

# VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ  
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION  
DEPARTMENT OF TELECOMMUNICATIONS

PARAMETRICKÉ A NEPARAMETRICKÉ STATISTICKÉ METODY A  
JEJICH APLIKACE

BAKALÁŘSKÁ PRÁCE  
BACHELOR'S THESIS

AUTOR PRÁCE  
AUTHOR

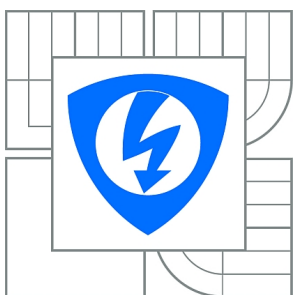
PETER MRAČKO

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH  
TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION  
DEPARTMENT OF TELECOMMUNICATIONS

## PARAMETRICKÉ A NEPARAMETRICKÉ STATISTICKÉ METODY A JEJICH APLIKACE

PARAMETRIC AND NON-PARAMETRIC STATISTICAL METHODS AND THEIR APPLICATION

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

PETER MRAČKO

VEDOUCÍ PRÁCE

SUPERVISOR

prof. Ing. ZDENĚK SMÉKAL, CSc.

BRNO 2013



VYSOKÉ UČENÍ  
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky  
a komunikačních technologií

Ústav telekomunikací

# Bakalářská práce

bakalářský studijní obor  
Teleinformatika

**Student:** Peter Mračko

**ID:** 138959

**Ročník:** 3

**Akademický rok:** 2012/2013

## NÁZEV TÉMATU:

**Parametrické a neparametrické statistické metody a jejich aplikace**

## POKYNY PRO VYPRACOVÁNÍ:

Parametrické statistické testy se zabývají testováním hypotéz o parametrech popř. o parametrických funkcích jednotlivých rozdělení pravděpodobnosti. Neparametrické metody zkoumají obecnější vlastnosti náhodných jevů. Cílem bakalářské práce se popsat a ověřit na příkladech z praxe nejznámější parametrické a neparametrické statistické metody.

## DOPORUČENÁ LITERATURA:

- [1] HÁTLE, J., LIKEŠ, J.: Základy počtu pravděpodobnosti a matematické statistiky. SNTL/ALFA, Praha, Bratislava, 1974. 04-311-74
- [2] ANDĚL, J.: Matematická statistika. SNTL, Praha, 1985. 04-03-85
- [3] SMÉKAL, Z.: Signály a soustavy (BASS). Elektronické texty VUT v Brně.

**Termín zadání:** 11.2.2013

**Termín odevzdání:** 5.6.2013

**Vedoucí práce:** prof. Ing. Zdeněk Smékal, CSc.

**Konzultanti bakalářské práce:**

**prof. Ing. Kamil Vrba, CSc.**

*Předseda oborové rady*

## UPOZORNĚNÍ:

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

## **ABSTRAKT**

Práca sa zaoberá parametrickými štatistickými metódami, ktoré spočívajú v testovaní hypotéz o parametroch jednotlivých rozdelení pravdepodobnosti. V práci sú teoreticky vysvetlené vybrané rozdelenia pravdepodobnosti a princípy testovania ich parametrov. Teoretické poznatky sú vysvetlené na konkrétnych príkladoch, ktoré sú riešené pomocou programu Matlab. Praktická časť spočíva vo využití metód pre štatistické vyhodnotenie parametrov pacientov, ktorí sú postihnutí Alzheimerovou chorobou.

## **KĽÚČOVÉ SLOVÁ**

parametrické štatistické metódy, testovanie hypotéz parametrov, parametre pravdepodobnostných rozdelení

## **ABSTRACT**

This thesis deals with parametric statistical methods, which consist in testing hypotheses about the parameters of probability distributions. The thesis theoretically explains chosen probability distribution principles and test their parameters. Theoretical knowledge is explained with real examples that are solved using Matlab. The practical part consists in the use of statistical methods for evaluating parameters of patients who are handicapped by Alzheimer's disease.

## **KEYWORDS**

parametric statistical methods, hypothesis testing parameters, parameters of probability distribution

MRAČKO, Peter *Parametrické a neparametrické statistické metódy a jejich aplikace*: bakalárska práca. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2013. 54 s. Vedúci práce bol prof. Ing. Zdeněk Smékal, CSc.

## PREHLÁSENIE

Prehlasujem, že som svoju bakalársku prácu na tému „Parametrické a neparametrické statistické metódy a jejich aplikace“ vypracoval samostatne pod vedením vedúceho bakalárskej práce, využitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú všetky citované v práci a uvedené v zozname literatúry na konci práce.

Ako autor uvedenej bakalárskej práce ďalej prehlasujem, že v súvislosti s vytvorením tejto bakalárskej práce som neporušil autorské práva tretích osôb, najmä som nezasiahol nedovoleným spôsobom do cudzích autorských práv osobnostných a/nebo majetkových a som si plne vedomý následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona č. 121/2000 Sb., o právu autorskom, o právach súvisejúcich s právom autorským a o zmene niektorých zákonov (autorský zákon), vo znení neskorších predpisov, vrátane možných trestnoprávných dôsledkov vyplývajúcich z ustanovenia časti druhej, hlavy VI. diel 4 Trestného zákoníka č. 40/2009 Sb.

Brno .....

.....

(podpis autora)

## POĎAKOVANIE

Rád by som poďakoval vedúcemu bakalárskej práce pánovi prof. Ing. Zdenku Smékalovi, CSc. za odborné vedenie, konzultácie, trpezlivosť a podnetné návrhy k práci.

Brno .....

.....

(podpis autora)

# OBSAH

<b>1</b>	<b>Úvod</b>	<b>10</b>
1.1	Postup práce . . . . .	10
<b>2</b>	<b>Úvod do štatistiky</b>	<b>11</b>
2.1	Základné pojmy štatistiky . . . . .	11
<b>3</b>	<b>Popisná štatistika</b>	<b>12</b>
3.1	Početnosť štatistického znaku . . . . .	12
3.2	Charakteristika polohy . . . . .	12
3.2.1	Priemery . . . . .	12
3.2.2	Modus a kvantilové miery . . . . .	13
3.3	Charakteristika variability . . . . .	13
<b>4</b>	<b>Náhodná veličina</b>	<b>15</b>
4.1	Diskrétna náhodná veličina . . . . .	15
4.2	Spojité náhodné veličiny . . . . .	16
<b>5</b>	<b>Rozdelenia pravdepodobnosti pre diskkrétne náhodné veličiny</b>	<b>17</b>
<b>6</b>	<b>Rozdelenia pravdepodobnosti pre spojité náhodné veličiny</b>	<b>19</b>
<b>7</b>	<b>Odhady charakteristík súboru</b>	<b>22</b>
7.1	Bodové a intervalové odhady parametrov . . . . .	22
<b>8</b>	<b>Testovanie štatistických hypotéz</b>	<b>24</b>
8.1	Princípy testovania štatistických hypotéz . . . . .	24
8.2	Testy parametrov normálneho rozdelenia . . . . .	25
8.3	Test parametru alternatívneho rozdelenia . . . . .	27
<b>9</b>	<b>Analýza rozptylu</b>	<b>28</b>
9.1	Testy normality . . . . .	28
9.2	Test zhody rozptylov súborov s normálnym rozdelením . . . . .	30
9.3	Logaritmickej transformácia . . . . .	31
<b>10</b>	<b>Neparametrické testy</b>	<b>33</b>
10.1	Wilcoxonov dvojvýberový test . . . . .	33
10.2	Neparametrická alternatíva analýzy rozptylu - Kruskalov-Wallisov test	33
<b>11</b>	<b>Praktická časť</b>	<b>35</b>

<b>12 Závěr</b>	<b>38</b>
<b>Literatúra</b>	<b>40</b>
<b>Zoznam príloh</b>	<b>41</b>
<b>A Príklady</b>	<b>42</b>
<b>B Zdrojový kód v Matlabe</b>	<b>54</b>

## ZOZNAM OBRÁZKOV

9.1	Q-Q plot výberu s normálnym rozdelením . . . . .	29
9.2	Q-Q plot výberu, ktorý nepochádza z normálneho rozdelenia . . . . .	30
9.3	Q-Q plot transformovaných dát . . . . .	32
11.1	Krabicový graf pre 24. parameter každej skupiny . . . . .	36
A.1	Koláčový graf relatívnej početnosti . . . . .	42
A.2	Stĺpcový graf absolútnej početnosti . . . . .	43
A.3	Pravdepodobnostná funkcia rozdelenia $Po(3)$ . . . . .	46
A.4	Pravdepodobnostná funkcia rozdelenia $B(6; 0, 86)$ . . . . .	48

## ZOZNAM TABULIEK

5.1	Charakteristiky Poissonovho rozdelenia . . . . .	17
5.2	Charakteristiky alternatívneho rozdelenia . . . . .	17
5.3	Charakteristiky binomického rozdelenia . . . . .	18
6.1	Charakteristiky rovnomerného rozdelenia . . . . .	19
6.2	Charakteristiky exponenciálneho rozdelenia . . . . .	20
6.3	Charakteristiky normálneho rozdelenia . . . . .	20
6.4	Charakteristiky normovaného normálneho rozdelenia . . . . .	21
8.1	Kritické obory, pri teste strednej hodnoty normálneho rozdelenia . . .	26
8.2	Kritické obory, pri teste rozptylu normálneho rozdelenia . . . . .	27
8.3	Kritické obory, pri teste alternatívneho rozdelenia . . . . .	27
11.1	Parametre s normálnym rozdelením . . . . .	35
11.2	Logaritmované parametre s normálnym rozdelením . . . . .	35
11.3	Signifikantné parametre zistené Wilcoxonovým testom . . . . .	36
11.4	Názvy signifikantných parametrov . . . . .	37
12.1	Názvy signifikantných parametrov . . . . .	38
A.1	Hodnoty pravdepodobnostnej a distribučnej funkcie pre $x = 0, 1, \dots, 8$	45

# 1 ÚVOD

## 1.1 Postup práce

Aby sme sa mohli zaoberať parametrickými štatistickými testami jednotlivých rozdelení pravdepodobností, musíme ich najprv pochopiť. Jednotlivé rozdelenia pravdepodobností sú charakterizované jedným alebo viacerými parametrami.

Cieľom tejto práce je zoznámiť sa so základnými pojmami štatistiky, ozrejmiť jednotlivé modely rozdelenia pravdepodobnosti pre diskrétne a spojité náhodné veličiny a vysvetliť princíp testovania hypotéz o parametroch. Praktické demonštrovanie príkladov je ukázané v programe Matlab, kde sú aj vysvetlené všetky použité funkcie.

Praktická časť práce spočíva vo využití metód pre štatistické vyhodnotenie parametrov pacientov, ktorí sú postihnutí Alzheimerovou chorobou. Jedná sa o cvičenia, kde pacient kreslí rôzne obrázky, píše úryvky textu a podobne, z týchto cvičení sú získavané parametre. Hlavnou úlohou je medzi týmito parametrami nájsť také, ktoré by dokázali signifikantne oddeliť skupinu zdravých, skupinu s ľahkým ochorením a skupinu s ťažkou formou ochorenia. V tomto prípade máme k dispozícii dáta kde je 79 pacientov, u ktorých bolo meraných 29 parametrov.

Tieto dáta boli získané v spolupráci s univerzitou v meste Haifa v Izraeli. Vzhľadom nato, že sú dôverné nemôžu byť v tejto práci uvedené.

## 2 ÚVOD DO ŠTATISTIKY

### 2.1 Základné pojmy štatistiky

Štatistický súbor je konečná neprázdna množina  $M$ . Je to skupina predmetov, vecí a podobne, ktoré sú zhromaždené na základe ich spoločných vlastností alebo znakov. Štatistická jednotka je základný prvok štatistického súboru.

Rozsah súboru (označenie  $n$ ) je počet všetkých prvkov množiny  $M$  a teda počet všetkých prvkov štatistického súboru. Znak súboru je spoločná vlastnosť jednotlivých prvkov súboru, ktorého zmeny sú predmetom skúmania. Napríklad počítače sú súborom a ich základné funkcie, ktoré sú u každého iné, sú znakom súboru. Hodnoty znaku sú štandardne označované  $x_1, x_2, \dots, x_n$ . Sú to jednotlivé údaje znaku. Môžeme ich vyjadriť kvantitatívne alebo kvalitatívne. Kvantitatívne znamená číslom. Je to ľubovoľná funkcia  $f$  (štandardne označovaná písmenom  $x$ ), ktorá každému prvku množiny  $M$  štatistického súboru priradí práve jedno reálne číslo. Kvalitatívne, čiže slovným popisom. Aspoň jedna hodnota tohto znaku nesmie byť reálne číslo.

Štatistické vyšetrenie resp. výskum je vyšetrenie hodnôt znaku a ich následne spracovanie danými štatistickými metódami. Štatistické triedenie je rozklad štatistického súboru na triedy podľa znaku. Trieda je množina jednotiek štatistického súboru, ktorým je priradená určitá hodnota alebo určitý interval znaku.[8]

## 3 POPISNÁ ŠTATISTIKA

V tejto kapitole budeme popisovať namerané a zistené dáta, ukážeme ako sa dajú spracovať a vyhodnotiť. Základným roztriedením dát je ich usporiadanie podľa veľkosti.

### 3.1 Početnosť štatistického znaku

Pri spracovaní štatistických údajov sa väčšinou zameriavame na tzv. početnosť štatistického znaku. Zavádzame pojmy absolútna a relatívna početnosť. Absolútna početnosť je daná počtom výskytu jednotlivého znaku v štatistickom súbore. Relatívna početnosť určuje podiel konkrétnej obmeny na celkovom počte nameraných alebo zistených hodnôt. Pri spracovaní jednotlivých početností využívame tabuľky a rôzne grafy, podľa jednotlivého typu dát (napr. koláčový alebo stĺpcový graf).<sup>1</sup>

### 3.2 Charakteristika polohy

Úroveň javu vyjadreného číselným znakom sa popisuje charakteristikou polohy. Táto miera meria obecnú veľkosť hodnôt znaku v súbore.

#### 3.2.1 Priemery

Snáď najznámejšou charakteristikou polohy je aritmetický priemer daný vzťahom

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (3.1)$$

kde  $n$  je celkový počet pozorovaní. Základné vlastnosti aritmetického priemeru sú[3]:

1.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0, \quad (3.2)$$

slovom vyjadrené to znamená, že súčet odchýliek od aritmetického priemeru je rovný nule.

2.

$$\frac{1}{n} \sum_{i=1}^n c = c, \quad (3.3)$$

ak sú jednotlivé hodnoty znaku  $x_i$  všetky rovnaké, potom ich aritmetický priemer je rovný tejto konštante.

---

<sup>1</sup>Pozri príloha príklad 1.

3.

$$\frac{1}{n} \sum_{i=1}^n (x_i + c) = c + \bar{x}, \quad (3.4)$$

ak pripočítame ku každej hodnote znaku  $x_i$  konštantu  $c$ , zvýši sa aj aritmetický priemer o túto konštantu.

4.

$$\frac{1}{n} \sum_{i=1}^n c \cdot x_i = c \cdot \bar{x}, \quad (3.5)$$

ak násobíme každú hodnotu znaku  $x_i$  konštantou  $c$ , je aj aritmetický priemer násobený touto konštantou.

V špeciálnych prípadoch sa používajú aj iné priemery ako aritmetický. Poznáme harmonický, geometrický a kvadratický priemer.

### 3.2.2 Modus a kvantilové miery

Modus  $\hat{x}$  je hodnota znaku s najväčšou početnosťou. Rozoznávame jednomodálne rozdelenie, tj. v prípade, kde je jednoznačne určený jeden modus alebo jeden modálny interval. V praxi sa môžeme stretnúť s dátovými súbormi, ktoré obsahujú viac ako jeden modus.

Kvantil  $x_p$  je hodnota znaku, pre ktorú platí, že  $100p\%$  jednotiek usporiadaného súboru má hodnotu menšiu alebo rovnú  $x_p$  a  $100(1-p)\%$  jednotiek má hodnotu väčšiu alebo rovnú  $x_p$ .

Kvantil  $x_{0,50}$  sa nazýva medián (často sa značí  $\tilde{x}$ ), kvantil  $x_{0,75}$  je horný kvartil, kvantil  $x_{0,25}$  je dolný kvartil. Kvantily  $x_{0,10}, x_{0,20}, \dots, x_{0,90}$  sa nazývajú decily a kvantily  $x_{0,01}, x_{0,02}, \dots, x_{0,99}$  sa nazývajú percentily.

## 3.3 Charakteristika variability

Charakteristiky polohy nám podávajú informácie len o polohe rozdelenia početnosti. My sa môžeme stretnúť s prípadmi, kde rozdelenia budú mať rovnakú polohu ale, aj napriek tomu budú odlišné. Stáva sa to v prípadoch, keď sa rozdelenia líšia variabilitou (premenlivosťou) dát. Najjednoduchšia charakteristika variability je variačné rozpätie  $R[3]$ . Je to rozdiel najväčšej a najmenej hodnoty znaku,

$$R = x_{max} - x_{min}. \quad (3.6)$$

V predchádzajúcej podkapitole sme sa zoznámili s pojmami horný a dolný kvartil, decil a percentil. S využitím týchto pojmov definujeme ďalšie miery variability, tj. kvantilové rozpätia

1. kvartilové rozpätie  $R_Q = x_{0,75} - x_{0,25}$

2. decilové rozpätie  $R_D = x_{0,90} - x_{0,10}$
3. percentilové rozpätie  $R_C = x_{0,99} - x_{0,01}$ .

Kvantilové rozpätia udávajú veľkosť intervalu, v ktorom leží určitá časť usporiadaných dát. Prakticky kvartilové rozpätie udáva dĺžku intervalu, v ktorom leží 50% prostredných hodnôt usporiadaného súboru. Podobne decilové 80% a percentilové 98%.

Ďalšou mierou variability je priemerná odchýlka  $\bar{d}_{\bar{x}}$  hodnôt  $x_1, x_2, \dots, x_n$  a je definovaná ako aritmetický priemer absolútnych odchýliek jednotlivých hodnôt znaku od aritmetického priemeru

$$\bar{d}_{\bar{x}} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|, \quad (3.7)$$

kde  $n$  je celkový počet pozorovaní.[3]

Ďalšou používanou mierou charakteristiky je rozptyl  $s_n^2$ , ktorý je daný aritmetický priemer štvorcov odchýliek jednotlivých hodnôt znaku od aritmetického priemeru

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (3.8)$$

kde  $n$  je celkový počet pozorovaní.

Odmocnina z rozptylu sa nazýva smerodajná odchýlka  $s_n$ ,

$$s_n = \sqrt{s_n^2}. \quad (3.9)$$

Jedná sa o kvadratický priemer odchýliek hodnôt znaku od ich aritmetického priemeru. Smerodajná odchýlka je narozdiel od rozptylu vyjadrená v rovnakých jednotkách ako sledovaný znak. <sup>2</sup>[5]

---

<sup>2</sup>Pozri príloha príklad 2.

## 4 NÁHODNÁ VELIČINA

Ak nie je výsledok ľubovlného pokusu alebo deja jednoznačne určený podmienkami, za ktorých sa odohráva, považujeme možné výsledky za priestor elementárnych javov  $\Omega$ . Jednotlivé možné výsledky budeme označovať  $\omega$  s prípadným indexom. O náhodnej veličine sa často hovorí ako o veličine, ktorá nadobúda svoje hodnoty vplyvom náhody. Náhodná veličina je reálna funkcia  $X(\omega)$  definovaná na množine elementárnych javov  $\Omega$ . Každému elementárnemu javu  $\omega$  z množiny elementárnych javov  $\Omega$  priraduje práve jedno reálne číslo  $X(\omega) = x$ . Obor hodnôt náhodnej veličiny  $X$  je množina  $M = x = X(\omega) : \omega \in \Omega$ .

Napríklad počet výrobkov vo vybranej dávke je 100 ks, náhodná veličina  $X$  označuje počet vadných výrobkov a môže nadobúdať hodnoty  $x = 0, 1, 2, \dots, 100$ . Zápisom  $X = x_0$  rozumieme, že náhodná veličina  $X$  nadobúda hodnotu  $x_0$ , zápisom  $x_1 < X < x_2$  vyjadrujeme, že náhodná veličina  $X$  nadobúda hodnoty z intervalu  $(x_1, x_2)$  a pod.

### 4.1 Diskrétna náhodná veličina

Náhodná veličina  $X$  sa nazýva diskrétna, ak jej obor hodnôt obsahuje len konečne veľa hodnôt, alebo aj nekonečne veľa hodnôt, ale tieto hodnoty sa dajú zoradiť do postupnosti. Hodnoty, ktoré diskrétna náhodná veličina môže nadobúdať, označíme  $x_1, x_2, \dots$  a ich počet označíme  $n$ , pričom  $n$  môže byť aj  $\infty$ . [4]

Pravdepodobnostná funkcia  $p(x_i)$  diskkrétnej náhodnej veličiny  $X$  priraduje každému reálnemu číslu  $x$  pravdepodobnosť, že náhodná veličina nadobudne túto hodnotu, teda [3]

$$p(x_i) = P(X = x_i). \quad (4.1)$$

Pre hodnoty pravdepodobnostnej funkcie platí

$$\sum_{i=1}^n p(x_i) = 1. \quad (4.2)$$

Pre každé reálne číslo  $x_i$  platí  $0 \leq p(x_i) \leq 1$ .

K popisu náhodných veličín sa využíva distribučná funkcia. Distribučná funkcia náhodnej veličiny  $X$  je funkcia  $F$ , ktorá je definovaná ako

$$F(x_i) = P(X \leq x_i). \quad (4.3)$$

Znamená to, že hodnota funkcie  $F$  v bode malé  $x_i$  je rovná pravdepodobnosti, že náhodná veličina  $X$  nadobudne hodnoty menšie než  $x_i$ , tj. hodnoty z intervalu  $(-\infty, x_i)$ . Distribučná funkcia má nasledujúce vlastnosti [3]:

1. Je neklesajúca, zľava spojitá.
2. Pre každé reálne číslo  $x_i$  platí  $0 \leq F(x_i) \leq 1$ .
3. Pre každú distribučnú funkciu platí

$$\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1.$$

Pre distribučnú funkciu diskkrétnej náhodnej veličiny platí

$$F(x_i) = P(X \leq x_i) = \sum_{t \in (-\infty, x_i] \cap \Omega} p(t) \quad (4.4)$$

Distribučná funkcia u diskkrétnej náhodnej veličiny má schodový charakter, jedná sa o funkciu, ktorá je po častiach konštantná.

## 4.2 Spojité náhodné veličiny

U spojitej náhodnej veličiny sa pravdepodobnosť, že náhodná veličina  $X$  padne do určitého intervalu  $(a, b)$ , počíta ako

$$P(X \in (a, b)) = \int_a^b f(x) dx, \quad (4.5)$$

kde funkcia  $f$  je tzv. hustota pravdepodobnosti náhodnej veličiny  $X$ . [4]

Funkcia hustoty pravdepodobnosti spojitej náhodnej veličiny  $X$  je nezáporná funkcia  $f(x)$  taká, že [3]

$$F(x) = \int_{-\infty}^x f(t) dt, \quad x \in \mathbb{R}. \quad (4.6)$$

## 5 ROZDELENIA PRAVDEPODOBNOTI PRE DISKRÉTNE NÁHODNÉ VELIČINY

### 1. Poissonove rozdelenie

Poissonove rozdelenie používame v prípade, keď sa pýtame na pravdepodobnosť, že nastane určitý počet udalostí za jednotku času, alebo pravdepodobnosť výskytu istých objektov v geometrickej oblasti o pevnej veľkosti  $v$ . K udalostiam resp. výskytom dochádza jednotlivo a nezávisle na sebe. Náhodná veličina  $X$  má Poissonove rozdelenie  $Po(\lambda)$ , práve keď má pravdepodobnostná funkcia tvar [3]

$$p(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{pro } x = 0, 1, 2, \dots, \\ 0 & \text{jinak.} \end{cases} \quad (5.1)$$

Skutočnosť, že náhodná veličina  $X$  má Poissonove rozdelenie s parametrom  $\lambda$ , zapisujeme  $X \sim Po(\lambda)$ . Vybrané číselné charakteristiky Poissonovho rozdelenia sú zobrazené v tab. 5.1.<sup>1</sup>

Tab. 5.1: Charakteristiky Poissonovho rozdelenia

$E(X)$	$D(X)$	$Mo(X)$
$\lambda$	$\lambda$	$\lambda - 1 \leq Mo(X) \leq \lambda$

### 2. Alternatívne rozdelenie

Ak je experiment takej povahy, že môžu nastať len dva rôzne výsledky, ktoré sa navzájom vylučujú: pokus je úspešný a pokus je neúspešný, náhodná veličina udávajúca počet úspechov v jednom pokuse sa nazýva alternatívna a označuje sa  $A(p)$ . Pravdepodobnosť úspechu je daná parametrom  $p$  ( $0 < p < 1$ ). Náhodná veličina  $X$  má alternatívne rozdelenie  $A(p)$ , práve keď má pravdepodobnostná funkcia tvar [3]

$$p(x) = \begin{cases} p^x(1-p)^{1-x} & \text{pro } x = 0, 1, \\ 0 & \text{jinak.} \end{cases} \quad (5.2)$$

Vybrané číselné charakteristiky alternatívneho rozdelenia zobrazíme v tab. 5.2.

Tab. 5.2: Charakteristiky alternatívneho rozdelenia

$E(X)$	$D(X)$
$p$	$p(1-p)$

<sup>1</sup>Pozri príloha príklad 3.

### 3. Binomické rozdelenie

Náhodná veličina, ktorej pravdepodobnostné chovanie môžeme charakterizovať pomocou binomického rozdelenia, udáva počet úspechov v postupnosti  $n$  nezávislých alternatívnych pokusov, pričom úspech nastáva s pravdepodobnosťou  $p$  ( $0 < p < 1$ ). Náhodná veličina má binomické rozdelenie  $B$  pravdepodobnosti s parametrami  $n, p$  a nadobúda hodnoty z množiny  $0, 1, 2, \dots, n$  s pravdepodobnosťou [3]

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{pro } x = 0, 1, \dots, n, \\ 0 & \text{jinak.} \end{cases} \quad (5.3)$$

Alternatívne rozdelenie je špeciálnym prípadom binomického rozdelenia pre  $n = 1$ .

Vybrané číselné charakteristiky binomického rozdelenia zobrazíme v tab. 5.3.<sup>2</sup>

Tab. 5.3: Charakteristiky binomického rozdelenia

$E(X)$	$D(X)$	$Mo(X)$
$np$	$np(1-p)$	$(n+1)p - 1 \leq Mo(X) \leq (n+1)p$

---

<sup>2</sup>Pozri príloha príklad 4.

## 6 ROZDELENIA PRAVDEPODOBNOTI PRE SPOJITÉ NÁHODNÉ VELIČINY

### 1. Rovnomerné rozdelenie

Rovnomerné rozdelenie[3] sa používa vtedy, keď má náhodná veličina konštantnú hustotu pravdepodobnosti na intervale  $\alpha, \beta$ , kde  $\alpha, \beta$  sú reálne čísla. Spojitá náhodná veličina  $X$  má rovnomerné rozdelenie  $R(\alpha, \beta)$ , práve keď funkcia hustoty pravdepodobnosti má tvar

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{pro } \alpha < x < \beta, \\ 0 & \text{jinak.} \end{cases} \quad (6.1)$$

Distribučná funkcia je popísaná rovnicami

$$F(x) = \begin{cases} 0 & \text{pro } x \leq \alpha, \\ \frac{x-\alpha}{\beta-\alpha} & \alpha < x < \beta, \\ 1 & x \geq \beta. \end{cases} \quad (6.2)$$

Vybrané číselné charakteristiky rovnomerného rozdelenia zobrazíme v tab. 6.1.

Tab. 6.1: Charakteristiky rovnomerného rozdelenia

$E(X)$	$D(X)$	kvantily $x_p$	$Me(X)$
$\frac{\alpha+\beta}{2}$	$\frac{1}{12}\beta - \alpha$	$\alpha + P(\beta - \alpha)$	$\frac{\alpha+\beta}{2}$

### 2. Exponenciálne rozdelenie

Náhodná veličina s exponenciálnym rozdelením [4] popisuje dobu čakania na ďalšiu udalosť alebo dĺžku časového intervalu medzi dvoma udalosťami, ak priemerne nastáva  $\lambda$  udalostí za jednotku času. Exponenciálne rozdelenie  $E$  je spojitým náprotivkom k Poissonovmu rozdeleniu  $Po(\lambda)$ . Spojitá náhodná veličina  $X$  má exponenciálne rozdelenie s parametrom  $\lambda > 0$ , ak funkcia hustoty pravdepodobnosti má tvar

$$f(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \lambda \cdot e^{-\lambda x} & x \geq 0. \end{cases} \quad (6.3)$$

Distribučná funkcia je popísaná rovnicami

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 1 - e^{-\lambda x} & x \geq 0. \end{cases} \quad (6.4)$$

Vybrané číselné charakteristiky exponenciálneho rozdelenia sú uvedené v tab. 6.2.

Tab. 6.2: Charakteristiky exponenciálneho rozdelenia

$E(X)$	$D(X)$
$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

### 3. Normálne rozdelenie

Normálne rozdelenie [3][4] sa používa vtedy, keď je kolísanie náhodnej veličiny spôsobené súčtom veľkého počtu vzájomne nezávislých vplyvov. Skutočnosť, že náhodná veličina má normálne rozdelenie so strednou hodnotou  $\mu$  a rozptylom  $\sigma^2$  zapisujeme ako  $X \sim N(\mu, \sigma^2)$ . Spojitá náhodná veličina  $X$  má tvar normálneho rozdelenia  $N(\mu, \sigma^2)$ , práve keď funkcia hustoty má tvar

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ pre } x \in \mathbb{R}. \quad (6.5)$$

Distribučná funkcia je definovaná rovnicou

$$F(x) = \int_{-\infty}^x f(t)dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \text{ pre } x \in \mathbb{R}. \quad (6.6)$$

Vybrané číselné charakteristiky normálneho rozdelenia zobrazíme v tab. 6.3.

Tab. 6.3: Charakteristiky normálneho rozdelenia

$E(X)$	$D(X)$	kvantily $x_p$	$Me(X)$	$Mo(X)$
$\mu$	$\sigma^2$	$\mu + \sigma u_p$	$\mu$	$\mu$

Normálne rozdelenie je významné, pretože za určitých podmienok formovaných centrálnou limitnou vetou sa k nemu blížia iné spojité a diskkrétne rozdelenia. Ak  $X_1, X_2, \dots, X_N$  sú nezávislé náhodné veličiny, ktoré majú všetky rovnaké rozdelenie so strednou hodnotou  $\mu$  a rozptylom  $\sigma^2$  tak pre súčtom týchto veličín je náhodná veličina  $X$ , pre ktorú platí:

$$E(X) = n\mu \quad D(X) = n\sigma^2, \quad (6.7)$$

pre priemer  $\bar{X}$  patí:

$$E(\bar{X}) = E\left(\frac{X}{n}\right) = \mu \quad D(\bar{X}) = D\left(\frac{X}{n}\right) = \frac{\sigma^2}{n} \quad (6.8)$$

Výpočet distribučnej funkcie  $F(x)$  normálneho rozdelenia sa prevádza len numericky, pretože Gaussov integrál nie je analyticky vyjadriteľný. Ak chceme stanoviť hodnotu distribučnej funkcie môžeme využiť vhodný počítačový program.<sup>1</sup>

<sup>1</sup>Pozri príloha príklad 5.

#### 4. Normované normálne rozdelenie

Ak chceme získať hodnotu distribučnej funkcie normálneho rozdelenia a nemáme možnosť využitia počítačového programu, používame štatistické tabuľky, ktoré sú však zostavené pre hodnoty distribučnej funkcie normovanej náhodnej veličiny  $U$  [4]. Pre náhodnú veličinu, ktorá má normované normálne rozdelenie platí, že stredná hodnota  $\mu = 0$  a rozptyl  $\sigma^2 = 1$ . Ak  $X$  je náhodná veličina s normálnym rozdelením so strednou hodnotou  $\mu$  a rozptylom  $\sigma^2$ , tak

$$U = \frac{X - \mu}{\sigma}. \quad (6.9)$$

Pre funkciu hustoty pravdepodobnosti normovaného normálneho rozdelenia platí [3]

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \quad \text{pre } u \in \mathbb{R} \quad (6.10)$$

a s distribučnú funkciu

$$\Phi(u) = \int_{-\infty}^u \phi(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{t^2}{2}} dt \quad \text{pre } u \in \mathbb{R}. \quad (6.11)$$

V tab. 6.4 sú uvedené vybrané hodnoty číselné charakteristiky normovaného normálneho rozdelenia.

Tab. 6.4: Charakteristiky normovaného normálneho rozdelenia

$E(X)$	$D(X)$	kvantily $x_p$	$Me(X)$	$Mo(X)$
0	1	tabelované	0	0

## 7 ODHADY CHARAKTERISTÍK SÚBORU

Hlavným cieľom štatistiky je odhad vlastností a charakteristík základného súboru pomocou informácií získaných na základe výberového šetrenia. Ak skúmame znaky u všetkých štatistických jednotiek základného súboru, hovoríme o úplnom šetrení. Pri veľmi veľkých základných súboroch je to však ekonomicky, organizačne a časovo náročné.

Preto v praxi využívame neúplné šetrenie, ktorého dôležitou formou je výberové šetrenie. Pri výberovom šetrení zisťujeme požadované vlastnosti len u vybraných jednotiek zo základného súboru. Pri výbere prvkov však musíme dbať na to, aby boli vyberané objektívne a reprezentovali celú škálu základného súboru. Úlohou výberového šetrenia je podať informáciu o neznámej hodnote charakteristiky základného súboru či o parametroch rozdelenia náhodnej veličiny.

Základným a najčastejším typom výberového šetrenia je náhodný (pravdepodobnostný) výber, tj. všetky prvky základného súboru majú určité pravdepodobnostné zahrnutie do výberového súboru. Pri náhodnom výbere predpokladáme, že každá jednotka základného súboru má rovnakú pravdepodobnosť dostať sa do výberového súboru a preto uplatňujeme v plnom rozsahu všetky princípy teórie pravdepodobnosti.

Charakteristiky náhodného výberu nazývame výberové charakteristiky alebo štatistiky. Medzi základné výberové charakteristiky patrí výberový priemer ( $\bar{X}$ ), výberový rozptyl  $S^2$  a pod.

### 7.1 Bodové a intervalové odhady parametrov

Odhadom parametru  $\tau$  môže byť štatistika  $T_n$ , ktorej hodnoty kolísajú okolo  $\tau$  alebo interval, ktorého krajné body sú štatistikami a ktorý s predom zvolenou pravdepodobnosťou pokryje skutočnú hodnotu parametru  $\tau$ . V prvom prípade hovoríme o bodovom, v druhom prípade o intervalovom odhade parametrickej funkcie. Ak je bodovým odhadom parametru  $\tau$  štatistika  $T_n$  zapíšeme to nasledovne

$$\hat{\tau} = T_n. \quad (7.1)$$

Nech  $x_1, x_2, \dots, x_n$  je náhodný výber z rozdelenia s hustotou pravdepodobnosti  $f(x, \tau)$ . Ak sú  $T_d(x_1, x_2, \dots, x_n)$  a  $T_h(x_1, x_2, \dots, x_n)$  štatistiky, pre ktoré platí

$$P(T_d < \tau < T_h) = 1 - \alpha, \quad (7.2)$$

kde  $\alpha \in (0, 1)$ , potom interval  $T_d, T_h$  sa nazýva  $100(1 - \alpha)\%$  interval spoľahlivosti pre parameter  $\tau$ . Číslo  $1 - \alpha$  je koeficient spoľahlivosti, číslo  $\alpha$  riziko odhadu.

Je zrejmé, že pre odhad parametru  $\tau$  je možné použiť celú radu štatistik. Aby sme mohli určiť, ktorá z štatistik je najvhodnejšia zavádzame pojem lepši alebo najlepší odhad parametru  $\tau$ . Kvalitu bodových odhadov budeme najčastejšie posudzovať pomocou nestrannosti a výdatnosti.

Štatistiku  $T_n$  nazveme nestranným odhadom parametru  $\tau$ , ak pre každú hodnotu parametru  $\tau$  platí [3]

$$E(T_n) = \tau \tag{7.3}$$

Ak štatistika  $T_n$  splňuje túto podmienku, odhad sa často označuje ako neskreslený alebo nevychýlený.

Ak máme dva nestranné odhady  $T_1$ , a  $T_2$ , pre ktoré platí  $E(T_1) = E(T_2) = \tau$ , považujeme za lepši nestranný odhad tú zo štatistik, ktorej rozptyl je menši. Ak platí

$$D(T_1) < D(T_2), \tag{7.4}$$

hovoríme, že  $T_1$  je lepši nestranný odhad ako  $T_2$ . Štatistiku  $T_n$ , ktorá má zo všetkých štatistik daného typu pre odhad parametru  $\tau$  najmenši rozptyl, nazývame najlepším nestranným odhadom parametru  $\tau$ .<sup>1</sup>[2]

---

<sup>1</sup>Pozri príloha príklad 6.

## 8 TESTOVANIE ŠTATISTICKÝCH HYPOTÉZ

### 8.1 Princípy testovania štatistických hypotéz

V predchádzajúcej kapitole sme sa zaoberali metódami bodového a intervalového odhadu parametru rozdelenia. V praxi nás často zaujíma či platí určité tvrdenie o parametroch rozdelenia, z ktorého náhodný výber pochádza alebo tvrdenie, že toto rozdelenie je konkrétnym typom rozdelenia (napr. normálne).

Takéto tvrdenie, týkajúce sa rozdelenia náhodných veličín sa nazývajú štatistické hypotézy.

Hypotézu, ktorej platnosť overujeme nazývame testovaná alebo nulová hypotéza a označujeme ju  $H$ . Proti nulovej hypotéze  $H$  staviame tzv. alternatívnu hypotézu, ktorú značíme  $A$ .

Test štatistickej hypotézy  $H$  proti alternatíve  $A$  je postupom, ktorý na základe náhodného výberu z daného rozdelenia vedie buď k zamietnutiu testovanej hypotézy (tj. k prijatiu alternatívy  $A$ ), alebo k nezamietnutiu testovanej hypotézy  $H$  (tj. k zamietnutiu alternatívy  $A$ ).[1]

Pri teste o parametre  $\tau$  sa väčšinou formuje testovaná hypotéza ako jednoduchá  $H : \tau = \tau_0$ . Proti tejto hypotéze sa najčastejšie volia:

- zložená dvojstranná alternatívna hypotéza  $A : \tau \neq \tau_0$
- pravostranná alternatívna hypotéza  $A : \tau > \tau_0$
- ľavostranná alternatívna hypotéza  $A : \tau < \tau_0$

Ak na základe informácií, ktoré sme získali z náhodného výberu, nestačia na zamietnutie hypotézy  $H$ , neznamená to ešte dokázanie jej platnosti.

Pri testovaní hypotézy  $H$  sa môžeme dopustiť jednej z dvoch chýb.

Hypotéza  $H$  platí, ale na základe výberových hodnôt ju zamietame. Túto chybu nazývame chybou prvého druhu. Ak hypotéza  $H$  neplatí, avšak na základe výberových hodnôt ju nezamietame, dopustíme sa chyby druhého druhu.[1]

Hladina významosti  $\alpha$  testu je pravdepodobnosť chyby prvého druhu. V praxi sa najčastejšie volí  $\alpha = 0,05$  alebo  $\alpha = 0,01$ . Pravdepodobnosť chyby druhého druhu označujeme  $\beta$ . Číslo  $1 - \beta$  sa nazýva sila testu. Sila testu vyjadruje pravdepodobnosť, že zamietneme testovanú hypotézu  $H$ , ak platí alternatívna hypotéza  $A$ .

K testu hypotézy  $H$  oproti  $A$  použijeme štatistiku  $T_n$ , ktorá má pri platnosti  $H$  známe rozdelenie. Štatistika  $T_n$  sa nazýva testovým kritériom.

Obor hodnôt, ktoré testové kritérium môže nadobúdať rozdelíme na dva disjunktné obory: na obor  $W_\alpha$ , ktorý nazveme kritický obor a na obor  $W_{1-\alpha}$ , ktorý sa nazýva obor prijatia.

Keď testované kritérium  $T_n$  nadobudne hodnoty z  $W_\alpha$ , zamietame nulovú hypotézu  $H$  a naopak, keď testové kritérium  $T_n$  nadobudne hodnoty z  $W_{1-\alpha}$ , nulovú hypotézu  $H$  nezamietame.

V tomto prípade hovoríme o teste hypotézy  $H$ , založenom na kritickom obore  $W_\alpha$ .

Ďalším postupom ako testovať štatistické hypotézy, je využiť intervaly spoľahlivosti. Ak daná hodnota testového parametru padne do intervalu spoľahlivosti, nezamietame na hladine významosti  $\alpha$  nulovú hypotézu  $H$ . Ak daná hodnota testového parametru nepadne do intervalu spoľahlivosti, zamietame na hladine významosti  $\alpha$  nulovú hypotézu  $H$ .

Štatistické počítačové programy väčšinou určujú rozhodnutie o teste hypotézy na základe tzv.  $p$ -hodnoty, tj. najmenšia hladina významosti, pri ktorej je možné ešte zamietnuť nulovú hypotézu  $H$ . Je daná obsahom plochy pod funkciou hustoty pravdepodobnosti testovacej štatistiky.

Ak je hladina významosti  $\alpha > p$ -hodnota, na hladine významosti  $\alpha$  hypotézu zamietame. Ak je hladina významosti  $\alpha < p$ -hodnota, na hladine významosti  $\alpha$  hypotézu nezamietame.<sup>1</sup>

## 8.2 Testy parametrov normálneho rozdelenia

Ako už vieme, normálne rozdelenie  $N(\mu, \sigma^2)$  je charakterizované dvoma parametrami a to strednou hodnotou  $\mu$  a rozptylom  $\sigma^2$ . Ak si chceme odvodiť test hypotézy o strednej hodnote a rozptyle tohoto rozdelenia, musíme si ozrejmiť vetu o normálnom rozdelení. Majme postupnosť normálne rozdelených nezávislých náhodných veličín  $X_1, X_2, \dots, X_n$  s parametrami  $\mu, \sigma^2$ . Nech

$$\bar{X} = \frac{1}{n} \sum_{x=1}^n X_i \quad a \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (8.1)$$

sú funkcie náhodných veličín, potom veličina

$$t = \frac{\bar{X} - \mu}{S} \sqrt{n} \quad (8.2)$$

má Studentovo rozdelenie  $t$  s  $n - 1$  stupňa voľnosti a veličina

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad (8.3)$$

má Pearsonove rozdelenie  $\chi$  s  $n - 1$  stupňom voľnosti. [3]

---

<sup>1</sup>Pozri príloha príklad 7.

Studentovo aj Pearsonove rozdelenie sa využívajú pri odvodení kritérií pre testovanie parametrov normálneho rozdelenia.

1. Test strednej hodnoty normálneho rozdelenia

Nech  $x_1, x_2, \dots, x_n$  označuje náhodný výber z  $N(\mu, \sigma^2)$ , kde parameter  $\sigma^2$  nie známy. Testujeme hypotézu, že parameter  $\mu$  je rovný hodnote  $\mu_0$ :

$$H : \mu = \mu_0$$

Potom testové kritérium

$$t = \frac{\bar{x} - \mu_0}{s} \sqrt{n} \quad (8.4)$$

má pri platnosti hypotézy  $H$  Studentovo rozdelenie  $t(\nu)$  s  $\nu = n - 1$  stupňom voľnosti. Kritické obory volíme podľa alternatívnej hypotézy a sú zobrazené v tab. 8.1.[3]

Tab. 8.1: Kritické obory, pri teste strednej hodnoty normálneho rozdelenia

alternatívna hypotéza	kritický obor
$A : \mu > \mu_0$	$W_a = \{t; t \geq t_{1-\alpha}(\nu)\}$
$A : \mu < \mu_0$	$W_a = \{t; t \leq t_{1-\alpha}(\nu)\}$
$A : \mu \neq \mu_0$	$W_a = \{t;  t  \geq t_{1-\alpha/2}(\nu)\}$

Kde  $\nu = n - 1$  je Studentovo rozdelenie s kvantilami  $t_{1-\alpha}(\nu), t_{1-\alpha/2}(\nu)$ .<sup>2</sup>

2. Test rozptylu normálneho rozdelenia

Nech  $x_1, x_2, \dots, x_n$  označuje náhodný výber z  $N(\mu, \sigma^2)$ , kde parameter  $\sigma^2$  nie známy. Testujeme hypotézu, že parameter  $\sigma^2$  je rovný hodnote  $\sigma_0^2$ :

$$H : \sigma^2 = \sigma_0^2$$

Potom testové kritérium

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \quad (8.5)$$

má pri platnosti hypotézy  $H$  Pearsonove rozdelenie  $\chi^2(\nu)$  s  $\nu = n - 1$  stupňom voľnosti. Kritické obory volíme podľa alternatívnej hypotézy a sú zobrazené v tab. 8.2. [3]

Kde  $\nu = n - 1$  je Pearsonove rozdelenia s kvantilmi  $\chi_{1-\alpha}^2(\nu), \chi^2(\nu), \chi_{1-\alpha/2}^2(\nu), \chi_{\alpha/2}^2(\nu)$ .<sup>3</sup>

---

<sup>2</sup>Pozri príloha príklad 8.

<sup>3</sup>Pozri príloha príklad 9.

Tab. 8.2: Kritické obory, pri teste rozptylu normálneho rozdelenia

alternatívna hypotéza	kritický obor
$A : \sigma^2 > \sigma_0^2$	$W_a = \{\chi^2; \chi^2 \geq \chi_{1-\alpha}^2(\nu)\}$
$A : \sigma^2 < \sigma_0^2$	$W_a = \{\chi^2; \chi^2 \leq \chi_{\alpha}^2(\nu)\}$
$A : \sigma^2 \neq \sigma_0^2$	$W_a = \{\chi^2; \chi^2 \leq \chi_{\alpha/2}^2(\nu) \vee \chi^2 \geq \chi_{1-\alpha/2}^2(\nu)\}$

### 8.3 Test parametru alternatívneho rozdelenia

Alternatívne rozdelenie  $A(p)$  je charakterizované jedným parametrom  $p$ . Pre dostatočne veľké  $n$  je možné použiť testové kritérium, ktoré má približne normálne rozdelenie. Majme náhodný výber  $x_1, x_2, \dots, x_n$  z alternatívneho rozdelenia  $A(p)$  a  $r$  je výberová relatívna početnosť sledovaného znaku, čiže  $r = \frac{1}{n} \sum_{i=1}^n x_i$ . Testujeme hypotézu

$$H : p = p_0$$

Potom testové kritérium

$$u = \frac{r - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} \quad (8.6)$$

má pri platnosti nulovej hypotézy  $H$  približne normálne rozdelenie  $N(0, 1)$ . Podľa alternatívnej hypotézy volíme nasledujúce kritické obory:

Tab. 8.3: Kritické obory, pri teste alternatívneho rozdelenia

alternatívna hypotéza	kritický obor
$A : p > p_0$	$W_a = \{u; u \geq u_{1-\alpha}\}$
$A : p < p_0$	$W_a = \{u; u \leq -u_{1-\alpha}\}$
$A : p \neq p_0$	$W_a = \{u;  u  \geq u_{1-\alpha/2}\}$

Kde  $u_{1-\alpha}, u_{1-\alpha/2}$  sú kvantily normovaného normálneho rozdelenia  $N(0, 1)$ . [3]

## 9 ANALÝZA ROZPTYLU

Analýza rozptylu (ANOVA) je v praxi využívaná keď potrebujeme porovnať viac skupín, ako v našom prípade, kde sledujeme pacientov zdravých (skupina 1), mierne chorých (skupina 2) a s ťažkou formou ochorenia (skupina 3). V prípade, že sú veľkosti všetkých výberových súborov rovnaké, hovoríme o vyváženom modeli analýzy rozptylu.

Namiesto analýzy rozptylu je možné použiť vhodný test pre dva výbery (napr.  $t$ -test) a otestujeme ako sa líšia zdraví od mierne chorých, mierne chorí od ťažko chorých a zdraví od ťažko chorých. Inak povedané vykonáme 3 testy pre dva výbery. Značná nevýhoda v tejto možnosti je v tom, že s narastajúcim počtom testovaných hypotéz rastie tiež pravdepodobnosť omylu a ukázania na štatisticky významný rozdiel tam, kde v skutočnosti nie je (chyba I. druhu). Preto v takomto prípade využívame analýzu rozptylu, ktorá je vhodná pre viac ako dva výbery.

ANOVA je základnou parametrickou metódou na porovnanie stredných hodnôt niekoľkých základných súborov. Nulová hypotéza je v tomto prípade stanovená ako rovnosť stredných hodnôt vo všetkých sledovaných skupinách, alternatívna hypotéza je daná nasledovne: najmenej jedna stredná hodnota je odlišná od ostatných.

V našom prípade testujeme nulovú hypotézu  $H : \mu_{skupina1} = \mu_{skupina2} = \mu_{skupina3}$  oproti alternatívnej hypotéze  $A$ : najmenej jedno  $\mu$  je odlišné od ostatných. [6][7]

Analýzu rozptylu môžeme použiť len vtedy, ak sú splnené nasledujúce podmienky:

- Výberové súbory pochádzajú zo základných súborov s normálnymi rozdeleniami.
- Nezávislosť pozorovaných hodnôt.
- Rovnaký rozptyl hodnôt vo všetkých porovnávaných skupinách.

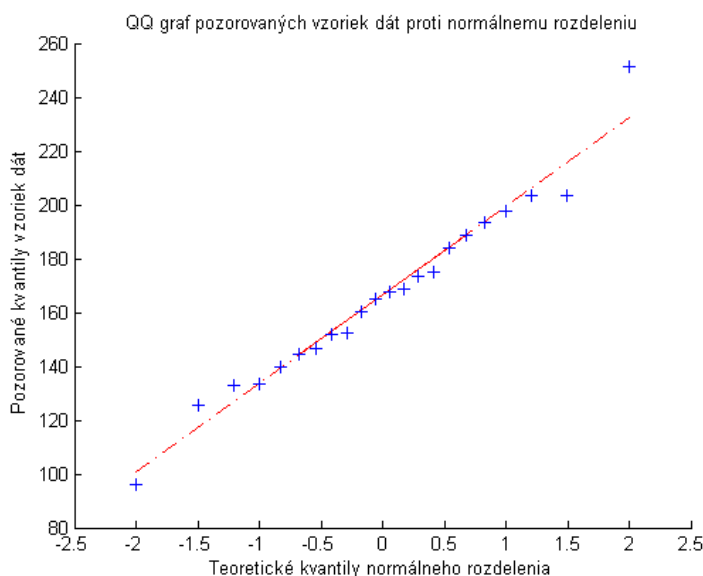
### 9.1 Testy normality

Normálne rozdelenie patrí medzi najdôležitejšie a v praxi najčastejšie používané pravdepodobnostné rozdelenia. Je to jedna zo základných podmienok pri použití ANOVY. V nasledujúcej časti uvedieme najpoužívanejšie testy normality.

#### 1. Grafické metódy

Posúdiť, či pozorované hodnoty pochádzajú z nejakého známeho rozdelenia pravdepodobnosti je možné aj pomocou Q-Q diagramu. Tento diagram proti sebe zobrazuje na ose  $x$  kvantily teoretického rozdelenia pravdepodobnosti (v našom prípade normálneho rozdelenia) a na ose  $y$  kvantily pozorovaných

hodnôt. [6] Ak tieto body ležia približne na priamke, môžeme usúdiť, že skúmaný výber pochádza z daného rozdelenia. Na obr. 9.1 je zobrazený Q-Q plot výberu, ktorý má normálne rozdelenie. Jedná sa o siedmy parameter z tretej skupiny pacientov. Na obrázku obr. 9.2 je Q-Q plot prvého parametru druhej skupiny, ktorý nepochádza z normálneho rozdelenia, pretože zobrazené body vytvárajú krivku odlišnú od priamky.



Obr. 9.1: Q-Q plot výberu s normálnym rozdelením

## 2. Šikmost a špicatosť ako kritérium normality

Medzi vlastnosti normálneho rozdelenia patrí že, má nulové koeficienty šikmosti  $\alpha_3 = 0$  a špicatosti  $\alpha_4 = 0$ .

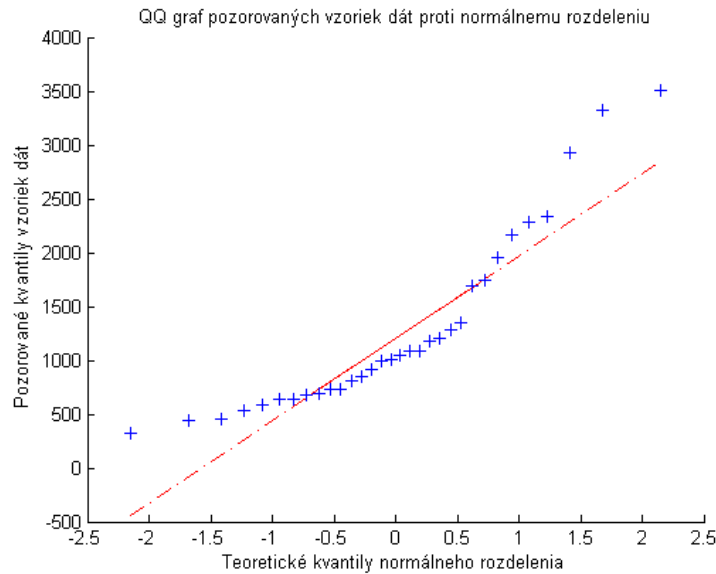
## 3. Kolmogorovov-Smirnovov (KS) test

Tento test je založený na porovnávaní distribučnej funkcie predpokladaného rozdelenia s výberovou (empirickou) distribučnou funkciou. Test je možné realizovať aj pre náhodné veličiny malých rozsahov. Na vopred zvolenej hladine významnosti  $\alpha$  testujeme hypozézu  $H : F(x) = F_0(x)$  proti alternatívnej hypotéze  $A : F(x) \neq F_0(x)$ , kde  $F$  je empirická distribučná funkcia výberového súboru a  $F_0$  je uvažovaná teoretická distribučná funkcia.

Testovacia štatistika  $D$  je definovaná ako najväčšia vzdialenosť medzi hodnotami  $F$  a  $F_0$ :  $D = \max|F_0(x) - F(x)|$ . Pre malé rozsahy výberu  $n$  použijeme presnú kritickú hodnotu z tabuľky kritických hodnôt tohoto testu, pre veľká  $n$  môžeme pre výpočet kritickej hodnoty použiť asymptotockú hodnotu  $\sqrt{-\frac{\ln \frac{\alpha}{2}}{2n}}$ , teda napr. pre  $\alpha = 0,05$  hodnotu  $\frac{1,3581}{\sqrt{n}}$ . [10]

V Matlabe môžeme pre tento test použiť funkciu *kstest*.

## 4. Lillieforsov test



Obr. 9.2: Q-Q plot výberu, ktorý nepochádza z normálneho rozdelenia

Lilieforsov test je modifikáciou Kolmogorovov-Smirnovov testu. Na rozdiel od KS testu, kde musela byť teoretická distribučná funkcia plne definovaná, pri tomto teste môžeme parametre  $\mu$  a  $\sigma$  odhadnúť pomocou výberových odhadov  $\bar{x}$  a  $s$ .

Rovnako ako v KS teste pre malé  $n$  použijeme presnú kritickú hodnotu z tabuľky Lilieforsovho testu, pre veľké  $n$  je možné použiť asymptotickú hodnotu  $\frac{0,89}{\sqrt{n}}$  pre  $\alpha = 0,05$ .

V Matlabe môžeme pre tento test použiť funkciu *lilietest*.<sup>[10]</sup>

## 9.2 Test zhody rozptylov súborov s normálnym rozdelením

Ako bolo uvedené vyššie, jednou z podmienok, ktoré musia byť splnené aby mohla byť ANOVA korektne použitá, je rovnosť rozptylov porovnávaných základných súborov.

Ak máme  $k$  nezávislých náhodných výberov, ktoré majú normálne rozdelenie  $N(\mu_i, \sigma_i^2)$ , kde  $\mu_i, \sigma_i^2$  sú neznáme parametre ( $i = 1, 2, \dots, k$ ). Testujeme nulovú hypotézu  $H : \sigma_1^2 = \dots = \sigma_k^2$  oproti aleternatívnej hypotéze  $A$ , že nie všetky rozptyly  $\sigma_i^2$  ( $i = 1, 2, \dots, k$ ) sa navzájom rovnajú. Neznámy rozptyl  $\sigma_i^2$  budeme odhadovať pomocou výberového rozptylu:

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2. \quad (9.1)$$

Bartlettov test

Označme  $S_i^2$  výberové rozptyly a  $n_i (i = 1, 2, \dots, k)$  rozsahy výberových súborov. Testovacia štatistika Bartlettovho testu je štatistika

$$\chi^2 = \frac{(n - k) \ln S^2 - [(n_1 - 1) \ln S_1^2 + \dots + (n_k - 1) \ln S_k^2]}{c}, \quad (9.2)$$

pričom

$$c = 1 + \frac{1}{3(k - 1)} \cdot \left( \frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + \dots + \frac{1}{n_k - 1} - \frac{1}{n - k} \right), \quad (9.3)$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \dots + (n_k - 1)S_k^2}{n - k}. \quad (9.4)$$

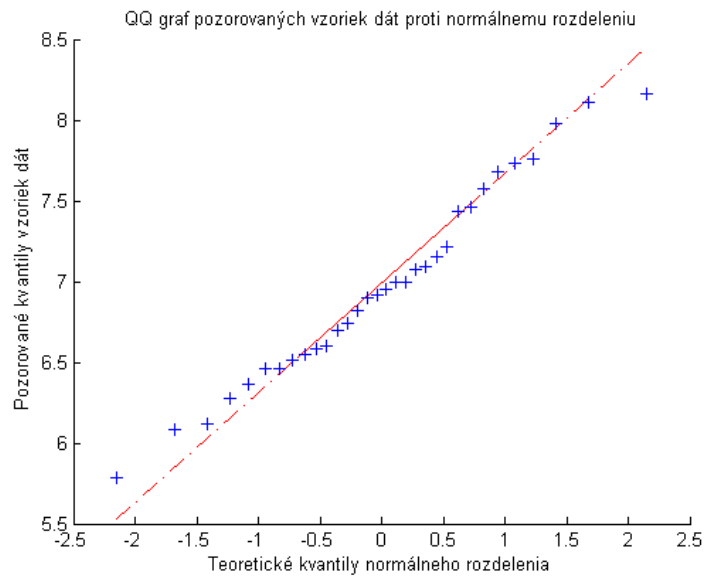
Štatistika  $\chi^2$  má za platnosti nulovej hypotézy asymptoticky  $\chi^2$  rozdelenie s  $k - 1$  stupňami voľnosti. Nulovú hypotézu  $H$  zamietame na hladine významnosti  $\alpha$ , ak pre hodnotu testovacej štatistiky  $\chi^2$  platí  $\chi^2 > \chi_\alpha^2(k - 1)$ , kde  $\chi_\alpha^2(k - 1)$  je kritická hodnota  $\chi^2$ -rozdelenia s  $k - 1$  stupňami voľnosti. [7]

### 9.3 Logaritmická transformácia

Keď rozloženie hodnôt v súbore nemá normálne rozdelenie, najčastejšia z možností ako pozorované hodnoty priblížiť normálnemu rozdeleniu je logaritmická transformácia a uplatňovať ďalšie štatistické postupy na takto transformovaných dátach. Zvyčajne sa používa transformácia pomocou prirodzeného logaritmu, menej časté pomocou dekadického logaritmu.

Ak overíme, že transformované dáta podmienku normálneho rozdelenia spĺňajú, môžeme na nich aplikovať parametrické štatistické metódy. Ak by sme výsledky štatistického spracovania chceli priviesť späť do pôvodných hodnôt, je nutné použiť inverznú funkciu a to exponenciálnu funkciu.

Praktický príklad možno vidieť porovnaním obr. 9.2, kde je zobrazený Q-Q diagram hodnôt, ktoré nemajú normálne rozdelenie (1. parameter druhej skupiny pacientov) a obr. 9.3, kde je zobrazený Q-Q plot týchto dát po logaritmickej transformácii. [9]



Obr. 9.3: Q-Q plot transformovaných dát

## 10 NEPARAMETRICKÉ TESTY

Parametrické testy sa zaoberajú testovaním tvrdení o neznámych parametroch rozdelení pravdepodobnosti, ktorým sa uvažovaná náhodná veličina riadi. Čo sa týka predpokladov, parametrické testy sú náročnejšie ako neparametrické, pretože u týchto testov sa predpokladá, že rozdelenie skúmanej náhodnej veličiny je známe až na jeden alebo viac parametrov a test sa tak týka týchto parametrov.

### 10.1 Wilcoxonov dvojvýberový test

Tento test sa používa namiesto dvojvýberového  $t$ -testu, ak nie je splnený predpoklad normality. Ekvivalent k tomuto testu je Mannov-Whitneyov test. Wilcoxonov dvojvýberový test umožňuje testovať hypotézu o zhode dvoch distribučných funkcií. Nech  $X_1, X_2, \dots, X_m$  je náhodný výber zo základného súboru so spojitou distribučnou funkciou  $F$  a nech  $Y_1, Y_2, \dots, Y_n$  je na ňom nezávislý náhodný výber zo základného súboru so spojitou distribučnou funkciou  $G$ . Testujeme nulovú hypotézu  $H : F = G$  oproti alternatívnej hypotéze  $A : F \neq G$ . [7]

Pri teste sa postupuje tak, že sa zoradia všetky údaje spoločne do neklesajúcej postupnosti. Vypočítame súčet poradí hodnôt pre obidve skupiny a označíme ho ako  $S_1$  a  $S_2$ . Hodnoty štatistík vypočítame podľa nasledujúceho vzťahu:

$$U_1 = m.n + \frac{m.(m+1)}{2} - S_1 \quad \text{a} \quad U_2 = m.n + \frac{n.(n+1)}{2} - S_2. \quad (10.1)$$

Ak  $\min(U_1, U_2) \leq W(\alpha)$ , kde  $W(\alpha)$  je tabelovaná kritická hodnota Wilcoxonovho dvojvýberového testu, nulovú hypotézu  $H$  zamietame na hladine významnosti  $\alpha$ . Označenie sa volí tak, aby platilo, že hodnota  $S_1$  je menšia z oboch súčtov.

### 10.2 Neparametrická alternatíva analýzy rozptylu - Kruskalov-Wallisov test

Neparametrickou alternatívou analýzy v prípade nesplnenia jej predpokladov, ktoré boli uvedené v predchádzajúcej kapitole, je Kruskalov-Wallisov test. Tento test je zovšeobecnením Wilcoxonovho dvojvýberového testu. Obidva testy netestujú zhodu konkrétnych parametrov, ale zhodu výberových distribučných funkcií porovnávaných súborov. Základná podmienka je nezávislosť pozorovaných hodnôt. [6]

Označme  $n_1, n_2, \dots, n_k$  rozsahy jednotlivých výberových súborov. Položme  $n = n_1 + n_2 + \dots + n_k$ . Všetkých  $n$  prvkov zoradíme do neklesajúcej postupnosti a každému priradíme jeho poradie. Označme  $S_i$  súčet poradí prvkov  $i$ -teho výberového súboru.

Testovacia štatistika, ktorú vypočítame podľa:

$$K = \frac{12}{n(n+1)} \cdot \sum_{i=1}^k \frac{T_i^2}{n_i} - 3(n+1), \quad (10.2)$$

má asymptoticky  $\chi^2$ -rozdelenia s  $k - 1$  stupňami voľnosti. [7]

Nulovú hypotézu  $H$  zamietame na hladine významnosti  $\alpha$ , ak  $K \geq K_\alpha$ , kde  $K_\alpha$  je kritická hodnota Kruskalovho-Wallisovho testu.

## 11 PRAKTICKÁ ČASŤ

Ako už bolo napísané v úvode, úlohou praktickej časti tejto práce je využiť štatistické metódy k vyhodnoteniu získaných parametrov pacientov a zistiť, ktoré sú signifikantné.

Dáta boli uložené v tabuľke  $N \times M$ , kde  $N$  odpovedalo počtu pacientov (79) a  $M$  počtu parametrov (29). V ďalšej tabuľke bol vektor, ktorý k poradovému číslu pacienta pridelil číslo od 1 do 3, ktoré určuje jeho zdravotný stav.

Aby sme mohli využívať štatistické metódy prvým potrebným krokom bolo rozdelenie pacientov do jednotlivých skupín podľa stupňa ochorenia.

Signifikantnosť parametrov sme sa pokúšali zistiť pomocou metódy typu ANOVA. Keďže metóda ANOVA je parametrická metóda, je potrebné overiť jej podmienky. Pre každý parameter každej skupiny sme overili normalitu dát pomocou Lillieforsovho testu. Lillieforsov test normality sa v Matlabe realizuje pomocou funkcie `[h,p]=lillietest(x)`, ak  $h = 0$  nulovú hypotézu na hladine významnosti  $\alpha$  nezamietame, to znamená, že dáta majú normálne rozdelenie.

V nasledujúcej tabuľke sú zobrazené parametre, ktoré prešli testom:

Tab. 11.1: Parametre s normálnym rozdelením

Skupina 1	1 2 3 5 17 23 24 27
Skupina 2	3 5 7 9 10 12 13 14 17 18 22 24 27
Skupina 3	3 5 7 8 9 12 14 21 24 27

Iba parametre číslo 3,5,24,27 prešli týmto testom v každej skupine.

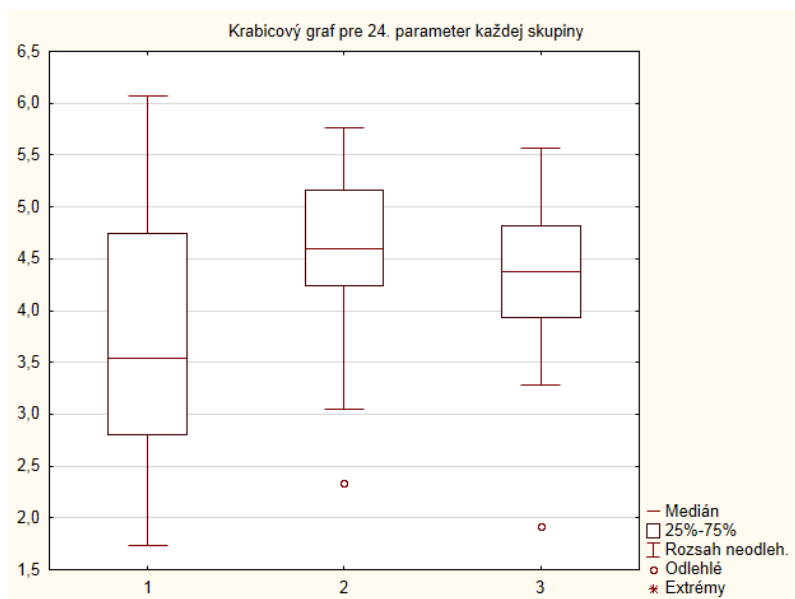
Pre tieto parametre bola overená druhá podmienka metódy ANOVA (zhoda rozptylov) pomocou Bartlettovho testu. Tento test sa v Matlabe vykoná pomocou funkcie `vartestn(y)`. Avšak ani pre jeden parameter nebol tento test úspešný. Výstupom tejto funkcie je aj krabicový graf, kde sú zobrazené jednotlivé rozptyly. Na obr. 11.1 je zobrazený krabicový graf pre 24. parameter každej skupiny.

Vzhľadom na to, že len tak málo parametrov prešlo testom normality, vykonali sme logaritmickú transformáciu pre ostatné parametre. Opäť sme realizovali Lillieforsov test normality a výsledky boli nasledujúce:

Tab. 11.2: Logaritmované parametre s normálnym rozdelením

Skupina 1	2 14 16 17 21 23 25
Skupina 2	1 2 9 10 12 14 22
Skupina 3	1 2 7 8 9 12 13 16 17 21 25

Len parameter číslo 2 prejavil normalitu v každej skupine.



Obr. 11.1: Krabicový graf pre 24. parameter každej skupiny

Keďže parametrov s normálnym rozdelením v každej skupine je veľmi málo, ukázala sa parametrická metóda ANOVA ako nevhodná pre tieto dáta. Preto je potrebné pre tieto údaje vykonať neparametrické testy.

Najprv sme vykonali Wilcoxonov dvojvýberový test, kde sme porovnali prvú skupinu s druhou, druhú s treťou a prvú s treťou. Test v Matlabe vykonáme pomocou funkcie  $[p,h]=\text{ranksum}(x,y)$ , ktorý vráti  $p$ -hodnotu tohto testu a ak  $h = 1$ , znamená to zamietnutie nulovej hypotézy  $H$  a teda signifikantnosť porovnávaných parametrov.

Signifikantné parametre sú zobrazené v nasledujúcej tabuľke.

Tab. 11.3: Signifikantné parametre zistené Wilcoxonovým testom

Medzi 1. a 2. skupinou	24 25
Medzi 2. a 3. skupinou	1 2 12 14
Medzi 3. a 1. skupinou	1 2 12 14 24 25

Najviac parametrov je logicky medzi 1. a 3. skupinou, keďže porovnáваме zdravých a pacientov s ťažkou formou ochorenia.

V tabuľke 11.4 sú zobrazené názvy signifikantných parametrov, ktoré sme získali týmto testom.

Ďalším neparametrickým testom, ktorý sme vykonali je neparametrická alternatíva analýzy rozptylu - Kruskalov-Wallisov test. Tento test porovnáva všetky 3 skupiny súčasne, v Matlabe je realizovaný funkciou  $p = \text{kruskalwallis}(X)$ . Výsledkom testu je  $p$ -hodnota, ktorá vypovedá o signifikantnosti parametrov. Ak je

Tab. 11.4: Názvy signifikantných parametrov

Číslo parametru	Názov parametru
1	Doba pohybu pera nad tabletom
2	Doba pohybu pera na tablete
12	Priemerná rýchlosť písania
14	Medián doby písania
24	Entropia tlaku na hrot pera
25	Priemerný tlak pri písaní

$p$ -hodnota blízko nuly (menej ako 0,05), znamená to, že aspoň jeden medián je signifikantne odlišný.

V našom prípade testom prešli parametre číslo:

2 12 14 24 25.

Sú to takmer rovnaké parametre ako pri Wilcoxonovom dvojvýberom teste pri porovnaní prvej a tretej skupiny.

Z toho vyplýva, že u vyššie uvedených parametrov sa prejavila štatistická odlišnosť medzi jednotlivými pozorovanými skupinami.

## 12 ZÁVER

Týmto sme sa zoznámili s princípmi testovania štatistických hypotéz. Hlavný postup pri testovaní hypotéz je nasledovný:

- Sformulujeme nulovú hypotézu  $H$  a alternatívnu hypotézu  $A$ .
- Zvolíme hladinu významosti  $\alpha$ .
- Využijeme jeden z vyššie popísaných štatistických postupov, na základe ktorého sformulujeme záver, teda na hladine významosti  $\alpha$  nulovú hypotézu zamietame alebo nezamietame.

Cieľom praktickej časti bolo využiť získané teoretické poznatky v praxi. Formou cvičení, kde pacienti postihnutí Alzheimerovou chorobou kreslili obrázky, prípadne písali úryvky textu boli získané parametre. Cieľom tejto práce bolo pomocou štatistických testov nájsť také parametre, ktoré by dokázali signifikantne oddeliť skupinu zdravých, skupinu s ľahkým ochorením a skupinu s ťažkou formou ochorenia.

Po rozdelení dát do skupín podľa stupňa ochorenia, sme sa snažili nájsť signifikantné parametre pomocou metódy typu ANOVA. Vzhľadom na to, že metóda ANOVA je parametrická, bolo nutné overiť jej podmienky. Normalita dát bola overená pomocou Lillieforsovho testu, zhoda rozptylov pomocou Barlettovho testu. Parametre, pri ktorých sa nepreukázala normalita, boli transformované pomocou logaritmickej transformácie. Keďže len veľmi málo parametrov splnilo podmienky metódy ANOVA, nie je táto metóda vhodná pre tieto dáta.

Neparametrické testy nie sú také náročné, čo sa týka predpokladov a podmienok pri ich použití. Na dátach bol vykonaný Wilcoxonov dvojvýberový test, kde boli porovnávané prvá skupina s druhou, druhá s treťou a tretia s prvou. Ďalším neparametrickým testom, ktorý bol vykonaný, je Kruskalov-Wallisov test. Tento test porovnáva všetky skupiny súčasne. V tabuľke 12.1 sú uvedené poradové čísla a názvy parametrov, ktoré sa prejavili ako signifikantne odlišné pomocou tohto testu.

Tab. 12.1: Názvy signifikantných parametrov

Číslo parametru	Názov parametru
2	Doba pohybu pera na tablete
12	Priemerná rýchlosť písania
14	Medián doby písania
24	Entropia tlaku na hrot pera
25	Priemerný tlak pri písaní

Znamená to, že vo vyššie uvedených parametroch pri porovnávaní medzi skupinami zdravých, mierne chorých a ťažko chorých sa nachádzajú štatisticky významné odlišnosti. Týmto bola splnená hlavná úloha bakalárskej práce.

Nájdenie signifikantných parametrov ma praktický význam v prevencii, rýchlejšom určení diagnózy, prípadne v určení úspešnosti liečby. Pri poznaní, ktoré parametre sú dôležité sa lekári môžu zamerať len na tieto a rovnakými cvičeniami ako boli parametre získané, môžu pacientov testovať a vyšetřovať.

# LITERATÚRA

- [1] ANDĚL, J. *Základy matematické statistiky*. 1. vyd. Praha : Matfyzpress, 2005, 358 s. ISBN 80-867-3240-1.
- [2] LIKEŠ, J., MACHEK, J. *Matematická statistika*. 2. vyd. Praha : SNTL, 1983, 178 s.
- [3] KŘÍŽ, O., NEUBAUER, J., SEDLAČÍK, M. *Základy statistiky: aplikace v technických a ekonomických oborech*. 1. vyd. Praha: Grada, 2012, 236 s. ISBN 978-80-247-4273-1.
- [4] HLAVIČKOVÁ, I., HLINĚNÁ, D. *Matematika 3 - Sbírka úloh z pravděpodobnosti*. Brno: UMAT FEKT VUT, 2007, 77 s.
- [5] FAJMON, B., RŮŽIČKOVÁ, I. *Matematika 3* Brno: UMAT FEKT, 2005, 255 s.
- [6] DUŠEK, L., PAVLÍK, T. *Biostatistika*. 1. vyd. Brno: Akademické nakladatelství CERM, 2012, 131 s. ISBN 978-80-7204-782-6.
- [7] MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. Nitra: UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE, 2011, 534 s.
- [8] PODMANICKÁ, P. *Štatistika – základné pojmy a typy priemerov* . [online]. 08.07.2011 [cit. 24.04.2013]. Dostupné z: [http://www.oskole.sk/?id\\_cat=2&clanok=16393](http://www.oskole.sk/?id_cat=2&clanok=16393)
- [9] JABOR, A. *Štatistické postupy v laboratorní medicíně*. [online]. [cit. 24.04.2013]. Dostupné z: <http://ciselniky.dasta.mzcr.cz/hypertext/200640/hypertext/AJDKR.htm>
- [10] *Transfer inovácií: Aplikácia štatistických testov dobrej zhody*. [online]. 2012 [cit. 24.04.2013]. ISSN 1337-7094. Dostupné z: [http://www.sjf.tuke.sk/transferinovacii/pages/archiv/transfer/23-2012/23\\_2012.htm](http://www.sjf.tuke.sk/transferinovacii/pages/archiv/transfer/23-2012/23_2012.htm)
- [11] Dokumentácia programu Matlab

# ZOZNAM PRÍLOH

A Príklady	42
B Zdrojový kód v Matlabe	54

## A PRÍKLADY

Riešené ukázkové príklady.

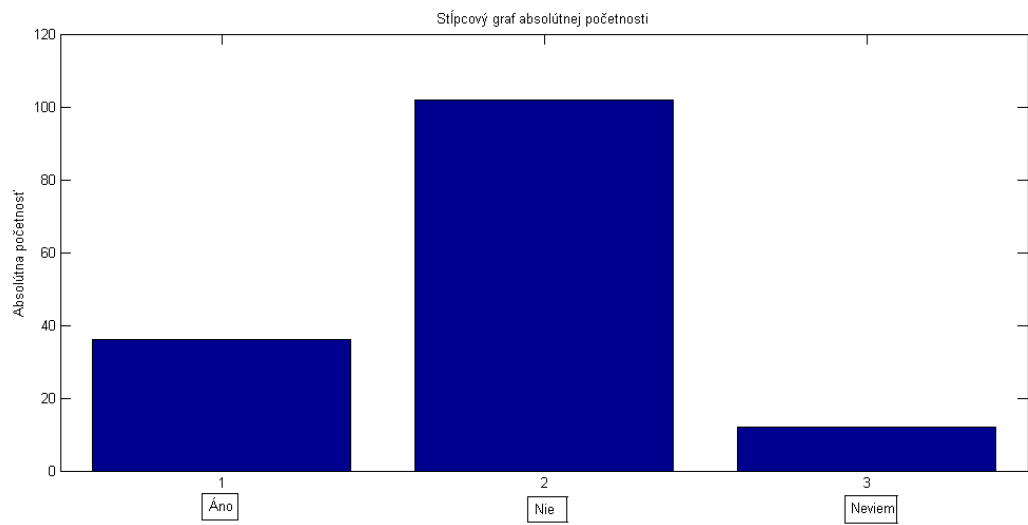
Príklad 1.

V náhodnej ankete bolo oslovených 150 respondentov. Každý vyjadril svoj názor v podobe odpovede áno, nie alebo neviem. Celkom 36 odpovedalo áno, 102 nie a zvyšných 12 neviem. Určte absolútnu a relatívnu početnosť jednotlivých odpovedí a výsledky vyjadrite v tabuľku, koláčovom a stĺpcovom grafe.

Odpovede	Absolútna početnosť $n_i$	Relatívna početnosť $p_i$
Áno	36	$\frac{36}{150} = 0,24$
Nie	102	$\frac{102}{150} = 0,68$
Neviem	12	$\frac{12}{150} = 0,08$
Spolu	150	1



Obr. A.1: Koláčový graf relatívnej početnosti



Obr. A.2: Stĺpcový graf absolútnej početnosti

Ako je možné vidieť z uvedených grafov koláčový graf je najvhodnejšie grafické znázornenie percent (relatívnej početnosti) a stĺpcový graf je vhodný pri grafickom znázornení absolútnej početnosti.

Príklad 2.

Štatistické charakteristiky

V Matlabe bolo vygenerovaných 20 nasledujúcich čísel v intervale 31 až 35: 33 35 31 32 34 32 31 34 31 32 33 32 32 34 35 32 30 34 33 35. Určte aritmetický priemer, modus, medián, horný a dolný kvantil, rozptyl, smerodajnú odchýlku, výberový rozptyl a výberovú smerodajnú odchýlku.

Charakteristika	Značka	Funkcia v MATLABe
Aritmetický priemer	$\bar{x}$	<code>M = mean(A)</code>
Modus	$\hat{x}$	<code>M = mode(X)</code>
Medián	$\tilde{x}$	<code>M = median(A)</code>
Dolný kvantil	$x_{0,25}$	<code>Y = prctile(X,25)</code>
Horný kvantil	$x_{0,75}$	<code>Y = prctile(X,75)</code>
Rozptyl	$s_n^2$	<code>V = var(X,1)</code>
Smerodajná odchylka	$s_n$	<code>s = std(X,1)</code>
Výberový rozptyl	$s^2$	<code>V = var(X)</code>
Výberová smerodajná odchylka	$s$	<code>s = std(X)</code>

Charakteristika	Značka	Funkcia v MATLABe
Aritmetický priemer	$\bar{x}$	32,750
Modus	$\hat{x}$	32,000
Medián	$\tilde{x}$	32,500
Dolný kvantil	$x_{0,25}$	32,000
Horný kvantil	$x_{0,75}$	34,000
Rozptyl	$s_n^2$	2,088
Smerodajná odchylka	$s_n$	1,445
Výberový rozptyl	$s^2$	2,197
Výberová smerodajná odchylka	$s$	1,482

Na jednoduchom príklade sú ukázané výpočty charakteristík polohy a variability pomocou programu Matlab.

Príklad 3.

Poissonove rozdelenie

K holičovi chodia priemerne 3 ľudia za hodinu. Určite pravdepodobnosť, že v priebehu hodiny:

- a) príde aspoň jeden
- b) prídu najviac dvaja
- c) príde jeden alebo dvaja.

Riešenie

Pri výpočte pravdepodobnostnej funkcie Poissonovho rozdelenia použijeme v Matlabe funkciu `poisspdf(X,lambda)` pre výpočet distribučnej funkcie použijeme funkciu `poisscdf(X,lambda)`, kde  $X$  je hodnota náhodnej veličiny  $X$  a  $\lambda$  je stredná hodnota tohto rozdelenia  $\lambda$ . Vypočítané hodnoty sú zobrazené v tab. A.1.

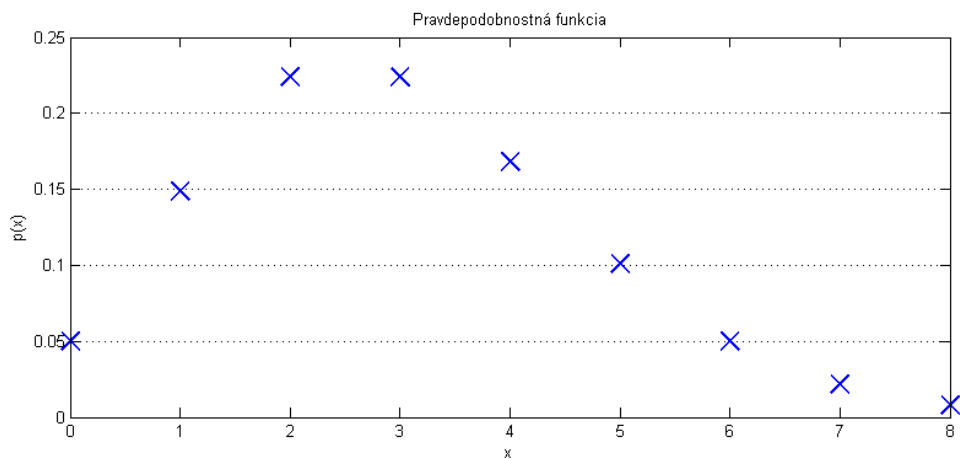
Tab. A.1: Hodnoty pravdepodobnostnej a distribučnej funkcie pre  $x = 0, 1, \dots, 8$

$x$	0	1	2	3	4	5	6	7	8
$p(x)$	0,050	0,149	0,224	0,224	0,168	0,101	0,050	0,022	0,008
$F(x)$	0,050	0,199	0,423	0,647	0,815	0,916	0,966	0,988	0,996

Za hodinu prídu priemerne 3 ľudia, preto pre časovú jednotku jednej hodiny platí  $\lambda = 3$ .

- a)  $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - p(0) \doteq 0,95$
- b)  $P(X \leq 2) = F(2) = P(X = 0) + P(X = 1) + P(X = 2) \doteq 0,423$
- c)  $P(X = 1 \vee X = 2) = P(X = 1) + P(X = 2) \doteq 0,373$

Z príkladu vyplýva, že Poissonovo rozdelenie sa používa v prípade, keď nás zaujíma pravdepodobnosť, že nastane určitý počet udalostí za jednotku času.



Obr. A.3: Pravdepodobnostná funkcia rozdelenia  $Po(3)$

No horizontálnej osi je index  $x$  určujúci počet udalostí, v našom prípade je to počet zákazníkov. Funkcia je definovaná len pre celé čísla  $x$ . Na vertikálnej osi je hodnota pravdepodobnosti realizácie konkrétnej udalosti.

Príklad 4.

Binomické rozdelenie

Pravdepodobnosť, že náhodne vybraný výrobok spĺňa výrobné normy, je 86%. Vyberáme náhodne 6 výrobkov. Aká je pravdepodobnosť, že medzi šiestimi po sebe vybranými výrobkami budú a) práve 2 chybné, b) najviac 3 chybné? Určte pravdepodobnostnú a distribučnú funkciu náhodnej veličiny udávajúcej počet kvalitných výrobkov medzi šiestimi po sebe vybranými výrobkami. Určite strednú hodnotu, rozptyl a smerodajnú odchylku danej náhodnej veličiny.

Riešenie

Pri výpočte pravdepodobnostnej funkcie Binomického rozdelenia použijeme v Matlabe funkciu `binopdf(X,N,P)` pre výpočet distribučnej funkcie použijeme funkciu `poisscdf(X,N,P)`, kde  $X$  je hodnota náhodnej veličiny  $X$ ,  $N$  je počet pokusov a  $P$  je pravdepodobnosť úspechu v jednom pokuse.

Hodnoty pravdepodobnostnej a distribučnej funkcie, pričom index  $x$  označuje počet kvalitných výrobkov:

$x$	0	1	2	3	4	5	6
$p(x)$	0,00001	0,00028	0,00426	0,03491	0,16082	0,39516	0,40457
$F(x)$	0,00001	0,00029	0,00455	0,03945	0,20027	0,59543	1,00000

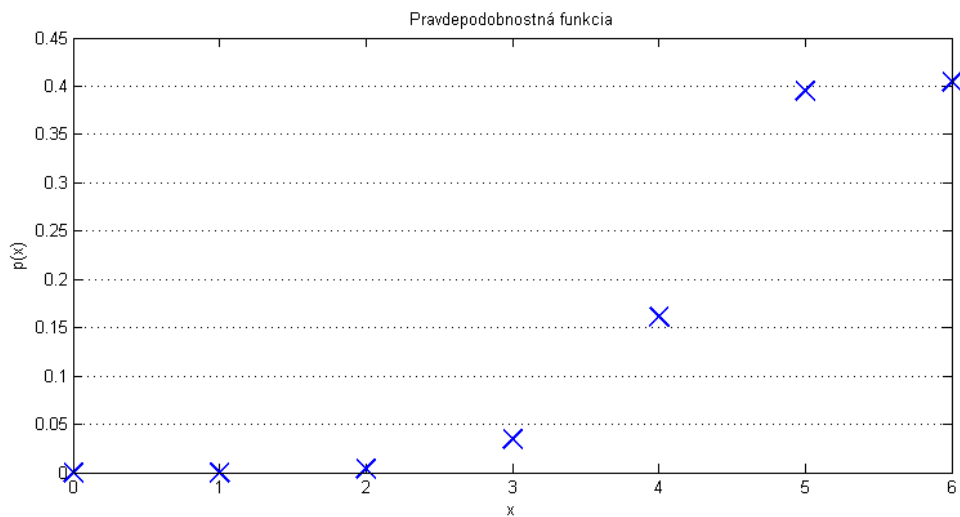
a) práve dva chybné, to znamená práve 4 dobré:  $P(X = 4) = p(x) = 0,16082$

b) najviac tri chybné  $P(X \leq 3) = F(3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = p(0) + p(1) + p(2) + p(3) = 0,39454$

Stredná hodnota:  $E(X) = n \cdot \pi = 6 \cdot 0,86 = 5,16$

Rozptyl:  $D(X) = n \cdot \pi(1 - \pi) = 6 \cdot 0,86 \cdot (1 - 0,86) = 0,7224$

Binomické rozdelenie sa používa v prípade, keď je experiment takej povahy, že môžu nastať len dva rôzne výsledky (úspech a neúspech), pričom náhodná veličina  $X$  udáva počet výskytu úspechu v postupnosti  $n$  nezávislých pokusov.



Obr. A.4: Pravdepodobnostná funkcia rozdelenia  $B(6; 0, 86)$

No horizontálnej osi je index  $x$  určujúci počet výskytov úspechu, v našom príklade počet kvalitných výrobkov. Funkcia je definovaná len pre celé čísla  $x$ . Na vertikálnej osi je hodnota pravdepodobnosti pri konkrétnom počte úspechov.

Príklad 5.

Normálne rozdelenie

Výsledok merania elektrického napätia je náhodná veličina  $X$  s normálnym rozdelením so strednou hodnotou  $\mu = 10$  V a rozptylom  $\sigma^2 = 4V^2$ . Vypočítajte pravdepodobnosť, že bude nameraná:

- a) menej než 13V
- b) viac ako 9V
- c) menej ako 12V, ale viac ako 5 V.

Riešenie

Pri výpočte pravdepodobnostnej funkcie normálneho rozdelenia použijeme v Matlabe funkciu `normpdf(X,mu,sigma)` pre výpočet distribučnej funkcie použijeme funkciu `normcdf(X,mu,sigma)`, kde  $X$  je hodnota náhodnej veličiny  $X$ ,  $\mu$  je stredná hodnota  $\mu$  a  $\sigma$  je smerodajná odchýlka  $\sigma$ .

Hodnoty distribučnej funkcie  $F(X)$ :

$F(13)$	0,933
$F(9)$	0,309
$F(12)$	0,841
$F(5)$	0,006
$F(12) - F(5)$	0,835

- a)  $P(X < 13) = F(13) \doteq 0,933$
- b)  $P(X > 9) = 1 - P(X < 9) = 1 - F(9) = 1 - 0,30854 \doteq 0,691$
- c)  $P(5 < X < 12) = F(12) - F(5) \doteq 0,835$

Normálne rozdelenie je jedno z najdôležitejších rozdelení spojitej náhodnej veličiny. Výpočet distribučnej funkcie  $F(x)$  normálneho rozdelenia sa prevádza len numericky, pretože Gaussov integrál nie je analyticky vyjadriteľný, preto je pri výpočte distribučnej funkcie jednoduché a vhodné využitie počítačového programu.

### Príklad 6.

#### Bodové a intervalové odhady parametrov

Pomocou funkcie `normrnd` bolo v Matlabe vygenerovaných 20 nasledujúcich čísiel s normálnym rozdelením: 9,49 9,71 9,02 9,55 9,45 9,18 9,33 9,46 10,01 9,87 9,17 9,92 9,52 9,39 9,52 9,37 9,38 9,65 9,64 9,64.

- Odhadnite strednú hodnotu tohto výberu a určte smerodajnú chybu tohto odhadu.
- Určite, v akých medziach možno so spoľahlivosťou 95 % očakávať strednú hodnotu tohto výberu.

#### Riešenie

- Platí, že výberový priemer  $\bar{x}$  (vypočítame pomocou funkcie `mean`) je nevychýleným bodovým odhadom strednej hodnoty  $\mu$  a výberová smerodajná odchýlka  $s$  (vypočítame pomocou funkcie `std`) je nevychýleným bodovým odhadom smerodajnej odchýlky  $\sigma$  základného súboru s normálnym rozdelením.

Bodový odhad strednej hodnoty je  $\hat{\mu} = \bar{x} = 9,514$  a odhad strednej chyby určíme zo vzťahu  $\widehat{SE}(\bar{x}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{0,251}{\sqrt{20}} \doteq 0,0561$ .

- Interval spoľahlivosti pre strednú hodnotu  $\mu$  pri neznámej smerodajnej odchýlke  $\sigma$  vypočítame v Matlabe pomocou funkcie `[h,p,ci]=ttest(x,m,alpha,tail)`, kde `x` sú namerané hodnoty, `m` je výberový priemer, `alpha` je zvolená hladina významnosti a `tail` určuje či sa jedná o obojstranný interval (`tail = 0/'both'`), pravostranný interval (`tail = 1/'right'`) alebo ľavostranný interval (`tail = -1/'left'`).

Do premennej `ci` sa uložia hranice intervalu spoľahlivosti pre strednú hodnotu  $\mu \in (9,396; 9,631)$ . Inými slovami môžeme povedať, že tento interval zahrňuje  $\mu$  s rizikom  $\alpha = 0,05$ .

Príklad 7.

Testovanie hypotéz pomocou  $p$ -hodnoty

Meraním hmotnosti 15 rožkov sme dostali priemernú hmotnosť 41,63 gramu a výberovú smerodajnú odchýlku 0,0416. Predpokladáme, že namerané údaje pochádzajú z normálneho rozdelenia. Na hladine významnosti 0,05 testujete hypotézu, že priemerná hmotnosť rožkov sa neodchyľuje od normy 41,6 gramu.

Výberové charakteristiky:

Priemer	$\bar{x}$	41,63
Výberová smerodajná odchýlka	$s$	0,0416
Stredná hodnota	$\mu_0$	41,6
Rozsah výberu	$n$	15

Riešenie

Testujeme nulovú hypotézu  $H : \mu = 41,6$  proti alternatíve  $A : \mu \neq 41,6$ . Testové kritérium  $t = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$ , má pri platnosti nulovej hypotézy Studentovo rozdelenie  $t(\nu)$  s  $\nu = n - 1$  stupňom voľnosti.

$P$ -hodnotu získame pomocou funkcie  $\text{tpdf}(\mathbf{X}, \mathbf{V})$ , ktorá vracia hladinu významnosti (pravdepodobnosť) pre funkciu Studentovho  $t$ -rozdelenia, kde  $\mathbf{X}$  je vypočítaná hodnota parametra  $t$ , pre ktorú sa zisťuje hladina významnosti a  $\mathbf{V}$  je celé číslo určujúce počet stupňov voľnosti  $n - 1$ .

$t$	2,793
$p$ -hodnota	0,014

Pretože  $0,05 > 0,014$ , zamietame na hladine významnosti 0,05 nulovú hypotézu  $H$ .

Príklad 8.

Test strednej hodnoty normálneho rozdelenia

Pre prevedenie testu použijete náhodný výber 20 hodnôt z príkladu 6. Na hladine významosti  $\alpha = 0,05$  určte pravdivosť tvrdenia, že stredná hodnota neprekročí hodnotu 9,7.

Základné charakteristiky náhodného výberu:

Rozsah výberu	$n$	20
Výberový priemer	$\bar{x}$	9,514
Výberová smerodajná odchýlka	$s$	0,251
Stredná hodnota	$\mu_0$	9,7
Hladina významosti	$\alpha$	0,05

Riešenie

Testujeme nulovú hypotézu  $H : \mu = 9,7$  proti alternatíve  $A : \mu < 9,7$ . Nulová hypotéza je  $\mu = \mu_0$ , alternatívna hypotéza je  $\mu < \mu_0$ .

Pri teste strednej hodnoty normálneho rozdelenia sa jedná o Studentov test pre jeden výber. Testové kritérium určíme podľa vzorca  $t = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$ . Pre výpočet kvantilu Studentovho rozdelenia sa v Matlabe používa funkcia  $X=TINV(P,V)$ . Táto funkcia vráti inverznú distribučnú funkciu Studentovho  $t$ -rozdelenia ako funkciu pravdepodobnosti (parameter P) a stupňov volnosti (parameter V).

$t$	-3,314
$t_{1-\alpha}(\nu)$	-1,729

Hodnota testového kritéria patrí do kritického oboru, pretože platí  $-3,314 \leq -1,729$ . Na základe tohto môžeme na hladine významosti 0,05 zamietnuť nulovú hypotézu, to znamená, že stredná hodnota s 95% spoľahlivosťou neprekročí hodnotu 9,7.

Príklad 9.

Test rozptylu normálneho rozdelenia

Pre prevedenie testu použijete náhodný výber 20 hodnôt z príkladu 6. Na hladine významnosti  $\alpha = 0,05$  určte pravdivosť tvrdenia, že rozptyl neprekročí hodnotu o 0,25 od priemernej hodnoty.

Základné charakteristiky náhodného výberu:

Rozsah výberu	$n$	20
Hladina významnosti	$\alpha$	0,05
Výberový rozptyl	$s^2$	0,063
Výberová smerodajná odchýlka	$s$	0,251
Rozptyl	$\sigma^2$	0,25

Riešenie

Testujeme nulovú hypotézu  $H : \sigma^2 = 0,25$  proti alternatíve  $A : \sigma^2 < 0,25$ . Nulová hypotéza je  $\sigma^2 = \sigma_0^2$ , alternatívna hypotéza je  $\sigma^2 < \sigma_0^2$ .

Pri teste rozptylu normálneho rozdelenia sa jedná o Pearsonov test. Testové kritérium určíme podľa vzorca  $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ . Pre výpočet kvantilu Pearsonovho rozdelenia sa v Matlabe používa funkcia  $X=\text{chi2inv}(P,V)$ . Táto funkcia vráti inverznú distribučnú funkciu Pearsonovho  $\chi^2$ -rozdelenia ako funkciu pravdepodobnosti (parameter P) a stupňov volnosti (parameter V).

$\chi^2$	4,788
$\chi_{\alpha/2}^2(\nu)$	10,117

Hodnota testového kritéria patrí do kritického oboru, platí  $4,788 \leq 10,117$ . Na základe tohto môžeme na hladine významnosti 0,05 zamietnuť nulovú hypotézu, to znamená, že rozptyl s 95% spoľahlivosťou neprekročí hodnotu o 0,25 gramov od priemernej hodnoty.

## **B ZDROJOVÝ KÓD V MATLABU**

Zdrojový soubor praktické části (testy s pacienty) v programu Matlab.