

PSY117/454

Statistická analýza dat v psychologii II

**Seminář 7 - 8**

---

# **Analýza rozptylu**

**Srovnávání více než dvou průměrů**

# Omezení $t$ -testu

---

$t$ -test umožňuje srovnání pouze dvou průměrů

- Více skupin ( $j$ )  $\gg$  mnoho porovnání:  $j(j-1)/2$

Více srovnání způsobuje strmý růst pravděpodobnosti chyby I. typu

- např. při  $\alpha=0,05$  a 20 testech  $p=0,64$  (1 nebo více chyb)
  - aplikace binomického rozložení
- Platí to pro jakýkoli statistický test (zejm. korelace)

Je *nevhodné* provádět velké množství testů na jedněch datech (cca  $>5$ )

- Zneužití se označuje jako rybaření v datech – capitalizing on chance
- Lze kompenzovat korekcí hladiny  $\alpha$  (Bonferroniho korekce), avšak za cenu značného snížení síly testu ( $1-\beta$ ).
  - Místo  $\alpha$  testujeme na hladině  $\alpha'=\alpha/N$ , kde  $N$  je počet prováděných testů.

# Řešení = Analýza rozptylu (ANOVA)

---

Testuje na více skupinách jen jednu hypotézu:

- Je někde mezi skupinovými průměry někde rozdíl?
  - Je rozdíl v životní spokojenosti adolescentů žijících s oběma rodiči (1), pouze matkou (2) a matkou a nevlastním otcem(3)?
  - $H_0: \mu_1 = \mu_2 = \mu_3$
- Je-li odpověď „**ano**“ ( $p < \alpha$ ), pak se můžeme podívat na jednotlivé rozdíly detailněji (**post-hoc testy**)
- Je-li odpověď „**ne**“ ( $p > \alpha$ ), pak bychom neměli (rybaření)

# 1. terminologická vložka - ANOVA

---

- ANOVA = ANalysis Of Variance = analýza rozptylu
    - i přes svůj název jde o srovnávání **průměrů**
  - ANOVA zjišťuje vztah mezi **kategoriální nezávislou a intervalovou závislou**.
    - kategoriální nezávislá = **faktor** (factor, „-way“)
    - hodnoty kategoriální nez. = **úrovně** (level, treatment)
  - Zjištěný rozdíl = efekt, účinek (effect)
-



# Princip ANOVY – $F$ -test

---

- Čím jsou si průměry podobnější, tím je rozptyl mezi skupinami nižší ( $MS_{\text{between}}$  se blíží 0)
- Čím nižší je rozptyl uvnitř skupin ( $MS_{\text{within}}$  se blíží 0), tím průkaznější se průměry mezi skupinami zdají být.
- Důležitý je **poměr těchto dvou odhadů rozptylu:**
$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$
- Čím vyšší je  $F$ -poměr, tím průkaznější jsou rozdíly mezi průměry (rozsah je 0 až  $\infty$  )
- $F$  -poměr má jako výběrová statistika  $F$  -rozložení

# Princip ANOVY – dělení rozptylu.

---

- Dělení variability (rozptylu) podle zdrojů **jako u lineární regrese**

$$X_{ij} = \mu + \alpha_j + e_{ij}$$

$$Y_i = a + b_1X_1 + b_2X_2 + \dots + b_{j-1}X_{j-1} + e_i$$

- $X_{ij}$  = skóre jedince ( $i$ -tý jedinec v  $j$ -té skupině)
- $\mu$  = průměr populace
- $\alpha$  = vliv příslušnosti ke skupině (vliv úrovně faktoru)
- $e_{ij}$  = chyba (vše, s čím nepočítáme, individuální prom.)

$$X_{ij} - m = (m - m_j) + (X_{ij} - m_j)$$

odchylka od celkového průměru = odchylka od skupinového průměru +  
odchylka skupinového průměru od celkového průměru

- ... odchylky umocněné na druhou = cesta k rozptylu

$$SS_{\text{Total}} = SS_{\text{Between (A, Model)}} + SS_{\text{Within (Error, R)}}$$

$$MS_{\text{Total}}; MS_{\text{Error}}; MS_A$$

---

# Velikost účinku (efektu)

---

- Podobně jako u regrese chceme vědět, jaká část rozptylu závislé je vysvětlená nezávislou
  - Ekvivalentem  $R^2$  je u anovy  $\eta^2$  (eta)
    - $\eta^2 = SS_{\text{Between}} / SS_{\text{Total}}$
    - Poněkud přesnější je  $\omega^2 = (SS_M - df_M MS_R) / (SS_T + MS_R)$
  - Pro konkrétní rozdíl průměrů  $d_{\text{Coh}} = m_1 - m_2 / \sqrt{MS_{\text{Within}}}$
  
  - Velikost účinku je vždy třeba uvádět
-



# Předpoklady použití ANOVY

---

- normální rozložení uvnitř skupin
  - při  $n_j > 30$  a  $n_1 = n_2 = \dots = n_j$  je ANOVA robustní
- stejné rozptyly uvnitř skupin:  
homoskedascita
  - do  $s_{\max}/s_{\min} < 3$  je ANOVA robustní, zvláště při  $n_1 = n_2 = \dots = n_j$
- nezávislost všech pozorování
  - při opakovaných měřeních je třeba použít ANOVU pro opakovaná měření

viz Hendl 343

---

# Dnešní program

---

- One-way ANOVA
    - kontrasty a post-hoc testy **ONEWAY**
  - Faktoriální (two-way, three-way...) ANOVA
    - interakce **UNIANOVA**
  - Analýza kovariance – ANCOVA
    - kontrola intervenující proměnné
  - ANOVA s více závislými - MANOVA
-

# Životní spokojenost a rodina

---

- Domníváme se, že kompletní rodina je základ životní spokojenosti.
    - $H_1: M_{\text{komplet}} > M_{\text{nekomplet}}$
  - Zajímá nás, zda se liší chybění otce a jeho nahrazení nevlastním otcem
    - $H_2: M_{\text{bez otce}} \neq M_{\text{nevlastní otec}}$
-

# Kontrasty

---

- I když můžeme srovnat všechny průměry se všemi ostatními, platíme za to velkou ztrátou síly
  - Řešením jsou předem plánovaná srovnání –  
**KONTRASTY**
  - Lze srovnat kterékoli 2 skupiny nebo skupiny skupin
    - např. 1. skupinu se průměrem všech ostatních, kontrolní skupinu se každou ze zbývajících skupin zvlášť
  - Realizuje se zvláštním kódováním
    - při platnosti nulové hypotézy je součet vážených průměrů 0
  - $H_1$ : 1. vs (2. a 3.) .....      -2   1   1
  - $H_2$ : 2. vs 3. ....                      0 -1   1
-

# Post-hoc testy (simultánní porovnávání)

---

- Po (a pouze po) prokázání „nějakých“ rozdílů mezi průměry obvykle chceme vědět, mezi kterými skupinami konkrétně rozdíly jsou: **post-hoc testy**
- Srovnáváme každou skupinu s každou způsobem, který nezpůsobí nárůst  $\alpha$ .
- Je-li důležité udržet  $\alpha$  pod kontrolou, je správnou volbou **Scheffe**ho test nebo **Tukey**ho HSD – volba pro *rybaření*
- Máte-li stejně velké skupiny (balanced design) - **REGWQ**
- Pokud to  $\alpha$  kritická a máte-li pár *kvazi*-hypotéz na mysli, pak je volbou **Student-Neuman-Keuls (S-N-K)**
- Extrémně „dajný“ a nepříliš vhodný pro více než 3 skupiny je **LSD** a proto se nedoporučuje.
- Při nesplnění homoscedascity – **Games-Howell**

# Faktoriální ANOVA

---

- více faktorů ... možnost **interakce** mezi nimi
- **fixed** vs. **random** faktory

Liší se výkonová motivace podle věku a pohlaví?

- INT: Jsou případné genderové rozdíly shodné v obou kohortách?

Liší se výkonová motivace mezi školami a podle pohlaví?

- INT: Liší se genderové rozdíly škola od školy?
-

# Analýza kovariance

---

Velká variabilita závislé může zastírat rozdíly.

Dokážeme-li část její variability vysvětlit nějakým prediktorem, můžeme hledat rozdíly pouze ve zbývající části rozptylu závislé.

- statistická kontrola – jako parciální korelace a regrese
  - Proměnnou, jejíž vliv chceme kontrolovat, vkládáme jako **kovariát**
-

# MANOVA

---

- Máme-li více závislých
  - Opatrně.
-



# Shrnutí

---

- ❑ ANOVA je pro situace s intervalovou závislou a více kategorickými nezávislými – porovnávání mnoha průměrů
  - ❑ Faktory mohou být **fixní** nebo **náhodné**
  - ❑ ANOVA je podobná regresi – pro interpretaci je dobré si vyžádat „**parametry**“, tj. regresní váhy
  - ❑ Lze testovat konkrétní hypotézy – **kontrasty**
  - ❑ Lze testovat všechny možné rozdíly průměrů – **post hoc**
  - ❑ Lze uvažovat o kombinovaném vlivu faktorů – **interakce**
  - ❑ Lze kontrolovat vliv intervenujících proměnných – **kovariáty - ANCOVA**
  - ❑ Lze mít i více závislých najednou – **MANOVA** - opatrně
-

