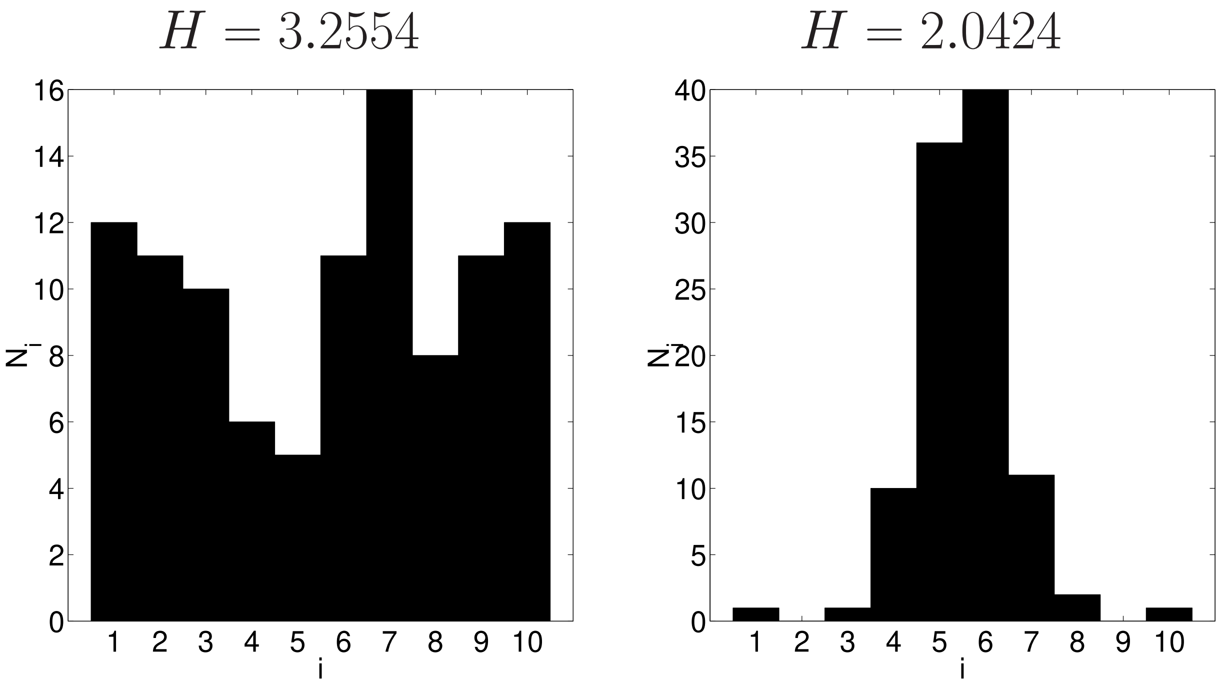
Entropie

1. Entropie jako míra informace.
2. Pokrytí množiny elementárních jevů.
3. Vlastnosti entropie.
4. Podmíněná entropie.
5. Vlastnosti podmíněné entropie.
6. Vzájemná informace.
7. Kvalita asociace: kontingenční test
8. Optimální histogramování

Entropie jako míra neuspořádanosti

Jev x_i: kulička padla do přihrádky i



$N = 100$ stejných objektů: $\{b_1, b_2, \ldots b_N\}$

$N!$ všech permutací

$N_i!$ permutací v každé přihrádce, $i = 1, 2, \ldots, m$

Celkový počet přerovnání objektů v histogramu je

$$W = \frac{N!}{\displaystyle\prod_{i=1}^{m} N_i!}$$

Entropie

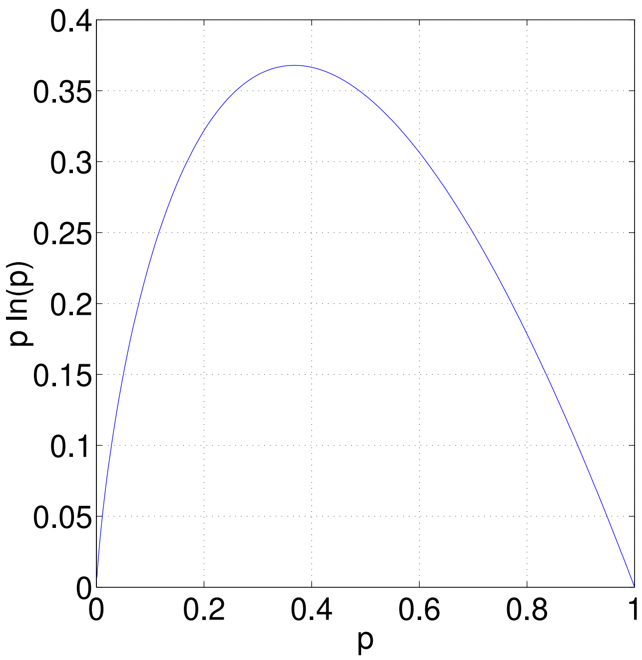
$$\lim_{N\to\infty} -\frac{1}{N}\ln W =$$

$$= \lim_{N\to\infty} -\sum_{i=1}^{m} \frac{N_i}{N} \ln \frac{N_i}{N} = -\sum_{i=1}^{m} p_i \ln p_i$$

Počet mikrostavů (rozdělení do m přihrádek) které dávají za vznik stejnému makrostavu (histogram).

- Entopie je malá pro úzká rozdělení
- Entropie je velká pro široká rozdělení

$p \ln p$



Entropie jako míra informace

Hledáme funkci $s(p)$:

1. spojitou
2. monotonně klesající s p
3. $s(1) = 0$ (nulová nejistota)
4. $s(p_A \cdot p_B) = s(p_A) + s(p_B)$

Hledaná funkce

$$s(p) = -s(\tfrac{1}{e}) \ln p$$

$$s(\tfrac{1}{e}) = 1$$ definujeme

Střední míra informace (v Natech) na množině jevů $\mathbf{x} = \{x_k\}$

$$H(\mathbf{x}) = -\sum_k p(x_k) \ln p(x_k)$$

Vlastnosti rozkladu množiny elementárních jevů na třídy ekvivalence

Množina elementárních jevů (MEJ): $\{A_1, A_2, \dots, A_n\}$

Definice

1. Rozklad je disjunktní pokrytí MEJ
2. Elementární rozklad $\mathcal{E} = \{A_1, A_2, \dots, A_n\}$
3. Zjemnění rozkladu

$$\mathcal{B} \preceq \mathcal{A} \quad \text{právě když} \quad \forall i\, \exists j : B_i \subseteq A_j$$

4. Součin rozkladů $\mathcal{C} = \mathcal{A} \cdot \mathcal{B} = \{A_i \cap B_j, \forall i, j\}$ je největší společné zjemnění

$$\mathcal{A} \cdot \mathcal{B} \prec \mathcal{A}, \quad \mathcal{A} \cdot \mathcal{B} \prec \mathcal{B}$$

Vlastnosti

1. $\mathcal{E} \prec \mathcal{A}$ pro každé $\mathcal{A}$
2. $\mathcal{A} \cdot \mathcal{B} = \mathcal{B} \cdot \mathcal{A}$ (komutativita)
3. $\mathcal{A} \cdot (\mathcal{B} \cdot \mathcal{C}) = (\mathcal{A} \cdot \mathcal{B}) \cdot \mathcal{C}$ (asociativita)
4. Jestliže $\mathcal{A}_1 \prec \mathcal{A}_2 \prec \mathcal{A}_3$ potom $\mathcal{A}_1 \prec \mathcal{A}_3$ (tranzitivita)
5. Jestliže $\mathcal{B} \prec \mathcal{A}$ potom $\mathcal{A} \cdot \mathcal{B} = \mathcal{B}$

Entropie na rozkladu MEJ

$\mathcal{A} = \{A_1, A_2, \dots, A_n\}$, $\quad P(A_i) \stackrel{\text{def}}{=} p_i$ (notace)

Entropie

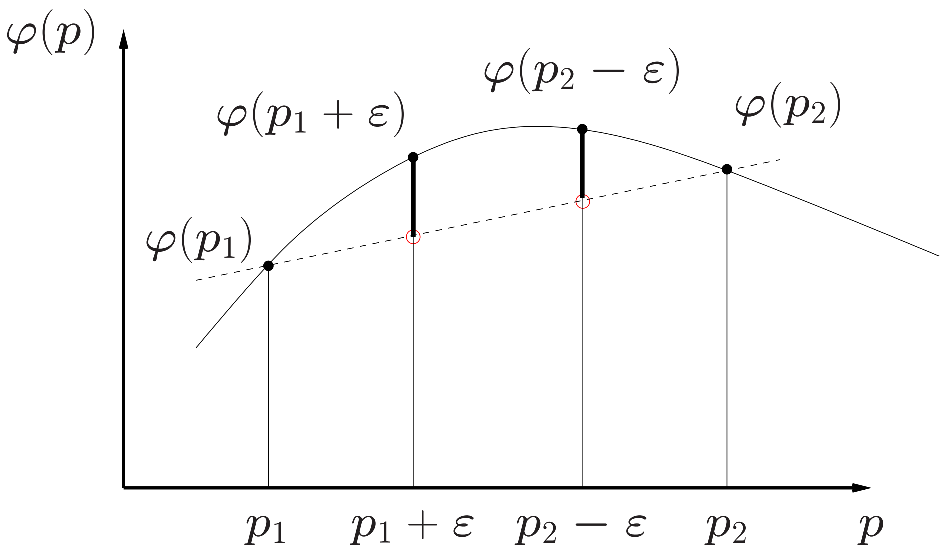
$$H(\mathcal{A}) \stackrel{\text{def}}{=} -\sum_{i=1}^{n} p_i \ln p_i = \sum_{i=1}^{n} \varphi(p_i)$$

$$\varphi(p_i) \stackrel{\text{def}}{=} -p_i \ln p_i$$

Pomocná věta:

$$\varphi(p_1 + p_2) \le \varphi(p_1) + \varphi(p_2) \le \varphi(p_1 + \varepsilon) + \varphi(p_2 - \varepsilon)$$

pokud $\quad p_1 < p_1 + \varepsilon \le p_2 - \varepsilon < p_2$



Vlastnosti

V1. Jestliže $\mathcal{B} \prec \mathcal{A}$ potom $H(\mathcal{B}) \geq H(\mathcal{A})$

Zjemněním rozkladu se entropie zvýší.

Pozn: zjemněním histogramu se entropie zvýší.

V2. Pro každé $\mathcal{A}$ platí $H(\mathcal{A}) \leq H(\mathcal{E})$

Entropie každého rozkladu je menší nebo rovna entropii elementárního rozkladu.

V3. Pro každé $\mathcal{A}$ a $\mathcal{B}$ platí

$$H(\mathcal{A}) \leq H(\mathcal{A} \cdot \mathcal{B}), \quad H(\mathcal{B}) \leq H(\mathcal{A} \cdot \mathcal{B}),$$

Pozn: $\mathcal{A} \cdot \mathcal{B} \prec \mathcal{A}, \quad \mathcal{A} \cdot \mathcal{B} \prec \mathcal{B}$

V4. Entropie $\mathcal{A}$ je maximální pokud všechny jeho prvky mají stejnou pravděpodobnost: $p_i = P(A_i) = p$

Statický systém v parlamentu

- podmnožina pěti vybraných poslanců $\{v_1, \ldots, v_5\}$
- někteří v_i hlasují podle v_j, $j \neq i$, ale nevíme kteří
- někteří hlasují nezávisle, nevíme kteří
- máme záznam 100 hlasování

v_1	v_2	v_3	v_4	v_5	$p(\mathbf{s})$
0	0	0	0	0	0.02
0	0	0	0	1	0.04
0	0	0	1	0	0.02
0	0	0	1	1	0.05
0	0	1	0	0	0.03
0	0	1	0	1	0.03
0	0	1	1	0	0.09
0	0	1	1	1	0.03
0	1	0	0	0	0.03
0	1	0	0	1	0.05
0	1	0	1	0	0.04
0	1	0	1	1	0.01
0	1	1	0	0	0.03
0	1	1	0	1	0.03
0	1	1	1	0	0.04
0	1	1	1	1	0.05

v_1	v_2	v_3	v_4	v_5	$p(\mathbf{s})$
1	0	0	0	0	0.01
1	0	0	0	1	0.03
1	0	0	1	0	0.03
1	0	0	1	1	0.03
1	0	1	0	0	0.01
1	0	1	0	1	0.03
1	0	1	1	1	0.01
1	1	0	0	0	0.03
1	1	0	0	1	0.03
1	1	0	1	0	0.03
1	1	0	1	1	0.04
1	1	1	0	0	0.01
1	1	1	0	1	0.04
1	1	1	1	0	0.03
1	1	1	1	1	0.05

Podmíněná entropie

Def. Nechť rozklady $\mathcal{A}$ a $\mathcal{B}$ jsou

$$\mathcal{A} = [A_1, A_2, \dots, A_{N_A}]$$
$$\mathcal{B} = [B_1, B_2, \dots, B_{N_B}]$$

Potom podmíněná entropie rozkladu $\mathcal{A}$ za předpokladu, že nastal jev B_j je

$$H(\mathcal{A} \mid B_j) = -\sum_{i=1}^{N_A} P(A_i \mid B_j) \ln P(A_i \mid B_j)$$

Podmíněná entropie $H(\mathcal{A} \mid \mathcal{B})$ je pak střední hodnota

$$\begin{aligned} H(\mathcal{A} \mid \mathcal{B}) &= \sum_{j=1}^{N_B} P(B_j)\, H(\mathcal{A} \mid B_j) \\ &= -\sum_{j=1}^{N_B} P(B_j) \sum_{i=1}^{N_A} P(A_i \mid B_j)\, \ln P(A_i \mid B_j) \\ &= -\sum_{i} \sum_{j} P(A_i\, B_j) \ln P(A_i \mid B_j) \end{aligned}$$

Pozn: Střední nejistota o $\mathcal{A}$ je-li pozorováno $\mathcal{B}$ je $H(\mathcal{A} \mid \mathcal{B})$.

Vlastnosti

Pro každé $\mathcal{A}$ a $\mathcal{B}$ platí

V5. Jestliže $\mathcal{B} \prec \mathcal{A}$ potom $H(\mathcal{A} \mid \mathcal{B}) = 0$.

Intuice: víme, které jevy v $\mathcal{A}$ nastaly.

Def. Rozklady $\mathcal{A}$ a $\mathcal{B}$ nezávislé $\Leftrightarrow$
$\forall A_i \in \mathcal{A},\ B_j \in \mathcal{B}\colon\ P(A_i\, B_j) = P(A_i) \cdot P(B_j)$.

V6. Jsou-li $\mathcal{A}$ a $\mathcal{B}$ nezávislé, potom

$$H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A}), \qquad H(\mathcal{B} \mid \mathcal{A}) = H(\mathcal{B})$$

Intuice: Jsou-li jevy nezávislé, potom pozorováním $\mathcal{B}$ nezískáme žádnou informaci o $\mathcal{A}$.

V7. $H(\mathcal{A} \cdot \mathcal{B}) \le H(\mathcal{A}) + H(\mathcal{B})$

Intuice: entropie složeného jevu není větší než entropie dílčích jevů.

V8. Jsou-li $\mathcal{A}$ a $\mathcal{B}$ nezávislé, potom

$$H(\mathcal{A} \cdot \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$$

V9. $H(\mathcal{A} \cdot \mathcal{B}) = H(\mathcal{B}) + H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B} \mid \mathcal{A})$

Intuice: Pozorujeme-li $\mathcal{B}$, získáme o $\mathcal{A}$ informaci $H(\mathcal{B})$, takže zbytková nejistota je $H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A} \cdot \mathcal{B}) - H(\mathcal{B})$.

V10. $H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{B}) - H(\mathcal{B} \mid \mathcal{A})$

V11. $H(\mathcal{B}) \leq H(\mathcal{A} \cdot \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B})$

V12. Pro každé $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ platí:

$$\begin{aligned} H(\mathcal{B} \cdot \mathcal{C} \mid \mathcal{A}) &= H(\mathcal{B} \mid \mathcal{A}) + H(\mathcal{C} \mid \mathcal{A} \cdot \mathcal{B}) \\ &= H(\mathcal{C} \mid \mathcal{A}) + H(\mathcal{B} \mid \mathcal{C} \cdot \mathcal{A}) \end{aligned}$$

Pozn: důsledek vlastnosti V9.

V13. $H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A})$

Intuice:

1. Pozorováním $\mathcal{B}$ se nejistota o $\mathcal{A}$ nemůže zvětšit.
2. Entropie podmnožiny rozkladu je menší.

V14. Jestliže $\mathcal{B} \prec \mathcal{C}$ potom $H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A} \mid \mathcal{C})$.

Intuice: Jemnějším rozkladem se dozvíme více o $\mathcal{A}$.

Vzájemná informace

Pozorování rozkladu $\mathcal{B}$ redukuje nejistotu o $\mathcal{A}$ z $H(\mathcal{A})$ na $H(\mathcal{A} \mid \mathcal{B})$ a získá tedy o $\mathcal{A}$ informaci $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B})$.

Def.

$$I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A} \cdot \mathcal{B}) \qquad \text{(plyne z V9)}$$

Vlastnosti

$$I(\mathcal{A}, \mathcal{B}) = I(\mathcal{B}, \mathcal{A}) \qquad \text{(symetrie)}$$

$$I(\mathcal{A}, \mathcal{B}) \geq 0 \qquad \text{(nonnegativita, plyne z V7)}$$

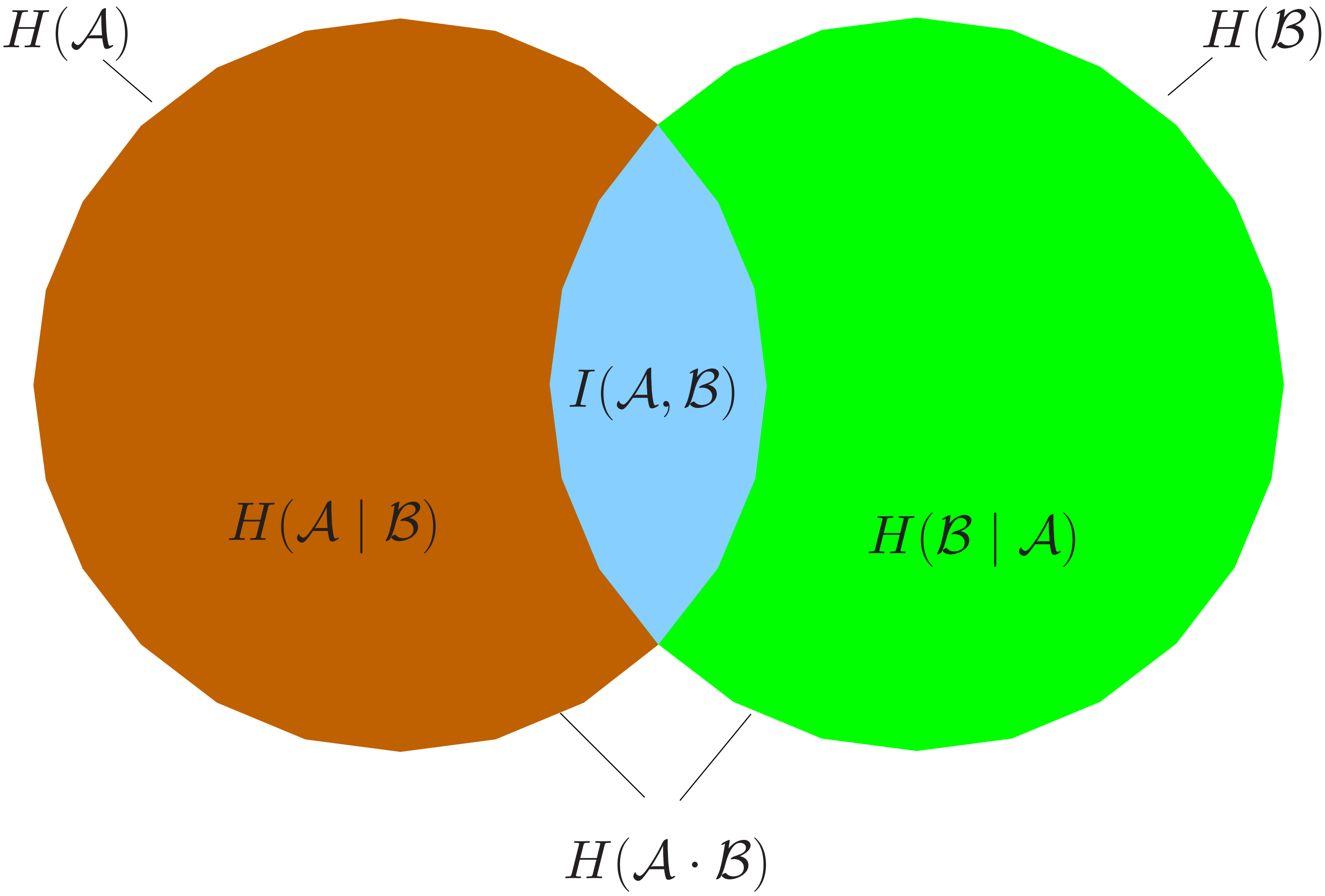
$\underline{I(\mathcal{A}, \mathcal{B})}$

1. Informace o $\mathcal{A}$ obsažená v $\mathcal{B}$
2. Informace o $\mathcal{B}$ obsažená v $\mathcal{A}$

Pro více rozkladů:

$$I(\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_k) = \sum_{i=1}^{k} H(\mathcal{A}_i) - H(\mathcal{A}_1 \cdot \mathcal{A}_2 \cdots \mathcal{A}_k)$$

Mnemotechnická pomůcka: Vzájemný vztah mezi podmíněnými druhy entropie



Čteme takto: platí aditivita plochy, příčemž levý kruh představuje $H(\mathcal{A})$, pravý kruh $H(\mathcal{B})$ a jejich sjednocení $H(\mathcal{A}\cdot\mathcal{B})$. Potom

$$H(\mathcal{A}) + H(\mathcal{B}\mid\mathcal{A}) = H(\mathcal{A}\cdot\mathcal{B})$$

$$H(\mathcal{A}\cdot\mathcal{B}) - I(\mathcal{A}\cdot\mathcal{B}) = H(\mathcal{A}\mid\mathcal{B}) + H(\mathcal{B}\mid\mathcal{A})$$

$$\ldots$$

Entropie náhodné proměnné

Nechť X je diskrétní náhodná proměnná nabývající hodnot $x_i \in R(X)$ s pravděpodobnostmi

$$P(X = x_i) \stackrel{\text{def}}{=} p_i$$

Sjednocení událostí $\{X = x_i\}$ tvoří rozklad $\mathcal{A}_X$ (tj. pokrytí oboru hodnot $R(X)$).

Def. 1

Entropie $H(X)$ diskrétní náhodné proměnné X je rovna

$$H(X) = H(\mathcal{A}_X) = -\sum_i p_i \ln p_i$$

Def. 2

Diferenciální entropie $H(X)$ spojité náhodné proměnné X je

$$H(X) \stackrel{\text{def}}{=} -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx$$

Pozn 1: události $\{X = x_i\}$ spojité X netvoří rozklad.

Pozn 2: pro spojitou X je $H(X) \in (-\infty, \infty)$

Sdružená entropie náhodné proměnné

$$H(X,Y) = H(\mathcal{A}_X \cdot \mathcal{A}_Y) = -\sum_i \sum_j p_{ij} \ln p_{ij}$$

Podmíněná entropie

$$H(X \mid Y) = H(\mathcal{A}_X \mid \mathcal{A}_Y)$$

Jestliže $y = f(x)$ je prosté zobrazení, potom

V15. $H(Y \mid X) = H(X \mid Y) = 0$

D: Rozklady $\mathcal{A}_X$ a $\mathcal{A}_Y$ jsou ekvivalentní, $\mathcal{A}_X = \mathcal{A}_Y$.

V16. $I(X,Y) = H(Y), \quad I(X,Y) = H(X).$

Parlament: Který poslanec měl největší vliv na výsledek hlasování?

Systém: $\mathbf{S} = \{v_1, v_2, v_3, v_4, v_5\}$

Pozorování:

$$\begin{aligned} H_1 &= H(v_2, v_3, v_4, v_5 \mid v_1) \text{ tím menší, čím větší je vliv } v_1 \\ &= H(v_1, v_2, v_3, v_4, v_5) - H(v_1) \end{aligned}$$

Výsledky:

v_i	H_i
v_1	2.6423
v_2	2.6292
v_3	2.6262
v_4	2.6310
v_5	2.6310

Pozn: $H(v_1, v_2, v_3, v_4, v_5) = 3.3191$ [Nat]

Normovaná střední vzájemná informace

$$i(X,Y) = \frac{I(X,Y)}{H(X,Y)}, \qquad 0 \le i(X,Y) \le 1$$

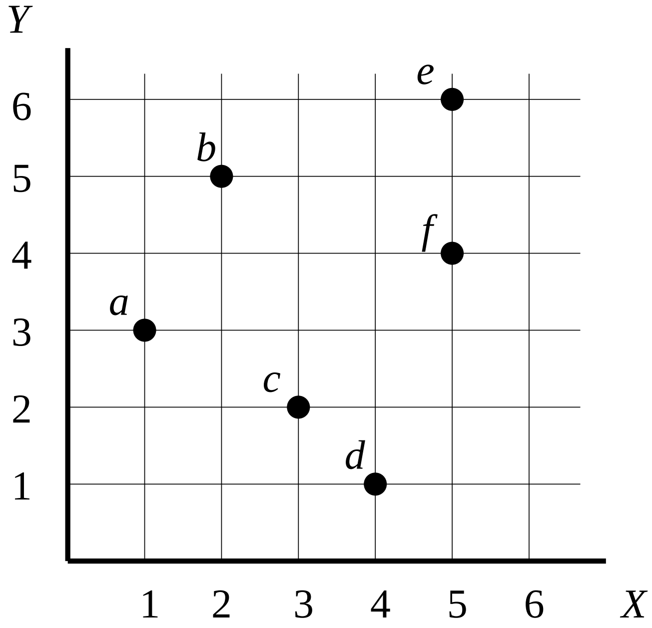
$i(X,Y) = 0$ pokud X, Y jsou nezávislé

$i(X,Y) = 1$ pokud mezi X, Y existuje jednoznačná deterministická funkční závislost

Dva aspekty relace mezi proměnnými

1. síla asociace $i(X,Y)$: jak moc jsou závislé?
2. statistická významnost této asociace: postačují data k takovému závěru?

Příklad



$$
\begin{aligned}
X &= \{1, 2, 3, 4, 5, 5\} \\
Y &= \{3, 5, 2, 1, 6, 4\} \\
X \cdot Y &= \{(1, 3), (2, 5), (3, 2), (4, 1), (5, 6), (5, 4)\} \\
H(X) &= 1.5607 \\
H(Y) &= 1.7918 \\
H(X, Y) &= 1.7918 \\
I(X, Y) &= 1.5607
\end{aligned}
$$

normovaná stř. vzájemná inf.	$i(X, Y) = 0.8710$
normalizovaný korelační koeficient	$\rho(X, Y) = 0.1964$
Spearmanův rankový koeficient	$r(X, Y) = 0.2609$

Entropie zobecněného dynamického systému

Stavový prostor $\mathbf{S} = S_1 \times S_2 \times \cdots \times S_k$

- Množina všech přípustných stavů $s \in \mathbf{S}$ tvoří elementární rozklad $\mathcal{S}$
- Každá proměnná s_i definuje rozklad na třídy ekvivalence $\mathcal{S}_i \ \ (= \mathcal{A}_{s_i})$, kde $\mathcal{S} \prec \mathcal{S}_i$.
- Entropie systému $\mathbf{T} = \{s_1, s_2, \ldots, s_k\}$ je sdružená entropie všech vzorkovacích proměnných

$$H(\mathbf{T}) = H(\mathcal{S}_1 \cdot \mathcal{S}_2 \cdots \mathcal{S}_k) = H(s_1, s_2, \ldots, s_k)$$

- Generativní nejistota systému $\mathbf{T} = \{\bar{\mathbf{G}}, \mathbf{G}\}$ je podmíněná entropie

$$H(\mathbf{G} \mid \bar{\mathbf{G}})$$

$\mathbf{G} = (s_1,\, s_2,\, s_n)$ množina generovaných vzorkovacích proměnných
$\bar{\mathbf{G}} = (s_{n+1},\, s_{n+2},\, s_k)$ množina generujících vzorkovacích proměnných

Pozn: $\{\mathbf{G}, \bar{\mathbf{G}}\}$ je rozklad množiny vzorkovacích proměnných (neplést s rozkladem stavového prostoru)

Odhad entropie z histogramu

Máme histogram $\{n_1, n_2, \ldots, n_k\}$ proměnné x se šířkou přihrádky $h > 0$.

Platí $n = \sum_{i=1}^{k} n_i$

2 případy:

1. diskrétní náhodná proměnná:

$$\hat{H}(x) = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

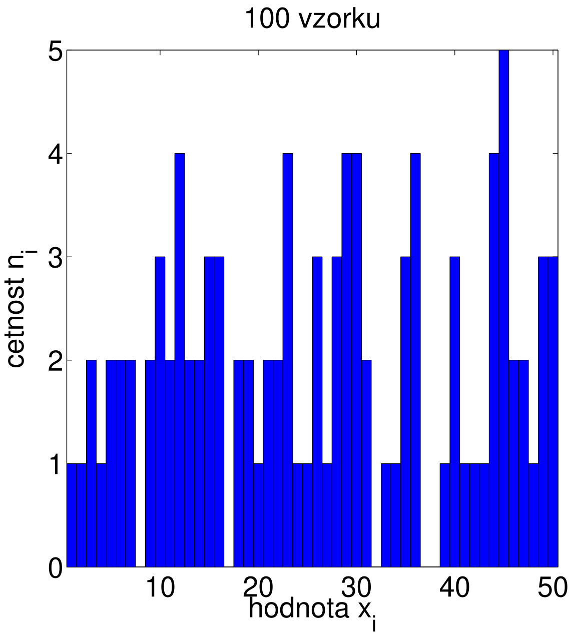
2. spojitá náhodná proměnná

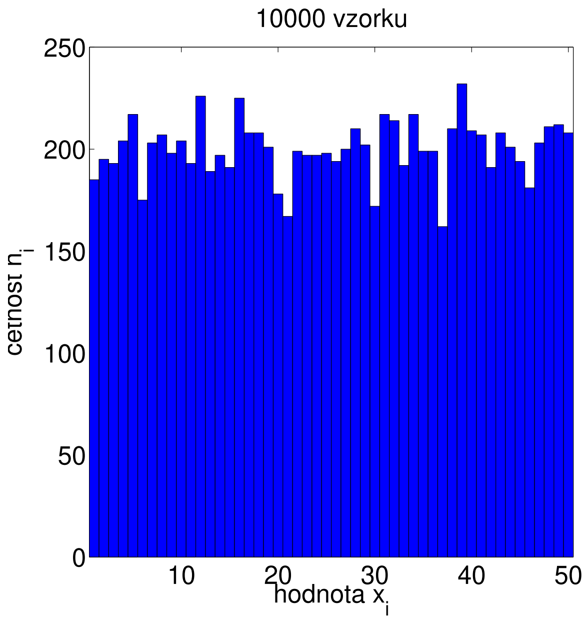
$$\hat{H}(x) = \log h - \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

Bez členu $\log h$ by hodnota statistiky rostla se zmenšováním rozlišení histogramu h

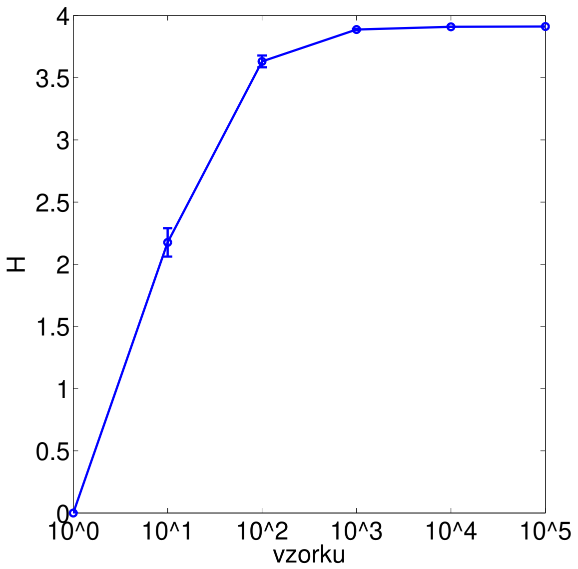
Je histogram kvalitním odhadem rozdělení psti?

$x \in \{1, \ldots, 50\}$: náhodná proměnná s rovnoměrným rozdělením





$\hat{H} = 3.6874$ nat $\qquad \hat{H} = 3.9095$ nat



- teoretická entropie $\log(50) = 3.9120$ nat
- hodnota četnosti je náhodná proměnná

Kontingenční analýza

Jaká je pravděpodobnost, že v i-té přihrádce histogramu bude n_i hodnot když celkem udělám n měření (pokusů)?

Sekvence, jejíž prvky jsou četnosti náhodných pokusů:

$$\left\{\mathcal{E}_i \text{ se vyskytne } n_i \text{ krát v daném pořadí}\right\}_{i=1}^{k}$$

pravděpodobnost takové sekvence

$$p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots p_k^{n_k}$$

$$\left\{\mathcal{E}_i \text{ se vyskytne } n_i \text{ krát v libovolném pořadí}\right\}_{i=1}^{k}$$

pst, že v 1. přihrádce je n_1 hodnot, ve 2. přihrádce n_2 hodnot,. . .

$$P(x_1 = n_1, x_2 = n_2, \ldots, x_k = n_k) = \frac{n!}{n_1!\, n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

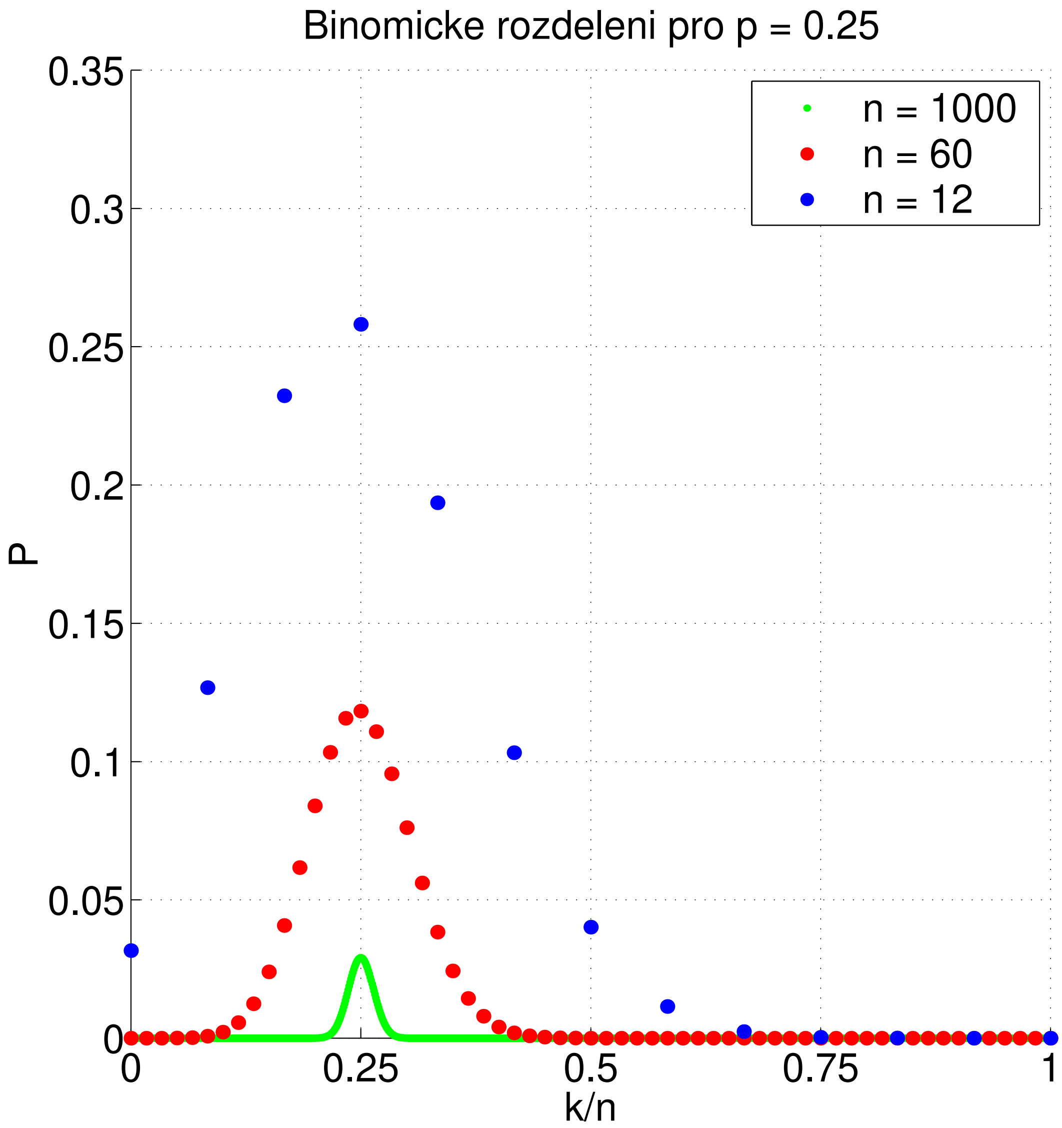
(permutace s opak.)

Multinomické rozdělení s parametry $n, p_1, p_2, \ldots, p_k$.

Příklad

Jaké hodnoty relativní četnosti k/n mohu očekávat v 1. přihrádce dvoupřihrádkového ‚histogramu' , když se hodnota vyskytuje s pravděpodobností $p_1 = 0.25$ (a druhá s $p_2 = 0.75$)?

$$P(x_1 = k) = \binom{n}{k} p_1^k \, (1 - p_1)^{n-k}$$



Vlastnosti multinomického rozdělení

Nechť $H = \{n_1, n_2, \ldots, n_k\}$ má multinomické rozdělení a $k-1$ je počet nezávislých prvků v H. Pak:

$$E\left(\frac{n_i}{n}\right) = p_i$$

$$\mathrm{var}\left(\frac{n_i}{n}\right) = \frac{p_i(1-p_i)}{n}$$

$$\mathrm{cov}\left(\frac{n_i}{n}, \frac{n_j}{n}\right) = -\frac{p_i p_j}{n}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i}$$

Veličina χ^2 má při $n \to \infty$ asymptoticky rozdělení χ^2_{k-1} s hustotou:

$$f_m(x) = \frac{x^{\frac{m}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{m}{2}}\ \Gamma\left(\frac{m}{2}\right)}$$

($m = k-1$: počet nezávislých prvků v $\{n_1, \ldots, n_k\}$)

Funkce inverzní k distribuční je $Q(\frac{m}{2}, \frac{x}{2})$.

Neúplná gama funkce

$$Q(a,x) = \frac{1}{\Gamma(a)} \int_x^\infty e^{-t}\, t^{a-1} dt, \quad a > 0$$

Kontingenční test

$$\begin{aligned} \text{Nulová hypotéza} \quad & H_0: \ \text{tvrzení } X \text{ platí} \\ \text{Alternativní hypotéza} \quad & H_1: \ \text{tvrzení } X \text{ neplatí} \end{aligned}$$

Chyba 1. druhu: zamítneme H_0, ač je správná

Chyba 2. druhu: nezamítneme H_0, ač není správná

Test hypotézy H_0 proti H_1: Pravděpodobnost chyby 1. druhu menší než $\alpha \in \langle 0, 1 \rangle$ (hladina testu)

Budeme například potřebovat:

- $H_0: \ p(\mathbf{a}, \mathbf{b}) = {}^1p(\mathbf{a}) \cdot {}^2p(\mathbf{b})$

  pro test nezávislosti subsystémů
- $H_0: \ p(\mathbf{a}, \mathbf{b}, \mathbf{c}) = {}^1p(\mathbf{a}, \mathbf{b}) \cdot {}^2p(\mathbf{c} \mid \mathbf{b})$

  pro test statistické významnosti rekonstrukce struktury systému

---

Pozn: $\{\mathbf{a}, \mathbf{b}\}$, $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ rozklady množiny vzorkovacích proměnných systému. Můžeme si představit, že $\mathbf{a}, \mathbf{b}, \mathbf{c}$ jsou vektorové proměnné.

Postup pro $p(\mathbf{a},\mathbf{b}) = {}^1p(\mathbf{a}) \cdot {}^2p(\mathbf{b})$

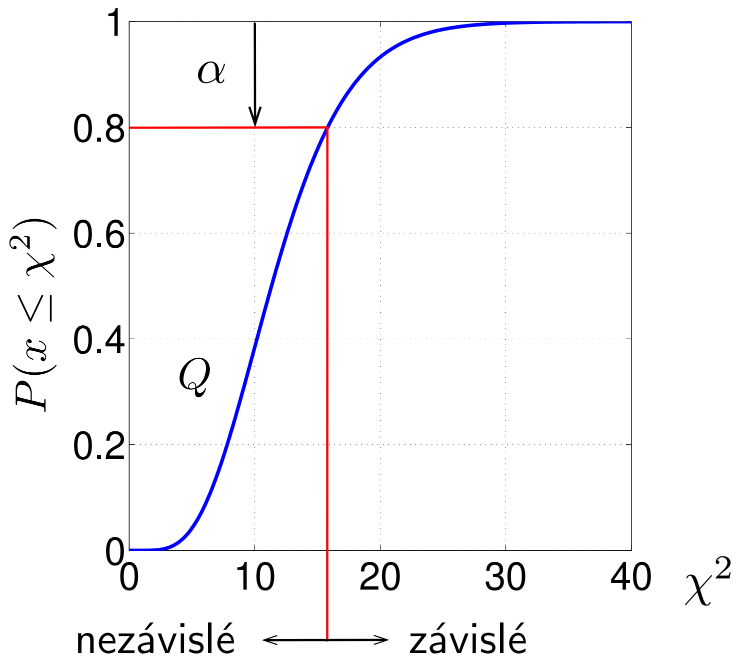
Z kontingenční tabulky vypočteme

skutečná četnost $n(a_i, b_j)$ četnost predikovaná modelem $n \cdot p(a_i) \cdot p(b_j)$

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

Jestliže

$$\chi^2 \geq \chi^2_{(r-1)(c-1)}(\alpha) = Q^{-1}\left(\frac{(r-1)(c-1)}{2}, \frac{\alpha}{2}\right)$$

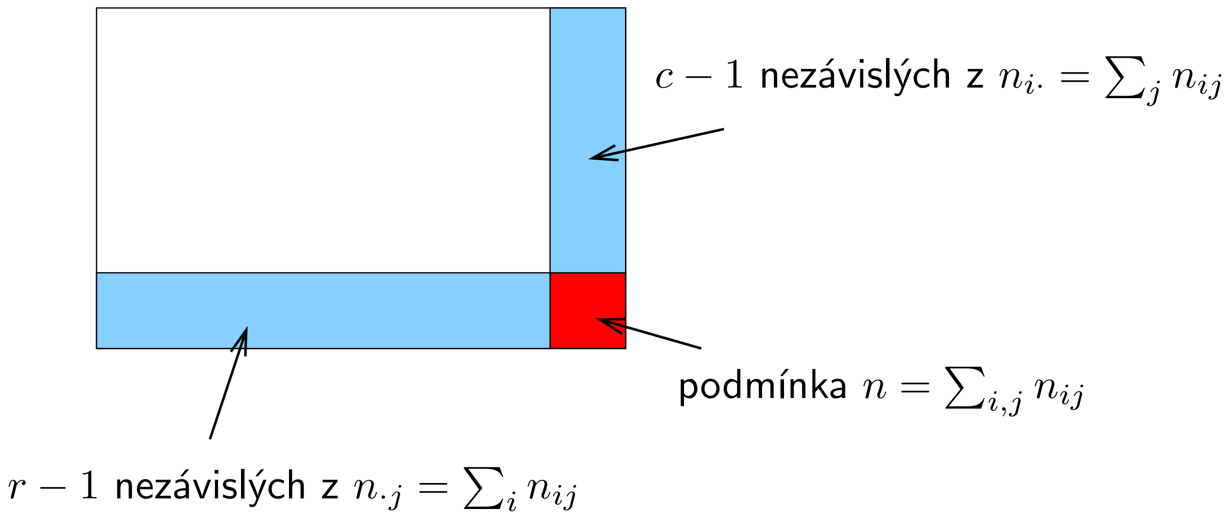


zamítneme hypotézu nezávislosti na hladině významnosti α typicky volíme $\alpha = 0.05$

(pokračování)

Počet stupňů volnosti

$rc - (r-1) - (c-1) - 1 = (r-1)(c-1)$



Anděl, J. Statistické metody, MATFYZPRESS Praha 1998.

Press, W.H. – Teukolsky S.A. –Vetterling, W.T. – Flannery, B.P. Numerical Recipes in C, Cambridge University Press. 1992.

Příklad

Q: Je okamžité chování racků R_1, R_2 nezávislé?

kontingenční tabulka

n_{ij}	1	2	3	4	n_j
0	1	0	3	1	5
1	4	1	3	1	9
2	3	1	2	4	10
3	0	0	6	1	7
4	0	1	1	12	14
n_i	8	3	15	19	$n = 45$

Volíme $\alpha = 0.01$

$$\chi^2 = 29.89 > 26.22 = \chi^2_{12}(0.01)$$

Na hladině významnosti 0.01 není okamžité chování obou jedinců nezávislé $\Rightarrow$ je nezávislé s pravděpodobností 0.01 nebo menší

Pozn:

- pozor na nulové sloupce nebo řádky
- v Matlabu použijeme `chi2inv(1-0.01,12)`
- $i(R_1, R_2) = 0.1506 \Rightarrow$ závislost je slabá, ale zjištěno s velkou jistotou

Příklad z parlamentu: která dvojice hlasuje nezávisle na ostatních?

1. Nalezení nezávislé dvojice $(i, j) \in \binom{N}{2}$

$$\arg\min_{i,j} I(\{s_i,\ s_j\}, \{s_k,\ s_l,\ s_m\})$$

2. Ověření statistické významnosti

$$H_0\colon\ \ p(s_1, s_2, s_3, s_4, s_5) = p(s_i, s_j) \cdot p(s_k,\ s_l,\ s_m)$$

s_i	s_j	I_{ij}
s_1	s_2	0.0277
s_1	s_3	0.0299
s_1	s_4	0.0308
s_1	s_5	0.0303
s_2	s_3	0.0294
s_2	s_4	0.0263
s_2	s_5	0.0300
s_3	s_4	0.0273
s_3	s_5	0.0293
s_4	s_5	0.0249

		s_4	0	0	1	1	
		s_5	0	1	0	1	
s_1	s_2	s_3					
0	0	0	2	4	2	5	13
0	0	1	3	3	9	3	18
0	1	0	3	5	4	1	13
0	1	1	3	3	4	5	15
1	0	0	1	3	3	3	10
1	0	1	1	3	0	1	5
1	1	0	3	3	3	4	13
1	1	1	1	4	3	5	13
			17	28	28	27	100

$\chi^2_{21} = 15.0677$, nezávislé s pstí 0.82

Volba šířky přihrádky histogramu

Systém $\mathbf{S} = \{s_1, s_2, \ldots, s_q\}$

q dimenze histogramu
n počet měření
$\hat{\sigma}_i$ odhad rozptylu proměnné s_i
h_i šířka přihrádky pro proměnnou s_i

Scottovo pravidlo

$$\frac{h_i}{\hat{\sigma}_i} \approx \frac{3.5}{\sqrt[2+q]{n}}$$

(předpoklad normálního rozdělení s diagonální kovarianční maticí)

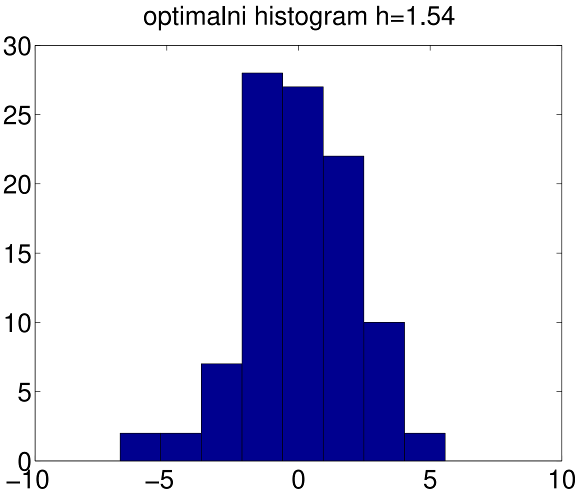
Scott, D. W. Multivariate Density Estimation: Theory Practice, and Visualization, John Wiley & Sons, Chichester 1992.

Příklad

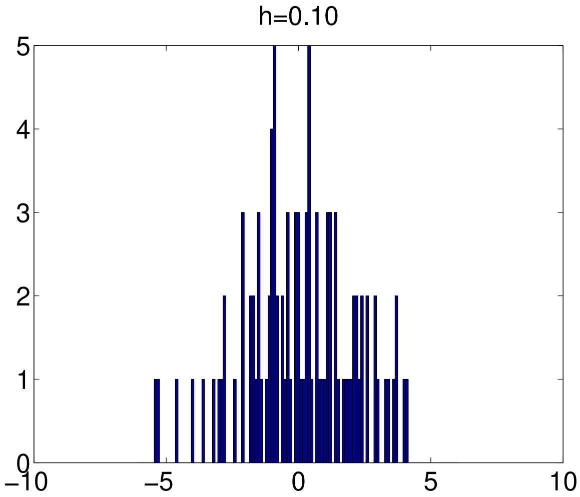
x – <u>spojitá</u> náhodná proměnná s rozdělením $N(0,2)$

$H(x) = \log \sigma\sqrt{2\pi e} = 2.1121$

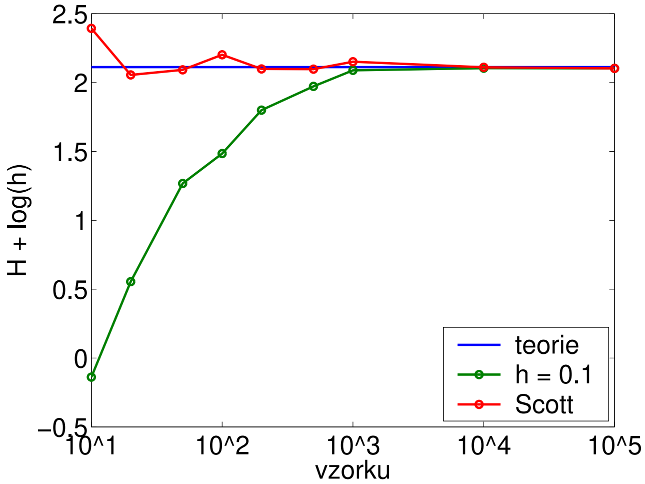
$n = 100$ vzorků, dimenze $q = 1$



$\hat{H} = 2.1279$ nat



$\hat{H} = 2.1028$ nat



- entropie z optimálního histogramu je lepším odhadem $H(x)$