TA Session 1

Chi-Yuan Fang

March 2, 2021

Contents

1 Introduction **1**
1.1 TA Information 1
1.2 TA Sessions Schedule 1
1.3 Reference 2

2 Large-Sample Approximations to Sampling Distributions **2**
2.1 The Law of Large Numbers and Consistency 2
2.2 The Central Limit Theorem 3

3 Empircial Exercise 3.1 **7**

1 Introduction

1.1 TA Information

TA: Chi-Yuan Fang

TA sessions: Tuesday 1:20 – 3:10 PM (SS 501)

Email: r09323017@ntu.edu.tw

Office hours: Friday 2:00 – 3:30 PM or by appointments (SS 643)

Class group on Facebook: Statistics (Fall 2020) and Econometrics (Spring 2021) https://www.facebook.com /groups/452292659024369/

Because screens are not clear in SS 501, I will provide the link of live streaming in the group.

1.2 TA Sessions Schedule

Week	TA Sessions	Quiz	Content	Remind
1	02/23: No class			
2	03/02: Class 1		Function, Confidence Interval, T test	03/10 Turn in HW1
3	03/09: Class 2			03/10 Turn in HW1, 03/16 Quiz 1
4	03/16: Class 3	Quiz 1		03/24 Turn in HW2
5	03/23: Class 4			03/24 Turn in HW2, 03/30 Quiz 2
6	03/30: Class 5	Quiz 2		04/14 Turn in HW3
7	04/06: No class			04/14 Turn in HW3
8	04/13: Class 6			04/14 Turn in HW3, 04/20 Quiz 3
9	04/20: Class 7	Quiz 3		**04/28 Midterm**
10	04/27: Class 8		Review and Q&A	**04/28 Midterm**, 05/05 Turn in HW4
11	05/04: Class 9			05/05 Turn in HW4, 05/11 Quiz 4

Week	TA Sessions	Quiz	Content	Remind
12	05/11: Class 10	Quiz 4		05/19 Turn in HW5
13	05/18: Class 11			05/19 Turn in HW5, 05/25 Quiz 5
14	05/25: Class 12	Quiz 5		06/02 Turn in HW6
15	06/01: Class 13			06/02 Turn in HW6, 06/08 Quiz 6
16	06/08: Class 14	Quiz 6	Review and Q&A	**06/16 Final Exam**
17	06/15: No class			**06/16 Final Exam**
18	06/22: No class			

1.3 Reference

Introduction to Econometrics with R

https://www.econometrics-with-r.org

R for Data Science

https://r4ds.had.co.nz

R Markdown

https://rmarkdown.rstudio.com

Introduction to R Markdown

https://rpubs.com/brandonkopp/RMarkdown

What is a good book on learning R with examples?

https://www.quora.com/What-is-a-good-book-on-learning-R-with-examples

2 Large-Sample Approximations to Sampling Distributions

2.1 The Law of Large Numbers and Consistency

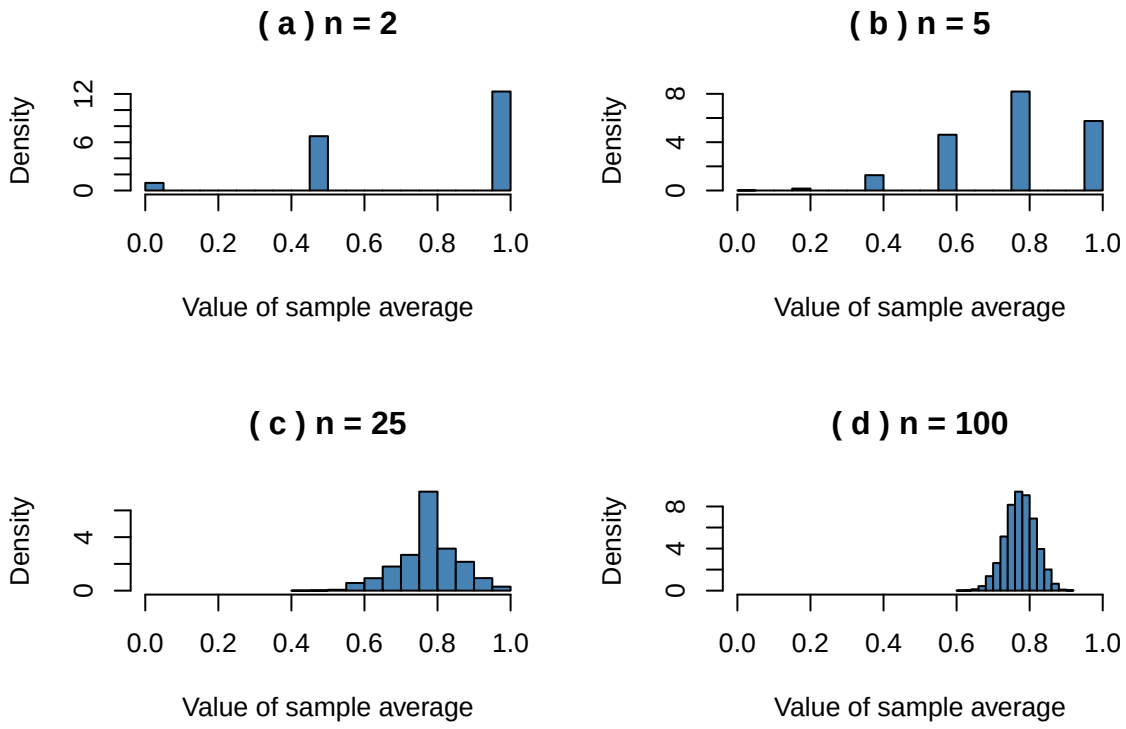
2.1.1 Figure 2.8 [p.87]

Figure 2.8: Sampling Distribution of the Sample Average of n Bernoulli Random Variables [p.87]

```r
# p.87 Figure 2.8

# set sample size and number of samples
Figure28 <- function(n, mu, reps, index){

  # n x reps sample matrix
  samples <- replicate(reps, rbinom(n, size = 1, prob = mu))

  # compute sample mean
  sample.avgs <- colMeans(samples)

  # histogram
  hist(sample.avgs,
       probability = TRUE, # freq = FALSE
       col = "steelblue",
       xlim = c(0, 1),
       xlab = "Value of sample average",
       main = paste("(", index, ") n =", n)) #
}
```

```r
# subdivide the plot panel into a 2-by-2 array
# https://www.r-bloggers.com/2012/06/two-tips-adding-title-for-graph-with-multiple-plots-add-significan
par(mfrow=c(2, 2),oma = c(0, 0, 2, 2))

# set the number of repetitions and the sample sizes
reps <- 10000
sample.sizes <- c(2, 5, 25, 100)

# set seed for reproducibility
set.seed(12345)

for (i in sample.sizes){
  Figure28(i, 0.78, reps, letters[which(sample.sizes == i)])
}

mtext("Figure 2.8", outer = TRUE, cex = 1.3, font = 2)
```

Figure 2.8



2.2 The Central Limit Theorem

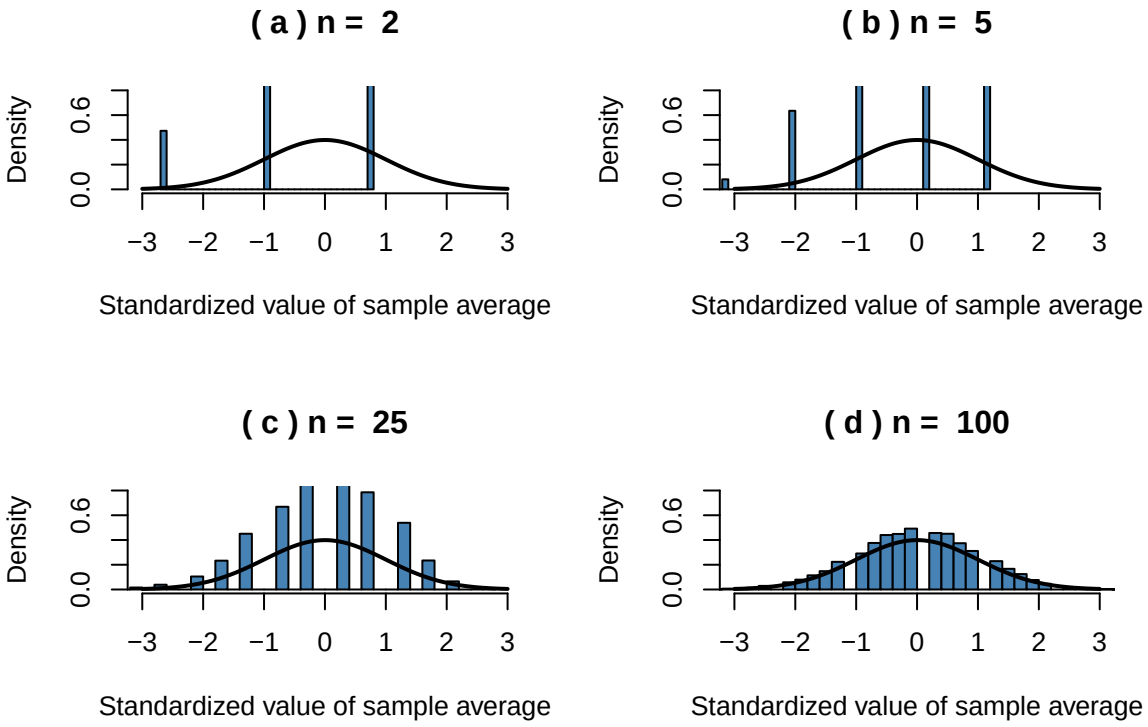
2.2.1 Figure 2.9 [p.88]

Figure 2.9: Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = 0.78$ [p.88]

```r
# p.88 Figure 2.9

Figure29 <- function(n, mu, reps, index){
  # initialize the vector of sample means
```

```r
  samplemean <- rep(0, reps) # 0 for reps times

  # initialize the vector of standardized sample means
  stdsamplemean <- rep(0, reps) # 0 for reps times

  # inner loop (loop over repetitions)
  for (i in 1:reps){
    # generate sample
    x <- rbinom(n, 1, mu) # mu = 0.78

    # sample mean
    samplemean[i] <- mean(x)

    # standardized
    stdsamplemean[i] <- (mean(x) - mu)/sqrt(mu*(1-mu)/n)
  }

  # plot histogram
  hist(stdsamplemean,
       probability = TRUE, # freq = FALSE
       col = "steelblue",
       breaks = 40,
       xlim = c(-3, 3), ylim = c(0, 0.8),
       xlab = "Standardized value of sample average",
       main = paste("(", index, ") n = ", n))

  # overlay the N(0,1) density for every iteration
  curve(dnorm(x), lwd = 2, col = "black", add = TRUE)

}

# subdivide the plot panel into a 2-by-2 array
# https://www.r-bloggers.com/2012/06/two-tips-adding-title-for-graph-with-multiple-plots-add-significan
par(mfrow=c(2, 2),oma = c(0, 0, 2, 2))

# set the number of repetitions and the sample sizes
reps <- 10000
sample.sizes <- c(2, 5, 25, 100)

# set seed for reproducibility
set.seed(12345)

for (i in sample.sizes){
  Figure29(i, 0.78, reps, letters[which(sample.sizes == i)])
}

mtext("Figure 2.9", outer = TRUE, cex = 1.3, font = 2)
```

Figure 2.9



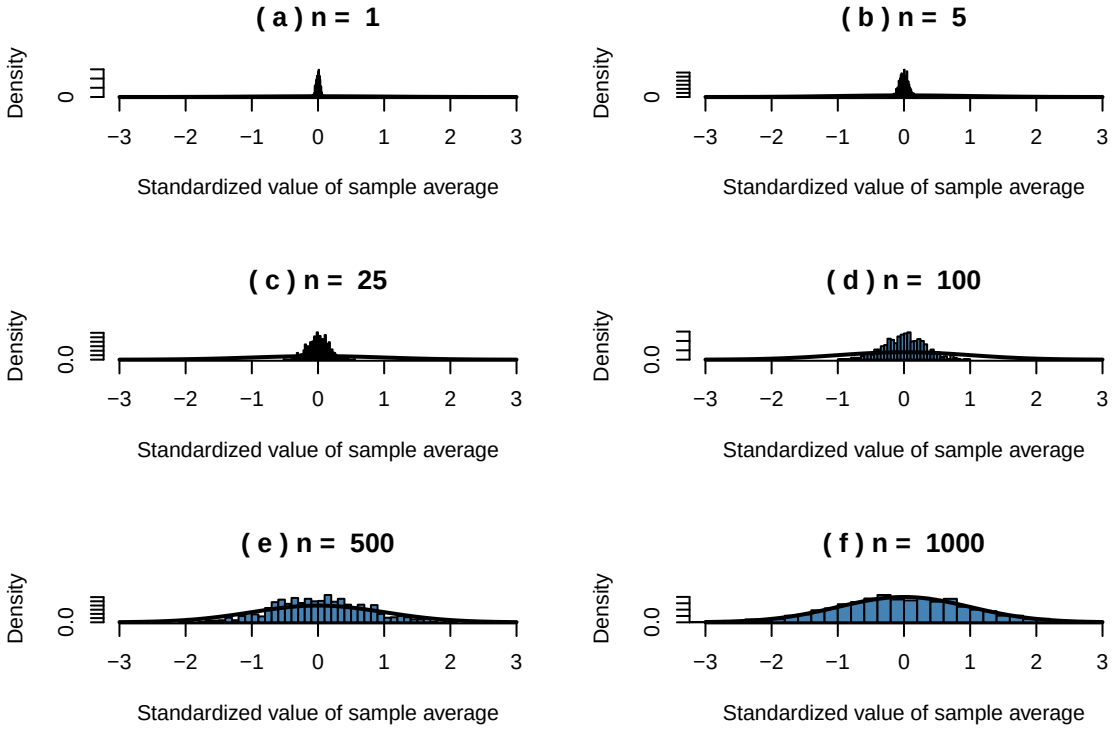
2.2.2 Figure 2.10 [p.90]

Figure 2.10: Distribution of the Standardized Sample Average of n Draws from a Skewed Population Distribution [p.90]

```
# p.90 Figure 2.10

# https://stackoverflow.com/questions/4807398/how-to-generate-distributions-given-mean-sd-skew-and-kurt

# https://stackoverflow.com/questions/1497539/fitting-a-density-curve-to-a-histogram-in-r

library("PearsonDS")

Figure210 <- function(n, mu, var, skew, kurt, reps, index){
  # set moments
  moments <- c(mean = mu, variance = var, skewness = skew, kurtosis = kurt)

  # initialize the vector of sample means
  samplemean <- rep(0, reps) # 0 for reps times

  # initialize the vector of standardized sample means
  stdsamplemean <- rep(0, reps) # 0 for reps times

  # inner loop (loop over repetitions)
  for (i in 1:reps){
    # generate sample
    x <- rpearson(reps, moments = moments)
```

```r
    # sample mean
    samplemean[i] <- mean(x)

    # standardized
    stdsamplemean[i] <- (mean(x) - mu)/sqrt(var/n)
  }

  # plot histogram
  hist(stdsamplemean,
       probability = TRUE, # freq = FALSE
       col = "steelblue",
       breaks = 40,
       xlim = c(-3, 3), #ylim = c(0, 0.8),
       xlab = "Standardized value of sample average",
       main = paste("(", index, ") n = ", n))

  # overlay the N(0,1) density for every iteration
  curve(dnorm(x), lwd = 2, col = "black", add = TRUE)

}

# subdivide the plot panel into a 3-by-2 array
# https://www.r-bloggers.com/2012/06/two-tips-adding-title-for-graph-with-multiple-plots-add-significan
par(mfrow=c(3, 2), oma = c(0, 0, 3, 3))

# set the number of repetitions and the sample sizes
reps <- 1000
sample.sizes <- c(1, 5, 25, 100, 500, 1000)

# set seed for reproducibility
set.seed(12345)

for (i in sample.sizes){
  Figure210(i, 0, 1, 0.1, 3, reps, letters[which(sample.sizes == i)])
}

mtext("Figure 2.10", outer = TRUE, cex = 1.3, font = 2)
```

Figure 2.10



3 Empircial Exercise 3.1

On the text website, http://www.pearsonglobaleditions.com, you will find the data file **CPS96_15**, which contains an extended version of the data set used in Table 3.1 of the text for the years 1996 and 2015. It contains data on full-time workers, ages 25–34, with a high school diploma or a B.A./B.S. as their highest degree. A detailed description is given in **CPS96_15_Description**, available on the website. Use these data to complete the following.

a. i. Compute the sample mean for average hourly earnings (AHE) in 1996 and 2015.
   ii. Compute the sample standard deviation for AHE in 1996 and 2015.
   iii. Construct a 95% confidence interval for the population means of AHE in 1996 and 2015.
   iv. Construct a 95% confidence interval for the change in the population means of AHE between 1996 and 2015.

Solution

```
# import data
library(readxl)
CPS96_15 <- read_xlsx("CPS96_15/CPS96_15.xlsx")

# data: 1996
CPS96 <- CPS96_15[CPS96_15$year == 1996,]

# data: 2015
CPS15 <- CPS96_15[CPS96_15$year == 2015,]

# i. sample mean # ii. standard error # iii. 95% CI
E31a <- function(x){
  # i. sample mean
```

```
  mu <- mean(x)

  # ii. standard error = sample standard deviation (s) / sqrt(n)
  se <- sd(x)/sqrt(length(x))

  # test
  test <- t.test(x,
                 alternative = c("two.sided"),
                 mu = 0, # H0
                 conf.level = 0.95) # alpha = 0.05

  # iii. 95% confidence interval
  lower <- round(test$conf.int[1], digit = 4)
  upper <- round(test$conf.int[2], digit = 4)
  CI <- paste(lower, "-"  ,upper)

  Table <- data.frame(mu, se, CI)
  colnames(Table) <- c("Mean", "Standard Error", "95% Confidence Interval")

  Table

}

tapply(CPS96_15$ahe, CPS96_15$year, E31a)
```

```
## $`1996`
##       Mean Standard Error 95% Confidence Interval
## 1 12.69326     0.08139908       12.5337 - 12.8528
##
## $`2015`
##       Mean Standard Error 95% Confidence Interval
## 1 21.23744      0.1439117       20.9553 - 21.5195
```

```
# iv. 95% confidence interval for difference

t.test(CPS15$ahe, CPS96$ahe,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS15$ahe and CPS96$ahe
## t = 51.677, df = 11049, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  8.220087 8.868268
## sample estimates:
## mean of x mean of y
##  21.23744  12.69326
```

b. In 2015, the value of the Consumer Price Index (CPI) was 237.0. In 1996, the value of the CPI was 156.9. Repeat (a), but use AHE measured in real 2015 dollars ($2015); that is,

adjust the 1996 data for the price inflation that occurred between 1996 and 2015.

Solution

```
# data: 1996
CPS96 <- CPS96_15[CPS96_15$year == 1996,]

# CPI
CPI_96 <- 156.9
CPI_15 <- 237

# adjusted 1996 AHE in $2015
ahe_adjust <- CPS96$ahe * (CPI_15/CPI_96)

# data: 1996 including adjusted 1996 AHE in $2015
CPS96 <- cbind(CPS96, ahe_adjust)

# i. sample mean
# ii. sample standard deviation
# iii. 95% confidence interval
# for adjusted 1996 AHE in $2015
E31a(CPS96$ahe_adjust)
```

```
##       Mean Standard Error 95% Confidence Interval
## 1 19.17338      0.1229546       18.9323 - 19.4144
```

```
# iv. 95% confidence interval for difference
t.test(CPS15$ahe, CPS96$ahe_adjust,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS15$ahe and CPS96$ahe_adjust
## t = 10.905, df = 13113, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.693038 2.435086
## sample estimates:
## mean of x mean of y
##  21.23744  19.17338
```

c. If you were interested in the change in workers' purchasing power from 1996 to 2015, would you use the results from (a) or (b)? Explain.

Solution

The results from part (b) adjust for changes in purchasing power. These results should be used.

d. Using the data for 2015:
   i. Construct a 95% confidence interval for the mean of AHE for high school graduates.
   ii. Construct a 95% confidence interval for the mean of AHE for workers with a college degree.
   iii. Construct a 95% confidence interval for the difference between the two means.

Solution

```
# data: high school in 2015
CPS15hs <- CPS15[CPS15$bachelor == 0, ]

# data: college in 2015
CPS15col <- CPS15[CPS15$bachelor == 1, ]

# i. 95% CI for AHE in high school # ii. 95% CI for AHE in college
tapply(CPS15$ahe, CPS15$bachelor, E31a)
```

```
## $`0`
##       Mean Standard Error 95% Confidence Interval
## 1 16.38111      0.1471396       16.0926 - 16.6696
##
## $`1`
##       Mean Standard Error 95% Confidence Interval
## 1 25.61503      0.2155545       25.1924 - 26.0376
```

```
# iii. 95% CI for difference
t.test(CPS15hs$ahe, CPS15col$ahe,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS15hs$ahe and CPS15col$ahe
## t = -35.381, df = 6463.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.745543 -8.722304
## sample estimates:
## mean of x mean of y
##  16.38111  25.61503
```

e. Repeat (d) using the 1996 data expressed in $2015.

Solution

```
# data: high school in 2015
CPS96hs <- CPS96[CPS96$bachelor == 0, ]

# data: college in 2015
CPS96col <- CPS96[CPS96$bachelor == 1, ]

# i. 95% CI for AHE in high school # ii. 95% CI for AHE in college
tapply(CPS96$ahe_adjust, CPS96$bachelor, E31a)
```

```
## $`0`
##       Mean Standard Error 95% Confidence Interval
## 1 16.26823      0.1299935       16.0134 - 16.5231
##
## $`1`
##       Mean Standard Error 95% Confidence Interval
```

```
## 1 23.03803       0.205452       22.6352 - 23.4409
```

```r
# iii. 95% CI for difference
t.test(CPS96col$ahe_adjust, CPS96hs$ahe_adjust,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS96col$ahe_adjust and CPS96hs$ahe_adjust
## t = 27.845, df = 4581.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.293168 7.246445
## sample estimates:
## mean of x mean of y
##  23.03803  16.26823
```

f. Using appropriate estimates, confidence intervals, and test statistics, answer the following questions:
   i. Did real (inflation-adjusted) wages of high school graduates increase from 1996 to 2015?
   ii. Did real wages of college graduates increase?
   iii. Did the gap between earnings of college and high school graduates increase? Explain.

Solution

i.

- **Prepare**

$H_0 : \overline{wage}^{hs}_{2015} - \overline{wage}^{hs}_{1996} = 0$ v.s. $H_1 : \overline{wage}^{hs}_{2015} - \overline{wage}^{hs}_{1996} \neq 0$

Let the significance level be 0.05.

- **Calculate**

Test statistics:

$$t = \frac{\left(\overline{w}^{hs}_{2015} - \overline{w}^{hs}_{1996}\right) - 0}{\sqrt{\frac{s^2_{2015,hs}}{n^{hs}_{2015}} + \frac{s^2_{1996,hs}}{n^{hs}_{1996}}}} \tag{1}$$

95% confidence interval:

$$\left(\overline{w}^{hs}_{2015} - \overline{w}^{hs}_{1996}\right) \pm Z_{0.025} \cdot \sqrt{\frac{s^2_{2015,hs}}{n^{hs}_{2015}} + \frac{s^2_{1996,hs}}{n^{hs}_{1996}}} \tag{2}$$

```r
E31fi <- function(x1, x2){
  # mean
  mu1 <- mean(x1); mu2 <- mean(x2)

  # s
  SD1 <- sd(x1); SD2 <- sd(x2)

  # n
```

```
  n1 <- length(x1); n2 <- length(x2)

  # difference in mean
  mu <- mu2 - mu1

  # difference in standard error
  se <- sqrt(SD1^2/n1 + SD2^2/n2)

  # 95% confidence interval
  lower <- round(mu - qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  upper <- round(mu + qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  CI <- paste(lower, "-"  ,upper)

  Table <- data.frame(mu, se, CI)
  colnames(Table) <- c("Mean", "Standard Error", "95% Confidence Interval")

  Table

}

E31fi(CPS96hs$ahe_adjust, CPS15hs$ahe)
```

```
##        Mean Standard Error 95% Confidence Interval
## 1 0.1128797      0.1963374        -0.2719 - 0.4977
```

```
t.test(CPS15hs$ahe, CPS96hs$ahe_adjust,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS15hs$ahe and CPS96hs$ahe_adjust
## t = 0.57493, df = 6714.1, p-value = 0.5654
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2720040  0.4977633
## sample estimates:
## mean of x mean of y
##  16.38111  16.26823
```

- **Conclude**

Because $p - value > 0.05$, we do not reject H_0. There is no evidence that real (inflation-adjusted) wages of high school graduates increase from 1996 to 2015.

ii.

- **Prepare**

$H_0 : \overline{wage}^{col}_{2015} - \overline{wage}^{col}_{1996} = 0$ v.s. $H_1 : \overline{wage}^{col}_{2015} - \overline{wage}^{col}_{1996} \neq 0$

Let the significance level be 0.05.

- **Calculate**

Test statistics:

$$t = \frac{\left(\overline{w}^{col}_{2015} - \overline{w}^{col}_{1996}\right) - 0}{\sqrt{\frac{s^2_{2015,col}}{n^{col}_{2015}} + \frac{s^2_{1996,col}}{n^{col}_{1996}}}} \tag{3}$$

95% confidence interval:

$$\left(\overline{w}^{col}_{2015} - \overline{w}^{col}_{1996}\right) \pm Z_{0.025} \cdot \sqrt{\frac{s^2_{2015,col}}{n^{col}_{2015}} + \frac{s^2_{1996,col}}{n^{col}_{1996}}} \tag{4}$$

```
E31fi(CPS96col$ahe_adjust, CPS15col$ahe)
```

```
##       Mean Standard Error 95% Confidence Interval
## 1 2.576997      0.2977822         1.9934 - 3.1606
```

```
t.test(CPS15col$ahe, CPS96col$ahe_adjust,
       alternative = c("two.sided"),
       mu = 0, # HO
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS15col$ahe and CPS96col$ahe_adjust
## t = 8.654, df = 6245.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.993241 3.160752
## sample estimates:
## mean of x mean of y
##  25.61503  23.03803
```

- **Conclude**

Because $p-value < 0.05$, we reject H_0. There is statistically significance evidence that real (inflation-adjusted) wages of college graduates increase from 1996 to 2015.

iii.

- **Prepare**

$H_0 : \overline{wage}^{diff}_{2015} - \overline{wage}^{diff}_{1996} = 0$ v.s. $H_1 : \overline{wage}^{diff}_{2015} - \overline{wage}^{diff}_{1996} \neq 0$

Let the significance level be 0.05.

- **Calculate**

Test statistics:

$$t = \frac{\left[\left(\overline{w}^{col}_{2015} - \overline{w}^{hs}_{2015}\right) - \left(\overline{w}^{col}_{1996} - \overline{w}^{hs}_{1996}\right)\right] - 0}{\sqrt{\left(\frac{s^2_{2015,col}}{n^{col}_{2015}} + \frac{s^2_{2015,hs}}{n^{hs}_{2015}}\right) + \left(\frac{s^2_{1996,col}}{n^{col}_{1996}} + \frac{s^2_{1996,hs}}{n^{hs}_{1996}}\right)}} \tag{5}$$

95% confidence interval:

$$\left[\left(\overline{w}^{col}_{2015} - \overline{w}^{hs}_{2015}\right) - \left(\overline{w}^{col}_{1996} - \overline{w}^{hs}_{1996}\right)\right] \pm Z_{0.025} \cdot \sqrt{\left(\frac{s^2_{2015,col}}{n^{col}_{2015}} + \frac{s^2_{2015,hs}}{n^{hs}_{2015}}\right) + \left(\frac{s^2_{1996,col}}{n^{col}_{1996}} + \frac{s^2_{1996,hs}}{n^{hs}_{1996}}\right)} \tag{6}$$

```r
E31fiii <- function(x11, x12, x21, x22){
  # mean
  mu11 <- mean(x11); mu12 <- mean(x12)
  mu21 <- mean(x21); mu22 <- mean(x22)

  # mu
  mu1 <- mu11 - mu12
  mu2 <- mu21 - mu22

  # SE
  SD11 <- sd(x11); SD12 <- sd(x12)
  SD21 <- sd(x21); SD22 <- sd(x22)

  # n
  n11 <- length(x11); n12 <- length(x12)
  n21 <- length(x21); n22 <- length(x22)

  SD1 <- sqrt(SD11^2/n11 + SD12^2/n12)
  SD2 <- sqrt(SD21^2/n21 + SD22^2/n22)

  # difference in mean
  mu <- mu2 - mu1

  # difference in standard error
  se <- sqrt(SD1^2 + SD2^2)

  # 95% confidence interval
  lower <- round(mu - qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  upper <- round(mu + qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  CI <- paste(lower, "-"  ,upper)

  Table <- data.frame(mu, se, CI)
  colnames(Table) <- c("Mean", "Standard Error", "95% Confidence Interval")

  Table
}

E31fiii(CPS96col$ahe_adjust, CPS96hs$ahe_adjust, CPS15col$ahe, CPS15hs$ahe)
```

```
##       Mean Standard Error 95% Confidence Interval
## 1 2.464117      0.3566828          1.765 - 3.1632
```

- **Conclude**

Because $0 \notin 95\%$ confidence interval, we reject H_0. There is statistically significance evidence that the gap between earnings of college and high school graduates increase.

g. Table 3.1 presents information on the gender gap for college graduates. Prepare a similar table for high school graduates, using the 1996 and 2015 data. Are there any notable differences between the results for high school and college graduates?

Solution

```r
E31g <- function(x){
  # sample mean
  mu <- mean(x)
```

```
  # standard deviation
  SD <- sd(x)

  # numbers of sample
  n <- length(x)

  Table <- data.frame(mu, SD, n)

  Table
}

# data: male in 1996
CPS96hsm <- CPS96hs[CPS96hs$female == 0, ]

# data: female in 1996
CPS96hsf <- CPS96hs[CPS96hs$female == 1, ]

# statistics: male and female in 1996
tapply(CPS96hs$ahe_adjust, CPS96hs$female, E31g)
```

```
## $`0`
##         mu       SD    n
## 1 17.78487 8.240476 2168
##
## $`1`
##         mu       SD    n
## 1 13.76968 5.830538 1316
```

```
# difference in 1996
t.test(CPS96hsm$ahe_adjust, CPS96hsf$ahe_adjust,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS96hsm$ahe_adjust and CPS96hsf$ahe_adjust
## t = 16.795, df = 3402, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.546458 4.483924
## sample estimates:
## mean of x mean of y
##  17.78487  13.76968
```

```
# statistics: male and female in 2015
tapply(CPS15hs$ahe, CPS15hs$female, E31g)
```

```
## $`0`
##         mu       SD    n
## 1 17.49846 9.026855 2222
##
## $`1`
```

```
##         mu       SD    n
## 1 14.20896 6.998409 1143
```

```r
# data: male in 2015
CPS15hsm <- CPS15hs[CPS15hs$female == 0, ]

# data: female in 2015
CPS15hsf <- CPS15hs[CPS15hs$female == 1, ]

# difference in 2015
t.test(CPS15hsm$ahe, CPS15hsf$ahe,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  CPS15hsm$ahe and CPS15hsf$ahe
## t = 11.665, df = 2857.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.736560 3.842432
## sample estimates:
## mean of x mean of y
##  17.49846  14.20896
```