

Testování hypotéz v programovém prostředí Matlab

Hypothesis testing with Matlab

Martin Kovářik

Bakalářská práce
2011

 Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně

Fakulta aplikované informatiky

akademický rok: 2010/2011

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Ing. Martin KOVÁŘÍK**

Osobní číslo: **A07551**

Studijní program: **B 3902 Inženýrská informatika**

Studijní obor: **Informační a řídicí technologie**

Téma práce: **Statistické testování hypotéz v programovém prostředí Matlab**

Zásady pro vypracování:

1. Popište statistické testování hypotéz a metody popisné statistiky.
2. Vypracujte vybrané statistické testy – parametrické i neparametrické.
3. Představte statistický toolbox pro Matlab.
4. Realizujte sadu vzorových příkladů.
5. Zpracujte případovou studii.
6. Provedte vyhodnocení prostřední Matalab pro daný účel.

Rozsah bakalářské práce:

Rozsah příloh:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

1. ANDĚL, J. **Základy matematické statistiky. 2. vyd. Praha: Matfyzpress, 2007. 358 s. ISBN 80-7378-001-1.**
2. HANOUSEK, J. CHARAZMA, P. **Moderní metody zpracování dat. Matematická statistika pro každého. 1. vyd. Praha: Edice EDUCA'99, 1992. 216 s. ISBN 80-85623-31-5.**
3. HEBÁK, P. **Vícerozměrné statistické metody. 1. vyd. Praha: Informatorium, spol. s.r.o., 2004. 236 s. ISBN 80-7333-025-3.**
4. HENDL, J. **Přehled statistických metod: analýza a metaanalýza dat. 3. vydání. Praha: Nakladatelství Portál, s.r.o., 2009. 687 s. ISBN 978-80-7367-482-3.**
5. MELOUN, M., MILITKÝ, J. **Kompendum statistického zpracování dat. 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2006. 982 s. ISBN 80-200-1396-2.**

Vedoucí bakalářské práce:

Ing. Petr Šilhavý, Ph.D.

Ústav počítačových a komunikačních systémů

Datum zadání bakalářské práce:

25. února 2011

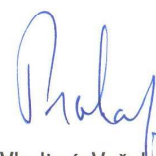
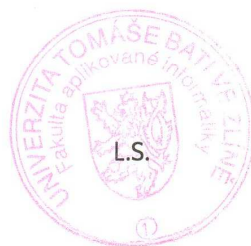
Termín odevzdání bakalářské práce:

7. června 2011

Ve Zlíně dne 25. února 2011



prof. Ing. Vladimír Vašek, CSc.
děkan



prof. Ing. Vladimír Vašek, CSc.
ředitel ústavu

ABSTRAKT

Cílem této bakalářské práce je popis vybraných statistických testů a jejich použití v programovém prostředí Matlab a vytvořit funkční programy, které budou sloužit jako studijní pomůcka při výuce statistických předmětů. Text práce začíná představením statistického toolboxu a úvodem do inferenční statistiky. Zbytek teoretické části je rozdělen na dvě hlavní části: parametrické a neparametrické testy. Použití jednotlivých testů je ilustrováno na příkladech s náhodně generovanými nebo reálnými daty. Jednotlivé příklady jsou prováděny v programu Matlab. V praktické části této práce bude v kapitole 5 popsán princip testování statistických hypotéz pomocí metody Monte Carlo. V kapitole 6 budou představeny vybrané statistické programy a aplikace vytvořené v Matlabu, které mohou sloužit jako studijní pomůcka při výuce matematické statistiky. Matlab jsem si zvolil z toho důvodu, že je všestranně zaměřený a není primárně určen jen pro jednu oblast použití. Díky statistickému toolboxu a vysokému výkonu systému se stává výborným pomocníkem při rozsáhlých analýzách.

Klíčová slova:

Matlab, statistický toolbox, inferenční statistika, parametrické a neparametrické testy, bootstrap, Box-Coxova transformace, Hallova transformace, normalita dat, aproximace rozdělení

ABSTRACT

The aim of this bachelor thesis is a description of selected statistical tests, their use in Matlab and to develop functional programs that will serve as a learning tool for teaching statistical courses. Thesis begins with the introduction of the statistics toolbox and an introduction to inferential statistics. The end of the theoretical part is divided into two main parts: parametric and nonparametric tests. Using different tests are illustrated with examples randomly generated and real data. Individual examples are implemented in Matlab. The practical part of this work will be described in Chapter 5 by the principle of testing hypotheses statistických using Monte Carlo. In Chapter 6 will be presented selected statistical programs and applications developed in Matlab, which can serve as a learning tool for teaching mathematical statistics. I chose MATLAB because it is broadly focused and is not primarily intended for only one application area. Due to the statistical toolbox and high performance of the system becomes a great help in large-scale analysis.

Keywords:

Matlab Statistics Toolbox, Inferential Statistics, Parametric and Nonparametric Tests, Bootstrap, Box-Cox Transformation, Hall Transformation, Data Normality, Distribution Approximations

Touto cestou bych rád poděkoval vedoucímu bakalářské práce panu Ing. Petru Šilhavému, Ph.D. za odborné vedení, cenné informace a připomínky, které mi při vypracování bakalářské práce poskytl.

Prohlašuji, že

- beru na vědomí, že odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na bakalářské práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně

.....

podpis diplomanta

OBSAH

ÚVOD	10
TEORETICKÁ ČÁST	11
1 POPIS STATISTICKÉHO TOOLBOXU	12
1.1 NADSTAVBA PRO STATISTICKÉ VÝPOČTY (STATISTICS TOOLBOX).....	12
1.2 PŘEHLED ZÁKLADNÍCH A NEJPOUŽÍVANĚJŠÍCH FUNKCÍ STATISTICS TOOLBOXU PŘI TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ.....	15
2 ZÁKLADNÍ POJMY Z TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ	16
2.1 POSTUP PŘI TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ.....	18
2.1.1 Definice P-hodnoty.....	22
2.1.2 Intervaly spolehlivosti pro střední hodnotu.....	24
3 VYBRANÉ PARAMETRICKÉ TESTY	27
3.1 TEST HYPOTÉZY O STŘEDNÍ HODNOTĚ μ PŘI ZNÁMÉM ROZPTYLU σ^2	27
3.2 TEST HYPOTÉZY O STŘEDNÍ HODNOTĚ μ PŘI NEZNÁMÉM ROZPTYLU σ^2	29
3.3 TEST HYPOTÉZY O ROZPTYLU.....	30
3.4 TEST HYPOTÉZY O SHODĚ DVOU STŘEDNÍCH HODNOT.....	32
3.5 PÁROVÝ T-TEST.....	36
3.6 TEST HYPOTÉZY O SHODĚ DVOU ROZPTYLŮ.....	37
4 VYBRANÉ NEPARAMETRICKÉ TESTY	39
4.1 χ^2 TEST V KONTINGENČNÍ TABULCE.....	40
4.2 JEDNOVÝBĚROVÉ NEPARAMETRICKÉ TESTY.....	43
4.2.1 Znaménkový test.....	43
4.2.2 Wilcoxonův test.....	44
4.3 NEPARAMETRICKÉ TESTY PRO DVA ZÁVISLÉ A PRO DVA NEZÁVISLÉ VÝBĚRY.....	46
4.3.1 Znaménkový test pro dva závislé výběry.....	46
4.3.2 Wilcoxonův test pro dva závislé výběry.....	46
4.3.3 Wilcoxonův dvouvýběrový test pro nezávislé výběry (Mannův- Whitneyův test).....	48
4.4 TESTY O TYPU ROZDĚLENÍ.....	50
4.4.1 χ^2 test dobré shody.....	50
4.4.2 Kolmogorovův-Smirnovův test.....	52
4.4.3 Lillieforsův test.....	53
4.5 NEPARAMETRICKÉ MÍRY TĚSNOSTI ZÁVISLOSTI.....	54
PRAKTICKÁ ČÁST	57
5 TESTOVÁNÍ HYPOTÉZ METODOU MONTE CARLO	58
6 UKÁZKY STATISTICKÝCH PROGRAMŮ	68

6.1	DVOUROZMĚRNÉ NORMÁLNÍ ROZDĚLENÍ.....	68
6.2	CENTRÁLNÍ LIMITNÍ VĚTY.....	69
	O centrální limitní větě.....	70
6.3	MÍRY ASYMETRIE.....	75
6.4	TESTOVÁNÍ NORMALITY DAT.....	78
6.4.1	Vícerozměrný test normality.....	80
6.4.2	Anderson-Darlinův test normality dat.....	81
6.5	TESTOVÁNÍ NORMALITY POMOCÍ EXPLORATORNÍCH GRAFŮ.....	83
6.5.1	Test normality pomocí kvantilové funkce.....	84
6.5.2	Test normality pomocí rankitového grafu s robustní přímkou.....	85
6.6	ODHAD INTERVALŮ SPOLEHLIVOSTI PRO STŘEDNÍ HODNOTU A ROZPTYL.....	86
6.6.1	Interval spolehlivosti pro střední hodnotu a rozptyl.....	87
6.6.2	Odhad intervalu spolehlivosti metodou bootstrap.....	88
6.7	TRANSFORMACE ZLEPŠUJÍCÍ ROZDĚLENÍ DAT.....	91
6.7.1	Zpracování transformovaných dat.....	92
6.7.2	Box–Coxova mocninná transformace.....	93
6.8	METODA BOOTSTRAP.....	98
6.8.1	Odhad z asymptotické normality.....	100
6.8.2	Percentilový odhad.....	100
6.8.3	Studentizovaný odhad.....	101
6.8.4	Vyhlazený odhad.....	101
6.8.5	Generace Bootstrap výběrů.....	101
6.8.6	Realizace postupu Bootstrap.....	102
6.9	ZPRACOVÁNÍ VÝBĚRŮ Z ASYMETRICKÝCH ROZDĚLENÍ.....	103
6.9.1	Omezení asymetrie rozdělení Studentovy statistiky.....	103
6.9.2	Výpočet korigovaného průměru.....	105
	ZÁVĚR.....	108
	ZÁVĚR V ANGLIČTINĚ.....	109
	SEZNAM POUŽITÉ LITERATURY.....	110
	SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK.....	111
	SEZNAM OBRÁZKŮ.....	112
	SEZNAM PŘÍLOH.....	114

ÚVOD

V současnosti nabývá na významu matematická (neboli pravděpodobnostní) statistika, tj. ta oblast statistiky, která nepracuje se základními statistickými soubory, ale pouze s výběry. Z nich pak metodami statistické indukce usuzuje na vlastnosti základních statistických souborů. Většina výzkumů pracuje s poznatky mnoha přírodních i společenských věd a využívá všech možností vyspělé výpočetní techniky a programového vybavení. V procesu sběru a analýzy získaných informací má statistika svou nezastupitelnou úlohu. Matematická statistika disponuje postupy a metodami, které umožňují za určitých podmínek zobecnit výsledky výběrového šetření na základní statistický soubor, a proto pracuje téměř vždy s výběrovými informacemi. V mnoha případech je tento způsob z technických, organizačních, ekonomických a jiných důvodů jediný možný.

Text práce začíná představením statistického toolboxu a úvodem do problematiky testování statistických hypotéz. Zbytek teoretické části je rozdělen na dvě hlavní části: parametrické a neparametrické testy. Použití jednotlivých testů je ilustrováno na příkladech s náhodně generovanými nebo reálnými daty. Jednotlivé příklady jsou prováděny v programu MATLAB 7.9 (R2009b).

V praktické části této práce bude v kapitole 5 popsán princip testování statistických hypotéz pomocí metody Monte Carlo. V kapitole 6 budou představeny vybrané statistické programy a aplikace vytvořené v Matlabu, které mohou sloužit jako studijní pomůcka při výuce matematické statistiky.

Při volbě vhodného jazyka pro řešení technických úloh je k dispozici řada možností. Ve třídě systémů pro technické výpočty je poměrně bohatá nabídka od řešičů úloh (DERIVE, TK Solver, EUREKA) přes kompaktnější systémy (MATHCAD, MuPAD) až k jazykům (MATLAB, S Plus, MATHEMATICA, MAPLE). Ke zpracování této bakalářské práce byl zvolen MATLAB z toho důvodu, jelikož má obecně jednoduché použití orientované na práci s poli a maticemi, je zde možnost interaktivní práce, má kvalitní grafiku (vědeckou), kvalitní numeriku (rychlá, robustní a přesná) a je zde možnost programování. Co je velmi důležité, tak MATLAB umožňuje orientaci na inženýrské problémy (toolboxy) a má rozšíření pro UNIX, DOS a Windows.

I. TEORETICKÁ ČÁST

1 POPIS STATISTICKÉHO TOOLBOXU

Otevřená architektura MATLABu vedla ke vzniku knihoven funkcí, nazývaných toolboxy, které rozšiřují použití programu v příslušných vědních a technických oborech. Tyto knihovny, navržené a v jazyce MATLABu napsané nejvýznačnějšími světovými odborníky, nabízejí předzpracované specializované funkce, které je možno rozšiřovat, modifikovat, anebo jen čerpat informace z přehledně dokumentovaných algoritmů. Statistics Toolbox nabízí rozsáhlý soubor nástrojů pro práci s daty. Zahrnuje funkce a interaktivní nástroje pro modelování dat, analýzu trendů, simulaci stochastických systémů a vývoj algoritmů pro statistiku. Statistics Toolbox podporuje širokou škálu úloh od výpočtů základní popisné statistiky až po vývoj a vizualizaci mnohorozměrných nelineárních modelů. Dále nabízí velké množství statistických grafů a interaktivních grafických nástrojů, jako je polynomiální prokládání a modelování výsledkových ploch. Veškeré funkce Statistics Toolboxu jsou napsány v otevřeném jazyce MATLABu, takže je možné algoritmy zobrazit, upravovat zdrojový kód nebo vytvářet vlastní uživatelské funkce. Z našeho hlediska je důležité, že jedním z komerčně dostupných toolboxů je programová nadstavba pro statistické výpočty - Statistics ToolBox. Této nadstavbě a jejímu použití pro vědecké i inženýrské účely bude věnována následující podkapitola.

1.1 Nadstavba pro statistické výpočty (Statistics ToolBox)

Nadstavba pro statistické výpočty Statistics ToolBox obsahuje více, než 200 *m*-souborů, které podporují výpočty v následujících oblastech.

1. *PROBABILITY DISTRIBUTIONS* – Statistics Toolbox podporuje 20 rozdělení pravděpodobnosti diskrétní a spojité náhodné veličiny. Pro každé rozdělení má 5 asociovaných funkcí, jsou to: pravděpodobnostní funkce (pdf), distribuční funkce (cdf), inverzní distribuční funkce, generátor náhodných čísel, střední hodnotu a rozptyl jako funkci parametru.
2. *DESCRIPTIVE STATISTICS* – stanovení statistických charakteristik souborů.
3. *LINEAR MODELS* – lineární regresní analýza, ANOVA.
4. *NONLINEAR MODELS* – funkce pro interaktivní predikci, nelineární regresní analýzu a vizualizaci pro vícerozměrná data.

5. *HYPOTHESIS TESTS* – testování statistických hypotéz, t -test, z -test aj.
6. *MULTIVARIATE STATISTICS* – metody pro statistickou analýzu vícerozměrných dat.
7. *STATISTICAL PLOTS* – statistické grafy např. boxplot, histogram aj.
8. *DEMOS* - demonstrační výukové úlohy.
9. *DATA* - demonstrační datové soubory.

Ve statistickém toolboxu MATLABu jsou implementovány funkce pro práci s následujícími 6-ti druhy rozdělení diskrétní náhodné veličiny: Binomické, geometrické, hypergeometrické, negativní binomické, Poissonovo, rovnoměrné diskrétní.

A funkce pro práci s následujícími 14-ti druhy rozdělení spojité náhodné veličiny: Beta, Pearsonovo chí-kvadrát, exponenciální, Fischerovo F -rozdělení, Gama, Gaussovo normální, Studentovo t -rozdělení, rovnoměrné spojité, Weibullovo, lognormální, Rayleighovo rozdělení, necentrováné chí-kvadrát, necentrováné F -rozdělení, necentrováné t -rozdělení.

Pro každý implementovaný typ rozdělení je možno zobrazit distribuční funkci a funkci rozložení hustoty pravděpodobnosti, provádět s nimi výpočty popř. vypočítat jejich charakteristiky. Je rovněž možno používat inverzní distribuční funkci, která stanoví hodnoty určitého rozdělení podle zadaných pravděpodobností. Dále se dají také zobrazit velmi jednoduše jednotlivé typy rozdělení dle zadaných parametrů. K použití se rovněž nabízejí generátory náhodných čísel pro každý typ rozdělení. V demonstračním bloku *DEMOS* je uvedena speciální funkce *distool*, která umožňuje grafickou demonstraci jednotlivých typů rozdělení. Je možno volit alternativní zobrazení distribuční funkce nebo funkce rozdělení hustoty všech typů implementovaných rozdělení. Je možno interaktivně měnit parametry studovaného rozdělení a zjišťovat jeho funkční hodnoty pro různé hodnoty nezávisle proměnné.

Ve statistickém toolboxu MATLABu jsou přímo k dispozici funkce, vypočítávající následující charakteristiky polohy: aritmetický průměr, geometrický průměr, harmonický průměr, medián, kvantily a aritmetický průměr bez extrémních hodnot a dále charakteristiky rozptýlení: rozptyl, směrodatnou odchylku, průměrnou odchylku, variační rozpětí a interkvartilové rozpětí aj.

Prakticky velmi důležité jsou možnosti grafické prezentace výsledků zpracování statistického souboru. Tak lze znázornit histogramy absolutních četností, absolutních kumulovaných četností, krabicový graf (prezentace polohy 1. kvartilu, mediánu a 3. kvartilu), odlehlé hodnoty, vrubový krabicový graf (s prezentací konfidenčního intervalu aritmetického průměru). Pro zjištění, zda výběrový soubor pochází ze základního souboru s normálním rozložením hustoty pravděpodobnosti, slouží graf normálního rozložení. K dispozici je funkce, umožňující zjistit, zda mají dva výběrové soubory stejné rozdělení (kvantil-kvantilový graf). Plnou čarou jsou spojeny 1. a 3. kvartily (dolní a horní kvartil). Výběry mají pravděpodobně stejné rozdělení, je-li závislost na první pohled lineární.

Jako demonstrační funkce pro generování náhodných hodnot s různými typy rozdělení a vykreslování histogramů četnosti je v demonstračním bloku *DEMOS* připravena funkce *randtool*. Při studiu daného rozdělení je možno interaktivně měnit parametry rozdělení a rozsah souboru, ukládat data do výstupních souborů aj.

Funkce MATLABu umožňují dále provádět analýzu lineárních regresních modelů. K dispozici jsou především funkce pro analýzu rozptylu (ANOVA - Analysis of Variance). Je možno je použít buď jako postup pro zjištění zdrojů variability u lineárních modelů, nebo jako samostatných technik. Ze statistického hlediska je možno tyto funkce chápat jako speciální případ regresní analýzy, kdy vysvětlující proměnné mají pouze binární charakter a mohou nabývat pouze hodnot 0 nebo 1. Při analýze zdrojů variability máme možnost vyšetřovat výběrový soubor při uvážení jednoho vlivu (faktoru) pomocí funkce pro jednofaktorovou (one-way) analýzu rozptylu, dvoufaktorová (two-way) analýza rozptylu umožňuje zkoumání vlivů dvou faktorů. Pro zkoumání vlivu faktorů na variabilitu se provádějí testy hypotéz o jejich významnosti. Funkce umožňují rovněž porovnávání dvou či více výběrů. Pro zobrazení výsledku analýzy je pak např. k dispozici okno se skupinou odpovídajících krabicových (vrubových krabicových) grafů, které umožňují evidentní posouzení shodnosti resp. rozdílu středních hodnot jednotlivých výběrů.

Funkce pro vícenásobnou lineární regresi umožňuje získat regresní závislost pro predikční účely. K dispozici je graf, znázorňující 95% konfidenční intervaly residuí.

Jako demonstrační funkce pro interaktivní polynomiální aproximaci souboru s možností predikce jeho hodnot je v bloku *DEMOS* k dispozici funkce *polytool*. Tato funkce vytváří interaktivní grafické prostředí pro křivkovou aproximaci polynomy různého stupně.

Významné jsou dále funkce, umožňující testování statistických hypotéz. Je možno provádět testy hypotéz o rozptylu (F -test), testy hypotéz o střední hodnotě (t -test), testy významnosti rozdílu párových hodnot a testy dobré shody. [8]

1.2 Přehled základních a nejpoužívanějších funkcí Statistics Toolboxu při testování statistických hypotéz

Není v možnostech této práce popsat veškeré funkce Statistics Toolboxu. Zaměřím se jen na ty, které se používají při testování statistických hypotéz. V příloze P I je uveden popis funkce včetně příkladu syntaxe použití. Výpis všech funkcí obsažených ve Statistics Toolboxu je uveden v souboru Contents.m v adresari ...\\Toolbox\\Stats. Bližší popis těchto funkcí je možno získat pomocí nápovědy (*help*) přímo v Matlabu, nebo z manuálu k Statistics Toolboxu.

2 ZÁKLADNÍ POJMY Z TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Statistickou hypotézou se rozumí určitý předpoklad o parametrech či tvaru rozdělení zkoumaného znaku. Tento předpoklad se může týkat charakteristik rozdělení náhodné veličiny v základním souboru nebo může být obecnější a vztahovat se pouze k zákonu rozdělení náhodné veličiny (k distribuční funkci, k pravděpodobnostní funkci nebo k hustotě pravděpodobnosti), k náhodnosti, nezávislosti apod.

Předpoklady, které tvoří statistickou hypotézu, se opírají o zkušenosti a dřívější informace a nevycházejí z náhodného výběru. Ten je základem k ověření statistické hypotézy a vlastnímu induktivnímu závěru.

Statistickou hypotézu mohou představovat např. výroky (věty):

- náhodný výběr x_1, x_2, \dots, x_n pochází ze základního souboru s normálním rozdělením
- všechny hodnoty v náhodném výběru x_1, x_2, \dots, x_n pocházejí z jednoho základního souboru s určitým rozdělením
- parametr Poissonova rozdělení má hodnotu 2
- rozptyly daných rozdělení v k základních souborech jsou stejné

Uvedené hypotézy mají v praxi bezprostřední interpretaci. Statistické hypotézy se formulují tak, aby měly interpretaci, která po ověření jejich platnosti umožní se rozhodnout. Jejich ověřování se děje pomocí testů. **Test statistické hypotézy** je pravidlo, které na základě výsledků zjištěných z náhodného výběru objektivně předepisuje rozhodnutí, má-li být ověřovaná hypotéza zamítnuta či nikoli. [6], [11]

Podstatu rozhodnutí můžeme formulovat takto:

Statistická hypotéza se týká základního souboru, který přesně neznáme. Ze základního souboru vezmeme náhodný výběr, který odráží poměry v základním souboru, a proto by se měl chovat podobně jako základní soubor specifikovaný hypotézou. Je-li chování náhodného výběru jiné, usuzujeme z toho, že pochází spíše z jiného základního souboru, než jaký specifikuje hypotézu, a proto tuto hypotézu zamítáme.

Při testu statistické hypotézy se rozlišuje **testovaná (nulová) hypotéza H_0** a **alternativní hypotéza H_1** (nebo někdy H_A). Testovaná hypotéza je hypotéza, o níž má test rozhodnout,

zda se zamítne či nikoli. Alternativní hypotéza je ta, kterou přijmeme, zamítneme-li hypotézu testovanou.

Je-li alternativní hypotéza H_1 formulována některou z nerovností

$$G > G_0, \quad G < G_0, \quad G_1 > G_2, \quad G_1 < G_2$$

znamená to, že je dána jedním intervalem hodnot G , a pak mluvíme o **jednostranném testu**. Má-li alternativní hypotéza formu nerovnosti $G \neq G_0$ (tj. buď $G > G_0$ nebo $G < G_0$), resp. $G_1 \neq G_2$ (tj. buď $G_1 > G_2$ nebo $G_1 < G_2$), znamená to, že je dána dvěma intervaly hodnot G a pak mluvíme o **oboustranném testu**. Jak budeme v případě jednostranných a oboustranných testů o vztahu charakteristiky (parametru) základního souboru G a konstanty G_0 , resp. o vztahu dvou charakteristik G_1 a G_2 , interpretovat nulovou hypotézu H_0 ? Chápeme-li hypotézu H_0 jako základ pro rozhodnutí, specifikujeme ji jako doplněk (opak) alternativní hypotézy H_1 , tj.

$$H_0: G \leq G_0 \quad \text{v případě } H_1: G > G_0$$

$$H_0: G \geq G_0 \quad \text{v případě } H_1: G < G_0$$

$$H_0: G = G_0 \quad \text{v případě } H_1: G \neq G_0$$

Protože při testování hypotézy jde o úsudek prováděný z údajů získaných náhodným výběrem, můžeme se ve svých úsudcích dopustit i chybných závěrů. Buď zamítneme nulovou hypotézu H_0 , ačkoliv ve skutečnosti platí, pak se dopouštíme tzv. **chyby prvního druhu**. Pravděpodobnost této chyby značíme α . Druhá možnost chybného závěru spočívá v tom, že přijmeme nulovou hypotézu H_0 , ačkoliv ve skutečnosti platí alternativní hypotéza H_1 , v tom případě se dopouštíme tzv. **chyby druhého druhu**, kterou zpravidla označujeme β . Pravděpodobnost $1-\beta$ se nazývá **síla testu**. Síla testu tedy vlastně vyjadřuje, s jakou pravděpodobností zamítneme nulovou hypotézu, platí-li alternativní hypotéza H_1 , jinak řečeno udává pravděpodobnost, že se nedopustíme chyby II. druhu. Klasický přístup k testování statistických hypotéz začíná tím, že si zvolíme tzv. hladinu významnosti v přijatelné výši (nejčastěji 5%). Testovací postup je odvozen tak, aby při dané hladině významnosti zajišťoval minimální pravděpodobnost chyby II. druhu a tím maximální sílu testu. [11]

Popis standardního testu zejména uvádí, jaké použít v dané situaci **testovací kritérium T** . Označme jej symbolem T . Množinu hodnot, jichž může testovací kritérium nabýt,

nazýváme výběrový prostor a označujeme S . Je docela logické, že dříve, než vypočteme hodnotu testovacího kritéria pro daný výběr nebo dokonce dříve, než výběr provedeme, chceme mít připraveno pravidlo umožňující rozhodnout ve prospěch H_0 nebo ve prospěch H_1 . To bude nejjednodušeji vyjádřeno, když rozdělíme výběrový prostor S na dva podprostory:

- a) podprostor V obsahující hodnoty svědčící ve prospěch H_0 , tzv. obor přijetí,
- b) podprostor W obsahující hodnoty svědčící ve prospěch H_1 , tzv. kritický obor.

Oba podprostory vyplňují zcela prostor S a nepřekrývají se, tedy

$$V \cup W = S$$

$$V \cap W = \emptyset$$

Hranice oddělující kritický obor a obor přijetí nazýváme **kritické kvantily**. [5]

2.1 Postup při testování statistických hypotéz

Klasický postup při testování statistických hypotéz bývá formálně členěn do šesti etap.

I. Formulace nulové hypotézy H_0 a alternativní hypotézy H_1

Formulujeme dvojici hypotéz H_0 a H_1 . Obě hypotézy se tím či oním způsobem týkají pravděpodobnostního rozdělení studovaného znaku X , nejčastěji parametrů tohoto rozdělení. Běžné pojetí testování hypotéz vyžaduje, aby nulová hypotéza byla **jednoduchá**, tj. jednoznačně specifikovala rozdělení studovaného znaku. Nejčastěji má podobu rovnice týkající se některého parametru. Dále se vyžaduje taková formalizace řešené úlohy, aby to, co chceme testem prokázat, bylo vyjádřeno v alternativní hypotéze H_1 . Ta pak bývá obvykle **složená**, tj. nespecifikuje již rozdělení zkoumaného znaku jednoznačně. Buď může všeobecně popírat platnost nulové hypotézy, nebo jde o nějakou nerovnici, týkající se některého parametru.

II. Volba hladiny významnosti α

Zpravidla volíme 5%, výjimečně i 1%.

III. Volba testového kritéria

Testové kritérium je statistika, tedy funkce výběru. Výpočet její hodnoty je při testování hypotéz cílem zpracování výběru. Ukáže se, že abychom mohli úspěšně provést další etapu

testování, sestrojít kritický obor, potřebujeme znát rozdělení testového kritéria při platnosti H_0 .

IV. Sestrojení kritického oboru a nalezení kritického kvantilu

Kritický obor W bude tak velký, aby bylo zajištěno, že chyby prvního druhu se dopustíme jen ve 100α % případů. Pravděpodobnost, že zpracování výběru by mohlo dát výsledek (hodnotu testového kritéria) v kritickém oboru za podmínky platnosti nulové hypotézy, má být rovna předem zvolené hladině významnosti α .

V. Výpočet hodnoty testového kritéria

Dosavadní etapy testování mohly být provedeny ještě před vlastním pořízením dat. Nyní předpokládejme, že k dispozici je již náhodný výběr a přistoupíme k jeho zpracování. Vzorec pro výpočet hodnoty testového kritéria je znám, takže jen zvolíme vhodný algoritmus, výpočetní prostředky a zjistíme jeho hodnotu.

VI. Formulace výsledků testu

- a) zamítnout H_0 (přijmout H_1), jestliže vypočtená hodnota testové charakteristiky padne do kritického oboru,
- b) nezamítnout H_0 , jestliže vypočtená hodnota testové charakteristiky nepadne do kritického oboru. [11]

Výše už jsme zmínili, co znamená, dopustíme-li se chyby 1. a 2. druhu. Nyní, pro lepší pochopení si pojdme tyto skutečnosti matematicky formulovat a graficky znázornit.

Mějme náhodný výběr $X = (X_1, X_2, \dots, X_n)$ a testujeme hypotézu H_0 proti alternativě H_1 (nebo také označovanou H_A) na hladině významnosti α . K testování hypotézy použijeme statistiku $T(X)$ založenou na náhodném výběru X . Nechť $T(x)$ je hodnota testové statistiky při dané realizaci $x = (x_1, x_2, \dots, x_n)$ náhodného výběru. Množinu hodnot, kterých může testová statistika nabýt, nazýváme výběrový prostor a označujeme V .

Obor zamítnutí W_α nulové hypotézy H_0 pro danou hladinu významnosti α je určen tak, aby

$$P(T(X) \in W_\alpha | H_0) = \alpha, \quad (2.1)$$

(tj. pravděpodobnost, že testová statistika nabude hodnoty z kritického oboru za platnosti nulové hypotézy, je rovna α). Pravděpodobnost chyby prvního druhu α je tedy definována předchozím vztahem.

Pravděpodobnost chyby druhého druhu β je pak

$$\beta = P(T(X) \notin W_\alpha | H_A). \quad (2.2)$$

poznámka: doplňkem oboru zamítnutí W_α je $\overline{W_\alpha}$ a značí obor přijetí nulové hypotézy H_0 .

Rozhodovací pravidlo $d(T(X))$ pro test nulové hypotézy je následující:

$$d_w(T(x)) = \begin{cases} 1 & \text{pokud } T(x) \in W_\alpha \\ 0 & \text{pokud } T(x) \notin W_\alpha. \end{cases} \quad (2.3)$$

Je-li hodnota rozhodovacího pravidla rovna 1, pak hypotézu H_0 zamítáme, je-li hodnota rozhodovacího pravidla rovna 0, pak říkáme, že hypotézu H_0 nelze zamítnout.

Předpokládejme, že známe rozdělení $F(t)$ testové statistiky T za platnosti H_0 . Pak kritický obor W_α pro zadanou pravděpodobnost α vymezují **kritické hodnoty (kvantily)** t_α rozdělení testové statistiky následujícím způsobem:

$$\alpha = P(T > t_\alpha) = 1 - F(t_\alpha). \quad (2.4)$$

Označíme-li nejmenší možnou hodnotu testové statistiky t_{\min} a největší možnou hodnotu t_{\max} , pak v případě pravostranného testu bude kritický obor

$$W_\alpha = (t_\alpha, t_{\max}), \quad (2.5)$$

v případě levostranného testu

$$W_\alpha = (t_{\min}, t_{1-\alpha}) \quad (2.6)$$

a nakonec v případě dvoustranného testu

$$W_\alpha = (t_{\min}, t_{1-\alpha/2}) \cup (t_{\alpha/2}, t_{\max}) = W_{1,\alpha/2} \cup W_{2,\alpha/2}. \quad (2.7)$$

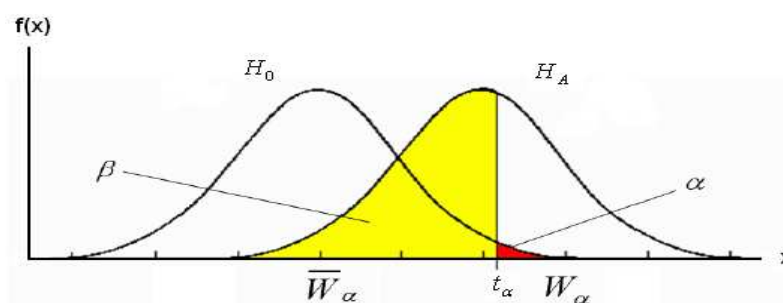
Obor přijetí $\overline{W_\alpha}$ je ve všech uvedených případech doplňkem kritického oboru $(W_\alpha \cup \overline{W_\alpha} = V)$. Pro jednoduchost budeme v dalším textu používat označení $T = T(X)$ pro

testovou statistiku a $t_c = T(x)$ pro její hodnotu vypočtenou z konkrétní realizace náhodného výběru. Hladina významnosti, tj. pravděpodobnost chyby prvního druhu α má ten praktický význam, že při mnoha opakovaných realizacích náhodného výběru (např. řádově v tisících) a současné platnosti testované hypotézy H se v přibližně $100\alpha\%$ testech této hypotézy zmýlíme, tedy zamítneme platnou hypotézu. Podobně, když hypotéza H neplatí, tak se v přibližně $100\beta\%$ testech zmýlíme a nezamítneme ji. Avšak snížením hladiny významnosti α se při nezměněném rozsahu statistického souboru n zvýší β a naopak, takže pro zvolenou hladinu významnosti α zajišťujeme snížení β zvýšením rozsahu n . Riziko chyb prvního i druhého druhu nelze v reálných úlohách eliminovat, pouze je můžeme snížit. [1]

Tab. 2.1. Výsledky testu hypotéz (skutečnost versus rozhodnutí)

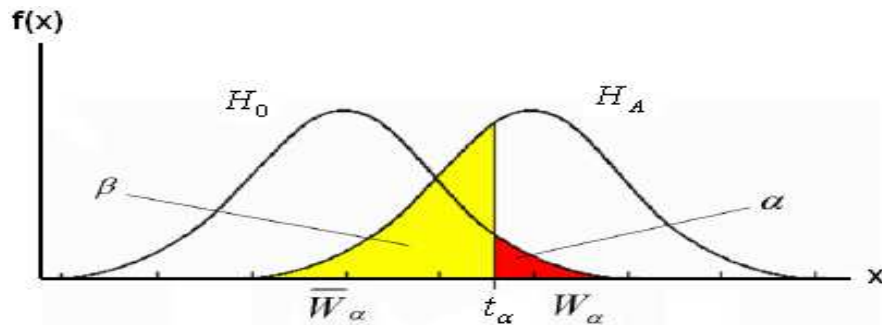
Skutečnost	Rozhodnutí	
	H_0 se nezamítá	H_0 se zamítá
H_0 je pravdivá	správné rozhodnutí Pravděpodobnost = $1 - \alpha$	chyba I. druhu pravděpodobnost = α
H_0 je nepravdivá	Chyba II. druhu Pravděpodobnost = β	správné rozhodnutí pravděpodobnost = $1 - \beta$

Vztah mezi α a β je ilustrováno na následujícím obrázku, kde pro jednoduchost je i alternativní hypotéza H_A jednoduchá. Na tomto obrázku křivky vlevo odpovídají hustotě (pravděpodobnostní funkci) testového kritéria T při platnosti hypotézy H_0 a křivky vpravo odpovídají hustotě (pravděpodobnostní funkci) testového kritéria T při platnosti hypotézy H_A .



Obr. 2.1. Vztah chyby prvního a druhého druhu

Ukážeme ještě, že zvětšíme-li v tomto případě rozsah náhodného výběru n , zmenší se při stejném α hodnota β (že se tedy zvětší síla testu $1-\beta$). Protože rozptyl výběrového průměru \bar{X} je $\frac{\sigma^2}{n}$, při zvětšení n se tento rozptyl zmenší, tedy se „zúží“ oba tvary hustot (viz následující obrázek). Vidíme, že pravděpodobnost chyby 2. druhu β se nyní oproti předchozímu obrázku značně zmenšila a je dokonce menší než α .



Obr. 2.2. Vztah chyby prvního a druhého druhu při zvětšení n

Poznámka: Zjistili jsme značně obecnou vlastnost statistických testů, že totiž zhruba řečeno při zvětšování rozsahu náhodného výběru se zvětšuje síla testu $1-\beta$ (při stejné hladině významnosti α). Zároveň se při zvětšování n zvětšuje kritický obor W pro zamítnutí nulové hypotézy, jak je vidět z porovnání předchozích dvou obrázků. Bohužel však je rozsah výběru téměř vždy limitován praktickými omezeními (přílišné finanční nebo časové náklady, přílišná pracnost, případně fakt, že výběr je již proveden, nemohli jsme jej ovlivnit a nelze jej opakovat). [3]

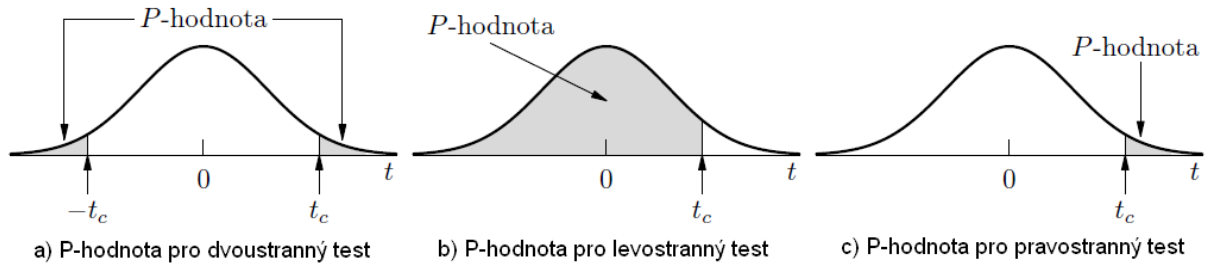
2.1.1 Definice P-hodnoty

Nechť T je testová statistika, t_c je pozorovaná hodnota testové statistiky. Pak p -hodnota testu hypotézy se rovná

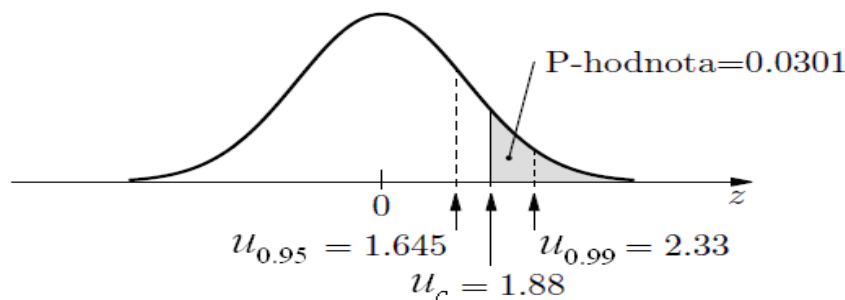
- $2 \cdot \min \{P(T \leq t_c), P(T \geq t_c)\}$ pro dvoustranný test,
- $P(T \leq t_c)$ pro levostranný test,
- $P(T \geq t_c)$ pro pravostranný test,

kde pravděpodobnosti jsou počítány za podmínky, že nulová hypotéza je správná.

Poznámka: Obvykle nemůžeme určit přesnou P -hodnotu pomocí odpovídající tabulky kritických hodnot, můžeme ji pouze odhadnout. Ke stanovení přesné P -hodnoty použijeme počítač.

Obr. 2.3. P -hodnota

P -hodnota může být interpretována jako **pozorovaná hladina** významnosti testu hypotézy. Ilustrujeme si to na příkladu. Uvažujeme pravostranný test založený na testové statistice, která má normované normální rozdělení. Předpokládejme, že hodnota testové statistiky je 1.88. Pak p -hodnota testu hypotézy je 0.0301, jak je znázorněno na následujícím obrázku.

Obr. 2.4. P -hodnota jako pozorovaná hladina významnosti

Jak vidíme z předchozího obrázku, nulová hypotéza by měla být zamítnuta na hladině významnosti $\alpha = 0.05$, ale neměla by být zamítnuta na hladině $\alpha = 0.01$. Ve skutečnosti, jak je zřejmé z obrázku, p -hodnota je přesně nejmenší hladina významnosti, na které by nulová hypotéza měla být zamítnuta.

P -hodnota jako pozorovaná hladina významnosti

P -hodnota testu hypotézy je rovna nejmenší hladině významnosti, na které nulová hypotéza může být zamítnuta, to je nejmenší hladině významnosti, při které výběrová data vedou k zamítnutí nulové hypotézy. S ohledem na předcházející skutečnost můžeme formulovat následující kritérium pro rozhodování, zda nulová hypotéza by měla být zamítnuta ve prospěch alternativní hypotézy.

Rozhodovací kritérium pro test hypotézy pomocí P -hodnoty

Jestliže p -hodnota je menší nebo rovna zadané hladině významnosti, pak zamítněte nulovou hypotézu, jinak nezamítejte nulovou hypotézu.

Obecná metoda testu hypotézy založená na P -hodnotě je uvedena v následujícím postupu, který budeme nazývat **přístup k testování hypotézy založený na P -hodnotě**.

1. Formulujte nulovou a alternativní hypotézu.
2. Zvolte hladinu významnosti α .
3. Vypočítejte hodnotu testové statistiky.
4. Určete P -hodnotu.
5. Jestliže $P \leq \alpha$ zamítněte H_0 , jinak nezamítněte H_0 .
6. Formulujte slovně závěr.

Poznámka: U autorů, kteří v popisech testů a dalších statistických metod používají kvantilů místo kritických hodnot, by např. místo kritické hodnoty z_α normovaného normálního rozdělení $N(0,1)$ byl kvantil tohoto rozdělení $u_{1-\alpha}$, místo kritické hodnoty $t_\alpha(n)$ by byl použit příslušný kvantil $t_{1-\alpha/2}(n)$ apod. V popisu a vzorcích pro jednotlivé testy se dále vyskytují kvantily některých rozdělení. O používání kvantilů místo těchto kritických hodnot platí v plném rozsahu to, co bylo řečeno pro intervaly spolehlivosti. Proto **vždy pozor**, co příslušný symbol (např. $t_{1-\alpha/2}(n)$ nebo nějaký jiný podobný) v té které publikaci znamená, zda kvantil nebo kritickou hodnotu (a i ta může být definována v některých publikacích jinak než v této publikaci, speciálně u t -rozdělení). [5]

2.1.2 Intervaly spolehlivosti pro střední hodnotu

Nyní budeme řešit problém sestavení intervalu spolehlivosti pro střední hodnotu rozdělení při zadaném koeficientu spolehlivosti. Zde využijeme výsledky uvedené v předchozí kapitole o asymptotickém rozdělení výběrového průměru. Předpokládejme, že máme náhodný výběr z rozdělení se střední hodnotou μ a rozptylem σ^2 . Dále předpokládejme,

že rozdělení je normální nebo rozsah výběru n je velký. Pak podle $M'_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ má

náhodná veličina $U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ (přibližně) normované normální rozdělení. Tudíž pro U platí

$$P(-u_{1-\alpha/2} < U < u_{1-\alpha/2}) = 1 - \alpha. \quad (2.8)$$

Pro připomenutí, $u_{1-\alpha}$ je taková hodnota náhodné veličiny U , pro kterou platí:

$$\int_{u_{1-\alpha}}^{\infty} \phi(u) du = \alpha.$$

Lze dokázat, že pro pozorovanou hodnotu \bar{x} náhodné veličiny \bar{X} lze použít vztah

$$P\left(\mu - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \quad (2.9)$$

Pak tento vztah přepíšeme pomocí algebraických operací na tvar

$$P\left(\bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \quad (2.10)$$

Jak je vidět z této rovnice, jakmile máme k dispozici pozorované hodnoty náhodného výběru, dostáváme následující interval

$$\left(\bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right), \quad (2.11)$$

což je $100(1-\alpha)\%$ intervalem spolehlivosti pro μ .

Dále pak pro levostranný interval spolehlivosti platí

$$P\left(\bar{x} - u_{1-\alpha} \frac{\sigma}{\sqrt{n}} < \mu\right) = 1 - \alpha \quad (2.12)$$

a pro pravostranný interval spolehlivosti

$$P\left(\mu < \bar{x} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (2.13)$$

Hodnoty $u_{1-\alpha/2}$, resp. $u_{1-\alpha}$ představují příslušné kvantily normálního normovaného rozdělení pro zvolenou spolehlivost.

Postup sestavení intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 , někdy také nazývaný **jednovýběrový interval** pro μ , je následující:

- *Předpoklady*
 - a) Normální rozdělení nebo velký rozsah výběru n ,
 - b) známý rozptyl σ^2 .

1. Pro koeficient spolehlivosti $1-\alpha$, najděte hodnotu $u_{1-\alpha/2}$ v tabulce kritických hodnot $N(0,1)$ -rozdělení.

2. Krajní body intervalu spolehlivosti jsou $\bar{x} \pm u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$,

kde $u_{1-\alpha/2}$ je hodnota, určená v 1. kroku, n je rozsah výběru a \bar{x} je vypočten ze zkoumané realizace náhodného výběru. V případě výběru z normálního rozdělení je koeficient spolehlivosti přesně roven $1-\alpha$, v případě výběru o velkém rozsahu z jiného než normálního rozdělení je koeficient spolehlivosti přibližně roven $1-\alpha$. [11]

Poznámky: Jedním z předpokladů pro použití tohoto postupu je, že výběr pochází z normálního rozdělení nebo rozsah výběru je velký. Tento postup je použitelný dokonce při výběru o malém nebo přiměřeně malém rozsahu z jiného než normálního rozdělení za předpokladu, že rozdělení se neliší příliš od normálního. Postupy, které nejsou citlivé na odchylky od předpokladů, na kterých jsou založené, se nazývají **robustní**. Tudíž postup pro sestavení intervalu pro parametr μ je robustní vůči malým odchýlkám od předpokladu normality.

3 VYBRANÉ PARAMETRICKÉ TESTY

V tomto odstavci budou uvedeny jednotlivé často používané statistické testy o neznámých středních hodnotách či rozptylech jednoho nebo dvou základních souborů. Předpokládá se přitom, že **základní soubory mají normální rozdělení**. Pro každý test budou vždy popsány situace, v nichž se test používá, nulová a alternativní hypotéza, dále pak testovací statistika a jejím prostřednictvím určený kritický obor testu, pro který se nulová hypotéza zamítá, a konečně jeden či více příkladů použití popisovaného testu. Před použitím testů je nutné se podrobněji seznámit s obsahem předchozího odstavce, kde jsou popsány základní pojmy a postup při testování hypotéz. U každého testu má testovací statistika při platnosti nulové hypotézy H_0 vždy jisté (obecně pokaždé jiné) rozdělení pravděpodobnosti. S kritickou hodnotou tohoto rozdělení je pak spočtena hodnota testovací statistiky porovnávána při rozhodování o zamítnutí či nezamítnutí hypotézy H_0 . Zmíněný předpoklad o (alespoň přibližně) normálním rozdělení základních souborů je společný všem testům o středních hodnotách a rozptylech uvedených v tomto odstavci, uvedených v následujícím odstavci. V případě, kdy o rozdělení základního souboru **nelze předpokládat, že je normální**, budeme používat neparametrické testy. [11]

3.1 Test hypotézy o střední hodnotě μ při známém rozptylu σ^2

Za předpokladu, že je znám rozptyl v základním souboru, zvolíme jako testovací kritérium veličinu

$$U = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}, \quad (3.1)$$

která má za platnosti nulové hypotézy (přibližně) normální rozdělení. Tento test hypotézy $H_0 : \mu = \mu_0$ při známém rozptylu σ^2 budeme nazývat **jednovýběrový z-test** pro μ nebo stručněji **z-test** pro μ .

- *Předpoklady*

- a) Normální rozdělení nebo velký rozsah výběru ($n \geq 30$).

- b) Známý rozptyl σ^2 .

- *Testová statistika:* $U = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \approx N(0,1)$ nebo $U \approx N(0,1)$


```
>> [h,p] = ztest(data,0,1,[],'right')
```

```
h = 1
```

```
p = 2.3195e-010
```

I tento pravostranný test zamítá nulovou hypotézu na hladině významnosti $\alpha = 0,05$. V případě levostranného testu bychom už nulovou hypotézu nezamítli.

3.2 Test hypotézy o střední hodnotě μ při neznámém rozptylu σ^2

Rozptyl rozdělení, z něhož výběr pochází, obvykle neznáme. Při odvození metody pro test hypotézy o střední hodnotě μ při neznámém rozptylu σ^2 , vyjdeme z tvrzení o rozdělení

normovaného tvaru výběrového průměru $U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$. Nyní si jej připomeneme. Je-li

k dispozici náhodný výběr o rozsahu n z normálního rozdělení se střední hodnotou μ , pak

náhodná veličina $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ má t -rozdělení s $n-1$ stupni volnosti. Můžeme tudíž provést

test hypotézy s nulovou hypotézou: $H_0 : \mu = \mu_0$ za pomoci testové statistiky

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \quad (3.2)$$

a s použitím tabulky kvantilů Studentova t -rozdělení určit kvantil t -rozdělení. Následující postup pro test hypotézy o střední hodnotě budeme nazývat **jednovýběrový t -test** nebo zkráceně **t -test** pro μ .

- *Předpoklady*

a) Normální rozdělení nebo velký rozsah výběru ($n > 30$).

b) Neznámý rozptyl σ^2 .

- *Testová statistika*: $T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \approx t(n-1)$ nebo $T \approx t(n-1)$

- *Kritické kvantily* H_0 : pro dvoustranný test: $\pm t_{1-\alpha/2}$

pro levostranný test: $-t_{1-\alpha}$

pro pravostranný test: $t_{1-\alpha}$

Test hypotézy je přesný pro normální rozdělení a pouze přibližný pro výběry z jiných než normálních rozdělení. Ačkoliv t -test byl odvozen za předpokladu, že máme výběry z normálního rozdělení, používá se i pro výběry o velkém rozsahu z jiných než normálních rozdělení. Test pracuje dobře i při poměrně malých výběrech z jiných než normálních rozdělení, pokud se rozdělení neliší příliš od normálního. Jinými slovy, t -test je *robustní* vůči malým odchylkám od předpokladu normality rozdělení. Co se týče odlehlých pozorování, mohou mít dokonce při velkém rozsahu výběru značný vliv na t -test, neboť výběrový průměr a výběrový rozptyl nejsou vůči nim rezistentní. [1], [11]

Příklad:

Pro příklad použijeme stejná data jako v případě z -testu, tedy náhodný výběr o 100 prvcích s normálním rozdělením pravděpodobnosti s parametry $\mu = 0,5$ a $\sigma = 1$.

V Matlabu provedeme oboustranný test předpokladu, že výběr pochází z normálního rozdělení se střední hodnotou $\mu = 0$ následovně:

```
[h,p,ci,stats] = ttest(data,0)
h = 1
p = 5.4342e-007
ci = 0.3924  0.8537
stats = tstat: 5.3603    df: 99    sd: 1.1624
```

V tomto případě je p -hodnota $\leq \alpha$ (0,05), tudíž zamítáme nulovou hypotézu, že střední hodnota je rovna nule. V hodnotě **ci** je vypočten interval spolehlivosti pro parametr μ a vzhledem k tomu, že testovaná hypotéza nespadá do tohoto intervalu, zamítáme tedy nulovou hypotézu.

3.3 Test hypotézy o rozptylu

Kromě rozhodnutí o neznámé střední hodnotě základního souboru je třeba někdy také činit rozhodnutí o tom, zda neznámý rozptyl σ^2 je nebo není roven konkrétní číselné hodnotě, resp., zda je nebo není menší než tato hodnota, apod. V tomto odstavci uvedeme postup pro test hypotézy $H_0 : \sigma^2 = \sigma_0^2$. Alternativní hypotéza je v případě dvoustranného testu $H_A : \sigma^2 \neq \sigma_0^2$. V případě jednostranných testů $H_A : \sigma^2 > \sigma_0^2$ nebo $H_A : \sigma^2 < \sigma_0^2$. Připomeňme, že v případě testů hypotéz o střední hodnotě normálního rozdělení nepoužíváme jako testovou statistiku výběrový průměr \bar{X} , ale normovaný tvar výběrového

průměru. Podobně, nepoužijeme ani v případě testu hypotézy o rozptylu normálního rozdělení náhodnou veličinu S^2 jako testovou statistiku, ale náhodnou veličinu

$$\chi^2 = \frac{n-1}{\sigma_0^2} S^2, \quad (3.3)$$

kteřá má χ^2 -rozdělení s $n-1$ stupni volnosti. Metodou testu hypotézy o rozptylu uvedenou níže budeme nazývat χ^2 -test (chí kvadrát test) o rozptylu.

- *Předpoklad*

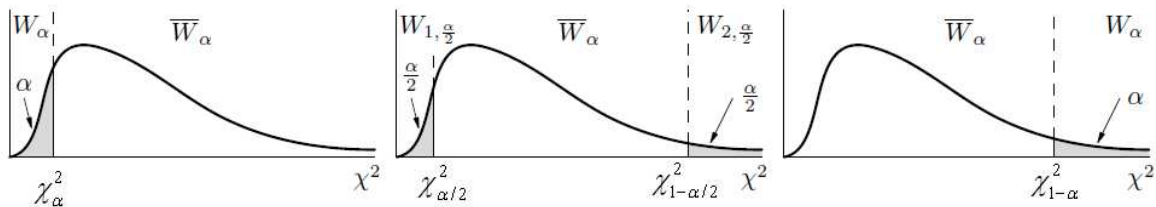
Normální rozdělení.

- *Testová statistika:* $\chi^2 = \frac{(n-1)}{\sigma_0^2} S^2 \approx \chi^2(n-1)$.

- *Obor zamítnutí H_0 :* pro levostranný test: $(0; \chi_\alpha^2)$,

pro dvoustranný test: $(0; \chi_{\alpha/2}^2) \cup (\chi_{1-\alpha/2}^2; \infty)$,

pro pravostranný test: $(\chi_{1-\alpha}^2; \infty)$.



- *P-hodnota testu H_0 :* pro levostranný test: $P(\chi^2 \leq \chi_c^2)$,

pro dvoustranný test: $2 \min\{P(\chi^2 \leq \chi_c^2), P(\chi^2 \geq \chi_c^2)\}$,

pro pravostranný test: $P(\chi^2 \geq \chi_c^2)$.

Na rozdíl od t -testu pro střední hodnotu, χ^2 -test pro rozptyl není robustní vůči odchylkám od předpokladu normality. Je dokonce tak nerobustní, že je doporučován pouze v případě výběru z normálního rozdělení nebo z rozdělení lišícího se nepatrně od normálního. Dříve než použijeme χ^2 -test je nutná předběžná analýza. [1], [11]

Příklad:

Použití testu budeme demonstrovat na vygenerovaném náhodném výběru $n = 100$ s normálním rozdělením s parametry $\mu = 0,1$ a $\sigma = 5$:

```
>> data = normrnd(0.1,5,100,1);
```

Budeme chtít testovat nulovou hypotézu $H_0: \sigma^2 = 20$ proti oboustranné alternativě:

```
>> [h,p,ci] = vartest(data,20)
```

```
h = 0
```

```
p = 0.0795
```

```
ci = 19.4676 34.0789
```

P-hodnota $> \alpha$ (0,05), tudíž na zvolené hladině významnosti tuto hypotézu nezamítáme. V proměnné ci se nachází i hodnota hypotézy σ^2 , která nezamítnutí hypotézy jen podporuje.

3.4 Test hypotézy o shodě dvou středních hodnot

Tento test patří mezi jeden z nejčastěji používaných, ať již v průmyslových aplikacích, v různých marketingových výzkumech apod. Je tomu tak proto, že umožňuje porovnávat různé situace ve výrobě, v prodeji, ve financování apod. Zcela obecně řečeno jde o případy, kdy neprovádíme úsudky pouze z jednoho náhodného výběru, ale porovnáváme mezi sebou výběry dva. Na základě porovnání těchto výběrů pak provádíme úsudky o dvou základních souborech, z nichž byly výběry provedeny. V dalším budeme předpokládat, že jde o nezávislé náhodné výběry, což je v praxi nejčastěji zajištěno tím, že v každém výběru jsou jiné jednotky. Tento test se zjednodušeně nazývá jako dvouvýběrový t -test.

Test hypotézy o rozdílu průměrů ve dvou základních souborech, z nichž byly pořízeny výběry, lze provádět za trojího předpokladu

1. Známe rozptyly v obou základních souborech.
2. Rozptyly v obou základních souborech jsou neznámé a shodné.
3. Rozptyly v obou základních souborech jsou neznámé a různé.

ad 1) Předpokládejme, že máme dva normálně rozdělené soubory se středními hodnotami μ_1 a μ_2 a rozptyly σ_1^2 a σ_2^2 . Z těchto základních souborů jsme provedli náhodné výběry o rozsahu n_1 a n_2 a stanovili výběrové průměry \bar{x}_1 a \bar{x}_2 .

Nulová hypotéza je $H_0: \mu_1 = \mu_2$. Alternativní hypotézu pak vymežíme podle povahy úlohy jako dvoustrannou nebo jednostrannou, tedy některým z těchto způsobů:

$$H_1: \mu_1 \neq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

$$H_1: \mu_1 < \mu_2$$

- *Předpoklady*

Známe rozptyly v obou základních souborech

- *Testová statistika:*
$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

- *Kritické hodnoty H_0 :* pro dvoustranný test: $\pm u_{1-\alpha/2}$

pro levostranný test: $-u_{1-\alpha}$

pro pravostranný test: $u_{1-\alpha}$

Za testové kritérium zvolíme statistiku

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad (3.4)$$

kteřá má za předpokladu nulové hypotézy normované normální rozdělení. Jako kritické hodnoty tedy zvolíme kvantily tohoto rozdělení $\pm u_{1-\alpha/2}$ u dvoustranné alternativy, resp. kvantily $-u_{1-\alpha}$ a $u_{1-\alpha}$ u jednostranných alternativ.

ad 2) Máme dva nezávislé výběry z normálního rozdělení. Neznáme-li rozptyly základního souboru, ale víme, že jsou stejné, tj. $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$ (tento předpoklad je nutné ověřit jiným testem, o kterém budeme mluvit dále), lze při testu shody dvou průměrů použít statistiku

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad (3.5)$$

kteřá má za platnosti H_0 rozdělení t s $n_1 + n_2 - 2$ stupni volnosti. s_1^2 a s_2^2 jsou výběrové rozptyly, které počítáme z jednotlivých pozorování x_{h_i} , $h = 1, 2$ (pořadí výběru), $i = 1, 2, \dots, n_h$ (pořadí pozorování v h -tém výběru) podle vzorce

$$s_h^2 = \frac{\sum_{i=1}^m (x_{h_i} - \bar{x}_h)^2}{n_h - 1}. \quad (3.6)$$

Při testování postupujeme stejně jako v bodě 1) s tím rozdílem, že jako kritické hodnoty použijeme místo kvantilů normovaného normálního rozdělení kvantily rozdělení t s $n_1 + n_2 - 2$ stupňů volnosti.

- *Předpoklady*

- Nezávislé výběry
- Rozptyly v obou základních souborech jsou neznámé a shodné
- Normální rozdělení nebo velké výběry

- *Testová statistika:* $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx t(n_1 + n_2 - 2)$

- *Kritické hodnoty H_0 :* pro dvoustranný test: $\pm t_{1-\alpha/2}$

pro levostranný test: $-t_{1-\alpha}$

pro pravostranný test: $t_{1-\alpha}$ [1], [11]

Příklad:

Vygenerujeme náhodné výběry s normálním rozdělením s parametry $\mu_1 = 0,1$ $\mu_2 = 0,2$ a stejnými směrodatnými odchylkami $\sigma_{1,2} = 1$:

```
>> x = normrnd(0.1,1,1000,1);
```

```
>> y = normrnd(0.2,1,1000,1);
```

V Malabu provedeme oboustranný test hypotézy $H_0 : \mu_1 = \mu_2$ následovně:

```
>> [h,p,ci,stats] = ttest2(x,y)
```

```
h = 1
```

```
p = 1.5177e-004
```

```
ci = -0.2571    -0.0819
```

```
stats = tstat: -3.7954 df: 1998 sd: 0.9988
```

P -hodnota $\leq \alpha$ (0,05) a testovaná hypotéza neleží uvnitř intervalu spolehlivosti, tudíž zamítáme hypotézu o shodě dvou středních hodnot na hladině významnosti $\alpha = 0,05$.

ad 3) Pokud se nacházíme v situaci, kdy nemáme ověřen některý z předpokladů v odstavcích ad 1), resp. ad 2) nebo - zcela obecně - nelze při neznámých rozptylech vůbec předpokládat jejich shodu, použijeme testového kritéria

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad (3.7)$$

keré má přibližně rozdělení t s ν stupni volnosti. Počet stupňů volnosti se přitom vypočítá podle vztahu

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1} \left(\frac{s_2^2}{n_2}\right)^2} - 2. \quad (3.8)$$

- *Předpoklady*

d) Nezávislé výběry

e) Rozptyly v obou základních souborech jsou neznámé a různé.

f) Normální rozdělení nebo velké výběry

- *Testová statistika:* $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t(\nu)$

- *Kritické hodnoty H_0 :* pro dvoustranný test: $\pm t_{1-\alpha/2}$

pro levostranný test: $-t_{1-\alpha}$

pro pravostranný test: $t_{1-\alpha}$ [1], [11]

Příklad:

Použijeme stejně vygenerovaných dat jako v předchozím příkladě, jen s různými rozptyly:

```
>> [h,p,ci,stats] = ttest2(x,y,[],[],'unequal')
```

```
h = 1
```

p = 1.5177e-004

ci = -0.2571 -0.0819

stats = tstat: -3.7954 df: 1.9980e+003 sd: [2x1 double]

Jelikož je p -hodnota $\leq \alpha$ (0,05) a testovaná hypotéza neleží v intervalu spolehlivosti, zamítáme na hladině významnosti $\alpha = 0,05$ nulovou hypotézu.

3.5 Párový t-test

Tento test použijeme v případě, že máme dva závislé náhodné výběry z dvourozměrné normální náhodné veličiny (X, Y) . To je situace například toho, kdy měříme objekt dvakrát (před pokusem a po něm) a chceme zjistit, zda měl pokus nějaký vliv na měřený objekt. Měřením tedy dostaneme dvojice (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) . Označme $\mu_1 = E(X)$ a $\mu_2 = E(Y)$. Na hladině významnosti α pak chceme testovat hypotézu $H_0: \mu_1 = \mu_2$. Postupujeme tak, že nejprve vypočteme rozdíly mezi párovými hodnotami $d_i = y_i - x_i$ pro $i = 1, 2, \dots, n$. Tyto hodnoty jsou realizací náhodné veličiny $Z = Y - X$ se střední hodnotou $\mu = \mu_2 - \mu_1$. Místo hypotézy $H_0: \mu_1 = \mu_2$ tak můžeme ekvivalentně testovat $H_0: \mu = 0$ a pro hypotézu pak lze použít z -test nebo t -test. [1]

Příklad:

Budeme testovat párová měření rychlosti zdolání okruhu tratě automobilových závodníků, kdy hodnoty z prvního měření byly pořízeny při tréninku a hodnoty druhého měření při závodě:

závodník	1	2	3	4	5	6	7	8	9	10	11	12	13
trenink	2,09	2,54	2,24	2,44	2,84	2,29	2,54	2,99	2,69	2,44	2,84	2,44	2,39
zavod	1,94	2,49	2,34	2,29	2,94	1,94	2,74	2,94	2,59	2,34	2,89	2,34	2,29

```
>> [h,p,ci,stats] = ttest(trenink,zavod)
```

h = 0

p = 0.1926

ci = -0.0311 0.1388

stats = tstat: 1.3806 df: 12 sd: 0.1406

Vzhledem k tomu, že p -hodnota $> \alpha$ (0,05), nezamítáme nulovou hypotézu na hladině významnosti $\alpha = 0,05$.

3.6 Test hypotézy o shodě dvou rozptylů

Při popisu konstrukce testu shody dvou průměrů jsme viděli, že při volbě testovacího postupu hraje důležitou roli, zda rozptyly základního souboru jsou stejné či nikoliv. Předpoklad o shodě dvou rozptylů se na základě výběrových dat může ověřit testem, který bude nyní stručně popsán. Předpokládejme tedy, že jsme provedli opět nezávislé náhodné výběry o rozsahu n_1 a n_2 . V těchto výběrech jsme vypočetli výběrové rozptyly s_1^2 a s_2^2 . Tento test, nazývaný **F-test**, mnoho statistiků tento test nedoporučuje z toho důvodu, že ačkoliv t -test je robustní vůči malým odchylkám: i když se rozdělení jen málo liší od normálního, F -test může dávat nespolehlivé výsledky. Statistik George E. P. Box řekl: „Testovat předem hypotézu o rozptylech je obdobné, jako kdybychom před tím, než zaoceánský parník vypluje z přístavu na širý oceán, spustili na moře člun, abychom si ověřili, že jsou vhodné povětrnostní podmínky pro vyplutí parníku. Nulovou hypotézu formulujeme ve tvaru $H_0: \sigma_1^2 = \sigma_2^2$. Dvoustrannou alternativní hypotézu jako $H_1: \sigma_1^2 \neq \sigma_2^2$, jednostranné alternativní hypotézy pak $H_1: \sigma_1^2 > \sigma_2^2$ a $H_1: \sigma_1^2 < \sigma_2^2$.

Za testové kritérium volíme statistiku

$$F = \frac{s_1^2}{s_2^2} \quad (3.9)$$

Za platnosti nulové hypotézy má testové kritérium rozdělení F s $\nu_1 = n_1 - 1$ a $\nu_2 = n_2 - 1$ stupni volnosti. V případě dvoustranné alternativní hypotézy je kritický obor vymezen nerovnostmi

$$F \geq F_{1-\alpha/2}[n_1 - 1; n_2 - 1], \quad (3.10)$$

$$F \leq F_{\alpha/2}[n_1 - 1; n_2 - 1] = \frac{1}{F_{1-\alpha/2}[n_2 - 1; n_1 - 1]} \quad (3.11)$$

Provádíme-li jednostranný test proti alternativní hypotéze $\sigma_1^2 > \sigma_2^2$, je kritický obor vymezen nerovností

$$F \geq F_{1-\alpha}(n_1 - 1; n_2 - 1), \quad (3.12)$$

v případě alternativní hypotézy $\sigma_1^2 < \sigma_2^2$ je kritický obor vymezen nerovností

$$F \leq F_\alpha(n_1 - 1; n_2 - 1) = \frac{1}{F_{1-\alpha}(n_2 - 1; n_1 - 1)} \quad (3.13)$$

[11]

Příklad:

Budeme uvažovat dva nezávislé výběry, pocházející z normálně rozdělené náhodné veličiny:

měření	1	2	3	4	5	6	7	8	9	10	11	12
skupina 1	5,9	4,9	6,4	7,5	7,4	7,9	4	6	7,3	5,1		
skupina 2	11,3	9,8	12,1	10,9	10,9	11,2	10,6	11,3	11,1	11,2	12	9,9

Výběrové rozptyly jsou $s_1^2 = 1,680$ a $s_2^2 = 0,482$, proto $s_1^2/s_2^2 = 3,49$. Pro $\alpha = 0,05$ a stupně volnosti $v_1 = 9$ a $v_2 = 11$ dostaneme kritickou hodnotu $F_{0,975} = 3,59$. Nulovou hypotézu $H_0: \sigma_1^2 = \sigma_2^2$ proti oboustranné alternativě tedy nezamítáme, protože $3,49 < 3,59$. Jinými slovy, testové kritérium neleží v kritickém oboru.

```
>> [h,p,ci,stats] = vartest2(data1,data2)
h = 0
p = 0.0551
ci = 0.9716 13.6378
stats = fstat: 3.4861    df1: 9    df2: 11
```

Vzhledem k tomu, že p-hodnota $> \alpha$ (0,05) a poměr dvou rozptylů jako testové kritérium s_1^2/s_2^2 leží v konfidenčním intervalu, tudíž nezamítáme nulovou hypotézu na hladině významnosti $\alpha = 0,05$.

4 VYBRANÉ NEPARAMETRICKÉ TESTY

Tato kapitola se bude zabývat skupinou speciálních statistických testů, které se nazývají neparametrické. Jejich použití je však vázáno na splnění určitých podmínek. Pokud tyto nelze splnit, nelze použít příslušný parametrický test. Proto byly statistiky vyvinuty tyto speciální testy. V této kapitole budou představeny nejdůležitější neparametrické testy. V předchozí části této práce můžeme nalézt některé testy hypotéz o hodnotách parametrů pravděpodobnostních modelů (např. testy hypotéz o parametrech μ a σ^2 normálního rozdělení). Tyto testy se nazývají **parametrické testy**. Jsou to, přísně vzato, testy o parametrech známých pravděpodobnostních modelů a jsou tedy použitelné v situacích, kdy se zkoumají veličiny, jejichž rozdělení lze alespoň přibližně popsat některým z pravděpodobnostních modelů (normálním, lognormálním, exponenciálním rozdělením aj.). V případech velkých výběrů lze při testech hypotéz o některých parametrech (například o parametrech, jež jsou středními hodnotami) postupovat stejně, ať výběr pochází z jakéhokoliv obvykle se vyskytujícího rozdělení, takže v tomto případě není znalost pravděpodobnostního modelu nutná. Při výzkumu trhu se poměrně často pracuje s menšími výběry. Přitom se žádají informace o určitých vlastnostech (např. poloze, variabilitě aj.) rozdělení neznámého typu. V některých z těchto situací nelze použít parametrické testy, lze ale využít tzv. neparametrické testy, jimž je věnována tato kapitola.

Neparametrické testy mají oproti parametrickým řadu **výhod**, mezi něž patří:

- Pravděpodobnostní závěry, které z nich získáme, jsou většinou nezávislé na tvaru rozdělení náhodných veličin v základním souboru, často se předpokládá pouze spojitost rozdělení základního souboru (distribuční funkce).
- Lze je použít i v případě, když neznáme tvar rozdělení základního souboru a rozsah výběru je malý.
- Lze je použít i tehdy, když výběry pocházejí ze základních souborů s různými rozděleními sledovaných náhodných veličin.
- Lze je použít i pro data, která mají charakter ordinálních (pořadových) nebo nominálních (slovních) proměnných nebo klasifikační charakter.
- Většinou jsou výpočetně poměrně jednoduché.

Z **nevýhod** neparametrických testů lze uvést následující výčet:

- informace z dat jsou méně efektivní – síla testu je tedy nižší (je zde větší pravděpodobnost chyby II. druhu β). Pokud jsou splněny podmínky adekvátního parametrického testu, měl by mít prioritu,
- při větším rozsahu výběru rostou nároky na výpočty a rovněž i na tabulky kritických hodnot (ty bývají poměrně těžko dostupné – ne každá statistická učebnice je uvádí). [6]

4.1 χ^2 test v kontingenční tabulce

Použití:

Při vyšetřování možné závislosti dvou nominálních proměnných. Výsledky pozorování zapisuje pro přehlednost do tzv. kontingenční tabulky. Kontingenční tabulka vznikne, třídíme-li soubor podle variant 2 kvalitativních znaků A a B, kdy A má r variant a B má s variant.

Tab. 4.1. Schéma kontingenční tabulky

A \ B	B ₁	B ₂	...	B _s	\sum_j
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.}
\sum_i	n _{.1}	n _{.2}	...	n _{.s}	n

Nulová hypotéza zní: proměnné A a B jsou **nezávislé**.

Testové kritérium má tvar:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}, \quad (4.1)$$

kde $n_{ij}^* = \frac{n_{i.} \cdot n_{.j}}{n}$ jsou teoretické četnosti. Kritický obor je vymezen nerovností:

$\chi^2 \geq \chi^2_{1-\alpha((r-1)(s-1))}$. Padne-li hodnota testového kritéria do kritického oboru, znamená to, že mezi proměnnými A a B byla prokázána závislost.

Míra síly závislosti nominálních proměnných se měří například pomocí Pearsonova kontingenčního koeficientu

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}} \in (0;1). \quad (4.2)$$

Jsou-li obě proměnné statisticky nezávislé, je tento koeficient rovný 0. Jeho horní mez je však závislá na velikosti čísla $h = \min((r - 1); (s - 1))$. S rostoucím h se hodnota Pearsonova koeficientu blíží 1. Okolnost, že horní mez tohoto koeficientu není ani při pevné závislosti rovná 1, ztěžuje interpretaci jeho hodnot. [6]

Příklad:

Při průzkumu prodeje sportovní obuvi byli respondenti dotázáni, zda při jejím nákupu preferují především její kvalitu, módnost či nízkou cenu. Dále byli dotázáni, kde si tuto obuv zakoupili: a) v prodejně obuvi, b) na tržišti, c) v obchodním domě. Průzkum měl mimo jiné ověřit, zda forma prodeje (vysvětlovaná proměnná) závisí na preferencích, které jsou pro zákazníky při nákupu rozhodující (vysvětlující proměnná). Jde tedy o posouzení závislosti dvou nominálních proměnných. Údaje zjištěné u 120 respondentů jsou uvedeny v tabulce.

Tab. 4.2. Zadání příkladu pro výpočet χ^2 testu o nezávislosti v kombinační tabulce

Hlavní kritérium při nákupu	Forma prodeje			Součty $n_{i.}$
	Prodejna obuvi	Tržiště	Obchodní dům	
Cena	10	20	15	45
Módnost	5	15	20	40
Kvalita	10	5	20	35
Součty $n_{.j}$	25	40	55	120

Příklad může být řešen v programovém prostředí Matlab následovně:

```
T = [10 20 15;
      5 15 20;
      10 5 20];
```

% kde radky odpovídají hl.kriteriu při nákupu a sloupce forme prodeje

```

% Testujte hypotézu H0: forma prodeje nezavisi na preferencich, ktere
jsou pro zakazniky pri nakupu rozhodujici.
% Reseni
[nr nc]=size(T); % rozmery tabulky
ss=sum(sum(T)); % celkovy soucet
t=T/ss; % normovan?
sc=sum(t); % norm. soucet pres radky (svisle)
sr=sum(t'); % norm. soucet pres sloupce (vodorovne)
tt=sr'*sc; % nezavisla normovana tabulka
TT=tt*ss; % nezavisla tabulka absolutni
o=T(:); % pozorovane cetnosti
e=TT(:); % teoreticke cetnosti
n=(nr-1)*(nc-1); % stupne volnosti
[pval ch2]=chisquare_test(o,e,n);
pearson = sqrt(ch2/(ch2+ss));
kritickaHodnota = chi2inv(0.95,n); % vypocet kriticke hodnoty
fprintf('Test nezavislosti\n\n');
fprintf('testove kriterium: %g\n', ch2);
fprintf('p-hodnota: %g\n\n', pval);
fprintf('Pearsonuv kontingencni koeficient: %g\n\n', pearson);
fprintf('Kriticka hodnota chi-kvadrat rozdeleni: %g\n\n',
kritickaHodnota);

```

Zvolíme obvyklou hladinu významnosti $\alpha = 0,05$, kritickou hodnotou bude kvantil χ^2 rozdělení o $(3 - 1) \cdot (3 - 1) = 4$ stupních volnosti. Tento kvantil je rovný 9,49. Kritický obor bude tedy vymezen nerovností $\chi^2 \geq 9,49$. V našem případě byla vypočtena hodnota testového kritéria $\chi^2 = 10,728$. Tato hodnota je v kritickém oboru, takže test zamítá na 5% hladině významnosti hypotézu o nezávislosti. S 5% rizikem omylu můžeme tedy říci, že volby formy prodeje závisí na preferencích, jež jsou pro zákazníky při nákupu rozhodující.

K posouzení, zda jde o silnou nebo slabou závislost, slouží Pearsonův kontingenční koeficient:

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}}; P \in (0;1), \text{ tedy } P = \sqrt{\frac{10,728}{10,728 + 120}} = 0,2865.$$

Tento koeficient je blízky nule, což naznačuje jen velmi slabou statistickou závislost volby formy prodeje na preferenci při nákupu v daném souboru 120 respondentů. Tato závislost je však statisticky významná.

4.2 Jednovýběrové neparametrické testy

4.2.1 Znaménkový test

Uvažujme náhodný výběr o rozsahu n ze spojitého rozdělení s mediánem \tilde{X} . Pravděpodobnost, že vybraná hodnota je menší než \tilde{X} je stejná jako pravděpodobnost, že hodnota je větší. $H_0: \tilde{X} = a$ (a je dané číslo); $H_1: \tilde{X} \neq a$. Testové kritérium označíme S^+ je počet rozdílů $(x_i - a)$ s kladným znaménkem. Platí-li H_0 , má S^+ $Bi(n; 1/2)$. Pro malá n jsou tabelována čísla k_1 a k_2 ve statistických tabulkách tak, že

$$P(S^+ \leq k_1) \leq \alpha/2 \text{ a } P(S^+ \geq k_2) \leq \alpha/2. \quad (4.3)$$

Jestliže $S^+ \leq k_1$ nebo $S^+ \geq k_2$, zamítáme nulovou hypotézu.

Pro velká n má testové kritérium tvar

$$U = \frac{2S^+ - n \pm 1}{\sqrt{n}} \quad (4.4)$$

s normovaným normálním rozdělením $N(0,1)$. Kritický obor je vymezen nerovností $|U| \geq u_{1-\alpha/2}$. Člen ± 1 v čitateli testového kritéria je tzv. oprava na spojitost, kterou používají některé statistické softwary. Je-li $S^+ < \frac{n}{2}$, je roven $+1$, je-li $S^+ > \frac{n}{2}$, je roven -1 .

Znaménkový test se používá v případě, že rozdělení náhodné veličiny X je značně asymetrické. Vzhledem k relativně malé síle testu se doporučuje volit větší n . Případy pro $\tilde{X} = a$ vypustíme a snížíme o ně počet pozorování n . Tento test se dá použít s malou úpravou i pro tzv. párová pozorování (tj. dva závislé výběry). Budeme se jím zabývat dále. [6]

Příklad:

Vyráběné ocelové tyče mají kolísavou délku s předpokládanou hodnotou mediánu 10 m. Náhodný výběr 10 tyčí poskytl následující výsledky:

9,83; 10,10; 9,72; 9,91; 10,04; 9,95; 9,82; 9,73; 9,81; 9,90.

Je předpoklad o hodnotě mediánu tyčí oprávněný?

$H_0: \tilde{X} = 10$ a $H_1: \tilde{X} \neq 10$. Pro použití znaménkového testu nejprve stanovíme odchylky od předpokládané hodnoty mediánu a dostáváme

-0,17; 0,10; -0,28; -0,09; 0,04; -0,05; -0,18; -0,27; -0,19; -0,10.

Hodnota testového kritéria je $S^+ = 2$. Kritické hodnoty z tabulek pro znaménkový test jsou $k_1 = 1$ a $k_2 = 9$. Zjištěná hodnota S^+ leží mezi nimi, tj. v oboru přijetí. Znamená to, že tímto testem nebyla prokázána neplatnost hypotézy, že medián vyráběných ocelových tyčí je 10 m. Použijeme-li aproximaci, dostaneme hodnotu testového kritéria:

$$|U| = \frac{2 \cdot 2 - 10 + 1}{\sqrt{10}} = 1,58 < 1,96. \text{ Není v kritickém oboru, nulovou hypotézu nezamítáme. [6]}$$

V Matlabu tento test provedeme jednoduše:

```
>> x = [9.83; 10.10; 9.72; 9.91; 10.04; 9.95; 9.82; 9.73; 9.81; 9.90]';
```

```
>> [p,h,stats] = signtest(x,10)
```

```
p = 0.1094
```

```
h = 0
```

```
stats = zval: NaN   sign: 2
```

Na hladině významnosti $\alpha = 0,05$ nezamítáme nulovou hypotézu, protože p -hodnota $> \alpha$ (0,05).

4.2.2 Wilcoxonův test

Používá se pro testování stejné hypotézy jako znaménkový test. Použijeme ho v případě symetričnosti rozdělení náhodné veličiny X .

Postup při testování:

1. Pro každé měření vypočítáme $|x_i - a|$. Je-li $|x_i - a| = 0$, měření vypustíme a snížíme n .
2. Uspořádáme tyto hodnoty vzestupně a přiřadíme jim pořadí.
3. Vypočítáme součet pořadí pro kladné rozdíly S^+ a součet pro záporné rozdíly S^- .
4. Testovým kritériem je pro oboustrannou H_1 veličina $S = \min(S^+, S^-)$.
5. Je-li S menší nebo rovno kritické hodnotě ze speciálních tabulek pro Wilcoxonův test (hodnoty w_n), zamítáme nulovou hypotézu H_0 .

Lze též pro větší n použít aproximace. Za platnosti nulové hypotézy má veličina

$$U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \quad (4.5)$$

asymptoticky rozdělení $N(0, 1)$. V případě $|U| \geq u_{1-\alpha/2}$ zamítáme hypotézu na hladině, která je asymptoticky rovna α .

Protože informace obsažená v pořadích je větší než informace obsažená jen ve znaménkách rozdílů má Wilcoxonův test větší sílu než test znaménkový. Wilcoxonův test lze použít podobně jako test znaménkový pro tzv. párová pozorování. Rovněž se jím budeme zabývat i v dalším výkladu. [6]

Příklad:

Při výzkumu trhu je třeba na hladině významnosti 0,05 rozhodnout, zda medián měsíčních příjmů rodiny v jisté oblasti je menší než 25 tis. Kč. Dotázáno bylo 20 náhodně vybraných rodin a byly získány následující měsíční příjmy v tis. Kč.

18,9; 20,3; 21,2; 22,5; 26,2; 19,3; 22,2; 25,3; 25,1; 23,9;

20,1; 17,5; 27,2; 24,9; 25,0; 28,0; 19,9; 33,0; 24,3; 28,0.

Předpokládá se, že rozdělení příjmů je symetrické kolem mediánu a cílem je posoudit hypotézy

$H_0: \tilde{X} = 25$ a $H_1: \tilde{X} \neq 25$.

Nulovou hypotézu zamítneme, bude-li hodnota testového kritéria nejvýše rovna kritické hodnotě z tabulek. Nejprve ovšem vypočítáme **pořadová čísla odchylek** jednotlivých pozorování od 25 tis. Kč (se znaménkem původní odchylky):

-17; -13; -12; -8; +6; -16; -9; +3; +1,5; -5;

-14; -18; +7; -1,5; +10,5; -15; +19; -4; +10,5,

přičemž rodina s příjmem 25 byla z analýzy vypuštěna (tedy $n = 19$). Součet $S^+ = 57,5$ a $S^- = 132,5$. Hodnota testového kritéria je $\min(57,5; 132,5) = 57,5$. Tabulková hodnota je rovna číslu 46. Jelikož hodnota testového kritéria je větší než kritická hodnota, znamená to, že nulovou hypotézu nezamítáme. Test neprokázal, že by se medián mezd významně lišil od 25 tis. Kč.

```
>> x = [18.9; 20.3; 21.2; 22.5; 26.2; 19.3; 22.2; 25.3; 25.1; 23.9;...
```

```
20.1; 17.5; 27.2; 24.9; 25.0; 28.0; 19.9; 33.0; 24.3; 28.0];
```

```
>> [p,h,stats] = signrank(x,25)
```

```
p = 0.1312
```

```
h = 0
```

```
stats = zval: -1.5094 signedrank: 57.5000
```

Na hladině významnosti $\alpha = 0,05$ nezamítáme nulovou hypotézu, protože p -hodnota $> \alpha$ (0,05).

4.3 Neparametrické testy pro dva závislé a pro dva nezávislé výběry

4.3.1 Znaménkový test pro dva závislé výběry

Znaménkový test popsany v předchozí části může být použit i pro párově získaná závislá pozorování k testování shody populačních mediánů dvou rozdělení. Použijeme ho tak, že obyčejný znaménkový test pro jeden výběr s $a = 0$ aplikujeme na rozdíly měření. Přesnější výsledky dostáváme pomocí následujícího Wilcoxonova testu. Příklad bude uveden po výkladu tohoto testu. [6]

4.3.2 Wilcoxonův test pro dva závislé výběry

Nechť X_1, X_2, \dots, X_n je náhodný výběr ze spojitého rozdělení s distribuční funkcí $F(x)$. Chceme testovat hypotézu, že F je symetrická kolem nuly v tom smyslu, že

$$F(x) = 1 - F(-x), \quad -\infty < x < \infty. \quad (4.6)$$

Seřadme X_1, X_2, \dots, X_n do rostoucí posloupnosti podle absolutní hodnoty, tj.

$$|X|^{(1)} < |X|^{(2)} < \dots < |X|^{(n)}. \quad (4.7)$$

Nechť R_i^+ je pořadí X_i při tomto uspořádání. Zavedme veličiny

$$S^+ = \sum_{X_i \geq 0} R_i^+, \quad S^- = \sum_{X_i < 0} R_i^+. \quad (4.8)$$

Přitom platí $S^+ + S^- = n(n+1)/2$, což lze užít pro kontrolu správnosti výpočtů. Je-li číslo $\min(S^+, S^-)$ menší nebo rovno tabelované kritické hodnotě, hypotéza se zamítá. Kritické hodnoty jsou v tabulkách.

Pro větší hodnoty n lze využít toho, že za platnosti hypotézy má veličina

$$U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \quad (4.9)$$

asymptoticky rozdělení $N(0, 1)$. V případě $|U| \geq u_{1-\alpha/2}$ zamítáme hypotézu na hladině, která je asymptoticky rovna α . [6]

Tyto testy jsou neparametrickou obdobou t -testu pro párová pozorování.

Příklad:

Studenti během semestru hodnotí přednášející pomocí bodů. K dispozici jsou výsledky 10 náhodně vybraných přednášejících. Úkolem je posoudit, zda přednášející získali od studentů více bodů letos než v minulém roce.

Tab. 4.3. Zadání pro výpočet Wilcoxonova testu pro dva závislé výběry

Přednášející	1	2	3	4	5	6	7	8	9	10
Min. rok	932	906	943	907	893	870	889	902	866	787
Letos	933	923	916	909	908	893	890	889	888	838
Rozdíly	1	17	-27	2	15	23	1	-13	22	51

H_0 : Medián bodů v letošním a minulém roce je shodný.

Pro náš příklad je $n = 10$, $S^+ = 8$, z tabulek V zjistíme $k_1 = 1$ a $k_2 = 9$. S^+ leží mezi nimi v oboru přijetí, tedy nejsme oprávněni zamítnout H_0 . Odlišnost dosažených bodů nebyla prokázána.

Použijeme-li aproximaci, dostáváme hodnotu testového kritéria:

$$|U| = \frac{2.8 - 10 - 1}{\sqrt{10}} = 1,58 < 1,96. \text{ Není v kritickém oboru, nulovou hypotézu nezamítáme.}$$

Stejně údaje testujme dále, nyní pomocí jednovýběrového Wilcoxonova testu.

Tab. 4.4. Hodnoty pro výpočet příkladu pomocí jednovýběrového Wilcoxonova testu

Uspořádané hodnoty $ X_i $	1	1	2	<u>13</u>	15	17	22	23	<u>27</u>	51
Pořadí	1,5	1,5	3	4	5	6	7	8	9	10

$n = 10$,

$$S^+ = 1,5 + 1,5 + 3 + 5 + 6 + 7 + 8 + 10 = 42,$$

$$S^- = 9 + 4 = 13.$$

$\min(S^+; S^-) = \min(13; 42) = 13$. Kritická hodnota je podle tabulek rovna 8. Jelikož $\min(S^+; S^-) = \min(13,42) = 13 > 8$, nejsme oprávněni zamítnout nulovou hypotézu.

Použijeme-li aproximaci, dostáváme dle (4.9):

$$|U| = \frac{42 - \frac{1}{4}10 \cdot 11}{\sqrt{\frac{1}{24}10 \cdot 11 \cdot 21}} = 1,48 < 1,96. \text{ Není v kritickém oboru, nulovou hypotézu nezamítáme.}$$

V Matlabu tento příklad vyřešíme následovně:

```
>> x = [932 906 943 907 893 870 889 902 866 787];
```

```
>> y = [933 923 916 909 908 893 890 889 888 838];
```

```
>> [p,h,stats] = signrank(x,y)
```

```
p = 0.1523
```

```
h = 0
```

```
stats = signedrank: 13
```

Na hladině významnosti $\alpha = 0,05$ nezamítáme nulovou hypotézu, protože p -hodnota $> \alpha$ (0,05).

4.3.3 Wilcoxonův dvouvýběrový test pro nezávislé výběry (Mannův-Whitneyův test)

Nechť X_1, X_2, \dots, X_m a Y_1, Y_2, \dots, Y_n jsou dva nezávislé výběry ze dvou spojitých rozdělení. Chceme testovat hypotézu, že distribuční funkce obou rozdělení jsou totožné.

Všech $n + m$ výběrových hodnot X_1, X_2, \dots, X_m a Y_1, Y_2, \dots, Y_n uspořádáme vzestupně podle velikosti. Zjistíme součet pořadí hodnot X_1, X_2, \dots, X_m a označíme ho T_1 . Obdobně T_2 je součet pořadí hodnot Y_1, Y_2, \dots, Y_n . Vypočítáme

$$U_1 = mn + \frac{m(m+1)}{2} - T_1, \quad U_2 = mn + \frac{n(n+1)}{2} - T_2. \quad (4.10)$$

Přitom platí $U_1 + U_2 = mn$. Pokud $\min(U_1, U_2)$ je menší nebo rovno kritické hodnotě uvedené v tabulkách, zamítáme hypotézu. Jsou-li hodnoty m a n velké, vypočteme veličinu:

$$U_0 = \frac{U_1 - \frac{1}{2}mn}{\sqrt{\frac{mn}{12}(m+n+1)}}, \quad (4.11)$$

kteřá má za platnosti hypotézy asymptoticky rozdělení $N(0; 1)$. V případě, že $|U_0| \geq u_{1-\alpha/2}$, zamítneme hypotézu na hladině asymptoticky rovné α . Tento Wilcoxonův test se používá nejčastěji místo dvouvýběrového t-testu. [6]

Příklad:

Bylo vybráno 10 polí stejné kvality. Na 4 z nich se zkoušel nový způsob hnojení, zbývajících 6 bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Výsledky jsou uvedeny v tabulce (v metrických centech na hektar). Je třeba zjistit, zda nový způsob hnojení má jiný vliv na průměrné hektarové výnosy než starý způsob.

Tab. 4.5. Zadání příkladu pro použití Mannova-Whitneyova testu

Hektarové výnosy při novém způsobu hnojení X_i	51	52	49	55	–	–
Hektarové výnosy při starém způsobu hnojení Y_i	45	54	48	44	53	50

Všechny hodnoty X_i a Y_i v tabulce 4.5 uspořádáme podle velikosti. Tím dostaneme první řádek tabulky, podtržená čísla patří do prvního výběru. Na druhém řádku jsou pořadí hodnot X a na dalším jsou pořadí hodnot Y .

Tab. 4.6. Upravené zadání pro použití Mannova-Whitneyova testu

Sdružený výběr	44	45	48	<u>49</u>	50	<u>51</u>	<u>52</u>	53	54	<u>55</u>
Pořadí hodnot X				4		6	7			10
Pořadí hodnot Y	1	2	3		5			8	9	

Odtud $T_1 = 4 + 6 + 7 + 10 = 27$, $T_2 = 1 + 2 + 3 + 5 + 8 + 9 = 28$

a $U_1 = 4 \cdot 6 + (4 \cdot 5) / 2 - 27 = 7$, $U_2 = 4 \cdot 6 + (6 \cdot 7) / 2 - 28 = 17$.

Kritická hodnota při $\alpha = 0,05$ pro $m = 4$, $n = 6$ je podle tabulek rovna 2. Protože $\min(7; 17) = 7 > 2$, nemůžeme zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy stejný vliv jako starý způsob.

V Matlabu se předchozí příklad vyřeší následujícím způsobem:

```
>> x = [51 52 49 55];
```

```
>> y = [45 54 48 44 53 50];
```

```
>> [p,h] = ranksum(x,y)
```

```
p = 0.3524
```

```
h = 0
```

Na hladině významnosti $\alpha = 0,05$ nezamítáme nulovou hypotézu, protože p -hodnota $> \alpha$ (0,05).

4.4 Testy o typu rozdělení

4.4.1 χ^2 test dobré shody

Použití tohoto testu je typické v následujících situacích ve dvou základních situacích:

a) H_0 předpokládá, že v konečném základním souboru roztříděném podle kvantitativního nebo kvalitativního znaku do k skupin jsou podíly jednotek rovny číslům $\pi_{0,1}, \pi_{0,2}, \dots, \pi_{0,k}$.

Za testové kritérium volíme statistiku

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\pi_{0,i})^2}{n\pi_{0,i}}, \quad (4.12)$$

kde n_i jsou pozorované výběrové četnosti a $n\pi_{0,i}$ jsou teoretické četnosti v i -té skupině. χ^2 (4.12) má za předpokladu dostatečně velkého výběru χ^2 rozdělení s $v = k - 1$ stupni volnosti. Kritický obor je vymezen nerovností $\chi^2 \geq \chi^2_{1-\alpha}(k-1)$.

b) H_0 předpokládá, že nekonečný základní soubor má rozdělení určitého typu. V případě, že H_0 udává nejen typ rozdělení, ale i jeho parametry, mluvíme o úplně specifikovaném modelu, jinak mluvíme o neúplně specifikovaném modelu. Postupujeme podobně jako v předchozím případě. Kritický obor je vymezen nerovností $\chi^2 \geq \chi^2_{1-\alpha}(k-s-1)$, kde s je počet parametrů příslušného rozdělení. [6]

Příklad:

V příkladu otestujeme pravidelnost hrací kostky. Jako data pro testování použijeme výsledky 4096 hodů 12 hracími kostkami. Při každém hodu se zjišťovalo, kolik šestek na kostkách padlo. Za předpokladu pravidelnosti hracích kostek je pravděpodobnost padnutí šestky na kostce rovna $1/6$ a protože hody různými kostkami jsou nezávislé, je počet šestek při hodu 12 kostkami náhodná veličina s rozdělením $Bi(12, 1/6)$. Hodnoty jsou shrnuty v následující tabulce:

Tab. 4.7. Zadání příkladu pro výpočet χ^2 testu dobré shody

Počet šestek	0	1	2	3	4	5	6	7 a víc	Celkem
X_i	447	1145	1181	796	380	115	24	8	4096
p_i	0,112	0,269	0,296	0,197	0,089	0,028	0,007	0,001	1,000
np_i	459	1103	1213	809	364	116	27	5	4095

Výpis v Matlabu vypadá následovně:

```
>> bins = 0:7;
```

```
>> [h,p,stats] = chi2gof(bins,'ctrs',bins,...
```

```
'frequency',namerene,...
```

```
'expected',ocekavane)
```

h = 0

p = 0.5619

stats = chi2stat: 5.8113 df: 7

edges: [-0.5000 0.5000 1.5000 2.5000 3.5000 4.5000 5.5000 6.5000 7.5000]

O: [447 1145 1181 796 380 115 24 8]

E: [459 1103 1213 809 364 116 27 5]

Vzhledem k tomu, že p -hodnota $> \alpha$ (0,05) nezamítáme hypotézu o pravidelnosti hrací kostky. [1]

4.4.2 Kolmogorovův-Smirnovův test

Tento test je založen na porovnávání distribuční funkce předpokládaného rozdělení s výběrovou (empirickou) distribuční funkcí. Jedná se o zcela obecný test pro jakýkoli typ rozdělení. Hypotézy formulujeme stejně jako v případě χ^2 -testu dobré shody. Na rozdíl od χ^2 -testu dobré shody lze test provádět i pro náhodné výběry poměrně malých rozsahů.

Testovací statistika D je definována jako největší vzdálenost mezi hodnotami výběrové distribuční funkce $F_n(x)$ a teoretické distribuční funkce $F_0(x)$:

$$D = \max |F_n(x) - F_0(x)|. \quad (4.13)$$

Pro malé rozsahy výběru n použijeme přesnou kritickou hodnotu z tabulky kritických hodnot Kolmogorova-Smirnova testu. Pro velká n můžeme pro výpočet kritické hodnoty použít přibližnou, asymptoticky platnou hodnotu $\sqrt{-\frac{1}{2} \ln(\alpha/2)}$, tedy např. pro $\alpha = 0,05$ hodnotu $1,358/\sqrt{n}$ a pro $\alpha = 0,01$ hodnotu $1,628/\sqrt{n}$. Pokud hodnota testovací statistiky překročí tuto hodnotu, zamítáme nulovou hypotézu, tedy že základní soubor nemá předpokládané rozdělení. [2], [3]

Příklad:

Pro simulaci testu v Matlabu si vygenerujeme náhodný výběr s normálním rozdělením, se střední hodnotou $\mu = 0$ a směrodatnou odchylkou $\sigma = 1$ a následně otestujeme, zda tento výběr skutečně pochází z tohoto rozdělení:

```
>> x=normrnd(0,1,1,1000);
```

```
>> [h,p,k,c] = kstest(x,[],0.05,0)
```

```
h = 0
```

```
p = 0.7146
```

```
k = 0.0219
```

```
c = 0.0428
```

Vzhledem k tomu, že p -hodnota vyšla větší než zvolená hladina významnosti $\alpha = 0,05$, nezamítáme hypotézu, že rozdělení pochází z normálního.

Příklad:

V dalším příkladu si vygenerujeme náhodný výběr pocházející z exponenciálního rozdělení s parametrem $\lambda = 1$ a následně otestujeme, zda tento výběr skutečně pochází z tohoto rozdělení:

```
>> x = exprnd(1,100,1);
```

```
>> [h,p] = kstest(x,[x expcdf(x,1)])
```

```
h = 0
```

```
p = 0.3800
```

Vzhledem k tomu, že p -hodnota vyšla větší než zvolená hladina významnosti $\alpha = 0,05$, nezamítáme hypotézu, že rozdělení pochází z exponenciálního.

4.4.3 Lillieforsův test

Tento test je určen pro testování normality dat a funguje na stejném principu jako předchozí Kolmogorovův-Smirnovův test, tzn. na porovnání teoretické distribuční funkce s výběrovou distribuční funkcí. Na rozdíl ale od Kolmogorovova-Smirnovova testu, kdy musela být předpokládána (teoretická) distribuční funkce plně definována, můžeme zde parametry μ a σ odhadnout pomocí výběrových odhadů \bar{x} a s . Stejně jako v případě Kolmogorovova-Smirnovova testu pro malá n použijeme přesnou kritickou hodnotu z tabulky kritických hodnot Lillieforsova testu, pro velká n lze použít přibližnou, asymptoticky platnou hodnotu $0,89/\sqrt{n}$ pro $\alpha = 0,05$, resp. hodnotu $1,04/\sqrt{n}$ pro $\alpha = 0,01$. [1]

4.5 Neparametrické míry těsnosti závislosti

Spearmanův koeficient pořadové korelace (obdoba jednoduchého koeficientu korelace) je dán vztahem

$$r_s = 1 - \frac{6 \sum (i_x - i_y)^2}{n(n^2 - 1)}, \quad (4.14)$$

kde i_x a i_y jsou pořadová čísla hodnot proměnných x a y , n – rozsah výběru.

Spearmanův koeficient pořadové korelace nabývá hodnot z intervalu $\langle -1; 1 \rangle$, přičemž hodnoty kolem 0 ukazují na nezávislost, hodnoty blízké 1 či -1 na existenci přímé či nepřímé závislosti.

Test hypotézy významnosti r_s

Testujeme nulovou hypotézu $H_0: \rho_s = 0$ proti alternativě $H_1: \rho_s \neq 0$. Pro výběry o rozsahu $n < 10$ je třeba kritickou hodnotu hledat ve speciálních tabulkách, pro $n \geq 10$ lze použít známého testového kritéria:

$$t = \frac{r_s}{\sqrt{1 - r_s^2}} \sqrt{(n - 2)}. \quad (4.15)$$

Za platnosti hypotézy H_0 má veličina t Studentovo t -rozdělení s $(n - 2)$ stupni volnosti.

Kritický obor je vymezen nerovností: $|t| \geq t_{1-\alpha/2}(n - 2)$.

Všimněme si ještě výpočtu r_s v případě, že se některé z hodnot x_i (resp. y_i) opakují a jsou jim přiřazeny průměry z pořadových čísel, která na ně připadají. V tomto případě r_s spočítáme podle vzorce

$$r_s = 1 - \frac{6 \sum (i_x - i_y)^2}{n(n^2 - 1) - C}, \quad (4.16)$$

kde pro opravný člen C platí

$$C = \frac{1}{2} \left[\sum_k (h_{x,k}^3 - h_{x,k}) + \sum_{k'} (h_{y,k'}^3 - h_{y,k'}) \right]. \quad (4.17)$$

V tomto vzorci značí $h_{x,k}$ četnost k -té skupiny stejných hodnot proměnné x a $h_{y,k'}$ četnost k' -té skupiny stejných hodnot proměnné y .

Spearmanův koeficient r_s se často používá jako charakteristika shody pořadí n jednotek podle dvou hledisek. Čím více se pořadí jednotek podle těchto hledisek shodují, tím je r_s bližší jedničce. [6]

Příklad:

V tabulce je uvedeno pořadí obratu zahraničního obchodu (y) a počtu obyvatel (x) 11 vybraných států. Údaje jsou uvedeny v následující tabulce.

Tab. 4.8. Zadání příkladu pro výpočet Spearmanova koeficientu pořadové korelace

Stát	i_x (obyvatelé)	i_y (obrat)	$(i_x - i_y)^2$
UK	7	8	1
USA	10	11	1
Francie	6	9	9
Nizozemí	4	6	4
Itálie	8	7	1
Japonsko	9	10	1
Norsko	1	1	0
CCCP	11	4	49
Španělsko	5	3	4
Švédsko	2	2	0
Belgie	3	5	4
Součet	66	66	74

$$r_s = 1 - \frac{6 \sum (i_x - i_y)^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 74}{11(11^2 - 1)} = 0,6636.$$

Test významnosti r_s :

$$t = \frac{0,6636 \cdot \sqrt{11-2}}{\sqrt{1-0,6636^2}} = 2,66 \geq t_{0,975}(11-2) = 2,26.$$

Hodnota t-kritéria převýšila kritickou hodnotu, znamená to, že r_s je statisticky významný. Existuje statisticky významná závislost mezi pořadím podle počtu obyvatel a pořadím podle velikosti obratu zahraničního obchodu.

V Matlabu můžeme vypočítat Pearsonův, Spearmanův a Kendallův koeficient korelace následujícím způsobem:

```
>> [RHO,PVAL] = corr(x,y,'type','Pearson','rows','all','tail','both')
```

```
RHO = 0.6636
```

```
PVAL = 0.0260
```

```
>> [RHO,PVAL] = corr(x,y,'type','Spearman','rows','all','tail','both')
```

```
RHO = 0.6636
```

```
PVAL = 0.0309
```

```
>> [RHO,PVAL] = corr(x,y,'type','Kendall','rows','all','tail','both')
```

```
RHO = 0.5636
```

```
PVAL = 0.0165
```

Ve všech případech použití testů neparametrické míry těsnosti závislosti vyšla p -hodnota $< \alpha$ (0,05). Zamítáme nulovou hypotézu o nevýznamnosti Spearmanova, Pearsonova a Kendallova korelačního koeficientu.

II. PRAKTICKÁ ČÁST

5 TESTOVÁNÍ HYPOTÉZ METODOU MONTE CARLO

Monte Carlo je třída algoritmů pro simulaci systémů. Jde o stochastické metody používající pseudonáhodná čísla. Typicky využívány pro výpočet integrálů, zejména vícerozměrných, kde běžné metody nejsou efektivní. Metoda Monte Carlo má široké využití od simulací experimentů přes počítání určitých integrálů až třeba řešení diferenciálních rovnic. Základní myšlenka této metody je velice jednoduchá, chceme určit střední hodnotu veličiny, která je výsledkem náhodného děje. Vytvoří se počítačový model toho děje a po proběhnutí dostatečného množství simulací se mohou data zpracovat klasickými statistickými metodami, třeba určit průměr, směrodatnou odchylku a míry asymetrie.

Postup – základy simulace Monte Carlo

1. Determinujeme pseudo-populaci nebo model, který bude reprezentovat skutečnou populaci.
2. Použijeme vzorkovací proceduru pro výběr z pseudo-populace nebo rozdělení.
3. Vypočítáme hodnoty statistik a uložíme je.
4. Opakujeme kroky 2 a 3 pro M iterací.
5. Použijeme M hodnot nalezených v kroku 4 ke studiu rozdělení dané statistiky.

Rychlost konvergence chyby výsledku k nulové hodnotě je u MMC rovna přibližně převrácené hodnotě odmocniny z počtu realizovaných pokusů N , z čehož plyne, že nepatří mezi metody nejefektivnější. Statistické zpracování výsledků v kroku 5 je míněno tak, že hledaná hodnota je zpravidla dána některým z momentů statistických veličin, nejčastěji střední hodnotou a rozptylem.

Postup – Testování hypotéz pomocí Monte Carlo (kritická hodnota)

1. Použijeme vhodný náhodný výběr velikosti n z dané populace, vypočítáme pozorované hodnoty testovací statistiky, t_0 .
2. Rozhodneme se o pseudo-populaci, která reflektuje charakteristiky skutečné populace pod nulovou hypotézou.
3. Získáme náhodný výběr velikosti n z pseudo-populace.

4. Vypočítáme hodnotu testovací statistiky pomocí náhodného výběru v kroku 3 a uložíme ji.
5. Opakujeme krok 3 a 4 pro M iterací. Nyní máme hodnoty t_1, \dots, t_M , které slouží jako odhady rozdělení testovací statistiky, T , kdy nulová hypotéza je pravdivá.
6. Získáme kritickou hodnotu pro danou hladinu významnosti α :

Levostranný test: dostaneme α -tý výběrový kvantil, \hat{q}_α , z t_1, \dots, t_M .

Pravostranný test: dostaneme $(1 - \alpha)$ -tý výběrový kvantil $\hat{q}_{1-\alpha}$, z t_1, \dots, t_M .

Oboustranný test: dostaneme výběrový kvantil $\hat{q}_{\alpha/2}$ a $\hat{q}_{1-\alpha/2}$ z t_1, \dots, t_M .

7. Pokud t_0 spadne do kritické oblasti, potom zamítáme nulovou hypotézu.

Následovat bude ukázka konceptu testování hypotéz pomocí metody Monte Carlo. Datový soubor *mcddata* zahrnuje 25 pozorování. Pro tento ukázkový příklad stanovíme následující nulovou a alternativní hypotézu.

$$H_0: \mu = 454; \quad H_1: \mu < 454.$$

Provedeme náš test hypotézy pomocí simulace k získání kritické hodnoty. Použijeme následující testovací statistiku

$$z = \frac{\bar{x} - 454}{\sigma / \sqrt{n}}.$$

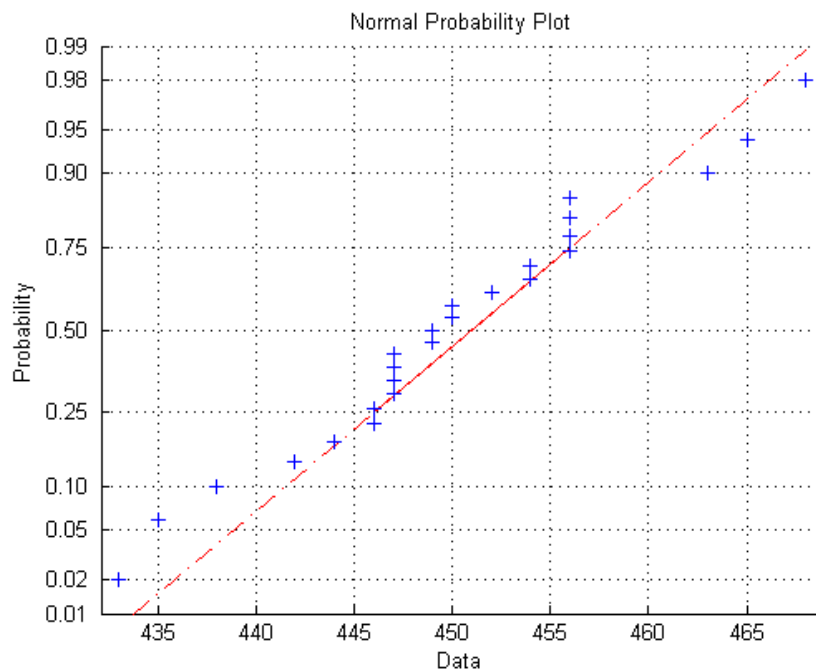
```
% Nahrajeme data.
load mcddata
n = length(mcddata);
% Populacni sm.odch je znama.
sigma = 7.8;
mi = 454;
sigxbar = sigma/sqrt(n);
% Ziskame pozorovane hodnoty testovane statistiky.
Tobs = (mean(mcddata)-mi)/sigxbar;
```

Pozorovaná hodnota testovací statistiky je $t_0 = -2.56$. Pro předpoklad použití parametrického testu hypotézy použijeme normální pravděpodobnostní graf. Výsledný graf na následujícím obrázku ukazuje, že můžeme použít normální rozdělení jako pseudo-populaci.

```
% Tento prikaz generuje normalni pravdepodobnostni graf.
% Je to funkce ve statistickem toolboxu Matlabu.
normplot(mdata)
```

Nyní jsme připraveni implementovat simulaci Monte Carlo. V tomto příkladu použijeme 1000 iterací. V každé iteraci, náhodně vybereme z rozdělení testovací statistiky pod nulovou hypotézou (normální rozdělení s $\mu = 454$ a $\sigma = 7.8$) a uložíme hodnotu testovací statistiky.

```
M = 1000; % Pocet iteraci Monte Carlo
% Uloziste pro testovaci statistiky z MC iteraci.
Tm = zeros(1,M);
% Zahajeni simulace.
for i = 1:M
    % Generovani nahodneho vyberu pod H_0,
    % kde n je velikost vyberu.
    xs = sigma*randn(1,n) + mi;
    Tm(i) = (mean(xs) - mi)/sigxbar;
end
```



Obr. 5.1. Normální pravděpodobnostní graf dat *mdata* ukazuje, že se jedná o normální rozdělení

Nyní, když máme odhadnuté rozdělení testovací statistiky obsažené v proměnné **Tm**, kterou můžeme použít pro stanovení kritického kvantilu pro levostranný test.

```

% Dostaneme kritickou hodnotu pro zvolenou alfu.
% Jedna se o levostranny test, takže dostavame
% alfa kvantil.
alpha = 0.05;
cv = csquantiles(Tm,alpha);

```

Odhadli jsme kritický kvantil -1.75. Odhadnutá hodnota testovací statistiky $t_0 = -2.56$ je menší než hodnota kritického kvantilu, tudíž zamítáme hypotézu H_0 .

Pozn.: Funkci pro nalezení kritického kvantilu můžeme napsat například následovně

```

function qhat = csquantiles(x,p)
% CSQUANTILES Vyberove kvantily.
%
% QHAT = CSQUANTILES(X,P) Vraci vyberove kvantily
% pro pravdepodobnosti zadane v P pomoci vyberu v X.

if ~isempty(find(p <= 0 | p >1))
    error('Pravdepodobnosti musí byt zadany v intervalu 0 a 1.')
    return
end
xs=sort(x);
qhat = zeros(size(p));
n=length(x);
phat = ((1:n)-0.5)/n;
%for i=1:length(p)
% if p(i)~=0
%     ind=find(p(i)<=phat);
%     qhat(i)=x(ind(1));
% elseif p(i)==0
%     qhat(i)=nan;
% end
%end
% k nalezeni kvantilu použijeme linearni interpolaci.
% Uvazujme na phat jako x-ove a poradove statistiky jako y-ove. Budeme
% pouzivat definici, ze j-ta poradova statistika je odhad (j-0.5) / n
% kvantilu. Chceme-li nejaky jiny kvantil, pak muzeme pouzit linearni
% interpolaci.
qhat=interp1(phat,xs,p);

```

Procedura testování hypotézy pomocí Monte Carlo používající přístup založený na p -hodnotě je podobný. Místo hledání kritické hodnoty ze simulovaného rozdělení testované statistiky, použijeme simulaci MC k odhadu p -hodnoty.

Postup – Testování hypotéz pomocí Monte Carlo (p -hodnota)

1. Použijeme vhodný náhodný výběr velikosti n z dané populace, vypočítáme pozorované hodnoty testovací statistiky, t_0 .
2. Rozhodneme se o pseudo-populaci, která reflektuje charakteristiky skutečné populace pod nulovou hypotézou.
3. Získáme náhodný výběr velikosti n z pseudo-populace.
4. Vypočítáme hodnotu testovací statistiky pomocí náhodného výběru v kroku 3 a uložíme ji.
5. Opakujeme krok 3 a 4 pro M iterací. Nyní máme hodnoty t_1, \dots, t_M , které slouží jako odhady rozdělení testovací statistiky, T , kdy nulová hypotéza je pravdivá.
6. Odhadneme p -hodnotu využívající rozdělení nalezeném v kroku 5, dle následujícího:

Levostranný test:

$$\hat{p}\text{-hodnota} = \frac{\#(t_i \leq t_0)}{M}; \quad i = 1, \dots, M$$

Pravostranný test:

$$\hat{p}\text{-hodnota} = \frac{\#(t_i \geq t_0)}{M}; \quad i = 1, \dots, M$$

7. Pokud $\hat{p}\text{-hodnota} \leq \alpha$, potom zamítáme nulovou hypotézu.

Nyní se vrátíme k případové studii testování hypotéz pomocí simulace Monte Carlo, avšak tentokrát za použití p -hodnoty. Jako testovací statistiku použijeme výběrový průměr.

```
% Zmenme testovaci statistiku na xpruh.
```

```
Tobs = mean(mcddata);
```

```
% Pocet iteraci Monte Carlo.
```

```
M = 1000;
```

```
% Zahajime simulaci.
Tm = zeros(1,M);
for i = 1:M
    % Vygenerujeme nahodny vyber pod H_0.
    xs = sigma*randn(1,n) + 454;
    Tm(i) = mean(xs);
end
```

Zjistili jsme odhadovanou p-hodnotu tím, že jsme načítali počet pozorování v T_m , které jsou nižší než hodnoty pozorované statistiky testu a vyděleny M .

```
% Nasledne ziskame p-hodnotu. Toto je pravostranny test.
% Najdeme vsechny hodnoty ze simulace, ktere jsou nizsi nez pozorovane
hodnoty % testovaci statistiky
ind = find(Tm <= Tobs);
pvalhat = length(ind)/M;
```

Dostaneme p-hodnotu 0.007. Pokud jsme zvolili hladinu významnosti $\alpha = 0.05$, tudíž zamítáme nulovou hypotézu.

Ohodnocení modelu testu hypotézy pomocí simulace Monte Carlo

Monte Carlo simulace může být použita pro ohodnocení výkonu modelu nebo testu hypotézy z hlediska chyby I. a II. typu. Pro některé statistické charakteristiky, jako je výběrový průměr, je možné tyto chyby stanovit například analyticky. Avšak, co se stane v případě, máme-li inferenční test, kde předpoklady standardních metod mohou být porušeny nebo analytické metody nelze vůbec použít? Například předpokládejme, že vybereme kritickou hodnotu pomocí normální aproximace (když naše testovací statistika není normálně rozložena), a následně musíme posoudit závěry a jejich relevanci? V těchto situacích můžeme použít simulaci Monte Carlo k odhadu chyby I. a II. druhu. Nejprve byl nastíněn postup pro odhad chyby I. druhu. Protože k chybě typu I. druhu dojde v situaci, když testem zamítneme nulovou hypotézu, ačkoliv ve skutečnosti je pravdivá, musíme vygenerovat vzorek z pseudo-populace, která reprezentuje H_0 .

Postup – posouzení chyby I. druhu pomocí simulace Monte Carlo

1. Vymezíme pseudo-populaci, kdy nulová hypotéza je pravdivá.
2. Vygenerujeme náhodný výběr o rozsahu n z této pseudo-populace.
3. Provedeme test hypotézy pomocí kritické hodnoty.

4. Určíme, jestli nastala chyba I. druhu. Jinak řečeno, jestli zamítáme nulovou hypotézu. Víme, že by neměla být zamítnuta, protože jsme prováděli výběry z rozdělení v souladu s nulovou hypotézou. Uložíme výsledek této iterace jako

$$I_i = \begin{cases} 1; & \text{Nastala chyba I. druhu} \\ 0; & \text{Nenastala chyba I. druhu.} \end{cases}$$

5. Opakujeme kroky 2 až 4 pro M iterací.
6. Pravděpodobnost, že se dopustíme chyby I. druhu je potom

$$\hat{\alpha} = \frac{1}{M} \sum_{i=1}^M I_i. \quad (5.1)$$

Všimněte si, že v kroku 6, je to stejné, jako výpočet falešně zamítnutých hypotéz z M pokusů. To poskytuje odhad hladině významnosti testu pro danou kritickou hodnotu. Pro odhad chyby I. druhu je postup velmi podobný. Chyba II. druhu nastává tehdy, když hypotéza zamítnuta není, přestože neplatí. Při daném rozsahu výběru obvykle nelze současně minimalizovat pravděpodobnosti obou druhů chyb.

Postup – posouzení chyby II. druhu pomocí simulace Monte Carlo

1. Vymezíme pseudo-populaci, kdy nulová hypotéza není pravdivá.
2. Vygenerujeme náhodný výběr o rozsahu n z této pseudo-populace.
3. Provedeme test hypotézy na hladině významnosti α a příslušnou kritickou hodnotu.
4. Určíme, jestli nastala chyba II. druhu. Jinak řečeno, nezamítáme nulovou hypotézu? Uložíme výsledek této iterace jako

$$I_i = \begin{cases} 1; & \text{Nastala chyba II. druhu} \\ 0; & \text{Nenastala chyba II. druhu.} \end{cases}$$

5. Opakujeme kroky 2 až 4 pro M iterací.
6. Pravděpodobnost, že se dopustíme chyby II. druhu je potom

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^M I_i. \quad (5.2)$$

Chyba II. druhu je odhadována pomocí poměru případů, kdy nulová hypotéza není zamítnuta (když by tato situace měla nastat) a M iterací.

Pro test hypotézy, v naší případové studii, máme kritickou hodnotu -1.645. Můžeme odhadnout hladinu významnosti testu pomocí následujících kroků:

```
M = 1000;
alpha = 0.05;
% Dostaneme kritickou hodnotu použitím z-statistiky jako testového
kriteriá.
cv = norminv(alpha,0,1);
% Zahajeni simulace.
Im = 0;
for i = 1:M
    % Vygenerujeme nahodny vyber pod H_0.
    xs = sigma*randn(1,n) + 454;
    Tm = (mean(xs)-454)/sigxbar;
    if Tm <= cv % potom zamitneme H_0
        Im = Im +1;
    end
end
alphahat = Im/M;
```

Kritická hodnota -1,645 v této situaci odpovídá požadované pravděpodobnosti chyby I. druhu 0.05. Z této simulace, dostaneme odhadovanou hodnotu 0,045, což je velmi blízko teoretické hodnotě. V tomto testu nyní zkontrolujeme chybu I. druhu. Všimněte si, že nyní musíme vybírat z alternativních hypotéz.

```
% Nyní zkontroluje pravdepodobnost chyby II. druhu.
% Ziskame nektere alternativni hypotezy:
mualt = 445:458;
betahat = zeros(size(mualt));
for j = 1:length(mualt)
    Im = 0;
    % Ziskame skutecnou stredni hodnotu.
    mu = mualt(j);
    for i = 1:M
        % Vygenerujeme vyber z H_1.
        xs = sigma*randn(1,n) + mu;
        Tm = (mean(xs)-454)/sigxbar;
        if Tm > cv % Potom nezamitame H_0.
            Im = Im +1;
        end
    end
    betahat(j) = Im/M;
end
```

```
% Ziskame odhadovanou silu testu.
```

```
powhat = 1-betahat;
```

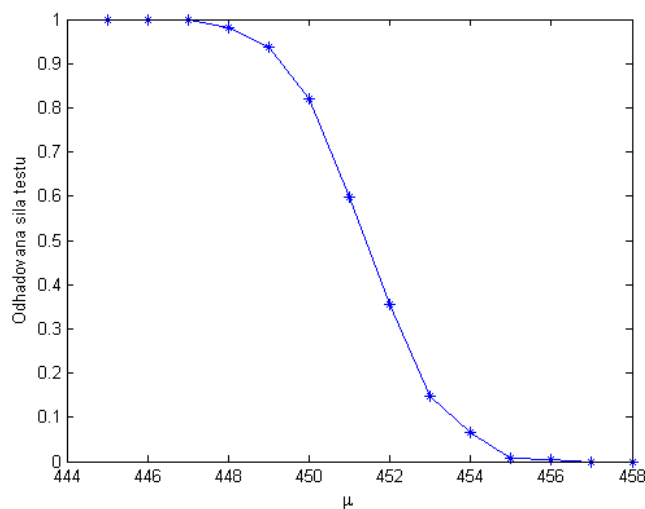
Vykreslíme odhadovanou sílu testu jako silofunkci (viz následující obrázek). Jak se dalo očekávat, skutečná hodnota μ se blíží k hodnotě 454 (střední hodnota pod nulovou hypotézou), síla testu klesá.

```
% Vystup muzeme videt na tomto obrazku.
```

```
plot(muallt, powhat, '*-')
```

```
ylabel('Odhadovana sila testu')
```

```
xlabel('\mu')
```



Obr. 5.2. Odhadovaná síla testu

Důležitým bodem, který je třeba mít na paměti o simulaci Monte Carlo popisované v této části je to, že experiment je použitelný pouze pro situaci, která byla simulována. Například, když jsme posuzovali chybu II. druhu v modelovém příkladu, je vhodná pouze pro dané alternativní hypotézy, velikost vzorku a kritické hodnoty. Jaká by měla být pravděpodobnost chyby II. druhu, pokud je nějaká jiná odchylka od nulové hypotézy, která se používá při simulaci? V jiných případech, bychom mohli potřebovat vědět, zda je rozdělení statistiky variabilní s velikostí vzorku nebo šikmosti v populaci, nebo nějaké jiné charakteristiky. Tyto varianty jsou snadno zkoumány pomocí několika Monte Carlo experimentů. Podstatou této simulace je generování velkého počtu, řádově tisíců budoucích situací (scénářů) a propočítání zvolených kritérií hodnocení pro každý scénář, což pak umožňuje stanovit rozdělení pravděpodobnosti těchto kritérií, číselné charakteristiky a příslušné statistiky pro jednotlivé posuzované situace. Cílem simulace Monte Carlo je

vytvořit takovou posloupnost hodnot náhodné veličiny, která odpovídá danému rozdělení pravděpodobnosti. Náhodná veličina je pak simulována v souladu s tímto rozdělením. K tvorbě posloupnosti hodnot náhodné veličiny se využívají tzv. náhodná čísla, což jsou čísla vybíraná náhodně ze souboru čísel s rovnoměrným rozdělením. Každé číslo souboru má stejnou pravděpodobnost, že bude vybráno. Zdrojem čísel mohou být tabulky náhodných čísel, generátory pseudonáhodných čísel na PC. Tvorba posloupnosti náhodné veličiny spočívá jednak v sestavení kumulativního rozdělení pravděpodobnosti a stanovení rozmezí náhodných čísel a za druhé v generování náhodných čísel a hodnot náhodné veličiny. [7]

6 UKÁZKY STATISTICKÝCH PROGRAMŮ

V této části jsou uvedeny vybrané případové studie, které souvisí s analýzou dat nejen v technických oborech. Pozornost je zaměřena především na prakticky orientované úlohy, které jsou v praxi běžně řešeny nekorektně nebo za nekorektních předpokladů. Každá studie je demonstrována programy v MATLABu, které lze přímo použít pro analýzu dat. Je vždy možné použít jak simulovaná, tak i reálná data.

6.1 Dvourozměrné normální rozdělení

Následovat bude ukázka interaktivního dvourozměrného normálního rozdělení pomocí programovací techniky Switched Board Programming. Tato technika používá příkazy Switch a Case a základem je využití jedné vlastnosti funkcí, a to možnosti volat sama sebe s různými parametry.

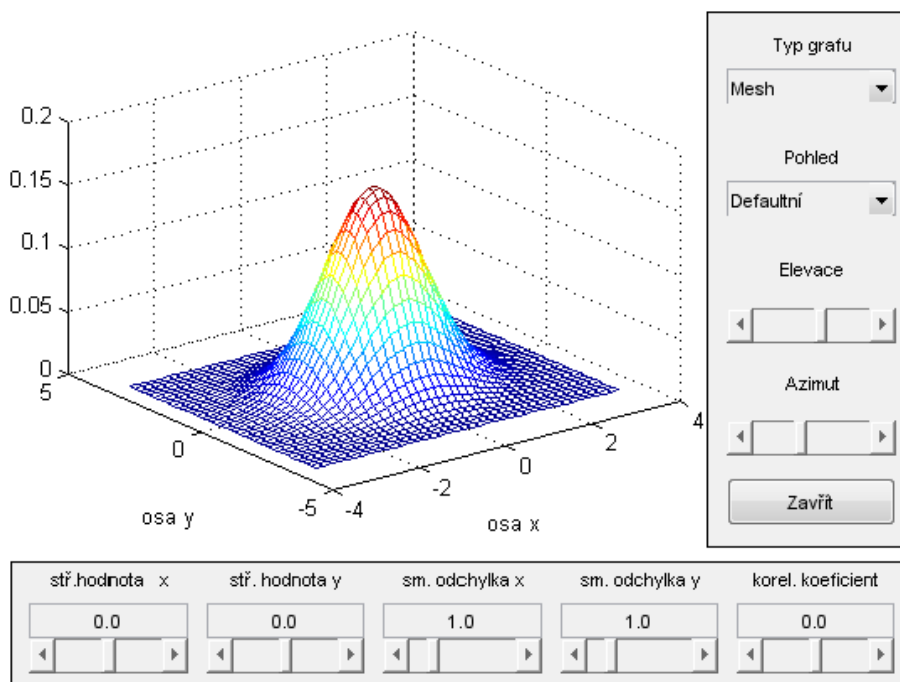
Nechť náhodný vektor (X, Y) má dvourozměrné rozdělení s vektorem středních hodnot μ , a kovarianční maticí Σ

$$\mu = (\mu_x, \mu_y)^T, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}, \quad (6.1)$$

jestliže jeho hustota $f(x, y)$ má tvar

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)\cdot(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right\}, \quad (6.2)$$

kde $(x, y) \in R^2$, a $\rho = \sigma_{xy} / \sigma_x\sigma_y$ je korelační koeficient složek X a Y náhodného vektoru (X, Y) . Pro $|\rho| = 1$ není hustota definována. Jestliže $\rho = 0$, pak veličiny X a Y jsou nejen nekorelované, ale v tomto případě také i nezávislé. [4]



Obr. 6.1. Program pro interaktivní dvourozměrné normální rozdělení

Pomocí tohoto programu může uživatel zadávat libovolnou střední hodnotu vektoru x a y , stejně tak i jejich směrodatné odchylky a korelační koeficient. Může si dále vybrat libovolný typ grafu, měnit pohled, elevaci a azimut. Tato interaktivní modifikace parametrů následně resultuje v graf dvourozměrné normální hustoty pravděpodobnosti. Zdrojový kód programu je uveden v příloze P II.

6.2 Centrální limitní věty

Centrální limitní věty tvrdí, že součty a tedy i průměru velkého počtu nezávislých náhodných veličin mají za velmi obecných podmínek přibližně normální rozdělení. Tyto věty vysvětlují, proč se v různých oborech setkáváme tak často s normálním nebo přibližně normálním rozdělením. Typickým příkladem jsou nepřesnosti při měření; výsledná chyba měření je složena z mnoha různých malých chyb. Centrální limitní věty nám umožňují předpokládat, že rozdělení chyb měření je normální. Proto se normálnímu zákonu rozdělení říká zákon chyb.

Poznámka: O náhodných veličinách, jejichž limitním zákonem je normální rozdělení říkáme, že mají **asymptoticky normální rozdělení**. Centrální limitní věty popisují limity pravděpodobností odchylek náhodných veličin od jejich střední hodnoty.

O centrální limitní větě

Normální rozdělení má pro své vlastnosti klíčový význam v mnoha aplikacích statistiky.

Jak bylo již uvedeno, má-li náhodná veličina $X \sim N(\mu, \sigma^2)$, potom náhodná veličina

$Y = aX + b$, $a \neq 0$, má opět normální rozdělení, $Y \sim N(a\mu + b, a^2\sigma^2)$. V předchozích

odstavcích jsme viděli, že k normálnímu rozdělení se pro velká n blíží rozdělení χ_n^2

a t -rozdělení. Další důležitou vlastností normálního rozdělení je to, že součet konečného

počtu nezávislých normálně rozdělených náhodných veličin má opět normální rozdělení.

Specielně pro X_1, X_2, \dots, X_n nezávislých náhodných veličin se stejným rozdělením

$N(\mu, \sigma^2)$ platí

$$\text{a) } E\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \frac{n\mu}{n} = \mu,$$

$$\text{b) } \text{var}\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

$$\text{c) } \text{je-li } Y = \frac{1}{n}(X_1 + X_2 + \dots + X_n), \text{ pak } Y \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

K normálnímu rozdělení se však přibližuje i součet nezávislých náhodných veličin

z jakéhokoliv rozdělení. Je to důsledek tzv. *centrální limitní věty*. Jsou-li X_1, X_2, \dots, X_n

vzájemně nezávislé náhodné veličiny téhož (ale jinak libovolného rozdělení) se střední

hodnotou μ a rozptylem σ^2 , pak pro každé reálné x platí

$$\lim_{n \rightarrow \infty} P\left[\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) < x\right] = \Phi(x). \quad (6.3)$$

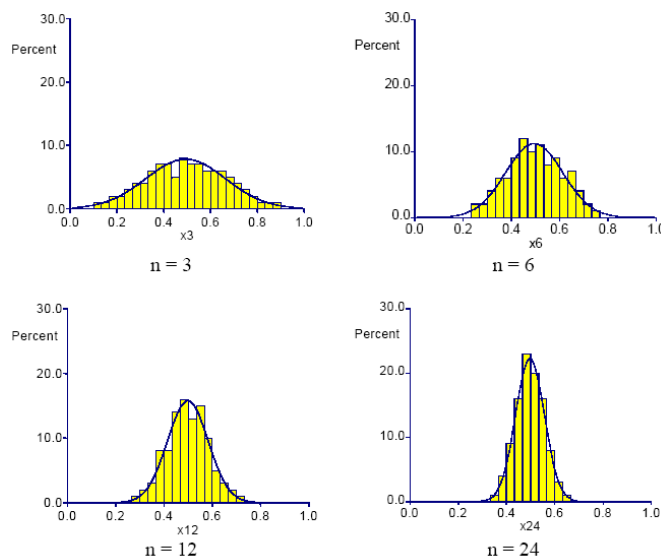
Tzn., že pro dostatečně velké n se distribuční funkce náhodné veličiny

$$Z_n = \frac{\sum_{i=1}^n X_i - E\left(\sum_{i=1}^n X_i\right)}{\sqrt{\text{var}\left(\sum_{i=1}^n X_i\right)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \quad (6.4)$$

jen nepatrně liší od distribuční funkce normovaného normálního rozdělení. Volně řečeno,

součet (a tedy i průměr) většího počtu nezávislých stejně rozdělených náhodných veličin

má přibližně normální rozdělení. Tuto skutečnost ilustruje následující příklad na následujících obrázcích, ve kterém jsou znázorněna empirická rozdělení hodnot získaných z 1000 nezávislých realizací náhodné veličiny $Y = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$, kdy náhodné veličiny X_i , $i = 1, 2, \dots, n$ měly rovnoměrné spojité rozdělení na intervalu $(0, 1)$, a n bylo postupně rovno 3, 6, 12 a 24. Z histogramů na obrázku vidíme, že s rostoucím n se empirické rozdělení stále těsněji blíží k normálnímu rozdělení a také se zmenšuje rozptyl.



Obr. 6.2. Rozdělení výběrových průměrů
z rovnoměrného rozdělení pro různé rozsahy výběru

Nejjednodušší případ centrální limitní věty je tzv. Moivreova-Laplaceova věta, která vyjadřuje konvergenci binomického rozdělení k rozdělení normálnímu a dává tak možnost aproximovat binomické rozdělení rozdělením normálním.

Moivreova-Laplaceova věta. Necht' X_1, X_2, \dots je posloupnost nezávislých stejně rozdělených náhodných veličin s alternativním rozdělením $A(p)$.

Položme $S_n = \sum_{i=1}^n X_i$ a $Z_n = (S_n - np) / \sqrt{np(1-p)}$. Potom platí

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x), \quad x \in \mathbb{R}. \quad (6.5)$$

Příklad: *Aproximace binomického rozdělení normálním rozdělením*

Student se podrobí zkoušce ve formě testu s 10 otázkami, na které odpovídá ANO nebo NE. Student hádá odpovědi na všechny otázky. Užijte binomické rozdělení ke stanovení

přesné pravděpodobnosti, že student odpoví na 7 nebo 8 otázek správně. Pak použijte aproximaci binomického rozdělení normálním rozdělením.

Řešení: Necht' S_{10} je počet správných odpovědí na 10 otázek. Protože student hádá odpovědi, je pravděpodobnost správné odpovědi $p = 0.5$, $S_{10} \sim B(10, 0.5)$. Z tabulky binomického rozdělení nebo přímým výpočtem dostaneme

$$P(S_{10} = 7 \vee 8) = P(7) + P(8) = 0.1172 + 0.0439 = 0.1611.$$

($X = 7 \vee 8$ označuje výrok X se rovná 7 nebo 8).

$E(S_{10}) = np = 10 \cdot 0.5 = 5$ a $D(S_n) = \sqrt{np(1-p)} = 1.58$. Protože n není příliš vysoké, je třeba při použití normální aproximace provést korekci pro nahrazení diskrétního rozdělení spojitým, tzv. *korekci na spojitost*. Úlohu lze totiž formulovat jako určení $P(6.5 \leq S_{10} \leq 8.5)$, neboť platí

$$\begin{aligned} P(6.5 \leq S_{10} \leq 8.5) &= P(S_{10} \leq 8.5) - P(S_{10} < 6.5) = \\ &= P(S_{10} \leq 8) - P(S_{10} \leq 6) = P(S_{10} = 8) + P(S_{10} = 7). \end{aligned}$$

Použitím Moivreova-Laplaceovy věty dostaneme

$$\begin{aligned} P\left(\frac{6.5-5}{1.58} \leq Z_{10} \leq \frac{8.5-5}{1.58}\right) &= P(0.95 \leq Z_{10} \leq 2.22) = \Phi(2.22) - \Phi(0.95) = \\ &= 0.9868 - 0.8289 = 0.1579 \end{aligned}$$

Porovnáním této hodnoty s hodnotou $P(S_{10} = 7 \vee 8)$ vidíme, že normální aproximace je velice dobrou aproximací binomického rozdělení. [1]

6.3.6 Linderbergova-Lévyho věta

Necht' X_1, X_2, \dots, X_n jsou nezávislé náhodné veličiny se stejným rozdělením, které mají konečnou střední hodnotu μ a rozptyl σ^2 . Položme $Y_n = \sum_{i=1}^n X_i$ a $Z_n = (Y_n - n\mu) / \sigma\sqrt{n}$.

Potom platí

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x), \quad x \in \mathbb{R} \quad (6.6)$$

Podle této věty konverguje distribuční funkce normovaných součtů k distribuční funkci $N(0, 1)$ -rozdělení pro libovolné výchozí rozdělení s konečnou střední hodnotou a konečným rozptylem. Jinak řečeno součet a tím i průměr n nezávislých náhodných

veličin, které mají stejné (libovolné) rozdělení s konečnou střední hodnotou a konečným rozptylem má pro dosti velké n přibližně normální rozdělení. [1]

6.3.7 Ljapunovova centrální limitní věta

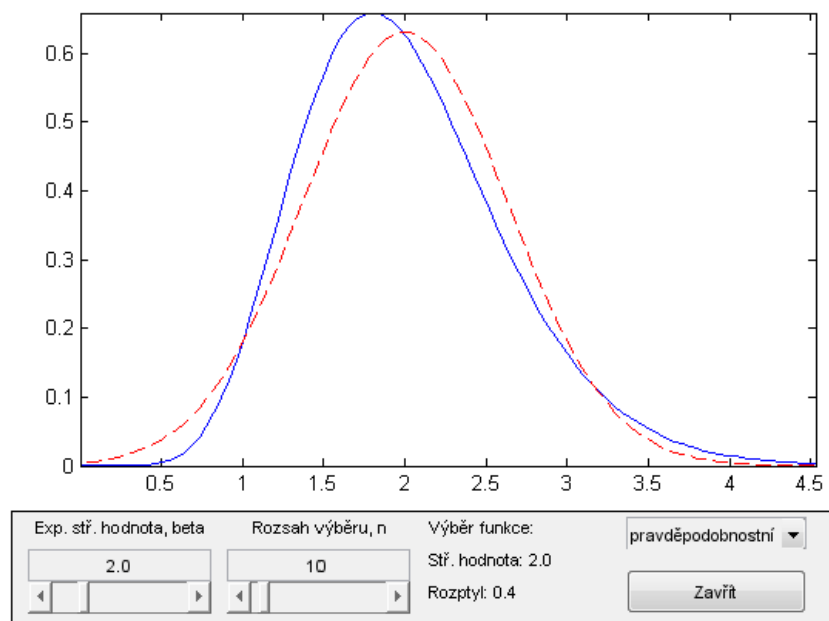
Tato věta nevyžaduje, na rozdíl od Lindebergovy-Lévyovy, stejné rozdělení u jednotlivých náhodných veličin.

Jsou-li X_1, X_2, \dots, X_n nezávislé náhodné veličiny, $s_n^2 = \text{var} \sum_{i=1}^n X_i = \sum_{i=1}^n E(X_i - EX_i)^2$ a je-li

$$\frac{\sum_{i=1}^n E|X_i - EX_i|^{2+\delta}}{s_n^{2+\delta}} \xrightarrow{n \rightarrow \infty} 0 \quad \text{pro nějaké } \delta > 0, \quad \text{potom pro každé } x \in R$$

$$P \left[\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n EX_i}{\sqrt{s_n^2}} < x \right] \rightarrow \Phi(x) \quad \text{při } n \rightarrow \infty, \quad \text{kde } \Phi \text{ je distribuční funkce standardního}$$

normálního rozdělení $N(0, 1)$. [1]



Obr. 6.3. Program pro aproximaci exponenciálního rozdělení normálním

Tento program slouží k demonstrační ukázké aproximace exponenciálního rozdělení rozdělením normálním. Uživatel může libovolně modifikovat parametr střední hodnoty v případě exponenciálního rozdělení a zvolit si libovolný rozsah výběru. Dále se zde nabízí výběr funkce, buď pravděpodobnostní, nebo distribuční. Graf resultuje v důkaz platnosti

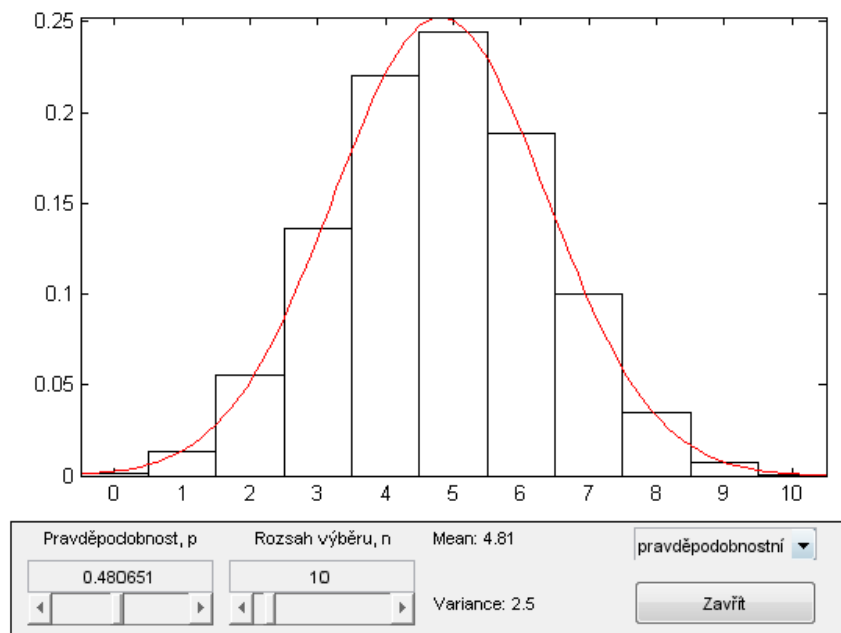
centrální limitní věty, která tvrdí, že součty a tedy i průměru velkého počtu nezávislých náhodných veličin mají za velmi obecných podmínek přibližně normální rozdělení. Zdrojový kód programu je uveden v příloze P III.

6.4.3 Aproximace binomického rozdělení normálním

Jestliže $np(1-p) > 9$, můžeme binomické rozdělení $Bi(n, p)$ náhodné veličiny X , aproximovat normálním rozdělením $N(\mu, \sigma^2)$, kde klademe $\mu = np$ a $\sigma^2 = np(1-p)$. Pro celá nezáporná čísla a, b potom je (s ohledem na skutečnost, že jde o diskrétní rozdělení)

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b + 0,5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0,5 - np}{\sqrt{np(1-p)}}\right). \quad (6.7)$$

Protože $Z_n = \sum_{i=1}^n X_i$ je náhodná veličina s rozdělením $Bi(n, p)$, plyne z Moivre-Laplaceovy věty, že lze rozdělení Z_n aproximovat rozdělením $N(np, np(1-p))$. Vhodnost aproximace je tím lepší, čím je p bližší 0.5. Zlepšuje se s rostoucím rozptylem a rovněž se zavádí oprava na spojitost. [1]



Obr. 6.4. Program pro aproximaci binomického rozdělení rozdělením normálním

Tento program slouží k demonstrační ukázce aproximace binomického rozdělení rozdělením normálním. Uživatel může libovolně modifikovat parametr p binomického rozdělení a rozsah výběru. Dále se zde nabízí výběr funkce, buď pravděpodobnostní, nebo

distribuční. Graf resultuje v důkaz platnosti Moivreovy-Laplaceovy věty, že při velkém počtu nezávislých pokusů konverguje binomické rozdělení k rozdělení normálnímu. Zdrojový kód programu je uveden v příloze P IV.

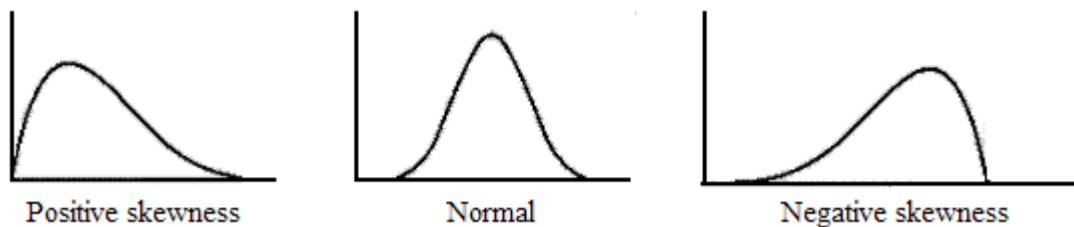
6.3 Míry asymetrie

Šikmostí náhodné veličiny X s nenulovým rozptylem rozumíme reálné číslo

$$A_3(X) = \frac{E([X - E(X)]^3)}{[\sigma(X)]^3} \text{ nebo } A_3(X) = \frac{M_3}{M_2\sqrt{M_2}}. \quad (6.8)$$

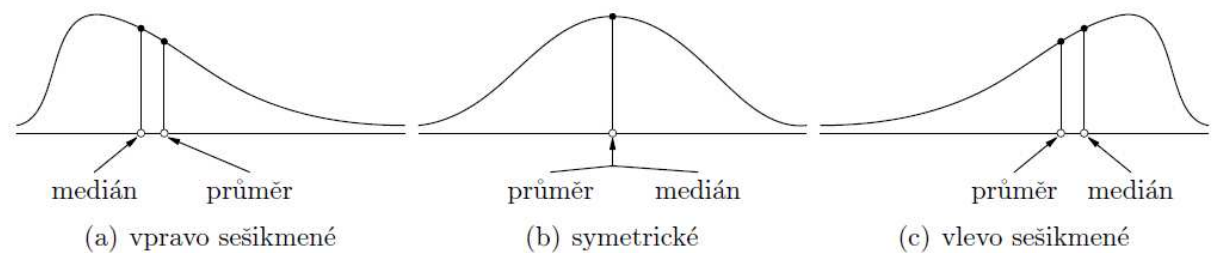
Vlastnosti: Pro šikmost náhodné veličiny X platí

1. $A_3(X) = 0$, je-li rozdělení náhodné veličiny X symetrické,
2. $A_3(X) < 0$, je-li rozdělení náhodné veličiny X doprava zešikmené,
3. $A_3(X) > 0$, je-li rozdělení náhodné veličiny X doleva zešikmené,
4. $A_3(aX + b) = A_3(X)$ pro všechna $a, b \in \mathbb{R}$, $a \neq 0$.



- a) Hustota při $A_3(X) > 0$ b) Hustota při $A_3(X) = 0$ c) Hustota při $A_3(X) < 0$

Míry šikmostí jsou založeny na porovnání stupně nahuštěnosti malých hodnot sledované náhodné veličiny se stupněm nahuštěnosti velkých hodnot této náhodné veličiny.



Obr. 6.5. Vzájemná poloha průměru a mediánu

Stejný stupeň hustoty malých a velkých hodnot se zpravidla projevuje v symetrii tvaru rozdělení četností. Větší stupeň nahuštěnosti malých hodnot v porovnání s hustotou velkých hodnot se projeví vpravo *zešikmeným tvarem rozdělení četností*, které označujeme také *kladně zešikmeným tvarem rozdělení* (A_3 je kladné číslo). Větší stupeň nahuštěnosti velkých hodnot ve srovnání s hustotou malých hodnot se projeví zpravidla *vlevo zešikmeným tvarem rozdělení četností*, které také nazýváme *záporně zešikmeným tvarem rozdělení* (A_3 je záporné číslo).

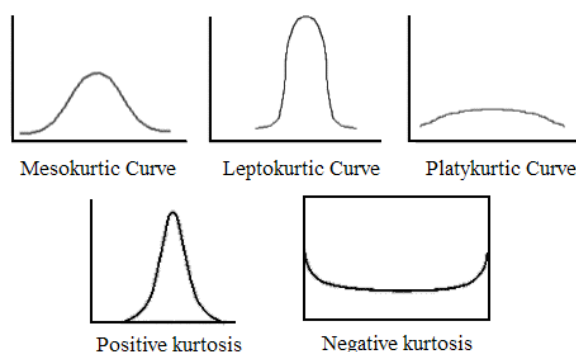
Špičatostí náhodné veličiny X s nenulovým rozptylem rozumíme reálné číslo

$$A_4(X) = \frac{E([X - E(X)]^4)}{[\sigma(X)]^4} - 3. \quad (6.9)$$

Možná překvapuje, že z rovnice pro špičatost odečítáme na pravé straně trojku. Důvod je ten, že špičatost vztahujeme k nejčastěji vyskytujícímu se rozdělení, k tzv. normálnímu rozdělení, u kterého je poměr $M_4/(M_2)^2$ roven 3. Špičatost je tedy vztažena ke špičatosti normálního rozdělení, kladná špičatost znamená špičatější rozdělení než normální, záporná špičatost znamená, že rozdělení pozorovaných hodnot je „placatější“ než normální.

Vlastnosti: Pro špičatost náhodné veličiny X platí

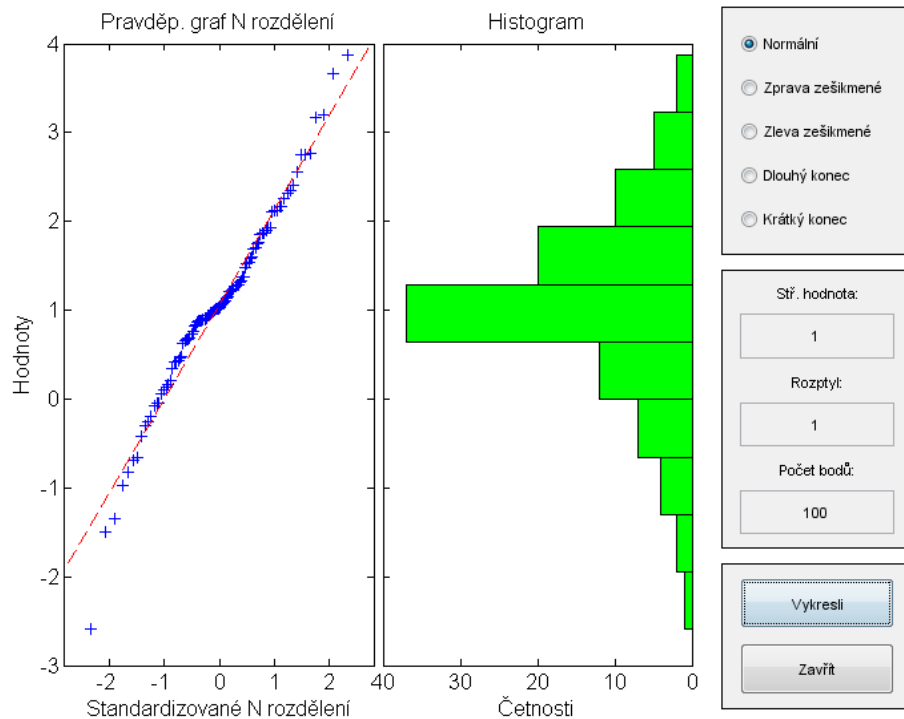
1. Špičatost náhodné veličiny s normálním rozdělením je $A_4(X) = 0$.
2. $A_4(aX + b) = A_4(X)$ pro všechna $a, b \in \mathbb{R}$, $a \neq 0$.



Obr. 6.6. Možné tvary špičatého rozdělení

Míry špičatosti jsou založeny na porovnání stupně nahuštěnosti hodnot střední velikosti se stupněm nahuštěnosti ostatních hodnot, respektive všech hodnot sledované náhodné veličiny. Jsou-li četnosti středních hodnot srovnatelné s četnostmi ostatních hodnot znaku, špičatost se zpravidla projevuje plochým tvarem rozdělení četností. Větší stupeň

koncentrace prostředních hodnot ve srovnání s četnostmi všech hodnot znaku se projeví špičatým tvarem rozdělení četností. Z vyšší číselné hodnoty míry A_4 se zpravidla usuzuje na špičatější rozdělení četností a tím zároveň na vyšší stupeň koncentrace prostředních hodnot ve srovnání s ostatními hodnotami sledovaného znaku. Často se používají různé modifikace míry šikmosti A_3 a míry špičatosti A_4 , které zde nebudu uvádět. [1]



Obr. 6.7. Program pro demonstraci symetrie a asymetrie rozložení dat

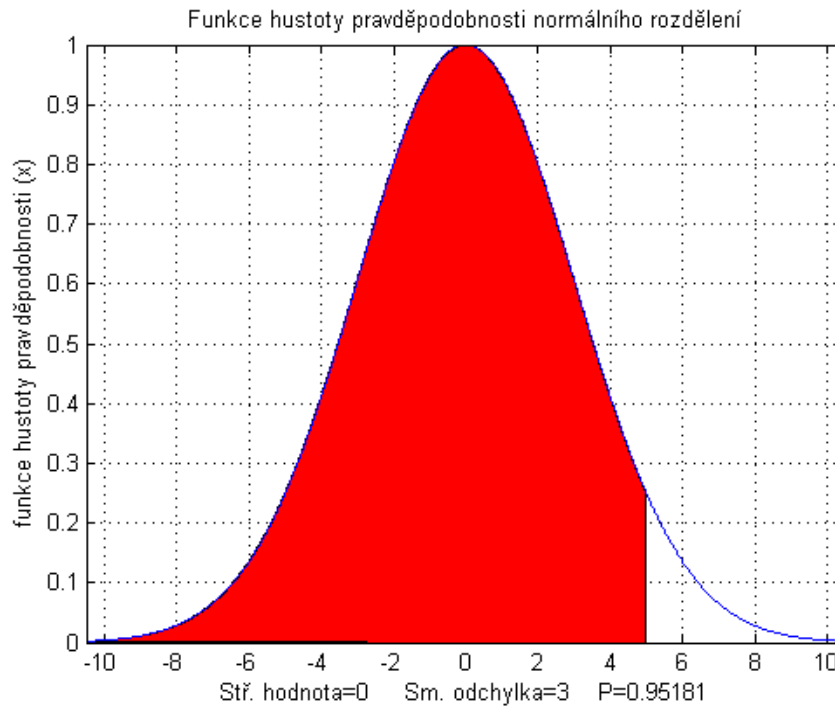
Tento program slouží pro grafickou vizualizaci symetrického a asymetrického rozdělení dat. Uživatel může libovolně modifikovat symetričnost a asymetričnost dat s libovolnými parametry střední hodnoty, rozptylu a velikosti výběru. Následné zadání těchto parametrů pak interaktivně resultuje v pravděpodobnostní graf normálního rozdělení a histogram.

Algoritmus počítá se standardizovanou kvantilovou funkcí $QI_e = \frac{(x_{(i)} - x_{0.5})}{2 \cdot (x_{0.75} - x_{0.25})}$, kdy

$|QI_e(P)| \geq 1$ jsou vybočující hodnoty (pro normální rozdělení) nebo ukazují na dlouhé konce. Šikmost je pak počítána dle vztahu $SQ = QI_e(0.25) + QI_e(0.75)$, délka konců $QI_e(0.5)$ resp. $QI_e(0.95)$ (má cenu jen pro symetrická rozdělení), kdy krátké konce $QI_e(0.95) < 0.5$, dlouhé konce $QI_e(0.95) > 1$, střední konce $0.5 < QI_e(0.95) < 1$. Zdrojový kód programu je uveden v příloze P V.

Program na výpočet pravděpodobnosti pod křivkou funkce hustoty pravděpodobnosti:

Tento program slouží pro ukázkou výpočtu plochy pod křivkou funkce hustoty pravděpodobnosti v případě normálního rozdělení $N(\mu; \sigma^2)$.



Obr. 6.8. Program na výpočet pravděpodobnosti pod křivkou funkce hustoty pravděpodobnosti

Vyvoláním funkce s parametry *normaldistribution(5, 0, 3)*, dostane uživatel vypočtenou pravděpodobnost 0.95181. Pro lepší představu jsou uvedeny parametry této funkce v následujícím tvaru: *normaldistribution(P(X ≤ 5), μ, σ)*. Zdrojový kód programu je uveden v příloze P VI.

6.4 Testování normality dat

Princip všech testů normality je týž. Testujeme hypotézu H_0 : náhodný výběr X_1, X_2, \dots, X_n pochází ze základního souboru s normálním rozdělením proti alternativě H_1 : výběr pochází ze základního souboru s jiným rozdělením. Vypočte se testovaná statistika a srovná se s příslušnou p -hodnotou. Pokud je p -hodnota větší než zvolená hladina významnosti α , pak nulovou hypotézu nezamítáme. V opačném případě se na hladině významnosti α přikloníme k alternativní hypotéze. Všechny hodnoty X_1, X_2, \dots, X_n náhodného výběru se

uspořádají podle velikosti $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ a vypočte se $U_{(i)} = \frac{X_{(i)} - M}{S}$, kde

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2$ je výběrový rozptyl a $M = \frac{1}{n} \sum_{i=1}^n X_i$ je výběrový průměr.

Testované statistiky jsou:

1. A-D (Anderson-Darling) test:

$$AD = -\frac{1}{n} \sum_{i=1}^n (2i-1) \left[\ln \Phi(U_{(i)}) + \ln(1 - \Phi(U_{(n-i+1)})) \right] - n, \quad (6.10)$$

kde $\phi_{(*)}$ je distribuční funkce standardizovaného normálního rozložení.

2. R-J (Ryan-Joiner) test:

$$R = \frac{\sum_{i=1}^n X_{(i)} q_i}{\sum_{i=1}^n (X_i - M)^2 * \sum_{i=1}^n q_i^2}, \quad (6.11)$$

kde kvantil q_i splňuje rovnici $\Phi_{(q_i)} = \frac{i-0,3}{n+0,4}$.

3. K-S (Kolmogorov-Smirnov) test:

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - \Phi(x)|, \quad (6.12)$$

kde $F_n(x)$ je empirická distribuční funkce. Tento test modifikoval Liliefors kritickými hodnotami stanovenými metodou Monte Carlo. Lilieforsova modifikace se používá v situacích, kdy neznáme parametry rozložení a odhadujeme je z dat.

4. S-W (Shapiro-Wilks) test:

$$W = \frac{\sum_{i=1}^m a_i^{(n)} [X_{(n-i+1)} - X_{(i)}]^2}{\sum_{i=1}^m (X_i - M)^2}, \quad (6.13)$$

kde $m = n/2$ pro n sudé a $m = (n-1)/2$ pro n liché. Koeficienty $a_i(n)$ jsou tabelovány. Na testovou statistiku W lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení. V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít W hodnota

1. hypotézu o normalitě tedy zamítneme na hladině významnosti α , když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení $N(0, 1)$. [1], [9]

V této části bude uveden test normality pro jednorozměrný a vícerozměrný vektor.

6.4.1 Vícerozměrný test normality

Tato grafická pomůcka slouží k porovnání rozdělení m -rozměrných dat s m -rozměrným normálním rozdělením pomocí Mahalanobisovy vzdálenosti převedené na veličinu s F -rozdělením. Jedná se o analogii Q-Q grafu pro jednorozměrná data. Postup konstrukce grafu je následující:

Pro každý řádek datové matice s m sloupci a n řádky (n m -rozměrných dat) vypočítáme hodnotu Z_i

$$Z_i = (\mathbf{x}_i - \mathbf{x}_{mi}) \mathbf{S}_j^{-1} (\mathbf{x}_i - \mathbf{x}_{mi})^T, \quad (6.14)$$

kde

$$\mathbf{S}_{ji} = \frac{1}{\mathbf{x}_{ci} \mathbf{S}^{-1} \mathbf{x}_{ci}^T} (\mathbf{S}^{-1} \mathbf{x}_{ci}^T \mathbf{x}_{ci} \mathbf{S}^{-1}) + \frac{n-2}{n-1} \mathbf{S}^{-1} \quad \text{a} \quad \mathbf{x}_{mi} = \frac{1}{n-1} (n\bar{\mathbf{x}} - \mathbf{x}_i).$$

\mathbf{x}_{ci} je centrováný i -tý řádek matice dat \mathbf{X} , $\mathbf{x}_{ci} = \mathbf{x}_i - \bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ je vektor (sloupcových) průměrů.

Získáme hodnoty Z_i , které mají F -rozdělení s m a $n-m$ stupni volnosti. Tyto hodnoty vyneseme proti teoretickým kvantilům rozdělení $F(m, n-m)$. Leží-li takto získané body grafu přibližně na přímce, můžeme považovat data za přibližně normální. [1], [2], [3]

V následujícím příkladu mějme následující trojrozměrný datový soubor, pro který se budeme snažit prokázat předpoklad vícerozměrné normality.

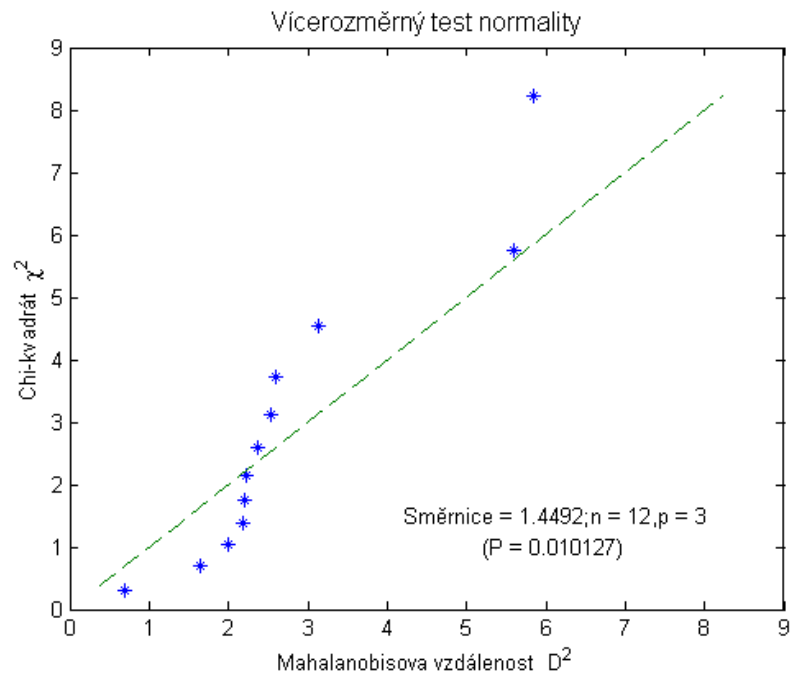
x1	x2	x3	x1	x2	x3
2.4	2.1	2.4	4.5	4.9	5.7
3.5	1.8	3.9	3.9	4.7	4.7
6.7	3.6	5.9	4.0	3.6	2.9
5.3	3.3	6.1	5.7	5.5	6.2
5.2	4.1	6.4	2.4	2.9	3.2
3.2	2.7	4.0	2.7	2.6	4.1

```
>> Multitest (X, 0.05)
```

Rozsah výběru	Proměnné	Směrnice	t	P
12	3	1.4492	2.7564	0.0101

Zadaná hladina významnosti je: 0.05

Předpoklad vícerozměrné normality není prokázán.



Obr. 6.9. Program pro test vícerozměrné normality

Body grafu neleží přibližně na přímce, proto nemůžeme považovat data za přibližně normální. Navíc p -hodnota $\leq \alpha$ (0,05), tudíž zamítáme hypotézu o vícerozměrné normalitě dat. Zdrojový kód programu je uveden v příloze P VII.

6.4.2 Anderson-Darlingův test normality dat

Jednorozměrnou normalitu můžeme otestovat například pomocí Anderson-Darlingova testu. Andersonův – Darlingův test je definován pro ověření hypotézy:

H_0 : data pocházejí ze základního souboru se specifikovaným rozdělením náhodné veličiny;

H_1 : data pocházejí ze základního souboru s jiným než specifikovaným rozdělením náhodné veličiny.

Testová statistika tohoto testu je definována následovně:

$$AD = -\frac{1}{n} \left[\sum_{i=1}^n (2i-1) \left\{ \ln F(x_{(i)}) + \ln (1 - F(x_{(n+1-i)})) \right\} \right] - n, \quad (6.15)$$

kde $x_{(i)}$ jsou podle velikosti vzestupně uspořádané napozorované hodnoty, n je rozsah výběru, F je distribuční funkce specifikovaného rozdělení sledovaného znaku jakosti.

Hypotéza H_0 se zamítá na hladině významnosti α , je-li vypočítaná hodnota testové statistiky AD větší, než její $(1-\alpha)$ kvantil. V případě, že specifikovaným rozdělením je v praxi nejčastěji normální rozdělení, potom pro velký rozsah výběru je přibližná hodnota 0,95 – kvantilu rovna¹

$$AD_{0,95} = 1,0348 (1 - 1,013/n - 0,93 / n^2), \quad (6.16)$$

V případě normálního rozdělení počítá např. MINITAB testovou statistiku na základě výběrového průměru a výběrové směrodatné odchylky místo na základě známých parametrů μ a σ . Počítá rovněž přibližnou p -hodnotu, která je citlivější pro rozhodnutí o přijetí, či zamítnutí testované hypotézy. Pomocí zjištěné hodnoty testové statistiky AD se vypočítá hodnota

$$A = AD \cdot \left(1 + \frac{0,75}{n} + \frac{2,25}{n^2} \right). \quad (6.17)$$

V závislosti na velikosti A se odhaduje příslušná p -hodnota²:

pokud	$13 > A \geq 0,6$,	je	$p = \exp(1,2937 - 5,709A + 0,0186A^2)$;
	$0,6 > A \geq 0,34$,	je	$p = \exp(0,9177 - 4,279A - 1,38A^2)$; (6.18)
	$0,34 > A \geq 0,2$,	je	$p = 1 - \exp(-8,318 + 42,796A - 59,938A^2)$;
	$0,2 > A$,	je	$p = 1 - \exp(-13,436 + 101,14A - 223,73A^2)$.

Například mějme následující datový vektor:

$X = [245.3; 245.6; 245.0; 244.7; 244.6; 245.9; 245.2; 245.5; 246.1; 246.5; 246.7; 246.0;$
 $245.8; 245.3; 245.3; 246.0; 245.5; 246.7; 245.3; 245.8; 245.8; 245.2; 245.7; 244.8; 246.7;$

¹ Hebák, P., Bílková, D., Svobodová A. - *Praktikum k výuce matematické statistiky II: Testování hypotéz*, VŠE 2002.

² D'Agostino, R. B., Stephens, M. A. - *Goodness-of-Fit Techniques*, Marcel Dekker (1986).

246.5; 245.5; 246.0; 244.8; 244.8; 244.4; 244.8; 245.6; 245.5; 244.0; 245.2; 244.8; 245.4; 246.1; 245.9; 245.6; 245.2];

V programovém prostředí Matlab byl sestaven skript pro test normality pomocí Anderson-Darlingova testu.

```
>> AnDartest(x)
```

Rozsah výběru: 43

Anderson-Darlingova statistika: 0.2858

Anderson-Darling upravená statistika: 0.2912

P-hodnota Anderson-Darlingovy statistiky = 0.6088

Se zvolenou hladinou významnosti = 0.050

Výběr pochází z normálního rozdělení.

Takže tento soubor pochází z normálního rozdělení se střední hodnotou a rozptylem = 245.4814 0.4139

Zdrojový kód programu je uveden v příloze P VIII.

6.5 Testování normality pomocí exploratorních grafů

Kvantilový graf (osa x: pořadová pravděpodobnost P_i , osa y: pořádková statistika $x_{(i)}$).

Umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, které může být symetrické, zešikmené k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce normálního rozdělení, $N_{P_i} = \hat{\mu} + \hat{\sigma}\mu_{P_i}$, pro $0 \leq P_i \leq 1$:

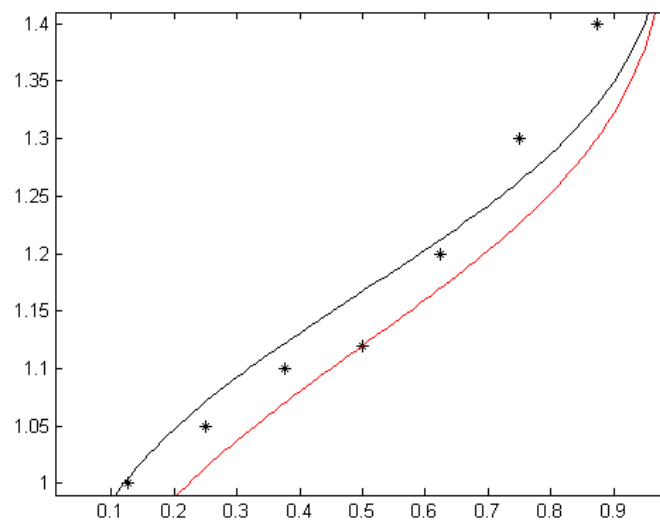
- 1) Klasických odhadů parametrů polohy a rozptýlení, tj. aritmetického průměru a směrodatné odchylky $\hat{\mu} = \bar{x}$ a $\hat{\sigma} = s$, a dále
- 2) Robustních odhadů, tj. mediánu M , $\hat{\mu} = M$ a $\hat{\sigma} = R_F / 1.349$, kde $R_F = F_H - F_D$ je interkvartilové rozpětí.

Kvantilově-kvantilový graf (graf Q-Q) (osa x: $Q_T(P_i)$, osa y: $x_{(i)}$). Umožňuje posoudit shodu výběrového rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$.

Jako odhad kvantilové funkce výběru se využívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením platí přibližná rovnost kvantilů $x_{(i)} \approx Q_T(P_i)$, kde P_i je pořadová pravděpodobnost, a závislost $x_{(i)}$ na $Q_T(P_i)$ je přibližně přímka. Pro porovnání rozdělení výběru s normálním rozdělením se graf Q-Q nazývá *grafem rankitovým*. Umožňuje také orientační zařazení výběrového rozdělení do skupin podle šikmosti, špičatosti a délky konců. [9], [10]

6.5.1 Test normality pomocí kvantilové funkce

Výběrová kvantilová funkce $(P, x_{(i)})$, kde $i = 1, 2, \dots, n$. Teoretická kvantilová funkce pro normální rozdělení – průměr a směrodatná odchylka (černě). Teoretická robustní kvantilová funkce pro normální rozdělení – medián a upravená interkvartilová odchylka (červeně). [9], [10]



Obr. 6.10. Výběrová kvantilová funkce

```
x = [1 1.2 1.05 1.1 1.3 1.12 1.4];
x1 = sort(x);
prum = mean(x);
rsm = sqrt(var(x));
iv = 1:size(x,2);
pi = iv./(size(x,2)+1);
plot(pi,x1,'k*');
di = .1.*max(diff(x1));
axis([.01 .99 x1(1)-di x1(size(x,2))+di]);
```

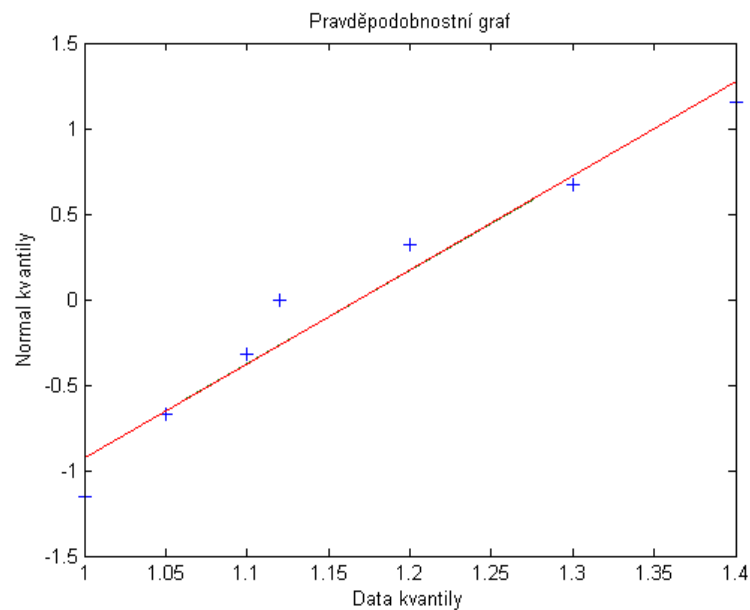
```

x2 = .01:.01:.99;
x25 = prctile(x,25);x50 = prctile(x,50);x75 = prctile(x,75); rf = (x75-
x25)/1.349;
for i = 1:size(x2,2)
nt(i) = norminv(x2(i),prum,rsm);
ntr(i) = norminv(x2(i),x50,rf);
end
hold on
plot(x2,nt,'k-');
plot(x2,ntr,'r-')

```

6.5.2 Test normality pomocí rankitového grafu s robustní přímkou

Účelem je vytvořit program pro kreslení Q-Q grafu pro normální rozdělení s robustní přímkou (z kvartilů) s využitím vektorizace kvantilové funkce normálního rozdělení.



Obr. 6.11. Rankitový graf s robustní přímkou

```

% rankitový graf s robustní přímkou (Q-Q graf pro normalitu)
% data v poli x se převádí na sloupec
x = [1 1.2 1.05 1.1 1.3 1.12 1.4];[rows,cols]=size(x);
if rows==1,x=x(:);rows=cols;cols=1;
end
xs = sort(x);minx=min(x);maxx = max(x);del = maxx-minx;
if del>0.025,minxa = minx-0.025*del;maxxa = maxx+0.025*del;
else
minxa = minx - 6.2500e-004;maxxa = maxx+6.2500e-004;
end;

```

```

eprob = 1./(rows+1):1./(rows+1):(rows./(rows+1));
y = norminv(eprob,0,1)';
minyaxis = norminv(0.25./(rows+1),0,1);
maxyaxis = norminv((rows+1-0.25)./(rows+1),0,1);
q1x = prctile(x,25);q3x = prctile(x,75);q1y = prctile(y,25);
q3y = prctile(y,75);qx = [q1x; q3x];qy = [q1y;q3y];dx = q3x - q1x;
dy = q3y-q1y;slope = dy./dx;centerx=(q1x+q3x)/2;centery=(q1y+q3y)/2;
maxx = max(x);minx = min(x);maxy = centery+slope.*(maxx-centerx);
miny = centery-slope.*(centerx-minx);mx=[minx maxx];my=[miny;maxy];
plot(xs,y,'+',qx,qy,mx,my);axis = ([minx maxx, minyaxis maxyaxis]);
xlabel('Data kvantily');ylabel('Normal kvantily');title('Pravděpodobnostní graf');

```

V obou dvou případech, simulovaná data vykazují mírnou asymetrii, resp. zprava sešikmené rozdělení. Můžeme konstatovat, že výběr nepochází z normálního rozdělení.

6.6 Odhad intervalů spolehlivosti pro střední hodnotu a rozptyl

Následující část se bude pro teoretické podklady odvolávat na kapitolu 2.1.2, kde byl popsán princip konstrukce intervalu spolehlivosti pro střední hodnotu.

Princip intervalu spolehlivosti pro střední hodnotu

```

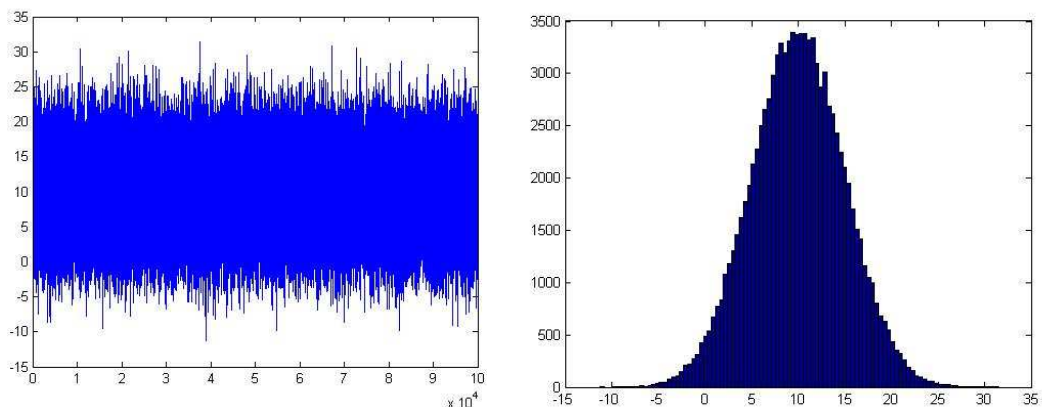
clc;
clear all;
close all;
% Vygenerování populace s normálním rozdělením.
Po=normrnd(10,5,[1,100000]);
SampleSize=10;
SampleNumber=10000;
% vykreslení Po a histogramu Po
figure
plot(Po);
figure
hist(Po, 100);
% získání výběrů, každý o rozsahu 'SampleSize'
Sa=[];
for i=1:SampleNumber
    for j=1:SampleSize
        index=round(abs(randn(1)*(length(Po)/10-1)))+1;
        Sa(i,j)=Po(index);
    end
end
end

```

```

% Výpočet 95% intervalu spolehlivosti každého výběru
mean=[];
interval=[];
for i=1:SampleNumber
    mean(i)=sum(Sa(i,:))/SampleSize;
    interval(i,1)=mean(i)-(1.96*5/sqrt(SampleSize));
    interval(i,2)=mean(i)+(1.96*5/sqrt(SampleSize));
end
% počet vzorků, které obsahují střední hodnotu = 10
count=0;
for i=1:SampleNumber
    if interval(i,1)>=10 || interval(i,2)<=10
        count=count+1;
    end
end
count/SampleNumber

```



Obr. 6.12. Demonstrace principu intervalu spolehlivosti pro střední hodnotu

Zvolíme 1000 vzorků, z nichž každý o velikosti 10, z $N(10; 5)$ rozdělení. Vypočítáme střední hodnotu a 95% interval spolehlivosti u každého vzorku. Spočítáme počet těchto intervalů spolehlivosti, které skutečně obsahují střední hodnotu rovnu 10. Výsledky se pohybují kolem hodnoty 0,05, což znamená, že tyto intervaly spolehlivosti mají 5% šanci, že nebudou obsahovat skutečnou střední hodnotu 10. To je důvod, proč jsou nazývány 95% intervaly spolehlivosti.

6.6.1 Interval spolehlivosti pro střední hodnotu a rozptyl

```

% Interval spolehlivosti pro středni hodnotu a rozptyl
function [xpruh, smodch,miconf, sigmconf] = ISproXpruh(x,alpha)
% intspol vypocet parametru a intervalu spolehlivosti pro data ve sloupci

```

```

% x.
if(nargin<2)|(isempty(alpha)),alpha=0.05; elseif
not((0<alpha)&(alpha<1)), alpha=.05;
end
[rows, cols] = size(x);if rows == 1, x = x(:);rows=cols;cols=1;end
xpruh=mean(x);smodch=std(x);if nargin==2, return; end
miconf=zeros(1,2);thor=tinv(1-alpha/2,rows-1);tdol=tinv(alpha/2,rows-1);
miconf=[(xpruh+tdol*smodch/sqrt(rows)),(xpruh+thor*smodch/sqrt(rows))];
if nargin<4, return; end
sigmconf=zeros(1,2);chid=chi2inv(1-alpha/2,rows-
1);chih=chi2inv(alpha/2,rows-1);
sigmconf=[(smodch*sqrt((rows-1)./chid)),(smodch*sqrt(rows-1./chih))];

```

Po zadání:

```
x = normrnd(10,3,20,1);
```

```
>> [pru smer inrpru intsigma] = ISproXpruh(x)
```

```
pru = 10.1446
```

```
smer = 2.3504
```

```
inrpru = 9.0446 11.2446
```

```
intsigma = 1.7874 10.4817
```

6.6.2 Odhad intervalu spolehlivosti metodou bootstrap

Metoda bootstrap bude blíže teoreticky popsána v kapitole 6.8. Následující program napsaný v jazyce MATLAB počítá interval spolehlivosti střední hodnoty z předpokladu normality, Studentizace a percentilové metody.

```

% Bootstrap odhad stredni hodnoty jednor. vyberu
clc;clear all;
kkk = menu('vstup dat','ze souboru','pokus');
if kkk==2
    ar=[0.0090, 0.0090, 0.0090, 0.0090, 0.0180, 0.0320, 0.0120, 0.0150,
0.0090, 0.0780, 0.0920, 0.0230, 0.0180, 0.0100]';
    % ar=not(le(ar,1)).*ar;
else
    s1=input('navez souboru:','s'); % jmeno s cislem
    s2=input('pripona:','s'); %bez tecky
    s2=strcat('.',s2);
    s5=num2str(i);
    nam=s1;

```

```

    nam1=strcat(nam,s2);
    load(nam1);
    ar=eval(nam);
end
alfa=.05;nq=norminv(1-alfa/2,0,1);
[c s]=size(ar);
if c==1
    ar=ar';c=s;
end
lidet=0.001;
s=(ar>lidet);ac=ar.*s;
b=800;mi=mean(ac);sig=var(ac);
B=ac(ceil(c*rand(c,b)));
pi=mean(B);ps=mean(pi);ss=var(pi);Pi=sort(pi);sic=var(B);
ti=(pi-ps)./sqrt(sic);Ti=sort(ti);
di=round(alfa*(b+1)/2);
hi=round((1-alfa/2)*(b+1));
pid=Pi(di);pih=Pi(hi);tid=Ti(di);tih=Ti(hi);
smezi=mi-nq*sqrt(sig/c);
hmezi=mi+nq*sqrt(sig/c);
smez=ps-nq*sqrt(ss); % /sqrt(b));
hmez=ps+nq*sqrt(ss);
fprintf('Klasicka normalita \n');
fprintf('95 proc. Konf. interval x %g. %g.\n',hmezi,smezi);
fprintf('Prumer = %g.\n',mi);
fprintf('Rozptyl = %g.\n',sig);
fprintf('Bootstrap normalita \n');
fprintf('95 proc. konf. interval %g. %g.\n',hmez,smez);
fprintf('PrumerBootstrap = %g.\n',ps);
fprintf('RozptylBootstrap = %g.\n',ss);
fprintf('Bootstrap pivot \n');
fprintf('95 proc. Konf. interval %g. %g.\n',pih,pid);
fprintf('Bootstrap Student \n');
dms = ps+tid*sqrt(ss);hms=ps+tih*sqrt(ss);
fprintf('95 proc. Konf. interval %g. %g.\n',hms,dms);
subplot(1,2,1);hist(pi);title('Boot pivot');
subplot(1,2,2);hist(ti);title('Boot Student');

```

Jako příklad byla vybrána ukázka naměřených hodnot monitorovacího přístroje obsahu škodlivých látek v ovzduší. Limita detekce je nastavena na hodnotě 0,009 a k nahrazení hodnot pod limitou detekce nulou můžeme využít příkaz $ar=not(le(ar,0.009)).*ar$. Účelem

bylo stanovit 95 procentní interval spolehlivosti střední hodnoty. Výstup z programu je následující:

Klasická normalita

95 proc. Konf. Interval: UC = 0.0364361. LC = 0.00613531.

Prumer = 0.0212857.

Rozptyl = 0.000836527.

Bootstrap normalita

95 proc. konf. Interval: UC = 0.0355396. LC = 0.00619627.

PrumerBootstrap = 0.0208679.

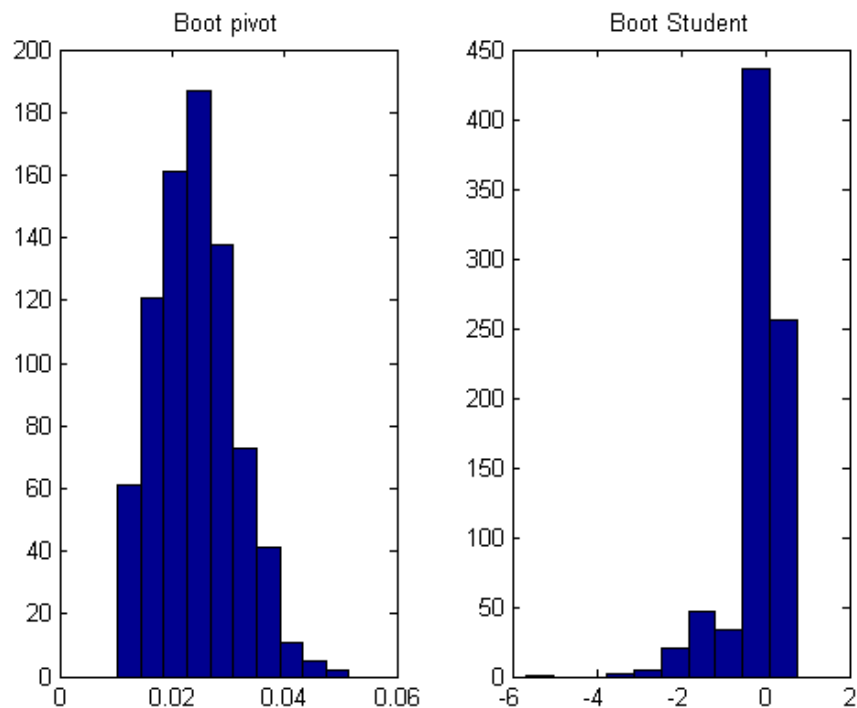
RozptylBootstrap = 5.60355e-005.

Bootstrap pivot

95 proc. Konf. Interval: UC = 0.0362857. LC = 0.00835714.

Bootstrap Student

95 proc. Konf. Interval: UC = 0.0240643. LC = 0.0111768.



Obr. 6.13. Rozdělení veličin p_i a t_i (nahrazení podlimitních hodnot nulou)

Na předchozím obrázku je uvedeno rozdělení veličin p_i a t_i . Jsou vidět odchylky od normálního rozdělení. Je patrné, že Studentizovaný Bootstrap poskytuje výrazně nižší horní mez UC než ostatní metody a nahrazení podlimitních hodnot nulou má za důsledek snížení všech horních mezí. Právě Studentizovaný Bootstrap je často považován za výhodný a doporučován pro komplexnější rozdělení dat.

6.7 Transformace zlepšující rozdělení dat

S transformací dat se při zpracování experimentů sekáváme velmi často. Podle příčin můžeme transformaci dělit do dvou základních skupin:

- A. Transformace zlepšující rozdělení dat. Zde je transformace žádána a přispívá ke zlepšení rozdělení dat (zjednodušuje jejich zpracování)
- B. Transformace jako důsledek matematických operací (obvykle realizace funkcí) s měřenými veličinami. To je případ, kdy známe u komplikovaných systémů vstupní náhodné veličiny, a zajímá nás výstupní náhodná veličina. Patří sem tedy všechny transformace, kdy na základě experimentálních výsledků počítáme jiné veličiny (např. z hodnot poloměru plochu kruhových elementů). Zde je vlastně transformace nežádaná, protože deformuje původní rozdělení dat.

V případě ad A) se hledá vhodná transformace. V případě ad B) se hledají vhodné postupy zpracování dat, které omezují vliv transformace. Tato dualita způsobuje, že oblasti transformace se v literatuře nevěnuje patřičná pozornost. Nadto vede ke stavu, kdy formálně shodné (matematicky správné) metody poskytují značně odlišné výsledky. Z uvedeného je zřejmé, že transformace může být buď "užitečným nástrojem", nebo "základní překážkou" při statistické analýze dat. Pro statistickou analýzu dat je ideální, pokud jsou prvky výběru náhodné vzájemně nezávislé veličiny se stejným normálním rozdělením. Reálné výběry se od tohoto stavu více či méně odlišují. V jednodušším případě mají delší konce (vyšší špičatost), než odpovídá normálnímu rozdělení. To je často důsledek přítomnosti vybočujících měření. Zde je při statistické analýze stále střed symetrie v místě módu, který je totožný s mediánem a střední hodnotou. Efektivní odhad polohy je medián (průměr \bar{x} má přibližně dvojnásobný rozptyl). Běžné statistické testy jsou vůči vyšší špičatosti dat poměrně robustní (to se týká zejména t -testu významnosti). Také valná většina robustních metod odhadu parametrů polohy a rozptýlení vychází z představy symetrického rozdělení dat, kontaminovaného jistým

podílem vybočujících dat. Komplikovanější je případ, kdy je rozdělení výběru zešikmené (obyčejně k vyšším hodnotám). Pak již není modus totožný s mediánem ani střední hodnotou a vlastní interpretace parametru polohy je ztížena. Efektivní odhad parametru polohy je možný jen při znalosti zákona rozdělení pravděpodobnosti (který však při analýze dat není přesně apriorně znám). Běžné statistické testy jsou vůči zešikmenému rozdělení dat obecně nerobustní. Také základní robustní metody odhadu parametrů polohy a rozptýlení zde nefungují dobře. Je tedy zřejmé, že již symetrizační transformace bude pro analýzu dat velmi užitečná. Původním zjevem u řady "nenormálně" rozdělených výběrů je nekonstantnost rozptylu (pouze pro normální rozdělení platí, že střední hodnota je nezávislá na rozptylu). Transformace stabilizující rozptyl je tedy zároveň transformací vedoucí k normalitě. [9], [10]

6.7.1 Zpracování transformovaných dat

Pokud vedle transformace dat k přibližné normalitě, lze pro veličiny $y = h(x)$ určit průměry x_p , rozptyl s_y^2 , konfidenční interval střední hodnoty $y_p \pm t_{1-\alpha/2} \cdot s_y / \sqrt{N}$ a případně provádět testy významnosti. V řadě případů je dosaženo tímto postupem adekvátních výsledků (i když je lépe použít t -testů vycházejících z \bar{d} -uřezaného průměru. I přes některé teoretické problémy, lze tedy v korektní transformaci provádět základní statistickou analýzu dat velmi snadno. Problém však je, že je často požadováno určit jak statistické charakteristiky, tak i konfidenční intervaly v původních proměnných. Při znalosti parametru transformaci λ lze vyčíslit střední hodnotu $E(x)$ původních dat jako nelineární funkci střední hodnoty μ_T a rozptylu σ_T^2 v transformaci.

$$E(x) = \int_{-\lambda/2}^{\infty} \frac{1}{\sigma} \sqrt{1 + \lambda y} * f_n \left(\frac{y - \mu_T}{\sigma_T} \right) dy \quad (6.19)$$

Zde f_n je hustota pravděpodobnosti normovaného normálního rozdělení. Pro $\lambda = 0$ vyjde po dosazení do rovnice pro $E(x)$ a integraci, že

$$E(x) = \exp(\mu_T + 0.5\sigma_T^2) \quad (6.20)$$

a pro $\lambda = 0.5$ je

$$E(x) = [0.5\mu_T + 1]^2 + 0.5\sigma_T^2 \quad (6.21)$$

Přesnější aproximace $E(x)$ pro logaritmickou transformaci má tvar

$$E(x) = \exp(\mu_T + 0.5\sigma_T^2) * \left[1 - \frac{\sigma_T^2(\sigma_T^2 + 2)}{4N} + \frac{\sigma_T^4(3\sigma_T^4 + 44\sigma_T^2 + 84)}{96N^2} \right] \quad (6.22)$$

Pro určení intervalu spolehlivosti lze využít asymptotické normality střední hodnoty v transformaci. Výsledný interval má tvar

$$h^{-1}(\mu_T - t_{1-\alpha/2}(N-1)\sigma_T/\sqrt{N}) \leq \mu \leq h^{-1}(\mu_T + t_{1-\alpha/2}(N-1)\sigma_T/\sqrt{N}) \quad (6.23)$$

Tento interval však již nemusí obsahovat uprostřed parametr polohy. Z uvedeného je zřejmé, že zpětná transformace je dosti komplikovaný problém. Většina odhadů střední hodnoty je vychýlená a mají také větší rozptyly. Proto je vždy výhodné pracovat jen s transformovanými hodnotami (pokud není nezbytně nutné znát charakteristiky původních veličin). Tento přístup vyhovuje zejména při realizaci testů významnosti, kde může být celá analýza v transformaci. Také při pravděpodobnostních úvahách lze pracovat pouze v transformaci. [9], [10]

6.7.2 Box–Coxova mocinná transformace

Nevýhody prosté mocinné transformace (zejména nespojitost v okolí nuly a nesrovnatelnost měřítek v transformaci) odstraňuje rodina Box–Coxových transformací $X^{(\lambda)}$, která je lineární transformací prosté mocinné transformace $X_p^{(\lambda)}$. Box–Coxova třída polynomických transformací má tvar

$$X^{(\lambda)} = \begin{cases} \frac{X^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln X & \text{for } \lambda = 0 \end{cases} \quad (6.24)$$

Pro posouzení kvality transformace, resp. nalezení optimálního λ je také možno použít grafu rozptýlení s kvantily (GRK), resp. kvantilových grafů (Q-Q grafů). Výhodnější je použití testů normality dat po mocinné transformaci. Známý Shapiro-Wilkův test je úměrný testu významnosti směrnice v Q-Q grafu, takže lze také posuzovat linearitu v Q-Q grafech. Tato rodina Box–Coxových transformací závisí na jediném parametru λ , který může být odhadnut metodou maximální věrohodnosti (MLE) nebo metodou nejmenších čtverců (LSE). Pro $\lambda = 1$ resultuje aditivní model měření a pro $\lambda = 0$ model multiplikativní. Nejprve se ze zvolené oblasti volí hodnota λ . Pro zvolené λ vypočítáme

$$L_{\max} = -\frac{1}{2} \ln \hat{\sigma}^2 + \ln J(\lambda, X) = -\frac{1}{2} \ln \hat{\sigma}^2 + (\lambda - 1) \sum_{i=1}^n \ln X_i, \quad (6.25)$$

kde

$$J(\lambda, X) = \prod_{i=1}^n \frac{\partial W_i}{\partial X_i} = \prod_{i=1}^n X_i^{\lambda-1}, \quad \text{pro všechna } \lambda, \text{ tak, aby } \ln J(\lambda, X) = (\lambda - 1) \sum_{i=1}^n \ln X_i.$$

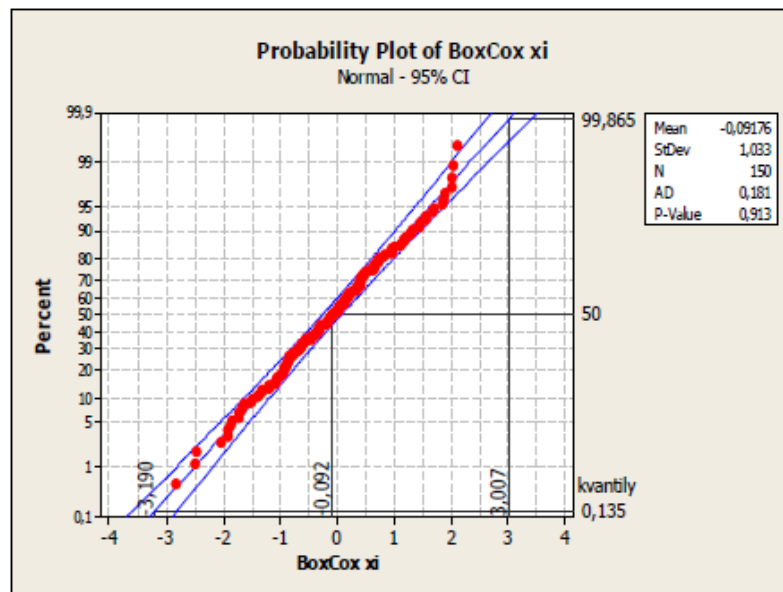
Odhad $\hat{\sigma}^2$ pro pevné λ je $\hat{\sigma}^2 = S(\lambda)/n$, kde $S(\lambda)$ je reziduální součet čtverců v analýze rozptylu $X(\lambda)$. Po vyčíslení $L_{\max}(\lambda)$ pro několik hodnot λ v rozsahu, mohou být hodnoty $L_{\max}(\lambda)$ vyneseny proti λ . Maximálně věrohodný odhad λ je získán z hodnoty λ , která maximalizuje funkci $L_{\max}(\lambda)$. S optimální hodnotou λ^* , každá X hodnota specifikačních mezí je transformována na normální veličinu pomocí rovnice (6.24). Box-Coxova transformace odhaduje hodnotu λ , která minimalizuje směrodatnou odchylku normalizované transformované proměnné. Software prozkoumá mnoho transformací, z nichž v následující tabulce jsou uvedeny jen ty, které jsou pro zaokrouhlené hodnoty λ . y je transformovaná hodnota proměnné (naměřeného znaku) x . [9], [10]

Tab. 6.1. Hodnoty λ a k nim transformované proměnné x [2]

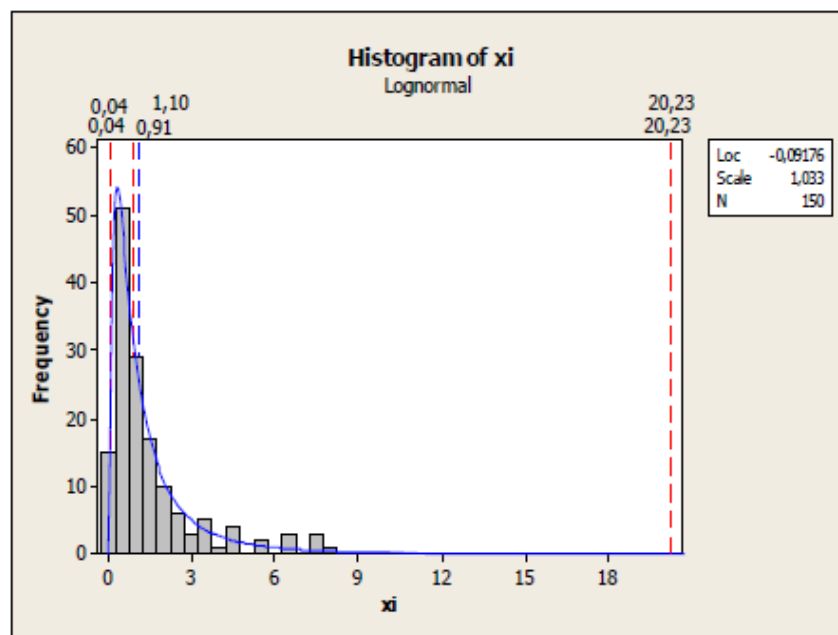
Hodnota λ	Transformace
$\lambda = 2$	$y = x^2$
$\lambda = 0.5$	$y = \sqrt{x}$
$\lambda = 0$	$y = \ln x$
$\lambda = -0.5$	$y = 1/(\sqrt{x})$
$\lambda = -1$	$y = 1/x$

Na následujícím obrázku je uveden příklad s použitím Box-Coxovy transformace provedené pomocí softwaru Minitab 14. Tato transformace se snaží převést data na nová data, která se již dají popsat normálním rozdělením. Problémem může být ten fakt, obdobně jako u Johnsonovy transformace, že sice vhodná transformace byla nalezena, ale nelze transformovat specifikační meze, protože jsou mimo definiční obor transformace. Software Minitab 14 postupuje tím způsobem, že hledá nejvhodnější transformaci volbou parametru λ v rozmezí od -5 do 5. Vybere se ta hodnota parametru λ , které dává největší

p -hodnotu pro test dobré shody. Pro zobrazení výsledku v řeči původních netransformovaných dat, je nutno použít zpětnou Box- Coxovu transformaci.



Obr. 6.14. Odhad kvantilů transformovaných dat vypočítaných po vložení příslušných procent do „Add - Percentil Lines“ na pravděpodobnostním grafu



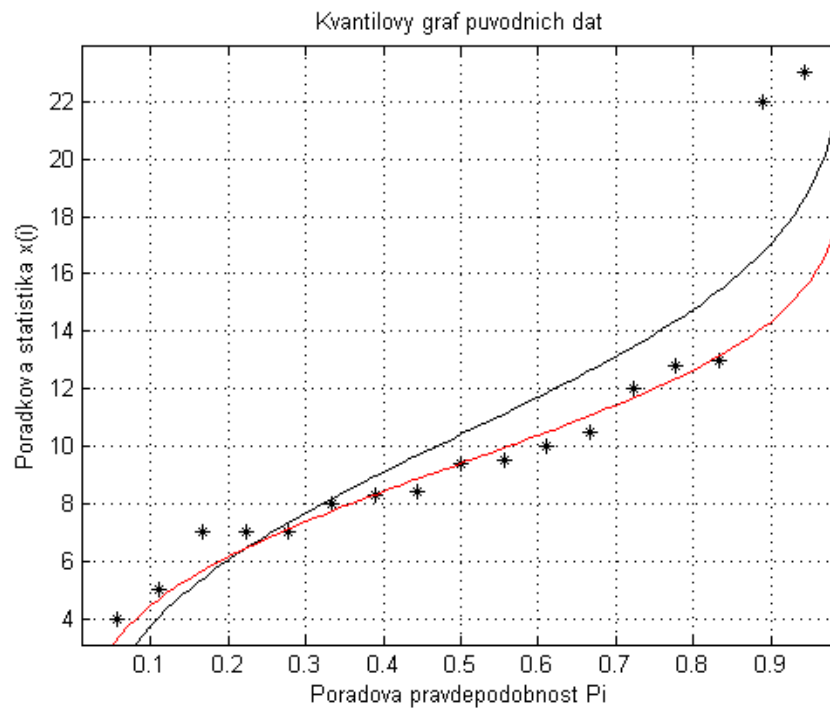
Obr. 6.15. Kvantily vypočítané z modelu lognormálního rozdělení (červené 0,04; 0,91; 20,23) a na základě zpětné BoxCox transformace (modré 0,04; 1,10; 20,23)

[2]

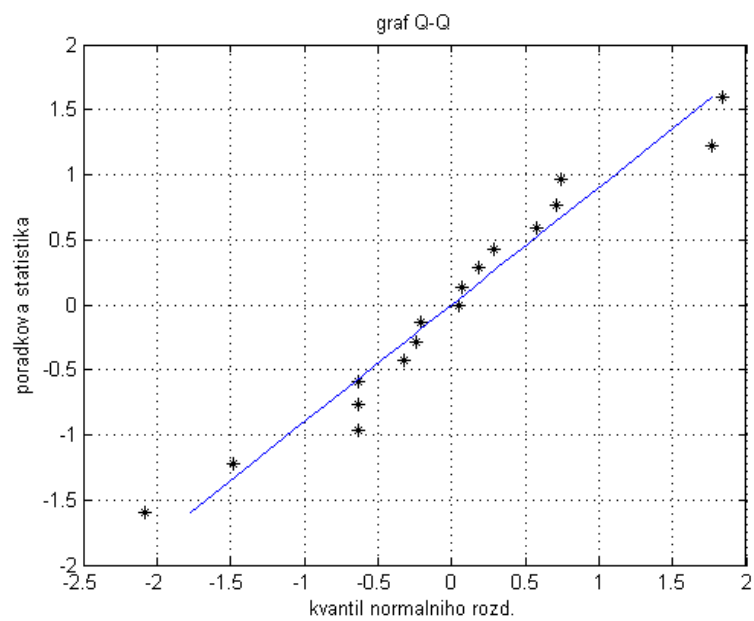
Program nazvaný *BCTransform.m*, jehož zdrojový kód je uveden v příloze P IX, umožňuje odhad parametru mocninné transformace metodou maximální věrohodnosti a maxima

korelace v Q-Q grafu. Je zde provedena retransformace dat a určen interval spolehlivosti střední hodnoty.

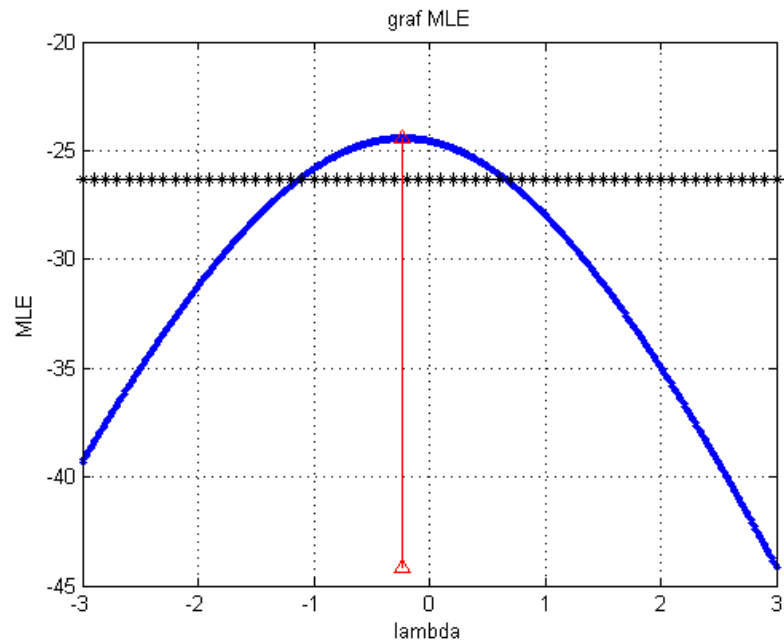
Výstup z Matlabu je následující:



Obr. 6.16. Kvantilový graf původních dat



Obr. 6.17. Q-Q graf transformovaných dat

Obr. 6.18. Odhad parametru λ pomocí metody maximální věrohodnosti

*****.

Puvodni data.

Prumer arit.=10.4059.

Prumer geom.=9.42055.

Median=9.4.

Rozptyl=26.8343.

Sikmost=1.27746.

Spicatost=3.78427.

Odhad lambda prec.=0.571042.

Odhad lambda=0.866339.

Odhad lambda rough=0.517177.

*****.

Optim lambda Box Cox-MLE -0.23.

Konf. interval-1.13.0.67.

Transformace.

Prumer=1.73925.

Rozptyl=0.0711426.

Sikmost=0.000514101.

Spicatost=2.70961.

Re transformace.

Prumer=10.4663.

Rozptyl=22.952.

Dmez=8.32249.

Hmez=13.3313.

*****.

Optim lambda Box Cox-SW -0.25.

smernice 1.11064.

Re transformace.

Prumer=11.3528.

Rozptyl=30.895.

Dmez=8.89367.

Hmez=14.7238.

.....

Původní kvantilový graf ukazuje na asymetrické rozdělení. Pomocí Box-Coxovy transformace došlo ke stabilizaci rozptylu a následné transformaci vedoucí k normalitě, jak ukazuje Q-Q graf, avšak optimální mocnina vyšla -0,23 s mezemi (-1.13, 0.67), kdy tento interval obsahuje nulu, proto lze provádět další analýzu v logaritmické transformaci resp. volit multiplikační model měření.

6.8 Metoda Bootstrap

Zvláštnosti dat v technických vědách se projevují na asymetrii výběrového rozdělení. Ta pak omezuje použití různých technik založených na průzkumové analýze a identifikaci vybočujících měření. Také robustní techniky obyčejně selhávají, protože eliminují extrémny, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat. Je známo, že pro konstrukci intervalu spolehlivosti populačního parametru p_s je třeba znát rozdělení $g(p)$ jeho odhadu p . Pro některá rozdělení (např. normální) a parametry (střední hodnota, rozptyl) jsou rozdělení odhadů nebo jejich funkcí známy a intervaly spolehlivosti je možné konstruovat relativně snadno. Pro odhad intervalu střední hodnoty z aritmetického průměru x_A a výběrového rozptylu s^2 není normalita tak striktní požadavek. Je známo, že pokud

zpracovávaný výběr velikosti N prochází z ne-normálního rozdělení se střední hodnotou μ a rozptylem σ^2 má tzv. Studentova náhodná veličina

$$t = \sqrt{N} * (x_A - \mu) / s \quad (6.26)$$

asymptoticky Studentovo rozdělení s $(N-1)$ stupni volnosti. Asymptotické Studentovo rozdělení veličiny t umožňuje konstrukci intervalu spolehlivosti střední hodnoty μ . Při tzv. frekventistickém přístupu je $100(1-\alpha)\%$ na interval spolehlivosti CI definován vztahem

$$P(LC \leq \mu \leq UC) = 1 - \alpha \quad (6.27)$$

Symbol P označuje pravděpodobnost a α je tzv. hladina významnosti. Obvykle se volí $\alpha = 0.05$ nebo $\alpha = 0.01$ s tím, že čím je α menší, tím je interval (LC, UC) širší. Pokud není σ^2 známo lze použít vztah

$$x_A - t_{1-\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \leq \mu \leq x_A - t_{\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \quad (6.28)$$

kde $t_{1-\alpha/2}(N-1) = -t_{\alpha/2}(N-1)$ jsou kvantily Studentova rozdělení s $N-1$ stupni volnosti. Pro případ normálního rozdělení má interval spolehlivosti pro střední hodnotu přesně $100(1-\alpha)\%$ -ní pokrytí střední hodnoty. To znamená, že jen v $100\alpha/2\%$ případů je *střední hodnota menší než CI* (nejistota NP zprava) a v $100\alpha/2\%$ případů je *větší než CI* (nejistota NL zleva). Pro případ nenormálního rozdělení platí tyto intervaly pouze asymptoticky tedy pro dostatečně vysoké N . Dostatečná velikost N závisí silně na šikmosti $g_1(x)$ rozdělení z kterého data pocházejí. Pro neznámé rozdělení výběru $x = (x_1 \dots x_N)$ a libovolný parametr ps lze s výhodou použít techniku Bootstrap, která umožňuje jak nalezení rozdělení výběrové statistiky p , tak i konstrukci intervalu spolehlivosti. Spočívá v generaci M -tice simulovaných výběrů $v_1 \dots v_M$ označovaných jako Bootstrap výběry. Jejich rozdělení odpovídá rozdělení původního výběru x , charakterizovaného hustotou pravděpodobností $g(x)$. Z těchto výběrů se určí M -tice odhadů $p_i = p(x)$ hledaného parametru ps . Z této M -tice hodnot lze počítat intervaly spolehlivosti pomocí celé řady metod. [9], [10]

6.8.1 Odhad z asymptotické normality

Jde o nejjednodušší postup založený na představě, že M je dostatečně veliké a $p_i, i=1..N$ lze zpracovat jako výběr z normálního rozdělení. Pro tzv. Bootstrap odhad střední hodnoty parametru p_B platí

$$p_B = \frac{1}{M} \sum_{i=1}^M p_i \quad (6.29)$$

a odpovídající rozptyl má tvar

$$s_B^2 = \frac{1}{M} \sum_{i=1}^M (p_i - p_B)^2 \quad (6.30)$$

Pro $100(1-\alpha)\%$ ní interval spolehlivosti parametru ps se pak použije známý vztah

$$p_B - u_{1-\alpha/2} * s_B \leq ps \leq p_B + u_{1-\alpha/2} * s_B \quad (6.31)$$

kde $u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení. [9], [10]

6.8.2 Percentilový odhad

Tento postup je založen na neparametrickém odhadu mezí intervalu spolehlivosti vycházejícím z pořádkových statistik $p_i \leq p_{(i+1)}$ jsou pořádkové statistiky, pro které platí, že jsou $d\%$ -ním kvantilem rozdělení odhadu p pro

$$d = \frac{i}{M+1} \quad (6.32)$$

Dolní mez $100(1-\alpha)\%$ ní intervalu spolehlivosti je pak

$$LC = p_{(k1)} \text{ kde } k1 = \text{int}[\alpha * (M+1)/2] \quad (6.33)$$

a pro horní mez platí

$$UC = p_{(k2)} \text{ kde } k2 = \text{int}[(1-\alpha/2) * (M+1)] \quad (6.34)$$

Zde $\text{int}(x)$ je celá část čísla x . [9], [10]

6.8.3 Studentizovaný odhad

Tento odhad vychází z jednoduché transformace vedoucí na Studentizovanou náhodnou veličinu t_i

$$t_i = \frac{P_i - P_B}{s_{Bi}} \quad (6.35)$$

kde s_{Bi} je výběrová směrodatná odchylka počítaná pro i – tý Bootstrap výběr v_i . Pro $100(1-\alpha)\%$ ní interval spolehlivosti pak platí

$$p_B - t_D * s_B \leq p_S \leq p_B + t_D * s_B \quad (6.36)$$

kde pořádková statistika $t_D = t_{(\text{int}[\alpha*(M+1)/2])}$ a pořádková statistika $t_H = t_{(\text{int}[(1-\alpha/2)*(M+1)])}$

[9], [10]

6.8.4 Vyhlazený odhad

Obecně lze na základě hodnot p_i sestavit odhad hustoty pravděpodobnosti jejich rozdělení $fe(p)$ např. s využitím histogramu nebo jádrového odhadu. Pro meze tohoto intervalu pak platí, že

$$\alpha/2 = \int_{-\infty}^{LC} fe(p) dp \quad (6.37)$$

a

$$\alpha/2 = \int_{UC}^{\infty} fe(p) dp \quad (6.38)$$

Podle typu odhadu fe může jít o úlohu numerické nebo analytické integrace. [9], [10]

6.8.5 Generace Bootstrap výběrů

Základním předpokladem úspěšnosti celého postupu je generace Bootstrap výběrů. Pro tento účel je třeba buď znát, nebo volit rozdělení $g(x)$. Standardní technika neparametrického Bootstrap vychází z neparametrického odhadu $g(x)$ ve tvaru

$$g(x) = \frac{1}{N} \delta(x - x_i) \quad (6.39)$$

kde Diracova funkce $\delta(x-x_i)=1$ pro $(x-x_i)$ a všude jinde je $\delta(x-x_i)=0$. Toto rozdělení pokládá pravděpodobnost $1/N$ v každém bodě. Simulované výběry se pak realizují jako náhodné výběry složené z prvků původního výběru x s vracením (tj. jeden prvek původního výběru se může v simulovaném výběru vyskytnout i opakovaně). Další možností je konstruovat vhodný parametrický model $g(x)$, odhadnout jeho parametry a generovat simulované výběry standardními postupy. Tento přístup naráží na celou řadu problémů souvisejících s možnou nehomogenitou, vybočujícími body, heteroskedasticitou a autokorelací. Bootstrap metody obecně poskytují informace o bodových odhadech, tak i intervalech spolehlivosti. Uvažujme standardní neparametrický Bootstrap (v_i jsou výběry s vracením) pro $p_s = \mu$, tj. jde o střední hodnotu a její interval spolehlivosti střední hodnoty. Lze snad určit, že v tomto případě je Bootstrap průměr totožný s aritmetickým průměrem původních dat a Bootstrap rozptyl je M -krát menší než rozptyl původních dat. Liší se však intervaly spolehlivosti zejména tam, kde se rozdělení dat výrazně odchyluje od normálního rozdělení. Kromě standardního Bootstrap lze použít také dvojitý Bootstrap (Bootstrap aplikovaný na výběry v_i), blokovaný Bootstrap (realizace výběru s vracením na bloky homogenních dat a sestavení celkového Bootstrap výběru spojením výsledků). [9], [10]

6.8.6 Realizace postupu Bootstrap

Z hlediska realizace metody Bootstrap na počítači je základem generace simulovaných výběrů. Velmi jednoduše se dá tato operace provést v jazyku MATLAB s využitím vektorového triku. Úsek programu má tvar

```
ard=load('conc.txt');[c s]=size(ar);b=800;
```

```
if c==1
```

```
ar=ar';c=s;
```

```
end
```

```
B=ar(ceil(c*rand(c,b)));
```

Předpokládá se, že n -tice dat je v souboru *conc.txt* a b – tice Bootstrap je v poli B . Pro výpočet odhadu p_i se používá standardních postupů. Výpočet intervalů spolehlivosti je pak závislý na volbě přístupu.

6.9 Zpracování výběrů z asymetrických rozdělení

6.9.1 Omezení asymetrie rozdělení Studentovy statistiky

Asymetrie rozdělení t statistiky je zřejmé z Edgeworthova rozvoje definovaného rovnicí (první člen Edgeworthova rozvoje)

$$P(t \leq x) = F_n(x) + \frac{g_1(x) * (2x^2 + 1)}{6\sqrt{N}} f_n(x). \quad (6.40)$$

Johnson navrhl nahradit čítecí rovnice

$$t = \sqrt{N} * (x_A - \mu) / s \quad (6.41)$$

několika členy inverzního Cornish Fisherova rozvoje.

$$t_J = \sqrt{N} * \left[(x_A - \mu) + \frac{g_1(x) * s}{6N} + \frac{g_1(x)}{3s} (x_A - \mu)^2 \right] / s \quad (6.42)$$

Pro tuto transformaci již přibližně platí, že

$$P(t_J \leq x) \approx F_n(x) \quad (6.43)$$

Johnsonova transformace t statistiky však není obecně ani monotónní ani v neupravené formě invertovatelná. Tyto problémy eliminují transformace navržené Hallem.

$$t_H = K + \frac{g_1(x) * K^2}{3} + \frac{g_1(x)^2 * K^3}{27} + \frac{g_1(x)}{6N} \quad (6.44)$$

resp.

$$t_{H1} = \frac{g_1(x)}{6N} + \frac{3 * \sqrt{N} * \exp\left(\frac{2 * K * g_1(x)}{3\sqrt{N}}\right)}{2 * g_1(x)}, \quad (6.45)$$

kde

$$K = \frac{x_A - \mu}{s} \quad (6.46)$$

Obě tyto transformace násobené faktorem $N^{-0.5}$ splňující rov (6.43) tj. vedou k přibližné normalitě (redukci šikmosti) a jsou invertovatelné. Inverzní forma statistiky t_H se zahrnutou násobivou konstantou má tvar

$$t_H^{-1}(y) = \frac{3 * \sqrt{N}}{g_1(x)} \left[\left(1 + g_1(x) * \left(\frac{y}{\sqrt{N}} - \frac{g_1(x)}{6N} \right) \right)^{1/3} - 1 \right] \quad (6.47)$$

Při sledování úrovně škodlivin je prakticky zajímavý pouze pravostranný interval spolehlivosti (jednostranný interval spolehlivosti zprava tj. horní hranici střední hodnoty). Tento interval se často používá u rozdělení zešikmených vpravo k určení povolené horní hranice např. znečištění pro horní mez pravostranného intervalu spolehlivosti pak platí, že

$$\mu \leq x_A + t_H^{-1}(z_{1-\alpha}) * \frac{s}{\sqrt{N}} \quad (6.48)$$

Místo normovaného normálního kvantilu z se doporučuje použít odpovídajícího kvantilu určeného z Bootstrap výběrů. Místo transformace definované rov. (6.47) lze použít zjednodušenou verzi

$$t_a^{-1}(y) = y - \frac{g_1(x) * (y^2 / 3 + 1/6)}{\sqrt{N}} \quad (6.49)$$

Tato transformace se pak dosadí do rov (6.48). Opět je možno použít Bootstrap kvantilů. Jak je patrné znalost šikmosti výběrového rozdělení je zde nezbytnou podmínkou pro použití korekcí. V práci (Boos D. D a Hughes-Oliver J. M.) byl na rozsáhlém simulačním experiment určen vztah mezi nejistotou pokrytí zleva, zprava a z obou stran. Nejistota pokrytí zprava NP vyjadřuje pravděpodobnost, že skutečná střední hodnota je nižší než meze intervalu spolehlivosti. Pro nejistotu pokrytí zleva NL se určuje pravděpodobnost, že skutečná střední hodnota je vyšší než meze intervalu spolehlivosti. Nejistota pokrytí z obou stran NC je pak sjednocení obou chyb pokrytí, tj. $NC = NP + NL$. Pro širokou třídu rozdělení bylo nalezeno, že

$$PR = \alpha / 2 + [-0.73 + 0.71 * \exp(-\alpha / 2)] * g_1 / \sqrt{N} \quad (6.50)$$

a

$$PL = \alpha / 2 + [0.19 + 0.026 * \ln(\alpha / 2)] * g_1 / \sqrt{N} \quad (6.51)$$

Z těchto rovnic se dá např. určit potřebná velikost výběru, aby byla zachována nejistota pokrytí jako rozdíl mezi požadovanou pravděpodobností pokrytí (např. 0.95) a dosaženou pravděpodobností pokrytí (např. 0.94). Další možnost použití výše uvedených vztahů je

fixovat nejistotu pokrytí na zvolené hodnotě a pro známé N i $g_{1(x)}$ nalézt pravděpodobnost α^* pro výpočet kvantilu Studentova rozdělení. Takto opravené kvantily se pak dosadí do rovnice pro parametrický intervalový odhad střední hodnoty při neznámé směrodatné odchylce. Klasický pravostranný interval spolehlivosti má tvar

$$\mu \leq x_A + t_{1-\alpha} (N-1) \frac{s}{\sqrt{N}} \quad (6.52)$$

Po dosazení do rov (6.51) za $PL = 0.05$ rezultuje výraz

$$0 = \alpha^* + \left[0.19 + 0.026 * \ln(\alpha^*) \right] g_1(x) / \sqrt{N} - 0.05 = f(\alpha^*) \quad (6.53)$$

Kořenem funkce $f(\alpha^*)$ je pak α^* , pro které se spočítá opravený kvantil Studentova rozdělení tj. hodnota $t_{1-\alpha^*} (N-1)$. [9], [10]

6.9.2 Výpočet korigovaného průměru

Jednoduchá možnost jak počítat korigovaný průměr pro stanovení intervalu spolehlivosti u asymetrických rozdělení je založena na Johnsonově transformaci. Opravený průměr x_0 má tvar

$$x_0 = \left(x_A + \frac{s^* g_1}{6N} \right) \quad (6.54)$$

Je patrné, že velikost korekce opět souvisí se šikmostí a počtem měření. Na rozdíl od předchozího postupu se však mění poloha centra. Další možností je použití odhadů minimalizujících penále za přecenění resp. nedocenení odhadu střední hodnoty. Chenová zavedla tzv. MCE odhad x_{MCE} ve tvaru

$$x_{MCE} = x_A + d * s, \quad (6.55)$$

kde d se počítá podle vztahu

$$d = 0.5 * \left[b - \frac{2\sqrt{N}}{g_1(x)} + \sqrt{4 - \frac{b^2}{3} + \frac{4 * N}{g_1(x)} + \frac{8 * \log(a) * \sqrt{N}}{b * g_1(x)}} \right] \quad (6.56)$$

Volba a a b souvisí se zvoleným penálem. Doporučuje se $a = 1$ a $b = 2$, i když na základě simulací vychází spíše $a = 10$ a $b = 3$. Zajímavé je použití koncepce vycházející

z kompromisu mezi vychýlením odhadu a pravděpodobností, že bude ležet nad střední hodnotou. Na tomto základě byl navržen penalizovaný průměr x_p , pro který platí, že

$$x_p = x_A + \frac{4.5 * s^2}{\sqrt{N}} f(x_A) [1 - F(x_A)] \quad (6.57)$$

Zde $f(x_A)$ resp. $F(x_A)$ jsou hodnoty hustoty pravděpodobnosti a distribuční funkce, které se nahrazují neparametrickými odhady. Pro určení $f(x_A)$ se doporučuje vztah

$$f(x_A) = \frac{\text{int}(\sqrt{N})}{2 * N * A(x_A)} \quad (6.58)$$

Zde $A(x_A)$ se bere jako k – tá nejmenší hodnota rozdílů $w_i = \text{abs}(x_i - x_A)$, kde $k = \text{int}(N^{0.5})$. Jde vlastně o k -tou pořádkovou statistiku. Hodnota distribuční funkce se počítá jako počet hodnot prvků výběru ležících pod x_A dělený N . Je možné použít i dalších neparametrických odhadů založených např. na pořádkových statistikách. Dalším zlepšením je použití upraveného výběru uvažujícího extrémů. V upraveném výběru se nejvyšší pořádková statistika $x_{(N)}$ nahrazuje hodnotou $x_A + 4.5 s$, pokud je větší. Tato modifikace se doporučuje pro silné zešikmená rozdělení, kde se vyskytují hodnoty, sice extrémně vysoké, ale patřící do výběru. Pro výpočet intervalů spolehlivosti z Hallovy transformace a korigovaného průměru byl sestaven následující program. [9], [10]

`% Zpracování výběrů z asymetrického rozdělení`

`clc;clear all;`

`pom=0;`

`pobr=1; kkk = menu('vstup dat','ze souboru','pokus');`

`if kkk==2`

`a=[1,2,3,5, 4, 1, 2, 3, 5, 4, 5, 8, 7, 4, 5, 2, 1, 4, 5, 8, 7, 4, 5,`
`2, 1, 4, 5, 2, 1, 4, 5,]';`

`% ar=not(le(ar,1)).*ar;`

`else`

`s1=input('navez souboru:','s'); % jmeno s cislem`

`s2=input('pripona:','s'); %bez tecky`

`s2=strcat('.',s2);`

`nam=s1;`

`nam1=strcat(nam,s2);`

`load(nam1);`

`a1=eval(nam);`

```

end
n=length(a);alpha=.05;za=1.98;
[rows,cols]=size(a);if rows==1,a=a(:);rows=cols;cols=1;end
xpruh=mean(a);smodch=std(a);sigm=smodch.^2;sigp=var(a);
si=mean((a-xpruh).^3);sik=si/(sigp*sqrt(sigp));
thor=tinv(1-alpha/2,rows-1);tdol=tinv(alpha/2,rows-1);
v1={(xpruh+tdol*smodch/sqrt(rows)) '<' xpruh '<'
(xpruh+thor*smodch/sqrt(rows))};
    %Hall zjednoduseny
    tdo=(3*sqrt(rows)/sik)*((1+sik*(za/sqrt(rows)-sik/(6*rows)))^(1/3)-
1);
    %tdo=za-sik*(za/3+1/6)/sqrt(rows);
    v2={(xpruh-tdo*smodch/sqrt(rows)) '<' xpruh '<'
(xpruh+tdo*smodch/sqrt(rows))};
    xop=xpruh+smodch*sik/(6*rows);
    v3={(xop+tdol*smodch/sqrt(rows)) '<' xop '<'
(xop+thor*smodch/sqrt(rows))};
    disp('intervaly spolehlivosti stredni hodnoty');disp('');
    disp('z predpokladu normality')
    v1
    disp('Hallova transformace')
    v2
    disp('opraveny prumer')
    v3

```

Výstup je následující:

intervaly spolehlivosti stredni hodnoty

z predpokladu normality

v1 = [3.0924] '<' [3.8387] '<' [4.5850]

Hallova transformace

v2 = [3.1408] '<' [3.8387] '<' [4.5366]

opraveny prumer

v3 = [3.0955] '<' [3.8418] '<' [4.5881]

ZÁVĚR

Cílem této bakalářské práce bylo popsat vybrané statistické testy včetně jejich použití v programovém prostředí Matlab a vytvořit funkční programy, které budou sloužit jako studijní pomůcka při výuce statistických předmětů.

Text práce začíná představením statistického toolboxu a úvodem do problematiky testování statistických hypotéz. Zbytek teoretické části je rozdělen na dvě hlavní kapitoly: parametrické a neparametrické testy. Použití jednotlivých testů bylo ilustrováno na příkladech s náhodně generovanými nebo reálnými daty. Jednotlivé příklady byly prováděny v programu MATLAB 7.9 (R2009b).

V praktické části této práce byl v kapitole 5 popsán princip testování statistických hypotéz pomocí metody Monte Carlo. V kapitole 6 byly představeny vybrané statistické programy a aplikace vytvořené v Matlabu, které mohou sloužit jako studijní pomůcka při výuce matematické statistiky a statistického řízení kvality. Tyto programy se týkaly aproximací rozdělení, demonstrace asymetrie dat, testů normality jednorozměrných a vícerozměrných datových souborů, intervalových odhadů střední hodnoty, jak klasickými metodami, tak i počítačově intenzivními metodami, jako je metoda bootstrap a v neposlední řadě transformačních metod. Z transformačních metod byla uvedena Box-Coxova transformace a výpočet korigovaného průměru pomocí Hallovy transformace.

K vypracování této bakalářské práce byl Matlab zvolen z toho důvodu, že je všestranně zaměřený a není primárně určen jen pro jednu oblast použití. Díky statistickému toolboxu a vysokému výkonu systému se stává výborným pomocníkem při rozsáhlých analýzách.

Je patrné, že statistické zpracování dat má celou řadu specifických zvláštností, které je třeba brát v úvahu. Je vždy výhodné začít exploratorní analýzou a porovnáním resp. selekcí modelů měření a až poté zvolit další přístupy jako je transformace, robustní metody a počítačové intenzivní metody k dosažení rozumných výsledků. Formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy bez hlubšího rozboru zde může vést ke katastrofickým výsledkům.

ZÁVĚR V ANGLIČTINĚ

The aim of this thesis was to describe selected statistical tests and their use in Matlab and develop functional programs that will serve as a learning tool for teaching statistics courses.

Thesis began with the introduction of the statistics toolbox and an introduction to inferential statistics. The end of the theoretical part was divided into two main parts: parametric and nonparametric tests. The use of tests was illustrated with examples randomly generated and real data. Individual examples have been implemented in MATLAB 7.9 (R2009b).

The practical part of this work was described in Chapter 5 by the principle of testing hypotheses using Monte Carlo method. In Chapter 6, were presented a selected statistics programs and applications developed in Matlab, which can serve as a learning tool for teaching mathematical statistics and statistical quality control. These programs were related to approximation of the distributions, demonstration of the asymmetry data, univariate tests for normality and multidimensional data sets, interval estimates of means, as traditional methods, and computer-intensive methods such as bootstrap method, ultimately, transformation methods. From transformation method was introduced the Box-Cox transformation and the calculation of corrected average using Hall's transformation.

I chose MATLAB because it is broadly focused and is not primarily intended for only one application area. Due to the statistical toolbox and high performance of the system becomes a great help in large-scale analysis.

It is obvious that the statistical data processing has a number of specific traits that should be taken into account. It's always good to start exploratory analysis and comparison respectively. Selection and measurement models and then choose other approaches such as transformation, robust methods and computer intensive methods to achieve reasonable results. Adapting to the needs of statistical data analysis without a deeper analysis here can lead to disastrous results.

SEZNAM POUŽITÉ LITERATURY

- [1] ANDĚL, J. *Základy matematické statistiky*. 2. vyd. Praha: Matfyzpress, 2007. 358 s. ISBN 80-7378-001-1.
- [2] FABIAN F., HORÁLEK V., KŘEPELA J., MICHÁLEK J., CHMELÍK V., CHODOUNSKÝ J., KRÁL J.: *Statistické metody řízení jakosti*. Praha, ČSJ, 2007, ISBN 978-80-02-01897-1.
- [3] HANOUSEK, J. CHARAZMA, P. *Moderní metody zpracování dat*. Matematická statistika pro každého. 1. vyd. Praha: Edice EDUCA '99, 1992. 216 s. ISBN 80-85623-31-5.
- [4] HEBÁK, P. *Vícerozměrné statistické metody*. 1. vyd. Praha: Informatorium, spol. s.r.o., 2004. 236 s. ISBN 80-7333-025-3.
- [5] HENDL, J. *Přehled statistických metod: analýza a metaanalýza dat*. 3. vydání. Praha: Nakladatelství Portál, s.r.o., 2009. 687 s. ISBN 978-80-7367-482-3.
- [6] KLÍMEK, P. *Aplikovaná statistika – přednášky*. Skripta pro 2. ročník denního studia Zlín: UTB, FaME, 2008. ISBN 978-80-7318-671-5.
- [7] MARTINEZ, W., MARTINEZ, A. *Computational Statistics Handbook with Matlab*. 2nd ed. London: Chapman & Hall/CRC, UK, 2008. 767 p. ISBN 978-1-58488-566-1.
- [8] MATLAB 7.9 (R2009b) – Help.
- [9] MELOUN, M., MILITKÝ, J. *Kompendium statistického zpracování dat*. 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2006. 982 s. ISBN 80-200-1396-2.
- [10] MELOUN, M., MILITKÝ, J. *Statistická analýza experimentálních dat*. 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2004. 953 s. ISBN 80-200-1254-0.
- [11] RYTÍŘ, V., STRŽÍŽ P., KLÍMEK, P., KASAL, Roman. *Přednášky z Metod statistické analýzy*. Zlín: FaME, 2005. ISBN 80-7318-353-6.

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

PDF	(Probability Distribution Function) Pravděpodobnostní funkce.
CDF	(Cumulative Distribution Function) Distribuční funkce.
ANOVA	(Analysis of Variance) Analýza rozptylu.
H_0	Formulace nulové hypotézy.
H_1 (H_A)	Formulace alternativní hypotézy.
V	Podprostor obsahující hodnoty svědčící ve prospěch H_0 , tzv. obor přijetí.
W	Podprostor obsahující hodnoty svědčící ve prospěch H_1 , tzv. kritický obor.
α	Hladina významnosti (pravděpodobnost chyby I. druhu).
β	Pravděpodobnost chyby II. druhu.
T(X)	Testové kritérium (testovací statistika).
P-hodnota	Nejmenší hladina významnosti α , na které zamítneme H_0 .

SEZNAM OBRÁZKŮ

Obr. 2.1. Vztah chyby prvního a druhého druhu.....	21
Obr. 2.2. Vztah chyby prvního a druhého druhu při zvětšení n	22
Obr. 2.3. P -hodnota.....	23
Obr. 2.4. P -hodnota jako pozorovaná hladina významnosti.....	23
Obr. 5.1. Normální pravděpodobnostní graf dat <i>mcd</i> data ukazuje, že se jedná o normální rozdělení.....	60
Obr. 5.2. Odhadovaná síla testu.....	66
Obr. 6.1. Program pro interaktivní dvourozměrné normální rozdělení.....	69
Obr. 6.2. Rozdělení výběrových průměrů z rovnoměrného rozdělení pro různé rozsahy výběru.....	71
Obr. 6.3. Program pro aproximaci exponenciálního rozdělení normálním.....	73
Obr. 6.4. Program pro aproximaci binomického rozdělení rozdělením normálním.....	74
Obr. 6.5. Vzájemná poloha průměru a mediánu.....	75
Obr. 6.6. Možné tvary špičatého rozdělení.....	76
Obr. 6.7. Program pro demonstraci symetrie a asymetrie rozložení dat.....	77
Obr. 6.8. Program na výpočet pravděpodobnosti pod křivkou funkce hustoty pravděpodobnosti.....	78
Obr. 6.9. Program pro test vícerozměrné normality.....	81
Obr. 6.10. Výběrová kvantilová funkce.....	84
Obr. 6.11. Rankitový graf s robustní přímkou.....	85
Obr. 6.12. Demonstrace principu intervalu spolehlivosti pro střední hodnotu.....	87
Obr. 6.13. Rozdělení veličin p_i a t_i (nahrazení podlimitních hodnot nulou).....	90
Obr. 6.14. Odhad kvantilů transformovaných dat vypočítaných po vložení příslušných procent do „Add - Percentil Lines“ na pravděpodobnostním grafu.....	95
Obr. 6.15. Kvantily vypočítané z modelu lognormálního rozdělení (červené 0.04; 0.91; 20.23) a na základě zpětné BoxCox transformace (modré 0.04; 1.10; 20.23).....	95
Obr. 6.16. Kvantilový graf původních dat.....	96
Obr. 6.17. Q-Q graf transformovaných dat.....	96
Obr. 6.18. Odhad parametru λ pomocí metody maximální věrohodnosti.....	97

SEZNAM TABULEK

Tab. 2.1. Výsledky testu hypotéz (skutečnost versus rozhodnutí)	21
Tab. 4.1. Schéma kontingenční tabulky	40
Tab. 4.2. Zadání příkladu pro výpočet χ^2 testu o nezávislosti v kombinační tabulce	41
Tab. 4.3. Zadání pro výpočet Wilcoxonova testu pro dva závislé výběry	47
Tab. 4.4. Hodnoty pro výpočet příkladu pomocí jednovýběrového Wilcoxonova testu	48
Tab. 4.5. Zadání příkladu pro použití Mannova-Whitneyova testu	49
Tab. 4.6. Upravené zadání pro použití Mannova-Whitneyova testu	50
Tab. 4.7. Zadání příkladu pro výpočet χ^2 testu dobré shody	51
Tab. 4.8. Zadání příkladu pro výpočet Spearmanova koeficientu pořadové korelace	55
Tab. 6.1. Hodnoty λ a k nim transformované proměnné x	94

SEZNAM PŘÍLOH

- Příloha P I: Funkce pro testování statistických hypotéz v programu Matlab
- Příloha P II: Interaktivní dvourozměrné normální rozdělení
- Příloha P III: Interaktivní centrální limitní teorém – aproximace exponenciálního rozdělení normálním
- Příloha P IV: Interaktivní aproximace binomického rozdělení normálním
- Příloha P V: Aplikace na porovnání normálního rozdělení s asymetrickým
- Příloha P VI: Výpočet obsahu plochy pod křivkou normálního rozdělení
- Příloha P VII: Test pro vícerozměrnou normalitu dat
- Příloha P VIII: Anderson-Darlingův test normality dat
- Příloha P IX: Box-Coxova transformace

PŘÍLOHA P I: FUNKCE PRO TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ V PROGRAMU MATLAB

Funkce	Název testu	Popis funkce	Syntaxe
ansaribradley	Ansari-Bradleyův test	Test hypotézy, že dva nezávislé výběry pocházejí ze stejného rozdělení, oproti alternativě, že pocházejí z rozdělení, které mají stejný tvar a střední hodnoty, ale různé rozptyly.	[h,p,stats] = ansaribradley(x,y,alpha,tail)
barttest	Bartlettův test	Lze jej využít k hodnocení homoskedasticity jak u vyvážených, tak i u nevyvážených souborů.	[ndim,prob,chisquare] = barttest(X,alpha)
canoncorr	Kanonická korelace	Kanonická korelace hledá obecný lineární vztah mezi dvěmi vícerozměrnými proměnnými X a Y s obecně různými dimenzemi m1, m2.	[A,B,r,U,V,stats] = canoncorr(X,Y)
chi2gof	Chí-kvadrát test dobré shody	Umožňuje ověřit, zda má náhodná veličina určité předem dané rozdělení pravděpodobnosti.	[h,p,stats] = chi2gof(X,name1,val1,name2,val2,...)
dwtest	Durbin-Watsonův test	Testuje nezávislost reziduí. Je-li výsledná hodnota blízká číslu 2, rezidua nejsou autokorelovaná a model byl zvolen správně.	[P,DW] = dwtest(R,X,method,tail)
friedman	Friedmanův test	Friedmanův test je obdobou analýzy rozptylu dvojného třídění s jedním pozorováním v každé podtřídě.	[p,table,stats] = friedman(X,reprs)

jbtest	Jarque-Berův test	Test normality, že vektor x pochází z normálního rozdělení s neznámou střední hodnotou a rozptylem.	[h,p,jbstat,critval] = jbtest(x,alpha,mctol)
kruskalwallis	Kruskal-Wallisův test	Používá se v případě, když se rozptyly souborů statisticky významně liší nebo v případě nenormálního rozdělení	[p,table,stats] = kruskalwallis(X,group)
kstest	Jednovýběrový Kolmogorov-Smirnovův test	Používá se pro ověření normality dat.	[h,p,ksstat,cv] = kstest(x,CDF,alpha,type)
kstest2	Dvouvýběrový Kolmogorov-Smirnovův test	Je metoda matematické statistiky, která umožňuje testovat, zda dvě jednorozměrné náhodné proměnné pocházejí ze stejného rozdělení pravděpodobnosti	[h,p,ks2stat] = kstest2(x1,x2,alpha,type)
lillietest	Lillieforsův test	Je dalším z používaných neparametrických testů nezávislosti.	[h,p,kstat,critval] = lillietest(x,alpha,distr)
linhypstest	Test linearity	K hodnocení linearity se používá téměř výhradně korelační koeficient. Korelační koeficient R měří stupeň korelace a ne linearitu.	p = linhypstest(beta,COVB,c,H,dfe)
ranksum	Wilcoxonův dvouvýběrový znaménkový test	Vykonává oboustranný test hypotézy, že data ve vektoru x a y jsou nezávislé výběry ze stejných spojitých rozdělení se stejnými mediány.	[p,h,stats] = ranksum(x,y,'alpha',alpha)
runstest	Run test náhodnosti	Toto je test nulové hypotézy, že hodnoty ve vektoru x jsou v náhodném	[h,p,stats] = runstest(...,param1,val1,param2,val2,...)

		pořadí, proti alternativní hypotéze, že nejsou.	
sampsizepwr	Rozsah výběru a síla testu	Tento test vrací velikost výběru oboustranného testu a sílu testu při zadané hladině významnosti.	$n = \text{sampsizepwr}(\text{testtype}, p_0, p_1, \text{power})$
signrank	Wilcoxonův jednovýběrový znaménkový test	Provede oboustranný test nulové hypotézy, že data ve vektoru x pochází ze spojitého, symetrického rozdělení s nulovou střední hodnotou, oproti alternativní hypotéze, že rozdělení nemá nulovou střední hodnotu.	$[p, h, \text{stats}] = \text{signrank}(\dots, 'alpha', \alpha)$
signtest	Znaménkový test	Provede oboustranný test nulové hypotézy, že data ve vektoru x pochází ze spojitého rozdělení s nulovým mediánem, oproti alternativní hypotéze, že rozdělení nemá medián roven nule.	$[p, h, \text{stats}] = \text{signtest}(\dots, 'alpha', \alpha)$
ttest	Jednovýběrový a párový t-test	Provede oboustranný test, že data ve vektoru x pocházejí z normálního rozdělení se střední hodnotou 0 a neznámým rozptylem, oproti alternativě, že střední hodnota se nerovná 0.	$[h, p, ci, \text{stats}] = \text{ttest}(\dots, \alpha, \text{tail}, \text{dim})$

ttest2	Dvouvýběrový t-test	Testuje hypotézu, že data ve vektoru x a y jsou nezávislé náhodné výběry z normálního rozdělení se shodnými středními hodnotami a se shodnými, ale neznámými rozptyly.	[h,p,ci,stats] = ttest2(x,y,alpha,tail,vartype,dim)
vartest	Chi-kvadrát test o rozptylech	Provede chi-kvadrát test, že data ve vektoru x pocházejí z normálního rozdělení se stejným rozptylem.	[H,P,CI,STATS] = vartest(X,V,alpha,tail)
vartest2	Dvouvýběrový F-test pro shodu rozptylů	Provede F-test, že dva nezávislé výběry ve vektoru x a y pocházejí z normálního rozdělení se stejnými rozptyly.	[H,P,CI,STATS] = vartest2(X,Y,alpha,tail)
vartestn	Bartlettův vícevýběrový test pro rovnost rozptylů	Provede Bartlettův test pro shodu rozptylů ve sloupcích matice X.	[p,STATS] = vartestn(X,group)
zscore	Standardizované Z-skóre	Provádí standardizaci vektoru x na tvar $z = (x - \text{mean}(x)) ./ \text{std}(x)$	[Z,mu,sigma] = zscore(X)
ztest	Z-test	Provádí z-test, že data ve vektoru x jsou náhodné výběry pocházející z normálního rozdělení se střední hodnotou m a směrodatnou odchylkou sigma, oproti alternativní hypotéze, že střední hodnota se nerovná m.	[h,p,ci] = ztest(x,m,sigma)

PŘÍLOHA P II: INTERAKTIVNÍ DVOUROZMĚRNÉ NORMÁLNÍ ROZDĚLENÍ

```
function bvnormal(option, first)

if nargin==0,

    option=0;
    figure('Name','Interaktivní dvourozměrné normální rozdělení', ...
        'Resize','off');

    border = 0.02;
    dimen1 = 0.20;

    uicontrol('Style','Frame', 'Units','normalized',...
        'Position',[border border 1-2*border 0.2-border]);

    uicontrol('Style','Frame', 'Units','normalized',...
        'Position',[0.76 0.2+border 0.24-border 1-0.2-2*border]);

    pbborder = 0.02;
    pbwidth = ( 1 - (2 * border) - 6*pbborder ) / 5;
    uicontrol('Style','text','String','stř.hodnota
x','Units','normalized',...
        'Position',[0.02+pbborder 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider', 'Max',5, 'Min',-5, ...
        'Callback','bvnormal(8);','value',0.0,...
        'tag','hmx', 'Units','normalized', ...
        'Position',[0.02+pbborder 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','0.0','Units','normalized',...
        'Position',[0.02+pbborder 0.09 pbwidth 0.05],...
        'tag','hmxedit','Callback','bvnormal(9);');

    uicontrol('Style','text','String','stř. hodnota
y','Units','normalized',...
        'Position',[0.02+2*pbborder+pbwidth 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider', 'Max',5, 'Min',-5, ...
        'Callback','bvnormal(8);','value',0.0,...
        'tag','hmy', 'Units','normalized', ...
        'Position',[0.02+2*pbborder + pbwidth 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','0.0','Units','normalized',...
        'Position',[0.02+2*pbborder + pbwidth 0.09 pbwidth 0.05],...
        'tag','hmyedit','Callback','bvnormal(9);');

    uicontrol('Style','text','String','sm. odchylka
x','Units','normalized',...
        'Position',[0.02+3*pbborder + 2*pbwidth 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider', 'Max',5, 'Min',0.1,...
        'Callback','bvnormal(8);',...
        'Value',1.0,'tag','hsx', 'Units','normalized',...
        'Position',[0.02+3*pbborder + 2*pbwidth 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','1.0','Units','normalized',...
        'Position',[0.02+3*pbborder + 2*pbwidth 0.09 pbwidth
0.05],...
        'tag','hsxedit','Callback','bvnormal(9);');

    uicontrol('Style','text','String','sm. odchylka
y','Units','normalized',...
        'Position',[0.02+4*pbborder + 3*pbwidth 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider', 'Max',5, 'Min',0.1, ...
```

```

        'Callback', 'bvnorm(8);', ...
        'Value', 1.0, 'tag', 'hsy', 'Units', 'normalized', ...
        'Position', [0.02+4*pbborder + 3*pbwidth 0.04 pbwidth 0.05]);
uicontrol('Style', 'Edit', 'String', '1.0', 'Units', 'normalized', ...
0.05], ...
        'Position', [0.02+4*pbborder + 3*pbwidth 0.09 pbwidth
        'tag', 'hsyedit', 'Callback', 'bvnorm(9);');

uicontrol('Style', 'text', 'String', 'korel.
koeficient', 'Units', 'normalized', ...
        'Position', [0.02+5*pbborder + 4*pbwidth 0.14 pbwidth 0.05]);
uicontrol('Style', 'Slider', 'Max', 0.99, 'Min', -0.99, ...
        'Callback', 'bvnorm(8);', 'value', 0, ...
        'tag', 'hrho', 'Units', 'normalized', ...
        'Position', [0.02+5*pbborder + 4*pbwidth 0.04 pbwidth
0.05]);
uicontrol('Style', 'Edit', 'String', '0.0', 'Units', 'normalized', ...
0.05], ...
        'Position', [0.02+5*pbborder + 4*pbwidth 0.09 pbwidth
        'tag', 'hrhoedit', 'Callback', 'bvnorm(9);');

oborder = 0.78;
owidth = 0.18;
oheight = 0.07;

uicontrol('Style', 'Text', 'String', 'Typ grafu', 'Units', 'normalized', ...
        'Position', [ oborder 0.96-0.08 owidth oheight] );

uicontrol('Style', 'Popup', 'String', 'Mesh|Surf|Černobílá', 'Units', 'normali
zed', ...
        'Position', [ oborder 0.96-2*0.08+0.03 owidth oheight], ...
        'tag', 'htypeplot', 'value', 1, ...
        'Callback', 'bvnorm(10);');

uicontrol('Style', 'Text', 'String', 'Pohled', 'Units', 'normalized', ...
        'Position', [ oborder 0.96-3*0.08 owidth oheight] );
uicontrol('Style', 'Popup', 'String', 'Defaultní|Vrchní|Z pohledu X|Z
pohledu Y', ...
        'tag', 'hview', ...
        'Callback', 'bvnorm(1);', 'Units', 'normalized', ...
        'Position', [ oborder 0.96-4*0.08+0.03 owidth oheight]);

uicontrol('Style', 'Text', 'String', 'Elevace', 'Units', 'normalized', ...
        'Position', [oborder 0.96-5*0.08 owidth oheight]);
uicontrol('Style', 'Slider', 'Max', 180, 'Min', -
180, 'Units', 'normalized', ...
        'Position', [ oborder 0.96-6*0.08+0.03 owidth 0.05], ...
        'value', 30.0, ...
        'tag', 'helevation', 'Callback', 'bvnorm(2);');

uicontrol('Style', 'Text', 'String', 'Azimut', 'Units', 'normalized', ...
        'Position', [oborder 0.96-7*0.08 owidth oheight]);
uicontrol('Style', 'Slider', 'Max', 180, 'Min', -
180, 'Units', 'normalized', ...
        'Position', [ oborder 0.96 - 8*0.08+0.03 owidth 0.05], ...
        'value', -37.5, ...
        'tag', 'hazimuth', 'Callback', 'bvnorm(3);');

uicontrol('Style', 'Pushbutton', 'String', 'Zavřít', 'Units', 'normalized', ...
        'Position', [ oborder 0.25 owidth oheight], ...
        'Callback', 'close(gcf);');

```

```

axes('Position',[0.08 0.30 0.65 0.65 ])
bvnorm(10, 1);

cwbuf = get(gcf,'WindowButtonUpFcn');
set(gcf,'WindowButtonUpFcn',[cwbuf,'; bvnorm(4);']);
set(gcf,'HandleVisibility','Callback');

end;

if option == 1, %Zmena pohledu

value = get( findobj('tag','hview'), 'Value');

if (value==1), %defaultni
view(-37.5, 30);
set( findobj('tag','hazimuth'),'value',-37.5);
set( findobj('tag','helevation'),'value',30.0);
elseif (value==2),
view(0, 90);
set( findobj('tag','hazimuth'),'value',0.0);
set( findobj('tag','helevation'),'value',90.0);
elseif (value==3), %podminka x
view(0, 0);
set( findobj('tag','hazimuth'),'value',0.0);
set( findobj('tag','helevation'),'value',0.0);
elseif (value==4), %podminka y
view(90,0);
set( findobj('tag','hazimuth'),'value',90.0);
set( findobj('tag','helevation'),'value',90.0);
end;

elseif (option==2), %zmena elevace

value = get( findobj('tag','helevation'), 'value');

[az, el] = view;
view(az, value);

elseif (option==3), %zmena azimutu

value = get( findobj('tag','hazimuth'), 'value');
if value>180,
value = 360-value;
set( findobj('tag','hazimuth'), 'value', value );
elseif value<-180
value = 360+value;
set( findobj('tag','hazimuth'), 'value', value );
end;

[az, el] = view;
view(value, el);

elseif (option==4), %nastaveni azimutu a elevace po 3D rotaci

[az, el] = view;
if az>180,
az = 360-az;
set( findobj('tag','hazimuth'), 'value', az );
elseif az<-180
az = 360+az;

```

```

        set( findobj('tag','hazimuth'), 'value', az );
    end;
    set( findobj('tag','hazimuth'),'value',az);

    set( findobj('tag','helevation'),'value',el);

elseif (option==8), %slidovani

    set( findobj('tag','hmxedit'), ...
        'string', num2str(get(findobj('tag','hmx'),'value')) );

    set( findobj('tag','hmyedit'), ...
        'string', num2str(get(findobj('tag','hmy'),'value')) );

    set( findobj('tag','hsxedit'), ...
        'string', num2str(get(findobj('tag','hsx'),'value')) );

    set( findobj('tag','hsyedit'), ...
        'string', num2str(get(findobj('tag','hsy'),'value')) );

    set( findobj('tag','hrhoedit'), ...
        'string', num2str(get(findobj('tag','hrho'),'value')) );

    bvnorm(10);

elseif (option==9), %Editace parametru

    entval = str2num( get(findobj('tag','hmxedit'),'string') );
    if (entval < -5 )
        set( findobj('tag','hmxedit'),'string', '-5.0');
        entval = -5.0;
    elseif (entval > 5)
        set( findobj('tag','hmxedit'),'string', '5.0');
        entval = 5.0;
    end
    set( findobj('tag','hmx'), 'value', entval );

    entval = str2num( get(findobj('tag','hmyedit'),'string') );
    if (entval < -5 )
        set( findobj('tag','hmyedit'),'string', '-5.0');
        entval = -5.0;
    elseif (entval > 5)
        set( findobj('tag','hmyedit'),'string', '5.0');
        entval = 5.0;
    end
    set( findobj('tag','hmy'), 'value', entval );

    entval = str2num( get(findobj('tag','hsxedit'),'string') );
    if (entval < 0.1 )
        set( findobj('tag','hsxedit'),'string', '0.1');
        entval = 0.1;
    elseif (entval > 5)
        set( findobj('tag','hsxedit'),'string', '5.0');
        entval = 5.0;
    end
    set( findobj('tag','hsx'), 'value', entval );

    entval = str2num( get(findobj('tag','hsyedit'),'string') );
    if (entval < 0.1 )
        set( findobj('tag','hsyedit'),'string', '0.1');
        entval = 0.1;
    elseif (entval > 5)

```

```

        set( findobj('tag','hsyedit'),'string', '5.0');
        entval = 5.0;
    end
    set( findobj('tag','hsy'), 'value', entval );

    entval = str2num( get(findobj('tag','hrhoedit'),'string') );
    if (entval < -0.99 )
        set( findobj('tag','hrhoedit'),'string', '-0.99');
        entval = -0.99;
    elseif (entval > 0.99)
        set( findobj('tag','hrhoedit'),'string', '0.99');
        entval = 0.99;
    end
    set( findobj('tag','hrho'), 'value', entval );

    bvnorm(10);

elseif (option==10), %Zmeny parametru

    if nargin==2,
        el = get( findobj('tag','helevation'), 'value');
        az = get( findobj('tag','hazimuth'), 'value');
    else
        [az, el] = view;
    end;

    meanx = get(findobj('tag','hmx'), 'Value');
    meany = get(findobj('tag','hmy'), 'Value');
    stdx = get(findobj('tag','hsx'), 'Value');
    stdy = get(findobj('tag','hsy'), 'Value');
    rho = get(findobj('tag','hrho'), 'Value');

    [zx, zy] = meshgrid( linspace(-3.5, 3.5, 40), linspace(-3.5, 3.5, 40)
);
    x = meanx + stdx*zx;
    y = meany + stdy*zy;

    Q = ( zx.^2 - 2*rho*zx.*zy + zy.^2 ) ./ ( 1 - rho^2 );

    z = exp( -0.5*Q ) ./ ( 2*pi*stdx*stdy*sqrt(1-rho^2) );

    how = get( findobj( 'tag','htypeplot' ), 'value');
    if (how==1), %mesh
        hplot = mesh(x, y, z);
        colormap default;
    elseif (how==2), %surface
        hplot = surf(x, y, z);
        colormap default;
    elseif (how==3), %cerna a bila
        hplot = surf(x, y, z);
        colormap gray; brighten(0.5);
    end

    view(az, el);
    rotate3d on;
    xlabel('osa x');
    ylabel('osa y');

    drawnow;

end

```

PŘÍLOHA P III: INTERAKTIVNÍ CENTRÁLNÍ LIMITNÍ TEORÉM – APROXIMACE EXPONENCIÁLNÍHO ROZDĚLENÍ NORMÁLNÍM

```
function cltexp(option)

if nargin == 0,

    option=0;
    figure('Name','Interaktivní centrální limitní teorém',...
        'Resize','off');

    uicontrol('Style','Frame','Units','normalized',...
        'Position',[0.02 0.02 1-2*0.02 0.2-0.02]);

    pbborder = 0.02;
    pbwidth = ( 1 - ( 2 * 0.02) - 6*pbborder ) / 4;

    uicontrol('Style','text','String','Exp. stř. hodnota, beta',...
        'Units','normalized',...
        'Position',[0.02+pbborder 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider', 'Max',10.0, 'Min',0.01, ...
        'Callback','cltexp(1);','value',2.0,...
        'tag','hcltemean', 'Units','normalized', ...
        'Position',[0.02+pbborder 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','2.0','Units','normalized',...
        'Position',[0.02+pbborder 0.09 pbwidth 0.05],...
        'tag','hcltmeantedit','Callback','cltexp(2);');

    uicontrol('Style','text','String','Rozsah výběru,
n','Units','normalized',...
        'Position',[0.02+2*pbborder+pbwidth 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider', 'Max',200, 'Min',1, ...
        'Callback','cltexp(1);','value',10,...
        'tag','hcltssize', 'Units','normalized', ...
        'Position',[0.02+2*pbborder+pbwidth 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','10','Units','normalized',...
        'Position',[0.02+2*pbborder+pbwidth 0.09 pbwidth 0.05],...
        'tag','hcltssizeedit','Callback','cltexp(2);');

    uicontrol('Style','text','String','Výběr funkce:',...
        'Units','normalized',...
        'Position',[0.02+3*pbborder+2*pbwidth 0.14 pbwidth 0.05],...
        'HorizontalAlignment','left');
    uicontrol('Style','text','String','Stř. hodnota:
2.0','Units','normalized',...
        'Position',[0.02+3*pbborder+2*pbwidth 0.09 pbwidth 0.05],...
        'tag','hcltmeantext','HorizontalAlignment','left');
    uicontrol('Style','text','String','Rozptyl:
0.4','Units','normalized',...
        'Position',[0.02+3*pbborder+2*pbwidth 0.04 pbwidth 0.05],...
        'tag','hcltvartext','HorizontalAlignment','left');

    uicontrol('Style','Popup','String','pravděpodobnostní|distribuční',...
        'Callback','cltexp(5);','Units','normalized',...
        'tag','hclttypedist',...
        'Position',[0.02+4*pbborder+3*pbwidth 0.12 pbwidth 0.07]);

    uicontrol('Style','Pushbutton','String','Zavřít',...
        'Callback','close(gcf);','Units','normalized',...
        'Position',[0.02+4*pbborder+3*pbwidth 0.04 pbwidth 0.07]);
```

```

        cltexp(5);

end;

if (option==1),

    ssize = round( get(findobj('tag','hcltssize'),'value') );
    set( findobj('tag','hcltssizeedit'), ...
        'string', ssize );

    emean = get(findobj('tag','hcltemean'),'value');
    set( findobj('tag','hcltemeanedit'), ...
        'string', emean );

    set( findobj('tag','hcltmeantext'), ...
        'String', ['Mean: ', num2str(emean)]);
    set( findobj('tag','hcltvartext'), ...
        'String', ['Var: ', num2str(emean^2/ssize)]);

    cltexp(5);

elseif (option==2),

    emean = str2num( get(findobj('tag','hcltemeanedit'),'string') );
    if (emean < 0.01 )
        set( findobj('tag','hcltemeanedit'),'string', '0.01');
        emean = 0.01;
    elseif (emean > 10.0)
        set( findobj('tag','hcltemeanedit'),'string', '10.0');
        emean = 10.0;
    end
    set( findobj('tag','hcltemean'), 'value', emean );

    ssize = round( str2num( get(findobj('tag','hcltssizeedit'),'string') )
);
    if (ssize < 1 )
        ssize = 1;
    elseif (ssize > 200)
        ssize = 200;
    end
    set( findobj('tag','hcltssize'), 'value', ssize );
    set( findobj('tag','hcltssizeedit'),'string', ssize);

    set( findobj('tag','hcltmeantext'), ...
        'String', ['Mean: ', num2str(emean) ]);
    set( findobj('tag','hcltvartext'), ...
        'String', ['Var: ', num2str(emean^2/ssize)]);

    cltexp(5);

elseif (option==5),

    delete(gca);
    axes('Position',[0.10 0.27 0.85 0.70],'tag','haxes');

    emean = get(findobj('tag','hcltemean'),'value');
    ssize = round(get(findobj('tag','hcltssize'),'value'));
    mn = emean;
    std = sqrt( emean^2/ssize );

    lower = max( 0.0001, mn - 4 * std);

```

```

upper = mn + 4 * std;
xg = linspace(lower, upper, 100);

if get(findobj('tag','hclttypedist'), 'value') == 1, %pdf

    ldenom = gammaln(ssize) + ssize*log(emean/ssize);
    ep = (ssize-1)*log(xg) + (-ssize*xg/emean);
    yg = exp( ep - ldenom );

    ep = -0.5*( (xg-mn)/std ).^2;
    yn = 1/(std*sqrt(2*pi)) * exp( ep );

    plot( xg, yg, 'b-', xg, yn, 'r--');

    axis( [lower upper 0 max([yg, yn])] );

else %cdf

    yg = gammainc( xg*ssize/emean, ssize );

    yn = ( erf( (xg - mn)/(sqrt(2)*std) ) + 1 ) / 2;
    plot( xg, yg, 'b-', xg, yn, 'r--');

    axis( [lower upper 0 1] );

end

set(gca, 'HandleVisibility','Callback');
set(gcf, 'HandleVisibility','Callback');

end

return;

```

PŘÍLOHA P IV: INTERAKTIVNÍ APROXIMACE BINOMICKÉHO ROZDĚLENÍ NORMÁLNÍM

```
function normbin(option)

if nargin == 0,

    option=0;
    figure('Name','Interaktivní aproximace binomického rozdělení
normálním',...
        'Resize','off');

    uicontrol('Style','Frame','Units','normalized',...
        'Position',[0.02 0.02 1-2*0.02 0.2-0.02]);

    pbborder = 0.02;
    pbwidth = ( 1 - (2 * 0.02) - 6*pbborder ) / 4;

    uicontrol('Style','text','String','Pravděpodobnost,
p','Units','normalized',...
        'Position',[0.02+pbborder 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider','Max',0.99, 'Min',0.01, ...
        'Callback','normbin(1);','value',0.2,...
        'tag','hprob','Units','normalized', ...
        'Position',[0.02+pbborder 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','0.2','Units','normalized',...
        'Position',[0.02+pbborder 0.09 pbwidth 0.05],...
        'tag','hprobedit','Callback','normbin(2);');

    uicontrol('Style','text','String','Rozsah výběru,
n','Units','normalized',...
        'Position',[0.02+2*pbborder+pbwidth 0.14 pbwidth 0.05]);
    uicontrol('Style','Slider','Max',100, 'Min',3, ...
        'Callback','normbin(1);','value',10,...
        'tag','hssize','Units','normalized', ...
        'Position',[0.02+2*pbborder+pbwidth 0.04 pbwidth 0.05]);
    uicontrol('Style','Edit','String','10','Units','normalized',...
        'Position',[0.02+2*pbborder+pbwidth 0.09 pbwidth 0.05],...
        'tag','hssizeedit','Callback','normbin(2);');

    uicontrol('Style','text','String','Stř. hodnota:
1.0','Units','normalized',...
        'Position',[0.02+3*pbborder+2*pbwidth 0.14 pbwidth 0.05],...
        'tag','hmeantext','HorizontalAlignment','left');
    uicontrol('Style','text','String','Rozptyl:
0.8','Units','normalized',...
        'Position',[0.02+3*pbborder+2*pbwidth 0.04 pbwidth 0.05],...
        'tag','hvartext','HorizontalAlignment','left');

    uicontrol('Style','Popup','String','pravděpodobnostní|distribuční',...
        'Callback','normbin(5);','Units','normalized',...
        'tag','htypedist',...
        'Position',[0.02+4*pbborder+3*pbwidth 0.12 pbwidth 0.07]);

    uicontrol('Style','Pushbutton','String','Zavřít',...
        'Callback','close(gcf);','Units','normalized',...
        'Position',[0.02+4*pbborder+3*pbwidth 0.04 pbwidth 0.07]);

    axes('Position',[0.10 0.27 0.85 0.70],'tag','haxes');
    normbin(5);
```

```

end;

if (option==1),

    prob = get(findobj('tag','hprob'),'value');
    set( findobj('tag','hprobedit'), ...
        'string', prob );
    ssize = round(get(findobj('tag','hssize'),'value'));
    set( findobj('tag','hssizeedit'), ...
        'string', ssize );

    set( findobj('tag','hmeantext'), ...
        'String', ['Mean: ', num2str(ssize*prob,3)]);
    set( findobj('tag','hvartext'), ...
        'String', ['Variance: ', num2str(ssize*prob*(1-prob),3)]);

    normbin(5);

elseif (option==2), %editace parametru

    prob = str2num( get(findobj('tag','hprobedit'),'string') );
    if (prob < 0.01 )
        set( findobj('tag','hprobedit'),'string', '0.01');
        prob = 0.01;
    elseif (prob > 0.99)
        set( findobj('tag','hprobedit'),'string', '0.99');
        prob = 0.99;
    end
    set( findobj('tag','hprob'), 'value', prob );

    ssize = round( str2num( get(findobj('tag','hssizeedit'),'string') ) );
    if (ssize < 3 )
        ssize = 3;
    elseif (ssize > 100)
        ssize = 100;
    end
    set( findobj('tag','hssize'), 'value', ssize );
    set( findobj('tag','hssizeedit'),'string', ssize);

    set( findobj('tag','hmeantext'), ...
        'String', ['Stř.hodnota: ', num2str(ssize*prob,3)]);
    set( findobj('tag','hvartext'), ...
        'String', ['Rozptyl: ', num2str(ssize*prob*(1-prob),3)]);

    normbin(5);

elseif (option==5), %plot

    delete(gca);
    axes('Position',[0.10 0.27 0.85 0.70],'tag','haxes');

    prob = get(findobj('tag','hprob'),'value');
    ssize = round(get(findobj('tag','hssize'),'value'));
    mn = ssize*prob;
    std = sqrt( ssize*prob*(1-prob) );

    %binomicke rozdeleni
    xb = [0:1:ssize];
    ncr = exp( gammaln(ssize+1) - gammaln(xb+1) - gammaln(ssize-xb+1) );
    ybinomial = ncr .* prob.^xb .* (1-prob) .^ (ssize - xb);

    %normalni rozdeleni

```

```

xn=linspace(-0.5,ssize+0.5);

if get(findobj('tag','htypedist'),'value') == 1, %pdf

    hbin = bar(xb,ybinomial,1,'w'); hold on;
    ep = -0.5*( (xn-mn)/std ).^2;
    ynorm = 1/(std*sqrt(2*pi)) * exp( ep );
    hnorm = plot(xn, ynorm, 'r-');

    axis( [-0.5 ssize+0.5 0 max([ybinomial, ynorm])]);

else %cdf

    sumy = cumsum(ybinomial);
    xbc = [-0.5 xb(1)];
    ybc = [0 0];

    for i=1:length(xb)-1
        xbc = [xbc xb(i) xb(i+1) ];
        ybc = [ybc sumy(i) sumy(i)];
    end

    xbc = [xbc xb(i+1) ssize+0.5];
    ybc = [ybc 1 1];

    ynorm = ( erf( (xn - mn)/(sqrt(2)*std) ) + 1 ) / 2;
    hnorm = plot(xbc, ybc, 'k-', xn,ynorm,'r-');

    axis( [-0.5 ssize+0.5 0 1]);

end

set(gca, 'HandleVisibility','Callback');
set(gcf, 'HandleVisibility','Callback');

end

```

PŘÍLOHA P V: APLIKACE NA POROVNÁNÍ NORMÁLNÍHO ROZDĚLENÍ S ASYMETRICKÝM

```
function normasym( option )

if nargin==0,

    hOverFit = figure('Name','Míry asymetrie', ...
        'tag','tagShowNorm');

    hNPPPlotist = axes( 'Position', [0.1 0.08 0.32 0.85], 'tag','tagNPPPlot'
);
    set(gca, 'FontSize',12);
    title('Pravděpodobnostní graf normálního rozdělení');
    hHist = axes( 'Position', [0.43 0.08 0.32 0.85], 'tag','tagHist' );
    set(gca, 'FontSize',12);
    title('Histogram');
    set( gca,'YTickLabel',[]);

    uicontrol('Style','Frame','Units','normalized',...
        'Position',[0.78 0.02 0.20 0.20] );

    uicontrol('Style','PushButton','Units','normalized',...
        'Position',[0.80 0.04 0.16 0.07], 'String','Zavřít',...
        'Callback','delete(gcf);');
    uicontrol('Style','PushButton','Units','normalized',...
        'Position',[0.80 0.13 0.16 0.07], 'String','Vykresli',...
        'tag','tagPauseBtn','Callback','nppdemo(1);');

    uicontrol('Style','Frame','Units','normalized',...
        'Position',[0.78 0.24 0.20 0.38] );

    uicontrol('Style','Text','Units','normalized',...
        'Position',[0.80 0.53 0.16 0.07], 'String','Stř. hodnota:');
    uicontrol('Style','Edit','Units','normalized',...
        'Position',[0.80 0.50 0.16 0.06], 'String','1',...
        'tag','tagMean','Callback','nppdemo(20);');
    uicontrol('Style','Text','Units','normalized',...
        'Position',[0.80 0.41 0.16 0.07], 'String','Rozptyl:');
    uicontrol('Style','Edit','Units','normalized',...
        'Position',[0.80 0.38 0.16 0.06], 'String','1',...
        'tag','tagVar','Callback','nppdemo(21);');
    uicontrol('Style','Text','Units','normalized',...
        'Position',[0.80 0.29 0.16 0.07], 'String','Počet bodů:');
    uicontrol('Style','Edit','Units','normalized',...
        'Position',[0.80 0.26 0.16 0.06], 'String','100',...
        'tag','tagNumPts','Callback','nppdemo(22);');

    uicontrol('Style','Frame','Units','normalized',...
        'Position',[0.78 0.64 0.20 0.34] );

    uicontrol('Style','Radio','Units','normalized',...
        'Position',[0.80 0.90 0.16 0.06], 'String','Normální', ...
        'Value',1,...
        'tag','tagShowNormal','Callback','nppdemo(10);');
    uicontrol('Style','Radio','Units','normalized',...
        'Position',[0.80 0.84 0.16 0.06], 'String','Zprava
zešikmené',...
        'tag','tagRightSkew','Callback','nppdemo(11);');
    uicontrol('Style','Radio','Units','normalized',...
```

```

        'Position',[0.80 0.78 0.16 0.06], 'String','Zleva
zešikmené',...
        'tag','tagLeftSkew','Callback','nppdemo(12);');
    uicontrol('Style','Radio','Units','normalized',...
        'Position',[0.80 0.72 0.16 0.06], 'String','Dlouhý
konec',...
        'tag','tagLongTail','Callback','nppdemo(13);');
    uicontrol('Style','Radio','Units','normalized',...
        'Position',[0.80 0.66 0.16 0.06], 'String','Krátký
konec',...
        'tag','tagShortTail','Callback','nppdemo(14);');

elseif nargin==1,

    if option==1, % vykreslit data

        mn = str2num( get( findobj( 'tag','tagMean' ), 'String' ) );
        vr = str2num( get( findobj( 'tag','tagVar' ), 'String' ) );
        np = str2num( get( findobj( 'tag','tagNumPts' ), 'String' ) );

        pt1 = 0;
        pt2 = 0;
        pt3 = 0;
        pt4 = 0;
        pt5 = 0;
        pt1 = get( findobj('tag','tagShowNormal'), 'Value');
        pt2 = get( findobj('tag','tagRightSkew'), 'Value');
        pt3 = get( findobj('tag','tagLeftSkew'), 'Value');
        pt4 = get( findobj('tag','tagLongTail'), 'Value');
        pt5 = get( findobj('tag','tagShortTail'), 'Value');

        if pt1 == 1, % Normalni data
            r = ( randn( np, 1 ) * vr ) + mn;
        end;
        if pt2 == 1, % Zprava sesikmene

            r = rand( np, 1);
            r = -mn * log( r );
        end;
        if pt3 == 1, % Zleva sesikmene

            r = rand( np, 1);
            r = -mn * log( r );
            r = max(r) - r;
        end;
        if pt4 == 1, % Dlouhy chvost

            r = rand( np, 1);
            r = sqrt(2) * ( r - 0.5 ) ./ ( sqrt( r - r.^2 ) );
            r = r + mn;
        end;
        if pt5 == 1, % Kratky chvost
            r = rand( [np, 1] );
            a = mn - sqrt( 3*vr );
            b = mn + sqrt( 3*vr );
            r = ( b - a ) * r + a;

        end;

        % Histogram
        axes( findobj('tag','tagHist') );
        [n, x] = hist( r );

```

```

barh( x, n, 1, 'g');
set( gca, 'tag', 'tagHist', 'View', [180 270]);
xlabel('Četnosti');
title('Histogram');
ax1 = axis;
set( gca, 'YTickLabel', []);

% Graf normality
axes( findobj('tag', 'tagNPPlot') );
limits = nprplot( r );
rlim = max(limits) - min(limits);
limits(1) = limits(1) - 0.1*rlim;
limits(2) = limits(2) + 0.1*rlim;
set( gca, 'tag', 'tagNPPlot');
ax2 = axis;
axis([ limits ax1(3:4) ] );

elseif option==10, % Zvoleni normality

set( findobj('tag', 'tagRightSkew'), 'Value', 0);
set( findobj('tag', 'tagLeftSkew'), 'Value', 0);
set( findobj('tag', 'tagLongTail'), 'Value', 0);
set( findobj('tag', 'tagShortTail'), 'Value', 0);

set( findobj('tag', 'tagVar'), 'Enable', 'on');

elseif option==11, % Zvoleni zprava sesikmene

set( findobj('tag', 'tagShowNormal'), 'Value', 0);
set( findobj('tag', 'tagLeftSkew'), 'Value', 0);
set( findobj('tag', 'tagLongTail'), 'Value', 0);
set( findobj('tag', 'tagShortTail'), 'Value', 0);

set( findobj('tag', 'tagVar'), 'Enable', 'off');

elseif option==12, % Zvoleni zleva sesikmene

set( findobj('tag', 'tagShowNormal'), 'Value', 0);
set( findobj('tag', 'tagRightSkew'), 'Value', 0);
set( findobj('tag', 'tagLongTail'), 'Value', 0);
set( findobj('tag', 'tagShortTail'), 'Value', 0);

set( findobj('tag', 'tagVar'), 'Enable', 'off');

elseif option==13, % Vyber dlouheho chvostu

set( findobj('tag', 'tagShowNormal'), 'Value', 0);
set( findobj('tag', 'tagRightSkew'), 'Value', 0);
set( findobj('tag', 'tagLeftSkew'), 'Value', 0);
set( findobj('tag', 'tagShortTail'), 'Value', 0);

set( findobj('tag', 'tagVar'), 'Enable', 'off');

elseif option==14, % Vyber kratkeho chvostu

set( findobj('tag', 'tagShowNormal'), 'Value', 0);
set( findobj('tag', 'tagRightSkew'), 'Value', 0);
set( findobj('tag', 'tagLeftSkew'), 'Value', 0);
set( findobj('tag', 'tagLongTail'), 'Value', 0);

set( findobj('tag', 'tagVar'), 'Enable', 'on');

```

```

elseif option == 20, % Zadani stredni hodnoty

    mn = get( findobj( 'tag','tagMean'), 'String');
    mn = str2num(mn);

    if isempty(mn), % zadani textu
        mn = 0;
    end;

    mn = real( mn );

    if ( abs(mn) > 100 ),
        mn = sign(mn)*100;
    end;

    set( findobj('tag','tagMean'), 'String', num2str(mn) );

elseif option == 21, % Vlozeni rozptylu

    vr = get( findobj( 'tag','tagVar'), 'String');
    vr = str2num(vr);

    if isempty(vr), % vlozeni textu
        vr = 1;
    end;

    vr = real( vr );

    if ( vr<= 0 ),
        vr = 1;
    end;

    if ( vr > 100 ),
        vr = 100;
    end;

    set( findobj('tag','tagVar'), 'String', num2str(vr) );

elseif option == 22, % počet hodnot

    np = get( findobj( 'tag','tagNumPts'), 'String');
    np = str2num(np);

    if isempty(np), % vlozeni textu
        np = 10;
    end;

    np = real( round(np) );

    if ( np <= 5 ),
        np = 5;
    end;

    if ( np > 1000 ),
        np = 1000;
    end;

    set( findobj('tag','tagNumPts'), 'String', num2str(np) );

end;
end;

```

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% SUBFUNKCE
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function limits = nprplot( data )
% Vykresleni pravděpodobnostního grafu pro data

sy = sort( data );
nd = inorm( (1:length(sy))/(length(sy)+1) );

% počet dat
plot(nd,sy,'b+');
xlabel('Standardizované N rozdělení');
ylabel('Hodnoty');
title('Pravděp. graf N rozdělení');

zx = [-3.5 0 3.5];
hold on;
cfplot = plot(zx, mean(sy) + zx*std(sy), 'r--');

limits = [ min(nd) max(nd) ];
hold off;

return

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%5

function x = inorm(p,mu,sigma)
%INVNORM inverzni distribucni funkce normálního rozdeleni

if nargin<3,
    sigma=1;
end;
if nargin==1,
    mu=0;
end;

%chybove hlasky
if ~(sigma>0),
    error('sigma musi byt kladna!');
end;
if size(p)~=size(mu) | size(p)~=size(sigma),
    error('Vstupni argumenty musi byt stejne velikosti!');
end;
%konec chybových hlasek

x = mu + sqrt(2)*sigma.*erfinv(2*p-1);

return;

```

PŘÍLOHA P VI: VÝPOČET OBSAHU PLOCHY POD KŘIVKOU NORMÁLNÍHO ROZDĚLENÍ

```
function prob=normaldistribution(x, mean, sigma)
resolution=0.01;
u=mean-3.5*sigma:resolution:mean+3.5*sigma; %array of points from -inf to
x
e=exp(1);
p=(1/(sqrt(2*pi)*sigma))*e.^(-((u-mean).^2)./(2*sigma^2));
mx=max(p);
p=p/mx; %normalizace na 1
k=find(u<x);
prob=sum(p(k))*resolution*mx;

%figure;
basey = min(0,min(p));
h = fill([u(1) u(k)], [p(k) basey], 'r'); % vyplneni červenou
hold on
plot(u,p);
v=gcf; h=get(v,'currentaxes');
set(h,'Xlim',[u(1) u(end)]);
set(h,'Ylim',[basey max(p)]);
title('Funkce hustoty pravděpodobnosti normálního rozdělení')
xlabel(['Stř. hodnota=' num2str(mean) ' Sm. odchylka='
num2str(sigma) ' P=' num2str(prob)]);
ylabel('funkce hustoty pravděpodobnosti (x)')
grid on
end
```

PŘÍLOHA P VII: TEST PRO VÍCEROZMĚRNOU NORMALITU DAT

```
function [Mulnortest] = Mulnortest(X,alpha)

if nargin < 2,
    alpha = 0.05; %(defaultni)
end;

if nargin < 1,
    error('Pozadovan nejmene jeden vstupni argument.');
```

end;

mX = mean(X); %stredni hodnoty z vektoru matice X.
[n,p] = size(X);
difT = [];

for j = 1:p;
 eval(['difT=[difT,(X(:,j)-mean(X(:,j)))]'];]);
end;

S = cov(X);
D2T = difT*inv(S)*difT';
D2 = sort(diag(D2T)); %Vzestupne ctverce Mahalanobisovych vzdalenessi.

Pr = [];
for i = 1:n;
 eval(['pr' num2str(i) '=(i-0.5)/n;']);
 eval(['x= pr' num2str(i) ';']);
 Pr = [Pr,x]; %Odpovidajici vyberove percentily.
end;

X2 = [];
for i = 1:n
 eval(['X2' num2str(i) '=chi2inv(pr' num2str(i) ',p);']);
 eval(['x= X2' num2str(i) ';']);
 X2=[X2,x]; %Ocekavane chi-kvadrat rozdeleni s p stupni volnosti
 %do vyberovych percentilu.
end;

X = D2;
Y = X2';
%Test primky metodou nejmenšich ctvercu.
X = [ones(size(X)) X];
b = inv(X'*X)*(X'*Y);
b1 = b(2,1); %Nestranny odhad smernice.
Ye = X*b; %Ocekavana hodnota Y.
e = Y-Ye; %Odhad fitovanych residui.
SCRes = e'*e; %Soucet ctvercu fitovanych residui.
[rb,cb] = size(b);
v2 = n-rb; %Stupne volnosti fitovanych residui.
CMRes = SCRes/v2; %Residua středních ctvercu (nahodny rozptyl).
varb = CMRes*inv(X'*X);
EEb = diag(sqrt(varb));
EEb1 = EEb(2,1); %Standardni chyba odhadu smernice.
t = abs((b1-1)/EEb1); % Pozorovane Studentovo rozdeleni t-statistiky za
predpokladu, ze sklon ma ocekavanou hodnotou 1,0.
P = 1-tcdf(t,v2); %Pravdepodobnost, ze nulova hypotéza je pravdiva

fprintf('-----\n');
disp(' Rozsah výběru Proměnné Směrnice t P ');
fprintf('-----\n');

```

fprintf('%8.i%13.i%14.4f%10.4f%8.4f\n',n,p,b1,t,P);
fprintf('-----\n');
fprintf('Zadaná hladina významnosti je: %.2f\n', alpha);

if P >= alpha;
    fprintf('Předpoklad vícerozměrné normality je prokázán.\n\n');
else
    fprintf('Předpoklad vícerozměrné normality není prokázán.\n\n');
end;
X = X(:,2);
plot(X,Y,'*',Y,Y,'--');
title('Vícerozměrný test normality','FontSize',12);
xlabel('Mahalanobisova vzdálenost D^2');
ylabel('Chi-kvadrat \chi^2');
text(4.2,1.5,['Směrnice = ',num2str(b1),';','n = ',num2str(n),','','p = ',num2str(p)]);
text(5.2,1,['(P = ',num2str(P),')']);
a = [X,Y];
for i = 1:n
    a(i,:) = ginput(1);
    gtext(num2str(a(i,:)),'FontSize',8);
end;
else
X = X(:,2);
plot(X,Y,'*',Y,Y,'--');
title('Vícerozměrný test normality','FontSize',12);
xlabel('Mahalanobisova vzdálenost D^2');
ylabel('Chi-kvadrát \chi^2');
text(4.2,1.5,['Směrnice = ',num2str(b1),';','n = ',num2str(n),','','p = ',num2str(p)]);
text(5.2,1,['(P = ',num2str(P),')']);
end;
clear all

```

PŘÍLOHA P VIII: ANDERSON-DARLINGŮV TEST NORMALITY

DAT

```
function [AnDartest] = AnDartest(x,alpha)

switch nargin
    case{2}
        if isempty(x) == false && isempty(alpha) == false
            if (alpha <= 0 || alpha >= 1)
                fprintf('Pozor: hladina vyznamnosti chyba; musí byt 0 <
alpha < 1 \n');
                return;
            end
        end
    case{1}
        alpha = 0.05;
    otherwise
        error(Pozadovan alespoň jeden vstupni argument.');
```

```
end
n = length(x);
if n < 7,
    disp(Rozsah vyberu musí byt vetsi jak 7.);
    return,
else
    x = x(:);
    x = sort(x);
    fx = normcdf(x,mean(x),std(x));
    i = 1:n;

    S = sum((((2*i)-1)/n)*(log(fx)+log(1-fx(n+1-i))));
    AD2 = -n-S;

    AD2a = AD2*(1 + 0.75/n + 2.25/n^2); %korekcni faktor pro maly rozsah
vyberu: pripad normalniho

    if (AD2a >= 0.00 && AD2a < 0.200);
        P = 1 - exp(-13.436 + 101.14*AD2a - 223.73*AD2a^2);
    elseif (AD2a >= 0.200 && AD2a < 0.340);
        P = 1 - exp(-8.318 + 42.796*AD2a - 59.938*AD2a^2);
    elseif (AD2a >= 0.340 && AD2a < 0.600);
        P = exp(0.9177 - 4.279*AD2a - 1.38*AD2a^2);
    else (AD2a >= 0.600 && AD2a <= 13);
        P = exp(1.2937 - 5.709*AD2a + 0.0186*AD2a^2);
    end
end

disp(' ')
fprintf('Rozsah výběru: %i\n', n);
fprintf('Anderson-Darlingova statistika: %3.4f\n', AD2);
fprintf('Anderson-Darling upravená statistika: %3.4f\n', AD2a);
fprintf('P-hodnota Anderson-Darlingovy statistiky = %3.4f\n', P);
fprintf('Se zvolenou hladinou významnosti = %3.3f\n', alpha);
if P >= alpha;
    disp('Výběr pochází z normálního rozdělení.');
```

```
    s = [mean(x),var(x)];
    fprintf('Takže tento soubor pochází z normálního rozdělení se střední
hodnotou a rozptylem = %6.4f  %8.4f\n',s);
else
    disp('Vyberovy soubor nepochazi z normalniho rozdeleni.');
```

```
end

return
```

PŘÍLOHA P IX: BOX-COXOVA TRANSFORMACE

```
% Box-Coxova transformace
clc;clear all;
pom=0;
pobr=1; kkk = menu('vstup dat','ze souboru','pokus');
if kkk==2
    a=[4,5,7,7,7,8,8.3,8.4,9.4,9.5,10,10.5,12,12.8,13,22,23]';
else
    s1=input('navez souboru:','s'); % jmeno s cislem
    s2=input('pripona:','s'); %bez tecky
    s2=strcat('.',s2);
    s5=num2str(i);
    nam=s1;
    nam1=strcat(nam,s2);
    load(nam1);
    a1=eval(nam);
end
n=length(a);
chtb=chi2inv(.95,1);tkr=tinv(.975,n-1);
i1=ones(n,1);
y1l=zeros(n,1);
y1t=zeros(n,1);
m1=zeros(601,1);
m1l=zeros(601,1);
xp=mean(a);
xg=(mean(log(a)));
xm=median(a);
xg=exp(xg);
y1=(a-i1*xg);
y2=(a-i1*xg).*(a-i1*xg);
y3=y1.*y2;
pom1=3*xg*sum(y1.*y2)+sum(y1.*y3);
pom2=3*sum(y2.*y2)+4*sum(y1.*y3);
lambp=1-pom1/pom2;
sigp=var(a);
sig1=mean((a-xp).^2);
si=mean((a-xp).^3);
sil=si/(sigp*sqrt(sigp));
sp=mean((a-xp).^4);
spl=sp/(sigp*sigp);
lamb=(sp-3*sig1*sig1)+3*sig1*si/xp+9*sig1*sig1*sig1/(4*xp*xp);
lamb=xp*si-lamb/3;
lal=7*(sp-3*sig1*sig1)+12*sig1*si/xp+6*sig1*sig1*sig1/(xp*xp);
lal=6*sig1*sig1+lal/3;
lamb=lamb/lal;
lamb=1-lamb;
lamba=1-(xp*si)/(6*sig1*sig1);
fprintf(1,'*****.\n');
fprintf('Puvodni data.\n');
fprintf('Prumer arit.=%g.\n',xp);
fprintf('Prumer geom.=%g.\n',xg);
fprintf('Median=%g.\n',xm);
fprintf('Rozptyl=%g.\n',sigp);
fprintf('Sikmost=%g.\n',sil);
fprintf('Spicatost=%g.\n',spl);
fprintf('Odhad lambda prec.=%g.\n',lambp);
fprintf('Odhad lambda=%g.\n',lamb);
fprintf('Odhad lambda rough=%g.\n',lamba);
fprintf('%g.\n');
fprintf(1,'*****.\n');
jp=sum(log(a));
cr=-999999999999;
crl=-999999999999;
```

```

lam=-99;
las=-99;
for i=1:n,
    pi=i/(n+1);
    zi(i)=norminv(pi,0,1);
end
for i=1:601,
    j=0.01*i-3.01;
    j=round(100*j)/100;
    if j==0
        z=log(a);
    else
        z=(a.^j-1)/j;
    end
    pom=var(z);
    pom=log(pom);
    ml(i)=-0.5*n*pom+(j-1)*jp;
    if ml(i)>=cr
        cr=ml(i);
        lam=j;
        ylt=z;
    end
    zl=(z-mean(z))/sqrt(var(z));
    zl=sort(zl);
    m11(i)=(zl.*zi)/(zi*zi.);
    if m11(i)>=cr1
        cr1=m11(i);
        y11=z1;
        las=j;
    end
end
dm=cr-0.5*chtb;
hm=0;
hmm=3;
dmm=-3;
for i=1:601,
    if ml(i)>dm && hm==0
        dmm=round(100*(0.01*i-3.01))/100;
        hm=1;
    end
    if ml(i)<dm && hm==1
        hmm=round(100*(0.01*i-3.01))/100;
        hm=2;
    end
end
xpt=mean(y1t);
sigt=var(y1t);
sit=mean((y1t-xpt).^3);
sit=sit/(sigt*sqrt(sigt));
spt=mean((y1t-xpt).^4);
spt=spt/(sigt*sigt);
fprintf('Optim lambda Box Cox-MLE %g.\n',lam);
fprintf('Konf. interval%g.%g.\n',dmm,hmm);
fprintf('Transformace.\n');
fprintf('Prumer=%g.\n',xpt);
fprintf('Rozptyl=%g.\n',sigt);
fprintf('Sikmost=%g.\n',sit);
fprintf('Spicatost=%g.\n',spt);
if lam==0
    xer=exp(xpt+.5*sigt);
    ser=xer*xer*sigt;
    ind=exp(xpt+.5*sigt-tkr*sqrt(sigt/n));
    inh=exp(xpt+.5*sigt+tkr*sqrt(sigt/n));
else
    pom=.5*sqrt(1+2*lam*(xpt+sigt)+lam*lam*(xpt^2-2*sigt));

```

```

    xr1=(.5*(1+lam*xpt)+pom)^(1/lam);
    xr2=(.5*(1+lam*xpt)-pom)^(1/lam);
    pom=median(y1t);poml=min(xr1-pom,xr2-pom);
    xer=poml+pom;
    ser=sigt*xer^(-2*lam+2);pom=xpt-.5*(lam-1)*sigt*xer^(-lam);
    ind=(1+lam*(pom-tkr*sqrt(sigt/n)))^(1/lam);
    inh=(1+lam*(pom+tkr*sqrt(sigt/n)))^(1/lam);
end
fprintf('Re transformace.\n');
fprintf('Prumer=%g.\n',xer);
fprintf('Rozptyl=%g.\n',ser);
fprintf('Dmez=%g.\n',ind);
fprintf('Hmez=%g.\n',inh);
fprintf('%g.\n');
fprintf(1,'*****.\n');
fprintf('Optim lambda Box Cox-SW %g.\n',las);
fprintf('smernice %g.\n',cr1);

if las==0
    xer=exp(xpt+.5*sigt);
    ser=xer*xer*sigt;
    ind=exp(xpt+.5*sigt-tkr*sqrt(sigt/n));
    inh=exp(xpt+.5*sigt+tkr*sqrt(sigt/n));
else
    pom=.5*sqrt(1+2*las*(xpt+sigt)+las*las*(xpt^2-2*sigt));
    xr1=(.5*(1+las*xpt)+pom)^(1/las);
    xr2=(.5*(1+las*xpt)-pom)^(1/las);
    pom=median(y1t);poml=min(xr1-pom,xr2-pom);
    xer=poml+pom;
    ser=sigt*xer^(-2*las+2);pom=xpt-.5*(las-1)*sigt*xer^(-las);
    ind=(1+las*(pom-tkr*sqrt(sigt/n)))^(1/las);
    inh=(1+las*(pom+tkr*sqrt(sigt/n)))^(1/las);
end
fprintf('Re transformace.\n');
fprintf('Prumer=%g.\n',xer);
fprintf('Rozptyl=%g.\n',ser);
fprintf('Dmez=%g.\n',ind);
fprintf('Hmez=%g.\n',inh);
fprintf('%g.\n');
fprintf(1,'*****.\n');

a=[4,5,7,7,7,8,8.3,8.4,9.4,9.5,10,10.5,12,12.8,13,22,23]';
b=a';
% Graf normality puvodnich dat
%figure('Name','Pstni graf puvodnich dat');
%normplot(a);
figure('Name','graf Q-Q puvodnich dat');
x1 = sort(b);
prum = mean(b);
rsm = sqrt(var(b));
iv = 1:size(b,2);
pi = iv./(size(b,2)+1);
plot(pi,x1,'k*');
di = .1.*max(diff(x1));
axis([.01 .99 x1(1)-di x1(size(b,2))+di]);
x2 = .01:.01:.99;
x25 = prctile(b,25);x50 = prctile(b,50);x75 = prctile(b,75); rf = (x75-
x25)/1.349;
for i = 1:size(x2,2)
nt(i) = norminv(x2(i),prum,rsm);
ntr(i) = norminv(x2(i),x50,rf);
end
hold on
plot(x2,nt,'k-');
plot(x2,ntr,'r-')

```

```

hold off
grid on
title('Kvantilovy graf puvodnich dat'),xlabel('Poradova pravdepodobnost
Pi');ylabel('Poradkova statistika x(i)');
% Grafy normality pro transformovana data
hold on
figure('Name','graf Q-Q');
plot(y11,zi,'k*');
hold on
plot(cr1*zi,zi,'-')
hold off
grid on
title('graf Q-Q'),xlabel('kvantil normalniho rozd. ');ylabel('poradkova
statistika');
pobr=pobr+1;
figure('Name','graf MLE');
plot(-3:0.01:3,m1, '.')
hold on
plot(-3:0.1:3,dm,'k*')
%hold off
plot([lam lam],[min(m1) cr], 'r^-')
grid on
title('graf MLE');xlabel('lambda');ylabel('MLE');
%end

```