Chapter 7 - Inference for Numerical Data

Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Working backwards, Part II. (5.24, p. 203) 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.

```
bot<-65
top<-77

n<-25
df<-n-1
```

```r
#if distribution is nearly normal, then the mean will be top-bot/2

x_bar<-((top-bot)/2)+bot

T_score<-abs(qt(.05,df=df))

#algebra to get SE

SE<-(x_bar-bot)/T_score

#algebra to get sample sd()

sigma<-SE*sqrt(n)

answer<-tribble(
  ~"sample mean, (x_bar)",~"margin of error (T_score*SE)",~"sample standard deviation (s)",
  x_bar,T_score*SE,sigma
)
print(answer)
```

```
## # A tibble: 1 x 3
##   ‘sample mean, (x_bar~ ‘margin of error (T_score*~ ‘sample standard deviation ~
##                   <dbl>                       <dbl>                        <dbl>
## 1                    71                           6                         17.5
```

SAT scores. (7.14, p. 261) 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?

```
sigma<-250
#estimate the average SAT score of students at this college. margin of error not more than +-25.
Z_score<-1.65
sd<-25

n<-round((1.65*250*1/25)^2)+1

#90% confidence interval

cat("the preferred sample size is ",n)
```

```
## the preferred sample size is  273
```

(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.

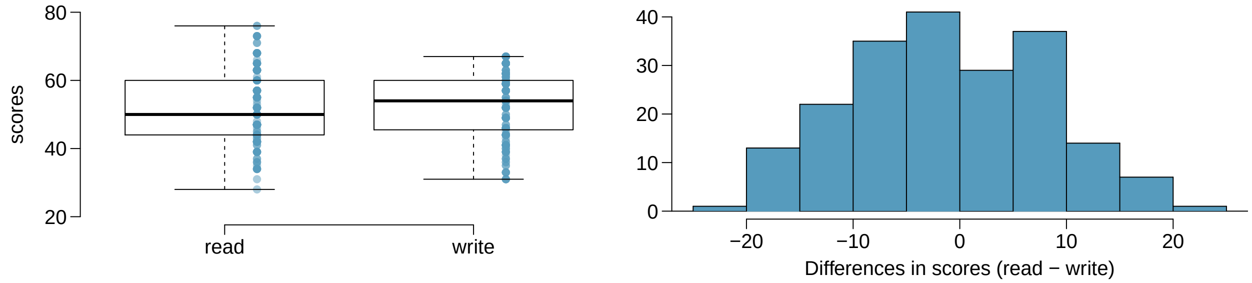
Because the Z_score is in the numerator, as it increases the sample size will increase. It also makes sense that the more accurate you want the test to be, the more data you need

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

```
Z_score<-qnorm(.99)
n<-round((Z_score*250*1/25)^2)+1

cat("the preferred sample size is ",n)
```

```
## the preferred sample size is  542
```

High School and Beyond, Part I. (7.20, p. 266) 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?

Not by visual inspection there isn't.

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

No they are not, each score is dependent on the student taking the test, so the student's two scores are dependent.

(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?

Hypothesis:

H_0: the mean difference between the student's two scores is zero (mean_diff)=0

H_A: There is a difference in the student's two scores

(mean_diff)!=0 (d) Check the conditions required to complete this test.

INDEPENDENCE:

The student's difference in score should be independent.

SAMPLE SIZE:

```r
nrow(as.data.frame(scores))
```

```
## [1] 400
```

400>30 so there is enough data to ignore skew.

(e) The average observed difference in scores is $\widehat{x}_{read-write} = -0.545$, 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?

```
x_bar<--.545
sigma<-8.887
n<-400
df<-n-1
SE_x<-sigma/sqrt(n)
null_mean<-0

T_score=abs((x_bar-null_mean)/SE_x)

p_value<-(1-pt(T_score,df=df))*2

cat("the p value for the mean difference in reading and writing scores is ",p_value,"which is greater th
```

```
## the p value for the mean difference in reading and writing scores is  0.2207299 which is greater than
```

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

If there is an error, we did not reject H_0 when H_A is true, so a type 2 error. (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.

Yes I do, because we did not reject the null hypothesis, so the confidence interval will contain 0 (the null hypothesis).

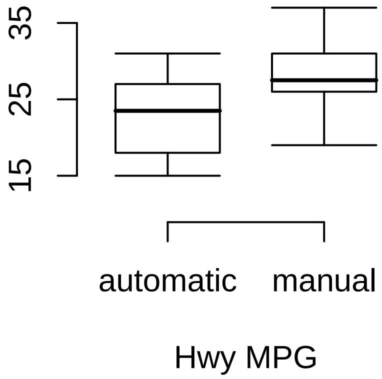
Fuel efficiency of manual and automatic cars, Part II. (7.28, p. 276) The table provides summary statistics on highway fuel economy of cars manufactured in 2012. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.

	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26

```r
point_estimates<-c("mean","sd","n")
automatic<-c(22.92,5.29,26)
manual<-c(27.88,5.01,26)
table<-data.frame(point_estimates,automatic,manual)

x_bar_diff<-table$automatic[1]-table$manual[1]

sigma_1<-table$automatic[2]
sigma_2<-table$manual[2]
n<-26
SE<-sqrt((sigma_1^2/n)+(sigma_2^2/n))
df<-n-1

T_score<-abs(qt(.01,df=df))

bot<-x_bar_diff- T_score*SE
top<-x_bar_diff+ T_score*SE

cat("the 98% confidence interval is ",bot,"to",top)
```

```
## the 98% confidence interval is  -8.510922 to -1.409078
```

There is a difference in the milage between an automatic and a manual transmission. This did not ask for a formal hypothesis test, but since zero is not in the 98% confidence interval we can say this is true.


35
25
15
automatic
manual
Hwy MPG

Email outreach efforts. (7.34, p. 284) 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%?

Sample: using normal survey method

control: normal survey method

experiment: new survey method

mean is 4 surveys

previously reported SD is 2.2 surverys

Test:

how likely is it that the mean of the null distribution will be shifted by .5 ?

```
null_mean<-4
null_sd<-2.2
n<-"?" #assuming its high
desired_effect<-.05


#assume 95% confidence interval to reject null

top<-4+ (1.96*null_sd)

#z score of getting new desired value
Z<-(4.5-4)/sd

(pnorm(Z))
```

```
## [1] 0.5079783
```

(this is power level is too low)

find the least amount of participants to raise power level.

work backwards from the desired power

```
p<-.8
(z_80<-qnorm(p))
```

```
## [1] 0.8416212
```

Therefore the center of the mean=4.5 is .84 sd from the cutoff region

also know top critical value is 1.96 sd from the center of the null distribution

means of the two distributions are (.842 +1.96)=2.8 sd from eachother

calculate standard error

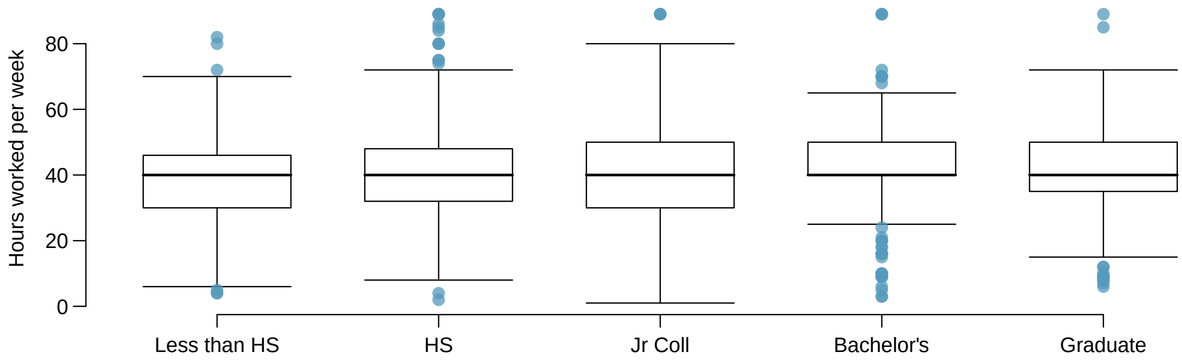
```
SE<-4.5/2.8

#standard deviation is known
sd<-2.2

n<-(2.2^2 +2.2^2)/SE^2

##OR
z_a<-1.96
z_b<-.84
sd<-2.2
d<-.5

#times 2 for two tails
n<-2*round(((z_a+z_b)^2*sd^2)/d^2)

cat("the smallest sample size to get 80% power from our test is ",n," in order to detect an effect size
```

```
## the smallest sample size to get 80% power from our test is  304  in order to detect an effect size o
```

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

```r
title<-c("mean","sd","n")
less_than_hs<-c(38.67,15.81,121)
hs<-          c(39.6,14.97,546)
jr_coll<-     c(41.39,18.1,97)
coll<-        c(42.55,13.62,253)
grad<-        c(40.85,15.51,155)
#total<-       c(40.45,15.17,1172)

table<-c(title, less_than_hs,hs,jr_coll,coll,grad)


library(psych)
```



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

H_0: the averages of hours does not vary across the five groups =0

H_A: the averages DO vary across groups !=0 (b) Check conditions and describe any assumptions you must make to proceed with the test.

1.

independence:

within groups: -the sampled observations are independent

-each sample is less than 10% of its population

approximate normality:

-the distributions look nearly normal on the boxplot

equal variance:

the groups look homoscedastic

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

```
#df
n_total<-1172
n_group<-5
df_g<-n_group-1
df_e<-n_total
df_tot<-(df_g+df_e)
df<-c(df_g,df_e,df_tot)

pf <- 0.0682
f_score <- qf( 1 - pf, df_g , df_e)

f_statistic<-c(f_score,NA,NA)

#get MSE from f_score=MSG/MSE

MSG<-501.24

MSE<-MSG/f_score

mean_square<-c(MSG,MSE,NA)
#get SSG from MSG=SSG/df_g

SSG<-MSG*df_g
SSE<-267382
sum<-SSG+SSE

sum_of_squares<-c(SSG,SSE,sum)

anova<-c("group","error","totals")

table1<-data.frame("anova"=anova,"Df"=df,"sum of squares"=sum_of_squares,"mean square"=mean_square,"f va

print(table1)
```

```
##    anova   Df sum.of.squares mean.square  f.value  P..F.
## 1  group    4        2004.96    501.2400 2.188904 0.0682
## 2  error 1172      267382.00    228.9913       NA     NA
## 3 totals 1176      269386.96          NA       NA     NA
```

(d) What is the conclusion of the test?