

Conclusion: As p-value is samller than 0.05, we reject the null hypothesis and conclude that light at night has relationship with glucose intolerance in mice.

If we use mean difference as test statistics and calculate the p-value via randomization test, results are as following:

Let 1 stands for GTT is Yes and 0 stands for GTT is No. Let test statistic $T_{obs} = \overline{Y}_1 - \overline{Y}_0$

```
set.seed(123)
GTT = ifelse(subset$GlucoseInt=="Yes", 1, 0)
T_obs = mean(GTT[subset$Light=="LL"])-mean(GTT[subset$Light=="LD"])
n.sim = 10000 #number of simulations
T.sim = rep(NA, n.sim)
for (i in 1:n.sim) {
  A.sim = sample(A)
  T.sim[i] = mean(GTT[A.sim==1], na.rm=TRUE) -  mean(GTT[A.sim==0],
na.rm=TRUE)
}
pval = mean(T.sim>=T_obs)
pval

## [1] 0.0063
```

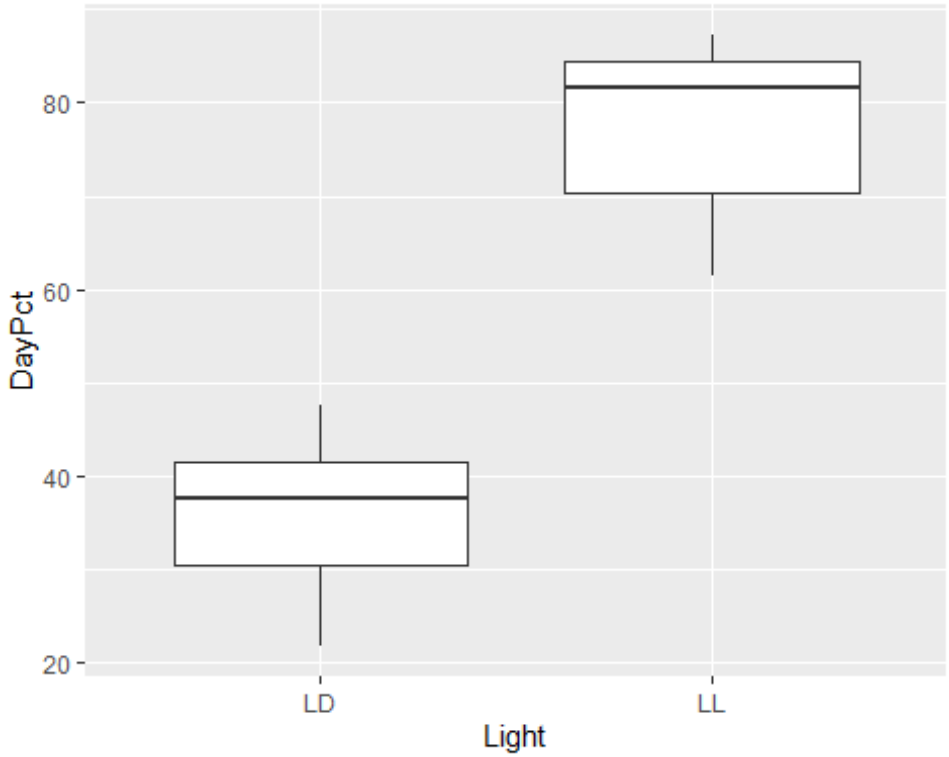
Conclusion: The p-value is 0.01. As p-value is samller than 0.05, we reject the null hypothesis and conclude that light at night has relationship with glucose intolerance in mice.

Question 7

Baseline BMI should not be significantly different on average, because baseline BMI might be a covariate. But if baseline BMI are significantly different, we still can make these two treatment group comparable by statification.

Question 8

```
light %>%
  filter(Light %in% c("LL", "LD")) %>%
  ggplot(aes(Light, DayPct)) +
  geom_boxplot()
```



According to the plot above, we can find out that the difference of percentages of calories consumed during the day on average between two treatment group is significant.

Null hypothesis: $\overline{DayPt_0} = \overline{DayPct_1}$. Alternative hypothesis: $\overline{DayPt_0} < \overline{DayPct_1}$

Test statistics: $T = \frac{\overline{DayPt_0} - \overline{DayPct_1}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}} = 9.26$

```
t.test(subset_LL$DayPct, subset_LD$DayPct, alternative = "greater")

##
##  Welch Two Sample t-test
##
```

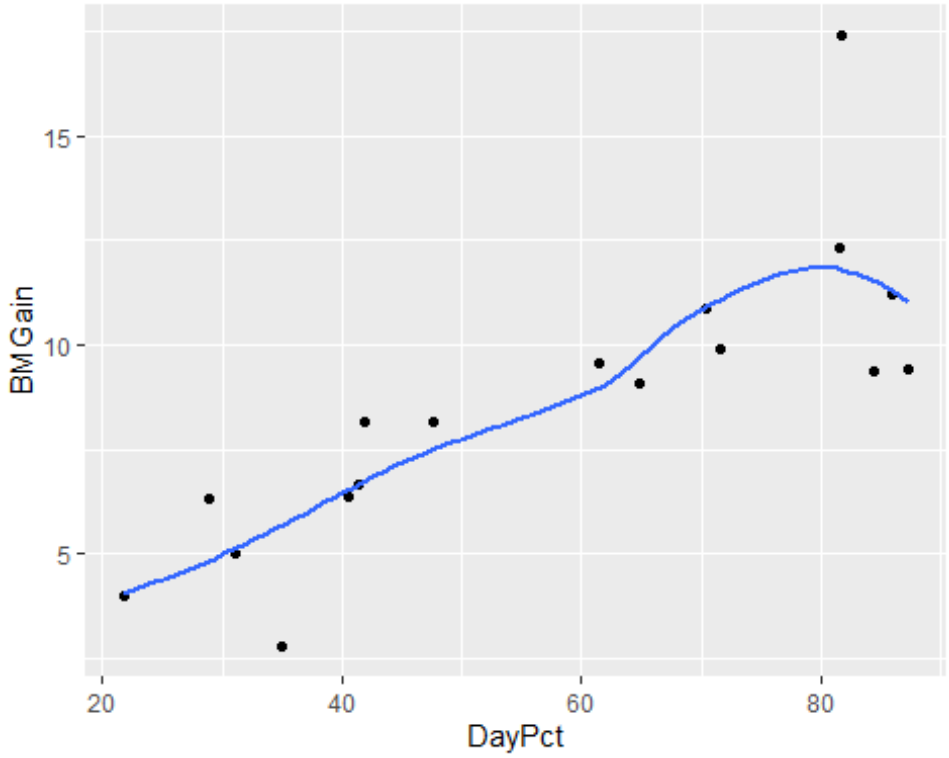
```
## data:  subset_LL$DayPct and subset_LD$DayPct
## t = 9.2616, df = 14.998, p-value = 6.796e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  32.87042      Inf
## sample estimates:
## mean of x mean of y
##  76.57322  36.02837
```

Related p-value is 6.796e-08.

Conclusion: As p-value is samller than 0.05, we reject the null hypothesis and conclude that $\overline{DayPt_0} < \overline{DayPct_1}$

Question 9

```
subset %>%
  ggplot(aes(x=DayPct, y=BMGain)) +
  geom_point()+
  geom_smooth(se=FALSE)
```



The plot of DayPct and BMGain is shown above. We can use correlation as test statistics.

Null hypothesis: The association between DayPct and BMGain isn't significant.

```
cor_obs = cor(subset$DayPct, subset$BMGain)
cor_obs
```

```
## [1] 0.8113803

# simulate the null
set.seed(123)
rdist = rep(NA, 1000)
for (i in 1:1000) {
  sim_DayPct = sample(subset$DayPct, 17, replace = TRUE)
  rdist[i] = cor(sim_DayPct, subset$BMGain)
}
pval = mean(rdist>=cor_obs)
pval

## [1] 0
```

So the test statistics cor_obs is around 0.811 and the related p-value is 0.

Conclusion: As p-value is samller than 0.05, we reject the null hypothesis and conclude that The association between DayPct and BMGain is significant.

Question 10

No. Currently, we only know exposed to light at night can cause weight gain and shift in calories consumption. So we can only conclude that weight gain and shift in calories comsumption are associated. However, we cannot get their casual relationship based on data because they might just associated because of exposed to light at night but not their casual relationship.

Appendix

```
# Input data
light = read.csv("./light.csv")
library(ggplot2)
library(tidyverse)
library(perm)
library(ri)
subset = light %>%
  filter(Light %in% c("LL", "LD"))
Y = subset$BMGain
A = rep(c(0,1),c(8,9))
set.seed(123)
A.sim = sample(A)
t_stat = mean(Y[A.sim==1])-mean(Y[A.sim==0])
round(t_stat, 2)
set.seed(123)
n.sim = 10000 #number of simulations
T.sim = rep(NA, n.sim)
for (i in 1:n.sim) {
  A.sim = sample(A)
  T.sim[i] = mean(Y[A.sim==1], na.rm=TRUE) -  mean(Y[A.sim==0], na.rm=TRUE)
}
hist(T.sim, xlab = "Test Statistic under Null", main="Randomization
```

```
Distribution")
T_obs = mean(Y[subset$Light=="LL"])-mean(Y[subset$Light=="LD"])
pval = mean(T.sim>=T_obs)
pval
subset_LL = subset %>%
  filter(Light == "LL")
subset_LD = subset %>%
  filter(Light == "LD")
set.seed(123)
outcome = ifelse(subset$Light=="LL", 1, 0)
perms = genperms(outcome)
probs = genprobexact(outcome)
low_ci = invert.ci(subset$BMGain,outcome,probs, perms,0.025)
high_ci = invert.ci(subset$BMGain,outcome,probs, perms,0.975)
z = qnorm(0.975)
low_q2 = T_obs - z*sqrt(var(subset_LL$BMGain)/9+var(subset_LD$BMGain)/8)
high_q2 = T_obs + z*sqrt(var(subset_LL$BMGain)/9+var(subset_LD$BMGain)/8)
q6_data = light %>%
  filter(Light %in% c("LL", "LD")) %>%
  group_by(Light, GlucoseInt) %>%
  summarize(n = n()) %>%
  spread(key = GlucoseInt, value = n) %>%
  mutate(Yes = ifelse(is.na(Yes), 0, Yes)) %>%
  ungroup()
q6_data %>%
  knitr::kable()
q6_data %>%
  select(No, Yes) %>%
  chisq.test()
#chisq.test(as.matrix(q6_data))
set.seed(123)
GTT = ifelse(subset$GlucoseInt=="Yes", 1, 0)
T_obs = mean(GTT[subset$Light=="LL"])-mean(GTT[subset$Light=="LD"])
n.sim = 10000 #number of simulations
T.sim = rep(NA, n.sim)
for (i in 1:n.sim) {
  A.sim = sample(A)
  T.sim[i] = mean(GTT[A.sim==1], na.rm=TRUE) -  mean(GTT[A.sim==0],
na.rm=TRUE)
}
pval = mean(T.sim>=T_obs)
pval
light %>%
  filter(Light %in% c("LL", "LD")) %>%
  ggplot(aes(Light, DayPct)) +
  geom_boxplot()
t.test(subset_LL$DayPct, subset_LD$DayPct, alternative = "greater")
subset %>%
  ggplot(aes(x=DayPct, y=BMGain)) +
  geom_point()+
```

```r
  geom_smooth(se=FALSE)
cor_obs = cor(subset$DayPct, subset$BMGain)
cor_obs
# simulate the null
set.seed(123)
rdist = rep(NA, 1000)
for (i in 1:1000) {
  sim_DayPct = sample(subset$DayPct, 17, replace = TRUE)
  rdist[i] = cor(sim_DayPct, subset$BMGain)
}
pval = mean(rdist>=cor_obs)
pval
```