Chapter 7 - Inference for Numerical Data

George Cruz

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.

```
l_bound <- 65
h_bound <- 77
s_mean <- (l_bound + h_bound)/2

m_e <- (h_bound - l_bound)/2

cv <- 1.64

t_df <- round(qt(c(.05, .95), df=24)[2], 3)
SE <- round((h_bound-s_mean)/t_df, 3)
std_d <- SE * sqrt(25)

c(mean=s_mean, me=m_e, SD=std_d)
```

```
##   mean     me     SD
## 71.000  6.000 17.535
```

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?

```
cv <- 1.64

n <- (10*1.64)**2

n
```

```
## [1] 268.96
```

She will need to collect a sample of 269 people.

(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.

If we want a higher confidence level, I would guess the sample size has to be larger.

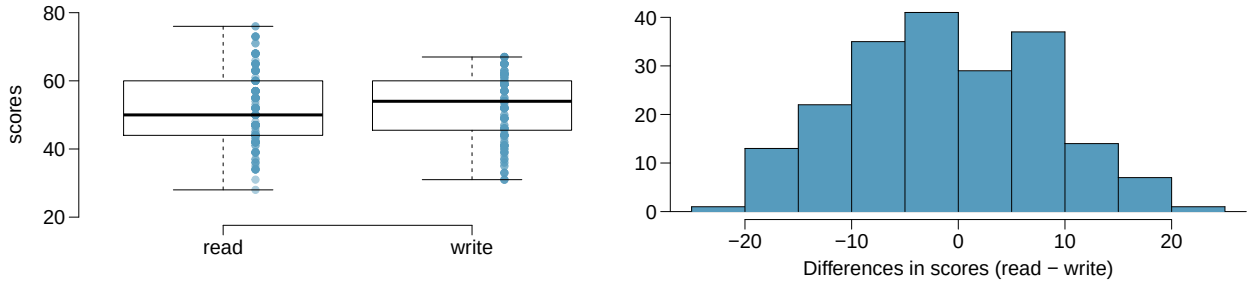
(c) Calculate the minimum required sample size for Luke.

```
cv <- 2.575

(10*2.575)**2
```

```
## [1] 663.0625
```

He would need a sample size of $\sim$ 663 people.

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. The distribution of differences appears normal.

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

Yes.

(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's no difference in the avg scores of students in reading/writing exam. H1: There is a difference in scores.

(d) Check the conditions required to complete this test. 1. Variables are independent 2. Distribution is normal 3. Sample size is > 30

(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?

```
std_diff <- 8.887
mu_diff <- -0.545
n <- 200

std_err_diff <- std_diff / sqrt(n)

t <- (mu_diff - 0) / std_err_diff

p <- pt(t, df = n-1)

p
```

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

Since the p value is > 0.05 then we reject H1.

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

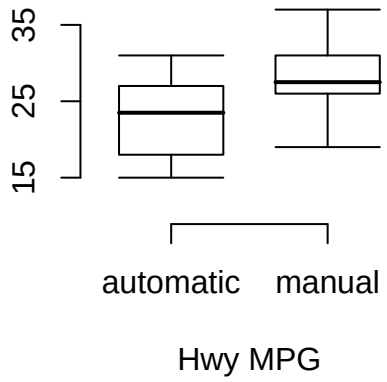
A Type II error is when we incorrectly reject the Alternate Hypothesis.

(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. If the confidence interval includes 0 then we must accept the Null Hypothesis.

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




```
z <- qnorm(0.99)  #for a 98% CI
avg_auto<-22.92
sd_auto<-5.29
n<-26

avg_manual<-27.88
sd_manual<-5.01

mean_diff <- avg_manual - avg_auto
se <- ( (sd_manual ** 2 ) / n) +
          ((sd_auto ** 2) / n) ** 0.5
me <-z * se

ci_95 <- c(low=mean_diff - me, high=mean_diff + me)

knitr::kable(ci_95)
```

	x
low	0.3006909
high	9.6193091

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%?

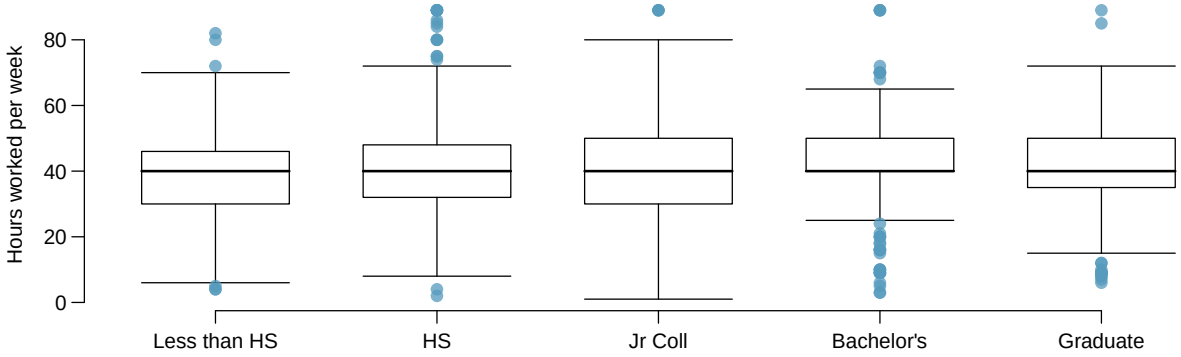
```
sd1 <- 2.2
sd2 <- 2.2
pl1 <- qnorm(0.99) # I could not use 1
pl2 <- qnorm(0.8)

std_err <- 0.5/(pl1+pl2)
n <- (sd1**2 + sd2**2)/std_err**2

n
```

```
## [1] 388.595
```

They would need 389 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.

- H0: The avg number of hours does not change across the 5 groups.
- H1: There is a variation of avg hours across the groups

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

The variables are independent of each other and the distribution of data appears normal.

(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					

```
mean <- c(38.67, 39.6, 41.39, 42.55, 40.85)
std_d <- c(15.81, 14.97, 18.1, 13.62, 15.51)
count <- c(121, 546, 97, 253, 155)
my_df <- data.frame (count, mean=mean, sd=std_d)

n <- sum(my_df$count)
k <- 5

df <- k - 1
residual <- n - k
```

```r
prf <- 0.0682
f_stat <- qf( 1 - prf, df , residual)

msq <- 501.54
msr <- msq / f_stat
ssq <- df * msq
ssr <- 267382
ss_total <- ssq + ssr
df_total <- df + residual

my_table <- data.frame(title=c("degree", "residuals", "total"),
                       df=c(df, residual, df_total),
                       sum_sq=c(ssq, ssr, ss_total),
                       mean_sq=c(msq, msr,""),
                       f_value=c(f_stat,"",""),
                       prf=c(prf,"",""))

knitr::kable(my_table)
```

title	df	sum_sq	mean_sq	f_value	prf
degree	4	2006.16	501.54	2.18893121413288	0.0682
residuals	1167	267382.00	229.125518774549		
total	1171	269388.16			

(d) What is the conclusion of the test?

Because the p-value is slightly greater than 0.05 we fail to reject the null hypothesis.

...