Chapter 7 - Inference for Numerical Data

Jai Jeffryes

10/26/2019

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 is the midpoint of the confidence interval.
s_mean <- (65 + 77) / 2
s_mean
```

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

```
# Margin of error is half the confidence interval.
me <- (77 - 65) / 2
me
```

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

```
# Figure out t* with degrees of freedom = 24
# Confidence interval is 90%. That means you need a cumulative .05 in each tail.
# (Is the name for this, density function?)
t_star <- abs(qt(0.05, 24))

# margin of error = t_star * (sample standard deviation / square root of n)
# solve for s.
s <- (me * sqrt(25)) / t_star
s
```

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

(whew, did I get it?)

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?

ANSWER

```
# Z score for 90% confidence: 1.65.
# 1.65 * (250 / square root of n) must be less than 25
n <- ((1.65 * 250) / 25)^2
n
```

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

```
# You need to round up.
ceiling(n)
```

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

(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 Larger, because the Z-score multiplier for a 99% confidence interval is larger than that of a 90% confidence interval, 1.96 vs. 1.65.

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

ANSER

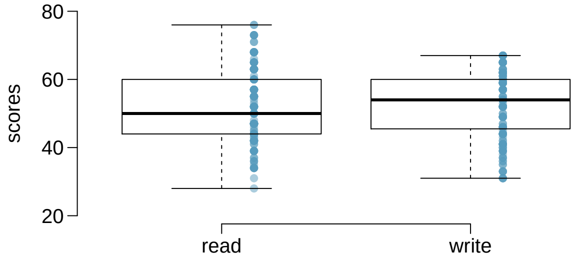
```
# Z score for 99% confidence: 1.96.
# 1.96 * (250 / square root of n) must be less than 25
n <- ((1.96 * 250) / 25)^2
n
```

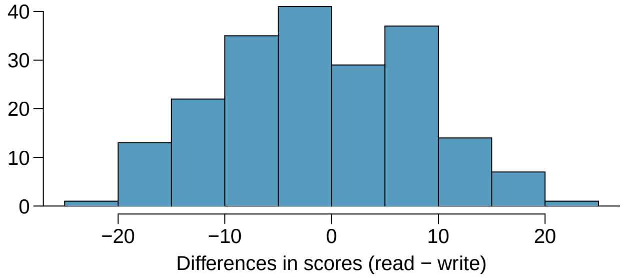
```
## [1] 384.16
```

```
# You need to round up.
ceiling(n)
```

```
## [1] 385
```

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

There is an observable difference. I don't know what "clear" means, though.

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

ANSWER

No, the scores are paired for each student.

(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: $\mu_{diff} = 0$. There is no difference in students' average reading and writing exam scores.

H_A: $\mu_{diff} \neq 0$. There is a difference.

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

ANSWER

The samples are random.

The sample size is 200. Plenty.

The histogram reveals no outliers. We're good to go.

(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?

```r
se <- 8.887 / sqrt(200)
samp_mean <- -0.545
t <- (samp_mean - 0) / se
df <- 199
p_val <- 2 * pt(t, df)
```

I come up with a p-value of 0.3868365, which is greater than 0.05. Therefore, I do not reject the null hypothesis.

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

ANSWER

It's possible I made a Type 2 error, failing to reject the null hypothesis when it was true. In that case, there could be a difference in test scores that was not simply due to chance.

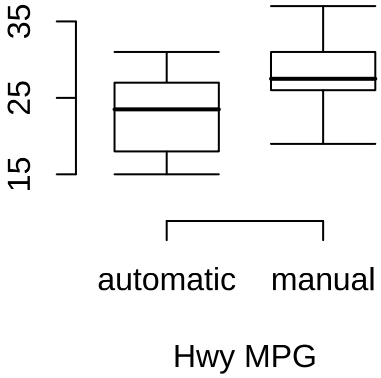
(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

ANSWER

Yes, the high p-value reflects a possibility that the population mean for a difference between test scores could be 0, which is what it means that 0 would be a member of 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



ANSWER

```
samp_diff <- 22.92 - 27.88
se <- sqrt((5.29^2 / 25) + (5.01^2 / 25))
t_star <-  abs(qt(0.025, 25))
me <- 1.95 * t_star
me
```

```
## [1] 4.0161
```

The 95% confidence interval is: -4.96 ± 2.0595386 * 1.95.

(-8.9761002, -0.9438998)

There is a 95% chance that automatic transmissions have lower mileage than manual transmission by an amount within the interval above.

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

```
z <- qnorm(.8)
# .84 * se + 1.96 * se = 2.8 * se
# .5 = 2.8 * sqrt((2.2^2 / n) + (2.2^2 / n))
n <- (2.8^2 / .5^2) * (2.2^2 + 2.2^2)
n
```

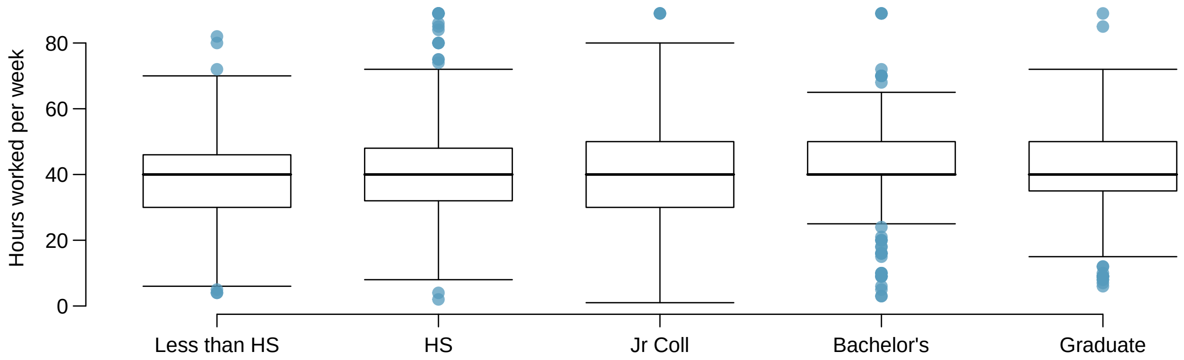
```
## [1] 303.5648
```

```
# Round up
ceiling(n)
```

```
## [1] 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.

ANSWER

H_0: The means of hours worked for all levels of education attainment are identical.

H_A: The mean hours worked for at least one pair of levels of education attainment differs.

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

ANSWER

We must assume that the subjects in each group are independent.

We must assume the distributions are normal. There is one extreme outlier for the category of Junior College, but we have over 1000 samples. We'll go ahead.

Constant variance. The standard deviation is close between the categories.

(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	4	□	□	501.54	□	0.0682
Residuals		□	267,382	□		
Total		□	□			

ANSWER

DEGREE: 4…

Sorry, Dr. Bryer, I have to throw in the towel at the end here. I almost made it! I really want to understand ANOVA because we use it a lot in my lab, but it's going to have to be another day.

(d) What is the conclusion of the test?