11 Analýza rozptylu jednoduchého třídění, ANOVA, Jednofaktorová analýza rozptylu

Testování normality

- Normalita = první předpoklad k provedení ANOVY
- testy normality
  - Shapirův-Wilkův ... shapiro.test()
  - Lillie-Forsův ... lillie.test()
  - Anderson-Darlingův test ... ad.test(),

Testování homogenity rozptylů u r náhodných výběrů

- homogenita = stejnorodost rozptylů
- druhý předpoklad k provedení ANOVY
- máme $r \geq 2$ náhodných výběrů
- testujeme $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$ oproti H_1 : alespoň jedna dvojice rozptylů se liší

  1. Levenův test
     - testovací statistika založena na odhadech středních hodnot
     - levene.test(y, group, location='mean') knihovna lawstat
       * y ... vektor dat
       * group ... typ skupiny
       * location='mean'
  2. Brownův-Forsytův test
     - je modifikací Levenova testu
     - testovací statistika založena na mediánech
     - → při větších rozsazích náhodných výběrů ($n_i > 20$) jej lze použít i na data, která nejsou z normálního rozdělení
     - levene.test(y, group, location='median') z knihovny lawstat
       * y ... vektor dat
       * group ... typ skupiny
       * location='median'
  3. Bartlettův test
     - bartlett.test(y, g) knihovna stat
     - používáme, pouze pokud jsou rozsahy všech výběrů ≥ 6

ANOVA - Jednofaktorová analýza rozptylu

- zkoumá závislost intervalové proměnné X na nominální proměnné A
- $A\ldots$ *faktor*; varianty $A\ldots$ *úrovně faktoru*
- motivační příklady
  - má metoda výuky (A) vliv na počet bodů (X) v závěrečném testu?
  - má typ potravy pračlověka (A) vliv na šířku stoliček (X)?
- trocha matematiky
  - předpokládáme, že faktor A má $r \geq 2$ úrovní $A_1, \ldots A_r$, přičemž i-té úrovni odpovídá n_i pozorování $X_{i1}, \ldots X_{in_i}$. Každý výběr $A_i \sim N(\mu_i, \sigma^2)$.
  - Celkový počet pozorování je $n = \sum_{i=1}^{r} n_i$.
  - 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
    * 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
      - 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
      - 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
- počet stupňů volnosti: $f_E = n - r$

∗ $S_T = S_A + S_E$.

Testování hypotéz o shodě středních hodnot

- $H_0 : \mu_1 = \cdots = \mu_r$; střední hodnoty všech výběrů jsou stejné
- $H_1 : \mu_i \neq \mu_j$ pro nějaké i, j; alespoň jedna dvojice středních hodnot se liší.
- Testovací statistika má tvar

$$F_A = \frac{S_A/f_A}{S_E/f_E} \sim F(r-1, n-r).$$

- $F_A \in \langle F_{1-\alpha}(r-1, n-r), \infty) \to H_0$ zamítáme na hl. význ. α
- nebo: p-hodnota$< \alpha \to H_0$ zamítáme na hl. význ. α
- přehledná tabulka výpočtů:

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	$f_T = n - 1$	-	-

Post-hoc metody mnohonásobného porovnávání

- zamítneme-li nulovou hypotézu o shodě středních hodnot, chceme zjistit, která dvojice středních hodnot se od sebe významně liší
- Scheffého metoda
  - vhodná i v případě, že rozsahy všech výběrů nejsou stejné
  - rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

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

  - $S_*^2 = \frac{S_E}{f_E}$
  - funkce `Scheffe(X, group, names, alpha)` z RSkriptu `AS-funkce.R`.
  - 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 nenajdou.

- POSTUP TESTOVÁNÍ ANOVY:
  1. ověření normality
  2. ověření rozptylu
  3. testování shody středních hodnot
  4. dojde-li k zamítnutí H_0 o shodě středních hodnot, použijeme *post-hoc metody*
- Poznámka: Mírné porušení normality nebo shody rozptylů ANOVĚ zas tak moc nevadí