Informační Bulletin

České Statistické Společnosti

č.3. říjen 2001, ročník 12

Statistické dny v Hradci Králové (druhá část)

Ve dnech 21. a 22. června 2001 přibyla další kapitola v historii Statistických dní České statistické společnosti. Dvoudenní setkání statistiků tentokrát hostil Hradec Králové a připravil jej prof. RNDr. PhDr. Zdeněk Půlpán, CSc. s podporou Katedry matematiky Pedagogické fakulty Univerzity Hradec Králové a Jednoty českých matematiků a fyziků, pobočky pro kraj Hradec Králové.

Česká statistická společnost sdružuje statistiky velice různorodého zaměření, a proto není překvapivé, že i v Hradci Králové se sešli odborníci z různých institucí od vysokých škol přes specializované vědecké ústavy až k Českému statistickému úřadu a že škála diskutovaných témat sahala od ryze teoretických až k vyloženě aplikačním. Účastníci tak např. vyslechli přednášky o Grayově kódu, o matematických metodách kolektivního rozhodování, o speciální regresní úloze, o problematice dotazníkových šetření, o analýze hydrologických dat, o kontingenčních tabulkách, o úmrtnostních tabulkách i o indexech spotřebitelských cen. Program prvního dne se protáhl až do pozdního večera, ale vzhledem k tomu, že byl nejdelší den v roce, zbyla ještě chvíle na příjemnou podvečerní procházku městem. Dopolední program druhého dne setkání uzavřel a ing. Z Roth jménem výboru České statistické společnosti poděkoval prof. Půlpánovi za pečlivou přípravu pěkné akce.

Pro ty, kteří se nemohli zúčastnit tohoto setkání, otiskujeme ve dvou letošních číslech našeho bulletinu třináct příspěvků, z nichž prvních pět bylo obsahem minulého čísla, další najdete zde.

Marek Malý

ÚMRTNOSTNÍ TABULKY A JEJICH PREZENTACE

Zuzana Škrabková

Ústav lékařské biofyziky, Lékařská fakulta UK, Šimkova 870, 500 01 Hradec Králové

Abstrakt:

Článek upozorňuje na odlišnosti ve způsobech prezentace a různé výpočty některých charakteristik v úmrtnostních tabulkách.

1. Úmrtnostní tabulka

Základem *úmrtnostní tabulky* je posloupnost čísel

$$l_0, l_1, l_2, \ldots, l_x, \ldots, l_\omega,$$

kde l_x je počet osob ve věku **x**, které zůstaly naživu ze souboru l_0 současně narozených jedinců. Dále úmrtnostní tabulka většinou obsahuje řadu dalších, systematicky uspořádaných údajů. Symbol **l** je odvozený z anglického living (žijící). V anglicky psané literatuře se používá optimističtější pojem – *life table*.

2. Základní dělení

V praxi se setkáváme se dvěma základními typy úmrtnostních tabulek

A **běžné (průřezové) tabulky**

- vycházejí z dané populace během krátkého (max. 10 roků) časového období
- hypotetický, uměle zkonstruovaný snímek života zvoleného počtu jedinců tak, jak by vypadal, kdyby se udrželo úmrtnostní chování populace z uvažovaného časového období
- používají se v demografii, pojišťovnictví

A1 - **úplné** - věkové intervaly o délce 1 rok

A2 - **zkrácené** - víceleté věkové skupiny (0-1, 1-5, 5-10 let, ...)

B **generační (kohortní) tabulky**

- opravdový záznam o průběhu života konkrétní populace současně narozených jedinců (např. skupina pacientů podstupujících současně určitou terapii nebo zkoumaný kmen bakterií)
- používají se např. v medicínském výzkumu

3. Ukázka běžné zkrácené úmrtnostní tabulky

1998-99 Východočeský kraj **Zkrácené úmrtnostní tabulky**

muži							
věk	qx	px	lx	dx	Lx	Tx	ex
0	0.004883	0.995117	100000	488	99551	7189720	**71.90**
1	0.001639	0.998361	99512	163	99430	7090169	**71.25**
5	0.001076	0.998924	99349	107	99295	6692449	**67.36**
10	0.001291	0.998709	99242	128	99178	6195973	**62.43**
15	0.004183	0.995817	99113	415	98906	5700086	**57.51**
20	0.005028	0.994972	98699	496	98451	5205555	**52.74**
25	0.005028	0.994972	98203	494	97956	4713301	**48.00**
30	0.006240	0.993760	97709	610	97404	4223523	**43.23**
35	0.008051	0.991949	97099	782	96708	3736503	**38.48**
40	0.014893	0.985107	96317	1434	95600	3252963	**33.77**
45	0.024179	0.975821	94883	2294	93736	2774962	**29.25**
50	0.042417	0.957583	92589	3927	90625	2306283	**24.91**
55	0.065173	0.934827	88661	5778	85772	1853158	**20.90**
60	0.098809	0.901191	82883	8190	78788	1424297	**17.18**
65	0.149268	0.850732	74693	11149	69119	1030355	**13.79**
70	0.226045	0.773955	63544	14364	56362	684761	**10.78**
75	0.337343	0.662657	49180	16591	40885	402950	**8.19**
80	0.486419	0.513581	32590	15852	24664	198525	**6.09**
85	0.661990	0.338010	16737	11080	11197	75208	**4.49**
90	0.829899	0.170101	5657	4695	3310	19220	**3.40**
95	0.944896	0.055104	962	909	508	2671	**2.78**
100	0.991338	0.008662	53	53	27	133	**2.50**

l_x - *počet dožívajících se věku x* - je to počet jedinců z l_0, kteří se dožijí věku x let (l_0 - kořen tabulky; používá se $l_0 = 100\,000$)

d_x - *počet zemřelých ve věku x* - počet jedinců zemřelých ve věku x;

$$d_x = l_x - l_{x+1}$$

q_x - *pravděpodobnost úmrtí ve věku x* - pravděpodobnost, že jedinec, který je naživu ve věku x let, zemře před dosažením věku x+1 :

$$q_x = d_x / l_x$$

p_x - *pravděpodobnost dožití* - je doplňkem pravděpodobnosti úmrtí a vyjadřuje pravděpodobnost, že jedinec ve věku x let v daném období nezemře a dožije se věku x+1:

$$p_x = 1 - q_x$$

$$p_x = l_{x+1} / l_x$$

L_x - *tabulkový počet žijících* - je průměrný počet žijících ve věku x let; počítá se (kromě věku 0, kde se zahrnuje zvýšená kojenecká úmrtnost) jako průměr ze dvou po sobě jdoucích tabulkových počtů dožívajících:

$$L_x = 1/2\ (l_x + l_{x+1})$$

$$L_0 = l_0\ (1 - 0{,}92\ q_0)$$

T_x – *počet zbývajících let života jedinců ve věku x* – je to pomocný ukazatel, vyjadřující počet let života, které má tabulková generace (nikoliv jednotlivec) v daném věku ještě před sebou (tj. celkový počet „člověkoroků", které do konce života prožije l_x osob:

$$T_x = L_x + L_{x+1} + \ldots + L_\omega$$

neboli

$$T_x = T_{x+1} + L_{x+1}$$

e_x - *střední délka života neboli naděje dožití* - udává počet let, který má naději prožít osoba právě x-letá:

$$e_x = T_x / l_x$$

4. Cenzorovaná pozorování

V generačních úmrtnostních tabulkách se kromě výše zmíněných pojmů vyskytuje pojem cenzorovaných pozorování. Jsou to taková pozorování, která vznikají v důsledku časového omezení studie nebo z důvodu ztráty údajů během trvání studie. Zajímá nás například délka trvání manželství nebo „přežívání" zubního implantátu. Studie trvá např. 5 roků a jejím výsledkem je soupis dob přežívání jednotlivých pozorování. Tato jsou buď neúspěšná (jejichž „život" skončil před ukončením studie) nebo úspěšná (trvající, přežívající, cenzorovaná). Jak taková úspěšná pozorování zahrnout do úmrtnostní tabulky? Jak zahrnout informaci, že „život" cenzorovaných pozorování není ukončen jejich selháním, ale přerušením sledování?

Jistě je rozumný předpoklad, že cenzorovaná pozorování mají stejnou pravděpodobnost selhání jako všechna ostatní pozorování. Jak ale započítat jejich vliv? Podívejme se na následující způsoby řešení.

5. Přístup k problému v programu STATISTICA

STATISTICA: Survival Analysis

Life Table & Survival Time Distribution Results

```
Total number of valid observations:    361
                        uncensored:     29 (  8.03%)
                          censored:    332 ( 91.97%)
```

Interval x	Interval Start	Mid Point	Interval Width	Number Entering n_x	Number Withdrwn w_x	Number Exposed n'_x
1	0	0.5	1	361	0	361
2	1	1.5	1	355	0	355
3	2	2.5	1	354	0	354
4	3	3.5	1	351	0	351
5	4	4.5	1	351	180	261
6	5	---	1	156	152	80

Interval	Number Dying d_x	Proportn Dead q_x	Proportn Surviving p_x	Cum. Prop Surviving l_x	Problty Density	Hazard Rate
1	6	.016620	.983379	1.0	.016620	.016760
2	1	.002817	.997183	.983379	.002770	.002821
3	3	.008475	.991525	.980609	.008310	.008511
4	0	0	1.0	.972299	0	0
5	15	.057471	.942529	.972299	.055879	.059172
6	4	.05	0.95	.916420	---	---

$n'_x = n_x - 1/2\, w_x$

$q_x = d_x / n'_x$

$l_x = l_0 p_0 p_1 \ldots p_{x-1} = l_{x-1} p_{x-1}$

Všimněme si nejprve odlišného značení sloupců proti úmrtnostní tabulce uvedené v bodě 3. Symbolem n_x je tady označen počet jedinců vstupujících do intervalu ve věku x, zatímco symbol l_x je použit pro kumulativní pravděpodobnost přežití.

Počty cenzorovaných pozorování w_x jsou ve sloupci *Number Withdrwn.* V tomto případě se jedná o pozorování, které vstoupily do studie o něco později než ostatní a nebylo možné je sledovat celých 5 roků.

Všimněme si, že je zde oproti běžné úmrtnostní tabulce navíc ještě sloupec n'_x (modifikovaný počet přežívajících), kde jsou zpracovány údaje o cenzorovaných pozorováních. Autor vytvořil vztah pro n'_x na základě **předpokladu, že jen polovina z těchto pozorování by přežila do konce sledovaného intervalu.**

Potom se pravděpodobnost úmrtí ve věku x a pravděpodobnost dožití (tj. q_x a p_x) počítají pomocí modifikovaného počtu přežívajících n'_x.

6. Přístup k problému v knize Riffenburgh (1999)

Interval	begin	died	lost	end	S (survival rate)
0 (outset)	361	0	0	361	1.0
>0-1	361	6	0	355	0.983379
>1-2	355	1	0	354	0.997183
>2-3	354	3	0	351	0.991525
>3-4	351	0	0	351	1.0
>4-5	351	15	0	336	0.957265
	336	0	180	156	
>5-6	156	4	0	152	0.974359

Tato tabulka obsahuje pro změnu cenzorovaná pozorování ve sloupci *lost* a sloupec *S* obsahuje pravděpodobnosti dožití (tedy p_x z předešlých tabulek).

Například *S* pro 1. interval se počítá jako 1.0 x (355/361). Když se vyskytnou cenzorovaná pozorování (jako zde v 5. intervalu), tak se tato informace do výpočtu pravděpodobnosti dožití konce daného intervalu vůbec neuvažuje. **Předpokládá se totiž, že daní jedinci přežívají až do konce intervalu**, ale do dalšího intervalu se již nezapočítávají.

Porovnáním vyznačených hodnot s předchozí tabulkou vidíme, že při tomto přístupu vycházejí pravděpodobnosti dožití o trochu vyšší, tedy výhodnější pro prezentaci.

7. Ukázky dalších přístupů

Table 13a. Cumulative success rates of 8 mm long implants

Interval (years)	Implants at start of interval	Drop-outs during interval	Implants under risk	Failures during interval	Success rate within period (%)	Cumulative success rate (%)
0–1	389	4	387.0	2	99.5	99.5
1–2	313	3	311.5	2	99.4	98.8
2–3	201	1	200.5	3	98.5	97.4
3–4	136	1	135.5	2	98.5	95.9
4–5	83	5	80.5	1	98.8	94.7
5–6	58	1	57.5	2	96.5	91.4
6–7	30	0	30.0	0	100.0	91.4
7–8	13	0	13.0	0	100.0	91.4

V této tabulce jsou kumulativní pravděpodobnosti přežití jakoby o 1 řádek posunuté nahoru (nezačínají číslem 1.0) a navíc jsou uváděny v procentech.

V dalších pracech, které využívají k prezentaci výsledků generační úmrtnostní tabulky, se vyskytují další varianty značení i další způsoby výpočtu charakteristik. Bohužel často bez podrobnějšího popisu, takže je, díky zaokrouhlování, mnohdy velmi obtížné se k použitým vzorcům dopracovat.

8. Závěr

Porovnáním uvedených přístupů k prezentaci úmrtnostních tabulek docházíme k závěru, že je potřeba postupovat velice obezřetně při porovnávání různých studií. Jejich výsledky mohou být především díky různým přístupům k cenzorovaným pozorováním neporovnatelné (nebo si pro porovnání musíme přepočítat tabulky podle jediného způsobu).

Z uvedených důvodů je vidět, že je nutné při prezentaci úmrtnostní tabulky uvádět přesné vzorce, které použijeme pro výpočet jednotlivých charakteristik.

9. Literatura

Armitage, P., Statistical Methods in Medical Research. Blackwell Scientific Publications, Oxford, 1971, s. 408-414.

Cipra, T., Pojistná matematika v praxi. HZ Praha s.r.o., Praha, 1994.

Hrubeš, J., Hrbáč, L., Pojistná matematika. Internetový studijní materiál, http://ws.vsb.cz/kat/k151/predmety/15115/info.htm, 2001.

Riffenburgh, R.H., Statistics in Medicine. Academic Press, San Diego, 1999, s. 165-167.

STATISTICA for Windows, Volume III (uživatelská příručka). 1999.

INDEXY SPOTŘEBITELSKÝCH CEN

Eva Čermáková
Odbor gesčního zpracování, Český statistický úřad, Myslivečkova 914,
500 03 Hradec Králové

Abstrakt:
Článek popisuje způsob výpočtu indexů spotřebitelských cen a složení spotřebního koše

Metodické vysvětlivky

Vývoj indexů spotřebitelských cen (životních nákladů) se od roku 2001 sleduje na nových spotřebních koších založených na souboru vybraných druhů zboží a služeb placených obyvatelstvem.

Ve spotřebních koších byl proveden výběr výrobků a služeb – cenových reprezentantů a konstruován váhový systém. Současně došlo k zásadní změně v členění úhrnných indexů spotřebitelských cen (životních nákladů), které vychází z mezinárodní klasifikace konečné spotřeby domácností (COICOP) a třídí výrobky a služby do dvanácti hlavních skupin. Došlo také ke změně ceny základního období na prosinec 1999.

Za cenové reprezentanty byly vybrány především výrobky a služby, které se významně podílejí na výdajích obyvatelstva a svým rozsahem pokrývají celou sféru spotřeby.

Počet cenových reprezentantů ve spotřebním koši pro výpočet indexů spotřebitelských cen je v roce 2001 následující:

Skupina zboží a služeb	Počet reprezentantů	Stálé váhy r. 99
Úhrnem	775	1000,0
1. Potraviny a nealkoholické nápoje	162	197,6
2. Alkoholické nápoje a tabák	16	79,2
3. Odívání a obuv	80	56,9
4. Bydlení, voda, energie, paliva	58	236,4
5. Bytové vybavení, zař. domácnosti, opravy	96	67,9
6. Zdraví	39	14,4
7. Doprava	90	101,4
8. Pošty a telekomunikace	15	22,5
9. Rekreace a kultura	111	95,5
10. Vzdělávání	11	4,5
11. Stravování a ubytování	47	74,2
12. Ostatní zboží a služby	50	49,5

Index spotřebitelských cen zboží a služeb charakterizuje cenový vývoj v celospolečenském průměru. Indexy spotřebitelských cen (životních nákladů) ukazují, jak se změny cen odrážejí ve vydáních jednotlivých sociálních skupin domácností. Tyto indexy se vypočítávají za domácnosti zaměstnanců, za domácnosti důchodců, domácnosti s dětmi v nízkém příjmovém pásmu a domácnosti žijící v hl. m. Praze.

Výběr reprezentantů pro všechny typy indexů je stejný, pouze u jednotlivých sociálních skupin domácností (zejména u domácností důchodců) nejsou některé výrobky a služby zastoupeny.

Ceny jednotlivých druhů zboží a služeb jsou zjišťovány měsíčně přímo ve vybraných prodejnách a provozovnách služeb (cca 10000) pracovníky statistických orgánů ve 41 vybraných okresech v celé ČR a hl.m. Praze.

U všech typů indexů byly váhy stanoveny na základě údajů o výdajích domácností podle výsledků statistiky rodinných účtů za rok 1999. U indexu spotřebitelských cen vychází váhy ze struktury průměrných výdajů všech domácností, u ostatních indexů ze struktury výdajů příslušné sociální skupiny domácností.

Způsob výpočtu indexů

Výpočet indexů spotřebitelských cen (životních nákladů) je prováděn na stálých vahách podle vzorce Laspeyresova

$$I_{1/0} = \frac{\sum \frac{p_1}{p_0} \cdot p_0 q_0}{\sum p_0 q_0} \cdot 100$$

p_1	*cena zboží (služby) ve sledovaném (běžném) období*
p_0	*cena zboží (služby) v základním období*
$p_0 q_0$	*stálá váha – výdaje domácností za zboží (službu) v základním období*
$\frac{p_1}{p_0}$	*individuální index ceny určitého druhu zboží či služby*

Základním obdobím pro výpočet indexů jsou váhy roku 1999 a ceny prosince 1999.

Míra inflace vyjádřená **přírůstkem indexu spotřebitelských cen k základnímu období** (prosinec 1999 = 100) vyjadřuje změnu cenové hladiny sledovaného měsíce příslušného roku proti prosinci 1999 (bazický index spotřebitelských cen).

Tato míra inflace je využívána pro analýzu dlouhodobých podrobných trendů (časových řad) vývoje cenových hladin a životních nákladů.

Míra inflace vyjádřená **přírůstkem indexu spotřebitelských cen ke stejnému měsíci předchozího roku** vyjadřuje procentní změnu cenové hladiny ve vykazovaném měsíci daného roku proti stejnému měsíci předchozího roku. Jedná se tedy o dosaženou cenovou úroveň, která vylučuje sezónní vlivy tím, že se porovnávají vždy stejné měsíce.

Tato míra inflace je vhodná ve vztahu ke stavovým veličinám, které měří změnu stavu mezi začátkem a koncem období bez ohledu na průběh vývoje během tohoto období. Bere se v úvahu při propočtech reálné úrokové míry, reálného zvýšení cen majetku, valorizací, apod.

Míra inflace vyjádřená **přírůstkem indexu spotřebitelských cen k předchozímu měsíci** vyjadřuje procentní změnu cenové hladiny sledovaného měsíce proti předchozímu měsíci.

Míra inflace vyjádřená **přírůstkem průměrného ročního indexu spotřebitelských cen** vyjadřuje procentní změnu průměrné cenové hladiny za 12 posledních měsíců proti průměru 12-ti předchozích měsíců (počítáno z bazických indexů).

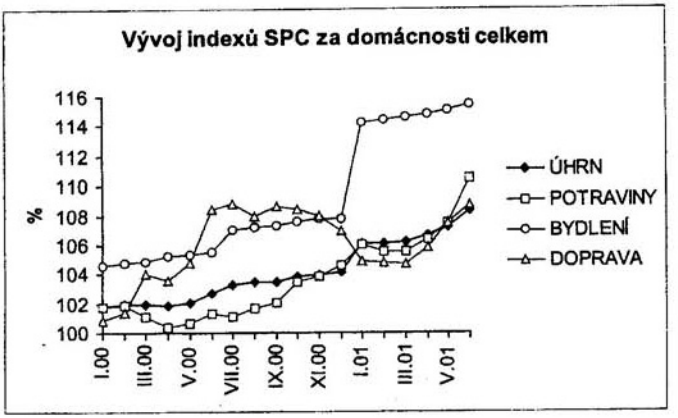
Tato míra inflace je vhodná při úpravách nebo posuzování průměrných veličin. Bere se v úvahu při propočtech reálných mezd, důchodů apod.

Pro potřeby uživatelů jsou ve statistických informacích publikovány indexy kromě toho též k základu průměr roku 2000 = 100.

Kromě toho se počítá **jádrová inflace**, t. j. meziměsíční přírůstek indexu spotřebitelských cen počítaný na celém spotřebním koši po vyloučení vlivu změn regulovaných cen, daňových úprav a jiných administrativních opatření .

Na základě koncepce ČNB se počítá tzv. **čistá inflace**, která se počítá na neúplném spotřebním koši, z něhož jsou vyloučeny položky s cenami ovlivňovanými jinými administrativními opatřeními. Položky, u kterých dochází k cenovým změnám z titulu daňových úprav, zůstávají součástí spotřebního koše, avšak vliv daňových úprav je eliminován.

Na následujícím grafu (Vývoj bazického indexu spotřebitelských cen za domácnosti celkem – k základu prosinec 1999) jsou kromě úhrnu zakresleny indexy ve skupinách s největší váhou ve spotřebním koši (Bydlení, Potraviny, Doprava). Na grafu je dobře vidět, jak index ve skupině Bydlení skokem roste v lednu (zvýšení cen energií) a v červenci (zvýšení regulovaného nájemného). Ve skupině Doprava je vidět, jak byl index zvýšen cenami pohonných hmot v létě 2000.



Literatura:
ČSÚ, CENY: Indexy spotřebitelských cen (životních nákladů) – podrobné členění, Kód 71 03 – 01
VŠE, Fak. Informatiky a statistiky – Úvod do sociálněhospodářské statistiky, Kol. 2000 (str. 61 – 77)
http://www.czso.cz/cz/novinky/inflace/inflace2.htm

PŘESNÝ TEST POMOCÍ KOEFICIENTU SOUHLASU

Josef Bukač
Ústav lékařské biofyziky, Lékařská fakulta v Hradci Králové
Šimkova 870, 50001 Hradec Králové

Abstrakt:
Výpočet p-hodnoty pro přesný test nezávislosti kontingenční tabulky proti alternativě závislosti měřené pomocí koeficientu souhlasu (concordance) je možné urychlit použitím kombinatorické identity.

1. Úvod, podmíněná pravděpodobnost

Předpokládejme, že řádkové součty R_i a sloupcové součty C_j tabulky typu m krát n jsou pevně dány. Za předpokladu nezávislosti sloupcových a řádkových klasifikací bude pravděpodobnost, že políčka tabulky budou obsahovat počty L_{ij}, rovna

$$P = \frac{\prod_{i=1}^{m} R_i! \prod_{j=1}^{n} C_j!}{T! \prod_{i=1}^{m} \prod_{j=1}^{n} (L_{ij})!},$$

kde $T = \sum_{i=1}^{m} R_i$.

Je-li zadána vstupní tabulka, vypočteme vstupní koeficient souhlasu (KS). Chceme generovat všechny tabulky za předpokladu nezápornosti četností a splnění podmínek daných řádkovými a sloupcovými součty, vypočítat jejich pravděpodobnosti a jejich KS porovnat se vstupním KS. Počítáme součty těchto pravděpodobností podle toho, zda je KS menší než vstupní KS, větší než vstupní KS, případně roven vstupnímu KS. Z těchto součtů pak vypočteme p-hodnotu.

Definice KS (coefficient of concordance) je vysvětlena v Agresti (1990), strana 20. Jako příklad uvedeme vstupní tabulku $(0,1,2/0,3,1/3,0,3)$, kde řádky jsou odděleny lomítky. Počet souhlasných párů SP je $SP = 13 = 0\times(3+1+0+3)+1\times(1+3)+0\times(0+3)+3\times 3$. Počet nesouhlasných párů NP je $NP = 27 = 1\times(0+3)+2\times(0+3+3+0)+3\times 3+1\times(3+0)$. Takže vstupní KS je $(SP-NP)/(SP+NP) = (13-27)/(13+27) = -0,35$. Jestliže například vygenerujeme tabulku $(2,1,0/1,2,1/0,1,5)$, bude její $KS = (40-2)/(40+2) = 0,905$ větší než vstupní KS a pravděpodobnost bude $3!4!6!3!4!6!/(13!2!2!5!) = 0,0036$.

Agresti (1990), strana 64, navrhuje použití rozdílu $SP-NP$ k výpočtu přesné p-hodnoty. Uvažujme raději podíl $(SP-NP)/(SP+NP)$.

2. Kombinatorická identita

Celkový počet dvojic počítáme dvojím způsobem a tím získáme kombinatorickou identitu.

Věta.

Jestliže řádkové a sloupcové součty jsou pevné, je možné součet počtu souhlasných párů a nesouhlasných párů napsat jako součet konstanty a jedné poloviny součtu čtverců políček tabulky.

Důkaz.

Předpokládejme, že máme celkem T prvků v tabulce

$$\begin{array}{llll} x_{11}, & x_{12}, & \ldots, & x_{1n} \\ x_{21}, & x_{22}, & \ldots, & x_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ x_{m1}, & x_{m2}, & \ldots, & x_{mn} \end{array}$$

typu m krát n s řádkovými součty $R_1, \ldots, R_m$ a sloupcovými součty $C_1, \ldots, C_n$. Nejprve vypočítáme počet dvojic prvků, když na pořadí ve dvojicích nezaleží, jako $T(T-1)/2$.

Jako druhý krok provedeme výpočty počtů dvojic rozdělených do následujících skupin

1) Počet dvojic s různými indexy jak sloupcovými tak řádkovými. Je jasné, že se jedná o součet počtu dvojic jak souhlasných, tak nesouhlasných, jejichž součet je $SP + NP$.

2) Počet dvojic se sloupcovými indexy různými, ale řádkovými indexy stejnými. Jestliže uvažujeme i-tý řádek, je počet takových dvojic roven

$$(\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{ik} - \sum_{j=1}^{n} x_{ij}^2)/2 = (R_i^2 - \sum_{j=1}^{n} x_{ij}^2)/2.$$

Jestliže sečteme tyto vzorce přes všechna i, dostaneme

$$(\sum_{i=1}^{m} R_i^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2)/2.$$

3) Počet dvojic se sloupcovými indexy stejnými, ale řádkovými indexy různými. Jestliže uvažujeme j-tý sloupec, dostaneme

$$(\sum_{i=1}^{m}\sum_{k=1}^{m} x_{ij}x_{kj} - \sum_{i=1}^{m} x_{ij}^2)/2 = (C_j^2 - \sum_{i=1}^{m} x_{ij}^2)/2.$$

Sečtení těchto vzorců podle j dává

$$(\sum_{j=1}^{n} C_j^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2)/2.$$

4) Počet dvojic se stejnými sloupcovými i řádkovými indexy je roven

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}(x_{ij}-1)/2 = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2/2 - T/2.$$

Součet počtů dvojic ve všech uvedených skupinách je roven $T(T-1)/2$. Z toho obdržíme identitu

$$SP + NP + \sum_{i=1}^{m} R_i^2/2 + \sum_{j=1}^{n} C_j^2/2 - \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2/2 - T/2 = T(T-1)/2.$$

Použijeme označení

$$K_{m,n} = T^2/2 - \sum_{i=1}^{m} R_i^2/2 - \sum_{j=1}^{n} C_j^2/2$$

a identita vypadá takto

$$SP + NP = K_{m,n} + \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2/2,$$

což ukončuje důkaz.

Někteří autoři, Agresti (1990), Somers (1962), uvádějí tento rozklad, protože pomáhá vytvářet další modifikace koeficientu souhlasu. My zde používáme tohoto rozkladu pro výpočet maximální a minimální hodnoty $SP + NP$.

Uvedená identita také ukazuje, že můžeme pro stejné řádkové a sloupcové součty dostat různé součty počtů souhlasných a nesouhlasných dvojic. Tento rozdíl způsobuje, že p-hodnoty získané pomocí $SP - NP$ mohou být různé od p-hodnot získaných pomocí KS.

3. Přesný test

Generujeme tabulky od dolního řádku. Pokud je dolní řádek hotov, vypočítáme jeho příspěvek jak k počtu souhlasných dvojic, tak i k počtu nesouhlasných dvojic, protože jejich počet se nezmění, ať je v řádcích nad dolním řádkem cokoliv.

Abychom zrychlili postup generování tabulek, chceme vědět, zda pro daný dolní řádek můžeme získat tabulku s menším nebo větším KS než je vstupní KS. Nevíme sice jak získat přesná maxima a minima, ale odvodíme užitečné odhady.

Podívejme se na jmenovatel $SP + NP$. Skládá se z příspěvku dolního řádku a také s příspěvku řádků nad ním, které mohou být různé. Protože mohou být zapsány jako

$$SP + NP = K_{m-1,n} + \sum_{i=1}^{m-1} \sum_{j=1}^{n} x_{ij}^2/2,$$

jsme schopni vypočítat minimum $Min(SP+NP)$ a maximum $Max(SP+NP)$, když jsou řádkové i sloupcové součty pevné.

Můžeme použít metodu, kterou navrhl Gross, pro výpočet přesného minima tabulek typu 2 krát n. Tuto metodu popisuje Saaty (1970), strana 184. Jako účelovou funkci definujeme $\sum \phi_j(x_j)$, kde $\phi_j(x_j) = x_j^2 + (C_j - x_j)^2$ je konvexní jako minimalizovaná funkce při podmínce $\sum x_j = R_1$.

K výpočtu maxima konvexní funkce, při omezení $\sum_j x_{ij} = R_i$ pro $i = 1, \ldots, m-1$ a $\sum_i x_{ij} = C_j$ pro $j = 1, \ldots, n$, stačí vyjmenovat vrcholy polyedru daného tímto omezením. Potřebnou teorii uvádí Hadley (1964). Jiné přístupy uvádí Joe (1988).

Jestliže chceme nalézt meze pro čitatele, musíme si uvědomit, že SP i NP jsou nezáporná. Pak $-Max(SP + NP) \leq SP - NP \leq Max(SP + NP)$. Pokud se ovšem chceme zabývat jen testem, který používá rozdíl $SP - NP$, nemusíme se jmenovatelem v KS zabývat.

Dosud jsme dostali meze, neboli intervaly, pro $SP + NP$ a $SP - NP$ a nyní můžeme použít intervalové aritmetiky k přičtení těchto intervalů k příslušným příspěvkům dolního řádku a tím dostat intervaly, ve kterých se musí nacházet čitatel a jmenovatel KS. Nyní již můžeme vydělit tyto intervaly a dostaneme interval, ve kterém se KS nachází.

Porovnání intervalu, ve kterém se KS nachází, se vstupním KS nám pomáhá rozhodnout, zda je nutno řádky nad dolním řádkem generovat

nebo ne. Pokud ne, použijeme součtový vzorec, který dá výsledek bez generování tabulek.

4. Příklad

Budeme pokračovat s tabulkou $(0,1,2/0,3,1/3,0,3)$, pro kterou $C_1 = 3$, $C_2 = 4$, $C_3 = 6$, $R_1 = 3$, $R_2 = 4$ a $R_3 = 6$. Předpokládejme, že je vygenerován dolní řádek $(0,1,5)$. Pak omezení pro zbývající tabulku typu 2 krát 3 jsou $C_1' = 3$, $C_2' = 3$, $C_3' = 1$, $R_1 = 3$ a $R_2 = 4$.

Příspěvek dolního řádku bude $SP_D = 3 \times (1+5) + 3 \times 5 = 33$ a $NP_D = 3 \times 0 + 1 \times (0+1) = 1$. Vypočteme $SP_D + NP_D = 34$ a $SP_D - NP_D = 32$.

Minimální a maximální počty součtu souhlasných a nesouhlasných dvojic zbývající tabulky typu 2 krát 3 jsou 8 a 12. Tím dostaneme interval $[34+8; 34+12] = [42; 46]$, ve kterém musí být součet $SP + NP$ pro celou tabulku 3 krát 3 s dolním řádkem $(0,1,5)$.

Interval, ve kterém musí být čitatel, je $[32-12; 32+12] = [20; 44]$. Podíl dvou intervalů dá interval pro KS jako $[20; 44]/[42; 46] = [20/46; 44/42] = [0,43; 1,05]$. Nyní víme, že KS musí být vetší než 0,43, je tudíž větší než vstupní KS. Nemusíme tudíž generovat zbytek tabulky a použijeme součtový vzorec, který dá 0,014.

Je zajímavé porovnat meze $[0,43; 1]$ s výsledky, které vzniknou při generování všech tabulek s dolním řádkem $(0,1,5)$. Nejmenší KS je $22/44 = 0,5$, zatímco největší KS je $44/46 = 0,975$. Můžeme nejen ověřit, že $[0,5; 0,975] \subset [0,43; 1]$, ale také vidíme, že rozdíl mezi dolními mezemi intervalů je jen 0,07. Tak úspěšný odhad se snadno zdůvodní tím, že uvažujeme relativní chyby.

5. Tabulka 2 krát 2

Generování tabulek provádíme od dolního řádku a nakonec nám zbývají tabulky 2 krát n. Dále pokračuje generování zprava, takže nakonec zbývají tabulky typu 2 krát 2. V takovém případě je možné výpočty zjednodušit.

Budeme uvažovat tabulku $(s,t/u,v)$ a minimalizovat nebo maximalizovat $sv - ut$, kde s,t,u,v jsou nezáporná, $s+t = R_1$, $u+v = R_2$ a $s+u = C_1$. Dostaneme $sv-ut = (R_1+R_2)s - C_1R_1$. Toto je lineární funkce a je snadné získat minimum nebo maximum na intervalu $[s_{min}, s_{max}]$, kde $s_{min} = R_1 - min(R_1, C_2)$ a $s_{max} = min(C_1, R_1)$.

Vypočítáme $sv + ut = 2s^2 + (R_2 - R_1 - 2C_1)s + C_1R_1$. Derivace funkce $f(s) = 2s^2 + (R_2 - R_1 - 2C_1)s + C_1R_1$ je $f'(s) = 4s + (R_2 - R_1 - 2C_1)$

a druhá derivace je $f''(s) = 4$. Vidíme, že $f(s)$ je konvexní. K nalezení maxima na $[s_{min}, s_{max}]$ stačí vzít vetší hodnotu z $f(s_{min})$ a $f(s_{max})$.

K nalezení minima nepostačí jen zkusit menší z hodnot $f(s_{min})$ a $f(s_{max})$, ale musíme také řešit rovnici $f'(s) = 0$ a výsledek zaokrouhlit nahoru a dolů k nejbližším celým číslům $s_1 = \lfloor (2C_1 + R_1 - R_2)/4 \rfloor$ a $s_2 = \lceil (2C_1 + R_1 - R_2)/4 \rceil$. Jestliže s_1 nebo s_2 je v intervalu $[s_{min}, s_{max}]$, musíme také uvažovat hodnoty $f(s_1)$ nebo $f(s_2)$.

6. Závěr

Naznačili jsme, že není třeba generovat všechny tabulky. Meze, které jsme použili, se mohou zdát hrubé, ale přesné meze je obtížné získat, protože se jedna o optimalizaci složité účelové funkce. Náš přístup se jeví z tohoto pohledu jako rozumný kompromis.

Výpočet se při metodě větvení a mezí dá urychlit nejen zpřesněním a zrychlením výpočtu mezí, ale také vhodnou strukturou dat a sdružováním částečně vygenerovaných tabulek dávajících stejné sloupcové součty.

Nelze samozřejmě provádět výpočty rychleji než u testu nezávislosti proti obecné alternativě závislosti. Hlavním důvodem je to, že u obecné alternativy nezáleží na pořadí sloupců a řádků. To umožňuje sdružování částečně generovaných tabulek do skupin a tím podstatné zrychlení. Tento přístup nemůže být použit v případě KS.

7. Literatura

Agresti, A., Categorical Data Analysis. John Wiley and Sons, New York, 1990.

Hadley, G., Nonlinear and Dynamic Programming. Addison-Wesley, Reading, MA, 1964.

Joe, H., Extreme probabilities for contingency tables under row and column independence with applications to Fisher's exact test. Communications in Statistics Theory and Methods, vol. 17, p. 3677-3685, 1988.

Saaty, T.L., Optimization in Integers and Related Extremal Problems. McGraw-Hill Book Company, New York, 1970.

Somers, R.H., A new asymmetric measure of association for ordinal variables. American Sociological Review, 27, p.799-811, 1962.

VZDÁLENOST POZOROVANÝCH HODNOT

Zdeněk Fabián

Ústav informatiky AV ČR,
Pod vodárenskou věží 2, 18200 Praha 8,
e-mail Zdenek@uivt.cas.cz

1 Vlivová funkce

Každý ví, jak byla sestrojena Celsiova teplotní stupnice. Vzdálenost bodu varu a bodu tání vody děleno stem a lineární extrapolace na obě strany. Extrapolace ve směru záporných teplot narazila na hradbu absolutní nuly v $t = -273°C$ a nějaké drobné.

Jaká je vzdálenost dejme tomu údajů $t_1 = 40°C$ a $t_2 = 50°C$? Odpověď 10°C nemá moc velkou výpovědní hodnotu. Hodnota dílku stupnice je totiž závislá na poloze dílku na stupnici. Vzdálenost mezi 40°C a 42°C je vzdáleností mezi životem a smrtí. Vzdálenost setiny dílku v blízkosti absolutní nuly je vyčíslitelná v miliónech dolarů. Naopak hodnota vzdálenosti mezi teplotami v nitrech dvou hvězd, vyjádřená statisíci stupni Celsia, je z lidského pohledu nezajímavá, protože rozdíl mezi jejich spektry je neměřitelný.

Lidské praxi by lépe vyhovovala teplotní stupnice *nelineární*. Vzdálenosti mezi teplotami t_1, t_2 by bylo lépe vyjádřovat ne pomocí obvyklého Euklidova vzorce, ale nelineárním vztahem

$$d_T(t_1, t_2) = |T(t_2) - T(t_1)|. \tag{1}$$

Jak by měla funkce T vypadat ? Absolutní teplotní nula je zřejmě minus nekonečno, zatímco teplotní plus nekonečno by byla nějaká konečná hodnota $T(\infty)$, ke které by se křivka $T(t)$ asymptoticky blížila. Funkci T nazveme třeba *vlivovou teplotní funkcí vzhledem k lidské aktivitě*.

S podobnou situací se setkáváme i ve statistice. Označme R reálnou přímku. Ve výběrovém prostoru $S_P \subseteq R$ náhodné veličiny X s rozdělením P absolutně spojitým vzhledem k Lebesguově míře měříme pravítkem, ačkoliv nějaká metrika odvozená od daného rozdělení by asi byla užitečnější. Nechť X znamená třeba hmotnost dospělého člověka. Vzdálenost $d(x_1, x_2) = 1kg$ mezi $x_1 = 30kg$ a $x_2 = 29kg$ znamená půlroční hladovění, zatímco k dosažení této vzdálenosti mezi $x_1 = 180kg$ a $x_2 = 179kg$ stačí, když si zápasník sumo utře pot. I zde by vzdálenosti mezi prvky datového souboru měly být měřeny pomocí (1) s nějakou vlivovou funkcí, v tomto případě řekněme funkcí kvantifikující úsilí potřebné ke zhubnutí.

Je patrně jasné, že problém obsažený v titulu článku není vlastně problémem matematické statistiky. To, oč běží, je *vzdálenost mezi body výběrového prostoru daného rozdělení* (statistika k tomu pouze dodá jak ji odhadovat z dat). Odpověď teorie pravděpodobnosti

byste však hledali marně. Je skutečně až s podivem, že množství pravděpodobnostních rozdělení nemá ve výběrovém prostoru odpovídající sadu metrik. Jediná vzdálenost, kterou teorie pravděpodobnosti přisuzuje výběrovému prostoru, je vzdálenost Euklidova. Nelineární metriky nejsou v módě ani jinde. I když se desetibojař přiblíží v některé disciplíně současné hranici lidských možností, naskakují mu body lineárně.

V tomto článku se pokusíme vhodnou vzdálenost zkonstruovat.

2 Core funkce na přímce

Odpověď na otázku kde začít je snadná: známe hustotu $f(x)$ náhodné veličiny X a nic dalšího.

Hustota $f(x)$ je reálná funkce definovaná na R. Předpokládejme ji regulární (hladkou). Nejdůležitější charakteristikou reálné funkce je její derivace. Protože $f(x)$ udává lokální relativní pravděpodobnost, je asi rozumné 'zrovnoprávnit' všechny hodnoty $f'(x)$ pomocí normování na $f(x)$. Funkci

$$T_f(x) = -\frac{f'(x)}{f(x)} \tag{2}$$

nazveme *core funkcí* pravděpodobnostního rozdělení (termín vlivová funkce je obsazen).

Core funkce rozdělení sice není v teorii pravděpodobnosti studována, ale je dlouhou dobu známa ve statistice pod jménem skórová funkce. Uvažujme náhodný výběr $\mathbf{X}_n = (x_1, \ldots, x_n)$ z parametrického rozdělení $f(x-\mu)$, kde μ je neznámý parametr polohy. V teorii odhadu se tento výběr transformuje na soubor $\mathbf{s}_n = (s^f_\mu(x_1-\mu), \ldots, s^f_\mu(x_n-\mu))$, kde $s^f_\mu(x)$ je věrohodnostní skór pro parametr μ, $s^f_\mu(x) = \frac{\partial}{\partial\mu} \ln f(x-\mu)$, který je v tomto případě shodný se skórovou i core funkcí rozdělení. Soubor $\mathbf{s}_n$ je sice jen 'latentní', ale μ je možno odhadnout na základě požadavku, že soubor $\mathbf{s}_n$ má mít nulový aritmetický průměr, což je požadavek metody maximální věrohodnosti.

Ačkoli se core funkce shoduje s věrohodnostním skórem jen v případě rozdělení bez parametru měřítka, je to shoda důležitá. Core funkce je ve speciálním případě identická s tím nejlepším, co statistika nabízí (maximálně věrohodný odhad s nejmenším možným rozptylem).

Zdá se, že jsme hotovi. Vzdálenosti v S_P můžeme měřit, podobně jako v případě teplotní vzdálenosti, pomocí neeuklidovské formule

$$d_T(x_1, x_2) = |T_f(x_2) - T_f(x_1)| \tag{3}$$

s vlivovou funkcí (2) jednoznačně určenou (předpokládaným) rozdělením, která ve výběrovém prostoru zavádí *'maximálně věrohodnou metriku' vzhledem k nejdůležitějšímu parametru rozdělení.*

3 Core funkce na polopřímce a intervalu

Je tu však problém. Uvedený postup není obecný. Výraz (2) ztrácí rozumný smysl, když se dostaneme do světa rozdělení s nosičem $S_P \neq R$, jako v případě náhodné veličiny hmotnost. Uveďme jednoduchý příklad. Hustota exponenciálního rozdělení na intervalu $S_P = (0, \infty)$, $f(x) = e^{-x}$, nemá v $(0, \infty)$ maximum, nemá ani parametr polohy, a pro skórovou funkci platí $-f'(x)/f(x) = 1$, což je pro úvahy o vzdálenosti nepoužitelné.

Ukázalo se (viz [2]), že podstatný není speciální tvar funkce (2), ale fakt, že (2) je věrohodnostínm skórem pro 'těžiště' rozdělení, kterým je na $S_Q = R$ mód. Označme jeho polohu y^*. Co však je 'těžištěm' rozdělení s nosičem $S_P \neq R$? Mód ani střední hodnota nemusí existovat a medián není geometrická charakteristika.

V [1] byl zobecněn a rozveden padesát let starý Johnsonův nápad (viz [3]). Buď P rozdělení s nosičem $S_P = (a, b)$ kde $-\infty \leq a < b < +\infty$ a $\varphi : S_P \to S_Q = R$ spojité ryze monotonní zobrazení. Najdeme Q s hustotou g tak, aby $P = Q\varphi$. Hustota f rozdělení P je pak zřejmě

$$f(x) = g(\varphi(x))\dot{\varphi}(x) \tag{4}$$

kde $\dot{\varphi}(x) = d\varphi(x)/dx$ je Jacobián transformace $\varphi : S_Q \to R$.

Jako φ vezmeme zobecněné Johnsonovo zobrazení

$$y = \varphi_{x_0}^{ab}(x) = \ln \frac{(b - x_0)(x - a)}{(x_0 - a)(b - x)}, \tag{5}$$

kde $a < x_0 < b$, které se pro nosiče $S_P = (0, \infty)$ a $x_0 = 0$, resp. $S_P = (a, b)$ a $x_0 = (a + b)/2$ redukuje na Johnsonovo, tj.

$$\varphi_0^{0\infty}(x) = \ln x, \qquad \varphi_{(a+b)/2}^{ab}(x) = \ln \frac{x - a}{b - x}. \tag{6}$$

Dá se ukázat ([2]), že takto lze chápat všechna známá modelová rozdělení na $S_P = (0, \infty)$ a většinu rozdělení na intervalu (s několika málo výjimkami, která nás zde nebudou zajímat). Jejich množinu označíme $\mathcal{J}$ a dále budeme značit $\varphi = \varphi_{x_0}^{ab}$.

A teď konečně vytáhneme králíka z klobouku. Je-li dáno $P \in \mathcal{J}$, sestrojíme $Q = P\varphi^{-1}$ a podle (2) určíme jeho core funkci T_g, jejíž obraz na S_P budeme považovat za 'vlivovou' funkci 'indukovaného' rozdělení P.

DEFINICE 1 *Funkci*

$$T_f(x) = T_g(\varphi(x)). \tag{7}$$

nazveme core funkcí rozdělení P.

T_f může být vyjádřena nezávisle na zdrojovém rozdělení pouze pomocí f a φ.

VĚTA 1 *Pro core funkci (7) platí*

$$T_f(x) = \frac{1}{f(x)} \frac{d}{dx}(-f(x)/\dot{\varphi}(x)). \tag{8}$$

DŮKAZ. Položme $y = \varphi(x)$. Užitím (7), (2) a (4)

$$T_f(x) = T_g(\varphi(x)) = -\frac{1}{g(y)} g'(y) = \frac{\dot\varphi(x)}{f(x)} \frac{d}{dx}(-f(x)/\dot\varphi(x)) \frac{dx}{dy}$$

a $(dy/dx) = \dot\varphi(x)$.

Poznamenejme, že vzorec (8) se pro $S_P = R$ redukuje na (2) při identickém zobrazení $\varphi(x) = x$ a je tedy zcela obecný.

Ekvivalentem 'těžiště' zdrojového rozdělení bude jeho obraz na S_P při zobrazení φ, t.j. $x^* = \varphi^{-1}(y^*)$. V parametrickém případě to znamená

$$\tau = \varphi^{-1}(\mu). \tag{9}$$

Parametr τ nazveme *Johnsonovým parametrem polohy.*

PŘÍKLAD. Buď $S_P = (0, \infty)$. Zde platí $\varphi(x) = \ln x$, $\dot\varphi(x) = 1/x$, $\tau = e^\mu$ a $T_f(x) = (-xf(x))'/f(x) = -1 - xf'(x)/f(x)$. Exponenciální rozdělení s hustotou $f_\tau(x) = \tau^{-1} e^{-x/\tau}$s Johnsonovým parametrem polohy τ má zdrojové rozdělení s hustotou $g_\mu(y) = e^{y-\mu} \exp(-e^{y-\mu})$ a skórovou funkcí $T_g(y) = e^{y-\mu} - 1$. Podle (7) je core funkcí exponenciálního rozdělení $T_f(x) = x/\tau - 1$.

Modelová parametrická rozdělení na polopřímce a intervalu vznikala v průběhu rozvoje statistiky a stalo se, že některá z nich Johnsonův parametr polohy mají a některá ne. Pro ta, která ho mají, platí

VĚTA 2 *Core funkce rozdělení je vnitřní částí věrohodnostního skóru pro Johnsonův parametr polohy.*

DŮKAZ. Buď $g(u)$, kde $u = (y - \mu)/\sigma$, hustota zdrojového rozdělení příslušného k rozdělení $P = Q\varphi^{-1}$ s hustotou $f_\theta(x)$. Užitím (4), (6) a (9) dostáváme $f_\theta(x) = g(w)\dot\varphi(x)$ kde

$$w = \frac{\varphi(x) - \varphi(\tau)}{\sigma}.$$

Podle (4) je věrohodnostní skór pro τ

$$l^f_\tau(x) = \frac{\partial}{\partial \tau} \ln f_\theta(x) = \frac{\partial}{\partial \tau} \ln(g(w)\dot\varphi(x)) = \frac{g'(w)}{g(w)} \frac{\partial w}{\partial \tau} = \frac{\dot\varphi(\tau)}{\sigma} T_g(w). \tag{10}$$

Core funkcí rozdělení je tedy vnitřní částí věrohodnostního skóru pro 'těžiště' rozdělení. V případě rozdělení bez Johnsonova parametru polohy nebo bez parametrů vůbec je to dosud neznámá funkce ('těžištěm' f je pak bod $x^* = \varphi^{-1}(y^*)$).

4 Vzdálenost pozorovaných hodnot

Core funkce regulárního pravděpodobnostního rozdělení $P \in \mathcal{J}$ je obecně definována vztahem (8) a zavádí do výběrového prostoru 'maximálně věrohodnou metriku' vzhledem k 'těžišti' rozdělení. Za vzdálenost bodů $x_1, x_2 \in S_P$ lze tedy považovat výraz

$$d_T(x_1, x_2) = |T_f(x_2) - T_f(x_1)| = \int_{x_1}^{x_2} \rho_P(x)\, dx, \tag{11}$$

kde $\rho_P(x) = dT_f(x)/dx$. Pokud je T_f spojitá rostoucí, je (S_P, ρ_P) Riemannův metrický prostor.

V Tabulce 1 jsou dány hustoty, core funkce a příslušné 'core' vzdálenosti bodů $x_1, x_2 \in S_P$. Některá rozdělení v tabulce tvoří páry zdrojového a indukovaného rozdělení, například normální a lognormální, Gumbelovo a Weibullovo (pro obě dvojice platí $\beta = 1/\sigma$), logistického a log-logistického, a konečně Cauchyho a log-Cauchyho. Poslední dvojice nemá core funkce monotónní a 'core' vzdálenost nelze zavést takto jednoduše. Je zajímavé, že vzdálenost (11) ve výběrovém prostoru je Euklidova nejenom v případě normálního rozdělení na $S_P = R$, ale i pro gamma rozdělení (a jeho speciální případ exponenciální rozdělení) na $S_P = (0, \infty)$ a pro beta rozdělení (a jeho speciální případ rovnoměrné rozdělení) na $S_P = (0, 1)$. Povšimněte si, že core funkce normálního rozdělení se liší od skórové funkce, která se obvykle chápe jako věrohodnostní skór pro μ, který je $s_\mu^N(x) = (x - \mu)/\sigma^2$.

TABULKA 1

Name	S_P	$f(x)$	$T_f(x)$	$\rho(x_1, x_2)$
normální	R	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$	$\frac{x-\mu}{\sigma}$	$(x_2 - x_1)/\sigma$
lognormální	$(0, \infty)$	$\frac{1}{\sqrt{2\pi}\,\sigma x} e^{-\frac{1}{2}\ln^2(x/\tau)^\beta}$	$\ln (x/\tau)^\beta$	$\ln(\frac{x_2}{x_1})^\beta$
Gumbelovo	R	$e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}}$	$e^{\frac{x-\mu}{\sigma}} - 1$	$e^{-\mu/\sigma}(e^{x_2/\sigma} - e^{x_1/\sigma})$
Weibullovo	$(0, \infty)$	$\frac{\beta}{x}\left(\frac{x}{\tau}\right)^\beta e^{-(x/\tau)^\beta}$	$\left(\frac{x}{/\tau}\right)^\beta - 1$	$\tau^{-\beta}(x_2 - x_1)^\beta$
extr. hodn II	$(0, \infty)$	$x^{-2}e^{-1/x}$	$1 - 1/x$	$\frac{x_2 - x_1}{x_1 x_2}$
logistické	R	$e^x/(1 + e^x)^2$	$\tanh(x/2)$	$\frac{\sinh((x_2-x_1)/2)}{\cosh(x_1/2)\cosh(x_2/2)}$
log-logistické	$(0, \infty)$	$1/(z + 1)^2$	$(z - 1)/(z + 1)$	$\frac{2(x_2-x_1)}{(x_1+1)(x_2+1)}$
Lomaxovo	$(0, \infty)$	$\alpha/(1 + x)^{\alpha+1}$	$(\alpha(x - 1)(x + 1)$	$\frac{(\alpha+1)(x_2-x_1)}{(x_1+1)(x_2+1)}$
gamma	$(0, \infty)$	$\frac{\gamma^\alpha}{\Gamma(\alpha)}x)^{\alpha-1}e^{-\gamma x}$	$\gamma x - \alpha$	$\gamma(x_2 - x_1)$
beta	$(0, 1)$	$\frac{1}{B(p,q)}x^{p-1}(1 - x)^{q-1}$	$(p + q)x - p$	$(p + q)(x_2 - x_1)$
Cauchyho	R	$1/\pi(1 + x^2)$	$2x/(1 + x^2)$	?
log-Cauchyho	$(0, \infty)$	$1/\pi(1 + \ln^2 x)$	$2\ln x/(1 + \ln^2 x)$	?

Na závěr se vrátíme k titulu článku. Za vzdálenost pozorovaných hodnot $x_1, x_2 \in S_P$ náhodného výběru $\mathbf{X}_n$ z rozdělení $P \in \mathcal{J}$ lze považovat vzdálenost (11), možná normovanou

příslušnou Fisherovou informací, po (parametrickém či neparametrickém) odhadu hustoty f rozdělení.

Poděkování. Autor děkuje I. Vajdovi za cenné rady a postřehy. Práci podpořil grant GA AVČR A1075101.

Odkazy.

[1] Fabián, Z. (1997). Information and entropy of continuous random variables. *IEEE Trans. on Information Theory*, 43, 1080-1083.

[2] Fabián, Z. (2001). Induced cores and their use in robust parametric estimation. *Commun. in Statistics, Theory Methods*, 30, 3, 537-556.

[3] Johnson, N.L. (1949). Systems of Frequency Curves Generated by Methods of Translations. *Biometrika* **36**, 149-176.

Statistické vyhodnocení změn srážko – odtokových vztahů vybraných povodí řeky Moravy

Ladislav Budík, Marie Budíková

Abstrakt: Příspěvek zkoumá změny mezi obdobími 1931-60 a 1961-90, k nimž došlo ve srážko-odtokovém vztahu v některých povodích řeky Moravy, a to jak v ročním, tak v měsíčním chodu. Z analýz provedených základními statistickými metodami vyplynulo, že ve většině povodí narůstá odtok, i když srážky klesají. Pravděpodobnou příčinou tohoto jevu je poškození vegetace spolu se snížením retence v kombinaci se změnami územního výparu. Tím je narušen malý vodní koloběh, neboť řeky odvádějí do oceánu větší podíl srážkové vody.

1. Úvod

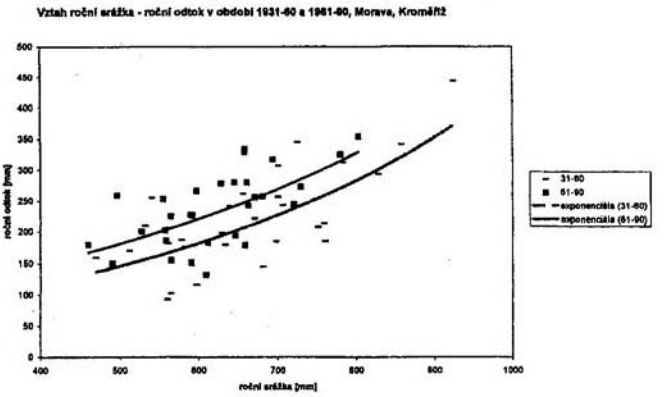
V působnosti brněnské pobočky Českého hydrometerologického ústavu se v současné době nachází 101 vodoměrných a 145 srážkoměrných stanic rozmístěných na území o rozloze 15548 km^2. Při analýze vztahů mezi srážkami a odtoky si hydrologové povšimli, že charakter závislosti odtoků na srážkách se na přelomu 50. a 60. let začíná měnit. Zdálo se, že ač klesá roční úhrn srážek, roste roční odtok. Toto zjištění dalo impuls k podrobnějšímu zkoumání ročních a měsíčních hodnot srážkových a odtokových charakteristik

Srážkou rozumíme množství vody, které za danou časovou jednotku spadne na jednotkovou plochu. Udává se v mm vodního sloupce. K měření srážek slouží srážkoměrné nádoby.
Průtok je objem vody, která proteče daným profilem za časovou jednotku. Udává se v m^3/s. Průtok lze přímo měřit hydrometrickou vrtulí.
Stav vody v daném říčním profilu se nepřetržitě měří limnigrafem a udává se v cm výšky vodní hladiny. Stav se převádí na průtok pomocí tzv. konsumpční křivky (vztah stav - průtok) získané z epizodických přímých měření průtoku při znalosti stavu v době měření pomocí hydrometrické vrtule.
Odtok (nebo též odtoková výška) je množství vody, které za danou časovou jednotku odteče z určitého území. Udává se v mm vodního sloupce. Mezi odtokem a průtokem existuje vzájemně jednoznačný vztah: odtok [mm] = průtok [l/s]*čas[s]/plocha [m^2].

Poměrně homogenní údaje o měsíčních srážkách a průtocích vesměs od r. 1931 jsou k dispozici na šesti profilech: Borovnice na Svratce, Dolní Bory na Oslavě, Dvorce na Jihlavě, Kroměříž na Moravě, Plumlov na Hloučele (zde jsou k dispozici údaje až od r. 1935) a Skalní Mlýn na Punkvě. S výjimkou Kroměříže se jedná o profily, které charakterizují horní části povodí.
V povodí Moravy bylo možno vyčlenit ještě dalších osm profilů a mezipovodí, kde změna charakteru ovlivnění antropogenní činností mezi obdobími 1931-1960 (1. období) a 1961-1990 (2. období) je zanedbatelná v ročním, ale nikoli měsíčním chodu. Jde o profily Moravičany na Moravě, Nové Sady na Moravě, Ptáčov na Jihlavě, Raškov na Moravě, Veverská Bitýška na Svratce, Vranov na Dyji a mezipovodí Trávní Dvůr – Ivančice – Židlochovice – Dolní Věstonice a Kroměříž – Strážnice.

2. Regresní analýza ročních hodnot

Pro uvedených 14 profilů a mezipovodí jsme zkoumali srážko – odtokový vztah v ročním chodu, a to zvlášť pro 1. a 2. období. Pro ilustraci uvádíme na obr. 1 dvourozměrné tečkové diagramy pro 1. a 2. období proložené exponenciálními křivkami v profilu Kroměříž. (Exponenciály poskytovaly vyšší indexy determinace než přímky či polynomické křivky.)



Obr. 1

V mnoha profilech - zvláště z vyšších poloh - se prokázalo, že v období 1961-90 je zhruba stejné rozdělení ročních odtoků jako v letech 1931-60, ale při nižších ročních srážkových úhrnech. Přitom nejvyšší srážkové úhrny dosažené v 1. období ve 2. období zcela chybí. Zároveň se ukazuje, že ve 2. období došlo k zešikmení srážek. S mnohem vyšší pravděpodobností se vyskytují menší úhrny a pravděpodobnosti výskytu úhrnů vyšších srážek klesají.

3. Analýza měsíčních hodnot

Nejprve jsme se zabývali měsíčními úhrny srážek a průměrnými měsíčními hodnotami odtoků. Informace o nárůstu či poklesu srážek a odtoků jsou uvedeny v tabulkách 1 a 2. Symbol + znamená nárůst, symbol – pokles.

Porovnání období 1931 – 1960 a 1961 – 1990, srážky

měsíc	profil					
	Borovnice	Dolní Bory	Dvorce	Kroměříž	Plumlov	Skalní Mlýn
I	-	+	+	-	-	+
II	-	-	+	-	-	-
III	+	+	+	-	+	+
IV	+	+	-	+	-	-
V	+	+	+	+	+	+
VI	+	-	-	+	+	+
VII	-	-	-	-	-	-
VIII	+	-	-	-	-	-
IX	+	+	+	-	+	-
X	-	-	-	-	-	-
XI	-	+	+	-	+	-
XII	+	+	+	-	-	+
celkem	5- 7+	5- 7+	5- 7+	9- 3+	7- 5+	7- 5+
rozdíl v mm	-1,1	8,8	-26,6	-39,3	-27,1	-19,5

Tabulka 1

S výjimkou Dolních Borů vykazují sledované profily ve 2. období oproti 1. období srážkový deficit, který je nejvýraznější v Kroměříži. Pokles srážek je zaznamenáván především v červenci, srpnu a říjnu.

Porovnání období 1931 – 1960 a 1961 – 1990, odtoky

měsíc	profil					
	Borovnice	Dolní Bory	Dvorce	Kroměříž	Plumlov	Skalní Mlýn
I	-	+	+	+	+	+
II	-	-	-	+	+	+
III	-	-	+	-	-	+
IV	+	-	-	+	+	+
V	+	+	+	+	+	+
VI	+	+	+	+	+	+
VII	+	-	+	+	-	+
VIII	-	-	-	+	+	+
IX	-	-	-	-	-	-
X	-	-	-	+	-	+
XI	-	-	-	-	-	-
XII	+	+	-	+	+	+
celkem	7- 5+	8- 4+	7- 5+	3- 9+	5- 7+	2- 10+
rozdíl v mm	-5,7	-8,0	3,6	14,5	17,0	24,0

Tabulka 2

Ve 2. období na všech profilech kromě Borovnice a Dolních Borů vzrostly odtoky, a to především v květnu, červnu a lednu.

Z hydrologického hlediska se jeví nejzajímavější tabulka 3, která obsahuje informace o vzájemném vztahu srážek a odtoků. Symbol + znamená, že současně vzrostly srážky i odtoky, symbol – znamená, že současně poklesly srážky i odtoky, symbol / značí, že s poklesem srážek došlo k nárůstu odtoků a symbol \ znamená, že s nárůstem srážek došlo k poklesu odtoků.

Porovnání období 1931 – 1960 a 1961 – 1990, srážko-odtokový vztah

měsíc	profil					
	Borovnice	Dolní Bory	Dvorce	Kroměříž	Plumlov	Skalní Mlýn
I	-	+	+	/	/	+
II	-	-	\	/	/	/
III	\	\	+	-	\	+
IV	+	/	-	+	/	/
V	+	+	+	+	+	+
VI	+	/	/	+	+	+
VII	/	-	/	/	-	/
VIII	\	-	-	/	/	/
IX	\	\	\	-	\	-
X	-	-	-	/	-	/
XI	-	\	\	-	\	-
XII	+	+	\	/	/	+
celkem	4+ 3\ 1/ 4-	3+ 3\ 2/ 4-	3+ 4\ 2/ 3-	3+ 0\ 6/ 3-	2+ 3\ 5/ 2-	5+ 0\ 5/ 2-

Tabulka 3

Rovněž nás zajímala intenzita lineárního vztahu mezi srážkami a odtoky. Měsíční úhrny srážek a průměrné měsíční odtoky však vykazují kladně zešikmené rozložení četností. Proto je nutné tyto hodnoty před dalšími analýzami transformovat tak, aby měly aspoň přibližně normální rozložení. Srážky jsou podrobeny odmocninové transformaci, průtoky logaritmické. Pro takto transformovaná data byly vypočteny výběrové koeficienty korelace a byla testována hypotéza o shodě korelačních koeficientů v 1. a 2. období. Případy, kdy nulová hypotéza byla zamítnuta na hladině významnosti 0,05, jsou zachyceny v tabulce 4.

profil	měsíc	r_{12} v 1. období	r_{12} ve 2. období	p-hodnota
Borovnice	III	0,27	0,67	0,05
Dolní Bory	V	0,38	0,78	0,02
Dvorce	I	0,15	0,68	0,01
	III	0,16	0,65	0,03
Kroměříž	IX	0,69	0,21	0,02

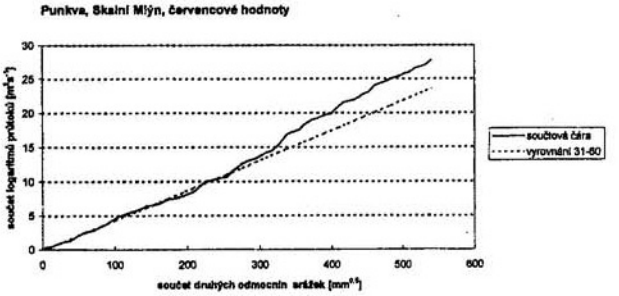
Tabulka 4

Pokles koeficientu korelace v září v Kroměříži lze vysvětlit tím, že v tomto povodí dlouhodobě klesají srážky v srpnu. Tím se uvolní vodní rezervoáry povodí a zářijové srážky jsou ve 2. období zadržovány mnohem intenzivněji než v 1. období.

4. Dvojné součtové čáry

V další části příspěvku jsme se zabývali zkoumáním odezvy průtoků na srážku.
V hydrologii se pro tyto účely používá dvojná součtová čára. Její konstrukce spočívá v tom, že na vodorovnou osu se vynášejí kumulované hodnoty srážek a na svislou osu kumulované hodnoty průtoků. Mají-li dvojné součtové čáry vykazovat co nejmenší rušivý šum spočívající ve výkyvech způsobovaných střídajícími se suchými a vlhkými obdobími, je nutné, aby kumulované veličiny byly přibližně homoskedastické. Je tedy zapotřebí transformovat průtoky logaritmickou transformací a srážkové úhrny odmocnit. Takto konstruovaná dvojná součtová čára má menší výkyvy.

Pro ilustraci uvádíme na obr. 2 ukázku dvojné součtové čáry pro červencové hodnoty srážek a průtoků na Punkvě ve stanici Skalní Mlýn. Hodnotami z období 1931– 60 je metodou nejmenších čtverců proložena přímka prodloužená i do následujícího období. Díky tomu je zřetelně vidět, že pro hodnotu asi 280 na vodorovné ose, která odpovídá roku 1959, dochází k výrazné změně srážko – odtokového vztahu, dál je větší nárůst odtoku při stejné srážce.

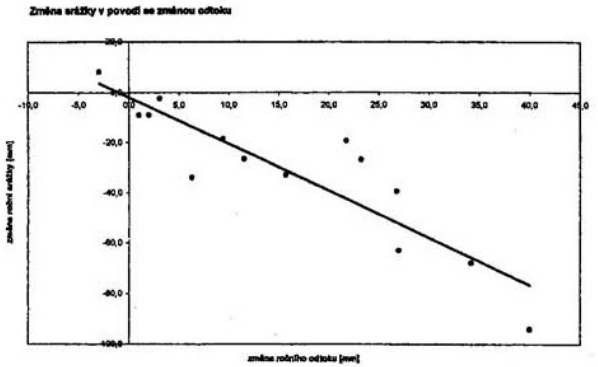


Obr. 2

V některých případech však odtok ve druhém období klesal. Na většině profilů se jednalo o měsíce září a listopad resp. prosinec. Nárůst odtoku naopak byl pozorován především v měsících červenci, srpnu, květnu a lednu, a to většinou i na profilech, kde v ročním chodu k pozorovatelným změnám nedošlo. To je patrné i v tabulce č. 3.

Změny ve srážko-odtokovém vztahu v 1. a 2. období lze dobře postihnout sledováním rovnoběžnosti proložených přímek v prvním a druhém období. Pokud se nezměnila odezva povodí na srážku, měla by směrnice přímky v 1. období být stejná jako směrnice ve 2. období. Znatelná změna rovnoběžnosti v obou obdobích prokládaných přímek byla v Borovnici 6x, Dolních Borech 4x, Dvorcích 8x, Kroměříži 10x, Plumlově 11x a Skalním Mlýně 10x. Z těchto 49 významnějších změn srážko-průtokového vztahu došlo ve 37 případech ke zvýšení směrnice proložené přímky. (Ve zbylých 23 případech méně znatelných změn nebo beze změny se směrnice zvýšila 5x.)

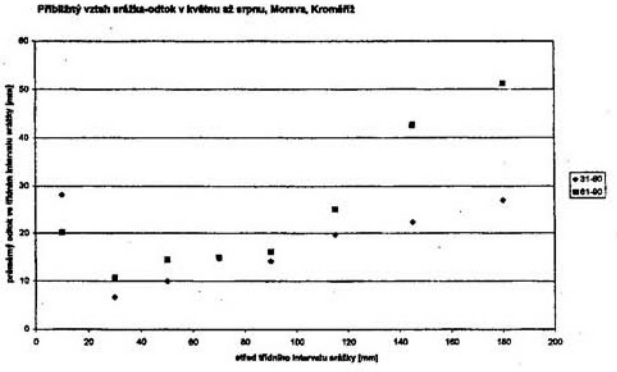
Pokusili jsme se tyto změny indikované ze součtových čar vyhodnotit i kvantitativně. Tento odhad jsme provedli zpětným výpočtem z odchylky změny směru součtové čáry z 2. období vůči 1. období. Z odchylek součtových čar zkonstruovaných pro 14 profilů a mezipovodí byly zpětně počítány změny odtoků ve 2. období oproti 1. období. Získané odchylky mezi obdobími byly vyneseny do obr. 3. Dvourozměrným tečkovým diagramem změn ročních hodnot jsme proložili regresní přímku, jejíž koeficient determinace činil 0,81.



Obr. 3

5. Rozdělení srážek do třídních intervalů

Z analýz měsíčních hodnot vyplynulo, že dochází ke změnám odtoků zvláště ve vegetační části roku a především při vyšších měsíčních srážkových úhrnech. Tuto změnu lze v datech lépe zviditelnit např. rozdělením srážky do třídních intervalů zhruba po 30-ti mm a v každém třídním intervalu spočítat průměrný odtok. To jsme provedli zvlášť pro1. období, zvlášť pro 2. období a pro porovnání vynesli do společného obrázku č. 4.



Obr. 4

Na něm je uveden příklad výsledku pro souhrn měsíců květen až srpen v profilu Kroměříž. V těchto čtyřech měsících jsou obecně největší změny směrem ke vzrůstu odtoku na všech šesti zkoumaných profilech. Změna směrem ke vzrůstu odtoku v tomto jarně-letním období je i v povodích, kde ke změně odtoku za celý rok nedochází. V těchto případech je ale mnohem menší.

6. Interpretace závěrů.

Předkládáme pravděpodobnou interpretaci výsledků (zatím jediná bezesporná vůči ostatním známým jevům i jiným jevům popsaným v [1] – na ty zde ale není prostor). Jelikož s úměrně narůstající změnou odtoku klesá srážka, viz obr 3, tj. při vzrůstajícím dlouhodobém odtoku klesá dlouhodobá průměrný srážkový úhrn a naopak, plyne z toho, že odtok je svázán se srážkou přes retenci (zadržení vody v povodí) a výpar. Pokud část vody není zadržena a zpětně výparem uvedena do koloběhu, sníží se srážka o hodnotu přibližně dvojnásobnou, jak je vidět z průběhu regresní přímky na obr. 3. Ve ztrátě je započteno především doplňování půdní vláhy, podzemních vod a výparu. Z dlouhodobého pohledu mají první dvě položky stálý objem a nemění se, pouze výpar převádí vodu z položky ztráta někam jinam, tj. do srážky. Tedy celou změnu srážko-odtokového vztahu můžeme přičíst výparu. Jelikož většinu výparu ovládá vegetace, vyplývá z toho, že snížení výparu a následně i srážky (a tedy i zvýšení odtoku) je způsobeno snížením výparu z vegetace.

Tento názor je podporován i výsledkem rozdělení srážek do třídních intervalů, kdy největší zvýšení odtoků je v měsících květnu až srpnu, a to při vyšších srážkách. Z toho plyne, že vegetace nezvládá vypařovat tak velká množství vody jako v létech 1931-60 v případě, že je vody nadbytek. Příčiny těchto jevů jsou pravděpodobně kombinované. A to jak poškození lesů ve vyšších polohách (horní část povodí Moravy), zmenšení retence (tento případ zřejmě je zvláště Drahanská vrchovina – povodí nad Plumlovem a nad Skalním Mlýnem ale i jinde), tak působení odvodnění, kdy rostliny vlhkých míst, které v zásadě neznají vodní deficit, jsou nahrazeny ornou půdou, která je část roku bez vegetačního krytu a podzemní voda se pro ně stane nedosažitelnou a odtéká drenáží do toku, aniž by se část přes vlhkomilné rostliny vypařila.

Předchozí rozbor lze shrnout následujícím způsobem: odtok se srážkou je svázán přes výpar, a to tak, že podle místních podmínek převod části vody ze srážek do odtoku vyvolá oslabení malého koloběhu, což se odrazí v poklesu srážek nebo obráceně, tj. zesílení malého koloběhu vyvolá zmenšení celkového odtoku. Důsledky tohoto jevu mohou být takové, že opětný výpar co největšího množství vody ze srážek s co možná minimálním celkovým odtokem je nutný k tomu, aby srážky mohly být transportovány co nejdále do nitra pevniny. Voda, která se řekami vrací předčasně do oceánu, ochuzuje o srážky oblasti, které jsou dále po směru proudění (v našem případě na východ). Proto zřejmě žádná rostlina, kromě xerofytů není stavěna na to, aby šetřila vodou, když je jí dostatek. Zřejmě pro globální rozvoj ekosystémů je nutné, aby bylo co nejvíce srážek na co největší ploše a rostliny, jež nejsou takto omezovány, jsou ve výhodě – vícekrát vodu použijí než se vrátí do oceánu. Pro tento koloběh má největší význam voda zachycená z velkých srážek ve velkých objemech. Stručně řečeno – na velikost srážek má vliv nejen množství vody, která přijde z oceánu přímo, ale i to, jestli vegetace v daných podmínkách je schopna zachytit a vypařit maximum ze srážek, vliv tedy má i půdní struktura, obsah humusu v půdě a dokonce i to, zda taková struktura vegetačního krytu je mezi zájmovým územím a oceánem proti směru obvyklé cirkulace (tedy jestli voda tam spadlá zbytečně neodtéká do oceánu bez výparu).

Literatura

[1] L. Budík, M. Budíková: Statistické zpracování měsíčních a ročních srážkových a odtokových charakteristik povodí řeky Moravy. Nakladatelství Českého hydrometerologického ústavu, 28. sešit edice Práce a studie, Praha 2001.

Adresy autorů

Ladislav Budík
ČMHÚ
Kroftova 43
600 00 Brno
e-mail: budik@chmi.cz

Marie Budíková
KAM PřF MÚ
Janáčkovo nám. 2a
600 00 Brno
e-mail: budikova@math.muni.cz

STATISTICKÁ ANALÝZA DAT OBJEKTŮ PODZEMNÍCH VOD

V.Sosna[1]

1. Úvod

V roce 1999 a v první polovině roku 2000 bylo provedeno základní statistické zpracování dat objektů podzemních vod na území Čech, které spravuje Český hydrometeorologický ústav (dále jen ČHMÚ). Výpočty provedl autor tohoto příspěvku pod metodickým vedením vedoucího Oddělení podzemních vod ČHMÚ RNDr.J.Kessla. Na přípravě a úpravě dat se podíleli pracovníci všech poboček ČHMÚ.

Cílem zpracování bylo získání souborů základních statistik dat podzemních vod na území České republiky. Důraz byl kladen na zjištění případných nehomogenit v řadách, na detekci existence dlouhodobých trendů, na studium dynamických charakteristik časových řad a extrémů. Výpočty byly provedeny pro standardní dvacetiletí 1971 - 1990, některé statistiky i pro třicetiletí 1969 - 1998.

2. Postup zpracování statistik

Zpracována byla data vydatnosti pramenů měřená v litrech za sekundu a hladin vody ve vrtech měřená v metrech nadmořské výšky. Využita byla měření pouze ze sítě mělkých vrtů, nikoliv hlubinných objektů. Měření vydatnosti pramenů a hladin vrtů bylo prováděno s týdenním krokem, zpravidla ve středu.

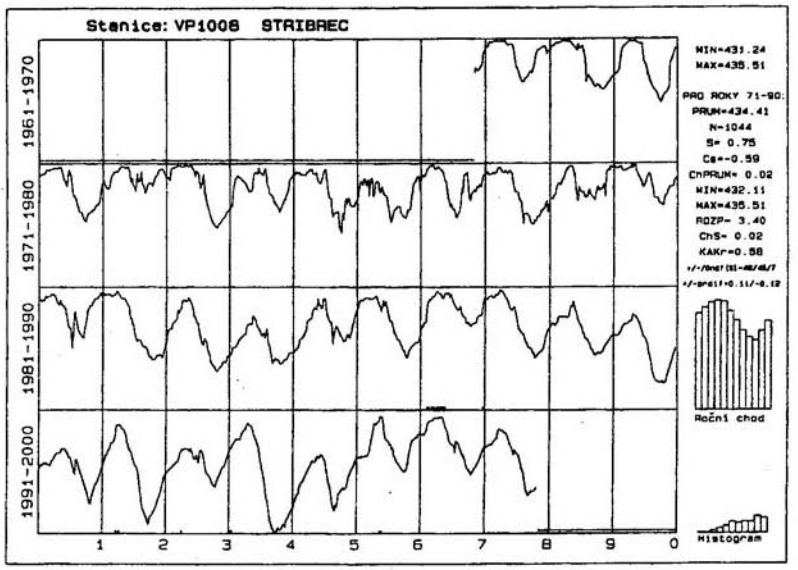
Před vlastním výpočtem statistik pracovníci poboček ČHMÚ data opravili, upravili a doplnili pomocí programu, který byl napsán pro tento účel [1]. Vznikly tak časové řady týdenních měření, které byly vytvořeny účelově pro statistické zpracování dat. Provedeny byly především tyto zásahy do dat: Doplnění chybějících dat při krátkodobém výpadku měření lineární regresí (metodou středního řešení) s pomocí jiného blízkého objektu podzemních vod a očištění řad od hodnot ovlivněných lidskou činností (např. čerpací zkoušky u vrtů). Dále byla provedena radikální selekce dat. Pro další výpočty byly použity pouze řady s kompletním nebo alespoň doplněným měřením ve standardním dvacetiletí 1971–1990. Byly vyřazeny objekty s evidentně nehomogenním měřením. Při posuzování homogenity byly používány grafy, jejichž ukázky jsou na obr.1 až obr.3. V případě pochybnosti při zjišťování nehomogenit byla použita metoda dvojné součtové čáry nebo byly využity testy změny pravděpodobnostního rozdělení, střední hodnoty, rozptylu nebo trendu, obsažené v programu CTPA [2]. Po provedení selekce bylo k dalšímu zpracování použito celkem 420 řad dat hladin mělkých vrtů a 75 řad dat vydatností pramenů.

Pro hromadný výpočet statistik a kreslení grafů byly v prostředí programovacích jazyků TurboPascal a Delphi vyvinuty programy, které dávkovým způsobem zpracovávají data vrtů a pramenů. Z týdenních měření byly vypočítány řady průměrů hydrologických roků vydatností pramenů a hladin vrtů, které byly použity pro výpočet parametrů a grafů dlouhodobých trendů. Dále byly vypočítány řady ročních extrémů, které byly využity pro výpočet N-letostí ročních extrémů na základě 3-parametrického logaritmicko-normálního rozdělení. Metoda použitá pro odhad parametrů LN3 rozdělení je popsána v článku [3]. Některé statistiky byly zpracovány do map pomocí programu ArcView GIS.

[1]Český hydrometeorologický ústav, pob. Plzeň, Denisovo nábř.14, 301 50 Plzeň, sosna@chmi.cz

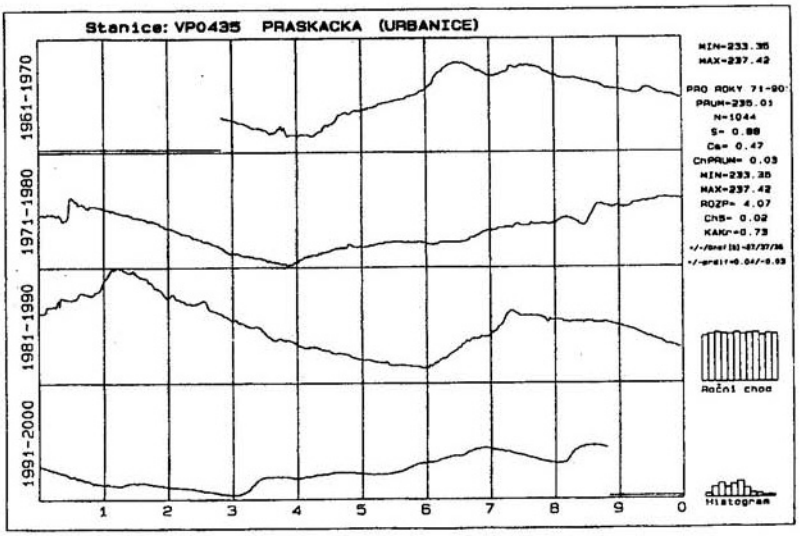
3. Základní statistiky a grafy

Pro každý objekt podzemních vod byl vytištěn přehledný obrázek, který obsahuje graf měřené veličiny v závislosti na čase, nejdůležitější číselné statistiky, schéma ročního chodu a histogramu. Ukázky jsou na obr.1 až obr.3. Celý obrázek je nadepsán číslem objektu a názvem stanice. Vlastní graf je rozdělen na 4 části (pruhy) po desetiletích 61-70, 71-80, 81-90, 91-00. Desetiletí jsou uvedena svisle vlevo. Ve spodní části jsou nadepsány jednotlivé roky, poslední číslice roku je umístěna vpravo od příslušného pruhu pod svislou dělící čarou. Pro nedostatek místa není na grafu uveden název veličiny, měřítko ani jednotky. U pramenů se vždy jedná o vydatnost v litrech za sekundu a u vrtů o nadmořskou výšku hladiny podzemní vody v metrech. Měřítko grafu je určeno pomocí variačního rozpětí všech dostupných dat v rozmezí let 1961-2000. Spodní hranice pruhu desetiletí odpovídá absolutnímu minimu ze všech dat, horní hranice pak maximu. Hodnoty absolutního minima a maxima jsou číselně uvedeny v pravém horním rohu obrázku. Neměřené období (výpadek v měření) je znázorněno vodorovnou čarou v blízkosti dolního okraje příslušného pruhu desetiletí. Pod nadpisem PRO ROKY 71-90 jsou uvedeny číselné statistiky, které jsou vypočítány jen pro standardní období měření v letech 1971 až 1990.

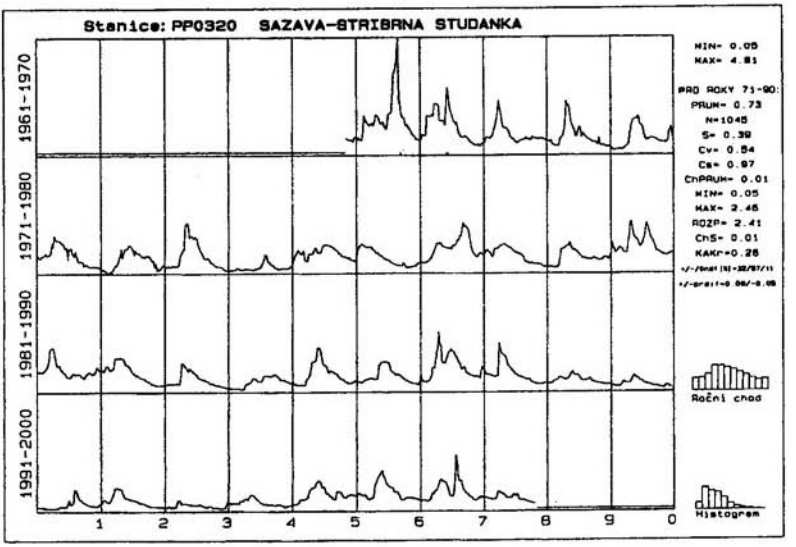


Obr. 1. Ukázka grafu hladiny vrtu (v m n.m.)

Základní statistiky typu průměru nebo směrodatné odchylky nemají význam při porovnávání vlastností objektů podzemních vod mezi sebou v ploše, jsou příliš závislé na místních podmínkách a nemá smysl je vynášet do map. Výjimku do jisté míry tvoří koeficient variace Cv u pramenů. Lze ho využít ke třídění objektů pramenů do kategorií podle kolísání vydatnosti. U námi zpracovaných pramenů kolísala hodnota od 0,05 do 1,27 a průměr byl 0,5.



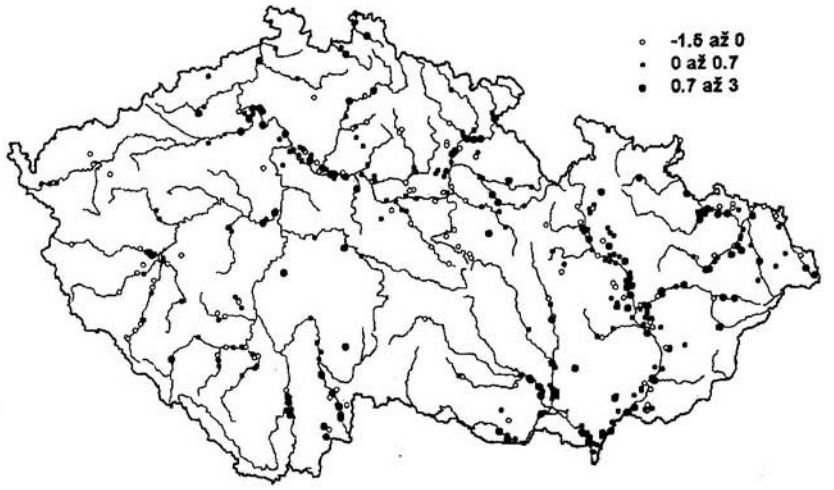
Obr. 2. Ukázka grafu hladiny vrtu (v m n.m.)



Obr. 3. Ukázka grafu vydatnosti pramene (v l/s)

Objekty podzemních vod lze také třídit podle typu ročního chodu. Na obrázku č.1 a č.3 jsou zobrazeny grafy s výrazným ročním chodem, u objektu na obrázku č.2 naopak roční chod chybí. Další zajímavou statistikou je koeficient šikmosti Cs, jehož hodnota úzce souvisí s tvarem histogramu. Histogram je důležitým grafem, který umožňuje názorně posoudit typ pravděpodobnostního rozdělení, kterým se řídí naměřená data (zde pro týdenní měření). V pravém dolním rohu obr.1 je histogram se záporným zešikmením, na obr.3 je ukázka kladného zešikmení. Obr.2 obsahuje histogram dat, která nemají unimodální rozdělení (histogram má dvě výrazná maxima). U pramenů se záporné zešikmení prakticky nevyskytuje. Cs se u vydatnosti pramenů pohybuje od hodnot blízkých nule až do maximální hodnoty okolo 5. Průměrná hodnota v oblasti ČR je 1,56.

U hladin vrtů se záporná šikmost vyskytuje v třetině případů, nejmenší hodnoty se pohybují okolo -1. Největší kladné hodnoty šikmosti nepřesahují 2. Průměrná hodnota je 0,3. Po zobrazení v mapě je patrné, že oblast se zápornou šikmostí hladin vrtů nebo hodnotou blízkou nule se nachází především na jihozápadě Čech (v horním povodí Ohře, v povodí horních přítoků Berounky a v povodí Otavy), dále v povodí Doubravy a Chrudimky a v povodí Labe v blízkosti Pardubic a Hradce Králové. Velké kladné hodnoty šikmosti nalezneme v povodí horního toku Vltavy a Lužnice, dále v mezipovodí Labe v úseku od Nymburka po soutok s Ohří a především se často vyskytují na celém území Moravy. Mapa koeficientu šikmosti hladin vrtů je na obrázku č.4.



Obr. 4. Mapa koeficientu šikmosti týdenního měření hladin vrtů pro roky 1971-90

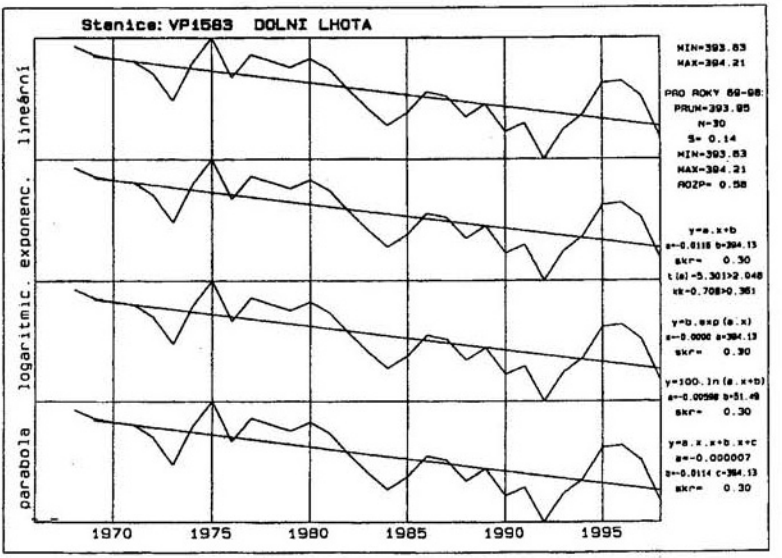
4. Diference

Pod pojmem diference se rozumí rozdíl dvou po sobě časově následujících hodnot hladin vrtů nebo vydatností pramenů (zde s týdenním krokem). Odečítá se vždy předcházející hodnota od následující. Diference je tedy kladná v případě, kdy měřená veličina roste, záporná když klesá. Diference se zavádějí pro popis dynamického chování časových řad.

Velikost kladné diference u podzemních vod charakterizuje rychlost plnění prostředí v okolí měřeného objektu vodou, záporná pak rychlost vyprazdňování. Pro přiřazování objektů do jednotlivých hydrogeologických rajonů mají větší význam poměrné statistiky, absolutní hodnoty (tj.průměrná kladná nebo záporná diference) mají význam pouze pro místní popis. Užitečná relativní statistika je poměr průměrné kladné ku záporné diferenci. Udává kolikrát větší je rychlost plnění prostředí vodou, než je rychlost vyprazdňování. Pro námi sledovanou oblast je průměrná hodnota tohoto poměru pro vydatnost pramenů i hladiny vrtů přibližně stejná a rovná číslu 1,5 . Hodnota pro konkrétní objekt je vždy větší než 1.

5. Trendy

Pro studium dlouhodobějších trendů v časových řadách dat podzemních vod byl pro každý objekt sestrojen graf průměrů hydrologických roků vydatnosti pramenů a hladin vrtů v závislosti na čase. Zároveň byla vypočítána lineární, exponenciální, logaritmická a kvadratická regrese. Parametry regresí byly odhadovány pomocí metody nejmenších čtverců. Bylo zjištěno, že se průběh exponenciální, logaritmické ani kvadratické (parabolické) regrese podstatně neliší od lineární pro dvacetiletí 1971-1990, ani pro třicetiletí 1969-1998. Pro tato období lineární regrese dostatečně dobře charakterizuje dlouhodobý trend. Ukázka grafu trendů je na obr.5.

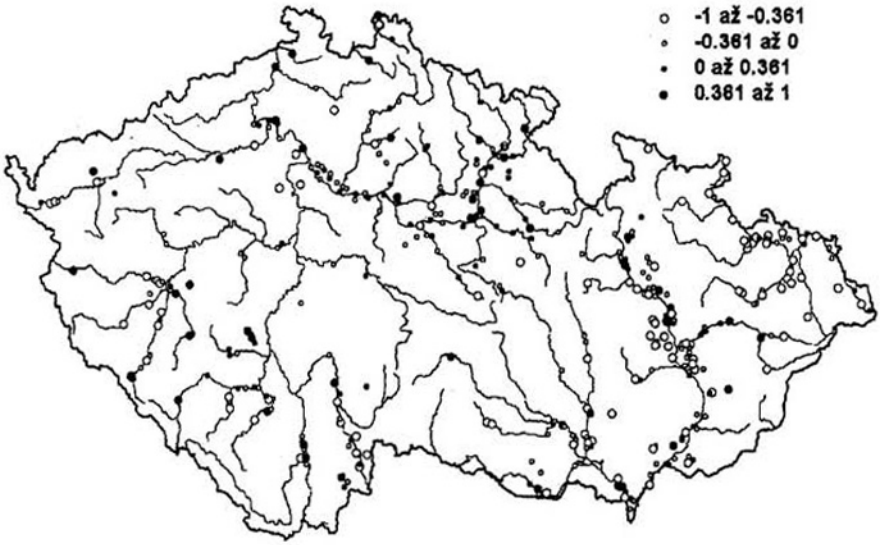


Obr. 5. Ukázka grafu trendů ročních průměrných stavů vrtu (v m n.m.)

Kromě grafů byly trendy zakresleny i do map, v podobě koeficientů korelace hladin vrtů a vydatností pramenů s časem. Při volbě intervalů v legendách map byly použity meze, při kterých se s 95%-ní jistotou zamítá hypotéza nulovosti koeficientu korelace. Tyto meze ovšem platí jen pro normální a nezávislá data. Oboje v našem případě není splněno. K objektům, které jsou v mapách zvýrazněny, nelze tedy mechanicky přiřazovat statisticky významný trend, zvláště pokud je takový bod

osamocený. V těchto souvislostech je také třeba chápat termín „výrazný trend", který je dále použitý v textu.

Na obrázku č.6 je znázorněna mapa koeficientu korelace průměrných ročních hladin vrtů s časem pro třicetiletí 1969-98. Plným velkým kolečkem jsou vyznačeny objekty s významným rostoucím trendem hladin, prázdným velkým kolečkem pak s významným klesajícím trendem hladin. Na mapě je patrný značný rozdíl v rozložení hodnot mezi Čechami a Moravou. V rámci Čech potom méně významný rozdíl mezi JZ území a SV. Pro třicetiletí 1969-98 byl na území Moravy zjištěn výrazný rostoucí trend jen u 8% případů, ale výrazný klesající trend u 37% vrtů. Za stejné období to bylo v oblasti Čech 15% vrtů s výrazným rostoucím trendem a 18% s výrazným klesajícím trendem.



Obr. 6. Mapa korelace ročních průměrů hladin vrtů s časem pro třicetiletí 1969-98

Literatura

[1] Sosna,V.: Návod k programu PODZVOD. ČHMÚ, Plzeň, 1997
[2] Procházka,M., Deyl,M.: Identifikace změny v časové řadě II. VH-TRES, České Budějovice, 1999
[3] Hostýnek,J., Lepka,Z., Sosna,V.: Zpracování N-letých ročních a měsíčních maxim denních úhrnů srážek v západních Čechách. Meteorologické zprávy 52/3, 1999, 73-77

Jackknife odhady vychýlení a standardní chyby

Bohdan Linda, Ústav matematiky, Univerzita Pardubice

Metoda Jackknife nalézá své uplatnění především při odhadování vychýlení a standardní chyby odhadů parametrů základního souboru. Používáme ji tehdy, když přesné určení těchto veličin je buď příliš složité, anebo dokonce nemožné.

Označme $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ náhodný výběr ze základního souboru X. Z tohoto náhodného výběru $\mathbf{X}$ vytvoříme postupně n nových náhodných výběrů $\mathbf{X}_i$ tak, že z původního výběru vynecháme vždy jednu hodnotu:

$$\mathbf{X}_i = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

Takto vytvořené výběry se nazývají Jackknife výběry.

Nechť dále Θ je nějaký parametr základního souboru X a nechť

$$T = q(\mathbf{X})$$

je jeho odhad. Výraz

$$T_i = q(\mathbf{X}_i) \quad i = 1, 2, \ldots, n$$

se nazývá i – tá jackknife replikace odhadu T.

Označme dále

$$\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$$

Jackknife odhad S(T) standardní chyby s(T) je definován jako statistika

$$S(T) = [(n-1)n^{-1}\sum (T_i - \overline{T})^2]^{\frac{1}{2}}$$

Jackknife metoda při odhadu standardní chyby s(T) odhadu T vychází z toho, že se na všechny jackknife replikace nedívá jako na jednotlivé naměřené hodnoty x_i, ale jako na n konkrétních výběrů $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Základem odhadu jsou potom rozdíly

$$T_i - \overline{T} \qquad i = 1, 2, \ldots, n$$

Koeficient (n-1) se nazývá „inflační faktor" a zavádí se proto, protože jackknife odchylky $(T_i - \overline{T})^2$ mají tendenci býti menší než odchylky při jiných odhadech. Přesný tvar inflačního faktoru byl odvozen z toho, abychom v případě $T = \overline{x}$ dostali nevychýlený odhad jeho standardní chyby.

Základem odhadu vychýlení b(T) jsou tzv. jackknife odchylky

$$T_i - \overline{T} \qquad i = 1, 2, \ldots, n$$

nazývané též někdy jackknife vlivové hodnoty. Vychýlení potom odhadujeme aritmetickým průměrem těchto hodnot

$$\frac{1}{n}\sum(T_i - T) = \overline{T} - T$$

Tato statistika však podceňuje skutečné vychýlení a proto odhad B(T) vychýlení b(T) získáme vynásobením tohoto výrazu inflačním faktorem n-1:

$$B(T) = (n-1)(\overline{T} - T)$$

M. Quenouille v r. 1949 v článku „Approximate tests of correlation in time series" poprvé uvedl myšlenku, provést odhad vychýlení odhadu na základě replikací náhodného výběru, které vzniknou z původního výběru vynecháním jednoho měření. J. W. Tukey v článku „Bias and confidence in not quite large samples" zkoumal možnosti této metody pro odhad standardní chyby a dal jí název Jackknife. Dalším rozvojem této metody se zabývali především R.G.Miller, H.L.Gray , W.R.Schucany, D.V.Hinkley, B.C.Wei, J.Shao a C.F.J.Wu.

Literatura:

[1] Efron,B.: Jackknife after Bootstrap standard errors and influence functions. Jour. Royal. Statist. Soc. B 54, 1992

[2] Gray,H.L., Schucany,W.R.: The Generalized Jacknife Statistics, Marcel Dekker, New York, 1972

[3] Miller,R.G.: The Jackknife – a Review. Biometrika. Vol. 61,1974

[4] Shao,J., Wu,C.F.J.: A General Theory for Jackknife variance estimation. Ann. Statist. Vol 17, 1989

[5] Thisted, R.A.: Elements of Statistical Computing. Chapman and Hall, London, 1986

DELTA METODA

Jana Kubanová, Bohdan Linda
Ústav matematiky, Fakulta ekonomicko správní,
Univerzita Pardubice

Předpokládáme, že T je odhad nějakého parametru Θ základního souboru X. Často bývá dosti obtížné zjistit rozdělení pravděpodobností tohoto odhadu T a vypočítat klasickým způsobem jeho charakteristiky.

V některých případech lze k odhadu parametrů s výhodou použít tzv. delta metodu, která obvykle nebývá popsána v učebnicích statistiky. Tato metoda je poměrně jednoduchá a její princip je založen na Taylorově rozvoji. Zpravidla se využívá pouze první člen, v některých případech první dva členy tohoto rozvoje a zanedbává se zbytek, který se s rostoucím rozsahem výběru poměrně rychle blíží nule. Delta metoda se využívá především u takových odhadů, jejichž formální vyjádření je ve tvaru součtu náhodných veličin. Rozdělení pravděpodobností takovýchto odhadů lze na základě centrální limitní věty aproximovat normálním rozdělením pravděpodobností

V parametrických analýzách je často odhad T nějakou funkcí základních statistik typu $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$, kde X_i i =1,2,...,n tvoří náhodný výběr z X. Těmito statistikami mohou být výběrové momenty $U_1,\ldots,U_m$. K získání aproximace některých charakteristik (především rozptylu) T lze pak s výhodou využít tzv. delta metody.

Předpokládejme tedy, že T je odhad parametru Θ nějaké jednorozměrné náhodné veličiny X. Dále předpokládejme, že tento odhad je funkcí skalární statistiky U z náhodného výběru o rozsahu n. Formálně potom píšeme

$$T = g(U).$$

Předpokládáme dále, že rozložení pravděpodobností náhodné veličiny U je přibližně normální, tzn.

$$U \doteq N(\mu, n^{-1}\sigma^2).$$

Potom můžeme veličinu U vyjádřit vztahem

$$U = \mu + n^{-0,5}\sigma Z + o_p(n^{-1}) \qquad (1)$$

kde náhodná veličina Z má N(0,1) rozdělení pravděpodobnosti.

Jestliže g je hladká funkce taková, že $g'(\mu) \neq 0$, potom na základě Taylorova rozvoje můžeme psát

$$T = g(U) = g(\mu) + (U - \mu)g'(\mu) + o_p(n^{-1/2}) \qquad (2)$$

Dosadíme-li do (2) za U výraz z (1) dostáváme

$$T = g(\mu) + n^{-0,5}g'(\mu)\sigma Z + o_p(n^{-1/2})$$

Náhodná veličina T = g(U) má přibližně normální rozložení.

Pro střední hodnotu náhodné veličiny T platí

$$E(T) = E(g(\mu) + n^{-0,5}g'(\mu)\sigma Z + o_p(n^{-1/2})) \approx g(\mu)$$

a podobně pro rozptyl náhodné veličiny T platí

$$D(T) = D(g(\mu) + n^{-0,5}g'(\mu)\sigma Z + o_p(n^{-1/2})) \approx (g'(\mu))^2\sigma^2$$

Tyto odhady se nazývají delta aproximace.

Můžeme shrnout, že rozdělení pravděpodobností náhodné veličiny T je přibližně normální s uvedenými parametry:

$$T \sim N(g(\mu), n^{-1}\{g'(\mu)\}^2\sigma^2)$$

Literatura:

[1] Efron,B.: The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference in Applied Mathematics, Philadelphia:SIAM 1982

[2] Fernholz,L.T.:Von Mises Calculus for Statistical Functionals, Lecture Notes in Statistics, Springer, NewYork 1983

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Předseda společnosti: Doc. RNDr. Jaromír Antoch, CSc., KPMS MFF UK Praha, Sokolovská 83, 186 75 Praha 8, e-mail: jaromir.antoch@karlin.mff.cuni.cz.
ISSN 1210 – 8022
Redakce: Doc. RNDr. Gejza Dohnal, CSc., Jeronýmova 7, 130 00 Praha 3, e-mail: dohnal@fsik.cvut.cz