Uporabna statistika

Gregor Dolinar

Fakulteta za elektrotehniko
Univerza v Ljubljani

6. november 2013

Posamezne verjetnosti dvorazsežnih slučajnih spremenljivk

Definicija

Naj bo $f_{XY}(x,y)$ gostota verjetnosti dvorazsežne slučajne spremenljivke (X,Y). Potem sta posamezni gostoti verjetnosti diskretnih slučajnih spremenljivk X in Y

- $f_X(x) = P[X = x] = \sum_{R_x} f_{XY}(x,y)$,
  R_x množica vseh točk za katere je $X = x$.
  $p_{i.} = P[X = x_i] = \sum_k p_{ik}$
- $f_Y(y) = P[Y = y] = \sum_{R_y} f_{XY}(x,y)$,
  R_y množica vseh točk za katere je $Y = y$.
  $p_{.k} = P[Y = y_k] = \sum_i p_{ik}$

Trditev

$E(X) = \mu_X = \sum_x x f_X(x) = \sum_x x(\sum_{R_x} f_{XY}(x, y)) = \sum_R x f_{XY}(x, y)$

$V(X) = \sigma_X^2 = \sum_R (x - \mu_X)^2 f_{XY}(x, y)$

Pogojna verjetnost

Primer

Prejeli smo 4 bite, prejeti biti so lahko dobri, sumljivi ali slabi.
Označimo: X število dobrih bitov ($p = 0.9$), Y število sumljivih bitov ($p = 0.08$).
$P[Y = 0|X = 4] = 1.$

$$P[Y = 2|X = 2] = P[X = 2, Y = 2]/P[X = 2]$$

$$= \frac{\frac{4!}{2!\cdot 2!}0.9^2 \cdot 0.08^2}{\binom{4}{2}0.9^2 \cdot 0.1^2} = 0.64$$

Definicija

Naj bo $f_{XY}(x,y)$ gostota verjetnosti dvorazsežne slučajne spremenljivke (X,Y). Potem je pogojna gostota verjetnosti diskretne slučajne spremenljivk Y pri pogoju $X=x$ enaka

$$f_{Y|x}(y)=P[Y=y|X=x]=\frac{f_{XY}(x,y)}{f_X(x)}.$$

Trditev

Naj bo $f_{XY}(x, y)$ gostota verjetnosti dvorazsežne slučajne spremenljivke (X, Y). Potem sta diskretni slučajni spremenljivki X in Y neodvisni natanko tedaj, ko velja katerakoli od trditev

- $f_{XY}(x, y) = f_X(x) f_Y(y)$
- $f_{Y|x}(y) = f_Y(y)$ *za vsak* x, y, $f_X(x) > 0$
- $f_{X|y}(x) = f_X(x)$ *za vsak* x, y, $f_Y(y) > 0$
- $P[X \in A, Y \in B] = P[X \in A] P[Y \in B]$.

Opomba

Če množica točk z neničelno verjetnostjo dvorazsežne slučajne spremenljivke (X, Y) ne tvori pravokotnika, potem spremenljivki X in Y nista neodvisni (pozitivna verjetnost ene spremenljivke omeji zalogo vrednosti druge spremenljivke). Če množica točk z neničelno verjetnostjo tvori pravokotnik, sta slučajni spremenljivki lahko neodvisni, vendar ne nujno (pravokotnik je potreben, ni pa zadosten pogoj za neodvisnost).

Večrazsežne diskretne slučajne spremenljivke

Primer

Digitalni prenos informacij. Prejeti bit karakteriziramo kot: zelo dober (p_1), dober (p_2), sumljiv (p_3), slab (p_4).
Prejeli smo n bitov. Označimo:

- X_1 število zelo dobrih
- X_2 število dobrih,
- X_3 število sumljivih
- X_4 število slabih

Definicija

Gostota verjetnosti n-razsežne slučajne spremenljivke $(X_1, \ldots, X_n)$ diskretnih slučajnih spremenljivk $X_1, \ldots X_n$ je

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = P[X_1 = x_1, \ldots, X_n = x_n].$$

Trditev

Diskretne slučajne spremenljivke $X_1, \ldots X_n$ so neodvisne natanko tedaj, ko je

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n).$$

Multinomska slučajna spremenljivka

Posplošitev binomske slučajne spremenljivke.

Primer

Digitalni prenos informacij. Prejeti bit:
zelo dober ($p = 0.6$), dober ($p = 0.3$), sumljiv ($p = 0.08$), slab ($p = 0.02$).
Prejeli smo 20 bitov. Kolikšna je verjetnost, da je od tega 15 zelo dobrih, 3 dobri, 2 sumljiva in noben slab?

$$P = \frac{20!}{15! \cdot 3! \cdot 2! \cdot 0!} \cdot 0.6^{15} \cdot 0.3^{3} \cdot 0.08^{2}.$$

Definicija

Multinomska verjetnostna porazdelitev. Denimo, da n-krat ponovimo poskus:

- rezultat vsakega poskusa pripada enemu izmed k razredov
- rezultat poskusa je v i-tem razredu z verjetnostjo p_i, $i = 1, \ldots, k$
- poskusi so med sabo neodvisni

Z X_i označimo slučajno spremenljivko, ki označuje število poskusov, ki so v i-tem razredu.

Definicija

Potem je

$$P[X_1 = x_1, \ldots, X_n = x_n] = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

za $x_1 + \cdots + x_k = n$ in $p_1 + \cdots + p_k = 1$.

Primer

Pri transportu elektronskih naprav so se 4 naprave prevrnile. V preteklosti je imelo 60 % prevrnjenih naprav večje poškodbe, 30 % manjše poškodbe, 10 % prevrnjenih naprav pa ni bilo poškodovanih.

- Kolikšna je verjetnost, da imata 2 večje poškodbe in 2 manjše poškodbe? 0.1944
- Kolikšna je verjetnost, da nobena naprava ni poškodovana? 0.0001

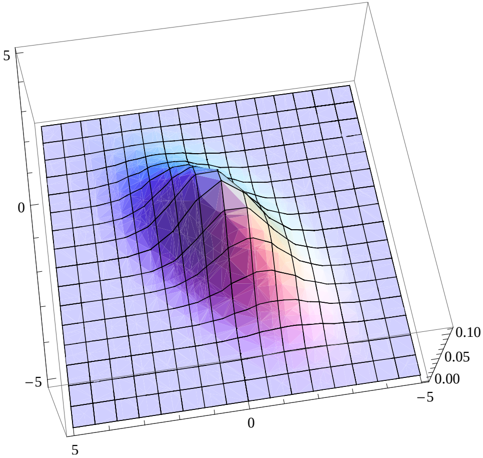
Primer

- Izdelek iz plastične mase. Opazujemo širino X in višino Y izdelka. Dimenziji npr. odvisni od pritiska, temperature.
- Čas X, da se računalnik poveže z glavnim računalnikom. Čas Y, da glavni računalnik avtorizira povezavo.

Definicija

Za gostoto verjetnosti $f_{XY}(x, y)$ dvorazsežne slučajne spremenljivke (X, Y) zveznih slučajnih spremenljivk X in Y velja

- $f_{XY}(x, y) \geq 0$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dxdy = 1$
- za vsako območje R dvodimenzionalnega prostora je $P[(X, Y) \in R] = \int \int_R f_{XY}(x, y) dxdy$.


5
0
−5
5
0
−5
0.10
0.05
0.00

Dvorazsežna normalna porazdelitev

Definicija

Dvorazsežna normalna porazdelitev

$$f_{XY} = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} e^{\frac{-1}{2(1-\rho_{XY}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}$$

Kovarianca in korelacija

Definicija

Kovarianca med slučajnima spremenljivkama X in Y je definirana s predpisom $\mathrm{cov}(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

$= E[XY] - \mu_X \mu_Y$.

Definicija

Korelacija med slučajnima spremenljivkama X in Y je definirana s predpisom

$$\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Trditev

$$-1 \leq \rho_{XY} \leq 1$$

Vzorčenje

Definicija

Populacija je množica vseh rezultatov poskusov oziroma vseh opazovanj pojava, ki ga obravnavamo.

Primer. Narodna pripadnost prebivalcev neke države.
Velikokrat kot matematični model za populacijo uporabimo verjetnostno porazdelitev.
Primer. Dolžina stranice sestavnega dela je normalno porazdeljena (populacija je normalno porazdeljena).

V večini primerov ne opazujemo celotne populacije.

Definicija

Vzorec je podmnožica opazovanj, ki jih izberemo iz populacije.

Na podlagi analize vzorca sklepamo o populaciji.
Vzorec mora biti reprezentativen, ne sme biti pristranski.

Slučajne spremenljivke $X_1, \ldots, X_n$ so slučajni vzorec velikosti n, če velja

- X_i so neodvisne slučajne spremenljivke
- vse spremenljivke X_i imajo enako verjetnostno porazdelitev

Primer

Zanima nas telesna teža prebivalcev Slovenije.
Populacija so vse možne telesne teže prebivalcev.
Izberemo del prebivalcev in izmerimo njihove telesne teže. Vzorec so potem telesne teže izbranih prebivalcev.

Definicija

Statistika je katerakoli funkcija na slučajnem vzorcu.

Primer. Povprečje vzorca

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

Povprečje vzorca je približek za povprečje celotne populacije.

Definicija

Varianca vzorca

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}.$$

Opomba. Varianca populacije

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

Zakaj $n-1$?

Predstavitev podatkov

Srednje vrednosti:

- mediana - vrednost, ki je na sredini ranžirne vrste izmerjenih vrednosti
- povprečna vrednost (aritmetična sredina)
- modus - vrednost, ki se največkrat pojavi

Primer

Telesne teže študentov: 56 kg, 92 kg, 60 kg, 82 kg, 81 kg, 74 kg, 79 kg, 75 kg, 78 kg, 52 kg, 75 kg, 84 kg.

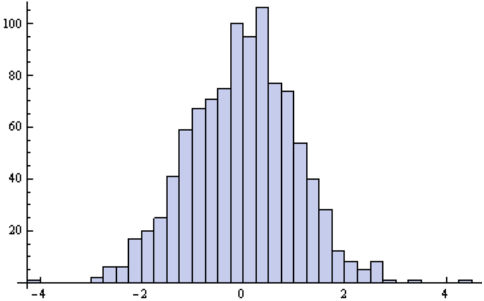
Mediana:

Povprečna vrednost:

Modus:

Podatke lahko predstavimo na več načinov:

- frekvenčna porazdelitev
- histogram



($\sqrt{n}$ razredov)

- škatlasti diagram
- časovne vrste
- verjetnostna mreža
- drevesasti grafikon (stem and leaf diagram)