P8130 Fall 2021: Biostatistical Methods I Homework III

Hao Zheng hz2770

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).

```r
population = read.csv("./ce8130entire.csv", encoding = "UTF-8") %>%
  mutate(
    sex = case_when(sex == 1 ~ "man", sex == 2 ~ "woman")
  )

set.seed(2000)
A = population %>%
  group_by(sex) %>%
  sample_n(100)

A
```

```
## # A tibble: 200 x 7
## # Groups:   sex [2]
##    X.U.FEFF.provnum sex    race smoker totchg   age  year
##               <int> <chr> <int>  <int>  <int> <int> <int>
##  1               47 man       0      0   7839    60  1994
##  2               47 man       0      0   2575    70  1993
##  3               18 man       0      0   7020    66  1992
##  4               41 man       0      0   2895    63  1993
##  5               42 man       0      0   4062    72  1990
##  6               12 man       0      0   6358    85  1994
##  7               47 man       0      0   2786    60  1991
##  8               32 man       0      0   4263    78  1995
##  9               23 man       0      0   3458    69  1995
## 10               18 man       0      0  10589    71  1993
## # ... with 190 more rows
```

Just draw a random Sample A.

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(2000)
B = population %>%
  group_by(sex) %>%
  sample_n(30)

B
```

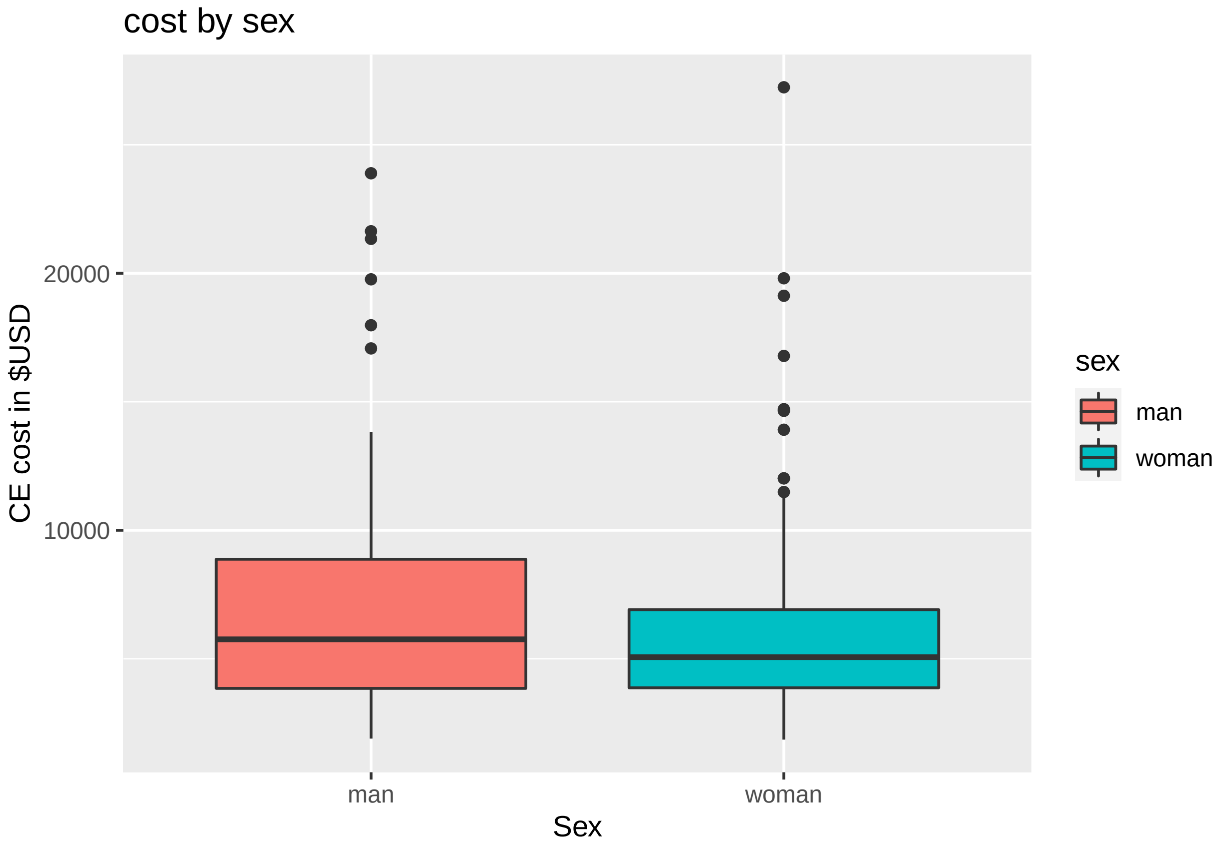
```
## # A tibble: 60 x 7
## # Groups:   sex [2]
##    X.U.FEFF.provnum sex    race smoker totchg   age  year
##               <int> <chr> <int>  <int>  <int> <int> <int>
##  1               47 man       0      0   7839    60  1994
##  2               47 man       0      0   2575    70  1993
##  3               18 man       0      0   7020    66  1992
##  4               41 man       0      0   2895    63  1993
##  5               42 man       0      0   4062    72  1990
##  6               12 man       0      0   6358    85  1994
##  7               47 man       0      0   2786    60  1991
##  8               32 man       0      0   4263    78  1995
##  9               23 man       0      0   3458    69  1995
## 10               18 man       0      0  10589    71  1993
## # ... with 50 more rows
```

Then draw Sample B.

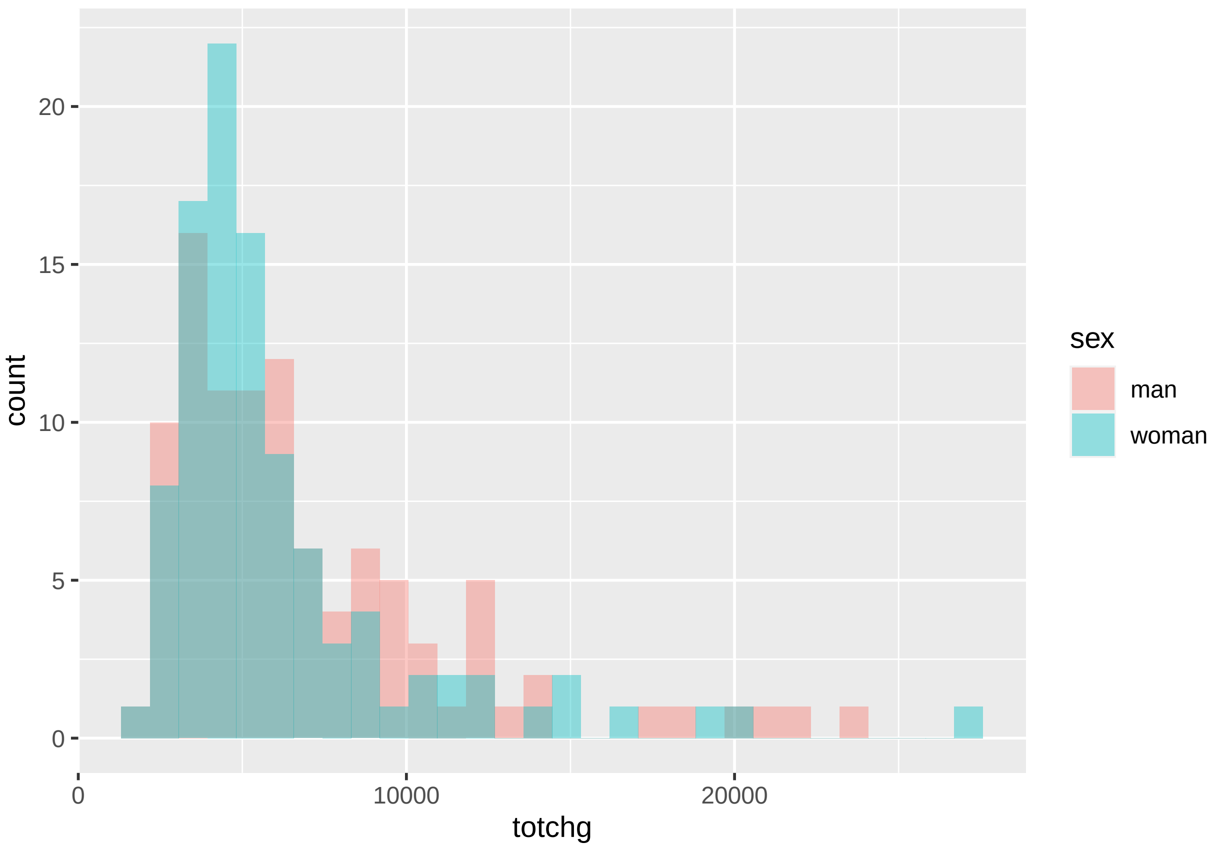
Problem 3 (3 points)

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.

```
# boxplot
A %>%
  ggplot(aes(x = sex, y = totchg, group = sex, fill = sex))+
  geom_boxplot()+
  labs(x = "Sex", y = "CE cost in $USD", title = "cost by sex")
```



```
# histogram
A %>%
  mutate(sex = recode(sex, `1` = "man", `2` = "woman")) %>%
  ggplot(aes(x = totchg, fill = sex)) +
  geom_histogram(position = "identity", bins = 30, alpha = 0.4)
```

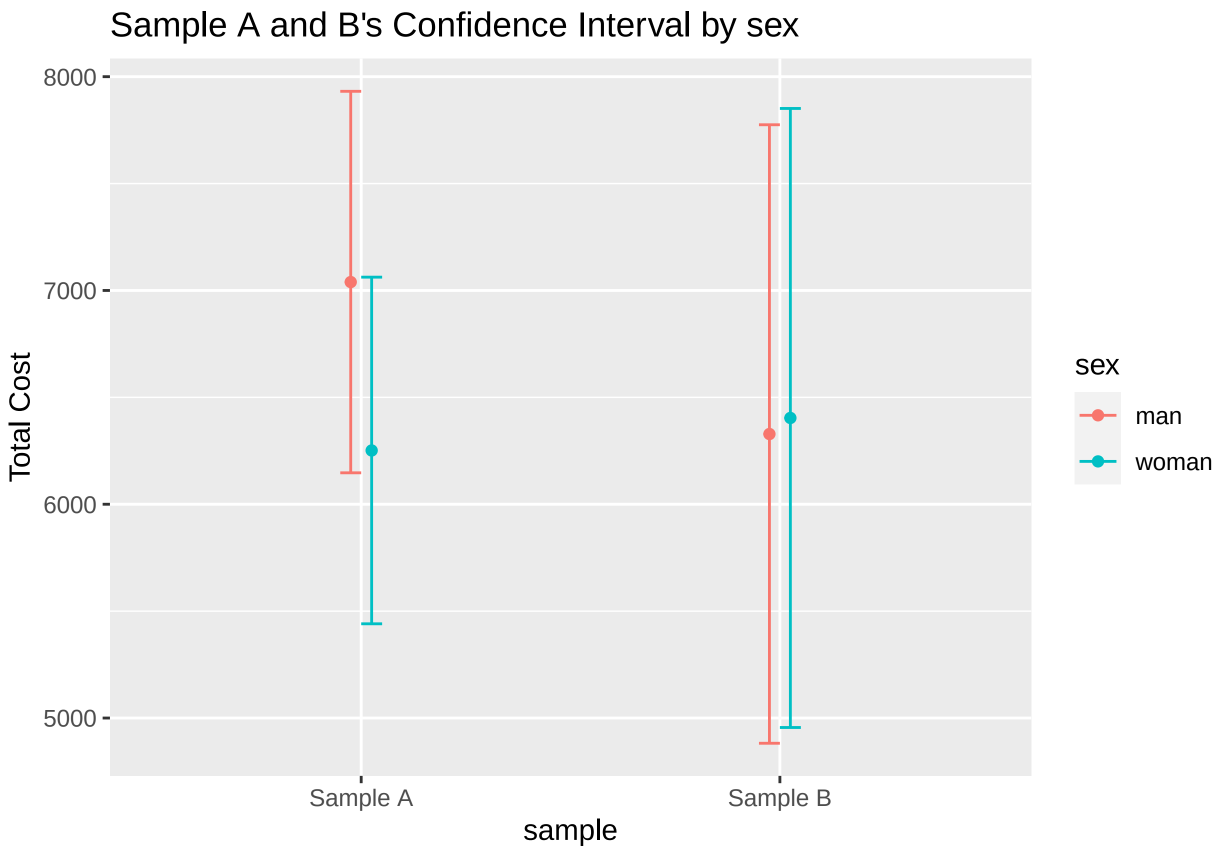


Problem 4 (6 points)

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.

```
A =
  A %>%
  mutate(sample = "Sample A")

B =
  B %>%
  mutate(sample = "Sample B")

A_B = rbind(A, B)

A_B_summary = summarySE(
  A_B, measurevar="totchg", groupvars=c("sex","sample")
  ) %>%
  mutate(sex = as.factor(sex))

p_dodge = position_dodge(0.1) # move them .05 to the left and right
```

```
plot =
  A_B_summary %>%
  ggplot(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) +
  labs(x = "sample", y = "Total Cost", title = "Sample A and B's Confidence Interval by sex")

plot
```



The confidence interval for Sample B is much wider compared to sample A. Because Sample A has larger sample size n than Sample B, and when we calculate the confidence interval, in the standard error term, n is in the denominator. Therefore, bigger value of n will lead to a smaller confidence interval.

Problem 5 (4 points)

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

```
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
```

```
Fcrit = qf(.975, df1 = 99, df2 = 99)
Fcrit
```

```
## [1] 1.486234
```

We conducted a F-Test to test whether the two variances are equal. We fail to reject the null hypothesis that the variances are the same since `F = 1.2113 < Fcrit = 1.486234`. In addition, the 95% confidence interval involves the value 1.

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.

From problem 5, we can assume same variances for man and woman in sample A. Then use the t-distribution to construct the confidence interval since we don't know the population variance.

```
sd_men_A =
  A %>%
  filter(sex == "man") %>%
  summarise(sd = sd(totchg)) %>%
  pull(sd)

sd_women_A =
  A %>%
  filter(sex == "woman") %>%
  summarise(sd = sd(totchg)) %>%
  pull(sd)

mean_men_A =
  A %>%
  filter(sex == "man") %>%
  summarise(mean = mean(totchg)) %>%
  pull(mean)

mean_women_A <- A %>%
  filter(sex == "woman") %>%
  summarise(mean = mean(totchg)) %>%
  pull(mean)

# Calculate the pooled standard deviation
std_pooled = sqrt(((100 - 1) * sd_men_A^2 + (100 - 1) * sd_women_A^2) / (100 + 100 - 2))
```

```
# calculate critical value
tcrit = qt(0.975, df = 100 + 100 - 2)

# Calculate the 95% CI
lower = mean_men_A - mean_women_A - tcrit * std_pooled * sqrt(1/100 + 1/100)
upper = mean_men_A - mean_women_A + tcrit * std_pooled * sqrt(1/100 + 1/100)

CI = c(lower, upper)
CI
```

```
## [1] -410.2963 1985.9163
```

Therefore, the 95% CI for the difference between the mean CE costs for men and women is [-410.2963, 1985.9163].

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.

```
# Conduct a hypothesis test for the difference in problem 6.
res = t.test(totchg ~ sex, data = A, var.equal = TRUE, paired = FALSE)
res
```

```
##
##  Two Sample t-test
##
## data:  totchg by sex
## t = 1.2967, df = 198, p-value = 0.1962
## alternative hypothesis: true difference in means between group man and group woman is not equal to 0
## 95 percent confidence interval:
##  -410.2963 1985.9163
## sample estimates:
##   mean in group man mean in group woman
##             7039.32             6251.51
```

The null hypothesis: `H0: mean_men_A = mean_women_A`; Alternative hypothesis: `H1: mean_men_A != mean_women_A`. We get p-value = 0.1962 > 0.05, so we cannot reject the null hypothesis. So we cannot say the mean CE costs for men and women in sample A is different.

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.

First, in order to decide whether there is a difference between different sex on average CE costs, we need to specify whether we can assume equal variance. So we conduct a F-test and find out that there in no evidence indicating unequal variances. We can assume equal variances for different sex regarding to average CE costs.

Next, since we don’t know the population variances, we use a t-test. The results shows: We cannot say the average CE costs for men and women are different, and we are 95% confident that the true difference between the average CE cost for men and women is between `-410.2963` and `1985.9163`. We also find the difference between the mean of the two group is 787.81.

We also inferred from sample A and B that larger sample size will lead to a narrower confidence interval. Therefore, for further study, if we want to obtain a narrower 95% CI, we can try to increase the sample size.

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?

```
# Calculate the actual mean for the whole population
true_mean =
  population %>%
  group_by(sex) %>%
  summarise(mean_CEcost = mean(totchg)) %>%
  pull(mean_CEcost)

# Calculate the actual mean CE costs for men and women
mean_men =
  population %>%
  filter(sex == "man") %>%
  summarise(mean_CEcost = mean(totchg)) %>%
  pull(mean_CEcost)

mean_women =
  population %>%
  filter(sex == "woman") %>%
  summarise(mean_CEcost = mean(totchg)) %>%
  pull(mean_CEcost)

# Obtain the true difference
true_diff = mean_men - mean_women

true_mean
```

```
## [1] 6946.585
```

```
mean_men
```

```
## [1] 6890.872
```

```
mean_women
```

```
## [1] 7014.377
```

```
true_diff
```

```
## [1] -123.5047
```

The actual mean CE cost for the whole population is 6946.58, for men and women in the population is 6890.87 and 7014.38. And the 95% CI for sample A and B both included the true actual mean CE costs for men and women. The true difference for the mean is -123.5.

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.