Chapter 7 - Inference for Numerical Data

Salma Elshahawy

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.

```
n <- 25 # the sample is below 30
df <- 24
upper <- 77
lower <- 65
s_hat <- (77 + 65) / 2
paste0("Sample mean= ",s_hat) #sample mean
```

```
## [1] "Sample mean= 71"
```

```
ME <- (77 - 65) / 2
paste0("Margin of error= ",ME) # margin of error
```

```
## [1] "Margin of error= 6"
```

```
t <- 1.711 # two tails t-tables
SE_hat <- (upper - s_hat) / t

sd_hat <- SE_hat * sqrt(n)
paste0("Sample standard deviation= ",sd_hat)
```

```
## [1] "Sample standard deviation= 17.5336060783168"
```

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_hat <- 250
ME <- 25
z <- 1.645 # alpha is 0.1 because of 90% confidence interval from the table
n <- ((z * s_hat) / ME) ^ 2
paste0 ("(a) Number of people needed for Raina sample is: ",round(n, digits = 0))
```

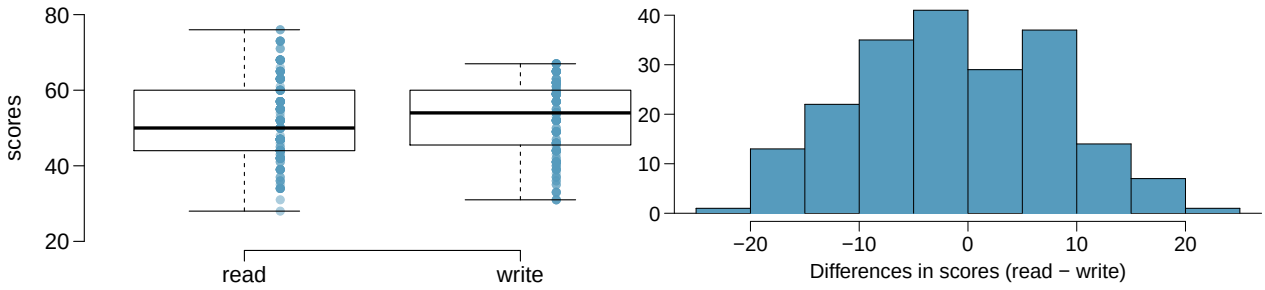
```
## [1] "(a) Number of people needed for Raina sample is: 271"
```

(b) Luke sampe size should be larger than Raina, because when we increase the confidence interval, the standard deviation increase,so as a result the number of samples should be bigger.

```
s_hat <- 250
ME <- 25
z <- 2.575 # alpha is 0.01 because of 99% confidence interval from the table
n <- ((z * s_hat) / ME) ^ 2
paste0 ("(c) Number of people needed for Luke sample is: ",round(n, digits = 0))
```

```
## [1] "(c) Number of people needed for Luke sample is: 663"
```

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?

The boxplots shows that there is some difference in reading and writing scores. The median for the writing appears to be higher. The 75th and 25th percentile line are somewhat close. The maximum for reading appears to be much higher than maximum score for writing.The reading - writing histogram show a spread from around -20 to 20. So there is a difference, but we can't tell if this difference is significant.

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

The sample is random and less than 10% of the total population. Thus we can assume that the scores of each other are independent

(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 the reading and writting scores HA: There is a difference in the reading and writting scores

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

Independence: yes, the scores of each student are independent of each other. Sample size is less than 10% of the total population Distribution: is a normal distribution

(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?

```
n <- 200
avg_red_wrt <- -0.545
sd_red_wrt <- 8.887
SE <- sd_red_wrt/sqrt(n)
SE
```

```
## [1] 0.6284058
```

```r
t_score <- (avg_red_wrt - 0)/SE
t_score
```

```
## [1] -0.867274
```

```r
P_value <- pt(t_score, n-1, lower.tail = TRUE)
paste0("The P_value is ",P_value)
```

```
## [1] "The P_value is 0.193418237099674"
```

This is about 0.20 or about 20% likelihood of seeing an average difference of -0.545 if the the null hypothesis is true that there is no difference.This is strong evidence to fail to reject the null hypothesis.

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

If there is a difference, then we would have made a type II error since we failed to reject the null hypothesis.It means that there is really a difference in the average scores between reading and writing and we failed to detect 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.

average difference between the reading and writing scores to include 0? Explain your reasoning. Based on the findings above, I would expect the confidence interval to include 0.

```r
SE <- sd_red_wrt/sqrt(n)
t_score <- qt(p=(0.05/2), n-1, lower.tail = FALSE)
t_score
```

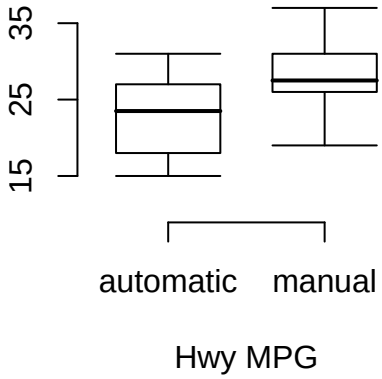
```
## [1] 1.971957
```

```r
upper <- avg_red_wrt + SE * t_score
lower <- avg_red_wrt - SE * t_score
c(upper, lower)
```

```
## [1]  0.6941889 -1.7841889
```

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	16.12	19.85
SD	3.58	4.51
n	26	26



H0: The average miles difference is equal to 0. HA: The average miles difference is not equal to 0.

```
n_man <- 26
n_auto <- 26
mean_auto <- 16.12
mean_man <- 19.85

sd_auto <- 3.58
sd_man <- 4.51

diff_mean <-  mean_man - mean_auto
diff_mean
```

```
## [1] 3.73
```

```
diff_sd <- sd_man - sd_auto
diff_sd
```

```
## [1] 0.93
```

```
se <- sqrt((sd_man^2/n_man) + (sd_auto^2/n_auto))
se
```

```
## [1] 1.12927
```

```
t <- (diff_mean - 0)/se
t
```

```
## [1] 3.30302
```

```
p <- pt(t, n-1, lower.tail = FALSE)
p
```

```
## [1] 0.0005670179
```

Since p-value < 0.05, reject H0. The data provide strong evidence that the there is a difference in the average miles.

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%?

$$n = \left(\frac{Z_{0.05} + Z_{0.8}}{estimated \quad size}\right)^2 \times 2SE^2$$

$$n = \left(\frac{1.96 + 0.84}{0.5}\right)^2 \times \sqrt{2.2^2 + 2.2^2}$$

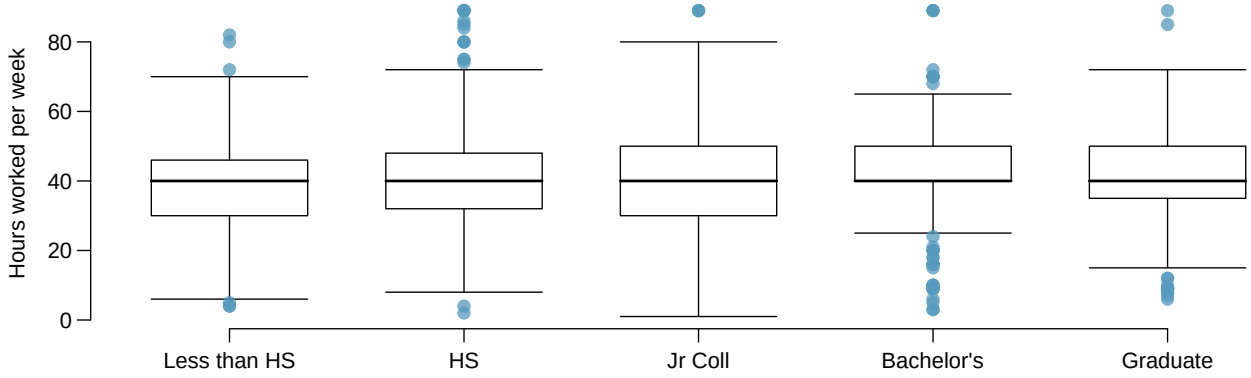
```
# alpha = 0.05, z0.8 = 0.84, z0.05 = 1.96
sd <- 2.2
es <- 0.5
n <- (((0.84 + 1.96)^2)/(0.5)^2)*(2 * 2.2^2)

paste0 ("Number of desired sample size to the power 80% is: ",round(n, digits = 0))
```

```
## [1] "Number of desired sample size to the power 80% is: 304"
```

We should target 304 survey in order to achieve 80% power at the 0.05 significance level for this context. The standard error difference of 2.8 × SE is specific to a context where the targeted power is 80% and the significance level is alpha = 0.05.

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 average number of hours worked on all 5 groups are the same. HA: The average number of hours worked on all 5 groups are ***not*** *the same*

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

Independence: yes, the observation parameters are independent of each other Sample size: the sample size is less than 10% of the total population The data follows the normal distribution

(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					

```
# we assume that the confidence interval is 95%, alpha is 0.05, p_value = 0.0682 which is > alpha, we f
k <- 5 # number of groups
n <- 1172
mean_srt <- 501.54
sum_srt <- 267382
p_value <- 0.0682

# get the Df for degree, residential, and total
```

```r
dfg <- k - 1
dfe <- n - k
dft <- dfg + dfe
df <- c(dfg, dfe, dft)
df
```

```
## [1]    4 1167 1171
```

```r
# find summation sqrt
sumation_rt_degree <- dfg * mean_srt
sumation_rt_resd <- sum_srt + sumation_rt_degree
total <- c(sumation_rt_degree, sum_srt, sumation_rt_resd)
total
```

```
## [1]   2006.16 267382.00 269388.16
```

```r
# find mean sqrt
mean_res <- sum_srt / dfe
mean_res
```

```
## [1] 229.1191
```

```r
# find the f-value
f_value <- mean_srt / mean_res
f_value
```

```
## [1] 2.188992
```

(d) What is the conclusion of the test?

P(>F): 0.0682 At significance level of 0.05, we fail to reject the null hypothesis that the means of the groups are different. There is about a 7% chance that this kind of variability is present when the means of the groups are the same.