Aplikovaná matematická statistika

Mgr. Martin Sebera, Ph.D.

Fakulta sportovních studií Masarykovy univerzity

Leden 2014

1. Úvod ... 3
2. Základní a výběrový soubor ANEB generování náhody je příliš důležité, než abychom ji mohli ponechat náhodě ... 4
3. Bodové a intervalové rozložení četností ANEB histogram není hysterie ... 8
4. Základní statistické charakteristiky ANEB není střední hodnota jako střední hodnota ... 12
5. Testování hypotéz, koncept věcné vs. statistické významnosti ANEB 0,05 nevládne ... 16
6. Testy normality ANEB normální rozdělení není až tak normální ... 20
7. Testy o rovnosti středních hodnot dvou výběrů ANEB t-testy nejsou protesty ... 25
8. Korelace ANEB korelace není kauzalita ... 29
9. Regresní analýza ANEB regrese mohla být reverse ... 34
10. Analýza rozptylu ANEB ANOVA-MANOVA-MANCOVA ... 40
11. Faktorová analýza ... 46
12. Závěr ANEB Statistický rozcestník ANEB co s daty ... 51
13. Použité zdroje ... 53
14. Anglicko-český slovník ... 54

Seznam tabulek

Tab. 1 Příklady typů proměnných ... 5
Tab. 2 Počty členů Českého atletického svazu v roce 2012 ... 5
Tab. 3 Bodové rozdělen četností ... 8
Tab. 4 Intervalové rozdělení četností ... 9
Tab. 5 Základní statistické charakteristiky ... 13
Tab. 6 Testování hypotéz ... 16
Tab. 7 Vybrané effect size koeficienty ... 17
Tab. 8 Výsledek t-testu, samostatný vzorek ... 25
Tab. 9 Data pro t-test, závislá pozorování ... 26
Tab. 10 Výsledky t-testu, závislá pozorování ... 26
Tab. 11 Test normality ... 26
Tab. 12 Wilcoxonův test ... 27
Tab. 13 Hodnoty jednoduchých korelačních koeficientů ... 31
Tab. 14 Hodnoty parciálních korelačních koeficientů ... 32
Tab. 15 Využití Pearsonova korelačního koeficientu (... 32
Hendl, 2004), p. 266 ... 32
Tab. 16 Logaritmická regrese ... 36
Tab. 17 Hyperbolická regrese ... 37
Tab. 18 Mocninná regrese ... 37

Tab. 19 ANOVA – popisné statistiky 41
Tab. 20 ANOVA - testy homogenity rozptylu 43
Tab. 21 ANOVA 43
Tab. 22 ANOVA – post-hoc testy (faktor věk) 43
Tab. 23 ANOVA – post-hoc testy (faktor pohlaví) 43
Tab. 24 ANOVA – post-hoc testy (interakce faktorů věk a pohlaví) 44
Tab. 25 Tabulka vlastních čísel u faktorové analýzy 47
Tab. 26 Výsledek faktorové analýzy 48

Seznam obrázků

Obr. 1 Náhodný výběr 4
Obr. 2 Stratifikovaný výběr 4
Obr. 3 Znázornění náhodné a systematické chyba 6
Obr. 4 Histogram 10
Obr. 5 Gaussova křivka normálního rozdělení 20
Obr. 6 Studentovo t-rozdělení 21
Obr. 7 Pearsonovo χ^2-rozdělení 21
Obr. 8 Fischerovo F-rozdělení 22
Obr. 9 Ověření normality 23
Obr. 10 Korelace – bodové grafy 31
Obr. 11 Bodový graf logaritmické regrese 35
Obr. 12 Bodový graf hyperbolické regrese 37
Obr. 13a Graf analýzy rozptylu 42
Obr. 13b Graf analýzy rozptylu 42
Obr. 14 Scree graf 48
Obr. 15 3D graf faktorů 49

1. Úvod

Předložený studijní materiál hodlá sloužit studentům a vědeckým pracovníkům k pochopení základních i rozšiřujících statistických metod vhodných k analýze dat v kinantropologickém výzkumu. Tento studijní materiál již předpokládá jistou znalost základních statistických pojmů. Přesto, pokud si čtenář nebude jistý významem termínů nebo probíranou problematikou, studijní materiál mu nabízí formou externích odkazů link na vysvětlení daného problému. Za základní literaturu považujeme knihu prof. Hendla Přehled statistických metod zpracování dat (Hendl, 2004), proto v seznamu zdrojů u každé kapitoly uvádíme i odkaz na konkrétní strany v této knize. Studijní materiál se úmyslně snaží zjednodušovat jednotlivé statistické metody, ač je zřejmé, že k jejich použití je nezbytné znát širší a podrobnější souvislosti.

Příklady zde použité pocházejí mnohdy z reálných výzkumů, někdy jsou data používána k doplnění výkladu dané statistické metody. Naším cílem je předložit studentům spíše materiál encyklopedického charakteru než čtivou beletrii o statistice. Tak, aby se čtenář vracel k jednotlivým kapitolám, podle svého aktuálního problému analýzy dat. Součástí příkladů je i řešení v sw Statistica firmy Statsoft, verze 12 CZ. Ačkoliv předkládáme učební test o statistice, nikoliv o řešení v software Statistica, obsahují řešení příklady i postup, jak se postupnými kroky dostat k požadovaným výsledkům a to konkrétně v tomto sw. **Tento postup je dále v textu graficky odlišen zelenou barvou.** Zároveň využíváme rozsáhlého elektronického manuálu firmy Statsoft, a v použitých zdrojích uvádíme link na relevantní stránky věnující se probírané tématice.

Na konci každé kapitoly jsou dodány další odkazy na anglické zdroje, které se zabývají danou problematikou. A to z důvodu, že někdy je dobré znát i anglickou terminologii vybraných statistických pojmů. Internet obsahuje mnoho zajímavých souhrnů, manuálů, učebnic a studijních materiálů z oblasti statistiky, ze kterých lze čerpat inspiraci. Na konci textu nabízíme studentovi anglicko-český slovník vybraných statistických pojmů.

K e-learningovému zpracování.

- Snažíme se držet základních doporučení, které má každý e-learningový materiál obsahovat. Velmi bojujeme s odhadem časových nároků na každou dílčí kapitolu, protože i samotným autorům trvá mnohdy méně, mnohdy více, než kapitolou projdou. Proto časový odhad neuvádíme.
- Uvědomujeme si, že statistika jako věda s matematickým základem, nemusí být vždy studenty oblíbená. Snažíme se v nadpisech kapitol kromě správného užití (odborně, terminologicky, ale i spisovně) nabídnout i odlehčenou variantu názvu kapitoly, která mnohdy více přiblíží studentovi probíranou problematiku nebo stručně popíše, kde je tzv. jádro pudla (= ustálené spojení, které použil ve svém díle Faust německý básník Johann Wolfgang Goethe)
- Každá kapitola je doplněna o externí linky na jiné www stránky s probíranou problematikou.
- Kontrolní otázky pak zjišťují základní pochopení dané kapitoly.

2. Základní a výběrový soubor ANEB generování náhody je příliš důležité, než abychom ji mohli ponechat náhodě

teorie

Statistika se zabývá hromadnými jevy. Jev se může mnohokrát opakovat. Pokud jev několikrát zopakujeme, přestává jej ovlivňovat vliv jedinečnosti zkoumaného objektu. Proto lze zkoumat u takových jevů zákonitosti a vztahy. Jednotlivé prvky se nazývají statistické jednotky, u nich sledujeme statistické veličiny (proměnné). Soubor veličin pak nazývejme data.

Základním souborem jsou tedy všechny statistické jednotky. **Výběrovým** souborem je (jakýmsi způsobem) vybraná část základního souboru. Naší snahou je najít takový výběrový soubor, jehož vlastnosti by nejvíce odpovídaly souboru základnímu.

Výběr může být:

- **náhodný** (losování, hod kostkou, generátory náhodných čísel). Jinými slovy je to takový výběr, kde každý prvek má stejnou pravděpodobnost, že bude vybrán
- **systematický**. Vybereme každý *n*-tý objekt, kde *n* získáme jako podíl velikosti základního souboru a velikosti výběrového souboru. Pokud hned první prvek vybereme náhodně, mluvíme o systematickém výběru
- **stratifikovaný**. Základní soubor rozdělíme podle předem jasně definovaných kritérií a poté v podskupinách postupujeme náhodným výběrem

Obr. 1 Náhodný výběr

Obr. 2 Stratifikovaný výběr

(http://alik.idnes.cz/mesto-nebo-kos-na-odpadky-0h0-/alik-alikoviny.asp?c=A090526_221914_alik-alikoviny_jtr, 2013)

(http://www.tezas.sk/index.php?pc=2, 2013)

Typy proměnných

Máme nadefinovaný **statistický soubor**, což je množina statistických jednotek. **Statistické jednotky** jsou prvky statistického souboru s minimálně jednou shodnou vlastností. A **statistický znak** je společná vlastnost několika jednotek statistického souboru. Rozdělme statistické znaky na **kvalitativní** (vyjádřena slovně) a **kvantitativní** (vyjádřena číslem). Kvantitativní znaky pak můžeme dělit na **spojité** (hodnota může být jakékoliv reálné číslo z daného intervalu) a **diskrétní** (jen celočíselné hodnoty). Můžeme-li proměnnou vyjádřit číslem, bude to proměnná **nominální**. Můžeme-li proměnné seřadit podle určitého znaku, pak se bude jednat o **ordinální** proměnné. Pokud u ordinálních proměnných můžeme konstatovat, o kolik se hodnoty liší, pak je nazýváme **intervalové**. Můžeme-li u ordinálních proměnných říct kolikrát je hodnota vyšší, pak je nazýváme **poměrové**.

Můžeme-li proměnnou zařadit do tříd, nazýváme ji **kategoriální**. Speciálním případem kategoriální proměnné je **dichotomická** proměnná, která nabývá jen dvou hodnot. V následující tabulce uvádíme typy proměnných společně s příklady.

Tab. 1 Příklady typů proměnných

typ proměnné	
spojitá	teplota, tlak, rosný bod, čas, délka, hmotnost
diskrétní	počet lidí
nominální	typ temperamentu, pohlaví, název výrobku
ordinální	výkon v běhu na 100 m, známky ve škole, cena výrobku
intervalové	délka, hmotnost, čas, rychlost, zrychlení
poměrová	délka, hmotnost, čas, rychlost, zrychlení
kategoriální	typ zaměstnání: sedavé, fyzické, fyzicko-sedavé. způsob dopravy do zaměstnání: pěšky, na kole, MHD, auto, vlastní doprava
dichotomická	pohlaví: žena, muž pravda, lež

Proměnné můžeme vzájemně transformovat. Např. proměnná BMI (body mass index)

hodnota BMI (nominální, spojitá, ordinální)

↓

podváha: < 18,5; ideální váha: 18,5–25; nadváha: 25,1–30; obezita: >30,1 (kategoriální)

↓

neobézní, obézní (dichotomická)

příklady

- základní soubor: všichni atleti v ČR, což je 56.874 členů Českého atletického svazu k 15. 2. 2013
- náhodný: z abecedního seznamu všech jmen použitím generátoru náhodných čísel (např. http://randomnumbergenerator.intemodino.com/cz/) nechám vygenerovat náhodná čísla, podle kterých provedu výběr
- systematický výběr. Pokud chceme vybrat 1000 atletů, pak náhodně vybereme první jméno v abecedním seznamu a poté každé 56 (56.874 / 1.000 = 56)
- stratifikovaný výběr. Rozdělíme Českou republiku podle krajů (Tab. 2) a v každém kraji provedeme náhodný výběr

zdroj: Výroční zpráva Českého atletického svazu 2012, http://www.atletika.cz/o-nas/publikace/

Tab. 2 Počty členů Českého atletického svazu v roce 2012

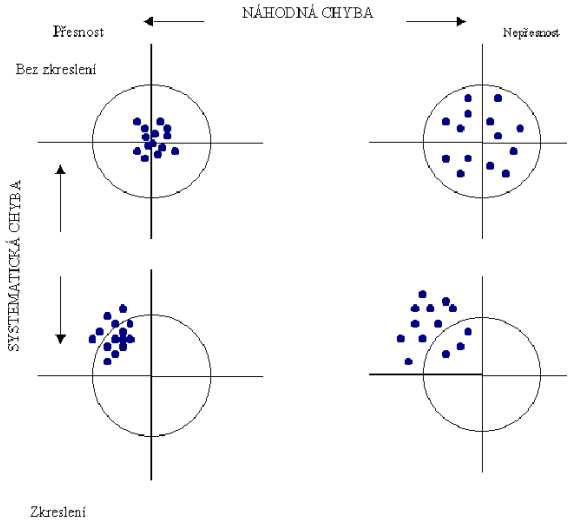
Praha	18 732
Středočeský	4 063
Jihočeský	2 438
Plzeňský	2 268
Karlovarský	1 108
Ústecký	2 492
Liberecký	3 574
Královéhradecký	2 573
Pardubický	3 021
Vysočina	1 717
Jihomoravský	5 553

Olomoucký	3 175
Moravskoslezský	4 640
Zlínský	1 520

shrnutí

Lidé, kteří se nezabývají ani vědeckými ani statistickými metodami se občas diví, jak je možné z náhodnosti utvářet reprezentativní závěry? Neboli se tvrdí, a navíc s propracovaným statistickým aparátem, že něco platí a přitom je to vše postaveno na náhodě? Ano, ale tuto otázku zodpovíme jednoduše. Platí totiž podmínka, že předem známe pravděpodobnosti našich jevů. Pak můžeme využít matematického aparátu, kde s rostoucím počtem měření výsledky konvergují ke skutečné hodnotě (Řezánková, Marek, & Vrabec, 2000).

Provést výběr dat, aby byl opravdu náhodný, je však v praxi složité. Pokud metodika výzkumu přesně neurčuje postup sběru dat, můžeme se dopustit náhodné nebo systematické chyby. Následující obrázek ilustruje, jaký je rozdíl mezi náhodnou a systematickou chybou



Obr. 3 Znázornění náhodné a systematické chyba

zdroj: http://ucebnice.euromise.cz/index.php?conn=0&section=epidem&node=node20, 2013

odkazy na další studijní zdroje

Easton, V. & McColl, J. (1997). *Statistics Glosary*. Retrieved July, 5, 2013, from http://www.stats.gla.ac.uk/steps/glossary/basic_definitions.html

Wikipedia. Retrieved July, 5, 2013, from http://en.wikipedia.org/wiki/Sampling_(statistics)

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Elementary-Statistics-Concepts/button/1

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 37-46.

kontrolní otázky

Určete typ proměnné: tepová frekvence
a) spojitá
b) diskrétní
c) dichotomická

Určete typ proměnné: temperament (sangvinik, cholerik, melancholik, flegmatik)
a) spojitá
b) ordinální
c) kategoriální

Je proměnná „pohlaví“ ordinální?
a) ano
b) ne

3. Bodové a intervalové rozložení četností ANEB histogram není hysterie

teorie

Prvním krokem, který většinou provedeme při náhledu na získaná data, je zjištění rozložení četností znaků. V rámci tohoto postupu získáme hned několik důležitých informací o našich datech. Mezi ně patří informace o chybějících datech a datech, která můžeme považovat za odlehlá od běžných či očekávaných hodnot. Z grafu četností můžeme odhadnout, zda data pocházejí z normálního rozdělení, což nám umožní vybrat následný postup.

Pro zjištění rozložení četností vytvoříme tabulku absolutních a relativních četností a k nim příslušné kumulativní četnosti (absolutní a relativní). V rozsáhlých datových souborech můžeme zkonstruovat intervalové rozdělení četností, které zpřehlední naše data. Označme: N – rozsah souboru. Dolní index $_i$ značí příslušnost k *i*-té skupině, n_i – absolutní četnost, r_i – relativní četnost, N_i – kumulativní absolutní četnost, F_i – kumulativní relativní četnost. Relativní četnost (též procentuální zastoupení) stanovíme vzorcem $f_i = \frac{n_i}{N}$. Lépe postup vysvětlíme na názorném příkladu, kde k tabulce přidáme i několik grafů, které pomohou s vizualizací dat.

příklady

Příklad 1

Máme 20 hodnot, ze kterých provedeme bodové rozdělení četností a stanovení absolutních a relativních četností a k nim příslušejících kumulativních četností.

Data: 18 19 19 20 20 20 20 20 20 20 20 21 21 21 21 21 21 22 22 22

Tab. 3 Bodové rozdělen četností

X	n_i	r_i	N_i	F_i
18	1	0,05 (= 1/20)	1	0,05
19	2	0,10 (= 2/20)	3	0,15
20	8	0,40 (= 8/20)	11	0,55
21	6	0,30 (= 6/20)	17	0,85
22	3	0,15 (= 3/20)	20	1,00
Celkem	20	1,00		

Příklad 2

Máme 93 hodnot z měření BMI (body mass index)

17,9 19,2 19,3 19,6 19,6 19,7 19,8 20,1 20,3 20,3 20,4 20,9 20,9 21,1 21,1 21,1 21,4 21,6 21,6 21,6 21,8 21,9 22,1 22,2 22,2 22,3 22,3 22,4 22,6 22,7 22,8 22,8 22,9 23,0 23,1 23,1 23,2 23,3 23,3 23,4 23,4 23,4 23,6 23,7 23,8 23,9 23,9 23,9 24,0 24,1 24,1 24,1 24,3 24,4 24,4 24,5 24,5 24,5 24,7 24,8 24,9 24,9 25,0 25,1 25,1 25,1 25,1 25,2 25,3 25,3 25,4 25,5 25,6 25,7 25,8 25,9 26,3 26,3 26,5 26,8 26,9 26,9 27,1 27,7 28,0 28,6 29,2 29,4 29,4 29,4 29,4 29,7 30,0

Sestavíme tabulku četností, kumulativních četností, relativních a kumulativních relativních četností

Statistiky → Základní statistiky → Tabulky četností

Zde by provést bodové rozdělení četností znamenalo vytvořit velkou a nepřehlednou tabulku, která by nám neposkytla žádné zajímavé informace. Vytvoříme proto intervaly a četnosti budeme sledovat uvnitř těchto intervalů.

Existuje mnoho způsobů, jak nastavit počet, resp. šířku intervalů. Např. tzv. Sturgesovo pravidlo navrhuje, aby počet intervalů byl roven hodnotě *k,* které se vypočítá přibližně jako *1 + 3.3 log n* (log je logaritmus ☺). V našem případě k ≈ 7,6. Takže bychom mohli mít 7 nebo 8 intervalů. Šířka intervalů se pak dodatečně určí, abychom pokryli všechny hodnoty. Šířka intervalů je shodná přes celé variační rozpětí.

SW Statistica nabízí několik možností nastavení počtu intervalů a to automaticky i ručně. Zajímavou nabídkou je „pěkné intervaly“, v anglické mutaci sw „neat intervals“, kdy zaokrouhluje hranice intervalů na desetinná čísla s poslední číslicí 0, 1 nebo 5 a to z důvodu snazší interpretace. Výsledkem je tabulka 4:

Tab. 4 Intervalové rozdělení četností

OD–DO	Tabulka četností: BMI			
	Četnost	Kumulativní četnost	Rel. četnost	Kumulativní rel. četnost
16 <x<=18	1	1	1,07527	1,0753
18 <x<=20	6	7	6,45161	7,5269
20 <x<=22	15	22	16,12903	23,6559
22 <x<=24	27	49	29,03226	52,6882
24 <x<=26	27	76	29,03226	81,7204
26 <x<=28	9	85	9,67742	91,3978
28 <x<=30	8	93	8,60215	100,0000
30 <x<=32	0	93	0,00000	100,0000
32 <x<=34	0	93	0,00000	100,0000
ChD	0	93	0,00000	100,0000

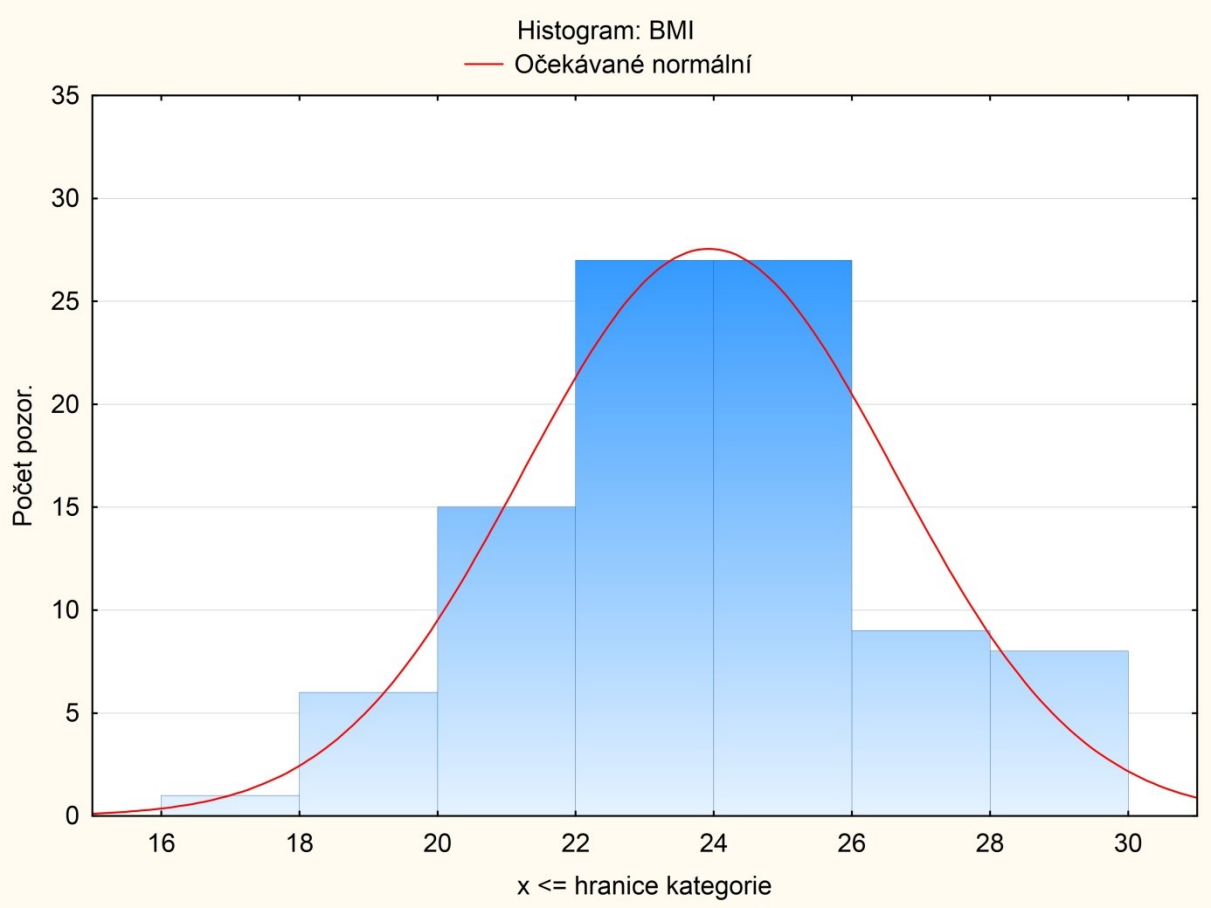
!!!! tady bych chtěl do tabulky zaznačit červené šipky pro ilustraci, jak se počítají kumulativní charakteristiky, viz obrázek níže !!!!

OD DO	Tabulka četností: BMI			
	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
16 <x<=18	1	1	1,07527	1,0753
18 <x<=20	6	7	6,45161	7,5269
20 <x<=22	15	22	16,12903	23,6559
22 <x<=24	27	49	29,03226	52,6882
24 <x<=26	27	76	29,03226	81,7204
26 <x<=28	9	85	9,67742	91,3978
28 <x<=30	8	93	8,60215	100,0000
30 <x<=32	0	93	0,00000	100,0000
32 <x<=34	0	93	0,00000	100,0000
ChD	0	93	0,00000	100,0000

Z tabulky můžeme vyčíst několik informací. Např. všech záznamů je celkem 93. Počet záznamů větších než 26 a menších rovno 28 je celkem 9, což tvoří 9,67 % všech hodnot. Pokud bychom data seřadili od nejnižší po nejvyšší, tak tento interval obsahuje 77. až 85. hodnotu a v pořadí vyjádřené procenty to je cca o 81,7 do 91,4 %. Kumulativní četnosti sečítají všechny předchozí četnosti společně s aktuální. Schematicky je výpočet naznačen v Tabulce 4.

Ke grafickému znázornění lze využít nejčastěji histogramu, což je graf, kde na ose X jsou vyneseny jednotlivé intervaly a na ose Y pak četnosti příslušejícím danému intervalu. Histogram může být doplněn ideální křivkou normálního rozdělení, kdy na základě podobnosti této křivky a

histogramu můžeme usuzovat, zda data pocházejí z normálního rozdělení (Obr. 4). Jde však jen o prvotní odhad, normalitu pak musíme testovat pomocí hypotéz, což bude probíráno v příštích kapitolách.



Obr. 4 Histogram

shrnutí

Tabulka absolutních i relativních četností pomáhá výzkumníkovi při prvním prozkoumání získaných dat. Společně s grafickým vyjádřením (např. pomocí histogramu) lze usuzovat na některé vlastnosti, které z dat vyplývají. Ale i tento relativně jednoduchý početní úkon v sobě skrývá záludnosti. Už jen tím, že záleží, jak široké nebo kolik intervalů zvolíte. Rázem se může tvar histogramu změnit…

odkazy na další studijní zdroje

Wikipedia. (2013). Retrieved July, 12, 2013, from http://en.wikipedia.org/wiki/Frequency_histogram

Green, L. (2008). *Elementary Statistics (Math 201)*. Retrieved September, 30, 2013, from http://www.ltcconline.net/greenl/courses/201/descstat/hist.htm

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Basic-Statistics#frequency%20tables

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 85–92.

kontrolní otázky

Data: 2 5 8 5 6 8 5 4 5 4 5 8 2 1 0 1 0 1 2 5 4 5
Která hodnota z následujících dat má největší absolutní četnost?
a) 2
b) 4
c) 5
d) 8

U které z hodnot ze stejných dosáhne kumulativní relativní četnost hodnoty 50 %?
a) 2
b) 4
c) 5
d) 8

Data představují body získané u písemného testu ze statistiky. Kolik studentů získalo bodové hodnocení 5 a vyšší?
a) 9
b) 10
c) 11
d) 14

Data představují body získané u písemného testu ze statistiky. Kolik procent studentů získalo bodové hodnocení 4 a nižší?
a) 36
b) 50
c) 51
d) 81

4. Základní statistické charakteristiky ANEB není střední hodnota jako střední hodnota

teorie

Základními statistickými charakteristikami rozumíme čísla, která nám o našich datech podávají určitou informaci. Dále je můžeme použít pro srovnání více souborů dat. Tyto charakteristiky můžeme rozdělit do několika skupin

charakteristiky:

- úrovně
    - střední hodnoty
        - Aritmetický (vážený) průměr – součet všech hodnot vydělených počtem hodnot. Často používaná charakteristika, která se snadno vypočítá, ale její intepretace je mnohdy nepřesná. Pokud je někde uváděn aritmetický průměr, vždy s ním musí být uvedena směrodatná odchylka a počet pozorování N. Jinak může dojít velmi lehce k výraznému zkreslení informací o datech.
        - Geometrický průměr – součin všech hodnot a odmocněn *n*-tou odmocninou. Používá se při analýze řad, časových řad, k identifikaci míry růstu nebo poklesu.
        - Modus – hodnota s nejčastějším výskytem
        - Medián – taková hodnota, která v uspořádaných datech podle velikosti, představuje střed a dělí tak data na dvě poloviny o stejném počtu hodnot. První polovina je ve svých hodnotách menší rovna mediánu, druhá polovina je ve svých znacích větší rovna hodnotě mediánu. Jedná se o střední hodnotu.
- variability
    - variační rozpětí – rozdíl mezi maximem a minimem
    - kvantily – dělí řadu hodnot na stejné části. Předpokládá se, že hodnoty jsou seřazeny od nejnižší po nejvyšší hodnotu.
        - kvartily – rozdělují hodnoty na 4 části. Dolní kvartil se nachází v první čtvrtině, horní kvartil ve třetí čtvrtině.
        - percentily – dělí řadu hodnot na 100 částí.
    - rozptyl – součet čtverců odchylek od aritmetického průměru vydělený počtem hodnot. Informuje o homogenitě hodnot, neboli, jak moc jsou hodnoty rozptýleny od aritmetického průměru.
    - směrodatná odchylka – odmocnina z rozptylu
    - variační koeficient – podíl aritmetického průměru a směrodatné odchylky. Umožňuje srovnat variabilitu souborů s nestejnými jednotkami.
    - směrodatná chyba průměru (střední chyba průměru) se vypočítá jako směrodatná odchylka vydělená odmocninou z n, kde n je počet hodnot. Tato charakteristika vyjadřuje rozptyl aritmetického průměru v souboru

Používáním jednotlivých statistických charakteristik ztrácíme mnoho cenných informací o původních datech. Tato skutečnost je jednou ze slabých mís používání aritmetického průměru. Na

jednoduchém příkladu níže ukážeme, že pokud už musíme původní data nahradit jejich statistickými charakteristikami, měli bychom to provádět s rozvahou.

příklady

Příklad 1

data: 1; 10; 22 průměr 11 směrodatná odchylka 10,53 n = 3
11; 11; 11 průměr 11 směrodatná odchylka 0 n = 3

Na tomto příkladu je zřejmé, jak aritmetický průměr, pokud by byl uveden samostatně, nevyjadřuje přesné informace o původních datech.

Příklad 2

Ve 2. čtvrtletí 2013 byla průměrná měsíční mzda 24 953,- Kč (vypočítáno pomocí aritmetického průměru) a 20 944,- Kč (vypočítáno pomocí mediánu)
Zdroj.: Český statistický úřad, http://www.czso.cz/csu/csu.nsf/informace/cpmz090613.doc

Medián představuje střední hodnotu, která není ovlivněna extrémními hodnotami (ať už maximy nebo minimy). Rozdíl 4 000,- Kč je značný, vždyť je to cca pětina průměrné mediánové mzdy.

Tento příklad ukazuje, jaký mocný nástroj dává statistika do rukou svému uživateli a jak jednoduše ji lze využít/zneužít pro zkreslenou interpretaci dat. Problém případných diskutujících, kteří se vzájemně přesvědčují, kdo z nich má pravdu při debatách o střední hodnotě, spočívá v tom, že většinou neuvádějí, jak ke střední hodnotě dospěli a kterou při výpočtech použili…

Příklad 3

Máme 93 hodnot z měření BMI

17,9 19,2 19,3 19,6 19,6 19,7 19,8 20,1 20,3 20,3 20,4 20,9 20,9 21,1 21,1 21,1 21,4 21,6 21,6 21,6 21,8 21,9 22,1 22,2 22,2 22,3 22,3 22,4 22,6 22,7 22,8 22,8 22,9 23,0 23,1 23,1 23,2 23,3 23,3 23,4 23,4 23,4 23,6 23,7 23,8 23,9 23,9 23,9 24,0 24,1 24,1 24,1 24,3 24,4 24,4 24,5 24,5 24,5 24,7 24,8 24,9 24,9 25,0 25,1 25,1 25,1 25,1 25,2 25,3 25,3 25,4 25,5 25,6 25,7 25,8 25,9 26,3 26,3 26,5 26,8 26,9 26,9 27,1 27,7 28,0 28,6 29,2 29,4 29,4 29,4 29,4 29,7 30,0

Vypočítejte základní statistické charakteristiky

Statistiky – Základní statistiky – Popisné statistiky – Detailní výsledky

Tab. 5 Základní statistické charakteristiky

N platných	93
Průměr	23,92903
Geometrický (Průměr)	23,78010
Medián	23,90000
Modus	Vícenás.
Četnost (modu)	4
Minimum	17,90000
Maximum	30,00000
Dolní (kvartil)	22,20000
Horní (kvartil)	25,30000
Rozptyl	7,250561
Sm. odch.	2,692687
Var. koef.	11,25280
Směrod. (chyba)	0,279219

shrnutí

Základní statistické charakteristiky představují poměrně vypracovaný pohled na data. Jedná se jen o tzv. jednorozměrné posouzení, zatím bez hledání závislostí a souvislostí. Pro prvotní posouzení kvality dat se jedná o zásadní krok, kterým začíná komplexní analýza dat.

Prakticky má význam tohoto kroku např. u hledání extrémních hodnot, které mohly vzniknout při opisování a přepisování dat do elektronické podoby. Včasná detekce případných chyb (i jinak přesný přístroj může chybou špatné kalibrace nebo nedodržením příslušné metodiky generovat data, která nemusí odpovídat realitě) je základem další analýzy dat.

odkazy na další studijní zdroje

Butterfield, A. E. (2013). *Descriptive Statistics*. Retrieved September, 22, 2013, from http://www.che.utah.edu/~tony/course/material/Statistics/12_descriptive.php

Green, L. (2008). *Elementary Statistics (Math 201)*. Retrieved September, 30, 2013, from http://ltcconline.net/greenl/courses/201/descstat/mean.htm

Emath zone (2013). *Mean Deviation and its Coefficient*. Retrieved September, 11, 2013, from http://www.emathzone.com/tutorials/basic-statistics/mean-deviation-and-its-coefficient.html

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Basic-Statistics#Descriptive%20statistics

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 93-113.

kontrolní otázky

Co znamená, pokud je směrodatná odchylka větší než aritmetický průměr?
a) data jsou poměrně homogenní
b) data pocházejí z normálního rozdělení
c) data jsou poměrně vzdálena od aritmetického průměru
d) nejedná se o heterogenitu dat

Jaká může být minimální směrodatné odchylka?
a) -1
b) 0
c) 1
d) 100

Jaká může být maximální směrodatné odchylka?
a) -1
b) ∞
c) 1
d) 100

Medián je
a) jiný název pro geometrický průměr
b) 50. kvantil
c) nejčastěji se vyskytující hodnota v datech
d) střední hodnota, která je vždy menší než aritmetický průměr

5. Testování hypotéz, koncept věcné vs. statistické významnosti ANEB 0,05 nevládne

teorie

Ve statistice (a připomeňme, že je to aplikovaná matematická věda) platí jen to, co jsme schopni doložit výpočtem. Konkrétně pro statistiku je typické testování hypotéz. Co je posléze danou hypotézou zamítnuto, o tom vlastně tvrdíme, že to neplatí. Co není statisticky významné, jakoby neexistovalo. A tak používáme koncept testování hypotéz pro rozhodování, jak dále nakládat s daty.

Postup testování hypotéz je poměrně jasný a jednoduchý. Vytvoříme hypotézu H_0, o které předpokládáme, že platí. Proti ní postavíme alternativu (H_A, což je obvykle naše výzkumná hypotéza). Ke každému našemu tvrzení, které tvoří prvotní myšlenku při výzkumu, sesbíráme data. A nyní potřebujeme najít věrohodný aparát, který nám pomůže při konstatování, zda domněnka platí nebo ne. Tímto aparátem bude statistický test.

Výsledkem testování jsou 2 možnosti, resp. 3 alternativy

- testování jsme provedli správně, výsledkem je tvrzení: hypotézu zamítneme nebo nezamítneme
- dopustili jsme se chyby
    - zamítli jsme hypotézu, která platí. Dopustili jsme se chyby 1. druhu, která se značí α a nazývá se **hladina významnosti** testu. Výraz 1 - α se nazývá pak **spolehlivost**.
    - přijali jsme hypotézu, která neplatí. Nastala chyba 2. druhu, značí se β. Výraz 1 - β se nazývá síla testu.

Obvyklé hodnoty pro spolehlivost jsou 0,95 nebo 0,99 pro sílu testu pak 0,8 nebo 0,9. Z čehož vyplývá, že můžeme (ale nemusíme) zvolit hladinu významnosti 0,05 nebo 0,01.

Tab. 6 Testování hypotéz

		výsledek testu	
		hypotéza H_0 platí	hypotéza H_A platí
reálná situace	hypotéza H_0 platí	správné rozhodnutí	chyba 1. druhu
	hypotéza H_A platí	chyba 2. druhu	správné rozhodnutí

To, že hypotézu H_0 nezamítáme, neznamená, že platí. Stejně jako u soudu se držíme tzv. presumpce neviny (Statsoft, Newsletter 10/12/2012. Retrieved from http://www.statsoft.cz/o-firme/archiv-newsletteru/newsletter-10122012/).

Hladina α je obvykle volena 0,05 (5 %). Často je další alternativou k α = 0,05 uváděna α = 0,01. Stejně tak je možné použít $\alpha = 0{,}1$ nebo α = 0,2 a to vyžadují-li to specifické podmínky kladeny na náš výzkum. Pokud tedy zamítneme na hladině **statistické významnosti** a naši hypotézu, ještě to vůbec nic neznamená pro naši vědeckou hypotézu, pro náš výzkum.

Jednou ze zásadních nevýhod statistické významnosti je závislost výsledku na počtu měření N. I minimální rozdíl může být pro velké N označen za statistický významný a naopak. Vcelku velký rozdíl může být pro malý počet pozorování označen za nevýznamný. Sigmundová & Sigmund (2012) uvádí příklad závislosti α na N na korelačním koeficientu.

Alternativou k statistické významnosti je posuzování tzv. **věcné významnosti** (effect size). Blahuš (2000) navrhuje stanovit:

- minimální hodnotu v absolutních hodnotách znamenající věcnou významnost a zároveň

- minimální vysvětlené procento rozptylu (relativní zhodnocení podílu ostatních faktorů – koeficient ω^2)

Pro jednotlivé testy lze v literatuře nalézt mnoho tzv. koeficientů věcné významnosti, které přistupují k stanovení významnosti odlišně od hladiny statistické významnosti α. Jednou z výhod konceptu věcné významnosti je nezávislost na počtu měření N.

Uvádíme vybrané koeficienty věcné významnosti s jejich použitím a interpretací Sigmundová, & Sigmund (2012)

Tab. 7 Vybrané effect size koeficienty

statistika	**koeficient**	**hodnocení efektu**
Chí kvadrát χ^2	r	r = 0,10 malý efekt r = 0,30 střední efekt r = 0,50 velká efekt
Korelační koeficient r	r^2 koeficient determinace	malý (nízký) efekt: r = 0,10–0,30 střední efekt: r = 0,31–0,70 velký (výrazný) efekt: r = 0,71–1
t-test, ANOVA	Cohenovo d	d = 0,20 malý efekt d = 0,50 střední efekt d = 0,80 velký efekt
F-test, t-test	ω^2	$\omega^2 \geq 0,1$ – významný efekt
Kruskal-Wallisův test, Friedmanova ANOVA	η^2	$\eta^2 = 0,01$ malý efekt $\eta^2 = 0,06$ střední efekt $\eta^2 = 0,14$ velký efekt

Velmi podrobné informace o statistické a věcné významnosti, jejich vztahu, reálné interpretaci a rozdílu mezi statistickou významností a vědeckou průkazností popisuje již jednou zmíněný Blahuš (2000).

příklad

Příklad 1

Uvažujme 3 měsíční tréninkovou intervenci na skupině sprinterů na 100 m s velmi slabou výkonností (cca 16 s). Po ukončení intervence u nich dojde k průměrnému zlepšení o 0,1 s. Jak se na toto zlepšení můžeme dívat?

- Vzhledem ke skutečnosti, že takové zlepšení v rámci kvality času, je zcela minimální, tak můžeme konstatovat, že ke zlepšení de facto vůbec nedošlo. Rozdíl 0,1 s totiž mohl být způsoben mnoha faktory. Přiznejme, že jedním faktorem mohl být opravdu i trénink ☺.
- Opakuje stejnou situaci, nyní však s elitními světovými sprintery (časy cca 10 s na 100 m). Pokud u nich dojde k lepšení o 0,1 s, pak mluvíme o naprosto nevídaném zlepšení, které je velmi významným počinem v tréninku sprinterů.

Příklad 2

Závislost hladiny α na počtu měření N.

Blahuš (2000) uvádí příklad z roku 1971–1972 s 80000 branci, u kterých byl změřen čas v běhu na 100 m a posléze se test o rok později zopakoval. Rozdíl, a to zhoršení, byl v průměru o 0,0003 s (tři desetitisíciny sekundy). Tento rozdíl je přesto statisticky významný, ačkoliv 0,0003 s de facto žádný rozdíl není.

shrnutí

Před vlastní výzkumnou prací bychom měli zvolit koeficient věcné významnosti a to v absolutních hodnotách/jednotkách, což bude znamenat určení, kdy budeme považovat změnu za významnou. Lze zvolit věcnou významnost i relativně v procentech vysvětlovaného rozptylu. Teprve poté zvolit hladinu statistické významnosti α. Pro konečný závěr nejprve posoudit věcnou významnost a teprve poté statistickou významnost. Uvedené kroky bychom měli provést přesně v pořadí, v jakém jsou popsány. Jinak se nevyhneme případnému podezření, že jsme hladinu významnosti stanovili až po ukončení výpočtů ve snaze dokázat a potvrdit „aspoň něco“…

odkazy na další studijní zdroje

Wikipedia-Effect size. Retrieved June, 19, 2013, from http://en.wikipedia.org/wiki/Effect_size

Coe, R. (2002). *It's the Effect Size, Stupid* . Retrieved June, 22, 2013, from http://www.leeds.ac.uk/educol/documents/00002182.htm

Ellis, P. (2010). *Effect Size FAQs.* Retrieved October, 23, 2013, from http://effectsizefaq.com/

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Elementary-Statistics-Concepts#How%20the%20%22level%20of%20statistical%20significance%22%20is%20calculated

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. s. 165–202.

kontrolní otázky

Koeficient věcné významnosti je na N (počet měření)
a) závislý
b) nezávislý

Koeficient determinace popisuje věcnou významnost k
a) t-testu
b) korelaci
c) analýze rozptylu
d) faktorové analýze

Cohenův koeficient d popisuje věcnou významnost k
a) t-testu
b) korelaci
c) shlukové analýze
d) faktorové analýze

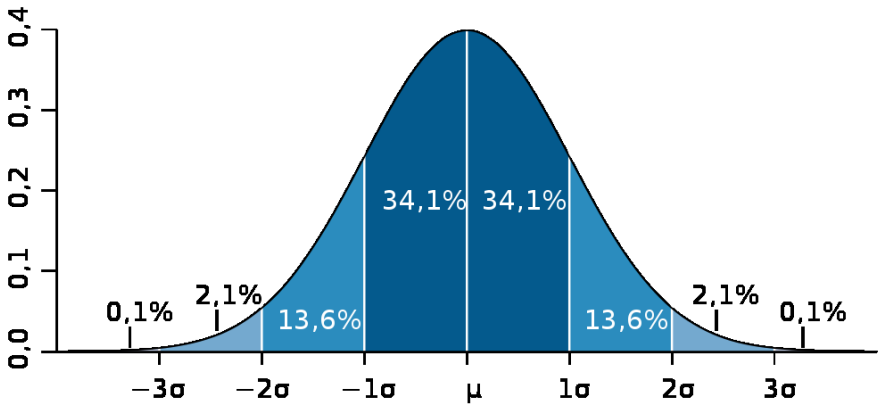
Řecké písmeno η (η^2 je koeficient věcné významnosti u neparametrické analýzy rozptylu) je
a) mí
b) ný
c) éta
d) fí

6. Testy normality ANEB normální rozdělení není až tak normální

teorie

Mnoho statistických testů vyžaduje jako základní předpoklad svého použití normalitu dat. Neboli že data pocházejí z normálního rozdělení. Pokud se řekne **normální rozdělení**, většina čtenářů si představí Gaussovu křivku. A mají pravdu. Gaussova křivka je symetrická s typickým zvonovitým tvarem. Bude-li histogram analyzovaných dat odpovídat tomuto tvaru, můžeme se domnívat, že data pocházejí z normálního rozdělení. Tvar Gaussovy křivky napovídá, že v případě takových dat je nejčastější hodnota rozmístěna kolem střední hodnoty, 2 třetiny se nacházejí v rozmezí ± jednonásobku směrodatné odchylky a cca 95 % všech hodnot je v rozmezí ± dvojnásobku směrodatné odchylky (viz Obr. 5).

Normální rozdělení má mezi ostatními rozděleními to vlastnost, že všechny ostatní rozdělení náhodné veličiny se za jistých podmínek (např. velký počet opakování) k normálnímu rozdělení blíží. Pokud bychom se drželi přesně významu slovního spojení normální rozdělení, ono až tak normální není ☺. Jeho výskyt je v reálných datech určitě v menšině oproti ostatním typům rozdělení. Navíc nikde se netvrdí, že ostatní rozdělení jsou nenormální. Takže čistě matematicky je normální rozdělení zvláštní snad jen tím, že se dá velmi dobře popsat matematickým aparátem.

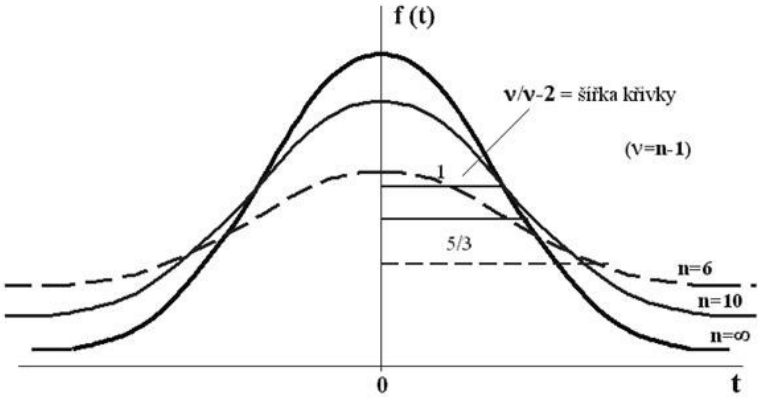


Obr. 5 Gaussova křivka normálního rozdělení

zdroj: http://www.scio.cz/o-vzdelavani/teorie-a-metodika-testu/statisticke-pojmy/, 2013

Dále budeme uvádět některá rozdělení spojitých náhodných veličin. Patří mezi ně např. Studentovo t-rozdělení, Pearsonovo χ^2 (chí-kvadrát) rozdělení a Fisherovo F-rozdělení.

Studentovo t-rozdělení



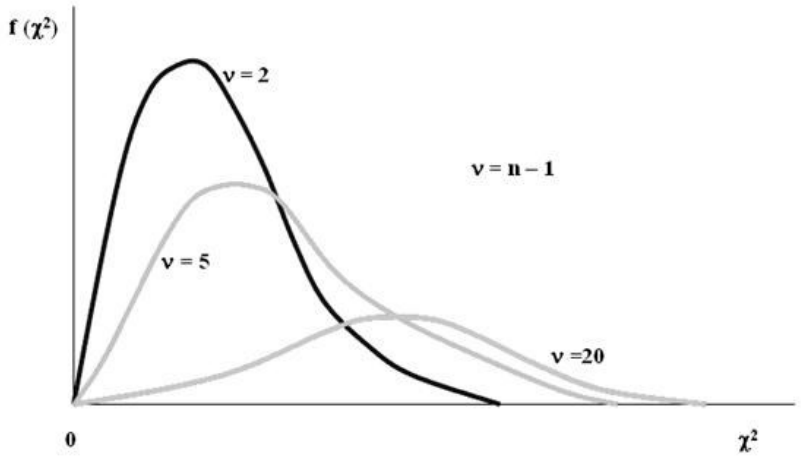
Obr. 6 Studentovo t-rozdělení

Zdroj: http://cit.vfu.cz/statpotr/POTR/Teorie/Predn2/rozdelVS.htm, 2013

Toto rozdělení má jeden parametr a to stupně volnosti ν. Ty se vypočítají jako n-1, kde n je počet měření. Studentovo rozdělení se používá při testování rozdílů středních hodnot 2 výběrů. Na Obr. 6 je rozdělení znázorněno v závislosti na parametru ν. S rostoucím n se Studentovo rozdělení blíží k normálnímu rozdělení.

Pozn.: jméno získalo toto rozdělení po W. S. Gossetovi, který pracoval jako sládek v pivovaru Guiness. Své domněnky o svých pozorováních publikoval pod pseudonymem Student.

Pearsonovo χ^2 (chí-kvadrát) rozdělení

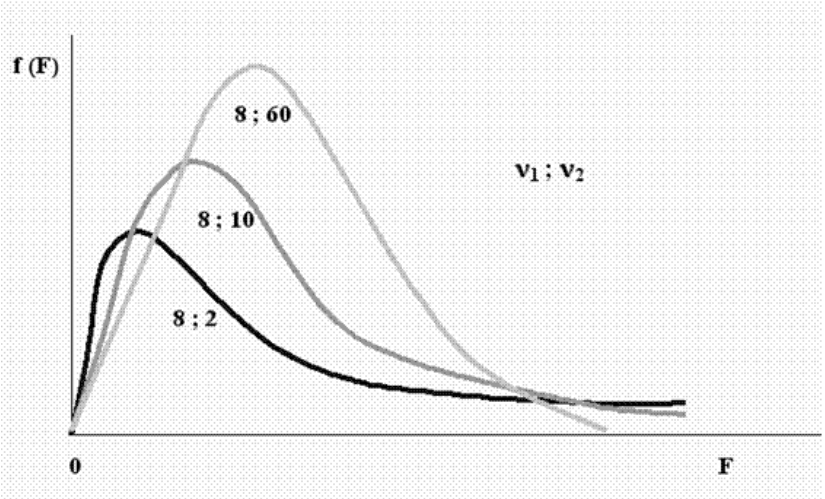


Obr. 7 Pearsonovo χ^2-rozdělení

Zdroj: http://cit.vfu.cz/statpotr/POTR/Teorie/Predn2/rozdelVS.htm, 2013

Toto rozdělení má též jeden parametr a používá se při zkoumání variability náhodné veličiny neb při zkoumání rozdílů četností.

Fisherovo F-rozdělení



Obr. 8 Fischerovo F-rozdělení

Zdroj: http://cit.vfu.cz/statpotr/POTR/Teorie/Predn2/rozdelVS.htm, 2013

Rozdělení se používá při testování dvou rozptylů. Tentokrát má rozdělení 2 parametry, ν_1 a ν_2, což jsou stupně volnosti dvou výběrových souborů.

Mezi všemi čtyřmi uvedenými rozděleními existují vztahy, kdy lze v jistých případech nahrazovat jedno rozdělení druhým…

Testování normality dat se v praxi děje několika testy.

První grafický odhad poskytne např. histogram, přesnější posouzení se pak provádí pomocí testů:

- Chí-kvadrát test dobré shody
- Kolmogorov-Smirnovův test a
- Shapiro-Wilkův test

Testy dobré shody předpokládají nulovou hypotézu H_0: naše data pocházejí z normálního rozdělení a alternativní hypotézu H_A: data nepochází z normálního rozdělení. Testy dobré shody porovnávají průběh distribuční funkce získané z dat (pro zjednodušení si pod tímto pojmem představme jistý průběh křivky) s normovanými distribučními funkcemi (např. normálního rozdělení).

Pokud neprokážeme normalitu dat, musíme použít **neparametrické testy**. Další z důvodů použití neparametrických testů je např. velmi malý rozsah našich dat nebo nemožnost transformace původních dat a dosažení tak normality. Výhodou těchto metod je použití v případě neznámého pravděpodobnostního rozdělení analyzovaných dat. Typickým postupem je převedení naměřených hodnot do jednoho pořadí a dále se již výpočty provádějí jen s těmito pořadími. Neparametrické testy pak mají menší sílu, což je ale na druhou stranu vyváženo vyšší robustností k extrémním hodnotám.

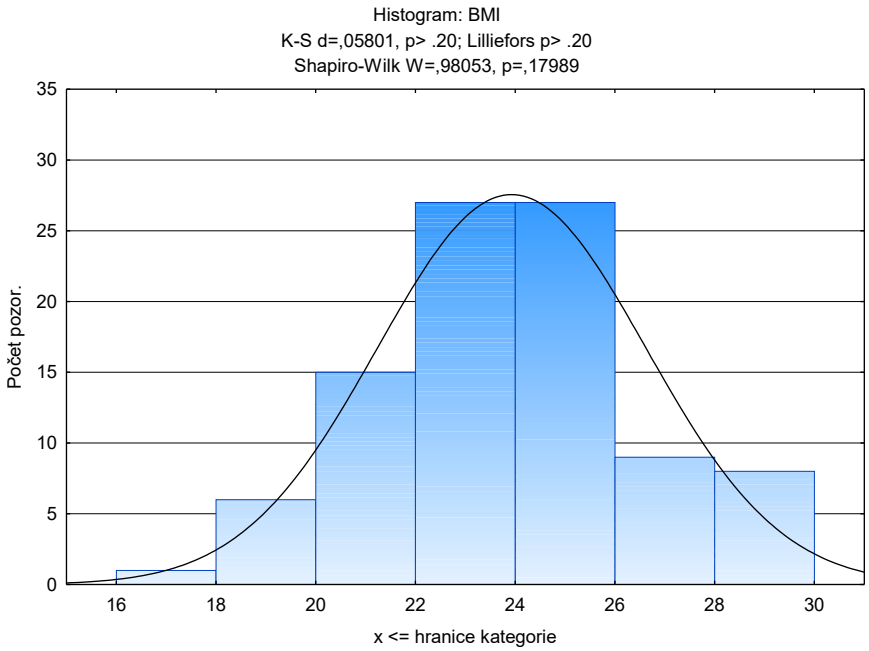
příklady

Máme 93 hodnot z měření BMI

17,9 19,2 19,3 19,6 19,6 19,7 19,8 20,1 20,3 20,3 20,4 20,9 20,9 21,1 21,1 21,1 21,4 21,6 21,6 21,6 21,8 21,9 22,1 22,2 22,2 22,3 22,3 22,4 22,6 22,7 22,8 22,8 22,9 23,0 23,1 23,1 23,2 23,3 23,3 23,4 23,4 23,4 23,6 23,7 23,8 23,9 23,9 23,9 24,0 24,1 24,1 24,1 24,3 24,4 24,4 24,5 24,5 24,5 24,7 24,8 24,9 24,9 25,0 25,1 25,1 25,1 25,1 25,2 25,3 25,3 25,4 25,5 25,6 25,7 25,8 25,9 26,3 26,3 26,5 26,8 26,9 26,9 27,1 27,7 28,0 28,6 29,2 29,4 29,4 29,4 29,4 29,7 30,0

Ověřte normalitu předložených dat

Statistiky – Základní statistiky a tabulky – Popisné statistiky - karta Normalita



Obr. 9 Ověření normality

Na základě histogramu, který vcelku věrně kopíruje křivku normálního rozdělení a na základě výsledků K-S, Lillieforsova a Shapiro-Wilkova testu (výsledky testů jsou vepsány v Obr. 9) konstatujeme, že předložená data pocházejí z normálního rozdělení. Správněji: nezamítáme hypotézu o normalitě dat.

shrnutí

Testování normality je základním krokem při další postupu analýzy dat. Na základě (ne)zjištění normality, volíme testy (ne)parametrické.

odkazy na další studijní zdroje

Ghasemi, A. & Zahediasl, S. (2012). *Normality Tests for Statistical Analysis: A Guide for Non-Statisticians*. Retrieved July, 21, 2013, from http://endometabol.com/?page=article&article_id=3505

GraphPad, Inc. (2013). *Statistics Guide*. Retrieved October, 2, 2013, from http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_interpreting_results_normality.htm

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Statistics-Glossary/N/button/0#Normality%20tests

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. s. 233–234.

kontrolní otázky

Je Gaussova křivka normální (v kontextu statistiky)?
a) ano
b) ne
c) nelze ji otestovat

Pro ověření normality dat lze použít test
a) Kruskal-Wallisův
b) Shapiro-Wilkův
c) Mann-Whitneyův

Proč se data testují, zda pocházejí z normálního rozdělení?
a) k zajištění normálních výsledků
b) testování, zda lze ve výpočtech dále pokračovat
c) k rozhodnutí, zda použít parametrických nebo neparametrických tesů

7. Testy o rovnosti středních hodnot dvou výběrů ANEB t-testy nejsou protesty

teorie

Po prvních krocích při analýze dat, kdy jsme zjistili základní statistické charakteristiky a otestovali normalitu, přichází další možnosti, jak dále postupovat. Po posuzování jednorozměrných dat nás bude zajímat testování, zjišťování a zkoumání závislostí mezi dvěma výběry. Zaměříme se na množinu statistických metod s názvem **t-testy**. Jsou to testy o shodě středních hodnot dvou výběrů. Jinými slovy budeme zkoumat, zda střední hodnoty dvou výběrů (souborů, skupin, proměnných) budou stejné nebo ne.

Jaký konkrétní t-test vybrat bude záležet na dvou skutečnostech:

- data
  - srovnáme s referenční nebo předem známou hodnotou, viz Příklad 1
  - jsou závislá (např. provedeme pretest a posttest na stejné skupině respondentů NEBO změření motorického test u každého respondenta provedeme dvěma způsoby), viz Příklad 2
  - jsou nezávislá (př. provedeme vybraný motorický test na dvou různých skupinách respondentů)
- varianta testu bude
  - parametrická – při nezamítnutí hypotézy normalitě dat
  - neparametrická – pří zamítnutí hypotézy o normalitě dat
    - Wilcoxonův test pro závislá pozorování, viz Příklad 3
    - Mann-Whitneyův test pro nezávislá pozorování

Nulová hypotéza H_0 předpokládá rovnost středních hodnot obou výběrů.

příklady

Příklad 1

Z předložených dat 93 respondentů BMI otestujte, zda hodnota BMI našeho výběru má hodnotu 23. Testování proveďte na 5% hladině statistické významnosti.

Normalitu, jako nutný předpoklad pro použití t-testu, jsme otestovali v předchozí kapitole. V sw Statistica vybereme následující postup

Statistiky – Základní statistiky a tabulky – t-test, samostatný vzorek

Výsledkem je tabulka:

Tab. 8 Výsledek t-testu, samostatný vzorek

Proměnná	Test průměrů vůči referenční konstantě (hodnotě)									
	Průměr	Sm. odch.	N	Sm. chyba	Int. spolehl. –95,000%	Int. spolehl. +95,000%	Referenční konstanta	t	SV	p
BMI	23,92903	2,692687	93	0,279219	23,37448	24,48358	0,00	85,70000	92	0,00

Výsledek: p-hodnota je menší, než hladina statistické významnosti 0,05, proto zamítáme nulovou hypotézu o rovnosti a tvrdíme, že naše skupina respondentů dosahuje statisticky vyšší hodnoty parametru BMI než předpokládaná referenční hodnota 23.

Pro kontrolu jsme vypočítali i 95 % interval spolehlivosti. V intervalu 23,37 – 24,48 se s 95% pravděpodobností bude pohybovat hodnota BMI. Naše referenční hodnota 23 zde není, což je ve shodě s výsledkem t-testu.

Příklad 2

Osm respondentů se zúčastnilo experimentu spojeného s diagnostikou a analýzou složení lidského těla pomocí 2 přístrojů různých výrobců. Zjistěte, zda mezi výsledky uvedených přístrojů je podstatný rozdíl. Uvedená data představují procentuální zastoupení tělesného tuku.

Tab. 9 Data pro t-test, závislá pozorování

Číslo	metoda 1	metoda 2
1	18,6	18,58
2	27,6	27,37
3	27,5	27,27
4	25,0	24,64
5	24,5	24,10
6	26,8	26,33
7	29,7	29,33
8	26,5	26,63

Tab. 10 Výsledky t-testu, závislá pozorování

Proměnná	t-test pro závislé vzorky. Označ. rozdíly jsou významné na hlad. p < ,05000									
	Průměr	Sm. odch.	N	Rozdíl	Sm. odch. rozdílu	t	sv	p	Int. spolehl. –95,000%	Int. spolehl. +95,000%
metoda 1	25,77500	3,316517								
metoda 2	25,53125	3,247916	8	0,243750	0,205352	3,357296	7	0,012129	0,072071	0,415429

Výsledek: p-hodnota je menší než hladina statistické významnosti 0,05, proto zamítáme nulovou hypotézu o rovnosti středních hodnot obou přístrojů.

95% interval spolehlivosti pro rozdíl průměrů je 0,07–0,41. Protože neobsahuje nulu, lze souhlasit se zamítnutím hypotézy o rovnosti průměrných hodnot změřených dvěma přístroji.

Spočítali jsme i Cohenův koeficient *d*, jakožto koeficient věcné významnosti pro t-test a to s výsledkem *d* = 0,07, což značí velmi malý efekt. Tedy podle věcné významnosti měří oba přístroje shodně. Pozn. výpočet Cohenova *d*, viz http://en.wikipedia.org/wiki/Effect_size#Cohen.27s_d

Až sem to bylo vcelku hezké použití t-testu, škoda, že nás nikdo nezastavil, že počítáme špatně. Proč? Neprovedli jsme test normality dat! Pokračování v dalším příkladu.

Příklad 3

Data z příkladu 2 jsme otestovali na normalitu

Tab. 11 Test normality

Proměnná	Testy normality					
	N	max D	K-S p	Lilliefors p	W	p
metoda 1	8	0,225326	p > .20	p > .20	0,862989	0,128578
metoda 2	8	0,222131	p > .20	p > .20	0,870221	0,151480

U obou proměnných nezamítáme hypotézu o normalitě dat, použití parametrického testu bylo správné. I přesto zkusme na stejná data použít neparametrický t-test pro závislá pozorování a to Wilcoxonův test

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků

Tab. 12 Wilcoxonův test

Dvojice proměnných	Wilcoxonův párový test Označené testy jsou významné na hladině p <,05000			
	Počet platných	T	Z	p-hodn.
metoda 1 & metoda 2	8	2,000000	2,240448	0,025063

Wilcoxonův párový test zamítl hypotézu o rovnosti středních hodnot a tvrdí, že přístroje měří různě.

Závěr příkladu: Dostali jsme se do situace, kdy musíme rozhodnout na základě několika odlišných výsledků. Parametrický t-test tvrdí, že metody měří různě. To potvrdil i neparametrický Wilcoxonův test. Věcně, podle Cohenova *d,* však je efekt malý. Konečná interpretace výsledku tohoto příkladu pak říká, že obě metody měří různým způsobem. Ovšem tento rozdíl není extrémně velký.

shrnutí

T-testy jsou množinou statistických metod, která je používána velmi často. Pro jejich použití je už nutné zhodnotit předpoklady jednotlivých testů a pak provést výběr parametrických nebo neparametrických metod. Aplikace několika metod (parametrický t-test a neparametrický t-test, koeficient effect size aj.) na stejná data nemusí vždy přinést shodné výsledky. V takovém případě je nutné zamyšlení a většinou i opatrná intepretace výsledků.

odkazy na další studijní zdroje

GraphPad, Inc. (2013). *Statistics Guide*. Retrieved October, 2, 2013, from http://www.graphpad.com/guides/prism/6/statistics/index.htm?analyses_of_one_grouping_varia.htm

Wikipedia-Student's t-test. September, 22, 2013, from http://en.wikipedia.org/wiki/Student%27s_t-test

Wikipedia-Wilcoxon signed-rank test. Retrieved September, 22, 2013, from http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

Wikipedia-Mann–Whitney U. Retrieved September, 22, 2013, from http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Basic-Statistics#t-test%20for%20independent%20samples

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Basic-Statistics#t-test%20for%20dependent%20samples

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Nonparametric-Statistics#brief

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 233–235.

kontrolní otázky

Wilcoxonův test je test pro
a) výběry pocházejí z normálního rozdělení, pozorování jsou závislá
b) výběry pocházejí z normálního rozdělení, pozorování jsou nezávislá
c) výběry nepocházejí z normálního rozdělení, pozorování jsou závislá
d) výběry nepocházejí z normálního rozdělení, pozorování jsou nezávislá

Mann-Whitneyův test je test pro
a) výběry pocházejí z normálního rozdělení, pozorování jsou závislá
b) výběry pocházejí z normálního rozdělení, pozorování jsou nezávislá
c) výběry nepocházejí z normálního rozdělení, pozorování jsou závislá
d) výběry nepocházejí z normálního rozdělení, pozorování jsou nezávislá

T-test pro závislá pozorování je test pro
a) výběry pocházejí z normálního rozdělení, pozorování jsou závislá
b) výběry pocházejí z normálního rozdělení, pozorování jsou nezávislá
c) výběry nepocházejí z normálního rozdělení, pozorování jsou závislá
d) výběry nepocházejí z normálního rozdělení, pozorování jsou nezávislá

8. Korelace ANEB korelace není kauzalita

teorie

Výraz korelace při náhledu do slovníků, i nestatistických, je definována jako vzájemný vztah mezi veličinami proměnnými, jevy. Korelace dokáže měřit vztah mezi dvěma i více proměnnými. Využívá k tomu různě definované koeficienty, které dokáží vystihnout sílu a případně i směr vztahu. Pokud tedy dostaneme za úkol analyzovat vztah mezi 2 proměnnými, opět začneme s grafickou interpretací dat. Graf nám pomůže ujasnit si, jaký vztah lze v datech hledat a jakým korelačním koeficientem tento vztah popsat. Jedním z velmi jednoduchých je bodový graf. Je to dvourozměrný graf, jednotlivé dvojice z analyzovaných dat zde vyneseme na osy X a Y.

Jednotlivé typy korelačních koeficientů se liší od sebe způsobem použití pro konkrétní typy proměnných. V drtivé většině tyto koeficienty mají stejnou vlastnost. Obvykle nabývají absolutních hodnot od 0 do 1, kde číslo blížící se nule většinou značí velmi malý nebo žádný vztah a naopak hodnota blížící se k jedné, pak vztah velmi silný.

Nulová hypotéza H_0 předpokládá nulovost korelačního koeficientu $r = 0$, alternativní hypotéza pak H_A: $r \neq 0$.

Pearsonův korelační koeficient

Obvykle se značí *r*. Nabývá hodnot od -1 do 1, znaménko pak rozhoduje, zda úměra je přímá (znaménko plus) nebo nepřímá (znaménko minus).

Omezení tohoto koeficientu spočívá v tom, že:

- předpokládá dvourozměrné normální rozdělení. Tedy velmi zjednodušeně řečeno, obě dvě proměnné pocházejí z normálního rozdělení.
- měří pouze vztahy lineární. Ostatní vztahy, ač je z bodového grafu zřejmá závislost, popsat nedokáže
- nerozeznává, která proměnná je závislá a která nezávislá. Nelze rozhodnout o příčinnosti vztahu mezi proměnnými
- interpretace je složitější, proto se dopočítávají dodatečné koeficienty, např. index determinace r^2, který udává, kolik procent z rozptylu jsme dokázali naším korelačním koeficientem vysvětlit.

Parciální korelační koeficient

Při znalosti tří korelačních koeficientů, můžeme vypočítat částečnou korelaci mezi zbývajícími proměnnými s vyloučením vlivu proměnné třetí. Jako bychom předpokládali, že třetí proměnná je konstantní. Vzorce pro případ parciální korelace mezi dvěma ze tří parametrů jsou uváděny např. v publikaci Kopřiva (2011).

Parciální korelační koeficient se značí např. $r_{12.3}$, kde za tečkou je proměnná, jejíž vliv chceme odstranit, přesněji za předpokladu konstantní úrovně proměnné za tečkou.

Mnohonásobný koeficient korelace

Tento koeficient popisuje celkový a společný vliv množiny nezávislých proměnných na proměnnou závislou. Lze tak např. určit, která proměnná má největší vliv. Hodnota tohoto koeficientu je vždy větší než nejvyšší jednoduchý korelační koeficient.

Značí se $r_{y.x1x2}$ kde *y* je závislá proměnná a x_i jsou nezávislé proměnné.

Spearmanův korelační koeficient

V případ porušení normality výběru, při malém počtu pozorování, nebo pokud chceme vyloučit vliv extrémních hodnot, můžeme použít neparametrický pořadový Spearmanův korelační koeficient.

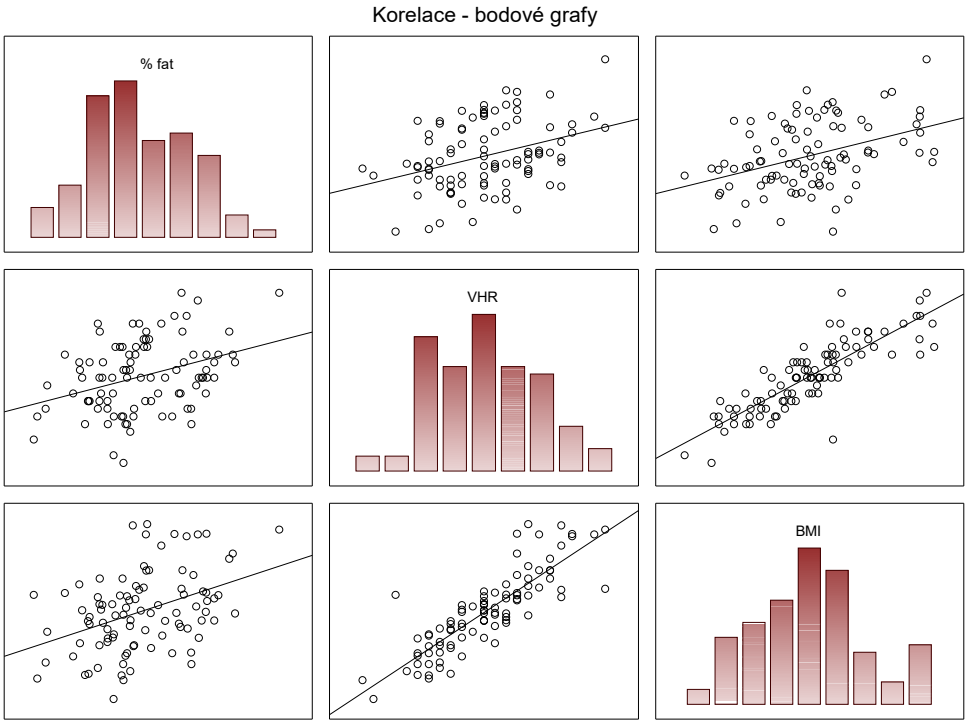
příklad

Máme k dispozici údaje, podle kterých lze popisovat obezitu (BMI-body mass index, % fat – procento tuku a WHR – poměr pasu a boků) od 93 respondentů. Proveďte výpočet a interpretaci korelačních koeficientů.

%Fat	WHR	BMI	%Fat	WHR	BMI	%Fat	WHR	BMI	%Fat	WHR	BMI
16,44	0,77	17,91	16,38	0,85	22,17	21,04	0,87	24,05	21,56	0,92	25,66
18,16	0,76	19,22	24,68	0,82	22,31	11,14	0,87	24,08	20,28	0,91	25,75
3,59	0,82	19,34	16,5	0,82	22,39	14,34	0,85	24,12	10,7	0,89	25,88
17,83	0,82	19,56	5,25	0,86	22,59	18,59	0,88	24,25	14,14	0,93	26,26
11,64	0,81	19,58	21,81	0,84	22,68	30,42	0,86	24,35	17,97	0,91	26,27
12,28	0,84	19,65	15,44	0,84	22,76	24	0,85	24,42	29,91	0,9	26,52
19,27	0,8	19,84	28,96	0,83	22,8	14,6	0,87	24,45	22,01	0,93	26,84
13,8	0,83	20,07	27,8	0,85	22,92	14,09	0,87	24,49	28,09	0,93	26,85
29,63	0,81	20,32	19,22	0,88	22,99	16,27	0,87	24,49	22,46	0,92	26,85
18,1	0,82	20,32	12,38	0,84	23,11	32,26	0,87	24,72	21,59	0,91	27,13
4,97	0,83	20,44	27,48	0,85	23,13	19,1	0,9	24,76	36	0,93	27,65
9,67	0,85	20,87	14,66	0,89	23,2	20,14	0,9	24,86	36,58	0,9	28
18,49	0,81	20,9	12,11	0,84	23,32	33,48	0,89	24,89	24,54	0,94	28,56
26,38	0,82	21,08	25,19	0,87	23,33	13,8	0,94	24,98	28,8	0,95	29,21
18,76	0,81	21,08	31,35	0,87	23,36	18,73	0,87	25,06	21,91	0,92	29,35
29,56	0,83	21,13	33,46	0,87	23,37	17,52	0,91	25,09	30,67	0,97	29,36
22,78	0,84	21,44	19,32	0,89	23,38	34,15	0,9	25,12	26,77	0,95	29,37
19,85	0,82	21,58	12,39	0,89	23,61	3	0,79	25,13	32,17	0,9	29,38
14,08	0,84	21,63	16,8	0,91	23,71	28,29	0,89	25,15	44,44	0,98	29,69
19,94	0,83	21,64	18,13	0,88	23,76	8,26	0,9	25,28	19,71	0,94	29,98
10,78	0,88	21,77	36,99	0,89	23,86	32,09	0,88	25,29	22,13	0,91	30,06
15,46	0,83	21,87	22,73	0,87	23,87	31,59	0,87	25,39			
15,48	0,81	22,11	27,05	0,88	23,9	20,8	0,94	25,49			
30,84	0,87	22,15	32,06	0,85	23,97	27,93	0,98	25,56			

Bodový graf a jednoduché korelační koeficienty vypočítáme v sw Statistica postupem:

Statistiky – Základní statistiky a tabulky – Korelační matice



Obr. 10 Korelace – bodové grafy

Na základě bodové grafu můžeme tušit přímkovou závislost mezi všemi třemi proměnnými s tím, že nejlepší korelace bude mezi WHR a BMI. Proč nejlepší korelace? Bodový graf přibližně kopíruje přímku, která udává směr závislosti. V extrémním případě, pokud by bodový graf zcela přesně kopíroval přímku, bude korelační koeficient roven 1.

Tab. 13 Hodnoty jednoduchých korelačních koeficientů

	Korelace , N=93 Označ. korelace jsou významné na hlad. p < ,05000				
Proměnná	Průměry	Sm. odch.	% fat	WHR	BMI
% fat	21,226	8,267	1,000	0,356	0,405
WHR	0,872	0,046	0,356	1,000	0,847
BMI	24,001	2,753	0,405	0,847	1,000

Nejvyšší **jednoduchý** korelační koeficient je mezi proměnnými BMI a WHR a to 0,847. Celkem vysvětluje 71,7 % procent celkové variability mezi těmi to proměnnými. K číslu 71,7 % jsme dospěli pomocí koeficientu determinace ($r^2 = 0{,}847^2 = 0{,}717$).

Výpočet **parciálních** korelačních koeficientů provedeme

Statistiky – Vícenásobná regrese – Detailní výsledky – Parciální korelace

V tomto postupu je nutné zvolit vždy jednu proměnnou jako závislou a další jako nezávislé, poté postup vyměnit a dopočítat zbývající parciální korelační koeficienty. Jako závislou proměnnou jsme nejprve zvolili BMI a nezávislé pak proměnné WHR a % fat.

Tab. 14 Hodnoty parciálních korelačních koeficientů

Proměnná	Proměnné obsažené v rovnici; ZP: BMI						
	b* v	Parciál. korelace	Semipar. korelace	Tolerance	R^2	t(90)	p-hodn.
% fat	0,118870	0,209067	0,111077	0,873181	0,126819	2,02820	0,045496
WHR	0,804854	0,822766	0,752089	0,873181	0,126819	13,73277	0,000000

Parciální korelační koeficienty můžeme přečíst následovně:

korelační koeficient	**jednoduchý**	**zápis**	**parciální**
BMI a %fat s vyloučením vlivu proměnné WHR hodnota klesne z 0,41 na 0,21	0,41	$r_{BMI\ \%fat.WHR}$	0,21
BMI a WHR s vyloučením vlivu proměnné %fat hodnota zůstává velmi podobná	0,85	$r_{BMI\ WHR.\ \%fat}$	0,82
%fat a WHR s vyloučením vlivu proměnné BMI hodnota klesá téměř na nulu	0,36	$r_{\%fat\ WHR.\ BMI}$	0,03

Výpočet **mnohonásobného** korelačního koeficientu provedeme v dialogu vícenásobné regrese a je to hodnota v R v záhlaví výstupu

Statistiky – Vícenásobná regrese – Základní výsledky – Výpočet: výsledky regrese

N = 93	Výsledky regrese se závislou proměnnou: BMI R = ,85443637 R2 = ,73006152 Upravené R2 = ,72406288 F(2,90) = 121,70 p<0,0000 Směrod. chyba odhadu: 1,4464

Hodnota mnohonásobného korelačního koeficientu je rovna $r_{BMI\ .\ WHR\ \%fat}$ = 0,85. Celková síla vztahu proměnných WHR a %fat na BMI je 0,85.

shrnutí

Ačkoliv má Personův korelační koeficient mnoho nevýhod, je často používán pro různé důvody v oblasti teorie měření. Využití je zřejmé z Tabulky 15.

Tab. 15 Využití Pearsonova korelačního koeficientu (Hendl, 2004), p. 266

Korelační koeficient r_{xy}		Aplikace/interpretace
X	Y	
měření v čase I	měření v čase II	odhad reliability
první polovina testu	druhá polovina testu	odhad reliability
paralelní forma testu I	paralelní forma testu II	odhad reliability
hodnocený test	cílové kritérium	souběžná validita
hodnocený test	měření kritéria v budoucnu	prediktivní validita
hodnotitel I	hodnotitel II	odhad objektivity

odkazy na další studijní zdroje

Wikipedia-Correlation and dependence. Retrieved June, 11, 2013, from http://en.wikipedia.org/wiki/Correlation_and_dependence

Sport Skeptic (2011). *Correlation and Partial Correlation.* Retrieved January, 22, 2013, from http://sportskeptic.wordpress.com/2011/07/18/correlation-and-partial-correlation/

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Basic-Statistics#Correlations

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 237-266.

kontrolní otázky

Jakých hodnot nabývá Pearsonův korelační koeficient
a) 0 až 1
b) –0,5 až 0,5
c) –1 až 1
d) 0 až ∞

Koeficient determinace má k Pearsonovu korelačnímu koeficientu vztah
a) je to odmocnina z Pearsonova korelačního koeficientu
b) je to Pearsonův korelační koeficient na druhou
c) je to polovina Pearsonova korelačního koeficientu
d) nemá žádný vztah

Vyberte správnou možnost.
a) Pearsonův korelační koeficient umí popsat jen lineární závislosti
b) Pearsonův korelační koeficient umí popsat i jiné než lineární závislosti
c) Pearsonův korelační koeficient neumí popsat žádnou závislost

Neparametrický korelační koeficient se nazývá
a) Pearsonův
b) Kendalův
c) Spearmanův
d) Kruskalův

9. Regresní analýza ANEB regrese mohla být reverse

teorie

Poprvé použil výraz regrese antropolog Francois Galton. Zabýval se ve své práci vztahem výšky otců a jejich synů. Přitom objevil vztah, kdy následujíc generace má tendenci návratu k průměru. Tento vztah původně nazval „reversion“, poté změnil na „regression“.

Zatímco v korelační analýze nám jde o popsání vztahů mezi dvěma a více proměnnými, pak v regresní analýze nám jde o víc. O popsání tvaru této závislosti a vytvořit tak model, který můžeme použít např. pro předpověď hodnoty závislé proměnné na několika nezávislých proměnných.
V našem studijním textu se omezíme jen na tvorbu lineárního regresního modelu. Nelineární regresní modely již vyžadují mnohem vyšší zkušenost výzkumníka, který data zpracovává a navíc neexistuje žádný universální způsob, jak by se model dal najít (kromě zkušeností s danou předmětnou oblastí).

Postup při tvorbě modelu obsahuje tyto kroky:

- Návrh modelu, kdy volíme vhodný tvar regresní funkce, která respektuje teoretický model závislosti. Není-li teoretický model znám, provádíme analýzu bodového diagramu a grafu podmíněných průměrů.
- Odhad regresních parametrů a testy jejich významnosti.
- Regresní diagnostika, kdy provádíme analýzu reziduí a identifikaci vlivných bodů.
- Konstrukce zpřesněného modelu, kdy vycházíme z výsledků regresní diagnostiky, např. vyloučíme vlivné body a podobně.
- Zhodnocení kvality modelu vychází ze statistických charakteristik, testů a regresní diagnostiky. Výsledkem je buď přijetí navrženého modelu, nebo návrh modelu dalšího.

Podrobnější informace o statickém modelování závislostí, vztahu regrese a korelace, tvorbou regresních modelů a jejich klasifikací, vyrovnávacích kritérií, bodovými odhady a intervaly spolehlivosti, analýzy reziduí a sedmi řešených příkladů s postupem v sw Statistica zmiňuje např. Sebera (2012).

Vzhledem ke komplexnosti kapitoly o regresi se odkazujeme na již vytvořené studijní materiály a tento materiál doplníme jen o další řešené příklady, které nejsou obsaženy v publikaci Sebery (2012). Doporučená odborná literatura k tvorbě regresních modelů je velmi bohatá, např. Hebák (2007).

V tomto studijním textu ukážeme postup při hledání regresního modelu u závislostí logaritmické a hyperbolické a poslední příklad pak ukáže postup při linearizaci modelu pomocí logaritmování. Připomeňme základní typy lineárních a nelineárních funkcí.

Nejčastěji používané ***funkce lineární z hlediska parametrů***

regresní přímka $Y = \beta_0 + \beta_1 x$,

regresní parabola $Y = \beta_0 + \beta_1 x + \beta_2 x^2$,

regresní log. funkce $Y = \beta_0 + \beta_1 \ln x$,

regresní hyperbola $Y = \beta_0 + \beta_1 \frac{1}{x}$.

Můžeme však použít pro tvorbu modelů **nelineární regresní funkce.** Například:

regresní exponenciální funkce $Y = \beta_0 \beta_1^x$,

regresní mocninná funkce $Y = \beta_0 x^{\beta_1}$,

posunutá exponenciální funkce $Y = \beta_0 \beta_1^x + \beta_2$.

Zde se většinou pokoušíme funkci linearizovat nějakou transformací. Většinou se nabízí logaritmování.

příklady

Příklad 1

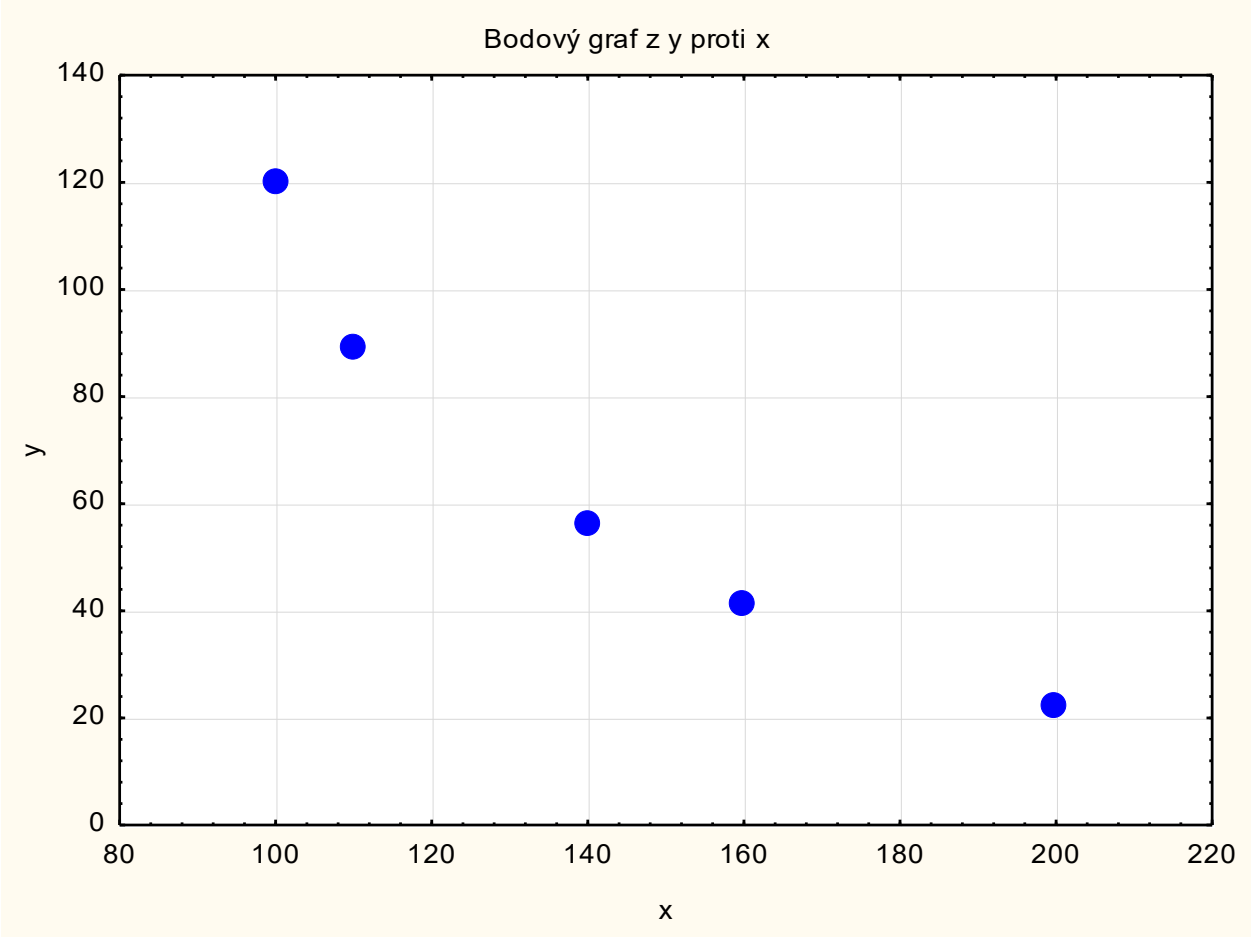
Logaritmická regrese je speciálním případem regrese lineární, kdy závislou proměnnou Y vysvětlujeme nezávislou proměnnou X, která je převedena logaritmickou funkcí. Model pak vypadá Y = a + b · ln(x)

Data v následující tabulce představují poptávku (Y) po určitém výrobku při různých cenách (X). Vyrovnejte data logaritmickou funkcí a odhadněte velikost poptávky při ceně 120 Kč.

x_i = cena v Kč	100	110	140	160	200
y_i = poptávka v tis. kusech	120	89	56	41	22

Protože se jedná o regresní funkci lineární v parametrech, je postup výpočtu stejný jako u přímkové regrese, hodnoty nezávislé proměnné x_i budou nahrazeny logaritmem, tedy $ln\ x_i$.

Bodový graf ukazuje, že závislost může být logaritmická.



Obr. 11 Bodový graf logaritmické regrese

Tvorba regresního modelu proběhne výběrem posloupností kroků:

Statistiky – Vícenásobná regrese – Základní výsledky

Tab. 16 Logaritmická regrese

N=5	Výsledky regrese se závislou proměnnou: y R = ,97575903 R2 = ,95210568 Upravené R2 = ,93614090 F(1,3) = 59,638 p<,00451 Směrod. chyba odhadu: 9,8726					
	b*	Sm. chyba z b*	b	Sm. chyba z b	t(3)	p-hodn.
Abs.člen			734,627	86,74530	8,46879	0,003456
ln x	–0,975759	0,126352	–135,866	17,59344	–7,72256	0,004514

Hodnota F-testu vede k zamítnutí nulové hypotézy o nulovosti regresních koeficientů, tedy můžeme konstatovat, že model je vhodný jako celek. Index determinace je velmi vysoký 0,95. Následují testy jednotlivých regresních koeficientů. Oba dva koeficienty - absolutní člen i koeficient u výrazu ln(x) - jsou statisticky významné.

Logaritmický regresní model má tvar: Y = 734,627 – 135,866 · *ln(x).*

Odhad velikosti poptávky provedeme

Statistiky – Vícenásobná regrese – Residua/předpoklady/předpovědi

POZOR nevkládejte hodnotu 120, ale *ln(120)*, tedy 4,787. Potom je předpověď *Y(120)* = 84,235 Kč.

Příklad 2

Hyperbolická regrese

Tento regresní model má tvar

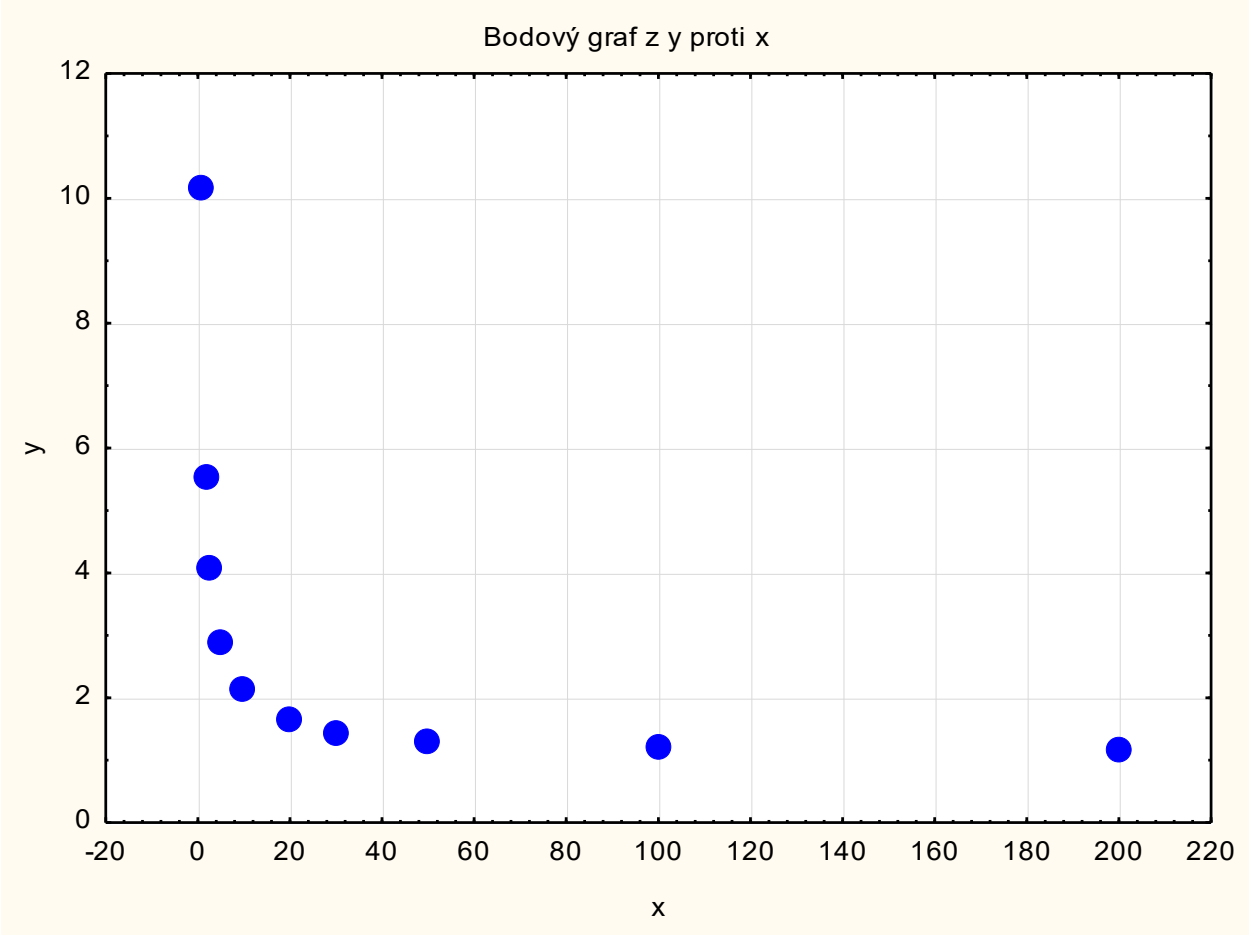
$$Y = a + b\frac{1}{x}$$

Vlastní výdaje (Z v desítkách Kč) na jeden exemplář knihy v závislosti na nákladu (X v tisících kusů), jsou charakterizovány následujícími údaji:

x_i	1	2	3	5	10	20	30	50	100	200
y_i	10,15	5,52	4,08	2,85	2,11	1,62	1,41	1,30	1,21	1,15

Odhadněte koeficienty regresní hyperboly.

Jako v předchozím případě budeme místo nezávislé proměnné x uvažovat proměnnou 1 / x.



Obr. 12 Bodový graf hyperbolické regrese

Tab. 17 Hyperbolická regrese

N=10	Výsledky regrese se závislou proměnnou: y R = ,99981292 R2 = ,99962588 Upravené R2 = ,99957911 F(1,8) = 21375, p<,00000 Směrod. chyba odhadu: ,05851					
	b*	Sm. chyba z b*	b	Sm. chyba z b	t(8)	p-hodn.
Abs. člen			1,118856	0,023097	48,4419	0,000000
1/x	0,999813	0,006839	8,976211	0,061396	146,2028	0,000000

Hodnota F-testu vede k zamítnutí nulové hypotézy o nulovosti regresních koeficientů, tedy můžeme konstatovat, že model hyperbolické regrese je vhodný jako celek. Index determinace je velmi vysoký 0,99. Následují testy jednotlivých regresních koeficientů. Oba dva koeficienty jsou statisticky významné. Výsledný hyperbolický regresní model $Y = 1{,}12 + 8{,}98\,\frac{1}{x}$.

Příklad 3

Mocninná regrese

Použijeme stejná data jako v Příkladu 1 v této kapitole.

x_i = cena v Kč	100	110	140	160	200
y_i = poptávka v tis. kusech	120	89	56	41	22

a vytvoříme mocninný regresní model.

Logaritmováním funkce $Y = \beta_0 x^{\beta_1}$ dostaneme lineární funkci $\ln Y = \ln \beta_0 + \beta_1 \ln x$

Tab. 18 Mocninná regrese

N=5	Výsledky regrese se závislou proměnnou: ln y R = ,99702251 R2 = ,99405389 Upravené R2 = ,99207185 F(1,3) = 501,53 p<,00019 Směrod. chyba odhadu: ,05916					
	b*	Sm. chyba z b*	b	Sm. chyba z b	t(3)	p-hodn.
Abs. člen			15,64629	0,519768	30,1024	0,000081
ln x	-0,997023	0,044520	-2,36083	0,105418	-22,3949	0,000195

Získáme tak model *ln Y = 15,646 – 2,361 ln x*.
Po zpětné transformaci

$$Y = e^{15,646-2,361\ln x}, \quad Y(120) = 76,917$$

Při ceně 120 Kč můžeme očekávat poptávku asi 77 tisíc kusů.

Vzhledem k tomu, že hodnoty regresních koeficientů byly odhadnuty pomocí výběrových (naměřených) hodnot, lze výsledky používat k odhadům pouze v rozsahu těchto naměřených hodnot!

shrnutí

V regresní analýze hledáme funkci, která by dostatečně popisovala vztah mezi dvěma nebo více proměnnými. Provádění regresní analýzy vždy předpokládá, že víme, která proměnná má být závisle proměnnou a která proměnná (nebo proměnné) má být nezávisle proměnnou. Zvolený typ regresní funkce musí především respektovat logické a věcné souvislosti jevů a jejich zákonitosti. Zároveň má být regresní funkce co nejjednodušší a její parametry snadno interpretovatelné. Teorie tvorby regresních modelů je natolik obsáhlá, že výše uvedené příklady lze chápat jako úvod do problematiky.

odkazy na další studijní zdroje

Wikipedia-Regression analysis. Retrieved July, 12, 2013, from http://en.wikipedia.org/wiki/Regression_analysis

TopBettingReviews, Inc. (2013). *Regression Analysis in Sports Betting Systems.* Retrieved July, 2, 2013, from http://www.topbettingreviews.com/regression-analysis-in-sports-betting-systems/

Morris, B. (2011). *The Case for Dennis Rodman*. Retrieved July, 3, 2013, from http://skepticalsports.com/?tag=regression

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from www.statsoft.com/Textbook/Multiple-Regression/button/2

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 266-295.

Sebera, M. (2012). *Vícerozměrné statistiky*. Retrieved January, 23, 2013, from http://www.fsps.muni.cz/~sebera/vicerozmerna_statistika/vicerozmerna_statistika-sebera-fsps-2011.pdf

kontrolní otázky

K čemu slouží lineární regresní modely?
a) popsat vztah mezi závislou a nezávislými proměnnými

b) testovat shodu středních hodnot mezi proměnnými
c) redukovat počet proměnných na základě lineárních vztahů

K čemu slouží bodové grafy v kontextu lineární regrese?
a) k zjištění přibližného vztahu v datech a volbě teoretického modelu
b) k odhadu regresních koeficientů
c) ke zjištění vlivných bodů

Co je reziduální rozptyl?
a) celková velikost odchylek experimentálních hodnot od hodnot daných modelem
b) průměrná hodnot reziduí
c) index determinace pro rezidua

Exponenciální závislost vyjádřená vztahem $Y = \beta_0 \beta_1^x$ je závislost
a) lineární
b) nelineární
c) kvadratická

Mocninná závislost vyjádřená vztahem $Y = \beta_0 x^{\beta_1}$ je závislost
a) lineární
b) nelineární
c) kvadratická

Jak převést nelineární modely na lineární?
a) logaritmováním
b) derivováním
c) integrováním

10. Analýza rozptylu ANEB ANOVA-MANOVA-MANCOVA

teorie

Pokud jsme se bavili v předchozí kapitole o t-testech, tak můžeme pro zjednodušení konstatovat, že t-test je speciální případ analýzy rozptylu (ANOVA), kdy srovnáváme 2 výběry. Připomínám, že t-test je test rovnosti středních hodnot dvou výběrů. Pokud máme výběrů / proměnných / skupin více než dvě, použijeme analýzu rozptylu. Tedy z druhé strany, analýza rozptylu je zobecnění t-testu pro více výběrů.

Častou otázkou je, zda při více proměnných / souborech dat / výběrech nepoužít jen párové t-testy. T-test zkoumá jen variabilitu mezi skupinami, nedokáže postihnout variabilitu uvnitř skupin. Dále nelze použít několik t-testů, protože se zvětšuje chyba 1. druhu.

Předpoklady pro použití parametrické ANOVY je normalita uvnitř jednotlivých skupin či výběrů a homogenita rozptylů. Druhou podmínku lze zmírnit na přibližnou shodu rozptylů. Shodu rozptylů lze provést testy Cochran, Hartley a Bartlett.

Principem ANOVY je rozdělení celkové variability (rozptylu) na rozptyl „uvnitř skupin“ a rozptyl mezi skupinami, což posléze testujeme pomocí F-testu. Při **jednofaktorové** analýze, kdy více proměnných ovlivňuje jeden faktor (např. zhodnocení BMI u více věkových skupin) předpokládá nulová hypotéza, že průměry všech výběrů jsou shodné. H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$. Pokud zamítneme nulovou hypotézu, obvykle nás zajímá, mezi kterými skupinami je statisticky významný rozdíl. K tomu slouží tzv. post-hoc testy. Softwary nabízejí několik post-hoc testů: např. Sheffého, Tukey, LSD. Každý se liší způsobem výpočtu, některé z nich jsou více přísné a konzervativní (Sheffé, Tukey – test nemusí označit rozdíl za staticky významný, ačkoliv ANOVA statisticky významný rozdíl detekovala) nebo liberální (LSD – snadněji označí rozdíl jako statisticky významný, i za cenu nesprávného označení). Doporučujeme spíše provádět konzervativní post-hoc testy.

V reálných datech z výzkumů můžeme najít situaci, kdy proměnnou ovlivňuje více faktorů. Potom mluvíme o **vícefaktorové** analýze rozptylu. Např. porovnání BMI v závislosti na věkových skupinách a pohlaví. Analýzou rozptylu pak můžeme zkoumat nejen působení jednotlivých faktorů na sledovanou proměnnou, ale i působení interakce faktorů na sledovanou proměnnou.

V případě nesplnění předpokladů normality či homogenity rozptylů nebo při velmi malých výběrech, lze použít **neparametrickou ANOVU**. Tyto testy mají nižší sílu, což znamená, spíše nezamítají nulovou hypotézu o rovnosti středních hodnot všech výběrů. Pro závislé výběry to je Friedmanova ANOVA, pro nezávislé výběry Kruskal-Wallisova.

Věcná významnost neboli počítání tzv. effect-size lze u parametrické ANOVY provést pomocí koeficientu eta-kvadrát (η^2 - viz předchozí kapitoly). Pro interpretaci lze použít následující doporučení: $\eta^2 = 0{,}01$ malý efekt; $\eta^2 = 0{,}06$ střední efekt; $\eta^2 = 0{,}14$ velký efekt

Zájemce může bližší informace najít v seznamu zdrojů na konci kapitoly.

Příklad

Ověřte na datech vliv dvou faktorů (věk a pohlaví) na hodnoty proměnné %fat.
Data:

Age group	Sex	% fat
40-59	women	16,44
40-59	women	18,16
18-39	men	3,59
18-39	women	17,83
18-39	men	11,64
18-39	men	12,28
18-39	women	19,27
40-59	women	13,8
18-39	women	29,63
18-39	women	18,1
18-39	men	4,97
>60	men	9,67
18-39	women	18,49
40-59	women	26,38
18-39	women	18,76
18-39	women	29,56
40-59	women	22,78
18-39	women	19,85
18-39	men	14,08
18-39	women	19,94
18-39	men	10,78
40-59	women	15,46
18-39	women	15,48
40-59	women	30,84
18-39	women	16,38
18-39	women	24,68
18-39	women	16,5
18-39	men	5,25
40-59	women	21,81
18-39	men	15,44
18-39	women	28,96

Age group	Sex	% fat
40-59	women	27,8
40-59	men	19,22
18-39	men	12,38
>60	women	27,48
18-39	men	14,66
18-39	men	12,11
18-39	women	25,19
18-39	women	31,35
18-39	women	33,46
40-59	men	19,32
18-39	men	12,39
40-59	men	16,8
40-59	men	18,13
18-39	women	36,99
18-39	women	22,73
40-59	women	27,05
40-59	women	32,06
40-59	men	21,04
18-39	men	11,14
40-59	men	14,34
18-39	men	18,59
>60	women	30,42
18-39	women	24
18-39	men	14,6
18-39	men	14,09
18-39	men	16,27
18-39	women	32,26
40-59	men	19,1
40-59	men	20,14
40-59	women	33,48
18-39	men	13,8

Age group	Sex	% fat
40-59	men	18,73
40-59	men	17,52
>60	women	34,15
40-59	men	3
18-39	women	28,29
18-39	men	8,26
>60	women	32,09
40-59	women	31,59
40-59	men	20,8
40-59	men	27,93
40-59	men	21,56
18-39	men	20,28
18-39	men	10,7
18-39	men	14,14
18-39	men	17,97
18-39	women	29,91
40-59	men	22,01
40-59	men	28,09
18-39	men	22,46
18-39	men	21,59
40-59	women	36
18-39	women	36,58
18-39	men	24,54
>60	men	28,8
18-39	men	21,91
>60	men	30,67
40-59	men	26,77
18-39	women	32,17
>60	women	44,44
18-39	men	19,71
18-39	men	22,13

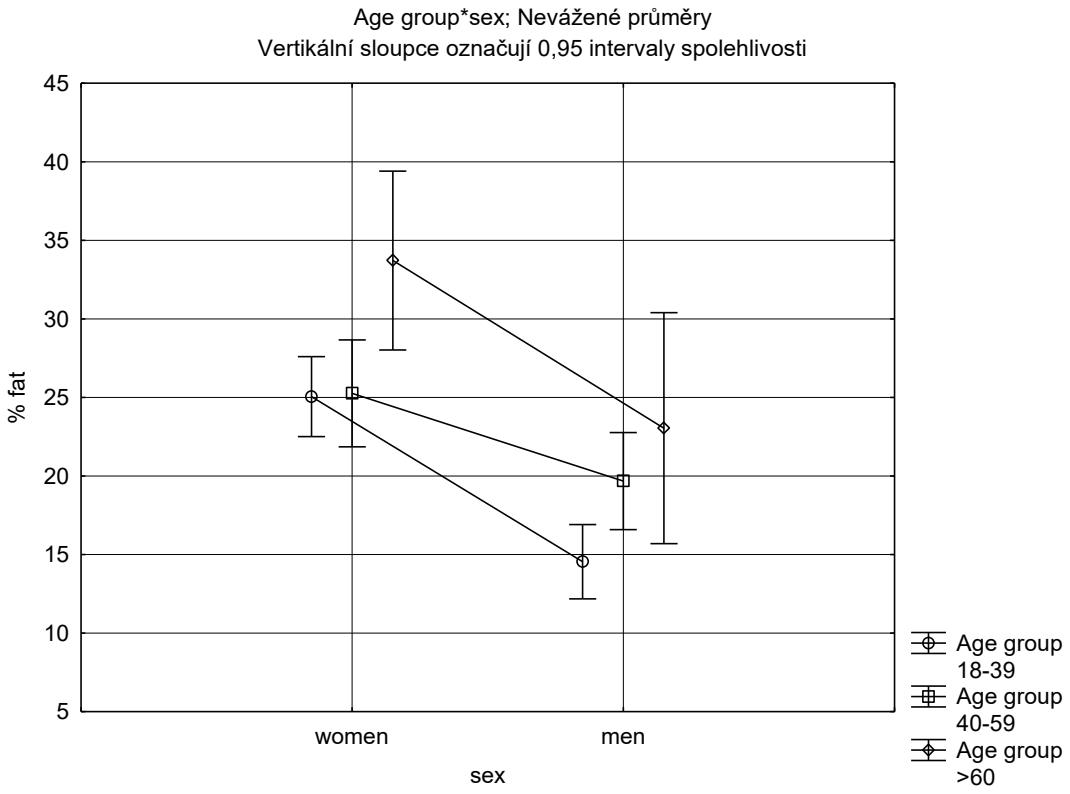
Statistiky – ANOVA – ANOVA s interakcemi

Vypočítáme základní statistické charakteristiky jednotlivých skupin a skupiny zobrazíme.

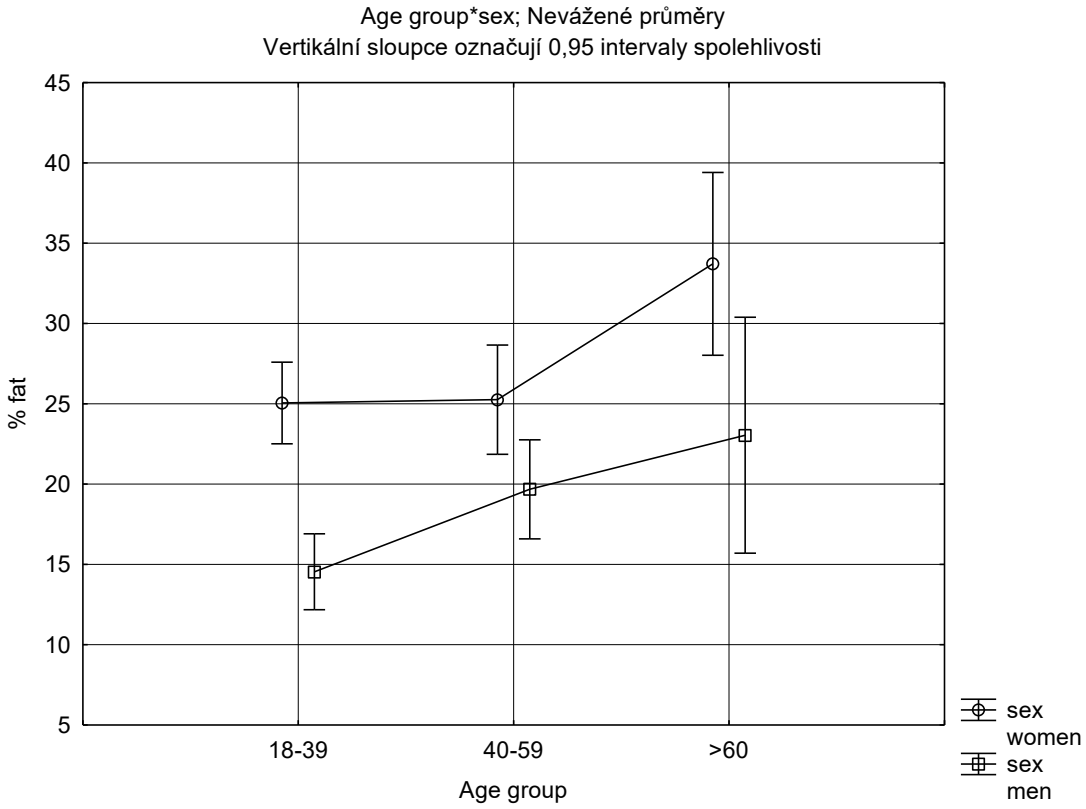
Tab. 19 ANOVA – popisné statistiky

Č. buňky	Age group*sex; Nevážené průměry Současný efekt: F(2, 87) = 1,5261, p = ,22315 Dekompozice efektivní hypotézy						
	Age group	sex	% fat (Průměr)	% fat (Sm. ch.)	% fat (-95,00%)	% fat (+95,00%)	N
1	18-39	women	25,05440	1,280551	22,50917	27,59963	25
2	18-39	men	14,54310	1,188962	12,17991	16,90629	29
3	40-59	women	25,26071	1,711208	21,85950	28,66193	14
4	40-59	men	19,67647	1,552896	16,58992	22,76302	17
5	>60	women	33,71600	2,863399	28,02468	39,40732	5
6	>60	men	23,04667	3,696632	15,69921	30,39412	3

Protože máme dva faktory, můžeme zobrazit data seskupená podle jednotlivých faktorů



Obr. 13a Graf analýzy rozptylu



Obr. 13b Graf analýzy rozptylu

ověříme předpoklady

Tab. 20 ANOVA - testy homogenity rozptylu

	Testy homogenity rozptylu Efekt: "Age group"*sex				
	Hartley. (F-max)	Cochran. (C)	Bartl. (Chí-kv.)	SV	p
% fat	4,592603	0,400195	4,277351	5	0,510213

Nezamítáme hypotézu o rovnosti rozptylů. Můžeme proto použít ANOVU.

Tab. 21 ANOVA

EFEKT	SČ	Stupně (volnosti)	PČ	F	p	Parciál. éta-kvadr.
Abs. člen	27050,27	1	27050,27	659,8388	0,000000	0,883509
Age group	538,30	2	269,15	6,5654	0,002210	0,131136
sex	970,59	1	970,59	23,6756	0,000005	0,213919
Age group*sex	125,12	2	62,56	1,5261	0,223150	0,033893
Chyba	3566,59	87	41,00			

Výsledkem ANOVY je tvrzení, že podle faktoru „věk“ existují statisticky významné rozdíly ve sledované proměnné %fat., což potvrzuje i vysoká hodnota éta-kvadrát (0,13). Pokud bychom provedli jednofaktorovu ANOVU jen podle grupovací proměnné „pohlaví“, pak i zde je zamítnuta hypotéza o nulovosti středních hodnot podvýběrů, jinými slovy hodnota % fat je odlišná i v závislosti na pohlaví. Totéž potvrzuje i hodnota éta-kvadrát (0,21). V interakci „věk x pohlaví“ však ANOVA nedetekuje statisticky významný rozdíl.

Provedením post Scheffeho post-hoc testu určíme, mezi kterými dvojicemi existuje statisticky významná změna.

Tab. 22 ANOVA – post-hoc testy (faktor věk)

Č. buňky	Scheffeho test; proměnná % fat Pravděpodobnosti pro post-hoc testy Chyba: meziskup. PČ = 40,995, sv = 87,000			
	Age group	1 19,409	2 22,198	3 29,715
1	18-39		0,160524	0,000274
2	40-59	0,160524		0,015375
3	>60	0,000274	0,015375	

Tab. 23 ANOVA – post-hoc testy (faktor pohlaví)

Č. buňky	Scheffeho test; proměnná % fat Pravděpodobnosti pro post-hoc testy Chyba: meziskup. PČ = 40,995, sv = 87,000		
	sex	1 26,104	2 16,845
1	women		0,000000
2	men	0,000000	

Tab. 24 ANOVA – post-hoc testy (interakce faktorů věk a pohlaví)

Č. buňky	Scheffeho test; proměnná % fat Pravděpodobnosti pro post-hoc testy Chyba: meziskup. PČ = 40,995, sv = 87,000							
	Age group	sex	1 25,054	2 14,543	3 25,261	4 19,676	5 33,716	6 23,047
1	18-39	women		0,000010	1,000000	0,222288	0,190333	0,998210
2	18-39	men	0,000010		0,000273	0,240488	0,000005	0,447450
3	40-59	women	1,000000	0,000273		0,331455	0,277727	0,997644
4	40-59	men	0,222288	0,240488	0,331455		0,004276	0,982096
5	>60	women	0,190333	0,000005	0,277727	0,004276		0,398702
6	>60	men	0,998210	0,447450	0,997644	0,982096	0,398702	

Zajímavá situace nastává v případě Scheffeho testu pro interakci obou faktorů, neboť detekuje statisticky významný rozdíl mezi skupinami „1 a 2“, „2 a 3“, „2 a 5“ a „4 a 5“. O které podskupiny se jedná, lze vyčíst z tabulky. Co se týče intepretace tohoto stavu, kdy úvodní F-test nezamítl hypotézu pro interakci grupovacích proměnných, ač post-hoc testy označili některé dvojice za statisticky významné, doporučuji přiklonit se k variantě, která je z pohledu věcného hlediska pro výzkumníka obhajitelnější.

shrnutí

Pro vyhodnocení experimentálních dat, kde zkoumáme vliv závislých proměnných na nezávislé, mluvíme o analýze rozptylu. Podle počtu faktorů, které ovlivňují naše data, pak hovoříme o jednofaktorové nebo vícefaktorové analýze rozptylu. V praxi pak zkoumáme, zda průměry mezi jednotlivými podskupinami jsou shodné nebo ne. Co se týče předpokladů použití ANOVY, při velkém počtu měření můžeme vynechat podmínky normality. Posledním krokem je pak aplikování post-hoc testů na zjištění statisticky významných rozdílů.

Na závěr bych vysvětlil nadpis kapitoly. ANOVA je zřejmá, MANOVA je vícerozměrná analýza rozptylu (Multivariate Analysis of Variance). ANCOVA je analýza kovariancí (Analysis Of Covariance) *MANCOVA* je pak (multivariate analysis of covariance). Existují ještě např. RMANOVA (Repeated Measures Analysis Of Variance). Vysvětlení těchto pojmů je již mimo předpokládaný záměr tohoto studijního textu.

odkazy na další studijní zdroje

Wikipedia-Analysis of variance. Retrieved June, 11, 2013, from http://en.wikipedia.org/wiki/Anova

Department of Psychology, University of Toronto (1997). *Statistica*. Retrieved September, 22, 2013, from http://www.psych.utoronto.ca/courses/c1/statistica/toc.htm

Oxford Brookes University (2013). *Statistical tests*. Retrieved September, 22, 2013, from http://www.brookes.ac.uk/services/upgrade/maths-stats/tests/anova.html

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/ANOVA-MANOVA/button/1

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 337-370.

Sebera, M. (2012). *Vícerozměrné statistiky*. Retrieved January, 23, 2013, from http://www.fsps.muni.cz/~sebera/vicerozmerna_statistika/vicerozmerna_statistika-sebera-fsps-2011.pdf

kontrolní otázky

Jakou nulovou hypotézu testujeme v analýze rozptylu?
a) střední hodnoty všech výběrů jsou shodné
b) střední hodnoty všech výběrů nejsou shodné
c) existuje alespoň jeden výběr, kde střední hodnota není rovna středním hodnotám ostatních výběrů

Cochran a Bartlett testy se používají
a) k testování homogenity rozptylů
b) k testování hypotézy o rovnosti středních hodnot výběrů
c) k zjištění síly vztahu mezi proměnnými

Sheffé a Tukey post-hoc testy jsou spíše
a) konzervativní
b) liberální
c) nejsou to post-hoc testy

Kruskal-Walisova ANOVA je
a) parametrickou analýzou rozptylu pro závislá pozorování
b) parametrickou analýzou rozptylu pro nezávislá pozorování
c) neparametrickou analýzou rozptylu pro závislá pozorování
d) neparametrickou analýzou rozptylu pro nezávislá pozorování

Friedmanova ANOVA je
a) parametrickou analýzou rozptylu pro závislá pozorování
b) parametrickou analýzou rozptylu pro nezávislá pozorování
c) neparametrickou analýzou rozptylu pro závislá pozorování
d) neparametrickou analýzou rozptylu pro nezávislá pozorování

11. Faktorová analýza

teorie

Faktorová analýza patří mezi vícerozměrné statistické metody. Mezi její hlavní úkoly patří redukce původního počtu proměnných, resp. hledání nových latentních proměnných. Vznikají tak nové proměnné – faktory, které shlukují původní proměnné, které spolu vysoce korelovali. Takto vzniklé faktory lze interpretovat na základě přítomnosti původních proměnných.

Faktorová analýza se potýká s několika metodologickými obtížemi:

1) kolik zvolit faktorů, aby dokázali dostatečně popsat původní proměnné, resp. aby dostatečně dokázali vysvětlit variabilitu původních proměnných
2) najít dostatečnou intepretaci a věcné zhodnocení vzniklé nové faktorové struktury
3) veškeré výpočty jsou založeny na lineárních kombinacích, tudíž existuje-li v datech vztah nelineární, faktorová analýza jej nezachytí
4) k optimalizaci se posléze provádějí tzv. rotace. Rotací existuje celá řada, což zvyšuje nejednoznačnost výsledků, neboť zlí jazykové tvrdí, že bychom mohli s faktory rotovat tak dlouho, až najdeme předpokládaný výsledek.

I přes tyto skutečnosti je faktorová analýza vyhledávanou statickou procedurou. Možnosti využití faktorové analýzy jsou z ideového pohledu dvě: explorační a konfirmační. **Explorační** přístup hledá v datech nové, latentní proměnné, které se výzkumník snaží vhodně interpretovat. U **konfirmační** faktorové analýzy má výzkumník předem danou představu o datech a struktuře v nich a faktorovu analýzu využívá jen pro potvrzení své domněnky.

Jednotlivé fáze faktorové analýzy:

a) Nejprve nalezneme prvotní faktorové zátěže. Např. pomocí metody zvané analýza hlavních komponent (Principal Component Analysis - PCA). To je postup, kdy hledáme lineární kombinace původních proměnných, které nejlépe vysvětlí variabilitu původních proměnných. Mohou nastat dvě extrémní situace. Všechny původní proměnné spolu vysoce korelují, tudíž lze vytvořit jednu jedinou komponentu, která dostatečně vysvětlí variabilitu původních dat. Druhým extrémem je situace, že původní proměnné spolu vůbec nekorelují, tudíž pro vysvětlení celkové variability je potřebné mít tolik komponent, kolik je původních proměnných. Obvykle k těmto extrémním situacím nedochází. Počet komponent se pak stanoví dobrým odhadem výzkumníka. Jako pomůcka může sloužit tvrzení, že hlavní komponenty by měly umět vysvětlit cca 70-80 % původní variability. Druhou pomůckou je pak sestrojení tzv. **scree** grafu (sutinový graf) a počet komponent je pak roven počtu **vlastních čísel** větších než 1.
b) Každou novou komponentu lze popsat jako lineární kombinaci původních proměnných. Těmto koeficientů se říká **faktorové zátěže** a popisují, jakou variabilitu původní proměnné popisuje nově vzniklá komponenta. Lineární kombinace původních proměnných lze optimalizovat vůči nějakému optimalizačnímu kritériu. Neboli nově vzniklou strukturou lze transformovat, otáčet, rotovat. Smyslem otáčení (**rotace**) je maximalizovat faktorové zátěže a tím najít co nejlepší interpretovatelnost, kdy původní proměnné jsou silně korelovány jen

s jediným faktorem a velmi slabě s ostatními faktory. Používané rotace jsou např. Varimax a Quartimax.

c) Hledání interpretace nově vzniklých faktorů a výpočet **faktorových skórů** (hodnoty faktorů popisující každého respondenta/měření)

Příklad

Přístroj InBody (www.inbody.cz) je analyzátor složení těla, který podává komplexní výsledky o měřených probandech a to formou mnoha testů. Výsledky jsou pak doprovázeny pomocnými kritérii, kterými lze hodnotit zdraví člověka, jeho kondici a případné doporučení pro optimální složení těla. Některé výsledky sledovaných proměnných jsou velmi motivační, neboť jejich zvyšování / snižování (pokud měříme za standardních podmínek) může identifikovat případnou změnu ve složení těla probanda. Pro náš příklad jsme vybrali několik proměnných. Jsou to:

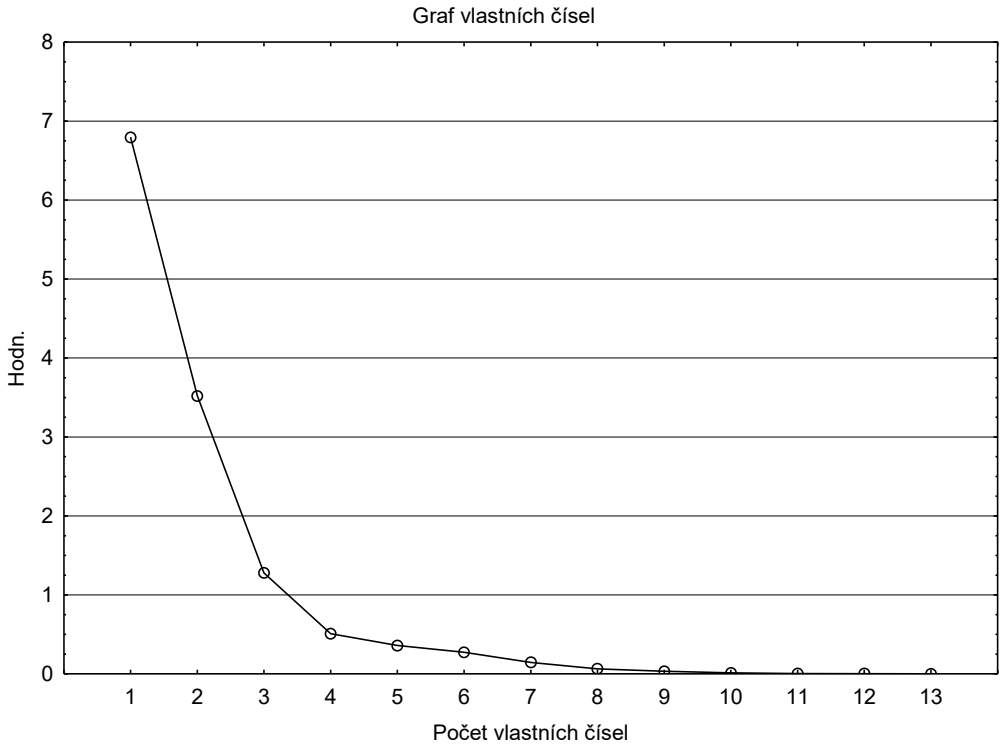
FS (fitness score), VFA (visceral fat area), Hmotnost, Množství kosterního svalstva, % Fat, WHR (waist-hip ratio), BMI (body fat mass index), Svalová hmota pravé ruky, Svalová hmota levé ruky, Množství svaloviny v trupu, Svalová hmota pravé nohy, Svalová hmota levé nohy, % muscle. Zajímá nás, s jakými proměnnými bude nejvíce korelovat proměnná FS (fitness score). Z výsledků 1412 probandů jsme provedli faktorovou analýzu.

Určení počtu faktorů:

V tabulce vlastních čísel vidíme celkem 3 vlastní čísla větší než 1. Celkem tyto 3 komponenty vysvětlují cca 89 % původní variability, což je dostatečné množství. Stejnou informaci nám podává Scree graf (graf vlastních čísel). Překlad sutinový graf lze popsat takto: pokud bychom seshora spustili kámen, tak v místě kde by se zastavil, tam je možné odhadnout počet faktorů. Na našem obrázku to je mezi 3 a 4 vlastním číslem, což je další pomůcka pro určení počtu faktorů.

Tab. 25 Tabulka vlastních čísel u faktorové analýzy

Hodn.	Vlastní čísla Extrakce: hlavní komponenty			
	Vl. číslo	% celk. rozptylu	Kumulativ. vlast. číslo	Kumulativ. %
1	6,796450	52,28038	6,79645	52,28038
2	3,522177	27,09367	10,31863	79,37405
3	1,278671	9,83593	11,59730	89,20998

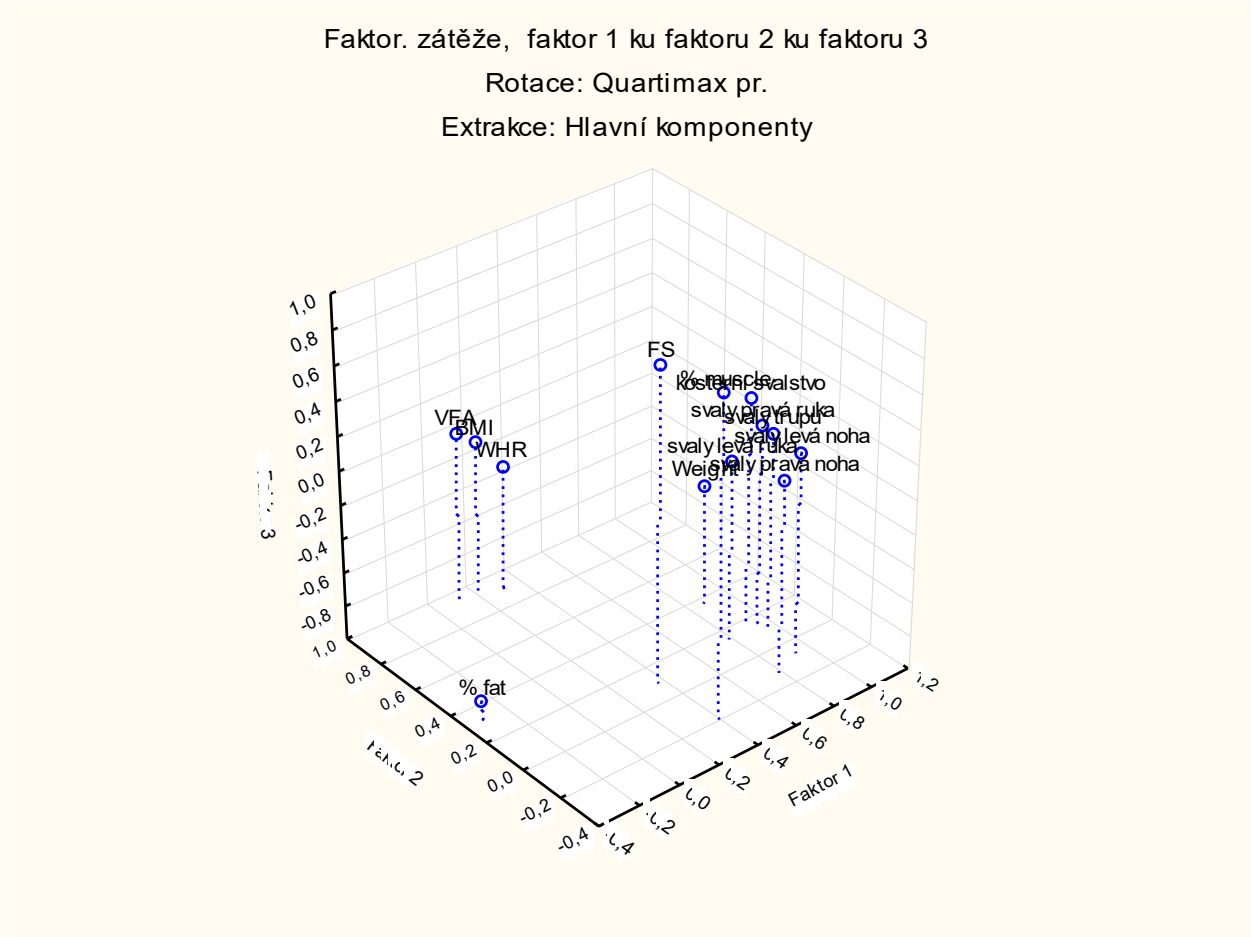


Obr. 14 Scree graf

Výsledkem faktorové analýzy je následující tabulka. Pro optimalizaci vzniklých faktorů jsme použili rotaci Varimax.

Tab. 26 Výsledek faktorové analýzy

Proměnná	Faktor. zátěže (Varimax pr.) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)		
	Faktor 1	Faktor 2	Faktor 3
FS	0,244025	0,115490	0,873887
VFA	0,065777	0,947183	-0,001216
Weight	0,876580	0,410105	-0,162577
kosterní svalstvo	0,853830	0,236441	0,440237
% fat	-0,216246	0,282879	-0,917894
WHR	0,286600	0,904598	-0,216518
BMI	0,174480	0,952787	-0,089275
svaly pravá ruka	0,901514	0,197570	0,309151
svaly levá ruka	0,756236	0,179916	0,164900
svaly trupu	0,924167	0,163201	0,283776
svaly pravá noha	0,761988	-0,084238	0,249347
svaly levá noha	0,892373	-0,024797	0,320149
% muscle	0,281951	-0,177441	0,924284
Výkl. roz	5,415794	3,077940	3,103564
Prp. celk	0,416600	0,236765	0,238736



Obr. 15 3D graf faktorů

Graficky lze znázornit rozvržení původních proměnných pomocí 3 rozměrného grafu.

Výsledkem jsou 3 nové faktory. První můžeme nazvat hmotnostní parametry (v absolutních jednotkách). Druhý faktor je reprezentován proměnnými, kterými lze popisovat obezitu. Třetí faktor je tvořen třemi proměnnými, které spolu vysoce korelují. Fitness score se přímo úměrně zvyšuje s relativním množstvím svalové hmoty a nepřímo úměrně s relativním množstvím tuku.

Pokud tedy chci zvýšit své fitness score (měřeno přístrojem InBody), musím se zaměřit na snížení obsahu tuku v těle a zvýšení relativního množství svalové hmoty v těle.

shrnutí

I přes uvedené metodologické nedostatky je použití faktorové analýzy důležitým způsobem při hledání skrytých datových struktur v původních proměnných. Ať už explorační nebo konfirmační přístup, pomůže tato analýza pochopit širší souvislosti, které jsou ukryty v analyzovaných datech.

odkazy na další studijní zdroje

Wikipedia-Factor analysis. Retrieved June, 19, 2013, from http://en.wikipedia.org/wiki/Factor_analysis

Rummel, R. J. (2002). *Understanding Factor Analysis*. Retrieved June, 19, 2013, from http://www.hawaii.edu/powerkills/UFA.HTM

Tucker, L. & MacCallum, R. (1997). *Exploratory Factor Analysis*. Retrieved September, 22, 2013, from http://www.unc.edu/~rcm/book/factornew.htm

StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Retrieved September, 22, 2013, from http://www.statsoft.com/Textbook/Principal-Components-Factor-Analysis/button/1

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál. p. 468-485.

Sebera, M. (2012). *Vícerozměrné statistiky*. Retrieved January, 23, 2013, from http://www.fsps.muni.cz/~sebera/vicerozmerna_statistika/vicerozmerna_statistika-sebera-fsps-2011.pdf

kontrolní otázky

Faktorová analýza slouží k
a) redukci počtu proměnných
b) testování rovnosti středních hodnot výběrů
c) testování homogenity rozptylů

Nově vzniklé faktory
a) shlukují původní proměnné, které spolu vysoce korelují
b) shlukuj původní proměnné s nejvyšší variabilitou
c) shlukují původní proměnné s nízkými korelacemi

Rotace faktorové struktury
a) zvyšuje interpretovatelenost výsledků
b) zvyšuje reziduální rozptyl faktorové struktury
c) snižuje střední hodnoty jednotlivých faktorů

Scree graf slouží k
a) určení počtu faktorů
b) určení vztahu mezi faktory
c) určení vztahu mezi původními proměnnými

12. Závěr ANEB Statistický rozcestník ANEB co s daty

Na závěr předkládáme jednoduchý rozcestník s nejběžněji používanými postupy při analýze dat. Tabulka je rozdělena do tří sloupců s popisem, který usnadní orientaci při hledání vhodné statistické metody.

1. příprava výzkumného šetření je nejdůležitější část
2. sběr a analýza dat slouží k zamítnutí/nezamítnutí předem stanovených úkolů práce a hypotéz
3. vždy mít na paměti věcné hledisko výzkumu, zejména v souvislosti s interpretací statistických výsledků. Statistika je dobrým sluhou, ale špatným pánem. Navíc v konečném důsledku to je jen a jen hra s čísly…

CHCI S DATY PROVÉST	ZPŮSOB	UMOŽNÍ MI ZJISTIT
První náhled na data	Základní popisná statistika • průměr, směrodatná odchylka, rozptyl, N, medián, kvartily a další míry polohy a variability • tabulky četností: absolutní, relativní, kumulativní • grafy: krabicový, histogram	• chybná měření, extrémy • homogenitu souboru • chybějící data
Otestovat normalitu	• Kolmogorov-Smirnov test, Shapiro-Wilks test	• rozhodnutí, zda použít parametrické nebo neparametrické testy
Zjistit, zda výběry/skupiny jsou shodné nebo ne	• 2 skupiny/proměnné: t-testy • 3 a více skupin/proměnných: Analýza rozptylu (ANOVA)	• konstatovat statisticky nebo věcně (size of effect) významný rozdíl Př. došlo ke zlepšení výbušné síly po intervenci?(pretest-posttest) Př. která ze dvou tréninkových metod je úspěšnější? Př. mezi kterými skupinami je statisticky významný rozdíl Př. byl zkoumán výsledný čas v motorickém testu v závislosti na typu suplementace sportovce (faktor A) a na způsobu tréninku (faktor B)

CHCI S DATY PROVÉST	ZPŮSOB	UMOŽNÍ MI ZJISTIT
Zjistit závislost více proměnných (spojité)	• korelace, index determinace • faktorová analýza	• těsnost lineárního vztahu mezi proměnnými • může existovat jasný vztah ale nelineární, který nezachytíme pomocí korelace nebo faktorové analýzy • korelace neznamená kauzalitu!!! Př. závisí výkon v běhu na 100 m s výkonem do skoku do dálky? Př. závisí ekonomika běhu na povrchu?
Zjistit závislost více proměnných (kategoriální-dotazník)	• test nezávislosti chí-kvadrát v kontingenční tabulce • vícerozměrné kontingenční tabulky - asociační stromy • shluková analýza	• sílu a směr vztahu Př. závisí bolestivost zad na věku a způsobu zaměstnání? Př. mezi kterými proměnnými z dotazníku existuje nejsilnější vazba?
Redukovat velký počet vstupních dat	• faktorová analýza	• zda za naměřenými daty není nějaká latentní struktura (POZOR na interpretaci) Př. lze 10 disciplín desetiboje popsat menším počtem faktorů?
Vysvětlit závislou proměnnou několika nezávislými, provést předpověď	• lineární regrese	• příspěvek jednotlivých nezávislých proměnných k popisu proměnné závislé Př. Popsat trend výkonnosti v atletických disciplínách a provést předpověď výkonů na olympiádě v Riu 2016

13. Použité zdroje

Blahuš, P. (2000). Statistická významnost proti vědecké průkaznosti výsledků výzkumu. In *Česká kinantropologie, 4*(2), 53–72.

Hebák, P. (2007). *Vícerozměrné statistické metody*. (2nd ed.) Praha: Informatorium.

Hendl, J. (2004). *Přehled statistických metod zpracování dat: analýza a meta analýza dat*. Praha: Portál.

Kopřiva, J. (2011). *Sport, matematika, počítač*. Brno: MU. Retrieved September, 22, 2013, https://is.muni.cz/auth/do/fsps/e-learning/sport-matematika/pdf/sport-matematika-pocitac.pdf

Řezanková, H., Marek, L., & Vrabec, M. (2000). *IASTAT - interaktivní učebnice statistiky*. Retrieved July, 19, 2013, http://iastat.vse.cz

Sebera, M. (2012). *Vícerozměrné statistiky*. Retrieved January, 23, 2013, from http://www.fsps.muni.cz/~sebera/vicerozmerna_statistika/vicerozmerna_statistika-sebera-fsps-2011.pdf

Sigmundová, D,. & Sigmund, E. (2012). Statistická a věcná významnost a použití koeficientů velikosti účinku při hodnocení dat o pohybové aktivitě. In. *Tělesná kultura, vol. 35, no 1*. Olomouc: FTK. Retrieved July, 9, 2013, http://www.telesnakultura.upol.cz/index.php/telesnakultura/article/viewFile/98/163.

Statsoft, Newsletter, Retrieved September, 22, 2013, from http://www.statsoft.cz/o-firme/archiv-newsletteru/newsletter-10122012/).

14. Anglicko-český slovník

A	
absolute deviation	absolutní odchylka
absolute error	absolutní chyba
absolute frequency	absolutní četnost
absolute increase	absolutní přírůstek
absolute moment	obecný moment
acceptance region	obor přijetí
Accuracy	přesnost
additive function	aditivní funkce
Adjusted	upravený
alternative hypothesis	alternativní hypotéza
analysis of covariance	analýza kovariance
analysis of variance	analýza rozptylu
approximate value	přibližná hodnota
ascending	vzestupný
arranging by size	uspořádání podle velikosti
assess	ocenit
assumption	předpoklad
average	průměr
asymptotic normality	asymptotická normalita
axis	osa
B	
balanced design	vyvážený pokus
bar chart	sloupkový graf
base period	základní období
base line	základní čára
basic	základní
bell-shaped curve	zvonovitá křivka
bias	vychýlení
biased estimator	vychýlený odhad
bivariate	dvourozměrný
boundary	hranice
box plot	krabicový graf
box and whiskers plot	krabicový graf („s vousy“)

C	
calculate	vypočítat
cartodiagram	kartodiagram
cartogram	kartogram
case	případ
central limit theorem	centrální limitní teorém
central moment	centrální moment
central tendency	obecná úroveň
chain base index	řetězový index
changed weights index	index proměnlivého složení
character	znak
chart	graf
chi-square distribution	rozdělení chi-kvadrát
chronological average	chronologický průměr
chunk sampling	živelný výběr
class	třída
class limits	hranice tříd
cluster sampling	výběr skupin
coefficient of association	koeficient asociace
coefficient of contingency	koeficient kontingence
coefficient of variation	variační koeficient
column	sloupec
comparing	srovnání
composite hypothesis	složená hypotéza.
composite index	souhrnný index
compound event	složený jev
compute	vypočítat
conclusion	rozhodnutí
condition	podmínka
conditional average	podmíněný průměr
conditional distribution	podmíněné rozdělení
conditional probability	podmíněná pravděpodobnost
confidence	spolehlivost
confidence belt	pás spolehlivosti
confidence interval	interval spolehlivosti
consumer price index	index spotřebitelských cen
contingency table	kontingenční tabulka
continuous variable	spojitá proměnná
correlation index	index korelace
correlation matrix	korelační matice
correlation ratio	korelační poměr
covariance matrix	kovarianční matice
critical region	kritický obor
critical value	kritická hodnota
crosstabulation	kombinační třídění
cumulative frequency	součtová četnost

curve fitting	vyrovnání křivkou
cutting-points	mezní hodnoty
cycle component	cyklická složka
D	
decomposition	rozklad
decile	decil
decision	rozhodování
decision tree	rozhodovací strom
definite integral	určitý integrál
degrees of freedom	stupně volnosti
density	hustota
density function	funkce hustoty pravděpodobnosti
dependence measurement	měření závislosti
descending	sestupný
descriptive statistics	popisná statistika
design of sample	výběrový plán
deviate square	čtvercová odchylka
difference	rozdíl, diference
discontinuous function	nespojitá funkce
discrete variable	diskrétní proměnná
dispersion	rozptyl
distance	vzdálenost
distribution function	distribuční funkce
distribution fitting	proložení optimálního rozdělení
distribution plotting	zobrazení distribuční funkce
E	
effcience	vydatnost
empiric value	empirická hodnota
empirical distribution	empirické rozdělení
equation	rovnice
error	chyba
error of estimation	chyba odhadu
error of measurement	chyba měření
estimate value	odhadovaná hodnota
estimation	odhad
exceed	převyšovat
exclusive events	neslučitelné jevy
expected value	očekávaná hodnota

exploratory analysis	průzkumová analýza
exponential curve	exponenciální křivka
exponential function	exponenciální funkce
extent of dispersion	variační rozpětí
extrapolation	extrapolace
F	
file	soubor
first-order	první řád
fixed base index	bazický index
fixed weights index	index stálého složení
forecasting	předpovídání
fraction	zlomek
frequency	četnost
frequency polygon	polygon četností
full-scope survey	vyčerpávající zjišťování
G	
Gauss' curve	Gaussova křivka
Gauss' normal equations	normální rovnice
general population	základní soubor
geometric mean	geometrický průměr
goodness-of fit	dobrá shoda
grouping	třídění
growth coefficient	koeficient růstu
growth curve	růstová křivka
H	
harmonic mean	harmonický průměr
histogram	histogram
hypothesis testing	testování hypotézy
I	

inconsistence	neslučitelnost
indefinite integral	neurčitý integrál
independence hypothesis	hypotéza o nezávislosti
independent event	nezávislý jev
index of shift in proportions	index struktury
interaction	interakce
interaction of events	průnik jevů
intercept	úsek, absolutní člen
interquartile range	kvartilové rozpětí
J	
judgment sample	záměrný výběr
K	
kurtosis	špičatost
L	
large sample	velký výběr
least squares	nejmenší čtverce
least squares method	metoda nejmenších čtverců
level	hladina
level of significance	hladina významnosti
likelihood	věrohodnost
line plot	spojnicový graf
linear correlation	lineární korelace
linear interpolation	lineární interpolace
logistic curve	logistická křivka
lottery sampling	výběr losováním
lower quartile	dolní kvartil
M	

main average	základní průměr
marginal distribution	marginální rozdělení
marginal frequency	marginální četnost
mean difference	střední diference
mean square error	střední kvadratická chyba
mean value	střední hodnota
measurable characteristic	měřitelný znak
median	medián
method of moments	momentová metoda
middle quartile	prostřední kvartil
midpoint	střed
missing value	chybějící hodnota
mode	modus
moving average	klouzavý průměr
moving series	klouzavá řada
multiple comparisons	vícenásobné porovnání
multiple correlation	mnohonásobná korelace
mutually exclusive events	vzájemně neslučitelné jevy
N	
nested sampling	vícestupňový výběr
notched box plot	vrubový krabicový graf
non replication sampling	výběr bez vracení
non-parametric method	neparametrická metoda
normal approximation	normální aproximace
normal curve	normální křivka
normalized moment	normovaný moment
normalized variable	normovaná proměnná
null hypothesis	nulová hypotéza.
O	
observation	pozorování
odd	lichý
one-factor analysis	jednofaktorová analýza
one-sample analysis	jednovýběrová analýza
one-tailed	jednostranný
opposite event	opačný jev
option	volba

order	pořadí
outcome	výsledek
outlier	odlehlý
P	
paired samples	párově uspořádané výběry
partial correlation	dílčí korelace
partial correlation coefficient	dílčí korelační koeficient
partial regression coefficient	dílčí regresní koeficient
patterned sampling	mechanický výběr
percentage	relativní četnost
percentile	percentil
periodical fluctuation	periodické kolísání
piechart	kruhový graf
point	bod
point estimation	bodový odhad
population size	rozsah základního souboru
power function	mocninná funkce
power of the test	síla testu
probability	pravděpodobnost
probability distribution	rozdělení pravděpodobnosti
probability of event	pravděpodobnost jevu
p-value	p-hodnota
Q	
quantiles	kvantily
quartiles	kvartily
R	
random error	náhodná chyba
random event	náhodný jev
random experiment	náhodný experiment

random fluctuation	náhodné kolísání
random function	náhodná funkce
random number	náhodné číslo
random variable	náhodná proměnná
range	rozpětí
rank	pořadí, postavení
rare event	vzácný jev
rate of growth	tempo růstu
reciprocal function	lomená funkce
region of variation	variační obor
regression curve	regresní křivka
regression function	regresní funkce
regression line	regresní čára
rejection region	obor zamítnutí
relative error	relativní chyba
relative frequency	relativní četnost
relative increase	relativní přírůstek
reliability	spolehlivost
replicated experiment	opakovaný pokus
replication sampling	výběr s vracením
representative sample	reprezentativní výběr
residual deviation	reziduální odchylka
residual variance	reziduální rozptyl
row	řádek
S	
sample	výběr
sample average	výběrový průměr
sample size	rozsah výběru
sample survey	výběrové zjišťování
sample total	výběrový úhrn
sample unit	výběrová jednotka
sample values	výběrová data
sampling	výběrová metoda
sampling characteristic	výběrová charakteristika
sampling error	výběrová chyba
sampling fraction	výběrový podíl
sampling frame	opora výběru
sampling interval	výběrový krok
sampling population	výběrový soubor
sampling variance	výběrový rozptyl
scatter plot	bodový graf
seasonal index	sezónní index
seasonal variation	sezónní kolísání

significance level	hladina významnosti
simple correlation	jednoduchá korelace
simple random sample	prostý náhodný výběr
single-stage sampling	jednostupňový výběr
skewness	špičatost
slope	směrnice
small sample	výběr malého rozsahu
smoothing	vyrovnávání
solution	řešení
standard deviation	směrodatná odchylka
standard error	směrodatná chyba
statistic population	statistický soubor
statistic unit	statistická jednotka
statistical analysis	statistická analýza
statistical hypothesis	statistická hypotéza
statistical inference	statistická indukce
statistical measurement	statistické měření
statistical survey	statistické šetření
stem and leaf diagram	diagram stonek s listy
stepwise regression	kroková regrese
sth-order	s-tého řádu
stochastic variable	náhodná veličina
stratified sampling	oblastní výběr
subset	podmnožina
sum of squares	součet čtverců
survey	pozorování, přehled
survey frequency	pozorovaná četnost
symmetrical distribution	souměrné rozdělení
systematic error	systematická chyba
systematic sampling	systematický výběr
T	
test of significance	test významnosti
testing	testování
theoretical frequency	teoretická četnost
time series	časová řada
time series correlation	korelace časových řad
total	úhrn
trend	trend
trend line	trendová čára
true value	skutečná hodnota
truncation error	chyba metody
two-factor analysis	dvoufaktorová analýza
two-sample analysis	dvouvýběrová analýza

two-stage sample	dvoustupňový výběr
two-tailed	dvoustranný
U	
unbiased estimate	nestranný odhad
uncertainty	nejistota
ungrouped data	netříděná data
uniform distribution	rovnoměrné rozdělení
unimodal distribution	jednovrcholové rozdělení
union of events	sjednocení jevů
unit of population	jednotka souboru
unit of sampling	jednotka zjišťování
unknown	neznámý
unweighted average	prostý průměr
upper quartile	horní kvartil
V, W	
value	hodnota
variance	rozptyl
weighted	vážený
weighted mean	vážený průměr
width	šířka