Problem Set 1

EC 303: Empirical Economic Analysis

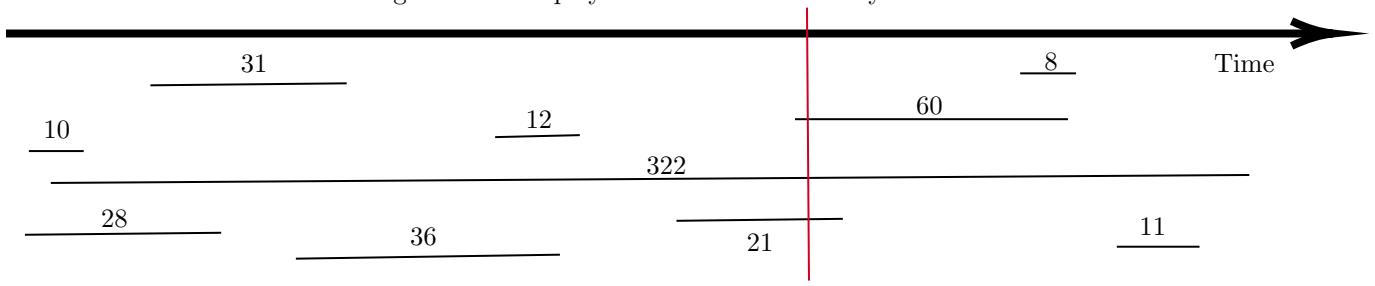
Due September 25, 2019 at the start of class.

1 Theoretical Problems

Problem 1.1: Summarizing Data & Length-Biased Sampling. In class, we discussed the main measures used to summarize data—however, even the way you **set up** a data set can affect these summary statistics. One of the ways this can be exploited is *length-biased sampling*, when the population of interest evolves over time, and that evolution is ignored.

To illustrate this issue, consider the classic economic problem of unemployment (see Clark & Summers, 1982). Suppose that there are 10 people in a community who are ever unemployed in a given year. This figure shows each person's unemployment spell, with each spell length listed (measured in days).

Figure 1: Unemployment in the Community



a. Compute the following measures of central tendency for the population of unemployed in this community: mean, median, mode, range. Comment on the difference between the mean and the median, and on the mode—what do these tell you?

b. Suppose that we randomly sample everyone who is unemployed in the community at the time denoted in the figure by the red line. If we follow each of these 3 individuals until they complete their spells and compute the average, what do we find? How does this compare to the overall average?

c. Now consider repeating this sampling procedure many times. That is, consider moving the red line horizontally to other random sampling times. (note you do not actually need to calculate new averages). Do you think that these averages tend to overstate or understate the true average spell length?

d. What do you think is driving this result? What recommendations do you have to fix the problem?

e. Use this framework to evaluate the following statement on recidivism: "About half of the prisoners released in any given year in the United States will end up back in prison within five years. Yet the proportion of prisoners ever released who will ever end up back in prison, over their whole lifetime, is just one third. How can this be?" (see *This Idea is Brilliant*, Edge 2017)

Problem 1.2: Math Review, Part 1. These problems focus on key concepts covered in our math review.

a. Solve $1 - e^{x/2} = 0.75$ for x.

b. Evaluate $\sum_{k=3}^{4} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$ when $n = 4$ and $p = 0.25$.

c. Simplify $\ln\left(\Pi_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$, where π, μ, σ and n are all constants.

d. Find $\frac{\partial f}{\partial x}$ when $f(x,y) = (1+\sin(y)+\ln(xy))^2$.

e. Choose a price p to maximize the following profit function:

$$\Pi = pq - 100q,$$

when the demand curve is such that $q = a - bp$ (a and b are parameters). Show that your solution is a maximum, and not a minimum. What is the profit-maximizing quantity?

f. Evaluate $\int_0^1 \int_0^{10} xydxdy$.

Problem 1.3: Math Review, Part 2. Answer each question as true (T) or false (F). If true, provide a proof; if false, provide a simple counterexample (*Hint:* for counterexamples, keep $n = 2$ for simplicity).

a. $\sum_{i=1}^{n} x_i y_i = (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$. Can a summation distribute through a product?

b. $\Pi_{i=1}^{n} a x_i = a\Pi_{i=1}^{n} x_i$ for some $a \in \mathbb{R}$. Can a coefficient come out of a product?

c. $\Pi_{i=1}^{n}(x_i + y_i) = (\Pi_{i=1}^{n} x_i) + (\Pi_{i=1}^{n} y_i)$. Can a product be distributed through a sum?

d. If $\overline{x}$ is the sample average, then $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, so that numbers are on average no higher or lower than the sample mean.

Problem 1.4: Proofs of some Probability Corollaries.** Prove each of the following consequences of the probability axioms, where E, F represent events in a sample space $\mathcal{S}$:

a. $P(E) = 1 - P(\overline{E})$

b. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$. *Hint:* Even if E and F aren't disjoint, can you write their union as a union of disjoint sets? A Venn diagram may help.

Problem 1.5: Basic probability problems.** These two problems are from the textbook:

a. Problem 2.24.

b. Problem 2.43, first two questions only.

Problem 1.6: Conditional Probabilities I.** Complete exercise 2.45 from the textbook.

Problem 1.7: Conditional Probabilities II. Prove or present a counter-example: if $P(C) > 0$ and $P(A) > P(B)$, then $P(A|C) > P(B|C)$.

Problem 1.8: Independence I.** Prove that if A and B are independent, then A is independent of $\overline{B}$.

Problem 1.9: Independence II. Two quick problems:

a. Are the events A and C in problem 1.6 above independent events? Why or why not?

b. Complete Exercise 2.83 in the textbook (**Note:** The answer in the back is slightly wrong).

2 Stata Exercises

Note: Each of these data sets comes from Wooldridge's econometrics textbook. To use them, first type ssc install bcuse to install a loading package. Then, type bcuse dataname to load the desired package.

Problem 2.1: Summarizing Data. For this problem, use the fertil1 data set, which is a data set detailing women and their fertility choices.

a. What does each observation represent?

b. Provide the following summary stats for the variable educ: min, first quartile, median, mean, standard deviation, third quartile, max, interquartile range, and range. Choose 2 or more or these measures and describe how their relationship impacts your interpretation of the data.

c. Make a frequency table for the variable kids. Now turn this into a relative frequency table. What stands out to you?

d. Find the 15th and 85th percentiles for meduc and feduc. How do they compare?

e. Make a histogram of age for all those with at least one child (*Hint:* Use the discrete option for the histogram to look better). If you're feeling ambitious, use the tools in the next problem to make this histogram more visually appealing. Is the shape of the histogram consistent with the interpretation that younger women are more likely to be parents than older women?

Problem 2.2: Data visualization. This problem will help you get under the hood in Stata's graph editor, to help you make data viz that looks good. We will use the hprice2 data set, which contains data on housing prices and neighborhood characteristics (each observation is a single house).

a. Create a scatter plot showing the relationship between price and rooms. To avoid outliers, ignore observations where price is equal to 50000.

b. **Labels and Titles.** Good data viz uses good labels and titles. Add labels using options like xlabel and a descriptive title using the title option. If we assume that price is measured in hundreds of dollars (e.g., so that 5000000 = \$5M), change your y-axis to show the price in millions of dollars.

c. **Colors.** Now lets add change the color scheme. No one likes the blue background Stata defaults to, so let's change that to white using the graphregion(color(white)) option. Also, change the color of the dots using a color option—make sure your chosen color is visible and not too garish. Keep the labels from before.

d. **Trends.** Oftentimes, a scatter plot is made better by showing a trend which fits the dots—this trend can be linear, quadratic, or any other kind of polynomial (although it's usually linear). Add a linear trend to the graph using the lfit command (note that this is a command, not an option—if you haven't been using graph twoway up until this point, you will need to here). Keep all of the options you've included so far, and make sure the line is a different color than your scatter plot dots.

e. **Saving.** The best way to save your graph is to add the graph export command. You **always** want to save your graph as a .pdf, as this allows rescaling without pixelation. Practice saving your graph. (Note that you don't have to respond to this part.)

f. Use the tools mentioned above and this data to make another interesting graph. You could summarize a variable, look at the relationship between two variables, or anything else you may think is interesting. *Do not make a simple histogram*, since we've already done that in this assignment.

Problem 2.3: Intro to Bayesian updating. This problem shows you how to simulate data in Stata, and practice Bayes' updating at the same time! For this problem, we are interested in the probability that a person has cancer given the results of their medical tests.

a. Begin with a cleared Stata window (use clear if you already have another data set open). Create a sample of 100000 patients using the command set obs 1000. We will be making random draws, so to make sure we all get the same results, use the command set seed 12345. This ensures our "random draws" are the same.

b. To each patient, assign a true cancer diagnosis ($c \in \{0, 1\}$). To do this first draw a random number for each patient using runiform(). Then, based on the random draw, assign $c = 1$ if the random draw is small enough and $c = 0$ otherwise. Assume that $c = 1$ with probability 0.00148.

c. Now we will assign the test results. For each patient, draw a new random variable using runiform(). If that patient has $c = 0$, assign the test to be negative ($t = 0$) with probability 0.99, and positive ($t = 1$) otherwise. If the patient has $c = 1$, the test should be negative with probability 0.07 and positive otherwise.

d. Based on your data, what is the (unconditional) probability that a patient both has cancer and a positive test? What is the probability that they do not have cancer but did test positive?

e. Given your answers to (d), what does Bayes' formula imply is the updated probability that a person has cancer given that they tested positively?

3 References

- Clark, Kim B. & Summers, Lawrence H. (1982). The dynamics of youth unemployment. In *The youth labor market problem: Its nature, causes, and consequences* (pp. 199-234). University of Chicago Press.
- Wrigley-Field, E. 2017. Length-Biased Sampling. In Brockman, J (ed.), *This Idea is Brilliant.* New York, NY: HarperCollins Publishers.