INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 35, číslo 2, červen 2024

ZPRÁVA O ČINNOSTI ČESKÉ STATISTICKÉ SPOLEČNOSTI (ČSTS) V ROCE 2023

CZECH STATISTICAL SOCIETY IN 2023

Ondřej Vencálek

E-mail: ondrej.vencalek@upol.cz

1. Členská základna

K 31. 12. 2023 měla ČStS **208** členů. V průběhu roku 2023 přibylo 7 nových členů, 2 členům bylo členství ukončeno, celkový počet členů se zvýšil o 5.

2. Členská schůze ČStS; volby výboru a předsedy společnosti

Členská schůze ČStS se uskutečnila dne 12. dubna 2023 v prostorách ČSÚ v Praze. Zúčastnilo se celkem 26 členů. Byla přečtena a schválena zpráva o činnosti za rok 2022 a zpráva o hospodaření za rok 2022. Členská schůze hlasováním odsouhlasila změnu stanov společnosti v článku 8, části (1), kde složení předsednictva výboru nově může zahrnovat více než jednoho místopředsedu. Původní formulace: „Předsednictvo je výkonným orgánem výboru ČStS. Je tvořeno těmito členy výboru: předsedou, místopředsedou, vědeckým tajemníkem, hospodářem a zástupci sekcí.“ Byla nahrazena formulací „Předsednictvo je výkonným orgánem výboru ČStS. Je tvořeno těmito členy výboru: předsedou, místopředsedy, vědeckým tajemníkem, hospodářem a zástupci sekcí.“

Proběhla volba výboru společnosti. Nově byla do výboru zvolena Zuzana Hübnerová z VUT Brno, která ve výboru nahradila Petra Volfa. Předseda společnosti poděkoval Petru Volfovi za jeho aktivní práci ve výboru. Předsedou společnosti byl na následující dvouleté období zvolen Ondřej Vencálek. Výbor společnosti odhlasoval následující další členy předsednictva: místopředsedy společnosti jsou pro následující dva roky Jaromír Antoch a Ondřej Vozár, vědeckým tajemníkem Martina Litschmannová, hospodářem Tomáš Löster.

V odborné části programu vystoupil doc. RNDr. Zdeněk Karpíšek, CSc., z Ústavu matematiky Fakulty strojního inženýrství VUT v Brně. Téma přednášky bylo fitování diskrétních rozdělení pravděpodobnosti pomocí kvazinorem.

3. Další akce ČStS

- **Statistické dny na břehu Tiché Orlice** se uskutečnily 19. – 21. května 2023 v Penzionu Mítkov na okraji Brandýsa nad Orlicí. Celkem 30 účastníků vyslechlo 15 odborných příspěvků. Tématem tohoto ročníku byly ozvěny epidemie nemoci covid-19.
- **Energy Days 2023** se uskutečnily 10. – 11. listopadu 2023. Semináře, který se konal v prostorách ČSÚ v Praze a ČVUT v Praze, se zúčastnilo celkem 82 účastníků, včetně dvou zahraničních hostů. Bylo předneseno 18 odborných příspěvků. ČStS finančně podpořila účast studentů na tomto semináři.
- **Mikuklášský statistický den** byl pořádán dne 5. prosince 2023 v respiriu budovy MFF v Praze-Kralíně. Zúčastnilo se 21 účastníků a bylo předneseno pět odborných příspěvků.

4. Spolupráce s Českým statistickým úřadem

Český statistický úřad je dlouhodobě partnerem ČStS a podporuje její činnost například zajištěním tisku Informačního bulletinu ČStS, poskytnutím sídla společnosti, prostor pro konání členských schůzí ČStS a publicity v časopisech ČSÚ (např. rozhovory s členy v časopise Statistika a My, informace o společných akcích) a na sociálních sítích.

5. Členství v mezinárodních organizacích

Naše společnost je od roku 2011 členem Federace evropských národních statistických společností FENStatS (The Federation of European National Statistical Societies). Členové ČStS jsou průběžně informováni o dění ve FENStatS. Na členský příspěvek ve výši 200 EUR jsme úspěšně čerpali dotaci prostřednictvím Rady vědeckých společností ČR. Podobně hodláme postupovat i v roce 2024.

6. Členství v Radě vědeckých společností ČR

ČStS je členem Rady vědeckých společností ČR (RVS). Plenární zasedání RVS se konalo 12. dubna 2023 v Praze. ČStS na jednání zastupoval Jakub Fischer. Do RVS byly přijaty další dvě společnosti. Na konci roku byla RVS odevzdána zpráva o činnosti naší společnosti.

7. Další činnost

- Byla vydána 4 čísla Informačního bulletinu ČStS.
- Byla pravidelně aktualizována webová stránka společnosti – bylo zveřejněno celkem 17 příspěvků v sekci novinky a byly provedeny úpravy na dalších stránkách.
- Předseda společnosti reprezentoval ČStS na konferenci České demografické společnosti, která se konala ve dnech 24. – 26. května 2024 v Hradci Králové.

8. Akce chystané v roce 2024

- Spoluorganizace Rakouských statistických dní ve Vídni, 3. – 5. 4. 2024, hlavní organizátor: Rakouská statistická společnost.
- Členská schůze ČStS v Praze, 16. 5. 2024.
- Spoluorganizace 12. ročníku konferencie OSSConf (Otvorený softvér vo vzdelávaní, výskume a v IT riešeniach) v Žilině, 2. – 4. 7. 2024 (`https://ossconf.soit.sk/`), hlavní organizátor: Spoločnosť pre otvorené informačné technológie.
- **Statistické dny 2024 v Perné, 24. – 26. 5. 2024.**
- **Robust 2024 v Bardějově, 8. – 13. 9. 2024.**
- Mikuklášský den ČStS v Praze, kolem 6. 12. 2024.

V Olomouci dne 8. 1. 2024, upraveno 14. 5. 2024.

Ondřej Vencálek
předseda ČStS

VÝBĚR PODLE KVÓT
SAMPLE BY QUOTA

Martin Anděl

Adresa: STEM/MARK, a.s., Smrčkova 2485/4, 180 00 Praha 8

E-mail: andel@stemmark.cz

Abstrakt: Při realizaci marketingových průzkumů se můžeme setkat s požadavkem na výběr jednotek z nějakého seznamu tak, aby tento výběr splňoval předepsané kvóty, které jsou dány jen nekříženě. Tento úkol lze formulovat jako úlohu celočíselného lineárního nebo kvadratického programování a řešit pomocí optimalizačních knihoven v prostředí R.

Klíčová slova: Kvótní výběr, celočíselné programování.

Abstract: In market research it is possible to solve the task about selection of units from the list in such a way, that this sample satisfies given condition on quotas, which are given only marginaly. This task is possible to formulate as the problem of integer linear or quadratic programming and solve with optimisation packages in R.

Keywords: Quota sampling, integer programming.

1. Úvod

Při výběrových šetřeních se realizují různé druhy výběrových plánů. Může to být obyčejný prostý náhodný výběr nebo některý z pokročilejších postupů jako jsou vícestupňový nebo stratifikovaný výběr. Chceme-li vybrat vyvážený výběr (balanced sample) a máme k dispozici celou oporu výběru, mohli bychom použít Tillého metodu kostky (cube method), viz [6], která je naprogramována např. v prostředí R v knihovně `sampling`. Někdy ale může nastat situace, že dostupná opora výběru je jen nějaký částečný seznam, který nemusí ve sledovaných znacích proporčně odpovídat zkoumané populaci. Úlohou je z tohoto seznamu vybrat výběr o velikosti n, který vyhovuje předepsaným kvótám ve sledovaných znacích. Uvažujme situaci, kdy tyto kvóty jsou stanoveny jen nekříženě. Chceme například dodržet předepsaný počet mužů a žen ve znaku pohlaví, předepsaný počet v jednotlivých věkových a vzdělanostních kategoriích. S touto úlohou se můžeme setkat v praxi například při doplňování panelu respondentů. Panelisté jsou v průběhu času postupně obměňováni nebo z vlastního rozhodnutí z panelu odejdou. Potřebujeme za tyto panelisty vybrat náhradu, a to v obdobné struktuře jako je struktura panelistů odešlých.

Představme si, že máme k dispozici $N = 500$ kontaktů a u každého z nich údaj o pohlaví, věkové kategorii a dosaženém vzdělání. Z databáze těchto kontaktů budeme chtít vybrat výběr velikosti $n = 100$ v požadované struktuře (tabulka 1).

Tabulka 1: Databáze a požadovaná struktura výběru

	Počty respondentů	
	Databáze	Požadovaná struktura výběru
Pohlaví		
Muž	156	44
Žena	344	56
Věková kategorie		
15–29 let	188	8
30–44 let	177	38
45–59 let	85	37
60+ let	50	17
Vzdělání		
Základní	32	2
Bez maturity	166	26
Maturita	220	43
VŠ	82	29
Celkem	500	100

2. Když výběr existuje

Označme $\mathbf{X}$ vstupní datovou matici 0–1 indikátorů sledovaných charakteristik v databázi. Matice $\mathbf{X}$ bude mít rozměry $(N \times p)$, kde N je počet kontaktů v databázi a p je sledovaný počet parametrů. V našem případě máme $N = 500$ a $p = 11$. Poslední jedenáctý parametr kontroluje celkový počet vybraných jedinců, v matici $\mathbf{X}$ mu odpovídá poslední sloupec obsahující všude hodnoty 1. Úlohu výběru se stanovenými kvótami můžeme formulovat jako úlohu binár-

ního lineárního programování. Informace o optimalizačních úlohách lze nalézt např. v [5]. Úlohu zapíšeme ve tvaru

$$\begin{aligned} &\min_{\mathbf{w}} \mathbf{1}^T\mathbf{w}, \\ &\mathbf{X}^T\mathbf{w} = \mathbf{s}, \end{aligned} \tag{1}$$

kde $\mathbf{w}$ je hledaný vektor binárních 0–1 proměnných indikujících zahrnutí daného kontaktu do výběru a vektor $\mathbf{s}$ je vektor s požadovanou strukturou výběru vyjádřenou absolutními počty kontaktů. Předepsané podmínky zahrnují v poslední rovnici požadavek na velikost výběru, součet složek vektoru $\mathbf{w}$ je tedy pevně dán a úlohu bychom mohli přeformulovat na maximalizaci stejné účelové funkce za shodných vedlejších podmínek.

Vstupní data úlohy si připravíme v prostředí R.

```
make.X <- function(){
  Xpocet <- matrix(
    c(1,1,1, 6,  1,1,2, 6,  1,1,3,26,  1,1,4,10,  1,2,1, 2,
      1,2,2,25,  1,2,3,17,  1,2,4, 9,  1,3,2,12,  1,3,3,16,
      1,3,4, 3,  1,4,1, 2,  1,4,2, 1,  1,4,3,12,  1,4,4, 9,
      2,1,1,16,  2,1,2,46,  2,1,3,67,  2,1,4,11,  2,2,1, 4,
      2,2,2,44,  2,2,3,50,  2,2,4,26,  2,3,2,20,  2,3,3,25,
      2,3,4, 9,  2,4,1, 2,  2,4,2,12,  2,4,3, 7,  2,4,4, 5),
    ncol=4, byrow=T)
  Xdf <- as.data.frame(Xpocet[rep(1:nrow(Xpocet),Xpocet[,4]),
                              1:3])
  Xdffac <- as.data.frame(lapply(Xdf,as.factor))
  names(Xdffac)<-c('sex','agecat','educat')

X <- with(Xdffac,{
  cbind(
    model.matrix(~sex-1),
    model.matrix(~agecat-1),
    model.matrix(~educat-1),
    total = 1)
 })
}

X <- make.X()
N <- dim(X)[1]
p <- dim(X)[2]
s <- c(44,56,8,38,37,17,2,26,43,29,100)
```

Funkce `make.X` nejprve definuje matici `Xpocet`, která obsahuje v prvních třech sloupcích postupně číselné hodnoty kategorií znaků pohlaví, věková skupina a vzdělání, přičemž např. u znaku pohlaví kód 1 odpovídá kategorii muž, hodnota 2 kategorii žena. Poslední čtvrtý sloupec matice `Xpocet` pak udává četnost výskytu dané kombinace hodnot znaků. Z této agregované podoby funkce vytvoří neagregovaná data a současně kódování jednotlivých znaků převede na 0–1 tvar pomocí příkazu `model.matrix`. Současně vytvoří sloupec samých jedniček, který odpovídá poslednímu parametru pro celkový počet vybraných jedinců. Na posledním řádku kódu jsou do vektoru `s` uloženy požadované počty jednotlivých kategorií ve výběru, přičemž poslední hodnota 100 udává celkový požadovaný počet jedinců ve výběru.

Úlohu můžeme řešit pomocí knihovny `Rsymphony` (viz [1]).

```
library('Rsymphony')
result <- Rsymphony_solve_LP(obj=rep(1,N),mat=t(X),
          dir=rep('==',p),rhs=s, types=rep('B',N), max=F)
```

Hledané `w` je uloženo v objektu `result$solution`. Nemusí jít o jediné možné řešení. Pokud bychom chtěli najít více řešení, mohli bychom použít například knihovnu `Rcplex` (viz [2]), která v jednom ze svých parametrů umožňuje nastavit počet hledaných řešení. Předpokladem ovšem je mít na počítači nainstalovaný optimalizační program CPLEX. Další možnost jak nalézt více řešení bude zmíněna v kapitole 6. Při hledání dalšího řešení různého od již nalezeného `w` bychom mohli postupovat také tak, že bychom přidali do úlohy další podmínku, která by zajistila různost řešení. Druhé řešení bychom dostali příkazem

```
result2 <- Rsymphony_solve_LP(obj=rep(1,N),mat=rbind(t(X),w),
           dir=c(rep('==',p),'<='),rhs=c(s,99),
           types=rep('B',N), max=F)
```

3. Když výběr neexistuje

Pokud bychom chtěli z databáze kontaktů uvedené v kapitole 1 udělat výběr o velikosti $n = 250$ s obdobnou podílovou strukturou v jednotlivých kategoriích, zjistili bychom, že řešení neexistuje, protože ve věkové skupině 45–59 let máme v databázi 85 kontaktů, ale ve výběru bychom chtěli nyní mít 92 kontaktů, viz tabulka 2, kde je pro úsporu místa již uveden i výběr, ke kterému dospějeme na konci kapitoly.

Musíme úlohu přeformulovat a rovnosti v podmínkách uvolnit na nerovnosti. Chtěli bychom najít takové nerovnosti, které se od rovností liší co

Tabulka 2: Výběr blízký požadované struktuře

	Počty respondentů		
	Výběr	Požadovaná struktura výběru	Databáze
Pohlaví			
Muž	111	111	156
Žena	139	139	344
Věková kategorie			
15–29 let	21	21	188
30–44 let	102	95	177
45–59 let	85	92	85
60+ let	42	42	50
Vzdělání			
Základní	5	5	32
Bez maturity	66	66	166
Maturita	107	107	220
VŠ	72	72	82
Celkem	250	250	500

nejméně. Úloha smíšeného binárního a celočíselného lineárního programování bude nyní vypadat takto:

$$\begin{gathered} \min_{(\mathbf{w},\mathbf{d})} \mathbf{1}^T(\mathbf{w},\mathbf{d}), \\ \mathbf{s}-\mathbf{d} \le \mathbf{X}^T\mathbf{w} \le \mathbf{s}+\mathbf{d}, \\ 0 \le d_i \le s_i, \quad i = 1 \dots p-1 \\ 0 = d_p \end{gathered} \tag{2}$$

kde $\mathbf{w}$ je vektor binárních 0–1 proměnných indikujících zahrnutí daného kontaktu do výběru a $\mathbf{d}$ je celočíselný vektor o kolik se může řešení lišit od požadované struktury. Poslední podmínka $0 = d_p$ říká, že požadovaný počet jedinců ve výběru $n = 250$ má být splněn přesně. Podmínky na vektor od-

chylek $\mathbf{d}$ můžeme případně upravit, mohli bychom například některé z nich chtít splnit přesně. Alternativou může být požadavek na splnění odchylky vyjádřené podílem z vektoru $\mathbf{s}$. Pokud by vektor $\mathbf{s}$ obsahoval v některé složce malé číslo, můžeme naopak horní mez zvýšit. Úlohu můžeme opět řešit pomocí R a knihovny `Rsymphony`.

```
s <- c(111,139,21,95,92,42,5,66,107,72,250)
obj <- c(rep(1,N),rep(1,p))
A <- matrix(0, nrow=2*p, ncol=p+N)
A[1:p,1:N] <- t(X)
A[1:p,(N+1):(N+p)] <- -diag(rep(1,p))
A[(p+1):(2*p),1:N] <- -t(X)
A[(p+1):(2*p),(N+1):(N+p)] <- -diag(rep(1,p))
op <- c(rep('<=',2*p))
rhs <- c(s,-s)
bounds <- list(
  lower = list(ind=c((N+1):(N+p)), val=c(rep(0,p))),
  upper = list(ind=c((N+1):(N+p)), val=c(s[1:p-1],0))
)
result <- Rsymphony_solve_LP(obj, A, op, rhs,
          types = c(rep('B',N),rep('I',p)),
          max = F, bounds=bounds)
w <- result$solution[1:N]
tabvysledek <- cbind(t(X)%*%w, s)
colnames(tabvysledek) <- c('vyber','s')
tabvysledek
```

Matice `A` v kódu slouží k zápisu nerovností

$$\mathbf{s} - \mathbf{d} \leq \mathbf{X}^T\mathbf{w} \leq \mathbf{s} + \mathbf{d}$$

ve funkci `Rsymphony_solve_LP`. Horní část matice `A` je pro horní nerovnosti $\leq$ a je tvořena dvojicí matic, v levé části je transponovaná matice `X`, což jsou koeficienty pro hledaný vektor indikující vybrané jednotlivce `w`, v pravé části je pak se záporným znaménkem jednotková diagonální matice dimenze `p` pro minimalizovaný vektor odchylek $\mathbf{d}$. Obdobně je pak ještě v kódu sestavena dolní polovina matice `A` pro dolní nerovnosti.

Výsledný výběr je uveden v tabulce 2. Vzhledem k nedostatečnému počtu kontaktů v databázi u věkové kategorie 45–59 let musíme nějak ve výběru navýšit ostatní věkové kategorie. V našem případě došlo k navýšení pouze u jedné z věkových kategorií, konkrétně u kategorie 30–44 let. U znaků pohlaví

a vzdělání se požadovanou strukturu výběru podařilo splnit zcela přesně. Nalezené řešení není jediné možné, k hledání více řešení bychom mohli opět použít např. program CPLEX nebo postup, který bude uveden na konci kapitoly 6. Také bychom mohli postupně přidávat další podmínky na různost řešení jak bylo naznačeno v předchozí kapitole, nyní ještě s ověřením, zda další řešení dosahuje stejného minima jako již nalezené první řešení.

4. Kvadratická minimalizace

I v této kapitole budeme vycházet ze stejné databáze kontaktů jako v předchozích úlohách. Navýšení sousední věkové kategorie v tabulce 2 by nám mohlo vyhovovat, můžeme ale také preferovat rozložení do více kategorií. Toho dosáhneme minimalizací součtu čtverců odchylek $\sum_{i=1}^{p} d_i^2$. Pak již ale nevystačíme s minimalizací lineární funkce, ale půjde o složitější úlohu smíšeného binárního a celočíselného kvadratického programování. Takovou úlohu můžeme řešit pomocí optimalizačních programů jako např. CPLEX nebo Gurobi. Tyto programy jsou komerční, existuje ale možnost zkušební nebo akademické licence.

Úlohu zformulujeme v následujícím tvaru.

$$\begin{gathered} \min_{(\mathbf{w},\mathbf{d})} (\mathbf{w},\mathbf{d})^T \mathbf{I}_{(N+p,N+p)} (\mathbf{w},\mathbf{d}), \\ \mathbf{s} - \mathbf{d} \le \mathbf{X}^T \mathbf{w} \le \mathbf{s} + \mathbf{d}, \\ 0 \le d_i \le s_i, \quad i = 1 \dots p-1, \\ 0 = d_p \end{gathered} \tag{3}$$

kde $\mathbf{w}$ jsou binární 0–1 proměnné a $\mathbf{d}$ jsou celočíselné proměnné, matice $\mathbf{I}_{(N+p,N+p)}$ je čtvercová matice s 1 na diagonále a 0 mimo diagonálu. Podmínku na velikost výběru chceme splnit přesně. Podmínky na vektor odchylek $\mathbf{d}$ můžeme případně upravit, jak již bylo zmíněno u úlohy 2. Řešení budeme hledat pomocí programu CPLEX, pro který existuje knihovna `Rcplex` (viz [2]).

```
library('Rcplex')
s <- c(111,139,21,95,92,42,5,66,107,72,250)
A <- matrix(0,nrow=2*p,ncol=p+N)
A[1:p,1:N] <- t(X)
A[1:p,(N+1):(N+p)] <- -diag(rep(1,p))
A[(p+1):(2*p),1:N] <- -t(X)
A[(p+1):(2*p),(N+1):(N+p)] <- -diag(rep(1,p))
rhs <- c(s,-s)
```

```
result <- Rcplex(cvec = c(rep(0,N+p)), Amat = A, bvec = rhs,
 Qmat = diag(c(rep(1,N),rep(1,p))),lb = 0,
 ub = c(rep(1,N),s[1:(p-1)],0),
 objsense = c('min'), sense = 'L',
 vtype = c(rep('B',N),rep('I',p)))
w <- result$xopt[1:N]
tabvysledek <- cbind(t(X)%*%w,s)
colnames(tabvysledek) <- c('vyber','struktura')
tabvysledek
```

Význam a sestavení matice `A` je zde obdobné jako v kódu pro úlohu 2. Řešení pomocí programu Gurobi (viz [3]) by se lišilo voláním optimalizační funkce a načtením knihovny `gurobi`.

```
library('gurobi')
...
model <- list(Q = diag(c(rep(1,N), rep(1,p))),
            obj = 0, A = A, rhs = rhs,
            sense = c(rep('<=',2*p)),
            lb = c(rep(0,N+p)), ub = c(rep(1,N),s[1:(p-1)],0),
            vtype = c(rep("B",N),rep("I",p)) )
params <- list(OutputFlag = 0)
result <- gurobi(model,params)
w <- result$x[1:N]
...
```

Dosažené optimální struktury výběru jsou uvedeny v tabulce 3. Řádek Celkem je proměnná jako ostatní řádky Muž, Žena, ..., nejde o typický součtový řádek, který by byl kontrolním součtem čísel uvedených výše. Konkrétně v naší ukázce se podařilo vybrat přesně 250 respondentů, odchylka je 0.

Nalezená řešení `w` nejsou jediná možná. Například funkci `Rcplex` můžeme v parametru `n` říci, kolik řešení chceme hledat. Pokud parametr nenastavíme, program hledá jedno řešení. Při nastavení `n = NA` program hledá všechna možná řešení, musíme však počítat s velkou časovou náročností.

Dále bychom mohli místo součtu čtverců odchylek minimalizovat veličinu $\chi^2 = \sum_{i=1}^{p} d_i^2/s_i$. V kódu bychom do parametru `Qmat`, resp. `Q`, zadali matici `diag(c(rep(1,N),1/s))`. Výsledek by se lišil jen u věkových kategorií (viz tabulka 4). I v tomto případě existuje více optimálních řešení hledaného vektoru `w`, tato řešení však dají shodné vybrané počty u všech kategorií.

Tabulka 3: Optimální výběr

	Počty respondentů					
	CPLEX		Gurobi		Požad. strukt. výběru	Data-báze
	Výběr	Čt. odch.	Výběr	Čt. odch.		
Pohlaví						
Muž	111	0	111	0	111	156
Žena	139	0	139	0	139	344
Věková kategorie						
15–29 let	24	9	23	4	21	188
30–44 let	97	4	98	9	95	177
45–59 let	85	49	85	49	92	85
60+ let	44	4	44	4	42	50
Vzdělání						
Základní	5	0	5	0	5	32
Bez maturity	66	0	66	0	66	166
Maturita	107	0	107	0	107	220
VŠ	72	0	72	0	72	82
Celkem	250	0	250	0	250	500

5. Největší možný výběr

Vidíme, že při $n = 250$ můžeme zkonstruovat jen výběr, který je požadované struktuře v nějakém smyslu nejbližší. Mohlo by nás ještě zajímat, jaký by byl největší možný výběr splňující danou strukturu s předepsanou přesností. Vyjdeme z předepsané struktury pro $n = 250$, pro ostatní n strukturu proporčně přepočteme, nebudou to tedy již pouze celá čísla. Z tohoto důvodu nebudeme požadovat nulovou odchylku, ale zkusíme nastavit pro každou z kategorií maximální možnou odchylku 0,9.

Tabulka 4: Výběr s minimálním χ^2

	Počty respondentů	
	Výběr min χ^2	Požadovaná struktura výběru
Věková kategorie		
15–29 let	22	21
30–44 let	99	95
45–59 let	85	92
60+ let	44	42

Optimalizační úloha bude ve tvaru.

$$\begin{aligned} &\max_{\mathbf{w}} \mathbf{1}^T\mathbf{w}, \\ -\mathbf{b} \leq &(\mathbf{X}^T - \mathrm{diag}(\mathbf{s}/\mathbf{250})\mathbf{1}_{(p,\ N)})\mathbf{w} \leq \mathbf{b}, \end{aligned} \tag{4}$$

kde $\mathbf{w}$ je vektor binárních 0–1 proměnných indikujících zahrnutí daného kontaktu do výběru a $\mathbf{b}$ je předepsaná odchylka o kolik se může řešení lišit od požadované struktury.

Řešení budeme hledat pomocí knihovny `Rsymphony`.

```
zaklad <- s[11]
r <- s/zaklad
obj <- c(rep(1,N))
A <- matrix(0, nrow=2*p, ncol=N)
J <- matrix(1, nrow=p, ncol=N)
dr <- diag(r) %*% J
A[1:p,1:N] <- t(X) - dr
A[(p+1):(2*p),1:N] <- t(X) - dr
op <- c(rep('>=',p),rep('<=',p))
b <- c(rep(0.9,p))
rhs <- c(rep(0,p)-b,rep(0,p)+b)
result <- Rsymphony_solve_LP(obj, A, op, rhs,
                             types = c(rep('B',N)),
                             max = T)
w <- result$solution
tabvysledek <- as.data.frame(cbind(c(t(X)%*%w,sum(w)),
```

```
  c(round(s/zaklad*sum(w),1),sum(w))))
colnames(tabvysledek) <- c('vyber','zadani')
tabvysledek
```

Výběr je uveden v tabulce 5. Našemu zadání tedy vyhovuje výběr o velikosti $n = 233$.

Tabulka 5: Maximální možný výběr

	Počty respondentů	
	Výběr	Zadání
Pohlaví		
Muž	103	103,5
Žena	130	129,5
Věková kategorie		
15–29 let	19	19,6
30–44 let	89	88,5
45–59 let	85	85,7
60+ let	40	39,1
Vzdělání		
Základní	4	4,7
Bez maturity	61	61,5
Maturita	100	99,7
VŠ	68	67,1
Celkem	233	233,0

6. Výběr s použitím knihovny ROI a jejích doplňků

Pokud nemáme k dispozici na počítači optimalizační programy, můžeme využít k řešení úlohy NEOS server. Přístup k NEOS serveru je zdarma, ale je nutné se zaregistrovat a poskytnout emailovou adresu. Postup si ukážeme na úloze z kapitoly 3 Využijeme k tomu obecné rozhraní `ROI` (viz [4]). Níže je pouze upravený kus kódu, do parametru `email` je pak nutno napsat vlastní registrovanou emailovou adresu.

```
library(ROI)
library(ROI.plugin.neos)
...
lp_bound <- V_bound(li=c((N+1):(N+p)),ui=c((N+1):(N+p)),
 lb=c(rep(0,p)), ub=round(c(s[1:p-1],0),0))
types <- c(rep('B',N),rep('I',p))
lp <- OP(objective = obj,
         L_constraint( L = A,
                       dir = op,
                       rhs = round(rhs,0)),
         bounds=lp_bound,
         types=types,
         maximum = FALSE)
sol <- ROI_solve(lp, solver = "neos", method = "cplex",
                 email = "emailova adresa")
w <- sol$solution[1:N]
...
```

V příkazu `ROI_solve` můžeme nastavit i jiné metody. Jedna z možností by mohla být nastavení `method="mosek"`.

Když nyní máme optimalizační úlohu zapsánu ve tvaru pro `ROI`, mohli bychom ještě zkusit hledat více řešení pomocí doplňku `ROI.plugin.msbinlp`. List s více řešeními bychom po načtení knihoven dostali příkazem

```
library(ROI.plugin.msbinlp)
library(ROI.plugin.symphony)
sol <- ROI_solve(lp, solver = "msbinlp", method = "symphony",
                 nsol_max = 4)
```

Zde jsme do parametru `nsol_max` uvedli, že chceme hledat 4 řešení. K jednotlivým hledaným vektorům `w` pak přistupujeme pomocí indexů, např. první řešení bychom dostali příkazem

```
w1 <- sol[[1]]$solution[1:N]
```

Můžeme použít i jiné optimalizační metody, např. `glpk`.

```
library(ROI.plugin.glpk)
sol <- ROI_solve(lp, solver = "msbinlp", method = "glpk",
                 nsol_max = 4)
```

Nastavení `nsol_max=Inf` hledá všechna řešení.

7. Závěr

Úlohu vytvoření výběru splňujícího předepsané marginální kvóty můžeme zformulovat jako úlohu binárního či smíšeného binárního a celočíselného lineárního nebo kvadratického programování. K řešení těchto úloh jsme zvolili prostředí R a jeho knihovny `Rsymphony`, `RCplex` a `gurobi`. Pro soubory o velikosti řádu stovek lze řešení nalézt pomocí volně dostupných knihoven, u větších souborů v řádu tisíců může být nutné k řešení zvolit některý z komerčních optimalizačních programů.

Základní zdrojový kód na vyzkoušení najde čtenář na Githubu:

`https://github.com/MartakAnd/Sample-by-quota`

Poděkování

Autor děkuje prof. RNDr. Jaromíru Antochovi, CSc., a dvěma anonymním recenzentům za cenné poznámky a návrhy, které text doplnily a učinily ho srozumitelnějším.

Literatura

[1] An R interface to the SYMPHONY solver for mixed-integer linear programs. `https://cran.r-project.org/web/packages/Rsymphony/`. *cit. 9*

[2] R Interface to CPLEX. `https://cran.r-project.org/web/packages/Rcplex/`. *cit. 9, 12*

[3] Gurobi optimisation. `https://www.gurobi.com/`. *cit. 13*

[4] ROI: R Optimization Infrastructure. `https://cran.r-project.org/web/packages/ROI/`. *cit. 16*

[5] Nemhauser, George L., Wolsey, Laurence A.: *Integer and combinatorial optimization.* Wiley-Interscience series in discrete mathematics and optimization. New York, N.Y.: John Wiley, 1999. ISBN 0-471-35943-2. *cit. 8*

[6] Tillé, Y.: *Sampling Algorithms,* Springer, 2006. *cit. 6*

PROBASTAT 2024
8TH INTERNATIONAL CONFERENCE ON PROBABILITY AND STATISTICS

Jaromír Antoch

E-mail: antoch@karlin.mff.cuni.cz

Po delší odmlce způsobené covidovou pandemií se ve dnech 20. – 24. května 2024 ve Smolenicích konala již patnáctá konference Probastat, tentokrát ročník 2024. Připomeňme, že série těchto úspěšných konferencí začala již v roce 1974 z iniciativy profesorů Andreje Pázmana a Silvie Pulmanové, pracovníků Slovenské akademie věd, takže jsme oslavili padesáté výročí od jejich zahájení. Profesor Pázman i tentokrát konferenci slavnostně otevřel a přiblížil účastníkům jeho historii.

Konference se zúčastnilo více než šedesát účastníků jak ze Slovenska a Česka, tak z celé Evropy. Klíčové přednášky byly věnovány následujícím tématům:

- Optimální plánování experimentů (W. Muller).
- Penalizovanému odhadování a jeho geometrii (U. Schneider).
- Dovýběrovému testování ve „zpětnovazebním učení“ (O. Okhrin).
- Detekci změn ve statistických modelech (M. Hušková).
- Aditivní regresi (G. Van Bever).
- Neklasickému zobecnění komplexních čísel s aplikacemi v pravděpodobnosti (W. – D. Richter).
- Historii pravděpodobnosti (Ch. Genest).

Uskutečnila se také soutěž studentů a doktorandů, které se zůčastnilo 11 účastníků, z nichž byli oceněni nejlepší tři:

- Matej Benko: Characteristic function and moment generating function of multivariate folded normal distribution.
- Iryna Zabaikina: Feedback on dilution in stochastic gene expression: a comparative study of single-cell and population frameworks, a,
- Pál Somogyi: A randomized exchange algorithm for optimal experimental design problems with general elementary information matrices.

Atmosféra byla tak jako vždy více než přátelská, a tak se lze jenom těšit na další Probastat a na Smolenice. Detaily a další informace lze nalézt na adrese https://www.um.sav.sk/probastat2024/.


CAIUS IOSEPHUS PUMPEY

POZVÁNKA NA IMPS 2024
INVITATION TO THE IMPS 2024 CONFERENCE

Patrícia Martinková

E-mail: martinkova@cs.cas.cz

Vážení kolegové,

přijměte pozvání na výroční konferenci Psychometric Society (IMPS 2024), která se uskuteční ve dnech 16. – 19. července v Praze. Abstrakty příspěvků (individuální přednášky, postery, symposia) lze podávat do konce února. Více informací naleznete na stránkách https://www.psychometricsociety.org/imps-2024.

Prosím přepošlete tuto informaci také svým kolegům, pro které by mohla být relevantní.

Se srdečným pozdravem

Patrícia Martinková
Local Host Committee Chair

Dear Colleagues,

Let me invite you to the annual International Meeting of the Psychometric Society (IMPS 2024) which will take place from 16–19 July in Prague. Abstracts of papers (oral presentations, poster presentations, symposia) can be submitted until the end of February. For more information, please visit https://www.psychometricsociety.org/imps-2024.

Please forward this information to your colleagues for whom it may be relevant.

Sincerely,

Patrícia Martinková
Local Host Committee Chair

POZVÁNKA NA TUG 2024
INVITATION TO THE TUG 2024 CONFERENCE

Michal Hoftich

E-mail: michal.h21@gmail.com

Vážené TEXistky, vážení TEXisté,

srdečně vás zveme na konferenci TUG 2024, která se uskuteční v Praze (Hotel Grandior) ve dnech 19. – 21. července 2024. Titul zní *Presentations covering the TEX world*, podtitul pak *Typography & typesetting, fonts & design, publishing and more.*

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 sázen v programu TEX, ve formátu LuaHBTEX s písmy balíku CSfonts.

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

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.