Chapter 7 - Inference for Numerical Data

Bryan Persaud

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.

```
(65 + 77) / 2
```

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

The sample mean is 71.

```
(77 - 65) / 2
```

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

The margin of error is 6.

```
# Calculate T-score
df <- 25 -1
p <- 0.90
x <- p + (1 - p) / 2
T_score <- qt(x, df)
SE <- 6 / T_score
Standard_deviation <- SE * sqrt(25)
Standard_deviation
```

```
## [1] 17.53481
```

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

(a)

```
Z_score <- 1.645
Standard_deviation_2 <- 250
Margin_error <- 25
Sample_size <- ((Z_score * Standard_deviation_2) / Margin_error)^ 2
Sample_size
```

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

If Raina wants to use a 90% confidence interval, she would need to collect a sample of 271.

(b)

If Luke wants to use a 99% confidence interval, he would need a larger sample size than Raina because he would be using a higher confidence interval, which means a higher z-score. So a larger sample is required to be more accurate.

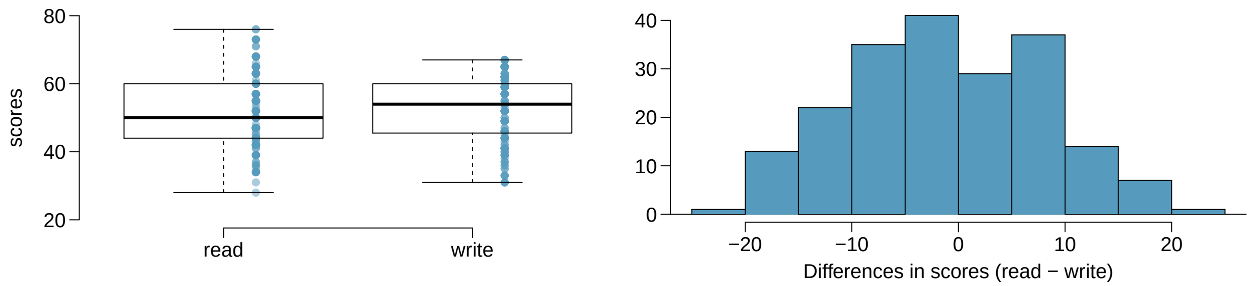
(c)

```
Z_score_99 <- 2.576
Standard_deviation_2 <- 250
Margin_error <- 25
Sample_size_2 <- ((Z_score_99 * Standard_deviation_2) / Margin_error)^ 2
Sample_size_2
```

```
## [1] 663.5776
```

The minimum required sample size for Luke is 664.

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?
(b) Are the reading and writing scores of each student 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?
(d) Check the conditions required to complete this 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?
(f) What type of error might we have made? Explain what the error means in the context of the application.
(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.

(a)

There is no clear difference in the average reading and writing scores based on the plots.

(b)

The reading and writing scores of each student are not independent of each other, they are paired.

(c)

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

(d)

The sample taken is random and we can assume that the sample size is less than 10% of the population, so we can the samples are independent. The sample size is large enough to assume a normal distribution.

(e)

```r
Standard_deviation_3 <- 8.887
Average_difference <- -0.545
Sample_size_3 <- 200
df_2 <- Sample_size_3 - 1
SE_2 <- Standard_deviation_3 / sqrt(Sample_size_3)
# Calculate T-score
T_score_2 <- (Average_difference - 0) / SE_2

# Use T-score to calculate p-value
P_value <- 2 * pt(T_score_2, df_2)
P_value
```

```
## [1] 0.3868365
```

P-value is 0.387. Since the p-value is greater than 0.05, and if we are assuming a 95% confidence interval, we fail to reject H0. Therefore, there is not enough data to provide convincing evidence of a difference between the average scores on the two exams.

(f)

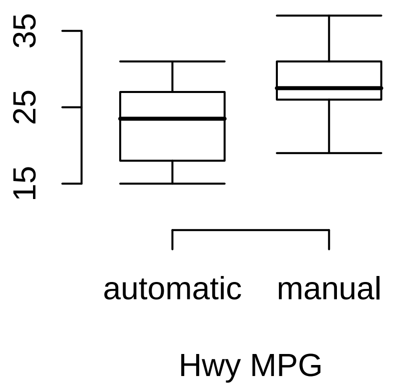
The type of error we might have made is a Type 2 error. This means that we may have incorrectly rejected HA.

(g)

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. This is because since we failed to reject H0, and it included the value of 0, we could also expect a confidence interval of 0.

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
SE_3 <- sqrt((5.29)^ 2 / 9 + (5.01)^ 2 / 9)
Average_difference_2 <- 22.92 - 27.88
# Calculate T-score
df_3 <- 9 -1
p_2 <- 0.98
x_2 <- p_2 + (1 - p_2) / 2
T_score_3 <- qt(x_2, df_3)
Lower_interval <- Average_difference_2 - T_score_3 * SE_3
Upper_interval <- Average_difference_2 + T_score_3 * SE_3
c(Lower_interval, Upper_interval)
```

```
## [1] -11.994429   2.074429
```

The 98% confidence interval for the difference between average highway mileage of manual and automatic cars is (-11.99, 2.07).

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%?

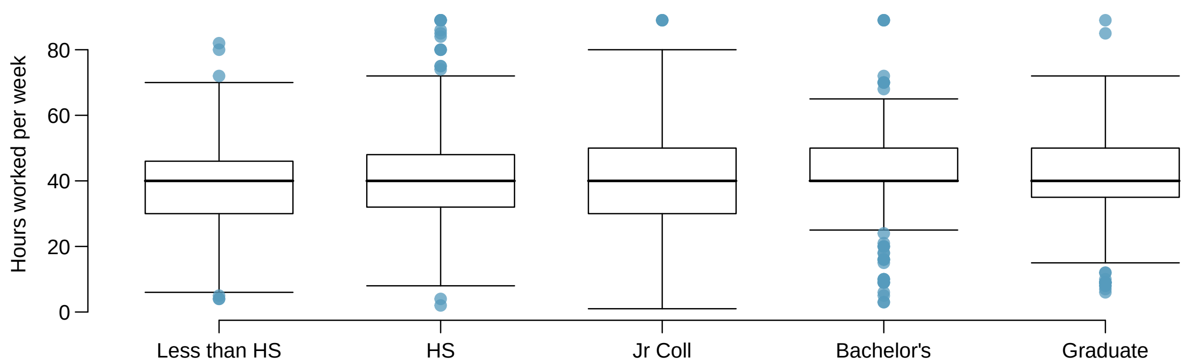
```
Z_score_80 <- 1.28
Standard_deviation_4 <- 2.2
Margin_error_2 <- 0.5
Sample_size_4 <- ((Z_score_80 * Standard_deviation_4) / Margin_error_2)^ 2
Sample_size_4
```

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

If the medical research group wants a desired power level of 80%, they would need to collect a sample of 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.
(b) Check conditions and describe any assumptions you must make to proceed with the test.
(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			501.54		0.0682
Residuals		267,382			
Total					

(d) What is the conclusion of the test?

(a)

HO: The average number of hours worked does no vary across the five groups. HA: The average number of hours worked does vary across the five groups.

(b)

We assume that the sample taken is random. The sample is less than 10% of the population so we can assume indendence. The sample size is large enough to assume a normal distribution.

(c)

```
df_degree <- 5 - 1
df_degree
```

```
## [1] 4
```

```
Sample_size_5 <- 1172
Mean_degree <- 501.54
Sum_resident <- 267382
df_resident <- Sample_size_5 - 5
df_resident
```

```
## [1] 1167
```

```
df_total <- df_degree + df_resident
df_total
```

```
## [1] 1171
```

```
Sum_degree <- df_degree * Mean_degree
Sum_degree
```

```
## [1] 2006.16
```

```
Sum_total <- Sum_resident + Sum_degree
Sum_total
```

```
## [1] 269388.2
```

```
Mean_resident <- Sum_resident / df_resident
Mean_resident
```

```
## [1] 229.1191
```

```
F_value <- Mean_degree / Mean_resident
F_value
```

```
## [1] 2.188992
```

(d)

The conclusion of the test is if we assume a 95% confidence interval, then the p-value = 0.0682 is greater than 0.05. This means that we would fail to reject HO. However, there is not enough evidence to accept HA.