STAT4051HW2

Mingming Xu

9/22/2019

1.Using the original sample to calculate the diffrence between two proportions.p_1 is the proportion of

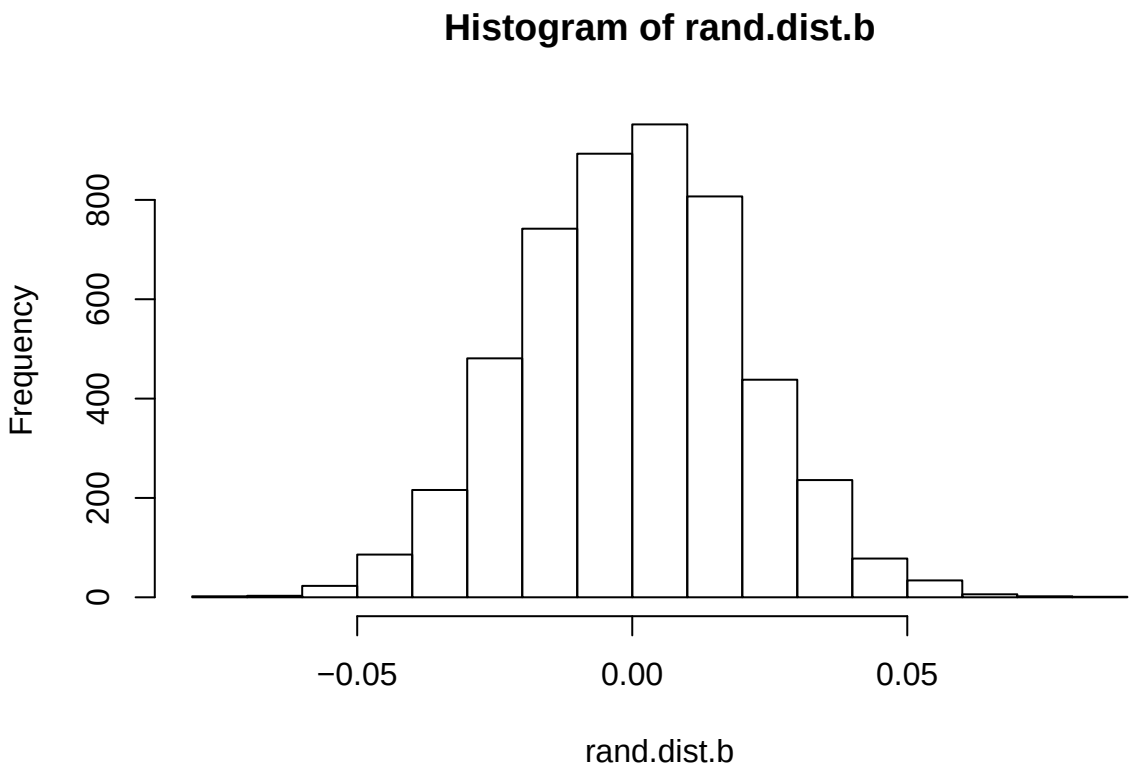
Americans who favor the death penalty today; p_2 is the proportion of Americans who favor the death penalty in

1990.$\hat{p_1} = 0.621$ and $\hat{p_2} = 1125/1500 = 0.75$,n1=1000,n2=1500. And the test hypotheses: $H_0 : p_1 - p_2 = 0$ vs.$H_a : p_1 - p_2 < 0$

```
##Under H0 is true
n1=1000
n2=1500
original.diff=0.621-0.75
```

```
n=5000
rand.dist.b=rep(NA,n)
for(i in 1 :n) {
  sample.vector=rbinom(2500,1,0.5)
  year_19=sample.vector[1:1000]
  year_90=sample.vector[1001:2500]
  rand.dist.b[i]=mean(year_19)-mean(year_90)
}

hist(rand.dist.b)
```



```
##pvalue
pval=mean(rand.dist.b<=original.diff)
pval
```

```
## [1] 0
```

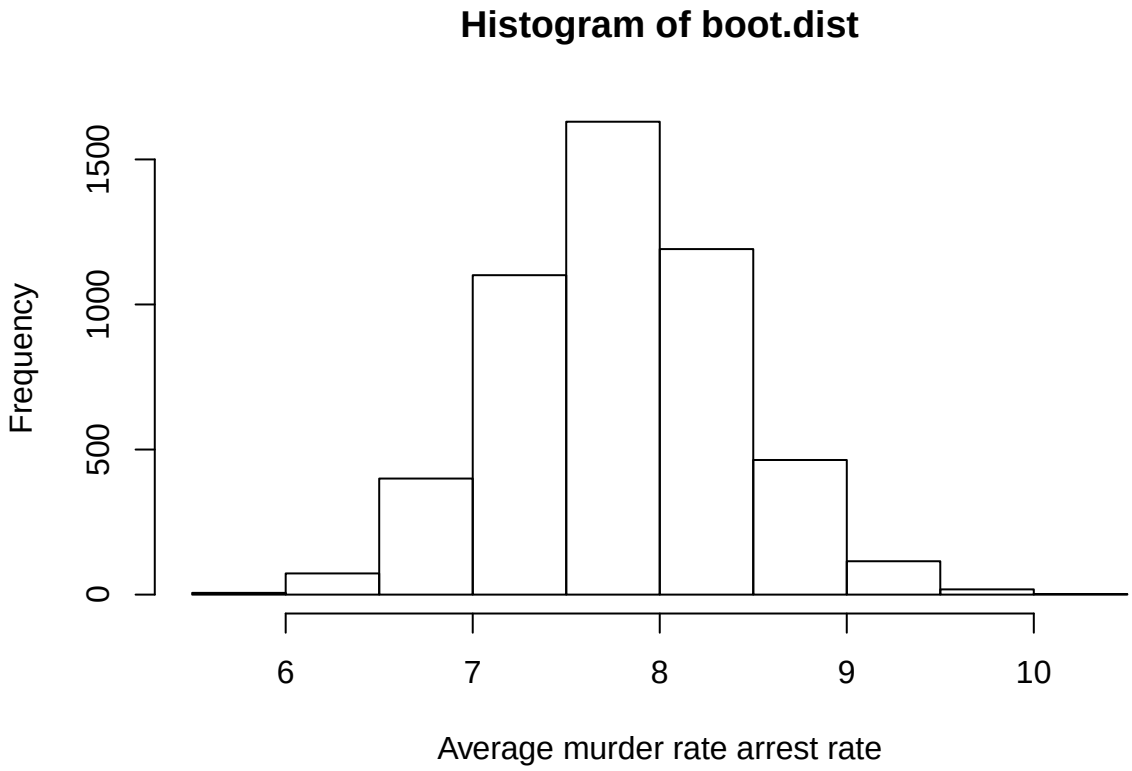
Because p-value is less than alpha=0.05, so we fail to reject the null hypothesis.

2.

```
##mean of original sample xbar
xbar=mean(USArrests$Murder)
sample.size=length(USArrests$Murder)
##repeatly sample with replacement n times
n=5000
boot.dist=rep(NA,n)

for(r in 1:n){
boot.sample=sample(USArrests$Murder,sample.size,replace=TRUE)
boot.dist[r]=mean(boot.sample)
}
##examine bootstrap distribution of mean to make sure that it is symmetric and bell-shaped.

hist(boot.dist,xlab="Average murder rate arrest rate")
```




```
##standard error of bootstrap distribution
sd=sd(boot.dist)

##create a 95% bootstrap t CI
xbar-qt(1-0.05,sample.size-1)*sd
```

```
## [1] 6.763455
```

So a 95% bootstrap t-confidence interval is (6.770782,100000),and it doesn't contain 5.

```
##construct a 95% percentile confidence interval
quantile(boot.dist,c(0,0.95))
```

```
##     0%    95%
## 5.5700 8.7921
```

So,a 95% percentile confidence interval contains 7.

3.

a.

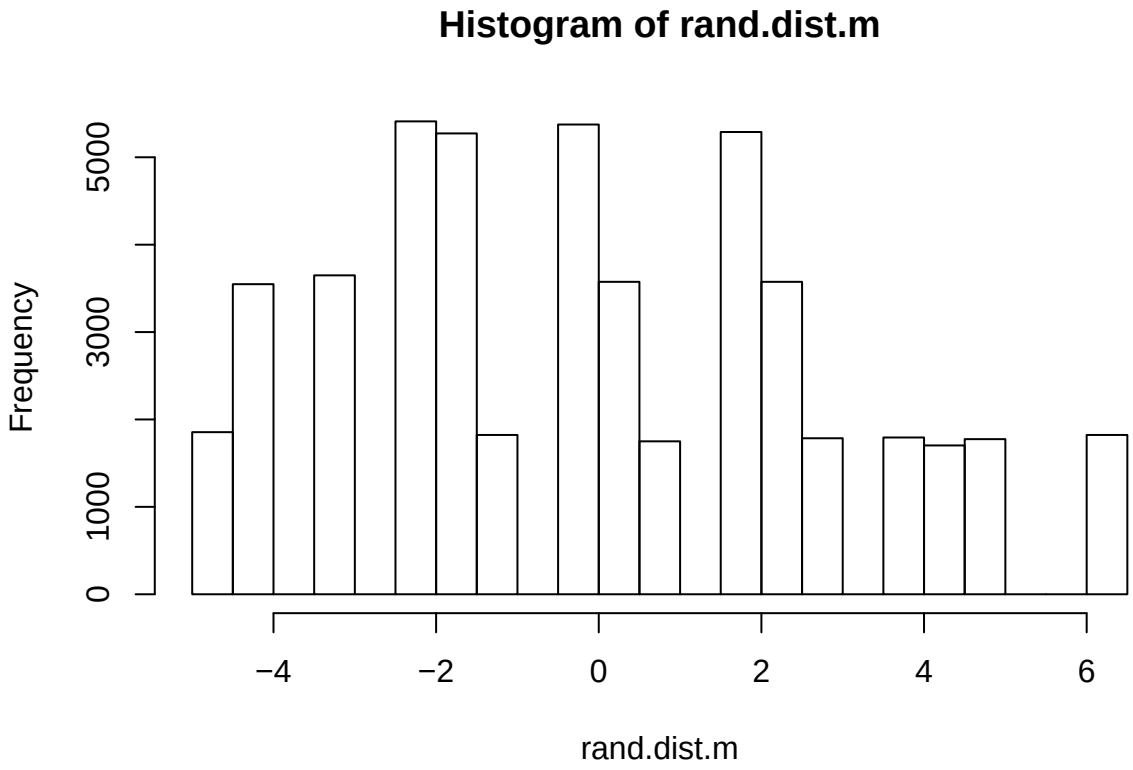
```
####Because it is always shows an error about install.package and combination functiaon when knitting
##Rmarkdown,I have to wirte these codes in text.
##install.packages("gtools")
##library(gtools)
##dominos=c(18, 20, 22, 24, 25, 25)
##papa=c(15, 21)
##mean.diff=rep(NA,28)
##output1=combinations(8,6,v=c(dominos,papa),set=FALSE)
##output2=combinations(8,2,v=c(dominos,papa),set=FALSE)
##j=28
##for (i in 1:28){
  ##mean.diff[i]=mean(output1[i,]) - mean(output2[j,])
  ##j=j-1
##}
##pval=mean(mean.diff<=mean(dominos)-mean(papa))
##pval
```

Because p-value=0.9285714 is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

b.

```
##original difference between two sample means
dominos=c(18, 20, 22, 24, 25, 25)
papa=c(15, 21)

ori.diff=mean(dominos)-mean(papa)
##construct the randomization distribution
deliver.time=data.frame(group=c(rep("Dominos",6),rep("Papa",2)),time=c(dominos,papa))
attach(deliver.time)
n=50003
rand.dist.m=rep(NA,n)
for(i in 1 :n) {
  sample.group=sample(group)
  dominos.m=time[sample.group=="Dominos"]
  papa.m=time[sample.group=="Papa"]
  rand.dist.m[i]=mean(dominos.m)-mean(papa.m)
}
hist(rand.dist.m)
```




```
##pvalue
pvalue.m=mean(abs(rand.dist.m)<abs(ori.diff))
pvalue.m
```

```
## [1] 0.7859328
```

Because p-value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

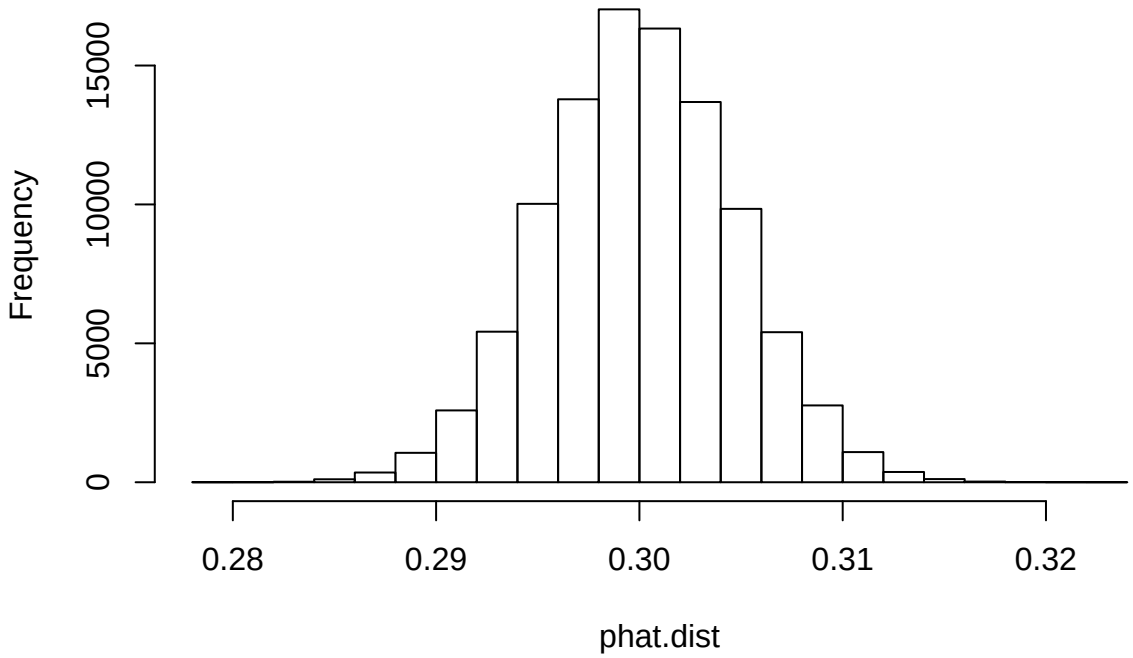
c. In HW1Q4,the p-value is 0.1424 which is smaller than the p-values in part a and b. The most reliable is the randmazition because it repeated many times.The least reliable is HW1Q4 bacause the sample size is too samll.

4.

Let p be the population proportion of firefighters who worked on the site for less than six months and had cardiovascular issues.So the sample mean $\hat{p} = \frac{3000}{9697}, p_0 = 0.3$

Hypothesis:$H_0 : p = 0.3$ vs.$H_a : p < 0.3$

```
##sample proportion for the data
phat=3000/9697
##To create the randomization distribution consistent with H0.By sampling replacement.
n=100000
sample_size=9697
rand.dist.f=rep(0,n)
for(i in 1:n){
  rand.dist.f[i]=rbinom(1,sample_size,0.3)
}
##compute the proortion of samples less than the original proportion phat
phat.dist=rand.dist.f/sample_size
hist(phat.dist,xlab="phat.dist")
```




```
p_value=mean(phat.dist<=phat)
p_value
```

```
## [1] 0.97801
```

```
##Bacuase the p-value is larger that \alpha=0.1, we fail to reject the null hypothesis.That is that we
```

The p-value is so closed to the p-value which is 0.9780137 I get in HW1 Q5.

****

Bonus

```
unemployment=read.csv("unemployment.csv")
attach(unemployment)
dim(unemployment)
```

```
## [1] 12  2
```

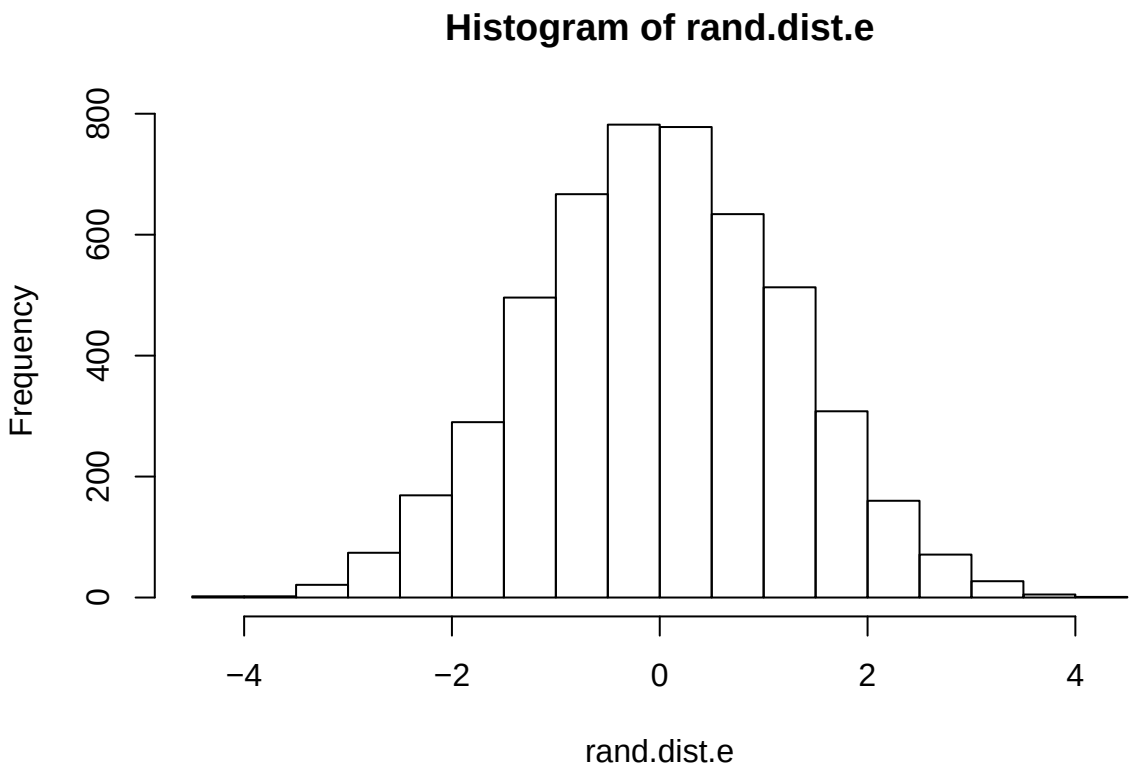
```
##the diffrence between Hischool and college based on each year
mean=unemployment$HiSchool-unemployment$College
##the sample diffence
mean.m=mean(mean)
```

1.

```
n=5000
rand.dist.e=rep(NA,n)
for(i in 1:n){
sign=sample(c(-1,1),12,replace=TRUE)
new.mean=mean*sign
rand.dist.e[i]=mean(new.mean)

}
```

```r
hist(rand.dist.e)
```




```r
##p value
pval.e=mean(abs(rand.dist.e)>abs(mean.m))
pval.e
```

```
## [1] 0
```

2.

```r
##original diffrence mean between two samples
xbar.h=mean(unemployment$HiSchool)
xbar.c=mean(unemployment$College)
orig.diff=xbar.h-xbar.c
##repeatly sample with replacement n times
sample.size.h=length(unemployment$HiSchool)
sample.size.c=length(unemployment$College)
n=5000
boot.dist.diff=rep(NA,n)

for (i in 1:n) {
  hi.sample=sample(HiSchool,sample.size.h,replace=TRUE)
  co.sample=sample(College,sample.size.c,replace=TRUE)
  boot.dist.diff[i]=mean(hi.sample)-mean(co.sample)
  }

sd(boot.dist.diff)
```

```
## [1] 0.4458782
```

```r
## a 95% bootrap t-confidence interval
(xbar.h-xbar.c)+c(-1,1)*qt(1-0.05/2,12)*sd(boot.dist.diff)
```

```
## [1] 3.136848 5.079819
```