Computer Project 3

Patrick Wong

12/07/2018

MATH 324 Computer Project 3

Exercise 1: Here, we look at how t critical values behave as their df (degrees of freedom) increases:

A) First, what is $z_{.05}$?

```
a <- qnorm(0.05) * (-1)
roundA <- round(a, 5)
paste0("Z(0.05): ", roundA)
```

```
## [1] "Z(0.05): 1.64485"
```

B) Second, if you look at $t_{.05,df}$ (t critical values for $\alpha = .05$) with df = 20, 40, 60, etc (continuing up by 20 each time), for what df does the t critical value first fall strictly within (e.g. $<$)

i. .05 of $z_{.05}$?

```
count1 <- 20
b1 <- qt(0.05, count1) * (- 1)
roundB1 <- round(b1, 5)

while (roundB1 > roundA){
  count1 = count1 + 20
  b1 <- (qt(0.05, count1) * (-1))
  roundB1 <- round(b1, 5)
}
paste0("DF: ", count1)
```

```
## [1] "DF: 1109780"
```

ii. .02 of $z_{.05}$?

```
aii <- qnorm(0.02) * (-1)
roundAii <- round(aii, 5)

count2 <- 20
b2 <- qt(0.02, count2) * (- 1)
roundB2 <- round(b2, 5)

while (roundB2 > roundAii){
  count2 = count2 + 20
  b2 <- (qt(0.02, count2) * (-1))
  roundB2 <- round(b2, 5)
}
paste0("DF: ", count2)
```

```
## [1] "DF: 439960"
```

iii. .01 of $z_{.05}$?

```r
aiii <- qnorm(0.01) * (-1)
roundAiii <- round(aiii, 5)


count3 <- 20
b3 <- qt(0.01, count3) * (- 1)
roundB3 <- round(b3, 5)

while (roundB3 > roundAiii){
  count3 = count3 + 20
  b3 <- (qt(0.01, count3) * (-1))
  roundB3 <- round(b3, 5)
}
paste0("DF: ", count3)
```

```
## [1] "DF: 523320"
```

C) What do you think the difference will be between $z_{.05}$ and $t_{.05,df}$ as df $\rightarrow \infty$?

By the data shown in Exercise 1B, it is displayed that as df approaches infinity, the $Z_{0.05}$ and $T_{0.05,df}$ will equal.

Exercise 2: A company with a large fleet of cars want to study the gasoline usage. They check the gasoline usage for 50 company trips chosen at random, finding a mean of 25.02 mpg and sample standard deviation is 4.83 mpg.

A) Which kind of confidence interval is appropriate to use here, z-interval or t-interval?

Since n = 50 > 40 It is a random sample test And the standard deviation is given

It is a z-interval.

B) What are the assumptions to check for the interval you chose?

Null hypothesis: $H_0 : \mu \geq 26$

Alternative hypothesis: $H_a : \mu < 26$

C) Please use R to find the critical value the company needs when constructing a (two-sided) 98% CI.

df = 50-1 df = 49

```r
crit2 <- qt(0.02/2, 49, lower.tail = FALSE) * (-1)
```

D) Please use R to construct a (two-sided) 98% CI for the mean of the general gasoline usage.

X $\pm Z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$

```r
mean <- 25.02
s <- 4.83
n <- 50
twoSided = function(mean, s, n){
error <- 2.33*(s/sqrt(50))  ###margin of error
upperLim <- mean + round(error,2)
lowerLim <- mean - round(error,2)
int = c(lowerLim, upperLim) ##print from low to high
return(int)
}
paste0("[", twoSided(mean, s, n), "]")
```

```
## [1] "[23.43]" "[26.61]"
```

E) Please use R to construct a 98% upper confidence bound for the mean of the gasoline usage.

```
upperConfd98 = function(mean, s, n){
error <- 2.05*(s/sqrt(50))  ###margin of error
upperLim <- mean + round(error,2)
lowerLim <- 25.02 - round(error,2)
int = c(lowerLim, upperLim) ##print from low to high
return(int)
}
paste0("[", upperConfd98(mean, s, n), "]")
```

```
## [1] "[23.62]" "[26.42]"
```

G) Apply the function you created in part (f) to demonstrate that larger sample size is required to achieve better accuracy (i.e, narrower CI width). Confidence level is fixed at 98%. Show at least three examples.

An first impression is to assume that the sample is a simple random sample and the information thatis gathered is usually provided in the problem. The sampling distribution is normally distributed.

Exercise 3: In a class survey, students are asked how many hours they sleep per night. In the sample of 22 students, the (sample) mean is 6.77 hours with a (sample) standard deviation of 1.472 hours. The parameter of interest is the mean number of hours slept per night in the population from which this sample was drawn, and the distribution of sleep for that population follows a normal distribution.

A) Which kind of confidence interval is appropriate to use here, z-interval or t-interval? Use T-interval because n = 22 < 30

B) What are the criteria to check in order to use the distribution you chose?

Check if it is a random sample. ; Check if n > 30 ; Check if the distribution is normal.

C) Please use R to find the critical value they need when constructing a 90% CI.

df = n-1 df = 21

```
crit3 <- round(qt(0.1/2, 21, lower.tail = FALSE), 4)
```

D) Please use R to find the 90% CI for the mean number of hours slept per night.

X $\pm Z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$

```
mean <- 6.77
s <- 1.473
n <- 22
slept = function(mean, s, n){
error <- crit3*(s/sqrt(50))  ###margin of error
lowerLim <- mean + round(error,2)
upperLim <- mean - round(error,2)
int = c(lowerLim, upperLim) ##print from low to high
return(int)
}
paste0("[", slept(mean, s, n), "]")
```

```
## [1] "[7.13]" "[6.41]"
```

Exercise 4: In the year 2001, the Youth Risk Behavior survey done by the U.S. Center for Disease Control reported that 747 out of 1168 female 12th graders said they always use seatbelts when driving. Let's construct a 95% confidence interval for the proportion of 12th grade females in the population who always use seatbelts when driving.

A) Use R to find the score CI for the proportion of 12th grade females in the population who always use seatbelts when driving.

Sample statistic: $\hat{p} = \frac{x}{n}$ Standard error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ X $\pm Z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$

```r
x <- 747
n <- 1168

p <- x/n
z <- 1.96

standard_error <- sqrt((p*(1-p))/n)
lowerLim <- round((p - 2*standard_error), 4)
upperLim <- round((p + 2*standard_error), 4)
paste0("CI: [", lowerLim , " ", upperLim, " ]")
```

```
## [1] "CI: [0.6115 0.6677 ]"
```

B) Assuming there is no prior information or past experience available, what is the sample size necessary to control the score 95% CI width to be within 0.01?

```r
error <- 0.01
control_score <- (z/error)^2
paste0("The sample size necessary to control the score is ", control_score)
```

```
## [1] "The sample size necessary to control the score is 38416"
```

C) How many times larger is the sample required in part b than the sample we have?

```r
multiple <- round((control_score/n), 2)
paste0("It is ", multiple,  " times larger than the original sample")
```

```
## [1] "It is 32.89 times larger than the original sample"
```

Exercise 5: Consider the problem in Exercise 4 again. The U.S. Center for Disease Control wants to conduct a test, with $\alpha = 0.05$, to see whether the proportion of 12th grade female seatbelt users is not 50%.

A) Write appropriate hypotheses. Use both the symbols and words. (Is the alternative hypothesis one-sided or two-sided? It should be clear in what you write.)

Null hypothesis: The proportion of female 12th grade seatbelt users is 50% Alternative hypothesis: The proportion of female 12th grade seatbelt users is not 50%

Null hypothesis: $H_0 : \mu = 50\%$

Alternative hypothesis: $H_a : \mu \neq 50\%$

B) What do you need to check to use your test? Please verify the conditions are met.

That the samples is randomly drawn. ; That n < 10 of the population ; The population proportion P and (1-P) > 10%

The conditions is satisfied, because samples is randomly drawn, 1168 < 10% of the population, and P(0.50) and (1-P) = 50%

C) Use R to compute the test statistic and obtain a p-value.

```
x <- 747 ###number of girls using seatbelts
n <- 1168 ###total amount of girls

p <- x/n
Popul_prop <- 0.50 ##population proportion
level_sign <- 0.05 ##level of significance
standard_error <- sqrt((p*(1-p))/n)

Z <- (p-Popul_prop/standard_error)
paste0("Test statistic: ", round(Z,2))
```

```
## [1] "Test statistic: -34.95"
```

```
pValue <- 2*(pnorm(Z, mean , 1) )
paste0("P-Value: ", pValue)
```

```
## [1] "P-Value: 0"
```

D) Make conclusions using the test statistic and p-value.

The test statistic is 9.57 and the P-value is 0. This is where we reject the null hypothesis(H_o : p = 0.5) because $\alpha < 0.05$.

E) There is a link between confidence intervals and p-values, but for now we will just answer: if you return to the confidence interval in 4a, is 50% included in the interval?

The population proportion value that is provided indicicates that the 50% is not included in the 95% confidence interval because of the sufficient evidence to believe that "the proportion of female 12th graders that uses seatbelts is not 50%"