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.

Since the sample mean is the midpoint of the lower and upper values of the confidence interval, we can calculate by:

```
sample_mean <- (65 + 77) / 2
paste0('The sample mean is: ', sample_mean)
```

```
## [1] "The sample mean is: 71"
```

To calculate the margin of error, we subtract the upper and lower bounds of the confidence interval from the sample mean:

```
ME <- (77 - 65) / 2
paste0('The margin of error is: ', ME)
```

```
## [1] "The margin of error is: 6"
```

To find the sample standard deviation, we'll use the following formula $ME = t^* \times SE$.

```
df <- 25 - 1
p <- 0.9
p_2tailed <- p + (1 - p)/2
t_value <- qt(p_2tailed, df)

# to solve for SE, we'll need to alter the above equation
SE <- ME / t_value

# then, with the formula for standard deviation
sd <- SE * sqrt(25)
paste0('The sample standard deviation is: ', round(sd, digits = 2))
```

```
## [1] "The sample standard deviation is: 17.53"
```

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?

In order to find the size of the sample needed to reach a 90% confidence interval:

- **Since** $ME = z \times SE$**, and** $SE = \frac{\sigma}{\sqrt{n}}$**, we have** $ME = z \times \frac{\sigma}{\sqrt{n}}$
- **Then, by solving for n, we have** $\frac{z \times \sigma}{ME} = \sqrt{n}$**, or** $(\frac{z \times \sigma}{ME})^2 = n$

```
# given that we have a 90% confidence interval
z_2 <- 1.65
ME_2 <- 25
sd_2 <- 250

n_2 <- ((z_2 * 250) / (ME_2))^2

paste0('The sample should be at least ', ceiling(n_2), ' students.')
```

```
## [1] "The sample should be at least 273 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.

Without doing the calculation, I would expect the sample to be larger than Raina's since the z-score would be larger, it would make the numerator in part (a) larger, making n larger.

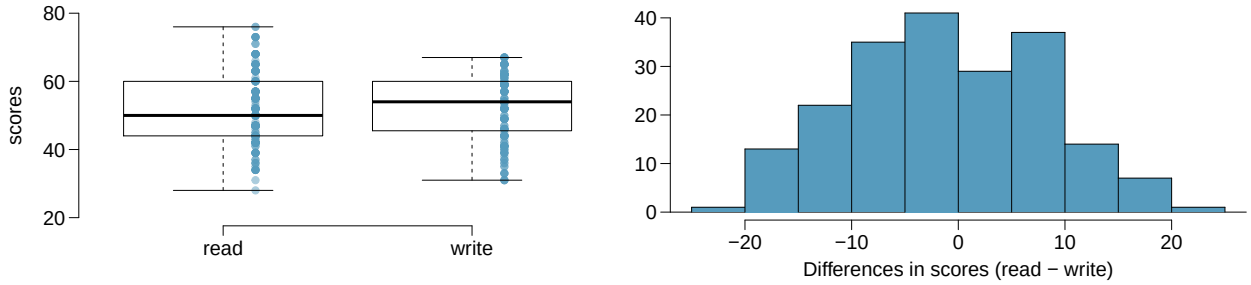
(c) Calculate the minimum required sample size for Luke.

Using the same formula as part (a):

```
# given that we have a 99% confidence interval
z_luke <- 2.575

n_luke <- ((z_luke * 250) / (ME_2))^2

paste0('The sample should be at least ', ceiling(n_luke), ' students.')
```

```
## [1] "The sample should be 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?

There doesn't appear to be a clear difference in the average reading and writing scores, but it does look like writing scores seem to trend higher than reading scores. However, the distribution of the differences in scores appears to be approaching normal, with a center around 0, which would indicate that there isn't a clear difference in scores.

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

The reading and writing scores of each student are not independent of each other since cognitive ability (reading/writing 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?

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

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

We would need to check the independence and normality conditions. Since the observations are based on a simple random sample, they are independent from one another. Additionally, we have a sample n of 200, and there do not appear to be any extreme outliers, so the normality of $\overline{x}$ is satisfied. Since both of these conditions are satisfied, we can move forward and complete the test.

(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?

To perform this calculation, we'll have to use the following formulas:

- $SE_{\overline{x}diff} = \frac{s_{diff}}{\sqrt{n_{diff}}}$
- $T = \frac{\overline{x}_{diff} - 0}{SE_{\overline{x}diff}}$

```r
n_stud <- 200
df_stud <- n_stud - 1

mean_difference <- -0.545
sd_stud <- 8.887
se_stud <- sd_stud / sqrt(n_stud)
t_score <- (mean_difference - 0) / se_stud
p_stud<- pt(t_score, df_stud)

paste0('The p-value is ', round(p_stud, digits = 4))
```

```
## [1] "The p-value is 0.1934"
```

Since the p-value is greater than 0.05, we do not reject the null hypothesis. These data do not provide convincing evidence of a difference between the average scores on the two exams.

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

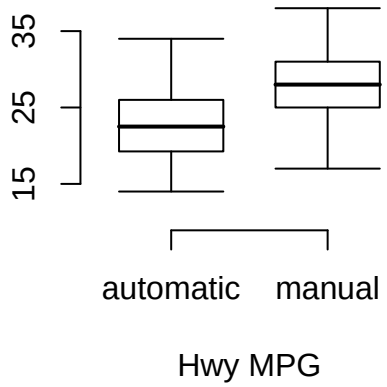
Because we fail to reject the hypothesis, we could be making a Type II error. In the context of the application, we would be stating that there is no difference between the average scores on the two exams when there actually is a difference between the average scores on the two exams.

(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.

Based on the results of the hypothesis test, I would expect a confidence interval for the average difference between the reading and writing scores to include 0. With a difference of 0 in the confidence interval, we are stating that there is strong likelihood that there isn't a difference between the average scores on the two exams.

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



- **H0: There is no difference between average highway mileage of manual and automatic cars.**
- **HA: There is a difference between average highway mileage of manual and automatic cars.**

First, we'll find the point estimate of the difference in the means between automatic and manual transmissions:

```
mean_diff <- 27.88 - 22.92
paste0('The mean difference is ', round(mean_diff, digits = 2))
```

```
## [1] "The mean difference is 4.96"
```

Then, in order to see if we can model this difference using a t-distribution, we'll need to make sure of the following:

- **Independence, extended: The data are independent within and between the two groups (e.g. the data come from independent random samples) – this is true.**
- **Normality: check the outliers rull of thumb for each group separately. – there do not appear to be any clear outliers, therefore this condition is also met. We can proceed.**

Next, we'll calculate the standard error using the sample standard deviations with this formula:

$SE = \sqrt{\frac{s_{aut}^2}{n_{aut}} + \frac{s_{man}^2}{n_{man}}}$

```
sd_man <- 5.01
sd_aut <- 5.29
n_samp <- 26

SE_fuel <- sqrt(((sd_man^2)/n_samp)) + ((sd_aut^2)/(n_samp))
paste0('The standard error is ', round(SE_fuel, digits = 2))
```

```
## [1] "The standard error is 2.06"
```

With $df = 25$, we can now calculate the confidence interval after determining the critical value of t_{25}^*, which is:

```
crit_t <- abs(qt(0.02, 25))
paste0('The critical value of t is ', round(crit_t, digits = 2))
```

```
## [1] "The critical value of t is 2.17"
```

And the confidence interval:

```
ci_upper <- mean_diff + crit_t * SE_fuel
ci_lower <- mean_diff - crit_t * SE_fuel

paste0('A 98% confidence interval is (', round(ci_lower, digits = 2),
       ', ', round(ci_upper, digits = 2), ').')
```

```
## [1] "A 98% confidence interval is (0.5, 9.42)."
```

To interpret this, we can reject the null hypothesis that there is no difference between the mean difference between average highway mileage of manual and automatic cars since the confidence interval does not bound to 0.

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%?

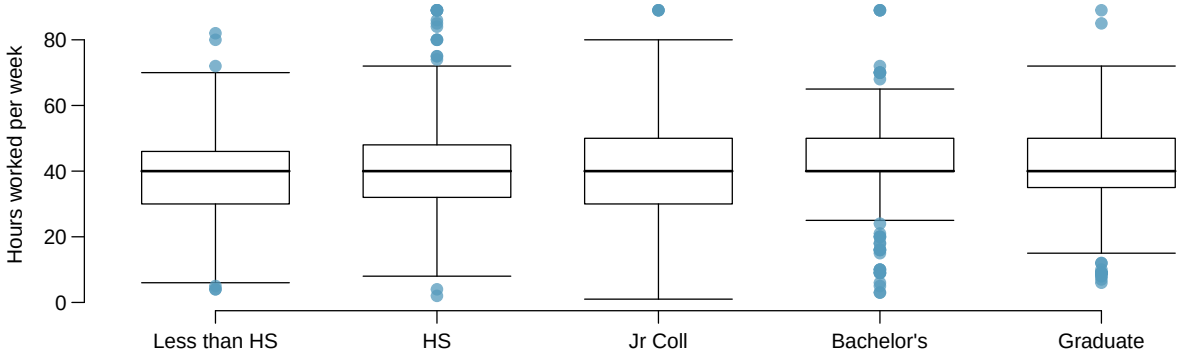
Since we have a significance level $\alpha = 0.05$, and a targeted power of 80% ($\beta = 0.80$), we'll need to do the following calculations:

```
# standard z-score for 95% confidence interval
z_05 <- 1.96

# standard z-score of lower tail of 80%
z_80 <- 0.84

effect_size <- 0.5
sd_surveys = 2.2
sample_size_needed <- (sd_surveys^2 / effect_size^2) * ((z_05 + z_80)^2)

paste0('The sample size needed is at least ', ceiling(sample_size_needed),
' new enrollees per interface.')
```

```
## [1] "The sample size needed is at least 152 new enrollees per interface."
```

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.

The hypotheses for this ANOVA test are:

- **H0: The average number of hours worked per week is not different across the five education groups.**
- **HA: The average number of hours worked per week is different across the five education groups.**

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

- **The observations are independent within and across groups – this is true, given that an individual cannot be in two different education groups, the observations will not overlap in the groups.**
- **The data within each group are nearly normal – this is true, looking at the box plots above, each seem to have a well defined center, and they seem to approach normalilty (they aren't skewed), and the sample sizes for all groups are quite large (greater than 30).**
- **The variability across the groups is about equal – this is also true when observing the standard deviations across the groups, they all seem to be similar (one group does not have a much higher standard deviation than others).**

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

```r
residuals_aov <- aov(hrs1 ~ degree, data = gss_sub)
summary(residuals_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	501.54	2.189	0.0682
Residuals	1167	267,382	229.1		
Total	1171	269,388			

(d) What is the conclusion of the test?

Because we have a p-value of 0.0682, which is greater than 0.05, we accept the null hypothesis that the average number of hours worked per week is not different across the five education groups.