CUNY MSDA 606 Spring 2016 Ch. 5 Graded Problems

James Topor

Chapter 5, Problems 5.6, 5.14, 5.20, 5.32, 5.48

_____

5.6
90% conf interval = (65, 77); n = 25

Calculate the sample mean, the margin of error, and the sample standard deviation.

```
uMean <- 65 + ((77 - 65)/2)
uMean
```

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

```
# find the margin of error
Me <- (77 - 65)/2
Me
```

```
## [1] 6
```

```
# find the t score using .95 (for 90% conf) and n-1 degrees of freedom
tScore <- qt(.95, 24)
tScore
```

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

```
# Then solve margin of error equation for stdDev:
# Me = tScore * s / sqrt(n) => (Me * sqrt(n)) / tScore = s

stdDev <- (Me * sqrt(25)) / tScore
stdDev
```

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

So we have mu = 71, ME = 6, and sd = 17.53

_____

5.14
sd = 250; ME <= 25

a) *How large of a sample for a 90% confidence interval?*

25 >= critical value * (250/sqrt(n)) =>
25 >= 1.64 * (250/sqrt(n)) =>

sqrt(n) >= 1.64 * 250 / 25 =>
sqrt(n) >= 1.64 * 10 =>
sqrt(n) >= 16.4
n >= 16.4^2

```
16.4^2
```

```
## [1] 268.96
```

So the sample size must be at least n = 269.

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 that required for the 90% interval, and explain your reasoning.*

The minimum sample size for a 99% confidence interval with a margin of error of no more than 25 will have to be larger than that required for the 90% interval derived above since the critical value required for a 99% confidence interval (2.58) is larger than that required for a 90% confidence interval (1.64). Plugging that larger critical value into the equations described above will obviously result in a larger value for the minimum sample size.

c) *The minimum sample size required for a 99% confidence interval is:*

25 >= critical value * (250/sqrt(n)) =>
25 >= 2.58 * (250/sqrt(n)) =>
sqrt(n) >= 2.58 * 250 / 25 =>
sqrt(n) >= 2.58 * 10 =>
sqrt(n) >= 25.8
n >= 25.8^2

```
25.8^2
```

```
## [1] 665.64
```

So we'd need a sample size of at least n = 666 for a 99% confidence interval with a margin or error of no more than 25.

_____

5.20
n = 200;

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

No there is not a clear difference in the average reading and writing scores. Both distributions have similar IQR's and appear to be nearly normal in shape.

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

No they are not independent. Since observations of reading and writing scores for a single student are dependent (i.e., the observations apply to the same individual student), we have paired data.

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?*

H0: mu(reading) - mu(writing) = 0: There is no difference between the average reading and writing score.

HA: mu(reading) - mu(writing) != 0: There is a difference between the average reading and writing score.

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

1. Samples must be randomly selected: TRUE
2. Sample size must be less than 10% of the population: TRUE (assuming that there were at least 2000 students included in the original survey)
3. The sample size must be > 30: TRUE
4. The distribution of original population cannot be heavily skewed. UNKNOWN - we aren't given any information on the distribution of the original population so we'll have to assume it is not strongly skewed

e) *The average observed difference in scores is -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?*

N = 200, mu = -0.545, sd = 8.887

The test statistic is:

zScore = (-0.545 - 0) / se

se = 8.887 / sqrt(200)

```
se <- 8.887 / sqrt(200)
se
```

```
## [1] 0.6284058
```

```
zScore <- (-0.545 - 0) / se
zScore
```

```
## [1] -0.867274
```

```
# now calculate the p-value
2 * pt(abs(zScore), df = 199, lower.tail = FALSE)
```

```
## [1] 0.3868365
```

So the p-value = .3868 for a 2-tailed test assuming a 95% confidence interval. As such, we fail to reject H0. In other words, there does not appear to be sufficient reason to state that the data provide evidence of a real difference between the average reading and writing scores.

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

If there was an error it would be a Type 2 error, i.e., we failed to reject the null hypothesis when in fact there is evidence that we should reject it. In other words, if a Type 2 error had occurred we should have rejected H0 since the data had provided evidence of a difference between the average reading and writing 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. Since we failed to reject the null hypothesis (muDiff = 0) we would expect the confidence interval to contain 0. We can verify this by constructing the confidence interval based on the data we were provided:

We can compute a 95% confidence interval: -0.545 +- tScore * (8.887/sqrt(200))

```
tScore <- qt(.975, 199)
```

```
se <- (8.887/sqrt(200))
se
```

```
## [1] 0.6284058
```

```
c((-0.545 - tScore * se), (-0.545 + tScore * se))
```

```
## [1] -1.7841889  0.6941889
```

The 95% confidence interval is (-1.78, 0.69), so it does contain 0.

_____

5.32

H0: muDiff = 0 There is no difference in the average fuel efficiencies of cars with manual and automatic transmissions

HA: muDiff != 0 There is a difference between the average fuel efficiencies of cars with manual and automatic transmissions

Automatic: mu = 16.12; sd = 3.58; n = 26 Manual: mu = 19.85; sd = 4.51; n = 26

First, calculate T

```
t <- (16.12 - 19.85 - 0) / sqrt( (3.58^2/26) + (4.51^2/26) )
t
```

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

df = Degrees of Freedom = min(26-1, 26-1) = 25

Now calculate the p value:

```
2 * pt(abs(t), df = 25, lower.tail = FALSE)
```

```
## [1] 0.002883615
```

A two-sided test at a yields a p-value of .002, so since .002 is less than a .05 signficance value for a 95% confidence interval, we should reject the null hypothesis and conclude that there is sufficient evidence to support the assertion that a difference exists between the average fuel efficiencies of cars with manual and automatic transmissions.

_____

5.48

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

H0: mu1 = mu2 = mu3 = mu4 = mu5 The average number of hours worked is the same across all five groups.

HA: At least one mean is different from the others

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

1. Observations must be independent within groups (random sample from less than 10% of population): We have no information on this so we have to assume that's the case if we are to proceed.

2. Observations must be independent between groups: Again, we have no information on this so we have to assume that is the case if we are to proceed.

3. Normality of the distributions:

We have zero-bounded data and boxplots that show what appear to be reasonably normal distributions, with the IQR's appearing fairly balanced around their corresponding medians. In any event, the sizes of the samples appear large enough to allows us to proceed assuming approximate normality.

4. Variability across the groups should be about equal: The standard deviations appear to be relatively similar, though not equal.

c) Fill in the blanks:

Group Degrees of Freedom = **4**

Total Degrees of Freedom = N - 1 = 1172 - 1 = **1171**

Residuals Degrees of Freedom = 1171 - 4 = **1167**

Sum of Squares (group): We have the mean squared error for the group as x / 4 = 501.54 => the sum of squares (group) = 501.54 * 4 = **2,006**

```
501.54 * 4
```

```
## [1] 2006.16
```

Sum of Squares (total): 2,006 + 267,382 = **269,388**

```
2006 + 267382
```

```
## [1] 269388
```

Mean squared Error (residuals) = 267,382 / 1167 = **229.12**

```
267382 / 1167
```

```
## [1] 229.1191
```

F = MSG / MSE = 501.54 / 229.12 = 2.189

```
501.54 / 229.12
```

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

Find p-value for F (ONE TAILED!!!) = .06812

```
1 - pf(2.189, 4, 1167)
```

```
## [1] 0.06819242
```

d) What is the conclusion of the test?

Since the p value of .06812 exceeds the significance value of .05, we fail to reject H0. The data do not provide evidence that the average number of hours worked varies across the five groups.