MEŠANE POISSONOVE PORAZDELITVE IN NJIHOVA UPORABA V KOLEKTIVNEM MODELU

ŽAN TAVČAR

Fakulteta za matematiko in fiziko
Univerza v Ljubljani

Mešane porazdelitve so sestavljene iz porazdelitvene funkcije odvisne od parametra, ki je tudi porazdeljen glede na neko porazdelitveno funkcijo. Ta ideja predstavlja alternativni model analize podatkov, njegova prednost pa je, da nam omogoča večjo fleksibilnost, saj je tudi parameter slučajna spremenljivka. Konflikt, ki nas privede do uporabe mešanih porazdelitev je ta, da točno ne vemo kakšen je parameter, zato tudi nanj gledamo kot na slučajno spremenljivko. Uporaba se kaže pri napovedi frekvence odškodninskih zahtevkov. Opisane bodo mešane Poissonove porazdelitve, ki nas privedejo do negativne binomske porazdelitve, ravno ta pa je v kolektivnem modelu uporabljena pri napovedi frekvence odškodninskih zahtevkov za populacijo, kjer tveganja niso homogena.

MIXED POISSON DISTRIBUTION AND ITS USE IN COLLECTIVE MODEL

Mixed distributions are basically distributions which parameter has its own distribution function. That idea gives us alternative model of data analysis. Its advantage is that we have more flexibility in our model, since parameter is also random variable. Conflict that lead us to use mixed distributions is that we are not sure what value does parameter take. We use mixed distributions to calculate distribution of number of claims. Typically, we use mixed Poisson distributions. They lead us to definition of negative binomial distributions, which is widely used in actuarial science, since it describes population which risk is not homogeneous.

1. Mešane porazdelitve

Definicija 1. Slučajna spremenljivka X ima mešano porazdelitev, če obstaja družina porazdelitvenih funkcij $F_X(x,\lambda)$ in podazdelitvena funkcija $G_\Lambda(\lambda)$, kjer je $F_X(x,\lambda) = P(X \leq x|\Lambda = \lambda)$. Po formuli za popolno verjetnost sledi

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{\infty} F_X(x,\lambda)\mathrm{d}G(\lambda). \tag{1}$$

Gostota mešane porazdelitve je enaka

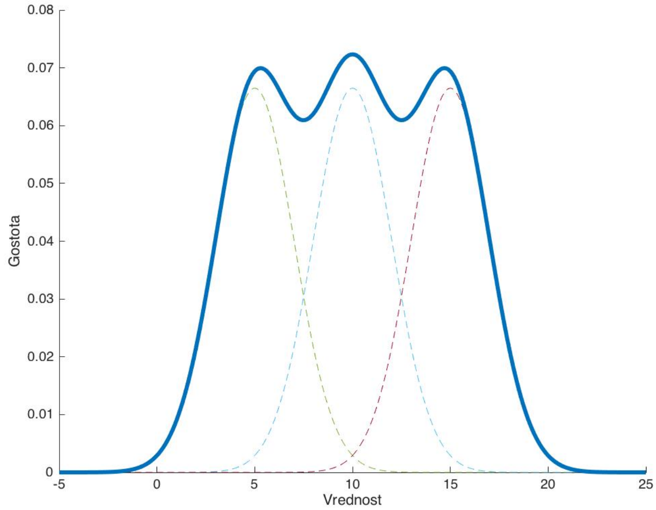
$$f_X(x) = \int_{-\infty}^{\infty} f_X(x,\lambda)g_\Lambda(\lambda)\mathrm{d}\lambda, \tag{2}$$

kjer predpostavljamo, da je gostota slučajne spremenljivke Λ definirana in jo označimo z $g_\Lambda(\lambda)$, izraz $f_X(x,\lambda)$ pa je pogojna gostota slučajne spremenljivke X pri $\Lambda = \lambda$.

Slučajna spremenljivka Λ, tj. slučajna spremenljivka parametra, je lahko porazdeljena diskretno ali zvezno. Zgornja definicija se nanaša na primer, ko je Λ zvezno porazdeljena, brez težav pa lahko prevedemo na diskreten primer, in sicer integral po vseh možnih izbirah slučajne spremenljivke Λ nadomestimo z vsoto. Porazdelitveno funkcijo mešane slučajne spremenljivke v tem primeru zapišemo z $F_X(x) = \sum_i F(x,\lambda_i)P(\Lambda = \lambda_i)$.

Primer 1. Naj bo dana mešana slučajna spremenljivka $X \sim N(\Lambda, 4)$, kjer je Λ porazdeljena diskretno, in sicer $\Lambda \sim \begin{pmatrix} 5 & 10 & 15 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$.

Denimo, da nas zanima gostota mešane porazdelitve. Črtkane črte na grafu predstavljajo gostoto porazdelitve vseh možnih izbir slučajne spremenljivke Λ, ki so še dodatno pomnožene s faktorjem $\frac{1}{3}$. Funkcija mešane gostote slučajne spremenljivke X ima sledečo obliko:



Slika 1. Graf funkcije gostote mešane normalne porazdelitve.

Primer 2. Naj bo $Y \sim Exp(\Lambda)$ in naj bo $\Lambda \sim EZ[0,1]$ slučajna spremenljivka parametra. Denimo, da nas zanima funkcija gostote mešane porazdelitve Y.

$$
\begin{aligned}
f_Y(x) &= \int_0^1 \lambda e^{-\lambda x} \mathrm{d}\lambda = \frac{\lambda}{x} e^{-\lambda x}\Big|_0^1 + \frac{1}{x}\int_0^1 e^{-\lambda x}\mathrm{d}\lambda \\
&= -\frac{1}{x}e^{-x} + \frac{1}{x}\Big(-\frac{1}{x}e^{x} + \frac{1}{x}\Big) \\
&= \frac{x - e^{-x}(x+1)}{x^2}, \quad x > 0.
\end{aligned}
$$

Z mešanjem eksponentne z enakomerno porazdelitvijo na $[0,1]$ smo med drugim pokazali tudi, da velja

$$
\int_0^\infty \frac{x - e^{-x}(x+1)}{x^2}\mathrm{d}x = 1, \text{ saj mora biti integral gostote enak 1.}
$$

Primer 3. Naj bo dana porazdelitev $N \sim Ber(\Lambda)$, kjer je $\Lambda \sim EZ[0,1]$, tj. Λ je enakomerno porazdeljena slučajna spremenljivka na intervalu $[0,1]$. Zanima nas porazdelitev N, velja naslednje

$$
\begin{aligned}
P(N=1) &= \int_0^1 P(N=1,\lambda)\mathrm{d}\lambda = \int_0^1 \lambda \mathrm{d}\lambda = \frac{1}{2}, \\
P(N=0) &= \int_0^1 P(N=0,\lambda)\mathrm{d}\lambda = \int_0^1 (1-\lambda)\mathrm{d}\lambda = (\lambda - \frac{\lambda^2}{2})\Big|_0^1 = \frac{1}{2}.
\end{aligned}
$$

Torej je $N \sim Ber(\frac{1}{2})$.

1.1 Momenti mešanih porazdelitev

Naj bo X mešana slučajna spremenljivka, Λ slučajna spremenljivka parametra z definirano gostoto $g_\Lambda(\lambda)$ ter $f_X(x,\lambda)$ funkcija pogojne gostote slučajne spremenljivke X pri $\Lambda = \lambda$. Če za vsak λ iz nosilca gostote g_Λ skoraj gotovo obstaja $E(X|\Lambda = \lambda)$, tj. skoraj gotovo velja da $\int_{-\infty}^{\infty} x f_X(x,\lambda)\mathrm{d}x$

absolutno konvergira, po formuli za popolno verjetnost sledi, da je pričakovana vrednost mešane slučajne spremenljivke X enaka

$$E(X) = E[E(X|\Lambda)] = \int_{-\infty}^{\infty} E(X|\Lambda = \lambda) g_\Lambda(\lambda) \mathrm{d}\lambda, \tag{3}$$

pod pogojem, da integral v (3) absolutno konvergira. (Nekonvergenca tega integrala je razlog neobstoja upanja v nadaljevanju v Primeru 2, kot je omenjeno v Opombi 1.)

Za varianco mešane slučajne spremenljivke X pa uporabimo znano formulo

$$Var(X) = E[Var(X|\Lambda)] + Var[E(X|\Lambda)], \tag{4}$$

pod pogojem, da navedene količine obstajajo.

V posebnem primeru, ko je Λ diskretna slučajna spremenljivka parametra, integral nadomestimo z vsoto. Pričakovana vrednost je enaka uteženi vsoti pričakovanih vrednosti po vseh izbirah Λ, ob dodatni predpostavki, da spodnja vrsta absolutno konvergira. Označimo $p_i = P(\Lambda = \lambda_i)$ ter $\mu_i = E(X|\Lambda = \lambda_i)$.

$$E(X) = \sum_i p_i \mu_i = \mu < \infty.$$

Če za vsak element iz zaloge vrednosti slučajne spremenljivke parametra, tj. Λ, poleg zgoraj navedenih oznak poznamo še varianco $\sigma_i^2 = Var(X|\Lambda = \lambda_i)$ velja

$$Var(X) = E[Var(X|\Lambda)] + Var[E(X|\Lambda)] = \sum_i p_i(\sigma_i^2 + \mu_i^2) - \mu^2.$$

Primer 4. Uporabimo prvi primer za izračun pričakovane vrednosti in variance mešane porazdelitve. Za vsak $\Lambda = \lambda_i$ poznamo $\mu_i = E[N(\lambda_i, 4)]$ ter $\sigma_i^2 = Var[E[N(\lambda_i, 4)]]$. Še enkrat označimo $p_i = P(\Lambda = \lambda_i)$, sledi

$$E(X) = \sum_{i=1}^{3} \mu_i p_i = \frac{1}{3}(5 + 10 + 15) = 10 = \mu,$$
$$Var(X) = \sum_{i=1}^{3} p_i(\sigma_i^2 + \mu_i^2) - \mu^2 = \frac{1}{3}(29 + 104 + 229) - 10^2 = \frac{62}{3} = \sigma^2.$$

Opomba 1. Pri drguem zgledu smo prišli do funkcije gostote mešane porazdelitve $X \sim Exp(\Lambda)$, kjer je bila $\Lambda \sim EZ[0, 1]$. Zlahka se prepričamo, da upanje ne obstaja. Ker je $E(X|\Lambda) = \frac{1}{\Lambda}$, je $E(X) = E(\frac{1}{\Lambda}) = \infty$.

2. Mešane Poissonove porazdelitve

Preden začnem z obravnavo mešanih Poissonovih porazdelitev je smiselno definirati Poissonovo porazdelitev ter nekaj njenih lastnosti. Čar Poissonove porazdelitve je v tem, da z njo lahko predstavimo dogodke, ki se sami zase zgodijo zelo redko, a imajo veliko možnosti, da se zgodijo. Naj bo $X \sim Pois(\lambda)$. Potem je

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \{0, 1, ...\}. \tag{5}$$

Posebnost Poissonove porazdelitve je v tem, da je pričakovana vrednost enaka varianci. Spodnja izpeljava nam bo koristila v nadaljevanju pri mešanih Poissonovih porazdelitvah.

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda,$$
$$E(X^2) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda\Big(\sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} + \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}\Big) = \lambda^2 + \lambda,$$
$$Var(X) = E(X^2) - E(X)^2 = \lambda.$$

Definicija 2. Pravimo, da je slučajna spremenljivka $X \sim Pois(\Lambda)$ mešana Poissonova porazdelitev, če je Λ nenegativna slučajna spremenljivka s porazdelitveno funkcijo G_Λ. V primeru, ko obstaja gostota $g_\Lambda(\lambda)$, sledi da je verjetnostna funkcija X dana s formulo (6).

$$P(X = k) = E[P(X = k)|\Lambda] = \int_0^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} g_\Lambda(\lambda)\mathrm{d}\lambda, \quad k \in \{0, 1, ...\}. \tag{6}$$

Ker za upanje mešane Poissonove porazdelitve velja $E(X|\Lambda = \lambda) = E[Pois(\lambda)] = \lambda$, sledi $E(X) = E[E(X|\Lambda)] = E(\Lambda)$. Tako vidimo, da je za obstoj $E(X)$ potreben in zadosten pogoj obstoj $E(\Lambda)$. Upanji sta kar enaki. Za varianco pri mešanem Poissonovem modelu pa velja naslednji izrek.

Izrek 1. *Naj bo Λ neka nenegativna slučajna spremenljivka ter $E[\Lambda] < \infty$. Označimo $N \sim Pois(\Lambda)$, kjer je Λ slučajna spremenljivka z definiranim upanjem in disperzijo. Za mešano Poissonovo porazdelitev vedno velja, da je varianca večja ali enaka od pričakovane vrednosti.*

Dokaz.

$$E(N) = E[E(N|\Lambda)] = E(\Lambda) = \mu < \infty$$
$$E(N^2) = E[E(N^2|\Lambda)] = E(\Lambda^2 + \Lambda) = E(\Lambda^2) + \mu$$
$$Var(N) = E(N^2) - E(N)^2 = E(\Lambda^2) + \mu - E(\Lambda)^2 = Var(\Lambda) + \mu \geq \mu = E(N)$$

Sledi, da je $Var(N) \geq E(N)$ in enakost nastopi le, če je Λ izrojena slučajna spremenljivka. Temu pojavu ($Var(N) \geq E(N)$) pravimo predisperziranje Poissonove porazdelitve. ■

Vpeljimo še momentno rodovno funkcijo mešane Poissonove porazdelitve. Ker je $E(e^{rPois(\lambda)}) = e^{\lambda(e^r-1)}$, sledi

$$M_N(r) = E(e^{Nr}) = E[E(e^{Nr})|\Lambda] = E(e^{\Lambda(e^r-1)}) = M_\Lambda(e^r - 1). \tag{7}$$

Trditev 2. *Naj bo $N \sim Pois(\Lambda)$, kjer $\Lambda \sim gamma(a, c)$. Rezultat mešanja Poissonove z gama porazdelitvijo je negativna binomska porazdelitev.*

Dokaz. Vemo, da $p_\Lambda(\lambda) = \frac{c^a}{\Gamma(a)}\lambda^{a-1}e^{-\lambda c}$. Po definiciji gostote mešanih Poissonovih porazdelitev sledi

$$\begin{aligned} P(N = k) = E[P(N = k)|\Lambda] &= \int_0^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \frac{c^a}{\Gamma(a)} \lambda^{(a-1)} e^{-\lambda c}\mathrm{d}\lambda \\ &= \frac{c^a}{\Gamma(a)k!} \int_0^{\infty} \lambda^{k+a-1} e^{-\lambda(c+1)}\mathrm{d}\lambda \\ = \frac{\Gamma(k+a)}{\Gamma(a)k!} \frac{c^a}{(c+1)^{k+a}} &= \frac{\Gamma(k+a)}{\Gamma(a)k!} \Big(\frac{c}{c+1}\Big)^a \Big(\frac{1}{c+1}\Big)^k, \end{aligned} \tag{8}$$

kjer smo pri prehodu iz druge v tretjo vrstico uvedli novo spremenljivko $t = \lambda(c+1)$ ter prepoznali funkcijo gama. Velja, da je $N \sim NegBin(a, \frac{c}{c+1})$. ■

Iz zgornje trditve je lepo vidna motivacija, ki sem jo predstavil na začetku. Negativna binomska porazdelitev se tako uporablja v kolektivnih modelih, in sicer kot predstavitev števila škodnih zahtevkov, kjer imamo opravka z vzorcem ljudi, kjer tveganja niso homogena. V nadaljevanju bom torej opisal kombiniran negativni binomski model, še prej pa je smiselno navesti osnove kolektivnega modela.

3. Osnove kolektivnega modela

Definicija 3. Naj bo $\{X_i\}_{i\in\mathbb{N}}$ zaporedje slučajnih spremenljivk (X_i predstavlja višino i-tega škodnega zahtevka), ki so med seboj neodvisne ter enako porazdeljene. Naj bo N slučajna spremenljivka, ki je neodvisna od velikosti škodnih zahtevkov. Skupno škodo definiramo kot

$$S = \sum_{i=1}^{N} X_i = X_1 + \cdots + X_N. \tag{9}$$

Očitno velja, da je $S = 0$, če je $N = 0$.

Skupna škoda, ki je vsota neodvisnih slučajnih spremenljivk, odvisna od števila ter višine škodnih zahtevkov, je tudi slučajna spremenljivka. Porazdelitvena funkcija slučajne spremenljivke S je

$$\begin{aligned} F_S(x) = P(S \le x) &= \sum_{n=0}^{\infty} P(S \le x | N = n)P(N = n) = \sum_{n=0}^{\infty} P(X_1 + \cdots + X_n \le x)P(N = n) \\ &= \sum_{n=0}^{\infty} H^{*n}(x)P(N = n), \end{aligned} \tag{10}$$

kjer je z zapisom $H^{*n}(x)$ označena n-ta konvolucija, tj. porazdelitvena funkcija vsote n neodvisnih in enako porazdeljenih slučajnih spremenljivk $X_1, \cdots, X_n$ s porazdelitveno funkvijo H. V splošnem je določitev porazdelitvene funkcije za S zahtevna, dovolj je, če poznamo nekaj karakteristik slučajne spremenljivke skupne škode.

Trditev 3. *Matematično upanje in disperzijo slučajne spremenljivke skupne škode izračunamo s pomočjo prve ter druge Waldove formule. Naj bo N porazdelitvena funkcija števila škodnih zahtevkov ter $\{X_i\}_{i\in\mathbb{N}}$ zaporedje slučajnih spremenljivk, ki so med seboj neodvisne, enako porazdeljene ter neodvisne od N. Označimo $S = \sum_{i=1}^{N} X_i$ in dodatno predpostavimo, da za slučajni spremenljivki X in N obstajata prva dva začetna momenta. Za matematično upanje in disperzijo slučajne spremenljivke S veljata naslednji formuli*

$$E(S) = E(N)E(X), \tag{11}$$

$$D(S) = D(N)E(X)^2 + E(N)D(X). \tag{12}$$

Dokaz. Dokaz obeh sledi s pomočjo pogojnega matematičnega upanja. V spodnjem dokazu je vpeljana oznaka $S_n = X_1 + \cdots + X_n$, kjer so X_i med seboj neodvisne, enako porazdeljene ter neodvisne od N.

$$
\begin{aligned}
E(S) &= E[E(S|N)] = \sum_{n=0}^{\infty} E(S|N=n)P(N=n) = \sum_{n=0}^{\infty} E(S_n)P(N=n) \\
&= \sum_{n=0}^{\infty} E(X_1+\cdots+X_n)P(N=n) = E(X)\sum_{n=0}^{\infty} nP(N=n) = E(X)E(N) \\
E(S^2) &= E[E(S^2|N)] = \sum_{n=0}^{\infty} E(S^2|N=n)P(N=n) = \sum_{n=0}^{\infty} E(S_n^2)P(N=n) \\
&= \sum_{n=0}^{\infty}[D(S_n)+E(S_n)^2]P(N=n) = D(X)\sum_{n=0}^{\infty} nP(N=n) + E(X)^2\sum_{n=0}^{\infty} n^2P(N=n) \\
&= D(X)E(N) + E(X)^2E(N^2)
\end{aligned}
$$

Iz zapisanega sledi $D(S) = E(S^2) - E(S)^2 = D(N)E(X)^2 + E(N)D(X)$. ■

Zapišimo še momentno rodovno funkcijo slučajne spremenljivke S. Velja

$$
\begin{aligned}
M_S(r) &= E(e^{rS}) = E[E(e^{rS}|N)] = \sum_{n=0}^{\infty} E(e^{rS}|N=n)P(N=n) = \sum_{n=0}^{\infty} E(e^{rS_n})P(N=n) \\
&= \sum_{n=0}^{\infty} E(e^{r(X_1+\cdots+X_n)})P(N=n) = \sum_{n=0}^{\infty} M_X(r)^n P(N=n) = W_N(M_X(r)),
\end{aligned} \tag{13}
$$

kjer je z $W_N(r)$ označena rodovna funkcija slučajne spremenljivke N. Zadnji korak lahko izpeljemo tudi na drugačen način, in sicer

$$M_S(r) = \sum_{n=0}^{\infty} M_X(r)^n P(N=n) = E\Big(e^{\log M_X(r) N}\Big) = M_N(\log M_X(r)) \tag{14}$$

Višje momente lahko izračunamo s pomočjo kumulantne funkcije, ki je definirana s predpisom $K_S(r) = \log M_S(r)$. Kumulanta n-tega reda je definirana z n-tim odvodom kumulantne funkcije pri $r = 0$. Matematično upanje je enako prvi kumulanti (prvemu odvodu kumulantne funkcije pri $r = 0$), varianca pa drugi kumulanti (drugemu odvodu kumulantne funkcije pri $r = 0$). Koeficient asimetrije slučajne spremenljivke S izračunamo z enačbo

$$A(S) = \frac{E[(S-E(S))^3]}{D(S)^{\frac{3}{2}}}, \tag{15}$$

kjer števec izraza izračunamo s pomočjo tretje kumulante, tj.

$$E[(S-E(S))^3] = \frac{\mathrm{d}}{\mathrm{d}r^3}\log(M_S(r))\Big|_{r=0} \tag{16}$$

3.1 Kombiniran mešani Poissonov model

Ko je govora o kombiniranih modelih, imamo v mislih, da za število škodnih zahtevkov vzamemo neko konkretno slučajno spremenljivko. V kombiniranem Poissonovem modelu tako velja, da je $N \sim Pois(\lambda)$. Kombiniran Poissonov model ima to omejitev, da $E(N) = D(N)$. Preprost način, da to veliko omejitev odpravimo, je ta, da tudi na parameter gledamo kot na slučajno spremenljivko. Porodi se ideja, da je N v kolektivnem modelu mešana Poissonova slučajna spremenljivka.

Iz dokaza prvega izreka vemo, da sta upanje in disprezija mešane Poissonove porazdelitve definirana z $E(N) = E(\Lambda)$ ter $Var(N) = Var(\Lambda) + E(\Lambda)$. Naj bo sedaj $S = \sum_{i=1}^{N} X_i$, kjer so $\{X_i\}_{i\in\mathbb{N}}$ neodvisne in enako porazdeljene, N je neodvisna od vseh X_i ter porazdeljena mešano Poissonovo (torej $N \sim Pois(\Lambda)$). S pomočjo prve in druge Waldove formule izpeljemo

$$E(S) = E(N)E(X) = E(\Lambda)E(X), \tag{17}$$

$$\begin{aligned} Var(S) = Var(N)E(X)^2 + E(N)Var(X) &= [Var(\Lambda) + E(\Lambda)]E(X)^2 + E(\Lambda)Var(X) \\ &= Var(\Lambda)E(X)^2 + E(\Lambda)E(X^2). \end{aligned} \tag{18}$$

V primeru kombiniranega mešanega Poissonovega modela je momentno rodovna funkcija

$$M_S(r) = M_N(\log M_X(r)) = M_\Lambda(M_X(r) - 1), \tag{19}$$

kjer smo upoštevali

$$M_N(r) = E(e^{Nr}) = E[E(e^{Nr}|\Lambda)] = E(e^{\Lambda(e^r-1)}) = M_\Lambda(e^r - 1).$$

Iz kombiniranega mešanega Poissonovega modela sledi podobna ideja kot v prvem delu besedila. Če je slučajna spremenljivka Λ gama porazdeljena, tj. $\Lambda \sim gamma(a, c)$, dobimo kombiniran negativni binomski model.

LITERATURA

[1] R. Kaas, M. Goovaerts, J. Dhaene in M. Denuit, *Modern actuarial risk theory using R, 2nd edition*, Springer (2008), (41–51).

[2] R. Jamnik, *Verjetnostni račun*, DMFA Slovenije(1987).

[3] D. Karlis, E. Xekalaki, *Mixed Poisson distributions*, International Statistical Review (2005).

[4] H. Schmidli, *Lecture notes on Risk Theory*, [ogled: 29.7.2016], dostopno na http://www.mi.uni-koeln.de/ schmidli/vorl/Risk/vorl.pdf.

[5] D. Ma, *Poisson Gamma mixture*, [ogled: 29.7.2016], dostopno na https://probabilityandstats.wordpress.com/tag/poisson-gamma-mixture.