BIOST 544 HW1

John Schoof

10/20/2020

Part 1: The probability a patient on TFD725+docetaxel will survive past 400 days by age

The following analysis describes and evaluates the effects of the treatment, TFD725, on the probability of surviving more than 400 days across age groups in a Phase II clinical trial of 188 patients with non-small cell lung cancer.

Methods: First, I remove the two patients younger than age 50 and then categorize patients into age groups by every five years (i.e. 50-54, 55-59, etc.). Then to determine point estimate for P(surviving > 400days), I divide the number of treatment patients who survived past 400 days by the total number of treatment patients within each age group. To calculate the 95% confidence intervals for each point estimate I use random binomial simulation.

I repeated the following process on each age group. I selected 101 candidates of the true population parameter for P(surviving > 400days), pi. This creates a vector of pi values (0, 0.01, 0.02...0.98, 0.99, 1.00). For each value in this vector, 10,000 random binomial simulations are run. The number of draws in each simulation is equal to the number of patients in the treatment group within that particular age group. We run 101 random binomial simulations, one with each candidate pi as the probability of success. We now have 101 sampling distributions. I determine at which percentile of each sampling distribution our observed pi falls. I create a vector of all 101 of those percentile values and then to get the confidence interval, I identify the values of pi hat are the percentiles greater than or equal to 0.025 and less than or equal to 0.975. We then find the min and max numbers of this vector and we have a 95% confidence interval.

Results: In Table 1, we see that the age group with the highest P(surviving > 400 days) is the 70+ group with four of five treatment patients surviving past 400 days. It is important to note that this age group includes the fewest number of observations of treated patients. The age group with the second highest P(surviving > 400 days) is the youngest group, 50-54, followed by the 65-69 age group. The 55-59 and 60-64 groups have the lowest probabilities with 0.472 and 0.464, respectively.

Table 1: Table 1

Age Group	N	Number Treated	P(Survival > 400 days)	95% CI
50-54	21	9	0.778	(0.53-0.97)
55-59	61	36	0.472	(0.31-0.61)
60-64	65	28	0.464	(0.28-0.62)
65-69	32	18	0.722	(0.47-0.86)
70 or older	7	5	0.800	(0.48-0.99)

Part 2: Is TFD725+docetaxel more effective than docetaxel alone after stratifying by age?

In this part of the analysis I identify and measure potential treatment effects within each age group. The analysis begins with the null hypothesis that the difference in proportion that survive past 400 days is exactly the same in both treatment and control groups, $H_o : \text{diff}_T = \text{diff}_C$.

Methods: We use random binomial simulation estimate the difference in P(surviving > 400days) within each age group. We run 10,000 random binomial simulations to estimate the difference between treatment and control groups using the overall probability in the age group as the probability of success. We use this pooled value in order to estimate what the difference would be in the absence of a treatment and so that we can see the entire distribution. We would expect the distribution of these 10,000 simulations to center around zero.

I find the difference between the observed difference and the mean of the simulated differences to estimate the difference in differences. This is the magnitude of the treatment effect.

Results: Figures 1-5 show the sampling distributions of the difference in P(surviving > 400days) within each age group. The red vertical line is the observed difference for that age group from this sample. In all age groups except for the 55-59 group, you can see that the red line lies in the far right tail of the distribution. This suggests that if this was the distribution of the true population parameter for the difference in P(surviving > 400days) under the null hypothesis, then it is unlikely that we would see our observed value.

We confirm this by finding the percentage of observations that fall to the right of our red line. This is the p-value of our hypothesis test at the $\alpha = 0.05$ significance level, we have evidence to reject the null hypothesis for all groups except for the 55-59 year olds. In other words, **the patients in the treatment group are more likely to survive more than 400 days than the patients in the control group in every age group except for the 55-59 group.**

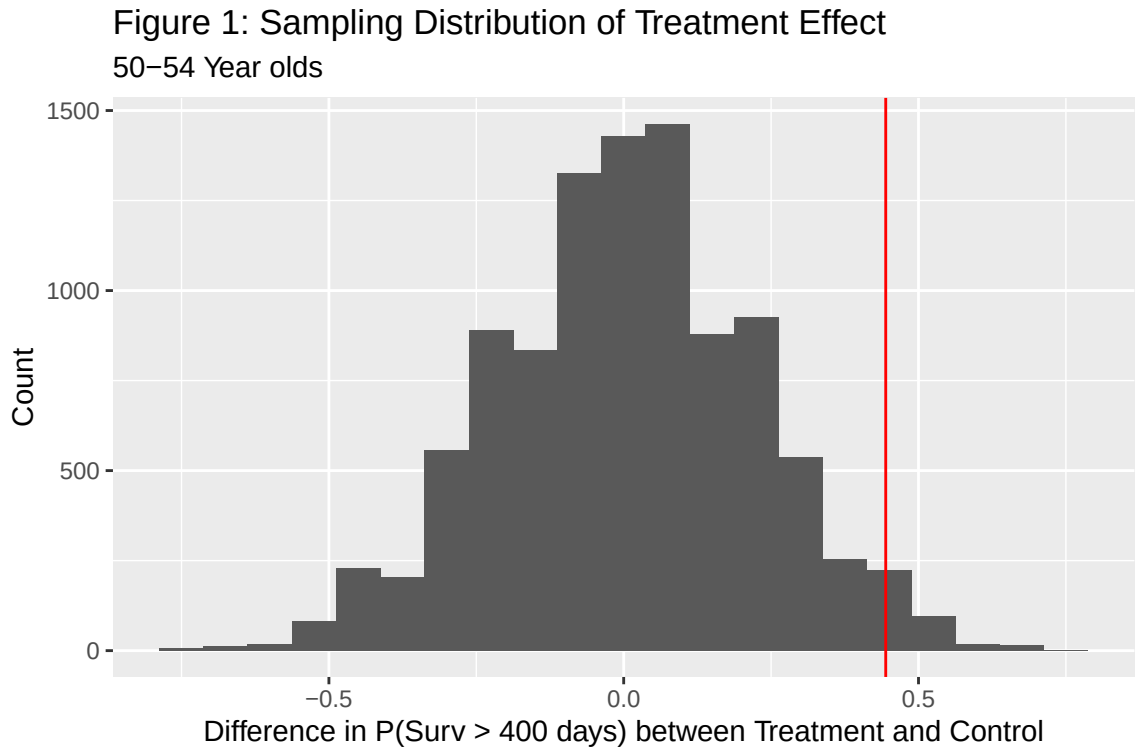

Figure 2: Sampling Distribution of Treatment Effect

55−59 Year olds

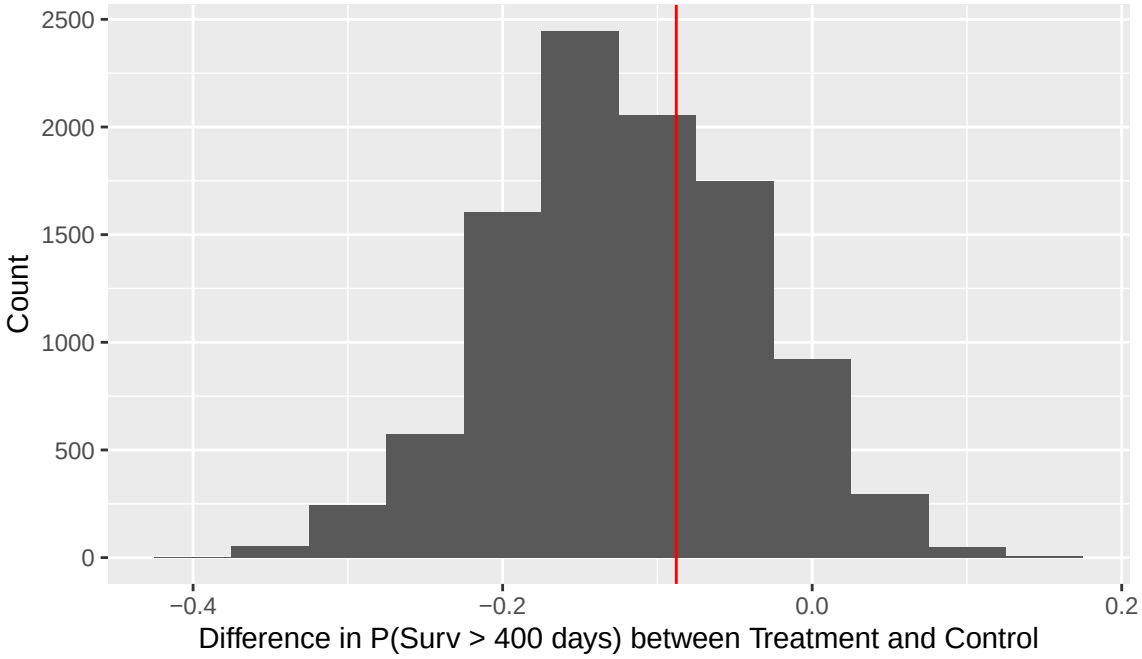

Figure 3: Sampling Distribution of Treatment Effect

60−64 Year olds

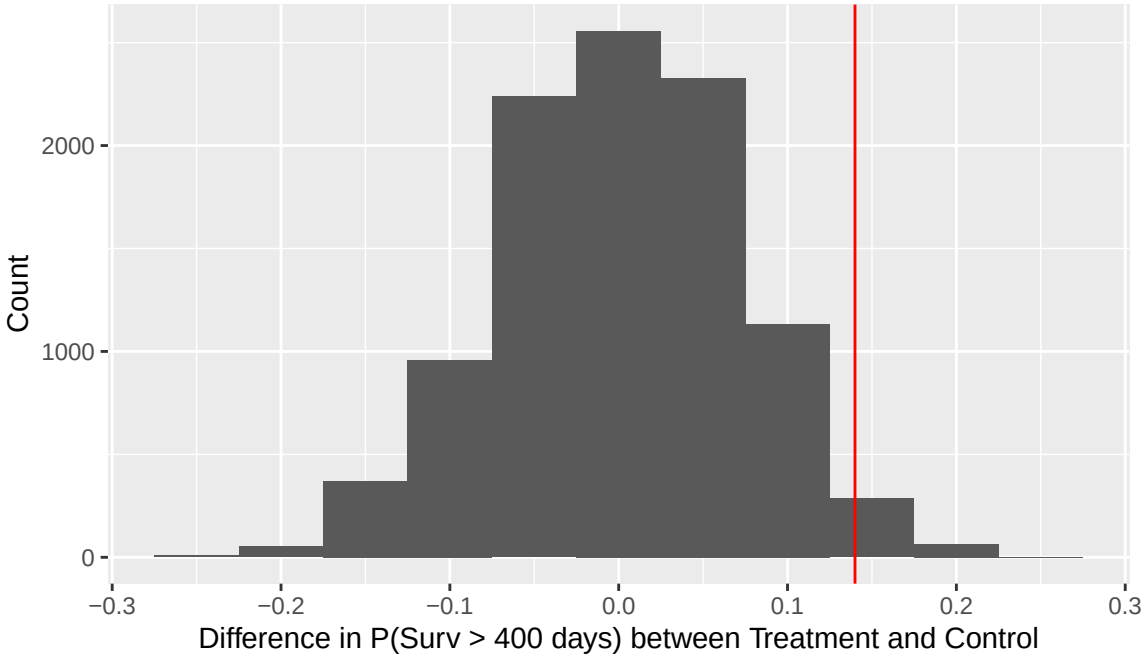

Figure 4: Sampling Distribution of Treatment Effect

65−69 Year olds

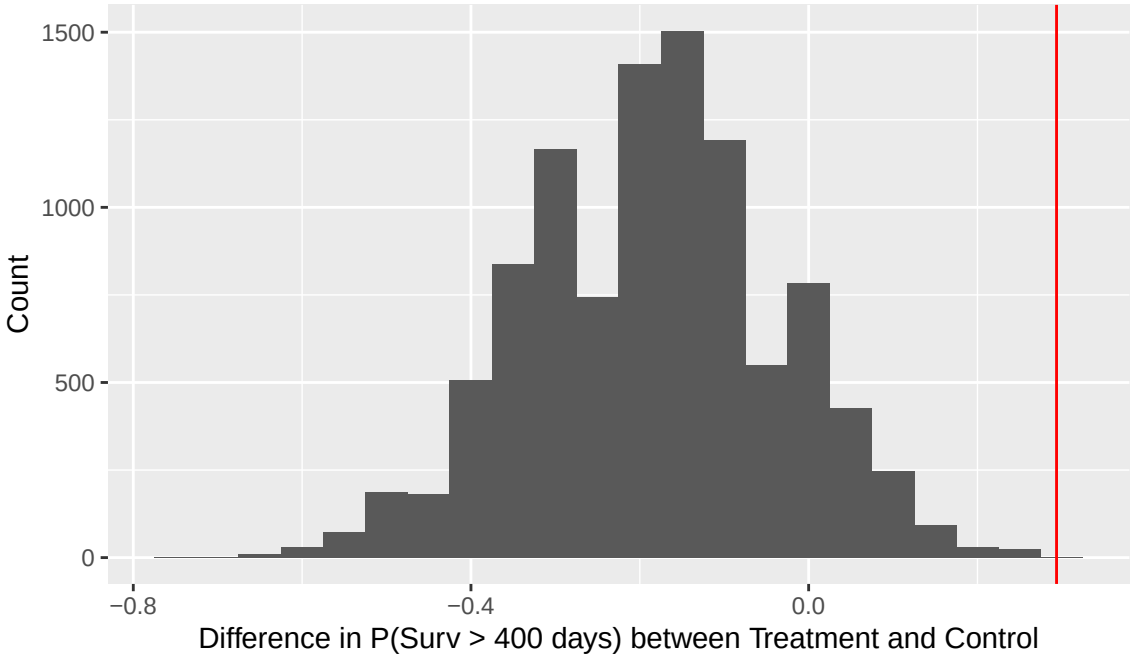

Figure 5: Sampling Distribution of Treatment Effect

70+ Year olds

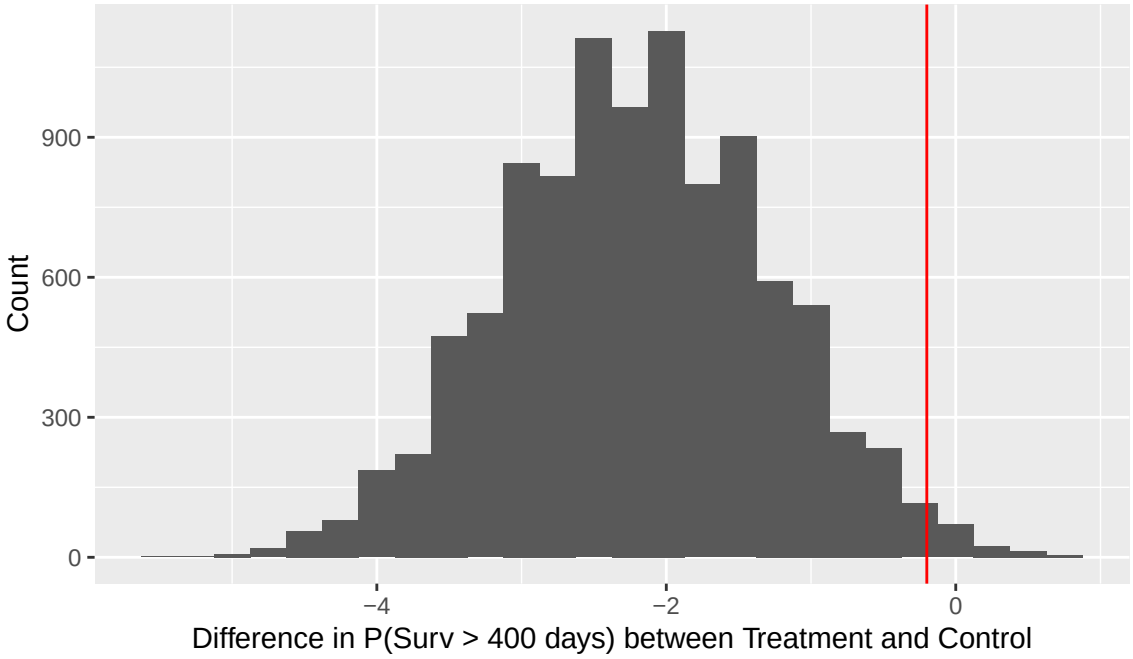

Table 2: Table 2

Age Group	Observed Difference	Simulated Mean Difference	Difference in Differences
50-54	0.444	0.003	0.441

Age Group	Observed Difference	Simulated Mean Difference	Difference in Differences
55-59	-0.088	-0.120	0.032
60-64	0.140	0.000	0.140
65-69	0.294	-0.186	0.479
70+	-0.200	-2.199	1.999

Part 3: Does the treatment effect appear to be substantively different across age?

Table 2 shows the observed difference in P(surviving > 400days) between treatment and control as well as the simulated mean difference in P(surviving > 400days in the absence of the treatment. Lastly, Table 2 also shows the difference between those two differences. This is the increase in the P(surviving > 400days) for those who receive the treatment. The treatment effect appears to largest in the 70+ group, but this group only includes five observations and our observed difference still suggests a small negative treatment effect. The 65-69 and 50-54 group have the next largest treatment effect. In these two age groups, patients who receive the treatment increase their probability of surviving 400 days by 48 and 45 percentage points, respectively.

Appendix: R Code

```
# Set up
knitr::opts_chunk$set(echo=TRUE, fig.width=6, fig.height=4)
knitr::opts_knit$set(root.dir = ("/Users/johnschoof/Documents/Analyses/Data Science"))

rm(list=ls())
library(tidyverse)
library(knitr)
options(digits = 3)
set.seed(1)


# Load data
nsclc <- read.table("datasets/nsclc-modified.txt", header = TRUE)
names(nsclc) ## Column Names
nsclc %>% summarize(n()) ## 188 observations
nsclc %>% glimpse() ## 14 variables
nsclc %>% head()


### QUESTION 1 #######
######################

# Create age group variable
nsclc$group.age <- cut(nsclc$age, c(50, 55, 60, 65, 70, Inf), right=F)

nsclc$group.age <- factor(nsclc$group.age,
                          labels = c("50-54", "55-59", "60-64",
                                     "65-69", "70 or older"))
nsclc <- nsclc %>% na.omit()
table(nsclc$group.age)
nsclc %>% group_by(tx, group.age) %>% summarise(n())
```

```r
# Estimate the probability of surviving past 400 days for treatment group in each age group in our samp
  (surv.prop <- nsclc %>%
                  filter(tx==1) %>%
                  group_by(group.age) %>%
                  summarise(prop = mean(survival.past.400)))

# Number treated in each age group in our sample
  (num.trt <- nsclc %>%
              filter(tx==1) %>%
              group_by(group.age) %>%
              summarise(num = n()))

# Function for probability of surviving over 400 days from 10,000 simulations
  calc_sample_dist <- function(prop, num){
                          nsamp <- 10000
                          sample_counts <- rbinom(nsamp, num, prop)
                          sample_prop <- sample_counts/num
                          return(sample_prop)
  }

  head(calc_sample_dist(0.778, 9))

# Loop to create sampling distribution for each of 101 candidate pi
  prop.tmp <- c(0.778, 0.472, 0.464, 0.722, 0.800)
  num.tmp <- c(9, 36, 28, 18, 5)

  candidate_pi_101 <- seq(from = 0, to = 1, length.out = 101)

  percentile.list <- vector("list", length(prop.tmp))

  for(i in 1:length(prop.tmp)){

    percentiles_101 <- c()

  for(pi in candidate_pi_101){
    samp_dist <- calc_sample_dist(pi, num.tmp[i])
    percentile <- mean(samp_dist <= prop.tmp[i])
    percentiles_101 <- c(percentiles_101,percentile)
  }
    percentile.list[i] <- percentiles_101

  }

  plot(candidate_pi_101, percentiles_101, ylim = c(0, 1))
  abline(h = 0.025, col = "red")
  abline(h = 0.975, col = "red")

# find confidence interval
  (consistent_pi <- candidate_pi_101[(percentiles_101 >= 0.025) & (percentiles_101 <= 0.975)])
  (lower_bound <- min(consistent_pi))
  (upper_bound <- max(consistent_pi))
```

```r
# Create table to show estimates and conf intervals
  table1 <- nsclc %>%
                    filter(tx==1) %>%
                    group_by(group.age) %>%
                    summarise(num = n(), prop = mean(survival.past.400))

  table1$ntotal <- nsclc %>%
                    group_by(group.age) %>%
                    summarise(ntotal = n()) %>%
                    .$ntotal

  table1$conf.int <- c("(0.53-0.97)","(0.31-0.61)", "(0.28-0.62)", "(0.47-0.86)", "(0.48-0.99)")

  table1 <- table1[ , c(1, 4, 2, 3, 5)]

  kable(table1, col.names = c("Age Group", "N", "Number Treated","P(Survival > 400 days)", "95% CI"),
                     caption = "Table 1", "pipe")


### QUESTION 2 #######
######################

# find proportion of survival among treat and control groups
  (resp.prop.overall <- nsclc %>%
                        group_by(group.age) %>%
                        summarise(prop = mean(survival.past.400)) %>%
                        .$prop)

  (resp.prop.treat <- nsclc %>%
                      filter(tx == 1) %>%
                      group_by(group.age) %>%
                      summarise(prop = mean(survival.past.400)) %>%
                      .$prop)

  (resp.prop.control <- nsclc %>%
                        filter(tx == 0) %>%
                        group_by(group.age) %>%
                        summarise(prop = mean(survival.past.400)) %>%
                        .$prop)

  (resp.prop.diff <- resp.prop.treat - resp.prop.control)

  (num.treat  <- nsclc %>%
                  filter(tx == 1) %>%
                  group_by(group.age) %>%
                  summarise(number = n()) %>%
                  .$number)

  (num.control  <- nsclc %>%
                  filter(tx == 0) %>%
                  group_by(group.age) %>%
                  summarise(number = n())%>%
                  .$number)
```

```r
# Function to run one trial
  simulate.trial <- function(pi.treat, pi.control, n.treat, n.control){
                      patients.treat <- rbinom(1, n.treat, pi.treat)
                      patients.control <- rbinom(1, n.control, pi.control)

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

                      return(prop.diff)
  }
  #test on 50-54
  #simulate.trial(0.778, .333, 9, 12)

# Replicate to run many trials
  ntrial <- 10000
  simulated.prop.diffs <- data.frame(t(replicate(ntrial,
                                    simulate.trial(resp.prop.overall,
                                                   resp.prop.overall,
                                                   num.treat,
                                                   num.control))))
  colnames(simulated.prop.diffs) <- c("diff.5054","diff.5559", "diff.6064","diff.6569","diff.70")
  head(simulated.prop.diffs)
  str(simulated.prop.diffs)

  # test on 50-54
  #simulated.prop.diffs <- replicate(ntrial,
  #                                  simulate.trial(0.778, .333, 9, 12))

# Histograms
ggplot(data = simulated.prop.diffs, mapping = aes(x=diff.5054)) +
    geom_histogram(binwidth=0.075) +
    geom_vline(aes(xintercept = resp.prop.diff[1]), color = "red") +
    labs(x = "Difference in P(Surv > 400 days) between Treatment and Control",
         y = "Count",
         title = "Figure 1: Sampling Distribution of Treatment Effect",
         subtitle = "50-54 Year olds")

ggplot(data = simulated.prop.diffs, mapping = aes(x=diff.5559)) +
    geom_histogram(binwidth=0.05) +
    geom_vline(aes(xintercept = resp.prop.diff[2]), color = "red") +
    labs(x = "Difference in P(Surv > 400 days) between Treatment and Control",
         y = "Count",
         title = "Figure 2: Sampling Distribution of Treatment Effect",
         subtitle = "55-59 Year olds")

ggplot(data = simulated.prop.diffs, mapping = aes(x=diff.6064)) +
    geom_histogram(binwidth=0.05) +
    geom_vline(aes(xintercept = resp.prop.diff[3]), color = "red") +
    labs(x = "Difference in P(Surv > 400 days) between Treatment and Control",
         y = "Count",
         title = "Figure 3: Sampling Distribution of Treatment Effect",
         subtitle = "60-64 Year olds")

ggplot(data = simulated.prop.diffs, mapping = aes(x=diff.6569)) +
```

```
    geom_histogram(binwidth=0.05) +
    geom_vline(aes(xintercept = resp.prop.diff[4]), color = "red") +
    labs(x = "Difference in P(Surv > 400 days) between Treatment and Control",
         y = "Count",
         title = "Figure 4: Sampling Distribution of Treatment Effect",
         subtitle = "65-69 Year olds")

ggplot(data = simulated.prop.diffs, mapping = aes(x=diff.70)) +
    geom_histogram(binwidth=0.25) +
    geom_vline(aes(xintercept = resp.prop.diff[5]), color = "red") +
    labs(x = "Difference in P(Surv > 400 days) between Treatment and Control",
         y = "Count",
         title = "Figure 5: Sampling Distribution of Treatment Effect",
         subtitle = "70+ Year olds")



# p values - terrible coding
  est1 <- mean(simulated.prop.diffs$diff.5054 <= resp.prop.diff[1])
  est2 <- mean(simulated.prop.diffs$diff.5559 <= resp.prop.diff[2])
  est3 <- mean(simulated.prop.diffs$diff.6064 <= resp.prop.diff[3])
  est4 <- mean(simulated.prop.diffs$diff.6569 <= resp.prop.diff[4])
  est5 <- mean(simulated.prop.diffs$diff.70 <= resp.prop.diff[5])

  (pval1 <- 1-est1)
  (pval2 <- 1-est2)
  (pval3 <- 1-est3)
  (pval4 <- 1-est4)
  (pval5 <- 1-est5)



# Difference in difference
  (sim.mean.5054 <- mean(simulated.prop.diffs$diff.5054))
  (sim.mean.5559 <- mean(simulated.prop.diffs$diff.5559))
  (sim.mean.6064 <- mean(simulated.prop.diffs$diff.6064))
  (sim.mean.6569 <- mean(simulated.prop.diffs$diff.6569))
  (sim.mean.70 <- mean(simulated.prop.diffs$diff.70))

  sim.mean <- c(sim.mean.5054, sim.mean.5559, sim.mean.6064, sim.mean.6569, sim.mean.70)

  resp.prop.diff

  (diff.mag <- resp.prop.diff-sim.mean)


# Table for question 2
  tab2 <- data.frame(age=c("50-54","55-59", "60-64", "65-69", "70+"))
  tab2$resp.prop.diff <- resp.prop.diff
  tab2$sim.mean <- sim.mean
  tab2$diff.mag <- diff.mag

  kable(tab2, col.names = c("Age Group", "Observed Difference","Simulated Mean Difference", "Difference
```

```
                      caption = "Table 2", "pipe")
```