Chapter 7 - Inference for Numerical Data

David Simbandumwe

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.

```
x1 <- 77
x2 <- 65
n <- 25
z <- 1.64
t <- 1.7109

(mean_s <- (x1+x2)/2)

## [1] 71

(ME <- (x1-x2)/2)

## [1] 6

(s <- (ME * sqrt(n)) / t)

## [1] 17.53463
```

- **sample mean is 71**
- **margine error 6**
- **standard deviation 17.53**

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?

$ME = z * \frac{s}{\sqrt{n}} n = \frac{z * s}{ME}$

```
z <- round(-qnorm((1-.9)/2),4)
s <- 250
ME <- 25

(z*s/ME)^2
```

```
## [1] 270.5696
```

- **n needs to be greater than 270.6 or 271**

(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 <- round(-qnorm((1-.99)/2),4)
```

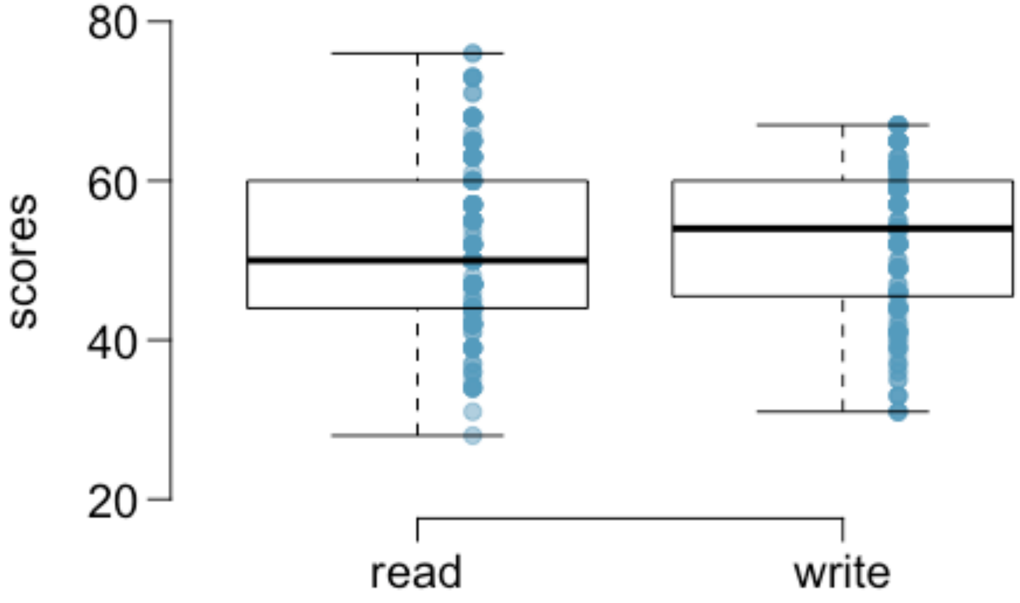
- **the z score for the 99% confidence interval is 2.58 compaired to 1.64 for the 90% confidence interval** $n = \frac{z*s}{ME}$
- **given the calculations for n Luke will require bigger sample size**

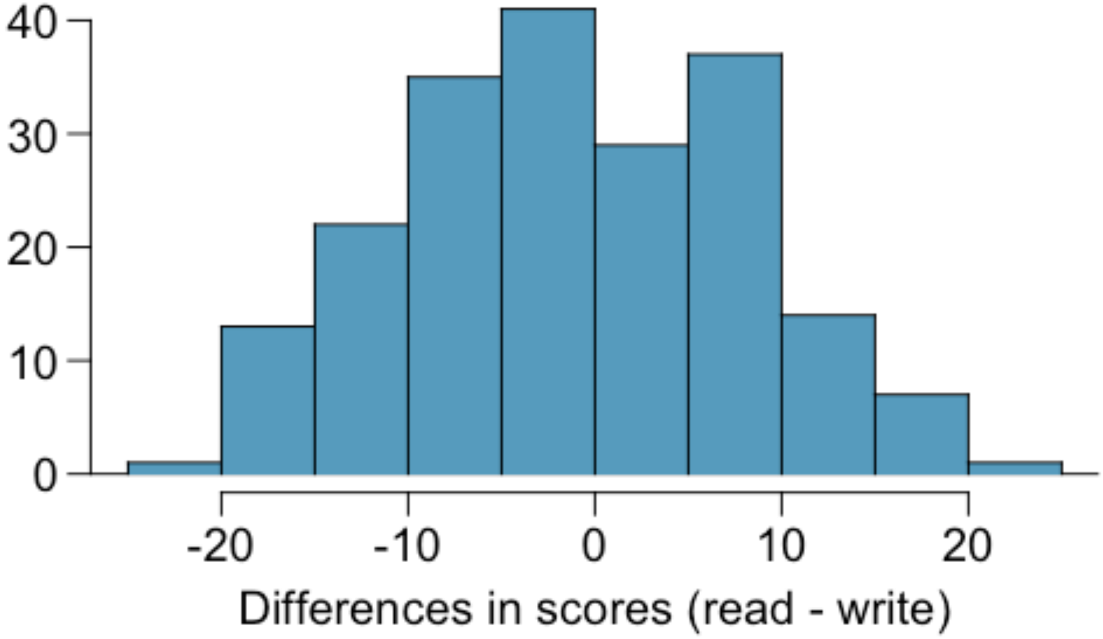
(c) Calculate the minimum required sample size for Luke.

```
z <- round(-qnorm((1-.99)/2),4)
s <- 250
ME <- 25

(z*s/ME)^2

## [1] 663.4746
```

- **n needs to be greater than 663.5 or 664**

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 there is on clear difference even though it looks like the median writing score is slightly higher than the median reading score and the writing score is less spread**

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

- **the reading and writing scores for an individual student will be dependent**

(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 average reading and writing exam scores for students**
- **H0 - there is a difference in the average reading and writing exam scores for students**

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

- **Independence - the scores of each student are independent**
- **Sample Size / skew - the samples size of 200 is greater than 30 and less than 10% of the population. the distribution does not look extremely skewed**

(e) The average observed difference in scores is $\hat{x}_{read-write} = -0.545$, 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?

```
x_hat <- -0.545
s <- 8.887
n <- 200

T <- ( x_hat - 0 ) / (s / sqrt(200))
df <- n - 1


(p_value <- 2 * pt(T, df))

## [1] 0.3868365
```

- **no - with p value of 0.386 at the .05 confidence interval we would fail to reject the null hypothesis**

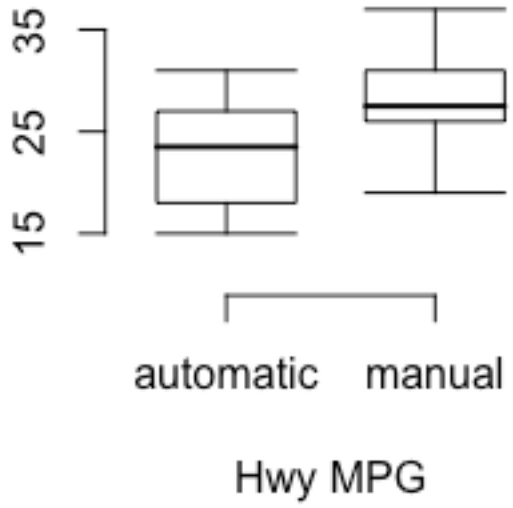
(f) What type of error might we have made? Explain what the error means in the context of the application.

- **Type 2 error - failed to rejected the null hypothesis when the alternate hypothesis is true**

(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 - the null hypothesis is that the difference between the means is 0 so we would expect 0 to be in the confidence interval**

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.



```
x_a <- 22.92
s_a <- 5.29
n_a <- 26
df_a <- n_a - 1

x_m <- 27.88
s_m <- 5.01
n_m <- 26
df_m <- n_m - 1


PE <- x_m - x_a
SE <- sqrt(s_a^2/n_a + s_m^2/n_m)

df <- min(df_a, df_m)
z <- qt(p = 0.98, df = 22)


(CI <- c(PE-z*SE, PE+z*SE))

## [1] 1.840907 8.079093
```

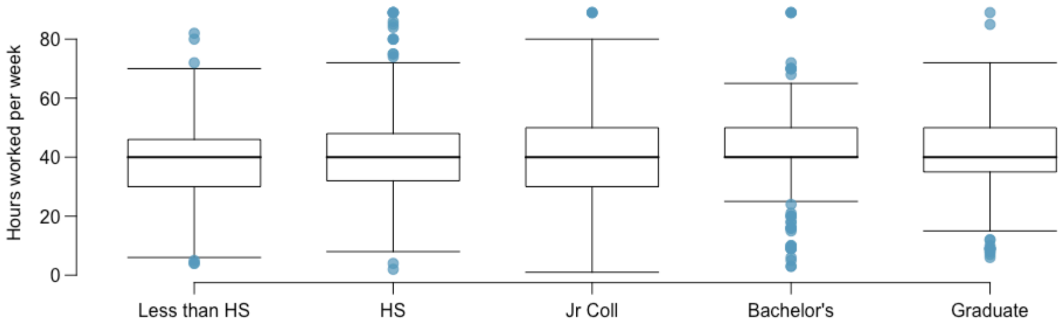
- **There difference in the average gas milage for automatic cars minus the average milage for manual cars will be between 1.184 and 8.079 with 98% confidence**

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%?

```
target <- 0.5
x <- 4
s <- 2.2
power <- 0.8

z_s <- qnorm(power)
z <- 1.96

SE <- target / (z + z_s)

(n <- (2 * s^2) / (SE^2))

## [1] 303.9164
```

- **at a confidence level or 95% they would require 304 enrollees to get the desired power level of 80%**

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.



(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

- **H0 - there is no difference in the average hours worked for individuals with Less than HS, HS, Jr Coll, Bachelor’s, Graduate**
- **H1 - there is a differnece in the average hours worked for individuals with Less than HS, HS, Jr Coll, Bachelor’s, Graduate. at least one mean is different**

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

- **independence (within / across) - each observation within the groups represents a single individual so the observations are independent. each group is mutually exclusive so the observations across groups are independent**
- **normal distribution - the data from each group approaches a normal distribution**
- **variability between groups is equal - based on the box plots its appears that the individual groups have similar variability**

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

```
one.way <- aov(hrs1 ~ degree, data = gss_sub)
summary(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
```

- **degree**
- df = 4
- sum sq = 2006
- F value = 2.189
- **residutal**
- df = 1167
- mean sq = 229.1
- **total**
- df = 1171
- sum sq = 269,388

(d) What is the conclusion of the test?

- **with a p-value of 0.0682 the is not enough evidence to reject the null hypothesis at a 0.05 level of confidence**