

PLSC 503: Problem Set 3 Solutions

Thad Dunning

Department of Political Science

Yale University

Spring 2010

Question 1: The average score on the GRE verbal examination between the years 2000 and 2003, among students planning to apply to political science and international relations graduate programs, was 520; the standard deviation was 105. The histogram for the exam scores should follow the normal curve reasonably well. About what proportion of exam-takers scored over 700 on the test? Explain your answer. It may help to consult Chapter 5 of FPP.

Solution: First, convert 700 to standard units: $700 - 520/105 \doteq 1.714$. Since the histogram follows the normal curve, we use a normal table (such as that on page A-104 of Freedman, Pisani, and Purves 2007) to find the area under the standard normal curve that is greater than $|1.714|$. This is around 91.25 ($z = 1.714$ lies between $z = 1.70$ and $z = 1.75$), so the desired area is $100 - 91.25 = 8.75$. We have to divide by 2 to get the area above 1.714 standard units; thus, around $8.75/2 = 4.375\%$ of students scored over 1.714 standard units on the verbal GRE, that is, above 700 in the original units.

Question 2: One ticket is drawn at random from a box containing six tickets: $\{1,2,3,4,5,6\}$. Then a second ticket is drawn, without replacement of the first ticket.

(a) What is the probability that the second ticket is 3?

Solution: The probability is $1/6$. (Read FPP pp. 226-7 if the logic is unfamiliar).

(b) What is the probability that the second ticket is 3, given that the first ticket is 2?

Solution: The probability is $1/5$: now there are five tickets in the box, one of them a 3.

(c) Is the unconditional probability the same as the conditional probability?

Solution: No; the unconditional probability is $1/6$, and the conditional probability (of drawing a 3, given that the first ticket is 2) is $1/5$.

(d) Is the value of the second ticket dependent or independent of the value of the first ticket?

Solution: The values are dependent. (The unconditional and conditional probabilities are different).

(e) What is the probability that the first ticket is 2 and the second ticket is 3?

Solution: The probability that the first ticket is 2 is $1/6$. The probability that the second ticket is 3, given that the first ticket is 2, is $1/5$. So the probability that the first ticket is 2 and the second ticket is 3 is $(1/6)(1/5)=1/30$.

(f) What is the probability that the first or the second ticket is 3?

Solution: These are mutually exclusive events, because we are drawing without replacement. Thus the probability is $(1/6) + (1/6) = 1/3$.

Question 3: Two tickets are drawn at random with replacement from a box containing two tickets: $\{0,1\}$.

(a) Construct a box model for this process, in which you draw once at random from a single box. (Hint: each ticket in the box should have two values).

Solution: There are four tickets in the box. Each ticket has two values; the tickets are labelled $\{0,0\}$, $\{0,1\}$, $\{1,0\}$, and $\{1,1\}$. Remember, you need both $\{0,1\}$ and $\{1,0\}$, because in the original experiment there are two ways of drawing one 1 and one 0: the 1 can come on the first draw or the second draw.

(b) What is the probability that at least one of the two tickets is 1?

Solution: Three out of four tickets in the box in (a) have at least one 1. So the probability is $3/4$.

(c) Now suppose this experiment of drawing twice at random with replacement from the original box is repeated 50 times. What is the expected value of the sum of the tickets?

Solution: The expected value of a single draw is the mean of the box: 0.5. We are drawing 100 tickets at random with replacement. So the expected value of the sum is $100 \times 0.5 = 50$.

(d) What is the standard error of the sum?

Solution: The usual formula for the standard error of the sum of n independent draws from a box applies: $\sqrt{n} \times SD$, where SD is the standard deviation of the box. Here, $n = 100$, and the SD of the box is $\sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}$. (See FPP p. 298 for this useful short-cut method of finding the standard deviation; here, the “big” number is 1 and the “small” number is 0). So the standard error of the

sum is $\sqrt{100} \times \frac{1}{2} = 5$.

(e) Use the normal approximation to find the probability that the sum of the tickets is less than 60.

Solution: The expected value of the sum is 50, and the standard error—the standard deviation of the sampling distribution of the sum—is 5. So 60 is two standard errors above the mean of the sampling distribution. According to the normal table on p. A-104 of FPP, about 95.45% of the area under the standard normal curve lies between -2 and 2 standard units from the mean. So $100 - 95.45 = 4.55\%$ lies above $z = 2$ or below $z = -2$. Thus, $4.55/2 = 2.275\%$ lies below $z = -2$. Thus, $95.45 + 2.275 = 97.725\%$ lies below 2 standard errors above the mean. The probability that the sum of tickets is less than 60 is therefore about 0.977.

(f) Why does the normal approximation apply?

Solution: The central limit theorem.

Question 4: A sample of citizens is drawn from a large population for a public opinion survey. Citizens are asked whether they support Candidate A or Candidate B. A news organization reports the “margin of error” of the survey. A little background: the margin of error is typically reported as plus or minus 2 standard errors, assuming that 50% of citizens support Candidate A and 50% support Candidate B.

(a) Construct a box model for the sampling process, assuming that 50% of citizens support Candidate A and 50% support Candidate B.

Solution: The box has two tickets, {0} and {1}. The {0} might be for Candidate A and the {1} for Candidate B.

(b) What is the expected percentage of survey respondents who support Candidate A?

Solution: 50%—that is, the mean of the box (0.5) multiplied by 100 to convert to percentages.

(c) What is the margin of error for this percentage, for a sample of size 100? What about for samples of size 200, 400, and 800?

Solution: For a sample of 100, we imagine drawing at random with replacement from the box

100 times. The standard error for the average of the sample is $\frac{SD}{\sqrt{n}}$, where SD is the standard deviation of the box, and the standard error for the percentage is $\frac{SD}{\sqrt{n}} \times 100$. Here, the SD of the box is $\sqrt{0.5 \times 0.5} = 0.5$. Thus, the standard error is $\frac{0.5}{\sqrt{100}} \times 100 = 5$. Since the reported margin of error is plus or minus 2 standard errors, it is plus or minus 10.

For sample sizes of 200, 400, and 800, the standard error is 3.54, 2.5, and 1.77, respectively. So the respective margins of error are plus or minus 7.08, plus or minus 5, and plus or minus 3.54.

(d) By about how much does doubling the sample size from 100 to 200 cut down the margin of error? How about from 400 to 800? Comment.

Solution: Doubling the sample size from 100 to 200 cuts down the margin of error from 10 to 7.08, or about $\frac{10-7.08}{10} = -29.2\%$. Doubling the sample size from 400 to 800 cuts down the margin of error from 5 to 3.54, or about $\frac{5-3.54}{5} = -29.2\%$. That is, doubling the sample size cuts down the margin of error not by a factor of 2 but instead by a factor of $\sqrt{2}$.

(e) What is the coverage of the confidence interval implied by a margin of error of plus or minus 2 SEs? Give an interpretation of this confidence interval. (That is, say what it means).

Solution: This is (roughly) a 95% confidence interval. The interpretation is this: if we drew many samples of size n (where $n = 100, 200, 400, 800, \dots$) and constructed the confidence interval—that is, plus or minus 2 SEs—the interval would cover the true percentage (that is, 50%) about 95% of the time.

Question 5: Consider the following regression equation:

$$Y = X\beta + \epsilon. \tag{1}$$

Here, Y is a $n \times 1$ column vector consisting of $(Y_1 \ Y_2 \ \dots \ Y_n)'$, and X is a $n \times 2$ design matrix with rank 2, where the first column of X is all 1's and the second column is $(X_{11} \ X_{12} \ \dots \ X_{1n})'$. The column vector $\epsilon_{n \times 1} = (\epsilon_1 \ \epsilon_2 \ \dots \ \epsilon_n)'$ is a vector of random errors, with $\epsilon \perp X$, $E(\epsilon) = 0$, and $\text{var}(\epsilon) = \sigma^2$. Here, $\beta = (\beta_0 \ \beta_1)'$ is a 2×1 parameter vector. Let $\hat{\beta} = (X'X)^{-1}X'Y$ be the OLS

estimator for β .

(a) What is another word for β_0 ? How about β_1 ?

Solution: Here, β_0 is the intercept, and β_1 is the slope. This is bivariate regression.

(b) How many equations are represented by (1) above?

Solution: There are n equations, one for each unit/subject.

(c) Show that the standard error of $\hat{\beta}_1$, the (2, 1) element of $\hat{\beta}$, is

$$SE_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{n} \sqrt{\text{var}(X_1)}}. \quad (2)$$

Solution: Recall that

$$\text{cov}(\hat{\beta}|X) = \sigma^2(X'X)^{-1} \quad (3)$$

The variance of $\hat{\beta}_2$ is given by the (2, 2) element of this matrix, and the standard error is the square root of the variance. Now, from problem set 2, we have that

$$(X'X)^{-1} = \frac{1}{n \sum_{i=1}^n (X_{1i})^2 - (\sum_{i=1}^n X_{1i})(\sum_{i=1}^n X_{1i})} \begin{pmatrix} \sum_{i=1}^n (X_{1i})^2 & -\sum_{i=1}^n X_{1i} \\ -\sum_{i=1}^n X_{1i} & n \end{pmatrix} \quad (4)$$

Thus, the (2, 2) element of $\text{cov}(\hat{\beta}|X)$ is

$$\begin{aligned} \frac{\sigma^2 \cdot n}{n \sum_{i=1}^n (X_{1i})^2 - (\sum_{i=1}^n X_{1i})(\sum_{i=1}^n X_{1i})} &= \frac{\frac{1}{n}\sigma^2}{\frac{1}{n} \sum_{i=1}^n (X_{1i})^2 - \frac{1}{n^2}(\sum_{i=1}^n X_{1i})(\sum_{i=1}^n X_{1i})} \\ &= \frac{\frac{1}{n}\sigma^2}{(\overline{X_1})^2 - \bar{X}_1^2} \\ &= \frac{\sigma^2}{n \text{var}(X_1)}, \end{aligned} \quad (5)$$

where in the first line of (5), we divided through by $\frac{1}{n^2}$; in the second line, we evaluated the terms in the denominator using the definition of the average; and in the third line, we used the alternate

definition of variance. Thus, taking square roots, we have

$$SE_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{n} \sqrt{\text{var}(X_1)}}.$$

Question 6: A data set has $n = 200$. A typical row of the $n \times 3$ design matrix is $[1 \ x_{1i} \ x_{2i}]$. Here, x_1 is a binary variable equal to 1 for men and 0 for women, and x_2 is a binary variable equal to 0 for men and 1 for women. Will the design matrix have full rank? Now suppose a typical row of the $n \times 2$ design matrix is $[x_1 \ x_2]$. Will this design matrix have full rank? Explain your answers.

Solution: The first design matrix will not have full rank (and thus $X'X$ won't be invertible), because the second and third columns are linear combinations of the first column. The second design matrix has full rank. Lesson: if you include the full set of dummy (binary) variables in the design matrix, don't include the constant as well.

1 Computer exercises

A .do file for the computer exercises is posted online, courtesy of Mario Chacón.