DATA 606 - Homework 5

Joshua Sturm

October 27, 2017

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.

Solution

Since the population distribution is approximately normal, the sample means will be nearly normal even if $n < 30$.
Formula for the sample mean: $\frac{x_1+x_2}{2}$.

```
n <- 25
x1 <- 65
x2 <- 77
samp_mean <- (x1+x2)/2
```

Formula for the margin of error: $\frac{x_2-x_1}{2}$.

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

Formula for the standard deviation: $SE = \frac{s}{\sqrt{n}}$.

```
df <- n-1
c <- 0.9
c2 <- c + (1-c)/2
t24 <- qt(c2, df)
```

$ME = t^*_{24}SE \rightarrow SE = \frac{ME}{t^*_{24}} \rightarrow s = \frac{ME\sqrt{n}}{t^*_{24}} \rightarrow \frac{6\sqrt{25}}{t24}$

```
s <- (moe*sqrt(25))/(t24)
```

The standard deviation for the sample is 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.
(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)

$ME = z * \frac{\sigma}{\sqrt{n}}, \quad ME = 25, \sigma = 250.$

```
# for 90% confidence interval, Z* = 1.645
z <- 1.645
me <- 25
sigma <- 250
# 25 = 1.645*(250/sqrt(n))
n <- (z*sigma/me)^2
```

Raina would need $270.6025 = 271$ students.

(b)

Since Luke wants a higher certainty, he'll need a higher z-value, which will result in a larger sample size.

(c)

```
# for 99% confidence interval, Z* = 2.58
z <- 2.58
n <- (z*sigma/me)^2
```

Luke would need a minimum of $665.64 = 666$ students.

5.20 High School and Beyond, Part I.

(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 $\bar{x}_{\text{read}-\text{write}} = -0.545$, 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)

There doesn't appear to be a clear difference. The means are slightly different in the box plot, but the histogram is nearly normal, with a center close to 0.

(b)

Because a student's reading abilities are related to their writing skills, I'd say that the two are not independent, but **paired**. If the question is asking between students, then yes, I'd say each student's abilities are independent.

(c)

$H_0 : \mu_{\text{diff}} = 0$. There is no difference in the average scores of students in the reading and writing exam.
$H_A : \mu_{\text{diff}} \neq 0$. There *is* a difference in average scores.

(d)

Independence: Since the sample size $n = 200 > 30$, and comprises less than 10% of high school students nationwide, we can assume that each student is independent. Skew: There is no strong skew evident in the histogram. Because both conditions are satisfied, we can apply the t-distribution.

(e)

```
n <- 200
df <- n-1
sd <- 8.886
avg_diff <- -0.545
se <- sd / sqrt(n)
tdf <- (avg_diff - 0)/(se)
p <- 2 * pt(tdf, df)
```

Since $p =0.3867831 > p = 0.05$, we can't reject the null hypothesis, and conclude that there is no difference in the average scores.

(f)

(From footnote 16 on page 235): It's possible we didn't detect a difference, and made a Type 2 Error. If we did make an error, we may have falsely accepted the null hypothesis.

(g)

Yes. Since we accepted the null hypothesis, which said that the difference is 0, then I'd expect 0 to be in a confidence interval.

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.

Solution

$\bar{x}_{\text{Automatic}} = 16.12, \quad s_{\text{Automatic}} = 3.58.$

$\bar{x}_{\text{Manual}} = 19.85, \quad s_{\text{Manual}} = 4.51.$
$n = 26.$
$df = n - 1 = 25.$

$H_0 : \mu_A = 0$. There is no difference in mpg between automatic and manual cars.
$H_A : \mu_M \neq 0$. There *is* a difference in mpg.

$\bar{x}_{\text{diff}} = \bar{x}_A - \bar{x}_M = 16.12 - 19.85 = -3.73$.

$SE = \sqrt{\frac{s_A^2}{n} + \frac{s_M^2}{n}} = \sqrt{\frac{(3.58)^2}{26} + \frac{(4.51)^2}{26}} \approx 1.12927$.

```
n <- 26
df <- n-1
meanA <- 16.12
sdA <- 3.58
meanM <- 19.85
sdM <- 4.51
xdiff <- meanA - meanM
se <- sqrt(((sdA^2)/n)+((sdM)^2)/n)
tdf <- (xdiff - 0)/(se)
p <- 2 * pt(tdf, df)
```

Since \$p =0.0028836 < p = 0.05\$, we reject the null hypothesis, and conclude that there *is* a difference in mpg between automatic and manual cars.

5.48 Work hours and education.

The General Social Survey collects data on demographics, education, and work, among 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 will be helpful in carrying out this analysis. (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. (d) What is the conclusion of the test?

(a)

$H_0 : \mu_1 = \mu_2 = ... = \mu_n$. The mean number of hours worked is the same across all groups.
$H_A : \mu_1 \neq \mu_2 \neq ... \neq \mu_n$. The mean number of hours worked is not the same.

(b)

Independence: $n = 1,172 > 30$, and comprises $< 10\%$ of the population, so we can assume they're independent. There doesn't appear to be any strong skew, so we can assume the data is nearly normal. Mean and standard deviation are similar for the most part, so we can assume variability across all groups is equal.

(c)

```
# All formulas taken from page 250 in the textbook

k <- 5   # Categories
df <- k - 1   # degrees of freedom
n <- c(121, 546, 97, 253, 155)              # array of totals for each category
tot_n <- sum(n)

df_E <- tot_n - k
```

```r
tot_df <- df + (tot_n - k)

# MSG = SSG / df_G

MSG <- 501.54 # (given)
# SSG = MSG * df_G
SSG <- df * MSG
# SSE = SST - SSG
SSE <- 267382
SST <- SSE + SSG

# MSE = SSE / df_E
MSE <- SSE / df_E

# F = MSG / MSE
F_stat <- MSG / MSE

Pr <- 0.0682

names <- c("df", "Sum Sq", "Mean Sq", "F Value", "Pr(>F")
row_names <-

col_df <- c(df, df_E, tot_df)
col_sq <- c(SSG, SSE, SST)
col_msq <- c(MSG, MSE, NA)
col_F <- c(F_stat, NA, NA)
col_Pr <- c(Pr, NA, NA)
dataf <- data.frame(col_df, col_sq, col_msq, col_F, col_Pr)
names(dataf) <- c("df", "Sum Sq", "Mean Sq", "F Value", "Pr(>F)")
rownames(dataf) <- c("degree", "Residuals", "Total")
dataf
```

```
##             df    Sum Sq  Mean Sq  F Value Pr(>F)
## degree       4   2006.16 501.5400 2.188992 0.0682
## Residuals 1167 267382.00 229.1191       NA     NA
## Total     1171 269388.16       NA       NA     NA
```

(d)

Since $P = 0.0682 > p = 0.05$, we can't reject the null hypothesis, and conclude that there's no difference in number of hours worked across demographics.