Homework 7

Irene Jacob

2020-10-16

Working backwards, Part II

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
m <- (65 + 77) / 2
paste0("Mean is ", m)
```

```
## [1] "Mean is 71"
```

```
ME <- 77-m
paste0("Margin of error  is ", ME)
```

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

```
z <- 1.645

sd <- ME*((n^0.5) / z)

paste0("Standard deviation  is ", sd)
```

```
## [1] "Standard deviation  is 18.2370820668693"
```

SAT scores

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 <- 25
sd <- 250
z <- 1.645
n <- (ME/(z*sd))^(-2)

paste0("The sample size should be: ", ceiling(n))
```

```
## [1] "The sample size should be: 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.

Luke's sample should be larger than Raina's if he wants a higher confidence in his estimate. A larger sample would give a more accurate estimate of the population.

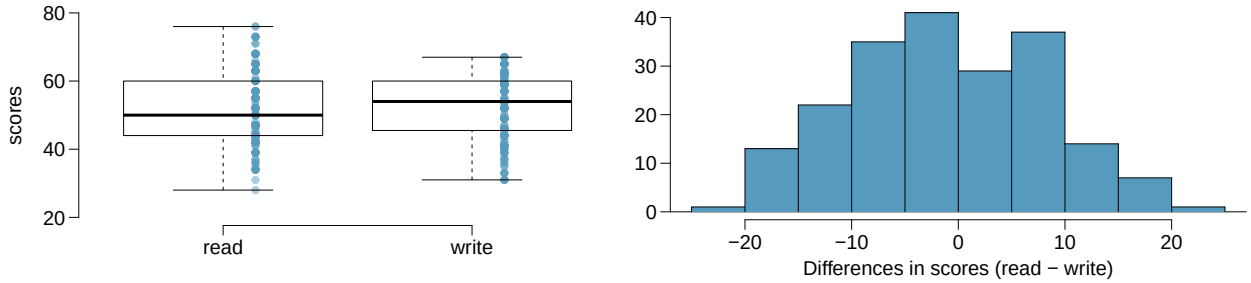
(c) Calculate the minimum required sample size for Luke.

```
ME <- 25
sd <- 250
z <- 2.576
n <- (ME/(z*sd))^(-2)

paste0("The sample size should be: ", ceiling(n))
```

```
## [1] "The sample size should be: 664"
```

High School and Beyond, Part I

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?

There is no clear difference in the average reading and writing scores.

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

The reading and writing scores of each student may not be independent of each other as it is the same student. The scores between different students are independant 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?

H0: There is no difference in the average scores of students in the reading and writing exams. HA: There is a difference in the average scores of students in the reading and writing exams.

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

The sample is random. The scores are normally distributed and independent. There are more than 10 cases.

(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?

```
mu <- -.545
n <- 200
df <- n-1
sd <- 8.887

SE <- sd/sqrt(n)

t <- (mu-0)/SE
```

```
p <- pt(t, df)
p
```

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

The p-value is greater than 0.05, so we fail to reject the null hypothesis. This means that there is not enough evidence to prove a difference between reading and writing scores.

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

We may have made a Type II error.

(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.

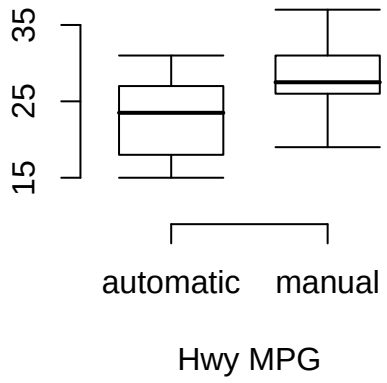
```
mu <- -.545
n <- 200
df <- n-1
sd <- 8.887
z<- 1.96

ME = z*sd/(n^0.5)

ci = mu + c(-ME,ME)
ci
```

```
## [1] -1.7766754  0.6866754
```

Fuel efficiency of manual and automatic cars, Part II

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




```
mu <- 22.92 - 27.88
SE <- (5.29^2/26 + 5.01^2/26)^0.5
Z <- 2.326
ME <- Z*SE

CI = mu + c(-ME,ME)
CI
```

```
## [1] -8.283576 -1.636424
```

Email outreach efforts

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%?

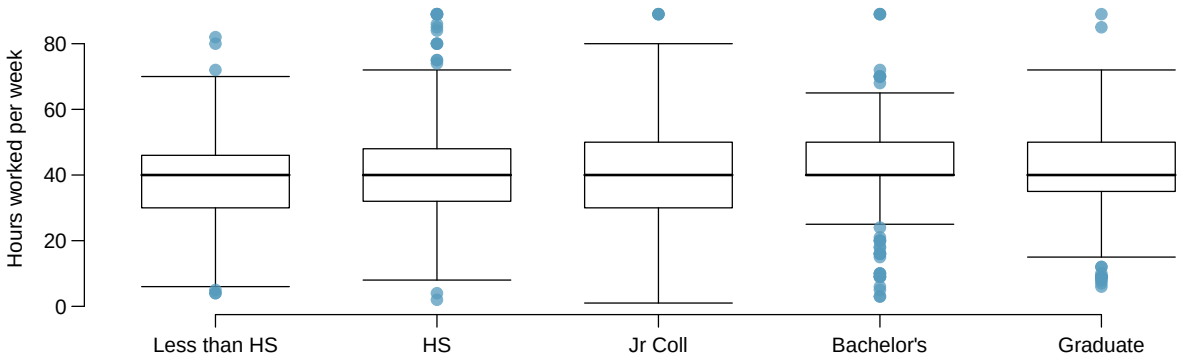
```
z <-  1.28
ME <- 0.5
sd <- 2.2

n <- round((( z * sd) / ME ) ^2 )

paste0('Sample size  is ',  n )
```

```
## [1] "Sample size  is 32"
```

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 work hours is the same within each group HA: The work hours vary across groups

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

Groups are independent as: The sample is random. There are more than 30 observations in each group. There are more than 10 cases.

(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					

```
n <- 1172
k <- 5
mean.total <- 40.45

df.total <- n - 1
df.degree <- k -1
df.residuals <- df.total -  df.degree

education.df <- data.frame(n=c(121, 546,97,253,155), sd=c(15.81,14.97,18.1,13.62,15.51)
                           , mean=c(38.67,39.6,41.39,42.55,40.85))

sum.sq.degree <- sum( education.df$n * (education.df$mean - mean.total)^2 )
sum.sq.total <-   sum.sq.degree + 267382
mean.sq.residuals<- (1 / df.residuals) * 267382
f.value.degree <- round( 501.54 / mean.sq.residuals , 4)


degree <- c(df.degree, sum.sq.degree,501.54,f.value.degree, 0.0682)
residuals <- c(df.residuals, 267382,mean.sq.residuals, NA,NA )
total <- c(df.total, sum.sq.total, NA, NA, NA)
```

```r
df <- data.frame(rbind(degree,residuals, total ))

colnames(df) <- c( 'Df','SumSq','MeanSq', 'F-value', 'Pr(>f)' )

df
```

```
##             Df      SumSq   MeanSq F-value Pr(>f)
## degree       4   2004.101 501.5400   2.189 0.0682
## residuals 1167 267382.000 229.1191      NA     NA
## total     1171 269386.101       NA      NA     NA
```

(d) What is the conclusion of the test?

The p value is greater than 0.5 so null hypothesis is not rejected.