# NEW PATHS IN THE TEACHING OF STATISTICS - WITH THE HELP OF SPREADSHEETS

Manfred Borovcnik
University of Klagenfurt, Austria
manfred.borovcnik@uni-klu.ac.at

*Situations, in which data form the basis of decisions, are abundant. The paper illustrates some concepts involved like "the correlation coefficient" and how it measures the degree of connections between several variables, or "remaining risk" and how it is possible to draw general statements from restricted data. To embed such notions in concrete manipulations of data and easily accessible diagrams facilitates understanding of statistics. The ideas may be worked out with the help of any spreadsheet, here EXCEL is used.*

## INTRODUCTION

Inferential statistics combines several types of knowledge, including a sound conception of remaining risk, probability, and some mathematical optimization procedures. Various approaches to facilitate understanding will be given below. In the first section an easy interpretation of the correlation coefficient will be disclosed by simple calculations with lists of data. In the second section sampling under well-defined conditions will be used to investigate useful patterns of randomness. While single observations still are 'irregular', a general pattern emerges, allowing for generalizations from data which hold with the exception of a remaining risk; a concept which will be illustrated herewith. In the third section resampling methods will be used to explore the size of sampling errors. This allows deriving confidence intervals or statistical tests from observed data to judge questions like "are there differences between two groups of patients, one getting a medical drug as treatment and one placebo?", or "is an observed correlation between variables due to mere fluctuation or due to true interconnections?"

## MODELLING

One of the statistician's tasks is to investigate, which variables influence substantially the value of a target variable. E.g., which variables influence the actual body weight of adult persons? During the phase of systems analysis, one searches for explaining variables like height, gender, body type, nutrition as a baby etc. As a resulting model – a simple example – one could try the following linear equation for the structure of relations between the variables

$$\text{weight} = a + b \times \text{body height}$$

It may be checked by data whether this equation yields a useful description, i.e. if it allows predictions for the weight if one knows the person's height. The influence of height is regarded as established if the coefficient b from the equation 'differs significantly from 0,' which is equivalent to 'the correlation coefficient between the two variables differs significantly from 0.' In either case, the question whether a variable has to be taken into a model equation or not is reduced to the technical question of some coefficient to *differ significantly from 0*.

The actual data will be superimposed by 'noise' and therefore deviate from the model equation for several reasons. It is the statistician's scope to separate the intrinsic signal from the noise in the data. By suitable methods it will be possible to "suppress" the noise in order to re-establish that pattern of the data, which is due to an influence that may be generalized. This general feature of data will be summarized by a model: data = model + residual

For some other splits of data and their educational advantage, see Borovcnik (2005). While the model and its implications are open to an explanation from the context, the residuals are not yet explainable. A model should be fitted to the data to make the residuals as small as possible, usually the least squares principle is applied but there are other criteria as well.

With a spreadsheet like EXCEL it is easy to plot the scattergram for the data with (not only) a linear trend function and to calculate $R^2$ – the square of the correlation coefficient. Hereby it is not necessary for the students to know anything about R, the criteria for drawing the trend line however may be easily explained without technical details. The following model equation for weight has been derived from an actual set of data for 15year olds

weight = 0.697 × height – 67.299,

This equation implies that a person with height = 170 cm should weigh ideally 0.697×170 – 67.299 = 51.2 kg. The following simple calculations establish an easy but important characteristic for the correlation coefficient: The model equation is simply applied to calculate the ideal weight values according to this model (the points on the trend line) and the resulting residuals (deviations of the data points from the trend line). Usual statistical measures of the columns of data, model data, and residuals are calculated, the results evince:

mean of residuals = 0
variance of data = variance of model data + variance of residuals
variance of data × $R^2$ = variance of model data
variance of data × (1-$R^2$) = variance of residuals

Remarkably, the equation for *single data* – data = model + residual also holds for the *variances*. Furthermore, the reduction of variance in the residuals, which is that part of data, which is still not explained by the model, equals 1-$R^2$. Thus, the bigger the correlation R, the smaller is the remaining unexplained variance of the target variable.

Predicting the *target* weight without knowledge of height is done by the 2 sigma rule:

mean weight ± 2×SD of weight

With the regression model, the prediction improves to model weight ± 2×SD of weight × √(1-$R^2$). The prediction intervals are cut by a factor, which is a function of the correlation coefficient. According to that, an R of 0.83 yields a shortening of prediction intervals by 0.56, i.e. by 44%. This establishes an important feature of the correlation coefficient, which is open to a clear-cut interpretation.

## SAMPLING

The calculus of probability has become obsolete by the growing simulation capacity of our computers. By simulating data on the basis of the assumptions of a probability model, any desired probability may be approximated by the relative frequencies of an event in question. For example, it is possible to evaluate a strategy for a game of pure chance by imitating the conditions of that game by suitably chosen random numbers. Or, it is possible to explore the question how and how much a correlation coefficient varies in a sample of fixed size if – on the basis of the chosen model the correlation coefficient *is* zero. This yields *limiting values* for the correlation in data; if these limits are exceeded in observed data, the decision is that the correlation is significant (significantly different from zero). Herewith associated is the concept of ‘remaining risk’ as the decision could be wrong. Even if the correlation *fluctuates* and could take any value between -1 and 1 in single data sets following the assumptions of the model, it makes sense to derive a *general pattern* from the sampling study: if the correlation truly is 0 then normally it fluctuates between -0.44 and 0.44 (for 20 data points), at least in 95% of repeated samples. Analogously one could be interested in such limiting values for a correlation that is known to be 0.875, for the limits of ‘normal’-behaving samples see Figure 1a and b.

Thus, if a dataset of 20 pairs has a correlation of e.g., 0.47, one could well state that the correlation between the variables in question is significant at the 5% level of remaining risk. Repeating the whole scenario of sampling may be easily done in EXCEL by pressing the function button F9. This will give a clear vision of the accuracy of the empirically measured variability of the correlation coefficient. For an implementation, see Borovcnik (2005) with the EXCEL files on the website of *Stochastik in der Schule*, the German sister journal of *Teaching Statistics*.

An analogous sampling study on this website explores the sampling error of the estimation of an unknown proportion (or a probability) by a series of Bernoulli trials with the proportion of ‘successes’ in a series of *n* trials. There are an unknown number of approaches to that central topic in elementary inferential statistics, starting with Kissane (1981). The resulting pattern is known as the “1 over square root of n – law” as the critical levels for the *normal*-behaving samples “converge” with the rate of 1/√*n* towards the “true” but unknown probability.

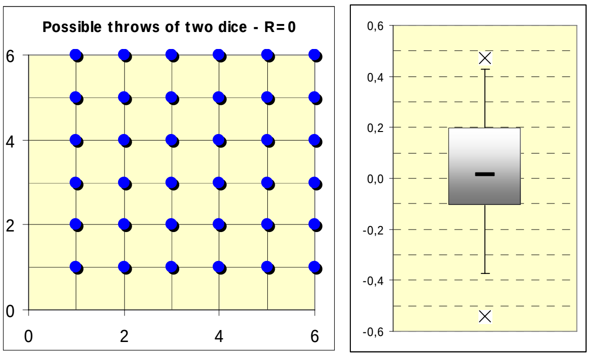


FIGURE 1a: Repeating a sample with 20 data if the 'true' R=0: usual random fluctuation of the correlation coefficient R lies between + or – 0.4 roughly. If one takes this statement as a prediction for the next series of 20, the remaining risk amounts to 5%.

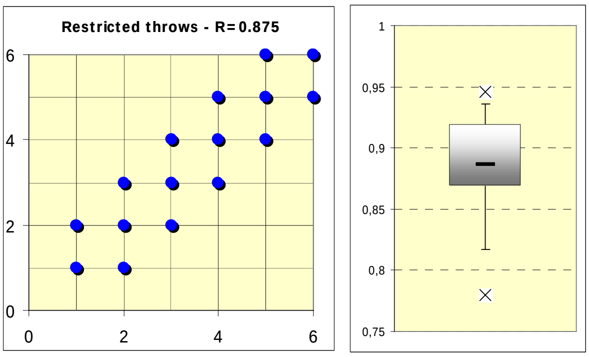


FIGURE 1b: Usual random fluctuation of R – now with true R=0.875 – implemented by restricted throws of two dice.

EXCEL allows simulating and investigating the data efficiently. There is room for the discussion of remaining risk as repeating the scenario will sometimes exceed the once calculated limits.

## RESAMPLING

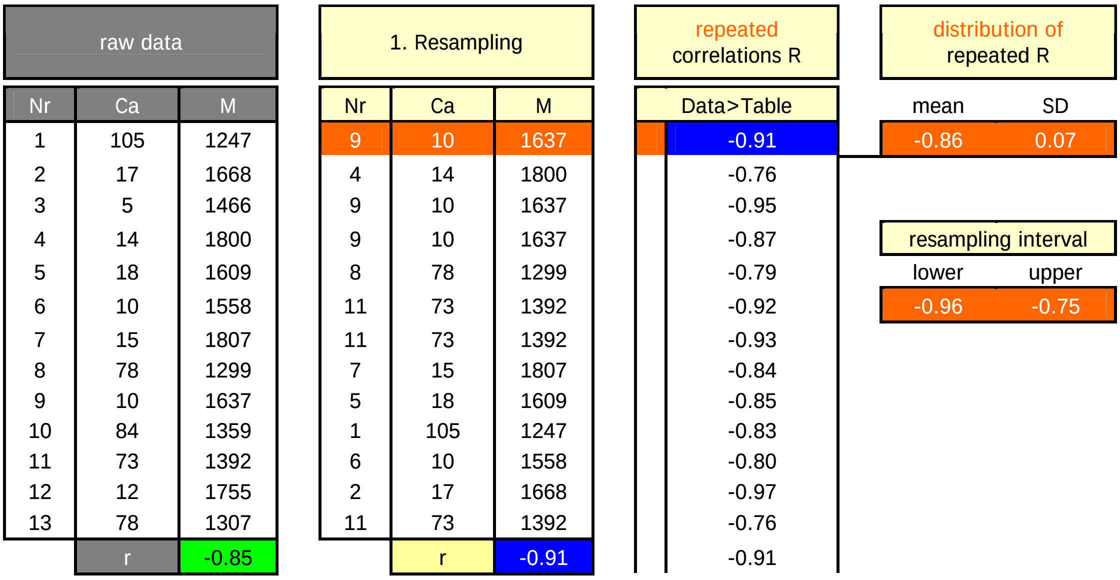
Resampling methods allow for investigating the random error of a statistic. Random error signifies the fluctuation of an investigated statistic (difference of means, R, etc.), which helps to judge a problem from the context. If, for example, data are related to the effect of a medical treatment – one group got the medication, the other a placebo (like the drug but without medically effective substances) – then the observed difference in means (time to heal, proportion of patients with complete cure, etc) allows judging the question whether the medical drug really is effective.

If one repeats the whole sample, i.e., get new patients and treat them alike, one gets further data about the effectiveness of the medical drug. The situation then is the same as in the sampling scenario of the previous section. It is possible to 'measure' the size of fluctuation due to randomness. Eliminating that, there remains the effect of treatment. However, usually it is not possible to replicate the whole study; the information has to be drawn from the original data. This is the case for methods of inferential statistics. The process of data 'generation' is modelled by distributions etc; from there the statistician derives the required inferences.

Re-sampling serves the purpose to replace this theoretical part and derive virtual data for the investigated statistics. This is achieved by repeatedly sampling a new sub sample *from the original data* and calculating the actual values of the statistic (e.g. the difference in treatment effect). The original sample yields an estimation of the theoretical distribution (function). Instead of sampling out of the *true* distribution, sampling is done out of an *approximation of it*.

How to draw a sample in EXCEL, especially one from the numbers 1 to 13 (in Table 1) is explained in Christie (2004). The first resampling gives a value of –0.91 for R, a first clue for the precision of R = –0.85 from the original data. By pressing F9 the sample is renewed easily, the first R of –0.91 changes to –0.76. Simultaneously this reveals the good precision of the original sample, as the fluctuation is not big. By repeatedly pressing F9 and storing the resampled values of the statistic one generates a data base to explore the precision of the estimation of the statistic. EXCEL offers an efficient short-cut to that by the Data>Table function; for this nice trick, see Christie (2004), or the *Stochastik in der Schule* website. The generated values of the statistic (of R) may then be analyzed by simple descriptive methods: 95% percent of resampled values lie between the 2.5% and the 97.5% quantile, which yields the resampling interval for the investigated statistic. This interval may be checked for accuracy by repeating the whole resampling by F9. Resampling works for any statistic from the original sample, e.g., for the difference of means between two treatment groups, or, here for R. It is possible to derive confidence intervals, or to test specific values for the statistic, e.g., to test whether the observed value of R is significantly different from 0, or not – in Table 1 it *is* significant (0 is not in the derived interval).

TABLE 1: Spreadsheet for the resampling study of a correlation from Christie (2004)



| raw data | | |
|---|---|---|
| Nr | Ca | M |
| 1 | 105 | 1247 |
| 2 | 17 | 1668 |
| 3 | 5 | 1466 |
| 4 | 14 | 1800 |
| 5 | 18 | 1609 |
| 6 | 10 | 1558 |
| 7 | 15 | 1807 |
| 8 | 78 | 1299 |
| 9 | 10 | 1637 |
| 10 | 84 | 1359 |
| 11 | 73 | 1392 |
| 12 | 12 | 1755 |
| 13 | 78 | 1307 |
| | r | -0.85 |

| 1. Resampling | | |
|---|---|---|
| Nr | Ca | M |
| 9 | 10 | 1637 |
| 4 | 14 | 1800 |
| 9 | 10 | 1637 |
| 9 | 10 | 1637 |
| 8 | 78 | 1299 |
| 11 | 73 | 1392 |
| 11 | 73 | 1392 |
| 7 | 15 | 1807 |
| 5 | 18 | 1609 |
| 1 | 105 | 1247 |
| 6 | 10 | 1558 |
| 2 | 17 | 1668 |
| 11 | 73 | 1392 |
| | r | -0.91 |

| repeated correlations R |
|---|
| Data>Table |
| -0.91 |
| -0.76 |
| -0.95 |
| -0.87 |
| -0.79 |
| -0.92 |
| -0.93 |
| -0.84 |
| -0.85 |
| -0.83 |
| -0.80 |
| -0.97 |
| -0.76 |
| -0.91 |

| distribution of repeated R | |
|---|---|
| mean | SD |
| -0.86 | 0.07 |

| resampling interval | |
|---|---|
| lower | upper |
| -0.96 | -0.75 |

## DISCUSSION

The author has used resampling in university courses for mathematics and business administration students. The students learned procedures like tests or confidence intervals, be it for the difference of means or the correlation coefficient. Mathematics was no more the exclusive key for understanding. Likewise the inductive logic, which was backed by hands-on activities, i.e., playing the scenarios of the conditions of the model and exploring its consequences. The idea of re-sampling another sample not from anew but from the original sample was well accepted.

In fact, the argument that the original sample may be biased and this ruins the resampling study as it is repeated, is also valid for the classical approach: biased samples always destroy the validity of conclusions that may be drawn from them. Resampling methods are more and more accepted by theoretical and practical statisticians as well. Statisticians implement required simulations by special languages like R, for teaching EXCEL is sufficient. A final remark on EXCEL – or any other spreadsheet: Not all questions are easy to transfer to a re-sampling-design. Some time is lost with trivial formatting etc. Working style is explorative and hands-on. The consequences of a model are transparent by playing ‚it'. The procedures of sampling and resampling are much easier than traditional concepts and let the learner witness what is going on.

## REFERENCES

Borovcnik, M. (2005). *Probabilistic and Statistical thinking*. CERME 4 – Working group on Stochastic thinking, http://cerme4.crm.es/Papers%20definitius/5/wg5listofpapers.htm.

Borovcnik, M. (2005). EXCEL-Files für den Unterricht in Stochastik, http://www.uni-klu.ac.at/stochastik.schule.

Borovcnik, M. and Peird, R. (1996). Probability. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick and C. Laborde (Eds.), *International Handbook of Mathematics Education*, (pp. 239-288). Dordrecht: Kluwer.

Christie, D. (2004). Resampling with EXCEL. *Teaching Statistics*, 6, 9-14.

Kissane, B. V. (1981). Activities in Inferential Statistics. In A. P. Shulte, and J. R. Smart (Eds.), *Teaching Statistics and Probability, 1981 Yearbook of NCTM*, (pp. 182-193). Reston, VA: NCTM.

Neuwirth, E. and Arganbright, D. (2004). *The Active Modeler: Mathematical Modeling with Microsoft EXCEL*. Belmont, CA: Brooks/Cole.