P8130 Fall 2021: Biostatistical Methods I

Homework III

Due Friday, 10/22 @5:00pm

Costs of Carotid Endarterectomy in Maryland

Scientific Background:

Carotid endarterectomy (CE) is a vascular surgical procedure intending to improve blood flow through the carotid artery, which ascends from the aorta to the brain. This surgery is designed to reduce the risk of stroke and sudden death. Approximately 2,000 CEs are performed each year at the more than 50 hospitals in the state of Maryland. Data on each procedure are routinely collected by the State of Maryland Health Services Cost Review Commission (HSCRC) and are publicly available.

An important question about carotid endarterectomy addressed by the HSCRC data is whether the risk of stroke or death after surgery decreases with increasing numbers of surgeries by the patient's physician and at the patient's hospital.

In this project, we will use the CE data from HSCRC to explore the distribution of procedure costs across a population of procedures conducted in Maryland for the period 1990 through 1995. An interesting question is how mean CE costs differ between men and women. We will be estimating mean costs for different strata and by using confidence intervals and tests of hypotheses to address the question of how the CE cost distribution differs between men and women. Here we have list of CE values for the entire population of Maryland so that we can directly calculate the "truth" (population means for men and women); in actual scientific studies, we have only a sample (subset). By pretending we don't know the true population values, we can see statistical inference in action.

Problem 1 (3 points)

Draw a random sample without replacement of 200 observations (100 men and 100 women) from the entire CE data set named ce8130entire.csv. Call this first sample "A" and save the sample. In "sex" variable, men are identified by "1", and women by "2". Note: To obtain the sample data set of approximately 200 observations, you can use the following code. Replace the "set.seed" number with an integer of your choice (3 points).

```
population = read.csv("./ce8130entire.csv")
library(dplyr)
```

```
##
##    'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
set.seed(2000) #replace 1234 with an integer
A = population %>%
  group_by(sex) %>%
  sample_n(100)
```

Problem 2 (3 points)

Now use the same seed as before but this time draw a random sample without replacement of 60 observations (30 men and 30 women) and call it sample "B" (Note that Sample "B" is more than 3 times smaller than sample "A"). Save it as a separate sample. Replace the seed number with the same seed number as you used above (3 points).

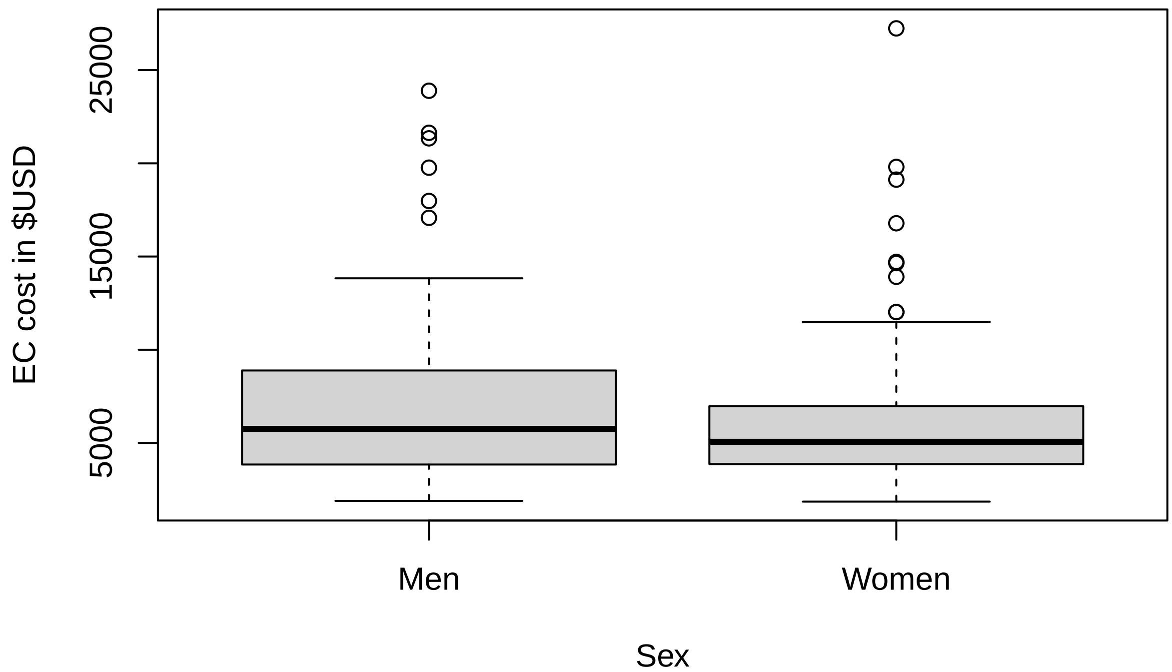
```
set.seed(1234) #replace seed number with the same integer you used above
B = population %>%
  group_by(sex) %>%
  sample_n(30)
```

Problem 3 (3 points)

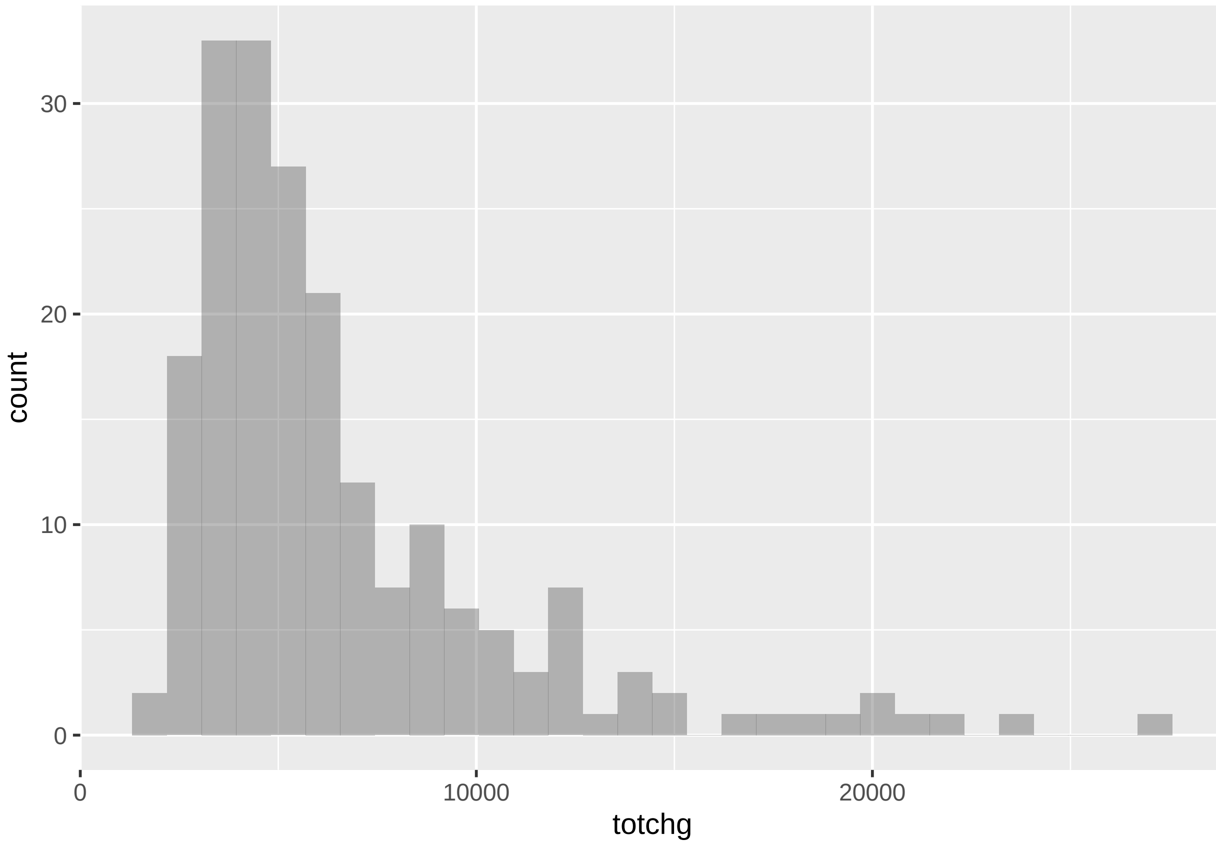
Using sample "A", display the distribution of CE cost in $USD (variable name: "totchg") separately for men and women using side-by-side boxplots and histograms. Label your figures appropriately.

```
#Suggested generic code is provided as a starting point.
#Feel free to modify or use any other function/s as deemed necessary.
boxplot(A$totchg ~ A$sex, names=c("Men","Women"), ylab ="EC cost in $USD", xlab ="Sex")

library(ggplot2)
```




```
ggplot(A, aes(x = totchg)) +
  geom_histogram(aes(color = sex, fill = sex),
                 position = "identity", bins = 30, alpha = 0.4)
```



Problem 4 (6 points)

Calculate the mean CE cost and 95% confidence interval separately for men and women in sample "A" as well as sample "B". Assume we don't know the population variance. Plot the sample "A" and sample "B" confidence intervals next to each other (by sex). How do they differ, which confidence intervals are wider? Explain why. ##Note: For the purposes of confidence interval estiamteion and hypothesis testing, let's assume that all the assumptions, including the assumption of normal distribution, are met.

```
#Suggested generic code is provided as a starting point.
#Feel free to modify or use any other function/s as deemed necessary.
A$sample = "Sample A"
B$sample = "Sample B"
A_B = rbind(A, B)
library(Rmisc)
```

```
##      lattice

##      plyr

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------

##
##    'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
A_B_summary <- summarySE(A_B, measurevar="totchg", groupvars=c("sex","sample"))
A_B_summary$sex <- as.factor(A_B_summary$sex)
p_dodge = position_dodge(0.1) # move them .05 to the left and right
plot = ggplot(A_B_summary, aes(x=sample, y=totchg, colour=sex)) +
    geom_errorbar(aes(ymin=totchg-ci, ymax=totchg+ci), width=.1, position=p_dodge) +
    geom_point(position=p_dodge)
```

Problem 5 (4 points)

Conduct test of equality of variance of CE cost among men vs women in sample A and interpret your results.

```
#Suggested generic code is provided as a starting point.
#Feel free to modify or use any other function/s as deemed necessary.
var.test(totchg ~ sex, data = A)
```

```
##
##  F test to compare two variances
##
## data:  totchg by sex
## F = 1.2113, num df = 99, denom df = 99, p-value = 0.3418
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8150345 1.8003222
## sample estimates:
## ratio of variances
##           1.211332
```

Problem 6 (5 points)

Using sample "A", calculate the difference between the mean CE costs for men and women (cost in men -cost in women). Calculate a 95% CI for this difference. Assume we don't know the population variance. Your decision of equal vs unequal variance should be based on your answer in Problem 5.

Problem 7 (7 points)

Now use sample "A" to test the hypothesis whether men and women have a different CE cost. State the null and alternative hypotheses and interpret your results.

```r
#Suggested generic code is provided as a starting point.
#Feel free to modify or use any other function/s as deemed necessary.
#res <- t.test(totchg ~ sex, data = A, var.equal = TRUE OR FALSE)
```

Problem 8 (11 points)

Use your results from Sample A: graphs, estimates, confidence intervals, and/or test results, to write a one paragraph summary of your findings regarding the average costs of CE for men and women. Write as if for an audience of health services researchers. Be quantitative and use health-services language, rather than statistical jargon in your write-up.

Problem 9 (4 points)

Now for the truth, which we have the luxury of knowing in this problem set. Compute the actual mean CE cost for men (μ_M) and for women (μ_W) for the whole population (CE8130entire.csv). Also calculate the difference ($\mu_M - \mu_W$). Do your 95% CIs include the true means?

Problem 10 (4 points)

If each student in a class of 140 calculates a 95% confidence interval for ($\mu_M - \mu_W$), how many of these intervals do you expect to contain the true population mean difference? Calculate the probability that all 140 will contain the true population mean difference.