Pravděpodobnost a statistika s programováním v R

Aleš Kozubík

Abstrakt: Příspěvek je věnován přípravě základního kurzu pravděpodobnosti a statistiky s podporou open source programovacího prostředí R. Stručně představuje předpokládanou obsahovou náplň připravovaného kurzu s důrazem na programování v R. Autor v článku uvádí zásadní důvody pro výběr právě tohoto prostředku. V druhé části pak ilustruje vyžití nástroje na práci s aktuálními reálnými daty.

Abstract: The paper is concerned with preparing an elementary course in probability and statistics with the support of the open-source programming environment R. It briefly presents the expected content of the planned curriculum with an emphasis on programming in R. The author presents the main reasons for choosing this particular tool. In the second part, he illustrates the use of this measure to work with actual real data.

1 Úvod

Znalost problematiky matematické statistiky je nezbytná pro všechny technické, ekonomické ale i přírodovědné či humanitní odbory studia. I když se ve většině případů snažíme přírodní zákonitosti postihnout deterministickými modely a exaktními zákonitostmi, nelze se statistickým šetřením vyhnout. Existuje totiž mnoho procesů, jenž nejsme schopni, ať již pro jejich složitost nebo velké množství faktorů, které je ovlivňují, popsat deterministickými zákonitostmi. V takových případech se odvoláváme na působení náhodnosti a na naše historické empirické zkušenosti. Ale i všechny přírodní zákonitosti, které se dnes učíme jako deterministické popsané matematickými vztahy se rodily na základě pozorování a experimentování. jsou tedy rovněž výsledkem statistických metod, ač si to již při jejich prezentaci neuvědomujeme. Jako příklady procesů, jimž stále přisuzujeme nahodilý charakter, uveďme alespoň několik z nich.

Typickým příkladem jsou modely rizika ve financích a pojišťovnictví. Zejména v pojištění, ať již životním nebo neživotním, se smlouvy vážou na události náhodného charakteru. Mohlo by se zdát, že otázka úmrtí pojištěnce je nenáhodného charakteru, ale je potřebné zdůraznit, že okamžik úmrtí má náhodný charakter. S tím je spojena i otázka očekávané délky života, což je jedna z důležitých demografických charakteristik s celoplošným dopadem v oblasti sociálního zabezpečení. Stejně tak i v neživotním pojištění je otázka počtu škodových událostí řešena jako náhodná a rovněž i velikosti vzniklé škody se přisuzuje náhodný charakter.

 DOI: https://doi.org/10.11118/978-80-7509-898-6-0040

Pro jiné příklady můžeme odskočit do oblasti dopravy. I zde dochází ke mnoha událostem, jenž jsou obtížně předvídatelné. Každý se již zcela jistě setkal se zpožděním vlaků, avšak nikdo nedovede apriori předpovědět kdy k němu dojde a zejména v jakém rozsahu. Lze je předpovědět při výlukových pracích nebo nepříznivé předpovědi počasí, ale reálnou časovou prodlevu lze jen odhadovat na základě zkušeností z minulosti. Již vůbec nelze spolehlivě předpovědět kdy dojde k dopravním zácpám na dálnici v důsledku nehody a jak velké zdržení budou představovat.

Podobně jako v dopravě, i v provozu na počítačových sítích lze pozorovat procesy, které mají náhodný charakter. Může jít o návštěvnost jednotlivých webových stránek, počet došlých požadavků na server za určitou časovou jednotku a pod. Samozřejmostí je dnes i personalizace reklamy v prostředí internetu, které se rovněž opírá o statistické analýzy navštěvovaných stránek či obsahu vyhledávání ve vyhledávacích službách.

Nelze opomenout ani oblast automatizace výrobních procesů. Délka trvání jednotlivých technologických procesů je nezbytným prvkem pro úspěšné řešení automatizace. Zrovna tak životnost výrobků a jejich poruchovost jsou předmětem statistických analýz a statistické kontroly jakosti.

Z humanitních oborů uveďme alespoň oblast medicíny, kde právě medicínské testy a hodnocení účinnosti léčiv se opírají o statistické metody. Konec-konců i mnohé diagnostické metody lze interpretovat jako aplikovanou statistiku. Klinická psychologie je rovněž založena na statistických analýzách a dokonce se v mnoha případech sama stala základem pro rozvoj statistiky jako takové. Bez statistiky se neobejde ani tak zdánlivě vzdálená disciplína jako je politologie. Dnes tolik populární průzkumy volebních preferencí nejsou ničím jiným, než výsledkem statistického šetření.

Úhrnem lze tedy konstatovat, že všude tam, kde je nějaká věda patří statistika. To ale klade požadavky na vzdělávání ve všech vědních odborech. Porozumění a zejména schopnost správné interpretace výsledků statistického šetření by měly patřit do profesní výbavy každého vysokoškolsky vzdělaného odborníka.

2 Obsah kurzu

Z předchozího úvodu plyne, že kurz základů statistiky by měl být zařazen do převážné většiny studijních programů na vysokých školách. V rámci řešeného mezinárodního projektu „Innovative Open Source courses for Computer Science curriculum“ se při tom zaměřujeme zejména na podporu výuky otevřenými softwarovými prostředky.

Samotná obsahová náplň předmětu nikterak nevybočuje z běžně vyučovaných kurzů statistiky. V úvodních lekcích se posluchači seznamují se základy teorie pravděpodobnosti. Cílem této části je zejména seznámit účastníky kurzu s rozděleními pravděpo-

dobnosti používanými při intervalových odhadech, statistických testech hypotéz a se zákony velkých čísel. Svým rozsahem se shoduje se zahraničními učebnicemi [7],[2] nebo s domácími učebnicemi [3], [6] a [4].

Druhá část kurzu je věnována samotné matematické statistice. V teoretické části představuje výběrové charakteristiky, základy teorie odhadu, parametrické a neparametrické testy hypotéz a základy korelační a regresní analýzy.

Praktická část připravovaného kurzu je pak zaměřena na seznámení s prvky programovacího prostředí jazyka R. V rámci laboratorních cvičení se absolventi nejprve seznamují s prvky jazyka R, zejména s implementovanými datovými typy a strukturami dat. Jako podporu pro výuku pravděpodobnosti se seznámí s vybranými diskrétními a spojitými rozděleními pravděpodobnosti, pro něž jsou v jazyce R implementovány distribuční funkce, hustoty, kvantilové funkce a také generátory náhodných hodnot z daného rozdělení.

Důležitým prvkem prostředí R je bohatá škála nástrojů pro vizualizaci dat. Proto je této části věnována mimořádná a rozsáhlá pozornost. Posluchači se postupně naučí vytvářet i pokročilé nástroje grafické prezentace dat, jako jsou histogramy, sloupcové a koláčové grafy, box ploty, kvantilové grafy. Součástí této kapitoly jsou i bohaté možnosti formátování výsledných grafů.

Následně se studenti naučí práci se vstupními a výstupními soubory a vytváření vlastních funkcí. Tím se vytvářeny předpoklady pro jejich možnost samostatné práce s dostupnými reálnými daty, dostupnými z internetu. v rámci analýzy těchto dat se naučí využívat implementované nástroje pro testování statistických hypotéz, které si rovněž mohou vyzkoušet experimentováním s reálnými daty. V samotném závěru si pak ověří i jednoduchost korelační analýzy při použití implementovaných funkcí a různé formy regresních modelů.

Praktická část kurzu se opírá o literární zdroje [1] a [8]. Pro seznámení se samotným jazykem R pak slouží například učebnice [5] nebo [9].

3 Proč R

Jedním z charakteristických rysů statistického šetření a analýzy získaných dat je hromadnost. V dnešní době si lze jen stěží představit jejich zpracování bez využití výpočetní techniky. Je proto přirozené požadovat, aby každý kurz statistiky byl organicky propojen s využitím vhodných softwarových nástrojů. Při řešení otázky, jaké nástroje použít máme na výběr vícero možností.

Jednou alternativou je spolehnout se na kancelářské balíky a tabulkové výpočty. Při tom se obvykle vychází z teze, že ten MS Excell každý už jaksi zná a tak zvládne jakékoliv výpočty. Avšak při práci s rozsáhlejšími soubory dat se už používání tohoto typu aplikací ukazuje jako nepraktické a mnohdy je daný objem dat nezvládnutelný.

Jinou volbou mohou být komerční nástroje pro řešení statistických úloh, jako je například *Statistica* nebo programovací jazyk S-plus. Jde sice o výkonné nástroje, tomu ale zodpovídá i jejich cena.

Nabízí se i třetí varianta a to použít svobodné otevřené nástroje. proto jsme se rozhodli pro programovací jazyk R, jenž lze označit za implementaci programovacího jazyka S pod svobodnou licencí. Tento jazyk svojí popularitou a počtem uživatelů již předstihl komerční S a stalo se faktickým standardem v řadě oblastí statistiky. Každý si ho může zdarma nainstalovat z archivu na adrese `https://cran.r-project.org`. Zde je dostupný pro všechny běžné platformy operačních systémů. Navíc, v případě zájmu lze k němu instalovat i GUI R-studio.

Samozřejmě, bezplatnost není jeho jedinou předností. Ty které nás nejvíce ovlivnily při výběru tohoto nástroje lze shrnout do několika jednoduchých bodů:

- je bezplatný, většina platforem statistického softwaru stojí tisíce dolarů,
- k programu je dostupné velké množství rozšiřujících balíčků,
- R dokáže snadno importovat údaje z různých zdrojů,
- R má implementovaných mnoho pokročilých statistických nástrojů,
- R poskytuje interaktivní platformu na analýzu údajů,
- prostředí R nabízí vizualizaci dat v podobě vysoce kvalitních a estetických grafů,
- je nezávislé na platformě, kompatibilní s většinou nejrozšířenějších operačních systémů,
- je kompatibilní s programovacími jazyky jako C,C++, Python, Java.

Nelze opomenout ani skutečnost, že student získává další programátorskou zručnost a základní znalost jazyka specializovaného na statistické výpočty. Navíc jeho použití není vázáno na zakoupení žádné licence, takže i v budoucnu ho lze bez obtíží používat resp. nainstalovat na jakýkoliv počítač. To přináší jistý typ svobody, jenž při vazbě na konkrétní komerční nástroj absentuje.

4 Ukázka práce v prostředí R

Jak už bylo zmíněno, programovací prostředí R poskytuje možnost interaktivní práce. Proto uvedeme ukázku jeho použití po jednotlivých příkazech. Tyto příkazy lze ale rovněž uložit do souboru s příponou `.R` a následně spustit jako skript jazyka R.

V současnosti je aktuální vizualizace a analyzování různých dat souvisejících se šířením onemocnění COVID-19. Ani my nebudeme v tomto směru výjimkou a ilustrujeme některé prvky jazyka R právě na těchto datech. Nejprve si načteme data o pandemii v rámci SR ze sídla, kde jsou uloženy ve formátu `.csv`. Toho dosáhneme pomocí následující funkce:

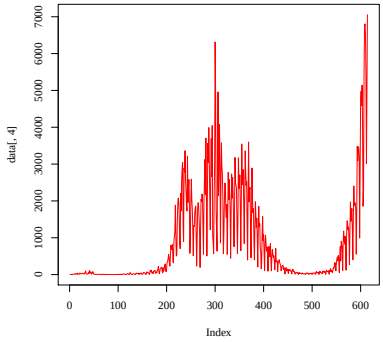
```
data<-read.csv("https://mapa.covid.chat/export/csv",header=T,sep=";")
```

Jejími argumenty jsou adresa vzdáleného zdroje dat, a informace o existenci záhlaví v datovém souboru a oddělovači hodnot, jímž je v tomto případě středník. O názvech jednotlivých položek struktury `data frame` se můžeme přesvědčit zavoláním funkce `names(data)`. Tak se dozvíme, že jednotlivé sloupce obsahují datum, počty potvrzených případů, počet vykonaných PCR testů, denní přírůstky a počet úmrtí. Chceme-li si jednoduše graficky prezentovat vývoj denních přírůstků, stačí zavolat funkci `plot()`. Jejími argumenty budou: čtvrtý sloupec struktury `data` a další argumenty pak upravují barevnost a typ zobrazené křivky v grafu.

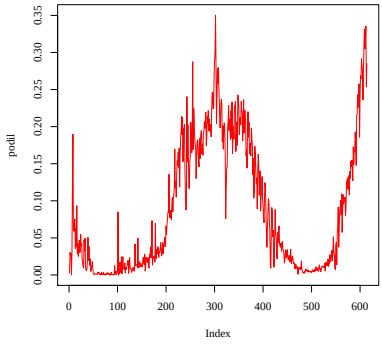
```
plot(data[,4],type="l",col="red")
```

Výsledek vidíme na obrázku 1. Absolutní velikost denních přírůstků však nemá zcela zásadní výpovědní hodnotu, protože počet případů je zcela jistě podmíněn počtem provedených testů. proto je lépe zobrazovat podíl pozitivních případů na provedených testech. ten si ovšem musíme nejprve přepočítat a až po té zobrazit. Výhodou v prostředí R je vektorová implementace funkcí, jež se vykonávají po složkách, což zjednodušuje zápis operace, jak ilustruje následující kód a jeho grafický výstup na obrázku 2.

```
podil<-data[,4]/data[,3]
plot(podil,type="l",col="red")
```



Obr. 1: *Graf přírůstků pozitivních PCR testů.*



Obr. 2: *Graf procentních podílů pozitivních na celkovém počtu testů.*

Tyto grafy však mají jisté nedostatky. Především si všimněme popisu jednotlivých os, kde bychom rádi doplnili informaci o údajích, jež graf prezentuje. Dále bychom jako značky na ose `x` asi rádi viděli datum. Toho lze dosáhnout potlačením vykreslení os ve funkci `plot()` nastavením na logickou hodnotu `FALSE`. Pomocí argumentů `xlab` a `ylab` definujeme popis jednotlivých os. Osy samotné pak vykreslíme pomocí funkce `axis()`.

Pro ukázku jsme z celého datasetu vybrali určitý fragment, který reprezentuje období, kdy na Slovensku proběhlo celoplošné testování a dodnes se vedou spory o tom, zda došlo k zastavení šíření infekce nebo naopak k navýšení pozitivních případů. Pro tento účel jsme vybrali úsek krátce před termínem testování, ke kterému došlo 1. 11. 2020 a druhé kolo 8. 11. 2020. Vývoj pak sledujeme až do konce roku 2020. Tyto údaje se nacházejí v data framu `data` na pozicích 230–300, jež musíme vyselektovat, např. definováním proměnné `pozice`, což ilustruje následující kód.

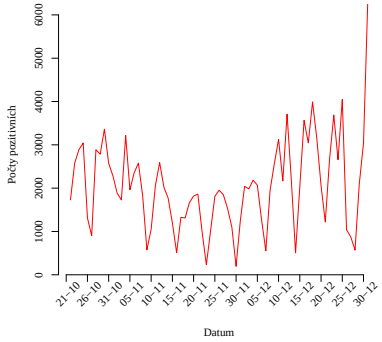
```
pozice<-seq(from=230,to=300,by=5)
plot(data[230:300,4],type="l",col="red",axes=FALSE,xlab="Datum",
ylab="Počty pozitivních",ylim=c(0,6000))
axis(2,at=seq(from=0,to=6000,by=1000),labels=seq(from=0,to=6000,by=1000))
axis(3,at=seq(from=0,to=70,by=5),labels=FALSE,pos=0)
```

Pro popis horizontální osy musíme nejdříve vytvořit příslušné popisy. Tyto datumy jsme uložili do proměnné `lablist`, přičemž jsme odstranili zbytečně dlouhý údaj o letopočtu, jenž je ve všech případech stejný, tady 2020. Popis pak přidáme k ose `x` pomocí funkce `text()`, která slouží k umístění textu na libovolnou pozici na kreslící ploše. Pomocí argumentu `srt` pak lze upravit i sklon textu. Výsledek vidíme na obrázku 3.

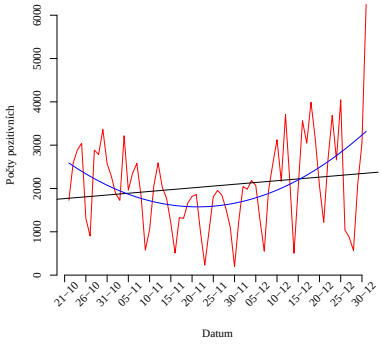
```
lablist<-substr(data[pozice,1],1,5)
text(x=seq(from=0,to=70,by=5),par("usr")[3] - 0.2,labels=lablist,srt=45,
pos=1,xpd = TRUE)
```

```
dny<-seq(0,70)
```

Důležitým nástrojem ve statistice je regresní analýza, jež se v základním kurzu vyučuje jen v rozsahu lineární regrese. Ta je v prostředí R implementována pomocí funkce `lm()`, což velmi usnadňuje její provedení. příslušnou regresní přímku pak lze v hotovém grafu přidat pomocí funkce `abline()`. Pokud bychom chtěli do regrese zařadit i vyšší mocniny, požijeme funkce `poly()`, kde v našem případě hodnota 2 udává, že jde o vyrovnání parabolou. Výslednou křivku pak zviditelníme pomocí funkce `curve()` s nastavením hodnoty parametru `add=T`, což opět ovlivní dokreslení křivky do již existujícího grafu.



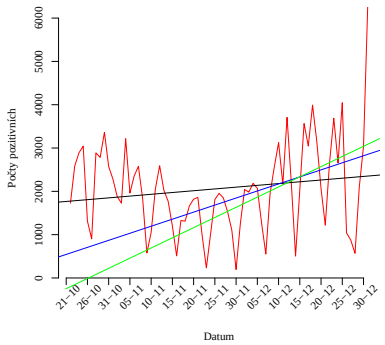
Obr. 3: *Graf podílů pozitivních PCR testů ve vybraném období roku 2020.*



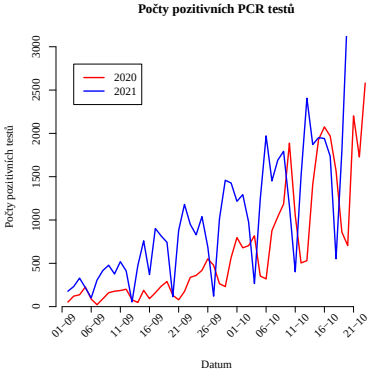
Obr. 4: *Graf podílů pozitivních PCR testů ve vybraném období roku 2020 s vyrovnáním přímkou a parabolou.*

```
rust_pocty<-lm(data[230:300,4]~dny)
abline(rust_pocty)
fit <- lm(data[230:300,4]~poly(dny,2,raw=TRUE))
curve(predict(fit,newdata=data.frame(dny=x)),add=T,col="blue")
```

Výsledný graf po přidání výsledných regresních křivek pak vidíme na obrázku 4. I když je z grafu zřejmé, že se nejedná o lineární vztah, nýbrž jsou zde výrazná sezónní minima víkendových dnů a nárůsty bezprostředně po nich, zejména parabola naznačuje výrazný nárůst pozitivních testů po uplynutí inkubační doby po termínech celoplošného testování.



Obr. 5: *Graf podílů pozitivních PCR testů ve sledovaném období se zobrazením trendů za několik období po plošném testování.*



Obr. 6: *Porovnání počtů pozitivních PCR testů ve stejných dnech v roce 2020 a 2021.*

Ještě lépe lze změny trendu v růstu počtu podílu pozitivních testů porovnat, pokud zobrazíme regresní přímky určené pro období před testováním (černá přímka na obrázku 6), a regresní přímky určené po uplynutí 10 dní od prvního testování (modrá přímka na obrázku 6) a 10 dní po druhém testování (zelená přímka na obrázku 6). Vývoj potvrzuje zrychlování v podílech pozitivních případů na celkovém počtu případů. Zmíněné regresní přímky přidáme do obrázku pomocí kódu

```
rust_podil<-lm(podil[250:300]~seq(20,70))
abline(rust_podil,col="blue")
rust_podil<-lm(podil[260:300]~seq(30,70))
abline(rust_podil,col="green")
```

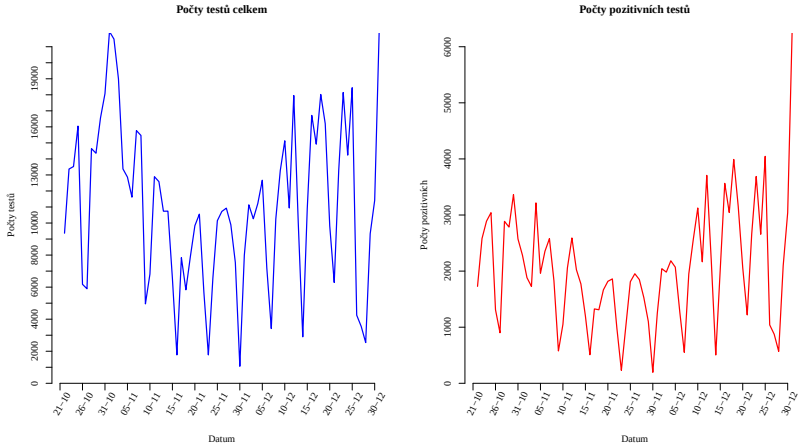
V současnosti je velice aktuální porovnávání vývoje letošní a loňské vlny. To lze provést zobrazením dvou křivek, odpovídajících stejným dnům v roce, do jednoho grafu. Nejprve definujeme proměnnou `pozice`, jež obsahuje údaje z období od 1.září 2020 do 21.října 2020. Do grafu zakreslíme příslušnou křivku červeně, přičemž uvedením proměnné `main` funkce `plot()` přidáme do grafu hlavní nadpis. K zobrazení údajů ze stejného období roku 2021 posuneme rozmezí indexů o 365 dní. Příslušnou křivku zobrazíme pomocí funkce `lines()`, čímž dojde k jejímu zobrazení ve stejném obrázku. K tomu požijeme následující kód:

```
pozice<-seq(from=180,to=231,by=5)
plot(data[180:231,4],type="l",col="red",lwd=2,axes=FALSE,xlab="Datum",
ylab="Počty pozitivních testů",ylim=c(0,3000),
main="Počty pozitivních PCR testů")
```

```
lines(data[545:593,4],type="l",col="blue",lwd=2)
axis(2,at=seq(from=0,to=3000,by=500),labels=seq(from=0,to=3000,by=500))
lablist<-substr(data[pozice,1],1,5)
axis(3,at=seq(from=0,to=50,by=5),labels=FALSE,tck=0.02,pos=0)
text(x=seq(from=0,to=50,by=5),par("usr")[3] - 0.2, labels = lablist,
srt = 45, pos = 1, xpd = TRUE)
```

Pro snazší orientaci ještě do grafu přidáme legendu pomocí funkce `legend()`, kde první dva parametry udávají pozici kde bude legenda umístěna. Konečný výsledek, jak je ilustrován na obrázku 6 dosáhneme přidáním kódu:

```
legend(2,2800, legend=c("2020", "2021"),col=c("red", "blue"), lty=1,lwd=2)
```



Obr. 7: *Graf počtu celkově provedených testů (vlevo) a počtu pozitivních případů (vpravo).*

Na závěr si ještě ilustrujme, jak lze porovnávat údaje tím, že je zobrazíme na dvou grafech vedle sebe, jak to vidíme na obrázku 4. Nejprve definujeme nové grafické zařízení a jeho rozměry. Vzhledem k tomu, že budeme chtít zobrazit dva grafy vedle sebe, zvolíme větší šířku grafického okna. Zadáme tedy kód:

```
dev.new(width=15, height=8, unit="in")
```

Použitím funkce `par()` a její proměnné `mfrow=c(1,2)` stanovíme, že výsledný obrázek bude mít jeden řádek a v něm dva sloupce, tedy dva grafy vedle sebe. Každé použití funkce `plot()` tedy vytvoří jeden nový graf. Celý kód pak vypadá takto:

```
dev.new(width=15, height=8, unit="in")
par(mfrow=c(1,2))
```

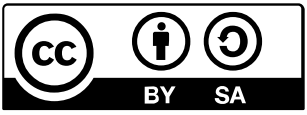
```
pozice<-seq(from=230,to=300,by=5)
plot(data[230:300,3],type="l",col="blue",lwd=2,axes=FALSE,xlab="Datum",
ylab="Počty testů",ylim=c(0,21000),main="Počty testů celkem")
axis(2,at=seq(from=0,to=21000,by=1000),labels=seq(from=0,to=21000,
by=1000))
lablist<-substr(data[pozice,1],1,5)
axis(3,at=seq(from=0,to=70,by=5),labels=FALSE,las=1,pos=0,tck=0.02)
text(x=seq(from=0,to=70,by=5),par("usr")[3] - 0.2, labels = lablist,
srt = 60, pos = 1, xpd = TRUE)
plot(data[230:300,4],type="l",col="red",lwd=2,axes=FALSE,xlab="Datum",
ylab="Počty pozitivních",ylim=c(0,6000),main="Počty pozitivních testů")
axis(2,at=seq(from=0,to=6000,by=1000),labels=seq(from=0,to=6000,
by=1000))
axis(3,at=seq(from=0,to=70,by=5),labels=FALSE,tck=0.02,pos=0)
text(x=seq(from=0,to=70,by=5),par("usr")[3] - 0.2, labels = lablist,
srt = 65, pos = 1, xpd = TRUE)
```

Literatura

1. Crawley, M. J. *Statistics: An Introduction Using R.* Addison-Wesley Publishing company, Boston, 2015.
2. DasGupta, A. *Fundamentals of Probability: A First Course*, New York, Springer-Verlag, 2010.
3. Holický, M. *Aplikace teorie pravděpodobnosti a matematické statistiky*, Praha, ČVUT, 2015.
4. Jakubowski, J., Sztencel, R. *Wstęp do teorii prawdopodobieństwa*, Warszawa, Script, 2010.
5. Kabacoff, R. *R in Action*, New York, Manning Publications, 2015.
6. Nánásiová, O., Kohnová, S.*Štatistika a pravdepodobnosť. Základy matematickej štatistiky a teórie pravdepodobnosti*, Bratislava, STU 2016.
7. Ross, S. M. *A first course in probability*, 10-th edition, Boston, Pearson, 2018.
8. Verzani, J. *Using R for Introductory Statistics.* Second edition, Boca Raton,CRC Press, Taylor & Francis Group, 2014.
9. Wickham, H., Grolemund, G. *R for Data Science*, Sebastopol, United States, O'Reilly Media, Inc, 2017.

Autor

RNDr. Aleš Kozubík, PhD., Katedra matematických metód a operačnej analýzy, Fakulta riadenia a informatiky, Žilinská univerzita v Žiline, Vysokoškolákov 8215/1, 010 26 Žilina, Slovenská republika, e-mail: alesko@frcatel.fri.uniza.sk



Open Access. This article is licensed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License, CC BY-SA 4.0 (`https://creativecommons.org/licenses/by-sa/4.0/`)