UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

PERMUTAČNÉ TESTY A INTERVALY SPOĽAHLIVOSTI

2011

László Pastorek

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

PERMUTAČNÉ TESTY A INTERVALY SPOĽAHLIVOSTI

Bakalárska práca

Študijný program: ekonomická a finančná matematika
Študijný odbor: 9.1.9 aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Školiteľ: Mgr. Ján Somorčík, PhD.
Evidenčné číslo: eef03aff-39f3-45b8-862e-2660aff140b9

Bratislava, 2011 **László Pastorek**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: László Pastorek
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov : Permutačné testy a intervaly spoľahlivosti

Cieľ : Naštudovať si z literatúry a pomocou počítača simulačne preskúmať správanie zopár permutačných štatistických metód.

Vedúci : Mgr. Ján Somorčík, PhD.

Dátum zadania: 27.10.2010

Dátum schválenia: 08.11.2010

.....
doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Dátum potvrdenia finálnej verzie práce, súhlas s jej odovzdaním (vrátane spôsobu sprístupnenia)

.....
vedúci práce

Prehlásenie

Čestne prehlasujem, že som predloženú bakalársku prácu vypracoval samostatne s využitím teoretických vedomostí a s použitím uvedenej literatúry.

Bratislava, 2.6.2011.

_____

podpis

Poďakovanie

Chcel by som poďakovať všetkým, ktorí mi akýmkoľvek spôsobom pomohli pri vypracovaní tejto bakalárskej práce. Moje poďakovanie patrí najmä vedúcemu práce, Mgr. Jánovi Somorčíkovi PhD., za vedenie a za cenné pripomienky pri záverečnom spracovaní práce.

ABSTRAKT

PASTOREK László: Permutačné testy a intervaly spoľahlivosti /bakalárska práca/, Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Ekonomická a finančná matematika, školiteľ: Mgr. Ján Somorčík PhD., Bratislava, 2011

Cieľom mojej bakalárskej práce je naštudovať si z literatúry a pomocou počítača simulačne preskúmať správanie sa zopár permutačných štatistických metód. V prvej časti práce sa venujem teoretickému základu a vysvetleniu najdôležitejších pojmov, ktoré sa v tejto problematike vyskytujú. Druhá časť práce sa zaoberá vyhodnotením permutačných štatistických metód a ich porovnávaním s inými štatistickými metódami. V poslednej časti práce sa zaoberám problémami, ktoré vznikli pri programovaní našich počítačových simulácií.

Kľúčové slová: permutácia, P(chyba 1.druhu), sila testu, p-hodnota

ABSTRACT

PASTOREK László: Permutation tests and confinence intervals (bachelor thesis), Comenius University of Bratislava, Faculty of matematics, physics and informatics, Economic and financial mathematics, Consultant: Mgr. Ján Somorčík,PhD., Bratislava, 2011

The aim of this thesis is to study the literature and using computer simulations to examine the behavior of some permutation of statistical methods. The first part is dedicated to the theoretical explanation and the most important terms that appear inthis issue. The second part deals with the evaluation of the permutation statistical methods and their comparison with other statistical methods. In the last part of the work deals with the problems that arose in the programming of our computersimulations.

Key Words: permutation, P(I.type error), power, p-value

OBSAH

1. Úvod

Testovanie hypotéz je jednou zo základných problematík matematickej štatistiky. Je to prístup, ktorý často používame aj v reálnom svete.

Poznáme viacero metód, pomocou ktorých sa dajú hypotézy otestovať. Veľakrát sa stane, že úspešná aplikácia týchto metód vyžaduje splnenie takých predpokladov, ktoré reálne dáta nie vždy spĺňajú (napríklad normalita dát, nekonečný počet dát a pod.). Preto sa môže stať, že použijeme testovaciu metódu, ktorej predpoklady nie sú splnené a výsledky našich testov budú nesprávne a sklamú naše očakávania.

Tento problém matematickej štatistiky vzbudil aj náš záujem, a preto sme sa rozhodli skúmať také metódy testovania hypotéz, ktoré nekladú na reálne dáta príliš silné, mnohokrát ani nesplniteľné predpoklady . Jedným z takýchto metód sú tzv. permutačné metódy . Táto práca je venovaná popisu týchto permutačných metód a ich využitiu pri testovaní hypotéz,

V prvej časti (kapitole) sú uvedené základné pojmy , ktoré sa v tejto problematike najčastejšie vyskytujú a oboznámime sa so základnými princípmi permutačných metód. V druhej časti práce sú spracované výsledky pomocou počítača urobených simulácií . Na základe týchto výsledkov sledujeme správanie sa našich permutačných metód v porovnaní s inými štatistickými metódami a vyhodnocujeme naše metódy pri rôznych typoch a počtoch dát respektíve poukážeme na ich výhody či nevýhody voči iným testovacím metódam.

Posledná časť práce sa zaoberá programátorskými záležitosťami. Sú v nej uvedené problémy, ktoré sa počas počítačových simulácií môžu vyskytnúť a ich možné riešenia.

2. Základné pojmy

V štatistike sa často stretávame s úlohami, ktoré sa zaoberajú testovaním štatistických hypotéz. Takáto hypotéza môže byť napríklad odhad strednej hodnoty nejakého merania (výška ľudí, čas cestovania). Hypotézu, ktorej platnosť overujeme nazývame *nulovou hypotézou* a označujeme ju H_0. Proti nej stojí tzv. *alternatívna hypotéza* H_1. Pravidlo, podľa ktorého rozhodneme o tom či testovanú hypotézu zamietneme, resp. nezamietneme nazývame štatistickým testom.

Postup pri testovaní hypotéz:

Nech máme dáta náhodného výberu $X_1,\dots,X_n$ (hodnoty merania) so strednou hodnotou μ a disperziou σ^2. Nami predpokladaná stredná hodnota je μ_0. Vo všeobecnosti máme 3 prípady testovania:

$H_0 : \mu = \mu_0$ testujeme proti obojstrannej alternatíve $H_1 : \mu \neq \mu_0$

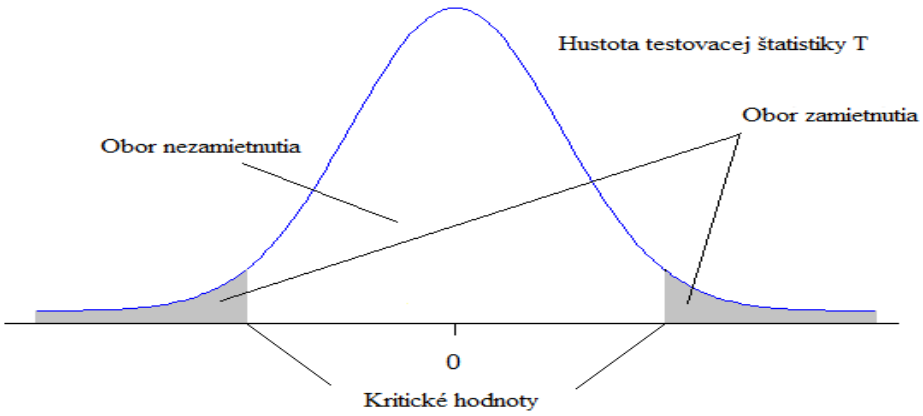
$H_0 : \mu \leq \mu_0$ testujeme proti pravostrannej alternatíve $H_1 : \mu > \mu_0$

$H_0 : \mu \geq \mu_0$ testujeme proti ľavostrannej alternatíve $H_1 : \mu < \mu_0$

Na testy hypotéz potrebujeme *testovaciu štatistiku*. To je funkcia $T = f(X_1,\dots,X_n)$, ktorá sa väčšinou skonštruuje tak, aby mala nejaké známe rozdelenie. Množinu hodnôt, ktoré testovacia štatistika nadobúda, rozdelíme na dve disjunktné množiny:

1. *Rejection region(R)* – obor zamietnutia testovanej hypotézy H_0

2. *Acceptance region (A)* – obor nezamietnutia testovanej hypotézy H_0.



Ak hodnota testovacej štatistiky T nadobudne hodnoty z oboru R, tak hypotézu H_0 zamietneme a prijmeme alternatívnu hypotézu H_1. Ak hodnota testovacej štatistiky T nenadobudne hodnoty z oboru R ale z oboru A, tak hypotézu H_0 nezamietneme. Hodnoty, ktoré oddeľujú obor A od oboru R, nazývame *kritické hodnoty (cut off points)*, ktoré sú určené vopred. Testy sú založené na náhodných dáta a preto môžu nastať dva druhy chýb:

chyba 1. druhu (type 1. error, alpha error) : H_0 platí, ale test ju zamietol

chyba 2. druhu (type 2. error, beta error) : H_0 neplatí, ale test ju nezamietol.

Pravdepodobnosť $P(\text{chyba 1.druhu}) = \alpha$ (alebo $P(\text{chyba 1.druhu}) \leq \alpha$ a žiadnou menšou α sa to ohraničiť nedá) a nazýva sa *hladina významnosti (level of significance).* Číslo α vyjadruje to, že ak H_0 platí a dlhodobo opakujeme náš test (veľakrát s inými dátami), tak test sa pomýli (zamietne H_0) v $100.\alpha\ \%$ prípadoch. Pri testoch žiadame , aby α bolo čo najmenšie, vo všeobecnosti volíme $0.05,\ 0.1$. Pomocou α určujeme aj kritické hodnoty a to tak, aby pod hustotou T mal obor nezamietnutia obsah $1 - \alpha$, a obor zamietnutia mal obsah α (je známe že celý obsah pod hustotou sa rovná 1). Pomocou pravdepodobnosti $P(\text{chyba 2.druhu})$ vypočítame *silu testu (power)*

$$1 - P(\text{chyba 2. druhu}) = 1 - P(test\ nezamietne\ H_0 | H_0 neplatí) =$$

$$= P(test\ zamietne\ H_0 | H_0 neplatí) = sila\ testu.$$

V súčasnosti sa na testovanie hypotéz používa aj postup, pri ktorom sa určí dosiahnutá hladina testu, t.j. najmenšia hladina významnosti, na ktorej ešte hypotézu H_0 zamietneme. Je to tzv. *p-hodnota (p-value)*. Inými slovami $p\text{-hodnota} = P(T\ nadobúda\ ešte\ extrémnejšiu\ hodnotu,\ než\ T = f(X_1,\ldots,X_n)\ /\ H_0\ platí)$. Jej malé hodnoty hovoria v neprospech hypotézy H_0:

$$H_0\ zamietneme\ ak\ p\text{-}value < \alpha$$

V skutočnosti je pravda, že $p\text{-hodnota}$ sa niekedy nedá vypočítať z viacerých dôvodov:

- T má takú hustotu, z ktorej sa ťažko ráta integrál
- Nepoznáme hustotu.

V takýchto prípadoch možno použiť permutačný test.

3. Princíp permutačného testu

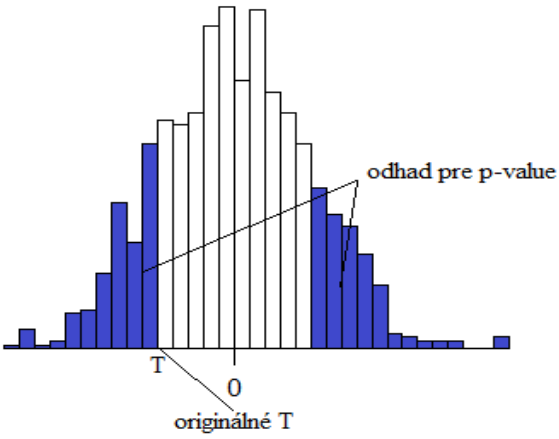
Permutačný test patrí medzi neparametrické testy. Pri teste nepotrebujeme, aby testovacia štatistika mala známe rozdelenie, ani to, aby testované dáta pochádzali z nejakého konkrétneho rozdelenia (pri testoch, kde testovacia štatistika má známe rozdelenie, je často nutné, aby dáta pochádzali z normálneho rozdelenia).

Princíp vysvetlíme na konkrétnom prípade:

Nech máme náhodný výber $X_1,\dots,X_n$ z nejakého rozdelenia s $E(X) = \mu_1$ a $D(X) = \sigma^2$ a náhodný výber $Y_1,\dots,Y_m$ z nejakého rozdelenia s $E(X) = \mu_2$ a $D(X) = \sigma^2$. Vypočítame testovaciu štatistiku $T = f(X_1,\dots,X_n, Y_1,\dots,Y_m)$. Teraz zoberieme náš náhodný výber dát $(X_1,\dots,X_n,Y_1,\dots,Y_m)$ a z $n + m$ dát vyberiem n to bude nové X a zo zvyšných m dát bude nové Y tým dostaneme novú testovaciu štatistiku T_2. Toto opakujeme N krát (je zrejmé, že v tomto konkrétnom prípade maximálny N je (n+m nad n)). Definujme funkciu $F(T) = \frac{poradie}{N}$, kde *poradie* označuje poradie T medzi $T_2,\dots, T_N$, ktoré sú usporiadané podľa veľkosti. Budeme rozlišovať dva prípady:

- $T < 0$
- $T > 0$

Keď platí $T < 0$, tak odhad pre *p-value* bude: $p\text{-value} = 2F(T)$. Tento prípad je znázornený na nasledovnom histograme:



Pre $T > 0$ bude odhad : $p\text{-value} = 2(1-F(T))$.

Jedna výhoda tejto metódy je, že za platnosti H_0 nemá permutovanie dát vplyv na rozdelenie testovacej štatistiky, a tým pádom za platnosti H_0 sú $T_1,...,T_N$ rovnako rozdelené a ich histogram je odhadom pre skutocnú hustotu T. Uvedený permutačný výpočet p-value potom znamená, že za platnosti H_0 bude mať permutačný test P(chyba 1.druhu) = 5%.

4. Porovnávanie štatistických metód

Pri porovnávaní testov budeme porovnávať ich dve vlastnosti:

1. *P(chyba 1. druhu)*
   Test je spoľahlivý ak $P(\text{chyba 1. druhu}) = \alpha$, test ktorý to nesplní budeme pokladať za zlý.
2. *Sila testu*
   Keď oba testy majú $P(\text{chyba 1. druhu}) = \alpha$, vtedy porovnáme ich sily a lepší bude ten, ktorý má väčšiu silu.

P(chyba 1.druhu) a silu testu zistíme pomocou počítačových simulácií. Náš postup bude nasledujúci:

1.krok: vygenerujeme náhodné dáta X, Y – pri porovnávaní budeme pracovať s totožnými dátami (porovnávané testy budú vykonané na tie isté dáta)

2.krok: na dátach vykonáme testy a pozrieme sa, či hypotéza H_0 bola prijatá alebo zamietnutá

3:krok : prvé dva kroky opakujeme 5000 krát (ešte prijateľná časová náročnosť, pri počítačových výpočtoch). Potom ak hypotéza H_0 platí

$$P(\text{chyba } 1.\text{druhu}) \approx \frac{\text{počet zamietnutia } H_0}{5000}$$

ak hypotéza H_0 neplatí

$$\text{sila} \approx \frac{\text{počet zamietnutia } H_0}{5000}$$

4.1.t-test

V tejto časti budeme porovnávať permutačnú a klasicku verziu Studentovho t-testu (obojstrannú dvoj-výberovú). Hypotéza, ktorú budeme testovať je nasledujúca:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2$$

kde μ_1 je stredná hodnota náhodného výberu $X = (X_1,...,X_n)$ a μ_2 je stredná hodnota náhodného výberu $Y = (Y_1,...,Y_m)$. Na porovnávanie testov simulujeme tri prípady:

- 1.prípad

Zoberieme dáta, z normálneho rozdelenia tak, aby mali rovnaké stredné hodnoty $E(X) = E(Y)$ a disperzie $D(X) = D(Y) = \sigma^2$ (disperzie vo všetkých troch prípadoch budú

rovnaké), teda bude platiť nulová hypotéza. To znamená, že ak dáta otestujeme, H_0 by sme nemali nikdy zamietnuť. Keďže sú to náhodné dáta, po viacerých opakovaniach sa naše testy niekoľkokrát pomýlia a nesprávne zamietnu H_0. Hladinu významnosti zvolíme $\alpha = 0.05$ a potom budeme očakávať, že naše testy budú mať P(chyba 1.druhu) blízko 5%.

.

- 2.prípad

Taktiež budeme mať normálne dáta ako v 1.prípade, len stredné hodnoty nebudú rovnaké $E(X) \neq E(Y)$. Keďže H_0 neplatí, lepší test by ho mal zamietnuť a vo väčšine prípadov prijať hypotézu H_1. Budeme porovnávať sily testov.

- 3prípad

V tomto prípade zoberieme dáta z Cauchyho rozdelenia, ktoré ale budú mať rovnaké stredné hodnoty a $E(X) = E(Y)$ a skúmať budeme P(chyba 1.druhu), podobne ako v prvom prípade.

4.1.1. Studentov t-test

Veta 1 Nech máme náhodný výber $X = (X_1,\dots,X_n)$ a náhodný výber $Y = (Y_1,\dots,Y_m)$, ktoré spĺňajú nasledujúce predpoklady:

a) $X_1,\dots,X_n \sim N(\mu_1, \sigma^2)$
b) $Y_1,\dots,Y_m \sim N(\mu_2, \sigma^2)$
c) X a Y sú nezávislé

potom testovacia štatistika T má Studentovo rozdelenie s $n+m-2$ stupňami voľnosti

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S} \sqrt{\frac{n.m}{n+m}} \sim t_{n+m-2}$$

kde

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

S_X^2 a S_Y^2 sú odhady pre σ^2.

Majme hypotézy:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2$$

ak platí hypotéza H_0, tak podľa Vety 1.:

$$T = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n.m}{n+m}} \sim t_{n+m-2}$$

potom

$$H_0 zamietneme, ak\ T > t_{n+m-2;\frac{\alpha}{2}}\ alebo\ T < -t_{n+m-2;\frac{\alpha}{2}}$$

Hore uvedený test sa nazýva *obojstranný-dvojvýberový-Studentovo t- test*.

Veta 2. Uvedený Studentov t-test má P(chyba 1.druhu) rovnú α.

Zdrojom tejto podkapitoly bola kniha [1].

4.1.2. Permutačný test

Permutačný test, ktorý budeme porovnávať so Studentovým t-testom je skonštruovaný nasledovne:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2$$

potom podľa Vety 1. testovacia štatistika T sa rovná:

$$T = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n.m}{n+m}}$$

Ak platí hypotéza H_0, tak $X_1,\ldots,X_n$ a $Y_1,\ldots,Y_m$ sú „rovnocenné“ (majú rovnaké rozdelenie, rovnakú strednú hodnotu a disperziu), čiže môžeme ich navzájom premiešať. Takto dostaneme dva nové súbory, pre ktoré vypočítame testovaciu štatistiku T_2. Potom ich znova pomiešame a dostaneme novú hodnotu T_3. V podstate nerobíme nič iné, len vytvoríme z $X_1,\ldots,X_n,Y_1,\ldots,Y_m$ nové kombinácie dát. ($z\ n+m$ dát vyberiem n to bude nové X a zo zvyšných m dát bude nové Y). Toto opakujeme N krát a potom sa pozrieme na poradie T medzi hodnotami T_1 až T_N a pomocou metódy uvedenej v 3.kapitole [11.strana] dostaneme odhad pre *p-value*.

$$H_0\ zamietneme, ak\ p - value <\ \alpha = 0.05$$

Z princípu permutačného výpočtu p-value zjavne plynie, že P(chyba 1. druhu) je α.

Počas testu budeme vytvárať kombinácie pomocou X a Y. Treba si uvedomiť, že počet všetkých možných kombinácií rastie veľmi rýchlo. Napríklad pre $n = m = 5$ je to *252*, pre $n=m=10$ už *184 756* kombinácií. Práve preto pri výpočtoch rozdelíme testy na test pre malý počet dát a test pre veľký počet dát. Pri malých dátach zoberieme všetky kombinácie a pri veľkých dátach len zopár (od *100* do *5000*) náhodne vybratých.

4.1.3. Výsledky počítačových simulácií

Najprv sme aplikovali testy na malý počet dát. Ich výsledky sú zhrnuté v prvých troch tabuľkách (v tabulkách máme označené: PT-permutačný test, ST-Studentov t-test, n-počet X dát, m-počet Y dát, N(0,1)-dáta v tomto stĺpci pochádzajú z rozdelenia N(0,1)).

Z prvej tabuľky vidieť, že obidva testy majú P(chyba 1. druhu) veľmi blízko *5%*, a navyše sú rovnako spoľahlivé

N(0,1)	**N(0,1)**	**P(chyba 1. druhu)**	
n	**m**	**PT**	**ST**
5	**5**	0.049	0.050
7	**5**	0.049	0.049
7	**7**	0.045	0.045
10	**5**	0.051	0.050
10	**7**	0.049	0.050

To, že ktorý z nich je lepší sa týchto hodnôt nedá posúdiť. Viac sa o tom možno dozvedieť z ďalšej tabuľky, ktorá obsahuje silu testov.

N(0,1)	**N(1,1)**	**Sila testu**	
n	**m**	**PT**	**ST**
5	**5**	0.296	0.280
7	**5**	0.351	0.351
7	**7**	0.405	0.407
10	**5**	0.395	0.397
10	**7**	0.459	0.461

Podobne ako pri prvej tabuľke, aj tu sú hodnoty veľmi blízke, sily testov sú skoro rovnaké. Možno však skonštatovať, že pri $n=m=5$ permutačný test (PT) má väčšiu silu, ale pri väčšom počte dát narastá sila Studentovho testu.

Pri normálnych dátach teda fungovali rovnako dobre oba testy. Teraz sa pozrieme, čo sa stane, ak dáta nebudú pochádzať z normálneho rozdelenia, teda nebudú splnené predpoklady vety1. Práve preto očakávame, že Studentov t-test by mal byť horší ako permutačný.

Cauchy	**Cauchy**	**P(chyba 1. druhu)**	
n	**m**	**PT**	**ST**
5	**5**	0.046	0.019
7	**5**	0.044	0.020
7	**7**	0.043	0.015
10	**5**	0.048	0.027
10	**7**	0.051	0.021

Naše očakávania potvrdili aj výsledky počítačových výpočtov. Pravdepodobnosť chyby 1. druhu je len okolo *0.02*. Testy ktoré majú nízku P(chyba 1. druhu) sa nazývajú konzervatívnymi testmi, lebo majú menšiu silu (vôľu) zamietnuť nulovú hypotézu a prijať novú, alternatívnu hypotézu. Hodnoty permutačného testu sú taktiež menej presnejšie ako predtým, ale stále sú na úrovni okolo hodnoty *0.05*. Väčšie odchýlky od hodnoty *0.05* možno vysvetliť tým, že robíme len 5000 simulácií (pri 50000000000 simuláciach by boli odchýlky od 5% oveľa menšie). Silu testov nebudeme porovnávať, lebo vieme, že Studentov test sa správa konzervatívne a práve preto stále bude mať menšiu ochotu zamietnuť nulovú hypotézu H_0.

Po malých počtoch dát, sa teraz pozrieme ako fungujú testy pre väčší počet dát. V tabuľkách číslo N označuje koľkokrát sme v permutačnom teste pomiešali dáta. Najprv sa pozrieme na pravdepodobnosť P(chyba 1. druhu). Jej hodnoty sú okolo *5%* podobne ako pri malom počte dát. Už N=100 permutacii prináša P(chyba 1. druhu) blízku 5%.

N(0,1)	**N(0,1)**	**P(chyba1.druhu)**					
n	**m**	**N=100**	**N=500**	**N=1000**	**N=2500**	**N=5000**	**ST**
10	**10**	0.051	0.054	0.050	0.051	0.048	0.054
15	**10**	0.050	0.054	0.048	0.049	0.056	0.046
15	**15**	0.048	0.048	0.055	0.050	0.050	0.044
20	**10**	0.054	0.051	0.045	0.050	0.050	0.046
20	**15**	0.051	0.050	0.046	0.051	0.048	0.042
20	**20**	0.047	0.054	0.051	0.050	0.052	0.048

V ďalšej tabuľke sú uvedené hodnoty síl jednotných testov. Vidíme, že sú veľmi podobné, ani jeden test nevyniká nad ostatné. Zvýšením počtu dát sa zvýši aj sila testov, čo je prirodzené, veď keď máme viac dát, vieme o danom jave viac povedať. Pri permutačnom teste možno si všimnúť, že narastaním počtu dát postupne väčšiu silu majú testy s vyšším počtom kombinácií N. Z toho sa dá predpokladať, že možno existuje nejaký optimálny počet kombinácií N (v závislosti od n a m), nad ktorým sa už sila testu nezvyšuje, teda netreba zobrať všetky kombinácie, ale len optimálny počet N.

N(0,1)	**N(1,1)**	**Sila testu**					
n	**m**	**N=100**	**N=500**	**N=1000**	**N=2500**	**N=5000**	**ST**
10	**10**	0.573	0.573	0.560	0.565	0.565	0.565
15	**10**	0.654	0.649	0.641	0.644	0.653	0.629
15	**15**	0.761	0.751	0.743	0.744	0.744	0.756
20	**10**	0.698	0.706	0.695	0.710	0.690	0.701
20	**15**	0.799	0.803	0.810	0.816	0.814	0.810
20	**20**	0.864	0.869	0.866	0.871	0.873	0.874

Pri malom počte dát zrušenie normality dát spôsobilo, že permutačný test zostal naďalej spoľahlivý ale Studentov test sa stal konzervatívnym . To isté sa stalo aj pri veľkom počte dát.

N(0,1)	N(0,1)	P(chyba1.druhu)					
n	**m**	**N=100**	**N=500**	**N=1000**	**N=2500**	**N=5000**	**ST**
10	**10**	0.050	0.048	0.047	0.047	0.050	0.019
15	**10**	0.047	0.044	0.044	0.051	0.056	0.023
20	**10**	0.047	0.053	0.054	0.051	0.054	0.020
15	**15**	0.048	0.046	0.047	0.044	0.051	0.027
20	**15**	0.053	0.046	0.050	0.045	0.052	0.020
20	**20**	0.044	0.050	0.056	0.052	0.054	0.018

Ako výsledok porovnávania Studentovho t-testu s permutačným testom možno konštatovať nasledovné veci:

- pri dátach, ktoré majú normálne rozdelenie sú Studentov t-test aj permutačný test rovnako spoľahlivé
- pri dátach z iného rozdelenia bol spoľahlivý len permutačný test
- pri permutačnom teste nepotrebujeme všetky kombinácie aj bez toho môžeme získať spoľahlivé testy s veľkou silou.

4.2. Testovanie vzájomnej korelácie dát

V predchádzajúcej časti sme porovnávali testy pomocou stredných hodnôt. Teraz budeme skúmať závislosť medzi dátami náhodného výberu $X = (X_1,\ldots,X_n)$ a dátami náhodného výberu $Y = (Y_1,\ldots,Y_n)$. Hypotéza, ktorú budeme testovať je nasledujúca:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

kde ρ je korelacný koeficient medzi dvojicami X_i a Y_i.

- 1.prípad

Budeme mať normálne dáta s rovnakou strednou hodnotou $E(X) = E(Y)$ a s rovnakou disperziou $D(X) = D(Y)$, pritom X a Y budú nezávislé. To znamená, že ak dáta otestujeme, hypotézu H_0 by sme nemali nikdy zamietnuť. Keďže sú to náhodné dáta, po viacerých opakovaniach sa naše testy niekoľkokrát pomýlia a nesprávne zamietnu hypotézu H_0. Hladinu významnosti zvolíme $\alpha = 0.05$ a potom budeme očakávať, že testy budú mať P(chyba 1.druhu) práve 5%.

- 2.prípad

Taktiež budeme mať normálne dáta ako v 1.prípade, len budú závislé. Keďže hypotéza H_0 neplatí, lepší test by ju mal zamietnuť a prijať hypotézu H_1 vo väčšine prípadov. Budeme porovnávať sily testov.

- 3prípad

V tomto prípade zoberieme nezávislé dáta z Cauchyho rozdelenia. Skúmať budeme P(chyba 1.druhu), podobne ako v prvom prípade.

Mieru závislosti medzi dátami X a Y budeme určovať pomocou troch korelačných koeficientov:

I. *Pearsonov korelačný koeficient*
II. *Spearmanov korelačný koeficient*
III. *Kendallov koeficient korelácie*

Testy ktoré budeme v tejto časti porovnávať budú permutačný test (PT) a tzv. klasický (KT). Pod slovom klasický test budeme rozumieť testy, ktoré pri testovaní budú porovnávať danú testovaciu štatistiku s kritickými hodnotami príslušného rozdelenia, a podľa toho rozhodnú či hypotézu H_0 zamietnu, alebo ju nezamietnu. Presné tvary klasických testov uvedieme neskôr, osobitne pri každom prípade.

4.2.1. Permutačný test

Pri všetkých testoch budeme pracovať s náhodnými výbermi dát $X_1,\dots,X_n$ a $Y_1,\dots,Y_n$. Keďže budeme merať závislosť, dáta usporiadame do dvojíc

$$\binom{X_1}{Y_1}, \binom{X_2}{Y_2}, \dots, \binom{X_n}{Y_n}$$

Potom podľa zvolenej metódy (Pearson, Spearman, Kendall; viď nižšie) pomocov uvedených vzorcov vypočítame hodnotu prislušného výberového korelačného koeficientu ρ. Teraz zoberieme dvojice a pomiešame poradie X dát, čím dostaneme nové dvojice, napríklad:

$$\binom{X_3}{Y_1}, \binom{X_n}{Y_2}, \dots, \binom{X_1}{Y_n}$$

Znovu vypočítame korelačný koeficient ρ_2. Toto opakujeme N krát (je zrejme, že maximálny počet permutácie je $n!$) a potom sa pozrieme na poradie ρ medzi hodnotami ρ_1 až ρ_N a pomocou metódy uvedenej v 2.kapitole [9.strana] dostaneme odhad pre $p\text{-}value$.

$$Hypotézu\ H_0 zamietneme, ak\ p - value < \alpha = 0.05$$

Uvedený permutačný test má $P(chyba\ 1.druhu) = \alpha$.

Počas testu budeme vytvárať permutácie z dát X. Treba si uvedomiť, že počet všetkých možných permutácií rastie veľmi rýchlo. Napríklad pre $n = 5$ je to 120, ale pre $n=10$ je to už $3\ 628\ 800$ permutácií! Práve preto pri výpočtoch zase rozdelíme testy na test pre malý počet dáta test pre veľký počet dát. Pri malých dátach aplikujeme všetky permutácie a pri veľkých dátach len zopár (od 100 do 5000) náhodne vybratých.

4.2.2. Pearsonov korelačný koeficient

4.2.2.1. Teória

Pearsonov korelačný koeficient dvoch premenných je definovaný ako podiel kovariancie a súčtu štandartných odchýlok :

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}.$$

Odhad Pearsonovho korelačného koeficienta :

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Veta 3. Nech nezávislé dáta *X,Y* pochádzajú z normálneho rozdelenia. Tak pre ľubovoľné *n* platí:

$$T = \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}}} \sqrt{n-2} \sim t_{n-2}.$$

Podľa vety 3. náš klasický test bude vyzerať nasledovne:

$$Hypotézu\ H_0\ zamietneme\ , ak\ T > t_{n-2;\frac{\alpha}{2}}\ alebo\ T < -t_{n-2;\frac{\alpha}{2}}$$

Zdrojom tejto podkapitoly bola kniha [1].

4.2.2.2. Výsledky počítačových simulácií

Najprv sa pozrieme ako fungujú naše testy pre malý počet dát. Prvé čo sme sledovali bola P(chyba 1.druhu).

P(chyba 1.druhu)		
n	PT	KT
5	0.041	0.046
6	0.053	0.052
7	0.048	0.048

Z tabuľky vidieť, že čísla sú blízko k hodnote *0.05,* čo potvrdzuje, že obidva testy sú spoľahlivé a nulovú hypotézu naozaj zamietajú v očakávaných *5 %*. Kvalitu klasického

testu ukazuje nasledujúca tabuľka, ktorá obsahuje silu testov. Nízku závislosť máme vtedy, keď $\rho < 0.5$ a vysokú keď $\rho > 0.5$ (ρ je vopred určená korelácia medzi X a Y).

	Nízká závislosť		**Vysoká závislosť**	
n	PT	KT	PT	KT
5	0.049	0.073	0.158	0.264
6	0.081	0.090	0.329	0.378
7	0.099	0.103	0.437	0.464

Vo všetkých prípadoch klasický test mal väčšiu silu zamietnuť nepravdivúhypotézu. Z tabuľky je však vidieť, že zvýšením počtu dát sa rozdiel medzi PT a KT stále zmenšuje. Možno to vysvetliť na základe vety 3. Keďže predpoklad normality máme splnený, klasický test funguje jednoducho dobre pre ľubovolný počet dát. V ďalšej tabuľke sú výsledky pre prípad, keď podmienka normality nie je dodržaná.

P(chyba 1.druhu)		
n	**PT**	**KT**
5	0.038	0.070
6	0.047	0.078
7	0.044	0.070

Tabuľka potvrdzuje citlivosť klasického testu, lebo pre dáta pochádzajúce z Cauchy-ho rozdelenia už nezamietne hypotézu H_0 v *5 %* (permutačný test pritom naďalej ostane na úrovni *5 %*). Testy, ktoré majú vysokú P(chyba 1. druhu) sa nazývajú *liberálnymi* testmi lebo majú veľkú silu (vôľu) zamietnuť nulovú hypotézu a prijať novú, alternatívnu hypotézu. Je preto zbytočné porovnávať silu testov, lebo nedostaneme reálny výsledok. Možno teda skonštatovať, že pri malom počte dát je klasický test lepší ako permutačný, ale len dovtedy, kým testované dáta pochádzajú z normálneho rozdelenia.

Teraz sa pozrieme, ako sa správajú naše testy pri veľkých počtoch dát (z normálneho rozdelenia N(0,1)).

P(chyba 1. druhu)						
n	**pp = 100**	**pp = 500**	**pp = 1000**	**pp = 2500**	**pp = 5000**	**KT**
8	0.054	0.052	0.047	0.047	0.048	0.048
9	0.050	0.046	0.052	0.049	0.050	0.052
10	0.055	0.051	0.043	0.049	0.050	0.053
15	0.047	0.051	0.046	0.050	0.052	0.049
20	0.043	0.051	0.050	0.044	0.053	0.048
25	0.050	0.053	0.047	0.051	0.053	0.055

Kvalita klasického testu sa nepokazí ani pri veľkom počte dát. V prípade permutačného výsledné hodnoty sa nám zdajú byť presnejšie. Kvalitu permutačného testu

s vyšším počtom permutácií potvrdzujú nasledujúce dve tabuľky, ktoré obsahujú silu jednotlivých testov.

Sila testu - Nízka závislosť						
n	**pp = 100**	**pp = 500**	**pp = 1000**	**pp = 2500**	**pp = 5000**	**KT**
8	0.095	0.101	0.102	0.111	0.110	0.105
9	0.092	0.115	0.116	0.119	0.124	0.116
10	0.113	0.130	0.128	0.137	0.131	0.140
15	0.166	0.188	0.183	0.199	0.190	0.197
20	0.197	0.240	0.239	0.251	0.257	0.254
25	0.261	0.301	0.310	0.310	0.315	0.316

Sila testu - Vysoká závislosť						
n	**pp = 100**	**pp = 500**	**pp = 1000**	**pp = 2500**	**pp = 5000**	**KT**
8	0.452	0.509	0.514	0.531	0.532	0.547
9	0.517	0.601	0.594	0.603	0.617	0.608
10	0.593	0.662	0.664	0.671	0.670	0.693
15	0.809	0.863	0.869	0.873	0.874	0.875
20	0.926	0.949	0.953	0.961	0.958	0.961
25	0.970	0.983	0.984	0.984	0.985	0.987

Ak použijeme pri permutačnom teste len *100* permutácií, má náš test vo všetkých prípadoch najmenšiu silu, naopak vo väčšine prípadoch „najsilnejší“ je test pri použití *5000* permutácií. Zvyšovaním počtu dát je tento rozdiel stále výraznejší. Porovnávanie permutačného testu (najmä, keď je pp veľké) s klasickým už nám nedá taký jednoznačný výsledok ako v prípade, keď bol počet dát malý. V niektorých prípadoch má väčšiu silu permutačný test a inokedy zase klasický, ale ani v jednom prípade nie je tento rozdiel výrazný.

Posledná tabuľka obsahuje P(chyba 1.druhu) pre nenormálne dáta. Permutačný test sa správa podobne ako pri normálnych dátach - vykazuje hodnoty okolo *0.05*. Klasický test rovnako ako pri menšom počte dát, zareaguje na „nenormalitu“ veľmi citlivo a jeho hodnoty sú ďaleko od *0.05* – stane sa liberálnym.

P(chyba 1. druhu)							
n	**pp = 100**	**pp = 250**	**pp = 500**	**pp = 1000**	**pp = 2500**	**pp = 5000**	**KT**
8	0.053	0.049	0.053	0.049	0.045	0.051	0.076
9	0.052	0.048	0.046	0.046	0.048	0.053	0.080
10	0.046	0.056	0.047	0.054	0.043	0.049	0.073
15	0.055	0.050	0.051	0.048	0.044	0.047	0.070
20	0.048	0.052	0.053	0.048	0.050	0.048	0.060
25	0.053	0.052	0.052	0.046	0.046	0.050	0.065

Na záver teda môžeme pri porovnávaní testov skonštatovať nasledovné

-v prípade Pearsonovho koeficienta nás náš permutačný test nesklamal: pri normálnych dátach bol podobne spoľahlivý ako klasický a aj pri zrušení normality zostal naďalej spoľahlivým.

-Klasický test pri normálnych dátach fungoval spoľahlivo , ale pri nenormálnych dátach jeho spoľahlivosť výrazne klesla.

4.2.3. Spearmanov korelačný koeficient

4.2.3.1.Teória

Nech dvojrozmerné dáta $(X_1, Y_1),\ldots, (X_n, Y_n)$ sú nezávislé a rovnako rozdelené, a $R(X) = (R_1,R_2,\ldots,R_n)$ a $Q(Y) = (Q_1,Q_2,\ldots,Q_n)$ sú prislúchajúce vektory poradia, tak

$$S_n = \sum_{i=1}^{n} R_i Q_i$$

sa nazýva *Spearmanova štatistika* a číslo

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(Q_i - \frac{n+1}{2})}{\sigma_R \sigma_Q}$$

kde

$$\sigma_R = \sigma_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(i - \frac{n+1}{2})^2}$$

sa nazýva *Spearmanov korelačný koeficient*.

Veta 4. Pre Spearmanov korelačný koeficient platí, že

$$\rho = 1 - \frac{6}{n(n^2 - 1)}\sum_{i=1}^{n}(R_i - Q_i)^2.$$

Veta 5. Ak $R = (R_1,R_2,\ldots,R_n)$, $Q = (Q_1,Q_2,\ldots,Q_n)$ sú náhodné vektory, ktoré sú nezávislé a rovnomerne rozdelené na $\mathcal{R}^n$, tak

$$E(\rho) = 0\,, \quad Var(\rho) = \frac{1}{n-1},$$

Spearmanov korelačný koeficient je symetricky rozdelený okolo nuly pre $n \to \infty$

$$\sqrt{n-1}\rho \sim N(0,1).$$

Testovaciu štatistiku T definujme nasledovne:

$$T = \sqrt{n-1}\rho,$$

potom podľa vety 5. klasický test skonštruujeme nasledovne:

$$H_0\ zamietneme\ , ak\ T > \mu_{\frac{\alpha}{2}}\ alebo\ T < -\mu_{\frac{\alpha}{2}}.$$

kde μ je kritická hodnota normálneho rozdelenia *N(0,1),* a α je hladina významnosti.

Zdrojom tejto podkapitoly bola kniha [2].

4.2.3.2. Výsledky počítačových simulácií

V simuláciach v tejto podkapitole sme použili iba dáta z normálneho rozdelenia. Prvé čo budeme pri testoch porovnávať bude P(chyba 1. druhu) pre rôzne počty dát. Z prvej tabuľky vidno, že pri malom počte dát dáva permutačný test výrazne lepšie hodnoty ako klasický test.

P(chyba 1. druhu)		
n	**PT**	**KT**
5	0.050	0.015
6	0.052	0.035
7	0.056	0.032
8	0.049	0.036

Klasický test má malé hodnoty (≈ 0.03). Takéto testy nazývame konzervatívne , lebo „ťažko“ zamietajú nulovú hypotézu. Tento výrazný rozdiel možno vysvetliť tým, že pri klasickom teste sme použili vzťah z vety 5., ktorý platí pre $n \to \infty$, čiže pri malom počte dát to nemusí platiť , a test môže byť nepresný. Možno teda očakávať, že zvýšením počtu dát by sa mal klasický test zlepšovať.

Výsledky počítačových simulácií potvrdili naše očakávaniach. Ako vidieť aj z nasledujúcej tabuľky, klasický test zamietol nulovú hypotézu H_0 v *5%.* Spoľahlivosť permutačného testu sa popri tom nezmenila

P(chyba 1. druhu)						
n	**pp=100**	**pp=500**	**pp=1000**	**pp=2500**	**pp=5000**	**KT**
10	0.053	0.052	0.045	0.048	0.050	0.048
20	0.042	0.056	0.050	0.047	0.055	0.044
30	0.048	0.051	0.048	0.047	0.049	0.052
40	0.052	0.057	0.049	0.052	0.056	0.047
50	0.051	0.053	0.052	0.046	0.047	0.052

Teraz porovnáme sily testov. Aj tu začneme najprv s malými počtami dát. Prvá tabuľka obsahuje výsledky permutačného testu a druhá výsledky klasického testu. Číslo rho označuje závislosť medzi X a Y (0.2 je najnižšia , 0.8 najvyššia). Predtým sme už zistili, že permutačný test je lepší. Kvalita permutačného testu sa prejaví aj pri porovnávaní síl. Permutačný test má vo všetkých prípadoch väčšiu silu zamietnuť nepravdivú hypotézu ako klasický test, čo potvrdzuje, že klasický test je konzervatívny.

Permutačný test				
n	**rho = 0.2**	**rho = 0.4**	**rho = 0.6**	**rho = 0.8**
5	0.035	0.037	0.059	0.133
6	0.052	0.077	0.143	0.315
7	0.067	0.123	0.241	0.499
8	0.068	0.135	0.279	0.596

Klasický test				
n	**rho = 0.2**	**rho = 0.4**	**rho = 0.6**	**rho = 0.8**
5	0.019	0.029	0.054	0.131
6	0.043	0.073	0.142	0.315
7	0.044	0.091	0.188	0.429
8	0.055	0.113	0.240	0.549

V prípade P(chyba 1. druhu) sa klasický test zvýšením počtu dát zlepšil, a preto očakávame, že pri väčšom počte dát bude sila klasického testu na úrovni permutačného testu.

Sila testu : rho = 0.2						
n	**pp=100**	**pp=500**	**pp=1000**	**pp=2500**	**pp=5000**	**KT**
10	0.065	0.077	0.071	0.078	0.074	0.073
20	0.091	0.115	0.116	0.123	0.130	0.110
30	0.137	0.172	0.156	0.161	0.170	0.159
40	0.169	0.208	0.201	0.221	0.211	0.216
50	0.222	0.241	0.257	0.252	0.248	0.258

Sila testu : rho = 0.4						
n	**pp=100**	**pp=500**	**pp=1000**	**pp=2500**	**pp=5000**	**KT**
10	0.142	0.169	0.169	0.177	0.174	0.169
20	0.301	0.360	0.352	0.373	0.378	0.376
30	0.475	0.532	0.531	0.555	0.546	0.536
40	0.617	0.667	0.679	0.674	0.684	0.682
50	0.719	0.767	0.783	0.780	0.785	0.790

Sila testu : rho = 0.6						
n	**pp=100**	**pp=500**	**pp=1000**	**pp=2500**	**pp=5000**	**KT**
10	0.326	0.381	0.383	0.394	0.394	0.386
20	0.682	0.746	0.752	0.760	0.765	0.771
30	0.875	0.922	0.925	0.920	0.923	0.921
40	0.958	0.971	0.977	0.981	0.976	0.981
50	0.987	0.992	0.996	0.996	0.994	0.993

Sila testu : rho = 0.8						
n	**pp=100**	**pp=500**	**pp=1000**	**pp=2500**	**pp=5000**	**KT**
10	0.679	0.736	0.738	0.752	0.752	0.742
20	0.969	0.987	0.983	0.989	0.985	0.988
30	0.998	0.999	1.000	1.000	1.000	0.999
40	1.000	1.000	1.000	1.000	1.000	1.000
50	1.000	1.000	1.000	1.000	1.000	1.000

Z tabuliek vidieť , že naše očakávania boli správne, a zvýšenie počtu dát vylepšil klasický test a zdá sa byť rovnako dobrým testom ako permutačný test s väčším počtom permutácii. Už v prípade Pearsonovho koeficienta sme zbadali, že pri permutačných testoch zvýšenie *pp* zvyšuje silu testu. Toto platí aj v prípade Spearmanovho koeficienta.

4.2.4. Kendallov koeficient korelácie

4.2.4.1. Teória

Nech (X, Y) je dvojrozmerný náhodný vektor. Ak sú vektory *(X_1,Y_1), (X_2,Y_2)* nezávislé a rozdelené ako *(X, Y)*, tak v označení

$$P^+ = P\big((X_1 - X_2)(Y_1 - Y_2) > 0\big), P^- = P\big((X_1 - X_2)(Y_1 - Y_2) < 0\big)$$

sa číslo

$$\tau = P^+ - P^-$$

nazýva Kendallovým koeficientom korelácie náhodných premenných *X, Y.*

Zaujímavé je, že v prípade normálne rozdelených vektorov existuje jednoznačná korešpondencia medzi Kendallovým τ a Pearsonovým korelačným koeficientom. Ak dvojrozmerná náhodná premenná *(X,Y)* má normálne rozdelenie $N(\mu,\sigma^2)$ s kladnými disperziami, tak Kendallov koeficient korelácie

$$\tau = \frac{2}{\pi}\arcsin(\rho),$$

kde arcsin nadobúda hodnoty z $\langle -\frac{\pi}{2}, \frac{\pi}{2}\rangle$ a ρ je Pearsonový korelačný koeficient.

Nech dvojrozmerné náhodné premenné $(X_1,Y_1),\ldots,(X_n,Y_n)$ sú nezávislé, a rozdelené rovnako, ako vektor *(X, Y)*. Potom štatistika

$$\tilde{\tau} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} sign(X_i - X_j)sign(Y_i - Y_j) =$$

$$= \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sign(X_i - X_j)sign(Y_i - Y_j)$$

sa nazýva Kendallovým koeficientom korelácie poradí.

Pri výpočte Kendallovho koeficienta korelácie poradí sa využíva aj pojem konkordantnosti a diskordantnosti poradí. Dvojice (x_1, y_1), (x_2, y_2) sa nazývajú konkordantné, ak

$$sign\big((x_1 - x_2)(y_1 - y_2)\big) = 1$$

a diskordantné, ak

$$sign\big((x_1 - x_2)(y_1 - y_2)\big) = -1\,.$$

Veta 6. Nech dvojrozmerné náhodné premenné $(X_1,Y_1),\ldots,(X_n,Y_n)$ sú nezávislé, a rozdelené rovnako, tak Kendallov koeficient korelácie poradí $\tilde{\tau}$ je symetricky rozdelený okolo nuly,

$$E(\tilde{\tau}) = 0\,, \quad Var(\tilde{\tau}) = \frac{2(2n+5)}{9n(n-1)}$$

a pre $n \to \infty$

$$\frac{\tilde{\tau}}{\sqrt{Var(\tilde{\tau})}} \sim N(0,1).$$

Testovaciu štatistiku T definujme nasledovne

$$T = \frac{\tilde{\tau}}{\sqrt{Var(\tilde{\tau})}}$$

potom podľa vety 6. klasický test možno skonštruovať nasledovne:

$$Hypotézu\ H_0\ zamietneme\ , ak\ T > \mu_{\frac{\alpha}{2}}\ alebo\ T < -\mu_{\frac{\alpha}{2}}.$$

kde μ je kritická hodnota normálneho rozdelenia N(0,1), a α je hladina významnosti.

Zdrojom tejto podkapitoly bola kniha [2].

4.2.4.2. Výsledky počítačových simulácií

V simuláciach v tejto podkapitole sme použili iba dáta z normálneho rozdelenia. Pravdepodobnosť chyby prvého druhu P(chyby 1. druhu) v prípade permutačného testu vyšla okolo 0.05. Klasický test sa správa veľmi zaujímavo: raz je test konzervatívny, inokedy zase liberálny.

P(chyba 1. druhu)		
n	**PT**	**KT**
5	0.050	0.015
6	0.039	0.061
7	0.050	0.030
8	0.046	0.062

Tento jav sa objavuje aj pri testovaný sily testu. Konzervatívny klasický test má menšiu silu ako permutačný test a liberálny klasický test má zase väčšiu silu ako permutačný test.

	Nízka Závislosť		**Vysoka Závislosť**	
n	**PT**	**KT**	**PT**	**KT**
5	0.035	0.019	0.058	0.054
6	0.036	0.074	0.088	0.210
7	0.047	0.039	0.173	0.172
8	0.051	0.083	0.223	0.331

Je teda jednoznačné, že klasický test je úplne nespoľahlivý. Túto vlastnosť klasického testu možno vysvetliť tým, že použitá testovacia štatistika má normálne rozdelenie (alebo aspoň približne) len pre veľký počet dát. Možno teda očakávať, že pri väčšom n sa klasický test bude zlepšovať, podobne ako v prípade Spearmanovho koeficienta

Pre väčšie n naozaj dostávame očakávané hodnoty. Permutačný aj klasický test má pravdepodobnosť chyby 1. druhu okolo 5%. Medzi testmi má najmenšie odchýlky od hodnoty 0.05 test s najvyšším počtom permutácií (viď. nasledujúcu tabuľku).

P(chyba 1. druhu)					
n	**pp = 100**	**pp = 500**	**pp = 1000**	**pp = 2500**	**KT**
10	0.053	0.056	0.047	0.049	0.049
15	0.049	0.055	0.053	0.049	0.050
20	0.043	0.057	0.047	0.049	0.042
25	0.053	0.049	0.050	0.049	0.055
30	0.048	0.050	0.047	0.047	0.052

Nasledujúce dve tabuľky ukazujú výsledky porovnania permutačného testu a klasického testu na základe ich sily.

Sila testu - Nízka závislosť					
n	**pp = 100**	**pp = 500**	**pp = 1000**	**pp = 2500**	**KT**
10	0.059	0.068	0.063	0.073	0.081
15	0.074	0.088	0.087	0.098	0.099
20	0.086	0.109	0.114	0.117	0.112
25	0.117	0.135	0.131	0.139	0.151
30	0.133	0.164	0.153	0.160	0.162

Sila testu - Vysoká závislosť					
n	**pp = 100**	**pp = 500**	**pp = 1000**	**pp = 2500**	**KT**
10	0.304	0.350	0.349	0.362	0.397
15	0.511	0.576	0.582	0.598	0.597
20	0.671	0.737	0.737	0.748	0.766
25	0.800	0.842	0.848	0.853	0.867
30	0.871	0.917	0.922	0.917	0.922

Nie je prekvapivé, že zvýšením počtu permutácií sa zvýši aj sila permutačného testu (platilo to aj pre Pearsona aj pre Spearmana). Zaujímavé je však v tomto prípade, že klasický test má, okrem dvoch prípadov, vždy väčšiu silu ako permutačný. Treba ale pripomenúť, že pri týchto testoch sme vynechali testy s pp=5000 , lebo výpočet Kendallovho koeficienta korelacie je numericky omnoho naročnejší ako v prípade Pearsona a Spearmana.

5. Programátorská časť

V tejto kapitole sú uvedené niektoré problémy a ich riešenia, ktoré sa vyskytli pri programovaní jednotlivých simulácií. Pri počítačových simuláciách sme používali programovací jazyk R.

5.1.Tvorba dát

Prvé, čo sme potrebovali, bol súbor náhodných dát. V R-ku existujú funkcie (rnorm, rcauchy) pomocou ktorých sa dajú vygenerovať nezávislé dáta normálneho a Cauchyho rozdelenia. Jediné čo sme potrebovali vyriešiť bolo vygenerovanie závislých dát. Problém sme riešili nasledovne:

Zvolili sme si kovariančnú maticu

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Pre kovariančnú maticu nezávislých dát X,Y platí:

$$Var\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, ak\ D(X) = D(Y) = 1.$$

Pomocou vlastnosti kovariančnej matice, sa závislé dáta U,V už ľahko vytvorili:

$$Var\begin{pmatrix} U \\ V \end{pmatrix} = Var(\ \Sigma^{1/2}\begin{pmatrix} X \\ Y \end{pmatrix}) = \ \Sigma.$$

Odmocninu z matice spravíme pomocou vťahu:

$$A^{1/2} = \ SD^{1/2}S^{-1}$$

kde D je diagonálna matica (na diagonále vlastné čísla matice A) a S je matica skladaná zo vlastných vektorov matice A.

Pri tejto metóde vieme pomocou premennej rho ľubovoľne zvoliť závislosť medzi dátami.

5.2.Vytvorenie permutácii

Ďalším problémom bolo vytvorenie permutácii. Pri testoch, kde sme použili len niekoľko permutácii, sme použili funkciu *sample*, ktorá náhodne vytvorí permutácie. Pri všetkých permutáciách sme pracovali s maticou, ktorá obsahovala všetky možne permutácie. Maticu permutácií sme vytvorili pomocou nasledujúcej idei:

Majme vytvorenú už maticu, ktorá obsahuje všetky permutácie n (pre znázornenie zoberme prípad n = 2)

$$\begin{matrix} 1 & 2 \\ 2 & 1 \end{matrix}$$

Keď pridáme ďalšiu hodnotu, tak stačí do už existujúcej matice uložiť novú hodnotu (ako stĺpec) na prvú, na druhú a nakoniec na poslednú priečku:

3 1 2
3 2 1

1 3 2
2 3 1

1 2 3
2 1 3

Takto získame maticu , ktorá už obsahuje všetky permutácie čísel 1,2,...,n.

5.3. Vytvorenie kombinácii

Podobne ako permutácie, je pre použitie 2-vyberoveho premutacneho t-testu je potrebné vytvorit aj kombinácie dát, čim myslíme všetky neusporiadané výbery "n" čísel z množiny {1,2,3,...,v}, ktorých je (v nad n).

Pri testoch, kde sme použili všetky kombinácii sme postupovali nasledovne:

```
KOM = function( n, v)
    if ( n > 1)
        G = KOM( n-1, v)
        prvi = 1
        m = v – n
        for i = 1 to m + 1
            for j=i to (v n − 1) .....kombinácia
            if  ( i < G_{j,1})
                A = ( i , j-ty riadok matice G)
                if (prvi = 0)  B = (B A)....matica, zložená z matic B a A
                if (prvi = 1)   B = A
                                prvi =0
    return(B)
    else
        P = (1, 2, 3,...., v)^T.....vektor
        return(P)
end
```

Pomocou uvedeného algoritmu dostaneme maticu, ktorá obsahuje všetky možné kombinácie.

Pri testoch, kde sme použili len niekoľko kombinácii, sme použili funkciu sample. Z pôvodných dát $X=(X_1,..., X_n)$, $Y=(Y_1,..., Y_m)$, sme permutáciou (spojenie –premiešanie-znovurozdelenie) vytvorili dva nové súbory dát $\dot{X}=(\dot{X}_1,..., \dot{X}_n)$, $\dot{Y}=(\dot{Y}_1,..., \dot{Y}_m)$. Je zrejmé, že takto môžu nastať aj také situácie, že súbory X a Y po permutácií budú

obsahovať tie isté dáta len v inom poradí (nevytvorili sme kombinácie ale variácie ,prípad a), čo je nedokonalosť tejto metódy)

$$P\hat{o}vodn\acute{y} : X = (a, b, c)\ \ Y = (d, e)$$

Po permutácií:

$$a)\ X = (c, a, b)\ Y = (e, d)\ toto\ je\ variácia$$

$$b) X = (c, e, b)\ Y = (a, d)\ toto\ je\ kombinácia$$

Napriek tomu, že táto metóda nie je dokonalá , vzhľadom na jej časovú jednoduchosť a dobré výsledky testov sme ju v našich simuláciách použili.

6. Záver

V tejto práci sme porovnávali vybrané štatistické testovacie metódy pomocou počítačových simulácií. Naším cieľom bolo nájsť take prípady, keď často používané štatistické metódy zlyhávajú a ukázať , že v týchto prípadoch možno ako alternatívu úspešne použiť permutačné metody. S takýmito situáciami sme sa stretli hlavne vtedy, keď testované dáta nepochádzali z normálneho rozdelenia alebo ich počet nebol dostatočne veľký. Na základe nami urobených počítačových simulácií môžme na riešenie týchto prípadov doporučiť použitie permutačných metód. Naše testy ukázali, že ich možno s úspechom použiť nielen vo výnimočných prípadoch, ale aj vtedy, keď je aj porovnávaná metóda dobrá, hoci v tomto prípade permutačné testy zľadom na ich časovo naročné výpočty nedoporučujeme.

Dalšim možným smerovaním tejto práce by mohlo byť podrobnejšie skúmanie týchto metód pri väčších počtoch opakovanií než 5000, čo bolo v tejto práci hornou hranicou vzhľadom na časovú náročnosť počítačových simulácií. Okrem toho by mohlo byť zaujímavé spraviť simulácie aj pre iné dáta než dáta pochadzajúce z normálneho a Cauchyho rozdelenia, ktoré boli v tejto práci použité.

Literatúra

[1] LAMOŠ, F. - POTOCKÝ, R. Pravdepodobnosť a matematická štatistika.

Bratislava, 1. vyd., Alfa 1989 ,2.vyd UK 1998, 343 s.

[2] RUBLÍK , F. 1993. Neparametrické metódy a štatistická kontrola akosti. 1. vyd. Bratislava : Univerzita Komenského, 1993. 184. S. ISBN 80 – 223 - 0540 – 5

[3] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.