Co potřebuje klinik vědět o (bio)statistice?

Jak popsat data

Pravoslav Stránský

Ústav lékařské biofyziky a Oddělení výpočetní techniky, Univerzita Karlova v Praze, Lékařská fakulta v Hradci Králové, Česká republika / Department of Medical Biophysics and Division of Computer Science, Charles University in Prague, School of Medicine at Hradec Králové, Czech Republic

Stránský P. Co potřebuje klinik vědět o (bio)statistice? Jak popsat data. Folia Gastroenterol Hepatol 2005; 3 (1): 42 – 46.

Souhrn. *Je uvedena definice základního a výběrového souboru a různé měrné stupnice, ve kterých mohou být popsány proměnné, které se v publikaci uvádějí. Je vysvětlena podstata užívaných charakteristik popisujících data a jejich proměnlivost.*

Klíčová slova: *populace, výběr, typy dat, charakteristiky polohy a proměnlivosti*

Stránský P. What should a clinician know about (bio)statistics? How to describe data. Folia Gastroenterol Hepatol 2005; 3 (1): 42 – 46.

Abstract. *Population and sample are defined together with different types of the measurement scales used for a description of data presented in a publication. A principle of statistics used for a characterisation of measures of location and variability is explained.*

Key words: *population, sample, data types, measures of location and variability*

Jedním z požadavků kladených na dobrou publikaci je uvedení charakteristik souboru, na jejichž základě jsme došli k závěrům, které čtenáři sdělujeme a které z údajů dělají informaci. Za tu považujeme jakoukoli formu sdělení ovlivňující naše chování či rozhodování. V úvodu je nutné vysvětlit určité pojmy, které učiní další výklad jednoznačným a srozumitelným.

Nejprve tedy k pojmu **soubor**. Ve statistice se rozlišují dva druhy souborů. První se nazývá **základní** neboli **populace** *(population)*, druhý se označuje jako výběrový neboli vzorek *(sample)*. Formálně by tedy bylo správné uvádět podstatné jméno soubor i s příslušným přídavným jménem. Protože by to však v řadě případů vedlo ke zbytečnému opakování, volí se jedna ze dvou možností podle toho, na co je kladen důraz: buď jde o výklad obecnější, teoretičtější, a pak rozšiřující přídavné jméno používáme pro výběrový soubor a samotné slovo soubor označuje populaci, nebo se práce týká šetření prováděného ve výběru, a pak rozšířené označení použijeme pro základní soubor a soubor bez adjektiva znamená, že máme na mysli soubor výběrový. Jelikož převážná většina pojmů a metod, o kterých se v tomto a dalších článcích zmíním, se bude vztahovat k vzorkům, budu používat možnost druhou. To, co bylo právě uvedeno, platí i pro všechny statistické charakteristiky, takže rozlišujeme např. výběrový a populační průměr, výběrový a populační rozptyl. K snadnějšímu rozlišení v textu se ujala téměř všeobecně používaná konvence: charakteristiky výběrové se označují písmeny abecedy latinské, charakteristiky populační písmeny abecedy řecké. V uvedeném případě tedy $\bar{x}$ a μ a s a σ.

Bez rozboru toho, čím je určen základní soubor, je pro další výklad nutné zdůraznit, že výběrový soubor musí být **reprezentativní**. To je základní podmínka splnění cílů každé vědecké práce, bez které nelze závěry zjištěné na vzorku rozšířit na celou populaci. Reprezentativnost výběru je zaručena splněním dvou podmínek. První je dána tím, že pravděpodobnost každého jedince v populaci dostat se do výběru je stejná. Toho lze dosáhnout různými způsoby a vytvořený

výběr se nazývá **náhodným** *(random sample)*. Nejlepším řešením, jak zajistit plnění této podmínky pro danou studii, je postup zkonzultovat se statistikem.

Známým příkladem, jak nesplnění tohoto požadavku ovlivní výsledek šetření, je předpověď výsledku prezidentských voleb v USA v roce 1936. Časopis, který šetření organizoval, kontaktoval vzorek 10 milionů voličů. Závěr průzkumu vyzněl jednoznačně ve prospěch republikánského kandidáta (dnes většině čtenářů tohoto článku zcela neznámého A. M. Londona) proti demokratickému kandidátovi, jímž byl F. D. Roosevelt. Zpětná analýza ukázala, že hlavním nedostatkem provedeného odhadu byla skutečnost, že výběr respondentů nebyl náhodný. Kontaktovány byly osoby uvedené v telefonním seznamu, v registru vlastníků automobilů a předplatitelů časopisu, který průzkum organizoval. Šlo o skupinu lidí, kteří v té době patřili k bohatší podmnožině všech voličů (1).

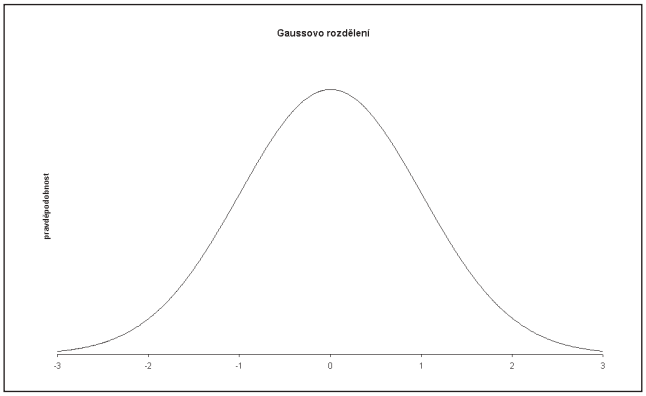
Druhá podmínka reprezentativnosti je dána dostatečnou velikostí souboru, počtem jedinců, na kterých bylo pozorování prováděno. Právě uvedený příklad je důkazem, že stanovení velikosti souboru není triviálním problémem. V našich podmínkách 10 milionů jedinců odpovídá zjišťování sledovaných údajů v celé populaci a je reálné pouze v případě dat zjišťovaných v rámci sčítání lidu. Pro výběrová šetření je nejmenší požadovaná velikost dána variabilitou, proměnlivostí zjišťovaného znaku a pravděpodobností, s jakou chceme daný závěr prokázat. Odpověď na tuto otázku není jednoduchá. V případě testování hypotéz se řeší pomocí určení tzv. síly testu *(power analysis)* a podrobnosti s uvedením příkladu pro nepárový *t*-test (asi v medicíně nejčastěji používaný) jsou v (3).

Zjišťování hodnot nějakého znaku (proměnné veličiny) znamená její měření a měření je to, čím se liší věda od umění. Právě tato skutečnost by měla pomoci v rozhodování o tom, zda medicína je více umění či věda. Podle toho, jaká je míra přesnosti určení hodnoty, rozlišujeme čtyři druhy stupnic, škál a data v nich měřená označujeme stejně (tedy pokud užijeme škálu ordinální, nazýváme hodnoty v ní určené daty ordinálními):

1. **nominální** – výsledkem měření je přiřazení určitého kódu zjištěnému nálezu, pokud je splněn určitý požadavek. Kód může být vyjádřen číslicemi (např. diagnóza podle Mezinárodní klasifikace nemocí), písmeny (krevní skupina) nebo jinými znaky (faktor Rh). Zvláštní skupinu na této škále představují data binární, která mohou nabývat pouze dvou hodnot (pohlaví). Základní vlastností nominálních dat je to, že i když použitý kód je číslo, nemá smysl je na základě přiřazené hodnoty řadit podle velikosti, zjišťovat jejich rozdíl či podíl;
2. **ordinální** – jejich seřazení podle velikosti přináší určitou informaci. Typickými příklady jsou škály používané k vyjádření velikosti potíží, výsledků léčby a také známky používané pro ohodnocení znalostí u zkoušky, body v různých testech. Určování rozdílu či podílu ordinálních dat žádný smysl nedává. V případě ordinální škály je žádoucí pečlivě zvážit definici jejích jednotlivých kategorií. Příkladem špatné volby je určení biometeorologické zátěže v předpovědi počasí: mírná, střední, vysoká zátěž. Z takto definované stupnice vyplývá, že počasí nejen že nemůže zdravotní stav neovlivňovat, ale ani zlepšovat. To je jistě v rozporu s každodenní zkušeností;
3. **intervalová** – jedná se o stupnici kvantitativní, na které má smysl určovat rozdíl mezi zjištěnými hodnotami. Vzhledem k tomu, že nula takovéto stupnice je stanovena na základě určité konvence (0 °C, 0 °F), nemá smysl zjišťovat podíl naměřených hodnot (rozdíl teplot mezi -10 °C a +10 °C je 20 °C stejně jako rozdíl mezi 30 °C a 50 °C, ale podíly -10 °C/10 °C a 30 °C/50 °C nemají žádný význam);
4. **poměrová** – nula v této stupnici má logický význam a rovněž tak podíl zjištěných hodnot. Hmotnost 100 kg je dvakrát tak velká v poměru k hmotnosti 50 kg, glykémie 10 mmol/l je dvojnásobná v porovnání s 5 mmol/l.

Typ dat zásadně ovlivňuje charakteristiky, které by se měly použít k popisu prezentovaných dat. To také úzce souvisí s jejich tzv. rozdělením. V medicíně se nejčastěji předpokládá, že data mají rozdělení nor-

Obr. 1 / Fig. 1
Gaussovo rozdělení
Gaussian Distribution



mální neboli Gaussovo. Jeho grafickým zobrazením je dobře známá zvonovitá křivka *(bell shaped curve, Glockenkurve)* obr. 1. Předpokladem, že data mohou mít toto rozdělení, je, že jsou spojitá a mohou nabývat hodnot v intervalu od $-\infty$ do $+\infty$.

Adjektivum *normální* má v této souvislosti zcela přesný význam. Stejné přídavné jméno se v medicíně používá pro označení výsledků vyšetření, která se nepovažují za patologická. Aby nedocházelo k záměně, bude se v dalším textu ve spojení s druhem rozložení dat používat spojení rozložení Gaussovo nebo gaussovské. Toto rozložení dat má ve statistice zásadní význam. Je předpokladem použití tzv. parametrických metod, které se používají nejčastěji. Je zřejmé, že zejména definiční obor hodnot tohoto rozdělení nemůže být v medicíně teoreticky nikdy splněn. Oprávnění použití parametrických testů je založeno na skutečnosti, že srovnáním rozložení experimentálních dat se na zvolené hladině významnosti neprokáže, že tato data se liší od rozložení teoretického, Gaussova.

K popisu výsledků měření v různých škálách se používají různé charakteristiky, které jednak vystihují, v jakém vztahu je daná charakteristika ke všem datům - nazývají se **míry polohy** znaku, a dále charakteristiky, které vyjadřují proměnlivost dat – **míry variability**. V medicíně nejčastěji používanými charakteristikami jsou **aritmetický průměr** z naměřených hodnot, $\bar{x}$, *(arithmetic mean)* a **směrodatná odchylka**, **s**, *(standard deviation)*. Důvodů, proč tomu tak je, je celá řada a v rámci výkladu tohoto článku není podstatné se jimi zabývat. Stačí jen připomenout, že ne vždy je aritmetický průměr jako vyjádření typické hodnoty nejvhodnější charakteristikou. Střední tlak krve není aritmetickým průměrem mezi tlakem systolickým a diastolickým, pro veličiny, jejichž průběh lze popsat sinovou funkcí (střídavý proud), je aritmetický průměr roven nule a míra polohy je nejlépe vyjádřena tzv. efektivní hodnotou, která se rovná maximální hodnotě dělené odmocninou ze dvou.

Další charakteristikou polohy znaku je hodnota, která se v souboru naměřených dat vyskytuje nejčastěji a která se nazývá **modus**, $\hat{x}$. Pokud je takováto hodnota jediná, označuje se rozložení dat jako unimodální. Existuje-li jich více, mluvíme o rozloženích bimodálních, resp. multimodálních. Z pohledu klinika je důležité si pamatovat, že pokud rozložení dat není unimodální, znamená to většinou, že data nejsou homogenní. Typickým příkladem je bimodální rozložení dat, jejichž nejčastěji se vyskytující hodnota je závislá na pohlaví a která byla zjišťována bez ohledu na tuto skutečnost. Modalita dat intervalových a poměrových také úzce souvisí s přesností jejich měření a počtem pozorování.

Aritmetický průměr i modus jsou charakteristiky polohy znaku, které můžeme vypočítat z hodnot neuspořádaných podle jejich velikosti. V případě další důležité charakteristiky, které se říká **medián**, $\tilde{x}$, (čti iks s tildou), musíme nejprve data seřadit od nejmenšího po největší. Z toho je zřejmé, že se jedná o charakteristiku nevhodnou pro data nominální.

Medián je hodnota, která dělí uspořádaná data na dvě stejné poloviny. Padesát procent všech naměřených hodnot je menších, než je hodnota mediánu, padesát procent je naopak větších. Domyslíte-li definici do konce, snadno odvodíte, že na rozdíl od modu se hodnota mediánu mezi naměřenými údaji nemusí vůbec vyskytovat (lichý počet měření, jejich prostřední hodnoty nejsou stejné). Výhodou mediánu je také skutečnost, že jeho hodnota není příliš (vůbec) ovlivněna extrémními hodnotami měření.

Medián je jednou hodnotou ze skupiny měr polohy znaku, které se označují jako **kvantily**. To jsou hodnoty dělící zjištěná pozorování do skupin, které zahrnují stejnou část, stejný podíl všech měřených hodnot. Pokud kvantily vymezují hodnoty obsahující čtyři stejné díly, říká se jim **kvartily** (první kvartil obsahuje prvních 25 % naměřených hodnot, čtvrtý hodnoty ležící mezi maximální hodnotou a větší než je 75 % všech naměřených hodnot). Jestliže je intervalů deset, mluvíme o **decilech**, je-li jich 100, říká se jim **percentily**. Medián je zřejmě druhý kvartil, pátý decil a padesátý percentil. Pochopení pojmu kvantilů je nesmírně důležité pro pochopení způsobu definice tzv. normálních, nepatologických hodnot výsledků nejrůznějších vyšetření.

V případě, že naměřená data jsou ve své podstatě spojitá a unimodální, mohou mít pro jejich popis význam ještě dvě další charakteristiky, které umožňují jejich srovnání s rozložením Gaussovým. Jde o **šikmost** *(skewness)* a **špičatost** *(kurtosis)*. Vyjadřují, v jakém vztahu jsou zjištěná experimentální data k rozložení teoretickému (gaussovskému). Pro něj platí, že $\bar{x}=\tilde{x}=\hat{x}$. Pro pravostranně sešikmené rozdělení (hodnot na levém konci grafu, který zobrazuje jejich četnost, je více než na konci pravém) je $\hat{x}<\tilde{x}<\bar{x}$, pro levostranně sešikmené rozdělení je tomu opačně $\hat{x}>\tilde{x}>\bar{x}$, obr. 2. Jestliže je šikmost kladná, jde o sešiku-

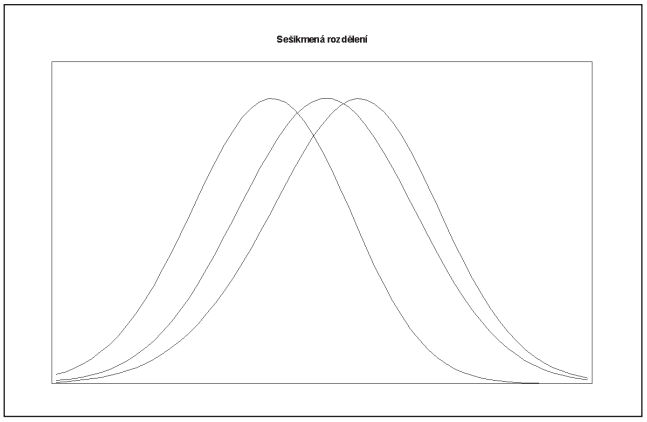
kmení pravostranné a obráceně. Kladná špičatost znamená, že experimentální rozdělení symetrických (nesešikmených) dat je „štíhlejší“ *(leptokurtické)* než rozdělení Gaussovo, záporná naopak, že jde o rozdělení „placatější“ *(platykurtické)*.

Tím se dostáváme k **mírám variability**. Nejčastěji užívanou charakteristikou je směrodatná odchylka. Její hlavní význam je v tom, že pro data s gaussovským rozložením leží 67 % naměřených hodnot v intervalu $\bar{x}\pm s$. U dat leptokurtických jich v tomto intervalu je více než 67 %, u dat platykurtických naopak méně. Už z tohoto faktu vyplývá, že směrodatná odchylka jako míra variability by měla být uváděna s ohledem na typ rozložení naměřených údajů.

Dříve než uvedeme další charakteristiky variability, řekneme si několik poznámek k směrodatné odchylce. Ta je definována jako druhá odmocnina **rozptylu** *(variance)*, $\boldsymbol{s}^2$. Populační rozptyl je σ^2, je aritmetickým průměrem druhých mocnin odchylek měření. Odchylka je definována jako rozdíl naměřené hodnoty a aritmetického průměru. Nevýhodou této veličiny je skutečnost, že její rozměr je jiný, než je rozměr měřené veličiny (např. pro hmotnost kg^2, pro tlak Pa^2), a představit si její vztah k dané proměnné je obtížné. Směrodatná odchylka má stejný rozměr jako zkoumaná data a její interpretace je jednoduchá. Za předpokladu gaussovského rozložení dat leží v intervalu $\bar{x}\pm 2s$ přibližně 95 % všech zjištěných hodnot (přesněji je násobitelem číslo 1,96), v intervalu $\bar{x}\pm 3s$ je to 99 % všech hodnot (přesněji je násobitel 2,57).

Vliv proměnlivosti zjišťovaných údajů, ať už z důvodů náhodných chyb či jejich biologické variability, na správnost vypočtené charakteristiky můžeme snížit zvýšením počtu pozorování a je nepřímo úměrný druhé odmocnině počtu pozorování.

Obr. 2 / Fig. 2
Sešikmené rozdělení
Skewed Distributions



Charakteristika vypočtená tímto způsobem se nazývá **standardní chyba** např. aritmetického průměru *(standard error of mean, SEM)*. Její význam je v tom, že umožňuje určit pravděpodobnost, se kterou se v určitém intervalu kolem vypočtené charakteristiky (v našem příkladu kolem aritmetického průměru) nachází skutečná hodnota (správná v případě měření fyzikálních, populační průměr v případě biologických proměnných). Tak např. $\bar{x} + 1{,}96\ SEM$ je **95% horní mez spolehlivosti** (95% upper confidence limit), $\bar{x} - 1{,}96\ SEM$ je **95% dolní mez spolehlivosti** *(95% lower confidence limit)*. Rozdíl horní a dolní meze je **interval spolehlivosti** (v uvedeném příkladu 95%).

Jde-li o charakteristiky výběrové, pak hodnota násobitele je obecně větší, než je uvedená hodnota (1,96 pro 95% mez spolehlivosti), a závisí na počtu provedených měření. Je to hodnota odpovídající příslušnému percentilu (pro uvedený příklad 97,5 percentilu) *t*-rozložení, které je známé jako rozložení Studentovo a o kterém se více zmíníme v dalším článku této série.

Další charakteristikou variability dat je **variační koeficient**, což je podíl směrodatné odchylky a aritmetického průměru. Z toho plyne, že jde o veličinu bezrozměrnou (s rozměrem 1) a po vynásobení 100 se udává v procentech. Hodnota do 30 % se v případě biologických dat považuje za přijatelnou, jestliže je výrazně vyšší než 30 %, je žádoucí vysvětlit důvod, který to způsobuje. **Variační rozpětí** *(range)* je rozdíl maximální a minimální hodnoty a zahrnuje 100 % všech zjištěných údajů. **(Inter)kvartilové rozpětí** je rozdíl třetího a prvního kvartilu a zahrnuje 50 % dat, **(inter)decilové rozpětí** je rozdíl devátého a prvního decilu a obsahuje 80 % hodnot. 95 % všech naměřených dat obsahuje interval údajů mezi 97,5 a 2,5 percentilem. Tento interval odpovídá tomu, co v případě dat s Gaussovým rozložením je dáno hodnotou $\bar{x} + 1{,}96s$ a je považováno za hodnoty normální, nepatologické. V případě, že předpoklad gaussovského rozložení našich výsledků nemůžeme přijmout (jak se to dělá, si řekneme příště), musíme k určení dat použít interval hodnot, který je dán rozdílem výše uvedených percentilů.

Z uvedených vlastností by mělo být zřejmé, že aritmetický průměr a směrodatná odchylka jsou vhodnými charakteristikami pouze za předpokladu, že rozložení popisovaných dat se neliší od gaussovského. Není-li tomu tak, potom v případě dat poměrových

a intervalových a rozhodně vždy u dat ordinálních jsou vhodnými charakteristikami polohy medián a modus (mody) a jako míra variability příslušné meze spolehlivosti mediánu. V případě dat nominálních má význam pouze modus. U nich je nejlépe uvést tabulku, ve které jsou relativní četnosti (v %) výskytu jednotlivých znaků. Z praktického hlediska vzhledem ke zvyklostem používaným v redakcích medicínských časopisů je nejlepším řešením pro data kvantitativní (poměrová a intervalová) použít k jejich popisu aritmetický průměr a směrodatnou odchylku a pokud víme, že jejich rozložení nelze za gaussovské považovat, tak i medián s mezemi spolehlivosti.

Použití aritmetického průměru a směrodatné odchylky u dat ordinálních je důkazem toho, že autor, recenzent a odpovědný redaktor nechápou podstatu a význam příslušných statistických charakteristik. Najdeme-li takovéto údaje v určité publikaci, měli bychom být velice kritičtí k závěrům, které autoři uvádějí. Příkladem může být abstrakt v (2), který mají asi všichni čtenáři k dispozici. Název práce je imponující, „randomizovaná, dvojitě slepá kontrolovaná studie“ jistě vyvolá ve čtenáři dojem, že nic lepšího už není možné. Při studiu výsledků v práci uváděných se však dostaví pochybnosti o správnosti tvrzení, která jsou z uvedených výsledků odvozována. Znamená např., pokud je skóre pro bolest 2±3,4, že až 67 % pacientů mělo skóre -1,6?

V této souvislosti si všimněme ještě dvou důležitých skutečností. Pokud jde o hodnotu před znaménkem ±, nejsou žádné pochybnosti. Vždy jde o aritmetický průměr. Zato v případě hodnoty za znaménkem u převážné většiny publikovaných údajů nevíme, co autoři uvádějí. Může to být *s*, *SEM* nebo $\bar{x}$% meze spolehlivosti. Z důvodů, které budou vysvětleny v dalších článcích této řady sdělení, by to měly být meze spolehlivosti dané charakteristiky.

Druhou skutečností je otázka, s jakým počtem platných číslic máme charakteristiky popisující data uvádět. V případě kvantitativních údajů by to mělo odpovídat přesnosti, s jakou jsme schopni daná data změřit. Pro většinu měřených veličin jsou to tři platné číslice, které odpovídají relativní přesnosti měření v řádu desetin procent (platná číslice je první číslice zleva nenulová; hodnota 0,0043 je uvedena s platností na dvě číslice, hodnota 123,4 na čtyři). Přesností měření v této souvislosti míníme minimální rozdíl hodnot, který můžeme při použité metodě zjistit (např. při měření výšky to jsou centimetry, při měření krevního tlaku rtuťovým nanometrem je to 5 mmHg).

Zvláštní pozornost je třeba věnovat počtu platných číslic při uvádění hodnot vyjádřených jako podíl či procento zjištěných údajů. Jejich počet by měl odpovídat velikosti souboru, ze kterého je údaj uváděn. Pokud je počet pozorování v řádu desítek, pak by uváděný údaj měl být vyjádřen v celých procentech. V případě, že počet hodnot je větší než 100 a menší než 1000, je oprávněné vyjádřit procentuální podíl s přesností na desetiny procenta a větší počet platných číslic používat jen v případě odpovídajícím velikosti souboru.

LITERATURA

1. Aczel AD. Complete Business Statistics (p 147 – 150). Irwin: Boston, 1989.
2. McAlindon T, Formica M, LaValley M, Lehmer M, Kabbara K. Effectiveness of glucosamine for symptoms of knee osteoarthritis: results from an internet-based randomized double-blind controlled trial. Am J Med 2004; 117: 643 – 649 (souhrn a komentář: Campus Medicorum, Účinky glukosaminu u osteoartritidy kolenních kloubů: výsledky z internetové randomizované, dvojitě slepé kontrolované studie, XII/2004, p 20).
3. Knížek J, Stránský P. Příspěvek k nápravě opomíjení významu síly testů ve studiích z experimentální medicíny. Čas Lék čes 2005; 144: 56 – 58.

Adresa pro korespondenci / correspondence to:
Prof. MUDr. Pravoslav Stránský, Ústav lékařské biofyziky a Oddělení výpočetní techniky, Univerzita Karlova v Praze, Lékařská fakulta v Hradci Králové, Šimkova 870, P.O. Box 38, 500 38 Hradec Králové, Česká republika / Czech Republic.
E-mail: str@lfhk.cuni.cz