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.

The given 90% confidence interval (65, 77) is equal to:

$$\text{point estimate } \pm z^{\star} SE$$

The sample mean is the average of the 90% confidence interval:

```
upper <- 77
lower <- 65
s.mean <- (upper+lower)/2
cat("The sample mean is:",s.mean)
```

```
## The sample mean is: 71
```

The margin of error:

$$z^{\star} \times SE$$

Is equal to half the the difference between the upper and lower confidence interval.

```
diff.u.l <- upper - lower
M.E. <- diff.u.l/2
M.E.
```

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

Knowing the margin of error and the z score for 90% confidence (1.64) we can calculate the standard deviation:

```
z.score.90pc <- 1.64
S.D <- (M.E./z.score.90pc) %>% round(2)
cat("The standard deviation is:", S.D)
```

```
## The standard deviation is: 3.66
```

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.

$$\text{MOE} = \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

(a) Raina wants to use a 90% confidence interval. How large a sample should she collect? From the Margin of Error equation:

$$n = \left(\frac{z^*(\sigma)}{\text{MOE}} \right)^2$$

```
SAT.sd <- 250
z.score.90pc <- 1.64
M.E.sd <- 25
n.SAT <- ((z.score.90pc*SAT.sd/M.E.sd)^2) %>% round(0)
cat("The sample size for a 90% confidence interval is:", n.SAT, "students.")
```

```
## The sample size for a 90% confidence interval is: 269 students.
```

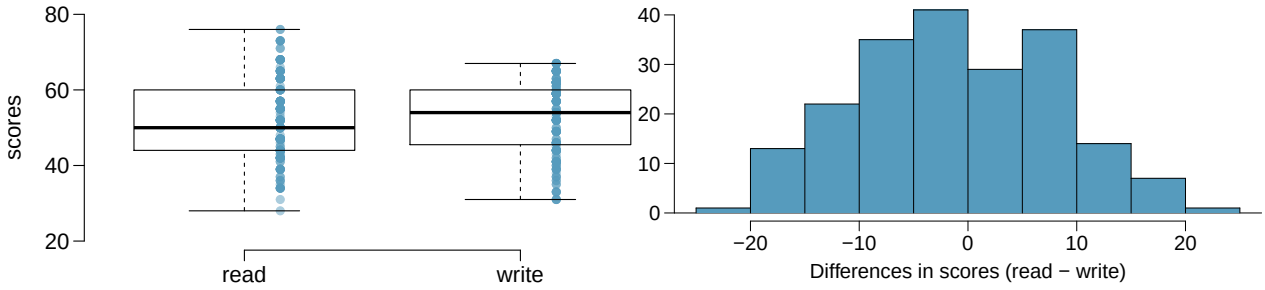
(b) Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina's, and explain your reasoning.

Z score is proportional to the square root of the sample size. The larger the Z score or confidence interval, the larger the sample size. (c) Calculate the minimum required sample size for Luke.

```
SAT.sd <- 250
z.score.99pc <- 2.58
M.E.sd <- 25
n.SAT <- ((z.score.99pc*SAT.sd/M.E.sd)^2) %>% round(0)
cat("The sample size for a 99% confidence interval is:", n.SAT, "students.")
```

```
## The sample size for a 99% confidence interval is: 666 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. The median reading and writing scores are different but there is a clear overlap between the Q1 and Q3 quartiles

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

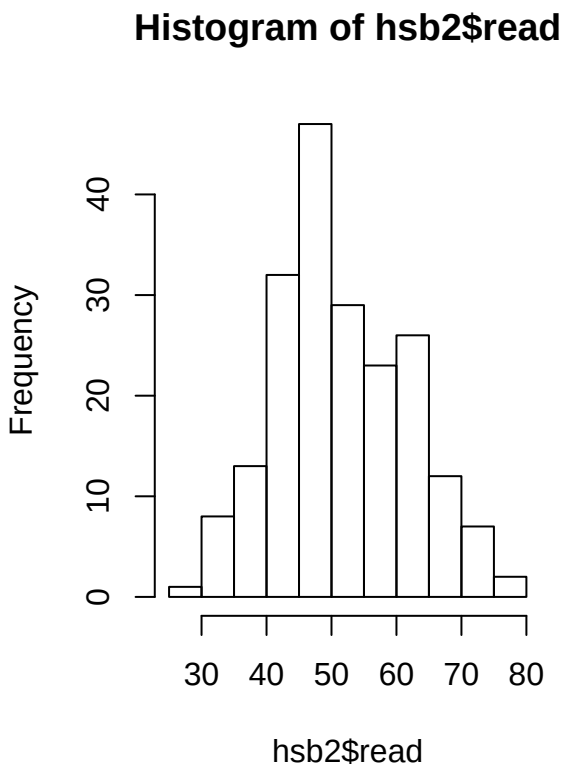
They are not completely independent. As a skill both are usually learned together.

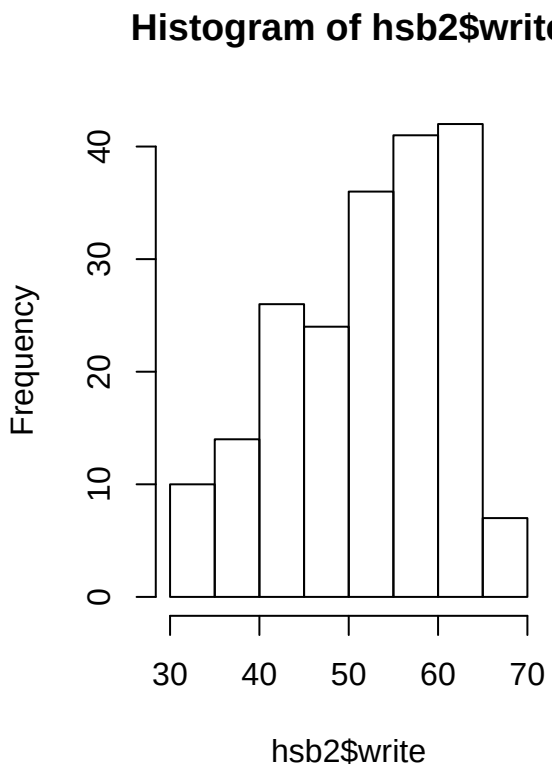
(c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

The Null Hypotheses is that the average scores of reading and writing are the same. The Alternative Hypotheses is that the average scores of students in the reading and writing exam are different.

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

```
par(mfrow=c(1,2))
hist(hsb2$read)
hist(hsb2$write)
```





The data for both the read and writing scores appears random, normal and independent.

(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?

We can compute the T score with the given values:

```
mean.diff <- -0.545
SD.diff <- 8.887
T.Read.Write <- ((mean.diff-0)/SD.diff) %>% round(2) %>% abs()
cat("The calculated T Value of", T.Read.Write, "correspond to a p-value of 0.75 in the T Value Table.\n"
```

```
## The calculated T Value of 0.06 correspond to a p-value of 0.75 in the T Value Table.
## This p-value score does not provide convincing evidence of a difference between the average scores o
```

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

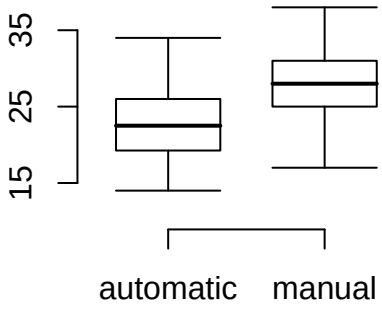
** Potentially we could have made a Type II error (false negative). Meaning that we did not discount the possibility that we erroneously fail to be reject the null hypotheses that there is an evident difference in the average scores of students in the reading and writing exam**

(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.

I would expect to include zero and it is included given that the verage observed difference in scores is $\widehat{x}_{read-write} = -0.545$, and the standard deviation of the differences is 8.887 points. We did not reject the null hypotheses that there is no difference between reading and writing exams. No difference would give a mean difference value 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



We can use T statistics to calculate the 98% confidence by first calculating the standard error for the difference of two means:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
s1 <- 5.29
s2 <- 5.01
n1 <- 26
n2 <- 26
SE.1.2 <- ( sqrt((s1^2/n1)+(s2^2/n2)) ) %>% round(2)
cat("The Standard Error is:", SE.1.2)
```

```
## The Standard Error is: 1.43
```

$$\text{Confidence Interval} \;=\; \text{point estimate} \;\pm t^{\star} SE$$

For the degree of freedom df we take them from smaller of $n_1 - 1$ or $n_2 - 1$

In this case df is equal to 25. Checking a t score table we find the value of 2.486 for 25 degrees of freedom and 98% confidence interval.

```
P.E <- (27.88-22.92)
T.score.98 <- 2.486
M.E.98 <- (T.score.98 * SE.1.2) %>% round(2)
cat("For a 98% confidence interval the Upper limit is", P.E+M.E.98, "the Lower limit is", P.E-M.E.98," a
```

```
## For a 98% confidence interval the Upper limit is 8.51 the Lower limit is 1.41  and with a margin of e
```

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%?

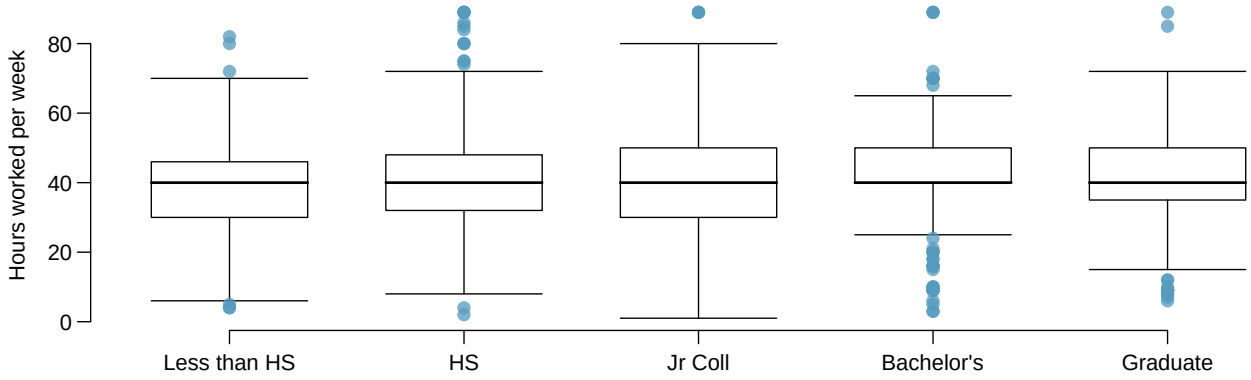
High Statistical Power: Small risk of committing Type II errors - e.g. a false negative

```
n.surveys <- 4
sd.survey <- 2.2
n <- ( ( (2.8^2)/(0.5^2) ) * (2.2^2 + 2.2^2) ) %>% round(0)
cat("We would need", n, "people to detect an effect size of 0.5 surveys per enrollee.")
```

```
## We would need 304 people to detect an effect size of 0.5 surveys per enrollee.
```

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 : The average number of hours worked is the same for all five groups.
H_A : The average hours varies by group. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the group averages than what we might expect from chance alone.

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

The distribution appears random, normal and independent.

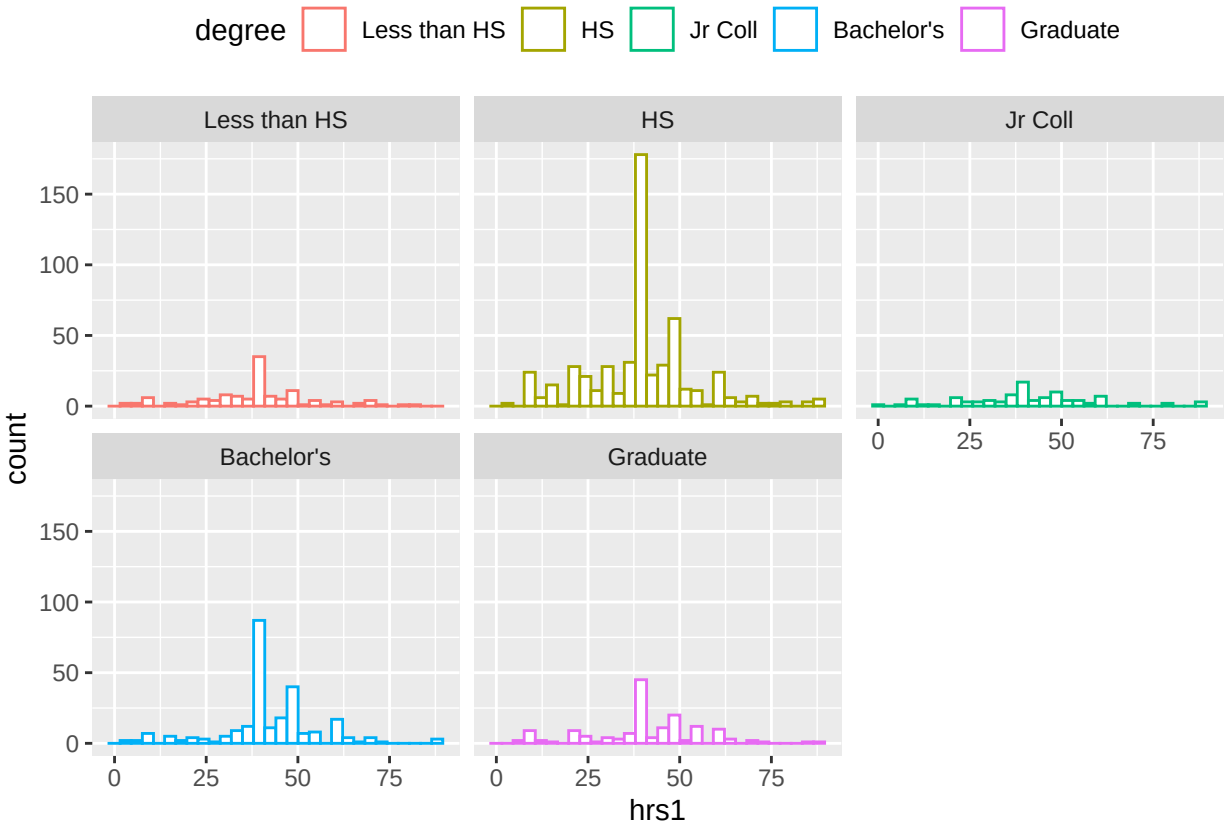
```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':
##
##     diamonds
```

```
ggplot(gss_sub, aes(x=hrs1,color=degree, scales = "free_y")) +
  geom_histogram(fill="white", position="dodge") +
  theme(legend.position="top")+
  facet_wrap(degree ~ .)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
anova_one_way <- aov(hrs1~degree,gss_sub)
summary(anova_one_way)
```

```
##               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
```

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree	4	2006.16	501.54	2.189	0.0682
Residuals	1167	267,382	229.1		
Total	1172	269388.16			

(d) What is the conclusion of the test?

The P (0.0682) value is does not reach the treshold significance value of 0.05. The results were not statistically significant. The average number of hours worked is the same for all five groups.