ECON 4130
2005 H
HG

Written Paper II 2005H (compulsary)

(To be submitted (in English or Norwegian), Monday 24 October, at Ekspedisjonskontoret 12th floor ES.)

Exercise 1 (Simulation of geometrically distributed data)

a. Suppose that $U \sim \text{uniform}(0,1)$, i.e., is uniformly distributed over the interval (0, 1). Let λ be a constant > 0. Show that $Y = -\frac{1}{\lambda}\ln(U)$ is exponentially distributed with parameter λ ($Y \sim \text{exponential}(\lambda)$)

b. Using **a.** there is a simple way to generate (simulate) observations from a geometric distribution. Let $X \sim \text{geometric}(p)$, with *pmf*: $f(x) = p(1-p)^{x-1}$, $x = 1,2,3,\ldots$ We want to simulate *n* independent observations of *X* for a given (known) *p*. As a first step, assume that $Y \sim \text{exponential}(\lambda)$. Show that

$$P(x-1 < Y \le x) = p(1-p)^{x-1} = P(X = x)$$

where λ is chosen as $\lambda = -\ln(1-p)$, and $x = 1,2,3,\ldots$

c. Introduce "[ ]" as a notation for truncation upwards, i.e., [*a*] means the smallest integer larger or equal to *a*. For example, [3,2] = [3,01] = [3,9] = 4, while [3] = 3. In STATA the function, `ceil`(x), does just that. Let $U_i \sim iid$ and $\text{uniform}(0,1)$, for $i = 1,2,\ldots,n$ Show that

$$X_i = \left\lceil \frac{\ln(U_i)}{\ln(1-p)} \right\rceil \sim \text{geometric}(p), \quad i = 1,2,\ldots,n$$

(**Hint:** Note that $[a] = x \quad \Leftrightarrow \quad x-1 < a \le x$ when *x* is an integer.)

Exercise 2

a. One game on a slot machine consists of initially paying a fixed amount, then push a button to make some figures spin before they stop in a random pattern. Certain predetermined patterns lead to a win, small or big, while the rest of the possible patterns lead to loss. Let *p* denote the probability of a win in a single game for a given slot machine (this *p* may vary between machines and is usually unknown for the player). Let *X* be the number of games (trials) played until the first win on this particular slot machine. We assume that $X \sim \text{geometric}(p)$. Why is this a reasonable model here?

b. We want to estimate p based on *n* independent observations of *X*, i.e., $X_1, X_2, \ldots, X_n$. Show that both the moment method estimator (mme) and the maximum likelihood estimator (mle) are equal to

$$\hat{p} = \frac{1}{\overline{X}} \quad \text{where} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Calculate the estimate for the data: $2, 13, 1, 9, 1$ (*n* = 5).

c. It is to be expected that the uncertainty of the estimate can be large since *n* is small. We will measure the uncertainty by calculating a 90% confidence interval (CI) for *p*. To find an exact 90% CI is not so easy here so we will utilize the asymptotic theory for mle, according to which

$$\sqrt{n}(\hat{p} - p) \xrightarrow[n\to\infty]{D} N(0, b(p))$$

where $b(p)$ is a certain continuous function of *p*. Show that

$$b(p) = p^2(1-p)$$

Use this to derive, along the lines of example 5 in the lecture notes to Rice chapter 5, an approximate 90% CI for *p*, and calculate the CI based on the data.

d. **Parametric Bootstrap:** A weakness with asymptotic methods is that they require large samples to apply. Given a finite sample we cannot always be sure that an interpretation of data based on asymptotic approximations is justified. Only in exceptional cases there exist analytical ways to evaluate the exact statistical properties of an estimator based on a finite small sample. Therefore the most common way to obtain insight into the small sample properties of estimators is by simulation techniques (of which the “Bootstrap” family has gained a lot of importance in the later years).

In the present case we have a quite small sample size (n = 5) which may give reason to believe that the CI in **c.** is not quite justified. To obtain evidence about this issue we arrange a parametric bootstrap experiment. Based on a simulated sample of estimates of p, we can calculate a so called bootstrap 90% CI for p and compare with the asymptotic CI. If there is no great difference this can be taken as evidence that the asymptotic methods worked well. If the difference is substantial, it is reasonable to discard the asymptotic CI's and report the bootstrap CI for p instead.

Read the text between example C and D in Rice section 8.5.3, in addition to example E. Use the method described there to calculate a 90% bootstrap CI for p based on a bootstrap sample of size B = 1000. You then need to generate 1000 data sets of size 5 drawn from a geometric distribution with success probability $\hat{p}$. For each of the simulated data sets, calculate the mle estimates of p . Thus you have obtained 1000 simulated observations of the mle's $\hat{p}$, that we may call $p_i^*,\ \ i = 1, 2, \ldots, 1000$. (Note that during the simulation, the true p has been approximated by the estimate $\hat{p}$.)

To draw geometric observations is not direct in STATA and you need a small program to do that which you can find in the appendix. Running that do-file produces a STATA data set containing the 1000 simulated p_i^* .

Compare the bootstrap CI with the asymptotic one in **c.** Also calculate the mean of the simulated p_i^* 's. If $\hat{p}$ is unbiased this mean should be close to the observed value of $\hat{p}$. So, if the difference is larger, this is evidence of a bias in $\hat{p}$. The difference itself can be taken as an estimate of the bias (i.e., $\mathrm{E}(\hat{p}) - p \ \approx\ \overline{p}^* - \hat{p}$, where $\overline{p}^*$ is the mean of the p_i^* 's). Calculate the estimate of the bias.

e. Let $\hat{p}_o$ denote the observed value of $\hat{p}$. According to the asymptotic theory $\hat{p}$ is approximately distributed as $N(p, b(p)/5)$, and the bootstrap sample consequently should be approximately a sample drawn from $N(\hat{p}_o, b(\hat{p}_o)/5)$. To study the quality of this approximation, make a histogram of the p_i^* 's with the density of $N(\hat{p}_o, b(\hat{p}_o)/5)$ drawn in the same graph. Comment on the graph.

> [**Hint:** To make two graphs on the top of each other in STATA you can achieve by the "Twoway"-command, or by means of the "Overlaid twoway graphs" on the graph menu. In addition you need to calculate the values of the relevant normal density for all the values of p_i^* . This can be done by the function, *normden(x,a,b)*, described by giving the command, *help normden*. Then use the line-graph for the density (don't forget to mark the *sort* option in the menu!)]

Appendix

Suppose we want 1000 observations of $\hat{p}$ based on 1000 samples of size 5 drawn from the geometric($\hat{p}_o$) distribution (I have used the value $\hat{p}_o = 0{,}192$ in the program, so change this if you have another value). The following small do-file (can be written directly into STATA by the do-editor or read into STATA from an ASCII-file from the do-editor), produces a STATA data file containing the 1000 observations. I have called the data file, “geomdat”, here but you can choose your own name of course. The file is stored under the name, geomdat.dta. For calculation of the p_i^*’s this file must be read into STATA by the use-command (e.g. use geomdat).

```
capture program drop geosim
program define geosim
      tempname sim
      postfile `sim’ p using geomdat
      quietly  {
            local i=1
            while `i’ <= 1000  {
                  drop _all
                  set obs 5
                  gen x=ceil(ln(uniform())/ln(1-.192))
                  gen z=sum(x)
                  post `sim’ (_N/z[_N])
                  local i=`i’+1
                        }
                  }
      postclose `sim’
      clear
end
```

Notes: The first line makes it possible to edit the program geosim and then run it again without STATA protesting. The single quotation mark ` I find on the top of my backslash (\) key. The closing single quotation mark, ’, I find under * on the *-key.

If z is a variable with numbers $z_1, z_2, \ldots, z_k$, using the function *sum(z)*, creates a new column with the cumulative sums, $z_1,\ z_1 + z_2,\ z_1 + z_2 + z_3, \ldots$ The total sum of the z_i’s is found as the last element of this column. Note that $\_N$ gives the number of observations in the current dataset, so $z[\_N]$ refers to the total sum.

1. Read the lines into the do-editor.
2. Run the do-file by pushing the run-key in the do editor (this only reads the program *geosim* into STATA but does not execute it).
3. Then run the program by writing *geosim* in the command window.
4. Then read into STATA the simulated data by the command, *use geomdat*.

When you start your STATA session your working directory will be m:\. (You can confirm this by giving the command, *cd*.) This means that all your STATA files will be together with all your other files on m:. It may be a good idea, before you start, to make a subdirectory named, for example, STATA1, to contain all your STATA-files. Then , the first thing you do after opening STATA, change the working directory by the command: *cd* STATA1. Then, when you save data files, do-files etc. from STATA, they end up in m:\STATA1.