Statistika i osnovna mjerenja

Dvodimenzionalne raspodjele i korelacija

M. Makek

2017/2018

Višedimenzionalne raspodjele

<u>Primjer</u>: u eksperimentu sudaramo 2 protona

- Prije sudara oni imaju energiju i impuls E_1, p_1; E_2, p_2
- Nakon sudara nastaje mnoštvo različitih čestica s energijama i impulsima E_k',p_k'
- Da bismo odredili energiju i impuls čestice u konačnom stanju mjerimo npr. njen impuls u magnetskom spektrometru i energiju u kalorimetru.
- U tom slučaju energija i impuls su kontinuirane slučajne varijable
- Pitanja: da li su nezavisne? Ako ne, kako su one povezane? Kako su povezane sa energijama i impulsima ostalih čestica?

- U eksperimentima se često opaža veći broj varijabli, koje mogu biti neovisne ili povezane, o čemu će ovisiti analiza rezultata i račun pogreške

Dvodimenzionalne raspodjele

- Uzmimo **dvije** diskretne slučajne varijable **X** i **Y** definirane na istom prostoru elementranih događaja
- <u>*Def*</u>: **združena raspodjela vjerojatnosti** p(x,y) za diskretne slučajne varijable X i Y je vjerojatnost da istodobno X poprimi vrijednost x i Y poprimi vrijednost y:

$$p(x,y) = P(X=x \text{ i } Y=y)$$

- <u>*Def*</u>: Rubne raspodjele vjerojatnosti za varijable X i Y označavamo s $p_X(x)$ i $p_Y(y)$, a dane su izrazima:

$$p_X(x) = \sum_{y \in D_Y} p(x,y) \qquad p_Y(y) = \sum_{x \in D_X} p(x,y)$$

→ vjerojatnost da pojedina varijabla iz združene raspodjele poprimi neku vrijednost bez obzira na drugu varijablu.

Dvodimenzionalne raspodjele

- Primjer: studenti prve godine razvrstani po ocjenama (varijabla X) i spolu (varijabla Y)

Ocjena	Muško	Žensko	UKUPNO
1	10	2	12
2	13	3	16
3	15	7	22
4	18	7	25
5	4	3	7
UKUPNO	60	22	**82**

- Na temelju raspodjele definiramo vjerojatnost a posteriori:
  - *Vjerojatnost* da student bude žensko i ima ocjenu 3 je 7/82, a vjerojatnost student bude muško i ima ocjenu 3 je 15/82.
  - *Rubna vjerojatnost* da bilo koji student ima ocjenu 3 je 22/82, a rubna vjerojatnost da student bude muško je 60/82

Dvodimenzionalne raspodjele

- Za kontinuirane slučajne varijable ***X,Y*** definiramo združenu funkciju gustoće vjerojatnosti *f(x,y)*, tako da za bilo koji dvodimenzionalni interval $A \subseteq \mathbb{R} \times \mathbb{R}$ vrijedi:

$$P((X,Y) \in A) = \iint_A f(x,y)dxdy$$

- Združena funkcija gustoće vjerojatnosti mora zadovoljavati uvjete: $f(x,y) \geq 0, \quad \forall x,y$ $\quad \iint_{\mathbb{R}\times\mathbb{R}} f(x,y)dxdy = 1$

- Rubne funkcije gustoće vjerojatnosti definiramo kao:

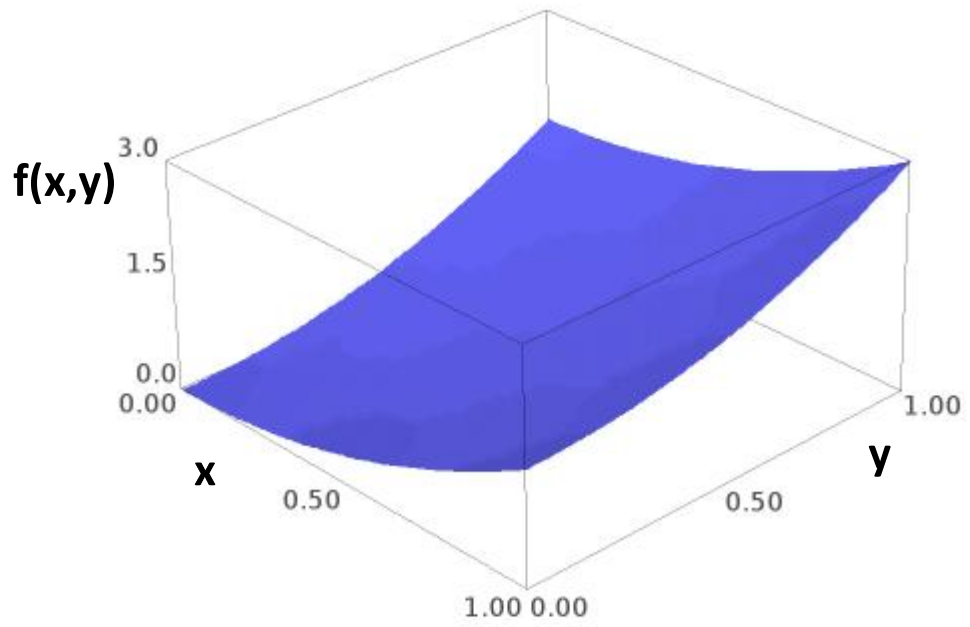
$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Dvodimenzionalne raspodjele

- Primjer: združena funkcija gustoće vjerojatnosti je:

$$f(x,y) = \begin{cases} \frac{3}{2}(x^2 + y^2), & x \in [0,1], y \in [0,1] \\ 0, & inače \end{cases}$$

- Koja je vjerojatnost da je X<1/4 i Y<1/4 ?
- Koja je vjerojatnost da je X< 1/4 bez obzira na Y ?



Podsjetnik: uvjetna vjerojatnost i nezavisnost događaja

- Definicija: događaji A i B su nezavisni onda i samo onda kad vrijedi:

  P (A|B) = P(A)

  (Vjerojatnost da se dogodi A ne ovisi o tome da li se dogodio B)

- Iz definicije uvjetne vjerojatnosti slijedi:

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

- Iz toga slijedi:

$$\boxed{P(A \cap B) = P(A)P(B)}$$

Uvjetna vjerojatnost

- Neka su ***X*** i ***Y*** kontinuirane slučajne varijable sa združenom funkcijom gustoće vjerojatnosti ***f(x,y)*** i rubnom funkcijom $f_X(x)$. Uvjetna funkcija gustoće vjerojatosti za varijablu ***Y***, ako je ***X=x*** je:

$$f_{(Y|X)} = \frac{f(x,y)}{f_X(x)}$$

  Uz uvjet da je $f_X(x) \neq 0$

- Analogno pomoću raspodjela vjerojanosti definiramo uvjetnu vjerojatnost za diskretne varijable:

$$p_{(Y|X)} = \frac{p(x,y)}{p_X(x)}$$

Nezavisne slučajne varijable

- Svaka slučajna varijabla predstavlja ishod nekog događaja, stoga možemo poopćiti definiciju nezavisnih događaja na nezavisne varijable
- Dvije slučajne varijable će biti nezavisne, ako *vrijednost* jedne ne uvjetuje *vrijednost* druge, tj. ako vrijedi:

$$p_{(X|Y)} = \frac{p(x,y)}{p_Y(y)} = p_X(x)$$

- Iz toga slijedi da za nezavisne varijable vrijedi:

$$p(x,y) = p_X(x)p_Y(y)$$

Nezavisne slučajne varijable

- Općenito, slučajne varijable ***X*** i ***Y*** su nezavisne ako za svaki par vrijednosti x i y vrijedi:

$$\boxed{p(x,y) = p_X(x)p_Y(y)}$$ (za diskretne varijable)

$$\boxed{f(x,y) = f_X(x)f_Y(y)}$$ (za kontinuirane varijable)

- Na primjeru raspodjele studenata po ocjeni i po spolu vidimo da te varijable *nisu nezavisne*

Nezavisne slučajne varijable

- Primjer: istovremeno bacanje kocke i novčića
  - slučajna varijabla X opisuje ishod na kocki
  - Slučajna varijabla Y opisuju ishod na novčiću

Ishod	1	2	3	4	5	6	UKUPNO
P	1/12	1/12	1/12	1/12	1/12	1/12	**1/2**
G	1/12	1/12	1/12	1/12	1/12	1/12	**1/2**
UKUPNO	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1**

- Vidi se da za svaki par vrijednosti x i y vrijedi:

$$p(x,y) = p_X(x)p_Y(y)$$

- Dakle varijable su nezavisne!

Momenti dvodimenzionalne raspodjele

- Pomoćni moment:

$$m_{rs} = \sum_{x \in D_X} \sum_{y \in D_Y} x^r y^s p(x,y)$$ Za diskretne varijable

$$m_{rs} = \iint_{\mathbb{R}\times\mathbb{R}} x^r y^s f(x,y)\,dx\,dy$$ Za kontinuirane varijable

- Centralni moment:

$$M_{rs} = \sum_{x \in D_X} \sum_{y \in D_Y} (x-\mu_X)^r (y-\mu_Y)^s p(x,y)$$ Za diskretne varijable

$$M_{rs} = \iint_{\mathbb{R}\times\mathbb{R}} (x-\mu_X)^r (y-\mu_Y)^s f(x,y)\,dx\,dy$$ Za kontinuirane varijable

Očekivanje

- Pomoćni momenti m_{01} i m_{10}:

$$m_{10} = \sum_{x \in D_X} \sum_{y \in D_Y} x\, p(x, y) = \sum_{x \in D_X} x p_X(x) = \mu_X$$

$$m_{01} = \sum_{x \in D_X} \sum_{y \in D_Y} y\, p(x, y) = \sum_{y \in D_Y} y p_Y(y) = \mu_Y$$

- Ovi pomoćni momenti odgovaraju **očekivanjima** diskretnih slučajnih varijabli ***X*** odn. ***Y***
- Analogno vrijedi i za kontinuirani slučaj

Varijanca

- Središnji momenti $\boldsymbol{M_{20}}$ i $\boldsymbol{M_{02}}$ odgovaraju varijancama:

$$M_{20} = \sum_{x \in D_X} \sum_{y \in D_Y} (x - \mu_X)^2 p(x, y)$$

$$= \sum_{x \in D_X} (x - \mu_X)^2 p_X(x) = V(X)$$

$$M_{02} = \sum_{x \in D_X} \sum_{y \in D_Y} (y - \mu_Y)^2 p(x, y)$$

$$= \sum_{y \in D_Y} (y - \mu_Y)^2 p_Y(y) = V(Y)$$

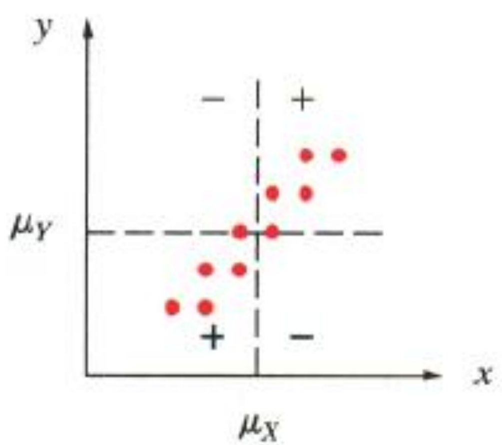
Kovarijanca

- Kovarijanca je veličina koja govori o zavisnosti varijabli ***X*** i ***Y***:

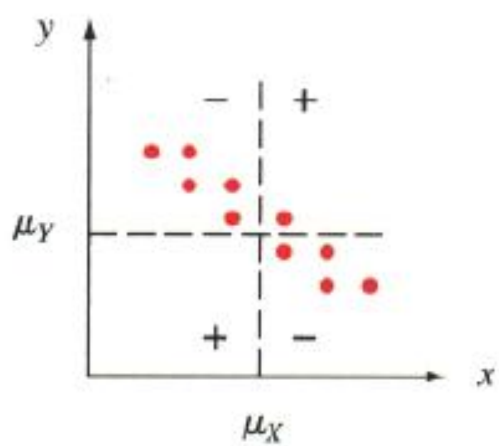
$$Cov(X,Y) = \sigma_{XY} = M_{11}$$

- Izražena pomoću pomoćnih momenata ova relacija postaje:

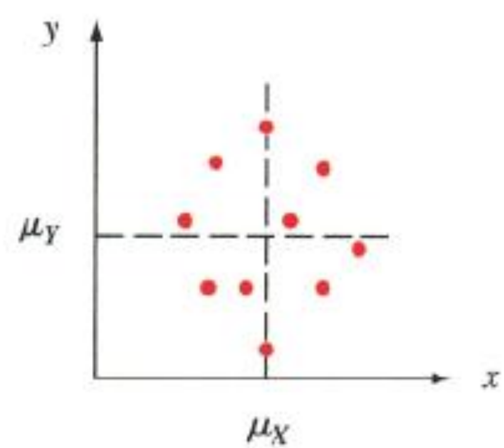
$$Cov(X,Y) = m_{11} - m_{10}m_{01} = E(XY) - E(X)E(Y)$$



Cov(X,Y) > 0



Cov(X,Y) < 0



Cov(X,Y) ~ 0

Kovarijanca nezavisnih varijabli

- Po definiciji za kovarijancu vrijedi:

$$\sigma_{xy} = E(XY) - E(X)E(Y)$$

- Prvi pribrojnik možemo pisati kao:

$$E(XY) = \sum_x \sum_y xy\, p(x,y)$$

- Za nezavisne varijable:

$$= \sum_x \sum_y xy\, p_x(x) p_y(y) = \sum_x x p_x(x) \sum_y y p_y(y)$$

- Slijedi: $E(XY) = E(X)E(Y)$
- Dobivamo da je kovarijanca nezavisnih varijabli: $\boxed{\sigma_{xy} = 0}$

Korelacija

- Kovarijanca ovisi o veličinama koje mjerimo i njihovim jedinicama → nije pogodno za usporedbu povezanosti različitih veličina
- Uvodimo bezdimenzionalnu veličinu – **korelaciju** – kao mjeru povezanosti, tj. zavisnosti varijabli
- **Koeficijent** ***linearne*** **korelacije** varijabli ***X*** i ***Y*** se definira kao:

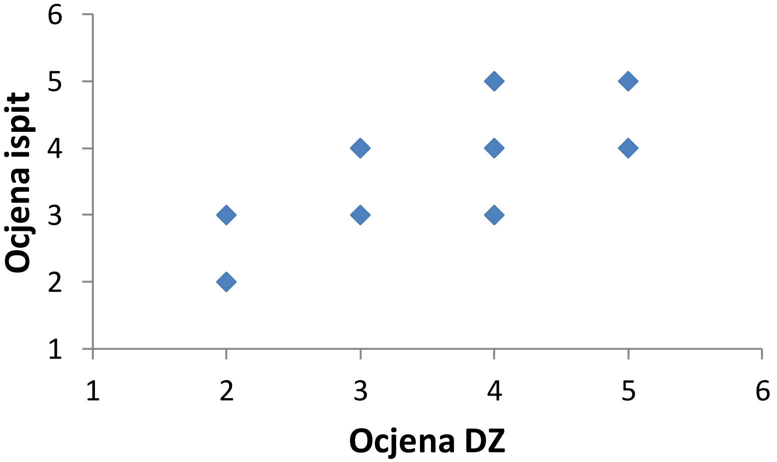
$$\boxed{Corr(X,Y) = \rho_{X,Y} = \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}}$$

(omjer kovarijance i produkta standardnih devijacija)

Korelacija

- Primjer 1: Za 10 studenata uspoređujemo ocjene iz domaće zadaće i ocjene na ispitu

Student	DZ	Ispit
1	5	5
2	4	5
3	4	4
4	2	3
5	5	4
6	2	2
7	3	4
8	3	3
9	4	3
10	3	4



$\sigma_{DZ} = 1,02$

$\sigma_{ISPIT} = 0,90$

$\sigma_{DZ\text{-}ISPIT} = 0,65$ (kovarijanca)

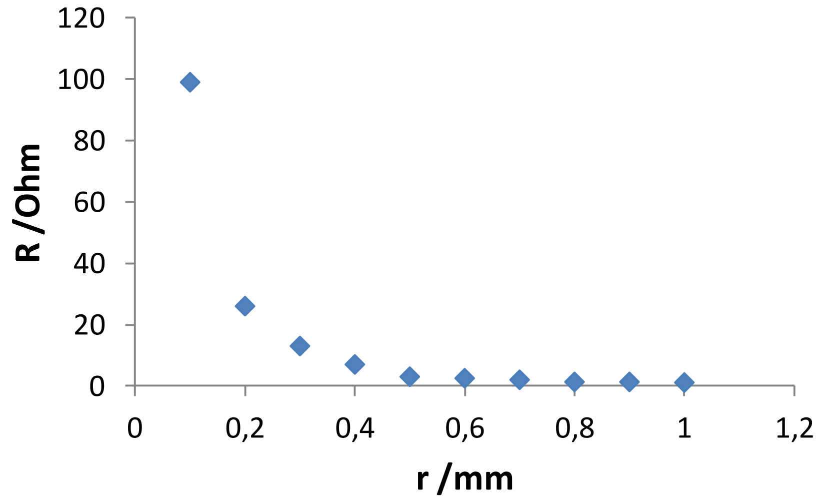
$\rightarrow$ **ρ = 0,7**

Vidimo naznaku linearne ovisnosti

Korelacija

- Primjer 2: Studenti mjere otpor cilindričnog vodiča u ovisnosti o njegovom radijusu:

r /mm	R /Ω
0,1	99
0,2	26
0,3	13
0,4	7
0,5	3
0,6	2,5
0,7	2
0,8	1,30
0,9	1,30
1	1,10



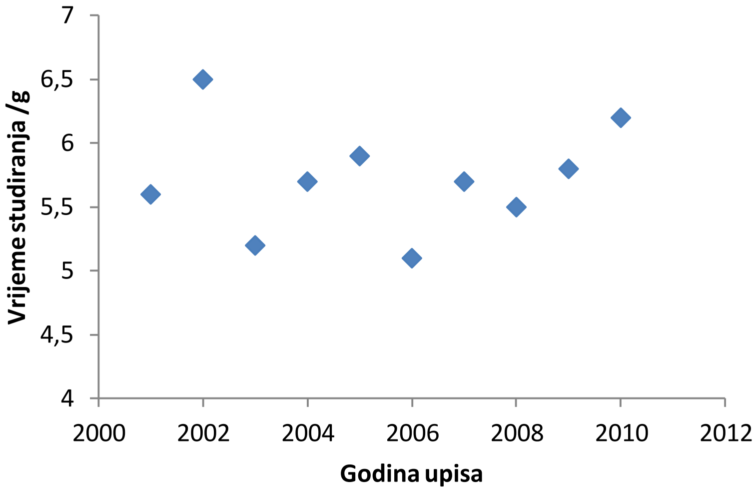
Koeficijent korelacije ρ = - 0,68

Vidimo izraženu obrnuto kvardatnu ovisnost, što teorijski i očekujemo: $R = \frac{1}{\sigma}\frac{l}{\pi r^2}$

Korelacija

- Primjer 3: promatramo vezu između godine upisa studija i trajanja studija

Godina	Trajanje /god
2001	5,6
2002	6,5
2003	5,2
2004	5,7
2005	5,9
2006	5,1
2007	5,7
2008	5,5
2009	5,8
2010	6,2



Koeficijent korelacije $\rho =$ 0,05

Ne uočava se izražena ovisnost!

Korelacija

- Svojstva:
  1. $-1 \leq \rho \leq 1$
  2. Ako su ***X*** i ***Y*** nezavisne varijable onda je ρ=0 → primjer 3 (ali obrat ne vrijedi)
  3. Ako je ovisnost varijabli linearna Y=aX+b onda i samo onda je **ρ=1** ili **ρ=-1** (za a>0 odn. a<0)
  4. Može se još pokazati da vrijedi:

     $$Corr(aX + b, cY + d) = Corr(X, Y)$$
- Općenito ako je **|ρ|<0,5** kažemo da je linearna korelacija slaba (primjer 3), a ako je **|ρ|>0,5** kažemo da je linearna korelacija izražena (primjeri 1 i 2)
- Iz primjera 2 vidimo da imamo izrazito jaku kvadratnu korelaciju, no koeficijent je -0.68 obzirom da je on mjera linearne korelacija

Suma slučajnih varijabli

- Definiramo slučajnu varijablu $\boldsymbol{Y=X_1+X_2}$
- **Očekivanje** slučajne varijable ***Y***:

$$E(Y) = E(X_1) + E(X_2)$$

- Dokaz:

$$\begin{aligned}
E(X_1+X_2) &= \sum_{x_1}\sum_{x_2}(x_1+x_2)p(x_1,x_2) \\
&= \sum_{x_1}\sum_{x_2}x_1p(x_1,x_2) + \sum_{x_1}\sum_{x_2}x_2p(x_1,x_2) \\
&= \sum_{x_1}x_1p_{x1}(x_1) + \sum_{x_2}x_2p_{x2}(x_2) \\
&= E(X_1) + E(X_2)
\end{aligned}$$

Suma slučajnih varijabli

- Definiramo slučajnu varijablu $\boldsymbol{Y=X_1+X}_2$
- **Varijanca** slučajne varijable ***Y***:

$$V(Y) = V(X_1) + V(X_2)+2\sigma_{x1x2}$$

- Dokaz:

$$\begin{aligned} V(X_1+X_2) &= E\{[(x_1+x_2)-(\mu_1+\mu_2)]^2\} \\ &= E\{[(x_1-\mu_1)+(x_2-\mu_2)]^2\} \\ &= E[(x_1-\mu_1)^2]+E[(x_2-\mu_2)^2] \\ &\qquad + 2E[(x_1-\mu_1)(x_2-\mu_2)]) \end{aligned}$$

Linearna kombinacija slučajnih varijabli

- Za slučajne varijable $X_1, X_2, \ldots, X_n$ možemo definirati novu slučajnu varijablu:

$$\boldsymbol{Y=a_1X_1+a_2X_2+\ldots+a_nX_n}$$

gdje su $a_1 \ldots a_n$ konstatne. Y se naziva linearnom kombinacijom varijabli X.

- **Poopćenje relacije za očekivanje:**

$$\boxed{E(a_1X_1+a_2X_2+\cdots+a_nX_n)=a_1E(X_1)+a_2E(X_2)+\cdots+a_nE(X_n)}$$

- Vrijedi bez obzira na to da li su varijable međusobno neovisne

Linearna kombinacija slučajnih varijabli

- **Poopćenje relacije za varijancu:**

$$\boxed{V(a_1X_1+a_2X_2+\cdots+a_nX_n)=a_1^2V(X_1)+a_2^2V(X_2)+\cdots+a_n^2V(X_n)+\sum_i\sum_{j\neq i}a_ia_jCov(X_i,X_j)}$$

- Posebno, ako su varijable nezavisne, vrijedi:

$$\boxed{V(a_1X_1+a_2X_2+\cdots+a_nX_n)=a_1^2V(X_1)+a_2^2V(X_2)+\cdots+a_n^2V(X_n)}$$

- <u>Važan primjer</u>: **varijanca razlike nezavisnih varijabli**

$$\boxed{V(X-Y)=V(X)+V(Y)=\sigma_x^2+\sigma_y^2}$$

Npr. mjerenje spektra s oduzimanjem pozadine.

Linearna kombinacija nezavisnih *normalnih* slučajnih varijabli

- Iz prethodnih razmatranja slijedi specijalni slučaj normalnih varijabli:

Neka su X_1, X_2,...,X_n nezavisne normalne slučajne varijable. Tada njihova linearna kombinacija

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

također ima normalnu raspodjelu s očekivanjem $\mu_Y = \sum_{i=1}^{n} a_i \mu_i$ i varijancom

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 .$$

- Dokaz preskačemo