IS606 Homework 5

Daniel Dittenhafer

October 22, 2015

5.6 Working Backwards, Part II (p257)

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.

First we can work out the sample mean, by finding the mid point of the confidence interval range. We also get the T * SE segment of the confidence interval equation out of these computations.

```
n <- 25
diff <- 77 - 65
marginOfError <- diff / 2

mean56 <- 65 + marginOfError
mean56
```

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

Next we need to resolve the t value for a 90% confidence interval of n=25, df=24. We use the `qt` function to look up the t value in the t distribution for the $1 - \alpha/2 = 0.95$ where $\alpha = 1 - CI = 0.10$:

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

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

With the t value, we can separate the standard error:

```
SE <- marginOfError / t
SE
```

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

With the standard error and the number of cases we can back into the sample standard deviation:

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

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

5.14 SAT Scores (p259)

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? Using the normal distribution Z score, we can compute the sample size needed as follows:

```
mE <- 25
S <- 250
z <- qnorm(0.95)
z
```

```
## [1] 1.644854
```

```
n <- ((S * z) / mE)^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. In order to achieve a 99% confidence interval while maintaining the margin of error of 25 points, a larger sample size will be needed sufficient to offset the increased Z score associated with the 99% CI.

c) Calculate the minimum required sample size for Luke. First we lookup the Z score associated with a 99% confidence interval. Then we can calcuate the sample size needed to acheive Luke's goal.

```
z <- qnorm(0.995)
z
```

```
## [1] 2.575829
```

```
n <- ((S * z) / mE)^2
n
```

```
## [1] 663.4897
```

5.20 High School and Beyond, Part I (p261)

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? I don't see a clear difference in the average reading and writing scores. The difference distribution is fairly normal around the zero difference, though there is a slight skew to the right.

b) Are the reading and writing scores of each student independent of each other? In the sample of 200 students from the survey, I would conclude that each student's scores are independent of other student's scores as a result of the simple random sampling technique. With that said, the reading and writing scores might be paired for each student and would not be independent of each other for a given student.

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? The question is asking for the difference in the average score of students, as opposed to the average difference in scores. As such, the hypotheses would be as follows:

$H_0 : \mu_r - \mu_w = 0$

$H_a : \mu_r - \mu_w \neq 0$

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 $\bar{x}_{read-write} = -0.545$, 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:

$H_0 : \mu_{diff} = 0$

$H_a : \mu_{diff} \neq 0$

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.

```
sDiff <- 8.887
xBar <- -0.545
nDiff <- 200
# Compute the standard error
seDiff <- sDiff / sqrt(nDiff)
# Compute T statistic
T <- (xBar - 0) / seDiff
df <- nDiff - 1
pVal <- pt(T, df=df)
pVal
```

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

The p-value is not less that 0.05, therefore I conclude that there is not convicing evidence of a difference in student's reading and writing exam scores.

On the otherhand, the question of a difference between average exam scores does not seem to be addressed by the data provided. I would need $\bar{x}_{read} - \bar{x}_{write}$ and the corresponding standard deviation to proceed with that test, and the data would have to be not paired.

f) What type of error might we have made? Explain what the error means in the context of the application. A Type I error is when we incorrectly reject the null hypotheis, while Type II is when we incorrectly reject the alternative hypothesis. In the case above, we may have made a type II error with incorrectly rejecting the alternative hypothesis. In other words, we might have wrongly concluded that there is not a difference in 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 spans 0 for this kind of hypothesis test, it indicates that the difference is not clearly on one side or the other of zero and therefore results is a failure to reject the null hypothesis of no difference.

5.32 Fuel efficiency of manual and automatic cares, Part I (p266)

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.

Statistic	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26

The hypotheses for this test are as follows:

$H_0 : \mu_a - \mu_m = 0$

$H_a : \mu_a - \mu_m \neq 0$

What is the point estimate of the population difference?

```
n <- 26
m_a <- 16.12
sd_a <- 3.58

m_m <- 19.85
sd_m <- 4.51

# Compute the difference in sample means
xDiff <- m_a - m_m
xDiff
```

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

What is the standard error of this point estimate?

```
seDiff <- sqrt( (sd_a^2 / n) + (sd_m^2 / n) )
seDiff
```

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

The t-statistic and p-value associated with the difference?

```
T <- (xDiff - 0) / seDiff
T
```

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

```
pVal <- pt(T, df=n-1)
pVal
```

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

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

5.48 Work hours and education (p272)

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.

Statistic	< 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	1172

a) Write the hypotheses for evaluating whether the average number of hours worked varies across the five groups. The hypotheses for this ANOVA test follow:

$H_0 : \mu_l = \mu_h = \mu_j = \mu_b = \mu_g$

$H_a :$ At least one mean is different

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

1. *The observations are independent within and across groups*: Given the nature of the survey, we will assume independence within and across the groups.
2. *The data within each group are nearly normal*: The box plots do not support nearly normal data within each group. Particularly, the Bachelor's distribution is skewed and has significant outliers. Each group has outliers though other groups appear closer to normally distributed.

3. *The variability across the groups is about equal*: There is general similarity of variabily, though the Bachelor's again stands out as deviating.

c) Below is part of the output associated with this test. Fill in the empty cells. The values have been included in bold 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**			

The following `R` code was used to compute the values:

```
n <- 1172
k <- 5

dfG <- k - 1
dfG
```

```
## [1] 4
```

```
dfE <- n-k
dfE
```

```
## [1] 1167
```

```
meanTotal <- 40.45
dfData <- data.frame(n=c(121, 546,97,253,155),
                     sd=c(15.81,14.97,18.1,13.62,15.51),
                     mean=c(38.67,39.6,41.39,42.55,40.85))
# Compute the SSG
SSG <- sum( dfData$n * (dfData$mean - meanTotal)^2 )
# Compute the MSG
MSG <- (1 / dfG) * SSG
# Compute the F statistic
F <- 501.54 / 229.12
F
```

```
## [1] 2.188984
```

d) What is the conclusion of the test? Given the p-value = 0.0682 is greater than 0.05, I conclude that there is not a significant difference between the groups and therefore I don't reject the null hypothesis.