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.

Sample Mean is 71

```
# sample mean
n <- 25
x1 <- 65
x2 <- 77

smean <- (x2 + x1) / 2
smean
```

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

Margin of Error is 6

```
# margin of error
n <- 25
x1 <- 65
x2 <- 77

me <- (x2 - x1) / 2
me
```

```
## [1] 6
```

Sample Standard Deviation is 17.54

```
# degrees of freedom
df <- 25 -1
df
```

```
## [1] 24
```

```
# T-value
t <-  1.71

# standard error
se <- (77 - 71) / 1.71
se
```

```
## [1] 3.508772
```

```r
#sample standard deviation
sample_sd <- se * sqrt(n)
sample_sd
```

```
## [1] 17.54386
```

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?

Raina should collect a sample of at least 272 to use a 90% confidence interval.

```
sd <- 250
me <- 25
z_90 <- 1.65

# sample size
n <- ((sd * z_90) / me)^2
n
```

```
## [1] 272.25
```

(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.

Luke's sample size should be larger than Raina's sample size at a 99% confidence interval. Since the confidence interval is narrower he needs to collect a larger sample to represent the population.

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

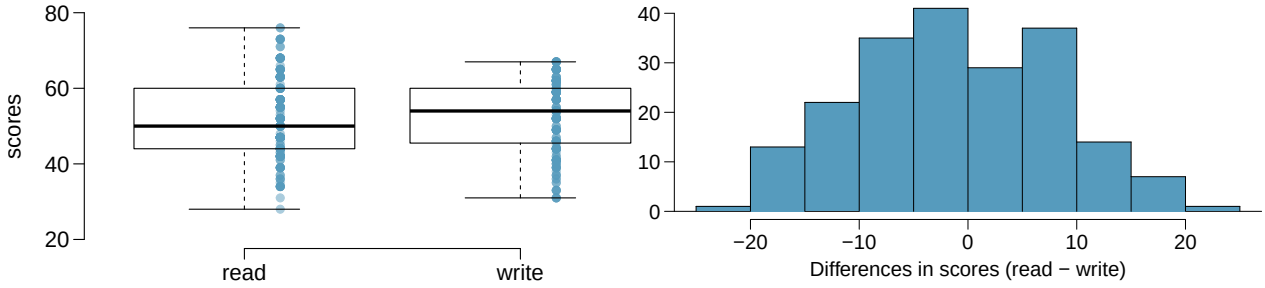
Luke's required sample size is at least 666

```
sd <- 250
me <- 25
z_99 <- 2.58

# sample size
n <- ((sd * z_99) / me)^2
n
```

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

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 seems to be a bit of a difference within the box plots. The mean for both box plots seems to be a bit off from each other. In the histogram, the distribution appears to be normal.

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

Being that it is a random sample of 200 students, it is reasonable to assume that reading and writing score of each student is 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?

Ho: There is no difference in the reading and writing scores

Ha: There is a difference in the reading and writing scores

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

The sample is random with a size of 200 which is 10% of the population. The reading and writing scores appear to be independent from each other and the distribution also appears to be normal.

(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?

P-value is 0.19 and not less than 0.05, therefore we can reject the alternative hypothesis. There is no difference in the reading and writing scores.

```
diff_sd <- 8.887
diff_mu <- -0.545
n <- 200

# standard error
diff_se <- diff_sd / sqrt(n)
diff_se
```

```
## [1] 0.6284058
```

```r
# T-Value
t_value <- (diff_mu - 0) / diff_se
t_value
```

```
## [1] -0.867274
```

```r
# degrees of freedom
df <- n - 1

#P-value
p_value <-  pt(t_value, df, lower.tail = TRUE)
p_value
```

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

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

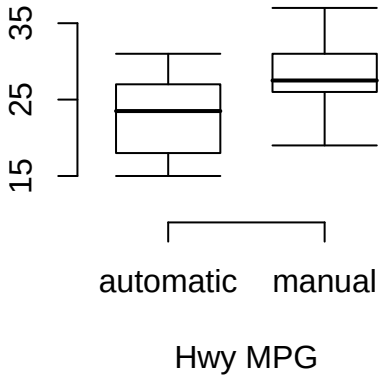
We would have made a Type II Error, incorrectly rejecting the alternative hypothesis. This would mean there is a difference in the average scores for reading and writing and we failed to identify it.

(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 would expect a confidence interval for the average difference between reading and writing scores to include 0 because it would indicate that the difference is not on either side.

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



Ho: There is no difference in average miles between manual and automatic cars.

Ha: There is a difference in average miles between manual and automatic cars.

The P-value is 0.011, being less than 0.05, therefore we can reject the null hypothesis

```
n <- 26

# Automatic
mean_auto <- 22.92
sd_auto <- 5.29

# Manual
mean_man <- 27.88
sd_man <- 5.01

# difference in sample means
mean_diff <- mean_auto - mean_man
mean_diff
```

```
## [1] -4.96
```

```
# standard deviation
se_diff <- ((sd_auto^2 / n) + (sd_man^2 / n))
se_diff
```

```
## [1] 2.0417
```

```r
# T-Value
t_value <- (mean_diff - 0) / se_diff

df <- n - 1

# P-Value
p_value <- pt(t_value, df, lower.tail = TRUE)
p_value
```

```
## [1] 0.01132343
```

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%?

There need to be at least 32 enrollee's each to detect an effect size of 0.5 surverys per enrolle at a 80% confidence interval.

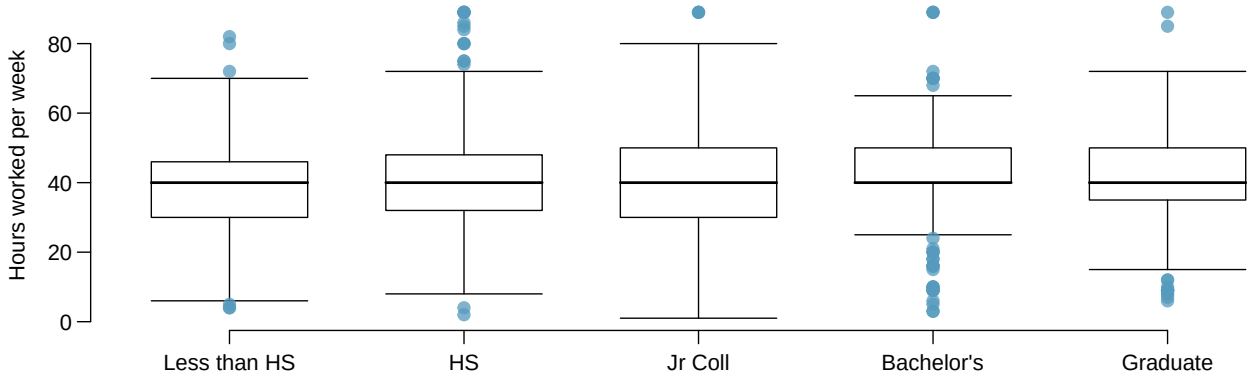
```
z_80 <- 1.28
sd <- 2.2
me <- 0.5

# sample size
n_4 <- ((sd * z_80) / me)^2
n_4
```

```
## [1] 31.71942
```

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 in the mean number of hours worked across the population.

Ha: There is a difference in the mean number of hours worked across the population

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

Based on the text the observations appear to be independent across the groups. The data within each group is nearly normal and by looking at the standard deviation there variability within the groups is almost equal.

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

ANOVA

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree	4	2004.1	501.54	2.19	0.0682
Residuals	1167	267,382	229.11		
Total	1171	269,377.73			

(d) What is the conclusion of the test?

Being that the p-value is 0.0684 it is higher than 0.05 and we reject our alternative hypothesis.