Chapter 7 - Inference for Numerical Data

Gabriel Campos

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 , p=.90 , q=1 − p=.10 $tails$=$p - \frac{q}{2}$
$degrees of freedom(df) = n - 1$ = 25-1 = 24
$t - value$=1.710882
$Margin Error$=$Standard Error \times t - value \therefore SE = ME/t - value$ =3.506963

ANSWER
$Sample Mean$= $\frac{CI_{upper}+CI_{lower}}{2}$ or $\frac{77+65}{2}$ =**71**
$Margin Error$= $\frac{CI_{upper}-CI_{lower}}{2}$ or $\frac{77-65}{2}$ =**6**
σ=$SE \times \sqrt{(n)}$ or **17.53481**

```
Q1n=25
Q1p=.9
Q1q=1-Q1p
Q1tails=Q1p+Q1q/2
Q1df<-25-1
Q1lower_CI<-65
Q1upper_CI<-77
Q1m<-(Q1upper_CI+Q1lower_CI)/2
Q1ME<-(Q1upper_CI-Q1lower_CI)/2
Q1tval<-qt(Q1tails,Q1df)
Q1tval
Q1se<-Q1ME/Q1tval
Q1se
Q1sd=Q1se*sqrt(Q1n)
Q1sd
```

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**.

σ=250 , $MarginError$= 25

(a)

Raina wants to use a **90% confidence interval**. How large a sample should she collect?
$Z - Score(90\%)$= 1.65
$\therefore n_{sample}$=$(\frac{Z-score \times \sigma}{ME})^2$ or $(\frac{1.65 \times 250}{25})^2$ = 270.5543

ANSWER:
Sample size should be 271

```
Q2sd<-250
Q2me<-25
Q2tail<-(1-.9)/2
Q2z<-qnorm(.9+Q2tail)
Q2z
Q2n<-((Q2z*Q2sd)/Q2me)^2
Q2n
```

(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.

ANSWER
A CI of 99% means the sample size needs to be bigger since n and confidence intervals have a positive correlation.

(c)

Calculate the minimum required sample size for Luke.

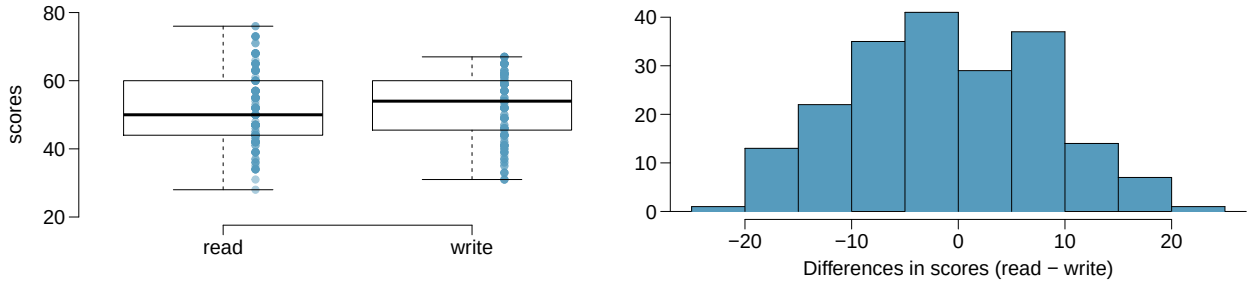
$Z - Score(99$= 2.58 4.$\therefore n_{sample}$=$(\frac{Z-score \times \sigma}{ME})^2$ or $(\frac{2.58 \times 250}{25})^2$ = 663.4897

ANSWER:
Sample size should be 664

```
Q2tail<-(1-.99)/2
Q2tail
Q2z<-qnorm(.99+Q2tail)
Q2z
Q2n<-((Q2z*Q2sd)/Q2me)^2
Q2n
```

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?
ANSWER:
No. There is not a clear difference in the average reading and writing scores.

(b)

Are the reading and writing scores of each student independent of each other?
ANSWER
Based on the similar symmetry, variability and mean, plus general context of the to attributes, I would say they are 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?
ANSWER
$H_0 := \mu_{reading.avg} = \mu_{writing.avg}$ $H_a :=\mu_{reading.avg} \neq \mu_{writing.avg}$

(d)

Check the conditions required to complete this test.
ANSWER
Data is not independent, they are actually paired, which is a fact supported by the *"Differenced in score"* histogram.

(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?

ANSWER
$p - value$ would still have to be calculated for sufficient evidence using
sd_{diff} **(8.887)**
μ_{diff} **(-.0545)**
$SE = \frac{sd_{diff}}{\sqrt{(n)}}$ =0.6284058
$t - value = \frac{\widehat{x}_{read-write}}{SE}$=-0.867274

Resulting with a $p - value$ =0.1934182

```
Q3sd <- 8.887
Q3m <- -0.545
Q3n <- 200
Q3se <- Q3sd/ sqrt(Q3n)
Q3se
Q3t_value <- (Q3m)/ Q3se
Q3t_value
Q3p_value <- pt(Q3t_value,Q3n-1)
Q3p_value
```

(f)

What type of error might we have made? Explain what the error means in the context of the application.
ANSWER:
Test might not be valid considering data is likely **NOT** independent.

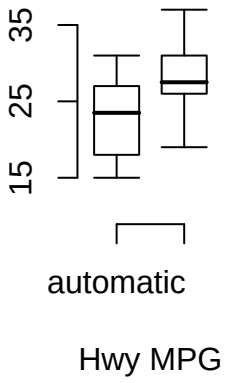
(g)

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
Makes sense for it to be 0 since we've noted close to no difference in the data.

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



$\mu_{diff} = \mu_{auto} - \mu_{manual}$
$SE_{diff} = \sqrt{(\frac{sd^2_{auto}}{n_{auto}} + \frac{sd^2_{manual}}{n_{manual}})}$

$Z - score = 2.326348$
CI_{lower}=$\mu_{diff} - SE_{diff} \times Z - score$=-8.284074
CI_{upper}=$\mu_{diff} + SE_{diff} \times Z - score$=-1.635926

ANSWER The CI doesn't contain 0, which is whats expected if no difference exist between manual and automatic ∴ a difference must exist.

```
an<-26
am<-22.92
asd<-5.29
mn<-26
mm<-27.88
msd<-5.01
mdiff<-am-mm
mdiff
sediff<-sqrt((asd^2/an)+(msd^2/mn))
sediff
Q4tail<-(1-.98)/2
```

```
Q4tail
Q4z<-qnorm(.98+Q4tail)
Q4z
Q4low_CI<-mdiff-sediff*Q4z
Q4low_CI
Q4up_CI<-mdiff+sediff*Q4z
Q4up_CI
```

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%?

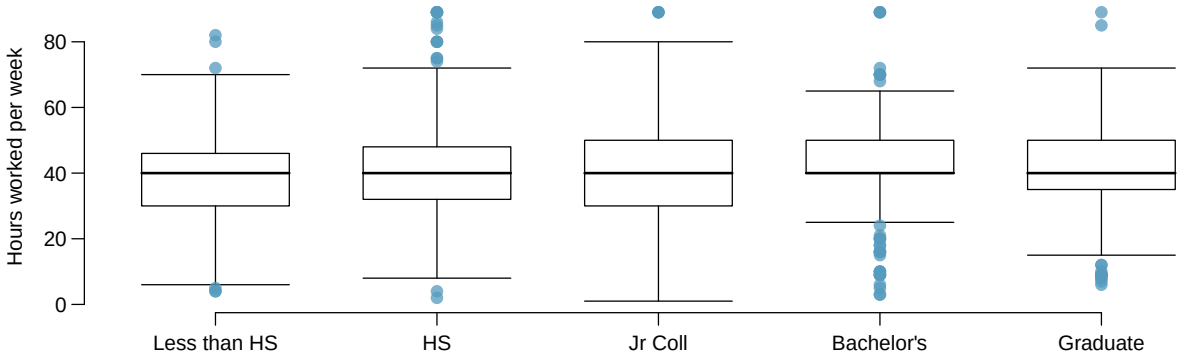
ANSWER:
$Z - score$=1.28 & $n = \frac{z-score \times sd}{ME}$=5.638827
$\therefore$ 6 enrollees is the minimum needed.

```
Q5m<- 4
Q5sd <- 2.2
Q5me <- 0.5
Q5tail<-(1-.8)/2
Q5tail
Q5z <- qnorm(.8+Q5tail)
Q5z
Q5n <- ((Q5z * Q5sd) / Q5me)
Q5n
```

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.

ANSWER
H_0: $\mu_{5groups} = \mu_{5groups}$ H_A: $\mu_{5groups} \neq \mu_{5groups}$

(b)

Check conditions and describe any assumptions you must make to proceed with the test.
ANSWER
Sample is taken randomly ✓, with independent observations✓ and n<30 is sufficient in size✓.

(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	4	2006.16	501.54	2.188931	0.0682
Residuals	1167	267,382	229.1255		
Total	1171	269388.2			

(d)

What is the conclusion of the test?

The high $p - value$ indicates that we cannot reject the Null hypothese and $\therefore$ cannot conclude there are no differences in the group.

```r
Q6m_list <- c(38.67, 39.6, 41.39, 42.55, 40.85)
Q6sd_list <- c(15.81, 14.97, 18.1, 13.62, 15.51)
Q6n_list <- c(121, 546, 97, 253, 155)
Q6table <- data.frame (mu = Q6m_list,sd = Q6sd_list,n = Q6n_list)
Q6n <- sum(Q6table$n)
Q6n
Q6k <- length(Q6table$mu)
Q6k
# DF x degree
Q6df <- Q6k - 1
Q6df
Q6residual <- Q6n - Q6k
#Df x Residual
Q6residual
# F-statistic using Pr(>F)
Prf <- 0.0682
f_statistic <- qf( 1 - Prf, Q6df , Q6residual)
#F-value x degree
f_statistic
# F-statistic = MSG/MSE
msg <- 501.54
mse <- msg / f_statistic
#Mean Sq. x Residuals
mse
# MSG = 1 / df * SSG
ssg <- Q6df * msg
#Sum Sq x degree
ssg
sse <- 267382
# SST = SSG + SSE, and df_Total = df + dfResidual
sst <- ssg + sse
#Sum Sq x Total
sst
dft <- Q6df + Q6residual
#Df Total
dft
```