STA 104 Take Home Project II

WEI LI (**998941394**)

JANET WEI (**999861062**)

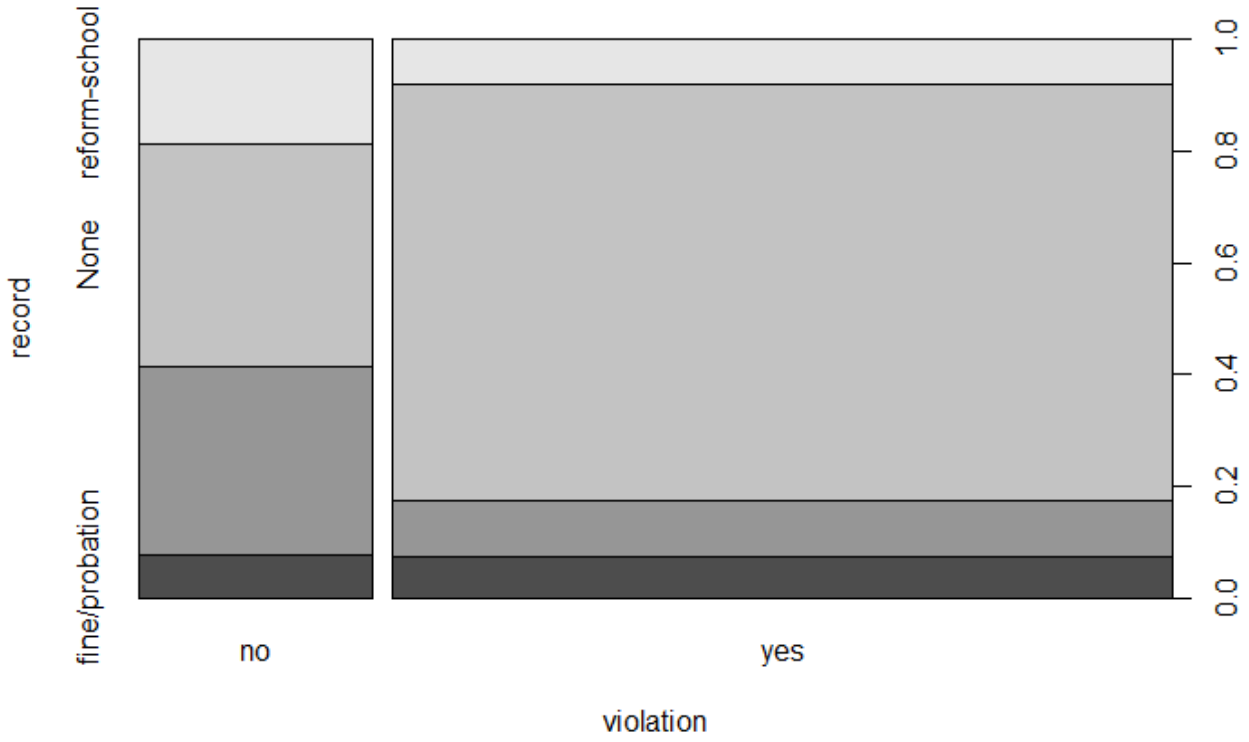
STA104 Nonparametric Statistics, Dr. Melcon
Department of Statistics, UC Davis
May 25th, 2016

Parole Data

I. **Introduction:** The question we are trying to answer is whether parole violation is dependent on their prior record or behavior (no prior record, went to reform school, served in jail, or received a fine/probation). This question is important to determine which factors may increase the likelihood of an individual violating parole, but also which ways may be more effective at preventing it. The approach we will take is permutation tests for contingency tables.

II. **Data Summary:** The frequency data is displayed in the table and figure below. There were approximately 3 times more probation violation than no violation, with no prior record having nearly 4 times as many individuals than the second most of jail time. The highest frequency between both variables is 132 in probation violation and no previous record, with the lowest of 4 for no probation violation and no previous fine/probation. Because of these extreme cell counts a nonparametric permutation test was conducted.

	fine/probation	jail	None	reform-school	
No violation	4	18	21	10	53
Violation	13	18	132	14	177
	17	36	153	24	230



III. **Analysis:** The null hypothesis is that there is no relationship between whether someone violates their parole and their prior record or behaviour, with the alternate being there is a difference. The observed parametric chi square test statistic was 26.9386 with a p-value of less than 0.001, suggesting there is a relationship between those two variables. For the nonparametric permutation test, the cutoff for a 0.05 significance level was determined to be 2.443378. The table below shows the test statistic for each comparison, with significant differences (in asterisks). The null hypothesis being that there is no difference for each prior record and whether they violate probation and the alternative that there is a relationship.

	jail	reform-school	None	fine/probation
yes vs. no	0.04944	4.181975*	-4.73031*	2.28925

IV. **Interpretation:** We found at the 0.05 significance level a difference between probation violation and prior record/behaviour with reform school resulting in more parole violation and no prior record resulting in less parole violations.

V. **Conclusion:** We conclude that reform schools increase the number of parole violations while having no previous record or behaviour leads to less parole violations. We recommend that reform schools cease if they want to reduce parole violations.

Appendix

```
df<-read.table("parole.csv",header=TRUE,sep=",")
df <- df[c("violation", "record")]
zee.table = table(df)
ni. = rowSums(zee.table)
n.j = colSums(zee.table)
the.test = chisq.test(zee.table,correct = FALSE)
eij = the.test$expected
chi.sq.obs = as.numeric(the.test$statistic)

R = 4000
r.perms = sapply(1:R,function(i){
  perm.data = df
  perm.data$violation =
sample(perm.data$violation,nrow(perm.data),replace = FALSE)
  chi.sq.i = chisq.test(table(perm.data),correct = FALSE)$stat
  return(chi.sq.i)
})
perm.pval = mean(r.perms >= chi.sq.obs)
n = sum(zee.table)
ni. = rowSums(zee.table)
n.j = colSums(zee.table)
all.pjG1 = zee.table[1,]/ni.[1] #all conditional probabilites
for row 1
all.pjG2= zee.table[2,]/ni.[2] #all conditional probabilites for
row 2
all.pbar = n.j/n #all probabilities regardless of group
all.Zij = c(all.pjG1 - all.pjG2)/sqrt(all.pbar*(1-
all.pbar)*(1/ni.[1] + 1/ni.[2])) #The z-test-statistics

r.perms.cutoff = sapply(1:R,function(i){
  perm.data = df
  perm.data$record =
sample(perm.data$record,nrow(perm.data),replace = FALSE)
  row.sum = rowSums(table(perm.data))
  col.sum = colSums(table(perm.data))
  all.pji = table(perm.data)[1,]/row.sum[1]
  all.pji.= table(perm.data)[2,]/row.sum[2]
  all.pbar = col.sum/sum(row.sum)
  all.Zij = c(all.pji - all.pji.)/sqrt(all.pbar*(1-
all.pbar)*(1/row.sum[1] + 1/row.sum[2]))
  Q.r = max(abs(all.Zij))
  return(Q.r)
})
alpha = 0.05
```

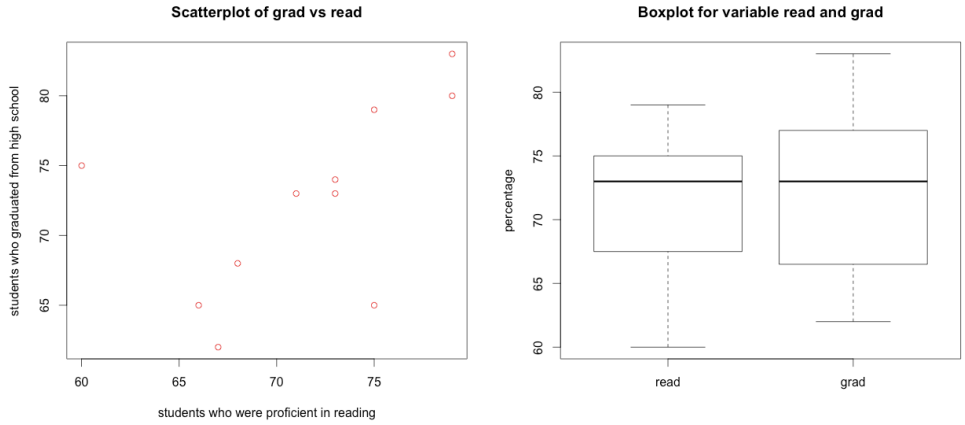
```
cutoff.q = as.numeric(quantile(r.perms.cutoff,(1-alpha)))

all.Zij = matrix(all.Zij,nrow=  1)
colnames(all.Zij) = c("jail","reform-
school","None","fine/probation")
rownames(all.Zij) = c("yes vs. no")
all.Zij
```

School Data

I. **Introduction:** We are interesting in the relationship between the percentage of student who were proficient in reading and the percentage who graduated high school. In this dataset, it contains three variables with the name of state, the percentage of student who were good at reading, and the percentage of the student who graduate from high school. We can assume that there is a linear relationship as when increasing the percentage of student who were proficient in reading, there will be increase the percentage of the high school graduation. Based on the dataset, we can get there are totally 11 observations for each variable, which is a small sample size. We will choose to use the permutation test and bootstrap method for this problem.

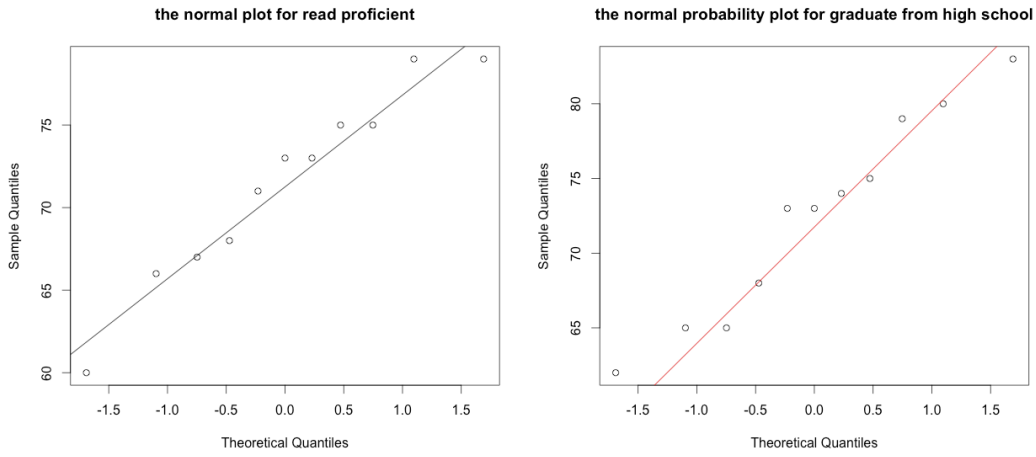
II. **Summary of the data:** Based on the basic knowledge, we get easily get the average percentage of the student who were proficient in reading is 71.45 with the standard deviation is 5.80; and the average percentage of student who graduate form high school is 72.45 with the standard deviation is 6.788. We also get the correlation between these two variables is 0.532. Then we can say there exists a positive relationship.

As we get the summary of the data in the figure below:



From the left scatterplot, we can find that the scatterplot of the percentage of the students who were proficient in reading and the percentage of the students who were graduated from high school exists a simple linear regression. And then, we also can say they have a positive correlation. From the right boxplot, we can conclude that the data are skewed.

Next, we do the qqplot below, as we can see the all plots are nearly to fall onto the qq line, we can say there is the normal probability in the percentage of the students who were proficient in reading and the percentage of the students who were graduated in high school. The assumption for this question is the data are in normality.



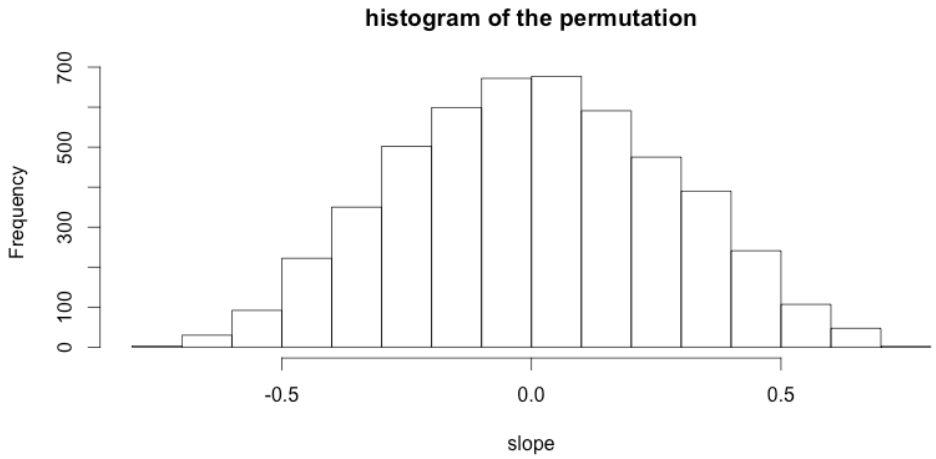
III. **Analysis:** From above, we do the null hypothesis and the alternative hypothesis:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

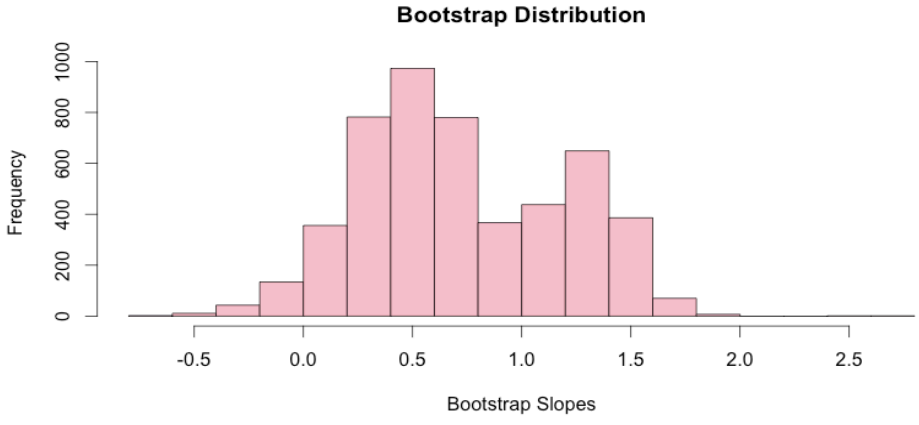
And, for these two variable, we were assume its exists a simple linear regression as: $Y = \beta_0 + \beta_1 X + \varepsilon$. From Introduction above, based on the small sample size, we do the permutation test and bootstrap method.

For permutation test, we do 5000 simulations. According to the distribution of the slope by using the permutation approach below:



It seems symmetric, and we can get the confidence interval at 95% significance level is [-0.4998569, 0.5203236], and the center of the confidence interval is 0.0102333 and the width is 1.020181.

For **bootstrap method**, for this dataset, the bivariate approach is the most appropriate because there is no fixed-X. We still do 5000 simulations. The distribution below, it seems symmetric. Then we get the confidence interval at 95% significance level is [-0.07513764,1.55035568], and the center of the confidence interval is 0.0737609, and the width is 1.625493.



IV. **Interpretation:** By the permutation approach, at 95% significance level, the confidence interval is [-0.4998569, 0.5203236], which means when the percentage of students who were proficient in reading increase 1%, and the percentage of students who were graduate from high school will also increase from -0.2998569 to 0.5203236. And for the bootstrap method approach, at 95% significance level, the confidence interval is [-0.07513764,1.55035568], which means when the percentage of students who were proficient in reading increase 1%, and the percentage of students who were graduate from high school will also increase from -0.07513764 to 1.55035568. And compare two width for two approaches, we find that the bootstrap method approach has the wider interval, we can say the bootstrap is more appropriate.

V. **Conclusion:** Based on the two confidence interval, we know that zero has contained in the interval, we failed to reject the null hypothesis, which means there should be no simple linear regression between the percentage of the students who were proficient in reading and the percentage of the students who were graduate from high school.

Appendix:

```
#load the data
setwd("~/Documents")
school = read.csv('school.csv')

#group the data
n = length(read)
meanr = mean(read)
rstandard_dev = sd(read)
meang = mean(grad)
gstandard_dev = sd(grad)
cor(read, grad)

# graphical
par(mfrow=c(1,2))
plot(read, grad, main='Scatterplot of grad vs read',
     xlab='students who were proficient in reading',
     ylab='students who graduated from high school',col =
"red")
boxplot(cbind(read, grad), main='Boxplot for variable read and
grad',
        xlab='', ylab='percentage')
par(mfrow=c(1,2))
qqnorm(read, main='the normal plot for read proficient')
qqline(read)
qqnorm(grad, main='the normal probability plot for graduate
from high school')
qqline(grad,col="red")

#permutation approach
set.seed(1000000)
OBS = lm(read ~ grad)$coefficients[2]
R = 5000
all.perm.slopes = sapply(1:R,function(i){
  the.data = school
  the.data$grad = sample(the.data$grad,n,replace = FALSE)
  slope.i = lm(read ~ grad, data = the.data)$coefficients[2]
  return(slope.i)
  })
lower.p = mean(all.perm.slopes < OBS)
upper.p = mean(all.perm.slopes > OBS)
two.p = mean(abs(all.perm.slopes) > abs(OBS))
all.slope.pval = c(lower.p,upper.p,two.p)
names(all.slope.pval) = c("lower tail","upper tail","two-
tailed")
```

```
all.slope.pval

#The histogram for each permutation
hist(all.perm.slopes, xlab='slope', main='histogram of the
permutation')

#95% CI for the permutation approach
alpha = 0.05
ci.percentile = as.numeric(quantile(all.perm.slopes,
c(alpha/2, 1-alpha/2)))
ci.percentile
#center for the CI percentile
mean(ci.percentile)
#the width
diff(ci.percentile)

#Bootstrap method
B = 5000
set.seed(9000000)
bi.boot  = sapply(1:B,function(i){
   boot.data =
school[sample(1:nrow(school),nrow(school),replace = TRUE),]
   boot.slope = lm(grad ~ read, data =
boot.data)$coefficients[2]
   return(boot.slope)
   })

#The histogram for bootstrap
hist(bi.boot, main = "Bootstrap Distribution",xlab =
"Bootstrap Slopes",col = "pink")

# use percentile to obtain CI
alpha = 0.05
ci.percentile = as.numeric(quantile(bi.boot, c(alpha/2, 1-
alpha/2)))
ci.percentile
# center
mean(ci.percentile)
# width
diff(ci.percentile)
```