

# PLSC 503: Problem Set 3

Thad Dunning

Department of Political Science

Yale University

Spring 2010

Due February 2, 2010

# 1 Theoretical/Conceptual Exercises

Before completing this problem set, it is recommended that you work all of the exercises in Freedman (2009, Chapter 4), including the discussion questions.

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.

(Data are from

[http : //www.wepapers.com/Papers/53972/GRE\\_Test\\_Percentage\\_Distribution\\_Of\\_Scores;](http://www.wepapers.com/Papers/53972/GRE_Test_Percentage_Distribution_Of_Scores;)  
see graduate major categories 1901 and 1902).

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?
  - (b) What is the probability that the second ticket is 3, given that the first ticket is 2?
  - (c) Is the unconditional probability the same as the conditional probability?
  - (d) Is the value of the second ticket dependent or independent of the value of the first ticket?
  - (e) What is the probability that the first ticket is 2 and the second ticket is 3?
  - (f) What is the probability that the first or the second ticket is 3?

Explain your answers. If this material is unfamiliar, read Chapters 13 and 14 of FPP.

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).
- (b) What is the probability that at least one of the two tickets is 1?
- (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?
- (d) What is the standard error of the sum?
- (e) Use the normal approximation to find the probability that the sum of the tickets is less than 60.
- (f) Why does the normal approximation apply?

Explain your answers. If this material is unfamiliar, read Chapters 16 and 17 of FPP.

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

- (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).

Explain your answers. If this material is unfamiliar, read Chapters 20 and 21 of FPP.

5. Consider the following regression equation:

$$Y = X\beta + \epsilon. \quad (1)$$

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  $2 \times 1$  parameter vector, Let  $\hat{\beta} = (X'X)^{-1}X'Y$  be the OLS estimator for  $\beta$ .

- (a) What is another word for  $\beta_0$ ? How about  $\beta_1$ ?
- (b) How many equations are represented by (1) above?
- (c) Show that the standard error of  $\hat{\beta}_1$ , the (2, 1) element of  $\hat{\beta}$ , is

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

(Hint: refer to problems 7 and 8 on problem set 2).

6. A data set has  $n = 200$ . A typical row of the  $n \times 3$  design matrix is  $[1 \ x_1 \ x_2]$ . 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.

## 2 Computer exercises

1. For this question, you may use Stata, Excel, Matlab, or another program.
  - (a) Simulate observations on 32 IID normal variables  $X_i$  with mean  $\mu = 15$  and variance  $\sigma^2 = 100$ .
  - (b) Calculate the sample mean  $\bar{X}$  and the sample SD  $\hat{\sigma}$  of the data.
  - (c) Repeat 1 and 2, 1000 times.
  - (d) Plot the histogram of the 1000  $\bar{X}$ 's.
  - (e) Plot a histogram of the 1000  $\hat{\sigma}$ 's.
  - (f) Plot a scatter diagram of the 1000 pairs  $(\bar{X}, \hat{\sigma})$ .
  - (g) Calculate the SD of the 1000  $\bar{X}$ 's. How does this compare to  $\frac{\sigma}{\sqrt{32}}$ ? Comment.
2. Open again the Yule data set you used in the previous problem set. Yule assumed

$$\Delta\text{Paup}_i = a + b\Delta\text{Out}_i + c\Delta\text{Old}_i + d\Delta\text{Pop}_i + \epsilon_i \quad (3)$$

for metropolitan unions  $i$ . Here, the errors  $\epsilon_i$  are IID, with mean 0 and variance  $\sigma^2$ .

- (a) In the last problem set, you used matrix commands to fit  $(X'X)^{-1}X'Y$ . Use this fit to estimate  $a$ ,  $b$ ,  $c$ , and  $d$ . Then use matrix commands to estimate  $\sigma^2$ .
- (b) Compute the SEs. (Do not use Stata's regress command or similar command in another package; use the matrix commands).
- (c) Are these SEs exact, or approximate?
- (d) Conduct a test of the null hypothesis that  $b = 0$ .
- (e) Plot the residuals against the fitted values. (This is often a useful diagnostic: if you see a pattern, something may be wrong with the model).

Now, set the parameters in Yule's equation above as follows:  $a = -40$ ,  $b = 0$ ,  $c = 0.2$ ,  $d = -0.3$ ,  $\sigma = 15$ . Fix the design matrix  $X$  as above. That is, a typical row of the design matrix is  $[1 \ \Delta\text{Out}_i \ \Delta\text{Old}_i \ \Delta\text{Pop}_i]$ , where you take the values of  $\Delta\text{Out}_i$ ,  $\Delta\text{Old}_i$ , and  $\Delta\text{Pop}_i$  from each row of the file `yule.dta` (or `yule.csv`).

- (f) Generate 32  $N(0, \sigma^2)$  errors and plug them into the equation

$$\Delta\text{Paup}_i = -40 + 0 \times \Delta\text{Out}_i + 0.2 \times \Delta\text{Old}_i - 0.3 \times \Delta\text{Pop}_i + \epsilon_i \quad (4)$$

to get simulated values of  $\Delta\text{Paup}_i$ , with  $i = 1, \dots, 32$ .

- (g) Regress the simulated  $\Delta\text{Paup}$  on an intercept,  $\Delta\text{Out}$ ,  $\Delta\text{Old}$ , and  $\Delta\text{Pop}$ . (You can use the conventional commands such as Stata's "regress" command to do this). Calculate  $\hat{b}$ ,  $\widehat{\text{SE}}$ , and  $t$ .
- (h) Repeat (f) and (g), 1000 times.
- (i) Plot a histogram for the 1000  $\hat{b}$ 's, a scatter diagram for the 1000 pairs  $(\hat{b}, \hat{\sigma})$ , and a histogram for the 1000  $t$ 's.
- (j) What is the theoretical distribution of  $\hat{b}$ ? Of  $\hat{\sigma}^2$ ? Of  $t$ ? How close is the theoretical distribution of  $t$  to normal?
- (k) Calculate the mean and SD of the 1000  $\hat{b}$ 's. How does the mean compare to the true  $b$ ? ("True" in the simulation). How does the SD compare to the true SE for  $\hat{b}$ ?