UNIVERZITA PALACKÉHO V OLOMOUCI

PŘÍRODOVĚDECKÁ FAKULTA

KATEDRA GEOINFORMATIKY

Jiří CHLEBNÍČEK

STATISTICKÉ INFERENCE VE ZDRAVOTNICKÝCH DATECH

Bakalářská práce

Vedoucí práce: Mgr. Pavel Tuček, Ph.D

Olomouc 2010

Prohlašuji, že jsem zadanou bakalářskou práci řešil sám a že jsem uvedl veškerou použitou literaturu.

Lesnice, 22. 5. 2010 …….

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

Katedra: Geoinformatiky **Školní rok:** 2008-2009

ZADÁNÍ BAKALÁŘSKĚ PRÁCE

pro J i ř í C H L E B N Í Č E K

obor **Geografie a geoinformatika**

Název tématu: **STATISTICKÉ INFERENCE VE ZDRAVOTNICKÝCH DATECH**

ENG: STATISTICAL INFERENCE IN MEDICAL DATASETS

Zásady pro vypracování:

Student na základě dostupné literatury a za pomoci své invence, s odbornou poradou vedoucího bakalářské práce a konzultanta provede analýzu datasetů poskytnutých LF UP. Dle konzultací zjistí signifikantní závislosti pomocí korelační analýzy a kontingenčních tabulek. Otestuje hypotézy o shodě různých veličin (střední hodnoty, rozptyly atd.) ve skupinách dělených dle různých atributů (pohlaví, místo bydliště atd.) V průběhu tvorby bakalářské práce se student seznámí s pokročilejší teorií statistického zpracování dat. Základní poznatky z této teorie již student absolvoval a tudíž o ní sepíše pouze krátký informační úvod. Složitější partie a teorie, která není standardním obsahem předchozího studia, bude sepsána v další kapitole bakalářské práce. Tato část může být po odevzdání práce dále rozvíjena a použita pro tvorbu složitějších analýz zdravotnických dat. V závěru své bakalářské práce student prokáže svou schopnost odborného vyhodnocení získaného datového souboru, prokáže schopnost signifikantní vizualizace.

O bakalářské práci student vytvoří internetovou stránku, která bude v den odevzdání práce umístěna na server UP. Na závěr práce připojí jednostránkové 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. V první řadě to budou jednotlivé grafické výstupy programu , které budou zařazeny do základní části bakalářské práce. Dále bude student analyzovat získaný dataset a vytvoří vizuální podobu svých výsledků tak, aby běžný uživatel resp. čtenář, který má základní znalosti z oboru, mohl na základě takto předložených vizualizací usoudit, jaký problém byl zpracovávám 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 30 - 40 stran vlastního textu, který bude obsahovat zejména tři stěžejní kapitoly, které budou obsahovat úvodní pojednání s rešerší dané problematiky, teoretické podklady s použitou teorií a vlastní zpracování praktického zadání. Dále samozřejmě bude práce obsahovat přílohy, které budou složeny z výstupů daného programového vybavení a z programového kódu, který si student sám napíše za účelem automatizace výpočtů apod.

Seznam odborné literatury:

Anděl, Jiří. Statistická analýza časových řad. Praha : SNTL, 1976

Anděl, Jiří. Základy matematické statistiky. Praha: MFF UK 2005

Cleophas, T., J. et al.. Statistics Applied to Clinical Trials. Kluwer Academic Publishers.2000

Liu, J.-Chow, S. Design and Analysis of Clinical Trials: Concepts and Methodologies. 1998

Machin, D. et al. Sample size tables for clinical studies. Blackwell Science. 1987

McFaccen, E. Management of Data in Clinical Trials. John Wiley and Sons. 1998

Meinert, C. L. Clinical Trials: Design, Conduct and Analysis. Oxford University Press. 1996

Norleans, M. X. Statistical methods for clinical trials. Marcel Dekker. 2001

Piantadosi, S. Clinical Trials: A Methodological Perspective. John Wiley and Sons. 1986

Shuster, J., J. Handbook of sample size guidelines for clinical trials. CRC Press. 1990

Wooding, W., M. Planning pharmaceutical clinical trials. John Wiley and Sons. 1994

Vedoucí bakalářské práce: Mgr. Pavel TUČEK

Konzultant bakalářské práce: Prof. RNDr. Vít VOŽENÍLEK, CSc.

Datum zadání bakalářské práce: červen 2008

Termín odevzdání bakalářské práce: duben 2009

Vedoucí katedry Vedoucí bakalářské práce……..

V Olomouci dne

OBSAH

ÚVOD

Téma bakalářské práce vzniklo v podstatě náhodou, když probíhala diskuze s lékařskou fakultou Univerzity Palackého v Olomouci ohledně poskytnutí dat k původně jinak zaměřené práci. Vzhledem k současné legislativě však téma nebylo možno zpracovat, tak byla navržena jiná varianta, která je náplní této práce.

Pojem statistické inference lze chápat jako odvozování, tvoření hypotéz. V této práci bylo odvozováno od zdravotnických dat reprezentujících projevy určité nemoci (incidence, smrt) a fyzickogeografickému nebo socioekonomickému jevu, který je nositelem předpokladu, že by mohl být s daty nemoci ve vzájemném vztahu.

Vzhledem ke geografickému zaměření studia jsou zpracovávaná data vázaná k určitým prostorovým jednotkám. Tato data byla volně dostupná ve veřejných statistických výkazech.

Bakalářská práce je zaměřena hlavně na nádorová onemocnění. Hlavním zaměřením práce je především jasná interpretace dat tak, aby výsledky statistických analýz byly jasně pochopitelné nejen odborně zaměřené veřejnosti.

V době, kdy se uzavíralo zadaní bakalářské práce, probíhala stále diskuze s Lékařskou fakultou Univerzity Palackého ohledně dat, kdy původní záměr na bakalářskou práci byl odlišný. Po uzavření zadání teprve došlo ke konkretizaci tématu a tedy i vstupních dat. V zadání se předpokládalo, že bude sledováno více znaků pro jedince, dostupná data však byla v sumarizované podobě, kdy byla sledována míra dané nemoci, která byla korelována s jevem, jenž s onemocněním vykazuje potencionální souvislost. Tento fakt vylučuje užití hypotéz o shodě veličin mezi znaky. Vzhledem k metodám, které bývají v rámci předmětu bakalářské práce užívány, byla závislost určována především pomocí korelačních koeficientů.

1. CÍLE PRÁCE

Hlavní náplní bakalářské práce je zhodnotit závislost mezi vybranými jevy a zdravotnickými daty reprezentující projev určité nemoci. Tomu však předchází několik kroků.

V prvé řadě je třeba konzultace s Lékařskou fakultou Univerzity Palackého v Olomouci (prof. Mihál), se kterým budou s ohledem na evidenci vybrána data, která budou analyzována. Základním předpokladem je, aby data vybraných jevů měla hypotetickou souvislost k danému projevu onemocnění.

Dále musí být u dat zajištěn předpoklad věkové standardizace zdravotnických dat, aby byl vyloučen jev různé věkové struktury.

Data budou zpracována do textových datasetů ve formě akceptovatelné softwarem R, ve kterém budou analýzy zpracovány.

Budou provedeny základní statistické charakteristiky, vyjadřující základní popis datasetů. Pomocí korelačních koeficientů a testováním nezávislosti bude provedena statistická analýza datasetů, kdy závěrem bude vyjádření o (ne)závislosti mezi daty.

Data budou vizualizována v grafech pro jejich jasnou schematičnost.

Závěrem budou vytvořeny internetové stránky, které budou umístěny na serveru katedry. Data budou zanesena do metainformačního systému MICKA.

2. TEORETICKÁ ČÁST

2.1. Epidemiologie

2.1.1. Charakteristika epidemiologie

Hlavní disciplínou, jež se zabývá metodikou použitou v této práci, je epidemiologie. Jedná se o interdisciplinární populační vědu.

Definice: Epidemiologie studuje rozložení (distribuci) a příčiny (determinanty) frekvence nemocí v lidské populaci. ([2], str. 8)

Obor epidemiologie předpokládá, že nemoc u jedince nevzniká náhodou a také existují příčinné (kauzální) a prevenční faktory, jež je možné identifikovat u různých populací, různých místech a v různých obdobích.

Hlavní úloha epidemiologie v medicíně lze formulovat v několika hlavních bodech:

- měřit frekvenci nemocí
- popisovat charakter výskytu nemocí
- vyšetřovat epidemie nemocí
- provádět surveillance u vybraných nemocí
- zhodnotit přesnost diagnostických testů
- určit příčiny vzniku nemocí
- vyhodnotit efektivitu léčby
- určit prognózu nemocí

([2], str. 5)

Z hlediska druhu nemocí je možno epidemiologii dělit na:

- epidemiologie infekčních nemocí
- epidemiologie neinfekčních nemocí

V minulosti byly hlavní náplní epidemiologie především infekční nemoci, ale s postupem doby a moderní lékařské péče se epidemiologie začínala zaměřovat na nemoci neinfekční, které jsou závislé kromě genetických vloh na faktorech životního stylu, životního prostředí, sociálních a pracovních podmínkách.

2.1.2. Charakter epidemiologických studií

- Deskriptivní studie
  - populační studie (korelační)
  - jednotlivci
  - kazuistika
  - popis série onemocnění
  - průřezové studie
- Analytické studie
  - observační
  - studie případu a kontrol
  - kohortové studie
  - intervenční
  - klinický pokus

([2], str. 15)

Bakalářská práce je zaměřena tedy na deskriptivní populační (korelační) studie.

2.1.3. Korelační studie

Popisuje nemoc ve vztahu k některému faktoru osoby, místa a času a používají údajů z celé populace. Mírou asociace mezi přepokládaným rizikem a nemocí je korelační koeficient, jehož hodnota se pohybuje od -1 do +1. ([2], str. 28)

Největší předností těchto studií je jejich relativní celková nenáročnost vůči ostatním epidemiologickým metodám, které jsou náročné především na časovou a personální složku. Data jsou obecně dostupná, a to i pro širokou veřejnost díky

statistickým výkazům, jenž jsou zdravotnickými, ale i jinými organizacemi publikovány v sítí Internet v rámci statistických databází.

Naopak nevýhodou korelačních studií je fakt, že data vstupující do analýz jsou sumarizována za populaci, tudíž jejich výsledky nemohou být vztahovány na jedince, jenž nemusí být nositelem asociativního vztahu i přes silnou korelaci. Toto tvrzení je objasněno v následující tabulce (Tabulka 1), kde je znázorněna struktura tří fiktivních populací vstupujících do korelační analýzy.

Tabulka 1.: Struktura fiktivní populace

	B – jednotlivec s brýlemi			R – jednotlivec s rakovinou		
označení jedince v populaci	1	2	3	4	5	6
populace 1	B	-	-	-	-	R
populace 2	B	B	-	-	R	R
populace 3	B	B	B	R	R	R

V první populaci je přítomna jedna osoba, která nosí brýle a jedna osoba, která je nemocna rakovinou, ve druhé populaci jsou dvě osoby s brýlemi a dvě s rakovinou, ve třetí populaci tři s brýlemi a tři s rakovinou. Pokud by tato data byla statisticky vyšetřována v sumarizované podobě, tak by závěr této studie byl, že je zcela přímá úměrnost mezi nošením brýlí a onemocněním rakovinou (korelační koeficient je roven +1). Z tabulky je ovšem patrné, že jedinci, kteří nosí brýle však rakovinu nemají. To znamená, že přestože by statisticky bylo možné prohlásit, že vztah mezi nošením brýlí a rakovinou je maximálně pravděpodobný, tak tato skutečnost v těchto fiktivních datech samozřejmě neplatí.

Případ zmiňovaný výše je samozřejmě uveden na extrémním případu a v reálném světě evidentně nesouvisejících datech, je z něj však patrná skutečnost, kterou je potřeba při těchto analýzách zohledňovat, zvláště pak při interpretaci získaných výsledků.

U dat, kde je uveden průměrný údaj za populaci (např. průměrná spotřeba alkoholu) nejsou známy poměry jeho konzumace - v populaci může pít velký počet jedinců z populace menší množství alkoholu nebo naopak úzká skupina může vypít velké množství, ale ve zprůměrovaném výsledku se může jednat o podobnou hodnotu. Efekt na zdravotní stav obyvatelstva v daných populacích to však bude mít pravděpodobně odlišný.

Dalším úskalím je vliv třetího faktoru. Vysvětlení je uvedeno na příkladu pozitivní korelace mezi příjmem masa a nádorovým onemocněním prsu. Výsledkem totiž nemusí být přímo konzumace masa, ale jiný faktor, který je skrytý (zvýšený příjem živočišných tuků, snížený příjem zeleniny, vyšší životní úroveň). [10]

Jak již bylo zmíněno výše, tak tyto studie jsou prováděny kvůli jejich jednoduchosti a je z nich možné tvořit hypotézy o možných deterministických faktorech, které mohou onemocnění vyvolat. Hypotézy na úrovni jednotlivce je možno ověřovat analytickými studiemi epidemiologie, převážně klinickými pokusy. Samozřejmě postupy pro získání výsledků jsou mnohem složitější ve všech jejich aspektech. Metody, které bývají v klinických pokusech aplikovány jsou uvedeny v literatuře [7][11][14][15][16][24].

2.2. POPISNÁ STATISTIKA

2.2.1. Míry polohy

2.2.1.1. Průměr

Průměr (aritmetický průměr) používáme, když čísla můžeme opravdu sčítat, tj. znaky jsou kvantitativní, měřené na číselné stupnici. Neměl by být používán pro ordinální znaky vzhledem k libovůli při volbě ordinální stupnice. Je rovněž velmi citlivý na odlehlé hodnoty. Průměr z hodnot ve výběru vypočítáme, jestliže součet všech hodnot dělíme rozsahem výběru (n). Máme-li tedy n pozorování, pak průměr počítáme následujícím způsobem. Součet pozorování se značí symbolem:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots x_n \tag{1}$$

Počet pozorování je n. Průměrem je:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

2.2.1.2. Modus

Modus je hodnota, která se v souboru dat vyskytuje nejčastěji. Důležitý je pro kvalitativní, zejména nominální znaky.

2.2.1.3. Medián

Máme-li pozorování uspořádána vzestupně nebo sestupně, potom *medián* je ta hodnota, která rozdělí pozorování na dvě stejně velké skupiny. Přesněji řečeno, máme-li lichý počet uspořádaných pozorování, pak mediánem je prostřední z nich.

U sudého počtu se mediánem rozumí obvykle průměr ze dvou prostředních pozorování. Medián využívá pouze informaci o pořadí hodnot, a proto ho má smysl používat pouze pro kvantitativní a ordinální veličiny.

2.2.2. Míry variability

2.2.2.1. Rozpětí

Jednou z měr variability je *rozpětí (variační šíře) R*, což je rozdíl mezi nejvyšší a nejnižší hodnotou v datech, tj.

$$R = x_{max} - x_{min} \tag{3}$$

2.2.2.2. Rozptyl

Rozptyl s^2 je průměr čtverců odchylek od průměru. Když však počítáme výběrový rozptyl, nedělíme většinou součet čtverců odchylek výrazem *n*, ale *(n-1)*, protože tím docílíme lepšího odhadu celkového rozptylu populace. Dělitel *(n-1)* se nazývá *počet stupňů volnosti* rozptylu. Obecný vzorec tedy vypadá takto:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4}$$

2.2.2.3. Směrodatná odchylka

Druhá odmocnina z rozptylu se nazývá *směrodatná odchylka s*. Směrodatná odchylka je používána častěji než rozptyl. Poznamenejme, že směrodatná odchylka *s* je ve stejných jednotkách jako původní hodnoty. ([4] str. 79-83)

2.2.3. Korelace

V bakalářské práci je užito korelace jako hlavní metody k určování statistické závislosti. Je však třeba rozlišovat pojmy korelace a lineární regrese, která je také obsahem praktické části této práce a to v podobě vizualizovaného naznačení vztahu mezi daty v grafech. Zjednodušeně lze regresi považovat za metodu, která vystihuje průběh závislosti mezi daty (nemusí být jen lineární), zatímco korelace (korelační koeficient) vystihuje sílu tohoto lineárního vztahu.

2.2.3.1. Korelační koeficient

Pearsonův korelační koeficient

Pro měření síly *lineární* závislosti mezi dvěma spojitými náhodnými veličinami se používá tzv. *(Pearsonův) korelační koeficient*. Počítá se podle vzorce:

$$r = \frac{s_{xy}}{\sqrt{s^2{}_x s^2{}_y}} \tag{5}$$

kde s_x^2 a s_y^2 jsou výběrové rozptyly a s_{xy} je kovariance.

Hodnota korelačního koeficientu se pohybuje od -1 do 1. Hodnoty ±1 nabývá tehdy, pokud všechny body $[x_i, y_i]$ leží na přímce. Nule je roven v případě, že veličiny jsou nezávislé. Korelační koeficient však může být nulový i v případě, že veličiny jsou funkčně závislé, ale závislost není lineární. Proto je při užití Pearsonova korelačního koeficientu vždy třeba posoudit, zda je jeho aplikace vhodná. Při měření lineární závislosti je znaménko korelačního koeficientu kladné, když obě veličiny *X* a *Y* zároveň rostou nebo obě zároveň klesají, a záporné, když jedna z veličin roste, zatímco druhá klesá. ([4], str. 178)

Spearmanův korelační koeficient

Neparametrické metody založené na pořadích rozšíříme nyní o metodu, která využívá pořadí při zjišťování závislosti dvou znaků. Dobrým ukazatelem takové závislosti je *Spearmanův korelační koeficient*, založený na pořadích jedinců

uspořádaných podle velikosti vzhledem ke dvěma sledovaným veličinám. Každému jedinci se přiřadí dvojice pořadí Q (pořadí podle první veličiny X) a R (pořadí podle druhé veličiny Y). Kdyby s rostoucími hodnotami X vzrůstaly i hodnoty Y, byla by zřejmě pořadí obou veličin shodná, tj. $Q = R$ pro každého jedince. Jestliže s rostoucími hodnotami X klesají hodnoty Y, jsou pořadí obou veličin právě opačná. Při nezávislosti jsou pořadí zpřeházená zcela náhodně. Pro n pozorovaných dvojic ve výběru se Spearmanův korelační koeficient počítá pomocí diferencí pořadí $d_i = Q_i - R_i$ jako:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2 - 1)} \tag{6}$$

Při shodném pořadí dosahuje koeficient r_S maximální hodnoty +1, při opačném pořadí minimální hodnoty -1. Hodnoty korelačního koeficientu blízké nule naznačují, že pořadí jsou náhodně zpřeházená, a mezi sledovanými veličinami tedy není závislost. Při platnosti nulové hypotézy o nezávislosti obou veličin jsou odchylky Spearmanova korelačního koeficientu od nuly jen náhodné. Když tedy absolutní hodnota Spearmanova korelačního koeficientu $|r_S|$ překročí 5% nebo 1% kritickou hodnotu, zamítá se nulová hypotéza o nezávislosti na příslušné hladině významnosti. ([4], str. 185)

2.2.3.2. Koeficient parciální korelace

Korelační koeficient měří závislost dvou náhodných veličin. Tato závislost však nemusí být příčinná, ale může promítat vliv dalších doprovodných veličin. Mějme dvě náhodné veličiny Y a Z a nějaký náhodný vektor $\mathbf{X} = (X_1,..., X_n)'$. Připusťme, že $\mathbf{X}$ může působit na Y i na Z. Ptejme se však, jaká by byla závislost mezi Y a Z, kdyby na ně $\mathbf{X}$ nepůsobil. Jednu z možností, jak to zjistit, je vytvořit takovou situaci, kdy vektor $\mathbf{X}$ zůstává konstantní. To však ve většině případů nelze, a tak nezbývá, než přistoupit k matematickému řešení. Nejlepším lineárním přiblížením k veličině Y je $\hat{Y} = \alpha + \beta'\mathbf{X}$. Proto $Y - \hat{Y}$ můžeme interpretovat jako tu část veličiny Y, která je očištěna od vlivu vektoru $\mathbf{X}$. Podobně nechť $\hat{Z} = \chi + \delta'\mathbf{X}$ je nejlepší lineární aproximace veličiny Z založená na $\mathbf{X}$.

Budiž $\mathbf{V}$ symetrická pozitivně semidefinitní matice typu $n \times n$. Pak pro libovolné *n*-rozměrné vektory $\mathbf{b}$ a $\mathbf{c}$ platí

$$(\mathbf{b}'\mathbf{V}\mathbf{c})^2 = (\mathbf{b}'\mathbf{V}\mathbf{b})(\mathbf{c}'\mathbf{V}\mathbf{c}) \tag{7}$$

$$\delta = \mathbf{V}^{-1}\operatorname{cov}(\mathbf{X}, Z), \quad \gamma = \mathsf{E}Z - \delta'\mathsf{E}\mathbf{X} \tag{8}$$

Tu část veličiny *Z*, kterou vektor $\mathbf{X}$ nevysvětlí, si můžeme představit jako $Z - \hat{Z}$. Proto závislost mezi Y a Z, je-li eliminován vliv vektoru $\mathbf{X}$, měříme korelačním koeficientem mezi $Y - \hat{Y}$ a $Z - \hat{Z}$. Říkáme mu koeficient parciální korelace mezi Y a Z při daném $\mathbf{X}$ a značíme ho $\rho_{Y,Z.\mathrm{x}}$.

Nechť náhodné veličiny $Y, Z, X_1, \ldots, X_n$ mají konečné druhé momenty a regulární varianční matici. Označme $\mathbf{P} = cor\mathbf{X}$. Pak platí

$$\rho_{Y,Z,\mathrm{x}} = \frac{\rho_{YZ} - cor(Y,\mathbf{X})\mathbf{P}^{-1}cor(\mathbf{X},Z)}{\sqrt{[1 - cor(Y,\mathbf{X})\mathbf{P}^{-1}cor(\mathbf{X},Y)][1 - cor(Z,\mathbf{X})\mathbf{P}^{-1}cor(\mathbf{X},Z)]}} \tag{9}$$

2.2.3.3. Koeficient mnohonásobné korelace

Mějme náhodnou veličinu Y a náhodný vektor $\mathbf{X} = (X_1, \ldots, Xn)'$ podobně jako u lineární regrese. Závislost mezi Y a celým vektorem $\mathbf{X}$ měříme pomocí koeficientu mnohonásobné korelace $\rho_{Y,\mathrm{x}}$, což je korelační koeficient mezi Y a veličinou $\alpha + \boldsymbol{\beta}'\mathbf{X}$, kde α a $\boldsymbol{\beta}$ jsou uvedeny v ([2], (36)) Je-li $\boldsymbol{\beta} = 0$, definuje se tento koeficient mnohonásobné korelace jako nula. Jde tedy o korelaci mezi veličinou Y a její nejlepší lineární náhradou založenou na $\mathbf{X}$. Někdy se místo $\rho_{Y,\mathrm{x}}$ píše $\rho_{Y,(X_1,\ldots,X_n)}$. Protože platí

$$\rho_{Y,\mathrm{x}} = \rho_{Y,\alpha+\beta\mathrm{x}} = \rho_{Y,\beta'\mathrm{x}}$$

a

$$\operatorname{cov}(Y, \boldsymbol{\beta}'\mathbf{X}) = \operatorname{cov}(Y,\mathbf{X})\boldsymbol{\beta} = \operatorname{cov}(Y,\mathbf{X})\mathbf{V}^{-1}\operatorname{cov}(X,Y) \geq 0,$$

je koeficient mnohonásobné korelace vždy nezáporný.

Označme $\mathbf{P} = cor\mathbf{X}$. Pak platí

$$\rho^2_{Y,\mathrm{x}} = cor(Y,\mathbf{X})\mathbf{P}^{-1}cor(\mathbf{X},Y). \tag{10}$$

2.3. Rešerše literatury

Studie [20] řeší v rámci 47 okresů státu Florida korelační závislost mezi rakovinou slinivky břišní a faktory: kouření, finanční příjem, objem pevného odpadu.

Jako zdroj dat za územní jednotky byl využit Florida Cancer Data System (FCDS) bílého obyvatelstva: muži, ženy, celkově za období 1981–1994, dále pak medián příjmu domácnosti, prevalence kuřáků a objem pevného odpadu na jednoho obyvatele. Faktor objemu pevného odpadu byl dále rozdělen na stavební odpad, jídlo, papír, rostlinný a nebezpečný odpad.

Z analýzy vyšla jako statisticky nesignifikantní korelace s nebezpečným odpadem ($r = 0,11$; $p = 0,40$). naopak jako statisticky významné byly korelace rakoviny slinivky (pro obě pohlaví zvlášť i dohromady) s prevalencí kuřáků, příjmem domácnosti i objemem pevného odpadu. V rámci skupiny pevného odpadu nebyly statisticky významné korelace se stavebním odpadem potravinovým odpadem. Naopak významné byly korelace s papírovým odpadem ($r = 0,25$) a rostlinným odpadem ($r = 0,42$).

Ve studii [8] je zkoumána závislost radiačního záření v souvislosti s nehodou jaderné elektrárny v Černobylu (Ukrajina) na rakovinu štítné žlázy v 25 administrativních regionech na Ukrajině a Krymu. Vyšetřován je vztah mezi mírou kontaminace v dané oblasti k incidenci. Nelineární podoba křivky závislosti je okomentována možným důvodem adaptace organismu člověka. Dále je ve studii uveden graf korelačních koeficientů kontaminace a incidence, kde bod v grafu reprezentuje korelační koeficient pro určitý věk v mužské populaci. V tomto grafu jsou zvlášť hodnoty pro oblasti se zamořením do 6 t/km^2 a nad tuto hranici.

Nejvýznamnější korelace byla zjištěna kolem dvanáctého roku v oblastech s vyšší kontaminací ($r = 0,67$). V oblastech s vyšší kontaminací vykazují vyšší korelační koeficienty skupiny okolo 37 let ($r = 0,6$) a skupiny věku 50–80 (koeficienty 0,6–0,7).

V rámci studie [12] je řešena úmrtnost na nádorová onemocnění v souvislosti s jevem chudoby v 25 provinciích státu Argentina. Jako zdroj dat sloužila národní databáze s daty z let 1980–1986. Tato data byla podrobena přímé standardizaci. Pro data určující míru socioekonomického statutu byl využit národní cenzus z roku 1980. Zde bylo dle kritérií za každou domácnost vyhodnoceno, zda domácnost vyhovuje určenému

standardu. Data, která vstoupila do analýzy představují poměr standardních a nestandardních (chudších) domácností v rámci provincie.

Nejprve byla analyzována korelace na úmrtnost na jednotlivá nádorová onemocnění.mezi pohlavími. Pozitivní signifikatní korelace vyšly u rakovinného onemocnění: tlustého střeva, jater, slinivky břišní, jícnu, žaludku, žlučníku a močového měchýře. Při korelaci standardizovaných úmrtností se socioekonomickým indikátorem (vyšší číslo = vyšší úroveň chudoby) vyšly jako negativní signifikatní korelace rakoviny: prsu a vaječníku u žen, slinivka břišní a tlusté střevo u obou pohlaví a močový měchýř u mužů. Pozitivní korelace byly zjištěny u nádorových onemocnění všech ženských orgánů.

Při užití metody parciální korelace mezi jednotlivými nádorovými onemocněními v rámci pohlaví, kdy jako třetí proměnná vstupuje socioekonomický status, došlo ke změnám hodnot parciálních korelačních koeficientů vůči Pearsonovým korelačním koeficientů, do kterých vliv socioekonomických statutů nevstupoval. Jediné hodnoty, které si statisticky odpovídaly i po odstranění faktoru socioekonomického statutu byly u rakoviny plic u mužů v kombinaci s rakovinami: močového měchýře, hrdla a slinivky břišní. Z této analýzy vyplynulo, že socioekonomický statut hraje v souvislosti mezi onemocněními významnou roli.

Korelační epidemiologická studie [1] vyšetřuje pomocí Pearsonova korelačního koeficientu vztah mezi kumulativní incidencí cholery ve státech latinské Ameriky (22 zemí včetně USA a Kanady) v letech 1991–1995 a socioekonomickými a demografickými faktory (kojenecká úmrtnost; Human Developement index (HDI) – hodnota založená na faktorech délky života, vzdělanosti a příjmu; hrubý domácí produkt (HDP) na obyvatele; vzdělanost u žen. Zjištěna byla pozitivní signifikantní korelace incidence cholery s kojeneckou úmrtností ($r = 0,55$) a signifikantní negativní korelace s HDI ($r = -0,38$). Faktory HDP a gramotnosti u žen se jevily jako nesignifikantní.

V rámci grantového projektu NR8480: Konstrukce socioekonomického deprivačního indexu pro analýzu rutinně sbíraných dat o zdravotním stavu populace s možností využití GIS (2005–2007) byla realizována studie [21], kde bylo v rámci 77 okresů ČR cílem identifikovat vztah mezi zdravotními charakteristikami a dostupnými socioekonomickými daty u obou pohlaví.

Jako zdravotnická data figurovala incidence pro všechny novotvary a pro diabetes, celková standardizovaná míra úmrtnosti (SMR) a standardizované míry úmrtnosti pro: všechna nádorová onemocnění, nádorů prsu žen, nádorů plic mužů; onemocnění kardiovaskulární, respirační a onemocnění gastrointestinálního traktu. Tyto charakteristiky byly analyzovány podle: váženého průměru úrovně vzdělání v každém okrese, složením domácností (proporce úplných a neúplných rodin nebo jednotlivců žijících osaměle) a hustotou bydlení, průměrných příjmů, celkové míry nezaměstnanosti a počtem lékařů na 1000 obyvatel. Zdravotní data jsou k roku 2001, stejně tak jako socioekonomické údaje, které byly získány v rámci Censu.

Jako statisticky signifikatní se jevily korelace: vzdělání–(celková SMR, SMR na nádory a nádory plic), rodinný stav–(celk. SMR, SMR na nádory a nádory prsu), míra nezaměstnanosti–(celková SMR a SMR na nádory, kardiovaskulární a onemocnění gastrointestinálního traktu a incidence nádorů). Naopak faktory: hustota bydlení, průměrný příjem a přístup ke zdravotní péči (počet lékařů na 1.000 obyvatel) nebyly statisticky signifikantní vůči zdravotnímu stavu.

Článek [5] řeší možnou asociaci standardizované úmrtnosti na cévní mozkové příhody s faktory: konzumace potravin (mléko, maso, vejce), příjem nasycených tuků, spotřeba cigaret, prevalence cukrovky, prevalence osob s vysokým tlakem, prevalence osob s BMI>30, míra nezaměstnanosti, míra negramotnosti, sedavé zaměstnání a konzumace vína v 50 provinciích Španělska. Španělsko vykazuje velké rozpětí hodnot standardizované úmrtnosti na cévní mozkové příhody v rámci provincií (70/100000–180/100 000).

Jako zdravotnická data byla využita standardizovaná úmrtnost ve věkové skupině 45–79 let z období 1989–1993. Pro všechny faktory byl vypočten Pearsonův korelační koeficient. Jako statisticky významné byly zaznamenány vztahy mezi cévní mozkovou příhodou a: mírou negramotnosti, sedavým zaměstnáním a konzumací vína. U posledního případu ovšem došlo k jevu, kdy v oblastech s nízkou konzumací byl faktor spotřeby vína jako protektivní faktor (negativní korelace), avšak v provinciích s vyšší spotřebou byla korelace pozitivní.

Studie [9] se zabývá vztahem mezi incidencí zhoubného nádoru kůže (melanomem) a zeměpisnou šířkou respektive UV indexem ve vybraných státech USA. Jako zdroj zdravotnických dat posloužila databáze SEER Národního nádorového

institutu z let 1992–2001. Věkově standardizovaná data jsou rozlišena podle pohlaví a etnické příslušnosti. Zeměpisná šířka byla stanovena dle největšího města ve státě. Hodnota UV indexu byla určena jako průměrná roční hodnota k danému státu. Byly vypočteny Pearsonovy korelační koeficienty pro onemocnění melaninem k rasové a pohlavní příslušnosti k faktorům zeměpisné šířky a UV indexu.

Jevy UV index a zeměpisná šířka spolu velmi silně statisticky významně korelovaly ($r = -0{,}97$; $p = 0{,}001$). Jako statisticky významná se s UV indexem jevila pozitivní korelace pouze u bílého obyvatelstva ($r = 0{,}85$; $p = 0{,}001$), statisticky nevýznamná vyšla korelace u původního obyvatelstva ($r = 0{,}42$; $p = 0{,}20$). Byly pozorovány také negativní ale nesignifikantní korelace u černošského obyvatelstva ($r = -0{,}53$; $p = 0{,}10$), hispánského ($r = -0{,}43$; $p = 0{,}19$) a asiatů ($r = -0{,}28$; $p = 0{,}41$). Vhledem k silné korelaci mezi zeměpisnou šířkou a UV indexem byla statisticky signifikantní negativní korelace u bělošského obyvatelstva a zeměpisnou šířkou ($r = -0{,}85$; $p = 0{,}001$).

3. PRAKTICKÁ ČÁST

3.1. Použitá data

Jedním z nejdůležitějších úkolů této bakalářské práce bylo vhodné zvolení dat, která budou vstupovat do analýz. Proto bylo toto téma konzultováno právě s Lékařskou fakultou Univerzity Palackého (prof. Mihál), který pomohl s výběrem dat.

Pro Českou republiku byla data čerpána z Českého statistického úřadu (ČSÚ), kde se jednalo o data znečištění ovzduší za kraje v roce 2005, dále byla z tohoto zdroje použita data úmrtnosti na nádorová onemocnění v okresech České republiky v roce 2006 a průměrný věk v okresech z roku 2006. Z dat na portálu Ústavu zdravotnických informací a statistiky (ÚZIS) byla vybrána standardizovaná incidence nádorových onemocnění plic za kraje z roku 2005.

Pro státy Evropy eviduje Světová zdravotnická organizace (WHO) za jednotlivé státy poměrně obsáhlou databázi jevů (databáze WHOSIS), jež se vztahují ke zdravotnictví. Zvolena byla data: prevalence dospělých kuřáků ve vybraných státech Evropy (zjišťováno v roce 2005), spotřeba cigaret na jednu osobu za jeden rok (databáze Tobacco control database) evidovaná ve většině zemí pro rok 2000 (vybrané státy Evropy) a standardizovaná úmrtnost na nádorová onemocnění plic (rok 2002, vybrané státy Evropy). Dále to byla průměrná spotřeba alkoholu na jednu osobu za rok 2003 (vybrané státy Evropy) a standardizovaná úmrtnost na ischemickou chorobu srdeční za rok 2003 (vybrané státy Evropy). Státy byly vybírány podle dostupnosti datových údajů.

Díky tomu, že WHO korektně uvádí již standardizovaná data (Světový standard), nemusela být v rámci práce řešena standardizace dat, která by ji značně zkomplikovala. Pro Českou republiku byla k dispozici standardizovaná data (Evropský standard) incidencí novotvarů plic za kraje.

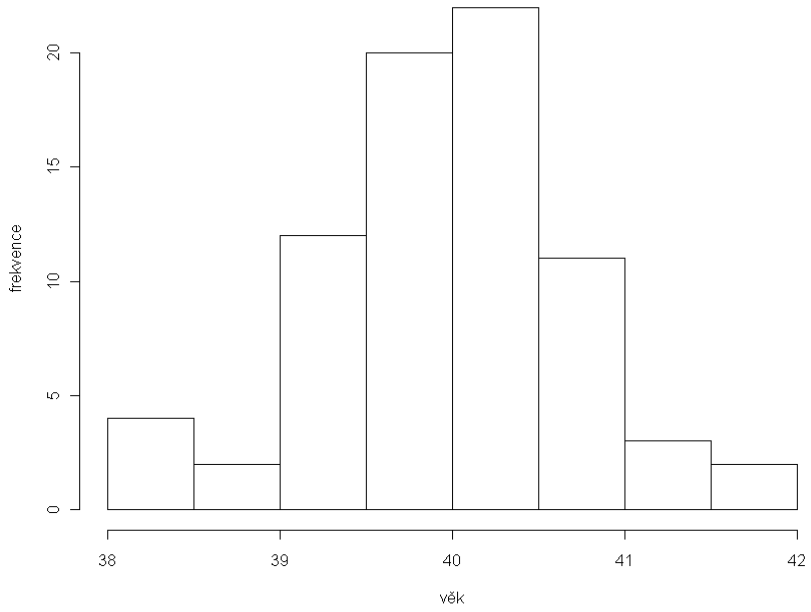
Další fakt, který plyne z výše uvedených charakteristik dat je ten, že není přesně sjednocen rok pořízení dat. Vzhledem k povaze korelačních analýz je tento fakt nepodstatný, protože jev většinou nevyvolá v daném roce nemoc (úmrtí na nemoc).

3.2. Korelace: průměrný věk–úmrtnost na nádorová onemocnění v okresech České republiky

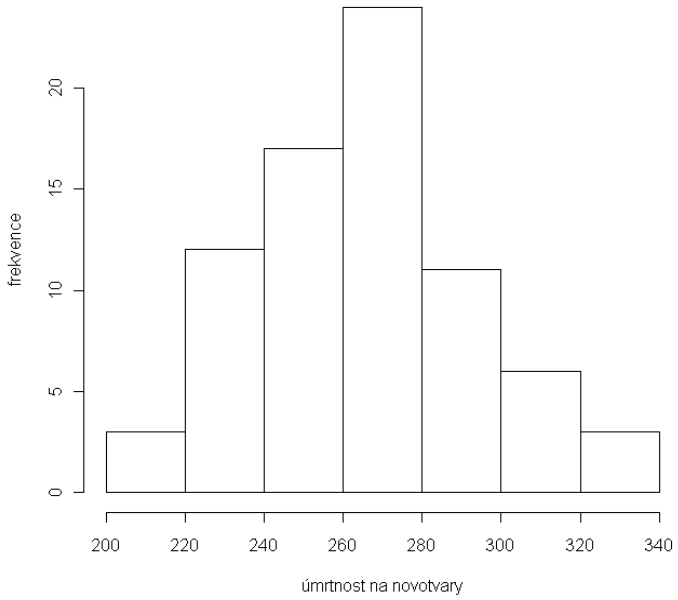
Vůbec jako první analýza byla zvolena statistická závislost věku na úmrtnost na nádorová onemocnění. Tato analýza potvrzuje důraz na standardizaci věkové struktury zdravotnických dat.

Jako vstupní data byla zvolena úmrtnost na novotvary v okresech České republiky v roce 2006, vzhledem k demonstrativnímu účelu nutnosti užití standardizace tato data logicky standardizována nejsou. Druhou veličinou je průměrný věk v okresech České republiky v roce 2006.

Obecně vhodnější by bylo provádět korelaci mezi věkem a hrubou míru úmrtnosti (počet zemřelých na 100 000 obyvatel), ale vzhledem k zaměření práce na nádorová onemocnění byla analýza prováděna právě nad těmito daty, aby bylo zřetelné, jak by mohla být data zkreslena, kdyby nebyla standardizována.



Obr. 1.: Histogram hodnot průměrného věku v okresech České republiky v roce 2006



Obr. 2.: Histogram hodnot úmrtností na novotvary v okresech České republiky v roce 2006

Před samotným výpočtem Pearsonova korelačního koeficientu bylo potřeba, vzhledem k jeho požadavkům, otestovat, zda-li vykazují charakteristiku normálního (Gaussova) rozdělení.

K základnímu optickému zhodnocení normality dat, slouží jako první histogram, z něhož je patrná četnost jednotlivých hodnot v datasetu (Obr. 1 a Obr. 2).

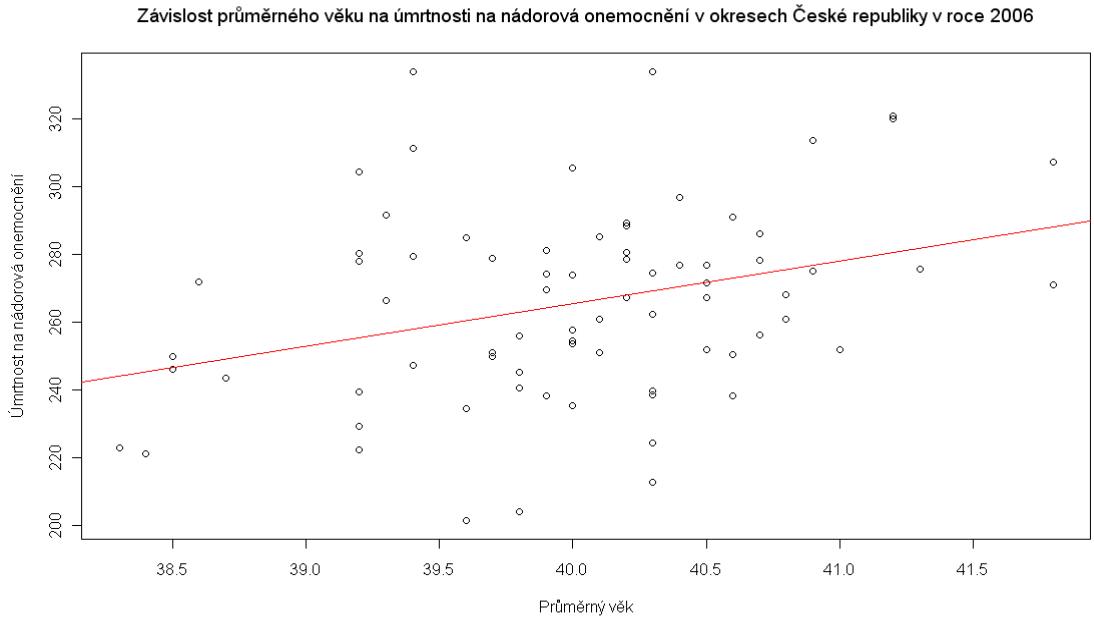
K testu normality slouží v softwaru R dva testy (Shapiro-Wilk test a Kolmogorov-Smirnov test). K hodnocení normality dat byl použit první ze jmenovaných, tedy Shapiro-Wilk test.

Tabulka 2.: Hodnoty Shapiro-Wilk testu (test normality dat)

Data	Shapiro-Wilk test	
	W-hodnota	p-hodnota
Průměrný věk	0,9818	0,3477
Úmrtnost na novotvary	0,9882	0,7068

První W-hodnota je analogií korelačního koeficientu. To znamená, že udává závislost mezi křivkou normálního rozdělení a hodnotami v požadovaném datasetu. Určující je však druhá p-hodnota, která udává chybu při zamítnutí nulové hypotézy: „ Data vykazují normální (Gaussovo) rozdělení.“

Z tabulky je zřejmé, že hodnoty jsou vyšší než (0,05), a tím pádem nedochází na hladině významnosti 5% k zamítnutí nulové hypotézy o normalitě dat a data mohla vstoupit do korelační analýzy Pearsonovým koeficientem.



Obr. 3: Graf závislosti průměrného věku na úmrtnost na novotvary v okresech České republiky v roce 2006 (lineární regrese).

Vzhledem k použití softwaru R, jenž respektuje jiné normy, je desetinná čárka u čísel v grafu reprezentována tečkou.

Tabulka 3.: Statistické charakteristiky

Statistická charakteristika	Dataset	
	Průměrný věk	Úmrtnost na novotvary na 100 000 obyvatel
Průměr	40,01	265,58
Medián	40,05	267,2
Maximum	41,8	334,0
Minimum	38,3	201,3
Směrodatná odchylka	0,7345	28,4473
Pearsonův korelační koeficient	0,3245	
p-hodnota (5% hladina významnosti)	0,00424	
Stupně volnosti	76	

Z grafu (Obr. 3) a vizualizované lineární regrese v datech je zřetelná závislost mezi průměrným věkem a úmrtností na novotvary. Podařilo se zamítnout na hladině významnosti 5% nulovou hypotézu o nezávislosti dat, jelikož p-hodnota je menší než (0,05), a byla přijata alternativní hypotéza, že data mezi sebou vykazují závislost. Pearsonův korelační koeficient má hodnotu 0,3245 (viz. Tabulka 3).

3.3. Parciální korelace: míra nezaměstnanosti–celková úmrtnost (s vyloučením faktoru věku)

Jinou alternativou, jak prokázat důležitost standardizace dat je užití metody parciální korelace, kdy do korelační analýzy vstupují tři veličiny, kdy třetí veličina je konstantní. Samotná metoda se z důvodu užívání standardizovaných měr spolu s faktorem věku neužívá, nicméně je na této metodě názorně vidět, jakým způsobem mohou věkové rozdíly ve sledovaném území zkreslit celkový výsledek analýzy. Na základě výsledků práce [21] byly z důvodu statisticky významné korelace vybrány faktory celkové míry úmrtnosti a míry nezaměstnanosti. U standardizovaných měr úmrtnosti a míry nezaměstnanosti (data z roku 2001) vyšly korelační koeficienty 0,64 mužů a 0,40 u žen (oboje statisticky významné).

Data použitá pro tuto analýzu jsou pro okresy ČR (77 záznamů) z roku 2006 (ČSÚ). Logicky, vzhledem k použití parciální korelace, zde není použita (na rozdíl od [21]) standardizovaná míra úmrtnosti, protože věk v analýze figuruje jako kontrolovaná veličina (neměnná). Základní charakteristiky datasetu jsou prezentovány v Tabulce 4, v Příloze 3 jsou ukázky dat v podobě boxplotů, na kterých je vizualizován medián, interkvartilový interval (mezi 1. a 3. kvartálem) a vymezení 1,5 násobkem interkvartilového intervalu. Hodnoty mimo tento interval jsou vyobrazeny samostatně.

Při výpočtu Pearsonova korelačního koeficientu pouze u veličin nezaměstnanosti a celkové úmrtnosti vyšel korelační koeficient 0,13 a na základě p-hodnoty 0,25 nebyla zamítnuta nulová hypotéza o závislosti mezi daty. Při použití metody parciální korelace však korelační koeficient vyšel 0,47 a na základě p-hodnoty (0,00018) byla nulová hypotéza zamítnuta a bylo tedy statisticky prokázáno, že nezaměstnanost má s vyloučením faktoru věku vliv na celkovou úmrtnost. Přestože byla oproti analýze se standardizovanými daty použita jiná metoda i data pořízena z jiného roku, lze korelační

koeficienty co se týče síly považovat za podobné (data nebyla odlišena podle pohlaví na rozdíl od [21]).

Tabulka 4.: Statistické charakteristiky

Statistická charakteristika	Dataset		
	Celková míra úmrtnosti	Míra nezaměstnanosti	Průměrný věk
Průměr	1015,4	8,06	39,93
Medián	1019,0	7,215	40,00
Maximum	1209,0	19,47	41,80
Minimum	837,0	0,06	38,20
Směrodatná odchylka	66,921	3,456	0,747

3.4. Korelace: měrné emise znečišťující ovzduší–standardizovaná incidence nádorových onemocnění plic

Český statistický úřad eviduje také údaje o životním prostředí, kde jsou dostupná i data o emisích znečišťujících ovzduší v krajích České republiky. Nevýhoda této konkrétní analýzy je, že jsou data dostupná právě jen za úroveň krajů. Znečištění bývá často i velmi lokálního charakteru, pokud je přítomen zdroj znečištění. Druhou nevýhodou je z toho vyplývající počet pouze třinácti záznamů vstupujících do analýzy, a tím pádem omezenější vypovídací schopnost datasetu.

Právě u tohoto rizikového faktoru je možné vhodně aplikovat nástroje geoinformačních systémů (GIS). Jeden ze záměrů práce byla extrakce údajů pro menší územní celky z rastrové vrstvy průměrného znečištění daným elementem, vytvořené metodikou SYMOS'97 (Systém modelování stacionárních zdrojů), ovšem nakonec tato vrstva nebyla Českým hydrometeorologickým ústavem (ČHMÚ) pro účely práce poskytnuta.

K dispozici jsou údaje o emisích pevných látek, oxidu siřičitého (SO_2), oxidů dusíku (NO_x) a oxidu uhličitého (CO_2) v tunách v přepočtu na jednotku plochy (km^2) za kraje České republiky v roce 2005. Jako zdravotnická data byla volena standardizovaná incidence nádorového onemocnění plic (evropský standard).

Vzhledem k počtu pouze třinácti údajů, které vstupují do analýzy nebyl proveden test normality dat a bylo přímo přistoupeno k metodě výpočtu korelačního koeficientu Spearmanovou metodou založenou na pořadí.

K identifikaci normality dat v datasetu může napovědět mimo histogramu, také základní popisná statistika v tabulce pod textem (Tabulka 5) a to porovnáním hodnot průměru a mediánu, protože v ideálním normálním rozdělení platí (průměr = medián = modus). Hodnoty se mezi sebou v porovnání s daty pro incidenci, která mají charakter normálního rozdělení, odlišují.

Tabulka 5.: Statistické charakteristiky

Statistická charakteristika	Dataset				
	pevné částice	oxid siřičitý (SO_2)	oxidy dusíku (NO_x)	oxid uhličitý (CO_2)	incidence novotvarů plic
Průměr	0,5071	3,157	2,464	3,407	34,36
Medián	0,4	1,75	0,85	1,5	34,49
Maximum	1,4	13,5	11,5	24,4	47,55
Minimum	0,2	0,5	0,4	0,8	20,82
Směrodatná odchylka	0,341	3,404	3,232	6,157	10,087

Tabulka 6.: Hodnoty korelačních koeficientů a p-hodnot

Data	Korelace s Incidencí novotvarů plic	
	Spearmanův korelační koeficient	p-hodnota (5% hladina významnosti)
pevné částice	0,4138	0,1413
oxid siřičitý (SO_2)	0,3934	0,1626
oxidy dusíku (NO_x)	0,3580	0,2088
oxid uhličitý (CO_2)	0,4093	0,1462

Tabulka 6 zobrazuje Spearmanovy korelační koeficienty a p-hodnoty korelace mezi jednotlivými druhy znečišťujících částic a incidencí novotvarů plic. Korelační koeficienty u všech druhů nabývají hodnoty kolem (0,4), avšak vzhledem nízkému počtu záznamů (třináct krajů, resp. dvanáct stupňů volnosti) se nepodařilo zamítnout nulovou hypotézu o nezávislosti dat. Statistický závěr tedy zní: „Incidence novotvarů plic není statisticky závislá na znečištění ovzduší.“

Toto tvrzení samozřejmě nemusí ve skutečnosti platit a je možné, že při provedení analýzy na menších územních jednotkách (okresy, obce s rozšířenou působností) by byla alternativní hypotéza přijata.

Data ke korelaci incidence novotvarů plic a pevných částic jsou vizualizovaná v podobě složeného pseudokorelačního kartogramu ([23], str. 76) v Příloze 4. Mapa byla vytvořena v softwaru ArcGIS 9.3. Každý atribut byl rozdělen do tří tříd stejnou metodou, kdy každá třída je definována podélnou/příčnou šrafou. Její hustota se stupňuje s rostoucí intenzitou jevu. Výsledkem je obdélníková síť, kdy husté křížení značí vysokou intenzitu obou jevů a naopak. Pokud by korelace mezi jevy byla velmi silná, síť by měla podobu čtverců (intervaly by se shodovaly). Chybou u tohoto příkladu je aplikace fyzickogeografického jevu na administrativní hranice, avšak tento příklad je pouze demonstrativní a tím splňuje hlavní účel.

3.5. Korelace: prevalence kuřáků a průměrná spotřeba cigaret–standardizovaná úmrtnost na nádorová onemocnění plic ve vybraných státech Evropy

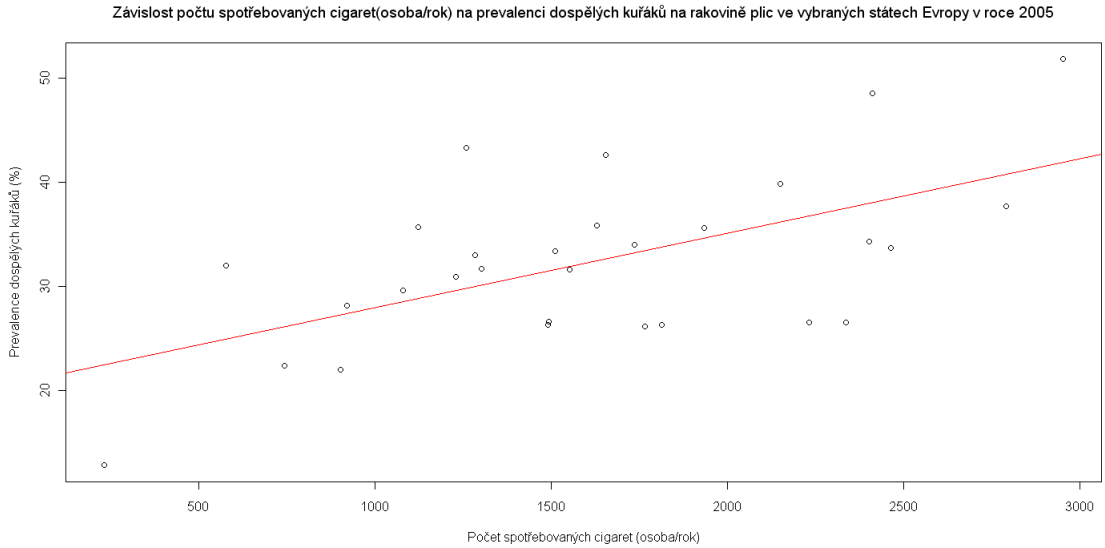
Fakt o škodlivém vlivu kouření na zdraví člověka je všeobecně již dlouhou dobu znám. Díky statistickým datům z World Health Organization (WHO), bylo možné tento fakt jednoduše statisticky ověřit.

Vstupními daty je prevalence kuřáků a počet cigaret na jednu osobu ve státech Evropy, kde byly tyto údaje zjištěny WHO. K dispozici jsou tedy dva odlišné typy dat, které se vztahují k jednomu rizikovému faktoru. Jako zdravotnická data byla zvolena standardizovaná úmrtnost na nádorová onemocnění plic. Data byla dostupná již ve standardizované podobě (světový standard).

První z analýz byla provedena nad prevalencí kuřáků a frekvencí nádorových onemocnění plic. Obecně lze předpokládat, že větší zastoupení kuřáků ve společnosti povede také k vyšší frekvenci nádorových onemocnění plic. Tento aspekt ovšem nezahrnuje styl ani množství vykouřených cigaret, tudíž teoreticky vzato mohou být v jedné populaci jedinci, jenž vykouří průměrně malé množství cigaret, což logicky povede také k nižší frekvenci nádorových onemocnění, zatímco v odlišné populaci bude situace naprosto opačná (průměrný počet vykouřených cigaret bude vysoký) a frekvence nádorových onemocnění plic vyšší.

Vzhledem k dostupnosti dat byla provedena analýza závislosti mezi prevalencí kuřáků a počtem spotřebovaných cigaret na osobu za rok (Obr. 4). Z grafu je patrné, že k výše zmíněnému jevu opravdu dochází. Kdyby byla spotřeba cigaret ve všech

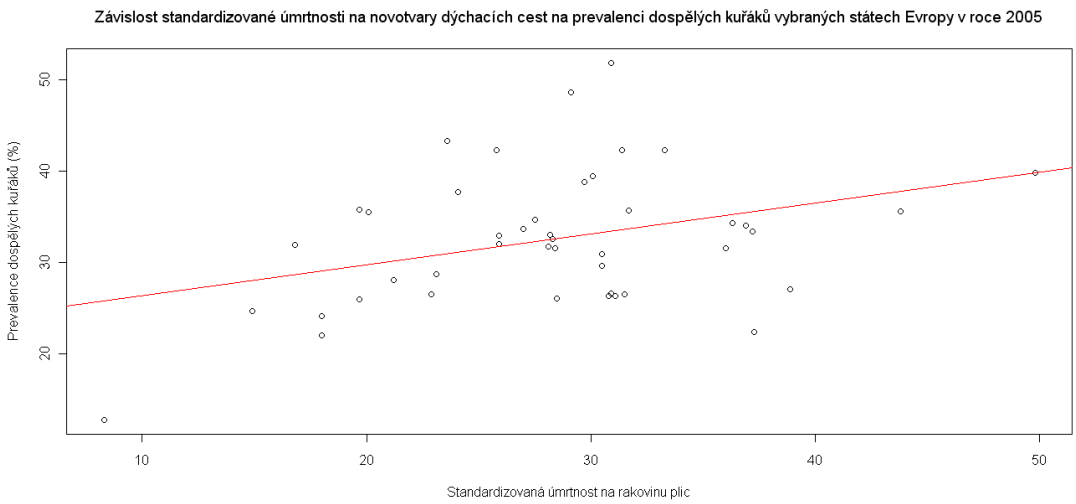
státech, již vstoupily do analýzy, stejná, tak by uskupení států, které jsou v grafu reprezentovány kroužky, mělo tvar přímky. Z grafu je zřejmé, že jsou přítomny odlehlejší elementy, již značí více či méně odlišné hodnoty od lineárního rostoucího trendu (p-hodnota [5%] = 0,0008276, Pearsonův korelační koeficient = 0,586).



Obr. 4.: Závislost prevalence dospělých kuřáků na počtu vykouřených cigaret na osobu za rok (lineární regrese)

3.5.1. Korelace: prevalence kuřáků–standardizovaná úmrtnost na nádorová onemocnění plic

Otestována byla závislost prevalence kuřáků na standardizované úmrtnosti na novotvary plic. Obecně lze tuto závislost přepokládat vzhledem k nebezpečnosti kouření. Již z grafu (Obr. 5) je pozorovatelný vzestupný trend mezi prevalencí a úmrtností. Výpočtem korelačního koeficientu a testováním hypotéz byla zamítnuta nulová hypotéza o nezávislosti dat a závislost mezi prevalencí kuřáků a úmrtností na novotvary plic byla statisticky prokázána, kdy p-hodnota byla 0,01845 (viz. Tabulka 8).



Obr. 5. : Závislost prevalence dospělých kuřáků na standardizované úmrtnosti na novotvary plic (lineární regrese)

Stejně jako u předchozích případů byla otestována normalita dat pro předpoklad užití Pearsonova korelačního koeficientu. Předpoklad normality je však zřetelný i z Tabulky 7 z hodnot minima, maxima, mediánu a průměru.

Tabulka 7.: Test normality dat

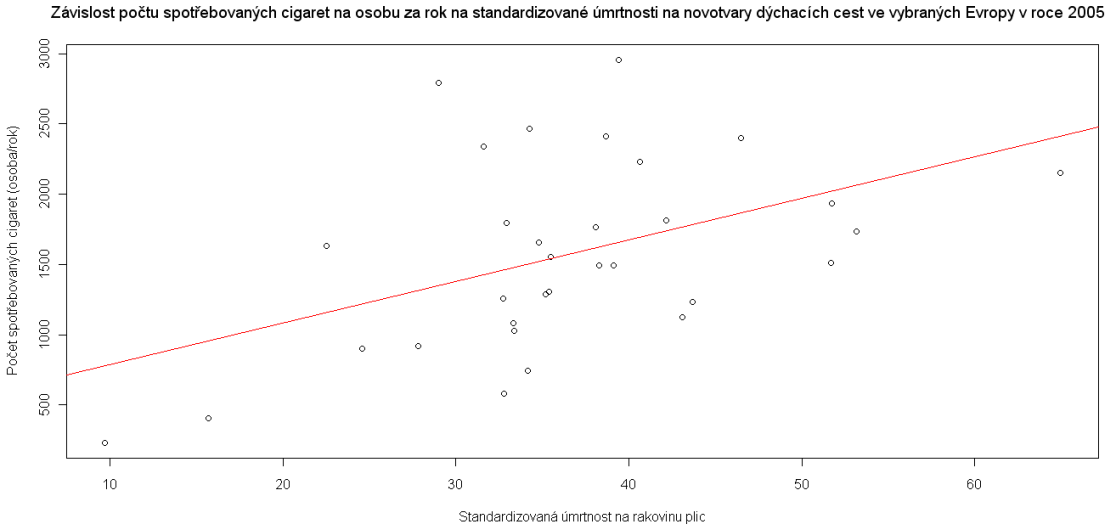
	Shapiro-Wilk test	
	W-hodnota	p-hodnota
Standardizovaná úmrtnost na novotvary plic	0,9795	0,6147
Prevalence dospělých kuřáků	0,9763	0,4935

Tabulka 8.: Statistické charakteristiky

Statistická charakteristika	Prevalence kuřáků (v %)	Standardizovaný počet úmrtí na novotvary plic na 100 000 obyvatel
Průměr	32,52	28,22
Medián	32,30	28,45
Maximum	41,80	49,80
Minimum	12,80	8,30
Směrodatná odchylka	7,364	7,741
Pearsonův korelační koeficient	0,3538555	
p-hodnota (5% hladina významnosti)	0,01845	
Stupně volnosti	42	

3.5.2. Korelace: průměrná spotřeba vykouřených cigaret–standardizovaná úmrtnost na novotvary plic

Další analýzou byla korelace počtu spotřebovaných cigaret v daném státě a úmrtnost na nádorová onemocnění plic. Oproti předchozí analýze byla výhoda v určení množství, jenž reprezentuje rizikový faktor, čímž je množství spotřebovaných cigaret na jednu osobu. Naopak nevýhodou je fakt, že vzhledem k různé prevalenci kuřáků v jednotlivých státech není vyjádřen počet cigaret, který skutečně spotřebují kuřáci, jelikož hodnota je přepočtena z celkového počtu obyvatelstva. Tato hodnota tedy také přesně nereprezentuje v analýze realitu, protože není splněn hlavní předpoklad, že počet vykouřených cigaret kuřákem může mít vliv úmrtí na novotvar plic.



Obr. 6: Graf závislosti počtu spotřebovaných cigaret na standardizovanou úmrtnost na novotvary plic (lineární regrese)

Z lineárního modelu grafu (Obr. 6) je jasně viditelná závislost, kdy s přibývajícím počtem cigaret také roste úmrtnost důsledkem novotvarů plic. V porovnání s předchozí analýzou je vyšší hodnota Pearsonova korelačního koeficientu (0,467) o (0,12), jež značí těsnější závislost.

Byla přijata alternativní hypotéza o závislosti mezi daty, kdy se ji podařilo výrazně zamítnout s p-hodnotou (5% hladina významnosti) 0,00694 (viz. Tabulka 9). Před analýzou byl otestován předpoklad normality vstupních dat.

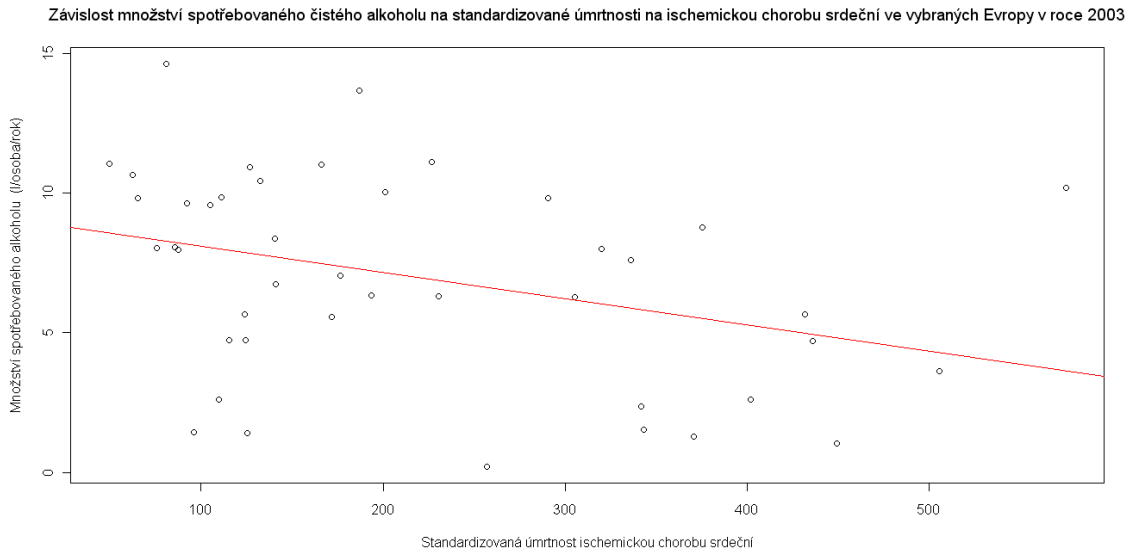
Tabulka 9.: Statistická charakteristika datasetu

Statistická charakteristika	Počet spotřebovaných cigaret na 1000 obyvatel/rok	Standardizovaný počet úmrtí na novotvary plic na 100 000 obyvatel
Průměr	1569	36,45
Medián	1533	35,29
Maximum	2954	64,96
Minimum	232,10	9,70
Směrodatná odchylka	674,0746	10,69244
Pearsonův korelační koeficient	0,467	
p-hodnota (5% hladina významnosti)	0,00694	
Stupně volnosti	30	

3.5.3. Porovnání analýz

V obou případech se podařilo statisticky prokázat závislost mez vstupními daty s tím, že v druhém případě byla závislost těsnější. Z dat spotřeby cigaret bylo zřetelnější, jaká je intenzita jevu, a tím pádem jeho potencionální dopad v souvislosti s onemocněním. Autor práce považuje druhý dataset za relevantnější vzhledem k dané analýze.

3.6. Korelace: spotřeba alkoholu–ischemická choroba srdeční ve vybraných státech Evropy



Obr. 7: Graf závislosti průměrné spotřeby alkoholu na 1 osobu za rok a standardizované úmrtnosti na ischemickou chorobu srdeční (lineární regrese)

Zatímco u předchozích analýz byl vždy vyšetřován deterministický faktor jako negativní vliv k vyvolání nemoci, tak zde je situace odlišná. Například podle [17] je statisticky vyšší riziko na ischemickou chorobu srdeční při abstinenci alkoholu než při dávce 20g za den, kdy je relativní riziko onemocnění nejnižší, avšak při vyšších dávkách alkoholu relativní riziko roste a při dávce kolem 72 g alkoholu za den se riziko rovná abstinenci alkoholu. Za tímto množstvím se riziko dále zvyšuje. Křivka závislosti má podobu písmena „J.“ Alkohol tedy v určitých případech působí jako preventivní faktor.

V datech vstupujících do korelační analýzy je uvedeno průměrné množství čistého alkoholu spotřebovaného v daném státě za rok. Nejvyšší hodnota v datasetu je 14,63 litrů čistého alkoholu na jednu osobu za rok (Lucembursko). V předchozím odstavci bylo zmíněno, že nejvyšší preventivní dávka alkoholu před onemocněním ischemické choroby srdeční činí 72 gramů za den odpovídající 26,3 kg za rok. Alkohol (ethanol) má nižší hustotu než voda (0,8 kg/l), což ve výsledku dává teoretickou spotřebu 32,88 litrů na osobu za rok. To je více než dvojnásobek maximální hodnoty v datasetu, a tudíž by teoreticky tato analýza měla vyjít kompletně jako preventivní faktor pro alkohol .

Z grafu (Obr. 7) zřetelně vyplývá předpoklad, který byl nastíněn v předchozím odstavci. S rostoucí spotřebou alkoholu ubývá případů úmrtí na ischemickou chorobu srdeční. Data jsou sice poměrně rozptýlena, nicméně klesající trend je viditelný.

Dataset v jednom z atributů (úmrtnost na ischemickou chorobu srdeční) nesplňoval parametry normálního rozdělení (viz. Tabulka 10), proto bylo užito Spearmanova neparametrického korelačního koeficientu namísto Pearsonova.

Tabulka 10: Test normality dat

	Shapiro-Wilk test	
	W-hodnota	p-hodnota
Počet úmrtí na ischemickou chorobu srdeční	0,897	0,001018
Průměrná spotřeba čistého alkoholu/rok	0,9616	0,1589

Z Tabulky 11 je viditelné, že Spearmanův korelační koeficient má hodnotu (-0,3784), tedy negativní korelaci. Testováním hypotéz se podařilo zamítnout nulovou hypotézu o nezávislosti mezi daty a byla přijata alternativní hypotéza: „Data mezi sebou vykazují závislost.“

Na tomto příkladě je zřetelný omyl, kterého by bylo možné se dopustit, kdyby se z této analýzy utvářely konkrétní závěry. Z analýzy by mohlo vyplynout, že čím více alkoholu člověk přijme, tím více se mu sníží relativní riziko na vznik ischemické choroby srdeční. V této populační studii sice takový výsledek vyšel, ale jak již bylo uvedeno na začátku kapitoly, v jiné epidemiologické analýze na úrovni dat od konkrétních jednotlivců byla konkluze odlišná.

Tabulka 11: Statistická charakteristika

	Standardizovaná úmrtnost na ischemickou chorobu srdeční	Spotřeba čistého alkoholu na 1 osobu/rok
Průměr	217,3	7,004
Medián	171,8	7,6
Maximum	575,1	14,63
Minimum	49,40	0,21
Směrodatná odchylka	137,8062	3,652462
Spearmanův korelační koeficient	-0,3784	
p-hodnota (5% hladina významnosti)	0,01281	
stupně volnosti	41	

3.7. Shrnutí výsledků

V rámci České republiky byly vyšetřovány vztahy: průměrný věk–standardizovaná úmrtnost na nádorová onemocnění v krajích České republiky, znečištění ovzduší–incidence novotvarů plic v krajích ČR (pevné částice, SO_2, NO_x, CO_2). U vyšetření korelace průměrný věk v okresech ČR–úmrtnost na nádorová onemocnění byla prokázána statistická závislost. Tento fakt upozorňuje na potřebu věkové standardizace dat, jenž může analýzy zkreslovat. Tato potřeba je také demonstrativně prezentována u dat míry nezaměstnanosti, celkové míry úmrtnosti a průměrného věku v okresech ČR, kdy je pomocí metody parciální korelace vyloučen vliv věku. Na rozdíl od neočištěných dat od průměrného věku se podařilo statisticky dokázat vliv míry nezaměstnanosti na celkové míře úmrtnosti U analýz znečištění ovzduší se ani u jednoho znečišťujícího elementu nepodařilo prokázat statistickou závislost. Výsledky těchto analýz by byly možná odlišné, pokud by byla k dispozici data za menší prostorové jednotky.

V analýzách na úrovni států Evropy byly vyšetřovány vztahy: průměrná spotřeba cigaret–standardizovaná úmrtnost na nádorová onemocnění plic, prevalence kuřáků–standardizovaná úmrtnost na nádorová onemocnění plic, průměrná spotřeba alkoholu–standardizovaná úmrtnost na ischemickou chorobu srdeční. Ve všech třech případech byla prokázána statistická závislost mezi jevem a úmrtností. U případu kouření byla u obou datových sad statisticky prokázána pozitivní korelace. U korelace standardizovaná úmrtnost na ischemickou chorobu srdeční–průměrná spotřeba alkoholu

byla jako u jediné analýzy statisticky prokázána negativní korelace a konzumace alkoholu tedy nevyšla jako jev, jenž nemoc způsobuje, ale alkohol byl z analýzy identifikován jako preventivní faktor.

4. DISKUZE

Nabízející se otázkou je, zda-li má význam vůbec tyto korelační analýzy provádět a jakou vůbec mají skutečnou vypovídací schopnost. Roli zde hraje velké množství faktorů. Jeden z nejdůležitějších je míra rizika, kterou potencionální deterministický faktor představuje.

Jev například může být v klinických pokusech vyhodnocen jako rizikový, avšak vzhledem k povaze agregovaných dat se tato skutečnost může ztratit v „šumu.“ V rámci bakalářské práci byly vybrány jevy, jež zasahují široké spektrum populace a například u kouření je jeho negativní dopad na lidské zdraví již dlouhá léta všeobecně znám a například u nádorového onemocnění plic se tento faktor podílí až z 90% [18], proto se statistickou závislost podařilo ve většině případů prokázat.

Dalším možným nedostatkem může být nevhodný výběr dat. Například v případě, že by byla vyšetřována souvislost mezi průměrným denním kalorickým příjmem a nádorovým onemocněním trávicí soustavy. Taková teze by mohla být statisticky neprokázaná, zatímco kdyby v jiném případě byl volen rozdíl mezi přijatou energií a výdejem a byl korelován se zdravotnickými daty, tak by se analýza stala statisticky signifikantní.

Nepřesnosti do dat mohou být vneseny samotnou evidencí dat. Ve vyspělých státech s moderní zdravotní péčí je i lékařská statistika vedena přesněji a uceleněji než u rozvojových zemích. K těmto datům je potřeba takto přistupovat. I kvůli tomuto faktu byla volena jako vstupující územní jednotka Evropa.

Korelační analýza může mít své úskalí již ze svého principu. Hodnota korelačního koeficientu je určována ze síly lineární závislosti. Data tedy mohou být závislá v jiném než lineárním vztahu, ale korelační koeficient tuto skutečnost nevyhodnotí. Samotná závislost mezi daty je zřetelná ze zhotoveného grafu, kde je možné tento fakt odhalit a zvolit jiný typ analýzy.

Zřejmě největším nedostatkem je skutečnost, že do analýzy vstupují zdravotnická data a data o intenzitě určitého jevu, která jsou vztažena ke konkrétnímu roku.

Taková data ve skutečnosti v žádném vztahu nejsou, protože nemoc se většinou projeví až po delším časovém období a ne v konkrétním roce při působení určitého deterministického faktoru. Analýza tedy přepokládá určitou časovou stabilitu dat.

Jak už bylo v textu zmiňováno, z agregovaných analýz by neměly být tvořeny konkrétní a všeobecně platné závěry, neboť skutečné vztahy mohou být mnohem složitější. Dobrým příkladem je tomu analýza v této bakalářské práci (korelace průměrné spotřeby alkoholu–standardizovaná úmrtnost na ischemickou chorobu srdeční). Tyto studie nehodnotí konkrétní riziko daného jevu, tím se zabývá jiné odvětví oboru epidemiologie.

Síla těchto studií spočívá právě v jejich jednoduchosti ve všech aspektech. Je z nich možné rychle utvářet hypotézy, které lze dále ověřovat jinými epidemiologickými metodami. Korelační studie je tedy možné považovat za vstupní bránu do procesu odhalování faktorů, které mohou ovlivňovat zdraví u lidí.

5. ZÁVĚR

Bakalářská práce se zabývala možnými závislostmi, které se mohou vyskytovat mezi vybranými daty reprezentující projevující se nemoc a příslušnými fyzickogeografickými či socioekonomickými jevy, jež jsou možnými nositely znaků, které mohou dané onemocnění vyvolávat. Tyto datasety byly utvořeny po konzultaci s prof. Vladimírem Mihálem (Lékařská fakulta UP Olomouc) s ohledem na jejich evidenci v dostupných veřejných statistikách, které vydává Světová zdravotnická organizace (WHO), Ústav zdravotnických informací a statistiky (ÚZIS) a Český statistický ústav (ČSÚ).

Vzhledem ke geografickému zaměření studia byl na data základní požadavek, aby se vztahovala k určité prostorové jednotce. Za tyto prostorové jednotky data vstupovala do korelačních analýz. Byly vybrány dva hlavní územní celky, pro které byla data vztažena, a tedy: Česká republika se základní prostorovou jednotkou kraj (čtyři analýzy) a okres (jedna analýza) a Evropa se základní prostorovou jednotkou stát (tři analýzy).

Práce je členěna do dvou hlavních částí. V první, teoretické části, jsou uvedeny obecné informace, které se vztahují k vlastnímu tématu bakalářské práce. Tato část je tematicky rozdělena do sekce epidemiologické, statistické a rešeršní. V epidemiologické sekci je především rozebrána charakteristika analýz, které byly v bakalářské práci provedeny. Ve statistické sekci jsou pak uvedeny konkrétní metody, které byly použity. Rešeršní část nastiňuje podobu odborných prací s podobným tématem, na základě kterých byly tvořeny postupy v bakalářské práci. Ve druhé, praktické, části bakalářské práce jsou uvedeny konkrétní výsledky z analýz provedenými nad daty spolu s prezentací těchto výsledků v grafech a tabulkách. Tyto výsledky jsou komentovány autorem práce.

6. POUŽITÉ ZDROJE

[1] ACKERS, Marta-Louise, et al. Are there national risk factors for epidemic cholera? The correlation between socioeconomic and demographic indices and cholera incidence in Latin America. *International Journal of Epidemiology* [online]. 1998, 27, [cit. 2010-05-21]. Dostupný z WWW: <http://ije.oxfordjournals.org/cgi/reprint/27/2/330.pdf>.

[2] ANDĚL, Jiří. *Matematická statistika*. Praha : SNTL - Nakladatelství technické literatury, 1985. 346 s.

[3] ANDĚL, Jiří. *Statistické metody*. Praha : Matfyzpress, 2003. 299 s.

[4] ANDĚL, Jiří. *Základy matematické statistiky*. Praha : Matfyzpress, 2005. 358 s.

[5] ARTALEJO, Fernando Rodríguez, et al. Socioeconomic Level, Sedentary Lifestyle, and Wine Consumption as Possible Explanations for Geographic Distribution of Cerebrovascular Disease Mortality in Spain . *Stroke* [online]. 1997, 28, [cit. 2010-05-21]. Dostupný z WWW: <file:///D:/Bakalarska%20prace/zdroje/zdroje/spain_vine.htm>.

[6] BENCKO, V. a kol.*: Základy statistiky pro biomedicínské obory.* Praha, Karolinum, 2003, 236 s.

[7] Cleophas, T., J. et al.. Statistics Applied to Clinical Trials. Kluwer Academic Publishers.2000

[8] DIKIY, N. P., et al. Some correlation aspects of thyroid cancer epidemiology in Ukraine after Chernobyl accident . *International Congress Series*. July 2002, 1236, s. 39-41.

[9] EIDE, Melody J.; WEINSTOCK, Martin A. Association of UV Index, Latitude, and Melanoma Incidence in Nonwhite Populations— US Surveillance, Epidemiology, and End Results (SEER) Program, 1992 to 2001. *ARCH DERMATOL* [online]. APR 2005, 141, [cit. 2010-05-21]. Dostupný z WWW: <http://archderm.ama-assn.org/cgi/reprint/141/4/477.pdf>.

[10] JANOUT, V.: *Základy epidemiologie. Olomouc,* Vydavatelství Univerzity Palackého v Olomouci, 1995, 93 s.

[11] Liu, J.-Chow, S. Design and Analysis of Clinical Trials: Concepts and Methodologies. 1998

[12] MATOS, Elena L.; LORIA, Dora I.; VILENSKY, Marta. Cancer Mortality and Poverty in Argentina: A Geographical Correlation Study. *Cancer Epidemiology : Bomarkers & Preventioni* [online]. April/May 1994, 3, [cit. 2010-05-21]. Dostupný z WWW: <http://cebp.aacrjournals.org/content/3/3/213.full.pdf+html>.

[13] MATOUŠEK, J.: Počasí podnebí a člověk: Bioklimatologie člověka, Praha, Avicenum, 1988, 293 s.

[14] Meinert, C. L. Clinical Trials: Design, Conduct and Analysis. Oxford University Press. 1996

[15] NORLEANS, M. X. Statistical methods for clinical trials. Marcel Dekker. 2001

[16] PIANTADOSI, S. Clinical Trials: A Methodological Perspective. John Wiley and Sons. 1986

[17] *Prevence nemocí a podpora zdraví: Alkohol a jeho vliv na zdraví – pít či nepít?* [online]. c2004, [cit. 2007-07-15].Dostupné z WWW: <http://www.cba.muni.cz/prevencenemoci/modules.php?name=Content&pa=showpage&pid=8>

[18] PROVAZNÍK, K. a kol.: Manuál prevence v lékařské praxi: I. Prevence poruch a nemocí. Praha, Fortuna, 1994, 137 s.

[19] STÁTNÍ ZDRAVOTNÍ ÚSTAV. *Systém monitorování zdravotního stavu obyvatelstva České republiky ve vztahu k životnímu prostředí* [online]. [cit. 2007-07-15]. Dostupné z WWW: <http://www.szu.cz/uploads/documents/chzp/souhrnna_zprava/Szu_08cz.pdf>

[20] SCHWARTZ G, Gary ; SKINNER G, Halcyon; DUNCAN, Robert. Solid waste and pancreatic cancer: an ecologic study in Florida, USA. *International Journal of Epidemiology* [online]. 1998, 27, [cit. 2010-05-21]. Dostupný z WWW: <http://ije.oxfordjournals.org/cgi/reprint/27/5/781.pdf>.

[21] Šlachtová, H., Tomášková, H., Skýbová, D., Polaufová, P., Tomášek, I., Šplíchalová, A.: *Socioekonomické nerovnosti ve zdraví obyvatel okresů České republiky*. Konference Životné podmienky a zdravie, vědecká konference s mezinárodní účastí, Štrbské Pleso, 11.-13.10.2006

[22] VOŽENÍLEK, V.: *Diplomové práce z geoinformatiky*. Olomouc: UP v Olomouci, 2002. 61 s. ISBN 80-244-0469-9.

[23] VOŽENÍLEK, Vít. *Aplikovaná kartografie I. : Tematické mapy*. Olomouc : Univerzita palackého v Olomouci, 2004. 187 s.

[24] WOODING, W., M. *Planning pharmaceutical clinical trials*. John Wiley and Sons. 1994

[25] ZVÁROVÁ, J.: *Základy statistiky pro biomedicínské obory*. Praha, Karolinum, 1998, 218 s.

SUMMARY

The theme of This Bachelor thesis is about associations between medical data and data of some potentional factor which could have relation to disease. There were two main groups of factor data (physical and social). Data were elected for regions of Czech republic and chosen states of Europe from public database of World health organization (WHO), Český statistický úřad (ČSÚ) and Ústav zdravotnických informací a statistiky (ÚZIS).

Data came into correlation analysis and testing of hypoteses were proceeded. By this criterium were decided if factor has relation to disease. Each significant relationship is visualized in graph. Graphs where one of the main parts of analysis because of patterns in data vizualization. Following data were obtained statistically significant:

- Average age in regions of Czech republic–Cancer mortality rate
- Prevalence of smoking adult in chosen states of Europe–Standardized mortality rate for lung cancer
- Consupmtion of ciggarettes per person per year–Standardized mortality rate for lung cancer
- Consumption of pure alcohol per person per year–Standardized mortality rate for ischemic heart disease

Correlation studies are one of the metodhs of epidemiology praxis. They have several disadvantages but their main advatage is that it is easy to proceed them. These analyses can be pilot analyses for advantage processing.

SEZNAM PŘÍLOH

Příloha 1

```
#nastavení pracovního adresáře
setwd("D:/Bakalarska prace/Datove podklady")

#načtení datového souboru
data = read.table("alkohol_isch.txt",header = T)

#načtení dat úmrtnosti ve 3. sloupci
umrtnost = data2[,3]

#zobrzení dat úmrtnosti
umrtnost

#načtení dat pro spotřeba alkoholu ve 4. sloupci
spotreba=data2[,4]

#zobrazení dat spotřeby
spotreba

#směrodatná odchylka
sd(umrtnost)
sd(spotreba)

#základní statistická charakteristika (průměr, medián, maximum, minimum, kvartily)
summary(umrtnost)
summary(spotreba)

#zobrazení grafu
plot(umrtnost,spotreba,ylab="průměrná spotřeba alkoholu na osobu za rok",xlab="standardizovaná úmrtnost na ischemickou chorobu srdeční" )

#uložení koeficientů do proměnné
b<-lm(spotreba~umrtnost)
#zobrazení lineárního modelu v grafu
abline(b, col = "red")


#výpočet korelačního koeficientu (Pearsonova)
cor.test(umrtnost,spotreba)
```