p8130_hw3_ym2813

Anna Ma

10/18/2021

```
library(tidyverse)
library(ggplot2)
```

Problem 1

Draw a random sample without replacement of 200 observations (100 men and 100 women) from the entire CE data set, where men are identified by "1", and women by "2" in the sex variable Call this first sample "A".

```
library(dplyr)
population = read.csv("./ce8130entire.csv")
set.seed(200)
A = population %>%
  group_by(sex) %>%
  sample_n(100)
```

Problem 2

Draw a random sample without replacement of 60 observations(30 men and 30 women) and call it sample "B"

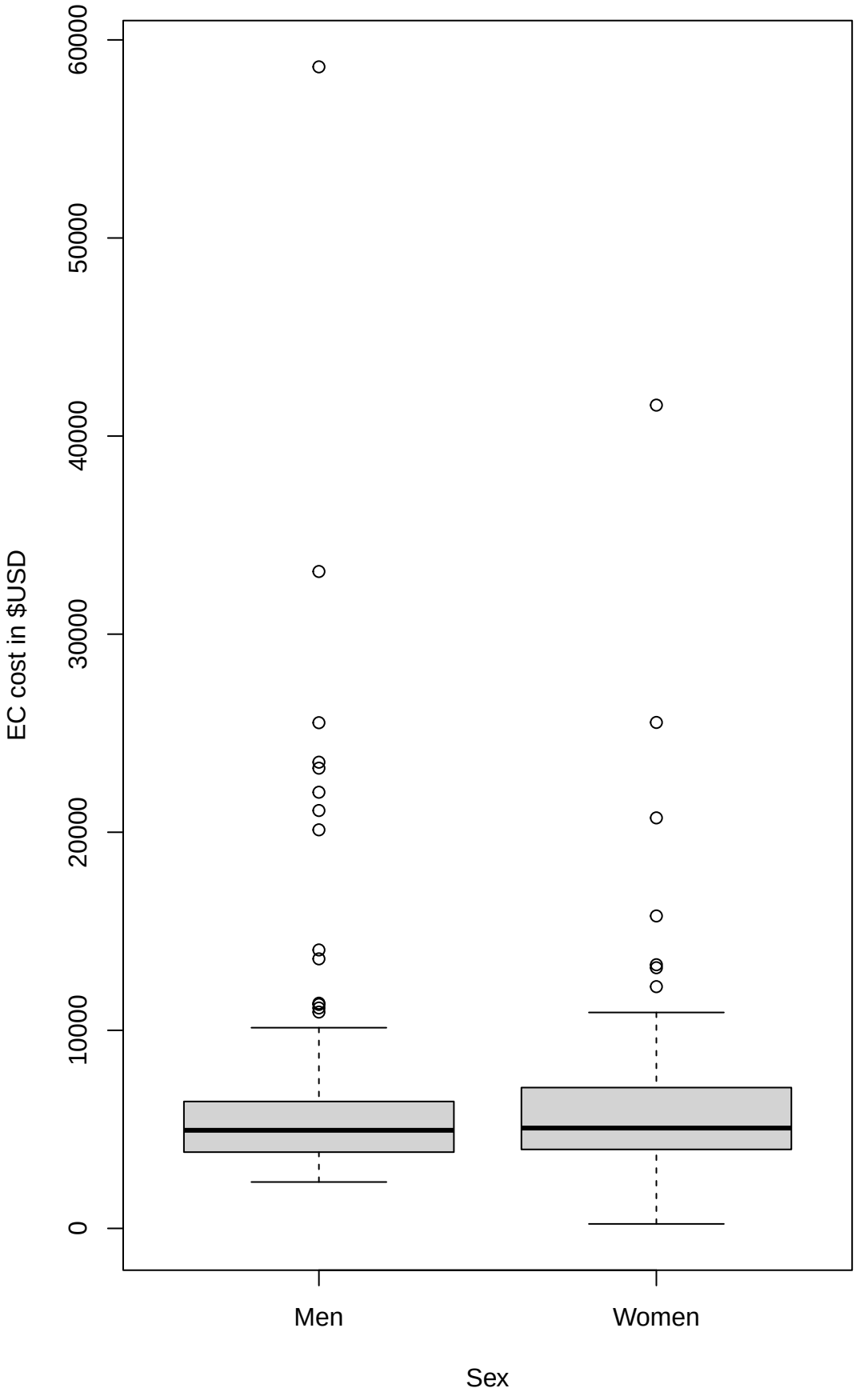
```
set.seed(200)
B = population %>%
  group_by(sex) %>%
  sample_n(30)
```

Problem 3

The distribution of CE cost in $USD for men and women.

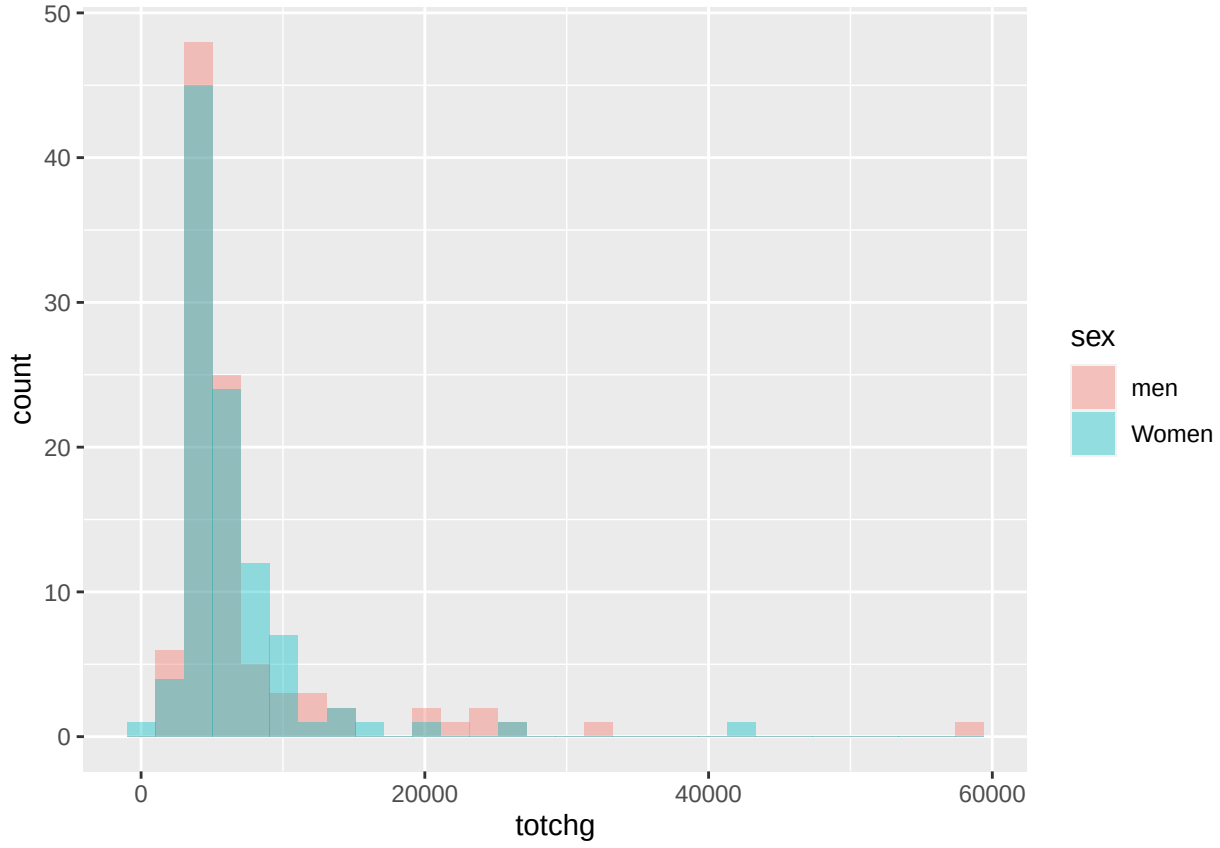
1) Side-by-side box plots

```
boxplot(A$totchg ~ A$sex, names = c("Men","Women"), ylab = "EC cost in $USD", xlab = "Sex")
```


EC cost in $USD
0
10000
20000
30000
40000
50000
60000
Men
Women
Sex

2) Histograms

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

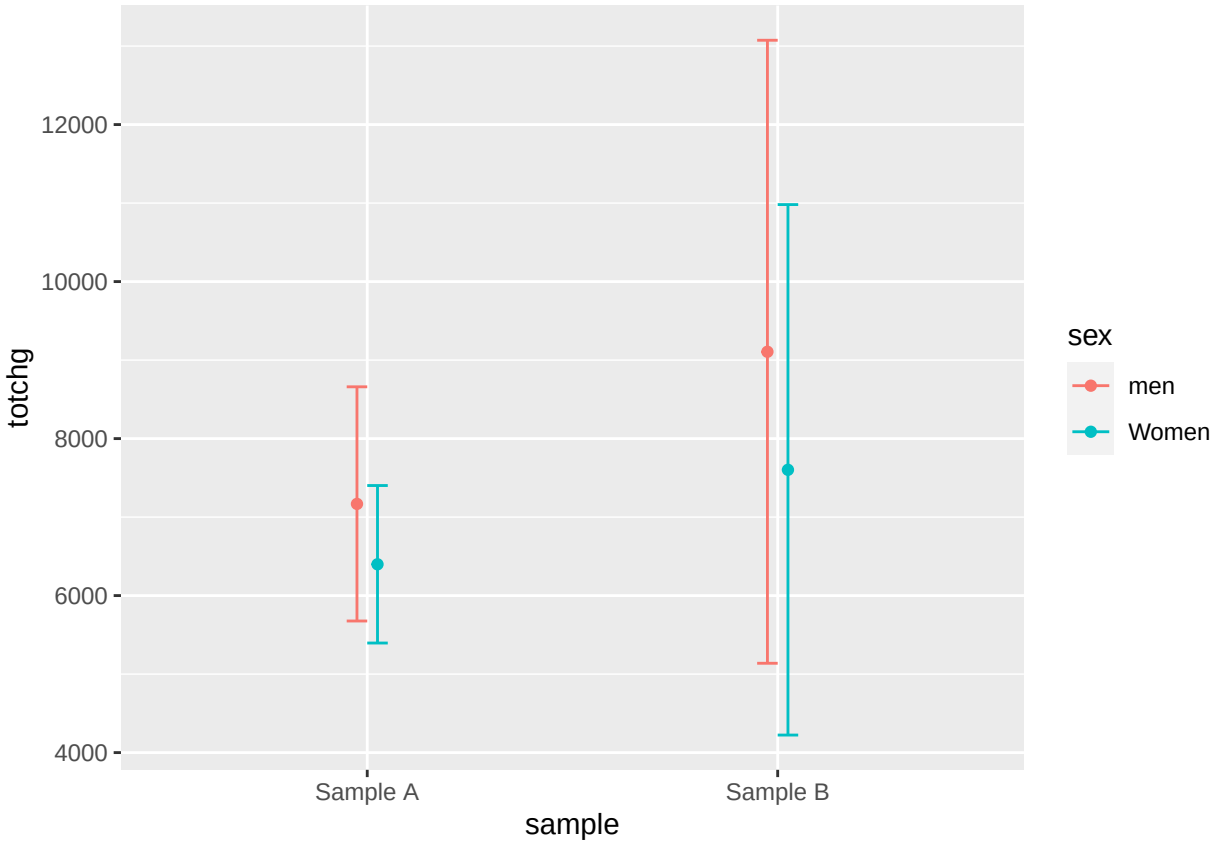


Problem 4

Mean CE cost and 95% confidence interval for men and women in sample "A" and sample "B" with unknown variance.

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

Sample_AB = rbind(A, B)

library(Rmisc)

A_B_summary <-
  summarySE(
    Sample_AB,
    measurevar = "totchg",
    groupvars = c("sex","sample")) %>%
  mutate(
   sex = recode(sex, `1` = "men", `2` = "Women"))
```

```r
A_B_summary$sex <- as.factor(A_B_summary$sex)

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

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)
```



The mean for men and women in sample A is close, whereas the mean in sample B are a bit more different. The CI for sample A has a closer and smaller range whereas in sample B, the CI has a wider and bigger range. Sample B has a wider confidence interval because it has a smaller sample n, which is in the denominator of the formula for CI. Therefore, a smaller sample will result in a wider CI.

Problem 5

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

$H_0 : \sigma^2_{men} = \sigma^2_{women}$, $H_a : \sigma^2_{men} \neq \sigma^2_{women}$

```r
var.test(totchg ~ sex, data = A,alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  totchg by sex
## F = 2.211, num df = 99, denom df = 99, p-value = 0.0001012
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  1.487670 3.286101
## sample estimates:
## ratio of variances
##           2.211025
```

The p-value of F-test is p = 0.0001012 which is smaller than the significance level 0.05. Therefore, we reject the null hypothesis that the two sample have the same variance. That is, we accept the alternative hypothesis that the two variances differs.

Problem 6

```
A_men = A %>% filter(sex == 1)
A_women = A %>% filter(sex == 2)

t.test(A_men$totchg, A_women$totchg, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  A_men$totchg and A_women$totchg
## t = 0.8491, df = 173.34, p-value = 0.397
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1018.745  2557.045
## sample estimates:
## mean of x mean of y
##   7168.00   6398.85
```

The difference between the mean is 7168.00 - 6398.85 = 769.15. The 95% confidence interval is (-1018.745, 2557.045)

Problem 7

Use sample "A" to test the hypothesis whether men and women have a different CE cost.

$H_0 : \mu_{men} = \mu_{women}$, $H_a : \mu_{men} \neq \mu_{women}$

```
t.test(totchg ~ sex, data = A, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  totchg by sex
## t = 0.8491, df = 173.34, p-value = 0.397
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -1018.745  2557.045
## sample estimates:
## mean in group 1 mean in group 2
##         7168.00         6398.85
```

The p value of the test is 0.397, which is greater than the significance level alpha = 0.05. So, we accept the null hypothesis. In conclusion, there is no significant difference between the mean cost for men and women.

Problem 8

The analysis of 100 men and 100 women randomly draw from the HSCRC record shows that there is no significant difference in the average cost for Carotid endarterectomy (CE) between men and women. From the boxplot graph, we can see that the mean cost for both men and women are nearly at the same level. From the hisogram, we can see that the cost for both groups overlaps a lot, meaning that the costs are similar and that the average of these cost would not differ significantly. To be more specific, we calculated the mean difference between men and women, and conducted hypothesis test to see if the mean have a difference and what that difference would be. In the 200 people sample we drew, the difference between the average cost for men and women is 769.15. Our test shows that the average cost between men and women in the entire population would not have a significant difference neither. We are 95 confident that the actual difference in the average cost of CE for men and women between 1990 through 1995 is between -1018.745 and 2557.045. So, we conclude that the average costs of CE for men and women is not significantly different.

Problem 9

```
population_df = population %>%
  group_by(sex)

mu_m_df = population_df %>% filter(sex == 1)
mu_m = mean(mu_m_df$totchg)

mu_w_df = population_df %>% filter(sex == 2)
mu_w = mean(mu_w_df$totchg)

pop_dif = mu_m - mu_w
```

The actual mean CE cost for men is μ_M = 6890.8719691 and for women μ_W = 7014.3766205. The difference $\mu_M - \mu_W$ = -123.5046513. The 95% CI (-1018.745, 2557.045) contains the true mean difference.

Problem 10

Since the event of ci containing the true mean difference has only two possible outcomes (yes and no) with a probability of 95%, the distribution of the ci containing the true mean difference follows a binomial distribution. The expectation is $E = np$ for binomial distributions. Therefore, we can expect $E = np = 140 \times 0.95 = 133$ intervals to contain the true population mean difference. The probability that all 140 will contain the true population mean difference is $0.95^{140} = 0.00076086$.