BIOST 544: Homework 1

Department of Biostatistics @ University of Washington

Alejandro Hernandez

October 10, 2024

Background

There is a belief that the effectiveness of the anti-angiogenesis agent TFD725 (evaluated in the **nsclc** dataset) may be different for older vs younger patients. We will use the nsclc dataset to attempt to evaluate this.

Analysis

1) As a first pass, we will consider a few subgroups of patients: Those 50 and older (50+), 55+, 60+, 65+, and 70+. Please estimate/evaluate the probability a patient on TFD725+docetaxel will survive past 400 days in each of those subgroups. Please also give an interval estimate for each of those probabilities.

For a patient in a subpopulation of the TFD725+docetaxel treatment arm, the probability of surviving past 400 days is estimated by the proportion of survivors in that subpopulation. Assuming a Binomial distribution, we can simulate outcomes using the sample size and observed proportion, then define a 90% confidence interval based on these simulations.

Table 1 shows the observed 400-day survival proportions and their confidence intervals for each age subgroup (note: subgroups are overlapping and share members).

Table 1: Estimated probability of 400-day survival within overlapping subsets of the treatment arm

Age	N	Estimate	90% CI
50 +	186	0.49	0.44 - 0.55
55 +	165	0.49	0.43 - 0.55
60 +	104	0.48	0.4 - 0.56
65 +	39	0.64	0.51 - 0.77
70 +	7	0.86	0.57 - 1

Table 1 and histograms of the simulated sampling distribution of P(Survived|Age) (see Supplementary) suggest that treated patients under 65 have similar survival chances, while those over 65 show higher survival rates, though older, smaller samples have wider confidence intervals.

To further explore age effects, we divided the cohort into non-overlapping subgroups, with 400-day survival proportions and confidence intervals shown below.

Table 2: Estimated probability of 400-day survival within subsets of the treatment arm

Age	N	Estimate	90% CI
[50,55)	9	0.78	0.56 - 1
[55,60)	36	0.47	0.33 - 0.61
[60,65)	28	0.46	0.32 - 0.61
[65,70)	18	0.72	0.56 - 0.89
[70,75]	5	0.80	0.4 - 1

Table 2 and histograms (see Supplementary) suggest treated patients aged 55-64 have similar survival chances, while those under 55 or over 65 show higher survival rates, though these subgroups have fewer members and greater uncertainty.

2) Now, in each of those subgroups evaluate whether TFD725+docetaxel is more effective than docetaxel alone (and the magnitude of any potential treatment effect). In addition, evaluate if the treatment effect appears to substantively and/or systematically differ across age (or if the data doesn't give a clear answer to this).

To test if TFD725+docetaxel is more effective than docetaxel alone, we can use a permutation test by repeatedly reassigning treatments to simulate a scenario where treatment has no effect (such that assignment is inconsequential). This generates a sampling distribution of survival differences, and comparison to the 95^{th} percentile assesses if our observed difference is unusually high given our assumption of no treatment effect.

Table 3 and histograms (see Supplementary) suggest that TFD725+docetaxel is more effective than docetaxel alone among the collection of patients aged 50+ years and 60+ years. However, across all non-overlapping age groups, there is insufficient evidence that TFD725+docetaxel is more effective than docetaxel alone.

Evaluating our histograms leads us to conclude that treatment effect does not appear to substantively and/or systematically differ across age. Our simulations suggest TFD725 treatment is slightly more effective, although in most age groups, not significantly more effective.

Table 3: Observed survival differences associated with treatment, stratified by age group

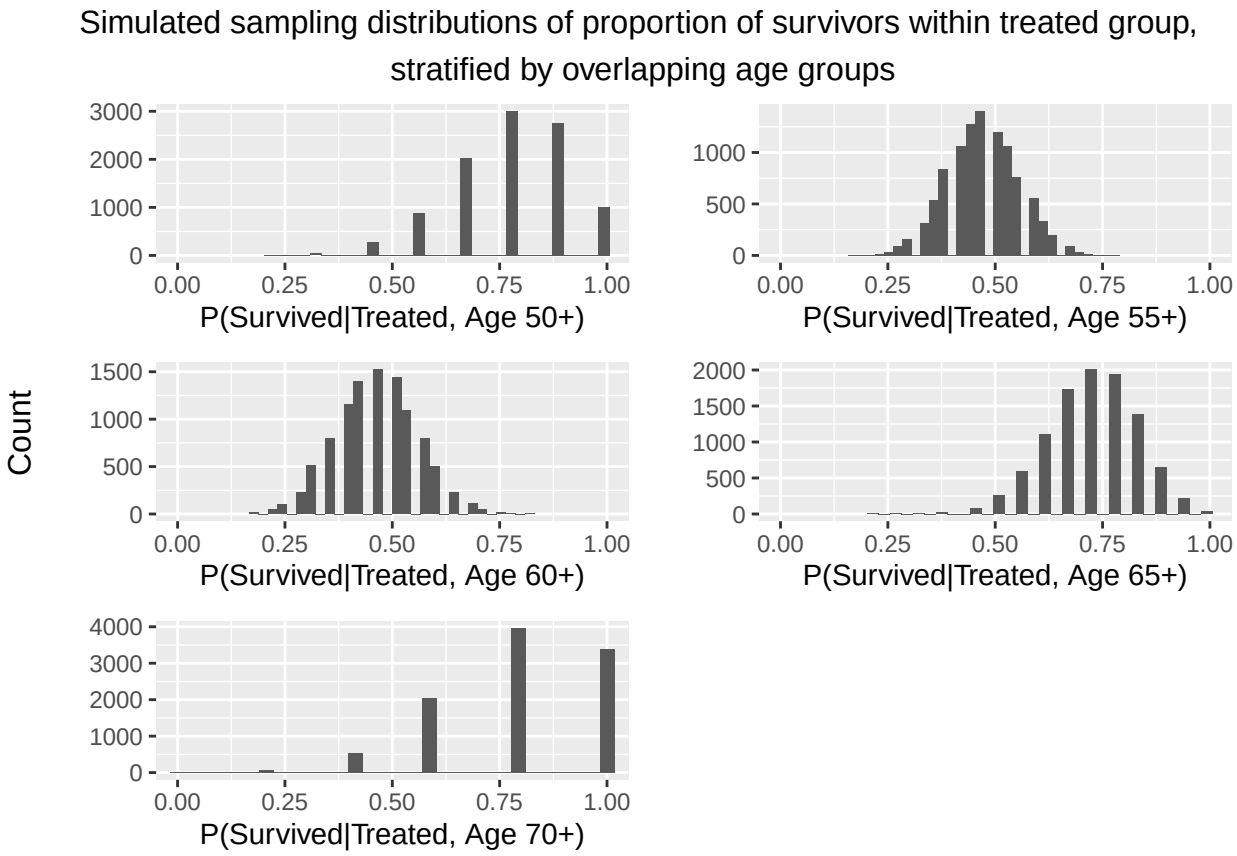
Age	Estimate	P.value	Age	Estimate	P.value
50+	0.14	0.04	50-54	0.44	0.05
55+	0.10	0.13	55-59	-0.09	0.83
60+	0.21	0.02	60-64	0.14	0.18
65+	0.24	0.12	65-69	0.29	0.10
70+	-0.20	1.00	70-74	-0.20	1.00

End of analysis. Supplementary begins on the next page.

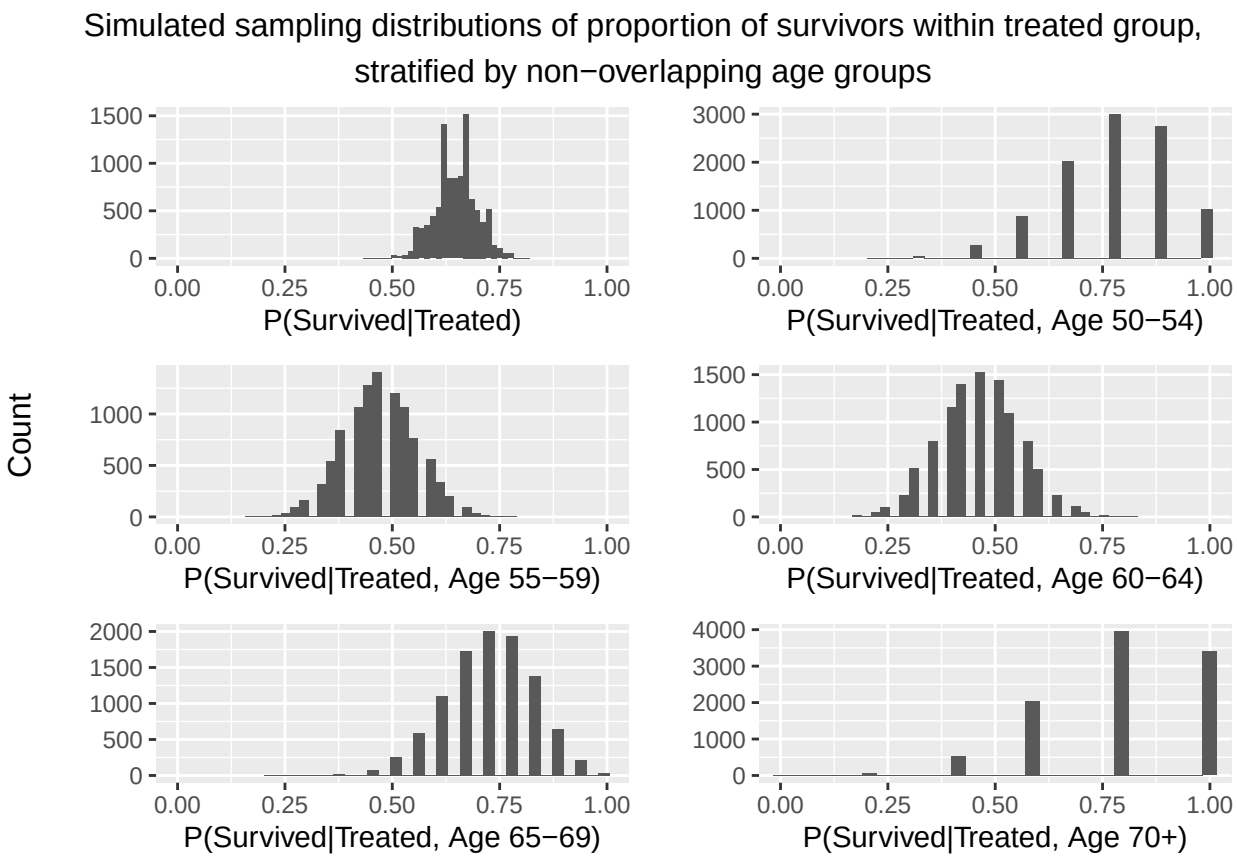
Supplementary

Question 1

Histograms of the simulated sampling distribution of P(Survived|Age), across overlapping age groups.

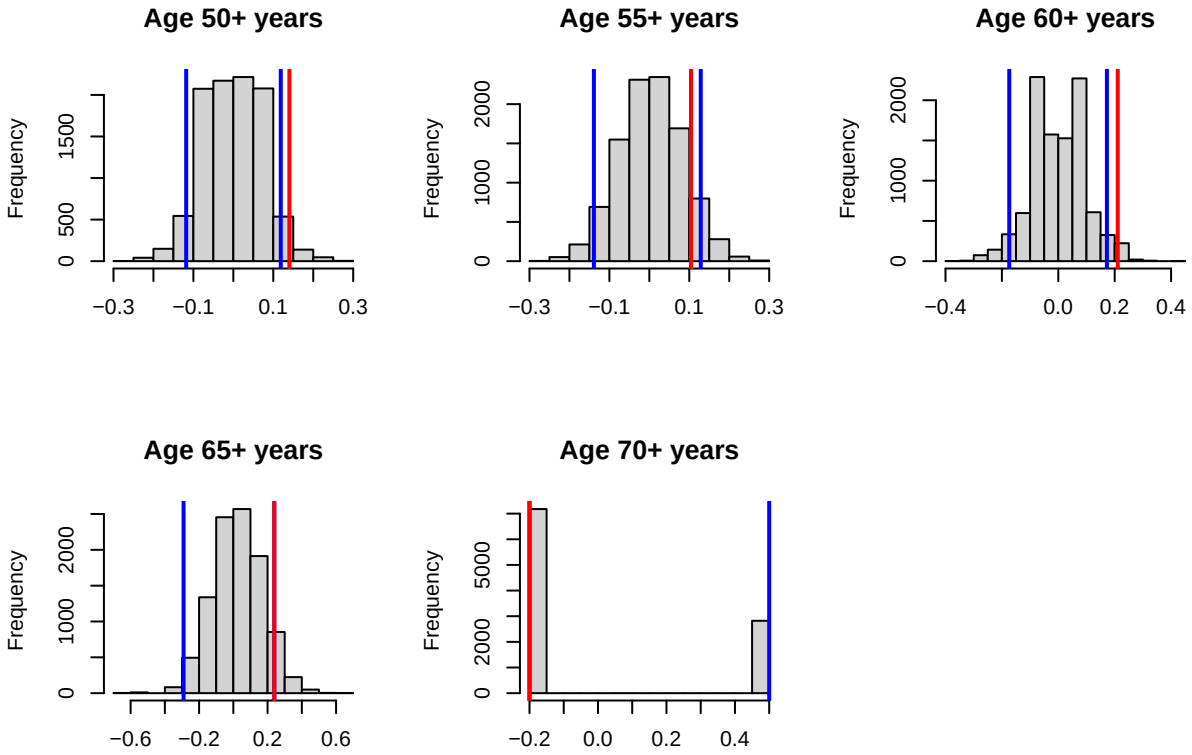


Histograms of the simulated sampling distribution of P(Survived|Age), across non-overlapping age groups.

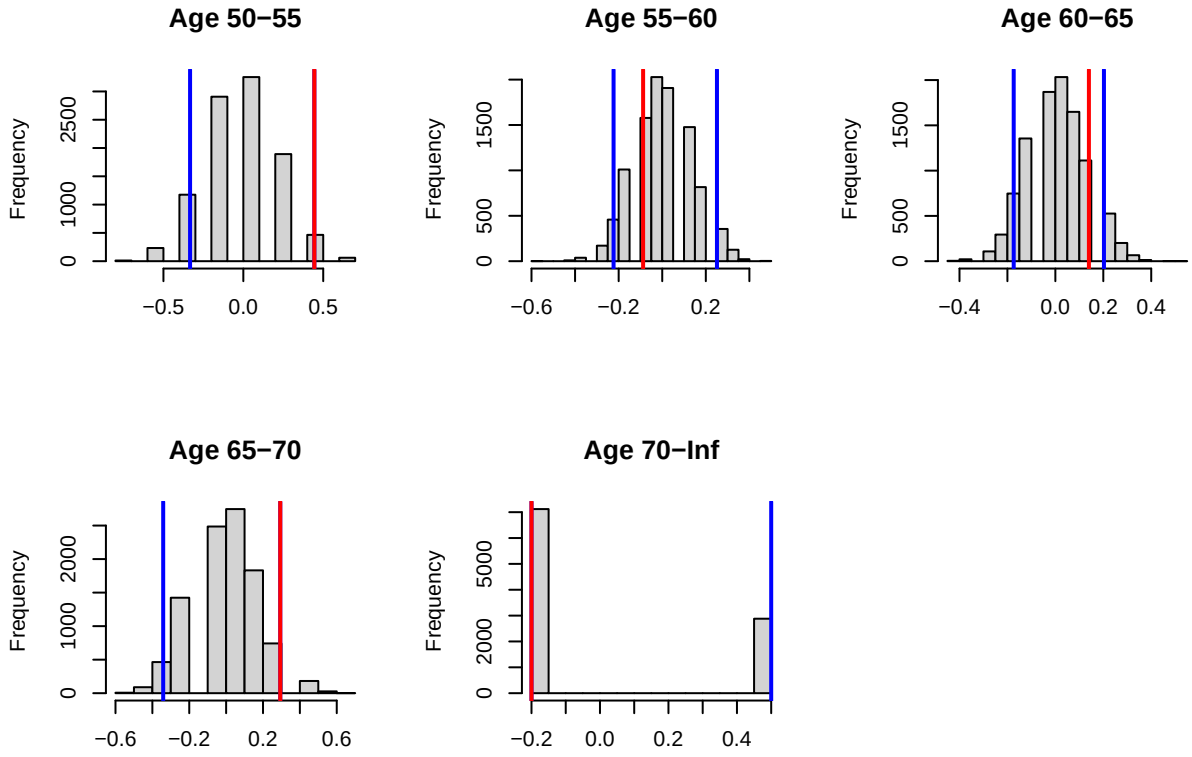


Question 2

Histograms of the simulated sampling distribution of P(Survived|TFD725+docetaxel) - P(Survived|docetaxel), across overlapping age groups. Blue lines denote 5% and 95% percentiles and define an inner 90% confidence interval under the hypothesis that TFD725 has no treatment effect; red line denotes our observed difference in survival associated with treatment.



Histograms of the simulated sampling distribution of P(Survived|TFD725+docetaxel) - P(Survived|docetaxel), across non-overlapping age groups. Blue lines denote 5% and 95% percentiles and define an inner 90% confidence interval under the hypothesis that TFD725 has no treatment effect; red line denotes our observed difference in survival associated with treatment.



End of report. Code appendix begins on the next page.

Code Appendix

```
# setup options
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
options(knitr.kable.NA = '-')
options(digits = 3)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
# clear environment
rm(list=ls())

# load relevant packages
library(dplyr)      # data frame manipulation
library(knitr)      # table formatting
library(ggplot2)    # plotting
library(gridExtra)  # assembling multiply plots on the same page

# load data
nsclc <- read.table("../data/nsclc-modified.txt")

## ===============
## Question 1
## ===============
simulate_sample_dist <- function(n, p){
  nsamp <- 10000
  sample_means <- rbinom(nsamp, n, p) / n
  return(list(
    simdata = sample_means,
    estimate = p,
    ci = quantile(sample_means, c(0.05, 0.95))))
}

## ============================
## Overlapping age groups
## ============================
# calculate and store proportion of success within treatment group
means.treated <- data.frame()
age_cutoffs <- seq(50, 70, 5)
for (c in age_cutoffs) {
  subset <- subset(nsclc, age >= c, select = survival.past.400)
  new_row <- data.frame(age = paste(c,"+"),
                        n = nrow(subset),
                        mean = mean(subset$survival.past.400))
  means.treated <- rbind(means.treated, new_row)
}

# simulate sampling distributions for each subpopulation of the treated group
res50 <- simulate_sample_dist(means.treated$n[1], means.treated$mean[1])
res55 <- simulate_sample_dist(means.treated$n[2], means.treated$mean[2])
res60 <- simulate_sample_dist(means.treated$n[3], means.treated$mean[3])
res65 <- simulate_sample_dist(means.treated$n[4], means.treated$mean[4])
res70 <- simulate_sample_dist(means.treated$n[5], means.treated$mean[5])
```

```r
# add 5% and 95% simulated percentiles to means table
means.treated$perc5 <-
  c(res50$ci[1], res55$ci[1], res60$ci[1], res65$ci[1], res70$ci[1])
means.treated$perc95 <-
  c(res50$ci[2], res55$ci[2], res60$ci[2], res65$ci[2], res70$ci[2])
# inspect estimates and 90% CIs
means.treated %>%
  mutate(CI = paste(round(perc5,2), "-", round(perc95,2))) %>%
  select(-c(perc5, perc95)) %>%
  knitr::kable(digits = 2,
               col.names = c("Age", "N", "Estimate", "90% CI"),
               caption = "Estimated probability of 400-day survival within overlapping subsets of the ti

# plot each simulated sampling distribution
gg50.2 <- ggplot(mapping=aes(res50$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 50+)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg55.2 <- ggplot(mapping=aes(res55$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 55+)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg60.2 <- ggplot(mapping=aes(res60$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 60+)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg65.2 <- ggplot(mapping=aes(res65$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 65+)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg70.2 <- ggplot(mapping=aes(res70$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 70+)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

## ============================
## Non-overlapping age groups
## ============================
# create a factor variable from patient age
nsclc$age2 <- cut(nsclc$age,
                  seq(50,75,5),
                  include.lowest = TRUE, right = FALSE)
                  # labels = c("50+","55+","60+","65+","70+")
# inspect subgroup sizes
# table(nsclc$age2)
# remove the few observations whose age was below 50 (their age2 is NA)
nsclc <- subset(nsclc, !is.na(age2))

# calculate and store proportion of success within treatment group
means.treated <- nsclc %>%
  filter(tx == 1) %>%
  group_by(age = age2) %>%
  summarize(n = n(), mean = mean(survival.past.400))
```

```r
# simulate sampling distributions for each subpopulation of the treated group
res <- simulate_sample_dist(sum(means.treated$n), mean(means.treated$mean))
res50 <- simulate_sample_dist(means.treated$n[1], means.treated$mean[1])
res55 <- simulate_sample_dist(means.treated$n[2], means.treated$mean[2])
res60 <- simulate_sample_dist(means.treated$n[3], means.treated$mean[3])
res65 <- simulate_sample_dist(means.treated$n[4], means.treated$mean[4])
res70 <- simulate_sample_dist(means.treated$n[5], means.treated$mean[5])

# add 5% and 95% simulated percentiles to means table
means.treated$perc5 <-
  c(res50$ci[1], res55$ci[1], res60$ci[1], res65$ci[1], res70$ci[1])
means.treated$perc95 <-
  c(res50$ci[2], res55$ci[2], res60$ci[2], res65$ci[2], res70$ci[2])
# inspect estimates and 95% CIs
means.treated %>%
  mutate(CI = paste(round(perc5,2), "-", round(perc95,2))) %>%
  select(-c(perc5, perc95)) %>%
  knitr::kable(digits = 2,
               col.names = c("Age", "N", "Estimate", "90% CI"),
               caption = "Estimated probability of 400-day survival within subsets of the treatment arm

# plot each simulated sampling distribution
gg <- ggplot(mapping=aes(res$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated)") + ylab("") + coord_cartesian(xlim = c(0,1))

gg50 <- ggplot(mapping=aes(res50$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 50-54)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg55 <- ggplot(mapping=aes(res55$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 55-59)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg60 <- ggplot(mapping=aes(res60$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 60-64)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg65 <- ggplot(mapping=aes(res65$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 65-69)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

gg70 <- ggplot(mapping=aes(res70$simdata)) + geom_histogram() +
  xlab("P(Survived|Treated, Age 70+)") + ylab("") +
  coord_cartesian(xlim = c(0,1))

## ===============
## Question 2
## ===============

calc_stat <- function (data) {
  diff.means <- with(data,
    mean(survival.past.400[tx==1]) - mean(survival.past.400[tx==0]))
  return(diff.means)
```

```
}

simulate_perm_trial <- function(data){
  # permute exposure status
  perm.data = data
  perm.data$tx = perm.data$tx[sample(1:nrow(perm.data), replace=FALSE)]

  ## return the difference in means from permuted data
  return(calc_stat(perm.data))
}

## ============================
## Overlapping age groups
## ============================

nsims <- 10000
age_cutoffs <- seq(50, 70, 5)
obsv.stats <- c()
pvals <- c()
for (c in age_cutoffs) {
  nsclc_subset <- subset(nsclc, c <= age)
  simdata <- replicate(nsims, simulate_perm_trial(nsclc_subset))
  obsv.stat <- calc_stat(nsclc_subset)
  obsv.stats <- c(obsv.stats, obsv.stat)
  # p-value
  pvals <- c(pvals, mean(simdata >= obsv.stat))
}

res1 <- data.frame(Age = paste(age_cutoffs, "+", sep=""),
                   Estimate = obsv.stats,
                   P.value = pvals)

## ============================
## Non-overlapping age groups
## ============================

nsims <- 10000
age_cutoffs <- c(seq(50, 70, 5), Inf)
obsv.stats <- c()
pvals <- c()
for (i in 1:5) {
  nsclc_subset <- subset(nsclc, age_cutoffs[i] <= age & age < age_cutoffs[i+1])
  simdata <- replicate(nsims, simulate_perm_trial(nsclc_subset))
  obsv.stat <- calc_stat(nsclc_subset)
  obsv.stats <- c(obsv.stats, obsv.stat)
  # p-value
  pvals <- c(pvals, mean(simdata >= obsv.stat))
}

res2 <- data.frame(Age = paste(age_cutoffs[-6], "-", age_cutoffs[-6]+4, sep=""),
                   Estimate = obsv.stats,
                   P.value = pvals)
```

```r
cbind(res1, res2) %>%
  kable(digits = 2,
        caption = "Observed survival differences associated with treatment, stratified by age group")
# plot each simulated sampling distribution
gridExtra::grid.arrange(gg50.2, gg55.2, gg60.2, gg65.2, gg70.2,
                        top = "Simulated sampling distributions of proportion of survivors within treat
                        left = "Count")
# plot each simulated sampling distribution
gridExtra::grid.arrange(gg, gg50, gg55, gg60, gg65, gg70,
                        top = "Simulated sampling distributions of proportion of survivors within treat
                        left = "Count")
nsims <- 10000
par(mfrow = c(2, 3))
age_cutoffs <- seq(50, 70, 5)
for (c in age_cutoffs) {
  nsclc_subset <- subset(nsclc, c <= age)
  simdata <- replicate(nsims, simulate_perm_trial(nsclc_subset))
  obsv.stat <- calc_stat(nsclc_subset)

  # simulated sampling distribution
  hist(simdata, main = paste("Age ", c, "+ years",  sep=""), xlab = "")
  # inner 95% percentiles
  abline(v = quantile(simdata, probs=c(0.05, 0.95), names = FALSE),
         col = "blue", lwd = 2)
  # observed value
  abline(v = obsv.stat, col = "red", lwd = 2)
}
nsims <- 10000
par(mfrow = c(2, 3))
age_cutoffs <- c(seq(50, 70, 5), Inf)
for (i in 1:5) {
  nsclc_subset <- subset(nsclc, age_cutoffs[i] <= age & age < age_cutoffs[i+1])
  simdata <- replicate(nsims, simulate_perm_trial(nsclc_subset))
  obsv.stat <- calc_stat(nsclc_subset)

  # simulated sampling distribution
  hist(simdata, main = paste("Age ", age_cutoffs[i], "-", age_cutoffs[i+1],  sep=""),
       xlab = "")
  # inner 95% percentiles
  abline(v = quantile(simdata, probs=c(0.05, 0.95), names = FALSE),
         col = "blue", lwd = 2)
  # observed value
  abline(v = calc_stat(nsclc_subset), col = "red", lwd = 2)
}
```

End of document.