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.

The mean is 71 since its halfway between 65 and 77.

The margin of error is 6 (the CI is ± 6).

To find the standard deviation, solve the following

$$6 \;=\; 1.645 * \frac{s}{\sqrt{25}}$$

```
n<-25
moe<-6
z_score<-qnorm(.95)        #  The z score of 90% (1.90/2) is 1.645

s<-sqrt(n)*moe/z_score

sprintf("The standard deviation is %.2f", s)
```

```
## [1] "The standard deviation is 18.24"
```

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?

To find the minimum sample size, solve the following

$$25 \ > \ 1.645 * \frac{250}{\sqrt{n}}$$

```
moe<-25
sigma<-250
z_score<-qnorm(.95)

n<-(z_score*sigma/moe)^2

n<-ceiling(n)      # raise n to the next integer

sprintf("Raina and Luke need to sample at least %d students. ", n)
```

```
## [1] "Raina and Luke need to sample at least 271 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.

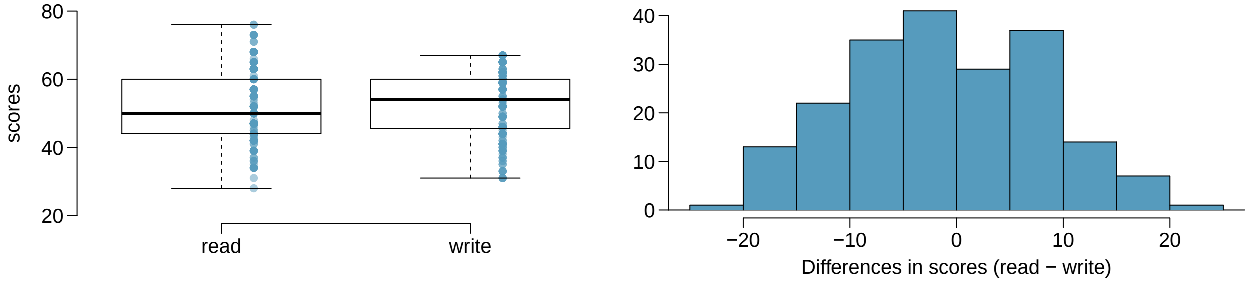
If he wants to keep the same Margin of Error, he will need to increase the sample size to satisfy the higher confidence level.

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

```
z_score<-qnorm(.995)

n<-(z_score*sigma/moe)^2

n<-ceiling(n)      # raise n to the next integer

sprintf("Luke now needs to sample at least %d students. ", n)
```

```
## [1] "Luke now needs to sample at least 664 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?

I can see most students dont have much difference. The frequency curve of diffs is fairly normalized. Maybe the writing scores are a bit higher.

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

No, I tend to think they would correlate although Im not sure how much. They are similar aptitudes.

Plus sudents with a high level of focus and determination will perform better across all disciplines.

(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?

$$\begin{aligned} H_0 : &\quad \bar{R} = \bar{W} \\ H_A : &\quad \bar{R} <> \bar{W} \end{aligned}$$

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

The scores of each student are independent.

The sample size is 200 so that is sufficient.

(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?

$$T = \frac{\bar{x}_{diff}}{SE_{\bar{x}_{diff}}}$$

```r
xdiff<- -0.545
s<-8.887

se<-s/sqrt(200)
df<-199
t_score<-xdiff/se
prob_diff<-pt(t_score, df=df)

sprintf("The T Score probability of the difference/standard deviation is %.2f ", prob_diff)
```

```
## [1] "The T Score probability of the difference/standard deviation is 0.19 "
```

It is within our CI so we cant reject the idea that there is no difference.

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

A Type 2 Error is when you fail to reject H_0 even though H_A is valid.

I would think Type 2 Errors are common in the interest of being careful.

So increased sampling might result in a different conclusion.

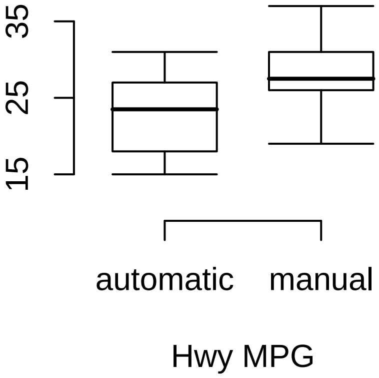
(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

Yes it appears to me that the average difference center at 0

i.e. if you kept sampling you may find half reading scores are better, and the other half writing scores are better.

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



The Standard Error of 2 Groups combined is

$$SE \ = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

```
x_diff<-22.92-27.88
s1<-5.29
s2<-5.01
n1<-26
n2<-26


se<-sqrt((s1^2/n1)+(s2^2/n2))
z_score<-qnorm(.99)        #  The z score of 98% (1.98/2) is 2.32
moe<-z_score*se

lcb<-x_diff-moe
ucb<-x_diff+moe

sprintf("The confidence interval is between %.2f and %.2f", lcb,ucb)
```

```
## [1] "The confidence interval is between -8.28 and -1.64"
```

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%?

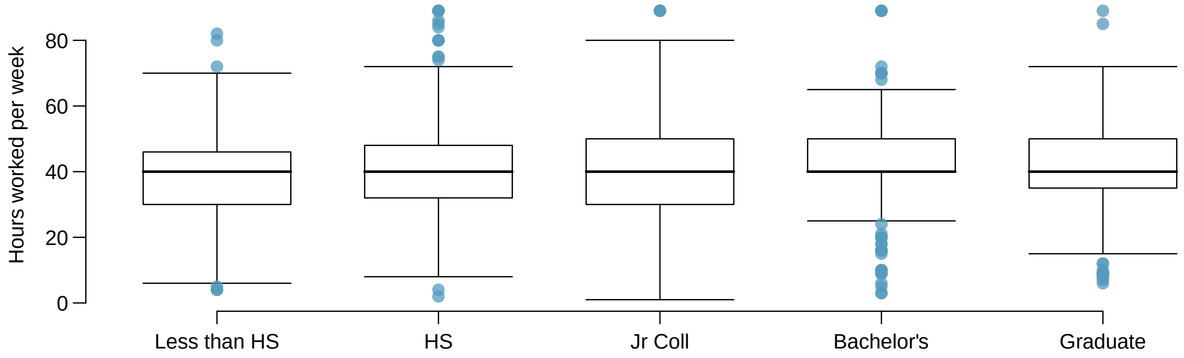
$$0.5 > Z_{80\%} \; * \; \frac{\sigma}{\sqrt{n}}$$

```
# surveys taken...
mu<-4        # not used

sigma<-2.2
desired_moe<-.5
z_score<-qnorm(.9)        #  The z score of 80% (1.8/2) is 1.28

n<-ceiling((z_score*sigma/desired_moe)^2)

sprintf("They need to observe at least %d enrollees", n)
```

```
## [1] "They need to observe at least 32 enrollees"
```

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.

$$H_0 : \hat{x}_{<HS} = \bar{x}_{HS} = \bar{x}_{JrCol} = \bar{x}_{Bchl} = \bar{x}_{Grad}$$

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

```
gss_sub %>%
  group_by(degree) %>%
  summarise(observations = n(),mean_hours = mean(hrs1))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 5 x 3
##   degree       observations mean_hours
##   <fct>               <int>      <dbl>
## 1 Less than HS          121       38.7
## 2 HS                    546       39.6
## 3 Jr Coll                97       41.4
## 4 Bachelor's            253       42.5
## 5 Graduate              155       40.8
```

When dealing with a comparison between groups, the normality check should include the independence **extended** *check*

I dont see any issue with dependent data between or within groups so Ill assume independence.

Sample sizes are over 30 with no extreme outliers so normality looks ok.

```
res.aov <- aov(hrs1~degree, data = gss_sub)
summary(res.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## degree         4   2006   501.5   2.189 0.0682 .
## Residuals   1167 267382   229.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
             # Df Sum Sq Mean Sq F value Pr(>F)
#degree         4   2006   501.5   2.189 0.0682 .
# Residuals   1167 267382   229.1
# F-statistic: 2.189 on 4 and 1167 DF,  p-value: 0.06819
```

(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	4	2006	501.54	2.189	0.0682
Residuals	1167	267,382	229.1		
Total	1171	269388			

(d) What is the conclusion of the test?

With only 5 groups to analyze, the distribution allows Degree of Freedom of just 4

meaning we make it more likely for random observations to fall outside the confidence interval

Still the p-value was .06 and the little dot next to the p-value in the summary tells me the null hypothesis is not rejected.

It does appear that higher education levels influence longer working hours but the evidence is not that strong.