Směruje použití mediánu ke statistickým neparametrickým postupům zpracování dat?

Okénko statistika – Analýza dat v neurologii – III. Nebojme se mediánu a robustních statistik

Vážený pane docente,
s potěšením sleduji Váš seriál o statistice na stránkách časopisu Čs. neurologie. Jednoduchou, srozumitelnou formou seznamujete se základními principy, které jsou však často lékařům cizí. Hned analyzují nějakými testy, ale vůbec se nezamýšlí nad strukturou dat. Proto považuji Vaše dílo za skvělé a moc za něj děkuji. V posledním čísle jsem také rád našel, že medián je výborný. Konečně to někdo napsal natvrdo. Laici ve statistice, a tedy i lékaři stále počítají průměry.

Mám dotaz. Když analýzy konzultuji s matematiky a statistiky, slýchávám nebo jsem to alespoň tak pochopil, že je vždy vhodnější se snažit mít normálně rozložená data, třeba i za cenu logaritmické transformace. To prý kvůli tomu, že se dají použít robustnější statistické testy jako t-test atd, tj. založené na normálním rozložení. Pokud jsem z Vašeho článku nabyl dojmu, že by bylo užitečné pracovat s mediánem (a já s tím souhlasím), musím však použít neparametrické statistické metody analýzy dat typu Mann-Whithney, Kruskal-Wallis apod. Takže už při klasifikaci dat na medián atd směruji další analýzy neparametrickým směrem. Jak to tedy je?

Ještě jednou díky za výuku nás lékařů-amatérů ve statistických metodách

MUDr. Aleš Bartoš, Ph.D.
AD Centrum, Psychiatrické centrum Praha, Ústavní 91, 181 03 Praha 8-Bohnice
a 3. LF UK a Neurologická klinika FNKV
e-mail: bartos@pcp.lf3.cuni.cz

Odpověď

Dobrý den, vážený pane doktore,
s radostí odpovídám, Vaše pochvala našeho snažení mne velmi potěšila. Jelikož dotaz, který vznášíte, je jeden z nejčastějších, věnuji mu zde určitý prostor. Dané téma následně podrobněji rozebereme přímo v některé kapitole našeho seriálu.

Výhodou normálního rozložení jednoznačně je, že má velmi dobře propracovaný a také laikům-nematematikům dostupný aparát hodnocení. Řada velmi standardních testů předpokládá existenci normálního rozložení. Jisté výsadní postavení normálního rozložení potom zajišťuje i styl výuky aplikované analýzy dat, kdy většina kurzů má čas právě na podrobné probrání tohoto fenoménu, ale již nemá dostatečný prostor pro jeho další typy. A tyto typy rozhodně existují, a to zcela legitimně. Bylo by možné jmenovat stovky velmi známých biologických parametrů, které ze své podstaty nemají normální rozložení. Metodicky lze konstatovat že:

- normální rozložení je u biologických znaků časté, nikoli ale univerzální
- s průkazností normálního rozložení bývá problém u menších vzorků dat
- výhodou normálního rozložení je dostupnost statistických testů, které mají větší sílu testu („power“) než hodnocení založená na pořadových charakteristikách
- na normální rozložení lze převést i jinak rozložené znaky tzv. transformací dat, nicméně to se nemusí vždy podařit a zdaleka ne vždy je to pro data přínosné

Výhodou práce s normálním rozložením tedy je jistá pohodlnost (vzorce jsou dány, testy všeobecně přijaty) a potom výkon (tzv. parametrické testy mají větší sílu, tedy schopnost rozpoznat neplatnost hypotézy; anebo jinými slovy – pro průkaz daného rozdílu je zde třeba o něco menší N než u neparametrických testů).

Neparametrické testy jsou naopak robustní, to je jejich hlavní výhoda. Tedy v drtivé většině nepředpokládají existenci nejen normálního, ale žádného typu rozložení. To je velmi praktické, neboť je tedy lze použít téměř vždy, odpadají starosti s odlehlými hodnotami, asymetrií rozložení apod.

K Vašemu dotazu, zda při volbě mediánu již nemohu použít např. t-test, zkusím zjednodušeně napsat v bodech toto:

- Neumožňuje-li rozložení výpočet aritmetického průměru nebo není-li normální (ani symetrické), pak musíme použít neparametrické statistiky. Z toho vyplývá, že následně i neparametrické testy, aplikace např. t-testu by byla velmi chybná.
- U normálního typu rozložení ovšem můžeme použít jak aritmetický průměr, tak i medián a další neparametrické statistiky. Taková data mohu sumarizovat všemi typy testů, parametrickými i neparametrickými. Logikou věci ale je, že pokud již máme normální rozložení a pracujeme s průměrem, pak je výhodnější parametrické testování, neboť má větší sílu testu. Povinné to ale není.
- A naopak: jsou-li data prezentována mediánem a percentily, je logické zpracovávat je pořadovými, tedy tzv. neparametrickými testy. U rozložení jiných než normálních je to nutné.
- Absolutně žádná volba pak není u ordinálních stupnic a skóre. Zde není aritmetický průměr ani definován a tedy jedinou možností jsou pořadové statistiky a neparametrické testy.

Vše tedy záleží na typu rozložení dat a dané situaci. Umím si představit prezentaci normálně rozložených dat pomocí mediánu, kvantilů i průměru a následně použití t-testu. I výběrové normální rozložení má svůj empirický 10% a 90% kvantil, své minimum a maximum a někdy je dobré je uvést pro představu o rozsahu primárních dat bez ohledu na metodiku dalšího testování. Jiným příkladem může být prezentace primárních asymetrických dat pomocí mediánu a řekněme 10%–90% kvantilů, následně normalizující transformace a statistické testování pomocí t-testu na transformovaných datech.

Ladislav Dušek