Chapter 7 - Inference for Numerical Data

Lab completed by Chad Bailey

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.

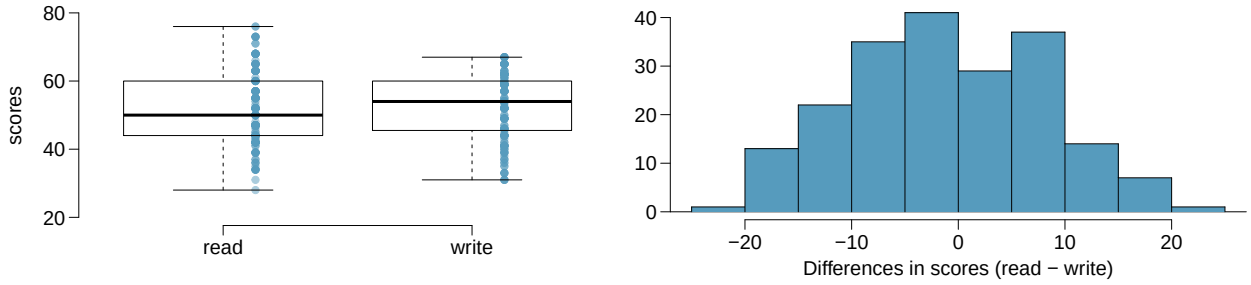
```
##Student response to problem 5.24

## known values:
      confidence  <- 0.90
      ci_lower    <- 65
      ci_upper    <- 77
      n           <- 25

## known formulas:
##    CI = xbar +/- ME
##    ME = t_df * SE
##    SE = s / sqrt(n)
##    df = n - 1

## derived values:
      me      <- (ci_upper - ci_lower) / 2  ## = (77-71)/2        = 6
      xbar    <- ci_lower + me              ## = 65 + 6           = 71

      p       <- (1-confidence)/2           ## = (1-0.90)/2       = 0.05
      df      <- n-1                        ## = 25-1             = 24
      t_df       <- abs(qt(p, df))          ##                    = 1.711

      se      <- me/t_df                    ## = 6/1.711          = 3.507
      s       <- se * sqrt(n)               ## = 3.507 * sqrt(25) = 17.53


## final answers:
##    sample mean     = 71
##    margin of error = +/- 6
##    sample sd       = 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.

```
##Student response to problem 7.14

## known values:
      sd           <- 250
      me           <- 25
      confidence1  <- 0.90
      confidence2  <- 0.99


## known formulas:
##    ME = z * SE
##    SE = s / sqrt(n)


## derived values:
      p_1       <- (1-confidence1)/2      ## = (1-0.09)/2          = 0.05
      z_1       <- abs(qnorm(p_1))        ## =                     = 1.64
      n_1       <- (sd * z_1 / me)^2      ## = (250 * 1.64 / 25)^2 = 269

      p_2       <- (1-confidence2)/2      ## = (1-0.99)/2          = 0.005
      z_2       <- abs(qnorm(p_2))        ## =                     = 2.58
      n_2       <- (sd * z_2 / me)^2      ## = (250 * 2.58 / 25)^2 = 666

## Final answers:
##  (a) For 90% confidence, Riana should use a sample of at least 269
##  (b) Luke's sample will need to be larger because he desires a higher
##      level of confidence
##  (c) For 99% confidence, Luke should use a sample of at least 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.

```
## Student response to problem 7.20

##  (a) No, there is not a clear difference between scores. There is an
##      observable difference but it is not clear if it is meaningful.

##  (b) NO, although the scores are from seperate tests they are for the same students.

##  (c) H0: The difference in average scores between reading and writing is 0.
##      HA: The difference in average scores between reading and writing is not 0.

##  (d) condition1: independence: No. Therefore use paired analysis.
##      condition2: sample size >= 30: yes; this sample has 200 cases
       nrow(hsb2)
```

```
## [1] 200
```

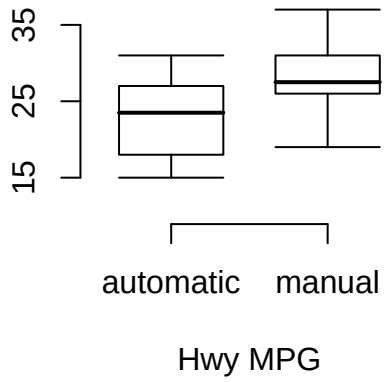
```
##  (e) No, these do no provide convincing evidence of a difference betweeen
##      average scores. The given the parameters of this distribution it would result
##      in a p-value of 0.39, which is greater than 0.05. Thus we would fail to reject
##      the null hypothesis.
       2*pnorm(-0.545, mean = 0, sd = 8.887/sqrt(nrow(hsb2)))
```

```
## [1] 0.3857919
```

```
##  (f) We could be making a Type II error. Failing to reject the null hypothesis (H0)
##      when there may be a real difference between reading and writing scores.
##
##  (g) Yes, based on this sample and the resulting hypothesis test, it would be
##      expected that a confidence interval would include 0. The p-value was below the
##      critical value by a large degree. Thus is would be expected that 0 would be
##      well within the confidence interval.
```

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



```
## Student response to problem 7.28

## known values:
  auto_mean   <- 22.92
  auto_sd     <- 5.29
  auto_n      <- 26

  man_mean    <- 27.88
  man_sd      <- 5.01
  man_n        <- 26

  confidence  <- 0.98


## known formulas:
##     ci = pe +/- t* x SE
##     SE = sqrt((sd_1^2)/n_1 + (sd_2)^2/n_2)
##     df = min((n_1 - 1), (n_2 - 1))
```

```r
## derived values:
      pe         <- man_mean - auto_mean

      p          <- (1-confidence)/2
      df         <- min(man_n -1, auto_n - 1)
      t          <- abs(qt(p = p, df = df))

      se         <- sqrt((man_sd^2)/man_sd + (auto_sd^2)/auto_n)

      ci_lower   <- pe - t*se
      ci_upper   <- pe + t*se


## Final answers:
##     (a) confidence interval: (-1.17, 11.09)
##     (b) interpretation: The confidence interval includes zero, so the
##                         evidence is not strong enough to conclude that
##                         the difference is real and not due to measurement
##                         error.
```

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%?

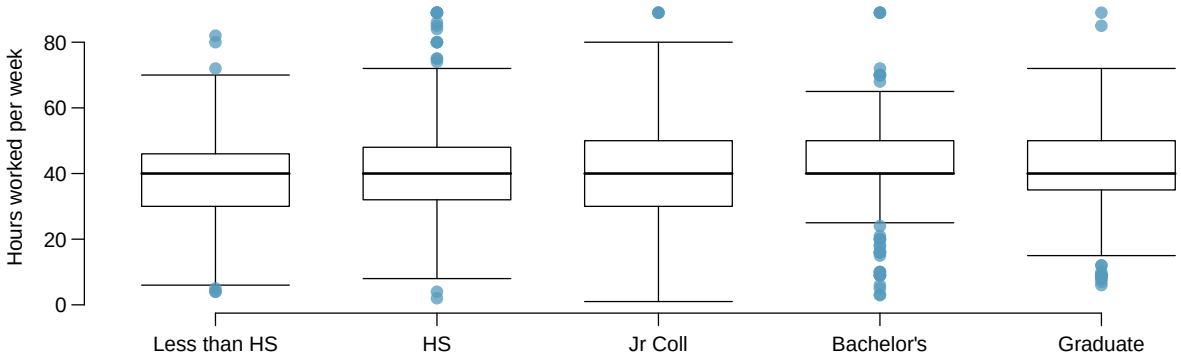
```
## Student response to problem 7.34

## known values:
  curr_mean   <- 4
  curr_sd     <- 2.2

  eff_size    <- -0.5
  pwr         <- 0.80

## assumed values:
  ## sample is greater than 30
  ## confidence is 95%
  ## alt_sd = curr_sd = 2.2


## known formulas:
##    SE = sqrt((sd_1^2)/n_1 + (sd_2^2)/n_2)
##    x  = z_confidence * SE
##    power  = (x - eff_size) / SE

## derived values:
      z_conf         <- qnorm((1-0.95)/2)   ## -.96

##    power = ((z_confidence * SE) - eff_size) / SE
##    power = (z_confidence * SE)/SE - eff_size/SE
##    power = z_confidence - eff_size / SE
##    eff_size / SE = z_confidence - power
##    eff_size = SE * (z_confidence - power)
##    SE = eff_size / (z_confidnece - power)
      se <- eff_size / (z_conf - pwr)

##    SE = sqrt(curr_sd^2 / n + curr_sd^2 / n)
##    SE^2 = (2*curr_sd^2) / n
##    n = (2 * curr_sd^2) / SE^2
      n <- (2* curr_sd^2) / se^2
      n
```

```
## [1] 294.9458
```

```
## Final answers:
##    A sample of at least 295 in each interface would be needed to detect
##    an effect size of 0.5 with a power of 80%.
```

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?

```
## Student response to work and education

##  (a) H0: The mean hours worked are the same for each education group.

##  (b) Conditions for an ANOVA test:
##      (1) independent observations: condition is met as each observation is a
##                                    seperate individual
##      (2) data are nearly normal within each group: condition appears to be met
##                                                    by visual inspection of the
##                                                    box plots
##      (3) variability across groups is about equal: it is assumed the difference
##                                                    in variance is not too great

##  (c) filling in the ANOVA output
##                    Df    Sum Sq    Mean Sq     F value     Pr(>F)
```

```
##         Degree      4     2006      501.54       2.189       0.0682
##         Residuals   1167  267382    229.12


##  (d) The conclusion of the test is that there is not sufficient evidence to reject
##      H0. Thus we continue with the assumption that all means are equal.
```