Review Questions

Econ 103

Spring 2018

About This Document

These questions are the "bread and butter" of Econ 103: they cover the basic knowledge that you will need to acquire this semester to pass the course. There are between 10 and 15 questions for each lecture. After a given lecture, and before the next one, you should solve all of the associated review questions. To give you an incentive to keep up with the course material, all quiz questions for the course will be randomly selected from this list. For example Quiz #1, which covers lectures 1–2, will consist of one question drawn at random from questions 1–10 and another drawn at random from questions 12–24 below. The review questions are straightforward. Most are taken directly from the lecture slides and nearly all of the rest involve calculations similar if not identical to those from the lecture. As such, we will not circulate solutions to review questions. Compiling your own solutions is an important part of studying for the course. We will be happy to discuss any of the review questions with you in office hours or on Piazza, and you are most welcome to discuss them with your fellow classmates. Be warned, however, that merely memorizing answers written by a classmate is a risky strategy. It may get you through the quiz, but will leave you woefully unprepared for the exams. There is no curve in this course: to pass the exams you will have to learn the material covered in these questions. Rote memorization will not suffice.

Lecture #1 – Introduction

1. Define the following terms and give a simple example: *population*, *sample*, *sample size*.
2. Explain the distinction between a *parameter* and a *statistic*.
3. Briefly compare and contrast *sampling* and *non-sampling* error.
4. Define a *simple random sample*. Does it help us to address sampling error, non-sampling error, both, or neither?

5. A drive-time radio show frequently holds call-in polls during the evening rush hour. Do you expect that results based on such a poll will be biased? Why?

6. Dylan polled a random sample of 100 college students. In total 20 of them said that they approved of President Trump. Calculate the margin of error for this poll.

7. Define the term *confounder* and give an example.

8. What is a randomized, double-blind experiment? In what sense is it a "gold standard?"

9. Indicate whether each of the following involves experimental or observational data.
   (a) A biologist examines fish in a river to determine the proportion that show signs of disease due to pollutants poured into the river upstream.
   (b) In a pilot phase of a fund-raising campaign, a university randomly contacts half of a group of alumni by phone and the other half by a personal letter to determine which method results in higher contributions.
   (c) To analyze possible problems from the by-products of gas combustion, people with with respiratory problems are matched by age and sex to people without respiratory problems and then asked whether or not they cook on a gas stove.
   (d) An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure rate of its pumps.

10. Based on information from an observational dataset, Amy finds that students who attend an SAT prep class score, on average, 100 points better on the exam than students who do not. In this example, what would be required for a variable to *confound* the relationship between SAT prep classes and exam performance? What are some possible confounders?

Lecture #2 – Summary Statistics I

11. For each variable indicate whether it is nominal, ordinal, or numeric.
   (a) Grade of meat: prime, choice, good.
   (b) Type of house: split-level, ranch, colonial, other.
   (c) Income

12. Explain the difference between a histogram and a barchart.

13. Define *oversmoothing* and *undersmoothing*.

14. What is an *outlier*?

15. Write down the formula for the sample mean. What does it measure? Compare and contrast it with the sample median.

16. Define *range* and *interquartile range*. What do they measure and how do they differ?

17. What is a boxplot? What information does it depict?

18. Write down the formula for variance and standard deviation. What do these measure? How do they differ?

19. Suppose that x_i is measured in inches. What are the units of the following quantities?
    (a) Sample mean of x
    (b) Range of x
    (c) Interquartile Range of x
    (d) Variance of x
    (e) Standard deviation of x

20. Evaluate the following sums:
    (a) $\sum_{n=1}^{3} n^2$
    (b) $\sum_{n=1}^{3} 2^n$
    (c) $\sum_{n=1}^{3} x^n$

21. Evaluate the following sums:
    (a) $\sum_{k=0}^{2} (2k+1)$
    (b) $\sum_{k=0}^{3} (2k+1)$
    (c) $\sum_{k=0}^{4} (2k+1)$

22. Evaluate the following sums:
    (a) $\sum_{i=1}^{3} (i^2+i)$

(b) $\sum_{n=-2}^{2}(n^2-4)$

(c) $\sum_{n=100}^{102} n$

(d) $\sum_{n=0}^{2}(n+100)$

23. Express each of the following using Σ notation:

(a) $z_1 + z_2 + \cdots + z_{23}$

(b) $x_1y_1 + x_2y_2 + \cdots + x_8y_8$

(c) $(x_1 - y_1) + (x_2 - y_2) + \cdots + (x_m - y_m)$

(d) $x_1^3 f_1 + x_2^3 f_2 + \cdots + x_9^3 f_9$

Lecture #3 – Summary Statistics II

24. Show that $\sum_{i=m}^{n}(a_i + b_i) = \sum_{i=m}^{n} a_i + \sum_{i=m}^{n} b_i$. Explain your reasoning.

25. Show that if c is a constant then $\sum_{i=m}^{n} cx_i = c\sum_{i=m}^{n} x_i$. Explain your reasoning.

26. Show that if c is a constant then $\sum_{i=1}^{n} c = cn$. Explain your reasoning.

27. Mark each of the following statements as True or False. You do not need to show your work if this question appears on a quiz, although you should make sure you understand the reasoning behind each of your answers.

(a) $\sum_{i=1}^{n}(x_i/n) = \left(\sum_{i=1}^{n} x_i\right)/n$

(b) $\sum_{k=1}^{n} x_k z_k = z_k \sum_{k=1}^{n} x_k$

(c) $\sum_{k=1}^{m} x_k y_k = \left(\sum_{k=1}^{m} x_k\right)\left(\sum_{k=1}^{m} y_k\right)$

(d) $\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{m} y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j$

(e) $\left(\sum_{i=1}^{n} x_i\right) / \left(\sum_{i=1}^{n} z_i\right) = \sum_{i=1}^{n} (x_i/z_i)$

28. Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Justify all of the steps you use.

29. Write down the formula for skewness. Why does this formula involve a cubic, and why do we divide by s^2?

30. How do we interpret the sign of skewness, and what is the "rule of thumb" that relates skewness, the mean, and median?

31. What is the distinction between μ, σ^2, σ and $\bar{x}, s^2, s$? Which corresponds to which?

32. What is the empirical rule?

33. Define *centering*, *standardizing*, and *z-score*.

34. What is the sample mean $\bar{z}$ of the z-scores $z_1, \ldots, z_n$? Prove your answer.

35. What is the sample variance s_z^2 of the z-scores $z_1, \ldots, z_n$? Prove your answer.

36. Suppose that $-c < (a - x)/b < c$ where $b > 0$. Find a lower bound L and an upper bound U such that $L < x < U$.

37. Compare and contrast *covariance* and *correlation*. Provide the formula for each, explain the units, the interpretation, etc.

38. Suppose that x_i is measured in centimeters and y_i is measured in feet. What are the units of the following quantities?
    (a) Covariance between x and y
    (b) Correlation between x and y
    (c) Skewness of x
    (d) $(x_i - \bar{x})/s_x$

Lecture #4 – Regression I

39. In a regression using height (measured in inches) to predict handspan (measured in centimeters) we obtained $a = 5$ and $b = 0.2$.
    (a) What are the units of a?
    (b) What are the units of b?

(c) What handspan would we predict for someone who is 6 feet tall?

40. Plot the following dataset and calculate the corresponding regression slope and intercept *without* using the regression formulas.

x	y
0	2
1	1
1	2

41. Write down the optimization problem that linear regression solves.

42. Prove that the regression line goes through the means of the data.

43. By substituting $a = \bar{y} - b\bar{x}$ into the linear regression objective function, derive the formula for b.

44. Consider the regression $\widehat{y} = a + bx$.
    (a) Express b in terms of the sample covariance between x and y.
    (b) Express the sample correlation between x and y in terms of b.

45. What value of a minimizes $\sum_{i=1}^{n}(y_i - a)^2$? Prove your answer.

46. Suppose that $s_{xy} = 30$, $s_x = 5$, $s_y = 3$, $\bar{y} = 12$, and $\bar{x} = 4$. Calculate a and b in the regression $\widehat{y} = a + bx$.

47. Suppose that $s_{xy} = 30$, $s_x = 5$, $s_y = 3$, $\bar{y} = 12$, and $\bar{x} = 4$. Calculate c and d in the regression $\widehat{x} = c + dy$.

48. A large number of students took two midterm exams. The standard deviation of scores on midterm #1 was 16 points, while the standard deviation of scores midterm #2 was 17 points. The covariance of the scores on the two exams was 124 points squared. Linus scored 60 points on midterm #1 while Lucy scored 80 points. How much higher would we predict that Lucy's score on the midterm #2 will be?

49. Suppose that the correlation between scores on midterm #1 and midterm #2 in Econ 103 is approximately 0.5. If the regression slope when using scores on midterm #1 to predict those on midterm #2 is approximately 1.5, which exam had the larger *spread* in scores? How much larger?

Extensions

50. Question about post-stratification, non-response bias, etc.

51. Treat regression to the mean in the extensions.

52. The *mean deviation* is a measure of dispersion that we did not cover in class. It is defined as follows:

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

    (a) Explain why this formula averages the absolute value of deviations from the mean rather than the deviations themselves.

    (b) Which would you expect to be more sensitive to outliers: the mean deviation or the variance? Explain.

53. Show that $\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2$

54. Using the preceding with $m = \bar{x}$, show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

55. Consider a dataset $x_1, \ldots, x_n$. Suppose I multiply each observation by a constant d and then add another constant c, so that x_i is replaced by $c + dx_i$.

    (a) How does this change the sample mean? Prove your answer.

    (b) How does this change the sample variance? Prove your answer.

    (c) How does this change the sample standard deviation? Prove your answer.

    (d) How does this change the sample z-scores? Prove your answer.

56. Assign them to read the chapter from "Thinking Fast and Slow" and write a one paragraph summary of *regression to the mean*.

57. Let

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x}, \quad \text{and} \quad z_{y_i} = \frac{y_i - \bar{y}}{s_y}.$$

    Show that if we carry out a regression with z_{y_i} in place of y and z_{x_i} in place of x, the intercept a will equal zero while the slope b will equal r, the sample correlation.

58. Let $\hat{y}$ denote our prediction of y from a linear regression model: $\hat{y} = a + bx$ and let r be the correlation coefficient between x and y.

    (a) Express b in terms of s_{xy} and s_x.

    (b) Express a in terms of b and the sample means of x and y.

    (c) Express r in terms of the s_{xy}, s_x and s_y.

    (d) Show that

$$\frac{\hat{y} - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right)$$

(e) (3 points) Using the equation derived in (d), briefly explain "regression to the mean."

59. Lothario, an unscrupulous economics major, runs the following scam. After the first midterm of Econ 103 he seeks out the students who did extremely poorly and offers to sell them "statistics pills." He promises that if they take the pills before the second midterm, their scores will improve. The pills are, in fact, M&Ms and don't actually improve one's performance on statistics exams. The overwhelming majority of Lothario's former customers, however, swear that the pills really work: their scores improved on the second midterm. What's your explanation?