BIOST 544 Homework 4

Ivy Zhang

12/3/2021

Data Loading and Cleaning

First, I load the data and try to do some cleaning. I decided to use the two BMD variables, smoking status variable(whether the participants ever smoked), age variables, alcohol in a month variable, and categorical income variables to do this project. Personally, without futuer research, I think age, alcohol and income should be confounding variables when we are trying to research the potential association between BMD and smoking status. I removed all observations without meaningful response, and doing complete case analysis.

```
#Data Loading and Cleaning
library(haven)
library(dplyr)
library(tidyr)
library(knitr)
library(ggplot2)
SWAN = read_dta("~/Desktop/R hw/28762-0001-Data.dta")
swan.use= SWAN%>%select(HPBMDT0,SPBMDT0, SMOKERE0, AGE0,INCOME0,ALCHMON0)
swan.use$income = swan.use$INCOME0
swan.use$income = as.factor(ifelse(swan.use$income %in% c(-9,-7,-8), NA, swan.use$income))
swan.use$smoke = swan.use$SMOKERE0
swan.use$smoke = as.factor(ifelse(swan.use$smoke %in% c(-9,-8), NA, swan.use$smoke))
swan.use = na.omit(swan.use)
swan.use$smoke = ifelse(swan.use$smoke == 1, "No","Yes")
swan.use$smoke01 = ifelse(swan.use$smoke == "Yes",1,0)
swan.use$alcohol.month = as.factor(swan.use$ALCHMON0)
#Probability of being each arm ignoring confounding
hpbmd.by.smoking = swan.use %>% group_by(smoke) %>% summarise(mean.hpbmd = mean(HPBMDT0))
kable(hpbmd.by.smoking)
```

smoke	mean.hpbmd
No	0.9524899
Yes	0.9729808

```
spbmd.by.smoking = swan.use %>% group_by(smoke) %>% summarise(mean.spbmd = mean(SPBMDT0))
kable(spbmd.by.smoking)
```

smoke	mean.spbmd
No	1.072005
Yes	1.080850

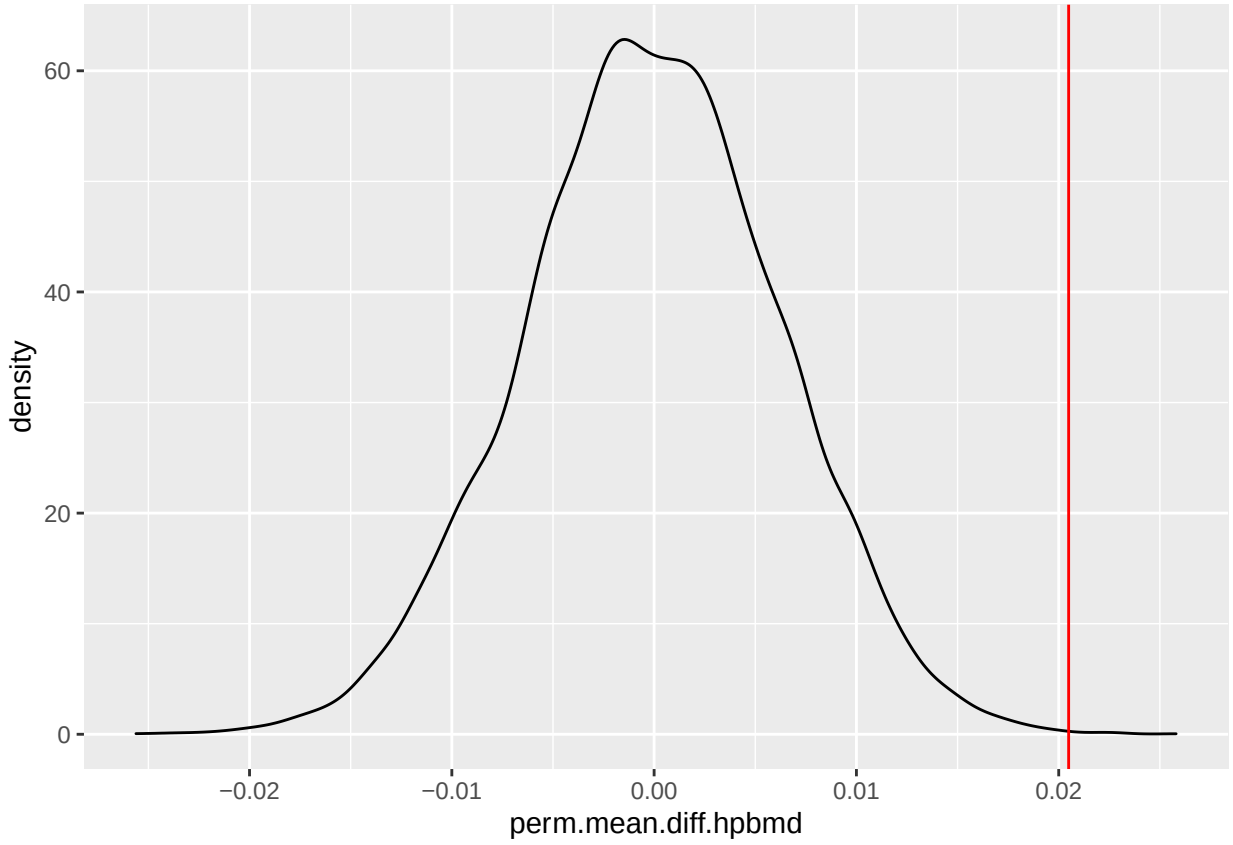
```
mean.diff.hpbmd = hpbmd.by.smoking $mean.hpbmd[2] - hpbmd.by.smoking$mean.hpbmd[1]
mean.diff.spbmd = spbmd.by.smoking$mean.spbmd[2] - spbmd.by.smoking$mean.spbmd[1]
```

In the previous two table, I show the mean hip BMD and the spine BMD for the two smoking status.

Permutation

Then I do the permutation test to see whether the difference in mean hip BMD values in the two smoking status, and the mean spine BMD values in the two smoking status are caused due to the chance.

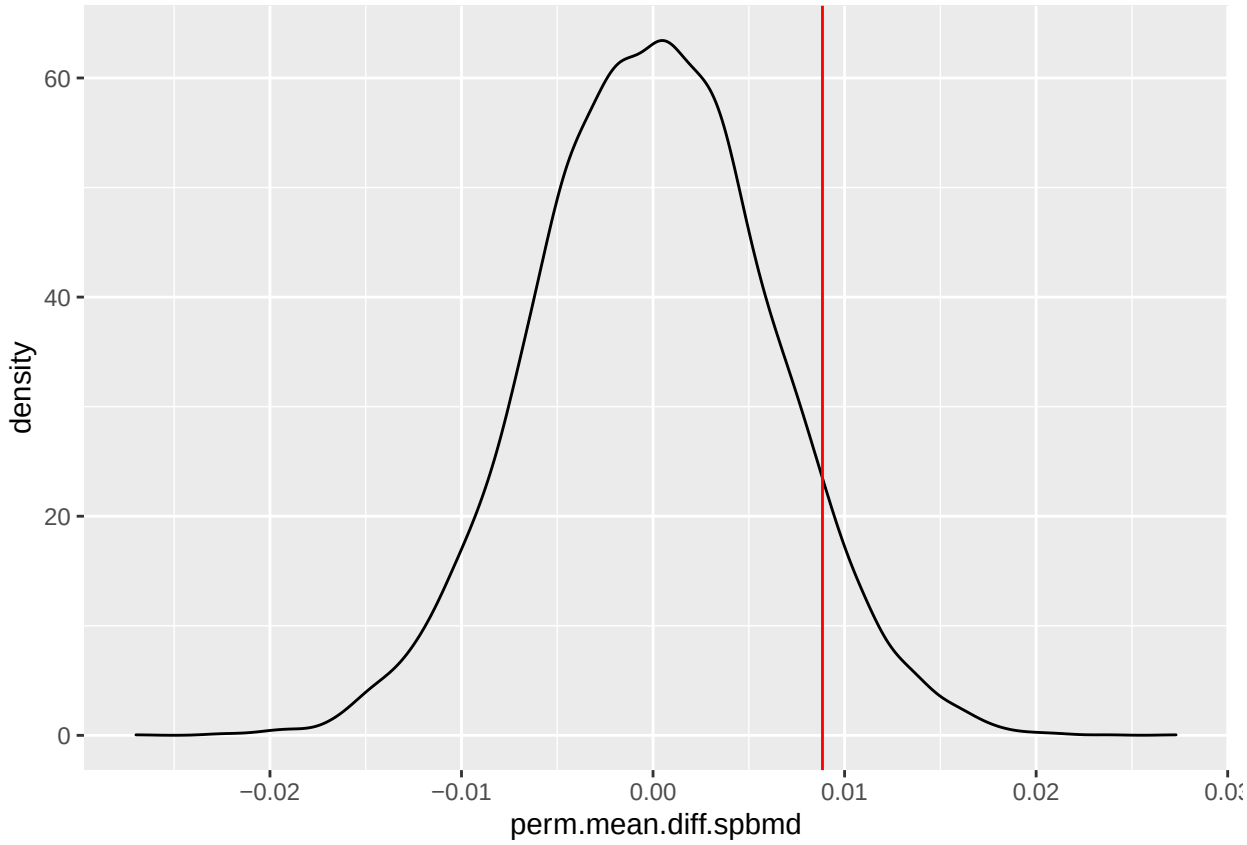
```
#Permutation
do.one = function(outcome, label){
  perm.label = sample(label)
  return(mean(outcome[perm.label == "Yes"]) - mean(outcome[perm.label == "No"]))
}
set.seed(1)
sampling.dist.hpbmd = with(swan.use, replicate(1e4, do.one(HPBMDT0,smoke)))
sampling.dist.spbmd = with(swan.use, replicate(1e4, do.one(SPBMDT0,smoke)))
#Permutation distribution of hip BMD
ggplot(data.frame(perm.mean.diff.hpbmd = sampling.dist.hpbmd),
       aes(x = perm.mean.diff.hpbmd, y = ..density..)) +
  geom_density() + geom_vline(xintercept = mean.diff.hpbmd, color = "red")
```




```
#P-value of our observed data
mean(abs(sampling.dist.hpbmd)>mean.diff.hpbmd)
```

```
## [1] 0.0013
```

```
#Permutation distribution of spine BMD
ggplot(data.frame(perm.mean.diff.spbmd = sampling.dist.spbmd),
       aes(x = perm.mean.diff.spbmd, y = ..density..)) +
  geom_density() + geom_vline(xintercept = mean.diff.spbmd, color = "red")
```




```
#P-value of our observed data
mean(abs(sampling.dist.spbmd>mean.diff.spbmd))
```

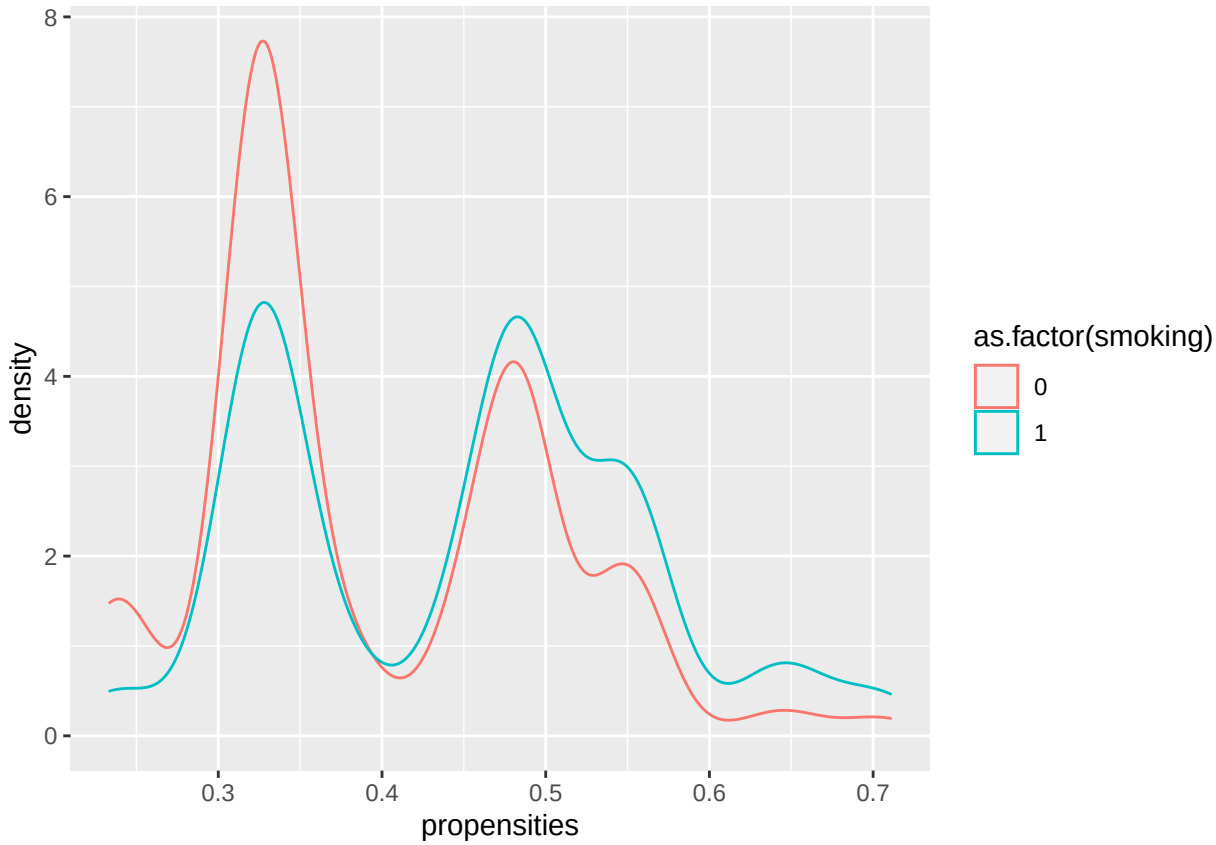
```
## [1] 0.073
```

The previous two plots show the sampling distribution of the permutation of the differences in the two BMD values if we randomize the smoker label. As we can see, the permutation result shows the difference is highly unlikely due to chance in hip BMD value and spine BMD value with p-value 0.0013 and 0.073.

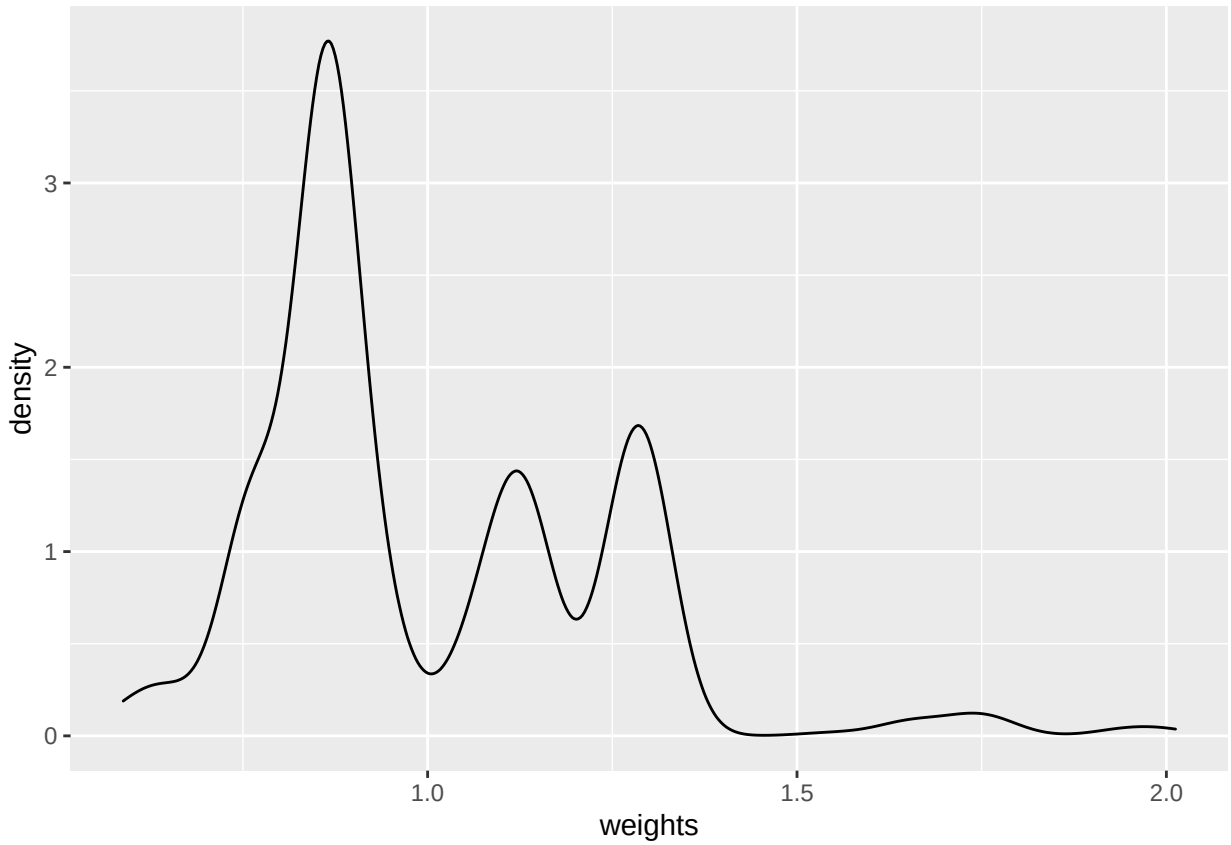
Confounding

I will use propensity scores/inverse probability weighting to help us address the confounding variable. We will use the logistic regression to build a model for the propensities.

```
#Propensity Model
propen.model = glm(smoke01~AGE0+income+alcohol.month, family = binomial, data = swan.use)
propensities = predict(propen.model,data = swan.use, type = "response")
ggplot(data.frame(propensities=propensities, smoking=swan.use$smoke01),
       aes(x = propensities, y = ..density.., color = as.factor(smoking))) + geom_density()
```




```
trunc.prop = propensities %>% pmax(0.05) %>% pmin(0.95)
npat = nrow(swan.use)
weights = rep(0,npat)
representative.propen = sum(swan.use$smoke == "Yes") /npat
actual.propen = trunc.prop
smoke.ind = which(swan.use$smoke == "Yes")
weights[smoke.ind] = representative.propen/actual.propen[smoke.ind]
weights[-smoke.ind] = (1 - representative.propen)/(1-actual.propen[-smoke.ind])
ggplot(data.frame(weights=weights),aes(x = weights, y = ..density..))+geom_density()
```




```
(smoking.prob.est.hpb = with(swan.use, mean((weights*HPBMDT0)[smoke.ind])))
```

```
## [1] 0.9715091
```

```
(nonsmoking.prob.est.hpb = with(swan.use,mean((weights*HPBMDT0)[-smoke.ind])))
```

```
## [1] 0.956533
```

```
(diff.est.hpb  = smoking.prob.est.hpb - nonsmoking.prob.est.hpb)
```

```
## [1] 0.01497607
```

```
(smoking.prob.est.spb = with(swan.use, mean((weights*SPBMDT0)[smoke.ind])))
```

```
## [1] 1.079851
```

```
(nonsmoking.prob.est.spb = with(swan.use,mean((weights*SPBMDT0)[-smoke.ind])))
```

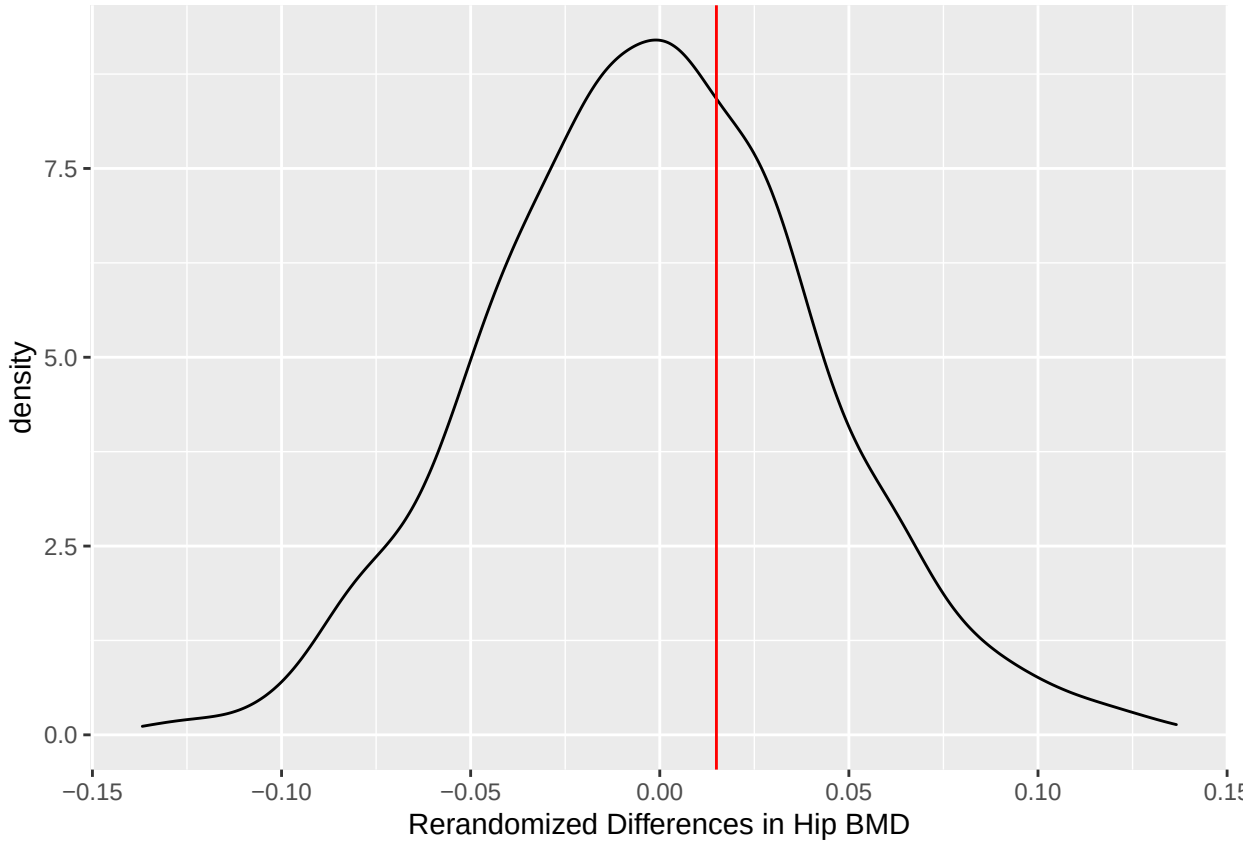
```
## [1] 1.07419
```

```
(diff.est.spb  = smoking.prob.est.spb - nonsmoking.prob.est.spb)
```

```
## [1] 0.005661793
```

```
do.one.propen = function(data,propen){
  n = nrow(data)
  label = rbinom(n,1,propen)
  weights = rep(0,n)
  representative = sum(label == 1) / n
  actual = propen
  ind.t = which(label == 1)
```

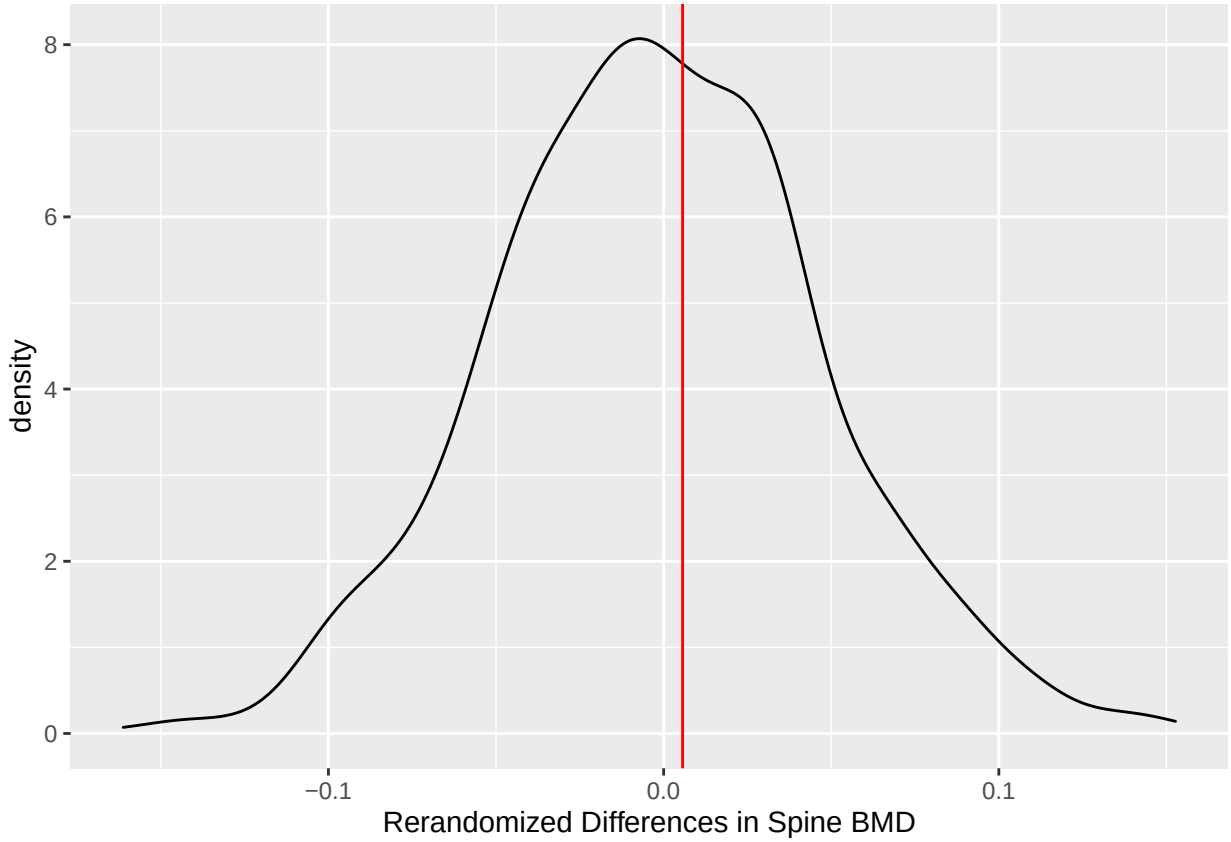
```
  weights[ind.t] = (representative/actual)[ind.t]
  weights[-ind.t] = ((1-representative)/(1-actual))[-ind.t]
  smoking.prob.est.hpb.p = with(data, mean((weights*HPBMDT0)[ind.t]))
  nonsmoking.prob.est.hpb.p = with(data,mean((weights*HPBMDT0)[-ind.t]))
  diff.est.hpb.p  = smoking.prob.est.hpb.p - nonsmoking.prob.est.hpb.p

  smoking.prob.est.spb.p = with(data, mean((weights*SPBMDT0)[ind.t]))
  nonsmoking.prob.est.spb.p = with(data,mean((weights*SPBMDT0)[-ind.t]))
  diff.est.spb.p  = smoking.prob.est.spb.p - nonsmoking.prob.est.spb.p

  return(c(diff.est.hpb.p,diff.est.spb.p))
}
set.seed(1)
randomized.diffs = matrix(replicate(1e3, do.one.propen(swan.use, trunc.prop)),
                          ncol = 2, nrow = 1e3, byrow = TRUE)
colnames(randomized.diffs) = c("Rerandomized Differences in Hip BMD",
                               "Rerandomized Differences in Spine BMD")
randomized.diffs = as.data.frame(randomized.diffs)
#Randomization differences in hip bmd after adjusting confounding
ggplot(randomized.diffs, aes(x = `Rerandomized Differences in Hip BMD`,
                             y = ..density..))+geom_density() +
  geom_vline(xintercept = diff.est.hpb,color = "red")
```




```
mean(abs(randomized.diffs$`Rerandomized Differences in Hip BMD`>diff.est.hpb))
```

```
## [1] 0.342
```

```
#Randomization differences in spine bmd after adjusting confounding
ggplot(randomized.diffs, aes(x = `Rerandomized Differences in Spine BMD`,
                             y = ..density..))+geom_density() +
  geom_vline(xintercept = diff.est.spb,color = "red")
```




```
mean(abs(randomized.diffs$`Rerandomized Differences in Spine BMD`>diff.est.spb))
```

```
## [1] 0.433
```

After adjusting the confounding, it seems that the differences in two smoking status don't vary too much in our data. I still do the randomization to see how our original value compared to do the randomization distribution for the two BMD values. The two sampling distribution have shown in the previous graph and the p-value are both very large compared to 0.05. It means that our data does not offer information to support the hypothesis that smoking status is related to the BMD.

Since this result is very different from our permutation test result, we may think the confounding working heavily in this data and this analysis.

Standardization

Then we will do the standardization to do the confirmatory analysis.

```
#Standardization
outcome.regression.hipbmd = lm(HPBMDT0~smoke01+AGE0+income+alcohol.month, data =swan.use)
outcome.regression.spbmd = lm(SPBMDT0~smoke01+AGE0+income+alcohol.month, data =swan.use)
swan.smoke = swan.use %>% mutate(smoke01 = 1 )
swan.nonsmoke = swan.use %>% mutate(smoke01 = 0)
```

```r
#Standardization difference in hip BMD
(standarlized.est.hip = mean(predict(outcome.regression.hipbmd, swan.smoke, type = "response")-
                          predict(outcome.regression.hipbmd,swan.nonsmoke, type = "response")))
```

```
## [1] 0.01542505
```

```r
#Standardization difference in spine BMD
(standarlized.est.sp = mean(predict(outcome.regression.spbmd, swan.smoke, type = "response")-
                          predict(outcome.regression.spbmd,swan.nonsmoke, type = "response")))
```

```
## [1] 0.006378691
```

Based on the standardization results, we think our data do not offer enough information to support the hypothesis that smoking is related to the BMD valuse.