Tadeusz Sozański

Statystyczne miary zmienności a kwantyfikacja nierówności społecznej

Notatka dla uczestników kursu „Podstawy statystyki”
Instytut Socjologii UJ, rok ak. 2005/2006
Listopad 2005

„Chociaż socjologowie od dawna interesowali się nierównością społeczna, niewielu próbowało sprecyzować sens tego terminu. Łatwo oczywiście odróżnić doskonałą równość od stanu nierówności, jednakże gdy dane są dwa różne, nierówne rozkłady jakiegoś dobra, natychmiast rodzi się pytanie, w jaki sposób ocenić, który z nich jest bardziej nierówny. Odpowiedź na to pytanie wydaje się warunkiem wstępnym budowy jakiejkolwiek teorii zajmującej się determinantami i konsekwencjami nierówności społecznej. […] Dopóki badania nad nierównością koncentrowały się na wyznacznikach indywidualnych osiągnięć, dopóty brak ścisłości nie powodował większych trudności. Dopiero najnowsze próby sprawdzania hipotez wyjaśniających dlaczego w pewnych społeczeństwach występuje silniejsza nierówność niż w innych zmusiły socjologów do zastosowania ścisłych miar takich jak indeks Giniego czy odchylenie standardowe. […] Uznanie jednego rozkładu za bardziej nierówny niż inny ma implikacje zarówno teoretyczne jak metodologiczne. W rzeczy samej, wybór miary nierówności należy traktować raczej jako wybór jednej z alternatywnych definicji nierówności niż jako wybór jednego z alternatywnych sposobów mierzenia jednego konstruktu teoretycznego” (Allison 1978: 865).

Wspomniany w cytowanym fragmencie najważniejszy współczynnik nierówności, zdefiniowany w 1912 r. przez włoskiego demografa i statystyka Corrado Giniego, wciąż jest mało znany polskim socjologom, o czym świadczy choćby monografia poświęcona nierównościom społecznym (M. Jarosz. *Nierówności społeczne*. Warszawa 1984), w której do oceny stopnia zróżnicowania dochodów stosuje się wyłącznie stosunek dochodu maksymalnego do minimalnego. W najnowszym podręczniku makrosocjologii (H. Domański. *Struktura społeczna*. Warszawa 2004) pojawia się (s. 33) wprawdzie zestawienie wartości współczynnika Giniego dla „dochodów rodzin na osobę“ w różnych krajach europejskich, poprzedzone krótkim wyjaśnieniem jak należy interpretować ów współczynnik, autor nie podaje jednak wzoru definicyjnego ani wzorów obliczeniowych, w związku z czym można odnieść mylne wrażenie, że temat daleko wykracza ponad niezbędne socjologowi minimum wiedzy statystycznej.Także podręczniki statystyki (w tym *Statystyka dla socjologów* Blalocka) oraz najpopularniejszy wśród socjologów program statystycznej analizy danych SPSS pomijają współczynnik Giniego (w języku komend SPSS można jednak napisać odpowiednią procedurę obliczeniową; patrz J. Górniak, J. Wachnicki. *SPSS PL for Windows. Pierwsze kroki w analizie danych*. Kraków 2000, s. 145).

Mój tekst, mający wypełnić tę lukę, zawdzięcza swoje powstanie Zbigniewowi Karpińskiemu (absolwentowi IS UJ obecnie na studiach doktoranckich w IFiS PAN). Studiując teorię Petera Blaua, klasyka socjologii XX wieku, autora monografii *Inequality and Heterogeneity* (New York 1997), odkrył on najpierw dla siebie współczynnik Giniego, a następnie zainteresował nim mnie, polecając mojej uwadze artykuł Paula Allisona “Measures of Inequality” (*American Sociological Review* 43, 1978, 865–880). Artykuł ten, którego początkowy fragment przytoczyłem na wstępie, wykorzystałem jako główne źródło, przygotowując niniejszą notatkę przeznaczoną dla uczestników kursu „Podstawy statystyki” (w roku akademickim 2005/2006 włączonego do kanonu studiów socjologicznych w UJ).

•

Po tym wprowadzeniu przejdę do bardziej systematycznego wykładu. Niech $x=(x_1,\ldots,x_n)$ oznacza ciąg wartości zmiennej x zaobserwowanych w pewnej n-elementowej zbiorowości. Zakładając, że $x_i \geq 0$ dla $i=1,\ldots,n$, wartość x_i będziemy interpretować jako kwotę pewnego *podzielnego, przekazywalnego* *dobra* w posiadaniu *i*-tej jednostki.

Niech *Sum*(*x*) oznacza *sumę wartości zmiennej x*, symbolicznie, $Sum\ (x) = \sum_{i=1}^{n} x_i$. Zgodnie z przyjętą interpretacją *x*, *Sum*(*x*) jest to całkowita ilość dobra w posiadaniu całej populacji. Dzieląc *Sum*(*x*) przez *n*, otrzymujemy *średnią arytmetyczną*, czyli ilość dobra przypadającą *przeciętnie* na jednostkę. Ten najważniejszy *parametr opisowy szeregu statystycznego*, znany także laikom, będę tu oznaczał symbolem M_x.

Miarami zmienności w szczególny sposób związanymi ze średnią arytmetyczna są *wariancja* oraz pierwiastek z niej zwany *odchyleniem standardowym*. Przypomnijmy odpowiednie wzory:

$$M_x = \frac{1}{n}\ Sum\ (x) = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\quad Var(x) = M_{(x-M_x)^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_x)^2,\quad s_x\ =\ \sqrt{Var(x)}$$

Wariancję można równoważnie zdefiniować za pomocą wzoru $\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$, w którym nie występuje M_x (według tego wzoru wariancja to połowa średniej arytmetycznej z kwadratów różnic wartości zmiennej wyznaczonych dla wszystkich uporządkowanych par jednostek). Wzór definiujący wariancję jako *średnią arytmetyczna z kwadratów odchyleń od średniej arytmetycznej* (*odchylenie od średniej* to różnica między wartością zmiennej a średnią) pozwala interpretować ten parametr jako miarę *rozproszenia* (*dyspersji*) wartości zmiennej wokół M_x. Dlaczego pod uwagę bierze się kwadraty odchyleń od średniej arytmetycznej, nie zaś od innej *miary tendencji centralnej* lub jakiejś innej wartości? Bo wówczas rozproszenie jest najmniejsze, dokładniej, $M_{(x-c)^2} \geq M_{(x-M_x)^2}$ dla dowolnego *c*.

W praktyce do opisu *szeregu statystycznego* wraz ze średnią arytmetyczną używa się odchylenia standardowego, parametru o uniwersalnym zastosowaniu i fundamentalnym znaczeniu w teorii statystyki.

•

Czy s_x nadaje się także do oceny stopnia zróżnicowania rozkładu dochodów pieniężnych lub innych zasobów? Odpowiedź na to pytanie zależy od tego, jakie warunki powinien spełniać współczynnik nierówności.

Najbardziej naturalnym warunkiem jest żądanie, aby współczynnik taki przyjmował *wartość minimalną* równą 0 wtedy i tylko wtedy gdy dobro rozdzielone jest równomiernie, tzn. $x_i=c$ dla każdego *i*, gdzie $c>0$ jest pewną liczbą (wówczas $Sum(x)=nc$, a stąd $M_x=c$). Odchylenie standardowe spełnia ten warunek, jako że $s_x=0$ wtedy i tylko wtedy gdy $x_i=M_x$ dla każdego *i*.

Drugą oczywistą własnością wymaganą od każdej miary nierówności jest zgodność z *zasadą transferów* (*principle of transfers*), która głosi, że przekazanie przez biedniejszego dowolnej części swoich zasobów bogatszemu zawsze pociąga za sobą wzrost nierówności w populacji.

Rozważmy rozkład dobra $x=(x_1,\ldots,x_n)$ taki, że $x_i \leq x_j$ dla dwu jednostek o ustalonych numerach *i*, *j*. *Transferem* od *i* do *j* o wielkości *d* ($0 \leq d \leq x_i$) nazywa się zmiana rozkładu dobra polegająca na tym, że *i*-ta osoba traci, a *j*-ta osoba zyskuje *d* jednostek dobra, tzn. $x'_i=x_i-d$, $x'_j=x_j+d$, gdzie *x*' oznacza nowy rozkład zasobów ($x'_h=x_h$ dla $h \neq i,j$, tzn. pozostałe osoby nie zmieniają stanu posiadania).

Zauważmy, że po transferze suma wartości zmiennej nie ulega zmianie, tzn. $Sum(x)=Sum(x')$. Wykorzystując ten fakt, łatwo wyprowadzić wzór $Var(x') = Var(x) + 2d^2 + 2d(x_j - x_i)$, z którego wynika, $Var(x') \geq Var(x)$, a stąd także $s_{x'} \geq s_x$, a więc odchylenie standardowe zachowuje zasadę transferów.

Naturalne jest także żądanie, by w zbiorze *n*-wymiarowych alokacji o tej samej sumie *u* (takich *x*, że $Sum(x)=u$) każdy współczynnik nierówności osiągał maksymalną w sytuacji, gdy całość zasobów jest w posiadaniu jednej osoby, tzn. $x_i=u$ dla pewnego *i* oraz $x_j=0$ dla każdego $j \neq i$. Warunku precyzującego, że najbardziej nierówne są *rozkłady najbardziej skoncentrowane*, nie potrzeba formułować osobno, gdyż jego spełnienie wynika już za zasady transferów. Istotnie, dowolny rozkład *x* taki, że *Sum*(*x*)=*u*, można zawsze przekształcić za pomocą odpowiedniej sekwencji transferów w rozkład maksymalnie skoncentrowany. Rzecz jasna dla wszystkich takich rozkładów, różniących się jedynie osobą monopolisty, dowolny współczynnik nierówności powinien przyjmować identyczną wartość, co z kolei wynika z *zasady anonimowości*, którą również zakładamy. Zasada ta, spełniona przez wszystkie parametry statystyczne, oznacza niezależność wartości parametru od numeracji jednostek analizy.

Odchylenie standardowe, podobnie jak średnią arytmetyczną, oblicza się przy założeniu *interwałowości* pomiaru zmiennej. Oba parametry są miarami „mianowanymi" wyrażonymi w tych samych jednostkach i dlatego nadają się do porównań międzypopulacyjnych jedynie wtedy, gdy zmienna reprezentuje to samo zjawisko w obu populacjach i w obu mierzona jest za pomocą tej samej skali. Dla zmiennych o wartościach nieujemnych opisujących stan posiadania rozmaitych zasobów, w tym pieniędzy, zakłada się mocniejszy od interwałowego typ pomiaru, mianowicie *pomiar stosunkowy* (*ilorazowy*), przy którym dopuszczalne przekształcenia skal mają postać $y=ax$, gdzie $a>0$.

Przekształcenie $y=ax$ może oznaczać zarówno zmianę skali pomiarowej, np. przeliczenie dochodu ze złotych na dolary, jak i zmianę wartości zmiennej mierzonej na tej samej skali, np. powiększenie ($a>1$) lub zmniejszenie ($0<a<1$) dochodu w tym samym stopniu dla każdej osoby.

Czy podwyżka płac o 10% ($a=1.1$), w wyniku której odchylenie standardowe rośnie w tym samym stosunku (z uwagi na wzór $s_{ax}=as_x$), pociąga za sobą także większą nierówność dochodów? Przypuśćmy, że dwie osoby zarabiają odpowiednio 1000 i 2000 zł. Po podwyżce pierwszy zarobi o 100 zł więcej a drugi o 200 zł więcej, a wówczas różnica ich płac, początkowo równa 1000 zł, zwiększy się do 1100 zł. Egalitaryści, którzy w ten sposób rozumują, dodaliby, że skoro do rozdysponowania jest łącznie 300 zł, należałoby raczej obu osobom podnieść pensję o 150 zł, bo wówczas nie zmieniłaby się różnica zarobków (przed i po podwyżce byłaby równa 1000 zł), a stosunek wyższej do niższej płacy, równy 2 przed podwyżką, spadłby do poziomu 2150/1150=1.87.

Praktykę podnoszenia płac o ten sam procent dla różnych kategorii pracowników uzasadnia się w ten sposób, że po podwyżce nie zmieni się udział każdej kategorii w funduszu płac. Osoba, która zarabiała 1000 zł, a teraz zarabia 1100 zł, zarówno przed jak i po podwyżce otrzymuje 1/3 całego funduszu płac, a druga osoba w obu przypadkach pobiera pozostałe 2/3.

Jeśli nierówność społeczną rozumieć jako nierówność *względnych* udziałów w sumie dobra, wówczas współczynnik nierówności powinien przyjmować tę samą wartość dla dwu rozkładów (1000, 2000) oraz (1100, 2200) różniących się jedynie wielkością „tortu" do podziału. Takie właśnie relatywne rozumienie nierówności przyjęło się w ekonomii i dlatego na miary nierówności nakłada się jeszcze jeden warunek, eliminujący odchylenie standardowe, *niezmienniczość* ze względu na przekształcenia skal właściwe dla pomiaru stosunkowego.

•

Najprostszą miarą nierówności spełniającą wszystkie 4 postulaty (interpretacja wartości minimalnej, zasada transferów, anonimowość, niezmienniczość) jest *współczynnik zmienności* zdefiniowany jako stosunek odchylenia standardowego do średniej arytmetycznej.

$$V_x = \frac{s_x}{M_x} \qquad \text{(V)}$$

Można wykazać, że $s_x \le M_x\sqrt{n-1}$, skąd wynika, że $V_x \le \sqrt{n-1}$. Maksymalna wartość współczynnika zmienności zależy zatem od wielkości populacji. Taka zależność wydaje się pożądana, jeśli uznać, że przechwycenie całego dobra przez członka małej grupy oznacza łagodniejszą nierówność, niż w sytuacji gdy „wykluczonych" jest wielu. Z drugiej strony do porównań międzypopulacyjnych bardziej nadaje się parametr, który przyjmuje wartości nie większe od 1 na mocy samej pierwotnej definicji, nie zaś wtórnej normalizacji (podzielenia przez maksimum zależne od *n*). Takim parametrem jest najpopularniejsza miara nierówności: *współczynnik Giniego*, zdefiniowany za pomocą wzoru:

$$G_x = \frac{\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{M_x} \qquad \text{(G1)}$$

Wielkość figurująca w liczniku (wyrażająca się wzorem podobnym do podanego wyżej równoważnego określenia wariancji) to połowa średniej arytmetycznej z bezwzględnych różnic wartości zmiennej obliczonych dla wszystkich uporządkowanych par (*i*,*j*) jednostek. Jeśli jednostki zostały ponumerowanej w ten sposób, że $x_1 \le x_2 \le \cdots \le x_n$, wzór (G1) jest równoważny wzorowi (G2), znacznie ułatwiającemu obliczenie współczynnika Giniego, także z pomocą SPSS.

$$G_x = \frac{2}{M_x n^2}\sum_{i=1}^{n} i x_i - \frac{n+1}{n} \qquad \text{(G2)}$$

Obliczanie *G* w SPSS PL dla zmiennej o nazwie X (określonej dla *n* przypadków) składa się z następujących kroków: (1) Najpierw obliczamy sumę wartości zmiennej X (*Analiza→Opis statystyczny→Statystyki opisowe*; do standardowego zestawu statystyk trzeba dodać sumę, zaznaczając odpowiednią opcję); (2) Tworzymy zmienną, której wartościami będą *rangi proste* (*Przekształcenia→Ranguj obserwacje*). Zmienną tę SPSS dołącza do listy zmiennych, nadając jej nazwę RX; (3) Tworzymy pomocniczą zmienną, nazywając ją np. Z (*Przekształcenia→Oblicz wartości*), za pomocą wzoru Z= RX*X; (4) Dla zmiennej tej obliczamy średnią arytmetyczną (*Analiza→*itd.); (5) Średnią dla Z mnożymy przez 2, a wynik dzielimy przez sumę X wyznaczoną w kroku 1, po czym od ilorazu odejmujemy (n+1)/n. Kto chce uniknąć wędrowania z okna do okna, zapisywania wyników i użycia na końcu kalkulatora, może przepisać i wykonać program podany przez Górniaka i Wachnickiego (2000, s. 145).

Wariant wzoru (G2) wraz z wzorem (G1) podaje encyklopedia matematyczna dostępna w Internecie (patrz http://mathworld.wolfram.com/GiniCoefficient.html).

Wzór (G2) można wykorzystać także do dowodu, że współczynnik Giniego zachowuje zasadę transferów. Załóżmy, że elementy populacji ponumerowano tak, że $x_1 \le x_2 \le \cdots \le x_n$. Niech *y* oznacza zmienną otrzymaną z *x* przez transfer o rozmiarze *d*>0 z *i*-tej do *j*-tej jednostki, gdzie *i*<*j*, a stąd $x_i \le x_j$. Zachodzi wówczas następująca nierówność

$$G_y - G_x \ge \frac{2}{Mn^2}(j-i)d,$$

($M=M_x=M_y$), z której wynika, że po transferze wartość *G* rośnie. Ponieważ tekst ten piszę dla studentów socjologii, opuszczę niezbyt trudny, acz nieco żmudny dowód powyższej nierówności jak również dokładny wzór na różnicę G_y i G_x, który udało mi się wyprowadzić.

Według Allisona (1978: 868) „Osobliwość współczynnika Giniego polega na tym, że jego wrażliwość na transfery zależy raczej od różnicy rang jednostek niż od wartości liczbowych”. Dalej czytamy, że równość $G_y - G_x = c(j-i)d$ (*c* zależy od *M* i *n*) „łatwo dowodzi się” ze wzoru oznaczonego tu (G2). Rzeczywiście, dowód jest trywialny, o ile założyć, że transfer zachowuje porządek wartości zmiennej, wszelako bez tego założenia równość nie zachodzi, co autor przyznał, odpowiadając na moją uwagę przesłaną listem elektronicznym.

Posługując się wzorem (G2), można bez trudu wykazać, że $G_x \le 1-(1/n)$, skąd wynika nierówność $G_x<1$. Ponieważ G oblicza się zwykle dla dużych prób, wartość maksymalną w praktyce można uznać za równą 1, jednak przy małym *n* warto pamiętać, jaki jest rzeczywisty kres górny (równy $^1/_2$ dla *n*=2).

Zadanie 1. Ze wzoru (G2) wyprowadzić wzór na współczynnik Giniego dla przypadku n=2, po czym obliczyć wartość G dla opisanego wyżej przykładu płacowego przed podwyżką (x_1=1000, x_2=2000) i po podwyżce proponowanej przez egalitarystę (x'_1=1150, x'_2=2150).

•

Omówię teraz jeszcze jedno równoważne określenie współczynnika Giniego, związane z tzw. krzywą Lorenza.

Niech $x^*_1,\ldots,x^*_k$ oznaczają *różne wartości* zmiennej x, ponumerowane w porządku wzrastania ($x^*_1<\cdots<x^*_k$), a $n_1,\ldots,n_k$ odpowiadające im *liczebności*; n_j to liczba przypadków, dla których zaobserwowano wartość x^*_j. Suma wartości zmiennej dla tych przypadków równa się $n_j x^*_j$. Liczba wszystkich przypadków oraz suma wszystkich wartości zmiennej x wyrażają się wtedy wzorami:

$$n = \sum_{j=1}^{k} n_j, \; Sum\ (x) = \sum_{j=1}^{k} Sum_j, \text{ gdzie } Sum_j = n_j x_j^*.$$

Zdefiniujemy teraz *liczebności skumulowane* oraz *skumulowane sumy* wartości zmiennej za pomocą następujących wzorów

$$n_j^c = \sum_{h=1}^{j} n_h \qquad Sum_j^c = \sum_{h=1}^{j} Sum_h .$$

Tak więc j-ta liczebność skumulowana jest to liczba przypadków, dla których badana zmienna przyjęła wartość mniejszą lub równą x^*_j, natomiast j-ta suma skumulowana to suma wartości zmiennej dla tych właśnie przypadków.

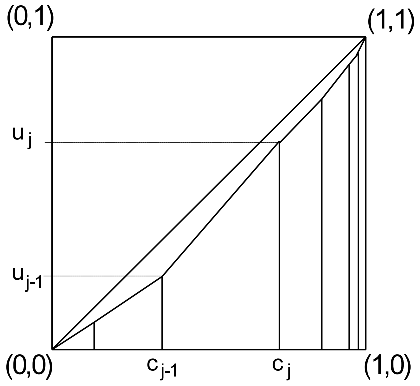
Dzieląc liczebność skumulowaną n^c_j przez liczbę wszystkich przypadków n otrzymujemy j-tą *skumulowaną częstość względną*: $c_j = n_j^c/n$. Podobnie określamy *skumulowany udział* w sumie wartości zmiennej: $u_j = Sum_j^c/Sum\ (x)$. Skumulowane częstości względne oraz skumulowane udziały tworzą ciągi rosnące: $c_1<\cdots<c_k$, $u_1<\cdots<u_k$. Zauważmy, że $c_k=u_k=1$.

Przyjmijmy dodatkowo $c_0=u_0=0$ i w dwuwymiarowym układzie współrzędnych na płaszczyźnie zaznaczmy punkty (0,0), $(c_1,u_1),\ldots,(c_{k-1},u_{k-1})$, (1,1). Punkty te leżą w kwadracie, którego bok ma długość 1, a wierzchołkami są punkty (0,0), (0,1), (1,1), (1,0). Łącząc odcinkami kolejne punkty, otrzymujemy łamaną zwaną *krzywą Lorenza*. Ponieważ $u_j<c_j$ dla $j=1,\ldots k-1$ (dowód tego faktu pomijam), łamana ta leży poniżej prostej przechodzącej przez punkty (0,0) i (1,1) zwanej *linią równego podziału* dobra, co ilustruje Rys. 1. Jeśli każdy posiada tyle samo dobra, wówczas linia ta pokrywa się z krzywą Lorenza.

Im bardziej nierówny podział dobra, tym większy obszar pomiędzy krzywą Lorenza a linią równego podziału. Stosunek pola tego obszaru do pola trójkąta o wierzchołkach (0,0), (1,0) i (1,1) może zatem służyć jako miara stopnia koncentracji dobra.

Obliczmy najpierw pole obszaru leżącego pod krzywą Lorenzą a nad osią poziomą. Obszar ten jest sumą k trapezów. Jak wiadomo, pole trapezu równe jest sumie boków równoległych pomnożonej przez połowę wysokości. Dla trapezu zaznaczonego na rysunku przez wskazanie współrzędnych pole wyraża się zatem wzorem $^1/_2(c_j-c_{j-1})(u_j+u_{j-1})$. Jeśli dodamy pola trapezów (pierwszy z nich redukuje się do trójkąta prostokątnego), sumę odejmiemy od $^1/_2$, czyli pola trójkąta o wierzchołkach (0,0), (1,0) i (1,1), a różnicę podzielimy przez $^1/_2$, dostaniemy wzór

$$G_x = 1 - \sum_{j=1}^{k} (c_j - c_{j-1})(u_j + u_{j-1}), \qquad \text{(G3)}$$



Rys. 1. Krzywa Lorenza

który, jak się okazuje, stanowi jeszcze jedną równoważną definicję współczynnika Giniego. Taką definicję podaje internetowa Wikipedia (http://en.wikipedia.org/wiki/Gini_coefficient).

•

Dla ilustracji pokażę teraz przykład liczbowy. Niech X1 oznacza zmienną określoną jako „minimalna płaca, jaką powinien otrzymywać absolwent wyższej uczelni". Pytanie o opinię na ten temat zadano w badaniach wykonanych w połowie lat 70. ubiegłego wieku na próbie złożonej z około 750 studentów 5 krakowskich uczelni. Ostatecznie w bazie przygotowanej na ćwiczenia ze statystyki znalazły się 704 przypadki.

Baza danych, zapisana pierwotnie na kartach dziurkowanych, po wczytaniu przez komputer (dostępny wówczas w międzyuczelnianym centrum obliczeniowym Cyfronet) została wydrukowana za pomocą dołączonej do niego drukarki. Gdy w 1998 roku przepisywałem tę bazę z papieru do pliku komputerowego (w celu zademonstrowania studentom zastosowania SPSS do obliczeń), nie udało mi się odczytać kilkunastu rekordów z wyblakłego wydruku. Ponadto, aby zapewnić jednorodność populacji, odrzuciłem nieliczne przypadki, w których badani podawali bardzo duże liczby (pensje powyżej 10000 ówczesnych złotych).

W zbiorze tym zmienna X1 przyjęła wartości w zakresie od 1.2 (1200 zł) do 7.0 (7000 zł), jednak poza przedziałem [2.0,5,0] znalazło się tylko 2% przypadków (jako ciekawostkę podam, że na stanowisku starszego asystenta zarabiałem wówczas 3600 zł). Średnia arytmetyczna wyniosła 2.98 a odchylenie standardowe 0.72. Tak więc współczynnik zmienności jest równy 0.72/2.98=0.24. Współczynnik Giniego obliczony za pomocą SPSS w sposób wyżej opisany wyniósł 0.127.

Zadanie 2. Ze wzoru (G2) obliczyć V i G dla zmiennych X1 („minimalna płaca po studiach") i X2 („maksymalna płaca po studiach") w 20-elementowej próbie losowej (przydzielonej każdemu na zajęciach).

W zbiorowości liczącej 704 jednostki zmienna X1 przyjęła 29 różnych wartości, jednak aż 90% zapytanych o pożądaną minimalną płacę po studiach, podało 7 „okrągłych" wartości: 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0. Zbiorowość złożona z tych 636 osób posłuży nam do prezentacji techniki obliczania współczynnika Giniego przy użyciu wzoru (G3). Liczby podane w kolumnach (c_j) i (u_j) Tabeli 1 wykorzystane zostały także do zilustrowania krzywej Lorenza na Rys. 1.

Najpierw obliczamy sumy mnożąc wartości zmiennej w kolumnie (x^*_j) przez ich liczebności podane w kolumnie (n_j). Po dodaniu sum zapisanych w kolumnie (Sum_j) wyznaczamy sumy skumulowane (Sum^c_j) oraz udziały (u_j). Na koniec wypełniamy dwie ostatnie kolumny (u_j+u_{j-1}) i ($n_j(u_j$+$u_{j-1})$).

Następnie sumę ostatniej kolumny dzielimy przez *n*, otrzymując 560.093/636=0.881. Odejmując tę liczbę od 1, dostajemy współczynnik Giniego równy 0.119. Jest to wartość nieznacznie mniejsza od obliczonej z danych surowych dla 704 jednostek.

Tabela 1. Obliczanie współczynnika Giniego dla zmiennej skokowej z użyciem wzoru (G3)

j	x^*_j	n_j	n^c_j	Sum_j	Sum^c_j	c_j (%)	u_j (%)	u_j+u_{j-1}	$n_j(u_j+u_{j-1})$
1	2.0	79	79	158.0	158.0	12.4	8.3	0.083	6.557
2	2.5	146	225	365.0	523.0	35.4	27.5	0.358	52.268
3	3.0	245	470	735.0	1258.0	73.9	66.0	0.935	229.075
4	3.5	78	548	273.0	1531.0	86.2	80.4	1.464	114.192
5	4.0	61	609	244.0	1775.0	95.8	93.2	1.736	105.896
6	4.5	10	619	45.0	1820.0	97.3	95.5	1.887	18.870
7	5.0	17	636	85.0	1905.0	100.0	100.0	1.955	33.235
		636		1905.0					560.093

Zadanie 3. Allison (1978: 868) twierdzi, że „Dla rozkładu dochodów o typowym kształcie, indeks Giniego wykazuje większą wrażliwość na transfery w obrębie środka rozkładu niż na transfery pomiędzy bardzo bogatymi bądź bardzo biednymi”. Zmodyfikujmy rozkład przedstawiony w Tabeli 1 w ten sposób, że dla 27 jednostek, które podały wartość 3.0, wykonujemy transfer rozmiaru 0.5 na korzyść 27 jednostek, które podały wartość 3.5. Po tej operacji liczba przypadków o wartości 2.5 będzie równa 146+27=173, liczba przypadków o wartości 3.0 spadnie do poziomu 245–27=218. Spadnie także o 27 liczba przypadków o wartości 3.5, osiągając liczebność 78–27=51, wzrośnie natomiast do poziomu 88=61+27 liczba przypadków o wartości 4.0. Nie zmieni się liczebność „najbiedniejszych” (wartość 2.0) i „najbogatszych” (wartości 4.5 i 5.0).

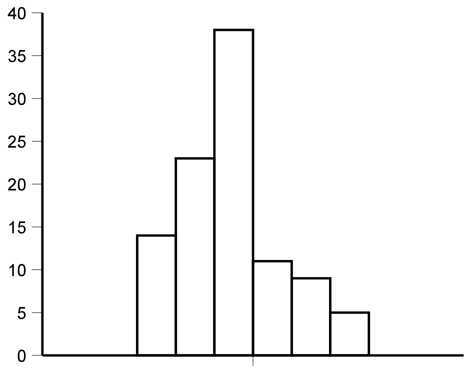
Rozważmy z kolei inną modyfikację rozkładu polegającą na transferach rozmiaru 0.5 w obrębie prawego „ogona” rozkładu: 27 z 61 jednostek o wartości 4.0 oddaje 0.5 punktu 27 jednostkom wartości ach 4.5 i 5.0. Nowe liczebności będą wtedy równe: 3.5: 78+27=105, 4.0: 61–27=34, 5.0: 10, 5.5: 17. Znika grupa 4.5, a grupy 2.0, 2.5 i 3.0 zachowują dotychczasowe liczebności.

Obliczyć współczynnik Giniego dla dwu nowych rozkładów, wypełniając tabelę analogiczną do Tabeli 1 (zainteresowani zaliczeniem na ocenę co najmniej dobrą niech narysują także krzywe Lorenza).

Choć zmienna X1 przyjmuje skokowe wartości dla większości przypadków, jej rozkład można badać także w sposób przyjęty dla *zmiennych ciągłych*, w szczególności można skonstruować *przedziały klasowe* i zilustrować rozkład za pomocą *histogramu*.

Tabela 2. Obliczanie współczynnika Giniego z danych pogrupowanych

j	x	n_j	n^c_j	Sum_j	Sum^c_j	c_j (%)	u_j (%)	u_j+u_{j-1}	$n_j(u_j+u_{j-1})$
1	–2.25	95	95	187.3	187.3	13.5	8.9	0.089	8.455
2	2.25–2.75	163	258	407.4	584.7	36.6	28.4	0.373	60.799
3	2.75–3.25	270	528	805.7	1400.4	75.0	66.8	0.952	257.040
4	3.25–3.75	80	608	280.2	1680.6	86.4	80.2	1.470	117.600
5	3.75–4.25	63	671	252.0	1932.6	95.3	92.2	1.724	108.612
6	4.25–	33	704	163.6	2096.2	100.0	100.0	1.922	63.426
		704		2096.2					615.932



Rys. 2. Ilustracja rozkładu w Tabeli 2.

W Tabeli 2 zastosowano przedziały o długości 0.5 i rozmieszczono je tak, by środki przedziałów pokrywały się z wartościami najczęściej wskazywanymi, gdyż wówczas środki będą się niewiele różnić od wartości średnich zmiennej w przedziałach (podobnie jest w sytuacji, gdy obserwacje rozkładają się równomiernie w przedziale). Przedział pierwszy i ostatni pozostawiono otwarte odpowiednio od dołu i od góry.

Do obliczenia współczynnika Giniego z danych pogrupowanych potrzebna jest znajomość sumy ogólnej (aby ją wyznaczyć wystarczy znać średnią arytmetyczną i liczebność populacji) oraz sum wartości zmiennej w przedziałach.

Jeśli nie znamy tych sum (np. gdy w raporcie z badań wykonanych przez kogoś innego podany jest tylko rozkład dochodów w przedziałach), możemy je oszacować, mnożąc środki przedziałów przez liczebności. Środek przedziału można wyznaczyć tylko wtedy, gdy znane są końce przedziału. Jeśli tylko jeden przedział skrajny ma nieokreśloną dolną/górną granicę, wówczas odpowiednią sumę dostaniemy, odejmując od sumy ogólnej sumę sum dla pozostałych przedziałów. Jeśli oba przedziały skrajne są półotwarte, proponuję wybrać jako reprezentanta pierwszego przedziału liczbę otrzymaną przez odjęcie od górnej granicy połowy długości drugiego przedziału.

Dalsze obliczenia przebiegają tak samo jak dla zmiennej skokowej, Sumę w ostatniej kolumnie dzielimy przez *n*, otrzymując w naszym przykładzie: 615.932/704=0.875, a stąd G=0.125. Jest to liczba o 0.002 mniejsza od wartości obliczonej z danych surowych.

•

Współczynnik Giniego niektórzy socjologowie skłonni są stosować także wtedy, gdy zmienna przyjmuje wartości nieujemne, nie dające się jednak interpretować jako przydziały pewnego dobra przekazywalnego. Co miałby oznaczać transfer dla zmiennej takiej jak wiek lub status? Co do wieku, mamy przynajmniej zapewnioną mierzalność na skali stosunkowej, lecz dla statusu pojęcie zera absolutnego nie ma znaczenia, co więcej, sama mierzalność tej zmiennej na skali mocniejszej niż porządkowa wydaje się problematyczna.

Blau zignorował ten problem, dopuszczając stosowanie współczynnika Giniego także w tym przypadku, chciał bowiem nadać sens ilościowy pojęciu nierówności społecznej, aby móc testować swoją teorię, a że nie znalazł miernika nierówności dostosowanego do słabszych poziomów pomiaru, zdecydował się na najbardziej popularny parametr, idąc śladem wielu socjologów, sądzących, że do obliczeń potrzebne są tylko liczby, a czas na interpretację przyjdzie wtedy, gdy zastosowanie parametru umożliwi wykrycie jakichś niebanalnych prawidłowości (mój stosunek do tej praktyki jest raczej tolerancyjny niż purystyczny).

Jak już wiemy, współczynnik nierówności, który spełnia warunek niezmienniczości ze względu na przekształcenia zmiennej postaci $y=ax$, gdzie $a>0$, nie mierzy zróżnicowania bezwzględnych kwot dobra przydzielonych jednostkom, lecz zróżnicowanie udziałów w puli niezależnie od jej rozmiaru. Dla V i G wyrażają to wzory $V_x = V_{x/Sum(x)}$ i $G_x = G_{x/Sum(x)}$ będące szczególnymi przypadkami wzorów: $V_x = V_{ax}$ i $G_x = G_{ax}$ dla dowolnego $a>0$.

Transformację $y=ax$ dla $a<1$ (np. $a=0.85$) można interpretować jako spadek dochodu wynikający z zastosowania *podatku liniowego* o stopie $1-a$ (np. 15%). Niezmienniczość V i G implikuje zatem ważną własność tej formy opodatkowania: po ściągnięciu podatku nierówność dochodów pozostaje na tym samym poziomie.

Zadanie 4. Czy *podatek progresywny* zmniejsza nierówność dochodową? Dla zbadania tego problemu określić hipotetyczną populację złożoną ze 100 jednostek, w której w odpowiednich proporcjach występują 3 kategorie podatników: o dochodzie niskim, średnim i wysokim (wskazać 3 częstości sumujące się do 100 oraz 3 liczby z przedziału [10,50] jako wartości zmiennej „dochód"). Dla każdej kategorii zaproponować stopę podatku (w zakresie od 10% do 50%) tak, by spełniony był warunek progresywności (im wyższy dochód tym wyższa stopa podatku), a następnie obliczyć współczynnik Giniego dla rozkładu dochodów przed i po opodatkowaniu. Osoby, które znają jakiś język programowania, niech spróbują napisać program wykonujący obliczenia dla dowolnego zestawu 9 liczb spełniającego warunki zadania.

Zauważmy jeszcze, że zwiększenie każdej osobie jej aktualnego stanu posiadania dobra o identyczną kwotę $c>0$ (wskutek czego suma dobra wzrasta o nc) pociąga za sobą spadek nierówności w stosunku równym M_X/M_X+c. Istotnie, ponieważ $s_{X+c}=s_X$ oraz $M_{X+c}=M_X+c$, mamy $V_{x+c} = s_x/(M_x+c) = (M_x/(M_x+c))V_x$, a stąd $V_{X+c}<V_X$. Identyczne wzory można wyprowadzić dla G_X, posługując się wzorem definicyjnym (G1), w którym dodanie c do x_i i x_j nie zmienia różnicy x_i-x_j.

Zmiana postaci $y=bx+c$, gdzie $b\geq 1$ i $c>0$ (przyrost dochodu proporcjonalny do aktualnego stanu posiadania plus premia o wysokości identycznej dla każdej osoby) także zmniejsza nierówność, ponieważ $G_{bx+c}<G_{bx}=G_x$. Okazuje się jednak, że rzeczywiste procesy społeczno-gospodarcze odbiegają od tego modelu: wzrostowi ogólnego dobrobytu z reguły towarzyszy wzrost nierówności. Oznacza to, że dochód nie rośnie w jednakowym stopniu w każdej grupie.

Czy stopa wzrostu jest tym wyższa, im wyższy dochód? Nie wiem. Gdy w latach 70-tych uczono mnie makrosocjologii, obowiązywała marksistowska teoria polaryzacji struktury społecznej, która podpowiada odpowiedź twierdzącą na postawione wyżej pytanie, jednak w podręcznikach trudno było wówczas znaleźć jakieś dane empiryczne na poparcie owej teorii. Gdy 25 lat temu sam po raz pierwszy podjąłem ten temat (T. Sozański. „Zmiany strukturalne a proces polaryzacji społeczeństwa". W: *Elementy socjologii dialektycznej*.Pod red. P. Sztompki. Warszawa-Poznań 1981), moja wiedza teoretyczna, metodologiczna i empiryczna o nierówności społecznej była minimalna. Lektura odpowiedniego hasła we współczesnej *Encyklopedii Socjologii* (B. Mach. „Równość i nierówność społeczna". t.3, Warszawa 2000) niewiele zmieniła ten stan rzeczy. Więcej informacji o rozwarstwieniu dochodów w różnych krajach można znaleźć w Internecie. Tak więc (podaję za Wikipedią) w Stanach Zjednoczonych współczynnik Giniego dla dochodów w latach spisowych 1970, 1980, 1990, 2000 był równy odpowiednio: 0.394. 0.403, 0.428, 0.462. W Polsce w latach 1996–98 był równy 0.33. Zainteresowanych socjologiczną problematyką nierówności odsyłam do wspomnianego wyżej podręcznika Domańskiego. Może ktoś zechciałby przygotować referat na ten temat na podstawie samodzielnie wyszukanej literatury?

Dla parametru przyjmującego wartości z przedziału [0,1], praktycy oczekują zwykle od teoretyków podzielenia zakresu jego wartości na interwały opisane za pomocą wyrażeń: „wartości niskie", „średnie" i „wysokie". Decyzja w tej materii należy jednak raczej do użytkowników statystyki niż teoretyków. Prof. Golinowska (podaję za *Polityką* nr 46 z 19/11/2005) uważa, że dla współczynnika Giniego wartością progową, oddzielającą strefę wartości umiarkowanych od strefy wartości wysokich, jest 0.40. Gdy G przekroczy tę wartość, nierówność dochodów staje się „problemem społecznym".

Sama wartość G nie mówi wszystkiego o postaci rozkładu dochodów. Dwa rozkłady o tej samej wartości G mogą znacznie się różnić kształtem krzywej Lorenza. Krzywe Lorenza dla dwu rozkładów mogą się przecinać, wszelako gdy jedna leży pod drugą, każdy współczynnik nierówności spełniający omówione wyżej postulaty przyjmie wyższą wartość dla tego rozkładu, dla którego krzywa położona jest niżej (twierdzenie to podaję za Allisonem, który z kolei powołuje się na publikacje innych autorów). W tej sytuacji nie dziwi popularność współczynnika Giniego, preferowanego ze względu na najbardziej „intymny” związek z krzywą Lorenza, prostotę i łatwość obliczania.

•

Poza współczynnikiem zmienności V, najpoważniejszym konkurentem dla G wydaje się współczynnik Theila, oparty na funkcji entropii, wprowadzonej przez Shannona w latach 40. XX wieku w kontekście *teorii informacji i kodowania*.

Niech $p=(p_1,\ldots,p_n)$ oznacza n-wymiarowy rozkład prawdopodobieństw, czyli ciąg liczb taki, że $p_i \geq 0$ dla każdego i oraz $\sum_{i=1}^{n} p_i = 1$. Liczby te można traktować jako *prawdopodobieństwa* parami rozłącznych *zdarzeń* $A_1, \ldots A_n$, których *suma* jest *zdarzeniem pewnym*, tzn. jedno z tych zdarzeń zawsze zachodzi. W epistemologii, a także *teorii decyzji*, $p_1,\ldots,p_n$ interpretuje się jako *prawdopodobieństwa subiektywne* przypisywane przez badacza/decydenta parami wykluczającym się *hipotezom/stanom świata*. *Entropia* rozkładu p to wielkość określona wzorem

$$H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log p_i,$$

w którym podstawa logarytmu może być dowolną liczbą dodatnią, np. e, 10 lub 2 (logarytm naturalny, dziesiętny, dwójkowy). Jeśli $p_i=0$, przyjmujemy dodatkowo, że $p_i \log p_i = 0$ (funkcja logarytmiczna jest określona tylko dla liczb dodatnich). Dalej potrzebne będą dwie ważne własności funkcji H:

(1) $H(p_1,\ldots p_n) \geq 0$, przy czym $H(p_1,\ldots p_n)=0$ wtedy i tylko wtedy gdy $p_j=1$ dla pewnego j (w konsekwencji $p_i=0$ dla każdego $i \neq j$).

(2) $H(p_1,\ldots p_n) \leq \log n$, przy czym $H(p_1,\ldots p_n)=\log n$ wtedy i tylko wtedy, gdy dla każdego i: $p_i={}^1/_n$.

Dzięki tym własnościom entropię można traktować jako miarę *niepewności* wyniku *doświadczenia losowego*, a przy „subiektywnym” rozumieniu prawdopodobieństwa jako stopień niepewności badacza, który ma zdecydować, która z n konkurencyjnych hipotez ma być przyjęta jako najbardziej *wiarygodna*. Jeśli doświadczenie ma tylko jeden możliwy wynik z prawdopodobieństwem 1 lub wiadomo, która hipoteza jest prawdziwa, niepewność jest równa 0. Gdy wszystkie wyniki (hipotezy) są jednakowo prawdopodobne (wiarygodne), niepewność jest największa i równa 1, gdy jako podstawę logarytmu wziąć n.

Rozważmy najprostsze doświadczenie losowe – rzut regularną monetą – lub dylemat, jaki ma badacz (sędzia), który uważa za jednakowo wiarygodne dwie sprzeczne odpowiedzi na dane pytanie dychotomiczne (np. czy podejrzany jest sprawcą zarzucanego mu przestępstwa). Przy zastosowaniu w definicji entropii logarytmu dwójkowego mamy wówczas $H({}^1/_2,{}^1/_2)=1$. Przez otrzymanie 1 *bita informacji* rozumie się *redukcję niepewności* w takiej właśnie sytuacji.

H nie jest jedyną funkcją rozkładu prawdopodobieństw osiągającą minimum dla rozkładów skupionych w jednym punkcie, a maksimum dla *rozkładu równomiernego*.

Inna taką funkcją, w statystyce znajdującą zastosowanie m.in. do konstrukcji miar siły zależności dla zmiennych nominalnych, jest funkcja dana prostszym wzorem: $\sum p_i(1-p_i)$. Zainteresowanych tym tematem, a nie lękających się matematyki, odsyłam do mojego artykułu ("Measures of Association for Nominal Variables." W: *Problems of Formalization in the Social Sciences*. Pod red. K. Szaniawskiego. Ossolineum 1977). W teorii informacji stosuje się miarę niepewności opartą na funkcji logarytmicznej ze względu na *addytywność* entropii dla rozkładów *niezależnych*. Dla wyjaśnienia rozważmy dwa rozkłady prawdopodobieństw, n-wymiarowy $p=(p_1,\ldots p_n)$ i m-wymiarowy $q=(q_1,\ldots q_m)$ i utwórzmy z nich rozkład nm-wymiarowy r, w którym prawdopodobieństwa dane są wzorem $r_{ij}=p_iq_j$. Addytywność entropii oznacza, że $H(r)=H(p)+H(q)$.

Po tym przygotowaniu nietrudno domyślić się, jak będzie wyglądać konstrukcja współczynnika nierówności Theila. Pomysł, polegający na obliczeniu entropii dla rozkładu p takiego, że $p_i=x_i/Sum(x)$, opiera się jedynie na *formalnej* analogii między n-wymiarowymi rozkładami prawdopodobieństwa a relatywnymi podziałami puli zasobów, nie ma jednak głębszego związku z *teorią informacji*. Tak określony współczynnik przyjmuje wartość *maksymalną* (równą log n) wtedy i tylko wtedy gdy $p_i=1/n$, czyli gdy $x_i=Sum(x)/n=M_x$, jest więc raczej miarą równości, skoro najwyższą wartość przyjmuje dla równego podziału. Aby otrzymać miarę nierówności, wystarczy jednak zastosować przekształcenie odwracające porządek: $T_x=H(x/Sum(x))-\log n$. Kto zna podstawowe własności logarytmu, łatwo już stąd wyprowadzi podany niżej wzór, za pomocą którego Theil zdefiniował współczynnik T:

$$T_x=\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{M_x}\log\frac{x_i}{M_x} \qquad \text{(T)}$$

Jego znormalizowaną wersję otrzymuje się, dzieląc T_x przez log n. Operacja ta, uważana przez wynalazcę za opcjonalną, wydaje się pożądana, gdyż nie tylko wprowadza maksimum równe 1 niezależne od n, lecz znosi równocześnie zależność parametru od arbitralnie wybranej podstawy logarytmu.

•

Post scriptum. Już po napisaniu tego tekstu zapoznałem się komentarzem Guillerminy Jasso do artykułu Allisona i repliką autora (G. Jasso. "On Gini's Mean Difference and Gini's Index of Concentration." *American Sociological Review* 44, 1979: 867–870; P. Allison. "Reply to Jasso." *Idem*: 870–872). Jasso (s. 869) także wytknęła Allisonowi błąd, o którym pisałem wyżej (uwaga zamieszczona na dole strony 4), zaś Allison (s. 871) przyznał jej rację w tym punkcie.

W swoim komentarzu Jasso zaproponowała także modyfikację współczynnika Giniego polegającą na pominięciu we wzorze (G1) par uporządkowanych postaci (i,i), gdyż dla każdej takiej pary różnica wartości zmiennej x automatycznie równa się 0. Liczba wszystkich par uporządkowanych (i,j), dla których trzeba zsumować bezwzględne różnice $|x_i-x_j|$ będzie wtedy równa $n^2-n=n(n-1)$ i przez tę właśnie liczbę zdaniem Jasso należy podzielić sumę, by uzyskać *średnią absolutną różnicę* wartości zmiennej. "Poprawiony" przez nią w ten sposób współczynnik Giniego (wzór (b) na s. 867) – oznaczmy go tu G' – okazuje się równy $(n/(n-1))G$, gdzie G dane jest wzorem (G1). G' pokrywa się zatem ze znormalizowaną wersją G i osiąga maksymalną wartość w tej samej sytuacji, tyle że równą 1 dla każdego n, co można uznać za plus tej propozycji. Wszelako, jak słusznie zauważył Allison, odpowiadając Jasso, taka modyfikacja ma też niepożądane konsekwencje. Po pierwsze, zaciera się związek z krzywą Lorenza. „Po drugie, wersja indeksu Giniego, podana przez Jasso, nie posiada pewnej narzucającej się własności, którą Sen (1973) nazywa aksjomatem symetrii populacji." (Allison 1979: 871).

Amartya Sen otrzymał nagrodę Nobla z ekonomii w 1998 roku przede wszystkim za badania nad nierównością ekonomiczną (*On Economic Inequality*, New York 1973), lecz doceniony został także jego wkład (odkrycie „paradoksu liberalizmu") do znanej mi bliżej „teorii wyboru społecznego". Zainteresowanych tą problematyką zapraszam na kurs „Modele formalne w polityce" (II semestr roku akademickiego 2005/2006)

Aby wyjaśnić sens tego aksjomatu, dwie populacje *n*-elementowe o identycznych rozkładach dochodów połączmy w jedną populację o 2*n* jednostkach. Po tej operacji podwojeniu ulegnie też suma dobra, gdyż każda wartość zmiennej będzie występować dwukrotnie częściej. *Częstości względne* będą jednak takie same. Postulat Sena głosi, że wówczas stopień nierówności też powinien pozostać niezmieniony. Oryginalny współczynnik Giniego zachowuje się w ten sposób,co wynika ze wzoru (G3), w którym $c_j - c_{j-1} = n_j/n$, $u_j = Sum^c{}_j / Sum(x)$. Połączenie dwu populacji spowoduje podwojenie n_j, n, Sum_j, $Sum^c{}_j$ i $Sum(x)$, lecz wielkości określone jako stosunki liczebności i stosunki sum nie zmienią się!

Współczynnik G' nie spełnia postulatu Sena. Przykładowo dla rozkładu maksymalnie skoncentrowanego (0,1) mamy $G={}^1/_2$, $G'=1$, a dla połączenia dwu egzemplarzy takiego rozkładu, czyli rozkładu (0,0,1,1), mamy $G={}^1/_2$, lecz $G'=({}^4/_3)({}^1/_2)={}^2/_3$.

Na zakończenie, do wszystkich, którzy znajdą ten tekst w Internecie, a mają większą ode mnie wiedzę i orientację w literaturze przedmiotu, kieruję prośbę o nadsyłanie uwag i informacji bibliograficznych, które pomogłyby mi ulepszyć wykład, a ewentualnie przygotować artykuł nadający się do druku.

http://www.cyf-kr.edu.pl/~ussozans/