ANOVA

Seminář 5

Původ dat, se kterými budeme pracovat

- Data **bicyclists.sav** pocházejí se studie, která se zabývala různými prediktory toho, v jaké vzdálenosti řidiči motorových vozidel předjíždějí cyklisty:
- Walker, I. (2007). Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender. *Accident Analysis & Prevention*, *39*(2), 417–425. https://doi.org/10.1016/j.aap.2006.08.010
- Autor sám najezdil na kole 320 km v různých britských městech.
  Jezdil v různé denní době od 7:00 do 18:00 a po různých silnicích.
- Zkoušel jezdit s helmou/bez helmy a v různé vzdálenosti od kraje silnice.
- Zaznamenával, v jaké vzdálenosti jej předjíždějí motorová vozidla (pomocí ultrazvukového snímače) a o jaké vozidlo se jednalo.

Stručný popis dat

Proměnná	Popis
id	Prostý numerický identifikátor události (předjetí)
vehicle	Typ vozidla, které cyklistu předjelo
color	Barva vozidla, které cyklistu předjelo
pass_distance	V jaké vzdálenosti cyklistu vozidlo předjelo / kolik místa mu nechalo (v metrech)
street	Typ silnice, po které cyklista jel na kole, když jej vozidlo předjelo.
time	Denní čas v hodinách + minutách, kdy došlo k předjetí.
hour	Denní čas (pouze hodina), kdy došlo k předjetí
helmet	Měl v době předjetí cyklista nasazenu cyklistickou přilbu?
kerb	V jaké vzdálenosti od kraje vozovky cylista na kole jel, když byl předjet.
bikeline	Jel v daném momentě cyklista v pruhu pro cyklisty?
city	Město, kde cyklista jel (Salisbury, Bristol, Porthmouth).
date	Datum

Zadání analýzy

- Nejprve si transformujeme závislou proměnnou tak, aby vyjadřovala "bezpečnost předjetí".
- Autor předpokládá, že bezpečnost předjetí závisí na vzdálenosti předjetí (VP) nelineárně: $bezpecnost = \ln(pass\_distance)$.
- Vysvětlení: Kdybyste jeli na kole a minuly Vás dvě auta, jedno ve vzdálenosti 10 cm a druhé ve vzdálenosti 50 cm, zřejmě byste cítili větší rozdíl v "bezpečnosti předjetí" od těchto vozidel, než kdyby Vás jedno vozidlo předjelo ve vzdálenosti 110 cm a druhé ve vzdálenosti 150 cm.

Analýza 1

- Pomocí one-way ANOVA nejprve otestuje H., že jsou rozdíly v bezpečnosti předjetí mezi typy vozidel, která cyklistu předjela (proměnná *vehicle*).
- Pomocí ortogonálních kontrastů pak ověřte:
  - Zda profesionální řidiči míjí cyklistu v bezpečnější vzdálenosti než ostatní řidiči. Předpokládejme, že profesionálové řídili tyto typy vozů: Large Goods Vehicle, Bus, Heavy Goods Vehicle, Taxi.
  - Zda řidiči jednostopých motorových vozidel (Powered Two Wheeler) míjí cyklistu v bezpečnější vzdálenosti než řidiči osobních aut (Ordinary Car, SUV/Pickup).
  - Zda profesionální řidiči ve větších vozidlech (Large Goods Vehicle, Bus, Heavy Goods Vehicle) objíždějí cyklistu v méně bezpečné vzdálenosti než řidiči taxíků.

Analýza 2

- Pomocí faktoriální ANOVA ověřte, zda řidiči míjí cyklistu v *méně* bezpečné vzdálenosti,
  - když má cyklista **helmu** (*helmet,* předpokládáme, že když má cyklista helmu, řidiči nemají tak silný pocit, že jej ohrozí těsnějším předjetím) a
  - když jede dále od okraje silnice (*kerb*, tj. blíže středu silnice).
- Počítejte s interakcí mezi nošením helmy a vzdáleností cyklisty od okraje silnice. Efekt nošení helmy by se měl projevit především tehdy, když cyklista jede blíže okraje silnice (protože když jede blíže středu silnice, řidiči nemají takový výběr v tom, v jaké vzdálenosti ho předjedou).
- Do modelu zahrňte také hlavní efekt typu vozidla (ale nikoli interakce typu vozidla s ostatními prediktory).

Opakování

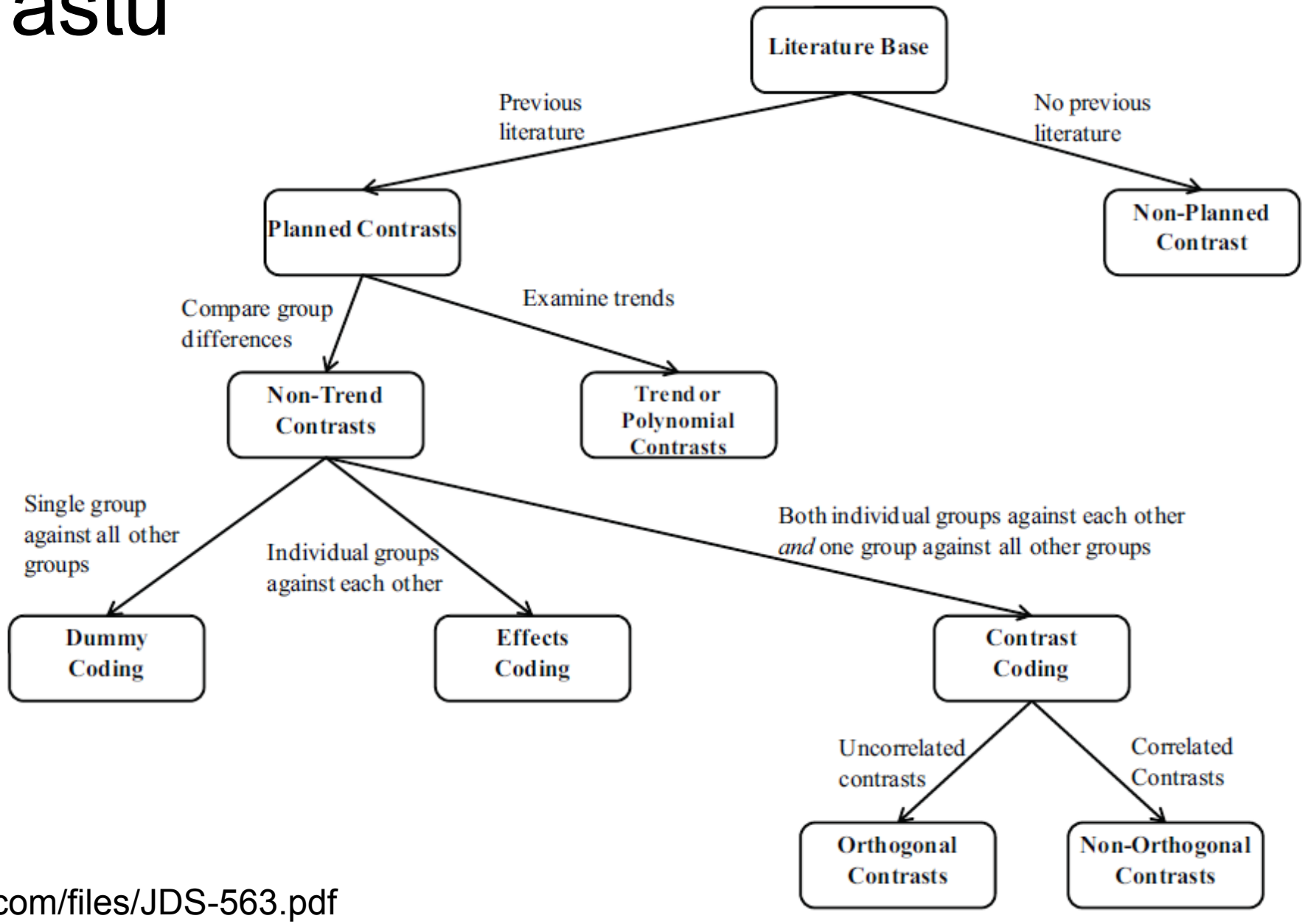
One-way ANOVA

- One-way ANOVA používáme, pokud chceme ověřit rozdíly mezi průměry dvou či více nezávislých skupin (tj. každý respondent patří do jedné a právě jedné skupiny, nepracujeme s opakovanými či párovými měřeními): např. mezi experimentálními podmínkami (je-li každý účastník vystaven pouze jedné z nich), mezi osobami různé úrovně vzdělání, mezi osobami s různými, ale vzájemně výlučnými diagnózami apod.
- Obvykle se používá, když máme více než dvě skupiny, protože v případě dvou skupin můžeme použít "obyčejný" nezávislý t-test. Ale je to spíše uzuální, protože čtverec t rozdělení s df stupni volnosti (t^2_{df}) odpovídá rozdělení $F(1; df)$
- Jedná se o "celkový" test toho, zda se průměry skupin liší, nulová hypotéza tedy je:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_n$$

- Neřekne nám, které ze skupin se signifikantně liší, od toho jsou zde kontrasty.

Typy kontrastů



http://www.jds-online.com/files/JDS-563.pdf

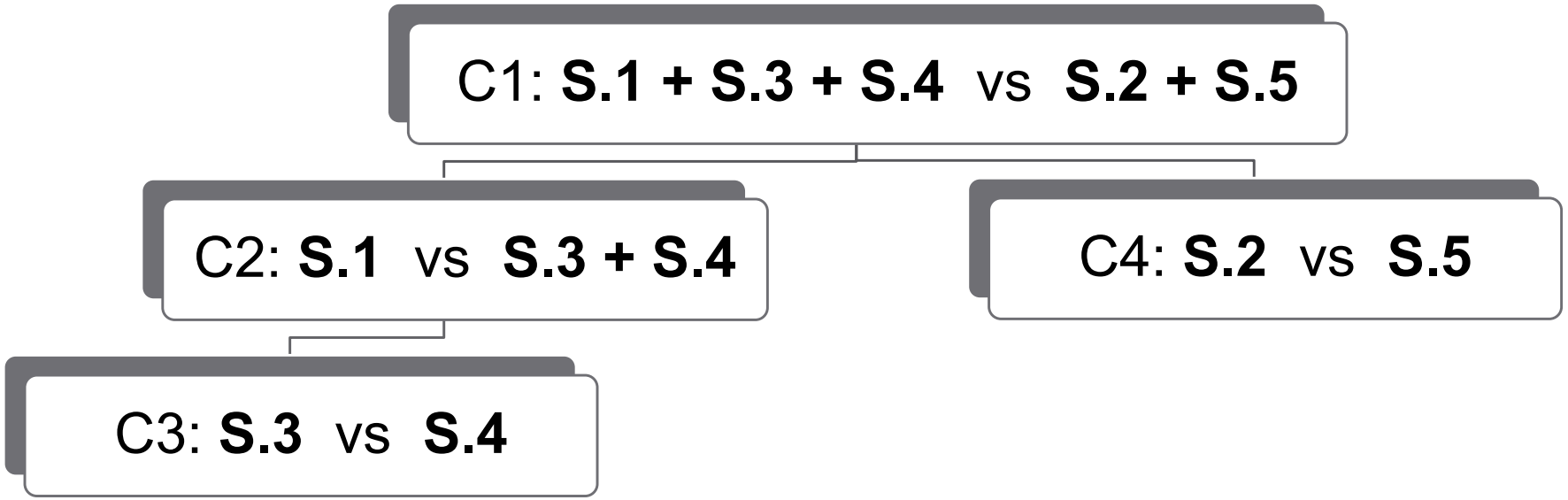
My se zaměříme na tzv. **ortogonální kontrasty**

Pravidla pro jejich vytváření:

1. Dopředu promyslete, jak chcete skupiny srovnávat.
2. Vytvořte si stromový diagram plánovaných kontrastů a nezapomeňte, že každá skupina by se měla objevit samostatně pouze v jednom kontrastu.
3. Maximální počet kontrastů = 1 – počet srovnávaných skupin.
4. Skupiny (resp. skupina), kterým přiřadíte v daném kontrastu pozitivní váhy (koeficienty), se porovnají se skupinami, kterým přiřadíte negativní váhy.
5. Pro každý jednotlivý kontrast by se součet vah měl rovnat nule.
6. Skupině, kterou nechcete do kontrastu zahrnovat, dejte váhu = 0.
7. V rámci jednoho kontrastu volte takové váhy, aby se pozitivní váhy nasčítaly do 1 a negativní váhy do -1 (výhodné v rámci ANOVa – hodnota kontrastu je pak rovna rozdílu srovnávaných průměrů.)

   V rámci jednoho kontrastu volte takové váhy, aby hodnoty pozitivních vah odpovídaly počtu skupin s negativními váhami a (obráceně) aby absolutní hodnota negativních vah odpovídala počtu skupin s pozitivními váhami (výhodné, když chceme kontrasty použít v regresi – hodnota kontrastu vynásobená počtem srovnávaných skupin je pak rovna rozdílu srovnávaných průměrů).

Příklad ortogonálních kontrastů



Kontrast	Skupina 1	Skupina 2	Skupina 3	Skupina 4	Skupina 5	Součet vah
C1	+2	–3	+2	+2	–3	0
C2	+2	0	–1	–1	0	0
C3	0	0	+1	–1	0	0
C4	0	+1	0	0	–1	0
Součin sloupce	0	0	0	0	0	0

Faktoriální ANOVA

- Podobně jako v případě one-way ANOVA ji používáme k testování rozdílů mezi průměry skupin.
- Na rozdíl od one-way ANOVA ale pracujeme s více než jednou nezávislou kategorickou proměnnou ("faktory") a více nás zajímá jejich interakce: můžeme např. očekávat větší rozdíly mezi muži a ženami při nižší úrovni vzdělání.
- Příklady: Chceme zjistit, jaký efekt má pohlaví a úroveň vzdělání na zájem o politické dění. Chceme zjistit, jaký efekt má typ pracovní pozice a věk (rozdělený do několika kategorií) na míru stresu. Chceme zjistit, jaký efekt má pohlaví, typ citové vazby a partnerský status na well-being atd.

Předpoklady mezisubjektové ANOVA

1. **Závislá proměnná je měřena minimálně na intervalové úrovni**.
2. **Nezávislá proměnná (proměnné) je kategorické povahy**.
3. **Dostatečný počet případů**. Měli bychom zkontrolovat četnosti v jednotlivých buňkách.
4. **Nezávislost pozorování**. Každý respondent byl měřen jen jednou a členství v kategoriích každé nezávislé proměnné je vzájemně výlučné.
5. **Absence silné multikolinearity (v důsledku nevyváženého designu)**. Stejně jako v případě předchozích analýz platí, že když spolu naše nezávislé proměnné (prediktory) souvisejí příliš silně, je obtížné identifikovat jejich jedinečný efekt (např. kdybychom analyzovali efekt pohlaví a vzdělání, ale naprostá většina žen z našeho vzorku by měla vysokou školu, zatímco naprostá většina mužů by měla pouze střední školu bez maturity). Zvolený typ součtu čtverců pak může zásadně ovlivnit odhad efektů jednotlivých prediktorů.
6. V rámci každé skupiny (v případě faktoriální ANOVA se tím myslí v rámci každé kombinace úrovní všech nezávislých proměnných) ověřujeme:
   a) **Normalitu rozdělení závislé proměnné**.
   b) **Absenci odlehlých případů**.
   c) **Shodu (homogenitu) rozptylů** (Welchova ANOVA tento předpoklad nevyžaduje).

Velikosti účinku

- Pro one-way ANOVA uvádíme η^2 nebo ω^2.
- η^2 je analogií R^2 a ω^2 je analogií adjustovaného R^2 z lineární regrese. Jedná se tedy o podíl rozptylu závislé proměnné vysvětlený nezávislou proměnnou, přičemž ω^2 zahrnuje korekci velikosti vzorku, protože η^2 má tendenci nadhodnocovat podíl rozptylu, který by model vysvětloval v populaci, zejm. u malých vzorků (čím větší máme vzorek, tím menší rozdíl bude mezi η^2 a ω^2, resp. mezi R^2 a adj. R^2).
- Pravidla palce: hodnoty η^2 či ω^2 0,01, 0,06 a 0,14 odpovídají malému, střednímu a silnému účinku.
- Rovnice pro výpočet (*SS* jsou součty čtverců, *df* jsou stupně volnosti a *MS* střední čtverec):

$$\eta^2 = SS_{effect}/SS_{total}$$

$$\omega^2 = \frac{SS_{effect} - df_{effect} \times MS_{error}}{SS_{total} + MS_{error}}$$

Velikosti účinku: kontrasty a srovnání dvou konkrétních skupin

- Pro kontrasty či srovnání konkrétních skupin obvykle uvádíme Cohenovo d, které vyjadřuje "standardizovaný" rozdíl mezi průměry skupin. Standardizováním zde máme na mysli to, že tento rozdíl dělíme směrodatnou odchylkou.
- Pokud se rozptyly v rámci skupin (a tedy ani SD) příliš neliší, má největší smysl ke standardizaci použít souhrnnou SD_{pooled}

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + \ldots + (n_k - 1)SD_k^2}{n_1 + n_2 + \ldots + n_k - k}}$$

- kde SD_1^2 až SD_k^2 jsou rozptyly v jednotlivých skupinách, n jsou velikosti skupin a k počet skupin.
- Pokud se rozptyly skupin výrazně liší, můžeme ke standardizaci použít SD jedné, "referenční" skupiny (např. v rámci experimentů to bývá kontrolní skupina).
- Pokud se rozptyly skupin výrazně liší a nedává smysl volit nějakou skupinu jako referenční, můžeme ke standardizaci použít průměr z odmocniny rozptylů dvou právě porovnávaných skupin:

$$SD_{avg} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

Příklady tabulek

Tabulka 1

Deskriptivní statistiky proměnné libido v závislosti na dávce

Dávka	*n*	*M*	*SD*
Vysoká	5	5,00	1,59
Nízká	5	3,20	1,30
Placebo	5	2,20	1,30

Tabulka 2

Deskriptivní statistiky proměnné libido v závislosti na dávce a Cohenovo d

Dávka	*n*	*M*	*SD*	1.	2.
1. Vysoká dávka	5	5,00	1,58		
2. Nízká dávka	5	3,20	1,30	1,24	
3. Placebo	5	2,20	1,30	1,93	0,77

Tabulka 3

ANOVA s fixními efekty a libidem jako závislou proměnnou

Prediktory	Součet čtverců	*df*	Střední čtverec	*F*	*p*	η^2
(Průsečík)	180,27	1	180,27	91,66	< 0,001	
Dávka	20,13	2	10,06	5,15	0,025	0,56
Chyba	23,60	12	1,97			

Příklady tabulek

Tabulka 4

Průměry a směrodatné odchylky vnímané atraktivity v závislosti na pohlaví a konzumaci alkoholu

Pohlaví	Konzumace alkoholu									Celkem		
	2 drinky			4 drinky			Žádný drink					
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Muži	8	62,5	6,7	8	57,5	7,1	8	60,6	4,9	24	60,2	6,3
Ženy	8	66,9	12,5	8	35,6	10,8	8	66,9	10,3	24	54,5	18,5
Celkem	16	64,7	9,9	16	46,6	14,3	16	63,8	8,5	48	58,4	12,4

Tabulka 5

ANOVA s fixními efekty a vnímanou atraktivitou jako závislou proměnnou

Prediktory	Součet čtverců (III. typu)	*df*	Střední čtverec	*F*	*p*	parciální η^2
(Průsečík)	163333	1	163333	1967,03	< 0,001	
Pohlaví	169	1	169	2,03	0,161	0,05
Alkohol	3332	2	1666	20,07	< 0,001	0,49
Pohlaví × Alkohol	1978	2	989	11,91	< 0,001	0,36
Chyba (Reziduum)	3487	42	83			