Analýza dat v neurologii

XVII. Neparametrické testy jako alternativa *t*-testu

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

Minulý díl seriálu jsme věnovali *t*-testu, který jsme označili za „zlatý standard statistického testování". A tím také skutečně je, alespoň viděno očima biologů a lékařů, kteří do analýzy dat pronikli. Málokoho *t*-test minul, je součástí úvodní výuky biostatistiky, najdete jej v každé učebnici. Méně zkušené kolegy jsme možná překvapili tím, že *t*-test má tři základní formy, které nelze zaměnit (pro dva nezávislé výběry, pro párové uspořádání experimentu a pro jeden náhodný výběr). Již samotný výběr správného *t*-testu tedy vyžaduje jistou znalost problematiky a bohužel nevystačíme pouze s hledáním vhodného příkazu v menu statistického software. Ještě závažnějším problémem ale mohou být předpoklady pro správnou aplikaci *t*-testu, které jsou nekompromisní. *T*-test pracuje s výběry ze základního normálního rozdělení, a tedy vyžaduje splnění předpokladu normálního rozdělení u sledované proměnné. Výrazně odlehlé hodnoty nebo asymetrické výběrové rozdělení znehodnotí práci, a aplikace *t*-testu vede k nesmyslným výstupům.

Řešení je jednoduché a v zásadě existují dvě možnosti. Jednak můžeme ověřit splnění předpokladů *t*-testu a v případě problémů s normalitou hodnot rozdělení normalizovat (např. vhodnou transformací) nebo zdůvodněně vylučovat odlehlé hodnoty. Tento postup může být

Příklad 1. Wilcoxonův test pro dva nezávislé výběry*.

1. Byly provedeny experimenty A a B, v nichž byly měřeny počty buněk u pokusných zvířat n_1(A) = 7, n_2(B) = 5.
2. Nulová hypotéza (oboustranná): Počty buněk u pokusných zvířat jsou mezi experimenty A a B shodné (nulová hypotéza *nesrovnává střední hodnotu experimentálních skupin, předpokládá shodu rozdělení sledované proměnné v obou skupinách)*.
3. Počtům buněk je přiřazeno pořadí bez ohledu na experiment, ze kterého pochází *(pořadí je možno přiřadit jak od nejmenší do největší hodnoty, tak opačně; vyskytne-li se více shodných hodnot, je jim přiřazen průměr pořadí, která na ně připadají)*.
4. Pro experiment A a B jsou spočteny sumy pořadí R_1(A) = 61, R_2(B) = 17.
5. Počty měření a sumy pořadí v jednotlivých experimentech jsou dosazeny do následujících vztahů

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U' = n_1 n_2 - U$$

$$U = 7*5 + \frac{7*8}{2} - 61 = 2 \qquad U' = 7*5 - 2 = 33$$

Experiment	Počet buněk	Pořadí počtu buněk
A	218 790	4
A	229 086	7
A	231 660	8
A	235 521	9
A	238 095	10
A	241 956	11
A	248 391	12
B	209 781	1
B	212 355	2
B	216 216	3
B	222 651	5
B	225 225	6

6. Pro porovnání s kritickou hodnotou testu je vybrána větší z hodnot *U* a *U'*, v tomto případě *U'* = 33.
7. Kritická hodnota testové statistiky *U* pro α = 0,05 je $U_{0.05(2),\,5,\,7}$ = 30 *(pro velké vzorky existuje aproximace testu na normální rozdělení, výsledky testu jsou poté porovnávány s kritickou hodnotou normálního rozdělení)*.
8. Hodnota *U'* = 33 je větší než nalezená kritická hodnota $U_{0.05(2),\,5,\,7}$ = 30, což vede k zamítnutí nulové hypotézy.
9. Závěr: Na hladině významnosti α = 0,05 byl prokázán statisticky významný rozdíl v počtu naměřených buněk mezi experimenty A a B.

* Tento test lze také velmi často nalézt pod alternativním názvem Mann-Whitney U test, podle autorů, kteří krátce po F. Wilcoxonovi navrhli vlastní kalkulaci statistiky U (nepočítá výše uvedenou sumu pořadí ve skupinách A a B, ale každé měření ze skupiny A je srovnáváno s každým měřením ze skupiny B). Oba výpočty jsou ve výsledku ekvivalentní, postup dle Manna a Whitneye nabízí poněkud širší pravděpodobnostní interpretaci.

Příklad 2. Wilcoxonův párový test pro dva závislé výběry.

Před pokusným zásahem	Po pokusném zásahu	Rozdíl	Pořadí absolutních hodnot rozdílů	Pořadí rozdílů s vyznačením směru
26 412	25 668	744	4,5	4,5
26 040	25 296	744	4,5	4,5
26 784	27 342	–558	3	–3
26 784	25 854	930	7	7
26 412	26 598	–186	1	–1
27 156	26 226	930	7	7
27 714	26 598	1 116	9,5	9,5
27 900	26 970	930	7	7
26 412	25 296	1 116	9,5	9,5
27 528	27 156	372	2	2

1. Počty buněk u pokusných zvířat byly změřeny před a po aplikaci pokusného zásahu; $n = 10$.
2. Nulová hypotéza (oboustranná): není rozdíl mezi počty buněk před a po pokusném zásahu .
3. Je spočítán rozdíl mezi hodnotami před a po pokusném zásahu *(nulové rozdíly jsou z výpočtu vyřazeny)*.
4. Absolutní hodnotě rozdílů je přiřazeno pořadí; vyskytnou-li se stejné hodnoty, je jim přiřazen průměr pořadí, která na ně připadají .
5. Je sečtena suma pořadí připadající na kladné a záporné rozdíly
$$T_+ = 4{,}5 + 4{,}5 + 7 + 7 + 9{,}5 + 7 + 9{,}5 + 2 = 51$$
$$T_- = 3 + 1 = 4$$
6. Pro porovnání s kritickou hodnotou testu je vybrána menší z hodnot T_+ a T_-, v tomto případě $T_- = 4$.
7. Kritická hodnota testové statistiky T pro $\alpha = 0{,}05$ je $T_{0.05(2),\,10} = 8$.
8. Hodnota $T_- = 4$ je menší než nalezená kritická hodnota $T_{0.05(2),\,10} = 8$, což vede k zamítnutí nulové hypotézy.
9. Závěr: na hladině významnosti $\alpha = 0{,}05$ byl prokázán statisticky významný rozdíl v počtu naměřených buněk před a po pokusném zásahu.

Příklad 3. Mediánový test pro dva nezávislé výběry.

1. Byly provedeny experimenty A a B, v jejichž rámci byly měřeny počty buněk u pokusných zvířat $n_1(A) = 7$, $n_2(B) = 5$.
2. Nulová hypotéza (oboustranná): mediány počtu buněk u pokusných zvířat jsou mezi experimenty A a B shodné *(na rozdíl od Mann-Whitney U testu, který srovnává překryv rozdělení hodnot, jde v tomto případě o srovnání odhadu střední hodnoty experimentálních skupin – mediánu)*.
3. Je spočítán společný výběrový medián ze sloučených dat obou experimentálních skupin, medián = 231 660.
4. Pro experiment A a B je zjištěn počet hodnot nad společným mediánem a počet hodnot shodných nebo menších než společný medián.
5. Počet hodnot nad a pod společným mediánem je vynesen do tabulky

	A	B	Celkem
Větší než společný medián	3	2	5
Rovno nebo menší než společný medián	4	3	7
Celkem	7	5	12

6. Statistická významnost vztahu mezi experimentálními skupinami a podílem hodnot nad/pod mediánem je testována χ^2 testem:
$$\chi^2 = \frac{n\left(\left|f_{11}f_{22} - f_{12}f_{21}\right| - \frac{n}{2}\right)^2}{C_1 C_2 R_1 R_2} = 0{,}245$$
Kde f_{xy} jsou četnosti v jednotlivých buňkách tabulky a C_x a R_y četnosti v sloupcích a řádcích tabulky.
7. Kritická hodnota testové statistiky χ^2 pro $\alpha = 0{,}05$ je $\chi^2_{0.05,1} = 3{,}841$.
8. Hodnota $\chi^2 = 0{,}245$ je menší než nalezená kritická hodnota $\chi^2_{0.05,1} = 3{,}841$ a nulovou hypotézu tedy není možné zamítnout.
9. Závěr: na hladině významnosti $\alpha = 0{,}05$ nebyl prokázán statisticky významný rozdíl v mediánu počtu naměřených buněk mezi experimenty A a B.

Experiment	Počet buněk
A	218 790
A	229 086
A	231 660
A	231 660
A	238 095
A	241 956
A	248 391
B	218 000
B	230 000
B	231 000
B	236 000
B	238 000

Příklad 4. Znaménkový test pro dva závislé výběry.

Před pokusným zásahem	Po pokusném zásahu	Rozdíl	Směr rozdílu
26 412	25 668	744	+
26 040	25 296	744	+
26 784	24 342	2 442	+
26 784	25 854	930	+
26 412	26 598	–186	–
27 156	26 226	930	+
27 714	26 598	1 116	+
27 900	26 970	930	+
26 412	25 296	1 116	+
27 528	27 156	372	+

1. Počty buněk u pokusných zvířat byly změřeny před a po aplikaci pokusného zásahu; $n = 10$.
2. Nulová hypotéza: Není rozdíl mezi počty buněk před a po pokusném zásahu (takto aplikovaný znaménkový test hodnotí symetrii rozdílů kolem hodnoty nula).
3. Je spočítán rozdíl mezi hodnotami před a po pokusném zásahu.
4. Rozdíly jsou převedeny na kladné (+) a záporné (–) $n_+ = 9$, $n_- = 1$ *(nulové rozdíly jsou z výpočtu vyřazeny)*.
5. V případě, že není rozdíl mezi počty buněk před a po pokusném zásahu, je v ideálním případě podíl kladných a záporných rozdílů shodně 0,5 (50 %); pro další testování je problém převeden na binomický test pro binomické rozdělení s $q = 0{,}5$.
6. Hledáme pravděpodobnost náhodného výskytu binomického rozdělení s $n = 10$ a $q = 0{,}5$ pro četnost jevu ≤ 1 (počet záporných diferencí) nebo ≥ 9 (počet kladných diferencí):
$$p(X \leq 1 \text{ nebo } X \geq 9) = 0{,}00098 + 0{,}00977 + 0{,}00977 + 0{,}00098 = 0{,}02$$
7. Hodnota $p = 0{,}02$ představuje pravděpodobnost, že náhodně vznikne stejné nebo extrémnější rozmístění rozdílů kolem nuly než v našem experimentu.
8. Hodnota $p = 0{,}02$ je menší než $\alpha = 0{,}05$, což vede k zamítnutí nulové hypotézy.
9. Závěr: na hladině významnosti $\alpha = 0{,}05$ byl prokázán statisticky významný rozdíl v počtu naměřených buněk před a po pokusném zásahu.

Příklad 5. Znaménkový test hodnoty mediánu.

1. V průběhu experimentu byly změřeny počty buněk u pokusných zvířat, $n = 10$.
2. Nulová hypotéza: není rozdíl v mediánu počtu buněk zjištěném v experimentu a mediánem cílové populace (26 100).
3. Je zjištěn počet hodnot nad (+) a pod (–) mediánem cílové populace $n_+ = 9$, $n_- = 1$ *(hodnoty shodné s mediánem cílové populace jsou z výpočtu vyřazeny)*.
4. Testování probíhá znaménkovým testem, který hodnotí symetrii rozdílů kolem zadané hodnoty (v tomto případě mediánu cílové populace).
5. V případě, že není rozdíl mezi mediánem experimentálních hodnot a mediánem cílové populace, je v ideálním případě podíl hodnot nad a pod mediánem cílové populace shodně 0,5 (50 %); pro další testování je využit binomický test pro binomické rozdělení s $q = 0{,}5$ a $n = 10$.
6. Hledáme pravděpodobnost náhodného výskytu binomického rozložení s $n = 10$ a $q = 0{,}5$, pokud je počet výskytu jevu ≤ 1 (počet hodnot pod mediánem cílové populace) nebo ≥ 9 (počet hodnot nad mediánem cílové populace):
$$p(X \leq 1 \text{ nebo } X \geq 9) = 0{,}00098 + 0{,}00977 + 0{,}00977 + 0{,}00098 = 0{,}02$$
7. Hodnota $p = 0{,}02$ představuje pravděpodobnost, že náhodně vznikne stejné nebo extrémnější rozmístění rozdílů kolem mediánu cílové populace než v našem příkladu.
8. Hodnota $p = 0{,}02$ je menší než $\alpha = 0{,}05$, což vede k zamítnutí nulové hypotézy.
9. Závěr: na hladině významnosti $\alpha = 0{,}05$ byl prokázán statisticky významný rozdíl v počtu naměřených buněk mezi experimentem a testovaným mediánem cílové populace.

Experiment	Hodnoty nad/pod testovaným mediánem
26 040	–
26 412	+
26 412	+
26 412	+
26 784	+
26 784	+
27 156	+
27 528	+
27 714	+
27 900	+

Cílová hodnota pro srovnání: 26 100

pro laika velmi pracný a hlavně u malých vzorků o $n < 20$ až nesmyslný, neboť každá vyloučená hodnota představuje nezanedbatelné procento z velikosti vzorku. Nadto – budeme-li dostatečně přísní – předpoklad normálního rozdělení není skoro nikdy přesně splněn, což opět platí především pro malé vzorky. Druhou možností je **aplikace neparametrických testů**, které nemají žádné nebo velmi minimální předpoklady na rozdělení hodnot náhodné proměnné. Pojmem neparametrický zde rozumíme nezávislý na rozdělení. Aplikací takového testu se tedy zbavíme trápení, o rozdělení náhodné proměnné nemusíme mít větší vědomosti. Je pouze nutné vědět, který typ testu je vhodný pro jakou situaci.

Obecně rozlišujeme tři základní typy neparametrických testů, takže někdy používané tvrzení, že jsou zcela univerzální, není úplně pravda. Jedním typem, tzv. permutačními testy jsme se již zabývali v díle 14 a 15 našeho seriálu. Připomeňme zde, že jde o testy skutečně velmi robustní, neboť nemají žádné předpoklady o rozdělení v populaci, a dokonce ani nevyžadují náhodný výběr. Tyto testy jsou postaveny na randomizaci získaných hodnot a jako takové mohou pracovat i s velmi malými vzorky. Jako učebnicový příklad jsme rozebírali např. Fisherův exaktní test (díl 14 seriálu).

V tomto díle se více zaměříme na další dva typy neparametrických testů, které se velmi často využívají právě v přírodních vědách jako alternativa *t*-testu. Postupy výpočtu všech zmíněných testů jsou demonstrovány na příkladech 1–5. Nikoli náhodou jsme příklady výpočtů připravili na experimentech, v nichž se hodnotí počty buněk. Právě počty („counts“), ať již buněk, nebo jevů, impulzů či událostí, velmi často vykazují nestandardní rozdělení četností a neparametrické testy se na ně velmi dobře aplikují.

- Prvním typem jsou testy, které pracují s pořadím hodnot neboli s ordinálními škálami. Ordinální data mohou být přímo výstupem měření nebo mohou být na pořadí převedena data kvantitativní, spojitá. Tím, že je převedeme na pořadí, děláme ovšem jakýsi krok zpět a hodnoty již nebudou vystupovat jako kvantitativní míra. Něco ztrácíme (informaci o kvantitě) a něco získáváme (svobodu od předpokladů testu). Převádíme-li takto např. řadu deseti hodnot, pak nejvyšší číslo bude mít vždy pořadí 10 a bude jedno, zda je za ním hodnota ve stovkách nebo v milionech. Takto je zcela odstraněn vliv odlehlých hodnot. Tyto, někdy také nazývané pořadové testy („rank tests“), reprezentuje Mann-Whitney U test a Wilcoxonův test (příklady 1 a 2).
- Druhý typ neparametrických testů vede k postupům, které pracují pouze s odchylkami od určité hodnoty a těmto přiřazují znaménko + nebo –, podle směru. Dále pracují s četnostmi odchylek, např. sledují, zda jsou kladné odchylky stejně četné jako záporné apod. Tento typ testu reprezentuje tzv. znaménkový test anebo mediánový test (příklady 3 a 4).

Čtenář se nyní může zeptat, jak je to tedy s onou slibovanou alternativou *t*-testu. Již z uvedených příkladů vyplývá odpověď. Mann-Whitney U test a mediánový test jsou alternativou *t*-testu pro dva nezávislé výběry; znaménkový test („sign test“) a Wilcoxonův test pro dva závislé výběry zastoupí párový *t*-test. Jelikož jsou tyto testy opravdu často využívány, věříme, že uvedené příklady nejsou zbytečné. Každý z testů má svůj specifický postup výpočtu, který určuje jeho využitelnost pro různé experimentální situace. Vlastní výpočet dnes samozřejmě provede statistický software, od uživatele se ale vyžaduje vědomá volba konkrétního testu.

Jestliže jde na stejná data použít dva rozdílné testy nebo i více testů, jistě to svádí k pokusům. Zvláště, když to ve věku výkonných osobních počítačů nestojí mnoho námahy. Uživatel ale nesmí být překvapen, když mu různé testy nabídnou poněkud různé výstupy. Často se může stát, že na stejných datech parametrický test povede k zamítnutí nulové hypotézy, ale neparametrický test ji potvrdí. Neparametrické testy mají totiž vždy o něco menší sílu než příslušné testy parametrické, hovoříme tedy o jejich nižší schopnosti rozpoznat neplatnou nulovou hypotézu. Tato skutečnost znamená, že pro prokázání statistické významnosti stejného rozdílu vyžadují větší velikost vzorku. Nižší síle neparametrických testů je nutné přizpůsobit plánování experimentů, které bude náplní některého z dalších dílů našeho seriálu. Zde pouze konstatujme, že u řady neparametrických testů nejde o velkou ztrátu a lze ji snadno kompenzovat zvýšením velikosti vzorku o 10–15 %.

S pojmem neparametrické testování je často spojován fakt, že netestujeme žádnou hypotézu o parametru nějakého modelového pravděpodobnostního rozdělení. Obecně tomu tak jistě je, nicméně při splnění určitých předpokladů se neparametrické testy používají k odhadům parametrů rozdělení nebo dokonce k hledání intervalů spolehlivosti těchto odhadů. Řada neparametrických testů také testuje hypotézy související s hodnotou mediánu. Jako typickou ukázku jsme zde zařadili aplikaci znaménkového testu pro hodnotu mediánu (příklad 5).

Pevně věříme, že po přečtení tohoto dílu se čtenáři nebudou obávat opustit *t*-test, zvláště při problémech se splněním jeho předpokladů. Znalost podstaty neparametrických testů umožní správný výběr pro správná data. Pouze doporučujeme jistou konzistentnost při psaní publikací. Jedna práce by měla používat na stejných datech buď parametrické, nebo neparametrické testy; jejich různé náhodné kombinace nesvědčí o promyšlené strategii a odpovědném plánování experimentu.

Na závěr bohužel vneseme do označení testů trochu zmatku. Již v legendě k příkladu 1 je uvedeno, že pro Wilcoxonův test pro dva nezávislé výběry existuje ekvivalentní Mann-Whitney U test. Významný matematik Frank Wilcoxon (1892–1965) je autorem dvou neparametrických testů, pro párové uspořádání (Wilcoxon Signed-Rank Test; Wilcoxon Matched-Pairs Ranks test) i pro dva nezávislé výběry (Wilcoxon Rank-Sum Test). Oba testy popsal v jediné práci z roku 1945. Mann-Whitney U test je modifikace výpočtu vzniklá dva roky poté v roce 1947 a je výsledkovým ekvivalentem Wilcoxonova testu pro dva nezávislé výběry. Zmatku často nelze zabránit. Zvláště nebezpečné je, že nepoučenému uživateli mohou splývat dvě varianty Wilcoxonova testu, a tudíž by si mohl splést test pro dva závislé a dva nezávislé výběry. Proto je především u nového software dobré se přesvědčit, jaký test se pod názvem skutečně skrývá.

Tímto závěrem jsme zabrousili hluboko do historie, která je ale, jak vidno, stále vlivná. Přes všechny komplikace je právě toto na výpočetních vědách krásné, tedy že nestárnou tak rychle jako my ☺.