Momenty – případ dvou proměnných

operátor očekávané hodnoty : $E[\ \ ]$

očekávaná hodnota náhodné proměnné x : $\mu_x \equiv E[x] = \iint x\ f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

očekávaná hodnota náhodné proměnné *y* : $\mu_y \equiv E[y] = \iint y\ f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

obecně : $E[g(x,y)] = \iint g(x,y)\ f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

Momenty – případ dvou proměnných

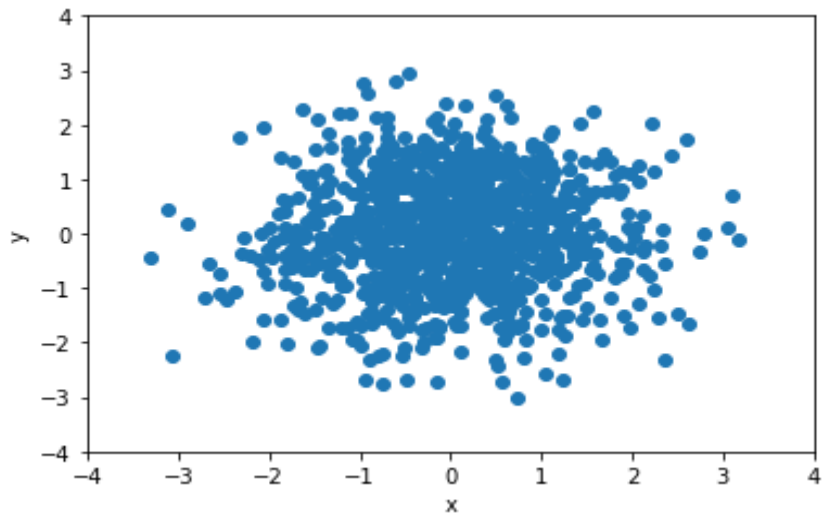
rozptyl náhodné proměnné x : $\sigma_x^2 \equiv V[x] = E\left[(x-\mu_x)^2\right] = \iint (x-\mu_x)^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

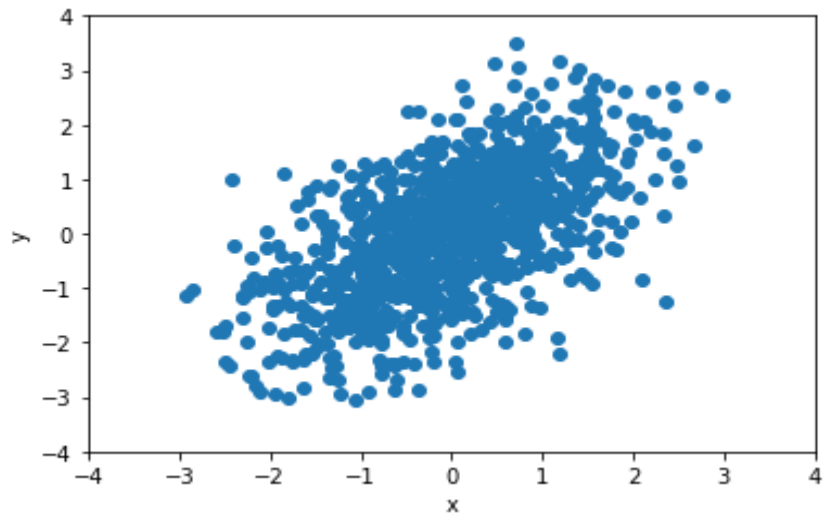
rozptyl náhodné proměnné y : $\sigma_y^2 \equiv V[y] = E\left[(y-\mu_y)^2\right] = \iint (y-\mu_y)^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

minimální smysluplná informace o náhodných proměnných x, y : μ_x, μ_y (míra polohy)

σ_x, σ_y (míra disperze)

4 čísla nestačí





Momenty – případ dvou proměnných

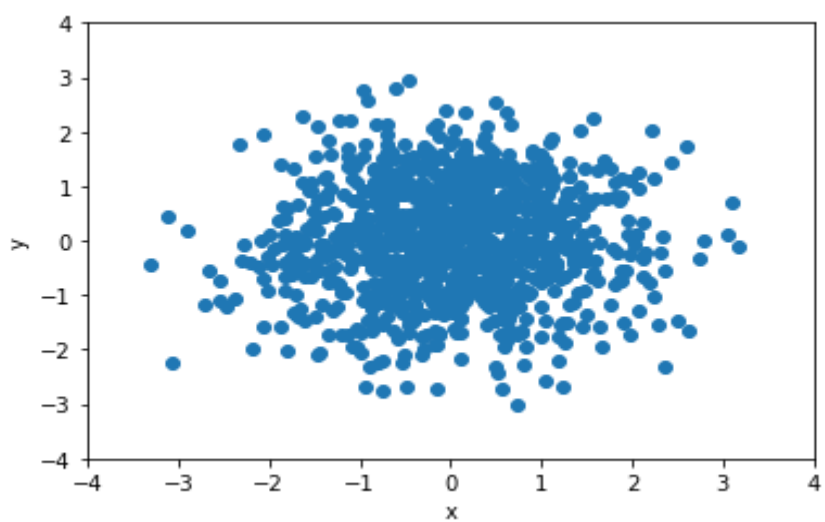
rozptyl náhodné proměnné x : $\sigma_x^2 \equiv V[x] = E\left[(x-\mu_x)^2\right] = \iint (x-\mu_x)^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

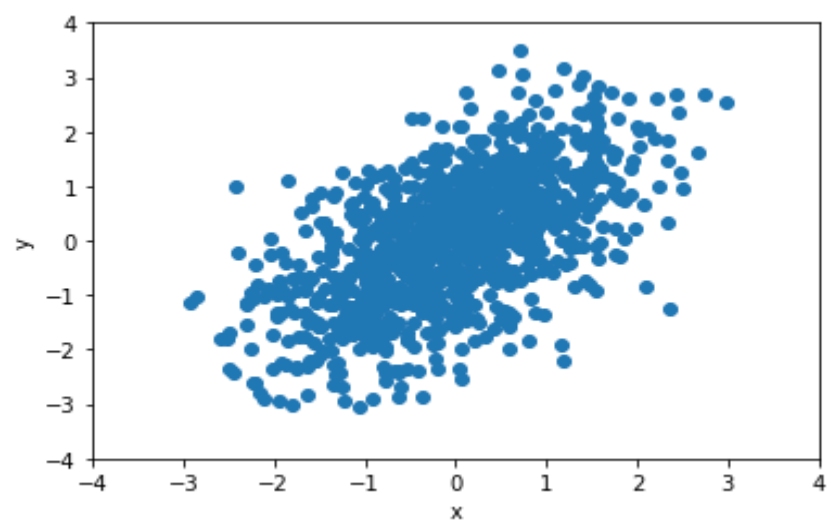
rozptyl náhodné proměnné y : $\sigma_y^2 \equiv V[y] = E\left[(y-\mu_y)^2\right] = \iint (y-\mu_y)^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

minimální smysluplná informace o náhodných proměnných x, y : μ_x, μ_y (míra polohy)

σ_x, σ_y (míra disperze)

kovariance cov (x,y)





Momenty – případ dvou proměnných

kovariance náhodných proměnných x a y :

$$\mathrm{cov}(x,y) = E\left[(x-\mu_x)(y-\mu_y)\right] = E[xy] - E[x]E[y]$$

$\mathrm{cov}(x,x) = E\,[(x-\mu_x)^2] = \sigma_x^2$

$\mathrm{cov}(y,y) = E\,[(y-\mu_y)^2] = \sigma_y^2$

kovarianční matice $$V = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_y^2 \end{pmatrix}$$

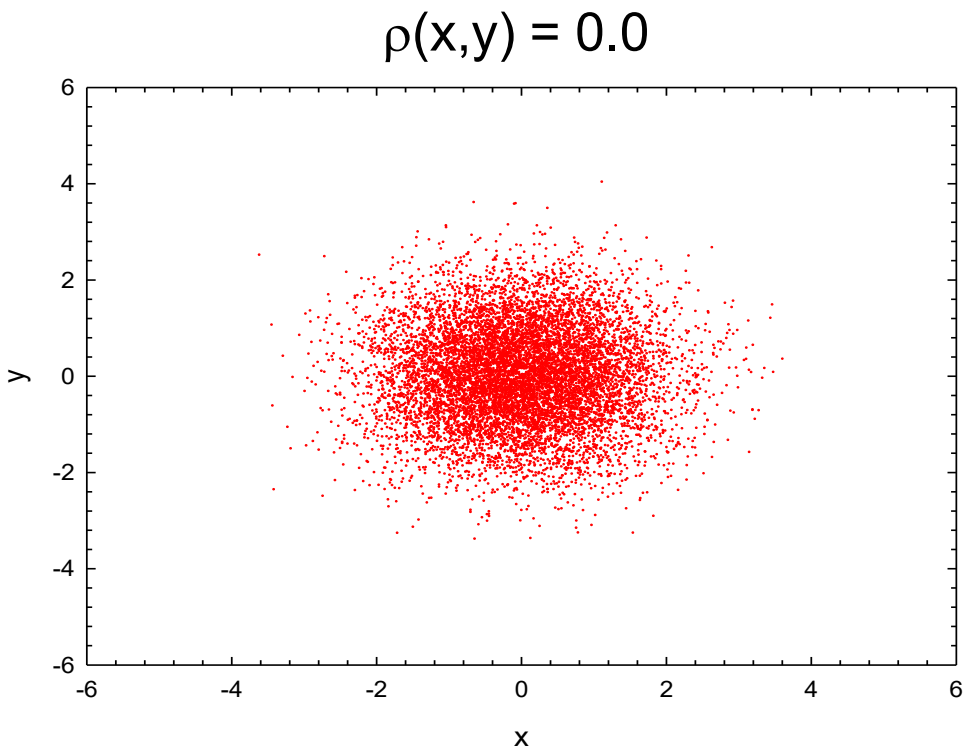
korelace náhodných proměnných x a y : $$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x\,\sigma_y}$$

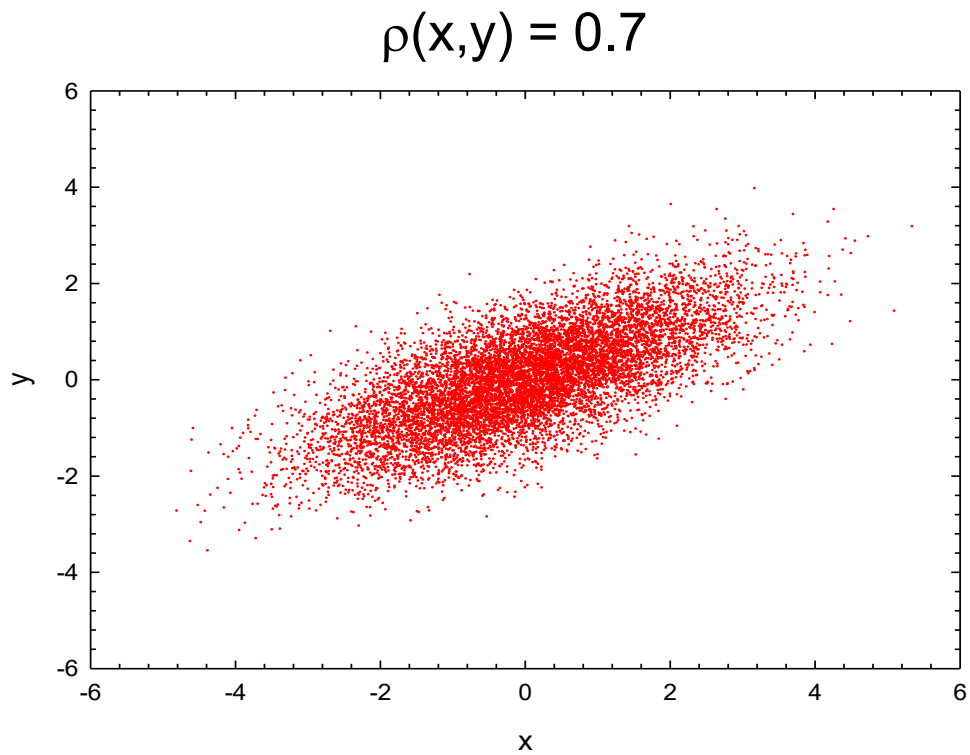
$\rho(x,x) = 1$

$\rho(y,y) = 1$

korelační matice $$R = \begin{pmatrix} \varrho(x,x) & \varrho(x,y) \\ \varrho(y,x) & \varrho(y,y) \end{pmatrix} = \begin{pmatrix} 1 & \varrho(x,y) \\ \varrho(x,y) & 1 \end{pmatrix}$$

Korelace náhodných proměnných

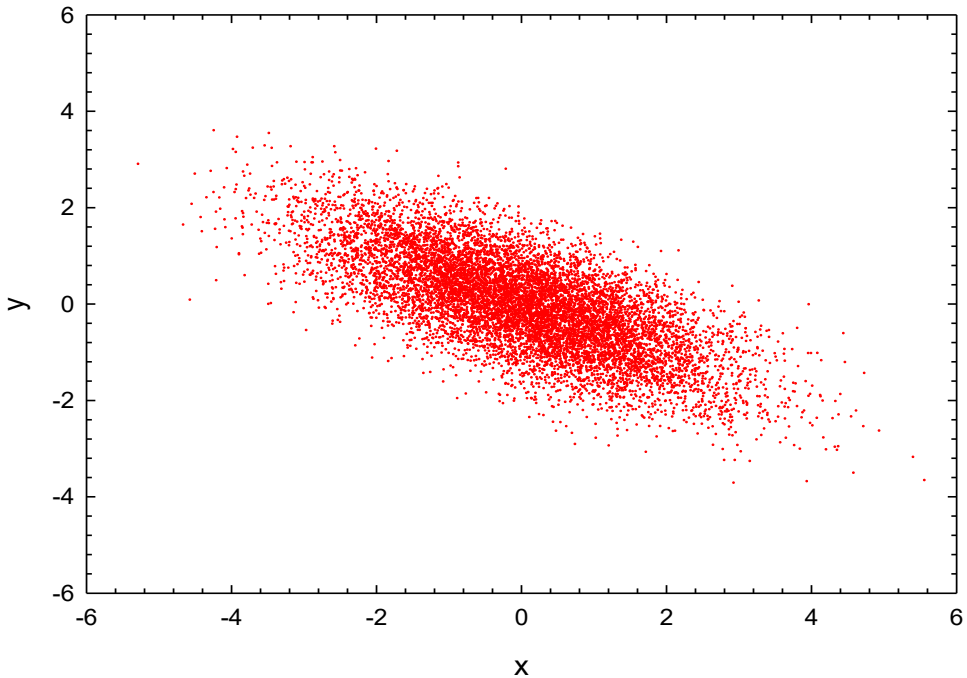




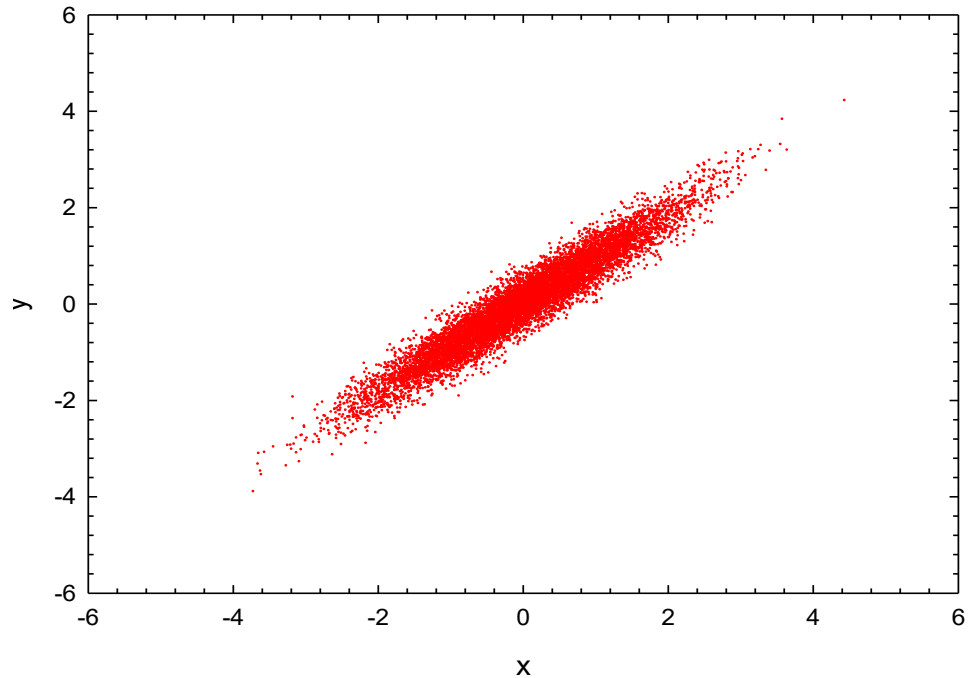
N = 10000

Korelace náhodných proměnných

$\rho(x,y) = -0.7$



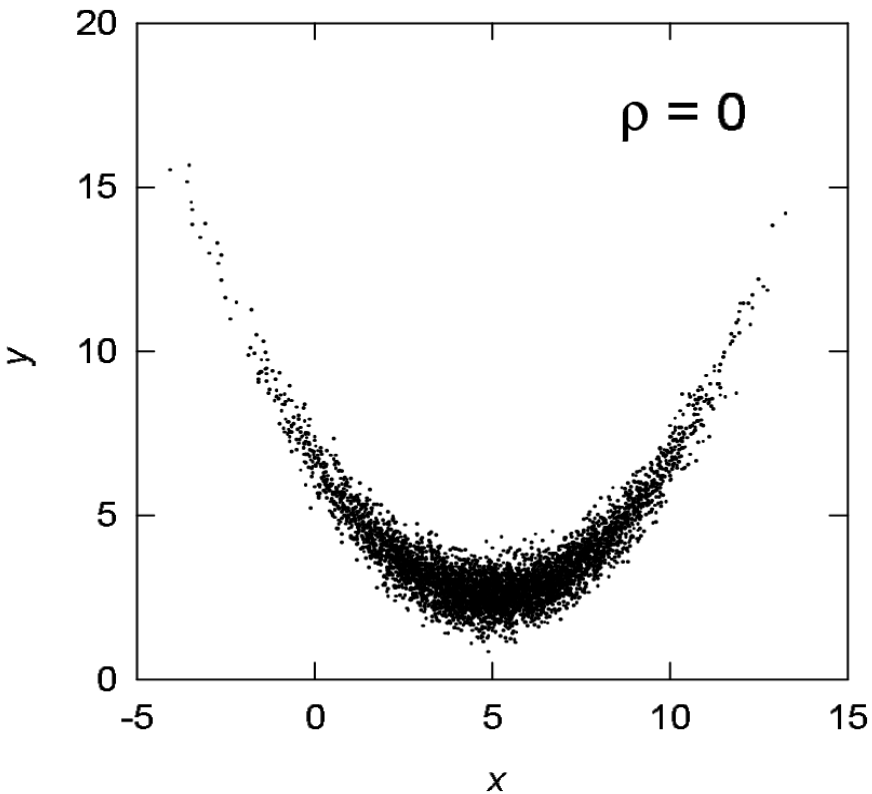
$\rho(x,y) = 0.96$

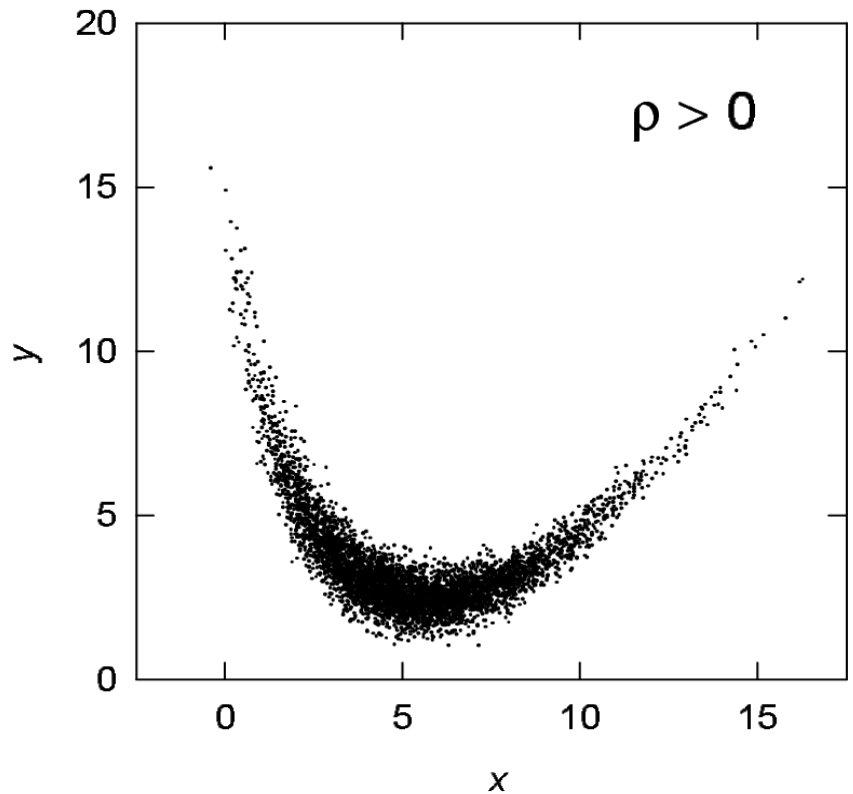


N = 10000

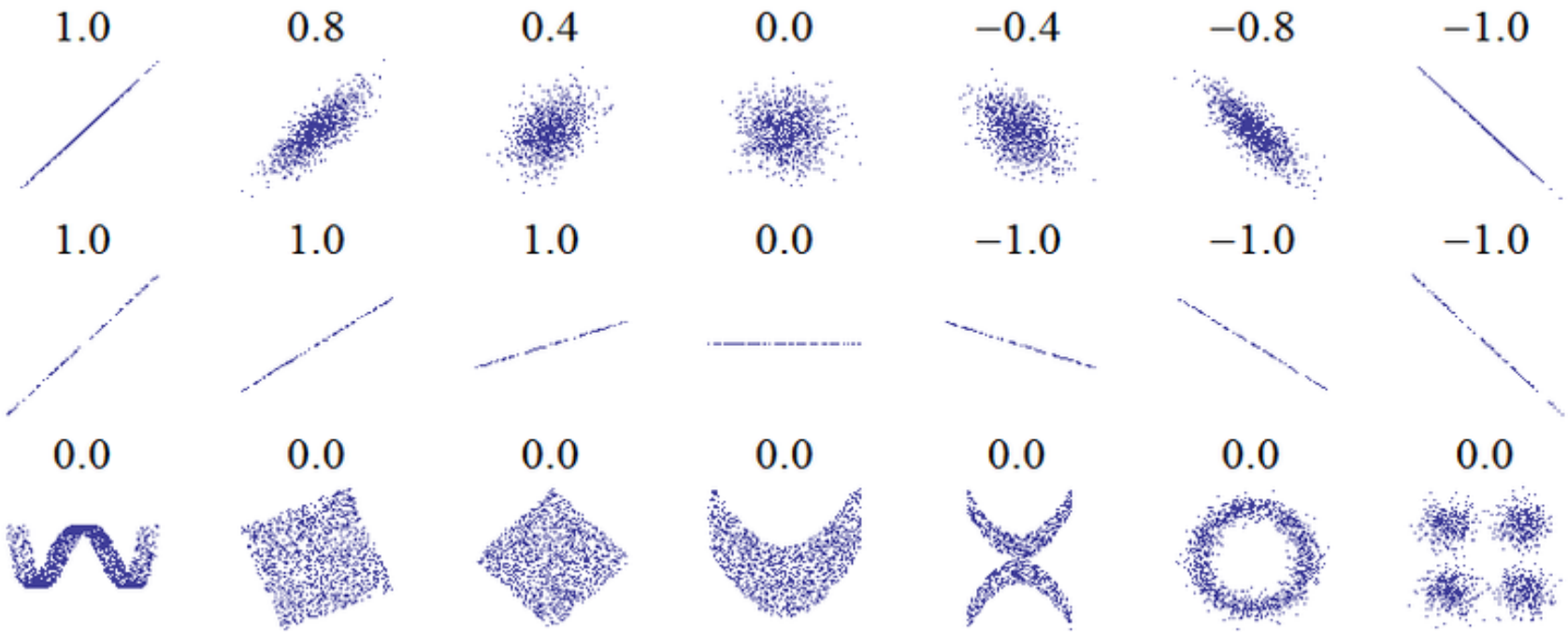
Korelace náhodných proměnných

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-a)^2}{2\sigma_x^2}\right) \exp\left[-\frac{\left(y-b-c(x-a)^2\right)^2}{2\sigma_y^2}\right]$$





Korelace náhodných proměnných



Momenty – případ dvou proměnných

nezávislé náhodné proměnné x,y : $$f(x,y) = f_x(x) f_y(y)$$

rozptyl náhodné proměnné x :

$$V[x] = E\left[(x-\mu_x)^2\right] = \int_{-\infty}^{\infty} (x-\mu_x)^2 f_x(x)\mathrm{d}x \int_{-\infty}^{\infty} f_y(y)\mathrm{d}y = \int_{-\infty}^{\infty} (x-\mu_x)^2 f_x(x)\mathrm{d}x$$

rozptyl náhodné proměnné y :

$$V[y] = E\left[(y-\mu_y)^2\right] = \int_{-\infty}^{\infty} (y-\mu_y)^2 f_y(y)\mathrm{d}y$$

$$\mathrm{cov}(x,y) = 0$$

kovarianční matice: $$\begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

Odhad kovariance a korelace

náhodné proměnné x, y

- naměříme $x_1, x_2, \ldots x_N$; $y_1, y_2, \ldots y_N$
- cov (x, y) ???

$$\mathrm{cov}(x,y) = E[xy] - E[x]E[y]$$

$$\boxed{\hat{\mathrm{cov}}(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle}$$

$$\langle xy \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i y_i$$

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\langle y \rangle = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$\hat{\varrho}(x,y) = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sigma_x \sigma_y}$$

$$\boxed{\hat{\varrho}(x,y) = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\hat{\sigma}_x \hat{\sigma}_y} \quad \sigma_\varrho \approx \frac{1-\varrho^2}{\sqrt{N-1}}}$$

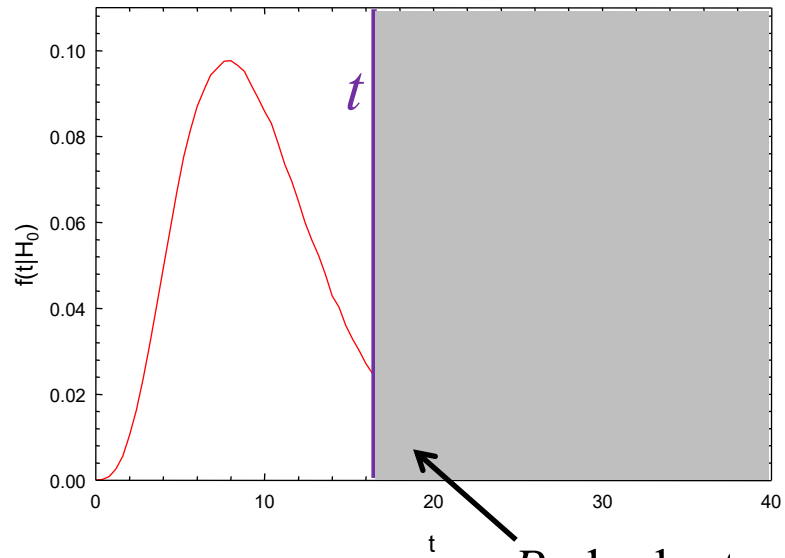
$$\hat{\sigma}_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2}$$

$$\hat{\sigma}_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \langle y \rangle)^2}$$

Test korelace – Fisherova transformace

náhodné proměnné x, y

- naměříme $x_1, x_2, \ldots x_N$; $y_1, y_2, \ldots y_N$ → dostaneme odhad korelace $\hat{\varrho}$
- je korelace statisticky významná?
- nulová hypotéza H_0 (zde nulová korelace)
- testovací statistika t
- známé rozdělení $f(t|H_0)$
- Fisherova transformace $F(\varrho) = \frac{1}{2} \ln \frac{1+\varrho}{1-\varrho}$



Pokud platí nulová hypotéza, tak má normální rozdělení s očekávanou hodnotou 0 a standardní odchylkou $\sigma = \frac{1}{\sqrt{N-3}}$

P - hodnota pokud je $P < P_\alpha$ odmítneme H_0

- z - proměnná $z = \frac{F(\varrho)}{\sigma}$

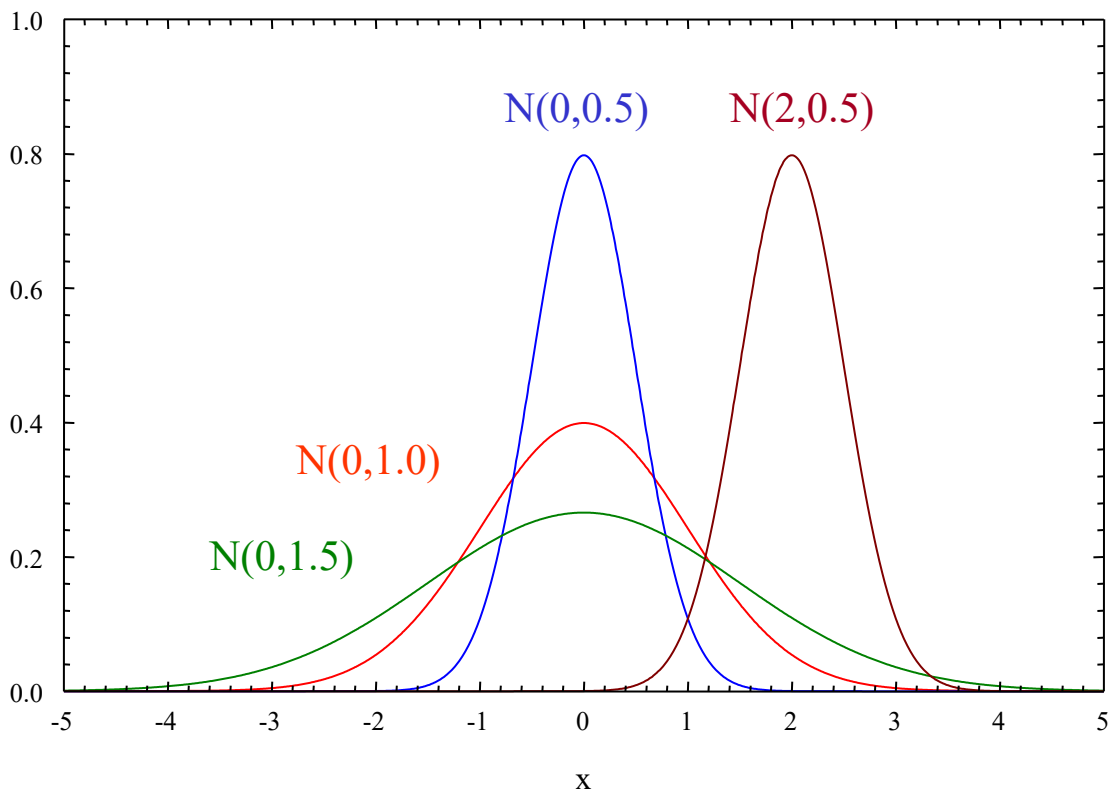
P_α hladina signifikance, buď 5% nebo 1%

- pokud platí nulová hypotéza $z \in N(0,1)$

$$P = 2\left(1 - F_{0,1}(z)\right) = 2\left(1 - \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right]\right)$$

Normální Gaussovo rozdělení

hustota pravděpodobnosti $f(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

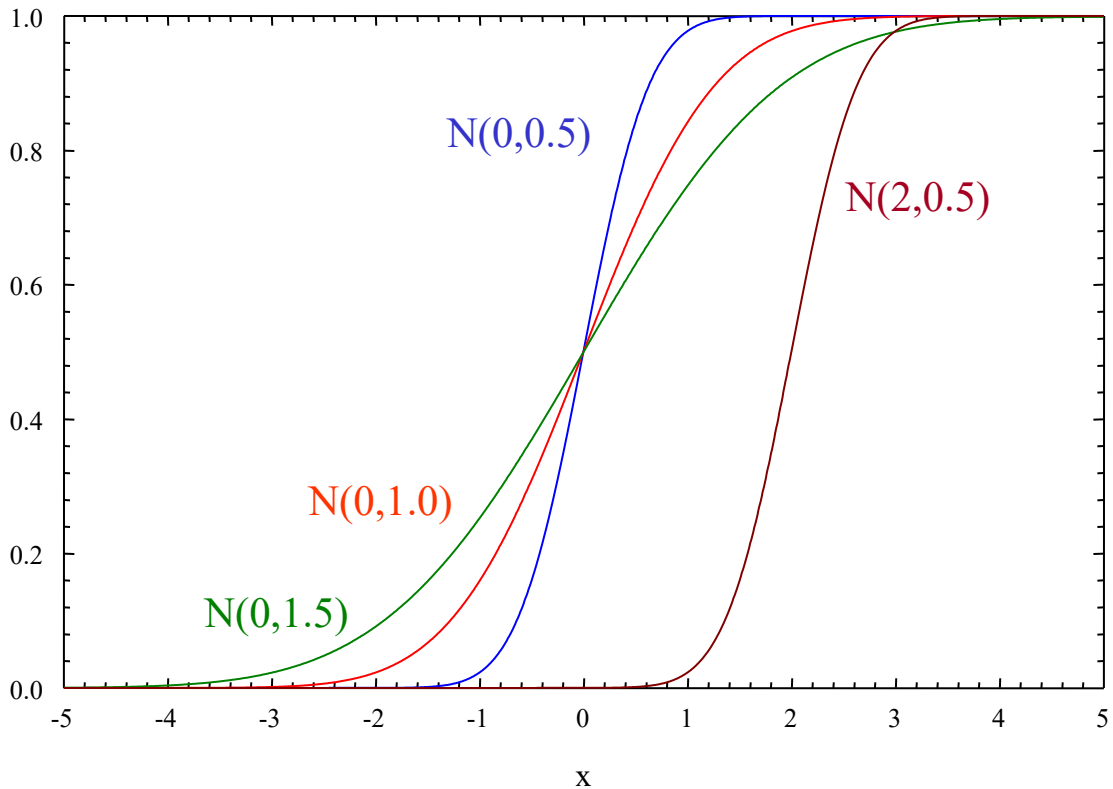


Normální Gaussovo rozdělení

Distribuční funkce $F_{\mu,\sigma}(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = F_{0,1}\left(\frac{x-\mu}{\sigma}\right)$

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

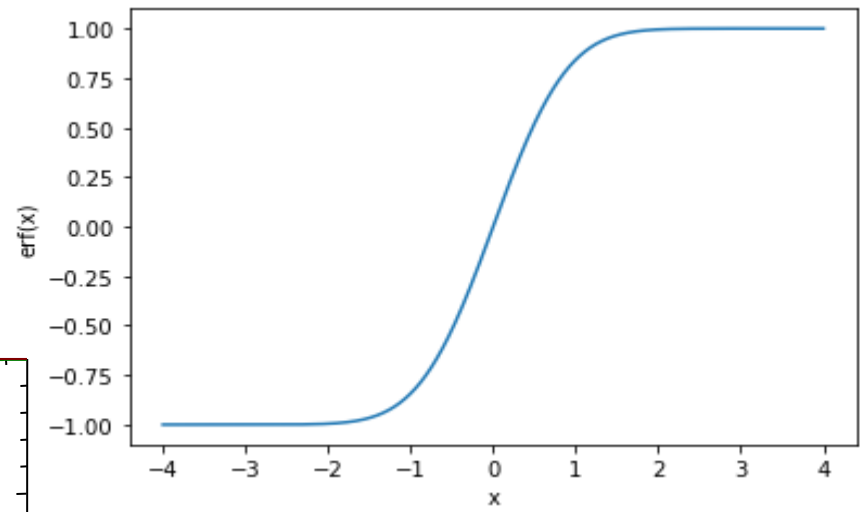
- error funkce: $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$

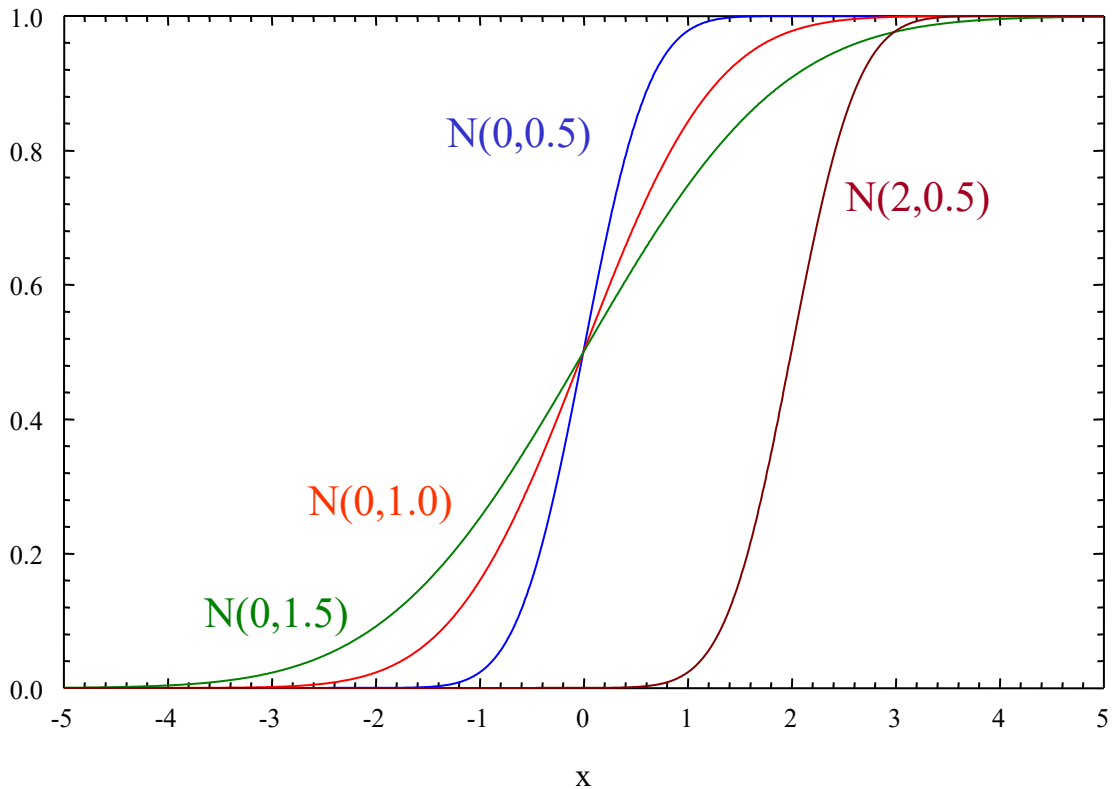


Normální Gaussovo rozdělení

Distribuční funkce $$F_{\mu,\sigma}(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = F_{0,1}\left(\frac{x-\mu}{\sigma}\right)$$

- error funkce: $$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\,dt$$





výpočet error funkce:

např.
Excel: `erf(x)`
ROOT: `ROOT::Math::erf(x)`
Matlab: `erf(x)`
Gnuplot: `erf(x)`
Python:

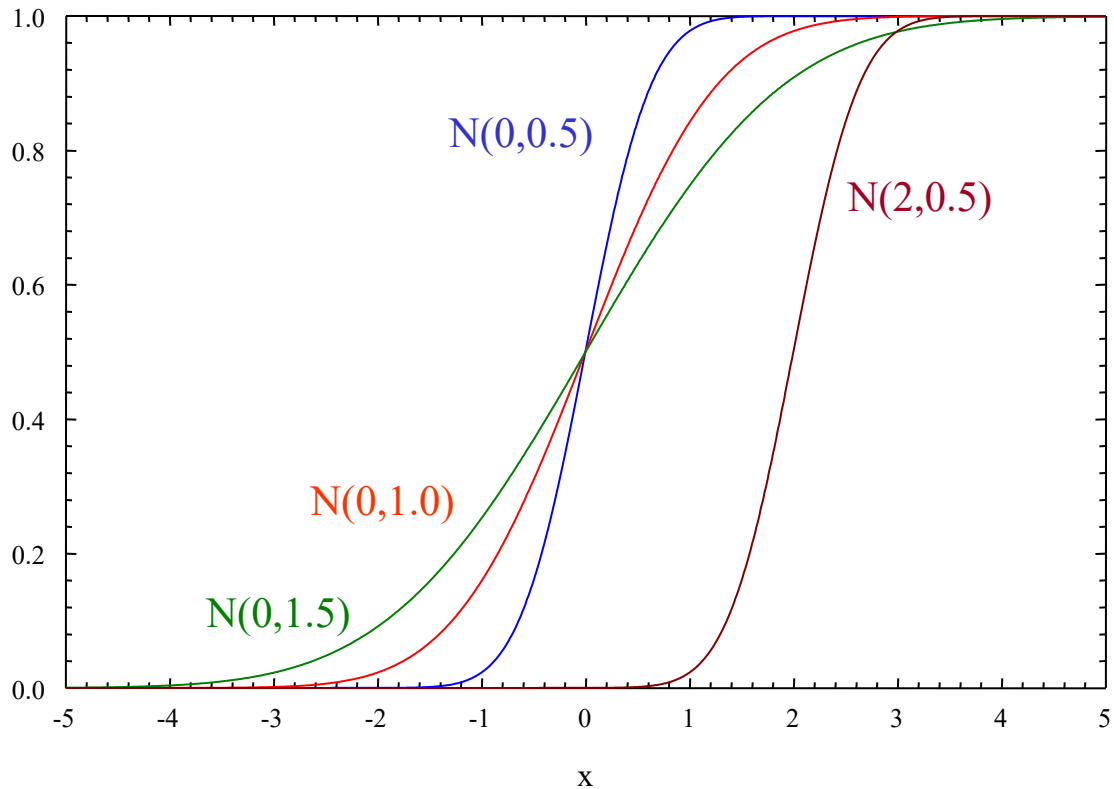
```
from scipy import special
special.erf(x)
```

Normální Gaussovo rozdělení

Distribuční funkce $F_{\mu,\sigma}(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = F_{0,1}\left(\frac{x-\mu}{\sigma}\right)$

- error funkce: $erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\,dt$ $\longrightarrow$ $F_{0,1}(x) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right)$

$$F_{\mu,\sigma}(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right)$$



výpočet error funkce:

např.
Excel: `erf(x)`
ROOT: `ROOT::Math::erf(x)`
Matlab: `erf(x)`
Gnuplot: `erf(x)`
Python:

```
from scipy import special
special.erf(x)
```

Test korelace – studentovo rozdělení

náhodné proměnné x, y

- naměříme $x_1, x_2, \ldots x_N$; $y_1, y_2, \ldots y_N$ $\rightarrow$ dostaneme odhad korelace $\hat{\varrho}$
- je korelace statisticky významná?
- nulová hypotéza H_0 (zde nulová korelace)
- testovací statistika $t = \varrho \sqrt{\frac{N-2}{1-\varrho^2}}$
- pokud platí nulová hypotéza $t \in T(x|N-2)$

Studentovo rozdělení s počtem stupňů volnosti $\nu = N - 2$

Studentovo *t* rozdělení

William Sealy Gosset

studentovo rozdělení s ν stupni volnosti

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

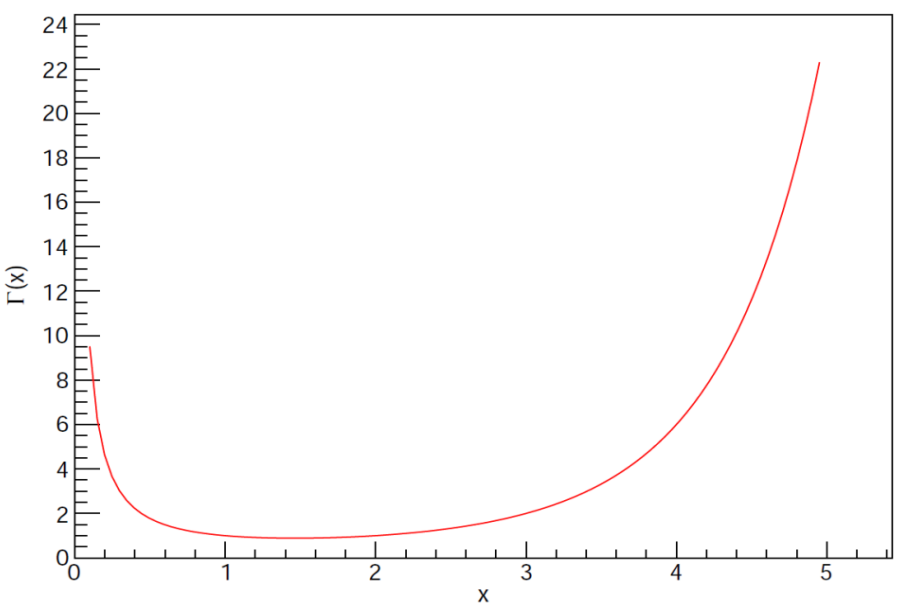
ν – počet stupňů volnosti

Gama funkce $\Gamma(x) = \int_0^\infty t^{x-1}\,e^{-t}\,dt \qquad x \geq 0$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma(n) = (n-1)! \qquad n \in \mathcal{N}$$

$$\Gamma(x+1) = x\Gamma(x) \qquad x \in \mathcal{R}$$



Python:

```
from scipy.special import gamma
```

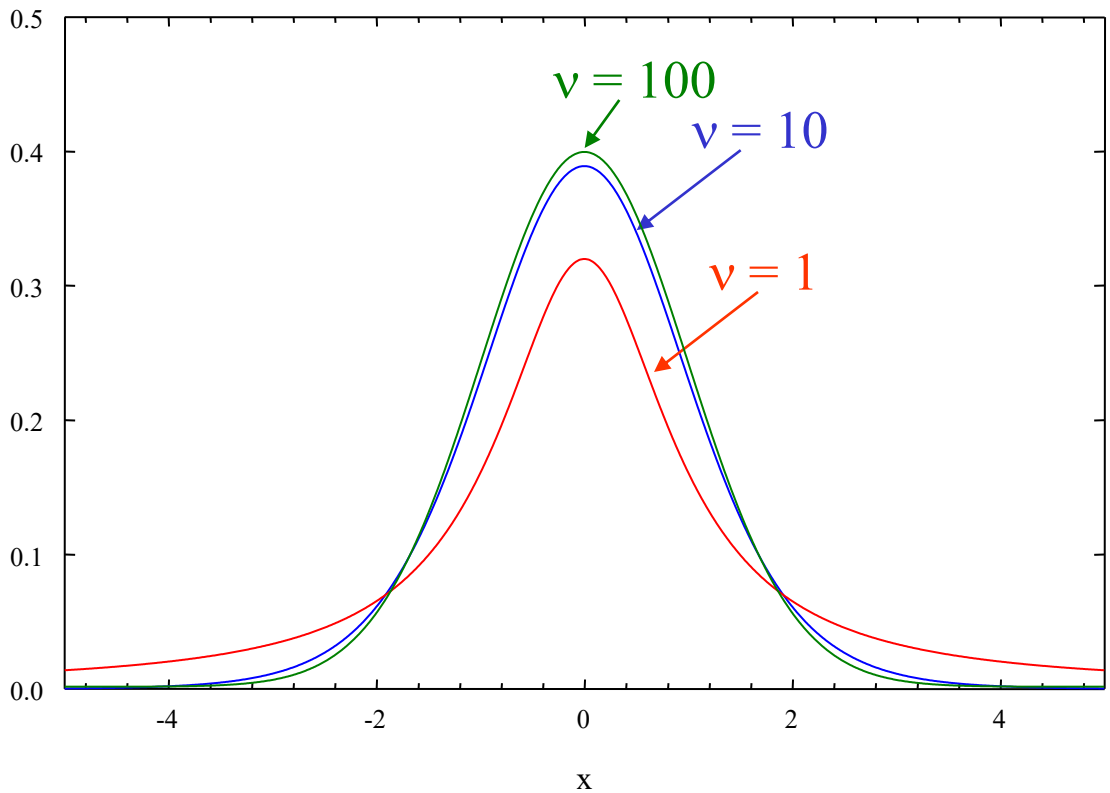
Studentovo *t* rozdělení

Python:

```
from scipy.stats import t
t.pdf(x,ν)
```

studentovo rozdělení s ν stupni volnosti

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



ν – počet stupňů volnosti

$$f(x|\nu=1) = \frac{1}{\pi}\frac{1}{1+x^2}$$

$$\nu \to \infty \quad f(x|\nu) \to N(0,1)$$

$$E[x] = \mu = 0$$

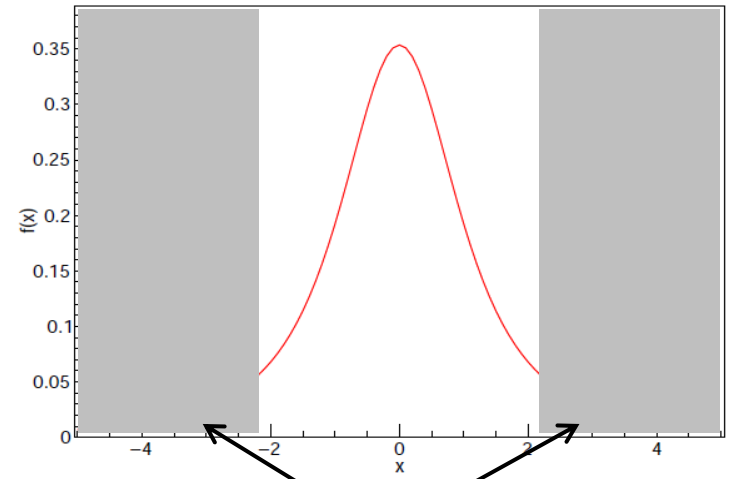
$$V[x] = \sigma^2 = \frac{\nu}{\nu-2} \quad \nu > 2$$

Test korelace – studentovo rozdělení

náhodné proměnné x, y

- naměříme $x_1, x_2, \ldots x_N$; $y_1, y_2, \ldots y_N$ → dostaneme odhad korelace $\hat{\varrho}$
- je korelace statisticky významná?
- nulová hypotéza H_0 (zde nulová korelace)
- testovací statistika $t = \varrho \sqrt{\frac{N-2}{1-\varrho^2}}$
- pokud platí nulová hypotéza $t \in T(N-2)$

Studentovo rozdělení s počtem stupňů volnosti $\nu = N - 2$



P_α

oboustranná P-hodnota

- P-hodnota: $P = 2(1 - T_\nu(t))$

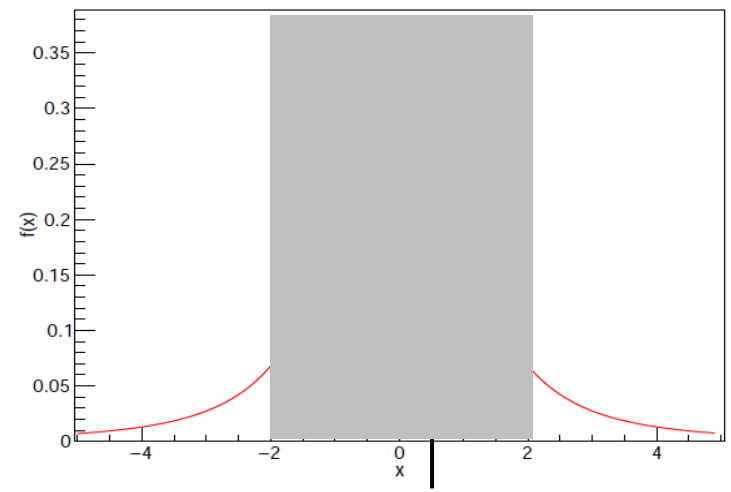
distribuční funkce studentova rozdělení

Test korelace – studentovo rozdělení

náhodné proměnné x, y

- naměříme $x_1, x_2, \ldots x_N$; $y_1, y_2, \ldots y_N$ → dostaneme odhad korelace $\hat{\varrho}$
- je korelace statisticky významná?
- nulová hypotéza H_0 (zde nulová korelace)
- testovací statistika $t = \varrho \sqrt{\frac{N-2}{1-\varrho^2}}$
- pokud platí nulová hypotéza $t \in T(N-2)$

Studentovo rozdělení s počtem stupňů volnosti $\nu = N - 2$



$1\text{-}P_\alpha$
konfidenční interval

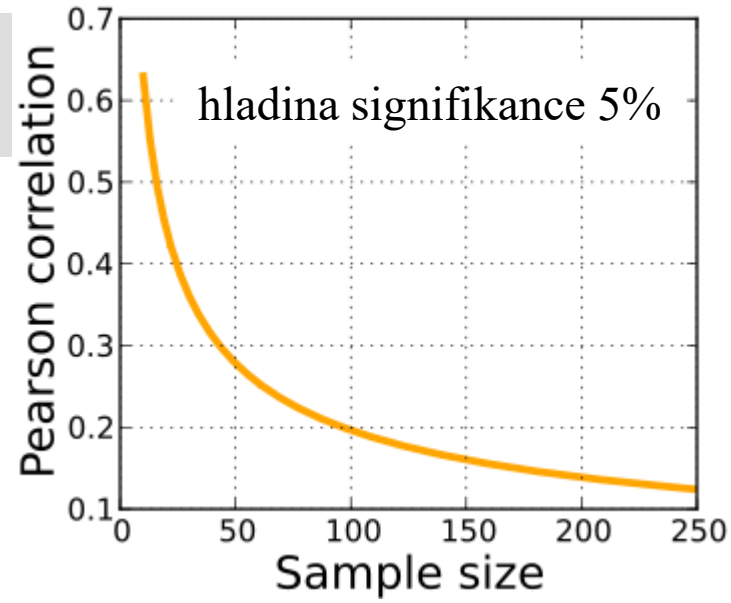
- P-hodnota: $P = 2(1 - T_\nu(t))$

distribuční funkce studentova rozdělení

- konfidenční interval: $\left(-T_\nu^{-1}\left(1-\frac{P}{2}\right), T_\nu^{-1}\left(1-\frac{P}{2}\right)\right)$

inverzní funkce k distribuční funkci studentova rozdělení

Odhad kovariance a korelace



- náhodné proměnné x, y
- naměříme $x_1, x_2, \ldots x_N$; $y_1, y_2, \ldots y_N$

$$\boxed{\hat{\varrho}(x,y) = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\hat{\sigma}_x \hat{\sigma}_y}}$$

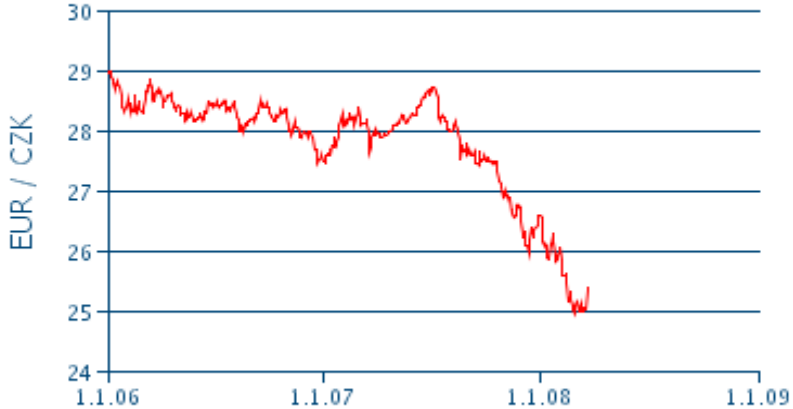
Za předpokladu, že korelace náhodných proměnných x,y je nula, má Pearsonův korelační koeficient rozdělení s hustotou pravděpodobnosti

$$f(\varrho|N) = \frac{(1-\varrho^2)^{\frac{N-4}{2}}}{B\left(\frac{1}{2}, \frac{N-2}{2}\right)}$$

kde $B(x,y)$ je Beta funkce $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$ $\qquad B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

P-hodnota: $P = 2F(-|\hat{\varrho}|, N) = 2\int_{-1}^{-|\hat{\varrho}|} f(r|N)\,dr$

Odhad korelace



$$\hat{\mathrm{cov}}(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

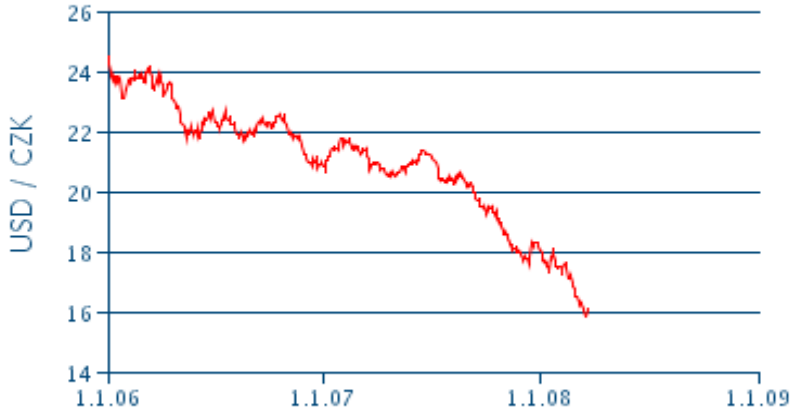
$$\hat{\varrho}(x,y) = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\hat{\sigma}_x \hat{\sigma}_y}$$

$\rho = 0.920 \pm 0.003$

$$\langle xy \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i y_i$$

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \hat{\sigma}_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2}$$

$$\langle y \rangle = \frac{1}{N} \sum_{i=1}^{N} y_i \qquad \hat{\sigma}_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \langle y \rangle)^2}$$



leden 2006 – březen 2008

Pearsonův korelační koeficient

- standartizovaná z – proměnná

$$z = \frac{x-\mu}{\sigma}$$

odhad $\rightarrow$ $$\hat{z} = \frac{x-\hat{\mu}}{\hat{\sigma}}$$

$$\hat{\mu} \equiv \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

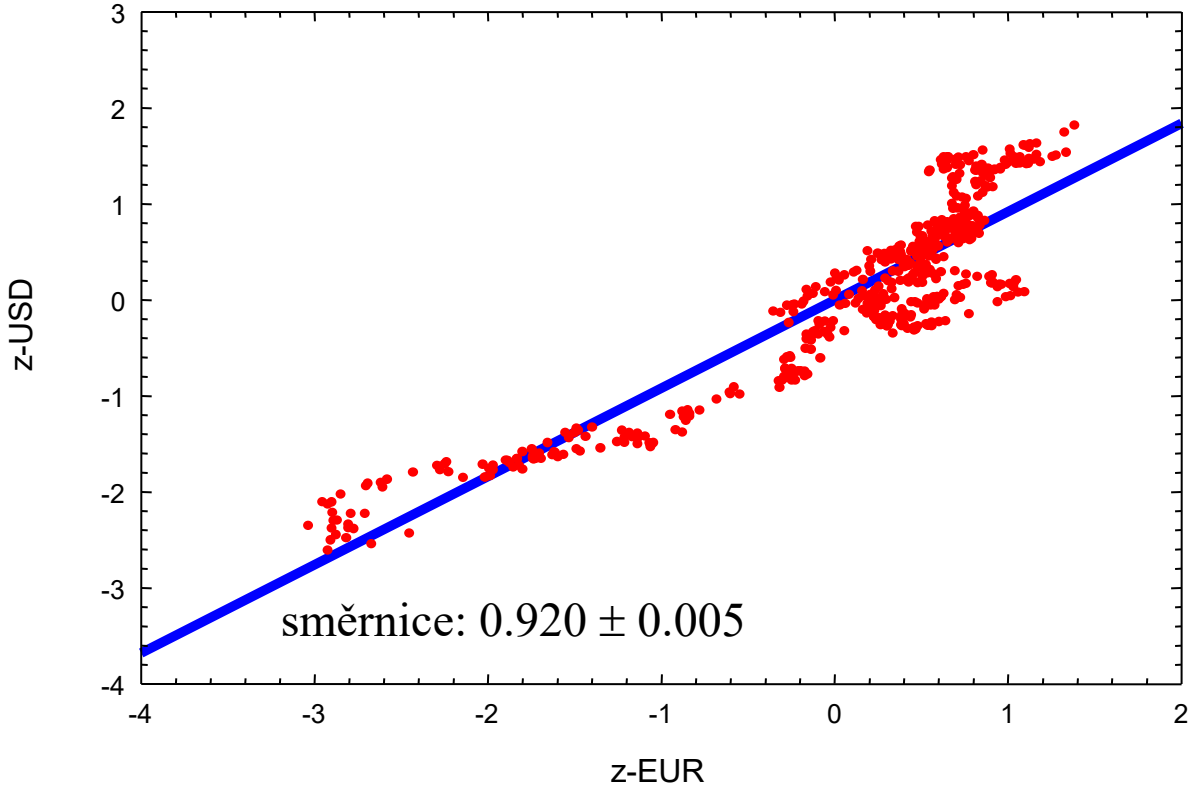
$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} \frac{\left(x_i - \bar{x}\right)^2}{N-1}}$$

Pearsonův korelační koeficient

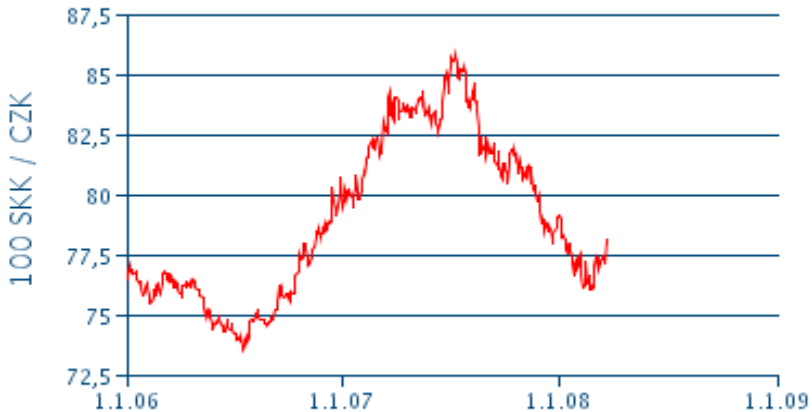
$$r = \frac{1}{N-1}\sum_{i=1}^{N} \frac{x_i - \bar{x}}{\hat{\sigma}_x^2}\frac{y_i - \bar{y}}{\hat{\sigma}_y^2} = \frac{1}{N-1}\sum_{i=1}^{N} \hat{z}_{x,i}\hat{z}_{y,i}$$

Odhad korelace

$\rho = 0.920 \pm 0.003$



Odhad korelace



$$\hat{\mathrm{cov}}(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

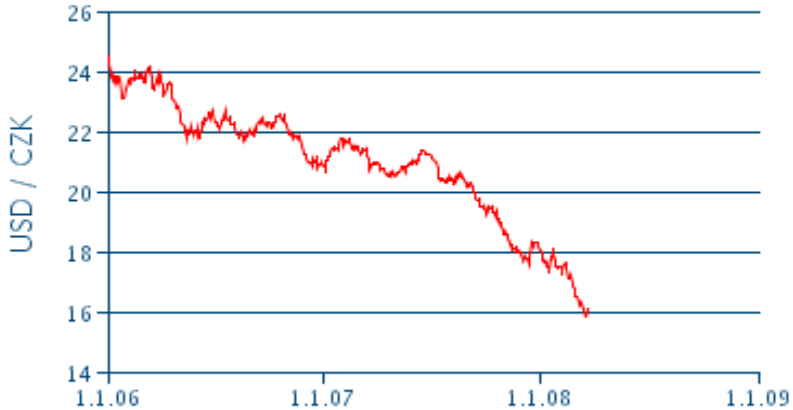
$$\hat{\varrho}(x,y) = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\hat{\sigma}_x \hat{\sigma}_y}$$

ρ = -0.30 ± 0.06

$$\langle xy \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i y_i$$

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \hat{\sigma}_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2}$$

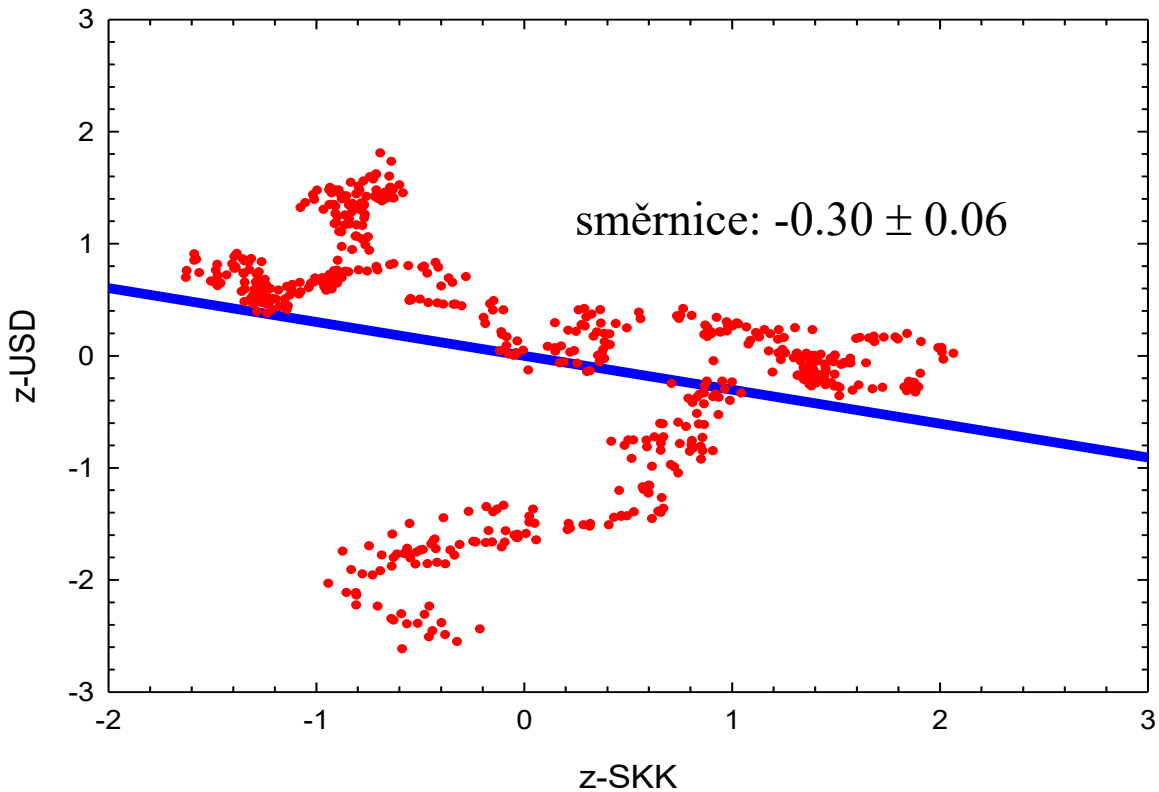
$$\langle y \rangle = \frac{1}{N} \sum_{i=1}^{N} y_i \qquad \hat{\sigma}_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \langle y \rangle)^2}$$



leden 2006 – březen 2008

Odhad korelace

$\rho = -0.30 \pm 0.06$



Autokorelace

- korelační funkce: $$R(t,t') = \rho(x_t, x'_{t'}) = \frac{E[(x_t - \mu_t)(x'_{t'} - \mu'_{t'})]}{\sigma_t \sigma_{t'}}$$

- stacionární proces 2. řádu: $\mu_t \equiv \mu$ (1. a 2. moment nezávisí na čase)

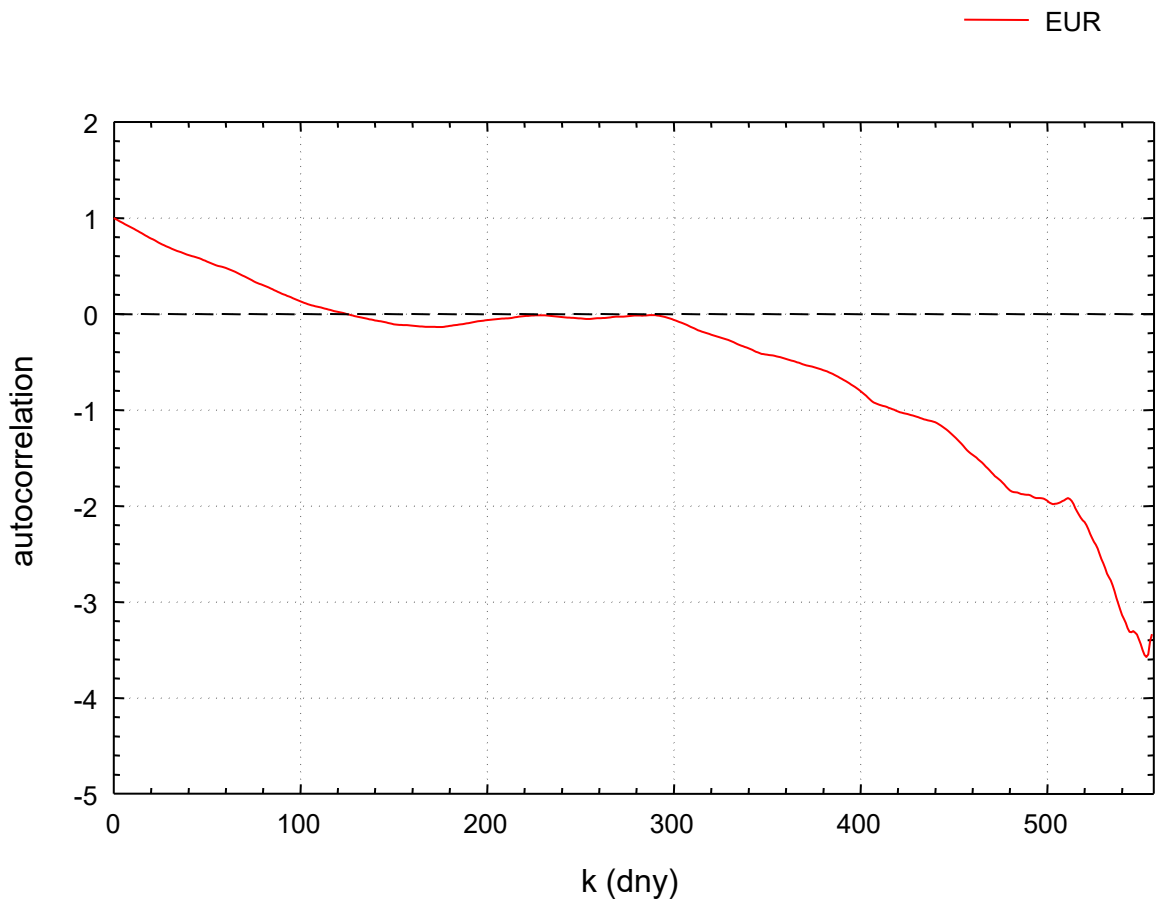
  $\sigma_t \equiv \sigma$

- autokorelace: $$R(\tau) = R(t, t+\tau) = \rho(x_t, x_{t+\tau}) = \frac{E[(x_t - \mu)(x_{t+\tau} - \mu)]}{\sigma^2}$$

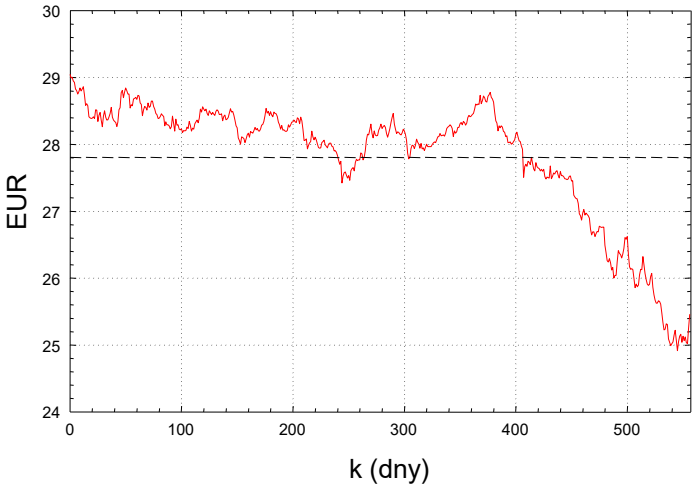
- diskrétní proces: $\{x_1, x_2, \ldots, x_N\}$

- odhad autokorelace: $$R(k) = \frac{1}{(N-k)\sigma^2} \sum_{i=1}^{N-k} (x_i - \mu)(x_{i+k} - \mu)$$

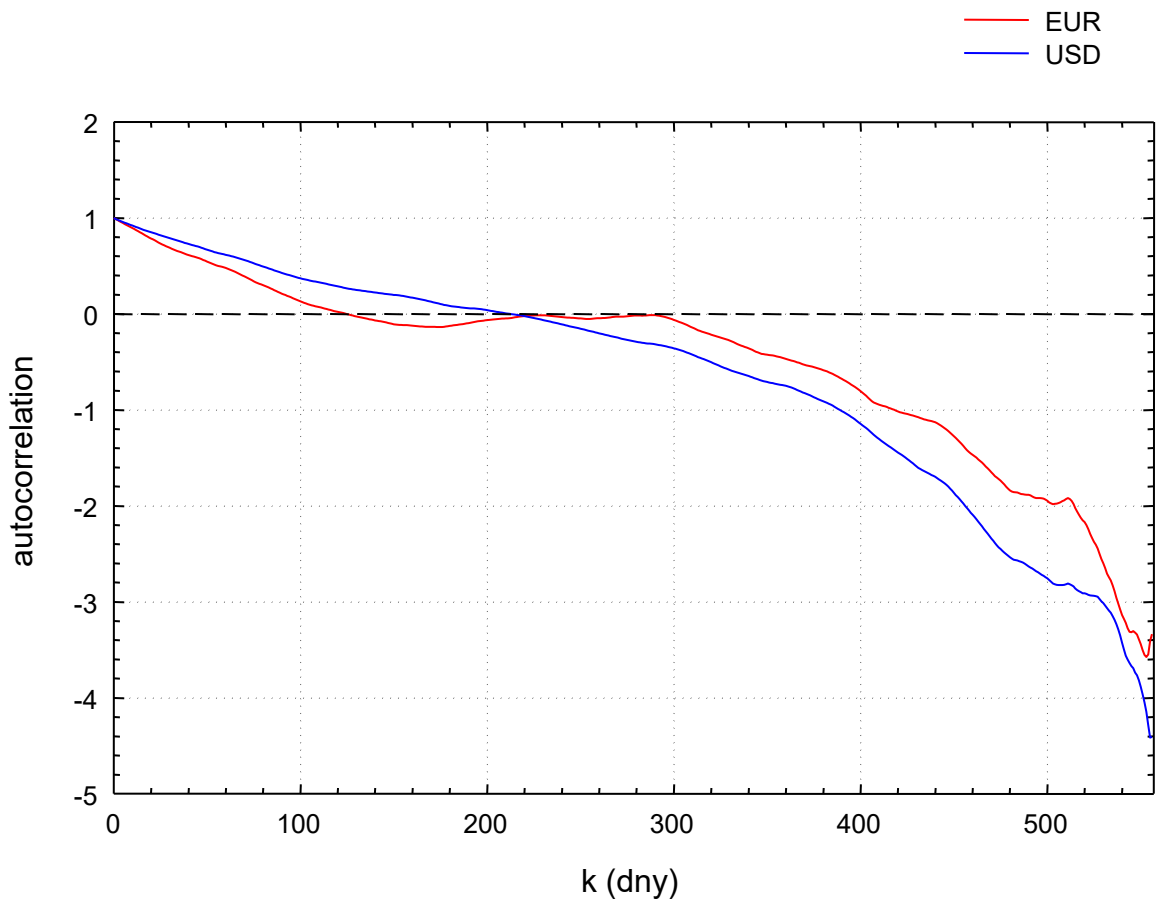
Autokorelace



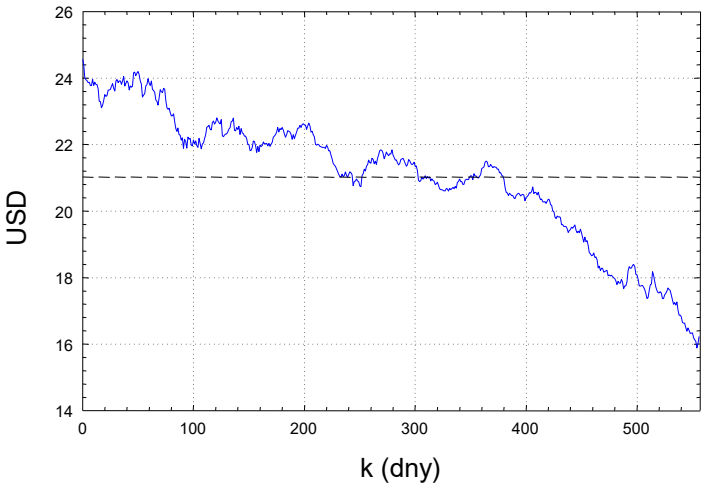
leden 2006 – březen 2008



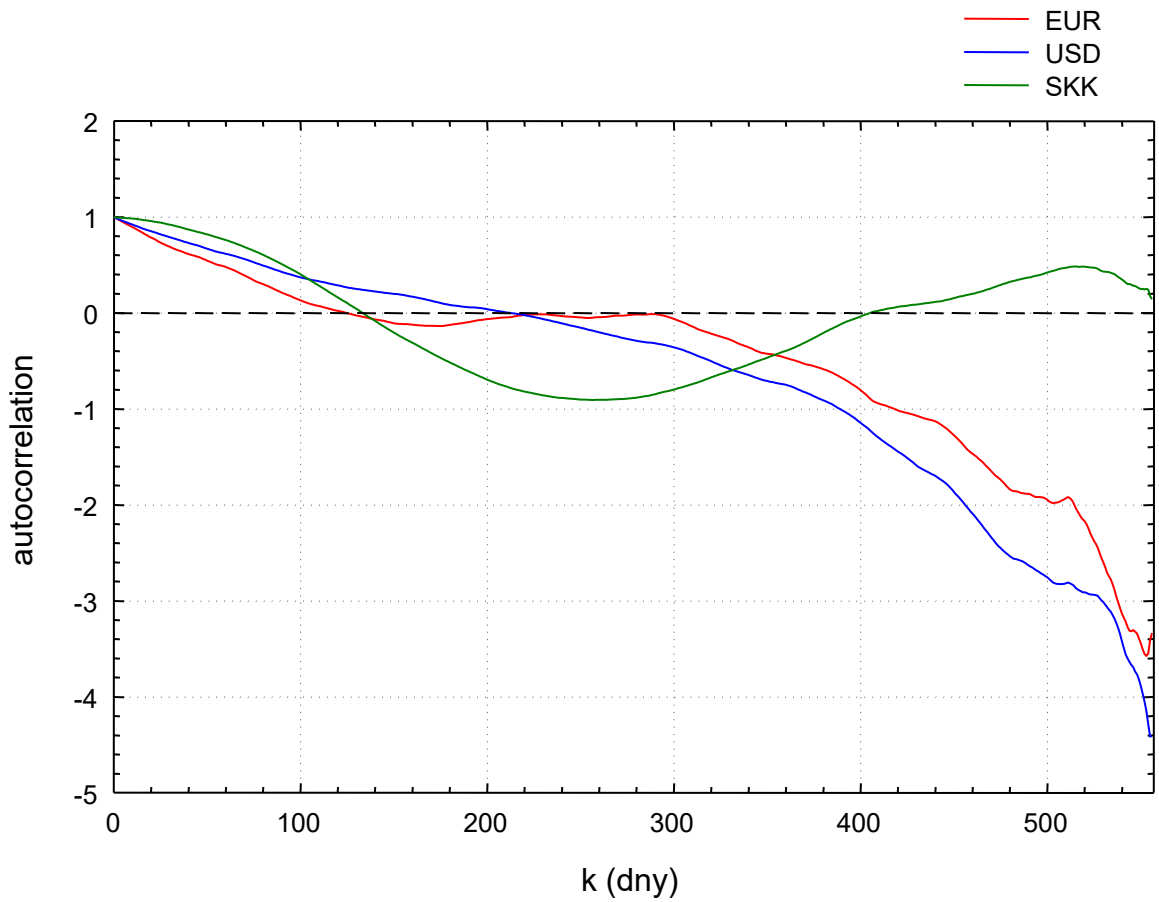
Autokorelace



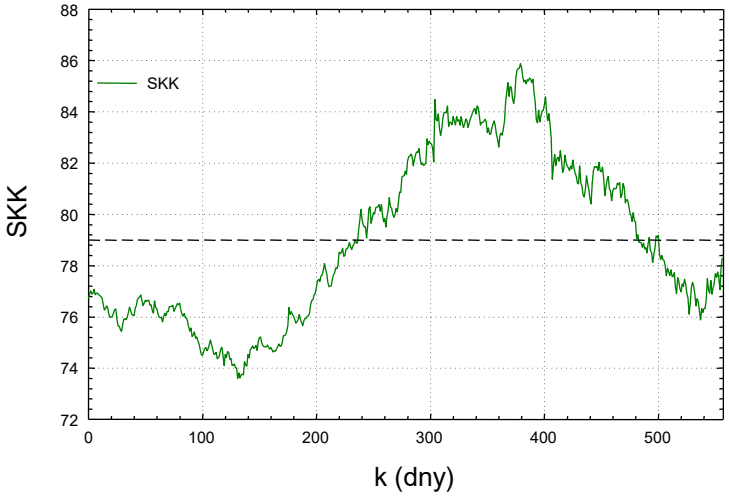
leden 2006 – březen 2008



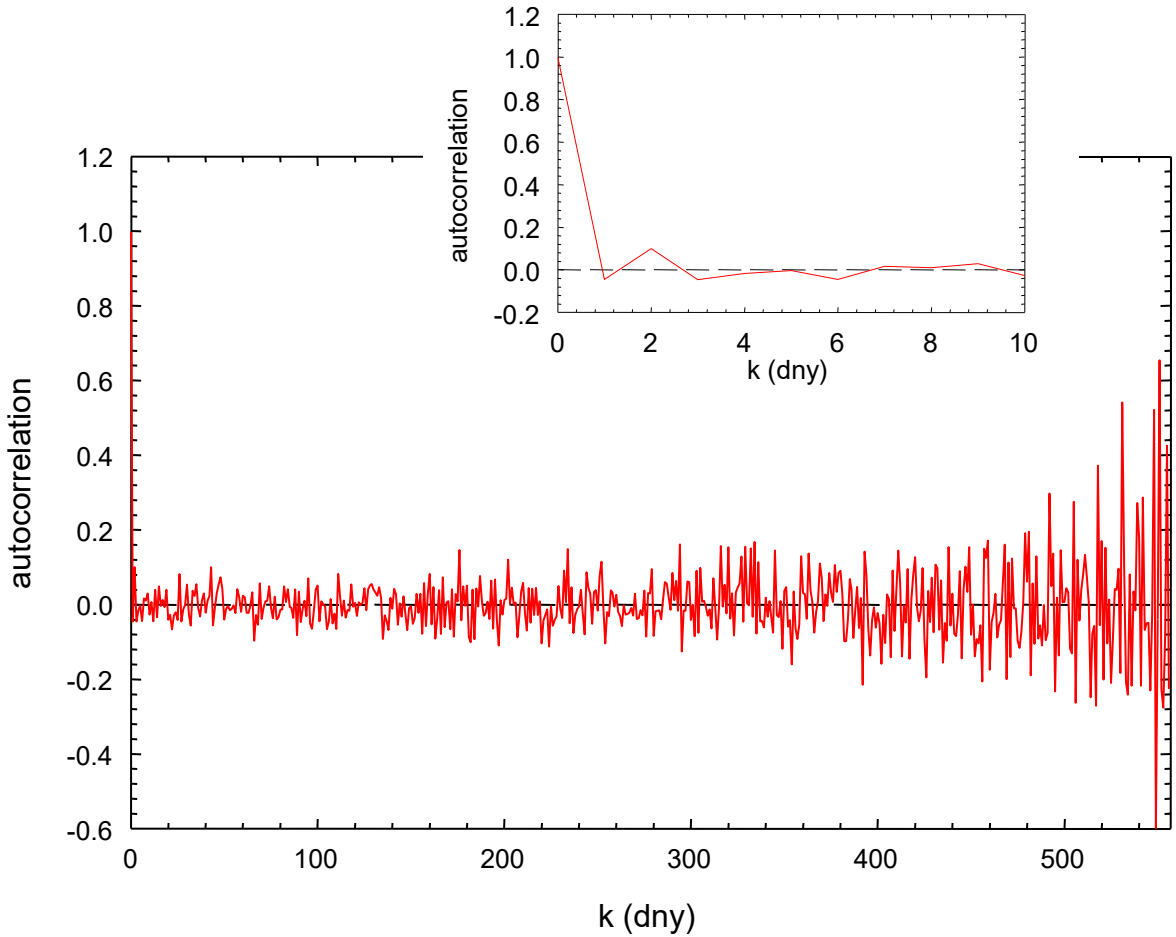
Autokorelace



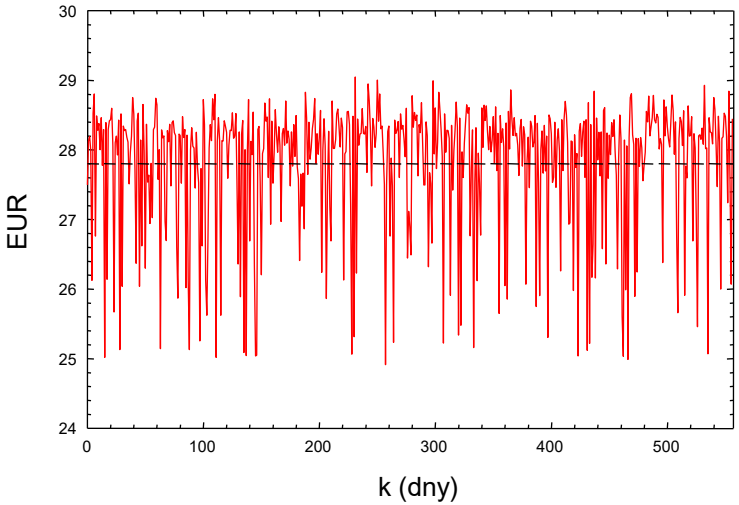
leden 2006 – březen 2008



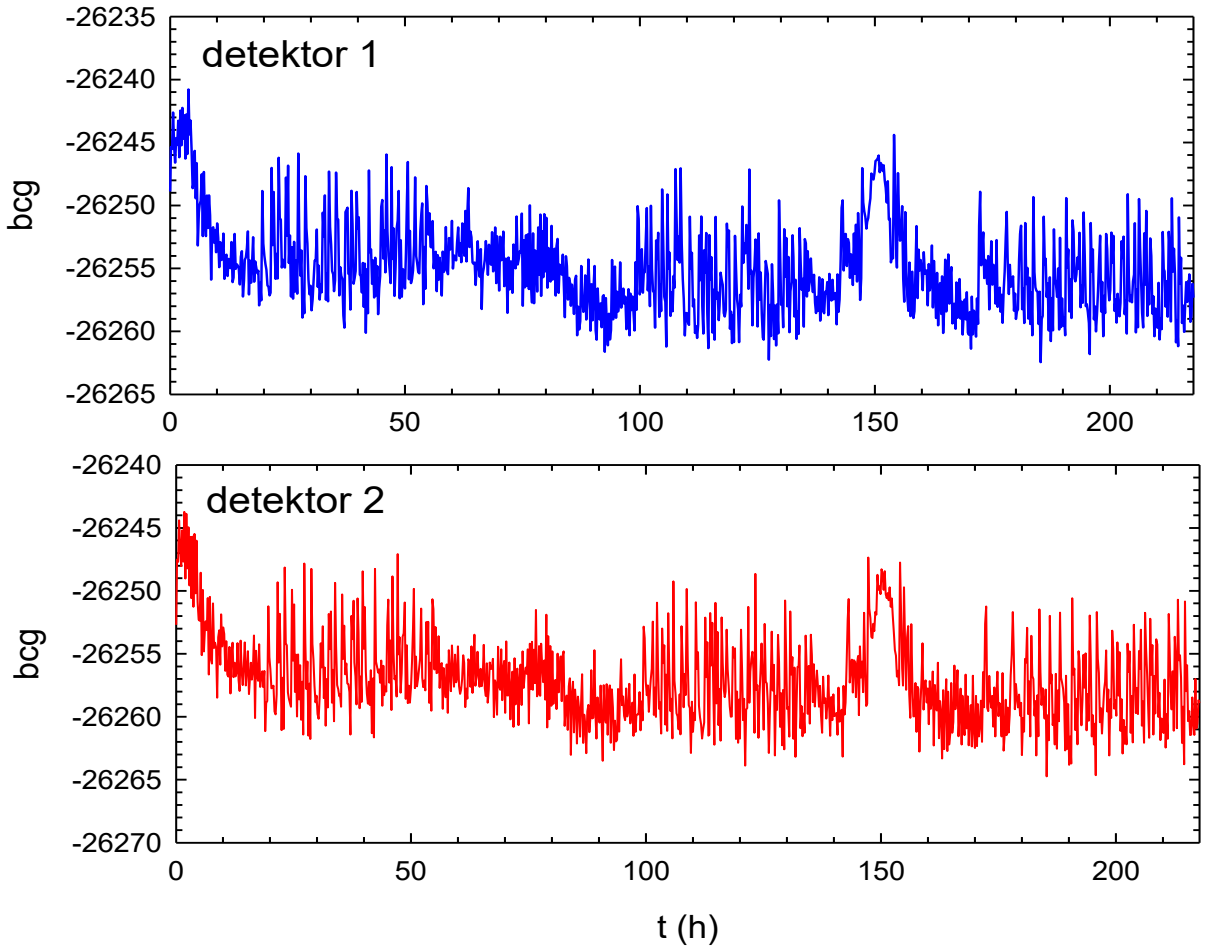
Autokorelace



leden 2006 – březen 2008

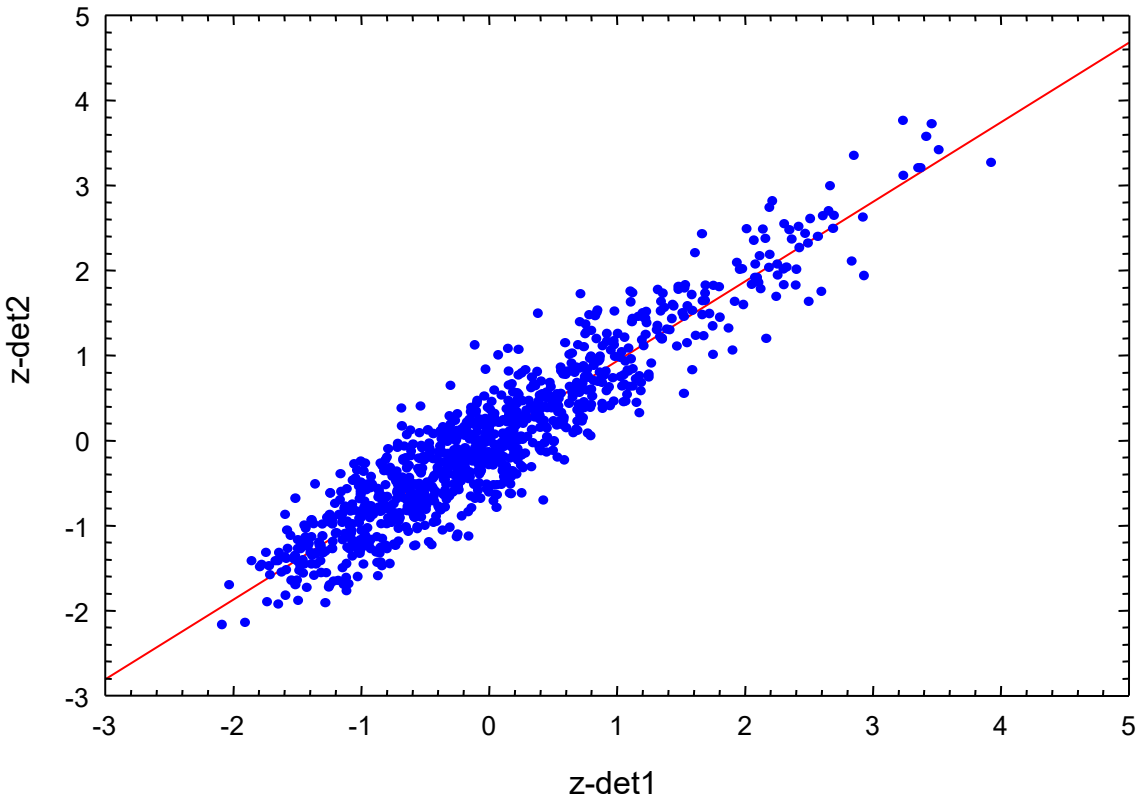


Odhad korelace

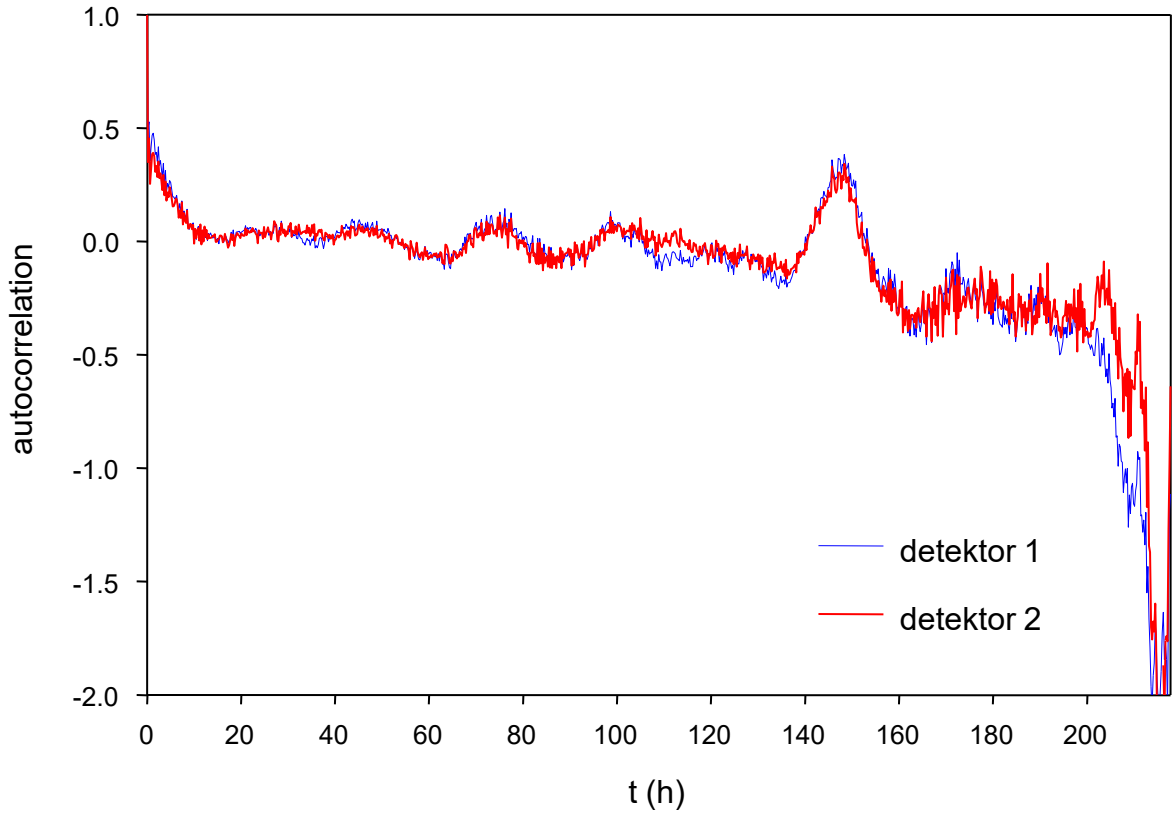


Odhad korelace

korelace $\rho = 0.936 \pm 0.002$

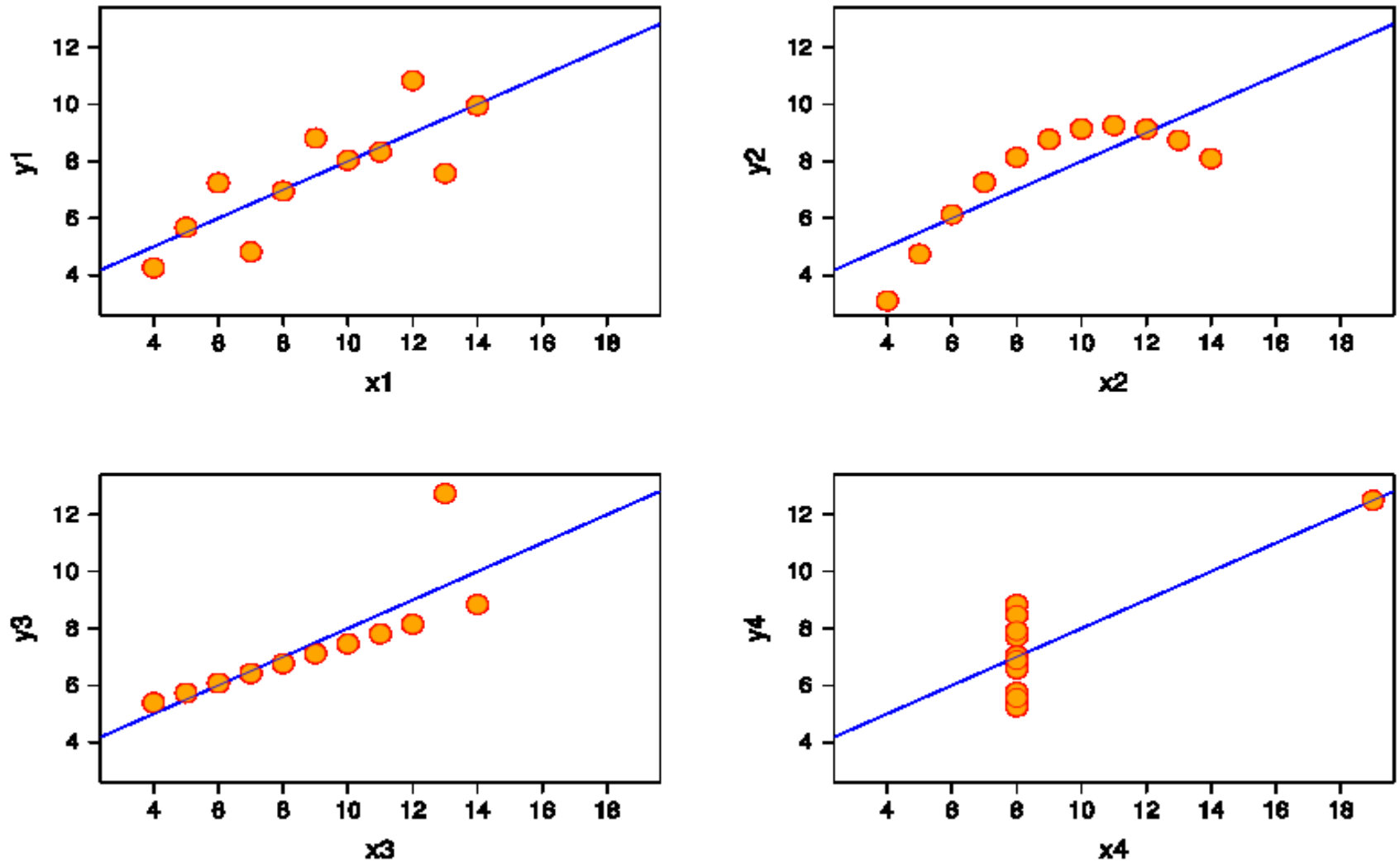


autokorelace



Korelace

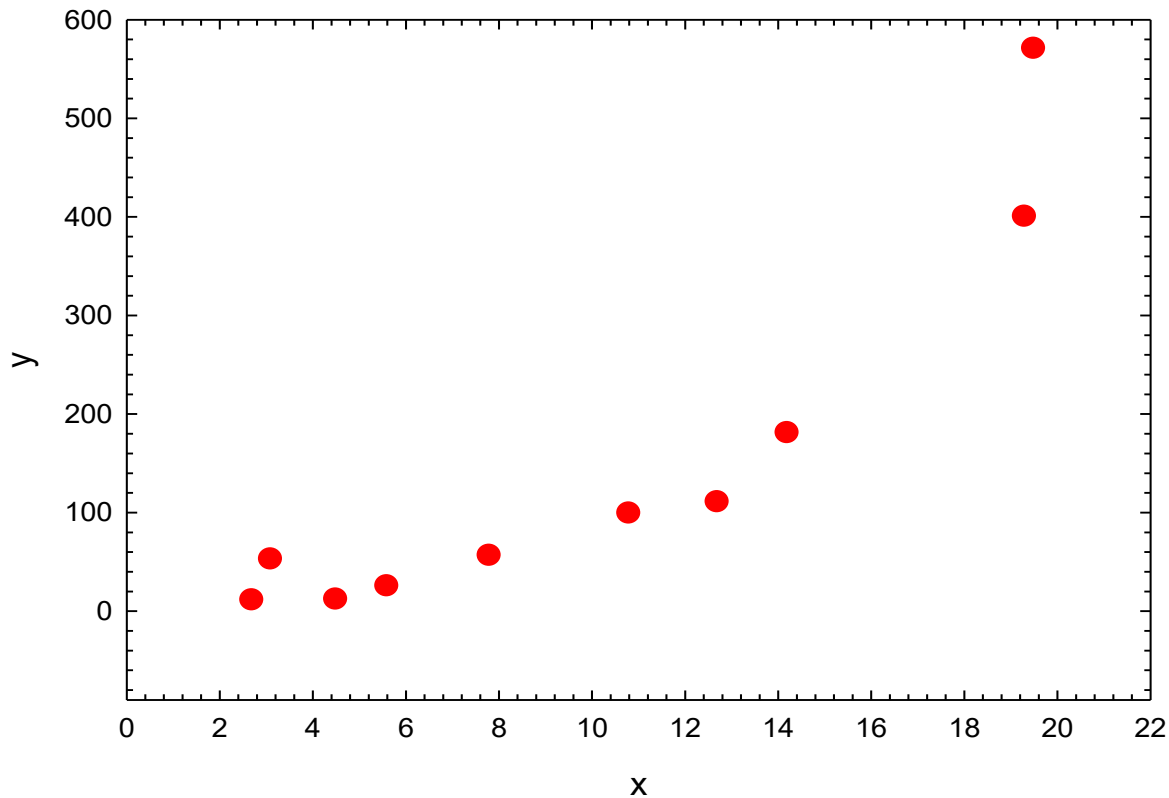
$$\hat{\mu}_y = 7.5 \qquad \hat{\sigma}_y = 4.1 \qquad \rho(x, y) = 0.82$$



Odhad korelace

- Spearmanův korelační koeficient

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)} \qquad d_i = rank(x_i) - rank(y_i)$$



Odhad korelace

x_i	y_i				
5.6	25.1				
14.2	180.7				
2.7	11.0				
3.1	52.5				
4.5	11.7				
10.8	99.1				
19.5	570.5				
19.3	400.0				
7.8	56.1				
12.7	110.5				

Odhad korelace

x_i	y_i	*rank-x*	*rank-y*		
5.6	25.1	4	3		
14.2	180.7	8	8		
2.7	11.0	1	1		
3.1	52.5	2	4		
4.5	11.7	3	2		
10.8	99.1	6	6		
19.5	570.5	10	10		
19.3	400.0	9	9		
7.8	56.1	5	5		
12.7	110.5	7	7		

Odhad korelace

x_i	y_i	*rank-x*	*rank-y*	d_i	d_i^2
5.6	25.1	4	3	1	1
14.2	180.7	8	8	0	0
2.7	11.0	1	1	0	0
3.1	52.5	2	4	-2	4
4.5	11.7	3	2	1	1
10.8	99.1	6	6	0	0
19.5	570.5	10	10	0	0
19.3	400.0	9	9	0	0
7.8	56.1	5	5	0	0
12.7	110.5	7	7	0	0

$$d_i = rank(x_i) - rank(y_i)$$

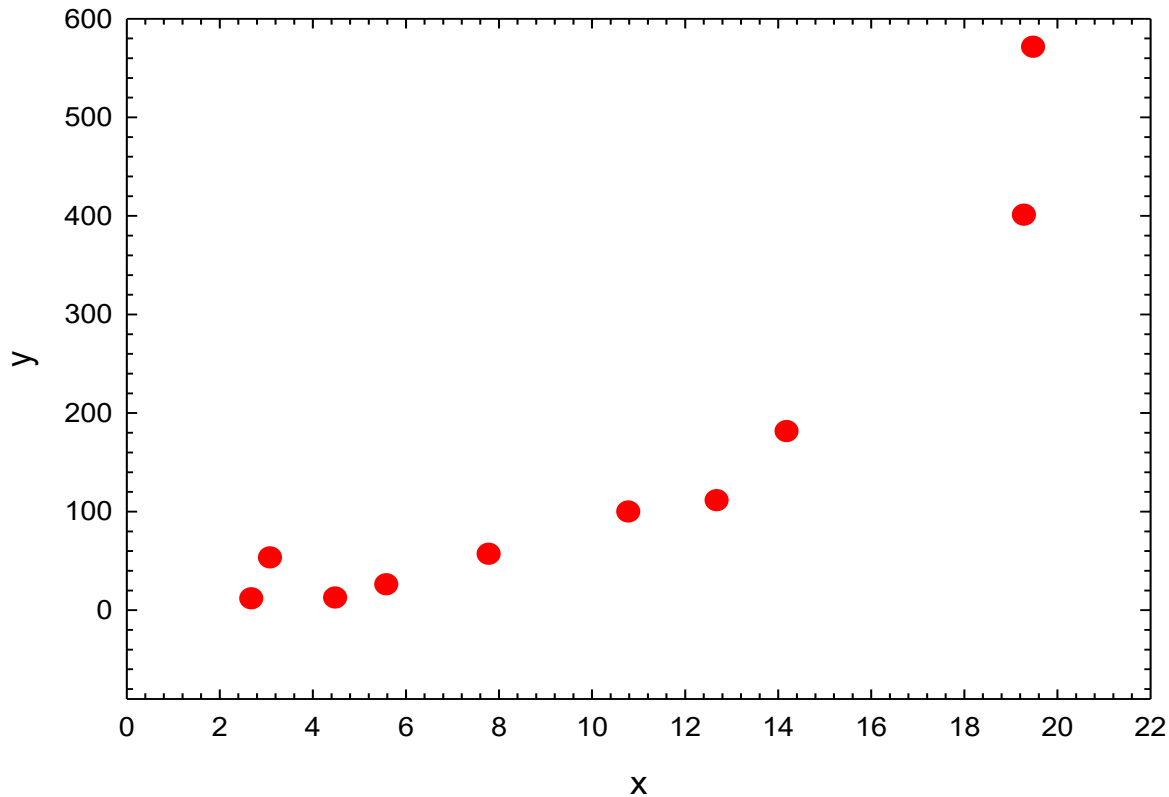
- Spearmanův korelační koeficient

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)}$$

Odhad korelace

Pearson: $\rho = \mathbf{0.88 \pm 0.05}$

Spearman: $\rho = \mathbf{0.96 \pm 0.02}$



Odhad korelace

- Spearmanův korelační koeficient

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)} \qquad d_i = rank(x_i) - rank(y_i)$$

- odhad spolehlivosti $F(\varrho) = \frac{1}{2} \ln \frac{1+\varrho}{1-\varrho}$ (Fisherova transformace)

- z - skóre $z = \sqrt{\frac{N-3}{1.06}} F(\rho)$

- pokud jsou x, y nezávislé je z výběrem ze standardního normálního rozdělení N(0,1)

$$P\left(|z| \geq z_0\right) = 2\left(1 - F_{0,1}\left(|z_0|\right)\right) = 1 - erf\left(\frac{|z_0|}{\sqrt{2}}\right)$$

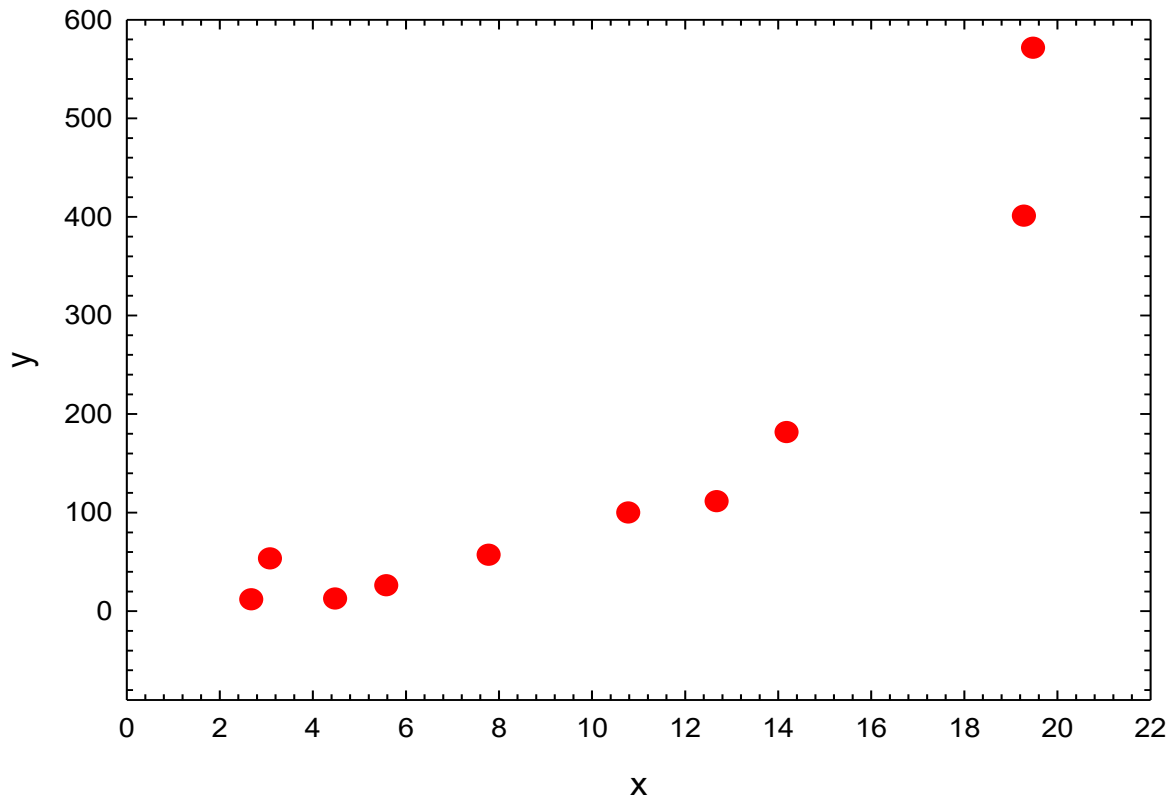
Odhad korelace

$z = 5.01$

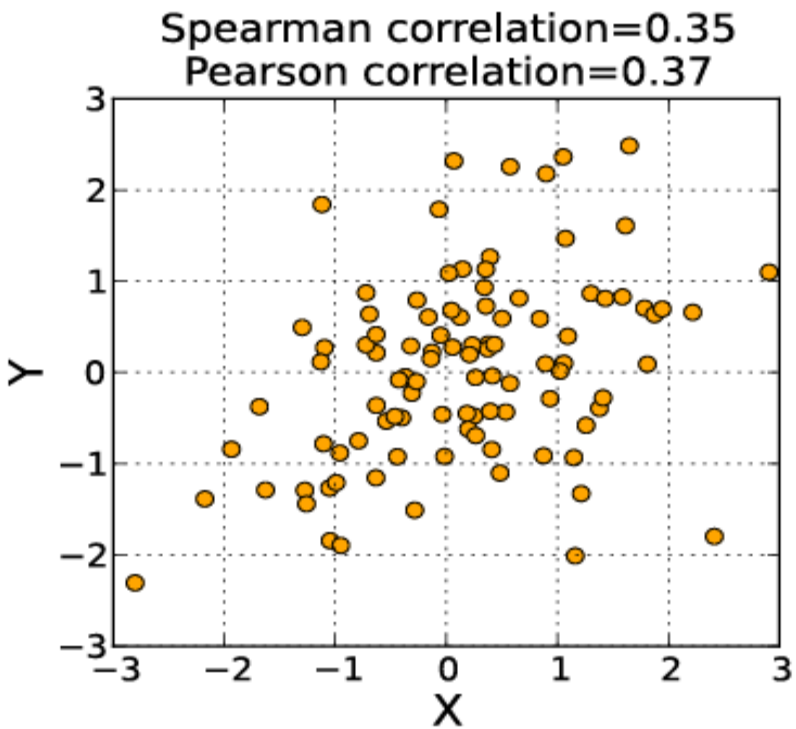
$P(|z| \geq 5.01) = 5 \times 10^{-7}$

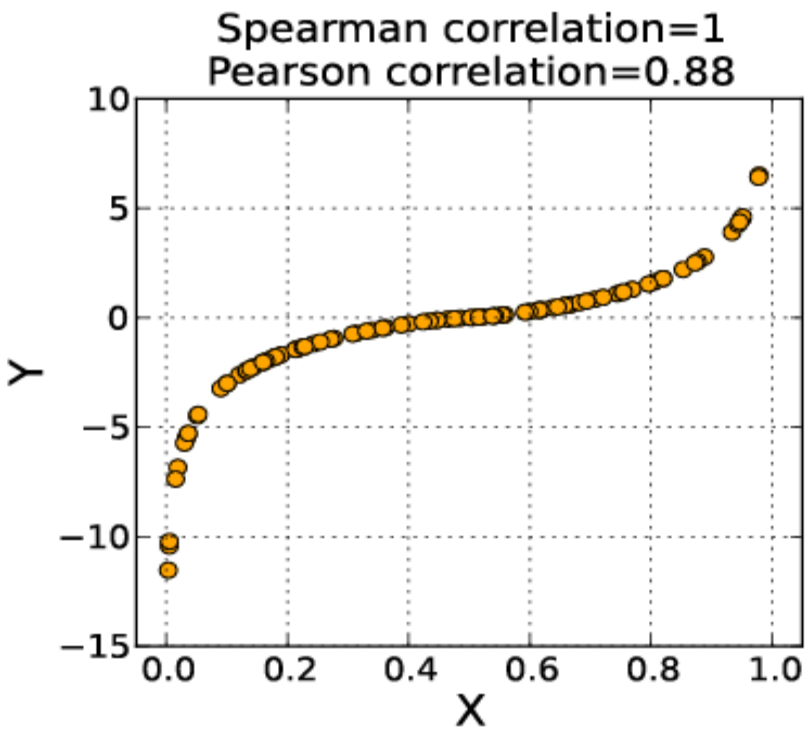
Pearson: $\rho = \mathbf{0.88 \pm 0.05}$

Spearman: $\rho = \mathbf{0.96 \pm 0.02}$

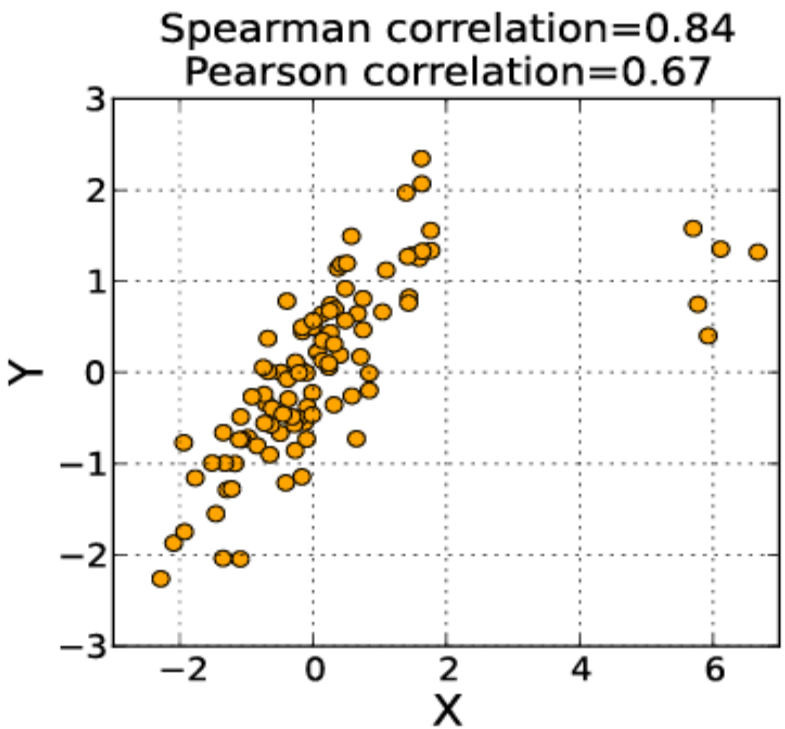


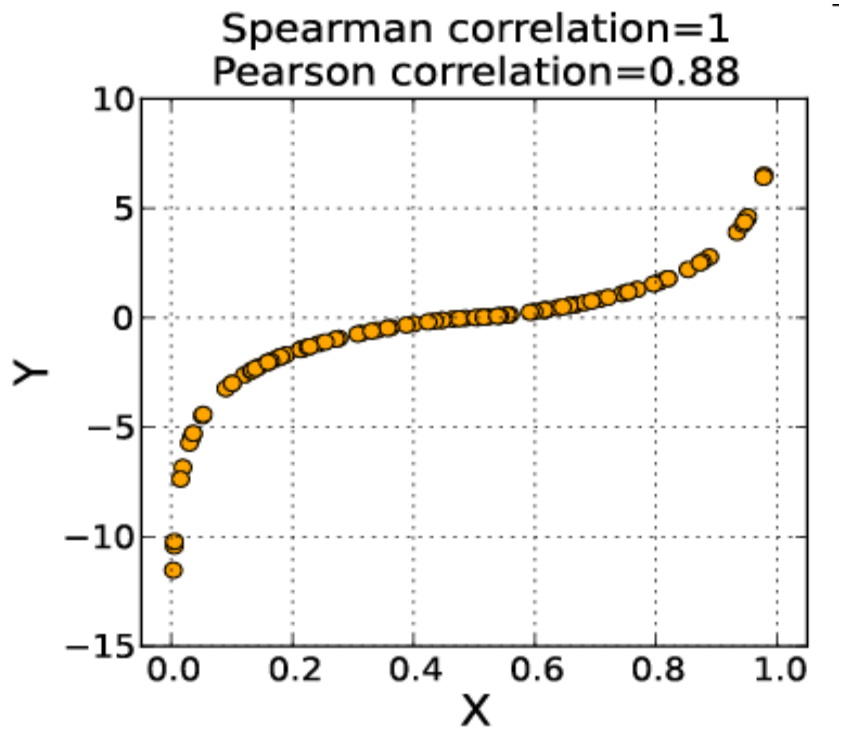
Odhad korelace



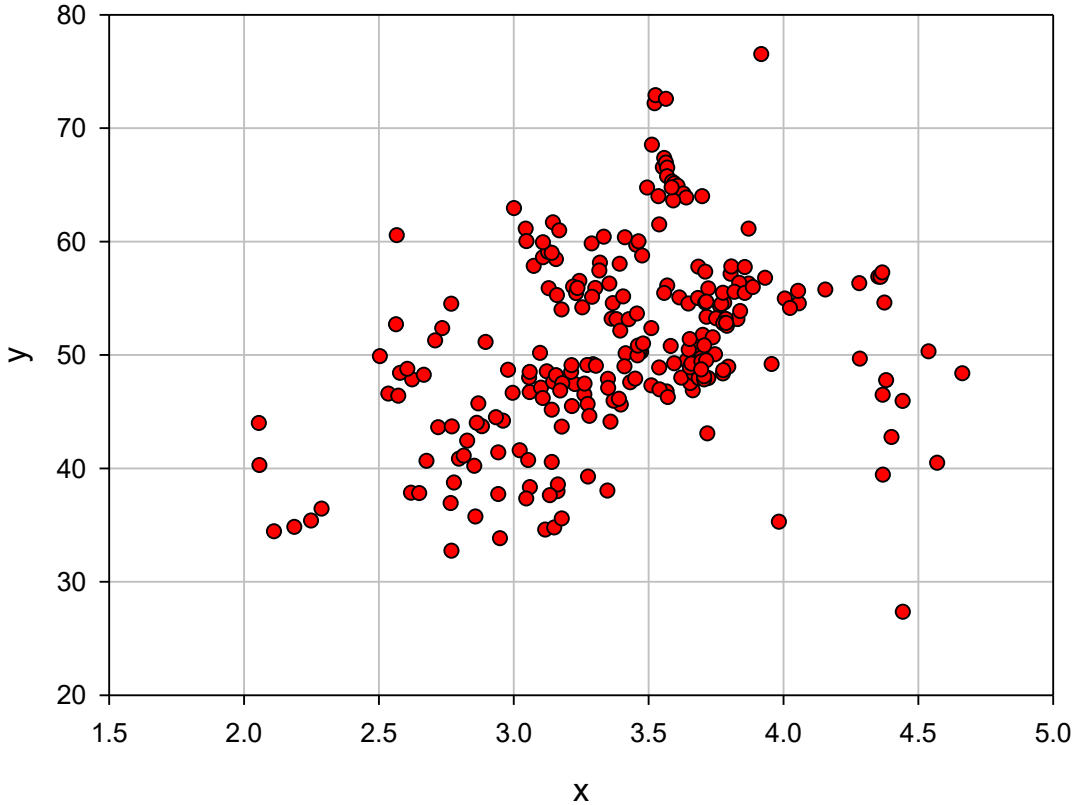


Odhad korelace





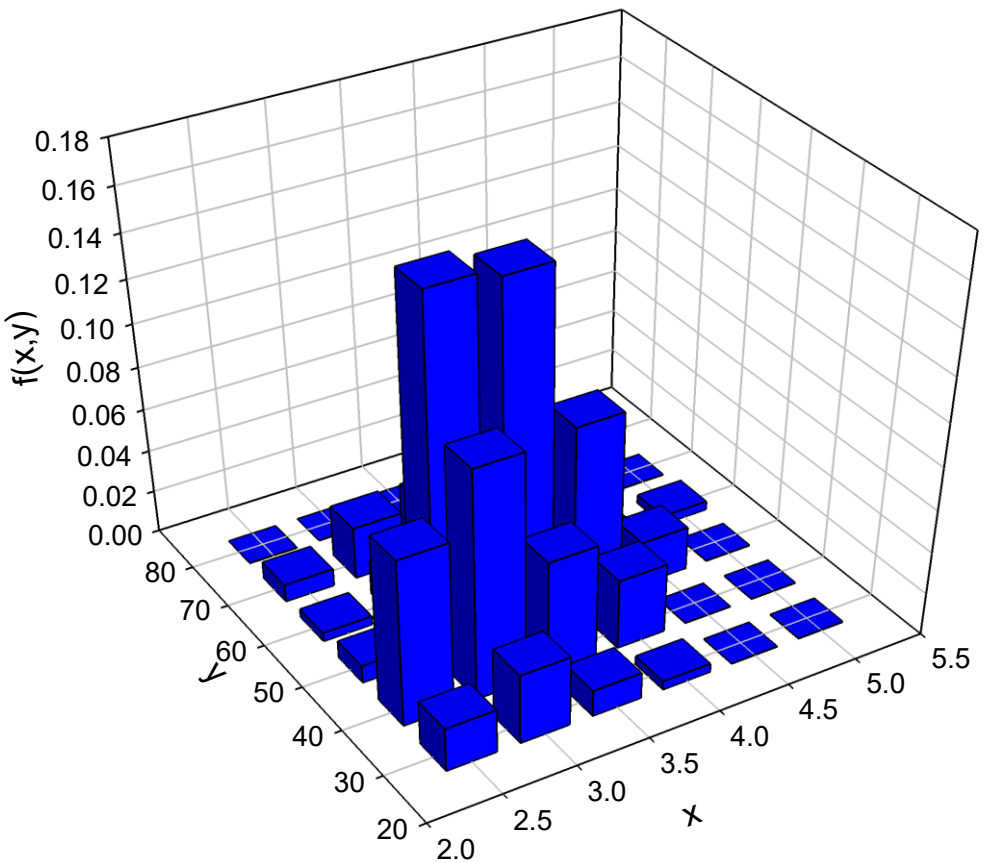
χ^2 test nezávislosti



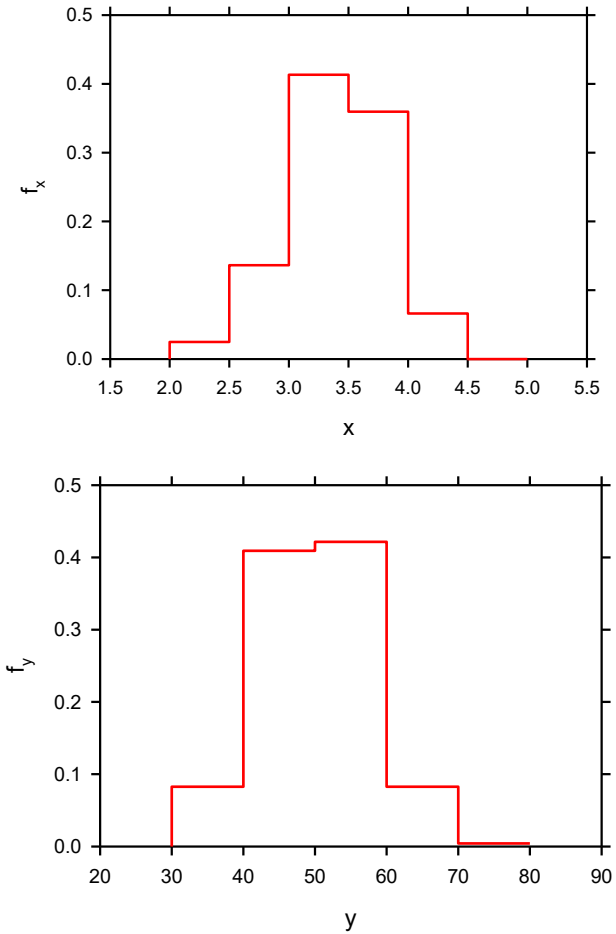
nulová hypotéza: x, y jsou nezávislé

χ^2 test nezávislosti

2D normovaný histogram → odhad $f(x,y)$



1D normované histogramy → $f_x(x)$, $f_y(y)$

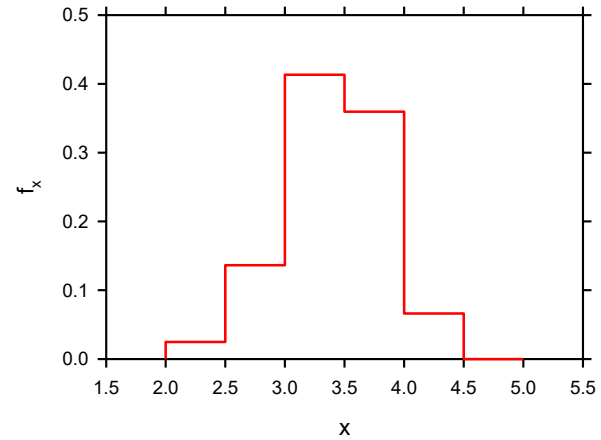


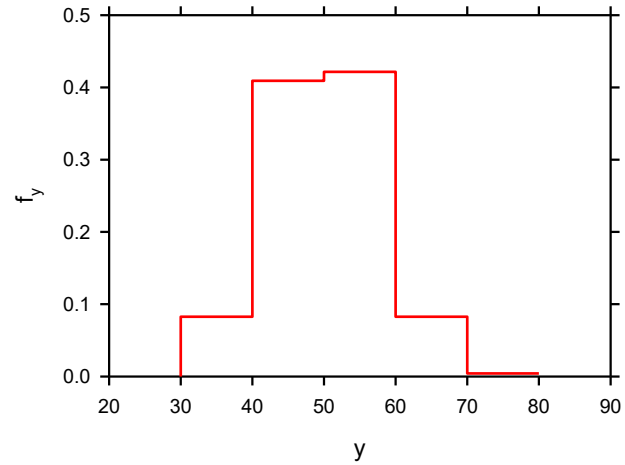
χ^2 test nezávislosti

nezávislé náhodné proměnné: $f(x,y) = f_x(x)\ f_y(y)$

$$\chi^2 = \sum_{i=1}^{m_c} \sum_{j=1}^{m_r} \frac{\left(f(x_i, y_i) - f_x(x_i) f_y(y_i)\right)^2}{f_x(x_i) f_y(y_i)}$$

m_c – počet sloupců 2D histogramu
m_r – počet řádků 2D histogramu





Rozdělení χ^2

$$f(y|N) = \frac{1}{2^{N/2}\Gamma(N/2)} y^{N/2-1} e^{-y/2} \qquad y \in \langle 0, \infty), \quad N = 1,2,\ldots$$

N – počet stupňů volnosti

$$z_i \in N(0,1)$$

$$y = \sum_{i=1}^{N} z_i^2 \longrightarrow y \in \chi^2(N)$$

$$x_i \in N(\mu, \sigma)$$

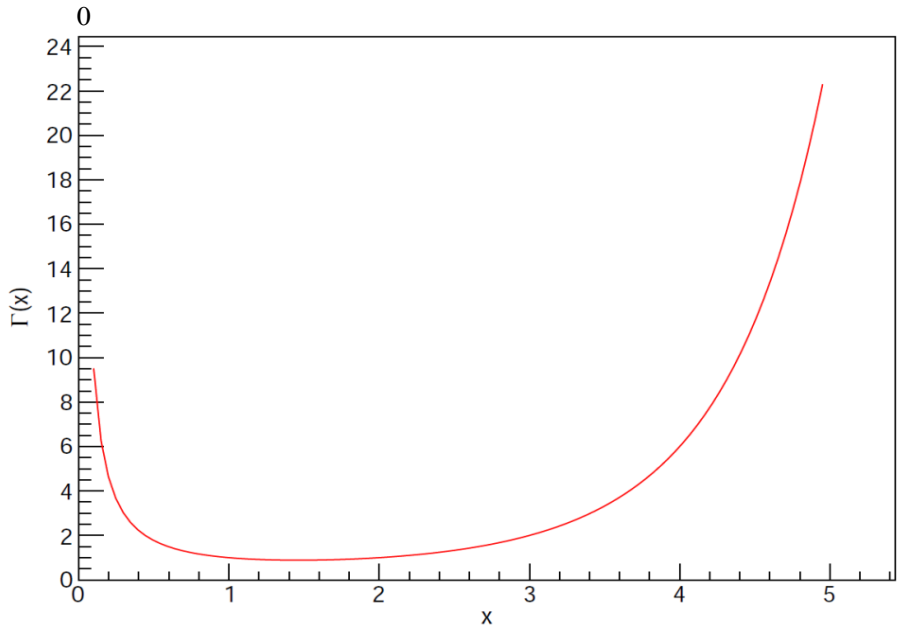
$$y = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \longrightarrow y \in \chi^2(N)$$

- gama funkce: $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, \mathrm{d}t$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(N) = (N-1)!$$



Rozdělení χ^2

$$f(y|N) = \frac{1}{2^{N/2}\Gamma(N/2)} y^{N/2-1} e^{-y/2} \qquad y \in \langle 0, \infty), \quad N = 1,2,\ldots$$

$$z_i \in N(0,1)$$

$$y = \sum_{i=1}^{N} z_i^2 \longrightarrow y \in \chi^2(N)$$

$$x_i \in N(\mu, \sigma)$$

$$y = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \longrightarrow y \in \chi^2(N)$$

N – počet stupňů volnosti

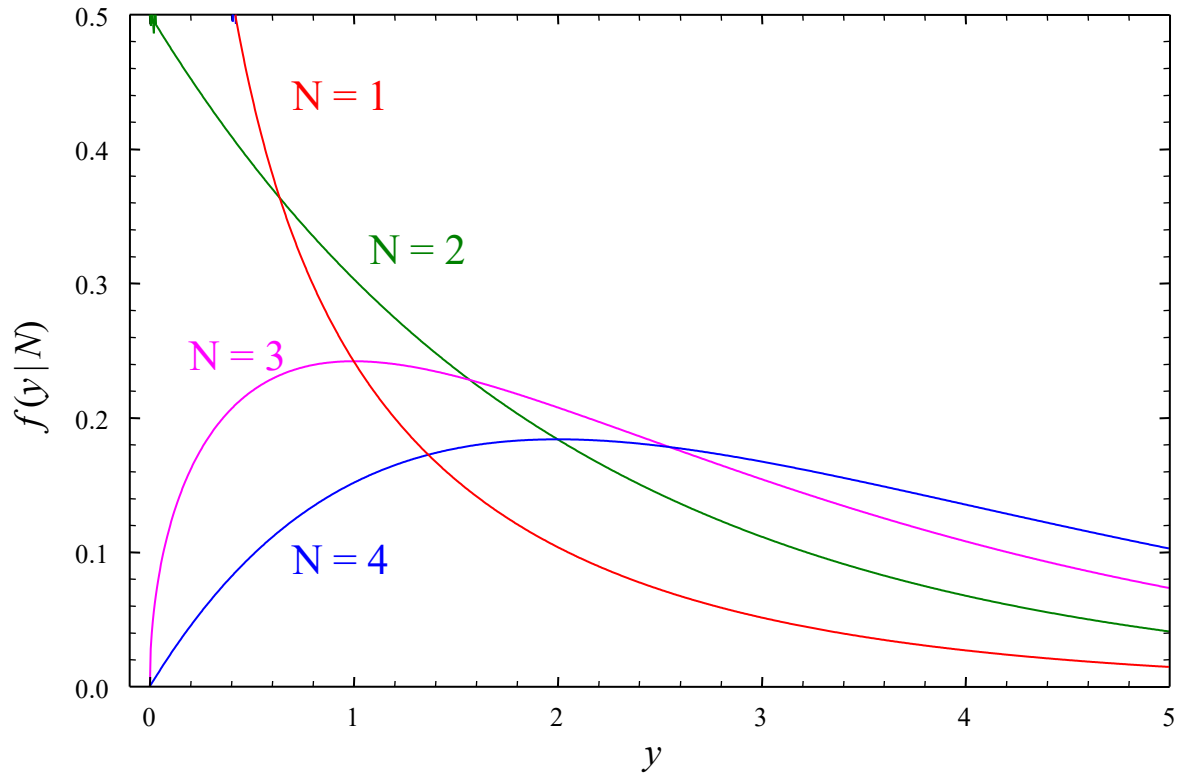
- gama funkce: $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, \mathrm{d}t$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(N) = (N-1)!$$

- momenty χ^2 rozdělení:

$$E[y] = N$$

$$V[y] = 2N$$