Binomska in Poissonova porazdelitev

Uvod

V tej vaji se bomo seznanili z binomsko in Poissonovo porazdelitvijo. Pogledali bomo, kako posamezni parametri vplivajo na potek porazdelitev. Prepričali se bomo, v katerih primerih je upravičeno za izračun verjetnosti binomske porazdelitve uporabiti približek s Poissonovo porazdelitvijo in kdaj lahko verjetnosti Poissonove porazdelitve izračunamo s približkom z normalno (Gaussovo) porazdelitvijo. V obeh primerih bomo primernost aproksimacije ovrednotili z izračunom relativne napake.

Teoretično ozadje

Binomska porazdelitev

Izhodišče binomske porazdelitve je Bernoullijev poskus, ki je definiran kot poskus, pri katerem sta možna le dva izida. Navadno ju imenujemo ugodni in neugodni izid, čeprav je kvalitativnost teh dveh oznak povsem poljubna. Verjetnost ugodnega izida označimo s p, verjetnost neugodnega izida je potem $q = 1 - p$. Izidi posameznih Bernoullijevih poskusov so med seboj neodvisni. Če Bernoullijev poskus ponovimo n-krat, pri čemer je verjetnost za vsakokratni ugodni izid p, potem se lahko vprašamo po verjetnosti za $k(k = 0, 1, 2, ..., n)$ ugodnih izidov med n vseh izidov. Te verjetnosti opisujejo binomsko porazdelitev oziroma porazdelitev diskretne naključne spremenljivke X, ki se podreja binomskemu zakonu. Verjetnost za k ugodnih izidov izračunamo po binomski formuli:

$$P_X(k) = P[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

Funkcijo porazdelitve verjetnosti za binomsko naključno spremenljivko pa definira izraz:

$$F_X(k) = P[X \leq k] = \sum_{k \leq x} \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

Kot vemo, omogoča funkcija porazdelitve verjetnosti izračun katerekoli verjetnosti v zvezi z dotično naključno spremenljivko.

Poissonova porazdelitev kot približek binomske porazdelitve

Pri binomski porazdelitvi postanejo z naraščanjem velikosti vzorca izračuni verjetnosti težavni. Pod določenimi pogoji lahko te izračune nadomestimo s približkom, ki upošteva Poissonovo porazdelitev. Pogoji za upravičenost aproksimacije binomske porazdelitve s Poissonovo so:

$$n \gg 1; p \ll 1; k \ll n$$

Tedaj velja:

$$\lim_{\substack{x\to\infty \\ p\to 0 \\ k\ll n}} \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{(np)^k}{k!} e^{-np}; a = np = \lambda\tau \tag{3}$$

Parameter a je Poissonov parameter in ima lahko različne interpretacije. Produkt np, ki izhaja iz binomske porazdelitve, je ena od teh interpretacij. Poissonova porazdelitev pa seveda ni le limita binomske porazdelitve. Poissonovo porazdelitev ima veliko naravnih pojavov. Pri naključnih spremenljivkah, ki se pokoravajo Poissonovemu zakonu, nimamo opravka s parametroma n in p, ampak s parametroma λ (gostota: število elementov v enoti časa, volumna, dolžine,...) in τ (interval: časa, volumna, dolžine,...). Kot primer Poissonove naključne spremenljivke omenimo število radioaktivnih razpadov nekega elementa (k) v nekem časovnem intervalu (τ), če poznamo povprečno število razpadov na časovno enoto (λ). Pri Poissonovi porazdelitvi se torej sprašujemo po verjetnostih za k elementov v intervalu τ, če je povprečna gostota elementov λ. To verjetnost torej izračunamo kot:

$$P_X = P[X = k] = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau} \tag{4}$$

Funkcijo porazdelitve verjetnosti za Poissonovo naključno spremenljivko pa definira izraz:

$$F_X(x) = P[X \leq x] = \sum_{k\leq x} \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau} \tag{5}$$

Gaussova (normalna) porazdelitev kot približek Poissonove porazdelitve

Računanje verjetnosti s Poissonovo porazdelitvijo lahko postane težavno ali nepraktično. Pod določenimi pogoji lahko te izračune nadomestimo s približkom, ki upošteva Gaussovo (normalno) porazdelitev. Pogoj za upravičenost aproksimacije Poissonove porazdelitve z Gaussovo je:

$$a \gg 1 \ (np \gg 1;\ \lambda\tau \gg 1)$$

Tedaj velja:

$$P[a \leq X \leq b] = \sum_{k=a}^{b} e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!} \simeq \frac{1}{\sqrt{2\pi}} \int_{l_1}^{l_2} e^{-\frac{1}{2}t^2} dt \tag{6}$$

Integracijski meji iz zgornjega izraza pa sta:

$$l_1 = \frac{a - \lambda\tau - 0.5}{\sqrt{\lambda\tau}};\ l_2 = \frac{b - \lambda\tau + 0.5}{\sqrt{\lambda\tau}} \tag{7}$$

Če želimo izračunati en sam člen iz vsote ($k = a = b$), je treba v izrazih za integracijski meji zamenjati a in b s k. Integral v enačbi 6 ni analitično

rešljiv, zato ga izračunamo numerično oziroma z uporabo tabel, ki vsebujejo izračunane vrednosti integralov tako imenovane funkcije napake $erf(x)$ (angl. *error funcion*). Velja:

$$\frac{1}{\sqrt{2\pi}}\int_{l_1}^{l_2} e^{-\frac{1}{2}t^2}dt = \frac{1}{\sqrt{2\pi}}\int_{0}^{l_2} e^{-\frac{1}{2}t^2}dt - \frac{1}{\sqrt{2\pi}}\int_{0}^{l_1} e^{-\frac{1}{2}t^2}dt = erf(l_2) - erf(l_1) \tag{8}$$

Naloge

1. Naloga 1 - Binomska verjetnost

   Statistično so ugotovili, da je zastopanost posameznih krvnih skupin (0, A, B in AB) v populaciji belcev približno sledeča[1]:

   $$p_0 = 0,47;\ p_A = 0,41;\ p_B = 0,09;\ p_{AB} = 0,03$$

   Denimo, da iz populacije belcev naključno izberemo n prostovoljcev. Izberimo si določene lastnosti in glede na te lastnosti definirajmo naslednje naključne spremenljivke:

   X_A: število oseb s krvno skupino A;

   X_{AB}: število oseb s krvno skupino AB;

   X_{neB}: število oseb, ki *nimajo* krvne skupine B;

   X_{0+AB}: število oseb s krvno skupino 0 ali AB.

   - Za tako definirane naključne spremenljivke izračunajte in grafično predstavite verjetnosti $P_{Xi}(k)$, da med $n = 30$ prostovoljci najdemo k oseb ($k = 0, 1, 2, ..., n$) z iskano lastnostjo. Oglejte si in komentirajte vpliv vrednosti elementarnih verjetnosti p_i na lego in obliko porazdelitvenih krivulj $P_{Xi}(k)$.
   - Izračunajte pričakovane vrednosti $E[X]$ naključnih spremenljivk in jih primerjajte s porazdelitvami $P_{Xi}(k)$. Kaj ugotovite?

   $$E[X] = \sum_{k=0}^{n} x_k P_x(k) = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}$$

2. Naloga 2 - Približek binomske z Poissonovo porazdelitvijo

   Statistično so ugotovili, da ima približno 15% populacije belcev negativen faktor Rh[2]. Če predpostavimo, da sta nastop posamezne krvne skupine (0, A, B ali AB) in negativnega faktorja Rh pri posamezniku neodvisna

[1] Vir fizioloških podatkov: A.C. Guyton and J.E. Hall: *Textbook of Medical Physiology*, 9th ed., WB Saunders Co., Philadelphia, 1996.

[2] Vir fizioloških podatkov: A.C. Guyton and J.E. Hall: *Textbook of Medical Physiology*, 9th ed., WB Saunders Co., Philadelphia, 1996.

dogodka, potem je na primer verjetnost, da ima naključno izbrani posameznik hkrati krvno skupino B in negativen faktor Rh enaka:

$$p = P[B \text{ in } Rh^-] = P[B] \cdot P[Rh^-]$$

Izberimo skupino $n = 100$ prostovoljcev in definirajmo novo naključno spremenljivko X, ki predstavlja število oseb, ki imajo obe lastnosti, torej hkrati krvno skupino B in negativen faktor Rh.

- Izračunajte in grafično predstavite verjetnosti $P_X(k)$ $(k = 0, 1, 2, ..., 20)$.
- Verjetnosti $P_X(k)$ izračunajte še s Poissonovo formulo in jih predstavite na istem grafu. Po potrebi uporabite logaritemsko merilo na osi y za boljšo primerjavo obeh porazdelitev. Glede na pogoje za veljavnost Poissonovega približka napovejte ali je tak približen ustrezen in če ja, za katere vrednosti k.
- Preverite svojo napoved v prejšnji točki, tako da določite relativno napako približka za vsak k. Relativno napako definiramo kot:

$$\eta = \frac{P_{X\ Poisson}(n) - P_{X\ binom}(n)}{P_{X\ binom}(n)}$$

- Sedaj fiksirajte vrednosti $k = 3$ in $p = P[B\text{in}Rh^-] = P[B] \cdot P[Rh-]$ ter spreminjajte velikost vzorca (vrednost n) med 5 in 100 s korakom 5. Z binomsko in Poissonovo formulo izračunajte verjetnosti $P_X(k)$ in jih podajte v odvisnosti od n. Podajte tudi relativno napako v odvisnosti od n.

  Komentirajte!
- Fiksirajte vrednosti $k = 3$ in $n = 100$ ter spreminjajte vrednost p med 0,01 in 0,25 s korakom 0,01. Z binomsko in Poissonovo formulo izračunajte verjetnosti $P_X(k)$ in jih podajte v odvisnosti od p. Podajte tudi relativno napako v odvisnosti od p.
- Preverite še kako se spreminja relativna napaka, če spreminjate velikost vzorca in verjetnost dogodka $P[B \text{ in } Rh^-]$!

  Kaj ugotovite?

3. Naloga 3 - približek Poissonove z Gaussovo porazdelitvijo

   Povprečna koncentracija različnih vrst belih krvnih celic v človeški krvi je približno naslednja[3]:

(65%)	granulociti:	$\lambda_{GRA} =$	$4550\mu l^{-1}$
(30%)	limfociti:	$\lambda_{LIM} =$	$2100\mu l^{-1}$
(5%)	monociti:	$\lambda_{MON} =$	$350\mu l^{-1}$
(100%)	vsi levkociti:	$\lambda_{LEV} =$	$7000\mu l^{-1}$

[3]Vir fizioloških podatkov: A.C. Guyton and J.E. Hall: *Textbook of Medical Physiology*, 9th ed., WB Saunders Co., Philadelphia, 1996.

Denimo, da vzamemo vzorec človeške krvi velikosti $10^{-3}\mu l$ in si ga ogledamo pod mikroskopom. Naključna spremenljivka X_i naj predstavlja število najdenih celic posamezne vrste v takem vzorcu.

- Izračunajte in grafično predstavite verjetnosti, da v vzorcu krvi velikosti $0,001\mu l$ najdemo k ($k = 0, 1, 2, 3, ..., 20$) levkocitov, granulocitov, limfocitov oziroma monocitov. Upoštevajte Poissonovo porazdelitev.
- Izračunajte in grafično predstavite približke za verjetnosti iz prejšnjega koraka z Gaussovo porazdelitvijo.
- Glede na pogoje za ustreznost približka napovejte za katere celice in za katere k je uporaba Gaussovega približka ustrezna.
- Preverite svojo hipotezo iz prejšnje točke, tako da izračunate in grafično prikažete absolutno in relativno napako, ki jo naredimo ob aproksimaciji Poissonove z Gaussovo porazdelitvijo, kot funkcijo parametra k za vse štiri skupine celic.

  Kaj ugotovite?

  Absolutna napaka:

$$\epsilon = P_{Xi\ Gauss}(k) - P_{Xi\ Poisson}(k)$$

  Relativna napaka:

$$\eta = \frac{P_{Xi\ Gauss}(k) - P_{Xi\ Poisson}(k)}{P_{Xi\ Poisson}(k)}$$

Priročni MATLAB ukazi

Spodaj imate navedene MATLAB ukaze, ki vam lahko zelo olajšajo delo pri vaji. Ne pozabite na pomoč (tipka F1)!

`n=factorial(k)`
Izračuna vrednost fakultete $n = k!$

`semilogy(x,y)`
Enako kot ukaz `plot(x,y)`, vendar uporabi logaritemsko skalo za os y.

`erf(x)`
Izračun določenega integrala normirane Gaussove porazdelitve. MATLABova definicija je nekoliko drugačna od definicije, ki smo jo vajeni, zato je treba spremeniti meje meje integracije. Definicija MATLAB:

$$erf_{MATLAB}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{9}$$

"Naša definicija":

$$erf(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt \tag{10}$$

Za pravilen izračun enačbe (8) z MATLABovo funcijo `erf(x)` moramo zato spremeniti integracijsko območje na naslednji način:

$$\frac{1}{\sqrt{2\pi}} \int_{l_1}^{l_2} e^{-\frac{1}{2}t^2} dt = erf(l_2) - erf(l_1) = \frac{erf_{MATLAB}(\frac{l_2}{\sqrt{2}}) - erf_{MATLAB}(\frac{l_1}{\sqrt{2}})}{2} \tag{11}$$