# 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 podle 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 $A_1, \ldots A_r$, přičemž $i$-té úrovni odpovídá $n_i$ pozorování $X_{i1}, \ldots X_{in_i}$. Tato pozorování tvoří náhodný výběr z $N(\mu_i, \sigma^2)$,$i = 1, \ldots, r$. 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 $\alpha$ 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 tstředních hodnot se liší.*
- vliv faktoru $A$ není významný, oproti alternativě, že vliv faktoru $A$ je významný.
- Testová 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 $\alpha$, pokud $F_A \in \langle F_{1-\alpha}(r-1, n-r), \infty)$
- případně $H_0$ zamítáme na hladině významnosti $\alpha$, pokud $p$-hodnota$< \alpha$.
- výsledky výpočtů lze zapsat do přehledné tabulky - na hodině jsme si jí neuváděli

| 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 $\alpha$

– 2 mětody: *Tukeyova*, *Scheffého*

– Tukeyova metoda

* používá se, mají-li všechny výběry týž rozsah $p$
* rovnost středních hodnot $\mu_l = \mu_k$ zamítneme na hladině významnosti $\alpha$, 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 $\mu_k$ a $\mu_l$ zamítneme na hladině významnosti $\alpha$, když

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

        * $S_*^2 = \frac{S_E}{f_E}$
        * metody mnohonásobného porovnávání jsou slabší, než ANOVA, proto se může stát, že ANOVOU zamítneme $H0$ 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
        - slabé porušení nevadí, anova na to není příliš citlivá
    2. ověření rozptylu
        - krabicový graf - je šířka krabic stejná?; + test
        - 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*

- Zajímavost k testování homogenity rozptylů:
  Parametr $\sigma^2$ 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 $\alpha$. 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 $\alpha$. Je-li totiž nulová hypotéza správná, pak každý t-test dá signifikantní výsledek, tj. zamítne hypotézu o shodě středních hodnot, s pravděpodobností $\alpha$. 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 $\alpha$.

**Příklad z hodiny:**
Ústav antropologie vypsal konkurz na přijetí nového antropologa do svých řad. Ředitel ústavu se rozhodl, že nedá na hezký obličejík a naučené fráze a vezme někoho, kdo je ve svém oboru zručný. Každý uchazeč měl za úkol provést v rámci pohovoru několik měření a byl mu stopován čas potřebný k měření. Konkurzu se zúčastnili tři kandidáti. Časy jejich měření v minutách jsou zaznamenány v tabulce:

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.antropolog: | 3.6 | 3.8 | 3.7 | 3.5 | | |
| 2.antropolog: | 4.3 | 3.9 | 4.2 | 3.9 | 4.4 | 4.7 |
| 3.antropolog: | 4.2 | 4.5 | 4.0 | 4.1 | 4.5 | 4.4 |

Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že rychlost měření těchto tří antropologů jsou stejné. Zamítnete-li nulovou hypotézu, určete, výkony kterých antropologů se liší na dané hladině významnosti $\alpha = 0.05$ a stanovte závěr, který by ředitele ústavu mohl zajímat.

*Poznámka:* Před samotným testováním **nezapomeňte ověřit, že všechny tři výběry pochází z normálních rozložení a že rozptyly těchto výběrů jsou shodné**. Jsou to důležité předpoklady, které musí být splněny, abychom mohli analýzu rozptylu použít. Normalitu otestujte pomocí vhodného testu (případně i graficky pomocí Q-Q plotu), shodu rozptylů potom ověřte pomocí Levenova testu a graficky pomocí krabicových diagramů. Proč nemůžeme k otestování shody rozptylů použít Bartlettův test?