Testy nieparametryczne

Wybrane testy nieparametryczne

1. Test chi-kwadrat zgodności z rozkładem oczekiwanym
2. Test chi-kwadrat niezależności dwóch zmiennych kategoryzujących
3. Test U Manna-Whitney'a
4. Test Kruskala-Wallisa

Test chi-kwadrat zgodności z rozkładem oczekiwanym

- Eksperyment składa się z n identycznych i niezależnych prób
- Każda z prób kończy się jednym z k zdarzeń
- Prawdopodobieństwo, że pojedyncze zdarzenie kończy się wynikiem i wynosi π_i, $i=1,\ldots,k$
- Interesuje nas, czy obserwujemy rozkład zgodny z pewnym oczekiwanym rozkładem π_i, $i=1,\ldots,k$

Wieloletnia standardowa terapia nadciśnienia daje następujące wyniki w 4 kategoriach

1. Znaczące obniżenie ciśnienia	50% pacjentów
2. Umiarkowane obniżenie ciśnienia	25%
3. Nieznaczne obniżenie ciśnienia	10%
4. Bez zmian lub podwyższenie ciśnienia	15%

Przygotowano nową terapię i przeprowadzono próbę kliniczną na 200 pacjentach. Wyniki:

1. Znaczące obniżenie ciśnienia	120 pacjentów
2. Umiarkowane obniżenie ciśnienia	60
3. Nieznaczne obniżenie ciśnienia	10
4. Bez zmian lub podwyższenie ciśnienia	10

Czy wyniki próby klinicznej wskazują na wyższość nowej terapii?

H0: $\pi_1 = 0.50$, $\pi_2 = 0.25$, $\pi_3 = 0.10$, $\pi_4 = 0.15$

HA: Przynajmniej jedno z prawdopodobieństw jest inne od wartości hipotetycznej

Karl Pearson 1900:

$$X^2 = \Sigma \left[(n_i - E_i)^2 / E_i \right]$$

n – obserwowana liczba obserwacji
E – oczekiwana liczba obserwacji
$i = 1, \ldots, k$

Odrzuć H0, jeżeli X^2 przekracza tabelaryczną wartość krytyczną dla poziomu istotności α oraz liczby stopni swobody $k-1$.

Pamiętaj, test chi-kwadrat jest przybliżony (ale b. praktyczny)

Wymagania wg. Cochrana: dla każdej kategorii liczba oczekiwana E wynosi co najmniej 1, dla nie więcej niż 20% kategorii E<5.

Test chi-kwadrat niezależności dwóch zmiennych kategoryzujących

Podejrzewa się, że pewien wariant genu DIO2 zwiększa ryzyko nadciśnienia. Zbadano polimorfizm A/T i obserwowano 200 osób. Czy dane potwierdzają takie przypuszczenia.

Tablica kontyngencji

Genotyp	wysokie	podwyższone	normalne	Razem
AA	30	15	15	**60**
AT	40	10	50	**100**
TT	10	5	25	**40**
Razem	**80**	**30**	**90**	**200**

Jeżeli zmienne są niezależne, to w populacji spodziewamy się 12% osób z genotypem AA i wysokim nadciśnieniem:

P(AA) = 60 / 200 = 0.3

P(wysokie) = 80 / 200 = 0.4

P(AA i wysokie) = 0.3 * 0.4 = 0.12 (niezależność zdarzeń)

Wśród 200 przebadanych osób powinno ich być ok. 24.

H0: Zmienne są niezależne
HA: Zmienne są zależne

$$X^2 = \Sigma \left[(n_{ij} - E_{ij})^2 / E_{ij} \right]$$

$i = 1,\ldots,k$
$j = 1,\ldots,w$

Odrzuć H0, jeżeli X^2 przekracza tabelaryczną wartość krytyczną dla poziomu istotności α oraz liczby stopni swobody ($(k-1)(w-1)$).

Wymagania Cochrana są ważne.

Ale stwierdzono, że jeżeli E w każdej komórce jest podobnej wielkości, a liczba kolumn i wierszy duża, to E_{ij} może być tylko 1 i jest OK! (duża elastyczność testu)

W innych przypadkach możesz stosować dokładny test Fishera (Fisher's exact test).

R

```
# wpisujemy dane sumaryczne
mytable = rbind(  c(30,15,15), c(40,10,50),  c(10,5,25)  )

# lub wczytujemy z pliku z surowymi obserwacjami
mydata <-read.table( )
attach( mydata )
# i uzyskujemy tabele kontyngencji
mytable <- table( A, B )
# A will be rows, B will be columns

# testujemy
chisq.test( mytable )
```

Test U Manna-Whitney’a

Mann–Whitney *U* test, Mann–Whitney–Wilcoxon, Wilcoxon rank-sum test

- Czy niezależne próby z jednej z dwóch populacji są wyższe niż z drugiej?
- Odpowiednik testu t (porównanie średnich)
- Często stosowany, gdy
  - rozkład w populacji znacznie odbiega od normalnego.
  - skala jest porządkowa, ale nie przedziałowa (większe lub mniejsze, ale nie wiadomo o ile, np. dst, dst+)
- Moc
  - 95% testu t, jeżeli rozkład normalny (stosuj test t)
  - Jeżeli rozkład znacznie odbiega od normalnego, moc dużo większa od testu t (duża skośność, obserwacje odstające)

Sprawdzano skuteczność zabiegu oczyszczania jeziora przy fermie. Badano ilość rozpuszczonego tlenu (w ppm (ang. parts per million))

12 prób przed oczyszczaniem

11.0	11.2	11.2	11.2	11.4	11.5
11.6	11.7	11.8	11.9	11.9	12.1

12 prób po oczyszczeniu

10.2	10.3	10.4	10.6	10.6	10.7
10.8	10.8	10.9	11.1	11.1	11.3

H0: Rozkład pomiarów przed i po oczyszczaniu jest taki sam

HA: Pomiary po oczyszczeniu mają tendencję do niższych wartości niż przed oczyszczeniem (zabiegi skuteczne)

Wartości rank

12 prób przed oczyszczeniem

11.0	11.2	11.2	11.2	11.4	11.5
10	**14**	**14**	**14**	**17**	**18**
11.6	11.7	11.8	11.9	11.9	12.1
19	**20**	**21**	**22.5**	**22.5**	**24**

12 prób po oczyszczeniu

10.2	10.3	10.4	10.6	10.6	10.7
1	**2**	**3**	**4.5**	**4.5**	**6**
10.8	10.8	10.9	11.1	11.1	11.3
7.5	**7.5**	**9**	**11.5**	**11.5**	**16**

Jeżeli oczyszczanie jest skuteczne, suma wartości rank dla próby 1 (T) powinna być wyższa.

rank	**grupa**	**t_j**
1	1	1
2	2	1
3	3	1
4.5 4.5	4	2
6	5	1
7.5 7.5	6	2
9	7	1
10	8	1
11.5 11.5	9	2
14 14 14	10	3
16	11	1
17	12	1
18	13	1
19	14	1
20	15	1
21	16	1
22.5 22.5	17	2
24	18	1

Suma rank (T) ma rozkład normalny z wartością oczekiwaną i wariancją:

$$\pi = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma^2 = \frac{n_1 n_2}{12}\left[(n_1 + n_2 + 1) - \frac{\sum t_j\left(t_j t_j - 1\right)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right]$$

Wymagania: n_1 i n_2 co najmniej po 10 obserwacji

Statystyka testowa:

$$z = \frac{T - \pi}{\sigma}$$

H0: Dwie populacje są identyczne

HA:
1. Populacja 1 ma przeciętnie wyższe wartości
2. Populacja 1 ma przeciętnie niższe wartości
3. Populacja 1 ma przeciętnie inne wartości od populacji 2

Dla przyjętego poziomu istotności odrzuć H0, jeżeli
1. $z > z_\alpha$
2. $z < -z_\alpha$
3. $|z| > z_{\alpha/2}$

R

```
# independent 2-group Mann-Whitney U Test
wilcox.test( y ~ A )
# where y is numeric and A is A binary factor


# independent 2-group Mann-Whitney U Test
wilcox.test(  y, x )
# where y and x are numeric


# dependent 2-group Wilcoxon Signed Rank Test
wilcox.test( y1, y2, paired=TRUE )
# where y1 and y2 are numeric
```

Zadanie	Testy nieparametryczne	Testy parametryczne
Porównanie 2 populacji	Test U	Test t
Porównanie 2 populacji-obserwacje sparowane	Wilcoxon’s signed-rank test	Test t dla sparowanych obserwacji
Porównanie wielu populacji	Test Kruskala-Wallisa;	Analiza wariancji (ANOVA)

Test Kruskala-Wallisa

H0: Rozkłady dla *k* populacji są takie same
HA: Nie wszystkie rozkłady są takie same

$$H = \frac{12}{n(n+1)} \sum \frac{T_i T_i}{n_i} - 3(n+1)$$

Odrzuć H0, jeżeli H przekracza wartość krytyczną chi-kwadrat dla ustalonego poziomu istotności i liczby stopni swobody *k*-1

Trzy grupy studentów weterynarii ankietowano pod kątem wiedzy o chorobach zakaźnych. Czy miejsce studiów ma znaczenie?

Poznań	Warszawa	Wrocław
32	32	28
30	32	21
30	26	15
29	26	15
26	22	14
23	20	14
20	19	14
19	16	11
18	14	9
12	14	8

R

```
# Kruskal Wallis Test One Way Anova by Ranks
kruskal.test( y ~ A )
# where y1 is numeric and A is a factor
```