Instatní statistika v jedné hodině (pro metodologii III)

1. přehled metod
2. kontingenční tabulky
3. porovnávání průměrů
4. korelace, regrese
5. faktorová analýza
6. shluková analýza

Přehled metod

- obě proměnné kategoriální: **kontingenční tabulka + Chí-kvadrát**
- nezávislá proměnná kategoriální, závislá kardinální: **porovnávání průměrů**
- obě proměnné kardinální: **korelace, regrese**

Test Chí-kvadrát

- chí-kvadrát pro testování nezávislosti proměnných se používá pro nominální nebo ordinální proměnné
- data jsou uspořádána do tzv. kontingenční tabulky (viz příklad)

Příklad

- zajímá nás, jak souvisí model manželství s jeho vydařeností
  - model manželství má kategorie: dominance žena, dominance muž, kooperace
  - vydařenost má 3 kategorie – vydařené, průměrné, nevydařené
- pozn.: jde o manželství rodičů respondentů, tak jak je posuzují oni (zdroj dat – výzkum doc. Plaňavy)

Příklad

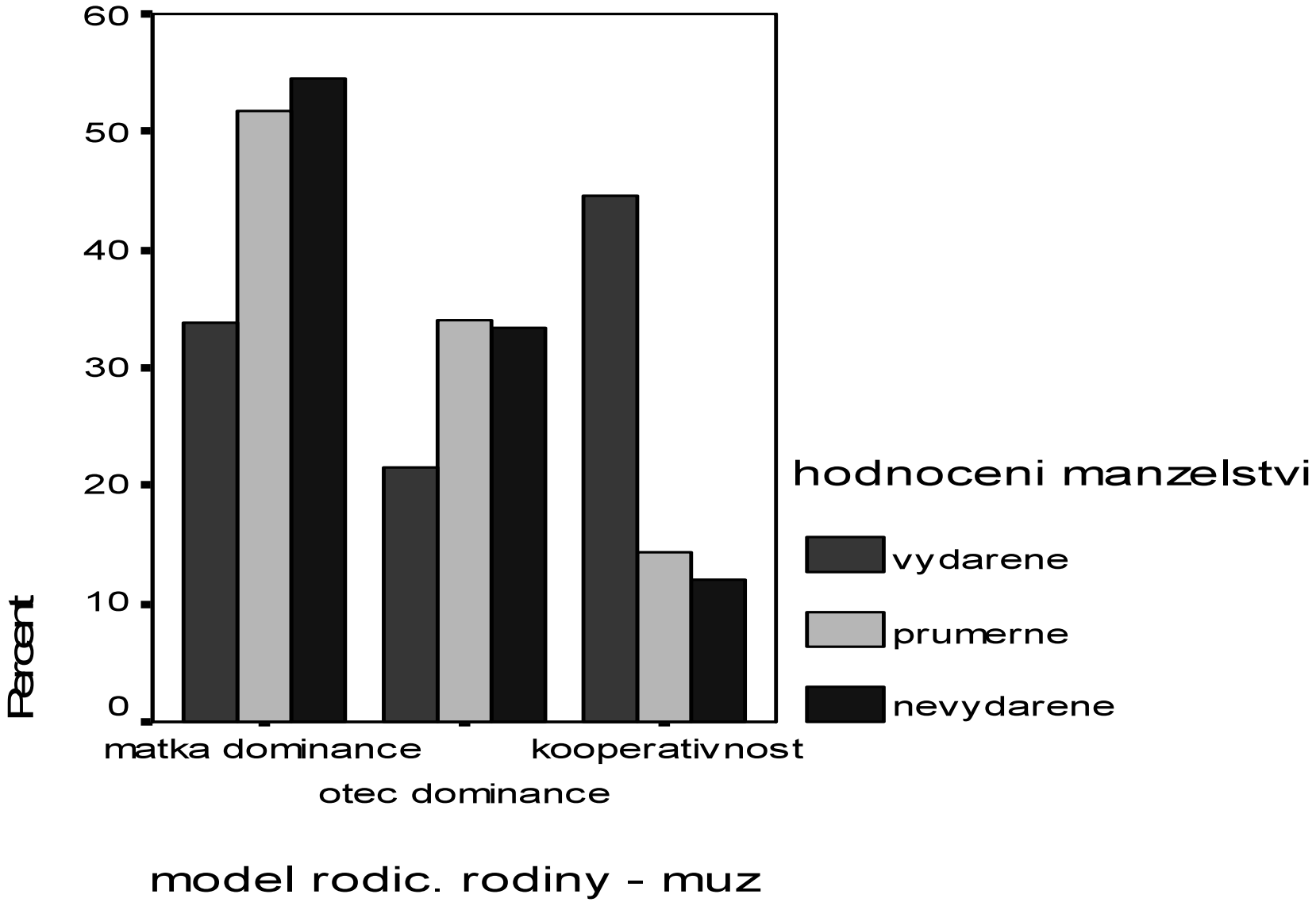
- otázka zní: liší se podíl vydařených, průměrných a nevydařených manželství u rodin, kde dominovala matka, rodin, kde dominoval otec a u rodin, kde nedominoval ani jeden z nich?

Kontingenční tabulka (SPSS)

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

Count

		hodnoceni manzelstvi rodicu - muz			
		vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	22	29	18	69
	otec dominance	14	19	11	44
	kooperativnost	29	8	4	41
Total		65	56	33	154



model rodic. rodiny - muz

Test Chí-kvadrát

- chí-kvadrát porovnává očekávané a pozorované četnosti
- očekávané jsou četnosti za předpokladu, že proměnné jsou nezávislé

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

			hodnoceni manzelstvi rodicu - muz			
			vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	Count	22	29	18	69
		% within model rodic. rodiny - mu	31,9%	42,0%	26,1%	100,0%
	otec dominance	Count	14	19	11	44
		% within model rodic. rodiny - mu	31,8%	43,2%	25,0%	100,0%
	kooperativnost	Count	29	8	4	41
		% within model rodic. rodiny - mu	70,7%	19,5%	9,8%	100,0%
Total		Count	65	56	33	154
		% within model rodic. rodiny - mu	42,2%	36,4%	21,4%	100,0%

Příklad

- v našem příkladu bylo 42,2% vydařených manželství
- pokud by proměnné (model a vydařenost manželství) byly vzájemně nezávislé, poměr vydařených manželství v jednotlivých modelech manželství by měl být přibližně stejný (a odrážet celkový podíl) – 42%
- podobně ostatní kategorie…

Test Chí-kvadrát

- očekávané četnosti – výpočet:

$$O_{ij} = (\check{r}_i \, s_j \,) / \, N$$

(pro každé políčko tabulky se vynásobí celkové četnosti z příslušného řádku se sloupcovými četnostmi a vydělí celkovým počtem osob)

Příklad

rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstab

Count

		hodnoceni manzelstvi rodicu - mu			
		vydarene	prumerne	nevydarene	Total
model rodic	matka dominan	22	29	18	69
rodiny - mu	otec dominance	14	19	11	44
	kooperativnost	29	8	4	41
Total		65	56	33	154

Očekávané četnosti

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

			hodnoceni manzelstvi rodicu - muz			
			vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	Count	22	29	18	69
		Expected Count	29,1	25,1	14,8	69,0
	otec dominance	Count	14	19	11	44
		Expected Count	18,6	16,0	9,4	44,0
	kooperativnost	Count	29	8	4	41
		Expected Count	17,3	14,9	8,8	41,0
Total		Count	65	56	33	154
		Expected Count	65,0	56,0	33,0	154,0

Test Chí-kvadrát

- chí-kvadrát porovná očekávané četnosti s pozorovanými

$$\chi^2 = \Sigma\ [(\text{pozor. četnosti} - \text{oček.})^2/\text{oček.}]$$

Test Chí-kvadrát v SPSS

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18,712[a]	4	,001
Likelihood Ratio	18,837	4	,001
Linear-by-Linear Association	11,482	1	,001
N of Valid Cases	154		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,79.

Porovnávání průměrů

- Možné typy problémů:
  - porovnáváme **průměr vzorku s průměrem populace** → jednovýběrový t-test
  - porovnáváme **průměry dvou vzorků** → t-test pro nezávislé výběry
  - porovnáváme **dva průměry jednoho vzorku** → t-test pro závislé výběry (tzv. párový t-test)
  - porovnáváme více průměrů → analýza rozptylu

T-test pro nezávislé výběry

- tento test používáme, pokud chceme porovnat průměry dvou skupin případů
- např.
  - průměrné skóre v neurocitismu u mužů a žen
  - průměr v indexu životní spokojenosti u extravertů a introvertů atd.

T-test pro závislé výběry

- označuje se někdy také jako t-test pro párované výběry
- v naprosté většině případů se používá pro porovnání dvou měření u stejných osob (tj. páru měření u jedné skupiny osob)
- někdy také pro porovnání průměrů u dvou skupin osob, které tvoří páry (např. manželské či podle jiného klíče – věku, pohlaví, nemoci atd.)

T-test pro závislé výběry - příklad

- Psychiatr chce vyhodnotit úspěšnost určitého způsobu terapie poruch příjmu potravy. Terapie se účastnilo 10 dívek. U každé z nich byla zaznamenána váha před a po terapii. Psychiatr si chce ověřit, zda jejich hmotnost průkazně vzrostla.

Porovnání výzkumných plánů

- t-test pro nezávislé výběry se používá většinou u výzkumných plánů s výzkumnou a kontrolní skupinou
- zatímco t-test pro závislé výběry většinou u výzkumných plánů s opakovaným měřením u stejných osob

Porovnávání průměrů

- t-testy jsou určeny pouze pro porovnávání dvojice průměrů
- v mnoha výzkumných plánech je však více skupin než dvě

Analýza rozptylu

- proto je vhodnější místo mnoha t-testů použít jinou statistickou techniku – analýzu rozptylu
- **an**alysis **o**f **va**riance –ANOVA
- umožňuje otestovat rozdíly mezi průměry více skupin najednou

Analýza rozptylu

- logika analýzy rozptylu
- výpočetní postup
- mnohonásobná porovnávání
- opakovaná měření
- faktoriální analýza rozptylu
- analýza kovariance
- vícerozměrná analýza rozptylu

Logika analýzy rozptylu

- analýza rozptylu nevyužívá pro testování rozdílu mezi průměry samotné průměry, ale **rozptyly**
- počítají se dva odhady:
  - rozptyl uvnitř skupin (within-groups nebo within-subjects variance)
  - rozptyl mezi skupinami (between-groups nebo between-subjects variance)

Logika analýzy rozptylu

- **rozptyl uvnitř skupin** je ukazatel celkové variability uvnitř skupin – tj. jak se od sebe vzájemně liší osoby v rámci jednotlivých skupin
- **rozptyl mezi skupinami** je měřítkem variability mezi skupinami – tj. jak se od sebe liší skupiny osob

Logika analýzy rozptylu

- poměr těchto dvou rozptylů je statistika F

$$F = \frac{\text{rozptyl mezi skupinami}}{\text{rozptyl uvnitř skupin}}$$

Logika analýzy rozptylu

- pokud nejsou mezi skupinami rozdíly, pak by měl být rozptyl mezi skupinami a uvnitř skupin velmi podobný (teoreticky shodný - F=1)
- pokud jsou mezi skupinami rozdíly, pak budou tyto rozdíly (between)větší než vzájemné rozdíly mezi osobami uvnitř skupin (within)

Logika analýzy rozptylu

- je-li F>1, pak kromě F musíme ještě spočítat pravděpodobnost, že bychom takto vysoké získali náhodou (tj. statistickou významnost)
- tabulka F rozdělení je vždy pro konkrétní hodnotu alfa; má v řádcích počet stupňů volnosti pro rozptyl uvnitř skupin a ve sloupcích pro rozptyl mezi skupinami

Výstup v SPSS

ANOVA

LATENCE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	332,986	2	166,493	6,173	,006
Within Groups	809,195	30	26,973		
Total	1142,182	32			

rozptyl mezi skupinami

rozptyl uvnitř skupin

hladina významnosti

Mnohonásobná porovnávání

- průkaznost F nám řekne, **zda** existují průkazné rozdíly mezi průměry
- ale **ne**dozvíme se tak, **mezi kterými** skupinami je průkazný rozdíl (která skupina se liší od které)
- je třeba provést tzv. **mnohonásobná porovnání** (multiple comparisons nebo post-hoc comparisons)

Mnohonásobná porovnávání

- jde vpodstatě o upravené t-testy
  - upravené vzhledem k počtu porovnávání
- existuje více různých typů mnohonásobných porovnávání, např. Fisherův LSD test, Bonferroniho test, Tukeyho test, Scheffeho test atd.

Opakovaná měření

- analýza rozptylu může být aplikována také na data z opakovaných měření
  - podobně jako t-test pro závislé výběry; analýza rozptylu se použije v případě, máme-li více než dvě měření
- např. v příkladu u t-testu – změna hmotnosti u dívek s PPP po terapii – hmotnost by mohla být měřena i několikrát v průběhu terapie

Opakovaná měření

- procedura se nazývá Analýza rozptylu pro opakovaná měření (Repeated measures)
- logika výpočtu je obdobná jako u analýzy rozptylu pro nezávislá data

Faktoriální analýza rozptylu

- **faktor** je v analýze rozptylu nezávislá proměnná
- máme-li faktorů (nezávislých proměnných) více, použijeme faktoriální ANOVu
- může jít o porovnání nezávislých výběrů, o opakovaná měření nebo obojí najednou (tzv. mixed design – se smíšenými efekty)

Faktoriální analýza rozptylu

- **příklad**: neuropsycholog zkoumá oblasti mozku odpovídající za tvorbu a porozumění řeči
- vyšetří speciálním testem 24 náhodně vybraných pacientů s poškozenou levou hemisférou mozku – polovina z nich jsou muži a polovina ženy
- kromě mezipohlavních rozdílů ho zajímá rovněž, zda bude rozdíl mezi praváky a leváky (těch je rovněž 12 a 12)

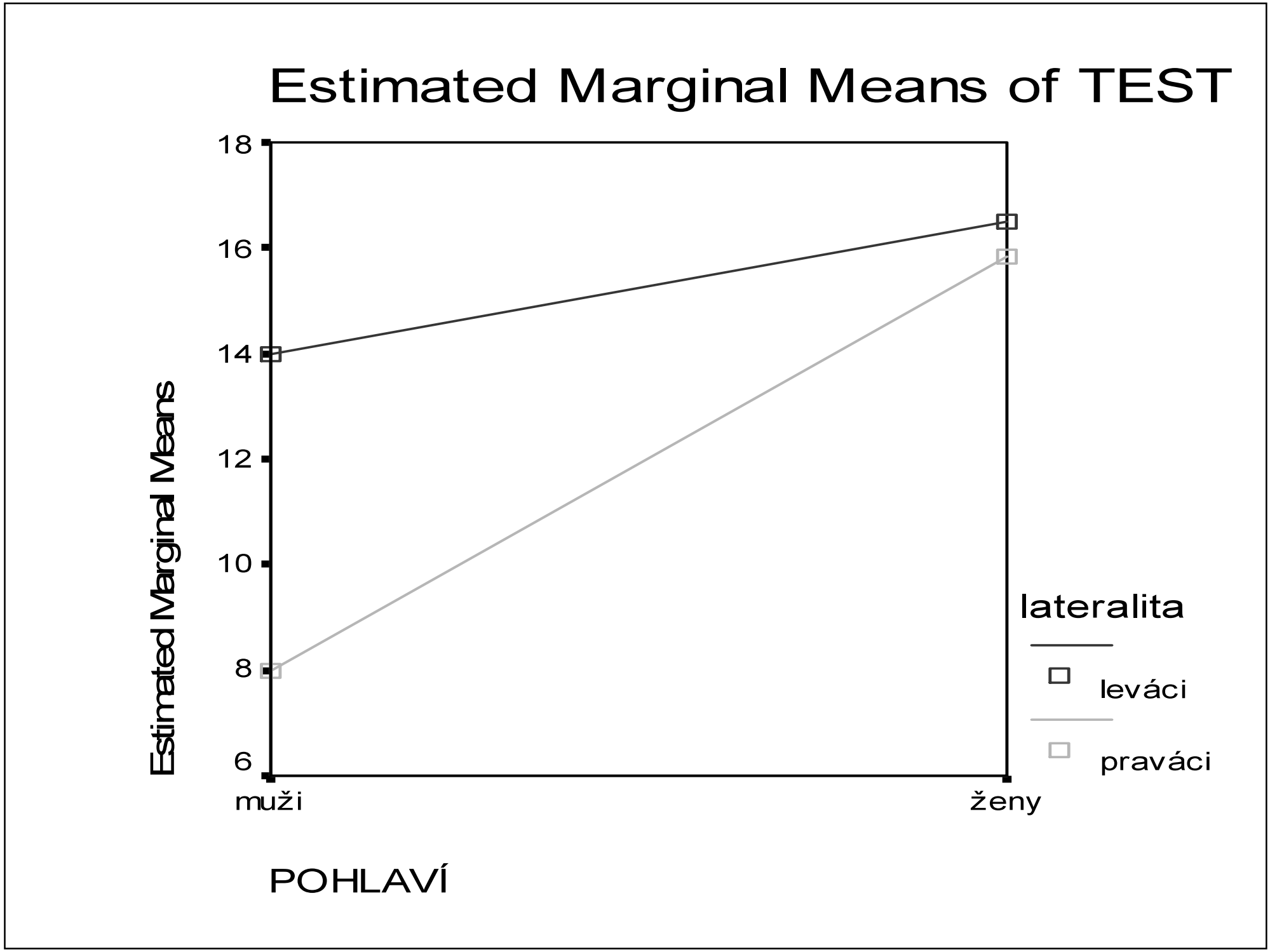
Faktoriální analýza rozptylu

- tento design se zapisuje 2x2 ANOVA
  - 2 kategorie pohlaví (muži x ženy)
  - 2 kategorie laterality (leváci x praváci)

Descriptive Statistics

Dependent Variable: TEST

POHLAVÍ	lateralita	Mean	Std. Deviation	N
muži	leváci	14,0000	2,89828	6
	praváci	8,0000	2,36643	6
	Total	11,0000	4,02266	12
ženy	leváci	16,5000	3,08221	6
	praváci	15,8333	2,63944	6
	Total	16,1667	2,75791	12
Total	leváci	15,2500	3,13702	12
	praváci	11,9167	4,73782	12
	Total	13,5833	4,28259	24


Estimated Marginal Means of TEST
Estimated Marginal Means
18
16
14
12
10
8
6
muži
ženy
POHLAVÍ
lateralita
leváci
praváci

Faktoriální analýza rozptylu

- faktoriální analýza rozptylu testuje
  - hlavní efekty
  - interakce

Faktoriální analýza rozptylu

- **hlavní efekt** (main effect) – vliv jedné nezávislé proměnné zprůměrovaný pro všechny úrovně ostatních nezávislých proměnných
- u faktoriální ANOVy jsou testovány hlavní efekty pro všechny faktory
- v příkladu testujeme hlavní efekt pro pohlaví a lateralitu

Tests of Between-Subjects Effects

Dependent Variable: TEST

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Mode	269,500[a]	3	89,833	11,794	,000
Intercept	4428,167	1	4428,167	581,379	,000
POHLAVÍ	160,167	1	160,167	21,028	,000
LATERAL	66,667	1	66,667	8,753	,008
POHLAVÍ * LAT	42,667	1	42,667	5,602	,028
Error	152,333	20	7,617		
Total	4850,000	24			
Corrected Total	421,833	23			

a. R Squared = ,639 (Adjusted R Squared = ,585)

Faktoriální analýza rozptylu

- průkazný (na hladině 1%) hlavní efekt pro faktor pohlaví
- ženy mají celkově vyšší skóry než muži (16,2 a 11,0)

Tests of Between-Subjects Effects

Dependent Variable: TEST

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Mode	269,500[a]	3	89,833	11,794	,000
Intercept	4428,167	1	4428,167	581,379	,000
POHLAVÍ	160,167	1	160,167	21,028	,000
LATERAL	66,667	1	66,667	8,753	,008
POHLAVÍ * LAT	42,667	1	42,667	5,602	,028
Error	152,333	20	7,617		
Total	4850,000	24			
Corrected Total	421,833	23			

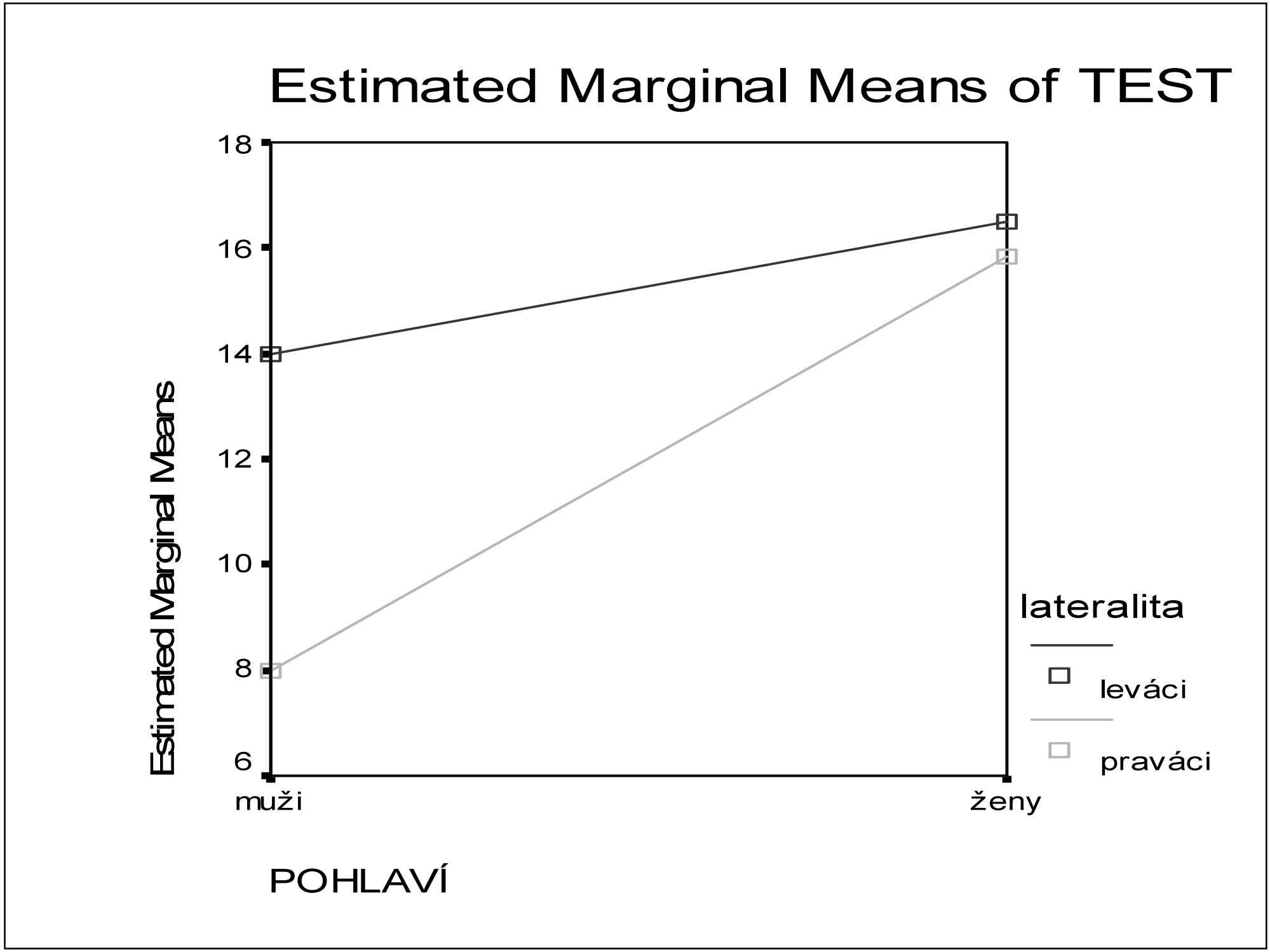
a. R Squared = ,639 (Adjusted R Squared = ,585)

Faktoriální analýza rozptylu

- průkazný (na hladině 1%) hlavní efekt pro faktor lateralita
- leváci mají celkově vyšší skóry než praváci (15,3 a 11,9)

Faktoriální analýza rozptylu

- **interakce** se projeví v případě, kdy vliv jedné nezávislé proměnné není stejný na všech úrovních druhé nezávislé proměnné
- v příkladu – je vliv laterality stejný u mužů a žen?
  - pokud ano, není zde interakce
  - pokud ne, je zde interakce


Estimated Marginal Means of TEST
Estimated Marginal Means
18
16
14
12
10
8
6
muži
ženy
POHLAVÍ
lateralita
leváci
praváci

Tests of Between-Subjects Effects

Dependent Variable: TEST

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Mode	269,500[a]	3	89,833	11,794	,000
Intercept	4428,167	1	4428,167	581,379	,000
POHLAVÍ	160,167	1	160,167	21,028	,000
LATERAL	66,667	1	66,667	8,753	,008
POHLAVÍ * LAT	42,667	1	42,667	5,602	,028
Error	152,333	20	7,617		
Total	4850,000	24			
Corrected Total	421,833	23			

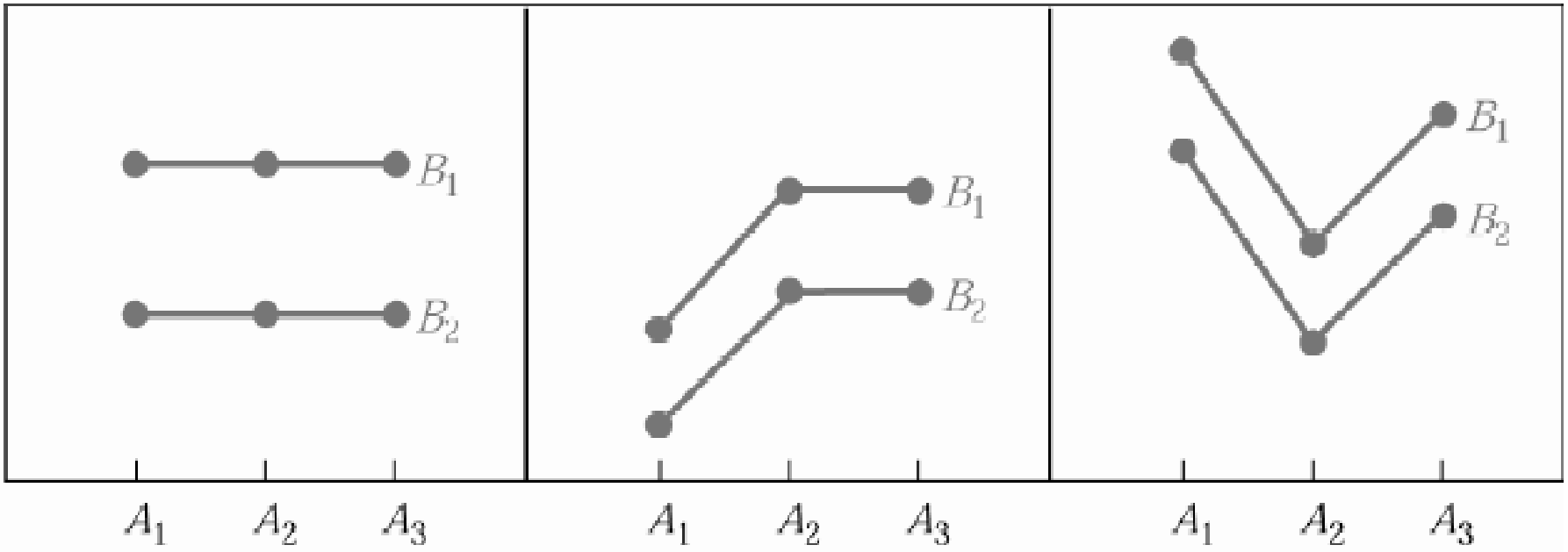
a. R Squared = ,639 (Adjusted R Squared = ,585)

Faktoriální analýza rozptylu

- interakce mezi pohlavím a lateralitou je průkazná (na 5% hladině významnosti)
- u žen nehraje lateralita pro výkon v testu roli – levačky a pravačky se neliší, zatímco u mužů leváci a praváci ano

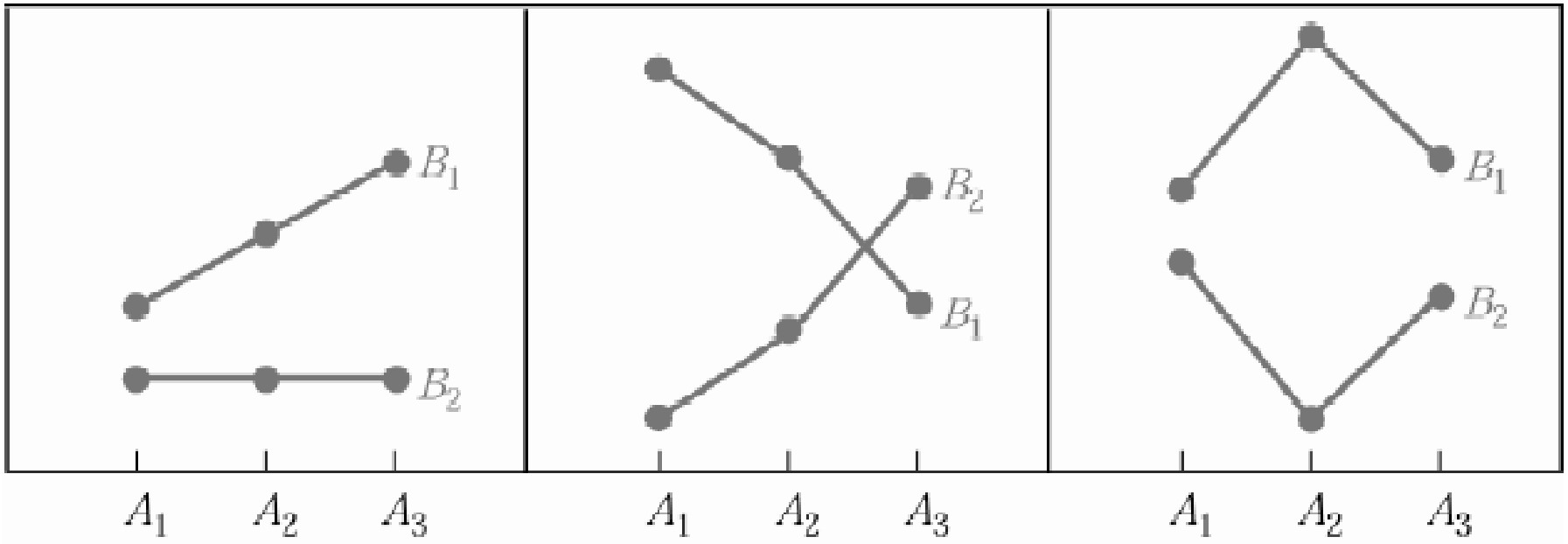
Faktoriální analýza rozptylu

- bez interakce – pouze hlavní efekty



Faktoriální analýza rozptylu

- interakce



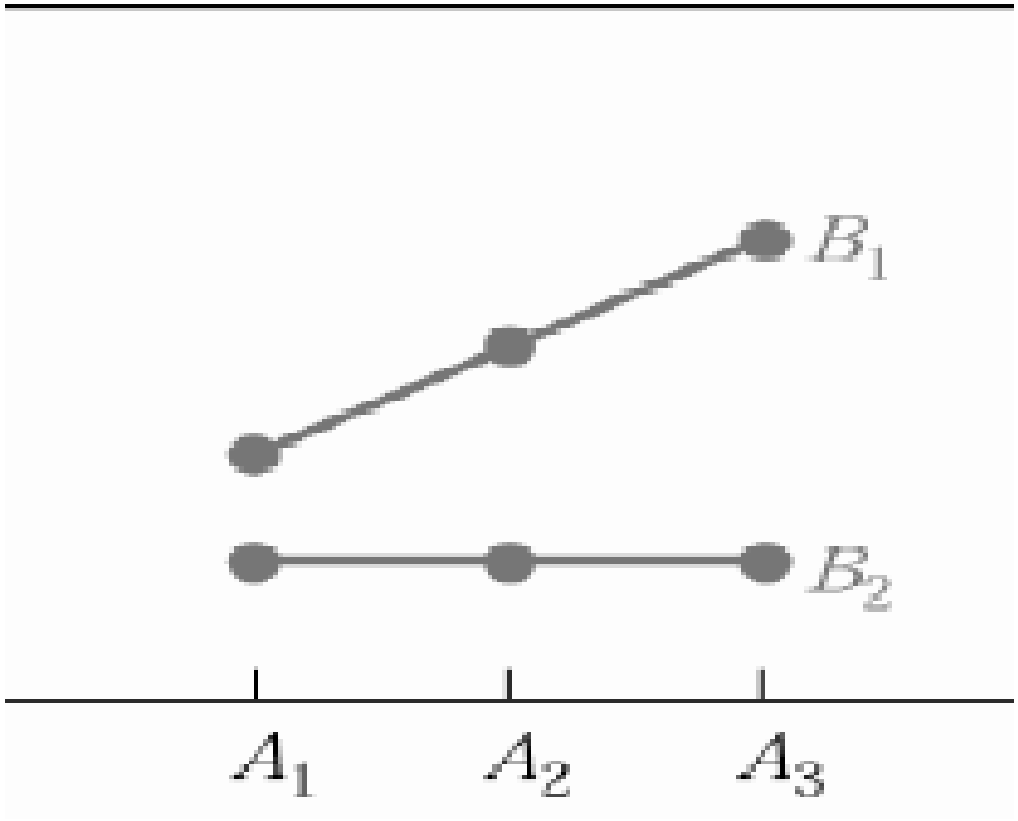
Opakovaná měření s další nezávislou proměnnou

- faktoriální design je možno uplatnit i u analýzy opakovaných měření
- interakce zde znamená, že jsou různě velké rozdíly mezi měřeními u jednotlivých kategorií nezávislé proměnné

Opakovaná měření s další nezávislou proměnnou

- **příklad**: psychiatr testující léčbu anorexie by mohl soubor rozdělit na dívky podstupující terapii dobrovolně a nedobrovolně
  - interakce by mohla vypadat třeba tak, že u motivovaných dívek by došlo k nárůstu hmotnosti, zatímco u nedobrovolných pacientek ke stagnaci

Opakovaná měření s další nezávislou proměnnou



Analýza kovariance

- kromě kategoriálních faktorů je možno do analýzy zařadit také spojitou nezávislou proměnnou – tzv. **kovariát**
- pak jde o analýzu kovariance (ANCOVA)

Analýza kovariance

- **příklad**: šéf firmy obdrží stížnost od zaměstnankyň, že ženy mají nižší platy než muži
- podle porovnání průměrů to tak vypadá, ale co kdybychom do analýzy zařadili jako další faktor (kovariát) délku praxe?

Multivariační analýza rozptylu

- ve všech předchozích příkladech jsme měli pouze **jednu závislou** proměnnou
- je však možno testovat také vliv jednoho či více faktorů na **několik závislých proměnných najednou**
- tato analýza se označuje jako MANOVA (multivariate analysis of variance)

Multivariační analýza rozptylu

- **příklad**: reklamní psycholog chce porovnat účinnost dvou typů TV reklam (emocionální x informativní)
- nechá respondenty hodnotit na 7-ti stupňové škále 3 aspekty účinnosti reklamy: zda je reklama zaujala, zda se jim líbí a jestli by uvažovali o koupi inzerovaného výrobku
- tyto 3 závislé proměnné pak porovná pro typ reklamy jako faktor

Vztahy mezi proměnnými

- obecná definice – síla a směr vztahu
- míry asociace pro nominální data
  - založené na chí-kvadrátu
  - PRE míry
- míry asociace pro ordinální data

Míry asociace

- míry asociace vyjadřují **těsnost vztahu proměnných** (a případně **směr** vztahu)
- z chí-kvadrátu se dozvíme pouze, **zda nějaký vztah mezi proměnnými existuje** (tj. zda se liší četnosti pozorované a četnosti očekávané za předpokladu, že proměnné jsou nezávislé)

Míry asociace

- **těsnost (síla) vztahu** – vyjádřena absolutní hodnotou koeficientu
- není shoda v tom, od jaké hodnoty je vztah považován za těsný (někdy uváděno >0.70, jindy >0.30), středně těsný či slabý

Míry asociace

- **směr vztahu** – pouze u ordinálních a kardinálních proměnných
- **pozitivní vztah** – čím vyšší hodnoty jedné proměnné, tím vyšší hodnoty druhé proměnné
- **negativní vztah** - čím vyšší hodnoty jedné proměnné, tím nižší hodnoty druhé proměnné

Míry asociace pro nominální data

- míry asociace pro nominální data ukazují pouze sílu vztahu dvou proměnných, nikoli směr či jiné informace o povaze vztahu
- rozlišujeme míry založené na chí-kvadrátu a míry PRE

Míry založené na chí-kvadrátu

- velikost hodnoty chí-kvadrát je ovlivněna velikostí výběru a počtem kategorií tabulky
- účelem koeficientů založených na chí-kvadrátu je eliminovat tyto vlivy

Míry založené na chí-kvadrátu

- rozsah koeficientů je obvykle mezi 0 a 1
  - čím vyšší hodnota, tím těsnější vztah
  - 0 – žádný vztah
  - 1 – absolutní vztah (z hodnot jedné proměnné můžeme předpovědět hodnoty druhé proměnné)
- pro koeficienty je možno spočítat statistickou významnost

Míry založené na chí-kvadrátu

- mezi nejčastěji užívané míry asociace založené na chí-kvadrátu patří koeficienty
  - Fí (Phi)
  - Cramerovo V (Cramer's V)
  - koeficient kontingence (Contingency Coefficient)

Míry založené na chí-kvadrátu

- **Fí koeficient** - užívá se pro tabulky 2x2 (tj. pro dichotomické proměnné, např. pohlaví)
- vypočte se tak, že se hodnota chí-kvadrátu vydělí počtem osob a výsledek se odmocní

Míry založené na chí-kvadrátu

- koeficient kontingence – užívá se někdy místo Fí pro tabulky větší než 2x2
- bohužel jeho max. hodnota je nižší než 1 (závisí na počtu políček tabulky)
- neužívá se proto příliš často

Míry založené na chí-kvadrátu

- Cramerovo V – podobný výpočet jako Fí; počet osob se navíc násobí počtem řádků - 1
  - (pokud je počet řádků menší než počet sloupců, jinak počtem sloupců – 1)
- používá se pro tabulky větší než 2x2

Míry PRE

- **PRE** je zkratka pro **Proportional Reduction in Error** (poměrná redukce chyby odhadu)
- princip PRE: porovnání odhadu hodnot závislé proměnné bez znalosti hodnot nezávislé proměnné a s její znalostí (o kolik se sníží chyba odhadu?)

Míry PRE

- **příklad** – jaký je vztah mezi pohlavím a užíváním rtěnky?*
- vypočítáme koeficient **lambda**
- pokud bychom měli odhadnout, zda náhodně vybraný respondent používá rtěnku: jaká je pravděpodobnost chybného odhadu?

- *převzat z Disman: Jak se vyrábí sociologická znalost

Míry PRE

- můžeme očekávat, že více lidí rtěnku nepoužívá než používá (naprostá většina mužů + některé ženy)
- takže bude výhodnější odhadnout, že náhodně vybraný respondent rtěnku nepoužívá
- pravděpodobnost chyby závisí na podílu lidí užívajících rtěnku

Míry PRE

RTĚNKA

		Frequency	Percent
Valid	nepoužívá	97	60,6
	používá	63	39,4
	Total	160	100,0

Míry PRE

- při tomto podílu osob je pravděpobnost chyby asi 40% (když budeme odhadovat, že nádodný respondent rtěnku neužívá)
- ze 160 případů bychom se zmýlili 63x

Míry PRE

- **o kolik by se chyba zmenšila, kdybychom znali pohlaví respondenta?**
- pro muže bychom odhadovali, že rtěnku nepoužívá, pro ženu naopak - že ji používá

Míry PRE

POHLAVÍ * RTĚNKA Crosstabulation

Count

		RTĚNKA		
		nepoužívá	používá	Total
POHLAVÍ	muži	78	2	80
	ženy	19	61	80
Total		97	63	160

Míry PRE

- pokud bychom znali pohlaví respondenta, zmýlili bychom se ve svém odhadu 21x (2 x u muže a 19x u ženy)
- **o kolik by se náš odhad zlepšil?**

Míry PRE

- chyby předtím – chyby teď = 63-21 = 42
- poměrná redukce chyby (tj. vzhledem k předchozím chybám) = **lambda** = 42/63 = **0,667**
- **chyba v odhadu užívání rtěnky se sníží asi o 67%, pokud známe pohlaví respondenta**

Míry PRE

- rozsah koeficientu lambda je od 0 do 1
- **0** znamená, že znalost hodnoty nezávislé proměnné vůbec nesníží chybu v odhadu hodnot závislé proměnné; **proměnné jsou vzájemně nezávislé**
- čím blíže **1**, tím lépe můžeme z hodnot nezávislé proměnné předpovědět hodnoty závislé proměnné

Míry PRE pro nominální data

- kromě koeficientu lambda se užívají také
  - Goodmanovo a Kruskalovo **tau** (nevyužívá při predikci nejčastější kategorii závislé proměnné jako lambda, ale rozdělení ve všech kategoriích závisle proměnné)
  - Cohenova **Kappa** – pro měření **shody dvou posuzovatelů**

Míry asociace pro ordinální data

- u ordinálních dat je výpočet založen na poměru souhlasných a nesouhlasných párů případů
- **souhlasný** pár případů – hodnota obou proměnných je vyšší (nebo nižší) u jednoho člena páru
- **nesouhlasný** pár případů – hodnota jedné proměnné je u jednoho člena páru vyšší a hodnota druhé proměnné je nižší

Míry asociace pro ordinální data

- koeficient gamma = počet souhlasných mínus počet nesouhlasných párů, tento rozdíl vzhledm k celkovému počtu souhlasných a nesouhlasných párů
- nerozhodné páry nebere gamma v úvahu

Míry asociace pro ordinální data

- pokud je většina párů souhlasných, je hodnota gamma kladná – tj. **pozitivní vztah** (až +1)
- pokud je většina párů nesouhlasných, je hodnota gamma záporná – tj. **negativní vztah** (až -1)
- pokud je počet souhlasných a nesouhlasných párů vyrovnán – gamma kolem 0

Míry asociace pro ordinální data

- gamma je symetrická míra – nedělá rozdíly mezi závislou a nezávislou proměnnou
- asymetrická varianta koeficientu gamma – **Somersovo D**
- **Kendallovo tau b**– bere v úvahu i nerozhodné páry (tzv. ties); ale hodnoty v rozsahu -1 až +1 mohou být získány pouze pro čtvercové tabulky (tj. stejný počet kategorií obou proměnných)

Shrnutí

- u nominálních dat hodnota míry asociace proměnných indikuje sílu vztahu – rozsah od 0 do 1
  - nejužívanější Fí nebo Cramerovo V; když víme, která proměnná nezávislá - lambda
- u ordinálních dat míry asociace indikují jak sílu vztahu (abs. hodnota koeficientu), tak směr vztahu

Korelační a regresní analýza

1. Pearsonův korelační koeficient
2. jednoduchá regresní analýza
3. vícenásobná regresní analýza

Pearsonův korelační koeficient

- u intervalových a poměrových dat můžeme jako míru asociace – vztahu mezi proměnnými použít **Pearsonův korelační koeficient**
- **korelace**
  - ko = s, spolu, vzájemně
  - relace = vztah
  - korelace = vzájemný vztah proměnných

Pearsonův korelační koeficient

- absolutní hodnota koeficientu vyjadřuje **sílu (těsnost) vztahu**
- znaménko (+ nebo -) **směr vztahu**
- **rozsah -1 až +1**
- označuje se **r**

Pearsonův korelační koeficient

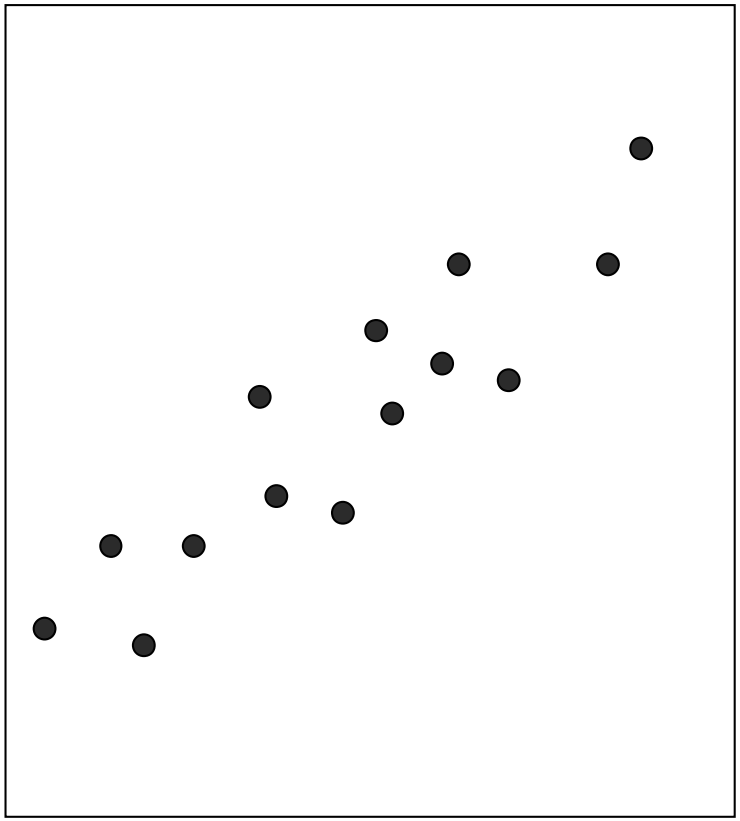
- sám o sobě je deskriptivní statistikou, ale podobně jako u ostatních měr asociace je možno spočíst **statistickou významnost**
- závisí na velikosti výběru – čím vyšší, tím nižší koeficient vychází průkazný

Pearsonův korelační koeficient

- je mírou **pouze pro lineární vztahy**
- před výpočtem je vhodné zobrazit vztah mezi proměnnými také graficky – tzv. **scatter** (dvourozměrný tečkový diagram)

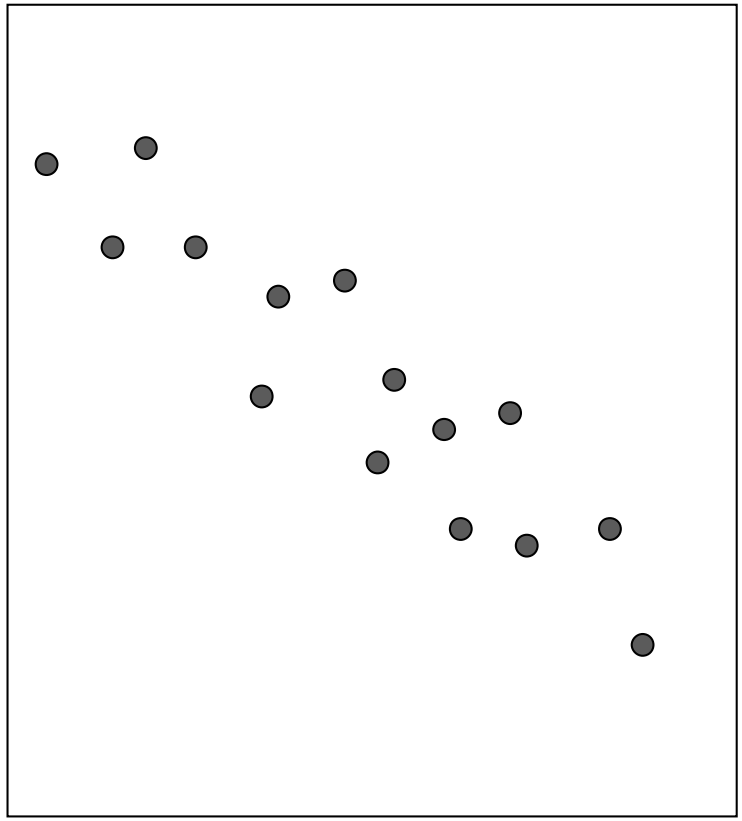
Scatter

- **pozitivní vztah** (přímá úměra) – čím vyšší hodnoty proměnné X, tím vyšší hodnoty proměnné Y
- r > 0



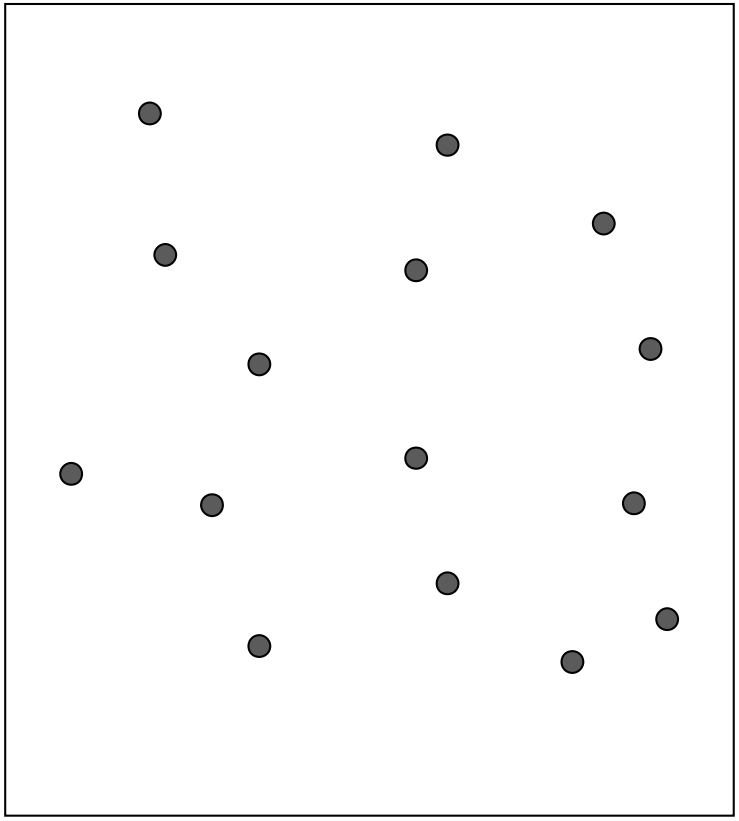
Scatter

- **negativní vztah** (nepřímá úměra) – čím vyšší hodnoty proměnné X, tím nižší hodnoty proměnné Y
- $r < 0$



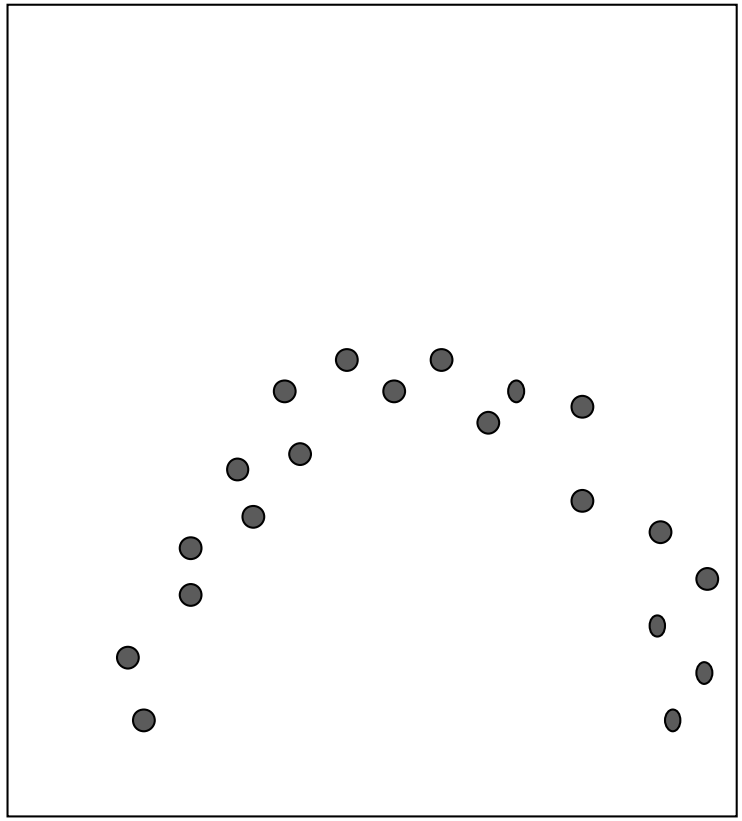
Scatter

- **žádný vztah** - hodnoty proměnné X nesouvisí s hodnotami proměnné Y
- r = 0



Scatter

- **nelineární vztah**
- r = 0



Interpretace r

- není shoda v tom, jaká hodnota r je považována za těsný vztah
- interpretace navržená Guilfordem:
  - <0.20 zanedbatelný vztah
  - 0.20-0.40 nepříliš těsný vztah
  - 0.40-0.70 středně těsný vztah
  - 0.70-0.90 velmi těsný vztah
  - >0.90 extrémně těsný vztah

Interpretace r

- pro lepší interpretaci je vhodné převést koeficient korelace na **koeficient determinace (r^2)**
- ukazuje, kolik rozptylu v jedné proměnné může být vysvětleno rozptylem ve druhé proměnné

Interpretace r

- **korelace neznamená příčinný vztah mezi proměnnými!!**
  - ten můžeme ověřovat pouze experimentem, kdy jsou všechny ostatní proměnné udržovány konstatní, proměnná X předchází Y v čase atd.

Faktory ovlivňující r

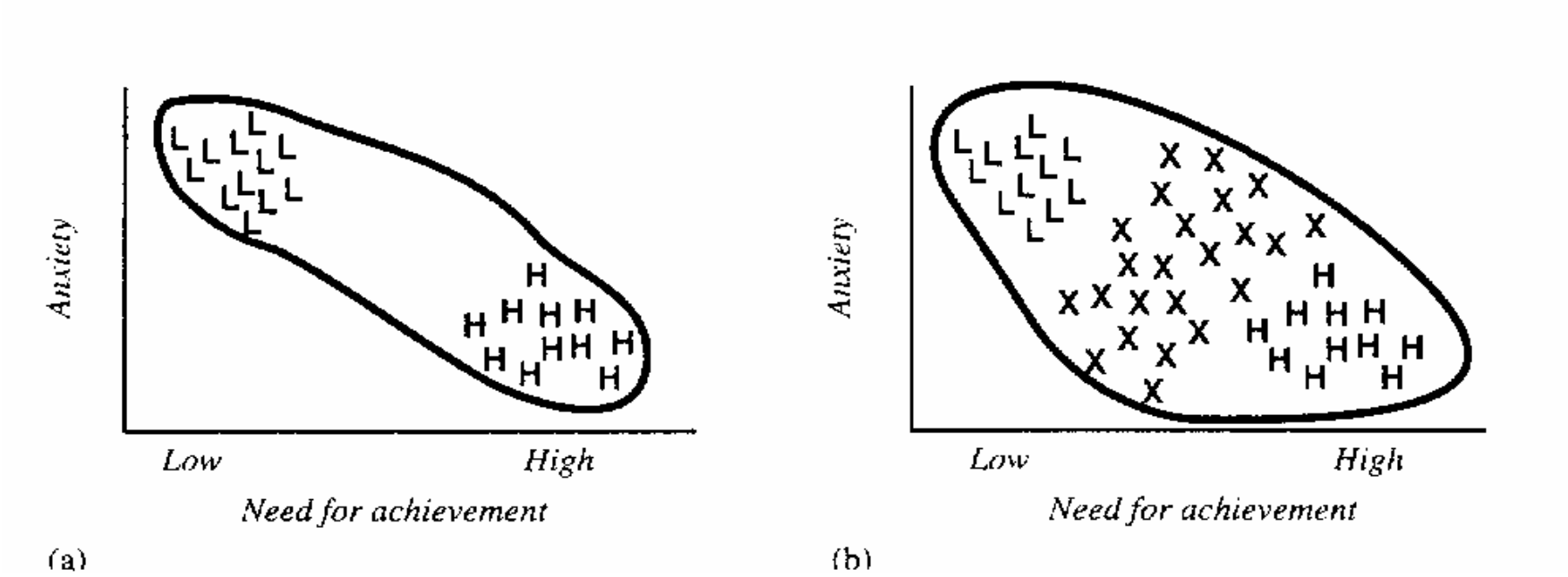
- omezený rozsah hodnot proměnné
- použití extrémních skupin
- nehomogenní soubor
- extrémní hodnoty (outliers)
- nelineární vztahy
- reliabilita použitých nástrojů

Omezený rozsah hodnot

- omezený rozsah hodnot jedné nebo obou proměnných snižuje hodnotu r
- stejně tak nízká variabilita (extrémní případ:pokud by všechny hodnoty 1 proměnné byly stejné, zákonitě r=0)

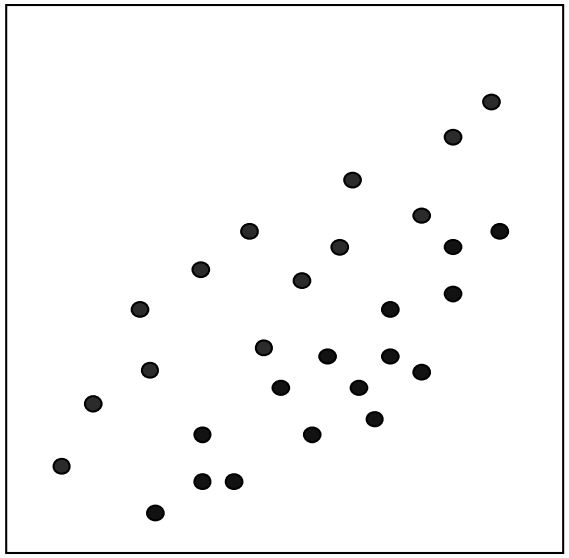
Použití extrémních skupin

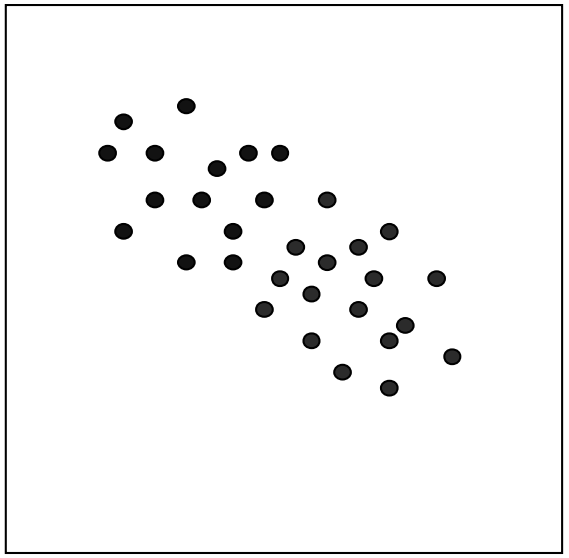
- použití extrémních skupin (např. jen osob s vysokým IQ) vede k vyššímu r



Nehomogenní soubor

- může zkreslit r jak směrem nahoru, tak dolů



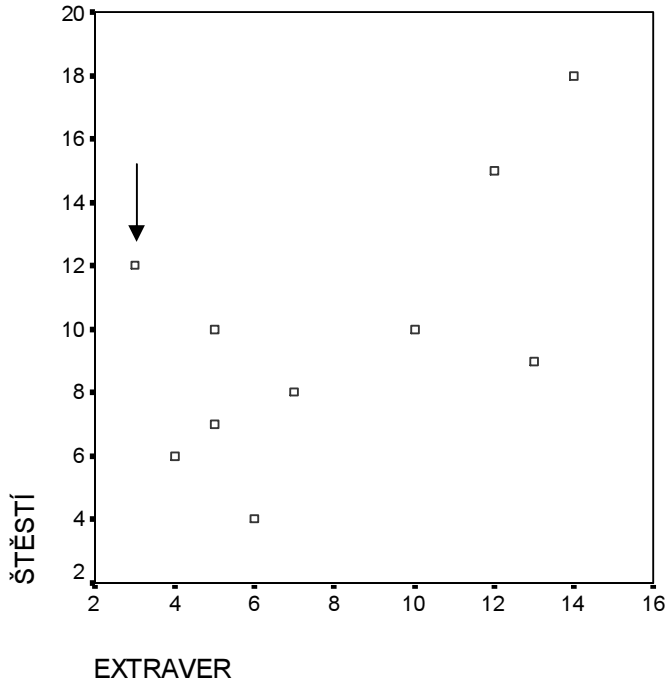


Extrémní hodnoty

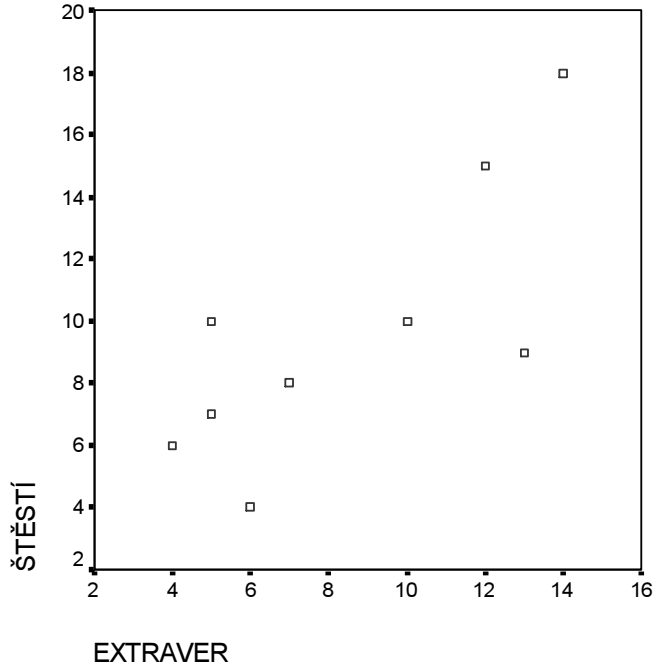
- extrémní hodnoty v jedné nebo obou proměnných mohou r výrazně zkreslit (nejen hodnotu, ale i směr), zvláště když je počet osob v souboru nízký

Extrémní hodnoty

- r= 0,606



- r= 0,766



Neparametrický koeficient

- pro ordinální data je možno spočítat **Spearmanův koeficient pořadové korelace** (ρ)
- počítá se tak, že
  - hodnoty obou proměnných se seřadí od nejnižší po nejvyšší a přidělí se jim pořadí
  - z pořadí se pak počítá Pearsonův koeficient korelace

Parciální korelace

- parciální korelace je taková korelace mezi dvěma proměnnými, kdy kontrolujeme vliv třetí proměnné na obě z nich
- např. chceme zjistit, jaký je vztah mezi prospěchem na SŠ a prospěchem na VŠ; obě proměnné jsou nejspíš ovlivněny IQ

Regresní analýza

- výsledkem regresní analýzy je **matematický model vztahu mezi** dvěma nebo více **proměnnými**
- snažíme se z jedné proměnné nebo lineární kombinace více proměnných predikovat hodnoty další proměnné

Regresní analýza

- dva typy proměnných: **predikovaná** (závislá) **proměnná** a **prediktory** (nezávisle proměnné)
- predikovaná proměnná se označuje **Y**, prediktory $\mathbf{X_1, X_2 \ldots X_n}$
- pouze 1 prediktor – **jednoduchá regrese**
- více prediktorů – **vícenásobná regrese**

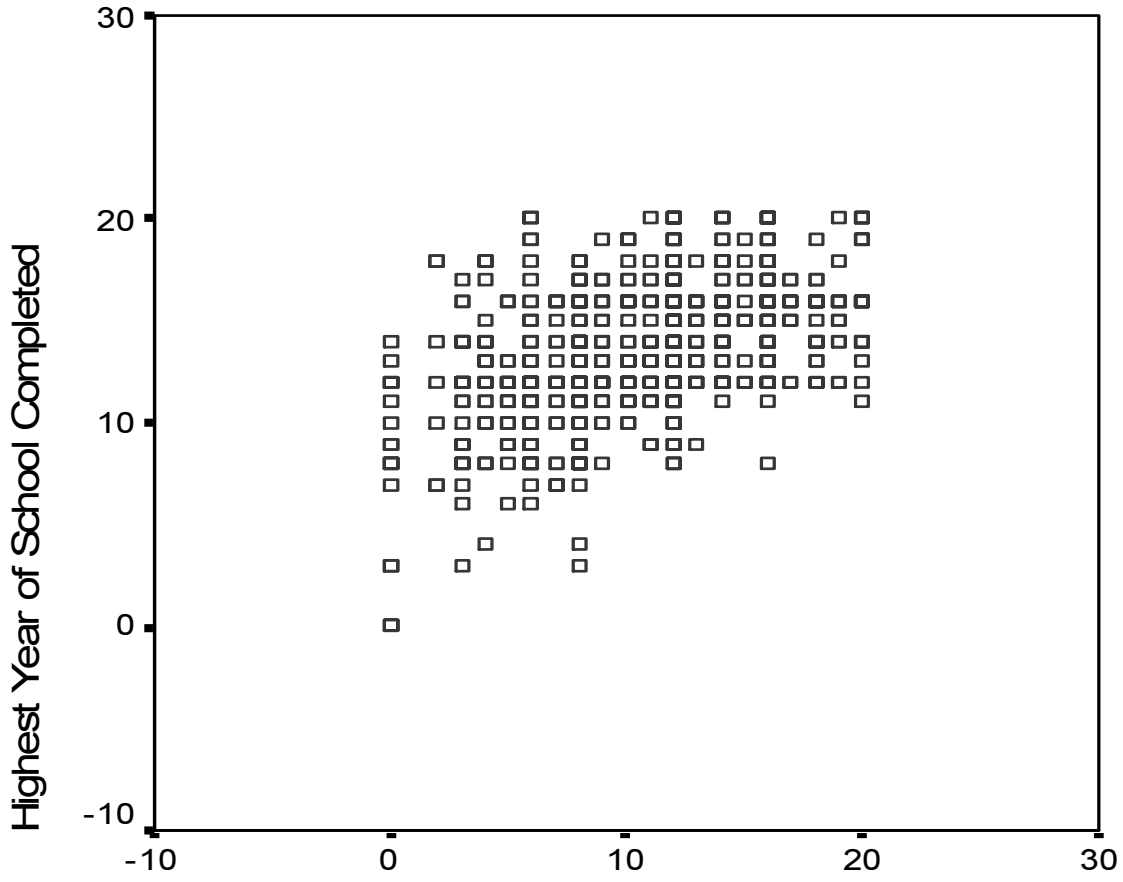
Regresní analýza

- regresní analýza umožňuje
  - porozumět vztahům mezi proměnnými,
  - predikovat hodnoty proměnné Y z hodnot proměnné X (s určitou přesností) – např. z hodnot známek na střední škole nebo z počtu bodů u přijímacího testu předpovědět úspěšnost na VŠ

Jednoduchá regresní analýza

- **příklad** – Jak souvisí vzdělání respondenta se vzděláním otce?
- tj. jak dobře můžeme předpovědět počet let formálního vzdělání respondenta z údaje o počtu let vzdělání jeho otce?

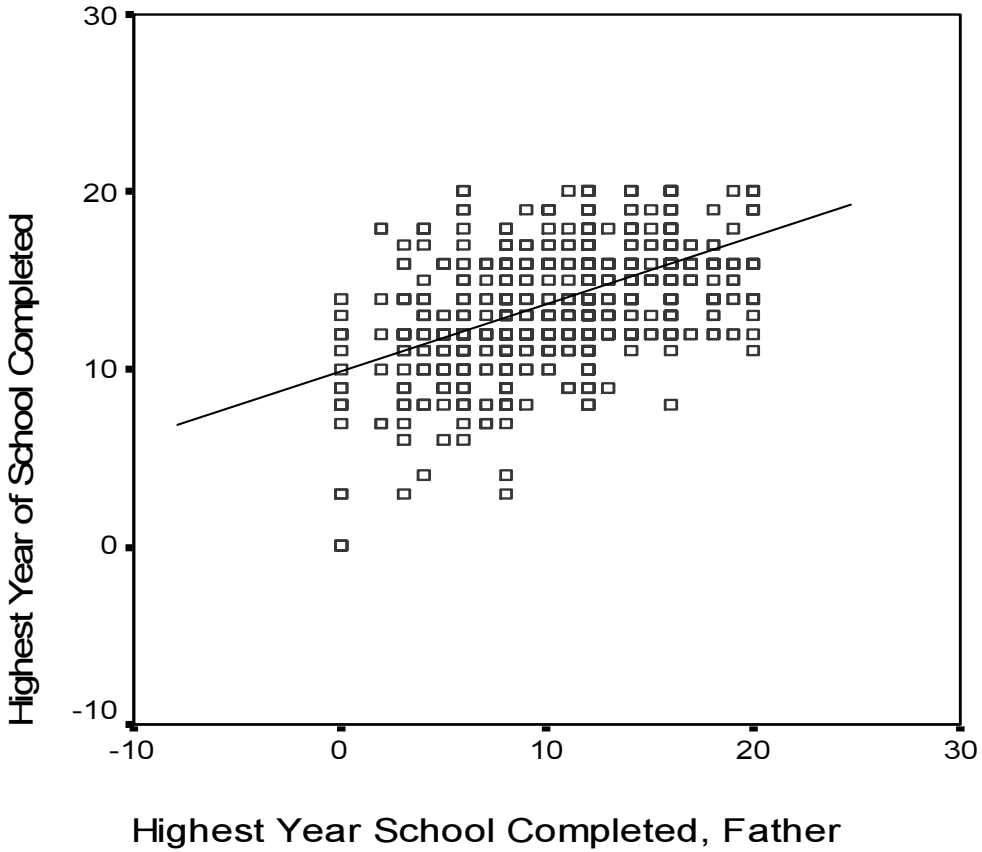
Jednoduchá regresní analýza



Jednoduchá regresní analýza

- snažíme se najít rovnici tzv. regresní přímky
- **regresní přímka** je taková přímka, od které je vzdálenost bodů (představujících naměřená data) co nejménší
- taková přímka, která nejlépe vystihuje data

Jednoduchá regresní analýza



Jednoduchá regresní analýza

- jednou z metod, jak regresní přímku nalézt, je **metoda nejmenších čtverců**
- je zvolena taková přímka, kdy platí, že součet čtverců vzdáleností jednotlivých bodů od přímky je minimální

Jednoduchá regresní analýza

- obecná rovnice regresní přímky

$$Y' = a + bX$$

- **a** je **konstanta** (predikovaná hodnota Y, když hodnota X je 0)
- **b** je **směrnice** regresní přímky (úhel přímky vzhledem k ose; kolikrát se Y zvětší s každou jednotkou X);

Jednoduchá regresní analýza

- v příkladu vychází rovnice regresní přímky ***Y' = 9,93 + 0,32*X***
- pro děti otců s 0 lety vzdělání předpovídáme necelých 10 let vzdělání
- s každým dalším rokem otcova vzdělání předpovídáme o 0,32 roku vzdělání respondenta více
  - např. pro děti otců s 12 lety vzdělání je predikovaná hodnota jejich vlastního vzdělání 13,8 let

Výstup v SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9,926	,219		45,260	,000
	Highest Year School Completed, Father	,322	,019	,463	17,050	,000

a. Dependent Variable: Highest Year of School Completed

Vícenásobná regresní analýza

- predikujeme závislou proměnnou z více prediktorů
- vliv každého z prediktorů na závislou proměnnou je **kontrolován** pro vliv všech ostatních prediktorů (jde tedy o vliv „očištěný od vlivů ostatních proměnných a tudíž počítáme **parciální** koeficienty)

Vícenásobná regresní analýza

- **příklad** – kromě vzdělání otce *(X_1)* může mít na dosažené vzdělání vliv také počet dětí v rodině *(X_2)*
- rovnice regresní přímky je

  $$Y' = a + b_1X_1 + b_2X_2$$

Vícenásobná regresní analýza

- $Y' = 10,68 + 0,30*X_1 - 0,13*X_2$
- vliv vzdělání otce (b=0,30) je o něco menší než u jednoduché regresní analýzy (b=0,32) – je kontrolován pro počet dětí v rodině, který je zřejmě ovlivněn také vzděláním otce
- vliv počtu dětí v rodině je záporný – tj. čím více dětí, tím nižší vzdělání

Vícenásobná regresní analýza

- vícenásobná regresní analýza nám umožní srovnat vliv všech prediktorů na závislou proměnnou
- můžeme dojít k závěru, že větší vliv na vzdělání respondenta má vzdělání otce než počet dětí v rodině?

Vícenásobná regresní analýza

- pokud chceme srovnávat vliv prediktorů měřených v různých jednotkách, je nutné použít tzv. **standardizované regresní koeficienty**
- ukazují, kolikrát vzroste hodnota závislé proměnné, pokud se změní hodnota prediktoru o 1 směrodatnou odchylku a hodnoty ostatních prediktorů přitom zůstanou konstatní

Výstup v SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10,675	,271		39,408	,000
	Highest Year School Completed, Father	,296	,019	,427	15,225	,000
	Number of Brothers and Sisters	-,128	,028	-,129	-4,595	,000

a. Dependent Variable: Highest Year of School Completed

Vícenásobná regresní analýza

- beta pro vzdělání otce je 0,43
- pro počet dětí v rodině -0,13
- větší vliv má tedy vzdělání otce než počet dětí v rodině

Vícenásobná regresní analýza

- kromě regresních koeficientů je počítán také tzv. **koeficient vícenásobné korelace** – korelace všech prediktorů se závislou proměnnou; ozn. **R**
- jde vlastně o korelaci mezi pozorovanými hodnotami závislé proměnné a hodnotami predikovanými na základě regresního modelu

Vícenásobná regresní analýza

- koeficient **vícenásobné determinace** – tzv. % vysvětleného rozptylu (závislé proměnné) lineární kombinací prediktorů; ozn. $\mathbf{R^2}$

Výstup v SPSS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,479[a]	,229	,228	2,512

a. Predictors: (Constant), Number of Brothers and Sisters, Highest Year School Completed, Father

b. Dependent Variable: Highest Year of School Completed

Vícenásobná regresní analýza

- u jednoduché regresní analýzy je **koeficient vícenásobné korelace** roven korelaci mezi oběma proměnnými

Testování hypotéz v regresní analýze

- jsou testovány 2 typy hypotéz
- 1) zda se R průkazně liší od 0
  - testuje se analýzou rozptylu (porovnává rozptyl vysvětlený regresním modelem a reziduální rozptyl)
- 2) zda se regresní koeficienty průkazně liší od 0
  - testuje se t-testem

Výstup v SPSS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1992,838	2	996,419	157,949	,000[a]
	Residual	6693,297	1061	6,308		
	Total	8686,134	1063			

a. Predictors: (Constant), Number of Brothers and Sisters, Highest Year School Completed, Father

b. Dependent Variable: Highest Year of School Completed

Výstup v SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10,675	,271		39,408	,000
	Highest Year School Completed, Father	,296	,019	,427	15,225	,000
	Number of Brothers and Sisters	-,128	,028	-,129	-4,595	,000

a. Dependent Variable: Highest Year of School Completed

Předpoklady regresní analýzy

- skóry v proměnných jsou nezávislé (nejde např. o opakovaná měření)
- dostatečná variabilita všech proměnných
- rozdělení hodnot proměnných je normální
  - u malých výběrů zkontrolovat extrémní hodnoty

Předpoklady regresní analýzy

- vztahy mezi Y a každou X jsou lineární
  - zkontrolovat scatterem
- vzájemné korelace mezi prediktory nejsou příliš vysoké (tzv. problém mulikolinearity)
  - pokud ano, je vhodné buď některou z nich vyřadit, nebo z nich vytvořit např. faktorovou analýzou jeden skór

Předpoklady regresní analýzy

- dostatečně velký počet osob ve výběru vzhledem k počtu prediktorů v modelu

Vybrané multivariační techniky

- faktorová analýza
- shluková analýza

Faktorová analýza

- **cílem** faktorové analýzy (exploratorní) je
  - 1) **redukce dat** – zmenšení počtu proměnných odstraněním nadbytečných proměnných (tj. těsně korelujících s ostatními proměnnými)
  - 2) **idetifikace struktury dat** – prozkoumat vztahy mezi proměnnými

Faktorová analýza

- **výsledkem** faktorové analýzy (exploratorní) je vytvoření několika hypotetických proměnných – **faktorů**
  - někdy bývají nazývány **latentní** proměnné
- faktory jsou lineárními kombinacemi původních proměnných
- vysvětlují vztahy mezi původními proměnnými

Faktorová analýza

- **extrakce** faktorů – na základě matice vztahů mezi proměnnými (např. korelační matice)
- **počet** extrahovaných faktorů – do značné míry závisí na rozhodnutí výzkumníka
  - cílem je vysvětlit co největší množství společného rozptylu co nejmenším počtem faktorů

Faktorová analýza

- **interpretace** faktorů – faktorová analýza sama o sobě nenabídne označení faktorů (to je opět na výzkumníkovi)
- faktor bývá označen na základě proměnných, které k němu mají nejtěsnější vztah (nejvyšší tzv. faktorové **náboje**)

Faktorová analýza

- **rotace** faktorového řešení – usnadní interpretaci faktorů
- rotace může být ortogonální (tj. předpokládá, že faktory jsou nezávislé) nebo šikmá (předpoklad korelace mezi faktory)

Faktorová analýza - příklad

- příklad aplikace FA:
- Osecká, L., Řehulková, O., Macek, P. (1998).
  Zdravotní stesky adolescentů:
  struktura a rozdíly mezi pohlavím.

  Sborník konference Sociální procesy a osobnost, MU Brno.

Faktorová analýza - příklad

- cílem studie bylo mj.vytvořit typologii adolescentů na základě jejich zdravotních obtíží
- adolescenti v dotazníku označili, jak často trpí každou z 18 nabídnutých zdravotních obtíží

Faktorová analýza - příklad

- ☐ bolesti hlavy
- ☐ dýchací potíže
- ☐ žaludeční potíže
- ☐ závratě
- ☐ nechutenství
- ☐ nervozita, neklid
- ☐ nespavost
- ☐ noční můry
- ☐ nesoustředěnost
- ☐ nevolnosti
- ☐ silný tlukot srdce
- ☐ třesení rukou
- ☐ náhlé zpocení
- ☐ průjem, zácpa
- ☐ bolesti v zádech
- ☐ krční bolesti
- ☐ bolesti na prsou
- ☐ bolesti v pánvi

Faktorová analýza - příklad

- typologie na základě 18 proměnných by byla příliš složitá – je třeba tento počet snížit
- autoři spočítali faktorovou analýzu a extrahovali 3 faktory (vysvětlovaly celkem 48% společného rozptylu)

Faktorová analýza - příklad

	F1	F2	F3
nevolnosti	**71**	17	22
nechutenství	**65**	23	10
závratě	**62**	14	30
žaludeční potíže	**60**	-15	**50**
bolesti hlavy	**58**	27	4
nervozita, neklid	**56**	**41**	12
třesení rukou	17	**69**	19
nespavost	38	**63**	-3
náhlé zpocení	-2	**61**	35
silný tlukot srdce	16	**60**	27
nesoustředěnost	37	**54**	4
noční můry	32	**49**	20
bolesti v pánvi	4	28	**69**
průjem, zácpa	21	-9	**65**
bolesti na prsou	16	36	**61**
krční bolesti	16	33	**52**
bolesti v zádech	15	36	**42**
dýchací potíže	**32**	21	**36**
procento rozptylu	***17***	***17***	***14***

Faktorová analýza - příklad

- první faktor nazvali **nevolnosti** – sytily ho především tyto potíže:
  - nevolnosti
  - nechutenství
  - závratě
  - žaludeční potíže
  - bolesti hlavy
  - nervozita, neklid

Faktorová analýza - příklad

- druhý faktor označili **vegetativní obtíže**– sytily ho především položky:
  - třesení rukou
  - nespavost
  - náhlé zpocení
  - silný tlukot srdce
  - nesoustředěnost
  - noční můry

Faktorová analýza - příklad

- třetí faktor označili **bolesti** – sytily ho především tyto potíže:
  - bolesti v pánvi
  - průjem, zácpa
  - bolesti na prsou
  - krční bolesti
  - bolesti v zádech

Faktorová analýza - příklad

- **místo původních 18** proměnných indikujících frekvenci zdravostních potíží měli **nyní 3 proměnné** (lineární kombinace původních proměnných) – nevolnosti, vegetativní potíže a bolesti
- s nimi pak pracovali při typologii (viz dále)

Shluková analýza

- slouží ke **klasifikaci** velkého počtu **osob** (na základě jejich dat v určitých proměnných) **do několika málo shluků**
  - anglické označení cluster analysis se někdy v českých textech překládá také jako clusterová analýza

Shluková analýza

- pro zájemce o podrobnosti o využití metod shlukové analýzy v psychologii doporučujeme publikaci:

  Osecká, L. (2001). Typologie v psychologii. Praha, Academia.

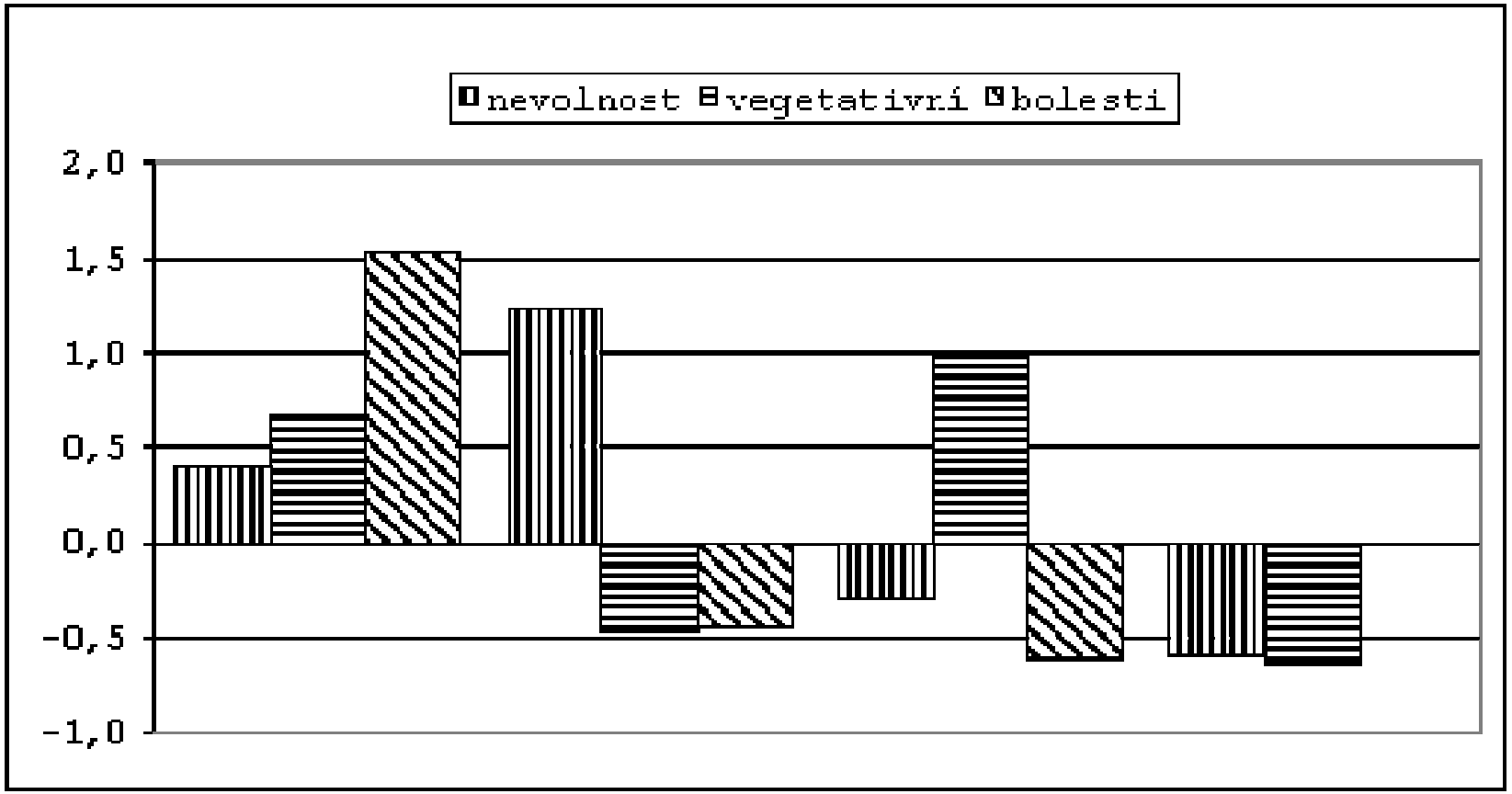
Shluková analýza - příklad

- navazuje na příklad aplikace faktorové analýzy
- autoři se snažili identifikovat skupiny (shluky) adolescentů, kteří jsou si podobní ve svých zdravotních obtížích
- použili 3 proměnné vytvořené na základě FA – nevolnosti, vegetativní potíže a bolesti

Shluková analýza - příklad

- bude uveden **výsledek pro 4 shluky** - v grafu na následujícím snímku jsou průměrná faktorová skóre v použitých 3 proměnných pro osoby klasifikované do daného shluku
  - výsledky pro vyšší počty shluků viz citovaná publikace Osecké, kapitola 14

Shluková analýza - příklad



Shluková analýza - příklad

- osoby v prvním shluku si stěžují především na **bolesti**, zčásti také na vegetativní potíže
- adolescenty ve druhém shluku trápí hlavně **nevolnosti**
- osoby ve třetím shluku trpí **vegetativními obtížemi**
- osoby ve čtvrtém shluku tvořily největší část souboru – trpěly pouze **do určité míry bolestmi** (průměrně), **úroveň ostatních zdravotních stesků u nich byla podprůměrná**