

NOWA MIARA ZALEŻNOŚCI STATYSTYCZNEJ WYBORÓW OD GRUPY¹

Grzegorz Lissowski*
Uniwersytet Warszawski

Streszczenie: Artykuł przedstawia propozycję nowego typu zależności statystycznej. Niezależność statystyczna tego rodzaju oznacza identyczność uporządkowania częstości wyborów kategorii we wszystkich grupach. Zgodność między wyborami będzie oceniana za pomocą miary różnorodności. Określona będzie nowa miara zależności statystycznej, jej własności i interpretacja.

Słowa kluczowe: zależność statystyczna, różnorodność, zgodność między wyborami.

NEW MEASURE OF STATISTICAL DEPENDENCE CHOICES ON GROUP

Abstract: The article proposes a new type of statistical dependence. Statistical independence of this type means the identity of the frequency order of category choices in all groups. The compatibility between choices will be assessed by means of a diversity measure. A new measure of statistical dependence, its properties and interpretation will be specified.

Keywords: statistical dependence, diversity, compatibility between choices.

1. WPROWADZENIE

Często zadawane są pytania o to, czy wybory dokonywane przez osoby należące do różnych grup są podobne, czy też różnią się, a także o to, jak duże są te ewentualne różnice. Tego rodzaju pytania są formułowane nie tylko wobec wyników wyborów powszechnych. W badaniach preferencji społecznych rzadko możliwe jest badanie

¹ Wczesna wersja tego artykułu była prezentowana na konferencji „Metodologiczne Inspiracje 2014. Badania Ilościowe w Naukach Społecznych – Wyzwania i Problemy” (Jabłonna, 24-26 września 2014), którą zorganizował Instytut Filozofii i Socjologii PAN.

* Grzegorz Lissowski, Instytut Socjologii, Uniwersytet Warszawski, ul. Karowa 18, 00-927 Warszawa, e-mail: gliss@is.uw.edu.pl

pełnej preferencji wobec zbioru kategorii i często ogranicza się do wskazania przez osobę badaną jednej z nich, a rozkład wskazań osób badanych traktuje się jako ocenę preferencji wobec tego zbioru.

W odniesieniu do zbioru kategorii, spośród których dokonuje się wyboru, można by zakładać, że jest to zbiór uporządkowany przez natężenie jakiejś własności bądź nawet że kategorie są mierzone na mocniejszej skali pomiarowej (np. dochody). W tym artykule nie przyjmuje się jednak tak mocnych założeń, lecz jedynie analizuje się uporządkowania tego zbioru – przez całą badaną zbiorowość lub przez poszczególne grupy (tj. podzbiory tej zbiorowości) – wyznaczone na podstawie częstości wyborów dokonanych przez osoby badane: od najczęściej do najrzadziej wybieranej kategorii.

Ocena zależności wyborów od grupy będzie szczególnym typem zależności statystycznej. Sytuacja pełnej niezależności będzie rozumiana jako identyczność uporządkowań częstości wyborów kategorii w różnych grupach, a nie jako identyczność rozkładów tych wyborów. Inaczej mówiąc, niezależność uporządkowania częstości wyborów od grupy będzie oznaczała pełną zgodność wyborów. Im większe będą różnice między uporządkowaniami tych wyborów w różnych grupach, tym mniejsza będzie ocena zgodności. Z maksymalną niezgodnością będziemy mieli do czynienia wtedy, gdy w każdej grupie wybrana będzie tylko jedna kategoria, przy czym wybory dokonane w różnych grupach będą różne. W tym wypadku dokonywane wybory będą w pełni zależne od grupy. Pojęcia te zostaną formalnie określone w sekcji 2.

Ocena zależności wyborów od grupy, a więc zgodności lub niezgodności tych wyborów, będzie wymagać wykorzystania miary różnorodności wyborów – zarówno w grupie, jak i między nimi. Taka miara różnorodności, wyprowadzona z klasycznej propozycji Corrado Giniego (1912), zostanie przedstawiona w sekcji 3. Podane zostaną pewne jej własności, które nie są znane w literaturze.

Proponowana miara zależności wyborów od grupy będzie skonstruowana w taki sam sposób, jak wszystkie miary zależności statystycznej. Będzie to jednak nowy typ zależności statystycznej, nieznan dotąd w literaturze. Temu zagadnieniu będzie poświęcona sekcja 4. Z kolei interpretacja wartości tej miary zostanie podana w sekcji 5. Źródłami tej interpretacji będą: miara proporcjonalnej predykcji (Goodman i Kruskal, 1954) oraz interpretacja współczynnika koncentracji Giniego w kategoriach gry (Pyatt, 1976). Zastosowania proponowanych miar do analizy wyników głosowania w wyborach prezydenckich będą przedstawione w sekcji 6.

2. ZGODNOŚĆ MIĘDZY WYBORAMI W RÓŻNYCH GRUPACH

Oznaczmy przez $K = \{k_1, k_2, \dots, k_n\}$ zbiór n kategorii, spośród których dokonuje się wyboru, natomiast przez $G = \{g_1, g_2, \dots, g_m\}$ zbiór m grup, na które podzielone są osoby dokonujące wyboru.

Niezależność uporządkowań częstości wyborów kategorii od grupy będzie oznaczać sytuację, w której uporządkowania częstości wyborów kategorii we wszystkich grupach są takie same, a więc identyczne jak w całej zbiorowości.

$$\forall g_h \in G, \forall k_i, k_j \in K :$$

$$P(K = k_i | G = g_h) < P(K = k_j | G = g_h) \Leftrightarrow P(K = k_i) < P(K = k_j)$$

$$P(K = k_i | G = g_h) = P(K = k_j | G = g_h) \Leftrightarrow P(K = k_i) = P(K = k_j)$$

$$P(K = k_i | G = g_h) > P(K = k_j | G = g_h) \Leftrightarrow P(K = k_i) > P(K = k_j)$$

Takie określenie niezależności jest nowe, nieznane w literaturze. Ma ono jednak dobre uzasadnienie. Abstrahuje bowiem od pewnych nieistotnych różnic między częstościami, a zamiast tego zwraca uwagę na zasadnicze podobieństwo między wyborami dokonanymi w poszczególnych grupach.

Warto przytoczyć ważną opinię Aleksandra A. Tschuprova sformułowaną w pracy *Zasady matematycznej teorii korelacji*. Jego istotnym wkładem było wyraźne odróżnienie niezależności stochastycznej między zmiennymi losowymi od niezależności korelacyjnej, którą zajmowała się Szkoła Angielska. Tschuprov pisał m.in.: „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.” (Tschuprov 1939: 41).

Przypomnijmy: niezależność stochastyczna oznacza, że rozkłady warunkowe jednej zmiennej są takie same we wszystkich grupach wyróżnionych ze względu na drugą zmienną. Niezależność korelacyjna oznacza natomiast, że warunkowe średnie jednej zmiennej są takie same we wszystkich grupach wyróżnionych ze względu na

drugą zmienną. Jest to więc warunek słabszy od niezależności stochastycznej. Nieskorelowanie liniowe jest jeszcze słabszym warunkiem niż niezależność korelacyjna.

Niezależność uporządkowań częstości wyborów od grupy jest warunkiem słabszym od niezależności stochastycznej. Wymaga jedynie identyczności uporządkowania częstości, a nie całych rozkładów warunkowych. Nie można jej porównywać z niezależnością korelacyjną, gdyż wymaga tylko pomiaru kategorii na skali nominalnej, co nie pozwala na obliczanie średnich.

Maksymalna zależność wyborów od grupy będzie oznaczać sytuację, w której wszystkie osoby w każdej z wyróżnionych grup wybierają tylko jedną, tę samą kategorię.

$$\forall g_h \in G, \exists k_i \in K : P(K = k_i | G = g_h) = 1$$

Ponadto zakłada się, że wybory dokonane w różnych grupach nie są jednakowe.

$$\exists g_h \neq g_i \in G, \exists k_i \in K : P(K = k_i | G = g_h) = 1 \wedge P(K = k_i | G = g_i) = 0$$

Dodatkowe założenie ma wykluczać identyczność wyborów we wszystkich grupach, która oznaczałaby niezależność szczególnie rodzaju uporządkowań częstości wyborów od grupy.

Maksymalna zależność wyborów od grupy jest własnością podobną do maksymalnej zależności stochastycznej, a także do innych typów maksymalnych zależności statystycznych związanych z regresjami pierwszego rodzaju.

Niezależność uporządkowań częstości wyborów od grupy i maksymalna zależność wyborów od grupy są jedynie dwiema skrajnymi sytuacjami związku między wyborami a przynależnością do grupy. Dla oceny zgodności wyborów w różnych grupach, a więc dla pomiaru siły zależności wyborów od grupy, niezbędne jest zastosowanie szczególnej miary różnorodności wyborów, której będzie poświęcona następująca sekcja.

3. MIARA RÓŻNORODNOŚCI WYBORÓW

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

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

Istnieją różne wersje tej miary, w których uwzględnia się różnice między wartościami zmiennej X dla wszystkich jednostek w badanej zbiorowości (liczącej n jednostek) bez powtórzeń (jak wyżej) bądź z powtórzeniami². Prace Giniego były publikowane głównie po włosku, lecz niestety te oryginalne wersje nie są dostępne w Polsce. Przeglądy artykułów badacza i ich pierwsze, częściowe tłumaczenia pojawiły się dopiero niemal sto lat później (Gini, 2005; Basulto i Gusto, 2010; Ceriani i Verme 2012).

Dla zmiennej mierzonej na skali nominalnej zamiast wartości zmiennej X można przyjąć częstości, z jakimi występują poszczególne wartości tej zmiennej.

Zapiszmy częstości wyborów poszczególnych kategorii w porządku niemalejącym $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_{n-1} \leq p_n$. Wówczas macierz D reprezentującą wartości bezwzględne różnic między wszystkimi parami częstości można zapisać następująco:

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

Wygodniej będzie przedstawić macierz D w postaci równoważnej, także zawierającej tylko nieujemne wyrazy, gdyż dla każdego i oraz q , jeżeli $q \geq i$, to $p_q \geq p_i$.

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

Oznaczmy przez

$$d_{ik} = |p_i - p_k|.$$

² Por. interesującą dyskusję między Jasso i Allisonem (1978-1979) na temat dwóch wersji miary MD(X) oraz współczynnika koncentracji Giniego.

W wyniku sumowania składników tej drugiej macierzy otrzymujemy:

$$\begin{aligned} \Delta &= \sum_{i=1}^n \sum_{k=1}^n d_{ik} = \sum_{i=1}^n \sum_{k=1}^n |p_i - p_k| = 2 \sum_{i=1}^n (i-1)p_i - 2 \sum_{i=1}^n (n-i)p_i = \\ &= 2 \sum_{i=1}^n (2i-1-n)p_i = 2 \sum_{i=1}^n (2i-2-2n+1+n)p_i = \\ &= 4 \sum_{i=1}^n (i-1-n)p_i + 2 \sum_{i=1}^n (1+n)p_i = \\ &= 2(n+1) - 4 \sum_{i=1}^n (n+1-i)p_i \end{aligned}$$

Warto zwrócić uwagę, że w ocenie wielkości bezwzględnych różnic między wszystkimi częstościami najmniejsze wagi przypisuje się największym częstościom, a największe – częstościom najmniejszym. (Odpowiada to – w wypadku pomiaru nierówności dochodów – przypisywaniu największych wag najniższym dochodom, a najmniejszych – najwyższym dochodom.) Ocena wielkości bezwzględnych różnic między wszystkimi parami częstości zależy więc od uporządkowania częstości.

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

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

Jeżeli rozkład jest jednopunktowy (tj. występuje w nim skrajna koncentracja), czyli w wypadku rozkładu, w którym jedna z kategorii występuje z częstością 1,

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

Maksymalna wartość miary $MD(K)$ w wypadku skrajnej koncentracji jest równa $\frac{2}{n}$, gdyż

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

Miara $MD(K)$ unormowana do przedziału $[0, 1]$ ma postać:

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

i przyjmuje wartość 0 – w wypadku rozkładu równomiernego, a największą wartość 1 – w wypadku rozkładu, w którym jedna z kategorii występuje z częstością 1. Jest to więc miara koncentracji.

Przyjęcie przeciwnego unormowania, tj. przypisanie wartości 0 rozkładowi jednopunktowemu, a wartości 1 – rozkładowi równomiernemu, przez odjęcie $MD(K)$ od 1, prowadzi do miary różnorodności $MDA(K)$, nazywanej analogiem średniej różnicy (*mean difference analog*).

$$\begin{aligned} MDA(K) &= 1 - \frac{\sum_{i=1}^n \sum_{j=1}^n |p_i - p_j|}{n-1} = 1 - \frac{n+1 - 2 \sum_{i=1}^n (n+1-i)p_i}{n-1} = \\ &= \frac{2 \left(\sum_{i=1}^n (n+1-i)p_i - 1 \right)}{n-1} = 1 - MD(K) \end{aligned}$$

Miara ta była analizowana przez Allena R. Wilcoxa (1973), proponowali ją wcześniej Maurice G. Kendall i Alan Stuart (1958: 47). W dalszej analizie przyjmujemy miarę $MDA(K)$ jako miarę różnorodności wyborów.

Warto zwrócić uwagę na dwie własności miary $MDA(K)$, które nie były wcześniej rozważane w literaturze.

Własność 1. Miara $MDA(K)$ spełnia postulat transferu Pigou-Daltona w sformułowaniu dostosowanym do porównywania rozkładów częstości.

Postulat transferu Pigou-Daltona jest najważniejszym i najczęściej akceptowanym postulatem dotyczącym oceny nierówności, a więc również różnorodności. W tym sformułowaniu oznacza on, że transfer niewielkiej części częstości od kategorii o większej częstości do kategorii o mniejszej częstości, ale takiej części, która nie zmienia relacji między ich częstościami, zmniejsza nierówność, a więc zwiększa różnorodność. Aby wykazać, że postulat ten jest spełniony, wystarczy rozważyć jego konsekwencje dla dwóch kategorii sąsiadujących ze sobą w uporządkowaniu ze względu na częstości. Transfer nie zmienia ich pozycji, a jedynie zmniejsza częstość kategorii zajmującej wyższą pozycję i zwiększa częstość kategorii zajmującej niższą pozycję. W konsekwencji, różnica między ich częstościami maleje, a różnorodność wzrasta.

Miara $MDA(K)$ spełnia jeszcze inne kryteria oceny nierówności (a więc również różnorodności), ale nie będą one tutaj rozważane.

Własność 2. Jeżeli uporządkowania kategorii ze względu na częstości wyborów we wszystkich grupach są jednakowe, to wartość średnia warunkowych miar różnorodności jest równa wartości tej miary w całej zbiorowości, tj.:

$$E[MDA(K|G)] = MDA(K).$$

Wystarczy zauważyć dwa fakty. Po pierwsze, uporządkowanie kategorii ze względu na częstości w całej zbiorowości jest w tym wypadku takie samo, jak w każdej

grupie. Po drugie, częstość każdej kategorii w całej zbiorowości jest równa wartości średniej częstości tej kategorii w poszczególnych grupach. Oczywiście, częstości tej kategorii w poszczególnych grupach mogą być i na ogół są różne.

Własność 3. Jeżeli uporządkowania kategorii ze względu na częstości wyborów we wszystkich grupach nie są jednakowe, to wartość średnia warunkowych miar różnorodności jest mniejsza od wartości tej miary w całej zbiorowości, tj.:

$$E[MDA(K|G)] < MDA(K).$$

Wystarczy rozważyć wymianę częstości wyborów między dwiema sąsiednimi kategoriami (np. o największych częstościach) w jednej grupie. Wymiana ta nie zmienia wartości miary różnorodności w tej grupie. Zmienia jednak częstości wyboru tych kategorii w całej zbiorowości, nawet jeżeli pozycje tych kategorii ze względu na częstości nie zmieniają się. Wartość miary różnorodności w całej zbiorowości zwiększy się.

Własności 2. i 3. są podobne do własności innych parametrów rozproszenia (zarówno zróżnicowania, jak i różnorodności) (por. Lissowski, Haman i Jasiński, 2011, rozdział 6). Zakłada się w nich identyczność lub różność parametrów przeciętnych bądź nawet całych rozkładów warunkowych.

4. PROPOZYCJA MIARY ZALEŻNOŚCI

Miara zależności wyborów od grupy $MZ_{K|G}$ będzie skonstruowana w ten sam sposób, co wszystkie miary zależności statystycznej (por. Lissowski, Haman i Jasiński, 2011, rozdział 9).

$$MZ_{K|G} = \frac{MDA(K) - E[MDA(K|G)]}{MDA(K)}$$

Miara zależności $MZ_{K|G}$ jest więc różnicą między różnorodnością w całej zbiorowości a średnią różnorodnością w grupach, unormowaną do przedziału $[0, 1]$. Miara ta nie jest określona w sytuacji, gdy wszystkie wybory w całej zbiorowości są jednakowe. Wartość tej miary jest oceną wielkości redukcji różnorodności w wyniku wykorzystania informacji o przynależności do grupy.

Analiza uporządkowań częstości wyborów kategorii w poszczególnych grupach wymaga wprowadzenia dodatkowych oznaczeń: rang częstości wyborów kategorii w grupach oraz warunkowych częstości wyborów kategorii. Rangi będą przyjmowały wartości od 1 do n , przy czym najwyższe rangi będą przypisywane kategoriom o najwyższych częstościach, a w wypadku kategorii o jednakowych częstościach rangi będą identyczne, równe średniej tych rang, które powinny być przypisane tym

kategoriom. W poprzedniej sekcji rangi były określone przez założenie, że częstości w całej zbiorowości są przedstawione w postaci niemalejącego ciągu.

$R(K = k_i) = r_i$ ranga częstości kategorii k_i w uporządkowaniu w całej zbiorowości

$R(K = k_i | G = g_h) = r_{ih}$ ranga warunkowej częstości kategorii k_i w uporządkowaniu w grupie g_h

Przy tych oznaczeniach, wykorzystując jedną z postaci miary $MDA(K)$, otrzymujemy:

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

$$MDA(K | G = g_h) = \frac{2 \left[\sum_{i=1}^n (n+1-r_{ih}) P(K = k_{r_n} | G = g_h) - 1 \right]}{n-1}$$

$$\begin{aligned} MZ_{K|G} &= \frac{MDA(K) - E[MDA(K|G)]}{MDA(K)} = \\ &= \frac{2 \left[\sum_{i=1}^n (n+1-r_i) P(K = k_r) - 1 \right]}{n-1} - \frac{\sum_{h=1}^m P(G = g_h) \cdot 2 \left[\sum_{i=1}^n (n+1-r_{ih}) P(K = k_{r_n} | G = g_h) - 1 \right]}{n-1} = \\ &= \frac{2 \left[\sum_{i=1}^n (n+1-r_i) P(K = k_r) - 1 \right]}{n-1} = \\ &= \frac{\sum_{i=1}^n (n+1-r_i) P(K = k_r) - \sum_{h=1}^m P(G = g_h) \left[\sum_{i=1}^n (n+1-r_{ih}) P(K = k_{r_n} | G = g_h) \right]}{\sum_{i=1}^n (n+1-r_i) P(K = k_r) - 1} \end{aligned}$$

Miara zależności $MZ_{K|G}$ przyjmuje następujące wartości:

$$0 \leq MZ_{K|G} \leq 1.$$

Własność 1. Miara zależności $MZ_{K|G}$ przyjmuje wartość 0 zawsze i tylko wtedy, gdy uporządkowania częstości wyborów kategorii są niezależne od grupy.

Własność ta wynika bezpośrednio z własności 2. i 3. miary $MDA(K)$ przedstawionych w poprzedniej sekcji. Łatwo zauważyć, że jeżeli – zgodnie z definicją niezależności uporządkowań częstości wyborów kategorii od grupy – w każdej grupie i dla każdej kategorii $r_{ih} = r_i$, to $E[MDA(K|G)]$ równa się $MDA(K)$.

W tabeli 1. uporządkowanie częstości w każdej grupie (tj. częstości warunkowych) jest identyczne, chociaż same częstości w poszczególnych grupach różnią się. Identyczne są więc rangi warunkowe. Zauważmy również, że częstość każdej kategorii w całej zbiorowości jest równa średniej ważonej częstości warunkowych w grupach. Wagami są oczywiście udziały grup w całej zbiorowości. Na przykład

$$P(K = k_1) = 0,40 \times 0,30 + 0,20 \times 0,40 + 0,40 \times 0,25 = 0,300.$$

Tabela 1
Obliczanie miary zgodności $MZ_{K|G}$ (dane fikcyjne)

K	G			Razem
	g_1	g_2	g_3	
k_1	0,30	0,40	0,25	0,300
k_2	0,25	0,25	0,21	0,234
k_3	0,20	0,20	0,20	0,200
k_4	0,15	0,15	0,19	0,166
k_5	0,10	0,00	0,15	0,100
Razem	1,00	1,00	1,00	1,000
$P(G = g_h)$	0,40	0,20	0,40	1,000
MDA	0,75	0,55	0,89	0,766

$$E[MDA(K|G)] = 0,40 \times 0,75 + 0,20 \times 0,55 + 0,40 \times 0,89 = 0,766 = MDA(K)$$

$$MZ_{K|G} = 0$$

Własność 2. Miara zależności $MZ_{K|G}$ przyjmuje wartość 1 zawsze i tylko wtedy, gdy wybory kategorii są maksymalnie zależne od grupy.

Ta własność miary zależności $MZ_{K|G}$ jest oczywista. Wystarczy zauważyć, że w tym wypadku wartość $E[MDA(K|G)]$ musi być równa 0. Tak może być jedynie wtedy, gdy w każdej grupie warunkowy rozkład wyborów jest zerojedynkowy i w konsekwencji dla każdego $g_h \in G$ zachodzi $MDA(K|G = g_h) = 0$.

Najważniejsza jest jednak interpretacja miary zależności $MZ_{K|G}$, która stanowi jej podstawowe uzasadnienie. Będzie jej poświęcona następną sekcja.

5. INTERPRETACJA MIARY ZALEŻNOŚCI

Inspirację dla interpretacji miary zależności $MZ_{K|G}$ stanowiły dwa źródła. Pierwszym była interpretacja miary proporcjonalnej predykcji (Goodman i Kruskal, 1954),

a drugim –interpretacja współczynnika koncentracji Giniego w kategoriach gry (Pyatt, 1976).

Inspiracja 1.

Proporcjonalna predykcja polega na „odgadywaniu” kategorii. W tym celu jednostkom w zbiorowości losowo przypisuje się różne kategorie z prawdopodobieństwami równymi częstościom $P(K = k_i)$, z jakimi występują w tej zbiorowości. Częstość popełniania błędu przy takim „odgadywaniu” jest określona za pomocą parametru nazywanego różnorodnością klasyfikacji³.

$$R(K) = 1 - \sum_{i=1}^n P(K = k_i)^2$$

Jeżeli przypisywanie dokonywane jest na podstawie informacji, że jednostka należy do grupy $g_h \in G$, to warunkowa częstość popełniania błędu jest równa

$$R(K|G = g_h) = 1 - \sum_{i=1}^n P(K = k_i | G = g_h)^2.$$

Miara zależności statystycznej kategorii od grupy w sensie proporcjonalnej predykcji ma postać

$$\tau_{K|G} = \frac{R(K) - E[R(K|G)]}{R(K)}.$$

Inspiracja 2.

Graham Pyatt sformułował interesującą interpretację współczynnika koncentracji Giniego w postaci gry czy też eksperymentu. Dla każdej osoby o określonym dochodzie wybiera się losowo dochód z prawdopodobieństwami określonymi w rozkładzie dochodów w danej zbiorowości. Jeżeli aktualny dochód osoby 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 Giniego ma interpretację jako wartość oczekiwana zysku w tej grze, tj. średnia wielkość zwiększenia dochodów.

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

gdzie wartości zmiennej X oznaczają dochody jednostek.

Niech $K = k_i$ oznacza kategorię wybraną przez osobę, a $K = k_q$ – kategorię przypisaną jej losowo, zgodnie z rozkładem rzeczywistych wyborów kategorii. Niech ocena porównania tych kategorii będzie równa 1, gdy kategoria przypisana osobie losowo k_q należy do kategorii rzadziej lub równie często wybieranych, jak wybrana przez

³ Przy pewnych założeniach parametr ten ma postać podobną do parametru różnorodności $MDA(K)$.

osobę kategorią k_i , natomiast 0 – gdy kategoria przypisana losowo k_q należy do kategorii wybieranej częściej niż kategoria k_i .

Strata ℓ w wyniku losowego przypisania osobie kategorii jest równa:

- 1, jeżeli $k_i \geq k_q$,
- 0, jeżeli $k_i < k_q$.

Ocena ta jest więc oceną braku zgodności między uporządkowaniem kategorii losowo przypisanej osobie i kategorii wybranej przez nią z uporządkowaniem wyznaczonym przez częstości rzeczywistych wyborów.

Porównując całe kategorie (wybraną k_i i przypisaną k_q), otrzymujemy

$$\ell(K = k_i, K = k_q) = \begin{cases} 0 & \text{gdy } k_i < k_q \\ P(K = k_q) & \text{gdy } (k_i \geq k_q) \end{cases}.$$

Sumując te oceny dla poszczególnych kategorii $K = k_i$ rzeczywiście wybranych przez osoby, otrzymujemy

$$\ell(K = k_i) = \sum_{q=1}^i P(K = k_q) = F_K(k_i),$$

gdzie $F_K(k_i)$ oznacza sumę częstości kategorii o mniejszych częstościach niż częstość kategorii k_i , a więc dystrybuantę rang przypisanych kategoriom ze względu na częstości.

Warto zauważyć, że

$$\sum_{i=1}^n \ell(K = k_i) = \sum_{i=1}^n F_K(k_i) = \sum_{i=1}^n (n+1-i)P(K = k_i).$$

To ostatnie wyrażenie jest znane z definicji $MDA(K)$.

Miara rozproszenia $MDA(K)$, analog średniej różnicy, jest więc równa

$$MDA(K) = \frac{2 \left(\sum_{i=1}^n F_K(k_i) - 1 \right)}{n-1} = \frac{2 \left(\sum_{i=1}^n F_K(r_i) - 1 \right)}{n-1}.$$

W tej ostatniej postaci $MDA(K)$ jest zapisana jako funkcja wszystkich dystrybuant rang kategorii ze zbioru K .

Zatem $MDA(K)$ jest oceną braku zgodności między uporządkowaniem przypisanej kategorii k_q i wybranej kategorii k_i z uporządkowaniem kategorii wyznaczonym przez rzeczywiste częstości ich wyborów.

Analogiczne wyniki otrzymujemy dla rozkładu warunkowego wyboru kategorii w każdej grupie $g_h \in G$.

$$MDA(K|G = g_h) = \frac{2 \left(\sum_{i=1}^n F_{K|G=g_h}(r_i) - 1 \right)}{n-1}$$

Miarę zależności wyboru kategorii od grupy $MZ_{K|G}$ można więc zapisać następująco:

$$\begin{aligned} MZ_{K|G} &= \frac{MDA(K) - E[MDA(K|G)]}{MDA(K)} = \\ &= \frac{\frac{2 \left(\sum_{i=1}^n F_K(r_i) - 1 \right)}{n-1} - \sum_{h=1}^m P(G = g_h) \frac{2 \left(\sum_{i=1}^n F_{K|G=g_h}(r_i) - 1 \right)}{n-1}}{\frac{2 \left(\sum_{i=1}^n F_K(r_i) - 1 \right)}{n-1}} = \\ &= \frac{\sum_{i=1}^n F_K(r_i) - \sum_{h=1}^m P(G = g_h) \sum_{i=1}^n F_{K|G=g_h}(r_i)}{\sum_{i=1}^n F_K(r_i)} \end{aligned}$$

Wartość miary $MZ_{K|G}$ zapisanej powyżej w innej postaci – jako funkcji dystrybuant rang kategorii w całej zbiorowości i (na ogół różnych) rang kategorii w poszczególnych grupach – oznacza stopień redukcji straty, tj. oceny braku zgodności między uporządkowaniem kategorii przypisanej i kategorii wybranej z uporządkowaniem kategorii wyznaczonym przez rzeczywiste częstości ich wyborów, w wyniku tego, że wykorzystana została informacja o przynależności osoby do określonej grupy.

6. PRZYKŁADY ZASTOSOWAŃ

Dla ilustracji proponowanych miar przedstawimy trzy przykłady. Wykorzystamy dane z wyborów kandydatów w I turze wyborów prezydenckich w latach 1995 i 2005.

Przykład 1. Wybory prezydenckie w 1995 roku

Państwowa Komisja Wyborcza zarejestrowała 17 kandydatów, ale czterech z nich wycofało swoje kandydatury przed wyborami (trzech zrezygnowało na rzecz Lecha Wałęsy, a jeden – na rzecz Jana Olszewskiego). Ostatecznie w wyborach uczestniczyło 13 kandydatów. W tabeli 2 podano odsetki głosów 7 kandydatów, którzy otrzymali powyżej 2,5% głosów. Informacja o pozostałych kandydatach podana została łącznie w kategorii „Inni”. Podstawą podziału wyborców na grupy było zamieszkanie na terenach dawnych zaborów.

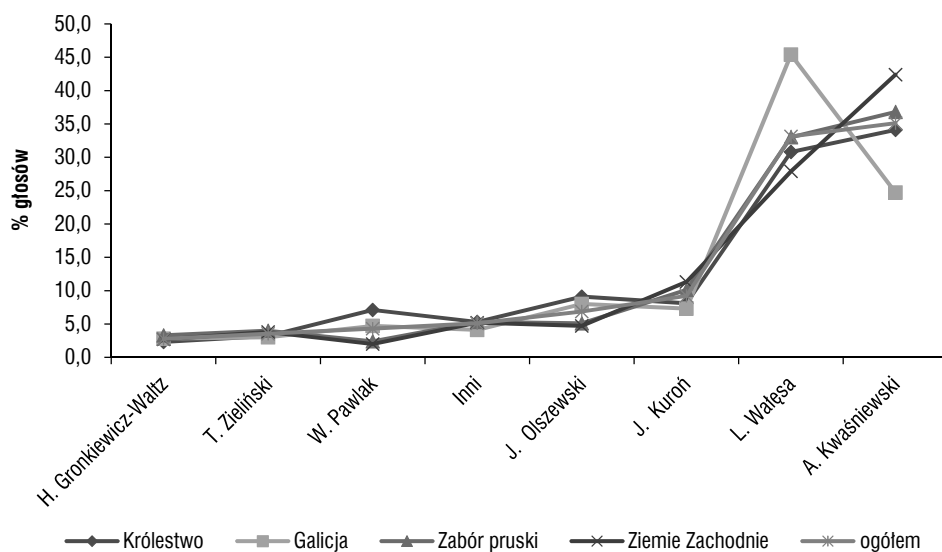
Tabela 2

Wyniki głosowania na najpopularniejszych kandydatów w wyborach prezydenckich 1995 roku (I tura) na terenach dawnych zaborów (w %)

Kandydat	Królestwo	Galicja	Zabór pruski	Ziemie Zachodnie	Ogółem
H. Gronkiewicz-Waltz	2,3	2,8	3,3	2,7	2,76
J. Kuroń	8,1	7,3	10,1	11,3	9,22
A. Kwaśniewski	34,1	24,7	36,8	42,4	35,11
J. Olszewski	9,1	8,0	5,1	4,7	6,87
W. Pawlak	7,1	4,7	2,4	2,0	4,31
L. Wałęsa	30,8	45,4	33,0	27,9	33,12
T. Zieliński	3,2	3,0	4,0	3,8	3,53
Inni	5,3	4,1	5,3	5,2	5,1
Razem	100,0	100,0	100,0	100,0	100,0
Udział grupy ^a	0,386	0,088	0,290	0,236	1,0
Miara różnorodności	0,750	0,681	0,700	0,666	0,725
Indeks rozbieżności	0,1046	0,2770	0,0758	0,1950	–

^a Wielkości przybliżone na podstawie danych GUS

Źródło: opracowanie własne na podstawie Raciborski (1997: tabela 5.4)



Wykres 1. Wyniki głosowania na kandydatów w wyborach prezydenckich 1995 roku (I tura) w uporządkowaniu według częstości wyborów na terenach dawnych zaborów (w %)

Wyniki głosowania warto przedstawić w postaci graficznej. Na wykresie 1 kandydaci zostali uporządkowani według rosnących częstości wyborów w całej zbiorowości wyborców. Bardzo zbliżone są odsetki otrzymanych głosów przez mniej popularnych kandydatów. Różnice ujawniają się głównie w głosowaniu na dwóch najpopularniej-

szych: L. Wałęsę i A. Kwaśniewskiego. Prawie identyczne jest uporządkowanie tych kandydatów według częstości głosów otrzymanych od wyborców mieszkających na terenach Królestwa Kongresowego i zaboru pruskiego. Wyraźnie większe różnice występują między wyborcami mieszkającymi na terenach Galicji i Ziemi Zachodnich.

Jednak odsetki głosów oddanych na poszczególnych kandydatów różnią się. Nieco bardziej wyrównane są odsetki głosów oddanych na kandydatów przez wyborców zamieszkałych na terenie Królestwa Kongresowego oraz zaboru pruskiego. W konsekwencji nieco większe są wartości miar różnorodności w tych dwóch grupach. Miara różnorodności dla wszystkich wyborców jest większa niż w poszczególnych grupach.

Wartość średnia miary różnorodności w grupach wynosi

$$E[MDA(K|G)] = 0,386 \times 0,75 + 0,088 \times 0,681 + 0,290 \times 0,700 + 0,236 \times 0,666 = 0,710.$$

Proponowana miara zgodności ma wartość

$$MZ_{K|G} = \frac{MDA(K) - E[MDA(K|G)]}{MDA(K)} = \frac{0,725 - 0,710}{0,725} = 0,0207,$$

czyli informacja o zamieszkiwaniu na terenach dawnych zaborów w niewielkim stopniu (2,07%) redukuje różnorodność zaobserwowaną w wyborach wszystkich wyborców i stanowi ocenę niewielkiego stopnia niezgodności między wyborami w poszczególnych grupach i w całej zbiorowości wyborców.

Większe różnice niż w miarach różnorodności można zaobserwować w odsetkach głosów wszystkich wyborców w porównaniu zwłaszcza z odsetkami głosów w Galicji i na Ziemiach Zachodnich. Miarą braku podobieństwa między rozkładami wyborców w poszczególnych grupach a rozkładem dla wszystkich wyborców jest indeks rozbieżności. Indeks ten jest równy połowie sumy modułów różnic między odsetkami w dwóch porównywanych rozkładach i przyjmuje wartości od 0 (w wypadku identycznych rozkładów) do 1 (w wypadku maksymalnej różnicy między nimi).

$$\text{indeks rozbieżności} = \frac{1}{2} \sum_{i=1}^n |P(X = x_i) - Q(X = x_i)|$$

Wartość tego indeksu wskazuje, jaki odsetek jednostek zbiorowości musiałby zmienić swoje wartości, aby rozkłady były identyczne.

Od wielu lat w rozmaitych dziedzinach obserwuje się znaczące i utrzymujące się różnice między terenami dawnych zaborów. Rozbieżności te stwierdza się również podczas głosowania, zwłaszcza w wyborach do Sejmu. Socjologiczną analizę wpływu dziedzictwa zaborowego na preferencje wyborcze przedstawił m.in. Jacek Raciborski (1997: 157–165).

Przykłady 2 i 3. Wybory prezydenckie w 2005 roku

Państwowa Komisja Wyborcza zarejestrowała 16 kandydatów, ale trzech zrezygnowało przed datą wyborów, a jeden zmarł. Ostatecznie w wyborach uczestniczyło 12 kandydatów. W tabelach 3 i 4 podano odsetki głosów 7 kandydatów, którzy otrzymali powyżej 1% głosów. Informację o pozostałych kandydatach zawarto łącznie w kategorii „Inni”.

Wyniki (podawane przez Państwową Komisję Wyborczą) nie zawierają żadnych informacji na temat sposobu głosowania przez wyborców. W tabelach 3 i 4 wykorzystano więc wyniki sondażu *exit poll* przeprowadzonego przez konsorcjum PBS i TNS OBOP dla TVP1 na losowej próbie 47 916 wyborców.

Podstawą podziału wyborców na grupy będą:

- zamieszkanie na terenie miast o różnej wielkości lub na wsi (w tabeli 3),
- sposób głosowania w wyborach do Sejmu w 2005 roku (w tabeli 4).

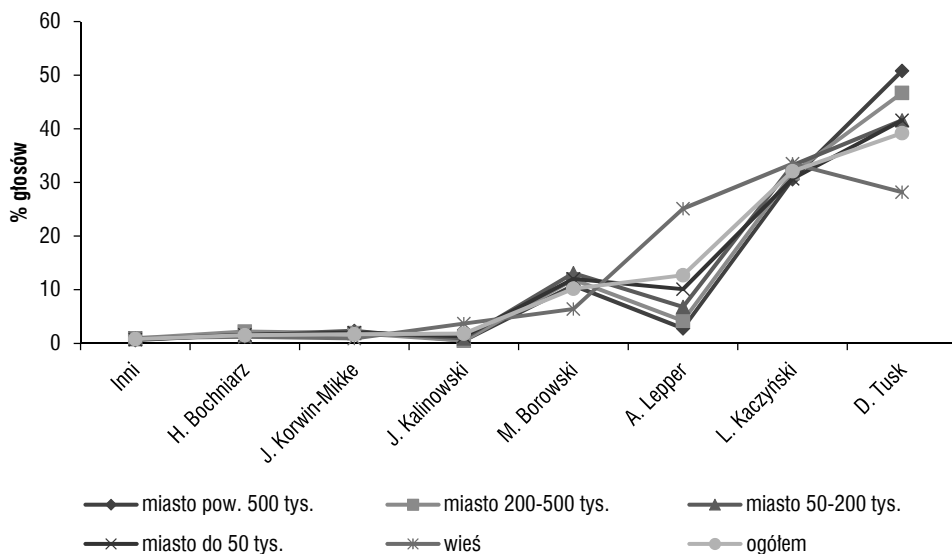
Tabela 3

Wyniki głosowania na najpopularniejszych kandydatów w wyborach prezydenckich 2005 roku (I tura) w miastach i na wsi (w %)

Kandydat	Typ miejscowości					Ogółem
	miasto pow. 500 tys.	miasto 200-500 tys.	miasto 50-200 tys.	miasto do 50 tys.	wieś	
H. Bochniarz	1,6	2,2	1,5	1,6	1,2	1,5
M. Borowski	10,8	11,9	13,1	12,0	6,4	10,2
L. Kaczyński	30,6	31,7	33,2	30,7	33,5	32,1
J. Kalinowski	0,5	0,5	0,9	1,3	3,7	1,8
J. Korwin-Mikke	2,3	1,9	2,1	2,0	0,9	1,7
A. Lepper	2,8	4,2	6,8	10,1	25,1	12,7
D. Tusk	50,8	46,7	41,6	41,6	28,2	39,2
Inni	0,6	0,9	0,8	0,7	1,0	0,8
Razem	100,0	100,0	100,0	100,0	100,0	100,0
Udział grupy	0,146	0,111	0,171	0,247	0,325	1,0
Miara różnorodności	0,524	0,555	0,588	0,608	0,656	0,621
Indeks rozbieżności	0,258	0,204	0,136	0,092	0,318	

Źródło: opracowanie własne na podstawie Miszkowski (2008: tabela 1.8)

Podobnie jak poprzednio, warto przedstawić wyniki głosowania w postaci graficznej. Na wykresie 2 kandydaci zostali uporządkowani w kolejności według rosnących częstości wyborów w całej zbiorowości wyborców.



Wykres 2. Wyniki głosowania na kandydatów w wyborach prezydenckich 2005 roku (I tura) w uporządkowaniu według częstości wyborów w miastach i na wsi (w %)

Warto zwrócić uwagę na prawie identyczne odsetki głosów, które kandydaci otrzymali od wyborców mieszkających w miastach powyżej 500 tys. i tych liczących 200-500 tys. mieszkańców. Odsetki te są również bardzo podobne w dwóch pozostałych typach miast: mających 50-200 tys. i do 50 tys. mieszkańców.

Odsetki głosów, a nawet uporządkowanie kandydatów według głosów otrzymanych od mieszkańców wsi, są znacząco inne. W tym przypadku znacznie niższe są wartości miar różnorodności. Wskazują one na większe zróżnicowanie częstości głosów w poszczególnych grupach. Mniejsza miara różnorodności wskazuje też na większe zróżnicowanie odsetków głosów w całej zbiorowości wyborców.

Wartość średnia miary różnorodności w grupach wynosi

$$E[MDA(K|G)] = 0,146 \times 0,524 + 0,111 \times 0,555 + 0,171 \times 0,588 + 0,247 \times 0,608 + 0,325 \times 0,656 = 0,608.$$

Proponowana miara zgodności ma wartość

$$MZ_{K|G} = \frac{MDA(K) - E[MDA(K|G)]}{MDA(K)} = \frac{0,621 - 0,608}{0,621} = 0,0210,$$

czyli informacja o typie miejscowości, w której mieszkają wyborcy, w niewielkim stopniu (2,1%) redukuje różnorodność zaobserwowaną w wyborach wszystkich wyborców i stanowi ocenę niewielkiego stopnia niezgodności między wyborami w poszczególnych grupach i w całej zbiorowości wyborców.

Tabela 4

Wyniki głosowania na najpopularniejszych kandydatów w wyborach prezydenckich 2005 (I tura) wg sposobu głosowania wyborcy w wyborach do Sejmu w 2005 roku (w %)

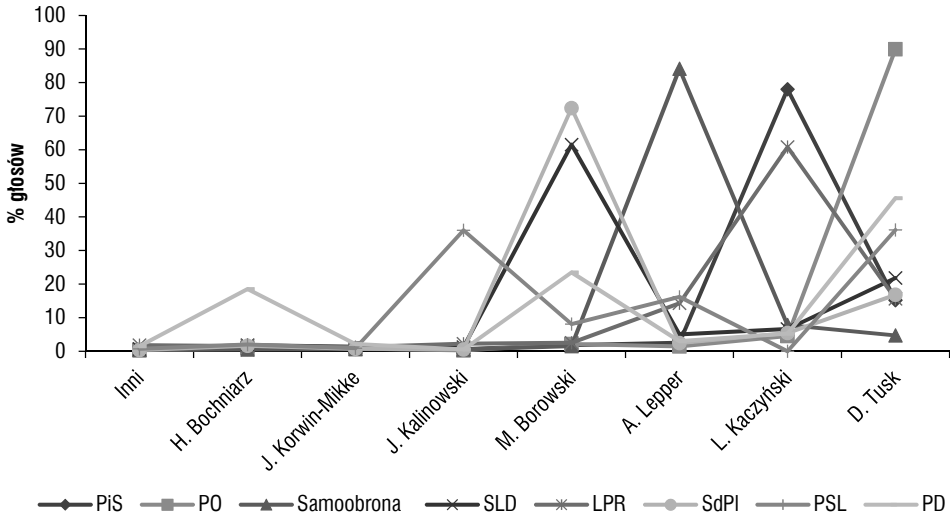
Kandydat	Sposób głosowania w wyborach do sejmu 2005											Ogółem
	PiS	PO	Samoo- obrona	SLD	LPR	SdPI	PSL	PD	Inne	Nie pamięta	Nie głosował	
H. Bochniarz	0,7	0,9	0,5	1,8	1,5	1,60	2,0	18,5	3,10	2,0	1,9	1,5
M. Borowski	1,8	2,1	1,5	61,5	2,5	72,40	8,1	23,5	9,60	11,1	10,4	10,2
L. Kaczyński	78,0	4,5	7,8	6,6	60,8	5,40	18,0	5,2	15,40	29,1	23,5	32,1
J. Kalinowski	0,6	0,2	0,5	1,2	2,2	0,50	36,0	0,5	1,70	2,2	1,5	1,8
J. Korwin-Mikke	0,7	0,8	0,6	1,3	1,2	0,70	0,9	2,1	27,90	1,3	1,9	1,7
A. Lepper	2,5	1,4	84,1	5,0	14,3	2,40	16,2	3,0	10,80	18,7	14,4	12,3
D. Tusk	15,2	89,9	4,7	21,8	15,6	16,8	36,1	45,6	26,70	32,5	45,4	39,6
Inni	0,5	0,3	0,3	0,7	1,8	0,30	0,7	1,5	4,80	3,0	1,1	0,8
Razem	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Udział grupy	0,280	0,244	0,078	0,063	0,039	0,030	0,028	0,019	0,026	0,023	0,170	1,000
Miara różnorodności	0,394	0,351	0,373	0,499	0,499	0,425	0,856	0,600	0,805	0,694	0,618	0,621
Indeks rozbieżności	0,459	0,503	0,718	0,517	0,322	0,623	0,346	0,375	0,318	0,105	0,089	

PiS – Prawo i Sprawiedliwość; PO – Platforma Obywatelska; SLD – Sojusz Lewicy Demokratycznej; LPR – Liga Polskich Rodzin; SdPI – Socjaldemokracja Polska; PSL – Polskie Stronnictwo Ludowe; PD – Partia Demokratyczna

Źródło: opracowanie własne na podstawie Miskowski (2008: tabela 1.2)

Rozkłady głosów w grupach wyborców, którzy w wyborach do Sejmu głosowali na poszczególne partie, wyraźnie różnią się od siebie i od rozkładu głosów w całej zbiorowości wyborców. Indeksy rozbieżności są znacząco wyższe niż w dwóch poprzednich przykładach. Naturalnie najwyższy i wysoki odsetek głosów otrzymali kandydaci, którzy byli przez daną partię zgłoszeni lub popierani. Jedynie rozkłady głosów wyborców, którzy nie pamiętają, na kogo głosowali, albo którzy nie głosowali w ogóle, niewiele różnią się od rozkładu głosów w całej zbiorowości wyborców.

Na wykresie 3 przedstawiono rozkłady głosów tylko tych wyborców, którzy w wyborach do Sejmu w 2005 roku głosowali na poszczególne partie. Wykres ten różni się wyraźnie od dwóch poprzednich.



Wykres 3. Wyniki głosowania na kandydatów w wyborach prezydenckich 2005 roku (I tura) w uporządkowaniu według częstości wyborów w grupach wyborców, którzy w wyborach do Sejmu w 2005 roku głosowali na poszczególne partie polityczne (w %)

Wartości miar różnorodności w poszczególnych grupach są znacznie niższe niż w poprzednich przykładach. Wynika to z występowania „wiodących” kandydatów w poszczególnych grupach i – w konsekwencji – większego zróżnicowania częstości głosów w grupach.

Wartość średnia miary różnorodności w grupach wynosi

$$E[MDA(K|G)] = 0,280 \times 0,394 + 0,244 \times 0,351 + 0,078 \times 0,373 + 0,063 \times 0,499 + 0,039 \times 0,499 + 0,030 \times 0,425 + 0,028 \times 0,856 + 0,019 \times 0,600 + 0,026 \times 0,805 + 0,023 \times 0,694 + 0,170 \times 0,618 = 0,466.$$

Proponowana miara zgodności ma wartość

$$MZ_{K|G} = \frac{MDA(K) - E[MDA(K|G)]}{MDA(K)} = \frac{0,621 - 0,466}{0,621} = 0,250,$$

czyli informacja o sposobie głosowania w wyborach do Sejmu w 2005 roku w większym stopniu (25,0%) redukuje różnorodność zaobserwowaną w wyborach wszystkich wyborców niż informacja o typie miejscowości, w której mieszkają wyborcy. Jest to ocena większego stopnia niezgodności między wyborami w poszczególnych grupach i w całej zbiorowości wyborców.

7. PODSUMOWANIE

Analiza wyborów dokonywanych przez ludzi i zależność tych wyborów od przynależności do grupy jest przedmiotem wielu badań społecznych i psychologicznych. Mogą to być wybory kandydatów lub partii w wyborach demokratycznych, wybory wartości lub ideałów społecznych, aspiracje i cele życiowe, a także preferencje ludzi lub grup społecznych.

Artykuł przedstawia propozycję nowego typu zależności statystycznej. Przede wszystkim *novum* stanowi określenie niezależności statystycznej jako niezależności częstości wyborów kategorii od grupy. Określenie to pozwala abstrahować od niewielkich różnic w częstościach wyborów w różnych grupach i podkreśla znaczenie tendencji ujawnianych przez uporządkowanie częstości tych wyborów. Nowa jest także proponowana miara zależności, której własności i interpretacja zostały opisane. Podstawą do konstrukcji tej miary jest nowa miara różnorodności. Wprawdzie miara ta jest wywiedziona z klasycznej propozycji miary rozproszenia, sformułowanej przez włoskiego statystyka Corrado Giniego ponad sto lat temu, zaadaptowanej tutaj do zmiennych mierzonych na skalach nominalnych, która czasem pojawiała się także w propozycjach *ad hoc* w ubiegłym wieku, ale jej własności (choć nie wszystkie, gdyż ograniczone przez cel tego artykułu) przedstawione zostały po raz pierwszy.

Zaprezentowana koncepcja i wszystkie pojęcia oraz proponowane parametry są zgodne z ogólną teorią opisu statystycznego. Miara zależności – jak wszystkie parametry siły zależności statystycznej – określa stopień redukcji błędu opisu statystycznego dzięki wykorzystaniu informacji o drugiej zmiennej (przynależności do grupy). Miara różnorodności ma takie same własności jak inne parametry rozproszenia (zróznicowania lub różnorodności).

Ocena trafności i przydatności przedstawionych propozycji możliwa będzie dopiero w przyszłości, gdy badacze zastosują je w swoich analizach. Obecnie można tylko ocenić zgodność proponowanej koncepcji z intuicją i potrzebami badaczy.

BIBLIOGRAFIA

- Allison, P.D. (1978). Measures of inequality. *American Sociological Review*, 43, 865–880.
- Allison, P.D. (1979). Reply to Jasso. *American Sociological Review*, 44, 870–872.
- Basulto, J., Busto, J. (2010). Gini's concentration ratio (1908–1914). *Electronic Journal for History of Probability and Statistics*, 6, No 1, 1–42.
- Ceriani, L., Verme, P. (2012). The origins of the Gini index: extracts from «Variabilità e Mutabilità» by Corrado Gini. *Journal of Economic Inequality*, 10, 421–443.
- Gini, C. (2005). On the measurement of concentration and variability of characters. *Metron – International Journal of Statistics*, 63, No. 1, 3–38.
- Goodman, L.A., Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732–764.
- Jasso, G. (1979). On Gini's mean difference and Gini's index of concentration. *American Sociological Review*, 44, 867–870.
- Kendall, M.G., Stuart, A. (1958). *The Advanced Theory of Statistics*, Vol. I. New York: Hafner Publishing Co.
- Lissowski, G., Haman, J., Jasiński, M. (2011). *Podstawy statystyki dla socjologów*. Warszawa: Wydawnictwo Naukowe SCHOLAR.
- Miszczkowski, M. (2008). Wybrane determinanty preferencji wyborczych w wyborach prezydenckich 2005 roku. W: J. Raciborski (red.), *Studia nad wyborami. Polska 2005–2006*. Warszawa: Wydawnictwo Naukowe SCHOLAR, 165–202.
- Pyatt, G. (1976). On the interpretation and disaggregation of Gini coefficient. *The Economic Journal*, 86, 243–255.
- Raciborski, J. (1997). *Polskie wybory. Zachowania wyborcze społeczeństwa polskiego 1989–1995*. Warszawa: Wydawnictwo Naukowe SCHOLAR.
- Tschuprov, A.A. (1939). *Principles of the Mathematical Theory of Correlation*. London: W. Hodge. Pierwsze wydanie w języku niemieckim z 1925 r.
- Wilcox, A.R. (1973). Indices of qualitative variation and political measurement. *The Western Political Quarterly*, 26, No. 2, 325–343.