Primijenjena i inženjerska matematika

Statistika - predavanje 3

Osnove vjerojatnosti

20.10.2023.

Motivacija

- cilj: na osnovu podataka donositi zaključke o cijeloj populaciji
- zaključci će imati dozu nesigurnnosti zbog nepotpunih podataka i slučajnog odabira uzorka
- želimo kontrolirati tu nesigurnost
- podatak iz baze podataka promatrat ćemo kao jednu vrijednost (realizaciju) obilježja koje proučavamo
- prvenstveno želimo znati **koliko je izvjesno** da to obilježje poprimi određene vrijednosti, tj. **kolika je vjerojatnost** da to obilježje poprimi određene vrijednosti

Motivacija

Kolika je vjerojatnost da slučajno odabrana osoba ima plave oči?

Kolika je vjerojatnost da pri bacanju novčića padne glava?

Kolika je vjerojatnost da se pri bacanju igraće kockice okrene šestica?

Određivanje vjerojatnosti

- određivanje vjerojatnosti u konkretnim problemima temelji se na dosadašnjem iskustvu u istraživanju i može biti vrlo složen postupak
- spomenut ćemo dva načina određivanja vjerojatnosti:
  - **klasična metoda** – pod uvjetom da obilježje može poprimiti konačno mnogo vrijednosti i da su sve vrijednosti jednako moguće
  - **statistička definicija vjerojatnosti**

Vjerojatnost - statistički pristup

- **slučajni pokus** - aktivnost za koju znamo skup svih mogućih ishoda, ali zbog nemogućnosti sagledavanja svih utjecaja, u svakom pojedinom provođenju te aktivnosti ne znamo koji će se ishod (od mogućih) realizirati
- **nezavisno izvođenje (ponavljanje) pokusa** - činjenica da se dogodio neki događaj prilikom izvođenja jednog od njih ne mijenja šanse za realizaciju bilo kojeg događaja drugog pokusa
  - primjer: bacanje novčića ili igraće kockice dva puta - nezavisno
  - primjer: izvlačenje drugog broja u igri loto - zavisno
- ako je pokus takav da ga možemo nezavisno ponavljati mnogo puta, relativna frekvencija pojavljivanja događaja A će se s povećanjem broja ponavljanja pokusa **stabilizirati** oko nekog broja koji predstavlja statistički određenu vjerojatnost pojavljivanja događaja A

Vjerojatnost - statistički pristup
Primjer - bacanje novčića

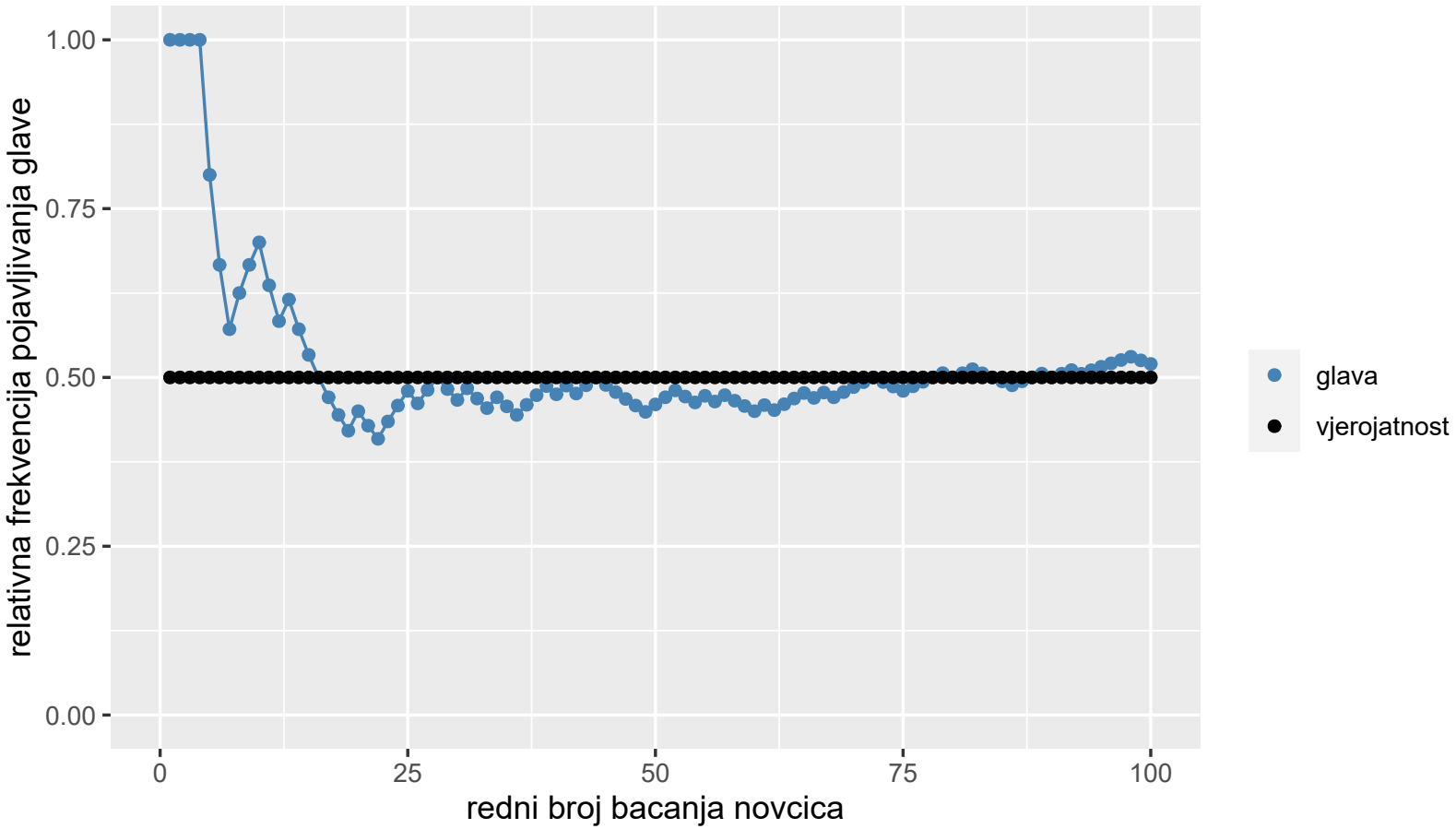

Figure 1: Relativne frekvencije pojavljivanja glave u 100 bacanja novčića.

Vjerojatnost - statistički pristup
Primjer - bacanje kockice

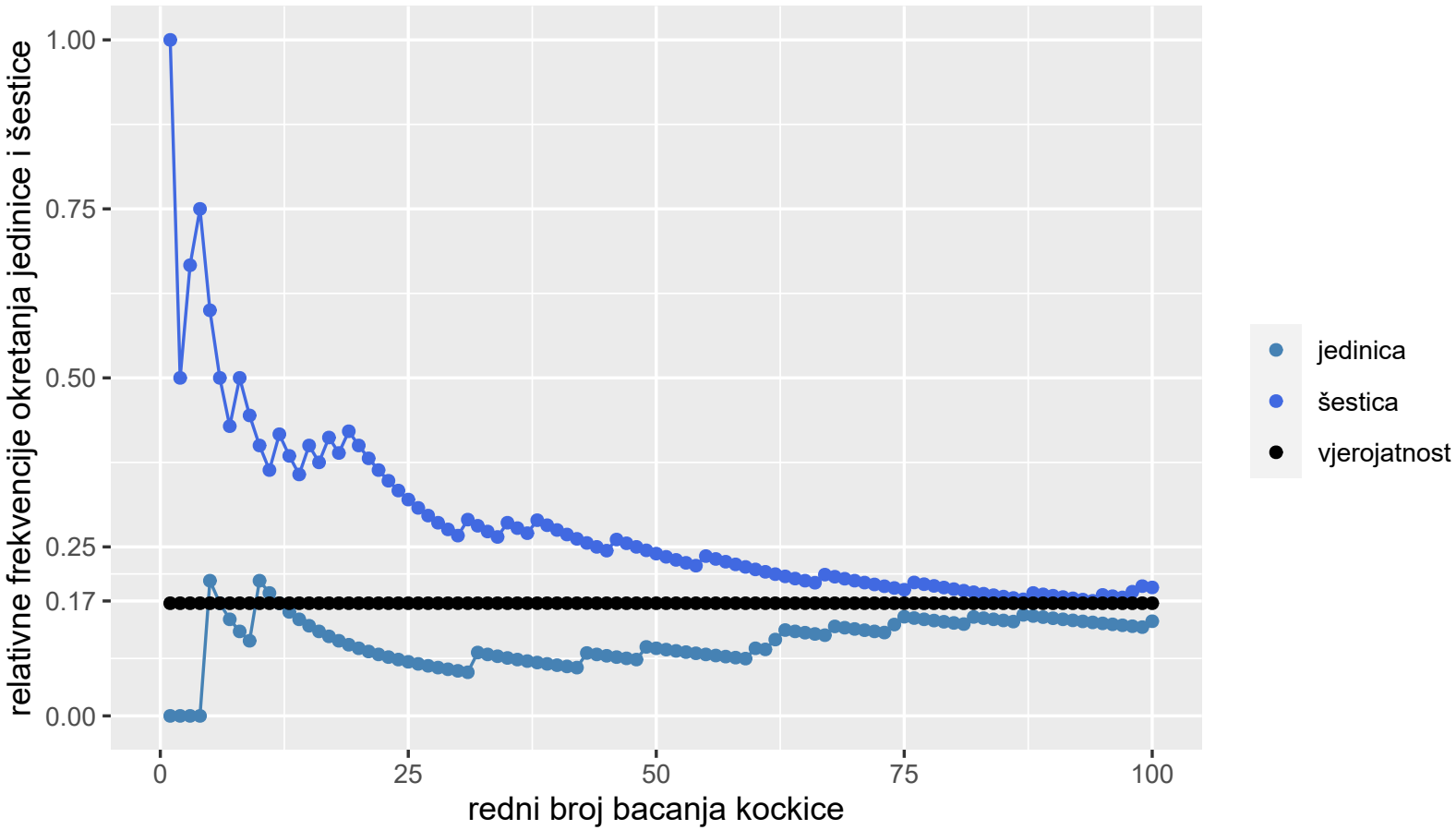

Figure 2: Relativne frekvencije pojavljivanja jedinice i šestice u 100 bacanja kockice.

Vjerojatnost - klasični pristup

Neka vrijede sljedeći uvjeti:

(1) skup svih mogućih ishoda slučajnog pokusa $\Omega \neq \emptyset$ ima konačno mnogo elemenata ($\Omega = \{\omega_1, \ldots, \omega_n\}, n \in \mathbb{N}$)

(2) svi jednočlani podskupovi od Ω su jednako vjerojatni, tj.

$$P(\{\omega_i\}) = P(\{\omega_j\}), \qquad \text{za sve } i, j \in \{1, \ldots, n\}$$

Tada vjerojatnost skupa (događaja) $A \subseteq \Omega$ definiramo na sljedeći način:

$$P(A) = \frac{\text{broj elemenata od } A}{\text{broj elemenata od } \Omega} = \frac{k(A)}{k(\Omega)},$$

gdje je $k(\cdot)$ oznaka za broj elemenata skupa

- primjeri 4.10, 4.11, 4.12
- Primjer. Pravilan novčić bacamo tri puta. Kolika je vjerojatnost da je pismo palo točno dva puta?

Vjerojatnost - motivacija

- neka je Ω neprazan skup – **skup elementarnih događaja** (bilo kakvih elemenata)
- elemente od Ω obično označavamo s ω – **elementarni događaj**
- primjer 4.8
- $\mathcal{F}$ – skup **događaja** (podskupova od Ω)
- **vjerojatnost** – mjera koja modelira stupanj vjerovanja da će se realizirati neki događaj vezan uz promatrani Ω

Vjerojatnost - definicija

Definition

Vjerojatnost (oznaka P) je funkcija koja svakom događaju $A \in \mathcal{F}$ pridružuje realan broj iz intervala $[0, 1]$ $(0 \leq P(A) \leq 1)$ tako da vrijede sljedeći zahtjevi:

V1. $P(\Omega) = 1$

V2. ako su A_1 i A_2 događaji iz $\mathcal{F}$ koji nemaju zajedničkih elemenata, tj. $A_1, A_2 \in \mathcal{F}$ i $A_1 \cap A_2 = \emptyset$, tada vrijedi

$$P(A_1 \cup A_2) = P(A_1) + P(A_2),$$

tj. vjerojatnost unije događaja A_1 i A_2 jednaka je zbroju vjerojatnosti $P(A_1)$ i $P(A_2)$ *

Svojstva vjerojatnosti

S1. **vjerojatnost suprotnog događaja** - ako je $A \in \mathcal{F}$, tada je $P(A^c) = 1 - P(A)$, $A^c = \Omega \setminus A$ komplement skupa A

S2. **vjerojatnost nemogućeg događaja** - $P(\emptyset) = 0$

S3. **monotonost vjerojatnosti** - ako su A i B skupovi iz $\mathcal{F}$ takvi da je $A \subseteq B$, tada je $P(A) \leq P(B)$; vrijedi i da je $P(B \setminus A) = P(B) - P(A)$

S4. **vjerojatnost unije** - ako su $A, B \in \mathcal{F}$ proizvoljni događaji (koji ne moraju biti disjunktni), tada je $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- primjer 4.14

Slučajna varijabla - definicija

- Neka je Ω skup elementarnih događaja, $\mathcal{F}$ skup događaja i P vjerojatnost na $\mathcal{F}$

Definition

Slučajna varijabla X je funkcija $X : \Omega \to \mathbb{R}$ takva da je

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}, \text{ za svaki } x \in \mathbb{R}.$$

- slučajna varijabla prevodi ishode pokusa od apstraktnog skupa Ω u skup realnih brojeva
- gornji zahtjev osigurava da možemo računati vjerojatnost skupova oblika $\{X \leq x\}$
- slučajne varijable označavat ćemo velikim slovima, npr. X, Y, Z itd.

Slučajna varijabla - definicija

- slučajna varijabla – moguće realizacije su realni brojevi, ali vrijednost koja će se realizirati u pojedinom eksperimentu nije jednoznačno određena uvjetima koje možemo sagledati prilikom istraživanja
- **skup svih mogućih realizacija** slučajne varijable ili **slika** slučajne varijable – oznaka $\mathcal{R}(X)$
- za skupove $C \subseteq \mathcal{R}(X)$ takve da je $\{X \in C\} \in \mathcal{F}$ možemo računati

$$P(X \in C)$$

- ako zadamo $\mathcal{R}(X)$ i vjerojatnosti oblika $P(X \in C)$ kažemo da smo zadali **razdiobu (distribuciju) slučajne varijable** X
- primjeri 4.9, 4.3 - 4.6, 4.11

Slučajne varijable i baza podataka

- Za slučajne varijable X i Y kažemo da su **jednako distribuirane** ako je $\mathcal{R}(X) = \mathcal{R}(Y)$ i

$$P(X \in C) = P(Y \in C)$$

  za sve $C \subseteq \mathcal{R}(X)$ takve da je $\{X \in C\} \in \mathcal{F}$
- varijable baze podataka mogu primiti mnogo različitih vrijednosti, a mi u trenutku njihovog proučavanja (mjerenja) ne možemo sa sigurnošću sagledati uvjete pod kojima će primiti neku od tih vrijednosti

Slučajne varijable i baza podataka

- jedna varijabla baze podataka sadrži podatke $x_1, \ldots, x_n$
- **svaki podatak iz varijable baze podataka je realizacija jedne slučajne varijable**

varijabla baze podataka	slučajne varijable
x_1	$\leftarrow X_1$
x_2	$\leftarrow X_2$
$\vdots$	$\vdots$
x_n	$\leftarrow X_n$

- niz slučajnih varijabli $X_1, \ldots X_n$ nazivamo **slučajan uzorak**

Slučajne varijable i baza podataka

- podaci varijable baze podataka $x_1, \ldots, x_n$ dolaze iz iste populacije
- zato pretpostavljamo da su $X_1, \ldots X_n$ sve **jednako distribuirane** i to kao neka slučajna varijabla X
- u tom smislu **varijabla baze podataka dolazi iz slučajne varijable** X
- na osnovu podataka želimo zaključivati o distribuciji slučajne varijable X
- primjer 4.1, zadatak 4.1

Slučajna varijabla - Tipovi

- prije smo uočili bitnu razliku u opisu numeričkih varijabli koje su diskretnog tipa od onih koje su neprekidnog tipa
- iste razlike vidljive su u načinu zadavanja slučajnih varijabli kojima modeliramo varijable u istraživanju
- razlikovat ćemo dva tipa slučajnih varijabli: **diskretne slučajne varijable** i **neprekidne slučajne varijable**

Diskretna slučajna varijabla - Definicija

- **Ukoliko je $\mathcal{R}(X)$ konačan ili prebrojiv skup kažemo da je slučajna varijabla X diskretna.**
- X može imati konačan skup svih mogućih realizacija - $\mathcal{R}(X) = \{x_1, x_2, \ldots, x_n\}$ ili prebrojiv skup svih mogućih realizacija - $\mathcal{R}(X) = \{x_1, x_2, x_3, \ldots\}$
- za svaku pojedinu realizaciju x_i definiramo realan broj

$$p_i = P(X = x_i)$$

Diskretna slučajna varijabla - Zadavanje

Distribucija diskretne slučajne varijable X je u potpunosti zadana skupom $\mathcal{R}(X)$ i pripadnim nizom $(p_i, i = 1, \ldots, n)$ (odnosno nizom $(p_i, i \in \mathbb{N})$ ako je $\mathcal{R}(X)$ prebrojiv skup) za kojega vrijede sljedeća svojstva:

1. $p_i \geq 0$ za sve pripadne $x_i \in \mathcal{R}(X)$
2. $\sum_{x_i \in \mathcal{R}(X)} p_i = 1$

- računanje vjerojatnosti da diskretna slučajna varijabla X primi vrijednosti iz nekog skupa $A \subseteq \mathcal{R}(X)$:

$$P(X \in A) = \sum_{x_i \in A} p_i$$

Diskretna slučajna varijabla - Tablica distribucije

Vjerojatnost poprimanja vrijednosti za diskretnu slučajnu varijablu se često prikazuje pomoću ta dva bitna niza u obliku **tablice distribucije** koju često zovemo samo **distribucija** diskretne slučajne varijable:

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

ili

$$X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix},$$

ako je $\mathcal{R}(X)$ beskonačan

Diskretna slučajna varijabla - Grafički prikaz distribucije

- distribuciju diskretne slučajne varijable X možemo slikovito prikazati **stupčastim dijagramom** - svaki stupić odgovara jednoj vrijednosti x_i koju ta slučajna varijabla može poprimiti (tj. jednom elementu iz $\mathcal{R}(X)$), a visina stupića je jednaka vjerojatnosti $p_i = P(X = x_i)$
- primjer 4.15, 4.16

$\leftrightsquigarrow$

Diskretna slučajna varijabla - primjeri

Bernoullijeva slučajna varijabla

- ako varijabla može poprimiti samo dvije vrijednosti (0 - "neuspjeh" ili 1 - "uspjeh") možemo je modelirati korištenjem Bernoullijeve slučajne varijable
- **Bernoullijeva slučajna varijabla** s parametrom $p \in (0,1)$ – svaka slučajna varijabla s tablicom distribucije

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

- parametar p – vjerojatnost da slučajna varijabla X poprimi vrijednost 1 ("vjerojatnost uspjeha")
- primjer 4.23, 4.24

Diskretna slučajna varijabla - primjeri

Binomna slučajna varijabla

- vezana uz n nezavisnih ponavljanja pokusa koji ima samo dva moguća ishoda – uspjeh (1) i neuspjeh (0)
- pri tome se u svakom izvođenju pokusa uspjeh realizira s vjerojatnošću $p \in (0, 1)$ (svako ponavljanje je jedna Bernoullijeva slučajna varijabla)
- **binomna slučajna varijabla** s parametrima $n \in \mathbb{N}$ i $p \in (0, 1)$ broji uspjehe u tih n nezavisnih ponavljanja pokusa
- oznaka $X \sim \mathcal{B}(n, p)$

Diskretna slučajna varijabla - primjeri

- njena distribucija zadana je tablicom

$$X \sim \begin{pmatrix} 0 & 1 & 2 & \cdots & n \\ (1-p)^n & \binom{n}{1}p(1-p)^{n-1} & \binom{n}{2}p^2(1-p)^{n-2} & \cdots & p^n \end{pmatrix},$$

gdje je

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{1\cdot 2\cdots(k-1)k}$$

- za $k \in \{0, 1, \ldots, n\}$

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

- primjer 4.25

Neprekidna slučajna varijabla - motivacija

- model za neprekidne numeričke varijable baze podataka – **neprekidne slučajne varijable**
- skup svih mogućih realizacija neprekidne slučajne varijable $\mathcal{R}(X)$ je interval realnih brojeva ili $\mathcal{R}(X) = \mathbb{R}$

Neprekidna slučajna varijabla - definicija

Definition

Za slučajnu varijablu X kažemo da je **neprekidna slučajna varijabla** ako postoji nenegativna realna funkcija f, definirana na skupu realnih brojeva, takva da je

$$P(a < X \leq b) = \int_a^b f(x)dx$$

- takvu funkciju f zovemo **funkcija gustoće** neprekidne slučajne varijable X
- ako je poznata funkcija gustoće tada poznajemo **razdiobu** ili **distribuciju** neprekidne slučajne varijable
- $P(a < X \leq b)$ je jednaka površini između osi x i grafa funkcije f nad intervalom $[a, b]$ (slika 3)

Neprekidna slučajna varijabla - Računanje vjerojatnosti

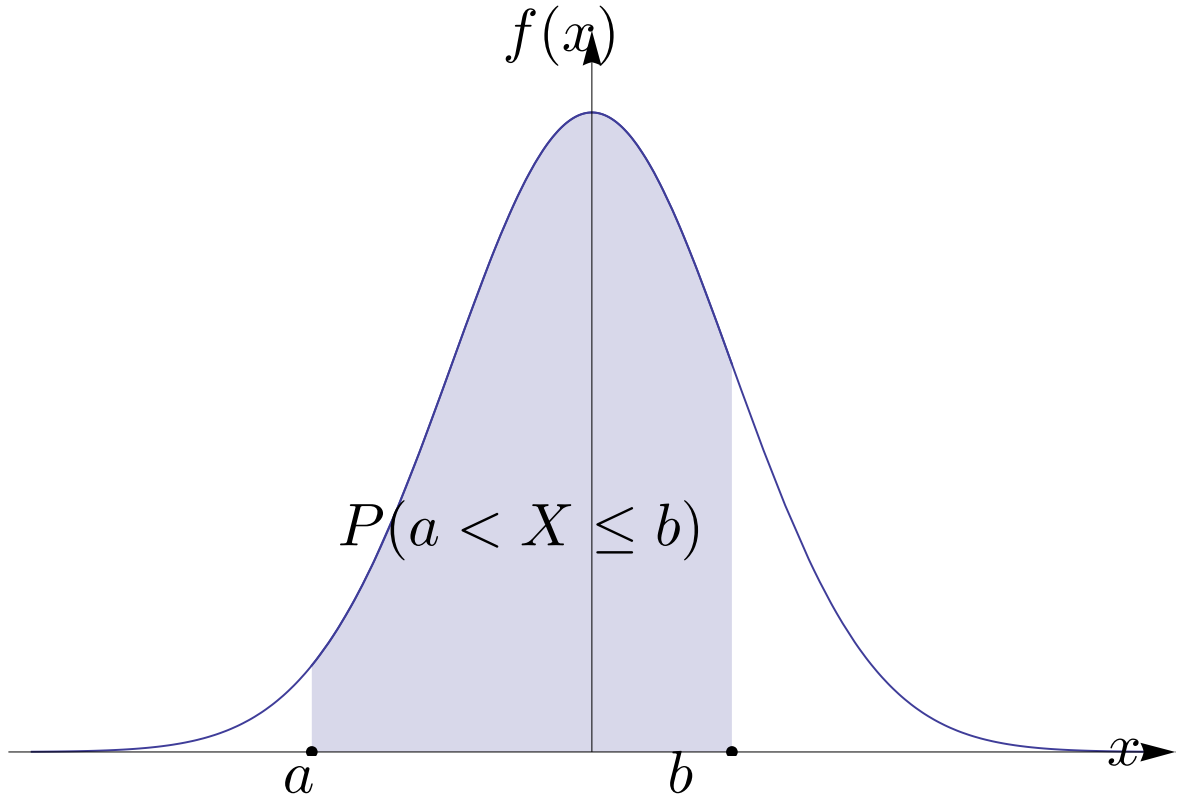

Figure 3: Vjerojatnost kao površina između osi x i grafa funkcije f nad intervalom $[a, b]$

- primjer 4.17

Neprekidna slučajna varijabla - Računanje vjerojatnosti

- za neprekidnu slučajnu varijablu X

$$P(X = x) = 0$$

za svaki $x \in \mathbb{R}$

- zato za neprekidnu slučajnu varijablu

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b)$$

odnosno

$$P\left(X \in (a, b]\right) = P\left(X \in [a, b]\right) = P\left(X \in (a, b)\right) = P\left(X \in [a, b)\right)$$

Normalna slučajna varijabla

- **normalna slučajna varijabla** - najvažnija neprekidna slučajna varijabla
- normalna slučajna varijabla (oznaka $X \sim \mathcal{N}(\mu, \sigma^2)$) je neprekidna slučajna varijabla za koju je

$$\mathcal{R}(X) = \mathbb{R},$$

a funkcija gustoće definirana je izrazom

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

gdje je μ bilo koji realan broj, a $\sigma > 0$

- za $\mu = 0$, $\sigma^2 = 1$ - normalnu slučajnu varijablu $X \sim \mathcal{N}(0, 1)$ nazivamo **standardna normalna slučajna varijabla**
- graf funkcije gustoće normalne slučajne varijable ovisi o izboru parametara μ i σ^2 (slika 4)

Normalna slučajna varijabla

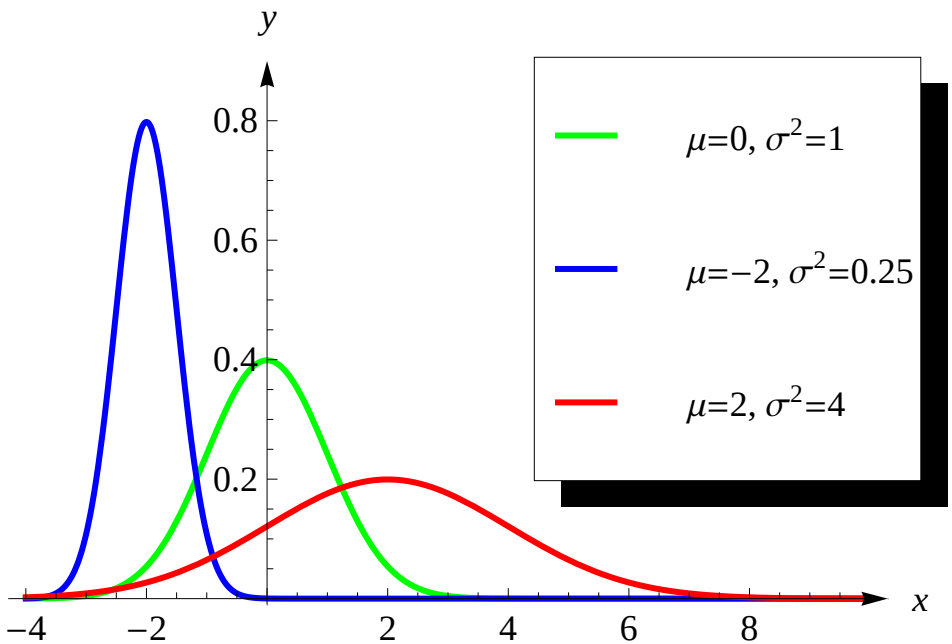

Figure 4: Graf funkcije gustoće normalne slučajne varijable za različite μ and σ^2.

- postupak standardizacije
- primjer 4.27

Očekivanje i varijanca slučajne varijable

diskretna slučajna varijabla

- **očekivanje** diskretne slučajne varijable X je realan broj (ako postoji)

$$\mu = EX = \sum_{x_i \in \mathcal{R}(X)} x_i p_i,$$

a njena **varijanca** realan broj (ako postoji)

$$\sigma^2 = Var\, X = \sum_{x_i \in \mathcal{R}(X)} (x_i - \mu)^2 p_i$$

- primjer 4.18
- primjer Bernoullijeva, binomna

Očekivanje i varijanca slučajne varijable

neprekidna slučajna varijabla s funkcijom gustoće f

- **očekivanje** neprekidne slučajne varijable X je realan broj (ako postoji)

$$\mu = EX = \int_{-\infty}^{\infty} x f(x) dx,$$

a njena **varijanca** realan broj (ako postoji)

$$\sigma^2 = Var\, X = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

- primjer normalna

Očekivanje i varijanca slučajne varijable

- drugi korijen iz varijance zovemo **standardna devijacija** slučajne varijable (oznaka σ)
- očekivanje - jedna od mjera centralne tendencije
- varijanca i standardna devijacija - mjere raspršenja vrijednosti slučajne varijable oko njenog očekivanja

$\leftrightsquigarrow$

Medijan slučajne varijable

- medijan slučajne varijable X je realan broj m za koji vrijedi da je

$$P(X \geq m) \geq \frac{1}{2} \quad \text{i} \quad P(X \leq m) \geq \frac{1}{2}$$

- medijan je jedna mjera centralne tendencije ali ne mora nužno biti jedinstven
- primjer 4.20, 4.21

Kvantili neprekidne slučajne varijable

- za $p \in (0,1)$, p-kvantil neprekidne slučajne varijable X je realan broj x_p za koji vrijedi da je

$$P(X \leq x_p) = p$$

- $x_{0.5}$ je medijan, $x_{0.25}$ donji kvartil i $x_{0.75}$ gornji kvartil slučajne varijable
- primjer kvantili normalne slučajne varijable

Empirijska distribucija - Diskretna slučajna varijabla

- u statističkom istraživanju bilježimo realizacije jedne diskretne numeričke varijable baze podataka u M promatranja
- podaci predstavljaju uzorak koji dolazi iz diskretne slučajne varijable X koja može primiti konačno mnogo vrijednosti $x_1, \ldots, x_n$
- distribucija od X je zadana tablicom distribucije

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix}$$

- niz vjerojatnosti p_i, $i = 1, \ldots, n$, ne znamo i želimo ga odrediti na temelju prikupljenih podataka

Empirijska distribucija - Diskretna slučajna varijabla

- **empirijska distribucija** - definira p_i kao relativnu frekvenciju pojavljivanja x_i u M ponavljanja mjerenja
- ako s f_i označimo frekvenciju pojavljivanja realizacije x_i u M ponavljanja mjerenja, onda je empirijska distribucija (slučajna varijabla) zadana tablicom

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ \frac{f_1}{M} & \frac{f_2}{M} & \dots & \frac{f_n}{M} \end{pmatrix}, \qquad f_1 + f_2 + \cdots f_n = M$$

- empirijska distribucija je **procjena** distribucije slučajne varijable X
- procjena će biti bolja što je broj promatranja (M) veći
- primjer

Empirijska distribucija - Općeniti slučaj

- općenito, neka je $v_1, \ldots, v_M$ uzorak iz neke slučajne varijable X koja može imati beskonačan skup vrijednosti ili biti neprekidna
- uočimo da je broj prikupljenih podataka mjerenjem vrijednosti slučajne varijable uvijek konačan
- među izmjerenim podacima može biti i jednakih - u nizu $v_1, \ldots, v_M$ pojavljuju se različite vrijednosti $x_1, \ldots, x_n$ s odgovarajućim frekvencijma $f_1, \ldots f_n$

Empirijska distribucija - Općeniti slučaj

- na temelju dobivenih podataka možemo definirati empirijsku distribuciju tablicom distribucije

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ \frac{f_1}{M} & \frac{f_2}{M} & \cdots & \frac{f_n}{M} \end{pmatrix}, \qquad f_1 + f_2 + \cdots f_n = M$$

- za neprekidne slučajne varijabla procjena vjerojatnosti $P(X = x_i)$ korištenjem empirijske distribucije nema smisla jer je $P(X = x_i) = 0$
- empirijsku distribuciju možemo koristiti za procjenu vjerojatnosti realiziranja X u nekom skupu ukoliko je M velik broj:

$$P(X \in [a, b]) \approx \begin{array}{l}\text{relativna frekvencija pojavljivanja} \\ \text{realizacija iz intervala } [a, b]\end{array}$$

Empirijska distribucija - Očekivanje i varijanca

- ako je S slučajna varijabla definirana empirijskom tablicom distribucije

$$S = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ \frac{f_1}{M} & \frac{f_2}{M} & \cdots & \frac{f_n}{M} \end{pmatrix}, \qquad f_1 + f_2 + \cdots f_n = M,$$

onda je

$$ES = \frac{1}{M} \sum_{i=1}^{M} v_i = \bar{v}_n$$

$$Var\, S = \frac{1}{M} \sum_{i=1}^{M} (v_i - \bar{v}_n)^2 = \bar{s}_n^2 = \frac{n}{n-1} s_n^2$$

- očekivanje empirijske distribucije odgovara aritmetičkoj sredini podataka
- varijanca empirijske distribucije odgovara varijanci podataka

Empirijska distribucija slučajne varijable - Pojam procjene

- empirijska distribucija dobivena na osnovu uzorka iz slučajne varijable X - **procjena** za njenu stvarnu distribuciju
- aritmetička sredina, varijanca i standardna devijacija tih podataka - **procjene** za očekivanje, varijancu i standardnu devijaciju slučajne varijable

- primjer 4.29