Proseminář z matematiky

Statistika 3

Test dobré shody

Z výsledků zkoušek dvou zkoušejících téhož předmětu rozhodněte, zda je distribuce známek u obou učitelů stejná.

	1	2	3	4	$\sum$
A	46	29	26	14	115
B	29	24	22	25	100

Data rozdělená do k kategorií a dvou souborů s četnostmi $A_1, \ldots, A_k$ a $B_1, \ldots, B_k$.
$N_A = \sum_{i=1}^{k} A_i$, $N_B = \sum_{i=1}^{k} B_i$ a $N = N_A + N_B$
Pro $i = 1, \ldots, k$ očekávané četnosti $EA_i = (A_i + B_i)\frac{N_A}{N}$ a $EB_i = (A_i + B_i)\frac{N_B}{N}$
$\chi^2 = \sum_{i=1}^{k} \frac{(A_i - EA_i)^2}{EA_i} + \sum_{i=1}^{k} \frac{(B_i - EB_i)^2}{EB_i}$
χ^2 rozdělení s $k - 1$ stupni volnosti

Test dobré shody

	A	B	$A+B$	EA	EB	$\chi^2 A$	$\chi^2 B$
1	46	29	75	40.12	34.88	0.8629	0.9924
2	29	24	53	28.35	24.65	0.0150	0.0172
3	26	22	48	25.67	22.33	0.0041	0.0047
4	14	25	39	20.86	18.14	2.2562	2.5947
$\sum$	115	100	215			3.1383	3.6090

Výsledná hodnota $\chi^2 = 6.7473$ při $k - 1 = 3$ stupních volnosti dává hodnotu $p = 0.0804$

Analýza rozptylu (ANOVA)

několika skupin podle (jedné nebo více) popisných proměnných a zkoumaná číselná náhodná proměnná na nich může záviset
Nulová hypotéza je, že rozdělení do skupin nemá vliv na hodnoty náhodné proměnné, tedy ve všech skupinách má zkoumaná náh. proměnná stejnou střední hodnotu
Předpoklady použití ANOVA testu:

1. náhodný výběr měření, pokud možno rovnoměrně přes všechny skupiny
2. v každé skupině je zkoumaná náhodná proměnná normálně rozdělená
3. a to se steným rozptylem

ANOVA – značení a výpočty

k počet skupin

n_i velikost jednotlivých skupin, tj. počet vzorků v i-té skupině, $i = 1, \ldots, k$

n celkový počet vzorků, $n = \sum_{i=1}^{k} n_i$

x_{ij} j-tý vzorek z i-té skupiny

$\overline{x}$ výběrová střední hodnota (celková), $\overline{x} = \sum_{i,j} x_{ij}$

$\overline{x_i}$ výběrová střední hodnota ve skupině i, $\overline{x_i} = \frac{1}{n} \sum_{j=1}^{n_i} x_{ij}$

s_i^2 výběrový rozptyl ve skupině i, $\overline{s_i}^2 = \frac{1}{n-1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$

SST sum of squres total, $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$

SSG sum of sq. between groups, $SSG = \sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x})^2$

SSE sum of sq. inside groups, $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2 = \sum_{i=1}^{k} (n-1) s_i^2$

df (stupně volnosti) $df(SSG) = k - 1$, $df(SSE) = n - k$, $df(SST) = n - 1$

ANOVA – značení a výpočty

MS mean of squares, $MS = SS/df$, tedy

$$\begin{aligned} MSG &= SSG/(k-1) \\ MSE &= SSE/(n-k) = \frac{(n_1-1)s_1^2 + \ldots + (n_k-1)s_k^2}{(n_1-1) + \ldots + (n_k-1)} \end{aligned}$$

F-statistika $F = MSG/MSE$, má tabulizovaný průběh pro dané stupně volnosti $k-1$ a $n-k$

P-value jednostranný odhad pravděpodobnosti pro F-statistiku, tj. pravděpodobnost, že F je větší rovna dané hodnotě, pokud je P-value menší než α, zamítáme nulovou hypotézu

Tabulka s výsledky ANOVA testu vypadá následovně.

faktor	SS	df	MS	F	P-value
Group/model	SSG	k-1	MSG	MSG/MSE	p
Error/residual	SSE	n-k	MSE		
Total	SST	n-1			

Vícefaktorová ANOVA

Posouzení míry vlivu dvou (případně i více) popisných proměnných
Lze pro každý faktor zvlášť, ale může se stát, že vliv jednotlivých faktorů se jednoduše nesčítá, ale kombinuje složitějším způsobem
Provedeme ANOVA test se skupinami určeným kartézským součinem oborů hodnout obou faktorů (je třeba dát pozor na podobnou velikost skupin, pokud se faktory ovlivňují navzájem, nemusí být předpoklad splněn)
V tabulce (příklad níže) uvažujeme, že faktor A má k skupin a faktor B má ℓ skupin, SSE se počítá společně podle rozdělení do $k\ell$ skupin. Obvykle faktory řadíme podle jejich vlivu na střední hodnoty ve skupinách, tento vliv můžeme odhadnout nebo použít pro každý z faktorů jednofaktorový ANOVA test.

faktor	SS	df	MS	F	P-value
A	SSA	k-1	MSA	MSA/MSE	p_A
B	SSB	ℓ-1	MSB	MSB/MSE	p_B
A×B	SSAB	(k-1)(ℓ-1)	MSAB	MSAB/MSE	p_{AB}
Error	SSE	n-k	MSE		
Total	SST	n-1			

Vícefaktorová ANOVA – příklad

Ilustrační příklad: analýzu závislosti platu na pohlaví a dosaženám vzdělání.
Faktor pohlaví má dvě možnosti, pro vzdělání máme tři hodnoty (bez maturity, s maturitou, vysokoškolské).
Konkrétní data uvádět nebudeme, ale základní charakteristiky jsou:
v každé ze šesti skupin je 5 hodnot, střední hodnota platu je 19.7, mezi ženami 18.9 a mezi muži 20.5, při rozdělení podle vzdělání jsou střední hodnoty 14.2 (bez maturity), 17.1 (s maturitou) a 27.8 (VŠ).
Test pro faktor pohlaví:

faktor	SS	df	MS	F	P-value
pohl.	17.6	1	17.6	0.437	0.515
Error	1130.7	28	40.4		

Test pro faktor vzdělání:

faktor	SS	df	MS	F	P-value
vzd.	1026	2	513	113	< 0.001
Error	122	27	4.52		

Vícefaktorová ANOVA – příklad

Dvoufaktorový test:

faktor	SS	df	MS	F	P-value
vzd.	1026.20	2	513.10	121.68	< 0.001
pohl.	17.63	1	17.63	4.78	0.052
oba	3.27	2	1.63	0.39	0.683
Error	101.20	24	4.22		

Interpretace pro volbu hladiny významnosti $\alpha = 0.05$:
Postupujeme odspodu a hledáme nejjednodušší, ale data vysvětlující model.
Protože P-value na řádku s kombinací obou faktorů je větší než α, lze přijmout nulovou hypotézu, respektive zamítnout alternativní hypotézu, že kombinace obou faktorů je nutná k popisu modelu.
Pro faktor pohlaví je P-value opět větší než α, můžeme též zamítnout vliv pohlaví na střední hodnotu platu.
Zbývá pouze faktor vzdělání s P-value menší než α, kde zamítneme nulovou hypotézu (tedy vzdělání má podstatný vliv na výši platu).

Chyba 1. a 2. druhu

Máme dva pytlíky s kuličkam
A: 80 bílých a 20 černých
B: 30 bílých a 70 černých
Máme jeden z nich, vylosujeme z něho 10 kuliček (s vracením) a pokud jich je alespň k bílých, přijmeme hypotézu, že je to pytlík A (pokud jich je méně než k, tuto hypotézu odmítneme).
V závislosti na parametru k určete pravděpodobnost chyby prvního a druhého druhu a najděte k pro které je test vyvážený, tj. pravděpodobnost chyby prvního a druhého druhu je (přibližně) stejná.

Chyba 1. a 2. druhu

Y náhodná veličina rovná počtu bílých kuliček, které v $n = 10$ tazích vylosujeme
Y má binomické rozdělení s parametry n, p, kde $p_A = 0.8$ a $p_B = 0.3$

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \ , \ y = 0, 1, \ldots, n$$

$$\begin{aligned} \alpha_k &= P(Y < k | H_0) = \sum_{i=0}^{k-1} \binom{n}{i} p_A^i (1-p_A)^{n-i} \\ \beta_k &= P(Y \geq k | H_1) = \sum_{i=k}^{n} \binom{n}{i} p_B^i (1-p_B)^{n-i} \end{aligned}$$

Chyba 1. a 2. druhu