Testy nieparametryczne dla prób zależnych

Do testów nieparametrycznych dla prób zależnych należą:

- test znaków
- test kolejności par Wilcoxona
- test McNemary

Testy te stanowią nieparametryczną alternatywę testu t-Studenta dla zmiennych powiązanych, a stosujemy je, gdy dysponujemy dwoma pomiarami (przed jakimś wydarzeniem i po) i chcemy dowieść, że pomiary te się różnią. Inaczej mówiąc, testy te są przeznaczone do sprawdzania istotności różnic między dwoma zależnymi pomiarami.
Te dwa zależne pomiary to albo dwie obserwacje u tej samej osoby (np. przed zabiegiem i po), albo obserwacje u par osób o tych samych właściwościach (tzw. równoważne dwójki).
Hipoteza zerowa mówi, że wyniki obu pomiarów są jednakowe. Testy te stosujemy również wtedy, gdy nie są spełnione założenia testu t dla zmiennych powiązanych. Za ich pomocą możemy stwierdzić, czy próby różnią się między sobą pod względem pewnych własności. Te łatwe w użyciu testy wymagają jedynie założenia, że wartości badanych zmiennych możemy uporządkować (są mierzalne na skali porządkowej)

Dla sprawdzenia, czy dwa pomiary różnią się między sobą stosujemy test znaków lub test Wilcoxona. Oba te testy dotyczą zmiennych zależnych, najczęściej są to pomiary pochodzące od tych samych osób.

Hipoteza zerowa mówi, że wyniki obu próbek są jednakowe.

Test znaków oparty jest na znakach różnic pomiędzy parami wyników. Liczba plusów i minusów jest zliczana i porównywana z wartością teoretyczną umieszczoną w odpowiednich tabelach. Tracimy informację niesioną przez liczbowe wartości różnic.

Test kolejności par Wilcoxona uwzględnia zarówno znak różnic, ich wielkość, jak i kolejność. Po uporządkowaniu różnic w sposób rosnący są im przypisywane rangi, a następnie sumowane osobno rangi różnic dodatnich i ujemnych. Ich suma po porównaniu z tabelą wartości teoretycznych decyduje o przyjęciu lub nie hipotezy zerowej.

Test Znaków

Test ten oparty jest na znakach różnic między kolejnymi parami wyników (badaniu czy są one ujemne, czy dodatnie). W ten sposób otrzymujemy ustaloną liczbę wyników z jednego zbioru, które są mniejsze od swych odpowiedników w drugim zbiorze oraz liczbę wyników z tego samego zbioru, które są większe od swych odpowiedników w drugim zbiorze. Tym samym dowiadujemy się, ile danych zostało niejako "przesuniętych" w jednym z kierunków w naszym eksperymencie.
Naturalnie gdy zaobserwujemy, iż wszystkie wyniki zostały przesunięte, to łatwo stwierdzić, że test daje wynik istotny. Jednakże w przeciwnym przypadku sytuacja nie jest już tak oczywista.
Dlatego też skupiamy się na wyznaczeniu prawdopodobieństwa związanego ze wszystkimi proporcjami wystąpienia znaków +/-, które mogłyby wystąpić. W konsekwencji znając prawdopodobieństwo każdej kierunkowej zmiany możemy ocenić czy nasze wyniki są istotne czy też nie.
Innymi słowy test znaków polega na ustaleniu liczby plusów oraz minusów, następnie porównanie ich z teoretyczną wartością - dostępną w odpowiednich tablicach. Test ten stosujemy więc przede wszystkim dla cech jakościowych. Wystarczy bowiem sprawdzić, że dana jednostka charakteryzuje się obecnością ("+") lub nieobecnością ("-") danego zjawiska. Dla danych mierzalnych nie uwzględniamy wartości różnic, lecz jedynie ich znaki. Różnice o wartości zero są pomijane.

Założenia

Pary próbek skorelowanych (xi,yi) są losowe i niezależne, tj. różnice zi=yi-xi są niezależne oraz pochodzą z tego samego rozkładu, o medianie θ.
Przestrzeń wartości powinna mieć przynajmniej zdefiniowany porządek (ordinal scale), a najlepiej gdyby była ciągła.

Hipoteza

H0: $\theta = 0$,
H1: $\theta \neq 0$(dwustronny)
H1': $\theta > 0$ lub $\theta < 0$ (jednostronny)

Test

Dwie populacje generalne o ciągłych rozkładach i dystrybuantach F1(x) i F2(x), z których wylosowano n parami odpowiadających sobie elementów.
Weryfikacja hipotezy H0 testem znaków przebiega następująco:

1. Badamy znak różnicy par wyników w obu próbach i znajdujemy liczbę tych znaków, których jest mniej (jeśli są w próbie pary o identycznych wartościach, to nie rozważamy ich w teście), tzn. r = min(r-,r+), gdzie r- i r+ oznaczają odpowiednio liczbę znaków ujemnych i dodatnich różnic rozważanych par wyników,
2. Z tablic rozkładu liczby znaków odczytujemy dla liczby par wyników n oraz przyjętego poziomu istotności taką wartość krytyczną r, że $P(r \leq r_\alpha) = \alpha$,
3. Obszar odrzucenia ma postać $W = \langle 0; r_\alpha \rangle$,

4 Jeżeli $r \in W$, to odrzucamy hipotezę H0 na rzecz hipotezy alternatywnej, w przeciwnym przypadku tzn. gdy $r \notin W$ brak podstaw do odrzucenia hipotezy, że obie próby pochodzą z jednej populacji.

Przykład w Matlabie

Zaczniemy od testu hipotezy zerowej dotyczącej mediany, tj. testując czy jest różna od zera. Generujemy przykładowe dane wykorzystując polecenie rng('default') (ustalenie losowych generatorów na tryb default) i wylosowanie próbki x=randn(1,25), czyli symetrycznego rozkładu z zerową medianą.
Następnie testujemy hipotezę, iż dane pochodzą z rozkładu o medianie różnej od zera:

```
[p,h,stats] = signtest(x,0)
```

otrzymujemy:

p = 0.1078, h = 0, oraz stats = zval: NaN; sign: 17

Zatem przy standardowym poziomie istotnośći (5%) rezultat h=0 oznacza, że test nie pozwala na odrzucenie zerowej hipotezy.

Jako drugi przykład sprawdźmy hipotezę, czy x-y ma zerową medianę. Rozpoczniemy od wygenerowania dwóch próbek danych

```
rng('default') % for reproducibility
x = lognrnd(2,.25,15,1);
y = x + trnd(2,15,1);
```

następnie testujemy hipotezę

```
[p,h,stats] = signtest(x,y)
```

p = 0.3018, h = 0, stats = zval: NaN; sign: 5

Zatem przy standardowym poziomie istotności (5%) rezultat h=0 oznacza, że test nie pozwala na odrzucenie zerowej hipotezy o zerowej wartości różnicy median.

UWAGA

Należy podkreślić, iż w teście znaków tracimy informację niesioną przez liczbowe wartości różnic. Ta informacja jest w pełni wykorzystywana przez test Wilcoxona. Staje się on więc w tym wypadku testem mocniejszym niż test znaków.

Należy także wspomnieć, iż test ten nie jest sugerowany dla małych zbiorów danych (n<6).

Ćwiczenie 1 Załóżmy, że przeprowadziliśmy badanie ciężaru ciała w grupie 20 kobiet przed 7-tygodniową dietą odchudzającą i po niej. Otrzymane dane przedstawiono:

w 1	88	69	86	59	57	82	94	93	64	91	86	59	91	60	57	92	70	88	70	85
w 2	73	68	75	54	53	84	84	86	66	84	78	58	91	57	59	88	71	84	64	85

Chcemy sprawdzić, czy otrzymane wyniki przeczą hipotezie, że dieta nie powoduje zmniejszenia ciężaru ciała. Na początku przeanalizuj dostępne dane za pomocą znanych Ci metod graficznych. Ponieważ nasze dane to dwie obserwacje u tej samej osoby, więc do weryfikacji postawionej hipotezy wykorzystamy testy nieparametryczne dla zmiennych powiązanych.

Ćwiczenie 2. Wczytaj plik czytelnictwo.csv. Dane w nim zawarte przedstawiają ilość czasu poświęcanego na codzienną lekturę prasy przed podjęciem pracy w firmie i po podjęciu tej pracy. Zbadaj, czy zatrudnienie w firmie miało wpływ na ilość czasu poświęcanego na lekturę prasy.

Wilcoxon rank-sum

Test kolejności par Wilcoxona uwzględnia znak różnic, ich wielkość, jak również ich kolejność (stąd nazwa). Po uporządkowaniu różnic w szereg rosnący przypisujemy im rangi. Następnie osobno sumujemy rangi różnic dodatnich i ujemnych. Mniejsza z otrzymanych sum to wartość testu Wilcoxona, która po porównaniu z odpowiednią wartością teoretyczną w tablicach decyduje o odrzuceniu hipotezy zerowej lub nie.

Ten nieparametryczny test można wykorzystać zarówno do sparowanych, jak i nie sparowanych zbiorów danych, i pozwala on na weryfikację czy dwie próby pochodzą z identycznego ciągłego rozkładu prawdopodobieństwa o takich samych medianach, względem alternatywy, iż rozkłady te nie mają takich samych median.

Założenia

Pary próbek skorelowanych (xi,yi) są losowe i niezależne, tj. różnice zi=yi-xi są niezależne oraz pochodzą z tego samego rozkładu - symetrycznego względem wspólnej mediany θ.
Przestrzeń wartości powinna mieć przynajmniej zdefiniowany porządek (ordinal scale), a najlepiej gdyby była ciągła.

Hipoteza

H0: $\theta = 0$,
H1: $\theta \neq 0$(dwustronny)
H1': $\theta > 0$ lub $\theta < 0$ (jednostronny)

Test

Algorytm wyliczania statystyki testu Wilcoxona rozpoczynamy od wyznaczenia różnic di pomiędzy kolejnymi parami pomiarów, tj. di = yi - xi, wyłączają wszystkie zera, tj. di=0. Następnie porządkujemy wartości bezwzględne różnic |d1| … |dn| od najmniejszej do największej i nadajemy rangi dla tak wyznaczonego zbioru. Rangi zaczynamy od R1=1, dla najmniejszego elementu, i nadajemy kolejno Ri - zaś wszystkie powtarzające się wartości

otrzymują rangę równą średniej arytmetycznej rang na jakich są rozpięte. Przykładowo dla d = [1,2,2,3] otrzymujemy rangi R1=1, R2=2.5, R3=2.5 oraz R4=4.
W dalszym kroku obliczamy statystykę W

$$W = |\sum_{i=1}^{m}[\mathrm{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]|$$

Ponieważ wraz ze wzrostem ilości próbek (par) o niezerowych różnicach pomiędzy próbkami (elementami pary) - Nr - rozkład statystyki W dąży do rozkładu normalnego, stąd możemy dla Nr > 10 wyznaczyć

$$z = \frac{W - 0.5}{\sigma_W}, \sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$$

wówczas jeżeli z > zcritical to możemy odrzucić H0, w przeciwnym przypadku gdy Nr < 10 należy porównać wartość W z tablicami i jeżeli W > Wcritical to możemy odrzucić H0.

Przykład w Matlabie

Zaczniemy od testu hipotezy zerowej dotyczącej mediany, tj. testując czy jest dla dwóch niezależny i nierówno licznych grup jest ona różna. Generujemy przykładowe dane wykorzystując polecenie rng('default') (ustalenie losowych generatorów na tryb default) i wylosowanie próbki x=unifrnd(0,1,10,1) oraz y=unifrnd(0.25,1.25,15,1), czyli różniących się o przesunięcie o 0.25. Następnie testujemy równość median dla x oraz y:

```
p = ranksum(x,y)
```

otrzymujemy

```
p = 0.0375
```

zatem wnioskujemy, iż test odrzuca zerową hipotezę o równości median przy 5% poziomie ufności.

UWAGA
Test t-Studenta sprawdza hipotezę zerową o równości średnich arytmetycznych w odpowiadających im populacjach, test Wilcoxona weryfikuje równość median.
Test Wilcoxona bazuje na różnicach pomiędzy wartościami cech z porównywanych zbiorów, jednakże w przeciwieństwie do testu t-Studenta, nie posiada założeń dotyczących rozkładu próby. Może zatem być używany w sytuacjach, gdy założenia testu t-Studenta nie są spełnione.

Ćwiczenie 3. Wybrano losowo kilka roślin chmielu i zapylono połowę każdej z nich. Otrzymane nasiona (masa nasion w g na 10 g chmielu) zawiera plik chmiel.csv. Na poziomie istotności 0,05 zweryfikuj hipotezę, że zapylenie ma wpływ na masę nasion.

Ćwiczenie 4. Sprawdź, czy średni czas poświęcany na czytanie prasy przez pracowników naszej firmy (czytelnictwo.csv) zmienił się po przyjęciu ich do pracy.

Ćwiczenie 5. Wczytaj plik Dane z koronografii.csv. Przy poziomie ufności 0,9 sprawdź, czy czas ćwiczenia zależy od stanu zdrowia.

UWAGA

Należy pamiętać, iż powyższe testy można także wykorzystać do weryfikacji dwu- lecz także jedno-stronnej hipotezy. To znaczy, że możemy również zastanawiać się na alternatywnymi hipotezami postaci H1: $\theta > 0$ lub $\theta < 0$ (gdzie θ oznacza medianę).