PŘEHLED HLAVNÍCH POJMŮ A VZORCŮ - KMA/PSE

Vysvětlení, použití, grafy, příklady, etc. budou na hodinách KMA/PSE .

21.9.2018, Z.Kobeda

****

ELEMENTÁRNÍ POČET PRAVDĚPODOBNOSTI

NÁHODNÝ JEV

Náhodný pokus - aspoň dva různé výsledky

Náhodné jevy - podmnožiny množiny Ω všech možných výsledků nějakého pokusu

Označení: $A \subset \Omega$,$B \subset \Omega, \ldots$ $\emptyset$ $\ldots$ nemožný jev Ω $\ldots$ jistý jev

Operace s jevy:

- sjednocení dvou jevů: $A \cup B$ (A nebo B)
- průnik dvou jevů: $A \cap B$ (A a B)
- negace jevu A: $\overline{A}$ (jev opačný neboli doplňkový k A)

Neslučitelné (disjunktní) jevy: Jevy A, B se nazývají neslučitelné, je-li $A \cap B = \emptyset$.

PRAVDĚPODOBNOST JEVU

$$A \longrightarrow P(A)$$

A...jev, P(A)...**pravděpodobnost (ppst) jevu A**

Axiomy ppsti:

A_1: $0 \leq P(A) \leq 1$;

A_2: pro neslučitelné jevy $A_1, A_2, \ldots$ platí: $P(\bigcup_i A_i) = \sum_i P(A_i)$;

A_3: $P(\Omega) = 1$, kde Ω je jev jistý?.

Statistická "definice"ppsti

Ppst $P(A)$ jevu A je limita relativní četnosti jevu A, zvětšujeme-li počet pokusů $n \to \infty$.

Klasická definice ppsti:

$$P(A) = \frac{N_A}{N} \quad .$$

Další možné definice ppsti : "geometrická", "axiomatická".

Základní pravidla pro ppst (lze je odvodit z axiomů A_1, A_2, A_3):

- $A \subset B \Rightarrow P(A) \leq P(B)$
- $P(\overline{A}\,) = 1 - P(A)$ (**ppst opačného jevu**)
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (**ppst sjednocení jevů**)

Nezávislost jevů: Jevy A, B jsou **nezávislé**, právě když $P(A \cap B) = P(A) \cdot P(B)$
(Podobně ppst průniku většího počtu jevů počítáme součinem, jde-li o jevy nezávislé.)

Ppst jevu A podmíněná jevem B: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (pro $P(B) \neq 0$) .

• Nechť pro jevy $B_1, B_2, \ldots, B_n$ platí: $B_i \cap B_j = \emptyset \ \ \forall\, i \neq j$, $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$, $P(B_i) > 0$ $\forall\, i = 1, 2, \ldots, n$, a nechť A je libovolný jev (tj. $A \subset \Omega$). Pak platí:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

(tzv. **věta o úplné ppsti**)

• Je-li navíc $P(A) > 0$, pak pro $k = 1, 2, \ldots, n$ platí:

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{P(A)}$$

(tzv. **Bayesova věta**, věta o inverzní ppsti)

****

NÁHODNÁ VELIČINA

Náhodná veličina (náhodná proměnná) je "vhodná" reálná fce definovaná na množině Ω

Označení: X, Y, Z ... **náhodné veličiny** (jsou to funkce)

$x,\ y,\ z\ \ldots$ jejich **realizace** (jsou to reálná čísla)

U náh. veličiny nutno určit též její **rozdělení ppsti**.

DISKRÉTNÍ náh. veličina: existuje konečná nebo spočetná množina reálných čísel $x_1, x_2, \ldots$ taková, že: $P(X = x_j) > 0 \ \ pro \ \ j = 1, 2, \ldots$ a $\sum_{x_j} P(x_j) = 1$

$P(X = x)$ (krátce $P(x)$) $\ldots$ **pravděpodobnostní funkce**

SPOJITÁ náh.veličina: **hustota ppsti** $\ldots$ fce $f(x)$, k níž se "blíží" histogram relativních četností při zjemňujícím se dělení.

Zřejmě: 1) $f(x) \geq 0$ pro $x \in (-\infty, +\infty)$ 2) $\int\limits_{-\infty}^{+\infty} f(x)\, dx = 1$

Popsat rozdělení ppsti náh. veličiny X (**diskr.** nebo **spoj.**) lze též pomocí tzv. **distribuční funkce** $F(x)$:

$$\boxed{F(x) = P(X \le x) \quad \text{pro} \quad x \in (-\infty, +\infty)}$$

- $0 \le F(x) \le 1$ pro $x \in (-\infty, +\infty)$;
- $F(x)$ je **neklesající** funkce ;
- $\lim\limits_{x \to -\infty} F(x) = 0$ a $\lim\limits_{x \to +\infty} F(x) = 1$.

Distribuční fce **diskrétní** náh. veličiny je **nespojitá**, je "schodovitá", tj. neklesající a po částech konstantní. Distribuční fce **spojité** náh. veličiny je **spojitá**.

STŘEDNÍ HODNOTA, píšeme $E(X)$:

Pro **diskrétní** náh. veličinu X : $\boxed{E(X) = \sum\limits_i x_i \cdot P(x_i)}$, kde $P(x)$ je ppstní fce.

Pro **spojitou** náhodnou veličinu X: $\boxed{E(X) = \int\limits_{-\infty}^{+\infty} x \cdot f(x)\, dx}$, kde $f(x)$ je hustota ppsti.

- Jsou-li $X_1, X_2, \ldots, X_n$ náh. veličiny, pak: $E(\sum\limits_{i=1}^{n} X_i) = \sum\limits_{i=1}^{n} E(X_i)$

Rozptyl: $\boxed{D(X) = E(X - E(X))^2}$

Roznásobení pravé strany a úprava $\Rightarrow$ tzv. **výpočetní tvar**: $D(X) = E(X^2) - E^2(X)$
kde $E^2(X) = (E(X))^2$ a kde $E(X^2) = \sum\limits_i x_i^2 \cdot P(x_i)$ pro diskr.náh.veličinu,

$$E(X^2) = \int\limits_{-\infty}^{+\infty} x \cdot f(x)\, dx \quad \text{pro spoj.náh.veličinu.}$$

- Jsou-li $X_1, X_2, \ldots, X_n$ *nezávislé* náh.veličiny, pak: $D(\sum\limits_{i=1}^{n} X_i) = \sum\limits_{i=1}^{n} D(X_i)$

Směrodatná odchylka: $\boxed{\sigma(X) = \sqrt{D(X)}}$

NĚKTERÁ DISKRÉTNÍ ROZDĚLENÍ

X ... **diskrétní náhodná veličina**

• **ALTERNATIVNÍ rozdělení** s parametrem $p \in (0,1)$: X nabývá jen hodnot 0 nebo 1,

přičemž $\boxed{P(0) = 1 - p\,, \quad P(1) = p}$ Píšeme: $X \sim A(p)$

Výpočtem: $E(X) = p, \quad D(X) = p\,(1-p)$

• **BINOMICKÉ rozdělení** s parametry $n \in \mathrm{N},\ p \in (0,1)$: X může nabývat pouze hodnot

$0, 1, \ldots, n$ a platí: $\boxed{P(k) = \binom{n}{k}\, p^k\, (1-p)^{n-k} \quad \text{pro } k = 0, 1, \ldots, n}$ Píšeme: $X \sim Bi(n,p)$

$Bi(n,p)$ je součtem n nezávislých veličin s rozd. $A(p)$, takže: $E(X) = n\,p\,,\ D(X) = n\,p\,(1-p)\,.$

• **HYPERGEOMETRICKÉ rozdělení** s param. $N, K, n \in \mathbf{N}, n \leq N, K \leq N:$

$\boxed{P(k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}}$ pro všechna $k \in \mathbf{N} \cup \{0\}$ taková, že $k \leq K,\ 0 \leq n - k \leq N - K$

Píšeme: $X \sim HG(N, K, n)\,.$

• **POISSONOVO rozdělení** s parametrem $\lambda > 0$: může nabývat jen hodnot $k = 0, 1, 2, \ldots$

a platí $\boxed{P(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!} \quad \text{pro} \quad k = 0, 1, 2, \ldots\,.}$ Píšeme: $X \sim Po(\lambda)$

Výpočtem: $E(X) = \lambda\,, \quad D(X) = \lambda$

$X_i \sim Po(\lambda_i)$, $i = 1, 2, \ldots, n$ **nezávislé**, $X = \sum\limits_{i=1}^{n} X_i \Longrightarrow X \sim Po(\lambda)$, kde $\lambda = \sum\limits_{i=1}^{n} \lambda_i$.

Pro $n \geq 30$ a $p \leq 0{,}1$ je $Bi(n,p) \approx Po(\lambda)$, kde $\lambda = n \cdot p$

Funkční hodnoty distribuční funkce $F(x)$ Poissonova rozdělení bývají **tabelovány** pro některá $\lambda \leq 10$. Pro $\lambda \geq 9$ používáme aproximaci normálním rozdělením.

NĚKTERÁ SPOJITÁ ROZDĚLENÍ

X ... **spojitá náhodná veličina**

• **ROVNOMĚRNÉ rozdělení** na intervalu (a,b) : X má hustotu ppsti
$$\boxed{f(x)=\begin{cases} \dfrac{1}{b-a} & \text{pro } x \in (a,b) \\ 0 & \text{jinde.} \end{cases}}$$

Píšeme: $X \sim R(a,b)$. Platí: $E(X) = \frac{a+b}{2}$, $D(X) = \frac{(b-a)^2}{12}$.

• **EXPONENCIÁLNÍ rozd.** s parametrem $\delta > 0$: X má hustotu
$$\boxed{f(x)=\begin{cases} 0 & \text{pro } x \in (-\infty, 0] \\ \dfrac{1}{\delta}\, e^{-\frac{x}{\delta}} & \text{pro } x \in (0, +\infty) \end{cases}}$$

Píšeme: $X \sim Exp(\delta)$. *Distribuční funkce* : $F(x) = \begin{cases} 0 & \text{pro } x \in (-\infty, 0] \\ 1 - e^{-\frac{x}{\delta}} & \text{pro } x \in (0, +\infty) \end{cases}$

Platí: $E(X) = \delta$, $D(X) = \delta^2$.

• **NORMÁLNÍ rozdělení** s parametry $\mu \in \mathrm{R}$, $\sigma^2 > 0$: X má pro $x \in (-\infty, +\infty)$ hustotu

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

Píšeme: $X \sim N(\mu, \sigma^2)$ Výpočtem: $E(X) = \mu$, $D(X) = \sigma^2$.

Používá se též název **GAUSSOVO rozdělení**.

$N(0,1)$ . . . **normované normální rozdělení**, hustota: $\varphi(u) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}}$, $u \in (-\infty, +\infty)$

její graf je symetrický kolem přímky $u = 0$, a proto **distribuční funkce** - píšeme $\Phi(u)$ - je **tabelována** jen pro $u \geq 0$, neboť platí: $\boxed{\Phi(-u) = 1 - \Phi(u)}$

Je-li $F(x)$ distribuční funkce rozdělení $N(\mu, \sigma^2)$, pak $\boxed{F(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)}$

APROXIMACE NORMÁLNÍM ROZDĚLENÍM

• Je-li $n \in N$ tak velké, že $\boxed{n\,p\,(1-p) \geq 9}$, pak $\boxed{Bi(n,p) \approx N(np,\, np\,(1-p))}$

• Je-li $\boxed{\lambda \geq 9}$, lze použít aproximaci: $\boxed{Po(\lambda) \approx N(\lambda,\, \lambda\,)}$

• Jsou-li $X_1, X_2, \ldots, X_n$ **nezávislé** náhodné veličiny, se stejným rozdělením,

$E(X_i) = \mu_0$, $D(X_i) = \sigma_0^2$, pak pro "dost velké n" platí: $\boxed{\sum\limits_{i=1}^{n} X_i \approx N(n\mu_0,\, n\sigma_0^2\,)}$

Odtud pak: $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i \approx N\left(\mu_0,\, \frac{\sigma_0^2}{n}\,\right)$

Mají-li $X_1, X_2, \ldots, X_n$ navíc normální rozdělení, má i jejich součet a průměr normální rozdělení.

KVANTILY SPOJITÝCH ROZDĚLENÍ

Nechť X je *spojitá* náhodná veličina, $p \in (0,1)$.

Reálné číslo x_p se nazývá **100 p%-NÍ KVANTIL**, je-li $\boxed{P(X \leq x_p) = p}$

$x_{0,5}$ ... **medián** $x_{0,25}$ ... **dolní kvartil** $x_{0,75}$ ... **horní kvartil**

Pomocí distribuční funkce $F(x)$ lze definici kvantilu x_p přepsat do tvaru $\boxed{F(x_p) = p}$

Určení kvantilů normálního rozdělení

Kvantily x_p rozdělení $N(\mu, \sigma^2)$ lze vyjádřit pomocí kvantilů u_p rozdělení $N(0,1)$, neboť pro $p \in (0,1)$ platí $\boxed{x_p = \mu + \sigma \cdot u_p}$

K určení u_p používáme **tabulky**, pro $p < 0,5$ navíc rovnost $\boxed{u_p = -u_{\,1-p}}$

[tedy např. $u_{0,05} = -u_{0,95}$] která platí, neboť hustota $\varphi(u)$ je sudá funkce.

ÚVOD DO MATEMATICKÉ STATISTIKY

POPISNÁ STATISTIKA

(statistický) **soubor**: $x_1, x_2, \ldots, x_n$ [$\approx$ hodnoty (diskr. nebo spoj.) náh. veličiny]

$x_i \in R \ldots$ prvek souboru $n \in N \ldots$ rozsah souboru

uspořádaný soubor: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

Charakteristiky polohy:

- **(aritmetický) průměr**: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- **medián** $\tilde{x}$ (uspořádaného) souboru je jeho prostřední hodnota (je-li n liché), resp. aritm. průměr dvou prostředních hodnot (je-li n sudé)
- **modus** $\hat{x}$ je hodnota(-y) s nejvyšší četností

Charakteristiky variability:

- **rozpětí** je rozdíl mezi největší a nejmenší hodnotou, tj. $x_{(n)} - x_{(1)}$
- **(výběrový) rozptyl**: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- **(výběrová) směrodatná odchylka**: $s = \sqrt{s^2}$

ODHADY PARAMETRŮ

náhodný výběr rozsahu n z rozdělení náhodné veličiny X je posloupnost *nezávislých* náh. veličin $X_1, X_2, \ldots, X_n$, které mají stejné rozdělení jako X

$x_1, x_2, \ldots, x_n$ je tzv. **realizace** náh. výběru (= statistický soubor)

neznámou hodnotu parametru v rozdělení X nahrazujeme vhodným reálným číslem, spočteným z realizace náh. výběru - tzv. **bodový odhad** parametru:

- $E(X) \approx \bar{x}$, např.: $p \approx \bar{x}$ pro $X \sim A(p)$, $\lambda \approx \bar{x}$ pro $X \sim Po(\lambda)$, $\mu \approx \bar{x}$ pro $X \sim N(\mu, \sigma^2)$.
- $D(X) \approx s^2$, např.: $\sigma^2 \approx s^2$ pro $X \sim N(\mu, \sigma^2)$

intervalové odhady (neznámého) parametru θ rozdělení náh. veličiny X:

pro $\alpha \in (0, 1)$ [často bývá $\alpha = 0.01,\ 0.05$ nebo 0.10]

interval (a, b) nazveme **100 (1−α) %-ní** (oboustranný) **interval spolehlivosti**, je-li $P(a < \theta < b) = 1 - \alpha$ (a, resp. b je tzv. dolní, resp. horní mez spolehlivosti)

Např.: Je-li $x_1, x_2, \ldots, x_n$ realizace náhodného výběru z rozdělení $X \sim N(\mu, \sigma^2)$, kde $n > 30$, pak $100\,(1-\alpha)$%-ní **interval spolehlivosti pro parametr** μ je:

$$\left(\bar{x} - u_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \,,\ \bar{x} + u_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right)$$

kde $u_{1-\frac{\alpha}{2}}$ jsou kvantily rozdělení $N(0, 1)$.

TESTOVÁNÍ HYPOTÉZ

Použitím naměřených hodnot testujme určitou (statistickou) hypotézu, označme ji H_0.

Náš závěr bude vždy jeden z následujících dvou: 1. Zamítáme H_0. 2. Nezamítáme H_0.

Přitom se můžeme dopustit chyb dvou druhů:

chyba 1. druhu: H_0 je pravdivá, ale test vede k zamítnutí H_0;

chyba 2. druhu: H_0 je nepravdivá, ale test vede k nezamítnutí H_0.

Ppst chyby 1. druhu se značí α, požaduje se malá a volí se před testem (0.05, 0.01, nebo 0.1)
$\alpha\ldots$ tzv. **hladina významnosti** Ppst chyby 2. druhu se označuje β (závisí na volbě α).

Rozhodnutí o případném zamítnutí H_0 provádíme pomocí testu zvaného **testové kritérium**.
Množina všech hodnot test. kritéria vedoucí k zamítnutí H_0, se nazývá **kritický obor**, označuje se W. Hodnota, která odděluje W od hodnot, které vedou k nezamítnutí H_0, je tzv. **kritická hodnota**.

Postup:

1) Stanovíme H_0 - tzv. **nulová hypotéza** (H_0 musí obsahovat rovnost).

2) Stanovíme H_1 - tzv. **alternativní hypotéza** (H_1 je negací H_0).

3) Zvolíme **hladinu významnosti** α.

4) Vybereme testové kritérium, určíme jeho rozdělení ppsti za předpokladu platnosti H_0. Toto rozdělení, hladina významnosti a formulace H_1 určují **kritický obor** W. Načrtneme graf rozdělení test. kritéria, vyznačíme W.

5) Hodnota test. kritéria $\in W \Rightarrow$ **zamítáme** H_0.
Hodnota test. kritéria $\notin W \Rightarrow$ **nezamítáme** H_0.

Známe-li typ rozdělení, z něhož pocházejí hodnoty v souboru, a testujeme jen neznámé hodnoty parametrů, říkáme, že testujeme tzv. **PARAMETRICKÉ HYPOTÉZY**, např.:

- **Test parametru p rozdělení $A(p)$**

Používáme testové kritérium: $u = \frac{\hat{p}-p}{\sqrt{p\,(1-p)}}\sqrt{n}\,,$

kde $\hat{p} = \bar{x}$ je bodový odhad parametru p.

p je testovaný parametr (daný v H_0)

n je rozsah souboru (test požaduje n tak velké, aby $n\hat{p}(1-\hat{p}) \geq 9$)

Je-li H_0 pravdivá, pak $u \approx N(0,1)$.

- **Test parametru μ rozdělení $N(\mu, \sigma^2)$**

Testové kritérium: $u = \frac{\bar{x}-\mu}{s}\sqrt{n}$ (test požaduje $n \geq 30$)

Je-li H_0 pravdivá , pak $u \approx N(0,1)$.

Pozn.: Pro malý rozsah n souboru ($n < 30$) používáme tzv. **t-test**. Pro stejné test. kritérium jsou pak kritickými hodnotami příslušné kvantily t-rozdělení ppsti s $n-1$ stupni volnosti.

Jiný druh hypotéz jsou tzv. **NEPARAMETRICKÉ HYPOTÉZY**, např:

- **Chí-kvadrát test dobré shody**

Testujeme hypotézu H_0: "Nejsou (významné) rozdíly mezi pozorovanými očekávanými četnostmi ."

$n \ldots$ **rozsah** souboru - tj. počet všech (ne nutně různých) naměřených hodnot experimentu

$k \ldots$ **všechny možné výsledky** experimentu pro $i = 1, 2, \ldots, k$ označme:

$n_i \ldots$ **pozorovaná četnost** i-tého výsledku,

$n_i^O \ldots$ **očekávaná četnost** i-tého výsledku.

testové kritérium: $\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i^O\right)^2}{n_i^O}$

Je-li n tak velké, aby $n_i^O \geq 5 \ \ \forall\, i = 1, \ldots, k$, pak $\chi^2 \approx \chi^2(\nu)$, kde $\nu = k - 1$.

Test je pravostranný, tj. kritickými hodnotami jsou "horní"kvantily $\chi^2_{1-\alpha}(\nu)$ rozdělení $\chi^2(\nu)$.

- **Chí-kvadrát test nezávislosti**

Testujeme hypotézu H_0: **Náhodné veličiny X, Y jsou nezávislé.**

Testové kritérium:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - n_{ij}^O\right)^2}{n_{ij}^O},$$

kde $n_{ij} \ldots$ **pozorované četnosti** náh. vektoru $\vec{X} = (X, Y)$, často zapisované ve formě tzv. **kontingenční tabulky**, *kontingence* ... "statistická"závislost X na Y

r je počet možných hodnot veličiny X

s je počet možných hodnot veličiny Y

$n_{ij}^O \ldots$ **očekávané četnosti** příslušné k n_{ij} (musí být vypočteny)

Za předpokladu platnosti H_0 je

$$n_{ij}^O = \frac{n_{i\cdot}\ n_{\cdot j}}{n},$$

kde $n_{i\cdot}$ je součet pozorovaných četností v i-tém řádku tabulky,

$n_{\cdot j}$ je součet pozorovaných četností v j-tém sloupci tabulky,

n je celkový počet pozorování v tabulce.

Jsou-li **všechny očekávané četnosti aspoň 5**, pak $\chi^2 \approx \chi^2(\nu)$, kde $\nu = (r-1)(s-1)$.

KORELACE

Jsou-li (x_i, y_i), $i = 1, 2, \dots, n$ realizace náh. výběru, definujeme číslo r takto:

$$r = \frac{\frac{1}{n}\sum x_i\, y_i - \bar{x}\,\bar{y}}{s_x \cdot s_y}$$

kde $\bar{x}$, resp. $\bar{y}$ jsou aritm. průměry, s_x, resp. s_y jsou výběr. směrodatné odchylky

$r \dots$ **výběrový korelační koeficient** (je bodovým odhadem tzv. koeficientu korelace ϱ)

Platí: $-1 \leq r \leq +1$,

$r \approx 0$ naznačuje, že x_i a y_i jsou lineárně nezávislé (resp. nezávislé - pro normální rozd.),

$r \to \pm 1$ naznačuje silnou lineární závislost (resp. závislost pro normální rozdělení)

K testu **nezávislosti** (při výběrech z normálního rozdělení) používáme testové kritérium

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

Platí-li $H_0 : \varrho = 0$, pak $t \sim t(\nu)$, kde $\nu = n - 2$.

(Alternativa $H_1 : \varrho \neq 0 \Rightarrow$ test je oboustranný.)

Zamítneme-li H_0 na hladině $\alpha = 0.05$, resp. $\alpha = 0.01$, říkáme, že r je **významný**, resp. **vysoce významný**, a hodnoty x_i a y_i považujeme za závislé.

REGRESE

dáno: $[x_i, y_i], i = 1, 2 \dots, n$

$x_i \dots$ dány "přesně", $y_i \dots$ určeny "nepřesně"($\approx$ realizace náh. veličiny)

$y_i \dots$ naměřené hodnoty **neznámé** funkce $y = f(x)$ v bodech x_i, tj. $y_i \approx f(x_i)$

cíl: určit "co nejlepší odhad"$\hat{f}(x)$ fce $f(x)$

Regresní přímka. Předpokládáme, že f je má tvar

$$f(x) = b_0 + b_1\, x$$

a odhady parametrů $\beta_0 \approx b_0$, $\beta_1 \approx b_1$ najdeme tzv. **metodou nejmenších čtverců** - hledáme, pro která β_0, β_1 je minimální součet čtverců

$$S = \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2.$$

Hledaná β_0, β_1 jsou řešením tzv. **soustavy normálních rovnic**

$$\beta_0 \cdot n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i\, y_i$$