DATA606_HW7_RJM

RJM

2019-11-04

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.

5.24

```
# The sample mean is 71.

s_mean <- (65 + 77) / 2

s_mean
```

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

```
# The margin of error is 6.
s_me <- (77 - 65) / 2

s_me
```

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

```
s_n <- 25

s_df <- s_n - 1

s_p <- 0.9

s_p2 <- s_p + (1 - s_p) / 2

s_t_score <- qt(s_p2, s_df)

s_t_score
```

```
## [1] 1.710882
```

```
s_se <- s_me / s_t_score

s_sd <- s_se * sqrt(s_n)

s_sd
```

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

```
# The standard deviation 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.

7.14 (a)

(a) Raina wants to use a 90% confidence interval. How large a sample should she collect?

```
# For 90% interval z-value will be 1.65.

sat_z_value_a <- 1.65

sat_me <- 25

sat_sd <- 250

sat_n <- ceiling(((sat_z_value_a * sat_sd) / sat_me)^2)

sat_n
```

```
## [1] 273
```

```
# The sample size should be 273.
```

7.14 (b)

(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.

The sample size would be larger as a larger z-value will be used in the denominator of the formula.

7.14 (c)

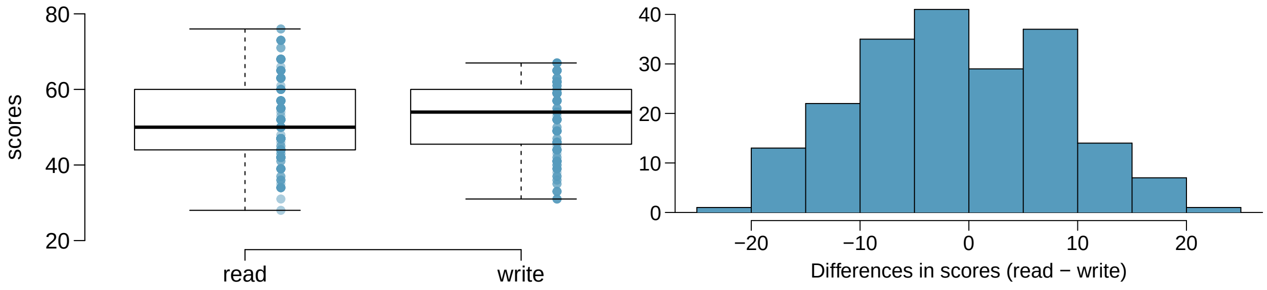
(c) Calculate the minimum required sample size for Luke.

```
# For 90% interval z-value will be 2.575.

sat_z_value_b <- 2.575

sat_me <- 25

sat_sd <- 250

sat_n <- ceiling(((sat_z_value_b * sat_sd) / sat_me)^2)

sat_n
```

```
## [1] 664
```

```
# The sample size should be 664.
```

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.



7.2 (a)

(a) Is there a clear difference in the average reading and writing scores?

There is no clear difference in the average reading and writing scores. The samples seem to be normally distributed with a very slight right skew.

7.2 (b)

(b) Are the reading and writing scores of each student independent of each other?

The scores seem to be independent of each other.

7.2 (c)

(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?

H_0: There is no evident difference in the average scores of students (mean_reading - mean_writing = 0).
H_a: There is an evident difference in the average scores of students mean_reading - mean_writing != 0.

7.2 (d)

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

The data should be independent which is known from (a). The distribution should be normal which is also observable from the boxplots and there are no outliers in the data as well.

7.2 (e)

(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?

```
n_students <- 200
avg_diff <- -.545
df_scores <- n_students - 1
sd_scores <- 8.887
se_scores <- sd_scores / sqrt(n_students)
t_value_s <- (avg_diff - 0) / se_scores
p_value_s <- pt(t_value_s, df_scores)
p_value_s
```

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

```
# With a p-value of 0.19, we cannot reject the null hypothesis. There is no significant evidence
# of difference between the average scores on two exams.
```

7.2 (f)

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

By not rejecting the null hypothesis, we are at a risk of making the type 2 error of wrongfully rejecting alternative hypothesis.

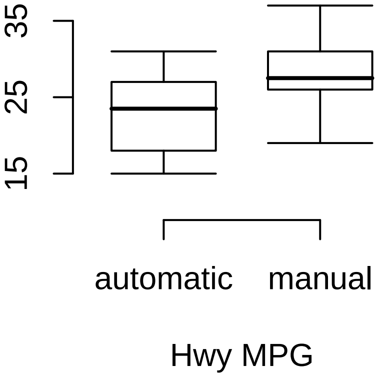
7.2 (g)

(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.

Yes, as there is a possibility of no difference between the average scores, the CI will include 0.

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



#7.28

```
#H_0: There is no difference in average mileage between automatic and manual cars.

#H_a: There is a difference in average mileage between automatic and manual cars.

n_cars <- 26

# For the automatic cars:
mean_auto <- 16.12
sd_auto <- 3.58

# For the manual cars:
mean_manual <- 19.85
sd_manual <- 4.51

avg_diff_cars <- mean_auto - mean_manual

se_cars <- ( (sd_auto ^ 2 / n_cars) + ( sd_manual ^ 2 / n_cars) ) ^ 0.5

t_value_cars <- (avg_diff_cars - 0) / se_cars
df_cars <- n_cars - 1
p_value_cars <- pt(t_value_cars, df = df_cars)
p_value_cars
```

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

```
# The p-value is lower than 0.05, so we can reject the null hypothesis.
# There is some evidence of difference between the average 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%?

7.34

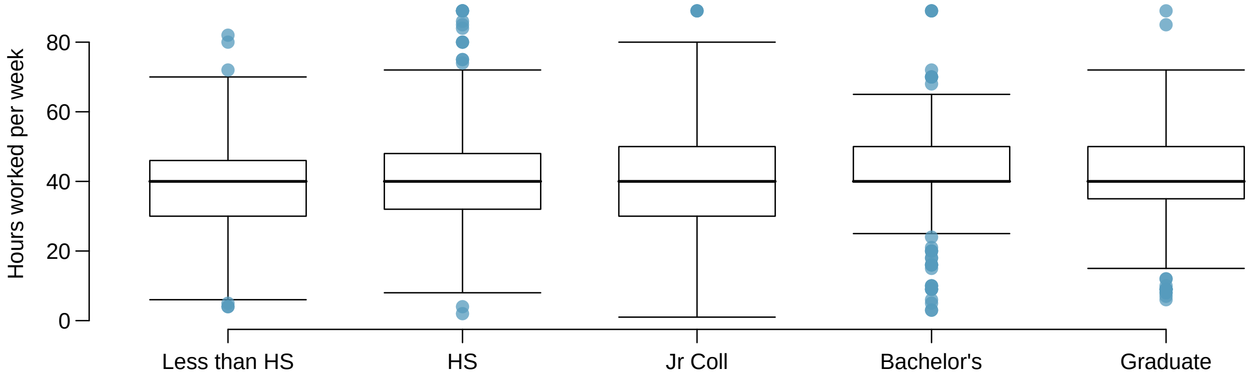
```r
# For 80% CI,

z_value_surveys <- 1.28

sd_surveys_current <- 2.2

me_surveys_desired <- 0.5

n_enrollees <- ceiling(((z_value_surveys * sd_surveys_current) / me_surveys_desired)^2)

n_enrollees
```

```
## [1] 32
```

```r
# 32 new enrollees are needed for the desired results.
```

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



Work hours and education (a)

(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

H_0: The average number of hours worked does not vary across the five groups. H_a: The average number of hours worked varies across the five groups (at least one is different from others).

Work hours and education (b)

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

The following assumptions are to be made:

1. The observations are independendent of each other.
2. The number of observations is large enough to have a normal distribution.
3. The equality of variability is established by the data provided (sd values).

Work hours and education (c)

(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	□	□			

```
# Missing df for degrees and residuals:

df_degree <- 5 - 1
df_degree
```

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

```
# DF for degrees is 4.

n_education <- 1172
mean_sq_degree <- 501.54
sum_sq_residuals <- 267382
df_residuals <- n_education - 5
df_residuals
```

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

```
df_total_a <- df_degree + df_residuals
df_total_a
```

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

```
# DF for degrees is 4.
# DF for residuals is 1167.
# DF for totals is 1171.
```

```
sum_sq_degree <- df_degree * mean_sq_degree
sum_sq_degree
```

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

```
# Sum sq of degrees is 2006.16.

sum_sq_total <- sum_sq_degree + sum_sq_residuals
sum_sq_total
```

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

```
# Sum sq total is 269388.20.
```

```
mean_sq_residuals <- sum_sq_residuals / df_residuals
mean_sq_residuals
```

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

```
# The mean sq of residuals is 229.1191.
```

```
f_value_d <- mean_sq_degree / mean_sq_residuals
f_value_d
```

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

```
# The f-value is 2.188992.
```

(d) What is the conclusion of the test?

The give p-value (0.0682) is greater than 0.05, so we reject the null hyposthesis. There is not enough evidence to prove that the average number of hours worked vary amongst the five groups.