O „TELESKOPICKÝCH“ ROZDĚLENÍCH PRAVDĚPODOBNOSTI, JEJICH DVOUROZMĚRNÉ VERZI I VYUŽITÍ

ON “TELESCOPIC” DISTRIBUTIONS OF PROBABILITY, THEIR BIVARIATE VERSIONS, AND APPLICATION

Petr Volf

Adresa: Institute of Information Theory and Automation AS CR, Pod vodárenskou věží 4, 182 08 Praha 8

E-mail: volf@utia.cas.cz

Abstrakt: V tomto příspěvku je popsána jedna zajímavá skupina diskrétních rozdělení pravděpodobnosti, která vznikla diskretizací rozdělení spojitých. Říká se jim také teleskopická rozdělení, což je inspirováno pojmem „teleskopická posloupnost“ v matematice. Po úvodních poznámkách se budu hlavně zabývat konstrukcí dvourozměrných teleskopických rozdělení vzniklých diskretizací dvourozměrných distribucí spojitých (konstruovaných zpravida pomocí vhodné kopuly). Na příkladě pak ukážu výhody a možnosti využití těchto modelů, mezi jinými i „teleskopické“ verze dvourozměrného negativně binomického rozdělení.

Klíčová slova: Diskretizace spojitých rozdělení pravděpodobnosti, teleskopické rozdělení, dvourozměrné rozdělení, kopula, negativně binomické rozdělení.

Abstract: In this paper, an interesting group of discrete probability distributions is described, which arose from the discretization of continuous distributions. They are called also telescopic distributions, which is inspired by the term “telescoping sequence” in mathematics. After the initial remarks, we will mainly deal with the construction of bivariate telescopic distributions resulting from a discretization of bivariate continuous ones (constructed via a suitable copula). We will show the advantages and possibilities of use of these models on an real data example. Among others, also a telescopic form of bivariate negative binomial distribution will be presented.

Keywords: Discretization of continuous probability distribution, telescopic distribution, bivariate distribution, copula, bivariate negative binomial model.

1. Úvod

Na pojem „teleskopické rozdělení pravděpodobnosti“ jsem narazil při recenzi jednoho článku (který zde neuvádím, protože nebyl dobrý). Na Wikipedii jsem našel k tomuto pojmu jen heslo „telescoping series“ [13]. Termín teleskopické rozdělení pravděpodobnosti používá (zatím) jen docela malá skupina statistiků. Zřejmě jeden ze základních článků o těchto diskrétních rozděleních pravděpodobnosti je [11]. Tam nabízená definice je následující:

Definice 1 (telekopické rozdělení pravděpodobnosti): Řekneme, že náhodná diskrétní nezáporná veličina X má teleskopické rozdělení, pokud se její pravděpodobnosti dají napsat ve tvaru

$$P(X = K) = q^{z(K;\theta)} - q^{z(K+1;\theta)}, \quad \text{pro } K = 0, 1, 2, ..., \tag{1}$$

kde $q \in (0, 1)$ a funkce $z(x; \theta)$ je pro každé θ (což je parametr rozdělení) ryze rostoucí v x, $z(0; \theta) = 0$, $\lim_{x\to\infty} z(x; \theta) = \infty$.

Tato formulace zahrnuje všechna nezáporná diskrétní rozdělení, pro která $P(X = K) > 0$ pro každé $K = 0, 1, 2, ...$ Označím $\overline{F}(x) = P(X \geq x)$ funkci přežití takovéhoto rozdělení (modifikovanou, neboť nerovnost je neostrá). Pak $\overline{F}(0) = 1$, $\lim_{x\to\infty} \overline{F}(x) = 0$, a pro $K = 0, 1, 2, ...$ platí

$$P(X = K) = \overline{F}(K) - \overline{F}(K + 1).$$

Pro libovolné $q \in (0, 1)$ definuju $z(x) = \ln(\overline{F}(x))/\ln(q)$. Po úpravě dostanu

$$q^{z(x)} = \mathrm{e}^{\ln(q)z(x)} = \mathrm{e}^{\ln(\overline{F}(x))} = \overline{F}(x),$$

a tedy

$$P(X = K) = q^{z(K)} - q^{z(K+1)}$$

(pro jednoduchost zápisu jsem vynechal parametr θ).

Protože se zde používají jen hodnoty distribuční funkce v bodech K, stejný výsledek dostaneme i pomocí distribuční funkce se stejnými hodnotami v nich, ale neklesající, spojité a integrovatelné mezi nimi. Tím se vlastně dostaneme k pojmu, který je již používán častěji, a to „diskretizace spojitého rozdělení pravděpodobnosti“.

Definice 2 (diskretizované spojité rozdělení pravděpodobnosti): Nechť Y je spojitá nezáporná náhodná veličina s distribuční funkcí $F(y)$, pak

příslušná diskretizovaná náhodná veličina X má rozdělení s pravděpodobnostmi

$$P(X = K) = F(K + 1) - F(K) = \overline{F}(K) - \overline{F}(K + 1), \tag{2}$$

kde $\overline{F}(y) = 1 - F(y)$ je funkce přežití veličiny Y.

Uvedené definice vlastně neříkají nic nového, ale zahrnují i ty skutečně zajímavé a užitečné (jak se ukáže) případy, kdy je diskrétní rozdělení odvozené z nějakého známého spojitého rozdělení, které má kompaktní tvar distribuční funkce.

Veličina Y ani nemusí být omezená na kladnou poloosu, stačí v Definici 2 položit $\overline{F}(0) = 1$, případně pravděpodobnosti vynormovat, je-li pro původní veličinu $P(Y < 0) > 0$, aby $\sum_{K=0}^{\infty} P(X = K) = 1$. Obdobně se diskretizují i spojitá rozdělení definovaná na celé reálné ose, např. v [12] je použito diskretizované normální rozdělení k odhadu dosažených tolerancí v analýze spolehlivosti.

V tomto příspěvku nejdříve okomentuji některé vlastnosti těchto rozdělení. Pak ukážu postup, jak obdobným způsobem – diskretizací dvourozměrných spojitých rozdělení pravděpodobnosti – konstruovat dvourozměrná diskrétní rozdělení. Ta jednak rozšiřují a zlepšují možnosti modelování dvourozměrných diskrétních dat, ale také zjednodušují práci s nimi. To bude ukázáno na jednoduchém příkladu s reálnými daty.

2. Poznámky k vlastnostem teleskopických rozdělení

Roknabadi et al. [11] se věnují hlavně otázce porovnání vlastností původních spojitých a z nich odvozených diskrétních rozdělení, a to při aplikaci v modelování a analýze spolehlivosti. Pohled tohoto příspěvku je širší, obecnější, cílem je metodou diskretizace obohatit nabídku diskrétních rozdělení pravděpodobnosti, a tedy i zvýšit flexibilitu modelování. V následujících poznámkách připomenu, že i některá standardní diskrétní rozdělení splňují Definici 2 pro nějakou funkci přežití, a tedy se s nimi dá pracovat jako s diskretizovanými spojitými distribucemi. To je, jak ukážu, výhodné dále při konstrukci jejich dvourozměrných verzí.

1. Termín „teleskopický“ (∼vysouvací, nastavovací, třeba typ dalekohledu nebo žebříku) se zjevně vztahuje k vlastnosti částečných součtů v posloupnosti pravděpodobností (1) či (2), a to sice že pro $M \leq N$ je

$$\sum_{k=M}^{N} P(X = k) = q^{z(M;\theta)} - q^{z(N+1;\theta)} = \overline{F}(M) - \overline{F}(N + 1).$$

Ještě zmíním jednu zajímavou vlastnost. Je snadné ukázat, že

$$EX = \sum_{K=0}^{\infty} K \cdot (\overline{F}(K) - \overline{F}(K+1)) = \sum_{K=1}^{\infty} \overline{F}(K) \leq \int_0^{\infty} \overline{F}(y)\,\mathrm{d}y = EY,$$

pokud je EY konečná. Pak poslední rovnost plyne z věty 1.5 v [1].

2. Jak již bylo řečeno, myšlenka diskretizace spojitého rozdělení jako v (2) zdaleka není nová. Je snadné si představit případ, kdy sice sledujeme realizaci spojité veličiny Y, ale známe jen počty hodnot v daných intervalech. Pak samozřejmě použijeme obdobu vztahu (2) a z relativních četností dostaneme odhad distribuční funkce F v hranicích intervalů. Příkladem může být odhad parametrů Weibullova rozdělení, vlastně jde o klasický způsob ještě z „před-počítačových“ dob. Uvedu zde k tomu příklad:

   Nechť z N výrobků uvedených do provozu v čase $t = 0$ je v časech $0 < t_1 < t_2 < ... < t_K$ v provozu ještě $N > N_1 \geq N_2 \geq ... \geq N_K$ výrobků. Předpokládejme, že Weibullovo rozdělení s funkcí přežití $\overline{F}(t) = \exp(-\alpha \cdot t^{\beta})$ je vhodným modelem pro dobu životnosti těchto výrobků. A tak porovnáme tuto modelovou funkci přežití s empirickou v bodech t_j, tj. $\exp(-\alpha \cdot t_j^{\beta}) \sim N_j/N = r_j$. Po dvojím logaritmování obou stran dostaneme lineární vztah

$$\ln(\alpha) + \beta \cdot \ln(t_j) \sim \ln(-\ln(r_j)), \quad j = 1, 2, ..., K.$$

   Z něj odhadneme parametry $\ln(\alpha)$ a β metodou nejmenších čtverců.

3. Pomocí (2) dostáváme velkou třídu diskrétních rozdělení pravděpodobnosti, v podstatě dost jednoduchých (tak jednoduchých, jako spojitá distribuce, ze které to diskrétní rozdělení odvodíme). Tím je tedy hodně obohacen náš výběr parametrických modelů pro diskrétní data. A tyto modely mají jednu výhodu, jsou vlastně zadány pomocí svojí distribuční funkce, zatímco u standardně používaných diskrétních rozdělení je zpravidla distribuční funkce dána prostě částečným součtem pravděpodobností, a to není příliš pohodlné. Ale je známo, že i mnohá standardní diskrétní rozdělení mohou být zároveň vyjádřena ve tvaru (2).

4. Je snadné ukázat, že geometrické rozdělení je teleskopického typu. Dostaneme je pomocí (2) z rozdělení exponenciálního:

   $P(X = K) = \exp(-\lambda K) - \exp(-\lambda(K+1)) = \exp(-\lambda K)(1 - \exp(-\lambda)) = (1-p)^K p$,

kde $p = 1 - \exp(-\lambda)$ je parametr odpovídající pravděpodobnosti úspěchu. Tady je tedy něco jako „styčný bod“ klasických a teleskopických rozdělení pravděpodobnosti.

5. I Poissonovo rozdělení může být vyjádřeno tímto „teleskopickým“ způsobem: Mějme homogenní Poissonův proces s parametrem λ, tj. koncové body postupně načítaných nezávislých exponenciálních náhodných veličin s tímto parametrem. Pak veličina $X =$ počet těchto bodů v intervalu $(0, 1)$ má Poissonovo rozdělení, zase s tímtéž λ. Zároveň veličina $T_K =$délka součtu K oněch exponenciálních veličin má gamma rozdělení $Gamma(K, \lambda)$, K je parametr tvaru, λ intenzity ($1/\lambda$ je parametr měřítka, gamma rozdělení takto vzniklému se také říká Erlangovo). Hustota tohoto rozdělení je $f_K(z) = \lambda^K z^{K-1} \mathrm{e}^{-\lambda z}/\Gamma(K)$, distribuční funkce $F_K(z) = \gamma(K, \lambda z)/\Gamma(K)$, kde $\Gamma(K) = \int_0^\infty t\mathrm{e}^{-t}\,\mathrm{d}t$ je úplná gamma funkce (připomeňme, že $\Gamma(K) = (K-1)!\ \forall K \in \mathbb{Z}$) a $\gamma(a, b) = \int_0^b t^{a-1}\mathrm{e}^{-t}\,\mathrm{d}t$ je neúplná (dolní) gamma funkce.

   Takže pro naši Poissonovu veličinu dostaneme, že

$$P(X = K) = P(T_K < 1) - P(T_{K+1} < 1) = \frac{\gamma(K, \lambda)}{\Gamma(K)} - \frac{\gamma(K+1, \lambda)}{\Gamma(K+1)}. \quad (3)$$

   Funkce $S(t; \lambda) = \gamma(t, \lambda)/\Gamma(t)$ je takzvaná regularizovaná (dolní) neúplná gamma funkce. Pro každé $\lambda > 0$ je jako funkce $t > 0$ nerostoucí, s $\lim_{t\to 0+} S(t) = 1$ a $\lim_{t\to\infty} S(t) = 0$, čili je to funkce přežití nějaké spojité náhodné veličiny. Takže (3) odpovídá definici teleskopického typu rozdělení pravděpodobnosti.

6. Důsledkem je, že i negativně binomické rozdělení je tohoto typu, protože toto rozdělení dostaneme mimo jiné i z rozdělení Poissonova, pokud předpokládáme, že jeho parametr λ je náhodný s gamma rozdělením (takže vlastně bayesovsky, nebo se tomu také říká compound, mixed, contagious Poisson distribution). Konkrétně, nechť λ Poissonovy veličiny X má apriorní $Gamma(r, p/(1-p))$ rozdělení (jeho hustotu označme $g(\lambda; r, p)$), pak výsledkem je náhodná veličina (označme ji Z) s negativně binomickým rozdělením s parametry r a p, a je:

$$\begin{aligned} P(Z = K; r, p) &= \int_0^\infty P(X = K; \lambda) g(\lambda; r, p)\,\mathrm{d}\lambda = \\ &= \int_0^\infty S(K; \lambda) g(\lambda; r, p)\,\mathrm{d}\lambda - \int_0^\infty S(K+1; \lambda) g(\lambda; r, p)\,\mathrm{d}\lambda = \\ &= I_{1-p}(K, r) - I_{1-p}(K+1, r), \end{aligned}$$

kde $I_x(a,b) = B(x;a,b)/B(a,b)$ je tzv. regularizovaná neúplná beta funkce, tedy $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$ je neúplná beta funkce a $B(a,b)$ je úplná beta funkce.

7. Podobným způsobem může být vyjádřeno i binomické rozdělení, jeho distribuční funkci lze opět napsat s pomocí regularizované beta funkce,

$$F(k) = P(X \leq k) = I_{1-p}(N-k, k+1).$$

Takto tedy můžeme vyjádřit distribuční funkce některých diskrétních rozdělení v kompaktním tvaru. Tyto zápisy můžeme najít i ve Wikipedii. Ač to nejsou příliš jednoduché funkce, tento tvar umožní s těmito distribucemi zacházet jako s diskretizovanými spojitými, a jak ukážu v následující části, vytvářet z nich s pomocí vhodně zvolených kopul různé varianty dvourozměrných rozdělení.

Ještě se nabízí i otázka, jestli takto diskretizovaným spojitým rozdělením říkat „teleskopická“, tj. například T-Weibull, či jenom D-Weibull jako „diskretizované“ Weibullovo rozdělení. Ale šikovný název je součástí marketingu, to jsme i v metodách analýzy dat už několikrát viděli (pojmy jako LASSO, MARS, i třeba genetický algoritmus, neuronová síť). I proto se asi ona určitá skupina badatelů toho termínu drží.

3. Dvourozměrná teleskopická rozdělení

Dá se říci, že teprve při práci s dvourozměrnými diskretními daty vynikne další výhoda těchto modelů. Konstrukce dvourozměrných verzí klasických diskrétních rozdělení je zpravidla dost nešikovná. Snadné je takto rozšířit Bernoulliho „0–1“ rozdělení, případně z něj zkonstruovat dvourozměrné geometrické rozdělení [3, 8], u dalších jde už o problém. Ještě je populární určitá verze Poissonova rozdělení, např. [5, 6]:

Nechť X_1, X_2, X_3 jsou vzájemně nezávislé náhodné veličiny s Poissonovým rozdělením s parametry $\lambda_1, \lambda_2, \lambda_3$, pak $Y = X_1 + X_3$ a $Z = X_2 + X_3$ tvoří dvourozměrné Poissonovo rozdělení.

Jejich korelace je $\lambda_3/\sqrt{(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}$ (tj. kladná a vázaná na hodnoty parametrů, to je tedy omezení této konstrukce), marginální rozdělení jsou Poissonova s parametry $\lambda_1+\lambda_3$ a $\lambda_2+\lambda_3$, jejich společná pravděpodobnostní funkce už není příliš jednoduchá. Obecnější způsob konstrukce mnohorozměrných diskrétních rozdělení ukazuje např. [10] na příkladě s negativně binomickým rozdělením. Předpis pro sdruženou pravděpodobnost je pak ještě značně složitější a samozřejmě opět závisí na „mixovacích“ parmetrech (jako bylo λ_3 pro uvedený Poissonův případ).

Na druhé straně, pokud použijeme metodu založenou na diskretizaci spojitých rozdělení, vyjdeme tedy z vhodného dvourozměrného spojitého rozdělení. To není problém s využitím vhodných marginálních spojitých rozdělení a vybrané kopuly. Samozřejmě i tento přístup se již objevil, v podstatě zase je snahou modelovat spojité rozdělení na základě diskretizovaných dat, obdobně jako je to ukázáno pro jednorozměrný případ Weibullova rozdělení v Poznámce 2. Takový přístup je předveden např. v Barbiero [2]. Ale zde se na tento problém díváme vlastně opačně, máme diskrétní data a snažíme se pro ně najít vhodný a jednoduchý model, a také navrhnout poměrně univerzání metodu takového modelování.

Připomenu zde ještě, že pokud $C(u, v, \theta)$ je kopula, θ její parametr, $\overline{F}_S(s)$, $\overline{F}_T(t)$ marginální funkce přežití dvou spojitých veličin S, T, tak sdružená funkce přežití je

$$P(S > s, T > t) = \overline{F}_2(s, t) = C(\overline{F}_S(s), \overline{F}_T(t), \theta). \tag{4}$$

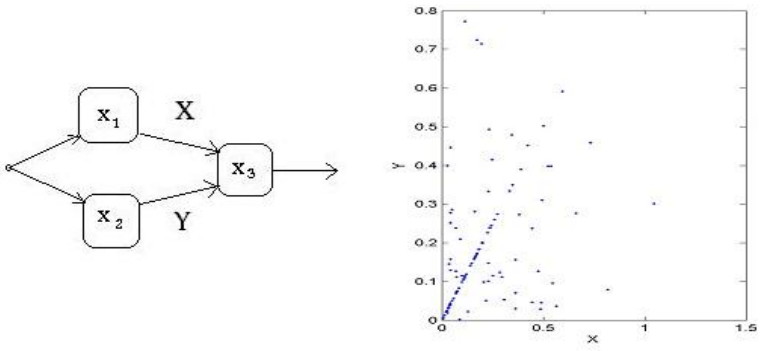
Samozřejmě, různé kopuly vedou k různým sdruženým modelům, i s různou korelací (některé kopuly vedou jen na korelaci v omezeném intervalu, ne v celém $[-1, 1]$), parametr θ je třeba odhadovat. Ale použití takového modelu už je dál poměrně jednoduché, jeho diskretizace pro popis rozdělení nějakých diskrétních náhodných veličin X, Y je nasnadě:

$$\begin{aligned} P(X = K, Y = L) = &\overline{F}_2(K, L) - \overline{F}_2(K, L + 1) - \\ &- \overline{F}_2(K + 1, L) + \overline{F}_2(K + 1, L + 1). \end{aligned} \tag{5}$$

Takto tedy opět poměrně snadno získáme rozsáhlou skupinu dvourozměrných diskrétních rozdělení. Také si všimněme, že vlastně marginální rozdělení pro X a Y nemusí být stejného typu. Poznámky na konci předchozí části navíc ukazují, že takto můžeme vytvářet i dvourozměrné verze některých standardních diskrétních rozdělení.

Volba vhodné kopuly je otázkou porovnání fitu modelů, zásadní je vzít model, který umožňuje adekvátně zachytit korelaci. Podstatné na použití kopul je také to, že dvourozměrný model zachová do něj vstupující marginální rozdělení. To není vždy samozřejmé, ukážeme to na následujícím příkladu s konstrukcí dvourozměrného rozdělení metodou, kterou použili Marshall a Olkin [7]. Ta vlastně souvisí i s výše uvedenou konstrukcí dvourozměrné Poissonovy distribuce.

Představme si systém tří nezávislých prvků (třeba ve výrobní lince) jako na Obrázku 1 vlevo a zkoumejme jeho spolehlivost (resp. dobu do poruchy). Nechť funkce přežití jednotlivých prvků jsou $\overline{F}_j(x)$, $j = 1, 2, 3$, náhodnou



Obrázek 1: Schéma jednoduchého systému. Generovaná data z Marshallova-Olkinova 2D exponenciálního rozdělení s $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3$.

dobu do poruchy pro horní větev, prvky 1 a 3, označíme X, pro dolní větev, prvky 2 a 3, Y. Takže vlastně $X = \min(X_1, X_3)$ a $Y = \min(X_2, X_3)$, marginální rozdělení veličin X a Y mají funkce přežití $P(X > x) = \overline{F}_1(x) \cdot \overline{F}_3(x)$, resp. $P(Y > y) = \overline{F}_2(y) \cdot \overline{F}_3(y)$, zatímco sdružené rozdělení je

$$P(X > x, Y > y) = \overline{F}_1(x) \cdot \overline{F}_2(y) \cdot \overline{F}_3(\max(x, y)).$$

Toto je tedy pravděpodobnost, že horní větev systému vydrží více než x a dolní více než y. Pravděpodobnost výdrže celého systému (označme příslušnou náhodnou veličinu T) by samozřejmě vypadala jinak:

$$P(T > t) = (1 - F_1(t) \cdot F_2(t)) \cdot \overline{F}_3(t).$$

Sdružené rozdělení (X, Y) již není spojité, má singulární část $X = Y$ s pravděpodobností $P(X_3 < \min(X_1, X_2))$. Nejznámější je tento model pro exponenciální rozdělení: Nechť X_j, $j = 1, 2, 3$, mají exponenciální rozdělení s parametry (intenzitami) λ_j, pak marginální rozdělení X a Y jsou exponenciální, s parametry $\lambda_1 + \lambda_3$, $\lambda_2 + \lambda_3$. Jejich korelace $\rho(X, Y) = \lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)$, čili opět nezáporná a omezená v $[0, 1)$. Také $P(X = Y) = \lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)$. Tuto singularitu ukazuje dobře i obrázek 1 vpravo, kde je z tohoto modelu s parametry $\lambda_1 = 1,\ \lambda_2 = 2,\ \lambda_3 = 3$ náhodně generováno a zobrazeno 100 hodnot. Zkusme ale v tomto modelu vzít rozdělení X_j třeba Weibullova, pak X a Y už Weibullovo rozdělení mít nebudou.

4. Ilustrace

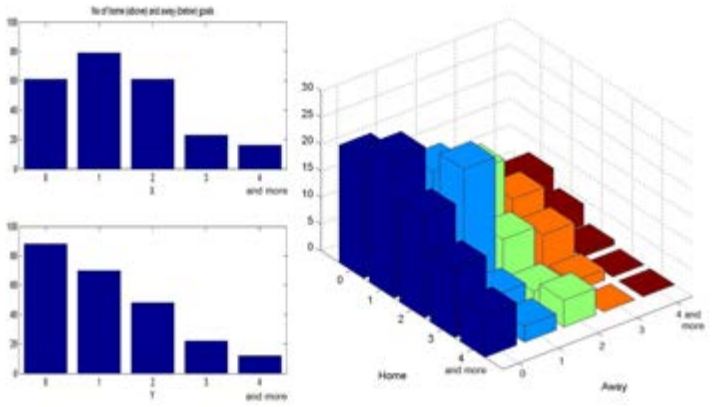
Využití konstrukce dvourozměrných teleskopických rozdělení v souladu s (5) ukážeme na následujícím příkladu. Budeme mít data ve formě („spárovaného“) náhodného výběru rozsahu N z dvojice diskrétních náhodných veličin X, Y. V prvním kroku bude naší snahou najít vhodná marginální rozdělení pro X i Y. Kromě několika diskretizovaných spojitých modelů vyzkoušíme i standardní diskrétní rozdělení (Poissonovo, geometrické, negativně binomické). Parametry rozdělení odhadneme metodou maxima věrohodnosti, dobrou shodu s daty otestujeme graficky i Pearsonovým chí-kvadrát testem. Data budou zpravidla omezena na několik hodnot či tříd, například $0, 1, 2, ..., M-1$ a $\geq M$. Nakonec i pro korektní použití chí-kvadrát testu se doporučuje sloučení tříd s malou pravděpodobností. Pak, i na základě odhadnuté korelace $\rho(X, Y)$, vybereme vhodnou kopulu a v souladu s (4) a (5) zkonstruujeme příslušné dvourozměrné diskrétní rozdělení. Přitom použijeme parametry marginálních rozdělení již odhadnuté, takže v tomto kroku je třeba odhadnout už jen parametr kopuly θ. Shodu opět otestujeme chí-kvadrát testem.

Základním požadavkem úspěchu bude tedy dobrá shoda jak pro dvourozměrný model, tak pro oba modely marginální.

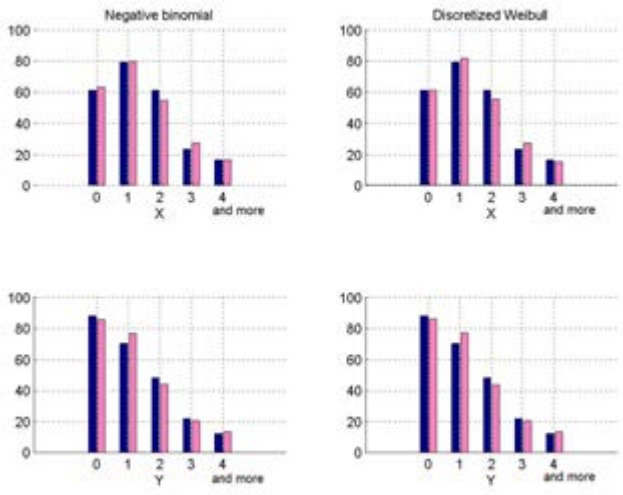
Příklad

I v tomto příkladu zpracovávám data z oblasti sportu. Jedná se o počty gólů vstřelených v jednotlivých utkáních české fotbalové Chance ligy v základní části ročníku 2024–2025. Veličiny X a Y jsou počty vstřelené domácím a hostujícím týmem. Data jsou stažena z webu https://www.sport.cz/sekce/fotbal-ceska-1-liga-vysledky-4 a obsahují tedy výsledky z 30 kol krát 8 utkání, tj. $N = 240$ dvojic hodnot. Obrázek 2 vlevo ukazuje hodnoty X a Y zvlášť. Součty jsou 345 a 282, maximum$(X) = 7$, maximum$(Y) = 5$, průměry jsou 1,438 a 1,175, rozptyly 1,661 a 1,3417, korelační koeficient $r(X, Y) = -0{,}1401$, což je statisticky významná hodnota. Tyto základní charakteristiky byly spočteny z úplných dat. Pravá část Obrázku 2 ukazuje dvourozměrný histogram výsledků. V grafech a pro další výpočty jsou již vysoká skóre sloučena do třídy „větší nebo rovno 4“. Pozorované četnosti $N\text{obs}_j$ jsme následně porovnávali s četnostmi očekávanými, $N\text{exp}_j$, predikovanými jednotlivými modely. Kritériem tedy byla veličina

$$C2 = \sum_j (N\text{obs}_j - N\text{exp}_j)^2 / N\text{exp}_j$$



Obrázek 2: Vlevo: grafy pozorovaných četností veličin X a Y. Vpravo: dvourozměrný graf hodnot (X, Y).



Obrázek 3: Graf shody četností pozorovaných (tmavá) a predikovaných modelem (světlá). Vlevo pro negativně binomické modely, vpravo pro diskretizované Weibullovy modely, nahoře pro veličinu X, dole pro Y.

porovnávaná s 95% kvantilem chí-kvadrát rozdělení s počtem stupňů volnosti rovným počtu tříd – 1 – počet odhadnutých parametrů modelu.

V prvním kroku jsme hledali vhodné 1-rozměrné modely zvlášť pro X i Y. Pro veličinu X vyhovovalo i Poissonovo rozdělení (pro Y již ne), ale daleko lepším se ukázalo pro obě veličiny jak negativně binomické, tak diskretizované Weibullovo rozdělení. Shoda obou s daty je ukázána na Obrázku 3. Pro negativně binomické modely byly obdrženy maximálně věrohodné odhady parametrů $p_x = 0{,}885,\ r_x = 10{,}964,\ p_y = 0{,}751,\ r_y = 3{,}584$, kriterium bylo $C2_x = 1{,}446,\ C2_y = 1{,}276$.

Pro Weibullovo rozdělení, s parametrizací ve tvaru $\overline{F}(t) = \exp(-\alpha \cdot t^{\beta})$, byly pak maximálně věrohodné odhady $\alpha_x = 0{,}292,\ \beta_x = 1{,}623,\ \alpha_y = 0{,}444,\ \beta_y = 1{,}354$ a $C2_x = 1{,}439,\ C2_y = 1{,}353$. Kritická hodnota byla pro všechny 4 testy tatáž, při $5 - 1 - 2 = 2$ stupních volnosti je $C2_{\text{crit}} = 5{,}992$.

Následně bylo z oněch dvou Weibullových marginálních rozdělení vytvořeno dvourozměrné, použili jsme Barnettovu kopulu, která je jednoduchá na výpočet a zaručuje zápornou korelaci. Její tvar je

$$C(u, v) = u \cdot v \cdot \mathrm{e}^{-\theta \ \ln(u)\ \ln(v)}.$$

Zbývá tedy odhadnout parametr θ, ten je v rozsahu $[0, 1]$, což odpovídá mezím pro korelaci $\rho(U, V)$ od nuly do zhruba $-0{,}52$. Pomocí této kopuly zkonstruoval vlastně dvourozměrné exponenciální rozdělení již Gumbel v [4] (aniž by použil pojem „kopula“). My jsme ji tedy využili pro vytvoření dvourozměrné Weibullovy funkce přežití

$$P(X > x, Y > y) = \mathrm{e}^{-\alpha_x \cdot x_x^{\beta}} \cdot \mathrm{e}^{-\alpha_y \cdot y_y^{\beta}} \cdot \mathrm{e}^{-\theta \alpha_x \alpha_y \, x_x^{\beta} \, y_y^{\beta}}.$$

Toto rozdělení bylo pak diskretizováno, dle (5). Maximaximálně věrohodný odhad parametru kopuly dal hodnotu $\theta = 0{,}225$. Odhadli jsme i teoretickou korelaci v příslušném spojitém modelu, $\rho(X, Y) = -0{,}171$. Kritérium pro test dobré shody vyšlo $C2 = 21{,}738$, nyní při $5 \times 5 - 1 - 5 = 19$ stupních volnosti, příslušný 95% kvantil je $C2_{\text{crit}} = 30{,}144$, p-hodnota testu 0,298.

Využití dvourozměrného negativně binomického rozdělení

V poznámce 6 v části 2 bylo ukázáno, že i negativně binomické rozdělení lze vyjádřit jako diskretizované spojité rozdělení, a to pomocí regularizovené neúplné beta funkce:

$$P(X = K) = I_{1-p}(K, r) - I_{1-p}(K + 1, r), \quad \text{kde } I_q(a, b) = B(q; a, b)/B(a, b),$$

přičemž $I_q(t, b)$, jako funkce $t > 0$, je funkcí přežití nějaké nezáporné spojité náhodné veličiny. Takže se pak dá pomocí (4) a (5) a vhodné kopuly

z daných marginálních negativně binomických rozdělení zkonstruovat dvourozměrné rozdělení. To jsem zkusil i v tomto příkladě. Odhadnuté parametry p_x, r_x, p_y, r_y jsou uvedeny výše, a opět jsem použil Barnettovu kopulu. Maximálně věrohodný odhad jejího parametru byl $\theta = 0{,}229$, kritérium testu dobré shody vyšlo $C2 = 21{,}246$, p-hodnota testu 0,323, výsledek je srovnatelný s výsledkem pro dvourozměrné diskretizované Weibullovo rozdělení.

5. Závěr

V příspěvku jsme připomněli skupinu diskrétních rozdělení pravděpodobnosti (také nazývaných „teleskopickými“ rozděleními), která vzniknou diskretizací rozdělení spojitých. Hlavní náplní pak byla obdobná konstrukce dvourozměrných diskrétních rozdělení. Zajímavou otázkou je interpretace takovýchto modelů. Původní ideu reprezentuje třeba příklad v poznámce 2 z úvodu, ale ve většině aplikací je spíš úkolem najít vhodný a početně jednoduchý kvantitativní model, interpretace či věcná motivace pro určitý typ modelu už je spíše nejasná. To ukazuje i zde uvedený příklad.

Nakonec, už i interpretace negativně binomického modelu s reálným parametrem r není úplně zřejmá, jedině snad jako mix Poissonových rozdělení. Proto je často používáno místo Poissonova rozdělení v případech, kdy je rozptyl dat významně větší než průměr, takže i na modelování skóre ve fotbalových zápasech. Například McHale a Scarf [9] konstruovali také dvourozměrné rozdělení pomocí kopuly, i když stále měli marginální distribuční funkce ve formě součtu. Ale uvažovali zároveň i závislost parametrů na kovariátách (jako je hodnocení týmu, indikátor domácí výhody), což je vlastně inspirace pro další vývoj modelů zkoumaných v tomto článku.

Literatura

[1] Anděl, J., 2005. *Základy matematické statistiky.* Matfyzpress, Praha. ISBN 80-86732-40-1. *cit. 13*

[2] Barbiero, A., 2022. *Discrete analogues of continuous bivariate probability distributions.* Annals of Operations Research, 312, 23–43, doi:10.1007/s10479-019-03388-8. *cit. 16*

[3] Basu, A. P. and Dhar, S., 1995. *Bivariate geometric distribution.* Journal of Applied Statistical Science, 2, 33–44. *cit. 15*

[4] Gumbel, E. J., 1960. *Bivariate Exponential Distributions.* J. Amer. Statist. Assoc. 55, 698–707. doi:10.1080/01621459.1960.10483368. *cit. 20*

[5] Johnson, N., Kotz, S., Balakrishnan, N., 1997. *Discrete Multivariate Distributions.* Wiley, New York. ISBN 978-0-471-12844-1. *cit. 15*

[6] Kocherlakota, S., Kocherlakota, K., 1992. *Bivariate Discrete Distributions,* Marcel Dekker, New York. doi:10.1201/9781315138480. *cit. 15*

[7] Marshall, A. W., Olkin, I., 1967. *A multivariate exponential distribution.* J. Amer. Statist. Assoc. 62, 30–44. doi:10.2307/2282907. *cit. 16*

[8] Marshall, A. W., Olkin, I., 1985. *A family of bivariate distributions generated by the bivariate Bernoulli distribution.* J. Amer. Statist. Assoc. 80, 332–338. doi:10.1080/01621459.1985.10478116. *cit. 15*

[9] McHale, I., Scarf, P., 2011. *Modelling the dependence of goals scored by opposing teams in international soccer matches.* Statistical Modelling 11, 219–236. doi:10.1177/1471082X1001100303. *cit. 21*

[10] Ng, C. M., Ong, S. H., Srivastava, H. M., 2010. *A class of bivariate negative binomial distributions with different index parameters in the marginals.* Appl. Mathematics and Computation 217, 3069–3087. doi:10.1016/j.amc.2010.08.040. *cit. 15*

[11] Roknabadi, A. H. R., Borzadaran, G. R. M, Khorashadizadeh, M., 2009. *Some Aspects of Discrete Hazard Rate Function in Telescopic Families.* Economic Quality Control 24, No. 1, 35–42. doi:10.1515/EQC.2009.35. *cit. 11, 12*

[12] Roy, D., 2003. *The Discrete Normal Distribution.* Comm. Statist. – Theory and Methods, 32, 1871–1883. doi:10.1081/STA-120023256. *cit. 12*

[13] Telescoping series, retrieved from `https://en.wikipedia.org/wiki/Telescoping_series`, 2/2025. *cit. 11*

INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 36, číslo 3, září 2025

Obsah

Vědecké a odborné články

Josef Bukač: Logaritmus exponenciálních a Cauchyho náhodných proměnných 3

Petr Volf: O „teleskopických“ rozděleních pravděpodobnosti, jejich dvourozměrné verzi i využití 10

Zprávy a informace

Pavel Stříž: Recenze knihy: Opt Art 23

Patrik Sleziak: Konferencia OSSConf 2025 (správa) 27

Redakce: Pozvánka na Energy Days 2025 29

Redakce: Pozvánka na Mikukláše 2025 31

Redakce: Pozvánka na Robust 2026: nové zprávy 33

Redakce: Pozvánka na České uživatelské fórum Copernicus a Dálkový průzkum Země 2025 34

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.

Předsedkyně společnosti: Ing. Martina Litschmannová, Ph.D., Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 2172/15, 708 33 Ostrava-Poruba, `Martina.Litschmannova@vsb.cz`.

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

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

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

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.