PLSC 503: "Multivariate Analysis for Political Research"

Exercise Zero
February 3, 2021

Introduction

The purposes of this (optional) exercise are several: to familiarize you with a couple statistical distributions, to develop your skills at generating, manipulating, and describing (graphically and in words) random variables using statistical software, and (most important) to begin to develop your skills at using that software to conduct simulation-based research.

Follow the instructions below. Text that appears in `typewriter` font denotes specific software commands. If no command(s) are specified, you can complete the fast however you would like. Also, be aware that many of the things I'm asking you to do here are for pedagogical purposes only, and are not necessarily the most efficient way to accomplish the tasks in question.

There are no ("real") data for this exercise. In writing up your homework, be sure to include all the code necessary to replicate your work *exactly*.

Hints

1. R has a somewhat steep learning curve. Some good resources (of various sorts) are:

   - Quick-R,
   - The R Cookbook,
   - William King's R tutorials for beginners,
   - Various R4Stats posts (e.g., this amazingly handily reference page),
   - R-Bloggers,
   - R Cheatsheets,
   - The R page on StackOverflow,
   - The Penn State R User Group,
   - And, of course, the mothership.

2. It's good practice to `set.seed()` before doing any simulations; this will ensure that every simulation gives you exactly the same results.

3. When – not if – you get errors, a simple but useful thing to do is to put the error message in quotes (e.g., "`Error in match.fun(FUN) : object 'U' not found`"), along with the letter "R," into our old friend Google. It is almost impossible to be the first person ever to have encountered a given error message. StackOverflow in particular is often a good place to search for debugging tips.

Exercise

1. **Generate Some Fake Data**

   (a) Generate 1000 independent random draws u_j from a uniform(0,1) distribution. Briefly describe (graphically and in words) the resulting "data," and compare them to the theoretical distribution of values from that distribution.

   (b) Repeat (a) 999 more times, saving each set of draws. You should wind up with 1,000,000 random $U(0,1)$ draws u_{ij}, organized into 1000 objects/variables (here, indexed by i), each of length 1000 (indexed by j).

   (c) For each of the 1000 U_i objects, create a new object V_i by adding the integer corresponding to the order of the observation to the value of u_{ij}. That is, to the value of the first observation in each data frame, add 1, to the second, add 2, etc., finishing by adding 1000 to the last listed observation in each data frame. *Be sure to retain the uniform variates as well.*

   (d) Generate an object containing the sums of the values of the observations in each of the 1000 Vs (that is, one where the ith entry is $A_i = \sum_{j=1}^{1000} V_{ij}$). Plot and describe the distribution of the resulting variable A, and explain why you see what you see.

   (e) Next, create a second object containing the sums of the jth observations across each of the 1000 Vs. That is, create an object of length 1000 where the jth entry is $S_j = \sum_{i=1}^{1000} V_{ij}$. Describe (graphically and in words) the distribution of the resulting variable S, and explain why.

2. **Gumbel Variates**

   A *Gumbel* distribution (here denoted $\mathcal{G}(\alpha, \beta)$) is a two-parameter ($\alpha \in \mathcal{R}, \beta > 0$) distribution with PDF $f(x) = \frac{1}{\beta} \exp\left[\frac{-(x-\alpha)}{\beta}\right] \times \exp\left\{-\exp\left[\frac{-(x-\alpha)}{\beta}\right]\right\}$ and CDF $F(x) = 1 - \exp\left\{-\exp\left[\frac{-(x-\alpha)}{\beta}\right]\right\}$. It is related to the standard uniform distribution U by $X = \alpha - \beta \log[-\log(U)]$.

   (a) Transform your 1000 bundles of 1000 random $U(0,1)$ draws into 1000 bundles of 1000 draws G_{ij} from a Gumbel(1,2) distribution.

   (b) Plot the empirical variances of G, and discuss how they compare to the theoretical variance of a $\mathcal{G}(1,2)$ distribution.

   (c) Plot the density of the values of the empirical 75th percentiles from those 1000 bundles of Gumbel variates.

3. **Association**

   (a) For each set of Gumbel draws, generate 1000 draws of Y corresponding to $Y_{ij} = -2G_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0,4)$.

   (b) Plot and briefly describe the distribution of the resulting 1000 Pearson r correlations between Y and G.

This is an optional exercise; it will not be graded.