Neformalan uvod u osnovne pojmove verovatnoće i statistike

Mladen Nikolić

Grupa za mašinsko učenje
`machinelearning.math.rs`

Matematički fakultet
Univerzitet u Beogradu

Ishodi i događaji

- Eksperimenti imaju ishode
  - Ishodi bacanja kockice su brojevi od 1 do 6
- Događaji su skupovi ishoda
  - Događaj može biti da je dobijen broj veći od 3, što je skup ishoda $\{4, 5, 6\}$
- Kažemo da se neki događaj desio ako se desio neki ishod iz tog događaja

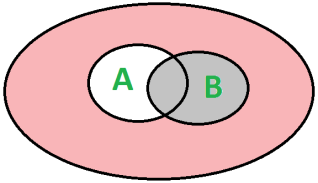
Verovatnoća događaja

- Verovatnoće su dugoročne frekvencije događaja
- Postoje drugačije interpretacije pojma verovatnoće (npr. Bajesova)
- Formalno, govorimo o verovatnosnoj meri P koja mora da zadovolji sledeće aksiome:
  - $P(\Omega) = 1$, gde je Ω skup svih ishoda
  - $P(A) \geq 0$, za svaki događaj $A \subseteq \Omega$
  - $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ ako su A_i disjunktni događaji

Uslovna verovatnoća

- $P(A|B)$ je verovatnoća događaja A pri uslovu B i definiše se kao

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Uslovna verovatnoća – primer

- Ako znamo da je broj koji je kockica dala neparan, koja je verovantoća da je to baš 3?

$$P(\{3\}|\{1,3,5\}) = \frac{P(\{3\})}{P(\{1,3,5\})} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Nezavisnost događaja

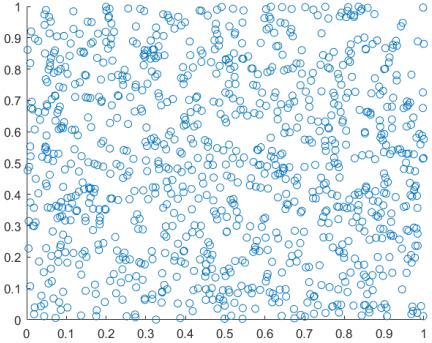
- Događaji A i B su nezavisni ako važi

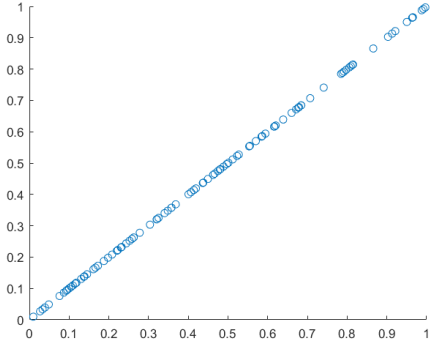
$$P(A|B) = P(A)$$

odnosno

$$P(A \cap B) = P(A)P(B)$$

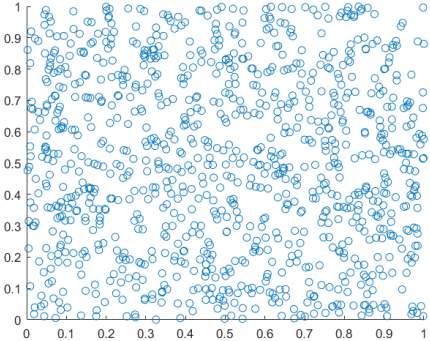
Nezavisnost događaja

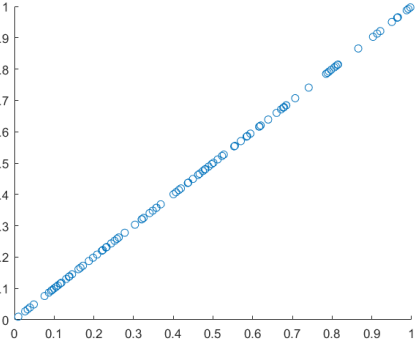




- Posmatrajmo događaje $x > 0.5$, $y > 0.5$ i $x > 0.5 \wedge y > 0.5$
- Koji slučaj odražava zavisne, a koji nezavisne događaje?

Nezavisnost događaja





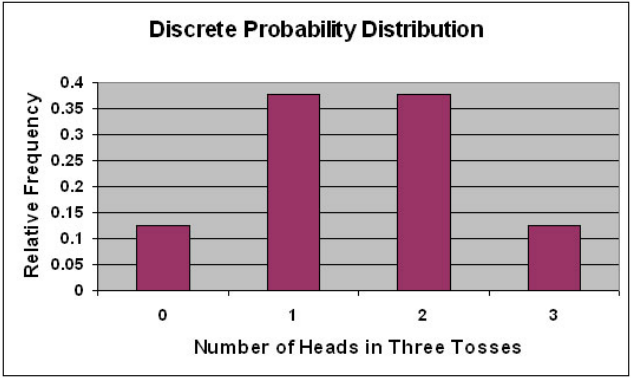
- Posmatrajmo događaje $x > 0.5$, $y > 0.5$ i $x > 0.5 \wedge y > 0.5$
- Koji slučaj odražava zavisne, a koji nezavisne događaje?
- U prvom se vidi nezavisnost, a verovatnoće su 0.5, 0.5 i 0.25
- U drugom se vidi zavisnost, a verovatnoće su 0.5, 0.5 i 0.5

Slučajna promenljiva

- Funkcije koje preslikavaju ishode u skup realnih brojeva nazivaju se slučajnim veličinama
- Na primer visina slučajno izabrane osobe
- Opisujemo ih pomoću pojma raspodele verovatnoće

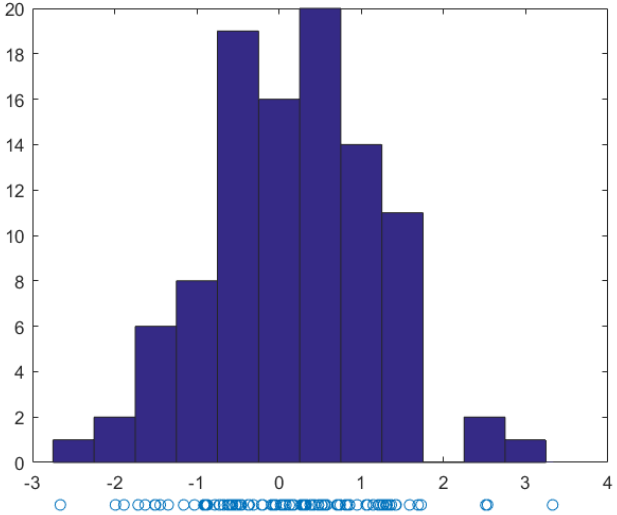
Diskretna raspodela verovatnoće

- Pridružuje verovatnoću svakom od prebrojivo mnogo vrednosti slučajne promenljive

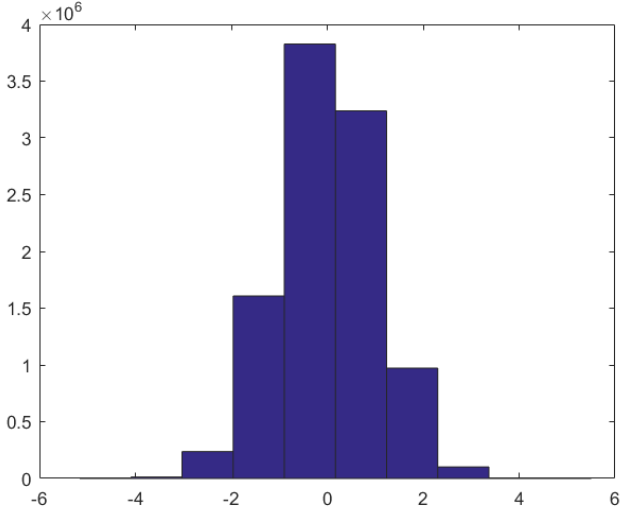


Slika: http://www2.cedarcrest.edu/academic/bio/hale/biostat/session10links/probgraph.html

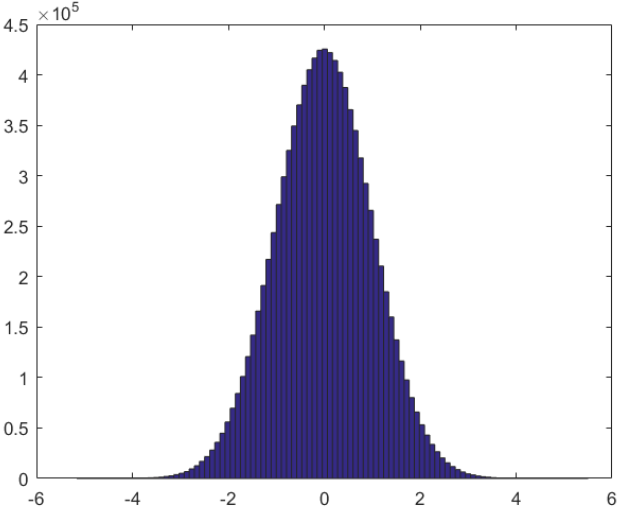
Histogram i neprekidna raspodela verovatnoće (1)



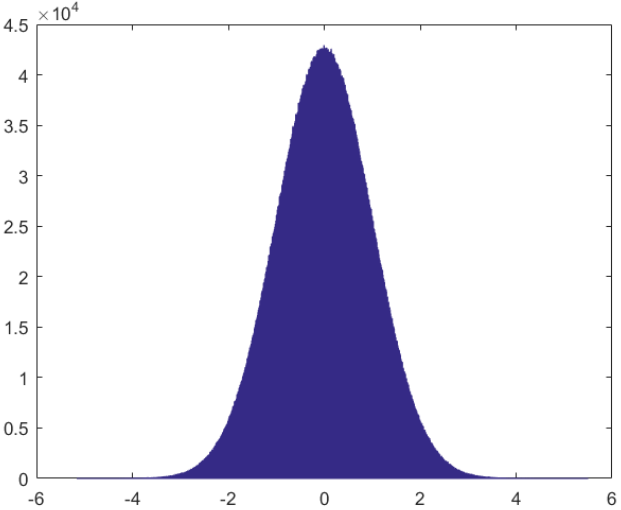
Histogram i neprekidna raspodela verovatnoće (2)



Histogram i neprekidna raspodela verovatnoće (3)



Histogram i neprekidna raspodela verovatnoće (4)



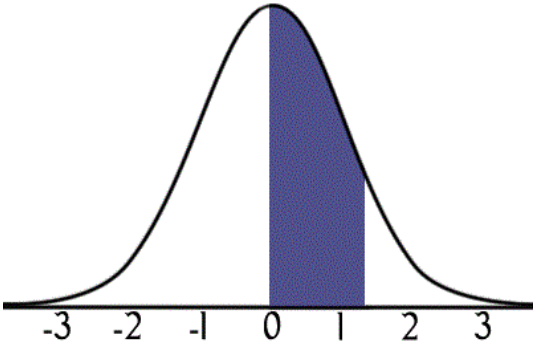
Gustina raspodele

- Histogram oslikava *gustinu raspodele*
- Mora da važi

$$\int p(x)dx = 1$$

Gustina raspodele

$$P([a,b]) = \int_a^b p(x)dx$$

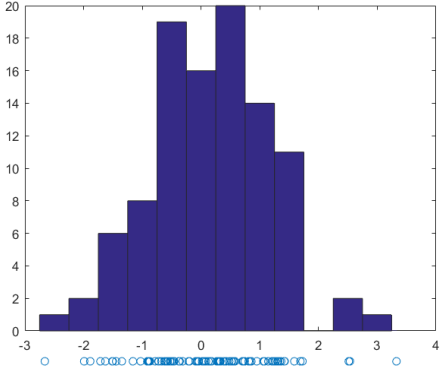


Slika: http://www.disfrutalasmatematicas.com/datos/distribucion-normal-estandar.html

Matematičko očekivanje

- Intuitivno, predstavlja srednju vrednost neke slučajne veličine
- Očekivanje slučajne veličine X se označava $\mathbb{E}[X]$
- Postoje i drugi načini da se definiše srednja vrednost (npr. medijana) koji mogu biti pogodniji u nekim kontekstima

Matematičko očekivanje



$$\mathbb{E}[X] \approx \sum_i x_i \frac{n_i}{N} = -2.5\frac{1}{100} - 2\frac{2}{100} - 1.5\frac{6}{100} - 1\frac{8}{100} - 0.5\frac{19}{100} + 0\frac{16}{100} + 0.5\frac{20}{100} + 1\frac{14}{100} + 1.5\frac{11}{100} + 2\frac{0}{100} + 2.5\frac{2}{100} + 3\frac{1}{100}$$

Matematičko očekivanje

- Za široku klasu raspodela važi (u praksi, praktično uvek):

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx$$

- Važi

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$$

Varijansa (disperzija)

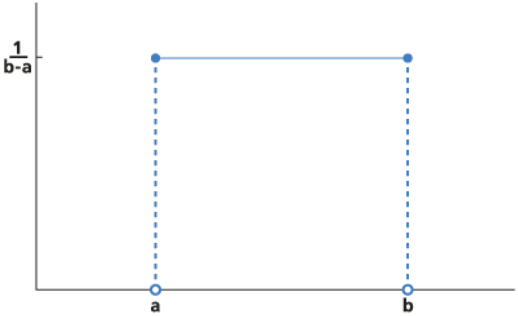
- *Varijansa* je mera raspršenosti slučajne promenljive i definiše se kao

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- Predstavlja očekivanje odstupanja slučajne promenljive od njenog očekivanja
- Koren varijanse se naziva *standardna devijacija*

Uniformna raspodela

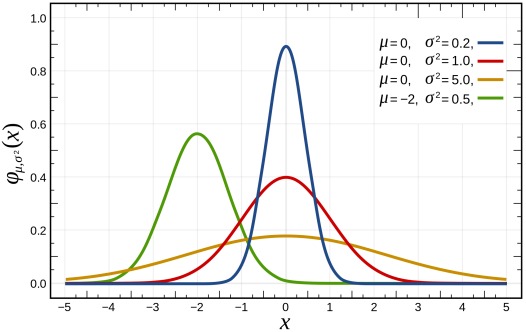
$$X \sim \mathcal{U}(a, b) \qquad p(x) = \frac{1}{b-a} \qquad \mathbb{E}[X] = \frac{a+b}{2} \qquad Var(X) = \frac{1}{12}(b-a)^2$$



Slika: http://pro.arcgis.com/en/pro-app/tool-reference/data-management/distributions-for-assigning-random-values.htm

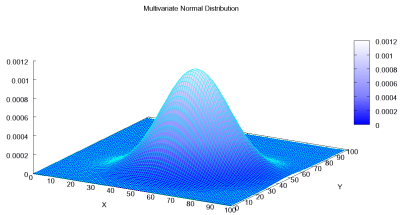
Normalna raspodela

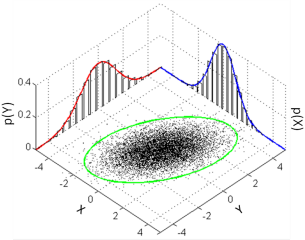
$$X \sim \mathcal{N}(\mu, \sigma) \qquad p(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mathbb{E}[X] = \mu \qquad Var(X) = \sigma^2$$

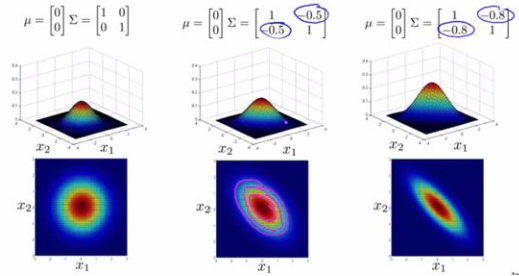


Slika: https://en.wikipedia.org/wiki/Normal_distribution

Visedimenziona normalna raspodela







Statistika kao disciplina

- Donosi zaključke o fenomenima na osnovu uzoraka iz iskustva
- Osnovni zadatak matematičke statistike je izabrati, iz skupa dopustivih verovatnosnih modela (npr. gustina raspodele), jedan koji najbolje odgovara uzorku podataka
- Često se ovaj, a i drugi problemi statistike, svode na ocenu nekih nepoznatih parametara verovatnosnog modela, na osnovu podataka

Statistika – primer

- Kako da izračunamo verovatnoću da slučajno izabrana osoba ima visinu veću od 180cm?
- Kako da izračunamo verovatnoću da slučajno izabrana osoba ima IQ veći od 120?
- Ovo su fenomeni koji prate normalnu raspodelu!
- Ako nekako procenimo prosek i varijansu normalne raspodele, onda površina ispod grafika te normalne raspodele desno od izabrane tačke daje željenu verovatnoću!

Statistika kao funkcija

- Funkcija koja slika uzorak u realan broj i *ima još neka lepa svojstva*, naziva se *statistika*
- Na primer – prosek uzorka

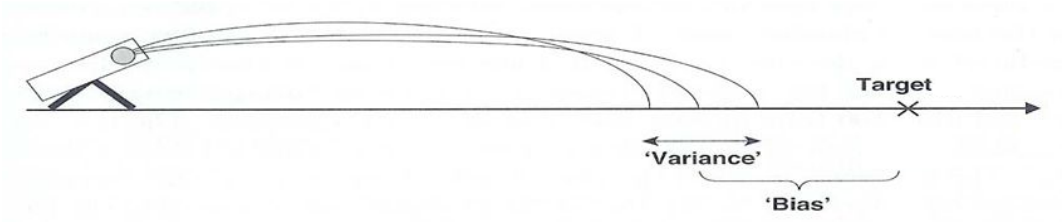
Poželjna svojstva statistika

- Neka je $\hat{\theta}$ ocena parametra θ
- Ocena $\hat{\theta}$ je *nepristrasna* (*centrirana*) ako važi

$$\mathbb{E}[\hat{\theta}] = \theta$$

- $\mathbb{E}[\hat{\theta}] - \theta$ je *sistematsko odstupanje* ocene $\hat{\theta}$
- Nepristrasna ocena koja ima manju varijansu je *bolja* od nepristrasne ocene koja ima veću varijansu

Ilustracija sistematske ocene i varijanse statistike



Slika: P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining. Modifikovano.