ANOVA

Peter Spáč

26.11.2020

ANOVA (ANalysis Of VAriance)

- Použití:
  - Měření závislosti kategorické (ne dichotomické) proměnné na kardinální proměnnou
  - Srovnání hodnot tří a více průměrů v rámci jedné proměnné

- Např. jak se liší průměrný příjem v závislosti na věku (věkových skupinách)

ANOVA - základy

- ANOVA testuje nulovou hypotézu, že průměry jednotlivých skupin jsou totožné
- Výsledkem je F-statistika:
  - Ta stanoví, zda jsou průměry totožné nebo ne
  - Nespecifikuje ale, jak se které průměry liší
- Identifikace odlišností mezi průměry se děje až v dalším kroku

ANOVA - základy

- Základní model, který se na data dá použít, je průměr
- Průměr vyjadřuje absenci efektu jiné proměnné (např. věku na příjem)
- Cílem je najít model, který naše data vystihuje lépe
- Pokud jsou rozdíly mezi skupinami dostatečně velké, bude model založený na více průměrech vhodnější

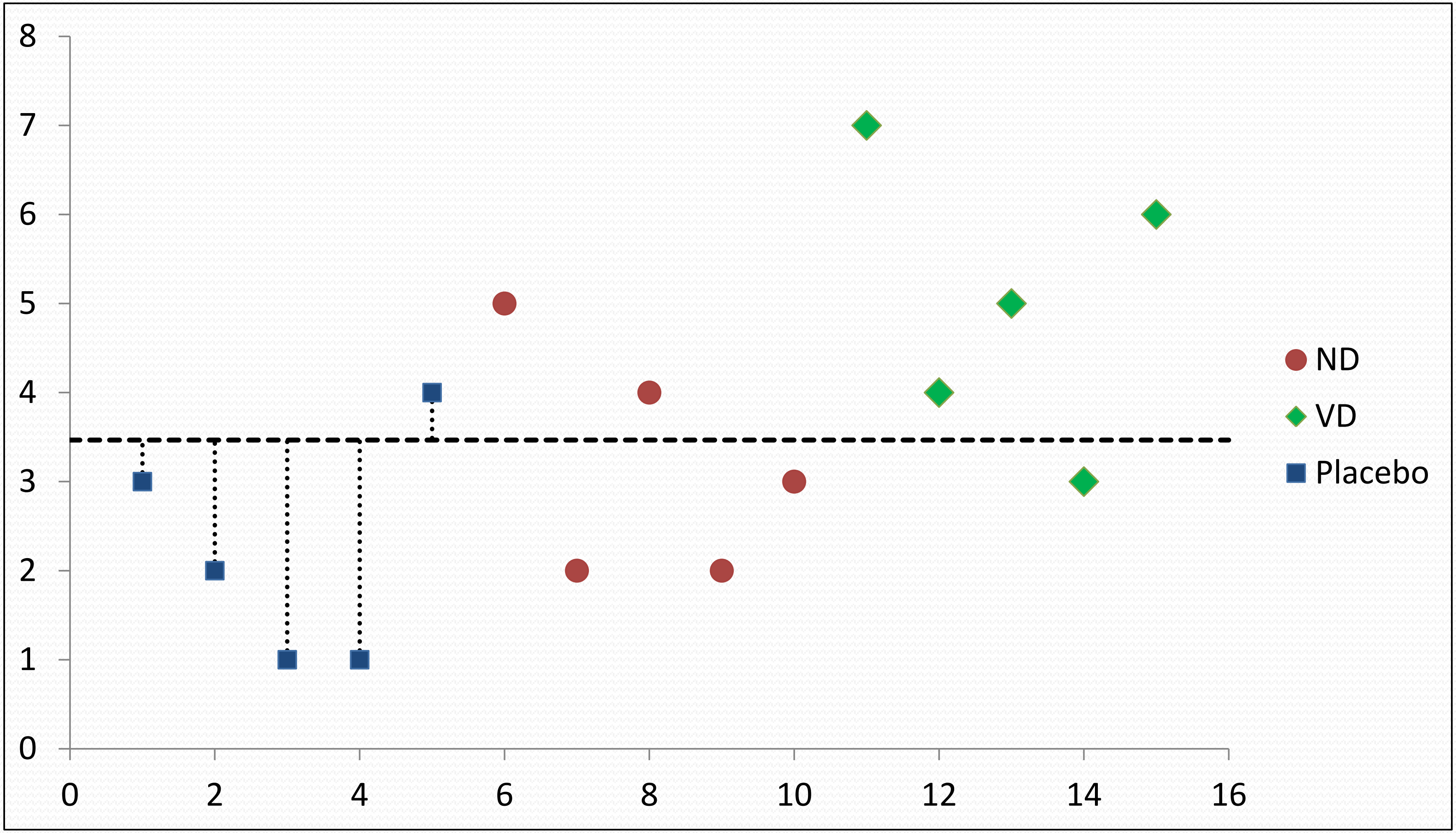
ANOVA - základy

- Jak zjistit, zda je nový model lepší?
- Odpověď – model musí představovat pokrok oproti vysvětlovací schopnosti starého modelu
- V případě průměru jsou vhodným ukazatelem jeho „nepřesnosti“ odchylky mezi modelem předpokládanými a skutečnými hodnotami

ANOVA – příklad (Field 2009: 350)

	Placebo	Nízká dávka	Vysoká dávka
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
Průměr	2,2	3,2	5
Celkový průměr	3,467		
Rozptyl	3,124		
Sm. odchylka	1,767		

ANOVA - základy

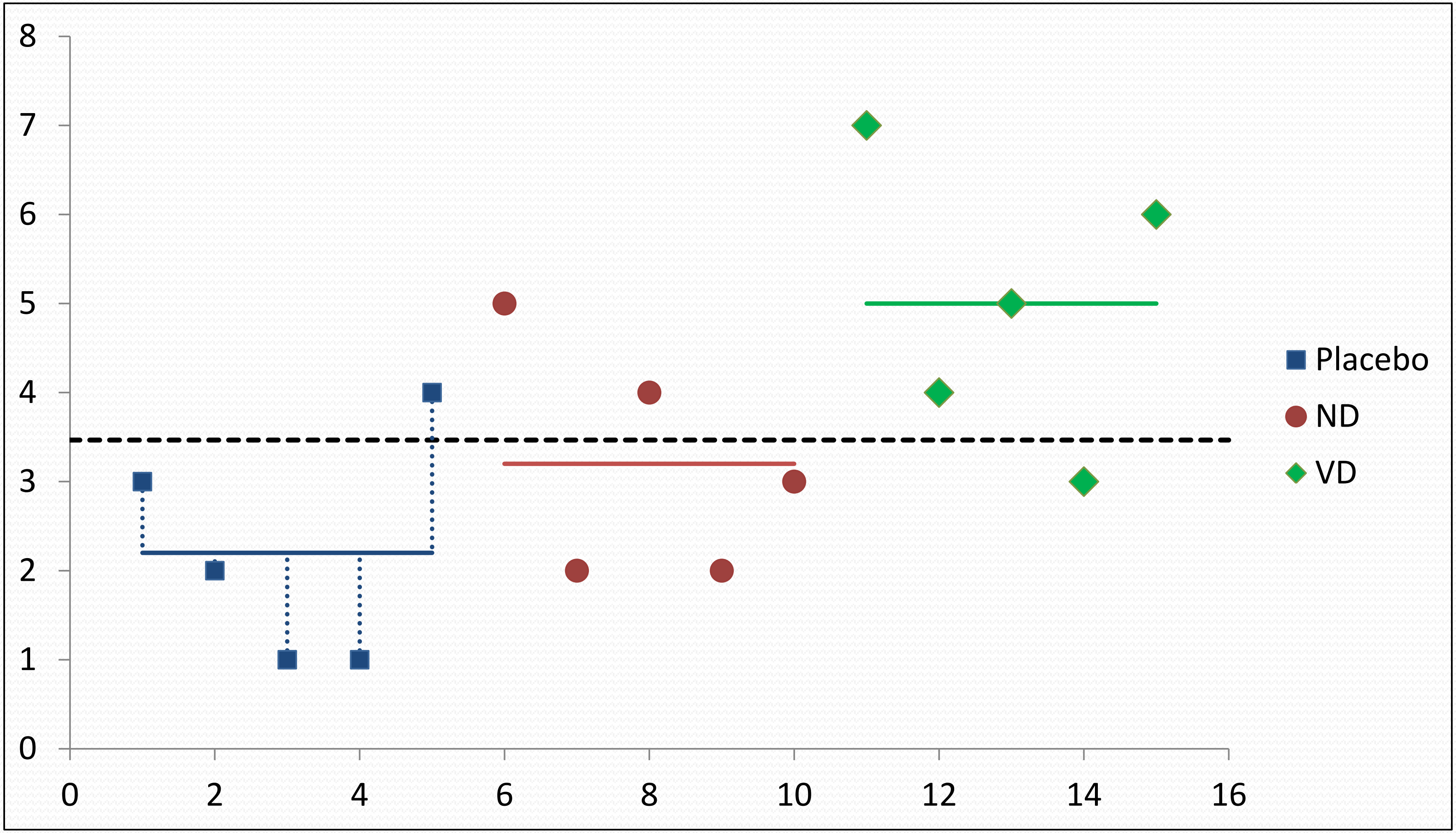


SS_T

- Celkový součet čtverců (Total Sum of Squares, SS_T)
  - Součet umocněných odchylek od celkového průměru
  - Čitatel zlomku výpočtu rozptylu
- $SS_T = s^2 (N - 1)$

Hodnoty	Průměr	Rozdíl	Po umocnění
3	3,467	-0,467	0,218089
2		-1,467	2,152089
1		-2,467	6,086089
1		-2,467	6,086089
4		0,533	0,284089
5		1,533	2,350089
2		-1,467	2,152089
4		0,533	0,284089
2		-1,467	2,152089
3		-0,467	0,218089
7		3,533	12,48209
4		0,533	0,284089
5		1,533	2,350089
3		-0,467	0,218089
6		2,533	6,416089
SS_T			**43,74**

ANOVA - základy



SS_R

- Součet čtverců reziduálů (Residual Sum of Squares, SS_R)
  - Součet umocněných odchylek od průměrů stanovených modelem
  - Vyjadřuje nepřesnost modelu (rozdíly, které model nedokáže vysvětlit)

$$SS_R = \sum (x_{ik} - \bar{x}_k)^2$$

$$SS_R = \sum s_k^2 (n_k - 1)$$

Hodnoty	Průměr skupiny	Rozdíl	Po umocnění
3	2,2	0,8	0,64
2		-0,2	0,04
1		-1,2	1,44
1		-1,2	1,44
4		1,8	3,24
5	3,2	1,8	3,24
2		-1,2	1,44
4		0,8	0,64
2		-1,2	1,44
3		-0,2	0,04
7	5	2	4
4		-1	1
5		0	0
3		-2	4
6		1	1
SS_R			**23,6**

SS_M

- Modelový součet čtverců (Model Sum of Squares, SS_M)
  - Součet umocněných rozdílů mezi hodnotami předpokládanými novým a starým modelem
  - Vyjadřuje pokrok nového modelu oproti modelu založeném na celkovém průměru

$$SS_M = \sum n_k \left(\bar{x}_k - \bar{x}_{\text{grand}}\right)^2$$

Průměr skupiny	Celkový průměr	Rozdíl	Po umocnění	Vynásobení velikostí skupiny
2,2	3,467	-1,267	1,605289	8,026445
3,2		-0,267	0,071289	0,356445
5		1,533	2,350089	11,750445
SS_M				**20,135**

Sumy čtverců

- SS_T – nepřesnost původního modelu
- SS_R – nepřesnost nového modelu
- SS_M – pokrok nového modelu oproti starému

- $SS_T = SS_R + SS_M$
- $43{,}74 = 23{,}6 + 20{,}135$

Sumy čtverců

- Význam pro nový model:
  - SS_M uvádí, kolik variability dat je model schopný vysvětlit (pokrok více průměrů oproti jednomu průměru)
  - SS_R naopak uvádí, co model není schopný vysvětlit (z důvodu vlivu dalších faktorů)

- Je potřebné, aby podíl vysvětlené variability byl vyšší než podíl variability nevysvětlené, a to čím víc, tím líp

Průměrné sumy čtverců

- SS_M = 20,135 / (3-1) = 20,135 / 2 = **10,068** = MS_M
- SS_R = 23,6 / (15 – 3) = 23,6 / 12 = **1,967** = MS_R

- Obě hodnoty je nutné srovnat na stejný základ, protože byli počítané jako součty z odlišného počtu prvků

- SS_M se dělí počtem skupin -1
- SS_R se dělí počtem prvků – počtem skupin

F-statistika

- $SS_M = 20{,}135 / (3-1) = 20{,}135 / 2 =$ **10,068** $= MS_M$
- $SS_R = 23{,}6 / (15 - 3) = 23{,}6 / 12 =$ **1,967** $= MS_R$

- F = vysvětlená variabilita / nevysvětlená variabilita
- $F = MS_M / MS_R$
- F = **5,12**

F-statistika

- Výstup analýzy ANOVA
- F-statistika (a její signifikantnost) jsou pouze prvním krokem (i když samotná ANOVA tím končí)
- Z F-statistiky lze poznat, že některé průměry se od sebe statisticky signifikantně liší, ale ne už které a jak
- Potřebný druhý krok – kontrasty nebo post hoc testy

ANOVA - předpoklady

- ANOVA je parametrický test
- Nezávislost pozorování, normální rozložení závislé proměnné (uvnitř skupin), homogenita rozptylu, závislá proměnná alespoň intervalová
- Za jistých okolností je ANOVA robustní = produkuje platné výsledky navzdory porušeným předpokladům

ANOVA - předpoklady

- **Porušení normality:**
  - Pokud jsou skupiny stejné, výsledky ANOVA by neměli být narušené
  - Pokud jsou skupiny různě velké, přesnost F-statistiky může být narušená
- **Porušení homogenity rozptylu:**
  - Stejně jako u porušení normality
  - Pokud mají větší skupiny vyšší rozptyl, hodnota F má tendenci být nižší (a naopak)
- **Porušení nezávislosti:**
  - Vážné navýšení pravděpodobnosti chyby I. typu

Post hoc testy

- Druhý krok, který následuje po zjištění hodnoty F-statistiky (pouze pokud ukazuje na výhodnost modelu)
- Post hoc testy porovnají všechny dvojice průměrů
- Využití spíše pro výzkumy bez hypotéz (není pravidlo)
- Více variant (v SPSS téměř dvě desítky)

Post hoc testy

- Kritéria použití:
  - Kontrola chyb I. typu
  - Kontrola chyb II. typu
  - Validní výstupy při porušení předpokladů ANOVA

- Konzervativní testy – nízká možnost chyby I. typu za cenu opatrnosti (neodhalí existující efekt)
- Liberální testy – nízká možnost chyby II. typu za cenu lehkovážnosti (odhalí se neexistující efekt)

Post hoc testy

- Co použít?
- Stejně velké skupiny a rozptyly – REGWQ nebo Tukey
- Konzervativní test – Bonferroni
- Rozdílná velikost skupin – Gabriel nebo GT2
- Narušena homogenita rozptylu – Games-Howell

ANOVA v SPSS

- Analyze → Compare Means → One-Way ANOVA
  - Závislou proměnnou vložit do *Dependent List*
  - Nezávislou proměnnou do *Factor*

- V *Options* možnost zvolit deskriptivní statistiky, Levenův test, Brown-Forsythe a Welch F

- V *Post Hoc* vybrat příslušné testy (při *Dunnett* skontrolovat další nastavení)

Test of Homogeneity of Variances

Libido

Levene Statistic	df1	df2	Sig.
,092	2	12	,913

ANOVA

Libido

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20,133	2	10,067	5,119	,025
Within Groups	23,600	12	1,967		
Total	43,733	14			

Robust Tests of Equality of Means

Libido

	Statistic[a]	df1	df2	Sig.
Welch	4,320	2	7,943	,054
Brown-Forsythe	5,119	2	11,574	,026

Multiple Comparisons

Dependent Variable: Libido

	(I) Dose of Viagra	(J) Dose of Viagra	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Placebo	Low Dose	-1,000	,887	,516	-3,37	1,37
		High Dose	-2,800*	,887	,021	-5,17	-,43
	Low Dose	Placebo	1,000	,887	,516	-1,37	3,37
		High Dose	-1,800	,887	,147	-4,17	,57
	High Dose	Placebo	2,800*	,887	,021	,43	5,17
		Low Dose	1,800	,887	,147	-,57	4,17
Games-Howell	Placebo	Low Dose	-1,000	,825	,479	-3,36	1,36
		High Dose	-2,800*	,917	,039	-5,44	-,16
	Low Dose	Placebo	1,000	,825	,479	-1,36	3,36
		High Dose	-1,800	,917	,185	-4,44	,84
	High Dose	Placebo	2,800*	,917	,039	,16	5,44
		Low Dose	1,800	,917	,185	-,84	4,44
Dunnett t (>control)[b]	Low Dose	Placebo	1,000	,887	,227	-,87	
	High Dose	Placebo	2,800*	,887	,008	,93	

Kruskal-Wallisův test

- Neparametrická alternativa k ANOVA
- Data seřadí a následně počítá (samotné hodnoty v rámci výpočtu nebere do úvahy)
- Výsledkem je statistika H
- Následně je možná obdoba post hoc testů (Mann-Whitney test) – ani zde se nebere ohled na hodnoty

Kruskal-Wallis v SPSS

- Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples
  - Zvolit Kruskal-Wallis H
  - Závislou proměnnou vložit do *Test Variable List*
  - Nezávislou proměnnou do *Grouping Variable* a stanovit minimální a maximální hodnotu

- Pro *Post Hoc*:
- Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples
  - Zvolit Mann-Whitney U
  - Stejný postup