CUNY SPS DATA606 HW7 - Inference for Numerical Data

Chinedu Onyeka

October 9th, 2021

Problem 1:

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.

Solution 1: $n = 25; df = n - 1 = 24$

$CI = \bar{x} + t^* * SE; SE = \frac{s}{\sqrt{n}}; t^* = 1.7109$ at 90% confidence level

1a)

Sample mean: sample mean $= \frac{upperlimit + lowerlimit}{2} = \frac{77+65}{2}$
Therefore, the sample mean is 71

1b)

The margin of error: $ME = \frac{upperlimit + lowerlimit}{2} = \frac{77-65}{2}$
Therefore, the margin of error is 6

1c)

Sample standard deviation: $ME = \frac{t^* * s}{\sqrt{n}}$

$=> s = \frac{ME * \sqrt{n}}{t^*} = \frac{6 * \sqrt{25}}{1.7109} = 17.53$

Therefore, the sample standard deviation is 17.53

Problem 2:

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.

Solution 2: $sd = 250; ME \leq 25; z^* = 1.645$ for 90% confidence interval.

2a)

$ME = \frac{z^* * s}{\sqrt{n}} \leq 25$
$=> n \geq (\frac{z^* sd}{25})^2$
$=> n \geq (\frac{1.645*250}{25})^2$
$=> n \geq 270.6$
Raina would need a sample size of at least 271.

2b)

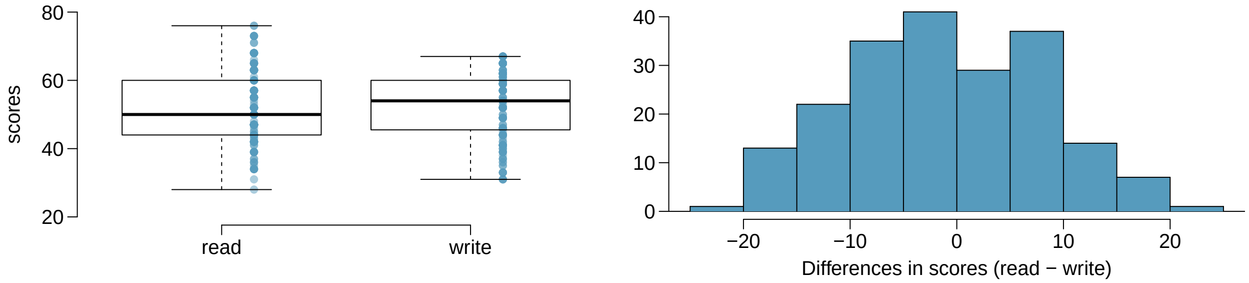
Luke will need a larger sample. n is directly proportional to the square of z^*. Increasing the confidence interval will lead to a higher value of z^* which will in turn lead to a higher value for n.

2c)

for Luke's sample size at 99% confidence level, $z^* = 2.5758$
$=> n \geq (\frac{2.5758*250}{25})^2$
$=> n \geq 663.47$
Therefore, Luke will need a sample size of at least 664.

Problem 3:

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.

Solution 3: 3a)

I do not see a clear difference between the reading and writing scores. Although the mean of the writing scores is slightly higher than the mean of the reading scores, but the spread of the writing scores is smaller than that of the reading scores. The histogram of differences in scores is nearly normal. Looking at the visualizations alone, there is no convincing evidence that a clear difference in the average reading and writing scores exists.

3b)

The sample is from a random sample. Hence, the scores of each student are independent of the scores of other students. However, reading and writing for each student would not be independent as they are paired.

3c)

Null Hypothesis, H_0: There is no difference in the reading and writing scores.$\mu_{read} - \mu_{write} = 0$
Alternative Hypothesis, H_1: There is some difference in the reading and writing scores.$\mu_{read} - \mu_{write} \neq 0$

3d)

Check conditions:
Independence: The samples come from a simple random sample. Hence, Independence is satisfied.
Normality: The sample size is large enough and the histogram shows a nearly normal distribution.
Since these conditions are satisfied, we can move forward with applying the t-distribution for this problem.

3e)

$\hat{x}_{read-write} = -0.545; sd_{diff} = 8.887; n = 200; df = n - 1 = 200 - 1$
$\mu_{read-write} = 0; SE_{diff} = \frac{sd_{diff}}{\sqrt{n}} = \frac{8.887}{\sqrt{200}} = 0.628$

Test statistic $T = \frac{\hat{x}_{diff}-\mu_{diff}}{SE_{diff}} = \frac{-0.545-0}{0.628} = -0.8678$
p value:

```r
alpha <- 0.05
p_value <- round(2*pt(-0.8678, 199), 4)
paste0("Since the p value is ", p_value, " which is greater than ", alpha,
       ", we do not reject the null hypothesis")
```

```
## [1] "Since the p value is 0.3865 which is greater than 0.05, we do not reject the null hypothesis"
```

Therefore, there is sufficient statistical evidence to support the null hypothesis that there is no difference in the reading and writing scores of the students.

3f)

There are two possible errors in hypothesis testing:
i) Type I error: Rejecting a True null hypothesis
ii) Type II error: Failing to reject a False null hypothesis
In this problem, we failed to reject the null hypothesis because the p value is greater than the level of significance. Here, we stand a chance of making a Type II error of failing to reject the null hypothesis if it is in fact False.

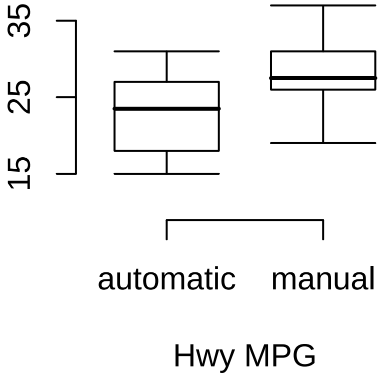
3g)

The null hypothesis is that the difference is zero which means that we would expect the confidence interval to include 0 in which case we do not reject the null hypothesis.

Problem 4:

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



Solution 4: $\bar{x}_a = 22.92; \bar{x}_m = 27.88; sd_a = 5.29; sd_m = 5.01; n_a = 26; n_m = 26$
$CI = (\bar{x}_m - \bar{x}_a) + t^* * SE$
Since $n_a = n_m$, we can use the pooled standard deviation to compute SE.
$s^2_{pooled} = \frac{s^2_a(n_a-1)+s^2_m(n_m-1)}{n_a+n_m-2} = \frac{5.29^2(26-1)+5.01^2(26-1)}{26+26-2}$
$s^2_{pooled} = 26.5421$
$SE = \sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_m}} = \sqrt{\frac{26.5421}{26} + \frac{26.5421}{26}}$
$SE = 1.428881$
$df = n_a + n_m - 2 = 26 + 26 - 2 = 50; t^* = 2.4$ for 50 df and 98% confidence level.
$CI = (27.88 - 22.92) \mp 2.4(1.428881) = (1.53, 8.39)$
Therefore, we are 98% confident that the difference between average highway mileage of manual and automatic cars is in the interval (1.53, 8.39)

Problem 5:

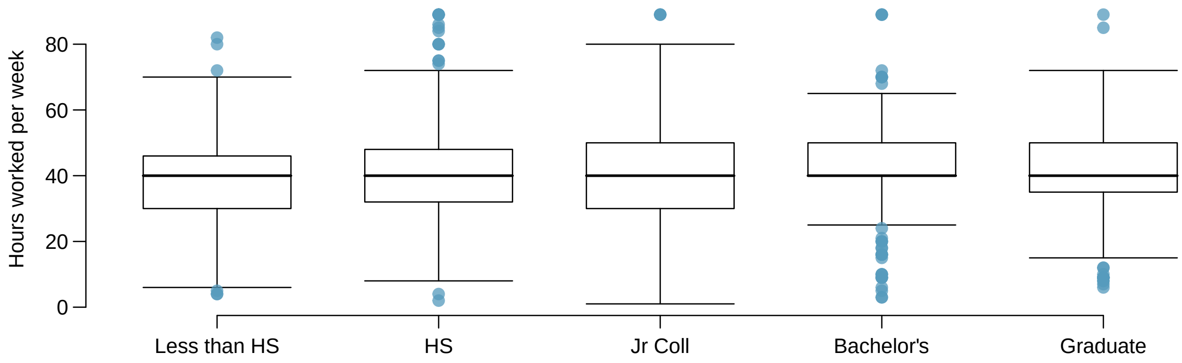
Email outreach efforts. (7.34, p. 284) A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks for information on a person's family history in regards to cancer. Another survey asks about what topics were discussed during the person's 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%?

Solution 5: s = 2.2; power = 80% = 0.8
z = 0.84 for 80%;
The target distance between the center of the null and alternative distributions in terms of the standard error is given by $0.84 * SE + 1.96 * SE = 2.8SE$. In this problem, the target distance = 0.5; Hence, $0.5 = 2.8 * \sqrt{\frac{2.2^2}{n} + \frac{2.2^2}{n}}$
$0.03188 = 2 * \frac{2.2^2}{n}; n = 303.5625$
Therefore, they would need at least 304 enrollees.

Problem 6:

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?

Solution 6: 6a)

H_0: The average number of hours worked is the same across all 5 groups
H_1: The average number of hours worked is not the same across all 5 groups

6b)
Check conditions/assumptions:
- The observations are independent within and across groups: The study has a sample size of about 1172 respondents which is sufficiently large.
- The data within each group are nearly normal: From the box plots, there are no extreme outliers and the sample sizes for each group are sufficiently large. Hence, I can assume that the data within each group are nearly normal.
- The variability across the group are equal. The variability for each group can be assumed to be about the same since they all have almost similar standard deviations.

6c)
Degree of freedom: Df;
degree = k - 1 = 5 - 1 = 4;
Residuals = n - k = 1172 - 5 = 1167;
Df Total = n = 1172
Sum of Sq
$MSG = \frac{1}{df_G} * SSG; SSG = MSG * df_G = 501.54 * 4 = 2006.16$
=> SSG = 2006.16;
$SST = SSG + SSE = 2006.16 + 267382 = 269388.16$
=> SST = 269388.16;
Mean Sq
$MSE = \frac{1}{df_E} * SSE = \frac{1}{1167} * 267382 = 229.119$
F-value
$F_{value} = \frac{MSG}{MSE} = \frac{501.54}{229.119} = 2.18899$

Create a dataframe:

```
headers <- c("Df", "Sum-Sq", "Mean-Sq", "F-value", "Pr(>F)")
row_names <- c("degree", "Residuals", "Total")
degree <- c(4, 2006.16, 501.54, 2.18899, 0.0682)
Residuals <- c(1167, 267382, 229.119, NA, NA)
Total <- c(1172, 269388.16, NA, NA, NA)
table <- data.frame(degree, Residuals, Total)
row.names(table) <- headers
table <- t.data.frame(table)
table
```

```
##             Df    Sum-Sq Mean-Sq F-value Pr(>F)
## degree       4   2006.16 501.540 2.18899 0.0682
## Residuals 1167 267382.00 229.119      NA     NA
## Total     1172 269388.16      NA      NA     NA
```

6d)
Conclusion: Since $p_{value} = 0.0682 > 0.05$, we do not reject the null hypothesis. Hence, there is no statistical evidence to support that there are differences across all groups.