Chapter 7 - Inference for Numerical Data

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.

```
n <- 25
p <- .9
df <- n-1
upper <- 77
lower <- 65

mean <- (upper+lower)/2

error <- (upper - lower)/2

t_val <- qt( (p+(1-p)/2)  ,df)

sd = (error/t_val )* sqrt(n)


c("Error"= error, "Mean" =mean,"Standard Deviation"= sd)
```

```
##              Error               Mean Standard Deviation
##            6.00000           71.00000           17.53481
```

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?

The margin of error is equal to Z * SE. Given that we want a margin of error no larger than 25, we can re-arrange the formula and get $n = \left(\frac{z \cdot sd}{ME}\right)^2$ in other words, $n = \left(\frac{1.645 \cdot 250}{25}\right)^2$

```
((1.645*250)/25)^2
```

```
## [1] 270.6025
```

The sample size should be no larger than 270 students.

(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.

He should use a larger sample. As the confidence interval increases, we will need more values to properly represent the population.

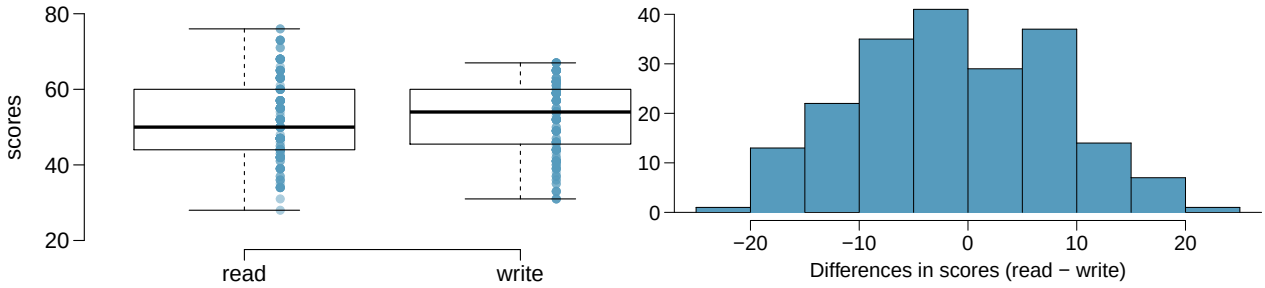
(c) Calculate the minimum required sample size for Luke.

```
((2.58*250)/25)^2
```

```
## [1] 665.64
```

We need at least 665 students.

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?

```
mosaic::favstats(hsb2$read)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##  min Q1 median Q3 max  mean       sd   n missing
##   28 44     50 60  76 52.23 10.25294 200       0
```

```
mosaic::favstats(hsb2$write)
```

```
##  min    Q1 median Q3 max   mean       sd   n missing
##   31 45.75     54 60  67 52.775 9.478586 200       0
```

While the median value is slightly different, the means for both values are relatively the same. There is no clear and obvious difference.

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

They are not independent of each other. Each value is paired as each student has both a reading and writing score. The students themselves are independent of each other.

(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?

H0: There is no difference in the average scores of students in the reading and writing exam.
Ha: There is a difference in the average scores of students in the reading and writing exam

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

We have a sufficiently large enough sample size, and the students used in this sample were randomly selected meaning they are all independent of one another. we can also surmise from the boxplot that the data is normally distributed.

(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?

```
se <- 8.887/sqrt(200)

t <- (-0.545 - 0)/se

df <- 200 - 1

pt(t,df , lower.tail = TRUE)
```

```
## [1] 0.1934182
```

Since our value is great than .1, we fail to reject the null hypothesis and conclude that there is no significant difference in the average scores of students in the reading and writing exam.

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

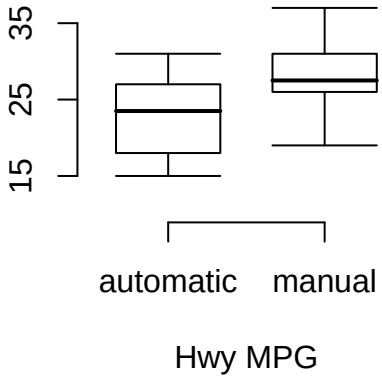
A type II error. In other words, there may actually be a significant difference in the scores for the reading and writing tests.

(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 we would. As we concluded that there is no significant difference, seeing 0 in the interval would further confirm this.

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




```
library(tidyverse)
```

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

## v tibble  3.1.0     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.0.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand()            masks Matrix::expand()
## x dplyr::filter()            masks stats::filter()
## x ggstance::geom_errorbarh() masks ggplot2::geom_errorbarh()
## x dplyr::lag()               masks stats::lag()
## x tidyr::pack()              masks Matrix::pack()
## x tidyr::unpack()            masks Matrix::unpack()
```

```
library(openintro)
library(infer)
```

```
## Warning: package 'infer' was built under R version 4.0.4
```

```
set.seed(2333)
fuel_eff %>%
  filter(transmission_desc %in% c("Automatic","Manual")) %>%
```

```
  group_by(transmission_desc) %>%
  slice_sample(n=26) %>%
   specify(hwy_mpg~transmission_desc ) %>%

  calculate(stat = "diff in means", order = c("Automatic","Manual"))  %>%
  get_ci(level = 0.98)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -5.65    -5.65
```

It is apparent that our confidence interval does not overlap 0, therefore can conclude that the difference in highway fuel efficiency for automatic and manual shifts is significantly different.

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%?

We can use the same formula from above, $n = \left(\frac{z \cdot sd}{ME}\right)^2$ with a t value of 1.2816 and a mean error of .5.

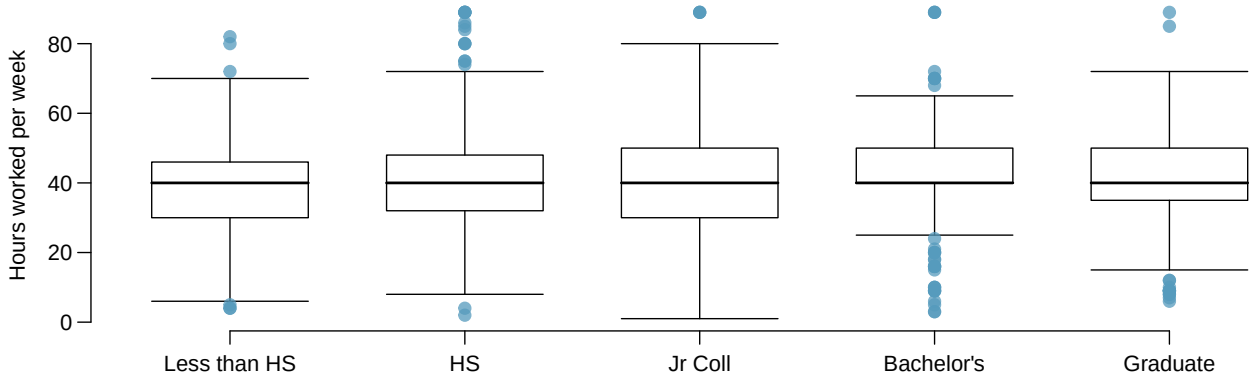
```
((1.2816 * 2.2)/.5)^2
```

```
## [1] 31.79877
```

Therefore, we need a sample of around 32.

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.

HO: There is no difference across the educational attainment levels, they are all the same.

HAL: The hours worked across educational attainment levels are different.

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

Our data is independent across groups and observations.The data in each group appears to be mostly normally distributed, with bachelors being the only exception. we also see roughly the same median value for each group.

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

```
n <- 1172

df <- 5 -1

df_r <-  n-5

df_t <- df + df_r

msq <- 501.54

ssq <- msq*df
```

```r
ssq_r <- 267382

ssq_t <- ssq_r+ssq


msq_r <- ssq_r/df_r

f<-msq/msq_r

cat(df,df_r,df_t,ssq,ssq_t,msq_r,f)
```

```
## 4 1167 1171 2006.16 269388.2 229.1191 2.188992
```

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree	4	2006.16	501.54	2.188992	0.0682
Residuals	1167	267,382	229.1191		
Total	1171	269388.2			

(d) What is the conclusion of the test?

Since we have a p value of 0.0682, we fail to reject the null hypothesis and conclude that there is no difference in hours worked per week across the educational attainment levels, they are all the same.