Študijsko gradivo ■

Poissonova porazdelitev – osnove, uporaba, nadgradnja

Gaj Vidmar

Izvleček. Gradivo celovito predstavlja Poissonovo porazdelitev. Izpeljana je iz limite binomske porazdelitve. Predstavljene so njene temeljne lastnosti – oblika, momenti, rodovna funkcija in konvolucija. Sledijo primeri uporabe s prostorskega in časovnega vidika, ocenjevanje parametra, preverjanje prileganja empiričnim podatkom (s statističnimi testi in grafično), pojem nad- in podrazpršenosti ter zgodovinski pregled. V zadnjem delu gradiva so kratko predstavljene izbrane nadgradnje: povezane porazdelitve; zmesi Poissonovih slučajnih spremenljivk; dvo- in večrazsežna Poissonova porazdelitev; statistični test za primerjavo dveh vrednosti iz Poissonove porazdelitve; Poissonova regresija; in kontrolne karte, povezane s Poissonovo porazdelitvijo. Gradivo spremlja obsežen interaktiven delovni zvezek v obliki Excel 2007/2010.

Poisson Distribution – Fundamentals, Applications, Extensions

Abstract. The tutorial comprehensively introduces the Poisson distribution. It is derived as a limit of the binomial distribution. Its fundamental properties are presented – shape, moments, moment generating function and convolution. Examples of its application in spatial and time framework are given, followed by parameter estimation, goodness-of-fit (via statistical tests and graphical methods), the concepts of under- and overdispersion, and a historical overview. The final part of the tutorial briefly presents selected extensions: related distributions; bi- and multivariate Poisson distribution; mixtures of Poisson random variables; a statistical test for comparing two Poisson counts; Poisson regression; and control charts related to the Poisson distribution. The tutorial is accompanied by a comprehensive and detailed interactive workbook in Excel 2007/2010 format.

■ **Infor Med Slov**: 2012; 17(2): 29-55

Instituciji avtorja / Author's institutions: Univerzitetni rehabilitacijski inštitut Republike Slovenije – Soča; Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biostatistiko in medicinsko informatiko.

Kontaktna oseba / Contact person: Gaj Vidmar, URI – Soča, Linhartova 51, SI-1000 Ljubljana. e-pošta / e-mail: gaj.vidmar@ir-rs.si.

Prejeto / Received: 30.10.2012. Sprejeto / Accepted: 17.12.2012. Recenzenta / Reviewers: prof. dr. Primož Ziherl in dr. Tim Vidmar.

Uvod

Poissonova porazdelitev je ena od osnovnih diskretnih verjetnostnih porazdelitev. V ogromni množici učnega gradiva s področja verjetnosti in statistike ter ved, kjer se verjetnost in statistiko uporablja, je praviloma na vrsti takoj za binomsko porazdelitvijo. Iz nje jo bomo kmalu tudi izpeljali, a da ne bi začeli s "suhoparno" matematiko, si najprej zastavimo dve vprašanji iz vsakdanjega življenja. – Kakšne rezultate bi dobili, če bi

- sedeli ob cesti (najsi bo prometni v mestu ali samotni v gozdu) in šteli, koliko vozil pripelje mimo na izbrano časovno enoto (npr. minuto v mestu ali uro v gozdu)?

- na travnik narisali kvadratno mrežo in šteli, koliko je v vsakem kvadratu neke cvetlice (najsi bo pogoste, kot je marjetica, ali redke, kot je štiriperesna deteljica)?

Izkazalo se bo, da nam na obe vprašanji pomaga odgovoriti Poissonova porazdelitev. Njena zgodovina je dolga, pestra in pomembna, a da bi jo lahko razumeli, moramo najprej spoznati matematične osnove. Še prej pa napotki za branje oziroma nadaljnje delo:

- gradivo spremlja dinamičen interaktiven delovni zvezek v obliki Excel 2007/2010 s prikazi porazdelitev in podatkovji. Dostopen je v obliki arhiva (ZIP) na naslovu http://ims.mf.uni-lj.si/archive/17(2)/31_s.zip. V njem je posebna pozornost namenjena pogojnemu oblikovanju (*Conditional Formatting*) s paličnimi grafikoni v celicah (*Data Bars*) in barvnimi merili (*Color Scales*), kar je priročno in učinkovito za prikaz podatkov s tabelografi. Formule, uporabljene v funkcijah, so izbrane tako, da se da delovni zvezek skoraj brez izgube funkcionalnosti uporabljati tudi z brezplačno elektronsko preglednico Calc iz odprtokodne zbirke LibreOffice;

- viri so navedeni v treh sklopih: učbeniki, članki iz Wikipedije in dodatni viri. Znotraj vsakega sklopa so navedeni po abecednem vrstnem redu. Kot je navada pri učbenikih in drugem pedagoškem gradivu, se besedilo[v obliki referenc] sklicuje le na nekatere vire;

- oštevilčene so le enačbe [v oglatih oklepajih], na katere se besedilo kasneje sklicuje.

Izpeljava

Poissonova porazdelitev je limitna oblika binomske, pri kateri je število poskusov (n) zelo veliko, verjetnost uspeha v vsakem posameznem poskusu (p) pa zelo majhna (zato je znana tudi kot *porazdelitev redkih dogodkov*). Če v obrazcu za binomsko porazdelitev (natančneje rečeno: verjetnostno funkcijo binomsko porazdeljene slučajne spremenljivke X, ki lahko zavzame vrednosti $k = 0,1,2,\ldots$)

$$X \sim Bin(n,p) \Leftrightarrow P(X=k) = \frac{k!}{n!(n-k)!}\, p^k (1-p)^{n-k}$$

vpeljemo $\lambda = np$ in torej $p = \lambda/n$, z nekaj preurejanja dobimo

$$P(k) = \left[\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdots\frac{n-k+1}{n}\right]\frac{\lambda^k}{k!}\left[\left(1-\frac{\lambda}{n}\right)^{n-k}\right].$$

Če sedaj n pošljemo v neskončnost, pri čemer ostaneta k in λ nespremenjena, bodo šli vsi ulomki v prvem oglatem oklepaju proti 1, izraz v drugem oglatem oklepaju pa bo šel proti $e^{-\lambda}$. Iz definicije Eulerjevega števila e kot limite izraza $(1+1/n)^n$ namreč izhaja, da če gre $n \to \infty$, velja $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$ in $\left(1-\frac{\lambda}{n}\right)^{-x} \to 1$.

Tako dobimo obrazec za verjetnostno funkcijo Poissonove porazdelitve, ki velja za $k = 0,1,2,\ldots$ pod pogojem $\lambda > 0$:

$$X \sim Pois(\lambda) \Leftrightarrow P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}. \qquad [1]$$

Pred predahom še dokažimo, da gre res za verjetnostno porazdelitev, torej da je vsota posameznih verjetnosti enaka 1. Iz definicije eksponentne funkcije $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ sledi, da je

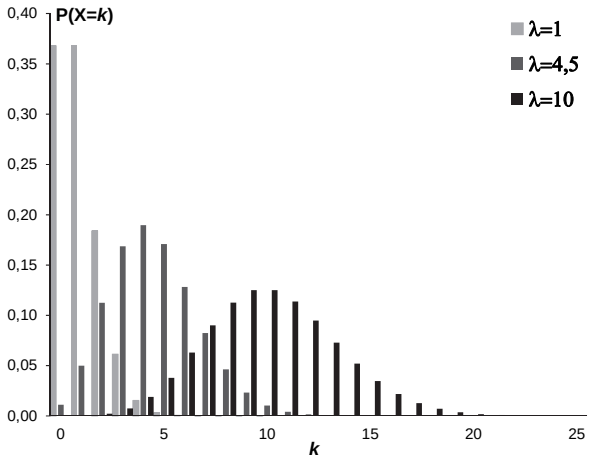
$$\sum_{k=0}^{\infty} P(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1 .$$

Lastnosti

Oblika

Kakšne oblike je Poissonova porazdelitev? Za različne vrednosti parametra λ je prikazana na sliki 1 in na 1. delovnem listu priloženega Excelovega delovnega zvezka. Nakazuje se, kar bodo kmalu potrdili izračuni:

- Poissonova porazdelitev je desno asimetrična (a vse manj z večanjem λ);
- Poissonova porazdelitev z večanjem λ (kmalu) postane podobna normalni;
- modus Poissonove porazdelitve je (približno) enak λ.



Slika 1 Verjetnostna funkcija Poissonove porazdelitve za tri izbrane vrednosti parametra λ.

Poglejmo si obrazec [1] za prve štiri vrednosti k:

$$P(0) = e^{-\lambda}$$
$$P(1) = \lambda e^{-\lambda} = \lambda P(0)$$
$$P(2) = \lambda^2 e^{-\lambda}/2 = (\lambda/2)\lambda e^{-\lambda} = (\lambda/2)P(1)$$
$$P(3) = \lambda^3 e^{-\lambda}/(3\cdot 2) = (\lambda/3)\left(\lambda^2 e^{-\lambda}/2\right) = (\lambda/3)P(2)$$

Hitro uvidimo splošno pravilo (ki ga sicer ni težko dokazati). Ker je celo za sodobne računalnike težko računati fakultete velikih števil, si je zato za računske potrebe potrebno zapomniti le, da verjetnost za $k = 0$ znaša $e^{-\lambda}$, in obrazec

$$P(k+1) = \frac{\lambda}{k+1} P(k) .$$

Ta rekurzivni obrazec nam tudi pojasni obliko Poissonove porazdelitve. Dokler je faktor $\frac{\lambda}{k+1}$ večji od 1, z naraščanjem k naraščajo tudi verjetnosti, ko pade pod 1, pa začno padati (in to vse hitreje). Porazdelitev je torej unimodalna, pri čemer je modus en, če λ ni naravno število, če je, pa sta modusa dve sosednji vrednosti k.

Momenti

Povprečje (pričakovano vrednost, matematično upanje, prvi moment) in varianco (disperzijo, drugi centralni moment) Poissonove porazdelitve je najpreprosteje izpeljati na enak način kot samo porazdelitev – z limito binomske porazdelitve:

- ker je povprečje binomske porazdelitve np, je v skladu z uvodno vpeljavo *povprečje Poissonove porazdelitve enako* λ;
- ker je varianca binomske porazdelitve $np(1-p)$ in ker gre $(1-p)$ proti 1, če gre p proti 0, je *varianca Poissonove porazdelitve* tudi *enaka* λ (standardni odklon pa je $\sqrt{\lambda}$).

Tudi izpeljava iz definicij ni pretežka. Pri izpeljavi povprečja upoštevamo že omenjeno definicijo eksponentne funkcije, dejstvo, da je $k/k! = 1/(k-1)!$,

in dejstvo, da je šteti x od 0 dalje enako kot šteti $x-1$ od 1 dalje. Pri izpeljavi variance poleg tega upoštevamo, da je $kA=(k-1)A+A$, in zaradi preglednosti označimo $k-2$ z i, $k-1$ pa z j:

$$E(X)=\sum_{k=0}^{\infty}kP(k)=\sum_{k=1}^{\infty}ke^{-\lambda}\frac{\lambda^k}{k!}=$$
$$=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda;$$

$$E(X^2)=\sum_{k=0}^{\infty}k^2e^{-\lambda}\frac{\lambda^k}{k!}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}k\frac{\lambda^{k-1}}{(k-1)!}=$$
$$=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\left((k-1)\frac{\lambda^{k-1}}{(k-1)!}+\frac{\lambda^{k-1}}{(k-1)!}\right)=$$
$$=\lambda e^{-\lambda}\left(\lambda\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}+\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}\right)=$$
$$=\lambda e^{-\lambda}\left(\lambda e^{\lambda}+e^{-\lambda}\right)=\lambda(\lambda+1)\Rightarrow$$

$$Var(X)=E(X^2)-E(X)^2=\lambda(\lambda+1)-\lambda^2=\lambda\,. \quad [2]$$

Dejstvo, da je varianca Poissonovo porazdeljene slučajne spremenljivke enaka njenemu povprečju, je najbolj znana lastnost Poissonove porazdelitve. Je tudi osnovno merilo za prepoznavanje Poissonove porazdelitve oziroma prvi kriterij pri presojanju, ali je Poissonova porazdelitev ustrezen model za dane empirične podatke.

Tretji in četrti centralni moment navedimo brez izpeljave (da prihranimo nekaj matematičnega zagona še za naslednji razdelek), vseeno pa nam bosta s svojim limitnim obnašanjem pomagala razjasniti obliko Poissonove porazdelitve. Asimetričnost je $m_3=1/\sqrt{\lambda}$ (torej vedno pozitivna oziroma desna, a se z večanjem λ približuje 0), sploščenost pa $m_4=3+(1/\lambda)$ (torej večja kot 3, kolikor znaša pri normalni porazdelitvi, a z večanjem λ razlika od normalne porazdelitve izginja).

Rodovna funkcija in konvolucija

Rodovna funkcija ni splošno znan pojem, saj presega gimnazijsko matematiko, ki smo se je doslej držali, a že angleški izraz (*moment generating function*) ga pomaga razjasniti. Rodovna funkcija je definirana kot pričakovana vrednost eksponentne funkcije produkta slučajne spremenljivke X in pomožne spremenljivke t. Uporabna je zato, ker če jo *r*-krat odvajamo glede na t in postavimo $t=0$, dobimo *r*-ti moment (surovi, tj. okrog nič) porazdelitve X. V primeru Poissonove porazdelitve je rodovna funkcija

$$M(t)=E(e^{tX})=e^{-\lambda}\sum_{k=0}^{\infty}e^{tk}\frac{\lambda^k}{k!}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!}=e^{-\lambda}e^{\lambda e^t}\,.$$

Z njo smo se spoznali zato, da bi ugotovili, kakšna je porazdelitev vsote (tj. v jeziku fizikov in inženirjev: konvolucija) Poissonovih slučajnih spremenljivk. To pot smo ubrali, ker za rodovno funkcijo vsote dveh neodvisnih slučajnih spremenljivk (X in Y) velja, da je enaka zmnožku rodovnih funkcij posameznih spremenljivk:

$$M(t)=E\left[e^{t(X+Y)}\right]=E(e^{tX}e^{tY})=$$
$$=E(e^{tX})\cdot E(e^{tY})=M_1(t)\cdot M_2(t)\,.$$

Rodovna funkcija vsote dveh Poissonovih slučajnih spremenljivk (s parametroma λ in ν) je zato

$$e^{-\lambda}e^{\lambda e^t}e^{-\nu}e^{\nu e^t}=e^{-(\lambda+\nu)}e^{(\lambda+\nu)e^t}\,,$$

kar je rodovna funkcija Poissonove slučajne spremenljivke s parametrom $\lambda+\nu$. Tako smo prišli do še ene zanimive in pomembne lastnosti Poissonove porazdelitve: *vsota dveh (in torej tudi več) Poissonovih slučajnih spremenljivk je zopet Poissonova slučajna spremenljivka*:

$$X\sim Pois(\lambda)\wedge Y\sim Pois(\nu)\Rightarrow X+Y\sim Pois(\lambda+\nu)\,. \quad [3]$$

Morda je koga zaskrbelo, da je to v nasprotju s centralnim limitnim izrekom, ki (poenostavljeno rečeno) pravi, da če vzamemo veliko slučajnih

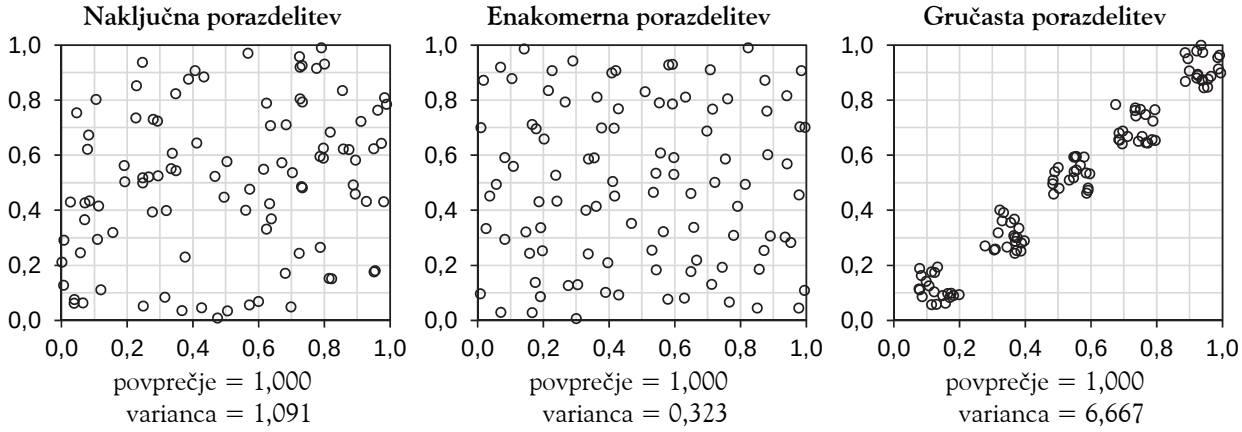
vrednosti iz neke porazdelitve, se njihova vsota porazdeljuje normalno. A da je ta skrb odveč, nas prepriča premislek: čim več Poissonovih spremenljivk seštejemo, tem večja bo vsota njihovih parametrov, ki je hkrati povprečje porazdelitve vsote, in večje kot je povprečje Poissonove porazdelitve, bolj je ta podobna normalni.

Prostorski in časovni vidik

Prostorski vidik

Premaknimo se od matematike v nekoliko oprijemljivejši svet (računalnikov, zlasti Excela) oziroma celo nazaj k naravi (na travnik s cvetlicami). Tri tipične možnosti dvorazsežne prostorske porazdelitve (npr. cvetlic, ki jih predstavljajo krožci, na travniku) prikazujeta slika 2 in 2. delovni list v priloženem Excelovem delovnem zvezku. V vseh treh primerih je zraslo 100 cvetlic s koordinatami med 0 in 1, mreža mejá kvadrantov pa ima vodoravni in navpični razmik 0,1 (torej je kvadrantov 100 in je povprečje števila cvetlic na kvadrant 1).

- Pri naključni porazdelitvi smo vsako dvojico koordinat izžrebali iz enakomerne porazdelitve (matematično rečeno: $x \sim U(0;1), y \sim U(0;1)$; po Excelovo rečeno: s funkcijo RAND).
- Pri enakomerni porazdelitvi smo cvetlice razporedili na mrežo kvadrantov in jih nato malo premaknili po naključju (ang. *jittering*).
- Pri gručasti porazdelitvi smo izhajali iz petih središč na glavni diagonali in nato uporabili naključne majhne premike.



Slika 2 Trije primeri dvorazsežnih prostorskih porazdelitev: naključna (kjer za število enot na kvadrant velja Poissonova porazdelitev), enakomerna (ki je v primerjavi s Poissonovo podrazpršena) in gručasta (ki je primerjavi s Poissonovo nadrazpršena).

Simulacija v Excelu je dinamična, saj se spremeni ob vsakem ponovnem izračunu (tj. če datoteko shranimo ali pritisnemo tipko F9). Poleg tega so razsevnim grafikonom dodani toplotni zemljevidi (ang. *heat maps*), v katerih so kvadranti obarvani po barvni lestvici glede na to, koliko cvetlic vsebujejo (od modre barve za najmanjše število preko sive za povprečno do rdeče za največje).

Pri naključni porazdelitvi je porazdelitev števila cvetlic na kvadrant Poissonova, o čemer priča enakost povprečja in variance (ki ga v Excelu z barvno asociacijo podkrepljuje zelena barva). Pri enakomerni porazdelitvi so cvetlice razpršene manj, kot bi pričakovali po Poissonovi porazdelitvi (na kar nas v Excelu asociira živo modra barva), kar označujemo s pojmom *podrazpršenost* (ang.

underdispersion). Pri gručasti porazdelitvi pa je varianca mnogo večja od povprečja (na kar nas v Excelu asociira živo rdeča barva), čemur z vidika Poissonove porazdelitve pravimo *nadrazpršenost* (ang. *overderdispersion*).

Časovni vidik

Poissonova porazdelitev izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem intervalu, če vemo, da se ti dogodki pojavljajo z znano povprečno pogostnostjo in neodvisno drug od drugega. Medsebojna neodvisnost dogodkov pomeni, da je čas od zadnjega dogodka do naslednjega neodvisen od časa, ki je pretekel od predzadnjega dogodka do zadnjega.

Primerov iz sodobnega sveta je na pretek:

- s področja storitev, proizvodnje oziroma poslovanja – število avtomobilov, ki pridejo na uro v avtopralnico; število strank, ki pridejo na uro na bančno okence; število kupcev, ki pokličejo servis zaradi okvare gospodinjskega aparata v garanciji, na mesec; število predlogov za stečajni postopek, vloženih na določeno sodišče, na mesec; število letal določenega modela v lasti določenega letalskega prevoznika, pri katerih pride do okvare motorja, na 100.000 ur letenja; število turističnih potovanj, na katera so se odpravili člani gospodinjstva v enem letu;

- s področja medicine – število pacientov, ki pridejo v nočnem času v ambulanto ali na kliniko, kjer ni naročanja (dežurna, urgenca ipd.), na uro; število mutacij določenega sklopa DNK na časovno enoto;

- s področja računalništva in telekomunikacij – število okvar določenega omrežja na dan; število okužb z virusi na določenem podatkovnem strežniku na 24 ur; število obiskov priljubljene spletne strani na minuto; število telefonskih klicev v klicni center na minuto …

Za Poissonovo porazdelitev v času je bistveno, da verjetnost za število dogodkov, ki se zgodijo v določenem časovnem obdobju, ki obsega t časovnih enot, izračunamo po obrazcu [1], v katerega namesto λ vstavimo λt :

$$P(k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}. \qquad [4]$$

To lahko hitro uporabimo na primeru iz biologije. Denimo, da ujeda ujame v povprečju eno miš na dan. Če na opazovanem območju živi 10 miši, kolikšna je verjetnost, da bo ujeda vse polovila v enem tednu? V obrazec [4] vstavimo $k = 10, \lambda = 1, t = 7$ in izračunamo, da znaša verjetnost dobrih 7%:

$$P(7) = \frac{e^{-7}\,7^{10}}{10!} = 0{,}071.$$

Nekoliko zahtevnejši in za urbanizirani svet nekoliko bolj življenjski je primer s komercialno telefonsko linijo, ki lahko zaradi omejenega števila zaposlenih in narave dela sprejme največ dva klica v petih minutah. V povprečju prejme 0,5 klica na minuto. Kolikšen je ocenjeni delež klicev, na katere osebje ne bo moglo odgovoriti? Če računamo "peš", po obrazcu [4] izračunamo verjetnosti za 0, 1 in 2 klica v petih minutah $(k = 0, 1, 2; \lambda = 0{,}5/\text{min}; t = 5\text{min})$ ter jih seštejmo, s čimer dobimo verjetnost, da bo osebje kos klicem (tj. da bo prejelo največ 2 v 5 minutah), nato pa to verjetnost odštejmo od 1:

$$\begin{aligned}
&P(\le 2) = P(0) + P(1) + P(2) = \\
&= e^{-\lambda t} + (\lambda t)e^{-\lambda t} + \frac{e^{-\lambda t}(\lambda t)^2}{2} = \\
&= e^{-\lambda t}\left(1 + \lambda t + \frac{(\lambda t)^2}{2}\right) = e^{-2{,}5}\left(1 + 2{,}5 + \frac{2{,}5^2}{2}\right) = \\
&= 0{,}082085 \cdot 6{,}625 = 0{,}544 \\
&\Rightarrow P(> 2) = 1 - P(\le 2) = 0{,}456.
\end{aligned}$$

V Excelu 2007/2010 je najpreprosteje, če v klicu funkcije POISSON.DIST uporabimo zadnji argument TRUE, s čimer takoj dobimo verjetnost za največ dva klica, kot je prikazano na 3.

delovnem listu priloženega Excelovega delovnega zvezka. Drugače povedano, gre za porazdelitveno funkcijo (pred katero je pridevnik "kumulativna" pri matematično natančnem izražanju odveč, za nematematike pa dobrodošla komunikacijska redundanca, ki zmanjšuje verjetnost napačnega razumevanja).

Če smo že pri časih med dogodki, se lahko vprašamo, kako so ti porazdeljeni? Odgovor, ki nas vodi k eni od osnovnih zveznih porazdelitev, ni zelo zapleten in sledi, a ker je povezan s pomembnimi raziskavami in znanimi ljudmi iz preteklosti, se bomo prej posvetili uvodoma obljubljeni zgodovini. Še prej pa si poglejmo, kako sploh ocenimo Poissonov parameter in kako ugotavljamo, ali se opažena porazdelitev sklada s Poissonovo.

Ocenjevanje in prileganje

Ocenjevanje parametra

Ker smo dokazali, da je povprečje Poissonove porazdelitve enako njenemu parametru λ, je naravno sklepati tudi v nasprotni smeri, torej da je najboljša cenilka za parameter Poissonove porazdelitve, ki se najbolje prilega opaženim podatkom, povprečje opaženih podatkov. Potrdimo to z metodo največjega verjetja (ang. *maximum likelihood*). Verjetje opaženih vrednosti $k_i, i = 1..n$ kot niza n realizacij Poissonove slučajne spremenljivke X s parametrom λ je verjetnost, da bi v slučajnem poskusu dobili opažene podatke. Ker gre za skupno verjetnost, posamezne verjetnosti, ki jih za vsak k_i izračunamo po obrazcu [1], med seboj zmnožimo:

$$L = e^{-n\lambda} \frac{\lambda^{k_1+\ldots+k_n}}{k_1!\ldots k_n!}.$$

Namesto verjetja je ekvivalentno, a lažje maksimirati njegov logaritem:

$$\ln L = -n\lambda + (k_1+\ldots+k_n)\ln\lambda - \ln(k_1!\ldots k_n!).$$

Odvod logaritma verjetja po parametru λ je

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{k_1+\ldots+k_n}{\lambda} = \frac{n}{\lambda}\left(\overline{X} - \lambda\right),$$

cenilka λ po metodi največjega verjetja pa je vrednost, za katero je odvod enak nič, torej $\overline{X}$ (kot smo slutili). Hkrati smo (v skladu z zahtevno matematiko, ki jo bomo izpustili – Rao-Cramérjevo oceno in teorijo učinkovitosti cenilk) dobili varianco cenilke, ki je obratna vrednost izraza pred tistim $\left(\overline{X} - \lambda\right)$, v katerem nastopa cenilka. *Standardna napaka* (tj. koren variance) ocene povprečja Poissonove porazdelitve je torej $\sqrt{\lambda/n}$. Isto dobimo iz splošnega obrazca za standardno napako ocene povprečja, vsem znanega $\sigma/\sqrt{n}$, če v skladu z [2] za standardni odklon vstavimo $\sqrt{\lambda}$. Kot pri vsaki cenilki po metodi največjega verjetja (za tem je spet zahtevna teorija, ki jo bomo izpustili), ima tudi naša cenilka za λ (asimptotično) normalno vzorčno porazdelitev. Zato lahko ocenimo *interval zaupanja* z množenjem standardne napake z vrednostjo iz standardne normalne porazdelitve, ki ustreza želeni stopnji zaupanja (1,96 za 95% interval zaupanja itd.).

Opisani pristop k ocenjevanju intervalov zaupanja je znan kot Waldova metoda (po Abrahamu Waldu, 1902-1950, pionirju statistične teorije odločanja in operacijskih raziskav). V primeru Poissonove porazdelitve je ustrezna le za velike vrednosti λ (vsaj 30), a k sreči je eksaktna metoda zelo preprosta za uporabo (če že ne izpeljavo, ki je še nekoliko zahtevnejša od izpeljave asimptotične metode). Temelji na enakosti Poissonove kumulativne porazdelitvene funkcije (tj. vsote verjetnosti v levem repu Poissonove porazdelitve) in komplementarne kumulativne porazdelitve χ^2 (tj. ploščine, ki bi jo izračunali z integralom, pod desnim repom porazdelitve χ^2). Ker je enota, na katero štejemo število dogodkov (v prostoru: osebkov ali predmetov), praviloma arbitrarna, torej lahko s spremembo enote v opazovano obdobje (v prostoru: področje)

izenačimo λ z opaženim k, oziroma ker λ dostikrat ocenimo na podlagi enega samega poskusa (štetja), se v praksi eksaktni interval zaupanja največkrat nanaša na opaženo število dogodkov (osebkov, predmetov). Meji intervala zaupanja (SM – spodnja, ZM – zgornja) za $\lambda = k$ sta odvisni le od stopnje zaupanja (α) in znašata:

$$SM = \frac{\chi^2\left(P = \frac{\alpha}{2}; df = 2k\right)}{2}$$
$$ZM = \frac{\chi^2\left(P = 1 - \frac{\alpha}{2}; df = 2(k+1)\right)}{2} \quad . \qquad [5]$$

Za izračun po eksaktni metodi torej potrebujemo le tabelo porazdelitve χ^2 (ki jo najdemo bodisi na koncu vsakega učbenika statistike bodisi na spletu) oziroma uporabimo programje, ki ima tabelo "vgrajeno" (v Excelu 2007/2010 uporabimo funkcijo CHISQ.INV). Eksaktni 95% interval zaupanja je za 0 do 7 opaženih dogodkov naveden v tabeli 1. Izračunan je po obrazcu [5] na 4. delovnem listu priloženega Excelovega delovnega zvezka, ki omogoča izračun za poljubno stopno zaupanja in poljubno število dogodkov (vnos v oranžno osenčeni celici).

Tabela 1 Eksaktni 95% interval zaupanja za Poissonovo porazdeljeno število dogodkov.

k oz. λ	spodnja meja	zgornja meja
0	0	3,6889
1	0,0253	5,5716
2	0,2422	7,2247
3	0,6187	8,7673
4	1,0899	10,2416
5	1,6235	11,6683
6	2,2019	13,0595
7	2,8144	14,4227

Eksaktna metoda je znana že dolgo,[34] a to – kot je v statistiki običajno – še zdaleč ni konec zgodbe. Najnovejši članek o intervalih zaupanja za povprečje Poissonove slučajne spremenljivke med seboj primerja kar devetnajst metod,[48] a za začetek bosta dve dovolj.

Preverjanje prileganja s statističnimi testi

Če imamo podatke iz več poskusov, jih lahko uredimo v frekvenčno porazdelitev, tj. ugotovimo opažene frekvence f_o za vse opažene vrednosti k. Če hkrati ocenimo λ, lahko izračunamo Poissonovo verjetnost, s tem pa tudi pričakovano frekvenco f_p za vsakega od možnih izidov. Na podlagi tega lahko ocenimo oziroma statistično testiramo prileganje (ang. *goodness-of-fit*) teoretične porazdelitve empirični (oziroma obratno – s praktičnega vidika je vseeno, kako rečemo, z epistemološkega pa ne, a v to razpravo se tu ne bomo spuščali). Ker gre za diskretno slučajno spremenljivko, je običajna izbira test χ^2, pri katerem se testna statistika

$$X^2 = \sum_k \frac{(f_o - f_p)^2}{f_p} \qquad [6]$$

pod ničelno domnevo porazdeljuje (asimptotično) po porazdelitvi χ^2 z dvema prostostnima stopnjama manj, kot je opaženih različnih vrednosti k. Eno prostostno stopnjo pobere dejstvo, da je pri danem skupnem številu podatkov (n) ob poznavanju ostalih frekvenc za eno celico tabele frekvenca v naprej določena; drugo pa pobere dejstvo, da smo na podlagi opaženih podatkov ocenili en parameter porazdelitve, ki jo prilegamo podatkom (λ).

Upoštevati je potrebno še, da morajo izidi pokriti celotno zalogo vrednosti opazovane spremenljivke in da naj ne bi bilo celic s pričakovano frekvenco pod 5 (oziroma naj bi jih bilo čim manj, sicer porazdelitev χ^2 ni dober približek za [6]), zato se najvišje vrednosti pogosto združi (oblikuje en izid oziroma celico za $k \geq$ neki vrednosti). Računski primer je predstavljen v tabeli 2 in na 5. delovnem listu priloženega Excelovega delovnega zvezka. Vidimo, da se porazdelitev števila nesreč, ki jih je imel posamezni pilot zaradi lastne napake,[46] praktično povsem ujema s Poissonovo, kar pomeni, da je naključje smiseln model za tovrstne nesreče. To govori v prid postopkov izbora, šolanja

in razporejanja pilotov, saj podatki kažejo, da med piloti ni bilo razlik v usposobljenosti in izpostavljenosti nesrečam.

Tabela 2 Primer preverjanja prileganja Poissonove porazdelitve opaženim podatkov s testom χ^2 – število nesreč zaradi napak pri vojaških pilotih.

k	f_o	P_p	$f_p = nP_p$	$(f_o - f_p)^2 / f_p$
0	662	0,691	661	0,001
1	242	0,256	245	0,026
2	47	0,047	45	0,069
3	6	0,006	*6	0,003
$n = 957$		$\sum P_p = 1$	$\sum f_p = n$	$X^2 = 0,010$
$\bar{k} = 0,370$				$df = 4 - 2 = 2$
$s_k = 0,369$				$p = 0,951$

* za $k \geq 3$

Sorodna statistika je *indeks razpršenosti* (ang. *index of dispersion*). Izpeljavo bomo izpustili, ker je predolga (čeprav ni težka), bistveno pa je, da indeksa razpršenosti ne računamo iz podatkov, urejenih v frekvenčno porazdelitev, pač pa iz vsake realizacije opazovane slučajne spremenljivke posebej (tj. iz posameznih $k_i, i = 1..n$). V primeru Poissonove porazdelitve ga imenujemo *Poissonov indeks razpršenosti* (z ang. kratico PID) in ima obliko

$$PID = \frac{\sum_{i=1}^{n} (k_i - \bar{k})^2}{\bar{k}}.$$

Ker se indeks razpršenosti (asimptotično) pod ničelno domnevo porazdeljuje kot χ^2 z $n-1$ prostostnimi stopnjami, z izračunom PID izvedemo *test Poissonove razpršenosti* (ang. *Poisson dispersion test*). Z njim primerjamo razpršenost opaženih podatkov z razpršenostjo, ki bi jo pričakovali pri Poissonovi porazdelitvi. Tudi pri tem testu velja omejitev glede asimptotičnega približka – primeren je le, če je povprečje opazovanj ($\bar{k}$ kot cenilka λ) večje od 3 (ali še raje od 5), je pa že Sir Ronald Aylmer Fisher (pionir statistike tudi glede permutacijskih metod) razvil njegovo eksaktno obliko (ki se jo najde v učbenikih in statističnem programju). Računski primer asimptotične oblike testa Poissonove razpršenosti je predstavljen v tabeli 2 in na 6. delovnem listu priloženega Excelovega delovnega zvezka. Podatki o številu prometnih nesreč v enem mesecu ne odstopajo statistično značilno od Poissonove porazdelitve, kar govori v prid interpretaciji, da se nesreče dogajajo slučajno.

Tabela 3 Primer preverjanja prileganja Poissonove porazdelitve opaženim podatkov s testom Poissonove razpršenosti – število prometnih nesreče v nekem mestu za zaporedne mesece.

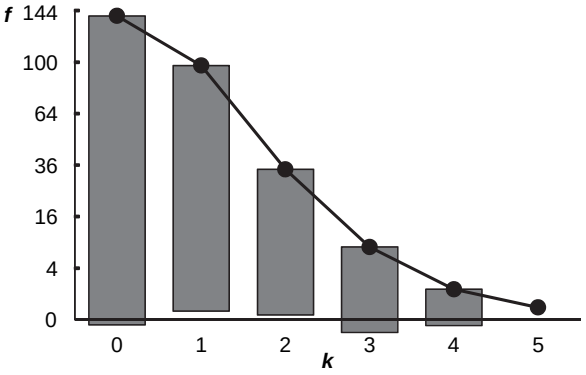
k	$(k_i - \bar{k})^2 / \bar{k}$
11	0,397
8	0,133
9	0,001
4	2,858
12	0,924
10	0,089
16	5,232
9	0,001
6	1,056
6	1,056
$\bar{k} = 9,100$	$PID = 1,747$
	$df = 10 - 1 = 9$
	$p = 0,228$

Podobno kot je veliko metod za ocenjevanje intervala zaupanja, je tudi testov prileganja Poissonove porazdelitve še mnogo in posvečene jim je veliko statistične literature.[29,40] Eden od tovrstnih testov temelji na testni statistiki Kolmogorova in Smirnova (tj. največji absolutni razliki med opaženo in pričakovano verjetnostjo izida). Skupina testov, ki je prav tako izpeljana iz testa za zvezne spremenljivke, temelji na testni statistiki Cramérja in von Misesa (tj. uteženi vsoti kvadriranih razlik med opaženo in teoretično pričakovano kumulativno porazdelitveno funkcijo). Nadaljnja možnost so testi na podlagi testa razmerja verjetij (ang. *likelihood ratio test*). Testov prileganja Poissonove porazdelitve empiričnim podatkom je še nekaj skupin, a bodi dovolj naštevanja – raje si oglejmo grafične metode (saj vsi poznamo tisto o sliki in tisoč besedah ali pa številkah).

Grafično preverjanje prileganja

Malo je področij sodobne statistike, ki niso povezana z imenom in delom Johna Wilderja Tukeya, 1915-2000, po mnenju mnogih (vključno z avtorjem tega gradiva) enega najpomembnejših in najbolj vsestranskih znanstvenikov vseh časov (med drugim je skoval izraza *bit* in *software* ter razvil *hitro Fourierovo transformacijo* – FFT). Med njegovimi prispevki k prikazu podatkov je najbolj znan *škatlasti grafikon kvantilov* (ang. *box-and-whiskers plot* ali skrajšano *boxplot*), za jasnejši pogled na Poissonovo porazdelitev pa si je Tukey izmislil *obešeni korenogram* (ang. *hanging rootogram*). Predstavil ga je v dveh epohalnih delih – članku o metodah za prikaz oziroma grafično analizo podatkov[59] in monografiji o eksploratorni analizi podatkov.[60] Gre za stolpčni grafikon, ki je "obešen" zato, ker dolžine stolpcev ne odmerimo od nič navzgor, pač pa od pričakovane frekvence navzdol (torej se stolpci končajo nad vodoravno osjo, če je opažena frekvenca manjša od pričakovane, in pod njo, če je večja). Na ta način so odstopanja jasnejša, saj vidimo njihov vzorec okoli vodoravne osi namesto okoli krivulje. Če grafikon uporabimo za Poissonovo porazdelitev, izračunamo pričakovane frekvence v skladu z njo, sicer pa gre za splošen princip, ki je uporaben za primerjavo dveh poljubnih porazdelitev (lahko tudi dveh vzorcev). "Korenskost" pomeni, da je navpična os v kvadratnokorenskem merilu (tj. potenčnem s potenco 0,5), kar poudari odstopanja v repih (tj. pri manjših pričakovanih frekvencah – podobno kot test χ^2). Tudi tu gre za splošni princip, za katerega je Tukeya sicer navdahnila Poissonova porazdelitev, a je uporaben za katerikoli prikaz porazdelitve s stolpci (kar vključuje histogram) ali s krivuljo (frekvenčni poligon, zglajeni grafikon gostote).

Obešeni korenogram na sliki 3 prikazuje podatke iz tabele 4, ki jih bomo kmalu spoznali v okviru zgodovinskega pregleda. Izdelava v Excelu (7. delovni list priloženega delovnega zvezka) zahteva nekaj "trikov" (trije podatkovni nizi namesto enega za prikaz opaženih frekvenc, "umetna" navpična os) in naprednejših veščin (kombinacija treh vrst grafikonov, spremenjene oznake podatkov). Večjih oziroma sistematičnih odstopanj od Poissonove porazdelitve ni opaziti.



Slika 3 Obešeni korenogram za preverjanje ujemanja empiričnih podatkov s Poissonovo porazdelitvijo (podatki iz tabele 4 – smrti zaradi konjskih brc v pruski vojski; povezane črne pike – teoretična porazdelitev, sivi stolpci – opažene frekvence).

Tukeyev študent in dolgoletni sodelavec David Caster Hoaglin (1944-) je razvil *grafikon poissonskosti* (ang. *Poissonness plot*).[37] Izhaja iz odnosa, ki ga dobimo iz obrazca [1] z logaritmiranjem obeh strani in nekaj preurejanja ob predpostavki, da so vse opažene frekvence (f_{o_k}) enake pričakovanim:

$$\ln \frac{k!\, f_{o_k}}{n} = -\lambda + k \ln \lambda \ .$$

Če levo stran zgornje enačbe označimo kot funkcijo $\varphi(f_{o_k}) = \ln(k!\, f_{o_k}/n)$, dobimo pravilo, da je za Poissonovo porazdelitev graf $\varphi(f_{o_k})$ v odvisnosti od k premica (z odsekom na ordinati $-\lambda$ in koeficientom $\ln \lambda$). Čim bližje so točke na takem grafikonu premici, tem bolj je opažena porazdelitev podobna Poissonovi. Hoaglin in Tukey[38] sta grafikon poissonskosti dodelala z oceno in prikazom intervalov zaupanja (IZ) za $\varphi(f_{o_k})$ in popravkom opaženih frekvenc. Če nekoliko poenostavimo (izpustimo možnost $f_o = 0$ in dodatni popravek IZ za $f_o = 1$ ter pozabimo na problem zmanjšanja stopnje zaupanja pri sočasnem ocenjevanju več intervalov zaupanja), priporočata

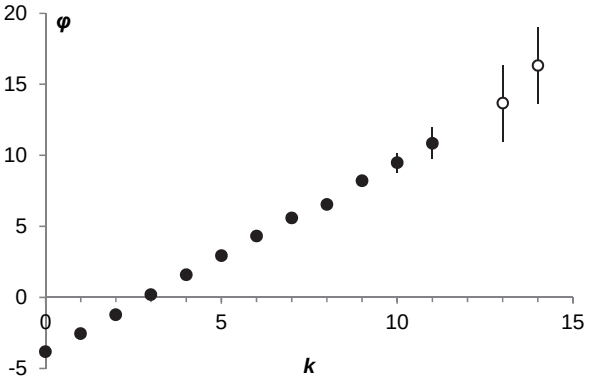
$$\varphi\left(f_{o_k}^*\right) = \ln\frac{k!\, f_{o_k}^*}{n}$$

$$f_{o_k}^* = \begin{cases} f_{o_k} - 0{,}67 - 0{,}8\, f_{o_k}/n, & f_{o_k} \geq 2 \\ 1/e, & f_{o_k} = 1 \end{cases}$$

$$95\%IZ = \varphi\left(f_{o_k}^*\right) \pm h(k) \qquad [7].$$

$$h(k) = \frac{1{,}96\sqrt{1 - f_{o_k}/n}}{\sqrt{f_{o_k} - \left(0{,}25\, f_{o_k}/n + 0{,}47\right)\sqrt{f_{o_k}}}}$$

Dodala sta še priporočilo, da se točke, kjer je $f_{o_k} = 1$, nariše z drugačno oznako, saj je variabilnost φ zaradi vzorčenja pri tako majhnih frekvencah velika in je zato tem točkam potrebno dati manjšo težo pri presojanju linearnosti. Na sliki 4 je prikazan tako dodelan grafikon poissonskosti za podatke o radioaktivnem razpadu polonija (tabela 6). Ujemanje s Poissonovo porazdelitvijo je dobro, saj ležijo točke praktično na premici. Samo pri $k = 8$ je opažena frekvenca nekoliko "prenizka" (kar je jasno razvidno tudi iz tabele 6); pri $k = 14$ ni pomembnega odstopanja, čeprav je točka nad "trendom", saj ta seka interval zaupanja.



Slika 4 Grafikon poissonskosti [7] (podatki iz tabele 6 – radioaktivni razpad polonija).

Tretji grafični pristop k preverjanju prileganja Poissonovi porazdelitvi je Ordov grafikon.[47] Temelji na linearni zvezi

$$kP_k / P_{k-1} = a + bk,$$

ki povezuje opaženi delež (kot oceno verjetnosti) za dve zaporedni vrednosti števila dogodkov ($k-1$ in k) pri štirih sorodnih diskretnih porazdelitvah (Poissonovi, binomski, negativni binomski in logaritemski). Pri Poissonovi porazdelitvi je koeficient b enak nič, ordinatni odsek a pa enak parametru λ. Kako je pri ostalih treh porazdelitvah, se ne bomo vprašali, in primera tudi ne bo, ker bomo porazdelitve, povezane s Poissonovo, sistematično obravnavali šele v zadnjem delu gradiva in ker presojanje o vrednostih a in b (pri čemer za orientacijo služi utežena linearna regresija) oziroma odločanje med alternativnimi modeli presega namen tega gradiva.

Zgodovina

Nekajkrat smo se že bežno ozrli v zgodovino (Wald, Fisher, piloti v 2. svetovni vojni, Tukey), sedaj pa se za nekaj časa povsem posvetimo ključnim osebnostim in podatkom iz zgodovine Poissonove porazdelitve.

Siméon Denis Poisson

Poissonova porazdelitev je dobila ime po francoskem matematiku in fiziku Siméonu Denisu Poissonu (1781-1840; slika 5). Poleg številnih prispevkov k različnim področjem znanosti je znan po reku, da je življenje dobro le za dve stvari – odkrivati matematiko in poučevati matematiko. Njegovi mentorji oziroma sodobniki so bili številni vsestranski pionirji znanosti – Lagrange, Laplace, Legendre in Fourier.[31]

Porazdelitev je opisal v svojem delu iz leta 1837 o sodnih odločitvah v kazenskopravnih in civilnopravnih zadevah.[50] Kljub Poissonovemu ugledu in vplivu delo ni pritegnilo širše pozornosti, zato so isto porazdelitev pomembni raziskovalci kasneje še nekajkrat ponovno "odkrili". Po drugi strani pa je njen poseben primer kratko opisal že Abraham de Moivre leta 1711 (v prvi znanstveni razpravi o verjetnosti dogodkov pri igrah na srečo[44]).

Slika 5 Siméon Denis Poisson (v duhu redkih dogodkov edini brez brkov med možmi na slikah v tem gradivu).

Prusija

V zgodovini znanosti veljajo za najbolj znan primer Poissonove porazdelitve smrti zaradi konjskih brc v pruski vojski, ki jih je kot primer *zakona majhnih števil* (nem. *Das Gesetz der Kleinen Zahlen*, kot je naslovil svojo knjigo o Poissonovi porazdelitvi iz leta 1898[27]) navedel Ladislaus von Bortkiewicz (slika 6). Podatki se nanašajo na 14 rodov vojske v obdobju 20 let. Vseh smrti je bilo 280. Kot je razvidno iz tabele 4 in 8. delovnega lista priloženega Excelovega delovnega zvezka, se porazdelitev števila smrti na leto ujema s Poissonovo.

Tabela 4 Porazdelitev števila smrti zaradi konjskih brc v letu v pruski vojski v obdobju 1875-1894 in njeno ujemanje s Poissonovo porazdelitvijo (tabelograf in test χ^2).

Št. smrti v letu	Opaženo	Pričakovano	$(O-P)^2/P$
0	144	139,0	0,177
1	91	97,3	0,412
2	32	34,1	0,125
3	11	7,9	1,171
4	2	*1,6	0,094

* za 4+

			Test χ^2:
Skupaj	280	X^2 =	1,979
Povprečje	0,700	p (df=5-2=3) =	0,577
Varianca	0,760		

Po von Bortkiewiczu (ki je bil poljskega rodu, a je večinoma živel in delal v Nečiji, kjer so tedaj njegov priimek pisali kot Bortkewitsch) se imenuje katedra za statistiko na Humboldtovi univerzi v Berlinu, kjer dandanes razvijajo najsodobnejše statistične in ekonometrične modele. Bil je učenec še enega nemškega pionirja statistike in ekonomije, Wilhelma Lexisa (slika 6), ki je razvil test Poissonove razpršenosti.

Slika 6 Ladislaus von Bortkiewicz (levo) in Wilhelm Lexis (desno).

Poleg smrti zaradi konjskih brc je von Bortkiewicz podrobno obravnaval še tri tragične pojave, ki se ravnajo po Poissonovi porazdelitvi: samomore otrok (število na leto, v obdobju 1869-1893), samomore žensk (število na leto za 8 nemških zveznih držav, v obdobju 1881-1894) in smrti zaradi delovnih nezgod (število na leto za 11 združenj za poklicno zavarovanje, v obdobju 1886-1894). A redki pojavi so lahko tudi dobrodošli, kot npr. uspešna večplodna nosečnost. Iz istega področja in obdobja kot von Bortkiewiczevi primeri so (iz drugega vira[52]) npr. znani tudi podatki o rojstvih četverčkov. V Prusiji jih je bilo v obdobju 69 let skupaj 109. Porazdelitev števila rojstev četverčkov na leto se prav tako zelo dobro ujema s Poissonovo (tabela 5 in 9. delovni list priloženega Excelovega delovnega zvezka).

Tabela 5 Porazdelitev števila rojstev četverčkov v Prusiji v obdobju 69 let in njeno ujemanje s Poissonovo porazdelitvijo (tabelograf in test χ^2).

Četverčkov v letu	Opaženo	Pričakovano	$(O-P)^2/P$
0	14	14,2	0,003
1	24	22,5	0,106
2	17	17,7	0,031
3	9	9,3	0,012
4	2		
5	2	*5,2	0,012
6	1		

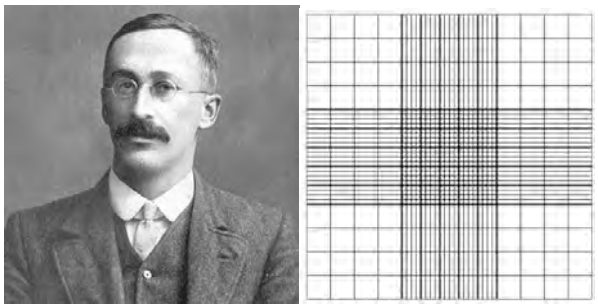
* za 4+

			Test χ^2:
Skupaj	69		
Povprečje	1,580	X^2 =	0,164
Varianca	1,722	p (df=5-2=3) =	0,983

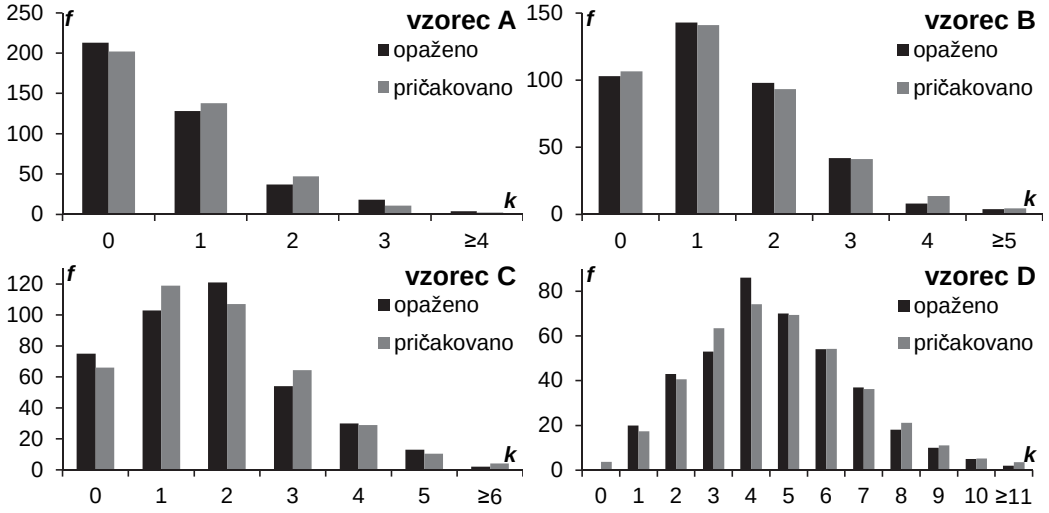
Student

Angleški matematik, kemik in pivovarski strokovnjak William Sealy Gosset (1876-1937; slika 7), ki je svoja statistična odkritja objavljal pod psevdonimom Student, je znan skoraj vsakemu študentu po testu t za primerjavo povprečij. Tehnološko delo v Guinessovi pivovarni, veselje do statistike in družinska povezanost z "očetom statistike" R. A. Fisherjem (ki je bil njegov tast) pa so Gosseta vodili tudi do prve uporabe Poissonove porazdelitve na področju biologije. Preučeval je napake pri štetju kvasovk s hemocitometrom (mikroskopom, pod katerim kanemo opazovano raztopino na kvadratno mrežo – slika 7).[57] Ugotovil je, da število celic na kvadratek sledi Poissonovi porazdelitvi (čeprav je ni poznal in jo je sam izpeljal), saj je verjetnost, da določena celica pade v določen kvadratek, majhna (1/400, saj je uporabljal hemocitometer z mrežo 20×20 kvadratkov), vseh celic v kapljici raztopine pa je veliko. Kot je razvidno iz njegovih štirih vzorcev[35] (A-D; slika 8 in 10. delovni list priloženega Excelovega delovnega zvezka), je ujemanje zelo dobro.

- V vseh štirih vzorcih sta si povprečje in varianca zelo podobna: A – 0,683 in 0,812; B – 1,323 in 1,283; C – 1,800 in 1,960; D – 4,680 in 4,485.
- Odstopanja podatkov od Poissonove porazdelitve (z izjemo vzorca A) niso statistično značilna.
  - Vrednosti p, dobljene s testom χ^2, so: za vzorec A 0,018; za vzorec B 0,531; za vzorec C 0,187; in za vzorec D 0,580.
  - Test Poissonove razpršenosti da za vzorec D (ki ima edini dovolj veliko povprečje za uporabo tega testa) $p = 0{,}734$.



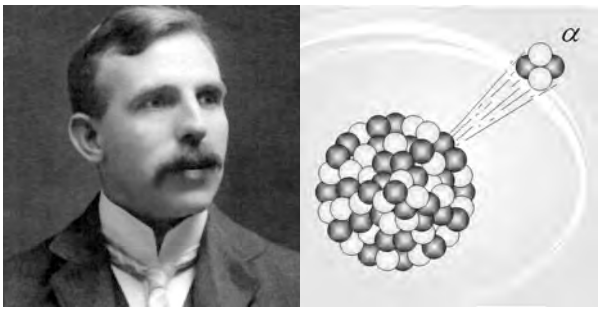
Slika 7 William Sealy Gosset – Student (levo) in mreža hemocitometra (desno).



Slika 8 Opažena porazdelitev števila celic kvasovk na kvadratek hemocitometra (z mrežo 200×200) in najbolje prilegajoča se Poissonova porazdelitev za štiri Studentove vzorce (A-D).

Radioaktivni razpad

Ernest Rutherford (1871-1937; slika 9), na Novi Zelandiji rojeni britanski kemik in fizik, ki velja za očeta jedrske fizike, si je prislužil Nobelovo nagrado (1908 za za kemijo), visok plemiški naziv in še mnoge druge najvišje časti, bil pa je tudi mentor štirim nobelovcem. S Poissonovo porazdelitvijo je povezano njegovo odkritje delcev, žarkov oziroma radioaktivnega razpada α in β.



Slika 9 Ernest Rutherford, 1. (in zadnji) baron Rutherford Nelsonski (levo), in shematičen prikaz radioaktivnega razpada alfa (desno).

Leta 1910 so Rutherford, Geiger in Bateman šteli število delcev α, ki jih je oddala tanka plast polonija v 2608 zaporednih intervalih dolžine ⅛ minute.[53] Menili so, da mora dobljena porazdelitev slediti Poissonovi (ki je sicer niso poznali in so jo izpeljali na novo), saj je bilo radioaktivnih atomov veliko, za vsakega od njih pa je verjetnost, da razpade v osmini minute, zelo majhna. Podatke in njihovo prileganje Poissonovi porazdelitvi (preverjeno s testom χ^2) prikazujeta tabela 6 in 11. delovni list priloženega Excelovega delovnega zvezka.

Omenimo še, da če namesto števila radioaktivnih razpadov na časovno enoto spremljamo čase med razpadi, dobimo eksponentno porazdelitev, a o tem več nekoliko kasneje.

Tabela 6 Porazdelitev števila izsevanih delcev α na osmino minute v poskusu Rutherforda, Geigerja in Batemana iz leta 1910 in njeno ujemanje s Poissonovo porazdelitvijo (tabelograf in test χ^2).

Št. delcev α	Opaženo	Pričakovano	$(O-P)^2/P$
0	57	54	0,167
1	203	210	0,233
2	383	407	1,415
3	525	525	0,000
4	532	508	1,134
5	408	394	0,497
6	273	254	1,421
7	139	141	0,028
8	45	68	7,779
9	27	29	0,138
10	10	11	0,091
11	4	4	0,000
12	0		
13	1	3	0,333
14	1		
15+	0		

Skupaj	2608		**Test χ^2:**
Povprečje	3,872	X^2 =	13,238
Varianca	3,695	p (df=13-2=11) =	0,278

Leteče bombe V-1

Pomemben in znan primer uspešne uporabe Poissonove porazdelitve je povezan z letečimi bombami V-1 (zaradi srhljivega hrupa pulznega reaktivnega motorja so jih Angleži poimenovali *buzz bombs*), s katerimi je nacistična Nemčija v 2. svetovni vojni napadala London (in še nekaj drugih mest, zlasti Antwerpen v Belgiji) od poletja 1944 do pomladi 1945 (slika 10).

Slika 10 Nemška leteča bomba V-1 iz 2. svetovne vojne (levo) in primer razdejanja, ki ga je padec take "hrupne bombe" povzročil v Londonu (desno).

Ključno vprašanje obrambe je bilo, ali V-1 padajo po naključju ali so vodene k določenim ciljem. Če bi veljalo prvo, bi morala biti dvorazsežna prostorska porazdelitev krajev zadetkov naključna, torej porazdelitev števila zadetkov na enoto

ploščine zelo podobna Poissonovi, v drugem primeru pa bi bila prostorska porazdelitev krajev zadetkov gručasta (kot je pojasnjeno v razdelku o prostorskem vidiku in prikazano na sliki 2). Razdelitev 144 km^2 velikega področja južnega Londona na 576 kvadrantov velikosti ¼ km^2 je potrdila naključnost oziroma Poissonovo porazdelitev,[32] kar je bilo zelo pomembno za strategijo obrambe (vključno z dezinformacijami o učinkih bomb, ki so jih dostavljali Nemcem preko mreže dvojnih agentov). Podatke in njihovo prileganje Poissonovi porazdelitvi prikazujeta tabela 7 in 12. delovni list priloženega Excelovega delovnega zvezka.

Tabela 7 Porazdelitev števila nemških letečih bomb V-1, ki so padle na južni London v zadnjem obdobju 2. svetovne vojne, po kvadrantih in njeno ujemanje s Poissonovo porazdelitvijo (tabelograf in test χ^2).

Bomb v sektorju	Opaženo	Pričakovano	$(O-P)^2/P$
0	229	226,7	0,022
1	211	211,4	0,001
2	93	98,5	0,311
3	35	30,6	0,626
4	7	7,1	0,003
5	0		
6	0	*1,6	0,206
7	1		

* za 5+

Skupaj	576		**Test χ^2:**
Povprečje	0,932	X^2 =	1,169
Varianca	0,969	$p(df=6-2=4)$ =	0,883

Nadgradnja

Povezavo Poissonove porazdelitve z binomsko porazdelitvijo in porazdelitvijo χ^2 smo že spoznali, v nadaljevanju pa bomo spoznali še nekaj zahtevnejših porazdelitev, povezanih s Poissonovo, vključno s porazdelitvami zmesi in večrazsežnimi porazdelitvami. Nato bomo spoznali, kateri statistični test velja za najpreprostejšega in zakaj. Dotaknili se bomo tudi točkovnih procesov, napovednih (regresijskih) modelov in kontrolnih kart.

Imen in pojmov, ki jih bomo pri tem srečali, bo veliko, navedeni pa so predvsem kot spodbuda, napotek in pripomoček za lažje iskanje nadaljnjih virov. Poissonova porazdelitev namreč predstavlja le ena vratca za vstop (ali vsaj okence za pogled) v ogromni svet statistike.

Povezane porazdelitve

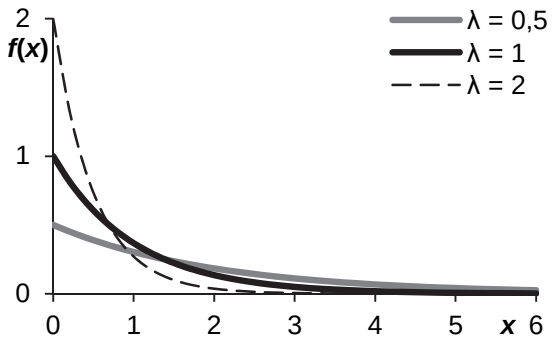
Če kje, potem v svetu verjetnostnih porazdelitev nedvomno velja, da je (skoraj) vse povezano s (skoraj) vsem,[42] zato je s Poissonovo povezanih veliko porazdelitev. Začnimo pri časovnem vidiku in napovedjo, da je s Poissonovo neločljivo povezana ena od temeljnih zveznih porazdelitev. To je *eksponentna porazdelitev* (imenovana tudi negativna eksponentna ali porazdelitev Poissonovega toka). Njena gostota verjetnosti je

$$X \sim Exp(\lambda) \Leftrightarrow f(x) = \lambda e^{-\lambda x}, x \geq 0,$$

(kumulativna) porazdelitvena funkcija pa

$$F(x) = P(x \leq X) = \int_0^x y e^{-\lambda u} du = \left[-e^{-\lambda u} \right]_0^x = 1 - e^{-\lambda x}. \quad [8]$$

Oblika eksponentne porazdelitve je vedno enaka, le njena razpršenost pada z naraščanjem λ (slika 11 in 13. delovni list priloženega delovnega zvezka). Njeno povprečje izračunamo iz definicije $E(x) = \int_0^\infty x f(x) dx$ s substitucijo $y = \lambda x$, s čimer dobimo $E(x) = 1/\lambda$. Povprečje torej narašča z manjšanjem parametra λ. Modus je vedno pri $x = 0$ (kjer je $f(x) = \lambda$), mediana pa je pri $F(x) = 1/2$, torej pri $(-\ln 1/2)/\lambda = \ln 2/\lambda = 0{,}693/\lambda$, kar je približno na dveh tretjinah razdalje med modusom in povprečjem. Momente eksponentne porazdelitve bi lahko izračunali z uporabo rodovne funkcije, a kar povejmo, da je varianca $1/\lambda^2$, asimetričnost pa $+2$. Izrazita desna asimetričnost se sklada s vrstnim redom mer srednje vrednosti (modus < mediana < aritmetična sredina).



Slika 11 Eksponentna porazdelitev za tri vrednosti parametra λ.

In kako je eksponentna porazdelitev povezana s Poissonovo? Vrnimo se k radioaktivnemu razpadu in se spomnimo, da se število razpadov v t sekundah porazdeljuje po Poissonovi porazdelitvi s parametrom λt. Po obrazcu [1] je verjetnost, da v t sekundah ne bo nobenega razpada, $P(0) = e^{-\lambda t}$, to pa je hkrati verjetnost, da bomo morali čakati več kot t sekund, da bo prišlo do razpada. Verjetnost, da bomo v času t dočakali radioaktivni razpad, je torej $P(T \le t) = 1 - e^{-\lambda t}$, kar pa je isto kot [8], tj. porazdelitvena funkcija eksponentne porazdelitve. Za čas med dvema razpadoma (T) torej velja eksponentna porazdelitev s parametrom λ. Jasen je tudi pomen parametra λ, ki je skupen obema porazdelitvama: več, ko je dogodkov, večje je povprečje njihovega števila na časovno enoto (tj. povprečje Poissonove porazdelitve) in manj časa v povprečju mine med dvema dogodkoma (manjše je povprečje eksponentne porazdelitve, ki je $1/\lambda$).

Eksponentni porazdelitvi smo namenili več pozornosti kot je bomo ostalim v nadaljevanju, saj s Poissonovo porazdelitvijo predstavljata "dve plati iste medalje" (ena v diskretnem svetu števila dogodkov, druga v zveznem svetu časov). Lahko bi rekli, da podobna dvojnost velja za statistiko in stohastične procese (imenovane tudi slučajni procesi, ang. *stochastic processes* oziroma *random processes*) kot pristopa k slučajnim pojavom. Podmnožica slučajnih procesov so točkovni procesi (ang. *point processes*), med katere sodi tudi *Poissonov proces*, ki smo ga pravkar spoznali. Zanj je namreč značilno, da je čas med dvema točkama eksponentno porazdeljen. Praktično vse, kar smo spoznali doslej v zvezi z Poissonovo porazdelitvijo, bi lahko izpeljali in tolmačili tudi z vidika slučajnih procesov, a za večino nematematikov (ne pa za tehnike, inženirje in računalničarje, ki jim je ta pristop bolj domač od statističnega) bi bilo to mnogo bolj zahtevno, saj ne bi šlo brez diferencialnih enačb.

Z eksponentno – in torej tudi s Poissonovo – je povezana *Erlangova porazdelitev*, ki si jo lahko zapomnimo kot porazdelitev vsote neodvisnih eksponentno porazdeljenih slučajnih spremenljivk. V jeziku čakalnih vrst to pomeni, da opisuje porazdelitev časa, ki preteče, da opazimo k dogodkov. Erlangova porazdelitev predstavlja poseben primer porazdelitve gama (Γ), pri kateri je parameter oblike (ang. *shape parameter*; taisti k) pozitivno celo število.

Pri prostorskem vidiku Poissonove porazdelitve smo spoznali pojem nadrazpršenosti. V resnici je bolj kot pri cvetlicah na travniku prisoten pri porazdelitvi nečesa oziroma nekoga mnogo manj prijetnega za ljudi in nam bližje sorodne živali – zajedavcev. Na področju parazitologije je namreč že dolgo znano, da se v nekaterih gostiteljih zadržuje mnogo več zajedavcev, kot bi jih pričakovali, če bi se zajedavci med gostitelje razporejali po naključju (in hkrati so potencialni gostitelji, prosti nadležne zalege, številčnejši, kot bi na ta način pričakovali). Kot model porazdelitve števila zajedavcev v različnih populacijah se zato uporablja *negativna binomska porazdelitev* (ang. *negative binomial distribution*), pri čemer parazitologi posebno pozornost namenjajo gručenju (ang. *crowding*).[51] Obrazec za verjetnostno funkcijo negativne binomske porazdelitve je podoben kot za binomsko porazdelitev, opisuje pa število uspešnih izidov (tj. enk) v nizu Bernoullijevih poskusov preden dosežemo določeno število neuspešnih izidov (ničel). Njen odnos do binomske porazdelitve je torej analogen odnosu eksponentne porazdelitve do Poissonove. Ena od alternativnih definicij negativne binomske porazdelitve pa pravi, da je to

Poissonova porazdelitev, katere parameter λ je slučajna spremenljivka, porazdeljena po posebni obliki porazdelitve gama.

Na področju aktuarstva je pomembna *sestavljena Poissonova porazdelitev* (ang. *compound Poisson distribution*). Definirana je kot porazdelitev vsote medsebojno neodvisnih in enako porazdeljenih (poljubno, čeprav praviloma z omejitvijo, da so nenegativne) slučajnih spremenljivk, pri čemer je število slučajnih spremenljivk, ki jih seštejemo, Poissonovo porazdeljena slučajna spremenljivka. Uporabna je pri ocenjevanju tveganja zaradi predvidenih odškodnin in s tem pri določanju zavarovalnin.[41] Zanimivo je, da je poseben primer sestavljene Poissonove porazdelitve negativna binomska porazdelitev, ki jo dobimo, če so slučajne spremenljivke, ki jih seštevamo, porazdeljene po logaritemski porazdelitvi (ki sodi med manj znane diskretne porazdelitve).

Mimogrede, porazdelitve imajo res "gosto socialno mrežo", mar ne? Ampak s porazdelitvijo števila povezav v analizi omrežij, ki je sicer včasih vsaj teoretično lahko Poissonova, se ne bomo ubadali. Pač pa spoznajmo neposredno razširitev Poissonove porazdelitve, ki sliši na ime Conway-Maxwell-Poissonova porazdelitev (skrajšano *COM-Poissonova porazdelitev*). V primerjavi s Poissonovo ima dodaten parameter ν, ki si ga lahko predstavljamo kot faktor, ki omogoča, da se hitrost (npr. radioaktivnega) razpada spreminja. Njena posebna primera sta Poissonova porazdelitev (če je $\nu = 1$) in geometrična porazdelitev (če je $\nu = 0$), limitni primer pa Bernoullijeva porazdelitev (če gre $\nu \to \infty$). O aktualnosti COM-Poissonove porazdelitve se bomo prepričali v zadnjih dveh razdelkih – o Poissonovi regresiji in o kontrolnih kartah.

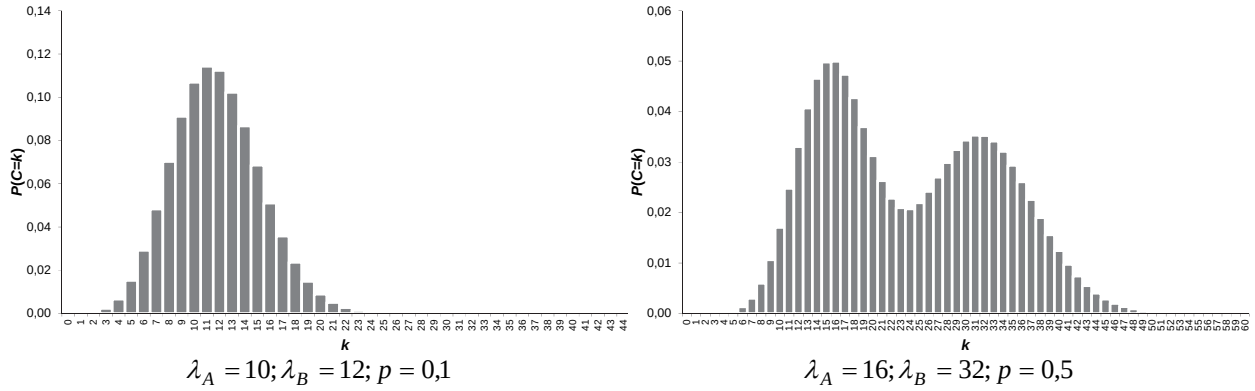
Poissonove zmesi

Najpreprostejši primer Poissonove zmesi je porazdelitev slučajne spremenljivke, ki jo dobimo tako, da se pri vsakem poskusu po slučaju – tj. z realizacijo Bernoullijeve slučajne spremenljivke z danim parametrom $p \in (0;1)$ – odločimo, ali bomo izžrebali slučajno vednost iz ene Poissonove porazdelitve (s parametrom λ_A) ali iz druge (s parametrom λ_B). Verjetnostna funkcija take slučajne spremenljivke C je

$$P(C=k) = \frac{pe^{-\lambda_A}\lambda_A^k + (1-p)e^{-\lambda B}\lambda_B^k}{k!}, \qquad [9]$$

povprečje in varianca pa[43]

$$E(C) = p\lambda_A + (1-p)\lambda_B,$$
$$Var(C) = p\lambda_A(\lambda_A+1) + (1-p)\lambda_B(\lambda_B+1) - [p(\lambda_A-\lambda_B)+\lambda_B]^2.$$



$\lambda_A = 10; \lambda_B = 12; p = 0{,}1$ $\lambda_A = 16; \lambda_B = 32; p = 0{,}5$

Slika 12 Dva primera porazdelitve zmesi dveh Poissonovih porazdelitev [9].

Z interaktivnim izračunom in prikazom na 14. delovnem listu priloženega Excelovega delovnega zvezka lahko ugotavljamo, kako se spreminja oblika porazdelitve C za različne kombinacije vrednosti λ_A (dopuščen je vnos med 0 in 20), λ_B (med 0 in 40) in p. Na sliki 12 sta prikazana primera, ki ilustrirata, da postaja z manjšanjem razlike med λ_A in λ_B porazdelitev zmesi unimodalna (zlasti, če je p blizu 0 ali 1), z večanjem razlike med λ_A in λ_B pa vse bolj izrazito bimodalna (zlasti pri $p \approx 0{,}5$). Modus porazdelitve zmesi ni analitično izračunljiv, zato ga je potrebno ugotoviti s pregledom $P(k)$ za možni razpon modalnega k, ki je med $round(\min(\lambda_A, \lambda_b))-1$ in $round(\max(\lambda_A, \lambda_b))+1$.

Zmes dveh (ali nekaj) Poissonovih porazdelitev je najpreprostejši primer rezultata *nehomogenega Poissonovega procesa*, tj. Poissonovega procesa, ki se mu parameter λ s časom spreminja. Taki procesi se sicer uporabljajo kot modeli številnih stohastičnih pojavov, npr. klicev v klicni center za nujno pomoč ali priletov letal v zračni prostor določenega letališča (kjer se gostota toka spreminja tekom dneva, pa tudi sezonsko v tednu in skozi leto).

Bivariatna Poissonova porazdelitev

Doslej smo se ves čas držali enorazsežnega sveta in obravnavali le enorazsežne slučajne spremenljivke. Seveda pa obstaja tudi *dvorazsežna* Poissonova porazdelitev, tj. skupna porazdelitev dveh Poissonovih slučajnih spremenljivk. Če imamo tri Poissonove slučajne spremenljivke (X_0, X_1, X_2 s parametri $\lambda_0, \lambda_1, \lambda_2$) in jih seštejemo v dve novi slučajni Poissonovi spremenljivki

$$\begin{aligned} X &= X_1 + X_0 \\ Y &= X_2 + X_0 \end{aligned},$$

je skupna porazdelitev teh dveh spremenljivk bivariatna Poissonova

$$(X,Y) \sim BP(\lambda_1, \lambda_2, \lambda_0)$$

z verjetnostno funkcijo

$$P(X = x, Y = y) = \\ = e^{-(\lambda_1+\lambda_2+\lambda_0)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{i=0}^{\min(x,y)} \binom{x}{i} \binom{y}{i} i! \left(\frac{\lambda_0}{\lambda_1 \lambda_2}\right)^i .$$

Robni porazdelitvi sta $X \sim Pois(\lambda_1 + \lambda_0)$ in $Y \sim Pois(\lambda_2 + \lambda_0)$, kovarianca X in Y pa je λ_0. Bivariatna Poissonova porazdelitev je uporabna kot model za napovedovanje športnih izidov, zlasti nogometnih tekem (kjer števili zadetkov, ki ju dosežeta moštvi na tekmi, med seboj nista neodvisni).[39]

Za napovedovanje modeliranje (in napovedovanje) športnih izidov je pomembna tudi porazdelitev razlike med X in Y. Verjetnostna funkcija spremenljivke $Z = X - Y$, ki predstavlja verjetnost za zmago (oziroma neodločen izid ali poraz) z določeno razliko zadetkov, je neodvisna od λ_0 in lažje izračunljiva (če uporabimo prilagojeno Besselovo funkcije prve vrste reda z, I_z; BESSELI v Excelu). Porazdelitev, ki jo dobimo, je znana kot Skellamova porazdelitev:

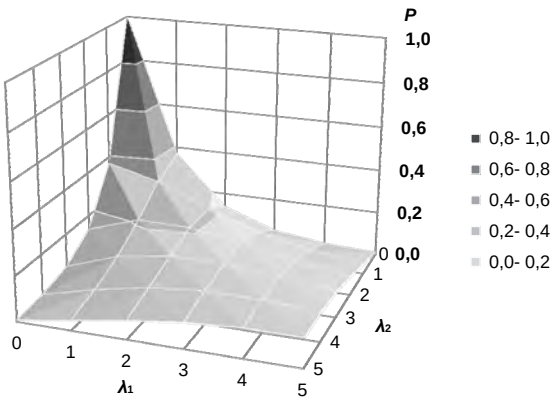
$$P(Z = z) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}\left(2\sqrt{\lambda_1 \lambda_2}\right). \qquad [10]$$

Če postavimo $z = 0$, dobimo z obrazcem [10] verjetnost neodločenega izida glede na pričakovano število zadetkov (kar pomeni "moč") obeh moštev. Če obdržimo konstantno moč nasprotnega moštva (λ_2, tj. pričakovano število zadetkov, ki jih bo prvo moštvo prejelo), pa dobimo z obrazcem [10] verjetnosti za posamezne razlike v zadetkih glede na moč prvega moštva (λ_1, tj. pričakovano število zadetkov, ki jih bo doseglo prvo moštvo). Ta dva trirazsežna grafikona, izdelana s 15. delovnim listom priloženega Excelovega delovnega zvezka, sta prikazana na slikah 13 in 14.

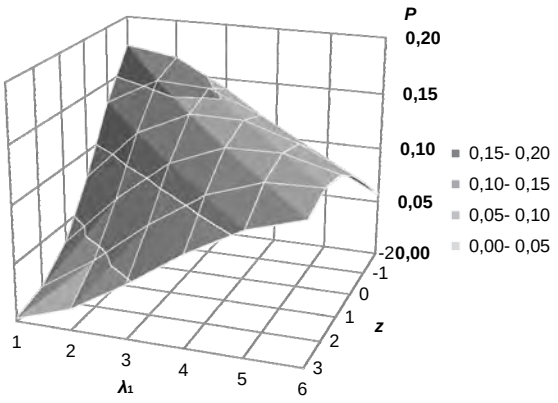
- Na sliki 13 vidimo, da je verjetnost neodločenega izida manjša pri večjem pričakovanem številu zadetkov, in manjša pri

večji razliki v moči med moštvoma. Porazdelitev "deluje" tudi za robni primer, ko sta obe povprečji 0, saj je tedaj neodločen izida neizbežen $(P=1)$.

- Na sliki 14 je primer $(\lambda_2 = 4)$, ki da med možnostmi v delovnem zvezku (za λ_2 lahko izberemo vrednost 1, 2, 3 ali 4; λ_1 je tabelirana od 1 do 6, z pa od -2 do 3) najnižje vrednosti P. Pri manjših vrednostih λ_2 se vrh pomakne proti levemu zgornjemu kotu slike in največji P naraste do 0,3.



Slika 13 Verjetnost neodločenega izida (*P*) med nogometnima moštvoma glede na povprečno število zadetkov, ki jih dosegata na tekmo (λ_1, λ_2; Skellamova porazdelitev).



Slika 14 Verjetnost izida (*P*) z razliko v zadetkih *z* med nogometnim moštvom, ki v povprečju dosega λ_1 zadetkov, in moštvom, ki v povprečju dosega 4 zadetke na tekmo (Skellamova porazdelitev).

Tako kot do običajnih (enorazsežnih) porazdelitev vodijo običajni slučajni processi, vodijo do večrazsežnih porazdelitev slučajna polja (ang. *random fields*). To zelo zahtevno in hkrati široko uporabno področje se je močno razvilo v zadnjih letih (npr. na področju slikanja možganov oziroma nevroznanosti, matematičnih modelov v ekologiji in računalniške grafike). Tu ga omenimo zaradi zanimive uporabe dvorazsežnih Poissonovih slučajnih polj na področju tekstilne tehnologije, kjer so z njimi razrešili zapleteni (tudi dobesedno) problem števila križanj vlaken v nètkanih mrežah vlaken.[58]

Posplošitev bivariatne na skupno porazdelitev več Poissonovih slučajnih spremenljivk je *večrazsežna* Poissonova porazdelitev. Že bivariatna je dovolj zapletena in zahtevna, zato verjetnostne funkcije za multivariatno Poissonovo porazdelitev sploh ne bomo zapisali (in itak ni primerna za izračun, ocenjevanje njenih parametrov pa je še težje). A kljub temu ji je sodobna statistika kos – ne le posamezni, pač pa tudi zmesem večrazsežnih Poissonovih porazdelitev, ki se uporabljajo npr. za združevanje v skupine (*clustering*). Primer s področja trženja je prepoznavanje tipov kupcev glede na to, koliko izdelkov katere vrste kupijo, pri čemer je število kupljenih izdelkov posamezne vrste ob enem obisku trgovine Poissonovo porazdeljeno.[28]

Najpreprostejši statistični test

Vrnimo se k preprostejšim problemom in si oglejmo primerjavo dveh ocenjenih Poissonovih parametrov na najpreprostejši način. Na področju biomedicine je sicer pereč problem nepotrebna uporaba starejših (predračunalniških) biostatističnih metod in zaostanek raziskovalne prakse za statističnimi dognanji se iz dneva v dan povečuje, a v tem primeru je drugače. Izkazalo se bo namreč, da je metoda, ki jo nekateri imenujejo *najpreprostejši statistični test*, presenetljivo praktično uporabna.

Najpomembnejši izid (ang. *primary outcome*) v številnih raziskavah v medicini, zlasti kliničnih poskusih, je izid bolezni. Praviloma gre za to, ali se je izbrani dogodek (npr. ozdravitev) zgodil ali ne. Predstavljajmo si randomiziran (to pomeni, da so udeleženci v eksperimentalne pogoje oziroma skupine razporejeni po naključju) klinični poskus z dvema (vsaj približno) enako velikima skupinama udeležencev, v katerem je izid, ki nas zanima, določen klinični dogodek. Za naš test moramo poznati le število udeležencev, ki jih je doletel dogodek – v eni skupini x_1, v drugi pa x_2. Skupini sta zaradi randomizacije neodvisni. Tedaj lahko testno statistiko za primerjavo med skupinama (natančneje: testiranje ničelne domneve o enaki pogostnosti dogodka v obeh populacijah, iz katerih smo vzorčili skupini) izračunamo po preprostem obrazcu[61]

$$z = \frac{x_1 - x_2}{\sqrt{x_1 + x_2}}. \qquad [11]$$

Izpeljava je prav tako preprosta:

- najboljša cenilka razlike med populacijskima povprečjema je razlika vzorčnih povprečij;
- varianca razlike dveh neodvisnih slučajnih spremenljivk je vsota njunih varianc;
- varianca Poissonove slučajne spremenljivke je enaka njenemu povprečju [2].

Dobljena vrednost z se pod ničelno domnevo, da imata obe zdravljenji enak učinek na tveganje za nastop dogodka, porazdeljuje približno po standardni normalni porazdelitvi, tj. Gaussovi porazdelitvi s povprečjem 0 in varianco 1.

Izračun je še preprostejši – zanj zadošča že računalo, ki je vgrajeno v vsak pametni telefon. Ker gre za približen test, ga lahko dodatno poenostavimo tako, da iskanje po statističnih tabelah oziroma klikanje nadomestimo s pravilom, da razlika po vsej verjetnosti ni slučajna, če smo (pri dvosmernem testiranju) dobili z večji od 2 (oziroma manjši od –2). Še preprosteje je, če za x_1 vedno vzamemo večje izmed obeh števil dogodkov, s čimer se izognemo negativnim vrednostim z. Če sta x_1 in x_2 enaka, pa seveda brez računanja sklenemo, da ničelno domnevo obdržimo ($p = 1$).

Poudariti je potrebno, da informacija znotraj vsake skupine leži le v številu deleža dogodkov, torej zgolj v številu udeležencev z dogodkom. Imenovalec ni pomemben – vseeno je, ali smo isto število dogodkov opazili v poskusu z večjima ali z manjšima skupinama. Od števila dogodkov pa je odvisna moč testa, ki je seveda večja pri večjem številu dogodkov. Poleg tega je test seveda pristranski, če se imenovalca deležev (t.j. velikosti skupin, ki ju primerjamo) nezanemarljivo razlikujeta.[45] Če je pogostnost dogodkov velika, postane test konservativen – daje vrednosti p, ki so večje, kot bi morale biti.

Kljub približnosti in omejitvam daje test v praksi večinoma zanesljive rezultate. Pri randomizaciji je namreč število udeležencev v obeh skupinah praviloma (skoraj) enako, kar velja tudi za čas spremljanja. Pogostnost kliničnega dogodka je praviloma majhna (manjša od 20%, pogosto pa še mnogo manjša), torej lahko predpostavimo, da se število udeležencev v izbrani skupini, ki jih doletel dogodek, porazdeljuje po Poissonovi porazdelitvi. Če skupno število dogodkov $(x_1 + x_2)$ ni premajhno (zadošča že okoli 20), približek z normalno porazdelitvijo dobro deluje.

Pogosto ta preprosti test vodi do praktično popolnoma enakih sklepov kot bolj zapleteni statistični testi oziroma modeli.[49] Tak primer je randomizirana študija učinkovine moksonidin v primerjavi s placebom pri srčnem popuščanju (objavljena leta 2004), pri kateri so preučevali umrljivost. Ob vmesni analizi je med 1860 pacienti prišlo do 46 smrti v skupini z moksonidinom in 25 smrti v skupini s placebom. Vrednost $z = (46 - 25)/\sqrt{(46 + 25)} = 2{,}49 \Rightarrow p = 0{,}013$ nudi močan dokaz za večjo umrljivost v skupini z monksonidinom. Zaradi tega ključnega podatka so klinični poskus predčasno ustavili. Z vključenimi dodatnimi 73 pacienti in skupaj 15 smrtmi so v končni analizi primerjali 54 smrti pri

monksonidinu z 32 pri placebu in s testom log rank (za primerjavo preživetja – za krnjene podatke) dobili praktično enako statistično značilnost $(p = 0{,}012)$. V tovrstnih študijah je za vmesno analizo pogosto nemogoče veljavno uporabiti metode analize preživetja, saj za nekatere udeležence sploh ni na voljo podatka o zadnjem datumu od vključitve, ko so bili še živi, zato je preprosti test še posebej priročen.

Podobno se je izkazalo v metaanalizi šestih študij revaskularizacije (objavljeni leta 2000). Študije so primerjale stente, ki izpirajo sirolimus ali paklitaksel. Grobo združeni podatki o 3669 pacientih iz vseh šestih študij skupaj so pokazali, da je do revaskularizacije prišlo pri 95 pacientih v skupini s sirolimusom in 142 pacientih v skupini s paklitakselom. S preprostim testom dobimo $z = (142 - 95)/\sqrt{(142 + 95)} = 3{,}05 \Rightarrow p = 0{,}002$. Rezultat se praktično povsem ujema z objavljenim stratificiranim testom Mantela in Haenszela, s katerim so dobili $p = 0{,}001$. Čeprav ne upoštevamo imenovalcev deležev v posameznih študijah in na "primitiven" način združimo vse podatke, nam da t.i. najpreprostejši test vseeno ustrezen odgovor.

Omenimo še, da lahko iz istovrstnih podatkov oziroma predpostavk kot za najpreprostejši test pridemo tudi do McNemarjevega testa razlike med odvisnima deležema, ki je številsko ekvivalenten. McNemarjev test izhaja iz dejstva, da se pod ničelno domnevo število dogodkov, ki od danega skupnega števila dogodkov odpade na vsako od skupin, porazdeljuje kot binomska slučajna spremenljivka z verjetnostjo posameznega dogodka 0,5. Testna statistika McNemarjevega testa je kvadrat obrazca [11]. Ekvivalentnost obeh testov uvidimo, če upoštevamo normalno porazdelitev kot asimptotični približek binomske ter vemo, da za kvadrat standardno normalno porazdeljene slučajne spremenljivke velja porazdelitev χ^2 z eno prostostno stopnjo.

Poissonova regresija

Področje statistike, kamor spada Poissonova regresija, so *posplošeni linearni modeli* (ang. *generalised linear models*).[33] Ker je zelo obsežno in matematično zahtevno, se ga bomo tu le dotaknili. Kot vsak regresijski model je tudi Poissonov lahko preprost, torej z eno neodvisno spremenljivko (prediktorjem, napovednim dejavnikom), ali multipli (z več neodvisnimi spremenljivkami); tu bomo zaradi splošnosti obravnavali slednjega. Seveda so lahko Poissonovi (in sorodni) regresijski modeli tudi multivariatni, kar pomeni, da napovedujemo multivariatno Poissonovo (ali sorodno) porazdeljeno skupino spremenljivk, a tako zapletenih stvari se ne bomo niti dotaknili. Uvodoma povejmo še, da se – kot pri vseh posplošenih linearnih modelih – pri Poissonovi regresiji parametre (tj. regresijske koeficiente) ocenjuje po metodi največjega verjetja.

Poissonov regresijski model predpostavlja, da je vzorec n vrednosti (opazovanj) y_i vzet iz medsebojno neodvisnih Poissonovih slučajnih spremenljivk Y_i s povprečji μ_i. Očitno je, da za tak model predpostavka običajne linearne regresije o enakosti varianc med opazovanji (t.i. homoscedastičnosti) ne drži, saj je za vsako Y_i varianca enaka njenemu povprečju. Zato je za napoved y_i na podlagi vektorja vrednosti neodvisnih spremenljivk $\mathbf{x}_i$ (ki vsebuje dodatno vrednost 1 zaradi regresijske konstante, ki je dodatna vrednost v vektorju regresijskih koeficientov $\boldsymbol{\beta}$) ustreznejši model

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} . \qquad [12]$$

Pri Poissonovi regresiji gre torej za posplošeni linearni model z logaritemsko vezjo (ang. *link function*) in Poissonovo porazdeljeno napako. Če obrazec [12] preuredimo z eksponentno funkcijo na obeh straneh enačaja, uvidimo, da gre za multiplikativen model. Povečanje neodvisne spremenljivke x_j za eno enoto prinese množenje y_j z e^{β_j}:

$$\mu_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}} .$$

Podatki, za katere je primerna Poissonova regresija, so pogosto na voljo le v združeni (agregirani) obliki. A to ne predstavlja težave, saj lahko zaradi lastnosti [3] podatke analiziramo bodisi v posamični (individualni, kakršne smo navajeni v uporabni statistiki), bodisi v združeni obliki. To si najlažje pojasnimo s primerom. Denimo, da imamo podatke o skupnem številu otrok po skupinah mater, ki so definirane glede na trajanje zakonske zveze (0-4 leta, 5-9 let itd.), okolje bivanja (urbano ali ruralno) in izobrazbo matere (stopnje od I do IX). Z Y_{ijkl} označimo število otrok, ki jih je rodila l-ta mati v skupini $(i; j; k)$, pri čemer i označuje trajanje zakonske zveze, j okolje bivanja in k stopnjo izobrazbe. Oznaka $Y_{ijk\bullet} = \sum_j Y_{ijkl}$ torej označuje skupno število otrok v posamezni celici tabele s podatki. Če je vsako posamezno opazovanje (tj. število otrok dane matere) realizacija Poissonove slučajne spremenljivke s povprečjem μ_{ijk}, je skupinska vsota realizacija Poissonove slučajne spremenljivke s povprečjem $n_{ijk}\mu_{ijk}$, pri čemer je n_{ijk} število mater v ustrezni skupini, tj. celici $(i; j; k)$ v podatkovni tabeli. V Sloveniji smo Poissonovo regresijo za agregirane podatkov uporabili npr. pri analizi povezanosti prezgodnje umrljivosti (smrti so redke, t.i. prezgodnje – tj. tiste pred starostjo za upokojitev – še bolj, živih ljudi pa je tudi v majhni Sloveniji dovolj, da s statističnega vidika predstavljajo zelo veliko množico) s socioekonomskimi dejavniki (regija, zakonski stan, materni jezik, stopnja izobrazbe).[25]

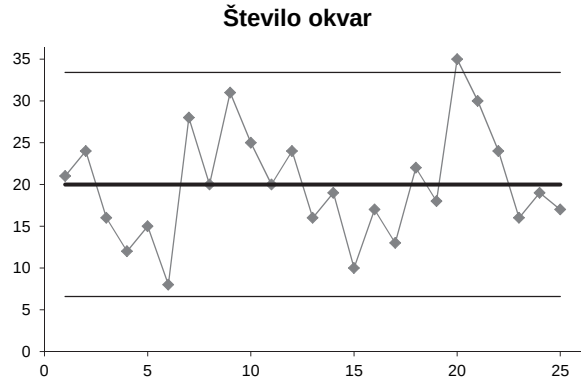
Nadaljevali ne bomo, saj si Poissonova regresija zasluži svoje gradivo. Omenimo le še njene razširitve oziroma alternative za primer pod- oziroma nadrazpršenosti odvisne spremenljivke. Prva možnost je regresijski model za napoved spremenljivke, ki ustreza Poissonovi porazdelitvi, le da ne more zavzeti vrednosti nič (ki je "odrezana" – ang. *zero-truncated Poisson model*). Druga možnost je *negativna binomska regresija*, o kateri je lani izšla obsežna monografija.[36] Tretja možnost je *COM-Poissonova regresija*, ki je v zadnjih letih zelo "vroča" statistična tema zlasti tam, kjer gre za izredno množične pojave, s katerimi je povezano izredno veliko denarja, npr. spletno oglaševanje (modeliranje števila obiskov spletnih strani), nakupovalne centre (modeliranje števila kupcev) ali spletne dražbe (modeliranje števila dražiteljev).[56] Za konec dodajmo, da spada Poissonova porazdelitev v široko družino *Tweediejevih porazdelitev*, ki kot posebne primere združuje različne zvezne (npr. normalno in gama) in diskretne porazdelitve (npr. Poissonovo) in se uporablja kot najsplošnejša porazdelitev napovedovanega odgovora v posplošenih linearnih modelih.

Kontrolne karte

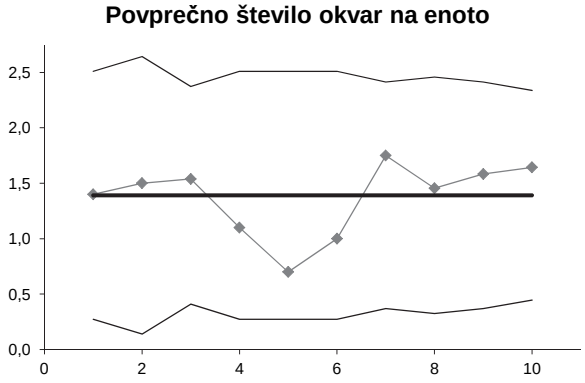
Kontrolne karte (in celotno področje *statističnega nadzora procesov* oziroma *kakovosti* – ang. s kratico *SPC* oziroma *SQC*) po eni strani ne sodijo v "glavni tok" verjetnosti in statistike (in potemtakem tudi ne v to gradivo), po drugi strani pa bi jih mirne duše lahko umestili že v razdelek o grafičnem preverjanju ujemanja podatkov s Poissonovo porazdelitvijo. Na koncu gradiva jih obravnavamo kot kompromis med tema pogledoma, pa tudi zato, da z vidika matematične zahtevnosti zaključimo nekoliko bolj lahkotno.

S Poissonovo porazdelitvijo sta povezni dve od osnovnih (Shewhartovih) kontrolnih kart za atribute (ang. *attributes control charts*): c-karta in u-karta. C-karta je namenjena nadzoru števila (ang. *count*) okvar (defektov, odstopanj od specifikacij). Za razliko od p-karte (za delež okvarjenih enot) dopušča več okvar na enoto, po drugi strani pa zahteva stalno velikost vzorca (za razliko od p-karte ali u-karte), da je središčna črta lahko konstantna. Primerna je za podatke, ki nastajajo s Poissonovim procesom (npr. število izdelkov, vrnjenih v trgovino zaradi reklamacije, na dan). Za vsak vzorec zabeležimo stopnjo (ang. *rate*) okvar c_j (ki je število okvar v izbrani enoti pregleda; enota pregleda je navadno en izdelek, lahko pa tudi več izdelkov). Središčna črta grafikona je pri vrednosti $\bar{c}$, ki je bodisi v naprej znana bodisi jo izračunamo kot povprečje vrednosti c_j. Ker predpostavimo Poissonovo porazdelitev podatkov, je standardni odklon

števila okvar $\sqrt{\bar{c}}$, torej narišemo 3σ meje nadzora pri $\bar{c} \pm 3\sqrt{\bar{c}}$. Če pade spodnja meja nadzora pod 0, jo postavimo na 0. Primer *c*-karte je na sliki 15 in na 16. delovnem listu priloženega Excelovega delovnega zvezka. Preden si ogledamo *u*-karto, ponovimo predpostavke za uporabo *c*-karte: priložnosti (lokacij) za potencialne okvare je veliko, verjetnost okvare na posamezni lokaciji je majhna in postopek pregleda je enak za vse vzorce.



Slika 15 Kontrolna *c*-karta za število okvar.



Slika 16 Kontrolna *u*-karta za povprečno število okvar na enoto.

U-karta je namenjena nadzoru stopnje okvar (ang. *rate*) na enoto. Podatki morajo ustrezati Poissonovi porazdelitvi, kot npr. število padcev bolnikov v bolnišnici na dan. Če središčna črta ni v naprej določena in se velikost vzorcev (n_j) spreminja, določimo središčno črto $\bar{u}$ kot dolgoročno povprečje, tj. skupno število okvar deljeno s skupno velikostjo vzorca. Meje nadzora (3σ) so tedaj pri $\bar{u} \pm 3\sqrt{\frac{\bar{u}}{n_j}}$ (v skladu s standardno napako ocene Poissonovega povprečja, ki smo jo spoznali v razdelku o ocenjevanju parametra). Seveda meje nadzora niso vodoravne, pač pa se ožijo in širijo glede na n_j. Primer *u*-karte je na sliki 16 in na zadnjem (17.) delovnem listu priloženega Excelovega delovnega zvezka.

Izračun meja nadzora in izdelava *c*- in *u*- kontrolne karte sta v elektronski preglednici zelo preprosta, zato ju puščamo za vajo bralcu oziroma bralki. Priporočljiva je uporaba podatkovnih tabel, da se lahko grafikoni samodejno prilagodijo, če dodamo podatke pod obstoječe. Kontrolni karti v priloženem delovnem zvezku pa sta izdelani z brezplačnim in javno dostopnim dodatkom (*add-in*) SPC Charts za risanje kontrolnih kart v Excelu, ki ga na spletu najdemo skupaj s spremljajočim člankom.[31] Dodatek deluje tako v starejših verzijah Excela (2003 in starejših) kot v sodobni (2007 in 2010 v okolju Windows, 2011 v okolju MacOS).

Prednost uporabe takega programja je, da lahko poleg meja nadzora samodejno upošteva tudi različna kontrolna pravila, kot npr. pravila Western Electric. Namen tovrstnih pravil je pravočasno odkrivanje procesov, ki niso pod nadzorom, na podlagi nakazujočih se trendov (2 od 3 zaporednih točk izven meja 2σ na isti strani središčne črte, 4 od 5 zaporednih točk izven meja 1σ na isti strani središčne črte, 8 zaporednih točk na isti strani središče črte ipd.). Če v dodatku SPC Charts izberemo to opcijo, so odstopajoče točke označene s številko pravila v stolpcu "Flag" (kaj pomeni katera številka, izvemo v dialogu za postavljanje pravil). Na predstavljeni *c*-karti (slika 15, 16. delovni list) sta dve odstopajoči točki (ena nad zgornjo mejo nadzora, tista za njo pa je odstopajoča po pravilu "2 od 3 zaporednih točk izven meja 2σ"), predstavljena *u*-karta (slika 16, 17. delovni list) pa prikazuje proces pod nadzorom (brez odstopajočih točk).

Obe kontrolni karti za Poissonove podatke smo spoznali v osnovni obliki, obstajajo pa tudi njuni

popravki in prilagoditve, ki naj bi izboljšali njune *operacijske značilnosti* (ang. *operating characteristics*), tj. ustreznost meja nadzora v smislu čim manj napačnih odločitev (ali v jeziku diagnostičnih testov: občutljivosti in specifičnosti). Poleg tega za podatke, ki ustrezajo Poissonovi porazdelitvi, obstajajo tudi naprednejše kontrolne karte na podlagi eksponentno uteženega drsečega povprečja (ang. exponentially weighted moving average – *EMWA*) in kumulativne vsote (*CUSUM*).[10,54] Najbolj svež dosežek SPC/SQC v zvezi s Poissonovo porazdelitvijo pa so robustne kontrolne karte za število dogodkov na podlagi COM-Poissonove porazdelitve, ki upoštevajo morebitno nad- ali podrazpršenost.[55]

Domača naloga

Detektivsko delo je zelo dobra analogija za uporabno statistiko oziroma statistično svetovanje. Zato naj bo domača naloga detektivska – seveda na neobvezen način, kot pri branju ali gledanju detektivskih zgodb, saj je rešitev na koncu gradiva.

Detektiv dobi nalogo prijeti serijskega avtomobilskega tatu, za katerega je znano, da deluje na določeni ulici v mestu. Tat se seveda želi izogniti prijetju, zato ulico obiskuje le ob slučajno izbranih dnevih v letu in ob slučajno izbranem času dneva, a v povprečju izpelje 10 uspešnih tatvin letno.

Detektiv se mora seveda hkrati ukvarjati z drugimi primeri, zato lahko posveti opazovanju te ulice le 3 ure na dan 3 dni v tednu. Kolikšna je verjetnost, da bo detektiv zasačil tatu v enem tednu, in kolikšna, da ga bo zasačil v enem letu?

Namesto zaključka

Gradivo je v avtorjevih mislih začelo nastajati pred dobrim desetletjem – *TEMPVS·FVGIT*. Bilo naj bi eno od ducata poglavij v knjigi z naslovom *Nematematikovi sprehodi po matematiki in statistiki* in vsako poglavje naj bi spremljal celovit in skrbno dodelan elektronski delovni zvezek v Excelu …

Vsaj eno od poglavij je torej skupaj z Excelovim delovnim zvezkom ugledalo luč sveta, ali bodo sprehodi kdaj nastali v celoti, pa ostaja odprto vprašanje, odvisno od mnogih dejavnikov. Vsekakor se še tako dolga pot začne s prvim korakom, ki je zdaj (do)končno narejen.

V uvodu bralec oziroma bralka ne naleti na opozorilo, kakšno predznanje zahteva gradivo. Lahko bi pisalo, da osnove statistike (tj. vsaj enosemestrski predmet na dodiplomskem oziroma prvostopenjskem univerzitetnem študiju, raje pa več in na višji ravni) in gimnazijsko matematiko (kot je omenjeno v razdelku o rodovni funkciji, ki jo edini presega). Toda predznanje ni bistveno – bistveno je veselje, zanimanje, radovednost, strast za statistiko, matematiko, znanost in nasploh za resnico!

Podobno velja za razpored in tok vsebine. Uvodno opozorilo, da bodo nekateri deli bolj zložni in nekateri zahtevnejši ter ponekod dovolj zmerna pozornost in ponekod potrebna popolna zbranost, bi bilo odveč. Seveda je tako – kot na vsakem sprehodu. Tisti, ki nas povabi na sprehod, nam tudi določi, kam bomo šli in usmerja pozornost na tisto, kar naj bi si ogledali – in tako je avtor gradiva nekoliko več pozornosti namenil tistemu, kar bolje pozna, se mu zdi bolj zanimivo oziroma meni, da je premalo znano. Tu se skriva tudi del razlage za izbor programja (tj. zakaj Excel in ne R ali IBM SPSS, Minitab ali kaj petega).

Seveda se je avtor po najboljših močeh trudil biti sistematičen in objektiven, a zanj je – in za bralca oziroma bralko naj bo – kot na sprehodu – bistvena želja po svežem in drugačnem ter iskanje lepote in pomena vsak trenutek in povsod. Kajti pot je cilj – do Poissonove porazdelitve in naprej.

Viri – učbeniki

1. Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research* (4th ed.). Oxford 2002: Blackwell.
2. Bulmer MG: *Principles of statistics*. New York 1985: Dover.

3. Cedilnik A: *Uvod v verjetnostni račun*. Ljubljana 2002: Fakulteta za družbene vede.
4. Davis G, Pecar B: *Business statistics using Excel*. Oxford 2010: Oxford University Press.
5. Eason G, Coles CW, Gettinby G: *Mathematics and statistics for the bio-sciences*. Chichester 1980: Ellis Harwood, Holstead Press.
6. Grinstead CM, Snell JL: *Introduction to probability* (2nd ed.). Providence 1997: American Mathematical Society.
7. Ivanović B: *Teorijska statistika* (2. izd.). Beograd 1979: Naučna knjiga.
8. Pavlić I: *Statistička teorija i primjena* (4. izd.). Zagreb 1988: Tehnička knjiga.
9. Rice JA: *Mathematical statistics and data analysis* (3rd ed.). Belmont 2007: Duxbury, Thomson Brooks/Cole.
10. Ryan TP: *Statistical methods for quality improvement* (3rd ed.). Hoboken 2001: John Wiley.
11. Suhov Y, Kelbert M: *Probability and statistics by example. Volume I. Basic probability and statistics*. Cambridge 2005: Cambridge University Press.
12. Vranić V: *Vjerojatnost i statistika* (3. izd.). Zagreb 1971: Tehnička knjiga.

Viri – Wikipedia

13. Compound Poisson distribution. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Compound_Poisson_distribution
14. Conway-Maxwell-Poisson distribution. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Conway%E2%80%93Maxwell%E2%80%93Poisson_distribution
15. Erlang distribution. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Erlang_distribution
16. Ernest Rutheford. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Ernest_Rutherford
17. Negative binomial distribution. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Negative_binomial_distribution
18. Poisson distribution. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Poisson_distribution
19. Poisson process. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Poisson_process
20. Poisson regression. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Poisson_regression
21. Siméon Denis Poisson. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Sim%C3%A9on_Denis_Poisson
22. Skellam distribution. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Skellam_distribution
23. Tweedie distributions. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/Tweedie_distributions
24. V-1 flying bomb. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/V-1_flying_bomb
25. William Sealy Gosset. In: *Wikipedia, The Free Encyclopedia*. San Francisco 2012: Wikimedia Foundation. http://en.wikipedia.org/wiki/William_Sealy_Gosset

Viri – dodatni

26. Artnik B, Vidmar G, Javornik JS, Laaser U. Premature mortality in Slovenia in relation to selected biological, socioeconomic, and geographical determinants. *Croat Med J* 2006; 47(1): 103-113.
27. von Bortkiewicz L: *Das Gesetz der Kleinen Zahlen*. Leipzig 1898: B.G. Teubner.
28. Brijs T, Karlis D, Swinnen G, *et al.*: A multivariate Poisson mixture model for marketing applications. *Stat Neerl* 2004; 58(3): 322-348.
29. Brown LD, and Zhao LH: A test for the Poisson distribution. *Sankhya Ser A* 2002; 64(3): 611-625.
30. Bru B: Poisson, the probability calculus, and public education. *J Electron Hist Probab Stat* 2005; 1(2): 1-25.
31. Buttrey SE: An Excel add-in for statistical process control charts. J Stat Soft 2009; 30(13). http://www.jstatsoft.org/v30/i13

32. Clarke RD: An application of the Poisson Distribution. *J Inst Actuar* 1946; 72: 481.
33. Dobson AJ, Barnett A: *An introduction to generalized linear models*. Boca Raton 2008: Chapman and Hall.
34. Garwood F: Fiducial limits for the Poisson distribution. *Biometrika* 1936; 28(3/4): 437-442.
35. Hand DJ, Daly F, Lunn D, McConway K, Ostrowski E: *A handbook of small data sets*. London 1994: Chapman & Hall.
36. Hilbe JM: *Negative binomial regression* (2nd ed.). Cambridge 2011: Cambridge University Press.
37. Hoaglin DC: A poissonness plot. *Am Stat* 1980; 34(3): 146-149.
38. Hoaglin DC, Tukey JW: Checking the shape of discrete distributions. In: Hoaglin DC, Mosteller F, Tukey JW (eds.), *Exploring data tables, trends and shapes*. New York: John Wiley 1985; 345-416.
39. Karlis D, Ntzoufras I: Analysis of sports data by using bivariate Poisson models. *Statistician* 2003; 52(3): 381-393.
40. Karlis D, Xekalaki E: A simulation comparison of several procedures for testing the Poisson assumption. *Statistician* 2000; 49(3): 355-382.
41. Komelj J: *Aktuarsko računanje agregatnih odškodnin in optimalnih parametrov pozavarovanja* (magistrsko delo). Ljubljana 2004: Univerza v Ljubljani, Ekonomska fakulteta.
42. Leemis LM, McQueston JT: Univariate distribution relationships. *Am Stat* 2008; 62(1): 45-53.
43. McLaughlin MP: *A Compendium of Common Probability Distributions* (Regress+ Tutorial Appendix A). McLean 1999: causaScientia. http://www.causascientia.org/math_stat/Dists/Compendium.html
44. de Moivre A: De mensura sortis seu; de probabilitate eventuum in ludis a casu fortuito pendentibus. *Phil Trans R Soc* 1711; 27(329): 213-264.
45. Ng HKT, Gu K, Tang ML: A comparative study of tests for the difference of two Poisson means. *Comp Stat Data Anal* 2007; 51(6): 3085-3099.
46. O'Gorman WD, Kunkle EC: Study of the relation between Minnesota multiphasic personality inventory scores and pilot error in aircraft accidents. *J Aviat Med* 1947; 18(1): 31-38.
47. Ord JK: Graphical methods for a class of discrete distributions, *J R Stat Soc A* 1967; 130(2): 232-238.
48. Patil VV, Kulkarni HV: Comparison of confidence intervals for the Poisson mean: some new aspects. *REVSTAT-Stat J* 2012; 10(2): 211-227.
49. Pocock SJ: The simplest statistical test: how to check for a difference between treatments. *BMJ* 2006; 332(7552): 1256-1258.
50. Poisson SD: Recherches sur la probabilité des jugements en matière criminelle et en matière civile. V: Procédés des Règles Générales du Calcul des Probabilitiés. Paris 1837: Bachelier.
51. Reiczigel J, Lang Z, Rózsa L, Tóthmérész B: Properties of crowding indices and statistical tools to analyze parasite crowding data. *J Parasitol* 2005; 91(2): 245-252.
52. Romanovskiy VI: *Primenenije matematičeskoj statistiki v opitnom dele*. Moskva, Leningrad 1947: Gostekhizdat.
53. Rutherford E, Geiger H (note by Bateman H): The probability variations in the distribution of α particles. *Phil Mag* 1910; 20(6): 698-704.
54. Ryan AG, Woodall WH: Control charts for poisson count data varying sample sizes. *J Qual Technol* 2010; 42(3): 260-275.
55. Sellers KF: A generalized statistical control chart for over- or under-dispersed data. *Qual Reliab Engng Int* 2012; 28(1): 59-65.
56. Sellers KF, Borle S, Shmueli G: The COM-Poisson model for count data: a survey of methods and applications. *Appl Stoch Model Bus Ind* 2012; 28(2): 104-116 (Rejoinder: 128-129).
57. Student: On the error of counting with a haemocytometer. *Biometrika* 1906; 5(3): 351-360.
58. Suh MW, Chun H, Berger RL, Bloomfield P: Distribution of fiber intersections in two-dimensional random fiber webs – a basic geometrical probability model. *Text Res J* 2010; 80(4): 301-311.
59. Tukey JW: Some graphic and semigraphic displays. In: Bancroft TA, Brown SA (eds.), *Statistical papers in honor of George W. Snedecor*. Ames 1972: Iowa State University Press; 293-316.
60. Tukey JW: *Exploratory data analysis*. Reading 1977: Addison-Wesley.
61. Vidmar G: Primer uporabe najpreprostejšega statističnega testa: ali se zahtevnost rehabilitacije bolnišničnih pacientov povečuje? *Rehabilitacija* 2008; 7(2): 8-11.

Rešitev domače naloge

Če predpostavimo, da se tat na ulici pojavlja ob slučajnih časih, porazdeljenih po Poissonovi porazdelitvi s povprečjem $\lambda = 10$ *na leto, potem dobimo verjetnost, da bo prišlo do vsaj ene tatvine v časovnem obdobju dolžine* L*, tako, da od ena odštejemo Poissonovo verjetnost, da ne bo prišlo do nobene:* $P = 1 - e^{-\lambda L}$.

Ker tri ure dnevno trikrat na teden ustrezajo približno 1/1000 leta opazovanja na teden $\left(3/24 \cdot 3/7 \cdot 1/52 = 9/8736\right)$, *je verjetnost, da bo detektiv zasačil tatu v* T *tednih opazovanja,* $P_T = 1 - e^{-T/100}$. *Za en teden opazovanja je torej verjetnost pičlih 1%* $\left(P_1 = 1 - e^{-1/100}\right)$, *v enem letu pa je verjetnost detektivovega uspeha že blizu polovični, saj znaša 41,5%* $\left(P_{52} = 1 - e^{-52/100}\right)$.