Solutions Autumn 2017

All questions are worth 10 points. Bullet points below each question give general grading instructions.

1 Solutions

1. Exercise 1

1a) Because of plagiarism Group A has very similar responses. This would results in a low variance. The variance is the squared deviations from the mean for every observation, so if all gave the same response the variance would be 0. If students would make some mistakes on the way, the variance would increase.

$$s^2 = \frac{\sum(X - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_n - \overline{X})^2}{n-1}$$

If $X = \overline{X}$ and we have identical answers, the variance will be 0. If one answer is a bit different, then $X \neq \overline{X}$, and the variance will grow. If there are many different answers, the variance will grow more.

Let's try to see this with a toy example. We have two sequences of 5 students:
1) 20,20,20,20,21 and 2) 19,20,21,22,20

The variance of the first sequence is 0.2. The variance of the second sequence is 1.3. In the first sequence we have mostly similar responses, with a smaller variance. In the second sequence the responses differ more, and the variance is higher.

- Long argument with no clear intuition -8

1b) This question should be solved with an F - test.

$$H_0 : \frac{Var(RQ1_B)}{Var(RQ1_A)} = 1$$

$$H_1 : \frac{Var(RQ1_B)}{Var(RQ1_A)} \neq 1$$

The test statistic is:

$$F = \frac{10.06^2}{1.16^2} = 75.21$$

which is F-distributed with 49,49 degrees of freedom. The cutoff value is $F_{45,45,0.005} = 2.19$. We reject the null of equality of variances.

Note: The larger variance goes in the numerator of the F-statistic. If the larger variance is in the denominator, then the student has to transform the cutoff value by taking the inverse and picking the cut-off value itself with flipped df - $\frac{1}{F_{49,49}} \approx \frac{1}{2.19} = 0.46$

The test indicates that maybe students of group A plagiarized.

- Correct test, but with calculation mistake -5
- Puts the smaller variance on top, without correcting the critical value -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5

1c) This question should be solved with T-test for unequal variances (based on the result of the F-test above):

$$H_0 : Mean(RQ1_B) = Mean(RQ1_A)$$

$$H_1 : Mean(RQ1_B) = Mean(RQ1_A)$$

The test statistic is:

$$T = \frac{20.1 - 18.8}{\sqrt{\frac{1.16^2+10.06^2}{50}}} \approx -0.9$$

which follows a Student t distribution with $\frac{(\frac{1.16^2+10.06^2}{50})^2}{\frac{(1.16^2/50)^2+(10.06/50)^2}{49}} \approx 50$ degrees of freedom. The cutoff value is $t_{50,0.005} = 2.678$, which means we can't reject the null hypothesis of no difference in means

The test indicates that the two groups of students performed similarly.

- Correct test, but does not calculate the correct df -3
- Correct test, but with calculation mistake -5
- [redacted]
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5

1d) This question should be solved with a test for equality between two proportions

$$H_0 : P(Rest_A) = P(Rest_B)$$

$$H_1 : P(Rest_A) \neq P(Rest_B)$$

The test statistic is:

$$z = \frac{p_A - p_B}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}} = \frac{0.98 - 0.58}{\sqrt{\frac{0.78*0.22*2}{50}}} = \frac{0.4}{0.083} \approx 4.81 > z_{0.005} = 2.57$$

with a different rounding:

$$z = \frac{0.4}{0.08} \approx 5 > z_{0.005} = 2.57$$

Students in Group A outperformed students in Group B, as they answered correctly a significantly larger proportion of the questions.

- Correct test, but does not calculate the correct cutoff value -5
- Correct test, but with calculation mistake -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5

1e) Performing a one sided test at the 1 percent significance level means that the new cutoff value would be lower than the one in the tests from point a. Implicated would be tests with a p-value between 0.01 and 0.005. None of the tests has such a p-value, therefore none of them is affected.

Alternatively, students can take the cutoff values corresponding to one sided test for each of the above tests:

a) $F_{45,45,0.01} = 2.02$
b) $t_{50,0.01} = 2.403$
c) $z_{0.01} = 2.32$

The new cut-off values lead to the same conclusion as the ones used in the tests above.

1f) There are 2 factors: Gender and Group. The factor levels for Gender are Female and Male. The factor levels for Group are A and B.

For the ANOVA, we observe 3 tests:

- Test for Differences between the Levels of Factor Group

$$\begin{aligned} H_0 : \overline{Total_A} &= \overline{Total_B} \\ H_1 : \overline{Total_A} &\neq \overline{Total_B}. \end{aligned}$$

- Test for the Differences between the Levels of Factor Gender

$$\begin{aligned} H_0 : \overline{Total_F} &= \overline{Total_M} \\ H_1 : \overline{Total_F} &\neq \overline{Total_M}. \end{aligned}$$

- Test for Interaction between Factor A and B

$$\begin{aligned} H_0 &: \overline{RQ1_{A,male}} = \overline{RQ1_{B,male}} = \overline{RQ1_{A,female}} = \overline{RQ1_{B,female}} \\ H_1 &: At \quad least \quad two \quad means \quad differ \\ H_2 &: \overline{RQ1_{A,male}} \neq \overline{RQ1_{B,male}} \neq \overline{RQ1_{A,female}} \neq \overline{RQ1_{B,female}} \end{aligned}$$

Both statements H_1 and H_2 are correct for specifying the alternative hypothesis

The condition for equal variances seems to be violated.

1g) We observe that there are significant differences between Group A and Group B in their total scores. We reject the null hypothesis for the first test at the 1 % level.

We observe that there are no significant differences between genders and we do not reject the null of no differences.

The third line means that there is no significant interaction between gender and cheating - in other words, females are not more or less likely to cheat than males.

1h) See Table 1. The ranked observations are as follows:

Group A: 134,133,131,131,127, (126) , 116,111. In the brackets we have the first value for Group B. Rank sum: 1+2+3+4+5+7+8=30

Group B: 126,107,92,84,81,69,67. Rank sum: 6+9+10+11+12+13+14=75

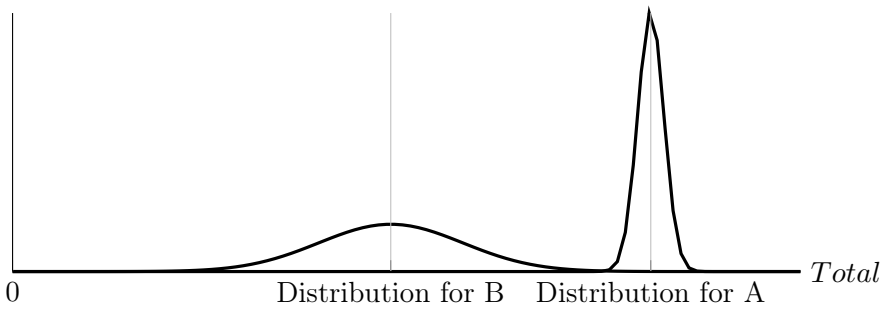
The test statistic has either a value of 30 or 75, which falls outside of the critical values, which are 37 and 68.

We reject the null of equal location between the two distributions. This means that students from group A are drawn from a different distribution than students of Group B, so they are not from the same distribution and the two groups are quite different.

Table 1: Wilcoxon Rank Sum Test

Rank	Group A	Group B
1	134	
2	133	
3	131	
4	131	
5	127	
6		126
7	116	
8	111	
9		107
10		92
11		84
12		81
13		69
14		67
Rank Sum	30	75

Figure 1: Distributions



- Correct test, but does not state hypothesis -5
- Correct test, but with calculation mistake -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5

1i) The mean total scores for the two groups of students are different, with different variances.The two distributions have a different location, the distribution of Group A is significantly to the right of the distribution of Group B. The distribution of Group A has a smaller variance than the distribution of Group B. Overall, students from Group A seem to outperform students from Group B, which is likely due to cheating. See Figure 1

2. Exercise 2

2a) The difference between panel data and cross-section data is that the panel includes repeated observations over time on the same units, while a cross-section is one observation in time on many units.

Given that the number of observations are 2,870, and we have 10 years, we can compute that there are $2870/10 = 287$ counties in the data.

Let the number of New Mexico counties be X. $X * 5 = 0.06 * 2870$. MML is equal to 1 for 6 percent of the observations, which is 5 years $\times$ all the New Mexico counties. So, $X = 0.06 * 2870/5 \approx 34$. The real answer is 33, but the MML variable is rounded upwards, so this is likely to generate a higher estimate.

2b) MML is likely an indicator variable, which takes a value 1 whenever the MML legislation is active and 0 otherwise. To find the significance level we can either calculate the t-statistic or use the calculated F-statistic that comes with the output- The F-statistic will have approximately 2870 (the number of obs)- 2 (the variables: MML and constant) degrees of freedom, giving us a p-value of close to 0. (or a significance level of 1 percent)

The t-statistic can be calculated as follows:

$$t = \frac{117.5 - 0}{16.76} \approx 7.01$$

This gives us a p-value below 0.005, as the cutoff is $t_{0.005,\infty} = 2.576$

In years with MML we observe 117 more crimes.

- Correct test, but with calculation mistake -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5
- No interpretation of MML -5
- Wrong interpretation of MML -7

2c) In column (3) we add county fixed effects, and the estimate turns negative. The fit increases to Rsquared of 0.67. When we add year fixed effects in column (4) we account for differences between counties and years, and we obtain -46.72 as an estimate. The additional year fixed effects increase the model fit to 0.674. The fixed effects account for all factors that are constant for a county such as geographic location, distance from Mexico, demographic variables, climate etc. The time fixed effects account for events such as the financial crisis or very cold years, whose effect is fixed in time. By adding implicitly all these variables to the model, we explain most of the variation in the crime rates.

- Long unclear answer -5
- Saying just "R^2 increases" with nothing more -2
- Mechanical reading, without demonstrating understanding: -5 or more

2d) My preferred model is in column (5), because it absorbs both year and county differences. The inclusion of control variables allows me to control for other factors that might influence crime. The effect of MML on crime is significantly negative for New Mexico, therefore introducing MML seems to lead to a decrease in crime. However, there are many other social costs that are also associated with MML, and this is just one effect.

Semi-elasticity which will be asked about in future exams: Crime decreases with 18 ($51.05/274.41 = 0.18$) percent after the introduction of MML.

- Long unclear answer -5
- Mechanical reading, without demonstrating understanding: -5 or more

2e) In column (2) we observe that an increase in the median income is associated with a decrease in crime, but the effect is not significant. This is consistent with economic theory. However, in model (5) we observe that the coefficient on median income is positive, so that an increase in median income will lead to an increase in crime, but the effect is not significant again. We conclude that the median income does not seem to have an effect on crime, contrary to the suggestion by the economic theory on crime.

The following 2 t-statistics will give the significance:

$$t = \frac{-5.365}{17.2} \approx -0.31$$

$$t = \frac{82.85}{63.61} \approx 0.19$$

Both values are below the most lenient one-sided cutoff value $t_{0.1,\infty} = 1.282$

- Correct test, but with calculation mistake -5
- Correct test, but wrong conclusion -7
- Correct test, but no conclusion -5
- No interpretation -5
- Wrong interpretation -7

2f) In column (5) we take the difference between two point predictions :

The first one is: $-51.05 * Mean(MML) - 670.2 * 0.02 + 82.85 * Mean(MedianIncome) + FixedEffects$

The second one is: $-51.05 * Mean(MML) - 670.2 * 0.97 + 82.85 * Mean(MedianIncome) + FixedEffects$

By decreasing the population of Hispanics we are likely to go from a high share Hispanics to a low share of Hispanics. We can bound the effect by subtracting from the predicted crime with the maximum the predicted crime with the minimum, so:

After canceling out the similar arguments we get $-670.2 * 0.97 - (-670.2 * 0.02) = -670.2 * 0.95 = -636.69$

So, counties with the maximum fraction of a Hispanic population have 636 less violent crimes than counties with the minimum fraction. Therefore, it is likely that crime could increase following the policy proposed by Trump, which would lead to a decrease in the population of Hispanics. This statement is not causal, it shows a correlation.

- Correct prediction, but with calculation mistake -5
- Correct prediction, but wrong conclusion -7
- Correct prediction, but no conclusion -5
- No interpretation -5
- Wrong interpretation -7