Višedimenzionalne slučajne varijable i Teorija pogrešaka

Višedimenzionalne slučajne varijable

Jednostavan primjer:
- možemo se pitati kolika je vjerojatnost da je koordinata x neke točke u intervalu (x_1, x_2), te tretirati taj slučaj 1D statističkim pristupom
- no, možemo se i pitati i kolika je vjerojatnost da je položaj neke točke u određenom dijelu ravnine, tj. da je $x_1 < x < x_2$ i $y_1 < y < y_2$
- slučajna varijabla tada je vektor $\vec{r} = x\hat{i} + y\hat{j}$, koji se sastoji od dvije nezavisne dimenzije x i y

Dakle:

$$P(x) \quad \leftrightarrow \quad P(\vec{r}) = P(x,y) \quad \leftrightarrow \quad P(\vec{r}) = P(x,y,z)$$

$$1D \qquad\qquad\qquad 2D \qquad\qquad\qquad 3D$$

$$\vec{r} = x\hat{i} + y\hat{j} \qquad\qquad \vec{r} = x\hat{i} + y\hat{j} + z\hat{k}$$

Višedimenzionalne slučajne varijable

Def.: neka su *X* i *Y* dvije diskretne slučajne varijable definirane na prostoru elementrarnih događaja nekog eksperimenta. Združena raspodjela vjerojatnosti $p(x,y)$ definirana je za svaki par točaka (x,y) kao vjerojatnost da istodobno varijabla X poprimi vrijednost x i varijabla Y poprimi vrijednost y:

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

Primjer

Na PMF-u studente razvrstamo prema područjima studiranja (varijabla *X*) i prema spolu (varijabla *Y*). Združenu raspodjelu vjerojatnosti za dva obilježja (područje i spol) možemo prikazati tablicom:

	M	Ž	
Matematika	0,151	0,105	
Fizika	0,125	0,079	
Kemija	0,069	0,074	
Biologija	0,062	0,153	
Geo-znanosti	0,124	0,058	

Višedimenzionalne slučajne varijable - rubne raspodjele vjerojatnosti

Def.: Rubne (marginalne) raspodjele vjerojatnosti za varijable *X* i *Y* označavamo $p_x(x)$ i $p_y(y)$, a dane su izrazima:

$$p_x(x) = \sum_{y \in D_y} p(x,y) \quad i \quad p_y(y) = \sum_{x \in D_y} p(x,y)$$

U gornjem primjeru je, npr., rubna vjerojatnost za matematiku $p_x(\text{mat}) = 0.256$, a rubna vjerojatnost za muški spol $p_y(M) = 0.531$.

	M	Ž	
Matematika	0,151	0,105	
Fizika	0,125	0,079	
Kemija	0,069	0,074	
Biologija	0,062	0,153	
Geo-znanosti	0,124	0,058	

Višedimenzionalne slučajne varijable - očekivanja

Očekivanja su dana sljedećim izrazima:

$$E(x)=\sum_{i=1}^{n} x_i P(x_i)=\sum_{i,j} x_i P(x_i,y_j),$$

$$E(y)=\sum_{j=1}^{n} y_j P(y_j)=\sum_{i,j} y_j P(x_i,y_j)$$

Neovisnost varijabli *X* i *Y*

- sjetimo se: dva su događaja neovisna ukoliko je $P(A_1 \cdot A_2) = P(A_1)P(A_2)$

Slično, dvije su varijable neovisne ako je:
$p(x_i, y_j) = P(x_i)P(y_j)$

Višedimenzionalne slučajne varijable - očekivanja

Neka važna svojstva očekivanja:

(1) $E(x+y) = E(x) + E(y)$

$$E(x+y)=\sum_{i,j}(x_i+y_j)p(x_i,y_j)=\sum_{i,j}x_i p(x_i,y_j)+\sum_{i,j}y_j p(x_i,y_j)=E(x)+E(y)$$

(2) Ako su x i y nezavisne varijable vrijedi:
$E(xy) = E(x)E(y)$

$$E(x\cdot y)=\sum_{i,j}(x_i\cdot y_j)p(x_i,y_j)=\sum_{i,j}x_i y_j P(x_i)P(y_j)=$$
$$=\sum_i x_i P(x_i)\sum_j y_j P(y_j)=E(x)E(y)$$

Višedimenzionalne slučajne varijable - momenti

Pomoćni moment reda (*r,s*) je: $\boxed{m_{rs}=\sum_{i,j}x_i^r y_j^s p\left(x_i,y_j\right)}$

$$1D: m_r=\sum_i x_i^r p\left(x_i\right)$$

Centralni moment reda (*r,s*) je: $\boxed{M_{rs}=\sum_{i,j}(x-\mu)^r(y-\nu)^s p\left(x_i,y_j\right)}$

gdje je μ = *E*(*x*), a ν = *E*(*y*)

$$1D: M_r=\sum_i\left(x_i-\mu\right)^r p\left(x_i\right)$$

Posebno: $\mu=E(x)=\sum_{i,j}x_i p\left(x_i,y_j\right)=m_{10},$

$$\nu=E(y)=\sum_{i,j}y_j p\left(x_i,y_j\right)=m_{01},$$

$$\sigma_x^2=V(x)=\sum_{i,j}\left(x_i-\mu\right)^2 p\left(x_i,y_j\right)=M_{20},$$

$$\sigma_y^2=V(y)=\sum_{i,j}\left(y_j-\nu\right)^2 p\left(x_i,y_j\right)=M_{02}.$$

Višedimenzionalne slučajne varijable - kovarijanca

Kovarijancom dviju varijabli nazivamo:

$$M_{11} = \sigma_{xy} = \sum_{i,j} (x_i - \mu)(y_i - \nu) p(x_i, y_j)$$

Jednostavnije je pisati:

$$\sigma_{xy} = \sum_{i,j} x_i y_j p(x_i, y_j) - \mu \underbrace{\sum_{i,j} y_j p(x_i, y_j)}_{\nu} - \nu \underbrace{\sum_{i,j} x_i p(x_i, y_j)}_{\mu} + \mu\nu \underbrace{\sum_{i,j} p(x_i, y_j)}_{1}$$

$$\boxed{\begin{aligned} &\sigma_{xy} = \sum_{i,j} x_i y_j p(x_i, y_j) - \mu\nu = E(xy) - E(x)E(y) \\ &M_{11} = m_{11} - m_{10} m_{01} \end{aligned}}$$

Kovarijanca je mjera toga koliko se se varijable zajedno mijenjaju

Višedimenzionalne slučajne varijable - kovarijanca

Primjer:

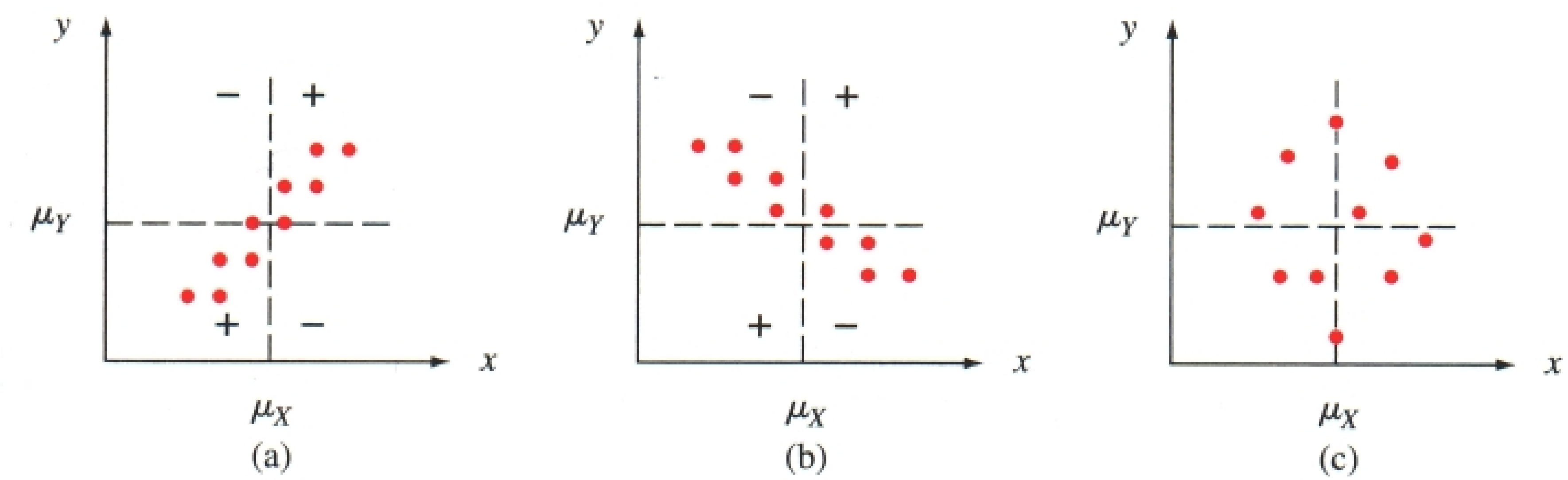


Figure 5.4 $p(x, y)$ = 1/10 for each of ten pairs corresponding to indicated points; (a) positive covariance; (b) negative covariance; (c) covariance near zero

Važno! Kovarijanca nezavisnih varijabli jednaka je nuli

$$\sigma_{xy} = \sum_{i,j} x_i y_j p(x_i, y_j) - \mu\nu = \sum_i x_i p(x_i) \sum_j y_j p(y_j) - \mu\nu = \mu\nu - \mu\nu = 0$$

To je vrlo važno svojstvo, jer nam kazuje da li postoji **KORELACIJA** danog skupa kao rezultat mjerenja, npr. jesu li visine majki i kćeri povezane

Višedimenzionalne slučajne varijable - kovarijanca

Neka važna svojstva varijanci:

$$V\left(x+y\right)=V\left(x\right)+V\left(y\right)+2\sigma_{xy}$$
$$V\left(ax+by\right)=a^2V\left(x\right)+b^2V\left(y\right)+2ab\sigma_{xzy}$$

Kovarijanca je veličina koja nam govori o povezanosti (zavisnosti) varijabli *X* i *Y*. Međutim, iznos kovarijance ovisi i o jedinicama u kojima mjerimo. Da bismo mogli nešto reći o stupnju zavisnosti, moramo tu povezanost izraziti bezdimenzionalnim veličinama. Stoga uvodimo pojam korelacije.

Višedimenzionalne slučajne varijable - korelacija

Def.: Koeficijent korelacije varijabli *X* i *Y* definiran je kao:

$$Corr(X,Y)=\rho_{x,y}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

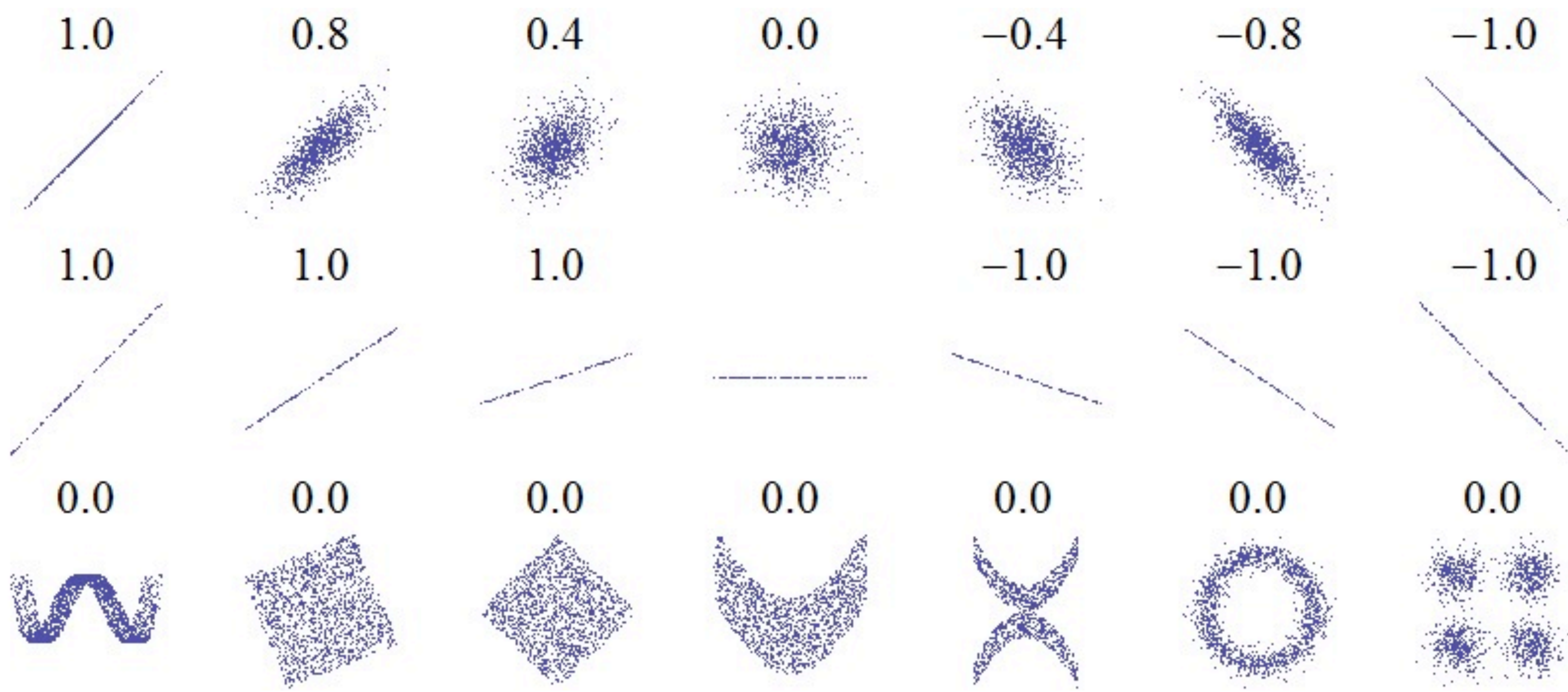
Svojstva:

1. *Corr(aX+b, cY+d)* = *Corr (X,Y)*
2. $-1 \leq \rho \leq 1$
3. Ako su *X* i *Y* nezavisne varijable onda je $\rho = 0$ (obrat ne vrijedi)
4. Ako je *Y* = *aX+b* onda i samo onda je $\rho = 1$ ili $\rho = -1$

Korelacija koja iščezava ne isključuje povezanost veličina, nego samo linearnu povezanost.
Korelacija blizu 1 ne znači nužno da porast veličine X uzrokuje porast veličino Y, nego samo da postoji veza među njima.

Višedimenzionalne slučajne varijable - korelacija

Koeficijenti korelacije:



Ne brkati korelaciju i ovisnost!

Višedimenzionalne slučajne varijable - kontinuirane varijable

Def.: Neka su X i Y dvije kontinuirane slučajne varijable. Funkcija f(x,y) je njihova združena funkcija gustoće vjerojatnosti ako za bilo koji dvodimenzionalni skup A⊆RxR vrijedi:

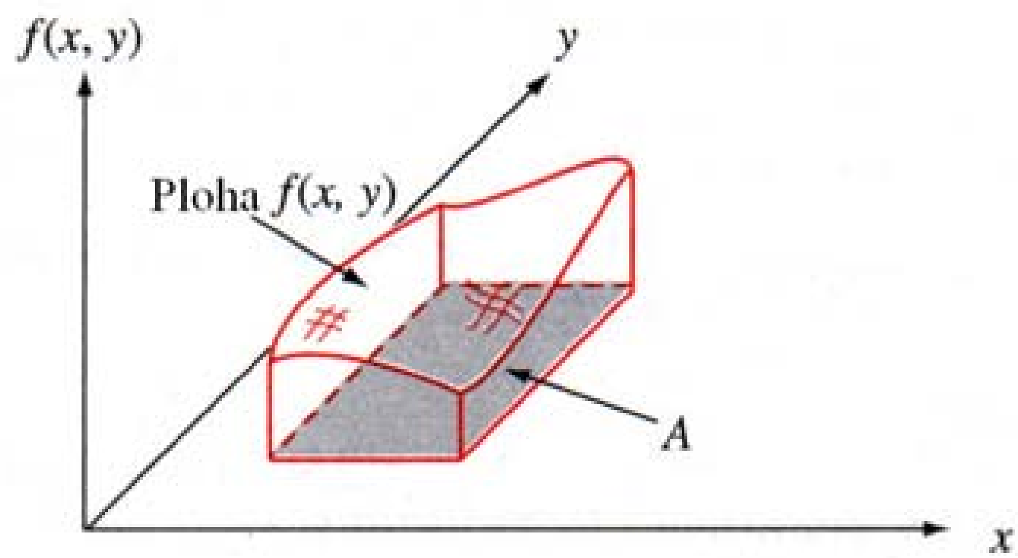
$$P\left(\left(X,Y\right)\in A\right)=\iint_A f\left(x,y\right)\mathrm{d}x\,\mathrm{d}y$$

Takva funkcija mora zadovoljavati uvjete:

$$f\left(x,y\right)\geq 0;\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(x,y\right)\mathrm{d}x\,\mathrm{d}y=1$$

Višedimenzionalne slučajne varijable - kontinuirane varijable

Zamislimo li plohu u 3D koordinatnom sustavu čija visina iznad točke $(x,y) \in A$ iznosi $f(x,y)$, tada je vjerojatnost $P[(x,y) \in A]$ određena volumenom ispod te plohe



Def.: Rubne (marginalne) funkcije gustoće vjerojatnosti za varijable X i Y označavamo $f_X(x)$ i $f_Y(y)$, a dane su izrazima:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \quad i \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

Višedimenzionalne slučajne varijable - kontinuirane varijable

Očekivane vrijednosti:

$$E(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x$$

$$E(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int_{-\infty}^{\infty} y p(y) \, \mathrm{d}x$$

E(x+y) i E(xy), kao i ostala opća svojstva ne ovise o tome je li raspodjela kontinuirana ili diskretna!

Višedimenzionalne slučajne varijable - kontinuirane varijable

$$m_{rs} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^r y^s f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$M_{rs} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu)^r (y-\nu)^s f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$\sigma_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,\mathrm{d}x\,\mathrm{d}y - \mu\nu$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} (y-\nu)^2 p(y)\,\mathrm{d}y$$

Teorija slučajnih pogrešaka

- znanstveni pokus, tj. mjerenje neke fizikalne veličine, uvijek rezultira nizom vrijednosti x_1, x_2, ..., koje se razlikuju od prave vrijednosti X, te međusobno

- pažljivi eksperimentator eliminira sistematske pogreške pa su gornje razlike rezultat slučajnih pogrešaka

- prva ocjena rezultata je računanje:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \quad \sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n}$$

- međutim, ove veličine ne dotiču se X i njene pogreške koje su obje nepoznate veličine

-teorija slučajnih pogrešaka ima za cilj opisati rezultat s obzirom na X na najbolji mogući način, koristeći eksperimentalno dostupne veličine

Gaussov zakon pogrešaka

- ako je *X* prava vrijednost neke veličine, onda se za najvjerojatniju vrijednost te veličine uzima:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

gdje su x_i mjerenja te veličine

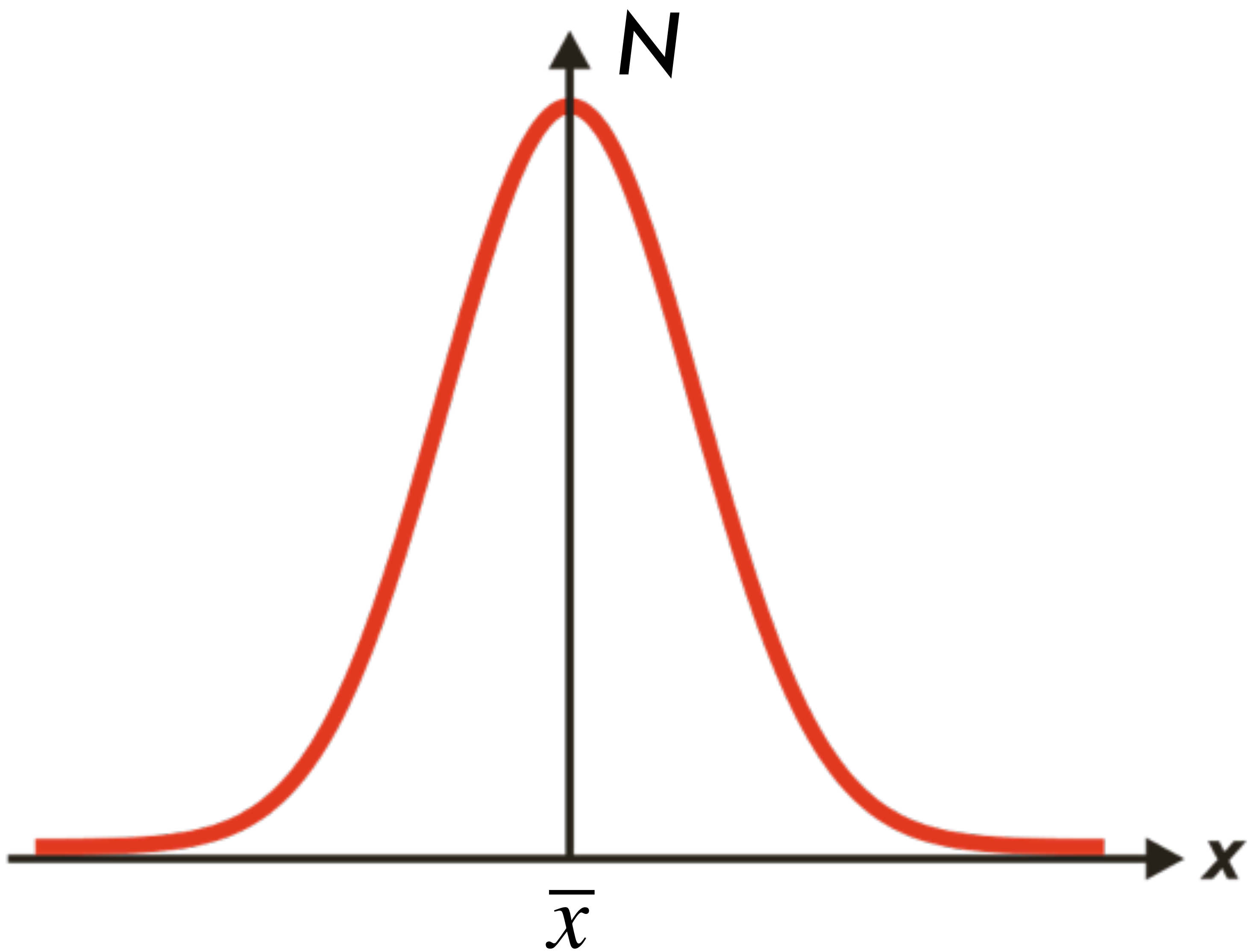
- neka je $v_i = \overline{x} - x_i$ onda je vjerojatnost za tu grešku jednaka:

$$\varphi(v)\mathrm{d}v = \frac{h}{\sqrt{\pi}} e^{-h^2 v^2} \mathrm{d}v$$

Gaussov zakon pogreške

- značenje parametra *h* odredit ćemo kasnije

Gaussov zakon pogrešaka



Značenje: izmjereni rezultati distribuiraju se približno po Gaussovoj raspodjeli oko $\overline{x}$

Zakon rasprostranjenja pogrešaka

Neka je $\varepsilon_i = X - x_i$ → ovo ne znamo jer ne znamo *X*!

Tada je srednja pogreška dana s:

$$m^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}$$

(**primijetite**: isto kao σ^2 samo $X \leftrightarrow \bar{x}$)

Neka je $X = X_1 + X_2$, gdje X_1 ima pogrešku m_1, a X_2 pogrešku m_2. Koliki je m za X?

$$m_1^2 = \frac{\sum_{i=1}^{n_1} \varepsilon_{i1}^2}{n_1}, \quad m_2^2 = \frac{\sum_{j=1}^{n_2} \varepsilon_{j2}^2}{n_2}$$

Zakon rasprostranjenja pogrešaka

Pogreška kod $X = X_1 + X_2$ mogu nastati svim zbrajanjem ε_{i1} i ε_{j2}, tj.

$$m^2 = \frac{\sum\limits_{i,j}\left(\varepsilon_{i1} + \varepsilon_{j2}\right)^2}{n_1 n_2} = \frac{\sum\limits_{i,j}\left(\varepsilon_{i1}^2 + \varepsilon_{j2}^2 + \overbrace{2\varepsilon_{i1}\varepsilon_{j2}}^{0}\right)}{n_1 n_2} = \frac{\overbrace{\sum\limits_{j=1}^{n_2}}^{n_2}\sum\limits_{i=1}^{n_1}\varepsilon_{i1}^2}{n_1 n_2} + \frac{\overbrace{\sum\limits_{i=1}^{n_1}}^{n_1}\sum\limits_{j=1}^{n_2}\varepsilon_{j1}^2}{n_1 n_2} =$$

$$= \frac{\sum\limits_{i=1}^{n_1}\varepsilon_{i1}^2}{n_1} + \frac{\sum\limits_{j=1}^{n_2}\varepsilon_{j2}^2}{n_2} = m_1^2 + m_2^2$$

Općenitije, vrijedi zakon rasprostranjenja pogrešaka!

Ako je: $X = X_1 \pm X_2 \pm X_3 \pm \ldots \pm X_n$ tada je:

$$m^2 = m_1^2 + m_2^2 + m_3^2 + \ldots + m_n^2$$

Napomena: *m* se naziva i 'pogreška jednog mjerenja'

Srednja pogreška, nepouzdanost

- imamo niz mjerenja, x_1, x_2, ..., x_n, te izračunamo očekivanu vrijednost

$$\bar{x} = \frac{\sum_i x_i}{n}$$

- **pitanje: kako interpretirati taj rezultat?**
- pretpostavka: vrijedi Gaussov zakon pogreške

- prije smo uveli $\varepsilon_i = X - x_i$
- sada uvodimo $v_i = \bar{x} - x_i$ $\longrightarrow$ TO ZNAMO!

- neka $\bar{x}$ odstupa od X za M, veličinom koju nazivamo **NEPOUZDANOST**

$$X = \bar{x} \pm M$$

$$\boxed{\varepsilon_i = X - x_i = \bar{x} \pm M - x_i = v_i \pm M}$$

Srednja pogreška, nepouzdanost

Srednja pogreška je:

$$m^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n} = \frac{\sum_{i=1}^{n} (v_i \pm M)^2}{n} = \frac{\sum_{i=1}^{n} v_i^2}{n} \pm 2M \frac{\overbrace{\sum_{i=1}^{n} v_i}^{0}}{n} + M^2 \frac{\sum_{i=1}^{n} 1}{n} = \sigma^2 + M^2$$

Pogledajmo kolika je POGREŠKA ARITMETIČKE SREDINE:

$$\bar{x} = \frac{x_1}{n} + \frac{x_2}{n} + \ldots + \frac{x_n}{n}$$

x_1 ima pogrešku *m*, x_2 ima pogrešku *m*, itd., pa svaki član gornje sume ima pogrešku *m/n*

Srednja pogreška, nepouzdanost

Prema zakonu o rasprostranjivanju pogrešaka:

$$m_{\bar{x}}^2 = \left(\frac{m}{n}\right)^2 + \left(\frac{m}{n}\right)^2 + \ldots + \left(\frac{m}{n}\right)^2 = n\left(\frac{m}{n}\right)^2 = \frac{m^2}{n}$$

- hipoteza: $m_{\bar{x}}$ ne razlikuje se bitno od *M*
- iako se ova pretpostavka čini vrlo *ad hoc*, u praksi se pokazuje zadovoljavajućom

$$m^2 = \sigma^2 + M^2 \approx \sigma^2 + \frac{m^2}{n}$$

$$m\sqrt{\frac{n-1}{n}} \approx \sigma$$

Srednja pogreška, nepouzdanost

STANDARDNA DEVIJACIJA

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

SREDNJA POGREŠKA ili POGREŠKA JEDNOG MJERENJA

$$m = \sqrt{\frac{n}{n-1}}\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

NEPOUZDANOST ili POGREŠKA ARITMETIČKE SREDINE

$$M = \frac{m}{\sqrt{n}} = \frac{\sigma}{\sqrt{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n\left(n-1\right)}}$$

Važno: $M = \frac{m}{\sqrt{n}}$

Srednja pogreška, nepouzdanost

- iskoristimo ove rezultate da nađemo *h* iz Gaussovog zakona pogrešaka
- za $n \to \infty$, $m \to \sigma$ ali uzimamo *m* (koja sadrži *X*) a ne σ kao relevantnu veličinu

$$m^2 \underset{n>>1}{\approx} \frac{\sum_{i=1}^{n} v_i^2}{n} \quad \to \quad \int_{-\infty}^{\infty} v^2 \varphi(v)\,\mathrm{d}v = \frac{1}{2h^2}$$

pa je stoga:

$$\boxed{h = \frac{1}{m\sqrt{2}}}$$

Srednja pogreška, nepouzdanost

Važno! Rezultat se uvijek izražava kao:

$$X = \bar{x} \pm M$$

- tako izražen rezultat kazuje da je *X* u intervalu $\left(\bar{x} - M, \bar{x} + M\right)$ s vjerojatnošću 68%, u intervalu $\left(\bar{x} - 2M, \bar{x} + 2M\right)$ s vjerojatnošću 95%, u intervalu $\left(\bar{x} - 3M, \bar{x} + 3M\right)$ s vjerojatnošću 99%.
- mjera "dobrote mjerenja" je relativna nepouzdanost

$R = \frac{M}{\bar{x}}\left(\cdot 100\%\right)$ i izražava se u postocima

- budući da je $M \propto \frac{1}{\sqrt{n(n-1)}}$ preporučljivo je uvijek načiniti ne manje od *n* = 10 mjerenja, ali ni puno više, jer je ovisnost *M* o *n za* n < *10* jaka, a za *n* > 10 slaba.

Srednja pogreška, nepouzdanost

- M se uvijek zaokružuje na jednu znamenku, iznimno na dvije (ako je zadnja znamenka 5). $\bar{x}$ se zaokružuje prema M

Primjer:

$$M = 0.0037624 \text{ cm}$$

$$\bar{x} = 0.0163274 \text{ cm}$$

$$M \rightarrow 0.004 \text{ cm}, \qquad \bar{x} \rightarrow 0.016 \text{ cm} \qquad R = \frac{0.004}{0.016} \cdot 100$$

3. znamenka 3. znamenka

Pišemo: x = (0.016±0.004) cm, R = 25% ili
$x = (1.6 \pm 0.4) \cdot 10^{-2}$ cm, R = 25%

Srednja pogreška, nepouzdanost

Primjer:

$$M = 0.153264 \text{ kg}$$
$$\bar{x} = 7.25434 \text{ kg}$$

$$M \rightarrow 0.15 \text{ kg}, \qquad \bar{x} \rightarrow 7.26 \text{kg}$$
$$\text{x=}\left(7.26 \pm 0.15\right)\text{kg}, \quad R = 2\%$$