Chapter 7 - Inference for Numerical Data

Jeff Shamp

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.

ME is half the distance from the upper to lower bounds of the CI, thus $ME = 6$. The sample mean is 71, since it is the midpoint of the CI. The sample std is found using the follwowing form; $77 = 71 \pm (1.708) \times \frac{s}{\sqrt{25}}$**,** $s = 17.56$**.**

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?

```
n<- ((1.645*250)/25)^2
n
```

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

Raina needs at least 271 students to sample.

(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.

To cover a wider band of data (99% CI) with the same ME, then he would need to sample more people.

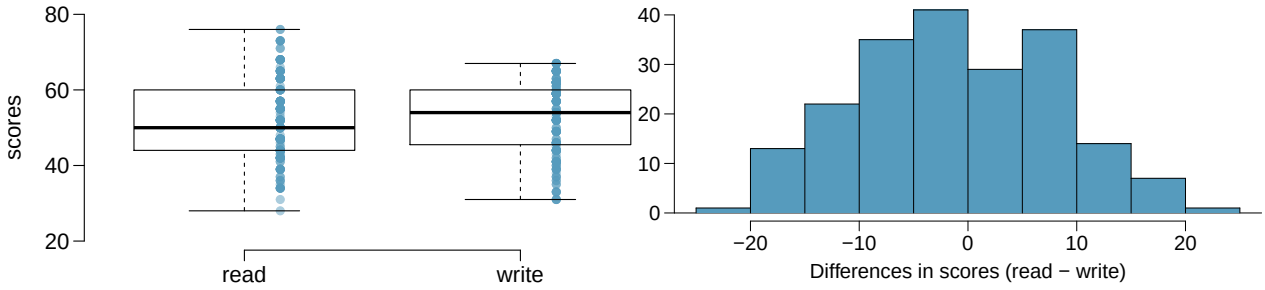
(c) Calculate the minimum required sample size for Luke.

```
n<- ((2.58*250)/25)^2
n
```

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

666 people to sample

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 does not appear to be a meaningful difference.

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

They are reasonably independent. Reading is a simple skill that most students should have learned years ago. Writing is a craft that takes year to develop.

(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 reading and writing mean score, $\mu_{diff} = 0$**. HA: There is a difference in reading and writing mean score,**$\mu_{diff} \neq 0$**.**

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

This data is definitely normal with no outliers. Data was collected by simple random sample. We are good to move forward.

(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?

```
SE<- 8.887/sqrt(200)
Z_score<-(-0.545-0)/SE
pnorm(Z_score)
```

```
## [1] 0.192896
```

This p-value would led us to not reject the null hypothesis, there is no difference between read and writing mean score.

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

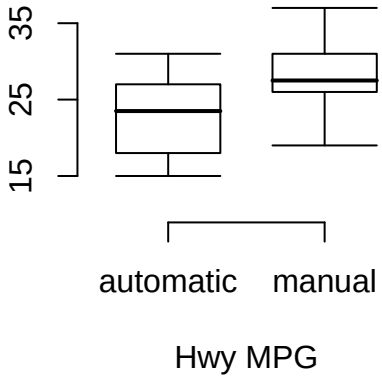
This could be a type 2 error, where the null is false, and we did not reject it.

(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.

Given these results, I would expect them to contain 0 as there is no difference between them.

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



Since this is a small sample (n<30) we will use the T-score paradigm with 98% CI

```
mean_diff<-22.78-27.88
SE<-sqrt(((5.29^2)/26)+((5.01^2)/26))
qt(.01,25,lower.tail = FALSE)
```

```
## [1] 2.485107
```

Now we can build our CI.

```
upper<-mean_diff+2.485*SE
lower<-mean_diff-2.485*SE
paste("98% CI:", lower, "-", upper)
```

```
## [1] "98% CI: -8.65076849181976 - -1.54923150818023"
```

We can be 98% confident that manual transmission cars have higher MPG.

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%?

**For a power of 80%, we need to calculate the needed Z score for the 20% tails.

```
qnorm(.8)
```

```
## [1] 0.8416212
```

Now we can compute the sample size with a critical value of 0.05 and a pooled SE

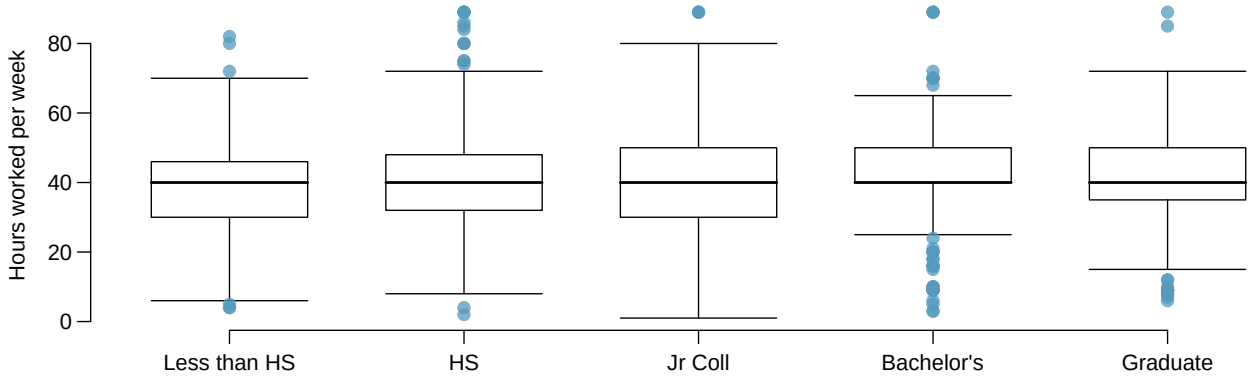
```
n<-((1.96+0.8416)^2 / 0.5^2 )*(2*2.2^2)
n
```

```
## [1] 303.9118
```

We would need to sample 303 people.

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.

H0: There is no difference between average hours worked relative to education. HA: There is a difference between average hours work by education level.

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

Independence: Random sample seems like a good assumption for this test. Normal: There don't appear to be an extreme outliers for these groups given the sample size. Constant Variance: These seem to be aligned as well with bachelor's group being a little more compressed.

(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?

```
mean<-40.45
n<-1172
k<-5

SSE <- sum((121-1)*15.81^2, (546-1)*14.97^2, (97-1)*18.1^2,
           (253-1)*13.62^2, (155-1)*15.51^2)
SSG <- sum(121*(38.67-mean)^2, 546*(39.6-mean)^2, 97*(41.39-mean)^2,
           253*(42.55-mean)^2, 155*(40.85-mean)^2)
SST <- SSE + SSG
MSG <- SSG / (k-1)
MSE <- SSE / (1172-k-1)
f = MSG/MSE
paste(SSG, f, MSE)
```

```
## [1] "2004.10059999999 2.1849398103003 229.308444854202"
```

Df : 4 Sum sq: 2004 F-value: 2.184 res Df: 1167 res Mean sq: 229.30 Total df: 1171 Total Sum sq: 269,386