HW5.Zunqiu.Wang

Zunqiu Wang

9/28/2021

Q1 a

```r
# null hypothesis: average person is IQ 100
```

$$H_0 = \mu = 100$$

b

```r
# Alternative hypothesis: average person IQ larger than 100
```

$$H_a = \mu > 100$$

c

```r
# although sample size 100 greater then 30, i will still use t statistics here
library(reshape2)
desc.val.table <- function(x.bar, mu, sd, n, CI) {
  se <- sd/sqrt(n)
  t.stat <- (x.bar-mu)/se
  t.crit.perc.low <- (1-CI)/2 # % for lower bound 2 tail
  t.crit.perc.high <- (1-CI)/2 + CI # % for higher bound 2 tail
  t.crit.1.tail.low <- qt(1-CI, n-1) # t critical lower bound 1 left tail
  t.crit.1.tail.high <- qt(CI, n-1) # t critical higher bound 1 right tail
  t.crit.2.tail.low <- qt(t.crit.perc.low , n-1) # t critical for lower bound 2 tail
  t.crit.2.tail.high <- qt(t.crit.perc.high, n-1) # t critical for higher bound 2 tail
  CI.low.1.tail <- x.bar - t.crit.1.tail.high*se # CI lower bound for 1 right tail
  CI.high.1.tail <- x.bar + t.crit.1.tail.high*se # CI higher bound for 1 right tail
  CI.low.2.tail <- x.bar - t.crit.2.tail.high*se  # CI lower bound for 2 tail
  CI.high.2.tail <- x.bar + t.crit.2.tail.high*se # CI higher bound for 2 tail
  p.val.1.tail <- 1- pt(t.stat, n-1) # p val right tail
  p.val.2.tail <- 2*(1- pt(t.stat, n-1)) # p val two tail
  df <- data.frame(t.stat=t.stat, t.crit.1.tail.high=t.crit.1.tail.high,
                   t.crit.2.tail.low=t.crit.2.tail.low,
                   t.crit.2.tail.high=t.crit.2.tail.high, CI.low.1.tail=CI.low.1.tail,
                   CI.high.1.tail= CI.high.1.tail, CI.low.2.tail=CI.low.2.tail,
                   CI.high.2.tail=CI.high.2.tail, p.val.1.tail=p.val.1.tail,
                   p.val.2.tail=p.val.2.tail)
  df <- melt(df)
  return(df)
}
desc.val.table(104, 100, 22, 100, 0.95)
```

```
## No id variables; using all as measure variables

##                 variable        value
## 1                 t.stat   1.81818182
## 2  t.crit.1.tail.high   1.66039116
## 3   t.crit.2.tail.low  -1.98421695
## 4  t.crit.2.tail.high   1.98421695
## 5        CI.low.1.tail 100.34713946
## 6       CI.high.1.tail 107.65286054
## 7        CI.low.2.tail  99.63472271
## 8       CI.high.2.tail 108.36527729
## 9          p.val.1.tail   0.03603016
## 10         p.val.2.tail   0.07206032
```

$$t_{stat} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = 1.818182$$

d

```
#I used one tail t test since the alternative hyphothsis is larger than null.
```

e $\alpha = 0.05$ which is a common cut off in most hypothesis testing and it implies that if mean/null hypothesis is true there will still have the probability as extreme as 5% of chance of getting $\bar{x}$what we observed. $t_{crit} = 1.66$ is set to be threshold under 1 tail t test using $CI = 95\%$.

f yes, according to above calculation results of $t_{stat} = 1.818$ and $t_{crit} = 1.66$, $t_{stat} > t_{crit}$ so we reject null hypothesis under 1 tail t test at 95% CI. Alternatively, we can look at $p_{val} = 0.036$, which is smaller then 0.05.

g No, according to above calculation results of $t_{stat} = 1.818$ and $t_{crit} = 1.984$, $t_{stat} < t_{crit}$ so we fail to reject null hypothesis under 2 tail t test at 95% CI.

h For 1 tail test, $CI = \bar{x} \pm t * \frac{s}{\sqrt{n}} = 104 \pm 1.66 * \frac{22}{10}\ So\ 100.347 < CI < 107.653$

i $p_{val} = 1 - pt(t.stat, n-1) = 1 - pt(1.818, n-1) = 1 - 0.964 = 0.036$

2 a H_0:skill level measured by mean score between men and women is same H_a:skill level measured by mean score between men and women is different Men sample: $se_1 = \frac{s_1}{\sqrt{n_1}} = \frac{200}{\sqrt{50}} = 28.29$ women sample: $se_2 = \frac{s_2}{\sqrt{n_2}} = \frac{200}{\sqrt{50}} = 28.29$ $se_{diff} = \sqrt{se_1^2 + se_2^2} = \sqrt{28.29^2 + 28.29^2} = 40$ $t_{stat} = \frac{\bar{x}_{men} - \bar{x}_{wm}}{se_{diff}} = \frac{1124-1245}{40} = -3.025$ since n and s are same for men and women samples, $df = 2n - 2 = 98$

b Consider $\alpha = 0.05$, \$t_{crit}:

```r
qt(c(0.025, 0.975),10)
```

```
## [1] -2.228139  2.228139
```

Since t stat lies in rejection region so we reject null hypothesis and accept alternate hypothesis. We can conclude that with 95% confidence that there is a statistically significant difference between women and men in the Tetris skills.

3 a $H_0 =$ Driking the night before exam helps improve exam performance $H_a =$ Driking the night before exam does not help improve exam performance treatment: $se_1 = \frac{s_1}{\sqrt{n_1}} = 10/\sqrt{50} = 1.414$ control: $se_2 = \frac{s_2}{\sqrt{n_2}} = 5/\sqrt{50} = 0.707$ $se_{diff} = \sqrt{se_1^2 + se_2^2} = \sqrt{1.414^2 + 0.707^2} = 1.581$ $t_{stat} = \frac{\bar{x}_{trt} - \bar{x}_{con}}{se_{diff}} = \frac{78-75}{1.581} = 1.8975$ since n

is same but s not same for men and women samples, we use general formula: $df = \frac{se_{diff}^4}{se_a^4/(n_a-1)+se_b^4/(n_b-1)} = 1.581^4/(1.414^4/(50-1)+0.707^4/(50-1)) = 72.07$

```r
# t crit
qt(c(0.025,0.975), 72.07)
```

```
## [1] -1.99343  1.99343
```

$t_{crit} = 1.99343$ and $t_{stat} = 1.8975$ since $t_{stat} < t_{crit}$ so we fail to reject null hypothesis at 95% confidence interval and conclude that riking the night before exam does not help improve exam performance.

4 a

```r
set.seed(1234)
data <- rt(100,99)
t.test(data)
```

```
##
##  One Sample t-test
##
## data:  data
## t = 0.52648, df = 99, p-value = 0.5997
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1417391  0.2441201
## sample estimates:
##  mean of x
## 0.05119051
```

```r
# Here I simulated the data using t distribution with 100 numbers.
# The t test null hypothesis is that true mean is 0 and alternate is
# not equal to 0. With confidence set at 95%, it performs essential
# calculations and returns t stat, df, p value and 95% CI and sample
# mean estmate. It is clear that p value is 0.5597, which is larger
# than 0.05 threshold. We fail to reject null hypothesis that mean is
# equal to 0.
```

b

```r
library(ggplot2)
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
```

```r
# diamonds
str(diamonds)
```

```
## tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

```r
levels(as.factor(msleep$vore)) # check vore levels after converting to factors
```

```
## [1] "carni"   "herbi"   "insecti" "omni"
```

```r
sub.df <- diamonds[diamonds$cut %in% c("Ideal", "Good"), ] %>% select(cut, price)

ideal.df <- sub.df %>% filter(cut %in% "Ideal") %>% select(price)
good.df <- sub.df %>% filter(cut %in% "Good") %>% select(price)

t.test(ideal.df, good.df, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  ideal.df and good.df
## t = -8.0409, df = 7484.7, p-value = 1.029e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -586.2251 -356.4198
## sample estimates:
## mean of x mean of y
##  3457.542  3928.864
```

```r
# Upon the calculation results from unpaired t test,
# p val is less than 0.05 threshold, so we reject null hypothesis
```

c

```r
# sleep
# null hypothesis: difference in means is equal to 0
# alternate hypothesis: difference in means is not equal to 0
t.test(extra~group, data = sleep, paired = TRUE, alternative = "two.sided")
```

```
##
##  Paired t-test
```

```
##
## data:  extra by group
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

```
# Upon the calculation results from paired t test,
# p val is less than 0.05 threshold, so we reject null hypothesis
```

d

```
x.bar <- mean(data)
x.bar
```

```
## [1] 0.05119051
```

```
sd <- sd(data)
sd
```

```
## [1] 0.9723213
```

```
n <- length(data)
n
```

```
## [1] 100
```

```
se <- sd/sqrt(n)
se
```

```
## [1] 0.09723213
```

```
mu <- 0
mu
```

```
## [1] 0
```

```
t.stat <- (x.bar-mu)/se
t.stat
```

```
## [1] 0.5264773
```

```
df <- n-1
df
```

```
## [1] 99
```

```r
CI.lower <- x.bar + qt(0.025, n-1)*se
CI.lower
```

```
## [1] -0.1417391
```

```r
CI.higher <- x.bar + qt(0.975, n-1)*se
CI.higher
```

```
## [1] 0.2441201
```

```r
p.val <- 2*(1-pt(t.stat, n-1))
p.val
```

```
## [1] 0.5997342
```

```r
# It verifies the result from part a that we fail to reject null hypothesis.
```