Analýza dat v neurologii

XV. Vyzkoušejte zvláštní typ neparametrického testování hypotéz: permutační testy – obecné aplikace

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 úvodu této části se vrátíme k Fisherovu exaktnímu testu, který jsme v předcházejícím díle představili jako užitečný příklad tzv. **permutačního testování**. Připomeňme, že tento test se používá k hodnocení závislosti (asociace) dvou znaků nabývajících pouze dvou hodnot. Typickým záznamem takových pozorování je 2 × 2 tabulka četností. Pokud zamítáme nulovou hypotézu, znamená to, že kombinace hodnot obou znaků nenastávají v sledované populaci náhodně a je mezi nimi závislost. Význam Fisherova exaktního testu spočívá v tom, že při hodnocení nulové hypotézy pracuje přímo s kombinacemi naměřených hodnot, kterým přiřazuje pravděpodobnost výskytu. Naměřená tabulka četností je tak simulačně přeskupována při zachování součtu řádků a sloupců a v testu hledáme pravděpodobnost takových kombinací, které jsou ještě více vzdáleny od platnosti nulové hypotézy (jsou „extrémnější“) než tabulka pozorovaná. Vlastní postup výpočtu opakuje příklad 1, příklad 2 uvádí aplikaci na reálných klinických datech.

Příklad 1

Fisherův exaktní test generuje náhodným procesem varianty pozorované tabulky četností při zachování součtů řádků i sloupců. V konečném výpočtu kalkuluje sumární pravděpodobnost výskytu variant, které jsou z hlediska platnosti nulové hypotézy ještě extrémnější než varianta pozorovaná. Tato pravděpodobnost je zároveň pravděpodobností chyby I. druhu při zamítnutí nulové hypotézy o náhodném vztahu řádků a sloupců tabulky. Termínem „extrémnější“ se rozumí varianty tabulky s menší pravděpodobností výskytu než varianta pozorovaná, ať již v jednom směru (jednostranný test), nebo v obou směrech (oboustranný test). Pokud máme například následující 2 × 2 kontingenční tabulku:

2	3
6	4

Potom všechny varianty tabulky při zachování součtů řádků a sloupců jsou:

0 5	1 4	2 3	3 3	4 1	5 0
8 2	7 3	6 4	5 5	4 6	3 7

K nim příslušející pravděpodobnosti výskytu pak:

0,007 0,093 0,326 0,392 0,163 0,019

Světle zelené tabulky představují varianty extrémnější než pozorovaná tabulka, a to ve stejném směru (ve smyslu asymetrie rozložení frekvencí v tabulce). „Stejný směr“ od pozorované tabulky je definován postupem, kdy od nejmenší hodnoty v pozorované tabulce odečteme 1 a dopočítáme zbývající četnosti buněk tabulky při zachování součtů sloupců a řádků. **Tmavě zelené tabulky** potom představují extrémnější varianty v opačném směru. Extrémnější varianty poznáme pomocí pravděpodobnosti, která je menší nebo stejná jako pro pozorovanou variantu. Pravděpodobnosti těchto extrémních variant v testu sčítáme.

Jednostranná p-hodnota pro hodnocenou tabulku je tedy:
0,326 + 0,093 + 0,007 = **0,426**.
Oboustranná p-hodnota je:
0,326 + 0,093 + 0,007 + 0,163 + + 0,019 = **0,608.**

Pravděpodobnost čtvrté varianty (0,392) není započítána, protože je méně extrémní (je pravděpodobnější) než pozorovaná tabulka.

Příklad 2

Jako příklad aplikace Fisherova exaktního testu uveďme příklad sledování počtu dvou typů nežádoucích příhod (NP I, NP II) u dvou různě léčených skupin pacientů (léčba 1 a léčba 2). Při celkových N = 10 pozorováních jsme u léčby 1 zachytili 5 NP typu I a 1 NP typu II. Léčba 2 vedla ve čtyřech případech k NP II, typ I nebyl pozorován.

	Léčba 1	Léčba 2	
NP I	5	0	Σ = 5
NP II	1	4	Σ = 5
	Σ = 6	Σ = 4	N = 10

Výpočet pravděpodobnosti *P* pro tuto tabulku pozorovaných četností je (viz též XIV. díl seriálu):

$$p = \frac{5!^2\ 6!\ 4!}{10!\ (5!\ 0!\ 1!\ 4!)} = 0{,}0238$$

Pro ostatní možné tabulky je *P*:

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} p = 0{,}2381 \qquad \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} p = 0{,}2381$$

$$\begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix} p = 0{,}4762 \qquad \begin{bmatrix} 1 & 4 \\ 5 & 0 \end{bmatrix} p = 0{,}0238$$

Celková suma pravděpodobností výskytu různých variant tabulky je 1. Suma pravděpodobnosti menších nebo rovných hodnotě *P* pozorované tabulky je 0,0476, což je hodnota menší než 0,05, a tedy indikující statisticky významný vztah mezi typem léčby a výskytem nežádoucích událostí.

Pokračujme nyní dále obecnějším vysvětlením permutačních (obecně randomizačních) testů. Tyto testy řadíme mezi tzv. neparametrické postupy, neboť nepracují s referenčními hodnotami teoreticky odvo-

zených distribučních funkcí, jako to dělají „klasické" testy parametrické, např. t-test. Klasickým postupem v testování hypotéz je výpočet hodnoty *p* odrážející platnost nulové hypotézy na základě teoretického rozdělení pravděpodobnosti odpovídajícího použité testové statistice. V mnoha případech však rozdělení pravděpodobnosti testové statistiky není známo, důvodem může být např. nedostatečná znalost problému nebo velmi malý počet dostupných pozorování. Testování může v takovém případě vést k mylným závěrům. Abychom se vyhnuli nutným předpokladům o rozdělení testových statistik, můžeme pro odhad hodnoty *p* použít tzv. **permutační algoritmus**, jednu z metod, která pracuje s opakovaným vzorkováním naměřeného souboru dat (dalšími příklady těchto metod jsou **bootstrap, jackknife a krosvalidace**, o kterých se zmíníme na konci kapitoly).

Permutační testování je založeno na opakovaném vzorkování neboli náhodném přeskupování naměřeného souboru. Cílem je posoudit variabilitu možných výsledků, které lze získat náhodným přeskupováním pozorovaného souboru dat. Přitom pracujeme stále pouze s naměřenými hodnotami. Celý postup získal jméno od pojmu **permutace**, což znamená přeskupení 1 až *n* čísel. Pro číselnou řadu 1 až 6 můžeme jako příklad uvést následující permutace:

(1, 2, 3, 4, 5, 6)
(1, 3, 2, 4, 5, 6)
(4, 5, 2, 6, 1, 3)
(3, 2, 1, 6, 4, 5)

Jde ovšem pouze o malou ukázku, neboť permutací *n* objektů je celkem *n!* („n faktoriál"), v našem případě tedy 6! = 720.

Nyní je zřejmé, že v permutačních testech jde o náhodné přeskupení již naměřených hodnot. Všechny permutace přitom považujeme za stejně pravděpodobné. Vlastní postup výpočtu potom provádíme tak, že z naměřených hodnot generujeme všechny možné nebo velmi mnoho permutací a pro každou variantu spočítáme příslušnou testovou statistiku. Principem je srovnání pozorované testové statistiky s testovými statistikami, které takto získáváme teoreticky ze stejného datového souboru, kde je přiřazení jednotlivých hodnot do studovaných skupin náhodné. Ještě jinak řečeno, permutační test je založen na výpočtu všech možných hodnot testové statistiky, které lze získat opakovaným přeskupením původního souboru dat tak, že v rámci každého opakování zůstane zachován jak celkový počet pozorování (celkové *N*), tak počet pozorování náležících do jednotlivých skupin. Máme-li uspořádání experimentu, v němž srovnáváme dvě skupiny hodnot (pozorování), pak si lze postup představit tak, že v jedné permutaci náhodně vybereme N_1 z celkového počtu *N* pozorovaných hodnot, které přiřadíme do první skupiny, a zbylých N_2 hodnot budeme považovat za hodnoty náležící do druhé skupiny (přitom vždy platí $N_1 + N_2 = N$).

Pro každé opakování tak dostaneme hodnotu testové statistiky (například statistiky *t*) $t_1, \ldots, t_M$ (celkem tedy *M* hodnot, kde *M* je celkový počet dostupných nebo provedených permutací). Výslednou hodnotu *p* pak vypočítáme jako podíl počtu testových statistik, které byly v absolutní hodnotě větší než původně pozorovaná testová statistika *t* (tedy představují extrémnější výsledky experimentu: $m = |t_i| \geq t$, kde $i = 1, \ldots, M$), ku celkovému počtu provedených permutací *(M)* : $p = m/M$.

Je patrné, že tento postup pouze ověřuje základní myšlenku, která je na pozadí každého statistického testu. Dosaženou hladinu významnosti totiž vždy čteme jako pravděpodobnost, že právě naměřený výsledek dostaneme náhodou, když vybíráme z jednoho nebo více základních souborů. Je-li tato pravděpodobnost malá, zamítáme nulovou hypotézu a výsledek testu označujeme za nenáhodný („významný"). Permutační postup nedělá nic jiného, než že tuto skutečnost simuluje a generuje mnoho takových náhodných kombinací a hodnotí možné výstupy.

Je-li počet pozorovaných hodnot *(N)* příliš velký, je nemyslitelné provést všechny dostupné permutace *(M)*. Potom pro výpočet uvažujeme pouze náhodnou podmnožinu permutací *B*, kde $B < M$. Velkou výhodou permutačního testování je fakt, že jej lze použít pro jakoukoliv testovou statistiku. Tu si tedy můžeme vybrat tak, aby nejlépe vyhovovala našim potřebám, a zároveň se nemusíme starat o její rozdělení pravděpodobnosti. Avšak i permutační testování má své limity. Zásadní podmínkou zde je předpoklad zaměnitelnosti pozorovaných hodnot v obou srovnávaných souborech. To jinými slovy znamená, že by oba soubory neměly mít výrazně odlišnou variabilitu a experimentální uspořádání by náhodnou zaměnitelnost nemělo vylučovat. Jistou nevýhodou je pak omezená aplikovatelnost permutačních postupů na velmi malých datových souborech, neboť při malém *N* (cca do 10) je poměrně malý také počet dostupných permutací, což může vést k nepřesnému odhadu hodnoty *p*. Příklad 3 přibližuje výpočet permutačního testu na situaci, kde je srovnávána hmotnost dvou nestejně velkých skupin pacientů. Interpretace výsledné hodnoty *p* je zde stejná jako pro klasický *t*-test.

Ještě před cca 15 lety bylo nasazení permutačního testování jistou formou úniku od matematických předpokladů parametrických testů. Avšak zcela nové využití pro permutační postupy nabídlo hodnocení dat molekulárně biologických experimentů, kde lze jen stěží předpokládat rozdělení naměřených hodnot, a použití permutačního algoritmu na mnohorozměrných datech zachovává při výpočtu jejich korelační strukturu. Aplikací permutačních testů je také zajištěna kontrola variability měření v rámci jednoho experimentu.

Jak již bylo uvedeno výše, permutační algoritmy mají velmi blízko i k dalším postupům založeným na opakovaném vzorkování původních dat. Dobře známou a používanou metodou je tzv. **bootstrap**, který se většinou využívá pro odhady intervalu spolehlivosti sledované charakteristiky, jako je např. průměr nebo medián. Bootstrap je založen na principu opakovaného vzorkování s vracením, kdy pro vytvoření nového vzorku dat může být každý prvek použit více než jednou, právě jednou anebo není použit vůbec (ovšem opět se zachováním celkové velikosti souboru *N* i velikosti jednotlivých skupin). Dalším obdobným postupem je tzv. **jackknife**, používaný též pro odhad variability měřených charakteristik. Zde je opakovaný výpočet sledované charakteristiky prováděn vždy s vynecháním právě jednoho pozorování. Tento postup nám stejně jako v případě metody bootstrap poskytuje představu o rozsahu hodnot, ve kterých se námi sledovaná charakteristika může pohybovat, budeme-li považovat naměřená data za reprezentativní vzorek z cílové populace. Třetí hojně používanou metodou z tohoto spektra je tzv. **krosvalidace**, která je nejčastěji používána pro validaci stochastických modelů. Jejím principem je opakované rozdělení datového souboru

Tab. 1. Ukázky možných permutací.

Skupina pacienta	Hmotnost pacienta (kg)	Pořadí permutace 1	2	3	…	6 435
A	91,5	A	B	B	…	B
A	79,8	B	B	B	…	B
A	66,2	A	A	A	…	A
A	70,7	A	B	A	…	B
A	63,4	B	B	A	…	A
A	77,7	B	B	B	…	A
A	71,9	B	A	A	…	B
B	83,9	A	B	A	…	A
B	92,2	B	B	A	…	A
B	85,4	A	A	B	…	A
B	99,2	A	A	B	…	A
B	77,5	A	A	A	…	B
B	80,8	B	A	B	…	B
B	91,6	B	B	B	…	B
B	86,2	B	A	B	…	B
Testová statistika	**2,900**	**0,429**	**0,341**	**3,106**	**…**	**0,798**

na dvě části: trénovací (obvykle větší část původního souboru) a testovací (obvykle menší část původního souboru) s tím, že v první fázi je model vytvořen na trénovacím souboru a následně je na testovacím souboru zjištěna jeho chybovost, tedy nepřesnost odhadu cílové proměnné. Opakování tohoto postupu nám dává užitečnou informaci o možné chybovosti modelu při nasazení v reálné praxi.

Příklad 3.

Příklad výpočtu permutačního testu pro srovnání hmotnosti dvou skupin pacientů.

Naměřená data:

- Skupina A: $N_1 = 7$, $\bar{x}_1 = 74{,}5$ kg
- Skupina B: $N_2 = 8$, $\bar{x}_2 = 87{,}1$ kg

Hodnota statistiky t (srovnání dvou nezávislých výběrů): $t = 2{,}900$. Tomu odpovídající hodnota p: $p = 0{,}015$.

Permutační test je zde použitelný, neboť je splněna podmínka teoretické záměny subjektů. K subjektům budeme simulačně přistupovat jako k jednomu základnímu souboru s celkovým N = 15 a jednotlivé permutace pomohou prověřit, zda je pozorovaná varianta extrémní, tedy málo pravděpodobná při platnosti nulové hypotézy. Nulovou hypotézou zde je rovnost obou skupin pacientů v průměrné hmotnosti.

Permutační test:

- Pro výpočet je použita statistika t pro dva nezávislé výběry.
- Pro $N_1 = 7$ a $N_2 = 8$ je možno provést celkem 6 435 jedinečných permutací.
- Ukázky možných permutací jsou uvedeny v tab. 1 i s výslednými hodnotami statistiky t.

Výpočet hodnoty p v permutačním testu:

- Hodnota původní statistiky $t = 2{,}900$
- Celkový počet provedených permutací $M = 6\,435$
- Počet permutací, pro které je absolutní hodnota testové statistiky: $|t_i| \geq t = 2{,}900$, je zde $m = 59$ **=> => $p = m/M = 59/6\,435 = 0{,}009$**

Permutační test tedy potvrdil statistickou významnost rozdílu průměrné hmotnosti v obou skupinách pacientů.