Teorija verovatnoće

Teorija verovatnoće

- Jedan od ključnih koncepata kod prepoznavanja šablona je nesigurnost (*uncertanty*). Izvori nesigurnosti:
  - šum u podacima (greške pri merenju, neuočene varijable,...)
  - konačna veličina uzorka (ograničen trening skup)

- Teorija verovatnoće (*probability theory*) nam omogućava da kvantifikujemo nesigurnost i da manipulišemo sa njom

- Kombinovanjem teorije verovatnoće i teorije odlučivanja (*decicion theory*) možemo davati optimalna predviđanja na osnovu dostupnih informacija, čak i ako su te informacije nepotpune ili dvosmislene

Ishodi i događaji

- Eksperimenti imaju ishode
  - Npr. ishodi bacanja kockice su brojevi od 1 do 6
- Događaji su skupovi ishoda
  - Npr. događaj može biti da je dobijen broj veći od 3, što odgovara skupu ishoda {4, 5, 6}
- Kažemo da se neki događaj desio ako se desio neki ishod iz tog događaja

- *A* je slučajna promenljiva ako označava događaj, takav da postoji određeni stepen nesigurnosti da li se događaj desio
  - Npr. *A* = prilikom bacanja kockice dobijen je broj veći od 3

Verovatnoća

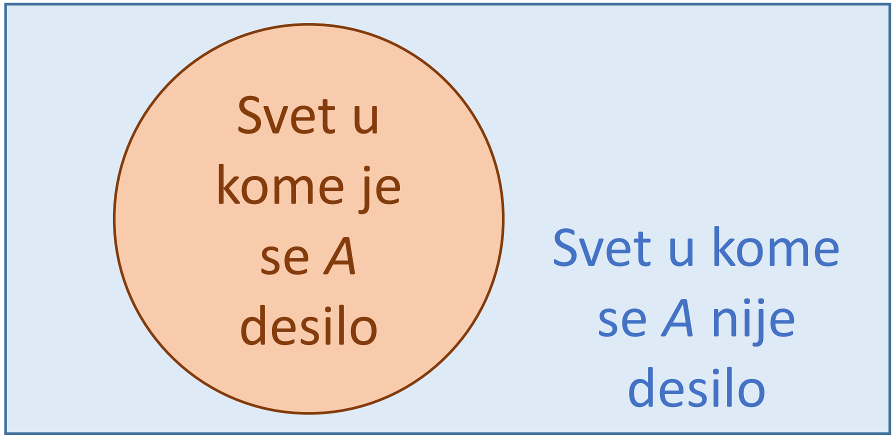
- Verovatnoća $P(A)$ predstavlja dugoročnu frekvencije događaja *A*

$$P = n/N$$

n – broj eksperimenata u kojima smo registrovali da se događaj *A* desio

N – ukupan broj izvedenih eksperimenata $N \to \infty$

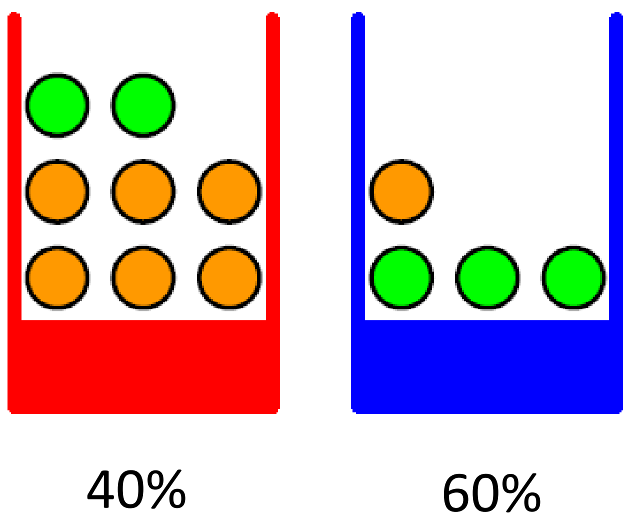
Prostor svih mogućih svetova



$P(A)$ - površina kruga

Verovatnoća događaja – primer

- Eksperiment:
  1. Na slučajan način biramo jednu od kutija
     - Odabranu kutiju označićemo *slučajnom promenljivom K.* Mogući ishodi (vrednosti *K*) su *c* (crvena) i *p* (plava)
  2. Iz odabrane kutije na slučajan način biramo jednu od voćki
     - Odabranu voćku označićemo *slučajnom promenljivom V.* Mogući ishodi (vrednosti *V*) su *j* (jabuka) i *n* (narandža)

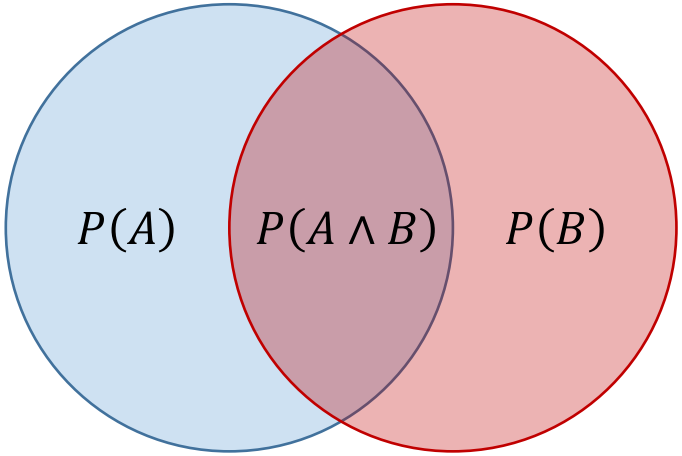


- Izveli smo mnogo eksperimenata ($N \to \infty$) i ispostavilo se da u 40% slučajeva biramo crvenu kutiju, a u 60% slučajeva plavu:

$$P(K = c) = 0.4$$
$$P(K = p) = 0.6$$

Osobine verovatnoće

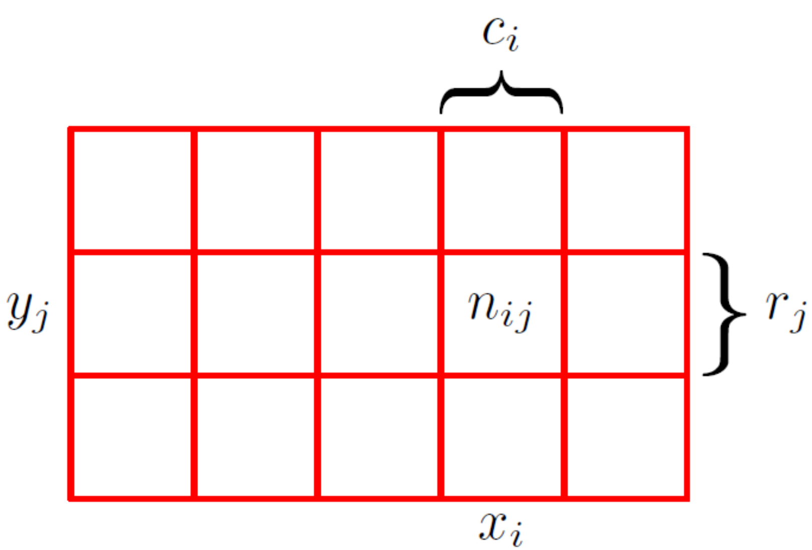
- Prema definiciji ($P = n/N$), verovatnoće leže u intervalu [0, 1]
- Verovatnosna mera *P* mora da zadovolji sledeće aksiome:
  - $P(\Omega) = 1$, gde je Ω skup svih ishoda
  - $P(A) \geq 0$, za svaki događaj $A \subseteq \Omega$
  - $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ ako su A_i disjunktni događaji
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Uslovna verovatnoća

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

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



N – ukupan broj pokušaja

n_{ij} – broj pokušaja u kojima je

$$X = x_i \wedge Y = y_j$$

c_i – broj pokušaja za koje važi $X = x_i$

r_j – broj pokušaja za koje važi $Y = y_j$

Združena verovatnoća (*joint probability*):

$$P(X = x_i, Y = y_j) = n_{ij}/N$$

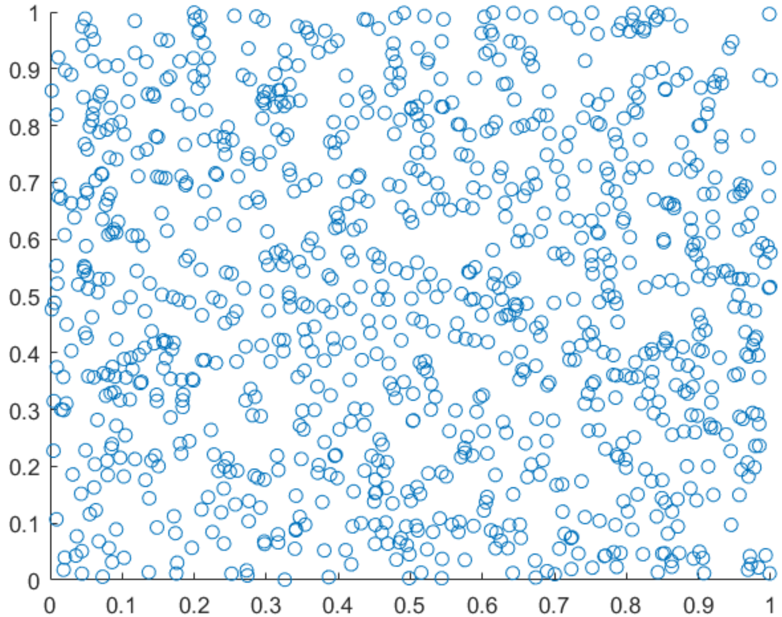
Uslovna verovatnoća (*conditional probability*):

$$\mathrm{P}(Y = y_j|X = x_i) = n_{ij}/c_i\,, \mathrm{P}(X = x_i|Y = y_j) = n_{ij}/r_j$$

Nezavisnost događaja

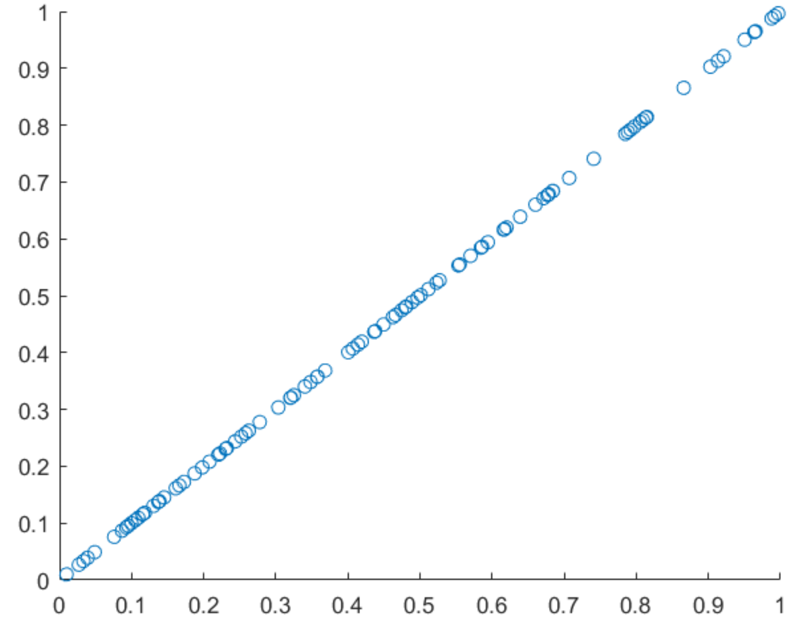
- Događaji *A* i *B* su nezavisni ako važi

$$P(A|B) = P(A), \text{ odnosno, } P(A \cap B) = P(A)P(B)$$



Nezavisni događaji

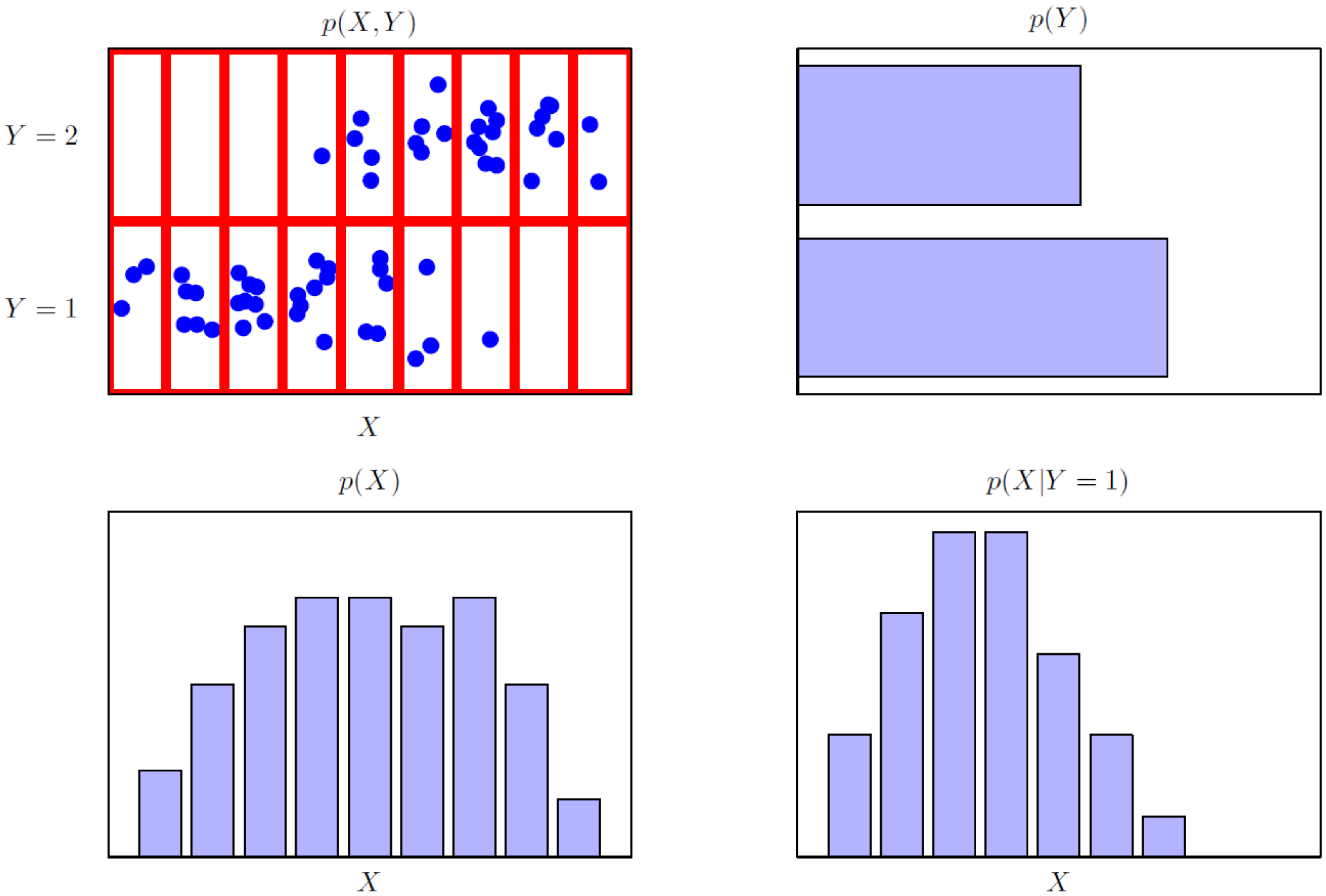
$$P(x > 0.5) = 0.5$$
$$P(y > 0.5) = 0.5$$
$$P(x > 0.5 \land y > 0.5) = 0.25$$



Zavisni događaji

$$P(x > 0.5) = 0.5$$
$$P(y > 0.5) = 0.5$$
$$P(x > 0.5 \land y > 0.5) = 0.5$$

Histogram



Pravilo sume i proizvoda

- Pravilo zbira (*sum rule*):

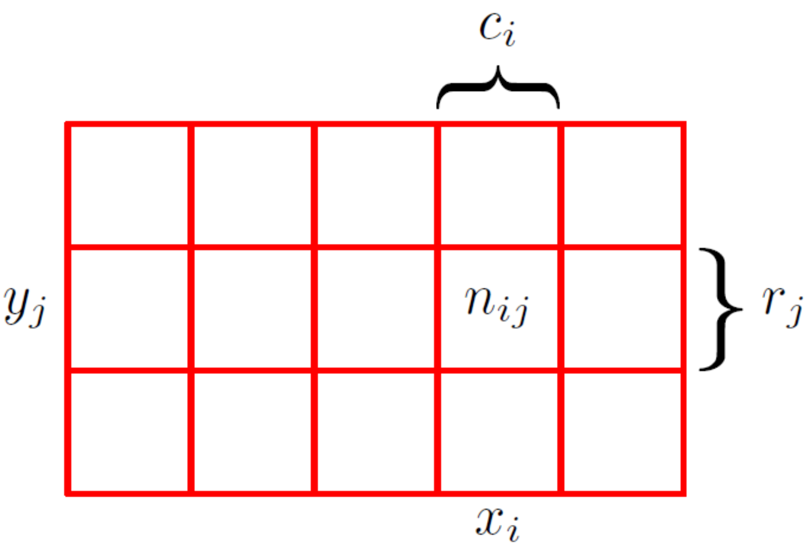
$$P(X) = \sum_{Y} P(X, Y)$$

- Pravilo proizvoda (*product rule*):

$$P(X, Y) = P(Y|X)P(X)$$

- Verovatnoće *P*(*X*) i *P*(*Y*) se nazivaju marginalne verovatnoće (*marginal probability*)

Podsetnik: teorija verovatnoće



N – ukupan broj pokušaja

n_{ij} – broj pokušaja u kojima je

$$X = x_i \wedge Y = y_j$$

c_i – broj pokušaja za koje važi $X = x_i$

r_j – broj pokušaja za koje važi $Y = y_j$

- Pravilo zbira:

$$P(X = x_i) = c_i/N, c_i = \sum_j n_{ij} \rightarrow P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$$

- Pravilo proizvoda :

$$P\left(X = x_i, Y = y_j\right) = n_{ij}/N = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P\left(Y = y_j \middle| X = x_i\right) \cdot P(X = x_i)$$

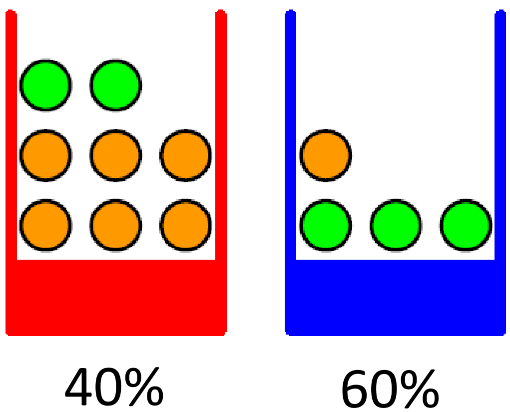
Bajesova teorema

- Iz pravila proizvoda sledi Bajesova teorema:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X) = \sum_Y P(X|Y)P(Y)$ - imenilac u Bajesovoj teoremi možemo posmatrati kao normalizacionu konstantu neophodnu da se uslovna verovatnoća $P(Y|X)$ za sve moguće vrednosti Y sabira na 1

Primer



$$P(K = c) = 0.4, \qquad P(K = p) = 0.6$$

$$P(V = j|K = c) = \frac{1}{4}, \qquad P(V = j|K = p) = \frac{3}{4}$$

$$P(V = n|K = c) = \frac{3}{4}, \qquad P(V = n|K = p) = \frac{1}{4}$$

$$P(V = j) = P(V = j|K = c)P(K = c) + P(V = j|K = p)P(K = p) = \frac{11}{20}$$

$$P(V = n) = 1 - P(V = j) = \frac{9}{20}$$

- Ako znamo da smo selektovali narandžu, koja je verovatnoća da je selektovana kutija crvena/plava?
  - $P(K = c|V = n) = \frac{P(V=n|K=c)P(K=c)}{P(V=n)} = \frac{2}{3}$
  - $P(K = p|V = n) = 1 - P(K = c|V = n) = \frac{1}{3}$

Primer

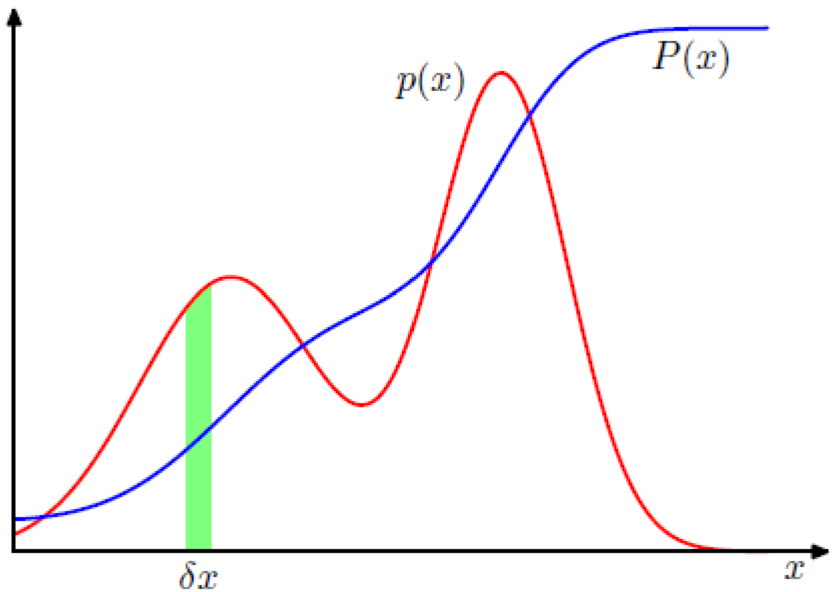
- Ako smo upitani koja je kutija odabrana *pre* nego što uočimo da je selektovana voćka narandža, najkompletnija informacija sa kojom raspolažemo je verovatnoća $P(B)$
  - $P(K)$ se naziva *apriorna verovatnoća* – verovatnoća dostupna *pre* opservacije vrste voća
  - $P(K = c) = 0.4 \rightarrow$ verovatnije je da smo izabrali plavu kutiju

- Ako smo upitani koja je kutija odabrana *nakon* što smo uočili vrstu voća, najkompletnija informacija sa kojom raspolažemo je $p(K|V)$
  - $p(K|V)$ se naziva *aposteriorna verovatnoća* – verovatnoća koju smo dobili nakon što smo uočili *V*
  - $p(K = c|V = n) = \frac{2}{3} \rightarrow$ (nakon što smo uočili da smo izvukli narandžu) verovatnije je da smo izabrali crvenu kutiju

Kontinualne varijable – gustina raspodele

- Do sada smo razmatrali diskretne događaje. Sada ćemo pričati o verovatnoćama *kontinualnih* varijabli
- Ako je verovatnoća da kontinualna varijabla *x* leži u intervalu $(x, x + \delta x)$ data sa $p(x)\delta x$ za $\delta x \to 0$, onda se $p(x)$ naziva *gustina* raspodele (*probability density*) promenljive *x*
- Verovatnoća da će (kontinualna) varijabla *x* ležati u intervalu (*a*, *b*) definisana je sa:

$$P\big(x \in (a, b)\big) = \int_a^b p(x)dx$$



$P(z) = \int_{-\infty}^{z} p(x)dx$ - kumulativna *funkcija* raspodele

Kontinualne varijable – gustina raspodele

- Gustina raspodele *p*(*x*) mora da zadovoljava dva uslova:

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

- Pravila zbira i proizvoda, kao i Bajesova teorema važe i u slučaju kontinualnih varijabli, kao i kombinacije kontinualnih i diskretnih varijabli:

$$\int p(x) = \int p(x,y)dy$$

$$p(x,y) = p(y|x)p(x)$$

Matematičko očekivanje

- Intuitivno, predstavlja srednju vrednost neke slučajne promenljive
- Ako je *X* *diskretna* slučana promenljiva sa raspodelom verovatnoće

$$\begin{pmatrix} x_1 & x_2 & \dots & x_N \\ p(x_1) & p(x_2) & \dots & p(x_N) \end{pmatrix}, p(x_1) + \cdots + p(x_N) = 1$$

  onda je *matematičko očekivanje* slučajne promenljive *X*:

$$E[X] = \sum_{n=1}^{N} x_n p(x_n)$$

- Ako je *X* *kontinualna* slučajna promenljiva sa gustinom verovatnoće *p*(*x*), onda je *matematičko očekivanje* slučajne promenljive *X*:

$$E[X] = \int_{-\infty}^{+\infty} x p(x) dx$$

Matematičko očekivanje - primer

- Recimo da imamo rulet sa 38 brojeva: nula (0), dupla nula (00) i brojevi 1,2,...,36
  - Verovatnoća da se loptica zaustavi na bilo kom broju je jednaka
  - Ako se loptica zaustavi na 0 plaćamo mušteriji \$5
  - Ako se loptica zaustavi na 00 plaćamo mušteriji \$10
  - Ako se loptica zaustavi na neparnom broju plaćamo mušteriji \$1
  - Ako se loptica zaustavi na parnom broju plaćamo mušteriji \$2

Koliko bismo trebali naplatiti pojedinačnu igru da bismo zaradili?

- Neka je *X* broj na kome se loptica zaustavi, a *u*(*X*) količina novca koji treba da isplatimo kada se loptica zaustavi na *X:*

$$E[u(x)] = 5\left(\frac{1}{38}\right) + 10\left(\frac{1}{38}\right) + 1\left(\frac{18}{38}\right) + 2\left(\frac{18}{38}\right) = 1.82$$

Na duže staze, u proseku moramo isplatiti \$1.82 po igri. Dakle, da ne bismo gubili novac, igru moramo naplatiti barem \$1.82.

Matematičko očekivanje uzorka

- Ako je dato N (konačno mnogo) tačaka iz neke distribucije X, možemo proceniti očekivanje:

$$E[X] \cong \frac{1}{N} \sum_{i=n}^{N} x_n$$

ova procena je jednaka tačnoj vrednosti očekivanja kada $N \rightarrow \infty$

Matematičko očekivanje – osobine

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

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

- Važi

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$$

Varijansa

- Srednja vrednost (centar raspodele) nije dovoljna za opisivanje pojava
  - Npr. ako kažemo da je prosečna temperatura nekog mesta 15 stepeni, imamo utisak prijatne klime. Međutim, možda je leti 40, a zimi -10 stepeni
- Treba znati kakva su odstupanja mogućih vrednosti slučajne promenljive *X* od njene srednje vrednosti *E*[*X*] – varijansu (disperziju)
- Varijansa (disperzija) slučajne promenljive *X* se definiše kao očekivanje odstupanja *X* od $E[X]$:

$$var[X] = E[(X - E[X])^2] = E[X^2] - (E[x])^2$$

- Koren varijanse se naziva standardna devijacija

Varijansa uzorka

- Ako je dato N (konačno mnogo) tačaka iz neke distribucije X, možemo proceniti varijansu:

$$var[X] = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

ova procena je jednaka tačnoj vredosti varijanse za $N \to \infty$

Određivanje očekivanja i varijanse uzorka

- Zadatak: odrediti prosečno vreme spavanja studenata
  - Cela populacija je prevelika – ne možemo zabeležiti vreme spavanja svakog studenta
  - Zato ćemo uzeti slučajan uzorak od 10 studenata. Studenti su prijavili sledeća vremena spavanja: 7, 6, 8, 4, 2, 7, 6, 7, 6, 5
  - Odredićemo matematičko očekivanje i varijansu uzorka koji će nam služiti kao ocena matematičkog očekivanja i varijanse cele populacije

- Matematičko očekivanje uzorka:

$$E[sleep\_time] = \frac{7+6+8+4+2+7+6+7+6+5}{10} = 5.8$$

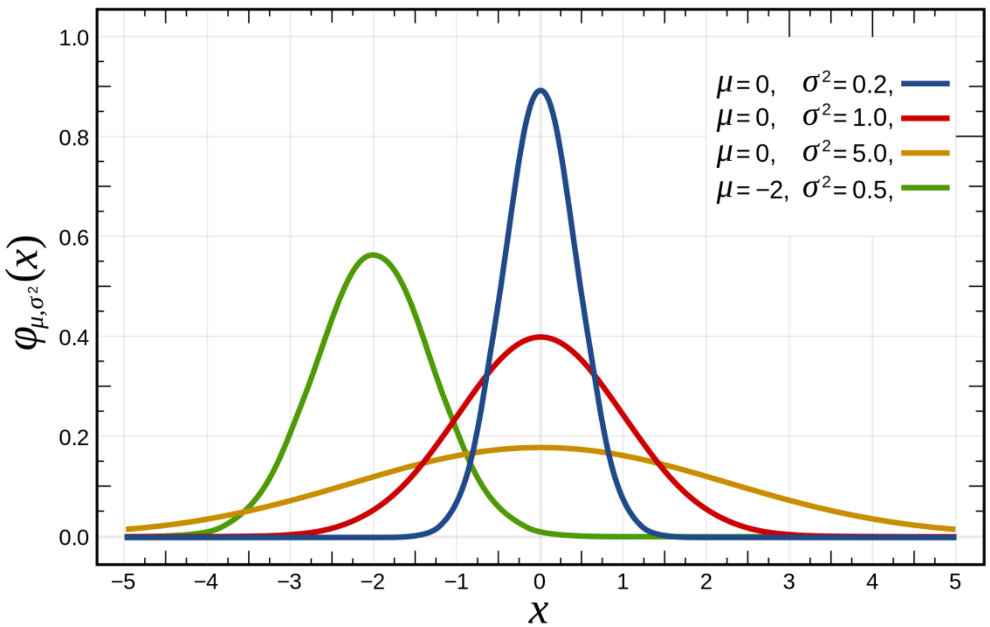
- Varijansa uzorka:

$$Var[sleep\_time] = \frac{(7-5.8)^2 + (6-5.8)^2 + \cdots + (5-5.8)^2}{10-1} = 3.067$$

- Standardna devijacija uzorka: $\sigma = \sqrt{3.067} = 1.75$

Normalna (Gausova) raspodela

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$



Zavisi od dva parametra:
μ – srednja vrednost (*mean*)
σ^2 – varijansa (*variance*)

$E[x] = \mu \qquad var[x] = \sigma^2$

σ: *standardna varijacija*
$\beta = 1/\sigma^2$: *preciznost*

Normalna (Gausova) raspodela zadovoljava uslove da bude validna gustina raspodele:

$$\mathcal{N}(x|\mu,\sigma^2) > 0 \text{ i } \int \mathcal{N}(x|\mu,\sigma^2) = 1$$

Normalna (Gausova) raspodela