INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 27, číslo 1, březen 2016

POUŽITÍ INTERAKTIVNÍCH VIDEÍ A TESTŮ PRO VÝUKU JAZYKA R VE STATISTICE

USE OF INTERACTIVE VIDEOS AND TESTS FOR TEACHING THE R LANGUAGE IN COURSES FOCUSED ON STATISTICS

Michal Běloch, Pavlína Kuráňová, Robert Skopal, Martina Litschmannová, Kateřina Janurová

Address: VŠB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, 17. listopadu 15, 708 33 Ostrava-Poruba

E-mail: michal.beloch@vsb.cz, pavlina.kuranova@vsb.cz, robert.skopal@vsb.cz, martina.litschmannova@vsb.cz, katerina.janurova@vsb.cz

Abstrakt: V současné době je stále složitější přilákat studenty k jakémukoliv předmětu. Vyučující proto musí řešit otázku, jak vyučovat svůj předmět a zároveň vzbudit ve studentech zájem o danou problematiku. Na VŠB-TUO je používán široký výběr komerčního software k výuce statistiky. Cílem projektu FRVŠ, který byl v roce 2015 na našem pracovišti řešen, je příprava výuky s využitím volně šiřitelného programového balíku R. Pro intuitivní využití metod z oblasti statistické analýzy využíváme grafické rozhraní RKWard. Interaktivní videa připravovaná v prostředí Adobe Captivate pomohou studentům učinit první krok při seznamování s RKWard. Tato videa využívají metody e-learningu s pokročilou interaktivitou pro přiblížení konkrétních příkladů diskutovaných během výuky statistiky. Videa po spuštění vyžadují aktivní účast studenta a bez provedení příslušných kroků ze strany studenta nepokračují. Pro lepší porozumění dané látce jsou vytvářeny testy ověřující orientaci v probírané problematice.

Klíčová slova: výuka, R, RKWard, Adobe Captivate, interaktivní videa.

Abstract: Currently, it is more and more difficult to attract the students and encourage them to study any subject. Contemporary educators face a difficult task: How to teach the students something and arouse their interest in the subject. The variety of commercial software is used at VŠB-TUO for teaching statistical courses. The aim of the FRVŠ project, that was solved at our department in 2015, is the preparation of the lectures using the freeware programming package R. To make the application of statistical analysis and its methods as intuitive as possible, we use the graphical interface of RKWard. Multimedia interactive videos in Adobe Captivate are being prepared to simplify the first steps with RKWard for the students. These videos

using e-learning content with advanced interactivity are designed for specific examples discussed during courses on statistical data analysis. After turning on, the videos require active participation of the student and they stop playing until the student performs certain steps. For better understanding of the subject, tests with randomly generated questions are created to verify the understanding of the practiced issue.

Keywords: teaching, R, RKWard, Adobe Captivate, interactive videos.

1. Introduction

Statistics is usually very difficult subject to grasp. The students have hard time understanding basic principles. Our focus is therefore to provide a plethora of tools to make learning this subject as easy as possible. Given current trends, using modern teaching methods and tools is required. It is also important to pass the latest methodology and tools onto the students. In the world of statistics, this translates to the use of programming language R, which has become the go to software for statistical data analysis. This approach provides another set of difficulties however. As mentioned, statistics is difficult in itself. Adding requirements to also learn a programming language complicate the matter even more. This is where we need to introduce the software in easy to use form. We approached this challenge in two steps. The first step is to introduce easy to use environment so the students do not need to concern themselves with syntax difficulties of R language [1], but give them the opportunity to access it, if they are interested. Second step is to use the interactive tools to provide a robust knowledge base so the students do not feel lost at any point of the statistical data analysis course. The answer for the first step is the interactive interface of RKWard [2], for the second step it is the collection of fully voice-commented e-learning videos created in Adobe Captivate further accompanied with the quizzes focused on testing of the theoretical knowledge of the students. The second step will be discussed in detail. It should also be mentioned that these videos are currently prepared for the Czech students and are available in this language only. We want to use these videos to teach the foreign students as well, therefore the conversion into the English language is planned too.

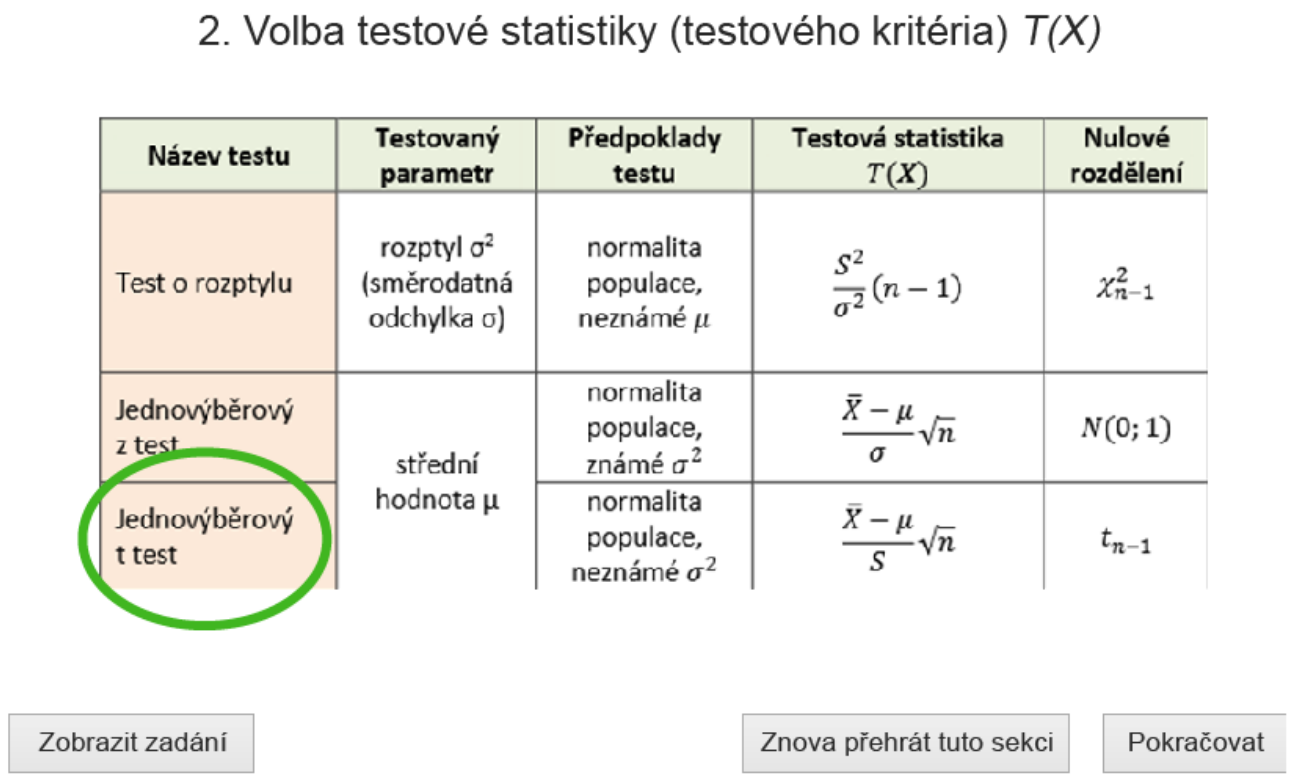
2. Creation of interactive videos and quizzes

The important consideration before the beginning of this project was to choose, which software is optimal for a video caption and editing. We ended up using the commercial version of Adobe Captivate 9, software specifically

designed for the creation of e-learning content. The advantage of Captivate produced files is the ability to view them on a vast array of devices. While the videos and quizzes are optimized to be viewed on a tablet and larger screens, it is still possible to access them via mobile phones.

2.1. Videos

Our goal is to make the educational videos as informative as possible, while remaining simple and easy to understand. To achieve this, it is important to choose a viable example to present. We also required the chosen example to be solvable "by hand". Usually when the students work with the software, they know where and what to click, but have problems understanding why they do so. Therefore in addition to providing the guidance through the RKWard environment for given example, the videos start with showing the principle of solution using "the old school" approach with the statistical tables. Both parts are fully voiced to provide commentary on the problem as well as giving the general guideline. Important facts are also graphically highlighted to draw the attention of the viewer. It is evident that before the recording itself, lots



2. Volba testové statistiky (testového kritéria) *T(X)*

Název testu	Testovaný parametr	Předpoklady testu	Testová statistika $T(X)$	Nulové rozdělení
Test o rozptylu	rozptyl σ^2 (směrodatná odchylka σ)	normalita populace, neznámé μ	$\frac{S^2}{\sigma^2}(n-1)$	χ^2_{n-1}
Jednovýběrový z test	střední hodnota μ	normalita populace, známé σ^2	$\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$	$N(0;1)$
Jednovýběrový t test		normalita populace, neznámé σ^2	$\frac{\bar{X}-\mu}{S}\sqrt{n}$	t_{n-1}

Zobrazit zadání

Znova přehrát tuto sekci

Pokračovat

Figure 1: Example of slide in Czech language. Buttons are used for easier navigation.

of the planning must be done. The scenario is required and its preparation is time consuming. Also the voice recording demands the attention. Finally the post production commands focus on details, as it is important to synchronize the voice and visual aspects of the video. This makes a time commitment to be the most noticeable disadvantage. On the flip side, the small video size and output in both SWF and HTML5 make the video easy to publish. Another high point is possible communication with the viewer.

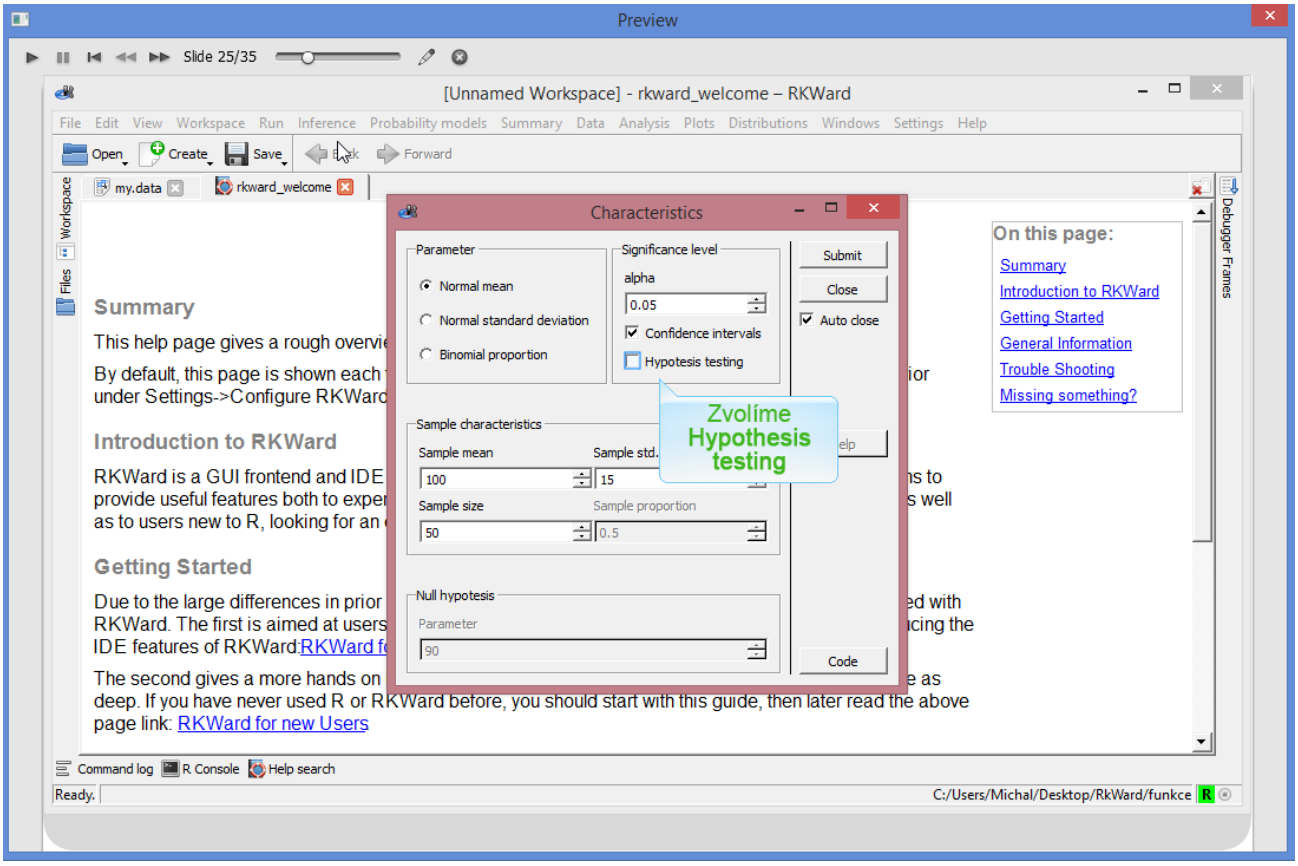


Figure 2: Example of the running video (comments are in Czech language). Currently discussed topic is highlighted to catch the attention of the viewer.

2.2. Quizzes

The quizzes started more as an afterthought before they were decided to become an important part of the project. The quizzes will be available to the students as a training material so they can verify their theoretical knowledge of various discussed topics and also as an entry test to each exercise.

The quizzes themselves consist of Multiple choice, True/False and Matching types of the questions. These questions are randomized from the predefined pool of the questions and offer variable experience each time the student

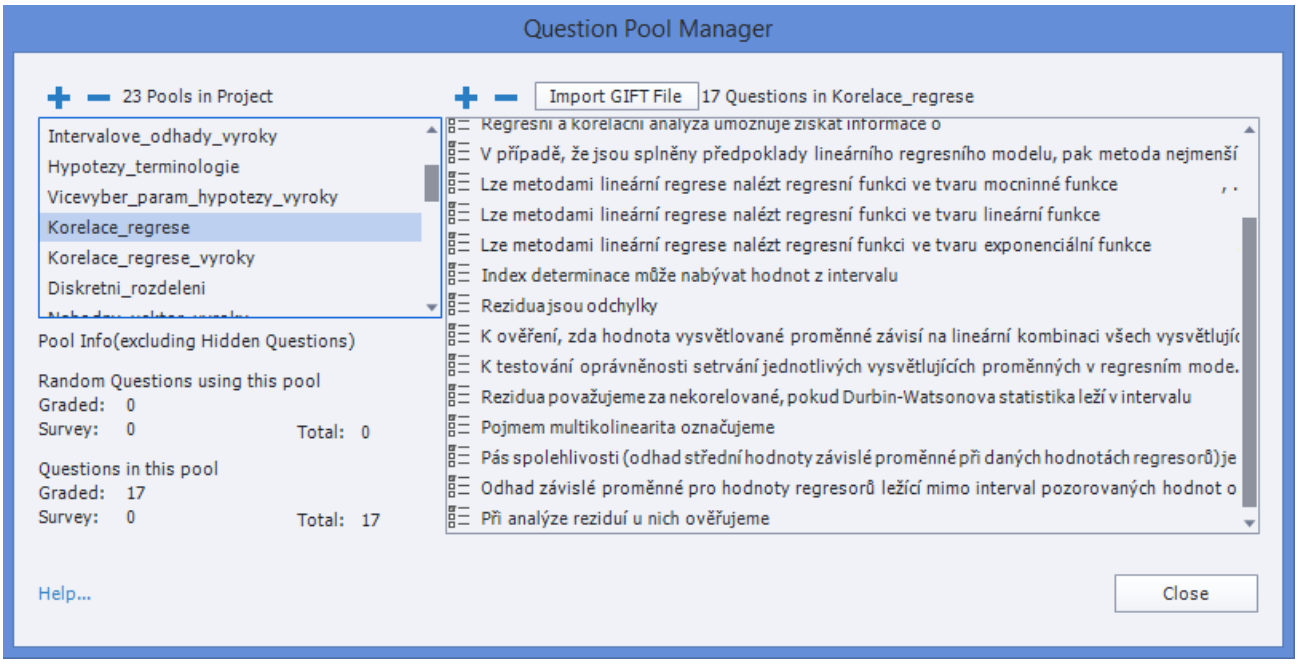


Figure 3: Question pool manager. The left column contains topics in a pool while the right column contains questions in a given topic (English version of the quizzes is currently being prepared).

decides to take a quiz. At the end of the quiz, the student is able to review his or her performance and take a look at his or her answers with the correct answers highlighted at places, where he or she made a mistake.

The main advantage is easy management of the questions. The questions – see Figure 3 – are organized in the topics and each topic can contain any number of the questions. It is also possible to set the minimum requirements for the student to pass. It is very important to think about this structure beforehand as a cleverly chosen structure can allow variable quizzes from the single topic or from any combination of the topics. Probably the greatest strength is the ability to link the quizzes with a Moodle learning system so the lecturer can review the performance of the students. The main disadvantage is a large room for errors so the creator of the questions must pay increased attention.

3. Conclusion

Our goal was to use the latest statistical software in our courses. The R language is currently the most favourite environment for a statistical data analysis. That is why we wanted to introduce it to our students. Because there is not any space left to teach the students a new programming language,

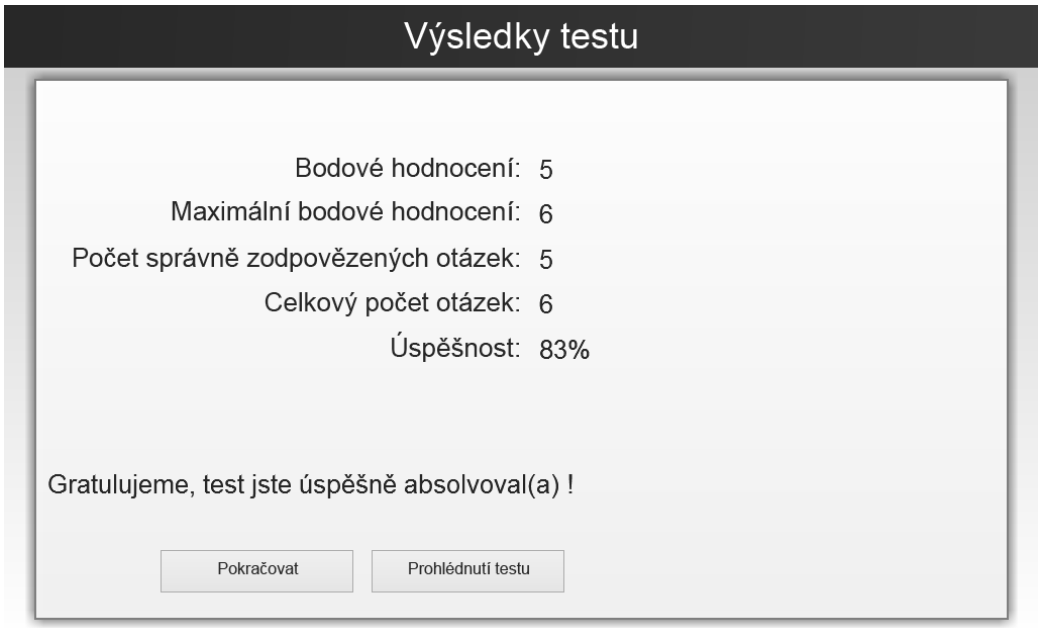


Figure 4: Review that is given at the end of each quiz (score, max score, correct questions, total questions, accuracy).

we instead opted for a much easier to use RKWard and we further optimized it for our purposes. To help the students with using the RKWard in solving the statistical problems, we created the interactive videos containing step-by-step instructions. These videos also show how to solve the said problems traditionally with the help of the statistical tables.

All of the available videos (in Czech) can be seen at `http://k470.vsb.cz/litschmannova/vyuka/statistika/studijni-materialy/videa/`.

Authors would be grateful for any remarks, comments and observations that could lead to improvements of these videos.

Acknowledgements

This work was supported by the SGS VSB-TUO Grant SP2015/84 and by the projects FRVS2015/57 and FRVS2015/81.

References

[1] R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2015, URL: `http://www.R-project.org`.

[2] RKWard: An Extensible IDE/GUI for R, RKWard Developer Team, 2015, URL: `https://rkward.kde.org/Main_Page`.

RKWARD: KOMPLEXNÍ PROSTŘEDÍ PRO VÝUKU STATISTIKY V JAZYKU R

RKWARD: A COMPLEX ENVIRONMENT FOR TEACHING STATISTICS WITH R

Robert Skopal, Kateřina Janurová, Michal Běloch, Pavlína Kuráňová, Martina Litschmannová

Address: VŠB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic

E-mail: `robert.skopal@vsb.cz`, `katerina.janurova@vsb.cz`, `michal.beloch@vsb.cz`, `pavlina.kuranova@vsb.cz`, `martina.litschmannova@vsb.cz`

Abstrakt: Během posledním několika let se programovací jazyk R stal jedním z nejoblíbenějších nástrojů pro statistické výpočty a vizualizaci dat. Největší překážkou ve využití tohoto jazyka při výuce statistiky je však absence GUI – grafického uživatelského rozhraní (současným stavem je rozhraní ovládané příkazovým řádkem). Kromě nového předmětu by tak studenti museli zvládnout i nový programovací jazyk, což by bylo během vyučovacích hodin značně časově náročné. V našem příspěvku představujeme možné řešení tohoto problému, bezplatné open-source prostředí RKWard. Jde o snadno přenosné a lehce rozšiřitelné GUI s pluginovou architekturou, která vyučujícím umožňuje přizpůsobit toto prostředí požadavkům konkrétního předmětu.

Klíčová slova: výuka statistiky, jazyk R, statistické GUI RKWard, pluginová architektura.

Abstract: Over the past years the R programming language has become more and more popular tool for statistical computing and data visualization. However, the absence of GUI – graphical user interface – represents significant disadvantage in using R for teaching statistics (the present status is a command line driven interface). Besides the new subject, the students would have to handle the new programming language as well. This could be quite time consuming during the lessons. We present in our paper a possible solution of this problem, free open-source RKWard environment for R developed by T. Friedrichsmeier, P. Ecochard, S. Roediger, P. Kapat and M. Michalke as an open source project. RKWard is portable and easily extensible GUI based on plugin architecture, what allows the teacher to adapt the GUI to the requirements of specific subject.

Keywords: teaching statistics, R language, statistical GUI RKWard, plugin architecture.

1. Background and motivation

R programming language is a software environment for statistical computing, data science and visualization that has been developed by Ross Ihaka and Robert Gentleman [1] and R Core Development Team [2] from 1996. R is available for free under an open-source license, platform independent, easily extensible, the most comprehensive and actively maintained by broad community. These and other benefits have lifted the R language to become one of the most popular tools used among statisticians and data miners.

On the other hand there are some significant disadvantages that prevent it from using in teaching statistics. It should be mentioned especially the absence of GUI – graphical user interface (at the present time R is a command-driven system) – that causes a steep learning curve for all beginners in statistical analysis. Another problem is the quality of some packages that are hardly perfect.

It follows that the GUI is needed for introductory statistical course using R language. The students should not have to remember an enormous amount of names and arguments of commands, thus eliminating the possibility of syntax and typing errors. The best GUI for teaching the beginners should be easy to use, with only the necessary functions (the commercial software usually includes a confusing labyrinth of commands), with analysis reproducible with one simple click (to reduce the time used for in lessons "clicking") and with extendable architecture. Simply the GUI adapted specifically for the needs of each statistical course.

The purposes of our paper are to present the GUI RKWard [3] which aims to meet all mentioned requirements, to encourage the teachers to use or even modify RKWard for their own teaching purposes and to direct the interested persons to the sources of further information [4, 5].

2. RKWard

RKWard is free open-source platform independent graphical user interface for the R programming language. It is easy to use, it can be easily modified for the specific needs of different courses of statistics from the introductory courses to the complex courses focused on a specific area of statistics and it combines the power of R language with the ease of use of commercial statistics environments.

RKWard contains:

- Spreadsheet-like data editor (with ability to copy and paste from other statistical software, e.g. MS Excel).
- Syntax highlighting, code folding and code completion.
- Data import (e.g. SPSS, Stata and CSV).
- Plot preview and browsable history.
- R package management.
- Workspace browser.
- GUI dialogs for a variety of statistical functions, analysis and plots.

Another advantage is easy installation process which is done just by unzipping the installation package into desired folder. The package contains the R language and all other packages and libraries necessary for RKWard itself which makes it easy to distribute the final software to students.

The detailed description of RKWard is given by Stefan Rödiger in [4] and by the web pages of RKWard Developer Team [3].

3. Modifying RKWard

3.1. Plugin framework

One of the key advantages of RKWard is the possibility of creating user plugins. These plugins are easy to made and the creation requires basic understanding of eXtensible Markup Language (XML) and JavaScript (or ECMAscript for technical precision). Using the procedure described in Section 3.2. one can arbitrarily extend the RKWard functionality.

3.2. Writing new plugin

The process of writing the new plugin is described in detail in [5] and can be summarized in these four steps:

1. Define new menu entry.
2. Create the graphical user interface (GUI).
3. Link GUI to R scripts/functions.
4. Write help page (optional).

1) First step is to define the new menu entry. In order to do so one has to define new so called *pluginmap* file. This file uses standard XML notation to define the structure of new menu entry. Figure 1 shows new menu entry called "Probability models" containing sets of Discrete and Continuous distributions added into RKWard.

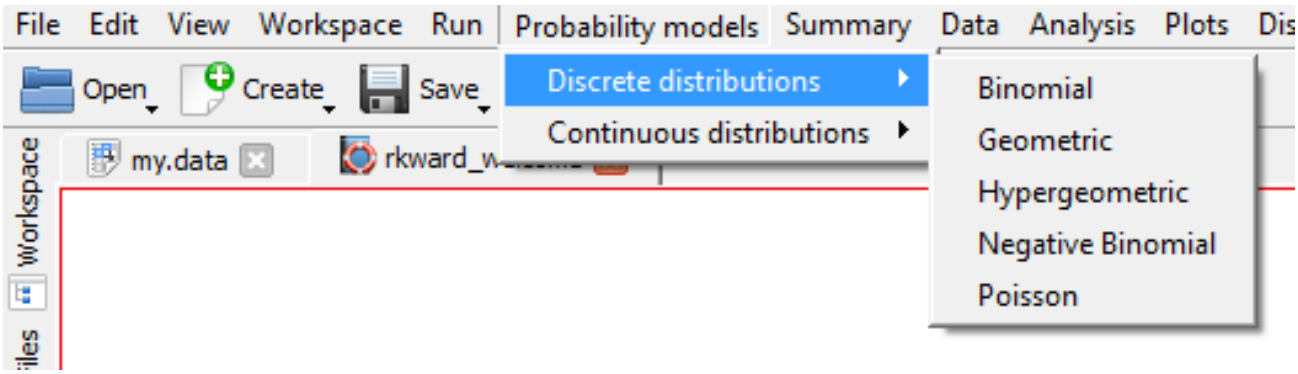


Figure 1: New menu entry – Probability models

In order do define menu entry structure the following (simplified) XML notation can be used in *pluginmap* file:

```
<menu label="Probability models">
<menu label="Discrete distributions">
<entry component="binomial_model"/>
<entry component="hypergeometric_model"/>
<entry component="negative_binomial_model"/>
<entry component="geometric_model"/>
<entry component="poisson_model"/>
</menu>
<menu label="Continuous distributions">
<entry component="exponential_model"/>
<entry component="weibull_model"/>
<entry component="normal_model"/>
<entry component="chisquare_model"/>
<entry component="students_model"/>
<entry component="fisned_model"/>
</menu>
</menu>
```

2) Next step after the menu entry definition is to create the GUI for every menu entry component. As in Step 1 the whole GUI is written in XML language. RKWard defines standard known GUI elements like `<frame>`,

`<tapbook>`, `<spinbox>` etc. Figure 2 shows our GUI for the Binomial distribution.

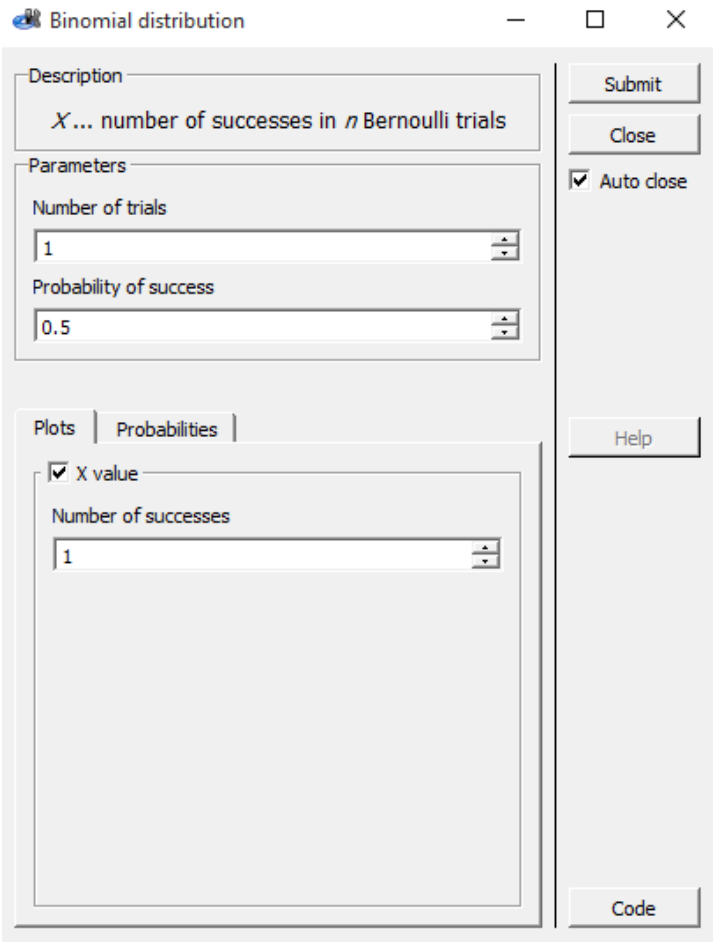


Figure 2: Binomial distribution GUI

For example the "Parameters" element in the Binomial distribution GUI is done using following (simplified) XML `<snippet>`:

```
<snippet id="dist_options">
<frame id="f_parameters_plots" label="Parameters">
<spinbox type="integer" initial="1" min="1"
       label="Number of trials"/>
<spinbox default_precision="4" type="real" initial="0.5"
        min="0" max="1" label="Probability of success"/>
</frame>
</snippet>
```

3) After the GUI is created the next part is to link the GUI to R scripts or functions. The JavaScript language is used for the linking part. The R

code can be called directly from the JavaScript using the `echo()` function. For example the code:

```
echo('X <- c(1,2,3)')
```

creates the standard one dimensional vector with entries 1, 2 and 3 in R. Nevertheless this approach is not convenient for more complex R computations because of the mixture of two programming languages. The much more convenient way is to write the R code in separate scripts and use the `source` function. For example the following code:

```
echo('source("myCode.R")')
```

invokes the R script stored in `myCode.R`.

The JavaScript file is organized into three parts (functions):

1. `preprocess()` part which collects and prepares input data,
2. `calculate()` part where all computations are done,
3. `printout()` part for the output.

Any of these parts is not strictly required. For example one can create a JavaScript file which only prints given output without any computations or preprocessing etc.

4) Last and optional step is to create a *rkhelp* file which is also coded in XML language. This file can contain both overall help for created plugin and help for individual GUI elements.

3.3. Modification of RKWard for the course of statistics at the VŠB-TU Ostrava

Currently we have created following plugins using procedure described in Section 3.2.:

Probability models – this plugin contains set of discrete (Binomial, Geometric, Hypergeometric, Negative-binomial, Poisson) and continuous (Exponential, Normal, Weibull, Chi-Squared, Student, Fisher-Snedecor) distributions. For every distribution the plugin is able to compute probabilities for given inputs and plot distribution function and probability mass/density functions. Figures 3 and 4 shows sample output for Binomial distribution.

Statistical inference – this plugin was created to provide confidence intervals and hypothesis testing for one or two samples. Both of these can be done for mean and standard deviation of normal distribution and Binomial proportion. In case of two samples there is a difference of Normal means

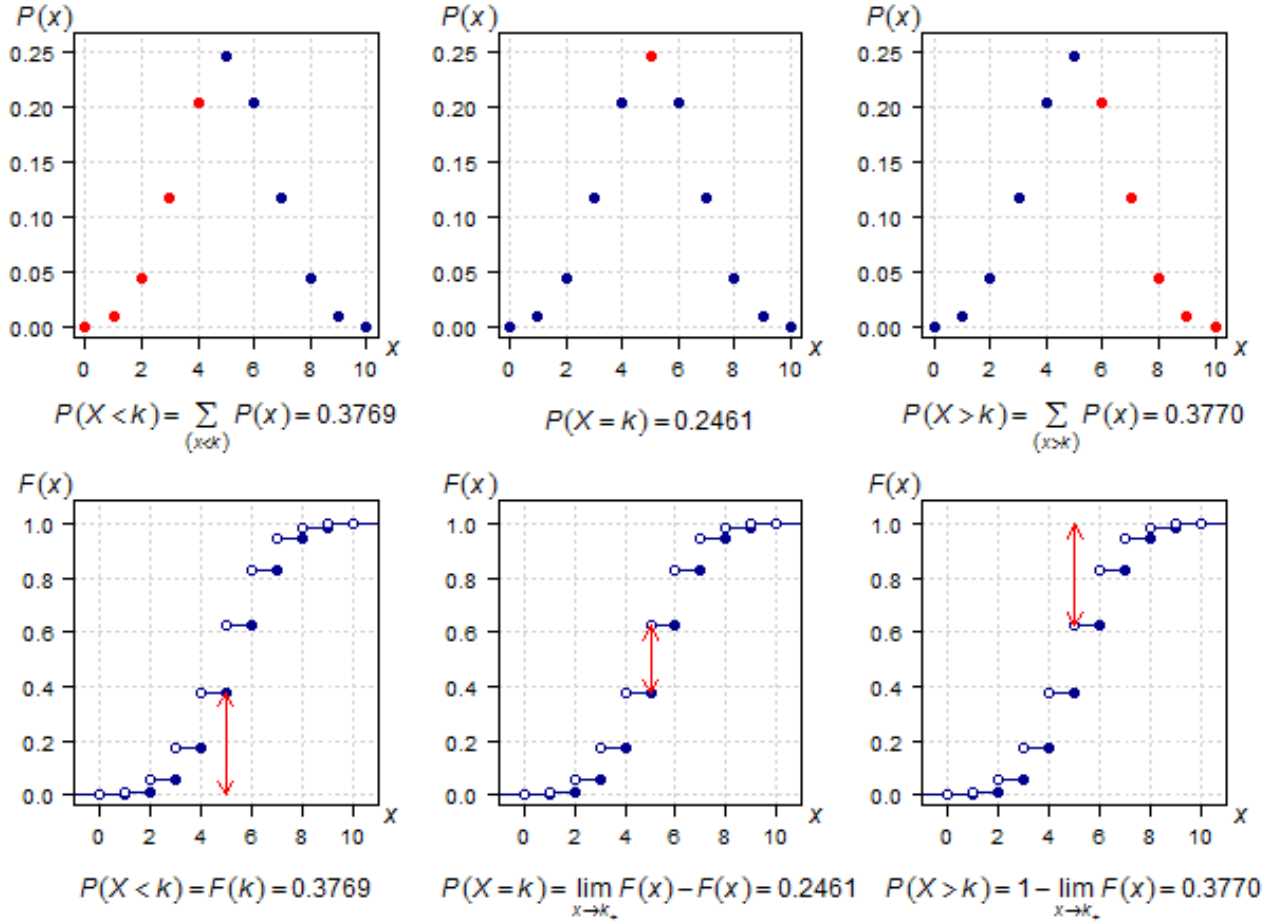


Figure 3: Sample graphic output for Binomial distribution

Probability

k	$P(X < k)$	$P(X = k)$	$P(X > k)$
1	0.0009	0.0098	0.9893
2	0.0108	0.0439	0.9453
3	0.0547	0.1172	0.8281

Figure 4: Sample table output for Binomial distribution

Hypotesis testing for mean of normal distribution

Null hypotesis: $\mu = 90.00$

Alternative hypotesis: $\mu < 90.00$

$x_{obs} = 4.71405$, $t_{krit} = -1.67655$, p-value = 0.99999

Alternative hypotesis: $\mu > 90.00$

$x_{obs} = 4.71405$, $t_{krit} = 1.67655$, p-value = 0.00001

Alternative hypotesis: $\mu \neq 90.00$

$x_{obs} = 4.71405$, $t_{krit1} = -2.00958$, $t_{krit2} = 2.00958$, p-value = 0.00002

Figure 5: Sample output for hypothesis testing

and Binomial proportions and ratio of Normal standard deviations. Figure 5 shows the output for the hypothesis testing for mean of Normal distribution.

4. Conclusion and future work

We presented in our paper the strengths and possibilities to use the free open-source graphical user interface for R programming language, the RKWard. We demonstrated its main advantage, the ability of easy extension, by example of some of our own modifications of RKWard, developed for the purposes of the introductory course of statistics at VŠB-TU Ostrava.

In the next few months we would like to finish the modification of RKWard by adding the plugins for non-parametric tests, contingency tables, ANOVA etc. to test this version by students in practice in Summer semester 2016.

Acknowledgements

This work was supported by the SGS VSB-TUO Grant SP2015/84 and by the projects FRVS2015/57 and FRVS2015/81.

References

[1] Ihaka, R., Gentleman, R.: R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5(3), pp. 299–314, 1996.

[2] R Core Development Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, 2004.

[3] RKWard Developer Team: *RKWard: An Extensible IDE/GUI for R*, 2015. URL: `https://rkward.kde.org/Main_Page`

[4] Rödiger, S., et al.: RKWard: A Comprehensive Graphical User Interface and Integrated Development Environment for Statistical Analysis with. *Journal of Statistical Software* 49(9), 2012.

[5] Friedrichsmeier, T., Michalke, M.: *Introduction to Writing Plugins for RKWard.* 2011. URL: `http://rkward.sourceforge.net/documents/devel/plugins/index.html`

DISKRÉTNÍ, SPOJITÉ A SMÍŠENÉ NÁHODNÉ VELIČINY VE VÝUCE STATISTIKY

DISCRETE, CONTINUOUS AND MIXED RANDOM VARIABLES IN STATISTICS LESSONS

Kamila Hasilová

Adresa: Katedra ekonometrie, Fakulta vojenského leadershipu, Univerzita obrany, Kounicova 65, 662 10 Brno

E-mail: `kamila.hasilova@unob.cz`

Abstrakt: Obvykle jsou ve výuce statistiky náhodné veličiny děleny na diskrétní a spojité a pouze okrajově se zmíní, že existují i jiné typy. Avšak v běžném životě se setkáváme i s jinými typy náhodných veličin, zejména s veličinami smíšeného typu, které lze zčásti popsat jako spojité a zčásti jako diskrétní veličiny. V příspěvku jsou proto uvedeny příklady s těmito typy náhodných veličin – jako například délka čekání na taxi či množství srážek. Některé příklady jsou uvedeny i s řešením, jiné jsou jen nastíněny. Tyto příklady mohou pomoci studentům lépe pochopit vztah mezi přednášenou teorií a postupy, s nimiž se mohou v praxi setkat. Dílčím cílem příspěvku je snaha poskytnout příklady, které budou blízké praktickým problémům. Studenti tak snáze uvidí aplikovatelnost teoretických postupů, čímž se může zvýšit jejich zájem o statistiku. Na druhou stranu vyučující mohou tento krátký souhrn příkladů využít jako jeden z podkladových materiálů ke své výuce.

Klíčová slova: náhodná veličina, smíšený typ, příklad, výuka, statistika.

Abstract: Usually, in statistics lessons, random variables are said to be discrete or continuous. Information about other types of random variables is only marginal. Nevertheless, in our daily life we come across other types of random variables – namely the mixed random variables, which can be described as a mixture of a discrete and continuous random variable. So we provide several examples with these types of random variables, such as the length of waiting for a taxi or the amount of rainfall. Some examples are given with solutions, others are only outlined. These examples can help students to understand better the relation between lectures and procedures they can encounter in practice. A partial aim of this paper is to provide examples that are close to practical problems. Students can easily see the applicability of theoretical methods, which can enhance their interest in statistics. On the other hand, teachers can use this short summary of examples as a part of the background material for their classes.

Keywords: random variable, mixed type, example, teaching, statistics.

1. Náhodné veličiny dělíme na diskrétní a spojité

Takto by se dala charakterizovat převážná většina výňatků ze skript, učebnic a přednášek ze statistiky. Občas je jednou či dvěma větami zmíněno, že existují i náhodné veličiny jiného typu. Na jednu stranu je to dáno výukou (někdy i jejím tempem), kdy se snažíme studentům předat základní statistické znalosti, a texty dostupnými na internetu, kdy se jejich autoři snaží podat látku co nejjednodušeji. Celkově se diskrétním a spojitým náhodným veličinám věnuje tolik pozornosti, až student získá dojem, že jde o jediné typy náhodných veličin, které existují. Na druhou stranu často slýcháváme větu „K čemu mi to v praxi bude?“. U matematiky je tato otázka sice častější, avšak i při výuce statistiky je na ni možné narazit. Měli bychom studentovi co nejvíce přiblížit probíranou teorii a motivovat jej.

K rozvíjení zájmu o předmět mohou sloužit úlohy blížící se reálným situacím. Snažila jsem se o souhrn několika úloh, které je buď možné jednoduše spočítat nebo je studentům alespoň představit jako takzvané intuitivní aplikace. [13] Typově jde o příklady smíšených a useknutých rozdělení náhodných veličin.

2. Příklady s kapkou teorie

Množství srážek je obecně považováno za spojitou náhodnou veličinu. Avšak při bližším pohledu na téměř libovolný záznam srážek si všimneme, že se některá pozorování opakují, což může být způsobeno například zaokrouhlovací chybou. Nicméně nelze přehlédnout, že hodnota nula se vyskytuje příliš často na to, aby šlo o pouhou chybu. Ze zkušenosti víme, že dnů, kdy neprší je mnoho, tedy můžeme psát $P(X = 0) > 0$, což odpovídá podílu dnů, kdy neprší. Tedy mluvíme-li o množství srážek jako o spojité veličině, nepřímo tak předpokládáme, že jde o deštivý den. Obecně je však množství srážek smíšenou náhodnou veličinou. [10]

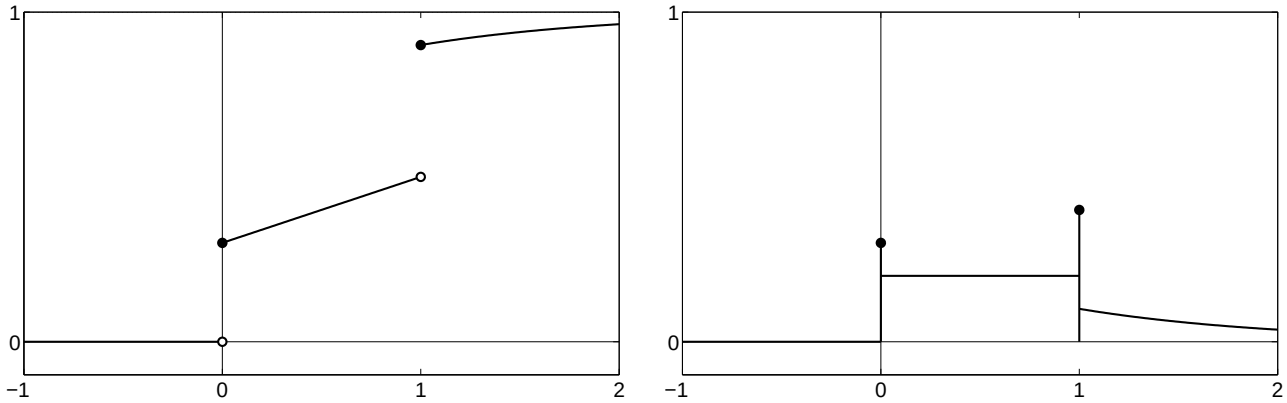
Podobně si můžeme představit situaci, že jsme právě přišli na poštu a chceme popsat, za jak dlouho se dostaneme na řadu. Tento časový úsek označme X. Jaké bude vhodné rozdělení pravděpodobnosti X? [8]

Nabízí se dvě možnosti: buď na poště není žádný zákazník, pak je čas čekání nulový, nebo je před námi několik zákazníků a musíme počkat. Je přirozené předpokládat, že doba čekání je spojitá a tudíž ji lze popsat vhodným spojitým modelem.

Vidíme, že čas čekání X není diskrétní veličinou, protože může nabýt libovolné reálné (kladné) hodnoty, avšak není ani spojitou veličinou, protože je

zde nenulová pravděpodobnost, že nebudeme čekat vůbec, tj. $P(X = 0) > 0$. Jde tedy o smíšené diskrétní-spojité rozdělení.

Je zřejmé, že obor hodnot smíšené náhodné veličiny obsahuje jak izolované body, tak i body ležící v nějakém intervalu. Příslušná distribuční funkce je spojitou funkcí s výjimkou jednoho či více konečných skoků (bodů nespojitosti prvního druhu) – viz obrázek 1(a). Hustotu pravděpodobnosti smíšené



(a) Distribuční funkce (b) Hustota pravděpodobnosti

Obrázek 1: Distribuční funkce a hustoty pravděpodobnosti smíšené náhodné veličiny

náhodné veličiny nelze spočítat přímo, jelikož derivace distribuční funkce neexistuje v bodech nespojitosti. Obvykle se k popisu hustoty pravděpodobnosti využívá Diracova delta funkce. Zejména s ohledem na časové možnosti výuky můžeme studenty vést tak, aby si všimli, že jde o součet diskrétní a spojité části. Pak už lze snadno dopočítat funkci hustoty pravděpodobnosti v obecném smyslu jako derivaci spojité části distribuční funkce a velikosti skoků pro diskrétní část náhodné veličiny – obrázek 1(b).

Na konkrétní praktické problémy je proto možné se dívat z různých úhlů. Například k vyjádření životnosti žárovky lze použít diskrétní rozdělení v případě, že nás zajímá, zda bude fungovat – poprvé se rozsvítí. Pro popis rozdělení pravděpodobnosti životnosti žárovky (doby do poruchy), za předpokladu, že se žárovka poprvé rozsvítila, pak využijeme spojité rozdělení. [6]

Příklad 1. Uvažujme náhodnou veličinu, která popisuje délku telefonního hovoru. Třetina hovorů ani nezačne (volaný nezvedá telefon či příchozí hovor odmítne) a délka započatého hovoru nabývá rovnoměrně hodnot od nula do tří minut. Popište délku hovoru pomocí distribuční funkce a určete průměrnou délku hovoru.

Podobně jako v předchozí úvaze se popis náhodné veličiny dá rozdělit na dvě části. Část diskrétní, kdy volaný nepřijímá hovor, $P(X = 0) = 1/3$, a část spojitou, kterou lze popsat rovnoměrným rozdělením na intervalu $(0, 3]$. Distribuční funkci délky hovoru lze odvodit s pomocí věty o celkové pravděpodobnosti. Označme p pravděpodobnost, že volaný nepřijímá hovor, pak můžeme pro nezáporná x psát

$$\begin{aligned} F(x) = P(X \leq x) &= p \cdot P(X \leq x | \text{volaný nepřijímá hovor}) \\ &\quad + (1-p) \cdot P(X \leq x | \text{volaný přijal hovor}) \\ &= \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot \frac{x}{3} \end{aligned}$$

a distribuční funkci můžeme zapsat takto

$$F(x) = \begin{cases} 0 & x < 0, \\ \dfrac{1}{3} + \dfrac{2x}{9} & 0 \leq x < 3, \\ 1 & x \geq 3. \end{cases}$$

Střední hodnotu dopočítáme buď pomocí zjištěné distribuční funkce

$$\mathsf{E}\, X = \int_0^\infty \big(1 - F(x)\big)\, \mathrm{d}x = \int_0^3 \left(\frac{2}{3} - \frac{2x}{9}\right) \mathrm{d}x = 1,$$

nebo můžeme studentům ukázat i intuitivní přístup pomocí váženého průměru diskrétní (X_d) a spojité (X_s) části

$$\mathsf{E}\, X = p \cdot \mathsf{E}\, X_d + (1-p) \cdot \mathsf{E}\, X_s = \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot \frac{3-0}{2} = 1.$$

Příklad 2. Doba čekání X zákazníka na taxi je nulová, je-li na stanovišti taxíků volné taxi, a rovnoměrně rozdělená v intervalu nula až jedna hodina, je-li stanoviště taxíků prázdné. Pravděpodobnost, že na stanovišti je přítomno taxi, když přijde zákazník, je rovna 0,8. Určete distribuční funkci doby čekání a spočítejte pravděpodobnost, že zákazník bude čekat méně než čtvrt hodiny. [14]

Obdobně jako u příkladu 1 určíme distribuční funkci

$$F(x) = \begin{cases} 0 & x < 0, \\ 0{,}8 + 0{,}2x & 0 \leq x < 1, \\ 1 & x \geq 1. \end{cases}$$

Pak už snadno spočítáme hledanou pravděpodobnost

$$P(X \leq 0{,}25) = F(0{,}25) = 0{,}8 + 0{,}2 \cdot 0{,}25 = 0{,}85.$$

Dalším příkladem může být poloha autobusu na jeho trase v náhodně zvoleném čase. Na polohu autobusu lze nahlížet jako na náhodnou veličinu řídící se rovnoměrným rozdělením od začátku do konce trasy, avšak výjimkou budou zastávky autobusu. [5] Náhodnou veličinu smíšeného typu lze také pozorovat při zaznamenávání počtu vyzářených fotonů vzhledem k jejich vlnové délce u xenonového ultrafialového světla. [9]

Rozdělení smíšeného typu se objevují přirozeně v případech, kdy je spojité rozdělení nějakým způsobem useknuté. Prohlédneme-li si vyúčtování záloh energií, můžeme si všimnout, že v případě přeplatku některé firmy převádí přeplatek menší než 50 Kč do dalšího období. Výše vrácené částky je spojitou náhodnou veličinou, ale její rozdělení je useknuto v bodě, který odpovídá 50 Kč.

Taktéž veličina T popisující životnost výrobku, která nabývá hodnot z intervalu $[0, \infty)$, má spojité rozdělení pravděpodobnosti. Při testování však nelze čekat do nekonečna, tudíž zvolíme kladnou konstantu h a budeme zaznamenávat novou veličinu U, která vznikne useknutím veličiny T v bodě h [11]:

$$U = \begin{cases} T & T < h, \\ h & T \geq h. \end{cases}$$

Se stejným principem se setkáváme i v pojišťovnictví, a to v případě pojistné události s maximální výší pojistného plnění.[1]

Příklad 3. Označme X částku, na kterou je vyčíslena škoda při pojistné události, a předpokládejme, že M značí limit pojistného plnění. Úkolem je popsat náhodnou veličinu Y, značící výši pojistného plnění vyplacenou v případě pojistné události, a spočítat průměrnou vyplacenou částku při pojistné události. [7]

Je zřejmé, že může nastat případ, kdy je škoda vyšší než maximální výše plnění, tedy předpokládáme, že platí $P(X > M) > 0$. Pak náhodnou veličinu Y popíšeme vztahem

$$Y = \begin{cases} X & X < M, \\ M & X \geq M. \end{cases}$$

[1] Pro případ pojištění odpovědnosti za škodu s maximální výší plnění a se spoluúčastí viz např. [3].

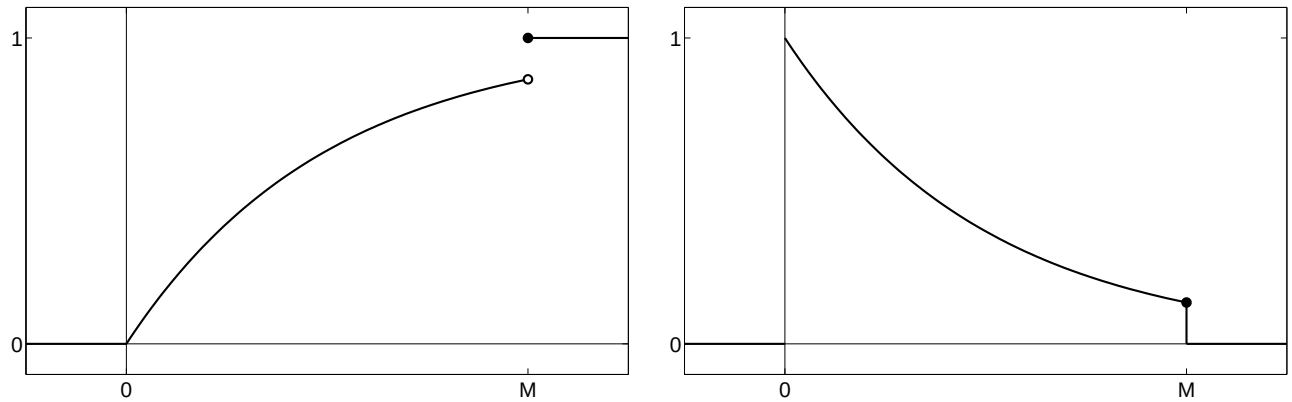
Jelikož výše ztráty X může nabývat libovolné kladné hodnoty, považujeme ji za spojitou náhodnou veličinu a pro jednoduchost uvažujeme, že se řídí exponenciálním rozdělením s parametrem λ. Pak výše pojistného plnění Y představuje smíšenou náhodnou veličinu, jejíž distribuční funkce je

$$F_Y(x) = \begin{cases} 0 & x < 0, \\ 1 - \exp(-\lambda x) & 0 \leq x < M, \\ 1 & x \geq M, \end{cases}$$

a funkci hustoty pravděpodobnosti dostáváme jako

$$f_Y(x) = \begin{cases} \lambda \exp(-\lambda x) & 0 < x < M, \\ \exp(-\lambda M) & x = M, \\ 0 & \text{jinak.} \end{cases}$$

Obě funkce jsou znázorněny na Obrázku 2.



(a) Distribuční funkce (b) Hustota pravděpodobnosti

Obrázek 2: Distribuční funkce a hustota pravděpodobnosti výše pojistného plnění

Vypočítáme-li střední hodnotu výše plnění

$$\mathsf{E}\,Y = \int_0^M \big(x\lambda \exp(-\lambda x)\big)\,\mathrm{d}x + M \cdot \exp(-\lambda M) = \frac{1}{\lambda}\big(1 - \exp(-\lambda M)\big),$$

vidíme, že je menší než střední hodnota výše škody $\mathsf{E}\,X = 1/\lambda$. Což není překvapující výsledek – maximální limit výše pojistného plnění snižuje riziko pojišťovny. Očekávaná výše plnění ze strany pojišťovny je zmenšena o částku $\exp(-\lambda M)/\lambda$ na jednu pojistnou událost.

Smíšené náhodné veličiny se vyskytují v technické praxi, obvykle jako transformované veličiny $Y = g(X)$, kde X je spojitá náhodná veličiny a funkce g, obvykle lineární, charakterizuje citlivost přístroje, kterým hodnoty veličiny zpracováváme. Příkladem může být zesilovač, který má lineární oblast $-\nu \leq x \leq \nu$, kde $g(x) = \alpha x$, avšak pro hodnoty mimo tento interval platí $g(x) = \alpha\nu$ pro $x > \nu$ a $g(x) = -\alpha\nu$ pro $x < \nu$. [1]

Dalším příkladem je měření viremie (množství viru v krvi), která se při onemocnění hepatitidou C obvykle pohybuje v milionech IU v mililitru krve[2] a jejíž spodní hranice měřitelnosti je kolem 50 IU/ml. To vede k problémům při rozlišení nízkého množství viru a jeho úplné nepřítomnosti. [2]

Také v hydrobiologii jsme omezeni spodní hranicí měřitelnosti přístroje. Například při měření obsahu rozpuštěného kyslíku ve vodě, což je pak potřeba brát v úvahu při případných výpočtech pravděpodobnosti kyslíkového deficitu a s ním souvisejícím úhynem ryb. [4]

3. Závěr

Smíšené náhodné veličiny jsou přirozeným, avšak ne jediným příkladem náhodných veličin, které nejsou ani diskrétní ani spojité. Příkladem mohou být náhodné vektory se smíšenými souřadnicemi. Případně i náhodné veličiny, jejichž oborem hodnot nejsou ani reálná ani celá čísla, ale jejich zobecnění na tzv. časové škály (time scales, viz např. [12]). Avšak jejich představení by už přesáhlo jak rozsah článku, tak i časové možnosti, které máme jako vyučující k dispozici pro výuku statistiky.

Cílem uvedených příkladů je pomoci studentům nahlédnout, že svět náhodných veličin není jen a pouze diskrétní a spojitý, ale že v praxi se mohou setkat i s veličinami jiného typu.

Poděkování

Článek byl podpořen Univerzitou obrany projektem Pokročilý automatizovaný systém velení a řízení (PRO PASVŘ2012).

Literatura

[1] Gubner, J. A.: *Probability and Random Processes for Electrical and Computer Engineers.* New York, Cambridge University Press, 2006.

[2] IU (International Unit) mezinárodní jednotka popisující množství ribonukleové kyseliny viru.

[2] Hepatitida C, 2013, URL: http://www.hepc.cz/index.php?page=2&faq=viremie.htm

[3] Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M.: *Modern Actuarial Risk Theory Using R.* Berlin, Springer, 2008.

[4] Kopp, R.: *Hydrochemie.* Mendelova univerzita v Brně, 2014. URL: http://web2.mendelu.cz/af_291_projekty2/vseo/ (nutné přihlášení)

[5] Larson, R. C., Odoni, A. R.: *Urban Operations Research.* New Jersey, Prentice Hall, 1981. URL: http://web.mit.edu/urban_or_book/www/book/chapter2/2.10.html

[6] Leitch, M.: *A Pocket Guide to Risk Mathematics: Key Concepts Every Auditor Should Know.* Chichester, Wiley, 2010.

[7] Ma, D.: An insurance example of a mixed distribution – I. *Applied Probability and Statistics in Actuarial Science and Financial Economics.* 2010, [blog]. URL: https://mathmodelsblog.wordpress.com/2010/01/05/an-insurance-example-of-a-mixed-distribution-i/

[8] Meester, R.: *A Natural Introduction to Probability Theory.* Birkhäuser, 2013.

[9] MIT OpenCourseWare: Probability Notes, Chapter 1, One Random Variable, 8.044 Statistical Physics I, Spring 2013. URL: http://ocw.mit.edu/courses/physics/8-044-statistical-physics-i-spring-2013/readings-notes-slides/

[10] Saha, A.: Lecture notes 26, STA 4321 – Introduction to Probability. University of Florida, 2010. URL: http://www.stat.ufl.edu/∼abhisheksaha/sta4321/

[11] Siegrist, K.: Random: Probability, Mathematical Statistics, Stochastic Processes, 1997–2015. URL: http://www.math.uah.edu/stat/

[12] Ufuktepe, Ü: Unification of probabilitu theory on time scales. *Advances in Difference Equations,* 211, pp. 1–13, 2012.

[13] Vágner, M., Hasilová, K.: Matematika a aplikace. *XXXIII International Colloquium on the Management of Educational Process: proceedings of abstracts and electronic versions of reviewed contributions on CD-ROM,* Brno, Univerzita obrany, 2015. ISBN 978-80-7231-995-4.

[14] Vosoughi, A.: Handout 5, ECE 270: Introduction to Probability. University of Rochester, 2008. URL: http://www.ece.rochester.edu/courses/ ECE270/

ZKUŠENOSTI S VÝUKOU PŘEDMĚTU ZPRACOVÁNÍ REÁLNÝCH DAT

EXPERIENCE WITH EDUCATION OF COURSE OF THE PROCESSING OF REAL DATA

Marie Budíková[1], Ivanka Horová[1], Kateřina Konečná[1], Petra Širůčková[1], Eva Bártová[2]

Adresa: [1]Ústav matematiky a statistiky Přírodovědecké fakulty Masarykovy univerzity, Kotlářská 2, 611 37, Brno
[2]Ústav biologie a chorob volně žijících zvířat Fakulty veterinární hygieny a ekologie Veterinární a farmaceutické univerzity Brno, Palackého tř. 1946/1, 612 42, Brno

E-mail: budikova@math.muni.cz, horova@math.muni.cz, xkonecn3@math.muni.cz, 324037@mail.muni.cz, bartovae@vfu.cz

Abstrakt: V rámci projektu podpořeného Fondem rozvoje Masarykovy univerzity byl na Oddělení aplikované matematiky Přírodovědecké fakulty Masarykovy univerzity ve spolupráci s Fakultní nemocnicí Brno a Veterinární a farmaceutickou univerzitou Brno vytvořen předmět „Zpracování reálných dat“. Příspěvek je zaměřen především na organizační stránku nově vzniklého předmětu a jeho obsahovou a edukační strukturu s ukázkou konkrétního zadání doprovodných úkolů určených k pochopení probíraných statistických metod. Poskytuje informace o závěrečném workshopu, na kterém studenti předvedli s využitím softwaru R a STATISTICA schopnost implementace nabytých teoretických poznatků v praxi. Součástí příspěvku je také ukázka jedné ze závěrečných studentských prezentací. V závěru jsou k tomuto předmětu shrnuty názory studentů, které byly získány v celoškolní předmětové anketě, rovněž jsou zde představeny plány do budoucna.

Klíčová slova: statistické programy STATISTICA a R, projekt Fond rozvoje Masarykovy univerzity, spolupráce s praxí.

Abstract: A course „The processing of real data“ has been created by the Department of Mathematics and Statistics, Faculty of Science, Masaryk University Brno in partnership with the University Hospital Brno and the University of Veterinary and Pharmaceutical Sciences Brno within the project supported by Masaryk University Development Fund. This contribution is focused especially on the organizational aspect of the new course, its content and educational structure with the specific task of illustrative excercises for the understanding of the discussed statistical methods. The contribution

provides information regarding the final workshop, where students demonstrated their ability to implement acquired theoretical knowledge in practice using R and STATISTICA software. The contribution also includes an example of final student presentation. In conclusion, students' opinions of the course which have been collected in the Course Opinion Poll are summarized as well as plans for the future are introduced.

Keywords: statistical software STATISTICA and R, Masaryk University Development Fund project, cooperation with practice.

1. Úvod

V současné době statistické metody podstatnou měrou přispívají k rozvoji přírodních a sociálních věd, medicíny i techniky. Oddělení aplikované matematiky Přírodovědecké fakulty Masarykovy univerzity (OAM PřF MU) má dosti bohaté zkušenosti se zpracováním statistických dat z různých oblastí praxe. V domácích i zahraničních časopisech jsme publikovali články a také jsme připravovali příspěvky na konference společně s klimatology, hydrology, ekonomy i psychology, nejčastěji však s lékaři. V rámci potřeby řešení výše zmíněných reálných problémů pomocí teoretického aparátu bylo naše oddělení v letech 2011–2014 zapojeno do rozsáhlého projektu A-Math-Net. Projekt mimo jiné zahrnoval spolupráci s Masarykovým onkologickým ústavem (MOÚ), přičemž spolupráce vedle odborné oblasti probíhala i v oblasti výuky.

V rámci tohoto projektu byl nově vytvořen předmět „Zpracování klinických dat“, v němž lékaři z MOÚ přednášeli našim studentům o problémech medicínských dat. Předmět vyvolal mezi studenty velký zájem. Z následné ankety vyplynulo, že ho považovali za velmi užitečný a přínosný. Studenti zvláště ocenili koncepci předmětu, v níž se stýkal statistický a medicínský pohled na problematiku. Tato pozitivní zkušenost nás motivovala ke spolupráci s odborníky i z jiných institucí. Tito odborníci by se podíleli na výuce, seznamovali by studenty s problémy, které potřebují s využitím statistických metod vyřešit, a v součinnosti by s vyučujícími z OAM připravovali dílčí úkoly pro studenty.

Abychom mohli tento záměr realizovat a inovovat tak předmět „Zpracování klinických dat“, požádali jsme spolu s Fakultní nemocnicí Brno (FN Brno) a Veterinární a farmaceutickou univerzitou Brno (VFU Brno) o grant z Fondu rozvoje Masarykovy univerzity. Navrhovatelem grantu byla RNDr. Marie Budíková, Dr. (OAM MU Brno), spolunavrhovateli RNDr. Marie Forbelská, Ph.D. (OAM MU Brno), prof. RNDr. Ivanka Horová, CSc. (OAM MU Brno), MUDr. Jiří Žurek, Ph.D. (primář Kliniky dětské anesteziologie a resuscitace FN Brno) a doc. MVDr. Eva Bártová, Ph.D. (Ústav biologie a chorob volně ži-

jících zvířat Fakulty veterinární hygieny a ekologie VFU Brno). Do projektu byly rovněž zapojeny doktorské studentky OAM MU Brno Mgr. Kateřina Konečná a Mgr. Petra Širůčková.

Grant byl získán a v jarním semestru školního roku 2014/15 již byl předmět „Zpracování reálných dat“ (MAZRD) vyučován.

2. Předmět Zpracování reálných dat

Cílem předmětu MAZRD bylo navázat již na dřívější teoretické i praktické znalosti studentů, poskytnout jim přehled dalších moderních statistických metod a naučit je správně interpretovat výsledky statistických analýz. Předmět zahrnoval přednášky, na které bezprostředně navazovala prakticky zaměřená cvičení, na nichž byla s využitím software STATISTICA [1], resp. R [2] zpracovávána konkrétní reálná data. Měl dvouhodinovou týdenní dotaci a byl zakončen kolokviem, které mělo formu závěrečného workshopu, na němž studenti prezentovali výsledky své práce.

Předmět si zapsalo 22 magisterských a bakalářských studentů ze studijních oborů Statistika a analýza dat, Aplikovaná matematika pro víceoborové studium a Matematická biologie. Na počátku semestru studenti navštívili Kliniku dětské anesteziologie a resuscitace FN Brno, kde vyslechli přednášku MUDr. Žurka o využití statistických metod při zpracování údajů o dětských pacientech uvedené kliniky a také si prohlédli kliniku. V polovině semestru pak studenti zavítali na Fakultu veterinární hygieny a ekologie VFU Brno, kde pro ně doc. Bártová připravila přednášku o parazitárních onemocněních zvířat a provedla je areálem VFU.

Osnova předmětu MAZRD zahrnovala následující témata:

- Typy statistických proměnných
- Příprava datového souboru
- Popis datového souboru
- Grafické znázornění dat
- Ověřování předpokladů modelu
- Parametrické a neparametrické testy
- Analýza kontingenčních tabulek
- Korelační analýza
- Regresní modely
- Zobecněné lineární modely
- Základy analýzy přežití

Ke každému tématu měli studenti připravené úkoly s návodem ke zpracování v systémech STATISTICA a R. Přitom mohli využívat konzultací s vyučujícími předmětu a doktorskými studentkami, které působily nejen jako konzultantky, ale podílely se i na tvorbě podpůrných materiálů.

3. Konkrétní úkoly

Následuje konkrétní ukázka úkolů zaměřených na analýzu kontingenčních tabulek. Datový soubor (viz tabulka 1) obsahuje údaje o 260 králících, kteří byli vyšetřeni pomocí sérologické metody nepřímé imunofluorescence (IFAT) na přítomnost protilátek proti parazitu *Toxoplasma gondii*. Jednalo se o samice chované na 13 farmách ve čtyřech provinciích v severní Itálii. Králíci nevykazovali žádné klinické příznaky onemocnění. Séra byla považována za pozitivní, pokud byl v IFAT detekován titr ≥ 50. Králíci byli rozděleni do skupin podle věku, plemene, provincie, velikosti farmy a počtu porodů.

Na hladině významnosti 0,05 budeme ověřovat, zda výskyt protilátek proti parazitu *Toxoplasma gondii* závisí na uvedených pěti faktorech. (Pokud některá kategorie není známa, do zpracování ji nezahrnujeme.)

Úkol 1. Kolik procent králíků bylo séropozitivních? Najděte 95% asymptotický interval spolehlivosti pro pravděpodobnost výskytu protilátek proti parazitu *Toxoplasma gondii*.

Úkol 2. Testujte hypotézu, že výskyt protilátek proti parazitu *Toxoplasma gondii* nezávisí na věkové kategorii králíků. Vytvořte kontingenční tabulku, kde roli řádkové proměnné bude hrát alternativní proměnná Y ($Y = 1$ pro pozitivního jedince, $Y = 2$ pro negativního jedince) a roli sloupcové proměnné pak věk. Spočtěte rovněž sloupcově podmíněné relativní četnosti. Uveďte hodnotu testové statistiky, příslušnou p-hodnotu a rozhodnutí o nulové hypotéze. Nezapomeňte ověřit splnění podmínek dobré aproximace. Vypočtěte Cramérův koeficient. Pro věkovou kategorii 7–12 měsíců sestrojte 95% interval spolehlivosti pro pravděpodobnost výskytu protilátek proti parazitu *Toxoplasma gondii*.

Úkol 3. Testujte hypotézu, že výskyt protilátek proti parazitu *Toxoplasma gondii* nezávisí na počtu porodů králíků. Postupujte analogicky jako v Úkolu 2.

Úkol 4. Testujte hypotézu, že výskyt protilátek proti parazitu *Toxoplasma gondii* nezávisí na plemenu králíků. Postupujte analogicky jako v Úkolu 2.

Úkol 5. V Úkolu 4 byla hypotéza o nezávislosti výskytu protilátek proti parazitu *Toxoplasma gondii* a plemena králíků zamítnuta na hladině významnosti 0,05. Z kontingenční tabulky sloupcově podmíněných relativních čet-

Tabulka 1: Charakteristiky králíků vyšetřených na protilátky proti parazitu *Toxoplasma gondii* (celkové počty a počty pozitivních zvířat)

Charakteristika	Celkový počet zvířat	Počet pozitivních
Věková kategorie (měsíce)		
2–4	59	4
$\geq$ 5–6	66	8
$\geq$ 7–12	74	10
$\geq$ 13–30	26	3
Neznáma	35	13
Počet porodů		
0	60	4
1–2	71	6
3–7	57	9
8–15	37	6
Mnoho, ale neznámo kolik	35	13
Plemeno		
Hyla	160	23
Hycole	40	12
Grimaud	40	3
Hybrid	20	0
Provincie		
Padova	20	0
Rovigo	20	3
Treviso	160	11
Verona	60	24
Velikost farmy		
Malé farmy ($<$ 1000 samic)	120	19
Velké farmy ($\geq$ 1000 samic)	140	19

ností je vidět, že nejvyšší séropozitivitu mají králíci plemena *Hycole*, a to 30 %. Pro toto plemeno tedy spočtěte výběrový podíl šancí na výskyt protilátek proti parazitu *Toxoplasma gondii* vůči všem ostatním plemenům dohromady. Výpočet doplňte 95% intervalem spolehlivosti pro podíl šancí.

Úkol 6. Testujte hypotézu, že výskyt protilátek proti parazitu *Toxoplasma gondii* nezávisí na provincii, v které králíci žijí. Postupujte analogicky jako v Úkolu 2.

Úkol 7. Testujte hypotézu, že výskyt protilátek proti parazitu *Toxoplasma gondii* nezávisí na velikosti farmy, na které králíci žijí. Postupujte analogicky jako v Úkolu 2.

4. Závěrečný workshop studentů

V posledních třech týdnech semestru měli studenti možnost samostatně statisticky zpracovat reálná data. Mohli tak zhodnotit své teoretické vědomosti, které nabyli během semestru. K dispozici bylo pět datových souborů, které se postupně týkaly těchto oblastí:

- Výskyt protilátek proti parazitům *Toxoplasma gondii* a *Neospora caninum* u psů.
- Výskyt protilátek proti parazitům *Toxoplasma gondii* a *Neospora caninum* u divokých prasat.
- Hladina proteinů v krvi při úrazech mozku dětských pacientů.
- Analýza přežití a zkoumání akutní a chronické toxicity radioterapie u pacientů s nádory mozku.
- Analýza přežití a zkoumání akutní a chronické toxicity radioterapie u pacientů s nádory rekta.

Ke každému datovému souboru byl sestaven komentář, který obsahoval vysvětlení významu jednotlivých proměnných, a seznam otázek, kterými se studenti měli zabývat. Studenti se rozdělili do skupin čítajících tři až pět členů. V těchto skupinách zpracovávali zadané úkoly a prokázali tak schopnost týmové spolupráce. Pro zpracování datových souborů bylo třeba využít analýzu kontingenčních tabulek, korelační analýzu, testy o středních hodnotách, logistickou regresi, analýzu přežití apod.

Na datové soubory bylo důležité nahlížet globálně – bylo potřeba se zaměřit na primární úpravu a průzkumovou analýzu dat, zvolit vhodný statistický aparát a prokázat analytické myšlení při interpretaci výstupů ze softwaru. Získané závěry studenti prezentovali na závěrečném workshopu, který se konal v posledním týdnu semestru. Studenti okamžitě obdrželi na podaný

výkon zpětnou vazbu od vyučujících předmětu, byli hodnoceni z hlediska správného použití statistických postupů, správnosti vyvozených závěrů a formální úrovně své prezentace.

Ukázkou jednoho z prezentovaných příspěvků je zpracování datového souboru, který obsahuje proměnné týkající se dětských pacientů s úrazem hlavy. Datový soubor obsahoval mnoho proměnných, při zpracování dat hrály nejdůležitější roli proměnné *věk*, *pohlaví*, *exitus*, hladina proteinu *S100* (protein mozkového poškození) a *NSE* (hladina proteinu mozkového poškození při příjmu pacienta) či typ zranění. Z důvodu omezeného rozsahu příspěvku uvedeme pouze stěžejní výsledky studentů. Celá prezentace studentů je k dispozici v Informačním systému MU [3].

Úkolem studentů bylo vybrat proměnné a pomocí nich vytvořit vhodný model, který bude s co nejvyšší přesností rozlišovat pacienty, u nichž dojde k úmrtí. Datový soubor obsahoval údaje o 63 pacientech, úplná data byla dostupná pouze pro 58 pacientů, z nichž 8 zemřelo.

Nejprve studenti provedli průzkumovou analýzu souboru. Byla testována závislost mezi proměnnými *pohlaví* a *exitus*, jde o binární proměnné s kategoriemi *muž* a *žena* u proměnné *pohlaví* a kategoriemi A (úmrtí nastalo) a N (úmrtí nenastalo) pro proměnnou *exitus*. Absolutní četnosti zkoumaných proměnných jsou uvedeny v tabulce 2. K analýze dat byl použit Fisherův

Tabulka 2: Závislost proměnných *pohlaví* a *exitus*

pohlaví	*exitus* = A	*exitus* = N
muž	5	33
žena	4	21

přesný test, jehož p-hodnota byla 1. Informaci o síle závislosti nám poskytuje Cramerův koeficient, jehož hodnota 0,0397 svědčí o zanedbatelné závislosti.

Dále byla ověřena nezávislost *věku* na *exitu*. Nejmladšímu pacientovi byly 2 měsíce, nejstaršímu 220 měsíců (18,$\bar{3}$ let), průměrný věk byl 106,6 měsíců. Kvůli porušení normality dat byl zvolen Mann-Whitneyho U test. Mediány věku v závislosti na proměnné *exitus* a p-hodnota testu jsou uvedeny v tabulce 3. Vysoká p-hodnota svědčí o tom, že statisticky významný rozdíl mezi skupinami nebyl prokázán.

Následně byla testována nezávislost proměnné *GCS* (Glasgow coma score, skórovací systém hodnotící stav pacienta při příjmu) na *exitu*. Data nevykazovala normální rozložení, proto byl pro testování rovněž zvolen Mann-Whitneyho U test. Mediány *GCS* v závislosti na proměnné *exitus* a p-hodnota

testu jsou uvedeny v tabulce 3. Nízká p-hodnota prokázala závislosti proměnných s rizikem omylu maximálně 0,05.

Tabulka 3: Test nezávislosti proměnných *věk* a *GCS* na *exitu*

	exitus = *A*	*exitus* = *N*	p-hodnota
věk	105	104	0,4494
GCS	4	8,5	0,0002249

Studenti poté porovnávali logaritmus hladin proteinu *NSE* v závislosti na skupinách proměnné *exitus*. Protože data vykazovala normální rozložení, k ověření shody středních hodnot obou vzorků byl použit Studentův t-test. Výběrové průměry obou skupin a p-hodnota testu jsou uvedeny v tabulce 4. Díky malé p-hodnotě zamítáme nulovou hypotézu o shodě středních hodnot, proto by mohla proměnná *NSE* hrát významnou roli v modelu, který bude rozlišovat pacienty do skupin N a A.

Tabulka 4: Srovnání logaritmu proměnné *NSE* v rámci skupin N a A pomocí dvouvýběrového t-testu

	exitus = *N*	*exitus* = *A*	p-hodnota
$\log(NSE)$	3,50	4,36	0,0003

Pro úplnost uvedeme hodnoty korelačních koeficientů mezi spojitými vysvětlujícími proměnnými *GCS*, *věk* a $\log(NSE)$ (logaritmická transformace proměnné *NSE*) a p-hodnoty odpovídajících testů (viz tabulka 5).

Tabulka 5: Hodnoty korelačních koeficientů mezi vysvětlujícími proměnnými

	Korelační koeficient	p-hodnota
GCS – $\log(NSE)$	−0,2165	0,1026
věk – $\log(NSE)$	−0,2861	0,0295
GCS – *věk*	0,1178	0,3785

Studenti rovněž testovali nezávislost spojitých proměnných na *pohlaví* pomocí Mann-Whitneyho U testu (všechny p-hodnoty byly větší než zvolená hladina významnosti 0,05).

Cílem studentů bylo vytvořit logistický regresní model pro proměnnou *exitus*, proměnné vstupující do modelu byly vybrány na základě průzkumové analýzy. Studenti vytvořili několik různých modelů, které vzájemně porovnali. V tomto příspěvku uvedeme pouze jeden z vytvořených modelů, roli vstupních proměnných v něm hrají *GCS*, $\log(NSE)$, *věk* a *pohlaví*. Označme regresní koeficienty β_0, β_1, β_2, β_3, β_4, β_0', β_1', β_2' a chybové vektory ε, ε'.

$$M_0\colon \mathit{exitus} \sim \beta_0 + \beta_1 GCS + \beta_2 \log(NSE) + \beta_3 \mathit{věk} + \beta_4 \mathit{pohlaví} + \varepsilon.$$

Pomocí stepwise procedury byl původní model zredukován na konečný model obsahující proměnné *GCS* a $\log(NSE)$:

$$M_1\colon \mathit{exitus} \sim \beta_0' + \beta_1' GCS + \beta_2' \log(NSE) + \varepsilon'.$$

Pro přesnější představu o výsledném modelu uvádíme výstup z jazyka R.

```
Call:
glm(formula = exitus ~ GCS + log(NSE.1), family = binomial,
    data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
     -1.54126  -0.20666  -0.10011  -0.03158   1.97760

Coefficients:
            Estimate Std. Error z~value Pr(>|z|)
(Intercept)  -3.0757     3.1140  -0.988   0.3233
GCS          -0.8233     0.3335  -2.469   0.0136 *
log(NSE.1)    1.5027     0.7839   1.917   0.0553 .
---
Signif. codes:  0 '***' 0.001 '**' 0. 01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.538  on 57  degrees of freedom
Residual deviance: 22.454  on 55  degrees of freedom
AIC: 28.454

Number of Fisher Scoring iterations: 7
```

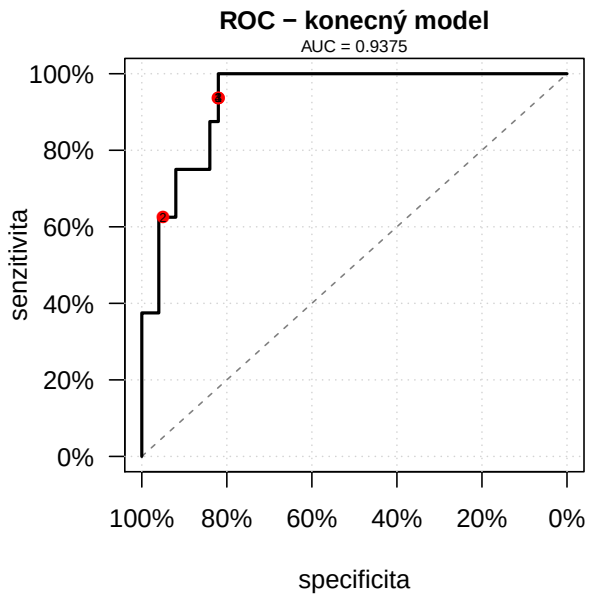
Pro posouzení kvality výsledného modelu studenti znázornili ROC křivku (viz obrázek 1). Pro daný model byla zvolena hodnota prahového bodu 0,1131,

tato hodnota představuje bod, ve kterém je součet senzitivity a specificity modelu maximální. Pro takto zvolený prahový bod senzitivita nabývá hodnoty 100 % a specificita 82 %. Na závěr byla modelována pravděpodobnost úmrtí pacientů s využitím výsledného modelu:

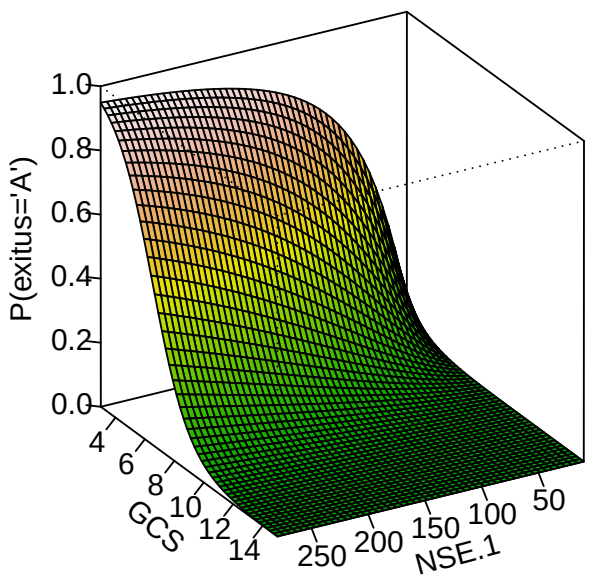
$$P\left(exitus = A\right) = \frac{\mathrm{e}^x}{1 + \mathrm{e}^x},$$

kde $x = -3{,}0757 - 0{,}8233\,GCS + 1{,}5027\log\left(NSE\right)$, graficky je znázorněna na obrázku 2.

Kromě popsaných analýz dále studenti detailně zpracovali průzkumovou analýzu všech proměnných vystupujících v souboru, zkoumali závislosti hladin proteinů na zranění, závislosti jednotlivých proměnných na proměnné *exitus* stejně jako např. dynamiku hladiny proteinů.



Obrázek 1: ROC křivka



Obrázek 2: Modelovaná pravděpodobnost úmrtí na základě *GCS* a *NSE*

5. Závěr

Na závěr každého semestru mají studenti možnost vyjádřit svůj názor na kvalitu zapsaných předmětů. Studentská anketa je jednou z možností zpětné vazby, která hodnotí nejen kvalitu výuky pedagoga, ale i obsahovou stránku předmětu jako takového. Níže přikládáme reakce a hodnocení studentů předmětu MAZRD:

„Předmět byl pro mě velmi přínosný. Je skvělé, že ÚMS se otevírá veřejnosti a snaží se navázat spolupráce s ostatními institucemi. Podle mě by takovýchto předmětů mělo být více :).“

„Látka probíraná na předmětu nebyla (většinou) nová, byla pro mě tedy spíše opakováním. Za největší (a velmi podstatný) přínos považuji závěrečný projekt, kdy jsme ve skupinkách zpracovávali reálná data z dětské nemocnice a VFU. Vyzkoušela jsem si týmovou práci, což byla krásná zkušenost, a samostatné zpracování dat. Moc mě těší, že jsme v rámci předmětu vypracovávali analýzu, jejíž výsledky mohou být využitelné v praxi. (Tuto možnost preferuji mnohem více než zpracování simulovaných dat).“

„Nejvíce si cením možnosti analyzovat reálná data ještě na vysoké škole. Obor Statistika by měl být (dle mého názoru) více orientován na praktické zpracování dat a absolventi by měli být vedeni k samostatné analýze, aby na ni po absolvování byli připraveni a měli s ní nějaké zkušenosti. Za tento předmět tedy moc děkuji a doufám, že bude otevřen ostatním studentům i v dalších letech :).“

Předmět „Zpracování reálných dat“ bude zařazen do studijního katalogu a nabízen studentům jako doporučený předmět i v dalších letech. Chtěli bychom pokračovat ve spolupráci s FN Brno i VFU Brno a rádi bychom získali odborníky i z dalších oborů.

Poděkování

Příspěvek byl podpořen projektem MUNI/FR/1584/2014 „Inovace a rozšíření náplně kurzu Zpracování klinických dat“. MUDr. Jiřímu Žurkovi, Ph.D., děkujeme za poskytnutí dat k prezentaci `kraniotraumata.pdf`.

Literatura

[1] StatSoft, Inc. (2013). *STATISTICA data analysis software system*, version 12. URL: `http://www.statsoft.com/`

[2] *The R Project for Statistical Computing*. URL: `http://www.r-project.org/`

[3] Studentská prezentace na téma „Hladina proteinů v krvi při úrazech mozku dětských pacientů“, 2015. URL: `http://is.muni.cz/el/1431/jaro2015/MAZRD/um/Kraniotraumata.pdf?lang=cs`

STAKAN 2015

Žaneta Miklová

Loňská vydařená konference STAtističní KANtoři se konala 9. – 11. října 2015 v hotelu Čarták na Soláni v malebných Vsetínských vrších. Akce byla organizována Českou statistickou společností a Slovenskou štatistickou a demografickou spoločnosti ve spolupráci s Katedrou aplikované matematiky Vysoké školy Báňské – Technické univerzity v Ostravě. O bezproblémový průběh akce se postarala hlavní organizátorka Martina Litschmannová se svým týmem.

Na konferenci zazněly příspěvky převážně českých a slovenských autorů. Tyto příspěvky byly v souladu s hlavními tématy letošní konference, a to:

- Výuka statistiky na středních a vysokých školách
- Aplikace statistických metod
- Nové směry ve vývoji statistických metod

První část konference se zaměřila právě na výuku statistiky, kdy autoři předávali ostatním účastníkům své zkušenosti s výukou statistiky nejen v České a Slovenské republice, ale také v zahraničí. Druhá část pak byla věnována jak novým, tak i současným statistickým metodám, jejich využití a hlavně zveřejnění dosažených výsledků. Mohli jsme vidět například aplikaci pokročilých metod na údajích o průtocích řek nebo na lékařských datech.

Atmosféru konference podpořila dobrá nálada všech zúčastněných. Za tuto skvělou akci musíme hlavně poděkovat organizačnímu a programovému výboru. Již teď se těšíme na další ročník, který bude určitě stejně kvalitní jako ten letošní.

Žaneta Miklová, Ostrava, říjen 2015

ČLENSKÁ SCHŮZE ČESKÉ STATISTICKÉ SPOLEČNOSTI V ROCE 2016

Hana Řezanková

Dne 18. ledna 2016 se konala členská schůze České statistické společnosti. Místem konání byla již tradičně zasedací místnost na Českém statistickém úřadu v Praze. Úvodem předsedkyně Společnosti prof. Řezanková přivítala místopředsedu ČSÚ Ing. Marka Rojíčka, Ph.D., který přivítal účastníky a připomněl dosavadní spolupráci ČSÚ a ČStS. Poté prof. Řezanková přednesla zprávu o činnosti České statistické společnosti v roce 2015 a Ing. Löster přednesl zprávu o hospodaření za rok 2015 a plán rozpočtu na rok 2016. Se stanoviskem revizorky, doc. Blatné, seznámila přítomné prof. Řezanková. Všechny přednesené zprávy byly členskou schůzí schváleny.

Poté byla zahájena zvaná přednáška RNDr. Patrície Martinkové, Ph.D., z Ústavu informatiky AV ČR na téma *Statistika v interdisciplinárním výzkumu: zkušenosti z University of Washington.*

Základní údaje o České statistické společnosti v roce 2015

V roce 2015 byl na členské schůzi České statistické společnosti konané dne 29. ledna zvolen nový výbor ČStS na dvouleté funkční období. Předsedkyní Společnosti byla do druhého funkčního období zvolena prof. Řezanková z VŠE v Praze. Na první schůzi výboru ČStS, která se konala ve stejný den, byli zvoleni další členové předsednictva, a to prof. Dohnal (FS ČVUT v Praze) – místopředseda, prof. Picek (TU Liberec) – vědecký tajemník a Ing. Löster (VŠE Praha) – hospodář. Dále členové výboru ČStS zastávali v roce 2015 následující funkce: prof. Dohnal – šéfredaktor časopisu *Informační bulletin České statistické společnosti* (IB ČStS), Mgr. Vencálek – odpovědný redaktor IB ČStS, Ing. Litschmannová – správce webových stránek ČStS. Činnost technického redaktora IB ČStS stejně jako v předchozích letech vykonával Ing. Pavel Stříž.

V prosinci 2015 měla Česká statistická společnost 205 členů. Od roku 2015 byl členský příspěvek stanoven na 300,– Kč, redukovaný členský příspěvek pro studenty a důchodce zůstal ve výši 100,– Kč. Plné znění výroční zprávy o činnosti ČStS v roce 2015 je k dispozici na webové stránce `http://www.statspol.cz/cs/wp-content/uploads/Vyrocni-zprava-2015.pdf`.

Strukturovaná roční zpráva odevzdaná Radě vědeckých společností je publikována na `http://rvs.paleontologie.cz/rocni/id-652/`.

Obsah

Vědecké a odborné články

Michal Běloch, Pavlína Kuráňová, Robert Skopal, Martina Litschmannová, Kateřina Janurová
Použití interaktivních videí a testů pro výuku jazyka R ve statistice 1

Robert Skopal, Kateřina Janurová, Michal Běloch, Pavlína Kuráňová, Martina Litschmannová
RKWard: Komplexní prostředí pro výuku statistiky v jazyku R 7

Kamila Hasilová
Diskrétní, spojité a smíšené náhodné veličiny ve výuce statistiky 16

Marie Budíková, Ivanka Horová, Kateřina Konečná, Petra Širůčková, Eva Bártová
Zkušenosti s výukou předmětu Zpracování reálných dat 24

Zprávy a informace

Žaneta Miklová
Stakan 2015 35

Hana Řezanková
Členská schůze České statistické společnosti v roce 2016 36

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server `http://www.vyzkum.cz/`.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

Předsedkyně společnosti: prof. Ing. Hana Řezanková, CSc., KSTP FIS VŠE v Praze, nám. W. Churchilla 4, 130 67 Praha 3, e-mail: `hana.rezankova@vse.cz`.

Redakce: prof. RNDr. Gejza Dohnal, CSc. (šéfredaktor), prof. RNDr. Jaromír Antoch, CSc., prof. Ing. Václav Čermák, DrSc., doc. Ing. Jozef Chajdiak, CSc., doc. RNDr. Zdeněk Karpíšek, CSc., RNDr. Marek Malý, CSc., doc. RNDr. Jiří Michálek, CSc., prof. Ing. Jiří Militký, CSc., doc. Ing. Iveta Stankovičová, PhD., doc. Ing. Josef Tvrdík, CSc., Mgr. Ondřej Vencálek, Ph.D.

Redaktor časopisu: Mgr. Ondřej Vencálek, Ph.D., `ondrej.vencalek@upol.cz`.
Informace pro autory jsou na stránkách společnosti, `http://www.statspol.cz/`.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.