Test niezależności chi-kwadrat i jego zastosowanie w interpretacji wyników badań klinicznych

The chi-square independence test and its application in the clinical researches

Przemysław J. Kwasiborski[1,2], Maria Sobol[1]

[1]Zakład Biofizyki i Fizjologii Człowieka Warszawskiego Uniwersytetu Medycznego
[2]Klinika Kardiologii Wojskowego Instytutu Medycznego Centralnego Szpitala Klinicznego Ministerstwa Obrony Narodowej w Warszawie

Kardiochirurgia i Torakochirurgia Polska 2011; 4: 550–554

Streszczenie

Praca ma na celu krótkie i praktyczne omówienie zasad stosowania testu χ^2, bez którego nie może obyć się analiza statystyczna ogromnej większości publikowanych prac medycznych. W badaniach klinicznych bardzo często ma się do czynienia z cechami niemierzalnymi, jakościowymi – chorzy klasyfikowani są pod względem różnych kategorii. Typowymi zmiennymi tego rodzaju są: płeć, występowanie określonej choroby czy zaszeregowanie chorych w różnych skalach – skali zaawansowania dławicy piersiowej opracowanej przez Kanadyjskie Towarzystwo Kardiologiczne (ang. *Canadian Cardiovascular Society* – CCS), skali służącej do klasyfikacji ciężkości objawów niewydolności krążenia zaproponowanej przez Nowojorskie Towarzystwo Kardiologiczne (ang. *New York Heart Association* – NYHA) czy klasyfikacji Killipa-Kimballa opracowanej do oceny niewydolności serca w świeżym zawale mięśnia sercowego. Często interesujący jest wpływ jednej cechy jakościowej na drugą cechę tego samego typu, np. wpływu płci na częstość występowania danej choroby lub na rozkład punktacji CCS/NYHA w populacji chorych. W takim przypadku należy posłużyć się odpowiednim aparatem obliczeniowym i testem statystycznym. Najczęściej stosowanym testem w takiej sytuacji jest opracowany w 1900 r. przez Karla Pearsona test niezależności χ^2. Służy on do weryfikacji hipotezy zerowej (H_0) – o braku różnic istotnych statystycznie między rozpatrywanymi cechami. Obszar stosowalności tego testu obejmuje nie tylko analizę częstości występowania cech jakościowych, ale również analizę zgodności cech ilościowych oraz przypadki badania zależności cechy jakościowej od ilościowej.

Analizowanie danych z wykorzystaniem testu χ^2 nie sprawia większych problemów – pod warunkiem, że badacz posiadł podstawową wiedzę z zakresu statystyki. Nawet nie będąc ekspertem w tej dziedzinie, lekarz prowadzący badania naukowe może stać się świadomym użytkownikiem pakietu staty-

Abstract

The chi-square test is one of most commonly used statistical tests in medical science. This test is used to verify the null hypothesis (H_0) which assumes that there is no significant difference between expected and observed data. The null hypothesis is rejected if chi-square value p is less than 5%. A comprehensive review of a chi-square test application and its limitations are presented.

Key words: chi-square test, Fisher exact test, Yates correction, McNemar test, Cochran-Armitage test.

Adres do korespondencji: dr Maria Sobol, Zakład Biofizyki i Fizjologii Człowieka Warszawskiego Uniwersytetu Medycznego, ul. Chałubińskiego 5, 02-004 Warszawa, e-mail: maria.sobol@wum.edu.pl

stycznego, co z pewnością ułatwi planowanie dalszych badań i analizę otrzymywanych wyników.
Słowa kluczowe: test χ^2, dokładny test Fishera, poprawka Yatesa, test McNemary, test Cochrana-Armitage.

Podstawy teoretyczne

Rewolucję w statystyce medycznej wywołało powszechne zastosowanie szybkich komputerów. Wraz z rozpowszechnieniem coraz bardziej rozbudowanych programów wspomagających obliczenia statystyczne analiza danych stała się możliwa do wykonania bezpośrednio przez lekarza prowadzącego badania naukowe. Łatwość, z jaką obecnie można otrzymać wynik, nawet najbardziej wyszukanego testu statystycznego, nie zwalnia od zapoznania się choćby z podstawowymi zasadami stosowania takiego testu.

Statystyka medyczna jest nauką żywą, która wciąż ewoluuje, jednak pewien kanon metod pozostaje niezmienny. Dlatego każdy, kto samodzielnie analizuje wyniki eksperymentów lub chce tylko czytać ze zrozumieniem współczesne prace naukowe, powinien poznać powyższy kanon.

Praca ma na celu krótkie i praktyczne omówienie zasad stosowania testu χ^2, bez którego nie może obyć się analiza statystyczna ogromnej większości publikowanych prac medycznych. W badaniach klinicznych często ma się do czynienia z różnymi typami zmiennych lub inaczej cech (tak określa się informację jednostkową w języku analizy statystycznej). Najogólniej wyróżnia się zmienne (cechy) jakościowe (nominalne) oraz zmienne ilościowe. Zmienne jakościowe pozwalają na klasyfikację chorych względem różnych i rozłącznych kategorii. Typowymi zmiennymi tego rodzaju są: płeć, występowanie określonej choroby czy zaszeregowanie chorych w różnych skalach – skali zaawansowania dławicy piersiowej opracowanej przez Kanadyjskie Towarzystwo Kardiologiczne (ang. *Canadian Cardiovascular Society* – CCS), skali służącej do klasyfikacji ciężkości objawów niewydolności krążenia zaproponowanej przez Nowojorskie Towarzystwo Kardiologiczne (ang. *New York Heart Association* – NYHA) czy klasyfikacji Killipa-Kimballa opracowanej do oceny niewydolności serca w świeżym zawale mięśnia sercowego. Oznacza to m.in., że zmienna jakościowa ma charakter odpowiedzi w postaci słownej i jest niemierzalna. Zmienne ilościowe mogą dostarczyć informacji np. o czasie pomiędzy dwoma zdarzeniami – wówczas mówi się o zmiennej ciągłej. Jeśli określa się liczbę zdarzeń, które miały miejsce w rozpatrywanym przedziale czasu, mówi się o zmiennej dyskretnej lub inaczej skokowej. Przykładem zmiennej ciągłej może być czas hospitalizacji pacjenta, wzrost czy masa ciała. Zmienną dyskretną będzie np. liczba przebytych operacji w dotychczasowym życiu. Tu odpowiedź może być tylko w postaci liczby całkowitej, a nie, jak w przypadku zmiennej ciągłej, gdzie może przyjmować dowolną wartość (czas pobytu w szpitalu może być teoretycznie określony z dowolną dokładnością). Często interesujący jest wpływ jednej cechy jakościowej na drugą cechę tego samego typu, np. wpływu płci na częstość występowania danej choroby lub na rozkład punktacji CCS/NYHA w populacji chorych. W takim przypadku należy posłużyć się odpowiednim aparatem obliczeniowym i testem statystycznym. Najczęściej stosowanym testem jest w takiej sytuacji opracowany w 1900 r. przez Karla Pearsona test niezależności χ^2. Obszar stosowalności tego testu obejmuje nie tylko analizę częstości występowania cech jakościowych, ale również analizę zgodności cech ilościowych oraz przypadki badania zależności cechy jakościowej od ilościowej.

Tak jak każdy test statystyczny, tak też χ^2 umożliwia oszacowanie prawdopodobieństwa błędu pierwszego rodzaju popełnianego przy odrzucaniu hipotezy zerowej (H_0). Podobnie jak w przypadku większości testów stosowanych w analizie danych medycznych, H_0 jest hipotezą zakładającą brak wpływu cechy jakościowej na wartość drugiej ocenianej cechy, np. brak wpływu płci na częstość występowania raka trzonu macicy. Odrzucając H_0, przyjmuje się hipotezę alternatywną, że płeć istotnie wpływa na częstość występowania raka szyjki macicy.

Błąd pierwszego rodzaju popełnia się wówczas, gdy na podstawie przeprowadzonej analizy statystycznej stwierdza się, że uzyskane wyniki są istotne statystycznie, podczas gdy w rzeczywistości jest wręcz odwrotnie. Jako wynik analizy statystycznej podaje się właśnie prawdopodobieństwo popełnienia błędu pierwszego rodzaju (p). Decyzję co do istotności danego prawdopodobieństwa podejmuje badacz, porównując jego wartość z założonym poziomem istotności (α). W naukach medycznych powszechnie przyjmuje się za istotną wartość $p < 0{,}05$, co oznacza, że akceptuje się jeden błąd pierwszego rodzaju na 20 analiz.

Wśród dwóch najważniejszych założeń stosowalności testu χ^2 należy wymienić ograniczenie dotyczące minimalnej liczebności próby oraz niezależność grup. Minimalna liczebność próby powinna wynosić nie mniej niż 5 osób. Związane jest to z faktem, że χ^2 testuje prawdopodobieństwa w poszczególnych komórkach. Oceny tych prawdopodobieństw dla liczebności poniżej 5 mogą być niewystarczająco precyzyjne [3–5]. Przy dużych badaniach spełnienie tego założenia nie stanowi problemu. Jednak przy małej liczbie osób biorących udział w badaniu może dojść do sytuacji, w której jedna z wydzielonych podgrup ma liczebność mniejszą niż wymagana wartość minimalna. Wówczas do używanego testu należy zastosować tzw. poprawki na ciągłość. W większości programów statystycznych są one uwzględniane automatycznie. Aby spełnić warunek niezależności między grupami, w zasadzie wystarczy zwrócić uwagę tylko na to, aby wynik otrzymany dla jednej osoby odzwierciedlał jedną cechę. Zastosowana klasyfikacja powinna być także wyczerpująca – czyli suma osób należących do wszystkich podzbiorów, na które została podzielona grupa, powinna obejmować cały rozpatrywany zbiór – oraz rozłączna, co oznacza,

że jeden element (pacjent) nie może znaleźć się w więcej niż jednej podgrupie. Problem z niezależnością między grupami pojawia się wówczas, gdy zadawane pytania zakładają możliwość wielokrotnej odpowiedzi, wtedy danego pacjenta można zaliczyć do wielu grup i niespełniony jest warunek niezależności pomiędzy nimi.

Do zobrazowania zastosowania testu niezależności χ^2 rozważone zostaną sytuacje, w których spośród populacji wybrano *n* osób do przeprowadzenia na nich badania. Zebrane wyniki dla tej grupy przedstawione będą w postaci tabeli, w której w wierszach/rzędach (r) przedstawione są dane dotyczące jednej cechy, a odpowiednio w kolumnach (k) drugiej. W ten sposób zostanie utworzona tablica liczności, która pozwala na weryfikację H_0, mówiącej, że przy przyjętym poziomie istotności α w populacji nie ma zależności między cechami. Do weryfikacji tej hipotezy stosuję się statystykę χ^2 (pakiety Statistica, SAS).

Modyfikacje testu chi-kwadrat

Należy pamiętać, że test χ^2 stosuje się w przypadku, gdy liczba badanych przekracza 40 osób, a liczebności w każdej podgrupie są nie mniejsze niż 10. W badaniach medycznych jednak bardzo często ma się do czynienia z mniejszą grupą osób lub z mniejszymi liczebnościami w odpowiednich podgrupach. W związku z tym należy zastosować odpowiednie modyfikacje (poprawki) do testu χ^2, tak aby dobrze odzwierciedlał rozpatrywany przypadek.

Poprawki często obserwuje się jako opcje, które można wybrać, wykonując analizę z wykorzystaniem pakietu statystycznego (np. Statistica). Ich mnogość może być kłopotliwa, dlatego poniżej umieszczono podstawowe informacje, kiedy i jaką poprawkę należy zastosować (tab. I).

Dodatkowo, jeżeli tabela, do której wprowadzono dane, składa się z trzech lub więcej rzędów lub kolumn oraz zmienna ma pewien naturalny porządek, np. klasą CCS można testować hipotezę o występowaniu trendu. Do tego celu służy test Cochrana-Armitage [6], który pozwala ocenić, czy występuje liniowa zależność pomiędzy proporcjami w poszczególnych kategoriach. Można go znaleźć np. w pakiecie Statistica czy MedCalc.

Natomiast aby określić istotność różnicy między wynikami w przypadku prób powiązanych, należy zastosować test χ^2 McNemary [3–5]. Test ten stosuje się, gdy przeprowadza się badanie dwukrotnie na tej samej grupie chorych, np. przed podaniem leku i po jego podaniu.

Tab. I. Zestawienie poprawek do testu chi-kwadrat, które należy uwzględnić zależnie od liczebności odpowiednio badanej grupy i podgrup, na które została podzielona

Liczba osób biorących udział w badaniu	Liczebności oczekiwane	Rodzaj testu
$n > 40$	> 10	χ^2
$n > 40$	którakolwiek < 10	test V-kwadrat
$n > 40$	którakolwiek < 5	χ^2 z poprawką Yatesa
$20 < n \leq 40$	wszystkie > 5	χ^2 z poprawką Yatesa
$n \leq 40$	którakolwiek < 5	dokładny test Fishera

Przykłady zastosowania testu chi-kwadrat

W celu głębszego zrozumienia zagadnienia warto posłużyć się konkretnym przykładem. Matusik i wsp. opisali grupę 93 osób, wśród których były 73 kobiety i 20 mężczyzn [1]. Autorzy m.in. podzielili badaną grupę ze względu na występowanie i niewystępowanie nadciśnienia tętniczego (NT). Do analizy zmiennych jakościowych w grupie z NT i bez NT [obecność cukrzycy, niewydolność serca, niewydolność nerek, zespół osłabienia, obecność przewlekłej obturacyjnej choroby płuc (POChP) lub astmy] użyto testu niezależności χ^2. W przypadku, gdy którakolwiek z liczebności oczekiwanych wyniosła 5 lub mniej, użyto testu χ^2 z poprawką Yatesa, dla liczebności poniżej 10 i powyżej 5 autorzy zastosowali test V-kwadrat. Matusik i wsp. podają, że wśród 73 przebadanych kobiet u 14 nie stwierdzono występowania NT (70%), podczas gdy wśród 20 przebadanych mężczyzn u 6 nie stwierdzono NT (30%), a u 14 (70%) choroba ta wystąpiła [1]. Korzystając z tablicy liczebności dla tych danych (tab. II), można obliczyć za pomocą pakietu Statistica wartość statystyki χ^2, która w tym przypadku wynosi $\chi^2 = 1{,}09$ ($p = 0{,}2967$).

Otrzymaną wartość prawdopodobieństwa porównuje się z przyjętym poziomem istotności $\alpha = 0{,}05$. Okazuje się, że dla związku płeć–NT wartość *p* nie przekracza wartości krytycznej, co pozwala przyjąć H_0 o braku wpływu płci na częstość występowania NT w badanej populacji. Podobnie przeprowadzona analiza dotycząca występowania niewydolności serca, niewydolności nerek, zespołu osłabienia w grupie pacjentów z NT i bez niego [1] prowadzi do analogicznych spostrzeżeń (tab. III i IV). W tych przypadkach jednak ze względu na liczebności w niektórych podgrupach nie większe niż 5, zastosowano test χ^2 z poprawką Yatesa.

Podsumowując, otrzymane wartości prawdopodobieństw są z przedziału $(1–\alpha)$, gdzie α oznacza przyjęty poziom istotności równy 0,05. Pozawala to przyjąć na poziomie istotności 0,05 (a taki stosują autorzy) H_0 o braku różnic w częstości występowania cukrzycy, niewydolności serca, niewydolności nerek czy zespołu osłabienia pomiędzy pacjentami bez NT i z NT w badanej grupie. Warto zwrócić uwagę, że różnic istotnych statycznie pomiędzy analizowanymi zmiennymi a występowaniem NT można byłoby się spodziewać, gdyby wystąpiły duże dysproporcje w częstości występowania np. cukrzycy w grupie bez NT i z NT.

Test niezależności χ^2 daje także możliwość, aby w prosty i niebudzący wątpliwości sposób analizować dane odsetkowe. Tego typu problemy występują w przypadkach klasyfikacji chorych wg licznych skal stosowanych w medy-

Tab. II. Występowanie nadciśnienia tętniczego w badanej grupie ($\chi^2 = 1{,}09$)

Nadciśnienie tętnicze	Kobiety	Mężczyźni	*p*
tak	59	14	0,2967
nie	14	6	

Tab. III. Występowanie zespołu osłabienia w badanej grupie z nadciśnieniem tętniczym i bez nadciśnienia tętniczego ($\chi^2 = 0,2$)

Zespół osłabienia	Pacjenci bez nadciśnienia tętniczego	Pacjenci z nadciśnieniem tętniczym	*p*
tak	15	51	0,6539
nie	5	22	

Tab. IV. Występowanie cukrzycy, niewydolności nerek i niewydolności serca w badanej grupie ($\chi_c^2 = 0,01$, $\chi_{nn}^2 = 1,76$, $\chi_{ns}^2 = 0,05$)

		Pacjenci bez nadciśnienia tętniczego *n* = 20	Pacjenci z nadciśnieniem tętniczym *n* = 73	*p*
cukrzyca	tak	4	16	0,9027
	nie	16	57	
niewydolność serca	tak	2	20	0,1852
	nie	18	53	
niewydolność nerek	tak	1	7	0,8174
	nie	19	64	

Tab. V. Klasa NYHA w badanej grupie [7]

Klasa NYHA	Torasemid (*n*)	Furosemid (*n*)
I–II	1	0
II	63	67
II–III	8	15
III	46	30
III–IV	1	1
IV	3	2

Tab. VI. Klasa NYHA w badanej grupie po sumowaniu populacji w klasach I i I–II (obecnie II), III i III–V (obecnie III)

Klasa NYHA	Totrasemid (*n*)	Furosemid (*n*)
II	64	67
II–III	8	15
III	47	31
IV	3	2

cynie klinicznej, takich jak skala NYHA czy skala CCS. Müller i wsp. przedstawiają dane dotyczące liczby chorych w kolejnych klasach NYHA, którzy byli leczeni dwoma różnymi lekami moczopędnymi [7]. Analizując otrzymane wyniki, napotkali na typowy w badaniach medycznych problem niewielkiej liczby pacjentów w podgrupach (tab. V).

W klasie I–II znajduje się tylko jeden chory w grupie leczonej torasemidem, brak natomiast chorych w grupie leczonej furosemidem, co uniemożliwia zastosowanie testu χ^2. Ponadto, w wydzielonej przez autorów klasie III–IV znajduje się zaledwie po jednym chorym w obu grupach. W tej sytuacji autorzy, chcąc porównać rozkład klas NYHA w obu grupach, zdecydowali się na procedurę uśredniania oraz analizy statystycznej „średniej klasy NYHA" wraz z jej odchyleniem standardowym w obu grupach. Jest to typowa nieścisłość, jaką można zaobserwować w licznych pracach z zakresu medycyny. Procedura uśredniania w przypadku tego typu danych (jakościowych) jest niedozwolona, klasy NYHA oznaczone są wprawdzie numerycznie, co zapewne prowadzi do popełnienia tego błędu, ale co w sytuacji, gdyby klasy oznaczone były literami alfabetu?

Rozwiązaniem tego problemu jest zawsze powiększenie grup lub połączenie kilku klas. W cytowanej pracy [7] wystarczy włączyć chorych z klasy I–II do grupy NYHA II, a klasę III–IV do klasy III lub IV, tak jak to zrobiono w tabeli VI.

Analizując tak zmienioną tabelę liczności, nie trzeba uciekać się do nieuprawnionego uśredniania, stosuje się bezpośrednio test χ^2 z poprawką Yatesa, a otrzymuje się w wyniku wartość $\chi^2 = 5,53$ przy 3 stopniach swobody. Wyliczone przez program prawdopodobieństwo takiego wyniku $p = 0,14$ przekracza założoną wartość α, co nie pozwala odrzucić H_0. Analizując wartości średnie NYHA, otrzymuje się $p = 0,198$, co także zmusza do wyciągnięcia analogicznego wniosku. Zatem w pracy nie popełniono błędu wnioskowania. Zastosowano natomiast źle dobraną metodę dającą akurat w tym przypadku prawidłowy wynik.

Podsumowanie

Analizowanie danych z wykorzystaniem testu χ^2 nie sprawia większych problemów pod warunkiem, że badacz posiadł podstawową wiedzę z zakresu statystyki. Nawet nie będąc ekspertem w tej dziedzinie, lekarz prowadzący badania naukowe może stać się w pełni świadomym, zaawansowanym użytkownikiem pakietu statystycznego. Wiedza ta z pewnością ułatwi planowanie dalszych badań i analizę otrzymywanych wyników.

Należy zwrócić uwagę, że im większa jest różnica pomiędzy wartościami doświadczalnymi, czyli faktycznie zaobserwowanymi przez badacza, a teoretycznymi, czyli takimi, jakich należałoby oczekiwać, gdyby zmienne były od siebie niezależne, tym silniejsza zależność pomiędzy grupami. Wprawdzie nie przesądza to jeszcze o istotności statystycznej, jednak pozwala wyrobić sobie pewną intuicję co do oczekiwanych wyników. Często analizując już same tylko tabele liczebności dla przeprowadzonego badania, można wstępnie ocenić, czy istnieje szansa, że zależności takie występują. Dodatkowo wiadomo będzie, czy do przeprowadzenia analizy statystycznej wystarczy użyć testu niezależności χ^2, czy którejś z jego modyfikacji. Jak widać, wstępna analiza jest bardzo prosta, nie wymaga znajomości zaawansowanego modelu matematycznego czy analizy statystycznej, pozwala natomiast uchronić się od popełnienia podstawowych błędów.

Piśmiennictwo

1. Matusik P, Nowak J, Tomaszewski K, Chmielowska K, Parnicka A, Dubiel M, Gąsowski J. Nadciśnienie tętnicze u osób w wieku podeszłym na przykładzie mieszkańców domów opieki. Polski Przegląd Kardiologiczny 2010; 12; 186-191.
2. Weiss SA, Blumenthal RS, Sharrett AR, Redberg RF, Mora S. Exercise blood pressure and future cardiovascular death in asymptomatic individuals. Circulation 2010; 121: 2109-2116.

3. Kendall M, Stuart A. The advanced theory of statistics. 4th ed. Griffin, London 1979.
4. Fienberg SE. The analysis of cross-classified categorical data. 2nd ed. Springer 216, New York 2007.
5. Bishop YMM, Fienberg SE, Holland PW. Discrete Multivariate Analysis: Theory and Practice. MA: MIT Press, Cambridge 1975.
6. Armitage P. Tests for linear trends in proportions and frequencies. Biometrics 1955; 11: 375-386.
7. Müller K, Gamba G, Jaquet F, Hess B. Torasemide vs. furosemide in primary care patients with chronic heart failure NYHA II to IV – efficacy and quality of life. Eur J Heart Fail 2003; 5: 793-801.