Chapter 7 - Inference for Numerical Data

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

The sample mean would be exactly in the middle between the upper and lower bounds of the confidence interval.

```r
s <- 25
# sample mean
mean = (65 + 77) / 2
paste0("Mean is ", mean)
```

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

```r
# margin of error
margin_error = 77 - mean
paste0("Margin of error  is ", margin_error)
```

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

```r
#  t-score
#df = s - 1 = 24
t <- round(qt(.05, df=24),3)
paste0("T score is ", t)
```

```
## [1] "T score is -1.711"
```

```r
# standard deviation
sd <- round(((margin_error  / t ) * sqrt(s)),3)
paste0("Standard deviation  is ", sd)
```

```
## [1] "Standard deviation  is -17.534"
```

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)

```
# 90% CI
z <- 1.65
n <- round(((z^2) / (25^2)) * 250^2)
paste0("Sample size for Raina is ", n)
```

```
## [1] "Sample size for Raina is 272"
```

(b)

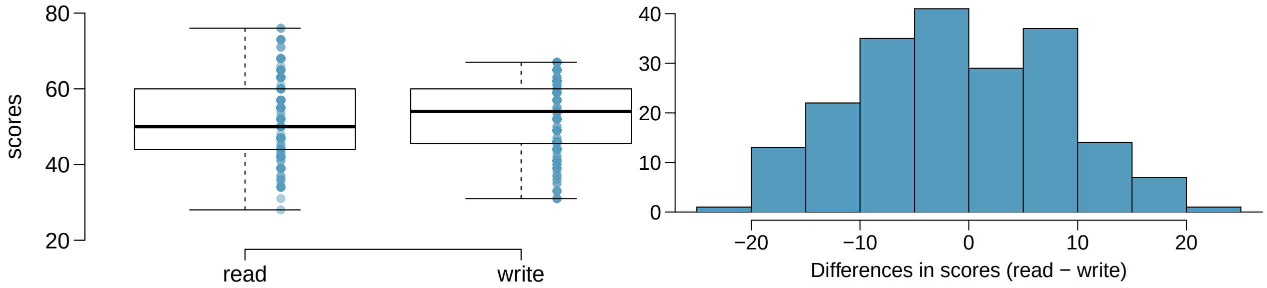
Luke's sample should be larger than Raina's if he wants a higher confidence in his estimate because a larger sample would give a more accurate estimate of the poopulation.

(c)

```
# 99% CI
z <- 2.58
n <- round(((z^2) / (25^2)) * 250^2)
paste0("Sample size for Luke is ", n)
```

```
## [1] "Sample size for Luke 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?
(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)

There is no clear difference in the average reading and writing scores from boxplots and histogram. The distribution looks normal centered around zero and the boxplots have similar medians and variances.

(b)

Yes,the reading and writing scores of each student are independent of each other.

(c)

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)

The difference in the average scores of students in the reading and writing exams were taken from a random sample and may be less than 10% of the population. Both are independent of each other. They seems to be normal distribution with no skew and the sample size is greater than 30. So, the normality condition is satisfied.

(e)

```
# calculate p value
mu <- -.545
df <- n-1
sd <- 8.887
n <- 200

Std_err <- sd/sqrt(n)
```

```r
t <- (mu-0)/Std_err

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

```
## [1] 0.1930524
```

The p-value is greater than 0.05, so we fail to reject the null hypothesis. There is not enough statistical convincing evidence of a difference between reading and writing scores.

(f)

We may have get a Type II error in rejecting the alternative hypothesis and wrongly concluded that there is no a difference in the average reading and writing scores.

(g)

I do expect a confidence interval for the average difference between the reading and writing scores to include 0. Due to p-value we fail to reject the null hypothesis.

```r
Std_err <- 8.887/sqrt(200)
cp <- qt(p=(.05/2), df=199, lower.tail=FALSE)
cp
```

```
## [1] 1.971957
```

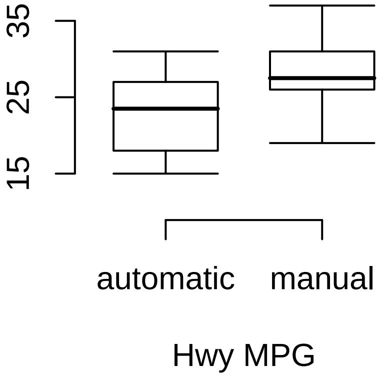
```r
avg_read_minus_write <- -0.545
c(avg_read_minus_write - Std_err * cp, avg_read_minus_write + Std_err * cp)
```

```
## [1] -1.7841889  0.6941889
```

95% confidence interval includes zero: (-1.7841889 0.6941889)

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

```
auto_mu <- 16.12
auto_s <- 3.58
man_mu <- 19.85
man_s <- 4.51
n <- 26

diff <- man_mu - auto_mu

# standard error
se <- sqrt((auto_s^2/n) + (man_s^2/n))

# T-score
t <- (diff - 0)/se
t
```

```
## [1] 3.30302
```

```
p = pt(q=t, df=n-1, lower.tail = FALSE)
p
```

```
## [1] 0.001441807
```

Since the p-value is 0.00144, there is strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage.

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%?

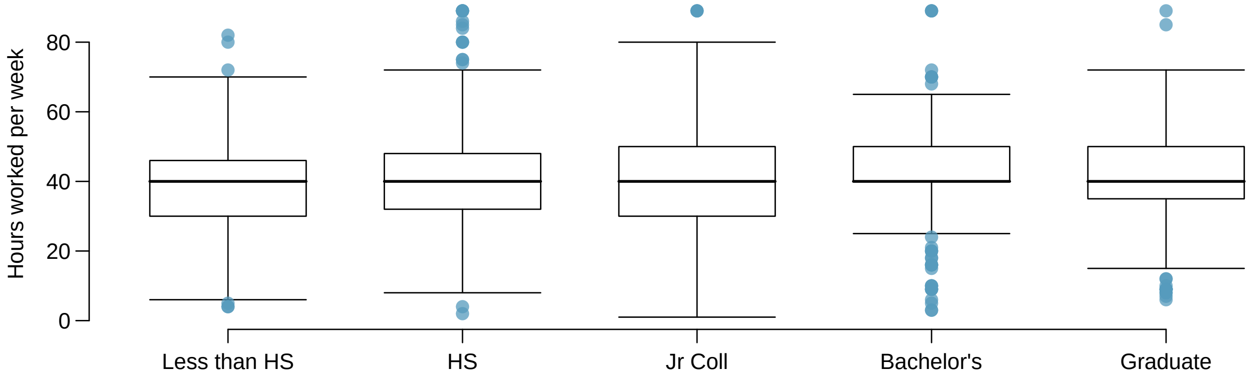
Answer

```r
# 80% CI
t <-  1.28  # 80% Confidence interval
ME <- 0.5
SD <- 2.2

n <- round((( t * SD) / ME ) ^2 )

paste('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.
(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: The mean hours worked is the same across all educational attainment groups. HA: There is a difference in the mean hours worked in at least one group.

(b)

The sample is assumed to be randomly selected and is less than 10% of the population and has greater than 30 observations in each group, so it can be consider to be independent. The boxplots appear to be mostly normal distribution. The Box plots and mean and SD value are somewhat similar between groups, so variability is about equal.

(c)

```r
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)

#final table
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)

df <- data.frame(rbind(degree,residuals, total ))
# name of columns
colnames(df) <- c( 'Df','SumSq','MeanSq', 'F-value', 'Pr(>f)' )
# put in table format
knitr::kable(df,  row.names = T)
```

	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)

The p-value is (0.0682) greater than 0.05, therefore we do not reject the null hypothesis and conclude that there is no significant difference between these 5 groups.