

Exploring Sampling

KEYWORDS:

Boxplots;
Estimation;
Regression.
Teaching;

James Nicholson

Belfast Royal Academy, UK.

Summary

This article attempts to integrate the crucial ideas of sampling and estimation in the teaching of linear regression. It won first prize in the recent PSI competition.

◆INTRODUCTION◆

SAMPLING, linear regression and correlation are topics which are encountered early on in most people's statistical experience, whether it be in a mathematical statistics course, or in applications of statistics in other subject areas. The least squares line of regression, and the product moment correlation coefficient are well defined functionally, and many calculators and computer software packages will generate them after elementary data entry procedures are followed.

I teach an A-level course which has 50% Statistics, in which these topics appear. I had been concerned that the pupils' perceptions of the regression line seemed almost to extend to a belief that this was the underlying relationship, and that the line, and 'predicted values' generated by the line of regression would be given to very high degrees of accuracy. These would be unwarranted even due to the accuracy to which the data had been recorded, before any consideration of the variability of the line due to sampling. I had also been concerned about the pupils' grasp of confidence intervals and particularly how dependent they are on the data used, and we were going to be looking at different sampling methods. We had already spent some time early in the course working with box and whisker plots, and had used them to make comparisons between sets of data with different centres and spreads.

Burghes (1994) contains a data set giving information on 200 trees on a piece of land that the owner wishes to sell. It is in a chapter dealing with data collection, and the activity suggested is for pupils to choose one of a number of sampling methods and construct an estimate of the average of various quantities such as value, girth and proportions of different types of tree. The class undertook this activity, but we then pooled all the results, and constructed boxplots of the data resulting from the various

(point) estimates of these quantities using random, systematic and stratified samples. These boxplots turned out to provide valuable insights into a number of different and difficult concepts.

◆SAMPLING METHODS◆

The nature of stratified sampling is illustrated by the diagram below, where the proportion of oaks in each sample remained the same since that was how the stratification had been constructed. The perfectly consistent prediction (of the true proportion) provided a focus for what makes a good estimator. In the same diagram the contrast between the profiles of the random and systematic sampling estimates opened up some worthwhile discussion as to why the systematic samples provided much more consistent estimates than the random samples in this case, and whether we would expect that in all cases.

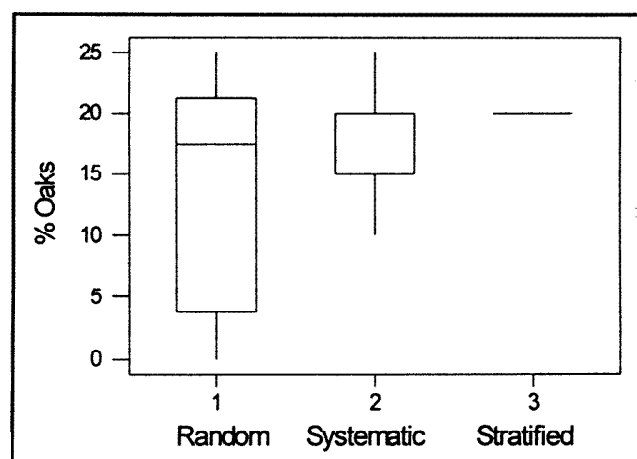


Fig. 1. Boxplot of proportion of oaks in different samples.

Figure 2 shows the estimates of the average value of the trees from the different sample types. Discussion was again forthcoming questioning

whether the size of our samples was sufficient to let us make strong statements concerning the relative merits and demerits of different sampling methods. Since the pupils had to do the work in producing the data from samples they showed a greater understanding than previous groups of the 'costs' involved in improving the quality of your conclusions by considering more data.

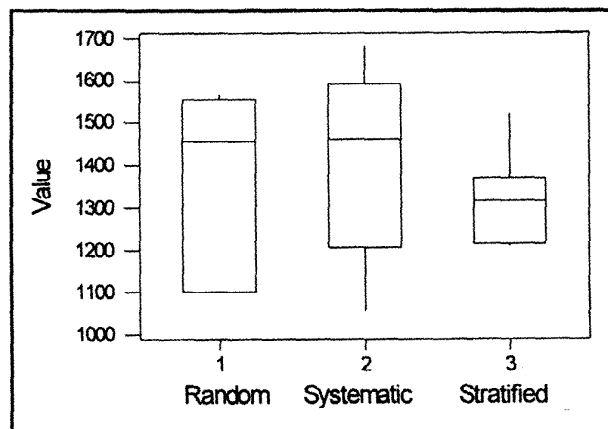


Fig. 2. Boxplot of average values from different samples.

◆INTERVAL ESTIMATION ◆

The ideas of interval estimation were able to be developed quite naturally from the experience of generating a number of different point estimates of the same quantity, and finding that they did not always produce the same value for the estimate, and that the consistency of estimate was affected by a wide range of factors including the size of sample used, the method of constructing the sample and the underlying variability of the quantity. This led on to some interesting investigations of other data sets trying to quantify some of these effects.

The process of repeating a number of samples using the same procedure to generate them, but obtaining different sets of data each time, and sometimes quite different estimated values, meant that they had an experience to draw on, which informed their intuition when dealing with subsequent situations involving sampling. Instead of a set of abstract rules which previous groups could learn, and apply, often very successfully, variance of estimated values varies inversely with the sample size, for example this group began to appreciate intuitively, based on experience, how the consistency of the estimates would vary.

◆REGRESSION ◆ AND CORRELATION

This led on to looking at the reliability of data used in other circumstances, such as in linear regression

and correlation. We took a set of data giving the body and heart masses of fourteen 10-month old male mice (Table 1), and looked at the regression lines and the predicted heart masses for certain body masses that would be generated by samples of the set of data. Figure 3 shows a scatterplot of the body and heart masses. Each of the data points was discarded in turn. Table 2 shows the values of the correlation coefficient, the coefficients of the equation of the line of regression of heart mass on body mass, and the 'predicted' values of heart mass for body masses of 20,

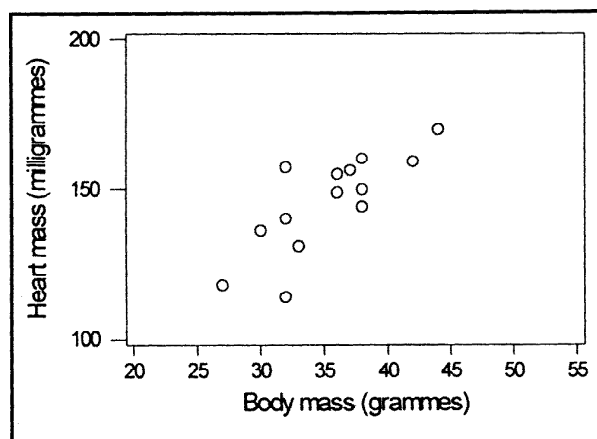


Fig. 3. Scatterplot of body and heart masses of mice.

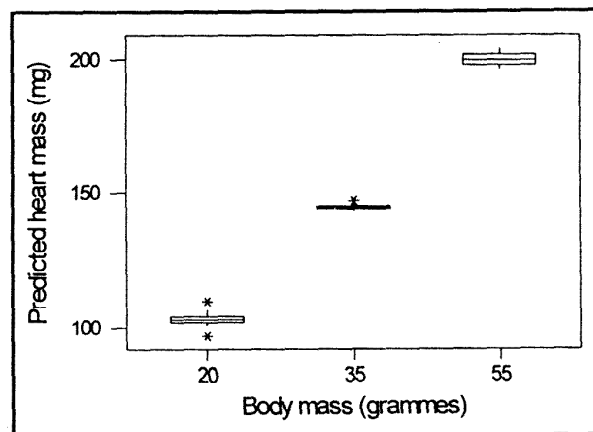


Fig. 4. Boxplot of predicted values from different samples.

35 and 55 grams. Figure 4 shows a boxplot of the predicted values obtained. Much greater variation arose when smaller samples (using 10 out of 14 mice for example) were chosen, and when other data sets were used where the correlation was weaker.

Again a number of important issues arose from this exercise:

- The regression line was appreciated much better as being an estimate of the underlying trend.
- The regression lines based on various samples were seen to be diverging as you move from the centre of the x-values, and the boxplot shows clearly the greater consistency of predicted val-

Table 1. Body and heart masses of 14 mice.

Mouse	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Body Mass	27	30	37	38	32	36	32	32	38	42	36	44	33	38
Heart Mass	118	136	156	150	140	155	157	114	144	159	149	170	131	160

Table 2. Regression statistics calculated from samples of 13 mice.

Mouse discarded	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Correlation	0.715	0.785	0.785	0.787	0.787	0.792	0.865	0.827	0.802	0.778	0.787	0.729	0.784	0.781
Intercept	56.0	43.8	49.2	47.5	46.4	48.4	36.9	60.8	45.8	43.5	48.4	49.2	51.7	50.4
Gradient	2.55	2.87	2.71	2.78	2.80	2.73	3.03	2.45	2.84	2.90	2.75	2.72	2.68	2.68
y for x = 20	107	101	103	103	102	103	97	110	103	102	103	104	105	104
y for x = 35	145	144	144	145	144	144	143	147	145	145	145	144	146	144
y for x = 55	196	202	198	200	200	199	204	196	202	203	200	199	199	198

- ues close to the middle of the range.
- The problems associated with extrapolation take on a new dimension. Not only may the existing fairly strong linear relationship not continue, but even if it does the predicted values become increasingly unreliable.

◆CONCLUSION ◆

On reflection, I was pleased at how coherently the different aspects of this work fitted together, and

indeed provided reinforcement for one another. In particular the pupils' intuitive understanding of some quite difficult and subtle concepts seemed to be more secure for being experientially based, rather than purely learnt by formal mathematical principles. theorem and proof. although I still regard these as an essential part of statistics.

Reference

Burghes, D. (1994) (Ed). *Statistics: AEB Mathematics for AS and A-Level*. Heinemann.
