BIOST 544 HOMEWORK1

Ivy Zhang

10/13/2021

(1)

We first calculate the estimate probability of each age sub-group by calculating the proportion of pariticipants who survive over 400 days.

```
nsclc <- read.csv("~/Desktop/R hw/nsclc-modified.txt", sep="")
nsclc$age_ca = "<50"
nsclc$age_ca[which(nsclc$age >=50)] = "50+"
nsclc$age_ca[which(nsclc$age >=55)] = "55+"
nsclc$age_ca[which(nsclc$age >=60)] = "60+"
nsclc$age_ca[which(nsclc$age >=65)] = "65+"
nsclc$age_ca[which(nsclc$age >=70)] = "70+"
resp.prop.overall = nsclc%>%group_by(age_ca)%>%summarise(prop = mean(survival.past.400),
                                                         n=n())
resp.prop.treatment = nsclc%>%filter(tx==1)%>%group_by(age_ca)%>%
  summarise(prop = mean(survival.past.400),n=n())
resp.prop.control = nsclc%>%filter(tx==0)%>%group_by(age_ca)%>%
  summarise(prop = mean(survival.past.400),n=n())
table_show = resp.prop.treatment[-1,c(1,2)]
colnames(table_show) = c("Age subgroups","Probability of survival past 400 days")
```

In order to have an interval estimate for each of those probabilities, we will use simulation to simulate sample distribution to help us. Our candidate probability is from 0 to 1, and we will simulate sample distribution on the hypothesis if the candidate probability is the true probability that the participants will survive past 400 days in each age subgroup. We will record down the candidate probability that our estimated probability have tailed probabilty between 0.025 to 0.975 as the interval of our estimated probability. We will show the lower bound and the upper bound in the following table.

```
num_sim = 10000
candidate_pi_101 = seq(from = 0, to = 1, by = 0.01)
calc_percentile = function(n,prop){
  percentiles_101 = c()
  for(p in candidate_pi_101){
    sample_counts = rbinom(num_sim, n, p)
    sample_means = sample_counts/n
    percentile = mean(sample_means <= prop)
    percentiles_101 = c(percentiles_101,percentile)
  }
  return(percentiles_101)
}
calc_confint = function(percentiles_101){
  consistent_pi <- candidate_pi_101[(percentiles_101 >= 0.025) & (percentiles_101 <= 0.975)]
  lower_bound <- min(consistent_pi)
```

```
  upper_bound <- max(consistent_pi)
  return(c(lower_bound, upper_bound))
}
set.seed(1)
table_show = as.data.frame(table_show)
table_show$`Lower bound` = NA
table_show$`Upper bound` = NA
for(age in 2:6){
  n =  as.numeric(resp.prop.treatment[age,3])
  prop = as.numeric(resp.prop.treatment[age,2])
  percentiles_101 = calc_percentile(n, prop)
  confint_int = calc_confint(percentiles_101)
  table_show[age-1,3] = confint_int[1]
  table_show[age-1,4] = confint_int[2]
}
colnames(table_show) = c("Age subgroups","Probability of survival past 400 days",
                         "Lower bound","Upper bound")
```

```
kable(table_show,
      caption = "Estimated probability and 95% CI of survival past 400 days for patient on TFD725+")
```

Table 1: Estimated probability and 95% CI of survival past 400 days for patient on TFD725+

Age subgroups	Probability of survival past 400 days	Lower bound	Upper bound
50+	0.778	0.52	0.97
55+	0.472	0.34	0.64
60+	0.464	0.32	0.66
65+	0.722	0.52	0.89
70+	0.800	0.48	0.99

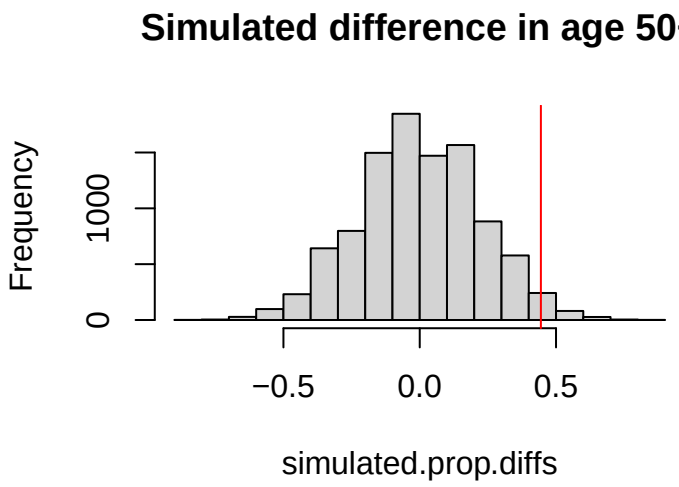
The previous table is showing the estimated probability and its 95 percent confidence interval of patients can survive past 400 days in each subgroup.

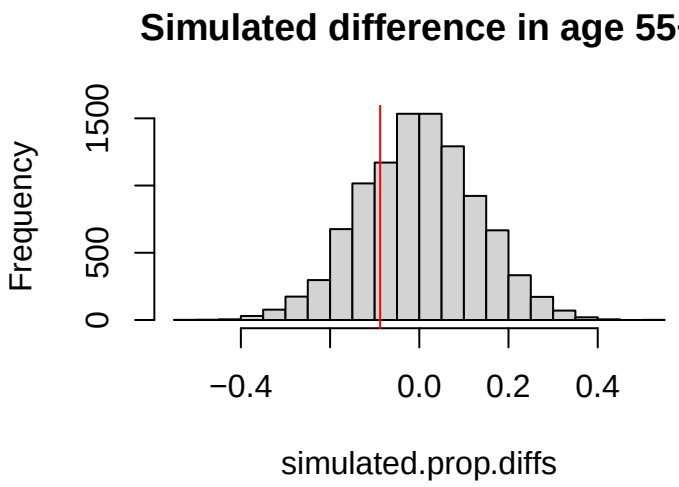
(2)

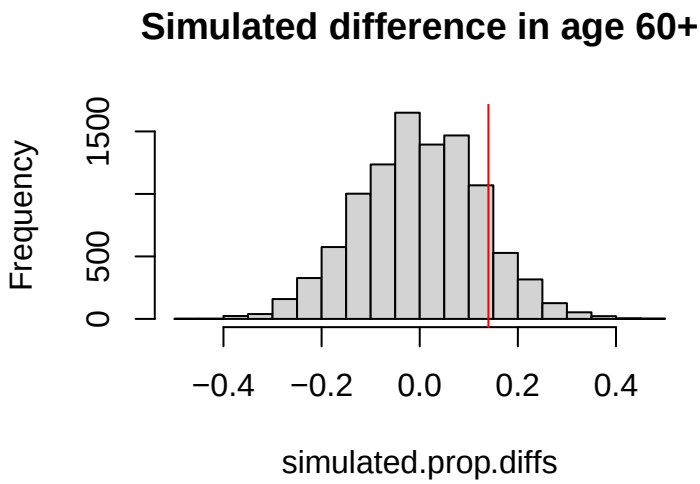
We first try to figure out whether TFD725+ docetaxel is more effective than docetaxel alone in each of those subgroups, we decided to apply the simulation for each subgroup of data to compare the proportions of patients who survive over 400 days in two treatment groups. The simulation is performing under the hypothesis that there is no different betweent the effects of two treatments and the true probability of the participants will survive past 400 days will be the pooled probability of the two treatment groups. If there is no difference between the treatments, our estimated difference from the data should not have a very low tail probability from the simulated sample distribution. The difference will be estimated by simply using the probability in the treatment group minus the probability in the control group.

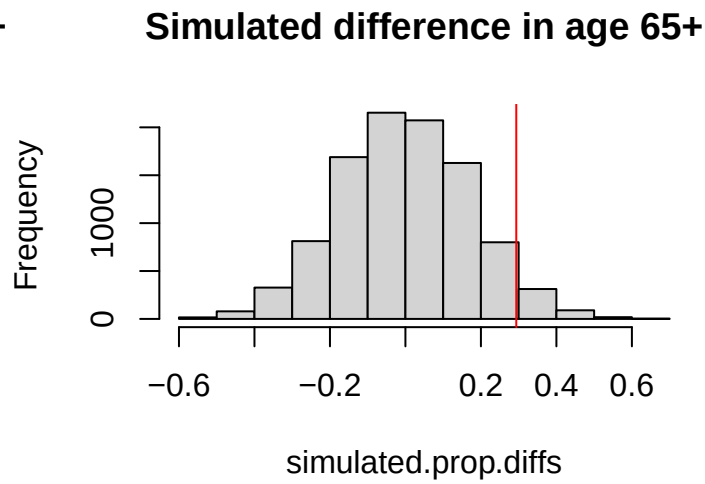
```
simulate.trial <- function(pooled.prob, n.treat, n.control){
    patients.treat <- rbinom(1,n.treat,pooled.prob)
    patients.control <- rbinom(1,n.control,pooled.prob)

    prop.diff <- patients.treat/n.treat - patients.control/n.control

    return(prop.diff)
```

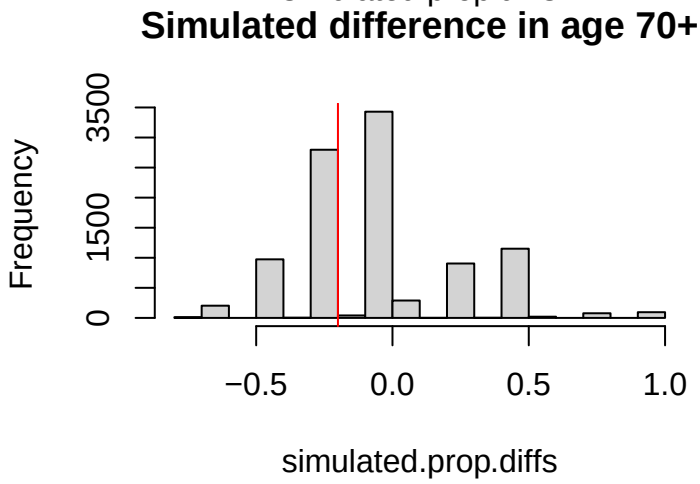
```
}

table_show_2 = matrix(NA, nrow = 5, ncol = 3)
colnames(table_show_2) = c("Age subgroups","Estimated Difference","Tail Probabilities")
table_show_2[,1] = table_show$`Age subgroups`
set.seed(1)

for(i in 1:5){
  pooled.prob = as.numeric(resp.prop.overall[i+1,2])
  n.treat = as.numeric(resp.prop.treatment[i+1,3])
  n.control = as.numeric(resp.prop.control[i,3])
  prop.diff = as.numeric(resp.prop.treatment[i+1,2]) - as.numeric(resp.prop.control[i,2])
  table_show_2[i,2] = round(prop.diff,3)
  simulated.prop.diffs = replicate(num_sim,
                                   simulate.trial(pooled.prob, n.treat,n.control))

  tail.prob = min(mean(simulated.prop.diffs >= prop.diff), mean(simulated.prop.diffs <= prop.diff))
  table_show_2[i,3] = round(tail.prob,3)
  par(mfrow= c(1,1))
  print(hist(simulated.prop.diffs, main =
               paste("Simulated difference in age",table_show_2[i,1])))
  print(abline(v = prop.diff, col = "red"))

}
```












```
kable(table_show_2)
```

Age subgroups	Estimated Difference	Tail Probabilities
50+	0.444	0.027
55+	-0.088	0.257
60+	0.14	0.132
65+	0.294	0.053
70+	-0.2	0.399

As the table and graph shows, the treatment seems to be effective in age 50-54, 60-64, and 65-69 by comparing the proportions of patients who survive over 400 days in two treatment groups. For age grou of 50+ and 60+, the effect of the treatment seems to have smaller tail probabilities. It means it is more unlikely for the treatment have no effect in those subgroups.

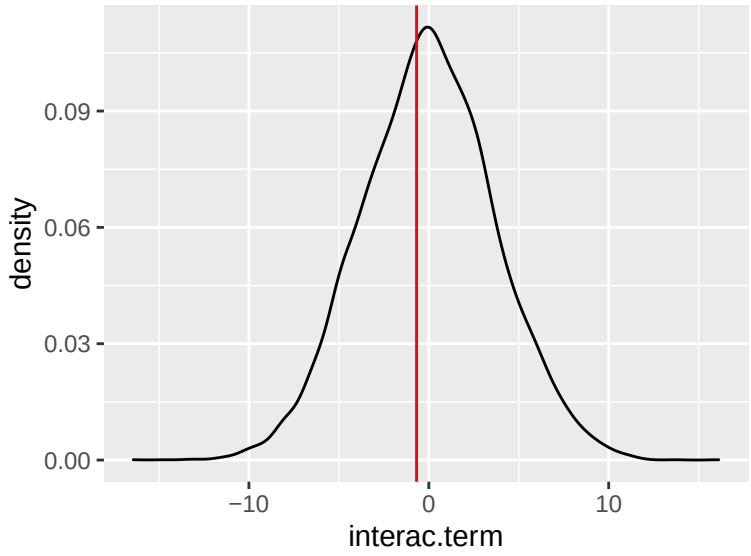
Then we try to evaluate if the treatment effects appears to substatively and/or systematically differ across age. We will use linear regression model with the interaction term to estiamte whether the systematically differ across age.

$$ObservationTime_i = I_{(Treatment=TFD795+)} + age_i + I_{(Treatment=TFD795+)} \times age_i$$

If the effect of TFD795+ is same across all age, then the interaction term should be 0 in the model. We only need to use permutation to help us figure out whether our result from the data is contrasting with the interaction term should equal to 0. Under the null hypothesis that the effect is same across all age, if we

shuffling the age variables, the effect of treatment in each age subgroup will still remail similar. Therefore, we do a simulation of shuffing the age 10,000 times and generate the sample distribution of the estimated coefficients of the interaction term to see whehter the results from our data is extreme in the sampling distribution. If so, it means the null hypothesis is highly possible to be not true.

```
simulate.perm.trial.mo = function(data){
  perm = sample(1:nrow(data), replace = FALSE)
  perm.data = data
  perm.data$age = data$age[perm]
  mod_age = lm(obstime~as.factor(tx)*age, data = perm.data)
  slope_age = coef(mod_age)[4]
  return(slope_age)
}
set.seed(1)
permuted.stats = data.frame(replicate(num_sim, simulate.perm.trial.mo(nsclc)))
diff_age = coef(lm(obstime~as.factor(tx)*age, data = nsclc))[4]
p_value_slope = min(mean(permuted.stats <= diff_age), mean(permuted.stats>=diff_age))
permuted.stats = as.data.frame(permuted.stats)
colnames(permuted.stats) = "interac.term"
ggplot(permuted.stats, aes(x=interac.term, y=..density..)) +
    geom_density()+
    geom_vline(xintercept=diff_age, colour = "red")
```



Based on our simulation, the tailed probability is 0.421, which is relatively large. We think it is very possible that treatment is not substantively/or systematically differ across age.