NDH802 - Hypothesis testing

Chapter 9 and 10

9.55

A random sample of 10 students contains the following observations, in hours, for time spent studying in the week before final exams:

28 57 42 35 61 39 55 46 49 38

Assume that the population distribution is normal.

a. Find the sample mean and standard deviation.

```
sample = c(28,57,42,35,61,39,55,46,49,38) #input the assumptions
mu_0 = mean(sample) #we learnt this from R live session 1
sd = sd(sample)  #we learnt this from R live session 1
```

b. Test, at the 5% significance level, the null hypothesis that the population mean is 40 hours against the alternative that it is higher.

$$H_o : \mu = 40$$
$$H_1 : \mu > 40$$

R shortcuts:

```
ttest = t.test(sample, mu = 40, alternative = "greater")
ttest$p.value
```

```
## [1] 0.08392533
```

Because the p-value is higher than $\alpha = 0.05$, we fail to reject the null hypothesis.

By hand:

```
n = 10
mu = 40

t = (mu_0 - mu)/(sd/sqrt(n)) #just plug in the numbers
t_score = qt(p = 0.95, df = n-1) #use the t-table as you prefer
```

Because `t_score` = 1.8331129 is higher than `t` = 1.5, we fail to reject the null hypothesis.

9.37

The probability of type II error is:

$\beta = P(\bar{x} \leq 5.041 | \mu =$ each of the mu given in the sub-questions)

power $= 1 - \beta$.

```r
#we calculate this similarly to the way we solve exercises in previous chapters
beta_a = pnorm(q = 5.041, mean = 5.1, sd = 0.1/sqrt(16))
power_a = 1 - beta_a

beta_b = pnorm(q = 5.041, mean = 5.03, sd = 0.1/sqrt(16))
power_b = 1 - beta_b

beta_c = pnorm(q = 5.041, mean = 5.15, sd = 0.1/sqrt(16))
power_c = 1 - beta_c

beta_d = pnorm(q = 5.041, mean = 5.07, sd = 0.1/sqrt(16))
power_d = 1 - beta_d
```

9.65

$$H_o : \mu \leq 40$$
$$H_1 : \mu > 40$$

```r
#summrize the assumption
n_965 = 125
xbar_965 = 40.9
s_squared_965 = 65

t_965 = (xbar_965 - 40)/sqrt(s_squared_965/n_965)
pvalue_965 = pt(q = t_965, df = n_965-1, lower.tail = FALSE)
```

p-value = 0.1071775 is not very small, hence the claim is not very strong.

10.4

$$H_o : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

```r
#load data into your Global environment
price <- read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Exercises/House_Se

#t.test on all data
t.test(x = price$Sale.1.Price,
       y = price$Sale.2.Price,
       #here we specify paired = TRUE because we want to do matched t-test
       paired = TRUE,
       #here we specify alternative = "greater" because that is our H1
       alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  price$Sale.1.Price and price$Sale.2.Price
## t = -53.675, df = 3999, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -22943.87       Inf
```

```
## sample estimates:
## mean of the differences
##               -22261.52
```

```
#t.test on Atlanta data
t.test(x = price[price$Atlanta == 1, "Sale.1.Price"],
       y = price[price$Atlanta == 1, "Sale.2.Price"],
       paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  price[price$Atlanta == 1, "Sale.1.Price"] and price[price$Atlanta == 1, "Sale.2.Price"]
## t = -31.439, df = 999, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -17431.42       Inf
## sample estimates:
## mean of the differences
##                  -16564
```

p-value in both cases are higher than the alpha we normally specify, therefore we fail to reject the null hypotheses.

10.34

Here I chose $\alpha = 0.05$ for example, feel free to choose other α as you reason.

```
nutrition <- read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Exercises/Food_
#summary(nutrition)
```

$$H_o : \mu_{adult-metro} = \mu_{adult-nonmetro}$$
$$H_1 : \mu_{adult-metro} \neq \mu_{adult-nonmetro}$$

```
#The following code is simply data manipulation
#"PCT_OBESE_ADULTS" of adults living in metro cities
adult_metro = nutrition[nutrition$metro == 1, "PCT_OBESE_ADULTS"]
#"PCT_OBESE_ADULTS" of adults living in non-metro cities
adult_nonmetro = nutrition[nutrition$metro == 0, "PCT_OBESE_ADULTS"]
#perform the t test
t.test(x = adult_metro,
       y = adult_nonmetro,
       #we specify alternative = "two.sided" because H1 is "not equal"
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  adult_metro and adult_nonmetro
## t = -6.8361, df = 2192.3, p-value = 1.051e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1920048 -0.6605671
## sample estimates:
```

```
## mean of x mean of y
##  27.68356  28.60985
```

p-value is smaller than alpha, we reject the null hypothesis.

$$H_o : \mu_{child-metro} = \mu_{child-nonmetro}$$
$$H_1 : \mu_{child-metro} \neq \mu_{child-nonmetro}$$

```
#This is similar, but on the children
child_metro = nutrition[nutrition$metro == 1, "PCT_Child_OBESITY"]
child_nonmetro = nutrition[nutrition$metro == 0, "PCT_Child_OBESITY"]
t.test(child_metro, child_nonmetro)
```

```
##
##  Welch Two Sample t-test
##
## data:  child_metro and child_nonmetro
## t = 0.89226, df = 2300.1, p-value = 0.3723
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1542429  0.4117915
## sample estimates:
## mean of x mean of y
##  14.26857  14.13980
```

p-value is higher than alpha, we fail to reject the null hypothesis.

10.48

$\alpha = 0.01$

$$H_o : \mu_{SalesO} = \mu_{SalesC}$$
$$H_1 : \mu_{SalesO} \geq \mu_{SalesO}$$

```
ole <- read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Exercises/Ole.csv")[

t.test(x = ole$Olesales,
       y = ole$Carlsale,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  ole$Olesales and ole$Carlsale
## t = 2.5171, df = 294.57, p-value = 0.006181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  475.0596      Inf
## sample estimates:
## mean of x mean of y
##  3791.128  2412.038
```

p-value < alpha, we reject the null hypothesis.

$$H_o : \mu_{PriceO} = \mu_{PriceC}$$
$$H_1 : \mu_{PriceO} \neq \mu_{PriceO}$$

```
t.test(x = ole$Oleprice,
       y = ole$Carlpric,
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  ole$Oleprice and ole$Carlpric
## t = -0.047883, df = 303.65, p-value = 0.9618
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02968326  0.02827300
## sample estimates:
## mean of x mean of y
## 0.8187820 0.8194872
```

p-value > alpha, we fail to reject the null hypothesis.

10.52

Let x_1, x_2 denote the diets of the immigrants in the first and second interview; y_1, y_2 denote the diets of the non-immigrants in the first and second interview.

Here are the hypotheses we want to test for the first interview. You can try the second interview yourself.

$$H_o : \mu_{x1} \leq \mu_{y1}$$
$$H_1 : \mu_{x1} > \mu_{y1}$$

```
#data prep
hei <- read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Exercises/HEI.csv")[

x1 = hei[hei$immigrant == 1 & hei$daycode == 1, "HEI2005"]
y1 = hei[hei$immigrant == 0 & hei$daycode == 1, "HEI2005"]
x2 = hei[hei$immigrant == 1 & hei$daycode == 2, "HEI2005"]
y2 = hei[hei$immigrant == 0 & hei$daycode == 2, "HEI2005"]

#difference in diet from the first interview
t.test(x = x1, y = y1, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and y1
## t = 12.586, df = 1352.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.740486      Inf
```

```
## sample estimates:
## mean of x mean of y
##  57.30373  50.69956
```

```
#difference in diet from the second interview
t.test(x = x2, y = y2, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  x2 and y2
## t = 12.698, df = 1228.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.156651      Inf
## sample estimates:
## mean of x mean of y
##  60.07289  52.99928
```

10.56

Let f_1, f_2 denote the daily cost of women in the first and second interview; m_1, m_2 denote the daily cost of men in the first and second interview.

Here are the hypotheses we want to test for the first interview. You can try the second interview yourself.

$$H_o : \mu_{f1} \geq \mu_{m1}$$
$$H_1 : \mu_{f1} < \mu_{m1}$$

```
#data prep
f1 = hei[hei$female == 1 & hei$daycode == 1, "daily_cost"]
m1 = hei[hei$female == 0 & hei$daycode == 1, "daily_cost"]
f2 = hei[hei$female == 1 & hei$daycode == 2, "daily_cost"]
m2 = hei[hei$female == 0 & hei$daycode == 2, "daily_cost"]

#difference in daily cost from the first interview
t.test(x = f1, y = m1, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  f1 and m1
## t = -13.007, df = 3956.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -1.03636
## sample estimates:
## mean of x mean of y
##  4.650782  5.837215
```

```
#difference in daily cost from the seond interview
t.test(x = f2, y = m2, alternative = "less")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  f2 and m2
## t = -12.089, df = 3658, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.9433771
## sample estimates:
## mean of x mean of y
##  4.490478  5.582471
```