Pravděpodobnost & statistika 1

shrnutí!

Nekonečné řady

Součet nekonečné řady

$\sum_{n=1}^{\infty}(a_n)$, má smysl pokud $\underbrace{\lim_{n\to\infty}(a_n)=0}_{\text{nutná podmínka konvergence řady}}$

řadu si rozdělíme na posloupnost částečných sčítanců, pokud má tato řada vlastní limitu, je tato limita = součtu řady

"částečni sčítanci"

$s_1 = a_1$

$s_2 = a_1 + a_2$

$\vdots$

$s_n = a_1 + \dots + a_n$

$$\boxed{\text{Pokud: } \lim_{n\to\infty}(s_n) = A \in \mathbb{R} \Rightarrow \sum_{n=1}^{\infty}(a_n) = A}$$

Geometrické řady

Součet nekonečné geometrické řady

$$\sum_{n=1}^{\infty}(a_n) = a_1 + a_1\cdot q + \dots + a_1\cdot q^k + \dots$$

, opět, pokud $\lim\limits_{n\to\infty}(a_n)=0$

→ platí pouze, pokud $q\in(-1,1)$

částeční sčítanci

$S_1 = a_1$

$S_2 = a_1 + a_1\cdot q$

$\vdots$

$S_n = a_1\cdot\dfrac{q^n-1}{q-1}$

v tomto výrazu se mění pouze čitatel, proto:

pro $|q|>1$ jde výraz $\to\infty$, proto musí $q\in(-1,1)$

pokud: $\lim\limits_{n\to\infty}\left(a_1\cdot\dfrac{q^n-1}{q-1}\right) = a_1\cdot\dfrac{0-1}{q-1} = a_1\cdot\dfrac{1}{1-q} = \dfrac{a_1}{1-q} \Rightarrow \underline{\underline{\sum(a_n)=\dfrac{a_1}{1-q}}}$

TODO:

1) střední hodnota [přednáška 4.]
2) rozptyl [přednáška 4.]
3.) spojité n. veličiny [přednáška 5.]
4.) podmíněná hustota a střední hodnota [přednáška 7.]

Statistika – předpoklady zkoumaných úloh

- vždy předpokládáme nezávislá měření – hodnoty n.n.v. $X_1, \dots, X_n \sim F$
- předpokládáme, že F patří do nějakého modelu = množiny distrib fcí s konečnou $\mathbb{E}$
- můžeme potkat parametrické / neparametrické modely

Param:

$$F \in \{ \text{d.f. } N(\mu, \sigma^2) \}$$

$$F \in \{ \text{d.f. } Exp(\lambda) \}$$

...

neparam:

empirická dist. fce

(*) víc si uvádět nebudeme

Statistika – značení

ϑ = malá theta
Θ = velká theta

$$\{\underbrace{F_\vartheta,}\ \vartheta \in \Theta\}$$

F_ϑ = model

ϑ = parametr

Θ = množina

Statistika – metody

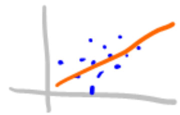
klasické → existuje neznámá konst. ϑ a my ji určujeme z náh. měření

Bayesovské

→ ϑ je náhodná, pomocí Bayesovy věty pak upravujeme naše poznání o distribuci ϑ

Statistika – cíle

- bodové odhady
  - -> určete ϑ
  - -> určete $g(\vartheta)$ pro nějakou fci g např. rozptyl, který je závislý na g
- intervalové odhady
  - -> odpovídáme intervalem, kterému náleží ϑ $\vartheta \in (a, b)$
- testování hypotéz
  - -> je mince spravedlivá?
- lineární regrese
  - -> data s dvěma číselnými hodnotami chápu jako souřadnice a prokládám je přímkou



  - -> zkoumáme jejich závislost $y = \underbrace{\beta_0}_{\text{průsečík s osou } y} + \underbrace{\beta_1}_{\text{náklon}} x$ $\begin{cases} >0 & \dots \; Z \\ =0 & \dots \; N \\ <0 & \dots \; Z \end{cases}$

Bodové odhady a jejich vlastnosti

Pozorování $X = (X_1, \dots, X_n)$, X_i je n.n.v. s dist. fcí F_ϑ

odhad je n.v. $\underbrace{\hat{\Theta} = \hat{\Theta}_n}_{\neq \Theta \ldots \text{ množina}} = g(X)$, pro vhodnou fci g

závisí na $\vartheta \in \Theta$

chyba odhadu

je n.v. $\underbrace{\tilde{\Theta} = \tilde{\Theta}_n}_{\text{chyba}} := \underbrace{\hat{\Theta}_n}_{\text{odhad}} - \underbrace{\vartheta}_{\text{skutečnost}}$

BIAS = „zaujatost"

je konstanta (pro pevnou ϑ) $\quad b_\vartheta(\hat{\Theta}_n) = \mathbb{E}(\tilde{\Theta}_n) \overset{\text{linearita } \mathbb{E}}{=} \mathbb{E}(\hat{\Theta}_n) - \vartheta$

říkáme, že odhad $\hat{\Theta}_n$ je

unbiased

nezaujatý $\iff b_\vartheta(\tilde{\Theta}_n) = 0, \ \forall \vartheta \in \Theta$ → nemá chybu v měření

asymptoticky nezaujatý $\iff b_\vartheta(\tilde{\Theta}_n) \to 0, \ \forall \vartheta \in \Theta$ → má chybu, která s postupem času klesá

konzistentní $\iff \hat{\Theta}_n \xrightarrow{P} \vartheta, \ \forall \vartheta \in \Theta$

$\iff P(|\hat{\Theta}_n - \vartheta| > \varepsilon) \to 0, \ \forall \varepsilon > 0$

Testování hypotéz – ilustrace

hypotéza $\begin{cases} H_0 \text{ nulová} \\ H_1 \text{ alternativní} \end{cases}$ $\Rightarrow$ výsledek $\begin{cases} \text{zamítneme } H_0 \text{ (platí } H_1) \\ \text{nezamítneme } H_0 \end{cases}$

chyba $\begin{cases} \text{1. druhu} & \text{chybné přijetí} \\ \text{2. druhu} & \text{chybné zamítnutí} \end{cases}$

Potřebujeme určit k takové, aby:

- $P(\text{chyby } 1) \leq \alpha$ (typicky 0,05)

- došlo k zamítnutí, pokud: $\underbrace{\left|S - \frac{n}{2}\right| > k}$

ilustrace na hodech mincí

\# úspěchů je víc než k daleko od očekávané hodnoty

Testování hypotéz - obecný postup

1) vybereme vhodný statistický model
2) zvolíme hladinu významnosti α (typicky $\alpha = 0{,}05$)
3) určíme testovou statistiku $S = h(X_1, \dots, X_n)$ h je fce, kterou aplikujeme při testu
4) určíme kritický obor - množinu W (kdy zamítáme H_0)
5) naměříme $x_1, \dots, x_n$ z $X_1, \dots, X_n$
6) určíme rozhodovací pravidlo

zamítneme H_0 pokud $h(x_1, \dots, x_n) \in W$

(*) často nejprve spočítáme p-hodnotu ... minimální α, kdy zamítneme H_0 místo toho, abychom určovali α "dopředu"

Bayesova věta

• pro jevy

$B_1, B_2, \dots$ rozklad Ω, $A \in \mathcal{F}$, $P(A)$ a $P(B_i) > 0$:

(*) $P(B_i)$ = apriorní
nic o B_i zatím nevíme
$P(B_i|A)$ = posteriorní
B_i ... vyvozujeme z naměřených dat A

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

(*) podíl "mého podm. jevu" ze $\forall$ podm. jevů

• pro diskrétní n.v.

X, Y diskrétní n.v.

$$P_{Y|X}(y|x) = \frac{P_{X|Y}(x,y)P_Y(y)}{\sum_{y'} P_{X|Y'}(x,y')P_{Y'}(y')}$$

• pro spojité n.v.