Callin Switzer – Stat 139 Midterm Corrections – 26 October 2014

1ci. (3 points) If you run a rank-sum test on the same data, will it also reject the null? Explain.
What my answer should have been:

Maybe. The rank sum test is less powerful than the t-tools.

My incorrect answer:

Yes.

Why Incorrect:

We're not sure if the rank-sum test would have been significant, because it has less power than t-tools.

1cii. Suppose that the rank-sum test also rejected the null hypothesis, would you expect a permutation test (using the difference in sample means as a test statistic) to reject the null as well? Explain.
What my answer should have been:

Probably, because the permutation test has more power than the rank-sum test.

My incorrect answer:

Yes.

Why Incorrect:

I didn't realize that the permutation test could have less power than the t-test.

1d. You were given two samples of the same size. The samples were collected independently. However, you mistakenly treated them as paired data and ran a paired t-test. Which of the following was (were) affected by this mistake: difference in sample means, its standard error, significance level of the test, or reference distribution? Would you have understated or overstated the p-value? Explain.

What my answer should have been:

Sample-means -> Same
Std err -> very close to the same
Ref. dist -> different
Alpha -> different

Overestimate p-value, because the sampling distribution for the test statistic will have larger tails for the paired t-test.

My incorrect answer:

Underestimate, because a paired t-test has more power.

Why Incorrect:

My original answer was incorrect because a paired t-test will only have more power if there is a dependence between the two groups…otherwise, it will have less power.

2a. What is the study unit?

What my answer should have been:

Home sales

My incorrect answer:

Homes in BCS that were sold b/w 10/2013 and 10/2014

Why Incorrect:

My answer was incorrect because homes could be sold multiple times.

2ai. Use this information to answer the question whether there is a difference in average sale prices between townhouses and single-family homes with 2,000 to 2,200 sq.ft. located in the BCS area. Name the statistical test that you are performing and specify the hypotheses that you are testing, defining and using symbols to represent population parameters. Show your calculation of a test statistic and specify a corresponding (approximate) p-value.

What my answer should have been:

$$S_p = \frac{27(0.40)^2 + 17*(0.46)^2}{28 + 18 - 2} = 0.179$$

$$T = \frac{1.17 - 0.87}{sqrt(0.179)*sqrt\left(\frac{1}{18}+\frac{1}{28}\right)} = 2.347$$

Df = 44

p-value = < 0.05

My incorrect answer:

I only missed the portion about pooling the variances.

$$S_p = \frac{(39)(0.92)^2 + 27(0.40)^2 + 17*(0.46)^2}{40 + 28 + 18 - 3}$$

Df = 83

Why Incorrect:

I used pooled variance for all three groups, when I should have just pooled the variance for the two groups I was testing (because the variances were not equal for all three groups).

2cii. Suppose you rejected the null hypothesis in part 2(c)ii. Formulate the conclusion that your colleague can present to her client. Comment on the scope of inference.

What my answer should have been:

There was a random sample and no random assignment. This means we can make inferences about the population, but not causal inferences.

My incorrect answer:

I thought that there was no random sample – I misread the question, where it stated that there was a random sample.

Why Incorrect:

I just didn’t read the part of the question that said a random sample was taken.

2di. Decide which R output is the most appropriate for this question and indicate its number. Specify which test is being performed and explain your choice.

What my answer should have been:

V.

My incorrect answer:

IV.

Why Incorrect:

I circled the correct answer, but accidentally wrote down the IV instead of V. I described the answer correctly.

2dii. Write down the hypotheses that are being tested in the test chosen in part 2(d)i in terms of parameters defined for distributions of list and sale prices. Specify the parameter that is being estimated and report the corresponding estimate. Also, report the test statistic and the p-value.

What my answer should have been:

H0: median(Xi/Yi) = 1
Ha: “ “ not = 0

My incorrect answer:

H0: median(Xi) – median(Yi) = 0
Ha: “” not = 0

Why Incorrect:

Log transformation is interpreted as the median of the ratio, rather than the difference in medians.

2div (4 points) What is the 95% confidence interval for the parameter estimated in part 2(d)ii? Interpret the interval in the context of the original question of interest.

What my answer should have been:

ln(Xi/Yi) = [0.009026, 0.0483]

Converted to original scale: exp(0.009026) = 1.009067, exp(0.0483) = 1.049485

I estimate the ratio of medians is between 1.009 and 1.049

My incorrect answer:

Diff in medians = [0.00926, 0.0483]

Why Incorrect:

I didn’t interpret on the original scale.

2dvii. (3 points) You noticed that the sample size in your assistant's analysis was smaller than the original one. Further investigation revealed that all records on short sales (i.e., homes sold out of necessity and, usually, for less than they are worth) are missing list prices and, therefore, were not included in the analysis. Comment on whether excluding these sales affected the analysis and, if so, how. Think of a situation when it may be acceptable to exclude them.

What my answer should have been:

Yes excluding would effect analysis. It would make the estimated prices higher than the actual mean. It would be acceptable to exclude these houses when the client wasn't interested in short sales – because they don't want to deal with crazy stuff happening with the sellers.

My incorrect answer:

Yes. Exclude for short-term reports

Why Incorrect:

I ran out of time and didn't include enough writing to get full credit.

3b. (2 points) Write down the estimator of the pooled variance, Sp2, that uses all three samples to estimate 2.

What my answer should have been:

$$\boldsymbol{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \ (n_3 - 1)S_3^2}{(n_1 + \ n_2 + n_3 - 3)}}$$

My incorrect answer: nothing.

Why incorrect: I ran out of time.

3c. (4 points) Using parts 3a and 3b, derive the sampling distribution of Sp2.

What my answer should have been:

Fact:

$$\frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} + \frac{(n_3-1)S_3^2}{\sigma^2}$$

$$= \chi^2_{n_1-1} + \chi^2_{n_2-1} + \chi^2_{n_3-1}$$

$$= \chi^2_{n_1+n_2+n_3-3}$$

So. . .

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{n_1+n_2+n_3-3}$$

$$= \frac{\sigma^2\chi^2_{n_1-1} + \sigma^2\chi^2_{n_2-1} + \sigma^2\chi^2_{n_3-1}}{n_1+n_2+n_3-3}$$

$$= \frac{\sigma^2\chi^2_{n_1+n_2+n_3-3}}{n_1+n_2+n_3-3} \leftarrow \text{sampling dist of } S_p^2$$

My incorrect answer: nothing.

Why incorrect: I ran out of time.

3d. (2 points) What is the expected value of Sp2?
What my answer should have been:

$$\boldsymbol{E[S_p^2] = \frac{(n_1-1)E[S_1^2] + (n_2-1)E[S_2^2] + (n_3-1)E[S_3^2]}{(n_1+n_2+n_3-3)}}$$

$$\boldsymbol{E[S_p^2] = \frac{(n_1-1)\sigma^2 + (n_2-1)\sigma^2 + (n_3-1)\sigma^2}{(n_1+n_2+n_3-3)}}$$

$$\boldsymbol{E[S_p^2] = \sigma^2}$$

My incorrect answer: I just wrote a random formula from my cheatsheet.
Why incorrect: I ran out of time.

3e. Using definitions of Sp2 and Si2, i = 1, 2,, and 3, show that this equality holds.
What my answer should have been:

$$S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

$$SSR = \sum_{j=1}^{n_1}(Y_{1j}-\bar{Y}_1)^2 + \sum_{j=1}^{n_2}(Y_{2j}-\bar{Y}_2)^2 + \sum_{j=1}^{n_3}(Y_{3j}-\bar{Y}_3)^2$$

$$SSR = (n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2$$

$$S_p^2 = \frac{\overset{\text{Fact}}{(n-1)S_1^2 + (n_2-1)S_2^2 + (n-1)S_3^2}}{(n_1+n_2+n_3-3)}$$

So. . .

$$\boxed{SSR = Sp^2(n_1+n_2+n_3- -3)}$$

My incorrect answer: nothing.
Why incorrect: I ran out of time.

4(f) (2 points) Use the equality in the previous part and the result in part 3c to determine the sampling distribution of SSR.

What my answer should have been:

we know $$S_p^2 \sim \frac{\sigma^2 \chi^2_{n_1+n_2+n_3-3}}{n_1+n_2+n_3-3}$$

$$SSR = S_p^2 \cdot (n_1+n_2+n_3-3)$$

So...

$$SSR \sim \sigma^2 \chi^2_{n_1+n_2+n_3-3}$$

My incorrect answer: nothing.
Why incorrect: I ran out of time.