Testování hypotéz – test korelace

- náhodné proměnné x a y
- naměříme hodnoty $x_1, x_2, \dots, x_N$ a $y_1, y_2, \dots, y_N$
- vypočítáme odhad korelace: $\hat{\rho}(x,y) = \dfrac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\hat{\sigma}_x \hat{\sigma}_y}$ s chybou: $\hat{\sigma}_\rho \approx \dfrac{1-\hat{\rho}^2}{\sqrt{N-1}}$
- **Je korelace statisticky významná?**

- použijeme **testovací statistiku** $f(t|H_0)$

  známá hustota pravděpodobnosti $f(t)$ a distribuční funkce $F(t)$

  transformace $\rho \to t$ (nová testovací proměnná t)

  nulová hypotéza H_0 (předpoklad nulové korelace proměnných x a y)

  hladina signifikance P_α (typicky 5 % nebo 1 %), pro $P < P_\alpha$ odmítneme H_0

  $$P = P\big(t > \hat{t}(\hat{\rho})\big) = 1 - F\big(\hat{t}(\hat{\rho})\big)$$

Test korelace – Fisherova transformace

- **Fisherova transformace**

transformace

$$\boxed{z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}}$$

Pokud platí nulová hypotéza, má proměnná z normální rozdělení $N(\mu, \sigma)$,

kde očekávaná hodnota $\mu = 0$ a standardní odchylka $\sigma = 1/\sqrt{N-3}$.

testovací proměnná

$$\boxed{t = \frac{z}{\sigma} = \frac{\sqrt{N-3}}{2} \ln \frac{1+\rho}{1-\rho}}$$

Pokud platí nulová hypotéza, má proměnná t normální rozdělení $N(0,1)$.

t-hodnota

$$\boxed{\hat{t}(\hat{\rho}) = \frac{\sqrt{N-3}}{2} \ln \frac{1+\hat{\rho}}{1-\hat{\rho}}}$$

Test korelace – Fisherova transformace

- **Fisherova transformace**

testovací statistika

$$\boxed{f(t|H_0) = N(0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)}$$

nulová hypotéza H_0 — Náhodné proměnné x a y jsou NEZÁVISLÉ.

hladina signifikance P_α — Pro $P > P_\alpha$ **přijmeme** nulovou hypotézu.

Pro $P < P_\alpha$ **odmítneme** nulovou hypotézu.

pravděpodobnost

$$\boxed{P = P(|t > \hat{t}(\hat{\rho})|) = 2[1 - F(|\hat{t}(\hat{\rho})|)] = 2\left\{1 - \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{|\hat{t}(\hat{\rho})|}{\sqrt{2}}\right)\right]\right\}}$$

Test korelace – Fisherova transformace

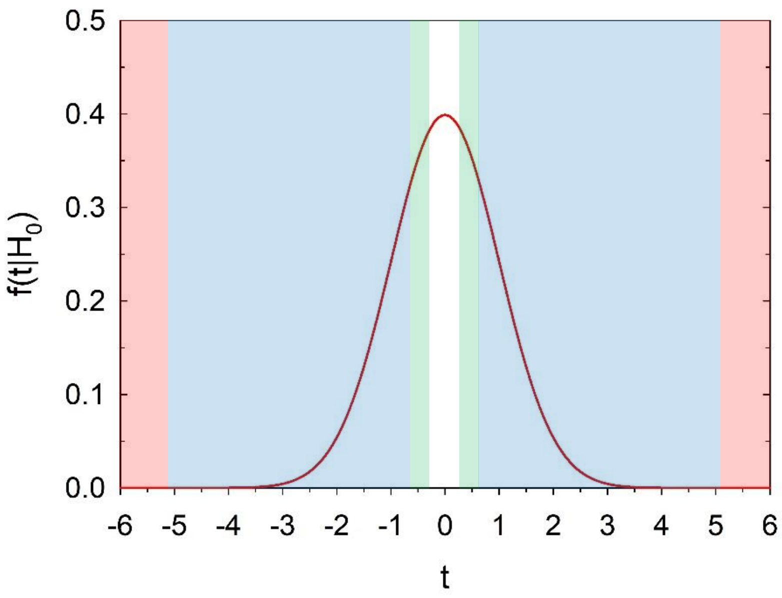
- **Fisherova transformace**

Příklad: *N* = 39, proměnné x, y, z

$\hat{\rho}(x,y) = 0.694$ $\left|\hat{t}_{x,y}\right| = 5.137$ $P_{x,y} = 3 \times 10^{-7}$

$\hat{\rho}(x,z) = -0.111$ $\left|\hat{t}_{x,z}\right| = 0.669$ $P_{x,z} = 0.50$

$\hat{\rho}(y,z) = 0.059$ $\left|\hat{t}_{y,z}\right| = 0.355$ $P_{y,z} = 0.72$



$P < 5\%$ **odmítneme** nulovou hypotézu

Proměnné x, y jsou **ZÁVISLÉ**.

$P > 5\%$ **přijmeme** nulovou hypotézu

Proměnné x, z a y, z jsou **NEZÁVISLÉ**.

Test korelace – studentovo rozdělení

- **studentovo rozdělení**

transformace $\boxed{t = \rho\sqrt{\frac{N-2}{1-\rho^2}}}$ ← testovací proměnná, t-hodnota

Pokud platí nulová hypotéza, má proměnná t studentovo rozdělení s $N-2$ stupni volnosti.

testovací statistika studentovo rozdělení s ν stupni volnosti

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

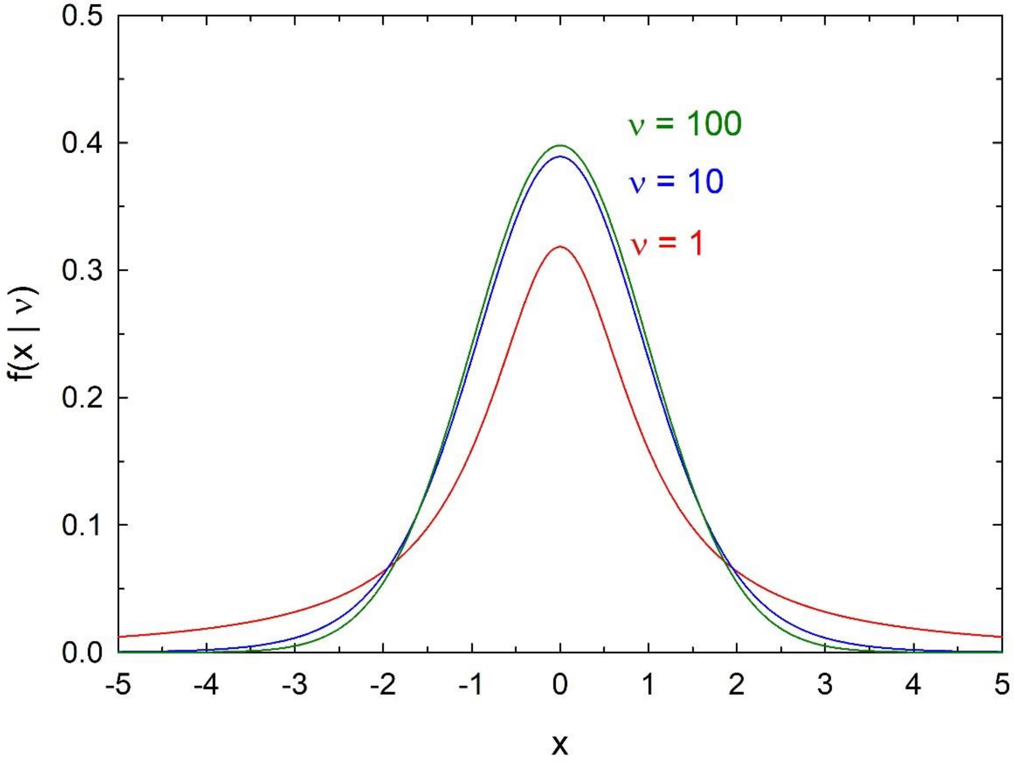
nulová hypotéza H_0 Náhodné proměnné x a y jsou NEZÁVISLÉ.

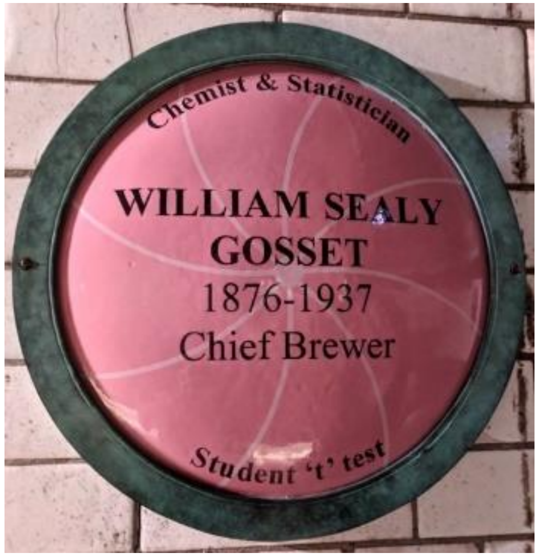
Studentovo rozdělení

- **studentovo rozdělení** s ν stupni volnosti

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

William Sealy Gosset („student“) → statistika na malém počtu vzorků





Studentovo rozdělení

gamma.py

- **studentovo rozdělení** s ν stupni volnosti

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

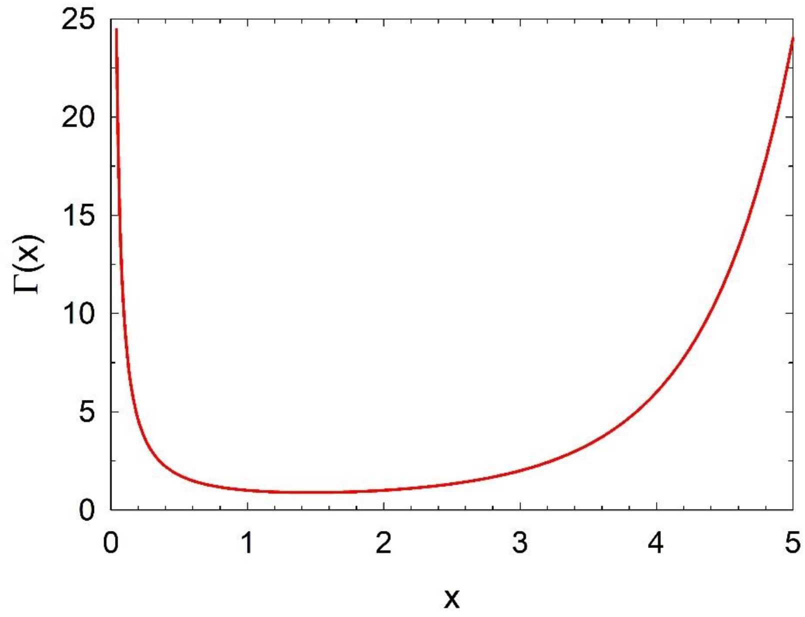
- **gama funkce**

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t \qquad x \geq 0$$

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

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

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



Excel `EXP(GAMMALN(x))`

ROOT `ROOT::Math::tgamma(x)`

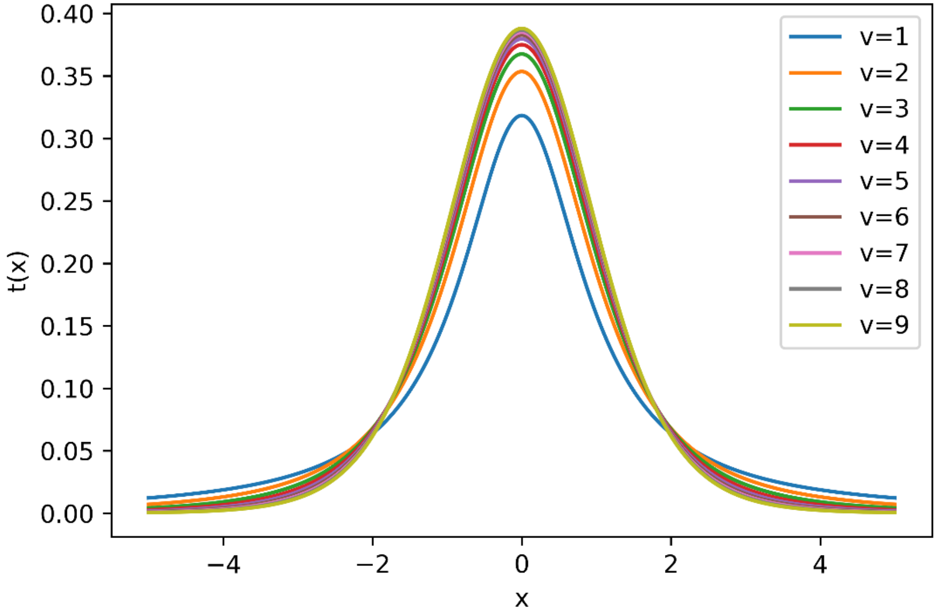
Python `from scipy.special import gamma`

Studentovo rozdělení

`student.py`

- **studentovo rozdělení** s ν stupni volnosti

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



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

$$f(x|\nu \to \infty) \to \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

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

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

Python
```
from scipy.stats import t
→ t.pdf(x,nu)
→ t.cdf(x,nu)
```

ROOT
```
ROOT::Math::tdistribution_pdf(x,nu)
```

Test korelace – studentovo rozdělení

- **studentovo rozdělení**

hladina signifikance P_α

Pro $P > P_\alpha$ **přijmeme** nulovou hypotézu

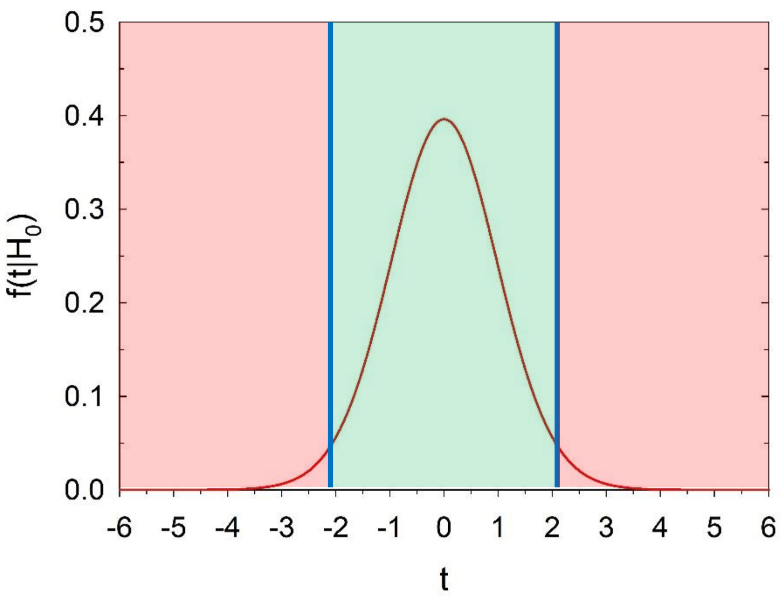
Pro $P < P_\alpha$ **odmítneme** nulovou hypotézu

pravděpodobnost $P = 2[1 - T_\nu(|\hat{t}(\hat{\rho})|)]$

distribuční funkce studentova rozdělení

konfidenční interval $\left(-T_\nu^{-1}(P_\alpha), T_\nu^{-1}(P_\alpha)\right)$

inverzní funkce k distribuční funkci studentova rozdělení



Test korelace – studentovo rozdělení

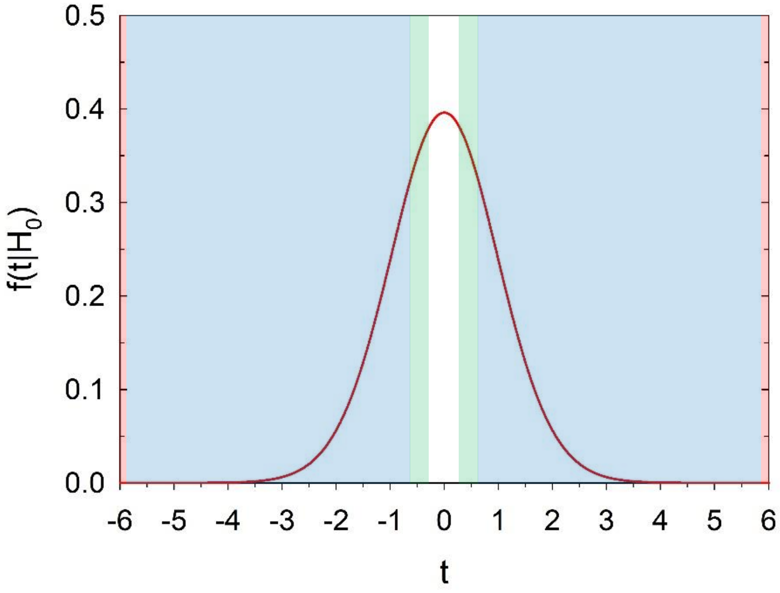
- **studentovo rozdělení**

Příklad: N = 39, proměnné x, y, z, $\nu = N - 2 = 37$

$$\hat{\rho}(x,y) = 0.694 \qquad \left|\hat{t}_{x,y}\right| = 5.867 \qquad P_{x,y} = 9 \times 10^{-7}$$

$$\hat{\rho}(x,z) = -0.111 \qquad \left|\hat{t}_{x,z}\right| = 0.679 \qquad P_{x,z} = 0.50$$

$$\hat{\rho}(y,z) = 0.059 \qquad \left|\hat{t}_{y,z}\right| = 0.360 \qquad P_{y,z} = 0.72$$



$P < 5\%$ **odmítneme** nulovou hypotézu

Proměnné x, y jsou **ZÁVISLÉ**.

$P > 5\%$ **přijmeme** nulovou hypotézu

Proměnné x, z a y, z jsou **NEZÁVISLÉ**.