Co potřebuje klinik vědět o (bio)statistice?

Neparametrická statistika

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 Praha, Faculty of Medicine at Hradec Králové, Czech Republic

Stránský P. Co potřebuje klinik vědět o (bio)statistice? Neparametrická statistika. Folia Gastroenterol Hepatol 2006; 4 (1): 38 – 40.

Souhrn. *Je vysvětlena podstata neparametrických statistických metod a důvody, proč by jim měla být věnována větší pozornost, než tomu v současnosti je. Na příkladu srovnání úspěšnosti terapie ve dvou různých nemocnicích je ukázáno, proč by hladina významnosti statistických testů neměla být automaticky 5 %.*

Klíčová slova: *neparametrická statistika, chí-kvadrát, čtyřpólová tabulka, pořadové statistiky, Spearmanův test pořadové korelace*

Stránský P. What should a clinician know about (bio)statistics? Non-parametric statistics. Folia Gastroenterol Hepatol 2006; 4 (1): 38 – 40.

Abstract. *Essence of non-parametric statistical methods and reasons for which they should be used is explained. Given example shows why significance level should not automatically equal to 5 %.*

Key words: *non-parametric statistics, chi-square, fourfold table, rank statistics, Spearman's tests of rank correlation*

Z důvodů tradice i názornosti interpretace výsledků se v testování hypotéz v medicíně nejčastěji užívají metody, které se souhrnně označují jako **parametrické**. Je tomu tak proto, že se závěry odvozují z rozložení dat ve zkoumané populaci, kde data jsou popsána pomocí veličin, kterým se říká parametry. Rozložení, které je základem pro metody vysvětlované v předešlé sérii (2-5) článků, je Gaussovo neboli normální rozložení dat.

Ve vztahu k datům získaným v klinických pozorováních je třeba připomenout, že ta v naprosté většině nemohou splňovat předpoklady, na jejichž základě je Gaussovo rozložení odvozeno. Těmito předpoklady jsou spojitost dat a dále to, že mohou nabývat hodnot od $-\infty$ do $+\infty$. Předpoklad spojitosti může být v řadě případů akceptován, horší to je s rozsahem hodnot, který veličina nabývá. U biologicky důležitých veličin existuje nějaká mez, většinou na levém konci jejich rozložení (směrem k nízkým hodnotám), která je slučitelná se životem, a hodnoty menší se v pozorovaných datech nevyskytují. Tyto veličiny mají tedy rozložení sešikmené (pozorný čtenář ví, že pravostranně) a předpoklad Gaussova rozložení dat může být a priori zamítnut[1].

Uvedená skutečnost se stávala s rozvojem aplikací statistiky v biologických vědách stále zřejmější a navíc se stále častěji (v psychologii, sociologii) objevoval požadavek testovat pozorování z dat, která nejsou dokonce ani kvantitativní, ale pouze pořadová (ordinální). Proto procedury, které nepředpokládaly, že data mají nějaké teoretické rozložení četností, že jsou tedy *distribution-free*, začaly být používány stále častěji. V kontrastu s klasickou statistikou se tyto způsoby prezentace dat a jejich testování označují

[1] Pokud jde o adjektivum „normální", tak ve statistickém textu je mu jednoznačně přiřazen výraz Gaussovo. V medicíně však toto slovo ve spojení se zjištěnými hodnotami (daty) používáme ve smyslu hodnoty fyziologické, nepatologické.

jako metody **neparametrické**. Patří sem metody dávající odpověď na otázky typu, zda pozorovaný rozdíl v četnosti výskytu nějakého jevu na určité hladině významnosti je rozdílný od rozdílu pouze náhodného. Chceme např. srovnat úspěšnost léčby stejného onemocnění ve dvou různých nemocnicích (je pozorovaný rozdíl statisticky významný?).

Základním přístupem k odpovědi na otázky tohoto druhu je porovnání pozorované četnosti jevu a četnosti očekávané.

Test, který odpovídá na tuto otázku, se nazývá χ^2 (chí-kvadrát, *chi-square*). Hodnota testového kritéria χ^2 se srovnává s odpovídající hodnotou daného (χ^2) rozložení. K pochopení podstaty testu je asi nejlepší kvantifikovat pravděpodobnost odpovědi na výsledek pokusu elementárního příkladu z výkladu o pojmu pravděpodobnost.

Představme si, že stokrát hodíme minci a zaznamenáváme, kolikrát padla panna a kolikrát lev. Vyloučíme-li možnost, že mince po dopadu bude stát na hraně (výsledek, který není teoreticky vyloučen, ale je v rozporu s našimi zkušenostmi), budeme očekávat, že četnost výskytu obou jevů bude stejná, tj. že pravděpodobnost každého p = 0,5 (50 %). Provedeme celkem 100 pokusů a zjistíme, že panna padla šedesátkrát a lev čtyřicetkrát. Je mince falešná (pozorovaný rozdíl četností není náhodný)? Odpověď na otázku dostaneme výpočtem testového kritéria χ^2 pro jednovýběrový test a srovnáním s oborem přijetí na **předem** zvolené hladině významnosti χ^2-rozdělení, např. p = 0,01 (= 1 % hladina významnosti)

Testové kritérium

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{pozorovaná} - \text{očekávaná})^2}{\text{očekávaná}}$$

kde k = 2. Po dosazení je

$$\chi^2 = \frac{10^2}{50} + \frac{-10^2}{50} = 4$$

Nulová hypotéza H_0 předpokládá, že není rozdíl mezi pozorovanými a očekávanými četnostmi, a přijmeme ji, pokud je testové kritérium menší nebo rovno hodnotě pro zvolenou hladinu významnosti a odpovídající počet stupňů volnosti. Ten se pro uvedený jednovýběrový test rovná počtu možných výsledků k – 1 (tedy v tomto příkladu jedné). Odpovídající hodnota χ^2 rozdělení pro oboustrannou hypotézu je 6,64, takže nulovou hypotézu přijmeme a prohlásíme, že mince není falešná. Pokud bychom zjišťovali, zda nehrajeme s falešnou kostkou, byla by očekávaná četnost padnutí jednoho z šesti čísel 1/6 a počet stupňů volnosti by byl pět.

Na uvedeném příkladu lze ukázat význam velikosti souboru, počtu pokusů a toho, proč je správnější uvádět četnosti výskytu pozorování jako četnosti absolutní, nikoliv relativní, např. v procentech. Provedeme-li deset hodů mince a panna padne šestkrát nebo 1000 hodů a panna padne 600x, je výsledek pokusu vždy vyjádřen 60 % ku 40 %. Testové kritérium v případě deseti hodů je 0,4, což odpovídá p = 0,527, v druhém případě je hodnota kritéria χ^2 rovna 40 s $p = 3\text{x}10^{-10}$. Při deseti hodech tedy můžeme prohlásit, že nemáme důkaz, že je mince falešná, v případě druhém je naopak oprávněné tvrzení, že s pravděpodobností hraničící s jistotou mince falešná je.

Použijeme stejný přístup, tj. srovnání pozorovaných a očekávaných četností výskytu výsledků určitého terapeutického zákroku ve dvou různých zařízeních k odpovědi na otázku, zda dosažené výsledky jsou stejné (není mezi nimi na zvolené hladině významnosti rozdíl). Výsledek takového pozorování zapíšeme do tabulky, které se říká čtyřpólová tabulka (*fourfold table*) nebo tabulka 2x2:

Nemocnice	Výsledek: úspěch	neúspěch
1	8	2
2	2	3

Tento typ srovnání snadno rozšíříme jak na více možných výsledků, tak na více zařízení, a dostaneme tak tabulku *k*x*r*.

Výpočet testového kritéria se rozšíří o druhý řádek:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{(\text{pozorovaná} - \text{očekávaná})^2}{\text{očekávaná}}$$

kde k (počet sloupců) = 2 a r (počet řádků) = 2. Očekávané četnosti vypočteme jako součin celkových četností v příslušném sloupci a celkových četností v daném řádku dělený celkovým počtem pozorování. V uvedeném příkladu jsou tedy očekávané četnosti (100/15), (50/15), (50/15) a (25/15). Testové kritérium má hodnotu 2,4, což odpovídá p = 0,121. Při použití obvyklé hladiny p = 0,05 musíme přijmout nulovou hypotézu a prohlásit, že v úspěšnosti léčby není mezi oběma zařízeními statisticky významný rozdíl.

Výsledek ukazuje na důsledek, jaký má hodnota zvolené hladiny významnosti v daném testu. Selský rozum nám říká, že budeme-li mít možnosti si vybrat, raději se necháme léčit v první nemocnici. Vysvětlení rozporu mezi závěry rozhodnutí statistického a rozu-

mového je prosté. Stanovení hladiny významnosti má být provedeno i s ohledem na důsledky rozhodnutí, ke kterému dospějeme.

Samostatnou skupinou neparametrických testů jsou **pořadové testy** (*ranking, order tests*). Název naznačuje, v čem se liší. V parametrických testech předpokládáme, že data jsou spojitá a kvantitativní. Znamená to, že s nimi můžeme smysluplně provádět aritmetické operace (např. různé transformace). U dat ordinálních, která nejsou ve své podstatě numerická (i když často výsledky získané na ordinálních datech kódujeme, označujeme pomocí číslic) a výsledky získané pomocí i jednoduchých početních úkonů zpochybňují předpokládané výsledky. Intuitivně cítíme, že aritmetický průměr vypočtený z pětihodnotové škály subjektivních potíží nedává smysl a srovnávat takto získané průměry je nesprávné. Různá „vážení" takovýchto indexů se snaží výsledkům dodat důvěryhodnost, kterou a priori nemají[2].

Podstatou těchto metod je, že zjištěná data uspořádáme podle velikosti a k dalšímu hodnocení použijeme pouze pořadové číslo příslušného pozorování. Nejlépe si tento přístup ověříme na příkladu. Máme soubor 10 uspořádaných hodnot systolického tlaku krve, které byly zjištěny při 10 návštěvách ambulance:
135, 140, 140, 150, 160, 165, 165, 165, 170, 200.

Po jejich převedení na data pro zpracování neparametrickými metodami získáme:
1, 2, 2, 4, 5, 6, 6, 6, 9, 10.

Na první pohled vidíme, že uvedený postup stírá velikost skutečných rozdílů. Rozdíl jedna po úpravě je 5 mmHg, ale i 30 mmHg. Z toho plynou dvě skutečnosti platné pro dané metody. Jsou méně citlivé na odlehlé hodnoty, což považujeme za jejich výhodu, ale na druhou stranu jsou rovněž méně citlivé na odhalení významných rozdílů, a to je většinou v rozporu s naším přáním (častěji přijímáme nulovou hypotézu).

Ze všech testů založených na pořadí je nejstarší a nejznámější Spearmanův test pořadové korelace. Je mírou vazby mezi veličinami, které jsou zjišťovány na ordinálních škálách anebo nemají gaussovské rozdělení. Stejně jako Pearsonův korelační koeficient leží v intervalu <-1, +1> a také jeho interpretace je stejná.

Vzhledem k tomu, že dnes ke všem běžně používaným parametrickým testům existují i jejich neparametrické obdoby uvedené v (1), stojí za to i u spojitých dat ověřovat jejich normalitu a v případě, že musíme nulovou hypotézu zamítnout, musíme se rozhodnout pro odpovídající metodu neparametrickou. V každém případě při prezentaci svých dat bychom měli kromě aritmetického průměru uvádět i medián s jeho mezemi spolehlivosti, a kde to má smysl i modus, pro popis variability dat kromě směrodatné odchylky i hodnoty prvního kvartilu a decilu, třetího kvartilu a devátého decilu.

[2] Právě v době psaní tohoto textu se v Lidových novinách objevily „žebříčky" fakult vysokých škol, které takovéto indexy používaly (www.lidovky.cz/zebricky).

LITERATURA

1. Siegel S. Nonparametric statistics for the behavioral sciences. Tokyo: McGraw-Hill, 1956.
2. Stránský P. Co potřebuje klinik vědět o (bio)statistice? Folia Gastroenterol Hepatol 2004; 2: 190-192.
3. Stránský P. Co potřebuje klinik vědět o (bio)statistice? Jak popsat data. Folia Gastroenterol Hepatol 2005; 3: 42-46.
4. Stránský P. Co potřebuje klinik vědět o (bio)statistice? Testování hypotéz. Folia Gastroenterol Hepatol 2005; 3: 74-76.
5. Stránský P. Co potřebuje klinik vědět o (bio)statistice? Korelační a regresní analýza. Folia Gastroenterol Hepatol 2005; 3: 110-113.

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, 500 38 Hradec Králové,
Česká republika / Czech Republic.
E-mail: str@lfhk.cuni.cz