Chapter 7 - Inference for Numerical Data

Samuel Kigamba

10/26/2019

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

Margin_of_Error <- ((77-65)/2)
Margin_of_Error
```

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

```
s_mean <- ((77+65)/2)
s_mean
```

```
## [1] 71
```

```
#since sample is 25, df is
df <- 25-1
tdf=round(qt(c(.05, .95), df=24)[2],3)
#using upper limit 77
serror = (77-s_mean)/tdf
serror
```

```
## [1] 3.506721
```

```
sd=serror*sqrt(25)
sd
```

```
## [1] 17.53361
```

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?

```
z.score <- 1.65
ME <- 25
SD <- 250

sample.size <- (((z.score*SD)/(ME))^2)
sample.size
```

```
## [1] 272.25
```

```
    The sample size should be 273
```

(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 Rainas, and explain your reasoning.

```
    Since we the z score for a 99% confidence is greater than that of 90% we will definitely be getting
```

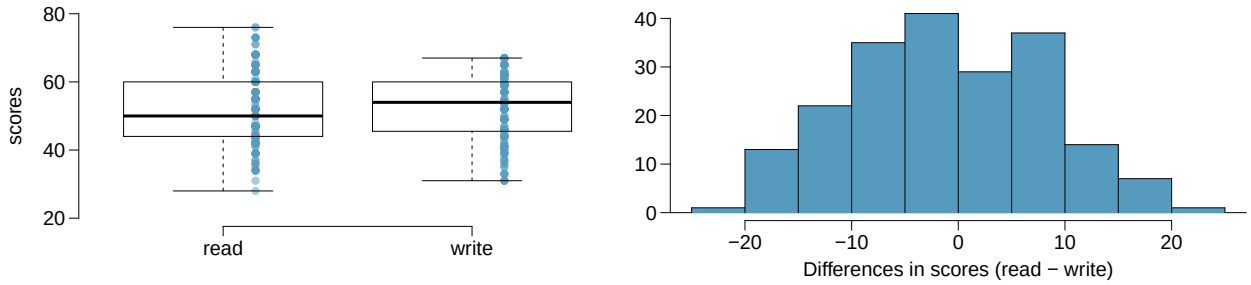
(c) Calculate the minimum required sample size for Luke.

```
zscore.Luke <- 2.58
ME <- 25
SD <- 250

samplesize.Luke <- (((zscore.Luke*SD)/(ME))^2)
samplesize.Luke
```

```
## [1] 665.64
```

```
    The sample size is 666.
```

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?

From the boxplot the mean seams abit different but the distribution of differences seems normal.

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

The scores are independent of each other although a student could have scores for both.

(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?

(H_0: mean_(read) - mean_(write) = 0) - There is no difference between the reading and writing scor
(H_A: mean_(read) - mean_(write) != equal 0) - - There is a difference between the reading and writ

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

The conditions required for this test are independence and normality.
(i) Under section (b) we noted the scores to be independent.
(ii) From the boxplot the distribution seems normal.

(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
mean.diff <- -.545
df <- n-1
SD <- 8.887
SE <- SD/sqrt(n)
T <- (mean.diff-0)/SE
pvalue <- pt(T, df)
pvalue
```

```
## [1] 0.1934182
```

Since the p value of 0.19 is greater than alpha of 0.05 we fail to reject H0 and by extension conclu

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

TypeII error - since we did not reject H0, there might have been a difference between the scores th

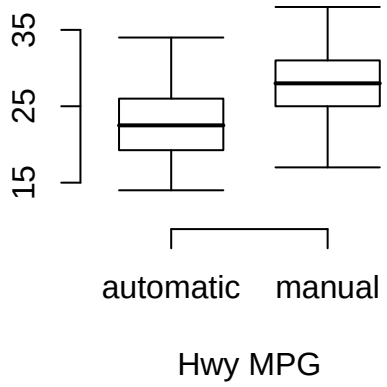
(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

Since there was no convincing evidence of a difference in average means we expect the confidence in

---

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




```
n <- 26
SD_aut <- 5.29
SD_manual <- 5.01
mean_aut <- 22.92
mean_manual <- 27.88

deg_fred <- 26 - 1

mean_diff <- mean_aut - mean_manual
mean_diff
```

```
## [1] -4.96
```

```
SE_diff <- sqrt(SD_aut^2/n + SD_manual^2/n)
SE_diff
```

```
## [1] 1.428881
```

```
t_df <- qt(.98, n-1)
t_df
```

```
## [1] 2.166587
```

```
Margin_of_Error_diff <- t_df * SE_diff
Margin_of_Error_diff
```

```
## [1] 3.095794
```

```
c(mean_diff - Margin_of_Error_diff, mean_diff + Margin_of_Error_diff)
```

```
## [1] -8.055794 -1.864206
```

Our 98% confidence interval is -8.056 and -1.864.

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 persons family history in regards to cancer. Another survey asks about what topics were discussed during the persons 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%?

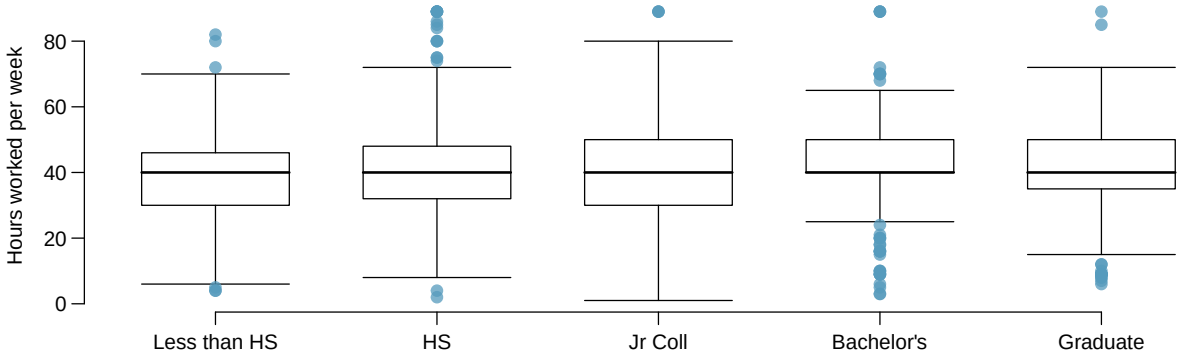
```
sd1 <- 2.2
sd2 <- 2.2


eff_size <- 0.5

z1 <- qnorm(0.975)
z2 <- qnorm(0.80)

n <- (z1 + z2)^2 * (sd1^2 + sd2^2)/eff_size^2
n
```

```
## [1] 303.9086
```

304 new enrollees are requird to achieve an effective 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.

H0: The difference of averages across all the 5 groups is equal.

HA: There is atleast one average that is not equal to the other averages.

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

we assume independence across observations within the groups.

We assume the data within each group are nearly normal.

We finally assume that the variability across the groups is about equal.

(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					

```
mu <- c(38.67, 39.6, 41.39, 42.55, 40.85)
sd <- c(15.81, 14.97, 18.1, 13.62, 15.51)
n <- c(121, 546, 97, 253, 155)
data_table <- data.frame (mu, sd, n)
n <- sum(data_table$n)
k <- length(data_table$mu)
```

Finding degrees of freedom

```
df <- k - 1
dfResidual <- n - k
dfResidual
```

```
## [1] 1167
```

Using the qf function on the Pr(>F) to get the F-statistic:

```
Prf <- 0.0682
F_statistic <- qf( 1 - Prf, df , dfResidual)
F_statistic
```

```
## [1] 2.188931
```

F-statistic = MSG/MSE

```
MSG <- 501.54
MSE <- MSG / F_statistic
MSE
```

```
## [1] 229.1255
```

MSG = 1 / df * SSG

```
SSG <- df * MSG
SSE <- 267382
```

SST = SSG + SSE, and df_Total = df + dfResidual

```
SST <- SSG + SSE
dft <- df + dfResidual
dft
```

```
## [1] 1171
```

(d) What is the conclusion of the test?

```
Since the p-value = 0.0682 is greater than 0.05, We conclude that there is not a significant differ
```