# Homework 2

Jun Ryu, UID: 605574052

2023-04-14

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

## Question 1

### a)

```
n <- 30 # sample size
xbar <- 23606 # sample mean
s <- 24757 # sample standard deviation

margin <- qt(0.975,df=n-1)*s/sqrt(n) # margin of error
LB <- xbar - margin # lower bound
UB <- xbar + margin # upper bound

c(LB,UB) # our 95% confidence interval
```

```
## [1] 14361.58 32850.42
```

### b)

We must assume that either the population distribution is normal or the sample size is large enough to yield good approximations.

### c)

No, the confidence interval is used to determine confidence about the true population mean, rather than another sample's mean.

### d)

Our null hypothesis is that the mean income of US residents (in 2000) was $25,000, and the alternate hypothesis is that it was not. Our derivation of the p-value is as follows:

```r
z_score <- (xbar - 25000)/(s/sqrt(n)) # get the z-score
2*pnorm(z_score) # turn it into two-tailed p-value
```

```
## [1] 0.757772
```

Since our p-value of 0.75772 is greater than the significance level of 5% (0.05), we fail to reject the null hypothesis.

### e)

The smallest significance level we could have used is approximately 76% ($0.76 > 0.757772$) to reject the null hypothesis.

## Question 2

### a)

Our formula for a 95% confidence interval given a population proportion p is: $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. So, in this case, when we increase the sample size (n), the expression inside the square root will yield a smaller value, leading to a smaller margin of error and a tighter confidence interval.

### b)

Now, when we adjust the confidence level to be lower, we observe that the number of SEs (standard errors) will be less than the current set value of 1.96. For example, a confidence level of 90% will result in 1.645, which is less than 1.96. Thus, having a smaller number out front will cause the margin of error to be smaller as well, corresponding to a tighter confidence interval.

## Question 3

```r
cdc <- read.csv("cdc.csv")
head(cdc)
```

```
##   state   genhlth physhlth exerany hlthplan smoke100 height weight wtdesire age
## 1    22      good        0       0        1        0     70    175      175  77
## 2    25      good       30       0        1        1     64    125      115  33
## 3     6      good        2       1        1        1     60    105      105  49
## 4     6      good        0       1        1        0     66    132      124  42
## 5    39 very good        0       0         1        0     61    150      130  55
## 6    42 very good        0       1         1        0     64    114      114  55
##   gender
## 1      m
## 2      f
## 3      f
## 4      f
## 5      f
## 6      f
```

### a)

```r
mean(cdc$weight[cdc$exerany == 1]) # mean weight of people who exercise
```

```
## [1] 169.0387
```

So, our null hypothesis is that people who do not exercise have the same weight as those of people that do exercise. In other words, $H_0$ : (average weight of people who do not exercise) = 169.0387. Alternative hypothesis is $H_a$ : (average weight of people who do not exercise) $\neq$ 160.0387.

### b)

```r
#perform a two-tailed t-test
t.test(cdc$weight[cdc$exerany == 0], cdc$weight[cdc$exerany == 1])
```

```
##
##  Welch Two Sample t-test
##
## data:  cdc$weight[cdc$exerany == 0] and cdc$weight[cdc$exerany == 1]
## t = 3.6842, df = 8024.9, p-value = 0.0002309
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.185482 3.881459
## sample estimates:
## mean of x mean of y
##  171.5722  169.0387
```

The value of the test statistic is 3.6842 here.

### c)

As seen in the above t.test, we have that the p-value is 0.0002309.

### d)

Since the p-value is less than our significance level of 5% (0.05), we have sufficient reasoning to reject the null hypothesis that the average weight of people who do not exercise do not differ from those of people that exercise.

**e)**

That is not correct. The p-value can be used to reject or fail to reject the null hypothesis, but this does not indicate the truth of the null (nor the alternative) hypothesis. What the p-value actually measures is how likely the observed differences between groups are according to chance. The lower this number is, the stronger evidence we have to reject the null since the observed differences are unlikely due to chance.

**f)**

The significance level indicates the likelihood that the events could have occurred due to chance. The lower this number is, we generally need stronger evidence to reject the "status quo" (the null) since we are setting a stricter bound on the events not occurring by chance.