Neparametrické metody - testy, rozdělení s kategoriálními proměnnými

Tato oblast obsahuje různé statistické techniky, které je možno rozdělit:

1. Testy shody (Goodness of Fit Tests) – obsahují testy shody skupinových a modelových četností, testy shody modelové distribuční funkce a empirické distribuční funkce, testy o empirické distribuční funkci, … .
2. Testy založené na pořadových a rankových statistikách, testy náhodnosti, testy nezávislosti, … .
3. Metody založené na teorii informace.
4. Metody založené na jádrových odhadech hustoty.
5. …. .

Opět v této přednášce budou uvedeny jen příklady a to některých metod.

χ^2 test dobré shody.

Mějme náhodnou veličinu ξ a náhodný výběr $\{x_1, x_2, ..., x_n\}$. Prostor možných hodnot této náhodné proměnné je rozdělen do M disjunktních skupin tvořících pokrytí definičního oboru této náhodné proměnné (jedná se tedy o rozklad definičního oboru). Předpokládejme, že jsou dostupné pravděpodobnosti $p_j, j = 1,...,M$ patření ke každé ze zavedených skupin. Dále: $n_j, j = 1,...,M$ je počet prvků náhodného výběru patřících do j-té skupiny. Evidentně platí: $\sum_{j=1}^{M} n_j = n; \quad \sum_{j=1}^{M} p_j = 1$. Zatímco n_j jsou pozorované četnosti, np_j jsou teoretické (očekávané) četnosti.

Četnosti $n_j, j = 1,...,M$ mají multinomické rozdělení

$$P(n_1, n_2, ..., n_M \,/\, p_1, p_2, ..., p_M) = \frac{n!}{(n_1)!(n_2)!...(n_m)!} p_1^{n_1} p_2^{n_2} ... p_M^{n_M}; \sum_{i=1}^{M} n_i = n; \sum_{i=1}^{M} p_i = 1.$$

Náhodná proměnná

$$\eta = \sum_{j=1}^{M} \frac{(n_j - np_j)^2}{np_j} = \sum_{j=1}^{M} \frac{n_j^2 + n^2 p_j^2 - 2nn_j p_j}{np_j} = \sum_{j=1}^{M} \left(\frac{n_j^2}{np_j} + np_j - 2n_j \right) = \sum_{j=1}^{M} \left(\frac{n_j^2}{np_j} \right) + n - 2n =$$

$$= \sum_{j=1}^{M} \left(\frac{n_j^2}{np_j} \right) - n$$

má asymptoticky χ^2 rozdělení s $M-1$ stupni volnosti, tedy $\chi^2(M-1)$. Důkaz tohoto tvrzení lze nalézt např. v: Jaroslav Hátle, Jiří Likeš. Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974. Str. 340-342. Za dostatečně velká n se považují taková n, pro které platí $np_j > 5, j = 1,...,M$.

Příklad:
Mějme náhodný výběr $\{x_1, x_2, ..., x_n\}$ z „neposunutého“ exponenciálního rozdělení s hustotou $f_\xi(x)=\frac{1}{\tau}e^{-\frac{x}{\tau}}$ a distribuční funkcí $F_\xi(x)=1-e^{-\frac{x}{\tau}}$. Dále mějme dělení $z_1, z_2, ..., z_{M-1}$ kladné reálné poloosy splňující následující podmínky:

$$nF_\xi(z_1)=n\left(1-e^{-\frac{z_1}{\tau}}\right)>5\Rightarrow\left(1-e^{-\frac{z_1}{\tau}}\right)>\frac{5}{n}\Leftrightarrow 1-\frac{5}{n}>e^{-\frac{z_1}{\tau}}\Leftrightarrow \lg\left(1-\frac{5}{n}\right)>-\frac{z_1}{\tau}\Leftrightarrow z_1>-\tau\lg\left(1-\frac{5}{n}\right), \textbf{(a)}$$

$$n\left(F_\xi(z_i)-F_\xi(z_{i-1})\right)>5; i=2,...,M-1 \text{ a} \quad \textbf{(b)}$$

$$n\left(1-F_\xi(z_{M-1})\right)>5 \quad \textbf{(c)}$$

Obvykle se takové dělení zvolí tak, že se nalezne nejmenší číslo z_1 splňující nerovnost (a), následně se volí čísla nejmenší $z_i, i=2,...$ splňující postupně nerovnosti (b) a to, tak dlouho, pokud platí $n\left(1-F_\xi(z_i)\right)>5$. Tím se nalezne nejemnější dělení kladné poloosy splňující podmínku $np_j>5, j=1,...,M$ a i počet skupin M při daném n. Potom:

$n_1=\left|\{x_i : x_i\in\{x_1,x_2,...,x_n\}; x_i<z_1\}\right|$

$n_j=\left|\{x_i : x_i\in\{x_1,x_2,...,x_n\}; z_{j-1}\le x_i<z_j\}\right|; j=2,...M-1$

$n_M=\left|\{x_i : x_i\in\{x_1,x_2,...,x_n\}; z_{M-1}\le x_i\}\right|; i=2,...M-1$

Testujeme hypotézu: pozorování z náhodného výběru $\{x_1, x_2, ..., x_n\}$ patří rozdělení s distribuční funkcí $F_\xi(x)=1-e^{-\frac{x}{\tau}}$ proti alternativě: pozorování z náhodného výběru $\{x_1, x_2, ..., x_n\}$ patří libovolnému jinému rozdělení se shodným definičním oborem.
Testové kritérium pak bude vypadat následovně:
$\eta=\sum_{j=1}^{M}\frac{(n_j-np_j)^2}{np_j}$, při pravděpodobnosti chyby prvního druhu α a jí odpovídající kritický obor: $W_\alpha=\{\eta\ge\chi^2_{1-\alpha}(M-1)\}$, kde $\chi^2_{1-\alpha}(M-1)$ je $1-\alpha$ kvantil χ^2 s $M-1$ stupni volnosti.

Námět: lze v tomto případě činit nějaké výroky o síle testu, případně i silofunkci?

Vzhledem k tomu, že rozdělení definičního oboru na skupiny je až na podmínku $np_j>5, j=1,...,M$ libovolné, je nezbytné si uvědomit to co je v základě jakéhokoliv testování hypotézy proti alternativě. Pro hypotézu musí být k dispozici ještě něco jiného (nestatistického) co ji preferuje před alternativou. Tedy výrok o zamítnutí hypotézy je výrokem s daleko větší váhou (u testů dobré shody) než výrok o přijmutí hypotézy!

Modifikovaný χ^2 test dobré shody.

Často potřebujeme řešit úlohu o tom, že daný náhodný výběr přísluší náhodné proměnné s nějakým typem parametrického rozdělení. Např. náhodný výběr je z „neposunutého“ exponenciálního rozdělení s hustotou $f_\xi(x)=\frac{1}{\tau}e^{-\frac{x}{\tau}}$ a distribuční funkcí $F_\xi(x)=1-e^{-\frac{x}{\tau}}$ a to pro jakoukoliv hodnotu parametru τ.

V tomto případě se pak bere za hodnotu daného neznámého parametru jeho „maximálně věrohodný odhad“ a k němu se určí odpovídající pravděpodobnosti patření do jednotlivých tříd.

Modifikovaný χ^2 test dobré shody – jedno-parametrická třída rozdělení

Předpokládáme, že máme k dispozici náhodný výběr $\{x_1, x_2, ..., x_n\}$ a rozdělení definičního oboru do M tříd a příslušné rozdělení pravděpodobnosti závisí jen na jednom parametru r. Potom pravděpodobnosti patření do jednotlivých tříd budou funkcí tohoto parametru. Daná konkrétní M-tice četností výskytu $n_1, n_2, ..., n_M$ se realizovala s pravděpodobností $p_1^{n_1} p_2^{n_2} \cdots p_M^{n_M}$. Hledáme odhad $\bar{r}$ neznámého parametru r pro který bude platit $p_1^{n_1} p_2^{n_2} ... p_M^{n_M} \underset{r}{\rightarrow} \max$. Toto maximum lze nalézt i maximalizací $\lg\left(p_1^{n_1} p_2^{n_2} ... p_M^{n_M}\right) \underset{r}{\rightarrow} \max$. $\lg\left(p_1^{n_1} p_2^{n_2} ... p_M^{n_M}\right) = \sum_{j=1}^{M} n_j \lg\left(p_j(r)\right)$. Pokud existuje řešení tohoto extremalizačního problému, bude nalezeno řešením normální rovnice

$$\frac{d}{dr} \sum_{j=1}^{M} n_j \lg\left(p_j(r)\right) = 0 \Leftrightarrow \sum_{j=1}^{M} \frac{n_j}{p_j} \frac{dp_j}{dr} = 0 . \qquad \textbf{(1)}$$

Zde je dobré připomenout:

1. Pro takový typ úlohy se věrohodnostní funkce konstruuje pro pozorování $n_1, n_2, ..., n_M$ nikoliv pro pozorování $x_1, x_2, ..., x_n$.
2. Tím se maximálně věrohodné odhady pro oba typy pozorování mohou lišit.
3. Jako řešení rovnice $\sum_{j=1}^{M} \frac{n_j}{p_j} \frac{dp_j}{dr} = 0$ nepotřebujeme obvykle hodnotu odhadu $\bar{r}$ ale hodnoty pravděpodobností $p_j, j = 1, ..., M$, přesněji jejich odhadů $p_j\left(\bar{r}\right), j = 1, ..., M$. Tyto pravděpodobnosti budou dále značeny $\overline{p_j} = p_j\left(\bar{r}\right), j = 1, ..., M$.
4. Řešení rovnice (1) může být velmi komplikované a tím i prakticky nepoužitelné. Proto se někdy používá slabšího testu shody:
    a. Určí se standardní metodikou maximálně věrohodné odhady neznámého parametru zvoleného typu rozdělení.
    b. Následně se provede klasický test shody.
        i. Vyloučení shody je pak významné a signifikantní.
        ii. Přijmutí shody je relativně slabším tvrzením.
        iii. Při přijmutí shody je zapotřebí pro definitivní potvrzení další dodatečné informace.

Námět: Formulujte výše uvedenou metodiku pro více-parametrickou třídu rozdělení.

Námět: Podrobně formulujte „Modifikovaný χ^2 test dobré shody“ pro exponenciální jedno-parametrické neposunuté rozdělení. Vyznačte vzniklé problémy.

χ^2 test homogenity výběrů

Mějme K náhodných výběrů a každému z nich četnosti pozorování příslušnosti k jedné z M tříd, stejně definovaných pro každý z $1,2,\ldots,K$ výběrů. Testujeme hypotézu (homogenity), že se všechny výběry řídí stejným pravděpodobnostním rozdělením.
Označíme:

$n_{i,j}$ počet pozorování, patřících v i-tém výběru j-té třídě,

$n_{*,j} = \sum_{i=1}^{K} n_{i,j}$ počet všech pozorování, bez ohledu na výběr, příslušných j-té třídě,

$n_{i,*} = \sum_{j=1}^{M} n_{i,j}$ počet všech pozorování v i-tém výběru,

$n = \sum_{i=1}^{K}\sum_{j=1}^{M} n_{i,j}$ počet všech pozorování.

Pokud jsou známy pravděpodobnosti p_j příslušnosti j-té třídě, bude mít náhodná proměnná $\chi_i^2 = \sum_{j=1}^{M} \frac{\left(n_{i,j} - n_{i,*}p_j\right)^2}{n_{i,*}p_j}$; $i = 1,\ldots,K$ za platnosti hypotézy homogenity přibližně $\chi^2(M-1)$ rozdělení. Podobně náhodná proměnná $\chi^2 = \sum_{i=1}^{K}\chi_i^2 = \sum_{i=1}^{K}\sum_{j=1}^{M} \frac{\left(n_{i,j} - n_{i,*}p_j\right)^2}{n_{i,*}p_j}$ má opět za platnosti hypotézy homogenity přibližně $\chi^2(K(M-1))$ rozdělení. V tomto případě je kritickým oborem $W_\alpha = \{\chi^2 : \chi^2 \geq \chi^2_{1-\alpha}(KM-K)\}$.

Pokud nejsou pravděpodobnosti p_j známy, může být jejich hodnota nahrazena odhadem $\hat{p}_j = \frac{n_{*,j}}{n}$. Potom náhodná proměnná $\chi^2 = \sum_{i=1}^{K}\chi_i^2 = \sum_{i=1}^{K}\sum_{j=1}^{M} \frac{\left(n_{i,j} - n_{i,*}\hat{p}_j\right)^2}{n_{i,*}\hat{p}_j} = n\left(\sum_{i=1}^{K}\sum_{j=1}^{M} \frac{{n_{i,j}}^2}{n_{i,*}n_{*,j}} - 1\right)$ má přibližně $\chi^2((K-1)(M-1))$ rozdělení a kritickým oborem bude $W_\alpha = \{\chi^2 : \chi^2 \geq \chi^2_{1-\alpha}((K-1)(M-1))\}$

Námět: Odvoďte výše uvedené stupně volnosti u χ^2 statistik. Odvození lze nalézt v: Jaroslav Hátle, Jiří Likeš. Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974. Str. 340-348.

Námět: Odvoďte podrobně vztah: $\chi^2 = \sum_{i=1}^{K}\chi_i^2 = \sum_{i=1}^{K}\sum_{j=1}^{M} \frac{\left(n_{i,j} - n_{i,*}\hat{p}_j\right)^2}{n_{i,*}\hat{p}_j} = n\left(\sum_{i=1}^{K}\sum_{j=1}^{M} \frac{{n_{i,j}}^2}{n_{i,*}n_{*,j}} - 1\right)$

Poznámka: Výše uvedený test homogenity je jedním ze základů pro studium vlastností a pojmů spojených s kontingenčními tabulkami a teorií výběrů z konečných souborů. Detaily lze nalézt v: Zuzana Prášková: Kontingenční tabulky, skripta MFFUK, Univerzita Karlova 1985 a Dana Vorlíčková: Výběry z konečných souborů, skripta MFFUK, Univerzita Karlova 1985.

Poznámka: (Důležitá) metodika χ^2 testů využívající četností příslušnosti k dané skupině nevyžaduje předpoklad náhodné (číselné) proměnné. Proto se takové testy užívají pro kategoriální (jevové – nečíselné) náhodné proměnné.

Námět (pro absolventy předmětu ZTI)**:** χ^2 kvadrát skóre lze chápat také jako divergenci dvou pravděpodobnostních modelů

$nD_{\chi^2}(p\|q)=\sum_{j=1}^{M}\frac{(np_j-nq_j)^2}{np_j}=n\sum_{j=1}^{M}\frac{(p_j-q_j)^2}{p_j} \quad \Rightarrow \quad D_{\chi^2}(p\|q)=\sum_{j=1}^{M}\frac{(p_j-q_j)^2}{p_j}$. Odvoďte odpovídající entropie a sdílené informace.

Testy shody založené na empirických distribučních funkcích.

Mějme náhodný výběr $x_1,x_2,...,x_n$ a k němu jeho „uspořádanou verzi“ $x_{(1)},x_{(2)},...,x_{(n)}$ z náhodné proměnné se spojitou distribuční funkcí $F(x)$. Potom pod empirickou distribuční funkcí budeme rozumět:

$$F_n(x)=\begin{cases}0 \Leftrightarrow x\le x_{(1)}\\ \frac{k}{n} \Leftrightarrow x_{(k)}<x\le x_{(k+1)};k=1,...,n-1\\ 1 \Leftrightarrow x>x_{(n)}\end{cases}$$

Jiná definice empirické distribuční funkce $F_n(x)=\frac{\text{pocet pozorovani mensich nez } x}{n}$

Námět: porovnejte obě definice.

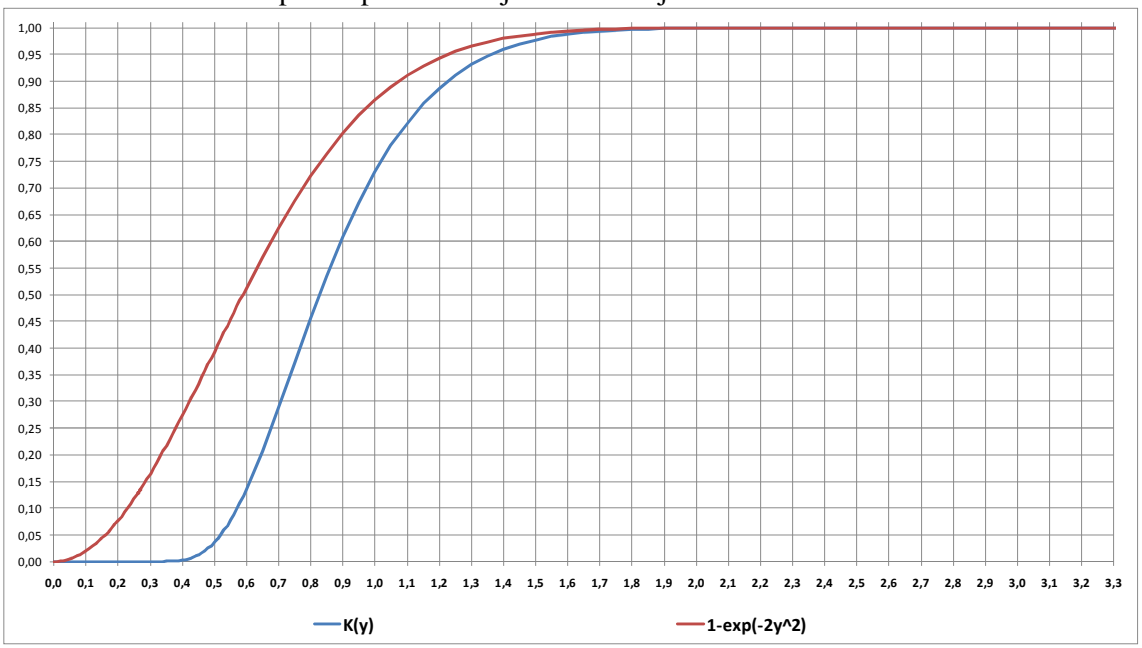
Pro empirické distribuční funkce platí:

1. $\lim_{n\to+\infty}P\left(\sqrt{n}\sup_{-\infty<x<+\infty}\left(F_n(x)-F(x)\right)<y\right)=\begin{cases}1-e^{-2y^2} \Leftrightarrow y>0\\ 0 \Leftrightarrow y\le 0\end{cases}$ Smirnov

2. $\lim_{n\to+\infty}P\left(\sqrt{n}\sup_{-\infty<x<+\infty}\left|F_n(x)-F(x)\right|<y\right)=\begin{cases}K(y) \Leftrightarrow y>0\\ 0 \Leftrightarrow y\le 0\end{cases}$ Kolmogorov,

kde $K(y)=\sum_{k=-\infty}^{k=+\infty}(-1)^k e^{-2k^2y^2}$. Tato řada má v definičním tvaru ***velice špatné numerické vlastnosti***. Pro její „numerickou realizaci“ viz: Miroslav Olehla, Vladimír Věchet, Josef Olehla: Řešení úloh matematické statistiky ve fortranu. Nakladatelství dopravy a spojů, Praha 1982. Str. 30-31.

Průběh obou limitních pravděpodobností je na následujícím obrázku:



Námět: Pokuste se interpretovat průběh Kolmogorovovy funkce. Mají smysl její hodnoty za jednotkou? Obdobně i pro Smirnovovu funkci $1-e^{-2y^2}$?

Zdroj: Alfréd Rényi: Teorie pravděpodobnosti, ACADEMIA, Praha 1972, str. 422-429

Tyto věty lze použít pro test shody rozdělení dat s některým modelovým rozdělením pro „dostatečně velká $\boldsymbol{n\equiv}$n>30 (empirie)“.

Pro testy shody dvou stejně rozsáhlých výběrů (tj. test shody dvou distribučních funkcí) je možné využít následujících tvrzení (Smirnov):

Mějme dva náhodné výběry $x_1, x_2, ..., x_n$ a $y_1, y_2, ..., y_n$ stejného rozsahu na stejném nosiči pak za platnosti **hypotézy shody** $\forall x \in D;\ F_x(x) = G_y(x)$ platí:

1. $$\lim_{n\to+\infty} P\left(\sqrt{\frac{n}{2}} \sup_{-\infty<x<+\infty} \left(F_n(x) - G_n(x)\right) < y\right) = \begin{matrix} 1-e^{-2y^2} \Leftrightarrow y>0 \\ 0 \Leftrightarrow y \le 0 \end{matrix}$$
2. $$\lim_{n\to+\infty} P\left(\sqrt{\frac{n}{2}} \sup_{-\infty<x<+\infty} \left|F_n(x) - G_n(x)\right| < y\right) = \begin{matrix} K(y) \Leftrightarrow y>0 \\ 0 \Leftrightarrow y \le 0 \end{matrix}$$

Opět tyto vztahy lze použít pro „dostatečně velká $\boldsymbol{n\equiv}$n>30“.

Pro konkrétní ***n*** lze použít následujících vztahů (Koruljuk, Gnědenko) opět za platnosti hypotézy shody: (Zdroj: Alfréd Rényi: Teorie pravděpodobnosti, ACADEMIA, Praha 1972, str. 426). Použijeme označení: $c = h(y\sqrt{2n})$ je horní celá část reálného čísla = nejbližší vyšší celé číslo k číslu $y\sqrt{2n}$.

$$0 \Leftrightarrow y \le 0$$

$$P\left(\sqrt{\frac{n}{2}} \sup_{-\infty<x<+\infty} \left(F_n(x) - G_n(x)\right) < y\right) = 1 - \frac{\binom{2n}{n-c}}{\binom{2n}{n}} \Leftrightarrow 0 < y \le \sqrt{\frac{n}{2}}$$

$$1 \Leftrightarrow y > \sqrt{\frac{n}{2}}$$

$$0 \Leftrightarrow y \le \frac{1}{\sqrt{2n}}$$

$$P\left(\sqrt{\frac{n}{2}} \sup_{-\infty<x<+\infty} \left|F_n(x) - G_n(x)\right| < y\right) = \frac{1}{\binom{2n}{n}} \sum_{i=-\lceil n/c \rceil}^{i=+\lceil n/c \rceil} (-1)^k \binom{2n}{n-kc} \Leftrightarrow \frac{1}{\sqrt{2n}} < y \le \sqrt{\frac{n}{2}}$$

$$1 \Leftrightarrow y > \sqrt{\frac{n}{2}}$$

Doporučená a zdrojová literatura:

Jiří Reif	Metody matematické statistiky, ZČU v Plzni 2004
Jaroslav Hátle, Jiří Likeš	Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974.
C. Radhakrishna Rao	Lineární metody statistické indukce a jejich aplikace, ACADEMIA, Praha 1978
Alfréd Rényi	Teorie pravděpodobnosti, ACADEMIA, Praha 1972
Mendenhall. W.	Introduction to Probability and Statistics. PWS-KENT, Publishing Company, Boston 1987.
Groebner, D.F., Shannon P. W.	Business Statistics. A Decision-Making Approach. Merrill Publishing Company, Columbus Ohio, 1989.
Jana Jurečková	Pořadové testy. Skripta MFFUK, SPN Praha 1981
Dagmar Blatná	Neparametrické metody. Testy založené na pořádkových a pořadových statistikách. Skripta VŠE, Praha 1996
Dagmar Blatná	Neparametrické metody II. Neparametrické odhady. Skripta VŠE, Praha 1999
Dana Vorlíčková	Výběry z konečných souborů, skripta MFFUK, Univerzita Karlova 1985.
Zuzana Prášková	Kontingenční tabulky, skripta MFFUK, Univerzita Karlova 1985.
Miroslav Olehla, Vladimír Věchet, Josef Olehla	Řešení úloh matematické statistiky ve fortranu. Nakladatelství dopravy a spojů, Praha 1982.