ACTIVITY-BASED STATISTICS, COMPUTER SIMULATION AND FORMAL MATHEMATICS

Joachim Engel
University of Education
Germany

Instructional methods involving students in activities for exploring statistical concepts have proven to be highly effective. Formal mathematics, on the other hand, constitutes the basis of inductive reasoning. This paper reports on an "ActivStats" class for college math majors that teaches statistical concepts as well as mathematical foundations. Its basis is a four-step procedure comprising problem analysis, student activities, computer simulation, and formal mathematical analysis

INTRODUCTION

Activity-based methods have been recognized to be a pedagogically extremely valuable approach to teaching statistical concepts (see, e.g. Scheaffer et al. 1997, Rossman, 1997). This approach is rooted in concepts of modern pedagogy emphasizing student's activities in problem-solving situations. Students learn mathematics and statistics with understanding when they actively build new knowledge from experience and prior knowledge. To develop a good conceptual understanding of statistics students need hands-on experience with data collection and data analysis. Perhaps the biggest problem with statistics instruction of "the old days" was the early introduction of formal mathematics. Probability, however, is a very powerful theory which constitutes the basis for inductive reasoning. For mathematically inclined students (such as future math teachers) probability instruction should not be limited to heuristics and simulation but should also include formal concepts of mathematics. The field of statistics provides ample opportunities to teach important concepts of applied mathematics and modeling. In fact, an activity-based statistics course may meet all the major demands that apply to modern mathematics instruction (NCTM, 2000), such as preparing students to

- represent and analyze real situations
- solve problems
- make decisions using mathematical reasoning
- communicate their thinking
- make connections

Approaching "real-world-problems" with statistical methods that are founded on formal mathematics provides compelling examples of the power and utility of mathematics. Most problems in probability can be solved either exactly by analytic methods or approximately through simulation. The latter may be the only option available in situations where the problem is too complex for an analytic approach. Simulations help students focus on conceptual understanding instead of being caught up in formal mathematical derivations. An important intermediate step between an activity or experiment and a mathematical analysis is developing a simulation model, represented by the computer. As supported by recent research, instruction that incorporates simulation promises to help students acquire conceptual, not merely mechanical, understanding of statistical concepts.

A FOUR-STEP-PROCEDURE

1. Introduction of a "real-world" problem involving some aspects of data analysis
2. Doing an activity related to understanding the dynamics of the problem. The activity somehow represents the problem and involves a simulation model and the generation of data with a physical random number generator
3. Representing the simulation model with a computer-based random number generator, generating data and trying to make sense of it, including simulation-based inferences
4. Mathematical analysis based on probability and mathematical statistics

There are excellent ideas for steps 1 and 2 in the literature (e.g. Scheaffer et al., 1997; Rossman, 1997) and these steps motivate students with some mathematical background to grapple with the mathematics involved in step 4. Step 3 requires a flexible and versatile programming environment that allows students to express their ideas about statistical models. Here my choice is LISP-STAT (Tierney, 1990), enhanced with macros provided by the instructor or downloaded from the WWW. Specific content that I cover in my ActivStat class include capture-recapture models, the regression effect and modeling scatterplot data, runs-up-and-down, Central Limit Theorem, the bootstrap, randomized response sampling, stratified sampling and design of experiments. Steps 1 and 2 of the following two examples are well explained in Scheaffer et al. (1997). Therefore, we concentrate on step 3 and 4.

EXAMPLE 1: CAPTURE-RECAPTURE

In light of international agreements on biodiversity and the protection of species, estimating animal abundance is of relevance not only for conservationists. One simple estimator is based on the capture-recapture method. A nice activity related to this problem is described in Scheaffer et al. (1997) asking students to draw (recapture) n Goldfish crackers (the "fishes") from a bowl or bag (the "lake") after a number m of "fishes" had been captured and marked before. Assuming that the fishes have been mixed well between the two captures, simple proportionality leads to the following estimator of the unknown population size

$$\hat{N} = \frac{mn}{X}, \qquad (1)$$

where X denotes the number of marked "fishes" in the recapture.
An analysis of the quality of this estimator is quite challenging because of the random variable X in the denominator. Alternatively, after physically simulating the problem with Goldfish crackers, a computer simulation may provide an evaluation of the estimator (1). First we define various parameters and represent the "fishes" by 0's, e.g., N = 500. Then mark m = 100 fishes (by changing 0 to 1) and recapture n = 80 fishes.

```
(def lake (repeat 0 500)),
(def m 100)
(def n 80)
```

The following LISP-STAT procedure marks m fishes in the "lake" by changing their label to 1:

```
(defun capture (m lake)
      (setf (select lake (iseq 1 m)) (repeat 1 m)))
```

The following recursive procedure produces a list of r estimators of the population size by repeating the simulation r times:

```
(defun recapture (r)
 (if (= r 0)( )
   (cons (/ (* m n)(sum (sample lake n)))
     (recapture (- r 1)))))
```

It is instructive to redo this experiment with different parameters to investigate questions like: "Is it better to mark more fish?" ,"Is it better to take a larger sample in the recapture phase?", "What happens when M or n are too small?", "Is it advantageous to modify the sampling plan and to keep on catching fishes in the recapture phase until a pre-defined number of marked fishes have been caught?" Investigations may be planned with an experimental design resulting in representations as in Figure 1, which shows four different boxplots containing 100 estimates each. Obviously, estimates are better the bigger m and n are. Moreover, the distribution is skewed to the right, i.e. the capture-recapture methods tend to overestimate the true population size as seen by the location of the medians and the long upper whiskers.

The problem of estimating the population's size contains almost everything on formal probability an introductory course in probability and statistics would want to offer: distributions (hypergeometric, binomial, negative binomial) and concepts for point estimation (bias, variance, maximum-likelihood, hypothesis testing, confidence intervals). Moreover, the problem invites one to question the modeling process: What problems might there be in practically applying this method and how does each problem affect the resulting estimate?

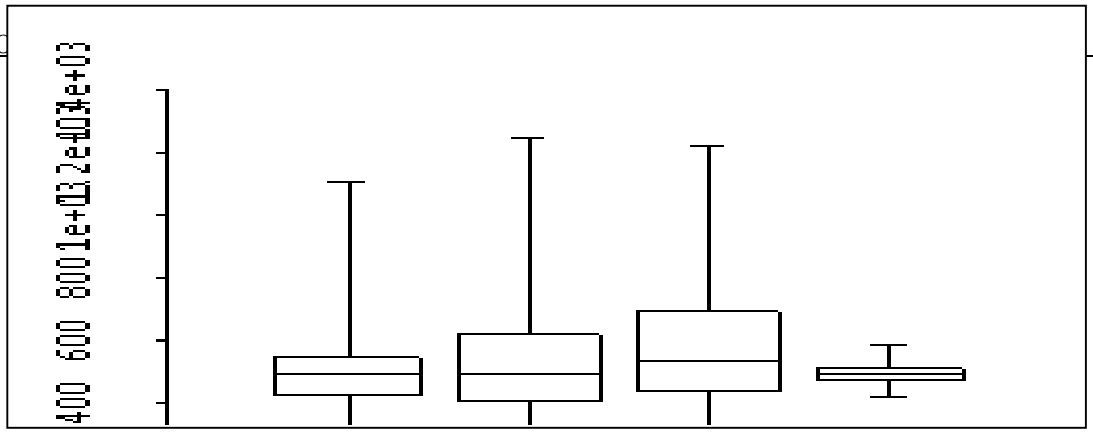

Figure 1: Comparative Boxplots from the Capture-Recapture experiment under different experimental conditions: far left: $m = 100$, $n = 100$; center left: $m = 50$, $n = 100$; center right: $m = 70$, $n = 70$; far right: $m = 200$, $n = 200$. Results are based on

EXAMPLE 2: REGRESSION-TOWARDS-THE-MEAN

The regression effect is a phenomenon that leads easily to misintepretation in studies following a test-retest design. A very instructive activity illustrating this effect is described by Levin (1993) where students draw two types of cards, one representing a score for skill and the other representing random influences, and add these two numbers to obtain an "observed" test score. Then the card representing the random influence (and only this card!) is replaced by a second draw. Its sum with the value of the "skill" card is the retest score. As a consequence one notices that the "observed" scores regress towards the mean, i.e. students with a high "observed" test score tend to have a "lower" retest scores and, vice-versa, students with low test scores tend to improve even though their "skill" card remains unchanged.

A possible representation of the regression effect in LISP-STAT is as follows: The following subroutine defines two scores $x_i = k \cdot a_i + e_i^{(1)}$ and $y_i = k \cdot a_i + e_i^{(2)}, i = 1,...,n$, where a_i, x_i, y_i represent the true score, the test, and the retest scores of the *i*-th person while $e_i^{(1)}$, $e_i^{(2)}$ is noise and *k* an intensity parameter allowing different levels for the signal-to-noise-ratio. The output is the slope of the regression line which tends to be less than 1.

```
(defun regression-effect-sub (k)
 (def a (/ (iseq 1 100) 100))
 (def e1 (normal-rand 100))
 (def e2 (normal-rand 100))
 (def x (+ a (* k e1)))
 (def y (+ a (* k e2)))
 (def regmodel (regression-model x y :print nil))
 (first (rest (send regmodel :coef-estimates))))
```

The following procedure allows *r* repetition of this experiment:

```
(defun regression-effect (k r)
 (if (= r 0)()
   (cons (regression-effect-sub k)
         (regression-effect k (- r 1)))))
```

Now the experiment can be repeated for varying intensity parameters *k* to compare the regression effect under different signal-to-noise-ratios.

Theoretical considerations require a model

$X = T + k \cdot \varepsilon, \quad Y = T + k \cdot \eta,$

where the random variable T, X, Y represent the true and observed scores. Simple considerations lead to a slope of the regression line of $m = m(k) = \dfrac{\text{var}(T)}{\text{var}(T) + k^2\ \text{var}(\varepsilon)}$, providing an explanation why the slope is always less than 1 and how the regression effect is influenced by the intensity parameter k.

CONCLUSION

One of the powerful aspects of mathematics is its use of abstraction. In many ways, this fact lies behind the success of mathematical applications and modeling. One of the prime goals of mathematics education in a technological era is to teach how to apply mathematics, i.e. how to translate information into usable knowledge with the help of mathematics. The activity-based approach to statistics provides an excellent pedagogical concept pursuing these goals.

The four-step procedure presented above has been used in a ActivStats class for third-year students at a teachers college. In prior math classes the students had been exposed to the SCHEME programming language (which belongs to the LISP family), and took a course on introductory probability. Compared to statistics classes of earlier years that had almost no hand-on activities and computer simulations, students learned with a lot more enthusiasm as measured by the amount of time they were willing to invest in preparing presentations and getting involved in discussions.

REFERENCES

Levin, J. (1993). An improved modification of a regression-towards-the-mean demonstration. *The American Statistician*, *47,* 24-26.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston: VA.

Rossman, A. (1997): *Workshop statistics. Discovery with data*. Springer: New York.

Scheaffer, R., Gnanadesikan, M., Watkins, A., & Witmer, J. (1997). *Activity-based statistics.* New York: Springer.

Tierney, L. (1990). LISP-STAT: *An object-oriented environment for statistical computing and dynamic graphics*. New York: Wiley.