Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Vizualizace výsledků statistického medicínského šetření

Plzeň 2016 Iva Ptáčková

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 5. května 2016

Iva Ptáčková

Abstract

Visualization of statistical medical research

This bachelor thesis solves visualization of statistical medical research that is based on the data from patients who suffered a stroke. The goal is to find optimal means of visualizing chosen statistical methods. The thesis contains basic theory about the phases of statistical data prosessing, about the register from which it originates and the principle of statistical methods. Solution includes recommendations of graphs for certain methods, on implementation and subsequent implementation of the assignment.

Vizualizace výsledků statistického medicínského šetření

Tato bakalářská práce řeší vizualizaci výsledků statistického medicínského šetření, které vychází z údajů o pacientech s prodělanou mozkovou příhodou. Jejím cílem je najít optimální způsob vizualizace vybraných statistických metod. Práce obsahuje základní teorii zabývající se fázemi statistického zpracování dat dat, registrem, z kterého data pochází, a principem statistických metod. Náplň cíle zahrnuje doporučení ohledně grafů u daných metod, implementaci a následnou realizaci zadání.

Obsah

1 Úvod

Cílem této bakalářské práce je navrhnout a následně implementovat a otestovat vhodné řešení pro vizualizaci výsledků statistického šetření medicínských dat, které bude usnadňovat práci při vyhodnocení získaný dat.

První tři kapitoly jsou čistě seznámení se s teorií. Obsahují informace o problematice vizualizace dat, kterou se tato práce v okrajové části medicínského výzkumu snaží zmírnit. Dále seznamuje čtenáře s fázemi statistického zpracování medicínských dat, které vedou k výsledným statistickým souborům dat, s možnostmi grafického vyhodnocení těchto dat, s ohledem na povahu dat. A také objasňuje původ dat, s kterými v dalších kapitolách pak pracuje. Tedy seznamuje čtenáře s registrem SITS.

V dalších kapitolách se pak práce začíná zabývat řešením. Je objasněn princip obvykle používaných metod v oblasti zpracování medicínských dat a následně navrhnuto grafické řešení. Vybraná možná řešení jsou pak dále implementována ve zvoleném programu spolu se statistickými metodami a následně otestována na dostupných datech.

2 Problematika vizualizace dat

Vizualizace je způsob vytváření obrazů nebo animací s úmyslem sdělení nějaké informace za pomoci abstrakce. Jde o jinou než textovou interpretaci.

Data získaná z medicínských přístrojů se již dají považovat za jistou interpretaci, ale bez příslušných znalostí nejsou tak čitelná a jednoznačná. V některých případech se jejich hodnota zvýší až v kombinaci s jiným údajem, či výstupem jiné analytické metody. Kontext dat je tedy důležitý a vizualizace je jeho doplněním.

Základní data, s kterými se pracuje při vizualizaci, jsou data získaná z papírových schématických dat. Tato data jsou vhodně přizpůsobena a přeformulována do elektronické podoby. Jedná se například o výstup počítačové tomografie známé také jako CT (Computer Tomography). Data jsou sice už v elektronické podobě, ale ne ve formě vhodné pro komplexnější zpracování [2].

Přeformulování a zadání dat do systému je součástí práce lékařského pracovníka. Lidský faktor v procesu sběru dat a informací tak ve výsledku zapříčinil snahu o jistou standardizaci postupu, který vede k vyhodnocení nebo diagnostice. Toto menší opatření vniklo i navzdory tomu, že míra znalostí, zkušeností a vlastní úsudek v tomto oboru značně ovlivňuje diagnostiku.

Ve skutečnosti a tedy i v praxi jde o vítaný faktor, nicméně při interakci zapisovatele a počítače to značně znesnadňuje přepis těchto dat do elektronické podoby. I tak jsou faktory jako vnímání, poznávání, komunikace člověka s počítačem a podpora znalostí podstatnými prvky v procesu, který vede k vizualizaci [2].

2.1 Výhody vizualizace medicínských dat

S růstem možných vyšetření a medicínských záznamů roste i množství informací o jednotlivých pacientech. Na jednoho pacienta je k dispozici zároveň i velké množství různorodých dat - textové, numerické, diagramové, apod. Jejich vizualizace je možným řešením nepřehlednosti, která může následně vést k uvědomění si komplex-

nosti anamnézy nebo jiných závislostí jednotlivých dat, které se ve větších objemech dat mohou ztratit. Při kreativním řešení je šance dosáhnout výsledků, jako oproštění od tradičního smýšlení a nalezení nových východisek při dané léčbě [2].

2.2 Data produkovaná lékařským systémem

Hlavní problematikou je tedy prezentace dat, která medicínský systém vyprodukuje. Typy dat jsou obvykle děleny do skupin:

- text,
- numerická hodnota,
- signálové data,
- obrazová data,
- zvuková data [23, 8].

Textovými daty jsou myšlena data, která mají za úkol popsat či ohodnotit situaci pacienta. Textové ohodnocení může být na základě speciálního kódování, které je standardizováno pro rychlejší vyhodnocení stavu pacienta. Numerická hodnota může také být speciálním kódem, pouze má jinou formu, jeho princip je ale stejný. Ostatní numerické hodnoty jsou pak většinou výsledkem měření z laboratoří nebo např. věk. Signálová data vznikají např. po použití elektrokardiografu (ECG). Obrazové formáty dat pochází většinou z tomografie (CT) nebo z ultrazvukového vyšetření. Zvukový charakter dat může být nahrávka srdeční odezvy nebo dokonce mluvený komentář.

Tato skladba datových typů činí sbíraná data z medicínského prostředí velice náročná na zpracování, primárně v případech, kdy je žádoucí nalézt různé závislosti a vazby. Dalším problémem je zacházení s těmito daty. Jejich znehodnocení (ztráta, fyzická újma dat apod.) může mít katastrofální následky pro pacienta. Proto by medicínská data měla splňovat tyto kritéria: jedinečnost a integritu, úspornost, přesnost a věrnost, veřejnost/ochranu, dostupnost autorizovaným osobám [8].

2.3 Kompletace dat do grafického 3D objektu

Jednou z myšlenek, jak data vizualizovat pro urychlení celkových diagnostik, je nechat data projít speciálním programem, který je zpracuje do specifických bloků. Tyto bloky jsou pak součástí jednoho grafického 3D objektu, který obsahuje vrstvy. Výsledkem je tedy virtuální variace pacienta [22].

Tato myšlenka vznikla na základě velkého nárůstu počtu informací, které jsme schopni zpracovat. Zvýšené množství dat se pak vyskytuje i např. v množství snímků, které během vyšetření jsou přístroje schopny vyprodukovat. Pro srovnání jak gigantický je to obrat oproti minulosti: 100 obrazů o zhruba 50MB v minulosti a 24000 obrazů o 20GB dnes. Další důvod proč je tento software vyvíjen, je časová náročnost prohlédnutí a zanalyzování všech těchto snímků [22]. Program má jednu nevýhodu, a to jsou velké požadavky na grafiku přístroje. Ale spolu s ním je vyvíjen i způsob práce s výsledným objektem. Pro jejich prohlížení existují dotykové pracovní plochy různých velikostí. Využití je možné jak ve vzdělávacích nebo výzkumných institutech tak i do budoucna v nemocnicíc [22].

2.4 Infografika

Jedna z technik zrychlení procesu vnímání a porozumění dat, informací a případně znalostí je používání *infografiky*. Tato metoda se ve světě využívá (např. v dopravě, kdy jednotlivé dopravní značky odpovídají jednomu prvku infografiky) řadu let, jedná se tedy spíš o relativně mladý pojem. V praxi jde tedy o zjednodušení informace pomocí např. symbolů, grafů a diagramů pro snadnější komunikaci. Proto se také spíše využívá ve vztahu k pacientovi jako pomocný komunikační nebo informační kanál [11].

3 Statistické zpracování medicínských dat

V medicíně mají statistické metody velký význam v oblasti rozhodování, kde jsou jejich výsledky využívány k zajištění nejlepší možné péče, k alokaci nákazy v případě epidemie apod. Je to neodmyslitelná součást medicínského výzkumu vedoucí k vývoji např. léčebných metod nebo jejich vylepšení [19].

Statistika v lékařství je rozdělena do 3 fází:

- sběr dat,
- analýza dat,
- statistické usuzování.

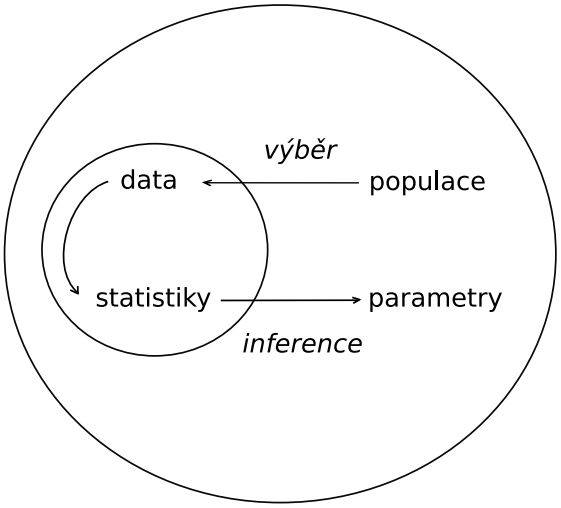
Během první fáze se nejdříve stanovuje soubor otázek, na které hledáme odpovědi, a určuje se cílová skupina (např. lidé postižení stejným defektem). Daný sběr dat se pak soustředí na kvalitu sledovaného vzorku. Od jeho kvality se odvíjí i kvalita výzkumu a přesnost jeho výsledků. Absence kvality je nejčastější bariéra při aplikaci statistické metody. Hlavním měřítkem při sběru vzorku je jeho náhodnost a reprezentativnost. Náhodnost vzorku zajišťuje software, který pomocí pseudonáhodných čísel vygeneruje jeden náhodný vzorek z celkového datového souboru [19].

Ve druhé fázi, tedy v analýze dat, je primárním úkolem vystihnutí podstaty sesbíraných dat, která plní funkci dostatečného shrnutí vlastností dat. Toho je docíleno vhodným statistický softwarem. Nejčastěji se využívají hodnoty deskriptivní statistiky, jako je medián, průměr, modus či směrodatná odchylka [19].

Ke zorientování se v datech nám slouží různá grafická vyjádření. Nejčastěji se při analýze dat používají grafy typu histogram, a pak krabicový nebo bodový graf. Jejich výběr je odvozen od použítých dat [19].

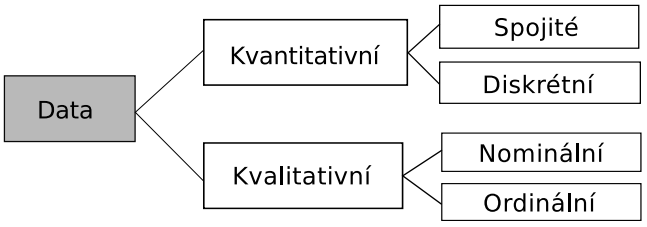
Poslední fází je tedy statistické usuzování (inference viz obr. 3.1). Jedná se o odhad skutečné pravděpodobnosti výskytu události. Hlavní úlohu zde hraje především

už zmiňovaná náhodnost vzorku (výběru). Z jeho zpracování se vyvozují závěry týkající se původce vzorku (množiny pozorovaných - populace). Nahodilost vzorků by měla být taková, že při novém zpracování člověk dojde opět k podobným výsledkům s minimálním rozdílem či ke stejné konstantě. Největší bariérou poslední fáze je schopnost objektivního analytického zhodnocení, ke kterému je potřeba rozumět metodám zkoumání (statistikám) a znát jejich předpoklady [19].



Obrázek 3.1: Proces statistického usuzování

Na konci první fáze jsou připraveny jednotlivé soubory dat, které jsou v rámci jednoho datového souboru heterogenní. Shromažďovaná data mohou být kombinací kvalitativních a kvantitativních dat (viz obr. 3.2). To je důsledek stylizace otázek, která je v oblasti medicínských dat rozličná.



Obrázek 3.2: Základní rozdělení datových typů

Kvantitativní data je možné dělit na spojité a diskrétní. U spojitých se objevují různá čísla z určitého intervalu. Interval může být nějaká množina čísel, která je lo-

gicky ohraničena. Např. věkové rozpětí, teplota, atd. Diskrétní data jsou vyjádřením libovolné četnosti celočíselnou hodnotou, nejsou omezena intervalem.

Kvalitativní data mohou být trojího druhu: binární, nominální a ordinární. V případě binárních dat jsou data reprezentována většinou *true/false* hodnotou, která může být dále vyjádřena např. *ano/ne* nebo *1 a 0*. Obecně jsou data schopna nabývat pouze dvou hodnot. Nominálními daty jsou označena data, v kterých je obsaženo více kategorií bez možnosti seřazení dle významu nebo hodnoty. U ordinálních dat je seřazení možné i přes kategoriální obsah. Jedná se většinou o data, která vyjadřují nějakou určitou stupnici nebo velikost.

3.1 Grafická interpretace statistických dat

Statistika nabízí obecnou množinu grafů nebo tabulek vhodných k vizualizaci výsledků statistického šetření. Tabulky a grafy se dají považovat za primární způsob znázornění celé množiny dat v kontextu, jedná se o způsob sdělení výsledků průzkumu. Tato interpretace může přispět k ucelenějšímu pohledu na danou problematiku (soubor otázek za jehož účelem byl sběr dat uskutečněn), získání nových informací skrze objevené souvislosti, dispozice nebo závislosti [7].

Statistická vizualizace se pak zabývá distribucí jednotlivých datových sad, porovnáváním těchto distribucí s distribucemi jiných datových sad a také samotným porovnáváním datových sad.

3.1.1 Bodový a spojnicový graf

Bodový i spojnicový graf používá k zobrazení hodnoty dvou proměnných kartézských souřadnic. Bodový graf znázorňuje vzájemný vztah dvou proměnných, který se v grafu zobrazí jako množina bodů. Rozložení množiny určuje první umístěný bod, který je zanesen do souřadnic (osa x, y) dle prvních hodnot první a druhé proměnné. Na základě tohoto rozložení je pak graf označován jako pozitivní nebo negativní korelace.

Pozitivní korelace nastává v případě, že přímka po interpolaci směřující od původního bodu do koncového bodu roste (koncový bod má vysoké hodnoty u obou hodnot proměnných). Příklad takové korelace je například srovnání hmotnosti a výšky jedince (viz obr. 3.3 vlevo). Negativní korelace se vyznačuje opačnými vlastnostmi, a to klesajícím charakterem. V tomto případě je hodnota y-souřadnice prvního bodu vysoká a hodnota x-souřadnice koncového bodu nízká [7, 12].

Mluví-li se o perfektní pozitivní nebo negativní korelaci, tak se jedná o proměnné se stálými hodnotami. Příkladem dokonalé pozitivní korelace může být srovnání počtu lidí a vybrané finanční částky za lístek do kina (v případě, že pro všechny platí stejné platební podmínky). Dokonalá negativní korelace nastane např. v případě, kdy je měřen čas strávený na cestě (osa x)a zbylé kilometry (osa y) do cíle (předpoklad konstantní rychlosti auta) [7, 12].

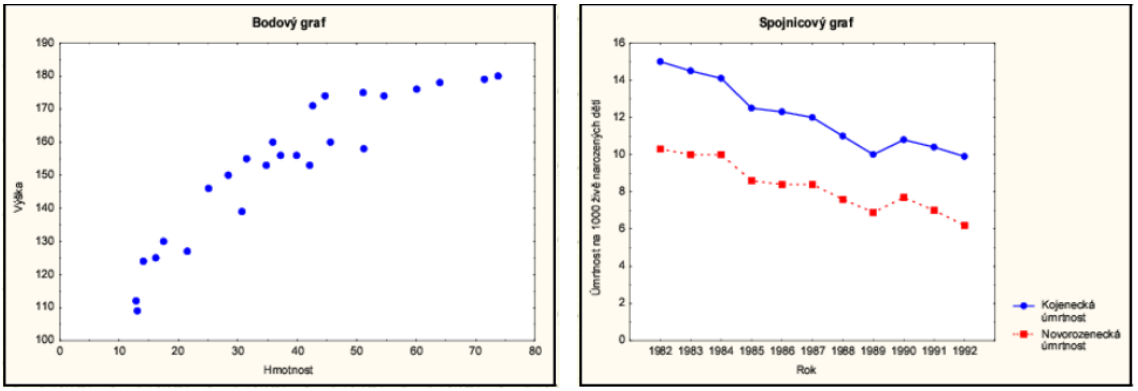
Hodnoty, které nám graf vrací, představují míru korelace. Perfektní pozitivní korelace vrací hodnotu 1. Perfektní negativní korelace vrací hodnotu -1. Pozitivní korelace pak vrací hodnotu v rozsahu $(0, 1)$ podle míry korelace, která může být nízká nebo vysoká. Negativní korelace pak vrací hodnoty v rozpětí $(-1, 0)$, kdy míra korelace roste opačným směrem, než u pozitivní, a to k -1. V případě nuly jde o nulovou korelaci.

Graf je vhodný pro velký objem dat numerického charakteru. Objem dat ovlivňuje přesnost měření, vetší množství dat vede k preciznějšímu výslednému grafu. V případě, že je zkoumáno více hodnot jiných skupin, mohou být použity jiné symboly při vykreslení [7, 12].

Spojnicový graf (obr. 3.3 vpravo) je určen převážně pro data, která jsou závislá na čase. Lze jej použít jako polygon četností, pokud jím je znázorněno rozdělení relativních a absolutních četností spojitého znaku. Tato varianta zobrazení je vhodná pro: zobrazení konkrétních hodnot dat - jedna proměnná může být určena druhou proměnnou, zřetelné zobrazení trendu, kdy je výrazně vidět jak jedna proměnná ovlivňuje druhou a případ, kdy je třeba menší predikce u dalšího dosud neznámého výsledku [7, 12].

Nevýhodou grafu je možnost úpravy jeho zobrazení, kdy může dojít i ke zkreslení výsledků. Upravený graf nemusí odpovídat skutečnosti, i když pracuje s kvalitními

daty.



Obrázek 3.3: Prezentace rozdílu bodového a spojového grafu [7]

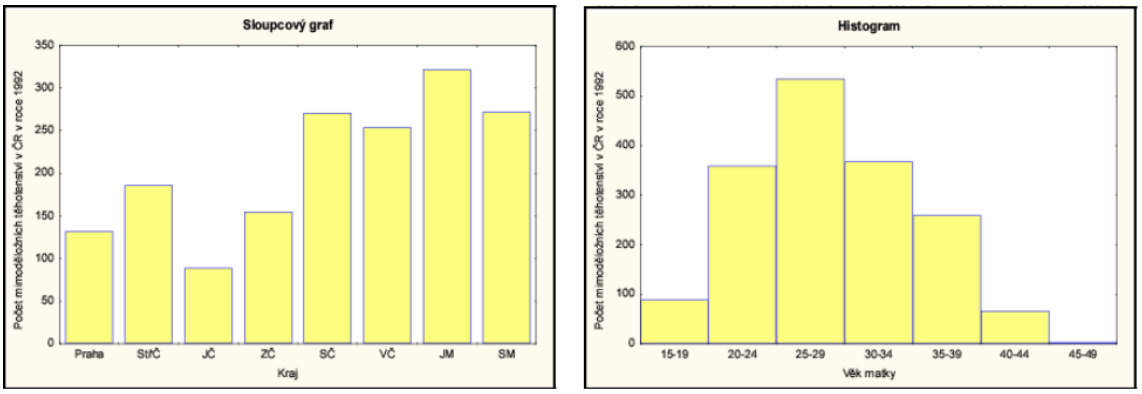
3.1.2 Sloupcový graf a histogram

Sloupcový graf (obr. 3.4 vlevo) zobrazuje data podobně jako dva předchozí grafy, a to za pomocí dvou os, jen je k definici hodnot použit obdélníkový sloupec, jehož výška je v odpovídající velikosti hodnot z množiny vizualizovaných dat. Sloupec může být vykreslen jak horizontálně tak vertikálně. Tento graf je vhodný pro srovnávání různých proměnných, také dokáže jasně zobrazovat trendy a je možné na základě jedné proměnné zjistit hodnotu dalších proměnných [7].

Histogram (obr. 3.4 vpravo) graficky znázorňuje distribuci kvantitativních nebo kategorických dat pomocí sloupců, kde konkretizuje odhad pravděpodobnostního rozdělení spojité proměnné. Sloupce jsou zanesené v grafu pomocí dvou os. Na ose x jsou zanesené dané intervaly (skupiny), které jsou konstantní a určují šířku sloupce. Tento interval je libovolný a je ve většině textů označován jako *"bin"*. Volba tohoto intervalu je důležitá, protože v případě kvantitativních dat může ovlivnit kvalitu informací vyplývajících z grafu. Výška sloupce odpovídá vrcholu dané množiny (intervalu) a je zanesena na ose y, jde o četnost v daném intervalu [7, 13].

Výhodou histogramu je možnost redukce velkého objemu dat na jeden graf, který zobrazí vrcholy (primární, sekundární a terciální) v datech a poskytne tak přehled statistické významnosti těchto vrcholů. Například lze pak z grafu zjistit, zda je prů-

měr či medián v daném datovém souboru signifikantní. V případě nejednoznačné odpovědi ale nelze vyvozovat pevné závěry [7, 13].



Obrázek 3.4: Prezentace rozdílu sloupcového grafu a histogramu [7]

3.1.3 Kruhový graf

Někdy označován také jako *koláčkový* nebo *výsečový graf*. Slouží k zobrazení procentuálního zastoupení různých částí v jednom celku. Kruh grafu odpovídá stům procentům jednoho celku, kdy jednotlivé části jsou zobrazené jako různě velké výseče z tohoto kruhu, které dohromady samozřejmě dávají opět kruh. Lze tak snadno zobrazit zastoupení nějakých skupin v rámci např. nějaké oblasti [7].

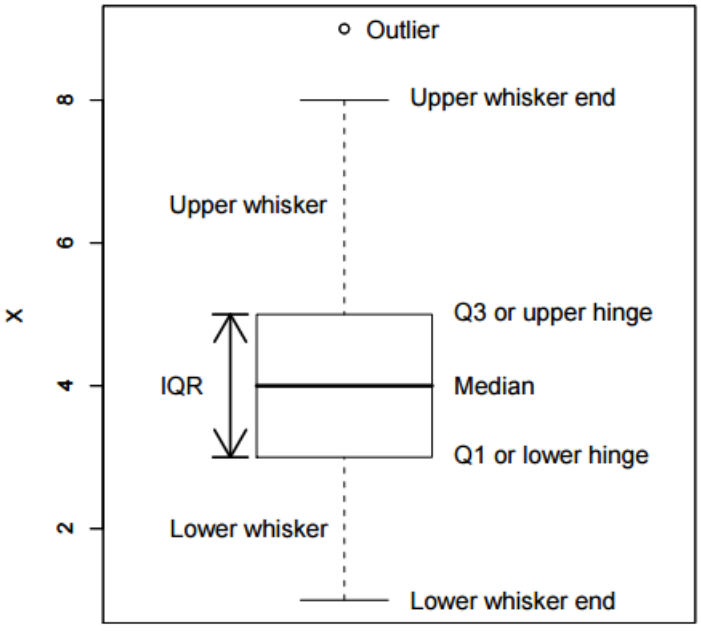
Existují dva případy, kdy tento graf může udávat zkreslené výsledky:

1. Za situace, že je vynechána jedna a více částí z celku.
2. Za situace, že je špatně (nebo není vůbec) definováno, co celý kruh představuje.

V prvním případě dojde ke zkreslení výsledků, protože procentuální zastoupení grafů nebude odpovídat realitě. Tedy procenta z vynechané části se automaticky rozdělí mezi ostatní části. Pokud je špatně definováno, co kruh představuje, tak nemůžeme ani vědět, co jednotlivé výseče představují.

3.1.4 Krabicový diagram

Krabicový diagram (viz obr. 3.5) je používán k vizualizaci číselných dat, a to pomocí jejich kvartilů, které rozdělují statistický soubor na třetiny. Dolní (první) kvartil Q_1 je střední hodnota mezi nejmenším číslem a mediánem. Druhý kvartil je označován jako medián, který rozděluje statistickou množinu dat na dvě množiny o stejné četnosti. Horní (třetí) kvartil Q_3 je střední hodnota mezi mediánem a nejvyšší hodnotou. Hodnota, která vyjadřuje rozdíl mezi třetím a prvním kvartilem, je označována jako mezikvartilové rozpětí (*IQR - Interquartile range*). Svislá osa obsahuje kvantitativní proměnné [7, 14].



Obrázek 3.5: Krabicový diagram [14]

Na obr. 3.5 lze uvést příklad. Jde o jednu množinu zkoumaných subjektů (např. studentů), se kterými souvisí nějaká data (např. počet úspěšně zakončených zkoušek). Medián odpovídá čtyřem zkouškám. To znamená, že polovina studentů má více jak čtyři zkoušky a druhá polovina má méně jak čtyři zkoušky hotové. Ovšem je to jen přibližná hodnota. Dále lze konstatovat, že čtvrtina studentů udělala více jak 5 zkoušek a další čtvrtina méně jak 3 zkoušky. Z toho můžeme odvodit, že polovina studentů je mezi třemi až pěti zkouškami (IQR). Z odlehlé hodnoty (*outliers*) lze

předpokládat, že celkový počet zkoušek je devět. Velikost vousků (*whiskers*) závisí na velikosti vzorku.

Krabicový diagram je vhodný pro neparametrická měření a v některých případech je možné jej použít i na jiné typy dat. V tom případě je pak ale medián chápán jako průměr a místo kvartilů jsou uvedeny násobky směrodatné chyby a extrémy pak uvádějí násobek směrodatné odchylky. [7]

4 Registr SITS

Safe Implementation of Treatments in Stroke dále jen SITS je registr užívaný k sběru dat ohledně pacientů s mrtvicí. Tato data jsou dále zpracovávána a použita k výzkumu a analýze pro zlepšení následné péče a léčby pacientů. Jedná se o mezinárodní registr a sběr dat je zcela anonymní.

Formulář pro sběr dat se skládá z více tzv. kapitol s podbody. Začíná identifikací pacienta a shrnutím informací o něm a jeho hospitalizaci. Ostatní otázky k vyplnění jsou již zaměřeny na průběh a následky mozkové příhody a reakce ošetřujícího lékaře na ni. Tj. informace o podaných lécích a zjištěné informace z vyšetření.

Každá položka ve formuláři představuje statisticky hodnotnou informaci. Vyplnění formuláře je součástí statistického šetření. Přesněji fáze analýzy dat. Formulář je sestaven tak, aby vyplňování nestálo ošetřujícího lékaře zbytečně příliš mnoho času. Téměř vše je v bodech s možností zaškrtnutí jedné varianty. Pro statistické vyhodnocování ideální, většinou číselně zhodnoceno nebo jsou zde kladné/záporné varianty. Některé hodnoty jsou pro porovnávání opět měřeny i s odstupem času. U dat jako jsou *modified rankin scale* a *national institute of health stroke scale* je možné přesně vidět, jak může být průběh sběru dat někdy složitý, a proto je u některých dat zjednodušen výběr anamnézy pouhou stupnicí. Ovšem i to někdy může být zavádějící.

4.1 Modified Rankin Scale (mRS)

Modifikovaná Rankinova stupnice je používána k zhodnocení celkového stavu pacienta po mrtvici. Stupnice je v rozmezí 1-6, kde ke každému číslu je přiřazen stav, ve kterém se pacient může nacházet [16]:

0 Bez symptomů
1 Bez výraznějšího omezení, schopen vykonávat všechny obvyklé denní potřeby a aktivity
2 Lehká invalidita: neschopnost vykonávat všechny dříve obvyklé aktivity, schopen vykonávat všechny své potřeby bez dopomoci
3 Mírná invalidita: vyžaduje pomoc, ale je schopen chůze bez pomoci
4 Středně těžká invalidita: neschopnost chůze bez dopomoci, neschopnost vykonávat tělesné potřeby bez dopomoci
5 Těžká invalidita: upoután na lůžko, inkontinentní, vyžaduje nepřetržitou péči
6 Smrt

Při vyhodnocování lze postupovat systematicky (graficky viz Příloha A). První položenou otázkou je, zda pacient je plně soběstačný a schopen žít sám. V případě že ano, následuje otázka, zda je pacient schopen vykonávat všechny aktivity jako před příhodou. Pokud ne, je mu přiřazeno v stupnici číslo 2, tedy lehká invalidita. Jestliže je však schopen těchto aktivit, následuje poslední otázka, zda je zcela bez defektu. Ano – je označen jako zcela bez symptomů, tedy hodnotou 0. Ne - je označen jako bez výraznějšího omezení, ve stupnici hodnota 1 [17].

Pokud na první otázku zní odpověď ne, je položena otázka, zda je schopen chůze bez pomoci jiné osoby. Z kladné odpovědi dostáváme, že pacientovy následky se dají označit jako mírná invalidita, čili na stupnici hodnota 3. Jestliže je odpověď ne, je dále na místě položení další otázky, zda je, či není upoután na lůžko. V přídě záporné odpovědi je osoba po mrtvici označena jako středně invalidní, tedy stupeň 4. A pokud je odpověď kladná, znamená to pro pacienta těžkou invaliditu, tedy 5. stupeň. Dalším, 6. stupněm je smrt [17].

4.2 National Institute of Health Stroke Scale (NIHSS)

NIH Stroke Scale je standardizované neurologické vyšetření sloužící k zjištění deficitu u pacienta s mozkovou příhodou. Byl vytvořen za účelem homogenního vy-

hodnocení stavu pacienta, a to z důvodu dalšího zpracování dat (porovnávání dat pacientů) [16, 17].

Vyšetření zahrnuje těchto 15 bodů:

1a. Level of Consciousness (*LOC* - úroveň vědomí)
1b. LOC Questions (odpovědi na otázky)
1c. LOC Commands (reakce na příkazy)
2. Best Gaze (schopnost pohybu očí)
3. Visual (zrak)
4. Facial Palsy (ochrnutí částí obličeje)
5a. Motor Right Arm (motorika pravé paže)
5b. Motor Left Arm (motorika levé paže)
6a. Motor Right Leg (motorika pravé nohy)
6b. Motor Left Leg (motorika levé nohy)
7. Limb Ataxia (poškození mozku)
8. Sensory (smyslové vjemy)
9. Best Language (jazyková dovednost, úroveň vyjadřování se)
10. Dysarthria (motorické schopnosti - řeč)
11. Extinction and Inattention - Neglect (orientace)

Postup je jednotný. Pokládají se postupně otázky, na které pacient odpovídá bez vměšování se vyšetřujícího. Otázka se vyhodnotí buď jako správná či nesprávná. První odpověď je ta, s kterou se pracuje, pokud se pacient později opraví, nebere se na to zřetel. Vyhodnocuje se pouze to, čeho je v danou chvíli pacient schopen [16, 17].

První bod *LOC* je povinný, protože jde o úroveň vědomí pacienta. Jiné body lze vynechat pokud nebyl zjištěn žádný nález. Stupnice hodnocení je pak u každého bodu rozdílná (viz. Příloha B).

4.3 Imaging-CT

4.3.1 CT ASPECTS Score

Jedná se o 10 bodové kvantitativní topografické hodnocení a je podotázkou v zobrazování CT (Imaging-CT). *Alberta Stroke Program Early CT Score* se používá jako nástroj k přesnějšímu hodnocení častých známek ischemie nebo-li infarktu, hlavním důvodem používání je sjednocení a větší spolehlivost výsledných závěrů. Je vyhodnocován z původních snímků z CT [16, 17].

5 Časté statistické metody v oblasti medicíny

Součástí práce jsou také statistické metody, které byly voleny na základě diplomové práce Statistické zpracování lékařských dat z roku 2012 [1].

5.1 Analýza dat

Analýza dat je oblast statistiky, která je známá také jako popisná statistika. Je to způsob charakterizace či prezentace dat. Analyzovaná data jsou zpracována většinou za účelem porozumění řešené problematiky. Tohoto cíle lze dosáhnout především je-li znám i kontext nasbíraných dat. Nemalou součástí procesu dospění k pravdivé odpovědi je pak i schopnost porozumět grafickým i výpočetním výstupům pocházejícím ze statistického zpracování dat. Tato skutečnost se ovšem týká celé oblasti statistiky [6, 7].

Pro zobrazování dat jsou využívána tabulková či grafová vyjádření. Tabulkové řešení je vhodné při předpokladu, že bude docházet ještě k dalšímu zpracování naměřených hodnot, nebo pokud je nutné zachovat výsledek statistického šetření v korektním tvaru. Grafické vyjádření je upřednostňováno hlavně v případě, když jsou hlavním zkoumaným faktorem dat kvalitativní vlastnosti. Obecně je grafické zpracování výsledných hodnot vždy přínosem pro pochopení globálního hlediska získaných výstupů [6, 7].

V téhle oblasti zpracování dat se jedná o naměřené veličiny vycházející z většího souboru dat. Jednou skupinou těchto veličin jsou střední hodnoty, resp. míry centrální tendence. Typicky je řeč o aritmetickém průměru, nebo jeho rozšíření jako váženého aritmetického průměru. A dále pak modus a medián, který je oproti aritmetickému průměru více nevšímavý k odlehlým hodnotám [6].

Jejich zobrazení může být pak různé. Většinou jde ale o normální graf např. histogram, který může poskytnout přibližnou informaci ohledně těchto statistických

veličin, ale pro přesnost je daný údaj zvýrazněn dodatečně.

5.2 Neparametrické testy

Neparametrické testy jsou vhodné porovnávání souborů dat, která nemají definované své rozdělení. Jedná se o metody univerzální povahy se sníženou statistickou efektivitou. Výstup těchto metod je pak nutné chápat z obecného hlediska.

5.2.1 Testování hypotéz

Při vyhodnocování statistických testů je nutné stanovení nulové a alternativní hypotézy. Nulová hypotéza (H_o) je tvrzení, které může být formulováno jak s potřebou kladné, tak záporné odpovědi. Záleží, zda chceme dosáhnout menšího okruhu možností, nebo se chceme dozvědět, zda toto určité tvrzení je správné. V konečném důsledku to ale má stejný význam. Alternativní hypotézou jsou pak všechna zbylá tvrzení. Tedy jen nám při svém potvrzení říká, že je H_0 zamítnuta. Nutnost formulace obou hypotéz je tedy bezpředmětná [7].

Výpočtem statistické metody je pak zjištěno, s jakou pravděpodobností bychom mohli dostat pozorovaná data, která by ještě více odporovala H_0, za předpokladu, že je H_0 pravdivá. Vypočtená pravděpodobnost je pak označena jako *dosažená hladina významnosti* p. Důvěryhodnost H_0 je pak závislá na velikosti p, s větší hodnotou p roste důvěryhodnost [7].

Pro rozhodnutí zda se H_0 zamítne či nikoliv, je nutné určit *hladinu významnosti*. Obvykle je to 0,1 nebo 0,05. Pokud je p menší, než je určená hladina významnosti, je H_0 zamítnuta. Tedy nulovou hypotézu zamítneme, pokud p překročí určitou mezní hodnotu [7].

5.2.2 Kruskal-Wallisův test

Kruskal-Wallisův test (dále jen KW) je neparametrickou verzí metody analýzy rozptylu jednoduchého třídění (ANOVA). Tento způsob testování dat je využíván, v případě výběrů z rozdělení, které je značně odlišné od normálního rozdělení. Je aplikován při testování shody zvoleného pravděpodobnostního rozdělení srovnávaných skupin. Data, s kterými pracuje, nevycházejí z normálního rozdělení, jsou na sobě nezávislá a jsou ordinálního typu. Jeden z předpokladů použití této metody jsou data, která obsahují dva a více naměřených údajů [3, 6, 1].

Obecné předpoklady použití KW:

- náhodné vzorky z populací,
- nezávislost každého vzorku a vztahů mezi nimi,
- data alespoň ordinální,
- buď identické distribuční funkce populace nebo populační tendence k dosahování větších hodnot jak u ostatních populací.

V principu jsou zvoleny skupiny (např. žena, muž) vztahující se k datům, která chceme testovat. Je zjištěn stupeň volnosti a zvolena kritická hodnota (χ^2-rozdělení). Data skupin jsou seřazena dle velikosti napříč skupinami, a následně je jim přiřazena hodnota pořadí (dále jen rank). V případě shodných naměřených hodnot se přechází k přiřazení průměru z pořadí. Data jsou nadále zpět rozřazena do svých skupin, ale reprezentována svojí rank hodnotou. Skupiny jsou pak sumarizovány a je určena četnost jejich dat. Po dosazení do vzorce (viz Příloha C) je výsledek porovnán s hladinou významnosti. H_0 je pak zamítnuta nebo přijata na základě tohoto porovnání. H_0 je definována jako skutečnost, že jsou všechny distribuční funkce stejné a alternativní hypotéza pak vyjadřuje skutečnost, že je alespoň jeden soubor z populace, který má tendenci k větším hodnotám než alespoň jedna z jiných populací.

KW k vizualizaci využívá krabicový diagram nebo histogram. V obou případech grafy umožňují porovnávání distribučního rozdělení. Krabicový diagram lze ale považovat za detailnější grafické zpracování obsahující více vyznačených informací. Například medián lze odhadnout i z histogramu, přestože v něm není individuálně

zvýrazněna jeho hodnota. Odhad se ale při menším objemu dat může od skutečnosti odchylovat.

5.2.3 Simultánní porovnávání

Toto porovnávání je zároveň také *post hoc* analýzou, která se používá v případě zamítnutí H_o u předešlé metody. Analýzu je možné provádět, aniž by tomu předcházela specifikace srovnání dat. Princip metody je postaven na porovnávání mediánů statisticky usuzovaných skupin. Pro výslednou hodnotu je potřeba porovnat navzájem všechny skupiny (vzorec viz Příloha C). Pro výslednou vizualizaci je samozřejmostí krabicový graf, který má vyznačený medián [6].

5.2.4 χ^2 test dobré shody

Tento test je neparametrickou metodou, která se používá v případě na sobě nezávislých dat. Základ této metody je v ověření shody usuzovaných četností s četnostmi, které byly vypozorovány (vzorec viz Příloha C). Použitá data musí být kategoriální nebo intervalová. Podle typu pak je možné data rozdělit do kategorií nebo intervalů. Metoda využívá χ^2-rozdělení (pravděpodobnostní rozdělení) [4, 5, 1].

V praxi je porovnávána nominální proměnná s dvěma a více hodnotami. Porovnávány jsou pak pozorované hodnoty s očekávanými hodnotami, které je možné vypočítat prostřednictvím nějakého teoretického očekávání (např. 1:1, kdyby šlo o pohlaví). Testovaná je H_0 - data pocházejí z očekávanáho pravděpodobnostního rozdělení. Alternativní hypotéza je tedy - data nepocházejí z tohoto rozdělení [10].

Pro přesnější výsledky se u této metody doporučuje větší množství dat. V opačné situaci mohou být výsledky nepřesné. Test je aplikovatelný na již zmíněné kategoriální nebo intervalové údaje. Tj. například pohlaví či typ údaje, který posouvá jedince do jisté kategorie nebo na jistou stupnici jako je mRS, která hodnotí stav pacienta [10].

V testu jsou hlavní zkoumanou veličinou četnosti (např. kolikrát se objevuje daný údaj u jedné skupiny za stejných podmínek), proto je logické zobrazování těchto dat

pomocí např. dvou histogramů. Podobnost očekávaných a pozorovaných četností je možné tímto způsobem vyhodnotit i namísto měření.

5.2.5 Randomizační test dobré shody

Je používán, pokud je jeden atribut nominální proměnné se třemi a více hodnotami a pro χ^2 test dobré shody je vzorek dat příliš malý - hodnoty četnosti u 1/4 očekávaných četností je menší jak 5. Test nepoužívá aproximaci χ^2-rozdělením, proto je vhodný v případě, že z jednoho testu dobré shody není možné pro malého množství očekávaných četností dojít správného výsledku. Aproximační vztah tak malého vzorku dat není přesný [9, 1].

Základem randomizační verze tohoto testu je pak opakované měření při ještě menším vzorku dat, kdy počítáme vždy jen s náhodně vybraným vzorkem dat z celého vzorku. Přitom je vždy dodržen poměr naměřených dat. Řešení grafického zpracování tohoto testu pak bude obdobné jako u χ^2 testu dobré shody [9].

6 Implementace v Matlabu

Původním návrhem zpracování praktické části této práce byla realizace zadání v tabulkovém procesoru, přesněji v Excelu, který je součástí souboru kancelářských nástrojů Microsoft Office. Excel je nástroj používaný k analýze dat s velkým množstvím grafů přizpůsobivých povaze analyzovaných dat a zároveň je pro běžného uživatele lehce ovladatelný. Tyto faktory jej staví do pozice atraktivního nástroje. Důvod odstoupení od jeho použití nesouvisí tedy s jeho funkčností ani nabídkou možností pro analýzu dat. Argument, na základě kterého bylo zvoleno jiné řešení, byl nakonec nízký rozsah jeho podpory na více operačních systémech.

Z tohoto důvodu bylo zvoleno interaktivní programové prostředí *Matlab* (verze Matlab R2014a) vyvíjené společností *MathWorks*, které používá vlastní skriptovací jazyk a je rozšířeno i na jiných operačních systémech než je Windows. Podporován je v rámci Windows (32-bit, 64-bit), Mac OS X (64-bit) a Linux (64-bit).

Matlab, přesněji *Matrix Laboratory*, je prostředí sloužící k numerickým výpočtům, grafickému znázornění a programování a jeho využitelnost se najde v mnoha odvětvích (výzkumné, technické, matematické, apod.). Jeho základní datovou strukturou jsou matice, na kterých je založen i jeho jazyk. Matlab je nástoj mnohých vlastností: analýza dat, tvorba algoritmů, simulace, inženýrská a vědecká grafika, inženýrské výpočty, tvorba aplikací apod. Pracovní prostředí je ovládáno interní příkazovou řádkou, která umožňuje okamžité zpracování příkazů nebo také sestavení souboru více příkazů i s funkcemi, které se v závěru provedou hromadně. To umožňuje skriptový soubor s příponou *.m. [20].

Jeho pracovní prostředí umožňuje, v rámci grafiky, vykreslování či zobrazování dvou a tří dimenzionálních grafů, obrázků a animací, a dále pak také vizualizaci dat a webový přístup. Pro práci s externími datovými zdroji je možný export i import textových souborů, tabulkových procesů i o velkém objemu dat [20].

Při pokročilém vývoji softwaru Matlab podporuje objektově-orientované programování a externí rozhraní ($Java, C/C++, .NET, ...$) [20].

6.1 Princip práce s Matlabem

Základem pro porozumění Matlabu je minimální znalost matic a vektorů. Výhoda maticové datové struktury je hlavně v možnosti selekce jednotlivých hodnot pomocí indexů, které také určují pozici hodnoty v datovém bloku. Každá hodnota je opatřena specifickým identifikátorem a lze ji snadno oddělit od zbytku dat.

V případě rozsáhlejších funkcí, které nejsou směrovány pouze na elementární data nebo jednoduché výpočty, je zapotřebí znát i způsoby práce s danými funkcemi. Tedy jaká data jakého typu do nich vstupují a jaké parametrické upřesnění je potřeba. V těchto případech je vhodné nahlédnout do dokumentace na webových stránkách MathWorks nebo použít příkaz *help <nazev přikazu/funkce>*, který vygeneruje nápovědu interním příkazovým řádkem i s odkazy pro podrobnější popis, které se zobrazují v interním webovém prohlížeči.

6.2 Import dat

Základem této práce je práce s daty ve formátu tabulkového procesoru *.xls případně *.xlsx, který obsahuje nasbíraná data. Poskytnutá data k vizualizaci jsou typická pro medicínský obor, tedy různá, ne příliš čistá a v mnohých případech jednoduše s chybějícími údaji. Čistota dat byla opomenuta a práce se jí tedy dále nevěnuje. Práce se zabývá vizualizací dat, ne jejich čištěním, které lze řešit dodatečně specializovanějším a vhodnějším softwarem.

R2014a je vybavena možností importu dat přes grafické uživatelské prostředí (dále jen GUI) Matlabu. Součástí této funkce je zároveň i generování skriptu pro import dat (viz Příloha E), který je možné pak dále použít ve vlastní aplikaci vytvořené v Matlabu. Při generování je možné nastavit i datový typ hodnot v tabulce. V případě že tabulka je naplněna nejen číselnými hodnotami, je vhodné nastavit ji přímo na *cell array*.

Cell array je strukturovaná buňka, která může obsahovat libovolný typ prvku v každé buňce, a je tedy vhodná pro heterogenitu importovaných dat. Cell array sice zachová formáty dat, ale jednotlivá měření je nutné následně převádět na potřebný

typ, protože ostatní jednoduché funkce pracující např. pouze s daty a nepoznají, že jde o cell array, v kterém je numerická hodnota. K tomu byly použity detekční a převodové funkce.

U převodních funkcí se jedná o funkce *cell2mat()* a *num2str()*, které například v tomto pořadí převedou numerická data typu *cell array* na data typu *char*. V případě detekčních funkcí jsou tu funkce *is*()* jako třeba *iscellstr()*, která zjistí, zda jsou data v *cell array* typu *string* nebo nikoliv. K dispozici je i funkce *isa()*, kterou lze upřesnit na vstupu daným datovým typem.

6.3 Statistické funkce

6.3.1 Kruskal-Wallis

Pro Kruskal-Wallisův test je připravena funkce *kruskalwallis()*, která pracuje s vloženými daty ve formě matice, skupinou a také je schopna rovnou údaje graficky zpracovat do podoby krabicového diagramu. Hodnoty, které jsou pak navráceny, jsou p-hodnota, ANOVA (*Analysis of variance*) tabulka tbl (*cell array*) a nebo struktura *stats*, na kterou je možné navázat v dalších testech.

```
[p,tbl,stats] = kruskalwallis(x,dataR,'off');
```

X jsou vstupní data, která podrobujeme hypotéze. Data musí být, jak už bylo řečeno, numerická. V případě, že je třeba použít ordinální data, která nejsou numerické povahy, nahrazují se textové řetězce číselnými kódy. *dataR* jsou kategorická data o stejné délce jako x a ze stejné populace. Tyto data slouží k rozdělení jednotlivých distribučních funkcí. *Displayopt*, který je třetím parametrem, je vypnut pro výrazně snadnější manipulaci s vlastním diagramem. Pokud je tento parametr *'on'* automaticky se vygeneruje krabicový diagram s tabulkou, které ale nelze dle potřeb přesměrovat nebo upravit. Tabulka je zachována v návratovém parametru funkce *tbl*.

Funkce *kruskalwallis()* pracuje v první řadě s jedním nebo s více kvantitativními vektory (soubory). Ale data, která je nutno zpracovat, jsou v cca devadesáti procen-

tech nenumerická (kvalitativní). Některá z nich mají ordinální charakter, proto je smysluplný jejich převod.

Tato data jsou tedy náležitě zpracována funkcí *unique*, která je primárně využívána k vyhledávání unikátních záznamů. Funkce obsahuje vícenásobné výstupy. V kombinaci s nastavením *sorted* je schopná přiřadit číselnou hodnotu, která plní substituční funkci pro původní data. Hodnota odpovídá pořadí po seřazení původních nenumerických dat.

```
[u,~,x] = unique(num,'sorted');
```

Pro povahu dat byla zohledněna možnost práce s věkovým údajem tedy i pro případy, že je nutné použít *dataR* obsahující spojitá data o velkém intervalu. Byla přidána možnost grupování údajů, která data seskupí dle zadaných parametrů.

```
e=begin:middle:endx;
labels = strcat(num2str((begin:middle:(endx-middle))',,'%d')
    ,{'s'});
dataset = ordinal(cell2mat(dataS),labels,[],e);
```

Vstupní parametr *e* pro funkci *ordinal()* je volena uživatelem pomocí funkce *inputdlg()*, tedy dialogovým oknem pro zadávání vstupních dat. Vstupními údaji jsou hraniční body rozpětí intervalu a specifikace seskupení, která je nejlépe volena dle velikosti intervalu. *e* pak může obsahovat hodnoty např. *0:10:100*. Další vstupní parametr *labels* označuje obsahuje názvy skupin, které vyplývají z podstaty *e* (např. 10s, 20s, 30s, atd.). Posledním hlavním vstupem je pak vybraný datový soubor, kterému jsou na výstupu funkce *ordinal()* přiřazeny řetězce z *labels*, které značí skupinu, do které daná hodnota náleží. Toto řešení využívají všechny použité statistické metody a lze je přeskočit. V případě, že není nutné grupování stačí ponechat vstupní hodnoty nevyplněny a potvrdit dialog.

Hodnoty na výstupu jsou vypisovány pomocí komponenty *static text*. Tento blok dat obsahuje informace o výsledku Kruskall-Wallisova testu a pravděpodobnost, na základě které uživatel sám rozhoduje zamítnutí/nezamítnutí H_0.

6.3.2 Simultánní porovnávání

U *post hoc* analýzy se používá funkce pro vícenásobné porovnávání *multcompare()*. Na jejím vstupu jsou nutná data z předešlé metody, v tomto případě se jedná o data pocházející z funkce *kuskalwallis()*. Funkce jako taková ještě není přesným vyjádřením simultánního porovnávání, předpokladem je nastavení *CType* na *scheffe*. Pro přesnější výsledek byla přidána ještě možnost nastavení *post hoc* analýzy na *Turkeyho metodu*. Její použití se omezuje na data se symetrickým tříděním.

```
[c,m,~] = multcompare(stats,'CType','scheffe','Display','off')
```

Data, která jsou během post hoc analýzy zpracována, jsou tedy jedním z výstupů Kruskal-Wallisova testu. Jde o výstup *stats*, který obsahuje strukturovaná data (informace o porovnávaných skupinách - mediány, jejich četnosti atd.) daného testu. Tento výstup metody je primárně určen k dalšímu zpracování v případě neuspokojivého výsledku – zamítnutí H_0.

Výčet získaných hodnot je zprostředkován pomocí komponenty *uitable*. Obsahuje jednotlivé pravděpodobnosti vycházející z porovnávání jednotlivých skupin, rozdíly jejich průměrů a jejich horní a dolní intervaly. Funkci předchází tedy KW test, na který pak volně navazuje s upřesnění typu kritické hodnoty CType.

6.3.3 χ^2 test dobré shody

Při použití χ^2 testu dobré shody je tu možnost využít funkci *chi2gof()*, která pracuje s daty, specifikací typu dat a nastavením jejich hodnoty. Na výstupu jsou pak hodnoty jako výsledná hypotéza, p-hodnota a struktura.

```
[tbl,chi2,p,labels] = crosstab(dataR,dataS);
```

Data, která zpracovávána v této funkci jsou nominálního typu, proto je třeba jejich úprava před vložením do funkce. Pro modifikaci zvolených dat byla vyžita funkce *crosstable()*. Funkce na výstupu poskytuje dva parametry s daty, která jsou dále zpracována jako vstupní data pro hlavní funkci testování. První data jsou obsahují vzájemné četnosti položek, další jsou jejich popisky tzv. labely. Data s četnostmi

jsou použita k zjištění pozorovaných (označení pro matlab je *Frequency*) a očekávaných (v matlabu *Expected*) četností, které se následně vkládají do hlavní funkce *chi2gof()*.

```
[h,p,st] = chi2gof(bins,'Ctrs',bins, 'Frequency',obsC, '
    Expected', expC, 'Alpha', hladina);
```

Data oznamující výsledek metody jsou informace stavu hypotézy, zda byla zamítnuta či nikoliv, a pravděpodobnost s hladinou významnosti, v rámci kterých se došlo předešlého statusu. *obsC* obsahuje data z výběru uživatele. Tedy je použita funkce *unique()*, která poskytne možnosti uživatelova výběru. Zvolená skupina je pak porovnávána očekávaným rozdělením s *expC*. Hladinu významnosti voli uživatel. Výsledné hodnoty jsou opět na výstupu jako statický text.

6.3.4 Randomizační test dobré shody

U této části se opakují všechny funkce z předchozího testování. Funkce náhodného výběru je zastoupena funkcí *randi()*. Počet opakování je nastaven na 1000 a provádí ji cyklus *for*. Výsledkem je soubor se stejným počtem pravděpodobností kolik je opakování. Tato data jsou následně porovnána s původní pravděpodobnostní hodnotou.

Na výstupu je standardně informace o zamítnutí/nezamítnutí H_0 a jaké soubory byly využity. Celý test je prováděn s hladinou významnosti 0,5.

6.4 Statistické grafy

Matlab je vybaven funkcemi s grafickým výstupem. Je tak možné svá data vyjádřit i v jiné než textové/číselné formě. V prostředí jsou obsaženy základní grafická vyjádření [21].

6.4.1 Kruhový graf

Kruhový graf vykresluje funkce *pie()*, která potřebuje na vstupu pouze nějaký datový blok, nejlépe kvalitativního charakteru. Na vstupu jsou tedy buď kvantitativní data nebo kvantitativní a kvalitativní data(ty pak slouží primárně k popisu grafu). Úprava vstupního souboru dat je ošetřena skrze funkci *tabulate()*,která obsahuje procentuální shrnutí dat. To je pak využito jako textový výstup ke analýze dat, která je další možností zpracování dat. Tabulka obsahuje ke každému unikátnímu záznamu jeho četnost výskytu a procentuální zastoupení v rámci zvolených dat.

6.4.2 Bodový graf

Bodový graf, který je znám také jako korelační diagram je zahrnut v možných funkcích řešení, ale není poskytnut uživateli. Jeho využitelnost i pro jeho možnou obměnu ve spojovací graf, je potenciálně přínosná u úplných dat. Bohužel i přesto, že byla čistota dat opomíjena, s poskytnutými daty nebyla nalezeno vhodné využití. V případě úplných časových údajů by bylo vhodné graf poskytnout jako jednu z uživatelových možností. Funkce pro bodový graf v Matlabu je *scatter*. Pro další práci s ním je nutná jistá úprava. Nabízí možnosti vykreslování daných bodů od koleček, po znaménka matematických operací nebo diamanty. Pro vyznačení např. určité hodnoty je možné použít funkci *refline()*, do které je nutné zadat jen souřadnice pro vykreslení linky.

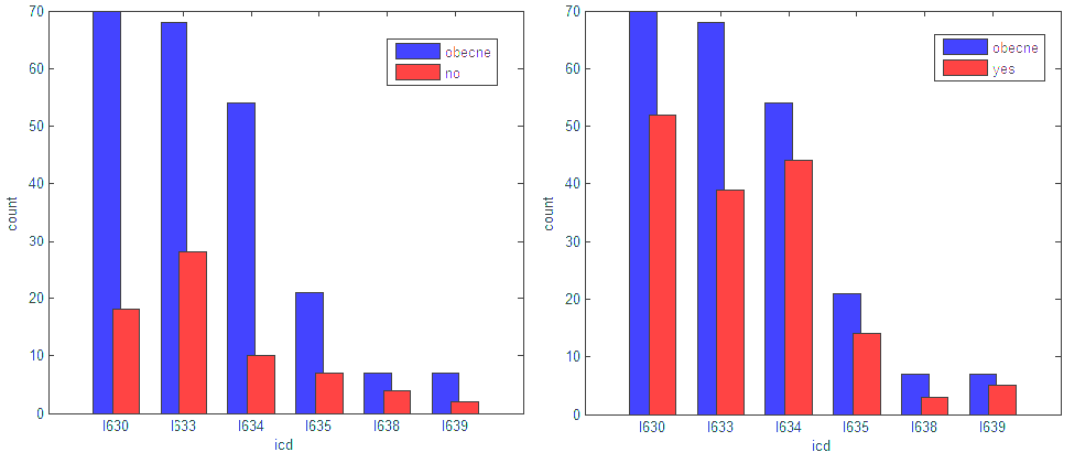
6.4.3 Histogram

Histogram je další velice variabilní možností grafického zpracování dat. V Matlabu je pod funkcí *hist()*. Je v něm možné nastavení, v kolika sloupcích budou daná data vykreslena. To je vlastnost, se kterou je docíleno přesnějšího vyjádření dat.

Pro vhodnější zpracování byla i přesto zvolena funkce *bar()*, která je schopna pracovat s kategoriemi. Data na vstupu jsou pouze kvantitativního typu. Pro zpracování dat na vhodný vstupní formát byla použita funkce *tabular()*, která ze souboru dat sumarizuje procentuální zastoupení jednotlivých položek v souboru. Je to taky

hlavní funkce analytického zpracování dat. Způsob převodu dat se v případě χ^2 liší pouze v počtu zpracovávaných kategorií. Hlavní kategorie, se kterou se v tomto případě porovnává, jsou celková data, a to se zvolenou kategorií. Tuto volbu provádí uživatel.

V případě χ^2 testu dobré shody je výstup, kdy zjišťujeme, zda má hypertenze (vysoký tlak) vliv na rozložení četností ICD skupin na Obr. 6.1. První graf vyjadřující data, která není zamítnuta pouze na hladině významnosti 0,5. Při hladině významnosti 0.1 už by byla zamítnuta. Totéž naznačuje graf vlevo, na kterém je větší rozdíl u prvního sloupce. Pokud by jsme si představily vrcholy daných četností jako křivku, tento menší rozdíl by byl očividnější.



Obrázek 6.1: Vliv hypertenze na ICD

U druhé skupiny na Obr. 6.1 vpravo je větší podobnost mezi četnostmi, pokud se pro srovnání použijí obecné nekategorizované četnosti souboru dat. Ty naznačují distribuční křivku, a tedy i jak by měli vypadat očekávané četnosti. Data (kategorie lidí s hypertenzí) jsou z pohledu na graf více podobna jejich předpokladu. To potvrzuje i fakt, že nulová hypotéza nebyla zamítnuta na hladině 0.01. Ovšem, vezme-li se v potaz skutečnost, že u některých četností se suma minima z celku dostává sotva na 10, bylo by vhodnější užití i jiné srovnávací metody. Data mohou být zkreslena menší objemností dat.

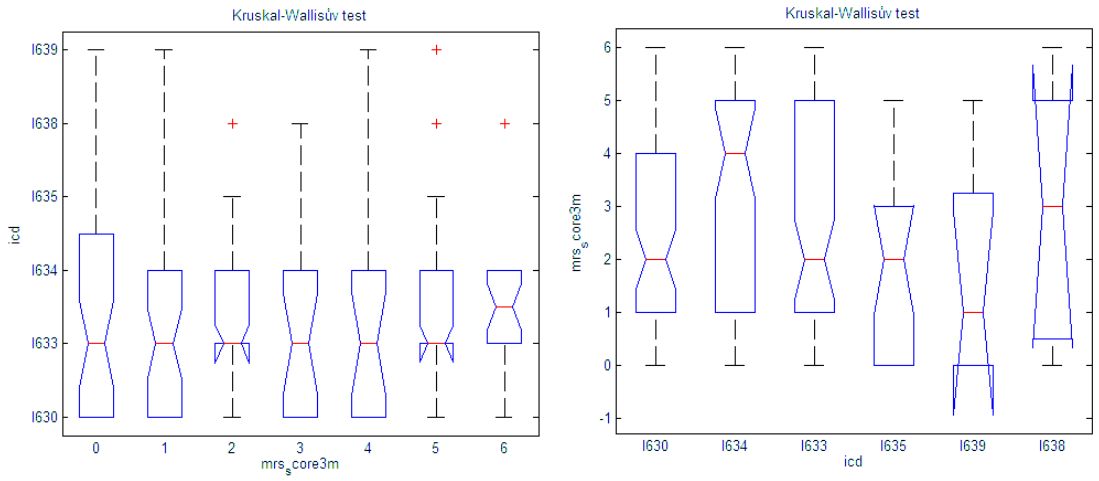
6.4.4 Krabicový diagram

Krabicový diagram je zde pod funkcí *boxplot)*. Výstup je rozdílný na základě vstupních dat (ne jen v případě různých hodnot nacházejících se v souboru dat), a to v množství výstupních diagramů. V případě souboru dat, který je maticí, je na výstupu tolik krabicových diagramů, kolik je sloupců v matici. Při datech o jedné sledované proměnné tedy o jednom vektoru je na výstupu jen jeden diagram.

Funkce je plněna daty (x), které mají různý výskyt. V případě textových dat byla použita pro zastoupení funkce *unique()*, která nevrací jen unikátní záznamy, ale také číselné kódování vyskytujících se řetězců.

```
boxplot(x,dataR,'notch','on');
```

V případě prvního grafu (Obr. 6.2) nelze odmítnout nulovou hypotézu, že je rozložení ICD u mRS skóre stejné, a to na hladině významnosti 0,5. Tuto skutečnost lze usuzovat i z levého grafu. Z druhého grafu je evidentní, že rozložení mRS skóre v rámci ICD skupin je více různorodé, a proto je možné zamítnou H_0 na hladině významnosti 0,1. Ovšem na hladině významnosti 0,01 ji nezamítáme. V případě přesnějšího výsledku, tedy které skupiny se liší natolik, že snižují celkovou pravděpodobnost, by bylo vhodné použít *post hoc* analýzu.



Obrázek 6.2: Ukázka dvou krabicových diagramu u Kruskal-Wallisova testu

6.5 Analýza dat

V případě analýzy dat je jednoduché vyhledání tří základních veličin statistiky, a to aritmetického průměru, mediánu a modusu s doplnění informace o procentuálním zastoupení hodnot vybraného souboru. Hlavní funkce zde jsou *mean()* pro aritmetický průměr, *median()* pro medián, *mode()* pro modus a *tabulate()* pro procenta.

6.6 Grafické uživatelské prostředí

V pracovním prostředí Matlab je možnost vytvoření si vlastní aplikace nebo GUI, které při používání vlastních či knihovních funkcí usnadňuje práci širšímu spektru uživatelů. Vytvoření GUI je snadné a přístupné ze základního pracovního prostředí. Při otevření nového souboru v návrhovém prostředí (dále jen GUIDE - *graphical user interface development environment*) se objeví možnosti čisté nebo již přednastavené plochy (základní ovládací prvky, graf s menu, přednastavený dialog; viz Příloha D). Souboru s GUI pak náleží přípona *.fig.

Okno pro nastavení GUI obsahuje panel s nástroji pro rychlejší práci. Panel obsahuje základní prvky (také viz Příloha D): Push Button, Slider, Radio Button, Check Box, Edit Text, Static Text, Pop-up Menu, Listbox, Tooggle Button, Table, Axes, Panel, Button Group, ActiveX Control.

Všechny prvky jsou velice variabilní a mohou být využity standardním i nestandardním způsobem dle uživatelových schopností. Neznamená to však, že všechny nápady lze zrealizovat. V případě propojení slideru lze narazit na omezení v podobě možnosti kombinace jen s komponentami vracejícími *string* hodnotu. Při sázení prvků na čistou plochu (*figure*) se automaticky generuje základní kód do skriptového souboru s příponou *.m, který pak obsahuje všechny inicializace komponent. Jejich propojení zajišťují vygenerované funkce s parametry *hObject*, *eventdata* a *handles* obsahující informace o komponentách.

V řešení bylo využito téměř všech komponent: *push button* pro potvrzování a provádění hromadnějších příkazů, *edit text* pro zadávání uživatelské modifikace, *table*

jako vizuální pomůcka při výběru proměnných, *list box* pro uživatelův výběr metody zpracování, *pop-up menu* pro upřesnění zpracování parametru a další výběr z nabízených možností aplikace, apod. Výstup grafů zachycuje tzv. axes, který je nastaven na *zoom on* a umožňuje tak přiblížení si jednotlivých bodů grafu. V případě, že výběr způsobu testování je *analýza dat* je možné přepínat pomocí *button group* mezi kruhovým grafem a histogramem.

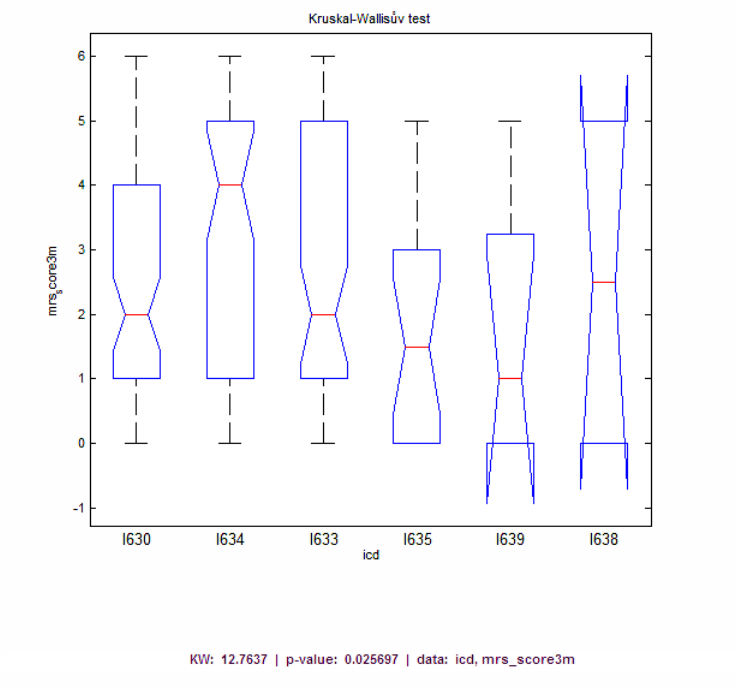
7 Testování

Testování je nutnou součástí vývoje softwaru. Pro tuto realizaci bylo zvoleno testování programátorem (kód), funkční testování (reakce GUI prostředí) a shodnost výpočtů s grafickým vyjádřením.

Testování kódu bylo prováděno průběžně po celou dobu vývoje. Všechny nalezené chyby byly odstraněni buď odstavením nefunkční části nebo její úpravou. Kontrola byla prováděna na používaných funkcích i na jejich vstupních proměnných.

Ostatní kontroly probíhaly paralelně vždy v rámci jedné metody. V rámci části funkčního testování byly nalezené chyby ihned opraveny. U části *analýza data* byla pro svoji triviálnost korektnost výpočtu ošetřena již při předchozích testech.

7.1 Kruskal-Wallisův test



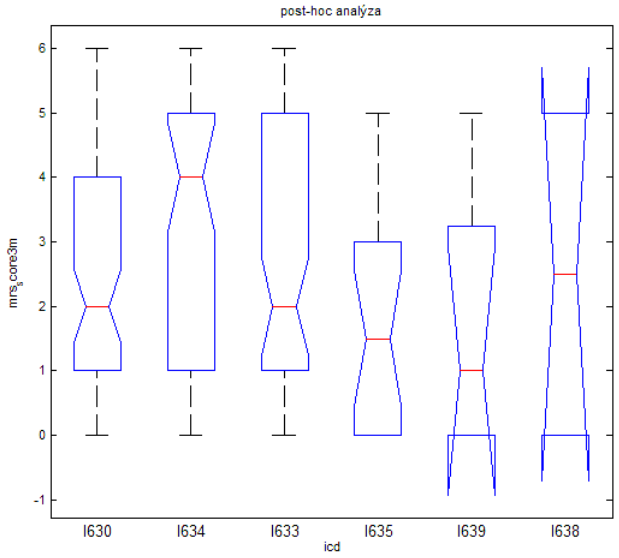
Obrázek 7.1: Výstup KW - ICD a mRS

Test údajů KW testu s daty ICD a mRS, ve kterém zjišťujeme zda mají všechny kódy ICD stejnou distribuční funkci.

Graf na Obr. 7.1 odpovídá se nerozchází s výsledky KW testu. Distribuční funkce jsou velmi podobné, ale přesto je na hladině významnosti 0.01 zamítáme a můžeme použít metodu post hoc.

7.2 Simultánní porovnávání

Test metody post hoc analýzy, která má dvě modifikace, s daty z předchozího testování.



Skupina	Skupina	P-value	Dolní interval	Rozdíl průměrů	Horní interval
1	2	0.1955	-69.7408	-31.3184	7.1041
1	3	0.9212	-49.3186	-13.0333	23.2521
1	4	0.9844	-40.3923	13.2279	66.8482
1	5	0.9652	-58.9448	24.7279	108.4007
1	6	0.9999	-96.8777	-7.1054	82.6669

Obrázek 7.2: Výstup post hoc metody (scheffe) - ICD a mRS

Na Obr. 7.2 je vidět jen část tabulky, která ukazuje porovnání I630 s ostatními kódy. Z obrázku je znát rozdílnost této první skupiny oproti ostatním a naměřené hodnoty tomu odpovídají.

U druhého typu post hoc analýzy *hsd* je graf stejný a pouze se liší údaje v tabulce.

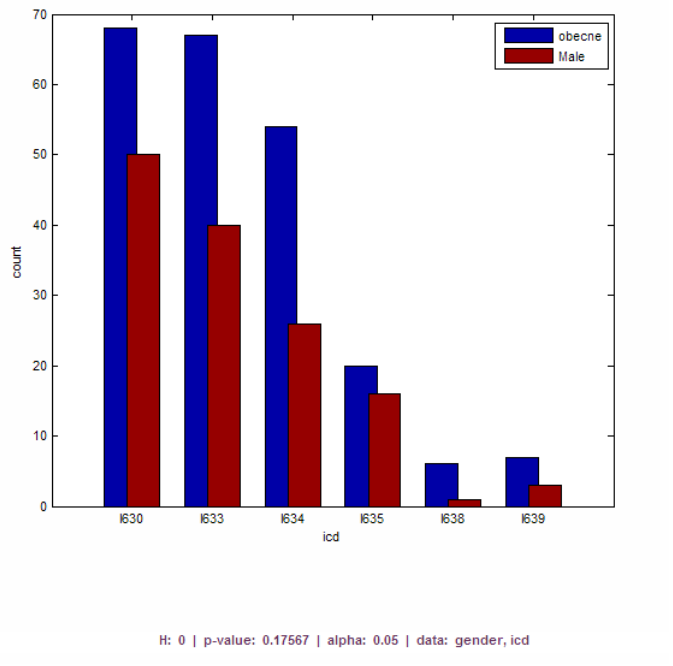
Skupina	Skupina	P-value	Dolní interval	Rozdíl průměrů	Horní interval
1	2	0.0728	-64.2264	-31.3184	1.5897
1	3	0.8394	-44.1109	-13.0333	18.0444
1	4	0.9638	-32.6966	13.2279	59.1525
1	5	0.9234	-46.9359	24.7279	96.3918
1	6	0.9998	-83.9934	-7.1054	69.7827

Obrázek 7.3: Tabulka post hoc metody (hsd) - ICD a mRS

Na základě Obr. 7.3 je viditelnější rozdíl oproti *scheffe* metodě při porovnávání první a druhé skupiny. Je větší pravděpodobnost podobnosti funkcí. Na hladině významnosti 0.1 nemusí být zamítnuta.

7.2.1 χ^2 test dobré shody a randomizační test dobré shody

Test metody na datech pohlaví a icd, kde je sledováno zda má pohlaví vliv na distribuční rozdělení icd.



Obrázek 7.4: Výstup χ^2 - gender, icd.

Z Obr. 7.4 vychází nulová hypotéza na hladině 0.05 jako zamítnutá. Z obrázku

je čitelná podobnost distribučního rozdělení a dle hodnoty *p-value* lze konstatovat, že na hladině 0.5 by hypotéza nebyla zamítnuta.

U randomizačního testu dobré shody, byly výsledky podobné jako u metody χ^2, hlavně z důvodu stejného vzorku dat.

8 Závěr

Během této práce byla stěžejní manipulace s daty. Pro základní znalosti obecného získávání dat pro statistické účely bylo nutné se seznámit s jednotlivými fázemi statistického šetření. Poskytnutá data byla původem z registru SITS. Z toho důvodu byl popsán význam tohoto registru, způsob zadávání těchto dat do systému, průběh získávání dat a samotný formulář. Formulář byl základním zdrojem informací, protože bylo možné na základě jeho obsahu lépe identifikovat poskytnutá data.

Součástí aplikace měla být možnost výpočtu a nejen grafického výstupu, proto bylo nezbytné se alespoň v principu seznámit s vhodnými metodami. Byly vybrány na základě předchozí práce, jež se zabývala vhodnými metodami v oblasti zpracování lékařských dat. V rámci implementace bylo náročnější částí pochopení těchto metod, a to celý proces zpracování dat do podoby, kterou by daná funkce byla schopna otestovat. A to i přesto, že předpokladem jsou čistá data. Hlavním úskalím proto byl obecně Matlab, který sice poskytuje velké množství funkcí, ale byla zde ve většině případů nezbytná jistá režie dat před jejich vstupem.

Program obecně splňuje všechny základní požadavky, které byly zadány v rámci této práce. V případě osvědčení jeho funkčnosti v rámci cílové skupiny je další vývoj s větší ovladatelností uživatele určitě na místě. Např. volba grafu v případě pokročilejšího uživatele (v programu pouze u analýzy dat), možnost jisté modifikace samotných metod, která by ve výsledku znamenala více prostoru pro statistické testování dat.

Přes splnění požadavků nepovažuji volbu Matlabu za vhodnou. Převážně z důvodu naražení na bariéry, které posunuli realizaci o velkou část práce opět zpět k začátku. Pro případný další vývoj by bylo na místě zvolit jinou variantu, ať už by šlo o jiný software nebo programovací jazyk. Nebylo by naškodu zvolit jiný způsob realizace této myšlenky.

Seznam obrázků

Literatura

[1] HEJNA, Miroslav. *Statistické zpracování lékařských dat.* Plzeň, 2012. Dostupné z: `https://otik.uk.zcu.cz/bitstream/handle/11025/3044/Hejna_DP.pdf`. Diplomová práce. Západočeská univerzita, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky.

[2] BUI, Alex A.T. a William HSU. *Medical imaging informatics.* New York: Springer, 2010. ISBN 1441903852-. Dostupné z: `http://www.mii.ucla.edu/~willhsu/pubs/bui.mii.ch4.pdf`

[3] ANDĚL, Jiří. Statistické metody. 4., upr. vyd. Praha: Matfyzpress, 2007, 299 s. ISBN 978-80-7378-003-6.

[4] HEBÁK, Petr, Jiří HUSTOPECKÝ, Eva JAROŠOVÁ a Ivana MALÁ. *Vícerozměrné statistické metody.* Vyd. 1. Praha: Informatorium, 2004-2005, 3 sv. ISBN 80-7333-025-3.

[5] ANTOCH, Jaromír a Dana VORLÍČKOVÁ. *Vybrané metody statistické analýzy dat.* 1. vyd. Praha: Academia, 1992, 279 s. ISBN 8020002049.

[6] HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat.* Vyd. 1. Praha: Portál, 2004, 583 s. ISBN 8071788201.

[7] ZVÁROVÁ, J. *Základy statistiky pro biomedicínské obory.* Vyd. 1. Praha: Karolinum, 2002, 218 s. ISBN 80-718-4786-0. Dostupné z: `http://new.euromise.org/czech/tajne/ucebnice/html/html/statist.html`

[8] DRGÁČ, Radim. *Metody zpracování medicínských dat.* Brno, 2006. Bakalářská práce. Masarykova univerzita. Vedoucí práce Mgr. Miroslav Kubásek. Dostupné z: `http://is.muni.cz/th/99227/fi_b/bakalarka.pdf`

[9] Haandbook of Biological Statistics: Randomization test of goodness-of-fit [online]. 2009. [cit. 2015-06-11]. Dostupné z: http://udel.edu/~mcdonald/statrand.html

[10] Chi-square test of goodness-of-fit: Haandbook of Biological Statistics [online]. 2014. [cit. 2015-06-09]. Dostupné z: http://www.biostathandbook.com/chigof.html

[11] Infografika *Wikisofia*[online]. 2013 [cit. 2016-03-24]. Dostupné z: https://wikisofia.cz/index.php/Infografika

[12] Scatter Plots *MSTE, University of Illinois* [online]. 2016 [cit. 2016-04-20]. Dostupné z: http://mste.illinois.edu/courses/ci330ms/youtsey/scatterinfo.html

[13] Histograms: Construction, Analysis and Understanding *Quarknet* [online]. 2002 [cit. 2016-04-20]. Dostupné z: http://quarknet.fnal.gov/toolkits/ati/histograms.html

[14] BROWN, Steven R. *Experimental design and analysis* [online]. Newbury Park: SAGE Publications, 2015 [cit. 2016-04-01]. Quantitative applications in the social sciences, 07-074. ISBN 08-039-3854-3. Dostupné z: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

[15] Histogram plot. *MathWorks: MATLAB and Simulink for Technical Commputing* [online]. 1994, 2015 [cit. 2015-06-10]. Dostupné z: http://www.mathworks.com/help/matlab/ref/hist.html

[16] MUDr: Lékařské klasifikace • Online kalkulačky • Skóre • Tabulky • MKN v.2 [online]. 2008-2009 [cit. 2014-11-16]. Dostupné z: http://www.mudr.org

[17] Diagnostický a terapeutický manuál cévních onemocnění mozku [online]. 2009, 2014 [cit. 2014-11-23]. Dostupné z: http://cmp-manual.wbs.cz/

[18] ABZ.cz: slovník cizích slov - on-line hledání [online]. 2005, 2014 [cit. 2014-11-27]. Dostupné z: http://slovnik-cizich-slov.abz.cz/

[19] BĚLÁŠKOVÁ, Silvie a Lenka BLAŽKOVÁ. *Moderní analýza medicínských dat.* [online]. 2010 [cit. 2014-11-07]. Dostupné z: http:

//www.systemonline.cz/it-pro-verejny-sektor-a-zdravotnictvi/moderni-analyza-medicinskych-dat.htm

[20] MATLAB documentation *MathWorks: MATLAB and Simulink for Technical Commputing* [online]. 1994, 2015 [cit. 2015-03-20]. Dostupné z: http://www.mathworks.com/help/matlab/index.html

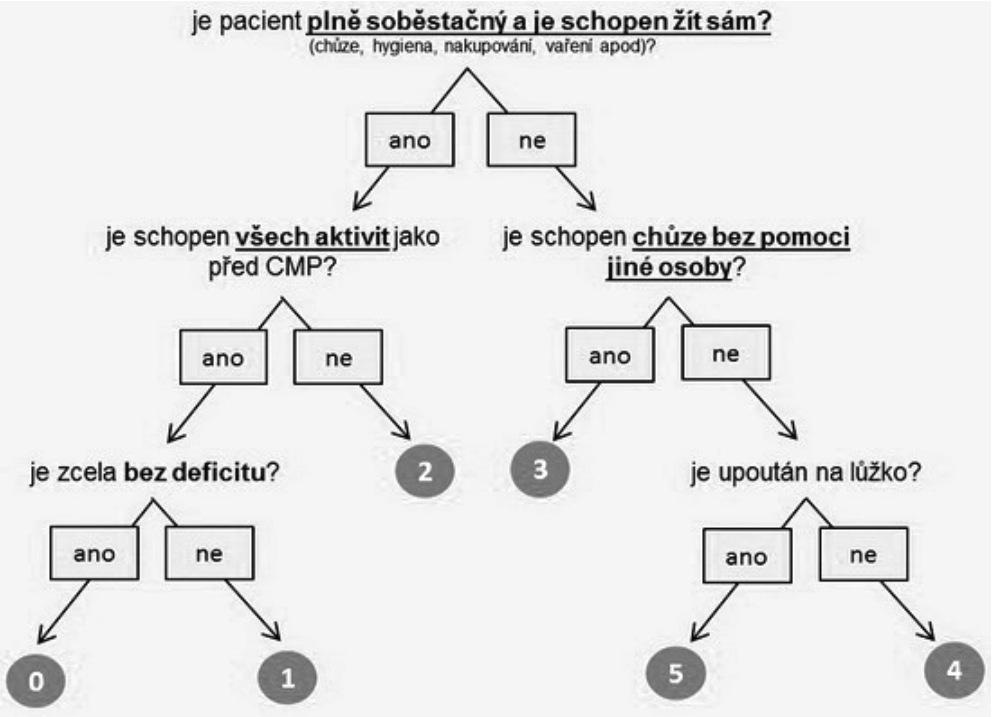
[21] Graphics *MathWorks: MATLAB and Simulink for Technical Commputing* [online]. 1994, 2015 [cit. 2015-03-24]. Dostupné z: http://www.mathworks.com/help/matlab/graphics.html

[22] Anders Ynnerman: Visualizing the medical data explosion *TED: Ideas worth spreading*[online]. 2010 [cit. 2015-06-16]. Dostupné z: http://www.ted.com/talks/anders_ynnerman_visualizing_the_medical_data_explosion/transcript?language=en

[23] SZOLOVITS, Peter. *Medical Informatics Computer: Computer Applications in Health Care* [online]. 1997 [cit. 2015-06-16]. Dostupné z: http://groups.csail.mit.edu/medg/courses/6872/96/notes/Tsien/

Přílohy

A Vyhodnocování mRS



B Stupnice vyšetřovaných bodů NIHSS

Level of Consciusness	0	plně při vědomí, spolupracující
	1	spavý, po mírné stimulaci poslechne, odpoví
	2	opakovaná stimulace k pozornosti, sopor
	3	koma (reflexní či žádná odpověď)
LOC Questions	0	obě odpovědi zcela správně
	1	jedna správně, těžká dysarthrie či jiná bariéra (OTI)
	2	obě špatně, afázie, kóma
LOC Commands	0	oba úkoly správně
	1	jeden úkol správně
	2	žádný správně, kóma
Best Gaze	0	bez patologie
	1	izol. paresa okohybného nervu, deviace či pohledová paresa potlačitelná OC manévry
	2	nepotlačitelná deviace či pohledová paresa
Visual	0	bez postižení
	1	částečná hemianopsie, fenomén extinkce
	2	kompletní hemianopsie
	3	oboustranná hemianopsie (slepota,včetně kortikální slepoty)
Facial Palsy	0	symetrický pohyb, bez postižení
	1	lehká paresa (např. asymetrie NL rýhy)
	2	úplná nebo částečná paréza dolní větve (centrální paresa)
	3	kompletní (perif.) paréza uni- či bilaterální, koma
Motor Arm/Leg	0	bez kolísání

	1	kolísání nebo pokles, bez úplného pádu na podložku
	2	určitý pohyb proti gravitaci, neudrží nad podložkou
	3	pohyb po podložce
	4	plegie, bez pohybu, koma (pro všechny konč.)
	9	amputace, ankylóza aj. příčiny patolog. nálezu nesouvisející s příhodou.
Limb Ataxia	0	nepřítomna, nebo jen důsledek paresy. Koma.
	1	na jedné končetině
	2	přítomna na více končetinách
	9	amputace, ankylóza aj.
Sensory	0	bez poruchy čití
	1	lehká a střední porucha sense (hypestezie,hypalgezie)
	2	těžká porucha sense až anestezie uni, či bilat. Koma.
Best Language	0	bez afázie
	1	lehčí fatická porucha, lze porozumět
	2	těžká fatická porucha
	3	globální afázie, mutismus, kóma
Dysarthria	0	nepřítomna
	1	setřelá řeč, je mu rozumět
	2	výrazně setřelá výslovnost, není rozumět, mutismus, kóma
	9	intubace, jiná bariéra
Extinction and Inattention - Neglect	0	nepřítomen
	1	neglektuje 1 kvalitu, anosognoze
	2	neglektuje více jak 1 kvalitu, kóma.

C Vzorce statistických metod

Krukal-Walllisuv test

$$H_{KW} = \frac{12}{n(n-1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3n(n+1)$$

n_i	četnost skupiny
R_i	suma hodnot skupiny

χ^2-test dobré shody

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i} \sim \chi^2_{k-1}$$

O_i	pozorované četnost
R_i	očekávané četnosti

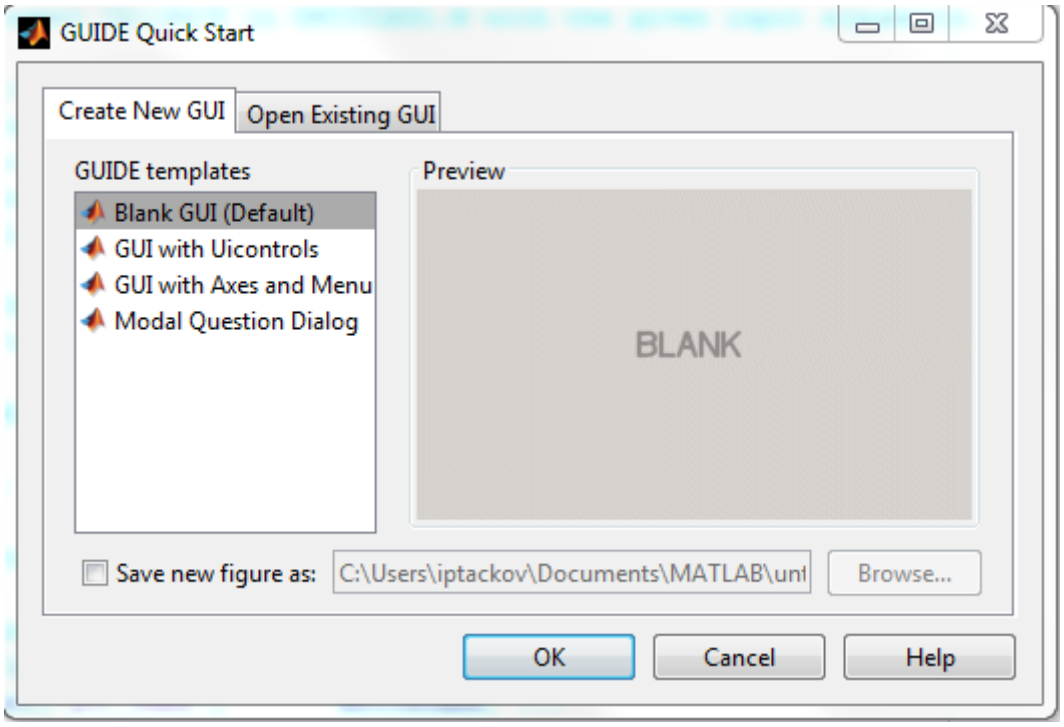
Simultánní porovnání

Počet porovnávání, které musíme provézt: $\frac{m(m-1)}{2}$

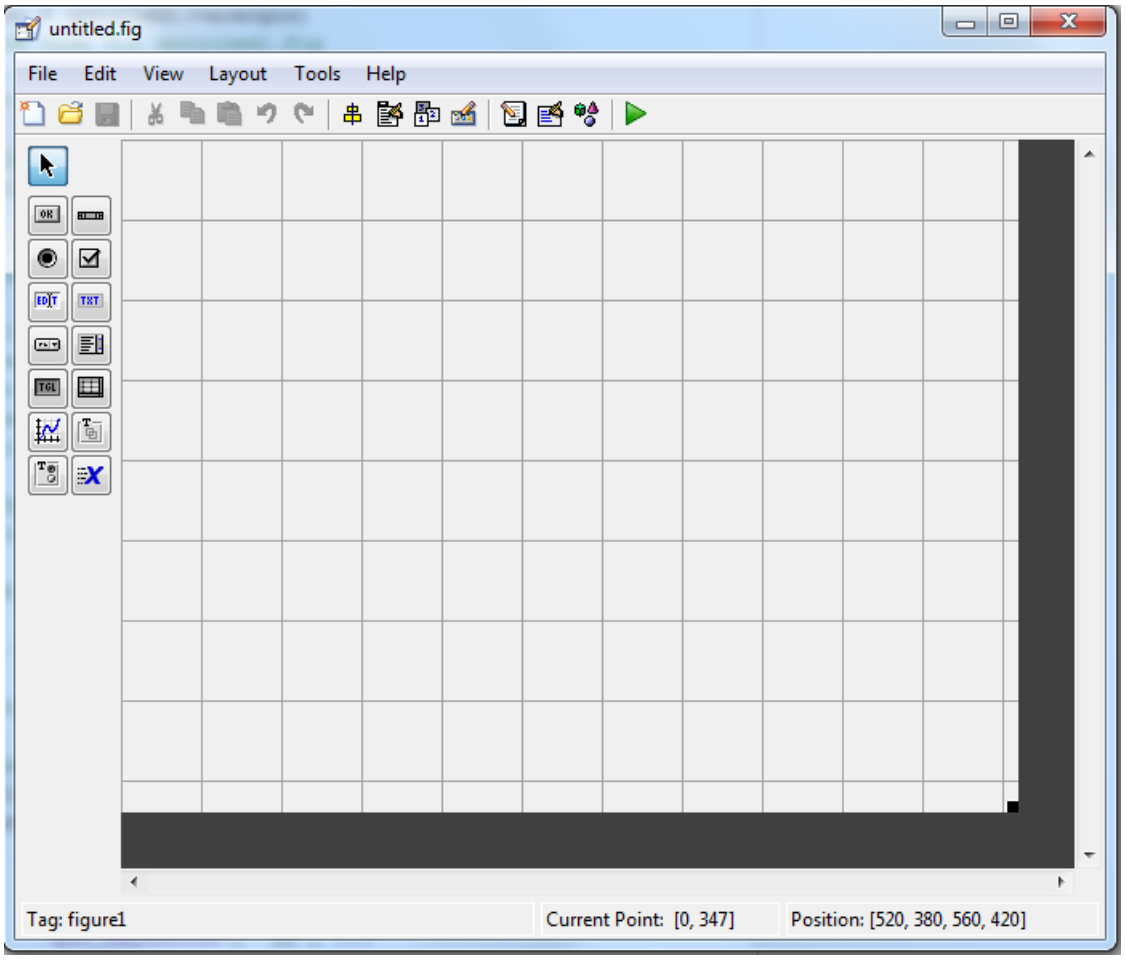
$$|\overline{R_i} - \overline{R_j}| \geq z_{\alpha/m(m-1)} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

$\overline{R_i}$	průměrná pořadí ve skupině (i,j)
$\overline{R_j}$	
m	počet skupin
n	suma počtu dat
n_i	počet dat v rámci skupiny i

D GUIDE



Obrázek D.1: Úvodni okno



Obrázek D.2: Pracovní prostředí

E Import dat

Ukázka vygenerovaného kódu z Matlabu, sloužícího k importu dat:

```
function data = importfile(workbookFile, sheetName, range)
%IMPORTFILE Import data from a spreadsheet
%   DATA = IMPORTFILE(FILE) reads data from the first worksheet in the
%   Microsoft Excel spreadsheet file named FILE and returns the data as a
%   cell array.
%
%   DATA = IMPORTFILE(FILE,SHEET) reads from the specified worksheet.
%
%   DATA = IMPORTFILE(FILE,SHEET,RANGE) reads from the specified worksheet
%   and from the specified RANGE. Specify RANGE using the syntax
%   'C1:C2', where C1 and C2 are opposing corners of the region.%
% Example:
%   sitsinfo = importfile('sits-info.xls','List1','A1:U716');
%
%   See also XLSREAD.

% Auto-generated by MATLAB on 2015/04/30 17:10:30

%% Input handling

% If no sheet is specified, read first sheet
if nargin == 1 || isempty(sheetName)
    sheetName = 1;
```

```
24 end
25
26 % If no range is specified, read all data
27 if nargin <= 2 || isempty(range)
28     range = '';
29 end
30
31 %% Import the data
32 [~, ~, data] = xlsread(workbookFile, sheetName, range);
33 data(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),
       data)) = {''};
```

F Uživatelská dokumentace

Import data

Program pracuje s údaji, které mají minimální nárok na strukturu svých dat. To znamená, že údaje obsahují popisek těchto dat (pohlaví,věk,..). V případě, že se jedná o data bez označení, program nijak nereaguje a bere první řádek jako popisek naměřených dat. Ve výsledku nebudou následně do testování nijak zapojována.

Samotný import dat: *Open File*$\longrightarrow$ vyberte soubor (např. *data.xls*)$\longrightarrow$ *Otevřít.*

Výběr metody

Kruskal-Wallisův test

Kruskal-Wallisův test je neparametrickou verzí metody analýzy rozptylu jednoduchého třídění. Tento způsob testování dat je využíván pokud jsou výběry z rozdělení, které je značně odlišné od normálního rozdělení. Je aplikován při testování shody zvoleného pravděpodobnostního rozdělení srovnávaných skupin. Data, s kterými pracuje, nevycházejí z normálního rozdělení, a jsou na sobě nezávislé. Jeden z předpokladů použití této metody je přítomnost dat, které obsahují dva a více naměřených údajů.

V principu jsou data rozdělena do skupin (např. žena, muž). Je zjištěn stupeň volnosti a zvolena kritická hodnota (χ^2-rozdělení). Data skupin jsou seřazena dle velikosti napříč skupinami, a následně je jim přiřazena hodnota pořadí (dále jen rank). V případě shodných naměřených hodnot se přechází k přiřazení průměru z pořadí. Data jsou nadále zpět rozřazena do svých skupin, ale reprezentována svojí rank hodnotou. Skupiny jsou pak sumarizovány a je určena četnost jejich dat. Po dosazení do vzorce je výsledek porovnán s hladinou významnosti. H_0 je pak zamítnuta nebo přijata na základě tohoto porovnání.

- **Použití:** Spojitá data v kombinaci s kategoriálními.

- **Hodnoty:** KW je výsledek daného testu. P-value je pravděpodobnost, že se jedná o stejnou distribuci. Zamítnutí nulové hypotézy závisí na pozorovateli.
- **Graf:** Krabicový diagram.

Volba dat

Výběr rozdělení:

- Data,která se promítnou na ose x u krabicového diagramu.
- Data, která určují porovnávané skupiny.
- V případě numerického bloku dat možnost *grupování*.

Soubor dat:

- Data obsahující hodnoty související s daty v rozdělení, která určují jejich zařazení.

Simultánní porovnávání

Toto porovnávání je zároveň také post hoc analýzou, která se používá v případě zamítnutí H_o u předešlé metody. Analýzu je možné provádět, aniž by tomu předcházela specifikace srovnání dat. Princip metody je postaven na porovnávání mediánů statisticky usuzovaných skupin. Pro výslednou hodnotu je potřeba porovnat navzájem všechny skupiny.

- **Použití:** Další krok v případě zamítnutí nulové hypotézy u KW testu. Je možné nastavení dvou metod. Scheffe a HSD, která je ale jen pro symetricky setříděná data.
- **Hodnoty:** Výsledné hodnoty jsou navzájem porovnávané skupiny: jejich pravděpodobnost, že jsou ze stejné distribuce a také rozdíly jejich průměrů. Vše v rámci výstupní tabulky.
- **Graf:** Krabicový diagram.

Volba dat

Totožná s volbou u Kruskal-Wallisova testu.

χ^2-test dobré shody

Tento test je neparametrickou metodou, která je používána v případě na sobě nezávislých dat. Základ této metody je v ověření shody usuzovaných četností s četnostmi, které byly vypozorovány. Data je možné rozdělit do kategorií nebo na intervaly. Záleží na typu dat, jestli jsou kategoriálního typu či intervalového typu.

V praxi je porovnávána nominální proměnná s dvěma a více hodnotami. Porovnávány jsou pak pozorované hodnoty s očekávanými hodnotami, které je možné vypočítat prostřednictvím nějakého teoretického očekávání (Např. 1:1, kdyby šlo o pohlaví). Pro přesnější výsledky se u této metody doporučuje větší množství dat. V opačném případě mohou být výsledky nepřesné. Test je aplikovatelný na již zmíněné kategoriální údaje. Tj. například pohlaví či typ údaje, který posouvá jedince do jisté kategorie.

- **Použití:** Data nominální, ordinální a diskrétní. Větší počet dat.
- **Hodnoty:** Rozhodnutí o hypotéze, pravděpodobnost, hladina významnosti a použité datové soubory.
- **Graf:** Histogram.

Volba dat

Výběr rozdělení:

- Data, která určují možnosti porovnávaní.
- V případě numerického bloku dat možnost *grupování*.

Soubor dat:

- Data,která se promítnou na ose x v histogramu.

- V případě numerického bloku dat možnost *grupování*.

Další možnosti:

- Volba porovnávané skupiny.

Volba dat

Totožná s volbou χ^2-testu dobré shody.

Randomizační test dobré shody

Je používán, pokud je nominální proměnná se třemi a více hodnotami a pro χ^2 test dobré shody je vzorek dat příliš malý. Test je prováděn v případě, že z jednoho testu dobré shody není možné pro malého množství očekávaných četností dojít správného výsledku. Aproximační vztah tak malého vzorku dat není přesný. Základem randomizační verze tohoto testu je pak opakované měření při ještě menším vzorku dat, kdy počítáme vždy jen s náhodně vybraným vzorkem dat z celého vzorku. Přitom je vždy dodržen poměr naměřených dat.

- **Použití:** Data nominální, ordinální a diskrétní. V případě malého vzorku dat.
- **Hodnoty:** Rozhodnutí o hypotéze, hladina významnosti (zde vždy 0,05) a použité datové soubory.
- **Graf:** Sloupcový graf (histogram).

Analýza dat

Analýza dat se týká základních statistických hodnot a grafů.

- **Procenta:** *Použití* pouze pro data, která v sobě mají kategorie nebo se aspoň opakují.
- **Aritmetický průměr a Medián:** *Použití* pouze u kvantitativních dat.

- **Medián a Modus:** *Použití* pro všechny typy dat. U mediánu, v případě nekvantitativních dat, může docházet k nerozhodnému výsledku, který zapříčiní vypsaní položky, která je v seznamu dále. Tento fakt může zapříčinit nepatrné zkreslení výsledku.

Graficky je podpořena kruhovým grafem a histogramem, mezi kterými je možnost přepnutí.

Další funkce

Grupování

Data numerického charakteru lze dle potřeby grupovat do skupin. Vstupní hodnoty pro provedení modifikace jsou zadána uživatelem:

- Dolní hranice: přibližná počáteční hodnota intervalu. Např. 0.
- Velikost skupiny: hodnota na základě, které dochází k rozdělení do skupin. Např. u intervalu od 0 do 100 při volbě velikosti skupiny 10 dojde k seskupení hodnot do těchto skupin: 10s, 20s, 30s, ...
- Horní hranice: přibližná konečná hodnota intervalu, v kterém se data nacházejí. Např. 100.

V případě, že není nutné datový blok grupovat stačí potvrdit dialog bez uvedených hodnot (vstupní pole musí být tedy prázdná).

Hladina významnosti

Při nezvolení se automaticky bere hodnota 0.05.