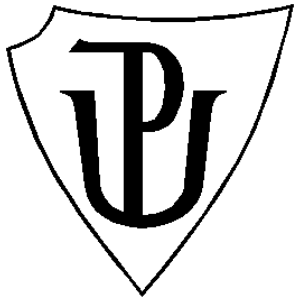
UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA GEOINFORMATIKY - DEPARTMENT OF GEOINFORMATICS

BAKALÁŘSKÁ PRÁCE

Neparametrické testování dvou a více náhodných výběrů z neznámého rozdělení pravděpodobností s využitím ESRI produktů.



Vypracoval: **Radek Brablec**

Vedoucí: **Mgr. Pavel Tuček, Ph.D.**

Olomouc 2010

Prohlášení

Prohlašuji, že jsem na základě zadání diplomové práce vytvořil tuto diplomovou práci samostatně, za vedení pana Mgr. Pavla Tučka, Ph.D., a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 25. května 2010

.....

Radek Brablec

Vysoká škola: Univerzita Palackého v Olomouci **Fakulta:** Přírodovědecká

Katedra: Geoinformatiky **Školní rok:** 2009-2010

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

pro R a d e k B R A B L E C

obor **Geoinformatika a Geografie**

Název tématu: Neparametrické testování dvou a více náhodných výběrů z neznámého rozdělení pravděpodobností s využitím ESRI produktů.

Název (Eng): Nonparametric testing of two or more random samples from an unknown probability distribution using ESRI products.

Zásady pro vypracování:

Student na základě dostupné literatury a za pomoci své invence, s odbornou poradou vedoucího bakalářské práce a konzultanta, vypracuje pojednání o využití neparametrického testování dvou a více náhodných výběrů z neznámých rozdělení pravděpodobností. Cíl spočívá jednak v sepsání krátké rešerše na téma neparametrické metody matematické statistiky. V další části své bakalářské práce se student zaměří na sofistikovaný popis používané teorie, kde bude rozebírat rovněž i dostupnost těchto metod v běžně používaných softwarech. Student provede všechny potřebné výpočty a tvorbu grafických vizualizací v doporučeném softwaru a vyhodnotí možnosti takovýchto analýz v jiném softwarovém řešení používaném na katedře geoinformatiky Univerzity Palackého v Olomouci. V závěru své bakalářské práce student prokáže svou schopnost odborného vyhodnocení poskytnutého datového souboru a prokáže také schopnost signifikantní vizualizace.

O bakalářské práci student vytvoří internetovou stránku, která bude týden před odevzdáním práce umístěna na server UP. Na závěr práce připojí třístránkové resumé v anglickém jazyce. Výstupy budou odevzdány v digitální podobě na CD – ROM. Student odevzdá údaje o všech datových sadách, které vytvořil nebo získal v rámci práce, pro potřeby zaevidování do Metainformačního systému katedry geoinformatiky ve formě vyplněného dotazníku. Práce bude zpracována podle zásad dle Voženílek (2002).

Rozsah grafických prací:

V průběhu sepisování bakalářské práce se bude student setkávat s grafickými výstupy jednotlivého používaného programového vybavení. Tyto budou přehledně řazeny v průběhu bakalářské práce. Větší formáty a mapy budou zařazeny výhradně jako příloha bakalářské práce. Student bude dále analyzovat získaná data, jejichž grafická podoba by měla být též připojena. Podstatou grafických výstupů je seznámit čtenáře bakalářské práce s výsledky tak, aby jakožto běžný uživatel, resp. čtenář, mající základní znalosti z oboru, mohl na základě takto předložených vizualizací usoudit, jaký problém byl zpracováván a jaké výsledky jsou mu předkládány. Další grafické výstupy budou zařazovány dle potřeby práce v průběhu její tvorby.

Rozsah průvodní zprávy:

Celá bakalářská práce by měla být v rozsahu 25 - 35 stran vlastního textu psaného pomocí typografického systému TEX, *který bude obsahovat zejména tří stěžejní kapitoly, které budou obsahovat úvodní pojednání společně s rešerší dané problematiky, teoretické podklady s použitou teorií (zejména analogii běžně používaných parametrických metod) a vlastní zpracování praktického zadání. Student na závěr své práce shrne vlastní přínos do dané problematiky a shrne dosažené výstupy. V neposlední řadě bude práce obsahovat přílohy, které budou složeny z výstupů daného programového vybavení a z okomentovaného programového kódu, který si studentka sama napíše za účelem automatizace výpočtů apod.*

Seznam odborné literatury:

Sprent, P.: *Applied Nonparametric Statistical Methods.* 2nd edition, Chapman & Hall, London, 1993

M. Meloun and J. Militký (1994) Statistické zpracování experimentálních dat. PLUS s.r.o. Praha.

P. Hebák, J. Hustopecký, E. Jarošová, I. Pecáková. Vícerozměrné statistické metody (1) Praha 2007

P. Hebák, J. Hustopecký, a kol. Vícerozměrné statistické metody. (2) Praha 2005

P. Hebák, J. Hustopecký. Vícerozměrné statistické metody (3) Praha 2007

Kunderová, P. Úvod do pravděpodobnosti a matematická statistika. Olomouc 1997

Voženílek, V. Diplomové práce z geoinformatiky. Olomouc 2002

Anděl, J.: Matematika náhody. Matfyzpress Praha 2000.

Anděl, J.: Statistické metody. Matfyzpress Praha 1998.

Antoch, J.,Vorlíčková, D.: Vybrané metody statistické analýzy dat. Academia Praha 1992.

Vedoucí bakalářské práce: Mgr. Pavel TUČEK, Ph.D.

Konzultant bakalářské práce: Mgr. Jaroslav MAREK, Ph.D.

Datum zadání bakalářské práce: Červen 2009

Termín odevzdání bakalářské práce: Květen 2010

L.S

UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA GEOINFORMATIKY
tř. Svobody 26, 771 46 Olomouc
-1-

Vedoucí katedry

Vedoucí bakalářské práce

V Olomouci dne

Poděkování

Rád bych na tomto místě poděkoval vedoucímu diplomové práce panu Mgr. Pavlu Tučkovi, Ph.D., za čas, který věnoval konzultacím a za jeho cenné připomínky při vzniku této práce. Dále své rodině za vynaloženou podporu při studiu a tvorbě bakalářské práce.

Obsah

ÚVOD

Matematická statistika je vědní disciplína na rozhraní popisné statistiky a aplikované matematiky. S využitím metod teorie pravděpodobnosti se snaží odhadnout vlastnosti rozdělení pozorovaných dat. Mezi tyto metody patří parametrické a neparametrické testování. Parametrické testy předpokládají konkrétní rozdělení dat s využitím daného parametru pro výpočet, ale pokud neznáme rozdělení dat, použijeme pro výpočet neparametrické testy. Neparametrické testování používáme také pro data ordinální stupnice. Nevýhodou neparametrických testů je menší vypovídající hodnota obzvláště u menšího počtu naměřených dat. Většina parametrických testů má svoji neparametrickou obdobu.

Pro bakalářskou práci byla využita data ze statistického průzkumu jízdních dokladů veřejné linkové dopravy Integrovaného dopravního systému [redacted] kraje města [redacted]. Veřejnou linkovou dopravu ve výše zmíněném obvodu zajišťují 4 dopravní společnosti [redacted]. Městská hromadná doprava je zajištěná dopravní společností [redacted]. Dle ustanovení Integrovaného systému [redacted], lze využívat platného jízdního dokladu pro cestu u jiného dopravce. V rámci průzkumu bylo šetřeno křížové využívání jízdních dokladů od různých dopravců, další součástí sledování bylo sčítání počtu nastupujících a vystupujících cestujících na jednotlivých zastávkách pro daný spoj. Výsledkem šetření je třeba zjistit, zdali se naměřená data mezi sebou shodují či nikoliv. K vyjádření výsledků jsme použili neparametrické testování.

Vzhledem k ochraně údajů a dat, které byly získány od dopravních společností a firmy zajišťující statistické šetření a vyhodnocování nelze poskytnout text a data bakalářské práce třetí osobě.

1 CÍL PRÁCE

Cílem práce bylo studium neparametrických testovacích metod. Hypotetická neparametrická data jsme srovnávali s praktickými výsledky naměřenými při tříd denním průzkumu veřejné linkové dopravy [ ]. Získané výsledky jsme dále použili pro vzájemné porovnání v rámci různých parametrů. Získaná data jsme vizualizovali pomocí produktu ESRI.

Současně bude vytvořena webová stránka, která bude umístěna na serveru katedry Geoinformatiky UP. Na závěr práce připojím třístránkové resumé v anglickém jazyce. Výstupy budou odevzdány v digitální podobě na CD - ROM.

2 TEORETICKÁ ČÁST

2.1 Historie

JAROSLAV HÁJEK

∗4.2.1926 (Poděbrady) 10.6.1974 (Praha)

Jaroslav Hájek [12] se narodil a vyrůstal ve známém lázeňském městě Poděbrady, kde chodil na základní školu, ale gymnázium vystudoval už v Praze.

V průběhu jeho studií na gymnáziu začala 2. světová válka a Hájek byl nucen pracovat pro německý zbrojní průmysl. Na čas přerušené studium na gymnáziu zakončil maturitou až po skončení 2. světové války a poté nastoupil na České vysoké učení technické (ČVUT), kde začal studovat statistické inženýrství. V roce 1947, pouhé 2 roky po maturitě, Hájek začal učit jako asistent v matematickém institutu na ČVUT. V roce 1949 obdržel diplom ve statistickém inženýrství a tentýž rok vydal svou první publikaci. Tato publikace byla na téma Výběrových přehledů. Hájek patřil mezi průkopníky v oblasti pravděpodobnosti a díky jeho příspěvkům k používání pomocných dat v oblasti odhadů veřejného mínění si vysloužil přezdívku "Hájek věštec"'.

Roku 1951 Hájek začal s jeho výzkumnými studiemi na matematickém institutu Československé akademie věd. Roku 1954 mu byl udělen doktorát a on dále pokračoval ve své práci na matematickém institutu jako výzkumný pracovník. Tuto výzkumnou práci prováděl po 12 let, což bylo jeho nejprodukti vnější období, napsal 20 spisů a 2 knihy:

- *Teorie pravděpodobnostního výběru s aplikacemi na výběrová šetření*
- *Pravděpodobnost ve vědě a technice*

V průběhu těchto 12 ti let si Hájek získal silnou mezinárodní reputaci, stal se uznávanou autoritou v několika různých sférách statistiky, zejména v neparametrických metodách a jejich asymptotické teorii.

Také lze zmínit ještě 2 další Hájkovy knihy a to:

- *Kurz neparametrické statistiky*
- *Výběr z konečné populace*

Ty byly ovšem vydány až v roce 1981, 7 let po Hájkově smrti.

Roku 1973 mu byla udělena státní cena Klementa Gottwalda za práce o asymptotické teorii pořadových testů. Zemřel ve věku 48 let po transplantaci ledvin.

Jacob Wolfowitze

∗19.3.1910 (Varšava) 16.7.1981 (Tampa, Florida)

Jacob Wolfowitze [13] se narodil ve Varšavě spadající pod tehdejší Ruské impérium (dnešní Polsko). V deseti letech emigroval za svým otcem do USA. Zde započal své studium na High School v New Yorku. Zatímco Wolfowitz byl v polovině svého vysokoškolského studia, započala velká hospodářská krize, která trvala od roku 1929 do roku 1932. Když Wolfowitz vystudoval v roce 1931, tak byla malá šance na dobré zaměstnání, proto pokračoval dále doktorským studie a dalších deset let strávil výukou matematiky na mnoha různých vysokých školách. První publikaci, kterou Wolfowitz napsal, byla ve spolupráci s AbrahamWaldem a nazývala se Nonparametric inference, v této publikaci v roce 1942 poprvé bylo zapsáno slovo "neparametrické". Dále v roce 1942 získal Wolfowitz doktorát na New York University a stál se součástí statistické výzkumné skupiny na Columbia University. V dalších letech působil také jako externista na mnoha univerzitách jak v USA, tak i v Evropě. Wolfowitz obdržel mnoho vyznamenání za jeho vynikající příspěvky ke statistikám. Byl zvolen do Národní akademie věd a Americké akademie umění a věd. Byl zvolen Fellow Ekonometrické společnosti, Technion, v Izraeli, mu udělila čestný titul v roce 1975. Zemřel na infarkt v Tampa, Florida, kde byl profesorem na University of South Florida. Wolfowitz příspěvky byly hlavně v oblasti teorie statistického rozhodování, neparametrické statistiky, sekvenční analýzy a informačních teorií.

2.2 Rešerše

STATISTICKÉ METODY
Jiří Anděl
Nakladatelství Matfyzpress, Praha 1998
ISBN 80 – 85863 – 27 – 8

Publikaci sepsal matematika a statistik prof. RNDr. Jiří Anděl, DrSc., profesor katedry pravděpodobnosti a matematické na Matematicko-fyzikální fakultě Univerzity Karlovy. Jedná se o první přepracování již publikovaného vydání z roku 1993, kde došlo k úpravám obrázků a rozšíření textů o další informace. Tato publikace vznikla na základě zjištění, že jednotlivé statistické metody jsou v jiných knihách nedostatečně popsány a uživatel tak zpracovává statistická data nevhodným způsobem. Kniha je psána na zcela obecných základech, kde jednodušší tvrzení jsou podrobně dokazována a u složitějších metod bývá spíše uváděna jejich motivace a příslušný algoritmus. Z hlediska neparametrických testů jsou zde přehledně uváděny jednotlivé metody jednovýběrových, dvouvýběrových a porovnání několika výběrových testů. Toto druhé vydání bylo podpořeno grantem GAČR 201/97/1176.

VYBRANÉ METODY STATISTICKÉ ANALÝZY DAT
Jaromír Antoch, Dana Vorlíčková
Nakladatelství Academia, Praha 1992
ISBN 80 – 200 – 0204 – 9

Na této knize se podíleli prof. RNDr. Jaromír Antoch, CSc., pracovník na katedře pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulty Univerzity Karlovy a externí pracovnice pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulty Univerzity Karlovy RNDr. Dana Vorlíčková, CSc.. Publikaci zaměřili na nestandardní statistické postupy aplikovatelné v každodenní praxi. Kladli při tom zřetel nejen na popis jednotlivých metod ale i na jejich výpočetní aspekty. K sepsání vedl rychlý rozvoj nových netradičních

postupů a stále více využívaná výpočetní technika obzvláště ve statistice. Na publikaci byla poskytnuta dotace MŠMT ČR v rámci matematicko-fyzikální fakulty Univerzity Karlovy.

ÚVOD DO TEORIE PRAVDĚPODOBNOSTI A MATEMATICKÉ STATISTIKY
Pavla Kunderová
Nakladatelství Univerzity Palackého, Olomouc 1997
ISBN 80-7067-710-4

Skripta, Úvod do teorie pravděpodobnosti a matematické statistiky sepsala doc. RNDr. Pavla Kunderová, CSc. pro potřeby doplnění teorie k přednáškám z teorie pravděpodobnosti a matematické statistiky. Ve skriptech se nachází kapitoly zabývající se axiomatické teorii pravděpodobností, dále kapitoly pojednávající o náhodné veličině a náhodném vektoru nebo klasické limitní věty z teorie pravděpodobností. Na závěr je kniha doplněna o některé definice a věty z teorie míry a z teorie integrálu. Doc. RNDr. Pavla Kunderová, Csc. působila dlouhá léta na katedře matematické analýzy a numerická matematiky, Přírodovědecké fakulty University Palackého a nyní působí jako externista na dané katedře.

PRAVDĚPODOBNOST A STATISTIKA
Petr Otipka, Vladislav Šmajstrla
Nakladatelství VŠB-TU, Ostrava 2006
ISBN 80 – 248 – 1194 – 4

Skripta vypracovala dvojce statistiků a matematiků Mgr. Petr Otipka a doc. PaedDr. Vladislav Šmajstrla v té době oba pracovníci katedry matematiky a deskriptivní geometrie na VYSOKÁ ŠKOLA BÁŇSKÁ - TECHNICKÁ UNIVERZITA OSTRAVA. V roce 2006 doc. PaedDr. Vladislav Šmajstrla přestoupil na Vysoká škola logistiky o.p.s. v Přerově. Skripta vypracovali pro potřeby studentů VŠB-TU s cílem zajistit ucelený podklad pro doplnění přednášek, ale hlavně

pro potřeby studentů kombinovaných a distančních forem studia, kteří se nemohou účastnit přednášek přímo. Skripta jsou členěna na dvě základní části. První z nich je věnována základům počtu pravděpodobnosti, druhá úvodu do problematiky matematické statistiky. Každá kapitola obsahuje příklady s podrobným řešením a v závěru sadu neřešených úloh s výsledky. Vytvořeno v rámci projektu Operačního programu Rozvoje lidských zdrojů CZ.04.1.03/3.2.15.1/0016 a dále byl spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

Archivum Mathematicum
Nakladatelství Masarykova Universita, Brno 1965 - do současnosti
ISSN 1212 – 5059 (elektronická edice)
ISSN 0044 – 8753 (tisková edice)

Jedná se o matematický časopis, který vydává výlučně vědecky matematických prací. Časopis byl založen Prof. RNDr. Otakar Borůvka, DrSc. v roce 1965, významným matematikem, zabývající se matematickou analýzou, teorií grafů, diferenciální geometrií, algebrou a diferenciálními rovnicemi. Byl to vysokoškolský pedagog a zakladatel Matematického ústavu ČSAV v Brně. Časopis je publikován ústavem matematiky a statistiky Přírodovědecké fakulty Masarykovi University a podílí se na jeho obsahu přední čeští matematici a statistici především s Masarykovi a Karlovy university a dále také spolupracují s předními matematiky a statistiky ze zahraničí a to s Universitat Wien a Universitat der Bundeswehr Munchen. Časopis vychází v ročníkových intervalech a od roku 1965 do posud, vyšlo již 45 výtisků.

2.3 Data a jejich získávání

Statistická data [3], v dnešní době dostupná obvykle v podobě počítačových databází, se dají zkoumat z různých hledisek. Data především mohou být úplná a zahrnovat celou základní populaci (čili základní soubor), tedy všechny objekty našeho zájmu. Častěji však máme k dispozici jen jejich podmnožinu, zvanou ve

statistice výběr, výběrový soubor, výběrová populace či vzorek. Počet objektů v této podmnožině se označuje n a nazývá rozsah výběru. Postupy získávání výběru zkoumá teorie výběru, která se zabývá mimo jiné tím, zda je výběr reprezentativní, tedy zda popisné charakteristiky výběru se až na náhodnou výběrovou chybu shodují s charakteristikami celé základní populace. Základním způsobem dosahování reprezentativnosti přitom jsou různé druhy pravděpodobnostního výběru, při nichž má každý prvek základní populace známou nenulovou pravděpodobnost, že bude obsažen ve vzorku. Není-li výběr reprezentativní, vzniká systematická chyba, která znemožňuje korektní zobecnění výsledků analýzy na celou základní populaci. Často však není pravděpodobnostní výběr možný a jsou k dispozici např. pouze data vzniklá "na základě příležitosti" (oportunitní), o jejichž reprezentativnosti není jasno - to se týká např. mnoha situací v astronomii nebo historických vědách. V takovém případě je k zobecnění potřeba přistupovat s velkou opatrností.

Zkoumají-li se kausální závislosti, tedy vliv různých zásahů, používá se experimentální design. Například některé náhodně vybrané prvky populace mohou být podrobeny zásahu, jejíž efekt se zkoumá, zatímco zbylé slouží jako kontrolní skupina. Rozdíl mezi ošetřenou a kontrolní skupinou pak lze až na výběrovou chybu interpretovat jako vliv zásahu. Do designu vytváření a sběru dat se může promítnout i čas, takže hovoříme o časových řadách a longitudinálních studiích. Data obsahují hodnoty sledovaných znaků (či - z hlediska datového souboru - proměnných), což mohou být hodnoty jak numerické (např. délka života pacienta po operaci), tak i nenumerické, kategoriální (např. umístění nádoru v těle). Podle toho, jakou interpretaci numerická data mají, tj. zda je např. lze pouze seřadit, nebo zda je lze i sčítat, hovoří se pak ještě o měřítku proměnné čili typu škály. Zvláštním problémem analýzy jsou chybějící údaje - data, která nebyla zjištěna, ztratila se anebo nejsou smysluplně definována.

Základní úlohou matematické statistiky je zobecnění (zvané v tomto oboru statistická inference, statistická indukce či statistické usuzování): zkoumá se, jak informace zjištěné o prvcích výběru zobecnit na celou populaci. Používané metody se opírají o zákon velkých čísel a příbuzné věty teorie pravděpodobnosti,

jako je například Glivenkova-Cantelliho věta; ty ukazují, že při rostoucím rozsahu reprezentativního výběru se výběrové odhady obvykle limitně blíží skutečným hodnotám na celé populaci. Matematická statistika zároveň stanovuje, jak přesný tento odhad pro daná data je (intervalový odhad), anebo testuje, zda vlastnosti vzorku jsou slučitelné s předpoklady o chování celé populace (testování statistických hypotéz).

2.4 Testování hypotéz

Testování hypotézy [3] je postup, který umožňuje na základě naměřených dat určit, zda náhodná veličina, jejímiž realizacemi data jsou, vykazuje určitou vlastnost. Například lze testovat, zda se střední hodnota náhodné veličiny liší od dané konstantní hodnoty - praktickou aplikací takového testu by mohlo být, zda je soustruh dobře seřízen a střední hodnota průměru jím vyráběných součástek se rovná hodnotě předepsané výkresem. V takovém případě je možné použít jednovýběrový t-test, jsou-li průměry součástek normálně rozděleny.

V klasické teorii testování se vychází z toho, že platí předpokládaná vlastnost zkoumaných náhodných veličin. Tento předpoklad, se označuje jako nulová hypotéza a značí H_0. Jelikož data jsou náhodná a náhoda může "pracovat proti nám", nelze obvykle závěry testování vyslovit s naprostou jistotou. Proto se zároveň předem stanoví hladina spolehlivosti α, což je míra rizika (pravděpodobnost) toho, že hypotézu H_0 zamítneme, ačkoliv ve skutečnosti platí (omyl označovaný jako chyba 1.*druhu*). Hladina spolehlivosti se tradičně stanovuje $0,05$ nebo $0,01$. Menší hladina spolehlivosti znamená větší jistotu při zamítání nulové hypotézy, ale zároveň také větší riziko chyby 2.*druhu*, jež spočívá v akceptování nulové hypotézy, ačkoli tato hypotéza ve skutečnosti neplatí.

Dále se z dat vypočítá takzvané testovací kritérium, jehož rozdělení podmíněné předpokládanou platností nulové hypotézy je známo. Vyjde-li hodnota testovacího kritéria typická pro toto známé rozdělení, nulovou hypotézu akceptujeme či přesněji řečeno nezamítáme na základě známých dat. Naopak vyjde-li hodnota extrémní, tedy v oblasti hodnot, do níž realizace přepokládaného roz-

dělení padají s pravděpodobností menší než α (tj. hodnota testovacího kritéria překročí kritickou mez), usoudíme, že testovací kritérium nejspíše nepochází z předpokládaného rozdělení a nulovou hypotézu zamítneme ve prospěch opačné tzv. alternativní hypotézy, označované H_1.

Zatímco dříve bylo třeba hledat kritické meze v tabulkách rozdělení příslušného testovacího kritéria, dnes statistické softwary vypisují takzvanou hodnotu významnosti (též zvanou signifikance nebo p-hodnota). Tato hodnota udává pravděpodobnost, že při platnosti nulové hypotézy vyjde testová statistika rovna naměřené nebo ještě extrémnější. Test se vyhodnocuje takto:

- *Je-li hodnota významnosti menší než hladina spolehlivosti* ($p < \alpha$), *pak zamítneme nulovou hypotézu a přijmeme alternativní hypotézu. Riskujeme chybu prvního druhu s pravděpodobností nanejvýš* α.
- *Je-li hodnota významnosti větší nebo rovna než hladina spolehlivosti* ($p \geq \alpha$), *pak nulovou hypotézu nezamítneme. Riskujeme chybu druhého druhu s pravděpodobností označovanou* α.

2.5 Testy χ^2 při neznámých parametrech

Často se stává, že pravděpodobnosti $p_1, \ldots, p_k$ závisejí na nějakém neznámém parametru $a = (a_1, \ldots, a_m)'$. Můžeme tedy psát $p_1 = p_1(a), \ldots, p_k = p_k(a)$. Pro každé a však musí platit

$$p_1(a) + \ldots + p_k(a) = 1.$$

Jsou-li funkce $p_i(a)$ dostatečně hladké, dostaneme odtud derivováním

$$\frac{\partial p_1(a)}{\partial a_j} + \ldots + \frac{\partial p_k(a)}{\partial a_j} = 0, \qquad j = 1, \ldots, m. \tag{1}$$

Tento vztah se uplatňuje při úpravách dalších vzorců. Místo máme nyní

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{k} \frac{\chi_i^2}{p_i(a)} - n. \tag{2}$$

K odhadu a se nabízí podobná myšlenka jako je metoda nejmenších čtverců v případě lineárního modelu. Nechť $a*$ je taková hodnota parametru a, která minimalizuje (2). Tu nazýváme odhadem minimálního χ^2.

Zpravidla ji získáme řešením soustavy rovnic

$$\frac{\partial \chi_2(a)}{\partial a_j} = -\frac{1}{n}\sum_{i=1}^{k} \frac{X_i^2}{p_i^2(a)} \frac{\partial p_i(a)}{\partial a_j} = 0, \qquad j = 1, \ldots, m. \tag{3}$$

Vzhledem k tomu, že se tato soustava většinou velmi obtížně řeší, hledala se jiná metoda. Kdybychom místo (2) derivovali podle aj vztah

$$\chi^2 = \sum_{i=1}^{k} \frac{[X_i - np_i(a)]^2}{np_i}, \tag{4}$$

dostali bychom soustavu

$$-\frac{1}{2}\frac{\partial \chi_2(a)}{\partial a_j} = \sum_{i=1}^{k} \left(\frac{X_i np_i(a)}{p_i(a)} + \frac{[X_i - np_i(a)]^2}{2np_i^2(a)} \right) \frac{\partial p_i(a)}{\partial a_j} = 0 \tag{5}$$

pro $j = 1, \ldots, m$. Ta je totožná se soustavou (3). Dá se však ukázat, že s rostoucím n je vliv druhého členu na pravé straně vzorce (5) čím dál tím menší. Pokud tento člen zcela vynecháme, dospějeme k nové soustavě rovnic

$$\sum_{i=1}^{k} \frac{X_i - np_i(a)}{p_i(a)} \frac{\partial p_i(a)}{\partial a_j} = 0, \qquad j = 1, \ldots, m.$$

Užijeme-li ještě vzat (1), získáme nakonec soustavu

$$\sum_{i=1}^{k} \frac{X_i}{p_i(a)} \partial p_i(a) \partial a_j = 0, \qquad j = 1, \ldots, m. \tag{6}$$

Řešením soustavy (6) označíme $\hat{a}$ a budeme ho nazývat odhad parametru a modifikovanou metodou minimálního χ^2.

Věta 2.1. *Nechť $m < k-1$ a nechť pro všechny body a nedegenerovaného konečného uzavřeného intervalu A z $\mathbb{R}_m$ platí:*

(1) $p_1(a) + \ldots + p_k(a) = 1$.

(2) Existuje takové $c > 0$, že $p_i(a) > c^2$ pro $i = 1, \ldots k$.

(3) Každá funkce $p_i(a)$ má spojité derivace.

$$\frac{\partial p_i(a)}{\partial a_j} \quad \text{a} \quad \frac{\partial^2 \mathrm{p_i(a)}}{\partial \mathrm{a_j} \partial \mathrm{a_s}} \quad \text{pro} \quad \mathrm{j,\ s = 1, \ldots, m.}$$

(4) Matice $\left(\frac{\partial p_i(a)}{\partial a_j} \right)$, která je typu $k \times\ m$, má hodnost m.

Nechť a^0 je vnitřním bodem A. Označme $p_i^0 = p_i(a^0)$. Nechť $X = (X1, \ldots, X_k)'$ má multinomické rozdělení s parametry $n, p_1^0, \ldots, p_k^0$. Pak v případě $n \to \infty$ existují takové posloupnosti kladných čísel $\epsilon_n \to 0$ a $\delta_n \to 0$, že soustava (6) má s pravděpodobností alespoň $1 - \epsilon_n$ právě jeden kořen $\hat{a}_n$ takový, že $|\hat{a}_n - a^0| < \hat{a}_n$. Dosadíme-li tento kořen do (7) [nebo zcela ekvivalentně do (2)], má veličina χ^2 při $n \to \infty$ asymptoticky χ^2_{k-m-1} rozdělení.

2.6 Test normalit

Nechť $\xi_1, \ldots, \xi_n$ je náhodný výběr. Chceme testovat hypotézu H_0, že jde o výběr z $N(\mu, \sigma^2)$, kde parametry μ a σ^2 nejsou známy. Nejprve vytvoříme třídy

$$(-\infty, b_1), [b_1, b_2), [b_2, b_3), \ldots, [b_{k-2}, b_{k-1}), [b_{k-1}, \infty),$$

kde $k \geq 4$. Pro stručnost označíme i-tou třídu symbolem J_i. Pravděpodobnost p_i, že daná veličina $\xi_j (j = 1, \ldots, n)$ padne do J_i, je rovna

$$p_i = p_i(\mu, \sigma) = \int_{J_i} f(x)dx,$$

kde

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Z (6) dostaneme po úpravě soustavu

$$\mu = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} xf(x)dx, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} (x-\mu)^2 f(x)dx,$$

která se zpravidla řeší iteračně. Nezapomeňme, že na μ i σ závisejí p_i i $f(x)$, které jsou na pravých stranách těchto rovnic. Řešením soustavy označme $\hat{\mu}$ a $\hat{\sigma}$. Pak vypočteme

$$\chi^2 = \sum_{i=1}^{k} \frac{[X_i - np_i(\hat{\mu}, \hat{\sigma})]^2}{np_i(\hat{\mu}, \hat{\sigma})}.$$

Pokud vyjde $\chi^2 > \chi^2_{k-3}(\alpha)$, zamítáme hypotézu H_0.

Často se také testuje normalita výběru pomocí šikmosti a_3 a špičatosti a_4. Platí-li hypotéza, že $\xi_1, \ldots, \xi_n$ je výběr z normálního rozdělení, pak a_3 a a_4 mají asymptoticky normální rozdělení s parametry

$$Ea_3 = 0, \qquad Ea_4 = 3 - \frac{6}{n+1},$$

$$var\ a_3 = \frac{6(n-2)}{(n+1)(n+3)} \qquad var\ a_4 = \frac{24n(n-3)/n-3)}{(n+1)^2(n+3)(n+5)}.$$

Přitom a_3 a a_4 jsou asymptoticky nekorelované.

Test proti alternativě, že výběr pochází z nějakého nesymetrického rozdělení, se založí na šikmosti a_3. Kritické hodnoty pro $n \leq 25$ lze najít v článku Mulholland (1977) a pro $n > 25$ v tabulkách Pearson a Hartley (1956, 1972). Kritické hodnoty rovněž tabelovali D´Agostino a Stephens (1986). Teprve pro velká n (v praxi pro $n \geq 200$) se dá využít asymptotické normality. Vypočte se veličina

$$U_3 = \frac{a_3}{\sqrt{var\ a_3}}.$$

Pokud $|U3| \geq u(\frac{\alpha}{2})$, zamítá se hypotéza, že jde o výběr z normálního rozdělení.

D´Agostino a kol. (1990) uvádějí podstatné vylepšen tohoto postupu. Položme

$$b = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}, \qquad W^2 = \sqrt{2(b-1)} - 1,$$

$$\sigma = \frac{1}{\sqrt{ln\ W}}, \qquad a = \sqrt{\frac{2}{W^2-n}}, \qquad Z_3 = \sigma\ ln\left[\frac{U_3}{a} + \sqrt{\left(\frac{U_3}{a}\right)^2 + 1}\right].$$

Veličina Z_3 má přibližně rozdělení $N(0,1)$. Pokud vyjde $|Z3| \geq u(\frac{\alpha}{2})$, zamítá se hypotéza, že jde o výběr z normálního rozdělení. Této aproximace se může použít už pro $n > 8$. Tento test byl odvozen v práci D´Agostino (1970).

Test proti alternativám, které se liší špičatostí, se založí na a_4. Kritické hodnoty pro $n \geq 50$ se najdou opět v tabulkách Pearson a Hartley (1956, 1972) a v knize D´Agostino a Stephens (1986). Pro $n \geq 500$ se již užívá limitních výsledků. Vypočte se

$$U_4 = \frac{a_4 - Ea_4}{\sqrt{var\ a_4}}$$

a hypotéza normalit se zamítá v případě, že $|U_4| \geq ?u(\frac{\alpha}{2})$. D´Agostino a kol. (1990) navrhují v tomto případě ještě dále vypočítat

$$B = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}, \qquad A = 6 + \frac{8}{B}\left(\frac{2}{B} + \sqrt{1+\frac{4}{B^2}}\right),$$

$$Z_4 = \frac{1 - \frac{2}{9A} - \sqrt[3]{\frac{1-\frac{2}{A}}{1+U_4\sqrt{\frac{2}{A-4}}}}}{\sqrt{\frac{2}{9A}}}.$$

Veličina Z_4 má přibližně $N(0,1)$ rozdělení. Hypotézu o normalitě výběru zamítáme v případě, že $|Z4| \geq u(\frac{\alpha}{2})$. Aproximace je použitelná pro $n \geq 20$. Test odvodili Anscombe a Glynn (1983).

Test založen na šikmosti a špičatosti zároveň je založen na veličině $U_3^2 + U_4^2$. Pokud vyjde $U_3^2 + U_4^2 \geq \chi_2^2(\alpha)$, zamítá se hypotéza o normalitě. V literatuře se doporučuje, aby tento test byl užíván jen pro rozsahy výběru $n > 200$. Místo toho lze však také test založit na $Z_3^2 + Z_4^2$. Hypotézu normalitě výběru pak zamítáme, když $Z_3^2 + Z_4^2 \geq \chi_2^2(\alpha)$. Nyní však stačí, aby platilo $n \geq 20$.

Gearyho test normality je založen na jakémsi porovnání průměrné odchylky kolem průměru a směrodatné odchylky s. Počítá se

$$g = \frac{1}{s\sqrt{n}} \sum_{i=1}^{n} |\xi_i - \bar{\xi}|.$$

Kritické hodnoty tohoto testu jsou v tabulkách Person a Hartley (1956, 1972). Některé balíky statistických programů počítají D´Agostinův test, který je založen na statistice

$$D_A = \frac{\sum_{i=1}^{n} i\xi_{(i)} - \frac{n(n+1)}{2}\bar{\xi}}{s\sqrt{n^3}}$$

(viz D´Agostino a Pearson 1973 a 1974). Test rovněž vyžaduje speciální tabulky kritických hodnot. Informace o testech normality lze nalézt v pracích Domański (1990), Gastwirth a Owens (1977), Schäbe (1987), Shenton a Bowman (1977).

2.7 Neparametrické testy

Dle [2] neparametrických testů se nepředpokládá konkrétní rozdělení náhodné veličiny, testy mají obecnou platnost. Obvyklým předpokladem je pouze to, že jde o veličinu spojitého typu. U jednotlivých testů mohou být specifikovány další předpoklady (např. nezávislost pozorování). K výhodám těchto testů patří kromě obecnosti také menší citlivost na odlehlá pozorování. Výhodou může být i to, že se některé neparametrické testy (např. pořadové testy) dají použít i k testování ordinálních dat. Mají tedy širší použití než parametrické testy. Na druhou stranu mají i své nevýhody - v případě platnosti určitého rozdělení mají menší sílu než příslušné parametrické testy k danému rozdělení. Pokud bychom chtěli dosáhnout stejné síly, museli bychom naměřit více hodnot, než by bylo nutné u parametrických testů.

2.7.1 Jednovýběrové testy

Jednovýběrové testy slouží k testování hypotézy, která tvrdí, že daný náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení s distribuční funkcí (x). Tuto hypotézu lze ověřovat hned několika různými testy, které mohou být založeny na výpočtu suprema vzdálenosti empirické a teoretické distribuční funkce v jednotlivých bodech, porovnání kvantil-kvantilového grafu s regresní přímkou proloženou danými body, nebo srovnávání očekávaných teoretických a empirických četností v daných třídicích intervalech.

Jednovýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozdělení s hustotou f, která symetrická kolem bodu a. Platí tedy $f(a+x) = f(a-x)$. Z toho plyne, že a musí být rovno mediánu $\tilde{x}$. Existuje-li konečná střední hodnota tohoto rozdělení, pak musí také pro každé i platit $EX_i = a$. Konečnost střední hodnoty se však obecně nepředpokládá. Jednovýběrový Wilcoxonův test je určen k testování hypotézy $H_0 : \tilde{x} = x_0$ proti alternativě $H_1 : \tilde{x} \neq x_0$.

Nejprve předpokládejme, že žádná z veličin X_i není rovna x_0. Položme $Y_i = X_i - x_0$. Veličiny Y_i seřadíme do neklesající posloupnosti podle jejich absolutní hodnoty

$$|Y|(1) \leq |Y|_{(2)} \leq \ldots \leq |Y|_{(n)}.$$

Budiž R_i^+ pořadí veličiny $|Y_i|$. Označme

$$S^+ = \sum_{Y_i \geq 0} R_i^+, \qquad S^- = \sum_{Y_i < 0} R_i^+.$$

Přitom platí $S^+ + S^- = n(n+1)/2$. Je-li číslo $\min(S^+, S^-)$ menší nebo rovno tabelované kritické hodnotě $w_n(\alpha)$, zamítáme H_0.

Z učiněných předpokladů vyplývá, že $Y_1, \ldots, Y_n$ jsou nezávislé stejně rozdělené náhodné veličiny, jejichž rozdělení je symetrické kolem nuly.

Věta 2.2. *Vektory* $(signY_1, \ldots, signY_n)'$ *a* $(|Y|_{(1)}, \ldots, |Y|_{(n)})'$ *jsou nezávislé.*

Důkaz. Jelikož veličiny Y_i jsou nezávislé, vektory $(signY_i, |Y_i|)'$ jsou také nezávislé. Ze spojitosti a ze symetrie rozdělení vyplývá, že

$$P(signY_i = 1) = P(signY_i = -1) = \frac{1}{2}.$$

Dále máme pro libovolné $y > 0$

$$P(sign\ Y_i = 1, |Y_i| < y) = P(0 < Y_i < y) = \frac{1}{2}P(-y < Y_i < y) =$$

$$\frac{1}{2}P|Y_i| < y)P(sign\ Y_i = 1)P(|Y_i| < y).$$

Proto veličiny sign Y_i a $|Y_i|$ jsou pro každé i nezávislé. Celkově dostáváme, že vektory $(sign\ Y_i, \ldots, sign\ Y_n)'$ a $(|Y_1|, \ldots, |Y_n|)'$ jsou nezávislé. Protože vektor $(|Y|_{(1)}, \ldots, |Y|(n))'$ je funkcí vektoru $(|Y_1|, \ldots, |Y_n|)'$, je tím věta dokázána.

Věta 2.3. *Označme* $S = \sum_{i=1}^n R_i^+ sign\ Y_i$ *atd.*

$$S^+ = \frac{1}{2}S + \frac{n(n+1)}{4}.$$

Důkaz. Platí $S^+ - S^- = S, S^+ + S^- = n(n+1)/2$. Odtud vypočteme S^+.

Věta 2.4. *Platí-li H_0, pak*

$$ES^+ = \frac{1}{4}n(n+1), \qquad var\ S^+ = \frac{1}{24}n(n+1)(2n+1).$$

Důkaz. Nejprve si všimneme, že $Esign\ Y_i = 0$ pro každé i. Z věty 1 dostaneme, že $E(R_i^+ sign\ Y_i) = (ER_i^+)(Esign\ Y_i)$, a tak

$$E(R_i^+ sign\ Y_i) = 0 \tag{7}$$

Odtud

$$ES = \sum_{i=1}^{n} E(R_i^+ sign\ Y_i) = 0$$

Vzhledem k (7) platí

$$var(R_i^+ sign\ Y_i) = E(R_i^+ sign\ Y_i)^2 = E(R_i^+)^2 E(signYi)^2$$

$$= E(R_i^+)^2 = 1^2\frac{1}{n} + 2^2\frac{1}{n} + \ldots + n^2\frac{1}{n} = \frac{1}{6}(n+1)(2n+1).$$

Obdobně se dokáže, že platí

$$cov(R_i^+ sign\ Y_i, R_j^+ sign\ Y_j) = 0 \quad proi \neq j.$$

Dá se dokázat, že S^+ má asymptoticky normální rozdělení. Testy hypotézy H_0 lze tudíž také založit na veličině

$$U = \frac{S^+ - ES^+}{\sqrt{var\ S^+}},$$

Kde ES^+ a $var\ S^+$ jsou uvedeny ve větě 2.4. Vyjde-li $|U| \geq u(\frac{\alpha}{2})$, zamítneme H_0 na hladině, která se s rostoucím n blíží číslu α.

Je třeba zdůraznit, že jedním z předpokladů jednovýběrového Wilcoxonova testu je i symetrie hustoty f kolem mediánu. K zamítnutí H_0 může tedy oprávněně dojít i tehdy, je-li medián roven x_0, ale hustota f je výrazně nesymetrická.

Je-li některá z veličin X_i rovna x_0, zpravidla se toto pozorování vynechává.

Znaménkový test

Nechť $X_1, \ldots, X_n$ je výběr ze spojitého rozdělení s mediánem $\tilde{x}$. Platí tedy

$$P(X_i < \tilde{x}) = P(X_i > \tilde{x}) = \frac{1}{2}, \qquad i = 1, \ldots, n.$$

Testujeme hypotézu $H_0 : \tilde{x} = x_0$, kde x_0 je dané číslo. Zabývejme se nejdřív oboustranným testem, kdy alternativou je $H_1 : \tilde{x} \neq x_0$. Utvoří se rozdíly $X_1 - x_0, \ldots, X_n - x_0$. Pokud je některý z těchto rozdílů s kladným znaménkem označme Y. Platí-li H_0, má Y binomické rozdělení $B_i(n, \frac{1}{2})$. Hypotézu H_0 zamítáme, bude-li Y blízké nule nebo blízké číslu n. Je-li n malé, používají se tabulky kritických hodnot k_1 a k_2 s vlastnostmi

$$P(Y \leq k_1) \leq \frac{\alpha}{2}, \qquad P(Y \geq k_2) \leq \frac{\alpha}{2}. \tag{8}$$

Přitom k_1 je největší a k_2 nejmenší z čísel, pro něž (8) platí. Vzhledem k symetrii rozdělení $B_i(n, \frac{1}{2})$ je $k_2 = n - k_1$.

Hypotézu H_0 tedy zamítneme, bude-li $Y \leq k_1$ nebo $Y \geq k_2$. Hladina testu je nejvýše rovna α. Obvykle je však značně menší než α, zejména při malých hodnotách n.

Zaveďme náhodné veličiny $\xi_1, \ldots, \xi_n$ tak, že $\xi_i = 0$ v případě $X_i - x_0 \leq 0$ a $\xi_i = 1$ v případě $X_i - x_0 > 0$. Tudiž $Y = \xi_1 + \ldots + \xi_n$. Protože $E\xi_i = \frac{1}{2}, var\ \xi_i = \frac{1}{4}$, podle centrální limitní věty má veličina

$$\frac{Y - \frac{n}{2}}{\sqrt{n}}$$

Asymptotické rozdělení $N(0, \frac{1}{4})$. Proto veličina

$$U = \frac{2Y - n}{\sqrt{n}} \tag{9}$$

Má asymptoticky rozdělení $N(0, 1)$. Hypotézu H_0 zamítneme, když $|U| \geq u(\frac{\alpha}{2})$. Hladina tohoto testu se s rostoucím n blíží číslu α. V praxi se tohoto postupu používá, je-li $n \geq 20$. Ekvivalentně se někdy místo toho používá modifikace, při níž se H_0 zamítá, když $U^2 \geq \chi_1^2(\alpha)$.

Další možností je využít transformací stabilizujících rozptyl, v první variantě můžeme použít veličinu

$$U1 = 2\sqrt{n}\left(arcsin\ \sqrt{\frac{Y}{n}} - arcsin\ \sqrt{\frac{1}{2}}\right)$$

A H_0 zamítáme, když $|U1| \geq u(\frac{\alpha}{2})$. Ve druhé variantě vypočítáváme

$$U2 = \sqrt{4n+2}\left(arcsin\sqrt{\frac{8Y+3}{8n+6}} - arcsin\sqrt{\frac{1}{2}}\right)$$

A H_0 zamítáme v případě $|U2| \geq u(\frac{\alpha}{2})$. Z testů založených na U, U_1 a U_2 lze doporučit první z nich, protože nejlépe zachovává hladinu chyby prvního druhu.

Znaménkový test používáme zejména v případě, kdy rozdělení veličin X_i je výrazně zešikmené. Jelikož tento test má poměrně malou sílu (pravděpodobnost chyby druhého druhu je ve srovnání s jinými testy dosti veliká), je žádoucí mít k dispozici vetší počet pozorování n.

Jestliže se z technických důvodů stane, že některé rozdíly $X_i - x_0$ jsou rovny nule (což teoreticky má nulovou pravděpodobnost, ale může k tomu dojít třeba vlivem zaokrouhlení chyb), pak se tyto hodnoty vynechají a o jejich počet se sníží číslo n. Jinak test proběhne beze změny.

Znaménkový test může být také jednostranný. Wilcoxonův test i znaménkový test řadíme mezi neparametrické testy, protože k jejich odvození nebylo nutné pro daný výběr specifikovat přesný typ rozdělení či dokonce jeho parametry.

2.7.2 Dvouvýběrové testy

Dvouvýběrové testy používáme pro testování dvou navzájem nezávislých náhodných výběrů. Jako příklad můžeme uvést porovnávání makroekonomických ukazatelů ve dvou různých zemích ve stejném období. (Pozor: Kdybychom chtěli porovnávat data v jedné zemi ve dvou různých obdobích, musíme užít testů párových.) Při rozhodování, který z dvouvýběrových testů použít, hraje opět klíčovou roli skutečnost, zda daná data pocházejí z nějakého známého rozdělení (v našem

případě normálního), či nikoliv. V závislosti na splnění či nesplnění podmínky normality dělíme testy na parametrické a neparametrické (tedy stejně jako u párového testování).

Dvouvýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_m$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F a nechť $Y_1, \ldots, Y_n$ je na něm nezávislý náhodný výběr ze spojitého rozdělení s distribuční funkcí G. Je třeba testovat hypotézu $H_0 : F = G$ proti alternativě $H_1 : F \neq G$.

Všech $m+n$ veličin $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ (tzv. sdružený výběr) uspořádáme vzestupně podle velikosti. Označme T_1 součet pořadí hodnot $X_1, \ldots, X_m$ a T_2 součet pořadí hodnot $Y_1, \ldots, Yn$. Je jasné, že

$$T_1 + T_2 = \frac{1}{2}(m+n)(m+n+1)$$

Nejdřív vyšetříme obecné vlastnosti testů tohoto typu. Položme pro stručnost $N = m+n$. Nechť R_i je pořadí i-té veličiny ze sdruženého výběru a nechť $a(i)$ je nějaká funkce definovaná pro $i = 1, \ldots, N$. Veličině

$$S = \sum_{i=1}^{N} c_i a(R_i)$$

se říká jednoduchá lineární pořadová statistika. Označme

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a(i), \qquad \bar{c} = \frac{1}{N}\sum_{i=1}^{N} c(i)$$

$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N}[a(i) - \bar{a}]^2, \qquad \sigma_c^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c})^2$$

Věta 2.5. *Platí-li H_0, pak*

$$ES = N\bar{a}\bar{c}, \qquad var\ S = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2$$

Důkaz. Platí-li H_0, pak R_i je náhodná veličina, která nabývá každé z hodnot $1, \ldots, N$ s pravděpodobnosti $1/N$. Proto

$$Ea(R_i) = \sum_{i=1}^{N} a(t)\frac{1}{N} = \bar{a}, \tag{10}$$

takže

$$ES = \sum_{i=1}^{N} c_i Ea(R_i) = \sum_{i=1}^{N} c_i \bar{a} = N\bar{a}\bar{c}.$$

Je-li $i \neq j$, pak za platnosti H_0 máme

$$P(R_i = s, R_j = t) = \frac{1}{N(N-1)} \quad \text{pro} \quad 1 \leq \mathrm{s} \neq \mathrm{t} \leq \mathrm{N}.$$

Užitím (10) dostaneme

$$var\ a(R_i = E[a(R_i) - Ea(R_i)]^2 = E[a(R_i) - \bar{a}]^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}[a(t) - \bar{a}]^2 = \sigma_a^2$$

Věta 2.6. *Platí-li H_0, pak*

$$ET_1 = \frac{1}{2}m(m+n+1), \qquad var\ T_1 = \frac{1}{12}mn(m+n+1).$$

Důkaz. Veličina T_1 se dostane z S, položíme-li $a(i) = i$ a

$$c_i = \begin{cases} 1, & \text{pro} \quad \mathrm{i} = 1, \ldots, \mathrm{m}, \\ 0, & \text{pro} \quad \mathrm{i} = \mathrm{m}+1, \ldots, \mathrm{m}+\mathrm{n}, \end{cases}$$

Položme znovu pro stručnost $N = m + n$. Postupně dostaneme

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2}, \quad \bar{c} = \frac{m}{N},$$

$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N} i^2 - \bar{a}^2 = \frac{1}{12}(N+1)(N-1),$$

$$\sigma_c^2 = \frac{1}{N}\sum_{i=1}^{N} c_i^2 - \bar{c}^2 = \frac{mn}{N^2}.$$

Podle věty 2.5 platí

$$ET_1 = N\bar{a}\bar{c} = \frac{1}{2}m(m+n+1),$$

$$var\ T_1 = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2 = \frac{1}{12}mn(m+n+1).$$

Místo veličiny T_1 se zpravidla používá veličina

$$U_1 = mn + \frac{1}{2}n(n+1) - T_1.$$

Testu založenému na U_1 se pak někdy říká Mannův-Whitneyův test. Zavede se dále označení

$$U_2 = mn + \frac{1}{2}n(n+1) - T_2.$$

Přitom platí $U_1 + U_2 = mn$. Pokud min (U_1, U_2) je menší nebo rovno tabelované kritické hodnotě uvedené tabulce, zamítá se nulová hypotéza. Přitom se označení výběrů volí tak, aby platilo $m \geq n$. Pro velká m a n se užije následující postup. Z věty 2.6 vyplývá, že

$$EU_1 = \frac{1}{2}mn, \qquad var\ U_1 = \frac{1}{12}mn(m+n+1). \tag{11}$$

Jelikož $U_2 = mn - U_1$, máme $EU_2 = EU_1, varU_2, varU_2 = varU_1$. Je dokázáno, že při $m \to \infty$ a $n \to \infty$ má veličina U_1 (i veličina T_1) asymptoticky normální rozdělení. Vypočte se tedy

$$U = \frac{U_1 - EU_1}{\sqrt{var\ U_1}}, \tag{12}$$

Přičemž se za EU_1 a $var\ U_1$ dosadí z (11). Pokud $\|U\| \geq u(\frac{\alpha}{2})$, zamítne se H_0 na hladině blížící se α. Test založený na (2.11) se dá užít už při $m > 10, n > 10$.
Ačkoliv je Wilcoxonův test formulován jako test proti obecné alternativě, je citlivý zejména na tzv. alternativu posunutí $H_1' : G(x) = F(x - \Delta), \Delta \neq 0$. Pro případné jiné alternativy, např. když se G liší od F spíše rozptylem nebo tvarem se raději doporučuje Kolmogorovů-Smirnovův test.

Dvovýběrový Kolmogorovův-Smirnovův test

Dříve, než přistoupíme k formulaci tohoto dvouvýběrového testu, uvedeme některá pomocná tvrzení.

Nechť $X_1, \ldots, X_m$ je náhodný výběr z rozdělení, které má distribuční funkci F. Budiž x dané reálné číslo. Zaveďme náhodné veličiny

$$\xi_i(x) = \begin{cases} 1, & \text{je} - \text{li } \mathrm{X_i} < \mathrm{x}, \\ 0, & \text{je} - \text{li } \mathrm{X_i} \geq \mathrm{x}, \end{cases}$$

Pro $i = 1, \ldots, m$. Položme

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} \xi_i(x). \tag{13}$$

Funkce $F_m(x)$ je empirická distribuční funkce. Při konkrétní realizaci výběru je totožná s empirickou distribuční funkcí, která byla zavedena. Ukážeme, že se s rostoucím m funkce $F_m(x)$ blíží skutečné distribuční funkci $F(x)$.

Věta 2.7. *Pro každé x platí*

$$F_m(x) \to F(x) \quad \text{skoro jistě pro} \quad \mathrm{m} \to \infty$$

Důkaz. Pro každé pevně zvolené x jsou veličiny $\xi_i(x)$ nezávislé a mají stejné rozdělení. Platí pro ně

$$P[\xi_i(x) = 1] = F(x), \qquad E\xi_i(x) = F(x).$$

Jelikož $F_m(x)$ je dána vzorcem (19), z Kolmogorovů věty ihned plyne dokazované tvrzení.

Dá se však dokázat ještě silnější tvrzení.

Věta 2.8. *(Glivenko). Označme $D_m = sup_x|F_m(x) - F(x)|$. Pak platí*

$$P\left(\lim_{m\to\infty} D_m\right) = 1.$$

Nyní již přejdeme k vlastnímu tématu tohoto odstavce. Nechť $X_1, \ldots, X_m$ je náhodný výběr z rozdělení se spojitou distribuční funkcí F a nechť $Y_1, \ldots, Y_m$ je na něm nezávislý náhodný výběr z rozdělení se spojitou distribuční funkcí G. Budeme se zabývat testem hypotézy $H_0 : F = G$ proti alternativě $H_1 : F \neq G$. Označíme F_m empirickou distribuční funkci prvního výběru a G_n druhého výběru. Z vět 2.7 a 2.8 vyplývá, že se funkce F_m a G_n při rostoucích m a n blíží

distribučním funkcím F a G.

Platí-li H_0, pak podle Glivenkovy věty $D_{m,n} \to 0$ skoro jistě při $m \to \infty, n \to \infty$. Přesnější výsledek, na němž se pak dá založit test, je popsán v následující větě.

Věta 2.9. *(Smirnov). Označme $M = mn/(m+n)$. Nechť*

$$K(\lambda) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}exp(-2k^2\lambda^2). \tag{14}$$

Pak pro každé λ platí

$$\lim_{m,n\to\infty} P(\sqrt{M}D_{m,n} < \lambda) = K(\lambda).$$

Rozdělení veličiny $D_{m,n}$ pro konečné hodnoty m, n je uvedeno v knize Hájek a Šidák (1967).

Funkce $K(\lambda)$ se aproximuje pomocí počátečních členů $1 - 2e^{-2\lambda^2}$ (viz Likeš a Laga 1978). Pak

$$P\left(D_{m,n} < \frac{\lambda}{\sqrt{M}}\right) \doteq 1 - 2e^{-2\lambda^2}. \tag{15}$$

Výraz na pravé straně je roven $1-\alpha$ pro $\lambda = \lambda_\alpha = \sqrt{\frac{1}{2}ln\frac{2}{\alpha}}$. Aproximativní kritická hodnota je tudíž

$$D^*_{m,n}(\alpha) = \frac{\lambda_\alpha}{\sqrt{M}} = \sqrt{\frac{1}{2M}ln\frac{2}{\alpha}}.$$

Praktické provedení Kolmogorovova-Smirnovova testu tedy spočívá v tom, že se z výběru $X_1, \ldots, X_m$ a $Y1, \ldots, Y_n$ vypočtou empirické distribuční funkce F_m a G_n a veličina $D_{m,n}$. Jsou-li čísla m a n malá, porovná se $D_{m,n}$ s přesnými kritickými hodnotami $D_{m,n}(\alpha)$. V případě větších hodnot m, n se využije věty 2.9. Položí se $\lambda_0 = \sqrt{M}\ D_{m,n}$ a vypočte se hodnota $K(\lambda_0)$. Pokud vyjde $K(\lambda_0) \geq 1-\alpha$, zamítne se H_0 na hladině, které se s rostoucími rozsahy výběru blíží číslu α. Přitom se při větších hodnotách m a n kritická hodnota pro veličinu $D_{m,n}$ obvykle aproximuje číslem $D^*_{m,n}(\alpha)$. Hypotéza H_0 se pak zamítá, když $D_{m,n} \geq D^*_{m,n}(\alpha)$.

Kolmogorovovův-Smirnovův test byl zobecněn na případ porovnání tří a více výběrů v článku Kiefer (1959). Kritické hodnoty tabelovali Wolf a Naus (1973). Viz též Domański (1990).

2.7.3 Porovnávání několika výběrů

Kruskal-Wallis (ův) test

Tento test je neparametrickou obdobou analýzy rozptylu jednoduchého třídění a je zobecněním dvouvýběrového Wilcoxonova testu. Bývá používán zejména tehdy, jde-li o výběry z rozdělení značně se lišících od normálního.

Nechť $Y_{i1}, \ldots, Y_{in_i}$ je výběr z nějakého rozdělení se spojitou distribuční funkcí $F_i, i = 1, \ldots, I$. Nechť všechny tyto výběry jsou na sobě nezávislé. Budeme testovat hypotézu

$$H_0 : F_1(x) = \cdots = F_I(x) \quad \text{pro všechna x}$$

Proti alternativě H_1, že H_0 neplatí. Všechny veličiny Y_{ij} dohromady vytvoří sdružený náhodný výběr o rozsahu $n = n_1 + \cdots + n_i$. Uspořádají se do rostoucí posloupnosti a určí se pořadí R_{ij} každé veličiny Y_{ij} ve sdružení výběru. Toto pořadí můžeme zapsat do schématu uvedeného v (Tabulka1)

Výběr	Pořadí veličin				Součet pořadí
1	R_{11}	R_{12}	$\ldots$	R_{1n_1}	T_1
2	R_{21}	R_{22}	$\ldots$	R_{2n_2}	T_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
I	R_{I1}	R_{I2}	$\ldots$	R_{In_i}	T_I

Tabulka 1: Pořadí sdruženého náhodného výběru

Celkový součet všech pořadí je $T_1 + \cdots + T_I = n(n+1)/2$. Jako testová statistika se použije

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

Nejprve vyšetříme obecný tvar statistiky tohoto typu. Přitom použijeme označení zavedené již u dvouvýběrového Wilcoxonova testu.

Věta 2.10. *Nechť $X_1, \ldots, X_N$ je náhodný výběr ze spojitého rozdělení. Nechť $g_1, \ldots, g_I$ je rozklad množiny $\{1, \ldots, N\}$ na disjunktní neprázdné podmnožiny. Nechť g_i má n_i prvků, $i = 1, \ldots, I$. Nechť R_i je pořadí $X_i, i = 1, \ldots, N$. Nechť $a(i)$ a c_i jsou daná čísla, $i = 1, \ldots, N$. Označíme*

$$\bar{a} = \frac{1}{N}\sum a(i), \qquad \sigma_a^2 = \frac{1}{N}\sum [a(i) - \bar{a}]^2.$$

Nechť

$$Q_j = \sum_{i \in g_j} a(R_i), \qquad Q = \frac{N-1}{N\sigma_a^2} \sum_{j=1}^{I} \frac{1}{n_j}(Q_j - EQ_j)^2.$$

Pak $EQ = I - 1$.

Důkaz. Máme

$$Q_j = \sum_{i=1}^{N} c_i(j) a(R_i),$$

kde $c_i(j) = 1$ pro $i \in g_j$ a $c_i(j) = 0$ pro $i \notin g_j$. Označme

$$\bar{c}(j) = \frac{1}{N}\sum_{i=1}^{N} c_i(j).$$

Podle věty 2.6 platí

$$var\ Q_j = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2,$$

kde

$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N}[c_i(j) - \bar{c}(j)]^2 = \frac{1}{N}\sum_{i=1}^{N} c_i^2(j) - [\bar{c}(j)]^2$$

$$= \frac{n_j}{N} - \left(\frac{n_j}{N}\right)^2 = \frac{n_j}{N} - \left(1 - \frac{n_j}{N}\right).$$

proto

$$EQ = \frac{N-1}{N\sigma_a^2}\sum_{j=1}^{I}\frac{1}{n_j} var\ Q_j$$

$$= \frac{N-1}{N\sigma_a^2}\sum_{j=1}^{I}\frac{1}{n_j}\frac{N^2}{N-1}\sigma_a^2\frac{n_j}{N}\left(1 - \frac{n_j}{N}\right)$$

$$\sum_{j=1}^{I}\left(1-\frac{n_j}{N}\right)=I-1$$

V případě Kruskalova-Wallisova testu je $a(i)=i, Q_j=T_j$ a $N=n$. Dále jsme vypočítali, že

$$\sigma_a^2=\frac{1}{12}(N+1)(N-1).$$

Podle věty 2.6 dále máme

$$EQ_j=N\bar{a}\bar{c}(j)=N\frac{N+1}{2}\frac{n_j}{N}=\frac{1}{2}(N+1)n_j,$$

takže

$$\begin{aligned}Q&=\frac{12(N-1)}{N(N+1)(N-1)}\sum_{j=1}^{I}\frac{1}{n_j}\left[T_j-\frac{1}{2}n_j(N+1)\right]^2\\&=\frac{12}{N(N+1)}\left[\sum_{j=1}^{I}\frac{T_j^2}{n_j}-\frac{1}{4}N(N+1)\right]\\&=\frac{12}{N(N+1)}\sum_{j=1}^{I}\frac{T_j^2}{n_j}-3(N+1)\end{aligned}$$

Položíme-li ještě $N=n$, dostaneme tím výraz uvedený na začátku.

Místo Q se někdy užívá označení H a mluví se pak o H testu. Je-li v datech více než 25% shod, používá se korigovaná statistika

$$Q_{korig}=\frac{Q}{1-(n^3-n)^{-1}\sum(t_j^3-t_i)'}$$

Kde $t_1, t_2, \ldots$ jsou počty shodných pozorovaní v jednotlivých skupinách veličin majících tutéž hodnotu.
Dá se dokázat (viz Hájek a Šidák 1967), že za platnosti H_0 má Q asymptotický χ^2 rozdělení, když všechna ni rostou nade všechny meze. Jelikož jsme dokázali, že $EQ=I-1$, půjde o asymptotické χ^2 rozdělení mající $I-1$ stupňů volnosti. Proto hypotézu H_0 zamítáme, když $Q\geq\chi^2_{I-1}(\alpha)$.

Kruskalův-Wallisův test je citlivý zejména na ty případy, kdy se jednotlivé distribuční funkce od sebe liší posunutím. Zamítne-li se H_0, je třeba obvykle rozhodnout, které dvojice výběrů se od sebe významně liší. U analýzy rozptylu

byla k tomu účelu použita Tukeyova metoda. U Kruskalova-Wallisova testu se postupuje následovně (viz Miller 1966). Označme $t_i = T_i/n_i, i = 1, \ldots, I$. Nechť $h_{I-1}(\alpha)$ je kritická hodnota Kruskalova-Wallisova testu na hladině α. Při malých rozsazích výběrů se $h_{I-1}(alpha)$ najde ve speciálních tabulkách a při větších rozsazích se použije výše zmíněné aproximace $h_{I-1}(\alpha) \doteq \chi^2_{I-1}(\alpha)$. Prohlásíme, že se distribuční funkce i-tého a j-tého výběru od sebe signifikantně liší, jakmile platí

$$|t_i - t_j| > \sqrt{\frac{1}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) n(n+1)h_{I-1}(\alpha)}. \tag{16}$$

Pravděpodobnost, že alespoň u jedné z $I(I-1)/2$ dvojic distribučních funkcí F_i, F_j bude vypočteno, že se F_i a F_j signifikantně liší, ačkoli ve skutečnosti platí hypotéza H_0, přitom nepřekročí α.

Je-li rozsah všech výběrů stejný, řekněme $n_1 = \cdots = n_I = m$ (takže jde o vyvážené třídění), lze použít Neményiovy metody založené na Tukeyově myšlence uplatněné již při analýze rozptylu (viz Neményi 1963 a Miller 1966). Pro menší hodnoty m a I jsou kritické hodnoty pro $|T_i - T_j|$ uvedeny v tabulce T15. Při větších hodnotách se užije následující postup. Nechť $q_i, \infty(\alpha)$ je kritická hodnota rozpětí I nezávislých náhodných veličin s rozdělením $N(0,1)$. Najde se v posledním řádku tabulky T11, resp. T12, a zavádí se takto. Je-li $\xi_1, \ldots, \xi_I$ je výběr z $N(0,1)$, označíme $R = \xi_{(I)} - \xi_{(1)}$ jeho rozpětí. Pak $q_1, \infty(\alpha)$ je číslo definované podmínkou

$$P[R \geq q_I, \infty(\alpha)] = \alpha.$$

Prohlásíme, že se F_i a F_j od sebe liší, když

$$|t_i - t_j| > q_I, \infty(\alpha)\sqrt{\frac{1}{12}I(Im+1)}. \tag{17}$$

Ačkoliv při vyváženém třídění lze užít obou zde uvedených metod, dává se přednost Neményiově metodě, protože je citlivější.

Další testy, které lze užít místo Kruskalova-Wallisova testu, popisují Bhapkar a Desphande (1968).

Je třeba upozornit na to, že přiřazování pořadí náhodným veličinám je transformace sice monotónní, ale nelineární. Právě tato nelinearita může někdy vést k paradoxním výsledkům.

Friedmanův test

Nechť Y_{ij} jsou nezávislé náhodné veličiny se spojitými distribučními funkcemi F_{ij} pro $i = 1, \ldots, J$. Fridmanovým testem se testuje hypotéza H_0, že F_{ij} nezávisí na j (zatímco na i záviset může).

Pro každé i zvlášť se určí pořadí R_{ij} veličiny Y_{ij}. Jde tedy jen o určení pořadí mezi veličinami $Y_{i1}, \ldots, Y_{ij}$. Za platnosti H_0 je splněna podmínka

H_0' : pro každé i je vektor $(R_{i1}, \ldots, R_{ij})'$ roven kterékoli permutaci čísel $1, \ldots, J$ se stejnou pravděpodobností $1/J!$ a všechny vektory $(R_{i1}, \ldots, R_{ij})'$ pro $i = 1, \ldots, I$ jsou na sobě nezávislé.

Protože všechny vlastnosti Friedmanova testu jsou odvozovány pouze z předpokladu platnosti H_0', lze často tento test použít za obecnějších podmínek, než vyplývá z jeho původní formulace. Teoretický tvar statistiky Friedmanova testu je

$$Q = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} \left[\sum_{i=1}^{I} R_{ij} - \frac{1}{2} I(J+1) \right]^2 . \tag{18}$$

Věta 2.11. *Platí-li H_0', pak $EQ = J - 1$*

Věta 2.12. *Platí*

$$Q = \frac{12}{IJ(J+1)} \sum_{j=1}^{J} \left(\sum_{i=1}^{I} R_{ij} \right)^2 - 3I(J+1). \tag{19}$$

Důkaz. Dostane se úpravou vzorce (2.11). Výpočty se provádějí podle vzorce (19). Hypotézu H_0' (a tedy také hypotézu H_0) zamítáme, když Q překročí kritickou hodnotu na hladině α. Při větších hodnotách I se za tuto kritickou hodnotu bere $\chi^2_{J-1}(\alpha)$. Podrobně o kritických hodnotách a jejich aproximacích pojednává

Michaelis (1971).

Zamítneme-li H_0, zajímá nás, pro které dvojice j a t se distribuční funkce F_{ij} a F_{it} od sebe významně liší. Označme

$$R_{.j} = \sum_{i=1}^{I} R_{ij}.$$

Jakmile $|R_{.j} - R_{.t}|$ je větší nebo rovno tabelované kritické hodnotě, zamítne se rovnost $F_{ij} = F_{it}$. Tato porovnání se dělají pro všechny dvojice $j < t$. Asymptoticky jsou kritické hodnoty pro tato mnohonásobná porovnání rovny

$$q_{J,\infty}(\alpha)\sqrt{\frac{1}{12}IJ(J+1)}. \tag{20}$$

Kritické hodnoty $q_{J,\infty}(\alpha)$ byly definovány v odstavci (Kruskal-Wallisuv test). Vzorec (20) se používá už při $I > 5$.

Někdy se místo Friedmenova testu užívá Andersonův-Kannemannův test , který nyní popíšeme.

Friedmanův test odpovídá situaci, kdy na každém z I objektů (resp. Bloků) je aplikováno J ošetření. Leckdy nejde o ošetření v pravém slova smyslu, ale prostě je nějaká veličina na každém sledovaném objektu zaznamenávána v J časových okamžicích. Opět označme R_{ij} pořadí, které je připsáno j-tému ošetření na i-tém bloku. Nechť D_{jm} je počet bloků, ve kterých ošetření j dostalo pořadí m. Matice

$$\mathbf{D} = \begin{pmatrix} D_{11} & \dots & D_{1J} \\ \dots & \dots & \dots \\ D_{J1} & \dots & D_{IJ} \end{pmatrix}$$

se nazývá incidenční. Označme

$$\chi^2_{AK} = \frac{J-1}{I}\sum_{j=1}^{J}\sum_{m=1}^{J}\left(D_{jm} - \frac{I}{J}\right)^2$$

Platí-li hypotéza H_0, pak χ^2_{AK} má asymptotický rozdělení $\chi^2_{(J-1)^2}$. V případě

$$\chi^2_{AK} \geq \chi^2_{(J-1)^2}(\alpha)$$

se H_0 zamítne. Je nutné upozornit na to, že D není kontingenční tabulka, a proto limitní rozdělení veličiny χ^2_{AK} muselo být v citované literatuře odvozeno jiným způsobem. Andersonův-Kannemannův test je citlivější proti vetší třídě alternativ než Friedmanův test.

Profilová analýza

Ve Friedmanově testu bylo popsáno neparametrické hodnocení jedné skupiny objektů, kde na každé z nich byla v J různých časových okamžicích měřena jistá veličina. Testovala se hypotéza, že se úroveň této veličiny v čase nemění. Nyní se budeme zabývat případem, kdy jde o několik skupin objektů. Půjde o to, zda případné změny úrovně sledované veličiny probíhají ve všech skupinách stejně.

Mějme tedy K skupiny objektů. Do k-té skupiny nechť je přitom zařazeno n_k objektů, $k = 1, \ldots, K$. Celkový počet objektů je tedy $N = n_1 + \cdots + n_K$. Sledujeme náhodné veličiny

$$X_{kit}, \quad k = 1, \ldots, K; \quad i = 1, \ldots, n_k; \quad t = 1, \ldots, T.$$

Předpokládejme, že pro tyto veličiny platí

$$X_{kit} = \mu_k + m_{ki} + \gamma_t + \delta_{kt} + e_{kit},$$

kde μ_k jsou pevné efekty skupin, m_{ki} jsou náhodné individuální efekty, γ_t jsou pevné efekty času a δ_{kt} jsou interakce skupin času. Nechť náhodné vektory

$$e_{ki} = (e_{ki1}, \ldots, e_{kiT})'$$

jsou nezávislé a mají stejné rozdělení s nulovou střední hodnotou a regulární variační maticí.
Nejprve budeme testovat hypotézu $H_0 : \delta_{kt} = 0$ pro všechna k a t. Platí-li H_0, pak střední hodnoty vektorů

$$K_i = (X_{ki1}, \ldots, X_{kiT})'$$

Liší jen posunutím. Říkáme, že vektor X_{ki} mají v tomto případě paralelní profily. Jako u Friedmanova testu se pro každou dvojici (k, i) zvlášť vytvoří posloupnost pořadí

$$R_{ki1}, \ldots, R_{kiT}.$$

Označme

$$R_{k.t} = \sum_{i=1}^{n_k} R_{kit}.$$

V práci Lehmacher (1978) byly k testování hypotézy H_0 navrženy veličiny

$$V_t = \frac{(N-1)\sum_{k=1}^{K} \frac{1}{n_k}\left(R_{k.t} - \frac{n_k}{N}\sum_{j=1}^{K} R_{j.t}\right)^2}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_{kit}^2 - \frac{1}{N}\left(\sum_{k=1}^{K} R_{k.t}\right)^2}, \quad t = 1, \ldots, T,$$

a bylo dokázáno, že za platnosti H_0 má každá z nich asymptotický χ^2_{K-1} rozdělení. Hypotézu H_0 zamítáme, jestliže alespoň pro jedno t platí

$$Vt \geq \chi^2_{K-1}\left(\frac{\alpha}{T}\right). \tag{21}$$

Pro ty okamžiky t, pro něž je splněna nerovnost (21), je současně statisticky prokázáno, že se v nich tvar profilů liší.

Je-li $K = 2$, lze postup trochu zjednodušit. Vypočtou se veličin

$$V_t^* = \sqrt{\frac{N-1}{n_1 n_2 N}} \frac{n_2 R_{1.t} - n1 R_{2.t}}{\sqrt{\sum_{k=1}^{2}\sum_{i=1}^{n_k} R_{kit}^2 - \frac{1}{N}(R_{1.t} + R_{2.t})^2}}, \quad t = 1, \ldots, T,$$

které mají za platnosti H_0 asymptoticky normální rozdělení. Tvary profilů se signifikantně liší v těch okamžicích t, pro něž platí

$$|V_t^*| \geq u\frac{\alpha}{2T}.$$

Platí-li tato nerovnost aspoň pro jedno t, zamítá se H_0. Tento test je konzervativní, neboť jeho asymptotická hladina je menší než α. Citlivější neparametrický test je popsán v článku Lehmacher (1979), ale ten zas neumožňuje stanovit ty

okamžiky t, v nichž se tvar profilů liší.

Žádný z těchto testů není pochopitelně citlivý vůči takovým rozdílům v profilech, které jsou eliminovány výpočtem pořadí. Nenajde se tedy rozdíl, jestliže ve všech K skupinách křivky rostou a nejsou rovnoběžné v geometrickém slova smyslu. Jiná situace v profilové analýze nastává tehdy, máme-li sice objekty zařazeny jen do jedné skupiny, ale zato se u každého objektu provádí měření v T okamžicích při jednom typu ošetření a pak po odeznění v analogických T okamžicích při jiném typu ošetření. Sledují se tedy náhodné veličiny

$$X_{ikt} = \mu + a_i + b_k + \gamma_t + (\beta\gamma)_{kt} + e_{ikt},$$

kde $i = 1, \ldots, I, k = 1, 2, t = 1, \ldots, T$. Přitom a_i jsou náhodné individuální efekty, b_k jsou náhodné efekty ošetření, γ_t jsou pevné efekty času a $(\beta\gamma)_{kt}$ jsou pevné interakce času a ošetření. Předpokládá se, že náhodné vektory

$$e_i = (e_{i11}, \ldots, e_{i1T}, e_{i21}, \ldots, e_{i2T})'$$

jsou nezávislé a mají stejné rozdělení s nulovou střední hodnotou a regulární variančí maticí. V úvahu přicházejí dvě možné nulové hypotézy. První je

$$H_0 : (\beta\gamma)_{kt} = 0 \qquad \text{pro všechna k a T.}$$

Tato hypotéza vyjadřuje homogenitu profilů, čili rovnoběžnost křivek. Přitom se počítá s možným efektem ošetření (např. při prvním ošetření mohou být křivky celkově položeny výš než při druhém ošetření).
Druhá možná nulová hypotéza je

$$H_0' : b_1 = 0, \quad b_2 = 0, \quad (\beta\gamma)_{kt} = 0 \;\; \text{pro všechna k a T.}$$

V tomto případě jde o homogenitu křivek. K rovnoběžnosti křivek přistupuje ještě to, že jejich hladina nezávisí na ošetření.

Testy uvedených hypotéz jsou popsány v článku Lehmacher (1980). V případě testu H_0 se i-té dvojici křivek přiřadí pořadí $R_{i1}, \ldots, R_{iT}$ (pro první křivku) a pořadí $Q_{i1}, \ldots, Q_{it}$ (pro druhou křivku). Názorně je to ukázáno v (Tabulka2).

	1.ošetření			2.ošetření		
Objekty	1	...	T	1	...	T
1	R_{11}	...	R_{1T}	Q_{11}	...	Q_{1T}
...	...	...	...	...	...	...
I	R_{I1}	...	R_{IT}	Q_{I1}	...	Q_{IT}

Tabulka 2: Pořadí u dvou typu ošetření

Označme $D_{it} = R_{it} - Q_{it}$. Platí-li H_0, pak každá z veličin

$$S_t = \frac{\sum_{i=1}^{I} D_{it}}{\sqrt{\sum_{i=1}^{I} D_{it}^2}}, \qquad t = 1, \ldots, T,$$

má asymptoticky rozdělení $N(0,1)$. Hypotézu H_0 zamítneme, platí-li alespoň pro jedno t nerovnost

$$|S_t| \geq u\left(\frac{\alpha}{2T}\right).$$

Přitom tvary profilů se signifikantně liší právě v těch okamžicích t, pro něž tato nerovnost platí.

Při testu H_0' z každé dvojice křivek vypočteme pořadí $R_{i1}, \ldots, R_{i,2T}$ jakožto ze spojeného výběru. Viz (Tabulka3)

	1.ošetření			2.ošetření		
Objekty	1	...	T	1	...	T
1	R_{11}	...	R_{1T}	$R_{1,T+1}$	...	$R_{1,2T}$
...	...	...	...	...	...	...
I	R_{I1}	...	R_{IT}	$R_{I,T+1}$	...	$R_{I,2T}$

Tabulka 3: Spojené pořadí u dvou typu ošetření

Označme $D_{it}^* = R_{it} - R_{i,T+t}$. Platí-li H_0' , pak každá z veličin

$$S_t^* = \frac{\sum_{i=1}^{I} D_i^* t}{\sqrt{\sum_{i=1}^{I} D^{*2n_i t}}}, \qquad t = 1, \ldots, T,$$

má asymptoticky rozdělení $N(0,1)$. Platí-li

$$|S_t^*| \geq u\left(\frac{\alpha}{2T}\right)$$

alespoň pro jedno t, zamítneme H_0'. Tyto okamžiky t současně indikují, kdy pozorování vedou k zamítnutí H_0'. Tento test je rovněž konzervativní.

2.8 Softwary podporující neparametrické testy

2.8.1 R-projekt

R podle [9] je jazykem a prostředím pro statistické výpočty a grafiku. Poskytuje nepřeberné množství statistických a grafických technik s možností o rozšíření dalších metod. R je dostupný, jako volně šiřitelný software (Free Software) při dodržení podmínek GNU General Public License nadace Free Software Foundation, což může představovat výraznou výhodu proti běžně dostupným komerčním softwarovým nástrojům pro analýzu dat a statistické výpočty, zejména vzhledem k možnostem modifikace programu a jeho další distribuce a dostupnosti zdrojového kódu. R běží pod celou řadou UNIXových platforem a dále pod operačními systémy Windows a MacOS.

R obsahuje veškeré typy neparametrických testů pod zadanými příkazy. Jednoduchými příkazy lze tak rychle provést statistické vyhodnocení. K využívání prostředí R je potřebná alespoň základní technická vybavenost uživatele - požadavek je třeba vyjádřit syntakticky správně v jazyce R, prostředí R stále nevyužívá grafické uživatelské rozhraní.

2.8.2 Microsoft Excel

Microsoft Excel je dostupný dle [10] a [11] v rámci kancelářského balíku Microsoft Office (verze 2007). Program nabízí přehledné tabulkové prostředí, jež je ovládán přes jednoduché menu. Obsahuje základní pro popisnou statistiku a testování hypotéz. Funkcemi pokrývá většinu parametrických testů, ale již neobsahuje mnoho funkcí pro neparametrické testy. Pro obsáhlejší neparametrické testy je proto třeba si zakoupit nadstavbu UNISTAT, který obsahuje veškeré základní i pokročilé statistické metody.

2.8.3 Matlab

Matlab podle [8] je numerické výpočetní prostředí a také i programovacím jazykem vydaným společností The MathWorks. Je ovládán stejně jako R pomocí zadávání příkazů do konzolového prostředí. Základní strukturou systému jsou matice, pro něž je celý sytém optimalizován. Samotný systém obsahuje jen základní funkce. Veškeré specializované metody jsou obsaženy v tak zvaných toolboxech. Pro oblast statistiky je třeba Statistitcs Toolbos, který obsahuje veškeré funkce pro statistické testování. Tento toolbox nabízí funkce pro většinu parametrických i neparametrických metod.

3 PRAKTICKÁ ČÁST

V praktické části jsme se zaměřili na postupy, které byly použity při zpracovávání dat. Dále zde uvádíme postup zpracování vybraných neparametrických metod a test normalit. K testování byly poskytnuty konkrétní datasety katedrou geoinformatiky z průzkumu jízdních dokladů veřejné linkové dopravy . Pro zpracování datasetu bylo využito programu Microsoft Excel 2007 a pro statistické výpočty jsme používali statistický software R.

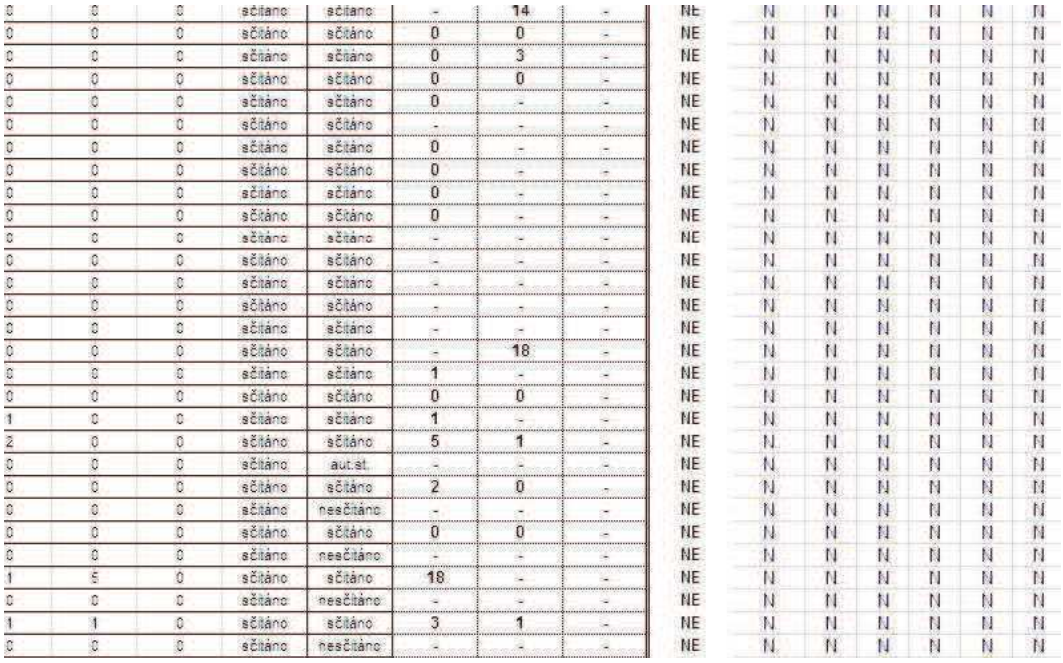
3.1 Zpracování a oprava poskytnutých dat

Průzkum veřejné linkové dopravy jízdních dokladů probíhal v obvodu a přilehlé okolí) dle členění v rámci Integrovaného dopravního systému kraje, a to ve dnech 13.4.2010 až 15.4.2010. Sčítání bylo prováděno studenty katedry geoinformatiky do předem připravených papírových formulářů. Do formulářů se zaznamenávalo číslo linky a jejího spoje, den sčítání a čas výjezdu spoje, počet nastupujících a vystupujících osob a celkový počet pasažérů v daném spoji a nakonec počet a typ křížových jízdních dokladů využívaných pro přepravu od jiného dopravce.Následně proběhlo zpracování formulářů do elektronické podoby v programu Microsoft Excel.

Data byla sbírána z terénního průzkumu sčítači, tudíž se musí počítat i s chybnými záznamy a v některých případech docházelo i k neúplnému záznamu požadovaných dat při statistickém průzkumu. Proto bylo třeba tyto data upravit nebo úplně tyto data odfiltrovat z příslušného datasetu.

Tato importovaná data (obr.1) obsahují již zmíněné chyby měření. Bylo třeba doplnit některé prázdné záznamy o pomlky nebo v případě nezměřeného spoje úplně daný záznam odfiltrovat.

Dále je zde v (obr.2) ukázany naměřené hodnoty počtu nastupujících a vystupujících cestujících u dopravců .



0	0	0	sčítáno	sčítáno	-	14	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	0	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	3	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	0	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	-	18	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	1	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	0	-	NE	N	N	N	N	N	N
1	0	0	sčítáno	sčítáno	1	-	-	NE	N	N	N	N	N	N
2	0	0	sčítáno	sčítáno	5	1	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	aut.st.	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	2	0	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	nesčítáno	-	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	sčítáno	0	0	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	nesčítáno	-	-	-	NE	N	N	N	N	N	N
1	5	0	sčítáno	sčítáno	18	-	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	nesčítáno	-	-	-	NE	N	N	N	N	N	N
1	1	0	sčítáno	sčítáno	3	1	-	NE	N	N	N	N	N	N
0	0	0	sčítáno	nesčítáno	-	-	-	NE	N	N	N	N	N	N

Obrázek 1: Ukázka primárních dat

3.2 Statistická analýza

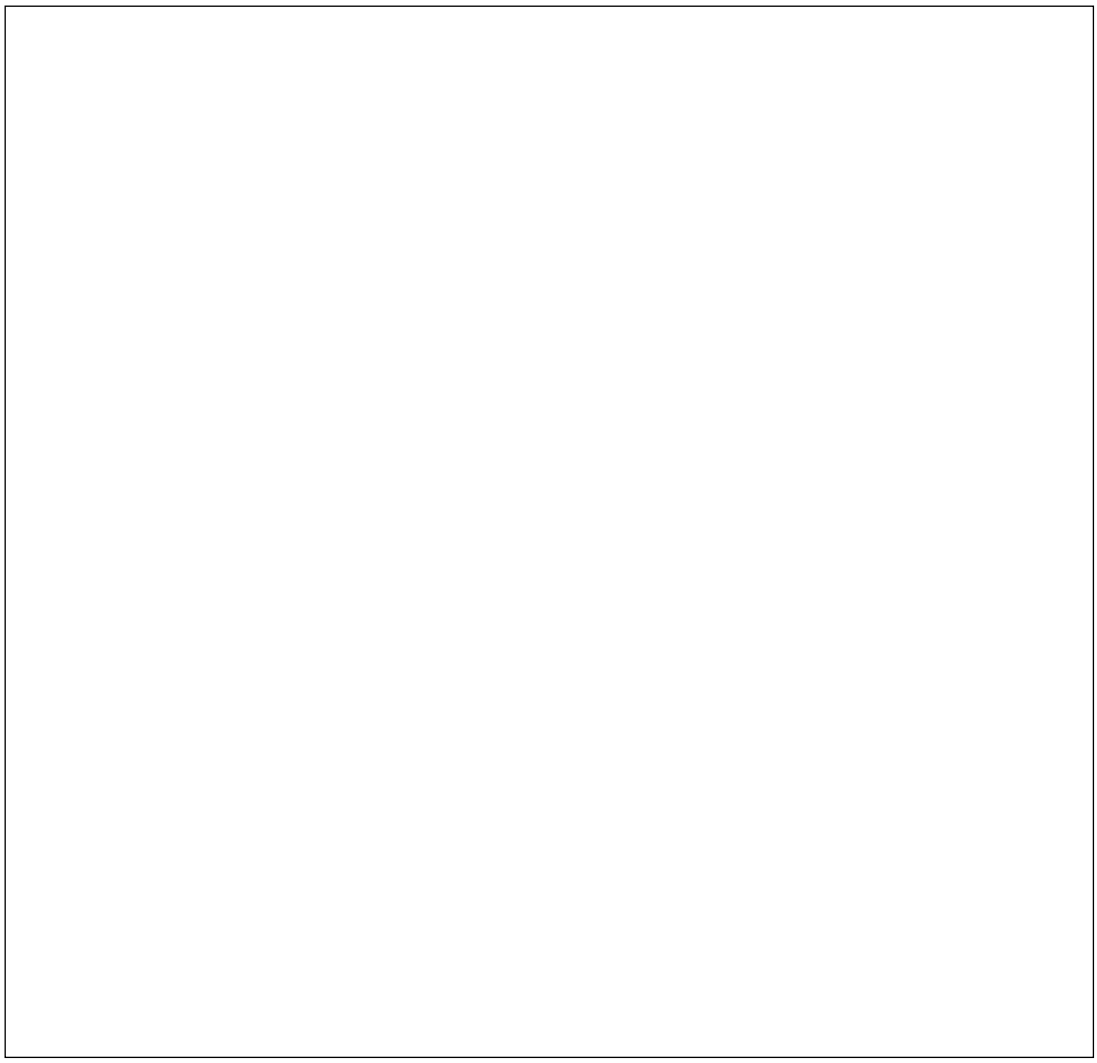
Pro veškeré statistické výpočty, bylo využito statistického programu R, který je silným nástrojem pro jakékoliv statistické výpočty a je na bázi freeware platformy.

3.2.1 Test normality

Pro Shapiro-Wilk test normality slouží v R funkce shapiro.test (). Funkce má jeden argument, kterým je numerický vektor. Výsledkem je hodnota W, která je ve své podstatě korelační koeficient, který nám říká, jak těsně naše data korelují s křivkou normálního rozdělení a p-value pak udává, jaké chyby se dopouštíme, pokud zamítneme nulovou hypotézu H_0.

a) Test normality pro sběr dat veřejné linkové dopravy [ ] nástup (pro 3 dny).

Shapiro-Wilk normality test



Obrázek 2: Ukázka naměřených hodnot

Data:
$W = 0.765, p - value = 9.356e - 09$

Shapiro-Wilk test naznačuje, že rozdělení hrubého skóru v tomto testu není v počtu nastupujících cestujících normální ($p = 9.356e - 09$). Pro relativně nízkou hodnotu p můžeme použít neparametrického testování hypotéz.

b) Test normality pro sběr dat veřejné linkové dopravy výstup (pro 3 dny).

Shapiro-Wilk normality test

Data: [redacted] $W = 0.3583, p - value = 7.362e - 14$

Shapiro-Wilk test naznačuje, že rozdělení hrubého skóru v tomto testu není v počtu vystupujících cestujících normální ($p = 7.362e - 14$). Proto pro relativně nízkou hodnotu p můžeme použít neparametrického testování hypotéz.

c) Test normality pro sběr dat veřejné linkové dopravy [redacted] nástup (pro 3 dny).

Shapiro-Wilk normality test

data: [redacted]
$W = 0.6319, p - value < 2.2e - 16$

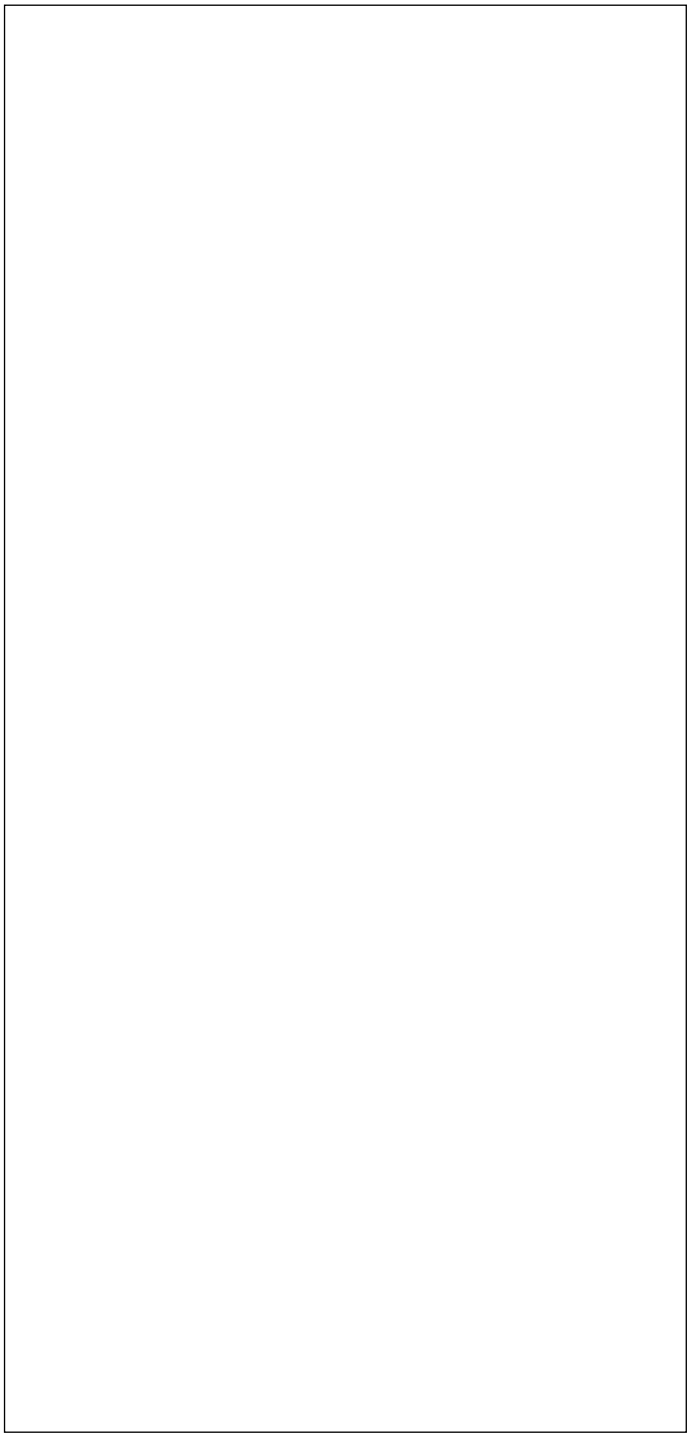
Shapiro-Wilk test naznačuje, že rozdělení hrubého skóru v tomto testu není v počtu nastupujících cestujících normální ($p = 2.2e - 16$). A proto pro relativně nízkou hodnotu p můžeme použít neparametrického testování hypotéz.

d) Test normality pro sběr dat veřejné linkové dopravy [redacted] výstup (pro 3 dny).

Shapiro-Wilk normality test

data: [redacted]
$W = 0.4633, p - value < 2.2e - 16$

Shapiro-Wilk test naznačuje, že rozdělení hrubého skóru v tomto testu není v počtu nastupujících cestujících normální ($p = 2.2e - 16$). Pro relativně nízkou hodnotu p -value lze použít neparametrické testování.



Obrázek 3: Výsledky testu normalit

3.2.2 Dvouvýběrový Wilcoxonův test

Pro dvouvýběrový Wilcoxonův test slouží v R funkce wilcox.test (). Funkce je jen použitelné pro dva testované vzorky. Jedná se o neparametrickou obdobu párového t-testu, srovnává tedy přes mediány a pořadí pozorování, nikoliv přes průměry. Je také známý pod označením ' Mann - Whitney ' test.

U dat bylo nejprve provedeno otestování na test normalit Kruskal-Wallisovým testem. Jelikož p-hodnota testovaných dat byla menší než $0,05$, zamítáme hypotézu o normalitě rozdělení dat a můžeme tedy přejít k samotnému dvouvýběrovému Wilcoxonovu test.

a) Dvouvýběrový Wilcoxon test pro sběr dat veřejné linkové dopravy [ ] [ ] pro nástup a výstup u autobusového dopravce [ ].

Wilcoxon rank sum test with continuity correction

Data: [ ]
$W = 2334, p - value = 0.0001500$ alternative hypothesis: true location shift is not equal to 0

Z výsledku vyplývá, že p-hodnota je nižší než je hladina významnosti (5%) a můžeme tedy zamítnout nulovou hypotézu říkající, že počet nastupujících cestujících je stejný jako počet vystupujících cestujících.

b) Dvouvýběrový Wilcoxon test pro sběr dat veřejné linkové dopravy [ ] [ ], pro nástup a výstup u autobusového dopravce [ ].

Wilcoxon rank sum test with continuity correction

data: [ ]
$W = 1058407, p - value < 2.2e - 16$ alternative hypothesis: true location shift is not equal to 0

Díky většímu počtu záznamů u daného dopravce můžeme výsledek lépe zhodnotit a potvrdit tak stejnou hypotézu jako v předešlém příkladu.

c) Dvouvýběrový Wilcoxon test pro sběr dat veřejné linkové dopravy
pro křížové využití jízdních dokladů u dopravce .

Wilcoxon rank sum test with continuity correction

data:
$W = 4629, p - value = 0.8133$
alternative hypothesis: true location shift is not equal to 0

Protože p-hodnota ukázala vyšší hodnotu, než je udána hladina významnosti, nelze tak zamítnout nulovou hypotézu. Počet cestujících využívajících křížových jízdních dokladů od obou dopravců je stejný. Z toho vyplívá, že pro přepravu v určité zóně ve spojích společnosti cestující využívají křížových jízdních dokladů od dopravce přibližně stejně jako od firmy .

d) Dvouvýběrový Wilcoxon test pro sběr dat veřejné linkové dopravy
, pro křížové využití jízdních dokladů u dopravce .

Wilcoxon rank sum test with continuity correction

Data:
$W = 4518020, p - value < 2.2e - 16$
alternative hypothesis: true location shift is not equal to 0

Z výsledku vyplývá, že p-hodnota je mnohem nižší, než je hladina významnosti (5%). Lze tak zamítnout nulovou hypotézu, která tvrdí, že počet cestujících využívajících křížových jízdních dokladů je stejný u obou testovaných dopravců.

3.2.3 Kruskal-Wallis(ův) test

Jedná se o neparametrickou obdobu jednocestné ANOVy a používá se pro porovnávání více výběrů.

a) Kruskal-Wallis(ův) test pro sběr dat veřejné linkové dopravy [redacted], pro nástup ve 3 období (dopoledne, odpoledne a večer) u autobusového dopravce [redacted].

Kruskal-Wallis rank sum test

data: [redacted]
$Kruskal - Wallischi - squared = 7.3122, df = 2, p - value = 0.02583$

Podle konečné hodnoty parametru p výsledek zamítáme a lze tvrdit, že počet nastupujících cestujících v jednotlivých denních dobách jsou rozdílné.

b) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [redacted], pro výstup ve 3 období (dopoledne, odpoledne a večer) u autobusového dopravce [redacted].

Kruskal-Wallis rank sum test

data: [redacted]
$Kruskal - Wallischi - squared = 14.9051, df = 2, p - value = 0.00058$

Podle hodnoty parametru p výsledek hypotézy zamítáme. Jednotlivé hodnoty vystupujících cestujících v obdobích dopoledne, odpoledne a večer jsou rozdílné tak, jak jsme dokázali v předchozím testu.

c) Kruskal-Wallis(ův) test pro sběr dat veřejné linkové dopravy [redacted], pro nástup ve 3 období (dopoledne, odpoledne a večer) u autobusového dopravce [redacted].

Kruskal-Wallis rank sum test

data: [redacted]
$Kruskal - Wallischi - squared = 4.8784, df = 2, p - value = 0.08723$

Výsledná hodnota parametru p byla získána vyšší než je hladina významnosti $0,05$ a nelze tak zamítnout tvrzení, že jednotlivé naměřené hodnoty počtu nastupujících cestujících v dopolední, odpolední a večerní době jsou obdobného typu.

d) Kruskal-Wallis(ův) test pro sběr dat veřejné linkové dopravy [redacted], pro výstup ve 3 období (dopoledne, odpoledne a večer) u autobusového dopravce [redacted].

Kruskal-Wallis rank sum test

data: [redacted]
$Kruskal - Wallischi - squared = 0.3658, df = 2, p - value = 0.8328$

Z hodnoty parametru p výsledek nelze zamítnout a tvrdíme, že jednotlivé naměřené hodnoty počtu vystupujících cestujících v denních dobách dopoledne, odpoledne a večer jsou podobné.

e) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [redacted], pro nástup ve 3 dnech (úterý, středa, čtvrtek) u autobusového dopravce [redacted].

Kruskal-Wallis rank sum test

data: [redacted]
$Kruskal - Wallischi - squared = 31.999, df = 2, p - value = 1.126e - 07$

Z hodnoty parametru p výsledek lze zamítnout a můžeme tvrdit, že jednotlivé hodnoty počtu nastupujících cestujících ve dnech úterý, středa a čtvrtek jsou

rozdílné.

f) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ], pro výstup ve 3 dnech (úterý, středa, čtvrtek) u autobusového dopravce [ ].

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 17.5363, df = 2, p - value = 0.0001556$

Podle hodnoty parametru p výsledek lze zamítnout a můžeme tak tvrdit, že jednotlivé naměřené hodnoty vystupujících osob ve dnech úterý, středa a čtvrtek jsou rozdílné.

g) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ], pro nástup ve 3 dnech (úterý, středa, čtvrtek) u autobusového dopravce [ ].

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 1.0323, df = 2, p - value = 0.5968$

Podle hodnoty parametru p se výsledek pohybuje na hladině významnosti $0,05$ a můžeme říci, že jednotlivé hodnoty naměřené při výstupu ve dnech úterý, středa a čtvrtek jsou podobné.

h) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ], pro výstup ve 3 dnech (úterý, středa, čtvrtek) u autobusového dopravce [ ].

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 1.4839, df = 2, p - value = 0.4762$

Podle hodnoty parametru p výsledek zamítáme a tvrdíme, že jednotlivé hodnoty počtu vystupujících cestujících ve dnech úterý, středa a čtvrtek jsou téměř identické.

i) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ], pro celkové křížové využití jízdních dokladů ve 3 období (dopoledne, odpoledne, večer) u autobusového dopravce [ ]

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 61.2047, df = 2, p - value = 5.123e - 14$

Z hodnoty parametru p výsledek zamítáme a lze říci, že jednotlivé naměřené hodnoty počtu cestujících využívajících křížových jízdních dokladů jiného dopravce v rozmezí dopoledne, odpoledne a večer jsou mezi sebou rozdílné.

j) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ], pro celkové křížové využití jízdních dokladů ve 3 období (dopoledne, odpoledne, večer) u autobusového dopravce [ ].

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 2.1562, df = 2, p - value = 0.3402$

Z hodnoty parametru p výsledek nelze zamítnou, a tudíž tvrdíme, že jednotlivé naměřené hodnoty počtu cestujících využívajících křížových jízdních dokladů jiného dopravce v rozmezí dopoledne, odpoledne a večer vyjadřují podobné naměřené hodnoty.

k) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ],

pro celkové křížové využití jízdních dokladů ve 3 dnech (úterý, středa, čtvrtek) u autobusového dopravce [ ].

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 0.4296, df = 2, p - value = 0.8067$

Z hodnoty parametru p usuzujeme, že výsledek nelze zamítnout, a tvrdíme tak, že jednotlivé naměřené hodnoty počtu cestujících využívajících křížových jízdních dokladů jiného dopravce v rozmezí úterý, středa a čtvrtek jsou mezi sebou srovnatelné.

l) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ], pro celkové křížové využití jízdních dokladů ve 3 dnech (úterý, středa, čtvrtek) u autobusového dopravce [ ].

Kruskal-Wallis rank sum test

data: [ ]
$Kruskal - Wallischi - squared = 1.0321, df = 2, p - value = 0.5969$

Z hodnoty parametru p výsledek nelze výsledek zamítnout, a jednotlivé naměřené hodnoty počtu cestujících využívajících křížových jízdních dokladů jiného dopravce v rozmezí úterý, středa a čtvrtek jsou si tak mezi sebou podobné.

m) Kruskal-Wallis (ův) test pro sběr dat veřejné linkové dopravy [ ] pro poměr nastupujících cestujících a celkovém počtu využitých křížových jízdních dokladů ve 3 dnech (úterý, středa, čtvrtek).

Kruskal-Wallis rank sum test

data: [ ]

$$Kruskal - Wallischi - squared = 2.059, df = 2, p - value = 0.3572$$

Hodnota parametru p vyjadřuje, že výsledek nelze zamítnout. Jednotlivé naměřené hodnoty počtu nastupujících cestujících v poměru celkového počtu využitých křížových jízdních dokladů jiného dopravce v rozmezí úterý, středa a čtvrtek jsou si mezi sebou podobné.

4 DISKUZE

Ke splnění cílů práce bylo zapotřebí nastudovat matematickou statistiku a především jednotlivé typy neparametrických testů. Pro potřeby statistického šetření jsme vybrali nejběžnější statistické metody, které jsme zakomponovali do praktických výpočtů. Největším úskalím bylo statistické šetření v terénu. Protože šetření bylo prováděno více brigádníky-studenty, výsledky byly zatíženy určitou odchylkou, která mohla vzniknout v důsledku špatné informovanosti či pochopení v zadání úkolu. Chybné či neúplné záznamy byly hodnoceny jako "NA", tedy nevalidní. Další problémy, které se vyskytly při průzkumu, zahrnovaly rozmanité zápisy do formulářů u jednotlivých sčítačů. Při následném přepisu do tabulek v programu Excel bylo třeba kontaktovat zmíněné sčítače pro objasnění a doplnění správných výsledků.

Zpracování dat proběhlo bez větších komplikací. Shapiro-Wilk test normality nám potvrdil, že naměřená data lze dále použít pro další testování. Potíže se vyskytly jen při zpracovávání skupiny dat využívající křížových jízdních dokladů v linkách společnosti ve večerních časech. Ve skupině bylo získáno malé množství vzorku s nulovými hodnotami. Ale při využití datasetu dle testování Kruskal-Wallise byla data zpracována v prostředí R bez větších problémů.

Pro dvě datové sady jsme využili dvouvýběrový Wilcoxonův test.Ve třech testováních ze čtyř byly naměřené hodnoty vyhodnoceny jako rozdílné. Výsledky zhodnotily, že počet nastupujících je rozdílný než počet vystupujících. Závěry se potvrdily u společnosti i u společnosti . Výsledek tohoto testu byl pro nás poněkud překvapující, jelikož jsme očekávali stejné počty nastupujících vůči vystupujícím cestujícím ve městě .Rozdíly byly patrné i u křížových jízdenek zakoupených od společnosti a , které byly využívány při jízdě linkami společnosti Dokládá nám to příloha č. 4, která vyjadřuje vyšší počet využívajících jízdenek cestujícími dopravce , který v zřizuje městskou hromadnou dopravu. Výjimku tvoří data získaná při studiu křížových jízdních dokladů společnosti nebo společnosti

využívajících autobusové přepravy. Bylo potvrzeno, že počty cestujících využívajících dvou výše zmíněných dopravců při přepravě u společnosti byly totožné.

K porovnání více datových sad jsme použili Kruskal-Wallisův test, který slouží pro porovnávání několika výběrů. Celkově byl tento test využit ve 13 případech s výsledkem 8 zamítnutých hypotéz a 5 případech, kdy nešlo hypotézu zamítnout. Z testů vyplynulo, že počty nastupujících a vystupujících cestujících v časových obdobích (dopoledne, odpoledne a večer) jsou u společnosti rozdílné, kdežto u společnosti byla naměřená data vyhodnocena jako shodná. Ke stejnému tvrzení jsme došli i u dalšího testovaní, kdy jsme srovnávali počet nastupujících a vystupujících v jednotlivých dnech (úterý, středa a čtvrtek). Při celkovém křížovém využívání jízdních dokladů v různých denních dobách a dnech byla data hodnocena jako srovnatelná. Jen v jednom případě byly naměřené hodnoty rozdílné, a to u křížového využívání jízdenek v linkách společnosti v jednotlivých částech dne (dopoledne, odpoledne a večer). U testování křížových jízdních dokladů linek společnosti jsme došli k hypotéze o shodnosti testovaných dat. Překvapením pro nás bylo, že test využívání křížových jízdenek v linkách společnosti ve dnech úterý, středa a čtvrtek jsou srovnatelné oproti ostatním testům uskutečněným na linkách společnosti . Důvodem může být i to, že tyto linky pravidelně využívají studenti a pracující.

Výsledky testů nejsou vždy objektivní z důvodů nízkého vzorku, např. při testování u společnosti nebo při sledování počtu spojů ve dnech středa a čtvrtek u dopravců . Důvodem takto malého vzorku v tyto dny bylo získání různých dat, protože první den průzkumu jsme se zaměřovali na všechny spoje, kdežto další dny jen na vybrané. Otázkou zůstává, jaké by byly výsledky testů, kdyby se sčítaly každý den všechny spoje.

Výsledné hodnoty testů by se daly aplikovat i na další města jako je nebo spadající pod kraje. . Zajímavé bude

sledovat, jak se projeví větší množství dopravních společností provozující své linky v rámci [redacted] ve městě [redacted] na statistickém průzkumu, nebo větší množství zastávek veřejné linkové dopravy v rámci městské hromadné dopravy v [redacted]. Dalším úskalím by mohlo být i získávání dat z obvodu [redacted], a to především z důvodu velikosti obvodu, množství spojů a většího počtu cestujících.

5 ZÁVĚR

Cílem práce bylo vypracovat studii o neparametrických testovacích metodách. Pomocí těchto metod jsme poté provedli srovnání s praktickými výsledky získanými při třídenním průzkumu veřejné linkové dopravy v obvodu [redacted]. Získaná data byla vizualizována pomocí produktu ESRI.

V teoretické části jsme se zaměřili na objasnění testu normalit, který nám předurčuje, zdali je možné využít datové sady pro další testování. Blíže jsme se zaměřili na teorie nejpoužívanějších jedno, dvou a více výběrových neparametrických testů. Závěrem teoretické části jsme se zaměřili na vybrané softwary využívající neparametrické testování. Tato část úkolu byla z hlediska časové vytíženosti nejnáročnější, jelikož bylo třeba prostudovat literaturu na dané téma a hlouběji proniknout do problému.

V praktické části jsme srovnávali naměřená data ze statistického šetření. Nejprve jsme provedli zápis a úpravu dat pomocí software MS Excel. Následně jsme data zpracovávali a rozřazovali dle parametrů, které jsme měli mezi sebou srovnávat. Výsledné datové sady byly uloženy do textového editoru. Nakonec proběhlo samotné testování v prostředí statistického softwaru R, kam byla načtena jednotlivá data z poznámkových bloků a proveden vstupní test normalit pomocí Shapiro-Wilk testu. Podle výsledků zmíněného testu jsme mohli použít datové sady pro neparametrické testování. Na závěr jsme použili dvouvýběrový Wilcoxonův test pro srovnávání dvou datových sad, pro více výběrové porovnávání jsme použili Kruskal-Wallisův test.

Neparametrické testy jsou vhodným nástrojem při srovnávání velkého množství dat mezi sebou. Zjistili jsme, že v případě srovnání menšího množství získaných údajů není neparametrické testování tak přesné jako u velkého vzorku.

Co se týče softwarové využití z hlediska ESRI produktů, chybí podpora pro neparametrické testování, a nelze tak přímo v tomto prostředí provádět testy, které by šly zpracovat do mapových podkladů.

6 LITERATURA

Reference

[1] ANDĚL, J.: *Statistické metody, Matematicko-fyzikální* fakulta Univerzity Karlovy, MATFYZPRESS, Praha, 1998, 274s

[2] ANTOCH, J., VORLÍČKOVÁ, D.: *Vybranémetody statistické analýzy dat*, Academia, Praha, 1992, 280s

[3] HENDL, J.: *Přehled statistických metod zpracování dat.*, Portál, Praha, 2004,

[4] KUNDEROVÁ, P: *Úvod do teorie pravděpodobnosti a matematické statistiky,*, Univerzita Palackého, Olomouc 1997, 194s

[5] OLŠÁK, P.: *Typografický sýstém TEX.*, CsTUG, 1995,

[6] RYBIČKA, J.: *LATEX pro začátečníky.*, 2. vydání, Konvoj, 1999,

[7] VOŽENÍLEK, V.: *Diplomové práce z geoinformatiky.*, Univerzita Palackého, Olomouc, 2002, 61s

[8] *MATLAB*, [online] 2010, [cit. 2010-11-05]. Dostupný z WWW:
<http://cs.wikipedia.org/wiki/MATLAB>

[9] *R-projekt*, [online] 2008, [cit. 2010-23-3]. Dostupný z WWW:
<http://www.r-project.cz/about.html>

[10] *Statistické funkce v Excelu*[online] 1999, [cit. 2010-23-3]. Dostupný z WWW:
<http://www.med.muni.cz/cvt/Excel97/Statistika_html/Statistika.htm>

[11] *UNISTAT*,[online] 2010,[cit.2010-23-3]. Dostupný z WWW:
http://www.unistat.cz/details.php?obsah=0}

[12] *Jaroslav Hájek*, [online] 2010. [cit. 2010-19-2]. Dostupný z WWW:

<http://www-groups.dcs.st-and.ac.uk/history/Biographies/Hajek.html>

[13] *Jacob Wolfowitze*, [online] 2010. [cit. 2010-19-2]. Dostupný z WWW: <http://www.ifp.illinois.edu/~junchen/jacob_wolfowitz.htm>

.

SUMMARY

Mathematical statistics is a scientific discipline which represents a boundary between the descriptive statistics and the applied mathematics. When using the method of probability theory it tries to estimate the features of the destribution of the observed data. Parametric and non-parametric testing belong to these methods. The parametric tests presuppose concrete data distribution using a given parameter for the calculation. However, if we do not know the data distribution, we use non-parametric tests for the calculation. Non-parametric testing is also used with the data of the ordinal scale. The weak point of the non-parametric tests is a smaller predicative value especially when having smaller amount of measured data. The aim of this work was to study non-parametric testing methods. We compared the hypothetical non-parametric data with the practical results we obtained during a three-day survey of the municipal bus operation in [ ]. The municipal bus operation in the above mentioned district is provided by 4 transport companies [ ]. The public transport in the town is provided by the transport company [ ]. According to the regulation of the [ ] it is possible to use a valid travel document when travelling with another transport company. As a part of the survey we counted the number of people getting on and getting off the bus at individual bus stops for a certain bus route or for a cross using of travel documents from different transport companies. The results we obtained were further used for a mutual comparison within various parameters in the enviroment of the software R. We visualized the obtained data with the help of the product ESRI.

In order to achieve the aims of our work it was necessary to study the mathematical statistics and above all the individual types of non-parametric tests. In the theoretical part we focused on the explanation of normality test, which predestines whether it is possible to use data sets for further testing. We also concentrated on the theories of the most used one, two or more selective non-parametric tests, which we included in the practical calculations. At the end of the theoretical part we focused on selected software that uses non-parametric testing. This part of the task was, concerning the time, the most demanding because we had to study

the literature related to the topic to understand the problematic thoroughly. In the practical part we compared the measured data from the statistical survey. As the survey was realized by more voluntary students, the outcomes were burdened with a certain statistical error. That could be caused by a lack of information or misunderstanding of the instructions.

The data processing went without major complications. At first, we carried out data logging and conditioning with the help of software MS Excel. Then we processed the data and grouped them according to the parameters, which we compared. The resulting data sets were later put in the text editor. Finally, the testing itself took place in the enviroment of the statistical software R, where the individual data items were loaded from the notebooks, and the initial normality test was carried out with the help of Shapiro-Wilk test. Shapiro-Wilk normality test proved to us that the measured data could be used for further testing.

A two-related-sample Wilcoxonův test was used for two data sets. The results showed that the number of passangers getting on was different from the number getting off. The findings proved true with both [ ] and [ ]. The result of this test was for us rather surprising because we expected the same numbers of getting on passangers and those getting off in the town of [ ]. Differences were apparent with cross tickets bought from the companies [ ] and [ ] [ ] These tickets were used on the buses of the company [ ]. It is illustrated in the appendix no. 4, which expresses a higher number of passangers using the tickets of the transport company [ ], which in [ ] runs the town public transport.

To compare more data sets we used Kruskal-Wallis test, which serves to compare several sets. The tests showed us that the numbers of getting on and getting off passangers in the time periods (in the morning, afternoon and evening) are different regarding the company [ ], whereas as for the company [ ] the obtained data were evaluated as identical. Further testing where we compared the number of passangers getting on and off on individual days (Tuesday, Wednesday and Thursday) led us to the same conclusion.

When general cross using of travel documents it various day times and days the data were evaluated as comparable. The measured data were different only in one case and that was when cross using of tickets on the buses of [ ] in individual parts of a day (in the morning, afternoon and evening).

When testing cross travel documents with bus routes of the company [ ] we came to the hypothesis about the identity of tested data. We were surprised that the results of the test of using cross tickets on the buses of the company [ ] on Tuesdays, Wednesdays and Thursdays are comparable with the other tests realized on the buses of [ ]. The reason might be that these buses are regularly used by students and workers.

The tests outcomes are not always objective and it is because of a small sample, e.g. testing in the company [ ] or observing the number of routes on Wednesdays and Thursdays with the companies [ ] and [ ]. The reason for such a small sample on these days was obtaining different data because during the first day of our survey we concentrated on all bus routes, while the next days we looked only at some. The question is what results we would have got, if all routes would have been calculated every day.

The tests outcomes could be applied in other towns as for example [ ] or [ ], which also belong to the [ ] region. Statistical survey concerning the public transport in the above mentioned towns is supposed to be carried out in the autumn and spring next calendar year. It will be interesting to observe how a bigger number of transport companies running their business within [ ] (in the town [ ]) will influence the statistical survey. Also a bigger number of bus stops of the municipality bus operation within the public transport in [ ] might have an impact on the surfvey. Another problem could be the process of collecting data from the district [ ] mainly because of the size of the district, the number of routes and a bigger number of passangers. In conclusion, non-parametric tests are a suitable tool to compare a large amount of data. We found out that when

comparing a smaller number of collected facts the non-parametric testing is not as exact as in the case of a big sample. As for the software application from the point of view of ESRI products the support for non-parametric testing is missing. Therefore, it is not possible to carry out the tests, which could be processed into background material for maps, directly in this environment.

As a part of our bachelor project we created a web site posted on the server of the Department of Geoinformatics UP. With respect to the data protection, which were obtained from the transport companies and the company providing statistical surveys and evaluation, we cannot allow a third party to read the text and the data of this work.

SEZNAM PŘÍLOH

Vázané přílohy:

1. Formulář pro sčítání veřejné linkové dopravy spojů v autobusech

Volné přílohy:

2. Vstupní a výstupní data (DVD-ROM)

Seznam mapových příloh:

3. NÁSTUP A VÝSTUP CESTUJÍCÍCH V ZASTÁVKÁCH VEŘEJNÉ LINKOVÉ DOPRAVY [ ], stav v roce 2010

4. KŘÍŽOVÉ VYUŽITÍ JÍZDNÍCH DOKLADŮ MEZI DOPRAVCI VEŘEJNÉ LINKOVÉ DOPRAVY NA ZASTÁVKÁCH [ ], stav v roce 2010

5. PODÍL NA OBSLUZE ZASTÁVEK VEŘEJNÉ LINKOVÉ DOPRAVY SPOJI DOPRAVNÍCH SPOLEČNOSTÍ [ ], stav v roce 2010

6. ZASTÁVKY A VEDENÍ LINEK VEŘEJNÉ LINKOVÉ DOPRAVY V [ ] INTEGROVANÉHO DOPRAVNÍHO SYSTÉMU [ ] [ ], stav v roce 2010

Příloha č.1 (Formulář pro sčítání veřejné linkové dopravy spojů v autobusech).