ANALÝZA ROZPTYLU

Doc. Mgr. Jiří Mazurek, Ph.D.

Analýza rozptylu (ANOVA)

- Často používaná metoda v marketingovém výzkumu i jiných oblastech datové analýzy.
- Metoda umožňuje posoudit vliv různých úrovní/kategorií nějakého kvalitativního nebo kvantitativního znaku na kvantitativní veličinu.
- ANOVA testuje, zda existují rozdíly v populačních průměrech kvantitativního znaku, které náleží různým úrovním znaku kvalitativního.
- Například dovoluje hodnotit účinky různých reklamních kampaní na velikost tržeb z prodeje konkrétního produktu. Různé reklamní kampaně v tomto případě reprezentují různé kategorie sledovaného kvalitativního znaku (znak = reklamní kampaň). Velikost tržeb je pak zmíněný kvantitativní znak.

Aplikace ANOVY

- Nejdůležitější aplikací ANOVY je test rovnosti tří a více výběrových průměrů.
- Máme-li dva (výběrové) soubory, testujeme rovnost jejich středních hodnot pomocí Studentova t-testu.
- Máme-li však tři a více souborů, musíme použít ANOVU.

Základní idea ANOVY

- Matematicky spočívá základní myšlenka analýzy rozptylu v rozkladu celkového rozptylu kvantitativního znaku na dílčí rozptyly (meziskupinový a vnitroskupinový) příslušející jednotlivým vlivům, které tuto variabilitu způsobují.

- Kromě dílčích rozptylů je složkou celkového rozptylu také reziduální rozptyl, způsobený nepostiženými vlivy.

Rozdělení ANOVY

- Podle počtu analyzovaných faktorů rozlišujeme:
- jednofaktorovou,
- dvoufaktorovou
- vícefaktorovou analýzu rozptylu.

- Hovoříme také o jednoduchém a dvojném třídění, případně o tříděních vyšší úrovně (trojném, čtverném a podobně).

Jednofaktorová ANOVA

- Často se vyskytuje situace, kdy máme *k* nezávislých náhodných výběrů, které obecně nepocházejí z jednoho základního souboru.

- Tyto výběry jsou rozsahu $n_1, n_2 \ldots n_k$, což jsou obecně různá přirozená čísla. Číslo *k* může být 2, 3,...

- V každém z těchto náhodných výběrů je znám výběrový průměr $\bar{x}_i$, a také výběrový rozptyl s_i^2. Index *i* = 1,2,..., *k* vyjadřuje, o který výběr jde.

Rozdělení podle statistického znaku

- Základní soubor rozdělíme podle určitého třídícího statistického znaku *X* do k skupin a z každé z těchto k populací vybíráme n_i samostatně prvků.

- Znak *X* se pak označuje jako **faktor**, jehož úrovně, respektive kategorie jsou předem stanoveny a hovoří se proto často o **faktoru kontrolovaném**, nebo **faktoru pozorovaném**, např. věková skupina, druh výrobku, typ reklamy, typ služby apod.

- Faktor *X* má *k* úrovní (kategorií) a potenciálně ovlivňuje statistický znak *Y*, jenž má **kvantitativní**, tedy číselnou povahu.

Princip výpočtu

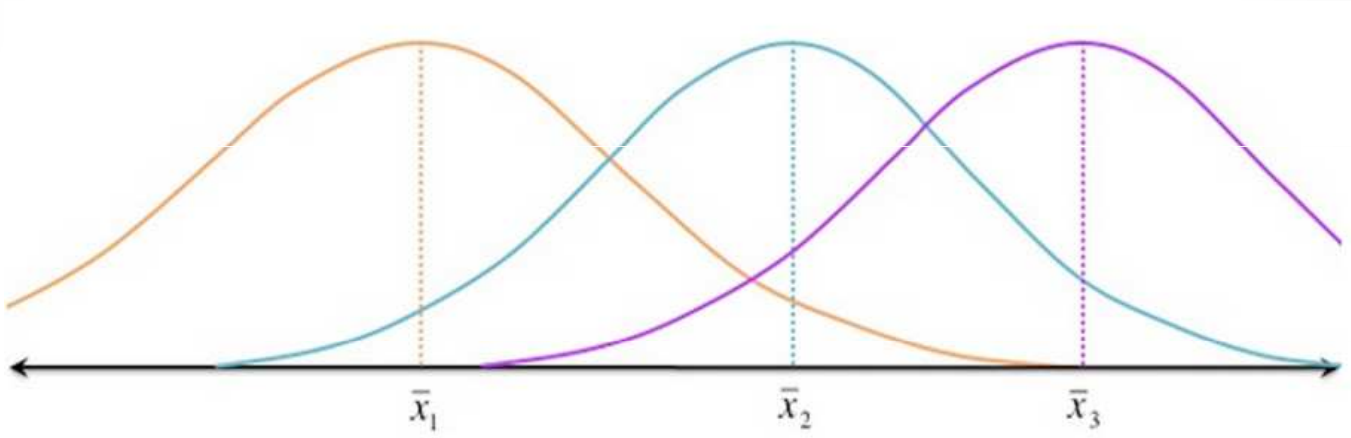
- Metoda analýzy rozptylu ANOVA spočívá v tom, že se celková variabilita měřená součtem čtverců odchylek zjištěných hodnot od celkového průměru rozdělí na variabilitu uvnitř jednotlivých výběrů a na variabilitu mezi jednotlivými výběry.

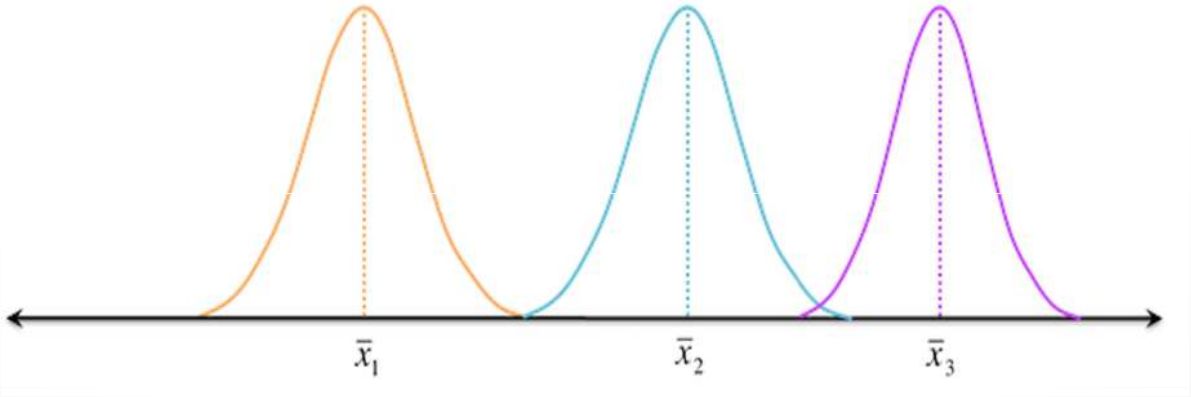
- **Analýza rozptylu je statistickým testem**.

- ANOVA má stejně jako i jiné statistické testy předpoklady svého použití. V případě ANOVA se předpokládá, že každý z k náhodných výběrů, s nimiž pracujeme, pochází z populace řídící se normálním rozdělením, že tato normální rozdělení mají stejný rozptyl a výběry jsou nezávislé.

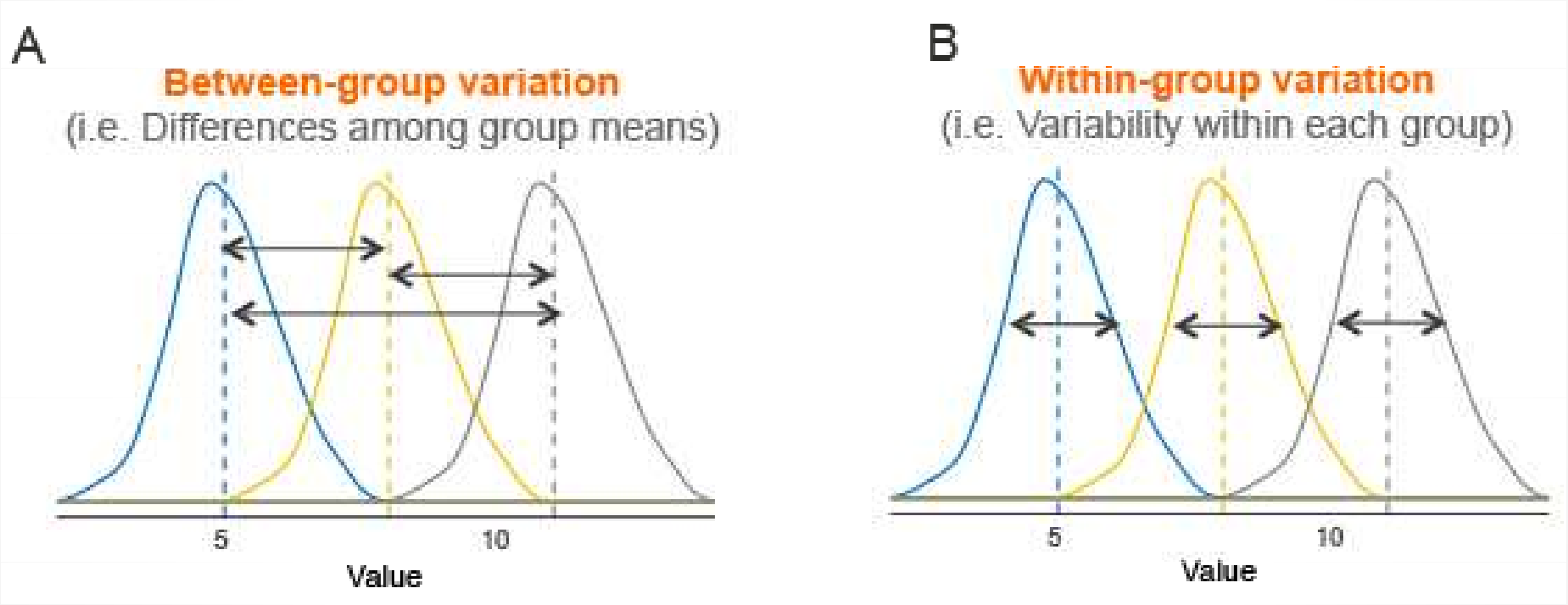
ANOVA

ANOVA





ANOVA


A
Between-group variation
(i.e. Differences among group means)
5
10
Value
B
Within-group variation
(i.e. Variability within each group)
5
10
Value

Postup testování: nulová hypotéza

- Testujeme nulovou hypotézu

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k$$

- Zkoumáme, zda střední hodnota (průměr) všech výběrů pochází ze stejné základní populace (základního souboru), což vzhledem k předpokladům učiněným pro ANOVA znamená, že si klademe otázku, zda střední hodnoty jsou stejné, respektive zda efekty jsou nulové.

- Alternativní hypotéza je negací nulové hypotézy.

Postup testování: testové kritérium

- Před vypočtením testového kritéria musíme zjistit hodnoty následujících veličin:
- ***Skupinové průměry:*** $\bar{y}_i = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$
- ***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, y_{ij} jsou zjištěné hodnoty a n_i počty prvků ve skupině *i*.

Postup testování: testové kritérium

- Dále musíme zjistit následující hodnoty:

- ***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:*** $S_y = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$

Postup testování: testové kritérium

- Platí: $S_y = S_{y,m} + S_{y,v}$

- V anglické literatuře nebo v softwarech je možné se setkat i s následujícím označením:
- $S_y = S_D$ (*D* z angl. Difference),
- $S_{y,m} = S_T$ (*T* z angl. Treatment),
- $S_{y,v} = S_R$ (*R* z angl. Residual).

Postup testování: testové kritérium

- Pro ověření nulové hypotézy použijeme statistiku:

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}}$$

která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F_{k-1,n-k}$.

Postup testování: kritická hodnota, výsledek

- Kritická hodnota je $F_{k-1,n-k}(\alpha)$, kde α je zvolená hladina významnosti.
- Kritický obor je dán intervalem:

$$C = \left(F_{\alpha}\left(k-1,\, n-k\right), \infty\right)$$

- Kritická hodnota testu pomocí funkce K = F.INV.RT() nebo v tabulkách.

Výpočet pomocí statistických programů

- *ANOVA tabulka*

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

Korelační poměr

- Na otázku „Jak silná je vazba mezi nezávislou nominální proměnnou a proměnnou číselnou?“, odpovídá hodnota *korelačního poměru*.

$$P = \sqrt{\frac{S_{y,m}}{S_y}}$$

Poměr determinace

- Pokud hodnotu korelačního poměru umocníme, dostáváme poměr determinace P^2.
- Hodnoty determinačního poměru blízké 1 svědčí o *vysoké závislosti mezi proměnnými.*
- Poměr determinace nabývá hodnot z intervalu [0,1]. Čím těsnější je závislost *Y* na *X*, tím více se hodnota poměru determinace blíží k jedné. Platí, že meziskupinový součet čtverců výrazně převažuje nad vnitroskupinovým součtem čtverců.
- Naopak, čím více se poměr determinace blíží k 0, tím menší část z celkového součtu čtverců připadá na meziskupinový součet čtverců, a tím menší je závislost znaku *Y* na *X*.

Příklad 1

Následující tabulka udává počet zákazníků, kteří navštívili 3 pobočky telefonního operátora během 5 pracovních dní. Našim úkolem je otestovat nulovou hypotézu, že průměrný počet zákazníků byl ve všech pobočkách stejný.

Pobočka 1	Pobočka 2	Pobočka 3
49	50	50
48	50	50
50	51	52
47	49	52
51	50	51

Tuto úlohu si vyřešíme Excelu. Určíme i korelační poměr a poměr determinace.

Řešení v Excelu

Anova: jeden faktor						
Faktor						
Výběr	Počet	Součet	Průměr	Rozptyl		
Sloupec 1	5	245	49	2.5		
Sloupec 2	5	250	50	0.5		
Sloupec 3	5	255	51	1		
ANOVA						
Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Mezi výběry (**S_{ym}**)	10	2	5	3.75	**0.05431**	3.885294
Všechny výběry (**Syv**)	16	12	1.333333			
Celkem (**S_y**)	26	14				

Závěr: H0 přijímáme.

Řešení v Excelu

Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Mezi výběry (**S_{ym}**)	10	2	5	3.75	**0.05431**	3.885294
Všechny výběry	16	12	1.333333			
Celkem (**S_y**)	26	14				

$$P^2 = \frac{S_{y,m}}{S_y} = \frac{10}{26} = 0{,}38.$$

$$P = 0{,}62.$$

Příklad 2

- Následující tabulka reprezentuje údaje získané nezávislými náhodnými výběry. Sledovaným faktorem je v tomto případě oktanové číslo pohonné směsi užívané v automobilech (90, 91, 95, 98). Máme tedy čtyři úrovně faktoru. Pro každou tuto úroveň byly náhodným výběrem čtyř řidičů zjištěny spotřeby automobilů. Zajímá nás otázka, zda oktanové číslo ovlivňuje (statisticky významně) úroveň spotřeby. Úlohu řešte v Excelu, alfa = 0,05.

	Faktor (oktanové číslo)			
	90	**91**	**95**	**98**
Spotřeba (červeně v litrech na 100 km)	8,1	7,7	7,6	7,5
	8	7,8	7,6	7,8
	7,9	7,9	7,5	7,6
	7,8	7,6	7,6	7,5

Příklad 3

- Pomocí ANOVY zjistěte, zda se mezi věkovými skupinami odlišuje čas strávený denně na internetu. Data 16 respondentů jsou uvedena v tabulce níže. Úlohu řešte v Excelu, alfa = 0,05.

	Věk			
Čas strávený na internetu (hod.)	15-24	25-34	35-44	45-54
	3	2	1,5	1
	4,5	1,5	2	1,5
	2,5	3,5	2	2
	3	4,1	2,5	1,5

Děkuji za pozornost