Chapter 7 - Inference for Numerical Data

Donald Butler

10/24/2021

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

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

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

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

```
(SD = ME / qnorm(.95) * sqrt(25))
```

```
## [1] 18.2387
```

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?

$ME = Z_* \times \frac{\sigma}{\sqrt{n}}$

```
(qnorm(.95) * 250 / 25)^2
```

```
## [1] 270.5543
```

n >= 271

(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.

Since the margin of error is higher with an increased confidence level, he must sample more students to maintain the margin of error to be no more than 25.

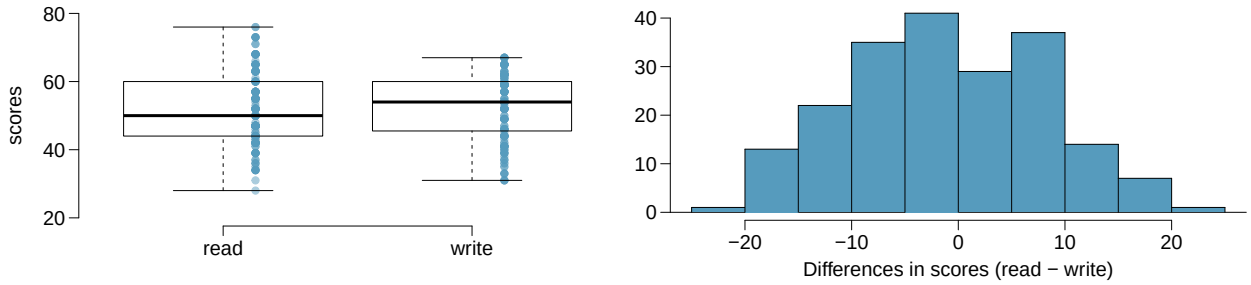
(c) Calculate the minimum required sample size for Luke.

```
(qnorm(.995) * 250 / 25)^2
```

```
## [1] 663.4897
```

He must 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?

No.

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

Not likely.

(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?

$H_0 : \mu_{diff} = 0$ There is no difference in reading and writing scores. $H_A : \mu_{diff} \neq 0$ There is a difference in reading and writing scores.

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

The observations are independent because it was taken from a simple random sample. There do not appear to be significant outliers and the sample size is 200, so the sample is nearly 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?

```
(T = (-0.545 - 0) / (8.887 / sqrt(200)))
```

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

```
(p_value = 2 * pt(T,199))
```

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

Since the p_value is greater than .05, we do not reject H_0 and conclude that there is no difference in reading and writing scores.

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

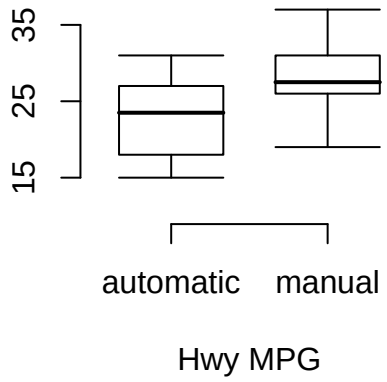
We may have made a Type II error by incorrectly failing to reject H_0. If there actually is a difference between reading and writing scores, but our sample failed to identify the difference.

(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, since we did not reject H_0, the confidence interval should contain the difference 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




```
mpg_diff <- 27.88 - 22.92
mpg_ME <- qt(.99,25) * sqrt(5.29^2/26 + 5.01^2/26)

c(mpg_diff - mpg_ME, mpg_diff + mpg_ME)
```

```
## [1] 1.409078 8.510922
```

We are 98% confident that the difference in fuel economy (mpg) between manual and automatic transmission vehicles is between 1.4 and 8.5 miles per gallon.

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%?

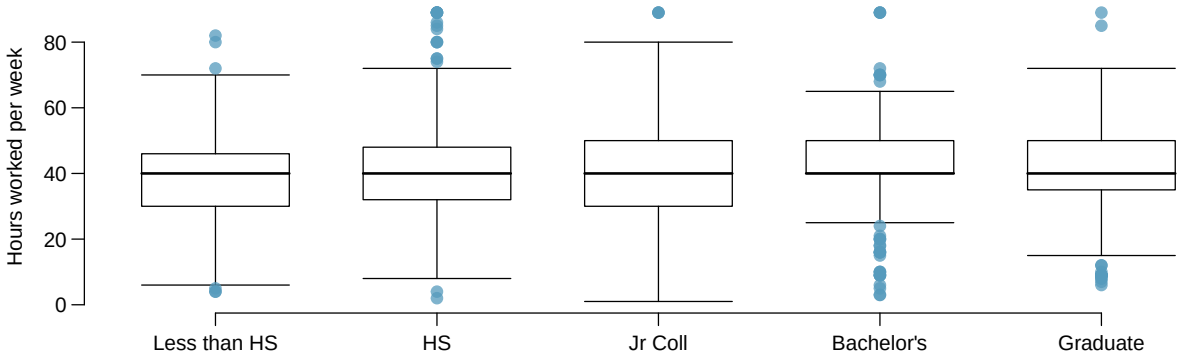
```
(n = (qnorm(.8) + qnorm(.975))^2/.5^2 * (2.2^2 + 2.2^2))
```

```
## [1] 303.9086
```

We would need 304 enrollees in each group to show an improvment in survey rates of at least .5 with a power level of 80%.

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 : There is no difference in the mean number of hours worked per week between groups based on their education level. H_A : There is a difference in the mean number of hours worked per week between groups based on their education level.

(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.

```
MS_G <- 501.54
SS_E <- 267382

df_G <- 5 - 1
df_T <- nrow(gss_sub) - 1
df_E <- df_T - df_G
SS_G <- MS_G * df_G
SS_T <- SS_G + SS_E
MS_E <- SS_E / df_E
F_value <- MS_G / MS_E
```

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree	4	2006.16	501.54	2.1889925	0.0682
Residuals	1167	267,382	229.1191088		
Total	1171	2.6938816×10^5			

(d) What is the conclusion of the test?

Since the p-value is greater than .05, we do not reject H_0 and conclude that there is no difference between hours worked and education level.