Computer Project 3: CI and Hypothesis Testing

Math 324: Probability and Statistics with Computing

Due Thursday 12/13 by 11:59 pm, but +1 EC if submitted by 10 pm (submit online; no late submissions allowed; contributes to the 20% computer project grade)

Purpose

1. Learn how to write a function in R.
2. See the relationship between t distributions and the z distribution.
3. More practice computing confidence intervals.
4. More practice performing hypothesis testing.

Please submit the pdf generated by your RMD file. Show all R code and output.

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

a. First, what is $z_{.05}$?

```
z_p = qnorm(.05)
z_p
```

```
## [1] -1.644854
```

we can not determine the t ctitical value. because we dont not have any degrees of freedom. if we ignor the degrees of freedom then we have .05 probability assuming that the data in normally distributed.

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. $<$)

```
qt(.05,20)-z_p
```

```
## [1] -0.07986462
```

```
#the difference is small enough
qt(.05,40)-z_p
```

```
## [1] -0.03899739
```

```
qt(.05,60)-z_p
```

```
## [1] -0.02579524
```

```
qt(.05,80)-z_p
```

```
## [1] -0.01927095
```

```
i. .05 of $z_{.05}$?
ii. .02 of $z_{.05}$?
iii. .01 of $z_{.05}$?
```

c. What do you think the difference will be between $z_{.05}$ and $t_{.05,df}$ as df $\to \infty$? as alpha goes to infinity the t value approaches zero. the t value will get closer to the mean if it is normally distributed. pretty much get closer to 0.

Exercise 2. A company with a large fleet of cars wants 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? *Z-interval

b. What are the assumptions to check for the interval you chose? 1)n>40 n =- 50 2)random sample

3) standard diviation is know
4) distribution is normal

the sample size is greater equal to 50, this means that we can use the z-interval.

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

1) we get the significance level, which is equal to alpha

```
alpha = (1-.98)/2
z_value = qnorm(alpha)

#critical values
z_value =-z_value
z_value
```

```
## [1] 2.326348
```

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

```
x_mean = 25.02
X_standard = 4.83
n = 50
#confidence interval values
ciUpper = x_mean+z_value*(X_standard/sqrt(n))
ciLower = x_mean-z_value*(X_standard/sqrt(n))
ciUpper
```

```
## [1] 26.60905
```

```
ciLower
```

```
## [1] 23.43095
```

e. Please use R to construct a 98% upper confidence bound for the mean of the general gasoline usage. $upperbound[-infinity, .98]$

```
upper_boundZ= qnorm(.02, lower.tail = FALSE)
upper_boundZ
```

```
## [1] 2.053749
```

```
x_mean+upper_boundZ*(X_standard/sqrt(n))
```

```
## [1] 26.42284
```

f. Create a R function whose argument is the width of CI, and the output is the sample size necessary to achieve such accuracy. The confidence level is fixed at 98%.

```
#function provide from the book page 273
n_size <-function(width)
{
  return ((2*z_value*X_standard)/width)^2
```

```
}

difference = ciUpper-ciLower
difference
```

```
## [1] 3.178094
```

```
n_size(difference)
```

```
## [1] 7.071068
```

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.

```
#width is narrower
n_size(.3)
```

```
## [1] 74.9084
```

```
n_size(.04)
```

```
## [1] 561.813
```

```
n_size(.004)
```

```
## [1] 5618.13
```

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? t-interval

b. What are the criteria to check in order to use the distribution you chose?

1) Random sample
2) n<40
3) normally distributed

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

```
u = 8
x_mean = 6.77
x_s = 1.472
sample_s = 22

df1 = (sample_s-1)
#t values
t=qt(.05,df1, lower.tail = FALSE)
t
```

```
## [1] 1.720743
```

```
-t
```

```
## [1] -1.720743
```

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

```
#t CI
 x_mean+t*(x_s/sqrt(sample_s))
```

```
## [1] 7.310023
```

```
 x_mean-t*(x_s/sqrt(sample_s))
```

```
## [1] 6.229977
```

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.

```
p=747/1168
q = 1-p

z_pval = qnorm(.025,lower.tail = FALSE)
z_pval
```

```
## [1] 1.959964
```

```
Pm = ((p+(z_pval^2))/(2*1168))/((1+z_pval^2)/(1168))
Pm
```

```
## [1] 0.4627751
```

```
Pm + z_pval*(sqrt(((p*q)/1168)+((z_pval^2)/(4*(1168)^2)))/(1+(z_pval^2/1168)))
```

```
## [1] 0.4902688
```

```
Pm - z_pval*(sqrt(((p*q)/1168)+((z_pval^2)/(4*(1168)^2)))/(1+(z_pval^2/1168)))
```

```
## [1] 0.4352815
```

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?

```
w= .01

new_p=.5
q=1-new_p
n1=((z_pval^2)*(2*new_p*q-w^2)+sqrt(4*(z_pval^4)*(p*q)*(new_p*q-w^2)+(w^2)*z_pval^4))/(.01^2)
ceiling(n1)
```

```
## [1] 40926
```

```
n = (4*(z_pval^2)*(p)*(q))/(.01^2)
ceiling(n)
```

```
## [1] 49137
```

c. How many times larger is the sample required in part b than the sample we have? is much more larger

```
ceiling(n1-1168)
```

```
## [1] 39758
```

```
ceiling(n-1168)
```

```
## [1] 47969
```

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.) $H_0 = p = .5$ $H_a = p! = .5$ our defult null hyposthesis we say is 50% of the female students do wear the seatbelt and for alternative we say the more than or less than 50% do not wear the seatbelt.

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

1) Random sample
2) np>10 -> 1168*(.5)>10
3) n(1-p)>10 -> 1168*(1-.5)>10

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

```
p_hat = 747/1168
z = (p_hat-.50)/sqrt((.5*(1-.5))/1168)
z
```

```
## [1] 9.538853
```

```
newp=pnorm(z, lower.tail = FALSE)
newp
```

```
## [1] 7.221002e-22
```

```
newp*2
```

```
## [1] 1.4442e-21
```

d. Make conclusions using the test statistic and p-value. our p values is too low comapred to the significance. there we reject the hypothese

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? it does not include it because the interval that I got is lower than 50% percent .49 and .43