ON TEACHING BASIC STATISTICS: A CAPTURE-RECAPTURE EXAMPLE

Flavia Mascioli[1], Carla Rossi[2] and Daria Scacciatelli[3]
[1]Università di Roma "La Sapienza", Italy
[2]Università di Roma "Tor Vergata", Italy
[3]CIBB, Università di Roma "Tor Vergata", Italy
flavia.mascioli@uniroma1.it

Nowadays many of us know that in an introductory statistics course great emphasis should be placed on problem-solving and active learning. This paper presents a real capture-recapture experiment used as a starting point to introduce fundamental concepts as experiment, sample variation, sampling distribution and bootstrap. It has been used in an introduction to statistics 2 course for undergraduate biology students; a particular attention has been given to their special interest in learning how to use statistics to solve problems and to their general negative attitude towards this discipline. A bootstrap application, based on the observed experimental data, has been carried out using the R package. Student feedback and reactions are also discussed.

INTRODUCTION

In the traditional lecture based approach, courses are often too abstract for science students, with the consequence that they do not really seize the key role of statistics in scientific work and real life. Furthermore, statistical concepts are not fully integrated with the students' discipline knowledge, so that not only are they often incapable of choosing the appropriate analysis in their research studies, but they also rapidly forget what they have been taught. The ideas of "learning by doing" (Boyle, 1999) and "working with data" (Moore et al., 1995) are now well established. "The practice of statistics involves a dialog with data rather than once-for-all analysis, contact with other disciplines, and a team approach, so the new style of teaching is easily accepted by those who want to bring teaching closer to practice" (Moore et al., 1995). With this purposes in mind we present here an example of a mark-recapture experiment which has been successfully used by two of us (Mascioli for two years and Rossi in this semester), in a second year one semester optional course, addressed to biology students. Students entering this course are familiar with statistical concepts such as confidence intervals and hypothesis tests, with the use of statistical software and have experienced simulation. Its purpose is to emphasize key concepts, to expand knowledge of statistics and to develop statistical reasoning to analyze data. Students (about 50) attend two weekly 100 minute lectures and one 90 minute lab session (with Scacciatelli). They are exposed to three assessments and one final group project.

Very often we ask ourselves how to get students to grasp basic concepts. As statistical ideas begin with data we decided to ask biologists for help and they provided data on their research (Angelini et al., 2009). The main objective of their experiment was to identify demographic parameters for the female population of *Salamandrina perspicillata*, a salamander unique to Italy. They monitored the population inhabiting a forest near Rome where there is a trough of about 2800 litres. They surveyed the site from 1998 to 2006 during the oviposition period considering only ovipositing females, which could be sexed for certain. Salamanders were marked by taking pictures of their ventral colouration pattern, which is unique to individuals and persistent.

After data presentation and bootstrap application it was possible to engage the class for discussions trying to solve different problems. Various topics were presented and discussed emphasizing two concepts which students have difficulty developing and which are "the heart of statistics and fundamental components of statistical thinking": variability and sampling distribution (Garfield et al., 2005).

HOW TO ESTIMATE THE SIZE OF THE FEMALE SALAMANDER POPULATION

Initial class discussion of this case involved the understanding of the research question and the choice of an appropriate experimental design which influences the selection of the statistical methods. Great emphasis was given to the basic ideas of statistical design. Students had a prior limited knowledge of capture-recapture methods. Capture-recapture studies were originally developed in the wildlife biology to estimate demographic parameters and trends in population studies. The classical problem of estimating the unknown size of a closed population is the main

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

issue of this case study. In 1998, biologists sampled, in the surveyed site, the ovipositing female population over 11 occasions. Only the oviposition period, which occurs in winter-early spring, was considered so that the population size remains fixed during the study time.

Individuals were captured, marked and then released and allowed to mix again with the general population. Subsequent recaptures were performed and the marked individuals were recorded. The recorded counts of capture-recaptures were: f_1=81, f_2=17, f_3=0, f_4=1, where f_k is the frequency of individuals captured exactly k times in the 11 trapping occasions. The maximum frequency for each individual is the number of occasions; n=99 is the number of distinct females caught in the experiment. The complete capture history for each female is expressed as a sequence of 0's and 1's, where 0 denotes absence and 1 denotes presence. So we have a 99x11 matrix $\mathbf{X} = (x_{ij})$, where x_{ij} = [the ith individual is caught (1) or not (0) in the jth sample] i = 1, 2,…, 99; j = 1, 2,…,11. The number f_0 of individuals never observed is unknown as is the population size: $N = f_0 + f_1 + f_2 + f_3 + f_4 = f_0 + n = f_0 + 99$. To estimate N, as most captured females were caught only a few times, we will use the non parametric estimator proposed by Chao (1984):

$$\hat{N}_C = n + f_1^2/(2f_2).$$

The estimator was introduced avoiding computational procedures, because this activity was focused on developing statistical thinking and reasoning.

Intuitively, the capture frequencies contain all the information to estimate the number of missing individuals in the samples. If the recaptures are few $\Rightarrow$ size is much higher than the number of distinct captures. If the recaptures are elevated $\Rightarrow$ we are likely to have caught most of the animals. Moreover, the emphasis on the lower frequency classes makes sense: females never observed are more similar to those captured a few times than to those captured many times. If $f_3, f_4, \ldots$ hold valuable information, $\hat{N}_C$ can be considered only as a lower bound of the population size. With the data of this experiment the value $\hat{N}_C = 292$ has been obtained. Unfortunately the distribution of $\hat{N}_C$ in practice is skewed to the right. It has been pointed out that this estimator is based on three assumptions: that the population is closed, that the individuals act independently and that their capture probabilities are not homogeneous. Students are encouraged to think that models should be generalizable to other situations (for example, with this model it is possible to estimate the size of a human population with a certain disease or one which is difficult to identify).

It can also be considered why in this situation the classical Lincoln-Petersen estimator, which is based on two capture occasions only, cannot be applied.

PROPERTIES OF THE ESTIMATOR: A NON PARAMETRIC BOOTSTRAP APPLICATION

We only have a sample from a non-parametric model and we want to calculate bias, variance and confidence interval of $\hat{N}_C$. To quantify the precision of this estimator we use the bootstrap which allows simulating the statistical model to fit the biological data; each trapping occasion represents one experimental trial of the model:

1) we draw a bootstrap sample with replacement from the original sample of the 99 capture histories, obtaining a bootstrap matrix 99x11 which represents the entire experiment

2) we calculate $\hat{N}_C^*$ based on this matrix

3) we repeat the above steps B times and obtain the empirical bootstrap distribution of $\hat{N}_C^*$ which approximates the unknown sampling distribution of $\hat{N}_C$

4) we calculate the mean of this bootstrap distribution obtaining a bootstrap estimate $\hat{N}_C^B$ of N

5) we also calculate the standard deviation $S\hat{D}_C^B$ of the bootstrap distribution and the bootstrap estimate of the bias of $\hat{N}_C$

6) we obtain a 95% bootstrap percentile confidence interval by finding the two bootstrap estimates that have, respectively, 2.5% of the bootstraps below and 2.5% above them. This method is not perfect and can be improved, but is good enough for our purposes.

Students should understand that we avoid the task of taking many samples from the population by instead taking many resamples from a single sample.

RESULTS AND DISCUSSION

We repeated the process previously described, using B = 250, 500, 1000 replications (Figure 1). From Figure 1 it is easily seen how the population estimates vary from sample to sample so that students' experience with variability and sampling distributions behaviour can be immediate and concrete. Students cope with a new scenario, with a different estimator from the ones they usually see. The skewness of the bootstrap distributions does not vary by increasing the number of replications. The bootstrap resampling process (using more resamples) introduces little additional variation. We can rely on a bootstrap distribution to inform us about the unknown sampling distribution.

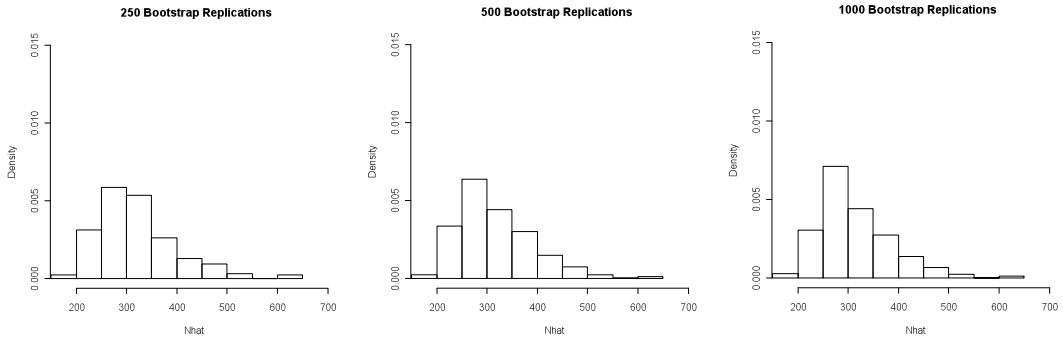

Figure 1. The bootstrap distributions of the estimated population size (year 1998)

Both the variances and the confidence intervals (that are not symmetric reflecting the skewness) are quite large; this is due to the possible occurrence of a very small number of recaptured individuals (f_2) and to a possible model mis-specification bias as well (Table 1).

Table 1. Bootstrap distributions: B= 250, 500, 1000 replications (year 1998)

Bootstrap replications	$\hat{N}_C^B$	$S\hat{D}_C^B$	Confidence interval
250	320	77.60	212-539
500	315	73.98	212-495
1000	312	71.56	208-486

A bootstrap estimate $\hat{b}$ of the bias of $\hat{N}_C$ can be computed as the difference between the average of the bootstrap distribution and the estimate of *N*: $\hat{N}_C^B - \hat{N}_C$. For the three bootstrap distributions we obtain respectively: $\hat{b}$ = 27.8; 23.19; 19.61, therefore $\hat{N}_C$ is a biased estimator of N.

For a comparison we also considered data regarding the ovipositing female population in 2000 (Figure 2).

The frequencies of capture-recaptures were f_1=160, f_2=79, f_3=27, f_4=7, f_5=2. The number of distinct females captured in 13 trapping occasions was n=275. We obtained a 275x13 matrix and we performed an analogous bootstrap study. Chao's estimate for the size of the population of females is $\hat{N}_C$ = 439.

Figure 2 shows that the three bootstrap distributions are only slightly skewed because of the great increase of recaptures compared to the year 1998. For the same reason the standard deviations of the bootstrap distributions are much lower (Table 2). Also the values of $\hat{b}$ are much lower: $\hat{b}$ = 4.04; 6.6; 5.1. For both years, as expected, the bootstrap distributions are approximately centered at the original statistics value.

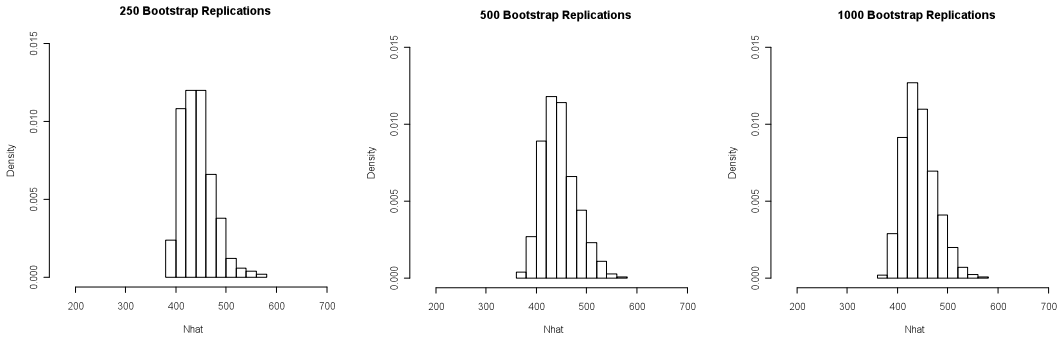

Figure 2. The bootstrap distributions of the estimated population size (year 2000)

Table 2. Bootstrap distributions: B= 250, 500, 1000 replications (year 2000)

Bootstrap replications	$\hat{N}_C^B$	SD_C^B	Confidence interval
250	443	32.15	392-518
500	446	34.07	391-522
1000	444	32.50	391-516

We have seen how the bootstrap goes one step further in comparison with simulation because not only does it allow us to clarify abstract and difficult concepts letting students to experiment with random samples, but it also permits statistical inference in all those situations where classical methods are not applicable. So bootstrapping proves to be an efficient tool in teaching basic concepts.

Discussions during the activity allowed the instructor to identify and focus on the most prevalent students' misconceptions. Most of the students had difficulties in understanding that there are two sources of variation among bootstrap distributions, that using more resamples does not introduce great variation and that the identification of the most appropriate model for the observed data is often not an easy task. On the other hand the students pointed out that the bootstrapping was important to see the variation of estimates and the approximation to the sampling distributions for non standard estimators. They also appreciated being able to apply statistical methods to problems in biology.

Assessment (homework and a test at the end of the activity) were used to evaluate students statistical thinking as described by Cobb (1992). Several questions were selected and adapted from delMas et al., (1999) and from the ARTIST Web site (https://app.gen.umn.edu/artist/). Only a few students exhibited a faulty reasoning with sampling variability and sampling distributions.

REFERENCES

Angelini, C., & Antonelli, D. & Utzeri, C. (2009). Capture-mark-recapture analysis reveals survival correlates in Salamandrina perspicillata. *Amphibia-Reptilia*, *00*, 1-6.

Boyle, C. R. (1999). A Problem-Based Learning Approach to Teaching Biostatistics. *Journal of Statistics Education*, *7*(1).

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, *11*, 265-270.

Cobb, G. (1992). "Teaching Statistics". In L. Steen (Ed.), *Heeding the Call for Change, MAA Notes No. 22, Washington: Mathematical Association of American* (pp. 3-34).

delMas, R. C., Garfield, J., & Chance, B. L. (1999). A model of classroom research in action: developing simulation activities to improve student's statistical reasoning. *Journal of Statistics Education*, *7*(3).

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*(1), 92-99.

Moore, D. S., Cobb, G. W., Garfield, J., & Meekar, W.A. (1995). Statistics Education Fin de Siecle, *The American Statistician*, *49*(3), 250-260.