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. 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?

Solution: They will likely be biased. People who are listening to the radio during rush hour are disproportionately likely to be commuters driving home from work. People who are employed and drive to work are not representative of the population at large.

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.

Solution: $2\sqrt{P(1-P)/n} = 2\sqrt{0.2 \times 0.8/100} = 0.08$

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.

   Solution: Observational

   (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.

   Solution: Experimental

   (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.

   Solution: Observational

   (d) An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure rate of its pumps.

   Solution: Observational

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?

Solution:

Lecture #2 – Summary Statistics I

11. For each variable indicate whether it is nominal, ordinal, or numeric.
    (a) Grade of meat: prime, choice, good.

    Solution: ordinal

    (b) Type of house: split-level, ranch, colonial, other.

    Solution: nominal

    (c) Income

    Solution: numeric

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. Two hundred students took Dr. Evil's final exam. The third quartile of exam scores was 85. Approximately how many students scored *no higher* than 85 on the exam?
17. Define *range* and *interquartile range*. What do they measure and how do they differ?
18. What is a boxplot? What information does it depict?

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

20. Suppose that x_i is measured in inches. What are the units of the following quantities?

    (a) Sample mean of x

    Solution: inches

    (b) Range of x

    Solution: inches

    (c) Interquartile Range of x

    Solution: inches

    (d) Variance of x

    Solution: square inches

    (e) Standard deviation of x

    Solution: inches

21. Evaluate the following sums:

    (a) $\sum_{n=1}^{3} n^2$

    Solution: $\sum_{n=1}^{3} n^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$

    (b) $\sum_{n=1}^{3} 2^n$

    Solution: $\sum_{n=1}^{3} 2^n = 2^1 + 2^2 + 2^3 = 2 + 4 + 8 = 14$

    (c) $\sum_{n=1}^{3} x^n$

Solution: $\sum_{n=1}^{3} x^n = x + x^2 + x^3$

22. Evaluate the following sums:

(a) $\sum_{k=0}^{2}(2k+1)$

Solution: $\sum_{k=0}^{2}(2k+1) = (2\times 0+1)+(2\times 1+1)+(2\times 2+1) = 9$

(b) $\sum_{k=0}^{3}(2k+1)$

Solution: $\sum_{k=0}^{3}(2k+1) = \left[\sum_{k=0}^{2}(2k+1)\right] + (2\times 3+1) = 9+7 = 16$

(c) $\sum_{k=0}^{4}(2k+1)$

Solution: $\sum_{k=0}^{4}(2k+1) = \left[\sum_{k=0}^{3}(2k+1)\right] + (2\times 4+1) = 16+9 = 25$

23. Evaluate the following sums:

(a) $\sum_{i=1}^{3}(i^2+i)$

Solution: $\sum_{i=1}^{3}(i^2+i) = (1^2+1)+(2^2+2)+(3^2+3) = 20$

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

Solution: $\sum_{n=-2}^{2}(n^2-4) = \left[(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2\right]-(4\times 5) = -10$

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

Solution: $\sum_{n=100}^{102} n = 100 + 101 + 102 = 303$

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

Solution: $\sum_{n=0}^{2} (n + 100) = (0 + 1 + 2) + 3 \times 100 = 303$

24. Express each of the following using Σ notation:

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

Solution: $\sum_{i=1}^{23} z_i$

(b) $x_1 y_1 + x_2 y_2 + \cdots + x_8 y_8$

Solution: $\sum_{i=1}^{8} x_i y_i$

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

Solution: $\sum_{i=1}^{m} (x_i - y_i)$

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

Solution: $\sum_{i=1}^{9} x_i^3 f_i$

Lecture #3 – Summary Statistics II

25. 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.

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

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

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

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

30. Re-write the formula for skewness in terms of the z-scores $z_i = (x_i - \bar{x})/s$. Use this to explain the original formula: why does it involve a cubic and why does it divide by s^3?

Solution:

$$\frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3$$

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

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

33. What is the empirical rule?

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

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

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

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

Solution: Rearranging,

$$-bc - a < -x < bc - a$$

and multiplying through by -1,

$$a - bc < x < a + bc$$

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

39. 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

    Solution: centimeters $\times$ feet

    (b) Correlation between x and y

    Solution: unitless

    (c) Skewness of x

    Solution: unitless

    (d) $(x_i - \bar{x})/s_x$

    Solution: unitless

Lecture #4 – Regression I

40. 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?

41. 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

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

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

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

45. 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.

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

47. 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 $\widehat{y} = a + bx$.

48. 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 $\widehat{x} = c + dy$. Note: we are using y to predict x in this regression!

49. 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?

50. 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?

Lecture #5 – Basic Probability I

51. What is the definition of probability that we will adopt in Econ 103?

52. Define the following terms:
    (a) *random experiment*
    (b) *basic outcomes*
    (c) *sample space*
    (d) *event*

53. Define the following terms and give an example of each:
    (a) *mutually exclusive events*
    (b) *collectively exhaustive events*

54. Suppose that $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 3\}$, $B = \{3, 4, 6\}$, and $C = \{1, 5\}$.
    (a) What is A^c?
    (b) What is $A \cup B$?
    (c) What is $A \cap B$?
    (d) What is $A \cap C$?
    (e) Are A, B, C mutually exclusive? Are they collectively exhaustive?

55. A family has three children. Let A be the event that they have less than two girls and B be the event that they have exactly two girls.
    (a) List all of the basic outcomes in A.
    (b) List all of the basic outcomes in B.
    (c) List all of the basic outcomes in $A \cap B$
    (d) List all of the basic outcomes in $A \cup B$.
    (e) If male and female births are equally likely, what is the probability of A?

56. Let $B = A^c$. Are A and B mutually exclusive? Are they collectively exhaustive? Why?

57. State each of the three axioms of probability, aka the *Kolmogorov Axioms*.

58. Suppose we carry out a random experiment that consists of flipping a fair coin twice.

    (a) List all the basic outcomes in the sample space.

    Solution: $S = \{HH, HT, TT, TH\}$

    (b) Let A be the event that you get at least one head. List all the basic outcomes in A.

    Solution: $A = \{HH, HT, TH\}$

    (c) List all the basic outcomes in A^c.

    Solution: $A^c = \{TT\}$

    (d) What is the probability of A? What is the probability of A^c?

    Solution: $P(A) = 3/4 = 0.75$ and $P(A^c) = 1/4$

59. Calculate the following:

    (a) 5!

    Solution: 120

    (b) $\dfrac{10!}{98!}$

    Solution: 9900

    (c) $\binom{5}{3}$

    Solution: 10

60. (a) How many different ways can we choose a President and Secretary from a group of 4 people if the two offices must be held by different people?

    (b) How many different committees with two members can we form a group of 4 people, assuming that the order in which we choose people for the committee doesn't matter.

61. Suppose that I flip a fair coin 5 times.

(a) How many basic outcomes contain exactly two heads?

(b) How many basic outcomes contain exactly three tails?

(c) How many basic outcomes contain exactly one heads?

(d) How many basic outcomes contain exactly four tails?

62. Explain why $\binom{n}{r} = \binom{n}{n-r}$.

63. Suppose I deal two cards at random from a well-shuffled deck of 52 playing cards. What is the probability that I get a pair of aces?

Solution: You can either solve this assuming that order doesn't matter:

$$\frac{\binom{4}{2}}{\binom{52}{2}} = \frac{4!/(2! \times 2!)}{52!/(50! \times 2!)} = \frac{6}{(52 \times 51)/2} = 6/1326 = 1/221$$

or that it does:

$$\frac{P_2^4}{P_2^{52}} = \frac{4!/2!}{52!/50!} = \frac{(4 \times 3)}{(52 \times 51)} = 12/2652 = 1/221$$

In either case, the answer is the same: $1/221 \approx 0.005$

64. Suppose that I choose two numbers at random from the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. What is the probability that both are odd?

Solution: This solution assumes that order doesn't matter. You could also assume that it does matter and get the same answer. There are $\binom{9}{2} = 36$ equally likely ways to choose 2 items from a set of 9. Of these, there are $\binom{5}{2} = 10$ ways to choose 2 of the 5 odd numbers. Hence the probability is $10/36 = 5/18$.

Extensions

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

66. Treat regression to the mean in the extensions.

67. 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.

Solution: As we showed in class, the average deviation from the sample mean is zero regardless of the dataset. Taking the absolute value is similar to squaring the deviations: it makes sure that the positive ones don't cancel out the negative ones.

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

Solution: The variance is calculated from squared deviations. When x is far from zero, x^2 is much larger than $|x|$ so large deviations "count more" when calculating the variance. Thus, the variance will be more sensitive to outliers.

68. Let m be a constant and $x_1, \ldots, x_n$ be an observed dataset.

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

Solution:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - m)^2 &= \sum_{i=1}^{n}(x_i^2 - 2mx_i + m^2)\\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2mx_i + \sum_{i=1}^{n} m^2\\ &= \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2\end{aligned}$$

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

Solution: Solving this requires two observations. First, note that $\bar{x}$ is a *constant*, i.e. that it does not have an index of summation. Second, note that

$\sum_{i=1}^{n} x_i = n\bar{x}$. Hence, taking $m = \bar{x}$ in the formula from the preceding part,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

69. 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.

Solution:

$$\frac{1}{n}\sum_{i=1}^{n}(c + dx_i) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} c + d\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = c + d\bar{x}$$

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

Solution:

$$\frac{1}{n-1}\sum_{i=1}^{n}[(c + dx_i) - (c + d\bar{x})]^2 = \frac{1}{n-1}\sum_{i=1}^{n}[d(x_i - \bar{x})]^2 = d^2 s_x^2$$

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

Solution: The new standard deviation is $|d|s_x$, the positive square root of the variance.

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

Solution: They are unchanged as long as d is positive, but the sign will flip if d is negative:

$$\frac{(c + dx_i) - (c + d\bar{x})}{ds_x} = \frac{d(x_i - \bar{x})}{ds_x} = \frac{x_i - \bar{x}}{s_x}$$

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

71. Define the z-scores

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

Show that if we carry out a regression with z_i in place of y_i and w_i in place of x_i, the intercept a^* will be zero while the slope b^* will be r_{xy}, the correlation between x and y.

Solution: All we need to do is replace x_i with w_i and y_i with z_i in the formulas we already derived for the regression slope and intercept:

$$a = \bar{y} - b\bar{x}, \quad b = \frac{s_{xy}}{s_x^2}$$

And use the properties of z-scores from class. Let a^* be the intercept for the regression with z-scores, and b^* be the corresponding slope. We have:

$$a^* = \bar{z} - b^*\bar{w} = 0$$

since the mean of the z-scores is zero, as we showed in class. To find the slope, we need to covariance between the z-scores, and the variance of the z-scores for x:

$$b^* = \frac{s_{wz}}{s_w^2}$$

But since sample variance of z-scores is always one, $b^* = s_{wz}$. Now, by the definition of the sample covariance, the fact that the mean of z-scores is zero, and the definition of a z-score:

$$\begin{aligned} s_{wz} &= \frac{1}{n-1}\sum_{i=1}^{n}(w - \bar{w})(z - \bar{z}) = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \\ &= r_{xy} \end{aligned}$$

72. 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.

Solution:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

Solution:

$$a = \bar{y} - b\bar{x}$$

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

Solution:

$$r = \frac{s_{xy}}{s_x s_y}$$

(d) Show that

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

Solution:

$$\begin{aligned}
\hat{y} &= a + bx \\
\hat{y} &= (\bar{y} - b\bar{x}) + bx \\
\hat{y} - \bar{y} &= b(x - \bar{x}) \\
\hat{y} - \bar{y} &= \frac{s_{xy}}{s_x^2}(x - \bar{x}) \\
\hat{y} - \bar{y} &= \frac{s_{xy}}{s_x}\left(\frac{x - \bar{x}}{s_x}\right) \\
\frac{\hat{y} - \bar{y}}{s_y} &= \frac{s_{xy}}{s_x s_y}\left(\frac{x - \bar{x}}{s_x}\right) \\
\frac{\hat{y} - \bar{y}}{s_y} &= r\left(\frac{x - \bar{x}}{s_x}\right)
\end{aligned}$$

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

Solution: The formula shows that unless r is one or negative one, perfect positive or negative correlation, our best linear prediction of y based on knowledge given x is closer to the mean of the y-observations (relative to the standard deviation of the y-observations) than x is to mean of the x-observations (relative to the standard deviation of the x-observations). If x is very large, for example, we would predict that y will be large too, but not as large.

73. 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?

Solution: This is an example of regression to the mean. The students Lothario seeks out were both unprepared for the midterm *and* got unlucky: the correlation between exam scores is less than one. It is very unlikely that they will be unlucky twice in a row, so their performance on the second exam will almost certainly be higher. Our best guess of their second score is closer to the mean than their first score.