Chapter 7 - Inference for Numerical Data

Chunjie Nan

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:

```
n <- 25
x1 <- 65
x2 <- 77

smean <- (x1 + x2) / 2
smean
```

```
## [1] 71
```

```
me <- (x2 - x1) / 2
me
```

```
## [1] 6
```

```
df <- 25 - 1
p <- 0.9
p_tails <- p + (1 - p)/2

t <- qt(p_tails, df)
se <- me / t
sd <- se * sqrt(n)
sd
```

```
## [1] 17.53481
```

Answer: The sample mean is 71, margin of error is 6, and the sample standard deviation is 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?

```
z <- 1.65
me <- 25
sd <- 250

n <- ((z * sd) / me ) ^ 2
n
```

```
## [1] 272.25
```

Answer: She needs at least 273 samples for 90% confidence interval.

(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: Luke needs larger sample to determine 99% confidence interval because according to the function, z score is higher for Luke at the numerator, as well as the number of sample size.

(c) Calculate the minimum required sample size for Luke.

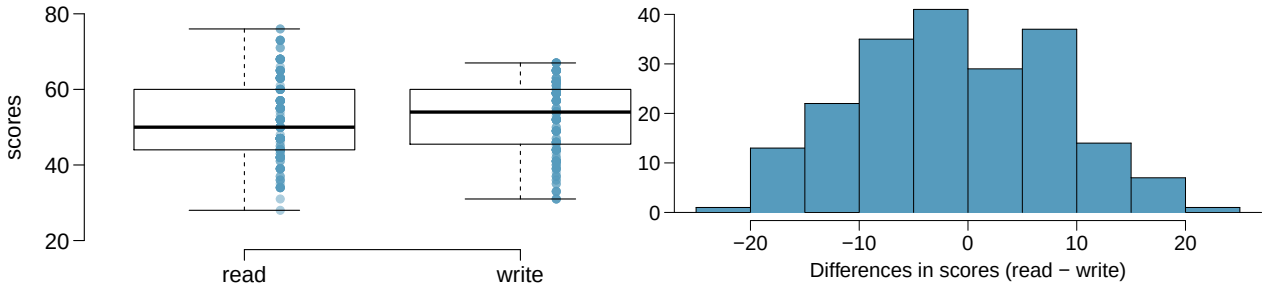
```
z <- 2.575
me <- 25
sd <- 250

n <- ((z * sd) / me ) ^ 2
n
```

```
## [1] 663.0625
```

Answer: Luke needs at least 664 sample for determine 99% confidence interval

---

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 much differences between reading and writing.

(b) Are the reading and writing scores of each student independent of each other? Answer: The student might be independent each other, but not sure about the writing and reading scores.

(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: H0: The difference of average in between reading and writing equal zero.

HA: The difference of average in between reading and writing are not equal to zero.

(d) Check the conditions required to complete this test.

Answer: Independence of observations

normal distribution: The box plot shows the data is normally distributed without outliers exist.

(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?

```
sd_dif <- 8.887
mu_dif <- -0.545
n <- 200

se_dif <- sd_dif / sqrt(n)

# Compute T statistic
t <- (mu_dif - 0) / se_dif

df <- n - 1

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

```
## [1] 0.1934182
```

Answer: The p-value is greater than 0.05, this implies that there is no evidence to show the difference of student's reading and writing exam scores.

(f) What type of error might we have made? Explain what the error means in the context of the application.

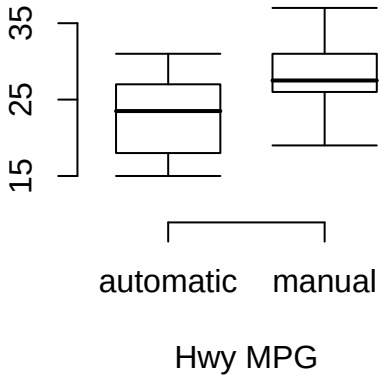
Answer: In the case, we may have made a type II error by rejecting the alternative hypothesis HA which is Incorrectly reject the alternative hypothesis.

(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: The confidence interval for the average difference between reading and writing scores should include 0 because when we make hypothesis test, it indicates that the difference is not in one side or another.

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




```
SE <- sqrt((5.29)^ 2 / 9 + (5.01)^ 2 / 9)
Average_difference<- 22.92 - 27.88
# Calculate T-score
df <- 9 -1
p <- 0.98
x <- p + (1 - p) / 2
t <- qt(x, df)
Low <- Average_difference - t * SE
Up <- Average_difference + t * SE
c(Low, Up)
```

```
## [1] -11.994429   2.074429
```

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%?

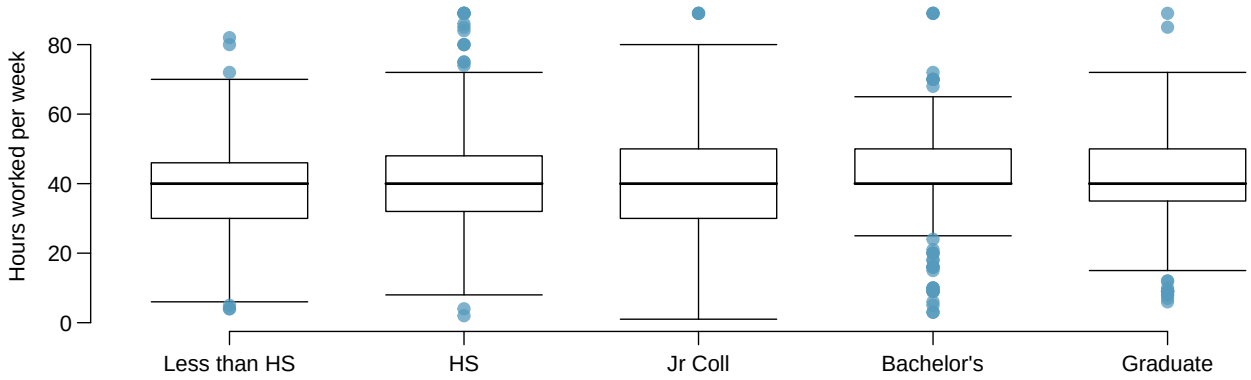
```
z <- 1.28
sd <- 2.2
me <- 0.5
n <- ((z * sd) / me)^ 2
n
```

```
## [1] 31.71942
```

Answer: They need at least 32 new enrollees to make the desired power level to 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.

Answer: HO: The average number of hours worked are the same for the five groups. HA: The average number of hours worked are not the same for the five groups.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

Answer: We assume the sample is randomly choosed, and the sample size is less than the 10% of the population and the sample size is large enough to proceed 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					

```
df_degree <- 5 - 1
df_degree
```

```
## [1] 4
```

```
Sample_size <- 1172
Mean_degree <- 501.54
Sum_resident <- 267382
df_resident <- Sample_size - 5
df_resident
```

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

```
df_total <- df_degree + df_resident
df_total
```

```
## [1] 1171
```

```
Sum_degree <- df_degree * Mean_degree
Sum_degree
```

```
## [1] 2006.16
```

```
Sum_total <- Sum_resident + Sum_degree
Sum_total
```

```
## [1] 269388.2
```

```
Mean_resident <- Sum_resident / df_resident
Mean_resident
```

```
## [1] 229.1191
```

```
F_value <- Mean_degree / Mean_resident
F_value
```

```
## [1] 2.188992
```

(d) What is the conclusion of the test?

Answer: The p-value shows as 0.0682 which is greater than 0.05 for 95 % confidence interval. This means that we would fail to reject our HO.