Stats101A, Spring 2023 - Homework 2

Luke Villanueva - 206039397

04/14/23

```r
# library setup
library(tidyverse)
```

Problem 1

a.

confidence interval = [mean - margin of error, mean + margin of error]

margin of error = t crit val of confidence * standard of error

standard of error = standard deviation / sqrt(n)

And so, to find the interval,

```r
# standard error
SE <- 24757 / sqrt(30)

# margin of error
MOE <- abs(qt(0.05, df = 29)) * SE

# vector of the 95% confidence interval
c(23606 - MOE, 23606 + MOE)
```

```
## [1] 15925.96 31286.04
```

So, the 95% confidence interval is (15925.96, 31286.04).

b.

Because only the sample standard deviation was known, the t value was taken instead of the the z value.

c.

No, because the 95% confidence level is describing the probability of the **true** average of the population being in that interval. The average of the sample is completely different than the true population average.

d.

Two tailed test:

Null hypothesis: The mean income in the US in the year 2000 is $25,000.

Alternative hypothesis: The mean income in the US in the year 2000 is not $25,000.

Doing a t-test, we need to calculate the t score, get the critical value for 5% significance, and reject or fail to reject null hypothesis:

```
# t score, abs because it' two tailed, so sign doesn't matter
t_score <- abs((25000 - 23606) / (24757 / sqrt(30)))
t_score
```

```
## [1] 0.3084078
```

```
# t crit val 5% sig, two tailed, divide 0.05 by 2, sig val is then 0.025
crit <- abs(qt(0.025, df = 29))
crit
```

```
## [1] 2.04523
```

```
# should we reject null hypothesis?
t_score > crit
```

```
## [1] FALSE
```

The sample data does not provide enough information to disprove the claim that the mean is $25,000.

e.

```
# brute force, find the t score
# checking if t_score can be rejected
abs(qt(0.38, df = 29))
```

```
## [1] 0.3083767
```

```
t_score
```

```
## [1] 0.3084078
```

```
# check
t_score > abs(qt(0.38, df = 29))
```

```
## [1] TRUE
```

```r
# since two tailed test, multiply by 2
2*0.38
```

```
## [1] 0.76
```

This means the smallest significance value would be 76% significance, which is really big and does not help with pointing out any statistically significant results.

2.

a.

The interval's width would decrease if the sample size were to grow just so the confidence level would stay the same.

b.

The interval's width would increase if confidence level decreased because there are more samples that can fit in the requirements of the interval's boundaries.

3.

```r
cdc <- read.csv(file = "C:/Users/lavil/Downloads/cdc.csv", header = TRUE)
glimpse(cdc)
```

```
## Rows: 20,000
## Columns: 11
## $ state    <int> 22, 25, 6, 6, 39, 42, 6, 48, 6, 48, 9, 26, 35, 25, 34, 6, 6, ~
## $ genhlth  <chr> "good", "good", "good", "good", "very good", "very good", "ve~
## $ physhlth <int> 0, 30, 2, 0, 0, 0, 0, 1, 2, 3, 4, 30, 0, 0, 3, 0, 0, 30, 0, 0~
## $ exerany  <int> 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1~
## $ hlthplan <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1~
## $ smoke100 <int> 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0~
## $ height   <int> 70, 64, 60, 66, 61, 64, 71, 67, 65, 70, 69, 69, 66, 70, 69, 7~
## $ weight   <int> 175, 125, 105, 132, 150, 114, 194, 170, 150, 180, 186, 168, 1~
## $ wtdesire <int> 175, 115, 105, 124, 130, 114, 185, 160, 130, 170, 175, 148, 2~
## $ age      <int> 77, 33, 49, 42, 55, 55, 31, 45, 27, 44, 46, 62, 21, 69, 23, 7~
## $ gender   <chr> "m", "f", "f", "f", "f", "f", "m", "m", "f", "m", "m", "m", "~
```

a.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu 2 \neq 0$$

where μ_1 and μ_2 are means of desired weight

b.

```r
# let x be all the desired weights of those who exercise and y be desired weights of those who don't
t.test(x = cdc$wtdesire[cdc$exerany == 1], y = cdc$wtdesire[cdc$exerany == 0], alternative = "two.sided
```

```
##
##  Welch Two Sample t-test
##
## data:  cdc$wtdesire[cdc$exerany == 1] and cdc$wtdesire[cdc$exerany == 0]
## t = 5.5844, df = 8742.8, p-value = 2.415e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.889004 3.932456
## sample estimates:
## mean of x mean of y
##  155.8340  152.9233
```

The t stat is 5.5844, which implies the difference between the means is very unlikely.

c.

The p-value of the t test is 2.415 * 10^-8, which is basically 0% probability.

d.

Based on the p-value, the means are different enough to conclude that there is a true difference in the means between the desired weights of the people who exercise and the desired weights of the people who do not exercise.

e.

The p-value can only say if the null hypothesis is can be rejected or failed to be rejected. This means, there is no way to fully "prove" a null hypothesis. In reality, the p-value allows us to statistical say "yes, the null hypothesis cannot be likely to occur., so the alternative hypothesis is strongly supported" or "we do not have enough evidence to deny the chances of the null hypothesis to occur randomly or if something else is affecting it".

f.

Significance level is a very subjective term in the world of statistics. It essentially is a level of how we perceive if something is really rare to occur by chance or not. Some people may perceive a 5% chance of occurring to be really low, but to others, it may be higher. Regardless of the number, the level is just a set threshold we make so we can conclude either: "yes, the result of the test is too unlikely to happen just by chance" or "no, the result seems to be likely to occur, so there is no evidence for other factors to influence this outcome".