Iwona Mróz,

Instytut Fizyki Doświadczalnej,

Uniwersytet Wrocławski

**ZMAD – materiały do wykładu nr 5 i 6 z dni 28.03 i 4.04.2022 r. – część 1**

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

**Badanie współzależności cech**

Pytanie o współzależność cech jest bardzo ważnym pytaniem, na które musimy odpowiedzieć m.in. przy budowaniu modeli regresyjnych. Pomiędzy cechami statystycznymi mierzonymi w różnych skalach mogą występować różne zależności. W przypadku cech ilościowych można mierzyć np. siłę związku liniowego miedzy cechami.

„Korelacja między dwiema losowymi zmiennymi X i Y jest miarą siły (stopnia) liniowego związku między tymi zmiennymi.”[[1]](#footnote-1)

Najpopularniejszą miarą siły takiego związku liniowego (korelacji) pomiędzy cechami statystycznymi o charakterze ilościowym jest współczynnik korelacji Pearsona, oznaczany dla populacji symbolem ρ, a dla próby *r*. Precyzyjna definicja współczynnika korelacji zostanie podana w dalszej części wykładu, w tej chwili podajemy wzory pozwalające na obliczenie wartości współczynnika.

Kowariancją zmiennych losowych X i Y nazywamy wyrażenie:

[[2]](#footnote-2) (13a.1)

Współczynnikiem korelacji w populacji nazywamy wyrażenie:

[[3]](#footnote-3) (13a.2).

Wartość współczynnika korelacji mieści się w przedziale <-1, 1>, przy czym wartości ujemne oznaczają korelację ujemną (wzrost wartości jednej zmiennej pociąga za sobą spadek wartości drugiej zmiennej), wartości dodatnie mówią o korelacji dodatniej, a zero oznacza brak korelacji liniowej. Przy badaniu korelacji nie określa się zmiennych zależnych i niezależnych.

Rozważny teraz n – elementową próbę, której realizacjami są pary , *i = 1,…,n*.

Estymator kowariancji wyliczony na podstawie próby ma postać:

[[4]](#footnote-4) (13a.4),

A współczynnik korelacji wyliczony z próby:

[[5]](#footnote-5) (13a.4)

Wartość współczynnika korelacji wyliczona z próby może mieć wysoką wartość, ale nie oznacza to, że jest ona istotna. Dlatego trzeba przetestować istotność współczynnika korelacji. Najbardziej popularny test badający istotność korelacji sprawdza, czy zmienne losowe X i Y są ze sobą skorelowane.[[6]](#footnote-6) Hipotezy tego testu mają postać:

H0: ρ = 0

H1: ρ ≠ 0,

a statystyka (sprawdzian hipotezy):

(13a.5)

ma przy założeniu prawdziwości H0 rozkład Studenta o .

Inną miarą, opartą o współczynnik korelacji Pearsona jest *współczynnik korelacji cząstkowej*.[[7]](#footnote-7) Wykorzystujemy go jeżeli w badanym układzie występuje więcej zmiennych losowych, a chcemy zbadać związek pomiędzy dwiema z nich, wyznaczamy *współczynnik korelacji cząstkowej*. Przykładowo, dla układu, w którym występuję trzy zmienne losowe (opisujące trzy cechy statystyczne): *X1, X2, X3*, chcemy zbadać związek liniowy pomiędzy zmiennymi *X1* i *X2* z wyłączeniem zmiennej losowej *X3*. Odpowiedni współczynnik korelacji cząstkowej ma postać:



**Korelacje nieparametryczne**

*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. 295-296 i 336-338.*

W przypadku danych jakościowych w skali porządkowej lub wtedy, gdy dla danych w skali mierzalnej nie można obliczyć współczynnika korelacji liniowej Persona, możemy ocenić korelacje („siłę związku”) pomiędzy zmiennymi przy pomocy metod nieparametrycznych.

Andrzej Stanisz wymienia kilka metod określania korelacji nieparametrycznych, z których najbardziej popularne wydaje się obliczanie współczynnika korelacji rang R Spearmana. Wspominamy też o współczynniku tau Kendalla, który jest używany jako jedna z metod obliczania wartości korelacji w R.

1. Współczynnik R Spearmana wymaga nadania poszczególnym obserwacjom rang (procedura rangowania została opisana przy testach nieparametrycznych dla dwóch grup niezależnych), więc może być stosowany dla danych w skali porządkowej. A. Stanisz podaje, że warto go również stosować dla danych ilościowych gdy mamy mało pomiarów, występują obserwacje odstające lub nie są spełnione założenia pozwalające zastosować współczynnik korelacji liniowej Persona (założenia te omówimy w dalszej części wykładu).

Współczynnik R Spearmana przyjmuje wartości .

Załóżmy, że mamy *n* obserwacji, dla każdej z nich podajemy wartości dwóch cech: X i Y. Porządkujemy obie cechy rosnąco i przypisujemy im rangi. Współczynnik R obliczamy według wzoru:

 (13a.6)

*di* oznacza różnicę rang nadanych wartościom cech X i Y dla *i-tej* obserwacji. Na przykład, dla czwartej obserwacji wartość cechy X ma rangę 2, a wartość cechy Y – rangę 5. Wtedy .

Wartość współczynnika rs, tak samo jak współczynnika r Pearsona jest wyliczona na podstawie danych z próby i należy sprawdzić, czy dla populacji jest ona statystycznie istotna, czyli różna od zera.

2. Współczynnik tau Kendalla „…opiera się na różnicy między prawdopodobieństwem tego, że dwie zmienne układają się w tym samym porządku (dla obserwowanych danych), a prawdopodobieństwem, że ich uporządkowanie się różni.” (A. Stanisz, str. 313-314) Dane muszą być wyrażone co najmniej w skali porządkowej. Wartości współczynnika przebiegają od -1 do 1, przy czym -1 odpowiada całkowitej przeciwstawności uporządkowania, 0 oznacza brak uporządkowania, a 1 – całkowite uporządkowanie. „Współczynnik Kendalla wskazuje więc nie tylko siłę, lecz również kierunek zależności.” (A. Stanisz, str. 314).

**„Siła związku” dla cech jakościowych**

*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. 325-326.*

W przypadku cech jakościowych także możemy zbadać „siłę związku” między cechami. Przypomnijmy, że do oceny, czy cechy statystyczne są od siebie niezależne, wykorzystujemy testy niezależności, np. test chi-kwadrat. Dla tego testu jest wyliczana wartość statystyki , która pozwala ocenić, dla danego poziomu istotności, czy rozpatrywane cechy są niezależne. Sama wartość statystyki nie pozwala na ocenę stopnia zależności pomiędzy cechami „…zależy ona bowiem od liczebności grupy N i rośnie wraz z jej wzrostem.”[[8]](#footnote-8) Jest ona jednak wykorzystywana do konstruowania miar siły związku. A. Stanisz wymienia:

1. Współczynnik Φ Yule’a, wykorzystywany dla tabel wielodzielczych 2×2. Wartości współczynnika Yule’a mieszczą się w przedziale <0, 1>, przy czym wartość 0 oznacza brak związku między cechami, a 1 oznacza pełny związek. Współczynnik Yule’a wyraża się wzorem:

,

gdzie to wartość statystyki , a N jest liczbą obserwacji.

2. Dla tabel wielodzielczych o wymiarach p×q popularną miarą siły związku dla cech jakościowych jest współczynnik V-Cramera, który może przybierać wartości <0, 1> tak jak współczynnik Yule’a. Współczynnik V-Cramera wyznaczamy ze wzoru:

3. Inną miarą stosowaną zwykle dla kwadratowych tabel wielodzielczych jest współczynnik kontyngencji Pearsona, wyliczany w postaci nieskorygowanej jako:

Wartość współczynnika C jest zawsze mniejsza od 1, przy czym minimalna wartość 0 oznacza brak związku między cechami. Można też obliczać skorygowany współczynnik C, którego wartości mieszczą się w przedziale <0, 1>.

1. Aczel, A.D., Statystyka w zarządzaniu”, Wydawnictwo Naukowe PWN, Warszawa 2000, s. 479. [↑](#footnote-ref-1)
2. Ibidem, s. 481. [↑](#footnote-ref-2)
3. Ibidem, s. 481. [↑](#footnote-ref-3)
4. Ibidem, s. 467. [↑](#footnote-ref-4)
5. Ibidem, s. 467. [↑](#footnote-ref-5)
6. Opis testu przytoczono za: Ibidem, s. 482. [↑](#footnote-ref-6)
7. Stanisz, A., Przystępny kurs statystyki z zastosowaniem STATISTICA.PL na przykładach z medycyny, StatSoft, Kraków 2006, t.1, s. 294. [↑](#footnote-ref-7)
8. Ibidem, s. 325. [↑](#footnote-ref-8)