Chapter 7 - Inference for Numerical Data

Joshua Registe

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.

critical value for 90% confidence interval is 1.64.

Standard error can be back calculated as

```
sample_mean <- mean(c(65,77))
paste0("The sample mean is ",sample_mean)
```

```
## [1] "The sample mean is 71"
```

```
error_margin<- (sample_mean-65)/1.64
paste0("the sample error is ",round(error_margin,2))
```

```
## [1] "the sample error is 3.66"
```

```
stdev<- error_margin*sqrt(25)

paste0("The sample standard deviation is ", round(stdev,2))
```

```
## [1] "The sample standard deviation is 18.29"
```

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.

a)

```
SE = 25/1.64
n = (250/SE)^2

paste0("The sample size needs to be at least ",ceiling(n))
```

```
## [1] "The sample size needs to be at least 269"
```

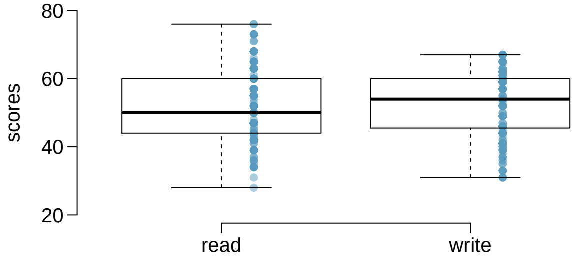
b)without calculating a sample size, the sample size to achieve a 99% confidence interval should be greater. based on the formula with standard error in the denominator decreasing

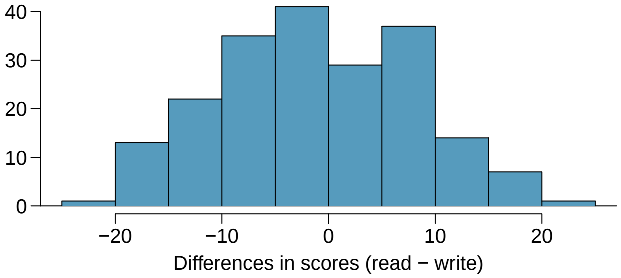
c

```
CV <- qnorm(.995)
SE = 25/CV
n = (250/SE)^2

paste0("The sample size for luke at a 95% confidence interval needs to be at least ",ceiling(n))
```

```
## [1] "The sample size for luke at a 95% confidence interval needs to be at least 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?
(b) Are the reading and writing scores of each student independent of each other?
(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?
(d) Check the conditions required to complete this test.
(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?
(f) What type of error might we have made? Explain what the error means in the context of the application.
(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.

a)There is a clear difference betwee nthe ranges and the medians of the two scores but its difficult to see if there is any significant difference betweeen the two distributions

b)Reading and writing scores are not independent of each other

c) null hypothesis: there is no difference in the average scores of students in the reading and writing exams alt hypothesis: there is a difference between the reading and writing score averages.

d)normality assumption holds true, proportions also hold true with $N>30$

e) we compute our t statistic as follows

```
SE <- 8.887/sqrt(200)

t_stat<- (-.545-0)/SE

2*pt(t_stat,199)
```

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

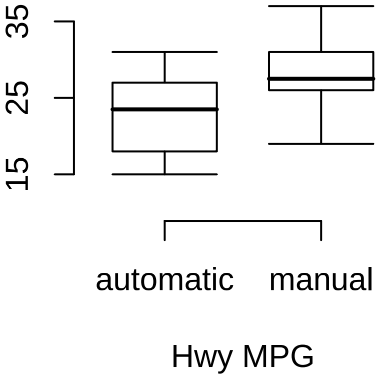
because our P value is > 0.5, there is insufficient evidence to reject the null hypothesis.

f) we may have made a type II error because we did not reject the null hypothesis although if more evididence is provided, we may be able to do so, or reduce our chances of making a type II error.

g)based on results of the hypothesis test, i would expect the confidence interval to include 0 because we did not reject that possibility.

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



```
mean_diff <- 22.92-27.88

sdiff<- 5.29-5.01

SE<- sqrt(5.29^2/26+5.01^2/26)

t_val<-qt(.99,25)

paste0("Confidence interval is represented by (", round(mean_diff- t_val*SE,2), ", ",round(mean_diff+t_
```

```
## [1] "Confidence interval is represented by (-8.51, -1.41)"
```

The interval shows that there is a difference between the mileage of the two vehicles since our interval does not contain our null hypothesis value of 0.

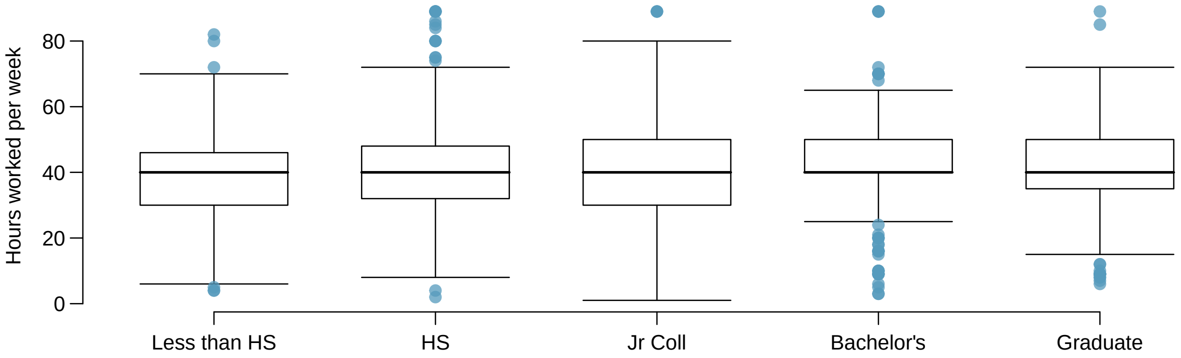
Email outreach efforts. (7.34, p. 284) A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks for information on a person's family history in regards to cancer. Another survey asks about what topics were discussed during the person's 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%?

```r
meansurv<- 4
stdsurv<-2.2

MarginError <- 0.5

SE<-MarginError/qnorm(0.9)


n <- (2.2/SE)^2

n
```

```
## [1] 31.79637
```

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.

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

(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					

(d) What is the conclusion of the test?

a) null hypothesis: There is no difference for the number of hours worked per week between educational groups alt hypothesis: there is a difference between groups for the number of hours worked

b) We can assume that the observations are independent accross groups since this is a smaller sample of a much larger population. based on the box pots, the data seems nearly normal with a slight skew on just the bachelor's level

c)

degrees of freedom = 5-1 = 4

sum sq = 2004

F value = 2.18

ResidualsDF = 1167

residuals MeanSq = 229

source: https://statpages.info/anova1sm.html

d) The conclusion of this test is that we do not reject the null hypothesis that and there is insufficient evidence to say there is a difference between groups.