Chapter 7 - Inference for Numerical Data

Alexis Mekueko

10/16/2020

Github link: https://github.com/asmozo24/DATA606_Homework7 Web Link : file:///C:/Users/Petit%20Mandela/Documents/R/DATA606_Homework7/DATA606_Homework7/DATA606_Homework7.html

```
library(tidyverse) #loading all library needed for this assignment
library(openintro)
library(infer)
library(gplots)

library(httr)
library(rvest)
library(xml2)
#head(fastfood)
#library(readxl)
#library(data.table)
#library(readr)
#library(plyr)
#library(dplyr)
#library(dice)
# #library(VennDiagram)
# #library(help = "dice")
#library(DBI)
#library(dbplyr)

#library(rstudioapi)
#library(RJDBC)
#library(odbc)
#library(RSQLite)
#library(rvest)
#library(stringr)
#library(readtext)
#library(ggpubr)
#library(fitdistrplus)
#library(ggplot2)
#library(moments)
#library(qualityTools)
library(normalp)
#library(utils)
#library(MASS)
#library(qqplotr)
library(DATA606)
```

```
##
```

```
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

```
getLabs()
```

```
##  [1] "Lab1"  "Lab2"  "Lab3"  "Lab4"  "Lab5a" "Lab5b" "Lab6"  "Lab7"  "Lab8"
## [10] "Lab9"
```

```
library(knitr)
#library(StMoSim)
```

Working backwards, Part II. A 90% confidence interval for a population mean is (65, 77). The

population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

Since we only know there is 25 observations and not the elements that define these observation, then the sample mean is midpoint or the average of the population mean = (65 + 77)/2 = 71 , point estimate + *t* x standard error(SE) = mean sample* + t* x (standard deviation(s)/sqrt(n observation))

margin of error = mean sample - Lower_Ci = 71 - 65 = 6....compute t* or t-value ...The t-value is a cutoff we obtain based on the confidence level and the t-distribution with df degrees of freedom. That cutoff is found in the same way as with a normal distribution: we find t-value such that the fraction of the t-distribution with df degrees of freedom within a distance t-value of 0 matches the con

dence level of interest. sample standard deviation = (margin of error x sqrt(n observation)) / t* = 6*sqrt(25)/1.71 = 17.544

```
help(pt)
```

```
## starting httpd help server ... done
```

```
n <- 25
t_value <- abs(qt(0.10/2, df = n-1))
t_value
```

```
## [1] 1.710882
```

7.14 SAT scores. The standard deviation of SAT scores for students at a particular Ivy League college is

250 points. Two statistics students, Raina and Luke, want to estimate the average SAT score of students at this college as part of a class project. They want their margin of error to be no more than 25 points.

#(a) Raina wants to use a 90% confidence interval. How large a sample should she collect? # Answer: 271

#(b) Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina's, and explain your reasoning. # Answer: at 99% CI, Luke need a large sample, there is high probability with large sample size.

(c) Calculate the minimum required sample size for Luke.

Answer: with 99% CI, samp . . . minimum required sample size is 663

```
#(0.99) + (1-0.99)/2= 0.995
sd <- 250
margin_error <- 25
z_value <- qnorm(0.995, 0 , 1 )
n <- ((sd * z_value)/ margin_error)^2
```

7.20 High School and Beyond, Part I. The National Center of Education Statistics conducted a survey of

high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

(a) Is there a clear difference in the average reading and writing scores?

Answer: The plot on scores shows a clear difference even it is not very significant (in 50 range)

(b) Are the reading and writing scores of each student independent of each other?

Answer: there are independent since it a simple random sample of 200 students who did not receive special instructions. The problem statement does not give me the option to think otherwise.

(c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

Answer: H_null = there is no difference betweeen the average scores in reading and writing

```
H_alt = There is a difference betweeen the average scores in reading and writing
```

(d) Check the conditions required to complete this test.

sample is independent, sample size is large , sample is normal

(e) The average observed difference in scores is mean_read_write = -0.545, and the standard deviation of the

differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? # Answer: p_value = 0.1934182 > alpha_value (0.05) , thus I cannot reject the H_null. based on this test, there is no difference betweeen the average scores in reading and writing

(f) What type of error might we have made? Explain what the error means in the context of the application.

Answer: error of type II, or type II error, which means we failed to reject the H_null.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

Answer: We failed to reject the H_null , which mean there is no difference, if there is a chance that the actual difference is 0 , this statement still holds based on our test above.

```
mean <- -0.545
sd <- 8.887
n <- 200
SE <- sd/sqrt(n)
t <- mean/SE
p_value <- pt(t, n-1)
```

##7.28 Fuel efficiency of manual and automatic cars, Part II. (7.28, p.276) 98% CI,

```
   HWy MPG
  Automatic  Manual
```

Mean 22.92 27.88 SD 5.29 5.01 n 26 26

H_null: there is no difference between average highway mileage of automatic and manual cars H_alt: there is difference between average highway mileage of automatic and manual cars Lower_CI = -8.055794, upper_CI = -1.864206 < alpha (0.02), thus we reject the H_null. meaning , there is difference between average highway mileage of automatic and manual cars

```
mean_Auto <- 22.92
mean_Manual <- 27.88
sd_Auto <- 5.29
sd_Manual <- 5.01
# n_auto = n_Manual =n = 26
n <- 26
df <- n -1
t_value <- qt(0.98, df)
point_estimate <-  mean_Auto - mean_Manual
SE <- sqrt( (sd_Auto^2)/n + (sd_Manual^2)/n)

lower_CI <- point_estimate - t_value * SE
upper_CI <- point_estimate + t_value * SE
lower_CI
```

```
## [1] -8.055794
```

```
upper_CI
```

```
## [1] -1.864206
```

7.34 Email outreach efforts. A medical research group is recruiting people to complete short surveys

about their medical history. For example, one survey asks for information on a person's family history in regards to cancer. Another survey asks about what topics were discussed during the person's last visit to a hospital. So far, as people sign up, they complete an average of just 4 surveys, and the standard deviation of the number of surveys is about 2.2. The research group wants to try a new interface that they think will encourage new enrollees to complete more surveys, where they will randomize each enrollee to either get the new interface or the current interface. How many new enrollees do they need for each interface to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%?

#Answer: they need n = 304 new enrollees for each interface to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%

```
sd <- 2.2
mean <- 4
power_level <- 0.80
effect_size <- 0.5
z_score80 <- qnorm(0.80)
z_score5 <- qnorm(0.975)
n <- 2*((z_score80 + z_score5)^2) * ((sd)^2/effect_size^2)
n
```

```
## [1] 303.9086
```

Work hours and education. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents.47 Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis. Educational attainment

	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172

(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

Answer: H_null there is no difference in the average number of hours worked varies across the five groups. H_alt there is difference in the average number of hours worked varies across the five groups.

#(b) Check conditions and describe any assumptions you must make to proceed with the test. Answer: the sample is independent, sample size is large, ditribution seems normal.

(c) Below is part of the output associated with this test. Fill in the empty cells.

```
    Df    Sum sq        Mean sq        F-value       Pr(>F)
```

degree XXXXX XXXXX 501.54 XXXXX 0.0682 Residuals XXXXX 267,382 XXXXX
Total XXXXX XXXXX

df_deg = 4 , df_Resi = 1167 , df_total = 1171, sumsq_deg = 2006.16, sumsq_total = 269388.2 , meansq_Resi = 229.12, F_value = 2.2

#(d) What is the conclusion of the test? #Answer: p_value = 0.0682 > alpha (0.05) , so based on the 95% CI , we reject the H_null, thus, there is difference in the average number of hours worked varies across the five groups

```
k <- 5
n  <- 121+546+97+253+155
df_deg <- k-1
df_deg
```

```
## [1] 4
```

```
df_Resi <- n-k
df_Resi
```

```
## [1] 1167
```

```
df_total <- df_Resi + df_deg
df_total
```

```
## [1] 1171
```

```
meansq_deg <- 501.54

sumsq_Resi <- 267382
sumsq_deg <- df_deg*meansq_deg
sumsq_deg
```

```
## [1] 2006.16
```

```
sumsq_total <- sumsq_Resi + sumsq_deg
sumsq_total
```

```
## [1] 269388.2
```

```
meansq_Resi <- sumsq_Resi/ df_Resi
meansq_Resi
```

```
## [1] 229.1191
```

```r
F_value <- meansq_deg/meansq_Resi
F_value
```

```
## [1] 2.188992
```

```r
p_value <- pf(F_value, df_deg, df_Resi, lower.tail = FALSE)
p_value
```

```
## [1] 0.06819325
```