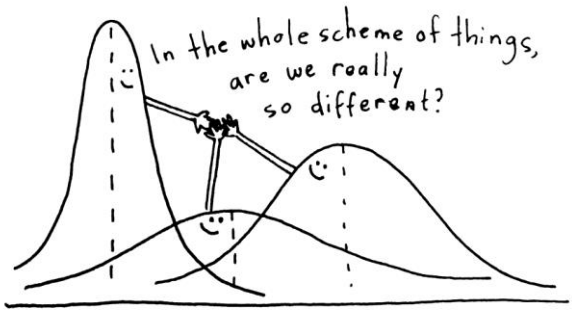
ANALYTICKÁ CHEMOMETRIE

Test normality a shodnosti rozptylů



Patrik Kania
Ústav analytické chemie, VŠCHT Praha

25. 11. 2025

první předpoklad pro ANOVu:

- srovnávané výběry pocházejí ze základních souborů s normálním rozdělením

Testování normality

- je nutné vždy, kdy vznikne pochybnost o normalitě základního souboru
- lze provést celou řadou testů (Shapiro-Wilkův test, testy založené na šikmosti a špičatosti, Kolmogorovův-Smirnovův test apod.)

Jeden z nejjednodušších testů je založen na šikmosti a špičatosti.

- má výhodu, že můžeme zjistit, která charakteristika tvaru (zda šikmost nebo špičatost) způsobuje odchylku od normality

U testů založených na šikmosti a špičatosti vycházíme z předpokladu, že výběr je z normálního rozložení. Potom výběrová šikmost *A* a špičatost *E* mají asymptoticky normální rozdělení s parametry:

$$\mathrm{E}(A) = 0 \qquad \mathrm{E}(E) = -\frac{6}{n+1}$$

$$\mathrm{D}(A) = \frac{6(n-2)}{(n+1)(n+3)} \qquad \mathrm{D}(E) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

H_0: výběr pochází ze základního souboru s normálním rozdělením
H_1: výběr pochází z nějakého nesymetrického či špičatého rozdělení

$$\mathrm{TK}(A) = \frac{A_{\text{výběr}} - \mathrm{E}(A)}{\sqrt{\mathrm{D}(A)}} \qquad \mathrm{TK}(E) = \frac{E_{\text{výběr}} - \mathrm{E}(E)}{\sqrt{\mathrm{D}(E)}}$$

pokud je alespoň jedno ze dvou $|TK| > z_{\alpha/2}$ hypotézu H_0 zamítáme
test je používán pro velký výběr (> 50 prvků)
demonstrace v Excelu „příklad na test normality.xlsx“

druhý předpoklad pro ANOVu:

- výběry pocházejí ze základních souborů se shodnými rozptyly

Testování shodnosti rozptylů

- pro test shody rozptylů (homoskedasticita) $k > 2$ náhodných výběrů z normálního rozdělení lze použít opět celou řadu testů (Bartlettův test, Hartleyův test, Cochranův test apod.)

Hartleyův test

- nejjednodušší test
- pokud nezjistíme rozdíl mezi největším a nejmenším výběrovým rozptylem, nebudou se statisticky významně lišit ani ostatní dvojice

Herman Otto Hirschfeld (1912–1980)

- německý a americký statistik
- Hirsch = Hart, Feld = field = lea = ley

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ H_1: existuje alespoň jeden rozdíl rozptylů

$$TK = F_{\max} = \frac{\max s_i^2}{\min s_i^2}$$

pokud je $TK > h_\alpha(k, n-1)$
hypotézu o rovnosti rozptylů zamítáme

kritické hodnoty $h_\alpha(k, n-1)$

$\alpha = 0{,}05$	k										
stupně volnosti ν	2	3	4	5	6	7	8	9	10	11	12
2	39	87,5	142	202	266	333	403	475	550	626	704
3	15,4	27,8	39,2	50,7	62	72,9	83,5	93,9	104	114	124
4	9,6	15,5	20,6	25,2	29,5	33,6	37,5	41,1	44,6	48	51,4
5	7,15	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9
6	5,82	8,38	10,4	12,1	13,7	15	16,3	17,5	18,6	19,7	20,7
7	4,99	6,94	8,44	9,7	10,8	11,8	12,7	13,5	14,3	15,1	15,8
8	4,43	6,00	7,18	8,12	9,03	9,78	10,5	11,1	11,7	12,2	12,7
9	4,03	5,34	6,31	7,11	7,8	8,41	8,95	9,45	9,91	10,3	10,7
10	3,72	4,85	5,67	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34
12	3,28	4,16	4,79	5,3	5,72	6,09	6,42	6,72	7,00	7,25	7,48
15	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,4	5,59	5,77	5,93
20	2,46	2,95	3,29	3,54	3,76	3,94	4,1	4,24	4,37	4,49	4,59
30	2,07	2,4	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39
60	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,3	2,33	2,36
∞	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

Příklad: Průběh reakce může být ovlivňován třemi různými biochemickými katalyzátory A, B a C. Zjistěte, zda výtěžek reakce (závisle proměnná) je významně ovlivňován některým z katalyzátorů. Před provedením ANOVA analýzy otestujte, zda je splněn požadavek rovnosti rozptylů.

úrovně faktoru									s_i^2	n_i
A	85	86	83	82	87	90	80	81	11,36	8
B	87	86	85	93	89	88	86	89	6,41	8
C	89	87	90	88	83	88	89	91	5,84	8

$$TK = F_{\max} = \frac{11{,}36}{5{,}84} = 1{,}94 \qquad h_{0{,}05}(3, 8-1) = 6{,}94$$

jelikož TK < kritická hodnota, hypotézu o rovnosti rozptylů nelze zamítnout

Bartlettův test

- nejsilnější možný test
- lze použít v případě, kdy výběry mají různý rozsah

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ H_1: existuje alespoň jeden rozdíl rozptylů

$$TK = B = \frac{\ln 10}{C}\left[(N-k)\log s^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2\right] \qquad N = \sum_{i=1}^{k} n_i$$

$$s^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)s_i^2 \qquad C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)$$

společný rozptyl účel korekce: zmírnit nadhodnocení *B* u malých vzorků.

Statistika *B* má při platnosti H_0 přibližně χ^2 rozdělení, pokud všechna $n_i \geq 6$. Lze nejdříve vypočítat hodnotu *B* pro *C* = 1. Je-li $B < \chi^2_{\alpha,k-1}$ hypotézu přijímáme. Je-li $B > \chi^2_{\alpha,k-1}$, počítáme ji znovu pro vypočítanou hodnotu *C*.

Příklad: Průběh reakce může být ovlivňován třemi různými biochemickými katalyzátory A, B a C. Zjistěte, zda výtěžek reakce (závisle proměnná) je významně ovlivňován některým z katalyzátorů. Před provedením ANOVA analýzy otestujte, zda je splněn požadavek rovnosti rozptylů.

úrovně faktoru									s_i^2	n_i
A	85	86	83	82	87	90	80		11,24	7
B	87	86	85	93	89	88	86	89	6,41	8
C	89	87	90	88	83	88	89	91	5,84	8

$$TK = B = \frac{\ln 10}{1}\left[(23-3)\log 7{,}659 - \sum_{i=1}^{3}(n_i - 1)\log s_i^2\right] = 0{,}843$$

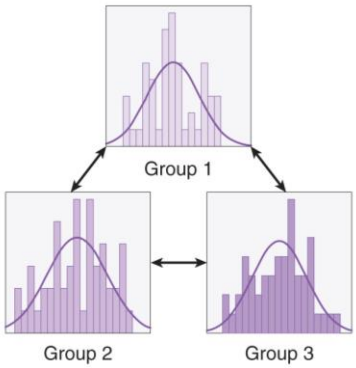
$$s^2 = \frac{1}{23-3}\sum_{i=1}^{3}(n_i - 1)s_i^2 = 7{,}659 \qquad \chi^2_{\alpha,k-1} = 5{,}99$$

jelikož TK < kritická hodnota, hypotézu o rovnosti rozptylů nelze zamítnout

Konec

ANALYTICKÁ CHEMOMETRIE

Neparametrická ANOVA



Patrik Kania
Ústav analytické chemie, VŠCHT Praha

25. 11. 2025

Neparametrická ANOVA (angl. *non-parametric ANOVA*)

použití v případě:

- kdy není zaručena normalita dat či homogenita rozptylů
- mají-li jednotlivé výběry velmi málo prvků a normalitu není možné spolehlivě stanovit

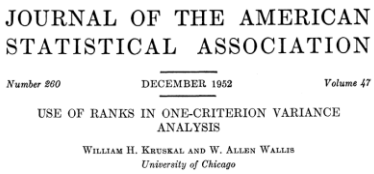
Neparametrická ANOVA má vždy nižší sílu testu (slabší schopnost zamítnout nulovou hypotézu) než parametrická ANOVA.

Při splnění předpokladů normality a homogenity má neparametrická ANOVA asi 95 % síly testu parametrické varianty.

Parametrická ANOVA sama je poměrně robustní a neparametrické metody používáme zpravidla při výrazném porušení předpokladů o normalitě dat či homogenitě rozptylů!

Kruskal-Wallisův test (K-W test) (angl. *Kruskal-Wallis test*)

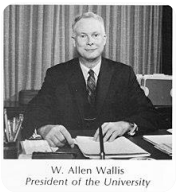
- neparametrická obdoba jednofaktorové ANOVy
- test založen na pořadí prvků


JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 260 DECEMBER 1952 Volume 47

USE OF RANKS IN ONE-CRITERION VARIANCE ANALYSIS

WILLIAM H. KRUSKAL AND W. ALLEN WALLIS
University of Chicago


W. Allen Wallis
President of the University

William Kruskal (1919–2005)

- americký statistik
- v roce 1950 zisk místa instruktora pro statistiku na Chicagské univerzitě pod Allenem Wallisem

W. Allen Wallis (1912–1998)

- americký ekonom a statistik
- poradce Eisenhowera, Nixona, Forda a Reagana

Kruskal-Wallisův test (K-W test) (angl. *Kruskal-Wallis test*)

- neparametrická obdoba jednofaktorové ANOVy
- test založen na pořadí prvků

H_0: mediány jednotlivých souborů jsou stejné

postup testu:

1) prvky všech *k* výběrů sloučíme do jednoho sdruženého výběru
2) prvky seřadíme od nejmenšího k největšímu
3) určíme pořadí prvků ve sdruženém výběru
4) provedeme součty R_i takto získaných pořadí ale v původních výběrech
5) vypočteme testovací kritérium *H*

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

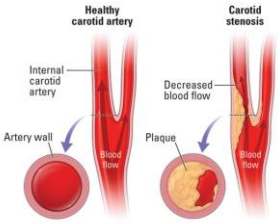
odvození vzorce pro Kruskal-Wallisův test

N - celkový počet prvků, n_i - počet prvků v *i-tém* výběru, R_i - součet pořadí v *i-tém* výběru

6) H_0 zamítáme, pokud je $H > \chi^2_{(\alpha,\, k-1)}$ použití krit. χ^2 hodnot jen pro $n_i > 5$

Příklad: Ateroskleróza je kornatění tepen, které vzniká v důsledku ukládání tukových látek do stěny tepny. Toto postupné ukládání způsobuje tvorbu překážky a následně tlakové rozdíly v krevním oběhu. Konečným stavem je prasknutí tepny. Při výzkumu pružnosti tepen byl studován vliv úrovně stenózy (%) na maximální průtok krve (ml/s), při kterém dojde ke kolapsu tepny. Pro každou úroveň stenózy byla provedena opakovaná měření a získaná data jsou uvedena v tabulce. Na hladině významnosti 5 % otestujte, zda úroveň stenózy má vliv na hodnoty maximálního průtoku.

úroveň stenózy		
0,78	0,71	0,65
10,6	11,7	19,6
9,7	12,7	15,1
10,2	12,6	15,6
14,3	16,6	19,2
10,6	16,3	18,7
11,0	14,7	13,0
13,1	19,5	18,9
10,8	16,8	18,0
14,3	15,1	18,6
10,4	15,2	16,6
8,3	13,2	
	14,8	
	14,4	
	17,6	



sdružení měření ze všech skupin

a určení pořadí hodnot

sdružený výběr		sdružený výběr	
hodnoty	pořadí	hodnoty	pořadí
8,3	1	14,8	19
9,7	2	15,1	20,5
10,2	3	15,1	20,5
10,4	4	15,2	22
10,6	5,5	15,6	23
10,6	5,5	16,3	24
10,8	7	16,6	25,5
11	8	16,6	25,5
11,7	9	16,8	27
12,6	10	17,6	28
12,7	11	18	29
13	12	18,6	30
13,1	13	18,7	31
13,2	14	18,9	32
14,3	15,5	19,2	33
14,3	15,5	19,5	34
14,4	17	19,6	35
14,7	18		

úroveň stenózy					
0,78		0,71		0,65	
hodnota	pořadí	hodnota	pořadí	hodnota	pořadí
10,6	5,5	11,7	9	19,6	35
9,7	2	12,7	11	15,1	20,5
10,2	3	12,6	10	15,6	23
14,3	15,5	16,6	25,5	19,2	33
10,6	5,5	16,3	24	18,7	31
11	8	14,7	18	13	12
13,1	13	19,5	34	18,9	32
10,8	7	16,8	27	18	29
14,3	15,5	15,1	20,5	18,6	30
10,4	4	15,2	22	16,6	25,5
8,3	1	13,2	14		
		14,8	19		
		14,4	17		
		17,6	28		
R_1	80	R_2	279	R_3	271
n_1	11	n_2	14	n_3	10

$$H = \frac{12}{35(35+1)}\left(\frac{80^2}{11} + \frac{279^2}{14} + \frac{271^2}{10}\right) - 3(35+1) = 20{,}438$$

Jelikož jsou mezi prvky i skupiny stejných hodnot a tedy stejných pořadí, je třeba provést korekci kritéria *H* – viz následující snímek.

sdružený výběr		sdružený výběr	
hodnoty	pořadí	hodnoty	pořadí
8,3	1	14,8	19
9,7	2	15,1	20,5
10,2	3	15,1	20,5
10,4	4	15,2	22
10,6	5,5	15,6	23
10,6	5,5	16,3	24
10,8	7	16,6	25,5
11	8	16,6	25,5
11,7	9	16,8	27
12,6	10	17,6	28
12,7	11	18	29
13	12	18,6	30
13,1	13	18,7	31
13,2	14	18,9	32
14,3	15,5	19,2	33
14,3	15,5	19,5	34
14,4	17	19,6	35
14,7	18		

$$H_c = \frac{H}{1 - \frac{\sum_{i=1}^{m}(t_i^3 - t_i)}{N^3 - N}}$$

t_i - počet stejných hodnot v *i-té* skupině stejných hodnot
m - počet skupin stejných hodnot

$$H_c = \frac{20{,}438}{1 - \frac{(2^3-2)+(2^3-2)+(2^3-2)+(2^3-2)}{35^3-35}} = 20{,}449$$

jelikož $H = 20{,}449 > \chi^2_{(0{,}05;3-1)} = 5{,}99$

H_0 neplatí

stupeň stenózy ovlivňuje max. průtok

Korekce *H* je významná pouze při velkém počtu rozsáhlých skupin stejných hodnot, jinak je její vliv zanedbatelný.

sdružený výběr		sdružený výběr	
hodnoty	pořadí	hodnoty	pořadí
8,3	1	14,8	19
9,7	2	15,1	20,5
10,2	3	15,1	20,5
10,4	4	15,2	22
10,6	6,5	15,6	23
10,6	6,5	16,3	24
10,6	6,5	16,6	26
10,6	6,5	16,6	26
11,7	9	16,6	26
12,6	10	17,6	28
12,7	11	18	29
13	12	18,6	30
13,1	13	18,7	31
13,2	14	18,9	32
14,3	15,5	19,2	33
14,3	15,5	19,5	34
14,4	17	19,6	35
14,7	18		

$$H_c = \frac{H}{1 - \frac{\sum_{i=1}^{m}(t_i^3 - t_i)}{N^3 - N}}$$

t_i - počet stejných hodnot v *i-té* skupině stejných hodnot
m - počet skupin stejných hodnot

$$H_c = \frac{20{,}438}{1 - \frac{(4^3 - 4) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3)}{35^3 - 35}} = 20{,}484$$

jelikož $H = 20{,}484 > \chi^2_{(0{,}05;3-1)} = 5{,}99$

H_0 neplatí

stupeň stenózy ovlivňuje max. průtok

MODIFIKACE PŘEDCHOZÍHO PŘÍKLADU

Dunnův test mnohonásobného porovnávání (angl. *Dunn's test*)

H_0: $\tilde{x}_A = \tilde{x}_B$ H_1: $\tilde{x}_A \neq \tilde{x}_B$

$$TK = Q = \frac{\bar{R}_B - \bar{R}_A}{SE}$$

$\bar{R}_i$ - průměrné pořadí v porovnávaných skupinách ($\bar{R}_i = R_i/n_i$)

$$SE = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum_{i=1}^{m}(t_i^3 - t_i)}{12(N-1)}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

pokud je $TK = Q > z_{\left(\frac{\alpha}{k(k-1)}\right)}$

$z_{\alpha/(k(k-1))}$ je kritická hodnota norm. normálního rozdělení pro hladinu významnosti $\alpha/(k(k-1))$
k je počet porovnávaných skupin

Konec