ZALEŻNOŚĆ STATYSTYCZNA UPORZĄDKOWANIA KATEGORII OD ZMIENNYCH NOMINALNYCH

Grzegorz Lissowski

Zmienne typu „wybór” (1)

Charakter danych

Często zamiast badania pełnej preferencji wobec zbioru opcji lub kategorii ogranicza się do wskazania przez osobę badaną jednej z nich. Niekiedy taki wybór jest ograniczony przez sytuację np. w czasie wyborów, gdy wyborca wskazuje jednego kandydata lub jedną partię (taki wybór nie musi zresztą oznaczać wyboru najwyżej ocenianej opcji). Podobnie jest w innych sytuacjach.

O zbiorze opcji lub kategorii możemy zakładać, że jest to zbiór uporządkowany przez natężenie jakieś własności, bądź nawet że są one mierzone na mocniejszej skali (np. dochody). Nie będziemy jednak przyjmować takich założeń, a jedynie uwzględniać uporządkowanie kategorii wyznaczone przez częstości wyborów dokonywanych przez osoby badane w całej zbiorowości lub w poszczególnych podzbiorowościach: od najczęściej do najrzadziej wybieranej opcji lub kategorii.

Zmienne typu „wybór” (2)

Sposób analizy

Do opisu tego typu zmiennych, które mają postać rozkładów częstości, bez uporządkowania kategorii, można stosować jedynie ograniczone metody:

- do opisu **poziomu wartości** – uwzględniające jedynie częstości wyborów, głównie modalną (dominantę),
- do opisu **rozproszenia**, a dokładniej **różnorodności**, można stosować więcej miar, np. różnorodność klasyfikacji, entropię itp.

Oprócz wymienionych wyżej istnieje wiele innych miar. Wiele z nich przedstawiono w książce Ph.B. Coultera „Measurement Inequality. A Methodological Handbook” (1989). Jedną z nich jest Corrado Gini średnia różnica (*mean difference* 1912) i współczynnik koncentracji (*concentration ratio* 1914).

Zmienne typu „wybór” (3)

Sposób analizy (c.d.)

Skrajne sytuacje w kategoriach ***różnorodności***

minimalna różnorodność	***maksymalna różnorodność***
rozkład jednopunktowy	*rozkład równomierny*

Skrajne sytuacje w kategoriach ***nierówności***

równość	***koncentracja***
wszystkie jednostki mają równe wartości	*tylko jedna jednostka ma wartość różną (i większą) od 0*

Oceny różnorodności i nierówności różnią się wyraźnie polaryzacją, tj. sposobem ukierunkowania. Maksymalna różnorodność odpowiada minimalnej nierówności i odwrotnie – minimalna różnorodność odpowiada maksymalnej nierówności.

Analiza różnorodności zakłada rozkład częstości zmiennej nominalnej, a nierówności - zmiennej stosunkowej.

Zmienne typu „wybór” (4)

Ocena zależności statystycznej zmiennej typu „wybór” od zmiennej nominalnej

Analiza zależności statystycznej zmiennej typu „wyboru” od innych zmiennych będzie ograniczona do rozkładów dwuwymiarowych: zmiennej „wybór” i zmiennej nominalnej, ale może być rozszerzona na większa liczbę zmiennych, mierzonych na mocniejszych skalach. Ponieważ nie zakłada się uporządkowania kategorii (a tym bardziej pomiaru na mocniejszej skali) wydaje się, że analiza zależności musi być ograniczona do zależności stochastycznej lub zależności modalnych.

Tak jednak nie jest.

Uwaga historyczna - problem pomiaru zależności statystycznych

„*Definicja niezależności, z jaką mamy do czynienia w teorii korelacji, może być również sformułowana w inny sposób. Można zaproponować szereg różnych sformułowań. W pewnych warunkach pewne z nich mogą oddać cenne usługi zakładając, że odpowiednie pojęcie jest dokładnie zdefiniowane i odróżnione od innych konkurencyjnych definicji niezależności. Dalej, pożądane jest wprowadzenie różnych technicznych terminów dla różnych definicji niezależności. (…) Przedstawiona wyżej definicja niezależności stochastycznej jest jednak najlepszą podstawą dla ogólnej teorii korelacji statystycznej. Jest ona najmocniejszą ze wszystkich formalnych definicji niezależności. Jeżeli, zgodnie z tą definicją, dwie zmienne losowe są wzajemnie niezależne, to są one również niezależne w sensie wszystkich innych formalnych definicji tego pojęcia*.” (s. 41).

A.A. Tschuprov. 1939. “Principles of the Mathematical Theory of Correlation”. London: W. Hodge.

Nowy rodzaj zależności statystycznej (1)

Niezależność statystyczna zmiennej X (typu „wybór”) od zmiennej Y (nominalnej) polega na tym, że takie same są uporządkowania częstości wyborów kategorii zmiennej X we wszystkich grupach badanych osób, wyróżnionych ze względu na zmienną Y.

Definicja 1.

Zmienna X jest ***niezależna statystycznie*** w powyższym sensie od zmiennej Y wtedy, gdy

$$\forall y_g \in Y, \forall x_i, x_k \in X:$$

$$P\left(X = x_i \middle| Y = y_g\right) < P\left(X = x_k \middle| Y = y_g\right) \Leftrightarrow P\left(X = x_i\right) < P\left(X = x_k\right)$$

$$P\left(X = x_i \middle| Y = y_g\right) = P\left(X = x_k \middle| Y = y_g\right) \Leftrightarrow P\left(X = x_i\right) = P\left(X = x_k\right)$$

$$P\left(X = x_i \middle| Y = y_g\right) > P\left(X = x_k \middle| Y = y_g\right) \Leftrightarrow P\left(X = x_i\right) > P\left(X = x_k\right)$$

Nowy rodzaj zależności statystycznej (2)

Maksymalna zależność statystyczna zmiennej X (typu „wybór”) od zmiennej Y (nominalnej) oznacza (jak zawsze), że wszystkie osoby w każdej z wyróżnionych grup ze względu na zmienną Y wybierają tę samą kategorię zmiennej X.

Definicja 2.

$$\forall y_s \in Y, \exists x_i \in X : P\left(X = x_i \middle| Y = y_s\right) = 1$$

Zaproponowana będzie ***nowa miara zależności statystycznej*** tego typu oraz jej uzasadnienie.

Miary MD i MDA i ich własności (1)

Miarę MD, nazywaną średnią różnicą (*mean difference*) zaproponował w 1912 r. Corrado Gini jako miarę rozproszenia. Jest ona ściśle związana, po zrelatywizowaniu do wartości średniej zmiennej X, z jego współczynnikiem koncentracji (1914) i krzywą Lorenza. Gini zakładał, że zmienna X jest mierzona na skali stosunkowej.

$$MD(X) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{n} |x_i - x_k|$$

Niech $p_1 \geq p_2 \geq p_3 \geq \ldots \geq p_{n-1} \geq p_n$.

Macierz D reprezentującą wartości bezwzględne różnic między wszystkimi parami częstości można zapisać jako:

$$D = \begin{bmatrix} |p_1 - p_1| & |p_1 - p_2| & |p_1 - p_3| & \ldots & |p_1 - p_n| \\ |p_2 - p_1| & |p_2 - p_2| & |p_2 - p_3| & \ldots & |p_2 - p_n| \\ |p_3 - p_1| & |p_3 - p_2| & |p_3 - p_3| & \ldots & |p_3 - p_n| \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ |p_n - p_1| & |p_n - p_2| & |p_n - p_3| & \ldots & |p_n - p_n| \end{bmatrix}$$

Miary MD i MDA i ich własności (2)

Wygodniej będzie przedstawić macierz D w postaci równoważnej.

$$D=\begin{bmatrix} 0 & (p_1-p_2) & (p_1-p_3) & \dots & (p_1-p_n) \\ (p_1-p_2) & 0 & (p_2-p_3) & \dots & (p_2-p_n) \\ (p_1-p_3) & (p_2-p_3) & 0 & \dots & (p_3-p_n) \\ \dots & \dots & \dots & \dots & \dots \\ (p_1-p_n) & (p_2-p_n) & (p_3-p_n) & \dots & 0 \end{bmatrix}$$

Oznaczmy przez

$$d_{ik}=\left|p_i-p_k\right|$$

wówczas

$$\Delta=\sum_{i=1}^{n}\sum_{k=1}^{n}d_{ik}=\sum_{i=1}^{n}\sum_{k=1}^{n}\left|p_i-p_k\right|=2\sum_{i=1}^{n}(n-i)p_i-2\sum_{i=1}^{n}(i-1)p_i=$$

$$=2\sum_{i=1}^{n}(n-2i+1)p_i=2\sum_{i=1}^{n}(n+1)p_i-4\sum_{i=1}^{n}ip_i=2(n+1)\sum_{i=1}^{n}p_i-4\sum_{i=1}^{n}ip_i=$$

$$=2(n+1)-4\sum_{i=1}^{n}ip_i$$

Miary MD i MDA i ich własności (3)

Jeżeli rozkład jest równomierny, tj. wszystkie częstości są równe,

$$p_1 = p_2 = p_3 = \ldots = p_{n-1} = p_n = 1/n, \quad \text{to } \Delta=0.$$

Jeżeli rozkład jest jednopunktowy, tj. występuje w nim skrajna koncentracja,

$$p_1 = 1,\ p_2 = p_3 = \ldots = p_{n-1} = p_n = 0, \quad \text{to } \Delta=2(n-1).$$

Maksymalna wartość miary MD w wypadku skrajnej koncentracji jest równa $2/n$, gdyż

$$\frac{2(n-1)}{n(n-1)} = \frac{2}{n}$$

Miara MD unormowana do przedziału [0, 1] ma postać:

$$\frac{\dfrac{2(n+1)-4\sum_{i=1}^{n} ip_i}{n(n-1)}}{\dfrac{2}{n}} = \frac{n+1-2\sum_{i=1}^{n} ip_i}{n-1}$$

Miary MD i MDA i ich własności (4)

Miara MD przyjmuje wartość 0 – w wypadku rozkładu równomiernego, a największą wartość 1 – w wypadku rozkładu, w którym jedna z kategorii występuje z częstością 1. Miara ta jest równa skorygowanej mierze koncentracji (Ray i Singer 1973).

Przyjęcie przeciwnego unormowania, tj. przypisanie wartości 0 rozkładowi jednopunktowemu, a wartości 1 – rozkładowi równomiernemu, czyli

$$1-\frac{n+1-2\sum_{i=1}^{n} ip_i}{n-1}=\frac{2\left(\sum_{i=1}^{n} ip_i-1\right)}{n-1}=$$
$$=1-\frac{\sum_{i=1}^{n-1}\sum_{k=i+1}^{k}|p_i-p_k|}{n-1}=MDA$$

prowadzi do miary różnorodności MDA nazywanej, analogiem średniej różnicy (*mean difference analog*), analizowanej przez Allena R. Wilcoxa (1973). Przyjmiemy miarę MDA jako miarę różnorodności wykorzystywaną w dalszej analizie.

Nowa miara zależności statystycznej (1)

Nowa miara zależności statystycznej zmiennej X od zmiennej Y będzie skonstruowana w ten sam sposób jak wszystkie miary zależności

$$\omega_{X|Y} = \frac{MDA(X) - E\left[MDA(X|Y)\right]}{MDA(X)}$$

Wprowadzimy następujące oznaczenia:

r_i ranga częstości kategorii x_i w uporządkowaniu nierosnącym w całej zbiorowości,

r_{ig} ranga warunkowej częstości kategorii x_i w uporządkowaniu nierosnącym w podzbiorowości $\{Y=y_g\}$.

Przy tych oznaczeniach, wykorzystując jedną z postaci miary MDA, mamy

$$MDA(X) = \frac{2\left[\sum_{i=1}^{n} r_i P(X = x_i) - 1\right]}{n-1}$$

Nowa miara zależności statystycznej (2)

$$MDA\left(X \middle| Y = y_g\right) = \frac{2\left[\sum_{i=1}^{n} r_{ig} P(X = x_i \middle| Y = y_g) - 1\right]}{n-1}$$

$$\omega_{X|Y} = \frac{MDA(X) - E\left[MDA(X|Y)\right]}{MDA(X)} =$$

$$= \frac{\frac{2\left[\sum_{i=1}^{n} r_i P(X = x_i) - 1\right]}{n-1} - \sum_{g=1}^{G} P(Y = y_g) \frac{2\left[\sum_{i=1}^{n} r_{ig} P(X = x_i \middle| Y = y_g) - 1\right]}{n-1}}{\frac{2\left[\sum_{i=1}^{n} r_i P(X = x_i) - 1\right]}{n-1}} =$$

$$= \frac{\sum_{i=1}^{n} r_i P(X = x_i) - \sum_{g=1}^{G} P(Y = y_g) \sum_{i=1}^{n} r_{ig} P(X = x_i \middle| Y = y_g)}{\sum_{i=1}^{n} r_i P(X = x_i)}$$

Miara zależności przyjmuje wartość 0, gdy zmienna X jest niezależna statystycznie od zmiennej Y w sensie definicji 1 oraz wartość 1, gdy jest maksymalnie zależna statystycznie.

Własności miary zależności statystycznej (1)

Twierdzenie 1.

Jeżeli zmienna X jest niezależna statystycznie od zmiennej Y w sensie definicji 1, to $\omega_{X|Y} = 0$.

Dowód.

$$\omega_{X|Y} = \frac{\sum_{i=1}^{n} r_i P(X = x_i) - \sum_{g=1}^{G} P(Y = y_g) \sum_{i=1}^{n} r_{ig} P(X = x_i \mid Y = y_g)}{\sum_{i=1}^{n} r_i P(X = x_i)}$$

Jeżeli - zgodnie z definicją 1 - dla każdego $y_g \in Y$: $r_{ig}=r_i$, to odjemnik powyższego wyrażenia upraszcza się do MDA(X).

$$\sum_{g=1}^{G} P(Y = y_g) \sum_{i=1}^{n} r_{ig} P(X = x_i \mid Y = y_g) =$$

$$= \sum_{i=1}^{n} r_i \sum_{g=1}^{G} P(X = x_i \wedge Y = y_g) - \sum_{g=1}^{G} P(Y = y_g) = \sum_{i=1}^{n} r_i P(X = x_i) - 1 = MDA(X)$$

W konsekwencji, $\omega_{X|Y} = 0$.

Własności miary zależności statystycznej (2)

Ilustracja.

Y / X	y_1	y_2	y_3	Razem
x_1	0,30	0,40	0,25	0,300
x_2	0,25	0,25	0,21	0,234
x_3	0,20	0,20	0,20	0,200
x_4	0,15	0,15	0,19	0,166
x_5	0,10	0,00	0,15	0,100
Razem	1,00	1,00	1,00	1,000
$P(Y=y_k)$	0,40	0,20	0,40	
MDA	0,75	0,55	0,89	0,766

$$E[MDA(X|Y)] = 0,40*0,75 + 0,20*0,55 + 0,40*0,89 = 0,766 = MDA(X)$$

$$\omega_{X|Y} = 0.$$

Niezależność statystyczna zmiennej X od zmiennej Y w sensie definicji 1 jest warunkiem wystarczającym na to, aby $\omega_{X|Y} = 0$. Nie jest jednak warunkiem koniecznym. Można wskazać szerszą klasę sytuacji, w których miara zależności przyjmuje wartość 0. Określa ją następujące twierdzenie.

Własności miary zależności statystycznej (3)

Twierdzenie 2.

Jeżeli dla każdej kategorii x_i zmiennej X zachodzi równość między rangą tej kategorii w całej zbiorowości a średnią rang tej kategorii w podzbiorowościach wyróżnionych ze względu na zmienną Y, to miara zależności przyjmuje wartość 0.

Dowód.

Licznik miary zależności w zapisie uproszczonym ma następującą postać.

$$\sum_{i=1}^{n} r_i P(X = x_i) - \sum_{g=1}^{G} P(Y = y_g) \sum_{i=1}^{n} r_{ig} P(X = x_i \,|\, Y = y_g) =$$

$$= \sum_{i=1}^{n} r_i P(X = x_i) - \sum_{i=1}^{n} \sum_{g=1}^{G} r_{ig} P(X = x_i \wedge Y = y_g) =$$

$$= \sum_{i=1}^{n} r_i P(X = x_i) - \sum_{i=1}^{n} \sum_{g=1}^{G} r_{ig} P(Y = y_g \,|\, X = x_i) P(X = x_i) =$$

$$= \sum P(X = x_i) \left[r_i - \sum_{g=1}^{G} r_{ig} P(Y = y_g \,|\, X = x_i) \right]$$

Wyrażenie w nawiasie kwadratowym jest różnicą między rangą kategorii x_i w całej zbiorowości a średnią rang tej kategorii w podzbiorowościach wyróżnionych ze względu na zmienną Y. Jeżeli dla każdej kategorii $x_i \in X$ jest ono równe 0, to $\omega_{X|Y} = 0$.

Własności miary zależności statystycznej (4)

Druga własność miary zależności , tzn. że przyjmuje ona wartość 1, gdy zmienna X jest maksymalnie zależna statystycznie od zmiennej Y w sensie definicji 2, jest oczywista. Wystarczy zauważyć, że w tym wypadku wartość E[MDA(X|Y)] musi być równa 0. Tak może być jedynie wtedy, gdy każdy warunkowy rozkład zmiennej X ze względu na zmienną Y jest zerojedynkowy i w konsekwencji dla każdego $y_g \in Y$: $MDA(X|Y=y_g)=0$.

Najważniejsza jest jednak interpretacja miary zależności $\omega_{X|Y}$, która stanowi jej podstawowe uzasadnienie.

Inspiracja do interpretacji miary zależności ma dwa źródła. Pierwszym jest miara proporcjonalnej predykcji (Goodman, L.A. i Kruskal, W,H. 1954. *Measures of association for cross classifications*. “Journal of the American Statistical Association” 49: 732-764, zwłaszcza strony 759-760), a drugim - interpretacja współczynnika koncentracji Gini w kategoriach gry (Pyatt, G. 1976. *On the interpretation and disaggregation of Gini coefficient*. “The Economic Journal” 86: 243-255).

Interpretacja miary zależności $\omega_{X|Y}$ (1)

Inspiracja 1.

Proporcjonalna predykcja polega na „odgadywaniu” wartości zmiennej X i w tym celu przypisuje się losowo obiektom w zbiorowości różne wartości x_i z prawdopodobieństwami równymi częstościom $P(X=x_i)$, z jakimi występują w tej zbiorowości. Częstość popełniania błędu przy tym przewidywaniu jest określona za pomocą parametru nazywanego różnorodnością klasyfikacji.

$$K(X) = 1 - \sum_{x_i \in X} P(X = x_i)^2$$

Jeżeli przewidywanie dokonywane jest na podstawie informacji, że obiekt ma wartość zmiennej Y równą y_g, to warunkowa częstość popełniania błędu jest równa

$$K(X \mid Y = y_g) = 1 - \sum_{x_i \in X} P(X = x_i \mid Y = y_g)^2$$

Miara zależności statystycznej zmiennej X od zmiennej Y w sensie proporcjonalnej predykcji ma postać

$$\tau_X = \frac{K(X) - E[K(X \mid Y)]}{K(X)}$$

Interpretacja miary zależności $\omega_{X|Y}$ (2)

Inspiracja 2.

Graham Pyatt sformułował interesującą interpretację współczynnika koncentracji Gini w postaci gry czy też eksperymentu. Dla każdej jednostki o określonym dochodzie wybiera się losowo dochód z prawdopodobieństwami określonymi w rozkładzie dochodów w danej populacji. Jeżeli aktualny dochód jednostki jest niższy od wybranego losowo dochodu, to otrzymuje ona dochód wylosowany, a jeżeli jest wyższy od wylosowanego, to zachowuje ona swój aktualny dochód. Współczynnik koncentracji Gini ma interpretację jako wartość oczekiwana zysku w tej grze, tj. średnia wielkość zwiększenia dochodów.

$$G = \frac{\left(\frac{1}{n^2}\right)\sum_{i=1}^{n}\sum_{k=1}^{n}\max(0,\, x_i - x_k)}{\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i}$$

gdzie wartości zmiennej X oznaczają dochody jednostek.

Interpretacja miary zależności $\omega_{X|Y}$ (3)

Kategorie zmiennej X są uporządkowane ze względu na częstość ich wyborów w danej zbiorowości: od kategorii wybieranej najczęściej x_1, do kategorii najrzadziej wybieranej x_n.

Częstości skumulowane do kategorii x_i będziemy oznaczali przez $F_X(x_i)$, tj.

$$F_X(x_i) = \sum_{k=1}^{i} P(X = x_k)$$

Analog średniej różnicy MDA można przedstawić w innej postaci.

$$MDA(X) = \frac{2\left(\sum_{i=1}^{n} iP(X = x_i) - 1\right)}{n-1} =$$

$$= \frac{2\left(\sum_{i=1}^{n} F_X(x_i) - 1\right)}{n-1} = \frac{2\left(\sum_{i=1}^{n-1} F_X(x_i)\right)}{n-1}$$

Interpretacja miary zależności $\omega_{X|Y}$ (4)

Porównajmy jednostkę ***a*** należącą do kategorii x_i z jednostką ***b*** należącą do kategorii x_k.

Niech ocena tego porównania będzie równa 0, gdy jednostka ***a*** należy do kategorii wybieranej częściej lub równie często jak kategoria do której należy jednostka ***b*** oraz równa 1, gdy jednostka ***a*** należy do kategorii wybieranej rzadziej od kategorii do której należy jednostka ***b***.

$$l(a,b) = \begin{cases} 0 & \text{gdy } X(a) = x_i \wedge X(b) = x_k \wedge (i \le k) \\ 1 & \text{gdy } X(a) = x_i \wedge X(b) = x_k \wedge (i > k) \end{cases}$$

Ocena ta jest więc oceną braku zgodności z uporządkowaniem kategorii wyznaczonym przez częstości ich wyborów.

Porównując całe kategorie x_i oraz x_k otrzymujemy

$$l(x_i, x_k) = \begin{cases} 0 & \text{gdy } (i \le k) \\ P(X = x_k) & \text{gdy } (i > k) \end{cases}$$

Sumując te oceny po wszystkich kategoriach x_k otrzymujemy

$$\sum_{k=1}^{n} l(x_i, x_k) = \sum_{k=1}^{i} P(X = x_k) = F_X(x_i)$$

Zatem

$$MDA(X) = \frac{2\left(\sum_{i=2}^{n} F_X(x_i)\right)}{n-1}$$

Interpretacja miary zależności $\omega_{X|Y}$ (5)

Analog średniej różnicy MDA jest funkcją oceny braku zgodności w rozkładzie zmiennej X z uporządkowaniem kategorii wyznaczonym przez częstości ich wyborów.

Analogicznie dla rozkładu warunkowego zmiennej X pod warunkiem, że zmienna Y przyjmuje wartość y_g.

$$MDA(X|Y=y_g)=\frac{2\left(\sum_{i=2}^{n}F_{X|Y=y_g}(x_i)\right)}{n-1}$$

Zatem

$$\omega_{X|Y}=\frac{MDA(X)-E\left[MDA(X|Y)\right]}{MDA(X)}=\frac{\frac{2\left(\sum_{i=2}^{n}F_X(x_i)\right)}{n-1}-\sum_{g=1}^{G}P(Y-y_g)\frac{2\left(\sum_{i=2}^{n}F_{X|Y=y_g}(x_i)\right)}{n-1}}{\frac{2\left(\sum_{i=2}^{n}F_X(x_i)\right)}{n-1}}=$$

$$=\frac{\sum_{i=2}^{n}F_X(x_i)-\sum_{g=1}^{G}P(Y=y_g)\sum_{i=2}^{n}F_{X|Y=y_g}(x_i)}{\sum_{i=2}^{n}F_X(x_i)}$$

Wartość tej miary oznacza stopień redukcji oceny braku zgodności w rozkładzie zmiennej X z uporządkowaniem kategorii wyznaczonym przez częstości ich wyborów w wyniku tego, że wykorzystana została informacja o wartości zmiennej Y.

Bibliografia

Allison, Paul D. 1978. *Measures of inequality*. “American Sociological Review” 43: 865-80.

Allison, Paul D. 1979. *Reply to Jasso*. “American Sociological Review” 44: 870-872.

Ceriani, Lidia; Verme Paolo. 2012. *The origins of the Gini index: extracts from «Variabilità and Mutabilità» by Corrado Gini*. “Journal of Economic Inequality” 10: 421-443.

Coulter, Philip B. 1989. *Measurement Inequality. A Methodological Handbook*. Boulder: Westview Press.

David, H.A. 1968. *Gini’s mean difference rediscovered*. “Biometrika” 55: 573-575.

Goodman, Leo A., Kruskal, William H. 1954. *Measures of association for cross classifications*. “Journal of the American Statistical Association” 49: 732-764.

Jasso, Guillermina. 1979. *On Gini's mean difference and Gini's index of concentration*. “American Sociological Review” 44: 867-870.

Kendall, Maurice G.; Stuart, Alan. 1958. *The Advanced Theory of Statistics*. Vol. I. New York: Hafner Publishing Co.

Pyatt, Graham. 1976. *On the interpretation and disaggregation of Gini coefficient*. “The Economic Journal” 86: 243-255.

Ray, James L.; Singer, David. 1973. *Measuring the concentration of power in the international system*. “Sociological Methods and Research” 1: 403-437.

Santos, Jesús Busulto; Guerrero, J. Javier Busto. 2010. *Gini’s concentration ratio (1908-1914)*. “Electronic Journal for History of Probability and Statistics” 6: 1-42.

Tschuprov, Alexander A. 1939. *Principles of the Mathematical Theory of Correlation*. London: W. Hodge. Pierwsze wydanie w języku niemieckim w 1925 r.

Yitzhaki, Shlomo. 2013. *More Than a Dozen Alternative Ways of Spelling Gini*. W: Yitzhaki, Shlomo; Schechtman, Edna (red.) *The Gini Methodology: A Primer on a Statistical Methodology*. Chapter 2. Springer Series in Statistics 272, New York: Springer.

Wilcox, Allen R. 1973*. Indices of qualitative variation and political measurement*. “Western Political Quarterly” 26: 325-343.

Wyjaśnienie

$$F_X(x_i) = \sum_{k=i}^{n} P(X = x_k)$$

i	**ip_i**								
1	$1p_1$		p_1			…			$F_X(x_1)$
2	$2p_2$		p_2	p_2		…			$F_X(x_2)$
3	$3p_3$		p_3	p_3	p_3	…			$F_X(x_3)$
…	…		…	…	…	…			
n-1	$(n-1)p_{n-1}$		p_{n-1}	p_{n-1}	p_{n-1}	…	p_{n-1}		$F_X(x_{n-1})$
n	np_n		p_n	p_n	p_n	…	p_n	p_n	$F_X(x_n)$