P8130 Fall 2021: Biostatistical Methods I (Homework 3)

TANVIR KHAN

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)
```

```
##
## Attaching package: '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(204) #replace 1234 with an integer
A = population %>%
  group_by(sex) %>%
  sample_n(100)
```

Analysis: In the code chunk above, we have obtained a sample of 200 observations (100 men and 100 women) from the data set.

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

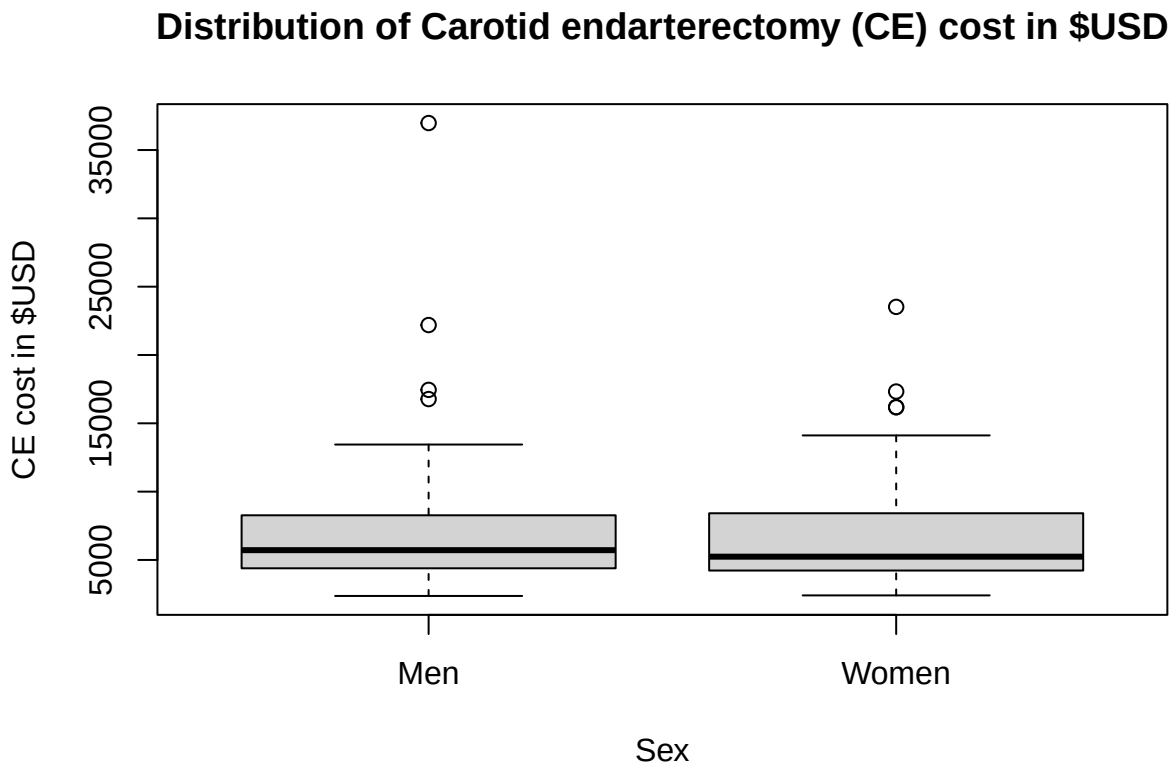
Analysis: In the code chunk above, we have obtained a sample of 60 observations (30 men and 30 women) from the data set.

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.

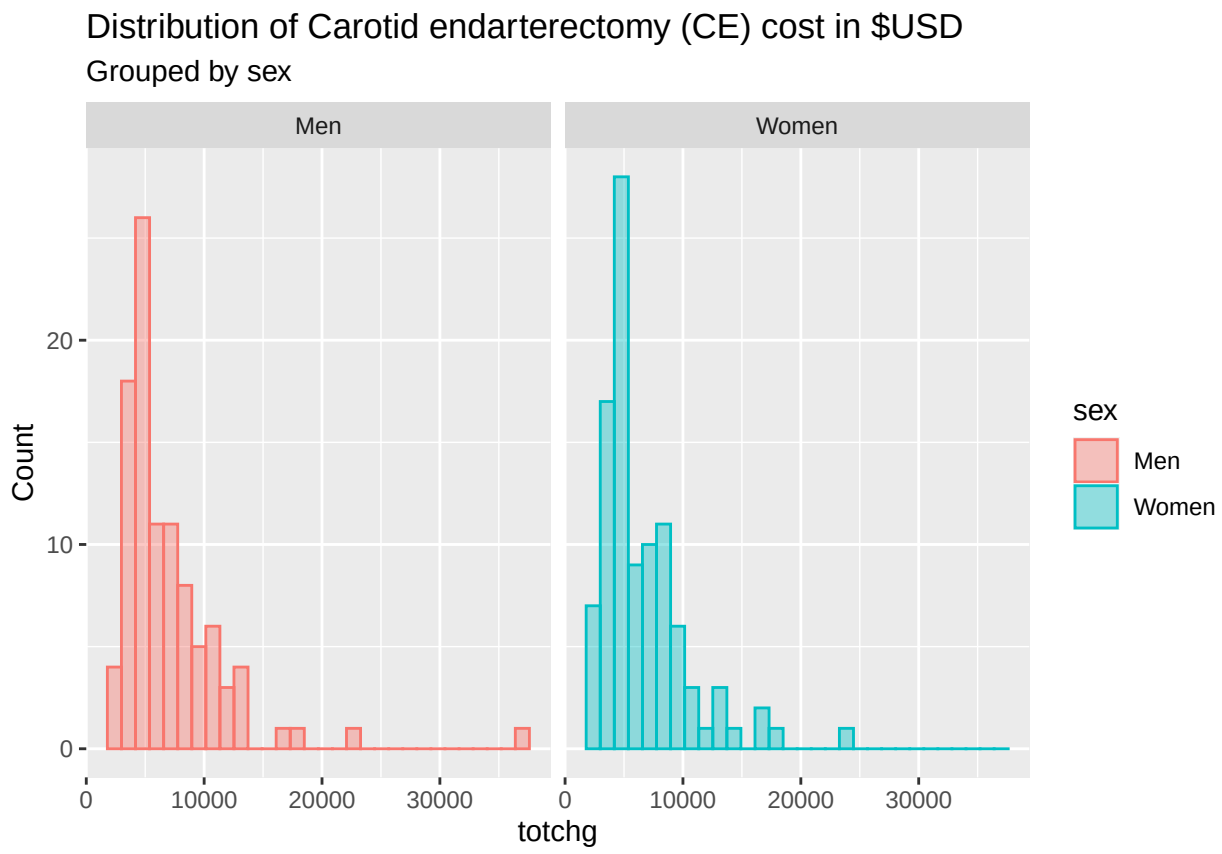
```
#Suggested generic code is provided as a starting point.
#Feel free to modify or use any other function/s as deemed necessary.

boxplot(A$totchg ~ A$sex,
        names = c("Men","Women"),
        main = "Distribution of Carotid endarterectomy (CE) cost in $USD",
        ylab = "CE cost in $USD",
        xlab = "Sex")
```



Analysis of boxplot: The side by side boxplot grouped by sex also shows that both men and women on average are towards the lower part of the CE cost. However in the men boxplot, it shows an extreme outlier (cost of CE being more than $25,000). In the women boxplot, it seems like there is two outliers but not as extreme as that one outlier in the men boxplot.

```
library(ggplot2)
A %>%
  mutate(sex = ifelse(sex == 1, "Men", "Women")) %>%
  ggplot(aes(x = totchg)) +
  geom_histogram(aes(color = sex, fill = sex),
                 position = "identity", bins = 30, alpha = 0.4) +
  labs(
      title = "Distribution of Carotid endarterectomy (CE) cost in $USD",
      subtitle = "Grouped by sex",
      x = "totchg",
      y = "Count") +
  facet_grid(~ sex)
```



Analysis of histogram plot: The stacked histogram shows the distribution of Carotid endarterectomy (CE) cost in United Stated of America between sex groups. Based of the stacked histogram, both men and women are shown towards the lower price of CE. However there are some data that shows men may deal with higher costs of CE.

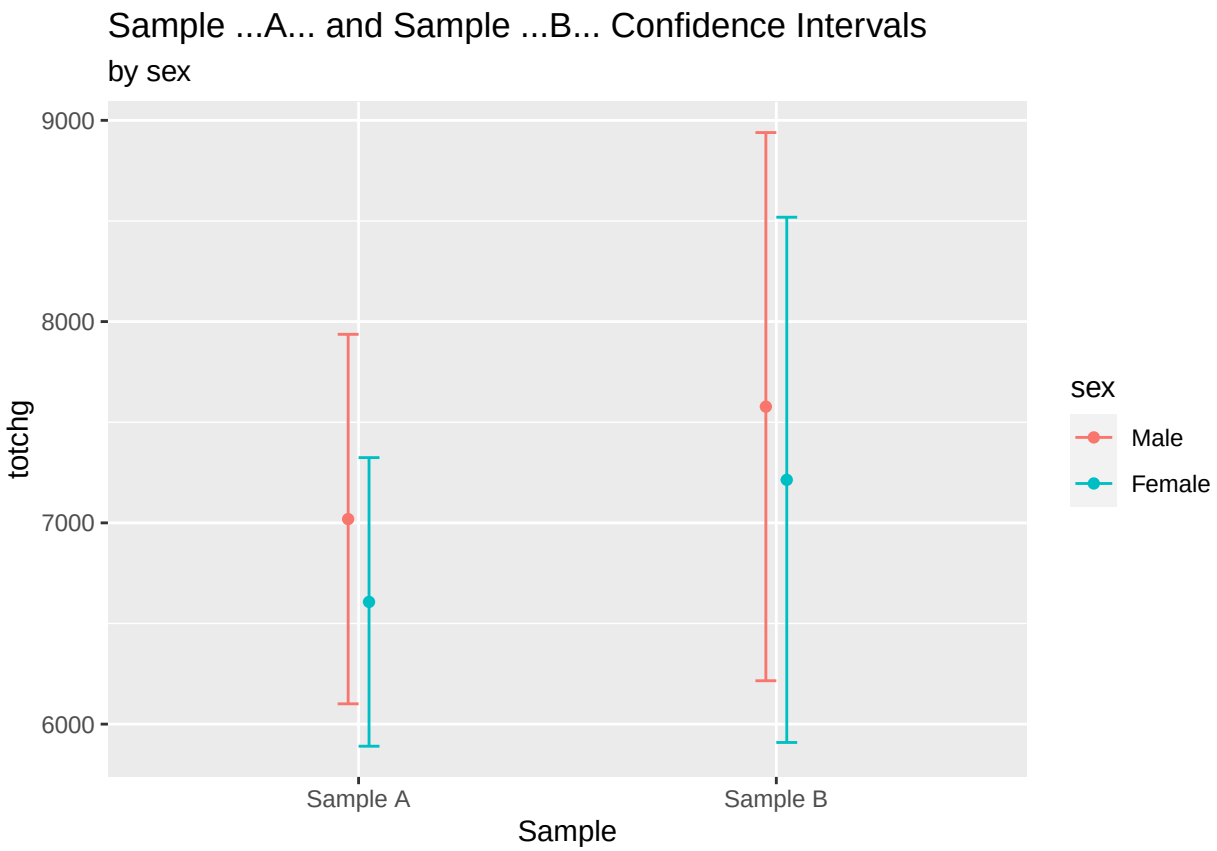
Problem 4 (6 points)

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)
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) +
    labs(
      title = "Sample "A" and Sample "B" Confidence Intervals",
      subtitle = "by sex",
      x = "Sample",
      y = "totchg") +
    scale_color_discrete(
      labels = c("Male", "Female")
    )
plot
```



We know that Sample A has larger sampling population (200 individuals) than sample B sampling population (60 individuals). A larger sample size or lower variability will result in a tighter confidence interval with a smaller margin of error. A smaller sample size or a higher variability will result in a wider confidence interval with a larger margin of error. If we analyze our confidence interval plot, we see that for sample B, which has a smaller sample population has a wider confidence interval, which means there is larger margin or error. However based on the graph, sample A which has larger sample population has a narrower confidence interval.

For sample A, we may state since men in sample A has a larger standard deviation than women in sample A, this will lead to a higher standard error and wider confidence interval for men.

For sample B, we may state since men in sample B has a larger standard deviation than women in sample B, this will lead to a higher standard error and wider confidence interval for men.

```r
# SAMPLE A
# Getting the Male and Female mean from sample population A
male_mean_A <- A_B_summary$totchg[1]
female_mean_A <- A_B_summary$totchg[3]

# Calculating the Confidence Interval
male_minimum = A_B_summary$totchg[1] - A_B_summary$ci[1]
male_maximum = A_B_summary$totchg[1] + A_B_summary$ci[1]

female_minimum = A_B_summary$totchg[3] - A_B_summary$ci[3]
female_maximum = A_B_summary$totchg[3] + A_B_summary$ci[3]

# Creating the confidence interval for male and female from sample population A
Confidence_interval_male_A <- c(male_minimum, male_maximum)
Confidence_interval_female_A <- c(female_minimum, female_maximum)

# SAMPLE B
### Getting the Male and Female mean from sample population B
male_mean_B <- A_B_summary$totchg[2]
female_mean_B <- A_B_summary$totchg[4]

male_minimum_b = A_B_summary$totchg[2] - A_B_summary$ci[2]
male_maximum_b = A_B_summary$totchg[2] + A_B_summary$ci[2]

female_minimum_b = A_B_summary$totchg[4] - A_B_summary$ci[4]
female_maximum_b = A_B_summary$totchg[4] + A_B_summary$ci[4]

# Creating the confidence interval for male and female from sample population B
Confidence_interval_male_b <- c(male_minimum_b, male_maximum_b)
Confidence_interval_female_b <- c(female_minimum_b, female_maximum_b)
```

The mean cost of Carotid endarterectomy (CE) surgical procedure for male in sample A is 7018.79. The mean cost of Carotid endarterectomy (CE) surgical procedure for female in sample A is 6607.23. The mean cost of Carotid endarterectomy (CE) surgical procedure for male in sample B is 7577.4666667. The mean cost of Carotid endarterectomy (CE) surgical procedure for female in sample B is 7213.7333333.

The confidence interval for male sample population A is `r`Confidence_interval_male_A`. The confidence interval for female sample population a is 5890.2380193, 7324.2219807. The confidence interval for male sample population B is`r` `Confidence_interval_male_b`. The confidence interval for female sample population B is 5908.7775451, 8518.6891216.

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 = 1.6394, num df = 99, denom df = 99, p-value = 0.01466
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.103054 2.436525
## sample estimates:
## ratio of variances
##           1.639395
```

```
F_statistics = A_B_summary$sd[1]^2 / A_B_summary$sd[3]^2
F_statistics
```

```
## [1] 1.639395
```

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

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

```
ifelse(F_statistics > F_critical, "reject", "fail to reject")
```

```
## [1] "reject"
```

Analysis:

Ho: the variances of CE cost among men vs women in sample A are equal **Ha**: the variance of CE cost among men vs women in sample A are not equal **alpha**: 0.05

The F test-statistics is greater than F-critical value under the null hypothesis, we **reject the null hypothesis**. We **have sufficient evidence** to reject the null hypothesis that the variances of CE cost among men vs women in sample A are equal.

For the next question, when I calculate a 95% confidence interval for two independent samples, I will be doing a two-independent samples (unequal variances).

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.

```
#pooled degrees of freedom
d1 = ((3705.969^2/100) + (4352.486^2/100))^2 / ((3705.969^2/100)^2/99) + ((4352.486^2/100)^2/99)

d2 = floor(d1)

#calculating the difference between the mean CE costs for men and women (for sample A)
diff <- A_B_summary$totchg[1] - A_B_summary$totchg[3]
diff
```

```
## [1] 411.56
```

```
tcrit = qt(0.975, d2)
tcrit
```

```
## [1] 1.959964
```

```
standard_error = sqrt((A_B_summary$se[1]^2) + (A_B_summary$se[3]^2))
standard_error
```

```
## [1] 587.0531
```

```
margin_error <- (tcrit * standard_error)

CI_difference <- c(diff - margin_error, diff + margin_error)
CI_difference
```

```
## [1] -739.0429 1562.1629
```

```
#mean difference confidence interval between male and female
```

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.

```
#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 = FALSE)

# Performing Test, Obtaining the the t-value test-statistics
tstat = (A_B_summary$totchg[1] - A_B_summary$totchg[3]) / sqrt((A_B_summary$sd[1]^2/100) + (A_B_summary$
tstat
```

```
## [1] 0.701061
```

```
#Obtaining the critical value
tcrit = qt(0.975, d2)
tcrit
```

```
## [1] 1.959964
```

```
ifelse(abs(tstat) > tcrit, "reject", "fail to reject")
```

```
## [1] "fail to reject"
```

```
res
```

```
##
##  Welch Two Sample t-test
##
## data:  totchg by sex
## t = 0.70106, df = 187.02, p-value = 0.4841
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -746.5369 1569.6569
## sample estimates:
## mean in group 1 mean in group 2
##         7018.79         6607.23
```

Hypothesis Testing:

Ho: The mean CE cost of Men and Women are equal. *Ha:* The mean CE cost of Men and Women are unequal.

Analysis: The absolute t-value (0.4841) is smaller than t-critical value (1.959964) under the null hypothesis, we fail to reject the null hypothesis. This indicates that we have insufficient evidence to the reject the null hypothesis in which the means of CE cost among men and women are equal.

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.

Today I will be focusing on this fascinating question, "Is there a significant/statistical difference in cost between men and women when it comes to getting Carotid endarterectomy (CE) vascular surgical procedure?" The data and the results that will be presented today was obtained from the Carotid endarterectomy (CE) data from HSCRC and with this data we are able to explore the distribution of procedure costs across a population of procedures conducted in Maryland for the period 1990 through 1995 between men and women. Before explaining our findings, let's take get a quick snapshot of Carotid endarterectomy (CE) in the state of Maryland. It has been reported that approximately 2,000 Carotid endarterectomy (CE) are performed each year at the more than 50 hospitals in the state of Maryland. We are fascinated to know if there is a significant difference in cost ($USD) for men and women when they have this surgical procedure. Now, I will be talking about my findings. After creating our hypothesis and analyzing the data from a sample population from 200 individuals (100 men and 100 women) from the state of Maryland, it appears that the average mean cost of Carotid endarterectomy (CE) vascular surgical procedure between men and women do not differ drastically. We may state that with 95% confidence, the average cost of Carotid endarterectomy (CE) vascular surgical procedure is between -972.70 and 2184.66. This indicates that there may be no difference in cost, or there may be case where males pay less than female or females pay more than male or vice versa. To further validate our results, we conducted a statistical test. From that statistical test, it was indicated that we do not have sufficient evidence to say that the average costs of Carotid endarterectomy (CE) vascular surgical procedure is significantly different between men and women. This statistical result aligns with the average cost range for the Carotid endarterectomy (CE) vascular surgical procedure and it also aligns with out statistical graphs that we have created. The statistical graphs also show the distribution of the cost for both men and women in which we may state there is no significant difference of cost between men and women. Based on the results. this is is positive sign that there is no significant difference of cost for Carotid endarterectomy (CE) vascular surgical procedure between men and women.

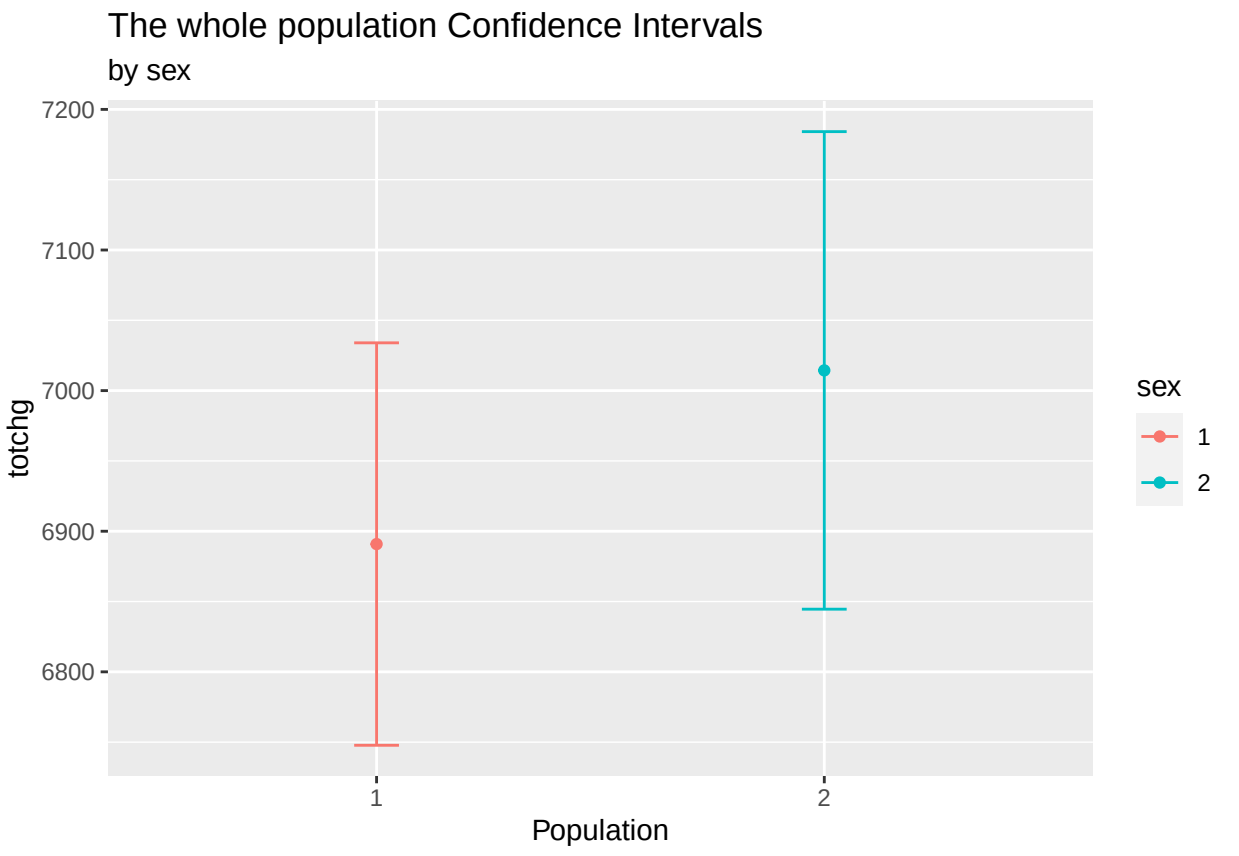
Problem 9 (4 points)

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?

```
whole_population <-
  population %>%
  group_by(sex)

whole_pop <- summarySE(whole_population, measurevar = "totchg", groupvars = c("sex"))
whole_pop$sex <- as.factor(whole_pop$sex)
p_dodge = position_dodge(0.1) # move them .05 to the left and right

plot =
    ggplot(whole_pop, aes(x = sex, y = totchg, colour = sex)) +
    geom_errorbar(aes(ymin = totchg - ci, ymax = totchg + ci), width = .1, position = p_dodge) +
    geom_point(position = p_dodge) +
    labs(
      title = "The whole population Confidence Intervals",
      subtitle = "by sex",
      x = "Population",
      y = "totchg")
plot
```



```r
true_mean <- c(whole_pop$totchg[1], whole_pop$totchg[2])
true_mean
```

```
## [1] 6890.872 7014.377
```

```r
true_mean_diff <- true_mean[1] - true_mean[2]
true_mean_diff
```

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

```r
male_min = A_B_summary$totchg[1] - A_B_summary$ci[1]
male_max = A_B_summary$totchg[1] + A_B_summary$ci[1]

female_min = A_B_summary$totchg[3] - A_B_summary$ci[3]
female_max = A_B_summary$totchg[3] + A_B_summary$ci[3]

CI_sample_male <- c(male_min, male_max)
CI_sample_female <- c(female_min, female_max)

CI_sample_female
```

```
## [1] 5890.238 7324.222
```

```r
CI_sample_male
```

```
## [1] 6100.762 7936.818
```

Yes the true mean 6890.8719691, 7014.3766205 is in the 95% Confidence Interval for male and female that was calculated for sample A 6100.7615111, 7936.8184889 and 5890.2380193, 7324.2219807. Yes the true mean 6890.8719691, 7014.3766205 is in the 95% Confidence Interval for male and female that was calculated for sample B 6215.5480595, 8939.3852738 and 5908.7775451, 8518.6891216. The true mean difference -123.5046513 is also in the 95% confidence interval for sample A mean difference -739.0429035, 1562.1629035.

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.

```
expected_number_of_students = 140*.95
expected_number_of_students
```

```
## [1] 133
```

```
p = dbinom(140,140,0.95)
p
```

```
## [1] 0.00076086
```

There will be 133 students who I expect to contain the true population mean difference. The probability that all 140 will contain the true population mean difference is 7.6085998×10^{-4}.