Chapter 7 - Inference for Numerical Data

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.

$65 = x - t * std/sqt(n) \rightarrow x = 65 + t * \sigma/sqrt(n)$

$77 = 65 + t * \sigma/sqrt(n) + t * \sigma/sqrt(n) \rightarrow \sigma = 12/t * sqrt(n) \rightarrow \sigma = 1.4 \rightarrow x = 65.4$

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?

$moe = z * s/sqrt(n) \rightarrow n = (z * s/moe)^2$

```
std = 250
moe = 25
z = 1.645
(z * std / moe)^2
```

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

(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.

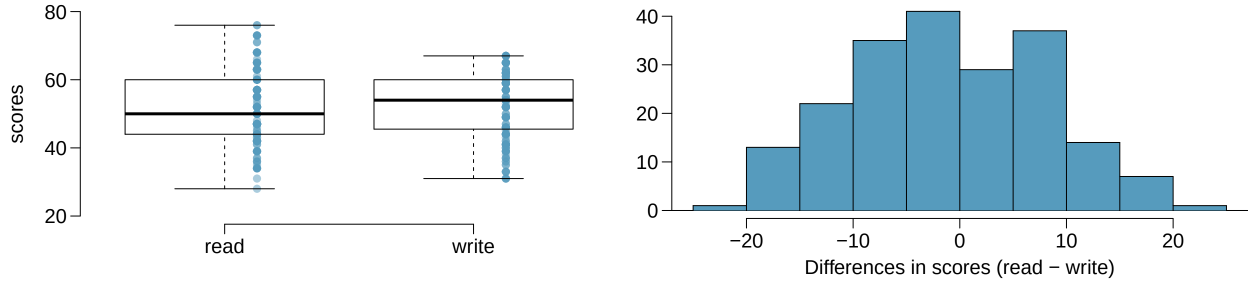
Increasing the confidence interval requirement will cause z to go up and this n.

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

```
z = 1.96
(z * std / moe)^2
```

```
## [1] 384.16
```

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?

Yes the average writing is higher than the average reading

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

Yes, one students skills are independent of anothers

(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?

Two tailed hypothesis $H_o = \bar{x}_{\text{read}} - \bar{x}_{\text{write}} = 0$

$H_a = \bar{x}_{\text{read}} - \bar{x}_{\text{write}} \neq 0$

(d) Check the conditions required to complete this test. n > 30 and and independance was verified in part b

(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?

```
udiff = -0.545
sdiff = 8.887
h_0 = 0
n = 200

t = (udiff - h_0) / (sdiff/sqrt(n))
pt(t, df=199) * 2
```

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

With such a large p value we accept the null hypothesis. The probability of obtaining a random sample of 200 students where the average difference between reading and writing is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.

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

Given type 1 error where we reject the null hypothesis when it’s true. type 2 error where we accept the null hypothesis when we are suppose to reject it

Since the null is true here we could have made a type 1 error.

(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.

```
t = 1.65
c(-0.545 - t * 8.887/sqrt(200), -0.545 + t * 8.887/sqrt(200))
```

```
## [1] -1.5818696  0.4918696
```

The confidence interval does contain 0.

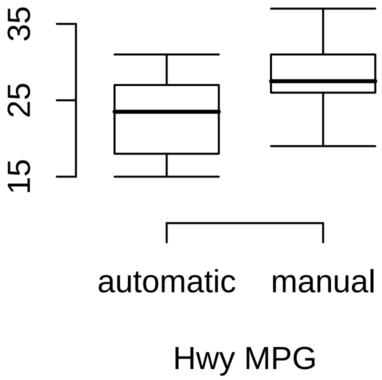
Fuel efficiency of manual and automatic cars, Part II. (7.28, p. 276) The table provides summary statistics on highway fuel economy of cars manufactured in 2012. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, 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
na=nm=26
dfa=dfm=25
t = 2.485
ma = 22.92
mm = 27.88
sa = 5.29
sm = 5.01
pe = mm - ma
se = sqrt(sm^2/nm + sa^2/na)
c(pe - t * se, pe + t * se)
```

```
## [1] 1.409232 8.510768
```

The difference of the average highway mileage of manual and automatic cars is between 1.4 and 8.5.



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%?

$$es = (z1 + z2) * se = (z1 + z2) * sqrt(s1^2/n + s2^2/n) \rightarrow n = (s1^2 + s2^2)/(es/(z1 + z2))^2$$

Identifying the Z-score that would give us a lower tail of 80%

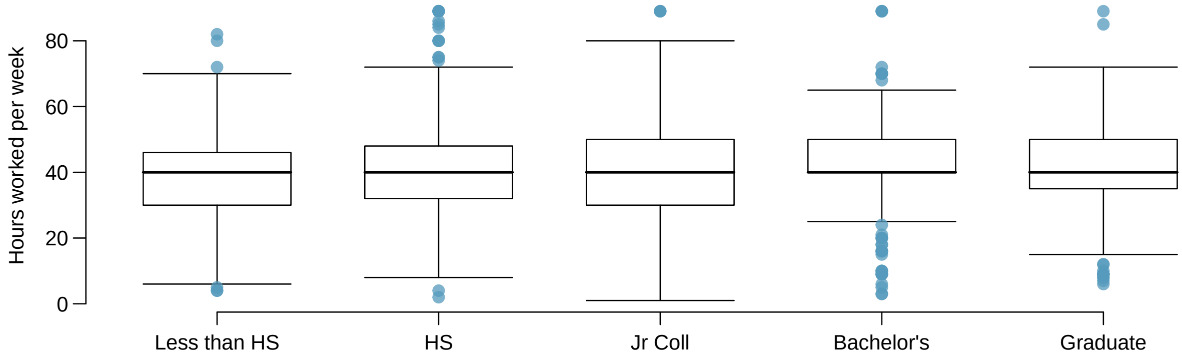
```
z1 = qnorm(0.8)
#alpha of 0.5 -> z=1.96
z2 = 1.96
s1 = s2 = 2.2
es = 0.5

n = (s1^2 + s2^2) / (es/(z1 + z2)) ^2
```

~304 new enrollees are needed for each interface to detect an effect size of 0.5.

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.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
H_A: The average varies across some or all groups

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

n = 1172 $>$ 30

Assume the observations are independent within and across groups and variability across the groups is about equal.

(c) Below is part of the output associated with this test. Fill in the empty cells.

```
k = 5
df_d = k - 1
df_r = 1172 - df_d - 1
df_t = df_d + df_r

ms_d = 501.54

ss_d = ms_d * df_d
ss_r = 267382
ss_t = ss_d + ss_r

ms_r = ss_r / df_r

f_val = ms_d / ms_r
```

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree	4	2006.16	501.54	2.189	0.0682
Residuals	1167	267,382	229.12		
Total	1171	269388.16			

(d) What is the conclusion of the test?

With a 5% significance level and a p_val of 0.0682 we accept the null hypothesis. However it should be noted how close those values are.