KGG/STG Statistika pro geografy

8. Analýza rozptylu

Mgr. David Fiedor
13. dubna 2015

Motivace

- dosud - maximálně dva výběry (jednovýběrové a dvouvýběrové testy)

Příklad

Na dané hladině významnosti $\alpha = 0{,}05$ testujte hypotézu, že všechny okresy Olomouckého kraje dosahují za období 2000–2012 průměrně stejné míry nezaměstnanosti.

t-test vs. analýza rozptylu

- máme m náhodných výběrů z normálních rozdělení s parametry $\mu_1, \mu_2, \ldots, \mu_m$ a společným rozptylem σ^2, který není znám
- testovat hypotézu o shodnosti středních hodnot $\mu_1 = \mu_2 = \cdots = \mu_m$
- nápad: utvořit si dvojice náhodných výběrů systémem „každý s každým", tedy $\dfrac{m(m-1)}{2}$ dvojic náhodných výběrů a na každou dvojici použít dvouvýběrový t-test na zvolené hladině významnosti α

t-test vs. analýza rozptylu

- zdá se, že pokud bychom našli aspoň jednu dvojici, u které bychom t-testem zamítli nulovou hypotézu o shodnosti středních hodnot $\mu_k = \mu_l$, mohli bychom zamítnout hypotézu $\mu_1 = \mu_2 = \cdots = \mu_m$
- v čem je však problém?

t-test vs. analýza rozptylu

- problém v tomto případě vzniká u pravděpodobnosti chyby I. druhu, která by zcela určitě nebyla rovna hodnotě zvolené hladiny významnosti α, ale byla by o mnoho větší – uvědomme si, že by stačilo u jednoho z $\frac{m(m-1)}{2}$ t-testů zamítnout nulovou hypotézu, aby byla zamítnuta celková nulová hypotéza o shodě středních hodnot

$\Rightarrow$ nutnost použití jiné metody - ANOVA (=Analýza rozptylu)

Analýza rozptylu

- jednofaktorová ANOVA - jádro této kapitoly
- neparametrická obdoba analýzy rozptylu
- analýza rozptylu dvojného třídění

Faktor

Faktorem nazýváme proměnnou, která má více variant a která je zpravidla nominálního typu, například národnost (česká x polská x $\cdots$).

Faktorem mohou být:

- různé metody práce (zjišťujeme, zda všechny metody vedou ke stejnému (či podobnému) výsledku, nebo zda na zvolené hladině významnosti bude některá metoda lepší a vhodnější pro danou práci)

Obecný model analýzy roztylu jednoduchého třídění

- jednofaktorová analýza rozptylu zkoumá závislost intervalové či poměrové proměnné na vybraném faktoru
- závislost proměnné na tomto faktoru se projeví tím, že existuje statisticky významný rozdíl ve středních hodnotách proměnné v náhodných výběrech, jenž vznikly „tříděním“ podle variant faktoru

Analýza rozptylu

Proč název analýza rozptylu?

- podstata spočívá v tom, že celkový rozptyl sledované proměnné se rozloží na *rozptyl uvnitř jednotlivých výběrů* a na *rozptyl mezi výběry*
- pokud je rozptyl mezi výběry příliš (nepravděpodobně) velký, bude tato situace svědčit o významném vlivu faktoru, podle kterého jsme dané třídění na jednotlivé náhodné výběry provedli
- situace vede k zamítnutí nulové hypotézy o shodě středních hodnot jednotlivých náhodných výběrů

Označení

- počet výběrů m, $m > 2$
- rozsahy těchto výběrů $n_1, n_2, \ldots, n_m$ nemusí být obecně stejné
- n celkový rozsah, $n = n_1 + \cdots + n_m$
- průměr $\bar{x}_j$ a rozptyl s_j^2, kde $j = 1, 2, \ldots, m$
- prvek x_{ij} pak označuje i-té pozorování v j-tém výběru

Označení

Tabulka: Označení základních charakteristik jednotlivých náhodných výběrů

	Skupina 1	Skupina 2	...	Skupina m
Měření 1	x_{11}	x_{12}	...	x_{1m}
Měření 2	x_{21}	x_{22}	...	x_{2m}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
Rozsah	n_1	n_2	...	n_m
Průměr	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_m$
Rozptyl	s_1^2	s_2^2	...	s_m^2

Teoretické vysvětlení

Každé pozorované x_{ij} (pro $i = 1, \ldots, n$ a $j = 1, \ldots, m$) se řídí modelem složeným z celkové střední hodnoty μ, skupinovým efektem α_j (efekt faktoru) a blíže nespecifikovanou náhodnou veličinou ε_{ij} s rozdělením $N(0, \sigma^2)$, které říkáme *náhodná chyba* ε_{ij}, tedy:

$$x_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Celkovou střední hodnotu μ, kterou neznáme, můžeme vyjádřit ze vztahu: $\mu_j = \mu + \alpha_j$ a přepsat tak výše uvedený model do tvaru:

$$x_{ij} = \mu_j + \varepsilon_{ij}$$

Teoretické vysvětlení

Parametry μ, α_j neznáme, avšak požadujeme, aby platila tzv. *reparametrizační rovnice* $\sum_{j=1}^{m} n_j\alpha_j = 0$.

Pokud mají výběry stejný rozsah, lze použít zjednodušenou podmínku ve tvaru $\sum_{j=1}^{m} \alpha_j = 0$.

Podstata analýzy rozptylu

- rozdělení celkového rozptylu S_T závisle proměnné do dvou částí (na variabilitu uvnitř jednotlivých výběrů S_E a variabilitu mezi jednotlivými náhodnými výběry S_A)

$$S_T = S_A + S_E$$

Podstata analýzy rozptylu

- variabilita S_E (reziduální) uvnitř jednotlivých výběrů popisuje, jak se každá z hodnot tohoto výběru liší od výběrového průměru, a to pomocí součtu čtverců těchto rozdílů, tj.:

$$S_E = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

s počtem stupňů volnosti $\nu_E = n - m$

Podstata analýzy rozptylu

- variabilitu S_A mezi jednotlivými náhodnými výběry charakterizuje skupinový[1] součet čtverců

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

s počtem stupňů volnosti $\nu_A = m - 1$

[1] Každý náhodný výběr budeme označovat též pojmeme skupina.

Podstata analýzy rozptylu

- celkový součet čtverců, který charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru, určíme ze vztahu:

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

s počtem stupňů volnosti $\nu_T = n - 1$

Předpoklady použití analýzy rozptylu

1. *Nezávislost jednotlivých náhodných výběrů jak uvnitř skupin, tak i mezi skupinami.* Tento předpoklad je velmi důležitý a musí být vždy splněn, jinak bychom mohli obdržet nesmyslné výsledky.
2. *Normalita dat.* Při mírném porušení tohoto předpokladu ještě stále můžeme použít parametrickou analýzu rozptylu, zvláště v případě výběrů s většími rozsahy. Při výraznějším porušení normality doporučujeme použít Kruskalův-Wallisův test, o kterém se ještě zmíníme v rámci neparametrických metod analýzy rozptylu.

Předpoklady použití analýzy rozptylu

3. *Shoda rozptylů jednotlivých náhodných výběrů.* Znovu platí, že mírné porušení shody rozptylů není překážkou a nebrání nám v použití parametrických metod analýzy rozptylu, zatímco při vážnějším porušení znovu doporučujeme použít Kruskalův-Wallisův test, či jemu podobný mediánový test. V následující podkapitole si předvedeme, jak lze testovat tento předpoklad, který v praxi ověřujeme až po zjištění, že je splněna podmínka normality dat.

Doporučený postup při testování analýzy rozptylu

1. ověření normality jednotlivých výběrů
   - potřeba ověřit všechny náhodné výběry
   - kombinace grafického ověření a testu normality
   - lehké porušení tohoto předpokladu lze akceptovat (v opačném případě nutnost použití neparametrického ekvivalentu analýzy rozptylu)
2. ověřování shody rozptylů jednotlivých náhodných výběrů
   - grafická možnost ověření - box plot (krabicový diagram)
   - testy - Levenův a Brownův-Forsytheův test
   - opět lze akceptovat mírné porušení shody rozptylů

Doporučený postup při testování analýzy rozptylu

3. jestliže jsou splněny výše uvedené předpoklady, lze přistoupit k samotnému testování (je vhodné spočítat si předem všechny průměry a rozptyly)
4. pokud zamítneme nulovou hypotézu o shodě všech středních hodnot, bude nás přirozeně zajímat, které dvojice výběrů se od sebe liší (metody mnohonásobného porovnávání)

Ověřování předpokladu o shodě rozptylů

- nulová hypotéza je vždy stejná – H$_0$: $\sigma_1^2 = \cdots = \sigma_m^2$
- alternativní hypotéza tvrdí, že *aspoň* jedna dvojice rozptylů se od sebe liší
- mimo samotné testy existuje také kritérium, jehož splnění postačuje ke splnění předpokladu, přičemž s_j označuje směrodatné odchylky měření v jednotlivých skupinách:

$$\frac{\max s_j}{\min s_j} \leq 3$$

Levenův test

- test velmi podobný samotné analýze rozptylu - i testová statistika se asymptoticky řídí stejným rozdělením se stejným počtem stupňů volnosti
- založen je na analýze rozptylu absolutních hodnot centrovaných pozorování

Brownův-Forsytheův test

- modifikace Levenova testu
- zjednodušeně řečeno v porovnání s Levenovým testem tento test využívá k určení testové statistiky mediány jednotlivých výběrů (Levenův test používá výběrové průměry)

Testování hypotézy o shodě středních hodnot

- na hladině významnosti α testujeme nulovou hypotézu H_0: $\mu_1 = \cdots = \mu_m$ proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice středních hodnot náhodných výběrů se liší
- nulová hypotéza nám jinými slovy říká, že vliv faktoru, podle kterého rozlišujeme výběry, není významný

Testování hypotézy o shodě středních hodnot

- testová statistika má tento tvar:

$$F_A = \frac{S_A/\nu_A}{S_E/\nu_E}$$

  a řídí se rozdělením $F(m-1, n-m)$, platí-li nulová hypotéza
- pokud realizace testovacího kritéria bude patřit do kritického oboru $W = \langle F_{1-\alpha}(m-1, n-m), \infty)$, zamítneme nulovou hypotézu na dané hladině významnosti α

Testování hypotézy o shodě středních hodnot

- výpočet testovacího kritéria je ve tvaru podílu míry variability[2] mezi skupinami a uvnitř jednotlivých skupin
- kritický obor nikdy nepokryje hodnoty menší než 1 (v tomto případě by totiž byla variabilita mezi skupinami dokonce menší než variabilita uvnitř skupin), není proto důvod pro hodnoty realizace testovacího kritéria menší než 1 zamítat nulovou hypotézu o shodě středních hodnot

[2] Mírou variability *MS* (průměrným čtvercem) rozumíme součty čtverců dělené odpovídajícím počtem stupňů volnosti.

Testování hypotézy o shodě středních hodnot

- čím větší bude realizace testovacího kritéria, tím pravděpodobněji budeme nuceni nulovou hypotézu zamítnout a přiklonit se k alternativní hypotéze
- přesné výsledky dostaneme až samotným porovnáním hodnoty testovacího kritéria a kritického oboru

Testování hypotézy o shodě středních hodnot

Tabulka: Ukázková tabulka výsledků analýzy rozptylu

Variabilita		ν	Míra var. *MS*	F_A
skupiny	S_A	$\nu_A = m - 1$	S_A/ν_A	$\frac{S_A/\nu_A}{S_E/\nu_E}$
reziduální	S_E	$\nu_E = n - m$	S_E/ν_E	–
celkový	S_T	$\nu_T = n - 1$	–	–

Metody mnohonásobného porovnávání

- v případě zamítnutí shody průměrů nás bude zajímat, které dvojice výběrů se liší
- tyto metody se často označují jako post-hoc (následné)
- těchto metod existuje celá řada - uvedeme si následující dvě (Bonferronniho metoda, Tukeyova metoda)
- ani tady není možno použít dvojice t-testů (stále se jedná o jednotlivé dvojice z mnoha náhodných výběrů)

Bonferonniho metoda

- smysl této metody spočívá v rozdělení hladiny významnosti α mezi všechna porovnání, přičemž na každou dvojici aplikujeme modifikovaný dvouvýběrový t-test
- např. pro 3 náhodné výběry bude hladina významnosti pro každý t-test rovna $\frac{\alpha}{3}$, jelikož pro tři výběry existují tři dvojice t-testů
- obecně platí pro testování m výběrů na hladině významnosti α, že nová hladina významnosti pro jednotlivé dvouvýběrové t-testy bude rovna $\dfrac{\alpha}{\frac{m(m-1)}{2}}$

Bonferonniho metoda

- modifikace testu nespočívá pouze ve změně hladiny významnosti - ve vzorci testové statistiky se zamění rozptyl variantou uvažující variabilitu všech výběrů, tj.:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_E}{\nu_E}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- za platnosti nulové hypotézy má testová statistika Studentovo t-rozdělení s ν_E stupni volnosti

Tukeyova metoda

- vhodná především pro vyvážené třídění (všechny výběry mají stejný rozsah), ale existuje i varianta pro nestejné rozsahy
- varianta pro výběry se stejným rozsahem
- na hladině významnosti α zamítneme nulovou hypotézu H_0 o shodě středních hodnot μ_k a μ_l, když $|\bar{x}_k - \bar{x}_l| \geq q_{1-\alpha}(m, n-m)\frac{s_*}{\sqrt{p}}$, kde kvantily $q_{1-\alpha}(m, n-m)$ studentizovaného rozpětí najdeme ve statistických tabulkách

Příklad: Počet rozvedených podle velikosti měst

Všechny obce jsme seřadili podle velikosti a rozčlenili do osmi skupin podle velikosti (strukturu rozdělení najdete v souboru *Obce podle velikosti*). Z těchto skupin jsme podle abecedního pořadí vybrali z každé skupiny 10 obcí, takže výsledný náhodný výběr má celkový rozsah 80. U obcí jsou uvedeny jejich velikosti (podle počtu obyvatel), počet ženatých a rozvedených mužů a vypočtený ukazatel, který je podílem počtu rozvedených a ženatých mužů. Testujte nulovou hypotézu na hladině významnosti $0,01$, že rozdíly mezi vypočteným ukazatelem ve skupinách měst sestavených podle velikosti jsou způsobeny pouze náhodnými vlivy.

Řešení

- ověření předpokladu normality dat
  - pro každou skupinu vytvoříme N-P plot
  - doplníme i testem normality: *Grafy–2D grafy–Normální pravděpodobnostní grafy*
  - zvolíme proměnnou, zaškrtneme S-W test a na kartě *Kategorizovaný* zaškrtneme u kategorie X *Zapnuto* a změníme proměnnou, podle které budeme kategorizovat, tj. *Velikost obce*
  - osm N-P plotů s hodnotami S-W testů

Řešení

- ověření předpokladu shody rozptylů
  - ověříme „těsně před“ samotným provedením analýzy rozptylu
- *Statistiky – Základní statistiky/tabulky – Rozklad & jednofakt. ANOVA – OK*
  - vybereme proměnné a potvrdíme
- na kartě *ANOVA & testy* máme vše, co potřebujeme, tj. testy homogenity rozptylů a samotnou analýzu rozptylu
  - také můžeme zvolit hladinu významnosti a meze intervalu spolehlivosti

Řešení

	Brown-Forsytheův test homogenity rozptylů (Počet rozvedených podle velikosti měst) Označ. efekty jsou význ. na hlad. p < ,01000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Počet rozvedených na počet ženatých	0,026292	7	0,003756	0,125568	72	0,001744	2,153695	0,048557

Obrázek: Výstup ve formě Brownova-Forsytheova testu homogenity rozptylů s vyznačenou *p*-hodnotou

- nelze zamítnout nulovou hypotézu o shodě rozptylů na dané hladině významnosti (předpoklad je tedy splněn)

Řešení

Proměnná	Analýza rozptylu (Počet rozvedených podle velikosti měst) Označ. efekty jsou význ. na hlad. p < ,01000 SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba
Počet rozvedených na počet ženatých	0,039490	7	0,005641	0,258730	72	0,003593

Obrázek: Výstup z provedení analýzy rozptylu

- porovnáním *p*-hodnoty s hladinou významnosti učiníme závěr, že jsme nedospěli k přesvědčení, že se střední hodnoty skupin liší
- nezamítáme nulovou hypotézu na dané hladině významnosti $\alpha = 0,01$

K-W test

- $m \geq 3$ nezávislých náhodných výběrů
- výběry pocházejí ze spojitých rozdělení (není potřeba znát konkrétní rozdělení)
- celkový rozsah $n = n_1 + \cdots + n_m$
- na dané hladině významnosti α testujeme hypotézu, že všechny tyto náhodné výběry pocházejí z téhož rozdělení proti alternativní hypotéze, která tvrdí, že existuje aspoň jedna dvojice výběrů, která se od sebe liší

Postup provedení K-W testu

a) Všech n hodnot seřadíme do rostoucí posloupnosti, přičemž dále u každé hodnoty určíme její pořadí (případně průměrné pořadí).

b) Podle zavedené symboliky určíme u každého výběru součet pořadí $\sum R_j$.

c) Platí-li nulová hypotéza H_0, pak testová statistika

$$H = \left(\frac{12}{n(n+1)} \sum_{j=1}^{m} \frac{\left(\sum R_j\right)^2}{n_j} \right) - 3(n+1)$$

se asymptoticky řídí rozdělením $\chi^2(m-1)$.

Postup provedení K-W testu

d) Kritický obor má tvar: $W = \langle \chi^2_{1-\alpha}(m-1), \infty)$.

e) Nulovou hypotézu H_0 zamítáme na asymptotické hladině významnosti α, jestliže $H \in W$. V případě interpretace p-hodnoty zamítáme nulovou hypotézu, když vypočtená p-hodnota je menší než zvolená hladina významnosti α.

K-W test

- v případě zamítnutí nulové hypotézy je potřeba zjistit, které dvojice výběrů se od sebe liší
- dva náhodné výběry pocházejí z různých rozdělení, jestliže platí

$$|r_k - r_l| > \sqrt{\frac{1}{12}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)n(n+1)h_\alpha(m-1)},$$

kde $h_\alpha(m-1)$ je kritická hodnota Kruskalova-Wallisova testu na hladině α s daným počtem stupňů volnosti a $r_k = \frac{\sum R_k}{n_k}$ a $r_l = \frac{\sum R_l}{n_l}$ (k, l jsou různé prvky z množiny $1, 2, \ldots, m$)

Příklad

Máme k dispozici soubor, kde jsou různé demografické údaje za rok 2010 pro vybrané obce České republiky (roztříděny jsou podle polohy, přičemž grupovací proměnná určuje příslušnost k dané skupině – kraj), mj. také údaje o počtu obyvatel a přirozeném přírůstku. Poslední proměnná, která je vypočítána jako podíl přirozeného přírůstku k celkovému počtu obyvatel, nazvaná *Přirozený přírůstek/Stav obyvatelstva* je uměle vytvořená. Testujte na hladině významnosti $\alpha = 0,05$ hypotézu, že geografická poloha nehraje žádnou roli v přírůstku obyvatelstva přepočítaného na celkový počet obyvatel.

Řešení

- otestování normality dat (*Grafy – 2D grafy – Normální pravděpodobnostní grafy*)
- po zamítnutí hypotézy o normalitě dat, přistoupíme ke K-W testu
- *Statistiky – Neparametrická statistika – Porovnání více nezávislých vzorků (skupiny) – OK*
- *Shrnutí: Kruskal-Wallis. ANOVA a mediánový test*

Řešení

Závislá: Přirozený přírůstek/Stav obyvatelstva	Kruskal-Wallisova ANOVA založ. na poř.; Přirozený přírůstek/Stav obyvatelstva (Přirozený přírůstek) Nezávislá (grupovací) proměnná : Grupovací proměnná Kruskal-Wallisův test: H (4, N= 100) =5,288236 p =,2590			
	Kód	Počet platných	Součet pořadí	Prům. Pořadí
1	1	20	930,000	46,50000
2	2	20	1207,500	60,37500
3	3	20	848,000	42,40000
4	4	20	1122,500	56,12500
5	5	20	942,000	47,10000

Obrázek: Výstup po provedení K-W testu

- p-hodnota větší než zvolená hladina významnosti, není důvod zamítat nulovou hypotézu
- v případě zamítnutí nulové hypotézy, by bylo potřeba provést obdobu metod mnohonásobného porovnávání, kterou zajistíme stiskem tlačítka *Vícenás. porovnání průměrného pořadí pro vš. sk.*

Motivace

- potřeba třídit celkový náhodný výběr podle více faktorů (třídicích znaků)
- např. geografie zemědělství - výnosy určité plodiny
  - závislost na typu půdy
  - závislost na druhu hnojiva, které bylo použito

Analýza rozptylu dvojného třídění

- máme dány faktory A a B, přičemž symbolem a označme, že faktor A má a úrovní (variant, kterých může nabýt), obdobně faktor B má b úrovní
- počet objektů ve výběru odpovídající i-té úrovni faktoru A a j-té úrovni faktoru B, označme symbolem n_{ij}
- zkoumáme tři páry hypotéz
  1. nulová hypotéza je ve tvaru: $H_{0_1} : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$, což znamená, že skupinové efekty faktoru A jsou všechny nulové (alternativní hypotéza tvrdí, že existuje aspoň jeden skupinový efekt různý od nuly)

Analýza rozptylu dvojného třídění

2. obdobně druhý pár hypotéz je ve tvaru $H_{0_2} : \beta_1 = \beta_2 = \cdots = \beta_a = 0$, což znamená, že skupinové efekty faktoru B jsou všechny nulové (alternativní hypotéza tvrdí, že existuje aspoň jeden skupinový efekt různý od nuly)
3. poslední hypotézy se týkají interakcí[3], kdy nulová hypotéza H_{0_3} tvrdí, že mezi faktory A a B není žádná interakce, všechny jsou rovny nule (alternativní hypotéza naopak říká, že některé interakce jsou nenulové. Tuto hypotézu o existenci interakcí testujeme jako první)

[3] To znamená, že faktory nepůsobí izolovaně, neboli nejsou nezávislé. Významné interakce způsobují, že jednotlivé faktory nevysvětlují veškerou variabilitu.

Analýza rozptylu dvojného třídění

Testová statistika F opět vychází z rozkladu variability na jednotlivé složky, tj.: $S_T = S_A + S_B + S_I + S_E$, kde S_T označuje celkovou variabilitu, S_A, S_B značí efekty faktoru A, B, S_E je variabilita uvnitř skupin a symbolem S_I označujeme interakce.

Analýza rozptylu dvojného třídění

Tabulka: Ukázková tabulka výsledků analýzy rozptylu dvojného třídění

Variabilita		ν	MS	F	H_0
faktor A	S_A	$\nu_A = a - 1$	S_A/ν_A	$\frac{S_A/\nu_A}{S_E/\nu_E}$	H_{0_1}
faktor B	S_B	$\nu_B = b - 1$	S_B/ν_B	$\frac{S_B/\nu_B}{S_E/\nu_E}$	H_{0_2}
interakce	S_I	$\nu_I = (a-1)(b-1)$	S_I/ν_I	$\frac{S_I/\nu_I}{S_E/\nu_E}$	H_{0_3}
reziduální	S_E	$\nu_E = ab(c-1)$	S_E/ν_E	–	–
celkový	S_T	$\nu_T = abc - 1$	–	–	–

Děkuji za pozornost...