Leland Randles DATA606

Homework, Chapter 5

5.6 Working backwards, Part II. 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.

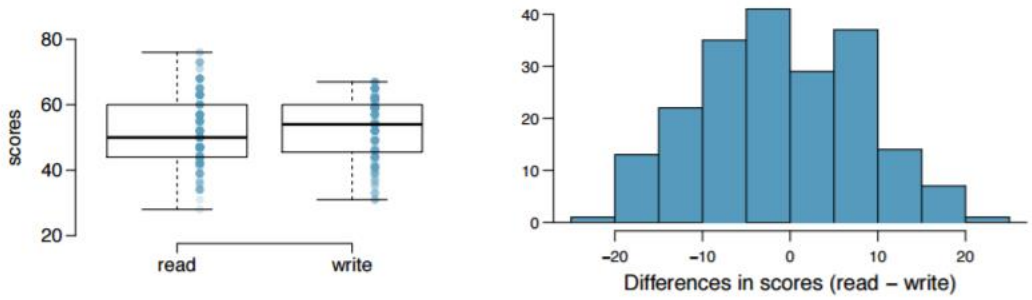
Mean = (65 + 77) / 2 = 71
Margin of Error = (77 – 65) / 2 = 6
71 – (z*1.644854) = 65 or 71 + (z*1.644854) = 77, solve for z to get standard error
SE = 3.64774
SD / sqrt(25) = 3.64774, solve for SD
SD = 18.2387

5.14 SAT scores. 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.

a) (250 / sqrt(n)) * 1.644854 = 25
   15.1989 * sqrt(n) = 250
   sqrt(n) = 16.448559
   n = 270.555 = 271
b) Luke's sample should be larger. A larger sample is needed for a more precise confidence interval because more confidence comes from being more like the population.
c) (250 / sqrt(n)) * 2.575829 = 25
   9.7056 * sqrt(n) = 250
   sqrt(n) = 25.7583
   n = 663.49 = 664

5.20 High School and Beyond, Part I. 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 $\bar{x}_{read-write} = -0.545$, 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.

a) No. There are minor differences in the boxplots, but no "clear" difference.

b) I would think that there's a relationship between a student's reading scores and writing scores, so I'd assume they are not independent.

c) H_0 : There is no difference between the averages scores of students in the reading and writing exam.
H_A : There is some difference between the averages scores of students in the reading and writing exam.

d) The data come from a simple random sample and consist of less than 10% of all cases, the observations are independent. There is no significant skew in the data.

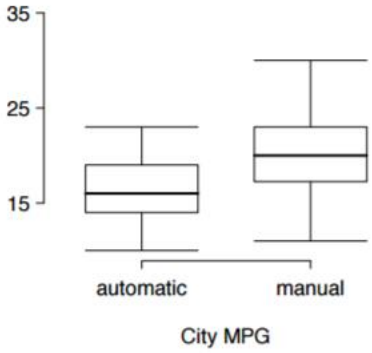
e) Since the standard deviation of the differences is 8.887, the standard error is 8.887 / sqrt(250) = 0.562. Thus, the T-score is 0.545 / 0.562 = 0.9696. p-value is 0.3324806. Therefore, we fail to reject the null hypothesis. There is not convincing evidence of a difference between the average scores on the two exams.

f) Type II. It could be that there is truly a difference between the averages for reading and writitng scores, and we are incorrectly failing to reject the null hypothesis.

g) Yes. The confidence interval should include the null hypothesis because we failed to reject the null hypothesis.

5.32 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[42]

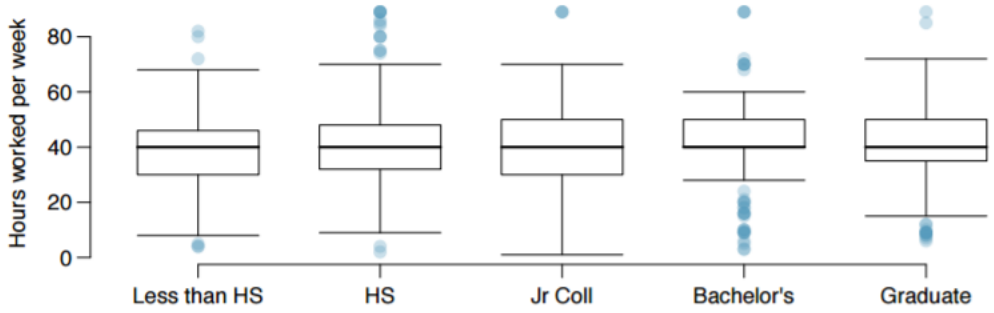
	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



The point estimate is 19.85 – 16.12 = 3.73
The standard error of the difference is the square root of ((4.51^2) / 26) + ((3.58^2) / 26), which is 1.129.
The T-score is 3.73 / 1.129 = 3.304, so the p-value is 0.000987. This p-value provides strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage.

5.48 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?

a) H_0 : There is no difference between average hours work for samples taken from the five educational attainment bins.
H_A : The average hours worked varies across some or all of the five educational attainment samples.

b) Are the observations independent within and across groups? Hard to access, but we'll assume so.
Are the data within each group nearly normal? Based on the boxplots, yes. There is some skew in a couple of groups, but the size of those groups negates concern.
Is the variability across the groups about equal? Yes.

c) Degrees of freedom for degree = 4. Df for Residuals is 1,167. Total Df is 1,171.
Sum Sq for degree is ?. Not sure how we can complete the rest without the data set.

d) The p-value is 0.0682, which is greater than our significance level of 0.05, so we fail to reject the null hypothesis.