Chapter 7 - Inference for Numerical Data

Cameron Smith

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

Please see output of code block below.

```r
merror <- (77 - 65) / 2
mean <- 65 + merror
n <- 25
z <- abs(qt(.05, n - 1))
se <- (77 - mean) / z
sd <- se * sqrt(n)

cat("The sample mean is:", mean, "\nThe margin of error is:", merror, "\nThe sample sd is:", round(sd, 2
```

```
## The sample mean is: 71
## The margin of error is: 6
## The sample sd is: 17.53
```

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) The minimum sample size is 271
b) His sample size will need to be higher (will need a bigger fishing net)
c) The minimum sample size is 663

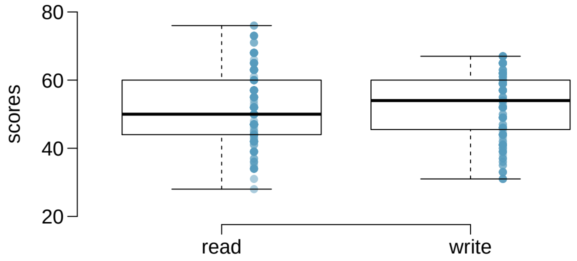
```
sd <- 250

# Calculation for a
z <- qnorm(.95, 0, 1)
round(((sd * z) / 25) ^2, 0)
```

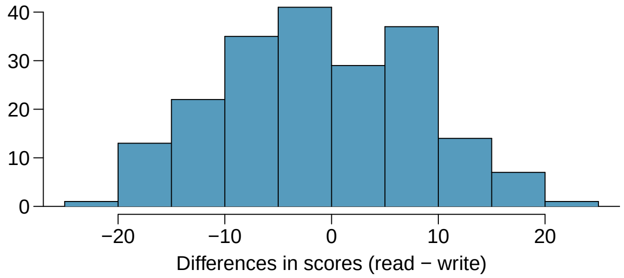
```
## [1] 271
```

```
# Calculatoin for c
z <- qnorm(.995, 0, 1)
round(((sd * z) / 25) ^2, 0)
```

```
## [1] 663
```

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

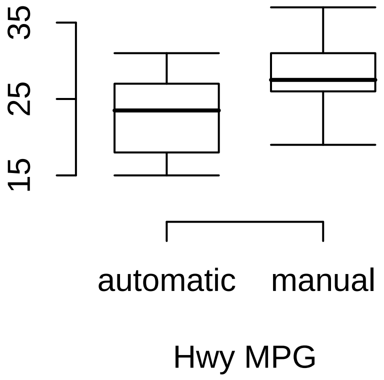
a) There is a clear difference in the mean, but it's not clear (at this point) whether the difference is statistically significant.
b) No, they are dependent as they are for the same student
c) Null hypothesis :There is no difference in the average scores of students in the reading and writing exam.
d) Independence (yes, both observations are based on an independent sample). Normality (yes, n = 200 and none of the outliers seem particularly extreme based on the box plots).
e) No, the p-value is .19 which is significantly higher than .05 and thus the difference is not considered statistically significant.
f) There could potentially be a type2 error, which means we failed to reject the null hypothesis.
g) Yes I would. In my (relatively short) experience when differences fail to be statistically significant 0 falls within the confidence interval, which means there's a good chance that the actual difference is actually 0.

```r
# Answer for e
sd <- 8.887
mean <- -.545
n <- 200
se <- sd / sqrt(n)
t <- mean / se
p <- pt(t, n - 1)
cat("The P-value is", round(p, 3))
```

```
## The P-value is 0.193
```

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

The interval is from -8.06 to -1.86 per the calculations below, which seems to indicate a statistically significant difference, i.e auto cars have lower gas mileage than manual.

```r
n_auto <- 26
n_manual <- 26
mean_auto <- 22.92
mean_manual <- 27.88
sd_auto <- 5.29
sd_manual <- 5.01
pt_estimate <- mean_auto - mean_manual
se_diff <- sqrt(((sd_auto^2) / n_auto) + ((sd_manual^2) / n_manual))
df <- n_auto - 1
t_value <- qt(.98, df)

lowertail <- round(pt_estimate - t_value * se_diff, 2)
uppertail <- round(pt_estimate + t_value * se_diff, 2)

cat("The 98% confidence interval is from", lowertail, "to", uppertail)
```

```
## The 98% confidence interval is from -8.06 to -1.86
```

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%?

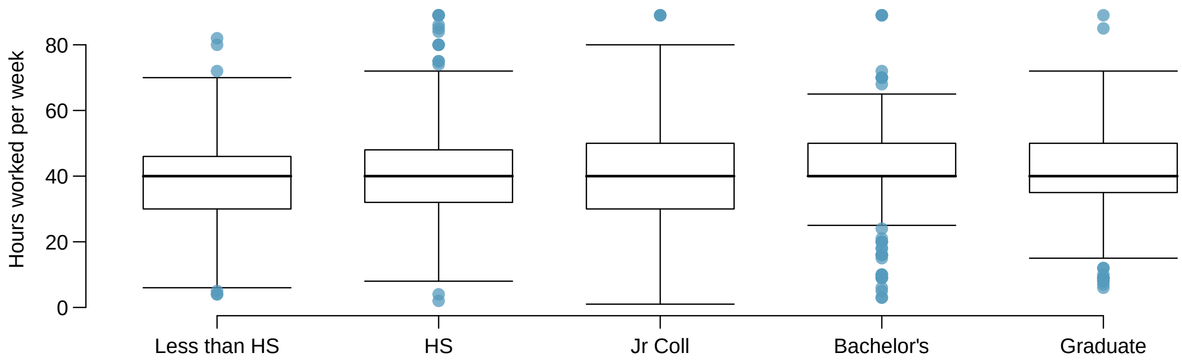
ANSWER

```r
mean <- 4
sd <- 2.2
effsize <- .5
powlevel <- .8
zscore_80 <- qnorm(.8)
zscore_5 <- qnorm(.975)
sample_size <- 2*((zscore_5 + zscore_80)^2) * 2.2^2/.5^2

cat("The sample size required for the desired power level is:", round(sample_size))
```

```
## The sample size required for the desired power level is: 304
```

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) Null hypothesis: the average number of hours worked does not vary across the 5 groups. Alternative hypothesis: the average number of hours worked varies across the 5 groups.
b) Independent within and across groups (assuming yes), data within each group are nearly normal (seems yes, all have n > 30 without too many outliers), and variability across the groups is about equal (yes, based on the box plots). All 3 conditions for ANOVA are satisfied.
c) Please see output of code block below.
d) The p-value is .07, which is greater than .05 and thus must reject the null hypothesis.

```
n_total <- sum(121, 546, 97, 253, 155)
k <- 5
df_deg <- k - 1
df_res <- n_total - k
```

```r
df_tot <- df_deg + df_res
ms_deg <- 501.54
ssq_res <- 267382
ssq_deg <- df_deg * ms_deg
ms_res <- ssq_res / df_res
f_deg <- ms_deg / ms_res
f_value <- ms_deg / ms_res
p_value <- pf(f_value, df_deg, df_res, lower.tail = FALSE)

cat("Df_degree:\t\t\t", df_deg,
    "\nDf_residuals:\t\t", df_res,
    "\nSumSq_degree:\t\t", ssq_deg,
    "\nMeanSq_Residuals:\t", ms_res,
    "\nF value_degree:\t\t", f_deg,
    "\n\nP-value:\t\t\t", round(p_value, 2))
```

```
## Df_degree:            4
## Df_residuals:         1167
## SumSq_degree:         2006.16
## MeanSq_Residuals:     229.1191
## F value_degree:       2.188992
##
## P-value:          0.07
```