12. prednáška

Neparametrické testy

Z celej množiny neparametrických testov sa budeme venovať dvom. Najskôr sa budeme venovať postupom, ktoré umožnia preveriť predpoklady o type rozdelenia skúmanej náhodnej premennej. Predpoklad o tom, že študované dáta pochádzajú z určitého teoretického (očakávaného) rozdelenia býva podložený buď informáciami o sledovanom jave, alebo odhadom teoretického rozdelenia na základě grafického zobrazenia výberového rozdelenia. Náš odhad však nemusí být správny, a preto ho v praxi overujeme tzv. **testom dobrej zhody** (tj. zhody medzi teoretickým a empirickým (pozorovaným, výberovým) rozdelením.) Nulovú a alternatívnu hypotézu môžeme v tomto prípade formulovat:

H_0 : Základný súbor má [názov] rozdelenie pravdepodobnosti
H_a : Základný súbor nemá [názov] rozdelenie pravdepodobnosti

Najpoužívanejší je **Perasonov χ^2–test dobrej zhody** (*Goodness of Fit test*). Dá sa použiť pre spojité i diskrétne rozdelenia. Myšlienka testu je nasledujúca: Náhodný výber sa rozdelí do r tried a posúdi sa miera zhody napozorovaných výsledkov s výsledkami, ktoré očakávame, ak by H_0 platila. Porovnávame tak empirické početnosti n_i s teoretickými početnosťami np_i. Testovacie kritérium je štatistika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

kde

- n je rozsah náhodného výberu,
- n_i je empirická početnosť i–tej triedy,
- r je počet tried,
- p_i je pravdepodobnosť $P(x_i < X_i \leq x_{i+1})$, resp. $P(X = x_i)$,
- np_i je teoretická (očakávaná) početnosť i–tej triedy.

Pre $n \to \infty$ má testovacie kritérium približne χ^2– rozdelenie pravdepodobnosti s $(r - k - 1)$ stupňami voľnosti $(\chi^2 \approx \chi^2(r - k - 1))$, pričom k je počet odhadovaných patrametrov, ktoré sme potrebovali pre výpočet pravdepodobnosti p_i.
Z tvaru testovacej štatistiky je zrejmé, že jej veľké hodnoty signalizujú nezhodu medzi nameranými a teoretickými početnosťami. Preto hypotézu H_0 na zvolenej hladine významnosti α zamietame, ak je hodnota testovacieho kritéria χ^2_{obs} väčšia ako $100(1 - \alpha)\%$–ný kvantil χ^2– rozdelenie pravdepodobnosti s $(r - k - 1)$ stupňami voľnosti, t.j.

$$W_\alpha =< \chi_{1-\alpha}; \infty), \text{ kde } F_{\chi^2(r-k-1)}(\chi_{1-\alpha}) = 1 - \alpha.$$

V praxi sa test používa pre $n \geq 50$ a triedy treba voliť tak, aby platilo $np_i \geq 5$. Ak to táto podmienka nie je splnená, vhodne zlúčime susedné triedy tak, aby platilo $np_i \geq 5$. Ďalej pri výpočte hodnoty testovacieho kritéria požadujeme, aby $\sum_{i=1}^{r} p_i = 1$. Počet tried sa odporúča viac ako 5 ale menej ako 30 ☺.

χ^2–test nezávislosti

Tento test je jedným z celého radu testov, ktoré využívajú, resp. sú založené na kontigenčných tabuľkách. Skôr, ako sa zmienime o teste, zaveďme pojem kontigenčnej tabuľky.

Uvažujme náhodný vektor (X, Y), ktorý má diskrétne rozdelenie. Náhodná premenná X nadobúda hodnoty $1, 2, \ldots, r$ a náhodná premenná Y nadobúda hodnoty $1, 2, \ldots, s$. Náhodné premenné X a Y predstavujú znak nejakého štatistického súboru (napr. pohlavie, dosiahnuté vzdelanie,.....)

Predopkladajme, že sa uskutočnil výber o rozsahu n. Počet prípadov, keď sa vo výbere vyskytla dvojica (i, j), t.j. u prvkov výberu sa zistil i–ty stupeň znaku X a j–ty stupeň znaku Y, označíme n_{ij}.

Kontigenčnú tabuľku potom definujeme ako maticu $(n_{ij})_{r\times s}$. Označme:

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij} \quad \text{a} \quad n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$$

$$\text{Zrejme} \quad n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} \quad \text{a platí} \quad n = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{s} n_{\bullet j}$$

Kontigenčná tabuľku vytvoríme v tvare:

$X\backslash Y$	1	2	...	s	$n_{i\bullet}$
1	n_{11}	n_{12}	...	n_{1s}	$n_{1\bullet}$
2	n_{21}	n_{22}	...	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet s}$	n

Ak máme dáta usporiadané do kontigenčnej tabuľky, kde kategórie jedného znaku určujú riadky a kategórie druhého znaku stĺpce, môžeme testovať hypotézu o nezávislosti náhodných premenných X a Y, t.j. o nezávislosti znakov X, Y.
Na každom prvku jedného súboru sledujeme dva znaky. Naším cieľom je otestovať hypotézu o nezávislosti sledovaných znakov, t.j.

H_0 : náhodné premenné X (1. znak) a Y (2. znak) sú nezávislé
H_0 : náhodné premenné X (1. znak) a Y (2. znak) sú závislé

Pre posúdenie, či empirické početnosti n_{ij} nie sú v rozpore s hypotézou H_0 o nezávislosti oboch znakov je potrebné skonštruovať tzv. **teoretické (očakávané) početnosti** e_{ij}. Pri konštrukcii teoretických početností sa vychádza z poučky o pravdepodobnosti prieniku nezávislých javov:

$$P(X = i \cap Y = j) = P(X = i)P(Y = j)$$

Pravdepodobnosti jednotlivých kategórií znaku sú odhadnuté relatívnymi početnosťami:

$$\text{i–ta kategória znaku X :} \qquad \widehat{P(X = i)} = \frac{n_{i\bullet}}{n}$$

$$\text{j–ta kategória znaku Y :} \qquad \widehat{P(Y = j)} = \frac{n_{\bullet j}}{n}$$

Pravdepodobnosť výskytu i–tej úrovne znaku X a j–tej úrovne znaku Y za predpokladu platnosti hypotézy H_0 je potom $\frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n}$ a očakávaná teoretická početnosť je $e_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$. Testovacie kritérium je založené na rozdieli empirickej a teoretickej početnosti a má tvar:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Za predpokladu platnosti nulovej hypotézy má testovacie kritérium asymptoticky χ^2–rozdelenie s $(r-1)\cdot(s-1)$ stupňami voľnosti, t.j. $\chi^2 \approx \chi^2_{(r-1)\cdot(s-1)}$. Je zrejmé, že vysoké hodnoty testovacieho kritéria znamenajú veľké rozdiely medzi skutočnými (empirickými) početnosťami a očakávanými (teoretickými) početnosťami, a to svedčí v prospech alternatívnej hypotézy H_a, teda pre závislosť medzi premennými.
Hypotézu H_0 zamietame, ak $\chi^2 \geq \chi_{(1-\alpha)}$, kde $F_{\chi^2_{(r-1)\cdot(s-1)}}(\chi_{1-\alpha}) = 1-\alpha$.
Pre zhodu s limitným rozdelením sa vyžaduje, aby všetky teoretické početnosti $e_{ij} = \dfrac{n_{i\bullet}n_{\bullet j}}{n} > 5$ pre $i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s$. Obvykle sa vyžaduje, aby všetky $e_{ij} > 1$ a aspoň 80% $e_{ij} > 5$. Ak tomu tak nie je, spájajú sa niektoré riadky alebo stĺpce.

PRÍIKLAD:
Budeme testovať nezávislosť medzi výsledkami z matematiky a študijným programom, na ktorý sa študenti hlásia. Študenti sa môžu hlásiť na bakalársky študijný program Finančná matematika, na 5–ročné štúdium učiteľstva matematiky pre ZŠ a 5–ročné štúdium učiteľstva pre SŠ.
Uchádzač môže získať z testu maximálne 80 bodov. Premenná X (výsledok testu z matematiky) nadobúda 4 hodnoty:

- 1 . . . počet získaných bodov 60 – 80
- 2 . . . počet získaných bodov 40 – 59
- 3 . . . počet získaných bodov 20 – 39
- 4 . . . počet získaných bodov 0 – 19

Študijné programy sú zoradené od najľahších po najťažšie. Premenná Y (zvolený študijný program) nadobúda 3 hodnoty:

- 1 . . . finančná matematika
- 2 . . . učiteľstvo pre ZŠ
- 3 . . . učiteľstvo pre SŠ

Kontigenčná tabuľka má tvar:

$X\backslash Y$	FM	$Z\check{S}$	$S\check{S}$	$n_{i\bullet}$
1	9	7	40	
2	10	31	58	
3	17	29	29	
4	14	25	19	
$n_{\bullet j}$				