Homework 7 : Inference for Numerical Data

Ramnivas Singh

04/22/2021

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.

Answer:

```
sampmean <- (77+65)/2
tdf <- round(qt(c(.05, .95), df=24)[2], 3)
SE <- round((77-sampmean)/tdf, 3)
sd <- SE * sqrt(25)
```

Our $\bar{x}$ is 71, the SE is 3.507 and the sample SD is roughly 17.535.

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.

Answer: (a) It would be good to use t_{df} instead of z, but since the sample size is in flux, we can't easily determine the df and use the t distribution. For safety, let's instead use the z score for the requested confidence interval.

```
# let's build a function to calculate using z scores
ztest <- function(z) {
    round((10 * z)^2, 0)
}

# estimate 90% using z
z90 <- 1.65
ztest(z90)
```

```
## [1] 272
```

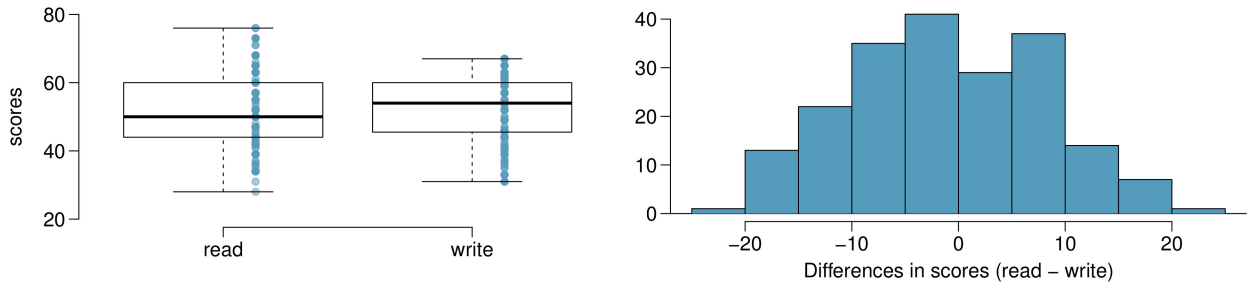
(b) Since we want need to be more confident, the required sample size will increase.

(c)

```
z99 <- 2.58
ztest(z99)
```

```
## [1] 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?
(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.

Answer :

(a) Not obviously. The means seem slightly different, but the distribution of differences looks quote normal.

(b) No. The scores of either reading or writing are independent, but one student is likely to have scores for both sides in the sample. The data appear to be paired.

(c)
H_O: $\mu_r - \mu_w = 0$, No difference between the average reading and writing scores.

H_A: $\mu_r - \mu_w \neq 0$, There is a difference.

(d) We are told that the 200 students in the sample were randomly collected. The box plots indicate a little enough skew that we can use the t distribution. Conditions are satisfied.

(e)

```
SE <- round(8.887 / sqrt(200), 3)
t <- round((-0.545-0)/SE, 3)
p <- round(2 * pt(t, 199), 3)
```

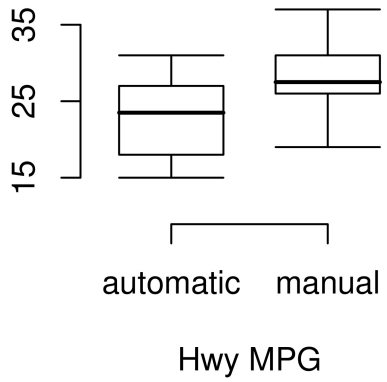
Our calculated p score is 0.386, which causes us to reject H_A. There is not convincing evidence of a difference between the average means.

(f) Since we have rejected H_A, if it were actually true, we would have made a type II error. That means that although we chose to accept H_0, there actually is convincing evidence of a difference in means. If we took a larger sample, we would decrease our chances of making a type II error.

(g) We would expect confidence intervals to include 0, as it has been determined that there is no convincing evidence of a difference in average means.

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



Answer : H0 :There is no difference between average highway mileage of manual and automatic cars. HA:There is a difference between average highway mileage of manual and automatic cars.

```
n <- 26
# Manual
mean_m <- 27.88
sd_m <- 5.01

# Automatic
mean_a <- 22.92
sd_a <- 5.29

# Standard Error
SE<- ( (sd_a ^ 2 / n) + ( sd_m ^ 2 / n) ) ^ 0.5

meandif<-mean_m-mean_a

# T value of 98% confidence interval
T<-qt(0.01,df=25)
T<--T

lower<-meandif-T*SE
higher<-meandif+T*SE
```

```
c(lower,higher)
```

```
## [1] 1.409078 8.510922
```

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%?

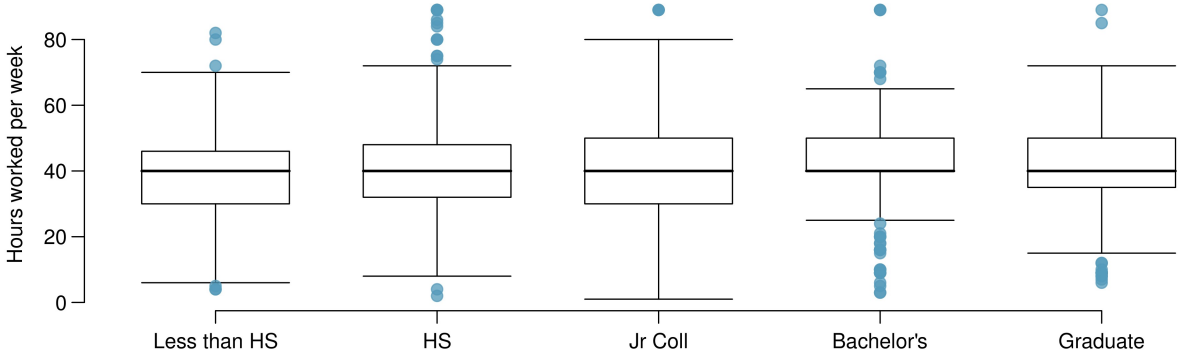
Answer: Z-score that give us a lower tail of 80% would be about Z=0.84. Additionally, the rejection region extends 1.96* SE from the center of the null distribution for sigma= 0.05. This allow us to calculate the target distance between the center of the null and alterative distributions in terms of the standard error: 0.84SE+1.96SE=2.8SE, so 0.5=2.8SE.

```
#SE<- sqrt( (2.2^2 / n) + ( 2.2^2 / n) )
n = (2.8^2/(0.5)^2)*(2.2^2+2.2^2)
n
```

```
## [1] 303.5648
```

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?

Answer:

(a) H0:Average number of hours worked is identical in all five groups HA:Average number of hours worked is various by groups

(b) The data from all 1,172 respondents are independent across groups. The data within each group are nearly normal and the variability across the groups is about equal.

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

```
mean <- c(38.67, 39.6, 41.39, 42.55, 40.85)
sd <- c(15.81, 14.97, 18.1, 13.62, 15.51)

n<-1172
k<-5
```

```r
df_G <- k - 1
dfResidual<- n - k
dfResidual
```

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

```r
Prf <- 0.0682
MSG <- 501.54
SSE <- 267382

#MSG=(1/df_G)*SSG
SSG<-MSG*df_G

#MSE=(1/dfResidual)*SSE
MSE<-SSE/dfResidual

#F value
f_value<-MSG/MSE

#SSE=SST-SSG
SST<-SSE-SSG

Df <- c(df_G, dfResidual,df_G+dfResidual)
Sum_Sq<- c(SSG, 267382, SSG+267382)
Mean_Sq <- c(501.54, round(MSE,2),NA)
F_value <- c(round(f_value,2), NA, NA)
Pr<- c(0.0682,NA,NA)

df<- data.frame(Df, Sum_Sq, Mean_Sq, F_value, "Pr>f" = Pr)
row.names(df) <- c("degree", "Residuals", "Total")
df
```

```
##              Df    Sum_Sq Mean_Sq F_value   Pr.f
## degree        4   2006.16  501.54    2.19 0.0682
## Residuals  1167 267382.00  229.12      NA     NA
## Total      1171 269388.16      NA      NA     NA
```

(d) Since the P value is 0.0682 and higher than the 0.05, we can't reject null hypothesis Average number of hours worked is identical in all five groups.