DATA 606 HW5

Ahsanul Choudhury

October 30, 2016

5.6 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 deviationis 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.

```
n <- 25
smean <- (65 + 77)/2
smean
```

```
## [1] 71
```

```
me <- (77-65)/2
df <- n-1
t<- qt(.95, df)

se <- round(me/t, 3)
se
```

```
## [1] 3.507
```

```
sd <- round(se*sqrt(n), 3)
sd
```

```
## [1] 17.535
```

Sample Mean = 71, Margin of of error= 3.507, standard deviation = 17.535

5.14 SAT scores. SAT scores of students at an Ivy League college are distributed with a standard deviation of 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?

We can use z score to find the sample size.

```
sd <- 250
me <- 25
z <- qnorm(0.95)
n <- round(((sd * z)/me)^2, 0)
n
```

```
## [1] 271
```

Sample size 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 yourreasoning.

For a higher confidence interval at 99% Luke will need a larger sample size.

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

```
z <- qnorm(0.995)
n <- round(((sd * z)/me)^2, 0)
n
```

```
## [1] 663
```

Luke's required sample size is 663

5.20 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 di???erences in scores are shown below.

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

No there is no clear diffrence in the average reading and writing scores, there are some diffrence visible in mean of boxplot but in the histogram the distribution seems quite normal

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

The reading and writing scores seem to be paired and not independent of each other for given student.

(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?

H_O: $\mu_r - \mu_w = 0$, Aaverage reading and writing scores has no diffrence.

H_A: $\mu_r - \mu_w \neq 0$, There is a difference in average reading and writing scores.

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

***The given sample size is large enough at 200 and are randomly selected, the boxplot indicates there is little skewness in the data so we can use t distribution.

(e) The average observed diference in scores is $\bar{x}_{read-write}$ = -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?

```r
n <- 200
xbar <- -0.545
sd <- 8.887
t <- xbar/(sd/sqrt(n))
round(pt(t, n-1), 3)
```

```
## [1] 0.193
```

p-value is above 0.05, we cannot reject null hypotheiss, average reading and writhing score has no difference.

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

We might have made a typeII error, if there is a diffrence we have failed to detect it because of a false negative result.

(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.

I would expect to a confidence interval for the average difference between the reading and writing scores to include 0, because there is no statisically significent evedience of diffrence in means.

5.32 Fuel effciency of manual and automatic cars, Part I. Each year the US EnvironmentalProtection Agency (EPA) releases fuel economy data on cars manufactured in that year.Below are summary statistics on fuel effciency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel effciency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.

```r
xbara <- 16.12
xbarm <- 19.85
sda <- 3.58
sdm <- 4.51
na <- 26
nm <- 26
se <- sqrt(sda^2/na + sdm^2/nm)
tstat <- (xbara-xbarm)/se

round(pt(tstat,na-1)*2, 4)
```

```
## [1] 0.0029
```

The p value is less than 0.05, there is strong evidence of a difference between the average fuel effciency of cars with manual and automatic transmissions in terms of their average city mileage.

5.48 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.

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

H_0: $\mu_{<hs} = \mu_{hs} = \mu_{jr} = \mu_b = \mu_g$.The mean hour worked is not diffrent. H_a: At least one of the means are different.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

The sample size is below 10% of United States population, observation is independent and we assume the data is truly random.

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

```
n <- 1172
k <- 5

dfg <- k - 1
dfe <- n-k
dfg
```

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

```
dfe
```

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

```
prf <- 0.0682
fstat <- qf(1-prf,dfg,dfe)
fstat
```

```
## [1] 2.188931
```

```
msg <- 501.54
mse <- msg/fstat
mse
```

```
## [1] 229.1255
```

```
ssg <- dfg*msg
sse <- dfe*mse
ssg
```

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

```
sse
```

```
## [1] 267389.5
```

```
sst <- ssg + sse
dft <- dfg + dfe
sst
```

```
## [1] 269395.6
```

```
dft
```

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

```
results <-  matrix(c("ANOVA","Df", "Sum Sq", "Mean Sq", "F value", "Pr(>F)","degree", dfg,ssg,501.54,fs

library(knitr)
kable(results)
```

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	4	2006.16	501.54	2.18893121413288	0.0682
Residuals	1167	267389.480409899	229.125518774549		
Total	1171	269395.640409899			

(d) What is the conclusion of the test?

The p-value is 0.0682, which is larger then 0.05, we therefore do not reject null hypthesis. The the mean of five groups are almost equal.