VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Statistická analýza pacientů s Crohnovou nemocí

Statistical analysis of patients with Crohn desease

2014 Štěpán Bernady

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání bakalářské práce

Student: **Štěpán Bernady**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 1103R031 Výpočetní matematika

Téma: Statistická analýza pacientů s Crohnovou nemocí
Statistical analysis of patients with Crohn desease

Zásady pro vypracování:

Etiologie Crohnovy choroby je stále nejasná a průběh onemocnění značně variabilní. U určité části pacientů je pak průběh onemocnění značně komplikovaný s nutností agresivní léčby a opakovaných chirurgických intervencí. Cílem práce je vyhodnotit možné rizikové faktory komplikovaného průběhu, zejména v souvislosti s dostupnými genetickými faktory a dalšími charakteristikami rozsahu a průběhu Crohnovy nemoci.
1. Studium základů statistického zpracování dat. Zdrojem dat bude FN Ostrava, dle instrukcí vedoucího práce.
2. Návrh uživatelského rozhraní pro sledování pacientů s Crohnovou nemocí.
3. Algoritmické zpracování metodiky pro zjištění vlivu genetických faktorů popř. dalších charakteristik na průběh léčby pacientů s Crohnovou nemocí.
4. Vyhodnocení testů pro sledovaný vzorek pacientů, dle instrukcí vedoucího.

Seznam doporučené odborné literatury:

- R.Briš, M.Litschmannová, Statistika I, elektronické skriptum VŠB TUO, FEI, 2004.
- Briš R., Litschmannová M., STATISTIKA II., E-learningový prvek pro podporu výuky odborných a technických předmětů,v rámci projektu CZ.O4.01.3/3.2.15.2/0326, VŠB TU Ostrava, 2007, ISBN 978-80-248-1482-7.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **prof. Ing. Radim Briš, CSc.**

Datum zadání: 01.09.2013

Datum odevzdání: 07.05.2014

VYSOKÁ ŠKOLA BÁŇSKÁ - TECHNICKÁ UNIVERZITA * Fakulta elektrotechniky a informatiky * OSTRAVA *

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 7. května 2014

Štěpán Bernady

Rád bych na tomto místě poděkoval prof. Ing. Radimu Brišovi, CSc., který mě při mé práci vedl, MUDr. Lubomíru Martínkovi, Ph.D., prof. MUDr. Petru Dítěti, DrSc. a dalším lékařům, kteří se na výzkumu Crohnovy nemoci podílejí, za jejich ochotu a vstřícnost.

Abstrakt

Tato práce zpracovává medicínská data pacientů s Crohnovou nemocí poskytnutá Fakultní nemocnicí v Ostravě (FNO). Jejím hlavním cílem je vytvoření vhodného uživatelského rozhraní, s jehož pomocí je možno provádět podrobné explorační analýzy dat pacientů i analyzovat závislosti mezi vybranými veličinami a genetickými faktory.

Klíčová slova: statistika, explorační analýza, testování hypotéz, statistické testy, analýza závislosti, Python

Abstract

This work processes medical data of patients with Crohn's desease which are provided by University Hospital of Ostrava (UHO). The main objective of this work is to create a suitable user interface through which is going to be possible to implement detailed exploratory analysis of patient's data and to analyse dependencies between selected variables and genetic factors.

Keywords: statistics, exploratory analysis, hypothesis testing, statistical tests, dependency analysis, Python

Seznam použitých zkratek a symbolů

CRP	–	C-reaktivní protein
CT	–	Computed Tomography
FNO	–	Fakultní nemocnice Ostrava
IBD	–	Inflammatory Bowel Disease
MAD	–	Median absolute deviation from the median
PDF	–	Portable document format

Obsah

Seznam tabulek

Seznam obrázků

Seznam výpisů zdrojového kódu

1 Úvod

V této bakalářské práci se zaměřujeme na statistické vyhodnocení lékařských dat pacientů s Crohnovou nemocí poskytnutých FNO. Ve FNO probíhá výzkum Crohnovy nemoci, lékaři se pokoušejí vyhodnotit možné rizikové faktory komplikovaného průběhu, zejména v souvislosti s dostupnými genetickými faktory a tato práce jim má pomoci tyto faktory určit.

Nejdříve si přiblížíme Crohnovu nemoc a popíšeme si její vlastnosti. Poté se budeme zabývat základními statistickými pojmy, které budeme později potřebovat. V následující kapitole budeme zkoumat statistické testy a analýzy závislostí mezi vybraným typem veličin.

Nakonec si přiblížíme program v jazyce Python, který je výstupem práce. Program zpracovává takové veličiny ze souboru dat pacientů, které si lékaři sami určili. Má dvě části, a to explorační analýzu s možností grafické vizualizace a analýzu závislosti. Není známo, že by podobný program již byl vytvořen a i kdyby ano, bylo nutno vytvořit nový, jelikož byly požadavky lékařů velmi specifické.

2 Crohnova nemoc

Crohnova nemoc (též Crohnova choroba nebo regionální enteritida) je chronické zánětlivé onemocnění, které se může projevit v jakékoli části trávicího ústrojí (jícen, žaludek, tenké a tlusté střevo), nejčastěji však v oblasti spojení tenkého a tlustého střeva. Zánět proniká celou stěnou, nezřídka je granulomatózní povahy.

Přestože příčina Crohnovy nemoci není známa, všeobecně se usuzuje, že se jedná o nemoc autoimunního charakteru. Sklon k onemocnění je ovlivněn geneticky, nemoc mohou vyvolat u náchylné osoby vlivy okolního prostředí. Crohnova nemoc patří do skupiny zánětlivých onemocnění označovaných IBD (anglicky Inflammatory Bowel Disease). Crohnova nemoc může být těžko rozeznatelná od jiných forem IBD, jako např. ulcerózní kolitidy (colitis ulcerosa).

Zánětlivá onemocnění střev popsal Giovanni Battista Morgagni (1682–1771), později polský lékař Antoni Leśniowski v roce 1904 a skotský lékař T. Kennedy Dalziel roku 1913. Nemoc byla pojmenována podle Burrilla Bernarda Crohna, amerického gastroenterologa. Spolu s Ginzbergem a Oppenheimerem ji popsal v roce 1932 u série čtrnácti pacientů v oblasti napojení tenkého střeva k tlustému (cékum).

2.1 Lokalizace

Důležitou specifikací Crohnovy nemoci je u pacienta lokalizace v rámci trávicího ústrojí. K tomu se používá Montrealská klasifikace – L (L1 – L4).

- **Ileitida (L1):** Crohnova nemoc obvykle napadá ileum, většinou konečnou část tenkého střeva před jeho vyústěním do tlustého (terminální ileitida). Ve 30 % případů je postižena pouze tato oblast.
- **Kolitida (L2):** V případě postižení tlustého střeva je velice komplikované odlišit Crohnovu nemoc od ulcerozní kolitidy. Ve 20 % případů je postižena pouze tato oblast tlustého střeva.
- **Ileokolitická forma (L3):** Je postiženo tenké i tlusté střevo. Tato forma se objevuje až v 50 % případů.
- **Perianální (L4):** Postižena je oblast konečníku a řitního otvoru, spojena s tvorbou píštělí či abscesů.
- **Ostatní:** Může být postižena jakákoli jiná oblast trávicího traktu (žaludek, dvanáctník, horní části tenkého střeva).

2.2 Klinické projevy

Řada pacientů s Crohnovou nemocí má její příznaky řadu let před stanovením diagnózy. Ve srovnání s ulcerózní kolitidou jsou počáteční symptomy této nemoci méně zřetelné. Vyskytují se tři různé formy této nemoci:

- **stenozující** – důsledkem zánětu je zesílení stěny a zúžení průsvitu střeva, z toho vyplývající omezení průchodnosti střev či jejich úplná neprůchodnost
- **fistulující** – zánět proniká celou tloušťkou stěny, dochází ke slepení střeva s okolími orgány a vzniku fistulí (píštělí), tedy patologických propojení mezi orgány (mezi střevními kličkami navzájem či střeva s okolními orgány jako močový měchýř, pochva) či vyústění na povrch těla
- **zánět omezený na stěnu bez vzniku zúžení či píštělí**

Mezi **projevy v trávicí soustavě** patří např. bolesti v oblasti břišní dutiny někdy spojené i s nevolností a zvracením, průjmy různé povahy, perianální projevy (svědění či bolest v oblasti konečníku), někdy krev ve stolici či postižení žaludku (bolest při polykání, bolest břicha či zvracení). Nemoc se také může projevovat tvorbou aft v dutině ústní nebo nadýmáním.

Crohnova nemoc se projevuje také na **celkovém stavu pacienta**, a to např. zvýšenou teplotou, která je většinou nižší než 38,5 °C, pokud ale dojde ke komplikacím, může být i vyšší, ztrátou hmotnosti v důsledku snížení příjmu potravy a s ní spojenou poruchou vstřebávání živin (malabsorpcí), nebo poruchami růstu u pacientů v období puberty.

Nemoc se projevuje i **mimo trávicí soustavu** např. postižením očí (iritida, iridocyklitida – zánětlivé onemocnění duhovky nebo celé uvey, episcleritida – vážné zánětlivé onemocnění očního bělma), postižením pohybového aparátu (např. artritida – postižení kloubů nebo páteře, osteoporóza v důsledku nedostatku vápníku, vitamínu D či celkově nedostatečné výživě), kožními projevy (Erythema nodosum – rudé uzlinky, Pyoderma gangrenosum – bolestivé hnisavé kožní projevy) či hematologickými projevy (trombóza, plicní embólie, autoimunní hemolytická anémie – stav, kdy imunitní systém napadá červené krvinky a dochází k jejich rozpadu).

2.3 Diagnostika

Diagnóza je obtížná a opírá se o celou řadu vyšetření. Při biochemickém vyšetření krevní obraz stanoví možnou chudokrevnost, která může být způsobena buďto ztrátou krve nebo nedostatkem vitamínu B12 typickým zejména při postižení ilea. Právě v ileu je vitamín B12 vstřebáván a proto hrozí jeho nedostatek při postižení ilea. K monitorování aktivity zánětu se používá především stanovení CRP, leukocytů a sedimentace. K vyšetření tenkého střeva se nejčastěji používá počítačová tomografie (CT enterografie). Vznikají

tak postupné obrazy tenkého střeva, které umožňují identifikaci míst s pravděpodobným zánětlivým onemocněním i zjištění dalších nitrobřišních komplikací způsobených Crohnovou nemocí, zvláště abscesů. Je také možná kolonoskopie či kapslová endoskopie.

Nejběžnější nemocí, která vykazuje stejné příznaky jako Crohnova nemoc, je colitis ulcerosa, neboť se v obou případech jedná o zánětlivé onemocnění střev, které mohou způsobovat stejné symptomy v oblasti tlustého střeva. Stanovit, o jakou z těchto nemocí se jedná, je důležité vzhledem k odlišnému způsobu léčby každé z nich. Tento proces se nazývá **diferenciální diagnostika**. V některých případech ovšem může dojít k tomu, že přesnou diagnózu střevního zánětu nelze určit. V takových situacích je nemoc klasifikována jako neurčitá kolitida (blíže neurčitelné střevní zánětlivé onemocnění).

2.4 Epidemiologie

Incidence Crohnovy nemoci je cca 4–9 případů na 100 000 obyvatel. Celkový počet onemocnění za období posledních dvaceti let stoupá. Obě pohlaví jsou postižena stejnou mírou. Zároveň byl prokázán vyšší výskyt v rámci příbuznosti v rodinách, či etnických skupinách. Jako příklad lze jmenovat vyšší výskyt nemoci u aškenázských židů. Lidé se světlou kůží mají riziko onemocnění dvojnásobně vyšší než lidé s kůží tmavou. V souvislosti s věkem se tato nemoc vyskytuje nejvýrazněji ve dvou kategoriích, a to u teenagerů a mladých po 20. roku života (zejména v kategorii 16 – 35) a potom opět zesiluje u 50 – 70letých.

Různé zdroje uvádějí různou míru výskytu, např. na základě populační studie, provedené v Norsku a v USA byla pozorována incidence Crohnovy nemoci v 6 až 7,1 případu na 100 000 obyvatel. Bylo zjištěno, že nemoc je rozšířenější v severních zemích a převažuje dokonce v severních oblastech v rámci stejné země. Výskyt Crohnovy nemoci je podle této studie 6 na 100 000 v Severní Americe, kde nemocí celkem trpí 400 000 až 600 000 lidí a podobná incidence se předpokládá v Evropě, nižší potom v Asii a Africe.

2.5 Patogeneze

I když příčina Crohnovy nemoci není přesně známa, patogeneze této nemoci zahrnuje genetické a environmentální faktory.

Abnormální projevy v **imunitním systému** často provázejí vznik Crohnovy nemoci. V souvislosti s touto nemocí existují hypotézy o cytokinární odezvě při zánětu. Také proto, že střevní prostředí obsahuje velký počet baktérií, může řada z nich, včetně Mycobacterium avium subspecies paratuberculosis vyvolávat infekce a být tak rizikovým faktorem, či příčinou vzniku Crohnovy nemoci.

Řada vlivů spojených s životním prostředím a životním stylem jako např. strava složená z velkého množství tučných, či předupravených jídel, kouření či hormonální antikoncepce mohou také zvyšovat riziko onemocnění Crohnovou nemocí.

2.6 Léčba

Terapeutický přístup při léčení je sekvenční: v první fázi je třeba léčit akutní onemocnění. Protože Crohnova nemoc je nemocí, kterou se často nedaří vyléčit absolutně, je ve fázi druhé třeba rozsah onemocnění udržovat na minimální úrovni (v remisi). Poté, co se v průběhu léčby dosáhne stádia remise, je cílem udržování tohoto stavu a zamezení nového propuknutí nemoci.

Při léčbě se nejprve nasazují protizánětlivé léky s cílem redukce zánětu. Běžně se používají aminosalicyláty (např. mesalazin), kortikosteroidy (jejich dlouhodobému užívání je třeba se vyvarovat), imunosupresiva (léky na potlačení funkce imunitního systému, např. azathioprin, methotrexát) a při biologické terapii infliximab či adalimumab. Antibiotika se podávají jen v případě infekčních komplikací. V případě srůstů nebo zánětlivých změn znemožňujících průchodnost střev, tvorby abscesů, či pokud organismus v přiměřené době nereaguje na léky, může dojít k nutnosti invazivní terapeutické intervence jako drenáž abscesu pod CT kontrolou nebo k nutnosti operačního výkonu v celkové anestezii.

Na základě objevu **helmintické imunomodulace** byl navržen nový alternativní způsob léčby pomocí kontrolované infekce pacientů tenkohlavcem prasečím (Trichuris suis). Úspěšná terapie těmito červy byla zaznamenána již v několika klinických studiích. Tento neobvyklý způsob se jeví jako vhodná alternativa nejen v léčbě Crohnovy nemoci, ale i dalších autoimunitních nemocí. Podobně jako použití nanotechnologií pro medikamentózní léčbu Crohnovy nemoci nebo nasazení některých speciálních postupů nepatří však zatím tato terapie ke standardním postupům.

Dosud neexistují důkazy o tom, že by způsob stravování ovlivňoval vznik nebo průběh Crohnovy nemoci. Mnozí pacienti ovšem pozorují, že požívání určitých druhů jídel zhoršuje jejich příznaky a naopak jejich nepožívání stav pacientů zlepšuje. V souvislosti s léčbou Crohnovy nemoci byla sestavena řada diet, které zlepšují příznaky nemoci, nicméně u žádné z nich nebyla prokázána schopnost nemoc efektivně vyléčit. Na stavu pacienta se může negativně projevit i stres. Snížení hladiny stresu může být pozorováno v bezprostředním zlepšení příznaků nemoci a celkového stavu pacienta.

2.7 Komplikace

Při onemocnění Crohnovou nemocí může dojít ke komplikacím jako neprůchodnost střeva, píštěle (fistule; patologická propojení střeva s jinou jeho částí nebo jinými orgány), abscesy, rakovina, hubnutí a nízký příjem živin (malnutrice) a dalším (ledvinové kameny, osteoporóza).

Významnou skutečností však zůstává, že v současnosti není možné v léčbě u konkrétního pacienta predikovat průběh onemocnění a zejména výskyt komplikací. Toto je důležité zejména z pohledu nežádoucích účinků a potenciální toxicity podávané léčby, která je nasazována často spíše na základě zkušenosti lékaře než podle medicíny založené na důkazech. Logickým důsledkem jsou pak na jedné straně pacienti s nedostatečně intenzivní léčbou a komplikovaným průběhem a na straně druhé nemocní léčeni nepřiměřeně agresivně se všemi nežádoucími a bohužel i zbytečnými důsledky.

[1]

3 Základní statistické pojmy

Dříve než se pustíme k samotnému zpracování dat, uveďme si některé pojmy, které jsou pro jeho pochopení klíčové.

3.1 Populace a výběry

Důvodem, proč byla vytvořena statistika jako věda, je zjišťování údajů o populaci na základě výběrového souboru. Pojem **populace** znamená souhrn všech existujících prvků, které při statistickém výzkumu sledujeme. Jelikož je rozsah populace obvykle vysoký, provádí se většinou tzv. **výběrová šetření**, ve kterých nezkoumáme celou populaci, ale její část (výběr, výběrový soubor). Cílem je určit takový výběr, aby jeho parametry byly dostatečně reprezentativní vzhledem k populaci. Existuje několik způsobů, jak výběr provést. Abychom se vyvarovali upřednostnění nebo opomenutí některých prvků populace, zvolíme tzv. **náhodný výběr**, ve kterém má každý prvek populace stejnou šanci na zařazení do výběru.

Je zřejmé, že výběrové šetření nikdy nemůže být tak přesné jako vyšetření celé populace. Existují ale rozumné důvody, které vysvětlují výhody výběrového šetření, např. úspora času, energie a financí, nedostupnost celé populace. Jedním ze základních šetření se nazývá **explorační (popisná) analýza**, ta bývá zpravidla prvním krokem k informacím o proměnných a jejich variantách ve výběrovém souboru. Popíše proměnné několika přehlednými hodnotami, které nám dají jakýsi souhrn informací, ze kterých si můžeme udělat obraz, jak asi daná proměnná vypadá a co od ní můžeme čekat. Dříve, než se k ní dostaneme, potřebujeme si dané proměnné více přiblížit.

3.2 Základní typy proměnných

Způsob zpracování proměnných závisí na jejich typu, uveďme si proto jejich základní dělení.

- **Proměnná kvalitativní (kategoriální)** je proměnná vyjádřená slovně, kterou nemůžeme měřit, můžeme ji pouze zařadit do tříd. Podle vztahu mezi jednotlivými kategoriemi se dělí na proměnné **nominální** a **ordinální**. Podle počtu variant, jichž proměnné mohou nabývat, ji dělíme na proměnné **alternativní** a **množné**.
- **Proměnná kvantitativní (numerická)** je proměnná vyjádřená číselně a dá se měřit. Dále ji dělíme na proměnnou **diskrétní**, nabývající konečného nebo spočetného množství variant a na proměnnou **spojitou** nabývající libovolných hodnot z $\mathbb{R}$ nebo nějaké její podmnožiny.

3.3 Explorační analýza kvalitativních proměnných

Kvalitativní proměnné dělíme na nominální a ordinální. Postupně jejich základní statistické charakteristiky prozkoumejme.

3.3.1 Nominální proměnná

Nominální proměnná nabývá různých avšak rovnocenných hodnot, které nelze seřadit a jejichž počet nebývá velký.

- **Četnost n_i** (též absolutní četnost, anglicky „frequency“) je definována jako počet výskytů dané varianty kvalitativní proměnné. Označme n rozsah hodnot a k počet variant, pak platí
$$n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i = n.$$
- **Relativní četnost p_i** (anglicky „relative frequency“) je definována jako podíl četnosti dané varianty ku celkovému počtu dat.
$$p_i = \frac{n_i}{n}, \quad \text{popř. } p_i = \frac{n_i}{n} \cdot 100[\%]$$
Pro relativní četnosti musí platit:
$$p_1 + p_2 + \cdots + p_k = \sum_{i=1}^{k} p_i = 1, \quad \text{popř. } 100[\%]$$

Při zpracování kvalitativní proměnné je vhodné četnosti i relativní četnosti uspořádat do tzv. **tabulky rozdělení četností** (anglicky „frequency table“) viz Tab. 1.

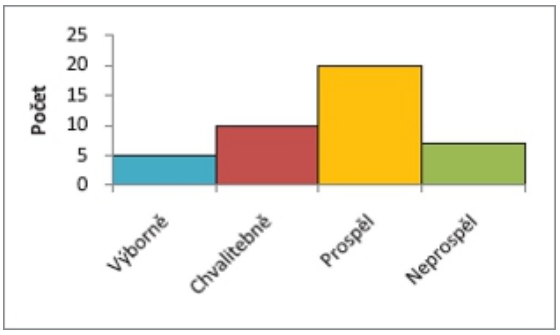
Tabulka rozdělení četností		
Hodnoty x_i	Absolutní četnosti n_i	Relativní četnosti p_i
x_1	n_1	p_1
x_2	n_2	p_2
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	p_k
Celkem	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$

Tab. 1: Tabulka rozdělení četností nominální proměnné

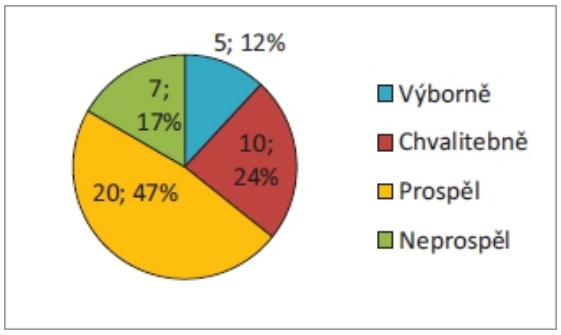
- **Modus $\hat{x}$** je definován jako název varianty proměnné vykazující nejvyšší četnost (typický reprezentant souboru). Vyskytuje-li se v souboru více variant s maximální četností, modus neurčujeme.

3.3.2 Grafické znázornění kvalitativních proměnných

Pro větší názornost analýzy proměnných se ve statistice často užívají **grafy**, které nám dávají vizuální přehled o datech. Pro nominální proměnnou užíváme **histogram** (také sloupcový graf, anglicky „bar chart“), ve kterém jsou četnosti jednotlivých variant zobrazeny jako výška sloupců, nebo **výsečový graf** (také koláčový graf, anglicky „pie chart“). Ve výsečovém grafu prezentujeme relativní četnosti variant proměnné a kromě nich uvádíme i absolutní četnosti pro úplnost (viz Obr. 1, 2).



Obr. 1: Histogram



Obr. 2: Výsečový graf

3.3.3 Ordinální proměnná

Ordinální proměnná stejně jako proměnná nominální nabývá v rámci souboru různých slovních variant, které však můžeme seřadit. Pro popis ordinální proměnné se používají stejné statistické charakteristiky a grafy jako pro popis nominální proměnné (četnost, relativní četnost a modus, histogram a výsečový graf) rozšířené o tyto další charakteristiky:

- **Kumulativní četnost m_i** (anglicky „cumulative frequency“) definujeme jako počet hodnot proměnné, které nabývají varianty nižší nebo rovné i-té variantě. Jsou-li jednotlivé varianty uspořádány podle své „velikosti“ („$x_1 < x_2 < \cdots < x_k$“), pak platí

$$m_i = \sum_{j=1}^{i} n_j.$$

- **Kumulativní relativní četnost F_i** (anglicky „cumulative relative frequency“) vyjadřuje, jakou část souboru tvoří hodnoty nabývající i-té a nižší varianty.

$$F_i = \sum_{j=1}^{i} p_j,$$

  což není nic jiného než relativní vyjádření kumulativní četnosti

$$F_i = \frac{m_i}{n}.$$

Pro ordinální proměnné (podobně jako pro nominální proměnné) můžeme prezentovat statistické charakteristiky pomocí tabulky rozdělení četností (viz Tab. 2). Ta navíc obsahuje hodnoty kumulativních a kumulativních relativních četností.

Tabulka rozdělení četností				
Hodnoty x_i	Absolutní četnosti n_i	Relativní četnosti p_i	Kumulativní četnosti m_i	Kumulativní relativní četnosti F_i
x_1	n_1	p_1	$m_1 = n_1$	$F_1 = p_1$
x_2	n_2	p_2	$m_2 = n_1 + n_2 = m_1 + n_2$	$F_2 = p_1 + p_2 = F_1 + p_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	p_k	$m_k = m_{k-1} + n_k = n$	$F_k = F_{k-1} + p_k = 1$
Celkem	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	—	—

Tab. 2: Tabulka rozdělení četností ordinální proměnné

3.3.4 Grafické znázornění ordinální proměnné

Ordinální proměnnou rovněž znázorňujeme pomocí histogramu a výsečového grafu. Ani jeden z těchto grafů však nezaznamenává uspořádání jednotlivých variant. K tomu nám slouží polygon kumulativních (resp. kumulativních relativních) – **Lorenzova křivka** (Obr. 3) nebo **Paretův graf**.

Obr. 3: Lorenzova křivka

3.4 Explorační analýza numerických proměnných

Pro popis numerické proměnné můžeme použít kromě statistických charakteristik pro popis ordinální proměnné také **míry polohy** a **míry variability**.

3.4.1 Míry polohy

Míry polohy určují typické rozložení hodnot proměnné na číselné ose.

- **Aritmetický průměr $\bar{x}$** (anglicky „mean“) je velmi citlivý na odlehlá pozorování. Vypočteme jej podle vztahu

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

kde n je rozsah výběru a x_i jednotlivé hodnoty proměnné. Existují i další typy průměrů viz [2]. Popišme si nyní **vlastnosti** aritmetického průměru.

1. Součet všech odchylek hodnot proměnné od jejich aritmetického průměru je roven nule:

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

2. Přičteme-li ke všem hodnotám proměnné stejné číslo, zvětší se o toto číslo i aritmetický průměr:

$$\forall a \in \mathbb{R} : \frac{\sum_{i=1}^{n} (a + x_i)}{n} = a + \bar{x}.$$

3. Vynásobíme-li všechny hodnoty proměnné stejným číslem, změní se stejným způsobem i aritmetický průměr:

$$\forall b \in \mathbb{R} : \frac{\sum_{i=1}^{n} (b \cdot x_i)}{n} = b \cdot \bar{x}.$$

Průměr je ovšem velmi citlivý na tzv. **odlehlá pozorování**, což jsou hodnoty, které se významně liší od ostatních a dokáží vychýlit průmer natolik, že přestává daný výběr reprezentovat.

- **Modus $\hat{x}$** je oproti průměru méně závislý na odlehlých pozorováních. Je jinak definován u diskrétních a spojitých proměnných. U **diskrétních proměnných** definujeme modus jako variantu proměnné s nejvyšší četností. U **spojitých proměnných** považujeme za modus $\hat{x}$ hodnotu, kolem níž je nejvetší koncentrace hodnot proměnné. K určení této hodnoty využijeme tzv. **shorth**, což je nejkratší interval, v němž leží alespoň 50% hodnot proměnné. Modus pak definujeme jako střed shorthu.

- **Výběrové kvantily** jsou rovněž odolné vůči odlehlým pozorováním. Výběrový kvantil je hodnota, která rozděluje výběrový soubor na dvě části (části, ve kterých jsou hodnoty buď menší, nebo větší či rovny danému výběrovému kvantilu). Pro jeho určení je nutno výběr uspořádat podle velikosti od nejmenších hodnot k největším. Nejznámějšími výběrovými kvantily jsou **kvartily**.

- **Dolní kvartil $x_{0,25}$** – 25% kvantil (25% hodnot je menších než tento kvartil, 75% pak větších nebo rovných)
- **Medián $x_{0,5}$** – 50% kvantil
- **Horní kvartil $x_{0,75}$** – 75% kvantil

Dalšími výběrovými kvantily jsou decily ($x_{0,1}, \ldots, x_{0,9}$) a percentily ($x_{0,01}, \ldots, x_{0,99}$).

3.4.2 Míry variability

Nyní se budeme zabývat statistickými charakteristikami umožňujícími popis variability výběrového souboru, neboli rozptýlenosti jednotlivých hodnot kolem „středu" proměnné. Zařazujeme zde i již dříve zmíněný **shorth**.

- **Variační rozpětí** (anglicky „range") je určeno rozdílem největší a nejmenší hodnoty výběru ($x_{max} - x_{min}$).
- **Interkvartilové rozpětí IQR** je definován jako vzdálenost mezi horním a dolním kvartilem.
$$IQR = x_{0,75} - x_{0,25}$$
- MAD (anglicky „median absolute deviation from the median"), čili medián absolutních odchylek od mediánu určíme takto: nejdříve výběrový soubor uspořádáme podle velikosti a určíme medián souboru. Poté pro každou hodnotu souboru určíme absolutní hodnotu její odchylky od mediánu, ty pak uspořádáme podle velikosti a určíme medián absolutních odchylek od mediánu, tj. MAD.
- **Výběrový rozptyl s^2** (anglicky „sample variance") určujeme podle vztahu
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$
- **Výběrová směrodatná odchylka s** (anglicky „sample standard deviation") je definována jako kladná odmocnina výběrového rozptylu.
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

3.4.3 Identifikace odlehlých pozorování

Jak jsme se již dříve zmínili, za odlehlá pozorování (anglicky „outliers") považujeme ty hodnoty proměnné, které se nějak významně liší od ostatních hodnot. Ukažme si pár způsobů, jak je identifikovat.

1. ***z-souřadnice* (*z-skóre*)** – za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota *z-souřadnice* je větší než 3, tj. hodnota, která je od průměru vzdálenější než $3s$, tedy

$$\textit{z-skóre}_i = \frac{x_i - \bar{x}}{s}$$

$$|\textit{z-skóre}_i| > 3 \Rightarrow \left|\frac{x_i - \bar{x}}{s}\right| > 3 \Rightarrow |x_i - \bar{x}| > 3 \cdot s \Rightarrow x_i \text{ je odlehlým pozorováním.}$$

2. ***$x_{0,5}$-souřadnice* (*$x_{0,5}$-skóre*)** – za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota mediánové souřadnice je větší než 3, tj. hodnota, která je od mediánu vzdálenější než $3 \cdot 1,483 \cdot MAD$, tedy

$$x_{0,5}\textit{-skóre}_i = \frac{x_i - x_{0,5}}{1,483 \cdot MAD}$$

$$|x_{0,5}\textit{-skóre}_i| > 3 \Rightarrow \left|\frac{x_i - x_{0,5}}{1,483 \cdot MAD}\right| > 3 \Rightarrow |x_i - x_{0,5}| > 3 \cdot 1,483 \cdot MAD \Rightarrow$$
$$\Rightarrow x_i \text{ je odlehlým pozorováním.}$$

Vidíme, že *z-souřadnice* se určuje pomocí průměru a výběrové směrodatné odchylky, což jsou charakteristiky závislé na odlehlých pozorováních. Naopak mediánovou souřadnici určujeme pomocí mediánu a MADu, které jsou proti odlehlým pozorováním odolné, proto je často vhodnější použít mediánovou souřadnici.

3.4.4 Grafické znázornění numerické proměnné

Pro znázornění numerické proměnné používáme např. histogram, empirickou distribuční funkci či krabicový graf.

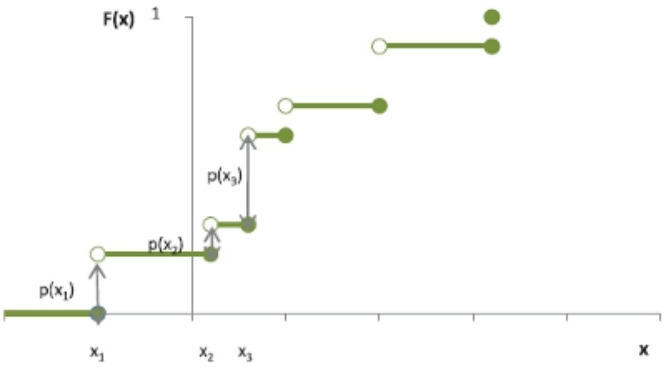
Empirická distribuční funkce $F(x)$, popř. **distribuční funkce kumulativní četnosti** je znázorněním setříděné proměnné (grafické nebo tabulkové) a příslušných kumulativních četností (viz Obr. 4). Označme si $p(x_i)$ relativní četnost hodnoty x_i seřazeného výběrového souboru $x_1 < x_2 < \cdots < x_n$. Pro empirickou distribuční funkci $F(x)$ pak platí:

$$F(x) = \begin{cases} 0 & \text{pro } x \leq x_i \\ \sum_{i=1}^{j} F(x) & \text{pro } x_j < x \leq x_{j+1}, 1 \leq j \leq n-1 \\ 1 & \text{pro } x_n < x \end{cases}$$

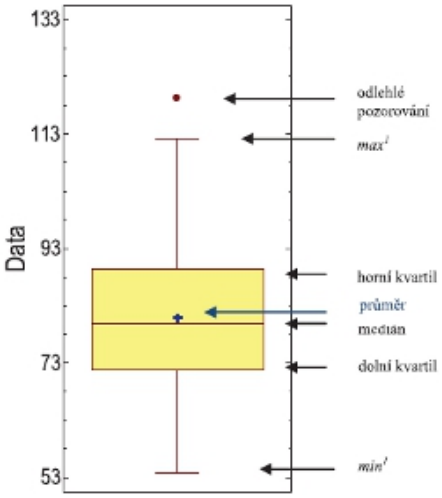
Empirická distribuční funkce je monotónně rostoucí, zleva spojitá funkce, která „skáče“ podle relativních četností příslušných jednotlivým hodnotám proměnné. Platí, že

$$p(x_i) = \lim_{x \to x_i} F(x) - F(x_i).$$

Krabicový graf (anglicky „box plot“) či **krabicový graf s vousy** (anglicky „box with whiskers plot“) viz. Obr. 5 znázorňuje mnoho statistických charakteristik, které jsme si výše popsali – odlehlá pozorování, nejmenší a největší hodnotu, kvartily a průměr. Tzv. krabice obsahuje hodnoty od dolního po horní kvartil a je rozdělená na dvě části mediánem. Tzv. vousy obsahují hodnoty mezi horním resp. dolním kvartilem a největší resp. nejmenší hodnotou. Odlehlá pozorování jsou znázorněna jako jednotlivé body mimo vousy.



Obr. 4: Empirická distribuční funkce



Obr. 5: Krabicový graf s vousy

3.5 Diskrétní náhodná veličina

Diskrétní náhodná veličina nabývá pouze hodnot z nějaké konečné nebo spočetné množiny, nejčastěji se jedná o celočíselné náhodné veličiny.

Definice 3.1 *Náhodná veličina X má diskrétní rozdělení pravděpodobnosti právě tehdy, když:*

1. $\exists$ *konečná nebo spočetná množina reálných čísel* $M = \{x_1, \ldots, x_n, \ldots\}$ *takových, že* $P(X = x_i) > 0 \quad$ *pro* $i = 1, 2, \ldots, n$
2. $\sum_i P(X = x_i) = 1.$

Funkce $P(X = x_i) = P(x_i)$ se nazývá **pravděpodobnostní funkcí** náhodné veličiny X. **Distribuční funkce** tohoto rozdělení je schodovitá a platí pro ni:

$$F(x) = \sum_{x_i < x} P(X = x_i).$$

3.6 Spojitá náhodná veličina

Náhodná veličina má spojité rozdělení, pokud může nabýt jakékoliv hodnoty z určitého intervalu. Těmto hodnotám ale nemůžeme přiřadit pravděpodobnostní funkci, protože ta je nulová. Pro její popis proto používáme distribuční funkci, tzn. stanovujeme pravděpodobnost výskytu náhodné veličiny v libovolném intervalu.

3.6.1 Distribuční funkce

Definice 3.2 *Nechť X je náhodná veličina. Reálnou funkci $F(t)$ definovanou pro všechna reálná $t, t \in \mathbb{R}$ vztahem $F(t) = P\{X \in (-\infty, t)\} = P(X < t)$ nazveme **distribuční funkcí** náhodné veličiny X.*

Poznámka 3.1 Za náhodnou veličinu považujeme proměnnou, jejíž hodnota je jednoznačně určena výsledkem náhodného pokusu.

Distribuční funkce je tedy funkce, která každému reálnému číslu přiřazuje pravděpodobnost, že náhodná veličina nabude hodnoty menší než toto reálné číslo. Má tyto vlastnosti:

- Distribuční funkce je nezáporné číslo menší nebo rovno jedné: $0 \leq F(x) \geq 1$, je neklesající, tj. $\forall x_1, x_2 \in \mathbb{R} : x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$, je zleva spojitá a platí:
- $\lim_{x \to +\infty} F(x) = 1$; $\lim_{x \to -\infty} F(x) = 0$
- $\forall a, b \in \mathbb{R}; a < b : P(a \leq X < b) = F(b) - F(a)$
- $P(X = x_0) = \lim_{x \to x_{0+}} F(x) - F(x_0)$

3.6.2 Hustota pravděpodobnosti

Hustota pravděpodobnosti je definována jako

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x}$$

a platí pro ni

$$\int_{-\infty}^{\infty} f(x) \mathrm{d}x = 1.$$

Jelikož je distribuční funkce spojité náhodné veličiny definována takto:

$$F(x) = \int_{-\infty}^{x} f(t) \mathrm{d}t \quad \text{pro } -\infty < x < \infty,$$

pak ve všech bodech, kde existuje derivace distribuční funkce, platí

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

Známe-li tedy distribuční funkci, můžeme určit hustotu pravděpodobnosti a naopak, známe-li hustotu pravděpodobnosti, můžeme spočítat distribuční funkci.

3.7 Některá spojitá rozdělení

Rozdělení spojité náhodné veličiny je dáno distribuční funkcí, popř. hustotou pravděpodobnosti. Nyní si některá uvedeme.

3.7.1 Normální rozdělení

Normální rozdělení je nedůležitějším pravděpodobnostním rozdělením, které popisuje chování velkého množství náhodných jevů, zvlášť, pokud na kolísání náhodné veličiny působí velký počet nepatrných vzájemně nezávislých vlivů. Lze podle něj aproximovat mnoho jiných spojitých i nespojitých rozdělení. Má dva parametry: $\boldsymbol{\mu}$ – **střední hodnotu** (je rovna mediánu i modu) a $\boldsymbol{\sigma^2}$ – **rozptyl**. Křivka hustoty pravděpodobnosti (Gaussova křivka) má zvonovitý tvar, maximum leží ve střední hodnotě a „šířka“ je úměrná směrodatné odchylce σ. Řídí-li se náhodná veličina normálním rozdělením, zapisujeme: $X \to N(\mu; \sigma^2)$.

Hustota pravděpodobnosti

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}; \quad -\infty < x < \infty$$

Distribuční funkce

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} \mathrm{e}^{-\left(\frac{t-\mu}{\sqrt{2}\sigma}\right)^2} \mathrm{d}t$$

3.7.2 Normované (standardizované) normální rozdělení

Normované normální rozdělení je speciálním typem normálního rozdělení, kde $\boldsymbol{\mu} = 0$ a $\boldsymbol{\sigma^2} = 1$ viz Obr. 6. Řídí-li se náhodná veličina Z tímto rozdělením, pak zapisujeme $Z \to N(0; 1)$. Toto rozdělení je velice důležité, protože distribuční funkci normálního rozdělení nelze vypočítat, naopak hodnoty distribuční funkce normovaného normálního rozdělení můžeme najít v tabulkách. Mezi distribučními funkcemi těchto dvou rozdělení je převodní vztah

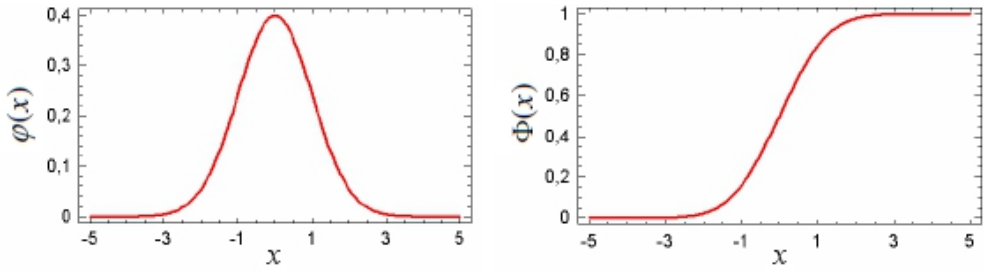
$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Hustota pravděpodobnosti

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\left(\frac{x^2}{2}\right)}; \quad -\infty < x < \infty$$

Distribuční funkce

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} \mathrm{e}^{-\left(\frac{t^2}{2}\right)} \mathrm{d}t$$



Obr. 6: Hustota pravděpodobnosti a distribuční funkce norm. normálního rozdělení

3.7.3 χ^2 rozdělení

Mějme nezávislé náhodné veličiny $Z_1, Z_2, \ldots, Z_n$, z nichž každá má normované normální rozdělení. Pak součet čtverců těchto náhodných veličin má rozdělení χ^2_n (chí-kvadrát) s n stupni volnosti (anglicky „degrees of freedom“).

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2$$

Má-li náhodná veličina X χ^2_n rozdělení, pak zapisujeme $X \to \chi^2_n$. Střední hodnota je rovna počtu stupňů volnosti: $\boldsymbol{E}\left(\boldsymbol{\chi^2_n}\right) = \boldsymbol{n}$ a rozptyl jejich dvojnásobku: $\boldsymbol{D}\left(\boldsymbol{\chi^2_n}\right) = \boldsymbol{2n}$. Hustota pravděpodobnosti pro n stupňů volnosti (obecný tvar) je značně komplikovaná, tudíž se jí nebudeme zabývat.

χ^2 rozdělení používáme např. u testu o rozptylu normálního rozdělení, testu dobré shody nebo χ^2 testu nezávislosti v kontingenční tabulce (o něm si povíme později).

[2], [3]

4 Statistické testy a analýzy

V této kapitole si povíme o statistických testech a analýzách závislostí mezi veličinami, ale nejdřív je třeba zmínit jiné statistické prvky, které s nimi souvisí.

4.1 Intervalový odhad

Potřebujeme-li odhadnout nějaký parametr Θ výběrového souboru, můžeme na základě znalosti výběrového souboru použít intervalový odhad, tzn. najdeme **interval spolehlivosti** $\langle T_D; T_H \rangle$, ve kterém se hledaný parametr vyskytuje s danou pravděpodobností (**spolehlivost odhadu**). Čím vyšší spolehlivost odhadu zvolíme, tím větší bude interval spolehlivosti a naopak. Označme spolehlivost odhadu $\mathbf{1 - \alpha}$, pak $\boldsymbol{\alpha}$ nazveme **hladinou významnosti**. V technické praxi se nejčastěji setkáme se spolehlivostí odhadu 95% nebo 99%, tedy s hladinou významnosti 5% nebo 1%. Pro interval spolehlivosti platí:

$$P(T_D \leq \Theta \leq T_H) = 1 - \alpha.$$

Intervalový odhad parametru Θ se spolehlivostí $1 - \alpha$ je interval $\langle t_D, t_H \rangle$, kde t_D, t_H jsou hodnoty statistik T_D, T_H na daném statistickém souboru $(x_1, \ldots, x_n)$.

Oboustranný interval spolehlivosti

Oboustranný interval spolehlivosti konstruujeme, pokud nás zajímají obě meze odhadu (dolní i horní). Většinou tyto meze určujeme tak, aby platilo, že pravděpodobnost, že parametr populace je menší než dolní mez, byla stejná jako pravděpodobnost, že hledaný parametr je větší než horní mez a byla rovna $\alpha/2$:

$$P(\Theta < T_D) = P(\Theta > T_H) = \frac{\alpha}{2}.$$

Tyto podmínky zaručují již známý vztah

$$P(T_D \leq \Theta \leq T_H) = 1 - \alpha.$$

Dvojice statistik T_D, T_H se pak nazývá **$\mathbf{100(1 - \alpha)\%}$ interval spolehlivosti pro parametr Θ**.

4.2 Shapirův-Wilkův test normality

Pomocí Shapirova-Wilkova testu (viz [4]) můžeme testovat hypotézu, zda náhodný výběr $x_1, \ldots, x_n$ pochází z normálního rozdělení ($X \rightarrow N(\mu; \sigma^2)$) s blíže nespecifikovanými parametry μ a σ^2. Test je určen pro menší výběry, zpravidla pro $n \leq 50$, pro výběry větších rozměrů se používá úprava Roystonovým algoritmem. Nechť $y_1, \ldots, y_n$ jsou seřazené

hodnoty náhodného výběru $x_1, \ldots, x_n$ takové, že $y_1 \leq \cdots \leq y_n$. Testová statistika W pak má tvar

$$W = \frac{b^2}{SS^2} = \frac{\left(\sum_{i=1}^{k} a_i(y_{n-i+1} - y_i)\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

kde a_i jsou tzv. tabelizované váhy, $\bar{y}$ je výběrový průměr a $k = \frac{n}{2}$, je-li n sudé, resp. $k = \frac{n-1}{2}$, je-li n liché. Čím je hodnota testové statistiky W blíže číslu 1, tím je lepší shoda mezi teoretickým a empirickým rozdělením. Pokud hodnota testové statistiky W nepřekročí tabelovanou kritickou hodnotu Shapirova-Wilkova testu (viz Tab. 9), zamítáme nulovou hypotézu, že výběr pochází z normálního rozdělení, na dané hladině významnosti α.

4.3 Testování hypotéz

Často se v praxi setkáváme se situacemi, kdy chceme srovnat nějaké veličiny (např. úmrtnost při různých typech operací, výsledky přijímacích zkoušek na vysokou školu u studentů z různých typů středních škol), chceme zjistit, jestli spolu nějak souvisí. Pak formulujeme tzv. **hypotézy** (např. úmrtnost je při laparoskopických operacích nižší než u operací konvenčních, výsledky přijímacích zkoušek na vysokou školu jsou lepší u studentů z gymnázií), o jejichž správnosti rozhodujeme na základě vhodného výběrového souboru, protože vyšetření celé populace je většinou velice složité či takřka neproveditelné. Tento proces nazýváme **testování hypotéz**.

Statistická hypotéza je výrok či tvrzení o rozdělení pozorované náhodné veličiny, které se zakládá na předchozích zkušenostech, na rozboru dosavadních znalostí nebo na pouhé domněnce. Pojednává-li statistická hypotéza o parametrech rozdělení náhodné veličiny (střední hodnotě, rozptylu, ...), mluvíme o **parametrické hypotéze**, týká-li se jiných vlastností náhodné veličiny (typu rozdělení, nezávislosti výběru, ...), nazýváme ji **neparametrickou hypotézou**.

Při rozhodování o správnosti hypotéz stojí proti sobě dvě tvrzení, a to **nulová** ($\boldsymbol{H_0}$) a **alternativní** ($\boldsymbol{H_A}$) hypotéza, která nulovou hypotézu nějakým způsobem popírá. Nulová hypotéza bývá vyjádřená rovností testovaného parametru $\boldsymbol{\Theta}$ s jeho očekávanou hodnotou $\boldsymbol{\Theta_0}$:

$$H_0: \quad \Theta = \Theta_0$$

Po získání výběrového souboru a formulaci nulové hypotézy zkonstruujeme alternativní hypotézu (zkráceně alternativu). Máme čtyři možnosti pro její formulaci. Výběr vhodné možnosti alternativy závisí na výběrovém souboru a měl by z něj vycházet.

a) $H_A : \Theta = \Theta_1$ (pro případ, kdy se rozhodujeme mezi dvěma hodnotami Θ_0 a Θ_1)

b) $H_A : \Theta \neq \Theta_0$ (popření H_0 bez bližší specifikace)

c) $H_A : \Theta < \Theta_0$ (popření H_0, testovaný parametr je menší než hodnota parametru z H_0)

d) $H_A : \Theta > \Theta_0$ (popření H_0, testovaný parametr je větší než hodnota parametru z H_0)

Při testování hypotéz máme možnost rozhodnout dvěma způsoby, buď zamítneme nulovou hypotézu H_0 ve prospěch alternativy H_A nebo nezamítneme nulovou hypotézu H_0. Obor hodnot testovaného parametru Θ se tedy dělí na dvě disjunktní množiny, které nazýváme **obor přijetí** (testované hypotézy H_0) $\boldsymbol{V}$ a **kritický obor** (obor zamítnutí hypotézy H_0) $\boldsymbol{W}$. Kritický obor W stanovujeme tak, aby pravděpodobnost výskytu pozorované hodnoty testovaného parametru Θ v něm byla velmi malá. Hranice mezi kritickým oborem a oborem přijetí se nazývá **kritická hodnota testu** $\boldsymbol{t_{krit}}$. Je-li tedy pozorovaná hodnota testovaného parametru Θ v kritickém oboru W, zamítáme H_0. Je-li pozorovaná hodnota v oboru přijetí V, hypotézu H_0 nezamítáme.

4.3.1 Testová statistika

Ke korektnímu testu statistické hypotézy máme nástroj nazývaný testovou statistikou (testovým kritériem), kterým je **výběrová charakteristika** $\boldsymbol{T(x)}$, jejíž rozdělení známe a která má vztah k nulové hypotéze.

Kritický obor W lze často popsat prostřednictvím kritického oboru W^* testové statistiky $T(X)$. Je-li pozorovaná hodnota $T(X)$ v kritickém oboru W^*, zamítáme H_0. V opačném případě H_0 nezamítáme.

4.3.2 Chyba I. a II. druhu

Pokud se takto budeme rozhodovat, nastane některý z případů uvedených v Tab. 3.

		Výsledek testu	
		Nezamítáme H_0	Zamítáme H_0
Skutečnost	Platí H_0	Správné rozhodnutí $1-\alpha$ (spolehlivost testu)	**Chyba I. druhu** α (hladina významnosti)
	Platí H_A	**Chyba II. druhu** β	Správné rozhodnutí $1-\beta$ (síla testu)

Tab. 3: Výsledky testování hypotéz

Pravděpodobnosti α a β, s nimiž chyby I. a II. druhu nastávají, rozhodují o kvalitě testu. Snažíme se minimalizovat obě pravděpodobnosti, tzn. zvýšit sílu testu snížením β při co nejmenší hladině významnosti α. Narážíme ale na problém, jelikož snížením jedné z pravděpodobností se zvýší druhá, proto je vhodné najít kompromis v požadavcích na α a β. Jako vstupní parametr testu volíme hladinu významnosti α (nejčastěji $\alpha = 0,05$), chybu II. druhu β můžeme snížit výběrem vhodného testu nebo zvětšením výběrového souboru (takto nezvýšíme α).

4.3.3 Přístupy k testování hypotéz

První možností, jak k testování hypotéz přistupovat, je použít **klasický test**. Při něm se postupuje takto: nejdříve formulujeme H_0 a H_A, zvolíme testovou statistiku $T(X)$ a stanovíme hladinu významnosti testu α. Poté sestrojíme kritický obor W^* testové statistiky $T(X)$ tak, aby pravděpodobnost, že $T(X)$ leží v kritickém oboru W^* za předpokladu platnosti H_0, byla rovna hladině významnosti α, tzn. $P\left(T(X) \in W^* | H_0\right) = \alpha$. Dále vypočteme pozorovanou hodnotu x_{OBS} testové statistiky $T(X)$ a zformulujeme závěr testu. Leží-li pozorovaná hodnota x_{OBS} v kritickém oboru W^*, zamítáme H_0 ve prospěch H_A. V opačném případě nezamítáme H_0.

Druhou možností je použít **čistý test významnosti**. Postup je podobný jako u klasického testu: nejdříve zformulujeme H_0 a H_A, zvolíme testovou statistiku $T(X)$, vypočteme pozorovanou hodnotu x_{OBS} testové statistiky $T(X)$ a *p-hodnotu*. ***P-hodnota*** (anglicky „*p-value*“) je nejnižší hladina významnosti, na níž můžeme zamítnout H_0 a zároveň nejvyšší hladina významnosti, na které se již H_0 nezamítá. *P-hodnotu* vypočteme v závislosti na tvaru alternativní hypotézy jednou z těchto definic:

a) $H_A : \Theta < \Theta_0 \Rightarrow$ *p-hodnota* $= F_0(x_{OBS})$

b) $H_A : \Theta > \Theta_0 \Rightarrow$ *p-hodnota* $= 1 - F_0(x_{OBS})$

c) $H_A : \Theta \neq \Theta_0 \Rightarrow$ *p-hodnota* $= 2min\left\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\right\}$ (pouze je-li nulové rozdělení symetrické)

Následně na základě vypočítané *p-hodnoty* rozhodneme, zda zamítneme nulovou hypotézu či nikoliv (viz. Tab. 4).

Známe-li hladinu významnosti α	
p-hodnota $< \alpha$	zamítáme H_0 ve prospěch H_A
p-hodnota $> \alpha$	nezamítáme H_0
Neznáme-li hladinu významnosti α	
p-hodnota $< 0,01$	zamítáme H_0 ve prospěch H_A
$0,01 <$ *p-hodnota* $< 0,05$	nemůžeme rozhodnout o zamítnutí H_0, doporučuje se rozšířit výběrový soubor a opakovat test
p-hodnota $> 0,05$	nezamítáme H_0

Tab. 4: Rozhodování na základě *p-hodnoty*

4.4 Kruskalův-Wallisův test

Tento test je neparametrickou obdobou jednofaktorové analýzy rozptylu (ANOVA). Užívá se tehdy, když chceme porovnávat střední hodnoty více než dvou nezávislých souborů na základě výběrů nesplňujících předpoklady pro použití ANOVY (zejména normalitu). Kruskalův-Wallisův test je **vícevýběrovým testem shody mediánů**.

Nechť je dáno k nezávislých výběrů $X_{11}, X_{12}, \ldots, X_{1n_1}$ atd. až $X_{k1}, X_{k2}, \ldots, X_{kn_k}$ z rozdělení se spojitou distribuční funkcí o rozsazích $n_1, n_2, \ldots, n_k$. Označme $n = n_1 + n_2 + \cdots + n_k$. Chceme testovat hypotézu

$$H_0 : \quad x_{0,5_1} = x_{0,5_2} = \cdots = x_{0,5_k}$$

vůči alternativě, že H_0 neplatí. Všech n pozorovaných hodnot veličiny X_{ij} seřadíme do rostoucí posloupnosti a určíme jejich pořadí R_{ij}. Tato pořadí uspořádáme do tabulky (viz Tab. 5) a určíme tzv. součty pořadí pro jednotlivé výběry T_i.

Výběr	Pořadí veličin X_{ij} v uspořádané rostoucí posloupnosti				Součty pořadí
1	R_{11}	R_{12}	$\cdots$	R_{1n_1}	T_1
2	R_{21}	R_{22}	$\cdots$	R_{2n_2}	T_2
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
k	R_{k1}	R_{k2}	$\cdots$	R_{kn_k}	T_k

Tab. 5: Pořadí veličin X_{ij} v uspořádané rostoucí posloupnosti a jejich součty

Celkový součet všech pořadí je $T_1 + \cdots + T_k = \frac{n(n+1)}{2}$. Jako testová statistika se používá

$$Q = -3(n+1) + \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i}.$$

Jsou-li rozsahy jednotlivých výběrů alespoň 5 prvků, má testová statistika Q v případě platnosti nulové hypotézy přibližně χ^2 rozdělení s $k-1$ stupni volnosti. Pak *p-hodnota* $= 1 - F_0(Q)$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k-1$ stupni volnosti. Jsou-li rozsahy jednotlivých výběrů menších než 5 prvků, pak je třeba považovat *p-hodnotu* jako nedokonalou aproximaci.

4.4.1 Post hoc analýza pro Kruskalův-Wallisův test

V případě zamítnutí nulové hypotézy nás zajímá, která dvojice výběrů se od sebe statisticky významně liší. Ukážeme si **Dunnové metodu**, která se používá pro mnohonásobné porovnávání. Nechť z_p je p kvantil normovaného normálního rozdělení, průměrné pořadí i-té skupiny $t_i = \frac{T_i}{n_i}$ a modifikovaná hladina významnosti $\alpha^* = \frac{\alpha}{\binom{k}{2}}$. Jestliže

$$|t_I - t_J| \geq \sqrt{\frac{1}{12}\left(\frac{1}{n_I} + \frac{1}{n_J}\right) n(n+1) z_{1-\alpha^*}},$$

pak se mediány I-tého a J-tého výběru statisticky významně liší.

4.5 Analýza závislostí v kontingenčních tabulkách

Pokud chceme zjistit, zda jsou dané dvě kategoriální veličiny na sobě závislé (např. pokud průběh nemoci závisí na typu použitých léků), použijeme analýzu závislosti v kontingenčních tabulkách.

4.5.1 Kontingenční tabulka

Kontingenční tabulka je tabulka, ve které jsou uvedeny výsledky šetření setříděné podle variant dvou kategoriálních znaků (označme X, Y). Nechť znak X nabývá variant $x_{[1]}, \ldots, x_{[r]}$ a znak Y $y_{[1]}, \ldots, y_{[s]}$. V kontingenční tabulce jsou uspořádány absolutní četnosti n_{ij} dvojice variant $\left(x_{[i]}, y_{[j]}\right)$, v hlavičce jsou uvedeny názvy jednotlivých variant znaků X a Y. Kontingenční tabulku často rozšiřujeme o další číselné charakteristiky:

- **celkový rozsah výběru n**,
- **marginální četnosti**, které udávají celkové četnosti jednotlivých variant znaku X, resp. Y. Zapisujeme je na okraj rozšířené kontingenční tabulky (viz Tab. 6).

$\boldsymbol{X \setminus Y}$	$\boldsymbol{y_{[1]}}$	$\boldsymbol{y_{[2]}}$	$\cdots$	$\boldsymbol{y_{[s]}}$	**Celkem**
$\boldsymbol{x_{[1]}}$	n_{11}	n_{12}	$\cdots$	n_{1s}	$\boldsymbol{n_{1\cdot}}$
$\boldsymbol{x_{[2]}}$	n_{21}	n_{22}	$\cdots$	n_{2s}	$\boldsymbol{n_{2\cdot}}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
$\boldsymbol{x_{[r]}}$	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$\boldsymbol{n_{r\cdot}}$
Celkem	$\boldsymbol{n_{\cdot 1}}$	$\boldsymbol{n_{\cdot 2}}$	$\cdots$	$\boldsymbol{n_{\cdot s}}$	$\boldsymbol{n}$

Tab. 6: Schéma rozšířené kontingenční tabulky

- **relativní četnosti**, které pro každé pole rozšířené kontingenční tabulky určíme jako podíl příslušné absolutní četnosti a celkového rozsahu výběru n,
- **řádkové a sloupcové relativní četnosti**, které udávají relativní četnosti znaku Y (resp. X) za předpokladu, že znak X (resp. Y) nabývá určité varianty, tzn. podíl příslušné absolutní četnosti a marginální četnosti v odpovídajícím řádku (resp. sloupci).

Grafickým zpracováním kontingenční tabulky je např. mozaikový graf, shlukový či kumulativní sloupcový graf.

4.5.2 χ^2 test nezávislosti v kontingenční tabulce

Po vyslovení domněnky na základě explorační analýzy, že znak Y závisí na X, můžeme otestovat toto tvrzení rozšířené na celou populaci. Testujeme H_0 vůči alternativě H_A:

H_0 : Znaky X a Y v kontingenční tabulce jsou statisticky **nezávislé**

H_A : Znaky X a Y v kontingenční tabulce jsou statisticky **závislé**

Pro tyto účely slouží χ^2 test nezávislosti v kontingenční tabulce, který je založen na **porovnávání empirických** (pozorovaných) **četností** s **četnostmi teoretickými**, tj. takovými, které bychom očekávali v případě nezávislosti znaků X a Y. Označme empirické četnosti

O_{ij} tak, že $O_{ij} = n_{ij}$. Očekávané četnosti E_{ij} určíme jako četnosti odpovídající součinu příslušných marginálních relativních četností:

$$E_{ij} = \left(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right) \cdot n = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}.$$

Jako testové kritérium používáme náhodnou veličinu K, která má v případě platnosti nulové hypotézy a za předpokladu splnění podmínek dobré aproximace přibližně χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti:

$$K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Podmínky dobré aproximace:

- žádná z očekávaných četností E_{ij} nesmí být menší než 2,
- alespoň 80% očekávaných četností E_{ij} musí být větších než 5.

Jsou-li splněny podmínky dobré aproximace, pak *p-hodnota* $= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

4.5.3 Yatesova korekce χ^2 testu nezávislosti v kontingenční tabulce

Nejsou-li splněny podmínky dobré aproximace nutné pro použití χ^2 testu nezávislosti v kontingenční tabulce, tzn. máme extrémně nízké očekávané četnosti, lze použít tzv. **Yatesovu korekci**. Efektem této korekce je snížení pozorované hodnoty testového kritéria, což znamená, že je obtížnejší zamítnout nulovou hypotézu. Snížíme tak pravděpodobnost chyby I. druhu, chyba II. druhu se však zvýší – test má menší sílu oproti χ^2 testu nezávislosti. Jako testové kritérium používáme náhodnou veličinu K_{Yates}, která má v případě platnosti nulové hypotézy přibližně χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

$$K_{Yates} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij} - 0,5)^2}{E_{ij}}$$

Pak *p-hodnota* $= 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti.

4.5.4 Měření síly závislosti

χ^2 test nezávislosti buď zamítá nebo nezamítá nulovou hypotézu o nezávislosti znaků X a Y, ale nevypovídá nic o síle vztahu. Pro zjištění síly vztahu používáme různé koeficienty. Jako první si uvedeme **koeficient kontingence CC**, který je mírou těsnosti závislosti.

$$CC = \sqrt{\frac{K}{K+n}}$$

Koeficient kontingence se pro čtvercové kontingenční tabulky $(r = s)$ vyskytuje v intervalu $(0, 1)$. Pro obdélníkové kontingenční tabulky $(r \neq s)$ je však maximální hodnota koeficientu kontingence

$$CC_{max} = \sqrt{\frac{\min(r, s) - 1}{\min(r, s)}},$$

proto se pro ně používá **korigovaný koeficient kontingence CC_{cor}** (exaktní korekce do intervalu $(0, 1)$).

$$CC_{cor} = \frac{CC}{CC_{max}}$$

Další mírou těsnosti závislosti je **Cramerův koeficient V** nazývaný též Cramerovo V . Rovněž Cramerův koeficient se vyskytuje v intervalu $(0, 1)$.

$$V = \sqrt{\frac{K}{n(\min(r, s) - 1)}}$$

Čím jsou tyto koeficienty blíže 1, tím je závislost mezi X a Y těsnější.

4.6 Analýza závislostí v asociačních tabulkách

Asociační tabulky jsou speciálním typem kontingenčních tabulek, používáme je ke sledování závislosti dvou dichotomických znaků, tj. kategoriálních znaků nabývajících pouze dvou variant (ano, ne; 0, 1). V medicínských aplikacích obvykle zkoumáme asociaci mezi sledovaným faktorem a výskytem onemocnění, mutací genu apod., proto se na ně více zaměříme. Absolutní četnosti označme takto: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$ (viz Tab. 7).

X (sledovaný faktor) $\setminus$ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Tab. 7: Asociační tabulka rozšířená o marginální četnosti v medicínských aplikacích

4.6.1 Poměr šancí

Jako míru asociace můžeme použít charakteristiku **poměr šancí** (anglicky „odds ratio“). Pozorovaný poměr počtu úspechů k počtu neúspechů za okolností E je $\frac{a}{c}$, za okolností $\bar{E}$ $\frac{b}{d}$, tedy kolikrát je např. vyšší šance výskytu nemoci u populace vystavené vlivu sledovaného faktoru ve srovnání s neexponovanou populací. **Odhad poměru šancí $\widehat{OR}$** (někdy označujeme **křížový poměr**, anglicky „cross-product ratio“) je pak

$$\widehat{OR} = \frac{ad}{bc}.$$

Populační poměr šancí OR nabývá kladných hodnot v intervalu $\langle 0, \infty)$ a při interpretaci poměru šancí rozlišujeme tyto možnosti (důležitá je hodnota 1): pokud $OR < 1$, resp. $OR > 1$, pak je u exponované populace (populace vystavené sledovanému faktoru) nižší, resp. vyšší šance výskytu nemoci, pokud $OR = 1$, pak šance výskytu onemocnění u exponované a neexponované populace jsou shodné. Je-li $OR \neq 1$, pak zpravidla musíme rozhodnout, zda je zjištěná asociace statisticky významná. Otestujeme tedy nulovou hypotézu H_0, že asociace mezi znaky X a Y neexistuje, proti alternativě H_A, že asociace existuje, pomocí **$100(1-\alpha)\%$ intervalu spolehlivosti pro OR**.

$100(1-\alpha)\%$ interval spolehlivosti pro OR

Meze intervalu spolehlivosti pro poměr šancí lze přímo určit pouze obtížně, a proto je aproximujeme. My se zaměříme na **Woolfovu metodu** založenou na aproximaci normálním rozdělením. Podle této metody je $100(1-\alpha)\%$ asymptotický intervalový odhad přirozeného logaritmu poměru šancí

$$\left\langle \ln\widehat{OR} - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}} ; \ln\widehat{OR} + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}} \right\rangle,$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$ kvantil normovaného normálního rozdělení. Na základě znalosti $100(1-\alpha)\%$ intervalového odhadu pro $\ln OR$ určíme **$100(1-\alpha)\%$ intervalový odhad OR**

$$\left\langle \widehat{OR} \cdot \mathrm{e}^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} ; \widehat{OR} \cdot \mathrm{e}^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle.$$

Jestliže $100(1-\alpha)\%$ intervalový odhad OR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y.

4.6.2 Relativní riziko

U asociační tabulky se uvádí jako další popisné statistiky **absolutní rizika** výskytu události (onemocnění, mutace genu, ...) v závislosti na okolnostech (přítomnosti sledovaného faktoru). Jde o vybrané řádkové relativní četnosti, které mohou nabývat hodnot z intervalu $(0, 1)$. Odhad absolutního rizika onemocnění u exponovaných respondentů je $\frac{a}{a+b}$, u neexponovaných respondentů $\frac{c}{c+d}$. Jako míru asociace mezi sledovanými okolnostmi a výskytem události používáme **relativní riziko RR** (anglicky „relative risk“). Odhad relativního rizika RR získáme jako poměr odhadů absolutních rizik:

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)}.$$

Relativní riziko nabývá kladných hodnot v intervalu $\langle 0, \infty)$ a při jeho interpretaci rozlišujeme tyto možnosti (důležitá je opět jako u poměru šancí hodnota 1): pokud $RR < 1$, resp. $RR > 1$, pak expozice snižuje, resp. zvyšuje riziko onemocnění, pokud $RR = 1$, pak mezi expozicí a onemocněním neexistuje žádná asociace. Je-li $OR \neq 1$, pak podobně jako při interpretaci poměru šancí rozhodneme, zda je zjištěná asociace statisticky významná, a to pomocí **$100(1-\alpha)\%$ intervalu spolehlivosti pro RR**.

$100(1-\alpha)\%$ interval spolehlivosti pro RR

Stanovení přesných mezí intervalu spolehlivosti pro relativní riziko je složité a výpočetně náročné. Použijeme proto **Katzovu metodu** založenou na aproximaci normálním rozdělením. Podle ní je $100(1-\alpha)\%$ asymptotický intervalový odhad přirozeného logaritmu relativního rizika

$$\left\langle \ln\widehat{RR} - \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}} ; \ln\widehat{RR} + \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}} \right\rangle,$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$ kvantil normovaného normálního rozdělení, jehož některé hodnoty nalezneme v tabulkách viz Tab. 8. Na základě znalosti $100(1-\alpha)\%$ intervalového odhadu pro $\ln RR$ určíme **$100(1-\alpha)\%$ intervalový odhad RR**

$$\left\langle \widehat{RR} \cdot \mathrm{e}^{-\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}}} ; \widehat{RR} \cdot \mathrm{e}^{\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle.$$

Jestliže $100(1-\alpha)\%$ intervalový odhad RR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y.

$\boldsymbol{\alpha}$	$0,1$	$0,05$	$0,025$	$0,01$	$0,005$	$0,001$	$0,0005$	$0,0001$
$\boldsymbol{z_\alpha}$	$1,2816$	$1,6449$	$1,96$	$2,3263$	$2,5758$	$3,0902$	$3,2905$	$3,7190$

Tab. 8: Vybrané kvantily normovaného normálního rozdělení ($z_{1-\alpha} = -z_\alpha$)

[2], [3]

5 Zpracování dat

V této sekci upustíme od teorie a budeme se věnovat praktické části práce – zpracování dat a programu.

5.1 Data

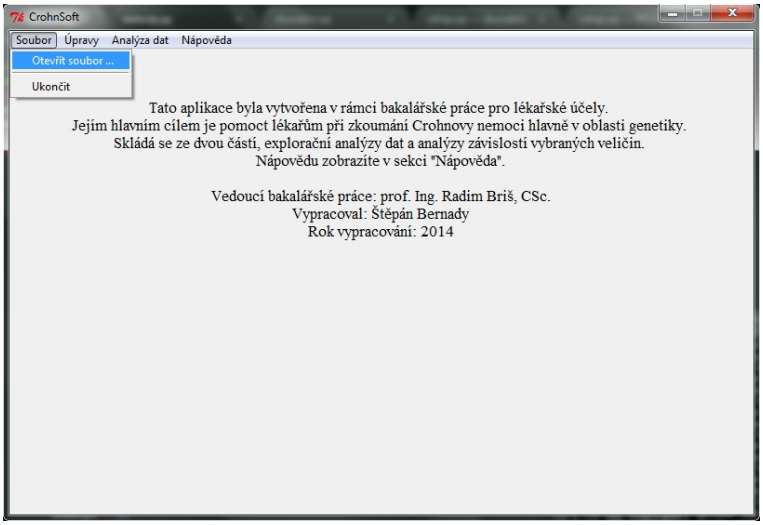
Lékařská data pacientů s Crohnovou nemocí zpřístupněná FNO jsou umístěna na internetu v Google Excel tabulce, která je přístupná jednak lékařům spojeným s výzkumem Crohnovy nemoci, jednak nám, kteří tato data zpracováváme. Jsou zde zastoupeny jak kvalitativní, tak kvantitativní veličiny. Tabulka není zaplněna celá, data se stále shromažďují a doplňují, ale pro naše zpracování stačí, jelikož nepotřebujeme znát údaje všech veličin, ale pouze těch, které lékaři chtějí zkoumat.

5.2 Program

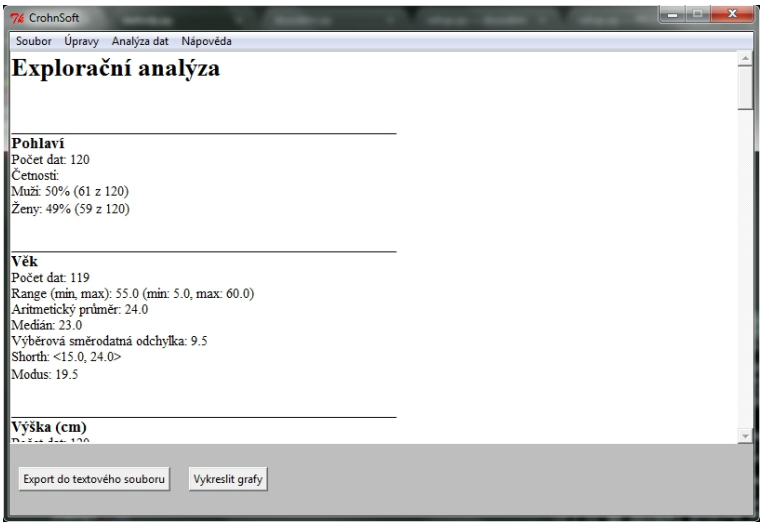
Nedílnou součástí práce bylo vytvoření vhodného uživatelského rozhraní, které bude sloužit ke statistickému zpracování dat. Zvolili jsme programovací jazyk Python verze 2.7.6, který je volně dostupný a dobře uživatelsky přístupný. Použili jsme různé balíčky např. xlrd pro práci s Excel souborem, Tkinter pro tvorbu uživatelského rozhraní, matplotlib pro tvorbu grafů, numpy a scipy pro pomocné matematické nebo statistické funkce, ... Program se skládá ze dvou souborů – uvodni_obrazovka.py a metody.py, z nichž první jmenovaný tvoří kostru programu a kterým program spouštíme, ve druhém jsou zastoupeny funkce pro různé výpočty a zobrazení výsledků.

Po spuštění programu se nám objeví úvodní okno (zobrazitelné také v sekci „O programu“). Jeho hlavní funkce jsou znepřístupněny až do té doby, než načteme Excel soubor s daty (viz Obr. 7). První a poslední řádek tabulky, ve kterých jsou uvedena data pacientů, musí být stejné s údaji v pomocném souboru vstup.txt. Pokud tomu tak není, tak je uživatel vyzván, aby zadal správné údaje. Tyto údaje lze změnit i za běhu programu, pak se data znovu načtou. Excel tabulka musí být pro načtení stáhnutá z internetu na lokálním disku. Načte-li se správně soubor, načtou se z něj i všechna data pacientů od zadaného prvního do posledního řádku a uloží do dvourozměrného pole, ze kterého bude program data nadále čerpat.

Prvním statistickým zpracováním je **explorační analýza dat** dostupná ze sekce „Analýza dat“. Zobrazí se textové pole (viz Obr. 8), v němž jsou rozebrány všechny veličiny, které lékaři chtěli mít zpracované. U všech veličin se zobrazí počet dat, u kvalitativních veličin potom absolutní a relativní četnosti, u numerických minimum, maximum a variační rozpětí (range), aritmetický průměr, medián, výběrová směrodatná odchylka, shorth a modus. Hodnoty buněk tabulky jsou různého charakteru nebo jsou některé buňky prázdné, proto pro každou veličinu načítáme „čistá“ data zvlášť do jiného pole, které lze snadněji zpracovat.

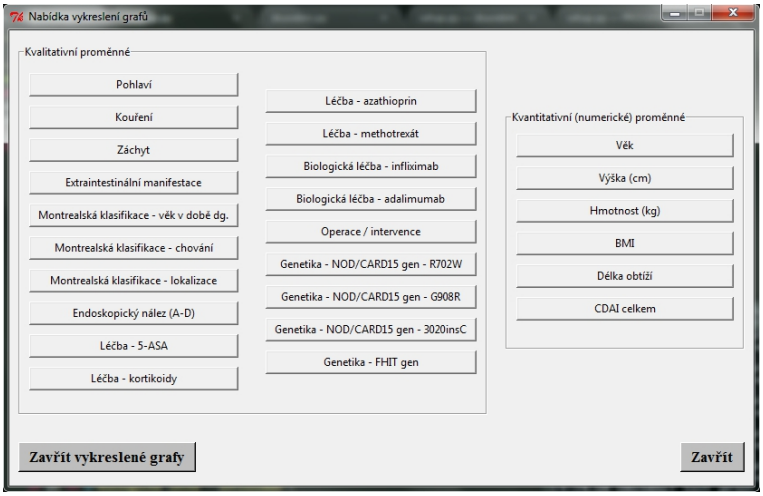


Obr. 7: Program – úvod



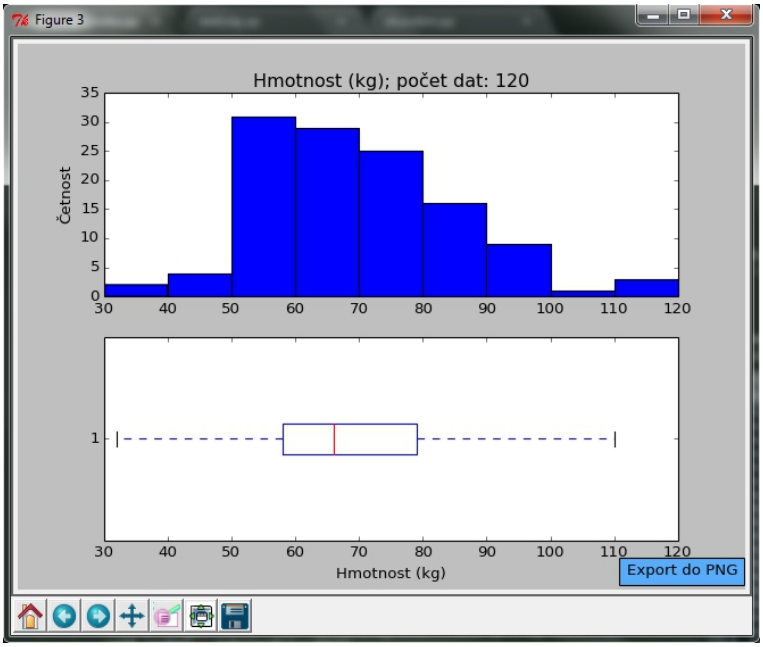
Obr. 8: Program – explorační analýza

Máme zde i možnost uložit výsledek explorační analýzy do textového souboru či vykreslit grafy jednotlivých veličin (viz Obr. 9).



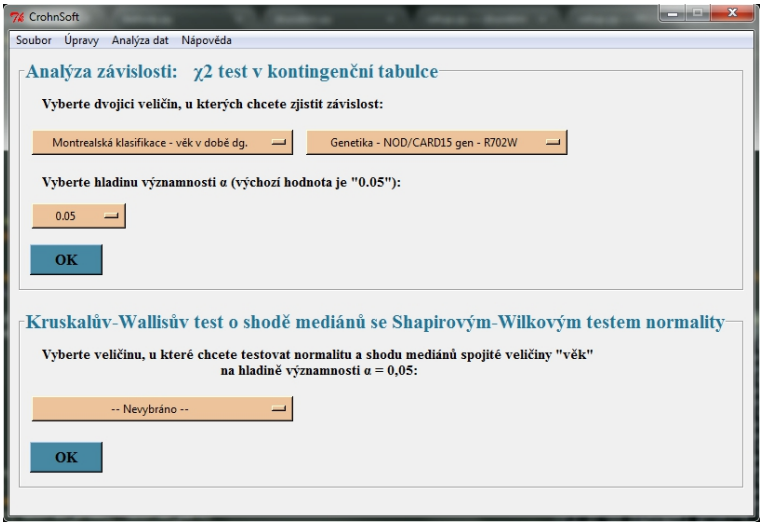
Obr. 9: Program – nabídka zobrazení grafů

Pro kvalitativní proměnné vykreslujeme koláčový graf, pro numerické histogram a boxplot (viz Obr. 10) a vykreslené grafy můžeme uložit jako obrázek s příponou *.png. Pokud bude počet dat menších než 30, v grafu i textovém poli na to budeme upozorněni.



Obr. 10: Program – grafy numerické proměnné

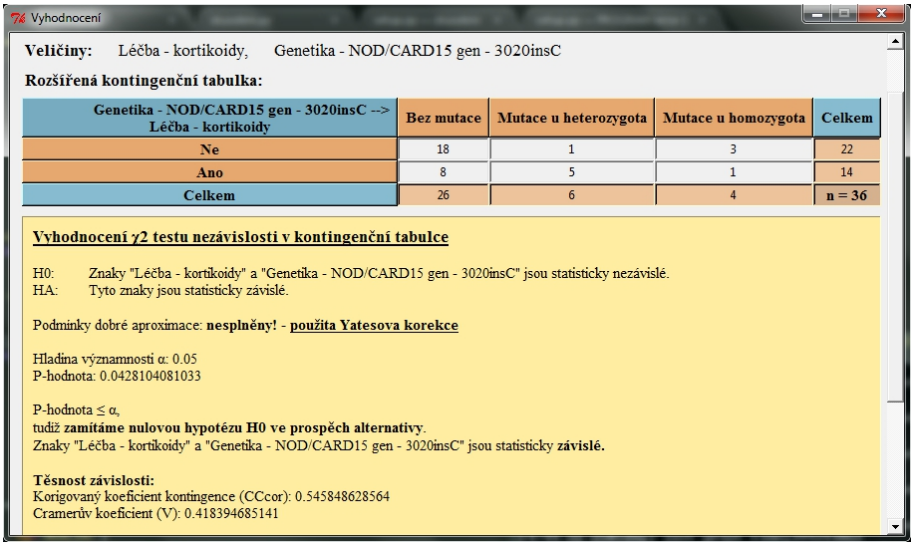
Pokud chceme provést **analýzu závislostí veličin v kontingenční tabulce**, dostaneme se k ní přes sekci „Analýza dat“. Nabídne se nám výběr dvou kategoriálních veličin, z nichž první se týká těch, u kterých chtějí lékaři zjistit souvislost s genetickými mutacemi, druhá se týká genů. Máme na výběr i hladinu významnosti α (viz Obr. 11).



Obr. 11: Program – nabídka pro analýzy závislostí

Jako první se vždy vyhodnocuje analýza závislostí v kontingenční tabulce, protože počet sloupců je vždy roven 3, takže asociační tabulku nemůžeme použít (viz Obr. 12). Pokud si tuto analýzu necháme zobrazit, vykreslí se nejprve rozšířená kontingenční tabulka a poté podrobné statistické vyhodnocení χ^2 testu nezávislosti. Pokud zamítneme nulovou hypotézu o nezávislosti daných veličin, zobrazí se ve výpisu 2 koeficienty síly závislosti – koeficient korelace a Cramerův koeficient, z nichž si uživatel může vybrat, podle kterého se bude řídit.

Při nedostatku dat, tzn. bude-li použita Yatesova korekce χ^2 testu nezávislosti, je uživatel na tento fakt upozorněn a je mu nabídnuta možnost znovuanalyzování závislosti daných veličin tentokrát se sloučenými variantami mutací genu. Tzn. zbudou u veličiny genu jen dvě varianty, buď „s mutací“ nebo „bez mutace“, to může sloužit pro potřeby zjištění závislosti, kdy nás zajímá jen, jestli daný gen zmutoval či nikoliv. V tomto znovuprovedení se při počtu řádků rovném 2 a pokud žádné pole tabulky není rovno 0, vyhodnotí analýza závislosti v asociační tabulce. Program nám dá možnost nahlédnout do asociační tabulky a poté do vyhodnocení, ve kterém nabídne dvě možnosti výsledku – poměr šancí a relativní riziko, z nichž si uživatel opět může vybrat, co je pro něj vhodnější. Pokud je počet řádků tabulky vyšší než 2, provede se analýza závislosti v kontingenční tabulce. Na konci každého vyhodnocení je zobrazen závěr, zda zamítáme nebo nezamítáme nulovou hypotézu H_0, tzn. jestli jsou dané veličiny závislé či nikoliv.



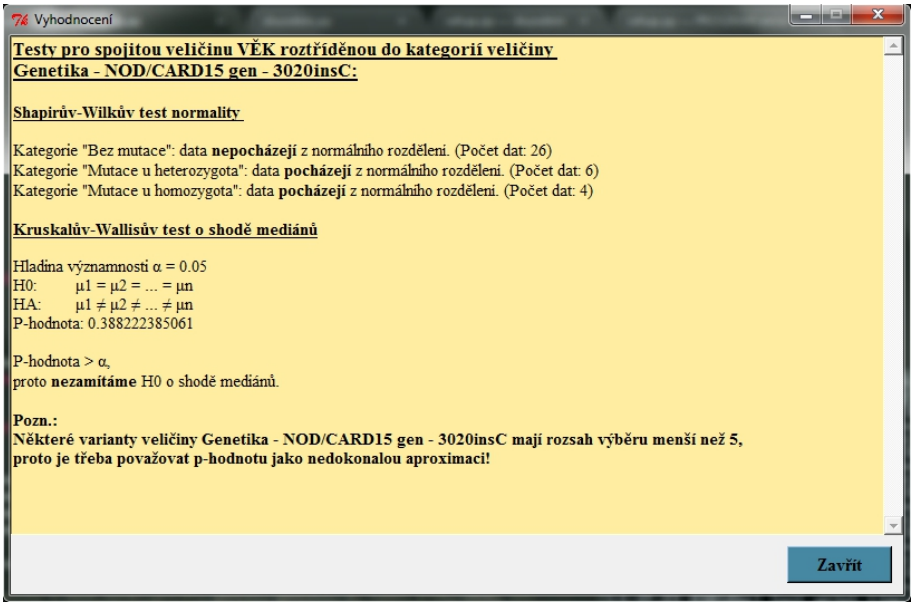
Obr. 12: Program – analýza závislostí v kontingenční tabulce

Chceme-li provést **Kruskalův-Wallisův test shody mediánů**, dostaneme se k němu přes sekci „Analýza dat“ a poté „Analýza závislostí“ (Obr. 11). Testování je provedeno s hladinou významnosti $\alpha = 0,05$. Toto testování je pro spojitou veličinu „věk“, která je rozdělena do kategorií vybrané veličiny. Nejprve se provede Shapirův-Wilkův test normality, abychom dostali informaci o normalitě dat. Pokud bude nedostatek dat ($n_i < 3$ pro varianty mutace genu), test se neprovede. Po Shapirově-Wilkově testu normality následuje Kruskalův-Wallisův test o shodě mediánů, jehož vyhodnocení dostaneme k nahlédnutí (viz Obr. 13). Při zamítnutí nulové hypotézy se provede post hoc analýza, kterou nám program také nabídne. Pokud rozsah výběru bude menší než 5, budeme na tento fakt upozorněni, jelikož pak je třeba považovat *p-hodnotu* jako nedokonalou aproximaci.

Uživatel má k dispozici i manuál (nápovědu) v sekci „Nápověda“, který mu dokáže pomoci orientovat se v programu a případně nastínit nějaké statistické pojmy. Pro správné zobrazení je nutné mít nainstalovaný dostatečně aktualizovaný prohlížeč dokumentů typu PDF.

5.3 Výpočet součtů pořadí v Kruskalově-Wallisově testu

Jako ukázku kódu si zde uvedeme výpočet součtů pořadí pro jednotlivé výběry T_i v Kruskalově-Wallisově testu. Všech n pozorovaných hodnot seřadíme do rostoucí posloupnosti a určíme jejich pořadí R_i, musíme si ale pamatovat, které pořadí je pro který výběr. Následně určíme součty pořadí pro jednotlivé výběry T_i. V cyklu procházíme všechny hodnoty seřazených výběrů a v každé iteraci porovnáváme danou hodnotu s předchozí. Pokud se nerovnají, uložíme pořadí předchozí hodnoty do součtu pořadí správného výběru, pokud se rovnají, pak postupujeme dále, dokud nenarazíme na odlišnou hodnotu.



Obr. 13: Program – vyhodnocení Shapirova-Wilkova a Kruskalova-Wallisova testu

Až na ni narazíme, tak všechna pořadí těchto stejných hodnot zprůměrujeme a tento průměr vložíme do součtů správných výběrů pro všechny tyto hodnoty. Nakonec ještě přidáme pořadí poslední hodnoty.

```
h = 1
T = [0, 0, 0]
for i in range(len(data) - 1):
    if data[i + 1] != data[i]:
        tmp = 0
        for j in range(h):
            tmp += i - j + 1
        for j in range(h):
            T[data2[i - j]] += (tmp / float(h))
        h = 1
    else:
        if i != len(data) - 2:
            h += 1
        else:
            tmp = 0
            for j in range(h):
                tmp += i - j + 1
            for j in range(h):
                T[data2[i - j]] += (tmp / float(h))
T[data2[len(data) - 1]] += len(data)
```

Výpis 1: Výpočet součtů pořadí v Kruskalově-Wallisově testu

6 Závěr

Práce plní všechny body zadání a dělí se na dvě hlavní části, a to programovou část provedenou v jazyce Python a teoretickou část zaměřenou na explorační analýzu a analýzu závislostí. Výstupy naimplementovaných procedur a vyhodnocení jsme kontrolovali programem R a ručním počítáním.

Zabývali jsme se Crohnovou nemocí a její problematikou, základními statistickými pojmy a explorační analýzou kvalitativní i numerické proměnné, statistickými testy a dalšími důležitými pojmy jako intervalový odhad či testování hypotéz, které jsme pak využili u analýzy závislostí, jenž byla jedním z hlavních cílů práce. Všechny tyto znalosti jsme převedli do vytvořeného programu, ve kterém zpracováváme a vyhodnocujeme data zpřístupněná FNO. V první fázi se zaměřujeme na explorační analýzu dat, zjištěné informace můžeme vykreslit v grafech nebo uložit do textového souboru pro vytisknutí nebo pozdější využití. V druhé fázi analyzujeme závislosti mezi kategoriálními veličinami, nechybí zde kontingenční či asociační tabulky a podrobné statistické vyhodnocení analýzy závislostí. Vzhledem k menšímu množství dat je při Yatesově korekci χ^2 testu nezávislosti nabízena možnost znovuanalyzování daných veličin se sloučenými variantami mutací genu, která může sloužit pro potřeby zjištění závislosti, kdy nás zajímá jen, jestli daný gen zmutoval či nikoliv. Je zde i možnost otestovat Kruskalovým-Wallisovým testem shodu mediánů pro spojitou veličinu *věk* roztříděnou do kategorií mutace genu.

Vzhledem k nízkému počtu dat nelze prozatím z této práce vyvozovat relevantní závěry. Po provedení testů závislostí ale můžeme přednostně sledovat veličiny, u nichž se dá očekávat, že by se v budoucnu mohly vyvíjet směrem buď k závislosti mezi jimi a genetickými mutacemi (léčba kortikoidy) nebo nezávislosti (nutnost chirurgické léčby).

Spolupráce s lékaři FNO byla velice příjemná a přínosná. Lékaři byli vstřícní nejen k doplňování chybějících údajů do databáze, ale také k možnostem konzultací, pokud byly potřeba. Požadavky na zpracování dat formulovali průběžně a v případě dalších požadavků bude program dodatečně doplněn. Aplikace bude sloužit lékařům i pro pozdější využití s ohledem na stále probíhající sběr dat a jejich doplňování do databáze. Není vyloučeno, že se mi v budoucnu naskytne příležitost podrobnějšího studia a analýzy Crohnovy nemoci nebo možnost podílet se i na jiných výzkumech a aktivitách FNO.

Díky předmětům, ve kterých jsem se učil základům programování, jsem se rychle adaptoval na programovací jazyk Python, s nímž jsem se v minulosti nesetkal. Studium látky potřebné pro vykonání této práce a samotný proces její tvorby mi pomohly prohloubit mé znalosti ohledně statistického zpracování dat, které budu v budoucnu potřebovat a dále rozvíjet. V praxi jsem si ověřil, že sběr vhodných dat je velice důležitý a mnohdy také zdlouhavý a složitý.

Štěpán Bernady

7 Reference

[1] Soukromé konzultace s odborníky z FNO – MUDr. Lubomírem Martínkem, Ph.D. a prof. MUDr. Petrem Dítětem, DrSc.

[2] LITSCHMANNOVÁ, Martina. *Úvod do statistiky* [online]. Ostrava, 2011 [cit. 24. dubna 2014]. Dostupné z: http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/uvod_do_statistiky.pdf

[3] Briš R., Litschmannová M., *STATISTIKA I. pro kombinované a distanční studium*, Elektronické skriptum VŠB TU Ostrava, 2004

[4] ZAIONTZ, Charles. *Real Statistics Using Excel* [online]. 2013 [cit. 4. května 2014]. Dostupné z: http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/shapiro-wilk-test/

A Tabulky

n	Hladina významnosti α		
	$\mathbf{0,1}$	$\mathbf{0,05}$	$\mathbf{0,01}$
3	$0,789$	$0,767$	$0,753$
4	$0,792$	$0,748$	$0,687$
5	$0,806$	$0,762$	$0,686$
6	$0,826$	$0,788$	$0,713$
7	$0,838$	$0,803$	$0,73$
8	$0,851$	$0,818$	$0,749$
9	$0,859$	$0,829$	$0,764$
10	$0,869$	$0,842$	$0,781$
11	$0,876$	$0,85$	$0,792$
12	$0,883$	$0,859$	$0,805$
13	$0,889$	$0,866$	$0,814$
14	$0,895$	$0,874$	$0,825$
15	$0,901$	$0,881$	$0,835$
16	$0,906$	$0,887$	$0,844$
17	$0,91$	$0,892$	$0,851$
18	$0,914$	$0,897$	$0,858$
19	$0,917$	$0,901$	$0,863$
20	$0,92$	$0,905$	$0,868$
21	$0,923$	$0,908$	$0,873$
22	$0,926$	$0,911$	$0,878$
23	$0,928$	$0,914$	$0,881$
24	$0,93$	$0,916$	$0,884$
25	$0,931$	$0,918$	$0,888$
26	$0,933$	$0,92$	$0,891$
27	$0,935$	$0,923$	$0,894$
28	$0,936$	$0,924$	$0,896$
29	$0,937$	$0,926$	$0,898$
30	$0,939$	$0,927$	$0,9$
31	$0,94$	$0,929$	$0,902$
32	$0,941$	$0,93$	$0,904$
33	$0,942$	$0,931$	$0,906$
34	$0,943$	$0,933$	$0,908$
35	$0,944$	$0,934$	$0,91$
36	$0,945$	$0,935$	$0,912$
37	$0,946$	$0,936$	$0,914$
38	$0,947$	$0,938$	$0,916$
39	$0,948$	$0,939$	$0,917$
40	$0,949$	$0,94$	$0,919$
41	$0,95$	$0,941$	$0,92$
42	$0,951$	$0,942$	$0,922$
43	$0,951$	$0,943$	$0,923$
44	$0,952$	$0,944$	$0,924$
45	$0,953$	$0,945$	$0,926$
46	$0,953$	$0,945$	$0,927$
47	$0,954$	$0,946$	$0,928$
48	$0,954$	$0,947$	$0,929$
49	$0,955$	$0,947$	$0,929$
50	$0,955$	$0,947$	$0,93$

Tab. 9: Kritické hodnoty Shapirova-Wilkova testu

B Příloha na CD

Obsah CD

- Složka „PROGRAM“, obsahem této složky je spustitelný soubor uvodni_obrazovka.exe a další soubory a složky potřebné k jeho běhu
- Složka „SKRIPTY“, ve které jsou uloženy skripty programovacího jazyku Python
- Textový soubor „README.txt“, kde jsou uloženy informace o programu