Analýza rozptylu

- V praxi často je potrebné porovnávať väčší počet nezávislých náhodných výberov z hľadiska úrovne, t. zn. zaujíma nás hypotéza:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \mu_i = \ldots \mu_m = \mu$$

$$H_1 : \mu_i \neq \mu$$ pre aspoň jeno ***i*** (i = 1, 2,…m)

pre ***m > 2,*** kde μ_i **,** ***i =1, 2, …m*** sú stredné hodnoty z normálne rozdelených základných súborov s rovnakým rozptylom σ^2 , t.j. $N(\mu, \sigma^2)$

- K overeniu tejto hypotézy sa používa dôležitá štatistická metóda, nazývaná **Analýza rozptylu**, skrátene **ANOVA (**resp**. AR)**

- Analýza rozptylu sa delí podľa sledovaných faktorov na jednofaktorovú, dvojfaktorovú, trojfaktorovú, atď.
- Aj keď je možné uvažovať neobmedzený počet faktorov a úrovní, v reálnych úlohách sa zvyčajne pracuje s 1 – 4 faktormi, ktoré sa uvažujú na 2 – 6 úrovniach, pričom počet opakovaní nebýva veľký.

Gymnázium	SPŠ	OU
55	54	47
54	50	53
58	51	49
61	51	50
52	49	46

- **Úrovňou faktora** budeme označovať:
  - určité množstvo kvantitatívneho faktora,

    napr. množstvo dávok čistých živín pri hnojení, rôzne príjmové skupiny domácností,
  - určitý druh kvalitatívneho faktora, napr. rôzne odrody tej istej plodiny, spôsoby umiestnenia výrobkov v predajni,
- AR je zovšeobecnením Studentovho t-testu pre nezávislé výbery
- AR zároveň skúma vplyv kvalitatívneho faktora (faktorov) na výsledný kvantitatívny znak - teda analyzuje vzťahy medzi znakmi

Hodnotenie	Typ školy	Lokalita
55	G	M
54	G	M
58	G	V
61	G	V
52	G	V
54	SPŠ	M
50	SPŠ	M
51	SPŠ	V
51	SPŠ	V
49	SPŠ	V
47	SOU	M
53	SOU	M
49	SOU	M
50	SOU	V
46	SOU	V

Jednofaktorová analýza rozptylu

- Namerané hodnoty xij sú zatriedené do r skupín (j = 1,...,r)
- Určenie α
- Stanovenie H_0: $\mu_1 = ... = \mu_r$, proti H_1: aspoň jedna dvojica stredných hodnôt sa líši
- Určenie TK: vypočítajú sa skupinové priemery , $\bar{x}_j$ j = 1, ...,r
- vypočíta sa celkový priemer $\bar{x} = \frac{1}{n}\sum_{j=1}^{r} n_j x_j$
- vypočítajú sa variability $S_T = S_E + S_A$

Hodnotenie	Typ školy
55	G
54	G
58	G
61	G
52	G
54	SPŠ
50	SPŠ
51	SPŠ
51	SPŠ
49	SPŠ
47	SOU
53	SOU
49	SOU
50	SOU
46	SOU

$$S_E = \sum_{j=1}^{r}(n_j - 1)S_j^{\,2} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2$$

$$S_A = \sum_{j=1}^{r} n_j(\bar{x}_j - \bar{x})^2$$

$$S_T = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(x_{ji} - \bar{x})^2$$

$$F = \frac{S_A/(r-1)}{S_E/(n-r)}$$

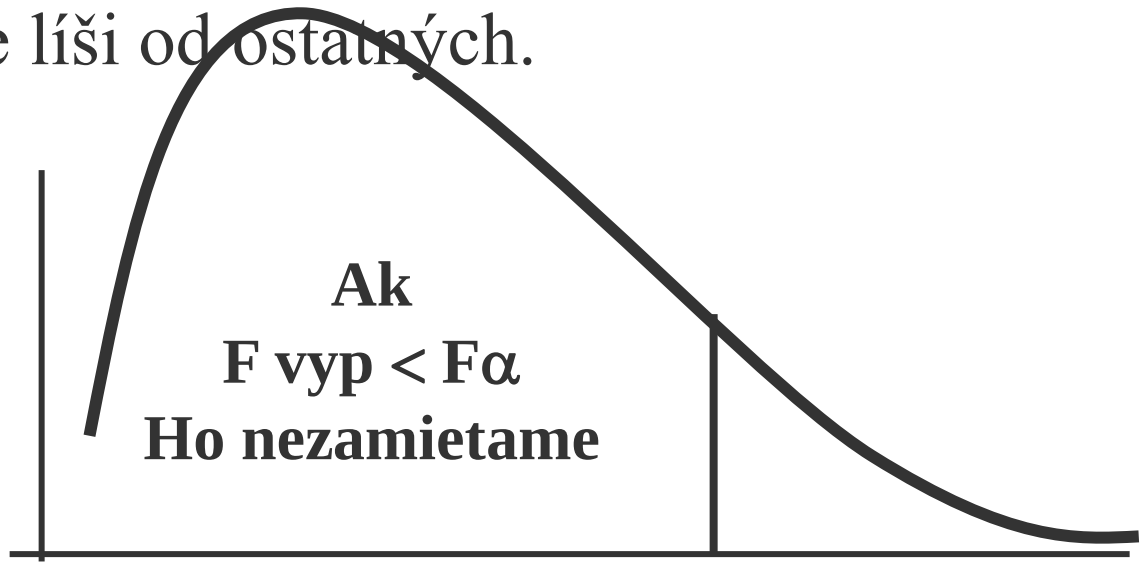
- S_E - reziduálny súčet štvorcov a charakterizuje variabilitu vo vnútri jednotlivých výberov
- S_A - skupinový súčet štvorcov a charakterizuje variabilitu medzi jednotlivými výbermi
- S_T - celkový súčet štvorcov a charakterizuje variabilitu jednotlivých pozorovaní okolo celkového priemeru

Zdroj variability	Súčet štvorcov	Stupne voľnosti	podiel	F
skupinový	S_A	r-1	$S_A/(r-1)$	$\frac{S_A/(r-1)}{S_E/(n-r)}$
reziduálny	S_E	n-r	$S_E/(n-r)$	-
celkový	S_T	n-1	-	-

Rozhodnutie

- Určenie KH: z tabuliek Fisherovho rozdelenia $F_{1-\alpha}(r-1, n-r)$,
- Ak TK< KH⇒Ho nezamietame,
- Ak TK≥ KH⇒Ho zamietame,

v takom prípade je aspoň efekt jednej úrovne faktora preukazný, teda priemerna úroveň ukazovateľa sa štatisticky významne líši od ostatných.



Ak nulovú hypotézu zamietame:

- Zistili sme len, že je preukazný vplyv faktora na skúmaný znak,
- ďalej je potrebné skúmať medzi ktorými úrovňami faktora je a medzi ktorými nie je preukazný rozdiel - k tomúto účelu sa používajú testy kontrastov – Post Hoc testy
- Medzi testy kontrastov patria: Duncanov test, Scheffeho test, Tuckey test a iné…..

Podmienky použitia AR:

- Výbery pochádzajú z normálnych rozdelení,

  narušenie tohto predpokladu nemá podstatnejší vplyv na výsledky AR

- štatistická nezávislosť náhodných chýb e_{ij}

- zhodné reziduálne rozptyly

  $\sigma_1^2 = \sigma_2^2 = \ldots. = \sigma^2$, t.j. $D(e_{ij}) = \sigma^2$

  pre všetky i = 1,2…., m, j=1,2, …n

  tento predpoklad je závažnejší a možno ho overovať Cochranovým, resp. Bartlettovým , Leveneovým testom

Leveneov Test

- Určenie α.
- Stanovenie H_0: $\sigma_1^2 = ... = \sigma_r^2$
- Určenie TK

$$z_{ij} = \left| x_{ij} - \bar{x}_j \right|$$

$$\bar{x}_{Zj} = \frac{1}{n_j}\sum_{i=1}^{n_j} z_{ij}, \bar{x}_Z = \frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n_j} z_{ij}, S_{ZE} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(z_{ij} - \bar{x}_{Zj}\right)^2, S_{ZA} = \sum_{j=1}^{r} n_j \left(\bar{x}_{Zj} - \bar{x}_Z\right)^2$$

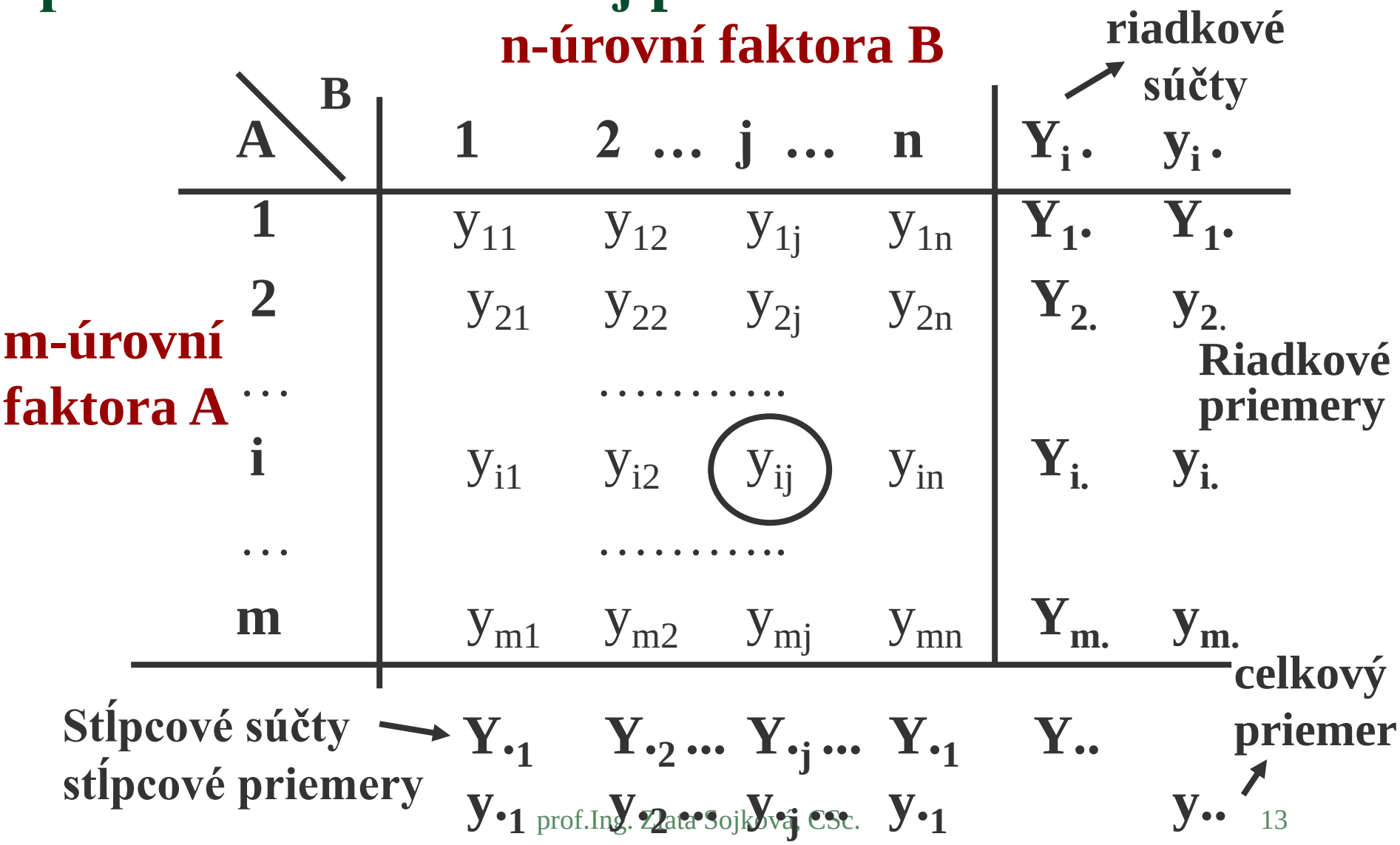
$$F_Z = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)}$$

- Určenie KH: z tabuliek Fisherovho rozdelenia F1-α(r-1, n-r)
- Rozhodnutie. Ak TK < KH platí H0.

Dvojfaktorová analýza rozptylu s jedným pozorovaním v každej podtriede ..DAR

- Uvažujme vplyv faktora A, ktorý skúmame na m - úrovniach, i = 1,2,….,m
- ďalej uvažujme faktor B, ktorý sledujeme na n - úrovniach , j = 1,2, …, n
- na každej i-tej úrovni faktora A a j-tej úrovni faktora B máme len jedno pozorovanie (opakovanie) y_{ij}
- overujeme tak vplyv dvoch nulových hypotéz

Schéma dvojfaktorového experimentu s jedným pozorovaním v každej podtriede DAR



Model pre skúmaný znak môžme zapísať

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Overujeme platnosť dvoch nulových hypotéz

Hypotéza pre faktor A:

$H_o^1: \alpha_1 = \alpha_2 = \ldots \alpha_i = \alpha_m = 0$

t.j. že efekty všetkých úrovni faktora A sú nulové, teda nepreukazné, oproti alternatívnej hypotéze

$H_1^1: \alpha_i \neq 0$ pre aspoň jedno ***i*** (i = 1,2…m)
efekt α_i aspoň jednej ***i*** - úrovne faktora je preukazný, významne odlišný od nuly

Hypotéza pre faktor B:

$H_o^2: \beta_1 = \beta_2 = \ldots \beta_j = \beta_n = 0$

t.j. že efekty všetkých úrovni faktora A sú nulové,
teda nepreukazné, oproti alternatívnej hypotéze

$H_1^2: \beta_j \neq 0$ pre aspoň jedno j (j = 1,2…m)
efekt β_j aspoň jednej j - úrovne faktora B je preukazný, významne odlišný od nuly

DAR Variabilita	1 Súčet štvorcov odchýlok	2 Stupne voľnosti	3 Priem. štvorec (1/2)	4 F-kritérium
Variabilita medzi riadkami	S_1	m-1	s_1^2	$F_1 = \frac{s_1^2}{s_r^2}$
Variabilita medzi stĺpcami	S_2	n-1	s_2^2	$F_2 = \frac{s_2^2}{s_r^2}$
Reziduálna variabilta	S_r	(m-1)(n-1)	s_r^2	
Celková variabilita	S_c	m.n -1		

Rozklad celkovej variability skúmaného znaku:

$S_c = S_1 + S_2 + S_r$

$$S_1 = n\sum_{i=1}^{m}(\bar{y}_i. - \bar{y}..)^2$$

Variabilita medzi riadkami, vplyv faktora A

$$S_2 = m\sum_{j=1}^{n}(\bar{y}._j - \bar{y}..)^2$$

Variabilita medzi stĺpcami, vplyv faktora B

$$S_r = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i. - \bar{y}._j + \bar{y}..)^2$$

Reziduálna variabilita

$$S_c = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}..)^2$$

Celková variabilita

Skúmanie vzťahov medzi štatistickými znakmi:

- Skúmanie vzťahov medzi kvalitatívnymi znakmi, napr. A⇔B , nazýme meranie asociácie
- skúmanie vzťahov medzi kvantitatívnymi štatistickými znakmi - regresná a korelačná analýza

Skúmanie asociácie

- Podkladom sú asociačné, resp. kontingenčné tabuľky,
- pre súhrné testovanie existencie štatisticky významného vzťahu medzi kvalitatívnymi znakmi sa používa **χ^2 - štvorcová kontingencia**

Ho: dva znaky A a B sú nezávislé

H_1: znaky A a B závisia

A znak má m - úrovní, obmien

B znak má k - úrovní , obmien

Formulovanie hypotéz

- Závislosť znakov sa prejaví v rozdielnych početnostiach
- napr. Skúmame či veľkosť balenia určitého výrobku je ovplyvnená početnosťou rodiny
- H_o : výber veľkosti balenia výrobku nezávisí od počtu členov v rodine
- H_1 : výber veľkosti balenia je ovplyvnená počtom členov v rodine
- test spočíva v porovnávaní empirických početností a teoretických, t,j, takých aké by mali empirické početnosti byť keby boli znaky A a B nezávislé

Simultánne početností, početnosti druhého stupňa $(a_i b_j)$

Marginálne početnosti (a_i) resp.(b_j)

Veľkosť balenia	Veľkosť rodiny 1-2 (b_1)	3-4 (b_2)	5 a viac (b_3)	Celkom
do 100g (a_1)	25 $(a_1 b_1)$	37 $(a_1 b_2)$	8	70
100-150g (a_2)	10	62	53	125
250g a viac (a_3)	5	41	59 $(a_3 b_3)$	105
Spolu	40	140	120	Celkový počet respondentov 300 **n**

Určovanie teoretických početností:

Vychádza z vety o nezávislosti náhodných javov A a B:

$P(A \cap B) = P(A) \,.\, P(B)$, teda ak znaky A a B sú nezávislé potom platí:

$P(a_i b_j) = P(a_i) \,.P(b_j)$

odhad na základe relatívnych početností :

$$\frac{(a_i b_j)_o}{n} = \frac{(a_i)}{n} \,.\, \frac{(b_j)}{n} \quad \Rightarrow \quad (a_i b_j)_o = \frac{(a_i) \,.(b_j)}{n}$$

Teoretické početnoesti

Výpočet teoretických početností

$(a_1b_1)_o = 70.40/300 = 9,33$

Veľkosť balenia	Veľkosť rodiny 1-2 (b_1)	3-4 (b_2)	5 a viac (b_3)	Celkom
do 100g	25	37	8	70
(a_1)	9.33	32,67	28.00	
100-150g	10	62	53	125
(a_2)	16.67	58.33	50	
250g a viac	5	41	59	105
(a_3)	14.00	49	42	
Spolu	40	140	120	Celkový počet respondentov 300 **n**

Výpočet testovacieho kritéria a rozhodnutie:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{((a_i b_j) - (a_i b_j)_o)^2}{(a_i b_j)_o}$$

Ak χ^2 vypočítané $\geq \chi^2$ pre hladinu významnosti α pre stupne voľnosti (m-1).(k-1)

$\Rightarrow$ H_o zamietame, t.zn. znaky A a B sú závislé

V našom prípade to znamená, že počet členov rodiny štatisticky významne ovplyvňuje výber veľkosti balenia výrobku. Ďalej by sme mali merať silu (tesnosť) závislosti.