University of Novi Sad
Center for Applied Statistics

Guide in Methodology of Teaching Applied Statistics

Selected Topics in Methodology of Teaching Applied Statistics

The publishing of this booklet is a part of the Tempus project “Master programme in applied statistics” MAS 511140-Tempus-1-2010-1-RS-Tempus-JPCR

“This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein."

CONTENT

TEACHING STATISTICAL CONCEPTS WITH SIMULATED DATA

Andrej Blejec
University of Ljubljana, 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.

1. 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.

2. 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 can not 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.

3. 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.

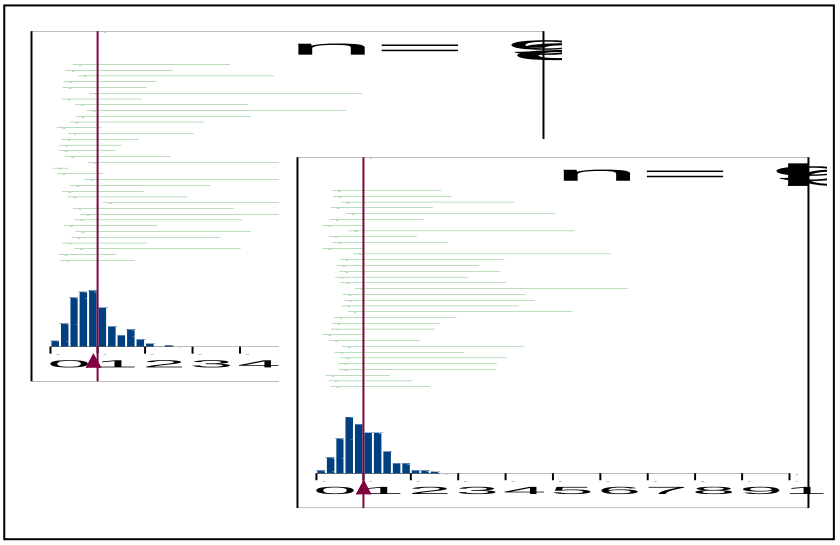

FIGURE 1. - EMPIRICAL SAMPLING DISTRIBUTION

4. 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.

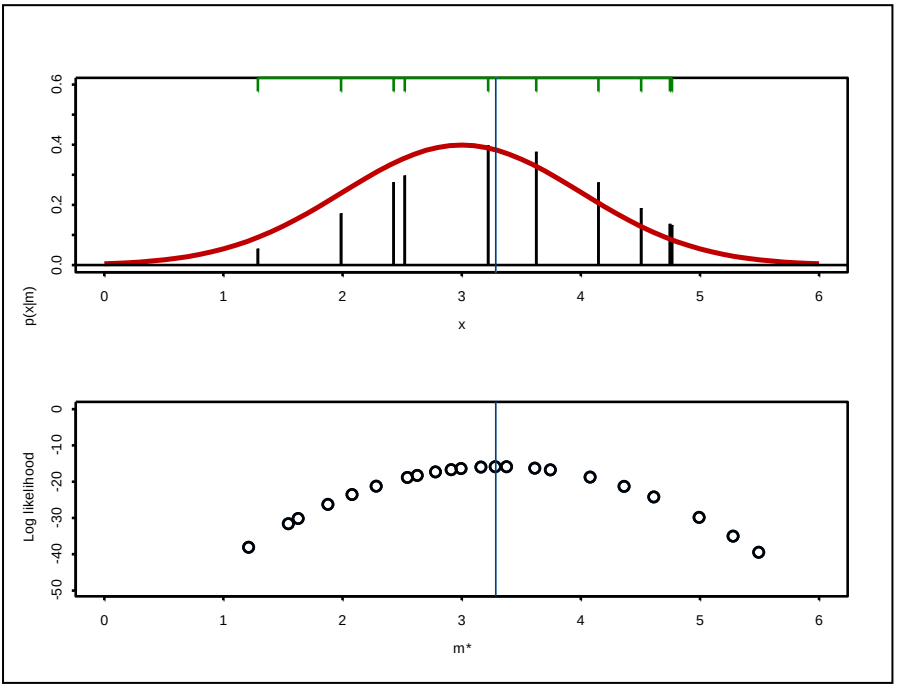

FIGURE 2. - MAXIMUM LIKELIHOOD ESTIMATION OF THE MEAN

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 ($\mu = 3$) from which sample (n=10) was taken.

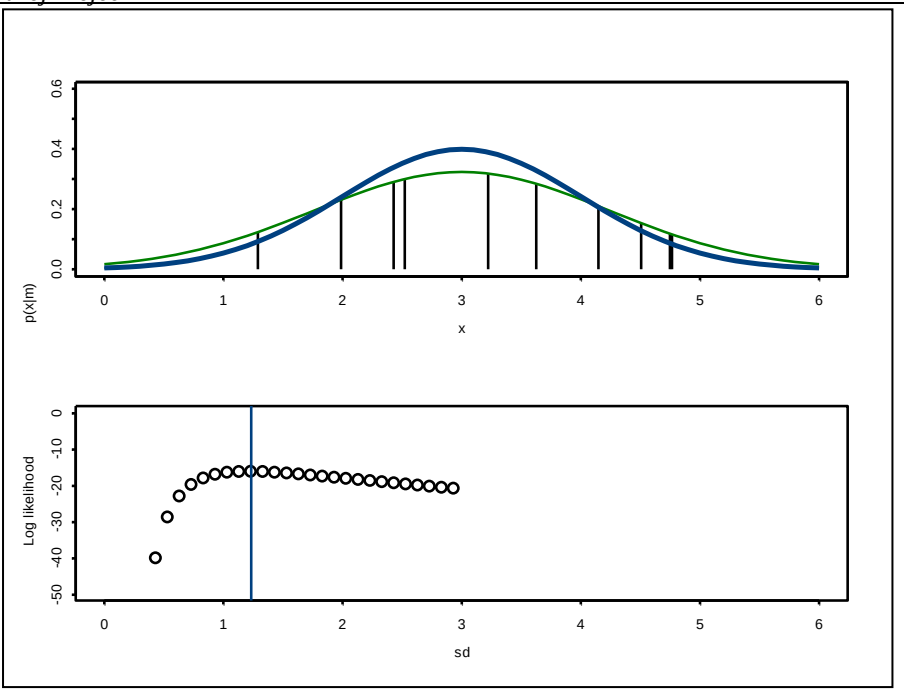

FIGURE 3. - MAXIMUM LIKELIHOOD ESTIMATION OF THE STANDARD DEVIATION

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.

5. 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.

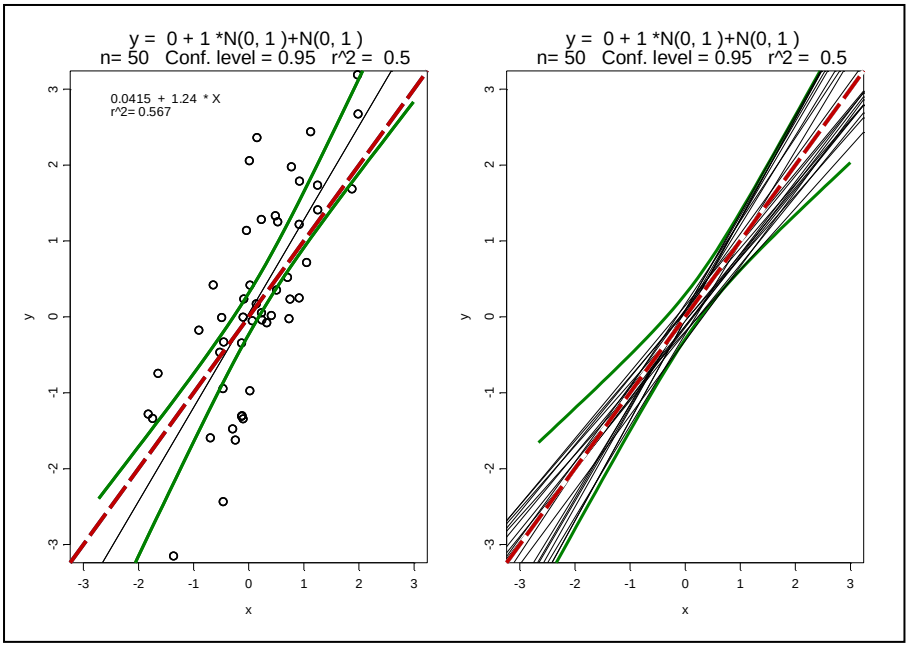

FIGURE 4. - LINEAR REGRESSION CONFIDENCE BAND

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.

6. 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.

7. REFERENCES

1. Fillebraun, S. (1994). Using projects in an elementary statistics course for non-science majors, *Journal of Statistics Education*, v.2, n.2.
2. Good P. (2001). *Resampling methods*, 2nd ed Berlin: Birkhauser.
3. Mackisack M (1994). What is the use of experiments conducted by statistics students?. *Journal of Statistics Education*, v.2, n.1
4. Mooney C. (1995). Conveying truth with the artificial: using simulated data to teach statistics in the social sciences, *SocInfo Journal 1*.
5. Moore S.D. (1996). New Pedagogy and New Content: The Case of Statistics. In: Phillips B (Ed.) *Papers on Statistical Education. ICME-8*, Ed. B. Phillips, pp 1-4. Swinburne: Swinburne University of Technology.

CONSTRUCTING MULTIPLE CHOICE TESTS TO CHECK BASIC KNOWLEDGE OF STATISTICS

Vesna Jevremović
Mathematical faculty, Belgrade

In the course Probability and Statistics for students of III year of the IT module, one of the students' duties were the preparation of multiple choice tests that would serve for checking the knowledge of the course material.

The idea was that through a detailed preparation of such a test, a student him/herself learn in detail at least this lesson. Therefore, the task was: to formulate 20 questions on selected lessons, for every question to prepare three answers, such that they are not obviously incorrect (e.g., the probability that would be negative and the like). They would make a computer program in which the test is implemented. At each start of the program, 10 out of 20 available questions would be randomly chosen, and offered answers would be randomly permuted. When the user answers all the questions, he/she would get a result (how many questions are answered correctly) and the opportunity to know and learn the correct answers to the questions that are not answered correctly. After the restart, the test would provide another 10 questions. In this way, the teacher has an opportunity to check all the students with the same test, given that there are $\binom{20}{10} = 184756$ various choices available, and when we consider the permutations of response, we have really more than enough tests for individual testing!

Students made open computer programs, so that the tests could be changed as needed and / or expanded. Overall results were more than satisfactory - tests have fulfilled their function, in the sense that their authors knew very well "their" lessons, it was inspiring, and are well prepared. Student prepared very interesting graphic designs, they used

their knowledge of software development and create an elegant and / or witty graphic designs.

Several examples of the tests can be viewed at the addresses below:

1. Student Martin Hofer worked on problems related to the solving the problems in connection to point estimation of population parameters. http://www.alas.matf.bg.ac.rs/~mi08005/test/
2. Student Bojana Kovacevic worked on the basic properties related to some well-known distributions - binomial, uniform, normal,...
   http://alas.matf.bg.ac.rs/~mi08131/vis/
3. Student Katarina Radović worked on the basic statistical concepts - population, marker, pattern...
   http://alas.matf.bg.ac.rs/~mi08222/vis/vis.php
4. Student Vladan Stanković worked on the problems related to parametric hypotheses testing in normal distribution.
   http://vaxter.neospindle.com/vis/

CENSUS OR SAMPLING?

Sanja Rapajić
Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad

Population stands for the entire collection of objects we want to study. If it is small enough, we can study it in its entirety. If, on the other hand, the population consists of a large number of elements, then studying it can be expensive, time-consuming, sometimes even destructive, and even more, impossible in principle. Exactly this is the reason for choosing the subset of population, called sample, which becomes an object of investigation. The idea is to try to draw the conclusion about the whole population based only on the analysis of the selected sample.

Sampling theory is a field of statistics which studies the sample selection and deals with estimating the relevant parameters of the whole population. There are many techniques and ways of sampling and, therefore, many various types of samples.

It should be noted that even when there is a possibility of testing the population as a whole, a researcher usually chooses only a sample. Firstly, because it is cheaper than testing the entire population, then because the control of data accuracy of a collection is simpler and easier on a sample than on the whole population, and lastly, because the feedback and results come much faster from the sample than from the population.

To make a relevant statistical inference about the general population based on a sample, it is essential to make the sample representative. A perfect sample is a scale-down version of the population, mirroring every characteristic of the whole population. Of course, no such perfect sample exists for complicated populations, and even if it does, we would not know it was a perfect one without measuring the whole population. For this reason, it is necessary to choose a sample which will display the characteristics of the whole population as

accurately as possible, i.e. which will be representative in a way that each sampled unit shows the characteristics of a known number of units in the population.

Errors which appear in any research study may be divided into two categories: sampling and non-sampling errors. Understanding the difference between them is essential when choosing between conducting a census and sampling, as well as when selecting the appropriate procedures in case we choose the sampling. It is necessary to minimize all types of errors in order to get as reliable as possible results.

The sampling error is the result of using a sample rather than investigating the whole population. It stands for the difference between a sample estimate and the true population value which is obtained from the population in total. Sampling error varies from sample to sample. If probability sampling is used, sampling error is the margin of error of the estimates. Sampling error may be minimized by selecting a large sample size and by implementing a stratified sample design.

In many surveys, a sampling error reported for the survey may be negligible compared to non-sampling errors. Sampling error affects the precision of estimates, while non-sampling errors affect the validity of estimates.

Non-sampling errors cannot be attributed to the sample to sample variability. These errors are due to the sampling procedures that produce estimates that systematically differ from the actual characteristics of the population. Non-sampling errors are research biased. They are classified according to the stages of the research process. There are three main categories: bias in selecting study elements, bias in collecting data from the selected elements and bias in analyzing the data collected.

Bias in selecting study elements consists of population specification bias, coverage bias and selection bias.

Population specification bias occurs due to ambiguity in defining a research problem or a poor definition of the target population. It may be minimized by clearly defining the target population and by a well-designed questionnaire with understandable questions.

The most common types of coverage bias are: under-coverage bias, over-coverage bias and multiple-coverage bias. Under-coverage is failing to include all of the target population in the sampling frame. Over-

coverage occurs when elements that are not members of the target population are listed in the sampling frame, while multiple-coverage occurs when elements of the population are listed more than once in the sampling frame. These kinds of biases may be minimized by thoroughly reviewing the sampling frame and target population.

Selection bias occurs when the chosen elements within a sample systematically differ from the non-selected elements. Example of this is a deliberate selection of a sample and using a sample-selection procedure which is dependant on certain characteristics associated to the investigated property. Conducting a census minimizes the selection bias. It can also be minimized by using probability sampling with equal probabilities, training the researchers and comprehensive quality control procedures.

Bias in data collection usually occurs in the form of non-responsiveness and inaccurate responses. Very often some elements of the population, which should make a sample, are not reachable, are incapable of responding, or unwilling to participate in the survey. Besides, the data collected from the sampled elements may be inaccurate, incomplete or inadequate for many reasons such as: poorly designed questionnaire, researchers' bias, reluctant respondents etc. Therefore, the bias occurs either due to the failure in data collection, or data insufficiency, or their quality. Measurement bias also falls into this group. Measurement bias occurs when the measuring instrument has a tendency to differ from the true value in one direction. Measurement bias is difficult to detect and may be imperceptible and insidious. Bias in data collection may be minimized by statistical adjustments, well-designed questionnaires, well-designed data collection procedures and instruments, and staff training.

Bias in data analysis is a difference between the actual population parameters and sample estimates which are the results of the data analysis procedures applied. It involves bias caused by data processing errors and bias caused by data analysis errors. These errors may be minimized by training the researchers, implementing quality control procedures and appropriate use of statistical tools.

Considering the purpose of the research, sampling and non-sampling errors, costs and other factors, the strengths and weakness of

taking a census and sampling may be compared. Although it seems that a census always has fewer errors than sampling, it may not be true. A census does eliminate sampling error but it may contain a great deal of non-sampling errors.

Conducting a census is recommended in the following situations: heterogeneous population, small-size population, importance of having credible results, detailed data analysis requirements, importance of minimizing sampling error and ethical and legal requirements. Sampling would be a better choice in the case of: large-size population, highly scattered and fragile population, need for a quick decision, need to regularly collect up-to-date information, need for a great deal of in-depth information, importance of minimizing inaccuracy of responses, importance of using easy operational procedures and limited resources (time, money, personnel). In deciding whether to choose census or sample it is necessary to take into account the objectives of the study, its importance, research design considerations, nature of the population, availability of resources, operational and data analysis procedures and ethical and legal requirements.

There are many poorly done surveys nowadays and this creates skepticism towards surveys in general. Some think that sampling is bad and insist that only a complete census, in which every element of the population is measured, has to be conducted. As previously said, for a small-sized population taking a census is not a problem – a complete census eliminates sampling error. However, it cannot exclude the non-sampling errors. By far the most frequent causes of errors in a survey are under-coverage, lack of response and sloppiness in data collection. In general, taking a complete census of a population consumes a great deal of time and money, but does not eliminate all errors. Sometimes, a census may be destructive. Even when we are able to examine the population as a whole, we often decide to select only a small part of it (sample), and make decisions related to the entire population. There are many reasons for that. Sampling can provide reliable information and yet be a lot cheaper than a census. Data can be collected more quickly, so the estimates can be published in a timely fashion. Finally, the estimates based on sample surveys are often more accurate than those based on a census, because of greater attention paid to data quality and personnel training, which results in minimizing errors.

In conclusion, it is by far better to perform quality measurements on a representative sample, than the unreliable or biased ones on the whole population.

THE METHODOLOGY OF DEVELOPING QUESTIONNAIRES USING DELPHI METHOD IN STATISTICS COURSES

Mirko Savić
Faculty of Economics in Subotica
savicmirko@ef.uns.ac.rs

Development of questionnaires and data collection using questionnaires is the segment that is often ignored in a number of basic courses in statistics. With this approach, students are trained to use statistical instruments for mathematical and statistical analysis of data while they do not know how to make a basic measuring instrument in the statistics and to collect data with it. Delphi method is a method of achieving consensus within the group of experts with regard to certain issues and as such can be easily and efficiently used for definition and selection of questions included in a survey. For this reason, it is possible to use this method to demonstrate to students the process of building statistical questionnaires.

Keywords: *Delphi method, questionnaire, course*

1. INTRODUCTION

When it comes to teaching statistics and other quantitative subjects at universities, as a rule, and with few exceptions, there is emphasis on the mathematical and statistical analysis in detail throughout the course. Planning of statistical surveys and, within it, a collection of raw data, remain in the background and are often completely ignored. That leads to a somewhat absurd situation in which students are trained to use statistical methods and analyse data but do not have even the basic knowledge of how to collect the data on which analysis should be performed.

Particularly sensitive point is the development of the questionnaire because the methodology for that issue is not the teaching subject on most courses. This problem is not present only at the universities in Serbia but also in many other countries, which can easily

be concluded by looking at the books by foreign authors. Also, this problem is particularly present on the faculties of economics and at the faculties with related disciplines.

As a consequence, it happens that in certain situations, students (or graduates) who should conduct the research are reluctant to collect data and do not know how to develop a measuring instrument (questionnaire) to observe the phenomenon of interest. It also happens that students are making a questionnaire that passed neither validity test of individual questions nor validity test of the questionnaire as a whole. Instead, the questions are simply thrown on the paper, without any consultation with professional literature, experts in the given field, etc. Data collected in this way cannot be considered relevant for the given object of study. Consequently, the results present a false and biased picture of the observed variables leading to erroneous conclusions and misinformation.

In many scientific fields, especially in social and economic science, one often needs to gather relevant data on observation units directly from the field, from the primary sources of data through questionnaires and interviews. When it comes to economy, for example, companies need to explore the market and collect data on the characteristics and behaviour of potential customers, business partners, competitors, employees and the like. Also, public and state enterprises need to monitor a large number of variables related to population, urban development, economy, transportation, etc.

For these reasons, there is a need to develop a methodology that would be implemented in the teaching of statistics at undergraduate level and at various faculties, and designed to deal with the problem of developing questionnaires. There is a range of different approaches to this subject, however, the emphasis here is on building questionnaires using Delphi method.

It is clear that questionnaire preparation methodology represents a systematic, extensive and elaborate procedure for which there is insufficient time during classes. This is a series of tests, panel studies and pilot studies that are being built in order to test the measuring instrument. Moreover, various disciplines favour different methodologies for developing questionnaires. Due to complexity, it is necessary to select

and implement a methodology for developing questionnaires that is not too time consuming and will give a sufficiently robust result, i.e., the questionnaire with relevant questions that are actually measuring the observed phenomenon. Of course, students must be informed that the creation of questionnaires is a serious and complex process and that they should not think that learning Delphi method automatically means mastering high quality surveys. That goal demands additional tests.

Additional benefit for the students who learn Delphi method is its very wide application in any situation when the goal is to achieve consensus on certain issues.

The subject of this paper is not to present in detail the Delphi method, but to show how Delphi method can be used to produce a good questionnaire and how it is possible to use this very popular method to achieve consensus in preparation of questionnaires. There is abundant literature on Delphi that comes from different scientific disciplines which mention all the disadvantages and advantages of this method. This paper shall present only a general discussion on Delphi method. Therefore, someone who is more interested in the topic should address several sources that are located in the references.

Finally, instead of presenting a complete statistical methodology for carrying out research, this paper shall focus on the development of the measuring instrument - the questionnaire, as one of its most sensitive phases. Lecturer on the course in statistics may decide whether to introduce this methodology separately or integrated into a broader presentation of the whole statistical research with all phases, from planning and research, through data collection, statistical analysis and evaluation research. Much depends on the time available and the volume of other materials that were intended to cover the course.

2. DELPHI METHOD

Delphi method is a method for achieving consensus on certain issues and as such can be used in a number of specific situations. When creating a questionnaire there is often a dilemma which questions should be included in a questionnaire and which not. It is necessary, therefore, to be sure that the questionnaire really measures the defined phenomenon that is the subject of research and that the selected questions are relevant. The person who designed the questionnaire, and even a team of

researchers, can never be quite sure if they have chosen the right questions. Instead, some kind of external evaluation is needed. It would be of great help to the designers of the questionnaire if a certain level of consensus on the selection of questions could be reached by people who are experts in the area of interest.

Delphi method offers the ability to precisely measure and achieve a satisfactory level of consensus on the choice of questions. Here we shall discuss the Delphi method that uses ranking. Procedure has several steps:

a) Selection of experts
b) Invitation to experts to cooperate and obtaining their consent
c) Submission of initial list of questions to the experts
d) Evaluation of questions by experts
e) Analysis of ranking from experts and summarization
f) Sending feedback to the experts with a modified list
g) Re-evaluation of questions by the experts
h) Analysis of re-evaluation and calculation of the coefficient of concordance
i) End of iteration

Steps d) through h) are reiterated maximum three times and the process can be stopped once the concordance coefficient exceeds 0.7. At that point, it is considered that there is a consensus among experts on the list of questions.

a) Selection of experts. This phase is the most sensitive part of the method and consists in the choice of experts in the field that is the subject of research. The group of experts do not include only scientists and researchers from a given field, but there might be included also other stakeholders. For example, if it is a survey that is related to the labour market, then as the experts can be involved the employees in the national employment services, members of trade unions, local government representatives, representatives of employers and the like. The key is to include the people who really have a high level of knowledge on the subject of research. In addition, it is recommended to include experts who know the observed area from different angles.

It is recommended that a group of experts is composed of 10 to 18 people. For purposes of presentation of Delphi method to the students it is not necessary to engage such a large number of experts. For a demonstration of the procedure, it is sufficient to engage 3 to 5 experts in order for the students to pass the evaluation process from start to finish.

It is also essential that none of the experts do not know who are the other members of the expert group. Only the administrator (the designer) of the whole procedure needs to know who are the experts in the group. This helps to ensure the impartiality of the proceedings, because experts can not influence each other.

b) Invitation to experts to cooperate and obtaining their consent. When the group of experts is formed, the administrator shall initiate the contact with them and invite them to cooperate and participate in the process of evaluation of the potential questions. The invitation must contain a brief and understandable explanation of why their participation is needed and also planned workload of experts in terms of time they should devote to rank the questions. In some cases, the written certificate of acceptance to participate in the evaluation is required. What is confirmed in the practice is that if someone accepts to be the expert there is a large possibility for him to stay until the end of the procedure. In other words, the dropout rate of the experts in the middle of the procedure is very small, which contributes to the sustainability of the whole procedure (Okoli & Pawlowski, p. 20).

c) Submission of an initial list of questions to the experts. Administrator is sending an initial list of questions to the experts with a clear explanation how to rank the issues and how to give their suggestions. If it's not the first iteration, but a subsequent, administrator sends a list of questions with hints and observations related to the previous evaluation.

d) Evaluation of questions by experts. According to the instructions they have received, experts rank the questions according to importance for the given problem. Rank 1 is awarded to the most important issue and so on. It is possible to assign the same priority for a few questions if the expert considered them equally important.

There is an expanded version of the method that enables experts to propose a modification of the offered questions and suggesting the new questions in a questionnaire.

If it's not the first iteration, but second or a third one, an expert has received a revised list of questions with hints and observations about how other experts have made the ranking, and eventually the new questions that have been proposed. Based on this expert is making the new observation and possibly adjusting his opinion.

At this stage it may be a problem if one of the experts was not sufficiently up to date and is late with his work assignment. Administrator in this case must make an extra effort to encourage given expert to complete the evaluation and send his results on time.

e) Analysis of ranking from experts and summarization. This phase is extremely sensitive and much depends on the experience and skills administrator has to properly examine the results obtained from the experts. We need to add new questions to the list, change the old ones (if they are suggested by experts), compare the ranking of experts, and to consider all their suggestions. Thus it is clear that the administrator has to be someone who knows well the field of research interest.

f) Sending feedback to the experts with a modified list. When summed up, the administrator sends a revised list of questions to the experts so that they could once again perform the ranking. With this revised list administrator must send a carefully worded opinions of other experts in the group in order to make a consensus.

g) Re-evaluation of the questions from the experts. Based on information received from the administrator, experts are ranking the revised list of questions and then they are giving an opinion on the current list, also commenting on the views of other experts, if necessary.

h) Analysis of re-evaluation and calculation of the coefficient of concordance. Administrator once again summarizes the views of experts and coordinates and calculates the coefficient of concordance. There are several ways to calculate this ratio, but the prevailing opinion is that the Kendall's correlation coefficient is the most suitable. The aforementioned ratio is calculated as follows:

The formula, if there are no common ranks:

$$r''_{12} = \frac{12}{m^2 \cdot n} \cdot \frac{n\sum_{i=1}^{n} S_i^2 - \left(\sum_{i=1}^{n} S_i\right)^2}{n^3 - n}$$

where is:

m – number of experts

n – number of questions

S_i, $i = 1, 2, ..., n$ – sum of ranks by rows.

The formula, if there are common ranks:

$$r''_{12} = \frac{12}{m^2 \cdot n} \cdot \frac{n\sum_{i=1}^{n} S_i^2 - \left(\sum_{i=1}^{n} S_i\right)^2}{\left(n^3 - n\right) - \sum_{i=1}^{k} T}$$

$$\sum_{i=1}^{k} T = \frac{1}{12} \sum_{i=1}^{k} \left(k^3 - k\right)$$

where is:

k – number of common ranks.

If the coefficient of concordance is greater than 0.7 than the consensus among experts regarding the list of questions has been reached. Otherwise there is no consensus and therefore the evaluation process must be continued.

i) End of iteration. Delphi method actually can have three outcomes:

- Consensus was reached because the concordance coefficient is greater than 0.7. This means that the list of questions for the questionnaire has been defined and the Delphi method ends successfully. The predetermined number of ranked questions enter the questionnaire, starting from the one with the highest average rank.

- The consensus was not reached and the administrator continues iterations until a desired level of consensus is reached. The procedure after the third iteration may continue only if the experts agree to continue with the evaluation.
- The consensus was not reached even after the third iteration and the administrator finds that there is no point to continue with the iterations. The definitive list of questions for the questionnaire cannot be made.

3. IMPLEMENTATION OF METHODOLOGY ON THE COURSE OF STATISTICS

It is assumed that students are familiar with the concept of statistical research and its phases. In addition, students should have basic knowledge about the different kinds of data collection. Only after students have acquired the knowledge mentioned above, it is possible to apply this methodology.

The methodology for the implementation of the development of the questionnaire through Delphi method includes the following steps:

1. Presentation of the Delphi method
2. Definition of the research objectives
3. Forming a team for research
4. Making the initial list of questions
5. Implementation of the Delphi method
6. Evaluation of the development process

The whole process requires some time and it is not possible to complete the implementation in a single lecture. Right approach would be to use one part of the class on the several consecutive classes to work on creating questionnaires.

3.1. Presentation of the Delphi method

The first step in implementing the methodology is to present the Delphi method in simple and understandable way to students in order to understand its significance in the whole procedure.

3.2. Definition of the research objectives

In order for students to adequately understand the methodology for the development of the questionnaire the best way is to conduct real small scale study in which the questionnaire will be used to collect data. It is necessary to determine the field of interest and research objectives related to some current problems within the scientific field. For example, at the faculty of economics, the research objectives could be connected with market prices, consumption, supplies and similar categories . It is important that topic is current, interesting and related to students' future profession.

At this stage the teacher can decide whether to make a complete design of statistical survey or focus only on the questionnaire design process. This paper explains the second option.

3.3. Forming a team for research

Several students were selected to join the research team that will closely work with a teacher. Team does not need to count more than three to five students. It would be helpful that everything that team has to do to be done in the class, so that all students could follow the procedure. The other option that would be less demanding in terms of time is that the student team at each stage refers at the class what have been done recently.

Out of the team members it is necessary to select one student who would be administrator of the Delphi method while other students will be responsible for direct communication with experts.

3.4. Making the initial list of questions

Through small brainstorming session students with the help of lecturer are suggesting the first list of questions that would be suitable for a questionnaire. They should also discuss the formulation of questions, how respondents should respond to them (type of variables) and what should be their order.

In this phase some time could be given to the students to consult the relevant literature, previous studies with the same topic and to bring their proposals at the next class.

The result of this phase is the initial list of questions that represents the input for the Delphi method. This means that evaluation of questions from experts starts with that list.

Number of questions in the initial list is not limited, but it is easier to control the procedure if there is small number of questions. It is enough that the list contains 10 to 15 questions of which in the final form of the questionnaire will remain 7 to 8 questions. Of course, to students should be informed that this is a textbook example and in practice the number of questions in the initial list may be several times higher.

List of questions should be placed in the table that has the following features:

Question	**Question's rank**	**Remarks**
1. Question		
2. Question		
3. Question		
...		
Suggestion for the new questions:		

In the column for the remarks expert may express an opinion on how it is possible to reformulate the question and also additional opinions on a given issue. In the last row of the table an expert can suggest additional questions that may be included in the survey.

3.5. Implementation of the Delphi method

Implementation of the Delphi method should go by the same procedure that was presented to the students at the start of the lecture.

Selection of the experts. Students choose members of a group of experts, and the best and easiest way is to choose among the faculty professors and teaching assistants. In this group, it is enough to find 3 members. In this way the whole process of evaluation will be done faster. Some students in the research team will be responsible for direct communication with experts. They collect their contact phone numbers and emails.

Inviting experts to cooperate and obtaining their consent. Making the first contact with experts and asking them for cooperation. Experience has shown that most teachers are very willing to participate in the evaluation process. Of key importance here is to determine whether

each expert will be available within the stipulated period of ten days. It is also important to explain how much time each evaluator should allocate to evaluation.

Submitting an initial list of questions to the experts. The initial list shall be submitted to the expert in the appropriate form (on paper or e-mail). The student in charge of contact provides expert with guidance on how to conduct the evaluation. It is recommended that the list of questions should be submitted in person and also it is useful to ask an expert to evaluate the questions immediately because it will further accelerate the whole process.

Evaluation of questions by experts. If expert received a list of questions in person there is a chance that he will immediately conduct the evaluation and that means that the administrators will very soon have the evaluation from him.

Analysis of ranking from experts and summarization. This is the most sensitive part of the procedure for students which requires significant help from the course leader. In this phase, the administrator may try to calculate the coefficient of concordance, and infer how far it is from the desired value of 0.7. The goal is through descriptive analysis to reveal where the greatest discrepancies occur for each question and also extreme values of rankings. Where the outliers were detected appropriate comments will be made and sent to the expert who made the outlier and indicate to him that his opinion differs significantly from the others.

At this stage, on the basis of suggestions, existing questions are modified while adding new questions is also possible. This assignment should only occur in the first and eventually second iteration of the Delphi method.

This phase can be divided into several smaller operations where each operation could be performed by another student. In this way, more students will be involved because one student will calculate the coefficient of concordance, other could do a descriptive analysis and another one could provide appropriate comments.

Sending feedback to the experts with a modified list. Students in charge of contacts with experts are sending to them a revised list with comments. The above mentioned students must be familiar with the

contents of the corrected list in order to provide additional information to experts if necessary.

Re-evaluation of the questions from the experts. Experts on the basis of comments received are repeating evaluation process and eventually providing additional explanations.

Analysis of re-evaluation and calculation of the coefficient of concordance. After collection of re-evaluated list the administrator calculates concordance coefficient and does the descriptive analysis. If the ratio is above 0.7 the procedure is completed and if not the new iteration will start.

End of iteration. Ideally, the agreement between experts is achieved after maximum three iterations and the list of relevant questions for the survey should be defined. Otherwise, it is necessary to do a few more iterations or end the Delphi method with conclusion that consensus has not been reached. Even if that happens, students will learn how to understand that methodology of questionnaire definition can be a serious problem in practice.

3.6. The evaluation of the process of development

Having completed the Delphi method, regardless of whether the questionnaire is defined or not, the recapitulation of the whole process should be done. Every stage from start to finish should be explained. During this presentation students need to see what was the look of list of questions after each iteration and also what were the comments and suggestions of experts and administrator.

The best solution at this stage would be if the students who have worked on designing the questionnaire explain to their colleagues what has been done and what were the specific problems encountered.

It is important to underline to the students that beside the Delphi method it is necessary to test the questionnaire in terms of both internal and external validity. Students need to know that if they have obtained the list of relevant questions, it does not mean that questions are homogeneous and fully functional as a whole. If there is enough space on the course, this part of developing the questionnaire should also be demonstrated.

4. CONCLUDING REMARKS

It is important for the students in the course of statistics to learn to collect data in different ways. Due to a large volume of material that must be taught, it is not possible to devote enough attention to this segment of statistical research. However, it is necessary to teach students at least in general to collect quality data and to get basic information on how to define questionnaire as an important measuring instrument in statistics.

This paper presents a methodology for defining questionnaires using Delphi method. Despite all its shortcomings this method allows students to learn the methodology for the selection of relevant questions that should be included in a survey.

5. REFERENCES

1. Gordon, J. (2009). *The Delphi Method.* The Millennium Project.
2. Holey, E., Feeley, J., Dixon, J., & Whittaker, V. (2007). An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. *BMC Medical Research Methodology* , 1-10.
3. Horvat, N., & Kos, M. (2010, May 10). *Development and Initial Validation of a Patient Satisfaction With Pharmacy Performance Questionnaire (PSPP-Q).* Retrieved September 19, 2011, from Evaluation & the Health Professions: http://ehp.sagepub.com/content/33/2/197
4. Legendre, P. (2004). *Species Associations: The Kendall Coefficient of Concordance Revisited.* Retrieved January 9, 2012, from http://www.bio.umontreal.ca/legendre/reprints/Kendall_W_paper.pdf
5. Okoli, C., & Pawlowski, S. (2009). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management* (42), 15-29.

6. Rayens, M., & Hahn, E. (2000). Building Consensus Using the Policy Delphi Method. *Policy, Politics, & Nursing Practice , 1* (4), 308-315.
7. Skulmoski, G., Hartman, F., & Krahn, J. (2007). The Delphi Method for Graduate Research. *Journal of Information Technology Education , 6*, 1-21.

HOW TO TEACH MATHEMATICS TO STUDENTS OF APPLIED STATISTICS

Branimir Šešelja
Department of Mathematics and Informatics Faculty of Sciences,
University of Novi Sad, Serbia

This text presents an experience in teaching mathematical topics to non-mathematicians. In particular, we deal with the master study program of Applied Statistics, namely, with the courses in Linear Algebra and Calculus. We analyze mathematical notions necessary for this study program and discuss didactical aspect of their presentation. After giving an overview of the course syllabi and their connection to other subjects of the program, we analyze the didactical aspects. These are related to the bachelor specialties of the students, to their motivation to understand and apply mathematics, and to some other conditions. We present some general conclusions about the role of Mathematics in the study of Applied Statistics.

1. INTRODUCTION

After Bologna changes in the high education programs, bachelors from one study program frequently enroll in master studies in another field. The reasons are various, sometimes connected with the situation on the labour market, or with the belief that it would be easier to get a certificate in the new field, or simply because the student realized that the previous studies were not her/his optimal choice. Apart from this reasons, there are also master-level interdisciplinary study programs, or those which connect or are related to several fields. To enroll in such master studies is a logical choice for bachelors from all fields which are in some way related to the topic of these studies.

The master program for applied statistics is addressed to bachelors in social sciences, economy, engineering, medicine etc. Obviously, these students need a considerable knowledge of mathematics. Since they enroll in the program after various basic (bachelor) studies, the levels of their previous knowledge of mathematics are uneven. This raises a problem not only in the choice of topics to be presented, but also in the way these topics should be elaborated during

the course. In addition, examples and exercises should be directly connected to applications in statistics and its usage.

The above aspects of teaching Mathematics for (applied) statisticians are elaborated in the present text.

2. SYLLABI AND HOW TO PRESENT THEM

After examining the curricula of Applied Statistics at the universities in Europe which have such or similar study programs, one could easily realize that the topics of the basic course in Mathematics more or less coincide, including:

- Some introductory topics from logic, sets, relations and functions;
- Elementary combinatorics;
- Basics of linear algebra (systems of linear equations, determinants, matrices);
- Calculus (structure of real numbers, real functions with one variable, differential and integral calculus), functions with several variables, partial derivations, optimization problems.

The logical approach should include revising on some general topics concerning deduction of mathematical notions and a few basic rules like modus ponens, contraposition, and similar, but only in an informal way and through concrete deductions. When revising (Boolean) algebra of sets and properties of set-theoretic operations, one should keep in mind the event algebra in Probability theory. This is also a good place to explain combinatorial notions. Binary relations should be connected with their characteristic functions, thus presented by tables; suitable applications of these are data bases. In addition, table representation enables a visual analysis of relations (reflexivity, symmetry, anti-symmetry, also equivalences and their block could be identified). Among special binary relations, equivalences and order should be particularly emphasized. It is sufficient to relate orderings with numbers and sets. Equivalence relations have to be connected with quotient sets, partitions. Suitable examples are e.g., fractions as classes of ordered pairs, or vectors in a plane; in both cases equality and operations could be explained in terms of class representatives.

Correspondences should be introduced analogously to binary relations, using diagrams and table (characteristic function) representations. Obviously, functions should be recalled as special correspondences, but also explained equivalently as "procedures" and "rules". In particular, operations should be introduced, and the distinction between these and relations indicated; well known examples of binary operations with numbers (addition, multiplication) are suitable.

Basic (secondary) school knowledge about numbers should be refreshed and systematized. It is not necessary to use structures like groupoid, semigroup, group, ring, integral domain or field; however, properties of all these are definitely supposed to be clearly revised and listed throughout known properties of operations with numbers. Methodologically speaking, there are two main approaches to number introduction and development: Either to start with natural numbers, then describe integers and finish with real and complex numbers, or to analyze the structure of (the field of) real (complex) numbers and identify other sets of numbers as their particular substructures. For the present course, we suggest the first option. There are several reasons. First, these topics are not new but should be revised and then systematized according to their properties and usage. For this purpose, for example, rather than locating integers in the field of real numbers, it is better to motivate the introduction of integers by the need to obtain solutions of the equation $a+x=b$ for all natural numbers a,b; similar holds true in other cases. Explicit set-theoretic construction of numbers (e.g., integers as equivalence classes of natural numbers) should be either avoided or reduced to some comments based on fractions, since they are well known. The other reason for presenting numbers according to inclusion is related to algebraic properties of the mentioned structures (groups, rings, etc.). Starting with naturals (as a semi-ring), it is easy to introduce the ring properties without explicitly defining the structure itself. It is similar with (fields of) rational and real numbers. Finally, taking into account the aim and the purpose of this study program, it is necessary to deal with arithmetical techniques through concrete problems and examples.

The basics of linear algebra are here represented by systems of linear equations, determinants and matrices. The vector space R^n over R need not be explicitly defined, though vectors, scalars and related operations in this context do appear and their properties should be

explained. As usual, determinants are introduced along with systems of two and three linear equations. In this framework it is convenient to introduce their properties. The general case can be introduced by the recursive formula, and the proper definition as a sum (using permutations) could be additionally given as an information. In matrix calculus it is important to underline properties of sum and product and consequently (implicitly) to introduce properties of the ring of square matrices: non-commutativity of multiplications, existence of the unit element (matrix) and zero divisors. In this context, inverse matrices should be given as inverse elements of the structure, and the corresponding construction explained on examples.

As usual, sequences of numbers are introduced by known examples and their limits by detailed geometric description over the real line; then comes the proper definition and techniques for determining and analyzing limits. General approach to real functions is standard in all courses of mathematics for non-mathematicians. What is specific here is that properties should be motivated and presented throughout concrete examples arising from managing various data, experiments, events. As an introduction to graphs of functions, it is convenient to use polygons, frequencies histograms, tables. After the standard presentation of elementary functions, differential calculus should be given both by means of a proper definition and by representing the derivation as an operator. The latter is even more important since it remains as a tool for all applications. In a complete presentation of real functions using differential calculus, it is necessary to analyze those which are connected with probability distributions. The same holds for integral calculus and partial derivations. In particular, use geometric interpretation when explaining basics of integral, still bearing in mind properties of distributions.

3. HOW TO TEACH

As already indicated, students enrolling these master studies have different knowledge of mathematics; some of them never had any bachelor-level mathematics, while others had. Therefore, it is not easy for the teacher to keep an appropriate tempo of delivering lectures, or to determine how detailed lectures should be. One way to overcome this

problem is to introduce each topic with a practical problem and to analyze it prior to strict definition of the involved mathematical notion.

When presenting a new topic, one should:

- Start with an example in which the new topic appears (implicitly);
- Then describe the new topic colloquially, without (too many) formulas;
- Again, present an example;
- Finally give the precise, correct mathematical definition, formulation of the theorem;
- If the theorem has to be proved, first illustrate the proof throughout an example;
- Then present the correct proof;
- Then again present the above example, now in the new framework.

For exercise, one should try to find problems in which students are given opportunity to discover the reason for using some mathematical tool. Present problems with several possible approaches and encourage students to explore and find an appropriate algorithm. Motivate students to understand topics, formulations, theorems.

Have permanent contact with colleagues who also teach statistical subjects.

Also important is to use general mathematical software (e.g., Mathematica) during practical work on problems and exercises. Thus students master get hands-on experience with that kind of application software which is a good preparation for the subsequent use of various statistical software.

4. LITERATURE

1. V. Jevremović, *Verovatnoća i statistika*, Matematički fakultet, Beograd, 2009.
2. J. Rosenblatt, *Basic Statistical Methods and Models for the Sciences*, Chapman&Hall/CRC, 2002.
3. Đ. Takači, Ar. Takači, Al. Takači, *Elementi više matematike*, Symbol, Novi Sad, 2008.

4. A. Tepavčević, Z. Lužanin, *Matematičke metode u taksonomiji*, Prirodno-matematički fakultet Novi Sad, 2006.

STATISTICS ON THE BACHELOR ACADEMIC STUDIES OF SOCIOLOGY

Valentina Sokolovska
University of Novi Sad
Faculty of Philosophy

After stressing the significance of statistical knowledge for the development of Sociology as a science, the paper discusses the chosen, most important issues, which occur in Statistics lectures at the Faculty of Humanities. Also, after the extraction of problems, possible solutions are proposed, which have been reached during all the previous research of this particular issue. The author concentrates on two groups of difficulties concerning courses on Statistics, on Sociology studies. The first ones are of methodical, and the second ones are of organisational nature.

Key words: *Sociology, statistics, courses on Statistics.*

1. INTRODUCTION

Sociology and Statistics have always been closely linked. Throughout history, however, their cooperation has been conditioned quite often, both by the manner of data collecting in Sociology, and by the development of statistical techniques which could satisfy the complexity of research of social occurences. For this reason, Clogg (1992) believes that the development of social methodology and quantitative sociology has always been linked with the development of statistical theory and methodology. He divides the application of quantitative methods in Sociology into two historical periods, before and after the Second World War. Due to the fragmentary data collection, the first period is characteristic for the application of descriptive statistics and simple methods; while, in the other period, as the quantity of data increased over time, so has the application of the more complex statistic methods.

In correspondence with such division, Raftery (2001) extracts three postwar periods in the application of statistical methods in Sociology and names them as: cross-tabulations, unit-level survey data, and newer data forms. The first period starts immediately after the

Second World War; the other starts in the early sixties of the previous century; whereas the third one begins in the late eighties. According to this author, characteristic for such extracted periods of the applied statistics in Sociology, is the fact that they are all quite current nowadays in sociological research.

It is obvious that Sociology uses statistical methods in the investigation of its own theories. However, the problem so often spotted here, is how frequently and correctly it has been performed. For this reason, many critics of Sociology have come from the very sociologists, and refer to application of statistical methods in Sociology. One of the critics appeared in the 1990s among the American sociologists. Cole (1994) explains one side of this issue. He believes that Sociology at the time: 1) has suffered due to the lack of theory which can be operationalized in research; 2) did not have a theoretical development that could be measured in accordance with the development of theories in Sciences; 3) did not have cognitive census; 4) and that such methodology is contradictory to research analyses. Similar was the reaction of Collins (1994). According to him, the advancement of statistical methodology did not reach high degree of uniformity in a way in which sociologists conduct their research, and created neither consensus nor hasty disclosure of essential issues.

Apart from all the verification of the theories, methodology development, processing precision and data analysis, Sociology recognises yet another advantage it can gain through statistics. It can be found in the ever present co-operation with other scientists during publications of scientific works. Conducting the analysis of co-author networks in Sociology, Moody (2004) concluded that authorship is more frequently found in Sociology. However, it is not equally distributed among sociological disciplines. It is often present in the areas in which quantitative analysis does not exist. Such report points to the fact that theory-inclined sociologists tend to recognise the significance of verification of its theoretical viewpoints by using statistical methods. However, for such a purpose, they require the help of others. Despite the fact that the professors of Statistics, hence the Statistics itself, are observed as "service providers" who provide services for other study programmes (Snelgar and Maquire, 2010), Ray (1974) believes that statistician, paired with a non-statistician, will never be as efficient as a

scientist who combines these skills. Only sufficient knowledge of sociological problems and statistical techniques can provide optimal combination of data characteristics and analytical methods.

2. POTENTIAL PROBLEMS IN ORGANISATION OF STATISTICS LECTURING PROGRAMME

Based on the records of the abovementioned works, the conclusion is imposed that for the contemporary calling of a sociologist, and even of Sociology as a science in the future, it is necessary to have adequate education, enriched with courses in Statistics. However, the question is what to offer on these courses and how to present statistical way of thinking and statistical methods and bring them closer to the students of Humanities. These are the questions which have been given much attention.

Even though the students of Sociology have a clear picture of what the research of mass phenomena are, they have very little or no statistical knowledge when enrolling a college. For this reason, Boynton (2004) warns that the lectures in Statistics are oftentimes inefficient, because they are too advanced for many students. What happens is the incorrect planning of courses in Statistics, conducted by the lecturers, because the knowledge of the students is not evaluated before the course starts. The consequence of unadjusted level of knowledge that we offer to the students is that they are not able to use the examples they went through in class, but they focus on memorising tests without being aware of the circumstances in which they ought to use the given examples.

The second, most common mistake in the Statistics lectures, is that many lecturers combine classical lectures with software for statistical analysis. Even though the usage of these programmes is of vital significance, Boynton (2004) believes that their premature usage can help the students to excel in the programme for statistical analysis, but without any knowledge about the test they used.

When statistics is presented as a group of tests, not connected with sociological problems, the students can be introduced to the formulas and data, but without the knowledge of how and in what way to interpret and understand the results. A completely different effect is

obtained when the data are connected to real examples, when the students are able to comprehend what is happening.

Boynton (2004) suggests that the lectures in methodology of research should connect with statistical courses. That way the students are encouraged to establish relationship between statistics and the research methods and understand why their choice of methods and research design can affect the future data analysis.

The problems of organisational nature appear in relation to the issues of content and methodical nature, and in education of sociologists. They are reflected in the unequal treatment of the place of statistics in the study programmes of Sociology. Our earlier research of study programmes of the academic studies of Sociology in Serbia (Sokolovska, 2011) show great inequality when it comes to focusing attention on the courses in Statistics. Therefore, we have programmes in which none of the courses are dedicated to statistical education, as well as those in which statistical courses take two compulsory and elective courses. Similar situation is reflected even in the neighbouring countries. Such discrepancy is the consequence of the ever present belief of creators of sociological programme, which implies that Sociology is primarily a theoretical science, which has no need for statistical education. The result of such approach is the myriad of simple, descriptive pointers in scientific works, and very few results of carefully planned, conducted and analysed sociological research. Unfortunately, such occurrences in Sociology will last until the belief that the role of statistical knowledge is a "service", which can be hired when needed, changes; and afterwards become easily forgotten. One of the ways to overcome the crisis, in which the Sociology as a science found itself, is the acceptance of Statistics as the integral part of the education of sociologists.

3. CONCLUSION

Beside the long-term cooperation between Sociology and Statistics, as well as recognising the benefits Sociology, as a science, can gain out of statistical knowledge, it has not yet become a constituency part nor the accepted part of higher education of sociologists.

The problems which might occur during that process can be divided into two groups. The first one is related to the content and methodical part of the offered statistical courses, in the sense that they

have been incorrectly planned and unadjusted to the students; that oftentimes mistakes are made in the manner of presenting statistical methods, stressing the usage of statistic software prematurely, which, consequently, leads to misunderstanding of the role of methods and their relation to research methodology; and many other reasons not enlisted in this work.

The second group of problems is defined as the organisational group. Our previous works show that the statistical courses are unequally represented in study programmes of Sociology. This discrepancy is spotted in the Universities in Serbia, as well as in certain countries in the region.

What is significant, apart from all the noticed difficulties, is the fact that multiple benefits from the application of Statistics in Sociology are recognised. Contemporary Sociology cannot be developed without the verification of its theoretical viewpoints, developed research methodology and the application of adequate statistical methods. Harmonious combination of theoretical, methodological and statistical knowledge also brought about the intensifying of cooperation between the sociologists and scientists from other areas of expertise.

4. LITERATURE:

1. Boynton, Petra (2004). Teaching statistics – the missing ingredients. *Radical Statistics*, 87: 19-30.
2. Clogg, C. Clifford (1992). The Impact of Sociological Methodology on Statistical Methodology. *Statistical Science*, 7(2): 183-207.
3. Cole, Stephen (1994). Introduction: What`s Wrong with Sociology? *Sociological Forum* 9(2): 129-131.
4. Collins, Rendall (1994). Why the Social Sciences Won`t Become High-Consensus. *Sociological Forum* 9(2): 157-177.
5. Moody, James (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohezion from 1963 to 1999. *American Sociological Review* 69(2): 213: 238.
6. Raftery, E. Adrian (2001). Statistics in Sociology, 1950-2000: A Selective Review. *Sociological Methodology*, 31, 1-45.

7. Ray, John (1974). Should Sociology Require Statistics? *The Pacific Sociological Review*, 17(3), 370-376.
8. Snelgar, Rosemary and Moira Maquire (2010). Assessing for success: An evidence-based approach that promotes learning in diverse, non-specialist student groups. In: P. Bidgoog at all (ed), *Assessment Methods in Statistical Education. An international perspective*. John Wiley & Sons Ltd.
9. Sokolovska, Valentina (2011). Position and Perspective of Statistics in Sociology. *Chinese Business Review* 10(10): 924-929.

VARIABLE SELECTION IN LOGISTIC REGRESSION

Antonio Lucadamo[1], Biagio Simonetti
Department of Economical, Juridical and Social Studies
University of Sannio
Via delle Puglie, 82, 82100, Benevento (Italy)
[1]alucadam@unisannio.it

In many fields, as for example in transportation system, in medicine, in chemiometry or in the evaluation of the Customer Satisfaction we deal with a dependent dichotomous character and with many explicative variables. In this situation the use of a linear regression model is not possible and the Logistic Regression Model is one of the possible solutions. Anyway, in many circumstances, one of the problems that can affect this model is the fact that there are few observations and many explicative variables. In this case the parameter estimation is computationally very expensive and the procedure can frequently lead to erroneous results. To solve this kind of problems, some possible solutions have been proposed in last years; in this paper we propose to use the Disco Coefficient to individuate the variables that can be significant for the Logistic Regression.

Keywords: *Logistic Regression, Disco Coefficient*

1. INTRODUCTION

In many applications related to medical, social and engineering sciences, it is of great importance to study the relationship between a dependent variable and several explanatory variables. In this context, the statistical tool most used is the multiple regression. However, when the dependent variable is dichotomous type, the model of logistic regression allows the study of dependence. In the research phase, the number of explanatory variables that influence the dependent variable can be very high, especially in relation to the number of available observations. To overcome this problem, a choose of non arbitrary number of explanatory variables can be used in the model. The objective of this paper is to apply a criterion to choose the number of explanatory variables to be used in logistic regression. The paper is organized as follows: paragraph 2 refers

to the theory of Binary Logit Model; section 3 presents an index proposed in the literature in discriminant analysis, the Disco coefficient; in section 4 the logistic regression on a suitable subset of variables selected by the results of Disco coefficient is performed; finally in section 5 a simulation study, showing that the proposed procedure allows to gain advantages in terms of estimated and computational cost is presented.

2. BINARY LOGIT MODEL

Many distribution functions have been proposed for the analysis of a dichotomous outcome variable. Logistic model is one of the most used and its popularity is due to the fact that the formula for logit choice probabilities is readily interpretable, particularly compared with other qualitative choice models (Train, 2003; Ben-Akiva & Lerman, 1985).

To understand how the model works we must consider, first of all, that the response is binary, assuming only two values that for convenience are coded as one or zero. For example we can write:

$$y_i = \begin{cases} 1 \\ 0 \end{cases}$$

So y_i is viewed as a realization of a random variable Y_i, that can take the values one and zero with probabilities π_i and 1-π_i respectively. The distribution of Y_i is a Bernoulli distribution with parameter π_i. The expected value and variance of Y_i are then

$$E(Y_i) = \mu_i = \pi_i$$

$$\text{var}(Y_i) = \sigma_i^2 = \pi_i(1-\pi_i)$$

It is easy to note that the mean and the variance depend on the underlying probability π_i.

Furthermore we would like to have the probabilities π_i depend on a vector of observed covariates x_i. The simplest idea would be to let π_i be a linear function of the covariates:

$$\pi_i = x_i^{'}\beta$$

where β is a vector of regression coefficients. The probability on the left-hand-side has to be between zero and one, but the linear predictor on the right side can take any real value and so, if we estimate the model using ordinary least squares, there is no guarantee that the predicted values will be in the correct range. A simple solution is to transform the probability to remove the range restrictions and model the transformation as a linear function of the covariates.

First of all it is important to move from the probability to the odds in the following way:

$$odds_i = \frac{\pi_i}{1-\pi_i}$$

The previous quantity is defined as the ratio of the probability to its complement. The second step is to take the logarithms, calculating the logit:

$$\eta_i = \log it(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$$

Solving for π_i we have:

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

In this way we are sure that the probability will assume a value between 0 and 1.

Furthermore now we can assume that the logit is a linear function of the predictors:

$$\eta_i = \log it(\pi_i) = x_i^{'}\beta$$

and so:

$$\pi_i = \frac{e^{x_i^{'}\beta}}{1+e^{x_i^{'}\beta}}$$

where x_i is a vector of covariates and β is the vector of regression coefficient.

The parameters of a logit model can be estimated via the Maximum Likelihood Estimation. Considering the log-likelihood function:

$$\log L(\beta) = \sum_i \left[y_i \log(\pi_i) + (\eta_i - y_i)\log(1-\pi_i) \right]$$

taking the first and expected second derivatives and developing a Fisher scoring procedure it is possible to obtain the estimates of the parameters.

The estimation of the model parameters can be obtained in different statistical software, but in some circumstances problems in the procedure can appear. In fact, in some fields, as for example in transportation system or in chemiometry, the number of independent variables is closed or larger than the number of observations. In this case the parameter estimation algorithms need many minutes and, furthermore, in many situations, strong correlation among predictors (multicollinearity) may exist and the estimation of the model parameters becomes inaccurate because of the need to invert nearsingular and ill-conditioned information matrices.

To provide an accurate estimation of the model parameters, Aguilera et al. (2006) proposed the Principal Component Logistic Regression and Camminatiello & Lucadamo (2010) the Principal Component Multinomial Regression. The two techniques overcome the

typical criticisms of the Principal Component Regression. In fact they don't consider in the analysis the components that explain better the variability in the data, but only the components significantly related with the dependent variable (generally is used a stepwise regression). Anyway, the number of components is very high and, also in this case, the stepwise regression may be computationally expensive. Disco index, introduced in discriminant analysis, could be a good tool to detect the variables that have the highest discriminatory power and that, in a second step of the analysis, can be used in a classical logistic model as regressors.

3. DISCO COEFFICIENT

Linear discriminant analysis (LDA - Fisher, 1936) is one of the most used method for classification. It captures the relationship between a categorical dependent variable and multiple independent variables. In the classical discriminant problem there are 2 classes and a single discriminant function is obtained. Fisher's approach is based on choosing linear combinations of the variables to maximize the ratio of the between-group to the within-group variances. The linear combinations of the original observations are given by

$$Z_i^{(1)} = \sum_{l=1}^{p} W_l X_{il}^{(1)}, \quad i = 1, \ldots, n_1$$

$$Z_j^{(2)} = \sum_{l=1}^{p} W_l X_{jl}^{(2)}, \quad j = 1, \ldots, n_2$$

where $X_{il}^{(1)}$ is the l-th variable of the i-th observation of Group 1, $X_{jl}^{(2)}$ is the l-th variable of the j-th observation of Group 2, n_1 and n_2 are the sizes of the two groups, p is the number of explicative variables and W_l are the weights that determine the discriminant function. Fisher's procedure maximizes $\frac{S_B^2}{S_w^2}$ where S_B^2 is the between-group variance of the linear combinations and S_w^2 is the pooled within group variance. This

approach is equivalent to maximizing Pearson's correlation ration $\eta^2 = \frac{S_B^2}{S_B^2 + S_w^2}$. This quantity can also be written as:

$$\eta^2 = \frac{N\left(\bar{Z}^{(1)} - \bar{Z}^{(2)}\right)^2}{N\left(\bar{Z}^{(1)} - \bar{Z}^{(2)}\right)^2 + S^2}$$

where $\bar{Z}^{(1)}$ and $\bar{Z}^{(2)}$ are the two arithmetic means of scores $Z_i^{(1)}$ and $Z_j^{(2)}$, $N = n_1 \cdot n_2 /(n_1 + n_2)$, $S^2 = \left[(n_1 - 1)S_1^2 + (n_2 - 1)S_1^2\right]/(n_1 + n_2 - 2)$ and S_1^2 and S_2^2 are the variances of the two distributions of scores.

In 1983 Raveh proposed a non-metric discriminant analysis (NDA), based on a separation rule different from that used in linear discriminant analysis.

In NDA the measure to be maximized is based on the set of the following inequalities:

$$Z_i^{(1)} \geq Z_j^{(2)}, \quad i = 1, \ldots, n_1, \quad j = 1, \ldots, n_2$$

The inequalities are equivalent to

$$\left(Z_i^{(1)} - Z_j^{(2)}\right) = \left|Z_i^{(1)} - Z_j^{(2)}\right| \quad \textit{for} \quad \textit{all} \quad i \quad \textit{and} \quad j$$

The index of the separation between the groups is given by:

$$IS = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(Z_i^{(1)} - Z_j^{(2)}\right)}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left|Z_i^{(1)} - Z_j^{(2)}\right|} = \frac{n_1 n_2\left(\bar{Z}^{(1)} - \bar{Z}^{(2)}\right)}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left|Z_i^{(1)} - Z_j^{(2)}\right|}$$

Raveh stated also that a generalized coefficient of discrimination for multiple groups will be given subsequently, but this was explicitly done by Guttman (1988) and then by Raveh (1989). Guttman named it disco index:

$$disco = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g n_h \left(\bar{Z}^{(g)} - \bar{Z}^{(h)} \right)}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h} \left| Z_i^{(g)} - Z_j^{(h)} \right|}$$

It is easily demonstrable that

$$0 \leq disco \leq 1$$

with $disco = 0$ if and only if all the samples have the same mean; $disco = 1$ when there is no overlapping between the scores of any two groups, except possibly at a single point.

Disco index is more robust to outliers than Fisher's index (L2 norm), because it is based on absolute values (L1 norm). Furthermore it is more informative and no distributional assumption are required for its application.

For these reasons, many application of Disco index in discriminant analysis have been developed in last years (Allard J. et al., 2000; Choulakian & Almhana, 2001; Simonetti & Choulakian, 2003)

4. LOGISTIC REGRESSION WITH DISCO

Considering the important proprieties of Disco index, our idea is to utilize it as a tool in logistic regression model.

As we said in previous sections, when there are many independent variables, the logistic stepwise regression need many minutes for the parameter estimation.

The procedure that we suggest is the following:

- Step 1: computation of the disco index for all the independent variables;
- Step 2: choosing only the variables that have a disco coefficient bigger than a defined values;

- Step 3: using these variables in a classical logistic regression model and estimating the parameters;
- Step 4: checking the obtained results.

Obviously one of the important step is the definition of the threshold value of the disco index to retain the variables in the analysis.

To define a proper value and to check the goodness of the results obtained in this way we consider a simulation procedure that we describe in the next section.

5. ANALYSIS AND RESULTS

The first step of our procedure, implemented in R-software (R Development Core Team, 2010), is to obtain a set of p explicative variables that we generated from a normal distribution. The second step is to fix a vector of real parameters β and to compute the real probabilities. Finally each value of response vector y is simulated from a binomial distribution. In this way we have the response vector and the matrix of the independent variables and we can fit a binary logit model.

As we said, the aim of the paper is to consider if the use of the disco index can reduce the computation time and if it is possible to define a threshold to choose the variables to maintain in the regression model. To reach this objective we repeated our simulation considering the same number of observation but changing the number of independent variables. We decided to fix the number of observations equal to 1000 while the number of variables varies from 5 to 115, with step 5.

For each simulation we considered, in a first phase of the analysis, the following points:

- Parameters estimation with the binary logit stepwise regression;
- Time required to estimate the previous model;
- Disco values for all the variables involved in the analysis.

At this point the problem was the definition of the value of the disco which can assure that all significant variables are preserved in the analysis.

To decide which value could be considered the right one, we compared the results obtained with all data set and we saw that a good value seems to be 0.1. Obviously this is a value obtained in empirical

way, considering only few simulations. More studies are necessary to detect a value that can be used as threshold for the disco index, but this one seems to be useful for our purposes.

In fact, in almost all the dataset, the variables that have a value bigger than 0.1 are the same that are significant in the stepwise regression. Only in some circumstances, a few variables that are significant, have a value lesser than 0.1. But, the important thing we can underline is that, in all the simulations, the variables not significant, have always a value closed to 0.

In this way we can perform a binary logit model (without stepwise procedure), considering only the independent variables selected by the disco index and we can calculate the time required to estimate this model.

As final step of the analysis a comparison between the time necessary to estimate the stepwise model and that needed with the second criterion is necessary and it can show the advantages obtainable using Disco index.

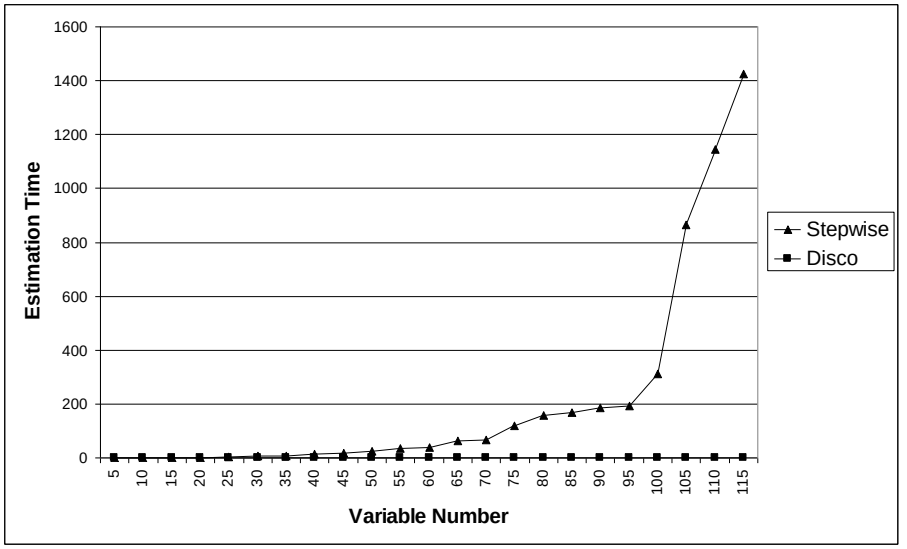

FIGURE 1 – COMPARISONS OF THE ESTIMATION TIME USING THE TWO PROCEDURES

The results about the estimation time are showed in the following table:

TABLE 1 – ESTIMATION TIME CONSIDERING THE TWO PROCEDURES

Number of variables	***Estimation time with stepwise regression***	***Estimation time with disco procedure***
5	0.16	0.02
10	0.17	0.03
15	1.14	0.04
20	1.35	0.04
25	3.77	0.05
30	5.56	0.05
35	7.00	0.06
40	13.76	0.06
45	18.18	0.08
50	23.33	0.09
55	36.69	0.09
60	38.33	0.12
65	63.25	0.12
70	67.36	0.14
75	120.49	0.18
80	156.78	0.18
85	168.45	0.42
90	186.23	0.43
95	191.95	0.50
100	310.64	0.51
105	866.44	0.56
110	1144.28	0.58
115	1424.42	0.62

It is easy to notice that, while with the disco procedure the estimation time is always less than a second, also when the number of variables increase, with the stepwise regression there is an exponential growth of the time. The situation is clearer if we consider the following figure:

In this simulations we considered a maximum number of variable equal to 115, but the advantages of using disco will be more evident if the number of independent variables increases. This situation is not unrealistic, almost in the fields that we already mentioned.

6. CONCLUSIONS

In this paper we have underlined the estimation problems that can arise in logistic regression when there are many independent variables. We described as the Disco index, used until now in Discriminant Analysis, could be an useful tool in Logistic Regression too. We showed as the time for the estimation of the parameters of the logistic model is

very low if we consider the disco index instead than a classical stepwise regression and furthermore the parameters estimated with the new procedure are very close to the real ones and there are no substantial differences with the values that we obtained using the logistic stepwise regression.

Further studies are obviously necessary, especially to define the threshold of the disco index. It has in fact defined in empirical way, but more simulations and deeper analysis are needed to define a more general criterion.

Finally other analysis to verify if the parameters estimated with the disco procedure leads to good results also for predictive aims, will be performed in future studies.

7. REFERENCES

1. Aguilera A.M., Escabias M., Valderrama M.J. (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data, *Computational Statistics & Data Analysis*, 50: 1905-1924.
2. Allard J., Choulakian V., LeBlanc R., MacNeill S., Mahdi S. (2000) Discriminant Anlysis of Seal Data, The Canadian Journal of Statistics, 28 (1), 205-212.
3. Ben-Akiva M., Lerman S. (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press, Cambridge, Ma.
4. Camminatiello I, Lucadamo A. (2010) Estimating multinomial logit model with multicollinear data, *Asian Journal of Mathematics and Statistics*, vol 3(2), 93-101, 2010, ISSN: 1994-5418
5. Choulakian V., Almhana J. (2001) An algorithm for nonmetric discriminant analysis, *Computational Statistics & Data Analysis*, 35, 253-264. Jan. 2001.
6. Fisher R.A.(1936) The use of multiple measurements in Taxonomic Problems, *Annals of Eugenics*, 7, 179-188

7. Guttman L. (1988) Eta, disco, odisco and F. *Psychometrika* 53, 393-405.
8. R Development Core Team (2010). R: A *language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
9. Raveh A. (1983) Preference structure analysis: A nonmetric approach. *Pattern Recognition* 16 (2), 253-259.
10. Raveh A. (1989) A nonmetric approach to lineat discriminant analysis. *J. Amer. Statist. Assoc*. 84, 176-183.
11. Simonetti B., Choulakian V. (2003) Discriminat analysis for spectroscopic data, Relazione invitata, *Atti del Convegno intermedio SiS 2003* "Analisi Statistica Multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia", Napoli 2003. ISBN: 88-8399-053-6.
12. Train K. (2003) *Discrete choice methods with simulation*. Cambridge University Press.

SAMPLING – WITH OR WITHOUT PROBABILITY?

Sanja Konjik
Department of Mathematics and Informatics
Faculty of Science, University of Novi Sad
Trg D. Obradovica 4, 21000 Novi Sad, Serbia
sanja.konjik@dmi.uns.ac.rs

1. SAMPLING INSTEAD OF TAKING A CENSUS

Suppose one needs to provide data on the total number of computers that are currently in use in all public elementary and secondary schools, the amount of expenses that is monthly spent on groceries per a household of a city, or the total number of overnight stays in all hotels and guest houses in a touristic region during some period. Access to all public schools, questioning all households in a city or checking guest lists of all hotels and guest houses in a touristic region would be too costly and time-consuming. Hence it seems natural to choose certain number of schools, households or hotels and guest houses, and for them collect all relevant information from which one can make conclusions about the total values that could be obtained only by examining all schools, households, or hotels and guest houses.

Every day we are witnesses, through the media, that various statistical agencies report quantitative data on national income, unemployment, political affiliation, education level or profit gained from import and export. Some statistics may be obtained from a census, but most are based on a representative sample. Thus, the information about a 100 million nation is obtained by examining a sample of only several thousand inhabitants.

In studies with the aim of collecting data about large number of people, companies or farms, the most commonly used methodology is to examine a sample instead of the entire population. A sample is a subset

of the population which is the target of research. If one makes a good sample selection, it guarantees accurate, reliable and useful data, while saving time, costs and effort. On the other hand, a bad sample may put under question the validity of research as well as the relevance of the obtained conclusions.

Therefore, a good selection of sample - which reflects as closely as possible the characteristics of interest of the entire population for a given survey - is one of the most important steps in the design and implementation of scientific research. In the literature one can find the following six steps in the process of sample selection:

1. Preparation of research;
2. Choice between taking a census or a sample;
3. Choice of sampling method – probability, nonprobability or combined;
4. Decision about the type of sample - probability, nonprobability or combined;
5. Determination of sample size;
6. Selection of sample.

Many scientists underline that a key to a successful research is preparation. The same applies to the process of selecting and examining a sample. Preparation should include the following: stating the main goals of the survey, deciding the significance of the expected research results, determining the desired accuracy, i.e., the error which can be tolerated, determining funds available for research, defining the target population, selecting research and sampling methods. In the second step it is necessary to decide whether to examine the entire population (taking census) or its subset (taking samples). This decision is often determined by resources and time that the researcher has. If it has been decided to examine a sample, the next third step is to choose between two basic methods of sampling – probability and nonprobability. In a probability sample each element of the target population has a positive probability of being selected. This form a theoretical framework for the study of properties of sampling estimators. If this condition is not fulfilled one deals with a nonprobability sample. In practice, these two methods are often combined. After the selection of sampling method one proceeds with choosing a specific form of sampling. Basic forms of probability

samples are simple random sample, stratified sample, cluster sample and systematic sample, while some of the types of nonprobability samples are purposive sample, availability sample, quota sample and snowball sample. Before the selection of the sample it remains to determine the sample size

In the sequel we shall discuss in more detail the third step in the above sample selection scheme, i.e., the choice between probability and nonprobability samples. We shall point out strengths and weaknesses of both methods, while emphasizing their mutual relationship.

2. THE THIRD STEP

After the second step, in which it has been decided to examine a sample instead of the entire population, the next step requires from researcher to determine whether to use probability sampling, nonprobability sampling or combine these two methods. Recall, if each element of the population can be in the sample with a known positive probability, while the main characteristic of the sample selection process is randomness, it is a probability sampling procedure. Otherwise, if certain elements of the population can not be selected in the sample, probabilities of selection of elements in the sample can not be accurately determined or selection of elements from the population in the sample is not random, it is the nonprobability sampling. The main advantage of probability sampling is reflected in the fact that the knowledge of selection probabilities of all elements of the population forms a mathematical framework that allows examination of sampling estimators, as well as estimates of sampling errors. On the other hand, if selection probabilities are not known it leads to necessarily subjective estimations, which is the greatest weakness of nonprobability sampling. And even when nonprobability sampling proved itself to be highly effective in some former studies, there is no guarantee that the same would happen again in similar future surveys. Thus a very natural question arises - why would anyone generally study and use nonprobability samples? We will try to answer this question in the sequel.

To begin with, let us first recall the definitions of the most important types of probability and nonprobability samples, in order to better understand their strengths and weaknesses. Assume that the

population is finite, i.e., that it contains N elements, and a sample of size n should be selected from the population.

The most important types of probability samples are:

- A simple random sample is the basic type of probability sample in which every possible subset of n elements in the population has the same probability of being selected for a sample. In most cases it is necessary to have a list of all elements of the population in order to be able to randomly select n units. For example, suppose that the aim of a survey is to hear the opinion from students at University of Novi Sad about implementation of the Bologna process in the teaching process, and that the sample size of 300 students has been determined. A simple random sample would be obtained by a random selection of 300 names from the list of all students of University of Novi Sad.
- A stratified sample is probability sampling procedure where the population is divided into subgroups called strata from which then simple random samples are taken. Strata are formed mostly of the elements that are similar with respect to the characteristics of interest, where each element of the population is in exactly one stratum. It is important to stress here that a simple random sample is taken from each stratum. Therefore, stratification in general increases precision. In the previous example, strata can be faculties of University of Novi Sad. A stratified sample would be obtained by taking a simple random sample of students from each faculty.
- A cluster sample is obtained by aggregating population units in clusters (groups), which become new sampling units, then selecting a simple random sample of clusters, and finally examining all elements of selected clusters or taking again simple random samples of elements within the selected clusters. Clusters are often naturally determined, e.g. classes in a school. Since a simple random sample is selected on the level of clusters, and then all or a subset of elements of selected clusters is examined, this type of sampling in general decreases precision. In our example natural clusters would be faculties or departments. A cluster sample would be obtained by selecting a simple random

sample of faculties or departments, and subsequently interviewing all students at selected faculties and departments, or choosing new simple random samples of students at selected faculties and departments.

It is worth mentioning that in all of the three cases the selection of units to be in a sample is random: in the case of a simple random sample elements are randomly selected from the population, in the case of a stratified sample random selection of units is made within strata, and in a cluster sampling one randomly chooses clusters.

Now we list some of the most frequently used nonprobability sampling procedures:

- A quota sample is a form of nonprobability samples in which the population is divided into disjoint subgroups according to characteristics of interest (as in a stratified sample), with established size of such subgroups, and then researchers determine the sample size and number of elements from each subgroup, i.e., quota, which should be included in the sample. The essential difference from a stratified sample is in the fact that the selection of the elements to be included in the sample in the last step is not made by the probability sampling techniques, but rather this choice is left to the free judgment of researchers. Going back to the example of a stratified sample we have given above, one would determine quota of students from each faculty, e.g. 10% of students from each faculty, but the method of selecting students is left to be decided by researchers.
- An availability sample is a sample that is available to the researcher, that is convenient. For example, professors who lead the research ask their students to complete a questionnaire on the implementation of Bologna process. It is obvious that this kind of sample in general is not representative, that most of the units of the population has no chance of being included in the sample, and that there is no theoretical basis for estimation of the error resulting from the conclusion based on this sample. On the other hand, availability samples are very easy for examination.
- A purposive sample is also a nonprobability sample in which the researcher selects those elements of the population in the sample

that in his/her opinion best fit most of the research objectives. Unlike the availability sample, units from the population are not selected in the sample due to their availability or convenience, but the researcher purposely chooses a sample of elements that meet the criteria that are important for the survey. In the example of the survey of students on the application of Bologna process at the University of Novi Sad, a purposive sample could be selected from the students who attend lectures regularly.

- A snowball sample belongs to the nonprobability sampling procedures that are used only in social surveys. First the researcher selects certain number of respondents who will further refer to new candidates that should be included in the sample. It is appropriate for studying characteristics of rare populations such as drug addiction, AIDS, crime, sexual orientation, prostitution, etc.

We may conclude that in probabilistic sampling the researcher's bias is excluded, and selection of elements to be included in the sample is independent of researcher's needs or desires. However, in the nonprobability sampling procedure the selection of sample is more or less influenced by the subjectivity of the researcher, thus excluding the possibility of using the probability and statistics theory in processing the collected data, drawing conclusions and estimating errors arising from the sampling instead of taking a census. We shall now list some main weaknesses of nonprobability sampling compared to probability one:

- not appropriate for a nonexploratory research
- does not allow determination of the precision of obtained estimates
- does not allow the use of mathematical tools in data processing
- selection bias
- loss of representativity
- subjectivity of researchers
- not appropriate in heterogeneous populations

Nevertheless, nonprobability samples are more appropriate than probability sampling procedures in the following cases:

- in exploratory researches

- when the funds available for the research are limited or small
- when the time available for the research is short
- when it is necessary to examine specific units of the population
- when the entire population is not known
- when the population is homogeneous
- when it is not necessary to have a representative sample
- when it is not necessary to minimize the selection bias
- when the research team has little or insufficiently trained researchers
- when the population is highly scattered
- for commercial surveys, public opinion surveys or for examination of rare characteristics of the population

Eventually, we summarize the advantages and disadvantages of probability and nonprobability sampling procedures in the following table:

TABLE 1 – THE ADVANTAGES AND DISADVANTAGES OF PROBABILITY AND NONPROBABILITY SAMPLING PROCEDURES

Characteristic of research	**Probability sample**	**Nonprobability sample**
Exploratory research	disadvantage	advantage
Nonexploratory research	advantage	disadvantage
Short available time	disadvantage	advantage
Representative sample	advantage	disadvantage
Precision estimation	advantage	disadvantage
Minimal selection bias	advantage	disadvantage
Homogeneous population	disadvantage	advantage
Heterogeneous population	advantage	disadvantage
Limited research funds	disadvantage	advantage
Public opinion surveys	disadvantage	advantage
Examination of rare populations	disadvantage	advantage
Commercial surveys	disadvantage	advantage

Similarly, we give tables of the strengths and weaknesses for particular types of probability and nonprobabilty samples:

TABLE 2 – THE STRENGTHS AND WEAKNESSES FOR PARTICULAR TYPES OF PROBABILITY SAMPLES

Probability samples	**Strengths**	**Weaknesses**
Simple random sample	simple	high costs, requires complete list of the population
Stratified sample	increases precision	complex, high costs
Cluster sample	low costs, simple	decreases precision

TABLE 3 – THE STRENGTHS AND WEAKNESSES FOR PARTICULAR TYPES OF NONPROBABILITY SAMPLES

Nonprobability samples	**Strengths**	**Weaknesses**
Quota sample	high representativity, lower sampling error	complex, high costs
Availability sample	simple, low costs, time-efficient	nonrepresentative, large selection bias
Purposive sample	good selection control	complex, nonrepresentative
Snowball sample	appropriate for examination of rare populations	time-consuming, complex

In practice, one often combines different types of probability and nonprobability sampling procedures in order to achieve greater efficiency, better estimates and cost reduction.

PROBABILITY TREE DIAGRAM, TOTAL PROBABILITY AND BAYES FORMULA

Marko Obradović
University of Belgrade
Faculty of
Mathematics

Opinions differ on how to teach mathematical topics to non-mathematics students. Some teachers maintain that it is not necessary to explain the essence to those students who, in their future work, will use mathematics only as a tool. It is considered enough just to present mathematics as a set of formulas which are to be memorized and applied as such. On the other hand, there are those who think that mathematics has to be explained, adapting the presentation to fit students' knowledge and needs. Considered in this paper are the two approaches based on some typical problems which can be found in any basic course on probability - namely, the problems that are solved using total probability and/or Bayes formula.

1. TOTAL PROBABILITY FORMULA

The problem that is solved using total probability formula can be understood as an experiment in stages. For example:

- First we choose an urn, and then randomly pick a ball from the urn
- First we randomly put two balls, then we randomly pick a ball

Let us consider two-stage problems. In such problems, the probabilities for the first stage of the experiment, and the conditional probabilities of outcomes from the second stage, given the ones from the first stage, are given or can be (easily) calculated. The goal is to find the probabilities from the second stage.

Let there be n outcomes in the first stage, and m outcomes in the second stage. Since an outcome from the second stage is presented with its conditional probabilities given the first stage outcomes, its probability is calculated as a sum of probabilities of its intersections with all the outcomes from the first stage, i.e.

$$P(B_j) = P(A_1 B_j) + P(A_2 B_j) + \cdots + P(A_n B_j), j = 1, \ldots, m. \quad (1)$$

As any intersection probability can be calculated as $P(AB) = P(A)P(B|A)$, we have

$$P(B_j) = \sum_{k=1}^{n} P(A_k) P(B_j|A_k), \quad (2)$$

which is known as *total probability formula*.

2. PROBABILITY TREE DIAGRAM

We can present the stages of such experiment graphically. The best diagram is probability tree diagram shown on Figure 1.

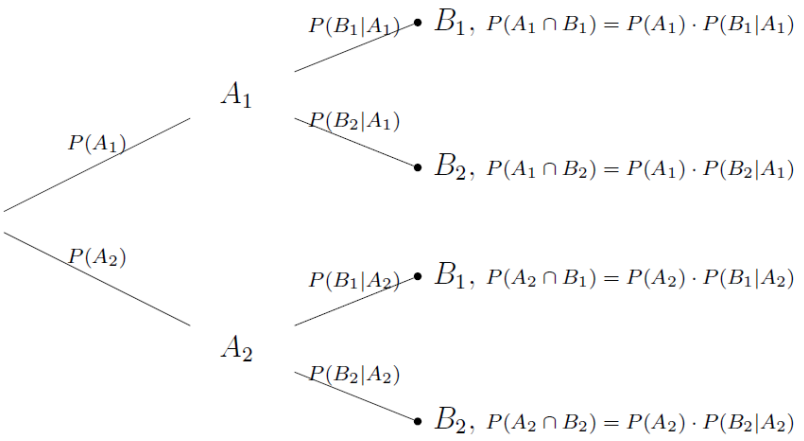

FIGURE 1. - PROBABILITY TREE DIAGRAM FOR n=m=2

On the branches of the tree there are probabilities that follow the particular branch, while the nodes represent the outcomes that the branch leads to. At the end of any path from the "root" to the "leaf" of the tree we get to the intersection probability of all the outcomes on the path.

Multiplying the probabilities on the used branches we obtain the probability from (1).

We can now calculate the probability of the outcome B_j by adding the intersection probabilities of all the paths that lead to B_j, i.e., by adding the probabilities of all the intersections in which there is B_j. Naturally, the obtained sum is identical to the one in (2).

Let us consider this on an example.

Example 1. In the first urn there are three white and one black ball, in the second there are two white and three black ones, and in the third there is one white and two black balls. An urn is chosen at random and the ball is randomly taken from it. Calculate the probability that it is white.

Solution. As the urn is chosen at random, all the probabilities in the first stage (urn choosing) are equal to $\frac{1.}{3}$.. All the conditional probabilities are shown on the diagram on Fugure 2.

We can now calculate the required probability as the sum of intersection probabilities on the branches that lead to B tj.

$$P(B) = \frac{1}{4} + \frac{2}{15} + \frac{1}{9} = \frac{89}{180}$$

Total probability formula yields following solution

$$P\,(B) = \sum_{k=1}^{3} P\,(K_k)\,P\,(B|K_{\,k}) = \frac{1}{3}\cdot\frac{3}{4} + \frac{1}{3}\cdot\frac{2}{5} + \frac{1}{3}\cdot\frac{1}{3} = \frac{89}{180}$$

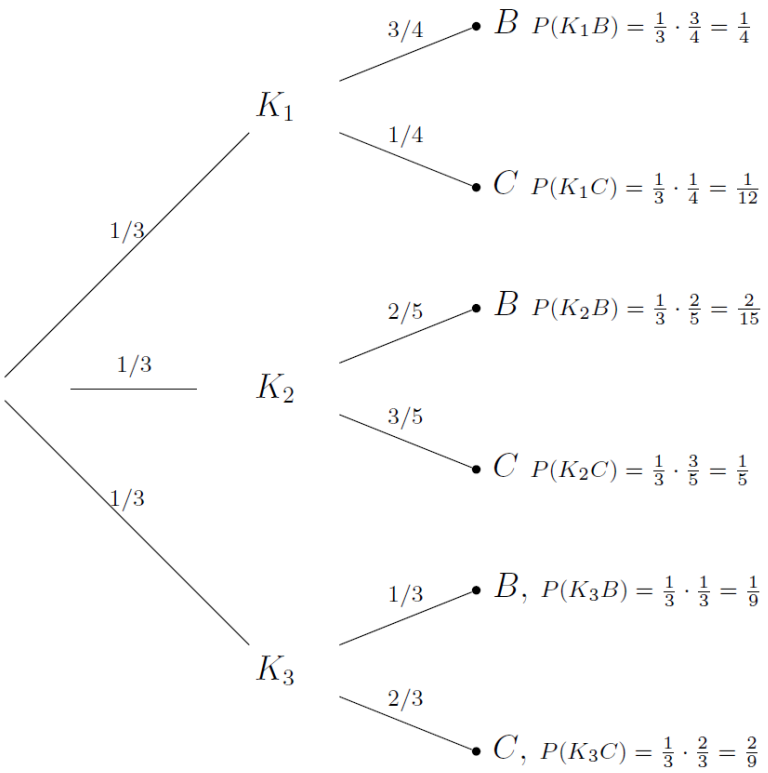

FIGURE 2. - PROBABILITY TREE FROM EXAMPLE 1, K -URN, B - WHITE, C-BLACK

3. BAYES PROBABILITIES

In this kind of problems, the goal is very often to find 'reverse' probabilities, i.e., the probabilities of the outcomes from the first stage given the outcomes from the second one. These probabilities are called Bayes probabilities.

The standard way to find these probabilities is by using Bayes formula :

$$P\left(A_i|B_j\right) = \frac{P(A_i)P\left(B_j|A_i\right)}{\sum_{k=1}^{n} P\left(A_k\right)P\left(B_j|A_k\right)} \tag{3}$$

We can find Bayes probabilities by constructing the tree of "reversed" experiment. From the diagram (Figure 2) we can calculate all

the intersection probabilities and the probabilities of all second stage outcomes B_j. Then by making a reversed probability tree diagram on the corresponding branches, Bayes probabilities will appear as the ratios of probabilities of intersections and the outcomes from the second stage. As the experiment is "reversed", we will reverse the tree as well (Figure 3).

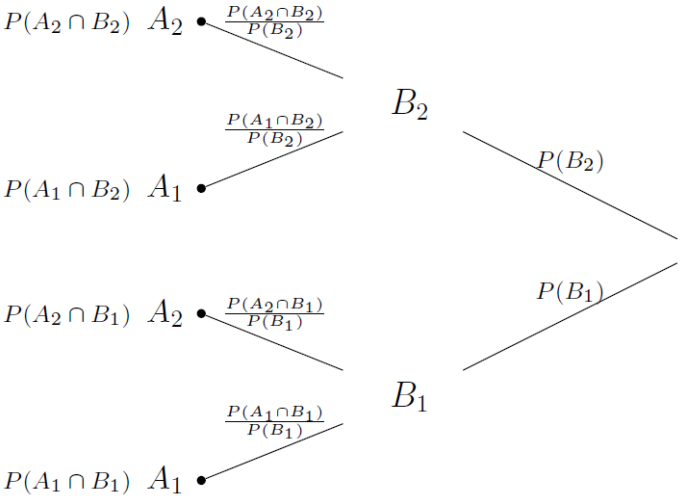

FIGURE 3. - REVERSE PROBABILITY TREE DIAGRAM FOR n=m=2

Let us again look at the example.

Example 2. Let us have the same experiment from the Example 1. Given the black ball was taken, what is the probability that it was taken from the first urn?

Solution. Using the calculations from the Example 1, we create the reverse tree diagram (Figure 4).

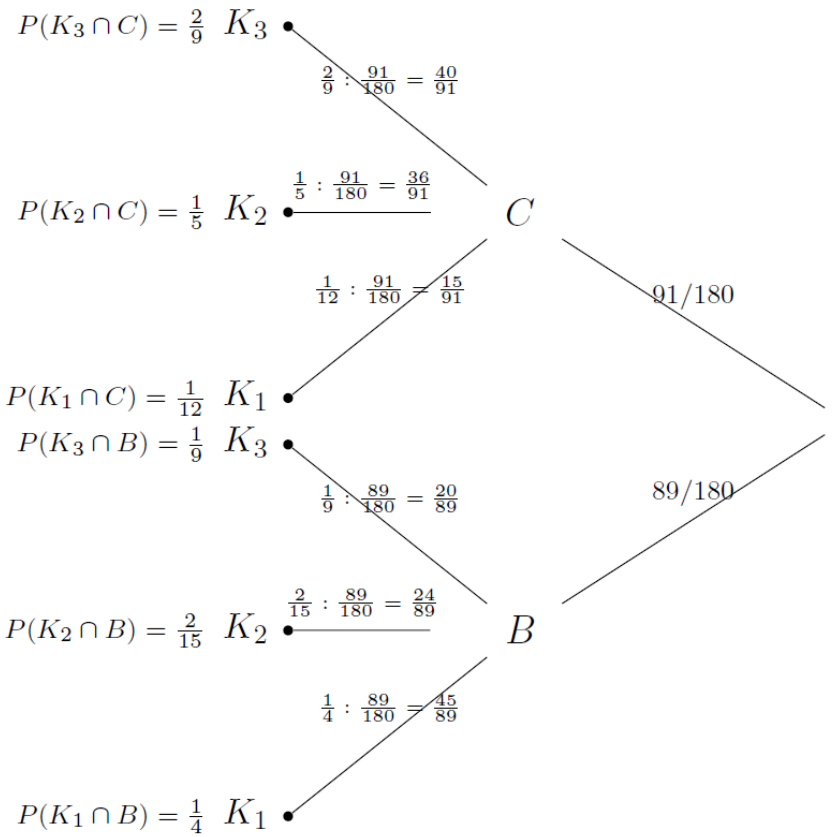

FIGURE 4. - REVERSE TREE DIAGRAM FROM EXAMPLE 2, K -URN, B -WHITE, C-BLACK

The asked probability is on the branch that links black ball and the first urn, $\frac{15}{19}$.

Using Bayes formula, the solution is:

$$P\,(K_1|C) = \frac{P(K_1)\cdot P(C|K_1)}{\sum_{k=1}^{3} P(K_k)\cdot P(C|K_k)} = \frac{\frac{1}{3}\cdot\frac{1}{4}}{\frac{1}{3}\cdot\frac{1}{4}+\frac{1}{3}\cdot\frac{1}{5}+\frac{1}{3}\cdot\frac{2}{9}} = \frac{15}{91}$$

4. CONCLUSION

Let us now state some advantages of these two methods.

The advantages of probability tree diagram are following:

- It is suitable for teaching examples and understanding the essence of the problem;
- The experiment can be seen in entirety;
- It is possible to calculate more than one probability using one diagram;
- Errors are reduced since there is no need to memorize the formula;
- The method can easily be applied to the problems involving more stages;

The advantages of classical method (formula) are given below:

- It is suitable for problems with large number of outcomes, as the formula does not grow in complexity
- The writing is shorter and the calculation is briefer;
- It is more “mathematical”;
- Has theoretical importance.

From the personal experience in teaching students of biology and meteorology, the author realized that some students have less difficulties when the diagrams are used, while others are more inclined to memorize the formula. This depends on both their individual affinities and their background which tends to be different from school to school. For this reason, and since both methods have some objective advantages, the author maintains that the best approach is, if possible, to combine them.

SAMPLE SIZE DETERMINATION

Dušan Rakić
Faculty of Agriculture
University of Novom Sadu

A wide variety of investigations leads to the necessity of studying the characteristics of groups of elements. The inability to access each element and time and material constraints, are the most common reasons why researchers are forced to examine the required characteristics in a subgroup of the initial group. The initial group is called population, the elements of the population are units (they can be people, plants, objects...) and by N the number of units in the population is denoted. The required subgroup is called sample and by n the number of units in the sample is denoted. Clearly, the chosen sample should be representable and accessible so a research can be done over it and, on the basis of these results, conclusions can be drawn for the whole population. Determination of the sample size (volume) is a very important decision in the research. Depending on whether it implies working with people or not, the sample size can significantly affect the results of a study and its validity.

For example, a very large sample increases the time and the cost of research, while using the small size of the sample can significantly reduce the accuracy and reliability of results. Although it seems that the determination of the sample size is the first step in a research, usually it is not the case, especially in well-designed and well-prepared research investigations. For example, it is wrong to begin a research with a question: "What percentage of the population has to be included in the sample?" because sometimes the number of units in the population is not known (for example, if you study certain characteristics of forest plants, the total number of plants throughout the forest most probably is unknown). Therefore, the "magic formula" for the size of the sample does not exist and the number of units that will be included in the study

depends on several factors. Well done research should include the following phases:

1. preliminary studies in which population is analyzed in relation to the required characteristic and also conditions under which research is carried out (time and budget constraints),
2. selection of the methods that will be used for data collection from the sample and the methods that will be used for processing data,

and only after them, a decision follows on the size of the sample. These phases are explained in more details in this text, particularly emphasizing how they affect the determination of the sample size.

1. PRELIMINARY STUDIES

First, the researcher should be thoroughly familiar with the specifics of the study and the conditions under which it is carried out. These studies are often crucial to the accuracy of final results and the cost of research and often they take longer than the later data collection and processing. Factors to be considered at this stage of work are to a large degree defined by the research itself and its characteristics. As a guideline we can list some of them:

1. the nature of the population and the units included in the research,
2. what financial resources are available and what are time constraints,
3. what are the objectives of the research,
4. do we have any ethical or legal restrictions at work.

2. NATURE OF THE POPULATION

Probably the main factor in determining the sample size is a knowledge of nature (characteristics) of the population and the units that it is consisted of. First of all, size of the population and the availability of its units. If it is very small and we have access to all units, then the question is whether to study whole population or to choose the sample. Further, assessment of homogeneity of the population is very important. If the population is homogeneous (units are fairly evenly distributed in comparison with the characteristic under study), then a small sample can

be determined, and also cheaper sampling methods could be used (for instance, cluster method). For example, if you have two pots, of 1 liter and 100 liter volume, and if the beans is mixed well, in both cases it is enough to take one teaspoon to check whether beans is well-flavored. On the other hand, if the population is heterogeneous, then a larger sample should be considered and sampling methods adjusted for this case (for example, stratified sampling). The spatial distribution of the population units can also affect the size of the sample. It is known that in the case of scattered population, in order to reduce the costs of research, the cluster method is often chosen although it is not characterized as accurate.

3. AVAILABLE RESOURCES

Some of the first information necessary for good planning of the research are financial resources that we have, estimated time for conducting the study, as well as how many and how skilled human resources are at our disposal. These parameters directly affect the quality of a research, especially if physically separated units are sampled, if the method of data collection is complex and requires special expertise (work with plants, animals, delicate questions in the surveys). Since the sample size is often directly proportional to the cost of processing one unit (then the sample size is simply determined by dividing the total funding at our disposal with the cost of processing one unit), it is recommended to observe more characteristics while examining one unit, and to perform more research that way with almost the same cost as one single study. Therefore, in order to rationalize the budget, researchers often use cluster method where the sample includes the entire subgroup of similar units (all students of a class, all residents of a building, all the plants in one part of the plot), although, on the other hand, the cluster method as one of inaccurate methods requires increasing the sample size in order to obtain data that can be considered as relevant.

4. OBJECTIVES OF THE RESEARCH

Sample size largely depends on the tasks and objectives of the research. If the research will not give the final decision on an issue, but will be a part of a study or a theoretical consideration, the sample may be small, while in the case when the research results can have serious financial and social consequences (banking, methods of treatment in

medicine, reforms in education...) observed sample has to be large in order to obtain more precisions results. Also, the sample size should be large if we want to get the answers to the question how subpopulations react to the observed characteristic, not just the total population (when examining the effect of a medication, it is important to know how it affects different ages, genders, patients suffering from other diseases and conditions...). Some subgroups that are of interest to us may be rare or hidden, and the larger sample is required to obtain as much information as possible about them. Therefore, great importance and responsibility of research together with the need for a more detailed and precise analysis will lead us to the large size of the sample.

5. ETHICAL STANDARDS

In the cases where research (usually when studying people) may violate the ethical and legal norms (invasion of privacy, disclosure of data or opinions that may, if anonymity is not preserved, violate participant's social status (sexual orientation, political issues), or cause problems at work (if the participant discloses data prohibited by the code of the company where he/she works)), then sample size should be limited to as small as possible that will be sufficient to achieve the required reliability of the results.

6. METHODS FOR DATA COLLECTION AND PROCESSING

After a detailed analysis of problems and work conditions, and before the determination of the sample size, we should decide which of sampling methods (or combination of them) from sampling theory and methodology to be used in research, how to collect and process data. It is important to know whether the data are obtained via phone, email, directly from the survey, and also whether probability or nonprobability sample design is going to be used. Only after all these preliminary studies and selection of methods the size of the sample can be determined. Depending on the chosen approach, total size of the sample can be determined at the start of work (if you use the method where a sample is fixed at the beginning and it does not change in the course of the study) or it is determined during the research (if the size of the sample changes based on some given rules and choices in the course of the research).

7. PROBABILITY AND NONPROBABILITY SAMPLING

When choosing a method for determination of the sample, the first decision to be made is whether to use a probability or nonprobability sampling method. The most frequent probability sampling methods (some of them are mentioned above) are the simple random sampling, stratified sampling and cluster method, while the most widespread nonprobability sampling methods are convenience, purposive and quota sampling.

In probability sampling the presence of formulas for calculating sample size can be expected and they are usually derived from the formulas of confidence interval and, roughly written, they have following forms:

$$n = \frac{z^2 \cdot s^2}{e^2},$$

where z determines the level of confidence (usually, 95% confidence is used), s^2 is the estimation of the population variance around the mean value, and e is margin of error tolerable in the research. Values z and e are based on the demands of accuracy and reliability of research results, and typical values for them are $z = 1.96$ (implies the 95% confidence level, following the rules of normal distribution), and $e = 0.03$, where special attention should be given to the fact whether e is given as an absolute or a relative margin of error. The hardest thing is to calculate a value of s^2, since at the beginning of the research we have no knowledge of the population variance around the mean value and it can be determined by one of the following:

- using the pilot sample size of 20 to 150 units that will represent the preliminary study for the main research from which we get the assessment of the values required for determining the sample size,
- using similar surveys, conducted earlier, which are available in the literature (it is very likely that someone else has already dealt with some similar issues),

- simply guess the value of the dispersion on the basis of experience and knowledge of the population characteristics (mostly homogeneity).

If we examine the proportion of some characteristics in relation to the entire population, we use following formula form:

$$n = \frac{z^2 \cdot p \cdot (1-p)}{e^2},$$

where p is the estimated proportion (obtained in one of the specified ways to assess dispersion). The following table shows the sample sizes for some typical values of the proportion and the margin of error when $z = 1.96$.

TABLE 1 – THE SAMPLE SIZES FOR SOME TYPICAL VALUES OF THE PROPORTION AND THE MARGIN OF ERROR WHEN $z = 1.96$

p \ e	0.01	0.03	0.05	1
0.01	380	42	15	4
0.1	3457	384	138	35
0.25	7203	800	288	72
0.5	9604	1067	384	96

If we are unable to estimate p, then it is safest to take that $p = 0.5$ since the function $p \cdot (1-p)$ has the maximum at $p = 0.5$ and value $n = 1067$ is often recommended for the size of the sample (when examining proportions, with the margin of error $e = 0.03$), whatever the number of the population is. It also should be noted that by examining the finite population correction factor

$$fpc = 1 - \frac{n}{N},$$

which participating in most of the formulas for evaluation of the parameters of the different sample designs, we can come to interesting conclusions about the sample size. Note that for small values of N the

correction factor is small (close to the value 0) and significantly affects the estimates, while if N has a great value and number of units in the sample is considerably smaller than N (the situation we usually have in public researches), then the correction factor is negligible and only becomes significant if the sample size is greater than 5% of the population size. Accordingly, absolute size of the sample is often more important, then its relative value. Let us show another example (simple random sample method) that if the population is very large then it is not a main factor in determination of the sample size. The variance of the sample mean, which directly determines the accuracy of our results, is calculated as follows:

$$V(\bar{y}) = \frac{S^2}{n} \cdot \left(1 - \frac{n}{N}\right),$$

where S^2 is the population variance around the mean value. We can see that for fixed values S^2 and $n = 100$ value $V(\bar{y})$ is approximately the same for $N = 100.000$ and $N = 100.000.000$, i.e. with a sample of 100 units we have the same precision in two cases where the size of the population differs significantly.

Probability sampling design in which the same unit is processed more than once within the specified time intervals requires a larger sample, knowing that we cannot be sure that each unit will be able to complete the entire cycle of research (an example might be the control of a medication effects on patients in the course of long-term treatment, where the inability of patients or in extreme cases patients' death, can keep some of them from undergo treatment at all its stages).

In nonprobability sampling only recommendations for the size of the sample can be expected in the form of specific values in relation to the type of research. These values can be reached through experience (it is very likely that someone has already conducted the same or similar research) over a long period of work on similar problems. For example, it is known that for an ethnographic study 35 to 50 units should be included in a sample, for market research 200 to 2.500 units, in the national research 10.000 to 15.000 units, etc..

From the facts given on the sample size and factors that influence it, we came to the conclusion that a good selection of sample size is made on the base of a detailed analysis of the population in relation to the characteristics under study, and a good selection of method used for collection and processing data in relation to the problems nature, financial and human resources we have.

ONE LECTURE – ONE HUNDRED STATISTICAL TERMS

Vesna Jevremović
Faculty of Mathematics,
University of Belgrade, Serbia

Uniform distribution is the simplest continuous distribution, but plays an important role in Statistics since it is indispensable in modelling random variables, and therefore in Monte Carlo methods. In this paper, which can be taught as one lecture, the basic properties of uniform distribution are given, as well as one nonparametric test of goodness of fit based on Lorenz's index of concentration which uses the uniform distribution.

1. INTRODUCTION

Teaching and learning Statistics requires knowledge from different areas of Mathematics: calculus (limits, series, integrals, functions in one and several variables), algebra (matrices), numerical analysis (solving equations) as well as the use of statistical software and programming skills.

There are lots of statistical terms in any lecture in Statistics, many of which are connected. That is why the understanding of the lecture and future application of the topics given are possible only if the student knows well the definitions of statistical terms, their basic properties and relationships between them.

Almost every lecture in statistics abounds with mathematical/statistical terms, but sometimes we can even be surprised with rather unexpected range of terms in one single lecture, as is the case with the following lecture about the uniform distribution. In order to draw attention to every statistical term introduced, their first appearances shall be italicized throughout this text.

2. BASIC PROPERTIES OF CONTINUOUS UNIFORM DISTRIBUTION

Basic properties of the *uniform distribution* (or *rectangular distribution*) are well known and we will briefly state them. The *random*

variable X is said to be uniformly distributed on the interval [a, b], and we write *X: U(a, b),* if its *probability density function* equals $f(x)=\frac{1}{b-a}$, $x\in[a,b]$, and 0 elsewhere. It follows that the *distribution function* is $F(x)=\frac{x-a}{b-a}$, $x\in[a,b]$.

The moments are $m_r=\frac{1}{r+1}\frac{b^{r+1}-a^{r+1}}{b-a}$, $r\in N$, while the *central moments* are $\mu_{2k-1}=0$, $\mu_{2k}=\frac{1}{2k+1}\left(\frac{b-a}{2}\right)^{2k}$, $k\in N$. Distribution *mode* is not unique, the *median* is obviously *(a+b)/2* because of *the symmetry* of the uniform distribution. From the symmetry it follows also that the *coefficient of asymmetry* is 0, while the *coefficient of skewness* is *−1.2*.

In the simple experiment where a number X is *chosen at random* from the interval (0,1), we have that *X: U(0, 1).*

Uniform distribution is a special case of the *beta distribution*. Namely, uniform distribution *U(0, 1)* is $B_2(1, 1)$ distribution. Moreover, if $(X_1,\ldots,X_n)$ is a *random sample* from the *U(0,1)* distribution, then the k^{th} order statistic $X_{(k)}$ has the beta distribution $B_2(k,n-k+1)$.

Characteristic function for *X: U(a, b)* is $\varphi(t)=\frac{e^{itb}-e^{ita}}{it(b-a)}$, $i=\sqrt{-a}$, while the *moment-generating function* is $M(t)=\frac{e^{tb}-e^{ta}}{t(b-a)}$.

3. DISCRETE AND CONTINUOUS UNIFORM DISTRIBUTION

A random variable (r.v) *X* that assumes a finite number of distinct values $x_1,\ldots,x_n$ each with the same *probability* $P(X=x_j)=1/n$, where *n* is a positive integer, is called a *discrete uniform distribution*.

Some relationships between discrete and continuous uniform distribution are given in the next two theorems.

Theorem 1. Let B: U(0,1) and Bj, j=1,2,...,N-1 are *independent* r.v. with P(Bj=0) = P(Bj=1) =1/2. *Then* $\sum_{j=1}^{N-1} Bj/2^j + B/2^{N-1} : U(0,1)$ *and* $\lim_{N\to\infty} \sum_{j=1}^{N-1} Bj/2^j : U(0,1)$.

The relationship between uniform distribution *U(0,1)* and discrete uniform distribution of the form

$$Y : \begin{pmatrix} 0 & 1 & 2 & \cdots & 9 \\ 0.1 & 0.1 & 0.1 & \cdots & 0.1 \end{pmatrix} \qquad (*)$$

is given by the following theorem

Theorem 2. Let $\gamma = 0.u_1u_2\ldots u_n\ldots$ be a *realization* of a random variable X: U(0,1). Then $u_1, u_2,\ldots,u_n,\ldots$ are independent realizations of a discrete uniform distribution Y, and vice versa.

The theorem 2 could be generalized:

Let us have a random number written down in base r-numeral system $\gamma = 0.\ v_1v_2\ldots v_n\ldots$, then $v_1, v_2,\ldots,v_n,\ldots$ are independent realizations of a discrete uniform distribution V having distribution

$$P(V=j)=1/r,\ j=0,1,\ldots,r-1,$$

and vice versa.

4. UNIFORM DISTRIBUTION AND THE MAXIMUM LIKELIHOOD

If *X: U(a,b)*, where *a* and/or *b* are unknown, we may *estimate* them using a sample $(X_1,\ldots,X_n)$ from this distribution. Since the uniform distribution is not *regular in the Rao-Cramer sense*, then the *estimators* based on the maximum likelihood method may have interesting properties.

Example 1. Let us first analyze the case of *U(0,b)*. Maximum likelihood method (ML) gives the estimator $\hat{b} = \max_{1\le j\le n} X_j = X_{(n)}$, and the method of moments (MM) gives $\bar{b}_n = 2\bar{X}_n$. We have

$$E(\hat{b}) = nb/(n+1), D(\hat{b}) = nb^2/(n+2)(n+1)^2$$

$$E(\bar{b}_n) = b, D(\bar{b}_n) = b^2/3n\,.$$

This way the ML estimator is *biased*, but *asymptotically unbiased* and *stable*. It is also *more efficient* then the *unbiased* MM estimator. The estimator $\frac{n+1}{n}\hat{b}$ is unbiased and its *variance* is still smaller than the variance of the MM estimator. In addition, the variance of this estimator is less than the *lower bound of variance* from the *Rao-Cramer inequality*, which is b^2/n. Furthermore the statistic $\hat{b} = \max\limits_{1\le j\le n} X_j = X_{(n)}$ is a *complete sufficient statistic* for the parameter *b*; it is also *consistent* and even *squared-error consistent* since $\lim\limits_{n\to\infty} E(\hat{b}-b)^2 = 0$. Using *order statistics* we can find other unbiased estimators, such as $\hat{b}_1 = (n+1)\min\limits_{1\le j\le n} X_j = (n+1)X_{(1)}$, or, in the case of odd *n*, *n=2k+1*, the estimator based on the *sample median* $\hat{b}_2 = 2X_{(k+1)}$. In addition, since the distribution of ML estimator is known, we can find the *confidence interval* for the parameter b, with the desired *level of confidence*.

Example 2. As a second example we take the distribution *X: U(a–1/2,a+1/2)*, for which the *ML estimator* is *not unique*, and every statistic *V* satisfying the inequality

$$\max_{1\le j\le n} X_j - \frac{1}{2} \le V \le \min_{1\le j\le n} X_j + \frac{1}{2},$$

could be taken as an ML estimator for *a*.

Example 3. Finally in the case of the *family of distributions X: U(–a,a), a>0*, we have to say that this *family is not complete*, and the *joint sufficient statistic* for *a* is $(X_{(1)},X_{(n)})$. The *coefficient of correlation* of r.v. $X_{(1)},X_{(n)}$ is 1/n, so their dependence is decreasing while n increase. The ML estimator $\hat{a}=max(-X_{(1)},X_{(n)})$ is also a *minimal sufficient statistic* for this parameter.

5. UNIFORM DISTRIBUTION AND THE MONTE CARLO METHODS

Monte Carlo methods use *pseudo-random numbers* in order to determine the properties of some function or set of functions, evaluate integrals, or study the outcome of some complex system. Pseudo-random numbers are computer-generated sequences of numbers that show *statistical properties* of a *sequence of independent identically distributed random variables* having uniform distribution, discrete or continuous.

Modelling random variables is the first step in Monte Carlo methods, and for this step uniform distribution is required. Methods for modelling random variables vary depending on the nature of random variables, but they all use modelled values of the uniform distribution. If this value comes from discrete uniform distribution (*) it is referred to as *"random digit".* Interestingly, transcendental numbers like e, π,… are generators of random digits, in the sense that the sequences of their digits have statistical properties of a sequence of independent, identically distributed random variables having uniform distribution (*).

A single value of the uniform distribution *U(0,1)* is referred to as *"random number",* and shall be denoted as *γ*. More precisely, Monte Carlo methods do not use random numbers, but use pseudo-random numbers, i.e. series of numbers from the interval *(0,1),* having statistical properties of random sample from the uniform distribution. These numbers are generated by computers using some appropriate formulae. The usual notation in computer programmes are RAND or RND for such numbers. The accuracy of the Monte Carlo methods generally improves with the increase of the pseudo-random numbers used.

When modelling random variables the properties of the uniform distribution given in the following theorem are often used.

Theorem 3.

1) *If X: U(0,1), then 1−X: U(0,1),*
2) *If the random variable X has the distribution function F(x), then random variable F(X) is uniformly distributed on the interval [0,1], i.e. F(X): U(0,1) and*
3) *If X: U(0,1), then a· X+b is a r.v. U(a,b).*

Methods for modelling random variables can be summarized as follows:

a) Let *X* be a discrete random variable (r.v.) with distribution

$$\mathrm{X}:\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ p_1 & p_2 & p_3 & \cdots & p_n \end{pmatrix},\ \sum_{k=1}^{n} p_k = 1. \qquad (1)$$

Let γ be one (pseudo)random number. If $\gamma \le p_1$, then we assume that the value x_1 of r.v. *X* is realized. If $p_1 < \gamma \le p_1 + p_2$, then we assume that the value x_2 of r.v. *X* is realized. If $p_1 + p_2 < \gamma \le p_1 + p_2 + p_3$, then we assume that the value x_3 of r.v. *X* is realized, etc. This way for every realization of r.v. *X* one (pseudo)random number is used.

The same idea could be applied to model realizations of some *random event* using the corresponding r.v. Let *p* be the probability of some event *A*, and let us define r.v. $I_A:\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, so called *indicator* or *dummy variable*. We can obtain the realizations of this r.v. as described earlier, and if we obtain 1, we assume that *event A occurs* in an experiment, and if we obtain 0, we assume that event A didn't occur.

This is the general approach for modelling discrete r.v., while there are many special solutions depending on the nature of r.v. to be modelled. If we have to obtain realizations of a r.v. which has *binomial distribution* B(n,p), then we shall use the representation of this variable as a sum of iid indicators, rather than calculate the probabilities and apply the general procedure like the one described above.

Example 4. "Rolling a die"

a) We can model the outcomes of this experiment using continuous uniform

distribution. If we have a sequence of random numbers, say: 0.23, 0.32, 0.98, 0.85,… they give the results: 2, 2, 6, 6, … (since $1/6<0.23\le 2/6$, the outcome is "2", etc…)

b) Using discrete uniform distribution is another way to model this experiment.

Let us have a sequence of random digits: 2, 3, 3, 2, 9, 8, 8, 5,…, then we obtain the results: 2, 3, 3, 2, 5,… (0,7,8 and 9 obviously cannot be used)

b) Let *X* be a *discrete r.v. with an infinite set of values*

$$X:\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n & \cdots \\ p_1 & p_2 & p_3 & \cdots & p_n & \cdots \end{pmatrix},\ \sum_{k=1}^{\infty} p_k = 1.$$

Let n be an integer such that $p_{n+1} + p_{n+2} + ... < \delta$, where δ is an arbitrary chosen small positive number. Then, instead of r.v. X we use

$$X_Z:\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ p_1^* & p_2^* & p_3^* & \cdots & p_n^* \end{pmatrix},$$

where $p_1^* = p_1, ..., p_{n-1}^* = p_{n-1}$, $p_n^* = 1-(p_1 + ... + p_{n-1})$. Realizations of this r.v. X_Z (*truncated r.v.* X) are modelled using the procedure described in a) and are taken as realizations of the r.v. *X*. In this way we could not obtain any of the values $x_{n+1}, x_{n+2}, ...$, but the probability for the r.v. *X* to have some of these values is less than δ, which is chosen to be small, say δ=0.001 or similar, depending on the purpose of the simulation.

c) Let *X* be a continuous r.v. with the distribution function *F(x)*, and let this function have its inverse. The realizations of the r.v. *X* could be modelled using the following theorem:

Theorem 4. *Let X be an r.v. which distribution function F(x) is continuous, strictly monotone with the inverse* F^{-1}. *Let the r.v. Y have a uniform distribution U(0,1). Then the r.v.* $F^{-1}(Y)$ *has the distribution function F(x).*

In this case one realization *x* of a modelled r.v. *X* is obtained using one (pseudo)random number γ, from the equation $x = F^{-1}(\gamma)$.

Example 5. As an example of application of the theorems 3 and 4, we give the formula for modelling *exponential distribution*. If X: $\varepsilon(\lambda)$, then its modelled value is $x = \frac{-1}{\lambda} ln(\gamma)$.

This method, with some small modifications, could be applied if the distribution function is not continuous or has constant value on some intervals.

d) Let X be continuous r.v. with probability density function (p.d.f.) g bounded and defined on a finite interval. Realizations for such an r.v. could be modelled using so-called *rejection method* based on the following theorem:

Theorem 5. *Let the p.d.f. g(x)of an r.v. X is defined on a finite interval* (α,β) *and let there is a positive number M such that* $g(x) \le M, x \in (\alpha, \beta)$. *Let* x_T *and* y_T *are modelled realizations of r.v. U(α,β) and U(0,M), in this order. If* $y_T < g(x_T)$ *holds, then* x_T *is realized value of r.v. X.*

If the inequality $y_T < g(x_T)$ does not hold, then another pair, x_T and y_T, should be modelled. The realizations of r.v. *U(α,β)* and *U(0,M)* could be obtained using the relationship $a \cdot \gamma + b$ which gives one value for a r.v. *U(a,b)* if γ is a value for r.v. *U(0,1)*.

The number of (pseudo)random numbers used in the rejection method procedure is not predictable.

The method is due to Neumann, and, with some changes could be applied for any p.d.f. The idea is to find another r.v. which coincides with the variable to be modelled, say r.v. X, with the probability close to 1. Let X have the p.d.f. $g(x)$, $a<x<b$, where a could be $-\infty$, and/or b could be ∞, and $g(x)$ must not be bounded on the interval (a,b). In any case we have $\int_a^b g(x)dx = 1$. Assume that there is a finite interval (a',b'), such that $(a',b') \subset (a,b)$ where $g(x)$ is bounded. If one interval of the kind is not enough we take a union of such intervals. And now let define r.v. X_Z

(truncated r.v. X) such that $X_Z = X$ on (a',b'), and 0 elsewhere. The p.d.f. $g_Z(x)$ of r.v. X_Z satisfies

$$g_Z(x) = c \cdot g(x) = \left[\int_{a'}^{b'} g(t)dt \right]^{-1} \cdot g(x),\ a' < x < b'.$$

The constant c is greater than 1, and $g_Z(x) > g(x)$ holds on (a',b'). The interval (a',b') is chosen in the way to satisfy $1 - \int_{a'}^{b'} g(x)dx < \delta$, where δ is an arbitrary small positive number.

The procedure of modelling the r.v. *X* is based on modelling values for X_Z using the rejection method, and taking these values as values of r.v. *X*, since *X* and X_Z coincide with the probability $1-\delta$.

e) Let *X* be a *normally distributed* r.v. The inverse function method cannot be applied, while the rejection method can, but is not usually used. The procedure of modelling values for a normally distributed r.v. is based on a *central limit theorem*, and uses (pseudo)random numbers, as we shall explain in what follows. Let $Y_1, Y_2, ...$ be independent r.v. with the uniform distribution *U(0,1)*. The sum $S_n = \sum_{j=1}^{n} Y_j$ has expectation and variance $E(S_n) = \frac{n}{2}$, $D(S_n) = \frac{n}{12}$, and following the central limit theorem for the r.v.

$$S_n^* = \frac{S_n - E(S_n)}{\sqrt{D(S_n)}} = \sqrt{\frac{3}{n}} \sum_{j=1}^{n} \left(2Y_j - 1\right)$$

holds

$$P\left\{S_n^* < x\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt,\ n \to \infty.$$

This *convergence in distribution* is fast and with *n=12* the difference between S_n^* and *N(0,1)* is small enough, so we can take $\xi^{(12)} = \sum_{j=1}^{12} \gamma_j - 6$ as a realized value for r.v. with *N(0,1)* distribution. In this way, one needs 12 (pseudo)random numbers for one realization of r.v. with *N(0,1)* distribution. It should be noticed that even *n=6* gives good result. More accurate normal distribution is obtained with: $S_n^* + \frac{1}{20n}\left((S_n^*)^3 - 3S_n^*\right)$, for all n.

There are many other methods for modelling normal distribution, and we give here one of these procedures, which uses uniform distribution:

Let α_1 and α_2 be independent r.v. with U(0,1). Then ξ and η :

$$\xi = \sqrt{-2\log\alpha_1}\cdot\cos(2\pi\alpha_2)$$
$$\eta = \sqrt{-2\log\alpha_1}\cdot\sin(2\pi\alpha_2)$$

are independent r.v. having N(0,1) distributions. This procedure could be generalized to the *n-dimensional case*.

6. SAMPLING AND RESAMPLING

When we have to obtain a *random sample with replacement* of *size n* from a finite population with N elements, then we can model n realizations of r.v (1) with values 1, 2, ..., N, as described before. If we need the *sample without replacement* we simply do not take more than once the values we obtain.

We use the method for modelling discrete random variables also in *bootstrap testing procedures*, when we have to obtain a random sample with replacement of a given size n_1 using given sample of size *n*. We take the elements of a given sample $x_1,\ldots,x_n$, as values of r.v. *X* in (1), no matter if there are repeated values, and all the probabilities p_j, *j=1,n* equal *1/n*.

Example 6. Let's have the sample 2, 3, 5, 4, 4. The corresponding discrete r.v. X has the distribution P(X=2)=P(X=3)=P(X=5)=1/5, and P(X=4)=2/5.

Let's have random numbers: 0.23, 0.42, 0.15, 0.78, 0.91, 0.32, 0.47, 0.11, 0.77, 0.08.

"New" sample of size 10 will be: 3, 5, 2, 4, 4, 3, 5, 2, 4, 2.

7. UNIFORM DISTRIBUTION AND TESTING PROCEDURES

If one wants to *test* the *hypothesis* that the data are from uniform distribution *U(a,b)*, where *a* and/or *b* are unknown then *nonparametric test* like *χ2 test* could be applied, or, if *a* and *b* are known, the *Kolmogorov test* is also applicable. It is possible also to use *parametric test* and compare the *(truncated r.v. X) power* of these tests.

In testing *random number generators* some of the tests use discrete uniform distribution as distribution to be fitted. This is the case of *test of frequencies* (when the distribution (*) is the distribution to be fitted), *test of pairs, test of triples*…

In the study of *income distribution* M.O. Lorenz introduced a new function, later named after him: Lorenz's line or Lorenz's curve. This curve and *the coefficient of concentration* which is related to the Lorenz's curve may be used in statistics as a base for *scale-free goodness of fit test*. Gail and Gastwirth developed in 1978 a scale-free goodness of fit test for the *Laplace distribution*, and they extended these results for testing exponentiality. Here we give the possibility to fit the given distribution not only to the *U(0,1)* distribution, but to any other distribution, using the properties of the uniform distribution.

Let *X* be a *nonnegative r.v.* with the values in *[0,a)*, where *a* could be infinite, and let the *expectation* $\mu=E(X)$ exist. Let *F(x)* denote the distribution function of *X*. We shall assume that function strictly increasing and *continuously differentiable*, and denote *G(x)* its inverse. Lorenz's curve is the function

$$L(p)=\frac{\int_0^{G(p)} t\,dF(t)}{\mu},\quad 0\le p<1.$$

One important property of the Lorenz's curve is that each distribution with finite mean uniquely determines its Lorenz's curve, up to a scale transformation (Thompson, 1976).

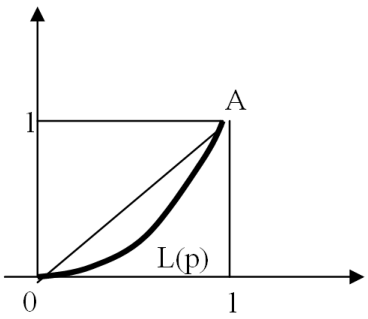

FIGURE 1 - LORENZ'S CURVE

The index of concentration for a population with the distribution *F(x)* and Lorenz's line *L(p)* equals the double surface between *L(p)* ant the segment *OA*, i.e.

$$C(X) = 2\int_0^1 (p - L(p))dp\ .$$

The corresponding *sample statistic*, often referred as the *Gini's index*, is

$$G = \frac{\sum_{i,j=1}^{n} \left| X_i - X_j \right|}{2n(n-1)\bar{X}_n},$$

where $(X_1,\ldots,X_n)$ is a random sample of size *n*, and $\bar{X}_n$ the *sample mean*.

The index of concentration for the uniform distribution is *1/3*, and if we want to test *the null hypothesis* H_o that the sample is drawn from *U(0,1)* distribution with a *significance level* α we may use the following procedure: first calculate the value *g* for the index of concentration based on the sample $(x_1,\ldots,x_n)$ and if $\left|g - 1/3\right| < \alpha$ do not reject the hypothesis

H_o. The *alternative hypothesis* H_1 is that the distribution is not uniform. Furthermore, if the r.v. X has the distribution function F, then the r.v. $Y=F(X)$ has the $U(0,1)$ distribution and the equality

$$P\big(X \in [a,b)\big) = P\big(F(X) \in [F(a), F(b))\big)$$

obviously holds for any real numbers a, b, $a<b$. This way, if we want to test the hypothesis H_o that X has the distribution function F_o based on the sample $(X_1,\ldots,X_n)$, then we can use the index of concentration and test the hypothesis that $F_o(X)$ has the uniform $U(0,1)$ distribution for a given level of significance α.

8. REFERENCES

1. Dagum, C. (1985) *Lorenz curve*, in Encyclopaedia of Statistical Sciences, Vol. 5, S. Kotz; N. L. Johnson and C. B. Read (editors), p. 156-161, New York, J. Wiley
2. Gail, M.H. and Gastwirth J.L. (1978) *A scale-free goodness of fit test for the exponential distribution based on Gini statistic*, J. R. Statist. Soc. B, 40, No. 3, p. 350-357
3. Gail, M.H. and Gastwirth J.L. (1978) *A scale-free goodness of fit test for the exponential distribution based on the Lorenz curve*, J. Amer. Statist. Assoc., 73, No. 364, p. 787-793
4. Jevremovic, V. (1998) *The index of concentration and the goodness of fit*, Compstat 98, XII Symposium on Computational Statistics, Bristol, England, 24-28 August 1998
5. Hogg, R.V., McKean J.W., Craig, A.T. (2005) *Introduction to Mathematical Statistics*, Pearson Education International, printed in USA
6. Larsen, R. J., Marx, M.L. (2006) *An Introduction to Mathematical Statistics and Its Applications*, Pearson International Edition International, printed in USA
7. Соболь, И. И. (1973) *Численные методы Монте Карло*, Наука, Москва (in russian)

8. Đorić, D., Jevremović, V. i drugi (2007) *Atlas raspodela*, Građevinski fakultet, Beograd (in serbian)
9. Jevremović, V. (2010) *Uniform Distribution in Statistics*, International Encyclopedia of Statistics (red. M. Lovrić), Springer