Motivacija za poslijediplomski iz matematike 2006.

U ovoj skici je uvod u pojam veličine (norme) vektora (složenih podataka, rezultata mjerenja), u pojam udaljenosti medju složenim podatcima i koeficijenta zavisnosti dvaju složenih podataka (skalarni produkt vektora). To je preliminarni materijal za razmatranje matričnih norma, na pr. prema G.M.Phillips,P.J.Taylor, Theory and Applications of Numerical Analysis, pogl. 10 i 11.

I. n-dimenzionalni prostor - koordinatni sustav

n-dimenzionalni koordinatni prostor - koordinatni sustav (za bilo koji prirodni broj n);
to je skup svih uredjenih n-torka realnih brojeva

$$(x_1, x_2, ..., x_n);\ x_1, x_2, ..., x_n \in \mathbf{R}$$

(oznaka $\mathbf{R}^n$).
Za $n = 1$ dobije se koordinatni pravac, za $n = 2$ koordinatna ravnina, za $n = 3$ koordinatni prostor itd.
Ideja je da se matematički zapiše vrijednost neke veličine koja, možda, ima više komponenata:
ako neka veličina ima samo jednu komponentu (na primjer, masa), njena se vrijednost zapisuje jednim brojem,
ako neka veličina ima dvije nezavisne komponente, njena se vrijednost zapisuje uredjenim parom brojeva, itd.
Na primjer, položaj slobodne čestice u prostoru opisuje se uredjenom trojkom.
Uredjene parove, trojke i, općenito, n-torke realnih brojeva možemo smatrati **vektorima**, što obrazlažemo u nastavku.

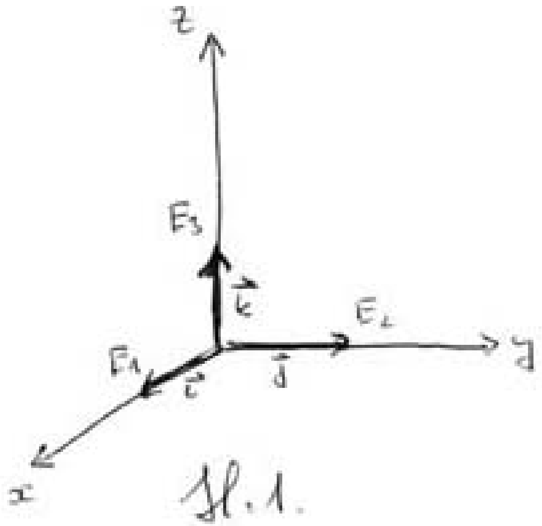
Jedinični vektori u trodimenzionalnom prostoru

Uočimo u koordinatnom prostoru tri točke (redom na pozitivnim djelovima x, y odnosno z osi, na jediničnoj udaljenosti od ishodišta):

$$E_1 := (1, 0, 0),\ E_2 := (0, 1, 0),\ E_3 := (0, 0, 1)$$

Te točke odredjuju tri jedinična vektora **(sl.1)**:

$$\overrightarrow{i} := \overrightarrow{OE_1};\ \overrightarrow{j} := \overrightarrow{OE_2};\ \overrightarrow{k} := \overrightarrow{OE_3};$$



Sl. 1.

Jedinični vektori zapisuju se i pomoću jednostupčanih **matrica**.

$$\vec{i} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} ; \; \vec{j} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} ; \; \vec{k} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Uočite da je to samo drukčiji zapis koordinata završnih točaka tih vektora.

Jedinični vektori u n-dimenzionalnom prostoru

Analogno jediničnim vektorima u ravnini i prostoru, definiraju se jedinični vektori u n-dimenzionalnom prostoru: to je n vektora:

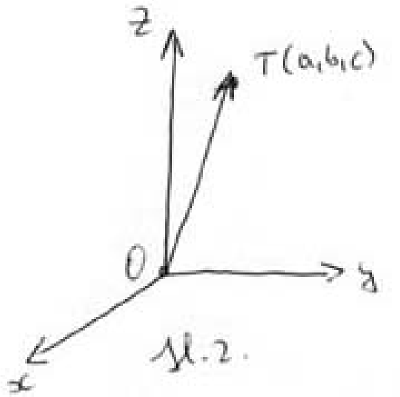
$$\mathbf{e_1}, \; \mathbf{e_2}, ..., \mathbf{e_n}$$

$$\mathbf{e_1} = \begin{bmatrix} 1 \\ 0 \\ \\ 0 \end{bmatrix}, \; \mathbf{e_2} = \begin{bmatrix} 0 \\ 1 \\ \\ 0 \end{bmatrix}, ..., \mathbf{e_n} = \begin{bmatrix} 0 \\ 0 \\ \\ 1 \end{bmatrix}$$

(kod $\mathbf{e_i}$ na i-tom je mjestu 1, a na ostalima je 0.

Radijus vektori - analitički prikaz vektora u koordinatnom prostoru

Točka $T(a, b, c)$ koordinatnog prostora prostora odredjuje jedinstven vektor $\overrightarrow{OT}$ s početkom u ishodištu i završetkom u T (**radijus vektor**) **(sl.2)**.



sl.2.

Vidimo da svaki vektor prostora možemo shvatiti kao radijus vektor. Vidimo, takodjer, da vrijedi:

$$\overrightarrow{OT} = a \cdot \overrightarrow{i} + b \cdot \overrightarrow{j} + c \cdot \overrightarrow{k}$$

Kažemo da smo vektor $\overrightarrow{OT}$ zapisali kao **linearnu kombinaciju** vektora $\overrightarrow{i}$, $\overrightarrow{j}$, $\overrightarrow{k}$ Tu linearnu kombinaciju zapisujemo i kao:

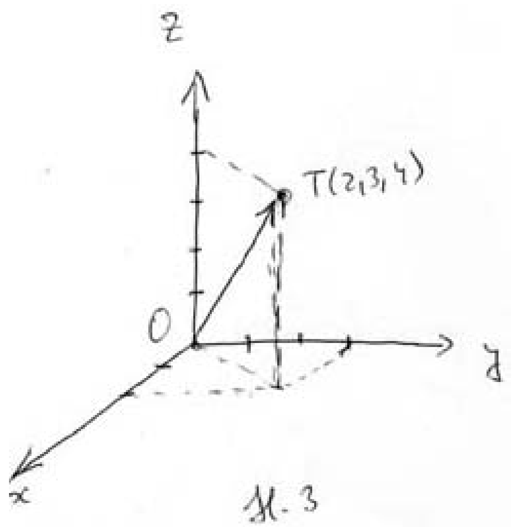
$$\overrightarrow{OT} = \begin{bmatrix} \text{a} \\ \text{b} \\ \text{c} \end{bmatrix}$$

(to je samo drukčiji zapis koordinata točke T). Na primjer, za $T(2,3,4)$ izravno iz slike **(sl.3)** vidi se da je

$$\overrightarrow{OT} = 2\overrightarrow{i} + 3\overrightarrow{j} + 4\overrightarrow{k}$$

odnosno da je

$$\overrightarrow{OT} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$$



sl. 3

Uočite:
Vektori u prostoru mogu se poistovjetiti s točkama u prostoru (tako da ta točka bude završetak, a ishodište početak), a točke u prostoru s jednostupčanim matricama sastavljenim od koordinata tih točaka, dakle:

$$\text{Skup vektora prostora} = \text{Skup matrica oblika} \begin{bmatrix} \text{a} \\ \text{b} \\ \text{c} \end{bmatrix}$$

gdje su a, b, c realni brojevi .
Kad vektor predočimo ovako ili kao linearnu kombinaciju jediničnih vektora $\overrightarrow{i}, \overrightarrow{j}, \overrightarrow{k}$, kažemo da smo ga predočili **analitički**.

U n-dimenzionalnom prostoru $\mathbf{R^n}$ za radijus vektor $\mathbf{v} = \overrightarrow{OT}$ gdje je $T(a_1, a_2, ..., a_n)$, vrijedi

$$\mathbf{v} = a_1\mathbf{e_1} + \mathbf{a_2e_2} + ... + \mathbf{a_ne_n}$$

Taj vektor, zapisujemo u matričnom obliku kao: $\mathbf{v} = \begin{bmatrix} a_1 \\ a_2 \\ \\ a_n \end{bmatrix}$

Konačno, vidimo da uredjenu n-torku $(x_1, x_2, ..., x_n)$ možemo shvatiti kao vektor u n-dimenzionalnom prostoru. Tada je, u pravilu, zapisujemo kao jednostupčanu matricu.

Beskonačno dimenzionalni vektorski prostor.
U matematici, a jednako tako i u primjenama, često dolaze beskonačno dimenzionalni vektorski prostori. Intuitivno, oni dolaze onda kad razmatramo veličine koje imaju beskonačno mnogo komponenata. Tipični primjeri takvih prostora jesu (i) prostor svih polinoma: njima je baza skup potencija

$$1, x, x^2, ...$$

Na primjer, svaki se polimom f trećeg stupnja jednoznačno prikazuje kao linearna kombinacija:

$$a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3$$

pa se može shvatiti kao vektor u četverodimenzionalnom prostoru $f = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$

(ii) prostor svih (neprekinutih) funkcija definiranih na segmentu $[a, b]$, samo što je tu teže doći do baze.
Intuitivno je jasno zašto je funkcija f na segmentu $[a, b]$ vektor u beskonačno dimenzionalnom prostoru - ona u sebi nosi beskonačno mnogo informacija.
Vektorski prostor ima strogu matematičku definiciju. Bitni su sljedeći zahtjevi:
(A) vektori se mogu zbrajati, a svojstva operacije zbrajanja su poput svojtava zbrajanja realnih brojeva.
(B) vektori se mogu množiti sa skalarima (realnim brojevima).

Algebarske operacije s vektorima u koordinatnom sustavu.

Vektore s analitičkim prikazom zbrajamo i množimo sa skalarom kao u primjeru.
Odredimo $2\overrightarrow{u} + 3\overrightarrow{v}$ ako je
$\overrightarrow{u} = 4\overrightarrow{i} - 2\overrightarrow{j} + \overrightarrow{k}$, $\overrightarrow{v} = 3\overrightarrow{i} - 5\overrightarrow{k}$

$2\overrightarrow{u} + 3\overrightarrow{v} = 2(4\overrightarrow{i} - 2\overrightarrow{j} + \overrightarrow{k}) + 3(3\overrightarrow{i} - 5\overrightarrow{k}) =$
$(8\overrightarrow{i} - 4\overrightarrow{j} + 2\overrightarrow{k}) + (9\overrightarrow{i} - 15\overrightarrow{k}) = 17\overrightarrow{i} - 4\overrightarrow{j} - 13\overrightarrow{k}$

Ako bismo se koristili zapisom pomoću jednostupčanih matrica (i ako bismo vektore zapisali masnim slovima, umjesto strjelicama), imali bismo:

$$\mathbf{2u} + \mathbf{3v} = 2\begin{bmatrix} 4 \\ -2 \\ 1 \end{bmatrix} + 3\begin{bmatrix} 3 \\ 0 \\ -5 \end{bmatrix} = \begin{bmatrix} 8 \\ -4 \\ 2 \end{bmatrix} + \begin{bmatrix} 9 \\ 0 \\ -15 \end{bmatrix} = \begin{bmatrix} 17 \\ -4 \\ -13 \end{bmatrix}$$

Slično je s operacijama na vektorima u n-dimenzionalnom prostoru.

II. Veličina-norma vektora, udaljenost, skalarni produkt.

Ako imamo posla s podatcima sastavljenim od n komponenata, tj. ako imamo posla s uredjenim n-torkama, sljedeća su pitanja vrlo prirodna:
1. Kako možemo uvesti pojam "veličine" složenog podatka tj. uredjene n-torke $(x_1, x_2, ..., x_n)$, tj. kako možemo usporedjivati takve podatke po veličini?

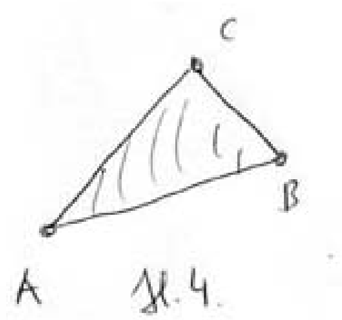
2. Kako možemo uvesti pojam "bliskosti", tj. pojam udaljenosti izmedju dvaju složenih podataka (uredjenih n-torka)? Drugim riječima, možemo li za dva podatka $(x_1, x_2, ..., x_n)$ i $(y_1, y_2, ..., y_n)$ reći je su li blizu ili daleko i kolika je ta njihova udaljenost?

3. Možemo li mjeriti razinu "nezavisnosti" dvaju složenih podataka, odnosno, geometrijskim jezikom govoreći, možemo li mjeriti kut izmedju dvaju složenih podataka?

Odgovor na 1. pitanje daje se uvodjenjem pojma **norme vektora**, na 2. pitanje, uvodjenjem pojma **udaljenosti** dvaju vektora (odnosno dviju točaka prostora), a odgovor na 3. pitanje uvodjenjem pojma **skalarnog produkta** dvaju vektora. Odgovori nisu jednoznačni ni matematički niti u primjenama. Na primjer, izbor pojma udaljenosti dvaju podataka u konkretnim okolnostima, ovisit će često o karakteru podataka (nije, možda, sve jedno jesu li to sociološki, politološki, ekonomski podatci, iz organske ili anorganske kemije, genetike, lingvistike, teorije informacija i sl.).

Pojam udaljenosti-metrike i veličine-norme na prostoru.
Od ovih pojmova primitivniji je pojam udaljenosti, pa ćemo s njim početi. Primitivnost se očituje u tome što za pojam udaljenosti podatci ne moraju činiti vektorski prostor. Uvesti konzistentan pojam udaljenosti medju elementima nekog skupa X, znači uvesti realnu funkciju d sa značenjem

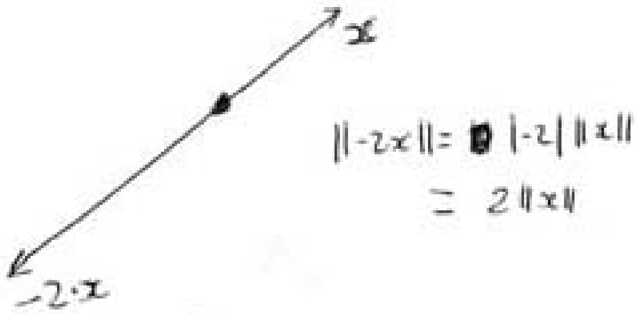
$d(A, B) :=$ udaljenost izmedju A i B, pri čemu d treba zadovoljavati sljedeća očita svojstva:
(D_1) (pozitivnost) $d(A, B) \leq 0$ za sve točke A, B zadanog skupa X; takodjer $d(A, B) = 0$ ako i samo ako je $A = B$.
(D_2) (simetričnost) $d(A, B) = d(B, A)$.
(D_3) (nejednakost trokuta) $d(A, B) + d(B, C) \geq d(A, C)$ (sl.4).



Sl.4.

Za pojam veličine (norme) tražimo da podatci čine vektorski prostor V (dakle, da se mogu zbrajati i oduzimati poput brojeva, i množiti sa skalarima). Norma je funkcija koja se ponaša poput apsolutne vrijednosti na realnim (ili kompleksnim) brojevima. Zato je njena uobičajena oznaka $\|\|$, prema uzoru na oznaku $||$ za apsolutnu vrijednost. Zato je prirodno tražiti da norma na vektorskom prostoru V zadovoljava ova svojstva:

(N_1) (pozitivnost) $||\mathbf{x}|| \geq 0$ za sve vektore $\mathbf{x}$; takodjer $||\mathbf{x}|| = 0$ ako i samo ako je $\mathbf{x} = 0$.
(N_2) (homogenost) $||c\mathbf{x}|| = |c| \cdot ||\mathbf{x}||$ za sve vektore $\mathbf{x}$ i sve brojeve c (sl.5.).



sl. 5.

(N_3) (nejednakost trokuta) $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ za sve vektore $\mathbf{x}, \mathbf{y}$ (sl.6).

sl. 6.

Svaka norma odredjuje neku metriku. Drugim riječima, ako znademo mjeriti veličinu podataka, znat émo mjeriti i udaljenost medju podatcima
Ta jednostavna, ali vrlo važna činjenica je prirodna generalizacija formule za udaljenost točaka na brojevnom pravcu (sl.7).

$$d(a, b) = |b - a|$$

Tako definiramo udaljenost vektora (zamišljamo udaljenost završnih točaka radijus vektora):

$$d(\mathbf{x}, \mathbf{y}) := ||\mathbf{y} - \mathbf{x}||$$

Sl.7

Veza izmedju norme i metrike predočena je na slici 8.

Sl.8

Primjeri nekih važnih norma i metrika.
Navodimo neke norme i metrike važne u matematici i primjeni. Uz svaku dajemo jednostavan primjer i crtamo jediničnu kružnicu, tj. skup svih točaka (podataka) kojima je udaljenost od ishodišta, u toj metrici, jednaka 1, tj. skup svih vektora jedinične duljine.

Stavimo $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix}$.

Obično, ako zamišljamo točke u prostoru, odnosno složene podatke, stavljamo: $\mathbf{x} = (x_1, x_2, ..., x_n)$, $\mathbf{y} = (y_1, y_2, ..., y_n)$. U skladu s tim, često točke u prostoru označavamo kao vektore: $\mathbf{x}$, $\mathbf{y}$ itd. (zamišljamo ih kao završne točke radijus vektora - početna je svima ishodište).

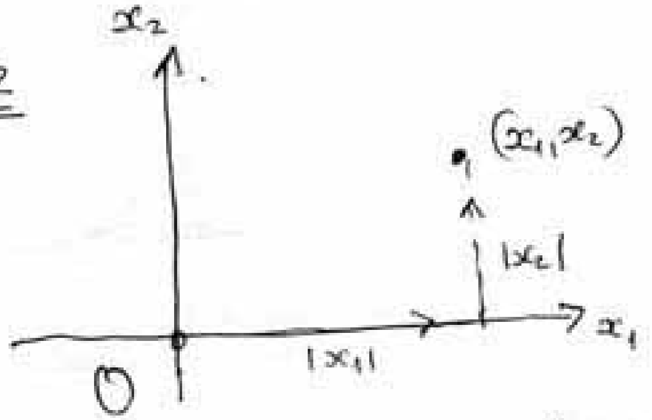
1 - norma : veličina podatka je zbroj apsolutnih vrijednosti komponenata.

$$\|x\|_1 := |x_1| + |x_2| + \dots + |x_n| = \sum_{i=1}^{n} |x_i|$$

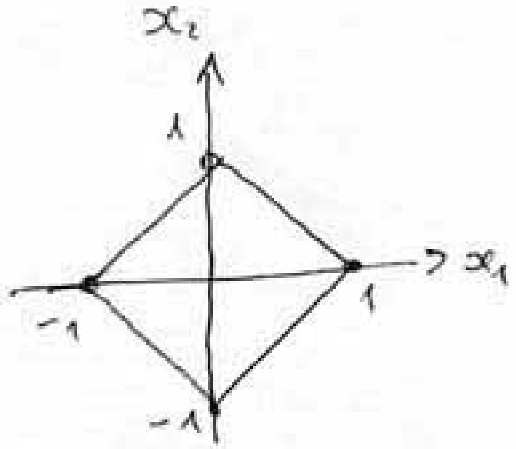
1 - metrika : udaljenost dvaju podataka je zbroj udaljenosti odgovarajućih komponenata.

$$d_1(x,y) := \|y-x\|_1 = |y_1-x_1| + |y_2-x_2| + \dots + |y_n-x_n| = \sum_{i=1}^{n} |y_i-x_i|$$

n = 2

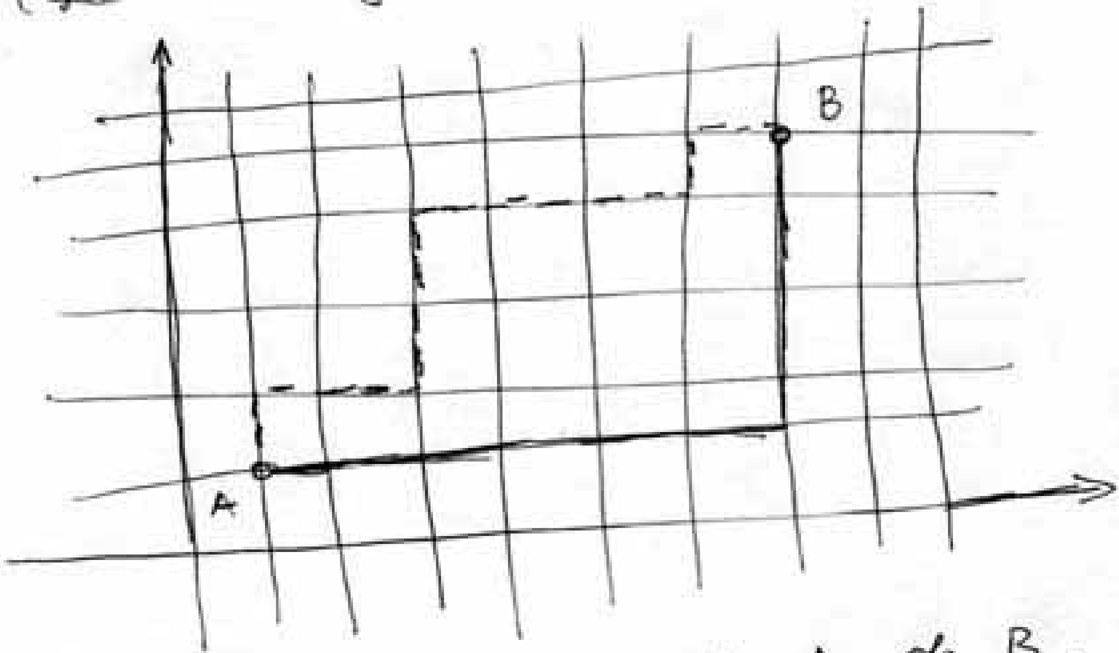


Najkraća spojnica ishodišta i točke (x_1,x_2) nije dužina, već unija horizontalnog i vertikalnog puta



Jedinična kružnica u 1 - normi

1 – metrika zove se i <u>Manhattan metrika</u> prema uzoru na horizontalne i vertikalne ulice u velegradu. Udaljenost dviju točaka je zbroj horizontalne i vertikalne ~~[illegible]~~ udaljenosti.



Dva najkraća puta od A do B. Svaki od njih ima 6 horizontalnih i 4 vertikalna pomaka. Tako je sa svakom dva najkraća puta između A i B.

$$d_1(A, B) = |x_B - x_A| + |y_B - y_A|$$

2 - norma (euklidska norma):

veličina podatka je duljina dužine koja spaja ishodište s tim podatkom.

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$
$$= \sqrt{\sum_{i=1}^{n} x_i^2}$$

2 - metrika (euklidska metrika):

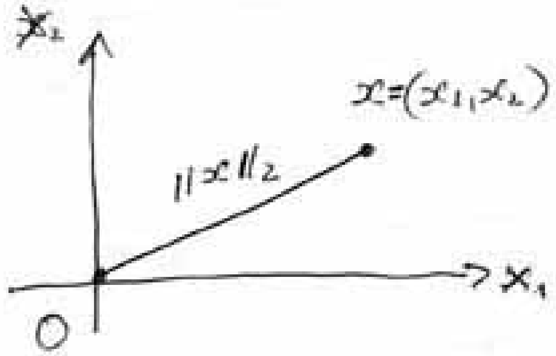
udaljenost dvaju podataka je duljina dužine koja ih spaja.

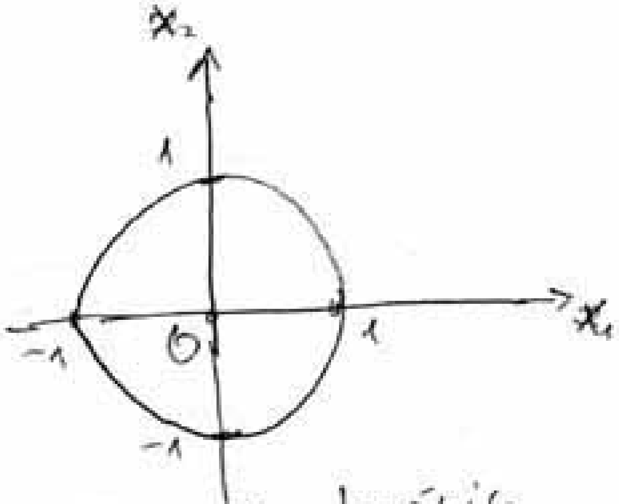
$$d_2(x,y) := \|y-x\|_2$$
$$= \sqrt{(y_1-x_1)^2 + (y_2-x_2)^2 + \dots + (y_n-x_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (y_i-x_i)^2}$$

Dakle, euklidska udaljenost je drugi korijen iz sume kvadrata pojedinačnih udaljenosti (udaljenosti odgovarajućih komponenata).

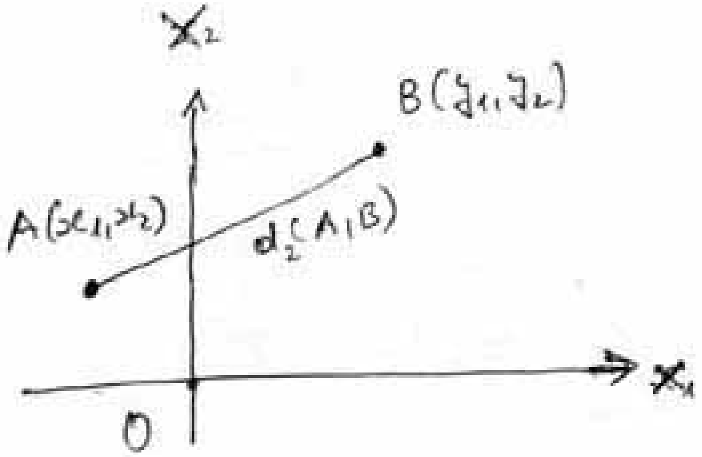
Euklidska norma i euklidska metrika ima najširu uporabu. Zato se obično nazivaju standardnom normom i standardnom metrikom.

$n = 2$





Jedinična kružnica u standardnoj normi.



U ravnini (za $n=2$) i u prostoru (za $n=3$) formule za euklidsku normu i euklidsku metriku proizlaze iz Pitagorina poučka, a za $n > 3$ iz analogona tog poučka.

max - norma (∞ - norma):

Veličina podatka je najveća od veličina komponenata koje sudjeluju u tom podatku.

$$\|x\|_{max} := \max\{|x_i|\} \quad i=1,2,\dots,n$$

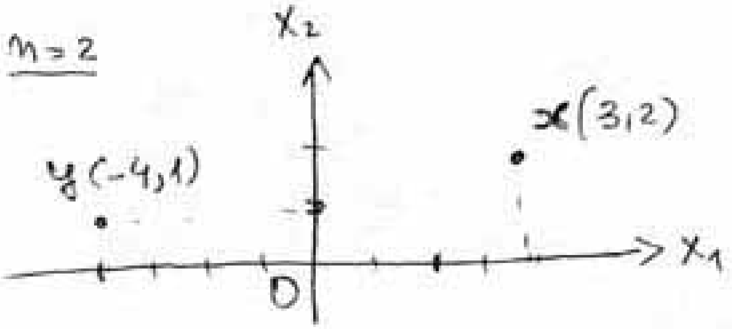
max - metrika (∞ - metrika):

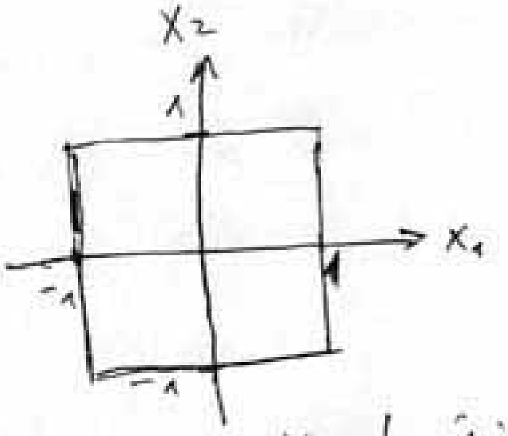
udaljenost dvaju podataka je najveća od udaljenosti odgovarajućih komponenata.

$$d_{max}(x,y) := \|y-x\|_{max}$$
$$= \max\{|y_i-x_i|\} \quad i=1,2,\dots,n$$

Umjesto $\| \, \|_{max}$ obično se piše $\| \, \|_\infty$, a umjesto d_{max} piše se d_∞.

n=2

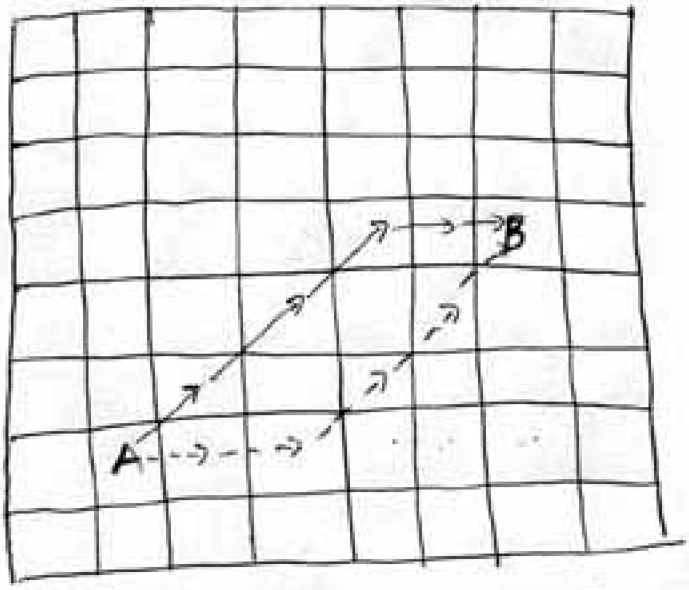




$$|x|_{max} := \max\{|3|,|2|\} = 3$$
$$|y|_{max} := \max\{|-4|,|1|\} = 4$$
$$d_{max}(x,y) := \max\{|-4-3|,|1-2|\} = 7$$

Jedinična kružnica u max-normi.

max-norma često se zove i Čebiševljeva norma), a max-metrika Čebiševljeva metrika), takodjer i šahovska metrika. Razlog za ~~posljednji~~ posljednji naziv je taj što se najkraći put ~~od drug~~ koji treba učiniti kralj da dođe ~~iz~~ iz jednog polja u drugo, računa prema toj metrici. Podsjetimo da kralj ide po jedno polje horizontalno, vertikalno ili dijagonalno.



Horizontalna udaljenost polja A i B je 5, vertikalna udaljenost 3. Zato kralju treba 5 poteza da iz A dođe u B. Na slici su predočena dva takva najkraća puta (jedan punim, a drugi iscrtkanim strjelicama).

Norme i metrike Minkovskog.

Za svaki realni broj $p \geq 1$ definiramo

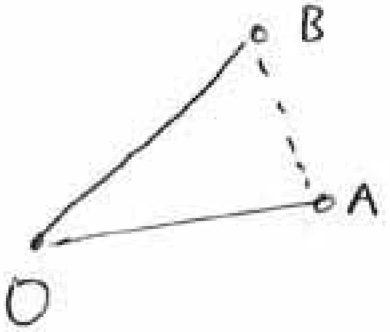
$$\|x\|_p :\!= \sqrt[p]{|x_1|^p + |x_2|^p + \dots + |x_4|^p}$$

$$d_p(x,y) :\!= \|y-x\|_p$$
$$= \sqrt[p]{|y_1-x_1|^p + |y_2-x_2|^p + \dots |y_4-x_4|^p}$$

Sve ove metrike imaju primjenu. Za $p=1$, dobije se 1-metrika, za $p=2$ dobije se 2-metrika, a pri limesu $p \to \infty$ (dakako, za $p=\infty$), dobije se max-metrika. To je razlog zašto se max-metrika zove i ∞-metrika.

Slično je za norme.

Poštanska metrika –



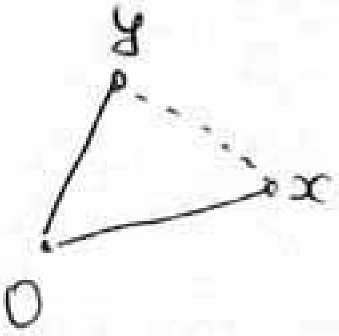
dobila je ime prema primjeru u poštanskom prometu sa središnjom poštom – oznaka O.

$d_{pošt}(A,B)$ označava duljinu puta što ga pomka treba proći da obide iz mjesta A u mjesto B. Tako je

$$d_{pošt.}(A,B) = d(O,A) + d(O,B)$$

gdje je $\underset{\sim}{d}$ obična udaljenost. ~~mjesto A~~

To vrijedi za $A \neq B$, dok je $d(A,A) = 0$.



Sve se može zapisati i pomoću vektora:

$$d_{pošt.}(x,y) = \|x\| + \|y\|$$

gdje $\|x\|$ znači duljinu vektora x (tj. udaljenost od njegova početka O i završetka).

Canberrina metrika - to je metrika koja uzima u obzir veličinu podataka. Ako su $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n)$ dva niza podataka, onda je

$$d_{Canb}(x, y) = \frac{|y_1 - x_1|}{x_1 + y_1} + \frac{|y_2 - x_2|}{x_2 + y_2} + \dots + \frac{|y_n - x_n|}{x_n + y_n}$$

Ta je metrika zanimljiva već za $n=1$. Na primjer.

$$d_{Canb}(1, 2) = \frac{|2-1|}{1+2} = \frac{1}{3}$$

$$d_{Canb}(10, 20) = \frac{|20-10|}{10+20} = \frac{10}{30} = \frac{1}{3}$$

$$d_{Canb}(100, 200) = \frac{1}{3}$$

dok je, na primjer, $d_{Can}(10, 11) = \frac{1}{21}$.

Vidimo da je ta metrika homogena. To znači da rezultat ne ovisi o izboru jedinice mjere. Na primjer, sve jedno je mo li rezultate mjerenja izrazili u metrima, decimetrima ili centimetrima.

Uočite da je u Canberrinoj metrici udaljenost od 1 do 2 veća od udaljenosti od 10 do 11 (iako su te udaljenosti u standardnoj metrici jednake). Dakle, Canberrina je metrika neka vrste relativne udaljenosti.

Hammingova metrika.
To je metrika koja mjeri udaljenost dviju jednostavnih informacaja (obično zadanih u obliku niza znakova 0 i 1 stalne duljine). Ona se definira kao broj različitih mjesta u porukama. Na primjer, ako je:
$\mathbf{x} := 10110001$ i $y := 11110000$, onda je $d_H(\mathbf{x}, \mathbf{y}) = 2$, jer se ove poruke ralikuju na dvjema mjestima: na drugom i osmom.

Ova se metrika poopćuje na poruke (riječi) različitih duljina i definira se kao najmanji broj izmjena slova (znamenaka) kojim se jedna riječ pretvara u drugu. Ta se metrika zove **Levenshteinova metrika**. Na primjer, ako je
$\mathbf{x} := 1011$ i $y := 11100$, onda je $d_H(\mathbf{x}, \mathbf{y}) = 3$.
Naime, da od $\mathbf{x}$ napravimo $\mathbf{y}$, treba 0 na drugom mjestu pretvoriti u 1, potom 1 na četvrtom mjestu pretvoriti u 0 i, konačno, treba dodati nulu. Uočite da su, takodjer, tri izmjene potrebne, da se od druge poruke dobije prva.
Ta se metrika često koristi u lingvistici. Na primjer, prema njoj, udaljenost riječi SLAVAN i GLAVA je 2, jer da se od prve dobije druga, treba G zamijeniti sa S i odbaciti N.
Levenstheinova se metrika proširuje dalje na Demerau-Levenstheinovu metriku koja zamjenu slova (znamenaka) na dvjema mjestima smatra jednim korakom.

Genetske metrike.
To je niz metrika kojima se mjeri genetska udaljenost dviju populacija ili jedinka. Precizan opis bilo koje od njih zahtijeva velike pripreme i uvodjenje novih pojmova. Zato ćemo ugrubo natuknuti najjednostavniju od formula. To je Cavalli-Sforza formula za udaljenost bez ikakvih bioloških predpostavaka (evolucije i sl.), ako gledamo dvije populacije S_1, S_2 i samo jedan lokus (recimo boju) za koji postoji k mogućih kodiranja (**allela**). Označimo, za $i = 1, 2, ..., k$:
s_{1_i} := relativna frekvencija i-tog allela u I. populaciji,
s_{2_i} := relativna frekvencija i-tog allela u II. populaciji.
Tada je udaljenost tih dviju populaciju s obzirom na to mjesto(lokus):
$d_{C-Sf}(S_1, S_2) := \sqrt{2 - 2[\sqrt{s_{1_1} s_{2_1}} + \sqrt{s_{1_2} s_{2_2}} + ... + \sqrt{s_{1_k} s_{2_k}}]} =$
$\sqrt{2 - 2\sum_{i=1}^{k} \sqrt{s_{1_i} s_{2_i}}}$.

Može se pokazati da se ta formula zasniva na pojmu mjerenja duljine najkraće spojnice dviju točaka na višedimenzionalnoj sferi, što je poznato kao geodetska ili geografska metrika.

Pseudometrike.
Često su objekti takvi da funkcija kojom mjerimo udaljenost ne zadovoljava neki od onih uvjeta koje smo postavili za metriku. Tada govorimo o pseudometrikama. Na primjer, može se dogoditi da za dva različita objekta udaljenost bude nula, što tumačimo tako da je razlika medju njima zanemariva. Takodjer, relacija može biti nesimetrična (na primjer u ireverzibilim procesima), a može izostati i nejednakost trokuta.
U takve se funkcije udaljenosti mogu ubrojiti i mnoge koje se pojavljuju u statistici (na primjer, nekoliko Pearsonovih formula za korelacijske koeficijente). Tu ćemo spomenuti samo funkciju mjere udaljenosti serije teoretskih podataka od serije eksperimentalnih podataka, koja se pojavljuje pri χ^2-testu:
Neka su: $f_1, f_2, ..., f_n$

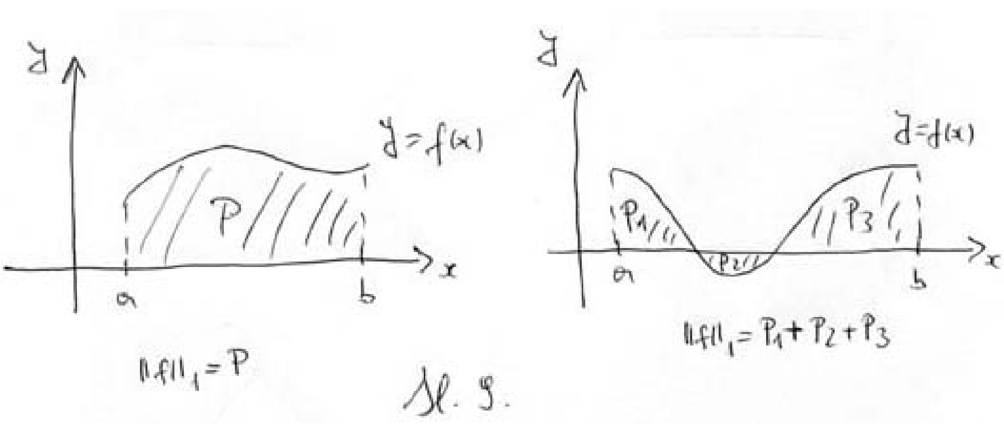
eksperimantalne frekvencije dobivene promatranjem nekog statističkog obilježja. Neka su $f_{t_1}, f_{t_2}, ..., f_{t_n}$
pripadne teoretske frekvencije. Tada je

$$\chi^2_{exp} := \frac{(f_1 - f_{t_1})^2}{f_{t_1}} + \frac{(f_2 - f_{t_2})^2}{f_{t_2}} + ... + \frac{(f_n - f_{t_n})^2}{f_{t_n}}$$

mjera udaljenosti tih dvaju podataka (zasnovani na χ^2 razdiobi). Može se precizno matematički odgovoriti kada tu udaljenost smatramo zanemarivom, a kad ozbiljnom.

Norma i metrika na prostoru funkcija.
Već smo rekli da, na primjer, funkcije definirane na nekom segmentu, recimo $[a, b]$ čine vektorski prostor. Postavlja se pitanje norme i metrike na tom prostoru. Radi jednostavnosti, ograničit ćemo se samo na **neprekinute funkcije**, iako se sve može provesti i općenitije (a to primjene i zahtijevaju). Ima više "razumnih" norma (dakle i metrika), a spomenut ćemo tri premu uzoru na 1-normu, 2-normu i normu *maks* na n-dimenzionalnom realnom vektorskom prostoru.

1-norma na neprekinutim funkcijama na segmentu (sl.9.)



$$||f||_1 := \int_a^b |f(x)| dx$$

Dakle, veličina funkcije je površina izmedju grafa funkcije i osi x.
Odavde se dobije formula za 1-metriku; udaljenost funkcija f, g je:

$$d_1(f, g) = \int_a^b |g(x) - f(x)| dx$$

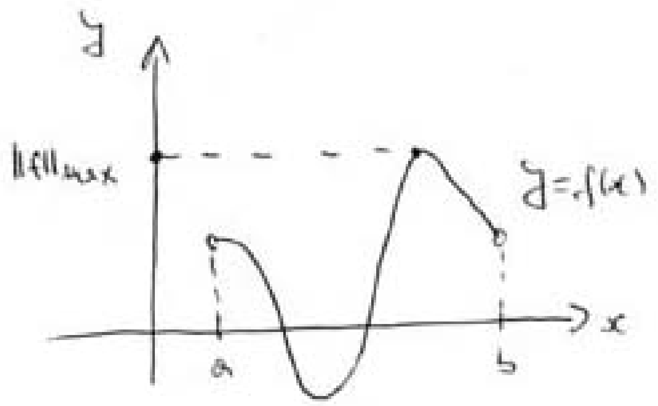
2-norma, euklidska norma na neprekinutim funkcijama na segmentu

$$||f||_2 := \sqrt{\int_a^b f^2(x) dx}$$

Odavde se dobije formula za euklidsku (2-metriku); udaljenost funkcija f, g je:

$$d_2(f,g) = \sqrt{\int_a^b [g(x) - f(x)]^2 dx}$$

maks-norma na neprekinutim funkcijama na segmentu (sl.10.)



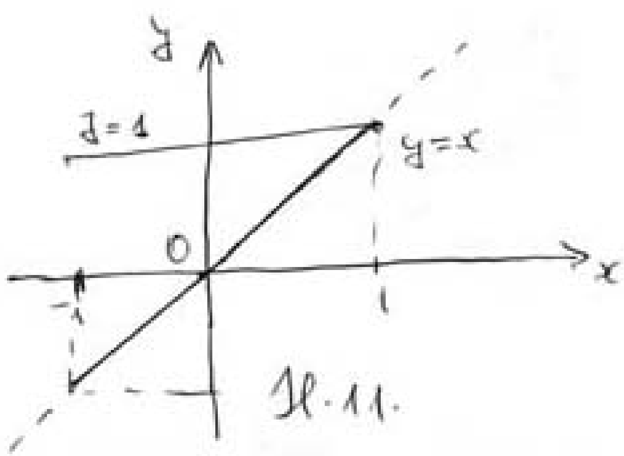
sl.10

$$||f||_{maks} := max_{x\in[a,b]}|f(x)|$$

Dakle, veličina funkcije je najveća vrijednost funkcije na intervalu.
Odavde se dobije formula za max-metriku; udaljenost funkcija f, g je:

$$d_{maks}(f,g) = max_{x\in[a,b]}|g(x) - f(x)|$$

Primjer. Neka je $[a, b] := [-1, 1]$, $f := 1$, $g(x) := x$. Tada je (sl.11.):



$||f||_1 = 2$, $||f||_2 = \sqrt{2}$, $||f||_{maks} = 1$
$||g||_1 = 1$, $||g||_2 = \sqrt{\frac{2}{3}}$, $||g||_{maks} = 1$
$d_1(1,x) = 1$, $d_2(1,x) = 2\sqrt{\frac{2}{3}}$, $d_{maks}(1,x) = 2$.

Kut medju vektorima - mjera zavisnosti dvaju podataka.

Prema uzoru na kut medju vektorima u ravnini i prostoru (trodimenzionalnom), uvodimo pojam kuta medju vektorima n-dimenzionalnog vektorskog prostora. Ako stavimo, kao i prije:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix},$$

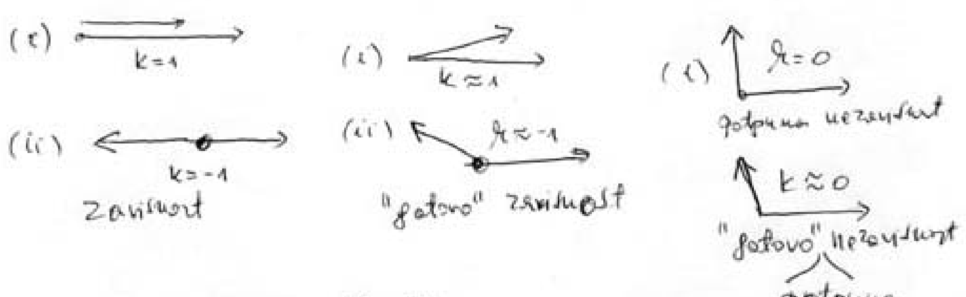
onda definiramo:
skalarni produkt vektora x, y kao zbroj umnožaka odgovarajućih komponenata:

$$\mathbf{x} \cdot \mathbf{y} := x_1y_1 + x_2y_2 + ... + x_ny_n = \sum_{i=1}^{n} x_iy_i$$

i kut α medju vektorima:

$$\cos\alpha := \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||_2||\mathbf{y}||_2} = \frac{x_1y_1 + x_2y_2 + ... + x_ny_n}{\sqrt{x_1^2 + x_2^2 + ... + x_n^2} \cdot \sqrt{y_1^2 + y_2^2 + ... + y_n^2}}$$

Tu smo se koristili euklidskom normom vektora. Napominjemo da se strogo matematički može dokazati da je to bilo forsirano, u smislu da takvo sto nimo mogli napraviti s 1-normom, max-normom, niti s nekom drugom normom Minkovskog, različitom od 2-norme. Dakle, za razumnu geometriju medju vektorima, potrebna nam je 2-norma.
Intuitivno, pojam kuta medju vektorima-složenim podatcima potreban nam je kao **mjera zavisnosti** (sl.12.).



Sl. 12.

Što je kut medju vektorima bliži nula kutu ili ispruženom kutu, podatci medjusobno "zavisniji", a što je bliži pravom kutu (skalarni produkt bliže nuli), podatci su to nezavisniji. Puna zavisnost je ako je kut nula (pozitivna kolinearnost) ili ispruženi kut (negativna kolinearnost).
Formula za kosinus kuta u analogiji je s pojmom **koeficijenta korelacije** cor_{xy}

dviju serija podataka $(x_1, x_2, ..., x_n)$ i $(y_1, y_2, ..., y_n)$ u statistici, samo što izvorne podatke treba zamijeniti s odstupanjima od srednje vrijednosti (aritmetičke sredine):

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}, \ \bar{y} = \frac{y_1 + y_2 + ... + y_n}{n}$$

Koeficijent korelacije definira se kao

$$cor_{xy} := \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + ... + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2} \cdot \sqrt{(y_1 - \bar{y})^2 + ... + (y_n - \bar{y})^2}}$$

Koeficijent korelacije je realan broj izmedju -1 i 1. Što je on bliže nuli, podatci su to nezavisniji, a što je bliži 1 ili -1 oni su to zavisniji.
Razmatrali smo odsutapanja od srednje vrijednosti jer smo zainteresirani za distribuciju tih podatka. U statistici još intuitivnije značenje imaju pojmovi **prosječne sume odstupanja** od srednje vrijednosti pa se definira:
kovarijanca Cov_{xy} dviju serija podataka, prema analogiji sa skalarnim produktu kao:

$Cov_{xy} := \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + ... + (x_n - \bar{x})(y_n - \bar{y})}{n}$

varijanca serije podataka $x_1, x_2, ..., x_n$ kao:

$Var_x = s_x^2 := \frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n}$

(dakle s_x je euklidska norma vektora-podatka $x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x}$ podijeljena s $\sqrt{n}$, tj. s drugim korijenom iz broja komponenata).
Sad se koeficijent korelacije može definirati i kao:

$cor_{xy} := \frac{Cov_{xy}}{s_x s_y}$.

Treba napomenuti da postoje analogni izrazi kod kojih je u nazivniku $n - 1$ umjesto n, što ima praktični i teoretski smisao.