# Rozdział 1. Analiza zależności zmiennych niemetrycznych

*Statistics is the grammar of science*.
[Statystyka jest gramatyką nauki.]
**Karl Pearson (1857–1936)**

## 1.1. Pojęcie zależności w statystycznej analizie danych

Jednym z podstawowych rodzajów badań i analiz, które są obecne niemal w każdej dziedzinie naukowej, są wielowymiarowe analizy porównawcze. Wielowymiarowość jest z jednej strony wielkim atutem, gdyż dostarcza pełnej informacji o zjawisku, z drugiej jednak strony – komplikuje i utrudnia proces pozyskania istotnych i syntetycznych informacji. Wielowymiarowy charakter zjawisk ekonomicznych nie stanowi jednak problemu, jeśli do analizy są wykorzystywane odpowiednie narzędzia i metody statystyczne, takie jak metody wielowymiarowej analizy danych. Metody te pozwalają na analizę struktury złożonych zjawisk, dostarczają o nim szerokiej wiedzy, upraszczają złożoną rzeczywistość poprzez redukcję wymiaru, a także pozwalają na interpretację otrzymanych wyników. Sięgają one do różnych dziedzin, takich jak filozofia, ekonomia, socjologia, psychologia, antropologia kulturowa, semiotyka, co świadczy o ich interdyscyplinarnym charakterze. Są uniwersalnym i praktycznym narzędziem analizy stosowanym niemal we wszystkich obszarach naukowych. Wykorzystywanie metod statystycznych w różnych dziedzinach nauki spowodowało, iż doczekały się one własnej terminologii, jak np.: demografia, ekonometria, psychometria, biometria, socjologia statystyczna czy też statystyka gospodarcza.

Metody wielowymiarowej analizy statystycznej z powodzeniem znajdują zastosowanie w różnych gałęziach naukowych takich, jak:

– ekonomia [Mynarski, 1973, 1987, 1990, 2000, 2001; Pluta, 1977; Kolonko, 1980; Gordon, 1981; Krzyśko, 1982; Lebart, Morineau, Warwick, 1984; Walesiak, 1985, 1993, 1996, 2002, 2004, 2011; Luck, Rubin, 1987; Sikorski, 1987; Green, Tull, Albaum, 1988; Jajuga, 1990; Kinnear, Taylor, 1991; McDaniel, Gates, 1991, 1993; Duliniec, 1994; Kinnear, Taylor, 1996; Venables, Ripley, 1997; Cameron, Trivedi, 1998; Gatnar, 1998a, 1998b; Ostasiewicz, 1998; Walesiak, Bąk, 2000, 2013; Everitt, Dunn, 2001; Franes, Paap, 2001;

Zaborski, 2001; Bąk, 2004, 2009, 2010, 2011, 2012; Gruszczyński, 2002; Kaczmarczyk, 2002, 2011; Rószkiewicz, 2002; Sagan, 2002, 2004a; Churchill, Iacobucci, 2004; Gatnar, Walesiak, 2004, 2011; Dobson, Barnett, 2008; Luszniewicz, Słaby, 2008; Powers, Xie, 2008; Panek, 2009; Kopczewska, Kopczewski, Wójcik, 2009; Sagan, Perek-Białas, 2011; Walesiak, Gatnar, 2009; Dudek, 2013; Balicki, 2013; Wiśniewski, 2013],

- socjologia [Pawłowski, 1971; Hays, 1973; Armor, 1974; Abell, 1975; Lissowski, 1977, 1984, 2005; Goodman, 1978; Koseła, Utzig, 1980; Heise, 1986; Sawiński, Domański, 1986; Nawojczyk, McCutcheon, 1996; Ananth, Kleinbaum, 1997; Agresti, 2002, 2010; Domański, Przybysz, 2007; Sawiński, 2010; Lissowski, Haman, Jasiński, 2011a, 2011b, 2011c],
- psychologia [McNemar, 1955; Siegel, 1956; Okóń, 1960; Tyler, 1967; Edwards, 1972; Coombs, Dawes, Twersky, 1977; Brzeziński, 1987, 1993; Heinen, 1996; Foxall, Goldsmith, 1998; Aron, Aron, 2002; Gravetter, Wallnau, 2006; Francuz, Mackiewicz, 2007; Silverman, 2007; Kwiatkowska, Stasiuk, 2008; Tabachnik, Fidell, 2013],
- biostatystyka [Pagano, Gauvreas, 2000; Newman, 2001; Watała, 2002; Brzeziński, 2011],
- informatyka [Lebart, Fénelon, 1973; Bakerman Robinson, 1994; Domański, 1996; Bąk, 1999; Walesiak, 2004; Gągolewski, 2014],
- zarządzanie [Hutcheson, Moutinho, 2008; Aczel, 2011],
- geografia [Chojnicki, 1977; Racine, Reymond, 1978],
- antropologia [Czekanowski, 1913],
- nauki polityczne [Garson, 1976].

Mynarski [2000] dokonuje podziału metod statystycznych na trzy podstawowe grupy. Pierwszą z nich stanowią metody analizy niezależności, do których należą testy istotności, pozwalające weryfikować hipotezę o nieistnieniu różnic między zmiennymi. Testy te umożliwiają wykrycie pewnych obszarów separowalności w zbiorach danych, które są traktowane jako odrębne próby pochodzące z różnych populacji porównywanych ze sobą. Przy danym poziomie istotności (prawdopodobieństwie), pozwalają odrzucić hipotezę zerową o niezależności na podstawie stwierdzenia, czy różnice pomiędzy badanymi grupami są istotne, czy też nie. Do testów tych należą m.in. testy nieparametryczne dla zmiennych mierzonych na słabych skalach pomiaru (test niezależności chi-kwadrat, test McNemara, test Cochrana, test Kołmogorowa, test serii), jak i testy parametryczne, przeznaczone dla zmiennych mierzonych na skalach mocnych (test $z$, test $t$, test $F$ ilorazu wariancji).

Druga grupa metod wielowymiarowej analizy statystycznej to opisowe metody analizy zależności (korelacji), polegające na znajdowaniu zależności pomiędzy zmiennymi bez podziału na zmienną zależną i zmienne niezależne. W metodach tych wykorzystuje się współczynniki przeznaczone zarówno dla zmiennych mierzonych na skali nominalnej ($\phi$ Yule'a, $Q$ Kendalla, $C$ Pearsona, $T$ Czuprowa, $V$ Cramera, $\lambda$ Goodmana i Kruskala), jak i porządkowej ($\tau$ Kendalla, $\gamma$ Goodmana i Kruskala, $d$ Somersa i $r$ Spearmana). Podobnego podziału metod analizy

zmiennych niemetrycznych dokonali także Reynolds [1977] i Blalock [1979], zaliczając do pierwszej grupy metod testy niezależności (*tests of independence*), a do drugiej analizę zależności (*analysis of association*).

Trzecia, najobszerniejsza i najbardziej dynamicznie rozwijająca się grupa metod, to statystyczna analiza wielowymiarowa. W metodach tych charakterystyczne jest to, iż jednoczesnej analizie są poddane pomiary na przynajmniej trzech zmiennych opisujących każdy obiekt badania [Walesiak, 1996, 2011]. Klasyfikację metod wielowymiarowych zaproponowano w: [Kendall, 1975; Borys, 1984; Jajuga, 1987; Hair, Andreson, Tacham, Black, 1995; Walesiak, 1996, 2002; Lattin, Carroll, Green, 2003; Gatnar, Walesiak, 2004, 2011]. Celem wielowymiarowych metod jest redukcja informacji do kilku podstawowych kategorii, otrzymanie jednorodnych grup obiektów ze względu na charakteryzujące je właściwości, a także wyjaśnienie struktury powiązań pomiędzy charakterystykami obiektów opisanych przez wiele zmiennych w celu redukcji wymiaru przestrzeni.

Wśród kryteriów klasyfikacji wielowymiarowych metod statystycznych wyróżnia się: rodzaj badania, skale pomiaru zmiennych, przedmiot badania oraz rodzaj metody.

Jeśli chodzi o rodzaj badania, podziału metod można dokonać ze względu na występowanie lub brak występowania w zbiorze danych zmiennej zależnej (zmiennych zależnych) od innych zmiennych (zmiennych niezależnych). Według tego kryterium można wyróżnić z jednej strony metody badania zależności (*dependence methods*), do których należą m.in. analiza regresji, drzewa klasyfikacyjne i analiza *conjoint*, z drugiej strony metody badania współzależności (*interdependence methods*), do których należą m.in. analiza korespondencji, skalowanie wielowymiarowe, analiza czynnikowa czy metody porządkowania liniowego.

Ze względu na drugie kryterium, jakim jest skala pomiaru, metody te można podzielić w zależności od rodzaju skali pomiaru zmiennej zależnej, jak i zmiennych niezależnych. Wyróżnia się metody przeznaczone do analizy zmiennych niemetrycznych (nominalnych, porządkowych) oraz metrycznych (przedziałowych i ilorazowych) [Stevens, 1951, 1959; Jajuga, 1993; Walesiak, 1990, 1991, 2002; Rusnak, 1998; Mynarski, 2000; Kaczmarczyk, 2002].

Trzecim kryterium podziału wielowymiarowych metod statystycznych jest przedmiot badania, którym mogą być obiekty (drzewa klasyfikacyjne, skalowanie wielowymiarowe, analiza korespondencji, porządkowanie liniowe) lub zmienne (analiza regresji, analiza *conjoint*, analiza czynnikowa).

Wielowymiarowe metody, mogą mieć charakter konfirmacyjny lub eksploracyjny. Metody konfirmacyjne w ujęciu popperowskiego paradygmatu rozwoju nauki polegają na sformułowaniu hipotezy dotyczącej prawidłowości na poziomie empirycznym, a następnie jej weryfikacji. Popper jest autorem skrajnej postaci redukcjonizmu, nazwanego falsyfikacjonizmem, który zalecał stosować we wszystkich naukach empirycznych, m.in. społecznych [Popper, 1959; Stachak 2003]. Metody eksploracyjne natomiast umożliwiają zidentyfikowanie natury i mechanizmów kształtujących badane zjawisko.

W statystycznej analizie danych w ramach metod jakościowych (*qualitative methods*) wyróżnia się dwa podstawowe nurty. Pierwszy dotyczy nauk społecznych i w nim przez metody jakościowe rozumie się badania bez przeprowadzania obliczeń [Nikodemska-Wołowik, 1999; Marshall, Kędzior, 2004; Silverman, 2007; Rossmann, 2009; Wolcott, 2010]. Drugi nurt, obejmujący m.in. statystykę, przez metody jakościowe rozumie analizę zmiennych mierzonych na słabych skalach pomiarowych (skala nominalna i porządkowa) [McFadden, 1973; Borys, 1980, 1984; Kolonko, 1980; Maddala, 1986; Walesiak, 1990, 1991, 1993, 2003, 2004; Gatnar, 2003; Bąk, 2004; Marzec, 2008; Kowal, 2009; Sawiński, 2010; Gatnar, Walesiak, 2011; Creswell 2013]. Niniejsza monografia jest poświęcona metodom statystycznym należącym do nurtu drugiego i obejmuje metody analizy zmiennych mierzonych na słabych skalach pomiaru.

Problem zależności, synonimicznie rozumiany jest jako korelacja (łac. *correlatio* – razem, łącznie, *relatio* – związek, relacja), asocjacja (łac. *accociatio* – połączenie), związek, relacja czy też powiązanie, jest jednym z najważniejszych i najtrudniejszych problemów występujących w nauce. Brak jednoznacznie określonego stanowiska co do tego pojęcia sprawia, że od starożytności toczy się spór o jego rolę, miejsce i znaczenie, a kontrowersje i dyskusje, które miały początek w czasach starożytnych, są kontynuowane do dziś. Liczne poglądy filozoficzne na przestrzeni dziejów dowodzą, że temat zależności i związku pomiędzy różnymi pojęciami, abstraktami czy też kategoriami filozoficznymi jest obecny w nauce od zawsze i zajmuje on badaczy wielu dyscyplin, m.in. przedstawicieli nauk społecznych, w tym ekonomistów.

Punktem wyjścia w analizie danych jest przede wszystkim ocena związku zachodzącego pomiędzy badanymi zmiennymi. W analizie korelacji wyróżnia się dwa rodzaje współzależności między badanymi zmiennymi. Pierwszą jest zależność funkcyjna, która polega na tym, że zmiana wartości jednej zmiennej powoduje ściśle określoną zmianę wartości drugiej zmiennej, a określonej wartości jednej zmiennej odpowiada tylko jedna wartość drugiej. Zależność funkcyjną prezentuje się zazwyczaj w postaci funkcji, gdzie $X$ jest zmienną zależną, a $Y$ zmienną niezależną. Druga to zależność stochastyczna (probabilistyczna), która występuje wówczas, gdy wraz ze zmianą wartości jednej zmiennej zmienia się rozkład drugiej zmiennej. Ze związkiem stochastycznym mamy do czynienia wtedy, gdy różne wartości jednej zmiennej współwystępują z tymi samymi wartościami drugiej zmiennej lub z tymi samymi kombinacjami wartości innych analizowanych zmiennych [Lissowski, Haman, Jasiński, 2011a]. Szczególnym przypadkiem takiej zależności jest zależność korelacyjna polegająca na tym, że określonym wartościom jednej zmiennej odpowiadają ściśle określone wartości drugiej zmiennej. Można dzięki temu ustalić, jak zmieni się, średnio rzecz biorąc, wartość zmiennej zależnej $Y$ w zależności od wartości zmiennej niezależnej $X$. W obrębie zależności stochastycznej wyróżnia się związki przyczynowo-skutkowe, symptomatyczne i pozorne. W praktyce zależność stochastyczna oznacza, że wpływ jednej zmiennej na drugą jest zależny również od czynników losowych

wspólnie działających na obie zmienne, oprócz tych, które działają na każdą z nich oddzielnie. Gdy obie zmienne wpływają na siebie jednocześnie, można mówić o ich współzależności.

## 1.2. Analiza zależności zmiennych nominalnych

Tablice, które stanowią podstawową formę zapisu zmiennych niemetrycznych, były znane w historii już ponad 2000 lat przed naszą erą. Babilończycy wykorzystywali je do przedstawienia zależności w pewnym systemie liczbowym. Matematycy chińscy używali tablic liczbowych w obliczeniach, które niewiele różniły się od znanej dziś tabliczki mnożenia [Crilly, 2008]. Część etymologów za źródłosłów terminu tablica uważa słowo stół (*table*) (w czasach średniowiecznych był on wykorzystywany do układania na nim należności podatkowych od obywateli państwa). W XVIII wieku, kiedy rozwinęła się statystyka państwowa, tablice były wykorzystywane do opisu zasobów państwa. Kluczowym okresem z punktu widzenia statystyki jako nauki jest przełom XIX i XX wieku, kiedy zaczęto analizować formalne własności tablic.

Tablica kontyngencji (*contingency table*), pełniąca w analizie danych jakościowych kluczową rolę, stanowi podstawową formę zapisu zmiennych niemetrycznych. Alternatywne nazwy tej tablicy to tablica krzyżowa (*cross–classified table, cross tabulation*) lub tablica wielowymiarowa (*multi–way table*). Początkowo prace z zakresu analizy danych niemetrycznych dotyczyły liczebności oczekiwanych [Galton, 1892]. Pionierem w tym obszarze był jednak Karl Pearson, który zainspirowany problemem losowości wyników ruletki w Monte Carlo zaproponował statystykę chi-kwadrat [Pearson, 1900], a następnie jako pierwszy zdefiniował pojęcie tablicy kontyngencji [Pearson, 1904a, 1904b]. Kolejną kluczową postacią w analizie danych niemetrycznych był George Udny Yule, który w latach 1900–1912 badał związki zachodzące pomiędzy zmiennymi i który jako pierwszy wprowadził współczynnik Yule'a odnoszący się bezpośrednio do ich liczebności [Yule, 1900, 1903, 1912]. Istotną postacią, mającą pokaźny wkład w analizę danych jakościowych, był też Ronald Aylmer Fisher, który zaproponował mierniki bazujące na współczynniku chi-kwadrat, pozwalające na analizę zależności. Jako pierwszy badał on interakcje występujące między zmiennymi i wprowadził pojęcie stopni swobody [Fisher, 1922, 1926].

### 1.2.1. Dwuwymiarowe tablice kontyngencji $2 \times 2$

Zagadnieniem dwuwymiarowych tablic kontyngencji zajmowali się: Galton [1892], Pearson [1900, 1904a, 1904b], Yule [1900, 1912], Pearson, Heron [1913], Fisher [1922, 1926], Yates [1934], Bartlett [1935], Deming, Stephan [1940], Norton [1945], Neyman [1949], Cochran [1954], McNemar [1955], Mantel, Haenszel [1959], Bennet, Hsu [1960], Hays [1963], Mood, Graybill, 1963], Conover [1968,

1974, 1980], Fienberg [1968, 1969, 1970, 1971], Mantel, Greenhouse [1968], Altham [1969, 1971], Gart [1969], Miettinen [1969], Fienberg, Gilbert [1970], Fleiss [1973], Camili, Hopkins [1978], Gail, Gart [1973], Kendall, Stuart [1973], Snee [1974], Cox [1970], Zelen [1971], Everitt [1977], Reynolds [1977], Hills, Armitage [1979], Upton [1982], Liebetau [1983], Barnard [1984], Robins, Breslow, Greenland [1986], Kenward [Jones, 1987], Little [1989], Agresti [2002].

W literaturze polskiej reguły budowy tablic kontyngencji wykorzystywali m.in. Hellwig [1975], Steczkowski, Zeliaś [1981], Walesiak, Gatnar [2009], Gatnar, Walesiak [2004, 2011], Sagan [2004a].

Tablica kontyngencji o wymiarach $2 \times 2$ jest tablicą kwadratową, dla której dwie zmienne $X$ i $Y$ mają odpowiednio po dwie kategorie $\{X_1, X_2\}$ oraz $\{Y_1, Y_2\}$ (tab. 1.1). Empiryczne (zaobserwowane) liczebności w $h$-tym wierszu i $j$-ej kolumnie to $n_{hj}$ $(h = 1, 2,\ j = 1, 2)$ – oznaczają one liczbę jednoczesnych wystąpień $h$-tej kategorii cechy $X$ oraz $j$-ej kategorii zmiennej $Y$.

**Tabela 1.1.** Tablica kontyngencji $2 \times 2$ dla dwóch zmiennych

| Zmienna $X$ | Zmienna $Y$ | | |
|---|---|---|---|
| | $Y_1$ | $Y_2$ | Suma |
| $X_1$ | $n_{11}$ | $n_{12}$ | $n_{1\bullet}$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | $n_{2\bullet}$ |
| Suma | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n_{\bullet\bullet}$ |

Źródło: opracowanie własne.

Całkowita liczebność dwuwymiarowej tablicy $2 \times 2$ wynosi

$$n = n_{\bullet\bullet} = \sum_{h=1}^{2} \sum_{j=1}^{2} n_{hj}. \tag{1.1}$$

Elementy $n_{h\bullet}$ oznaczają liczebności brzegowe wierszy, które są zdefiniowane następująco:

$$n_{h\bullet} = \sum_{j=1}^{2} n_{hj}, \tag{1.2}$$

natomiast elementy $n_{\bullet j}$ są liczebnościami brzegowymi kolumn i są wyznaczane za pomocą wzoru

$$n_{\bullet j} = \sum_{h=1}^{2} n_{hj}. \tag{1.3}$$

Dla tak określonej tablicy można wyznaczyć tablicę rozkładu prawdopodobieństwa **P**. Rozkład ten, ze względu na jednoczesne uwzględnienie dwóch badanych zmiennych, nosi nazwę łącznego rozkładu prawdopodobieństwa (*joint distribution*) (tab. 1.2).

**Tabela 1.2.** Tablica rozkładu prawdopodobieństwa $2 \times 2$ dla dwóch zmiennych

| Zmienna $X$ | Zmienna $Y$ | | |
|---|---|---|---|
| | $Y_1$ | $Y_2$ | Suma |
| $X_1$ | $p_{11}$ | $p_{12}$ | $p_{1\bullet}$ |
| $X_2$ | $p_{21}$ | $p_{22}$ | $p_{2\bullet}$ |
| Suma | $p_{\bullet 1}$ | $p_{\bullet 2}$ | 1 |

Źródło: opracowanie własne.

W tabeli 1.2 elementy $p_{hj}$ oznaczają prawdopodobieństwa wystąpienia $h$-tej kategorii zmiennej $X$ oraz $j$-ej kategorii zmiennej $Y$, tj. elementu $n_{hj}$. Częstości brzegowe wierszy są wyznaczane jako

$$p_{h\bullet} = \sum_{j=1}^{2} p_{hj} = \sum_{j=1}^{2} \frac{n_{hj}}{n}, \tag{1.4}$$

natomiast częstości brzegowe kolumn są wyznaczane według formuły

$$p_{\bullet j} = \sum_{h=1}^{2} p_{hj} = \sum_{h=1}^{2} \frac{n_{hj}}{n}, \tag{1.5}$$

gdzie $n$ oznacza ogólną liczebność tablicy (por. wzór (1.1)).

W przypadku tablicy kontyngencji opisującej dwie zmienne można wyznaczyć tablicę liczebności oczekiwanych $m_{hj}$ (tab. 1.3).

**Tabela 1.3.** Tablica $2 \times 2$ liczebności oczekiwanych dla dwóch zmiennych

| Zmienna $X$ | Zmienna $Y$ | | |
|---|---|---|---|
| | $Y_1$ | $Y_2$ | Suma |
| $X_1$ | $m_{11}$ | $m_{12}$ | $n_{1\bullet}$ |
| $X_2$ | $m_{21}$ | $m_{22}$ | $n_{2\bullet}$ |
| Suma | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n$ |

Źródło: opracowanie własne.

Niech $p_{h\bullet}$ oznacza prawdopodobieństwo w populacji, że zmienna wierszowa należy do $h$-tej kategorii, natomiast $p_{\bullet j}$ prawdopodobieństwo, że zmienna kolumnowa należy do $j$-ej kategorii. Niezależność między dwiema zmiennymi w populacji występuje wtedy, gdy

$$p_{hj} = p_{h\bullet} \cdot p_{\bullet j}. \tag{1.6}$$

Liczebności teoretyczne są zdefiniowane w następujący sposób:

$$m_{hj} = n \cdot p_{hj} = \frac{n_{h\bullet} \cdot n_{\bullet j}}{n}. \tag{1.7}$$

Wartości prawdopodobieństw $p_{hj}$ są wartościami nieznanymi w populacji, jednak mogą być one estymowane za pomocą liczebności zaobserwowanych

w próbie. Estymatorami prawdopodobieństw $p_{h\bullet}$ oraz $p_{\bullet j}$ wyznaczonymi za pomocą metody największej wiarygodności są odpowiednio $\hat{p}_{h\bullet}$ oraz $\hat{p}_{\bullet j}$ [Everitt, 1977]:

$$\hat{p}_{h\bullet} = \frac{n_{h\bullet}}{n}, \tag{1.8}$$

$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}. \tag{1.9}$$

Wykorzystanie tych estymatorów pozwala także na estymację wartości oczekiwanych $m_{hj}$ (tab. 1.3). Jeśli zmienne są niezależne, spełnione jest równanie

$$\hat{m}_{hj} = n \cdot \hat{p}_{h\bullet} \cdot \hat{p}_{\bullet j} = \frac{n_{h\bullet} \cdot n_{\bullet j}}{n}, \tag{1.10}$$

i liczebności empiryczne oraz teoretyczne nie powinny się różnić. Jeśli natomiast zmienne nie są niezależne, różnica pomiędzy nimi jest duża.

### Iloraz szans

Podstawową miarą niezależności zmiennych niemetrycznych, która w analizie logarytmiczno-liniowej pełni kluczową rolę, jest szansa. Słowo to jest pojmowane jest dwojako. Pierwsze rozumienie odpowiada angielskiemu *chance* i jest traktowane jako synonim słowa prawdopodobieństwo, drugie zaś odpowiada angielskiemu *odds*, które określa szansę zajścia pewnego zdarzenia w stosunku do innego zdarzenia.

Jeśli zmienna wierszowa $X$ w tablicy kontyngencji o wymiarach $2 \times 2$ jest zmienną dychotomiczną i może przyjąć tylko dwie kategorie, wówczas występuje odpowiednio w pierwszej lub drugiej kategorii. Szansa w znaczeniu *odds* jest rozumiana jako iloraz sukcesu (wystąpienia w pierwszej kategorii) i porażki (wystąpienia w drugiej kategorii). Jeśli $p$ jest prawdopodobieństwem sukcesu, a $1 - p$ jest prawdopodobieństwem porażki, wówczas szansa (bez uwzględnienia kategorii zmiennej kolumnowej) jest zdefiniowana jako iloraz tych zdarzeń [Agresti, 2002]

$$\Omega = \frac{p}{1-p}. \tag{1.11}$$

Jeśli sukces jest bardziej prawdopodobny niż porażka, wtedy szansa $\Omega > 1$. Szansa, określona jako stosunek prawdopodobieństwa wystąpienia zjawiska do prawdopodobieństwa jego niewystąpienia, łączy się z pojęciem prawdopodobieństwa. Jeśli szansa jest większa od 1, wówczas prawdopodobieństwo wystąpienia zjawiska jest większe od 0,5. Jeśli natomiast szansa jest mniejsza od 1, wówczas prawdopodobieństwo zdarzenia jest mniejsze od 0,5 [Christensen, 1997].

Prawdopodobieństwo tak określonej szansy jest równe

$$p = \frac{\Omega}{1+\Omega}. \tag{1.12}$$

Szanse dla zmiennej wierszowej $X$ w tablicy o wymiarach $2 \times 2$ (tab. 1.2), odpowiednio dla wiersza pierwszego oraz drugiego, są zdefiniowane następująco:

$$\omega_1 = \frac{p_{11}}{p_{12}}, \tag{1.13}$$

$$\omega_2 = \frac{p_{21}}{p_{22}}. \tag{1.14}$$

Iloraz szans jest podstawową miarą opisu stopnia związku zmiennych w tablicy kontyngencji $2 \times 2$, która jest zdefiniowana jako

$$\theta = \frac{\omega_1}{\omega_2} = \frac{\frac{p_{11}}{p_{12}}}{\frac{p_{21}}{p_{22}}} = \frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}}. \tag{1.15}$$

Estymatorem ilorazu szans $\theta$ jest statystyka $\hat{\theta}$ zdefiniowana jako

$$\hat{\theta} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}. \tag{1.16}$$

Iloraz szans $\theta$ przyjmuje zawsze wartości dodatnie, przy czym $\theta = 1$ oznacza, że zmienne są niezależne. Jeśli $1 < \theta < \infty$, to prawdopodobieństwo wystąpienia (sukces) podmiotu w pierwszym wierszu jest większe niż prawdopodobieństwo wystąpienia podmiotu w drugim wierszu. Jeśli natomiast $0 < \theta < 1$, wówczas prawdopodobieństwo wystąpienia podmiotu w pierwszym wierszu jest mniejsze niż prawdopodobieństwo wystąpienia podmiotu w drugim wierszu. Wartość ilorazu szans $\theta$ znacznie przekraczająca 1 oznacza silną zależność pomiędzy zmiennymi. Iloraz szans jest zatem miarą odchylenia od niezależności. Jeśli $1 < \theta < \infty$, wówczas zależność między zmiennymi jest zgodna co do kierunku, jeśli natomiast $0 < \theta < 1$, zależność między zmiennymi jest odwrotna. W miarę wzrostu liczebności próby rozkład ilorazu szans staje się coraz bardziej skośny, w związku z tym wygodnie stosować zamiast ilorazu szans $\theta$ jego logarytm (*log–odds–ratio,* LOR). Logarytm ilorazu szans należy do przedziału $(-\infty, \infty)$, przy czym $\ln(\theta) = 0$ oznacza niezależność zmiennych. W przypadku dużych prób $\ln(\theta)$ ma w przybliżeniu rozkład normalny. Wartość logarytmu ilorazu szans jest miarą symetryczną względem 0, co oznacza, że zamiana wierszy z kolumnami w tablicy kontyngencji powoduje tylko zmianę jego znaku.

W przypadku występowania niewielkiej liczby obserwacji, aby sztucznie zwiększyć liczebność tablicy kontyngencji, miarę tę można skorygować poprzez dodanie do niewielkich liczebności stałej, np. 0,5:

$$\hat{\theta} = \frac{(n_{11} + 0{,}5)(n_{22} + 0{,}5)}{(n_{12} + 0{,}5)(n_{21} + 0{,}5)}. \tag{1.17}$$

Iloraz szans jest miarą wyróżniającą się pożądanymi własnościami, których nie posiadają inne współczynniki wykorzystywane do pomiaru zależności zmiennych nominalnych:

1. Jeśli zmienne w tablicy $2 \times 2$ zostaną zapisane w odwrotnej konfiguracji, tzn. wiersze staną się kolumnami, a kolumny wierszami, wówczas wartość ilorazu szans $\hat{\theta}$ pozostaje niezmienna [Knoke, Burke, 1980; Rudas 1998; Agresti, 2002]
$$\hat{\theta}^* = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \frac{n_{22} \cdot n_{11}}{n_{21} \cdot n_{12}} = \hat{\theta}. \tag{1.18}$$
2. Jeśli każda z liczebności tablicy $n_{hj}$ pomnożona zostanie przez dowolną stałą $c$, iloraz szans dla nowej tablicy kontyngencji jest taki sam, jak dla tablicy pierwotnej
$$\hat{\theta}^* = \frac{c \cdot n_{11} \cdot c \cdot n_{22}}{c \cdot n_{12} \cdot c \cdot n_{21}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \hat{\theta}. \tag{1.19}$$
3. Jeżeli liczebności zmiennej w wierszu pierwszym zostaną pomnożone przez stałą $c$, a liczebności w drugim przez inną stałą $d$, wówczas
$$\hat{\theta}^* = \frac{c \cdot n_{11} \cdot d \cdot n_{22}}{c \cdot n_{12} \cdot d \cdot n_{21}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \hat{\theta}. \tag{1.20}$$

Iloraz szans jest miarą niepodatną na zmiany rozkładów brzegowych [Yule, 1900].

Przy założeniu, że próba jest wystarczająco duża $(n > 25)$ i w tablicy brak jest zerowych liczebności, tzn. wszystkie $n_{hj} > 0$, błąd standardowy logarytmu ilorazu szans $SE(\ln(\hat{\theta}))$ jest zdefiniowany jako [Reynolds, 1977]

$$\sigma = SE(\ln(\hat{\theta})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{1.21}$$

Iloraz szans wykorzystuje się do testowania hipotezy o niezależności zmiennych:

$H_0$: $\theta = 1$,

$H_1$: $\theta \neq 1$.

Dla wystarczająco dużych prób statystyka testowa $z$ ma postać

$$z = \frac{\ln(\hat{\theta})}{SE(\ln(\hat{\theta}))} \tag{1.22}$$

i ma ona rozkład normalny standaryzowany. Jeśli $|u| > u_\alpha$, hipotezę $H_0$ odrzucamy, co świadczy o istnieniu zależności między zmiennymi. Jeśli natomiast $|u| \leqslant u_\alpha$, wówczas brak podstaw do odrzucenia hipotezy zerowej i wnioskujemy, że zmienne są niezależne.

### Współczynnik korelacji $Q$ Yule'a

Współczynnik korelacji (*correlation coefficient*) [Yule, 1900] jest miarą zależności zbliżoną do współczynnika korelacji Pearsona, zdefiniowaną jako

$$Q = \frac{p_{11} \cdot p_{22} - p_{12} \cdot p_{21}}{p_{11} \cdot p_{22} + p_{12} \cdot p_{21}} = \frac{\frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}} - 1}{\frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}} + 1} = \frac{\theta - 1}{\theta + 1}. \tag{1.23}$$

Estymatorem współczynnika $Q$ Yule'a jest statystyka

$$\hat{Q} = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{n_{11} \cdot n_{22} + n_{12} \cdot n_{21}} = \frac{\hat{\theta} - 1}{\hat{\theta} + 1}. \tag{1.24}$$

Współczynnik ten przyjmuje wartości z przedziału $[-1, 1]$, gdzie 0 oznacza niezależność zmiennych [Reynolds, 1977].

### Współczynniki koligacji $Y$ Yule'a

Kolejną miarą przeznaczoną do badania zależności zmiennych nominalnych w tablicy kontyngencji, wykorzystującą iloraz szans, jest współczynnik koligacji (*coefficient of colligation*) [Yule, 1912]

$$Y = \frac{\sqrt{p_{11} \cdot p_{22}} - \sqrt{p_{12} \cdot p_{21}}}{\sqrt{p_{11} \cdot p_{22}} + \sqrt{p_{12} \cdot p_{21}}} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1}. \tag{1.25}$$

Estymatorem współczynnika koligacji $Y$ jest statystyka

$$\hat{Y} = \frac{\sqrt{n_{11} \cdot n_{22}} - \sqrt{n_{12} \cdot n_{21}}}{\sqrt{n_{11} \cdot n_{22}} + \sqrt{n_{12} \cdot n_{21}}} = \frac{\sqrt{\hat{\theta}} - 1}{\sqrt{\hat{\theta}} + 1}. \tag{1.26}$$

Własności oraz zakres występowania wartości współczynników koligacji $Q$ i $Y$ są identyczne. W przypadku tablicy $2 \times 2$, w której prawdopodobieństwo wystąpienia każdej z kategorii badanych zmiennych wynosi 0,5, wartość współczynnika koligacji $Y$ mówi o różnicy w prawdopodobieństwach będących na przekątnej oraz poza nią. Współczynnik ten przyjmuje wartości z przedziału $[-1, 1]$, gdzie 0 oznacza niezależność zmiennych [Reynolds, 1977].

### Współczynnik korelacji dla tablicy $2 \times 2$

Miernikiem korelacji przeznaczonym do badania zależności między dwiema zmiennymi o charakterze metrycznym jest współczynnik korelacji liniowej Pearsona. W literaturze istnieje również jego odpowiednik, który jest stosowany w przypadku zmiennych niemetrycznych [Bishop, Fienberg, Holland, 1975; Reynolds, 1977]:

$$\rho = \frac{p_{11} \cdot p_{22} - p_{12} \cdot p_{21}}{\sqrt{p_{1\bullet} \cdot p_{2\bullet} \cdot p_{\bullet 1} \cdot p_{\bullet 2}}}. \tag{1.27}$$