1 10-Analýza rozptylu jednoduchého třídění, ANOVA, Jednofaktorová analýza rozptylu

1.1 Nová látka

1.1.1 Testování homogenity rozptylů u r náhodných výběrů

- homogenita (stejnorodost) rozptylů u většího množství náhodných výběrů je důležitým předpokladem, který musí být splněn, abychom mohli provést tzv. ANOVU - jednofaktorovou analýzu rozptylu (viz dále).
- předpokládejme, že máme $r \geq 2$ náhodných výběrů
- testujeme nulovou hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$ oproti alternativní hypotéze H_1 : alespoň jedna dvojice rozptylů se liší
- testy rozptylu
  1. Levenův test
     - levene.test(D,K) knihovna lawstat ; $D\ldots$ vektor dat, $K\ldots$ typ skupiny
     - je založen na analýze rozptylu absolutních hodnot centrovaných pozorování
     - výpočet je založen na 'hraní si' s odhady středních hodnot
  2. Brownův-Forsytův test
     - je modifikací Levenova testu
     - je založen na mediánu (namísto střední hodnoty)
     - při větších rozsazích náhodných výběrů ($n_i > 20$) jej lze použít i na data, které nejsou z normálního rozdělení
     - v Rku ho používat nebudeme, ale je dobré, abyste o něm aspoň slyšeli
  3. Bartlettův test
     - bartlett.test(D,K) knihovna stat
     - můžeme jej použít, pouze pokud jsou rozsahy všech výběrů větší než 6
     - nelze jej použít, pokud je více náhodných výběrů z výrazně nenormálního rozložení

1.1.2 ANOVA - Jednofaktorová analýza rozptylu

- zkoumá závislost intervalové/poměrové proměnné X na nominální proměnné A, které má alespoň dvě varianty
- $A\ldots$ *faktor*; varianty $A\ldots$ *úrovně faktoru*
- závislost X na A se projeví tím, že existuje statisticky významný rozdíl v průměrech proměnné X v náhodných výběrech, které vznikly tříděním podlě variant proměnné A.
- motivační příklady
  - má metoda výuky (faktor A) vliv na počet bodů (intervalová proměnná X) dosažených studenty v závěrečném testu?

- má typ potravy pračlověka (A) vliv na šířku stoliček (X)?
- má způsob života (A: na stromu-šplh; na zemi - šplhá málo) vliv na intenzitu svalových úponů na rukou (X)?
- má pohlaví (A) vliv na hmotnost člověka (X), nebo na šířku očnic (X)?

- trocha matematiky
  - předpokládáme, že faktor A má $r \geq 2$ úrovní $A_1, \dots A_r$, přičemž i-té úrovni odpovídá n_i pozorování $X_{i1}, \dots X_{in_i}$. Tato pozorování tvoří náhodný výběr z $N(\mu_i, \sigma^2)$,$i = 1, \dots, r$. Celkový počet pozorování je $n = \sum_{i=1}^{r} n_i$. Jednotlivé náhodné výběry jsou stochasticky nezávislé.
  - !!! Před samotnou ANOVOU musíme vždy ověřit předpoklady normality všech výběrů (r testů) a homogenity rozptylů (1 hromadný test)
  - Tečková anotace
    * součet hodnot v i-tém výběru
$$X_{i.} = \sum_{j=1}^{n_i} X_{ij}$$
    * výběrový průměr v i-tém výběru
$$M_{i.} = \frac{1}{n_i} X_{i.}$$
      - klasický aritmetický průměr dat z i-té skupiny, jen trochu jinak zapsaný
    * součet hodnot všech výběrů
$$X_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$$
    * celkový průměr všech r výběrů
$$M_{..} = \frac{1}{n} X_{..}$$
      - klasický aritmetický průměr všech dat
    * *celkový součet čtverců*
$$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{..})^2$$
      - charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru
      - stejný princip jako výběrový rozptyl, akorát ho nedělíme počtem pozorování
      - má počet stupňů volnosti $f_T = n - 1$
    * skupinový součet čtverců
$$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2$$
      - charakterizuje variabilitu mezi jednotlivými náhodnými výběry
      - má počet stupňů volnosti $f_A = r - 1$

* reziduální součet čtverců
$$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{i.})^2$$
- charakterizuje variabilitu uvnitř jednotlivých výběrů - má počet stupňů volnosti $f_E = n - r$

– lze dokázat $S_T = S_A + S_E$.

1.1.3 Testování hypotéz o shodě středních hodnot

- na hladině významnosti α testujeme nulovou hypotézu, která tvrdí že všechny střední hodnoty jsou stejné $H_0 : \mu_1 = \cdots = \mu_r$, proti alternativní hypotéze H_1 : *Alespoň jedna dvojice středních hodnot se významně liší.*
- jiná definice nulové a alternativní hypotézy (z pohledu faktoru A): H_0: Vliv faktoru A není významný; H_1: Vliv faktoru A je významný.
- Testovací statistika má tvar
$$F_A = \frac{S_A/f_A}{S_E/f_E} \sim F(r-1, n-r). \tag{1}$$
- H_0 zamítáme na hladině významnosti α, pokud $F_A \in \langle F_{1-\alpha}(r-1, n-r), \infty)$
- případně H_0 zamítáme na hladině významnosti α, pokud p-hodnota$< \alpha$ (testování H_0 přes p-hodnotu jsme na hodině nedělali).
- výsledky výpočtů lze zapsat do přehledné tabulky:

Zdroj variability	součet čtverců	stupně volnosti	průměrný čtverec	F_A
skupiny	S_A	$f_A = r-1$	S_A/f_A	$\frac{S_A/f_A}{S_E/f_E}$
reziduální	S_E	$f_E = n-r$	S_E/f_E	-
celkový	S_T	$f_T = n-1$	-	-

1.1.4 Post-hoc metody mnohonásobného porovnávání

– zamítneme-li nulovou hypotézu o shodě středních hodnot, chceme zjistit, která dvojice středních hodnot se významně liší na hladině významnosti α

– 2 mětody: *Tukeyova*, *Schefféhо*

– Tukeyova metoda

* používá se, mají-li všechny výběry týž rozsah (tento rozsah značíme p)
* rovnost středních hodnot $\mu_l = \mu_k$ zamítneme na hladině významnosti α, pokud
$$|M_{k.} - M_{l.}| \geq q_{1-\alpha}(r, n-r)\frac{S_*}{\sqrt{p}}, \tag{2}$$
kde kvantily $q_{1-\alpha}$ najdeme ve statistických tabulkách a S_* je z minulé hodiny známý vážený průměr výběrových rozptylů. Lze jej ale zjednodušeně vypočítat podle vzorce $S_*^2 = \frac{S_E}{f_E}$.

    * existuje i modifikace Tukeyovy metody pro nestejné rozsahy výběrů tzv. *Tukey HSD metoda*

  - Scheffého metoda
    * používá se, pokud nejsou rozsahy všech výběrů stejné
    * rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq S_* \sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(r-1, n-r)}. \tag{3}$$

    * $S_*^2 = \frac{S_E}{f_E}$
    * metody mnohonásobného porovnávání jsou slabší, než ANOVA, proto se může stát, že ANOVOU zamítneme H_0 o shodě středních hodnot ale metody mnohonásobného porovnávání u žádné dvojice významný rozdíl nenajsou.
    * dochází tomu tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti

- POSTUP TESTOVÁNÍ ANOVY:
  1. ověření normality
     - Q-Q plot + test (Shapiro, Lillie, Ad)
     - slabé porušení nevadí, anova na to není příliš citlivá
  2. ověření rozptylu
     - krabicový graf - je šířka krabic stejná?; + test (Levenův, Bartlettův)
     - na slabé porušení homogenity rozptylu není anova příliš citlivá
  3. testování shody středních hodnot
  4. dojde-li k zamítnutí H_0 o shodě středních hodnot, použijeme *post-hoc metody* (Tukeyova, Scheffého)

- Zajímavost k testování homogenity rozptylů:
  Parametr σ^2 není znám a je třeba testovat hypotézu $H_0 : \mu_1 = \cdots = \mu_r$. Na první pohled by se zdálo, že tento problém lze snadno převést na testování dvou nezávislých výběrů, a to tak, že vytvoříme dvojice souborů a na každou dvojici aplikujeme dvouvýběrový t-test na hladině významnosti α. Jestliže alespoň jedna dvojice dá signifikantní výsledek (tedy zamítáme hypotézu o shodnosti středních hodnot vybrané dvojice), zdá se, že můžeme zamítnout hypotézu H_0. A současně hned vidíme, které dvojice se od sebe signifikantně liší. Tento postup však nesplňuje podmínku, že pravděpodobnost chyby prvního druhu má být α. Je-li totiž nulová hypotéza správná, pak každý t-test dá signifikantní výsledek, tj. zamítne hypotézu o shodě středních hodnot, s pravděpodobností α. My však chceme H_0 zamítnout, když alespoň jeden ze všech testů dá signifikantní výsledek. Takže pravděpodobnost zamítnutí H_0, je-li správná, bude při $I \geq 3$ větší než α.