TEACHING STATISTICAL CONCEPTS WITH SIMULATED DATA

Andrej Blejec
National Institute of Biology
Slovenia

Different kinds of data are used in teaching statistics. In applied statistics courses we usually use real life data related to the main subject matter of our students. Such data are interesting for students and motivate final interpretation of statistical results. For demonstration of statistical concepts, computer simulated data with known statistical properties can be used. The advantage of such data is that results of analysis can be compared with known and pre-defined properties of data. Many important statistical concepts and procedures can be obviously shown with computer simulations and dynamic graphics. Such simulations can sometimes be more convincing than proofs and are appreciated by students.

INTRODUCTION

One of the goals of statistics teaching is to show students how to apply statistical methods. We try to attract their attention by application of statistics to real life problems or problems from their specific field of studies. Such examples, usually connected to a story that describes the problem, motivate students to interpret statistical results according to the problem context (Fillebrown, 1994). While such method gives students a possibility to see and get familiar with what we call *"statistical thinking"*, it is sometimes difficult to see the analytical potential and limitations of the applied statistical method.

To understand and interpret statistical results one has to adopt many statistical concepts. Some are as simple as the central tendency or variation of natural and social phenomena. Some are less obvious and are often described and presented in a way that needs some mathematical insight or abstract thinking. In such cases, non-mathematically oriented students feel very uncomfortable and are unable to understand the meaning of such concepts. Generations of students have problems with understanding important concepts as, for example, confidence interval, standard error or true meaning of p-values.

For correct interpretation of, let's say confidence intervals, one has to understand its meaning. Without that, reporting the confidence interval for the mean is merely the calculation drill or even just another mouse click in a statistical package, despite the interesting project in which it is applied. Sometimes, the real life project data are too complex and one cannot say if the real interrelations are described (Mackisack, 1994). For better understanding, the meaning of certain concepts and methods can be demonstrated and presented by the use of simulated data with known statistical properties. In such cases, one can say whether the tested method can reveal real property or data relation.

REAL, INVENTED AND SIMULATED DATA

Though the ultimate goal of statistical investigation is making decisions in context sphere, it is not necessary that complete learning is performed only by context related problems. Though the use of *real data* is attractive and motivates students, the problems are often too complex in structure and sometimes require deep knowledge of subject matter if we wish to make reasonable interpretation. Since the real data structure is unknown, we cannot be sure if the applied method revealed it. To avoid the problem of complexity, problems are simplified, sometimes even oversimplified. They become in a sense similar to *invented data*, sets of raw numbers used just to practice statistical calculations. They have no background story and the results are just numbers whose correctness can be checked on answer pages in the textbook.

To demonstrate statistical properties and concepts we can use *simulated data* with known statistical properties. They are samples from distributions with known type and parameters, for example normal with known mean μ and variance σ^2. With such data one can see whether the applied method (e.g. arithmetic mean of a sample) can reveal the imposed property (e.g. true population mean μ). Or, for demonstration of properties and power of linear regression, we can construct variables with known linear relationship: samples of normally distributed variable

$X \sim N(\mu, \sigma_X^2)$ and error term $\varepsilon \sim N(0, \sigma^2)$, with known variation of X and ε, combined into variable Y by linear relation $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 and β_1 are known model constants. Such data are usually generated by computer, using the random number generators present in every computing program or programming language. Computers are essential for demonstration of statistical concepts via re-sampling *i.e.*, generation of large number of samples with the same predefined statistical properties and comparison of statistical results on such sets of samples.

RESAMPLING

One of the basic methods for computer-supported demonstration of statistical concepts is resampling (Good, 2001). Sample after sample is taken from the population with known parameters. To each sample, the considered statistical procedure is applied and the distribution of results is inspected. A typical example of resampling procedure is demonstration of sampling distribution of a mean, standard error, and confidence intervals. Results can be presented by dynamic computer graphics in attractive and obvious way. Since "seeing is believing", students can get the feeling for such concepts as central limit theorem, influence of the sample size on sampling distribution shape and variability of estimates. Because the true mean value is known, students can check how many confidence intervals include the true mean and get the insight into the real meaning of confidence interval and confidence level.

Using the same procedure for estimation of variance, one can demonstrate properties of "divide by n-1" rule, which confuses many students in elementary statistics courses. Plot of the estimates of biased estimator (divisor n) and their sampling distribution for small samples (Figure 1a) shows the skewed distribution with expected value, which is smaller than the true value. The bias disappears if the unbiased estimator (divisor n-1) is used (Figure 1b). The sampling distribution is still skewed to the right, which means that, for the variance σ^2 one should not construct symmetric confidence intervals (based on normal distribution properties) but rather asymmetric ones, based on the χ^2 distribution.

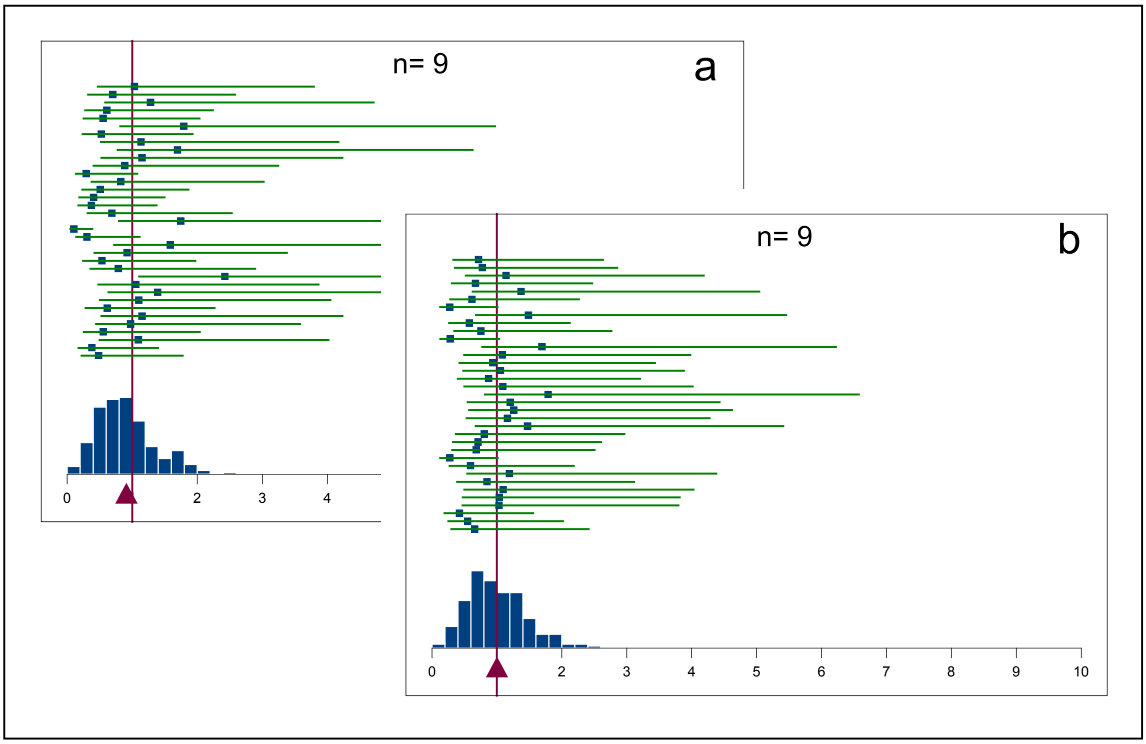

Figure 1. Empirical sampling distribution (histograms, 200 samples) for biased (a) and unbiased (b) variance estimators. Dots: estimates of variance (sample size n=9), horizontal lines: confidence intervals, vertical line: true variance value σ^2, triangle: sampling distribution (histogram) mean value. Note that the mean value (triangle) in (a) is smaller than the true value.

MAXIMUM LIKELIHOOD ESTIMATION

Many students have difficulties in understanding of maximum likelihood estimation. Using the interactive dynamic computer graphics it can be shown that it is essentially an educated guessing procedure. For that purpose, first a sample from known population is taken and plotted as shown by tick marks on upper panel of Figure 2. Students can be asked to guess what the mean value would be. Next, using the mouse or other pointing device, the proposed parameter (e.g. mean) value is selected. The individual data likelihood, according to the proposed distribution, are plotted as vertical segments (Figure 2, upper panel) showing the sketch of the proposed distribution. The log likelihood for proposed value is plotted in the lower panel of Figure 2. After selection of some values and inspection of different situations, the shape of log likelihood function leads to an observation, that the best proposal is at the maximum of log likelihood function. The situation with the best estimate for given sample and true distribution curve is plotted for comparison and observation of the lack of fit.

In a similar way, the estimation of standard deviation can be illustrated (Figure 3, the same data as in Figure 2). This time the proposed parameter values change the spread of distribution. In the same manner as in previous example, we can observe that some guesses, next to the maximum of a log likelihood function in lower panel of Figure 3, make more sense than those far away. The least squares estimation can be illustrated in similar fashion, making clear the concepts as deviation from the mean and minimum sum of squared deviations principle.

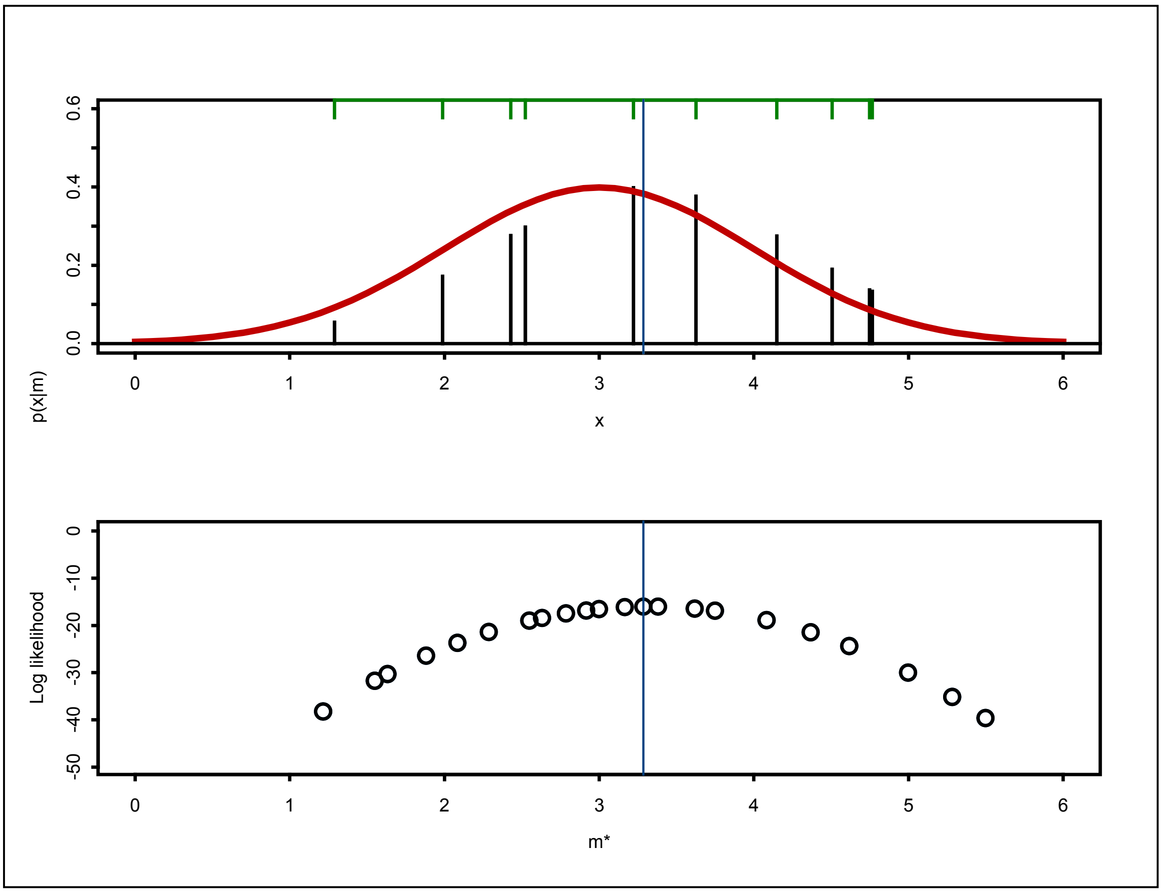

Figure 2. Maximum likelihood estimation of the mean. Log likelihood for some proposed values for μ (circles) with the best estimate for given sample (vertical line) are plotted in lower panel. Individual data are presented as ticks (upper panel). Vertical segments are individual data likelihood for the best estimate. Curve represents parent distribution (μ = 3) from which sample (n=10) was taken.

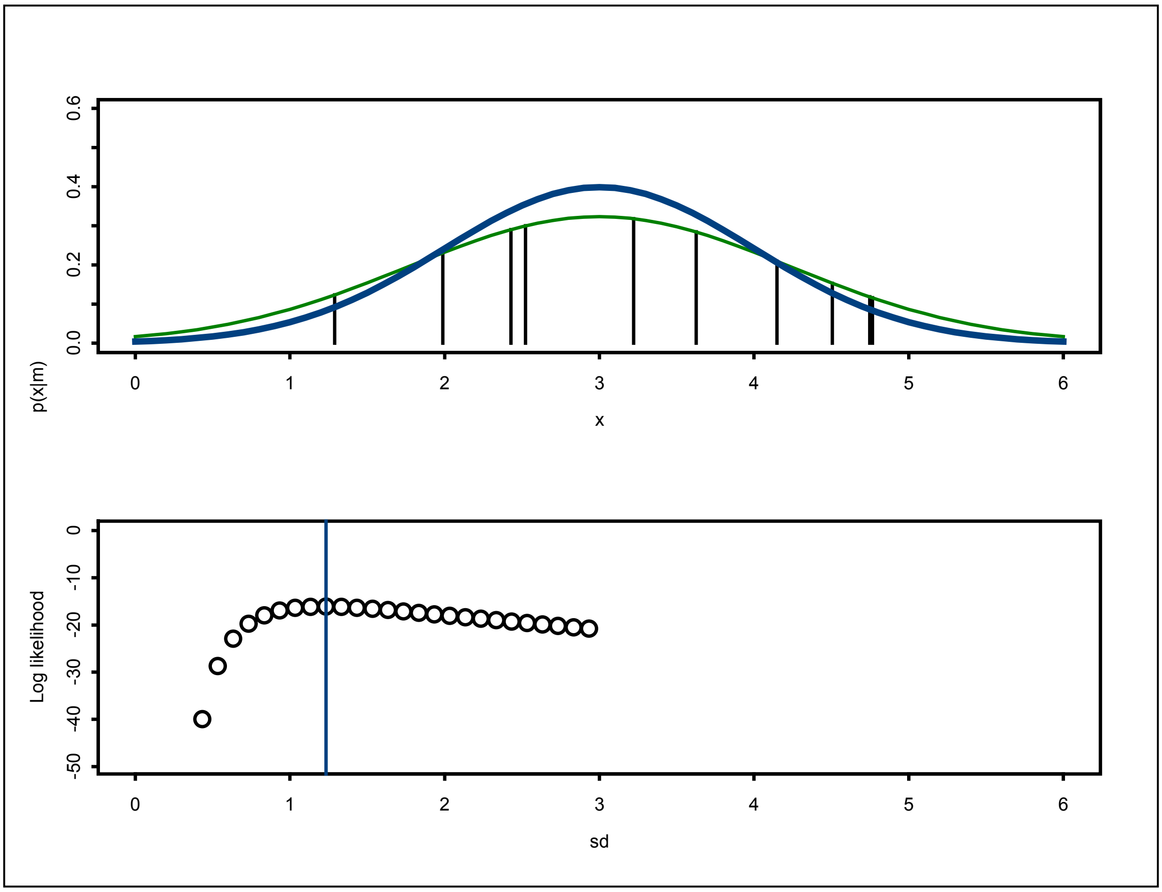

Figure 3. Maximum likelihood estimation of the standard deviation (σ=1). Upper panel: modeled distribution curve(thick line), estimated distribution curve (thin line) with vertical segments at individual data. Lower panel: log likelihood for some proposed values (dots) and best estimate (vertical line)for given sample.

LINEAR REGRESSION

For demonstration of linear regression properties, data based on linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ are simulated. All parameters of linear model are known: β_0 and β_1 are selected model constants, X and ε are normally distributed $X \sim N(\mu, \sigma_X^{\ 2})$ and $\varepsilon \sim N(0, \sigma^2)$ with selected parameters. With some resampling it can be shown, that the distribution of Y is normal with mean value $\mu_Y = \beta_0 + \beta_1 \mu$ and variance $\sigma^2_Y = \beta_1^{\ 2} \sigma_X^{\ 2} + \sigma^2$. To get the feeling for closeness of sample picture and model picture, we can generate series of samples and plot the data and regression lines. Students can get the notion of the influence of error term variation and how it is connected to the coefficient of determination $r^2 = 1 - \sigma^2/\sigma^2_Y$. A series of regression lines and model line are plotted in Figure 4. The regression lines differ from the model line due to the variation of the error term. Students easily notice that the lines are embedded in the curved regression line prediction band around the model line (Figure 4, right panel), which can be transformed into confidence band for particular regression situation (Figure 4, left panel). Looking at the results of regression for many simulated samples and comparing the estimates of model parameters and coefficient of determination (in example from Figure 4: 0.0415, 1.2400, 0.567) to the model parameters (0, 1, 0.5) students can learn to what extend the method can show the true data structure. Getting familiar with the power of the method on simple and pure simulated data the students are prepared for inspection of real life data in which they will be able to interpret lack of fit or understand the meaning of confidence band.

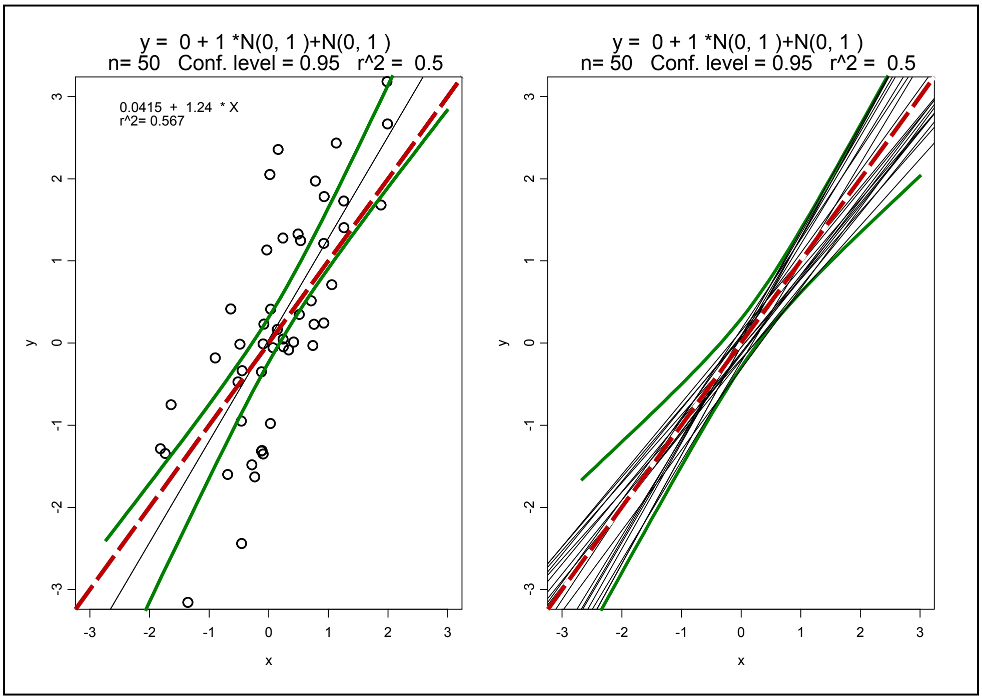

Figure 4. Linear regression confidence band. Left panel: regression line (thin line), based on a sample (dots) with n=50 with model line (dashed) for modeled data (on top) incorporated in 95% confidence band (thick curves). Right panel: regression lines from 15 different samples (thin lines) with 95% prediction band (thick curves) - essentially the envelope of the set of regression lines.

DISCUSSION

Computer simulated data have many advantages in teaching statistical concepts. Their statistical properties are known and one can see the connection of data properties and results of analysis. It is easy to change the properties and observe the influence of such changes for the analysis. Many users of statistics feel uncomfortable to select appropriate statistical method since they are not sure if necessary assumptions for specific method are met by their data. It is easy to simulate the data not meeting the assumptions (for example non-constant variance of error term) and show possible fallacy of results.

Graphically supported simulations can - to some extent - replace proofs, usually not understandable for non-mathematics majors. Maybe they can answer Moore's question: "*If an audience is not convinced by proof, why do proof?* (Moore, 1996). Simulated data have to be combined with real life data and projects (Mooney, 1995). They serve as pure and simple data on which we can train our perception for statistical results and learn what patterns and properties in data can be revealed by applied method. After such preparation students will be able to interpret the real life and subject matter data in all their complexity.

REFERENCES

Fillebraun, S. (1994). Using projects in an elementary statistics course for non-science majors, *Journal of Statistics Education*, *2*(2).

Good, P. (2001). *Resampling methods* (2nd edn). Berlin: Birkhauser.

Mackisack, M. (1994). What is the use of experiments conducted by statistics students? *Journal of Statistics Education*, *2*(1).

Mooney, C. (1995). Conveying truth with the artificial: using simulated data to teach statistics in the social sciences. *SocInfo Journal, 1.*

Moore, S.D. (1996). New pedagogy and new content: The case of statistics. In B. Phillips (Ed.) *Papers on statistical education: ICME-8* (pp. 1-4). Melbourne: Swinburne University of Technology.