

The BioSS Challenge – A demonstration of sampling bias

KEYWORDS:

Teaching;
Sampling;
Randomisation;
Bias.

Trevor S Smart

Biomathematics & Statistics Scotland,
Macaulay Land Use Research Inst, Aberdeen,
Scotland.

e-mail: trevorsmart@sandwich.pfizer.com

Summary

A sampling problem is described which was given to the general public to illustrate the problems of bias and the need for randomisation when sampling.

◆INTRODUCTION◆

AS part of the Macaulay Land Use Research Institute's Open Day, we set up a sampling problem for the general public. The problem introduced several statistical concepts that could be approached at many levels, from the simple calculations of averages and bias to more complicated ideas on sampling strategies and distributional assumptions. Each participant drew a sample which was entered into a computer to give clear graphical summaries of its properties and comparison with samples chosen by previous participants and samples chosen at random.

◆BACKGROUND◆

The general public does not have a clear understanding of the role of statistics in research, so when we were asked to contribute to the Macaulay Land Use Research Institute's Open Day, it gave us an excellent opportunity to explain this in greater depth. As part of BioSS's environmental modelling unit, we provide statistical advice for all aspects of the Macaulay Institute's research into the physical, environmental and social consequences of land use. Many view the work of a statistician as just summarising data. This is only a small part of what we do. Statisticians are involved in the experimentation right from the outset. We have an important role to play in the design of efficient experiments for which the results can be analysed to give meaningful results. As part of the Open Day we set up the 'BioSS Challenge', a sampling problem. The problem introduced several statistical concepts and could be approached at many levels.

We had several objectives that we were trying to meet. We wanted something interactive, so that the visitors could participate. It needed to be interesting and enjoyable with some educational value. The visitors on the open day would come from a wide background with differing levels of statistical knowledge, so our contribution had to be approachable at many levels, from primary school children to scientists. We also wanted to emphasise the need for statistics at all stages of an experiment.

◆THE CHALLENGE◆

The problem we asked people to solve was a simple sampling problem. A field has several patches of clover in it (figure 1) and we need to sample eight of the patches to estimate the average area covered by a clover patch. This was motivated by research at the institute into management practices on grazing pastures and their effect on clover patch size. The clover patches were of varying size, coming from a skewed distribution: the majority were small, with a few very large patches. Having chosen a sample of eight patches the participant was told the corresponding patch number and entered these into a computer. Graphs were drawn making several comparisons (figure 2):

1. A graph comparing the mean area of the eight chosen patches with the overall mean. The distribution of estimated means obtained from sampling eight patches at random was also given. This showed how accurate a participant's estimate was.
2. A graph comparing the participant's mean area, the true mean and the distribution of everyone else's sample mean. This enabled the participant to see how well they did in comparison with everyone else.
3. A graph comparing the inferences based on every-

one's samples with the true mean. This used the central limit theorem on the mean areas from all of the participants who had already taken part. If there was no bias then the true mean should lie well within the Normal curve.

This was then repeated by the computer selecting the patch numbers at random (without replacement). Each patch had an equal chance of being selected. The results from random selection were compared with the participants' selections. Finally a plot of the field showing the most popular clover patches was drawn and each participant was able to see which patches were most popular and whether they chose them.

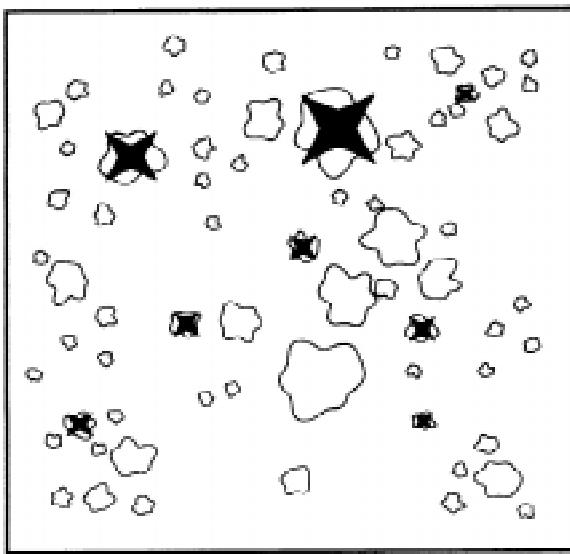


Fig 1. The clover patches in the field with a sample of eight chosen to estimate the mean area of a patch. The field is simulated with a deliberately skewed distribution, but such a picture could have come from aerial photography

◆RESULTS◆

There was bias in the sampling, with big patches being by far the most popular and small patches around the edge being the least popular (figure 3). This emphasised the need for randomisation and an unbiased sampling strategy, where each patch was equally likely to be chosen, or a method that accounted for the bias. The results clearly show how human selection can introduce bias. Some of the more statistically astute were able to see that a sample size of eight was too small to obtain a precise estimate of the mean area from such a skewed distribution.

The public participated with enthusiasm and enjoyed making simple comparisons of how they did in comparison to everyone else and how close their estimate was to the true mean. They also found it very revealing seeing which patches were popular, and to many it was a great surprise to find out the bias in people's sampling. Most people could clearly see the need for random selection. For a few the challenge led on to further discussions on sample sizes, and different sampling strategies.

Several people devised sampling strategies which they hoped would remove the bias. These included drawing grid lines, taking line transects, choosing some big and some small patches. Most of these strategies still introduced positive bias because they favoured the clover patches that covered a larger area. There was less bias when a queue had built up. People overheard what was said to previous participants about the big patches being very popular and tried to compensate in their selection. Some people over-compensated and chose only

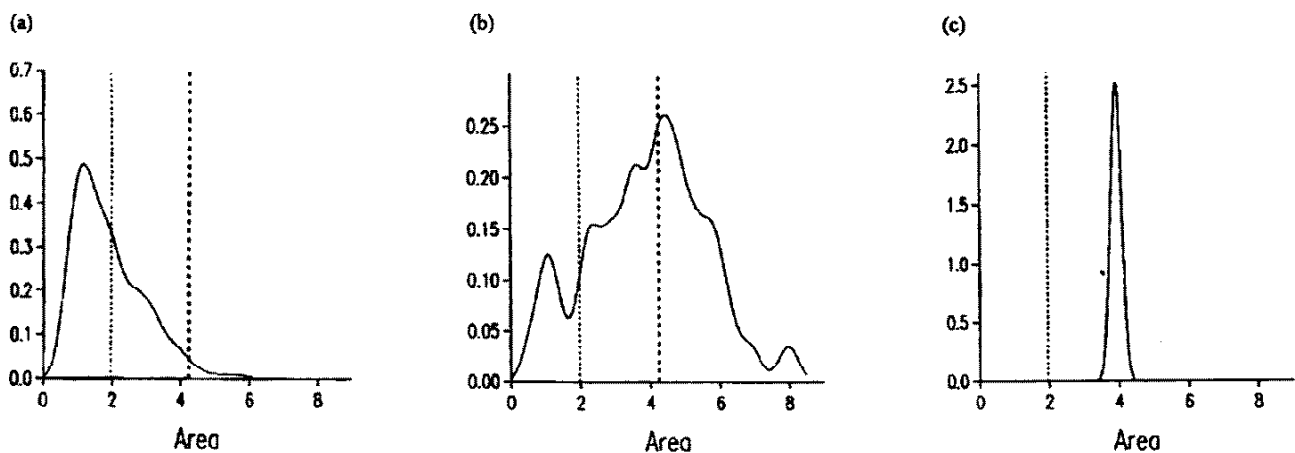


Fig 2. The participant's sample mean and inferences based on this are compared with the true mean and all the preceding sample means. (a) The participant's mean (—), the true mean (...) and the probability density function (p.d.f.) of sample means if the samples were chosen at random (—). (b) The participant's mean (—), the true mean (...) and the p.d.f. for the distribution of all previous sample means (...). (c) The true mean (...) and the p.d.f. of the Normal distribution for the estimated mean area based on the central limit theorem applied to all preceding sample means

small patches. One of the most accurate estimates came from someone who said he was taking the route a cow would take entering the field where he imagined there was a gate.

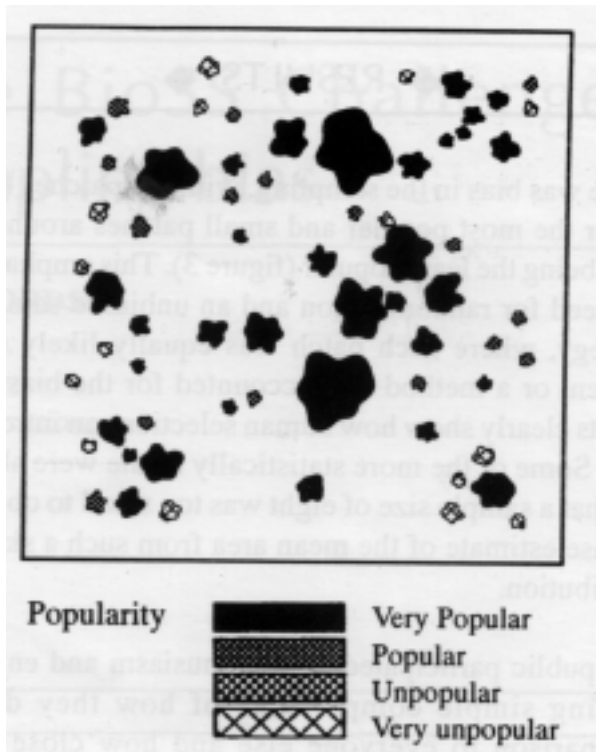


Fig 3. The clover patches, showing the popularity of each patch.

◆DISCUSSION ◆

The challenge was successful with the participants showing enthusiasm and competitions between family members ensued on several occasions. This sort of challenge of sampling from a skewed distribution could easily be used at schools and universities to introduce students to some of the statistical concepts and the need for randomisation when sampling. It could lead on to students trying to develop sensible sampling strategies discussing bias and the effects of skewed distribution on the sample size needed for the central limit theorem to hold.

The method of sampling a selection of clover patches is not the only unbiased way to estimate the average area of a patch and in practice other methods may be used. There are also methods of accounting for bias in selection, so it is not always necessary to ensure that each patch is equally likely to be selected.

Acknowledgement

This work has been done under funding from the Scottish Office Agriculture, Environment and Fisheries Department.

Footnote

The author's current address is Pfizer Ltd. Sandwich, Kent, England.