Homework Assigment 5

CUNY MSDA DATA 606

Duubar Villalobos Jimenez mydvtech@gmail.com

April 2, 2017

Chapter 5 - Inference for Numerical Data

```
Practice: 5.5, 5.13, 5.19, 5.31, 5.45
Graded: 5.6, 5.14, 5.20, 5.32, 5.48
```

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.

Answer

Sample Mean

Since we know that the sample mean is $\frac{(x_2+x_1)}{2}$ where the confidence interval is (x_1, x_2)

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

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

The sample mean is 71.

Marging of Error

Since we know that the margin of error is $\frac{(x_2-x_1)}{2}$ where the confidence interval is (x_1, x_2)

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

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

The margin of error is ME = 6.

Sample standard deviation

To calculate the sample standard deviation we use $ME = t^* \cdot SE$ by using the **qt()** function and **df = 25 - 1**.

```
df <- 25 - 1
p <- 0.9
p_2tails <- p + (1 - p)/2

t_val <- qt(p_2tails, df)
```

```r
# Since ME = t * SE
SE <- ME / t_val

# Since SE = sd/sqrt(n)
sd <- SE * sqrt(n)
```

The standard deviation is sd = 17.5348146.

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.

Answer

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

For this, I will use as follows: $ME = z \cdot SE$ and since $SE = \frac{sd}{\sqrt{n}}$

we have as follows: $ME = z \cdot \frac{sd}{\sqrt{n}}$ at the end we obtain: $\frac{ME}{z} = \frac{sd}{\sqrt{n}}$

$n = \left(\frac{z \cdot sd}{ME}\right)^2$

```r
z <- 1.65 # due to 90% Confidence interval
ME <- 25
sd <- 250

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

The sample size should be 273 students.

(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.

Luke's sample should be larger since it will require a higher **z** number multiplied by the standard deviation and then squared.

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

```r
z <- 2.575 # due to 99% Confidence interval
ME <- 25
sd <- 250

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

The sample size should be 664 students.

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.

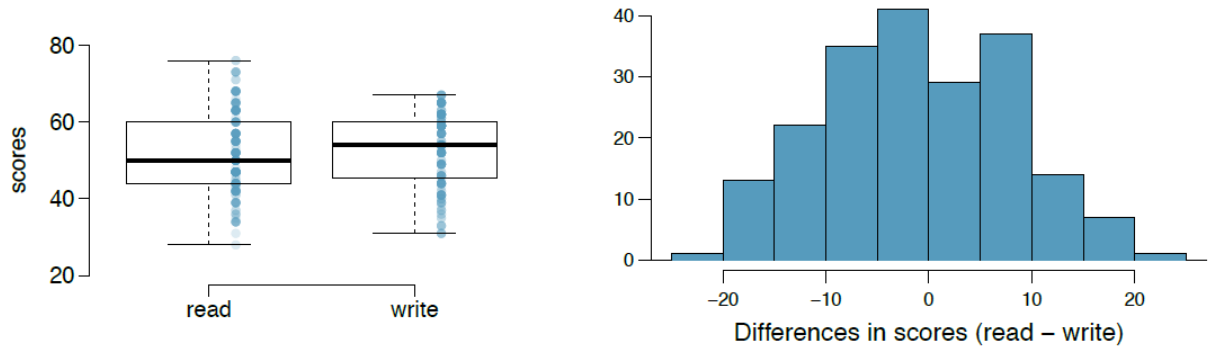

Figure 1:

Answer

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

I do not see a clear difference in the average of the reading and writing scores. The difference distribution is fairly normal around the zero difference, though it seems to be a slight skew to the right.

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

I would say that the scores are independent of each student but not of each score, that is reading and writing scores are not independent of each other 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?

Since the question is referring for the difference in the average score of students, and not referring to the average difference in scores. The hypotheses could be as follows:

H_0: The difference of average in between reading and writing equal zero. That is: $\mu_r - \mu_w = 0$

H_A: The difference of average in between reading and writing does NOT equal zero. That is: $\mu_r - \mu_w \neq 0$

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

1. *Independence of observations*: The difference histogram suggested the data are paired. If paired, then they wouldn't be independent.
2. *Observations come from nearly normal distribution*: The box plot provided in the text suggests the data are reasonably normally distributed and no outliers exist.

(e) The average observed difference in scores is $\bar{x}_{read} - \bar{x}_{write} = -0.545$, 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?

The hypotheses for the average difference test are:

H_0: The difference of average scores is equal to zero. That is: $\mu_{diff} = 0$

H_A: The difference of average scores is NOT equal to zero. That is: $\mu_{diff} \neq 0$

The paired data is presumably from less than 10% of the population of senior high schoolers, and from a simple random sample. We noted that the differences are nearly normally distributed, so the conditions are met in order to apply the t-distribution.

```r
sd_Diff <- 8.887
mu_Dif <- -0.545
n <- 200

SE_Diff <- sd_Diff / sqrt(n)

# Compute T statistic
t_value <- (mu_Dif - 0) / SE_Diff

df <- n - 1

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

Since the p-value is not less that 0.05, this implies that there is not convincing evidence of a difference in student's reading and writing exam scores maintaining our NULL hypothesis.

The above conclusion need to be analyzed with further detail since the data need to be independent and currently is not.

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

Type I error: Incorrectly reject the null hypothesis.

Type II error: Incorrectly reject the alternative hypothesis.

In the case, we may have made a type II error by rejecting the alternative hypothesis HA. that is, we might have wrongly concluded that there is not a difference in the average student reading and writing exam scores.

(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, I would expect a confidence interval for the average difference between reading and writing scores to include 0.

When the confidence interval include 0 for this kind of hypothesis test, it indicates that the difference is not in one side or another.

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

Answer

The hypotheses for this test are as follows:

H_0: The difference of average miles is equal to zero. That is: $\mu_{diff} = 0$

H_A: The difference of average miles is NOT equal to zero. That is: $\mu_{diff} \neq 0$

From the text we have as follows:

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

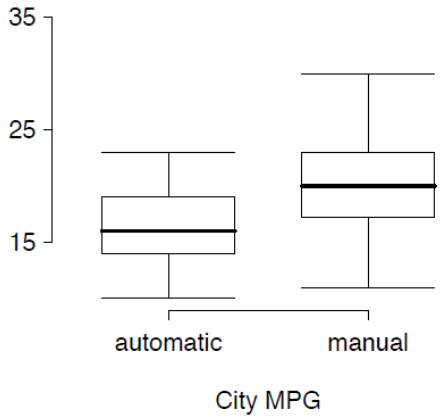

Figure 2:

```
n <- 26
# Automatic
mu_a <- 16.12
sd_a <- 3.58
# Manual
mu_m <- 19.85
sd_m <- 4.51

# difference in sample means
mu_Diff <- mu_a - mu_m

# standard error of this point estimate
SE_Diff <- ( (sd_a ^ 2 / n) + ( sd_m ^ 2 / n) ) ^ 0.5

t_val <- (mu_Diff - 0) / SE_Diff
df <- n - 1
p <- pt(t_val, df = df)
p
```

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

Since the **p-value** is less than 0.05, we reject the null hypothesis H0 and conclude that there is strong evidence of a difference in fuel efficiency between manual and automatic transmissions.

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.

Answer

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

	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

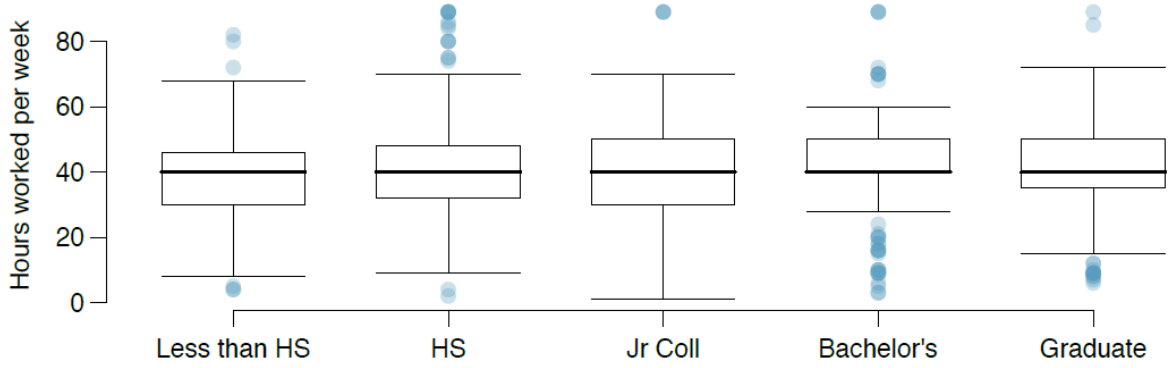

Figure 3:

The hypotheses for this ANOVA test follow:

H_0: The difference of ALL averages is equal. That is: $\mu_l = \mu_h = \mu_j = \mu_b = \mu_g$

H_A: There is one average that is NOT equal to the other ones.

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

- *The observations are independent within and across groups*: I will assume independence within and across the groups based on the nature of the provided data.
- *The data within each group are nearly normal*: The box plots do not support nearly normal data within each group. Each group has outliers some groups seem to follow a normal distribution.
- *The variability across the groups is about equal*: There seems to be a similarity of variability in between some of the groups just by observing the standard deviations.

(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					

Figure 4:

```
mu <- c(38.67, 39.6, 41.39, 42.55, 40.85)
sd <- c(15.81, 14.97, 18.1, 13.62, 15.51)
n <- c(121, 546, 97, 253, 155)
data_table <- data.frame (mu, sd, n)
```

```r
n <- sum(data_table$n)
k <- length(data_table$mu)

# Finding degrees of freedom
df <- k - 1
dfResidual <- n - k

# Using the qf function on the Pr(>F) to get the F-statistic:

Prf <- 0.0682
F_statistic <- qf( 1 - Prf, df , dfResidual)

# F-statistic = MSG/MSE

MSG <- 501.54
MSE <- MSG / F_statistic

# MSG = 1 / df * SSG

SSG <- df * MSG
SSE <- 267382

# SST = SSG + SSE, and df_Total = df + dfResidual

SST <- SSG + SSE
dft <- df + dfResidual
```

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	**4**	**2006.16**	501.54	**2.188984**	0.0682
Residuals	**1167**	267,382	**229.12**		
Total	**1171**	**269388.16**			

(d) The independence assumption can be relaxed when the total sample size is large.

Since the p-value = 0.0682 is greater than 0.05, We conclude that there is not a significant difference between the groups and the null hypothesis does not get rejected.