PSY117
Statistická analýza dat v psychologii
Přednáška 12 2018

Analýza rozptylu

Srovnávání více než dvou průměrů

If your experiment needs statistics, you ought to have done a better experiment.

Ernest Rutherford

Omezení *t*-testu (i jeho nPar alternativ)

t-test umožňuje srovnání pouze dvou průměrů

- Více skupin (j) >> mnoho porovnání: j (j -1)/2

Více srovnání způsobuje strmý růst pravděpodobnosti chyby I. typu

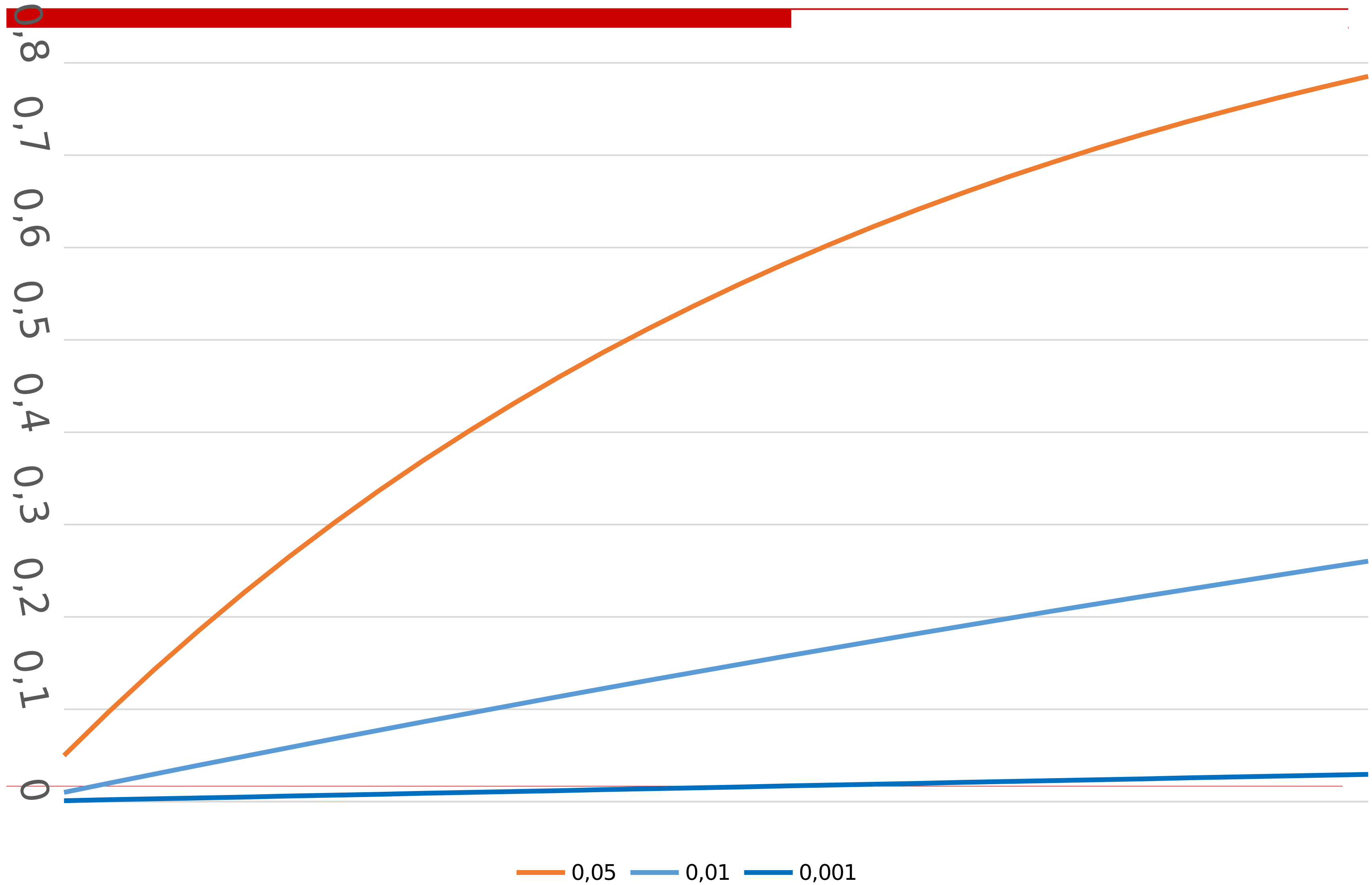
- např. při α=0,05 a 20 testech p=0,64 (1 nebo více chyb)
  - aplikace binomického rozložení
- Platí to pro jakékoli statistické testy (zejm. korelace)

Je ***problematické*** provádět mnoho testů na jedněch datech (cca >5)

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

AJ : multiple tests, capitalizing on chance, fishing, Bonferroni correction, statistical power

obnost 1 nebo více falešných pozitiv při dané hladině a a p



Řešení = Analýza rozptylu (ANOVA)

Testuje na více skupinách jen jednu hypotézu:

- Je mezi skupinovými průměry někde rozdíl?
  - Je mezi Pražáky, Brňáky a Ostraváky rozdíl v průměrné lakotě?
  - H_0: μPražáci = μBrňáci = μOstraváci
- 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)

AJ : ANalysis Of Variance, post-hoc tests (multiple comparisons)

Terminologická vložka - ANOVA

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

Opáčko

- Výběrové rozložení průměru má $s_m = s/\sqrt{N}$
- $s = s_m\sqrt{N}$ nebo $s^2 = s^2_m N$

- Když vybereme ze stejné populace 50 vzorků o 100 lidech a rozptyl těch 50 průměrů bude 25, jaká je směrodatná odchylka proměnné?

Platí-li H_0...

- Rozptyl každé skupiny je jedním nezávislým odhadem populačního rozptylu
  - Zprůměrováním těch odhadů se odhad ještě zpřesní
- Rozptyl vypočítaný z rozptylu skupinových průměrů by také měl být odhadem populačního rozptylu
- Tyto dva odhady by měly být stejné, až na výběrovou chybu
- Je-li rozptyl vypočítaný z rozptylu skupinových průměrů vyšší, asi se průměry skupin liší

Princip ANOVY 1.

- rozptyl = MS = mean square = SS/df = $(\sum(x-m))/(n-1)$
- MSwithin : variabilita uvnitř skupin (MSe, *error*)
  - MSwithin = SSwithin/$n - j$
  - SSwithin = $\sum_j \sum_i (x_i - m_j)2$
  - MSbetween : $s2$ spočítaný ze skupinových průměrů, variabilita uvnitř skupiny je ignorována (též MSA, B, treatment)
  - MSbetween = SSbetween/j -1
  - SSbetween = $\sum_j (n_j (m_j - m) 2)$

Platí-li $H0$, jaký čekáme vztah mezi Mbetween a Mwithin ?

Princip ANOVY – *F* -test

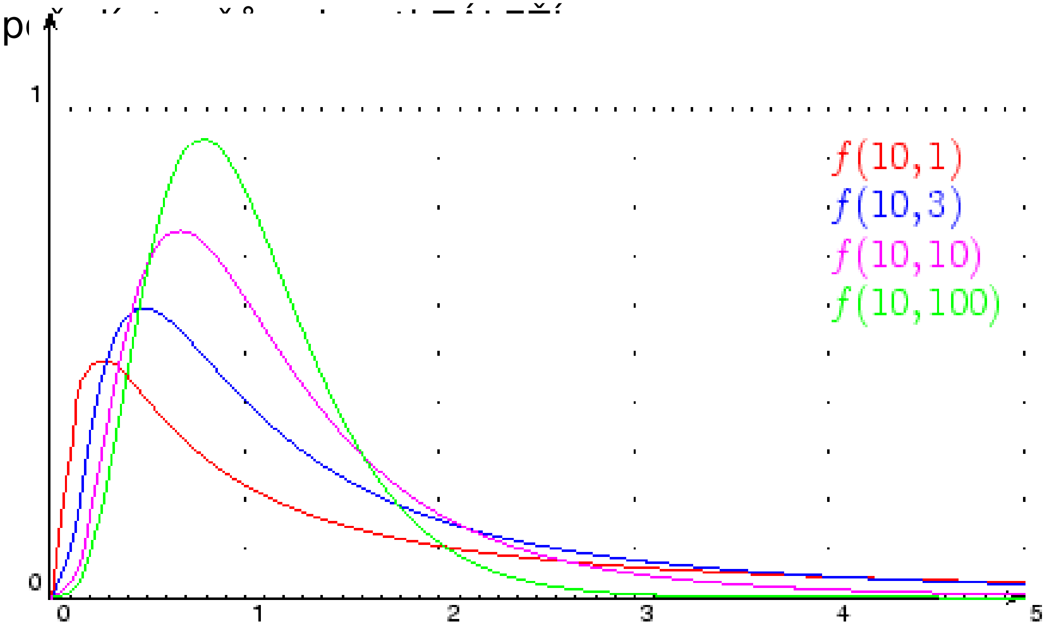
- Čím jsou si průměry podobnější, tím je rozptyl mezi skupinami nižší (Platí-li *H*0, *MS*between se blíží *s*2)
- Čím nižší je rozptyl uvnitř skupin (*MS*within se blíží 0), tím průkaznější se průměry mezi skupinami zdají být.
- Důležitý je **poměr těchto dvou odhadů rozptylu:**

$$F = \frac{MS_{between}}{MS_{within}}$$

- Čím vyšší je *F*-poměr, tím průkaznější jsou rozdíly mezi skupinovými průměry (rozsah je 0 až ∞)
- *F* -poměr má při platnosti *H*0 jako výběrová statistika ***F* –rozložení** s ***(df1,df2)***, které má průměr přibližně 1 (přesně df2/(df2-2))

Fisherovo-Snedecorovo *F*-rozložení

- Podobně jako ***t***-rozložení, je ***F***-rozložení vlastně rodina mnoha rozložení mírně se lišící svým tvarem $(F(1;\ \nu)= t(\nu)2)$
- Tato rozložení se liší tentokrát dvěma parametry – stupni volnosti
  - $\nu 1$ = ***počet skupin – 1*** : stupně volnosti čitatele - ***MS***between
  - $\nu 2$ = ***počet lidí – počet skupin*** : stupně volnosti jmenovatele - ***MS***within
  - na p



http://www.econtools.com/jevons/java/Graphics2D/FDist.html

Princip ANOVY – dělení rozptylu.

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

$$Y_i = a + bX_i + e_i$$

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

  - X_{ij} = skóre jedince (i-tý jedinec v j-té skupině)
  - μ = průměr populace
  - α = 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 Total = ***SS*** Between (A, treatment) + ***SS*** Within(Error)

MS Total; ***MS*** A; ***MS*** Error

Velikost účinku (efektu)

- Podobně jako u regrese chceme vědět, jaká část rozptylu závislé je vysvětlená nezávislou
- Ekvivalentem R^2 je u anovy η^2 (eta)
  - $\eta^2 = SS_{Between}/SS_{Total}$
  - Poněkud přesnější je $\omega^2 = (SS_{Between} - df_{Between} \cdot MS_{Within})/(SS_{Total} + MS_{Within})$
- Pro konkrétní rozdíl průměrů $d_{Coh} = m_1 - m_2/\sqrt{MS_{Within}}$

- Velikost účinku je vždy třeba uvádět

Předpoklady použití ANOVY

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

viz Hendl 343

Post-hoc testy (simultánní porovnávání)

- Po <u>(a pouze po)</u> 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 α.
- Je-li důležité udržet α pod kontrolou, pak je správnou volbou **Scheffe**ho test – volba pro ***rybaření***
- Pokud to není tak kritické 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.

Další varianty a rozšíření ANOVA

- ANOVA pro opakovaná měření (jako párový t-test)
- ANOVA s 2 a více faktory (faktoriální ANOVA)
- MANOVA – s více závislými proměnnými

To vše v SPSS skryto pod GLM – general linear model

- Pořadovou (neparametrickou) alternativou ANOVY jsou
  - **Kruskal-Wallis H:** $H_0: Md_1 = Md_2 = \ldots = Md_j$ $H_1: Md_1 \neq Md_2 \neq \ldots \neq Md_j$
  - **Jonckeheere-Terpstra Test:** $H_1: Md_1 \leq Md_2 \leq \ldots \leq Md_j$

AJ : repeated measures ANOVA, two(three..)-way ANOVA,(factorial ANOVA)