Statistika

Studijní opora

Jaroslav Stuchlý

2017
České Budějovice

2. vydání

ISBN 978-80-7468-021-2

Vydala: Vysoká škola technická a ekonomická v Českých Budějovicích, Okružní 10, 370 01 České Budějovice

Za obsahovou a jazykovou správnost odpovídá autor.

Cíl předmětu

Cílem předmětu je seznámit studenty se základními postupy z oblasti statistické indukce, metodami analýzy kvalitativních i kvantitativních znaků a s elementy analýzy časových řad.

Výstupy z učení

Student po absolvování předmětu umí definovat základní postupy z oblasti statistické indukce, umí charakterizovat a aplikovat metody analýzy kvalitativních i kvantitativních znaků a elementy analýzy časových řad. Absolvent umí shromažďovat, třídit, zpracovávat a prezentovat statistická data.

Základní okruhy studia

1. Metody popisné statistiky;
2. Základní statistické charakteristiky;
3. Pravděpodobnost a rozdělení pravděpodobností a jejich číselné charakteristiky;
4. Základní pravděpodobnostní modely;
5. Výběrová šetření, rozdělení výběrových charakteristik a základy statistické indukce;
6. Testování statistických hypotéz;
7. Dvouvýběrové testy;
8. Další testy a analýza rozptylu;
9. Jednoduchá lineární regrese a korelace;
10. Statistická indukce v regresním modelu;
11. Vícerozměrná regrese a prognostická aplikace regrese;
12. Úvod do analýzy časových řad.
13. Analýza časových řad.

Povinná literatura

MAREK, Luboš. *Statistika v příkladech*. Druhé vydání. Praha: Kamil Mařík - Professional Publishing, 2015, 425 stran. ISBN 978-80-7431-153-6.

STUCHLÝ, Jaroslav. *Statistika: studijní opora pro kombinované studium.* 1. vyd. České Budějovice: Vysoká škola technická a ekonomická v Českých Budějovicích, 2012. 197 s. ISBN 978-80-7468-021-2.

Studijní průvodce

 **- Klíčové pojmy**

 **- Cíle kapitoly**

 **- Čas potřebný ke studiu kapitoly**

 **- Výklad**

 **- Úkoly k zamyšlení a diskuzi**

 **- Klíč k řešení otázek**

 **- Studijní materiály**

Kapitola 1: Metody popisné statistiky

Klíčové pojmy:

popisná statistika, statistické jednotky, znaky, proměnné, základní a výběrový soubor, rozsah souboru, klasifikace proměnných, statistická šetření, rozdělení četností, histogram, sloupkový a výsečový diagram, polygon, skupinové rozdělení četností, vícerozměrné rozdělení četností, modus, medián, koeficient mutability, nominální a ordinální variance

Cíle kapitoly:

- **pochopení základních pojmů z popisné statistiky;**
- **seznámení s etapami statistického zkoumání;**
- **znalost pojmů tabulka rozdělení četností, intervalové rozdělení četností, dvourozměrné tabulky rozdělení četností a příslušné statistické grafy;**
- **popisování rozdělení nominální a ordinální proměnné číselnými charakteristikami.**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly

Vznik a význam statistiky
Základní statistické pojmy
Etapy statistických prací
Elementární zpracování dat (tabulky a grafy)
- o kvalitativní (nominální) proměnné;
- o pořadové (ordinální) proměnné;
- o kvantitativní (numerické) proměnné s malým a velkým počtem obměn;
- o vícerozměrné proměnné.

Statistická analýza nominální a ordinální proměnné.

Život nedal nic lidem, co by nezaplatili velkou námahou
Horatius

Vznik a význam statistiky

Poznání stále pronikavěji zasahuje do všech stránek našeho života. Každých 10-15 let se množství znalostí zdvojnásobuje. Orientace v oborech lidské činnosti a jejích výsledcích je stále náročnější. Vzrůstá rozsah informací (údaje o hromadných jevech), ale i jejich cena (informace jsou zbožím). Jejich zkoumání a vyhodnocování se stalo důležitou náplní praktické i teoretické statistiky.

- První použití statistických operací – Čína, Řím před naším letopočtem: sčítání obyvatel, zařazení do daňových skupin.

- První statistické analýzy – 17. století: Anglie - J. Graunt, W: Petty (zpracování údajů z matriky).

- Termín „statistika“ vznikl v 18. století (G.Achenwall):

    - Latinsky status = stav → status rei publicae = stav věci veřejné → italské státo = stát → státistico = statistický, statistik → státistica = statistika.
    - Statistikové byli vzdělaní muži, kteří byli velmi ceněni.

- Další rozvoj statistiky 19. a 20. století: Bernoullli, Laplace, Gauss, Pearson, Fisher, Janko, Hájek.

Moderní statistika 20. století vznikla z úředních zjišťování, univerzitní státovědy, politické aritmetiky a teorie pravděpodobnosti.

Lze ji chápat nejméně ve 3 pojetích: Jako

- číselné údaje o hromadných jevech;
- praktickou činnost spočívající ve sběru, zpracování a vyhodnocování statistických údajů;
- teoretickou disciplínu, zabývající se metodami zkoumání hromadných jevů.

Nachází široké uplatnění ve všech oblastech lidské činnosti:

- biologie, medicína, fyzika, technické disciplíny,
- ekonometrie, marketing, sociálně-ekonomické vědy,
- významná podpora pro manažerské rozhodování.

Základní statistické pojmy

Hromadné jevy (opak individuálních jevů):

- masově se vyskytují a mohou se libovolně opakovat;
- měříme je u prvků, které nazýváme statistické jednotky;
- to co měříme, nazýváme statistické znaky či proměnné.

Statistický soubor:

- určitá věcně, prostorově a časově vymezená množina všech zkoumaných statistických jednotek, u kterých zjišťujeme hodnoty sledovaných statistických znaků;
- jednorozměrný, dvourozměrný, vícerozměrný;
- základní soubor (populace) a výběrový soubor (vzorek) – obsahuje všechny nebo jen vybrané jednotky.

Rozsah souboru – počet jeho statistických jednotek:

- Základní: N;
- Výběrový: n.

Rozdělení statistických znaků

Zdroj: vlastní

Klasifikace podle stupnice, na které jsou znaky měřeny:

- Nominální (kvalitativní, slovní),
- ordinální (pořadové),
- kardinální znaky (kvantitativní, číselné).

Etapy statistických prací

Statistická šetření (zjišťování):

- použití sekundárních dat (publikovaná – ČSÚ, ČNB apod.),
- primární data – získaná od zpravodajských jednotek nebo respondentů (přímé pozorování, dotazník, anketa; z výkazů).
    - Rozhodná doba u intervalových proměnných, rozhodný okamžik u okamžikových proměnných.
    - Šetření úplné = vyčerpávající (cenzus) a neúplné = dílčí (výběr, zatíženo výběrovou chybou); expediční, korespondenční (telefonické nebo přes internet).
    - Výběr pravděpodobnostní = náhodný (reprezentativní) a nenáhodný (záměrný = úsudkový, kvótní aj.).
- Statistické zpracování (na počítači):
    - kontrola dat, tabulka, třídění a shrnování dat, číselné charakteristiky.
- Statistické vyhodnocování (rozbor) a prezentace dat:
    - slovní text, prezentační tabulka nebo graf, v prezentačním programu na počítači.

Elementární zpracování dat

O nominální a ordinální proměnné:

- tabulka variant a jejich četností (frekvencí)
    - absolutních n_i - počet výskytů i-té varianty,
    - relativních $p_i = n_i/n$ - poměr výskytů i-té varianty;
- graf rozdělení četností:

- sloupcový diagram (histogram);
    - vertikální = svislý nebo horizontální = vodorovný,
    - jednoduchý nebo skupinový;
- polygon rozdělení četností (lomená čára);
- výsečový, koláčový diagram;
- plošný graf.

- Tvary rozdělení: jednovrcholové, vícevrcholové, symetrické, různě šikmé a špičaté.

O pořadové nebo kvantitativní proměnné:

- Jednostupňové třídění do tabulky rozdělení četností.
- U diskrétní proměnné s malým počtem obměn vyjadřujeme:
    - hodnoty obměn x_i (třídní znak),
    - absolutní četnosti (frekvence) n_i,
    - relativní četnosti $p_i = n_i/n$,
    - kumulované absolutní četnosti $N_i = n_1+n_2+\ldots+n_i$,
    - kumulované relativní četnosti $M_i = p_1+p_2+\ldots+p_i$:
        - představují tzv. empirickou distribuční funkci,
        - násobené stem udávají, jaké % rozsahu souboru má menší nebo rovnou hodnotu proměnné než je x_i,
        - kumulované absolutní (relativní) četnosti udávají počty (podíly) statistických jednotek, pro které je uvažovaná proměnná rovna nejvýše x_i (tj. $X \leq x_i$).

- Rozdělení znázorňujeme obvykle sloupkovým diagramem (histogramem) nebo polygonem.

Příklad – viz Stuchlý (1999a), s. 32-33.

U diskrétní kvantitativní proměnné s velkým počtem obměn nebo u spojité kvantitamivní proměnné vyjadřujeme třídění do tabulky třídního (intervalového) rozdělení četností.

- Počet tříd nejčastěji určuje Sturgesův vzorec: $k = 1+3{,}322 \log_{10}(n)$ (zaokrouhlíme na celé číslo).

- Obvyklý počet tříd k 5 – 20 (při menším – přílišná redukce dat a při větším – nepřehlednost výsledků).

- Délka třídního intervalu $h = R/k$ (zaokrouhlíme), variační rozpětí $R = x_{max}-x_{min}$.

Příklad – viz Stuchlý (1999a), s. 33-34.

Elementární zpracování dat o vícerozměrné proměnné:

- více kontigenčních tabulek – podle dvojic znaků (v R);

- kontingenční tabulka s hierarchickou strukturou (větvení v řádcích, resp. v sloupcích, viz Excel);

Grafy: dvourozměrné histogramy, skupinový diagram, bodový (rozptylový) diagram.

Speciální tabulky:

- asociační tabulky – podle 2 kvalitativních znaků;

- korelační tabulky – podle 2 kvantitativních znaků.

Příklad – viz Stuchlý (1999a), s. 34-35.

Statistická analýza nominální proměnné

- Kategorie seřazujeme obvykle podle velikosti četností nebo podle abecedy.
- Pokud se u určitých otázek objevuje více odpovědí, nedostaneme tabulku rozdělení četností, ale jen tabulku počtu voleb (vyhodnocení procenty z počtu obměn nebo z rozsahu výběru).

- Poloha – modus (hodnota s nejvyšší četností).
- Variabilita:
  - Koeficient mutability $M = \frac{n^2 - \sum n_i^2}{n(n-1)}$
    - Je $0 \leq M \leq 1$, přitom $M = 0$ znamená 1 obměnu a $M = 1$ je n obměn.
  - Nominální variance: nomvar = $\frac{k}{k-1}\left(1 - \sum_{i=1}^{k} p_i^2\right)$
    - Interpretace je obdobná jako u M.

Příklad – viz Stuchlý (2011), s. 40.

Statistická analýza ordinální proměnné

- Poloha – medián (prostřední hodnota) a modus.
- Variabilita - diskrétní ordinální variance $\text{dorvar} = \frac{4}{k-1}\sum_{i=1}^{k} M_i(1 - M_i)$,

  kde M_i, resp. F_i jsou kumulativní relativní četnosti.

Příklad – viz Řezanková-Löster (2009), s. 22.

Podrobnější popis metod popisné statistiky najdeme zejména v učebnici Cyhelský (2001), s. 13-55.

Tabulky rozdělení četností a jejich grafy lze získat v Excelu pro kvalitativní proměnnou pomocí prostředku Kontingenční tabulka a pro numerickou proměnnou pomocí nástroje Histogram v Analýze dat - viz Řezanková-Löster (2009), s. 39-42. Výpočty je také možné provést v interaktivní nabídce R-Commanderu (viz řešení následujících úkolů).

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 11-29. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE Praha 1999. S. 30-36. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 7-26. ISBN 80-245-0178-3.

BÍNA V. a kol. *Jak na jazyk R.* J. Hradec: FM VŠE, 2006.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 13-55. ISBN 80-7261-003-1.

GIBILISCO, S. *Statistika bez předchozích znalostí.* Brno: Computer Press, 2009. s. 35-46. ISBN 978-80-251-2465-9.

HINDLS, R. a kol. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000. S. 11-17. ISBN 80-7261-013-9.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 11-20, 37-44. ISBN 978-80-86446-40-5.

MINAŘÍK, B. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita, 1995. S. 9-58. ISBN 80-7157-166-0.

ŘEZANKOVÁ, H. a T. LÖSTER. *Úvod do statistiky*. Praha: Oeconomica, 2009. S. 7-22, 39-44, ISBN 978-80-245-1514-4

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 9-29. ISBN 80-7187-058-7.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/)

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 51-85. ISBN 80-7169-089-9.

? Otázky a úkoly

1) Pracujte se souborem byty.xls. Úkoly:

   a) Načíst data do Excelu a charakterizovat typ jednotlivých proměnných;

   b) vytvořit tabulku rozdělení absolutních a relativních četností podle proměnné čtvrť a znázornit je graficky histogramem, resp. sloupcovým diagramem nebo výsečovým diagramem (použít kontingenční tabulky a grafy);

   c) vytvořit tabulku rozdělení všech četností podle proměnné počet obyvatel a znázornit je graficky sloupkovým diagramem nebo histogramem (použít z analýzy dat histogram);

   d) vytvořit tabulku rozdělení všech četností podle proměnné obytná plocha a znázornit je graficky sloupkovým diagramem (použít z Analýzy dat Histogram v Excelu);

   e) vytvořit kontingenční tabulku pro proměnné čtvrť, obytná plocha a vybavení telefonem

2) Pro proměnnou a) čtvrť b) kategorie ze souboru byty.xls určete charakteristiky úrovně a variability a interpretujte výsledky.

3) Načtěte do programu R data ze souboru studenti.dat a určete v tomto programu a) tabulku rozdělení četností a její graf pro proměnnou „doprava“, b) tabulku rozdělení absolutních a relativních četností a histogram pro proměnnou „výška“. c) dvojrozměrnou tabulku rozdělení četnosti pro proměnné „pohlaví“ a „výška“ a znázorněte je graficky.

? Úkoly k zamyšlení a diskuzi

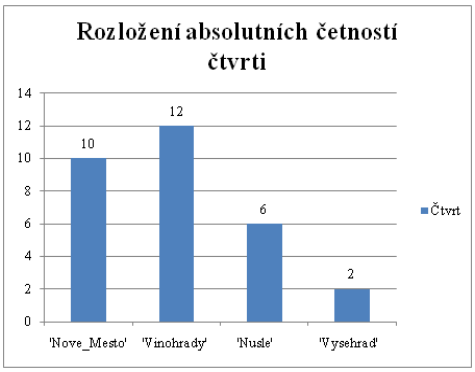
1) Jaká kritéria budete považovat za důležitá při sestavování reprezentativního výběru osob pro a) předvolební průzkum, b) marketingový průzkum prodeje aut, c) průzkum ohrožení populace cévními chorobami? Vyberte z těchto navrhovaných: věk, krevní tlak, pohlaví, barva očí, národnost, velikost obce bydliště, členství v politické straně, tělesná výška, náboženské vyznání. Která z nich jsou nejdůležitější v bodě a), b), c)? Zkuste navrhnout další kritéria!

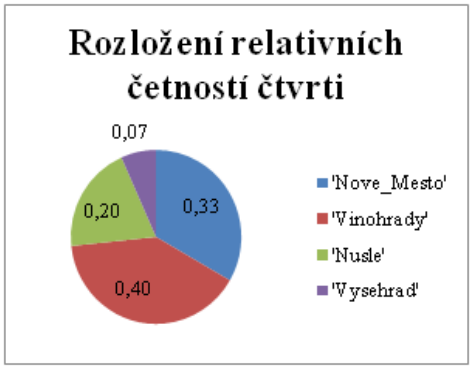
2) Sestavte statistický soubor ze svých přátel a známých a roztřiďte je současně podle pohlaví a podle toho, zda jsou kuřáci či nekuřáci. Sestavte asociační tabulku a znázorněte ji graficky.

Klíč k řešení otázek:

1) a) Klasifikace statistických znaků a popisné statistiky v Excelu: Nominální množná: Čtvrť, nominální alternativní: Telefon, ordinální: Kategorie, numerická diskrétní: Počet obyvatel, numerická spojitá: Obytná plocha a Nájemné.

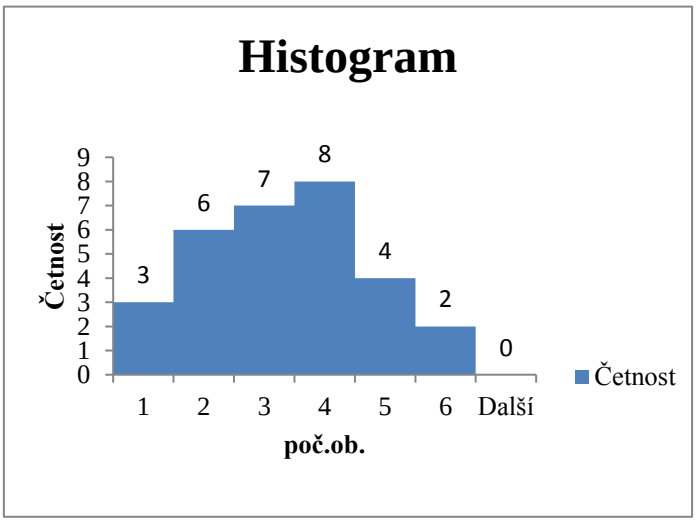
b) Sloupcový diaagram a koláčový diagram čtvrti, tabulka rozdělení četností a histogram počtu obyvatel:





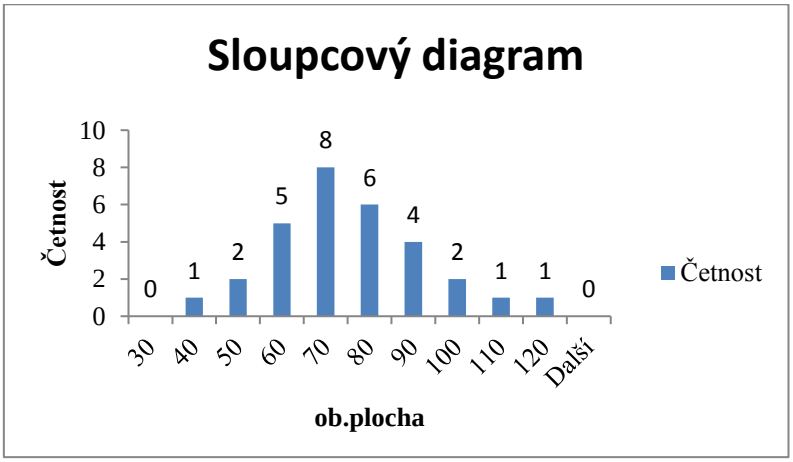
c)

poč.ob.	četnost
1	3
2	6
3	7
4	8
5	4
6	2
Součet	30



d) Variační rozpětí je R = 116,0 - 34,8 = 34,8. Podle Sturgesova vzorce je počet tříd k = 1 + 3,322 log 30 = 5,907. Délka třídního intervalu je h = 81,2/5,907 = 13,746. Pro lepší přehlednost zaokrouhlíme h = 10. Třídní rozdělení četností a sloupcový diagram:

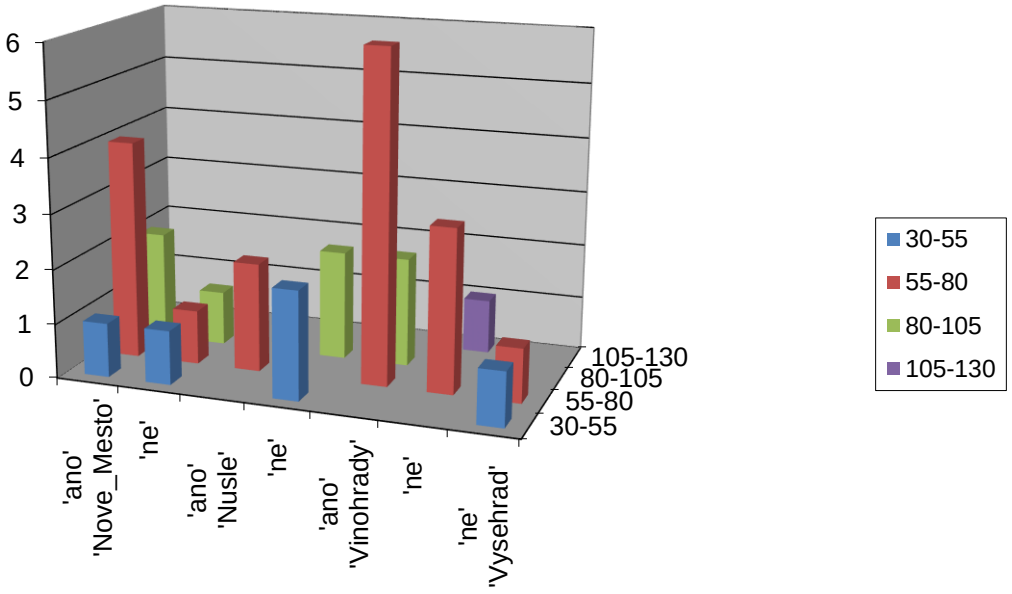
Obytná plocha (v m^2)	x_i	n_i	p_i	N_i	M_i
31 - 40	35	1	0,033	1	0,033
41- 50	45	2	0,067	3	0,1
51 - 60	55	5	0,167	8	0,267
61 - 70	65	8	0,267	16	0,534
71 - 80	75	6	0,2	22	0,734
81 - 90	85	4	0,133	26	0.867
91 - 100	95	2	0,067	28	0,934
101 - 110	105	1	0,033	29	0,967
111 - 120	115	1	0,033	30	1
Součet	×	30	1	×	×



e) Kontingenční tabulka s hierarchickou strukturou a její dvourozměrný histogram:

Počet z Ctvrť		Ob.plocha				
Ctvrť	Telefon	30-55	55-80	80-105	105-130	Celkový součet
'Nove_Mesto'	'ano'	1	4	2		7
	'ne'	1	1	1		3
Celkem z 'Nove_Mesto'		2	5	3		10
'Nusle'	'ano'		2			2
	'ne'	2		2		4
Celkem z 'Nusle'		2	2	2		6
'Vinohrady'	'ano'		6	2		8
	'ne'		3		1	4
Celkem z 'Vinohrady'			9	2	1	12

'Vysehrad'	'ne'	1	1			2
Celkem z 'Vysehrad'		1	1			2
Celkový součet		5	17	7	1	30



2) a) Jde o nominální proměnnou. Pro výpočet použijeme tabulku

Čtvrt	n_i	p_i	n_i^2	p_i^2
'Nove_Město'	10	0,333	100	0,111
'Vinohrady'	12	0,400	144	0,160
'Nusle'	6	0,200	36	0,040
'Vyšehrad'	2	0,067	4	0,004
Součet	30	1	284	0,316

Úroveň popisuje modus = „Vinohrady“, variabilitu popisuje koeficient mutability

$$M = \frac{n^2 - \sum n_i^2}{n(n-1)} = \frac{30^2 - 284}{30(30-1)} = 0{,}708,$$

nebo nominální variance

$$\text{nomvar} = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} p_i^2\right) = \frac{4}{4-1}(1 - 0{,}316) = 0{,}913.$$

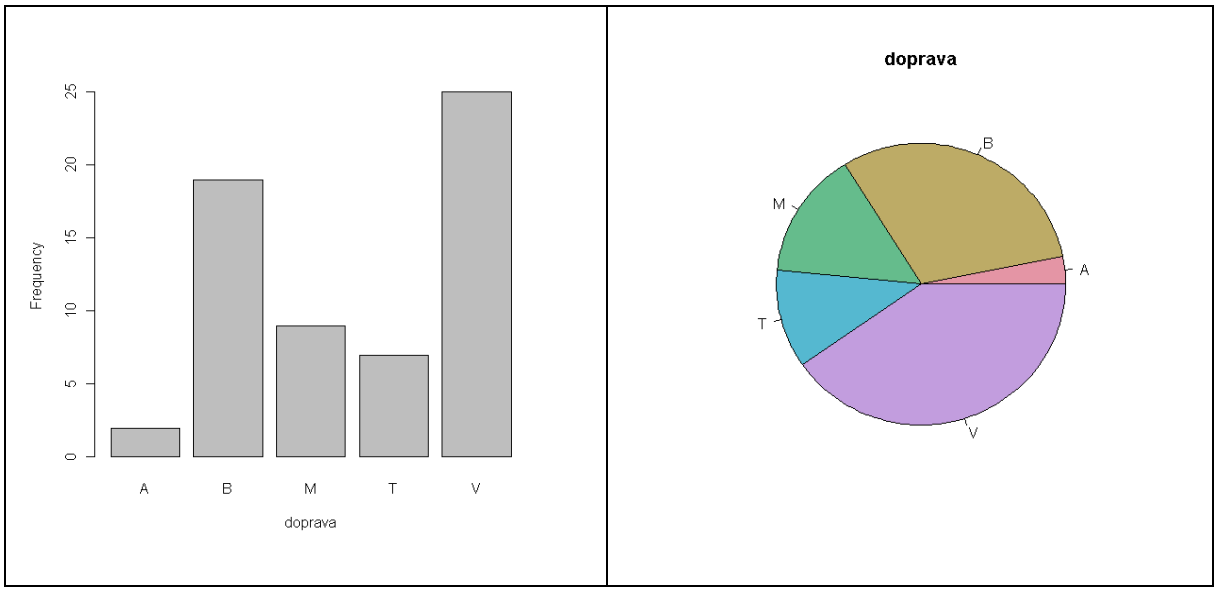
Hodnoty blízké 1 ukazují na velkou variabilitu (měnlivost) hodnot nominální proměnné „čtvrť“.

b) Jde o ordinální proměnnou. Úroveň charakterizuje modus nebo medián, variabilitu ordinální variance. Dále postupujeme obdobně.

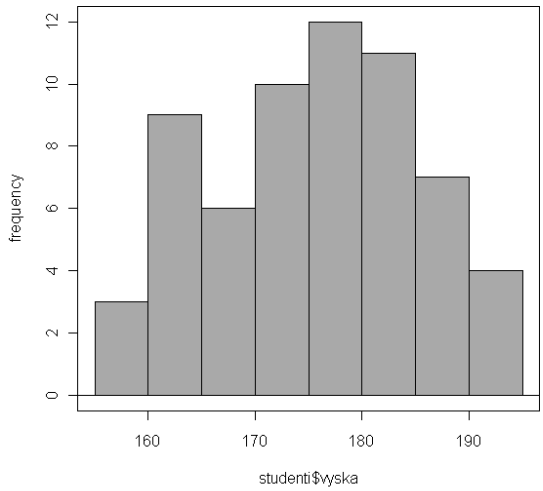
3) R je pro potřeby výuky volně šiřitelný program. Instalce je možná z Internetu nebo jen zkopírováním na učebně. Program R otevřeme pomocí příslušné ikonky „R“ (otevře se *R Konzola*). V *R Konzole* napíšeme příkaz *library(Rcmdr)* a odešleme ho *Enterem*. Tím se otevře nadstavba *R Commander*. Obsahuje vstupní okno (*Script Window*) na vkládání příkazů, výstupní okno (*Output Window*), kde dostáváme výsledky výpočtů a úplně dole dialogové okno (*Messages*), v kterém se budou zobrazovat chybová hlášení či jiné informace. Základní poznatky o práci s systémem R jsou v Bína a kol. (2006) nebo můžeme použít *Help v R Commanderu*: *Introduction to Rcmdr*. Datový soubor studenti.dat načteme z nabídky *Data* pomocí *Import data from text file*. (v dialogovém okně vypíšeme do okna *Enter name for data set*: studenti a po stlačení *OK* zadáme cestu k umístnění datového souboru). Tabulku načteného datového souboru si můžeme prohlédnout stlačením tlačítka u *Data set* (v našem případě se toto tlačítko označilo „studenti“). Tlačítko Edit data set otvírá editor dat (po provedení editace ho zavíráme křížkem X v pravém horním rohu). Program rozlišuje velká a malá písmena a používá desetinné tečky. Většinu výpočtů je možno provádět v R interaktivně (s využitím nabídky *R Commanderu*), avšak některé musíme provádět příkazově (vypsáním a odesláním příslušného příkazu ve vstupním okně). Přehled příkazů najdeme v Stuchlý (2011). Modifikátory k těmto příkazům lze vyhledat v nápovědě.

   a) Popisná statistika v R:Určení tabulky rozdělení absolutních a relativních četností a jejích grafů pro proměnou doprava použijeme v nabídce *Statistics* položku *Summaries Frequency distribution* (označíme proměnnou „doprava“) a z nabídky *Graphs* položku *Bar graph*, resp. *Pie Chart*. Na výstupu se objeví tabulky rozdělení četností a grafy (grafy se automaticky kreslí do *R Konzoly*):

```
> .Table  # counts for doprava

A  B  M  T  V
2 19  9  7 25

> 100*.Table/sum(.Table)  # percentages for doprava
A        B        M        T        V
3.225806 30.645161 14.516129 11.290323 40.322581
```



b) Protože R-ko kreslí histogram jen pro kvantitativní proměnné a tabulku rozdělení počítá jen pro kategoriální proměnné, nakreslíme nejdříve histogram pomoci nabídky *Graphs* a položky *Histogram* a proměnou „výška“:



Z grafu vidíme, že optimální třídní rozdělení četností (založené na Sturgesově vzorci) je do 8 tříd délky 5 na intervalu od 150 do 200 cm. Budeme proto kategorizovat výšky do těchto tříd v *Data - Manage variables in active data set - Compute new variable* (vyplníme *New variable name*: vyska_k, *Expression to compute*: cut(studenti$vyska, breaks=seq(150,200,by=5)), v editoru se objeví nová kategoriální proměnná vyska_k,

zavřeme editor a použijeme pro ni *Statistics Summaries Frequency distribution* a dostane rozdělení absolutních relativních četností:

```
> .Table  # counts for vyska_k

(150,155] (155,160] (160,165] (165,170] (170,175] (175,180] (180,185] (185,190]
        1         2         9         6        10        12        11         7
(190,195] (195,200]
        4         0

> 100*.Table/sum(.Table)  # percentages for vyska_k

(150,155] (155,160] (160,165] (165,170] (170,175] (175,180] (180,185] (185,190]
 1.612903  3.225806 14.516129  9.677419 16.129032 19.354839 17.741935 11.290323
(190,195] (195,200]
 6.451613  0.000000
```

Kumulované četnosti se počítají pomocí příkazu

cumsum (table(studenti$vyska_k))

a kumulované relativní četnosti (v %) pomocí příkazu

cumsum(100*table(studenti$vyska_k))/sum(table(studenti$vyska_k))

c) Použijeme *Statistics Contigency tables* - Two-way table, zaškrtneme pohlavi a vyska_k a No percentages (pro absolutní četnosti) nebo Percentages of total (pro relativní četnosti v %) a Chi-square test of independence zatím odškrtneme.

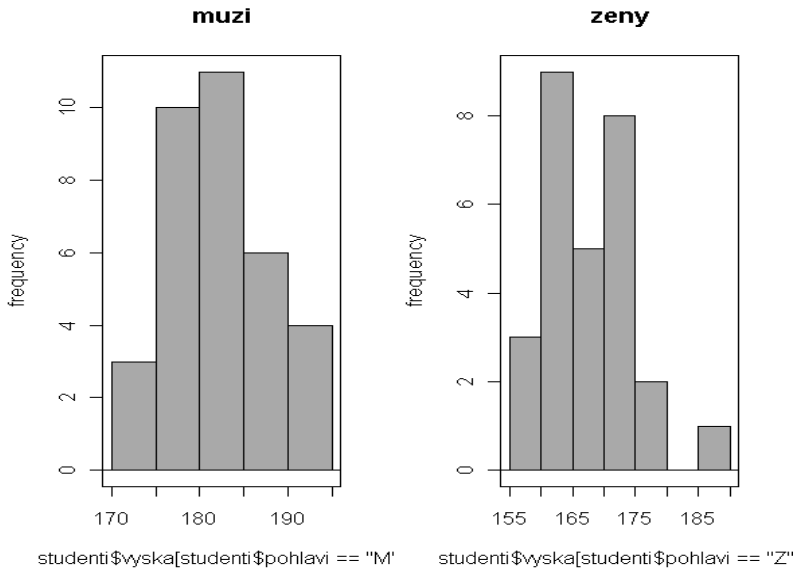
```
       vyska_k
pohlavi (150,155] (155,160] (160,165] (165,170] (170,175] (175,180] (180,185]
      M         0         0         0         1         2        10        11
      Z         1         2         9         5         8         2         0
       vyska_k
pohlavi (185,190] (190,195] (195,200]
      M         6         4         0
      Z         1         0         0
```

Graf dostaneme vypsáním programu:

```
par(mfrow=c(1,2))
```

```
Hist(studenti$vyska[studenti$pohlavi=="M"], scale="frequency", main="muzi", col="darkgray")
Hist(studenti$vyska[studenti$pohlavi=="Z"], scale="frequency", main="zeny", col="darkgray")
```

do vstupního okna a jeho odesláním pomocí *Submit*:



Kapitola 2: Základní statistické charakteristiky

Klíčové pojmy:

číselné charakteristiky (míry) statistického znaku (proměnné), prostý a vážený aritmetický průměr, vlastnosti průměru, modus, medián, kvantily, kvartily, odlehlá hodnota, variační a kvartilové rozpětí, prostý a vážený rozptyl, populační a výběrový rozptyl, vlastnosti rozptylu, meziskupinový a vnitroskupinový rozptyl, směrodatná odchylka, variační koeficient, absolutní a relativní kvartilová odchylka, koeficient asymetrie (šikmosti) a špičatosti, kovariance, korelační koeficient, kovarianční a korelační matice, krabicový diagram

Cíle kapitoly:

- **pochopení významu jednotlivých číselných charakteristik statistického znaku;**
- **porozumění vlastnostem aritmetického průměru a rozptylu;**
- **znalost výpočtu a věcné interpretace jednotlivých číselných charakteristik pomocí vhodného software (Excel, R).**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly.

Charakteristiky numerické proměnné

- Charakteristiky polohy
- Charakteristiky variability
- Charakteristiky tvaru rozdělení

Kvantilové charakteristiky

Grafické znázornění číselných charakteristik

Charakteristiky vícerozměrné veličiny

Struktura výkladu

Charakteristiky numerické proměnné

Charakteristiky polohy (úrovně)

- Popisují obecnou úroveň znaku
- Udávají střed rozdělení četností (střední hodnoty)
- Známe již modus a medián

Průměry:

Motto: Statistika je předmět, který nás učí, že když stojíme jednou nohou na rozpálené plotně a druhou máme v mrazáku, jsme na tom v průměru dobře.

Prostý aritmetický průměr

- populační $\mu=\frac{1}{N}\sum_{i=1}^{N} x_i$

- výběrový $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$

- Interpretace aritmetického průměru – jaká část z celkového úhrnu připadne na jednu jednotku;
    - fyzikálně: těžiště.
- Z dat agregovaných v tabulce rozdělení četností dostaneme vážený průměr:
    - váhy – absolutní četnosti n_i: $\bar{x}=\frac{1}{n}\sum_{i=1}^{k} x_i n_i$
    - váhy - relativní četnosti p_i : $\bar{x}=\sum_{i=1}^{k} x_i p_i$

Vlastnosti aritmetického průměru:

a) Přičteme-li k jednotlivým hodnotám znaku konstantu, zvýší se o tuto konstantu i aritmetický průměr.
b) Aritmetický průměr konstanty je opět roven konstantě.
c) Násobíme-li jednotlivé hodnoty znaku konstantou, je touto konstantou násoben i průměr.
d) Součet jednotlivých odchylek od průměru je nulový.
e) Součet čtverců odchylek hodnot znaku od jeho aritmetického průměru je minimální.
f) Je-li statistický soubor rozdělen na k dílčích podsouborů, v nichž známe jednotlivé dílčí průměry $\bar{x}_i$ a počty pozorování (absolutní četnosti) n_i, i = 1,…,k, potom celkový aritmetický průměr se rovná váženému aritmetickému průměru dílčích průměrů s vahami rovnými absolutním četnostem.

- Uvedené vlastnosti je možno zapsat pomocí následujících vzorců:

a) $\overline{x \pm c} = \bar{x} \pm c$ b) $\bar{c} = c$;

c) $\overline{c.x} = c.\bar{x}$; d) $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$;

e) $\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - a)^2$; f) $\bar{x} = \frac{1}{\sum_{i=1}^{k} n_i}\sum_{i=1}^{k} \bar{x}_i\, n_i$

- Kromě aritmetického průměru používáme v některých situacích harmonický, geometrický nebo kvadratický průměr – viz Hindls (2007), s. 32-34.

Příklad:

- Doba pobytu pacientů v nemocnici je 5 9 6 6 9 8 9 6 38 5 9.
- Aritmetický průměr je

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{11}(5+9+6+6+9+8+9+6+38+5+9)=\frac{110}{11}=10.$$

- Je aritmetický průměr za 1. týden dostatečně spolehlivou (typickou) střední hodnotou, když deset z jedenácti pacientů strávilo v nemocnici kratší dobu než 10 dní? (Vliv extrémní hodnoty 38 dní.)
- Spolehlivější je zde medián (prostřední hodnota): 5 5 6 6 6 **8** 9 9 9 9 38, medián = 8.

Charakteristiky variability:

- Popisují měnlivost (rozptýlenost=variabilitu) hodnot znaku (vyrovnanost, homogenita hodnot znaku).
- Malá variabilita znamená malou vzájemnou různost hodnot znaku, v tomto případě je průměr dobrou mírou.
- Vysoká variabilita značí velkou vzájemnou odlišnost hodnot znaku, pak průměr není dobrá míra.
- Známe již míru variační rozpětí $R = x_{max} - x_{min}$.

Absolutní míry variability:

- Rozptyl prostý (variance)
  - populační: $\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2,$
  - výběrový: $s_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ nebo $s'^2_x=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$
- Charakterizuje rozptýlenost hodnot znaku kolem aritmetického průměru.
- Platí $s^2=\frac{n-1}{n}s'^2$ a $s'^2=\frac{n}{n-1}s^2.$

- Pro ruční výpočet je vhodnější používat následující výpočetní vzorce místo předcházejících definičních:
- $$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2, \text{ resp. } s'^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$
- Pro data shrnutá do tabulky rozdělení četností používáme:
- Výběrový rozptyl vážený $s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$, resp. $s^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \bar{x}^2$.
- Rozptyl měříme ve čtvercích měrných jednotek hodnot znaku. Proto je vhodnější místo rozptylu používat jeho odmocninu. Dostaneme míru:
- Směrodatná odchylka
  - populační σ;
  - výběrová s (resp. s').
- Absolutní odchylka $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$.
- Vlastnosti rozptylu:
  1. Rozptyl konstanty je roven nule, tj. $s_c^2 = 0$
  2. Rozptyl je vždy nezáporný tj. $s_x^2 \geq 0$.
  3. Přičteme-li ke všem hodnotám znaku konstantu, rozptyl se nezmění, tj. $s_{x+c}^2 = s_x^2$.
  4. Násobíme-li všechny hodnoty znaku konstantou, rozptyl je násoben čtvercem této konstanty, tj. $s_{cx}^2 = c^2 s_x^2$.
  5. Předpokládejme, že statistický soubor o rozsahu n je rozdělen do k dílčích podsouborů kde známe dílčí rozptyly s_i^2, dílčí průměry $\bar{x}_i$ a dílčí četnosti n_i. Potom rozptyl celého souboru je dán součtem rozptylu skupinových průměrů (meziskupinový rozptyl) a váženému průměru ze skupinových rozptylů (vnitroskupinový rozptyl), tj. $s^2 = s_{\bar{x}}^2 + \overline{s^2}$, kde

$$s^2_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i, \qquad \overline{s^2} = \frac{\sum_{i=1}^{k} s_i^2 n_i}{\sum_{i=1}^{k} n_i}$$

Meziskupinový rozptyl měří variabilitu mezi skupinami a vnitroskupinový rozptyl variabilitu vnitroskupinovou. Na principu tohoto rozkladu je založena analýza rozptylu.

Relativní míry variability:

- Jedná se o variabilitu vztaženou na jednotku znaku.

- Variační koeficient $V = \frac{\sigma}{\mu} \cdot 100\%, \quad V(x) = \frac{s_x}{\bar{x}} \cdot 100\%.$

  - Udává, kolik procent průměru činí směrodatná odchylka. Je-li V > 50%, je soubor silně nesourodý a není vhodné používat $\bar{x}$.

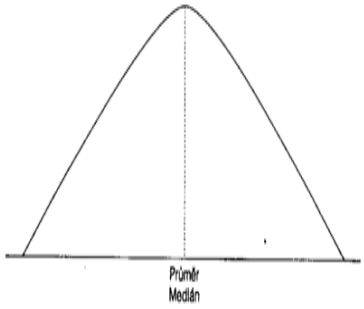
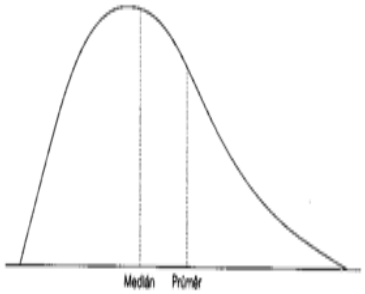
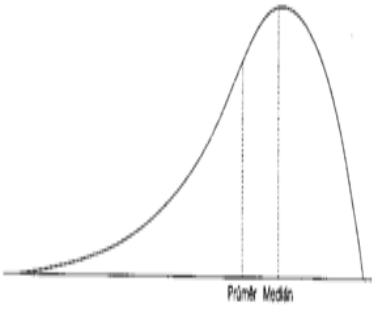
  - Platí: $V(x \pm c) = \frac{s}{\bar{x} \pm c}; \quad V(c \cdot x) = \frac{c.s}{c.\bar{x}} = \frac{s}{\bar{x}} = V(x).$

- V řešeném příkladu je V = 8,99.100/10 = 89,8% (silně nehomogenní data).

Charakteristiky tvaru rozdělení četností:

- Koeficient asymetrie (šikmosti) a špičatosti rozdělení četností
  - Charakteristiky jsou založeny na srovnání stupně koncentrace malých a velkých hodnot pozorovaného znaku.
  - Je-li stejný počet podprůměrných a nadprůměrných hodnot je rozdělení symetrické – levý obrázek (průměr = mediánu).
  - Převažují-li velké hodnoty, jde o rozdělení s kladnou šikmostí – prostřední obrázek (průměr > medián).
  - Převažují malé hodnoty, jde o rozdělení se zápornou šikmostí – pravý obrázek (průměr < medián).

Grafy symetrických a asymetrických rozdělení:

		
Symetrické rozdělení	Kladná šikmost	Záporná šikmost

Číselné charakteristiky tvaru rozdělení:

- Koeficient šikmosti (asymetrie) $a_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns_x^3}.$
    - Rozdělení symetrické: $a_x = 0$, zleva zešikmené $a_x > 0$, zprava zešikmené $a_x < 0$.
- Koeficient špičatosti $b_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns_x^4} - 3.$
    - Špičatost jako u standardního normálního rozdělení: $b_x = 0$, špičatější $b_x > 0$, méně špičaté $b_x < 0$.

Kvantilové charakteristiky

Úroveň popisujeme kvantily. Z nich nejvíce používaný je medián.

- p kvantil x_p – bod, který dělí hodnoty seřazené podle velikosti přibližně v poměru p:(1-p).
- Přesnější zápis:
    - (relativní četnost jednotek x_i, pro něž je $x_i \leq x_p$) $\geq$ p;
    - (relativní četnost jednotek x_i, pro něž je $x_i > x_p$) $\leq 1 - p$.
- Výpočet lze provádět z tabulky kumulovaných relativních četností.
- Medián $x_{0,5}$ je prostřední hodnota v posloupnosti dat srovnaných podle velikosti při lichém počtu měření a průměr z prostředních dvou měření při sudém počtu měření.
    - Jinak řečeno: (aspoň polovina hodnot je $\leq x_{0,5}$ a nejvýše polovina je $> x_{0,5}$)

- Jiná označení $x_{50,}$ $\tilde{x}$.

- Kvartily $x_{0,25}$, $x_{0,50}$, $x_{0,75}$.
- Decily $x_{0,1}$, $x_{0,2}$,…,$x_{0,9}$.
- Percentily $x_{0,01}$, $x_{0,02}$, …, $x_{0,99}$.

Absolutní variabilitu popisuje:

Kvartilové rozpětí $R_q = x_{0.75} - x_{0,25}$ a kvartilová odchylka $Q = (x_{0,75} - x_{0,25})/2$.

Odlehlé hodnoty jsou hodnoty nižší než $x_{0,25} - 1{,}5\ R_q$ nebo vyšší než $x_{0,75} + 1{,}5\ R_q$.

Relativní variabilitu popisuje relativní kvartilová odchylka $Q_{rel} = (x_{0,75} - x_{0,25})/ (x_{0,75} + x_{0,25})$

Příklad: Budeme charakterizovat data z řešeného příkladu o pobytu pacientů v nemocnici rozptylem a směrodatnou odchylkou. Výpočet provedeme v tabulce:

data x_i	průměr $\bar{x}$	odchylka $x_i - \bar{x}$	čtv.odchylky $(x_i - \bar{x})^2$	čtv.dat x_i^2
5	10	-5	25	25
5	10	-5	25	25
6	10	-4	16	36
6	10	-4	16	36
6	10	-4	16	36
8	10	-2	4	64
9	10	-1	1	81
9	10	-1	1	81
9	10	-1	1	81
9	10	-1	1	81
38	10	28	784	1444
Součet		0	890	1990

- Počítáme populační rozptyl $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{11}890 = 80{,}9091\ \text{dní}^2.$

- Směrodatná odchylka je s = 8,99 dní – velmi vysoká variabilita způsobená jednou extrémní hodnotou.

Grafické znázornění číselných charakteristik

Hodnoty kvantilových charakteristik znázorňujeme krabicovým diagramem v R. Krabicový diagram – obdélník, 2 vousy a body. Dolní a horní hrana obdélníku představují dolní a horní

kvartil, dělicí čára uvnitř představuje medián. Dolní vous představuje menší z hodnot x_{min} a $x_{0,25} - 1{,}5\ R_q$ a horní vous představuje vyšší z hodnot x_{max} a $x_{0,75} + 1{,}5\ R_q$. Body představují odlehlé (extrémní) hodnoty.

Celé rozdělení znázorňujeme histogramem nebo polygonem v Excelu nebo R. Z těchto grafů je také možné odhadnout přibližnou hodnotu číselných charakteristik.

Výpočet číselných charakteristik v Excelu (viz Řezanková-Löster 2009, s. 39-42):

- pomocí statistických funkcí;
- pomocí popisné statistiky z Analýzy dat.

Výpočet číselných charakteristik v R Commanderu:

- interaktivně: Statistics-Summaries-Active data set, Statistics-Summaries-Numerical summaries…, Statistics-Summaries-Table of statistics… (počítá charakteristiky podmíněné hodnotami nominální proměnné), Statistics-Summaries-Correlation matrix…, Graphs-Histogram, Graphs-Boxplot nebo pomocí příkazů (viz Stuchlý 2011).

Příklady: Viz Stuchlý (1999a), s. 41 (prosté číselné charakteristiky), s. 42 (vážené číselné charakteristiky)), s. 45 (rozklad rozptylu), s. 46 (charakteristiky tvaru rozdělení), s. 47-48 (grafy).

Charakteristiky vícerozměrných proměnných

Podmíněné průměry a rozptyly (počítané v závislosti na hodnotách kategoriální proměnné).

- Výpočet v Excelu – vyfiltrováním dat podle kategoriální proměnné a přímo v R.

Kovariance $s_{xy} = \mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

- Obecně platí $s_{x\pm y}^2 = s_x^2 \pm s_{xy} + s_y^2$,
- Jsou-li znaky x, y nezávislé, je $s_{x\pm y}^2 = s_x^2 + s_y^2$.

Korelační koeficient $r_{xy} = \mathrm{cor}\,(X,Y) = \frac{s_{xy}}{s_x s_y}$.

- Měří těsnost lineární závislosti. Platí:

$r_{yx} = r_{xy}$,

$-1 \leq r_{xy} \leq 1$,

$r_{xy} = 1 \Leftrightarrow$ mezi proměnnými je přímá funkční lineární závislost,

$r_{xy} = -1 \Leftrightarrow$ mezi proměnnými je nepřímá funkční lineární závislost,

$r_{xy} = 0 \Leftrightarrow$ proměnné jsou nekorelované.

Jsou-li proměnné x, y lineárně nezávislé, je $s_{xy} = r_{xy} = 0$.

Pro více proměnných počítáme kovarianční matici (na diagonále rozptyly, mimo kovariance) a korelační matici (na diagonále 1 mimo korelační koeficienty).

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 29-47. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE Praha, 1999. S. 37-50. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Skripta VŠE, Praha 2003, s. 27-53, ISBN 80-245-0178-3

BÍNA V. a kol. *Jak na jazyk R.* J. Hradec: FM VŠE, 2006.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 56-81. ISBN 80-7261-003-1.

GIBILISCO, S. *Statistika bez předchozích znalostí.* Brno: Computer Press, 2009. S. 46-5. ISBN 978-80-251-2465-9.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada Publishing, 1999. S. 12-40. ISBN 80-7169-255-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 21-37. ISBN 978-80-86446-40-5.

MINAŘÍK, B. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita, 1995. S. 61-93. ISBN 80-7157-166-0.

ŘEZANKOVÁ, H. a T. LÖSTER. *Úvod do statistiky*. Praha: Oeconomica, 2009. S. 22-45, ISBN 978-80-245-1514-4.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 33-51. ISBN 80-7187-058-7.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/).

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 87-130. ISBN 80-7169-089-9.

WONNACOT, T.H. a R.J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 33-48. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Pracujte se souborem byty.xls. Řešte v Excelu:
    a) Několika vhodnými způsoby charakterizujte polohu proměnných obytná plocha a počet obyvatel. Použijte vhodné funkce Excelu.
    b) Několika vhodnými způsoby charakterizujte absolutní a relativní variabilitu proměnných obytná plocha a počet obyvatel. Použijte vhodné funkce Excelu. Interpretujte výsledky.
    c) Řešte úkoly a), b) pomocí popisné statistiky v Analýze dat.
    d) Určete koeficient asymetrie a špičatosti a ověřte, zda jsou získané výsledky v souladu s grafem rozdělení obou proměnných.
    e) Určete meze pro odlehlé hodnoty pro proměnnou obytná plocha.
    f) Určete decily pro proměnnou obytná plocha

g) Pomocí filtru rozdělte data proměnné obytná plocha podle kategoriální proměnné vybavení telefonem a určete příslušné podmíněné průměry a výběrové směrodatné odchylky.
h) Určete kovarianční a korelační matici pro proměnné obytná plocha, počet obyvatel a nájemné a interpretujte výsledky.

2) Načtete do R data ze souboru studenti.dat. Úkoly:

a) Pro proměnnou výška vypočítat průměr, standardní odchylku a kvartily a znázornit výsledek krabicovým diagramem.
b) Pro proměnnou výška vypočítat podmíněný průměr, standardní odchylku a kvartily podle pohlaví a znázornit výsledek krabicovým diagramem.
c) Určete kovarianční a korelační matici pro proměnné *vyska*, *vaha*, *test* a interpretujte výsledky.

3) Byla vypočtena průměrná mzda 21037 Kč a rozptyl mezd 360000. Určete průměrnou mzdu a směrodatnou odchylku mezd, pokud

a) každý pracovník dostane přidáno 500 Kč,
b) každý pracovník dostane 1,5 násobek platu,
c) každý pracovník dostane přidáno 5% ze stávajícího platu.

? Úkoly k zamyšlení a diskuzi

1) Představte si, že nějaký test píše velký počet lidí a že každý jednotlivý žák dosáhne přesně polovinu správných odpovědí. V tomto případě bude směrodatná odchylka (vyberte a zdůvodněte správnou odpověď):

    a) rovna průměru,
    b) rovna mediánu,
    c) rovna nule,
    d) směrodatnou odchylku nebude možné určit bez více odpovědí.

2) Zamyslete se nad tím, jak dokážeme platnost výpočetního vzorce pro rozptyl.

Klíč k řešení otázek:

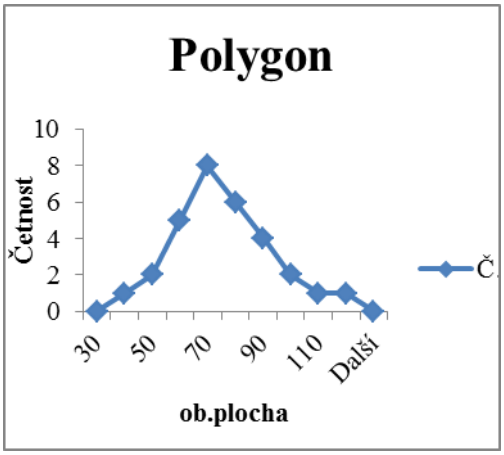
1) Číselné charakteristiky v Excelu: Vložíme data do sloupců v Excelu.

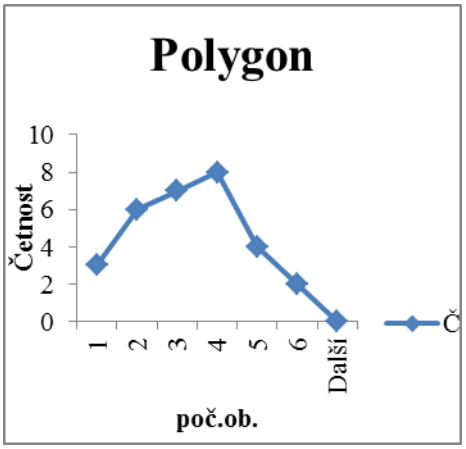
    a) Aplikujeme na příslušné sloupce v Excelu statistické funkce průměr a medián a dostaneme pro obytnou plochu: průměr = 70,46, medián = 69,35 a pro počet obyvatel: průměr = 3,33, medián = 3.
    b) Pro absolutní variabilitu použijeme SMODCH.VÝBĚR a kvartilovou odchylku Q (pro její výpočet použijeme Q = (QUARTIL(D2:D31;3) - QUARTIL(D2:D31;1))/2 dostáváme pro obytnou plochu s = 17,63, Q = 10,5 a pro počet obyvatel s = 1,40, Q = 1. Relativní variabilitu budeme charakterizovat variačním koeficientem V (V = SMODCH.VÝBĚR/ PRŮMĚR) a relativní kvartilovou odchylku Q_{rel} (Q_{rel} = (QUARTIL(D2:D31;3) – QUARTIL (D2:D31;1))/(QUARTIL(D2:D31;3) + QUARTIL(D2:D31;1)). Dostáváme pro obytnou plochu V = 0,25, Q_{rel} = 0,15 a pro počet obyvatel V = 0,41, Q_{rel} = 0,33. Obě proměnné jsou homogenní, homogennější je obytná plocha.
    c) Na kartě *Data* stlačíme tlačítko *Analýza dat* (instalace viz Řezanková-Löster 2009, s. 41) a vybereme položku *Popisná statistika*. Ve vstupním okně vyplníme

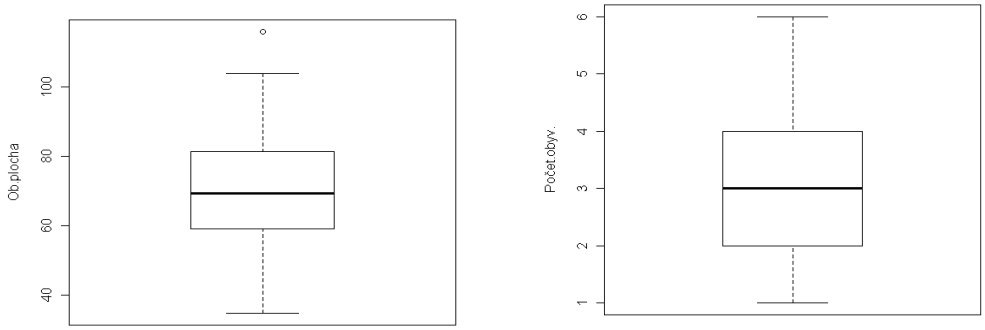
Vstupní data (sloupce obou proměnných), zaškrtneme popisky v 1. řádku a *Celkový přehled*. Dostaneme následující tabulku výsledků, kde ještě dopočteme V. Kvartilové odchylky je lepší počítat pomocí statistických funkcí.

Ob.plocha		*Počet obyvatel*	
Stř. hodnota	70,457	Stř. hodnota	3,333
Chyba stř. hodnoty	3,2191	Chyba stř. hodnoty	0,255
Medián	69,35	Medián	3
Modus	93	Modus	4
Směr. odchylka	17,632	Směr. odchylka	1,398
Rozptyl výběru	310,89	Rozptyl výběru	1,954
Špičatost	0,5753	Špičatost	-0,654
Šikmost	0,4713	Šikmost	0,084
Variační rozpětí	81,2	Variační rozpětí	5
Minimum	34,8	Minimum	1
Maximum	116	Maximum	6
Součet	2113,7	Součet	100
Počet	30	Počet	30
Variační koeficient	0,2503	Variační koeficient	0,419

d) Výsledky jsou v předchozí tabulce. Polygony získáme z vkládání grafů v Excelu a krabicové diagramy v R:







e) Odlehlé hodnoty: Dolní odlehlá mez = $x_{0,25}$-1,5.R_q = 59,78-1,5.21 = 28,28, horní odlehlá mez = $x_{0,75}$+1,5.R_q = 80,78+1,5.21 = 112,28 (dílčí hodnoty počítáme pomocí statistických funkcí).

f) Pro p = 0,1 použijeme funkci =PERCENTIL($A$2:$A$31;B2) a potahováním za pravý dolní růžek výsledkového okna dostaneme další decily. Výpočet je v následující tabulce:

Ob.plocha	p	x_p
82,6	0,1	52,18
57,3	0,2	56,84
70,4	0,3	62,13
65	0,4	65,54
48,4	0,5	69,35
103,8	0,6	71,68
73,6	0,7	77,8
43,5	0,8	82,92
66,1	0,9	93
93	1	116

g) Vyfiltrujeme proměnnou Obytná plocha podle kategoriální proměnné Telefon a překopírujeme ji do dvou sloupců na nový list. Výsledky získáme pomocí funkcí PRŮMĚR a SMODCH.VÝBĚR. Pro byty s telefonem je průměrná obytná plocha 71,19 a sm.odchylka = 13,19 a byty bez telefonu 69,5 a 22,75. Byty bez telefonu mají v průměru menší obytnou plochu ale vyšší variabilitu.

h) Použijeme z Analýzy dat nástroj Kovariance a Korelace a dostaneme kovariančí matice a korelační matice:

	Ob.plocha	*Počet obyv.*	*Nájemné*
Ob.plocha	300,523789		
Počet obyv.	12,9744444	1,88888889	
Nájemné	6950,89233	49,1666667	588580,82

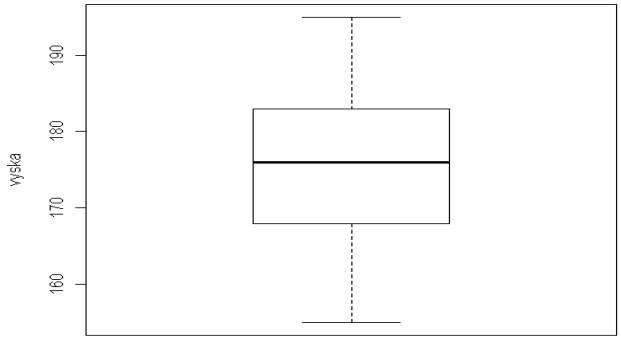
	Ob.plocha	*Počet obyv.*	*Nájemné*
Ob.plocha	1		
Počet obyv.	0,5445605	1	
Nájemné	0,5226345	0,0466299	1

Mezi Počtem obyvatel a Obytnou plochou a Nájemným a Obytnou plochou je středně silná lineární přímá závislost, mezi Nájemným a Počtem obyvatel je velmi slabá lineární závislost.

2) Načteme soubor studenti.dat do R Commanderu.

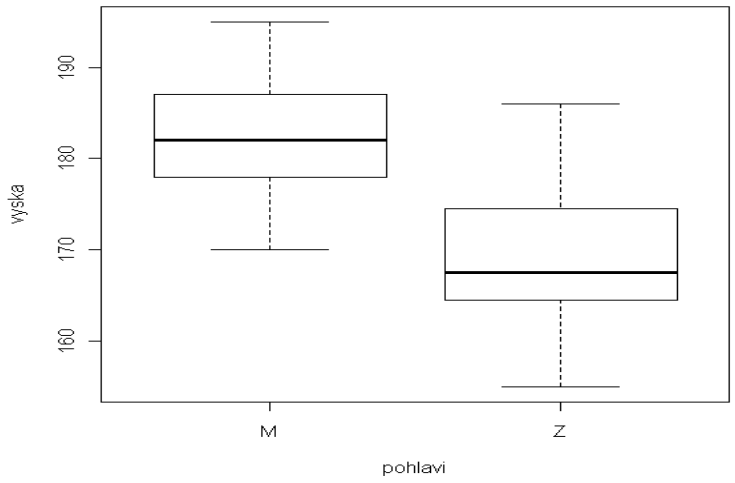
a) Číselné charakteristiky v R: Užijeme z nabídky *Statistics-Summaries-Numerical summaries*. Ve vstupním okně označíme proměnnou výška. Krabicový diagram dostaneme z *Graphs-Boxplot.* Výsledky:

```
    mean       sd   0%   25% 50% 75%   100% n
176.2903 9.48923 155 168.25 176 183  195  62
```



b) Použijeme *Statistics-Summaries-Table of statistics*. Ve vstupním okně označíme proměnnou výška a pohlaví. Graf dostaneme opět z *Graphs-Boxplot* jen ve vstupním okně po označení výška stiskneme tlačítko *Plot by Groups…* a označíme pohlaví. Podmíněné průměry a rozptyly:

```
      mean       sd  0%    25%   50%    75% 100%  n
M 182.7059 6.032887 170 178.25 182.0 186.75  195 34
Z 168.5000 6.647194 155 164.75 167.5 174.25  186 28
```



c) Pro korelační matici užijeme *Statistics–Summaries–Correlation matrix* a pro kovarianční matici příkazu *cov(studenti[,c("test","vaha","vyska")])*. Výsledky:

```
>cor(studenti[,c("test","vaha","vyska")],use="complete.obs")
```

```
test      vaha      vyska
test  1.00000000 0.1833762 0.06266131
vaha  0.18337621 1.0000000 0.88564944
vyska 0.06266131 0.8856494 1.00000000

> var(studenti[,c("test","vaha","vyska")], use="complete.obs")
            test      vaha      vyska
test  265.769236  41.57787   9.693548
vaha   41.577869 193.43443 116.885246
vyska   9.693548 116.88525  90.045479
```

Silná lineární přímá závislost je jen mezi váhou a výškou.

3) Příklad vychází z vlastností aritmetického průměru a z vlastností rozptylu. Přidání 5% vyjádříme jako vynásobení konstantou 1,05.

   a) $\bar{x} = 21037 + 500 = 21537$; $s_x = \sqrt{(360000)} = 600$.

   b) $\bar{x} = 21037.1{,}5 = 31555{,}5$; $s_x = \sqrt{(1{,}5^2.360000)} = 1{,}5.600 = 900$,

   c) $\bar{x} = 21037.1{,}05 = 22088{,}85$; $s_x = \sqrt{(1{,}05^2.360000)} = 1{,}05.600 = 630$.

Kapitola 3: Pravděpodobnost a náhodná veličina

Klíčové pojmy:

význam pravděpodobnosti a její historický vývoj, náhodný pokus, náhodný jev, operace s jevy, elementární jev, základní jevový prostor, úplný systém jevů, klasická definice pravděpodobnosti, kombinatorika, variace, permutace, kombinace, vlastnosti pravděpodobnosti, podmíněná pravděpodobnost, složená, úhrnná a úplná pravděpodobnost, náhodná veličina a její rozdělení pravděpodobností, diskrétní a spojitá náhodná veličina, distribuční a pravděpodobnostní funkce, hustota pravděpodobnosti, střední hodnota, rozptyl a kvantil náhodné veličiny, vícerozměrná náhodná veličina a její rozdělení, kovarianční a korelační matice

Cíle kapitoly:

- **pochopení základních pojmů z počtu pravděpodobnosti;**
- **porozumění pojmu náhodná veličina X a její rozdělení pravděpodobnosti;**
- **znalost výpočtu a vlastností číselných charakteristik náhodné veličiny.**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly.

- Náhodný pokus a náhodný jev
    - Operace s náhodnými jevy
- Pravděpodobnost náhodného jevu
- Opakování kombinatoriky
- Pravidla pro počítání s pravděpodobnostmi
- Úplná pravděpodobnost
- Náhodná veličina
- Systém náhodných veličin a jejich rozdělení pravděpodobnosti
- Číselné charakteristiky náhodných veličin
    - Kvantily
- Číselné charakteristiky dvourozměrné náhodné veličiny

Struktura výkladu

Motto: Kdyby bylo vše jasné, tak by vám všechno přišlo úplně marné. Nuda by zaplavila svět bez stínů jakýmsi necitelným životem tvořeným nekvašenými dušemi. Naděje, která září na temný práh, nevychází ze světa nadměrné jistoty

Marcel Proust, spisovatel

Náhodný pokus a náhodný jev

Úvod:

- Teorie pravděpodobnosti studuje jevy a procesy, ve kterých se uplatňují prvky náhody. Představuje statistickou možnost kvantifikovat neurčitost, s kterou se setkávají firmy, podnikatelé i manažeři.
- Pravděpodobnost je jazykem neurčitosti.

- Neurčitost působí manažerům při rozhodování nemalé problémy. Kdyby manažer dokázal identifikovat přesně důsledky svých rozhodnutí, jistě by volil vždy tu nejlepší alternativu. Přesto musí manažer odhadnout důsledky alternativních možností a učinit jednoznačné rozhodnutí. K tomu musí umět situace popsat pomocí pravděpodobností.
- Pravděpodobnost hraje důležitou roli v marketingovém výzkumu. Princip technik marketingového výzkumu spočívá v tom, že shromáždí data jen o výběrovém souboru (např. zákazníků) a pomocí metod pravděpodobnosti přenáší závěry na celou populaci (statistická indukce - inference). Teorie pravděpodobnosti tvoří takto most mezi popisnou statistikou a statistickou indukcí.
- Historické začátky pravděpodobnostních zkoumání spadají do 17. století v souvislosti s řešením úloh z oblasti hazardních her.
- Další rozvoj následoval v 19. století a byl podmíněn prudkým rozvojem přírodních věd. Teoretické základy pravděpodobnosti jako vědy vybudovali matematici Bernoulli, Laplace, Gauss, Poisson, Čebyšev aj. Ve 30. letech našeho století vypracoval A. N. Kolmogorov matematickou teorii výstavby pravděpodobnosti.
- Pravděpodobnost má velký význam v přírodních a technických vědách a ve statistice. Buduje modely, které lze aplikovat ve všech oborech ekonomické teorie a praxe.
- Teorie pravděpodobnosti se nejdříve zabývá studiem náhodných jevů. Při zavádění tohoto pojmu vycházíme z tzv. náhodného pokusu. Pokusy, jejichž výsledky se mění, i když zachováváme stejné experimentální podmínky, nazýváme náhodné pokusy.
    - Např. hod kostkou, hod mincí, výběr kuliček z osudí, přesné měření tloušťky destičky ap.
- Náhodné jevy – jednotlivé výsledky náhodného pokusu nebo množiny těchto výsledků.
- Označení A, B, resp. A_1, A_2,…,A_n,…
- Jistý jev E, nemožný jev $\varnothing$.
- S jevy je možno pracovat jako s množinami, tj. můžeme zavést pojmy $A \subset B$, $A \supset B$, $A = B$ i složené jevy.
- Operace s náhodnými jevy:
- Jev $\bar{A}$ nazýváme opačný jev nebo komplementární k jevu A. Složené jevy $A \cup B$, $A \cap B$, A-B aj. znázorňujeme pomocí Vennových diagramů.
- Pro tyto operace platí pravidla, která známe z teorie množin, např. de Morganova pravidla $\overline{A \cup B} = \bar{A} \cap \bar{B}$, $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

- Jevy A, B budeme nazývat disjunktní (neslučitelné jevy), když $A \cap B = \emptyset$.
- Elementární jev e - nedá se dále rozložit na sjednocení podjevů.
- Základní jevový prostor E je množina všech možných jevů.
- Jevy $A_1,...,A_n$ tvoří úplný systém jevů, když platí:
- a) $A_i \cap A_j = \emptyset$, $\forall i, j=1,...,n, i \neq j$,
- b) $A_1 \cup A_2 \cup ... \cup A_n = E$.

Příklad l. Za náhodný pokus vezmeme hod kostkou. Potom:

- a) Elementární jev je např. $e_6=\{6\}$.
- b) Základní jevový prostor $E = \{1,2,3,4,5,6\}$.
- c) Náhodnými jevy jsou např. $\emptyset$, E, $A=\{2,4,6\}$ - padne sudé číslo, $\overline{A} =\{1,3,5\}$ - padne liché číslo, $B=\{5,6\}$ - padne číslo větší než 4. Platí

$$E = \bigcup_{i=1}^{6} e_i \,, A = e_2 \cup e_4 \cup e_6 \,, B = e_5 \cup e_6 \,.$$

- Při opakovaných náhodných pokusech provádíme pokus několikrát za sebou a při každém pokusu sledujeme uskutečnění jevu A.
    - Podle techniky provedení pokusů dělíme pokusy na
    - Nezávislé pokusy: v těchto pokusech není v daném opakování pokusu uskutečnění jevu A závislé na výsledcích předchozích pokusů. Nezávislými pokusy jsou např. ty pokusy, při nichž postupně vybíráme ze souboru určité prvky a ty před dalším výběrem (opakováním) vracíme zpět do souboru – tzv. výběr s vracením;
    - Závislé pokusy: v těchto pokusech je v daném opakování uskutečnění jevu A závislé na výsledcích předchozích pokusů; závislými pokusy jsou např. pokusy, při nichž postupně vybíráme ze souboru určité prvky a ty před dalším výběrem (opakováním) již nevrátíme zpět do souboru – tzv. výběr bez vracení.

Pravděpodobnost náhodného jevu

- Pravděpodobnost náhodného jevu A je číslo P(A), které můžeme interpretovat jako míru možnosti nastoupení (realizace) náhodného jevu.

- Existuje několik definic pravděpodobnosti. Historicky se způsob zavádění pravděpodobnosti vyvíjel od statistické pravděpodobnosti, přes klasickou pravděpodobnost (založenou na kombinatorických úvahách), geometrickou pravděpodobnost až po axiomatickou pravděpodobnost, která všechny předcházející způsoby zahrnuje a zobecňuje.

Klasická definice pravděpodobnosti:

- Nechť základní jevový prostor E je konečná n prvková množina, přičemž všechny elementární jevy jsou stejně možné. Nechť náhodný jev A má právě m příznivých případů. Potom pravděpodobnost jevu A definujeme vztahem $P(A) = \frac{m}{n}$.

- Jiná terminologie: P(A) je poměr počtu případů příznivých jevů A ku celkovému počtu všech možných výsledku náhodného pokusu.
- V příkladu 1 je $P(A) = \frac{1}{2}$, $P(\overline{A}) = \frac{1}{2}$, $P(B) = \frac{2}{6} = \frac{1}{3}$.

- Při výpočtu P(A) v případě závislých i nezávislých pokusů využijeme kombinatoriku. Proto si ji zopakujeme.

Kombinatorika:

- Je nauka o skupinách (množinách) prvků. Nejjednodušší skupiny vzniknou, vyberemeli z množiny n prvků podmnožiny k prvků (k-tici); k ≤ n. Podle způsobu výběru rozlišujeme:
- Variace k-té třídy z n prvků; vznikají v případě, že při výběru záleží na pořadí vybraných prvků. Variace dělíme na:
    - variace bez opakování: žádný již vybraný prvek se v k-tici nesmí opakovat; počet variací k-té třídy bez opakování z n prvků $V_k(n)$ je dán vzorcem

      $V_k(n) = \frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1)$;

    - variace s opakováním: vybrané prvky se v k-tici mohou opakovat. Platí $V_k(n) = n^k$;
    - permutace je variace n té třídy z n prvků, $P(n) = n! = n.(n-1)...2.1$ (tzv. faktoriál);
- Kombinace k-té třídy z n prvků; vznikají v případě, že při výběru nezáleží na pořadí vybraných prvků. Variace dělíme na:

- kombinace bez opakování: žádný již vybraný prvek se v k-tici nesmí opakovat; počet kombinací k-té třídy bez opakování z n prvků (kombinační číslo) $C_k(n)$ je dán vzorcem

$$C_k(n) = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 2.1};$$

- kombinace s opakováním: vybrané prvky se v k-tici mohou opakovat; počet kombinací k-té třídy s opakováním vybírané z n prvků $C'_k(n)$ je dán vzorcem

$$C'_k(n) = \binom{n+k-1}{k}.$$

- Vlastnosti kombinačních čísel:

$$\binom{n}{k} = \binom{n}{n-k}, \quad \binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}, \quad \binom{n}{k+1} = \frac{n-k}{k+1}\binom{n}{k}.$$

  - Excel umožňuje v sestavě svých matematických funkcí počítat i faktoriály a kombinační čísla.

Příklad 2. (klasická definice pravděpodobnosti) Ze 75 zaměstnanců provozovny, mezi nimiž je 50 mužů a 25 žen, bylo vybráno 10 zaměstnanců. Jaká je pravděpodobnost jevu A, že byli vybráni samí muži?

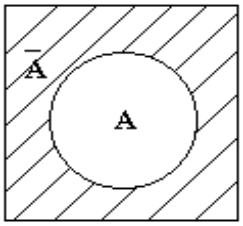
- *Řešení.*
- Vybíráme 10 zaměstnanců ze 75, ve skupině nezáleží na pořadí ani se nemohou zaměstnanci opakovat. Celkový počet možných výběrů $n = C_{10}(75) = \binom{75}{10}.$
- Podobně počet příznivých výběrů k jevu A je $m = C_{10}(50) = \binom{50}{10}.$
- Hledaná pravděpodobnost je $P(A) = \frac{m}{n} = \frac{\binom{50}{10}}{\binom{75}{10}} = 0{,}01239.$

- Ve statistice se často používá statistická definice pravděpodobnosti jako relativní četnost v sérii dostatečně velkého počtu n nezávislých náhodných pokusů. Obě uvedené definice pravděpodobnosti nejsou použitelné pro jevy, jež nelze aspoň za podobných podmínek opakovat. V těchto případech můžeme použít

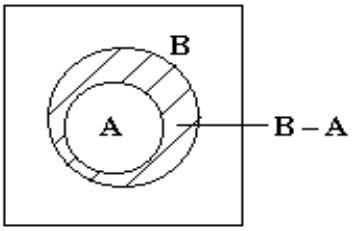
definice tzv. subjektivní pravděpodobnosti jako stupně důvěry jednotlivce ve výskyt uvažovaného jevu.

- V exaktní teorii pravděpodobnosti se používá definice axiomatická.

Základní vlastnosti pravděpodobnosti:



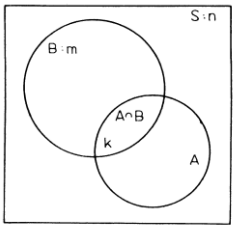
- Z uvedených definic dostaneme
  - a) $0 \le P(A) \le 1$,
  - b) $P(\varnothing) = 0$, $P(E) = 1$,
  - c) $P(A \cup B) = P(A) + P(B)$, jsou-li A, B disjunktní jevy.
- Odtud lze odvodit další vlastnosti, např.
  - d) $P(\bar{A}) = 1 - P(A)$ (pravděpodobnost opačného jevu)
  - e) $A \subset B \Rightarrow P(A) \le P(B)$ (monotónnost),
  - f) $A \subset B \Rightarrow P(B-A)=P(B)-P(A)$ (subtraktivnost).



Podmíněná pravděpodobnost:

- Podmíněná pravděpodobnost jevu A za podmínky, že již dříve nastal jev B, se definuje vztahem $P(A|B) = P(A \cap B) / P(B)$, pro $P(B) \neq 0$.

Příklad 3. V telefonní ústředně je ze 120 drátů 75 modrých a z nich je 54 zapojených. Vybereme náhodně modrý drát. Jaká je pravděpodobnost, že je zapojený?



- *Řešení:*
- Označíme jevy: A - drát je zapojený, B - drát je modrý. Počítáme:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{k}{n}}{\frac{m}{n}} = \frac{\frac{54}{120}}{\frac{75}{120}} = \frac{54}{75} = 0{,}72.$$

- Přímý výpočet podle klasické definice (místo základního jevového prostoru S uvažujeme B): $P(Z|M) = \frac{54}{75} = 0{,}72.$

Pravidla pro počítání s pravděpodobnostmi

Násobení pravděpodobností (složená pravděpodobnost):

- Z definice podmíněné pravděpodobnosti dostaneme

  $P(A\cap B) = P(A)P(B|A) = P(B)P(A|B)$

- Matematickou indukcí získáme zobecnění:

  $P(A_1\cap A_2\cap...\cap A_n)=P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1\cap A_2)...P(A_n \mid A_1\cap A_2\cap...\cap A_{n-1})$

- Nezávislost náhodných jevů:
- Říkáme, že jevy A,B jsou nezávislé, když platí $P(A\cap B) = P(A).P(B)$.

- Jsou-li jevy A, B nezávislé, je

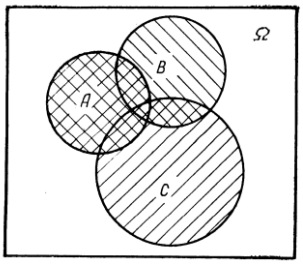
  $P(A \mid B) = P(A)$, $P(B \mid A) = P(B)$.

- O n jevech $A_1,...,A_n$ říkáme, že jsou nezávislé, když pro každou podmnožinu r jevů z množiny jevů $A_1,A_2,...,A_n$, $2\le r\le n$ (tj. pro každou dvojici, trojici,...,n-tici z jevů A_1, $A_2,...,A_n$) platí $P(A_{k_1}\cap A_{k_2}\cap...\cap A_{k_r})=P(A_{k_1})P(A_{k_2})...P(A_{k_r})$.
  - Jsou-li jevy $A_1,...,A_n$ nezávislé, jsou i po dvou nezávislé. Opačné tvrzení neplatí. K nezávislým jevům jsou nezávislé i jejich doplňky.
- Jsou-li jevy $A_1,...,A_n$ nezávislé, potom platí:

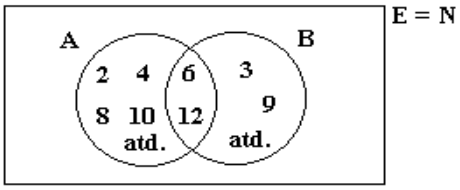
  $P(A_1\cap A_2\cap...\cap A_n)=P(A_1)P(A_2)...P(A_n)$

Sčítání pravděpodobností (úhrnná pravděpodobnost):

- Platí $P(A\cup B) = P(A) + P(B)$, jsou-li A, B neslučitelné jevy a
  $P(A\cup B) = P(A) + P(B) - P(A\cap B)$, jsou-li jevy A, B slučitelné.
- Zobecnění pro 3 slučitelné jevy:
  $P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$.
- Pro nezávislé náhodné jevy platí
  $P(A_1 \cup A_2 \cup... \cup A_n) = 1 - P(\bar{A}_1)P(\bar{A}_2)...P(\bar{A}_n)$



Příklad 4. Jaká je pravděpodobnost, že náhodně vybrané přirozené číslo je dělitelné 2 (jev A) nebo 3 (jev B)?

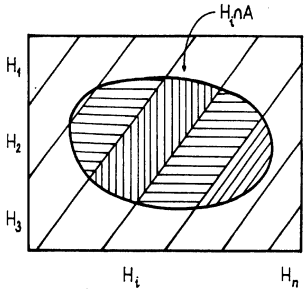


- Platí P(A) = 1/2, P(B) = 1/3, P(A∩B) = 1/6.
- Jevy jsou slučitelné. Proto
  P(A∪B) = P(A) + P(B) - P(A∩B) = 1/2 + 1/3 - 1/6
  = (3+2-1)/6 = 2/3.
- Číslo je dělitelné 2 nebo 3 s 66,7% pravděpodobností.

Úplná pravděpodobnost:

- Je-li $A \subset \bigcup_{i=1}^{n} H_i$,

- $P(H_i)>0$, i=1,...,n a jevy H_i (náhodné hypotézy) tvoří úplný systém jevů (viz obr.). Potom platí:
  $$P(A) = \sum_{i=1}^{n} P(H_i)P(A|H_i).$$



- *Důkaz* $A = \bigcup_{i=1}^{n}(H_i \cap A)$, $H_i \cap A$

  jsou neslučitelné =>

  $$P(A) = \sum_{i=1}^{n} P(A \cap H_i) = \sum_{i=1}^{n} P(H_i)P(A|H_i).$$

Příklad 5. Viz Stuchlý (1999a), s. 53.

- Složené pravděpodobnosti je možno počítat i pomocí tzv. pravděpodobnostních stromů (viz Wonnacot, T. H. a Wonnacot, R. J. 1993, s. 76-100 a Stuchlý 2004, s. 84-86).

Náhodná veličina

- Výsledkem většiny náhodných pokusů jsou reálná čísla (např. počet poruch, celkový počet padnutých bodů či minimum z počtu dosažených bodů při hodu 2 kostkami, pří-

jem či vydání čtyřčlenné rodiny apod.). Tedy výsledek náhodného pokusu, daný reálným číslem, můžeme považovat za hodnotu veličiny, kterou nazveme náhodná veličina. Náhodné veličiny označujeme velkými písmeny a jejich hodnoty odpovídajícími malými písmeny z konce abecedy a dělíme je na diskrétní a spojité náhodné veličiny.

- Náhodná veličina X je diskrétní, nabývá-li konečného nebo spočetného počtu hodnot. Náhodná veličina X je spojitá, může-li nabývat všech hodnot z konečného nebo nekonečného intervalu.

Rozdělení pravděpodobnosti náhodné veličiny

- Náhodná veličina je z pravděpodobnostního hlediska plně popsána, je-li známé její rozdělení pravděpodobnosti.
- Rozdělení pravděpodobností je vztah mezi hodnotami náhodné veličiny (pro diskrétní náhodné veličiny), resp. intervaly hodnot (pro spojité náhodné veličiny) a jejich pravděpodobnostmi.

Rozdělení pravděpodobnosti diskrétní náhodné veličiny:

- Pro popis rozdělení diskrétní náhodné veličiny používáme pravděpodobnostní funkci $p(x_i)$, která je určena zadáním pravděpodobností $P(X = x_i)$, $i = 1,2,\ldots,n$, že náhodná veličina nabude této hodnoty. Tyto pravděpodobnosti obvykle zapisujeme do tabulky:

x_i	x_1	x_2	…	x_n	Σ
$P(x_i)$	$P(x_1)$	$P(x_2)$	…	$P(x_n)$	1

- Grafickým zobrazením tabulky je polygon rozdělení pravděpodobnosti.
- Příklad viz Hindls a kol. (2007), s. 61-62.
- Základní formou popisu rozdělení pravděpodobnosti je distribuční funkce. Pro každé reálné číslo x udává pravděpodobnost, že nahodná veličina X nabývá hodnot ≤ než x. Distribuční fukci značíme F(x) a definujeme ji vztahem
  $F(x) = P(X \leq x)$.
- Vlastnosti distribuční a pravděpodobnostní funkce:
- F(x) je neklesající, zprava spojitou funkcí, nabývající hodnot od 0 do 1.

- Platí $F(-\infty) = 0$, $F(\infty) = 1$.

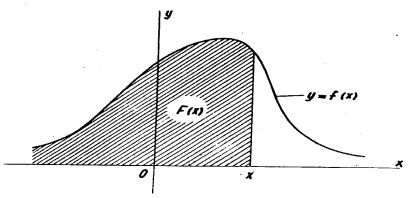
  $P(a < X \leq b) = F(b) - F(a) = \sum_{a<x_i \leq b} p(x_i)$.

  $F(x) = \sum_{x_i \leq x} p(x_i)$.

- Distribuční funkce diskrétní náhodné veličiny je schůdkovitá funkce.
- Příklad: Viz Marek a kol. (2007), s. 65-66.
  - V starší české literatuře se obvykle uvádí modifikovaná definice distribuční funkce s ostrou nerovností. Vlastnosti je potom třeba upravit.

Rozdělení pravděpodobnosti spojité náhodné veličiny:

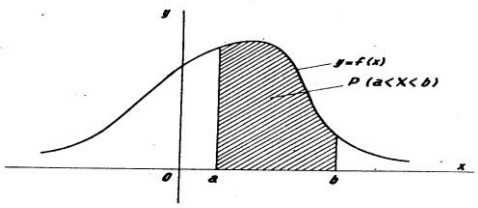
- Distribuční funkce spojité náhodné veličiny je spojitá rostoucí funkce, $0 \leq F(x) \leq 1$.
- Derivováním distribuční funkce dostaneme tzv. hustotu pravděpodobnosti f(x).



- Platí $f(x) = \frac{dF(x)}{dx}$, $F(x) = \int_{-\infty}^{x} f(t)dt$.
- Vlastnosti hustoty pravděpodobnosti:
- Platí $f(x) \geq 0$ pro $\forall x \in R$,

  $\int_{-\infty}^{\infty} f(x)dx = 1$,

  $P(a<X<b) = \int_{a}^{b} f(x)dx$.



  - Důsledek: P(X=k) = 0.

Systém náhodných veličin (vícerozměrná náhodná veličina) a jeho rozdělení pravděpodobnosti

- Dvourozměrná náhodná veličina – uspořádaná dvojice (X, Y).
- Sdružené rozdělení pravděpodobností:
  - Distribuční funkce $F(x,y) = P(X \leq x, Y \leq y)$.

    - Pravděpodobnostní funkce $p(x_i,y_j) = P(X=x_i, Y=y_j)$.
    - Hustota pravděpodobnosti $f(x,y) = F''_{xy}(x,y)$.
- Marginální rozdělení – rozdělení složek, např.
    - $F_1(x) = F(x,\infty)$, $F_2(y) = F(\infty,y)$.
- Podmíněná rozdělení – rozdělení jedné veličiny při podmínce, že druhá nabyla předem zvolené hodnoty.
    - $F(x|y)=F(x,y)/F_2(y)$ a podobně zavádíme podmíněnou pravděpodobnostní funkci a podmíněnou hustotu.
- Veličiny X, Y jsou nezávislé, je-li jejich sdružené rozdělení rovno součinu příslušných marginálních rozdělení.
- Příklad: Viz Hindls a kol. (2007), s. 70-71.

Číselné charakteristiky náhodných veličin

- Podobně jako u rozdělení četností poskytují charakteristiky náhodných veličin koncentrovanější a přehlednější informaci o rozdělení.
- Střední hodnota pro diskrétní a spojitou náhodnou veličinu X se zavádí vztahy

$$E(X) = \begin{cases} \sum_i x_i \, p_i \,, \\ \int_{-\infty}^{\infty} x \; f(x)dx. \end{cases}$$

- Vlastnosti střední hodnoty:

  $E(aX \pm b) = aE(X) \pm b$, $E(k) = k$,

  $E(X \pm Y) = E(X) \pm E(Y)$,

  $E(XY) = E(X).E(Y)$, když X,Y jsou nezávislé náhodné veličiny (tj. veličiny X,Y jsou výsledkem nezávislých náhodných pokusů).
- Rozptyl pro diskrétní a spojitou náhodnou veličinu X se zavádí vztahy

$$D(X) = E[X - E(X)]^2 = \begin{cases} \sum_i [x_i - E(X)]^2 \, p_i \,, \\ \int_{-\infty}^{\infty} [x - E(X)]^2 \; f(x)dx. \end{cases}$$

- Vlastnosti rozptylu:

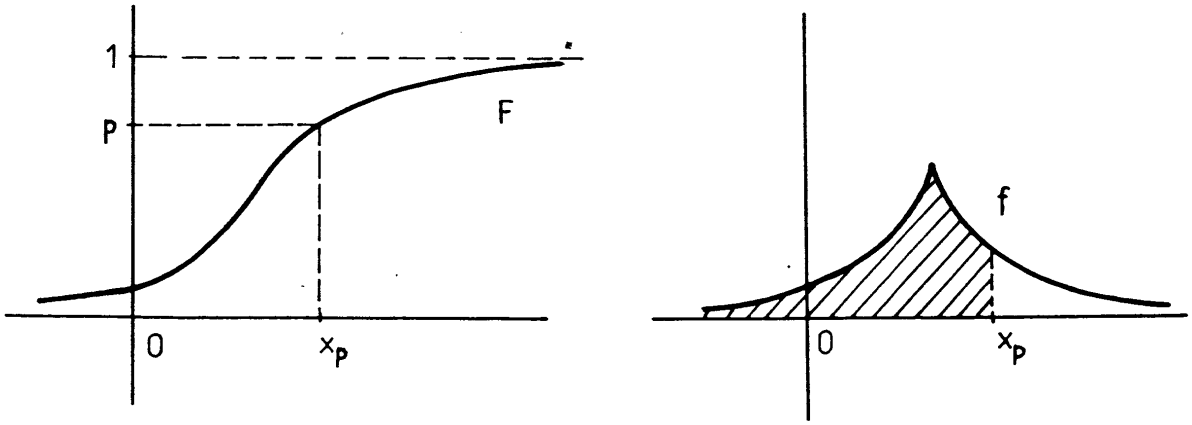
  $D(X) \geq 0$,

  $D(k) = 0$, $D(a \pm bX) = b^2 D(X)$,

  $D(X \pm Y) = D(X) + D(Y)$, jsou-li X, Y nezávislé náhodné veličiny,

  $D(X) = E(X^2) - [E(X)]^2$.
- Směrodatná odchylka $\sigma(X) = \sqrt{D(X)}$.
- p-kvantil spojité náhodné veličiny X je číslo x_p, které při daném p ($0<p<1$) rozděluje množinu hodnot náhodné veličiny X tak, že platí $P(X \leq x_p) : P(X > x_p) = p : (1-p)$
- Pro spojitou náhodnou veličinu platí
$$P(X < x_p) = F(x_p) = \int_{-\infty}^{x_p} f(x)\,dx = p.$$

- Příklady: viz Stuchlý (1999a), str. 57-59.



  - $\mathrm{cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)\,E(Y)$.
  - Kovarianční matice
$$\Sigma = \begin{pmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{pmatrix}.$$
  - Koeficient korelace
$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$
  - Platí $-1 \leq \rho(X, Y) \leq 1$.
  - Jsou-li X, Y nezávislé $\Rightarrow \mathrm{cov}(X, Y) = 0$ a $\rho(X, Y) = 0$.
  - Platí $D(X \pm Y) = D(X) + D(Y) \pm \mathrm{cov}(X, Y)$.
- Marginálních a podmíněných rozdělení zavádíme obvyklým způsobem.
  - Závislost podmíněné střední hodnoty na proměnné v podmínce nazýváme regresní funkcí

- Příklady: Viz Stuchlý (1999a), s. 65-66.

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 51-76. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE, 1999. S. 51-73. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 55-86. ISBN 80-245-0178-3.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 83-121, 129-142. ISBN 80-7261-003-1.

GIBILISCO, S. *Statistika bez předchozích znalostí.* Brno: Computer Press, 2009. S. 57-74, 81-86. ISBN 978-80-251-2465-9.

HEBÁK, P. a J. KAHOUNOVÁ. *Počet pravděpodobnosti v příkladech.* Praha: Informatorium, 1994. ISBN 80-85427-48-6.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 51-81. ISBN 978-80-86446-40-5.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 55-76. ISBN 80-7187-058-7.

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S.131-145, ISBN 80-7169-089-9

WONNACOT, T. H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 73-121. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Volejbalového turnaje se zúčastní 8 družstev. Kolik kombinací medailistů (tj. prvních, druhých a třetích umístnění) celkem existuje?
2) Stroj vyrobil 20 výrobků, z toho 5 vadných. Všechny výrobky jsou v jedné krabici. Určete, kolika způsoby lze vybrat 4 výrobky tak, aby a) ani jeden nebyl vadný, b) byly všechny vadné, c) ve výběru byly 2 vadné a 2 dobré, d) byl nejvýše jeden vadný, e) byl aspoň jeden vadný.
3) Určete, kolika způsoby je možné rozmístnit sedm stejných kuliček do tří krabiček.
4) Stroj vyrobil 20 výrobků, mezi nimi 3 nevyhovující. Jaká je pravděpodobnost, že a) mezi 4 vybranými nebude ani jeden zmetek, b) dva výrobky budou nevyhovující, c) mezi třemi vybranými výrobky bude nejvýše jeden vadný?
5) V populaci jsou 3% lidí s inteligencí na úrovni geniality. 1% lidí z populace je současně geniální a věnuje se vědě. Určete pravděpodobnost, že náhodně zvolený génius je vědec.
6) V osudí je 10 papírku, z toho 5 modrých a 5 červených. Jaká je pravděpodobnost, že a) první vytažený papírek je modrý a druhý červený, b) první vytažený papírek je modrý a druhý rovněž modrý?
7) Na lince vzniká 5% zmetků. U výstupní kontroly je jako dobrý výrobek označeno 5% z nich. Naopak kontrola neprávem označí za zmetek 2% dobrých výrobků. Jaká je pravděpodobnost, že kontrola vyřadí výrobek?
8) Na základě údajů o prodeji v posledních 4 týdnech bylo spočítáno, že počet zákazníků, kteří během jedné hodiny zakoupí novou polévku v plechovce od firmy Poleko (náhodná veličina X), má rozdělené pravděpodobností dané tabulkou

x	0	1	2	3	4	5	6
P(x)	0,15	0,16	0,20	0,18	0,15	0,10	0,06

Vypočítejte a) $P(X \leq 4)$, b) $P(2 \leq X < 6)$, c) $P(X > 2)$, d) $E(X)$, e) $\sigma(X)$.

9) V masokombinátu jsou zásoby čerstvého masa skladovány v chladírnách, maximálně však po dobu 5 dnů. Doba skladování (tedy doba od uložení do expedice) je určena poptávkou a z minulosti je známo, že se jedná o náhodnou veličinu (měřenou ve dnech, označme ji X) s následujícím rozdělením pravděpodobnosti

$$P(x) = \begin{cases} \frac{6-x}{15} & \text{pro x} = 1,2,3,4,5, \\ 0 & \text{jinak.} \end{cases}$$

Napište tabulku a) rozdělení pravděpodobností, b) distribuční funkce, c) vypočítejte $P(X > 2)$.

10) Náhodná veličina X se řídí pravděpodobnostním rozdělením daným hustotou pravděpodobnosti $f(x) = 3x^2$ pro $0 < x < 1$.Určete a) $P(0 < X < 0,5)$, b) $P(X > 0,75)$, c) $P(X = 0,9)$, d) $x_{0,9}$, e) $E(X)$, f) $F(x)$ pro $0 < x < 1$.

? Úkoly k zamyšlení a diskuzi

1) Diskutujte o tom, jakým způsobem budeme počítat pravděpodobnost v situacích, kdy nejsou splněny podmínky pro použití klasické definice pravděpodobnosti.

2) Zamyslete se nad tím, jaké poznatky z počtu pravděpodobnosti může využívat manažer ke zkvalitnění manažerského rozhodování.

Klíč k řešení otázek:

1) Variace: $V_3(8)=8.7.6 = 0,336$.

2) Kombinace: a) $C_4(15)=15.14.13.12/4.3.2.1 = 1365$, b) $C_4(5)=C_1(5) = 5$, c) $C_2(15).C_2(5)= 1050$, d) $C_4(15)+C_3(15).C_1(5)= 3640$, e) $C_4(20)-C_4(15) = 3480$.

3) Kombinace s opakováním: $C'_7(3)=C_7(9)= 36$.

4) Klasická definice pravděpodobnosti: a) $C_4(17)/C_4(20) = 0{,}4912$; b) $C_2(3)/C_2(20) = 0{,}0158$; c) $[C_1(3).C_2(17)+C_3(17)] / C_3(20) = 0{,}9544$.

5) Podmíněná pravděpodobnost: Označme A -vědec, B – genius. $P(A|B) = P(A\cap B)/P(B) = 0{,}01 / 0{,}03 = 0{,}333$.

6) Násobení pravděpodobností: Označme A_1 - 1. je modrý, A_2 - 2. je modrý, B_2 - 2. je červený. Potom a) $P(A_1\cap B_2) = P(A_1).P(B_2|A_1) = 5/10.5/9 = 0{,}2778$; b) $P(A_1\cap A_2) = P(A_1).P(A_2|A_1) = 5/10.4/9 = 0{,}2222$.

7) Úplná pravděpodobnost: Označme V - kontrola vyřadí výrobek, D - výrobek je dobrý, Z - výrobek je zmetek. Potom $P(V)=P(V|D)P(D) + P(V|Z)\,P(Z) = 0{,}02.0{,}95 + 0{,}95.0{,}05 = 0{,}0665$.

8) Pravděpodobnostní rozdělení diskrétní náhodné veličiny: Z tabulky rozdělení pravděpodobností dostáváme a) $P(X \le 4) = 0{,}84$ nebo $P(X \le 4)= 1 - P(X > 4)=1-0{,}10-0{,}06 = 0{,}84$; b) $P(2 \le X < 6) = 0{,}63$; c) $P(X > 2) = 0{,}49$; d) $E(X) = \Sigma xp(x) = 2{,}56$; e) $D(X) = \Sigma x^2p(x) - [E(X)]^2 = 3{,}0864$; $\sigma(X) = 1{,}7568$.

9) a) Pravděpodobnostní funkce:

x	1	2	3	4	5	Součet
p(x)	0,333	0,267	0,2	0,133	0,067	1

b) Distribuční funkce:

x	$(-\infty.1)$	$<1,2)$	$<2,3)$	$<3,4)$	$<4,5)$	$<5,\infty)$
F(x)	0	0,333	0,6	0,8	0,933	1

c) $P(X > 2) = 0{,}2+0{,}13+0{,}07 = 0{,}4$.

10) Pravděpodobnostní rozdělení spojité náhodné veličiny: a) $P(0 < X < 0{,}5 = \int_0^{0,5} 3x^2\,dx = \left[\frac{3x^3}{3}\right]_0^{0,5} = 0{,}125;$

c) $P(X > 0{,}75) = \int_{0,75}^{1} 3x^2\,dx = \left[\frac{3x^3}{3}\right]_{0,75}^{1} = 1 - 0{,}75^3 = 0{,}5781;$

d) $P(X = 0{,}9) = 0$;

e) $0{,}9 = \int_0^{x_{0,9}} 3x^2 dx = x_{0,9}^3 \Rightarrow x_{0,9} = \sqrt[3]{0{,}9} = 0{,}9655;$

f) $E(X) = \int_0^1 x.3x^2 dx = \left[\frac{3}{4}x^4\right]_0^1 = 0{,}75;$

g) $F(x) = \int_0^x 3t^2 dt = \left[\frac{3t^3}{3}\right]_0^x = x^3.$

Kapitola 4: Základní pravděpodobnostní modely

Klíčové pojmy:

rozdělení diskrétní a spojitá, alternativní, binomické, hypergeometrické, normální, normované (standardizované) normální, chi-kvadrát, Studentovo t, Fisher-Snedeckerovo F, Poissonovo, binomické, záporně binomické, rovnoměrné, logaritmicko-normální, exponenciální, multinomické, vícerozměrné normální, centrální limitní věty

Cíle kapitoly:

- **umět aplikovat binomické rozdělení;**
- **umět aplikovat normální rozdělení;**
- **porozumění základním centrálním limitním větám;**
- **získat základní poznatky o rozděleních chi-kvadrát, t a F.**

Čas potřebný ke studiu kapitoly: 9 hodin

Výklad:

Nastínění obsahu kapitoly.

Alternativní rozdělení

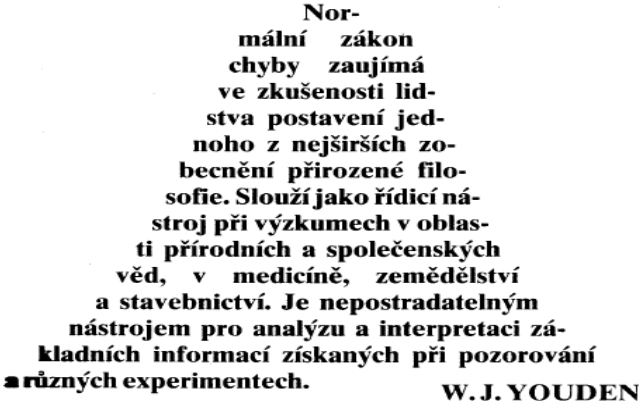
Binomické rozdělení

Normální rozdělení

Centrální limitní věty

Další rozdělení diskrétní a spojité náhodné veličiny

Struktura výkladu



- U často vyskytujících se rozdělení:
  - funkce rozdělení jsou tabelovány a graficky znázorněny
    - v statistických tabulkách (některé i v učebnicích statistiky),
    - v sofwareových prostředcích (např. Excel, R);
  - to usnadňuje provádění výpočtů;
    - viz statistické funkce v Excelu,
    - viz nabídka Distributions v R Commanderu.

Alternativní rozdělení A(π)

- Popis dichotomní populace.
  - Rozdělení nula-jedničkové veličiny – kvantifikuje výsledek náhodného pokusu. X má rozdělení A(π) - stručně zapisujeme: $X \sim A(\pi)$.
  - Nastane-li sledovaný jev A, je X = 1 a nenastane-li jev A, je X = 0 (hod mincí).
  - Rozdělení: X = 1 s pravděpodobností π a X = 0 s pravděpodobností 1 - π, kde $0<\pi< 1$ je parametr rozdělení, tj. pravděpodobností funkce je

$$p(x) = \pi^x (1-\pi)^{1-x},\ x = 0, 1.$$

- Platí $E(X) = \pi$, $D(X) = \pi(1-\pi)$.
    - Důkaz: $E(X)=1\pi + 0\ (1-\pi) = \pi$, $E(X^2)= \pi$,

      $D(X)= E(X^2)-[E(X)]^2= \pi- \pi^2 = \pi(1-\pi)$.

Binomické rozdělení Bi(n;π)

- n,π jsou parametry rozdělení;
- Je jedním z nevíce používaných rozdělení.

- Odvozeno z procesu známého jako Bernoulliho pokus.
    - Švédský matematik James Bernoulii (1654-1705).
    - Příklady situací vedoucích k binomickému rozdělení:
        - Házíme n krát mincí. Výsledek hodu je „pana“ nebo „orel“ - Bi(1;0,5).
        - Výzkumná laboratoř vyvíjí nový lék proti vysokému tlaku – má obavy z určitých škodlivých vedlejších účinků. Ověření na vzorku 80 pacientů - u 12 vedlejší účinky, u 68 ne - Bi(80; 12/80).
        - Firma vyrábí fotoaparáty vybavené elektronickým zařízením pro automatické nastavení správné rychlosti závěrky. Pro kontrolu spolehlivosti této elektroniky firma ověřuje její funkci na náhodně vybraných 20 fotoaparátech z výrobní linky. Z testovaných 20 přístrojů jeden nefunguje správně - Bi(20;1/20).

    - Jde o experimenty, u nichž jsou vždy možné dva výsledky U (úspěch) a N (neúspěch). S nimi jsou spojeny pravděpodobnosti $\pi = P(U)$, $1-\pi = P(N)$.

    - Příklady Bernoulliho pokusů:
        - házení mincí – „pana“ – „orel“: $\pi = 1-\pi = 1/2$.
        - vývoj nového léku - vedlejší účinky léku proti vysokému tlaku: $\pi = 12/80$, $1-\pi = 68/80$.
        - Zkouška fotoaparátu - vadná závěrka: $\pi = 1/20$ (= 0,05), $1-\pi = 19/20$ (= 0,95).

    - Nutné podmínky pro binomické rozdělení:

- Experiment sestává z *n* Bernoulliho pokusů (pokusů, které mají jen dva možné výsledky).
- Pravděpodobnost úspěchu π je stejná pro všechny pokusy.
- Pokusy jsou vzájemně nezávislé (výběr s vracením = nahrazováním vybraných).

- Pravděpodobnostní funkce, tj. pravděpodobnost, že v sérii n nezávislých opakovaných pokusů se úspěch U (= jev A) objeví právě x-krát je

$$p(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} = p_x,$$

kde x = 0,1,2,...,n, $0<\pi<1$ (Bernoulliho vzorec).

  - distribuční funkce vznikne nasčítáváním p(x),
  - výpočet a grafy většiny rozdělení: Excel a R Commander.
- Výpočtem dostaneme číselné charakteristiky:

$E(X) = n\,\pi$,

$E(X^2) = n^2\,\pi^2 + n\,\pi(1-\pi)$,

$D(X) = E(X2) - [E(X)]^2 = n\,\pi(1-\pi)$,

$\sigma(X) = \sqrt{[n\,\pi(1-\pi)]}$.

Příklad: – Viz Stuchlý (1999), s. 82-83.

Pravděpodobnostní a distribuční funkci a jejich grafy počítáme v Excelu pomocí funkce BINOMDIST nebo v R Commanderu v *Distributions-Discrete distribution-Binomial distribution* (zde získáme i kvantilovou funkci).

Hypergeometrické rozdělení H(N, M, n)

- Používáme ho při výběru bez vracení – závislé výběry (např. sportka). Má-li v populaci o rozsahu N sledovaný znak M jednotek, potom pravděpodobnost, že ve výběru n jednotek bez nahrazování bude se nacházet právě k jednotek se sledovaným znakem (a zbývajících n-k jednotek bez sledovaného znaku), je

$$p_k = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \text{ kde } k = \max(0,\, n - N + M),\ldots,\min(M, n).$$

- Platí $E(X) = n\frac{M}{N},\ D(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}.$

- Pro velká N, n a pro n podstatně menší než N (n/N < 0,05) lze hypergeometrické rozdělení H(N, M, n) aproximovat binomickým rozdělením Bi(n, M/N).
- Platí pak E(X) = nπ, D(X) = nπ(1-π)(N-n)/(N-1), kde π = M/N**.**

Příklad: – viz Stuchlý (1999), s. 84.

Distribuční funkci počítá Excel pomocí statistické funkce HYPGEOMDIS (zadáváme pak parametry v pořadí k, n, M, N). R Commander počítá všechny funkce a jejich grafy v *Distributions-Discrete distribution-Hipergeometric distribution* (dále zadáváme parametry v pořadí k, M, N-M, n).

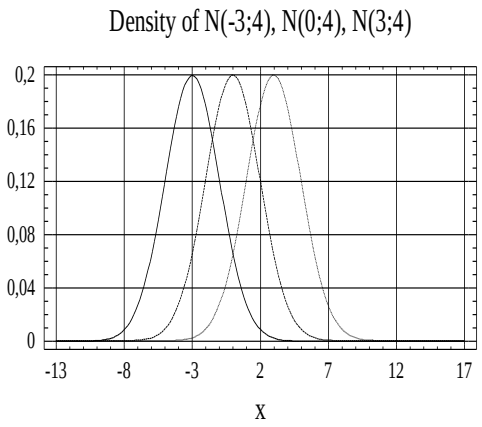
Normální (Gaussovo) rozdělení N(μ;σ²)

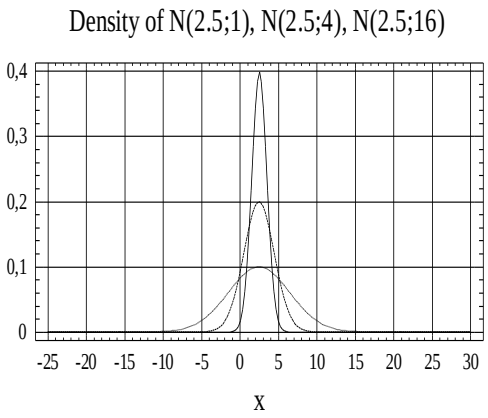
Je nejdůležitější a nejčastější používané rozdělení spojité náhodné veličiny.

- Podle centrální limitní věty k němu za určitých podmínek konvergují jiná rozdělení.
- Představuje pravděpodobnostní model chování velkého množství jevů v technice, přírodních vědách i ekonomii.
- Používá se tam, kde kolísání náhodné veličiny je způsobeno součtem velkého počtu nepatrných vzájemně nezávislých vlivů.
    - Např. v teorii chyb.
    - Bylo zavedeno v roce 1733 Abrahamem de Moivre (1667-54).
    - Je spojeno i se jmény Laplace a Gauss.
- Hustota pravděpodobnosti - grafem zvonovitá funkce (Gaussova křivka).
    - Maximum (medián) je v bodě x = μ a inflexní body v x = μ±σ.

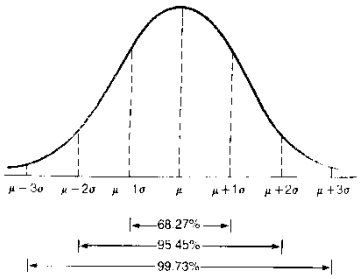
    $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ -\infty < x < \infty$

- Grafy (pro různé hodnoty parametrů):





- Platí: $E(X) = \mu$, $D(X) = \sigma^2$
  - $P(\mu - \sigma < X < \mu + \sigma) = 0{,}6827$ (pravidlo jednoho sigma);
  - $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{,}9545$ (pravidlo dvou sigma: 95,5% populace leží v tomto intervalu);
  - $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{,}9973$ (pravidlo tří sigma).

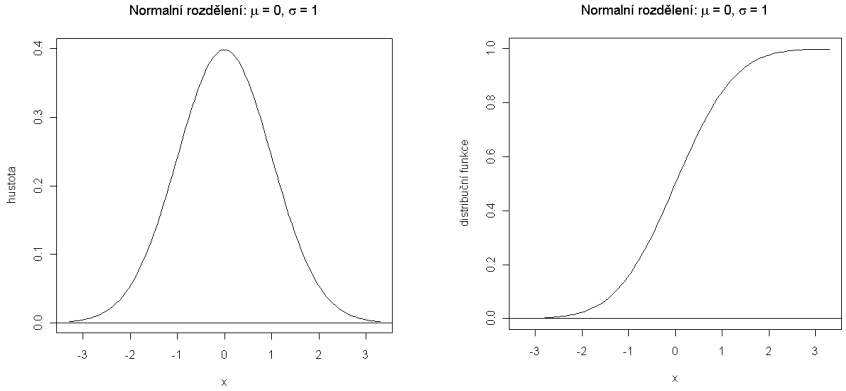


Normované (standardní) normální rozdělení N(0;1):

- Je-li $X \sim N(\mu,\sigma^2) \Rightarrow$ Standardizovaná veličina je $U = \frac{X - \mu}{\sigma} \sim N(0;1)$.

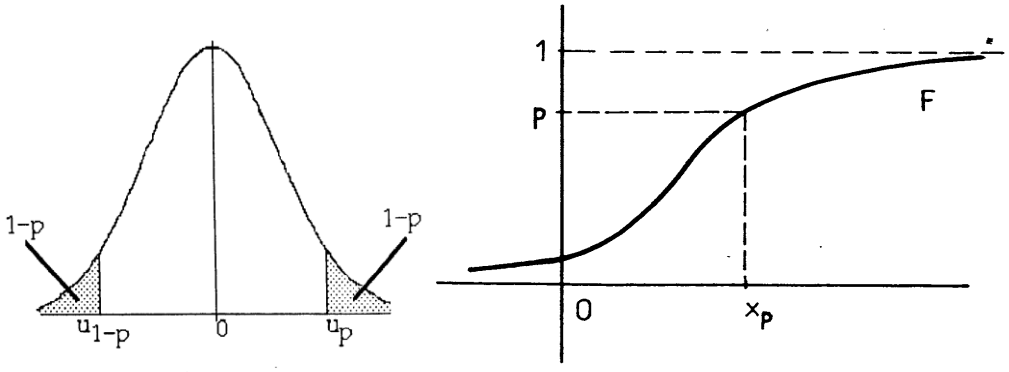
- Platí: $E(U) = 0$, $D(U) = 1$.
- Hustota pravděpodobnosti je $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.

- Distribuční funkce (Laplaceova funkce) je $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt$.
  - Je tabelována pro $u \geq 0$ (viz tab. I. v dodatku).
  - Pro $u < 0$ je $\varphi(u) = \varphi(-u)$, $\Phi(u) = 1 - \Phi(-u)$.
- Grafy hustoty a distribuční funkce standardního normálního rozdělení N(0;1):



- Platí $P(a<X<b) = P\left(\frac{a-\mu}{\sigma}<U<\frac{b-\mu}{\sigma}\right)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$.

  - Kvantil (kvantilová funkce) u_p je důležitým nástrojem ve statistice;
  - Jsou tabelované v tab. II. dodatku (počítá je Excel i R) a definované vztahem $P(U< u_p) = \Phi(u_p) = p$, pro $0<p<1$.



  -
  - Kvantilová funkce je inverzní k distribuční funkci.
  - Platí $u_p = - u_{1-p}$,

    $x_p = \mu + \sigma u_p$ ($u_{0,975} = 1,96$, $u_{0,025} = -1,96$).

- Distribuční a kvantilovou funkci normovaného normálního rozdělení počítáme v Excelu pomocí funkcí NORMSDIST a NORMSINV (lze je také najít v statistických tabulkách v dodatku). Distribuční a kvantilovou funkci obecného normálního rozdělení počítáme v Excelu pomocí funkcí NORMDIST a NORMINV. Tyto funkce a jejich

grafy lze také najít v R Commanderu (v *Distributions-Continuous distribution-Normal distribution*).

Centrální limitní věty

a) Moivre-Laplaceova věta:

- Vyjadřuje konvergenci binomického rozdělení k rozdělení normálnímu
- Je-li X~Bi(n, π), potom pro n velké platí

$$U = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0;1), \text{ tj } \lim_{n\to\infty} P(U \le u) = \Phi(u).$$

- Důsledek: Pro n velké lze rozdělení Bi(n, π) aproximovat normálním rozdělením N(nπ; nπ(1-π)).
- Aproximace je dobrá, je-li $n > 9/[\pi(1-\pi)] \Leftrightarrow \min\{n\pi; n(1-\pi)\} > 5$.

b) Lindebergova-Lévyho věta:

- Součet $X = \sum X_i$, resp. průměr $\bar{X}$, nezávislých stejně rozdělených náhodných veličin (s konečnými stejnými středními hodnotami $E(X_i)=\mu$ a konečnými stejnými rozptyly $D(X_i)=\sigma^2$) má asymptoticky normální rozdělení N(nμ; nσ2), resp. $N(\mu; \sigma^2/n)$.
- Tj. pro n velké platí

$$U = \frac{X - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sqrt{\sigma^2}}\sqrt{n} \sim N(0;1), \text{ tj } \lim_{n\to\infty} P(U \le u) = \Phi(u)$$

- Odtud je $P(a < U < b) \approx \Phi(b) - \Phi(a)$.

Příklady: – Viz Hindls a kol. (2007), s. 90-100 a Stuchlý (1999a), s. 89.

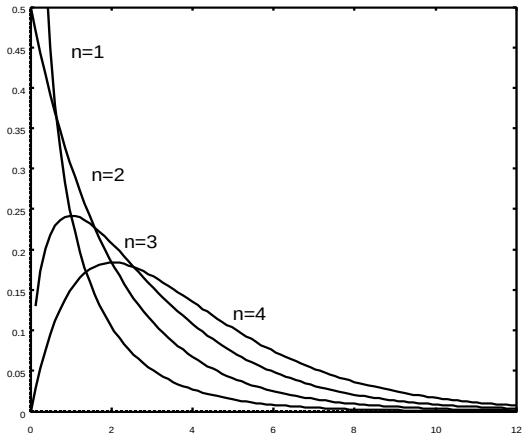
Od normálního rozdělení se odvozují další tři typy rozdělení, která jsou často používána ve statistice.

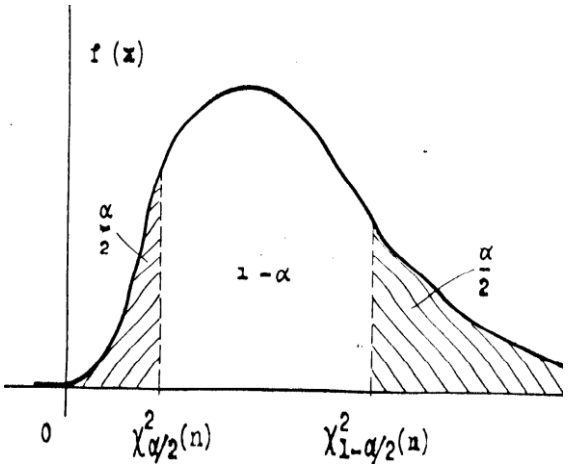
χ^2-rozdělení (chi-kvadrát nebo Pearsonovo rozdělení) $\chi^2(n)$

- Jsou-li $X_1,...,X_n$ nezávislé náhodné veličiny s rozdělením N(0;1) potom $S = X_1^2 + X_2^2 +...+ X_n^2$ má rozdělení $\chi^2(n)$.
  - n nazýváme stupně volnosti.

- Jde o asymetrické rozdělení (hustota – viz obrázek), které se pro velké n (alespoň 30) blíží k rozdělení N(0;1).
- Platí E(S) = n, D(S) = 2n.
- V statistice jsou důležité kvantily chi-kvadrát rozdělení. Označujeme je $\chi_\alpha^2(n)$ a jsou tabelované v tab. III. dodatku pro $n \leq 100$ (Excel i R Commanter je počítá).
- Znázornění kvantilů uvádí další graf.
- Pro n>30 počítáme kvantily pomoci asymptotického vzorce

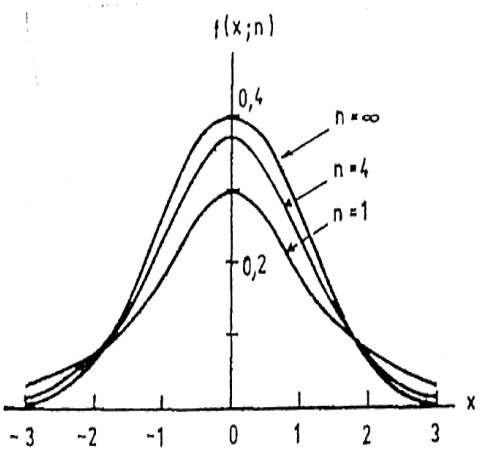
$$\chi_\alpha^2(n) \approx \frac{1}{2}\left(\sqrt{2n-1} + u_\alpha\right)^2$$





Studentovo t-rozdělení t(n)

- Nechť X_1, X_2 jsou nezávislé náhodné veličiny s rozdělením N(0,1) a $\chi^2(n)$. Potom náhodná veličina $T = \frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t(n)$.

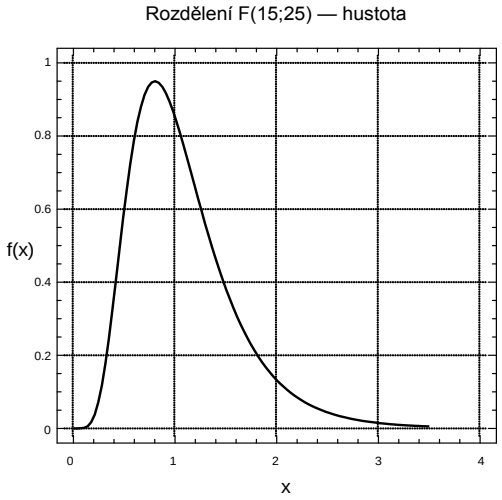


  - n představuje opět stupně volnosti.

- Platí E(T) = 0 (pro n > 1), D(T) = n/(n-2) (pro n > 2).
- Hustota rozdělení t(n) je
    - symetrická,
    - graf podobný rozdělení N(0;1) jen je plošší.
- Pro velká *n* se graf blíží grafu rozdělení N(0;1).
- Důležité jsou kvantily t-rozdělení $t_\alpha(n)$ a $t_{1-\alpha}(n) = -t_\alpha(n)$ jsou tabelované v tab. IV. dodatku (je možno je určit v Excelu i R).
- Pro n velké (n>30) je můžeme přibližně nahradit u_α.

Fisherovo-Snededecorovo F-rozdělení F(*n*,*m*)

- Jsou-li X_1, X_2, nezávislé náhodné veličiny, přičemž $X_1 \sim \chi^2(n)$ a $X_2 \sim \chi^2(m)$, potom náhodná veličina
$$F = \frac{\frac{X_1}{n}}{\frac{X_2}{m}} \sim F(n,m).$$
    - Zde n jsou stupně volnosti čitatele a m stupně volnosti jmenovatele.
- Na obrázku je uveden graf hustoty pravděpodobnosti.
- Kvantily F-rozdělení $F_\alpha(n,m)$ jsou pro obvyklé $\alpha \geq 0{,}5$ tabelované v tab. V. dodatku (najdeme je v Excelu i v R). Jsou definované obdobně jako u rozdělení $\chi^2(n)$.
- Pro $\alpha < 0{,}5$ je $F_\alpha(n,m) = \frac{1}{F_{1-\alpha}(m,n)}$



Z diskrétních rozdělení se ještě často používají Poissonovo rozdělení P(λ) – popisuje počet jevů v prostorové jednotce nebo počet událostí v časové jednotce, geometrické rozdělení Ge(π) a záporné binomické rozdělení ZBi(n,π) popisující počet neúspěchu do 1. úspěchu, resp. do n-tého úspěchu. Viz Cyhelský (2001), s. 157-159.

Ze spojitých rozdělení se ještě často používají rozdělení rovnoměrné rozdělení R(a,b) v simulačních metodách, logaritmicko-normální rozdělení $LN(\mu;\sigma^2)$ v teorii spolehlivosti a účetnictví,

exponenciální rozdělení E(A,δ) v teorii spolehlivosti a v hromadné obsluhy a další rozdělení. Viz Hindls a kol. (2007), s. 89-92 a Stuchlý (1999a), s. 80-81.

Z diskretních vícerozměrných se používá multinomické rozdělení, jako zobecnění rozdělení Bi(n,π) (viz Cyhelský 2001, s. 161-163) a ze spojitých vícerozměrné normální rozdělení (viz Stuchlý 1999, s. 81-82, 85 a Cyhelský .2001, s. 172-175).

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 76-103. ISBN 978-80-86946-43-6.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 82-118. ISBN 978-80-86446-40-5.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE, 1999. S. 74-90. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 87-115. ISBN 80-245-0178-3.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 149-185, 190-193. ISBN 80-7261-003-1.

HEBÁK, P. a J. KAHOUNOVÁ. *Počet pravděpodobnosti v příkladech.* Praha: Informatorium, 1994. ISBN 80-85427-48-6.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 77-100. ISBN 80-7187-058-7.

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada Publishing, 1996. S. 145-172. ISBN 80-7169-089-9.

WONNACOT, T. H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 133-140. ISBN 80-85605-09-0.

? Otázky a úkoly

1) U přijímacích zkoušek na vysokou školu z angličtiny musí student v písemném testu zaškrtnout odpověď u 20 otázek. U každé otázky jsou 4 možnosti a pouze jedna je správná. Aby v testu uspěl, musí student zaškrtnout nejméně polovinu odpovědí správně. Přestože se student poctivě připravoval, test je náročný a on si je jist správnou odpovědí pouze u 8 otázek. U dalších 12 zaškrtává zcela náhodně. Jaká je pravděpodobnost, že a) v testu uspěje, b) zodpoví správně všechny otázky, c) zodpoví správně právě 15 otázek, d) zodpoví správně alespoň 15 otázek, e) v testu neuspěje?

2) Pouze 5 pracovníků určitého ministerstva používá na Internetu komunikační program ICQ. Z 50 pracovníků ministerstva, kteří využívají ke své práci Internet, náhodně vybereme 10 pracovníků. Jaká je pravděpodobnost, že z těchto 10 pracovníků program ICQ používají a) právě 2 pracovníci, b) 0 pracovníků, c) více než 3 pracovníci?

3) Náhodná veličina U má normované normální rozdělení. Určete pravděpodobnosti a kvantily: a) $P(U < 0)$, b) $P(U > 2)$, c) $P(U=0)$, d) $P(-1<U<1)$, e) $u_{0,975}$, f) $u_{0,05}$, h) $u_{0,95}$.

4) Pravděpodobnost, že při daném výrobním procesu bude na určitém stroji vyroben vadný výrobek, je 0,04. Jaká je pravděpodobnost, že z 250 vyrobených výrobku počet vadných výrobků V bude alespoň 5, ale nejvýš rovný 15. Proveďte a) přesný výpočet, b) výpočet pomocí centrální limitní věty.

5) Z chovného rybníka bylo vyloveno 10 kaprů, kteří byli zváženi a puštěny zpět. Byla vypočítána jejich průměrná hmotnost 2,852 kg. Na základě naměřených hodnot byla odhadnuta směrodatná odchylka hmotnosti kapra na 0,6 kg. Přitom do rybníka bylo vysazeno 1500 ks kapřího plůdku a počítá se s 20% úmrtností. Jaká je pravděpodobnost,

že a) náhodně vylovený kapr bude vážit méně než 2 kg, b) výlov celého rybníka přesáhne co do hmotnosti vylovených kaprů 3400 kg?

6) Nechť X a Y jsou náhodné veličiny s χ^2 rozdělením s 10 a 8 stupni volnosti. Která z následujících pravděpodobností je menší P(X < 3) nebo P(Y < 2)?

? Úkoly k zamyšlení a diskuzi

1) Diskutujte o tom, které poznatky o pravděpodobnostních rozděleních lze využít při regulační kontrole jakosti a jak budeme postupovat?
2) Zamyslete se nad tím, jaký je vztah mezi kvantily rozdělení t(n) a N(0;1).
3) Odlehlé hodnoty z dat pocházejících z normální populace lze určovat pomocí normované hodnoty znaku. Zamyslete se nad tím, jaké poznatky použijeme a jak toto určení provedeme.

Klíč k řešení otázek:

1) Binomické rozdělení: a) 0,842, b) 0, c) 0,011, d) 0,014, e) 0,158. Podrobněji Marek a kol. (2007), s. 83-84.
2) Hypergeometrické rozdělení: a) 0,21, b) 0,311, c) 0,004. Podrobněji Marek a kol. (2007), s. 92.
3) Standardní normální rozdělení: a) 0,5; b) 0,023; c) 0; d) 0,683; e) 0,317; f) 1,96; g) -1,645; h) 1,645. Podrobněji Marek a kol. (2007), s. 97-102.
4) Normální rozdělení: a) 0,9278; b) 0,9203. Podrobněji Hindls a kol. (2007), s. 99-100.
5) Označme váhu kapra X, a váhu všech vylovených kaprů $Y = \sum_{i=1}^{n} X_i$. Potom X ~ N(2,852;0,62) a podle centrální limitní věty Y ~ N(nμ;nσ^2), kde n = 0,8.1500 = 1200, tedy Y ~ N(1200.2,852; 1200.0,6^2) = N(3422,4; 20,7846^2). a) P(X<2)= P((X-

2,852)/0,6<(2-2,852)/0,6) = P(U<-1,42) = Φ(-1,42)=1-Φ(1,42) = 1-0,9222 = 0,0778; b) P(Y≥3400) = 0,8594, neboť výstup rychlejšího výpočtu v R Commanderu je:

```
> pnorm(c(3400), mean=3422.4, sd=20.7846, lower.tail=FALSE)
[1] 0.8594209
```

6) Chi-kvadrát rozdělení: X ~ $\chi^2(10)$, Y ~ $\chi^2(8)$, P(X<3)=$F_X(3)$=0,018576 < P(Y<2)= $F_Y(8)$=0,018988 (distribuční funkci určíme v Excelu).

Kapitola 5: Výběrová šetření, rozdělení výběrových charakteristik a základy statistické indukce

Klíčové pojmy:

výběrová šetření, statistická indukce, reprezentativní výběr, prostý náhodný výběr, systematický a kvótní výběr, záměrné výběry, výběrové charakteristiky, výběrová rozdělení charakteristik, standardní chyba průměru, výběrový průměr, poměr a podíl, bodové odhady, nestrannost, výdatnost, konzistence a eficience odhadu, intervalové odhady, koeficient spolehlivosti, intervaly spolehlivosti pro normální výběr, určování rozsahu výběrového souboru, asymptotické intervaly spolehlivosti

Cíle kapitoly:

- **popis metod výběrových šetření;**
- **porozumění principu získávání reprezentativního odhadu;**
- **znalost základních výběrových charakteristik a jejich vlastností;**
- **stanovit a interpretovat bodový a intervalový odhad.**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly.

Základní pojmy z výběrových šetření

Základní výběrové charakteristiky a jejich výběrová rozdělení

Odhady parametrů

Bodový odhad a jeho vlastnosti

Intervalové odhady

Intervaly spolehlivosti pro normální výběr

Asymptotické intervaly spolehlivosti

Určování optimálního rozsahu výběru

Struktura výkladu

Nelze jíst celého vola jenom proto,
abychom poznali, že to jde ztuha.
Samuel Johnson

Základní pojmy z výběrových šetření

Některé pojmy jsou již vysvětleny v úvodní části textu. Statistickým šetřením rozumíme získávání informací (poznatků, dat) o statistických jednotkách. Tyto informace poskytují manažerům podniků a organizací podporu při jejich rozhodování. Takovéto informace mají jednu věc společnou. Jsou získávány obvykle pouze od vybraného souboru, a nikoli od celé populace. Vycházejí pouze z poznání části určitého celku.

Statistická indukce se zabývá zobecňováním úsudků o vlastnostech základního souboru založených na informacích získaných z výběrového souboru.

Získávání údajů o všech jednotkách základního souboru je často velmi pracné, zdlouhavé, nákladné a ve většině případů to není ani možné (destrukční zkoušky). Proto postupujeme tak, že provedeme výběr určitého počtu jednotek ze základního souboru, u nich zjistíme potřebné údaje a poznatky o rozdělení nebo o parametrech výběrového souboru a přenášíme je indukcí na celý základní soubor (např. zjišťování názoru zákazníků na nový výrobek, předvolební průzkumy preferencí jednotlivých kandidátů, statistická kontrola kvality výroby). Výběrový soubor musí byt určen tak, aby reprezentoval celou populaci, tj. musí být reprezentativní (věrná zmenšenina základního souboru). Nejčastěji ho získáme náhodným výběrem (např. losováním, pomocí tabulek náhodních čísel nebo simulací těchto náhodných čísel na počítači).

Techniky šetření:

- Vyčerpávající šetření (census) – nákladné, ne vždy možné (např. sčítání bytů a obyvatel prováděná Českým statistickým úřadem).
- Výběrové šetření – ekonomičtější, výsledky zatíženy výběrovou chybou (lze ji odhadnout).
    - Nereprezentativní výběry – např. anketa, metoda základního masivu, záměrný výběr (viz Hindls a kol. 2007, s. 109-110). Obsahují navíc nevýběrové chyby, které nelze odhadnout.
    - Reprezentativní výběry – založeny na náhodném výběru a použití metod počtu pravděpodobnosti.

Prostý náhodný výběr – nejjednodušší a nejčastěji používaná metoda.

- Každá jednotka základního souboru, ale i každá n-tice měření musí mít stejnou pravděpodobnost, že bude vybrána. Jednotlivé výsledky výběru musí být nezávislé.
- Získáme ho výběrem s vracením. Při velkém rozsahu základního souboru (je alespoň 20 krát větší než výběrový soubor) je rozdíl mezi výběrem s vracením (nezávislé výběry - řídí se binomickým rozdělením) a výběrem bez vracení (závislé výběry – řídí se hypergeometrickým rozdělením) zanedbatelný. Z praktických důvodů používáme obvykle výběr bez vracení (analýzy ale provádíme podle technik odvozených pro výběr s vracením).

- Technika pořízení prostého náhodného výběru:

- Vytvoříme nejdříve tzv. oporu výběru, tj úplný seznam jednotek základního souboru a provedeme jejich očíslování.

- Náhodný výběr zabezpečí reprezentativnost výběru a tím i dobrou kvalitu získaných výsledků statistické indukce.

- Je možno používat i složitější upořádání výběru (oblastní, skupinový, vícestupňový – viz Hindls a kol. 2007, s. 113-115).

- Při obtížném pořízení opory můžeme použit systematický výběr.

    - Spočívá ve výběru každé j-té jednotky, počínaje od prvé, která byla vybrána náhodně;
        - např. každé páté, má-li se vybrat 20% populace;
        - podmínka pro použití této techniky: jednotky z populace tvoří náhodnou posloupnost nezávislou na sledovaném znaku;
        - pozor na periodicity – např. při výběru novin.

- V marketinkových a sociologických výběrech se používá i kvótní výběr (např. respondenty vybíráme podle kvót stanovených na pohlaví, věk, vzdělání).

Metody statistické indukce se využívají např. v marketingovém výzkumu trhu.

- Testování nových výrobků u určité skupiny zákazníků před jejich zavedením na trh.
- Prodejní organizace mají zájem na tom, aby věděli, jak zákazníci vnímají a oceňují jejich výrobky.
- Organizace poskytující veřejné služby se stále více zajímají i o to, jak jejich aktivity hodnotí občané.

Metody statistické indukce zahrnují:

- bodové a intervalové odhady
- statistické testy (parametrické, neparametrické).

Základní výběrové charakteristiky a jejich výběrová rozdělení

Výsledkem náhodného výběru o rozsahu n jsou hodnoty $x_1,x_2,...,x_n$ nějakého statistického znaku, které můžeme považovat za realizace n nezávislých stejně rozdělených náhodných veličin $X_1,X_2,...,X_n$. Toto rozdělení budeme nazývat rozdělením základního souboru. Z uvedených hodnot náhodného výběru počítáme různé výběrové charakteristiky neboli statistiky, které jako funkce náhodného výběru jsou též náhodnými veličinami s určitým rozdělením pravděpodobností závislým na rozdělení základního souboru. Jako náhodné veličiny je budeme značit velkým písmenem (např. $\overline{X}$, S_x^2) a jejich realizace odpovídajícím malým písmenem $\overline{x}$, s_x^2). Dále si uvedeme přehled nejpoužívanějších výběrových charakteristik a jejich vlastností.

A) Výběrový průměr (*Sample Mean*)

Označme $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ výběrový průměr z náhodného výběru n pozorování vybraných ze základního souboru se střední hodnotou μ a rozptylem σ^2. Potom platí:

a) Výběrové rozdělení statistiky $\overline{X}$ má střední hodnotu

$E(\overline{X}) = \mu$.

b) Výběrové rozdělení statistiky $\overline{X}$ má rozptyl

$$D(\overline{X}) = \frac{\sigma^2}{n}$$

a standardní odchylku

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Odhad standardní odchylky (σ nahradíme s) nazýváme standardní chybou průměru.

c) Není-li rozsah výběrového souboru n podstatně menší než rozsah základního souboru N, potom pro standardní odchylku platí

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

d) Je-li rozdělení základního souboru normální, potom standardní náhodná veličina

$$U = \frac{\overline{X} - \mu}{\sigma(\overline{X})} = \frac{\overline{X} - \mu}{\sigma} \sqrt{n}$$

má standardní normální rozdělení.

e) Není-li rozdělení základního souboru normální ale rozsah výběrového souboru n je velký, potom podle centrální limitní věty má veličina U přibližně standardní normální rozdělení.

B) Výběrový poměr (podíl)

Označme X počet úspěchů v alternativním výběrovém souboru n pozorování, kde pravděpodobnost úspěchu je π. Potom statistiku představující poměr úspěchů

$$P = \frac{X}{n}$$

ve výběrovém souboru nazýváme výběrovým poměrem a statistiku X nazýváme výběrovým úhrnem.

Potom platí:

a) Výběrové rozdělení výběrového poměru P má střední hodnotu

$E(P) = \pi.$

b) Výběrové rozdělení statistiky P má rozptyl

$$D(X) = \frac{\pi(1-\pi)}{n}$$

a standardní odchylku

$$\sigma(P)=\sqrt{\frac{\pi(1-\pi)}{n}}.$$

Veličinu σ(P) nazýváme standardní odchylkou výběrového poměru P.

c) Není-li rozsah výběrového souboru n podstatně menší než rozsah základního souboru N, je

$$\sigma(P)=\sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}.$$

d) Je-li rozsah výběrového souboru n velký, má náhodná veličina

$$U=\frac{P-\pi}{\sigma(P)}=\frac{P-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$$

přibližně standardní normální rozdělení.

Pro výběrový úhrn X platí

$E(X) = n\pi$, $D(X) = n\pi(1-\pi)$.

C. <u>Výběrový rozptyl</u> (*Sample Variance*)

Označme $S_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ výběrový rozptyl z náhodného výběru n pozorování vybraných ze základního souboru se střední hodnotou μ a rozptylem σ^2. Potom platí:

a) Výběrový rozptyl S_x^2 má střední hodnotu

$E(S_x^2) = \sigma^2$.

b) Rozptyl výběrového rozptylu závisí na rozdělení základního souboru. Je-li toto rozdělení normální $N(\mu;\sigma^2)$, potom

$$D(S_x^2) = \frac{2\sigma^4}{n-1}.$$

c) Je-li rozdělení základního souboru $N(\mu;\sigma^2)$, potom náhodná veličina $Y = \frac{(n-1)S_x^2}{\sigma^2}$ má rozdělení $\chi^2(n-1)$ a náhodná veličina $T = \frac{\overline{X} - \mu}{S_x}\sqrt{n}$ má Studentovo t-rozdělení t(n-1).

d) Pro dva nezávislé náhodné výběry $X_1,...,X_n$, $Y_1,...,Y_m$, vybrané ze základních souborů o rozsazích n a m s rozděleními $N(\mu_1;\sigma_1^2)$, $N(\mu_2;\sigma_2^2)$, mají náhodné veličiny

$$F = \frac{\frac{S_x^2}{\sigma_1^2}}{\frac{S_y^2}{\sigma_2^2}} \quad \text{a} \quad \overline{X} - \overline{Y}$$

Fisherovo F-rozdělení F(n-1,m-1) a normální rozdělení $N(\mu_1 - \mu_2; \sigma_1^2/n + \sigma_2^2/m)$.

Předpoklady o normalitě rozdělení základního souboru jsou v tvrzeních b)-d) podstatné. Ani při velkém rozsahu výběrového souboru je nelze vynechat.

Příklady: Viz Stuchlý (1999a), s. 93-94.

Odhady parametrů

Pro rozhodování manažera je důležité získávat informace a využít je na odhady parametrů. Např.: Vládu zajímá odhad množství zboží ze zahraničního obchodu, nebo odhad preferencí (úhrn a poměr); obchodníky zajímá odhad úrovně trhu s akciemi; spotřebitele zajímají průměrné ceny určitého zboží apod.

Jednou ze základních úloh statistické indukce je odhad neznámých parametrů základního souboru pomocí náhodného výběru. Existují dva způsoby odhadu:

Bodový odhad - neznámý populační parametr (populační charakteristiku) odhadujeme jedním číslem vypočítaným z hodnot výběrového souboru.

Intervalový odhad - najdeme interval, v kterém daný parametr s velkou pravděpodobností leží.

Bodový odhad a jeho vlastnosti

Bodovým odhadem odhadujeme neznámý parametr základního souboru pomocí jedné hodnoty neboli bodu. Je potřebné rozlišovat mezi dvěma významy pojmu bodový odhad: odhadem jako funkcí náhodného výběru, tj. náhodnou odhadovou funkcí (*Estimator*) a jeho realizací, která udává číselnou hodnotu této náhodné veličiny (*Estimate*).

Předpokládejme, že je daný náhodný výběr $X_1,...,X_n$ ze základního souboru popsaného určitým rozdělením pravděpodobností. Neznámý parametr základního souboru θ odhadujeme vhodnou funkcí náhodného výběru $T(X_1,...,X_n)$. Zapisujeme

$$\hat{\theta} = T(X_1,...,X_n),$$

a výběrovou charakteristiku $\hat{\theta}$ nazýváme bodovým odhadem parametru θ. Např. odhad populačního průměru a populační směrodatné odchylky je

$$\hat{\mu} = \overline{X} = \frac{X_1 + \cdots + X_n}{n}, \quad \hat{\sigma} = S = \sqrt{\frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}}.$$

Aby byl použitý odhad dobrý, musí mít určité vlastnosti. Mezi důležité vlastnosti kvalitních statistických odhadů zařazujeme nestrannost, vydatnost, konzistentnost a postačitelnost.

a) Nestranný odhad:

Nestrannými neboli nezkreslenými odhady (*Unbiased Estimator*) parametru θ jsou ty, jejichž střední hodnota se rovná tomuto parametru, tj.

$$E(\hat{\theta}) = \theta.$$

Tato vlastnost zaručuje, že nedochází k systematickému podhodnocování nebo nadhodnocování skutečné hodnoty parametru. Protože např. platí

$$E(\overline{X}) = \mu,\ E(S_x^2) = \sigma^2,\ E(P) = \pi,$$

jsou výběrový průměr, výběrový rozptyl a výběrový poměr nestrannými odhady svých populačních protějšků. Odhad, který nesplňuje podmínku nestrannosti, nazýváme vychýlený (*Biased*). Funkci

$$b(\hat{\theta},\theta) = E(\hat{\theta}) - \theta$$

nazýváme vychýlením či zkreslením (*Bias*) odhadu $\hat{\theta}$. Odhady splňující podmínku

$$\lim_{n\to\infty} E(\hat{\theta})=\theta,$$

nazýváme asymptoticky nestranné. Např. pro rozptyl

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

platí $\lim_{n\to\infty} E(S^2)=\lim_{n\to\infty}\frac{n-1}{n}\sigma^2=\sigma^2$ a proto je S^2 asymptoticky nestranným odhadem σ^2.

b) Vydatný odhad:

Druhou požadovanou vlastností odhadu je, aby se rozdělení výběrové statistiky $\hat{\theta}$ s největší pravděpodobností koncentrovalo blízko odhadovaného parametru θ. To je zaručeno, když požadujeme, aby rozptyl odhadu $D(\hat{\theta})$ byl minimální. Odhad, který splňuje oba dva uvedené požadavky, nazýváme vydatný neboli optimální. Takové odhady nemusí vždy existovat nebo je lze v některých případech získat jen obtížně. Lze ukázat, že statistiky $\overline{X}$, P jsou v případě normality základního souboru vydatnými odhady svých populačních protějšků.

c) Konzistentní odhad:

Nestrannost odhadu zabezpečuje jen, aby jeho střední hodnota se rovnala odhadovanému parametru, nedává však odpověď na otázku, jak se odhad přibližuje k hodnotě tohoto parametru. Odhad, který se v pravděpodobnosti s rostoucím rozsahem výběru n blíží k hodnotě odhadovaného parametru, nazýváme konzistentní. Matematicky lze konzistenci odhadu $\hat{\theta}$ zapsat vztahem

$$\lim_{n\to\infty} P(|\hat{\theta}-\theta|<\varepsilon)=1,$$

tj. pro velké n konverguje odhad $\hat{\theta}$ v pravděpodobnosti k parametru θ. Postačující podmínkou pro konzistenci nestranného odhadu je

$$\lim_{n\to\infty} D(\hat{\theta})=0 .$$

d) Postačující odhad:

Odhad je postačující (*Sufficient*), když v sobě zahrnuje veškerou informaci o odhadovaném parametru, kterou poskytuje náhodný výběr. Znamená to, že žádný jiný odhad nemůže o odhadovaném parametru dodat více informace.

Výběrové statistiky: výběrový průměr, výběrový úhrn, výběrový podíl a výběrový rozptyl splňují uvedené vlastnosti a proto je můžeme považovat za nejlepší odhady odpovídajících parametrů základního souboru.

Intervalové odhady

Intervalový odhad spočívá v nalezení intervalu spolehlivosti neboli konfidenčního intervalu (T_d,T_h), který pokrývá neznámý parametr θ s pravděpodobností 1-α. Tato pravděpodobnost 1-α se nazývá spolehlivostí neboli koeficientem či úrovní spolehlivosti (*Level of Confidence*) intervalového odhadu. Pokud výběr mnohokrát opakujeme, potom právě ve 100(1-α)% případů bude parametr θ obsažen ve vypočteném intervalu spolehlivosti. Uvedený interval nazýváme 100(1-α)%-ním intervalem spolehlivosti pro parametr θ. Zapisujeme

$$P(T_d < \theta < T_h) = 1-\alpha.$$

Číslo α volíme obvykle malé (nejčastěji α = 0,05 nebo 0,01). Pokud jsou obě meze intervalu spolehlivosti konečné, nazýváme tento interval dvojstranný. Je-li jedna z těchto mezí nevlastní (nekonečno), hovoříme pak o jednostranném intervalu spolehlivosti. Speciálně interval spolehlivosti určený vztahem $P(T_d < \theta) = 1-\alpha$, nazýváme levostranný interva spolehlivosti a interval určený vztahem $P(\theta < T_h) = 1-\alpha$, nazýváme pravostranný interval spolehlivosti. Meze intervalu spolehlivosti závisí na odhadovaném parametru, použitém náhodném výběru a zejména na jeho výběrovém rozdělení. V dalším si naznačíme postup, jak odvodit vzorce pro dolní mez T_d a

horní mez T_h intervalů spolehlivosti pro nejčastěji používané parametry v případě malých výběrů pocházejících z normálně rozdělených základních souborů a v případě velkých výběrů.

Intervaly spolehlivosti pro normální výběr

Předpokládejme nejdříve, že náhodný výběr $X_1,...,X_n$ pochází z normálního rozdělení $N(\mu,\sigma^2)$, kde μ je odhadovaný parametr střední hodnoty a rozptyl σ^2 je známý. Potom statistika

$$U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

má rozdělení N(0;1). Proto platí

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < u_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

kde $u_{1-\frac{\alpha}{2}}$, $-u_{1-\frac{\alpha}{2}} = u_{\frac{\alpha}{2}}$ jsou kvantily rozdělení N(0;1). Ekvivalentními úpravami této nerovnosti dostaneme

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

což představuje vzorec pro 100(1-α)%-ní interval spolehlivosti pro populační průměr μ (u-interval). Tedy meze tohoto intervalu jsou $T_d = \overline{X} - u_{1-\frac{\alpha}{2}}\sigma(\overline{X})$, $T_h = \overline{X} + u_{1-\frac{\alpha}{2}}\sigma(\overline{X})$, kde $\sigma(\overline{X})$ je standardní chyba výběrového průměru. Interval můžeme psát ve tvaru ($\overline{X}$ -d, $\overline{X}$ +d), kde $d = u_{1-\frac{\alpha}{2}}\sigma(\overline{X})$ nebo ve tvaru $\mu = \overline{X} \pm$ d.

Ve většině reálných situací je parametr rozptylu σ^2 neznámý. Potom ho musíme nahradit odhadem S_x^2 a místo statistiky U pak dostaneme statistiku

$$T = \frac{\overline{X} - \mu}{S_x}\sqrt{n},$$

která má za uvedených předpokladů rozdělení t(n-1). Stejným způsobem jako dříve dostáváme 100(1-α)%-ní interval spolehlivosti pro populační průměr μ (t-interval) ve tvaru

$$P\left(\overline{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{S_x}{\sqrt{n}} < \mu < \overline{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{S_x}{\sqrt{n}} \right) = 1-\alpha ,$$

kde $t_{1-\frac{\alpha}{2}}(n-1), -t_{1-\frac{\alpha}{2}}(n-1) = t_{\frac{\alpha}{2}}(n-1)$ jsou kvantily rozdělení t(n-1).

Podobně dostaneme 100(1-α)%-ní pravostranný interval pro populační poměr μ ve tvaru

$$P\left(\mu < \overline{X} + t_{1-\alpha}(n-1)\frac{S_x}{\sqrt{n}} \right) = 1-\alpha$$

a 100(1-α)%-ní levostranný interval pro parametr μ

$$P\left(\overline{X} - t_{1-\alpha}(n-1)\frac{S_x}{\sqrt{n}} < \mu \right) = 1-\alpha .$$

K odvození intervalu spolehlivosti pro parametr populačního rozptylu σ^2 použijeme statistiku

$$\frac{(n-1)S_x^2}{\sigma^2},$$

která má za předpokladu normality rozdělení χ^2(n-1). Proto platí

$$P\left(\chi^2_{\frac{\alpha}{2}}(n-1) < \frac{(n-1)S_x^2}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2}}(n-1) \right) = 1-\alpha ,$$

kde $\chi^2_{\frac{\alpha}{2}}(n-1)$, $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ jsou kvantily rozdělení χ^2(n-1). Úpravou nerovností pomocí ekvivalentních úprav odtud dostaneme 100(1-α)%-ní interval spolehlivosti pro populační rozptyl σ^2 ve tvaru

$$P\left(\frac{(n-1)S_x^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S_x^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right) = 1-\alpha\,.$$

Odmocněním uvedených nerovností dostaneme odtud interval spolehlivosti pro standardní odchylku σ. Intervaly spolehlivosti pro parametr střední hodnoty jsou symetrické se středem v bodě $\overline{X}$ a jejich délka 2d s rozsahem souboru n klesá a se zvyšováním hladiny významnosti roste. Hodnotu d lze interpretovat jako statistickou chybu průměru. Počítá ji Excel a nazývá jí „konfidence“ (ve statistických funkcích pro u-interval a v Analýze dat pro t-interval). Interval spolehlivosti pro rozptyl σ^2 symetrický není.

Určování rozsahu souboru.

Jestliže polovina délky intervalu spolehlivosti pro parametr μ nesmí překročit hodnotu Δ, musí být v případě známého rozptylu splněna podmínka

$$u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \Delta\,.$$

Řešením této nerovnosti dostaneme k tomu požadovaný rozsah souboru

$$n \ge u^2_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\Delta^2}\,.$$

Veličinu Δ nazýváme přípustná chyba. V případě neznámého rozptylu nahradíme σ^2 odhadem S_x^2.

Asymptotické intervaly spolehlivosti

Mějme náhodný výběr $X_1,...,X_n$ z libovolného rozdělení s neznámými parametry střední hodnoty μ a rozptylu σ^2. Nechť rozsah souboru n je velký (n > 30). Potom k odvození intervalu spolehlivosti pro parametr μ můžeme použít statistiky

$$U = \frac{\overline{X}-\mu}{S_x}\sqrt{n}\,,$$

která má podle centrální limitní věty rozdělení N(0;1). Odtud dostaneme asymptotický 100(1-α)%-ní interval spolehlivosti pro populační průměr μ (u-interval) ve tvaru

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}} \right) = 1-\alpha \,.$$

Potřebný rozsah souboru n (pro danou přípustnou chybu Δ) určíme podle vzorce

$$n \geq u^2_{1-\frac{\alpha}{2}} \frac{S_x^2}{\Delta^2} \,.$$

Podobně můžeme s pomocí výběrového poměru P odvodit 100(1-α)%-ní interval spolehlivosti pro populační poměr π ve tvaru

$$P\left(P - u_{1-\frac{\alpha}{2}} \frac{\sqrt{P(1-P)}}{\sqrt{n}} < \pi < P + u_{1-\frac{\alpha}{2}} \frac{\sqrt{P(1-P)}}{\sqrt{n}} \right) = 1-\alpha$$

Požadovanýý rozsah souboru n určíme podle vzorce

$$n \geq u^2_{1-\frac{\alpha}{2}} \frac{P(1-P)}{\Delta^2} \,.$$

Zde P je výběrový poměr, který získáme předvýběrem (popř. hodnotou 0,5).

V případě, že rozsah základního souboru N není podstatně větší než rozsah výběrového souboru n, musíme vzorce pro parametry μ a π opravit tak, že standardní chybu výběrového průměru σ($\overline{X}$) nebo výběrového poměru σ(P) násobíme opravným faktorem $\sqrt{\frac{N-n}{N-1}}$.

V systému R se intervaly spolehlivosti pro průměr a rozptyl dostaneme interaktivně současně s prováděním parametrických testu (viz následující kapitola). Přesné (pro normální výběr) i asymptotické intervaly dostaneme po aktivování balíku *vsePackage* (Komárek 2012) příkazy *estim.mean(x, type="two.sided" [„less", „greater"], conf.level=), estim.var(x, type=" „,*

conf.level=). Pro poměr počítáme interval spolehlivosti ručně podle uvedených vzorců. A obdobně počítáme i optimální rozsah výběru. Ověřování normality dat probereme v následující kapitole.

Příklady: – Viz Hindls a kol. (2007), s. 131-132 a Stuchlý (1999), s. 101-105.

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 107-133. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE, 1999. S. 91-98, 100-105, 107-109. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 117-139. ISBN 80-245-0178-3.

BÍNA V. a kol. *Jak na jazyk R*. J. Hradec: FM VŠE, 2006.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 197-214, 227-235, 237-238. ISBN 80-7261-003-1.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada Publishing, 1999. S. 57-68. ISBN 80-7169-255-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 121-131, 166-167. ISBN 978-80-86446-40-5.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 103-127. ISBN 80-7187-058-7.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011 (v elektronické podobě – viz https://is.vstecb.cz/auth/www/6384/).

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada Publishing, 1996. S. 195-208. ISBN 80-7169-089-9.

WONNACOT, T. H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 199-283. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Hypermarket Hyper chce pro zkvalitnění služeb poskytovaných zákazníkům zkrátit dobu jejich čekání u pokladen. Náhodně bylo proto vybráno 10 zákazníků a byla změřena doba jejich čekání u pokladny (předpokládáme normalitu rozdělení doby čekání). Výsledky šetření (v sekundách): 50, 65, 30, 45, 45, 35, 55, 70, 65, 50. a) Určete bodový odhad průměrné doby čekání a ohodnoťte ji standardní chybou průměru. b) V jakých mezích lze s pravděpodobností 0,95 očekávat průměrnou dobu čekání zákazníka na obsluhu? c) Jaká je horní hranice doby čekání, která nebude s pravděpodobností 0,95 překročena? d) Odhadněte bodově a v jakých mezích lze s pravděpodobností 0.95 očekávat rozptyl (resp. směrodatnou odchylku) doby čekání na obsluhu?
2) Z provozních důvodu sledujeme dobu životnosti žárovek od určitého dodavatele. Chceme zjistit, kolik žárovek musíme vybrat, abychom odhad střední hodnoty životnosti provedli s 95% spolehlivostí, jestliže jsme ochotni připustit maximální možnou chybu ve výši ±35 hodin.
3) Při výrobě určitých komponentů jsme dosud používali některé komponenty dovážené ze země Z. Ale obchod s touto zemí se velmi zkomplikoval a my jsme byli nuceni změnit dodavatele. Zajímá nás, zda změnou dodavatele nedošlo i ke změně kvality našich výrobků. Zatímco dříve bylo mezi našimi výrobky v průměru 5% zmetků, zjistila výstupní kontrola mezi 250 nově vyrobenými výrobky 16 nevyhovujících. Na základě 95% intervalu spolehlivosti rozhodněte, zda došlo ke změně kvality výrobků.

4) Jaký minimální rozsah výběru pro odhad podílu chybně zaúčtovaných položek musíme navrhnout, chceme-li při 90% spolehlivosti zajistit přípustnou chybu 3%. O možném podílu chybných položek nemáme při prováděném auditu žádnou informaci.

? Úkoly k zamyšlení a diskuzi

1) Když zvětšíme rozsah výběru čtyřikrát, co se stane se standardní chybou průměru?
2) Jaké budou hodnoty kvantilu u u-intervalu pro hladinu významnosti 90% a 99%?
3) Komentujte intervaly spolehlivosti z hlediska vlivu změny hladiny spolehlivosti.
4) Na čem všem záleží velikost optimálního rozsahu výběrového souboru?

Klíč k řešení otázek:

1) Vložíme v R data do souboru cekani proměnné doba (pomocí editoru, tj. z nabídky vybereme Data – New data set…, do vstupního okna vyplníme: cekani, OK, v Data Editor ťukneme na var1, ve Variable editor přepíšeme *var1* na doba a zaškrtneme *type*: numeric a uložíme křižkem v pravém horním rohu, potom zapíšeme do 1.sloupce tabulky data a vše opět uložíme křížkem v pravém horním rohu).
   Po aktivování balíku vsePackage lze bodový a oboustranný intervalový odhad (u-interval a t-interval) pro populační průměr E(X) = μ dostat příkazem *estim.mean(cekani$doba)*, napíšeme ho do *Script Windows* a odešleme ho pomocí *Submit*. Pokud nezadáme typ intervalu a hladinu významnosti, počítá R obvyklý oboustranný interval na hladině významnosti 95%. Výstup z počítače je:

```
> estim.mean(cekani$doba, type="two.sided", conf.level=0.95)
Data: cekani$doba
   Estimate of E(X)  = 51
   Estimate of sd(X) = 13.08094
   Sample size       = 10
   95% confidence interval for E(X) based on the asymptotical approximation:
       (42.89250,  59.1075)
```

```
   95% confidence interval for E(X) based on the assumption of normality:
       (41.64246,  60.35754)
```

Interpretace: a) Bodový odhad průměrné doby čekání je 51 sekund, standardní chyba průměru je 13,08/√10 = 4,14 sekund, b) intervalový odhad: S 95% spolehlivostí se průměrná doba obsluhy pohybuje v intervalu od 41,64 do 60,36 sekund. c) Určujeme pravostranné intervaly spolehlivosti. Napíšeme do vstupního okna příkaz *estim.mean(cekani$doba, type="less")* a dostaneme:

```
> estim.mean(cekani$doba, type="less",conf.level=0.95)
Data: cekani$doba
   Estimate of E(X)  = 51
   Estimate of sd(X) = 13.08094
   Sample size       = 10
   95% confidence interval for E(X) based on the asymptotical approximation:
       (-Inf,  57.80403)
   95% confidence interval for E(X) based on the assumption of normality:
       (-Inf,  58.58278)
```

Interpretace: Horní hranice doby čekání, která nebude s 95% pravděpodobností překročena je 58,58 sekundy. d) Počítáme bodový odhad a oboustranný interval spolehlivosti pro rozptyl (resp. směrodatnou odchylku) doby čekání. Dostaneme ho příkazem: *estim.var(cekani$doba)*. Výstup:

```
Data: cekani$doba
   Estimate of var(X)       = 171.1111
   Estimate of std. dev.(X) = 13.08094
   Sample size              = 10
   95% confidence interval for var(X) based on the assumption of normality:
       (80.95562,  570.2881)
   95% confidence interval for std. dev.(X) based on the assumption of norma-
lity:
       (8.997534,  23.88071)
```

Interpretace: Bodový odhad rozptylu doby čekání je 171,11 s^2, směrodatné odchylky 13,08 sekund. S 95% spolehlivostí se bude rozptyl pohybovat v mezích od 80,95 do 570,29 s^2 a směrodatná odchylka od 9,00 do 23,88 s.

2) Požadovaný rozsah souboru: n = 35 (výpočet viz Marek a kol. 2007, s. 128).

3) $0{,}034 < \pi < 0{,}094$; interval spolehlivosti pro poměr obsahuje hodnotu 0,05; dá se tedy předpokládat, že změna dodavatele neměla za následek změnu kvality našich výrobku (výpočet viz Marek a kol. 2007, s. 130).

4) Při řešení využijeme vztah $n \geq u_{1-\alpha/2}^2 \frac{p(1-p)}{\Delta^2} = 1{,}645^2 \frac{0{,}5(1-0{,}5)}{0{,}03^2} = 751{,}7.$ I za nejméně příznivých okolností nám rozsah souboru 752 účetních položek zajistí požadovanou spolehlivost a přesnost odhadu.

Kapitola 6: Testování statistických hypotéz

Klíčové pojmy:

statistický test, nulová a alternativní hypotéza, jednostranné a dvoustranné testy, testové kritérium, testovací statistika, hladina významnosti, chyba 1. a 2. druhu, síla testu, kritický obor, kritická hodnota, obor přijetí, věcná interpretace testu, parametrické a neparametrické testy, jednovýběrový u-test o průměru a t-test o průměru, jednovýběrový test o rozptylu a o poměru, p-hodnota testu, testování pomocí intervalu spolehlivosti, Shapirův-Wilkův test, grafické metody ověřování normality, Wilcoxonův jednovýběrový test, jednovýběrové testy v R

Cíle kapitoly:

- **pochopení základních pojmů o testování statistických hypotéz;**
- **porozumění strategie provádění klasických testů proti metodám používání p-hodnoty;**
- **zvládnout postup provádění parametrických i neparametrických testů v běžných situacích s využitím počítačových programů.**

Čas potřebný ke studiu kapitoly: 13 hodin

Výklad:

Nastínění obsahu kapitoly.

Úvod

Základní pojmy

Základní rozdělení testů

Testování hypotéz o parametrech normálního rozdělení

- Testy o průměru a rozptylu
- p-hodnota testu

Asymptotické testy

Shapirův-Wilkův test normality a grafické ověřování normality dat

Neparametrické testy

Testy v R

Struktura výkladu

Tak Vám nevím, jestli to její mrknutí bylo významné nebo ne.
Hanousek J., Charamza P.: Moderní metody zpracování dat

Úvod

Testování statistických hypotéz patří mezi základní metody statistické indukce a mezi nejjednodušší metody kvantitativní teorie rozhodování.

- Při řešení testovacího problému hledáme odpověď např. na otázky:

    - Způsobuje kouření rakovinu?
    - Ovlivní reklamní kampaň postoj spotřebitele k nově zaváděnému výrobku?
    - Zvyšují vhodné dávky daného hnojiva úrodu brambor?
    - Vede nový technologický postup ke změně jakosti výrobku?
- V pozadí těchto otázek stojí parametry (podíl osob postižených rakovinou, podíl zájemců o nový výrobek, průměrná velikost sklizně, podíl nekvalitních výrobků), jichž se otázka dotýká. Přejeme si posoudit, zda se tento parametr (označme jej obecně jako θ) nějakým systematickým způsobem změní (např. vroste), když v uvažované situaci dojde k nějaké zásadnější změně (osoba začala kouřit, proběhla reklamní kampaň, bylo použito intenzivnější hnojení, byl použit nový technologický postup).
- Odpověď na uvedené otázky získáváme z výběru (ne z celé populace) – statistickou indukcí.

Základní pojmy

Statistickou hypotézou je určité tvrzení o parametrech základního souboru (nebo o parametrech více souborů)

Testem statistické hypotézy nazýváme postup, jímž na základě výsledku zjištěných z náhodného výběru ověřujeme, zda statistickou hypotézu o populaci lze pokládat za správnou či nikoliv

- Např. testování změny poměru či průměru v předchozích případech.

Postup statistického testování:

- Formulace hypotéz.
- Výběr testového kritéria a jeho rozdělení.
- Volba hladiny významnosti.
- Vymezení kritického oboru.
- Výpočet hodnoty testového kritéria z měřených hodnot.
- Formulace závěru testu a jeho věcná interpretace.

Formulace hypotéz:

- Stavíme proti sobě nulovou hypotézu H_0: $\theta = \theta_0$ proti alternativní hypotéze H_1 (či H_A): $\theta \neq \theta_0$, kde θ_0 je hypotetická hodnota parametru.
    - Např. průměrná doba potřebná na určitou pracovní operaci je H_0: $\mu = 5$ min., H_1: $\mu \neq 5$ min.
- V tomto případě hovoříme o dvoustranné alternativě, resp. o dvoustranném testu.
- Jednostranné alternativy:
    - levostranná H_1: $\theta < \theta_0$,
    - pravostranná H_1: $\theta > \theta_0$.
- Levostranný test: H_0: $\theta \geq \theta_0$ proti H_1: $\theta < \theta_0$,
- Pravostranný test: H_0: $\theta \leq \theta_0$ proti H_1: $\theta > \theta_0$.

Výběr testového kritéria a jeho výběrové rozdělení:

- Testové kritérium - testovací statistika je funkce výběru $T = T(x_1,...,x_n)$, jejíž rozdělení je známé Nejčastěji: standardizovaný průměr, standardizovaný poměr apod.
    - Obvyklá rozdělení: N(0;1), Studentovo, chi-kvadrát aj.

Volba hladiny významnosti:

- Rozhodování o hypotézách se řídí rozhodovací tabulkou

Naše rozhodnutí	Skutečná situace	
	H_0 platí	H_0 neplatí
Nezamítneme H_0	správné rozhodnutí pravděpodobnost $1 - \alpha$	chyba II. druhu pravděpodobnost β
Zamítneme H_0	chyba I. druhu pravděpodobnost α hladina významnosti	správné rozhodnutí pravděpodobnost $1 - \beta$ síla testu

- Chyba 1. druhu - nesprávné zamítnutí nulové hypotézy.
- Chyba 2. druhu - chybné přijetí (nezamítnutí) nulové hypotézy.
- Ohodnocení těchto chyb pomocí pravděpodobností:

- Pravděpodobnost chyby 1. druhu:
    - $P(H_1 \mid H_0) = \alpha$.-. hladina významnosti (vyjadřuje se v %).
- Pravděpodobnost chyby 2. druhu:
    - $P(H_0 \mid H_1) = \beta$.
- Pravděpodobnost, že se nedopustíme chyby 2. druhu:
    - $P(H_1 \mid H_1) = 1 - \beta$ - nazýváme sílou testu (silofunkcí)

Vymezení kritického oboru (kritické oblasti):

- Obor hodnot testového kritéria dělíme na dvě disjunktní oblasti:
- a) Kritický obor W - množina hodnot kriteria T, které jsou při platnosti hypotézy H_0 málo věrohodné (chvosty rozdělení statistiky T) a má pravděpodobnost α , tj.:

$$P(T \in W \mid H_0) = \alpha$$ - pravděpodobnost chyby 1. druhu

    - Kritickým oborem W u pravostranného testu je pravý chvost, u levostranného - levý chvost a u dvoustranného - oba chvosty rozdělení
- b) Obor přijetí (akceptování, nezamítnutí) H_0 - množina hodnot V kritéria T, která při platnosti hypotézy H_0 je hodně věrohodná - má pravděpodobnost $1 - \alpha$, tj. platí:

$$P(T \in V \mid H_0) = 1 - \alpha.$$

- Pravděpodobnost chyby 2. druhu potom lze zapsat:

$$P(T \in V \mid H_1) = \beta.$$

    - Sílu testu počítáme z podmínky

$$P(T \in W \mid H_1) = 1 - \beta.$$

    - Kritické hodnoty testu – body oddělující kritický obor W od oboru přijetí V (= kvantilům rozdělení testového kritéria).
    - Snahou je, aby hodnoty α i β byly co nejmenší. Ale protože se zmenšováním hladiny významnosti roste chyba 2. druhu (viz následující obrázek), postupujeme takto:
    - Volíme co nejmenší α (0,05 nebo 0,01) a vybereme test, který při zvolené hladině významnosti maximalizuje sílu testu $1 - \beta$. Takový test se nazývá nejsilnější (určuje se v matematické statistice). α představuje riziko nesprávného zamítnutí nulové hypotézy. Chceme-li toto riziko snížit volíme α menší.

- Jediný způsob současného snížení α i β je zvětšení rozsahu výběru n.

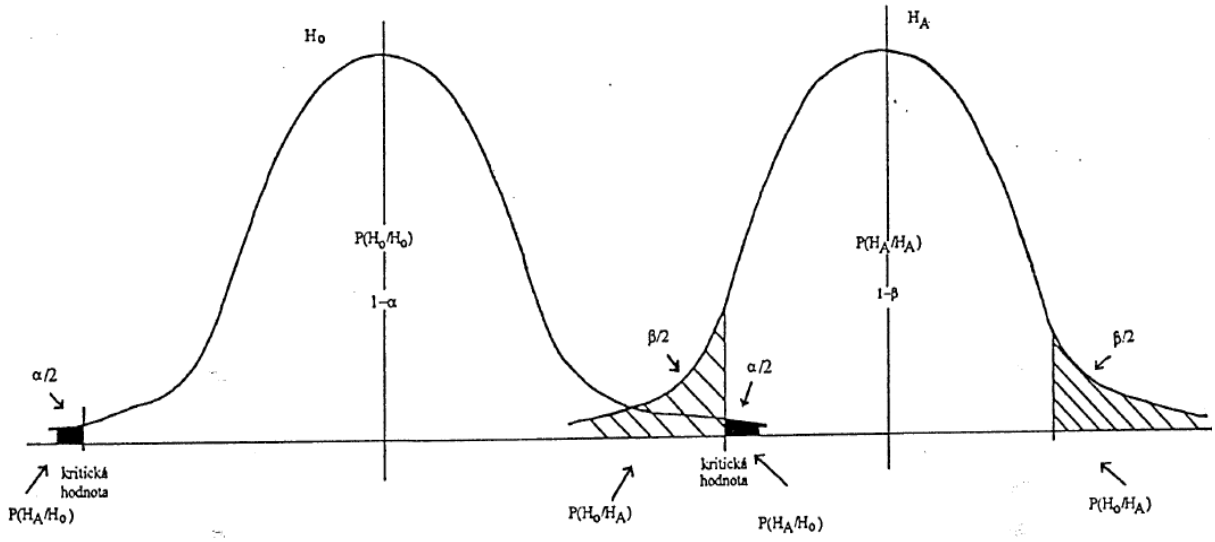
Výpočet hodnoty testového kritéria:

- Provedeme náhodný výběr n měření $x_1,...,x_n$ a z těchto měření určíme hodnotu testového kritéria - jde o nejpodstatnější část statistického testování.

Formulace závěru testu a jeho věcná interpretace:

- Rozhodnutí provedeme na základě následujícího pravidla:
- Rozhodovací pravidlo:
  - Je-li $T \in W$, zamítneme nulovou hypotézu H_0 na hladině významnosti α (na α%-ní hladině). Máme zaručeno, že chyba takového rozhodnutí nepřekročí α.
  - Je-li naopak $T \in V$ (tj. $T \notin W$) nezamítneme nulovou hypotézu H_0 na hladině významnosti α (budeme jí akceptovat). Skutečnost, že test nepotvrdil H_0 není na závadu, neboť za H_0 volíme obvykle tvrzení, které chceme zamítnout.
  - Můžeme se přitom dopustit chyby 2. druhu, která může mít dosti velkou pravděpodobnost β.
  - Proto raději neříkáme, že přijímáme H_0.
  - Obvykle je jen třeba zvýšit rozsah výběru n, aby se snížila chyba 2. druhu β, a test se stal průkaznější.
- Výsledek rozhodnutí věcně interpretujeme
  - Např. reklamní kampaň přinesla významné zvýšení poměru zájemců o nový výrobek, intenzivnější hnojení přineslo významné zvýšení průměrné úrody apod.
- Postup statistického testování můžeme porovnat s postupem u soudu

Na následujícím obrázku graficky znázorňujeme základní pojmy ze statistického testování hypotéz.



Zdroj: Čermáková 1995

Základní rozdělení testů

Podle předpokladů o rozdělení sledovaného statistického znaku:

1) Parametrické;
2) Neparametrické.

- Parametrické testy jsou založené na předpokladech o charakteru rozdělení statistického znaku a týkají se výhradně hodnot jednoho nebo několika parametrů daného rozdělení (např. středních hodnot, rozptylů apod.).
    - Nejčastěji předpokládáme normalitu rozdělení.
    - Jedná se o početně náročnější, avšak silné testy.
- Neparametrické testy - nevyžadují splnění téměř žádných předpokladů o charakteru rozdělení statistického znaku. Netýkají se parametrů rozdělení, tj. hypotézy neobsahují žádná tvrzení o průměrech či rozptylech, ale týkají se jiných charakteristik (např. mediánu).
    - Výhoda: mohou být použity pro studium jak kvantitativních tak kvalitativních znaků a po výpočetní stránce jsou jednoduché a rychlé.

- Nevýhoda: mají menší sílu.

Testování hypotéz o parametrech normálního rozdělení

Testy o populačním průměru:

- Předpoklad: $x_1,...,x_n \sim N(\mu;\sigma^2)$, kde σ^2 je známý parametr.
- Pravostranný test:
  - Testujeme H_0: $\mu = \mu_0$ proti alternativě H_1 : $\mu > \mu_0$, (μ_0 je známá hodnota)
  - Testové kritérium $U = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \sim N(0;1)$ při H_0
  - Hypotézu H_0 zamítáme na hladině α, když $U > u_{1-\alpha}$

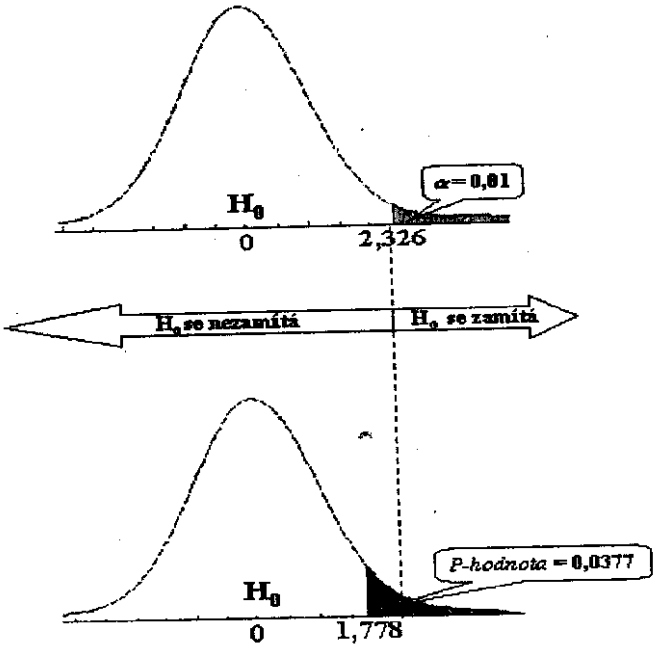
Příklad (Arltová a kol. 2003, s. 151-152):

- Lze vyrobit 1 mil. součástek určitého elektronického zařízení se střední životností 900 hodin a směrodatnou odchylkou 225 hodin. Vývojové oddělení ve svých dílnách vyrábí experimentálně tyto součástky novou technologií a tvrdí, že tak dosáhne vyšší průměrnou životnost. K ověření tohoto tvrzení byl ze součástek vyrobených novou technologií pořízen náhodný výběr 100 ks, u nichž průměrná životnost činila 940 hodin. Jeví se nová technologie na základě těchto výsledku lépe než původní? Nejedná se jen o náhodu?

Řešení:

- Testujeme hypotézu H_0: $\mu = 900$ proti pravostranné alternativě H_1: $\mu > 900$. Vypočítáme
$$U = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \frac{940 - 900}{225}\sqrt{100} = 1{,}778.$$

- Protože U=1,778 > $u_{0,95}$ = 1,645, zamítáme H_0 na 5%-ní hladině významnosti.
- Protože U=1,778 < $u_{0,99}$ =2,326, nezamítáme H_0 na 1%-ní hladině významnosti (viz obr.).
- Statistické programy provádějí rozhodnutí pomocí tzv. p-hodnoty testu, tj. pravděpodobnosti dosažení ještě extrémnější (více ve chvostu rozdělení) hodnoty než je vypočtená hodnota testového kritéria.



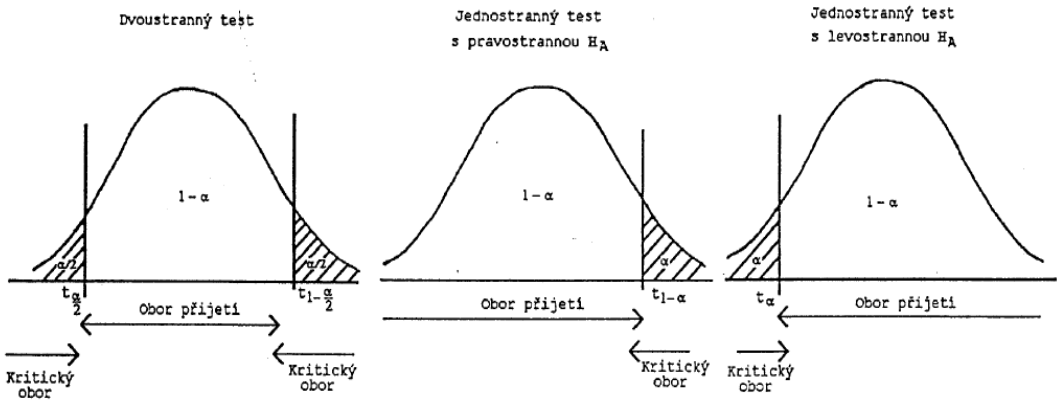
- p-hodnota = P(U ≥ 1,778 | H_0) = 0,0377 = minimální hladině, na které H_0 zamítáme.
- H_0 zamítáme, je-li p-hodnota < α.
- p-hodnotu lze totiž interpretovat jako pravděpodobnost nesprávného zamítnutí H_0. Protože jsme si ji předem zadali jako α, H_0 můžeme zamítnout, jen když p-hodnota nepřekročí α.
- Levostranný test:
  - Testujeme H_0: $\mu = \mu_0$ proti H_1 : $\mu < \mu_0$
  - H_0 zamítneme, když $U < u_\alpha$ nebo, když p-hodnota P(U ≤ vypočítaná .hodnota u | H_0)< α.
- Dvoustranný test:
  - Testujeme H_0: $\mu = \mu_0$, proti alternativní hypotéze H_1: $\mu \neq \mu_0$.
  - Použijeme testové kritérium $U = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \sim N(0,1)$ při H_0
  - H_0 zamítneme na hladině významnosti α na kritické oblasti $W = \{U: U \leq -u_{1-\alpha/2} \vee U \geq u_{1-\alpha/2}\}$, tj. když $|U| \geq u_{1-\alpha/2}$.
  - Nebo H_0 zamítneme na hladině významnosti α, je-li p-hodnota = P(|U| ≥ vyp.hodnota | H_0) < α.
    - Podobně používáme p-hodnoty i u dalších testů.

Pod pojmem klasický statistický test rozumíme testování pomocí kritických hodnot a kritických oborů.

Případ neznámého rozptylu:

- Použijeme testové kritérium $T = \frac{\bar{x} - \mu}{s / \sqrt{n}}$, které má při splnění H_0 rozdělení t(n-1).

- Pro vymezení kritických oblastí W používáme proto místo kvantilů rozdělení N(0;1) kvantily rozdělení t(n-1).
    - Hovoříme zde o t-testech na rozdíl od dřívějších u-testů.
- Rozhodování v klasických t-testech je znázorněno na následujícím obrázku.



Zdroj: Arltová 2003

Pro hypotetický průměr μ_0 platí: $$P\left(\bar{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{\sigma}{\sqrt{n}} < \mu_o < \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- Testování lze provést i pomocí tohoto intervalu spolehlivosti.
- Nulovou hypotézy H_0: $\mu = \mu_0$ zamítneme a přijmeme opačnou oboustrannou alternativu H_1, když μ_0 padne mimo tento interval spolehlivosti.

Příklad:

- Testujte hypotézu, že průměrný denní výtěžek určité chemikálie ve farmaceutické továrně je $\mu = 880$ tun proti alternativě, že $\mu \neq 880$ tun. Na vzorku z 50 dní (n=50) jsme zjistili, že výběrový průměr $\bar{x} = 871$ a výběrová směrodatná odchylka s = 21 tun.

Řešení:

- Testujeme nulovou hypotézu H_0: $\mu = 880$ proti oboustranné alternativě H_1: $\mu \neq 880$ na hladině významnosti $\alpha = 0,05$.
- Platí $T = (\bar{x} - m_0)/(s/\sqrt{n}) = (871-880)/(21/\sqrt{50}) = -3,0305$
- Tedy $|T| = 3,0305 > t_{0,975}(49) = 2,010$, tj. H_0 na 5%-ní hladině významnosti zamítáme a tvrdíme, že denní výtěžek se významně liší od 880 tun.

Test o populačním rozptylu:

- Předpokládejme, že náhodný výběr pochází z normálního rozdělení $N(\mu;\sigma^2)$, kde oba parametry jsou neznámé.
- Testujeme nulovou hypotézu H_0: $\sigma^2 = \sigma_0^2$, kde σ_0^2 je určitá předem zvolená hodnota, proti alternativě H_1: $\sigma^2 \neq \sigma_0^2$.
- Použijeme testové kritérium $\chi^2 = \frac{(n-1)S_x^2}{\sigma_0^2}$ které má při splnění nulové hypotézy rozdělení $\chi^2(n-1)$
- Kritický obor testu je $W = \{\chi^2 : \chi^2 \leq \chi^2_{\frac{\alpha}{2}}(n-1) \vee \chi^2 \geq \chi^2_{1-\frac{\alpha}{2}}(n-1)\},$

  kde v nerovnostech na pravé straně jsou příslušné kvantily rozdělení $\chi^2(n-1)$.
  - Podobně je možno zavést i jednostranné testy

Příklad: – Viz Hindls a kol. (2007), s. 142.

Asymptotické testy

Test o populačním průměru:

- Při velkém rozsahu souboru ($n > 30$) nemusí být splněn předpoklad o normalitě výběru ze základního souboru.
- Používáme testové kritérium $U = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, jehož normalita je zaručena z centrální limitní věty.

- Rozhodnutí proto provádíme pomocí kvantilů rozdělení N(0;1).

Příklad: – Viz Hindls a kol. (2007), s. 139.

Test o populačním poměru:

- Testujeme hypotézu H_0: $\pi = \pi_0$ proti alternativě H_1: $\pi \neq \pi_0$.
- Použijeme testové kritérium $U = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}.$

  Toto má při H_0 podle Moivreovy-Laplaceovy limitní věty přibližně rozdělení N(0;1).
- Kritickým oborem je $W = \{U : |U| \geq u_{1-\frac{\alpha}{2}}\}.$

Příklad – viz Hindls a kol. (2007), s. 140.

Shapiro-Wilkův test (SWT) a grafické ověřování normality

Test budeme potřebovat na ověření normality dat vždy při malém počtu měření ($n \leq 30$).

Testujeme H_0: X má normální rozdělení proti opačné alternativě.

Test (včetně rozdělení testové statistiky), navržený v práci Shapiro a Wilk (1965), využívá k ověření normality testové statistiky $W = \frac{S^2_{x,Norm}}{S^2_x}$, jenž je podílem dvou odhadů rozptylů: klasického S_x^2 a odhadu $S_{x,Norm}^2$, který je platným odhadem rozptylu za předpokladu, že data jsou normálně rozdělena. V případě, že platí nulová hypotéza normality, máme jak v čitateli, tak ve jmenovateli statistiky W platný odhad rozptylu a W tedy musí být blízké hodnotě 1. Pokud nulová hypotéza normality neplatí, potom je W výrazněji vzdáleno od 1. Přesný význam slova „výrazněji“ přitom závisí na rozsahu výběru n. Rozdělení statistiky W při platnosti H_0 umožňuje výpočet p-hodnot a určení tabulek kritických hodnot. p-hodnoty poskytují statistické programy (např. R) a tabulku kritických hodnot (včetně podrobnějšího popisu testu) poskytuje Hebák a kol. (2004).

Graficky je možno provádět testování normality dat z vyhodnocení krabicového diagramu (symetrie, malé množství odlehlých hodnot), z porovnání histogramu s křivkou příslušného normálního rozdělení (symetrie, unimodalita) a z qq-diagramu, do kterého zakreslujeme empirické kvantily a teoretické kvantily normálního rozdělení N(μ; σ2), počítané podle vztahu

$$F^{-1}_{\mu,\sigma^2}(p) = \mu + \sigma\, u_{p.}$$

Za parametry dosazujeme jejich odhady. Body v grafu by ideálně měly ležet na přímce.

Příklady – viz úkoly.

Neparametrické testy

- Nemá-li X normální rozdělení a počet měření je malý (do 30 měření) používáme místo t-testu Wilcoxonův test.
- Za charakteristickou hodnotu úrovně používáme obvykle medián Me(X) místo průměrů μ

Wilcoxonův jednovýběrový test úrovně (WJT)

- Testujeme hypotézu H_0: $Me(X) = m_0$ proti alternativní hypotéze H_1: $Me(X) \neq m_0$
- Wilcoxonův test
  - Počítáme pořadí od nejmenších k největším číslům $|x_i - m_0|$.
  - R^+ a R^- označuje součet těchto pořadí pro kladné nebo záporné $x_i - m_0$.
    - Nulové hodnoty vynecháváme.
    - K stejným hodnotám počítáme průměrná pořadí.
- Testové kritérium: $T = \min(R^+, R^-)$
  - Příznivé alternativě – nízká hodnota T.
- Kritická oblast: $W = \{ T: T \leq T_{\alpha/2}\}$, kde $T_{\alpha/2}$ je $100(\alpha/2)\%$ kvantil jednvýběrové Wilcoxonovy statistiky T (v R *qsignrank(p, n))* – viz tab. VI. v dodatku.
  - Pravostranný test H_1:$Me(X)>m_0$, $W = \{T=R^-: R^- \leq T_\alpha\}$

- Levostranný test H_1:Me(X)< m_0 , W = {T=R^+: $R^+ \leq T_\alpha$}
- Podrobnější zavedení testu a příklady - viz Stuchlý (2004), s. 35-37, 167 nebo Blatná (1996), s. 86-91, 179.

Jednovýběrové testy v R

- SWT (test normality)
  - Provádíme v *Statistics+Summaries+Shapiro-Wilks* test of normality
- Grafické ověření normality dat provedeme v R Commanderu takto:
  - Vybereme z nabídky *Graphs*
    - *Boxplot*
    - *Quantile-comparison plot*
    - Histogram (zaškrtneme *densities*)
      - Za příkaz histogramu připíšeme
      - x<-a:b
      - lines(x, dnorm(data$pr, mean(data$pr), sd(data$pr))
    - Zde data$pr je proměná „pr“ z datového souboru „data“, v kterém je proměnná „pr“ uložena a interval a:b je celočíselné rozmezí, v kterém se „pr“ pohybuje.
    - Histogram dostaneme také příkazem:
    - hist(data$pr, scale="density", breaks="Sturges", col= "darkgray", ylim =c(0,0.045))
- t-test střední hodnoty
  - Při normalitě dat provádíme v *Statistcs+Means+Single-sample t-test*
  - Asymptotický test po aktivaci balíku *vsePackage* příkazem:
    - asymp.mean.test(x, mu=),
    - resp.: asymp.mean.test(x, mu= , conf.level=)
    - Pro pravostranný test: asymp.mean.test(x, mu= , type ="greater")
    - Pro levostranný test: asymp.mean.test(x, mu= , type =„less")
- Wilcoxonův jednovýběrový test
  - Při malém výběru a nenormalitě dat

    - Provádíme příkazem: wilcox.test(x, mu=)
- Test variability
    - Provádíme příkazem: onesample.var.test(x, sd=) nebo onesample.var.test(x, var=) po aktivaci balíku *vsePackage.*
- Test o poměru u alternativní proměnné
    - Např. podíl mužů proti ženám
    - Provádíme v *Statistics+Proportion+Single-sample proportion test*, popř. příkazem prop.Z.test(x, n, p= , alternative=" ", conf.level=) po aktivaci balíku *vsePackage*
    - Zavedení podmnožiny dat podle určité proměnné (odpovídá filtrování v Excelu): *Data+Active data set+Subset active data set*.

Excel jednovýběrové testy nezahrnuje.

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 133-142. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE, 1999. S. 111-115, 117-122, 125-127. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 140-158, 171-173. ISBN 80-245-0178-3.

BÍNA V. a kol. *Jak na jazyk R.* J. Hradec: FM VŠE, 2006.

BLATNÁ, D. *Neparametrické metody.* Praha: VŠE 1996. ISBN 80-7079-607-3.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 256-263. ISBN 80-7261-003-1.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 69-73. ISBN 80-7169-255-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 132-139, 167-168. ISBN 978-80-86446-40-5.

SEGER, J. a HINDLS, R. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 127-137. ISBN 80-7187-058-7.

STUCHLÝ, J. *Statistické metody pro manažerské rozhodování*. J. Hradec: VŠE, 2004. S. 35-37. ISBN 80-245-0153-8.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/).

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 209-222, 226-222, 226-230. ISBN 80-7169-089-9.

WONNACOT, T. H. a WONNACOT, R. J. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 310-350. ISBN 80-85605-09-0.

? Otázky a úkoly

Pro úkoly v této kapitole budeme používat datový soubor studenti.dat (popř. studenti.csv), který je v studijních materiálech na IS VŠTECB

1) Na hladině významnosti 5% otestujte normalitu rozdělení výšky studentů studujících na pražských fakultách VŠE. Grafickými metodami ověřte získaný výsledek.
2) Pomocí vhodného testu zjistěte, zda průměrná výška studentů studujících na pražských fakultách VŠE je různá od 178 cm. Rozhodnutí proveďte pomocí a) kritického oboru,

b) intervalu spolehlivosti, c) p-hodnoty. Při rozhodování použijte hladinu významnosti 5%. Změní se rozhodování na hladině významnosti 1% či 10%?

3) Pomocí vhodného testu zjistěte, zda průměrná výška studentů studujících na pražských fakultách VŠE je nižší než 180 cm. Rozhodnutí proveďte pomocí a) kritického oboru, b) intervalu spolehlivosti, c) p-hodnoty. Při rozhodování použijte hladinu významnosti 5% a 1%.

4) Pomocí vhodného testu zjistěte, zda průměrná váha studentů na VŠE je 75 kg.

5) Pomocí vhodného testu zjistěte, zda směrodatná odchylka výšky studentů studujících na pražských fakultách VŠE je a) 11,5 cm, b) nižší než 11.5. Při rozhodování použijte hladinu významnosti 5% a 10%. Při volbě vhodného testu nezapomeňte ověřit jeho předpoklady. Pokud jste zvolili parametrický test, určete odpovídající interval spolehlivosti.

6) Pomocí vhodného testu zjistěte, zda u studentů cestujících vlakem je typická vzdálenost od školy a) 220 km, b) méně než 220 km. Při rozhodování použijte hladinu významnosti 5% a 10%. Při volbě vhodného testu nezapomeňte ověřit jeho předpoklady. Pokud jste zvolili parametrický test, určete odpovídající interval spolehlivosti.

7) Testujte hypotézu, že typický podíl studentů cestujících vlakem je 60% proti alternativě, že je menší.

? Úkoly k zamyšlení a diskuzi

1) Diskutujte o analogii statistického testování s rozhodováním u soudu.

2) Zamyslete se nad tím, jak souvisí statistická přejímací kontrola s testováním statistických hypotéz?

Klíč k řešení otázek:

1) Testujeme nulovou hypotézu H_0: výška má normální rozdělení proti opačné alternativě H_1. Použijeme v R Commanderu Shapiro-Wilkův test.

Výstup:

```
> shapiro.test(studenti$vyska)
        Shapiro-Wilk normality test
data:  studenti$vyska
W = 0.9816, p-value = 0.4751
```

Závěr: Nezamítáme H_0. Výška studentů se řídí normálním rozdělením.

Graficky ověříme normalitu proměnné „vyska“ v R pomocí grafu histogramu, krabicového diagramu a qq-diagramu. Použijeme příkazy (nebo postupujeme interaktivně z nabídky Graphs):

hist(studenti$vyska, scale="density", breaks="Sturges", col="darkgray")

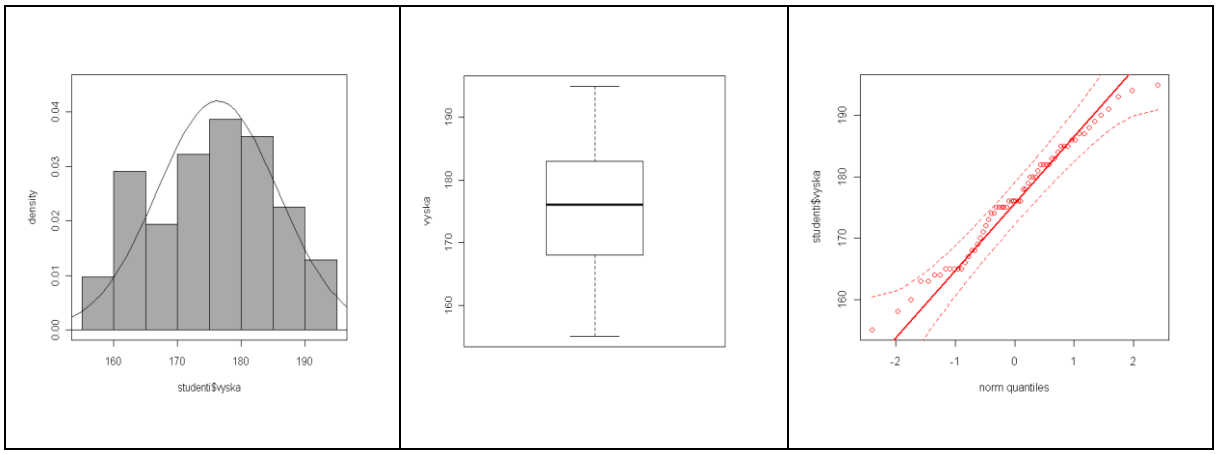
x<-150:200

lines(x,dnorm(x, mean(studenti$vyska),sd(studenti$vyska)))

boxplot(studenti$vyska, ylab="test")

qq.plot(studenti$test, dist= "norm", labels=FALSE)

a dostaneme následující grafy,



což potvrzuje normalitu proměnné výška.

2) Testujeme H_0: $\mu=178$ proti H_1: $\mu\neq178$. Použijeme oboustranný jednovýběrový t-test.

Výstup:

```
>  t.test(studenti$vyska,  alternative='two.sided',  mu=178,  conf.level
=.95)
 One Sample t-test
data:  studenti$vyska
t = -1.4187, df = 61, p-value = 0.1611
alternative hypothesis: true mean is not equal to 178
```

```
95 percent confidence interval:
 173.8805 178.7001
sample estimates:
mean of x
 176.2903
> qt(c(0.975), df=61, lower.tail=TRUE)
[1] 1.999624
```

Závěry:

a) $|t|=1{,}42<t_{0{,}975}(61) = 2{,}00$

b) $178 \in (173{,}88\ 178{,}70)$

c) p-hodnota = $0.16 > \alpha=0{,}05$

H_0 nezamítáme na 5% hladině významnosti, tj. test neprokázal, že průměrná výška studentů je odlišná od 178 cm. Závěr platí i na hladinách 1% a 10% (podle c).

3) Testujeme H_0: $\mu = 180$ proti H_1: $\mu < 180$. Použijeme levostranný jednovýběrový t-test.

Výstup:

```
> t.test(studenti$vyska, alternative='less', mu=180, conf.level =.95)
 One Sample t-test
data:  studenti$vyska
t = -3.0782, df = 61, p-value = 0.001559
alternative hypothesis: true mean is less than 180
95 percent confidence interval:
     -Inf 178.3032
sample estimates:
mean of x
 176.2903
> qt(c(0.05), df=61, lower.tail=TRUE)
[1] -1.670219
```

Závěry:

a) $t=-3{,}08<-1{,}67$

b) $180 \notin (-\infty, 178{,}3)$

c) p-hodnota= $0{,}00156 < \alpha = 0{,}05$

H_0 na 5% hladině významnosti zamítáme, tj. test prokázal, že průměrná výška studentů je menší než 180 cm. Na 1% hladině významnosti dostaneme stejný závěr (p-hodnota < 0,01).

4) Ukážeme nejdříve SWT, že váha studentů se neřídí normálním rozdělením:

```
> shapiro.test(studenti$vaha)
        Shapiro-Wilk normality test
data:  studenti$vaha
W = 0.9221, p-value = 0.0007557
```

Máme 62 měření, můžeme použít asymptotický test o průměru. Testujeme hypotézu H_0: $\mu = 75$ proti opačné hypotéze H1. Odešleme příkaz asymp.mean.test(studenti$vaha, mu=75)

Výstup:

```
> library(vsePackage)
> asymp.mean.test(studenti$vaha, mu=75)
Asymptotical test for the expected value
    Alternative hypothesis: true mean is not equal to 75
Data var1
Z = -2.54766,   p-value = 0.01084481
Estimate of the true mean = 70.5
95% confidence interval: (67.03806,  73.96194)
```

Závěr: Na 5% hladině významnosti zamítneme H_0, tj. průměrná váha studentů na VŠE není 75 kg.

5) a) Pro výšku studentů testujeme hypotézu H_0: $\sigma = 11{,}5$ proti opačné alternativě H_1. Použijeme test o populačním rozptylu (směrodatné odchylce). Podmínky na jeho použití (normalita výšek) je splněna. V R používáme příkaz *onesample.var.test(studenti$vyska, sd=11.5)*

Výstup:

```
> library(vsePackage)
> onesample.var.test(studenti$vyska, sd=11.5)
One-sample test for the variance of normal data
    Alternative hypothesis: true variance is not equal to 132.25
                            true std. dev. is not equal to 11.5
Data: studenti$vyska
SS = 41.53326,   p-value = 0.05324797
Estimate of the true variance = 90.04548
   95% confidence interval: (65.02138,  132.9868)
Estimate of the true std. dev.= 9.48923
   95% confidence interval: (8.063583,  11.53199)
```

Závěr: H_0 na 5% hladině i 10% významnosti nezamítáme, tj. směrodatná odchylka výšky studentů není 11,5; 95% interval spolehlivosti je (8,06;11,53).

b) Pro výšku studentů testujeme hypotézu H_0: $\sigma = 11{,}5$ proti alternativě H_1: $\sigma < 11{,}5$. Použijeme příkaz *onesample.var.test(studenti$vyska, sd=11.5, alternative="less")*

Výstup:

```
> library(vsePackage)
> onesample.var.test(studenti$vyska, sd=11.5, alternative="less")
One-sample test for the variance of normal data
    Alternative hypothesis: true variance is less than 132.25
                            true std. dev. is less than 11.5
Data: studenti$vyska
SS = 41.53326,   p-value = 0.02662398
Estimate of the true variance = 90.04548
   95% confidence interval: (0,  124.7284)
Estimate of the true std. dev.= 9.48923
   95% confidence interval: (0,  11.16819)
```

Závěr: H_0 na 5% i 10% hladině významnosti zamítáme, tj. směrodatná odchylka výšky studentů je nižší než 11,5. 95% interval spolehlivosti pravostranný: (0; 11,17).

6) Nejdříve zavedeme podsoubor studentů, kteří cestují do školy vlakem takto:
V nabídce *Data – Active data set – Subset active data set* vyplníme *Subset expression:* doprava=="V" a *Name for data set:* vlak (filtrování v R) a testujeme SWT normalitu proměnné bydliště.

Výstup:

```
 > vlak <- subset(studenti, subset=doprava=="V")
> shapiro.test(vlak$bydliste)
 Shapiro-Wilk normality test
data:  vlak$bydliste
W = 0.7851, p-value = 0.0001296
```

Proměnná není normálně rozdělená. Testujeme H_0: Me(bydliště)=220 proti opačné H_1. Použijeme oboustranný Wilcoxonův jednovýběrový test příkazem *wilcox.test(vlak$bydliste, mu=220)*

Výstup:

```
> wilcox.test(vlak$bydliste, mu=220)
 Wilcoxon signed rank test with continuity correction
data:  vlak$bydliste
```

```
V = 141, p-value = 0.808
alternative hypothesis: true location is not equal to 220
```

Závěr: Na 5% hladině významnosti nezamítáme H_0, tj. test neprokázal, že typická vzdálenost bydliště studentů, dopravujících se do školy vlakem, se liší od 220 km.

(b) Testujeme H_0: Me(bydliště) = 220 proti pravostranné H_1: Me(bydliště) < 220. Použijeme levostanný WJT. Odešleme příkaz *wilcox.test(vlak$bydliste, mu=220, alternative= "less")*

Výstup:

```
> wilcox.test(vlak$bydliste, mu=220, alternative="less")
 Wilcoxon signed rank test with continuity correction
data:  vlak$bydliste
V = 141, p-value = 0.404
alternative hypothesis: true location is less than 220
```

Závěr: H_0 nezamítáme na 5% hladině významnosti, tj. test neprokázal, že typická vzdálenost bydliště studentů, dopravujících se do školy vlakem, je menší než 220 km. Oba závěry platí i na 10% hladině významnosti.

7) Testujeme hypotézu H_0: $\pi = 0{,}6$ proti alternativě H_1: $\pi < 0.6$. Použijeme asymptotický test o populačním poměru. Pomocí *Statistics – Summaries – Active dataset* zjistíme, že z 62 studentů jezdí jich vlakem 25 (proměnná „doprava“ nabývá hodnoty „T“). K provedení testu odešleme příkaz: *prop.Z.test(25, 62, p=0.6 , alternative ="less")*

Výstup:

```
> library(vsePackage)
> prop.Z.test(25, 62, p=0.60 , alternative="less")
1-sample proportions test based on asymptotical normality
    Alternative hypothesis: true pi is less than 0.6
Data:  25 successes out of 62 trials
Z = -3.162703,   p-value = 0.0007815595
Estimate of the proportion of successes:  0.4032258
Estimated SE of the estimate:             0.06229932
95 % confidence interval for the proportion of successes:
    (0,  0.5076619)
```

Závěr: Zamítáme H_0 na 5% hladině významnosti. Test prokázal na 5% hladině významnosti, že typický poměr studentů jezdících do školy vlakem je menší než 60%.

Kapitola 7: Dvouvýběrové testy

Klíčové pojmy:

dvouvýběrové testy, dvouvýběrový F-test o shodě rozptylů, dvouvýběrový u-test o shodě průměrů, dvouvýběrový t-test o shodě průměrů, zobecněný dvouvýběrový test o shodě průměrů, dvouvýběrový asymptotický u-test o shodě průměrů a o shodě poměrů, dvouvýběrový t-test o shodě průměru pro závislé výběry, dvouvýběrový Wilcoxonův test pro závislé a pro nezávislé výběry, Mannův-Whitneyův test, Kolmogorovův-Smirnovův dvouvýběrový test

Cíle kapitoly:

- **pochopení pojmu dvouvýběrový test jako prostředek analýzy numerické proměnné na proměnné alternativní;**
- **naučit se používat dvouvýběrové testy parametrické i neparametrické;**
- **umět rozlišit, kdy je který test potřebný použít.**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly.

Testy o shodě parametrů dvou nezávislých i závislých normálních souborů

- Testy o shodě rozptylů a středních hodnot

Asymptotické testy o shodě parametrů dvou souborů

- Testy o shodě průměru a poměrů

Neparametrické testy

- Mannův-Whitneyův a Wilcoxonův dvouvýběrový test
- Kolmogorovův-Smirnovův dvouvýběrový test

Dvouvýběrové testy v R a Excelu

Struktura výkladu

Hypotézy jsou lešením, které se staví před budovou a pak se strhává, je-li budova postavena. Jsou nutné pro vědeckou práci, avšak skutečný vědec nepokládá hypotézy za předmětnou pravdu, podobně jako nelze pokládat lešení za stavbu samu.

J.W.Goethe

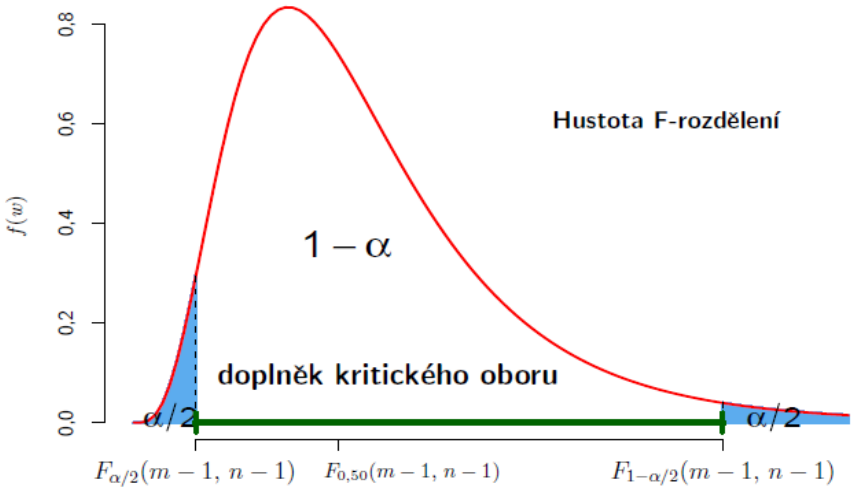
Testy o shodě parametrů dvou normálních soubor.

Předpokládejme nejdříve, že jsou dané dva nezávislé náhodné výběry $x_1,...,x_m$ a $y_1,...,y_n$, které pocházejí z normálních rozdělení $N(\mu_1,\sigma_1^2)$ a $N(\mu_2,\sigma_2^2)$

Dvouvýběrový F-test o shodě populačních rozptylů:

- Testujeme hypotézu H_0: $\sigma_1^2=\sigma_2^2$ proti alternativě H_1: $\sigma_1^2\neq\sigma_2^2$
- Použijeme testové kritérium $F=s_1^2/s_2^2 \sim F(m-1,n-1)$, při H_0.
- H_0 zamítáme, když $F<F_{\alpha/2}(m-1,n-1)$ nebo $F>F_{1-\alpha/2}(m-1,n-1)$.
    - Výrazy uvedené na pravých stranách nerovností jsou kvantily Fisherova rozdělení F(m-1, n-1).
    - Tento test je důležitý pro správné vymezení, který test o průměrech použijeme.

Grafické znázornění F-testu (testování shody rozptylů):



Test o shodě populačních průměrů:

Předpokládejme nejdříve, že rozptyly : σ_1^2, σ_2^2 jsou známé.

- Testujeme hypotézu H_0: $\mu_1 = \mu_2$ proti alternativě H_1: $\mu_1 \neq \mu_2$.
- Použijeme testové kritérium $U = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} \sim N(0;1)$
- při platnosti H_0
- H_0 zamítáme, když $|U| > u_{1-\frac{\alpha}{2}}$.

- Testujeme-li hypotézu H_0: $\mu_1 = \mu_2$ proti pravostranné alternativě H_1: $\mu_1 > \mu_2$, zamítáme H_0, když $U > u_{1-\alpha}$.
- Testujeme-li hypotézu H_0: $\mu_1 = \mu_2$ proti levostranné alternativě H_1: $\mu_1 < \mu_2$, zamítáme H_0, když $U < -u_{1-\alpha}$.

Jsou-li populační rozptyly neznámé ale stejné použijeme kritérium

$$T = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{mn}{m+n}}, \text{ kde } S^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}$$

- Při platnosti H_0 je $T \sim t(m+n-1)$. Rozhodnutí proto provedeme pomocí příslušných kvantilů tohoto rozdělení – jde o klasický dvouvýběrový t-test.

Příklad: – Viz Stuchlý (1999a), s. 122.

Jsou-li populační rozptyly neznámé a různé použijeme kritérium

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{S_x^2}{m} + \dfrac{S_y^2}{n}}} \sim t(\nu), \text{ kde } \nu = \frac{117\left(\dfrac{S_x^2}{m} + \dfrac{S_y^2}{n}\right)^2}{\dfrac{1}{m-1}\left(\dfrac{S_x^2}{m}\right)^2 + \dfrac{1}{n-1}\left(\dfrac{S_y^2}{n}\right)^2}$$

při H_0 a k rozhodnutí proto použijeme kvantily tohoto rozdělení – jde o zobecněný dvouvýběrový t-test (Welchova aproximace).

Příklad: – viz Stuchlý (1999a), s. 123.

Jsou-li oba výběry normální závislé s m = n (párová měření), počítáme $D_i = x_i - y_i$ a test provádíme jako jednovýběrový test o parametru průměru aplikovaném na tyto rozdíly – jde o párovy dvouvýběrový t-test.

Příklad – Viz Stuchlý (1999a), s. 124.

Asymptotické dvouvýběrové testy

Jsou-li rozsahy obou souborů velké (m > 30, n > 30), můžeme upustit od předpokladu normality obou souboru.

Asymptotický u-test o shodě populačních průměrů:

- V důsledku centrální limitní věty testové kritérium $$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

  má rozdělení N(0;1). Tudíž k rozhodnutí používáme u-kvantily.
  R tento test neobsahuje. Proto zde používáme t-test.

Asymptotický dvouvýběrový test o populačních poměrech:

Předpokládejme, že máme dva velké výběry $x_1,...,x_m$ a $y_1,...,y_n$ (m, n jsou velké), které pocházejí z alternativních rozdělení $A(\pi_1)$ a $A(\pi_2)$, kde π_1, π_2 jsou neznámé parametry, představující populační poměry. Označme p_1, p_2 odpovídající výběrové poměry.

- Testujeme hypotézu H_0: $\pi 1 = \pi 2$ proti alternativě H_1: $\pi_1 \neq \pi_2$. Použijeme testové kritérium
$$U = \frac{p_1 - p_2}{\sqrt{p^*(1-p^*)(\frac{1}{m} + \frac{1}{n})}}, \quad \text{kde } p^* = \frac{p_1 m + p_2 n}{m+n},$$

- které má při H_0 rozdělení N(0;1).

- Nulovou hypotézu H_0 proto zamítáme na kritickém oboru

  $W = \{U: |U| > u1\text{-}\alpha/2\}$.
- Podobně postupujeme při jednostranných testech.

Příklad: – Viz Stuchlý (1999a), s. 125.

Neparametrické testy

Budeme se nejprve zabývat testy o shodě úrovní.

Porovnávání úrovní při neparametrických testech se obvykle provádí porovnáváním mediánů $Me(X_i)$ místo průměrů μ_i.

Rozlišujeme:

- Závislé výběry – stejné rozsahy výběrů, pro 2 výběry – párová měření.
- Nezávislé výběry – mohou být různé rozsahy výběrů.

Testy úrovně pro dva závislé výběry:

- Testujeme hypotézu H_0: $Me(X) = Me(Y)$ proti alternativní hypotéze H_1: $Me(X) \neq Me(Y)$

Wilcoxonův párový test:

- Počítáme pořadí od nejmenších k největším číslům $|x_i - y_i|$,
- T^+ a T^- označuje součet těchto pořadí pro kladné nebo záporné $x_i - y_i$.
- Nulové hodnoty vynecháváme.
- K stejným hodnotám počítáme průměrná pořadí.

  - Testové kritérium: $T = \min(T^+, T^-)$.

  - Kritická oblast: $W = \{ T: T \leq T_{w;\alpha/2}\}$, kde $T_{w;\alpha/2}$ je $100\alpha/2$-procentní kvantil jednovýběrové Wilcoxonovy statistiky T_w (viz tab. VI. v dodatku).

    - Pravostranný test H_1:$Me(X)>Me(Y)$, $W = \{T^-: T^- \leq T_{w;\alpha}\}$.
    - Levostranný test H_1:$Me(X)<Me(Y)$, $W = \{T^+: T^+ \leq T_{w;\alpha}\}$.

Příklad:

Majitel firmy Okula prodává fialové a růžové sluneční brýle. Zajímá ho, jak jdou na odbyt tyto barvy. Chce vědět, zda jsou rozdíly v prodaných množstvích těchto dvou barev nebo ne. V 16 náhodně vybraných dnech měsíce zjistil počty prodaných brýlí uvedené v následující tabulce:

Den i	Poč. prodaných fialov. brýlí x_i	Poč. prodaných růžov. brýlí y_i	Rozdíl $d_i = x_i - y_i$	Pořadí $\|d_i\|$	Pořadí kladných d_i	Pořadí záporných d_i
1	48	35	13	10	10	
2	55	67	-12	8		8
3	85	54	31	14	14	
4	81	68	13	10	10	
5	35	22	13	10	10	
6	51	48	3	2	2	
7	45	35	10	7	7	
8	48	55	-7	4,5		4,5
9	57	58	-1	1		1
10	75	68	7	4,5	4,5	
11	85	93	-8	6		6
12	25	25	0			
13	77	62	15	12	12	
14	93	56	37	15	15	
15	48	42	6	3	3	
16	75	50	25	13	13	
Součet	×	×	×	×	$T^+ = 100,5$	$T^- = 19,5$

Testujeme H_0: Me(X) = Me(Y) proti opačné alternativě. Výpočty jsou provedeny v tabulce. Testové kritérium $T = 19,5 < T_{w.,0,025} = 25$ a proto H_0 zamítáme. V počtu prodaných brýlí jsou významné rozdíly.

Testy úrovně pro dva nezávislé výběry:

Testujeme hypotézu H_0: Me(X) = Me(Y) proti alternativní hypotéze H_1: Me(X) ≠ Me(Y)

Mannův-Whitneyův test (MWT):

- Smícháme výběry a k jednotlivým měřením určíme pořadí.
- Označme R_1 součet pořadí výběru menšího rozsahu (X) a R_2 součet pořadí výběru většího rozsahu (Y), tj.rozsahy výběrů označíme $n \leq m$.
- Testové kritérium: $T = \min(T_1, T_2)$, kde
  - platí $T_1 + T_2 = mn$.
- Kritický obor: $W = \{T: T < k_{\alpha/2}\}$, kde $k_{\alpha/2}$ je kritická hodnota.
  - Viz tabulka VII. v dodatku.
  - Pravostranná alternativa H_1:Me(X) > Me(Y), $W = \{ T_1: T_1 < k_\alpha\}$
  - Levostranná alternativa H_1:Me(X) < Me(Y), $W = \{ T_2: T2 < k_\alpha\}$

Určitou modifikací MWT je dvouvýběrový Wilcoxonův test (DWT), který používá R.

- Jeho testovací statistika je $W = R_1 - \frac{n(n+1)}{2}$.
- Při H_0 má W rozdělení W(n,m), jehož kvantily jsou tabelované (a počítá je i R pomocí přikazu *qwilcox(p,n,m)*).
- Asymptotická verze DWT používá testové kritérium $Z = \frac{W\sqrt{12}}{\sqrt{nm(n+m+1)}}$,

  které má při H_0 rozdělení N(0,1).

Příklad:

- Mezinárodní korporace plánuje otevřít svoji pobočku v Řecku. Zabezpečení jejího provozu bude vyžadovat, aby se do Řecka přestěhoval větší počet pracovníků. Vedení korporace se rozhodlo nabídnout pracovníkům, kteří přicházejí do úvahy, intenzívní program výuky řečtiny. Při předcházejícím kurzu italštiny využili program poskytnutý firmou Lingua. Podle názoru jednoho z ředitelů efektivnější výuku cizích jazyků poskytuje společnost Trend. Proto se rozhodli otestovat nulovou hypotézu, že oba programy jsou stejně efektivní proti alternativě, že studenti, kteří absolvovali program poskytovaný

společností Trend, dosahují lepší výsledky. Náhodným výběrem vybrali výsledky závěrečných testů 14 studentů, kteří absolvovali program společnosti Trend a 15 studentů, kteří absolvovali program společnosti Lingua. Závěrečný test byl v obou skupinách stejný a jeho výsledky jsou následující:

- Trend x_i 85 87 92 98 90 88 75 72 60 93 88 89 62 73 (bodů).
- Lingua y_i 65 57 74 43 39 88 62 69 70 72 59 60 80 83 50 (bodů).
- Potvrzují výsledky výběrového šetření tvrzení jednoho z ředitelů, že program společnosti Trend je lepší než program společnosti Lingua?

Řešení:

- Jedná se o nezávislé náhodné výběry, proto použijeme MWT. Nejdříve uspořádáme výsledky závěrečného testu vzestupně podle velikostí a přiřadíme jim pořadová čísla.
- Trend 60 62 72 73 75 85 87 88 89 90 92 93 98 88.
- Lingua 39 43 50 57 59 60 62 65 69 70 72 74 80 83 88.
- Protože hodnota 60 je v pořadí na 6. a 7. místě, její pořadové číslo bude průměr z těchto dvou pořadí (6+7)/2 = 6,5. Podobně pořadí hodnoty 62 je (8+9)/2 = 8,5, pořadí hodnoty 72 je (13+14)/2 = 13,5 a pořadí hodnoty 88 je (22+23+24)/3 = 23. Pořadí hodnot v obou skupinách jsou následovné:
- Trend 6,5 8,5 13,5 15 17 20 21 23 23 25 26 27 28 29.
- Lingua 1 2 3 4 5 6,5 8,5 10 11 12 13,5 16 18 19 23.
- Součet pořadí pro společnost Trend je $R_1 = 282,5$ a pro společnost Lingua $R_2 = 152,5$.
- Dále vypočítáme:

$$T_1 = mn + \frac{m(m+1)}{2} - R_1 = 14.15 + 14.15/2 - 282,5 = 32,5,$$

$$T_2 = mn + \frac{n(n+1)}{2} - R_2 = 14.15 + 15.16/2 - 152,5 = 177,5.$$

- Testujeme hypotézu H_0: Me(X) = Me(Y) proti pravostranné alternativě H_1: Me(X) > Me(Y), použijeme testové kritérium $T = T_1 = 32,5$. V tab. VII v. dodatku najdeme pro $\alpha = 0,05$ kritickou hodnotu $k_\alpha = k_{0,05} = 67$. Protože $T_1 = 32,5 < k_{0,05} = 67$, zamítáme na 5% hladině významnosti hypotézu H_0 proti pravostranné alternativě H_1. Znamená to, tvrzení jednoho z ředitelů korporace je správné.

- Při použití DWT počítáme testové kritérium
- Horní kvantil Wilcoxonova dvojvýběrového rozdělení $w_{0,95}(14,15) = 143$ (v R použijeme příkaz *qwilcox(0.95,14,15)*). Testovací statistika ho překročí, proto H_0 zamítáme a přijímáme pravostrannou alternativu.

$$W = R_1 - \frac{m(m+1)}{2} = 282{,}5 - \frac{14.15}{2} = 177{,}5.$$

- Asyptotická verze DWT používá testové kritérium

$$Z = \frac{W\sqrt{12}}{\sqrt{mn(m+n+1)}} = 7{,}747,$$

- které překročí hodnotu $u_{0,95} = 1{,}645$, tj. dostáváme stejný závěr.
- Výpočet pomocí DWT nabízí R. Výstup z počítače je:

```
> wilcox.test(test ~ společnost, alternative="greater", data=jazyk)
   Wilcoxon rank sum test with continuity correction

data:  test by společnost
W = 177.5, p-value = 0.0008303
alternative hypothesis: true location shift is greater than 0
> qwilcox(0.95,14,15)
[1] 143
```

Kolmogorovův-Smirnovův dvouvýběrový test (KSDT) – test shody rozdělení

- Testujeme hypotézu H_0: dva výběry $x_1,\ldots,x_m$ a $y_1,\ldots y_n$ pocházejí ze stejných rozdělení proti opačné alternativě.
- Srovnáme všechna měření do neklesající posloupnosti $z_1,\ldots,z_{m+n}$.
- Z výběrů vypočítáme empirické distribuční funkce $F_m(z)$ a $G_n(z)$ – tj. kumulové relativní četnosti jednotlivých výběrů.
- Testové kritérium je $D = \sup_z |F_m(z) - G_n(z)|$

- Kritický obor: $W = \{D: D \geq d_{1-\alpha}\}$, kde $d_{1-\alpha}$ jsou kvantily KSDT (viz tabulka VIII. v dodatku).

Příklad:

Bylo vybráno 13 polí stejné kvality. Na 5 z nich se zkoušel nový způsob hnojení, zbývajících 8 bylo ošetřeno běžným způsobem. Výnosy pšenice uvedené v tunách na hektar jsou označeny x_i u nového a y_i u běžného způsobu hnojení.

x_i: 5,0 4,5 4,2 5,4 4,4

y_i: 5,7 5,5 4,3 5,9 5,2 5,6 5,8 5,1

Testujte hypotézu H_0: oba výběry pocházejí ze stejného rozdělení proti opačné alternativě H_1

Řešení.

Použijeme KST pro dva výběry. Potřebné výpočty jsou provedeny v následující tabulce. Hodnota testového kritéria je $D = \sup_x |F_n(x) - G_m(x)|$ = 0,675. V tab. VIII. v dodatku najdeme pro n = 5, m = 8, α = 0,05 odpovídající kvantil $d_{0,95}$ = 0,75. Protože D < $d_{0,95}$, nezamítáme hypotézu H_0, že oba výběry pocházejí ze základních souborů se stejnými distribučními funkcemi.

Dvouvýběrové testy je možno používat jako prostředek analýzy závislostí numerické proměnné na alternativní.

Výnosy z_i	Četnost x_i	Četnost y_i	Kumulovaná četnost x_i	Kumulovaná četnost y_i	$F_n(z_i)$	$G_m(z_i)$	$\lvert F_n(z_i)-G_m(z_i)\rvert$
4,2	1	0	1	0	0,2	0	0,2
4,3	0	1	1	1	0,2	0,125	0,075
4,4	1	0	2	1	0,4	0,125	0,275
4,5	1	0	3	1	0,6	0,125	0,475
5,0	1	0	4	1	0,8	0,125	0,675
5,1	0	1	4	2	0,8	0,25	0,55
5,2	0	1	4	3	0,8	0,375	0,425
5,4	1	0	5	3	1	0,375	0,625

5,5	0	1	5	4	1	0,5	0,5
5,6	0	1	5	5	1	0,625	0,375
5,7	0	1	5	6	1	0,75	0,25
5,8	0	1	5	7	1	0,875	0,125
5,9	0	1	5	8	1	1	0
Součet	5	8	×	×	×	×	×

Řešení v R:

Načteme data do souboru vynosy.dat. Použijeme příkaz: *ks.test(vynosy$x,vynosy$y)*. Výstup:

```
> ks.test(vynosy$x,vynosy$y)

        Two-sample Kolmogorov-Smirnov test

data:  vynosy$x and vynosy$y
D = 0.675, p-value = 0.07925
alternative hypothesis: two-sided
```

Dvouvýběrové testy v R a v Excelu

Pro nezávislé výběry:

1) Paramerické testy:
   t.test(x, y, mu= , var.equal=T)
   t.test(x, y, mu= , var.equal=F)
   t.test(y~factor, mu= , var.equal=T)
   t.test(y~factor, mu= , var.equal=F)

   var.test(x, y, ratio=)
   var.test(y~factor, mu=)

2) Neparametrické testy

wilcox.test(x, y, mu=)
wilcox.test(y~factor, mu=)

ks.test(x, y)

Pro závislé výběry

3) Parametrický test:
   t.test(x, y, mu= , paired=T)

4) Neparametrický test:
   wilcox.test(x, y, mu= , paired=T)

Dvouvýběrový asyptotický test o poměrech (relativních četnostech):

Po aktivaci balíku *vsePackage* lze provádět příkazem

prop.diff.test(x, n, diff= , alternative=)

Viz Otázky a úkoly č. 5.

To jsou příkazy, pomocí kterých je možno jednotlivé testy vykonávat. Většina dvouvýběrových testů v R je možno provádět interaktivně přímo z nabídek.

Dvouvýběrové testy v Excelu: Excel nabízí v Analýze dat všechny dvouvýběrové parametrické testy úrovně. Neparametrické testy neuvádí.

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 144-150. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE, 1999. S. 115-119, 122-125, 128-129. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 159-178, 171-173. ISBN 80-245-0178-3.

BÍNA V. a kol. *Jak na jazyk R.* J. Hradec: FM VŠE, 2006.

BLATNÁ, D. *Neparametrické metody. Testy založené na pořádkových a pořadových statistikách.* Praha: VŠE, 1996. S. 94-98, 102-117. ISBN 80-7079-607-3.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 268-274, 283-286, 289-290. ISBN 80-7261-003-1.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 73-79. ISBN 80-7169-255-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 140-154, 167-168. ISBN 978-80-86446-40-5.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 138-147. ISBN 80-7187-058-7.

STUCHLÝ, J. *Statistické metody pro manažerské rozhodování*. J. Hradec: VŠE, 2004. S. 37-43, 53, 57-58, 60. ISBN 80-245-0153-8.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/)

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 223-226. ISBN 80-7169-089-9.

? Otázky a úkoly

1) Použijeme data ze souboru studenti.dat. Pomocí vhodného testu zjistěte, zda výška studentů závisí na jejich pohlaví.

2) Použijeme data ze souboru studenti.dat. Pomocí vhodného testu ověřte, zda ženy jsou v průměru o 20 kg lehčí než muži.

3) U 10 dvojčat byla zjištěna následující porodní váha (v gramech)

starší	2440	3500	2820	2540	2650	2690	2750	2750	2650	2200
mla-dší	2700	3080	2200	2700	2550	2350	3500	2500	2420	2520

Pomocí vhodného testu zjistěte, zda porodní váha u staršího z dvojčat je vyšší než porodní váha mladšího z dvojčat.

4) Použijeme data ze souboru studenti.dat. Pomocí vhodného testu zjistěte, zda typický rozdíl výšky a váhy studentů studujících na VŠE je 90.

5) V souvislosti s kontrolováním své osobní váhy získaly v posledních letech na popularitě dietní nápoje. Inzerenti těchto nápojů se domnívají, že muži dávají přednost nedietním nápojům mnohem častěji než ženy. K ověření této domněnky byl vybrán náhodný výběr n = 300 mužů, kteří pijí kolu, a bylo zjištěno, že 192 z nich pije obyčejnou kolu a zbývajících 108 dietní kolu. V obdobném souboru 300 žen pije 144 obyčejnou kolu a 156 dietní kolu. Ověřte předpoklad inzerentů na hladině významnosti 0, 05.

6) Použijeme data ze souboru studenti.dat. Pomocí vhodného testu zjistěte, zda bodové rozložení ve statistickém testu je stejné pro angličtináře i neangličtináře.

? Úkoly k zamyšlení a diskuzi

1) Diskutujte o podmínkách používání jednotlivých testů.

2) Zamyslete se nad tím, jak používat jednotlivé testy v manažerské praxi.

Klíč k řešení otázek:

1) Nejdříve dvouvýběrovým F-testem (*Statistics – Variance - Two-variances F-test*) zjistíme, zda rozptyly výšek u mužů a žen se liší, tj. testujeme H_0: $\sigma_1^2 = \sigma_2^2$ proti opačné alternativě H_1.

   Výstup:

```
> tapply(studenti$vyska, studenti$pohlavi, var, na.rm=TRUE)
       M        Z
36.39572 44.18519
>var.test(vyska~pohlavi,alternative='two.sided',conf.level=.95,data=studenti)
       F test to compare two variances
data: vyska by pohlavi
F = 0.8237, num df = 33, denom df = 27, p-value = 0.5908
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3902931 1.6919592
sample estimates:ratio of variances    0.8237087
```

   Závěr: Nezamítáme H_0, tj. výstup ukazuje, že rozptyly výšek jsou stejné.

   K testování shody průměrných výšek použijeme dvojvýběrový t-test (*Statistics – Means - Independent samples t-test*) a zaškrtneme *vyska* a *Assume equal variances?: Yes*. Testujeme H_0: $\mu_1 = \mu_2$ proti opačné alternativě H_1.

   Výstup:

```
>t.test(vyska~pohlavi,alternative='two.sided',conf.level=.95,var.equal=TRUE,
data=studenti)
       Two Sample t-test
data: vyska by pohlavi
t = 8.8125, df = 60, p-value = 2.044e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.98137 17.43039
sample estimates:
mean in group M mean in group Z
       182.7059        168.5000´
```

Závěr: H_0 zamítáme na 5% hladině významnosti, tj. výstup ukazuje, že průměrné výšky se liší a tedy výška studentů závisí na pohlaví. (Normalitu výšek v obou souborech prokážene SWT – viz úkol 1 v předchozí kapitole.) Stejné výsledky dostaneme i pomocí Analýzy dat v Excelu.

2) SWT lze ověřit, že váhy v souboru mužů i v souboru žen se neřídí normálním rozdělením. Proto použijeme Wilcoxonův dvouvýběrový test v *Statistics – Nonparametric tests – Two-sample Wilcoxon test*. Po jeho provedení doplníme ještě do příkazu mu=20. Testujeme H_0: Me(vaha mužů) - Me(vaha žen) = 20 proti opačné alternativě H_1.
Výstup:

```
wilcox.test(vaha ~ pohlavi, mu=20, alternative="two.sided", data=studenti)
      Wilcoxon rank sum test with continuity correction
data:  vaha by pohlavi
W = 602, p-value = 0.07542
alternative hypothesis: true location shift is not equal to 20
```

Závěr: Test těsně nulovou hypotézu nezamítnul. Na 5% hladině významnosti jsme neprokázali, že typický rozdíl vah studentů a studentek se liší od 20 kg. Stejné výsledky dostaneme i pomocí Analýzy dat v Excelu.

3) Vložíme data do souboru *dvojcata* proměnných *starsi* a *mladší* pomocí editoru. Testujeme nejdříve normalitu dat pomocí SWT. Výstup:

```
> shapiro.test(dvojcata$mladsi)
      Shapiro-Wilk normality test
data:  dvojcata$mladsi
W = 0.8578, p-value = 0.0719

> shapiro.test(dvojcata$starsi)
      Shapiro-Wilk normality test
data:  dvojcata$starsi
W = 0.8787, p-value = 0.1262
```

Závěr: Oba soubory se řídí normálním rozdělením a jsou závislé. Použijeme proto dvouvýběrový párový t-test (*Statistics – Means – Paired t-test*). Testujeme hypotézu H_0: $\mu_{st} - \mu_{ml} = 0$ proti H_1: $\mu_{st} - \mu_{ml} > 0$

Výstup:

```
>
 t.test(dvojcata$mladsi, dvojcata$starsi, alternative='greater',
+   conf.level=.95, paired=TRUE)
       Paired t-test
data:  dvojcata$mladsi and dvojcata$starsi
t = 0.3596, df = 9, p-value = 0.3637
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -192.6179       Inf
sample estimates:
mean of the differences
                     47
```

Závěr: H_0 nezamítáme. Znamená to, že jsme na 5% hladině významnosti neprokázali, že starší dvojče má těžší porodní váhu než dvojče mladší. Stejné výsledky dostaneme i pomocí Analýzy dat v Excelu.

4) Nejdříve ověříme SWT, že uvažovaný rozdíl se neřídí normálním rozdělením: použijeme příkaz *shapiro.test(studenti$vyska-studenti$vaha).*

   Proto použijeme Wilcoxonův párový test (*Statistics - Nonparametric tests – Paired-samples Wilcoxon test*) a do příkazu ještě dopíšeme *mu=90*. Testujeme H_0: Me(vyska) – Me(vaha) = 90 proti opačné H_1.

   Výstup:

```
> wilcox.test(studenti$vyska, studenti$vaha, mu=90, alternative='two.sided',
paired=TRUE)
       Wilcoxon signed rank test with continuity correction
data:  studenti$vyska and studenti$vaha
V = 1953, p-value = 7.603e-12
alternative hypothesis: true location shift is not equal to 90
```

   Závěr: Na 5% hladině významnosti zamítáme H_0. Výstup ukazuje, že typický rozdíl není 90.

Testujeme H_0 π_m-π_z = 0 proti alternativě H_1: π_m-π_z > 0. Použijeme asymptotický dvouvýběrový test o poměrech. Řešení získáme v R Commanderu po aktivaci balíku *vsePackage* příkazem:

prop.diff.test(x=c(144,192), n=c(300,300),diff=0,alternative="greater")

Výstup:

```
> library(vsePackage)
> prop.diff.test(x=c(144,192), n=c(300,300),diff=0,alternative="greater" )
Difference of proportions test based on asymptotical normality
Data:  c(144, 192)
Alternative hypothesis: true pi(1) - pi(0) is greater than 0
    Success = x
Z = 4,   p-value = 3.167124e-05
Sample estimates of proportions of successes:
   0    1
0.48 0.64
Estimate of the difference of proportions:  0.16
Estimated SE of the estimate:               0.04
95% confidence interval for the difference of proportions:
    (0.09420585,  1)
> qnorm(c(0.95), mean=0, sd=1, lower.tail=TRUE)
[1] 1.644854
```

Závěr: Zamítáme H_0, tj. potvrdila se domněnka inzerentů, že muži vypijí více nedietních nápojů než ženy.

5) Pomocí příkazů:

skup1 <- subset(data, data$jazyk=="A")
skup2 <- subset(data, data$jazyk!="A")

Rozdělíme studenty na skupinu1 angličtinářů a skupinu2 neangličtinářů. K testování shody rozdělení použijeme Kolmogorovův-Smirnovův dvouvýběrový test. Testujeme H_0: bodové rozdělení ve statistickém testu je stejné v skupině angličtinářů jako ve skupině neangličtinářů proti opačné alternativě. Aktivujeme balík vsePackage příkazem *library(vsePackage)*. Samotný test provedeme příkazem: *ks.test(skup1$test, skup2$test).*

Výstup:

```
> ks.test(skup1$test, skup2$test)
        Two-sample Kolmogorov-Smirnov test
data:  skup1$test and skup2$test
D = 0.224, p-value = 0.523
alternative hypothesis: two-sided
```

Závěr: Nezamítáme H_0, tj. test neprokázal významný rozdíl rozdělení bodového hodnocení ve statistickém testu mezi skupinou angličtinářů a neangličtinářů.

Kapitola 8: Další testy a analýza rozptylu

Klíčové pojmy:

chi-kvadrát testy, testy dobré shody, testování shody empirických a teoretických četností, testování nezávislosti v kontingenční tabulce, kontingenční koeficienty, Kolmogorovův-Smirnovův jednovýběrový test, jednofaktorová analýza rozptylu, celková, meziskupinová a vnitroskupinová suma čtverců, poměr determinace, homoskedasticita, Bartletův test, tabulka analýzy rozptylu, Kruskalův-Wallisův test

Cíle kapitoly:

- **pochopení principu chi-kvadrát testů;**
- **provádění dalších neparametrických testů;**
- **princip a používání analýzy rozptylu.**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly.

Úvod

Chi-kvadrát testy

- Testování shody empirického rozdělení s rozdělením teoretickým
- Testování nezávislosti v kontingenční tabulce

Kolmogorovův-Smirnovův test pro jeden výběr

Příkazy pro další testy v R

Jednofaktorová analýza rozptylu

Kruskalův-Wallisův test

Struktura výkladu

Život je umění vytvářet uspokojivé závěry na základě nedostatečných předpokladů
Samuel Butler

Úvod

V dosud probraných parametrických testech vycházíme vždy ze znalosti rozdělení základního souboru, z kterého jsme pořídily náhodný výběr (např. předpoklad normality).

Musíme proto umět tento předpoklad ověřit:

- Testujeme shodu mezi předpokládaným rozdělením a rozdělením empirickým.

Používáme k tomu testy dobré shody.

- Např. shodu empirického rozdělení s normálním rozdělením ověřujeme SWT.

Patří do velmi početné skupiny neparametrických testů.

Předpoklady na použití neparametrických testů jsou menší.

- Neparametrické testy jsou robusnější, tj. kvalita výsledků je méně závislá na povaze konkrétních dat a na narušení předpokladů kladených na tato data.

Síla neparametrických testů je obvykle slabší, tj. dochází častěji k chybnému nezamítnutí nepravdivé nulové hypotézy.

Chi-kvadrát testy

Používáme je nejčastěji jako testy dobré shody

- Mají širší použití:
    - Testy nezávislosti dvou znaků (v kontingenční tabulce).
    - Testy homogenity (shody) rozdělení 2 výběrových souborů.
    - Testy o shodě 2 nebo více populačních poměrů.

a) Chi-kvadrát test dobré shody

- Ověřujeme jím předpoklad, že rozdělení základního souboru, z něhož byl výběr pořízen, je určitého konkrétního typu.
- Testujeme hypotézu H_0, že náhodný výběr
    - pochází z předpokládaného rozdělení (normálního, Poissonova aj.), které má $r \geq 1$ neznámých parametrů (tzv. neúplně specifikovaný model), popř. toto rozdělení je určeno i s parametry (tzv. úplně specifikovaný model),
    - nebo tvoří určité intuitivně formulované teoretické rozdělení (viz příklad 2),

  proti opačné alternativě H_1.

Postup chi-kvadrát testu dobré shody:

- Náhodný výběr o rozsahu n roztřídíme do k tříd.
- Označme n_i absolutní empirické četnosti těchto tříd.
- Při splnění H_0 je znám tvar rozdělení sledovaného znaku X.
- Odhadneme parametry tohoto rozdělení.
- Potom určíme pravděpodobnosti jednotlivých tříd π_i pomocí odhadů
- $p_i = P(X \in$ i-té třídy).
- Z nich určíme teoretické (očekávané) četnosti $n_i' = np_i$.
- Vypočteme testové kritérium: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$.

- Při platnosti H_0 má testové kritérium χ^2 rozdělení χ^2(k-r-1), kde k je počet tříd a r je počet odhadovaných parametrů.
- Kritický obor testu je $W=\{\chi^2 : \chi^2 > \chi^2_{1-\alpha}(k-r-1)\}$.
- Pro korektní použití testu je požadováno splnění podmínek:

  n je dostatečně velké ($n \geq 50$) a $np_i = n_i' > 5$ pro $i = 1,\ldots,k$.
- Není-li splněna 2. podmínka, je potřebné některé třídy spojit.
- Test používáme nejčastěji na testování hypotéz:

  H_0: $X \sim N(\mu,\sigma^2)$, kde odhadneme $\hat{\mu}=\bar{x},\hat{\sigma}^2=s_x^2$;

  H_0: $X \sim Po(\lambda)$, kde odhadneme $\hat{\lambda}=\bar{x}$;

  H_0: empirické rozdělení četností je shodné se zadaným rozdělením.
- Chi-kvadrát test je rovnocenný s u-testem o shodě poměrů.

Příklad 1: Viz Stuchlý (1999a), s. 131-133.

Příklad 2: Třetí sloupec tabulky udává strukturu korunových úvěrů klientů ČR v roce 1995 v členění podle účelu. Jeden z bankovních ústavů poskytující korunové úvěry potřebuje operativně znát, zda i v jeho klientele je rozložení shodné s celostátní strukturou.

Provedl náhodný výběr 253 úvěrových smluv a ověřuje nulovou hypotézu o shodě. Údaje i potřebné propočty jsou v tabulce. Počet tříd k = 6. Testové kritérium $\chi 2 = 9{,}476$. Kritický obor je omezen zdola kvantilem $\chi^2_{0,95}(5) = 11{,}1$. Protože $\chi 2 < 11{,}1$, nezamítáme na 5% hladině významnosti nulovou hypotézu. Test neprokázal rozdíly v struktuře úvěrů.

Korunové úvěry klientů podle účelu	Počet úvěrů n_i	Celostátní úvěrová struktura p_i	$n_i' = n\,p_i$	$\frac{(n_i - n_i')^2}{n_i'}$
- provozní	92	35,6% = 0,356	90,068	0,041
- investiční	63	26,2% = 0,262	66,286	0,163
- hypotekární	4	0,4% = 0,004	1,012	8,822
- privatizační	11	3,9% = 0,039	9,867	0,130
- na přechodný nedostatek zdrojů	24	9,0% = 0,090	22,770	0,066
- ostatní	59	24,9% = 0,249	61,997	0,254
Celkem	253	100% = 1,000	253,000	9,476

b) Test nezávislosti dvou znaků

Provedeme dvoustupňové třídění do kontingenční tabulky:

A \ B	**B_1**	**B_2**	. . . . .	**B_s**	**Součet**
A_1	n_{11}	n_{12}	. . . . .	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	. . . . .	n_{2s}	$n_{2.}$
. . .	. . . . .				. . .
A_r	n_{r1}	n_{r2}	. . . . .	n_{rs}	$n_{r.}$
Součet	$n_{.1}$	$n_{.2}$	. . . . .	$n_{.s}$	n

přičemž znak X třídíme do r skupin $A_1,\ldots,A_r$ a znak Y do s skupin $B_1,\ldots,B_s$. Tabulka obsahuje absolutní sdružené četnosti n_{ij} a součtové (marginální) četnosti $n_{i.}$ (součty řádků), i = 1,…,r a $n_{.j}$ (součty sloupců), j = 1,…,s.

Testujeme nulovou hypotézu H_0: Znaky X, Y jsou nezávislé proti opačné alternativě H_1.

Testové kritérium $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

při dostatečně velkém rozsahu souboru $n = \sum_i \sum_j n_{ij}$

a při dostatečně velkých očekávaných četnostech $n_{ij}' = n\frac{n_{i.}}{n}\frac{n_{.j}}{n}$

(požadujeme, aby $n_{ij}' \geq 1$) má při H_0 rozdělení $\chi^2((r-1)(s-1))$.

Odtud dostaneme kritický obor testu $$W = \{\chi^2 : \chi^2 > \chi^2_{1-\alpha}((r-1)(s-1))\}.$$

Podobně testujeme homogenitu výběrů (tj. předpoklad, že výběry pocházejí ze stejného rozdělení pravdepodobnosti).

Míry síly závislosti kvalitativních proměnných:

Pearsonův kontingenční koeficient $$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

- Při nezávislosti znaků je C = 0, velké C znamená silnou závislost.

Cramérův kontingenční koeficient $V = \sqrt{\frac{\chi^2}{n(m-1)}} =$

- Zde m = min(r,s), r je počet řádků, s je počet sloupců kontingenční tabulky.
- Platí 0 <V < 1, V blízké 0 ukazuje na slabou a V blízké 1 na silnou závislost.

Příklad 3: – Viz Stuchlý (1999), s. 135.

Testy o shodě dvou nebo více populačních poměrů

Test o shodě dvou poměrů: Testujeme hypotézu H_0: $\pi_1 = \pi_2$ (dvě kategorie jsou v populaci stejně zastoupeny) proti alternativě H_1: $\pi_1 \neq \pi_2$. K testování lze použít jednovýběrový test o poměru nebo chi-kvadrát test (viz Pecáková 2008, str.123).

V případě dostatečné velikosti parametrů četností n_1, n_2 lze použít testové kriterium

$$U = \frac{n_1 - n_2}{\sqrt{n_1 + n_2}},$$

které má při platnosti H_0 rozdělení $N(\mu; \sigma^2)$. Pro dvoustrannou alternativu lze použít testové kriterium $\chi^2 = \frac{(n_1 - n_2)^2}{n_1 + n_2}$,

které má při H_0 přibližně chí-kvadrát rozdělení s jedním stupněm volnosti.

Test o shodě více populačních poměrů: Testujeme hypotézu H_0: $\pi_1 = \pi_2 = \ldots = \pi_m = 1/m$ (m kategorií je v populaci stejně zastoupeno) proti opačné alternativě H_1. Použijeme statistiku

$$\chi^2 = \frac{\sum_{i=1}^{m} (n_i - \bar{n})^2}{\bar{n}}$$

kde $\bar{n} = \frac{\sum_{i=1}^{m} n_i}{m}$. Nulovou hypotézu zamítneme, když $\chi^2 > \chi^2_{1-\alpha}(m-1)$.

Příklad 4: Viz Pecáková (2008), str. 124 – 126.

Kolmogorovův-Smirnovův jednovýběrový test (KSJT)

Při ověřování dobré shody mezi empirickým a teoretickým rozdělením dáváme tomuto testu přednost před chi-kvadrát testem v případech výběru malého rozsahu. Má totiž větší sílu a

hlavně se vyhneme komplikacím, které při chi-kvadrát testu přinášejí omezující podmínky na rozsah tříd. V případě dokonce velmi malého rozsahu výběru je KSJT jedinou možností. Jeho další předností je, že vychází z původních dat a nikoliv z údajů setříděných do tříd. Tím nedochází ke ztrátě informace obsažené ve výběru.

Test používáme k ověření hypotézy H_0: pořízený výběr pochází z rozdělení se spojitou distribuční $F_0(x)$, která je úplně specifikována včetně všech parametrů proti opačné alternativě H_1.

Testovým kritériem D je největší zjištěná vzdálenost distribuční funkce $F_0(x)$ rozdělení, ze kterého náhodný výběr pochází, od výběrové (empirické) distribuční funkce $F_n(x)$, tj.

$$D = \sup_x |F_n(x) - F_0(x)| .$$

Uvažujeme-li náhodný výběr $x_1, x_2, ..., x_n$ z rozdělení spojitého typu, potom označme $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ je týž náhodný výběr uspořádaný vzestupně podle velikostí. Funkce

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{i}{n}, & x_{(i)} \le x < x_{(i+1)}, \quad i = 1,2,\ldots,n-1, \\ 1, & x \ge x_{(n)} \end{cases}$$

se nazývá empirickou distribuční funkcí.

V případě, že jsou data setříděna do tabulky rozdělení četností, představuje empirická distribuční funkce kumulované relativní četnosti.

Grafem nespojité empirické funkce $F_n(x)$ je schodovité funkce. Maximální absolutní odchylka D může být nalezena buď jako vzdálenost křivky $F_0(x)$ od vrcholu schodu $|F_n(x_i)-F_0(x_i)|$, nebo naopak jako vzdálenost křivky $F_0(x)$ od sedla schodu $|F_n(x_{i-1})-F_0(x_i)|$. Největší z nich je pak hodnotou testového kritéria D.

Závěrečné vyhodnocení testu: V dodatku v tabulce IX. jsou kvantily $d_{n;1-\alpha}$ pro KSJT (pro hladiny významnosti α = 0,10; 0,05; 0,01 a pro daný rozsah výběru n), které tvoří kritické hodnoty testu. Nulovou hypotézu zamítáme, je-li $D \ge d_{n;1-\alpha/2}$.

Příklad 5: Viz Hindls a kol. (2007), s. 158-159.

Příkazy pro další testy v R

Testy shody v R:

chisq.test (x=c( ), p=c( )) , x jsou empirické četnosti a p očekávané relativní četnosti;
ks.test(x, "pnorm", mean(), sd()), lze použít i na testování jiných rozdělení.
Chi-kvadrát test lze provést v R Commanderu v nabídce *Statistics – Summaries – Frequency Distributions*, když ve vstupním okně zaškrtneme položku *Chi-square goodnes- of-fit test*.

Vytváření kontingenční tabulky:

table(x,y), tabulka sdružených četností proměnných x, y;
addmargins(table(x,y)), tabulka sdružených a marginálních četností;
prop.table(table(x.y)), tabulka sdružených relativních četností;
tab -> matrix(c(), r, s, byrow=T[F]), zadání tabulky maticí typu r x s po řádcích [sloupcích].

Testování nezávislosti v kontingenční tabulce:

chisq.test (tab), testování nezávislosti v kontingenční tabulce;
pearson.indep.test(tab), testování nezávislosti v kontingenční tabulce po aktivaci balíku *vse-Package*, počítá i koeficienty kontingence (tabulku lze zadat v nabídce *Statistics – Contingency tables – Enter and analyse two-way table*).

Analýza rozptylu (AR)

Úvod

Analýza rozptylu zkoumá, zda číselná veličina Y (odezvová veličina) závisí na kategoriálních (kvalitativních) proměnných X_i (faktory).

Rozhodnutí se provádí na základě rozkladu rozptylu, resp. odpovídajícího součtu čtverců.

AR byla zavedena R. A. Fisherem (v r. 1912) k sledování vlivů různých úrovní určitého faktoru na úrodu uvažované plodiny.

Anglický název: Analysis of variance (ANOVA).

Jednofaktorová analýza rozptylu (JAR)

Sledovaná statistická veličina Y je ovlivňována jen jedním faktorem X uvažovaným na k úrovních.

- Např. závislost úrody na hnojivu, tržby a prodavači, hodinové mzdy na kvalifikační třídě, investic na vzdělání respondenta.

Podle úrovní daného faktoru X jsou pozorování znaku Y rozdělena do k skupin o n_i pozorováních, $\sum n_i = n$:

Skupina	Hodnoty znaku y	Průměry skupin
1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$\bar{y}_1$
2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$\bar{y}_2$
. . .	. . . . .	. . . .
k	$y_{k1}, y_{k2}, \ldots, y_{kn_k}$	$\bar{y}_k$

Základní myšlenka JAR:

Rozklad rozptylu veličiny Y:

- na meziskupinový a vnitroskupinový.
- Místo rozptylu používáme v AR jen příslušné součty čtverců.

Součet čtvercových odchylek n hodnot veličiny Y od jejich průměru (celkový součet S_y) rozkládáme na součet meziskupinový $S_{y,m}$ (rozptyl skupinových průměrů) a vnitroskupinový (reziduální = zbytkový) $S_{y,v}$, tj.

$S_y = S_{y,m} + S_{y,v}$,

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}^2, \text{ kde } \bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

$$S_{y,m} = \sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 n_i = \sum_{i=1}^{k}\bar{y}_i^2 n_i - n\bar{y}^2,$$

$$S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{\ 2} - \sum_{i=1}^{k}\bar{y}_i^2 n_i$$

Interpretace těchto součtů:

celkový součet čtverců S_y charakterizuje celkovou měnlivost (variabilitu) hodnot sledovaného znaku Y;

meziskupinový součet čtverců $S_{y,m}$ - měnlivost mezi skupinami (vliv jednotlivých úrovní faktoru x);

vnitroskupinový součet čtverců $S_{y,v}$ - měnlivost v skupinách (tj. nevysvětlená variabilita, způsobená náhodnými vlivy). Nazýváme ho také reziduální součet čtverců S_R.

- Mírou těsnosti (síly) závislosti Y na x je tzv. poměr determinace $P^2 = S_{y,m}/S_y$.
- Platí: $0 \leq P^2 \leq 1$. Čím silnější je závislost (čím větší podíl na celkové variabilitě má meziskupinová variabilita) tím více se P^2 blíží k 1 (samotné P nazýváme korelační poměr – představuje neobecnější míru síly závislosti).
    - Hodnota $P^2 = 0$ odpovídá rovnosti všech skupinových průměrů (nulové meziskupinové variabilitě) a $P^2 = 1$ nulové vnitroskupinové variabilitě.

K jednotlivým součtům čtverců můžeme definovat tzv. stupně volnosti.

- Počet stupňů volnosti součtu čtverců m veličin je určen tím, kolik z těchto veličin je nezávislých. Existuje-li mezi m veličinami c lineárních vztahů, má součet čtverců těchto m veličin m – c stupňů volnosti.

Lze ukázat, že S_y má $\nu = n - 1$, $S_{y,m}$ má $\nu_1 = k - 1$ a $S_{y,v}$ má $\nu_2 = n - k$ stupňů volnosti (platí $\nu = \nu_1 + \nu_2$).

Předpoklady použití ANOVA:

- Výběry ve skupinách musí být nezávislé a pochází ze základních souborů s rozdělením $N(\mu_i;\sigma_i^2)$, které mají stejné rozptyly, tj. platí, že $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, tzv. homoskedasticita.
- Homoskedaticitu ověřuje Bartlettův test (viz Seger a Hindels 1995, s. 162-163) nebo Levenův test v R a normality SWT nebo grafickými metodami. Aplikujeme je na rezidua (odhady chyb měření v modelu).

Test hypotézy o neexistenci vlivu faktoru (neboli o nezávislosti znaku Y na zkoumaném faktoru x) umožní zobecnit závěr o rozdílnosti či podobnosti skupinových průměrů na celou populaci.

Pomocí JAR testujeme nulovou hypotézu H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ proti opačné alternativě H_1

- Jiná interpretace JAR: H_0: odezvová veličina nezávisí na faktorech.

Použijeme testové kritérium $F = \frac{S_{y,m}/(k-1)}{S_{y,v}/(n-k)}$,

které má při H_0 rozdělení F(k-1;n-k).

Alternativní hypotéze jsou příznivé vysoké hodnoty F.

Proto H_0 zamítáme, když $F > F_{1-\alpha}(k-1;n-k)$, kde $F_{1-\alpha}(k-1;n-k)$ je kvantil příslušného F-rozdělení a interpretujeme to tak, že faktor x působí významně na odezvovou veličinu Y (resp. kvantitativní veličina Y závisí na hodnotách kvalitativní proměnné x).

Hodnoty náhodné veličiny (odezvy) Y lze vyjádřit ve tvaru $y_{ij}=\mu+\alpha_i+\varepsilon_{ij}$, $i=1,\ldots,k$, $j=1,\ldots,n_i$, (model analýzy rozptylu)

kde y_{ij} je j-té pozorování odezvy Y při i-té úrovni faktoru x, $\mu = E(Y)$,
α_i je efekt (vliv) i-té úrovně faktoru na odezvovou veličinu Y,
ε_{ij} jsou náhodné chyby.
μ odhadujeme výběrovým průměrem $\bar{y}$
α_i rozdílem skupinového a celkového průměru $\bar{y}_{i.} - \bar{y}$.
Hodnoty y_{ij} odhadujeme vyrovnanými hodnotami $\hat{y}_{ij}$ (v R *fittted.values(model)*),
chyby ε_{ij} odhadujeme rozdílem empirických (naměřených) a vyrovnaných hodnot $e_{ij} = y_{ij} - \hat{y}_{ij}$ (tzv. residua – v R *residuals(model)*).

Výpočet provádíme obvykle do následující tabulky ANOVA:

Zdroj měnlivosti	Součet čtverců	Stupně volnosti	Průměrný součet čtverců	Testové kritérium
Faktor	$S_{y,m} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	k - 1	$S_{y,m}/(k-1)$	$F = \frac{S_{y,m}/(k-1)}{S_{y,v}/(n-k)]}$
Rezidua	$S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	n - k	$S_{y,v}/(n-k)$	×
Celkový	$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	n - 1	×	×

Při zamítnutí H_0 provádíme vícenásobná porovnávání např. Tukeyovo metodou (zjišťujeme, které dvojice úrovní faktorů způsobily zamítnutí H_0) a můžeme počítat také intervaly spolehlivosti pro jednotlivé třídy – viz R.

Podobným způsobem můžeme provádět i vícefaktorovou analýzu rozptylu.

Poznámky:

Předpoklad normality rozdělení se ověřuje obvykle SWT na rezidua nebo některou z počítačových grafických metod. Odchylky skutečného rozdělení znaku Y od normálního rozdělení nemají velký vliv na rozdělení statistiky F, kromě případu výskytu extremních hodnot v jednotlivých výběrech.

Ověření homoskedasticity (shody rozptylů): Pokud nelze pro nedostatečně obsazené skupiny použít Bartlettův test, můžeme použít Levenův test nebo použijeme k ověření bodový diagram (závislost reziduí na úrovních faktorů) popř. odhadneme nesplnění zhruba posouzením hodnot výběrových rozptylů s_i^2.

Při nejistotě splnění předpokladů normality a homoskedesticity dat, můžeme místo AR použít **Kruskalův-Wallisův test** (KWT) – viz Stuchlý (2004), s. 44-46.

Analýza rozptylu v R:

1. ANOVA (v nabídce *Models* je podrobná analýza číselná i grafická v modelu):

model <- aov(y~x), uložení výsledků AR do objektu model (nebo interaktivně v *Statistics-Means-One-way ANOVA*);
factor(x, levels=c(), labels=c()), zadání a označení úrovní faktoru
fitted.values(model), výpis vyrovnaných hodnot z AR
TukeyHSD(model), provedení Tuckeyovo vícenásobného porovnávání

2. rezidua a grafy reziduí:

residials(model), výpis reziduí AR;
resplot(model, type="e-yhat", xterm= , lowess= F, hline=T), graf závislosti reziduí na vyrovnaných hodnotách;
resplot(model, type="e-x", xterm= , lowess= F, hline=T), graf závislosti reziduí na faktoru;

resplot(model, type="e-time", xterm= , lowess= F, hline=T), graf závislosti reziduí na čase.

3. Barttletův a Levenův test homoskedasticity:

bartlett.test(y ~ x);
levene.test(y ~ x,data =).

4. Kruskalův-Wallisův test:

kruskal.test(y ~ x, data =);
kruskal.test(y , x, data =).

Příklad 6 (Stuchlý 1999a): V následující tabulce jsou uvedeny měsíční tržby tří prodavačů v tis. Kč. Na hladině významnosti 0.05 testujte hypotézu o shodě průměrných měsíčních tržeb u všech tří prodavačů proti opačné alternativě. Odhadněte celkovou průměrnou tržbu a efekty jednotlivých prodavačů na průměrné tržbě. Intenzitu závislosti charakterizujte korelačním poměrem. Ověřte podmínky potřebné pro použití analýzy rozptylu.

Řešení. Použijeme výpočty v následující tabulce:

Prodavač číslo	Měsíční tržby y_{ij}	$\bar{y}_{i.}$	$\bar{y}_{i.}^2$	$\sum_{j=1}^{n} y_{ij}^2$
1	15 10 9 5 16	11	121	687
2	15 10 12 11 12	12	144	734
3	19 12 16 16 17	16	256	1306
Součet	×	39	521	2727

Testujeme nulovou hypotézu H_0: $\mu_1 = \mu_2 = \mu_3$ proti opačné alternativě H_1. Při ručním výpočtu je výhodné přepsat si zavedené sumy čtverců tak, jak je uvedeno dále

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_{i.} = \frac{1}{3}\,39 = 13,$$

$$S_{y,m} = n\sum_{i=1}^{m} \bar{y}_{i.}^2 - mn\bar{y}^2 = 5.521 - 3.5.13 = 2605 - 3.5.13^2 = 70,$$

$$S_{y,v} = \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2 - n\sum_{i=1}^{m} \bar{y}_{i.}^2 = 2727 - 5.521 = 122,$$

$$S_y = \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2 - mn\bar{y}^2 = 2727 - 3.5.13^2 = 192.$$

Protože platí

$$F = \frac{S_{y,m}/(m-1)}{S_{y,v}/[m(n-1)]} = \frac{70/2}{122/3.4} = \frac{35}{10{,}14} = 3{,}44 < F_{0,95}(2,12) = 3{,}89,$$

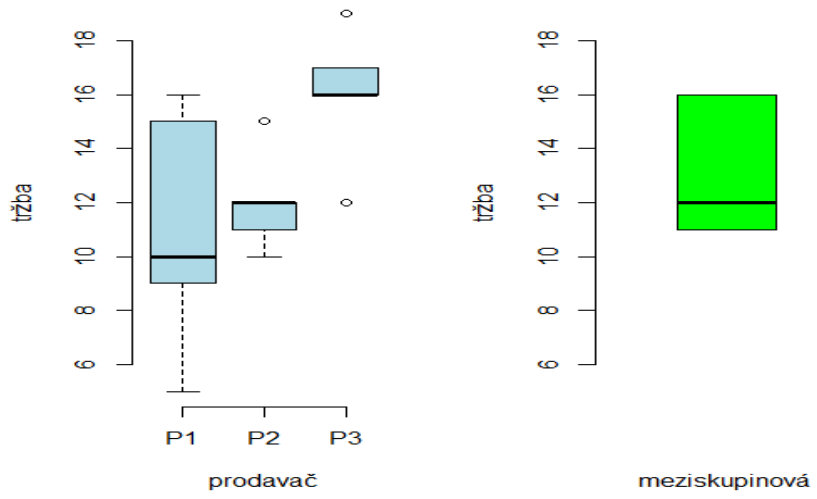
nezamítáme nulovou hypotézu H_0. Znamená to, že v tržbách jednotlivých prodavačů není statisticky významný rozdíl, tj. tržby nezávisí na faktoru prodavač.

Odhadneme ještě parametry modelu:

$\hat{\mu} = \bar{y}$ = 13, $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}$ = 11-13 = -2, $\hat{\alpha}_2 = \bar{y}_{2.} - \bar{y}$ = 12-13 = -1, $\hat{\alpha}_3 = \bar{y}_{3.} - \bar{y}$ =16-13 =3.

Celkový průměr tržeb je 13 tis. Kč a efekty, jakými se jednotlivý prodavači podílejí na celkovém průměru jsou - 2, -1 a 3 tis. Kč.

Určení meziskupinové sumy čtverců přibližuje její grafické znazornění v následujícím grafu. V levé části grafu jsou pomocí mediánů znázorněny skupinové průměrné tržby a v pravé části grafu je krabicovým diagramem znázorněna jejich variabilita, představující meziskupinovou variabilitu.



Úlohu je možno řešit na počítači např. pomocí programu R. Dostáváme:

```
> AnovaModel.1 <- aov(trzba ~ prodavac, data=trzby)

> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value Pr(>F)
prodavac     2     70   35.00   3.443 0.0658 .
Residuals   12    122   10.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>    numSummary(trzby$trzba    ,    groups=trzby$prodavac,    statis-
tics=c("mean", "sd"))
   mean       sd % data:n
P1   11 4.527693 0      5
P2   12 1.870829 0      5
P3   16 2.549510 0      5

> sqrt(70/(70+122))
[1] 0.6038074
```

Korelační poměr P = 0,604. Tedy na 5% hladině významnosti není závislost tržeb na prodavačích významná, ale intenzita závislosti je více jak středně silná (malá variabilita dat).

Načteme k datům rezidua a ověříme, zda jsou splněny podmínky pro použití analýzy rozptylu:

```
> trzby$residuals <- with(trzby, residuals(AnovaModel.1))

> shapiro.test(trzby$residuals)
     Shapiro-Wilk normality test
W = 0.9659, p-value = 0.7931

> bartlett.test(residuals ~ prodavac, data=trzby)
     Bartlett test of homogeneity of variances
Bartlett's K-squared = 2.9245, df = 2, p-value = 0.2317

> dwtest(trzba ~ prodavac, alternative="two.sided", data=trzby)
     Durbin-Watson test
DW = 2.2377, p-value = 0.9372
alternative hypothesis: true autocorrelation is not 0
```

Testy ukazují na to, že požadované podmínky jsou splněny.

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 151-165, 210-212. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE, 1999. S. 130-140, 142-145, 147-148. ISBN 80-7079-754-1.

Doporučené studijní zdroje:

ANDĚL, J. *Matematická statistika*. Praha: SNTL/ALFA, 1985. S. 147-157, 209-217, 231-2.

ARLTOVÁ, M. a kol. *Příklady k předmětu Statistika A*. Praha: VŠE, 2003. S. 178-185. ISBN 80-245-0178-3.

BLATNÁ, D. *Neparametrické metody. Testy založené na pořádkových a pořadových statistikách.* Praha: VŠE, 1996. S. 117-125. ISBN 80-7079-607-3.

CYHELSKÝ, L. a kol. *Elementární statistická analýza.* Praha: Management Press, 2001. S. 279-283. ISBN 80-7261-003-1.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 79-81, 102-105, 112-122. ISBN 80-7169-255-7.

HINDLS, R. a kol. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000. S. 14-19, 22-27, 37-43. ISBN 80-7261-013-9.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 155-165, 169-170, 181-192, 203-210. ISBN 978-80-86446-40.

MINAŘÍK, B. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita v Brně, 1995. S. 137-142. ISBN 80-7157-166-0.

PECÁKOVÁ, I. *Statistika v terénních průzkumech.* Praha: Professional Publishing, 2008. S. 116-128. ISBN 978-80-86946-74-0.

ŘEZANKOVÁ, H. a T. LÖSTER. *Úvod do statistiky*. Praha: Oeconomica, 2009. S. 50-56. ISBN 978-80-245-1514-4.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 147-163, 216-219. ISBN 80-7187-058-7.

STUCHLÝ, J. *Statistické metody pro manažerské rozhodování*. J. Hradec: VŠE, 2004. S. 44-47. ISBN 80-245-0153-8.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/)

WONNACOT, T.H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 352-364. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Použijte data ze souboru casopis.dat. Zjistěte, zda výběr respondentů odpovídá ohledně vzdělání a) celostátnímu údaji, tj. že podíl základoškoláků, středoškoláků a vysokoškoláků je v poměru 7:9:4, b) je v stejném poměru.
2) Použijeme opět data ze souboru časopis.dat. Zjistěte, zda zájem o časopis závisí na vzdělání. V případě, že ano, určete koeficienty kontingence a vhodně je okomentujte.
3) Použijeme data ze souboru vydaje.dat. Rozhodněte, zda výše výdajů za zboží A závisí na vzdělání respondenta (neopomeňte ověřit předpoklady testu). Pokud ano, proveďte hlubší analýzu pomocí metody mnohonásobného porovnávání.

? Úkoly k zamyšlení a diskuzi

1) Zamyslete se nad tím, jak souvisí uspořádání údajů v kontingenční tabulce se závislostí jednotlivých proměnných.
2) Diskutujte o podmínkách na použití AR.

Klíč k řešení otázek:

1) Použijeme test o shodě poměrů. Příkaz: *pearson.test(x=c(), p=x()).*

   Použijeme příkazy:

   summary(casopis)

   Výstup:

```
id              zajem      vzdelani  pohlavi         vek           vekint
Min.   :   1.0   ano: 167   SS: 433   muz : 951   Min.   :15.00   (0,25] :310
1st Qu.: 500.8   ne :1833   VS: 426   zena:1049   1st Qu.:31.00   (25,40]:698
Median :1000.5              ZS:1141               Median :40.00   (40,60]:800
Mean   :1000.5                                    Mean   :40.81   (60,85]:192
```

```
3rd Qu.:1500.2                                        3rd Qu.:51.00
Max.   :2000.0                                        Max.   :85.00
```

table(casopis$zajem,casopis$vzdelani)

```
          SS   VS   ZS
  ano    63   69   35
  ne    370  357 1106
```

a) Testujeme H_0: $\pi_{SS} : \pi_{VS} : \pi_{ZS} = 9:4:7$ proti opačné alternativě H_1.

Test provedeme příkazem:

pearson.test(x=c(63,69,35), p=c(9/20,4/20,7/20))

Výstup a závěr:

```
Pearson's chi-squared test
Data:  c(63, 69, 35)
    Hypothetical probabilities: 0.45,  0.2,  0.35
X2 = 49.31737,  df = 2,   p-value = 1.953755e-11
Observed counts:
     63,  69,  35
Expected counts:
     75.15,  33.4,  58.45
Estimated probabilities:
     0.3772455,  0.4131737,  0.2095808
Pearson residuals:
     -1.401560,  6.159944,  -3.06726
Pearson squared residuals:
     1.964371,  37.94491,  9.408084
```

Zamítáme H_0, tj. výběr respondentů vzhledem ke vzdělání neodpovídá celostátnímu rozložení.

Po vyfiltrování podmnožiny studentů, kteří mají zájem o časopis, lze test provést i v *Statistics-Summaries-Frequency distribution* (a zaškrtnutím *Chi-squared-goodnes-of-fit test* a zadáním očekávaných četností pro jednotlivé kategorie vzdělání).

b) Obdobně dostaneme:

```
X-squared = 11.8323, df = 2, p-value = 0.002696
```

H_0 zamítáme, tj. výběr respondentů vzhledem ke vzdělání není ve stejném poměru.

2) Jde o test nezávislosti v kontingenční tabulce: Kontingenční tabulku vytvoříme v interaktivní nabídce (*Statistics – Contingency tabeles – Two-way table* + označíme proměnné a název tabulky: *.Table*). Současně se provede i chi-kvadrát test. Pokud chceme ještě určit koeficienty kontingence, použijeme po aktivizaci balíku *vsePackege* příkaz: *pearson.indep.test (.Table).*

Výsledky:

```
>.Table <- xtabs(~zajem+vzdelani, data=casopis)
> .Table
     vzdelani
zajem   SS   VS   ZS
  ano   63   69   35
  ne   370  357 1106
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
       Pearson's Chi-squared test
data:  .Table
X-squared = 97.6307, df = 2, p-value < 2.2e-16
```

Pomocí příkazu: *pearson.indep.test(.Table)* dostaneme mj.:

```
Pearson's chi-squared test of independence
Data:  .Table
X2 = 97.6307,  df = 2,   p-value = 6.306132e-22
Contingency coefficients:
           Pearson:  0.215739
 Pearson (maximum):  0.7071068
            Cramer:  0.2209420
```

Závěr: Zamítáme nulovou hypotézu H_0 o nezávislosti zájmu o nový časopis na vzdělání. Závislost je významná, ale intenzita vyjádřená koeficienty kontingence této závislosti je nízká.

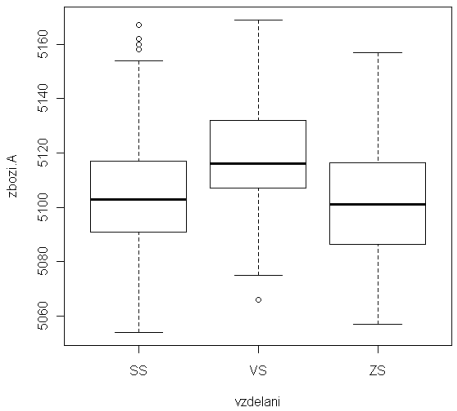
3) Analýzu provedeme pomocí jednofaktorové analýzy rozptylu (JAR) - závislost kvantitativní proměnné (odezva) na kvalitativní (faktor, ošetření). JAR provádíme interaktivně (*Statistics – Means – One-way ANOVA*, označíme faktor a odezvu). Výsledky se uloží pod názvem modelu a současně pomocí *summary(model)* jsou vypsané základní výsledky. V nabídce *Model* můžeme provádět další rozsáhlou výpočetní i grafickou analýzu – testy a grafy na ověřování podmínek (převážně aplikovanou na residua). Pokud

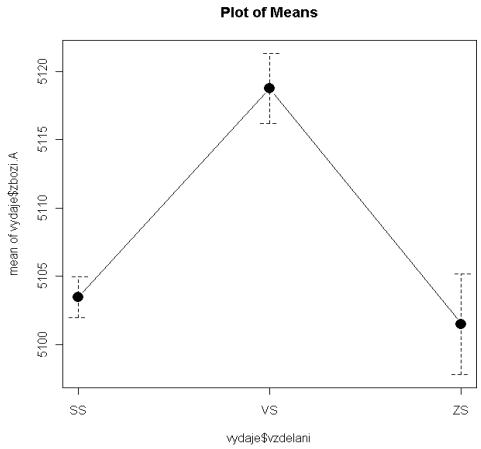
H_0 zamítneme, provádíme příkazem *TukeyHSD(model)* Tukeyovo vícenásobné porovnávání. Normalitu testujeme SWT a homoskedasticitu Bartlettovým testem (interaktivně z *Variance*). Pokud nejsou splněny podmínky, lze použít Kruskalův-Wallisův test též z interaktivní nabídky (z *Nonparametric tests*). Data musí být uspořádána do 2 sloupců (v jednom odezva a v druhém faktor).

Načteme data. Popsaným způsobem dostaneme model závislosti výdajů za zbozi A na vzdělání ve tvaru:

```
> AnovaModel.1 <- aov(zbozi.A ~ vzdelani, data=vydaje)
> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value    Pr(>F)
vzdelani      2  38936   19468  51.125 < 2.2e-16 ***
Residuals   997 379650     381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> numSummary(vydaje$zbozi.A , groups=vydaje$vzdelani, statistics=c("mean",
+   "sd"))
       mean       sd   n
SS 5103.487 19.82091 690
VS 5118.764 18.37047 199
ZS 5101.495 19.57358 111
```

Závěr: Nulovou hypotézu o nezávislosti těchto výdajů na vzdělání zamítáme. Krabicový diagram na následujícím obrázku potvrzuje výsledky testu. Výrazněji zamítnutí shody potvrzuje skupinový krabicový diagram a graf průměrů Výdajů za zboží A (*Graphs – Plot of means*, označime vzdelani a zbozi A, zaskrtneme *Conf.intervals*).





Příkazem *TukeyHSD(AnovaModel.1)* provedeme ještě Tukeyovo vícenásobné porovnávání. Výstup:

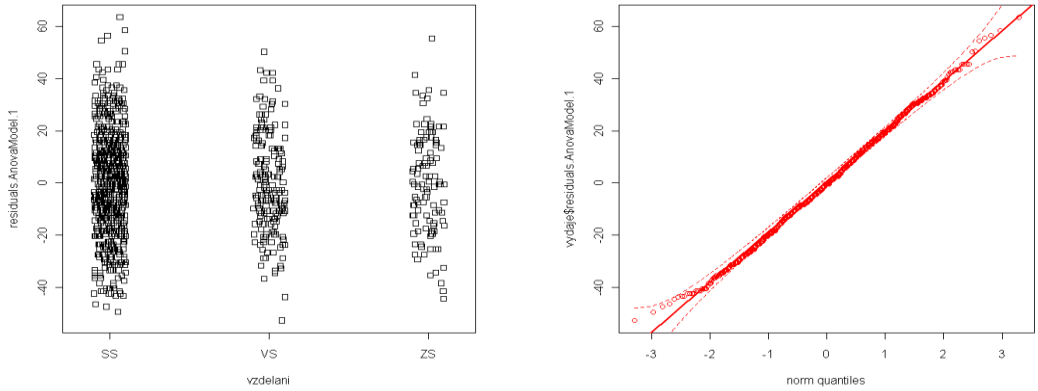
```
> TukeyHSD(AnovaModel.1)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = zbozi.A ~ vzdelani, data = vydaje)
$vzdelani
             diff        lwr        upr     p adj
VS-SS  15.276863  11.591363  18.962362 0.0000000
ZS-SS  -1.991461  -6.675568   2.692646 0.5783367
ZS-VS -17.268324 -22.694443 -11.842204 0.0000000
```

Znamená to, že významné rozdíly jsou mezi průměrnými výdaji základoškoláků a vysokoškoláků a mezi výdaji vysokoškoláků a středoškoláků (p-hodnoty jsou nulové).
Ověření podmínek pro AR: Pomocí *Data – Manager variables in activ data set – Compute new variable* (a vyplněním *New variable name: residuals*; *Expression to compute: residuals(AnovaModel.1)* přidáme k datům sloupec reziduí. Na ně aplikujeme SWT o normalitě a Bartlettův test o homoskedasticitě. Výstupy:

```
shapiro.test(vydaje$residuals)
       Shapiro-Wilk normality test
data:  vydaje$residuals
W = 0.9977, p-value = 0.1867

> bartlett.test(residuals.AnovaModel.1 ~ vzdelani, data=vydaje)
       Bartlett test of homogeneity of variances
data:  residuals.AnovaModel.1 by vzdelani
Bartlett's K-squared = 1.7246, df = 2, p-value = 0.4222
```

Normalita a homoskedasticita reziduí nebyla zamítnuta.

Poslední výsledky ověříme ještě graficky. Z *Graphs – Strip chart* – zaškrtneme *Jiter* a dostaneme graf závislosti reziduí na faktoru vzdělání (homoskedasticita). V Graph – Quantile-comparision dostaneme qq-diagram (normalita):



Grafy potvrzují splnění předpokladů potřebných k použití AR. Na doplnění ještě provedeme KWT:

```
> kruskal.test(zbozi.A ~ vzdelani, data=vydaje)

        Kruskal-Wallis rank sum test

data:  zbozi.A by vzdelani
Kruskal-Wallis chi-squared = 88.5283, df = 2, p-value < 2.2e-16
```

Závěry jsou stejné jako testu v JAR. Základní výsledky JAR lze získat i v Excelu použitím jednofaktorové analýzy z Analýzy dat (odezvová veličina musí být zapsána ve zvláštních sloupcích pro každou úroveň faktoru).

Kapitola 9: Jednoduchá lineární regrese a korelace

Klíčové pojmy:

regresní přímka, závislost funkční a statistická, regrese, korelace, elementární popis závislostí, bodový diagram, graf podmíněných průměrů, teoretická a empirická regresní přímka, metoda nejmenších čtverců, systém normálních rovnic, korelační a regresní koeficient, koeficient determinace a jeho interpretace, predikce, interpretace odhadnutých regresních parametrů, nelineární regrese

Cíle kapitoly:

- **pochopení principů jednoduché lineární regrese;**
- **být schopni odhadnout a interpretovat parametry regresní přímky;**
- **umět vypočítat a interpretovat ukazatele síly jednoduché lineární závislosti;**
- **naučit se využívat regresi k analýze a k predikci.**

Čas potřebný ke studiu kapitoly: 10 hodin

Výklad:

Nastínění obsahu kapitoly.

Úvod

Elementární popis závislostí

Regresní přímka a její odhad

Metoda nejmenších čtverců

Síla lineární závislosti

Interpretace odhadnutých regresních parametrů

Další typy regresních funkcí

Vyrovnání regresní přímky v Excelu a v R

Struktura výkladu

Regresní a korelační analýza umožňuje výrazně
rozvinout způsob vašeho statistického myšlení a
navršit kvalitu a efektivnost práce s ekonomickými daty
R.Hindls

Úvod

Budeme se zabývat studiem závislostí statistických veličin.

Závislost funkční (pevná, deterministická):

- v matematice, fyzice, technické praxi;
- vzájemné jednoznačné přiřazení;
- např. objem koule $V = \frac{4}{3}\pi r^3$, dráha volného pádu $s = \frac{gt^2}{2}$.

Závislost statistická (volná, nedeterministická):

- obecnější závislost studovaná ve statistice;

- při změnách jedné veličiny dochází ke změnám podmíněných středních hodnot druhé veličiny;
- Např. závislost středních výdajů rodiny na počtu členů rodiny, poptávky na ceně apod.
- Závislost je ovlivňována řadou dalších nekontrolovatelných vlivů i chyb (příjmy a velikost rodiny, její návyky apod.).

Studiem statistických závislostí se zabývá regresní a korelační analýza

- Pojem regrese - zaveden F. Galtonem koncem 19. století.
    - Vyšetřoval závislost výšky synů na výšce jejich otců;
    - zjistil tendenci jit ve výšce zpět k celkovému průměru;
    - původní význam slova "regression" byl proto návrat zpět.
- Regresní analýza - zkoumání průběhu statistické závislosti, tj. závislosti změn podmíněných průměrů vysvětlované proměnné na změnách vysvětlující proměnné.
    - Vysvětlujících proměnných může být více;
    - hledáme tvar tzv. regresní funkce, jejímž grafem je odpovídající regresní křivka;
    - na základě náhodného výběru najdeme empirickou regresní funkci, která představuje její odhad, a provádíme její analýzu.
- Korelační analýza - určování stupně síly (intenzity) s jakou se statistická závislost projevuje a vypočítat a interpretovat číselné charakteristiky (míry) této závislosti.
    - Obě disciplíny se vzájemně prolínají a budeme je probírat souběžně.

Elementární popis závislostí

Metody popisu:

- korelační tabulka a její graf;
- bodový (rozptylový) diagram;
- graf podmíněných průměrů (popř. i rozptylu).

Příklad 1: Viz Stuchlý (1999b), s. 8-9.

Regresní přímka a její odhad

Statistickou lineární závislost vysvětlované náhodné veličiny Y (regresand) na jedné vysvětlující veličině X (regresor) zapisujeme rovnicí (teoretický neboli populační regresní model)

$$E(Y|x) = \beta_0 + \beta_1 x,$$

kde β_0 a β_1 jsou regresní parametry (absolutní člen a směrnice). Podmínku v střední hodnotě obvykle vynecháváme.

- Např. závislost průměrné poptávky Y na ceně x.

Jde o lineární regresní funkci a jejím grafem je regresní přímka.

Hlavní úloha: odhad regresních parametrů.

Použijeme k tomu dvourozměrný náhodný výběr dvojic n pozorování (x_1,y_1), (x_2,y_2),....,(x_n,y_n).

Rovnici přepíšeme do stochastického tvaru:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = E(Y|x_i) + \varepsilon_i \text{ , } i = 1,\ldots,n,$$

kde ε_i představuje náhodné složky (chyby měření).

Regresní koeficienty odhadneme metodou nejmenších čtverců (MNČ). Jejich odhady označíme b_0 a b_1

Odhadnutá regresní funkce je

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = b_0 + b_1 x_i, \; i = 1,...,n,$$

což představuje výběrový (empirický) regresní model

Jiný tvar zápisu modelu

$$y_i = b_0 + b_1 x_i + e_i,$$

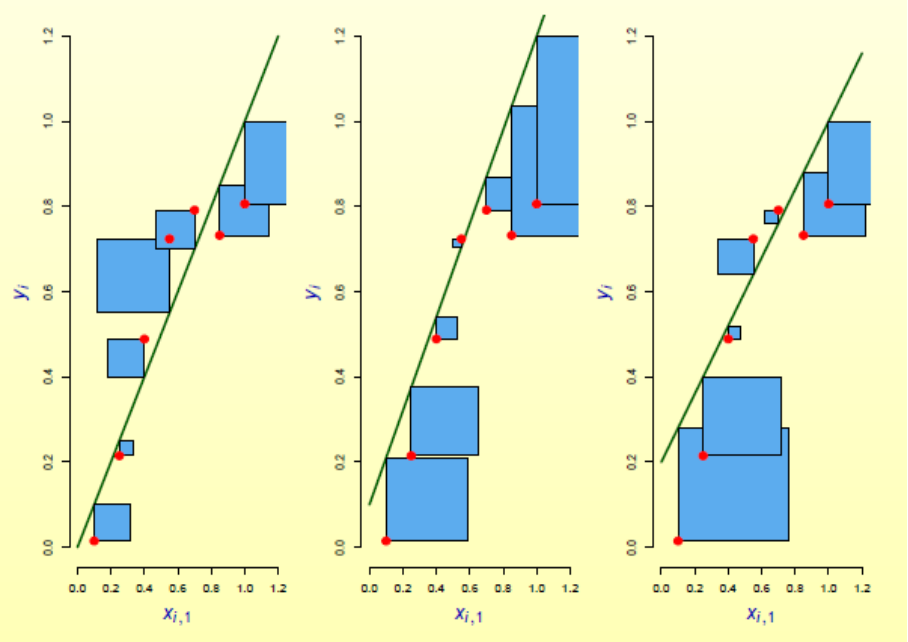
kde $e_i = y_i - (b_0 + b_1 x_i)$ jsou odhady náhodné složky - tzv. rezidua = empirická (naměřená) hodnota minus vyrovnaná hodnota;

residuals = empirical – fitted values.

Grafem je odhadnutá (empirická) regresní přímka

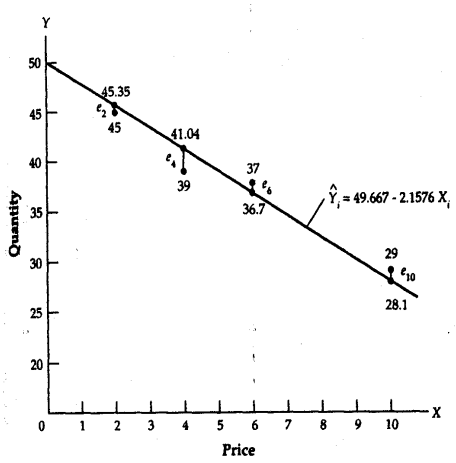
Metoda nejmenších čtverců

Grafické znázornění vyrovnání MNČ: Za optimální vyrovnání volíme to, které minimalizuje součet čtverců reziduí (naznačené čtverce).



Zdroj: Komárek 2007a

Matematický princip MNČ:



b_0, b_1 dostaneme minimalizací funkce

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Představuje součet čtverců rozdílů mezi empirickými a vyrovnanými hodnotami regresandu.

Derivováním podle proměnných b_0, b_1, položením těchto rovnic nule a úpravou dostaneme systém normálních rovnic (SNR) pro neznámé parametry

$$b_0\, n \quad + b_1 \Sigma\, x_i \quad = \Sigma\, y_i\,,$$

$$b_0 \Sigma\, x_i \; + b_1 \Sigma\, x_i^2 \; = \Sigma\, x_i\, y_i\,.$$

Řešením SNR dostaneme MNČ-odhady regresních parametrů

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}, \quad b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$$

S využitím kovariance a dalších výběrových charakteristik odtud dostaneme hledané odhady

$$b_1 = \frac{s_{xy}}{s_x^2},\; b_0 = \bar{y} - b_1 \bar{x},\; \hat{y} = \bar{y} + b_1 (x - \bar{x}),$$

kde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ jsou výběrové průměry, $s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ je výb. rozptyl a $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$ je výběrová kovariance.

Síla lineární závislosti:

Jako míra síly (intenzity) lineární závislosti Y na X se používá korelační koeficient

$$r_{yx} = \frac{s_{xy}}{s_x s_y}$$

Zde s_x , s_y jsou výběrové směrodatné odchylky a

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

je výběrová kovariance. Význam a interpretace r_{yx} je znám z popisné statistiky. Pro ruční výpočet lze použít vzorec

$$r_{yx}=\frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right\}\left\{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n\right\}}},$$

Síla obecné závislosti se obvykle také popisuje koeficientem determinace R^2, který u regresní přímky je roven čtverci korelačního koeficientu.

Interpretace R^2:

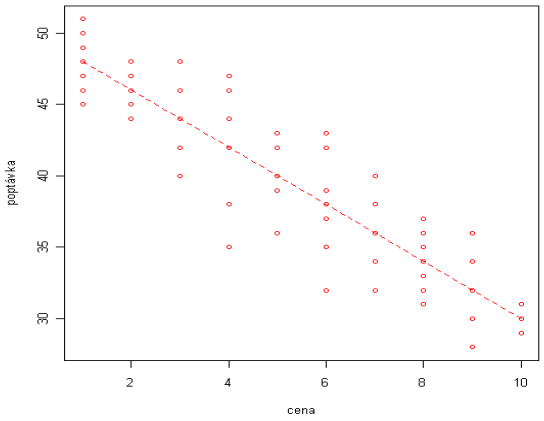
- Udává, jakou část změn vysvětlované proměnné je možno vysvětlit změnami vysvětlující proměnné (obvykle se vyjadřuje v %).

Regresní rovnici lze použít na předpověď (predikci) hodnoty vysvětlované proměnné y, když za x dosadíme do regresní rovnice zadanou hodnotu vysvětlující proměnné.

Příklad 2: V určité obci s 55 obyvateli byl proveden průzkum poptávky *Y* po určité zbožní komoditě v závislosti na ceně x_i. Při ceně 1 Kč byla poptávka u 7 obyvatel v množstvích 45, 46, 47, 48, 49, 50 a 51 kusů, při ceně 2 Kč byla poptávka u 5 obyvatel v množstvích 44, 45, 46, 47, 48 kusů. Další údaje o této poptávce v závislosti na ceně jsou uvedeny v 1. a 2. sloupci následující tabulky. Doplníme do tabulky podmíněné hodnoty poptávky v závislosti na ceně $E(Y \mid x_i)$. Znázorníme do jednoho obrázku bodový diagram závislosti poptávky na ceně a vypočítané podmíněné průměry. Spojíme tyto průměry populační regresní čarou (regresní přímka). Řešení provedeme do tabulky:

Cena [Kč] x_i	Poptávka [kusů] Y	Počet zákazníků	$E(Y \mid x_i)$
1	45 46 47 48 49 50 51	7	48
2	44 45 46 47 48	5	46
3	40 42 44 46 48	5	44
4	35 38 42 44 46 47	6	42
5	36 39 40 42 43	5	40
6	32 35 37 38 39 42 43	7	38
7	32 34 36 38 40	5	36
8	31 32 33 34 35 36 37	7	34
9	28 30 32 34 36	5	32
10	29 30 31	3	3
Součet	×	55	×

Závislost poptávky Y na ceně X:



Příklad 3: Pro data z předcházejícího příkladu byl proveden náhodný výběr. Jeho výsledek je v 1. a 2. sloupci následující tabulky. Odhadneme rovnici příslušné výběrové regresní funkce.

Data a výpočty jsou v tabulce:

x_i	y_i	$x_i y_i$	x_i^2	$\hat{y}_i$	e_i	e_i^2	y_i^2
1	49	49	1	47,509	1,4909	2,2228	2401
2	45	90	4	45,352	-0,3515	0,1236	2025
3	44	132	9	43,194	0,8061	0,6497	1936
4	39	156	16	41,036	-2,0364	4,1468	1521
5	38	190	25	38,879	-0,8788	0,7723	1444
6	37	222	36	36,721	0,2788	0,0777	1369
7	34	238	49	34,564	-0,5637	0,3177	1156
8	33	264	64	32,406	0,5940	0,3528	1089
9	30	270	81	30,248	-0,2484	0,0617	900
10	29	290	100	28,091	0,9091	0,8265	841
55	378	1901	385	378,000	0	9,5515	14682

Základní číselné charakteristiky:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 5{,}5, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 37{,}8,$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = 38{,}5 - 5{,}5^2 = 8{,}25, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = 14682 - 37{,}8^2 = 39.36$$

$s_x = 2{,}87,$ $s_y = 6{,}27,$

$$s_{yx} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = 190{,}1 - 5{,}5.37{,}8 = -17{,}$$

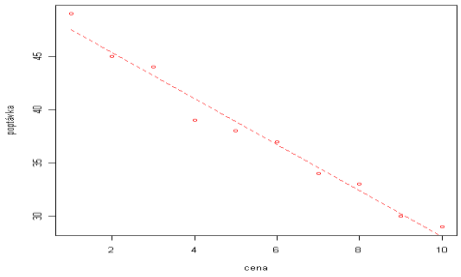
Odhady regresních parametrů:

$$b_1 = \frac{s_{yx}}{s_x^2} = \frac{-17,8}{8,25} = -2,1576, \quad b_0 = \bar{y} - b_1 \bar{x} = 37,8 - (-2,1576).5,5 = 49,6670.$$

Odhad regresní přímky:

$$\hat{y} = \bar{y} + b_1(x - \bar{x}) = 37,8 - 2,1576(x - 5,5) = 49,6770 - 2,1576\ x.$$

Výběrová závislost poptávky Y na ceně x:



Síla lineární závislosti

Korelační koeficient:

$$r_{yx} = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right\}\left\{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n\right\}}} =$$

$$= \frac{1901 - 55/10}{\sqrt{385 - 55^2/10}\sqrt{14682 - 378^2/10}} = \frac{-17,8}{2,87.6,27} = -0,989.$$

Znamená to, že síla lineární závislosti je velká a nepřímá (s růstem ceny klesá poptávka) – viz obrázek.

Koeficient determinace $R^2 = (-0,989)^2 = 0,978$

Znamená to, že změnami cen je vysvětleno 97,8% změn poptávky.

Interpretace odhadnutých regresních parametrů

Důležitá je směrnice regresní přímky b_1. Nazýváme jí regresním koeficientem.

Udává, o kolik se změní vysvětlovaná proměnná, když se vysvětlující proměnná změní o jednotku.

Absolutní člen b_0 má význam jen v některých situacích.

V uvedeném př. 3: je rovnice regresní přímky E(Y)=49,7-2,16x.

Interpretace b_1: Zvětší-li se cena o 1 Kč, sníží se poptávka v průměru o 2,16 ks.

Interpretace b_0: Poptávka při nulové ceně je v průměru 49,7 ks.

Predikce: Při ceně x=5,50Kč je předpověď poptávky y=49,7-2,16.5,50=37,8 ks.

Další typy regresních funkcí

Pokud vidíme z bodového diagramu, že mezi proměnnými je nelineární statistická závislost, můžeme často i tuto nelineární regresní funkci pomocí vhodné transformace převést na lineární a tuto potom vyrovnat MNČ.

Např. funkci $y = a + b/z + e$ převedeme na regresní přímku transformací $x = 1/z$.

Exponenciální závislost $y = e^{a+bz+e}$ nebo mocninnou $y = az^b$ převedeme na lineární logaritmováním této rovnice.

Příklad 4: Viz Stuchlý (1999b), s. 41-42.

Vyrovnání regresní přímky v Excelu a v R

Regresní přímka v Excelu:

- Vyrovnání lineární regresní funkce i korelační koeficient: *Analýza dat – Regrese*.
- Korelační koeficient dostaneme i pomocí statistické funkce *Corel*.

Regresní přímka v R:

- V *Statistics – Fit model - Linear regression* vytvoříme i odhadneme model.
- Graf dostaneme v *Graphs – Scatterplot* (necháme zaškrtnuté jen *Least-squares line*.
- Bodovou předpověď dostaneme příkazem:

  predict(model, newdata=data.frame(prom=).

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 169-210. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika II Cvičení ze statistických metod pro manažery*. J. Hradec: VŠE, 1999. S. 5-15, 21-22, 25-26, 41-43. ISBN 80-7079-035-0.

Doporučené studijní zdroje:

GIBILISCO, S. *Statistika bez předchozích znalostí.* Brno: Computer Press, 2009. S. 152-177, 216-225. ISBN 978-80-251-2465-9.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 122-132, 146-154. ISBN 80-7169-255-7.

HINDLS, R. a kol. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000. S. 19-32, 44-59, 72-77. ISBN 80-7261-013-9.

JAROŠOVÁ, E. *Statistika B. Řešené příklady.* Praha: VŠE, 1994. S. 9-14, 31-32, 37-39, 54-57. ISBN 80-7079-328-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 213-215, 222-236, 249-250. ISBN 978-80-86446-40.

MINAŘÍK, B. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita, 1995. S. 94-112. ISBN 80-7157-166-0.

ŘEZANKOVÁ, H. a T. LÖSTER. *Úvod do statistiky*. Praha: Oeconomica, 2009. S. 56-58. ISBN 978-80-245-1514-4.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 167-187, 202-214. ISBN 80-7187-058-7.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/).

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 309-325. ISBN 80-7169-089-9.

WONNACOT, T. H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 388-407, 487-500, 514-522. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Pracovník personálního oddělení určitého podniku zkoumá, zda existuje vztah mezi počtem dní absence v práci a věkem pracovníka. Náhodně vybere pracovní záznamy 10 pracovníků a získá údaje o jejich věku x_i (v letech) a počtu dní y_i, v kterých nenastoupili do práce v době jednoho kalendářního roku. Údaje jsou v následující tabulce:

x_i	27	61	37	23	46	58	29	36	64	40
y_i	15	6	10	18	9	7	14	11	5	8

Určete: a) bodový odhad regresní přímky (napište i systém normálních rovnic), b) charakteristiky popisující sílu této závislosti a interpretujte jejich význam, c) interpretujte odhadnutý regresní koeficient, d) odhadněte průměrný počet dní absence pro 26-ti letého pracovníka.

2) Hodláte prodat auto, které má najeto 30000 km, a chcete si udělat představu o jeho prodejní ceně. V bazaru stojí 50 aut téže značky, údaje o ceně a počtu najetých kilometru naleznete v datovém souboru ojetiny.dat, resp. ojetiny.csv. V souboru jsou následující

údaje: id (identifikační číslo ojetého auta), cena (cena ojetého auta v tis. Kč), najeto (počet najetých kilometrů v tis. km). Pomocí vhodného obrázku a charakteristiky popište míru závislosti mezi cenou ojetého auta a počtem najetých km.

3) Pro zadání z předchozího úkolu na základě modelu regresní přímky proveďte následující kroky: a) Odhadněte průměrnou cenu nového auta. b) Odhadněte, jak se průměrná cena auta změní s každými 10000 najetými kilometry. c) Pomocí vhodné charaktristiky posuďe vhodnost modelu. d) Rádi byste prodali vaše auto za 150000 Kč. Odpovídá vaše představa cenám ojetin v bazaru? Své rozhodnutí zdůvodněte.

4) Vyrovnejte data ze souboru ojetiny.csv regresní hyperbolou. Porovnejte kvalitu tohoto vyrovnání s vyrovnáním regresní přímkou.

? Úkoly k zamyšlení a diskuzi

1) Při pokusu najít přímku proloženou MNČ v bodovém diagramu použijeme počítačový program. Ten nám ovšem ukáže, že žádná taková přímka neexistuje. Když si graf prohlédneme, zjistíme, že body jsou rozprostřeny po celém prostoru. Korelace mezi dvěma proměnnými se očividně a) nachází mezi 0 a +1, b) se rovná 0, c) nachází mezi -1 a 0, d) rovná -1. Které z uvedených tvzení platí?

2) Uvažujme lineární regresní závislost mezi ziskem a prodejem. Jaké hodnoty mohou v tomto případě nabývat regresní parametry: negativní nulové nebo pozitivní? Jaká je jejich ekonomická interpretace?

🔑 Klíč k řešení otázek:

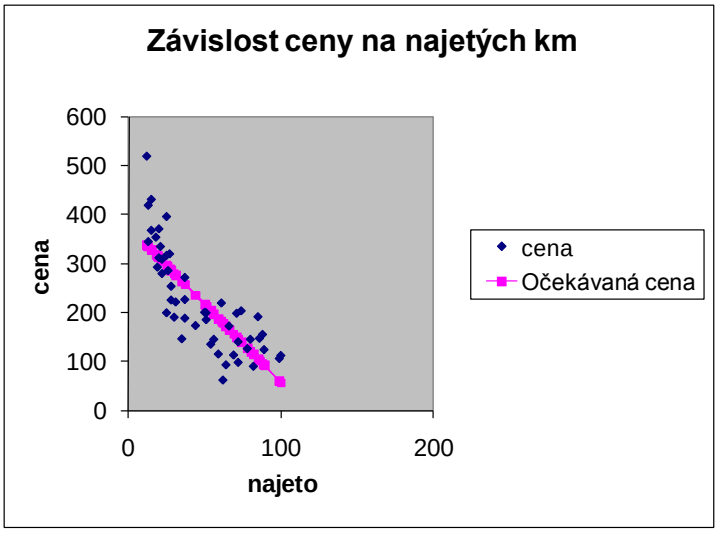
1) Regresní přímka: a) 21,578 - 0,268x, ($10b_0+421b_1=103$, $421b_0+19661b_1=3817$); b) r = -0,9325, R^2 = 0,8692; c) Zvýší-li se věk pracovníka o 1 rok, sníží se průměrná roční absence o 0,268 hodin; d) 14,6 dní. Podrobné řešení viz Stuchlý (1999b), s. 25-26.

2) Regresní přímka v Excelu: V *Analýze dat* použijeme nabídku *Regrese*. Ve vstupním okně vypíšeme: *Vstupní oblast Y*: B1:B51 (odkaz na ceny); *Vstupní oblast X*: B1:B51 (odkaz na najeto). Zaškrtneme *Popisky* a *Graf regresní přímky*. Z Výstupu:

Regresní statistika

Násobné R	0,808076	korelační koeficient
Hodnota spolehlivosti R	0,652988	koeficient determinace
Nastavená hodnota spolehlivosti R	0,645758	korigovaný koeficient determinace
Chyba stř. hodnoty	62,03298	residuální standartní chyba
Pozorování	50	počet měření

	Koeficienty
Hranice	374,7484
najeto	-3,18673



Na obrázku je bodový diagram a odhadnutá regresní přímka. Korelační koeficient lze určit v *Analýze dat* z nabídky *Korelace* (přejedeme oba sloupce dat).

Dostaneme r = -0,808. Interpretace: Mezi cenami a počtem najetých kilometrů je silná nepřímá lineární závislost.

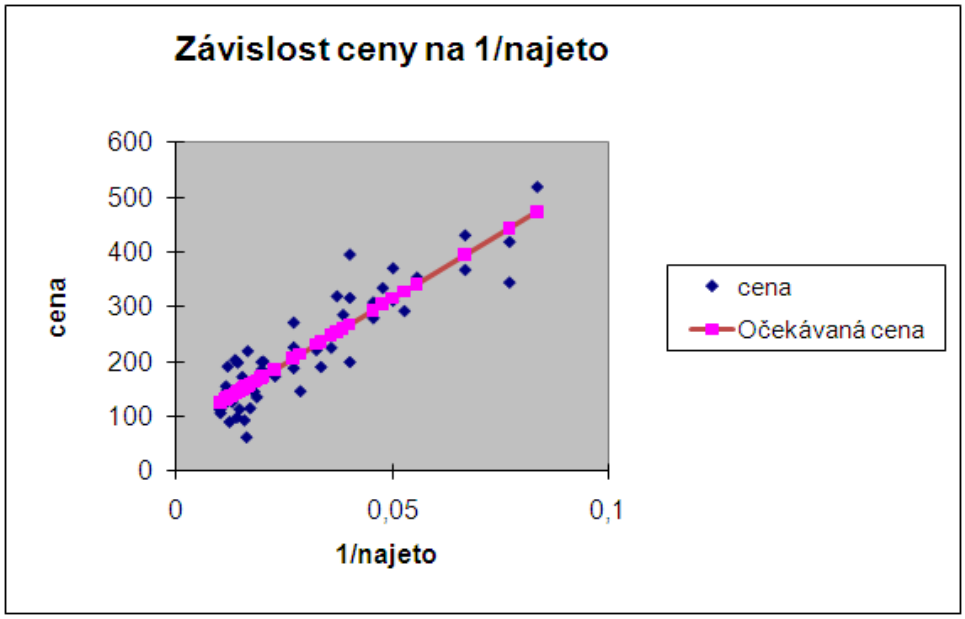
3) Z počítačového výstupu k 2. úkolu dostáváme: a) Průměrná cena nového auta je 375 tis. Kč; b) S každými 10 000 najetými km klesne průměrná cena o 32 tis. Kč; c) Koeficient determinace je $R^2 = 0,653$. Jeho interpretace: Změnami v počtu najetých km je lineárním modelem vysvětleno 65,3% změn cen; d) Předpověď dostaneme dosazením za najeto = 30 tis. Kč do regresní rovnice: průměrná cena = 374,7 – 3,2.najeto = 374,7 – 3,19.30 =

279 tis. Kč. Tedy vaše cena 150 tis. Kč představuje značné podhodnocení ceny prodávaného auta.

4) Regresní přímka v R: Odhadujeme model střední cena = $\beta_0+\beta_1$.(1/najeto). Postupujeme obdobně jako v úkolu 2, jen místo proměnné najeto použijeme její převrácenou hodnotu, kterou dopočítáme v Excelu. Výstup z Excelu:

Regresní statistika	
Násobné R	0,902415
Hodnota spolehlivosti R	0,814352
Nastavená hodnota spolehlivosti R	0,810484
Chyba stř. hodnoty	45,37278
Pozorování	50

	Koeficienty
Hranice	78,52338
1/najeto	4752,269



Rovnice odhadnuté závislosti: střední cena = 78,5 + 4752,3.(1/najeto). Koeficient determinace R^2 = 0,902 i obrázek ukazují, že regresní hyperbola představuje lepší vyrovnání než regresní přímka.

Kapitola 10: Statistická indukce v regresním modelu

Klíčové pojmy:

statistická indukce v regresním modelu, residuální rozptyl, standardní chyba odhadu, standardní normální model, standardní chyby regresních parametrů, intervaly spolehlivosti a testy pro regresní parametry, index determinace, bodová a intervalová předpověď, predikční chyba, homoskedasticita, heteroskedasticita, autokorelace, Levenův test, Breutch-Paganův test, Durbinův-Watsonův test, residuální analýza

Cíle kapitoly:

- **uvědomit si možnosti provádění statistické indukce v regresi;**
- **umět ověřovat předpoklady pro použití statistické indukce v regresi;**
- **naučit se interpretovat výsledky statistické indukce v regresi.**

Čas potřebný ke studiu kapitoly: 11 hodin

Výklad:

Nastínění obsahu kapitoly.

Odhady náhodné složky

Standardní normální regresní model (SNRM)

Vlastnosti odhadů v SNRM

Statistická indukce v SNRM

- Intervaly spolehlivosti
- Testy

Míry síly závislosti

Použití modelu na předpověď

Ověřování podmínek SNRM

Struktura výkladu

Suave est ex magno tollere acervo
Milo jest bráti z velkého množství
Horatius

Odhady náhodné složky

Náhodné složky ε_i , (i = 1, …,n) odhadujeme pomocí reziduí $e_i = y_i - \hat{y}_i$

- Tedy rezidua jsou rozdíly empirických a vyrovnaných hodnot.

Nestranný odhad rozptylu náhodné složky: $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = s^2.$

- Je to residuální součet čtverců dělený odpovídajícími stupni volnosti n-2.
- Odmocněním dostaneme s - standardní chybu odhadu (SEE).
- Charakterizuje přesnost odhadu regresního modelu.

Standardní normální regresní model

Též klasický regresní model popsaný rovnicí $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, v kterém jsou plněny podmínky:

- náhodné složky ε_i jsou nezávislé,
- mají rozdělení $N(0; \sigma^2)$.

Vlastnosti odhadů v SNRM

- V standardním normálním regresním modelu pro n=1,…,n platí: $y_i = b_0 + b_1 x_i + e_i$,

  $b_0 \sim N(\beta_0; \frac{\sigma^2 \sum x_i^2}{n^2 s_x^2})$, $b_1 \sim N(\beta_1; \frac{\sigma^2}{n s_x^2})$, $\frac{1}{\sigma^2} \sum_{i=1}^{n} e_i^2{}^2 \sim \chi^2(n-2)$.

- Rozptyly odhadnutých parametrů b_0 a b_1 obsahují neznámý parametr σ^2. Po jeho náhradě odhadem s^2 a odmocnění dostaneme <u>standardní chyby</u> odhadnutých regresních parametrů:

  $s(b_o) = s \sqrt{\frac{\sum x_i^2}{n^2 s_x^2}}$, $s(b_1) = s \sqrt{\frac{1}{n s_x^2}}$.

- Představují odhadnuté směrodatné odchylky odhadů parametrů b_0 a b_1.
- Charakterizují přesnost odhadnutých regresních parametrů.

Statistická indukce v SNRM

<u>Intervaly spolehlivosti</u> pro regresní parametry (i = 0, 1):

$P(b_i - t_{1-\alpha/2}(n-2)s(b_i) \le \beta_i \le b_i + t_{1-\alpha/2}(n-2)s(b_i)) = 1-\alpha.$

Interval spolehlivosti pro rozptyl náhodných složek:

$$P\left(\frac{s^2(n-2)}{\chi^2_{1-\alpha/2}(n-2)} \le \sigma^2 \le \frac{s^2(n-2)}{\chi^2_{\alpha/2}(n-2)} \right) = 1-\alpha.$$

<u>Statistické testy v regresním modelu:</u>

Testování významnosti regresních parametrů:

- Testujeme hypotézu H_0: $\beta_i = 0$ proti alternativní hypotéze H_1: $\beta_i \neq 0$ (i = 0, 1) na hladině významnosti α.
- Hypotézu H_0 zamítáme na kritickém oboru $W = \{T = b_i/s(b_i): |T| > t_{1-\alpha/2}(n-2)\}$.
- Je-li předem známé, že $\beta_i > 0$ nebo $\beta_i < 0$, používáme potom jednostranné testy.

Zobecnění testů významnosti:

- Testujeme hypotézu H_0: $\beta_i = \beta_i^*$, kde β_i^* je určitá předem zvolená konstanta, proti H_1: $\beta_i \neq \beta_i^*$. Testování provádíme stejným způsobem, jen místo dřívějšího testového kritéria používáme kritérium $T = (b_i - \beta_i^*)/s(b_i)$.

Míry síly závislosti

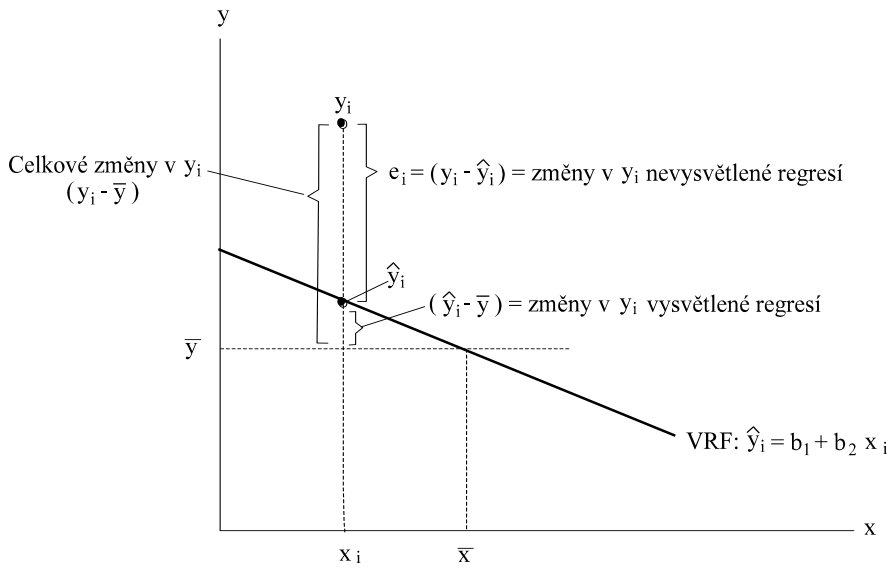
Koeficient determinace: Definujeme vztahem $$R^2 = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y}.$$

- Zde

$S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$ je celkový součet čtverců,

$S_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ teoretický součet čtverců neboli součet čtverců vysvětlený regresí,

$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ reziduální součet čtverců neboli součet čtverců regresí nevysvětlený



Koeficient korelace: $r_{yx}=\frac{s_{yx}}{s_x s_y}$

Použití regresního modelu na předpověď

- Předpověď bodová:
  - dostaneme jí dosazením za x do předpovědní rovnice.
- Předpověď intervalová:
  - Predikční interval (pro Y):

$$P(y_P - t_{1-\alpha/2}(n-2)s(y_P) \le Y \le y_P + t_{1-\alpha/2}(n-2)s(y_P)) = 1-\alpha$$

  - Predikční chyba: $s(y_P)=s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{ns_x^2}}$
  - Je možno počítat i přesnější konfidenční interval pro E(Y).
  - Konfidenční chyba: $s(y_c)=s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{ns_x^2}}$

Příklad: Statistická indukce v modelu regresní přímky: Z tabulky u příkladu 3 (závislost poptávky q na ceně p) z kap. 9 dostáváme:

a) Odhad rozptylu σ^2 náhodných složek a všech standardních chyb:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}e_i^2 = \frac{1}{10-2}9.5515 = 1.1939, \quad s = 1{,}0926,$$

$$s(b_0) = \sqrt{\frac{\sum x_i^2}{n^2 s_x^2}} = 1.0926\sqrt{\frac{385}{100.8{,}25}} = 0.7464,$$

$$s(b_1) = \sqrt{\frac{1}{ns_x^2}} = 1.0926\sqrt{\frac{1}{10\,8.25}} = 0.1203.$$

b) Intervaly spolehlivosti pro regresní parametry:

$$P(b_i - t_{1-\alpha/2}(n-2)s(b_i) \le \beta_i \le b_i + t_{1-\alpha/2}(n-2)s(b_i)) = 1-\alpha, \quad i = 0, 1.$$

P(-2,1576-2,306.0,1203≤ β_1 ≤-2,1576+2,306.0,1203) = 0,95,

tj. P(-2,435≤ β_1 ≤ -1,880)=0,95,

P(49,667-2,306.0,7464≤ β_0 ≤49,667+2,306.0,7464) = 0,95,

tj. P(47,946≤ β_0 ≤ 51,388)=0,95,

c) Test významnosti regresních parametrů:

$T = b_1/s(b_1)$ = -2,1576/0,1203 = -17,94,

|T| = 17,97 > $t_{0,975}(8)$ = 2,306,

$T=b_0/s(b_0)$=49,667/0,7464=66.54

|T| = 66.54 > $t_{0,975}(8)$=2,306,

tj. oba koeficienty jsou statisticky významné

d) Koeficient (index) determinace a korelační koeficient:

$R^2 = I_{yx}^2 = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y}$ = 1 - 9,5515/(10.39,39) = 0,9757,

$r_{yx} = \frac{s_{yx}}{s_x s_y}$ = -17,8/(2,87.6,27) = 0,989.

e) Prezentaci výsledků:

$\hat{y}$ = 49,6670 - 2,1576 x_i ; R^2 = 0,9757

se = (0,7464) 0,1203) , s.v. = 8

t = (66,538) (-17,935)

f) Předpověď (predikci):

a) bodovou: pro x = 5,5 je y = 49,6670 - 2,1576.5,5 = 37,799,

b) intervalovou:

predikční chyba: $s(y_P)=s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{ns_x^2}} = 1{,}0926\sqrt{1+\frac{1}{10}+\frac{(5{,}5-5{,}5)^2}{10.8{,}25}} = 1{,}146,$

$$P(y_P - t_{1-\alpha/2}(n-2)s(y_P) \le Y \le y_P + t_{1-\alpha/2}(n-2)s(y_P)) = 1-\alpha$$

predikční interval:

$P(37{,}799-2{,}306.1{,}146 \leq Y \leq 37{,}799+2{,}306.1{,}146) = 0{,}95,$

tj. $P(35{,}156 \leq Y \leq 40{,}442) = 0.95$

Základní výsledky je možno získat i v Excelu:

Regresní statistika					Test významnosti:	
Násobné R	0,987792	korelační koeficient			H_0: β_1=0 vs. H_1: β_1≠0	
Hodnota spolehlivosti R	0,975733	koeficient determinace			|T|=17,9>	2,6850108
Nastavená hodnota spolehlivosti	0,9727	korigovaný koeficient determinace			p-hod=	9,576E-08
Chyba stř. hodnoty	1,092675	s			H_0 zamítáme	
Pozorování	10	n			q závisí významně na p	
ANOVA						
	Rozdíl	*SS*	*MS*	*F*	*Významnost F*	
Regrese	1	384,0484848	384,05	321,665	9,57605E-08	
Rezidua	8	9,551515152	1,1939			
Celkem	9	393,6				
	Koeficienty	*Chyba stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	49,66667	0,746439359	66,538	2,9E-12	47,94537442	51,387959
p	-2,157576	0,120299595	-17,935	9,58E-08	-2,43498712	-1,880164
Předpověď q pro p=5,5:	37,8					
Interpretace b_1	Vzroste-li cena p o 1 Kč, klesna poptávka q v průměru o 2,17 ks					

Ověřování podmínek SNRM

Normalita chyb:

- SWT aplikovaným na rezidua.
- QQ-diagram reziduí.

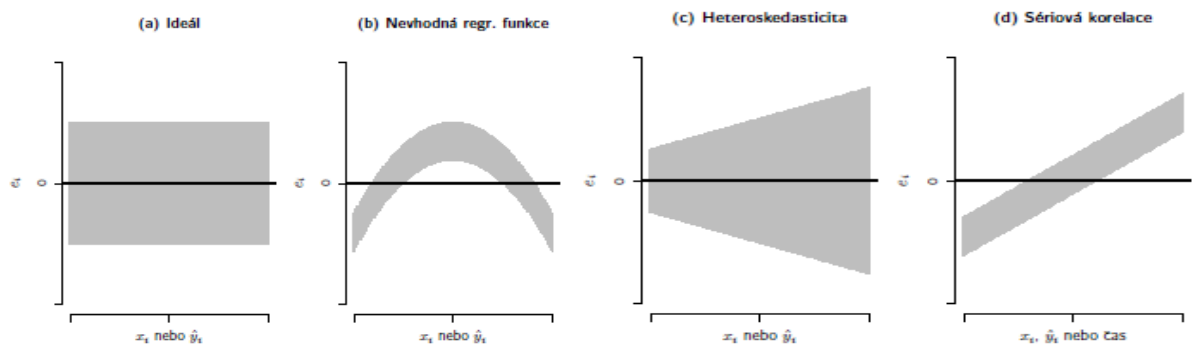
Homoskedasticita chyb (rozptyl se nemění s i)

- Levenovým nebo Breusch-Paganovým testem (v R).
- Grafem závislosti reziduí na pořadí měření nebo na hodnotách vysvětlované proměnné, krabicový diagram.

Nezávislost chyb (nepřítomnost autokorelace = sériová závislost).

- Durbinův-Watsonovým testem (DWT) reziduí - viz Hindls (2007), s. 320.
- Grafem závislosti reziduí na pořadí měření nebo proměnných.

Z grafu reziduí je možno usuzovat na následující problémy v regresním modelu (residuální analýza):



Zdroj: Komárek 2007a

Odstraňování problémů v regresním modelu:

Nenormalita chyb:

- Použijeme jiný model nebo transformovaný model.

Heteroskedasticita chyb:

- Odstraníme odlehlá pozorování.
- Místo MNČ použijeme metodu vážených nejmenších čtverců (viz Stuchlý 2000).

Porušená nezávislost chyb:

- Použijeme pokročilejší metody odhadu (zobecněná metoda nejmenších čtverců, ARIMA metody, metoda maximální věrohodnosti – viz Stuchlý 2000).

R-kové příkazy:

- Regresní modely (v *Statistics - Linear model…*):

  lm(y~x, data=);
  lm(y~I(1/x), data=);
  lm(y~I(log(x),data=);
  lm(y~I(sqrt(x), data=);
  lm(y~x+I(x^2), data=).

- Odhady a testy regresních parametrů (po aktivaci balíku *vsePackage*):

  lmbeta.test(model, beta.null= , alternative= , conf.level=)

- Předpovědi (predikční a konfidenční):

  predict(model, newdata=data.frame(x=), interval=„prediction“, level=);
  predict(model, newdata=data.frame(x=), interval=„confidence“, level=).

- Levenův test:

  skupiny <- (data$x >= median(data$x));
  levene.var.test(residuals(model)~skupiny).

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 226-234. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika II Cvičení ze statistických metod pro manažery*. J. Hradec: VŠE, 1999. S. 17-22, 25-27. ISBN 80-7079-035-0.

Doporučené studijní zdroje:

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 132-138, 140-142. ISBN 80-7169-255-7.

HINDLS, R. a kol. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000. S. 59-68. ISBN 80-7261-013-9.

JAROŠOVÁ, E. *Statistika B. Řešené příklady.* Praha: VŠE, 1994. S. 39-46. ISBN 80-7079-328-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 215-222. ISBN 978-80-86446-40.

MINAŘÍK, B. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita, 1995. S. 114-118, 120-123. ISBN 80-7157-166-0.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 193-197, 236-243. ISBN 80-7187-058-7.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/).

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 325-330. ISBN 80-7169-089-9.

WONNACOT, T.H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 408-429, 500-512. ISBN 80-85605-09-0.

? Otázky a úkoly

1) V úkolu 1 kap. 9 jsme odhadli závislost mezi počtem dní absence y a věkem pracovníka x (v letech) lineární regresní funkcí tvaru E(Y) = 21,59 – 0,27x a sílu lineární závislosti popsali korelačním koeficientem r = -0,933. a) Testujte významnost regresního koeficientu a určete příslušný interval spolehlivosti. b) odhadněte bodově i intervalově průměrný počet dní absence pro 26-ti letého pracovníka, c) odhadněte bodově i intervalově počet dní absence v letech pro jednoho 26-ti letého pracovníka.
2) V úkolech 2-3 kap. 9 jsme řešili v Excelu základní zadání z regrese a korelace. Nyní si vyřešte v R tato zadání, rozšířená o statistickou indukci. Tedy hodláte prodat auto, které má najeto 30000 km, a chcete si udělat představu o jeho prodejní ceně. V bazaru stojí 50 aut téže značky, údaje o ceně a počtu najetých kilometru naleznete v datovém souboru ojetiny.dat. a) Pomocí vhodného obrázku a charakteristiky popište míru závislosti mezi cenou ojetého auta a počtem najetých km. b) Odhadněte bodově a intervalově vprůměrnou cenu nového auta. c) Odhadněte bodově a intervalově, jak se průměrná cena auta změní s každými 10000 najetými kilometry. d) Otestujte, zda cena auta závisí

na počtu najetých kilometrů. e) Otestujte, zda cena auta klesá s počtem najetých kilometrů. f) Odhadněte bodově a intervalově průměrnou cenu aut, které mají najeto stejně jako vaše auto, tj. 30000 km. g) Rádi byste prodali vaše auto za 150000 Kč. Odpovídá vaše představa cenám ojetin v bazaru? Své rozhodnutí zdůvodněte. h) Ověřte předpoklady regresní analýzy.

3) Porovnání modelů. Vyberte pro data ze souboru ojetiny2.dat nejvhodnější jednovýběrovou regresní funkci pro závislost ceny ojetého auta na počtu najetých km. Použijeme tyto regresní funkce: a) přímku, b) odmocninovou funkce, c) hyperbolu, d) logaritmickou funkci, e) kvadratickou funkci. Rozhodování provedeme pomocí R^2, s a bodového diagramu. Řešte úlohu v R.

? Úkoly k zamyšlení a diskuzi

1) Uvažujte o analogii mezi jednovýběrovým t-testem o průměru a testy o regresních parametrech.
2) Pokuste se řešit předcházející řešený úkol 2 pomocí regresní hyperboly.

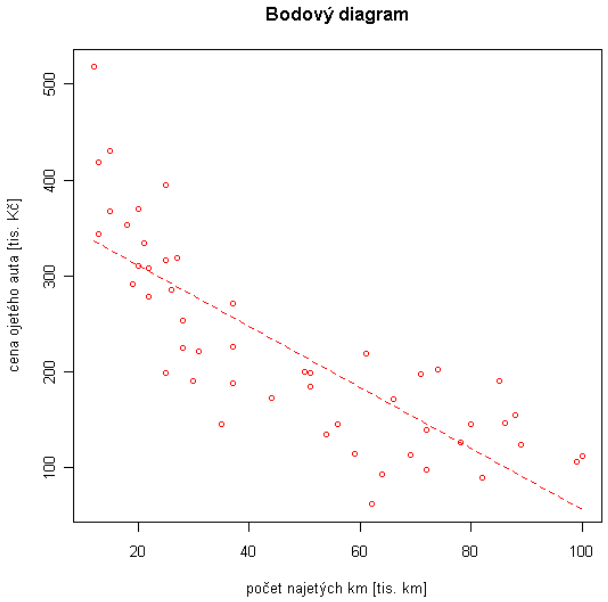
Klíč k řešení otázek:

1) Statistická indukce v regresním modelu: a) Testujeme hypotézu H_0: $\beta_1 = 0$ proti jednostranné alternativě H_1: $\beta_1 < 0$ na hladině významnosti $\alpha = 0,05$. K tomu účelu použijeme testové kritérium $T = b_1/s(b_1) = -0,2681/0,0367 = -7,305$. Protože platí $T < -t_{0,95}(8) = -1,860$, zamítáme nulovou hypotézu a tvrdíme, že lineární vztah mezi počtem dní absence pracovníka a věkem pracovníka je na 5%-ní hladině významnosti statistiky významný; $P(-0,353 \leq \beta_1 \leq -0,183) = 0,95$. b) $E(Y|x=26) = 14,62$, $P(12,81 \leq E(Y|x = 26) \leq 16,42) = 0,95$; c) $P(10,48 \leq Y_n \leq 18,75) = 0,95$. Podrobněji viz Stuchlý (1999b), s. 26.

2) Úlohy budeme řešit v R-ku pomocí příkazů (většinu kroků je možno řešit interaktivně, tj. z nabídky). Aktivujeme balík *vsePackage* a načteme data ze souboru ojetiny.dat.

a) Bodový diagram s regresní přímkou dostaneme nabídky *Graphs-Scatterplot* (označíme cena a najeto a zrušíme *Marginal boxplot* a *Smooth Line*, napíšeme do *x-axis label:* počet najetých km [tis. km] a do *y-axis label:* cena ojetého auta [tis. Kč] nebo použijeme příkaz (pokud chceme mít i hlavní nadpis):

scatterplot(cena~najeto, reg.line=lm, smooth=F, main="Bodový diagram", xlab="počet najetých km [tis. km]", ylab="cena ojetého auta [tis. Kč]", boxplot=F, span=0.5, data=ojetiny)



Korelační koeficient příkazem *cor(ojetiny$najeto, ojetiny$cena)*:

```
> cor(ojetiny$najeto, ojetiny$cena)
[1] -0.8080765
```

Regresní přímku dostaneme v nabidce *Statistics-Fit models-Linear regression…* (zaškrtneme cenu a najeto a stiskneme OK):

```
> RegModel.1 <- lm(cena~najeto, data=ojetiny)

> summary(RegModel.1)

Call:
lm(formula = cena ~ najeto, data = ojetiny)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
-117.21299  -45.94419   -0.09883   39.69985  181.49233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 374.7484    18.1188  20.683  < 2e-16 ***
najeto       -3.1867     0.3353  -9.504 1.30e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.03 on 48 degrees of freedom
Multiple R-squared: 0.653,Adjusted R-squared: 0.6458
F-statistic: 90.32 on 1 and 48 DF,  p-value: 1.304e-12
```

Testy a interval spolehlivosti pro regresní parametry dostaneme příkazem *lmbeta.test(RegModel.1):*

```
> library(vsePackage)
> lmbeta.test(RegModel.1)
              Estimate Std. Error Conf. Alternative Estim. Low  Estim. Up
(Intercept) 374.748379 18.1188041  0.95   two.sided 338.318082 411.178676
najeto       -3.186725  0.3353081  0.95   two.sided  -3.860908  -2.512543
            Beta H0   t value      p value
(Intercept)       0   20.682843   1.490125e-25
najeto            0   -9.503872   1.304310e-12
```

Interpretace regresních parametrů:

b) Průměrná cena nového auta je 375 tis. Kč, intervalově od 338 do 411 tis. Kč

c) S každými 10 000 najetými km klesne cena o 32 tis. Kč, intervalově od 25,1 do 38,6 tis. Kč

d) Testujeme hypotézu H_0: $\beta_1=0$ vs.H_1: $\beta_1\neq0$. |T| = 9,5, p-hod.=$1{,}3.10^{-12}$. H_0 zamítáme, tj. regresní koeficient je významný, proto cena auta závisí na počtu najetých km.

e) Testujeme hypotézu H_0: $\beta_1=0$ vs. H_1: $\beta_1<0$. Pro levostranný test použijeme příkaz *lmbeta.test(RegModel.1, beta.null=0, alternative="less")*:

```
> lmbeta.test(RegModel.1, beta.null=0, alternative="less")

              Estimate Std. Error Conf. Alternative Estim. Low  Estim. Up Beta
H0   t value      p value
(Intercept) 374.748379 18.1188041  0.95         less        -Inf 405.137676
0       20.682843    1.000000e+00
```

```
najeto        -3.186725   0.3353081   0.95         less        -Inf   -2.624339
0       -9.503872    6.521548e-13
> qt(0.05,48)
[1] -1.677224
>
```

Závěr: T = -9,5 < -1,67, p-hod.=6,5.10^{-13}, proto H_0 zamítáme, což znamená, že cena auta významně klesá s počtem najetých kilomertů.

f) Bodovou a intervalovou předpověď střední ceny při 30 tis. najetými km dostaneme příkazem *predict(RegModel.1, newdata=data.frame(najeto=30), interval= "* confidence*")*:

```
> predict(RegModel.1, newdata=data.frame(najeto=30), interval=" confidence )
          fit      lwr      upr
[1,] 279.1466 258.0078 300.2854
```

Závěr: Průměrná cena aut s najetými 30 tis.km je 279 tis.Kč, intervalově od 258 do 300 tis. Kč.

g) Bodovou i intervalovou predikci ceny vašeho auta dostaneme příkazem *predict(RegModel.1, newdata=data.frame(najeto=30), interval="prediction")*:

```
> predict(RegModel.1, newdata=data.frame(najeto=30), interval="prediction")
          fit      lwr      upr
[1,] 279.1466 152.6423 405.6509
>
```

Závěr: Cena Vašeho auta by měla být 279 tis.Kč, intervalově od 152,6 do 405,6 tis.Kč. Interval je širší (méně přesný). Vaše představa o ceně je podhodnocená.

h) Ověření předpokladu pro korektnost použití statistické indukce v regresi: Výpočetně testujeme normalitu reziduí SWT, homoskedasticitu reziduí Levenovo nebo Breusch-Paganovým testem a nezávislost reziduí Durbinovo-Watsonovým testem. Příkazy:

shapiro.test(residuals(RegModel.1))

skupiny <- (ojetiny$najeto >= median(ojetiny$najeto))

levene.var.test(residuals(RegModel.1)~skupiny)

bptest(cena ~ najeto, studentize=FALSE, data=ojetiny)

dwtest(cena ~ najeto, alternative="two.sided", data=ojetiny)

Výstupy:

```
> shapiro.test(residuals(RegModel.1))
        Shapiro-Wilk normality test
data:  residuals(RegModel.1)
```

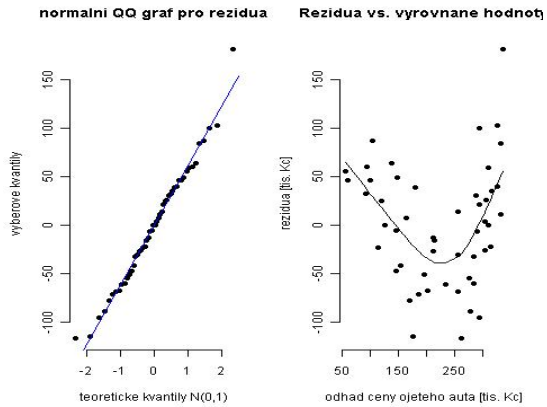
```
W = 0.9848, p-value = 0.7648

> skupiny <- (ojetiny$najeto >= median(ojetiny$najeto))
> levene.var.test(residuals(RegModel.1)~skupiny)
        Levene test of homogeneity of variances
data:  residuals(RegModel.1) by skupiny
Levene's F = 0.4992, num df = 1, denom df = 48, p-value = 0.4832
> qf(0.95,1,48)
[1] 4.042652

> bptest(cena ~ najeto, studentize=FALSE, data=ojetiny)
       Breusch-Pagan test
data:  cena ~ najeto
BP = 2.0503, df = 1, p-value = 0.1522

> dwtest(cena ~ najeto, alternative="two.sided", data=ojetiny)
       Durbin-Watson test
data:  cena ~ najeto
DW = 2.2445, p-value = 0.3847
alternative hypothesis: true autocorelation is not 0
```

p-hodnoty všech těchto testů jsou vysoké. Proto nezamítáme nulové hypotézy o normalitě, homoskedasticitě a nezávislosti reziduí. Podmínky na použití statistické indukce v regresi jsou splněny. Z grafických ověřovacích metod použijme QQ-diagram a bodový diagram závislosti reziduí na vyrovnaných hodnotách. Použijeme příkazy:

```
par(mfrow=c(1,2), bty="n")
qqnorm(residuals(RegModel.1), main="normalni QQ graf pro rezidua", xlab="teoreticke kvantily N(0,1)", ylab="vyberove kvantily", pch=16)
abline(a=mean(residuals(RegModel.1)), b=sd(residuals(RegModel.1)), col="blue")
resplot(RegModel.1,"e-yhat", lowess=T, main="Rezidua vs. vyrovnane hodnoty", xlab="odhad ceny ojeteho auta [tis. Kc]", ylab="rezidua [tis. Kc]", pch=16)
```

Výstup:

QQ-diagram potvrzuje normalitu a bodový diagram reziduí ukazuje, že na vyrovnání dat nebyla použita optimální regresní funkce (residua by měla náhodně kolísat okolo nuly). Lepší výsledky než regresní přímka by dala regresní hyperbola.



3) Porovnání regresních modelů: Načteme data do R. Následujícími příkazy vytvoříme a vypíšeme výsledky pro jednotlivé regresní modely:

```
model01 <- lm(cena~najeto, data=ojetiny2)
model02 <- lm(cena~I(sqrt(najeto)), data= ojetiny2)
model03 <- lm(cena~I(1/najeto), data= ojetiny2)
model04 <- lm(cena~I(log(najeto)), data= ojetiny2)
model05 <- lm(cena~najeto+I(najeto^2), data= ojetiny2)

summary(model01)
summary(model02)
summary(model03)
summary(model04)
summary(model05)
```

Z jednotlivých výstupů můžeme shrnout tyto výsledky pro odhadnuté funkce:

a) $y = 458,5 – 5,8x + e$, $R^2 = 0,762$, $s = 65,4$

b) $y = 634,7 - 67,5\sqrt{x} + e$, $R^2 = 0,850$, $s = 51,9$

c) $y = 130,4+2689,7/x + e$ $R^2 = 0,791$, $s = 61,27$

d) $y = 851,5-169,7\ln(x)+e$ $R^2 = 0,890$, $s = 44,42$

e) $y = 589,9-15,8x+0,134x^2+e$ $R^2 = 0,913$, $s = 39,9$

K sestrojení bodových diagramů použijeme příkazy:

```
par(mfrow=c(3,2))
usek <- seq(5,75,by=0.1)
```

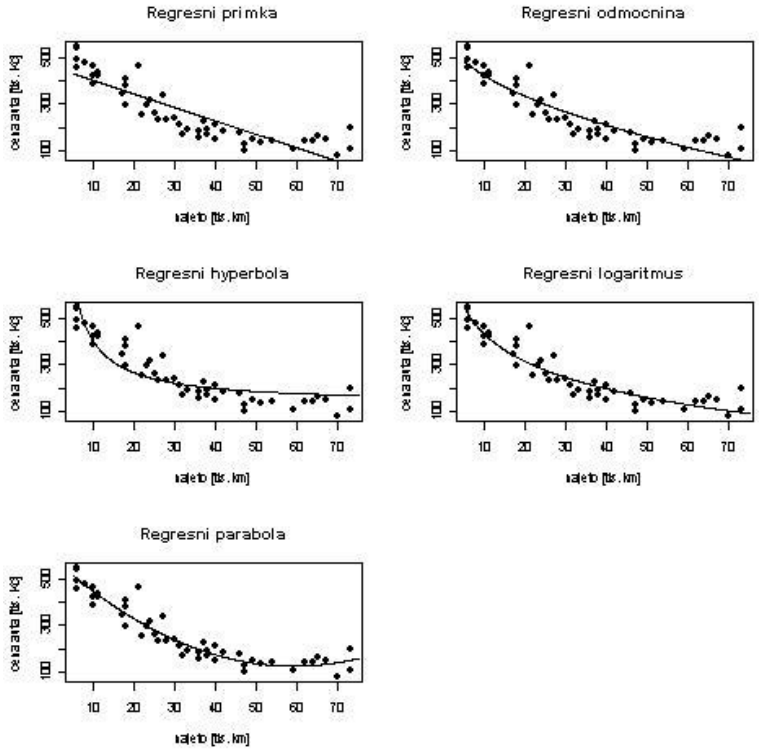
plot(ojetiny2$cena~ ojetiny2$najeto, pch=16, main="Regresni primka", ylab="cena auta [tis. Kc]", xlab="najeto [tis. km]")
lines(usek, predict(model01, newdata=data.frame(najeto=usek)))

plot(ojetiny2$cena~ ojetiny2$najeto, pch=16, main="Regresni odmocnina", ylab="cena auta [tis. Kc]", xlab="najeto [tis. km]")
lines(usek, predict(model02, newdata=data.frame(najeto=usek)))

plot(ojetiny2$cena~ ojetiny2$najeto, pch=16, main="Regresni hyperbola", ylab="cena auta [tis. Kc]", xlab="najeto [tis. km]")
lines(usek, predict(model03, newdata=data.frame(najeto=usek)))

plot(ojetiny2$cena~ ojetiny2$najeto, pch=16, main="Regresni logaritmus", ylab="cena auta [tis. Kc]", xlab="najeto [tis. km]")
lines(usek, predict(model04, newdata=data.frame(najeto=usek)))
plot(ojetiny2$cena~ojetiny2$najeto, pch=16, main="Regresni parabola", ylab="cena auta [tis. Kc]", xlab="najeto [tis. km]")
lines(usek, predict(model05, newdata=data.frame(najeto=usek)))

Výstup:



Závěr: Nejlepší vyrovnání poskytuje regresní parabola (kvadratická funkce).

Kapitola 11: Vícerozměrná regrese

Klíčové pojmy:

vícerozměrná regrese, obecný lineární model, absolutní a relativní pružnosti, elasticity, B-keficienty, reziduální rozptyl, standardní normální model, F-testy v regresních modelech, koeficient (index) mnohonásobné determinace, koeficient parciální a mnohonásobné korelace, korigovaný koeficient determinace, intervaly spolehlivosti pro korelační koeficient, testy o korelačních koeficientech, multikolinearita, umělé proměnné v regresi

Cíle kapitoly:

- **pochopení pojmu vícenásobná regrese a korelace;**
- **porozumění základním cílům strategie analýz závislostí numerických proměnných;**
- **znalost metod odhadů vícerměrných modelů;**
- **naučit se pomocí umělých proměnných zahrnovat mezi regresory i nominální proměnné.**

Čas potřebný ke studiu kapitoly: 13 hodin

Výklad:

Nastínění obsahu kapitoly.

Vícerozměrná regrese

Interpretace regresních koeficientů

Statistická indukce v regresní analýze

Použití modelu na předpověď

Ukazatele síly vícerozměrné lineární závislosti

Statistická indukce v korelační analýze

Ověřování podmínek SLRM

Umělé proměnné v regresi

Struktura výkladu

Durante causa, durant effectus – cesante causa, cessant effectus
Dokud trvá příčina, trvá důsledek – ustává-li příčina, ustává také důsledek
Bacon

Vícerozměrná regrese

Obecný lineární model:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_p x_{pj} + \varepsilon_j,\ j = 1,2,...,n;$$

- obsahuje p vysvětlujících proměnných (regresorů), p+1 neznámých parciálních regresních parametrů a náhodnou složku;
- $\beta_1, \beta_2, \ldots, \beta_p$ nazýváme dílčí (parciální) regresní koeficienty (tzv regresní nadrovina).

Příklad: Pro p = 2 - závislost výnosů y na ceně x_1 a nákladech na reklamu x_2.

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_j, j = 1,2,...,n,$$

- Jde o tzv. regresní rovinu.
- Přesnější označení regresních koeficientů: $\beta_{yx1.x2}$, $\beta_{yx2.x1}$.

Regresní parametry odhadujeme opět MNČ.

- Jejich ruční výpočet je komplikovaný (opírá se o maticový zápis).
- Používáme proto vhodný statistický program.

Odhadnutý regresní model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

Interpretace dílčích parciálních regresních koeficientů b_j a použití výsledků k analýze:

- Když se zvětší x_i (i = 1,…,p) o jednotku a ostatní vysvětlující proměnné zůstanou stejné, udává b_i, o kolik se v průměru změnila hodnota vysvětlované proměnné y.
- Ekonomický význam
  - Udávají empirické absolutní pružnosti: $b_i = \frac{\partial y}{\partial x_i}$.
  - Lze pomocí nich počítat okamžité relativní pružnosti (elasticity):

    $$E_i = \frac{\partial y}{\partial x_i} : \frac{y}{x_i} = \frac{\partial \ln y}{\partial \ln x_i} = \beta_i \frac{x_i}{y}, i = 1,..., p.$$

  - Protože definici relativní pružnosti lze numericky aproximovat vzorcem (tzv. empirické elasticity):

    $$E_i \approx \frac{\Delta y}{\Delta x_i} : \frac{y}{x_i} = \frac{\frac{\Delta y}{y} 100}{\frac{\Delta x_i}{x_i} 100}, i = 1,..., p,$$

    plyne odtud, že relativní pružnosti vyjadřují separovanou procentuální změnu vysvětlované proměnné y odpovídající jednotkové procentuální změně vysvětlující proměnné X_i.

<u>B-koeficienty:</u>

- Pro i= 1,...,p jsou definovány vzorci $B_{yx_i} = b_i \frac{s_{x_i}}{s_y}$.

- Představují též dílčí korelační koeficienty.
- Lze je také počítat jako
    - parciální regresní koeficienty mezi odpovídajícími standardizovanými proměnnými,
    - nebo pomocí párových korelačních koeficientů.
- Používáme je k výpočtu veličin $\frac{|B_{yx_i}|}{\sum_{i=1}^{p}|B_{yx_i}|}$,

  které udávají, jak se podílí změny jednotlivých vysvětlujících proměnných na variabilitě (změnách) vysvětlované proměnné.
    - Intenzitu vlivu jednotlivých proměnných lze vyjádřit v procentech.
    - Příklad na interpretaci - Hindls (2007), s. 218-219 nebo Stuchlý (1999b), s. 52.

Odhad rozptylu náhodné složky:

- Provádíme opět reziduálním rozptylem $s^2 = \frac{1}{n-p-1}\sum_{j=1}^{n} e_j^2$

  kde e_j jsou rezidua (rozdíly naměřených a odhadnutých hodnot vysvětlované proměnné).
- Tento odhad je nestranným odhadem.
- Další regresní analýza se provádí obdobně jako v modelu regresní přímky.
    - V dalším upozorníme na případné rozdíly.

Statistická indukce v regresní analýze

Standardní regresní model:

Splňuje podmínky standardizace, tj. pro j = 1,2,...,n platí:

1) náhodné složky ε_j mají normální rozdělení (normalita),

2) $E(\varepsilon_j) = 0$ (vhodnost lineárního modelu - kolísání chyb kolem nuly),

3) $D(\varepsilon_j) = \sigma^2$ (homoskedasticita),

4) $\text{cov}(\varepsilon_j,\varepsilon_k) = 0 \ \forall \ j \neq k$ (nezávislost, resp. nekorelovanost chyb),

5) proměnné x_i (i = 0,1,...,p) a x_0 (vektor jedniček) jsou nenáhodné a vzájemně lineárně nezávislé (v modelu není multikolinearita).

Potom MNČ-odhad je nejlepší lineární nestranný odhad (BLUE) regresních parametrů a statistickou indukci (intervaly spolehlivosti a testy) můžeme provádět obdobně jako v modelu regresní přímky.

100(1-α)% intervaly spolehlivosti pro regresní parametry:

$$P(b_i - t_{1-\alpha/2}(n-p-1)\, s(b_i) \le \beta_i \le b_i + t_{1-\alpha/2}(n-p-1)\, s(b_i)) = 1 - \alpha, \quad i = 0,1,...,p.$$

Testy o regresních parametrech:

- Pro i = 0,1,...,p testujeme hypotézy H_0: $\beta_i = 0$ proti alternativním hypotézám H_1: $\beta_i \neq 0$ na hladině významnosti α.
- H_0 zamítáme na kritickém oboru $W = \{T = b_i/s(b_i): |T| > t_{1-\alpha/2}(n-p-1)\}$.
  - Nezamítnutí H_0 interpretujeme jako statistickou nevýznamnost regresního parametru β_i.
  - Znamená to, že na proměnnou Y buď nepůsobí významně proměnná X_i (je jí třeba z modelu vypustit) nebo nemáme vhodná data.
- Testy je možno zobecnit.

Celkový F-test o regresním modelu:

Testujeme hypotézu H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$ proti alternativní hypotéze H_1, že aspoň jeden z těchto koeficientů je nenulový. Jde o celkový F-test.

Testové kritérium:
$$F = \frac{\frac{S_T}{p}}{\frac{S_R}{n-p-1}} = \frac{\sum(\hat{y}_j - \bar{y})^2 / p}{\sum(\hat{y}_j - y_j)^2 / (n-p-1)}$$

- Zde S_T je teoretický či regresí vysvětlený a S_R je reziduální součet čtverců. Jsou počítány v speciální analýze rozptylu.
- Kritický obor: $W = \{F: F > F_{1-\alpha}(p;n-p-1)\}$.

Na postupném vynechávání nevýznamných regresorů je založena tzv. metoda stupňovité regrese.

Zobecnění celkového F-testu:

V modelech s více vysvětlujícími proměnnými je možno uvedený test zobecnit tak, že testujeme hypotézu, že jen posledních *r* parametrů je statisticky nevýznamných, tj. hypotézu H_0: $\beta_p = \beta_{p-1} = ... = \beta_{p-r+1} = 0$ proti opačné alternativě H_1.

Označme R_r^2 koeficient determinace redukovaného modelu.

K testování použijeme testové kritérium

$$F = \frac{(R^2 - R_r^2)/r}{(1-R^2)/(n-p-1)} \sim F(r, n-p-1)$$

Kritický obor: $W = \{F: F > F_{1-\alpha}(r; n-p-1)\}$.

Předpovědi v regresním modelu

Bodová předpověď pro $X_1=x_{1*}$, $X_2=x_{2*}$,…, $X_p=x_{p*}$ (kde hvězdičkované hodnoty jsou hodnoty regresorů, v kterých počítáme předpověď) je

$$y_* = b_0+b_1x_{1*}+b_2x_{2*}+...+b_px_{p*}$$

Predikční interval pro jednotlivé hodnoty Y, resp. konfidenční interval pro E(Y), počítáme pomocí statistického programu jako u regresní přímky, jen musíme zadat hodnoty všech regresorů (hvězdičkové hodnoty).

Ukazatelé síly vícenásobné lineární závislosti

Koeficient (index) mnohonásobné determinace

$$R^2 = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2} = 1 - \frac{\sum e_j^2}{\sum(y_j - \bar{y})^2}.$$

Po převedení na procenta udává, kolik procent změn vysvětlované proměnné lze vysvětlit změnami vysvětlujících proměnných.

Koeficient mnohonásobné korelace r:

Dostaneme ho odmocněním koeficientu mnohonásobné determinace. Platí $0 \leq r \leq 1$.

Vícenásobná lineární závislost se popisuje i dílčí korelační koeficienty - viz Hindls (2007), s. 220-222.

Statistická indukce v korelační analýze

Test významnosti koeficientu determinace:

Je ekvivalentní s testem o významnosti celého modelu.

Testujeme hypotézu H_0: $R^2 = 0$ proti alternativě H_1: $R^2 \neq 0$.

Testové kritérium: $$F = \frac{R^2 / p}{(1 - R^2)/(n - p - 1)}.$$

Kritický obor: $W = \{F: F > F_{1-\alpha}(p; n-p-1)\}$.

Korigovaný (adjustovaný) koeficient determinace:

$$\overline{R}^2 = 1 - \frac{\sum_{j=1}^{n} (y_j - \hat{y})^2 / (n - p - 1)}{\sum_{j=1}^{n} (y_j - \bar{y})^2 / (n - 1)}$$

Platí $\overline{R}^2 = R^2 - (1 - R^2)\frac{p}{n - p - 1}$ a odtud dostáváme, že platí $\overline{R}^2 \leq R^2$.

Při zvyšování počtu vysvětlujících proměnných se automaticky zvyšuje i R^2, i když kvalita modelu se nemusí zlepšovat. Proto při porovnávání kvality modelů s různým počtem proměnných je lepší používat korigovaný koeficient determinace.

Intervaly spolehlivosti pro korelační koeficient:

Viz Hindls (2007), s. 230-232.

Testy o korelačních koeficientech:

Viz Hindls (2007), s. 234-238.

Příklady: Viz Stuchlý (1999b), s. 50-58.

V ekonomické praxi potřebujeme často odhadnout i nelineární regresní model (např. Cobb-Douglasovu produkční funkci). Obvykle používáme k tomu logaritmickou transformaci. Viz Hindls (2007), s. 223-224.

- Pokud nelze model linearizovat – používáme nelineární MNČ (např. v R).

Ověřování podmínek SNRM

Ověřování provádíme obdobně jako u regresní přímky. Navíc musíme ověřit, zda v modelu není multikolinearita.

Pokud je mezi regresory lineární závislost, říkáme, že v modelu je perfektní multikolinearita. Model MNČ pak nelze odhadnout (závislý regresor musíme vynechat).

Jsou-li regresory silně skorelované, je v modelu silná multikolinearita a odhad získaný MNČ nemá dobré statistické vlastnosti. Viz Hindls (2009), s. 224-226.

Umělé proměnné v regresi

Vícerozměrná regrese s kvantitativními i kvalitativními regresory:

Používá se v analýze dat ke zkoumání závislosti numerické proměnné na numerických i nominálních proměnných.

Pokud dáváme do lineárního modelu více vysvětlujících proměnných (regresorů), rozhoduje o tom, které proměnné do modelu zařadit, příslušný párový korelační koeficient mezi vysvětlovanou proměnou (Y) a regresorem (určíme ho z korelační matice).

Při zařazování kvalitativní proměnné (např. pohlaví, vzdělání) používáme umělé proměnné (UmP) pro její úrovně. Nabývají hodnoty 1, pokud kvalitativní proměnná nabude této úrovně a hodnotu 0 v opačném případě. Aby v regresním modelu nebyla multikolinearita (lineární závislost regresorů), musí být počet UmP roven počtu úrovní minus jedna.

Úroveň s vynechanou umělou proměnnou nazýváme referenční (obvykle první nebo poslední úroveň).

Odhadnuté regresní koeficienty u umělých proměnných udávají, o kolik se změní průměrná hodnota Y, když úroveň UmP přešla z referenční úrovně na uvažovanou úroveň. Např. vzdělání se změnilo ze ZS na SS.

Do modelu můžeme zahrnovat i interakce (spolupůsobení) kvantitativní s kvalitativní proměnnou (odhadnutý koeficient u intraktivní proměnné se interpretuje jako změna směrnice při dané úrovni kvalitativní proměnné).

Příklady: Viz Stuchlý (2000), s. 49-53 a úkoly řešené na konci kapitoly.

Vícerozměrná regrese v R

Bodový diagram s vyrovnáním MNČ (viz Úkol 3)

```
hodnoty <- 0:cislo
plot(data$x,data$y,pch=16,xlab=„“,ylab=„“,main=„ “,ylim=c( , ),col=„barva“)
lines(hodnoty, predict(model, newdata=data.frame(x1= ,x2= )), col=„barva“)
points(data$x,y,pch= ,xlab=„ “,ylab=„ “,main=„ “,ylim=c( , ),col=„barva“), le-
gend( , ,legend=c(„ “,„ “,…), col=c(„ “,„ “,…),pch=16)
```

Regresní rovina:

```
lm(y~x1+x2, data=název)
```

Model regresní roviny s interakcemi:

```
lm(y~x1*x2, data= )
```

Vícerozměrná regrese:

```
lm(y~x1+x2+x3+…, data= )
```

Testování podmodelu:

```
anova(submodel,model)
```

Durbinův-Watsonův test:

```
durbin.watson(residuals(model))
```

Reziduální grafy:

```
resplot(model, „e-yhat“,…)
```

```
resplot(model, „e-x“, xterm=„x1“,…)
```

Všechny regresní modely lze také získat interaktivně ze *Statistics-Fit model-Linear model…*

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 213-226, 230-241. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika II Cvičení ze statistických metod pro manažery*. J. Hradec: VŠE, 1999. S. 23-25, 47-60. ISBN 80-7079-035-0.

Doporučené studijní zdroje:

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 138-146, 154-160. ISBN 80-7169-255-7.

HINDLS, R. a kol. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000. S. 77-85. ISBN 80-7261-013-9.

JAROŠOVÁ, E. *Statistika B. Řešené příklady.* Praha: VŠE, 1994. S. 46, 76-87, 106-126. ISBN 80-7079-328-7.

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 239-258, 265-275. ISBN 978-80-86446-40.

MINAŘÍK, B. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita, 1995. S. 124-136. ISBN 80-7157-166-0.

SEGER, J. a R. HINDLS. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 219-236, 240-241, 244-253. ISBN 80-7187-058-7.

STUCHLÝ, J. *Ekonometrie*. J. Hradec: VŠE, 2000. S. 31-41, 49-56.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/).

WONNACOT, T. H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 431-470, 537-556. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Lékař léčí určitou nemoc dvěma druhy léku. Domnívá se, že pokud budou pacienti užívat oba léky společně ale v různých dávkách, potom se zkrátí počet hodin léčby. Lékař se rozhodne ověřit svůj předpoklad, a aby zachoval stejné podmínky experimentu, umístí v nemocnici 16 náhodně vybraných pacientů s danou nemocí a začne podávat léky formou injekcí se stanovenými dávkami v ml. Tyto údaje pečlivě zaznamenává spolu s celkovým počtem hodin léčby, po kterých je pacient opět zdravý. Údaje jsou uvedeny v souboru lecba.xlsx. a) MNČ odhadněte závislost počtu hodin léčby z na množství prvního léku x v ml a množství druhého léku y v ml a interpretujte získané regresní koeficienty. b) Určete a interpretujte standardní chybu modelu. c) Testujte statistickou významnost obou parciálních regresních koeficientů, interpretujte jejich standardní chyby a testujte model jako celek. d) Určete intervaly spolehlivosti pro parciální regresní koeficienty. e) Určete, jak se na léčbě podílely jednotlivé léky. f) Určete a interpretujte koeficient mnohonásobné determinace počtu hodin léčby. Vypočítejte i adjustovaný koeficient determinace. g) Proveďte bodovou a intervalovou predikci počtu hodin léčby pacienta a průměrného počtu hodin léčby pacienta při dávce prvního léku x = 1 ml a dávce druhého léku y = 2 ml.

2) Použijeme data ze souboru ojetiny2.dat. Na základě modelu vícenásobné regrese, kdy vysvětlovaná proměnná cena v tisících Kč a obě vysvětlující proměnné (počet najetých kilometrů v tisících km a stáří auta v měsících) vstupují do modelu lineárně, proveďte následující kroky: a) Odhadněte bodově i intervalově průměrnou cenu nového auta. b) Odhadněte bodově i intervalově, jak se liší průměrná cena stejně starých aut, pokud jedno auto má najeto o 5000 km více než druhé. c) Otestujte zda cena auta klesá (i)

s počtem najetých km (po vyloučení vlivu stáří auta); (ii) se stářím auta (po vyloučení vlivu najetých km). d) Rozhodněte, zda průměrná cena aut, která po zakoupení stojí pouze v garáži, klesá s každým rokem o 40 tisíc Kč. e) Odhadněte bodově i intervalově cenu vašeho auta, které má najeto 30 tisíc km a je staré 2 roky. f) Ověřte předpoklady regresní analýzy.

3) V souboru *platy.dat* jsou k dispozici údaje o platech (výše měsíční mzdy v Kč) u 100 náhodně zvolených zaměstnanců velké firmy. Kromě výše platu se v datovém zaznamu uvádí rovněž počet odpracovaných let u firmy a dosažené vzdělání zaměstnance (ZS – základoškolské, SS – středoškolské a VS - vysokoškolské). a) Odhadněte funkční předpis závislosti platu zaměstnance (i) se ZS vzděláním, (ii) se SS vzděláním, (iii) s VS vzděláním. b) Odhadněte bodově i intervalově průměrný přírůstek platu za každý odpracovaný rok za předpokladu, že zaměstnanec již při zaměstnání nestuduje. c) Otestujte, zda plat ve firmě po vyloučení vlivu vzdělání roste s počtem odpracovaných let ve firmě. d) Otestujte, zda je rozdíl mezi průměrnými platy ZS a (i) SS, (ii) VS je statisticky významný. V případě že ano, odhadněte bodově i intervalově tento rozdíl. e) Otestujte, zda plat ve firmě po vyloučení vlivu odpracovaných let závisí na vzdělání zaměstnance.

? Úkoly k zamyšlení a diskuzi

1) Zamyslete se nad tím, jak zapsat výsledky ve vícenásobné regresi a korelaci pomocí matic.

2) Uvažujte o souvislosti JAR a jednoduché lineární regrese, která má za vysvětlující proměnnou kvalitativní veličinu.

Klíč k řešení otázek:

1) Regresní rovina v Excelu: a) $\hat{z} = 46{,}8973 - 1{,}4528x - 1{,}3702y$, b) 2,52, c) významné, d) $-2{,}68 \leq \beta_1 \leq -0{,}22$, $-2{,}00 \leq \beta_2 \leq -0{,}74$, e) 35,3% a 64,7%, f) 0,68 a 0,63, g) $42{,}70 \pm 2{,}99$.

Podrobný výpočet – viz Stuchlý (1999b), s. 50-53. Základní výsledky získáme v Excelu použitím Regrese z Analýzy dat:

x_i	y_i	z_i
1	2	46
1	4	40
1	6	38
1	8	35
2	2	41
2	4	36
2	6	35
2	8	33
3	2	39
3	4	34
3	6	30
3	8	33
4	2	42
4	4	33
4	6	34
4	7	35

VÝSLEDEK

Regresní statistika

Násobné R	0,824923
Hodnota spolehlivosti R	0,680498
Nastavená hodnota spolehlivosti R	0,631344
Chyba stř. hodnoty	2,527853
Pozorování	16

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	176,9294604	88,46473	13,844156	0,000601283
Rezidua	13	83,07053959	6,390042		
Celkem	15	260			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*
Hranice	46,89726	2,153913927	21,77304	1,297E-11	42,24401215
xi	-1,45276	0,565669912	-2,56822	0,0233725	-2,67482033
yi	-1,3702	0,292232475	-4,68872	0,0004238	-2,001527006

2) Regresní rovina v R: Odhadujeme regresní rovinu s rovnicí $E(Y) = \beta_0+\beta_1x_1+\beta_2x_2$, kde Y cena auta (tis. Kč), x_1 je najeto (roků) a x_2 je stáří auta (měsíců). Načteme měření do datového rámce *data*. Obvykle začínáme výpočtem korelační matice (*Statistics-Summaries-Correlation matrix*; podržíme Ctrl a zaškrtneme cena, najeto, stari):

```
> cor(data[,c("cena","najeto","stari")], use="complete.obs")
             cena     najeto      stari
cena    1.0000000 -0.8728738 -0.6733943
najeto -0.8728738  1.0000000  0.3666705
stari  -0.6733943  0.3666705  1.0000000
```

Závěr: Mezi cenou a oběma regresory (najeto a stáří) je dosti silná lineární nepřímá závislost (-0,873; -0,673). Mezi oběma regresory není významná multikolinearita (r = 0,367).

Interpretace regresních parametrů: Použijeme příkazy *library(vsePackage)*, *model.v1 <- lm(cena~najeto+stari, data=data)* a *lmbeta.test(model.v1))*:

```
> lmbeta.test(model.v1)
              Estimate Std. Error Conf. Alternative Estim. Low  Estim. Up
(Intercept) 525.725902 13.8625756  0.95   two.sided 497.837997 553.613807
najeto       -4.815436  0.3198214  0.95   two.sided  -5.458834  -4.172039
stari        -3.611565  0.4249370  0.95   two.sided  -4.466428  -2.756702
            Beta H0   t value      p value
(Intercept)       0  37.92411 6.719919e-37
najeto            0 -15.05664 1.321060e-19
```

```
stari             0  -8.49906 4.631383e-11
```

a) Průměrná cena nového auta je 525,7 tis.Kč, tj. od 498,8 do 553,6 tis. Kč.

b) O 5x4,82 = 24,1 tis. Kč.

c) Levostranné testy o regresních koeficientech: Použijeme příkazy *lmbeta.test(model.v1, beta.null=0, alternative="less")* a *qt(0.95,47)*:

```
             Estimate Std. Error Conf. Alternative Estim. Low  Estim. Up
(Intercept) 525.725902 13.8625756  0.95        less       -Inf 548.986289
najeto       -4.815436  0.3198214  0.95        less       -Inf  -4.278800
stari        -3.611565  0.4249370  0.95        less       -Inf  -2.898552
            Beta H0   t value      p value
(Intercept)       0  37.92411 1.000000e+00
najeto            0 -15.05664 6.605298e-20
stari             0  -8.49906 2.315692e-11
> qt(0.95,47)
[1] 1.677927
```

Závěr: (i) Testujeme H_0: $\beta_1=0$ vs. H_1:$\beta_1<0$, T= -15, p-hodnota = $6{,}61.10^{-20}$, (ii) Testujeme H_0: $\beta_2=0$ vs. H_1:$\beta_2<0$, T=-8,5, p-hodnota = $2{,}32.10^{-11}$. V obou případech H_0 zamítáme, tj, cena významně klesá s počtem najetých km i se stářím auta.

d) Testujeme hypotézu H_0: $\beta_2 = -40/12$ (měsíční pokles ceny v tis. Kč) proti H_1: $\beta_2 \neq -3{,}3$. Použijeme příkaz *lmbeta.test(model.v1, beta.null=-40/12)*:

```
> lmbeta.test(model.v1, beta.null=-40/12)
             Estimate Std. Error Conf. Alternative Estim. Low  Estim. Up
(Intercept) 525.725902 13.8625756  0.95   two.sided 497.837997 553.613807
najeto       -4.815436  0.3198214  0.95   two.sided  -5.458834  -4.172039
stari        -3.611565  0.4249370  0.95   two.sided  -4.466428  -2.756702
              Beta H0    t value      p value
(Intercept) -3.333333 38.1645699 5.038805e-37
najeto      -3.333333 -4.6341587 2.863838e-05
stari       -3.333333 -0.6547604 5.158137e-01
```

Závěr: T=-0,655, p-hodnota = 0,51, tj. H_0 nezamítáme, tj. cena auta s každým rokem klesá o 40 tis.Kč (s každým měsícem klesá o 40/12 = 3,3 tis. Kč).

e) Bodová a intervalová predikce: Použijeme příkaz *predict(model.v1, newdata = data.frame(najeto=30, stari=24), interval="prediction")*:

```
>predict(model.v1,newdata=data.frame(najeto=30,stari=24),interval=      "pre-
diction")
       fit      lwr      upr
1 294.5852 210.1766 378.9938
```

Závěr: Předpověď ceny auta, které má najeto 30 000 km a je staré 2 roky je 294,6 tis. Kč, tj. od 210,2 do 379 tis. Kč.

f) Ověření předpokladů regresní analýzy: Použijeme příkazy

shapiro.test(residuals(model.v1))
skup1 <- (data$najeto >= median(ojetiny2$najeto))
skup2 <- (data$stari >= median(ojetiny2$stari))
levene.var.test(residuals(model.v1)~skup1)
qf(0.95,1,48)
levene.var.test(residuals(model.v1)~skup2)

Výstupy:

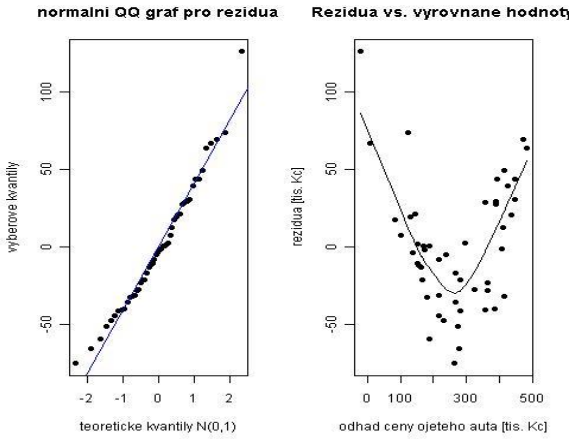
```
> shapiro.test(residuals(model.v1))
       Shapiro-Wilk normality test
data:  residuals(model.v1)
W = 0.9725, p-value = 0.2904

> skup1 <- (ojetiny2$najeto >= median(data$najeto))
> skup2 <- (ojetiny2$stari >= median(data$stari))
> levene.var.test(residuals(model.v1)~skup1)
       Levene test of homogeneity of variances
data:  residuals(model.v1) by skup1
Levene's F = 0.0144, num df = 1, denom df = 48, p-value = 0.905

> levene.var.test(residuals(model.v1)~skup2)
       Levene test of homogeneity of variances
data:  residuals(model.v1) by skup2
Levene's F = 1.9953, num df = 1, denom df = 48, p-value = 0.1642

> dwtest(cena ~ najeto + stari, alternative="two.sided", data=data)
       Durbin-Watson test
data:  cena ~ najeto + stari
DW = 1.8534, p-value = 0.6151
alternative hypothesis: true autocorelation is not 0
```

Závěr: Podmínky SNLM jsou splněny. Grafické ověření:

prumer.r1 <- mean(residuals(model.v1))

odchylka.r1 <- sd(residuals(model.v1))

par(mfrow=c(1,2))
qqnorm(residuals(model.v1), main="normalni QQ graf pro rezidua", xlab="teoreticke kvantily N(0,1)", ylab="vyberove kvantily", pch=16)
abline(a=prumer.r1, b=odchylka.r1, col="blue")
resplot(model.v1,"e-yhat",lowess=T, main="Rezidua vs. vyrovnane hodnoty", xlab="odhad ceny ojeteho auta [tis. Kc]", ylab="rezidua [tis. Kc]", pch=16)
dev.off()



Grafický výstup potvrzuje normalitu a homoskedasticitu reziduí a ukazuje, že použití lineárního modelu není optimální.

3) Jde o model vícerozměrná regrese s kvantitativní i kvalitativní proměnnou (umělé proměnné). Plat je vysvětlován kvantitativní proměnnou odpracováno a kvalitativní proměnnou vzdělání (3 úrovně, použijeme 2 umělé proměnné). Můžeme uvažovat 2 typy modelu: I. Model bez interakcí kvantitativní a kvalitativní proměnné II. model s interakcemi těchto proměnných. Budeme nejdříve uvažovat model I. Ohadujeme model $E(Y) = \beta_0+\beta_1x_1+\beta_2x_2+ \beta_3x_3$, kde Y je plat, x_1 odpracováno, x_2 a x_3 jsou umělé proměnné pro vzdělání SS a VS. Načteme data do souboru *platy* a aktivujeme balík *vsePackage*. K řešení použijeme program:

platy$vzdelani <- factor(platy$vzdelani, levels=c("ZS","SS","VS"))
zs <- subset(platy, platy$vzdelani=="ZS")
ss <- subset(platy, platy$vzdelani=="SS")
vs <- subset(platy, platy$vzdelani=="VS")
model bez interakce

```
### a)
modelbez <- lm(plat~odpracovano+vzdelani, data=platy, x=TRUE)
summary(modelbez)

hodnoty <- 0:36
par(mfrow=c(1,1))
plot(zs$odpracovano, zs$plat, pch=16, xlab="odpracovano [roky]", ylab="plat [Kc]", main="Model bez interakce", ylim=c(20000,34000), col="blue")
lines(hodnoty, predict(modelbez, newdata=data.frame(odpracovano=hodnoty, vzdelani="ZS")), col="blue")

points(ss$odpracovano, ss$plat, pch=16, col="darkgreen")
lines(hodnoty, predict(modelbez, newdata=data.frame(odpracovano=hodnoty, vzdelani="SS")), col="darkgreen")

points(vs$odpracovano, vs$plat, pch=16, col="red")
lines(hodnoty, predict(modelbez, newdata=data.frame(odpracovano=hodnoty, vzdelani="VS")), col="red")

legend(5, 32000, legend=c("VS", "SS", "ZS"), col=c("red", "darkgreen", "blue"), pch=16)

### b)
lmbeta.test(modelbez)

### c)
lmbeta.test(modelbez, beta.null=0, alternative="greater")

### d)
lmbeta.test(modelbez)

### e)
modelbez.kvant <- lm(plat~odpracovano, data=platy)
summary(modelbez.kvant)
anova(modelbez.kvant,modelbez)
```

Výstupy: a)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     19739.256    120.273  164.12   <2e-16 ***
odpracovano       270.677      6.186   43.76   <2e-16 ***
vzdelani[T.SS]   1620.304    150.903   10.74   <2e-16 ***
vzdelani[T.VS]   4663.840    143.069   32.60   <2e-16 ***
```

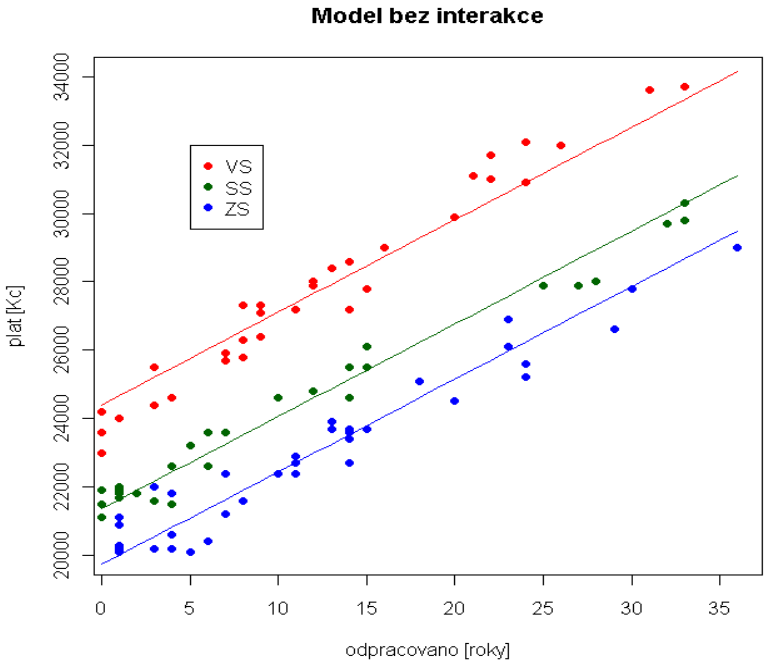
$E(Y)=19739,3+270,7x_1+1\ 620,3x_2+4\ 663,8x_3$

ZS: $E(Y|x_2=0, x_3=0)=19739,3+270,7x_1$

SS: $E(Y|x_2=1,x_3=0)=19739,3+270,7x_1+1620,3=21359,6+270,7x_1$

VS: $E(Y|x_2=0,x_3=1)=19739,3+270,7x_1+4\ 666,8=24403,1+270,7x_1$

Graf:



b)

```
lmbeta.test(modelbez)
                Estimate Std. Error Conf. Alternative Estim. Low  Estim. Up
Beta H0   t value       p value
(Intercept)     19739.2556 120.273059   0.95    two.sided 19500.5155 19977.9957
0 164.12034 2.136450e-119
odpracovano       270.6767   6.186181   0.95    two.sided   258.3973   282.9562
0  43.75506  3.239238e-65
vzdelani[T.SS]   1620.3036 150.903039   0.95    two.sided  1320.7634  1919.8438
0  10.73738  3.911054e-18
vzdelani[T.VS]   4663.8399 143.068946   0.95    two.sided  4379.8503  4947.8295
0  32.59855  1.016458e-53
```

Roční přírůstek platu je 270,7 Kč, intervalově od 258,4 do 283 Kč.

c)

```
> lmbeta.test(modelbez, beta.null=0, alternative="greater")
                Estimate  Std. Error  Conf. Alternative Estim. Low Estim. Up
Beta H0   t value       p value
(Intercept)     19739.2556 120.273059   0.95      greater 19539.4963        Inf
0 164.12034 1.068225e-119
odpracovano       270.6767   6.186181   0.95      greater   260.4022        Inf
0  43.75506  1.619619e-65
vzdelani[T.SS]   1620.3036 150.903039   0.95      greater  1369.6715        Inf
0  10.73738  1.955527e-18
vzdelani[T.VS]   4663.8399 143.068946   0.95      greater  4426.2194        Inf
0  32.59855  5.082288e-54
```

Testujeme H_0: $\beta_1=0$ vs. H_1:$\beta_1>0$, T=43,8, p-hod.=1,62.10^{-65}, H_0 zamítáme, plat ve firmě po vyloučeni vlivu vzdělání roste

d)

SS: Testujeme H_0:$\beta_2=0$ vs. H_1:$\beta_2>0$, T=10,7, p-hod.=1,96.10^{-18}, H_0 zamítáme;
VS: Testujeme H_0:$\beta_3=0$ vs. H_1:$\beta_3>0$, T=33,0, p-hod.=5,08.10^{-54}, H_0 zamítáme.
Je významný rozdíl mezi platy SS a ZS a také je významný rozdíl mezi platy ZS aVS.

e)

```
> modelbez.kvant <- lm(plat~odpracovano, data=data)
anova(modelbez.kvant,modelbez)
Analysis of Variance Table

Model 1: plat ~ odpracovano
Model 2: plat ~ odpracovano + vzdelani
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1     98 431841761
2     96  35225594  2 396616167 540.45 < 2.2e-16 ***
```

Testujeme H_0: $\beta_2=\beta_3=0$ proti opačné alternativě. Je F=540,5, p-hod.=2,2.10^{-16}, H_0 zamítáme, plat po vyloučení vlivu odpracovaných let závisí významně na vzdělání.

Uvedeme dále ještě stručně výstupy řešení pro model II. Odhadujeme regresní model s interakcemi $E(Y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1x_2+\beta_5x_1x_3$.

```
> modelint <- lm(plat ~ odpracovano*vzdelani, data=platy)
> summary(modelint)

Call:
lm(formula = plat ~ odpracovano * vzdelani, data = platy)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                19981.130    134.813 148.213  < 2e-16 ***
odpracovano                  249.101      9.246  26.940  < 2e-16 ***
vzdelani[T.SS]              1567.071    194.035   8.076 2.21e-12 ***
vzdelani[T.VS]              3871.227    201.409  19.221  < 2e-16 ***
odpracovano:vzdelani[T.SS]     3.910     12.956   0.302    0.763
odpracovano:vzdelani[T.VS]    67.697     13.575   4.987 2.80e-06 ***
---
Residual standard error: 531.4 on 94 degrees of freedom
Multiple R-squared: 0.9774,      Adjusted R-squared: 0.9762
F-statistic: 813.8 on 5 and 94 DF,  p-value: < 2.2e-16
```

Regresní funkce pro model s interakcemi:

$E(Y)=19981,1+249,1x_1+1567,1x_2+3871,2x_3+3,9x_1x_2+67,7x_1x_3$

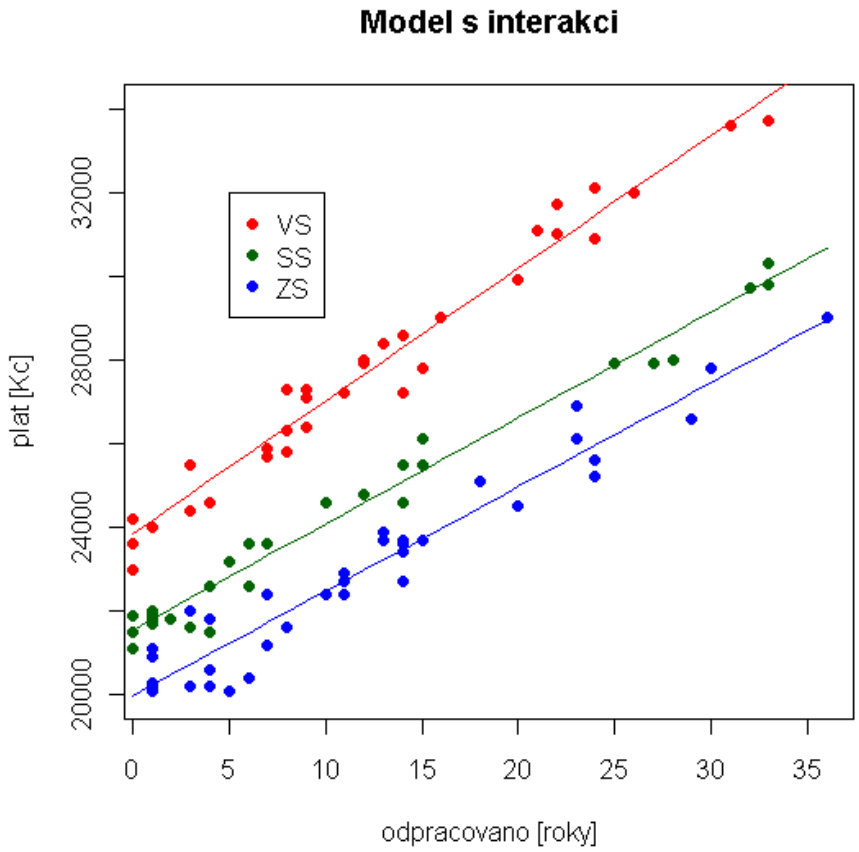
ZS: $E(Y|x_2=0, x_3=0)=19981,1+249,1x_1$

SS: $E(Y|x_2=1,x_3=0)= 19981,1+249,1x_1+1567,1+3,9x_1 = 21548,2+253x_1$

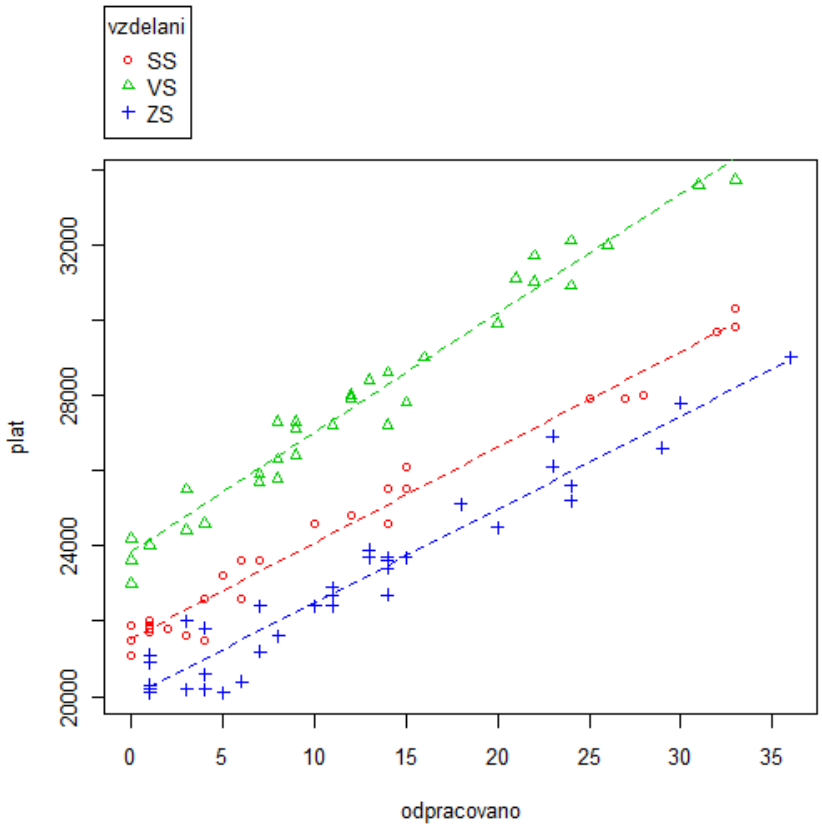
VS: $E(Y|x_2=0,x_3=1)= 19981,1+249,1x_1+3871,2+67,7x_1 = 23852,2+316,8x_1$

Graf:

```
plot(zs$odpracovano, zs$plat, pch=16, xlab="odpracovano [roky]", ylab="plat
[Kc]", main="Model s interakci", ylim=c(20000,34000), col="blue")
lines(hodnoty, predict(modelint, newdata=data.frame(odpracovano=hodnoty, vzde-
lani="ZS")), col="blue")
points(ss$odpracovano, ss$plat, pch=16, col="darkgreen")
lines(hodnoty, predict(modelint, newdata=data.frame(odpracovano=hodnoty, vzde-
lani="SS")), col="darkgreen")
points(vs$odpracovano, vs$plat, pch=16, col="red")
lines(hodnoty, predict(modelint, newdata=data.frame(odpracovano=hodnoty, vzde-
lani="VS")), col="red")
legend(5, 32000, legend=c("VS", "SS", "ZS"), col=c("red", "darkgreen", "blue"),
pch=16)
```



Graf (jen pro model s interakcemi) můžeme získat jednodušším způsobem (interaktivně) v nabídce *Graphs – Scatterplot* po vyplnění vstupního okna: *x-variable*: odpracovano, *y-variable*: plat, odškrtnout: *Marginal Box* a *Smooth line*, stisknout*: Plot by groups…*a potom *OK*. Výstup v *R-Console* má tvar:



Porovnáními modelu bez interakcí s modelem s interakcemi:

```
> anova(modelbez,modelint)
Analysis of Variance Table

Model 1: plat ~ odpracovano + vzdelani
Model 2: plat ~ odpracovano * vzdelani
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     96 35225594
2     94 26539416  2   8686177 15.383 1.662e-06 ***
---
```

Testujeme hypotézu H_0: $\beta_4 = \beta_5 = 0$ proti opačné alternativě H_1. Testovací statistika F = 15,4 a p-hodnota = $1{,}67.10^{-6}$. Tedy H_0 zamítáme, tj. interakce v modelu jsou významné. Model má i vyšší korigovaný koeficient determinace (modelem je vysvětleno 97,6% změn platů).

Kapitola 12: Úvod do analýzy časových řad

Klíčové pojmy:

časová řada, intervalová a okamžiková časová řada, diference, absolutní a relativní přírůstky, tempa růstu, řetězové a bazické indexy, modely časových řad, trendová, sezónní, náhodná složka, trendové funkce, vyrovnání tredové funkce, lineární, kvadratický, exponenciální, modifikovaný exponenciální, logistický trend a gompertzova křivka, střední kvadratická chyba (MSE), prosté a centrované klouzavé průměry

Cíle kapitoly:

pochopení pojmu časová řada a její číselné charakteristiky;

porozumění základním cílům strategie modelování časových řad;

naučit se metodám vyrovnávání časové řady;

být schopni elementárního prognózování časové řady.

Čas potřebný ke studiu kapitoly: 8 hodin

Výklad:

Nastínění obsahu kapitoly.

Časové řady

Číselné charakteristiky časových řad

Složené cenové indexy

Modelování časových řad

Trendové funkce

Klouzavé průměry

Struktura výkladu

Inflace je jako zubní pasta – jakmile jednou vyleze z tuby,
těžko se někomu podaří nacpat ji zpátky.
Karl-Otto Pohl

Časové řady

Časová řada y_t, $t = 1,2,\ldots,n$:

- je posloupnost věcně a prostorově srovnatelných pozorování, která jsou jednoznačně uspořádána v čase.
- Příklady z ekonomie:
    - vývoj HDP, míry inflace, nezaměstnanosti a počtu volných míst, kurzu měny, peněžních zásob, cen akcií, obchodování s akciemi apod.
    - časové řady publikované státní statistikou
        - v statistických ročenkách, statistických přehledech a bulletinech apod.
- Analýza časových řad – soubor metod, které slouží k jejich popisu nebo předvídání jejich budoucího chování.
- Dělení časových řad podle časového hlediska:
    - intervalové (měřené v určitých časových intervalech u stejného objektu),
    - okamžikové či průřezová data (měřené k určitému časovému okamžiku u různých objektů).

Číselné charakteristiky časových řad

Pro intervalové časové řady:

- používáme součty (úhrny) a průměry,
- očisťujeme je od kalendářních variací (Příklad: – Viz Hindls 2007, s. 247-248).

Pro okamžikové časové řady používáme:

- Chronologický průměr: (Příklad: Viz Hindls 2007, s. 248-249)
    - Zpřehlednění časové řady – graf časové řady
- Diference (přírůstky):

$\Delta y_t = y_t - y_{t-1}, \ t = 2,3,...,n$

$\Delta^{(2)} y_t = \Delta y_t - \Delta y_{t-1} = y_t - 2y_{t-1} + y_{t-2}, \ t = 3,4,...,n,$ atd.

- Relativní přírůstky

$$\delta_t = \frac{\Delta y_t}{y_{t-1}}, \ t = 2,3,\cdots,n,$$

- Koeficienty neboli tempa růstu (řetězové indexy)

$$k_t = \frac{y_t}{y_{t-1}}, \ t = 2,3,\cdots,n.$$

- Bazické indexy

$$i_t = \frac{y_t}{y_0}, \ t = 1,2,...,n,$$

kde y_0 je hodnota srovnávané veličiny ve výchozím (bazickém) období (např. indexy cen, inflace apod.).

Průměrné charakteristiky:

- Průměrný absolutní přírůstek $\quad \overline{\Delta} = \frac{1}{n-1}\sum_{t=2}^{n} \Delta y_t = \frac{y_n - y_1}{n-1}.$
- Průměrný koeficient růstu $\quad \bar{k} = \sqrt[n-1]{k_2 k_3 \cdots k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}.$

Další charakteristiky:

- klouzavé úhrny a průměry;
- Příklady – Viz Stuchlý (1999b), s. 63-65.

Složené cenové indexy

Složené cenové indexy jsou objemově vážené indexy.

Laspeyresův index – používá váhy (množství) z běžného období

$$I_p^{(L)} = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

Paascheův index – používá váhy z běžného období

$$I_p^{(P)} = \frac{\sum p_1 q_1}{\sum p_0 q_1}.$$

Zde p jsou ceny a q množství.

Příklad. Racionalizace práce firmy v oblasti úklidu, zásobování a pomocných kancelářských prací.

Hod. sazba v zákl. období v € p_0	Odprac. hodiny v zákl. období q_0	Hod. sazba v běžn. období v € p_1	Odprac. hodiny v běžn. období q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
2	4000	1,9	5500	8000	7600	11000	10450
2,5	2000	3	2100	5000	6000	5250	6300
3,5	6000	3,75	7000	21000	22500	24500	26250
Součet				34000	36100	40750	43000

Laspayresův index hodinových sazeb

$$\frac{\sum p_1q_0}{\sum p_0q_O} = \frac{36100}{34000} = 1{,}062$$

.

Kdyby u firmy bylo v běžném období odpracováno na uvedených pracích stejné množství hodin jako v základním období, pak by náklady na tyto práce stouply v důsledku zvýšení hodinových sazeb o 6,2% (použili jsme jako váhy počty odpracovaných hodin v základním období).

Paasheův index hodinových sazeb

$$\frac{\sum p_1q_1}{\sum p_0q_1} = \frac{43000}{40750} = 1{,}055.$$

Kdyby v běžném období bylo odpracováno na uvedených pracích množství hodin z běžného období, pak by náklady na tyto práce vzrostly o 5,5% (použili jsme jako váhy počty odpracovaných hodin v běžném období). Kompromisem je Fisherův index (geom.průměr).

Modelování časových řad

Modely:

- Aditivní model:
  $y_t = T_t + S_t + C_t + \varepsilon_t$,
  - T_t je trendová, S_t je sezónní, C_t je cyklická a ε_t je náhodná složka.

- Multiplikativní model
  $y_t = T_t S_t C_t \varepsilon_t$.
  - Multiplikativní model lze převést logaritmováním na aditivní model.

Volbu vhodné trendové funkce lze provádět pomocí R.

Trendovou funkci lze použít k predikci (předpovědi) hodnot časové řady.

Trendové funkce

Model pro časovou řadu bez sezónní a cyklické složky:

$$Y_t = T_t + \varepsilon_t \ (t = 1,2,\ldots,n)$$

Trend T_t budeme modelovat vhodnou matematickou křivkou.

Nejčastěji uvažované trendové funkce (pro t = 1,2,…,n):

- a) Lineární trend $T_t = \alpha_0 + \alpha_1 t.$
- b) Kvadratický trend $T_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2.$
- c) Exponenciální trend $T_t = \alpha_0 \alpha_1^t \ (\alpha_1 > 0).$
- d) Modifikovaný exponenciální trend $T_t = k + \alpha_0 \alpha_1^t, \ (\alpha_1 > 0).$
- e) Logistický trend $T_t = \dfrac{k}{1 + \alpha_0 \alpha_1^t}, \ (k>0, \ \alpha_1>0).$
- f) Gompertzova křivka $T_t = k\, \alpha_0^{\alpha_1^t}, \ (\alpha_0 > 0, \ \alpha_1 > 0).$

Parametry α_0, α_1, k jsou neznámé, nutno je odhadnout z dat – metody odhadu: Viz Stuchlý (1999b), s. 67-69.

Výběr vhodného modelu trendové funkce provedeme:

a) pomocí věcně ekonomických kritérií (volba na základě teoretických znalostí a zkušeností zkoumaného ekonomického jevu),

b) pomocí analýzy grafu zobrazované časové řady (vizuální analýza, jejíž nevýhodou je subjektivní přístup),

c) s využitím některých regresních kritérií jako např. reziduálního součtu čtverců, indexu korelace, F-statistiky používané k celkovému testu modelu aj.

Kromě toho lze využít testů založených na jednoduchých charakteristikách časové řady (viz následující tabulka).

V statistických programech je možno se setkat s následujícími mírami vhodně zvolené trendové funkce:

střední kvadratická chyba odhadu MSE (*Mean Squared Error*)

$$MSE = \frac{\sum_{t=1}^{n} (y_t - \hat{T}_t)^2}{n},$$

střední absolutní chyba odhadu MAE (*Mean Absolute Error*)

$$MAE = \frac{\sum_{t=1}^{n} | y_t - \hat{T}_t |}{n},$$

Trend	Test
lineární	první diference jsou přibližně konstantní, druhé diference jsou přibližně nulové.
kvadratický	druhé diference jsou přibližně konstantní, třetí diference jsou přibližně nulové.
exponenciální	podíl relativních diferencí $\Delta y_t / \Delta y_{t-1}$ nebo koeficienty růstu jsou přibližně konstantní.
logistický	křivka prvních diferencí se podobá křivce hustoty normálního rozdělení, podíly $(1/y_{t+2} - 1/y_{t+1})/(1/y_{t+1} - 1/y_t)$ jsou přibližně konstantní.
Gompertzova křivka	podíly $(\ln y_{t+2} - \ln y_{t+1})/(\ln y_{t+1} - \ln y_t)$ jsou přibližně konstantní.

střední absolutní chyba procentuální MAPE (*Mean Absolute Percentage Error*)

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{| y_t - \hat{T}_t |}{y_t} .100,$$

střední chyba procentuální MPE (*Mean Percentage Error*)

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \hat{T}_t)}{y_t}.100,$$

kde y_t značí pozorovanou hodnotu časové řady v okamžiku t a $\hat{T}_t$ značí vyrovnanou hodnotu trendu v okamžiku t. Z uvedených kritérií se nejčastěji používá střední kvadratická chyba MSE. Obecně dáváme přednost modelu, u něhož je hodnota MSE nejnižší.

Příklady: Viz Stuchlý (1999b), s. 71-80.

Klouzavé průměry (Moving Averages)

Použití k:

- vyrovnání časové řady,
- k předpovědi.

Průměry počítané klouzáním po časové řadě

Prosté klouzavé průměry pro lichou délku klouzavé části

$$\bar{y}_t = \frac{y_{t-p} + y_{t-p+1} + \cdots + y_{t+p-1} + y_{t+p}}{2p+1} = \frac{1}{2p+1}\sum_{i=-p}^{p} y_{t+i}, \text{ pro } t = p+1, p+2, \ldots, n-p.$$

Takto vypočítané klouzavé průměry jsou nevhodné pro sudou délku klouzavé části z hlediska porovnávání skutečných hodnot časové řady s hodnotami klouzavých průměrů a proto je třeba je centrovat, tj. počítat průměr z každých dvou po sobě následujících klouzavých průměrech. Dostáváme potom tzv. centrované klouzavé průměry (*Centered Moving Average*). Obecně můžeme centrované klouzavé průměry počítat podle vzorce:

$$\bar{y}_t = \frac{1}{4p}(y_{t-p} + 2y_{t-p+1} + \cdots + 2y_{t+p-1} + y_{t+p}), \text{ pro } t = p+1, p+2, \ldots, n-p.$$

Problémy při používání klouzavých průměrů:

- Prvních a posledních p hodnot není vyrovnáno (chybí).
- U časových řad bez trendové a sezónní složky představuje klouzavý průměr vždy předpověď na následující období (takto používá klouzavé průměry Excel v *Analýze dat*).

Příklady: Viz Stuchlý (1999b), str. 84-86.

Časové řady v R (po aktivizaci balíku *vsePackage*):

- 1. Časová řada

ts(y, start =)
plot(objekt)

2. Trendová složka
ts.explore(objekt)
ts.fit.trend(objekt, trend="linear")
ts.fit.trend(objekt, trend="quadratic")
ts.fit.trend(objekt, trend="exponential")
ts.fit.trend(objekt, trend="modified.exponential")
ts.fit.trend(objekt, trend="logistic")
ts.fit.trend(objekt, trend="gompertz")

3. Náhodná složka
resplot(model)
write.noise.test(residuals(model))

4. Predikce
predict(model, step=)
plot(model, step=)

5. Klouzavé průměry
ts.masooth(objekt, order= , length=)

🕮 Studijní materiály:

Základní literatura:

HINDLS, R. a kol. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. S. 245-302. ISBN 978-80-86946-43-6.

STUCHLÝ, J. *Statistika II Cvičení ze statistických metod pro manažery*. J. Hradec: VŠE, 1999. S. 67-86. ISBN 80-7079-035-0.

Doporučené studijní zdroje:

BÍNA V. a kol. *Jak na jazyk R.* J. Hradec: FM VŠE, 2006.

HINDLS, R. a kol. *Analýza dat v manažerském rozhodování*. Praha: Grada, 1999. S. 164-181, 185-189. ISBN 80-7169-255-7.

HINDLS, R. a kol. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000. S. 89-126, 137-144. ISBN 80-7261-013-9.

JAROŠOVÁ, E. *Statistika B. Řešené příklady.* Praha: VŠE, 1994. S. 127-171, ISBN 80-7079-328-7

MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing, 2007. S. 279-303. ISBN 978-80-86446-40-5.

ŘEZANKOVÁ, H. a T. LÖSTER. *Úvod do statistiky*. Praha: Oeconomica, 2009. S. 59-64, 67-72. ISBN 978-80-245-1514-4.

SEGER, J. a R. HINDLS, R. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing, 1995. S. 257-310. ISBN 80-7187-058-7.

STUCHLÝ, J. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice, 2011. (v elektronické formě - https://is.vstecb.cz/auth/www/6384/).

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. S. 268-276, 283-308. ISBN 80-7169-089-9.

WONNACOT, T. H. a R. J. WONNACOT. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing, 1993. S. 754-772. ISBN 80-85605-09-0.

? Otázky a úkoly

1) Pro časovou řadu hodnot průměrné měsíční mzdy pracovníků státního a družstevního sektoru národního hospodářství v ČR v letech 1981-1990: 2 692, 2 757, 2 808, 2 858, 2 901, 2 944, 3 005, 3 070, 3 138, 3 247 vypočítejte a interpretujte a) absolutní přírůstky a průměrný absolutní přírůstek, b) koeficienty růstu a průměrný koeficient růstu, c) 2. diference.
2) K dispozici jsou údaje o stavu skotu v ČR v letech 1989-1997 (tis. kusů):

   3506, 3360, 2950, 2512, 2161, 2030, 1989, 1866, 1690

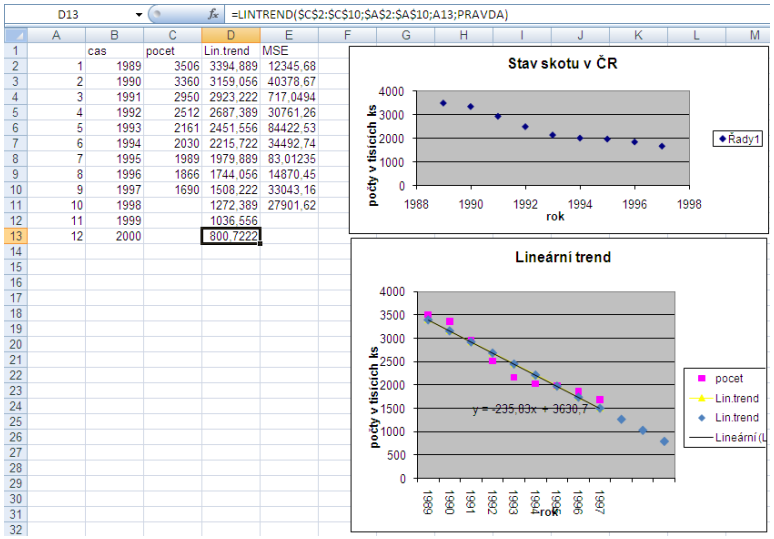
   Vyrovnejte data lineární trendovou funkcí, pomocí MSE vyhodnoťte přesnost vyrovnání a proveďte předpověď stavu skotu na roky 1998-2000.
3) Řešte předcházející úlohu pomocí klouzavých průměrů a centrovaných klouzavých průměrů. Předpověď počítejme pro rok 1998. Porovnejte výsledky.
4) Řešte úkol 2 v R výběrem nejvhodnějšího trendu a proveďte předpověď stavu skotu na roky 1998-2000.

? Úkoly k zamyšlení a diskuzi

1) Diskutujte o významu indexů a rozdílů pro ekonoma.
2) Zamyslete se nad možnostmi využití časových řad k předpovědím.

Klíč k řešení otázek:

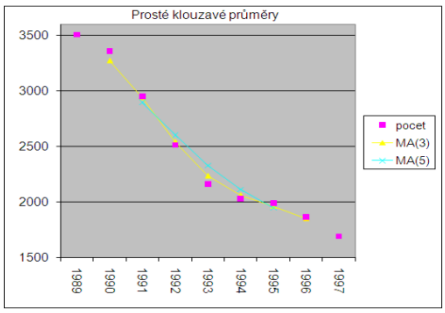
1) Číselné charakteristiky časové řady: a) 61,67, b) 1,021 (řešení - viz Stuchlý 1999b, s. 63-64).
2) Lineární trend: Znázornění, vyrovnání i předpověď je možno provést v Excelu. Předpověď pro rok 1998 je 1293 (pro rok 2000 je 800,7). MSE je 27902. Výstup:



	A	B	C	D	E
1		cas	pocet	Lin.trend	MSE
2	1	1989	3506	3394,889	12345,68
3	2	1990	3360	3159,056	40378,67
4	3	1991	2950	2923,222	717,0494
5	4	1992	2512	2687,389	30761,26
6	5	1993	2161	2451,556	84422,53
7	6	1994	2030	2215,722	34492,74
8	7	1995	1989	1979,889	83,01235
9	8	1996	1866	1744,056	14870,45
10	9	1997	1690	1508,222	33043,16
11	10	1998		1272,389	27901,62
12	11	1999		1036,556	
13	12	2000		800,7222	

3) Klouzavé průměry v Excelu. Nejdříve pomocí tříčlenných a pětičlenných průměrů. Výstupy:

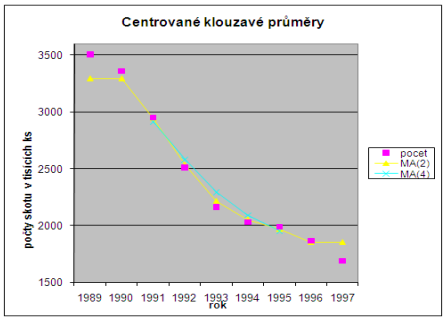
cas	pocet	MA(3)	MA(5)	MSE(3)	MSE(5)
1989	3506				
1990	3360	3272		7744	
1991	2950	2940,667	2897,8	87,11111	2724,84
1992	2512	2541	2602,6	841	8208,36
1993	2161	2234,333	2328,4	5377,778	28022,76
1994	2030	2060	2111,6	900	6658,56
1995	1989	1961,667	1947,2	747,1111	1747,24
1996	1866	1848,333		312,1111	
1997	1690				
				2287,016	9472,352



Lepší vyrovnání poskytují tříčlenné klouzavé průměry. Předpověď pro rok 1998 (určená poslední členem vyrovnané řady) je 1848 a MSE = 2287.

Použijme dále dvojčlenné a čtyřčlenné centrované klouzavé průměry. Výstup:

cas	pocet	MA(2)	MA(4)	MSE(2)	MSE(4)
1989	3506				
1990	3360	3294		4356	
1991	2950	2943	2913,875	49	1305,016
1992	2512	2533,75	2579,5	473,0625	4556,25
1993	2161	2216	2293,125	3025	17457,02
1994	2030	2052,5	2092,25	506,25	3875,063
1995	1989	1968,5	1952,625	420,25	1323,141
1996	1866	1852,75		175,5625	
1997	1690				
				1286,446	5703,297



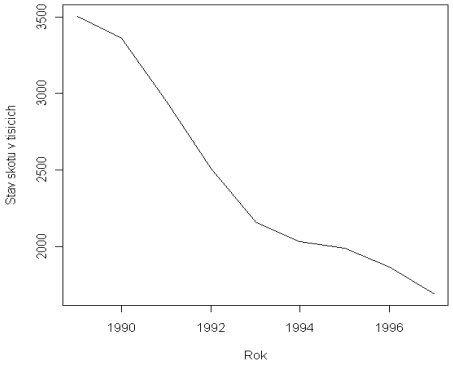
Ještě lepší vyrovnání poskytuje dvojčlenný centrovaný průměr. Předpověď na rok 1998 je 1852,8 a MSE = 1286,5. Úkol je možno řešit i v R.

4) Trendové funkce: Vstup dat do R (po aktivaci *vsePackage*):

cas <- 1989:1997
pocet <- c(3506, 3360, 2950, 2512, 2161, 2030, 1989, 1866, 1690)
data <- data.frame(cas, pocet)

Ulozeni a zobrazeni dat ve forme casove rady:

skot <- ts(data$pocet, start=1989)
skot
par(mfrow=c(1,1))
plot(skot, xlab="Rok", ylab="Stav skotu v tisicich")



Výstup:

```
> skot
Time Series:
Start = 1989
End = 1997
Frequency = 1
[1] 3506 3360 2950 2512 2161 2030 1989 1866
1690
```

Hledani nejvhodnejsiho trendu:

ts.explore(skot, xlab="t")
vyber <- ts.fit.trends(skot)
vyber
plot(vyber)

Výstup:

```
> vyber

    Time series: analysis of trend

                          MSE       MAE             ME          MPE      MAPE
linear              27901.617 144.81481  5.810691e-13  0.366022519 6.312749
quadratic            8281.953  83.27561  1.010523e-13 -0.002979555 3.372092
exponential         13653.084 104.09502  1.913640e+00  0.341035107 4.578573
modified.exponential 9686.729  85.02633 -1.727843e-03  0.010563399 3.316746
logistic            12107.499  90.17628 -2.418045e+00 -0.241446602 3.421633
gompertz            10200.831  88.10475 -1.022910e-02  0.018052998 3.461254

> plot(vyber)
```

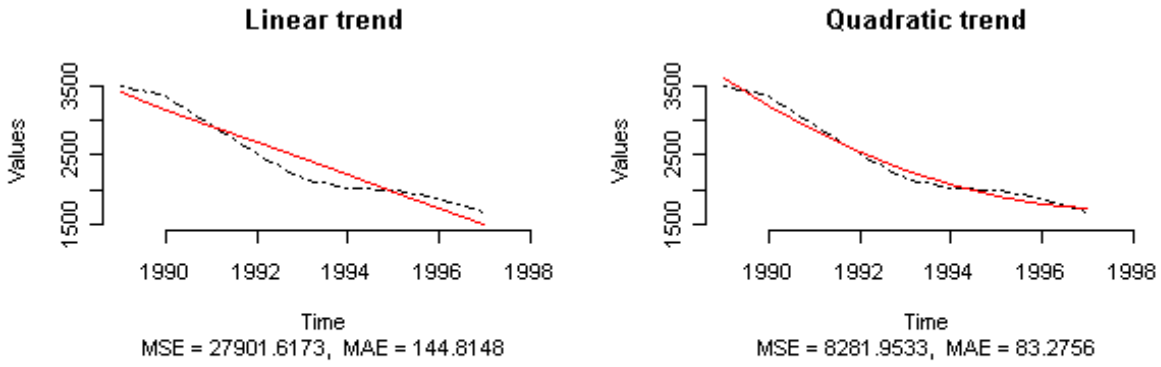
Nejlepší výsledky dává kvadratický trend, kde je MSE nemenší = 8281,9. Srovnatelné výsledky dává i modifikovaný exponenciální trend s MSE = 9686,7. Grafy jednotlivých trendových funkcí jsou na následujím obrázku.

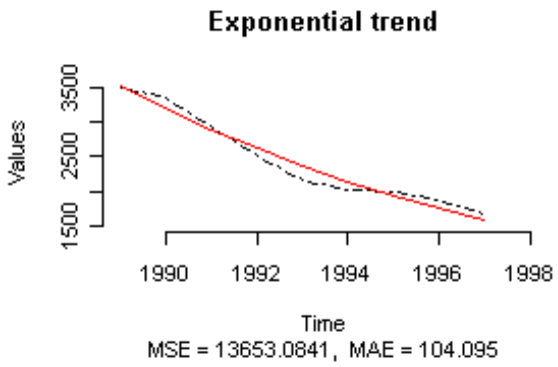
Odhad parametru pro nejvhodnejsi trend:

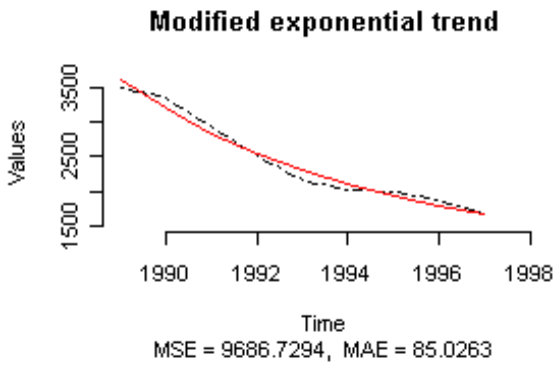
model01 <- ts.trend(skot, trend="quadratic")
model02 <- ts.trend(skot, trend="modified.exponential")
model01
model02

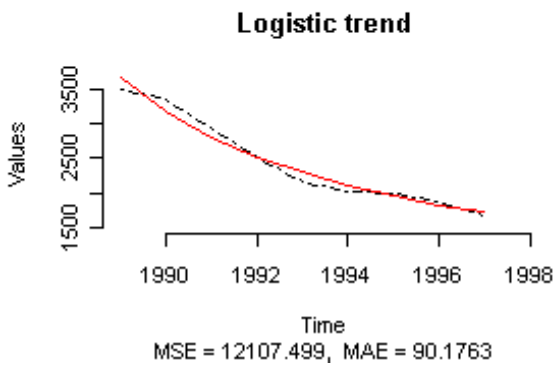
Výstup:

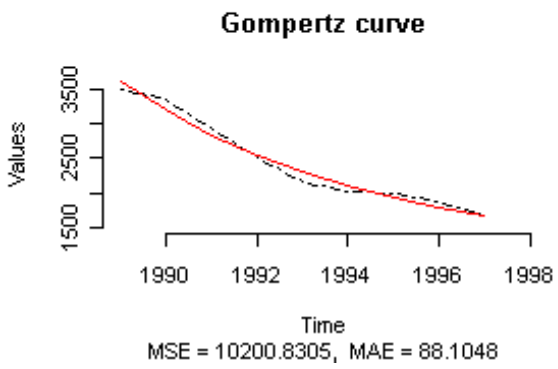
```
> model01
    Time series: analysis of trend

Trend = quadratic  (alpha + beta*t + gamma*t^2)

    Fitted coefficients:
       alpha      beta    gamma
    4069.690 -475.2706 23.94372

             Sum of Squared Errors (SSE): 74537.58
                Mean Squared Error (MSE): 8281.953
              Mean Absolute Error (MAE): 83.27561
                         Mean Error (ME): 1.010523e-13
  Mean Absolute Percentage Error (MAPE): 3.372092%
             Mean Percentage Error (MPE): -0.002979555%

    Fitted values:
Time Series:
Start = 1989
End = 1997
Frequency = 1
[1] 3618.364 3214.924 2859.372 2551.708 2291.931 2080.041 1916.039 1799.924
[9] 1731.697
```












```
> model02
    Time series: analysis of trend

Trend = modified.exponential  (gamma + alpha*beta^t)

    Fitted coefficients:
       alpha      beta     gamma
    3006.715 0.8290591 1131.099

             Sum of Squared Errors (SSE): 87180.56
                Mean Squared Error (MSE): 9686.73
              Mean Absolute Error (MAE): 85.02633
                        Mean Error (ME): -0.001727843
  Mean Absolute Percentage Error (MAPE): 3.316746%
             Mean Percentage Error (MPE): 0.0105634%

    Fitted values:
Time Series:
Start = 1989
End = 1997
```

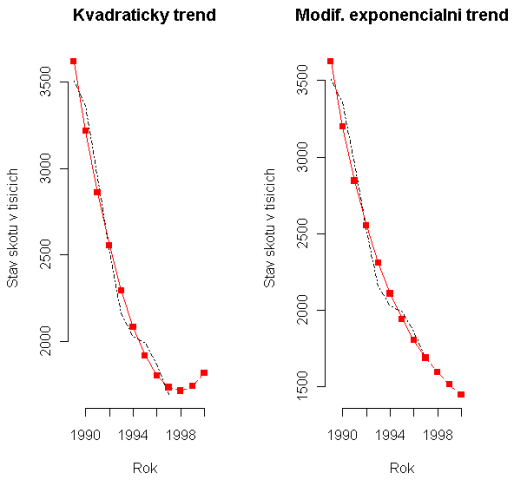
```
Frequency = 1
[1] 3623.844 3197.732 2844.459 2551.576 2308.758 2107.448 1940.550 1802.182
[9] 1687.466
```

Předpověd pro roky 1997-2000 a grafické znázornění:

predict(model01, step=1:3)
predict(model02, step=1:3)
par(mfrow=c(1,2))
plot(model01, step=1:3, xlab="Rok", ylab="Stav skotu v tisicich", main="Kvadraticky trend")
plot(model02, step=1:3, xlab="Rok", ylab="Stav skotu v tisicich", main="Modif. exponencialni trend")

Výstup:

```
> predict(model01, step=1:3)
       1        2        3
1711.357 1738.905 1814.340

> predict(model02, step=1:3)
       1        2        3
1592.360 1513.512 1448.142
```



Závěr: Optimální odhad pro rok 1998 kvadratickou trendovou funkcí je 1711,4 s MSE = 8282.

Dodatky

Statistické tabulky

I. Distribuční funkce standardního normálního rozdělení N(0;1)

Jsou tabelovány hodnoty Laplaceovy funkce $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$. Platí $\Phi(-x) = 1 - \Phi(x)$.

x	ø(x)	x	ø(x)	x	ø(x)	x	ø(x)
0,00	0,5000	0,30	0,6179	0,60	0,7257	0,90	0,8159
0,01	0,5040	0,31	0,6217	0,61	0,7291	0,91	0,8186
0,02	0,5080	0,32	0,6255	0,62	0,7324	0,92	0,8212
0,03	0,5120	0,33	0,6293	0,63	0,7357	0,93	0,8238
0,04	0,5160	0,34	0,6331	0,64	0,7389	0,94	0,8264
0,05	0,5199	0,35	0,6368	0,65	0,7422	0,95	0,8289
0,06	0,5239	0,36	0,6406	0,66	0,7454	0,96	0,8315
0,07	0,5279	0,37	0,6443	0,67	0,7486	0,97	0,8340
0,08	0,5319	0,38	0,6480	0,68	0,7517	0,98	0,8365
0,09	0,5359	0,39	0,6517	0,69	0,7549	0,99	0,8389
0,10	0,5398	0,40	0,6554	0,70	0,7580	1,00	0,8413
0,11	0,5438	0,41	0,6591	0,71	0,7611	1,01	0,8438
0,12	0,5478	0,42	0,6628	0,72	0,7642	1,02	0,8461
0,13	0,5517	0,43	0,6664	0,73	0,7673	1,03	0,8485
0,14	0,5557	0,44	0,6700	0,74	0,7703	1,04	0,8508
0,15	0,5596	0,45	0,6736	0,75	0,7734	1,05	0,8531
0,16	0,5636	0,46	0,6772	0,76	0,7764	1,06	0,8554
0,17	0,5675	0,47	0,6808	0,77	0,7794	1,07	0,8577
0,18	0,5714	0,48	0,6844	0,78	0,7823	1,08	0,8599
0,19	0,5753	0,49	0,6879	0,79	0,7853	1,09	0,8621
0,20	0,5793	0,50	0,6915	0,80	0,7881	1,10	0,8643
0,21	0,5832	0,51	0,6950	0,81	0,7910	1,11	0,8665
0,22	0,5871	0,52	0,6985	0,82	0,7939	1,12	0,8686
0,23	0,5910	0,53	0,7019	0,83	0,7967	1,13	0,8708
0,24	0,5948	0,54	0,7054	0,84	0,7995	1,14	0,8729
0,25	0,5987	0,55	0,7088	0,85	0,8023	1,15	0,8749
0,26	0,6026	0,56	0,7123	0,86	0,8051	1,16	0,8770
0,27	0,6064	0,57	0,7157	0,87	0,8078	1,17	0,8790
0,28	0,6103	0,58	0,7190	0,88	0,8106	1,18	0,8810
0,29	0,6141	0,59	0,7224	0,89	0,8133	1,19	0,8830

Pokračování tabulky I.

x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$
1,20	0,8849	1,60	0,9452	2,00	0,9772	2,40	0,9918	2,80	0,9974
1,21	0,8869	1,61	0,9463	2,01	0,9778	2,41	0,9920	2,82	0,9976
1,22	0,8888	1,62	0,9474	2,02	0,9783	2,42	0,9922	2,84	0,9977
1,23	0,8907	1,63	0,9484	2,03	0,9788	2,43	0,9925	2,86	0,9979
1,24	0,8925	1,64	0,9495	2,04	0,9793	2,44	0,9927	2,88	0,9980
1,25	0,8944	1,65	0,9505	2,05	0,9798	2,45	0,9929	2,90	0,9981
1,26	0,8962	1,66	0,9515	2,06	0,9803	2,46	0,9931	2,92	0,9982
1,27	0,8980	1,67	0,9525	2,07	0,9808	2,47	0,9932	2,94	0,9984
1,28	0,8997	1,68	0,9535	2,08	0,9812	2,48	0,9934	2,96	0,9985
1,29	0,9015	1,69	0,9545	2,09	0,9817	2,49	0,9936	2,98	0,9986
1,30	0,9032	1,70	0,9554	2,10	0,9821	2,50	0,9938	3,00	0,9986
1,31	0,9049	1,71	0,9564	2,11	0,9826	2,51	0,9940	3,10	0,9990
1,32	0,9066	1,72	0,9572	2,12	0,9830	2,52	0,9941	3,20	0,9993
1,33	0,9082	1,73	0,9582	2,13	0,9834	2,53	0,9943	3,30	0,9995
1,34	0,9099	1,74	0,9591	2,14	0,9838	2,54	0,9945	3,40	0,9996
1,35	0,9115	1,75	0,9599	2,15	0,9842	2,55	0,9946	3,50	0,9998
1,36	0,9131	1,76	0,9608	2,16	0,9846	2,56	0,9948	3,60	0,9998
1,37	0,9147	1,77	0,9616	2,17	0,9850	2,57	0,9949	3,70	0,9999
1,38	0,9162	1,78	0,9625	2,18	0,9854	2,58	0,9951	3,80	0,9999
1,39	0,9177	1,79	0,9633	2,19	0,9857	2,59	0,9952	3,90	0,9999
1,40	0,9192	1,80	0,9641	2,20	0,9861	2,60	0,9953		
1,41	0,9207	1,81	0,9649	2,21	0,9864	2,61	0,9955		
1,42	0,9222	1,82	0,9656	2,22	0,9868	2,62	0,9956		
1,43	0,9236	1,83	0,9664	2,23	0,9871	2,63	0,9957		
1,44	0,9251	1,84	0,9671	2,24	0,9873	2,64	0,9959		
1,45	0,9265	1,85	0,9678	2,25	0,9878	2,65	0,9960		
1,46	0,9279	1,86	0,9686	2,26	0,9881	2,66	0,9961		
1,47	0,9292	1,87	0,9693	2,27	0,9884	2,67	0,9962		
1,48	0,9306	1,88	0,9699	2,28	0,9887	2,68	0,9963		
1,49	0,9319	1,89	0,9706	2,29	0,9890	2,69	0,9964		
1,50	0,9332	1,90	0,9713	2,30	0,9893	2,70	0,9965		
1,51	0,9345	1,91	0,9719	2,31	0,9896	2,71	0,9966		
1,52	0,9357	1,92	0,9726	2,32	0,9898	2,72	0,9967		
1,53	0,9370	1,93	0,9732	2,33	0,9901	2,73	0,9968		
1,54	0,9382	1,94	0,9738	2,34	0,9904	2,74	0,9969		
1,55	0,9394	1,95	0,9744	2,35	0,9906	2,75	0,9970		
1,56	0,9406	1,96	0,9750	2,36	0,9909	2,76	0,9971		
1,57	0,9418	1,97	0,9756	2,37	0,9911	2,77	0,9972		
1,58	0,9429	1,98	0,9761	2,38	0,9913	2,78	0,9973		
1,59	0,9441	1,99	0,9767	2,39	0,9916	2,79	0,9974		

II. Kvantily standardního normálního rozdělení

Jsou tabelovány kvantily standardního normálního rozdělení x_p určené vztahem $P(X < x_p) = p$, kde $X \sim N(0;1)$.

p	x_p	p	x_p	p	x_p	p	x_p
0,50	0,000	0,75	0,674	0,950	1,645	0,975	1,960
0,51	0,025	0,76	0,706	0,951	1,655	0,976	1,977
0,52	0,050	0,77	0,739	0,952	1,665	0,977	1,995
0,53	0,075	0,78	0,772	0,953	1,675	0,978	2,014
0,54	0,100	0,79	0,806	0,954	1,685	0,979	2,034
0,55	0,126	0,80	0,842	0,955	1,695	0,980	2,054
0,56	0,151	0,81	0,878	0,956	1,706	0,981	2,075
0,57	0,176	0,82	0,915	0,957	1,717	0,982	2,097
0,58	0,202	0,83	0,954	0,958	1,728	0,983	2,120
0,59	0,228	0,84	0,994	0,959	1,739	0,984	2,144
0,60	0,253	0,85	1,036	0,960	1,751	0,985	2,170
0,61	0,279	0,86	1,080	0,961	1,762	0,986	2,197
0,62	0,305	0,87	1,126	0,962	1,774	0,987	2,226
0,63	0,332	0,88	1,175	0,963	1,787	0,988	2,257
0,64	0,358	0,89	1,227	0,964	1,799	0,989	2,290
0,65	0,385	0,90	1,282	0,965	1,812	0,990	2,326
0,66	0,412	0,905	1,311	0,966	1,825	0,991	2,366
0,67	0,440	0,910	1,341	0,967	1,838	0,992	2,409
0,68	0,468	0,915	1,372	0,968	1,852	0,993	2,457
0,69	0,496	0,920	1,405	0,969	1,866	0,994	2,512
0,70	0,524	0,925	1,440	0,970	1,881	0,995	2,576
0,71	0,553	0,930	1,476	0,971	1,896	0,996	2,625
0,72	0,583	0,935	1,514	0,972	1,911	0,997	2,748
0,73	0,613	0,940	1,555	0,973	1,927	0,998	2,878
0,74	0,643	0,945	1,598	0,974	1,943	0,999	3,090

III. Kvantily rozdělení chi-kvadrát

Jsou tabelovány kvantily chi-kvadrát rozdělení $\chi^2_p(n)$ určené vztahem $P(X < \chi^2_p(n)) = p$, kde $X \sim \chi^2(n)$.

n \ p	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990
1	0,0002	0,0010	0,0039	0,0158	2,71	3,84	5,02	6,63
2	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21
3	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3
4	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3
5	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1
6	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8
7	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5
8	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1
9	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7
10	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2
11	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7
12	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2
13	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7
14	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1
15	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6
16	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0
17	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4
18	7,01	8,23	9,39	10,9	26,0	28,9	31,5	34,8
19	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2
20	8,26	9,56	10,9	12,4	28,4	31,4	34,2	37,6
21	8,90	10,3	11,6	13,2	29,6	32,7	35,5	38,9
22	9,54	11,0	12,3	14,0	30,8	33,9	36,8	40,3
23	10,2	11,7	13,1	14,8	32,0	35,2	38,1	41,6
24	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43,0
25	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3
26	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6
27	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47,0
28	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3
29	14,3	16,0	17,7	19,8	39,1	42,6	45,7	49,6
30	15,0	16,8	18,5	20,6	40,3	43,8	47,0	50,9
40	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7
50	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2
60	37,5	40,5	43,2	46,5	74,4	79,1	83,3	88,4
70	45,4	48,8	51,7	55,3	85,5	90,5	95,0	100,4
80	53,5	57,2	60,4	64,3	96,6	101,9	106,6	112,3
90	61,8	65,6	69,1	73,3	107,6	113,1	118,1	124,1
100	70,1	74,2	77,9	82,4	118,5	124,3	129,6	135,8

IV. Kvantily Studentova *t* - rozdělení

Jsou tabelovány kvantily t-rozdělení $t_p(n)$ definované vztahem $P(X < t_p(n)) = p$, kde $X \sim t(n)$.

v	P				
	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

V. Kvantily F-rozdělení

Jsou tabelovány kvantily F-rozdělení $F_p(\nu_1;\nu_2)$ definované vztahem $P(X < F_p(\nu_1;\nu_2)) = p$ pro $p = 0{,}95,\ 0{,}975,\ 0{,}99,\ 0{,}995$, kde $X \sim F(\nu_1;\nu_2)$. Platí $F_p(\nu_1;\nu_2) = 1/F_{1-p}(\nu_2;\nu_1)$.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
	648	800	864	900	922	937	948	957	963	969
	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4
	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4
	198	199	199	199	199	199	199	199	199	199
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4
	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2
	55,6	49,8	47,5	46,2	45,4	44,8	44,4	44,1	43,9	43,7
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5
	31,3	26,3	24,3	23,2	22,5	22,0	21,6	21,4	21,1	21,0
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1
	22,8	18,3	16,5	15,6	14,9	14,5	14,2	14,0	13,8	13,6
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
	18,6	14,5	12,9	12,0	11,5	11,1	10,8	10,6	10,4	10,2
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
	16,2	12,4	10,9	10,0	9,52	9,16	8,89	8,68	8,51	8,38
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
	14,7	11,0	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
	13,6	10,1	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42

Pokračování tabulky V.

ν_2 \ ν_1	11	12	15	20	24	30	40	60	120	∞
1	243	244	246	248	249	250	251	252	253	254
	973	977	985	993	997	1001	1006	1010	1014	1018
	6083	6106	6157	6209	6235	6261	6287	6313	6339	6366
	24325	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	19,5	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99,5
	199	199	199	199	199	199	199	199	199	199
3	8,76	8,74	8,70	8,66	8,63	8,62	8,59	8,57	8,55	8,53
	14,4	14,3	14,3	14,2	14.1	14,1	14,0	14,0	13,9	13,9
	27,1	27,1	26,9	26,7	26,6	26,5	26,4	26,3	26,2	26,1
	43,5	43,4	43,1	42,8	42,6	42,5	42,3	42,1	42,0	41,8
4	5,94	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	8,79	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
	14,4	14,4	14,2	14,0	13,9	13,8	13,7	13,7	13,6	13,5
	20,8	20,7	20,4	20,2	20,0	19,9	19,8	19,6	19,5	19,3
5	4,71	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
	6,57	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
	9,96	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
	13,5	13,4	13,1	12,9	12,8	12,7	12,5	12,4	12,3	12,1
6	4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
	5,41	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
	7,79	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
	10,1	10,0	9,81	9,59	9,47	9,36	9,24	9,12	9,00	8,88
7	3,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
	4,71	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
	6,54	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
	8,27	8,18	7,97	7,75	7,65	7,53	7,42	7,31	7,19	7,08
8	3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
	4,24	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
	5,73	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
	7,10	7,01	6,81	6,61	6,50	6,40	6,29	6,18	6,06	5,95
9	3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
	3,91	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
	5,18	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
	6,31	6,23	6,03	5,83	5,73	5,62	5,52	5,41	5,30	5,19

Pokračování tabulky V.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
	12,8	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,86
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
	11,8	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
	10,8	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64
	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
	9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
	8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	3,22	3,12
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
	7,08	4,98	4,14	3,65	3,34	3,12	2,95	2,82	2,72	2,63
	8,49	5,80	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16
	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,92	2,81	2,71
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83
	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05
	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32
	7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,74	2,62	2,52

Pokračování tabulky V.

ν_2 \ ν_1	11	12	15	20	24	30	40	60	120	∞
10	2,94 3,66 4,77 5,75	2,91 3,62 4,71 5,66	2,85 3,52 4,56 5,47	2,77 3,42 4,41 5,27	2,74 3,37 4,33 5,17	2,70 3,31 4,25 5,07	2,66 3,26 4,17 4,97	2,62 3,20 4,08 4,86	2,58 3,14 4,00 4,75	2,54 3,08 3,91 4,64
12	2,72 3,32 4,22 4,99	2,69 3,28 4,16 4,91	2,62 3,18 4,01 4,72	2,54 3,07 3,86 4,53	2,51 3,02 3,78 4,43	2,47 2,96 3,70 4,33	2,43 2,91 3,62 4,23	2,38 2,85 3,54 4,12	2,34 2,79 3,45 4,01	2,30 2,72 3,36 3,90
15	2,51 3,01 3,73 4,33	2,48 2,96 3,67 4,25	2,40 2,86 3,52 4,07	2,33 2,76 3,37 3,88	2,29 2,70 3,29 3,79	2,25 2,64 3,21 3,69	2,20 2,59 3,13 3,59	2,16 2,52 3,05 3,48	2,11 2,46 2,96 3,37	2,07 2,40 2,87 3,26
20	2,31 2,72 3,29 3,76	2,28 2,68 3,23 3,68	2,20 2,57 3,09 3,50	2,12 2,46 2,94 3,32	2,08 2,41 2,86 3,22	2,04 2,35 2,78 3,12	1,99 2,29 2,69 3,02	1,95 2,22 2,61 2,92	1,90 2,16 2,52 2,81	1,84 2,09 2,42 2,69
24	2,21 2,59 3,09 3,50	2,18 2,54 3,03 3,42	2,11 2,44 2,89 3,25	2,03 2,33 2,74 3,06	1,98 2,27 2,66 2,97	1,94 2,21 2,58 2,87	1,89 2,15 2,49 2,77	1,84 2,08 2,40 2,66	1,79 2,01 2,31 2,55	1,73 1,94 2,21 2,43
30	2,13 2,46 2,91 3,25	2,09 2,41 2,84 3,18	2,01 2,31 2,70 3,01	1,93 2,20 2,55 2,82	1,89 2,14 2,47 2,73	1,84 2,07 2,39 2,63	1,79 2,01 2,30 2,52	1,74 1,94 2,21 2,42	1,68 1,87 2,11 2,30	1,62 1,79 2,01 2,18
40	2,04 2,33 2,73 3,03	2,00 2,29 2,66 2,95	1,92 2,18 2,52 2,78	1,84 2,07 2,37 2,60	1,79 2,01 2,29 2,50	1,74 1,94 2,20 2,40	1,69 1,88 2,11 2,30	1,64 1,80 2,02 2,18	1,58 1,72 1,92 2,06	1,51 1,64 1,80 1,93
60	1,95 2,22 2,56 2,82	1,92 2,17 2,50 2,74	1,84 2,06 2,35 2,57	1,75 1,94 2,20 2,39	1,70 1,88 2,12 2,29	1,65 1,82 2,03 2,19	1,59 1,74 1,94 2,08	1,53 1,67 1,84 1,96	1,47 1,58 1,73 1,83	1,39 1,48 1,60 1,69
120	1,87 2,10 2,40 2,62	1,83 2,05 2,34 2,54	1,75 1,95 2,19 2,37	1,66 1,82 2,03 2,19	1,61 1,76 1,95 2,09	1,55 1,69 1,86 1,98	1,50 1,61 1,76 1,87	1,43 1,53 1,66 1,75	1,35 1,43 1,53 1,61	1,25 1,31 1,38 1,43
∞	1,79 1,99 2,25 2,43	1,75 1,94 2,18 2,36	1,67 1,83 2,04 2,19	1,57 1,71 1,88 2,00	1,52 1,64 1,79 1,90	1,46 1,57 1,70 1,79	1,39 1,48 1,59 1,67	1,32 1,39 1,47 1,53	1,22 1,27 1,32 1,36	1,00 1,00 1,00 1,00

VI. Kritické hodnoty Wilcoxonova jednovýběrového testu (WJT)

Jsou tabelovány kritické hodnoty Wilcoxonova testu T_α, WJT určené vztahem $P(T \leq T_\alpha) \leq \alpha$.

n	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}01$	$\alpha = 0{,}005$
5	1			
6	2	1		
7	4	2	0	
8	6	4	2	0
9	8	6	3	2
10	11	8	5	3
11	14	11	7	5
12	17	14	10	7
13	21	17	13	10
14	26	21	16	13
15	30	25	20	16
16	36	30	24	19
17	41	35	28	23
18	47	40	33	28
19	54	46	38	32
20	60	52	43	37
21	68	59	49	43
22	75	66	56	49
23	83	73	62	55
24	92	81	69	68
25	101	90	77	68
26	110	98	85	76
27	120	107	93	84
28	130	117	102	92
29	141	127	111	100
30	152	137	120	109
31	163	148	130	118
32	175	159	141	128
33	188	171	151	138
34	201	183	162	149
35	214	195	174	160
36	228	208	186	171
37	242	222	198	183
38	256	235	211	195
39	271	250	224	208
40	287	264	238	221
41	303	279	252	234
42	319	295	267	248
43	336	311	281	262
44	253	327	297	277
45	371	344	313	292
46	389	361	329	307
47	408	379	345	323
48	427	397	362	339
49	446	415	380	356
50	466	434	398	373

VII. Kritické hodnoty pro Mannův-Whitneyovův test

Jsou tabelovány kritické hodnoty Mannova-Whitneyova testu k_p definované vztahem $P(T \leq k_p) \leq p$, kde n v řádcích a sloupcích představují rozsahy jednotlivých souborů.

		n																		
n	P	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0,025							1	1	1	1	2	2	2	2	2	3	3	3	3
	0,050				1	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
3	0,025				1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
	0,050		1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
4	0,025			1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	14	15
	0,050		1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
5	0,025		1	2	3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
	0,050	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
6	0,025		2	3	4	6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
	0,050	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
7	0,025		2	4	6	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
	0,050	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
8	0,025	1	3	5	7	9	11	14	16	18	20	23	25	27	30	32	35	37	39	42
	0,050	2	4	6	9	11	14	16	19	21	24	27	29	32	34	37	40	42	45	48
9	0,025	1	3	5	8	11	13	16	18	21	24	27	29	32	35	38	40	43	46	49
	0,050	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
10	0,025	1	4	6	9	12	15	18	21	24	27	30	34	37	40	43	46	49	53	56
	0,050	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
11	0,025	1	4	7	10	14	17	20	24	27	31	34	38	41	45	48	52	56	59	63
	0,050	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
12	0,025	2	5	8	12	15	19	23	27	30	34	38	42	46	50	54	58	62	66	70
	0,050	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
13	0,025	2	5	9	13	17	21	25	29	34	38	42	46	51	55	60	64	68	73	77
	0,050	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
14	0,025	2	6	10	14	18	23	27	32	37	41	46	51	56	60	65	70	75	79	84
	0,050	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
15	0,025	2	6	11	15	20	25	30	35	40	45	50	55	60	65	71	76	81	86	91
	0,050	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
16	0,025	2	7	12	16	22	27	32	38	43	48	54	60	65	71	76	82	87	93	99
	0,050	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
17	0,025	3	7	12	18	23	29	35	40	46	52	58	64	70	76	82	88	94	100	106
	0,050	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
18	0,025	3	8	13	19	25	31	37	43	49	56	62	68	75	81	87	94	100	107	113
	0,050	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
19	0,025	3	8	14	20	26	33	39	46	53	59	66	73	79	86	93	100	107	114	120
	0,050	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
20	0,025	3	9	15	21	28	35	42	49	56	63	70	77	84	91	99	106	113	120	128
	0,050	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139

VIII. Kvantily dvouvýběrového Kolmogorovova - Smirnovova testu

Jsou tabelovány kvantily dvouvýběrového Kolmogorova-Smirnovova testu $d_{1-\alpha}$ definované vztahem $P(D \leq d_{1-\alpha}) \geq 1 - \alpha$ pro $\alpha = 0{,}05$.

n \ m	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2					$\frac{8}{8}$	$\frac{18}{18}$	$\frac{10}{10}$	$\frac{22}{22}$	$\frac{12}{12}$	$\frac{26}{26}$	$\frac{14}{14}$	$\frac{28}{30}$	$\frac{15}{16}$	$\frac{32}{34}$	$\frac{17}{18}$	$\frac{36}{38}$	$\frac{19}{20}$
3		$\frac{15}{15}$	$\frac{6}{6}$	$\frac{21}{21}$	$\frac{21}{24}$	$\frac{8}{9}$	$\frac{27}{30}$	$\frac{30}{33}$	$\frac{10}{12}$	$\frac{33}{39}$	$\frac{36}{42}$	$\frac{12}{15}$	$\frac{39}{48}$	$\frac{42}{51}$	$\frac{15}{18}$	$\frac{45}{57}$	$\frac{48}{60}$
4	$\frac{4}{4}$	$\frac{20}{20}$	$\frac{10}{12}$	$\frac{24}{28}$	$\frac{7}{8}$	$\frac{28}{36}$	$\frac{15}{20}$	$\frac{33}{44}$	$\frac{9}{12}$	$\frac{39}{52}$	$\frac{21}{28}$	$\frac{44}{60}$	$\frac{12}{16}$	$\frac{48}{68}$	$\frac{25}{36}$	$\frac{53}{76}$	$\frac{15}{20}$
5		$\frac{5}{5}$	$\frac{24}{30}$	$\frac{28}{35}$	$\frac{30}{40}$	$\frac{35}{45}$	$\frac{8}{10}$	$\frac{39}{55}$	$\frac{43}{60}$	$\frac{45}{65}$	$\frac{46}{70}$	$\frac{11}{15}$	$\frac{54}{80}$	$\frac{55}{85}$	$\frac{60}{90}$	$\frac{61}{95}$	$\frac{13}{20}$
6			$\frac{5}{6}$	$\frac{30}{42}$	$\frac{17}{24}$	$\frac{13}{18}$	$\frac{20}{30}$	$\frac{43}{66}$	$\frac{8}{12}$	$\frac{52}{78}$	$\frac{27}{42}$	$\frac{19}{30}$	$\frac{30}{48}$	$\frac{62}{102}$	$\frac{12}{18}$	$\frac{70}{114}$	$\frac{36}{60}$
7				$\frac{6}{7}$	$\frac{40}{56}$	$\frac{42}{63}$	$\frac{46}{70}$	$\frac{48}{77}$	$\frac{53}{84}$	$\frac{56}{91}$	$\frac{9}{14}$	$\frac{62}{105}$	$\frac{64}{112}$	$\frac{68}{119}$	$\frac{72}{126}$	$\frac{76}{133}$	$\frac{79}{140}$
8					$\frac{6}{8}$	$\frac{46}{72}$	$\frac{21}{40}$	$\frac{53}{88}$	$\frac{15}{24}$	$\frac{62}{104}$	$\frac{32}{56}$	$\frac{67}{120}$	$\frac{10}{16}$	$\frac{77}{136}$	$\frac{40}{72}$	$\frac{82}{152}$	$\frac{22}{40}$
9						$\frac{6}{9}$	$\frac{53}{90}$	$\frac{59}{99}$	$\frac{21}{36}$	$\frac{65}{117}$	$\frac{70}{126}$	$\frac{25}{45}$	$\frac{78}{144}$	$\frac{82}{153}$	$\frac{10}{18}$	$\frac{89}{171}$	$\frac{93}{180}$
10							$\frac{7}{10}$	$\frac{60}{110}$	$\frac{33}{60}$	$\frac{70}{130}$	$\frac{37}{70}$	$\frac{16}{30}$	$\frac{42}{80}$	$\frac{89}{170}$	$\frac{46}{90}$	$\frac{94}{190}$	$\frac{11}{20}$
11								$\frac{7}{11}$	$\frac{72}{132}$	$\frac{75}{143}$	$\frac{82}{154}$	$\frac{84}{165}$	$\frac{89}{176}$	$\frac{93}{187}$	$\frac{97}{198}$	$\frac{102}{209}$	$\frac{107}{220}$
12									$\frac{7}{12}$	$\frac{81}{156}$	$\frac{43}{84}$	$\frac{31}{60}$	$\frac{24}{48}$	$\frac{100}{204}$	$\frac{18}{36}$	$\frac{108}{228}$	$\frac{29}{60}$
13										$\frac{7}{13}$	$\frac{89}{182}$	$\frac{96}{195}$	$\frac{101}{208}$	$\frac{105}{221}$	$\frac{110}{234}$	$\frac{114}{247}$	$\frac{120}{260}$
14											$\frac{8}{14}$	$\frac{98}{210}$	$\frac{53}{112}$	$\frac{111}{238}$	$\frac{58}{126}$	$\frac{121}{266}$	$\frac{63}{140}$
15												$\frac{8}{15}$	$\frac{114}{240}$	$\frac{116}{255}$	$\frac{41}{90}$	$\frac{127}{285}$	$\frac{27}{60}$
16													$\frac{8}{16}$	$\frac{124}{272}$	$\frac{64}{144}$	$\frac{133}{304}$	$\frac{35}{80}$
17														$\frac{8}{17}$	$\frac{133}{306}$	$\frac{141}{323}$	$\frac{146}{340}$
18															$\frac{9}{18}$	$\frac{142}{342}$	$\frac{76}{180}$
19																$\frac{9}{19}$	$\frac{160}{380}$
20																	$\frac{9}{20}$

Pokračování tabulky VIII. pro $\alpha = 0{,}01$

n \ m	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2															$\frac{38}{38}$	$\frac{20}{20}$
3					$\frac{9}{9}$	$\frac{30}{30}$	$\frac{33}{33}$	$\frac{12}{12}$	$\frac{39}{39}$	$\frac{42}{42}$	$\frac{11}{15}$	$\frac{45}{48}$	$\frac{48}{51}$	$\frac{17}{18}$	$\frac{54}{57}$	$\frac{57}{60}$
4		$\frac{12}{12}$	$\frac{28}{28}$	$\frac{8}{8}$	$\frac{36}{36}$	$\frac{18}{20}$	$\frac{40}{44}$	$\frac{11}{12}$	$\frac{48}{52}$	$\frac{24}{28}$	$\frac{52}{60}$	$\frac{14}{16}$	$\frac{60}{68}$	$\frac{30}{36}$	$\frac{64}{76}$	$\frac{17}{20}$
5	$\frac{5}{5}$	$\frac{30}{30}$	$\frac{35}{35}$	$\frac{35}{40}$	$\frac{40}{45}$	$\frac{9}{10}$	$\frac{45}{55}$	$\frac{50}{60}$	$\frac{52}{65}$	$\frac{56}{70}$	$\frac{12}{15}$	$\frac{64}{80}$	$\frac{68}{85}$	$\frac{70}{90}$	$\frac{71}{95}$	$\frac{16}{20}$
6		$\frac{6}{6}$	$\frac{36}{42}$	$\frac{20}{24}$	$\frac{15}{18}$	$\frac{24}{30}$	$\frac{54}{66}$	$\frac{10}{12}$	$\frac{60}{78}$	$\frac{32}{42}$	$\frac{23}{30}$	$\frac{36}{48}$	$\frac{73}{102}$	$\frac{14}{18}$	$\frac{83}{114}$	$\frac{44}{60}$
7			$\frac{6}{7}$	$\frac{48}{56}$	$\frac{49}{63}$	$\frac{53}{70}$	$\frac{59}{77}$	$\frac{60}{84}$	$\frac{65}{91}$	$\frac{11}{14}$	$\frac{75}{105}$	$\frac{77}{112}$	$\frac{84}{119}$	$\frac{87}{126}$	$\frac{91}{133}$	$\frac{93}{140}$
8				$\frac{7}{8}$	$\frac{55}{72}$	$\frac{30}{40}$	$\frac{64}{88}$	$\frac{17}{24}$	$\frac{72}{104}$	$\frac{38}{56}$	$\frac{81}{120}$	$\frac{11}{16}$	$\frac{88}{136}$	$\frac{47}{72}$	$\frac{98}{152}$	$\frac{26}{40}$
9					$\frac{7}{9}$	$\frac{63}{90}$	$\frac{70}{99}$	$\frac{25}{36}$	$\frac{78}{117}$	$\frac{84}{126}$	$\frac{30}{45}$	$\frac{91}{144}$	$\frac{99}{153}$	$\frac{12}{18}$	$\frac{107}{171}$	$\frac{111}{180}$
10						$\frac{8}{10}$	$\frac{77}{110}$	$\frac{40}{60}$	$\frac{84}{130}$	$\frac{45}{70}$	$\frac{20}{30}$	$\frac{50}{80}$	$\frac{106}{170}$	$\frac{54}{90}$	$\frac{113}{190}$	$\frac{13}{20}$
11							$\frac{8}{11}$	$\frac{86}{132}$	$\frac{91}{143}$	$\frac{96}{154}$	$\frac{102}{165}$	$\frac{106}{176}$	$\frac{110}{187}$	$\frac{118}{198}$	$\frac{122}{209}$	$\frac{127}{220}$
12								$\frac{8}{12}$	$\frac{95}{156}$	$\frac{52}{84}$	$\frac{36}{60}$	$\frac{29}{48}$	$\frac{119}{204}$	$\frac{21}{36}$	$\frac{130}{228}$	$\frac{35}{60}$
13									$\frac{9}{13}$	$\frac{104}{182}$	$\frac{115}{195}$	$\frac{121}{208}$	$\frac{127}{221}$	$\frac{131}{234}$	$\frac{138}{247}$	$\frac{143}{260}$
14										$\frac{9}{14}$	$\frac{123}{210}$	$\frac{63}{112}$	$\frac{134}{238}$	$\frac{70}{126}$	$\frac{148}{266}$	$\frac{76}{140}$
15											$\frac{9}{15}$	$\frac{133}{240}$	$\frac{142}{255}$	$\frac{49}{90}$	$\frac{152}{285}$	$\frac{32}{60}$
16												$\frac{10}{16}$	$\frac{143}{272}$	$\frac{77}{144}$	$\frac{160}{304}$	$\frac{42}{80}$
17													$\frac{10}{17}$	$\frac{164}{306}$	$\frac{166}{323}$	$\frac{175}{340}$
18														$\frac{10}{18}$	$\frac{176}{342}$	$\frac{91}{180}$
19															$\frac{10}{19}$	$\frac{187}{380}$
20																$\frac{11}{20}$

IX. Kvantily pro Kolmogorův - Smirnovovův jednovýběrový test (KSJT)

Jsou tabelovány kvantily Kolmogorova-Smirnova testu $d_{1-\alpha}$ definované vztahem $P(D \leq d_{1-\alpha}) \geq 1-\alpha$ pro uvedené hodnoty $1-\alpha$.

n	0,90	0,95	0,99
1	0,950	0,975	0,995
2	0,776	0,842	0,929
3	0,636	0,708	0,829
4	0,565	0,624	0,734
5	0,509	0,563	0,669
6	0,468	0,519	0,617
7	0,436	0,483	0,576
8	0,410	0,454	0,542
9	0,387	0,430	0,513
10	0,369	0,409	0,489
11	0,352	0,391	0,468
12	0,338	0,375	0,449
13	0,325	0,361	0,432
14	0,314	0,349	0,418
15	0,304	0,338	0,404
16	0,259	0,327	0,392
17	0,286	0,318	0,380
18	0,279	0,309	0,371
19	0,271	0,301	0,361
20	0,265	0,294	0,352
21	0,259	0,287	0,344
22	0,253	0,281	0,337
23	0,247	0,275	0,330
24	0,242	0,269	0,323
25	0,238	0,264	0,317

n	0,90	0,95	0,99
26	0,233	0,259	0,311
27	0,229	0,254	0,305
28	0,225	0,250	0,300
29	0,221	0,246	0,295
30	0,218	0,242	0,290
31	0,214	0,238	0,285
32	0,211	0,234	0,281
33	0,208	0,231	0,277
34	0,205	0,227	0,273
35	0,202	0,224	0,269
36	0,199	0,221	0,265
37	0,196	0,218	0,262
38	0,194	0,215	0,258
39	0,191	0,213	0,255
40	0,189	0,210	0,252
41	0,187	0,208	0,249
42	0,185	0,205	0,246
43	0,183	0,203	0,243
44	0,181	0,201	0,241
45	0,179	0,198	0,238
46	0,177	0,196	0,235
47	0,175	0,194	0,233
48	0,173	0,192	0,231
49	0,171	0,190	0,228
50	0,170	0,188	0,226

Pro velká n přibližně:

$d_{0,9} = 1,22/(n)^{1/2}$ $\quad d_{0,95} = 1,36/(n)^{1/2}$ $\quad d_{0,99} = 1,63/(n)^{1/2}$

Použitá literatura

ANDĚL, J., 1985. *Matematická statistika*. Praha: SNTL/ALFA.

ANDĚL, J., 2007. *Matematitika náhody.* Praha: Matfyzpres. ISBN 80-7378-004-6.

ANDĚL, J., 1993. *Statistické metody.* Praha: Matfyzpres.

ANDĚL, J., 2007. *Základy matematické statistiky.* Praha: Matfyzpres. ISBN 80-7378-001-1.

ARLTOVÁ, M. a kol., 2003. *Příklady k předmětu Statistika A*. Praha: VŠE. ISBN 80-245-0178-3.

BLATNÁ, D., 1996. *Neparametrické metody. Testy založené na pořádkových a pořadových statistikách.* Praha: VŠE. ISBN 80-7079-607-3.

CYHELSKÝ, L. a kol., 2001. *Elementární statistická analýza.* Praha: Management Press. ISBN 80-7261-003-1.

ČERMÁKOVÁ, A. a F. STŘELEČEK, 1995. *Statistika I*. České Budějovice: JU zemědělská fakulta. ISBN 80-7040-126-5.

GIBILISCO, S., 2009. *Statistika bez předchozích znalostí.* Brno: Computer Press. ISBN 978-80-251-2465-9.

GUJARATI, D.N., 1992. *Essentials of Econometrics*. New York: Mc Grow-Hill. ISBN 0-07-112624-4.

HEBÁK, P., 1995. *Testování statistických hypotéz.* Praha: VŠE Praha. ISBN 80-7079-294-9.

HEBÁK, P. a kol., 2004. *Praktikum k výuce matematické statistiky II. Testování hypotéz* Praha: Oeconomica. ISBN 80-245-0721-8.

HEBÁK, P. a J. KAHOUNOVÁ, 1994. *Počet pravděpodobnosti v příkladech.* Praha: Informatorium. ISBN 80-85427-48-6.

HINDLS, R. a kol., 1999. *Analýza dat v manažerském rozhodování*. Praha: Grada. ISBN 80-7169-255-7.

HINDLS, R. a kol., 2000. *Metody statistické analýzy pro ekonomy*. Praha: Management Press. ISBN 80-7261-013-9.

HINDLS, R. a kol., 2007. *Statistika pro ekonomy*. Praha: Professional Publishing. ISBN 978-80-86946-43-6.

JAROŠOVÁ, E., 1994. *Statistika B. Řešené příklady.* Praha: VŠE. ISBN 80-7079-328-7.

KOMAREK, A. 2012. *Package 'vsePackage' (manual)* Praha: http://www.karlin.mff. cuni.cz/ ~komarek.

KOMÁREK, A. a KOMÁRKOVÁ, L. 2007. *Statistická analýza závislostí s příklady v R*. Praha: VŠE Nakladatelství Oeconomica. ISBN 978-80-245-1226-6.

KOMÁRKOVÁ, L. a kol., 2007. *Základy analýzy dat a statistického úsudku s příklady v R*. Praha: VŠE Nakladatelství Oeconomica. ISBN 978-80-245-1227-3.

MAREK, L. a kol., 2007. *Statistika pro ekonomy – aplikace*. Praha: Professional Publishing. ISBN 978-80-86446-40-5.

MINAŘÍK, B., 1995. *Statistika I pro ekonomy a manažery*. Brno: Mendelova zemědělská a lesnická universita. ISBN 80-7157-166-0.

NEWBOLD, P., 1991. *Statistics for business and economies.* New York: Prentice-Hall Int. Englewood Clifis. ISBN 0-13850645-0.

PECÁKOVÁ, I. *Statistika v terénních průzkumech.* Praha: Professional Publishing, 2008. ISBN 978-80-86946-74-0.

ŘEZANKOVÁ, H. a T. LÖSTER, 2009. *Úvod do statistiky*. Praha: Oeconomica. ISBN 978-80-245-1514-4.

SEGER, J. a R. HINDLS, 1995. *Statistické metody v tržním hospodářství*. Praha: Vicoria Publishing. ISBN 80-7187-058-7.

SHAPIRO, S. S. a M. B. WILKS, M. B., 1965. An analysis of variance test for normality (complete samples). *Biometrica.* 52(3-4), 591-611. doi:10.1093/biomet/52.3-4.591.

STUCHLÝ, J., 2000. *Ekonometrie*. J. Hradec: VŠE.

STUCHLÝ, J., 2011. *Referenční karta pro systém R*. České Budějovice: VŠTE Č. Budějovice. (v elektronické formě – viz https://is.vstecb.cz/auth/www/6384/).

STUCHLÝ, J., 1999a. *Statistika I. Cvičení ze statistických metod pro managery.* Praha: VŠE. ISBN 80-7079-754-1.

STUCHLÝ, J., 1999b. *Statistika II Cvičení ze statistických metod pro manažery*. J. Hradec: VŠE. ISBN 80-7079-035-0.

STUCHLÝ, J., 2004. *Statistické metody pro manažerské rozhodování*. J. Hradec: VŠE. ISBN 80-245-0153-8.

SVATOŠOVÁ, L. a M. PRÁŠILOVÁ, 2009. *Statistické metody v příkladech.* Praha: ČZU. ISBN 978-80-213-1673-7.

WISNIEWSKI, M. *Metody manažerského rozhodování*. Praha: Grada, 1996. ISBN 80-7169-089-9.

WONNACOT, T. H. a R. J. WONNACOT, 1993. *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing. ISBN 80-85605-09-0.

Rejstřík a česko-anglický slovník