

P8130_hw3_wl2829

Wentong

10/15/2021

Problem 1

Read the dataset and save the sample.

```
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(1000)
A = population %>%
  group_by(sex) %>%
  sample_n(100)
```

Problem 2

Save a separate sample.

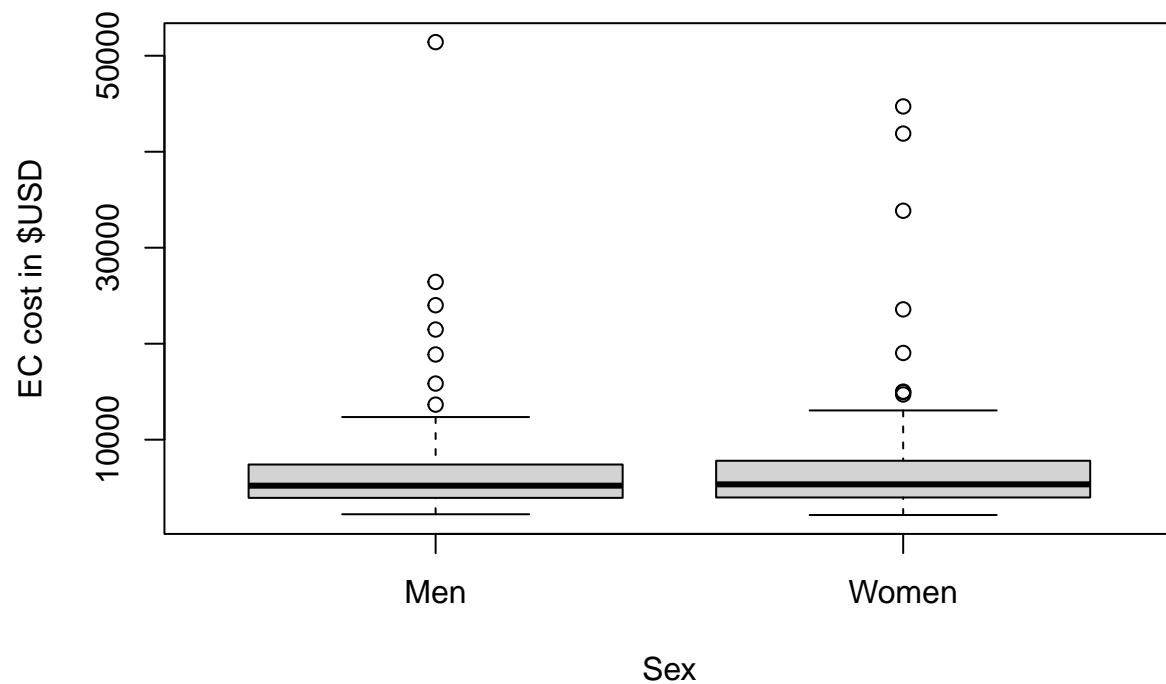
```
set.seed(1000)
B = population %>%
  group_by(sex) %>%
  sample_n(30)
```

Problem 3

Display the distribution of CE cost in \$USD separately for men and women using side-by-side boxplots and histograms.

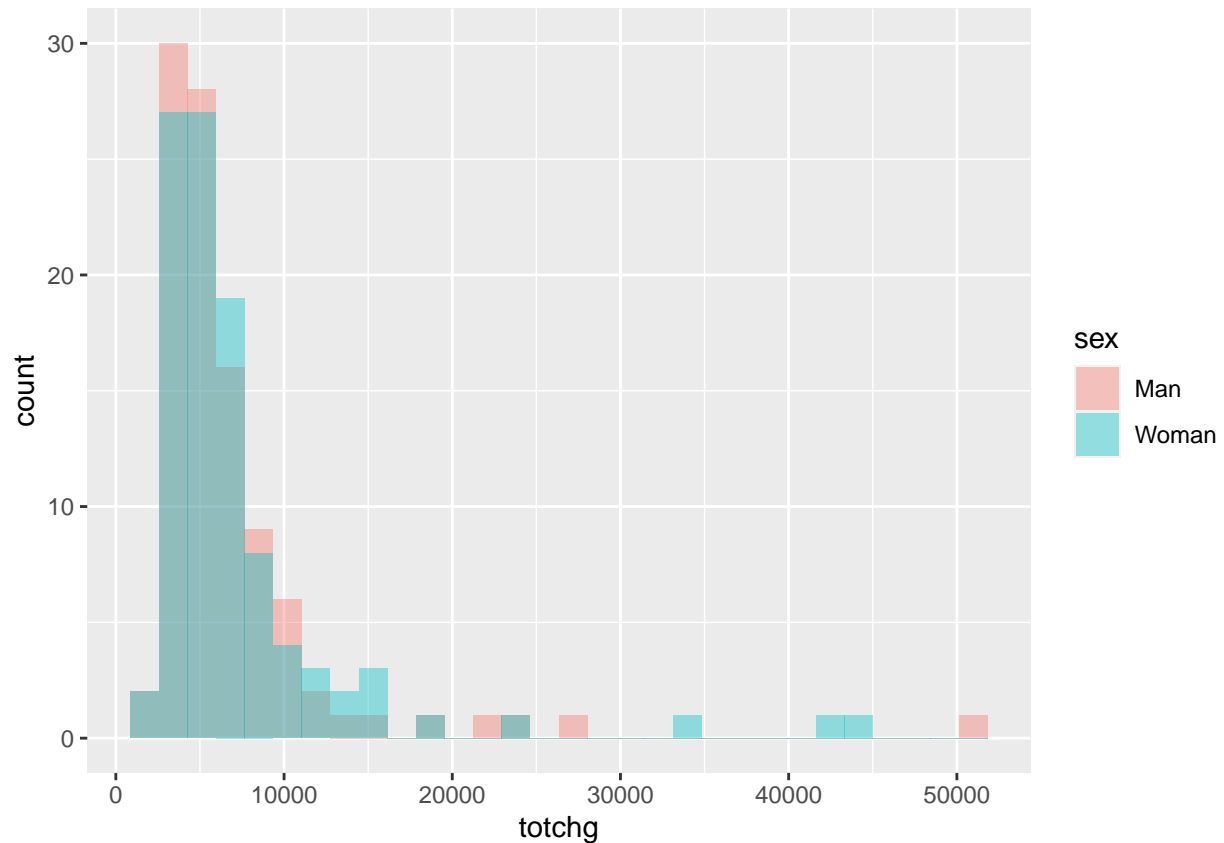
Side-by-side box plots

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



Histograms

```
library(ggplot2)
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

Calculate the mean CE cost and 95% confidence interval separately for men and women in sample “A” as well as sample “B”.

```
A$sample = "Sample A"
B$sample = "Sample B"
A_B = rbind(A, B)

library(Rmisc)

## Loading required package: lattice

## Loading required package: 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)

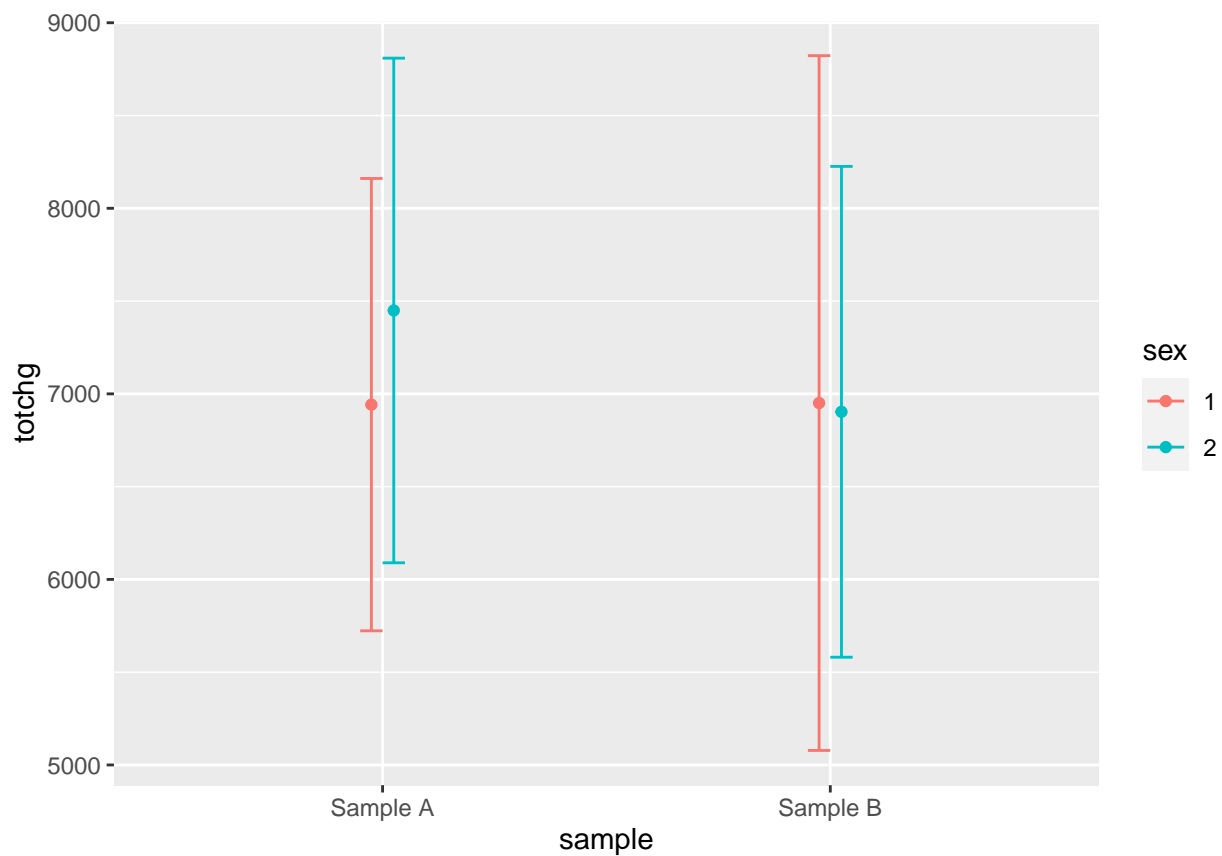
## -----
```

```
##
## Attaching package: '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)

plot
```



The confidence interval of B is wider especially man, considering it has a smaller sample.

Problem 5

Conduct test of equality of variance of CE cost among men vs women in sample A

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

```
##
## F test to compare two variances
##
## data:  totchg by sex
## F = 0.80342, num df = 99, denom df = 99, p-value = 0.2779
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.540577 1.194076
## sample estimates:
## ratio of variances
##      0.8034238
```

p-value = 0.2797 > 0.05 We cannot reject the null hypothesis at significance level 0.05. Thus, two samples have equal variance.

Problem 6

Using sample “A”, calculate the difference between the mean CE costs for men and women.

```
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 = TRUE)
```

```
##
## Two Sample t-test
##
## data:  A_men$totchg and A_women$totchg
## t = -0.55142, df = 198, p-value = 0.582
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2322.415 1307.435
## sample estimates:
## mean of x mean of y
##  6941.84  7449.33
```

The difference between the means is $6941.84 - 7449.33 = -507.49$ The 95% confidence interval is $(-2322.547, 1307.567)$

Problem 7

```
res <- t.test(totchg ~ sex, data = A, var.equal = TRUE)
res
```

```
##
## Two Sample t-test
##
## data: totchg by sex
## t = -0.55142, df = 198, p-value = 0.582
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
## -2322.415 1307.435
## sample estimates:
## mean in group 1 mean in group 2
## 6941.84 7449.33
```

Problem 8

We 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. We draw a random sample without replacement of 200 observations (100 men and 100 women) from the entire CE data. According to the sample analysis, the average CE cost for men is 6941.84 and the average CE cost for women is 7449.33. It can be shown that the CE cost for women has higher variation than men, according to the test result. So the data of women has a greater degree of dispersion. Meanwhile, we are 95% confident that the difference of the mean CE cost falls between (-2322.547, 1307.567). However, although the difference exist, the average cost are not statistically different since the high variation within the samples. In summary, the distribution of CE cost of the men is the same with the women.

Problem 9

```
population_man =
  population %>%
  filter(sex == 1)

population_woman =
  population %>%
  filter(sex == 2)

man_mean = mean(population_man$totchg)
woman_mean = mean(population_woman$totchg)
mean_dif = man_mean - woman_mean
```

The mean for men is 6890.8719691 and for women is 7014.3766205. The difference is -123.5046513. The difference is in the 95% CI (-2322.547, 1307.567).

Problem 10

This question could be regarded as a binomial distribution. The number of the intervals which contain true mean difference is equal to the expectation of the binomial distribution. $E = np = 140 \times 0.95 = 133$ About 133 students' intervals contain the true population mean difference.

```
p140 = dbinom(140, size=140, prob=0.95)
```

the probability that all 140 will contain the true population mean difference is 7.6085998×10^{-4}