SOCIETY. INTEGRATION. EDUCATION
Proceedings of the International Scientific Conference. Volume IV, May 22th -23th, 2020. 602-611

DATA GENERATION FOR THE ACQUISITION OF THE UNIVERSITY COURSES OF INFORMATICS AND STATISTICS IN THE HEALTH CARE SPECIALITIES

Oskars Rasnacs
The EKA University of Applied Sciences, Latvia

Maris Vitins
University of Latvia, Faculty of Computing, Latvia

Abstract. *The authors of the present article are investigating the influence of the data generation in the university courses of informatics and statistics (UCIS) on the acquisition of both the UCIS and other study courses in the health care specialities. First of all, the authors inquired students in order to find out their attitude to the UCIS. The inquiry results show evidence that an important role in the acquisition of UCIS has the work/study material – data. The UCIS work/study material can be associated with various branches, including health care. The data of the health care patients are of a special status. Their use is strictly regulated/limited.by legislation. The authors of the present article offer an acceptable solution – data associated with the health care patients are generated from the parameters of statistics of scientific publications. Investigations were performed in the Red Cross Medical College of Rīga Stradiņš University (RCMC of RSU). It was found out that the data generation resulted in higher marks both in UCIS and many other study courses. In this article, the authors present proposals how to generate data comparatively easy, apply traditional MS Excel generation tools as well as tools of Goal Seek and Solver.*
Keywords: *data generation, health care, study courses, university informatics and statistics courses.*

Introduction

The authors' investigations show that the student attitude to UCIS is neither expressed positive nor expressed negative – it is rather neutral. The authors are sure that the student attitude becomes more positive if the UCIS work/study material – data – is associated with the field of health care; whereas investigations of other scientists show evidence that the marks in UCIS has an important influence on the marks in other study courses. All that indicates the importance of data generation.

Research questions:

What is the UCIS achievement of the set results?

What is the comprehensibility of the UCIS study content?

http://dx.doi.org/10.17770/sie2020vol4.4941

What is the student's contribution to the UCIS acquisition?

What is the influence of data generation on the assessments in other study courses?

What is the influence of the UCIS assessments on the marks in other study courses?

Material and methods

In order to find out the students attitude to UCIS, the authors asked them 31 questions during the period of time from 2012 to 2019. Twenty-one students were from the RCMC of RSU, 10 students from the Riga Technical University. There is specialty "Medical physics and engineering" in the Riga Technical University. The graduates of this specialty will work with medical technology.

Research questions:

What is the UCIS achievement of the set results in the 5-point scale? (Self-evaluation)

What is the comprehensibility of the UCIS study content in the 5-point scale? (Self-evaluation)

What is the student's contribution to the UCIS acquisition in the 5-point scale? (Self-evaluation)

Regarding the data generation and the influence of UCIS on the knowledge evaluation of other courses, the authors surveyed the RCMC of RSU graduates database with the knowledge evaluation of 366 graduates of treatment and 427 graduates of nursing speciality study courses.

Results

A median table was obtained after inquiring 31 students (Table 1). The obtained results show that it is necessary to improve comprehension of the study content and to achieve a more substantial student contribution to the course acquisition.

Table 1 ***Medians of variables***

Variable	Summary	Health care (n=21)	Medical technology (n=10)
Achieving the results set by UCIS	4	4	4
UCIS comprehensibility	3	3.5	3
Student contribution to UCIS acquisition	3	3	3

n- number

The following results were obtained after processing the database of the RCMC of RSU graduates. The treatment and nursing specialities have many different study courses that is why the calculation of correlation was done for each speciality separately depending on the number of generated data sets in the course. In the first column (Table 2), the Spearmen correlation coefficient r is given, in the second column there is the Spearmen coefficient p value, and in the third column the number of pairs involved in the Spearman correlation n. Statistically significant and positive correlations are printed in bold because the authors were interested in which study courses the evaluation increased by increasing the number of generated data sets. For example, by increasing the number of generated data sets in the treatment speciality, the evaluation of knowledge increases in the humanities (r=0.199, p<0.001, n=346). In total, there are many study courses, therefore only some of the correlation results are shown.

Table 2 ***Spearmen correlation coefficient between the numbers of generated data sets in the treatment speciality***

Variables	r	p	N
Humanities	**.199**	**<.001**	**346**
Medical terminology, English	.064	.266	302
Latin in medicine	**.388**	**<.001**	**194**
Sociology	.044	.421	336
Pedagogy	**-.110**	**.044**	**337**
Psychology	**.211**	**<.001**	**330**
Information technologies, statistics	-.007	.899	336
Anatomy, cytology and genetics	.001	.981	358

r - Spearmen correlation coefficient
p – p value of Spearmen correlation coefficient
n- number

For example, by increasing the number of generated data sets in the nursing speciality, the evaluation of knowledge increases in philosophy of care taking (r=0.363, p<0.001, n=422). In total, there are many study courses, therefore only some of the correlation results are shown (Table 3).

The level of comprehension correlates with the level of contribution (Spearmen correlation coefficient r=0.568, p=0.001, n=31). By increasing the level of comprehension, the student would be more interested in working, and the contribution level would also increase. A comparatively lower level of comprehension of students of medical technology speciality can be explained by the fact that programming is included in the course for all students. Programming for students of health care is actual only then if the student qualifies for a higher mark than 8.

Table 3 ***Spearmen correlation coefficient between the numbers of generated data sets in the nursing speciality***

Variables	r	p	N
Philosophy of care taking	.363	<.001	422
Natural sciences	-.145	.003	420
Psychology and sociology	.202	<.001	416
Entrepreneurship	.162	.010	255
Information technologies and statistics	.165	.001	409
Clinical procedures in practice of medical nurses, radiology	.144	.003	417
Medical terminology, English	.181	<.001	366
Latin in medicine	.214	<.001	415

r - Spearmen correlation coefficient
p – p value of Spearmen correlation coefficient
n- number

The data of the health care patients are of a special status. Their use is strictly regulated/limited.by legislation. The authors of the present article offer an acceptable solution – data associated with the health care patients are generated from the parameters of statistics of scientific publications (Aviñó, Ruffini, & Gavaldà, 2018; Hartmane, Mikazans, Ivdra, & Derveniece, 2018; Janssen & Dundurs, 2018; Kalnina, Selga, Sauka, & Larins, 2018; Mickevica, Margaliks, & Mamaja, 2018). In order to increase the comprehension level, the authors recommend to generate the data. Generated data can be used in the study process for both health care and medical technology students. For generating the data you can use the traditional MS Excel tool Data/Data Analysis/Random Number Generation. The most popular data distributions acquired in the statistics course are Discrete and Normal (Nelson & Nelson, 2014).

First of all, let us generate the discrete random variable data according to the article (Heidemann et al., 2019). In the article, 74% of respondents are men, and 26% are women. We will generate data for 3,000 imaginary patients. Allocate the MS Excel A column for generating the data. Design the MS Excel sheet (Figure 1).

	A	B	C	D
1	Data	Gender	Labels	Percent
2		Male	0	0.74
3		Female	1	0.26
4				=SUM(D2:D3)

Figure 1 ***MS Excel sheet design for generating the discrete random variable***

Fill in the Data/Data Analysis/Random Number Generation window (Figure 2).

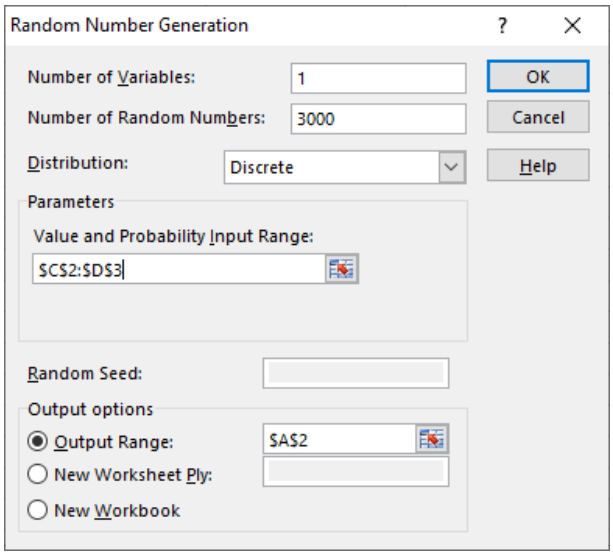

Figure 2 ***Filling in the Data/Data Analysis/Random Number Generation window for generating the discrete random variable***

In the column A, 0 and 1 values are generated, in total 3,000.

Let us look at normally distributed data generation according to the article (Tanwi, Shashank, & Kishwar Hayat, 2013). Almost all LDL values are in the interval of [25; 200] mg/dL. By the 3 sigma rule, almost all normally distributed data values are in the mean interval of ± 3 std. deviations. So the desired mean of the generated data is (25+200)/2=112.5 and std. deviation is (200-25)/6=29.2. Fill in the Data/Data Analysis/Random Number Generation window (Figure 3).

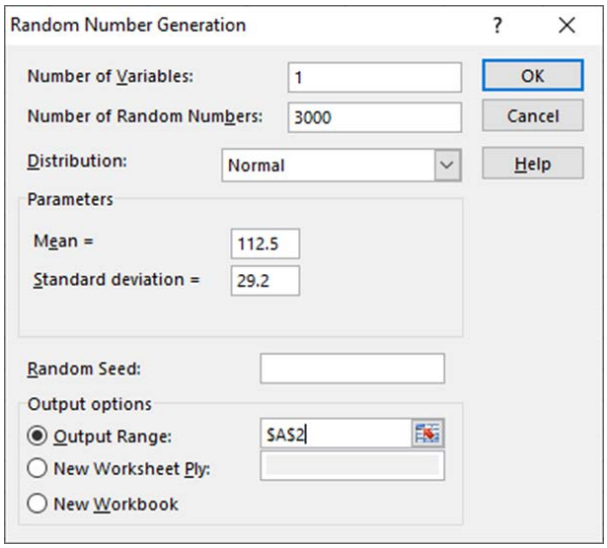

Figure 3 ***Filling in the Data/Data Analysis/Random Number Generation window for normal distributed variable generation***

In the column A, totally 3,000 values are generated. It is recommended to round up the data values for further use with the function =ROUND(A2;0) (if to integer number). In order to avoid false – negative values, a minimum value must be calculated for the generated data =MIN(A2:A3001). If the minimum is negative, then the minimum values should be replaced by positive numbers. For further analysis, copy data as values – Paste Special/Values.

Data can also be generated by using the MS Excel tools Goal Seek and Solver. The Goal Seek tool is used to find x value for the one-argument equation f(x)=y, if x is known. The Solver tool has many uses. It can solve systems of equations and inequalities, and optimize functions. In the simplest case, you can also solve the same tasks as Goal Seek.

Let us solve the equation 2x+6=0 with Goal Seek and Solver. First, fill in the MS Excel sheet (Figure 4). In the cell B1, enter a random start value for x, for example 4 (Saleh & Latif, 2008; Guerrero, 2010; Chandrakantha, 2014; Wray, 2015; Ezeokwelume, 2016; 15.053. Excel Solver [15053ES]; Excel Goal Seek Function [EGSF]; An Introduction to Spreadsheet Optimization Using Excel Solver [ISOUES]).

	A	B
1	x	4
2	y	=2*B1+6

Figure 4 ***Design of the MS Excel sheet for solving the equation 2x+6=0***

After that, fill in the Goal Seek window. Data/What-If Analysis/Goal Seek (Figure 5). In the cell B1, the solution -3 is recorded (Guerrero, 2010; EGSF).

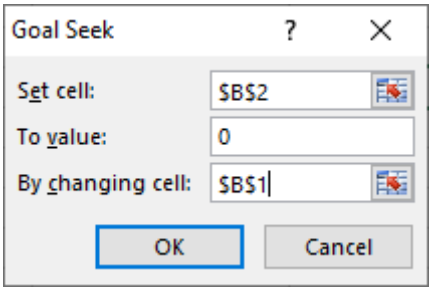

Figure 5 ***Filling in the Data/What-If Analysis/Goal Seek window for solving the equation 2x+6=0***

The Solver tool sheet is designed in the same way as Goal Seek (Figure 5). Then, fill in the Solver window. Data/Solver (Figure 6). If a negative solution to the problem is possible, then the option “Make Unconstrained Variables Non-

Negative" should be turned off. To solve linear problems, select "Simplex LP" from Select a Solving Method list. In the cell B1, the solution -3 is recorded (Saleh & Latif, 2008; Guerrero, 2010; Chandrakantha, 2014; Wray, 2015; Ezeokwelume, 2016; 15053ES; ISOUES).

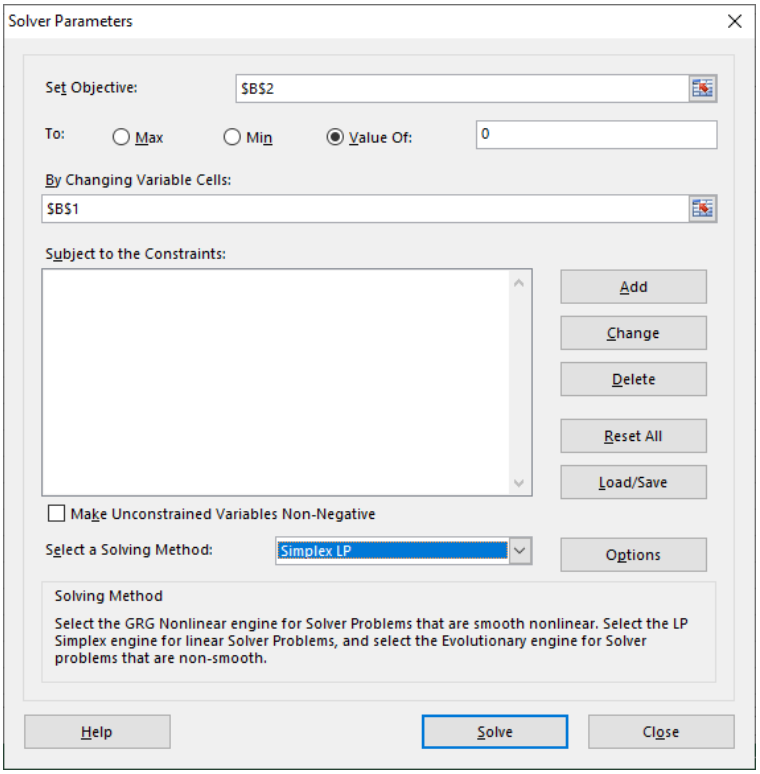

Figure 6 ***Filling in the Data/Solver window for solving the equation 2x+6=0***

Now, use Goal Seek to generate the data. Let us generate 3,000 LDL data, starting with a value of 25 and ending with a value of 200. Let us suppose the data change linearly with a step x. The task is to find the step x. Design the MS Excel sheet (Figure 7). In the column A, the row of formulas continues to the cell A3001. In the cell B, like before, enter a random start value for x, for example 4 (Guerrero, 2010; EGSF).

	A	B	C
1	Data	x	4
2	25	y	=MAX(A2:A3001)
3	=A2+C1		

Figure 7 ***Designing the MS Excel sheet for generating data with Goal Seek***

Fill in the Goal Seek window (Figure 8). The step is calculated approximately 0.06. Similar to generating normally distributed data, column A should be rounded to the desired number of characters (Guerrero, 2010; EGSF).

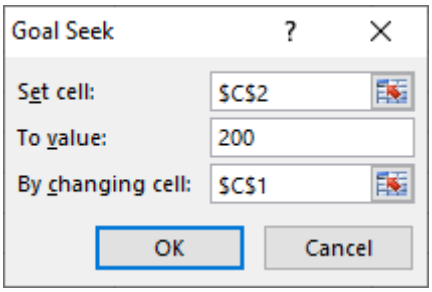

Figure 8 ***Filling in the Data/What-If Analysis/Goal Seek window for data generation***

Design the Solver tool sheet in the same way as for Goal Seek (Figure 7). After that, fill in the Solver window Data/Solver (Figure 9) (Saleh & Latif, 2008; Guerrero, 2010; Chandrakantha, 2014; Wray, 2015; Ezeokwelume, 2016; 15053ES; ISOUES).

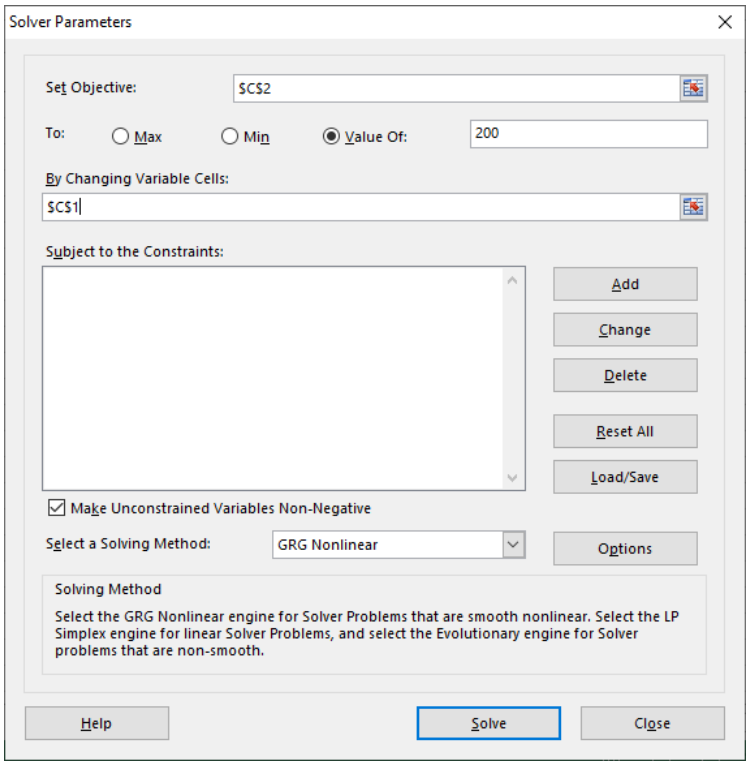

Figure 9 ***Filling in the Data/Solver window for data generation***

The authors of the present article have discussed the simplest solutions of data generation. These solutions are possible to make more complex in various ways, for example:

- Introduce a set cell and minimize it so that the parameters of statistics do not differ much from the desired ones.
- For the non-linear data sequences, introduce the non-linear functions in data generation.

Conclusions

The authors' research has proved that:

- the data generation improves the evaluation of other study courses;
- the data generation improves the UCIS evaluation.

Generating larger amounts of data improves students' attitude towards UCIS, which in turn contributes to the achievement of the set goals. A larger number of generated data sets allow the student to practise more and understand the study content.

The data generating is necessary to make the UCIS more interesting and understandable, and also to improve the students' knowledge evaluation in other study courses.

In the nursing speciality, the number of generated data sets has a more positive influence on the evaluations of courses than those of the treatment speciality. The authors could not find explanation other than the fact that there were more data on the nursing graduates.

Acknowledgements

The research was developed under the University of Latvia contract no. AAP2016/B032 "*Innovative information technologies*".

References

15.053. Excel Solver. Retrieved from http://web.mit.edu/15.053/www/Excel_Solver.pdf

An Introduction to Spreadsheet Optimization Using Excel Solver. Retrieved from http://www.meiss.com/download/Spreadsheet-Optimization-Solver.pdf

Aviñó, L., Ruffini, M., & Gavaldà, R. (2018). *Generating Synthetic but Plausible Healthcare Record Datasets*. arXiv preprint arXiv:1807.01514.

Chandrakantha, L. (2014). Using Excel Solver In Optimization Problems. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/267557388_USING_EXCEL_SOLVER_IN_OPTIMIZATION_PROBLEMSExc el Goal Seek Function. Retrieved from http://www.ce.memphis.edu/1112/notes/excel/Excel_goal_seek.pdf

Ezeokwelume, O.V. (2016). Solving Linear Programming Problems and Transportation Problems using Excel Solver. *International Journal of Scientific & Engineering Research, 7,* 9. Retrieved from https://www.ijser.org/researchpaper/Solving-Linear-Programming-Problems-and-Transportation-Problems-using-Excel-Solver.pdf

Guerrero, H. (2010). *Excel Data Analysis. Chapter 9 Solver, Scenarios, and Goal Seek Tools.* Springer – Verlag Berlin Heidelberg. Retrieved from http://jlarrosa.tripod.com/solver.pdf

Hartmane, I., Mikazans, I., Ivdra, I., & Derveniece, A. (2018). Evaluation of Clinical Efficacy and Safety in Treatment of Patients with Moderate and Severe Forms of Psoriasis with Combined Low Dose Methotrexate and Narrow Band UVB Therapy. *Rīga Stradiņš University, Collection of Scientific Papers 2017, Research articles in medicine & pharmacy.* Retrieved from https://www.rsu.lv/sites/default/files/book_download/rsu_research_articles_med_pharm_2017.pdf

Heidemann, B.E., Koopal, C., Bots, M.L., Asselbergs, F.W., Westerink, J., & Visseren, F L.J. (2019). Remnant cholesterol increases the risk for recurrent vascular events independent of LDL-cholesterol in patients with clinical manifest vascular disease. *European Heart Journal, 40,* Issue Supplement_1. DOI: https://doi.org/10.1093/eurheartj/ehz746.0013

Janssen, I.J., & Dundurs, J. (2018). Influence of Noise in Ambulance Vehicles on Emergency Service Personnel in North Germany and Latvia. *Rīga Stradiņš University, Collection of Scientific Papers 2017, Research articles in medicine & pharmacy.* Retrieved from https://www.rsu.lv/sites/default/files/book_download/rsu_research_articles_med_pharm_2017.pdf

Kalnina, L., Selga, G., Sauka, M., & Larins,V. (2018). Assessment of Fat Mass Index and Fat-Free Mass Index in Young Athletes. *Rīga Stradiņš University, Collection of Scientific Papers 2017, Research articles in medicine & pharmacy*. Retrieved from https://www.rsu.lv/sites/default/files/book_download/rsu_research_articles_med_pharm_2017.pdf

Mickevica, E., Margaliks, M., & Mamaja, B. (2018). Safety and Efficacy of Narcotrend Controlled Sedation with Dexmedetomidine vs. Propofol during Elective Colonoscopy. *Rīga Stradiņš University, Collection of Scientific Papers 2017, Research articles in medicine & pharmacy*. Retrieved from https://www.rsu.lv/sites/default/files/book_download/rsu_research_articles_med_pharm_2017.pdf

Nelson, L.S., & Nelson, E.C. (2014). *Excel Data Analysis for Dummies*. Retrieved from http://excelpro.ir/wp-content/uploads/2015/12/Excel-Data-Analysis-for-Dummies.pdf

Saleh, S.A., & Latif, T.I. (2008). Solving Linear Programming Problems By Using Excel's Solver. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/332513460_Solving_Linear_Programming_Problems_By_Using_Excel's_Solver

Tanwi, P., Sashank, M., & Kishwar Hayat, K. (2013). Cholesterol: Genetic, Clinical and Natural Implications. *Research Journal of Pharmaceutical, Biological and Chemical Sciences*. Retrieved from https://www.researchgate.net/publication/257429583_Cholesterol_Genetic_Clinical_and_Natural_Implications/link/00b49525440cf8676d000000/download

Wray, B. (2015). *Excel Solver Tutorial: Wilmington Wood Products. Gebauer/Matthews*: *MIS 213 Hands-on Tutorials and Cases, Spring*. Retrieved from http://csbapp.uncw.edu/sibonac/mis313/docs/project10.pdf