Problem Set 3

EC 303: Empirical Economic Analysis

Due October 30, 2019 at 3:30pm.

1 Theoretical Problems

Problem 1.1: Continuous Joint Random Variables.** Complete Exercise 5.9 from the textbook. Note that the solution in the back of the textbook has a typo.

Problem 1.2: Properties of the Covariance.** Prove these properties about the covariance function:

a. The shortcut formula: $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X \mu_Y$

b. $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$

c. $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$

Problem 1.3: Working with Correlations.** a. Use the properties you proved in the previous problem, along with variance properties, to prove that $\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$ when a and c have the same sign.

b. What happens to the formula in (a.) when a and c have opposite signs?

c. For two random variables X and Y, let Z_X and Z_Y be the standardized variables (e.g., $Z_X = (X - \mu_X)/\sigma_X$. Prove that $\text{Corr}(X, Y) = \text{Cov}(Z_X, Z_Y)$.

Problem 1.4: Linear Combinations. Let X_1 and X_2 be two random variables, and let $a_1, a_2 \in \mathbb{R}$. Consider the linear combination $Y = a_1 X_1 + a_2 X_2$.

a. What is the variance $\mathbb{V}[Y]$ in terms of $\mathbb{V}[X_1], \mathbb{V}[X_2]$, and $\text{Cov}(X_1, X_2)$?

b. How does the equation for $\mathbb{V}[Y]$ simplify when X_1 and X_2 are independent random variables?

c. If you know that $\text{Cov}(X_1, X_2) = 0$, can you conclude that the two variables are independent? If you think so, prove the statement; if not, provide a simple counterexample.

d. Now suppose that you have n random variables: $\{X_1, ..., X_n\}$. How do your answers to (a) and (b) generalize to the case of $\mathbb{V}\left[\sum_{i=1}^n a_i X_i\right]$? There is no need to prove your answer.

Hints for (d.):

- Try a simple case of $n = 3$ or 4 to fix your intuition.
- Begin with the simpler case where your collection of random variables are all independent.
- Try to write the formula as simply as possible, using summation notation (you'll need a double sum to handle the covariances).

Problem 1.5: Joint and Conditional Moment Generating Functions. Since the joint and conditional distributions are valid pmfs/pdfs, we can work with joint and conditional **moment generating functions** as well. To see how this works, consider a restaurant with two phones to receive take-out orders. The waiting time between calls for each phone is exponentially distributed with a mean of 1 minute. Suppose that X_1 and X_2 are jointly distributed with the joint pdf

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 2e^{-(x_1+x_2)} & x_1, x_2 \in (0, \infty) \\ 0 & \text{otherwise.} \end{cases}$$

a. The *joint* moment generating function is defined as:

$$M(s,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{sx+ty} f(x,y)dxdy,$$

for constants s and t. What is the moment generating function for this joint pdf? (*Note:* The joint MGF will be defined whenever $s, t < 1$).

b. What is the conditional distribution of $X_2|X_1 = x$?

c. The conditional distribution is also a valid pdf. Derive its MGF (called the *conditional* MGF), and use it to find the conditional expectation $\mathbb{E}[X_2|X_1 = x]$ and conditional variance $\mathbb{V}[X_2|X_1 = x]$. What does the conditional expectation suggest to you? Notice that this method saves you from having to do integration by parts, which is pretty handy.

Problem 1.6: Sampling Distributions I.** Complete Exercise 6.5 from the textbook.

Problem 1.7: Sampling Distributions II. a. Complete Exercise 6.23 from the textbook.

b. Complete Exercise 6.24 from the textbook.

Extra Credit Theory Problem

Problem 1.8: Independence of $\overline{X}$ and S^2. We are frequently interested in estimating the mean and variance from a sample. This problem helps you prove that these two random variables are independent, a useful fact in a lot of econometric proofs.

First, recall that the square of a standard normal follows a χ^2_ν distribution (where ν denotes the degrees of freedom), which has a moment generating function given by $M_X(t) = (1-2t)^{-\nu/2}$.

a. Prove that the sum of two independent χ^2 variables with degrees of freedom ν_1 and ν_2, respectively, is also distributed as $\chi^2_{\nu_1+\nu_2}$. (*Hint:* Use the fact that for independent random variables, the MGF of their sum is the same as the product of the individual MGFs).

b. How does your result in (a) generalize to the sum of n independent random variables with a χ^2 distribution? Try to prove this result (*Hint*: Induction) and then use it to show that for an i.i.d. sample of standard normals, $\{Z_1, ..., Z_n\}$, the statistic $S = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$.

Thus, the χ^2 distribution is preserved under sums, and we can add standard normals together to get a χ^2 distribution. Now recall the formula for the sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})$. However, if the true mean μ is known, the statistic can be written as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2.$$

c. Assume that each X_i is i.i.d. normally distributed with mean μ and standard deviation σ. Show that if one transforms the statistic $\hat{\sigma}^2$ by an appropriately chosen constant c, the result can be written as the sum of squared standard normal distributions. What is the distribution of $\hat{\sigma}^2$?

d. Expand the sum of squared normals to show that

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

(*Hint:* Add and subtract $\overline{X}$ to each numerator of the sum, and then expand the quadratic using $(X_i - \overline{X})$ and $(\overline{X} - \mu)$.)

e. Write the first term of the right-hand side of this equation in terms of S^2.

This equation then says that the sum of S^2 and a χ^2 random variable is also χ^2. If we can show that the two terms on the right-hand side of the equation are independent, we will have therefore established that S^2 is a χ^2 random variable.

f. Show that $\overline{X}$ is uncorrelated with any observation's deviation from the mean. That is, show that $\text{Cov}(X_i - \overline{X}, \overline{X}) = 0$.

g. What is the joint distribution of $(X_i - \overline{X}, \overline{X})$? Why does knowing this distribution help to show that the two are independent (using your answer from (e))?

Given this result, we conclude that $\overline{X}$ and S^2 are independent, and that $S^2 \sim \chi^2$.

2 Stata Exercises

Problem 2.1: Covariance, Correlation, and Graphing. For this problem, use the Excel file called Uninsured (PS3) on Blackboard. This data set contains information on 20 municipalities in Massachusetts. For each municipality, the fraction of people without health insurance (frac_uninsured) and the fraction of people declaring bankruptcy (frac_bankrupt) are reported. Read this into Stata (the first row consists of variable names).

a. What is the covariance between these two variables? Make a nice-looking scatterplot of the variables' relationship. How does the covariance you reported jibe with the graph? Why do you think this is?

b. Create new variables for both bankruptcy and (un-)insurance that is measured in people (rather than percentages). Use the population variable to do so. Does this change the linear relationship? What is the new covariance? What does this teach you about covariance and data viz?

c. As discussed in class, the correlation is a unitless measure that resolves some of the problems discussed above. What is the correlation between the two original variables? Does this correlation change when you use the new variables (based on people, not percentages) instead? Why (or why not)?

d. What is the correlation between frac_uninsured and your count of uninsured? What explains this? Why doesn't this cause a problem in calculating the correlation between your new variables? (A scatter plot—or multiple plots—may be useful here.)

e. However, even looking at the correlation can be misleading. Create a data set of 1000 observations and 2 variables: X that ranges continuously over the interval $[0, 5]$ and Y given by $y = -x(x - 5)$. Create a scatter plot of this relationship. What economic variables may have this relationship? What is the correlation between X and Y, and what drives this result? When should I be careful of looking only at the correlation coefficient ρ?

Problem 2.2: Regression to the Mean. We are all susceptible to mistaken interpretations of data analysis. One of the most common is the **regression fallacy**, in which we observe data moving from abnormal to normal values over time, and assume that this was caused by a particular action. This problem illustrates the flaws behind that reasoning, based on a classic randomized control trial evaluating the effectiveness of a treatment for hypertension (Reader, 1979).

Suppose that we want to run an experiment measuring how effective a new treatment is at reducing high diastolic blood pressure, D. For simplicity, suppose that there are two types of patients: those with regular blood pressure, for whom $D \sim \mathcal{N}(90, 16)$ and those with high blood pressure, for whom $D \sim \mathcal{N}(100, 16)$. The fraction of those in the population with truly higher than usual blood pressure is 10%.

Suppose that you have a randomly selected sample of $n = 1000$ patients. For each patient, you measure their initial blood pressure by drawing from their corresponding distribution.

a. Use the rnormal() command to simulate a blood pressure reading for 1000 patients, 10% of which have high blood pressures and 90% of which do not. Based on your measurement, assign patients to the following categories: (i) DBP measured between 90 and 99, (ii) DBP measured between 100 and 104, and (iii) DBP measured at 105 or higher. Make a table that reports the number of patients observed in each category and the mean blood pressure within the group.

- Set the seed to 122333 so that we have replicable results.

b. Your first pass at the experiment is to use only patients in each of these categories. You administer the treatment to each person and then re-measure their blood pressure after the treatment has ended. For now, suppose that the treatment was ineffectual (so that underlying distributions did not change), and re-sample blood pressures for all those in the three categories above. Add columns to your table with the new in-group average as well as the difference between averages.

c. What do you find? If this were the data you were given, what would you conclude about the drug's effectiveness? What do you think is *actually* going on here? (*Hint:* The phenomenon above is called *regression toward the mean*; if you like, you can Google the phenomenon to get a clearer picture of what your last answer might be.)

d. Suppose that we want to isolate a true treatment effect from this regression toward the mean. What would we need in order to do this? (*Hint:* Think about a typical experiment setup. What do researchers do to strip out effects that aren't related to the treatment?)

e. For each of the three categories, split your sample randomly into treatment and control groups Calculate the estimated treatment effect using the following steps:

  - For each of the three categories, use runiform() < 0.5 to select the treated group. *Do not* reassign or move people between the three categories. All others are control patients.
  - Based on the initial measurement used to categorize patients, report the within-group average for each of the 6 groups (category 1 control, category 1 treatment, category 2 control, etc.).
  - For each of the 6 groups, take a second measurement and report the within-group average, as well as the difference for each group.
  - For each category (1, 2, and 3), the estimated average treatment effect is (the difference in the treated group) − (the difference in the control group). Report the estimated average treatment effects.

  What do you find? Did including the control group help?

f. Now suppose that there *is* a treatment effect: those who take the drug and have high blood pressure have their new distribution almost perfectly centered around a normal blood pressure: $D_{\text{treat}} \sim \mathcal{N}(90, 1)$. Those whose blood pressures are normal are unaffected by the drug. Estimate the treatment effects for each of the three categories in this case, and discuss your results.

Extra Credit Stata Problem

Problem 2.3: Simulation I. This problem asks you to perform simulations as in Chapter 6.

a. 6.10 (revised): Carry out a simulation experiment to study the sampling distribution of $\overline{X}$ when the population distribution is log-normal with $\mathbb{E}[ln(X)] = 3$ and $\mathbb{V}[ln(X)] = 1$. Consider the four sample sizes $n \in \{10, 20, 30, 50\}$, and in each case use 500 replications.

  - For each of the 4 simulations, provide a histogram of the sampling distribution.
  - Provide 1 table comparing the means, medians, and standard deviations of the sampling distributions across the 4 simulations. Are the sampling distributions symmetric? Are the results consistent with the LLN? Discuss.
  - For which of these sample sizes does the $\overline{X}$ sampling distribution appear to be approximately normal?
  - *Hint:* You will need to use the command set matsize 500 at the beginning of the code for this problem in order to use $k = 500$ in your MATA code.

b. 6.26 (revised): A friend commutes by bus *to and from* work 5 days/week. Suppose that waiting time is uniformly distributed between 0 and 10 min, and that all waiting times are independent of each other.

- What is the approximate probability that total waiting time for an entire week is at most 60 min? Use $k = 500$ replications.
- The idea of this problem is that even for an n as small as 10, T_0 and $\overline{X}$ should be approximately normal when the parent distribution is uniform. What do you think?

3 References

- Reader, R. (1979). Initial Results of the Australian Therapeutic Trial in Mild Hypertension: Report by the Management Committee.