Iwona Mróz,

Instytut Fizyki Doświadczalnej,

Uniwersytet Wrocławski

**„Podstawy statystyki i analizy danych” – materiały do wykładu nr 13 z dnia 18.01.2021 r – część 2**

Niniejsze materiały mają charakter roboczy. Bardzo proszę o zgłaszanie zauważonych błędów, braków, niedociągnięć i niejasności. Prośba dotyczy też przypisów. Z góry dziękuję za pomoc☺.

**Analiza danych jakościowych - testy niezależności i test McNemara**

***Opracowano na podstawie podręcznika Andrzeja Stanisza „Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny”, StatSoft Polska Sp. z o.o., Kraków 2006, tom 1, str. 321-336.***

Analiza danych jakościowych, wyrażonych w skali nominalnej lub porządkowej, jest bardzo ważnym działem analizy danych. Za danymi jakościowymi mamy do czynienia bardzo często, np. w badaniach społecznych, marketingowych czy przyrodniczych. Jednym z najczęściej zadawanych pytań związanych z danymi jakościowymi jest pytanie o niezależność cech statystycznych. Przykładowo, pytamy czy występowanie pewnych chorób zależy od stylu życia, czy preferencje kulinarne wiążą się ze sposobami spędzania wolnego czasu itp. Dane jakościowe pozyskuje się często przy pomocy ankiet, które trzeba odpowiednio przygotować, co wymaga dużej wiedzy oraz wiąże się z nakładem czasu i środków finansowych.

**1. Test niezależności χ2 Persona (*A. Stanisz, str. 321-323*)**

Najbardziej popularnym testem do badania niezależności cech jakościowych X i Y o skali nominalnej lub porządkowej jest test χ2 Persona. Test „bada” tabelę o k wierszach (reprezentujących k wartości cechy X) i p kolumnach (reprezentujących p wartości cechy Y).

Przykładowo, w poniższej tabeli k = 4, p = 3. nij oznaczają liczbę obserwacji o wskazanych wartościach cech X i Y.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y1 | Y2 | Y3 |  |
| X1 | n11 | n12 | n13 |  |
| X2 | n21 | n22 | n23 |  |
| X3 | n31 | n32 | n33 |  |
| X4 | n41 | n42 | n43 |  |
|  |  |  |  |  |

Test porównuje obserwowane („observed”) wartości nij z tzw. liczebnościami (częstościami) oczekiwanymi, które wystąpiłyby w tabeli gdyby cechy X i Y były niezależne.

Aby dla komórki ij tabeli obliczyć jej częstość oczekiwaną Eij (od „expected”) korzystamy ze wzoru:



Innymi słowy, dla danej komórki jej Eij = (suma i-tego wiersza)(suma j-tej kolumny)/liczba wszystkich obserwacji.

Hipotezy testu χ2:

H0: Cechy X i Y są niezależne.

H1: Cechy X i Y są zależne.

Statystyka testu χ2:



O to obserwowana wartość, a E to częstość oczekiwana komórki.

Przy założeniu prawdziwości hipotezy zerowej statystyka ta ma asymptotyczny rozkład χ2 o df = (k-1)(p-1).

**2. Poprawki do testu χ2 (*A. Stanisz, str. 335*)**

Andrzej Stanisz podkreśla, że wartość testu χ2 zależy od liczby obserwacji i liczby komórek w tabeli. W szczególności wskazuje, że stosowanie testu nie jest zalecane gdy w tabeli pojawi się częstość oczekiwana mniejsza od 5. Ponadto, test jest bardzo czuły na małe różnice pomiędzy wartościami obserwowanymi a częstościami oczekiwanymi jeżeli liczebność próbki jest duża. Ze względu na te ograniczenia testu χ2, w niektórych przypadkach lepiej jest stosować poprawki lub tzw. dokładny test Fishera. Poprawki są opracowane dla tabel 2×2!

Zasady są takie:

Niech N oznacza całkowita liczebność próby.

1. Test χ2 (bez poprawek) stosujemy gdy N>40 i wszystkie częstości oczekiwane są większe od 10.

2. Test V-kwadrat stosujemy gdy N>40 i którakolwiek z częstości oczekiwanych jest mniejsza od 10.

3. Test χ2 z poprawką Yatesa stosujemy gdy:

- N>40 i którakolwiek z częstości oczekiwanych jest mniejsza od 5

- 20<N≤40 i wszystkie częstości oczekiwane są większe od 5.

4. Dokładny test Fishera stosujemy gdy:

- 20<N≤40 i którakolwiek z częstości oczekiwanych jest mniejsza od 5

- N≤20.

**3. Test McNemara (wariant A/D) (*A. Stanisz, str. 327 i 336*)**

Test McNemara stosowany jest gdy mamy do czynienia z danymi jakościowymi pochodzącymi z grup zależnych. Występuje on w dwóch wariantach: A/D i B/C. Omówimy szczegółowo wariant A/D.

Test McNemara jest nazywany testem istotności zmian. Mianowicie, „…służy on do określania istotności różnic w wynikach, które zaszły pod wpływem jakiegoś oddziaływania.” (A. Stanisz, str. 327).

Przeprowadzając test McNemara konstruujemy tabelę 2×2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Po oddziaływaniu | | Suma |
| - | + |
| Przed oddziaływaniem | + | A | B | A+B |
| - | C | D | C+D |
| Suma | | A+C | B+D | N |

Gdzie:

A – liczba obserwacji, dla których w wyniku oddziaływania doszło do zmiany z + na -,

D – liczba obserwacji, dla których w wyniku oddziaływania doszło do zmiany z - na +,

B i C – liczby obserwacji, dla których nie doszło do zmiany w wyniku oddziaływania.

N – całkowita liczebność próby.

Hipotezy testu McNemara:

H0: Pod wpływem oddziaływania nie doszło do zmiany liczebności obserwacji z cechą +

H1: Pod wpływem oddziaływania doszło do zmiany liczebności obserwacji z cechą +.

Innymi słowy, hipoteza zerowa mówi, że liczebności w komórkach A i D są takie same.

Statystyka testu wyraża się wzorem (w źródłach można znaleźć też postaci bez jedynki w liczniku):



Przy założeniu prawdziwości hipotezy zerowej statystyka ta ma w przybliżeniu rozkład χ2

z df = 1.