Illapani_Homework_HW5

Srini Illapani

October 30, 2015

Inference for numerical data

5.6 Working Backwards, Part II

A 90% confidence interval for a population mean is (65, 77). The population distribution is approx. 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.

```
n <- 25
diff <- 12

# margin of error (moe)
moe <- diff / 2
mean <- 65 + moe
mean
```

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

```
# calculate t
df <- n - 1
t <- qt(.95, df)
t
```

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

```
# calculate Standard Error SE
SE <- moe / t
SE
```

```
## [1] 3.506963
```

```
# calculate sample Standard Deviation SD

SD <- SE * sqrt(n)
SD
```

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

5.14 SAT Scores

SAT Scores of students at an Ivy League college are distributed with a standard deviation of 250 points. Two statistitcs 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 of a sample should she collect?

```
SD <- 250
moe <- 25

# calculate z
z <- qnorm(0.95)
z

## [1] 1.644854

# calculate the sample size
n <- ((SD * z) / moe)^2
n

## [1] 270.5543
```

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.

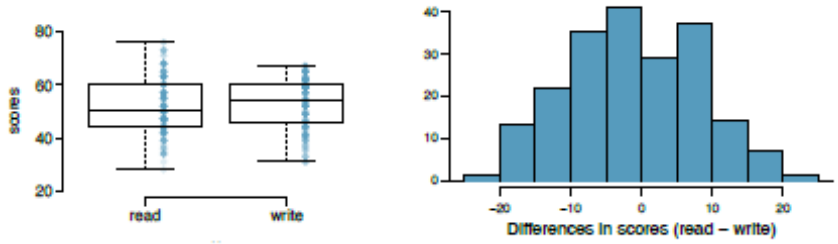
- Luke should have a larger sample size to compensate for the increased Z score for a 99% CI.

c) Calculate the minimum required sample size for Luke.

```
# calculate z for a 99% CI
z <- qnorm(0.995)
z

## [1] 2.575829

# calculate the sample size
#
n <- ((SD * z) / moe)^2
n

## [1] 663.4897
```

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.



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

- No. Based on the above graphic, there is no clear difference in the average reading and writing scores.

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

- From the sample of 200 students, Yes, each student's readinga and writing scores are independent of other student's scores.

c) Create hypotheses appropriate for the following research question: is there an evident difference in the average score of students in the reading and writing exam?

Null Hypothesis H0: u_r - u_w = 0 Alternate Hypothesis Ha: u_r - u_w != 0

d) Check the conditions required to complete this test.

```
- Independence of observations
- Observations show normal distribution
```

e) The average observed difference in scores iss 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?

The paired data is presumably from less than 10% of the population of high schoolers, and from a simple random sample. We've already see the differences are nearly normally distributed, so the conditions are met to apply the t-distribution.

```
sD <- 8.887
x <- -0.545
n <- 200

# Calculate the standard error SE
SE <- sD / sqrt(n)

# Calculate t
t <- (x - 0) / SE

# Calculate the p-value
df <- n - 1
p <- pt(t, df=df)
p
```

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

The p-value of 0.19 is greater than 0.05, hence one can conclude that there is no convicing evidence of a difference in student's reading and writing scores.

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

- Type II error: It is made when we incorrectly reject the alternative hypothesis. In the above question, we may have made a type II error.

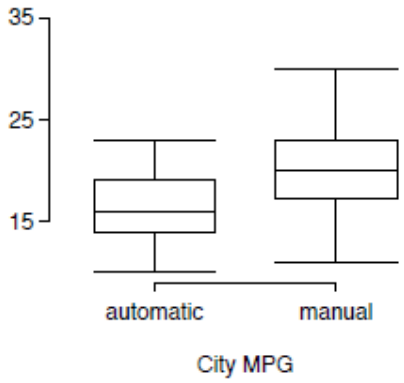
g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

```
- Yes, one can expect a confidence interval for the average difference
between reading and writing scores to include 0. It         indicates that
the difference is not clearly on either side of zero and therefore results in
a failure to reject the null        hypothesis.
```

5.32 Fuel efficiency of manual and automatic cares, Part I

Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cares manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cares 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.

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



The hypotheses for this test are:

Null Hypothesis H0: u_a - u_m = 0
Alternate Hypothesis Ha: u_a - u_m != 0

```
n <- 26

m_a <- 16.12
```

```r
SD_a <- 3.58
m_m <- 19.85
SD_m <- 4.51

# Calculate the difference in sample means
x <- m_a - m_m
x
```

```
## [1] -3.73
```

```r
# Calculate the Standard Error SE
SE <- sqrt( (SD_a^2 / n) + (SD_m^2 / n) )
SE
```

```
## [1] 1.12927
```

```r
# Calculate the t-statistic and the p-value
t <- (x - 0) / SE
t
```

```
## [1] -3.30302
```

```r
p <- pt(t, df=n-1)
p
```

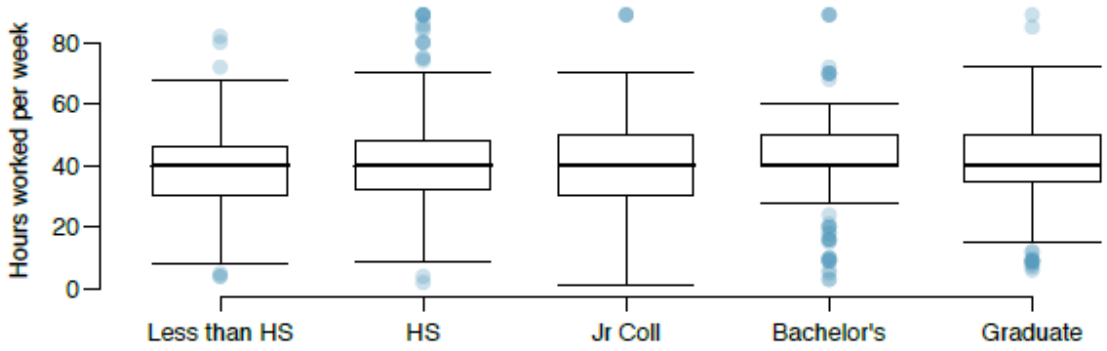
```
## [1] 0.001441807
```

The p-value is less than 0.05, so we can reject the null hypothesis and say that there is strong evidence of a difference in fuel efficiency between auto and manul transmission cars.

5.48 Work hours and education

The General Social Survey collects data on demographics, education, and work amoung many other characteristics of US residents. 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 ill 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 the hypotheses for evaluating whether the average number of hours worked varies across the five groups.

The hypotheses for this ANOVA test are:

Null Hypothesis H0: u_lhs = u_hs = u_jc = u_b = U_g = 0
Alternate Hypothesis Ha: atleast one of the means is different

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

The observations are independent within and across groups
The data within each group are nearly normal
The variability across the groups is about equal

c) Below is part of the output associated with this test. Fill in the empty cells.

The values have been filled in bold italics in the table below.
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) What is the conclusion of the test?

Since the p-value of 0.0682 is slightly greater than 0.05, one can conclude that there is not a significant difference between the groups and hence do not reject the null hypothesis.