

**Náhodná (stochastická)** proměnná přiřazuje **pravděpodobnost/hustotu pravděpodobnosti** možnému **diskrétnímu/spojitému** jevu z **diskrétní/spojité** množiny jevů.

● diskrétní příklad: hod kostkou:  $p_i = 1/6$  pro  $i \in \{ \square, \square \cdot, \square \cdot \cdot, \square \cdot \cdot \cdot, \square \cdot \cdot \cdot \cdot, \square \cdot \cdot \cdot \cdot \cdot \}$

● spojitý příklad: čas rozpadu jádra:  $p(t) = ke^{-kt}$

Spojitou náhodnou veličinu v 1D (tj.  $x \in \mathbb{R}$ ) popisuje **distribuční funkce** (hustota pravděpodobnosti, rozdělení/rozložení pravděpodobnosti, *probability distribution function (PDF)*  $p(x)$ ):

$p(x)dx$  je pravděpodobnost, že nastane jev  $x \in [x, x + dx)$

Ve dvou dimenzích definujeme hustotu pravděpodobnosti  $p(x, y)$  tak, že jev  $x \in [z, x + dx)$  a zároveň  $y \in [y, y + dy)$  nastane s pravděpodobností  $p(x, y)dxdy$ .

Uzavřený interval značím  $[a, b]$ , aby se nepletl se střední hodnotou.

Normalizace:

$$\sum_i p_i = 1 \quad \text{nebo} \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

Kumulativní (integrální) distribuční funkce = pravděpodobnost, že padne náhodná hodnota  $x \leq x$ :

$$P(x) = \int_{-\infty}^x p(x')dx'$$

**Varování.** Ve fyzice a technice nepřesně a volně zaměňujeme symbol  $\mathbf{x}$  pro náhodnou veličinu a  $x$  pro její hodnotu (např. při integraci).

**Střední hodnota** (též *expectation value*, očekávaná hodnota; slovo průměr budeme rezervovat pro aritmetický průměr, tj. střední hodnotu výběru)

$$E(\mathbf{x}) \equiv \langle \mathbf{x} \rangle \equiv \langle x \rangle_{\mathbf{x}} \stackrel{\text{volně}}{=} \langle x \rangle = \int x p(x) dx \text{ nebo } \sum_i x_i p_i$$

**Variance**, rozptyl, fluktuace, střední kvadratická odchylka, *mean square deviation (MSD<sup>\*</sup>)*, disperze

$$\text{Var}(\mathbf{x}) \stackrel{\text{volně}}{=} \text{Var } x = \langle (x - \langle x \rangle)^2 \rangle = \langle \Delta x^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \text{ kde } \Delta x = x - \langle x \rangle$$

**Směrodatná odchylka**, standardní odchylka, *standard deviation*, *root mean square deviation*, *RMSD*

$$\sigma(\mathbf{x}) = \sqrt{\text{Var}(\mathbf{x})}$$

též  $\sigma(x)$ ,  $\sigma_x$ ,  $\delta x$ ,  $s(x)$  (odhad  $\sigma$ ) ...

\*also Mean Square Displacement

Ověřte normalizaci a vypočtěte střední hodnotu a varianci pro následující rozdělení:

- a) Rovnoměrné rozdělení v intervalu  $[0, 1)$ ; na počítači např. `rnd(0)`. viz mmfch5.mw

$$\int p(x)dx = \int_0^1 1dx = 1, \quad \langle x \rangle = \int_0^1 xdx = \frac{1}{2}, \quad \text{Var}(x) = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12}, \quad \sigma(x) = \sqrt{\frac{1}{12}}$$

- b) Exponenciální rozdělení ( $t$  interpretujeme jako čas):

$$p(t) = \begin{cases} ke^{-kt} & \text{pro } t > 0, \\ 0 & \text{pro } t < 0 \end{cases}$$

kde  $k$  je kladná konstanta;  $p(t)$  je hustota pravděpodobnosti, že atom se rozpadne v čase  $t$

$$\int_0^{\infty} ke^{-kt} dt = 1, \quad \langle t \rangle = \int_0^{\infty} tke^{-kt} dt \stackrel{\text{per partes}}{=} \frac{1}{k} = \text{střední doba života } \tau$$

$$\text{Var}(t) = \int_0^{\infty} (t - \tau)^2 ke^{-kt} dt = \frac{1}{k^2} = \tau^2$$

Další trik vhodný pro výpočet integrálů:  $\int_0^{\infty} e^{-kt} dt = 1/k$  derivujeme podle parametru  $k$

# Funkce náhodné veličiny

Mějme reálnou náhodnou veličinu  $x$  s rozdělením  $p(x)$  a reálnou funkcí  $f(x)$ . Veličina (pozorovatelná)  $f(x)$  má rozdělení (sčítá se přes všechny kořeny):

$$p_f(y) = \sum_{x:f(x)=y} \frac{p(x)}{|f'(x)|}$$

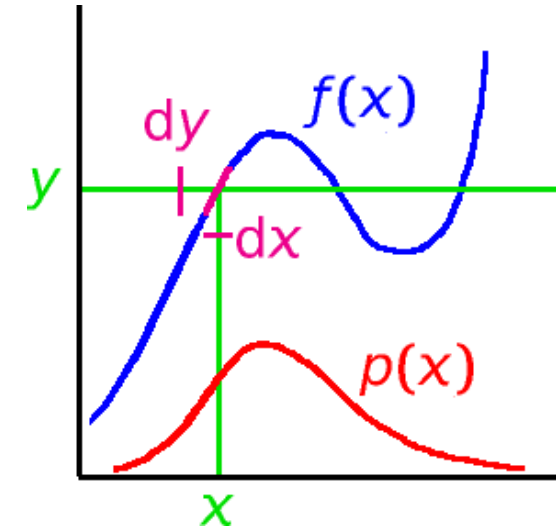
**Příklad 1.** Necht'  $x$  má rovnoměrné rozdělení v intervalu  $[0, 1)$ . Jaké rozdělení má  $y = -\ln x$ ?

```
> restart;
> with(Statistics):
> rectf := t->piecewise(t<0,0, t<1,1, 0);
> Rect := Distribution(PDF=(rectf));
> X := RandomVariable(Rect);
> Mean(X); StandardDeviation(X);
> PDF(-log(X),x);
```

$$y = f(x) \equiv -\ln x \Rightarrow x = e^{-y} \quad (1 \text{ kořen})$$

$$f'(x) = -1/x$$

$$\begin{aligned} p_f(y) &= \sum_{x:f(x)=y} \frac{p(x)}{|f'(x)|} \\ &= \frac{1}{|-1/x|} = |x| = e^{-y} \end{aligned}$$



**Příklad 2.** Necht'  $x$  má rovnoměrné rozdělení v intervalu  $[-1, 1)$ . Jaké rozdělení má  $y = \arcsin(x)$ ?

```
seq 1 1000000 | tabproc 'asin(rnd(-1))' | histogram -.031416 | plot -:x:y:o '[99]:x:cos(x)/2:-' z/(Λ)soo = (Λ)fd
```

Míra nerovnosti příjmu. Příjem  $x$  s hustotou pravděpodobnosti  $p(x)$ ,  $x \geq 0$ .

$$G = \frac{1}{2\langle x \rangle} \int_0^{\infty} p(x) dx \int_0^{\infty} p(y) dy |x - y|, \quad G \in [0, 1]$$

Jihoafrická republika: 65 % ... USA  $\approx$  Čína: 48 % ... ČR: 26 % ... Ukrajina: 25%

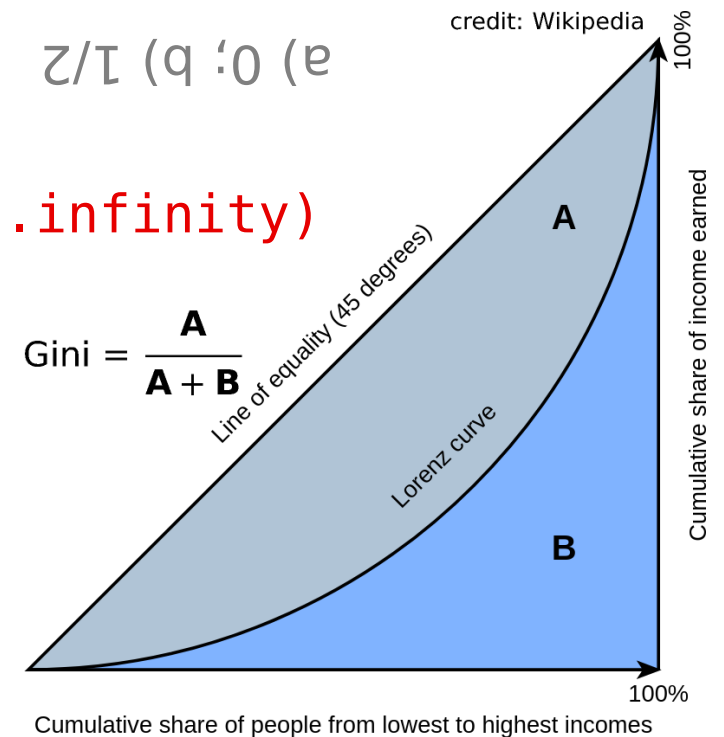
**Příklad.** Vypočtete Giniho koeficient pro:

- Diracovu delta-distribuci (všichni berou stejně),
- exponenciální rozdělení příjmů.

rozměr  $x$  je \$  
rozměr  $p(x)$  je  $\$^{-1}$



```
> restart;  
> Gini:=p->int(p(x)*int(p(y)*abs(x-y),y=0..infinity),x=0..infinity)  
    /2/int(p(x)*x,x=0..infinity)  
> assume(a>0);  
> p:=x->Dirac(x-a);  
> int(p(x),x=0..infinity);  
> Gini(p);  
> p:=x->a*exp(-x*a);  
> int(p(x),x=0..infinity);  
> Gini(p);
```



Střední hodnota funkce  $f(x)$  vzhledem k náhodné veličině  $\mathbf{x} \in \mathbb{R}$  s distribuční funkcí  $p(x)$ :

$$\langle f \rangle = \int f(x)p(x)dx \quad (1)$$

nebo z nové náhodné proměnné  $\mathbf{f} = f(\mathbf{x})$ :

$$\langle f \rangle = \int yp_f(y)dy, \quad p_f(y) = \sum_{\mathbf{x}:f(\mathbf{x})=y} \frac{p(\mathbf{x})}{|f'(\mathbf{x})|} \quad (2)$$

Obě střední hodnoty jsou stejné:

$$\langle f \rangle = \int f(x)p(x)dx \stackrel{\text{subst. } y=f(x)}{=} \int \frac{yp(x)}{f'(x)}dy = \int yp_f(y)dy$$

kde v 2. integrálu  $x =$  řešení rovnice  $f(x) = y$ , které zde pro jednoduchost uvažujeme jen jedno a také předpokládáme, že funkce  $f$  je rostoucí.

**Stat-mech příklad** v  $3N$ -dimenzionálním prostoru:  $f = P$  (tlak),  $\boldsymbol{\tau} = (\vec{r}_1, \dots, \vec{r}_N) \in \mathbb{R}^{3N}$ :

$$p(\boldsymbol{\tau}) = \frac{e^{-E(\boldsymbol{\tau})/k_B T}}{\int e^{-E(\boldsymbol{\tau})/k_B T} d\boldsymbol{\tau}}, \quad \langle P \rangle = \int P(\boldsymbol{\tau})p(\boldsymbol{\tau})d\boldsymbol{\tau}$$

Obecně a jednotně  $\langle f \rangle_\mu = \int f(x)d\mu(x)$ , kde  $\mu$  je pravděpodobnostní míra,  $x \in$  měřitelná množina.

- Kovariance  $x \in \mathbb{R}$  a  $y \in \mathbb{R}$  dvojrozměrného rozdělení  $p(x, y)$

$$\text{Cov}(\mathbf{x}, \mathbf{y}) = \langle \Delta x \Delta y \rangle = \int \Delta x \Delta y p(x, y) dx dy$$

- Kovariance dvou veličin  $f(x)$  a  $g(x)$ ; obdobně u diskrétního či vícerozměrného rozdělení ( $\mathbf{x} = \boldsymbol{\tau}$ ):

$$\text{Cov}(f, g) = \langle \Delta f \Delta g \rangle = \int \Delta f \Delta g p(\boldsymbol{\tau}) d\boldsymbol{\tau}$$

## Nezávislé náhodné veličiny

Náhodné veličiny  $\mathbf{x}$  (s rozdělením  $p_1(x)$ ) a  $\mathbf{y}$  (s rozdělením  $p_2(y)$ ):

$$p(x, y) = p_1(x)p_2(y) \tag{3}$$

V diskrétním případě (např. dva hody kostkou,  $p_{ij} = 1/36$ ):

$$p_{ij} = p_{1,i} p_{2,j}$$

Je-li  $\text{Cov}(\mathbf{x}, \mathbf{y}) = 0$  pak  $\mathbf{x}$  a  $\mathbf{y}$  jsou nekorelované, ale nemusí být nezávislé

Kovariance nezávislých náhodných veličin je nula:

$$\text{Cov}(\mathbf{x}, \mathbf{y}) = \langle \Delta x \Delta y \rangle_{\mathbf{x}+\mathbf{y}} = \int dx \int dy \Delta x p_1(x) \Delta y p_2(y) = \langle \Delta x \rangle_{\mathbf{x}} \langle \Delta y \rangle_{\mathbf{y}} = 0$$

$$r(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$

**Příklad.** Necht'  $u_1$  a  $u_2$  jsou dvě nezávislá rovnoměrná rozdělení v  $[0,1]$ . Vypočtete:

a)  $r(u_1, -u_1)$

b)  $r(u_1^2, u_1^2)$

c)  $r(u_1, u_2 + u_1)$

viz [mmpc5.mw](#) Correlation Coefficient

a)  $-1$ , b)  $1$ , c)  $1/\sqrt{2}$

```
seq 1 1000000 | tabproc "rnd(0)" "rnd(0)" | tabproc A A+B | lr
```



Necht'  $(x, y)$  jsou dvě spojité náhodné proměnné s rozdělením  $p(x, y)$ . Rozdělení součtu  $x + y$  je

$$p_{x+y}(z)dz = \int \int_{x+y \in (z, z+dz)} p(x, y) dx dy \stackrel{y:=z-x}{=} \int p(x, z-x) dx dz$$

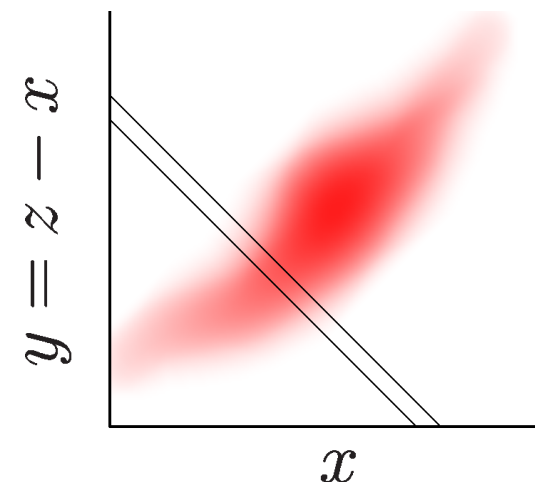
⇒

$$p_{x+y}(z) = \int p(x, z-x) dx$$

Necht' nyní  $p(x, y) = p_1(x)p_2(y)$ . Pak

$$p_{x+y}(z) = \int p_1(x)p_2(z-x) dx \equiv (p_1 * p_2)(z)$$

$p_1 * p_2$  se nazývá **konvoluce**



# Diskrétní příklad

Hodíme dvojicí kostek. Jaké rozdělení má součet ok?

$$p_2(2) = p(\ominus)p(\ominus) = 1/36$$

$$p_2(3) = p(\ominus)p(\odot) + p(\odot)p(\ominus) = 2/36$$

$$p_2(4) = p(\ominus)p(\odot\odot) + p(\odot\odot)p(\odot) + p(\odot\odot)p(\ominus) = 3/36$$

$$p_2(5) = p(\ominus)p(\odot\odot\odot) + p(\odot\odot)p(\odot\odot) + p(\odot\odot\odot)p(\odot) + p(\odot\odot\odot)p(\ominus) = 4/36$$

$$p_2(6) = p(\ominus)p(\odot\odot\odot\odot) + p(\odot\odot)p(\odot\odot\odot) + p(\odot\odot\odot)p(\odot\odot) + p(\odot\odot\odot\odot)p(\odot) + p(\odot\odot\odot\odot)p(\ominus) = 5/36$$

$$p_2(7) = p(\ominus)p(\odot\odot\odot\odot\odot) + p(\odot\odot)p(\odot\odot\odot\odot) + p(\odot\odot\odot)p(\odot\odot\odot) + p(\odot\odot\odot\odot)p(\odot\odot) + p(\odot\odot\odot\odot\odot)p(\odot) + p(\odot\odot\odot\odot\odot)p(\ominus) = 6/36$$

$$p_2(8) = p(\odot\odot)p(\odot\odot\odot\odot) + p(\odot\odot\odot)p(\odot\odot\odot\odot) + p(\odot\odot\odot\odot)p(\odot\odot\odot) + p(\odot\odot\odot\odot\odot)p(\odot\odot) + p(\odot\odot\odot\odot\odot\odot)p(\odot) = 5/36$$

$$p_2(9) = p(\odot\odot\odot)p(\odot\odot\odot\odot) + p(\odot\odot\odot\odot)p(\odot\odot\odot\odot) + p(\odot\odot\odot\odot\odot)p(\odot\odot\odot) + p(\odot\odot\odot\odot\odot\odot)p(\odot\odot) = 4/36$$

$$p_2(10) = p(\odot\odot\odot\odot)p(\odot\odot\odot\odot) + p(\odot\odot\odot\odot\odot)p(\odot\odot\odot) + p(\odot\odot\odot\odot\odot\odot)p(\odot\odot) = 3/36$$

$$p_2(11) = p(\odot\odot\odot\odot\odot)p(\odot\odot\odot) + p(\odot\odot\odot\odot\odot\odot)p(\odot\odot) = 2/36$$

$$p_2(12) = p(\odot\odot\odot\odot\odot\odot)p(\odot\odot) = 1/36$$

$$\sum_{i=2}^{12} p_2(i) = 1$$

```
seq 1 100000 | tabproc "rnd(6)+rnd(6)+2" | histogr 1.5 12.5 1 | plot -
```

# Spojité příklad

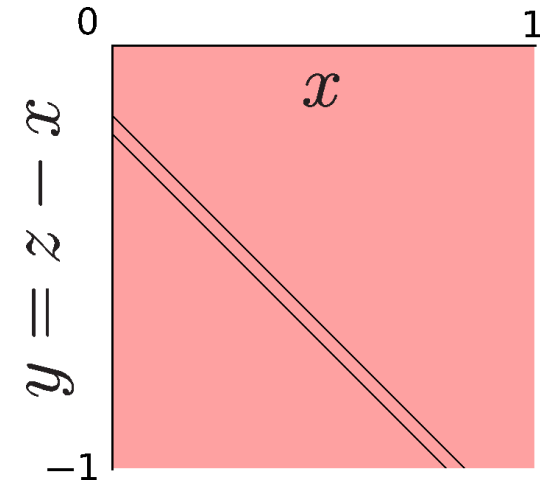
Jaké rozdělení má  $u_1 - u_2$ , jsou-li  $u_i$  nezávislá náhodná čísla v intervalu  $[0, 1)$ ?

$$p(x) = \begin{cases} 1 & \text{pro } 0 < x < 1 \\ 0 & \text{jindy} \end{cases}$$

$$p_2(z) = \int_{-\infty}^{\infty} p(x+1)p(z-x)dx = \int_{-1}^0 p(z-x)dx = \int_0^1 p(z+x)dx$$

$$= \begin{cases} \int_0^{1-z} p(z+x)dx = \int_0^{1-z} 1dx = 1-z & \text{pro } 1 > z > 0 \\ \int_0^{1+z} p(z+x)dx = \int_0^{1+z} 1dx = 1+z & \text{pro } -1 < z < 0 \\ 0 & \text{jindy} \end{cases}$$

$$p_2(z) = \begin{cases} 1 - |z| & \text{pro } |z| < 1 \\ 0 & \text{jindy} \end{cases}$$



```
seq 1 100000 | tabproc "rnd(0)-rnd(0)" | histogr -1.5 1.5 .1 | plot -
```

Střední hodnota i variance součtu nezávislých náhodných veličin jsou aditivní. Přímo z (3):

$$\begin{aligned}\langle \mathbf{x} + \mathbf{y} \rangle &= \int p_1(x)p_2(y)(x + y)dxdy \\ &= \int p_1(x)p_2(y)x dxdy + \int p_1(x)p_2(y)y dxdy = \int p_1(x)x dx + \int p_2(y)y dy = \langle \mathbf{x} \rangle + \langle \mathbf{y} \rangle\end{aligned}$$

Pomocí konvoluce distribucí:

$$\begin{aligned}\langle \mathbf{x} + \mathbf{y} \rangle &= \int zp_{\mathbf{x}+\mathbf{y}}(z)dz = \int zp_1(x)p_2(z-x)dxdz \\ &\stackrel{y:=z-x}{=} \int (x+y)p_1(x)p_2(y)dxdy = \langle \mathbf{x} \rangle_1 + \langle \mathbf{y} \rangle_2 = \langle \mathbf{x} \rangle + \langle \mathbf{y} \rangle\end{aligned}$$

A variance:

$$\text{Var}(\mathbf{x} + \mathbf{y}) = \langle (\Delta x + \Delta y)^2 \rangle_{\mathbf{x}+\mathbf{y}} = \langle (\Delta x)^2 \rangle_{\mathbf{x}+\mathbf{y}} + 2\langle \Delta x \Delta y \rangle_{\mathbf{x}+\mathbf{y}} + \langle (\Delta y)^2 \rangle_{\mathbf{x}+\mathbf{y}} = \text{Var}(\mathbf{x}) + \text{Var}(\mathbf{y})$$

# Centrální limitní věta I

Součet  $n$  stejných nezávislých rozdělení s konečnou střední hodnotou a konečnou variancí je pro velké  $n$  rovno Gaussovu rozdělení (normálnímu rozdělení) se střední hodnotou  $n\langle x \rangle$  a variancí  $n\text{Var} x$ .

Distribuční funkce normálního rozdělení se střední hodnotou  $\mu$  a směrodatnou odchylkou  $\sigma$ :

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

Raďte si ověřit, že:

$$\int p(x) dx = 1$$

$$\langle x \rangle = \int x p(x) dx = \mu$$

$$\text{Var} x = \int (x - \mu)^2 p(x) dx = \sigma^2$$

# Centrální limitní věta II

Součet  $n$  stejných nezávislých rozdělení s konečnou střední hodnotou a konečnou variancí je pro velké  $n$  rovno Gaussovu rozdělení se střední hodnotou  $n\langle x \rangle$  a variancí  $n \text{Var } x$ .

**Příklad.** Uvažujme diskrétní rozdělení  $\mathbf{b}$ :  $p(-1/2) = p(1/2) = 1/2$ . Aproximujte součet  $n$  takových rozdělení.

$$n = 1 \quad p(-1/2) = 1/2, \quad p(1/2) = 1/2, \quad \text{Var } \mathbf{b} = 1/4$$

$$n = 2 \quad p(-1) = 1/4, \quad p(0) = 1/2, \quad p(1) = 1/4, \quad \text{Var } \mathbf{b}^2 = 2/4$$

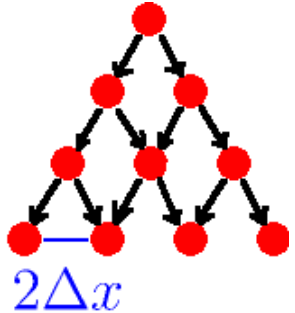
$$n = 3 \quad p(\pm 3/2) = 1/8, \quad p(\pm 1/2) = 3/8, \quad \text{Var } \mathbf{b}^3 = 3/4$$

$$\tau = 0$$

$$\tau = \Delta\tau$$

$$\tau = 2\Delta\tau$$

$$\tau = 3\Delta\tau$$



Pro jednoduchost uvažujme jen sudé  $n$ . Pak pro  $k = -n/2..n/2$ :

$$p(k) = \binom{n}{n/2 + k} 2^{-n} \approx \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{k^2}{2\sigma^2}\right), \quad \sigma^2 = \text{Var}(\mathbf{b}^n) = \frac{n}{4}$$

Důkaz: potřebujeme Stirlingův vzorec ve tvaru  $n! \approx n^n e^{-n} \sqrt{2\pi n}$ , nebo viz další stránka

Distribuční funkce normálního rozdělení se střední hodnotou  $\mu$  a směrodatnou odchylkou  $\sigma$ :

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

$$\binom{n}{\frac{n}{2} + 1} = \frac{n!}{(\frac{n}{2} - 1)! (\frac{n}{2} + 1)!} = \frac{n!}{(\frac{n}{2})! / (\frac{n}{2}) \cdot (\frac{n}{2})! (\frac{n}{2} + 1)} = \binom{n}{\frac{n}{2}} \times \frac{\frac{n}{2}}{\frac{n}{2} + 1}$$

$$\ln p(\frac{n}{2}, 1) = \ln p(\frac{n}{2}, 0) + \ln \frac{\frac{n}{2}}{\frac{n}{2} + 1} \approx \ln p(\frac{n}{2}, 0) - \frac{2}{n}$$

Další člen

$$\ln p(\frac{n}{2}, 2) = \ln p(\frac{n}{2}, 1) + \ln \frac{\frac{n}{2} - 1}{\frac{n}{2} + 2} \approx \ln p(\frac{n}{2}, 1) - \frac{6}{n}$$

a obecně

$$\ln p(n, k) \approx \ln p(n, 0) - 2 \sum_{j=1}^k \frac{2j-1}{n}, \quad \sum_{j=1}^k (2j-1) \approx \int_0^k (2k-1) dk = k(k-1) \approx k^2$$

Obdobně pro záporná  $k$ . V limitě velkých  $k$  a  $n$  tedy

$$p(n, k) \approx p(n, 0) \exp\left(-\frac{k^2}{n/2}\right)$$

Po normalizaci dostaneme kýžené

Pro náhodnou proměnnou  $\mathbf{x}$  s normálním rozdělením platí:

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt$$

$$\operatorname{prob}(|\mathbf{x} - \langle \mathbf{x} \rangle| \geq t\sigma(\mathbf{x})) = 2 \int_t^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = \operatorname{erfc}(t/\sqrt{2})$$

např.:  $\operatorname{prob}(|\mathbf{x} - \langle \mathbf{x} \rangle| \geq 2\sigma(\mathbf{x})) = 0.0455 \approx 5\%$

**Čebyšovova nerovnost:** Pro obecnou náhodnou proměnnou  $\mathbf{x}$  s konečným průměrem i variancí platí:

$$\operatorname{prob}(|\mathbf{x} - \langle \mathbf{x} \rangle| \geq t\sigma(\mathbf{x})) \leq \frac{1}{t^2}$$

např.:  $\operatorname{prob}(|\mathbf{x} - \langle \mathbf{x} \rangle| \geq 2\sigma(\mathbf{x})) = 25\%$

**Důkaz.** Definujme (jako v C/C++):  $(x \leq 1) = \begin{cases} 1 & \text{pro } x \leq 1, \\ 0 & \text{jindy.} \end{cases}$

$$\begin{aligned} \operatorname{prob}(|\mathbf{x} - \langle \mathbf{x} \rangle| \geq t\sigma(\mathbf{x})) &= \langle |\mathbf{x} - \langle \mathbf{x} \rangle| \geq t\sigma(\mathbf{x}) \rangle \\ &= \left\langle \left( \frac{\mathbf{x} - \langle \mathbf{x} \rangle}{t\sigma(\mathbf{x})} \right)^2 \geq 1 \right\rangle \leq \left\langle \left( \frac{\mathbf{x} - \langle \mathbf{x} \rangle}{t\sigma(\mathbf{x})} \right)^2 \right\rangle = \frac{1}{t^2} \end{aligned}$$

rovnost pro:  $X = \begin{cases} -1, & p = \frac{1}{2t^2} \\ 0, & p = 1 - \frac{1}{t^2} \\ +1, & p = \frac{1}{2t^2} \end{cases}$

$\operatorname{prob}(|\mathbf{x} - \langle \mathbf{x} \rangle| \leq t\sigma(\mathbf{x}))$

t	normální	nejhorší
1	68.27 %	0 %
2	95.45 %	75 %
3	99.73 %	88.89 %
5	99.999943 %	96 %



Názvosloví kolísá podle oboru...

**Statistika**, *statistic*, *estimator*, odhad, „statistický algoritmus“, (úžeji) „statistický funkcionál“, v metrologii „měřicí funkce“, *measurement function*, je vzorec/algoritmus, podle kterého počítáme výsledek z vzorku náhodných veličin (v metrologii z dat). Statistika je také náhodnou veličinou.

**Další dělení:** bodový odhad (*point estimation*): výsledkem je číslo, intervalový odhad: výsledkem je interval, kde s jistou pravděpodobností leží výsledek.

**Příklady:** aritmetický průměr, parametry modelu při fitování metodou nejmenších čtverců.

**Standardní chyba** bodové statistiky = směrodatná (standardní) odchylka (odmocnina variance) rozdělení (rozdělovací funkce) této statistiky.

**Nejistota** (*uncertainty*) v metrologii zahrnuje kritické posouzení systematických, náhodných, diskretizačních aj. chyb. Obdobně „standardní nejistota“.

**Angličtina rozlišuje:**

● estimation (whole process), statistic = estimator (formula, algorithm), estimate (final number)

● statistics = field of mathematics

Mějme **vzorek** (výběr, *sample*) náhodné veličiny

Příklady:

- velikosti bot 1000 lidí
- 100× hodíme kostkou
- tlak v průběhu simulace

Aritmetický průměr (vzorku, výběru), výběrový průměr, *sample mean*

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

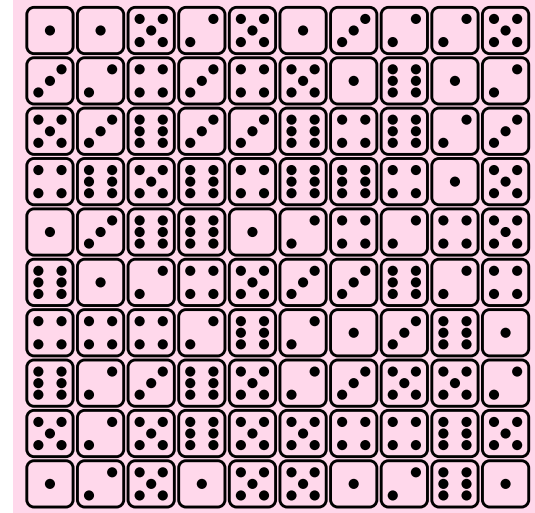
Toto je **nestranný** (nevychýlený, *unbiased*) odhad  $\langle x \rangle$ , protože

$$\langle \bar{x}_n \rangle = \langle x \rangle$$

Spočtíme varianci veličiny  $\bar{x}_n$ :

$$\text{Var}(\bar{x}_n) = \langle (\bar{x}_n - \langle x \rangle)^2 \rangle = \left\langle \left( \frac{1}{n} \sum_{i=1}^n \Delta x_i \right)^2 \right\rangle = \frac{1}{n^2} (n \text{Var} x) = \frac{\text{Var} x}{n} \equiv \frac{\sigma(x)^2}{n}, \quad \Delta x_i = x_i - \langle x \rangle$$

kde jsme předpokládali, že  $x_i$  jsou nezávislé,  $\langle \Delta x_i \Delta x_j \rangle = 0$  pro  $i \neq j$ .



pro jednoduchost  
píši  $\langle x \rangle$  místo  $\langle \mathbf{x} \rangle$

$$\sigma(x) \equiv \sqrt{\text{Var} x}$$

Jak odhadnout rozptyl  $\text{Var } x = \sigma(x)^2$ ? Neznáme střední hodnotu  $\langle x \rangle$ , ale jen její odhad,  $\bar{x}_n$ .

$$\sigma^2(x) = \langle (x - \langle x \rangle)^2 \rangle \approx \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^n \left( x_i - \frac{1}{n} \sum_{j=1}^n x_j \right)^2$$

$$= \frac{1}{n} \left[ \left( 1 - \frac{1}{n} \right) x_1 - \frac{1}{n} x_2 - \frac{1}{n} x_3 - \dots \right]^2 + \text{dalších } n-1 \text{ členů}$$

$$\left\langle \left[ \left( 1 - \frac{1}{n} \right) \Delta x_1 - \frac{1}{n} \Delta x_2 - \dots \right]^2 \right\rangle = \left[ \left( 1 - \frac{1}{n} \right)^2 + (n-1) \frac{1}{n^2} \right] \sigma(x)^2 = \frac{n-1}{n} \sigma(x)^2$$

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

● Vzorec v  $\langle \cdot \rangle$  je nestranný odhad variance  $\sigma(x)^2$

● Nestranný odhad  $\sigma(\bar{x}_n)^2$  dostaneme vydělením  $n$ :

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

**Výběrový rozptyl** (*corrected sample variance*): **Výběrový rozptyl aritmetického průměru:**

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

kde **1** = počet stupňů volnosti. Protože platí

$$\left\langle \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right\rangle = \sigma^2(x) = \text{Var } x$$

je to nestranný (nevychýlený) odhad rozptylu.

Ale odmocnina výběrového rozptylu (výběrová směrodatná odchylka) je vychýlený odhad  $\sigma(x)$ .

$$\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

jeho odmocnina je vychýleným odhadem standardní chyby aritmetického průměru

“Korekci” **-1** zavedl Friedrich Wilhelm Bessel, bez korekce máme (*uncorrected*) *sample variance*, český termín neznám.

**Poznámky k výpočtu:**

$$\begin{aligned} \langle (x - \langle x \rangle)^2 \rangle &= \langle x^2 \rangle - \langle x \rangle^2 \\ \frac{1}{n} \sum_{i=1}^n \left( x_i - \frac{1}{n} \sum_{j=1}^n x_j \right)^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2 \end{aligned} \quad \text{nevhodné, pokud } \sigma(x) \ll x_i$$

Pro zpracování nekorelovaných dat metodou aritmetického průměru, se stejnými vahami dat:

- Směrodatná (standardní) odchylka náhodné proměnné  $x$  = standardní chyba jednoho měření

$$\sigma(x) = \sqrt{\langle (x - \langle x \rangle)^2 \rangle}$$

je aproximována vzorcem

$$s_n(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

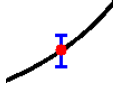
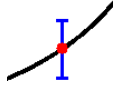
- Směrodatná (standardní) chyba aritmetického průměru z  $n$  nezávislých měření náhodné proměnné  $x$  = standardní chyba (nejistota), se kterou  $\bar{x}_n$  aproximuje  $\langle x \rangle$ , je

$$\sigma(\bar{x}_n) = \frac{\sigma(x)}{\sqrt{n}}$$

a my ji počítáme (= aproximujeme) vzorcem

$$s_n(\bar{x}_n) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Jak udávají nejistotu měřených hodnot různé obory:

- **Fyzika:**  $Q = 123.4 \pm 0.5 \equiv 123.4(5) \equiv 123.4_5$  ,  
kde  $0.5 = \sigma(Q)$  = odhadnutá směrodatná/standardní chyba/nejistota statistiky  $Q$  (např.  $Q = \bar{X}$ ) počítané z výběru (*sample*),  
**také:** standardní/směrodatná odchylka (rozumí se aritmetického průměru či jiné statistiky)  
**nepřesně jen:** (odhadnutá) chyba/nejistota, standardní/směrodatná odchylka  
**V případě normálního rozdělení** s pravděpodobností 68 % platí  $\langle Q \rangle \in 123.4 \pm 0.5$
- **Biologie, ekonomie, inženýrství, politologie, farmakologie:**  $Q = 123.4 \pm 1.0$  ,  
 $\pm 1.0 = \pm 2\sigma(Q)$  = interval spolehlivosti (*confidence interval*) na hladině (spolehlivosti) 95 %  
nepřesně jen:  $\pm 1.0$  = interval spolehlivosti, 1.0 = chyba/nejistota, ...  
**V případě normálního rozdělení** s pravděpodobností 95 % platí  $\langle Q \rangle \in 123.4 \pm 1.0$
- **Chemie:** často ignorováno; pokud udáno, tak nikdo neví, jaká je hladina spolehlivosti 🙄
- **„Fyzikální jistota“** začíná na  $\pm 5\sigma(Q)$  (hladina spolehlivosti 0.999 999 43)

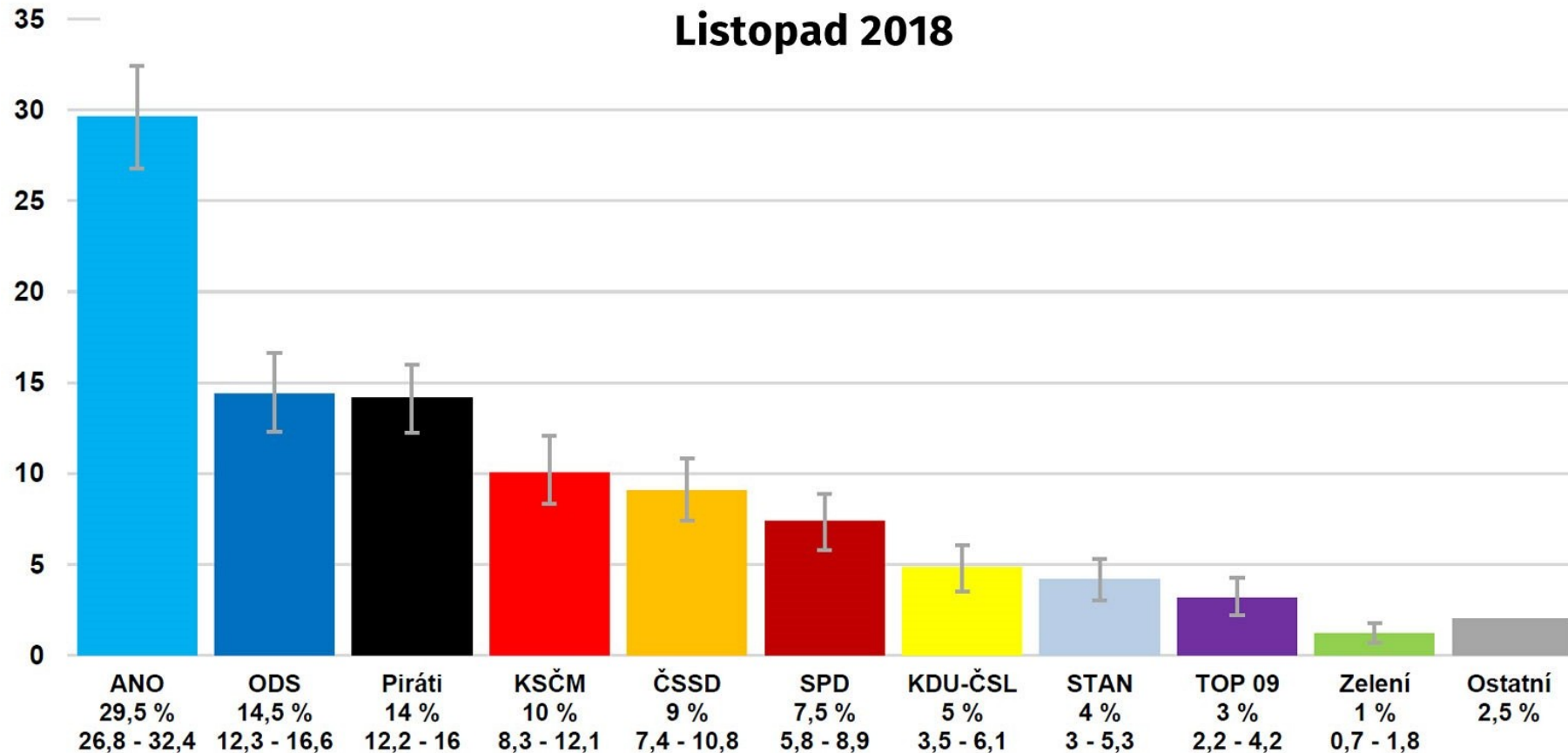
Vždy nutno udat typ chyby/nejistoty resp. hladinu spolehlivosti

$\alpha$  = hladina významnosti (*significance level*), často 5 %

$1 - \alpha$  = hladina spolehlivosti (*confidence level*), často 95 %

## Volební model podle agentury CVM

Listopad 2018



V průzkumu volebních preferencí bylo dotázáno 1080 lidí. Ve výsledcích jsou udány intervaly spolehlivosti, neznáme však použitou hladinu spolehlivosti (tj. s jakou pravděpodobností je skutečná hodnota uvnitř intervalu). Odvod'te tuto hladinu z dat.

Rada: vypočtete nejprve varianci náhodné proměnné  $x$ , která je 1 s pravděpodobností  $p$  a 0 s pravděpodobností  $1 - p$ .

$$d(1-p) \pm 1.96 \sqrt{p(1-p)}$$

$$\mathbf{x} = \begin{cases} 1 & \text{s pravděpodobností } p \\ 0 & \text{s pravděpodobností } 1 - p \end{cases}$$

$$\langle \mathbf{x} \rangle = 1 \times p + 0 \times (1 - p) = p$$

$$\text{Var } \mathbf{x} = \langle (\mathbf{x} - p)^2 \rangle = (1 - p)^2 \times p + (0 - p)^2 \times (1 - p) = p(1 - p)$$

Předpokládáme platnost centrální limitní věty. Měřená veličina je

$$P_{\text{partaj}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x} \quad (N = 1080)$$

$$\sigma^2 = \text{Var } P_{\text{partaj}} = \frac{\text{Var } \mathbf{x}}{N} = \frac{p(1 - p)}{N}$$

$$p_{\text{ANO}} = 0.295, \sigma_{\text{ANO}} = \sqrt{\frac{p(1 - p)}{N}} = 0.0139$$

$$\text{interval spolehlivosti: } \pm \frac{0.324 - 0.268}{2} = 0.028 = t\sigma, \quad t = 2.02$$

$$\text{erf}(t/\sqrt{2}) = 0.956 \approx \underline{\underline{95\%}}$$

ANO  
29,5 %  
26,8 - 32,4



**Nulová hypotéza**,  $H_0$ : Hypotéza, že vlastnost (určitá hodnota veličiny [statistiky], rozdíl. aj.) odvozená ze vzorku dat je vysvětlitelná chybou vzorkování nebo experimentálními chybami a odchylka není signifikantní: „není efekt“, „není rozdíl“, „žádná změna“, „lék je neúčinný“.

**Alternativní hypotéza**,  $H_1$ : Hypotéza, že naměřená odchylka od nulové hypotézy je signifikantní: „efekt existuje“, „lék je účinný“. Abychom ji přijali, potřebujeme dostatečně silný důvod (*evidence*), což se vyjadřuje:

- hladinou spolehlivosti (*confidence level*)  $1 - \alpha$ : alternativní hypotéza platí s pravděpodobností větší než  $1 - \alpha$ ; často  $1 - \alpha = 95\%$
- hladinou významnosti (*significance level*)  $\alpha$ ; často  $\alpha = 5\%$

## Výsledky testu:

- Zamítneme (*reject*)  $H_0$ : máme dost silné důvody pro  $H_1$ . „Lék je účinný.“ Můžeme se mýlit s pravděpodobností menší než  $\alpha$ : **chyba I. druhu**, falešně pozitivní (přijetí alternativní hypotézy, *false positive*).
- Nezamítneme (*fail to reject*)  $H_0$ : nemáme dost silné argumenty pro přijetí  $H_1$ . „Nemáme dost silné důvody k tvrzení, že lék je účinný“, „lék je asi nedostatečně účinný“. Můžeme se mýlit (**chyba II. druhu**), falešně negativní (přijetí alternativní hypotézy, *false negative*).

**Příklad:** Studentky si měří tep (PR, *pulse rate*). Ze  $n = 100$  měření jsme dostali:  $\overline{PR}_n = 73.6(9) \text{ min}^{-1}$ ; tj.  $s_n(\overline{PR}_n) = 0.9$ .

**a)** Souhlasí tento údaj s literaturou, která udává průměrnou hodnotu 72 pro ženy tohoto věku?

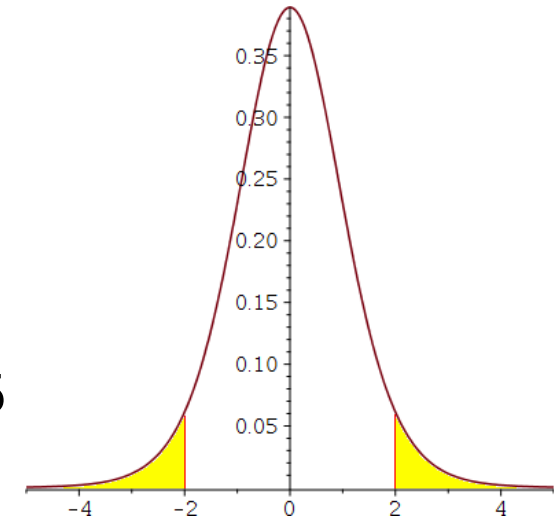
● Nulová hypotéza:  $\langle PR \rangle = 72$

● Alternativní hypotéza:  $\langle PR \rangle \neq 72$

Pro  $n = 100$  můžeme předpokládat, že rozdělení  $\overline{PR}_n$  je normální a  $s_n(\overline{PR}_n)$  je dostatečně přesné (centrální limitní věta)

$$t = \frac{\overline{PR}_n - \langle PR \rangle_{\text{null}}}{s_n(\overline{PR}_n)} = \frac{73.6 - 72}{0.9} = 1.78 \quad (\text{"1.78}\sigma\text{"})$$

$$p = 2 \int_{|x|>t} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = \text{erfc}(k/\sqrt{2}) = 0.075 > \alpha = 0.05$$



Na hladině významnosti 5 % nemůžeme zamítnout nulovou hypotézu. To, že se měření odchyluje od 72, může být náhoda. Pokud se mýlíme, je to chyba II. druhu.

Viz [mmpc5.mw Normal distribution example](#)

**Příklad:** Studenti si měří tep (PR, *pulse rate*). Ze  $n = 100$  měření jsme dostali:  $\overline{PR}_n = 73.6(9)$ ; tj.  $s_n(\overline{PR}_n) = 0.9$ .

**b)** Jsou studenti nervózní? (Platí  $PR > 72$ , kde 72 je střední hodnota je udávaná hodnota pro muže tohoto věku?)

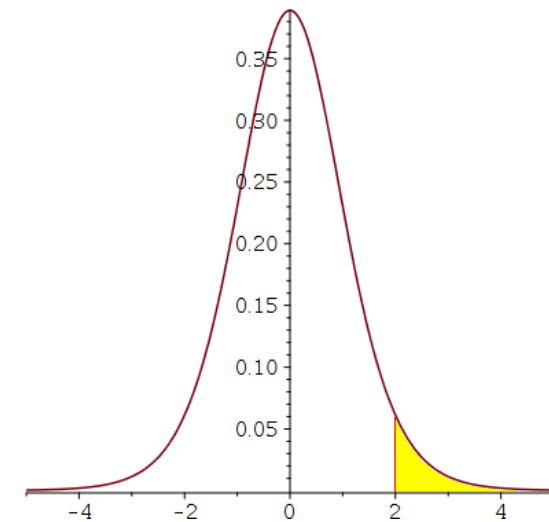
● Nulová hypotéza:  $\langle PR \rangle \leq 72$

● Alternativní hypotéza:  $\langle PR \rangle > 72$

Pro  $n = 100$  můžeme předpokládat, že rozdělení  $\overline{PR}_n$  je normální a  $s_n(\overline{PR}_n)$  je dostatečně přesné (centrální limitní věta)

$$t = \frac{\overline{PR}_n - \langle PR \rangle_{\text{null}}}{s_n(\overline{PR}_n)} = \frac{73.6 - 72}{0.9} = 1.78$$

$$p = \int_t^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} = \frac{\text{erfc}(k/\sqrt{2})}{2} = 0.038 < \alpha = 0.05$$



Na hladině významnosti 5 % nulovou hypotézu zamítneme. Studenti jsou nervózní. Pokud se mýlíme, je to chyba I. druhu. Ale spíš (na 96 %) se nemýlíme.

Ukázali jsme, že náhodná proměnná  $\bar{x}_n$  má Gaussovo rozdělení se střední hodnotou  $\langle \bar{x}_n \rangle = \langle x \rangle$  a směrodatnou odchylkou  $\sigma(\bar{x}_n) = \sqrt{\text{Var}x/n}$ . Ale známe jen jejich odhady, takže nemůžeme tvrdit, že  $\bar{x}_n$  je v mezích  $\pm$  odhadnutého  $\sigma(\bar{x}_n)$  s pravděpodobností 68 %.

Definujme Studentovo rozdělení  $t$  s parametrem  $\nu$  (počet stupňů volnosti) jako rozdělení následující náhodné proměnné:

$$\frac{\bar{x}_{\nu+1} - \langle x \rangle}{\sigma(\bar{x}_{\nu+1})}$$

Distribuční funkce je

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

Limita pro velké vzorky je normalizované normální rozdělení

$$\lim_{\nu \rightarrow \infty} t_{\nu}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Bacha,  $t_1(x)$  má nekonečný rozptyl a (striktně) nedefinovanou střední hodnotu.

$$\begin{aligned}\Gamma(x) &= \int_0^{\infty} x^{n+1} e^{-x} dx, \\ \Gamma(n) &= (n-1)!, \\ \Gamma\left(n + \frac{1}{2}\right) &= \sqrt{\pi} \cdot \frac{1}{2} \cdot \frac{3}{2} \cdots \left(n - \frac{1}{2}\right)\end{aligned}$$

Osm důchodců před odběrem vzorku ze sliznice nosohltanu mělo následující hodnoty tepu:

$$[91, 83, 67, 79, 86, 87, 72, 75]$$

Průměrný tep v tomto věku je 75. Jsou důchodci nervózní?

$$\overline{PR}_n = 80, s_n(\overline{PR}_n) = 2.91, PR = 80.0(29)$$

● Nulová hypotéza:  $\langle PR \rangle \leq 75$

● Alternativní hypotéza:  $\langle PR \rangle > 75$

$$t = \frac{\overline{PR}_n - \langle PR \rangle_{\text{null}}}{s_n(\overline{PR}_n)} = \frac{80 - 75}{2.91} = 1.719, \quad p = \int_t^{\infty} t_{n-1}(x) dx = 0.065 > 0.05$$

Závěr: Nemáme dostatečný důvod k tvrzení, že důchodci jsou nervózní

Pozn.: v případě (nesprávného) použití normálního rozdělení místo Studentova bychom dostali  $p = 0.043 < 0.05$  a tvrdili bychom neoprávněně, že důchodci jsou nervózní. Tento rozdíl mezi normálním a Studentovým výsledkem by se zvětšoval pro hladiny spolehlivosti velmi blízko 1.

Máme dvě sady měření takové, že můžeme předpokládat, že očekávané rozptyly v obou sadách jsou stejné (alespoň přibližně).

Porovnáváme 2 výběry ( $n$  a  $m$  dat) ze stejného souboru.

**Tvrzení.** Náhodná veličina

$$t = \frac{\bar{x}_n - \bar{y}_m}{s \sqrt{1/n + 1/m}}, \quad \text{kde } s^2 = \frac{(n-1)[s_n(x)]^2 + (m-1)[s_m(y)]^2}{n+m-2}$$

má Studentovo rozdělení s  $\nu = n + m - 2$ .

●  $s_n$  je výběrová směrodatná odchylka (tj. s Besselovou korekcí)

**Nulová hypotéza** (oboustranný test *two-tailed*):  $\langle x \rangle = \langle y \rangle$

**Nulová hypotéza** (jednostranný test *one-tailed*):  $\langle x \rangle > \langle y \rangle$

**Applety** např.:

● <https://stattrek.com/online-calculator/t-distribution.aspx>

● <http://www.statskingdom.com/t-student.html>

● <https://planetcalc.com/5019/>

**Excel, LibreOffice:** T.TEST(array1, array2, tails, type) vrátí  $p$

tails={1=one tail, 2=two tails}

type={1=paired, 2=two samples, equal variances, 3=two samples, unequal variances}

(Welschův  $t$ -test) Porovnáváme 2 výběry ( $n$  a  $m$  dat) z různého souboru, takže nemůžeme předpokládat rovnost variancí

**Tvrzení.** Náhodná veličina

$$t = \frac{\bar{x}_n - \bar{x}_m}{s_{\Delta}}, \quad \text{kde } s_{\Delta}^2 = \frac{[s_n(x)]^2}{n} + \frac{[s_m(x)]^2}{m}$$

má **přibližně** Studentovo rozdělení s počtem stupňů volnosti

$$\nu = \frac{\left( \frac{[s_n(x)]^2}{n} + \frac{[s_m(x)]^2}{m} \right)^2}{\frac{[s_n(x)]^4}{n^2(n-1)} + \frac{[s_m(x)]^4}{m^2(m-1)}}$$

**Nulová hypotéza:** Rovnají se střední hodnoty?

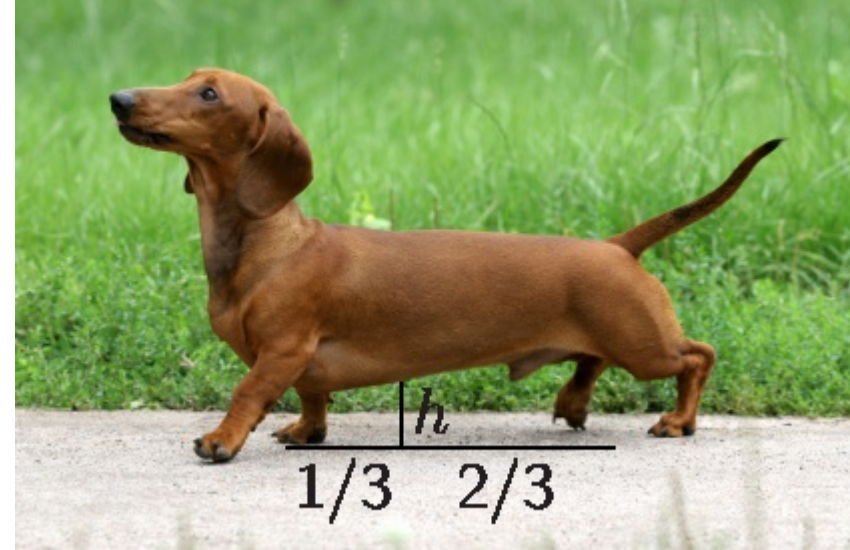
Nepoužívat  $F$ -test (zda variance dvou výběrů jsou stejné) k rozhodnutí, zda aplikovat Studentův nebo Welschův test!

Firma vyrábí podpěry pro příliš dlouhé jezevčíky. Zadala dvěma agenturám měření spodní výšky jezevčíka.

Firma SmileyDog:  $x/\text{cm} = [12.1, 20, 15.1, 20.8, 19.7]$

Firma HappyDog:  $y/\text{cm} = [18.9, 10.1, 12.1, 9.2, 12.4, 16.7, 12.7]$

- a) Jsou oba výsledky v souladu (na hladině spolehlivosti 95 %)?
- b) Jaký je nejlepší odhad výšky podpěry?



a) předpokládajíce stejně rozptily:  
 $t = 2.08, p = 0.064 \Rightarrow$  obě sady měření asi souhlasí  
(důvod odmítnout toto tvrzení není dost pádný)  
b) 15.0(12) cm



# Pilulka na COVID-19 umutuje virus k smrti

Podle J. Pazdery (OSEL): Molnupiravir snížil riziko hospitalizace nebo úmrtí přibližně o 50 %; 7,3 % pacientů, kteří dostávali molnupiravir, bylo buď hospitalizováno, nebo zemřelo do 29. dne po randomizaci (28/385), ve srovnání se 14,1 % pacientů léčených placebem (53/377);  $p = 0,0012$ . Ověřte hodnotu  $p$ .

## Problémy:

- Data jsou diskrétní  $[0,1,0,0,1,\dots]$ , kde 0 = pacient se uzdravil, 1 = pacient hospitalizován/zemřel, zatímco standardní  $t$ -test je odvozen v  $\mathbb{R}$ .
- Variance nejsou stejné, ale je mezi nimi vztah.

$p$	metoda
0.0011723	Studentův test pro diskrétní data, stejné variance
0.0012116	Studentův test pro diskrétní data, různé variance
0.0011521	Obě $\sigma$ vypočteny z binomického rozdělení, pak použijeme normální rozdělení
0.0011920	Obě $\sigma$ vypočteny z binomického rozdělení, pak Studentovo rozdělení (nejlepší ze snadno dostupných metod – jediná aproximace je „data jsou diskrétní“)
0.0011878(8)	„Přesně“ (MC simulace, viz binbin.c)