ANALÝZA ROZPTYLU

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.

Základní idea ANOVY

- Matematicky spočívá základní myšlenka analýzy rozptylu v rozkladu celkového rozptylu kvantitativního znaku na dílčí rozptyly 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 a 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 $n_1, n_2 \dots n_k$, 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 1 2 , ,...., k 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.

Postup testování: nulová hypotéza

Testujeme nulovou hypotézu

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

- Zkoumáme, zda všech k 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:
- *Podmíněné průměry*

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

- pro $i = 1,2,\dots,k$, kde y_{ij} jsou zjištěné hodnoty.
- *Celkový průměr*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i$$

- kde *n* je celkový rozsah souboru.

Postup testování: testové kritérium

- Dále musíme zjistit hodnoty
- *Meziskupinový součet čtverců*

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

- Kde n_i je počet měření v jednotlivých skupinách, $\bar{y}_i$ je výběrový průměr v jednotlivých skupinách.
- *Vnitroskupinový součet čtverců*

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

- *Celkový součet čtverců* .

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

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:

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

- 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

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

- Kritická hodnota testu pomocí funkce K = FINV() 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, tím více se také meziskupinový součet čtverců blíží k celkovému součtu čtverců, přičemž vnitroskupinový součet čtverců se blíží k nule. 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

H1	H2	H3
49	50	50
48	50	50
50	51	52
47	49	52
51	50	51

Řešení

Anova: jeden faktor							
Faktor							
Výběr		Počet	Součet	Průměr	Rozptyl		
Anova: jeden faktor							
Faktor							
Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*			
H1	5	245	49	2,5			
H2	5	250	50	0,5			
H3	5	255	51	1			
ANOVA							
Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*	
Mezi výběry	10	2	5	3,75	0,054310001	3,885293835	
Všechny výběry	16	12	1,3333333 33				
Celkem	26	14					

Děkuji za pozornost