ECOM90025 Advanced Data Analysis

Tutorial 2

Zheng Fan

The University of Melbourne

Introduction

Zheng Fan

- Ph.D. student in Economics
- Email: fan.z@unimelb.edu.au

Seek help?

- Ed discussion board
- Consultations: refer to Canvas for details

Shooting Stars

Consider a simple linear model

$$y_i = x_i\beta + \varepsilon_i,$$

where x's dimension is $p = 100$ (hence β). The true value of β is a vector of 0's. Such a setting assumes all variables of x are noises. The error term has a standard normal distribution, $\varepsilon_i \sim N(0, 1)$.

Carry out a simulation study to learn how likely some variables are falsely classified to have a statistically significant impact on the output variable y.

- Significance level $\alpha = 0.05$. Assume a sample size of $n = 300$.
- Let each element x_{ik} for the ith observations and kth variable to be independent and randomly drawn from a standard normal distribution $N \sim N(0, 1)$.
- Discuss simulation results and their implication for empirical works.

Simulation as a Tool

Simulation is a cost-effective approach if you want a quick understanding of a model or a random variable. For example, if we do not know what a chi-square distribution looks like, we can simulate many variables from standard normal and square them to have some visual impression.

1. Draw $B = 10000$ standard normal random variables.
2. Sqaure them.
3. Make a histogram and a kernel density plot for the simulated and transformed data.
4. What is the above distribution?

Order statistics

Definition (recap):

- Sample $X_1, X_2, \ldots, X_n$
- Arrange them in increasing order:

$$\begin{aligned} X_{(1)} &= \text{ Smallest of the } X_i \\ X_{(2)} &= \text{ 2nd smallest of the } X_i \\ &\vdots \\ X_{(n)} &= \text{ Largest of the } X_i \end{aligned}$$

- These are called the order statistics

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$$

- $X_{(k)}$ is called the kth order statistic of the sample

Order statistics from the Uniform distribution

There are $p = 10$ uniformly random variables denoted by $p_1, ..., p_{10}$ on the interval $[0, 1]$. Use a simulation method to investigate the order statistics $p^{(1)}, ...p^{(10)}$ (ascendingly ordered).

1. What are the means of these statistics?
2. Draw these means via a scatter plot with a proper reference line.
3. Discuss this exercise's implication for significance tests.

Order statistics from the Uniform distribution

In theory:

- The pdf of kth order statistic is

$$g_k(x) = k \begin{pmatrix} n \\ k \end{pmatrix} x^{k-1}(1-x)^{n-k}$$

  There is a formula for this. Google it if you are interested

- This is a beta distribution,

$$F\left(X_{(k)}\right) \sim \text{Beta}(k, n-k+1)$$

- So the theoretical mean is

$$\mathbb{E}\left(X_{(k)}\right) = \frac{k}{n+1}$$

$$\text{mode}\left(X_{(k)}\right) = \frac{k-1}{n-1}$$

The end

Thanks for your attention!