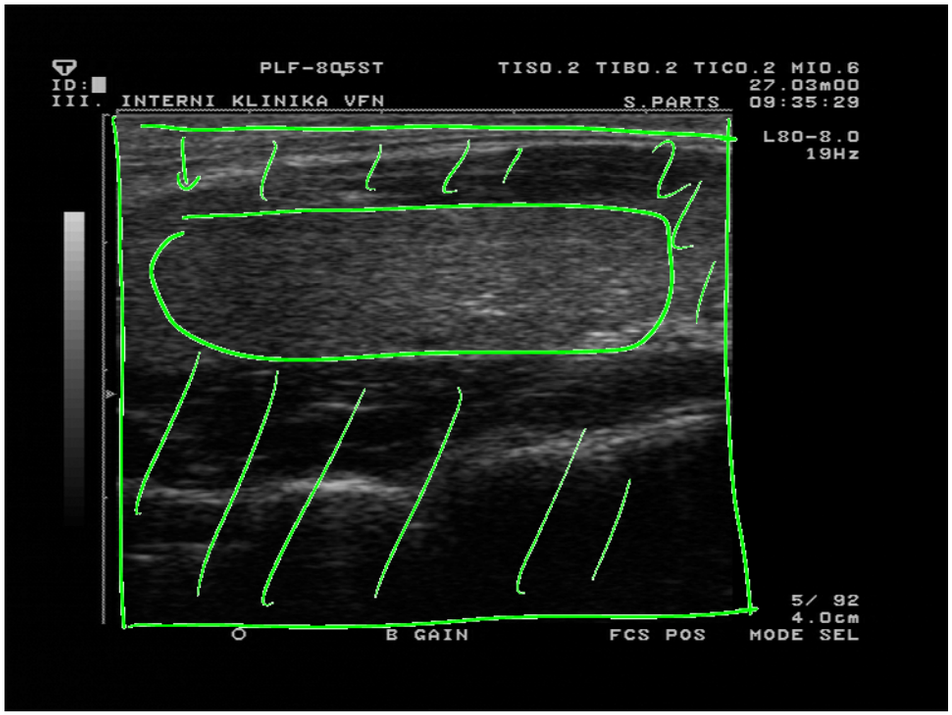
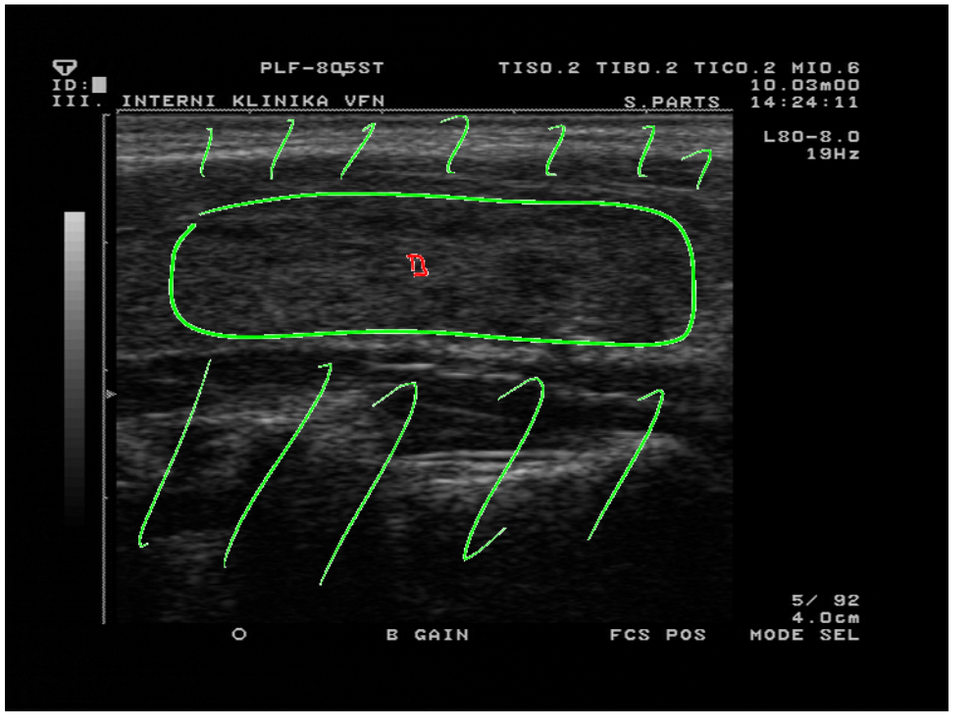
Aplikace 2: Hledání informativních příznaků pro rozpoznávání

Sonogram štítné žlázy v podélném řezu

zdravá



lymfocitická thyroitida



Zajímá nás, kolik se lze z dat dozvědět o třídě c a „kde“ ta informace je.

Příznaky $\mathbf{x}_i$, $i = 1, 2, \ldots n$.

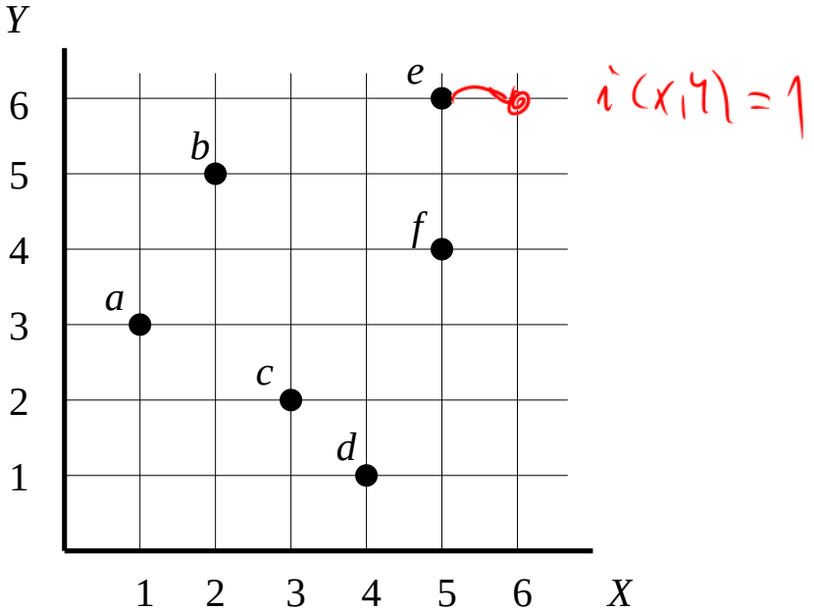
Informace o třídě v příznaku $\mathbf{x}_i$ je podmíněná entropie $H(c \mid \mathbf{x}_i)$.

Aplikace 3: 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 bijekce



$i(X,Y) = 1$

$$X = \{1,2,3,4,5\}, \quad Y = \{3,5,2,1,6,4\}$$

$$X \cdot Y \supseteq \{(1,3),(2,5),(3,2),(4,1),(5,6),(5,4)\}$$

$$H(X) = 1.5607 \text{ nat}, \quad H(Y) = 1.7918 \text{ nat}$$

$$H(X,Y) = 1.7918 \text{ nat}, \quad I(X,Y) = 1.5607 \text{ nat}$$

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

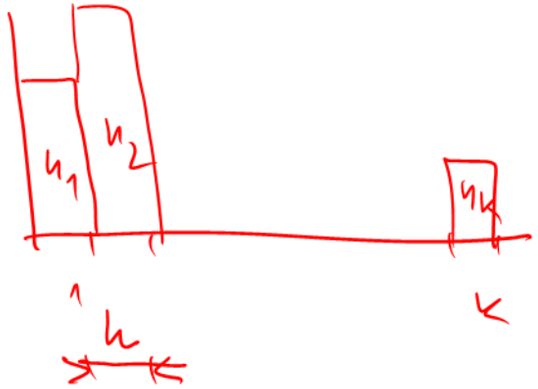
Dva aspekty relace mezi proměnnými

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

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:

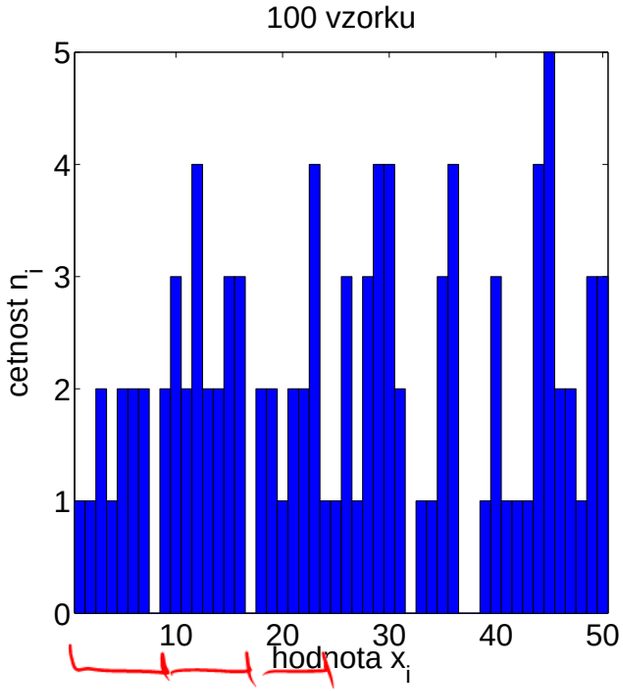
1. diskrétní náhodná proměnná: chceme entropii v přirozeném rozlišení
2. kvantizovaná spojitá náhodná proměnná: chceme entropii původní spojité proměnné

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

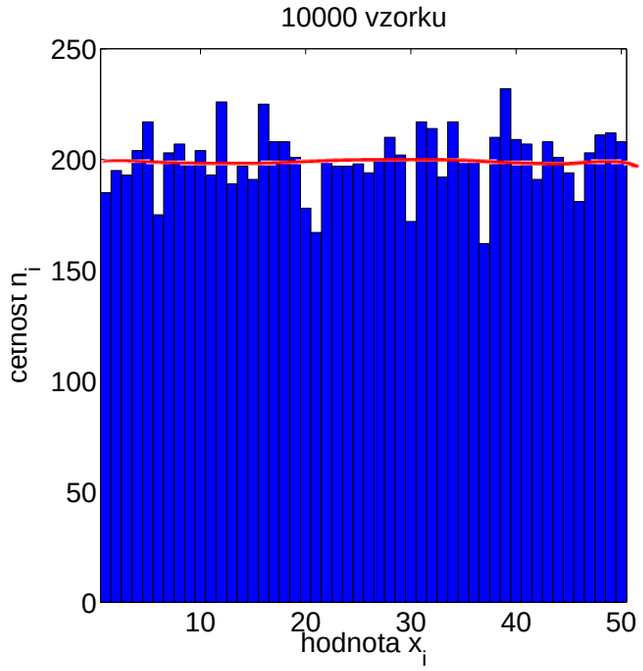
- h . . . v přirozených jednotkách oboru hodnot
- Bez členu $\ln h$ by hodnota statistiky rostla se zmenšováním rozlišení histogramu h

Je histogram kvalitním odhadem rozdělení psti?

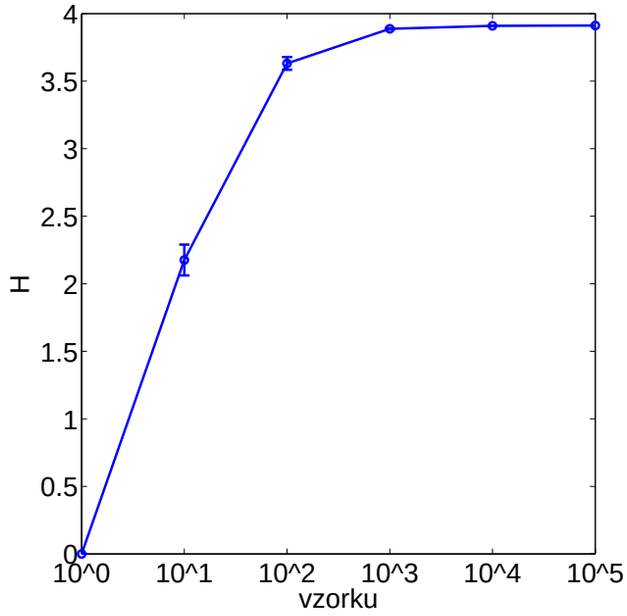
$x \in \{1, \ldots, 50\}$: náhodná proměnná s rovnoměrným rozdělením



$\hat{H} = 3.6874$ nat



$\hat{H} = 3.9095$ nat



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

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

Předpoklad normálního rozdělení s diagonální kovarianční maticí

$$f(\underline{\mathbf{x}}) = \frac{1}{\sqrt{\det(2\pi\mathbf{S})}}\, e^{-\frac{1}{2}(\underline{\mathbf{x}}-\bar{\underline{\mathbf{x}}})^\top \mathbf{S}^{-1}(\underline{\mathbf{x}}-\bar{\underline{\mathbf{x}}})}$$

$$\mathbf{S} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2)$$

Scottovo pravidlo

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

obyčejný histogram $h = \dfrac{3.5\,\hat{\sigma}}{\sqrt[3]{n}}$

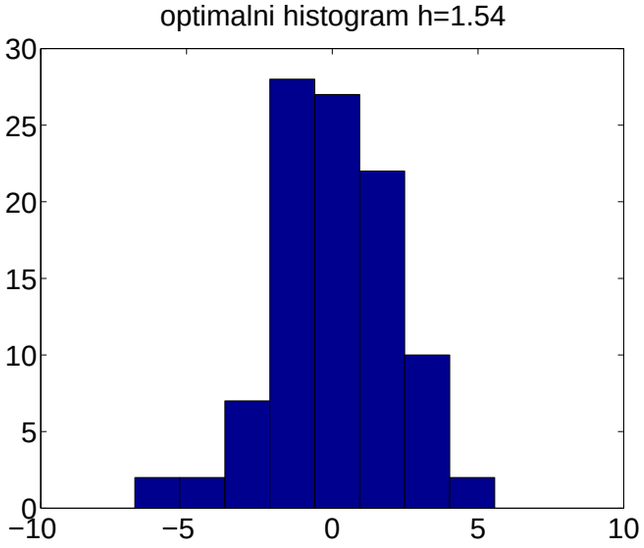
Scott, D. W. Multivariate Density Estimation: Theory Practice, and Visualization, John Wiley & Sons, Chichester 1992.

Příklad

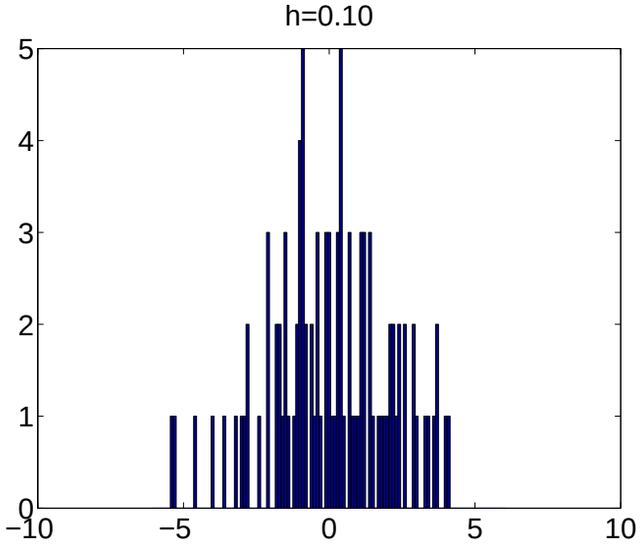
x – spojitá náhodná proměnná s rozdělením $N(0, 2)$

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

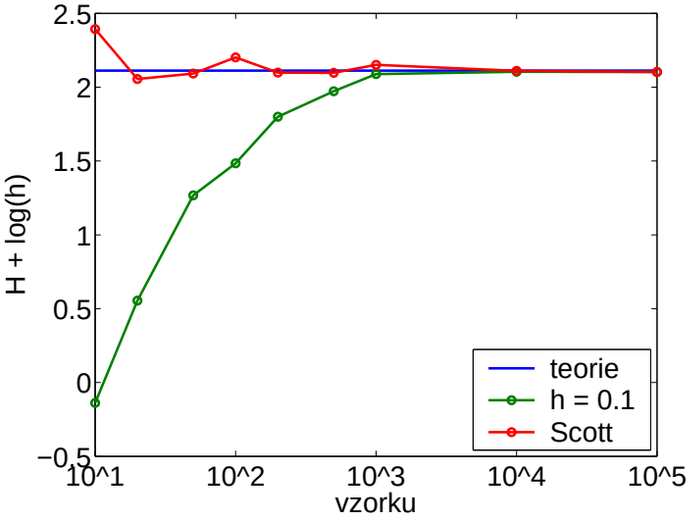
$n = 100$ vzorků, dimenze $q = 1$, takže $h = \frac{3.5 \cdot 2}{\sqrt[3]{100}} \approx 1.51$



$\hat{H} = 2.1279$ nat



$\hat{H} = 1.3026$ nat



Scott: pokaždé vypočteme novou hodnotu h a přepočteme histogram

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

Estimátor entropie bez histogramování (Kozačenko-Leoněnko)

Dáno: množina vektorových měření $\{\underline{\mathbf{x}}_i,\ i = 1, \ldots, n\}$ z neznámého spojitého rozdělení pravděpodobnosti

Cíl: výpočet entropie bez diskretizace a histogramování

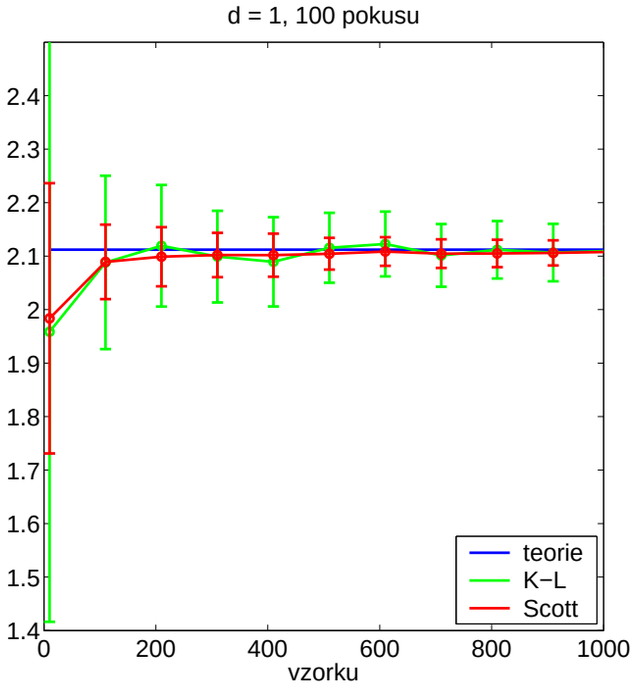
q dimenze vektoru měření
n počet měření
r_i euklidovská (L_2) vzdálenost k nejbližšímu sousedu $\underline{\mathbf{x}}_i$
γ Euler-Mascheroniho konstanta ($\gamma \approx 0.5772156649$)

$$H = \frac{q}{n} \sum_{i=1}^{n} \ln r_i + \ln \frac{(n-1)\pi^{\frac{q}{2}}}{\Gamma(1 + \frac{q}{2})} + \gamma$$

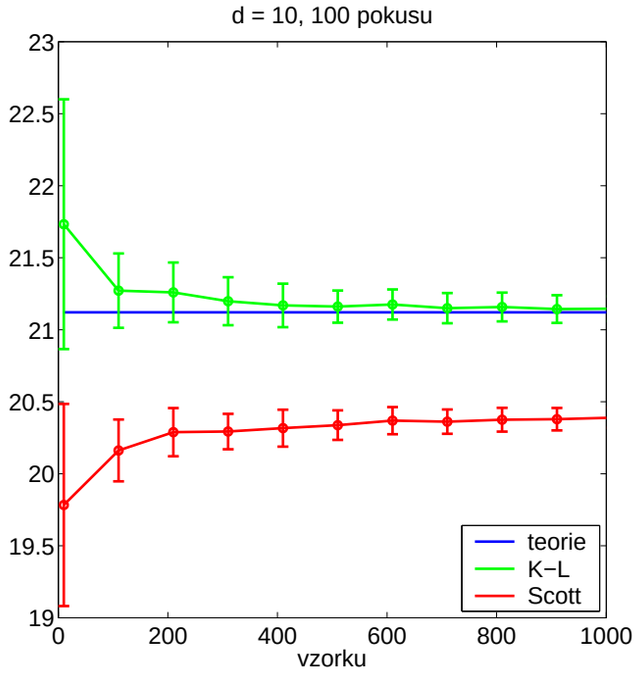
Poznámky

- množina nejbližších sousedů pro všechna $\underline{\mathbf{x}}_i$ lze teoreticky nalézt za dobu $O(c^q\, n \log n)$ pro libovolné q.
- degenerovanost: použít $\ln \max(r_i, \frac{1}{\sqrt{n}})$ místo $\ln r_i$.
- vhodné pro velká q (viz příklad)
- vhodné pro multimodální rozdělení psti
- $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,\ x \in R, \Gamma(k) = (k-1)!,\ k \in N$

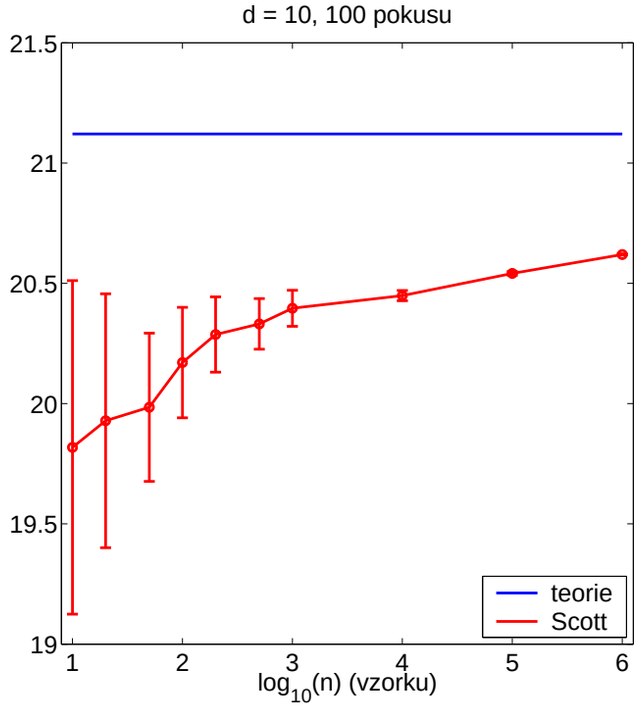
Příklad



$\underline{\mathbf{x}} \sim N(0, 2),\ q = 1$



$\underline{\mathbf{x}} \sim N(0, 2),\ q = 10$



- malá dimenze q: histogramovací metoda má lepší rozptyl
- velká dimenze q: KL je méně vychýlený, rozptyly srovnatelné
- velká q: histogramová metoda konverguje (pro velmi velká n)

Odhad entropie a její chyby

Dáno: množina $\mathcal{D} = \{\underline{\mathbf{x}}_i,\ i = 1, 2, \ldots, n\}$

Cíl: odhad entropie $\hat{H}(\mathcal{D})$ včetně chyby $\mathrm{var}[\hat{H}(\mathcal{D})]$

Jackknife

1. Pro $i = 1, 2, \ldots, n$ dělej:
   a. zkonstruuj $\mathcal{D}_i = \mathcal{D} \setminus \{\underline{\mathbf{x}}_i\}$ vynecháním jednoho bodu
   b. odhadni $\hat{H}_i = \hat{H}(\mathcal{D}_i)$ z $\mathcal{D}_i$
2. Vypočti odhad entropie $\hat{H}$ a chyby $\mathrm{var}[\hat{H}]$:

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n} \hat{H}_i, \qquad \mathrm{var}[\hat{H}] = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{H}_i - \hat{H}\right)^2$$

Duda, Hart, Stork Pattern Recognition

Poznámky

- jackknife může být použit na jakoukoliv statistiku, nejen entropii, např. na medián,. . .
- Je to metoda *Resampling Theory*

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

$$p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots p_k^{n_k}$$ víme, v jakém pořadí padaly kuličky do přihrádek

---

$$\left\{\mathcal{E}_i \text{ se vyskytne } n_i \text{ krát v libovolném pořadí}\right\}_{i=1}^{k}$$ nevíme, v jakém pořadí padaly do přihrádek

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

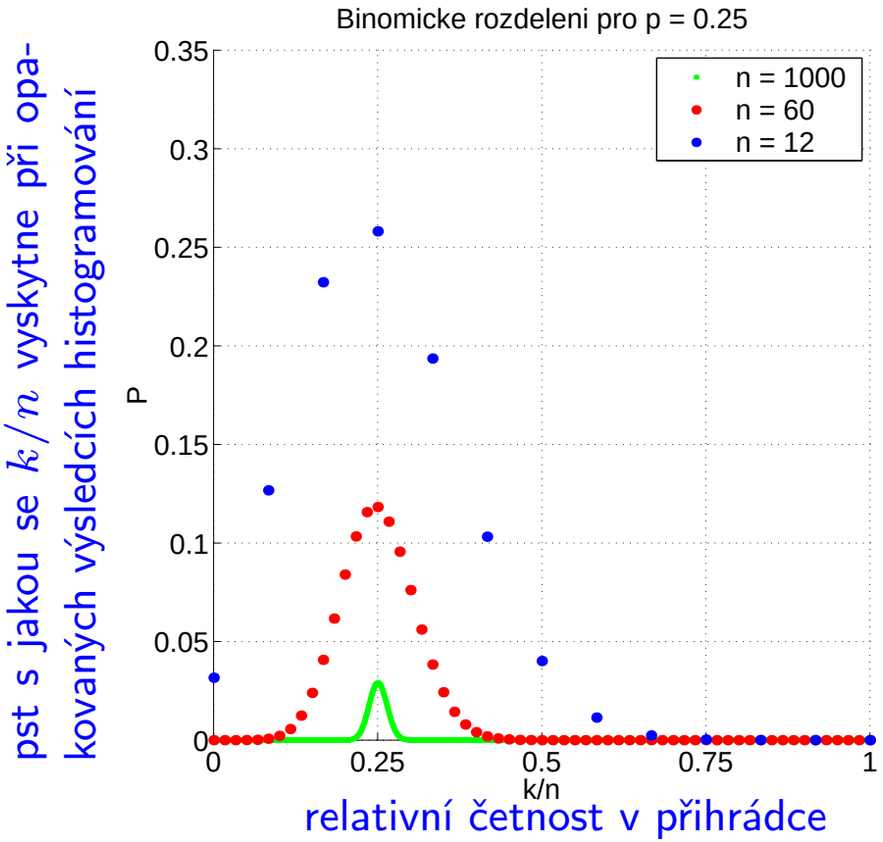
$$P(x_1 = n_1, x_2 = n_2, \ldots, x_k = n_k) = \frac{n!}{n_1!\, n_2! \cdots n_k!}\, p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$ permutace s opak.

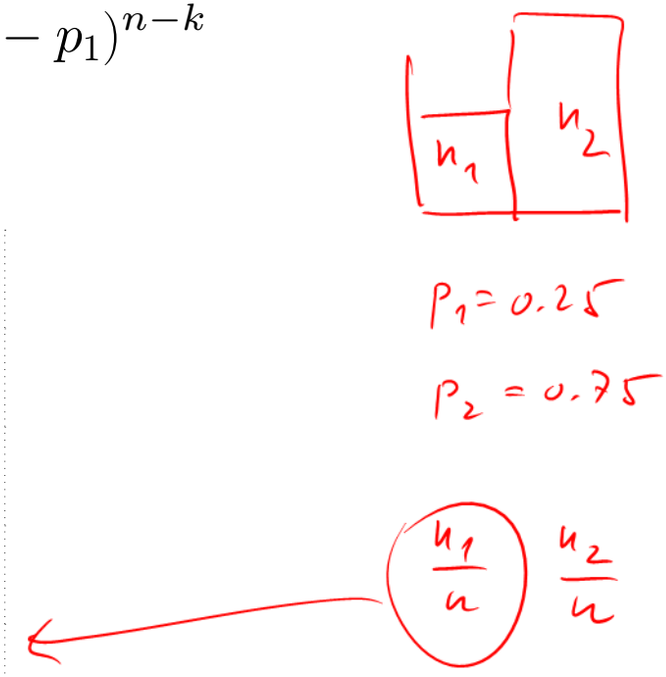
To je **multinomické rozdělení** s parametry $n, p_1, p_2, \ldots, p_k$.

Příklad

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

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





- nejistotu musíme brát v úvahu, když činíme nějaký závěr z relativních četností

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:

$$E\left(\frac{n_i}{n}\right) = p_i \qquad \text{var}\left(\frac{n_i}{n}\right) = \frac{p_i\,(1-p_i)}{n}$$

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

$$q = \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i} \qquad \text{(Pearson)}$$

Veličina q 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 $Q(\frac{m}{2}, \frac{x}{2})$, tj. neúplná gama funkce

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

Pearsonova Statistika

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

n_i – nahistogramované četnosti
p_i – model — tj., co v histogramu očekáváme

Př:

1. 2-D histogram nezávislých veličin x, y, pak $p_{ij} = p_i \cdot p_j = p(x = x_i) \cdot p(y = y_j)$ a my použijeme odhad p_{ij} modelu

$$p_{ij} = \frac{n_i}{n} \cdot \frac{n_j}{n}$$

$n_i = \sum_j n_{ij}$ $\quad n_j = \sum_i n_{ij}$

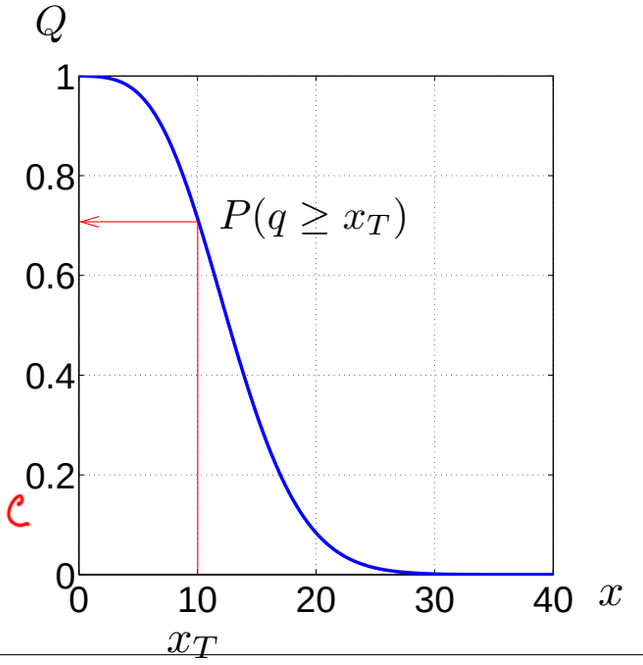
2. $p_i = p(x_i \mid \Theta)$ (možné, ale nepraktické)

$$q = \sum_i \sum_j \frac{\left(n_{ij} - n\,\frac{n_i}{n} \cdot \frac{n_j}{n}\right)^2}{\frac{n_i\,n_j}{n}}$$

Potom

$$Q\left(\frac{m}{2}, \frac{x}{2}\right) = P(q \geq x)$$

je pravděpodobnost, že změřená hodnota statistiky q je větší než práh x, za předpokladu platnosti modelu.



m – počet přihrádek minus počet dodatečných podmínek, které musí splňovat soubor $\{n_i\}$ a které jsou potřeba k výpočtu hodnoty p_i
(např. $\sum_{i=1}^{k} n_i = n$ a $\sum_{i=1}^{k} n_{ij} = n_i$ pro Př. 1).

$m = r \cdot c - r - c + 1$

Standardní kontingenční test

Nulová hypotéza H_0: tvrzení X platí

Chyba: „Odmítnu H_0, a (ale) H_0 ve skutečnosti platí“ — chyba 1. druhu

Cíl: $P(\text{error}) \leq \alpha$ — α: hladina významnosti

α: penále, které musím zaplatit, když udělám chybu.

Řešení: Procedura statistického testu

1. vyslov H_0 — např. složky $\underline{\mathbf{x}}$ jsou statisticky nezávislé
2. změř n hodnot $\mathcal{D} = \{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_n\}$
3. vypočti nějakou statistiku q z $\mathcal{D}$ — např. Pearsonovu statistiku
4. zvol (malé) α — typicky $\alpha = 0.01$ nebo $\alpha = 0.05$
5. odmítni H_0 když $q \geq T(\alpha)$ — musím znát $T(\cdot)$, $T = Q^{-1}$

- q roste s rostoucí odchylkou dat od H_0
- T monotonicky klesá s α — $\longrightarrow$ dostanu velké T pro malé α
- Což znamená: ‘Jsem velmi tolerantní a odmítnu H_0 jen, když je ve zřejmém rozporu s daty.’

Náš problém

Například:

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

Procedura testu

1. vyslov H_0
2. změř n hodnot $\mathcal{D} = \{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_n\}$
3. vypočti Pearsonovu statistiku q z $\mathcal{D}$
4. vypočti α takové, že $q = T(\alpha)$
5. je-li dáno $\mathcal{D}$, H_0 platí s pravděpodobností alespoň α
   - odmítnutím H_0 udělám chybu α: $P(\text{odmítnu} \wedge \text{platí}) = \alpha$
   - malé $\alpha \Rightarrow$ mohu odmítnout
   - velké $\alpha \Rightarrow$ nemohu odmítnout $=$ musím přijmout
   - $P(H_0 \text{ platí}) = P(\text{přijmu} \wedge \text{platí}) + P(\text{odmítnu} \wedge \text{platí}) \geq \alpha$

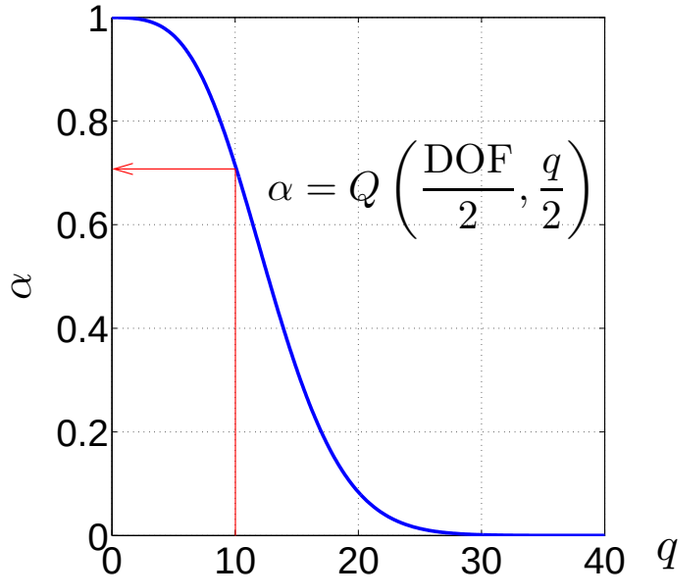
Postup pro $p(\mathbf{a},\mathbf{b}) = {}^1p(\mathbf{a}) \cdot {}^2p(\mathbf{b})$

1. Z kontingenční tabulky vypočteme

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

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

2. stupně volnosti: $\mathrm{DOF} = (r-1)(c-1)$
3. vypočteme $\alpha = Q\left(\frac{\mathrm{DOF}}{2}, \frac{q}{2}\right)$
4. vyjde-li malé α, pak tvrdím, že $\mathbf{a}$ a $\mathbf{b}$ závislé
5. vyjde-li velké α, pak tvrdím, že $\mathbf{a}$ a $\mathbf{b}$ jsou nezávislé s pravděpodobností alespoň α

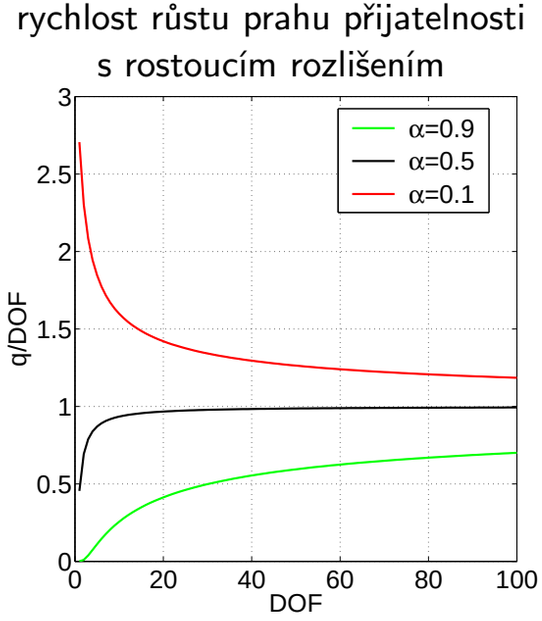


poznámky

Počet stupňů volnosti $\mathrm{DOF} = rc - (r + c) + 1 = (r - 1)(c - 1)$

- Máme rc prvků v tabulce, ale použili jsme dodatečné vztahy $n_{.j} = \sum_i n_{ij}$ a $n_{i.} = \sum_j n_{ij}$, kterých je dohromady $r + c$.
- Ale tyto podmínky nejsou nezávislé, protože $\sum_j n_{.j} + \sum_i n_{i.} = n$, odečteme 1.

$$\frac{T(\alpha, \mathrm{DOF})}{\mathrm{DOF}} :$$



mnoho měření ⇒ velký DOF ⇒ $T \approx \mathrm{DOF}$

Anděl, J. Statistické metody, MATFYZPRESS Praha 1998.

Press, WH. – Teukolsky SA. et al. Numerical Recipes in C, Cambridge University Press. 1992.