Chapter 7 - Inference for Numerical Data

Zachary Palmore

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.

```r
# At 90% confidence level z = 1.645
z <- 1.645
n <- 25
upci <- 77
loci <- 65
# Finding x_bar - the sample mean
x_bar <- (upci+loci) / 2
# Find the standard error
SE <- (upci-x_bar)/z
# Validating - SE and SE_lo should be identical
SE_lo <- (loci-x_bar)/(-z)
# Find the sample standard deviation - s
s <- SE * sqrt(n)
# Finding the margin of error - me
me <- z*(s/(sqrt(n)))
```

The sample mean is 71, the margin of error is 6, and the sample 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?
(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.

```
s <- 250
me <- 25
z_90 <- 1.645
z_99 <- 2.58
n_raina <- (z_90^2*s^2)/me^2
```

The sample size for Raina should be at least 270.6025. The minimum sample size for Luke should be larger because he wants a higher level of confidence in the average SAT score. Increasing the sample size will increase the accuracy of the estimate.

```
s <- 250
me <- 25
z_90 <- 1.645
z_99 <- 2.58
n_luke <- (z_99^2*s^2)/me^2
```

Luke's minimum sample size is 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.
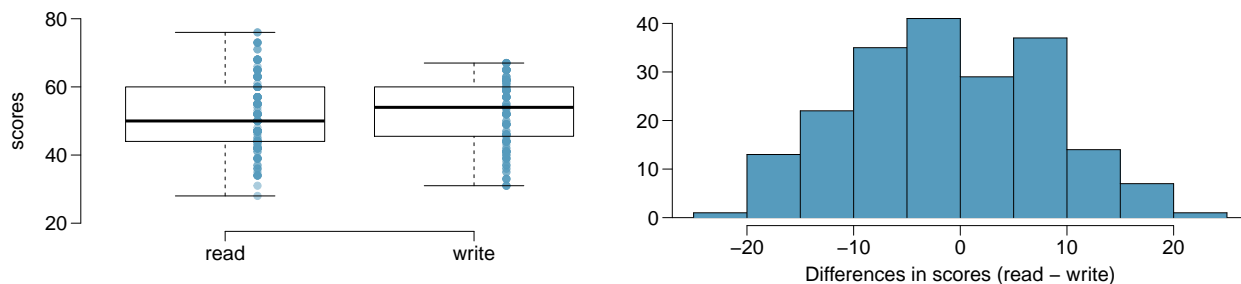
(a) Is there a clear difference in the average reading and writing scores?

No, there may be a slightly higher average writing score but the difference is not clear.

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

Yes, each student took the test individually and presumably has little to no effect on other student's scores. The scores of one student gives no indication of the scores of the other students.

(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?

Null hypothesis: There is no evidence to show the average scores of students in the reading and writing exam are different.

Alternative hypothesis: There is evidence to show the average scores of students in the reading and writing exam are different.

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

There are two conditions, independence, which was confirmed in part b, and normality. There is a sample size greater than 30 and there are no particularly extreme values. The data appears normally distributed. Thus, the required conditions are satisfied.

(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?

```
s <- 8.887
obs_diff <- -0.545
n <- 200
p <- 2 * pnorm(obs_diff, mean=0, sd=s/sqrt(n))
```

No, at a significance level of 0.05 the data do not provide evidence that there is a statistically significant difference between the average reading and writing scores. Our p-value, 0.3857919, is greater than 0.05 indicating that we should reject the null hypothesis.

(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 1 error where we falsely rejected the null hypothesis. In other words, if the 95% confidence interval includes zero, there is a good chance that there is no difference in the average scores. We can see this in a 95% confidence interval calculation below.

```
z <- 1.96
obs_diff + z * (s/sqrt(n))
```

```
## [1] 0.6866754
```

```
obs_diff - z * (s/sqrt(n))
```

```
## [1] -1.776675
```

At 95% confidence, the interval spans zero.

(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.

There is a chance that the average difference between the reading and writing scores would include zero. Specifically, if we repeated this simple random sample with 200 students and if the true average difference is zero, the probability of finding the observed average difference to be 0.545 in both directions is about 38.6%. In this case, the confidence interval is included in part f. The lower interval is -1.7766754 and the upper interval is 0.6866754 which includes zero.

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
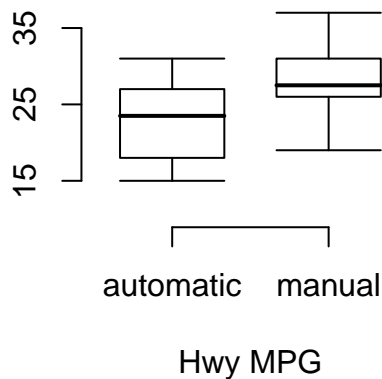

```
xa <- 22.92
sa <- 5.29
na <- 26
xm <- 27.88
sm <- 5.01
nm <- 26
# Critical value for t
df = (nm - 1)
t <- 2.479 # from t-score table for one-way
# Standard error
se <- sqrt(((sm^2)/nm)+((sa^2)/na))
# Confidence intervals
(xm - xa) + t*se
```

```
## [1] 8.502195
```

```
(xm - xa) - t*se
```

```
## [1] 1.417805
```

We are 98% confident that the average difference between highway mpg for automatic and manual cars is between 1.4178048 and 8.5021952.

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%?

Using a significance level of 0.05 where the distribution is approximately normal,

```
s <- 2.2
p1 <- 0.84
p2 <- 1.96
p <- p1 + p2
efsize <- 0.5
enrollees <- 2*(s^2)*(p/efsize)^2
```

they will need at least 304 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
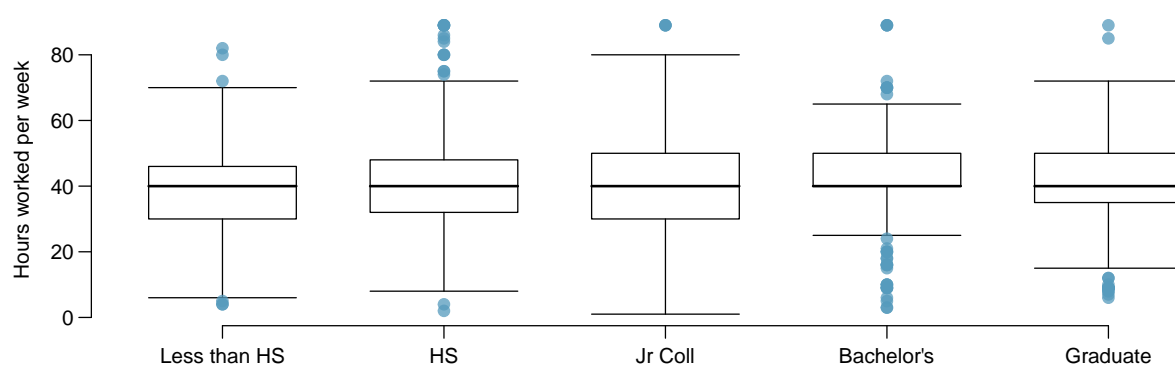
(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

Null hypothesis: The average number of hours worked does not vary across the five education groups.

Alternative hypothesis: The average number of hours worked varies across the five education groups.

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

There is an assumption that the observations are independent within and across groups. The data also appears nearly normal and the variability across the groups is about equal. With this, the conditions for testing are met.

(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					

The statistics can be found through the anova below.

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

```
##              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
```

(d) What is the conclusion of the test?

The results give a p-value of 0.0682. At a significance level of 0.05, we fail to reject the null hypothesis. There is not enough evidence to support the alternative.