Analýza dat v neurologii

XII. Úvod do statistického usuzování – postupy a terminologie

L. Dušek, T. Pavlík, J. Koptíková
Institut biostatistiky a analýz
Masarykova univerzita, Brno

✉
doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz
Masarykova univerzita, Brno
e-mail: dusek@cba.muni.cz

V minulých dvou dílech seriálu jsme se stručně seznámili s metodologií tzv. statistického usuzování, zjednodušeně tzv. statistického testování. V poněkud teoretickém výkladu stále dlužíme ilustrativní příklad výpočtu statistického testu, a proto takový uvádíme pro jednu z nejjednodušších experimentálních situací na obr. 1. Jde o provedení jednoho náhodného výběru z populace (n = 100) a o ověření hypotézy, zda se aritmetický průměr tohoto výběru rovná nějaké stanovené hodnotě, která v příkladu vystupuje jako konstanta μ (může jít například o klinicky hraniční hodnotu, normou danou hodnotu apod.). Vidíme tedy konkrétní problém, experimentální situaci a použitý konkrétní test. Každého jistě v tuto chvíli napadne otázka: takových typových situací jsou stovky, ne-li tisíce, kdo tedy rozhoduje o správném testu, který má být aplikován? A jak se v tom vyznat?

Ponechme stranou teoretické mudrování a podívejme se na statistické testování realisticky. V „reálném světě" sedí uživatel, předpokládejme bez matematického vzdělání, před statistickým softwarem na obrazovce počítače a v podstatě provádí statistické testování mačkáním klávesy enter. Pojďme nyní stručně a pragmaticky sumarizovat minimální „know-how", které je nutné mít:

1. Uživatel musí alespoň obecně znát základní typy uspořádání experimentů a pro ně vhodné testy. Z toho bohužel nejde ustoupit, bez znalosti názvů testů si totiž nevybere ani z nabídky software. Naštěstí není typologie experimentálních situací nijak složitá; my se jí budeme zabývat v následujících dvou dílech seriálu a tento prostor nám postačí k vysvětlení kompletní sady všeobecně používaných testů.
2. Analytik si musí být vědom skutečnosti, že statistický test není nic víc než matematický vzorec aplikovaný na konkrétní data, který má pravděpodobnostně ověřit platnost stanovené hypotézy. Existuje nenulová pravděpodobnost, že výsledek bude chybný (viz minulý díl seriálu a výklad chyb α a β).
3. Statistický test sám o sobě také nemůže rozhodovat o tom, zda je pozorovaný výsledek věcně významný. V minulých dvou dílech jsme uvedli příklady, které dokumentovaly, jak lehce získáme statisticky významný výsledek pro hodnoty, které nemají věcný význam.
4. Téměř každý statistický test má nějaké předpoklady týkající se vstupních dat. Tyto předpoklady nesmí analytik nikdy ignorovat, neznalost se neomlouvá. Platí pravidlo, že smysl má pouze aplikace „správného testu na správná data". Následující text sumarizuje hlavní faktory, které správnou aplikaci testu určují.

Ověřovaná hypotéza

Je logické, že stanovená hypotéza je cílem testu a zároveň vyjadřuje smysl celého snažení. Formuluje otázku, kvůli které byl experiment jako takový proveden. Příklad na obr. 1 testuje hypotézu „rovnosti výběrového odhadu průměru a stanovené hodnoty μ". Lidštěji řečeno se ptáme, do jaké míry je naměřený rozdíl mezi odhadnutým průměrem (odhadnut měřením parametru u 100 osob) a hodnotou μ náhodný; nebo ještě jinými slovy, zda hodnota μ patří do rozdělení dat naměřených u 100 lidí nebo nikoli. Jelikož zde nepředjímáme jednostrannou nerovnost (při neplatnosti hypotézy může být naměřený průměr vyšší i nižší než hodnota μ), pak takovou hypotézu nazýváme obecně oboustrannou. Jednostranná hypotéza by přímo vymezovala, že naměřený průměr je nižší nebo vyšší než hodnota μ. Smysluplným příkladem jednostranné hypotézy je např. testování obsahu karcinogenních látek v nějaké tkáni, kdy testujeme jednostrannou hypotézu, zda je obsah nenulový, resp. vyšší než je mez detekce analytické metody. Obsah jakékoli látky logicky nemůže být menší než nula a testujeme tedy jednostrannou hypotézu.

Testovaná hypotéza určuje volbu testu, ovlivňuje potřebný rozsah experimentu a samozřejmě také interpretaci výsledku. V terminologii klinických studií se místo výše popsaného členění oboustranných a jednostranných hypotéz používá poněkud jiná, exaktnější terminologie, kterou uvádí tab. 1.

Splnění předpokladů statistického testu

Téměř každý statistický test má ve výbavě nějaké předpoklady, které podmiňují jeho správnou aplikaci. Splnění předpokladů musíme umět doložit, často i samostatně otestovat. Vzniká tak poněkud úsměvná situace, kdy aplikaci správného testu podmiňuje jiný statistický test, který ověřuje předpoklady hlavního

Řešený problém:

Naměřeno n = 100 hodnot hmotnosti pacientů určité skupiny, sumarizováno jako aritmetický průměr ($\bar{x}$) a směrodatná odchylka (s). Otázkou je, zda se průměr této skupiny liší od očekávané hodnoty (μ).

Nulová hypotéza:

H_0: $\bar{x} = \mu$ ..... oboustranná hypotéza

Statistický test a jeho předpoklady:

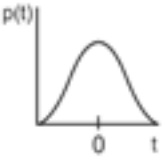
t-test pro hodnocení jednoho výběru ("one sample test"); test vhodný pokud hodnoty ve výběru 100 subjektů vykazují normální rozdělení.

Nastavená hladina významnosti testu pro ověření hypotézy H_0: α = 0,05

Testová statistika:

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

.....vzorec "testové statistiky t":
.....velikost rozdílu ($\bar{x} - \mu$) při uvážení variability znaku (hodnota s)
.....hodnota má známé rozdělení pro případ platnosti H_0 : Studentovo rozdělení



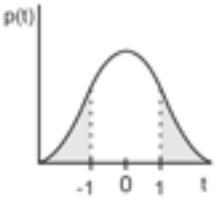
Výsledek:

(příklad ve třech variantách)
(¤ konkrétní možný číselný výstup testu)

Čím více se t jako číselný výstup testu vzdaluje od středu rozdělení (možno oběma směry), tím klesá pravděpodobnost platnosti H_0 a klesá hodnota p. Klesne-li p pod 0,05 → zamítáme H_0 na této hladině významnosti p < 0,05.
Jelikož H_0 je oboustranná, sleduje indikátor t obě strany rozdělení a hodnota p je dělena na každé straně.

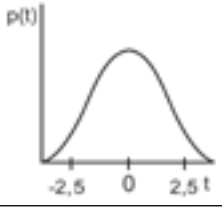
Příklad 1 (potvrzující H_0)
výsledek výpočtu:

n = 100; $\bar{x}$ = 62,0 kg; s = 10,0 kg; μ = 63,0 kg
t = –1
p = 0,319
Výsledek testu je pravděpodobný při platnosti H_0, nebyla překročena kritická mez nastavená jako α = 0,05.



Příklad 2 (zamítající H_0)
výsledek výpočtu:

n = 100; $\bar{x}$ = 62,0 kg; s = 10 kg; μ = 64,5 kg
t = –2,5
p = 0,014
Výsledek je velmi málo pravděpodobný při platnosti H_0. Hypotézu zamítáme.



Obr. 1. Příkladem dokumentovaný standardní postup testování hypotéz.

Tab. 1. Základní typy hypotéz pro konfirmační klinická hodnocení.

Typ hypotézy	Definice/Poznámky
Superiorita	Experimentální rameno je lepší než kontrolní Nejzávaznější hypotéza, jejíž potvrzení zřejmě povede k modifikaci léčebného postupu. Potvrzení hypotézy musí znamenat jednoznačnou nerovnost měřených parametrů ve prospěch experimentální skupiny.
Non-inferiorita	Experimentální rameno není ve výsledku horší než kontrolní Opět prokazujeme nerovnost, ale potvrzení hypotézy zde může znamenat i rovnost obou srovnávaných skupin.
Ekvivalence	Experimentální rameno je ve výsledku stejné jako kontrolní Testy non-inferiority nebo ekvivalence mohou označit za splnění cílů studie i situaci, kdy se experimentální rameno statisticky významně neliší od kontroly. Je zřejmé, že v takovém případě musí být doloženo, že neprůkaznost rozdílu není ovlivněna špatným nebo nedostatečným statistickým testováním. Musí být předem číselně nastaven nejmenší klinicky podstatný rozdíl v sledovaném znaku, pro který má být deklarován rozdíl. Studie musí být velmi korektně optimalizována včetně nastavení potřebné velikosti vzorku.
Odpověď jako funkce dávky	Různé hypotézy týkající se optimalizace dávky léku nebo zkoumající kombinované účinky různých terapeutik.

testu. Dále rozebereme nejpodstatnější typy předpokladů a následně uvedeme jednoduché řešení situace:

- typ dat, na která má být test aplikován. Jsou-li předepsána data ordinální nebo nominální, nemůžeme daný test použít na hodnocení spojitých hodnot. Někdy chybě zabrání již sám matematický vztah, do kterého při špatné volbě nejde dosadit, to ale neplatí vždy.
- normalita rozdělení naměřených hodnot je téměř učebnicový předpoklad. Jelikož mnoho běžně používaných testů tento předpoklad má, velmi často vzniká až dojem, že normalita rozdělení podmiňuje jakékoli úspěšné testování. To samozřejmě není pravda, obecně takto může být předpokladem existence jakéhokoli modelového rozdělení. Na ověření těchto předpokladů existují samostatné testy.
- homogenita rozptylu srovnávaných skupin. Tento předpoklad znamená, že daný test vyžaduje, aby rozptyl ve skupinách byl přibližně stejný, resp. přesněji řečeno, aby se rozptyl nebo směrodatná odchylka mezi skupinami statisticky významně nelišily. Jde o závažný předpoklad, a to rozhodně nejen z matematického hlediska. Srovnáváme-li dvě skupiny pacientů v nějakém znaku a obě skupiny se podstatně liší ve variabilitě, tak to vyžaduje vysvětlení. Příčinou může být špatně provedený výběr nebo odlehlá hodnota. Nicméně je také možné, že například ovlivnění experimentální skupiny změnilo variabilitu znaku a více „rozrůznilo" měřené hodnoty, například proto, že někteří pacienti na podnět odpověděli, jiní nikoli. Takové situace vyžadují samostatný rozbor. Na hodnocení homogenity rozptylu existují také statistické testy.
- vyrovnané počty opakování srovnávaných skupin. Předpoklad, který samozřejmě souvisí s podobnou přesností a spolehlivostí odhadů v obou srovnávaných skupinách. Pokud to experimentální situace dovoluje, měly by být přibližně stejné počty opakování standardem.

Jak postupovat při problematickém typu rozdělení hodnot?

Otázka se váže na situace, kdy data nemají normální rozdělení nebo existují problémy s rozptylem, případně s odlehlými hodnotami. Zde můžeme čtenáře potěšit, a to hned dvakrát. Zaprvé, sám předpoklad normality je v praxi málokdy naplněn a nemáme-li soubor dat o velikosti alespoň 100 opakování, stejně normalitu exaktně neprokážeme. Celá diskuze o normalitě se redukuje na prověření odlehlých hodnot a symetrie rozdělení (alespoň přibližná rovnost mediánu a průměru, pravidlo ± 3 s). Zvláště u malých souborů je v datech často vidi-

telný problém, který brání nasazení testů vyžadujících normalitu rozdělení. A zde přichází slíbené druhé pozitivum – jednoduchým řešením je nasazení tzv. neparametrických testů.

- Parametrické testy jsou založeny na matematickém vztahu, který provádí závěr o hodnotě parametru nějakého modelového pravděpodobnostního rozdělení. Ukázkově nám opět poslouží příklad na obr. 1, kde hodnotíme aritmetický průměr jako parametr normálního rozdělení. Pokud námi naměřený soubor toto rozdělení nebude mít, bude celý test znehodnocen: např. průměr nebude využitelný jako odhad středu rozdělení, neboť bude ovlivněn odlehlými hodnotami; nebo odlehlé hodnoty silně zvýší numericky směrodatnou odchylku s, která je ve jmenovateli vztahu pro výpočet statistiky t. Při vysokém jmenovateli s nebude vztah schopen citlivě reagovat na rozdíl naměřeného průměru a hodnoty μ v čitateli apod.
- Neparametrické testy naopak nepotřebují předpoklad o modelovém rozdělení, jsou na něm nezávislé. Jsou založeny buď na hodnocení četnosti odchylek hodnot, nebo převádějí hodnocená data na pořadí a s ním dále pracují (někdy se jim říká pořadové testy, „rank"). I tyto testy sice mají své předpoklady (např. shoda typu výběrového rozdělení srovnávaných sou-

Tab. 2. Nejběžněji užívané parametrické a neparametrické testy pro jednoduché experimentální situace.

Testy o odhadu aritmetického průměru (parametrické testy)

Předpoklady: Normální rozdělení všech výběrů, které jsou odhadem průměru sumarizovány.

(1) Test pro odhad jednoho výběru („one-sample" t-test)
Testuje hypotézy o hodnotě jednoho odhadu průměru, srovnáním s předpokládanou hodnotou (μ). *Příklad:* obr. 1.

(2) Srovnání průměru dvou nezávislých výběrů („two-sample t-test: independent")
Srovnává dva nezávislé výběry, předpokládá u obou výběrů normální rozdělení a homogenitu rozptylu. *Příklad:* srovnání průměrné hodnoty znaku X u pacientů a u zdravých kontrol.

(3) Srovnání dvou párově uspořádaných výběrů („two-sample t-test: paired")
Test hypotéz o hodnotě průměrného rozdílu mezi dvěma závislými (= párovými) výběry; testuje, *o kolik* se oba výběry liší a předpokládá normální rozdělení diferencí původních hodnot. Párové uspořádání určuje cíl experimentu, tj. hodnocení změny hodnot. *Příklad:* změna hmotnosti u operovaných pacientů – srovnáním hodnot před výkonem a po něm; je testován průměrný rozdíl hmotnosti bez ohledu na rozdělení původních hodnot před operací.

Neparametrické testy srovnávající dva výběry

Testy převádějící spojitá čísla na ordinální (pořadové) škály, i extrémní hodnoty jsou tak převedeny na pořadí náležející jim v daném souboru. Představují univerzálnější alternativu k parametrickým testům.

(1) Mann-Whitney test (MW test; U test)
Pořadový test určený k srovnání dvou nezávislých výběrů, alternativa nezávislého t-testu.

(2) Wilcoxonův test (W test)
Pořadový test určený k srovnání dvou párově uspořádaných výběrů, alternativa párového t-testu. Analyzuje diference dvou párových výběrů a testuje hypotézy typu *o kolik* se liší.

(3) Mediánový test („median test")
Test pro srovnání dvou nezávislých výběrů. Převádí spojitá čísla na binární kód podle toho, zda se vyskytují pod anebo nad společným mediánem obou výběrů. Následně hodnotí vyrovnanost výskytu hodnot v obou výběrech podle jejich pozice k celkovému mediánu.

(4) Znaménkový test („sign test")
Neparametrický test pro srovnání párových experimentů s nominálním výstupem nebo s výstupem typu (ano/ne). Hodnotí shodu párových výběrů v nominálních kategoriích. Je alternativou párového t-testu, ale má výrazněji sníženou sílu testu a vyžaduje větší vzorky (n > 50).

borů), ale jejich naplnění je v praxi mnohem jednodušší než u testů parametrických. Při běžných analýzách nabídka neparametrických procedur plně kryje nabídku parametrických testů, stačí tedy pouze tyto testy znát a umět je nalézt v menu statistických software. Často jsou totiž umístěny na jiném místě nebo v jiném modulu než testy parametrické, takže uživatel musí vědět, co hledá a proč to hledá.

O neparametrických testech se často hovoří jako o robustnější alternativě testů parametrických. To znamená, že jsou méně náchylné na ovlivnění odlehlými hodnotami a jinými odchylkami v datech. Jejich aplikaci můžeme skutečně výrazně doporučit. Především laický uživatel, který si není jistý splněním předpokladů parametrických testů, neudělá použitím této alternativy žádnou chybu. Snad jedinou viditelnější nevýhodou je, že neparametrické testy ve srovnání s parametrickými jsou tzv. konzervativní. To znamená, že mají menší tendenci zamítat nulovou hypotézu, neboli mají menší sílu testu. U řady neparametrických testů ale není tento rozdíl oproti parametrickým nijak dramatický, a lze mu tedy jednoduše čelit mírným navýšením velikosti vzorku. Tyto postupy rozebereme v některém z dalších dílů seriálu.

Kromě neparametrické alternativy lze samozřejmě řešit problémy s rozdělením hodnot i jinak, např. vhodnou normalizující transformací. Transformacím dat jsme věnovali V. díl našeho seriálu. Např. často používaná logaritmická transformace nás zbaví asymetrie rozdělení zprava, anebo zajistí homogenitu rozptylu ve srovnávaných souborech. Někdy ale může být použití transformace problematické nebo svou složitostí zastíní i vlastní cíl výzkumu. Za těchto okolností lze opět doporučit nasazení neparametrických metod jako jednoduché a přímočaré řešení. Základní přehled nejběžnějších testů uvádí tab. 2, detailnímu popisu testování se budeme věnovat v dalších dílech seriálu.