Analýza rozptylu (ANOVA)

Analýza rozptylu s jedním faktorem (Jednoduché třídění)

Tuto metodu použijeme, jestliže potřebujeme ověřit významnost rozdílu mezi výběrovými průměry většího počtu náhodných výběrů.

Příklad 1.

Testujeme tři automobily stejné značky pro závod do vrchu. Vlivem nestejných podmínek (různí řidiči, povrch vozovky, atd.) se naměřené časy liší. Naším úkolem je zjistit, zda rozdíly v různých *průměrných časech* při opakovaných jízdách jsou způsobeny rozdílnou *kvalitou automobilů* nebo je lze přičíst na vrub náhodným vlivům.

A) Test proveďte na hladině významnosti alfa = 0,05.
B) Test proveďte na hladině významnosti alfa = 0,01.

Automobil	**Doba jízdy (min.)**			
A	5,3	5,2	5,4	4,9
B	5,9	5,8	5,7	
C	5,2	4,2	5,1	5,2

Řešení.

Faktor typ automobilu má 3 skupiny, tzn. $k = 3$, s následujícími četnostmi ve skupinách $n_1 = ...$, $n_2 = ...$, $n_3 = ...$ Celkový počet měření $n = …..$.

Testujeme nulovou hypotézu $H_0: \ \mu_1 = \mu_2 = \mu_3$, tj., že **doba jízdy nezávisí na automobilu.**

Před vypočtením testového kritéria musíme zjistit hodnoty následujících veličin:

- *Podmíněné průměry* $\bar{y}_i = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$, pro $i = 1,2...,k$, kde: y_{ij} jsou zjištěné hodnoty.
- *Celkový průměr* $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \dfrac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i$, kde: n je celkový rozsah souboru.
- *Meziskupinový součet čtverců* $S_{y,m} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$, kde: n_i je počet měření v jednotlivých skupinách, $\bar{y}_i$ je výběrový průměr v jednotlivých skupinách.
- *Vnitroskupinový součet čtverců* $S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$.
- *Celkový součet čtverců* $S_y = S_{y,m} + S_{y,v}$.

Pro ověření nulové hypotézy použijeme statistiku:
$$F = \frac{\dfrac{S_{y,m}}{k-1}}{\dfrac{S_{y,v}}{n-k}}, \qquad (*)$$

která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F(k-1, n-k)$. Kritická hodnota je $F_\alpha(k-1, n-k)$, kde: α je zvolená hladina významnosti. Kritický obor je dán intervalem $C = (F_\alpha(k-1, n-k), \infty)$.

Výpočet

Hodnoty výše uvedených veličin jsou:

$$\bar{y}_1 = \frac{5{,}3+5{,}2+5{,}4+4{,}9}{4} = \ldots. \qquad \bar{y}_2 = \frac{5{,}9+5{,}8+5{,}7}{3} = \ldots\ldots \qquad \bar{y}_3 =$$

$$\bar{y} =$$

$$S_{y,m} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 =$$

$$S_{y,v} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 =$$

$$S_y =$$

Testové kritérium F =

A) Kritická hodnota z tabulek $F_{0,05}(\ldots\ldots\ldots,\ldots\ldots\ldots\ldots) =$

Závěr:

B) Kritická hodnota z tabulek $F_{0,01}(\ldots\ldots\ldots,\ldots\ldots\ldots\ldots) =$

Závěr:

ANOVA tabulka

Zdroj měnlivosti	**Součty čtverců odchylek**	**Počty stupňů volnosti**	**Průměrné čtverce**	**Testové kritérium F**
Faktor *x* (meziskupinová variabilita)	S_{ym}	$k - 1$	$S_{ym}/(k-1)$	(*)
Reziduální (vnitroskupinová variabilita	S_{yv}	$n - k$	$S_{yv} / (n-k)$	
Celkový	S_y	$n - 1$		

Zdroj měnlivosti	Součty čtverců odchylek	Počty stupňů volnosti	Průměrné čtverce	Testové kritérium F
Mezi výběry (meziskupinová variabilita)	1,34	$k-1=2$	0,67	6,1
Všechny výběry (vnitroskupinová variabilita)	0,86	$n-k=8$	0,11	
Celkový	2,2	$n-1=12$		

A) Hodnota statistiky (*) $F=6{,}1$, kritická hodnota $F_{0,05}(2;8)=......$, kritický obor je $C=(.......;\infty)$. Hodnota F leží v kritickém oboru, proto zamítáme na hladině významnosti 5% nulovou hypotézu. **Doba jízdy závisí na automobilu.**

B) Hodnota statistiky (*) $F=6{,}1$, kritická hodnota $F_{0,01}(2;8)=......$, kritický obor je $C=(.......;\infty)$. Hodnota F leží v oboru přijetí, proto nelze zamítnout nulovou hypotézu. Na hladině významnosti 0,01 **nebylo prokázáno**, že by **doba jízdy závisela na automobilu.**

Příklad 2.
Pomocí determinačního poměru zjistěte těsnost závislosti doby jízdy na použitém automobilu.

Řešení.

Na otázku „Jak silná je vazba mezi nezávislou nominální proměnnou (typ automobilu) a proměnnou číselnou (doba jízdy) ?“, odpovídá hodnota *korelačního poměru*.

Korelační poměr $$P=\sqrt{\frac{S_{y,m}}{S_y}}, \qquad (**)$$

kde: $S_{y,m}$ je meziskupinový součet čtverců,

S_y je celkový součet čtverců.

Dosadíme-li do vztahu (**) dostaneme $P=\sqrt{\frac{1{,}34}{2{,}2}}=\sqrt{0{,}609}=0{,}78$.

Pokud hodnotu korelačního poměru umocníme, dostáváme poměr determinace $P^2=0{,}609$. Hodnoty determinačního poměru blízké 1 svědčí o *vysoké závislosti kvantitativního znaku na kvalitativním znaku.*

Příklad 3.
Na hladině významnosti 0,05 (0,01) testujte, zda názory respondentů na spokojenost s nakupováním v OD závisí na jejich příjmech.

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
H.CELK * PRIJEM	Between	(Combined)	87,955	4	21,989	12,697	,000
	Within Groups		238,982	138	1,732		
	Total		326,937	142			

Řešení.
Protože signifikantní hodnota je 0,00; což je menší než 0,05 (než 0,01); zamítáme nulovou hypotézu o nezávislosti příjmu a hodnocením spokojenosti s nakupováním v OD.

Analýza rozptylu se dvěma faktory (Dvojné třídění)

Příklad.
Bylo vybráno 6 řidičů , z nichž každý absolvoval s každým typem benzínu jednu jízdu. Na hladině významnosti 0,05 testujte, je-li průměrná spotřeba paliva závislá na typu použitého benzínu a na tom, který řidič s vozem jel.

	Řidiči						
Typ benzínu	A	B	C	D	E	F	**Průměry**
Aral	7,5	6,9	7,9	7,3	6,9	7,8	**7,38**
Shell	7,6	7,2	7,5	8	7,3	8,2	**7,63**
Benzina	7,2	8,1	7,8	7,6	7,8	6,9	**7,57**
Slovnaft	7	7,3	7,2	7,5	8,2	7,7	**7,48**
Průměry	**7,33**	**7,38**	**7,6**	**7,6**	**7,55**	**7,65**	**7,5**

Řešení.
Zkoumáme tedy závislost průměrné spotřeby (znak *Y*) na typu použitého benzínu (znak *X1*) a na řidiči (znak *X2*). (V tabulce jsou již doplněny podmíněné průměry a celkový průměr.)
Znak *X1* má $k = 4$ skupiny, znak *X2* má $r = 6$ bloků.

Pro faktor *X1* formulujeme hypotézu:
H_0 faktor *X1* neúčinkuje,
H_1 faktor *X1* účinkuje; tj. průměrná spotřeba závisí na použitém druhu benzínu.

Pro faktor *X2* formulujeme hypotézu:
H_0 faktor *X2* neúčinkuje,
H_1 faktor *X2* účinkuje; tj. průměrná spotřeba závisí na řidiči, který s vozem jel.

Výpočet jednotlivých součtů:

$$S_{ym} = r\sum_{i=1}^{4}(\bar{y}_{i.} - \bar{y})^2 = 6\left[(7{,}38 - 7{,}5)^2 + \ldots(7{,}48 - 7{,}5)^2\right] = 0{,}21$$

$$S_{yb} = k\sum_{j=1}^{6}(\bar{y}_{.j} - \bar{y})^2 = 4\left[(7{,}33 - 7{,}5)^2 + \ldots(7{,}38 - 7{,}5)^2\right] = 0{,}36$$

Potřebujeme znát i hodnotu součtu S_{yv}, z praktického hlediska je však výhodnější vypočítat hodnotu součtu S_y. Součet S_{yv} pak dopočteme ze vztahu $S_y = S_{ym} + S_{yv} + S_{yb}$.

$$S_y = \sum_{i=1}^{4}\sum_{j=1}^{6}\left(y_{ij} - \bar{y}\right)^2 = (7{,}5 - 7{,}5)^2 + (6{,}9 - 7{,}5)^2 + \ldots + (8{,}2 - 7{,}5)^2 + (7{,}7 - 7{,}5)^2 = 3{,}79$$

Po dosazení $S_{yv} = 3{,}22$.

Testové kritérium pro 1. hypotézu: $F = \dfrac{\dfrac{S_{ym}}{k-1}}{\dfrac{S_{yv}}{(k-1)(r-1)}} = \dfrac{\dfrac{0{,}21}{3}}{\dfrac{3{,}22}{3.5}} = 0{,}33$

V tabulce kritických hodnot F-rozdělení najdeme $F_\alpha(k-1; (k-1)(r-1))$ $F_{0,05}(3, 15) = 3{,}29$. Protože $0{,}33 < 3{,}29$, nelze zamítnout H_0, což znamená, že **použitý typ benzínu nemá na průměrnou spotřebu vliv**.

Testové kritérium pro 2. hypotézu: $F = \dfrac{\dfrac{S_{yb}}{r-1}}{\dfrac{S_{yv}}{(k-1)(r-1)}} = \dfrac{\dfrac{0{,}36}{5}}{\dfrac{3{,}22}{3.5}} = 0{,}34$

V tabulce kritických hodnot F-rozdělení najdeme $F_{0,05}(5, 15) = 2{,}9$. Protože $0{,}34 < 2{,}9$, nelze zamítnout H_0, což znamená, že **volba řidiče nemá na průměrnou spotřebu vliv**.

Domácí úkol:
Proveďte analýzu rozptylu u dvojného třídění.

Faktor / Blok	**1**	**2**
A	7	13
B	36	44
C	2	18

Výsledek: Hodnota Y závisí na typu faktoru, ale nezávisí na typu bloku. (pro 0,05)