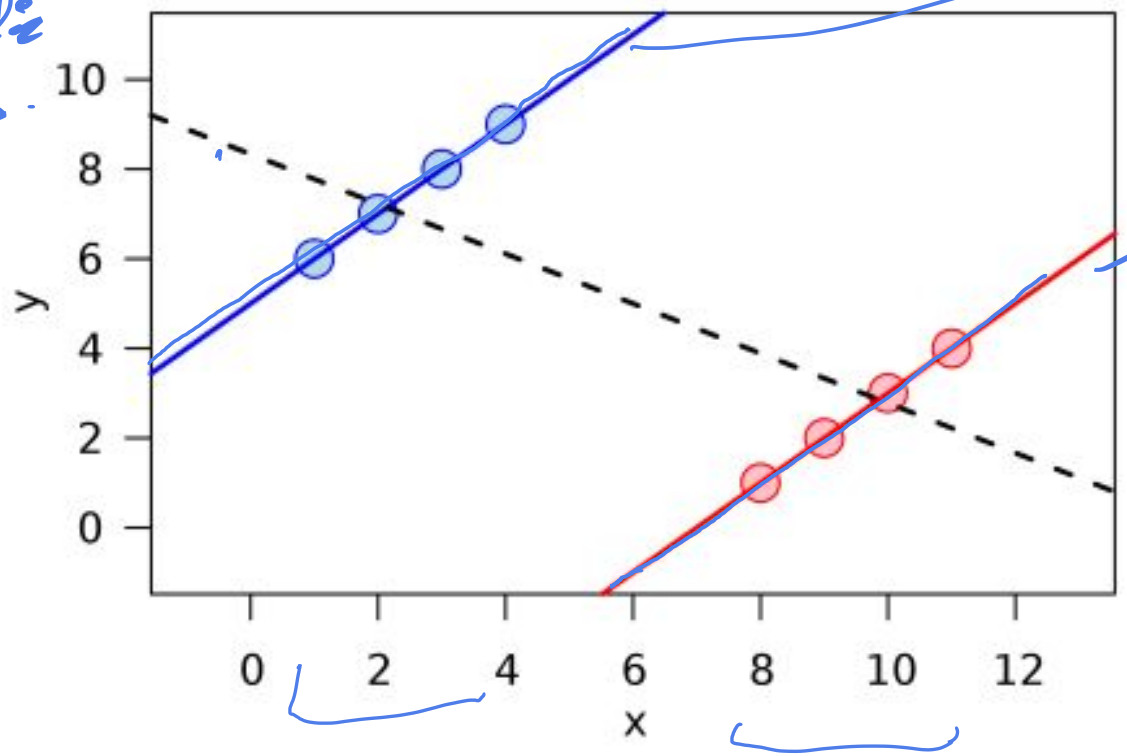
NMAI059 Pravděpodobnost a statistika 1
13. přednáška

Robert Šámal

Simpsonův paradox

Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



Přehled

Situace

- Máme k dispozici dvě sady nezávislých náhodných veličin (náhodné výběry):
- $X_1, \ldots, X_n \sim F_X$ a $Y_1, \ldots, Y_m \sim F_Y$
- Chceme rozhodnout, zda platí $H_0 : F_X = F_Y$ nebo $H_1 : F_X \neq F_Y$
- Příklady: doba běhu programu před/po vylepšení, hladina cholesterolu u lidí co jedí/nejedí Zázračnou Superpotravu^{TM}, frekvenci krátkých slov v textu autora X a Y.
- Nevíme nic o vlastnostech F_X, F_Y (zejména nečekáme, že je normální)

Postup

- Zvolíme vhodnou statistiku, např.

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = |\bar{X}_n - \bar{Y}_m|$$

- $t_{\text{obs}} := T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$
- Za předpokladu H_0 jsou „všechny permutace stejné“: X_i i Y_j se generovaly ze stejného rozdělení.
- Náhodně zpermutujeme zadaných $m + n$ čísel a pro každou permutaci vyčíslíme T – dostaneme čísla $T_1, T_2, \ldots, T_{(m+n)!}$ (každé stejně pravděpodobné).
- Jako p-hodnotu vezmeme pravděpodobnost, že $T > t_{\text{obs}}$, neboli

$$p = \frac{1}{(m+n)!} \sum_j I(T_j > t_{\text{obs}}).$$

- To je pravděpodobnost chyby 1. druhu, neboli H_0 zamítneme, pokud je $p < \alpha$ (pro naši zvolenou hodnotu α, např. $\alpha = 0.05$).

Vylepšení

- Zkoušet všechny permutace může trvat moc dlouho. Vezmeme tedy jen vhodný počet B nezávisle náhodně vygenerovaných permutací a spočítáme jenom B hodnot $T_1, \ldots, T_B$.
- Jako p-hodnotu vezmeme odhad pravděpodobnost, že $T > t_{\text{obs}}$, neboli

$$\frac{1}{B} \sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

- Pro dostatečně velké m, n dává podobné výsledky jako testy založené na CLV, vhodné je tedy zejména pro středně velké počty.

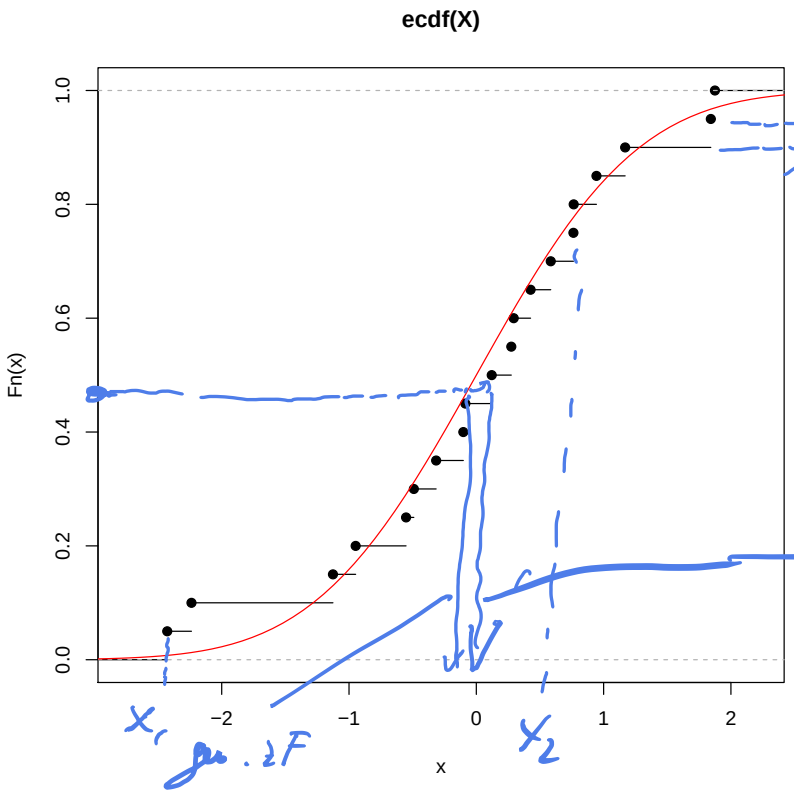
Přehled

Empirická distribuční funkce – připomenutí

- $X_1, \ldots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

kde $I(X_i \leq x) = 1$ pokud $X_i \leq x$ a 0 jinak.



Boostrap – základní idea

- z naměřených dat $X_1 = x_1, \ldots, X_n = x_n \sim F$ vytvoříme $\widehat{F}_n$
- další data můžeme samplovat z $\widehat{F}_n$
- to se dělá tak, že vybereme uniformně náhodné $i \in \{1, \ldots, n\}$ a řekneme x_i

Bootstrap – základní použití

- $T_n = g(X_1, \ldots, X_n)$ nějaká statistika (funkce dat)
- chceme odhadnout $var\, T_n$
- nasamplujeme $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$ (viz minulá strana)
- spočteme $T_n^* = g(X_1^*, \ldots, g_n^*)$
- opakujeme B-krát, dostaneme $T_{n,1}^*, \ldots, T_{n,B}^*$
- odhad rozptylu:

$$\frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{k=1}^{B} T_{n,k}^*\right)^2$$

... sm. rozptyl

jeden měř.

výb. střed.

Přehled

Srovnání dvou přístupů ke statistice

Frekventistický/klasický přístup

- Pravděpodobnost je dlouhodobá frekvence (z 6000 hodů kostkou padla šestka 1026-krát). Je to objektivní vlastnost reálného světa.
- Parametry jsou pevné, neznámé konstanty. Nelze o nich říkat smysluplné pravděpodobnostní výroky.
- Navrhujeme statistické procedury tak, aby měly žádané dlouhodobé vlastnosti. Např. 95 % z našich intervalových odhadů pokryje neznámý parametr.

Bayesovský přístup

- Pravděpodobnost popisuje, jak moc věříme nějakému jevu, jak moc jsme ochotní se vsadit. (Pravděpodobnost, že Thomas Bayes měl 18. prosince 1760 šálek čaje, je 90 %.)
- Můžeme vyslovovat pravděpodobnostní výroky i o parametrech (třebaže jsou to pevné konstanty).
- Spočítáme distribuci ϑ a z ní tvoříme bodové a intervalové odhady, atd.

Bayesovská metoda – základní popis

- neznámý parametr považujeme za náhodnou veličinu Θ
- zvolíme *apriorní distribuci (prior distribution)*, neboli hustotu pravděpodobnosti $f_\Theta(\vartheta)$ nezávislou na datech.
- zvolíme statistický model $f_{X|\Theta}(x|\vartheta)$, který popisuje, co naměříme (s jakou pravděpodobností), v závislosti na hodnotě parametru
- poté, co pozorujeme hodnotu $X = x$, spočítáme *posteriorní distribuci (posterior distribution)* $f_{\Theta|X}(\vartheta|x)$
- z té pak odvodíme, co potřebujeme např. najdeme a, b, aby $\int_a^b f_{\Theta|X}(\vartheta|x)d\vartheta \geq 1 - \alpha$
- $\vartheta = \theta$ malá théta, Θ je velká théta

Bayesova věta

Věta (Bayesova pro diskrétní náhodné veličiny)

X, Θ jsou diskrétní n.v.

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')p_\Theta(\vartheta')}.$$

(sčítance s $p_\Theta(\vartheta') = 0$ považujeme za 0).

Věta (Bayesova pro spojité náhodné veličiny)

X, Θ jsou spojité n.v., které mají hustotu f_X, f_Θ i sdruženou hustotu $f_{X,\Theta}$

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_\Theta(\vartheta)}{\int_{\vartheta' \in \mathbb{R}} f_{X|\Theta}(x|\vartheta')f_\Theta(\vartheta')d\vartheta'}.$$

(sčítance s $f_\Theta(\vartheta') = 0$ považujeme za 0).

Bayesovské bodové odhady – MAP a LMS

MAP – Maximum A-Posteriori

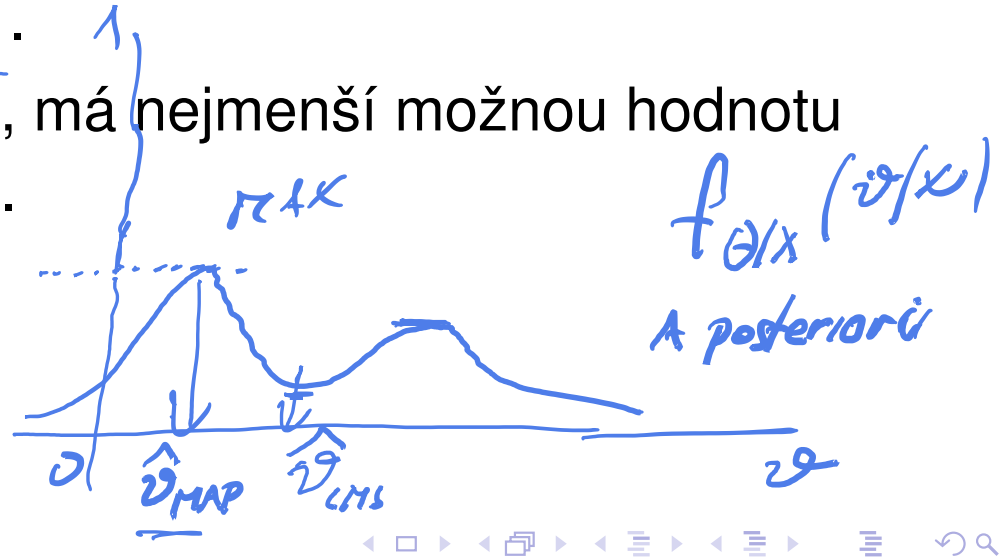
Volíme $\hat{\vartheta}$ tak, aby maximalizovalo

- $p_{\Theta|X}(\vartheta|x)$ v diskrétním případě
- $f_{\Theta|X}(\vartheta|x)$ ve spojitém případě
- Podobné metodě ML v klasickém přístupu, pokud bychom volili „flat prior“ – uniformní $p_\Theta(\vartheta)$.

LMS – Least Mean Square

Též metoda podmíněné střední hodnoty.

- Volíme $\hat{\vartheta} = \mathbb{E}(\Theta \mid X = x)$.
- Nestranný bodový odhad, má nejmenší možnou hodnotu LMS: $\mathbb{E}\big((\Theta - \hat{\vartheta})^2 | X = x\big)$.



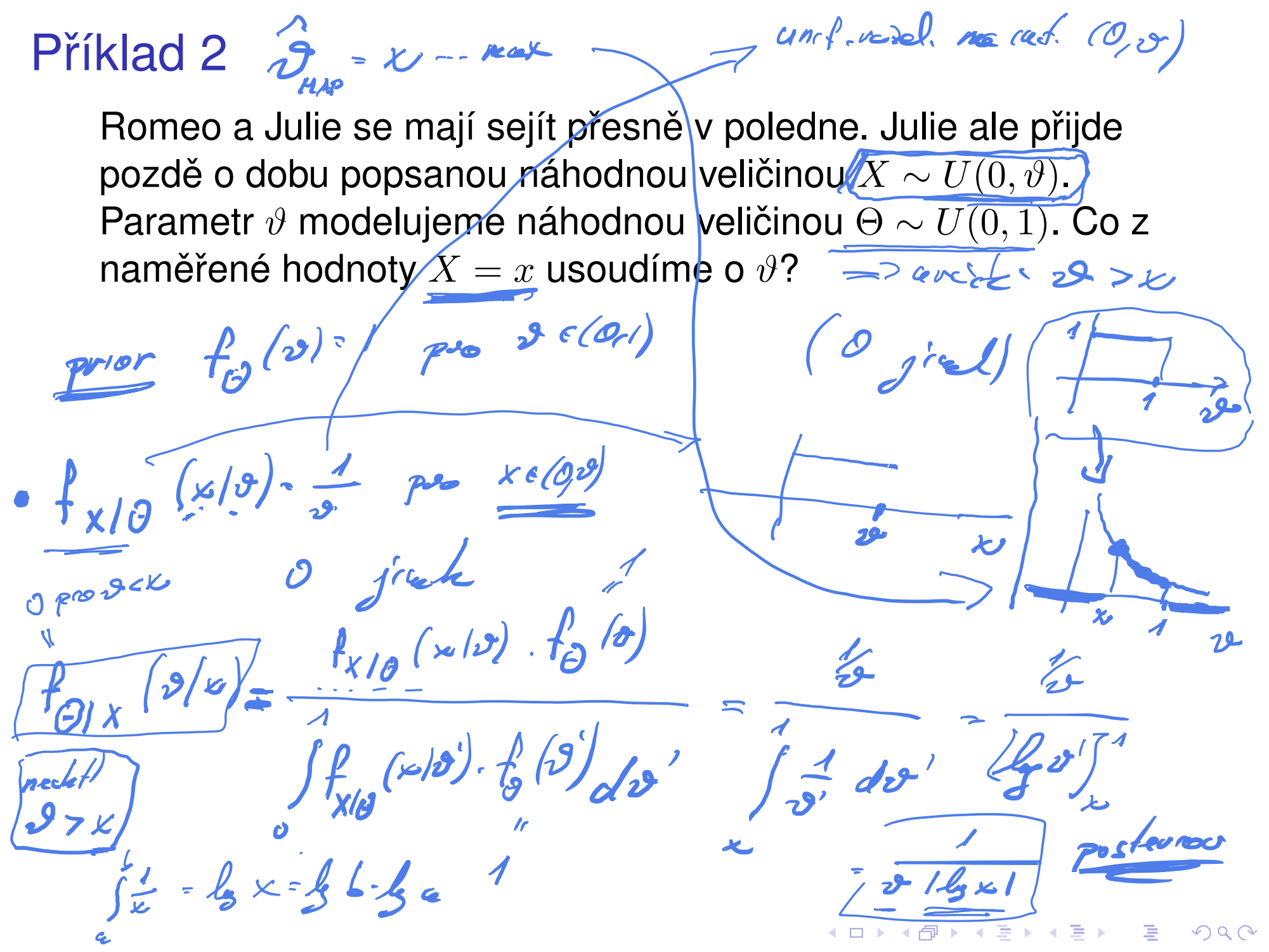
Příklad 1

Bayesovský klasifikátor spamů:

- vytvoříme seznam podezřelých slov (money, win, pharmacy, ...)
- N.v. X_i popisuje, zda email obsahuje podezřelé slovo w_i.
- N.v. Θ popisuje, zda email je spam $\Theta = 1$ nebo ne $\Theta = 0$.
- Z předchozích emailů získáme odhady $p_{X|\Theta}$ a p_Θ
- Použijeme Bayesovu větu na výpočet $p_{\Theta|X}$

Příklad 2

Romeo a Julie se mají sejít přesně v poledne. Julie ale přijde pozdě o dobu popsanou náhodnou veličinou $X \sim U(0, \vartheta)$. Parametr ϑ modelujeme náhodnou veličinou $\Theta \sim U(0, 1)$. Co z naměřené hodnoty $X = x$ usoudíme o ϑ?



LMS, pdf. stó. bocha.

$$\hat{\vartheta}_{LMS} = \mathbb{E}(\Theta / X = x) = \int_{-\infty}^{\infty} \vartheta \, f_{\Theta/X}(\vartheta/x) \, d\vartheta$$

$$= \int_{x}^{1} \vartheta \frac{1}{\vartheta \ln x} \, d\vartheta = \int_{x}^{1} \frac{1}{\ln x} \, d\vartheta$$

$$= \frac{1-x}{\ln x}$$

Příklad 3

Pozorujeme náhodné veličiny $X = (X_1, \ldots, X_n)$, předpokládáme $X_i \sim N(\vartheta, \sigma_i^2)$ a ϑ je hodnota náhodné veličiny $\Theta \sim N(x_0, \sigma_0)$. Co z naměřených hodnot $X = x = (x_1, \ldots, x_n)$ usoudíme o ϑ?

Příklad 4

Házíme mincí, pravděpodobnost, že padne panna je ϑ. Z n hodů padla panna v $X = k$ případech. Pokud naše apriorní distribuce byla $U(0, 1)$, jaká bude distribuce posteriorní?

Přehled

Základní metoda (inverse transformation method)

Věta

Nechť F je funkce „typu distribuční funkce“: neklesající zprava spojitá funkce s $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to+\infty} F(x) = 1$.
Nechť Q je odpovídající kvantilová funkce.
Nechť $U \sim U(0,1)$ a $X = Q(U)$
Pak X má distribuční funkci F.

- Funguje dobře, když umíme vyčíslit Q, třeba pro exponenciální nebo geometrické rozdělení.
- Gamma rozdělení je součet několika exponenciálních – tak ho tak i vygenerujeme.

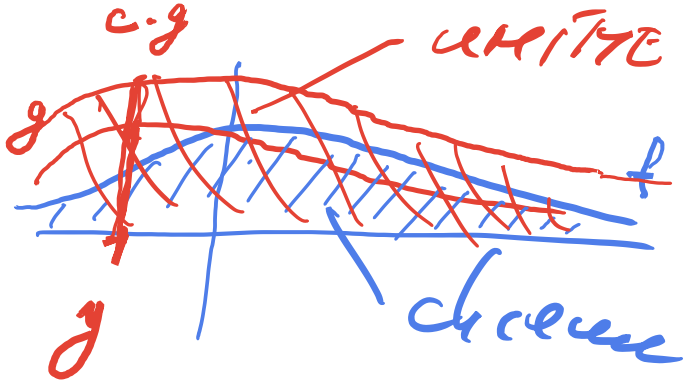
Varianta základní metody pro diskrétní proměnné

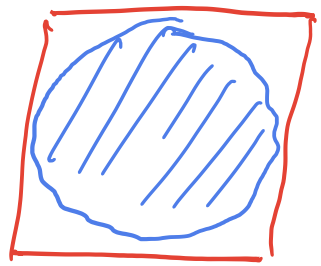
- Chceme n.v. X, která nabývá hodnot $x_1, x_2, \ldots$ s pravděpodobnostmi $p_1, p_2, \ldots$. ($\sum_i p_i = 1$).
- Vygenerujeme $U \sim U(0,1)$.
- Najdeme i takové, že $p_1 + \cdots + p_{i-1} < U < p_1 + \cdots + p_i$.
- Položíme $X := x_i$.

- Funguje hezky když máme vzorec pro $p_1 + \cdots + p_i$ (např. geometrické rozdělení).
- Binomické rozdělení je lepší simulovat jako součet n nezávislých Bernoulliových veličin.
- Na další (Poisson) jsou speciální triky).

Zamítací metoda (rejection sampling)

- Chceme vygenerovat n.v. s hustotou f.
- Umíme vygenerovat n.v. s hustotou g (která je „podobná“).
- $\frac{f(y)}{g(y)} \leq c$ pro nějakou konstantu c.
- Postup
  1. Vygenerujeme Y s hustotou g, a $U \sim U(0,1)$.
  2. Pokud $U \leq \frac{f(Y)}{cg(Y)}$, tak $X := Y$.
  3. Jinak hodnotu Y, U zamítneme a opakujeme od bodu 1.
- Zdůvodnění: vygenerovat náhodnou hodnotu X s hustotou f je totéž, jako vygenerovat náhodný bod pod grafem funkce f, jehož vodorovná (x-ová) souřadnice je X (a svislá je uniformně náhodná mezi 0 a X).





Navazující přednášky

- Pravděpodobnost a statistika 2 - NMAI073
- Úvod do aproximačních a pravděpodobnostních algoritmů - NDMI084
- Úvod do strojového učení v Pythonu|systému R - NPFL129|NPFL054
- a mnoho magisterských přednášek