# SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

## SLOVAK STATISTICS and DEMOGRAPHY

1/2016
ročník/volume 26


Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.


Článok/Article: 5

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 82 – 91

Dátum vydania/Publication date: 15. január 2016/January 15, 2016


**Michal PÁLEŠ**
**Katedra matematiky a aktuárstva Fakulty hospodárskej informatiky, Ekonomická univerzita v Bratislave**

# GRAFICKÁ PODPORA JAZYKA R PRI ŠTATISTICKÝCH ANALÝZACH

# GRAPHICAL SUPPORT FOR THE R LANGUAGE IN STATISTICAL ANALYSIS

## ABSTRAKT
Fenoménom v oblasti softvéru na aktuárske analýzy sa v súčasnosti stáva jazyk R. Jeho funkcionalita sa prezentuje jednak na rôznych vedeckých fórach a rovnako sa jeho výučba zaraďuje do študijných programov na renomovaných univerzitách ekonomického, technického i humanitného zamerania. Príspevok sa zameriava na stručné predstavenie kľúčových výhod použitia tohto softvéru a niektorých jeho neštandardných grafických výstupov. Špecificky sa venuje analýze šikmosti a špicatosti údajov s využitím Cullenovho-Freyovho grafu. Grafické výstupy sú doplnené o syntax príkazov pre danú analýzu.

## ABSTRACT
The R language becomes a phenomenon in software for actuarial analysis. Its functionality is presented both on the various scientific forums and it has been incorporated into various study programmes at prestigious universities (economic, technical and humanities). The aim of this paper is to briefly describe the main benefits of using this software and some of its non-standard graphical outputs. It is specifically devoted to the analysis of the skewness and kurtosis of data with the use of the Cullen-Frey graph. The graphical outputs are complemented by the command syntax for specific analysis.

## KĽÚČOVÉ SLOVÁ
jazyk R, štatistické analýzy, grafy, šikmosť, špicatosť, regresná a korelačná analýza, aktuárstvo

## KEY WORDS
R language, statistical analysis, graphs, skewness, kurtosis, regression and correlation analysis, actuarial science

## 1. ÚVOD
R je programovací jazyk špecializovaný predovšetkým na štatistické výpočty a grafiku. Ide o projekt GNU podobný jazyku S, ktorý vyvinul *v Bell Laboratories* (predtým AT&T, teraz *Lucent Technologies*) John Chambers so svojím kolektívom. V roku 2009 New York Times zverejnil článok, v ktorom jazyk R získal veľké uznanie medzi dátovými analytikmi a naznačil, že predstavuje vážnu hrozbu pre komerčný softvér. R možno považovať za inú implementáciu jazyka S. Jazyk R je voľne dostupný a využíva ho akademická, vedecká i komerčná sféra. Štatistické analýzy možno realizovať už v štandardnej verzii, prípadne po inštalácii konkrétnych podporných balíčkov (*packages*), kde je implementované veľké množstvo pokročilých funkcií. Samozrejmosťou je aj možnosť vytvárať vlastné funkcie a skripty.

Jazyk R sa dá v aktuárskych analýzach využiť pri maticovom počte, simuláciách, vytváraní rôznych rizikových scenárov, maximálne vierohodných odhadoch, práci s rozdeleniami pravdepodobnosti, modeloch kredibility, výpočte technických rezerv, zovšeobecnených lineárnych modeloch a pod. Medzi hlavné výhody systému R patrí:

- **dostupnosť** – open-source programovací jazyk zaradený v rámci projektu GNU nadácie *Free Software Foundation*,
- **kompatibilita** – kompatibilný s operačnými systémami Windows, Linux, Mac,
- **množstvo analytických nástrojov** – pre štatistiku, demografiu, biometriu, taxonómiu, genetiku, geografiu, finančné analýzy a pod. s využitím mnohých vyvinutých doplňujúcich balíčkov s knižnicami funkcií na rôzne typy analýz,
- **aktuálnosť** – rýchle reakcie na vývoj nových metód v štatistike, obsahuje často metódy, ktoré ešte nie sú implementované do klasického komerčného softvéru,
- **grafické výstupy** – štandardné aj nové moderné grafické výstupy s interaktívnym zásahom používateľa vrátane možnosti doplniť matematické vzorce a symboly,
- **študijné materiály** – voľne dostupné materiály a manuály na prácu s jazykom R šírené na webových lokalitách a rovnako aj publikácie vydávané v renomovaných vydavateľstvách,
- **programovanie** – pre pokročilých používateľov ako objektovo orientovaný programovací jazyk na vlastné (náročné) štatistické analýzy.

Je zrejmé, že opísať celú funkcionalitu jazyka R s jeho viacerými aplikáciami v rôznych oblastiach nie je možné. Za cieľ sme si preto stanovili stručne predstaviť niektoré jeho grafické možnosti.

Grafická prezentácia údajov nielen v aktuárstve je významná a prezentovať údaje na relevantnom i zaujímavom grafe by malo byť prioritou každého analytika. Zvoliť sofistikované nové druhy grafov, resp. docieliť lepšiu finálnu vizualizáciu grafov, ktoré sa často používajú, umožňuje práve jazyk R. Nasledujúce kapitoly predstavujú možnosti grafického zobrazenia v analýze portfólia, regresnej a korelačnej analýze a pri analýze šikmosti a špicatosti. V poslednej spomenutej oblasti predstavíme málo známy Cullenov-Freyov graf, ktorý môže byť zaujímavý pri odhadoch šikmosti a špicatosti analyzovaných údajov. Je potrebné si uvedomiť, že každá aplikácia sa v jazyku R štandardne nespúšťa pomocou kontextového menu, ale voľbou naprogramovaných príkazov. V príspevku uvádzame príklady týchto príkazov na zobrazenie grafov. Inštaláciu prostredia jazyka R môže používateľ uskutočniť prostredníctvom odkazu [12].

## 2. NIEKTORÉ GRAFICKÉ MOŽNOSTI JAZYKA R

Po nainštalovaní a spustení jazyka R sa pracuje v konzolovom okne – *R Console*, kde používateľ zadáva funkcie, vkladá objekty a objavujú sa tam aj základné výstupy. Používateľské rozhranie jazyka R disponuje samostatným oknom na prácu s grafikou – *R Graphics*. V tomto okne sa zobrazujú grafy, ktoré sa spustia v konzolovom okne. Používateľ má na výber rôzne 2D, 3D grafy, z ktorých je značné množstvo obsiahnuté už v základnej verzii programového prostredia. Ak sú tieto grafy nedostačujúce, môžeme použiť doplňujúce balíčky, napríklad *grid, lattice, iplots, misc3D, scatterplot, corrgram maps.*

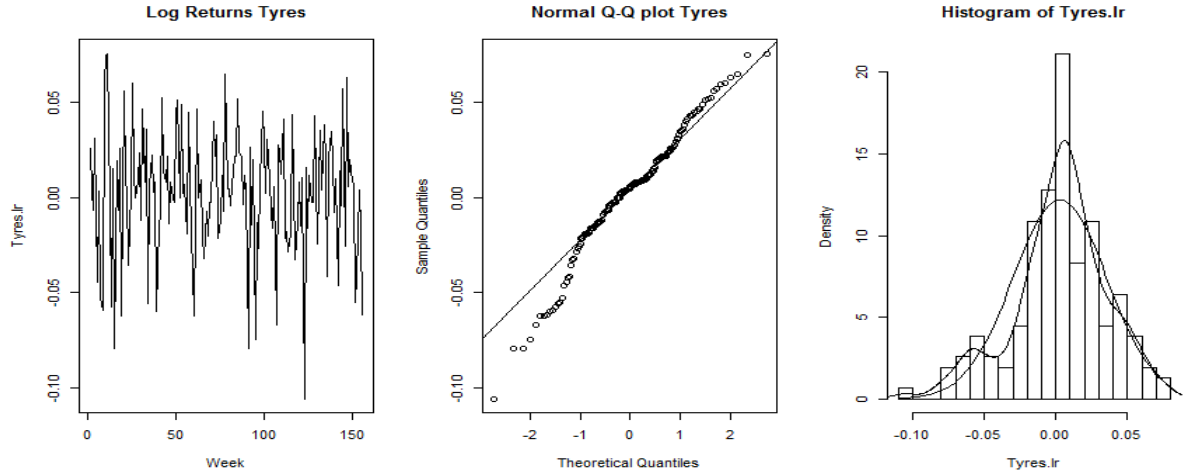
Grafy zobrazujeme pomocou základného príkazu *plot* určeného na ich vytváranie. Tento príkaz obsahuje niekoľko parametrov. Základným je zdrojový objekt, ktorého graf ideme vytvoriť. Pri zadaní len tohto argumentu jazyk R nastaví všetky ostatné

argumenty implicitne a v grafickom okne sa zobrazí výstup. Jednotlivé príkazy, parametre obsahujú názov grafu (*main*), popis grafu (*sub*), názvy osí grafu (*xlab*, *ylab*), hodnoty osí (*xlim*, *ylim*) a typ zobrazovaného grafu (*type*). Jednotlivé písmená parametra *type* zobrazia konkrétny typ grafu (napr. p pre bodový graf).

Objavilo sa mnoho pokusov o vytvorenie grafických rozhraní od editorov kódu (ako *RStudio*) až po plnohodnotné GUI rozhranie (ako je *RCommander*). Viac základných informácií o jazyku R je verejne dostupných na jeho webových lokalitách [12].

Ďalej si ukážeme príklady troch grafických výstupov z rôznych oblastí štatistických analýz. Obrázok 1 opisuje grafickú analýzu vývoja cien akcií, pričom výstup zobrazuje tri grafy: graf vývoja výnosov portfólia za dané časové obdobie, grafické posúdenie normality špecifickým Q-Q grafom (*Q-Q normal plot*) a histogram logaritmov výnosov akcií s krivkou hustoty normálneho rozdelenia (so strednou hodnotou a štandardnou odchýlkou analyzovaných údajov) a skutočnou hustotou pravdepodobnosti analyzovaných údajov. Veľkou výhodou v jazyku R je tiež ľubovoľné poradie a počet zobrazovaných grafov, ktoré chceme použiť vo finálnom výstupe, napr. vo vedeckom článku (na obrázku č. 1 v zobrazení 1 : 3). Vzhľadom na rozsiahlejšiu analýzu zdrojový kód (funkcie pre analýzu a samotný grafický výstup) neuvádzame. Pre viac informácií pozri napríklad [4], [6].

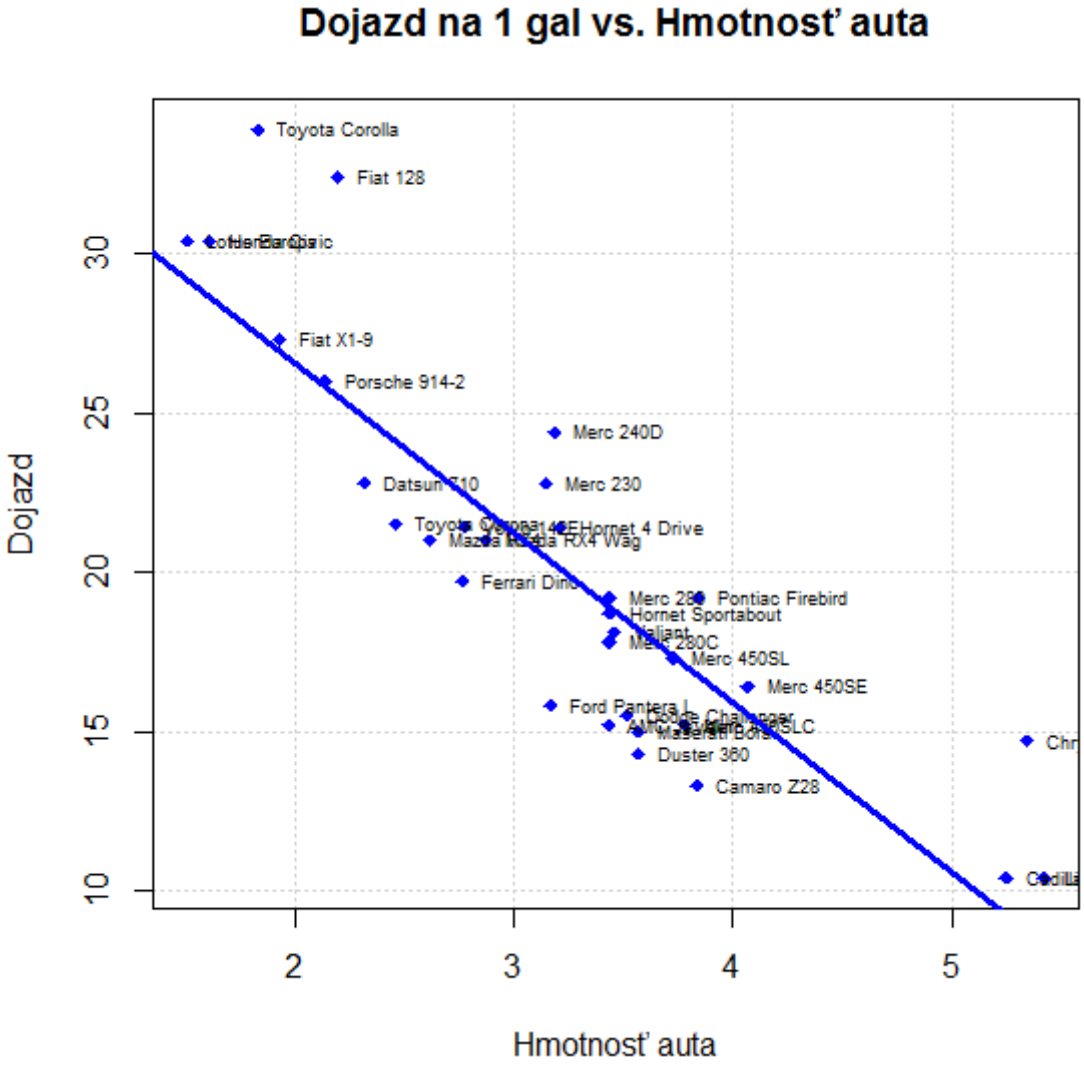
***Obrázok č. 1: Ukážka grafickej analýzy portfólia CP v jazyku R***




***Zdroj: [4]***

Zaujímavým doplnkom (obrázok č. 2) pri regresnej analýze a konštrukcii lineárneho regresného modelu môže byť doplnenie textu (funkcia *text*) do grafu lineárneho regresného modelu k jednotlivým závislostiam.

***Obrázok č. 2: Zobrazenie textu v grafe závislostí v lineárnom regresnom modeli***




***Zdroj: vlastné spracovanie***

Výstupom obrázka č. 2 je zobrazenie dojazdu automobilu na 1 galón v závislosti od jeho hmotnosti s priamkou vyrovnaných hodnôt. Použitý súbor (údajový súbor *mtcars*) je štandardnou súčasťou jazyka R a obsahuje reálne údaje z magazínu *Motor Trend USA* z roku 1974, ktoré môže používateľ pri analýzach použiť. Popis týchto údajov sa dá získať príkazom v tvare *??mtcars*. Ďalej uvádzame ukážku zdrojového kódu pre daný výstup:

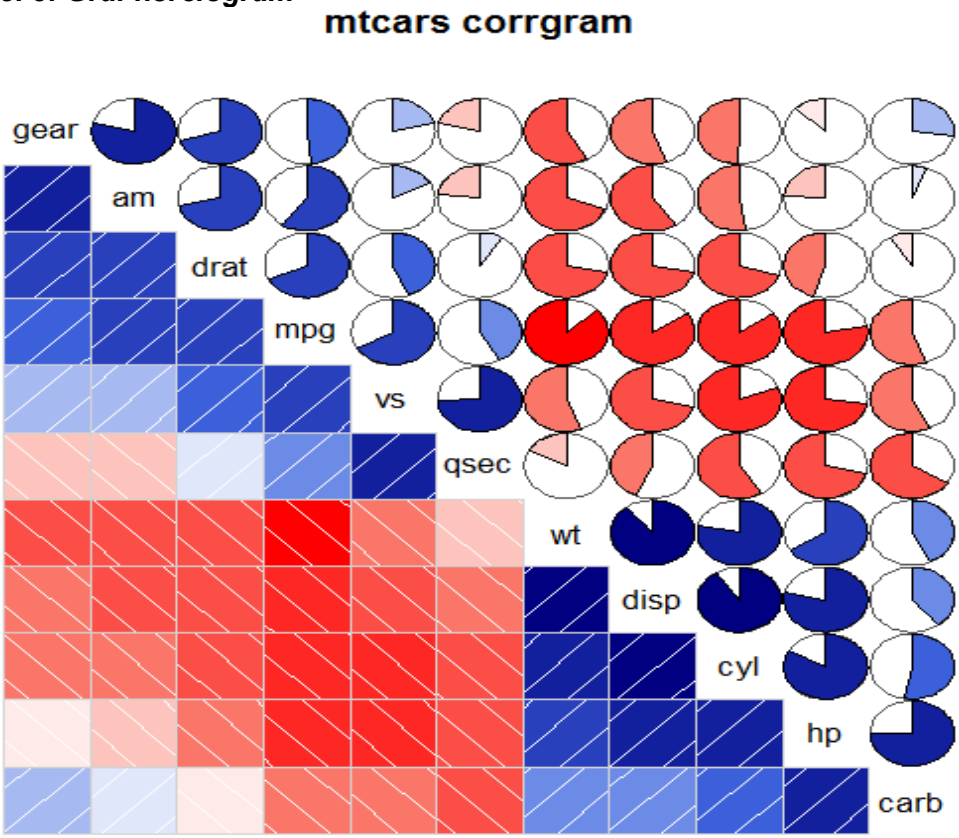
```
attach(mtcars)
mtcars.lm<-lm(mpg~wt,data=mtcars)
plot(wt,mpg,main="Dojazd na 1 gal vs. Hmotnosť auta", xlab="Hmotnosť auta",
ylab="Dojazd",pch=18,col="blue",grid())
text(wt,mpg,row.names(mtcars),cex=0.6,pos=4,col="black")
abline(mtcars.lm,lwd=3,col=4)
```

Posledným výstupom (obrázok č. 3) je korelogram pre identický údajový súbor *mtcars*, ktorý je relatívne novým nástrojom na vizualizáciu údajov v korelačnej analýze. Ak chceme interpretovať tento graf, začnime s hodnotami vľavo dole. Čiary s modrou farbou, ktoré smerujú z ľavého dolného rohu do horného pravého rohu, predstavujú pozitívnu koreláciu medzi dvoma premennými. Naopak, červená farba a

čiary, ktoré vedú z ľavého horného rohu do pravého dolného rohu, predstavujú negatívnu koreláciu. Čím tmavšia a sýtejšia farba, tým je korelácia väčšia. [3] Podrobný opis výstupu pozri v [11]. Na jeho zobrazenie je už potrebná inštalácia doplnkového balíčka *corrgram* a následne píšeme zdrojový kód s využitím rovnomennej funkcie *corrgram:*

```
attach(mtcars)
library(corrgram)
corrgram(mtcars,order=TRUE,lower.panel=panel.shade,upper.panel=panel.pie,text.panel=pa
nel.txt,main="mtcars corrgram")
```

***Obrázok č. 3: Graf korelogram***




***Zdroj: vlastné spracovanie***

## 3. CULLENOV-FREYOV GRAF PRI ANALÝZE ŠIKMOSTI A ŠPICATOSTI

Analýza šikmosti a špicatosti má nezastupiteľné miesto v aktuárskych analýzach (napríklad analýza výnosov finančných aktív, výber portfólia, oceňovanie aktív, analýza extrémnych škôd, analýza rezerv v neživotnom poistení a pod.). V tejto kapitole ukážeme, ako sa dá analyzovať šikmosť a špicatosť pomocou Cullenovho-Freyovho grafu (uvedený v roku 1999, [1]) dostupného v jazyku R. Cullenov-Freyov graf možno použiť pri výbere vhodného rozdelenia skúmaných údajov.

Na analýzu šikmosti a špicatosti náhodnej premennej *X* štandardne využívame koeficient šikmosti a špicatosti, prípadne grafické techniky. Pre rozdelenia s ťažkým chvostom tieto koeficienty nemusia existovať. Pomocou funkcie *descdist* z balíčka *fitdistrplus* vygenerujeme Cullenov-Freyov graf, ktorý zobrazuje empirické odhady koeficientu šikmosti a špicatosti pre základné rozdelenia pravdepodobnosti (triedy gama pre spojité rozdelenia a Poissonovej triedy pre diskrétne rozdelenia).

Rozdelenia len s jednou možnou hodnotou koeficienta šikmosti a špicatosti (napr. normálne, rovnomerné, logistické, exponenciálne rozdelenia) sú tieto koeficienty súradnicovo zobrazené bodkou na ploche grafu (obrázok č. 4). Pri iných rozdeleniach (napríklad gama a lognormálne rozdelenie) sa použili kontúry alebo väčšie plochy (napríklad pre rozdelenie beta). Weibullovo rozdelenie nie je zastúpené v grafe, ale jeho orientáciu naznačuje legenda grafu. Ak v syntaxe funkcie zvolíme *discrete = TRUE,* zobrazia sa rozdelenia diskrétne a naopak. Túto analýzu možno doplniť aj bootstrapovou analýzou pri neistote odhadu. Pre ďalšie informácie o tejto funkcii pozri napr. [10].

Demonštrujme jednoduchú ukážku so zdrojovým kódom v jazyku R. Prvotne nasimulujme 10 000 údajov, ktoré sa riadia normálnym rozdelením, teda napríklad $X \sim N(0;2^2)$. Uložíme ich do objektu *x*, pričom využijeme funkciu *rnorm,* kde r = random, norm = normal distribution; *x<-rnorm(10000,0,2),*následne po spustení balíčka *fitdistrplus* a funkcie *descdist*

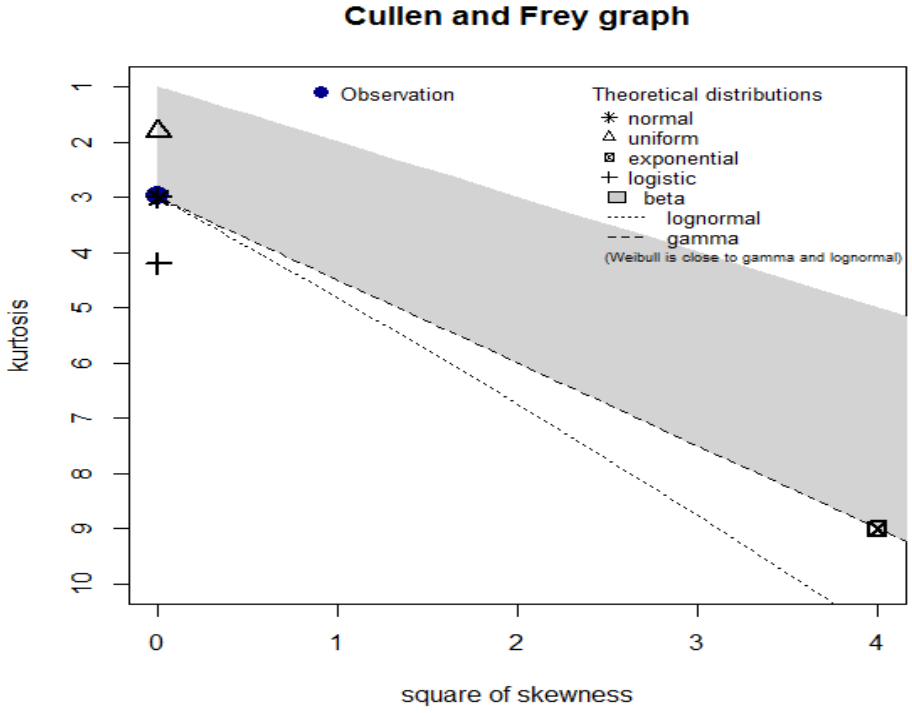
```
library(fitdistrplus)
descdist(x)
```

dostávame tento textový výstup:

```
summary statistics
------
min:  -6.707957   max:  7.964508
median:  -0.01665936
mean:  0.01913099
estimated sd:  2.00055
estimated skewness:  0.03302886
estimated kurtosis:  2.963445,
```

ktorému zodpovedá grafický výstup obrázka č. 4.

***Obrázok č. 4: Cullenov-Freyov graf, ak X~N(0;4)***




***Zdroj: vlastné spracovanie***

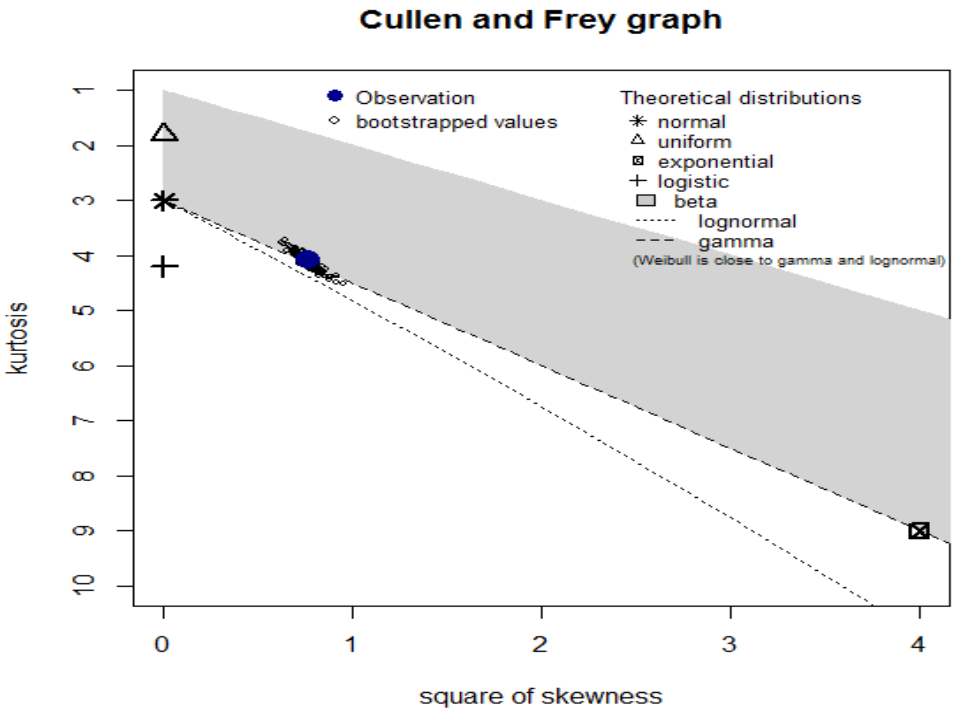
Je zrejmé, že pre normálne rozdelenie platí: koeficient šikmosti sa rovná 0 a špicatosti 3 (resp. v závislosti od literatúry 0), čo potvrdzuje modrá bodka (•) v grafe. To, že ide o normálne rozdelenie, zobrazuje poloha bodky pri značke hviezdička (*).

Komparatívne nasimulujme rovnaký počet hodnôt gama rozdelenia, teda napríklad $X \sim \Gamma(5;0,1)$. Dostávame výstup (obrázok č. 5), v ktorom je poloha bodky posunutá k značke priamky gama rozdelenia. Pri tejto situácii sme navyše zvolili aj ukážku bootstrapového odhadu (značka (o)):

```
descdist(rgamma(10000,5,0.1),boot = 100, boot.col = "black")

summary statistics
------
min:  4.122676   max:  189.4649
median:  46.70345
mean:  49.78119
estimated sd:  22.08738
estimated skewness:  0.8691677
estimated kurtosis:  4.166519 .
```

***Obrázok č. 5: Cullenov-Freyov graf, ak X~ Γ(5;0,1)***



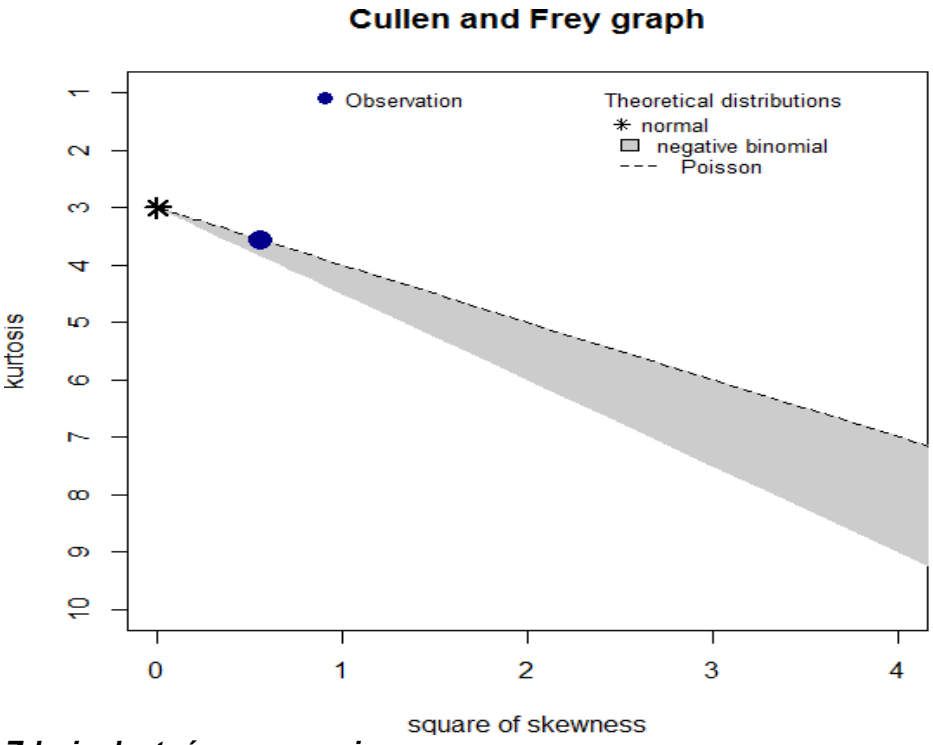

***Zdroj: vlastné spracovanie***

Analýzu uzavrieme ukážkou diskrétneho rozdelenia, keď napr. $X \sim Po(2)$ graf vyvoláme príkazom:

```
descdist(rpois(10000,2),discrete=TRUE).
```

Dostávame nasledujúci grafický výstup (obrázok č. 6).

***Obrázok č. 6: Cullenov-Freyov graf, ak X~ Po(2)***




***Zdroj: vlastné spracovanie***

summary statistics
---
min: 0 max: 10
median: 2
mean: 1.9931
estimated sd: 1.433896
estimated skewness: 0.7470107
estimated kurtosis: 3.585507 .

Textový výstup odhadu koeficienta šikmosti a špicatosti, ktorý vidíme súradnicovo v grafe č. 6, môžeme v jazyku R vyjadriť aj výpočtom. Ak $X \sim \text{Po}(\lambda)$, tak pre koeficient šikmosti platí $\lambda^{-½}$ (0,7071 ≈ 0,7470) a koeficient šikmosti $\lambda^{-1}$ (0,5 ≈ 3,5855 – 3).

## 4. ZÁVER

Aktuár sa vo svojej činnosti zaoberá najmä oblasťou analýzy rizika, a to tak v neživotnom, ako aj v životnom poistení. Budovanie osobnostného vedomostného aparátu je podmienené jeho poznatkami z inžinierskej matematiky, teórie pravdepodobnosti, štatistiky, ale tiež všeobecnej ekonomickej teórie, práva a účtovníctva. Analýzy sa často nezaobídu bez využitia rôzneho softvéru – či už komerčného (licenciovaného), alebo tzv. open-source. Vývojárske spoločnosti si konkurujú zvyšovaním atraktivity používateľského prostredia, grafickými výstupmi a najmä výpočtovými možnosťami softvérov. Popri balíku MS Office, z ktorého MS Excel je dlhodobo najpoužívanejším a najrozšírenejším kalkulátorom na aktuárske výpočty vôbec, sa na softvérovom trhu objavujú aj ďalšie produkty vhodné pre aktuárov, ako napr. SAS (SAS® for Insurance), ModelRisk, @Risk, DFA, EMB, Oracle, Matlab, z nekomerčných spomeňme napr. jazyk R. Tieto špecifické softvéry bývajú vyvinuté pre rôzne špecifické oblasti poisťovne (rezervy, finančné

modelovanie, cenotvorba, zaistenie, štatistické spracovanie a vyhodnotenie dát, nástroje na meranie a analýzu rizík a pod.).

V tomto príspevku sme predstavili práve open-source jazyk R. V prvej časti sme čitateľa oboznámili s niektorými dôležitými výhodami jazyka R a ďalej sme prezentovali najmä niektoré jeho nie celkom bežné grafické nástroje vhodné na štatistické analýzy v komerčnej, vedeckej i akademickej sfére. Väčšiu pozornosť sme venovali analýze šikmosti a špicatosti pomocou Cullenovho-Freyovho grafu, pričom sme na ukážku využili nasimulované údaje z rôznych spojitých a diskrétnych rozdelení. Jazyk R dokáže spájať grafy, ktoré si používateľ vyberie v rôznych zoskupeniach v riadku alebo stĺpci. Grafy možno ľubovoľne upravovať a škálovať, čo pri komerčnom softvéri je často prednastavené. Grafické výstupy sú tiež v kvalitnom rozlíšení. Používateľ môže do grafov aj pridávať rôzne ďalšie doplňujúce objekty (text, krivky, matematické symboly, vzorce, popisky a pod.), čo môže len prispieť k lepšej vypovedacej schopnosti daného grafu. Analýza špicatosti a šikmosti Cullenovým-Freyovým grafom je zase príkladom použitia už vytvoreného doplňujúceho balíčka, po ktorého nainštalovaní používateľ získava nové možnosti (analytické i grafické) na svoje analýzy. Tieto balíčky vytvárajú rôzni odborníci z praxe a ich dostupnosť je pre používateľa jazyka R bezplatná. Ku každému balíčku je vytvorený manuál a používatelia majú možnosť s ich autormi viesť elektronické diskusie.

## LITERATÚRA


[1] CULLEN, A. – FREY, H.: Probabilistic Techniques in Exposure Assessment. New York: Springer, 1999. ISBN 978-0306459573.

[2] CHARPENTIER, A.: Computational Actuarial Science with R. New York: CRC Press, 2015. ISBN 978-1-4665-9259-9.

[3] KIŠOŇ, B.: Spracovanie údajov využitím systému R: Bakalárska práca. Bratislava: Ekonomická univerzita v Bratislave, 2015.

[4] PÁLEŠ, M.: Využitie software pri výučbe predmetov z oblasti aktuárstva. In: MITAV 2014. Brno: Univerzita obrany v Brně, 2014. SBN 978-80-7231-961-9.

[5] PÁLEŠ, M.: Generating a Pseudo-Random Automobile Insurance Portfolio in R. In: Managing and Modelling of Financial Risks. 7th International Scientific Conference: 8th-9th September 2014 Ostrava, Czech Republic. Ostrava: VŠB – Technical University of Ostrava, 2014. ISBN 978-80-248-3631-7.

[6] PÁLEŠ, M.: R language graphical support for the analysis of portfolio. In: Praktické využívanie softvérovej podpory v oblasti aktuárskych vied. Vedecká konferencia 30. júna – 2. júla 2014, vzdelávacie zariadenie EU, Virt, Slovensko [elektronický zdroj]. Bratislava: Vydavateľstvo EKONÓM, 2014. ISBN 978-80-225-3872-5.

[7] PÁLEŠ, M.: Aproximácia binomického rozdelenia normálnym a príklad jej aplikácie v aktuárstve s využitím jazyka R [elektronický zdroj]. Bratislava: Ekonomická univerzita v Bratislave, 2015; http://fhi.sk/files/katedry/km/veda-vyskum/prace/2015/Pales3.pdf.

[8] PÁLEŠ, M.: Panjerove rekurentné vzťahy v prostredí jazyka R. In: Ekonomika a informatika: vedecký časopis FHI EU v Bratislave a SSHI [elektronický zdroj]. Bratislava: Ekonomická univerzita v Bratislave, 2015, č. 1. ISSN 1336-3514.

[9] http://www.R-project.org/, prístup k 2015-08-06.

[10] http://www.jstatsoft.org/v64/i04/, prístup k 2015-08-06.

[11] WRIGHT, K.: corrgram: Plot a Correlogram. R package version 1.8 http://CRAN.R-project.org/package=corrgram, prístup k 2015-08-06.
[12] http://cran.at.r-project.org, prístup k 2015-08-06.

**RESUME**
The paper provides basic information about graphical interface of the R language. We used this software for practical applications in the portfolio, regression and correlation analysis. We presented options of the specific graphical outputs of this open-source system and in a separate part we introduced the Cullen-Frey graph for the skewness and kurtosis analysis. Outputs for various probability distributions were described on the basis of simulated data and the results were also confirmed by manual calculations. Our goal was to demonstrate the use of the R language in various statistical or actuarial analyses of skewness within which the presented analysis plays an important role.

***PROFESIJNÝ ŽIVOTOPIS***
***Ing. Michal Páleš, PhD.,*** *od roku 2012 pôsobí ako odborný asistent (sekcia aktuárskych vied) a tajomník Katedry matematiky a aktuárstva Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. V rámci pedagogickej činnosti vyučuje cvičenia k predmetom matematika, vybrané kapitoly z matematiky, teória pravdepodobnosti, teória rizika v poistení a programovacie techniky pre aktuárov. Vo svojej vedeckej práci sa orientuje na využitie matematickoštatistických metód v ekonómii a teóriu rizika v neživotnom poistení (Panjerove rekurentné vzťahy, rozdelenia pravdepodobnosti využívané v aktuárskej praxi, softvérová podpora riadenia rizík, najmä jazyk R).*

**KONTAKT**
pales.euba@gmail.com