Probability Models for Beginners

Bowman, Adrian
The University of Glasgow
School of Mathematics and Statistics
University Gardens
Glasgow G12 8QQ, UK
E-mail: adrian.bowman@glasgow.ac.uk

1. Introduction

I used to be uncertain – now I'm not so sure.

The concepts of probability, and the tools of probability modelling, are to many people simultaneously intriguing, rather inaccessible and often counter-intuitive. This is true for those taking introductory courses at undergraduate level, those who need the tools of probability and statistics for research and professional work in other subjects, and indeed for the public at large. All these groups are 'beginners' and, while the approaches to learning and the appropriate levels of technical detail may be different, there is a common need to understand the motivation for modelling uncertainty, to understand something of the thinking involved, and to gain some intuition on the meaning of random variation.

2. Motivation

In studying any subject for the first time, motivation is important. For probability, some might be intrigued by the logical framework it provides but for many it is the relevance to contextual problems, and sometimes the appearance of surprising answers, which stirs up interest. Here are some well known problems which are guaranteed to provoke discussion among those who have not met them before.

Birthday problem

> On a football field there are 22 players plus a referee. What is the probability that two (or more) of these people have the same birthday?

Rare disease

> A test for the presence of a contaminant gives the right answer 90% of the time, both when samples have the contaminant and when they haven't. The contaminant is present in 10% of the samples which are taken from a particular site. If a sample tests positive in the lab, what is the probability that the contaminant is present?

Monty Hall

> A prize is hidden in one of three boxes and a contestant chooses one of the three. The host opens one of the other boxes to show it is empty. Is it a better strategy for the contestant to move to the remaining box or stick with the one she has chosen?

Contamination

> Bacterial vaccines have to pass a test for microbiological sterility. These tests are carried out on a bulk test of the vaccine and are then repeated on the individual ampoules or vials, which are filled aseptically before being released as a finished product. Regulations require that 20 filled ampoules are taken at random from each batch. The contents of each ampoule are tested for bacteriological contamination. The batch of ampoules will pass the test if each of the 20 taken are found to be free of living bacteria. If, due to a fault, each ampoule has a 3% chance of being contaminated with microbes, what is the probability that the test detects this?

We have probably all had the experience of seeing perplexity, scepticism or downright disbelief in the reactions of some, when hearing the answers. This provides an excellent opportunity to communicate the role of probability in providing a framework where problems of uncertainty can be worked through logically, defending against the danger of adopting 'intuitive' or 'obvious' answers.

3. Elementary ideas: thinking clearly

For beginners, a significant part of the learning process involves exposure to the logical arguments presented by others, exemplified by the lecture process. However, understanding often comes through grappling with these arguments for oneself. Where students struggle, the most helpful role of a tutor is often to nudge thinking in the right direction rather than simply to present the correct answer. This is a rather time-intensive process. However, there are some situations where the process is amenable to a degree of automation.

The task of identifying probability models, even in the very simple form of appropriate distributions, from a problem context is a skill which requires practice. Technology can help by providing a supportive environment in which students can be prompted with some simple guidance, to help in developing the skill of identifying appropriate distributions through the appropriate characteristics. Some 'Model Choice' software, developed by John McColl and others in Glasgow, will be described.

4. Elementary ideas: random variation

The level of technical detail at which a formal framework for probability is developed will, of course, depend on the context. However, there are more general common issues which face all beginners. One is the tendency to over-interpret the apparent structure exhibited in random variation. Students often see non-linearity in scatterplots, or bimodality and outliers in single samples, which suggests a lack of understanding of the meaning of random variation and a lack of intuition in assessing it.

Simulation is a very simple device to assist with this. For example, repeated simulation of data from three identical normal populations shows the extent to which individual samples can exhibit apparent differences, even when we know the data generating mechanism to be the same for each group. Figure **??** illustrates this, using interactive tools available in the `rpanel` package (**?**) for `R` (**?**). Of course, this kind of demonstration can be constructed fairly easily in many standard computing environments. However, the ability to use graphical controls and the 'packaged' presentation with

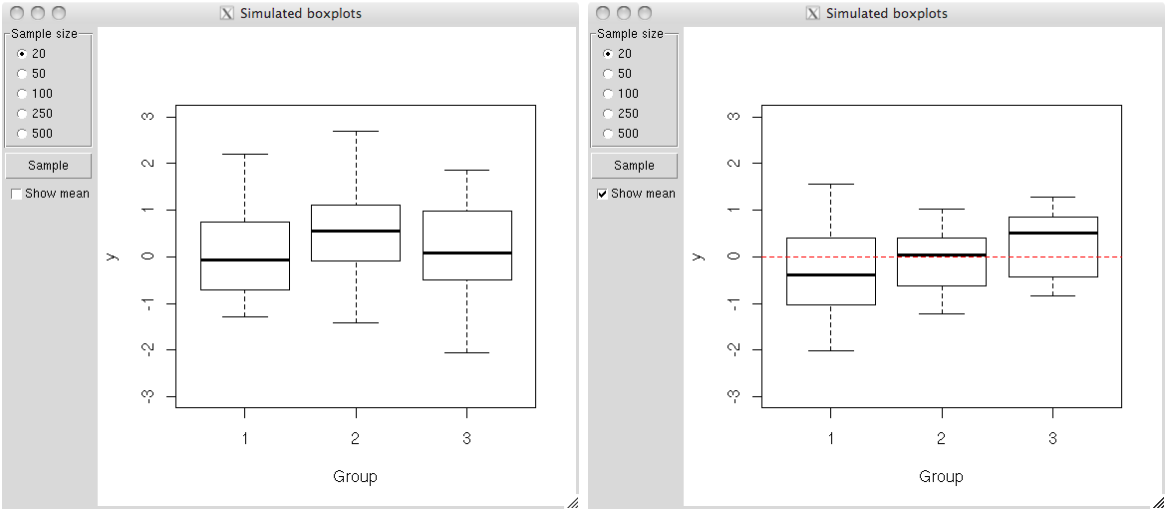

Figure 1: Repeated simulation of data from three identical normal populations.

hidden computer code has the helpful effect of focussing attention on the educational issues rather than those of practical construction.

Another useful example is in simulating q-q plots for normality, where the degree of non-linearity is the key indicator. Repeated simulation, aided by a quantitative measure of linearity such as the correlation coefficient, can lead to very useful discussions about how to compare observed data with the patterns exhibited by q-q plots from normal data. This can help to prepare the ground for more statistical ideas such as hypothesis tests, and indeed a very effective test of normality is based on the straightness of the q-q plot. However, it also helps to develop a more general appreciation of the nature of random variation and the need to quantify it.

Graphical methods are, of course, extremely helpful in displaying random variation. However, it may sometimes be the case that the precise definitions of medians etc., and the precise and hard edges of displays such as boxplots, communicate a message which is at odds with the idea of random variation as imprecise and 'fuzzy'. **?** propose the use of 'memory' and the idea of remaining in the same visual space in order to counteract this. Along similar lines, density strips (**?**) and density images can provide very useful representations of random variation in one or two dimensions. Although the methods of construction involve more complex ideas, the end result is more in keeping with the notion of random variation as fluid and imprecise. Figure **??** illustrates this on some principal component scores from data on aircraft (**?**).

5. More advanced concepts

Some 'beginners' are experienced scientists whose research work requires sophisticated statistical techniques but whose statistical training is limited. An intuitive approach, with a strong element of graphics and simulation, can help in understanding quite complex probability models.

For example, environmental scientists often need to use spatial methods, but the concept of a spatial process is a sophisticated one which can be difficult to understand at a technical level. **?** describe software in the `rpanel` package for `R` which simulates spatial processes, allowing interactive

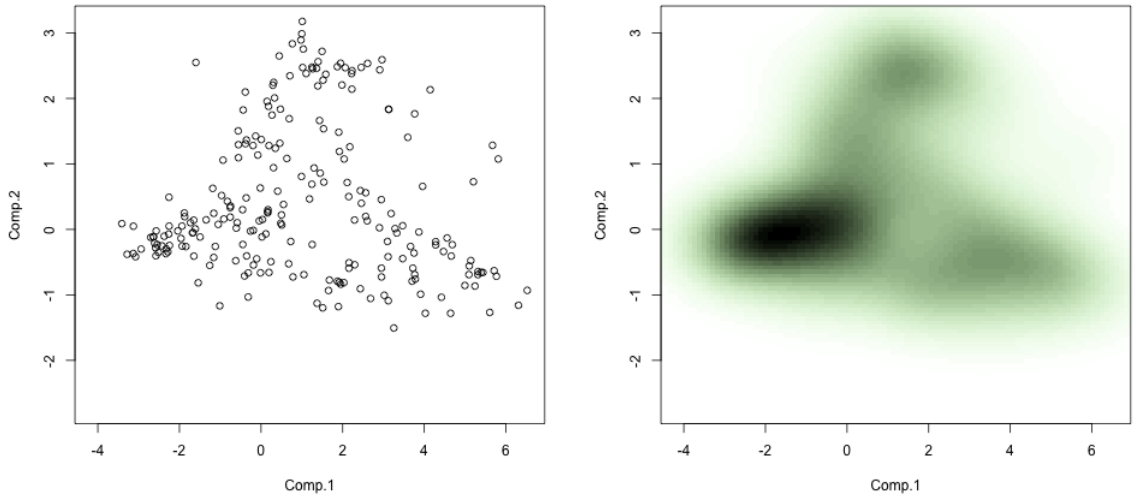

Figure 2: Scatterplot and density image of aircraft data.

control of the parameters involved. This is a useful starting point from which issues of sampling and the collection of point data can be discussed. Figure **??** illustrates the controls and the kinds of display which can be produced. This is a case where starting from the underlying probability model can prepare the ground well for an understanding of how the data arise and therefore how they might be modelled.

6. Tools for teachers

Technology has a very helpful supportive role to play in helping beginners, at any level, to grapple with the thinking involved in probabilistic argumentsand to gain intuition about random variation and associated probability models. Informative feedback can be programmed in some special situations while simulation and graphics have major roles to play in communicating the meaning of random variation and the nature of probability models.

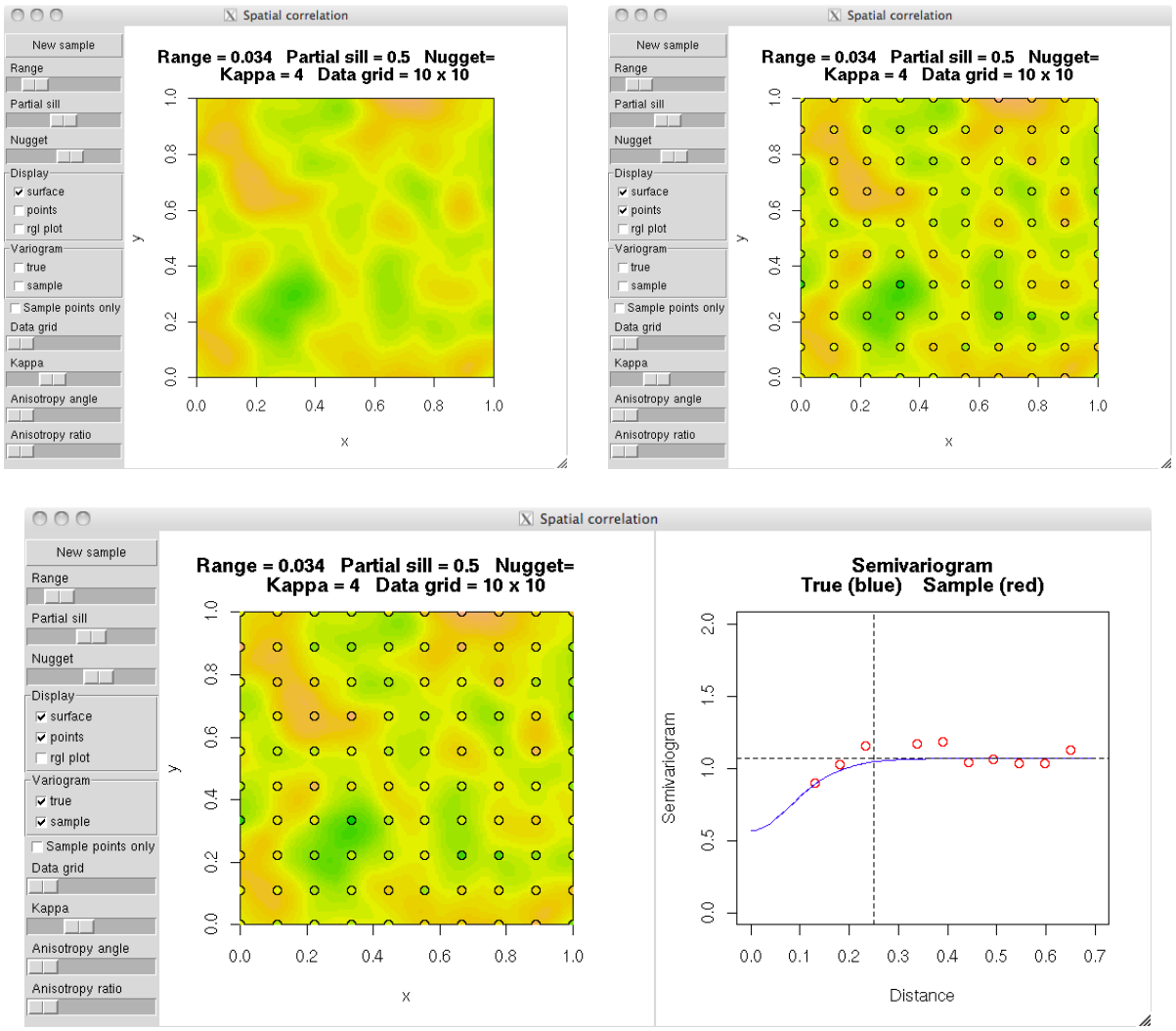

Figure 3: Spatial data.