ECOM20001: Econometrics 1

Tutorial 2: Visualising and Describing Data in R, Conditional Density, Probability, Summation Problems

---

Overview

This tutorial consists of two parts. In part 1, you study visualising data with R. Part 2 provides practice problems with the summation operator.

Part 1: Visualising and Describing Data in R

A. Getting Started

Please create a Tutorial2 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following files into the Tutorial2 folder:

- tute2.R
- tute2_crime.csv

The first file is the R code for tutorial 2, the second file is the .csv file that contains the dataset for the tutorial.[1] The dataset has the following 5 variables:

- stateid: identifier for a US state
- vio: violent crime rate — incidents per 100,000 people
- rob: robbery rate — incidents per 100,000 people
- dens: population per square mile of land
- avginc: real per capita personal income in the state

With the R file and data downloaded into your Tutorial2 folder, you are ready to proceed with the tutorial. Please go to the tute2.R file to continue with the next 2 sections of the tutorial, which are Visualising Data and Descriptive Statistics.

[1] The reference research article for these data is: Donohue, J., Ayres, I. (2003): "Shooting Down the 'More Guns Less Crime' Hypothesis," *Stanford Law Review*, 55, pp. 1193-1312.

B. Questions

Having worked through the tute2.R code and graphs, please answer the following:

1. Discuss the sample means, standard deviations, min and max for each of the four main variables in the dataset: vio, rob, density, avginc.
    - What does a "typical" state look like in the dataset? Focus on sample means in describing a typical state. Be sure to state the units of a variable to accurately describe what a typical state looks like.
    - Discuss how the min and max of each variable, highlighting the range of values that each variable takes on. Just how varied is the degree of violent crimes and robbery rates, and population densities and per capita incomes in the sample? How violent and robbery-filled is the worst state compared to the best state?
2. How do the respective probability densities of vio, rob, density, avginc look?
    - Focus on their mean, standard deviation, and skewness
3. Comment on the 3 scatter plots listed below; these should be saved in your working directory for tutorial 2. Visually, does a relationship appear exist in each graph? If so, offer an **economic explanation**[2] for why the relationship might exist. There may be multiple explanations, so you may offer various explanations if you wish. But just one explanation is fine.
    - Robbery vs Violence: fig_nice_scatter_rob_vio.pdf
    - Robbery vs Per Capita Income: fig_nice_scatter_avginc_vio.pdf
    - Robbery vs People per Square Mile: fig_nice_scatter_dens_rob.pdf

[2] **Economic explanations focus on the costs and benefits of a particular behaviour for explaining empirical patterns.** Here's an example economic explanation for a positive relationship between robbery rate and person per square mile: a robber might find it easier (or less costly) to rob people in a dense city than in a sparse regional town because there is less distance between the robber and people (e.g., the robber's targets) in cities. This could be an economic explanation for a positive relationship between the robbery rate and people per square mile (or urban density). Other possible economic explanations to consider: what if people are richer in urban places (affects robbers benefit from robbing people)? What is police are less effective in identifying robbers in dense places than in regional places (affects the robbers expected cost of robbing people)? Are there other economic explanations you could think of that affect robbery cost and benefits? Similarly for violent crime offender costs and benefits?

Note: Part 2 of this tutorial contains extra practice exercises and will potentially only be partially covered in the tutorial, depending on time remaining. Solutions will be provided for students to work through and follow-up on in consultations.

Part 2: Practice Problems

Conditional Distributions

1. Suppose you have a random variable X that is i.i.d distributed from a N(mu$_X$,1) distribution, and a separate random variable Y that is defined as follows:

$$Y=2+2X$$

a. What is the distribution of Y?

b. Graphically plot the conditional distribution of Y if mu$_X$=2, if mu$_X$=5, and if mu$_X$=10. What is happening to the conditional distribution of Y for these different X values?

c. Suppose Y was instead distributed as

$$Y=2+4X.$$

- What is the distribution of Y now?
- Again, graphically plot the conditional distribution of Y if mu$_X$=2, if mu$_X$=5, and if mu$_X$=10 and compare your results to what you found in part b.
- What can you conclude about the magnitude of how shifts in the conditional distribution of Y as a function of different X values as the magnitude of the slope in the linear function that defines Y increases?

Note: In answering this question, you may simply draw the conditional distributions of Y by hand, or you may use R to plot the different conditional distributions of Y given X.

Conditional Probability

2. Consider the following table which describes the joint probability distribution for all combinations of studying and performance. The outcome space for Studying (Y) and Performance (X) is:

- Y - Studying: Study Hard, Study Sometimes, Study Never
- X - Performance: High Grade, Medium Grade, Low Grade

	High Grade	Medium Grade	Low Grade	Total
Study Hard	0.20	0.10	0.02	0.32
Study Sometimes	0.07	0.30	0.10	0.47
Study Never	0.01	0.05	0.15	0.21
Total	0.28	0.45	0.27	1

a. What is the marginal distribution for studying?

b. What is the marginal distribution for performance?

c. What is the probability distribution of Performance, conditional on Studying hard?

d. What is the probability distribution of Performance, conditional on Studying Sometimes?

e. What is the probability distribution of Studying, conditional on Medium Grade?

f. What is the probability distribution of Studying, conditional on Low Grade?

g. Using an example from the table above, show that Studying and Performance are not independently distributed.

Summation Practice Problems

1. Show the following equality is true:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

2. Show the following equality is true:

$$n\bar{x} = \sum_{i=1}^{n} x_i$$

3. Show the following equality is true:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

4. Show the following equality is true:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$