Univerzitetni rehabilitacijski inštitut
Republike Slovenije - Soča

UL100 Univerza v Ljubljani
Medicinska fakulteta

Kaj imajo skupnega konjske brce, rojstva **četverčkov**, pivske kvasovke in klicni centri? Poissonovo porazdelitev.

izr.prof.dr. *Gaj Vidmar*

23. 10. 2019

Primera iz vsakdanjega življenja

- Sedimo ob cesti (prometni v mestu ali samotni v gozdu) in štejemo, koliko vozil pripelje mimo na izbrano časovno enoto (npr. minuto v mestu ali uro v gozdu)?
- Na travnik položimo kvadratno mrežo in štejemo, koliko je v vsakem kvadratu neke cvetlice (pogoste, kot je marjetica, ali redke, kot je štiriperesna detelja)?

Izpeljava iz binomske porazdelitve

- Poissonova porazdelitev je limitna oblika binomske, pri kateri je število poskusov (n) zelo veliko, verjetnost uspeha v posameznem poskusu (p) pa zelo majhna, zato je znana tudi kot **porazdelitev redkih dogodkov**
  - obrazec za binomsko porazdelitev je $X \sim Bin(n,p) \Leftrightarrow P(X=k)=\frac{k!}{n!(n-k)!}p^k(1-p)^{n-k}$
  - vpeljemo $\lambda = np$ oz. $p = \lambda / n$
  - upoštevamo, da če gre $n \to \infty$, velja $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$ in $\left(1-\frac{\lambda}{n}\right)^{-x} \to 1$
- Dobimo verjetnostno funkcijo Poissonovo porazdeljene slučajne spremenljivke X, ki velja za $k = 0,1,2,\ldots$ pod pogojem $\lambda > 0$:

$$X \sim Pois(\lambda) \Leftrightarrow P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$$

- Res gre za verjetnostno porazdelitev, ker ni težko dokazati, da je vsota posameznih verjetnosti enaka 1

Oblika

- Excel 1
  - desno asimetrična (vse manj z večanjem λ)
  - z večanjem λ (kmalu) postane podobna normalni
  - modus je (približno) enak λ
- Izhajamo iz
  - $k = 0 \Rightarrow P = e^{-\lambda}$
  - rekurzivnega obrazca $P(k+1) = \frac{\lambda}{k+1} P(k)$
- Dokler je faktor $\frac{\lambda}{k+1}$ večji od 1, z naraščanjem k verjetnosti naraščajo; ko pade pod 1, pa začno padati (vse hitreje)
- Porazdelitev je torej unimodalna
  - če λ ni naravno število, je modus en, in sicer največje celo število, ki je ≤ λ
  - če je λ naravno število, sta modusa λ-1 in λ

Momenti

- Najpreprostejša je izpeljava z limito binomske porazdelitve
  - povprečje binomske porazdelitve je $np \Rightarrow$ **povprečje Poissonove porazdelitve je λ**
  - varianca binomske porazdelitve je $np(1\text{-}p)$ in $(1\text{-}p)$ gre proti 1, če gre p proti 0 $\Rightarrow$ **varianca Poissonove porazdelitve je tudi λ** (standardni odklon je torej $\sqrt{\lambda}$)
- Asimetričnost je $m_3 = 1/\sqrt{\lambda}$
  - torej vedno pozitivna oziroma desna, a se z večanjem λ približuje 0
- Sploščenost je $m_4 = 3 + (1/\lambda)$
  - torej večja kot 3, kolikor znaša pri normalni porazdelitvi, a z večanjem λ razlika od normalne porazdelitve izginja)

Konvolucija

- **Vsota dveh** (in torej tudi več) Poissonovih slučajnih spremenljivk **je zopet Poissonova slučajna spremenljivka**:

  $$X \sim Pois(\lambda) \wedge Y \sim Pois(\nu) \Rightarrow X + Y \sim Pois(\lambda + \nu)$$

  - Je to v nasprotju s centralnim limitnim izrekom?
  - Ne, ker več Poissonovih spremenljivk ko seštejemo, večja bo vsota njihovih parametrov, ki je hkrati povprečje porazdelitve vsote, in večje kot je povprečje Poissonove porazdelitve, bolj je ta podobna normalni.

Prostorski vidik

- Cvetlice na travniku: **Excel 2**
  - pri naključni porazdelitvi je vsaka dvojica koordinat izžrebana iz enakomerne porazdelitve (po matematično: $x \sim U(0;1), y \sim U(0;1)$; po Excelovo: s funkcijo RAND)
  - pri enakomerni porazdelitvi so cvetlice razporejene na mrežo kvadrantov in nato malo premaknjene po naključju (*jittering*) $\Rightarrow$ **podrazpršenost** (*underdispersion*)
  - pri gručasti porazdelitvi izhajamo iz petih središč na glavni diagonali in nato uporabimo naključne majhne premike $\Rightarrow$ **nadrazpršenost** (*overdispersion*)

Časovni vidik

- Verjetnost števila dogodkov v danem časovnem intervalu, če se dogodki pojavljajo z znano povprečno pogostnostjo in neodvisno drug od drugega
- Storitve, proizvodnja, poslovanje:
  - 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
- Medicina:
  - število pacientov, ki pridejo v nočnem času v dežurno ambulanto oz. kliniko
  - število mutacij določenega sklopa DNK na časovno enoto
- IKT:
  - 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

Časovni vidik – računska primera

- Če časovno obdobje obsega t časovnih enot, v obrazec Poissonove porazdelitve namesto λ vstavimo λt: $P(k)=\dfrac{e^{-\lambda t}(\lambda t)^k}{k!}$
- Igra mačke z mišjo
  - Mačka ujame v povprečju eno miš na dan.
  - Na opazovanem območju živi 10 miši.
  - Kolikšna je verjetnost, da bo mačka polovila vse miši v enem tednu?

$$k=10, \lambda=1, t=7$$

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

- Komercialna telefonska linija (090) – **Excel 3**
  - Sprejme lahko največ dva klica v petih minutah.
  - V povprečju prejme 0,5 klica na minuto.
  - Kolikšen je pričakovani delež klicev, na katere osebje ne bo moglo odgovoriti?

$$P(\leq 2)=P(0)+P(1)+P(2)$$
$$P(>2)=1-P(\leq 2)$$

Ocenjevanje parametra

- Intuitivno: najboljša **cenilka za parameter λ je opaženo povprečje**
  - ustreza kriteriju največjega verjetja (*maximum likelihood*)
  - standardna napaka ocene povprečja, ki je v splošnem $\sigma/\sqrt{n}$, je torej $\sqrt{\lambda/n}$
  - **interval zaupanja** torej dobimo z množenjem standardne napake z vrednostjo iz standardne normalne porazdelitve, ki ustreza želeni stopnji zaupanja (1,96 za 95% itd.)
  - to je Waldova metoda, uporabna za velike vrednosti λ (vsaj 30)
- Eksaktni interval zaupanja temelji na enakosti Poissonove kumulativne porazdelitvene funkcije in komplementarne kumulativne porazdelitve χ^2
  - Enota, na katero štejemo število dogodkov oz. osebkov/predmetov, je praviloma arbitrarna, zato jo lahko spremenimo v celotno opazovano obdobje oz. področje
  - Enako je, če ocenimo λ na podlagi enega samega poskusa oz. štetja
  - **Meji** esaktnega **intervala zaupanja** za opaženo število dogodkov oz. osebkov/predmetov $\lambda = k$ sta odvisni le od stopnje zaupanja – **Excel 4**

$$SM = \frac{\chi^2\left(P = \frac{\alpha}{2}; df = 2k\right)}{2}$$

$$ZM = \frac{\chi^2\left(P = 1 - \frac{\alpha}{2}; df = 2(k+1)\right)}{2}$$

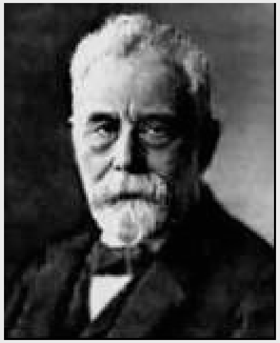
  - Sodobna statistika pozna vsaj še 17 drugih pristopov

Prileganje – **nesreče**

- Običajen pristop je test hi-kvadrat
  - testna statistika $X^2 = \sum_k \frac{(f_o - f_p)^2}{f_p}$ se pod ničelno domnevo porazdeljuje (približno) kot χ^2 z dvema prostostnima stopnjama manj, kot je opaženih različnih vrednosti
  - izidi morajo pokriti celotno zalogo vrednosti opazovane spremenljivke in celic z ničelno oz. majhno f_p naj ne bi bilo, zato se najvišje vrednosti pogosto združi ($k \geq$)
  - Primer: nesreče vojaških pilotov – **Excel 5**
- Indeks razpršenosti
  - računamo ga iz surovih podatkov
  - asimptotično se porazdeljuje kot χ^2 s k–1 prostostnimi stopnjami
  - Poissonov $PID = \frac{\sum_{i=1}^{n} (k_i - \bar{k})^2}{\bar{k}}$

  $\Rightarrow$ test Poissonove razpršenosti
  - Primer: prometne nesreče v mestu – **Excel 6**
- Drugi pristopi
  - testna statistika Kolmogorova in Smirnova ali Cramérja in von Misesa
  - testi na podlagi testa razmerja verjetij (*likelihood ratio*) idr.

Prileganje – **grafični pristopi**

- Primer: smrti zaradi konjskih brc v pruski vojski
  - Ladislaus von Bortkiewicz (učenec Wilhelma Lexisa)
  - podatki iz 14 rodov vojske v obdobju 20 let
  - **Excel 7** in **Excel 8**
- Obešeni korenogram (*hanging rootogram*)
  - J. W. Tukey (skoval/izumil *bit*, *software*, *FFT*, *boxplot*, ...)
  - stolpce merimo navzdol od pričakovane frekvence
  - korenska navpična os poudarja odstopanja v repih
- Grafikon poissonskosti (*Poissonness plot*)
  - D. Hoaglin (Tukeyev učenec in sodelavec)
  - če so opažene frekvence enake pričakovanim, je $\varphi(k) = \ln \frac{k!\, f_{o_k}}{n} = -\lambda + k \ln \lambda$
  - graf φ v odvisnosti od k je torej za Poissonovo porazdelitev premica
  - Tukey in Hoaglin dodelala (izpustimo $k=0$, popravek opaženih frekvenc, dodani intervali zaupanja, manjša teža točkam z opaženo frekvenco 1)

Rojstva četverčkov

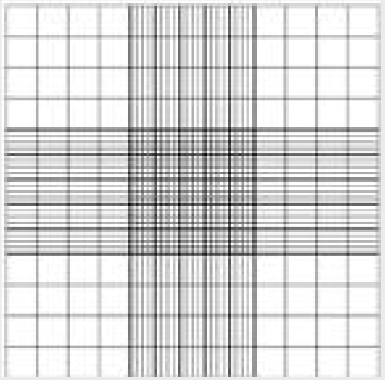
- Von Bortkiewicz je analiziral tudi
  - samomore otrok (na leto v obdobju 1869-1893)
  - samomore žensk (na leto za 8 nemških zveznih držav v obdobju 1881-1894)
  - smrti zaradi delovnih nezgod (na leto za 11 združenj za poklicno zavarovanje v obdobju 1886-1894)
- Iz istega področja in obdobja – **Excel 9**

Pionir

- Siméon Denis Poisson (1781-1840)
  - "Življenje dobro le za dve stvari – odkrivati matematiko in poučevati matematiko"
  - mentorji oz. sodobniki Lagrange, Laplace, Legendre in Fourier
  - porazdelitev je opisal v svojem delu iz leta 1837 o sodnih odločitvah v kazenskopravnih in civilnopravnih zadevah

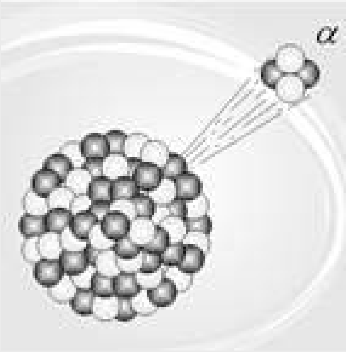
Kvasovke

- William Sealy Gosset – Student (1876-1937)
  - glavni tehnolog v pivovarni Guiness, tast "očeta statistike" R. A. Fisherja
  - preučeval napake pri štetju kvasovk s hemocitometrom
  - mreža 20×20=400 kvadratkov
  - Poissonove porazdelitve ni poznal, sam izpeljal
  - **Excel 10**



Radioaktivni razpad

- Ernest Rutherford (1871-1937)
  - rojen na Novi Zelandiji
  - pionir jedrske fizike, Nobelova nagrada za kemijo 1908
  - odkritje delcev, žarkov oziroma radioaktivnega razpada α in β
  - 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
  - Poissonove porazdelitve niso poznali, izpeljali na novo
  - **Excel 11**
  - časi med razpadi sledijo eksponentni porazdelitvi

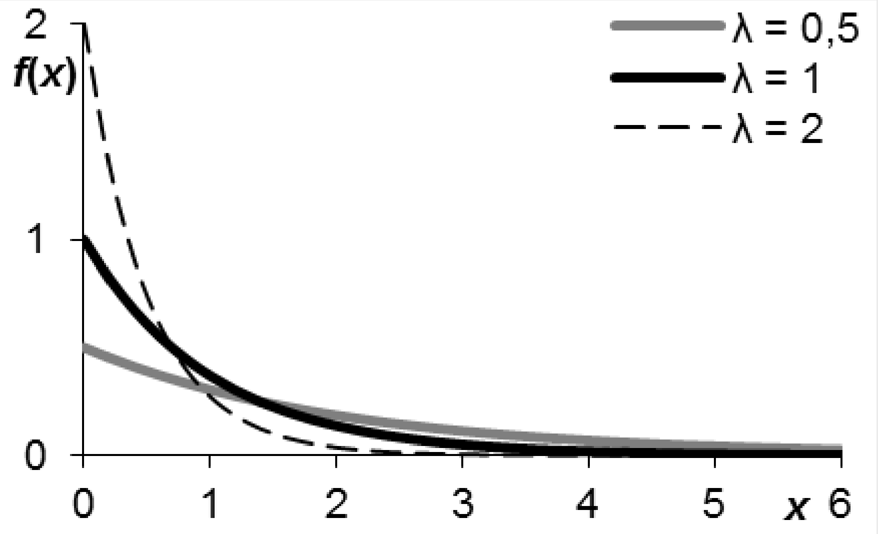


Leteče bombe V-1

- Z njimi je nacistična Nemčija v 2. svetovni vojni napadala London (in še nekaj drugih mest, zlasti Antwerpen) od poletja 1944 do pomladi 1945
- Zaradi hrupa pulznega reaktivnega motorja so jim rekli *buzz bombs*
- Vprašanje obrambe: ali V-1 padajo po naključju ali so vodene k ciljem?
  - če bi veljalo prvo, bi morala biti dvorazsežna prostorska porazdelitev krajev zadetkov naključna, torej porazdelitev na enoto ploščine zelo podobna Poissonovi
  - v drugem primeru bi bila prostorska porazdelitev krajev zadetkov gručasta
- Razdelitev 144 km^2 velikega področja južnega Londona na 576 kvadrantov velikosti ¼ km^2 je potrdila naključnost
- **Excel 12**


Eksponentna porazdelitev

- Gostota verjetnosti $X \sim Exp(\lambda) \Leftrightarrow f(x) = \lambda e^{-\lambda x}, x \geq 0$
- Porazdelitvena funkcija $F(x) = P(x \leq X) = 1 - e^{-\lambda x}$
- Oblika je vedno enaka, le razpršenost pada z naraščanjem parametra λ (asimetričnost je konstantno 2)
- Če se število dogodkov v t sekundah porazdeljuje po Poissonovi porazdelitvi (s parametrom λt), sledi verjetnost, da bomo v času t dočakali dogodek, eksponentni porazdelitvi (s parametrom λ)
- 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 $1/\lambda$)



Druge povezane porazdelitve

- Negativna binomska porazdelitev
  - v primerjavi s Poissonovo nadrazpršena
  - primer: zajedavcev na glodavca
  - odnos do binomske porazdelitve analogen odnosu eksponentne porazdelitve do Poissonove
- Sestavljena Poissonova porazdelitev (*compound Poisson distribution*)
  - porazdelitev vsote medsebojno neodvisnih in enako porazdeljenih (poljubno, čeprav praviloma nenegativnih) slučajnih spremenljivk, pri čemer je število slučajnih spremenljivk, ki jih seštejemo, Poissonovo porazdeljena slučajna spremenljivka
  - uporabna v aktuarstvu pri ocenjevanju tveganja zaradi predvidenih odškodnin in s tem pri določanju zavarovalnin
  - njen poseben primer je negativna binomska porazdelitev
- Conway-Maxwell-Poissonova porazdelitev (COM-Poissonova)
  - ima dodaten parameter (ν), ki si ga lahko predstavljamo kot faktor, ki omogoča, da se hitrost poissonovega procesa spreminja
  - njena posebna primera sta Poissonova porazdelitev ($\nu=1$) in geometrična porazdelitev ($\nu=0$), limitni primer pa Bernoullijeva porazdelitev ($\nu\to\infty$)

Poissonove zmesi

- Zmes dveh Poissonovih slučajnih spremenljivk: pri vsakem poskusu po slučaju (tj. z realizacijo Bernoullijeve slučajne spremenljivke s parametrom p) odločimo, ali bomo izžrebali slučajno vrednost iz ene Poissonove porazdelitve (s parametrom λ_A) ali iz druge (s parametrom λ_B)

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

- Z manjšanjem razlike med parametroma postaja porazdelitev zmesi unimodalna (zlasti, če je p blizu 0 ali 1), z večanjem razlike med parametroma pa vse bolj izrazito bimodalna (zlasti pri p blizu 0,5)
- **Excel 14**
- Zmes dveh (ali nekaj) Poissonovih porazdelitev je najpreprostejši primer nehomogenega Poissonovega procesa, kjer se λ s časom spreminja
  - modeli številnih stohastičnih pojavov, npr. klicev v klicni center za nujno pomoč ali priletov letal v zračni prostor določenega letališča
  - gostota toka se spreminja tekom dneva, pa tudi sezonsko v tednu in skozi leto

Dvorazsežna Poissonova porazdelitev

- Če imamo tri Poissonove slučajne spremenljivke in jih seštejemo v dve novi slučajni Poissonovi spremenljivki $\begin{matrix} X = X_1 + X_0 \\ Y = X_2 + X_0 \end{matrix}$
  je skupna porazdelitev teh dveh spremenljivk bivariatna Poissonova:
  $$(X,Y) \sim BP(\lambda_1, \lambda_2, \lambda_0)$$
- Robni porazdelitvi sta $X \sim Pois(\lambda_1 + \lambda_0)$ in $Y \sim Pois(\lambda_2 + \lambda_0)$
  Kovarianca X in Y je λ_0
- Uporabna kot model za napovedovanje športnih izidov, zlasti nogometnih tekem (števili zadetkov, ki ju dosežeta moštvi, med seboj nista neodvisni)
- Za modeliranje športnih izidov pomembna tudi porazdelitev razlike X in Y
  - verjetnost za zmago (oziroma neodločen izid ali poraz) z določeno razliko zadetkov je neodvisna od λ_0 in je znana kot Skellamova porazdelitev
  - izračunamo lahko verjetnost neodločenega izida glede na pričakovano število zadetkov ("moč") obeh moštev
  - če obdržimo konstantno moč nasprotnega moštva (tj. pričakovano število zadetkov, ki jih bo prvo moštvo prejelo), dobimo verjetnosti za posamezne razlike v zadetkih glede na moč prvega moštva – **Excel 15**

Najpreprostejši statistični test

- Najpomembnejši izid v številnih raziskavah v medicini je izid bolezni
  - ali se je izbrani dogodek (npr. ozdravitev) zgodil ali ne
- Randomiziran klinični poskus z dvema enako velikima skupinama
  - število udeležencev, ki jih je doletel dogodek: v eni skupini x_1, v drugi x_2
  - skupini sta zaradi randomizacije neodvisni
- Testna statistika za testiranje ničelne domneve o enaki pogostnosti dogodka v obeh populacijah, iz katerih smo vzorčili skupini:

$$z = \frac{x_1 - x_2}{\sqrt{x_1 + x_2}} \sim N(0,1)$$

- Izpeljava:
  - 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
- Primer: randomizirana študija učinkovine moksonidin v primerjavi s placebom pri srčnem popuščanju (objavljena 2004), izid umrljivost
  - ob vmesni analizi je med 1860 pacienti prišlo do 46 smrti v skupini z moksonidinom in 25 smrti v skupini s placebom
  - $z = (46 - 25) / \sqrt{(46 + 25)} = 2{,}49 \Rightarrow p = 0{,}013$
  - zato so klinični poskus predčasno ustavili

Poissonova regresija

- Spada med posplošene linearne modele (*generalised linear models*)
  - logaritemska vez (*link function*)
  - Poissonovo porazdeljena napaka
- Kot vsak regresijski model je tudi Poissonov lahko preprost, torej z eno neodvisno spremenljivko (prediktorjem, napovednim dejavnikom), ali multipli (z več neodvisnimi spremenljivkami)
- Model je multiplikativen: povečanje neodvisne spremenljivke x_j za eno enoto prinese množenje odvisne spremenljivke z e^{β_j}
- Podatki so lahko v združeni (agregirani) ali posamični obliki (individualni, običajni v uporabni statistiki)
- Socioekonomski primer z agregiranimi podatki:
  - skupno 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)

Kontrolne karte

- Statistični nadzor procesov oz. kakovosti (*SPC* oz. *SQC*)
- Shewhartove kontrolne karte
  - za spremenljivke (tj. številske podatke)
  - za atribute (tj. štetje diskretnih izidov; sem sodita Poissonovi)
- *c*-karta za nadzor števila okvar (*count*)
  - Za razliko od binomske *p*-karte (za delež okvarjenih enot) dopušča več okvar na enoto, po drugi strani pa zahteva stalno velikost vzorca (za razliko od *p*-karte in *u*-karte)
  - primer: število izdelkov, vrnjenih v trgovino zaradi reklamacije, na dan
  - meje nadzora pri $\bar{c} \pm 3\sqrt{\bar{c}}$
  - **Excel 15**
- *u*-karta za nadzoru stopnje okvar na enoto (*rate*)
  - primer: število padcev bolnikov v bolnišnici na dan
  - meje nadzora (se spreminjajo glede na n_j) pri $\bar{u} \pm 3\sqrt{\frac{\bar{u}}{n_j}}$
  - **Excel 16**
- Lahko upoštevamo dodatna kontrolna pravila (npr. Western Electric)

Poissonov detektiv

- Detektiv želi prijeti serijskega avtomobilskega tatu, za katerega je znano, da deluje na določeni ulici v mestu
  - Tat ulico obiskuje ob slučajno izbranih dnevih v letu ob slučajno izbranem času dneva; v povprečju izpelje 10 uspešnih tatvin letno
  - Detektiv lahko posveti opazovanju te ulice 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?
- Rešitev:

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 $\left(\frac{3}{24}\cdot\frac{3}{7}\cdot\frac{1}{52}=\frac{9}{8736}\right)$*, 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%* $\left(P_{52} = 1 - e^{-52/100}\right)$.