Matematická statistika

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

- diskrétní příklad: hod kostkou: $p_i = 1/6$ pro $i \in \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$
- spojitý příklad: čas rozpadu jádra: $p(t) = k\mathrm{e}^{-kt}$

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

$p(x)\mathrm{d}x$ je pravděpodobnost, že nastane jev $\boldsymbol{x} \in [x, x + \mathrm{d}x)$

Ve dvou dimenzích definujeme hustotu pravděpodobnosti $p(x, y)$ tak, že jev $x \in [z, x + \mathrm{d}x)$ a zároveň $y \in [y, y + \mathrm{d}y)$ nastane s pravděpodobností $p(x, y)\mathrm{d}x\mathrm{d}y$.

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

Normalizace:

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

Kumulativní (integrální) distribuční funkce = pravděpodobnost, že padne náhodná hodnota $\boldsymbol{x} \le x$:

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

Rozdělení pravděpodobnosti

Varování. Ve fyzice a technice nepřesně a volně zaměňujeme symbol $\boldsymbol{x}$ 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)

$$\mathrm{E}\,(\boldsymbol{x}) \equiv \langle \boldsymbol{x} \rangle \equiv \langle x \rangle_{\boldsymbol{x}} \overset{\text{volně}}{=} \langle x \rangle = \int x p(x)\mathrm{d}x \quad \text{nebo} \quad \sum_i x_i p_i$$

Variance, rozptyl, fluktuace, střední kvadratická odchylka, *mean square deviation (MSD*[*]*)*, disperze

$$\mathrm{Var}\,(\boldsymbol{x}) \overset{\text{volně}}{=} \mathrm{Var}\,x = \langle (x - \langle x \rangle)^2 \rangle = \langle \Delta x^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \quad \text{kde } \Delta x = x - \langle x \rangle$$

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

$$\sigma(\boldsymbol{x}) = \sqrt{\mathrm{Var}\,(\boldsymbol{x})}$$

též $\sigma(x)$, σ_x, δx, $s(x)$ (odhad σ) ...

[*] also Mean Square Displacement

Příklady

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`

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

b) Exponenciální rozdělení (t interpretujeme jako čas):

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

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

$$\int_0^\infty k\mathrm{e}^{-kt}\mathrm{d}t = 1, \quad \langle t \rangle = \int_0^\infty t k \mathrm{e}^{-kt}\mathrm{d}t \overset{\text{per partes}}{=} \frac{1}{k} = \text{střední doba života } \tau$$

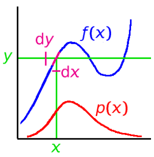
$$\mathrm{Var}\,(t) = \int_0^\infty (t - \tau)^2 k\mathrm{e}^{-kt}\mathrm{d}t = \frac{1}{k^2} = \tau^2$$

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

Funkce náhodné veličiny

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

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



Příklad 1. Nechť $\boldsymbol{x}$ 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);
```

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

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

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

Příklad 2. Nechť $\boldsymbol{x}$ 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:-'` $p_f(y) = \cos(y)/2$

Příklad: Giniho koeficient (index)

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

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

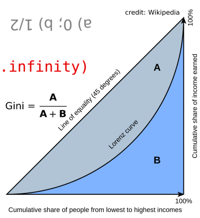
Jihoafrická republika: 65 % ... USA ≈ Čína: 48 % ... ČR: 26 % ... Ukrajina: 25 %

Příklad. Vypočtěte Giniho koeficient pro:

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

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

a) 0; b) 1/2




```
> 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);
```

Funkce náhodné veličiny: střední hodnota

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

$$\langle f \rangle = \int f(x)p(x)\mathrm{d}x \tag{1}$$

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

$$\langle f \rangle = \int y p_f(y)\mathrm{d}y, \quad p_f(y) = \sum_{x: f(x) = y} \frac{p(x)}{|f'(x)|} \tag{2}$$

Obě střední hodnoty jsou stejné:

$$\langle f \rangle = \int f(x)p(x)\mathrm{d}x \overset{\text{subst. } y = f(x)}{=} \int \frac{y p(x)}{f'(x)}\mathrm{d}y = \int y p_f(y)\mathrm{d}y$$

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), $\boldsymbol{\tau} = (\vec{r}_1, \ldots, \vec{r}_N) \in \mathbb{R}^{3N}$:

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

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

Kovariance

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

$$\mathrm{Cov}\,(\boldsymbol{x}, \boldsymbol{y}) = \langle \Delta x \Delta y \rangle = \int \Delta x \Delta y p(x, y)\mathrm{d}x\mathrm{d}y$$

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

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

Nezávislé náhodné veličiny

Náhodné veličiny $\boldsymbol{x}$ (s rozdělením $p_1(x)$) a $\boldsymbol{y}$ (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 $\mathrm{Cov}\,(\boldsymbol{x}, \boldsymbol{y}) = 0$ pak $\boldsymbol{x}$ a $\boldsymbol{y}$ jsou nekorelované, ale nemusí být nezávislé

Kovariance nezávislých náhodných veličin je nula:

$$\mathrm{Cov}\,(\boldsymbol{x}, \boldsymbol{y}) = \langle \Delta x \Delta y \rangle_{\boldsymbol{x}+\boldsymbol{y}} = \int \mathrm{d}x \int \mathrm{d}y\, \Delta x p_1(x) \Delta y p_2(y) = \langle \Delta x \rangle_{\boldsymbol{x}} \langle \Delta y \rangle_{\boldsymbol{y}} = 0$$

Korelační koeficient

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

Příklad. Nechť $\boldsymbol{u}_1$ a $\boldsymbol{u}_2$ jsou dvě nezávislá rovnoměrná rozdělení v [0,1]. Vypočtěte:

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`

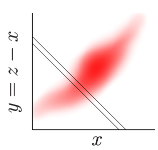
Součet náhodných proměnných

Nechť $(\boldsymbol{x}, \boldsymbol{y})$ jsou dvě spojité náhodné proměnné s rozdělením $p(x, y)$. Rozdělení součtu $\boldsymbol{x} + \boldsymbol{y}$ je

$$p_{\boldsymbol{x}+\boldsymbol{y}}(z)\mathrm{d}z = \int\int_{x+y \in (z, z+\mathrm{d}z)} p(x, y)\mathrm{d}x\mathrm{d}y \overset{y := z - x}{=} \int p(x, z - x)\mathrm{d}x\mathrm{d}z$$

$\Rightarrow$

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



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

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

$p_1 * p_2$ se nazývá **konvoluce**

Diskrétní příklad

Hodíme dvojicí kostek. Jaké rozdělení má součet ok?

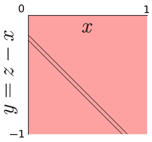
$p_2(2) = p(⚀)p(⚀) = 1/36$
$p_2(3) = p(⚀)p(⚁) + p(⚁)p(⚀) = 2/36$
$p_2(4) = p(⚀)p(⚂) + p(⚁)p(⚁) + p(⚂)p(⚀) = 3/36$
$p_2(5) = p(⚀)p(⚃) + p(⚁)p(⚂) + p(⚂)p(⚁) + p(⚃)p(⚀) = 4/36$
$p_2(6) = p(⚀)p(⚄) + p(⚁)p(⚃) + p(⚂)p(⚂) + p(⚃)p(⚁) + p(⚄)p(⚀) = 5/36$
$p_2(7) = p(⚀)p(⚅) + p(⚁)p(⚄) + p(⚂)p(⚃) + p(⚃)p(⚂) + p(⚄)p(⚁) + p(⚅)p(⚀) = 6/36$
$p_2(8) = p(⚁)p(⚅) + p(⚂)p(⚄) + p(⚃)p(⚃) + p(⚄)p(⚂) + p(⚅)p(⚁) = 5/36$
$p_2(9) = p(⚂)p(⚅) + p(⚃)p(⚄) + p(⚄)p(⚃) + p(⚅)p(⚂) = 4/36$
$p_2(10) = p(⚃)p(⚅) + p(⚄)p(⚄) + p(⚅)p(⚃) = 3/36$
$p_2(11) = p(⚄)p(⚅) + p(⚅)p(⚄) = 2/36$
$p_2(12) = p(⚅)p(⚅) = 1/36$

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

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

Spojitý příklad

Jaké rozdělení má $\boldsymbol{u}_1 - \boldsymbol{u}_2$, jsou-li $\boldsymbol{u}_i$ nezávislá náhodná čísla v intervalu $[0,1)$?



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

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

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

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

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


Součet nezávislých náhodných proměnných

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

$$\langle \boldsymbol{x}+\boldsymbol{y}\rangle = \int p_1(x)p_2(y)(x+y)\mathrm{d}x\mathrm{d}y$$

$$= \int p_1(x)p_2(y)x\mathrm{d}x\mathrm{d}y + \int p_1(x)p_2(y)y\mathrm{d}x\mathrm{d}y = \int p_1(x)x\mathrm{d}x + \int p_2(y)y\mathrm{d}y = \langle \boldsymbol{x}\rangle + \langle \boldsymbol{y}\rangle$$

Pomocí konvoluce distribucí:

$$\langle \boldsymbol{x}+\boldsymbol{y}\rangle = \int z p_{\boldsymbol{x}+\boldsymbol{y}}(z)\mathrm{d}z = \int z p_1(x)p_2(z-x)\mathrm{d}x\mathrm{d}z$$

$$\overset{y:=z-x}{=} \int (x+y)p_1(x)p_2(y)\mathrm{d}x\mathrm{d}y = \langle x\rangle_1 + \langle y\rangle_2 = \langle \boldsymbol{x}\rangle + \langle \boldsymbol{y}\rangle$$

A variance:

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

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 $n\,\mathrm{Var}\,x$.

Distribuční funkce normálního rozdělení se střední hodnotou μ a směrodatnou odchylkou σ:

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

Račte si ověřit, že:

$$\int p(x)\mathrm{d}x = 1$$

$$\langle x\rangle = \int x p(x)\mathrm{d}x = \mu$$

$$\mathrm{Var}\,x = \int (x-\mu)^2 p(x)\mathrm{d}x = \sigma^2$$

Centrální limitní věta II

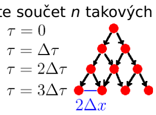
Součet n stejných nezávislých rozdělení s konečnou střední hodnotou a konečnou variancí je pro velké n rovno Gaussovu rozdělení se střední hodnotou $n\langle x\rangle$ a variancí $n\,\mathrm{Var}\,x$.

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

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

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

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



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

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

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 μ a směrodatnou odchylkou σ:

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

Ověření centrální limitní věty z binomického rozdělení

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

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

Další člen

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

a obecně

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

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

Gaussovo rozdělení a Čebyšovova nerovnost

Pro náhodnou proměnnou $\boldsymbol{x}$ s normálním rozdělením platí: $\quad \mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\mathrm{d}t$

$$\mathrm{prob}\Big(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\Big) = 2\int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{d}x = \mathrm{erfc}(t/\sqrt{2})$$

např.: $\mathrm{prob}\Big(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge 2\sigma(\boldsymbol{x})\Big) = 0.0455 \approx 5\,\%$

$\mathrm{prob}\Big(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \le t\sigma(\boldsymbol{x})\Big)$

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

Čebyšovova nerovnost: Pro obecnou náhodnou proměnnou $\boldsymbol{x}$ s konečným průměrem a variancí platí:

$$\mathrm{prob}\Big(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\Big) \le \frac{1}{t^2}$$

např.: $\mathrm{prob}\Big(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge 2\sigma(\boldsymbol{x})\Big) = 25\,\%$

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

$$\mathrm{prob}\Big(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\Big) = \Big\langle |\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\Big\rangle$$

$$= \left\langle\left(\frac{\boldsymbol{x}-\langle\boldsymbol{x}\rangle}{t\sigma(\boldsymbol{x})}\right)^2 \ge 1\right\rangle \le \left\langle\left(\frac{\boldsymbol{x}-\langle\boldsymbol{x}\rangle}{t\sigma(\boldsymbol{x})}\right)^2\right\rangle = \frac{1}{t^2}$$

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

Matematická statistika a metrologie

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.

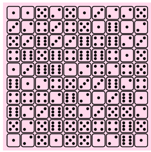
Nejistota (*uncertainty*) v metrologii zahrnuje kritické posouzení systematických, náhodných, diskretizačních aj. chyb. Obdobně „standardní nejistota".

Angličtina rozlišuje:

- estimation (whole process), statistic = estimator (formula, algorithm), estimate (final number)
- statistics = field of mathematics

Aritmetický průměr jako příklad statistiky

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*

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

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

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

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

$\sigma(x) \equiv \sqrt{\mathrm{Var}\,x}$

Spočtěme varianci veličiny $\overline{x}_n$:

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

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

Směrodatná (standardní) odchylka jako příklad statistiky

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

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

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

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

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

- Vzorec v $\langle\cdot\rangle$ je nestranný odhad variance $\sigma(x)^2$
- Nestranný odhad $\sigma(\overline{x}_n)^2$ dostaneme vydělením n: $\quad \sigma(\overline{x}_n) \approx s_n(\overline{x}_n) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}$

Směrodatná (standardní) odchylka jako příklad statistiky

Výběrový rozptyl (*corrected sample variance*):

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

kde 1 = počet stupňů volnosti. Protože platí

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

je to nestranný (nevychýlený) odhad rozptylu.

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

Výběrový rozptyl aritmetického průměru:

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

jeho odmocnina je vychýleným odhadem standardní odchylky aritmetického průměru
"Korekci" -1 zavedl Friedrich Wilhelm Bessel, bez korekce máme (*uncorrected*) *sample variance*, český termín neznám.

Poznámky k výpočtu:

$$\Big\langle(x-\langle x\rangle)^2\Big\rangle = \langle x^2\rangle - \langle x\rangle^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2$$

nevhodné, pokud $\sigma(x) \ll x_i$

Souhrn

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
$$s_n(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$
- Směrodatná (standardní) chyba aritmetického průměru z n nezávislých měření náhodné proměnné x = standardní chyba (nejistota), se kterou $\overline{x}_n$ approximuje $\langle x \rangle$, je
$$\sigma(\overline{x}_n) = \frac{\sigma(x)}{\sqrt{n}}$$
a my ji počítáme (= aproximujeme) vzorcem
$$s_n(\overline{x}_n) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$

Zvyky a zlozvyky

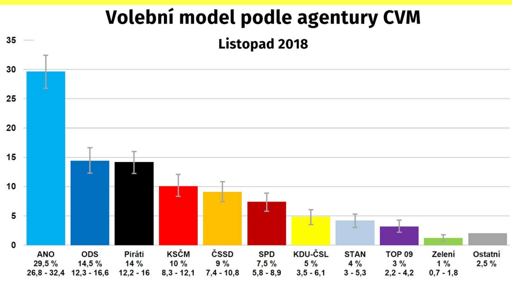
Jak udávají nejistotu měřených hodnot různé obory:

- **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. $Q = \overline{X}$) 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: ± 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

α = hladina významnosti (*significance level*), často 5 %
$1 - \alpha$ = hladina spolehlivosti (*confidence level*), často 95 %

Příklad



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). Odvoďte tuto hladinu z dat.

Rada: vypočtěte nejprve varianci náhodné proměnné $\mathbf{x}$, která je 1 s pravděpodobností p a 0 s pravděpodobností $1 - p$.

$p(1-p)$; 95 %

Řešení příkladu

$$\mathbf{x} = \begin{cases} 1 & \text{s pravděpodobností } p \\ 0 & \text{s pravděpodobností } 1-p \end{cases}$$
$$\langle \mathbf{x} \rangle = 1 \times p + 0 \times (1-p) = p$$
$$\mathrm{Var}\,\mathbf{x} = \langle (\mathbf{x} - p)^2 \rangle = (1-p)^2 \times p + (0-p)^2 \times (1-p) = p(1-p)$$

Předpokládáme platnost centrální limitní věty. Měřená veličina je

$$P_{\text{partaj}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x} \quad (N = 1080)$$
$$\sigma^2 = \mathrm{Var}\,P_{\text{partaj}} = \frac{\mathrm{Var}\,\mathbf{x}}{N} = \frac{p(1-p)}{N}$$
$$p_{\text{ANO}} = 0.295, \sigma_{\text{ANO}} = \sqrt{\frac{p(1-p)}{N}} = 0.0139$$
$$\text{interval spolehlivosti: } \pm\frac{0.324 - 0.268}{2} = 0.028 = t\sigma, \ \ t = 2.02$$
$$\mathrm{erf}(t/\sqrt{2}) = 0.956 \approx \underline{95\,\%}$$

Testování hypotéz

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*) α; 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 α: **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 1 – oboustranný odhad

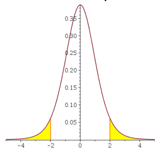
Příklad: Studentky si měří tep (PR, *pulse rate*). Ze $n = 100$ měření jsme dostali: $\overline{\mathrm{PR}}_n = 73.6(9)\,\mathrm{min}^{-1}$; tj. $s_n(\overline{\mathrm{PR}}_n) = 0.9$.

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

- Nulová hypotéza: $\langle \mathrm{PR} \rangle = 72$
- Alternativní hypotéza: $\langle \mathrm{PR} \rangle \neq 72$

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

$$t = \frac{\overline{\mathrm{PR}}_n - \langle \mathrm{PR} \rangle_{\text{null}}}{s_n(\overline{\mathrm{PR}}_n)} = \frac{73.6 - 72}{0.9} = 1.78 \quad (\text{“}1.78\sigma\text{”})$$
$$p = 2\int_{|x|>t}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = \mathrm{erfc}(k/\sqrt{2}) = 0.075 > \alpha = 0.05$$



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 2 – jednostranný odhad

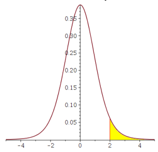
Příklad: Studenti si měří tep (PR, *pulse rate*). Ze $n = 100$ měření jsme dostali: $\overline{\mathrm{PR}}_n = 73.6(9)$; tj. $s_n(\overline{\mathrm{PR}}_n) = 0.9$.

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

- Nulová hypotéza: $\langle \mathrm{PR} \rangle \le 72$
- Alternativní hypotéza: $\langle \mathrm{PR} \rangle > 72$

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

$$t = \frac{\overline{\mathrm{PR}}_n - \langle \mathrm{PR} \rangle_{\text{null}}}{s_n(\overline{\mathrm{PR}}_n)} = \frac{73.6 - 72}{0.9} = 1.78$$
$$p = \int_t^{\infty}\frac{e^{-x^2/2}}{\sqrt{2\pi}} = \frac{\mathrm{erfc}(k/\sqrt{2})}{2} = 0.038 < \alpha = 0.05$$



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

Studentovo t-rozdělení

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

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

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

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

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

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

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

Opět tep

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?

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

- Nulová hypotéza: $\langle \mathrm{PR} \rangle \le 75$
- Alternativní hypotéza: $\langle \mathrm{PR} \rangle > 75$

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

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.

Porovnání dvou výběrů – stejné variance

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

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

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)` vrací p
tails={1=one tail, 2=two tails}
type={1=paired, 2=two samples, equal variances, 3=two samples, unequal variances}

Porovnání dvou výběrů – různé variance

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

Tvrzení. Náhodná veličina

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

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

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

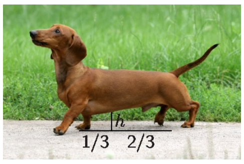
Nulová hypotéza: Rovnají se střední hodnoty?

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

Příklad (viz mmpc5.mw)

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) předpokládající stejné rozptyly:
$t = 2.08$, $p = 0.064 \Rightarrow$ 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,...], 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 σ vypočteny z binomického rozdělení, pak použijeme normální rozdělení
0.0011920	Obě σ 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)