PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1163 — 2007
Statystyka w praktyce społeczno-gospodarczej

Walenty Ostasiewicz
Akademia Ekonomiczna we Wrocławiu

MODELE STATYSTYCZNE BADAŃ SONDAŻOWYCH

1. Wstęp

Nauki społeczne przeżywają obecnie w Polsce okres bujnego rozkwitu. Pojawia się coraz więcej publikacji. Niestety, ilość nie idzie w parze z jakością. Ta druga wyraźnie nie nadąża za pierwszą. Jakość wymaga rozwagi, a ta z kolei potrzebuje czasu. Żyjemy przecież w pośpiechu i przyśpieszamy niemal wszystko, jak uświadamia nam J. Gleick w swojej książce *Szybciej*. Czasu na rozwagę więc brakuje, a pisać oraz publikować trzeba i wypada, jest przecież m.in. zapotrzebowanie społeczne na literaturę o badaniach społecznych.

Jedną z głównych metod badawczych stosowanych od dawna jest metoda badań sondażowych. W języku angielskim określa się je mianem *survey research*. Dawniej w Polsce określano je mianem *lustracji społecznych*. Niestety, ten piękny termin, wprowadzony niegdyś przez F. Znanieckiego, został ostatnio tak zohydzony przez niemądrych polityków, że nikt nie zechce go używać jako pojęcia naukowego. W zamian stosowane są przeróżne dziwolągi językowe stanowiące bezmyślne przeróbki słów angielskich. Coraz częściej w publikacjach w języku polskim, wydawanych nawet przez Wydawnictwo Naukowe PWN, stosowane są określenia świadczące o takim zubożeniu kulturowym języka, które wymagałoby przeprowadzenia specjalnych badań społecznych w celu zatrzymania korozji języka polskiego. Terminy z literatury anglojęzycznej importowane są zupełnie bezmyślnie bez dostosowania ich do pisowni polskiej, a także do wymowy. Badania sondażowe określane są mianem „surveye". Pytania zawarte w ankiecie nazywa się „pozycjami" lub „itemami", a formaty odpowiedzi na nie, typu: „tak", „nie", „nie wiem", nazywa się „skalami itemizowanymi" [10]. Z kolei różne aspekty zjawiska społecznego lub psychologicznego określa się mianem „dymensja" – tu przynajmniej wiadomo, jak to słowo trzeba wymówić. Równie niesamowite są definicje tych terminów. W pracy [15] podano np., że „przez badanie surveyowe, ogólnie mówiąc, rozumie się systematyczne studium życia zbiorowego oparte na ilościowej charakterystyce zjawisk społecznych oraz na interrogatywnym (poprzez zadawanie pytań) zbieraniu informacji od wytypowanych w tym celu jednostek (respon-

dentów)". Pewnym typem takich badań, który „nosi wszelkie znamiona kwestionariuszowych badań surveyowych łącząc w sobie także elementy badań pollingowych" są badania trackingowe [15].

2. Pojęcia podstawowe

Ograniczmy się do intuicyjnego rozumienia wyjaśnienia najważniejszych pojęć dotyczących badań sondażowych. Potraktujmy te badania jako metodę zbierania danych za pomocą kwestionariuszy. Kwestionariusz lub ankieta jest to zestaw pytań kierowanych do osób z prośbą o udzielenie odpowiedzi. Osoby odpowiadające nazywane są respondentami. Pytania zawarte w ankiecie mogą mieć gramatyczną formę zdań pytających, na przykład „ile ma Pani lat?", „czy jest Pan dumny z osiągnięć swoich dzieci?" itp. Mogą to być także pytania formułowane w postaci zdań kategorycznych. Na przykład „wojna humanitarna jest to najgłupsze określenie zabijania ludzi w Jugosławii wymyślone przez polskiego ministra", „Stany Zjednoczone są krajem miłującym pokój" itp. Stwierdzenia tego typu traktujemy również jako pytania, gdyż respondent jest proszony o udzielenie *odpowiedzi* w postaci aprobaty lub dezaprobaty dotyczącej tego typu stwierdzeń.

W ankiecie podawany jest zwykle sposób odpowiedzi nazywany formatem odpowiedzi, np. przez wyszczególnienie wariantów: „zgadzam się", „nie zgadzam się", lub też w przypadku więcej niż 2 wariantów: „całkowicie się zgadzam", „zgadzam się", „nie zgadzam się", „absolutnie się nie zgadzam" itp.

W niektórych przypadkach możliwe warianty odpowiedzi można uporządkować ze względu na charakter stwierdzeń, których one dotyczą. Warianty odpowiedzi są wówczas kodowane za pomocą kolejnych liczb naturalnych. Format udzielania odpowiedzi nazywa się czasem skalą. Może to prowadzić do poważnych nieporozumień i błędnych interpretacji, szczególnie wówczas gdy wybór wariantu odpowiedzi na pytanie niektórzy niepoprawnie nazywają skalowaniem jednowymiarowym. Pojęcie „skala" ma wiele znaczeń, przy tym bardzo różnych, czasem niemających ze sobą prawie nic wspólnego poza samą nazwą.

W tzw. reprezentacyjnej teorii pomiaru skalą nazywa się homomorfizm empirycznego systemu relacyjnego w system teoretyczny (liczb rzeczywistych). Jeżeli homomorfizm jest określany z dokładnością do dowolnego ściśle rosnącego przekształcenia, to skalą nazywa się skalę porządkową. Tę definicję ścisłej, matematycznej teorii pomiaru często bezkrytycznie stosuje się w wielu przypadkach, w których w ogóle nie można w sposób rozsądny określić pojęcia homomorfizmu.

W przypadku kwantyfikacji cech psychicznych oraz społecznych przez pojęcie skali rozumiane są bardzo różne rzeczy. Bardzo często skalą nazywa się zestaw wszystkich pytań zawartych w jednej ankiecie. Skalą nazywa się też zestaw wariantów odpowiedzi na jedno pytanie. Dość często skalą nazywa się całą procedurę kwantyfikacji określonej cechy lub zjawiska. W tym przypadku skalę utożsamia się z pojęciem narzędzia pomiarowego, nazywanym także instrumentem pomiarowym. Wreszcie, najbardziej zgodnie z intuicyjnym rozumieniem, skalą

nazywa się podziałkę na osi liczbowej lub całe kontinuum liczbowe. Skalowaniem czegoś nazywa się wówczas umieszczanie tego czegoś na tej osi liczbowej. Jeśli skalowaniem objęci są respondenci, to każdemu respondentowi przyporządkowywana jest pewna liczba, czyli jest on „umieszczony" na skali w miejscu odpowiadającym tej liczbie. Zamiast respondentów mogą być skalowane rzeczy lub zjawiska.

Należy bardzo wyraźnie podkreślić to, że skalowanie jest to pewien szczególny sposób pomiaru. Z tego względu dotyczy on zawsze jakiejś cechy. Zwykle jest to cecha, której istnienie jest jedynie postulowane. Cechy takie nie są oczywiście bezpośrednio obserwowalne. Przykładami tego typu cech są: pacyfizm, szowinizm, sarmatyzm itd.

W niniejszym artykule będzie rozpatrywana szczegółowo cecha nazwana wrażliwością na zło społeczne. Do skalowania takich cech stosowane są modele statystyczne odpowiedzi ankietowych. Tu omówiono bardzo szczegółowo dwa typy takich modeli.

3. Sformułowanie problemu i specyfika modelu

Załóżmy, że chcemy zbadać, co najbardziej irytuje Polaków w sposobie rządzenia krajem. To znaczy chcielibyśmy się dowiedzieć, z czego obywatele Rzeczypospolitej są najbardziej niezadowoleni oraz jaki jest stopień tego niezadowolenia. Najlepszym sposobem uzyskania takiej wiedzy jest zwrócenie się do społeczeństwa z prośbą o opinię na interesujący problem.

Przeprowadźmy więc badanie sondażowe. Wybieramy w tym celu w odpowiedni sposób pewną grupę osób, do których kierujemy ankietę złożoną z odpowiednio przygotowanych pytań. Przyjmijmy, że pytania przygotowane będą w postaci stwierdzeń kategorycznych. Respondenci zaś są proszeni o udzielenie odpowiedzi w postaci aprobaty lub dezaprobaty, czyli negacji tych stwierdzeń.

Załóżmy dla przykładu, że „pytania", czyli stwierdzenia są następujące:

P1 – władza wcale nie liczy się z opinią publiczną,

P2 – cynizm i obłuda to najbardziej typowe cechy trzymających władzę,

P3 – władza bardziej dba o swój interes niż o interes narodu.

Ponieważ *a priori*, tzn. przed zwróceniem się do respondenta, nie znamy jego odpowiedzi, to potraktujmy każde pytanie jako niewiadomą. Zgodnie z tradycją matematyczną oznaczmy te niewiadome symbolami Y_1, Y_2, Y_3.

Jeżeli respondent w pełni zgadza się ze stwierdzeniem Pi, $i = 1, 2, 3$, zapiszemy to w postaci $Y_i=1$. Jeśli się nie zgadza, to zapiszemy jako $Y_i = 0$.

Odpowiedzi na wszystkie 3 pytania udzielone przez i-tego respondenta traktujemy więc jako wektor zero-jedynkowy:

$$y_i = (y_{i1}, y_{i2}, y_{i3},) \quad y_{ij} \in \{0,1\}.$$

Jeżeli do respondentów skierujemy nie 3, lecz p pytań, to odpowiedzi udzielone przez i-tego respondenta będą stanowić wektor p-wymiarowy:

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{ip}).$$

Jeśli pytania te skierujemy do n respondentów, to uzyskane odpowiedzi można przedstawić w postaci macierzy:

$$\begin{bmatrix} y_{11}, y_{12}, \ldots, y_{1p} \\ y_{21}, y_{22}, \ldots, y_{2p} \\ \vdots \\ y_{n1}, y_{n2}, \ldots y_{np} \end{bmatrix}$$

gdzie y_{ij} oznacza odpowiedź udzieloną przez i-tego respondenta na j-te pytanie.

Przyjmijmy w dalszej części, że wiersze tej macierzy będą oznaczane jako y_1, y_2, ..., y_n. Mając zebrane dane, trzeba dokonać ich analizy i wyciągnąć odpowiednie wnioski.

Zastosujmy do rozwiązania tak sformułowanego problemu podejście statystyczne. Rozpatrzmy najpierw samą istotę tego podejścia. Najprościej rzecz ujmując, podejście statystyczne polega na sformułowaniu „mechanizmu", czyli skonstruowaniu maszyny, która z dużą dozą pewności mogłaby, po jej uruchomieniu, „wyprodukować" te dane, które zostały uzyskane w wyniku przeprowadzenia sondażu. Stanowczo można stwierdzić, że nie jest możliwe skonstruowanie takiej maszyny, która mogłaby wyprodukować dokładnie te same dane. Wiadomo przecież, że gdybyśmy prosili o informację innych ludzi, uzyskalibyśmy prawdopodobnie inne wyniki. A być może gdybyśmy drugi raz zwrócili się nawet do tych samych osób, to też mogliby odpowiadać inaczej niż za pierwszym razem.

Istota podejścia statystycznego polega na tym, aby skonstruować taką maszynę, której działanie nie byłoby deterministyczne, lecz mogłoby być modyfikowane losowo.

Używając języka potocznego i posiłkując się intuicją, można powiedzieć, że podejście takie polega na skonstruowaniu p egzemplarzy loterii (kół fortuny), tak aby na każdej można wygrać wielkość y_j, $j = 1, 2, \ldots, p$, ale żeby te wygrane

$$y_j,\ j = 1, 2, \ldots, p$$

były jak najbardziej zbliżone (podobne) do tych danych, które uzyskaliśmy w wyniku przeprowadzonego sondażu. Używając zaś języka statystycznego, to samo można wyrazić bardziej precyzyjnie. A mianowicie, chcemy zdefiniować taki wektor zmiennych losowych $(Y_1, Y_2, \ldots, Y_p)$, aby ich realizacje były podobne do uzyskanych odpowiedzi. Oznacza to, że chcemy, aby zarówno odpowiedź twierdząca na pytanie Y_i, czyli $Y_i = 1$, jak i odpowiedź negująca, czyli $Y_i = 0$, były traktowane jako zdarzenia losowe. Przez zdefiniowanie wektora zmiennych losowych rozumiemy tu określenie rozkładu tego wektora. Oznaczmy go symbolem $f(y_1, y_2, \ldots, y_p)$ lub bardziej zwięźle jako $f(y)$, pamiętając, że $y = (y_1, y_2, \ldots, y_p)$. Wyrażenie matematyczne (wzór) określające ten rozkład nazywamy modelem statystycznym odpowiedzi na pytanie $Y_1, Y_2, \ldots, Y_p$. Określenie tego wyrażenia

nazywa się specyfikacją modelu. Po dokonaniu specyfikacji modelu należy dokonać jego estymacji. Następnie należy sprawdzić, czy skonstruowany mechanizm (model statystyczny) „produkuje" dane zbliżone do tych, które uzyskano w sondażu. Używając języka statystycznego, mówi się, że trzeba dokonać weryfikacji modelu. W przybliżeniu oznacza to określenie szans, czyli prawdopodobieństwa tego, że dane pochodzące z sondażu mogą stanowić realizację próby losowej pochodzącej z rozkładu $f(y_1, y_2, ..., y_p)$.

Zauważmy ponadto, że odpowiedzi na poszczególne pytania są zależne zarówno od samych pytań, jak i od respondentów.

Każde pytanie zawiera bowiem w sobie zarówno ładunek emocjonalny, jak i ładunek intencjonalny, formułowane jest przecież z intencją uzyskania informacji na określony temat. Z drugiej zaś strony każdy respondent jest w mniejszym lub większym stopniu wrażliwy na badany problem społeczny. Intencjonalność zawarta w pytaniu (stwierdzeniu) może więc być różnie odbierana przez różnych respondentów. Odpowiedź na zadane pytanie można więc potraktować jako funkcję wrażliwości.

Mimo iż intuicyjnie wydaje się oczywiste, że różni ludzie mają różną wrażliwość na przejawy zła społecznego, nie jest jednak tak bardzo oczywiste to, że wrażliwość tę można wyrazić ilościowo w postaci liczb. Przyjmijmy więc *hipotezę*, że wrażliwość można skalować, tzn. osobom o różnej wrażliwości można przypisać różne wartości liczbowe na skali „wrażliwość". Ponieważ jest to tylko hipoteza, czyli przypuszczenie, to należy ją dokładnie zweryfikować. Weryfikacji takiej dokonuje się po dokonaniu estymacji modelu.

Oprócz postulatu, że wrażliwość ludzką można mierzyć, przyjmijmy jeszcze drugą hipotezę, że każde pytanie można również scharakteryzować za pomocą wielkości liczbowej. Tę wielkość liczbową będziemy traktować jako miernik stresu, jaki wywołuje odpowiedni aspekt zła społecznego. Kolejna ważna hipoteza, którą trzeba przyjąć i uzasadnić, dotyczy losowości, którą trzeba „wbudować" (wprowadzić) do formułowanego modelu. Chcemy przecież aby „$Y_i = y_i$", $y_i \in \{0,1\}$ traktowane były jako zdarzenia losowe. To trzeba jakoś uzasadnić. Istnieją 2 odmienne sposoby uzasadnienia [13; 16].

Pierwszy z nich opiera się na przekonaniu, że nawet jedna i ta sama osoba może różnie reagować na to samo pytanie. Zależy to od jej nastroju, humoru, postawy wobec samego badania itp. W takim przypadku zapis $P(Y_i = 1)$ oznacza prawdopodobieństwo, że ustalona osoba udzieli odpowiedzi twierdzącej na pytanie Y_i.

Drugi sposób uzasadnienia opiera się na przekonaniu stwierdzającym, że każda osoba jest dostatecznie konsekwentna i jeśli np. uważa, że udział Polski w wojnie irackiej jest złem, to niezależnie od nastroju i humoru zawsze odpowie „tak", ilekroć zostanie poproszona o udzielenie odpowiedzi na pytanie „czy udział Polski w wojnie irackiej jest złem?".

Zapis $P(Y_i = 1)$ interpretuje się w tym drugim przypadku jako prawdopodobieństwo zdarzenia, że nie ustalona, lecz losowo wybrana osoba udzieli odpowiedzi „tak" na pytanie Y_i. Natomiast to, czy odpowie „tak" czy też „nie", zależy od jej „wrażliwości". Przyjmijmy jako pewnik, że ludzie charakteryzują się różną wrażliwością na zło społeczne, a jako założenie, że stopień tej wrażliwości może być określony za pomocą liczby rzeczywistej.

Przyjmijmy teraz kolejną hipotezę, twierdząc, że wrażliwością ludzie zostali obdarzeni losowo. Oznacza to, że wrażliwość osoby wybranej na chybił-trafił traktujemy jako realizację zmiennej losowej. Oznaczmy ją symbolem Z, a jej rozkład jako $h(z)$.

Jeżeli wrażliwość traktujemy jako wielkość losową określającą rodzaj odpowiedzi, to zarówno odpowiedź twierdząca na pytanie Y, tzn. $Y = 1$, jak i odpowiedź negatywna, $Y = 0$, są zdarzeniami losowymi.

Niezależnie od sposobu uzasadnienia losowości odpowiedzi, stwierdzenie Y traktujemy więc jako zmienną losową zero-jedynkową o rozkładzie:

$$\pi = P(Y = 1),\ 1 - \pi = P(Y = 0). \tag{1}$$

Odpowiedź na pytanie, w danym przypadku jest to reakcja na stresor społeczny, zależy od wrażliwości respondentów, więc prawdopodobieństwo π możemy traktować jako funkcję tej wrażliwości. Liczbę π traktujemy więc jako funkcję $\pi(z)$. Czyli, precyzując powyższy rozkład zmiennej losowej Y, otrzymujemy:

$$\pi(z) = P(Y = 1|z),\ 1 - \pi(z) = P(Y = 0|z). \tag{2}$$

Funkcję $\pi(z)$ w literaturze psychometrycznej nazywa się funkcję charakterystyczną pytania. W danym przypadku można ją nazwać funkcją charakteryzującą dany stresor społeczny lub krótko funkcją charakterystyczną stresora. Skoro $\pi(z)$ oznacza prawdopodobieństwo, to wiemy, że funkcja ta jest ograniczona od dołu zerem oraz jednością od góry.

Zauważmy też, że im bardziej wrażliwa jest jakaś osoba na zło społeczne, tym łatwiej zgodzi się ze stwierdzeniem określającym jakiś aspekt tego zła. W ten sposób wnioskujemy, że funkcja $\pi(z)$ jest funkcją niemalejącą.

Jeżeli rozpatrywany stresor Y traktujemy jako zmienną losową, to jego rozkład zapisać można następująco:

$$P(Y = y|z) = \pi(z)^y (1 - \pi(z))^{1-y},\ y \in \{0,1\}. \tag{3}$$

Jeżeli danych jest p stwierdzeń określających p aspektów zła społecznego $Y_1, Y_2, ..., Y_p$ opisywanych za pomocą rozkładów

$$P(Y_1 = y_1|z),\ P(Y_2 = y_2|z), ...,\ P(Y_p = y_p|z),$$

to należy je połączyć we wspólny rozkład określający łączne prawdopodobieństwo warunkowe:

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_p = y_p|z).$$

Przyjmując kolejną istotną hipotezę o tzw. lokalnej niezależności, ten łączny rozkład definiuje się następująco:

$$g(y|z) = g(y_1, y_2, ..., y_p|z) = P(Y_1 = y_1,\ Y_2 = y_2, ..., Y_p = y_p|z) = \\ = \prod_{j=1}^{p} \pi(z)^{y_j} (1-\pi(z))^{1-y_j},\ y_j \in \{0,1\}. \tag{4}$$

Uwzględniwszy rozkład zmiennej losowej Z, określimy rozkład bezwarunkowy zmiennej losowej $(Y_1, Y_2, ..., Y_p)$ następująco:

$$f(y) = f(y_1, y_2, ..., y_p) = \int g(y|z)\, h(z)\, dz. \tag{5}$$

Określenie tego rozkładu można traktować jako określenie mechanizmu generującego dane.

Zauważmy bowiem, że dane uzyskane z obserwacji, czyli wektory postaci

$$y = (y_1, y_2, ..., y_p),$$

traktujemy jako realizacje zmiennej losowej p-wymiarowej

$$(Y_1, Y_2, ..., Y_p).$$

Jeśli więc potrafimy dobrać taki rozkład $f(y_1, y_2, ..., y_p)$ tej zmiennej, aby wyniki próby losowej z tego rozkładu były zgodne z obserwacjami, to rozkład ten nazywamy modelem statystycznym. W języku mniej technicznym, rozkład ten nazywa się mechanizmem generującym dane. Istota podejścia statystycznego polega więc na tym, aby mechanizm taki skonstruować, a następnie go „uruchomić" i analizować wyniki bez konieczności przeprowadzania realnych eksperymentów.

Wprowadzając oznaczenie

$$g_j(y_j|z) = \pi(z)^{y_j} (1-\pi(z))^{1-y_j}, \tag{6}$$

przedstawmy uzyskaną postać modelu następująco:

$$f(y) = \int g(y|z)\, h(z)\, dz = \int \prod_{j=1}^{p} g_j(y_j|z)\, h(z)\, dz. \tag{7}$$

Dalsza specyfikacja modelu polega na określeniu postaci dwóch podstawowych funkcji: $h(z)$ oraz $g(y|z)$.

Pierwsza z nich określa rozkład wrażliwości w potencjalnej populacji wszystkich możliwych respondentów, druga zaś określa wpływ tej wrażliwości na reakcję respondentów na stresory społeczne. Specyfikacja tych funkcji stanowi jedno z trudniejszych zadań tego etapu modelowania statystycznego. Trudno podać formalne, a nawet intuicyjne kryteria wyboru tych funkcji. Jako uzasadnienie podanych niżej funkcji przyjmijmy więc odwołanie się do autorytetów i do doświadczenia. Do określania funkcji $g(y|z)$ potrzebna jest znajomość funkcji $\pi(z)$.

Jako funkcję $\pi(z)$ przyjmijmy funkcję, którą po raz pierwszy zaproponował G. Rasch w 1960 r., czyli następującą funkcję [1; 18]:

$$\pi_j(z) = \frac{\exp(z-\alpha_j)}{1+\exp(z-\alpha_j)}, \tag{8}$$

gdzie α_j oznacza parametr charakteryzujący j-ty badany stresor. Funkcja $g(y|z)$ przybierze wówczas następującą postać:

$$g(y|z) = \prod_{j=1}^{p} \pi_j(z)^{y_j} (1-\pi(z))^{1-y_j} = \prod_{j=1}^{p} \left(\frac{\exp(z-\alpha_j)}{1+\exp(z-\alpha_j)} \right)^{y_j} \left(\frac{1}{1+\exp(z-\alpha_j)} \right)^{1-y_j}. \tag{9}$$

Przyjmijmy wreszcie, że zmienna losowa Z ma rozkład normalny z parametrami μ oraz σ^2, oznaczmy go jako $h(z;\mu,\sigma^2)$:

$$h(z;\mu,\sigma^2) = \left(\sqrt{2\pi}\sigma\right)^{-1} \exp\left(-\frac{1}{2}(z-\mu)^2/\sigma^2 \right).$$

Podstawiając wyspecyfikowane funkcje, otrzymujemy:

$$f(y;\alpha,\mu,\sigma^2) = \int \prod_{j=1}^{p} \left(\frac{\exp(z-\alpha_j)}{1+\exp(z-\alpha_j)} \right)^{y_j} \left(\frac{1}{1+\exp(z-\alpha_j)} \right)^{1-y_j} h(z;\mu,\sigma^2)\, dz. \tag{10}$$

W ten sposób została dokonana specyfikacja modelu statystycznego.

Modelem tym jest wyrażenie matematyczne (wzór) określające rozkład wektora losowego $(Y_1, Y_2, ..., Y_p)$. Wzór ten ma następującą ogólną postać [6; 8]:

$$f(y) = \int g(y|z)\, h(z)\, dz.$$

Sformułowanie takiego modelu oznacza przyjęcie pewnych hipotez, zwanych częściej założeniami. Podsumujmy najważniejsze z nich.

1. Sądzimy, czyli przyjmujemy hipotezę, że różne przejawy lub aspekty życia społecznego, z których ludzie nie są zadowoleni, można sformułować w postaci stwierdzeń kategorycznych, które są nazywane stresorami.

2. Każdy stresor może być scharakteryzowany za pomocą jednej liczby, którą można interpretować jako stopień irytacji, jaką on wywołuje w społeczeństwie.

3. Reakcja ludzi na różne przejawy zła społecznego zależy od pewnej wewnętrznej ich cechy, którą nazywa się tu wrażliwością.

4. Twierdzimy, że wszyscy ludzie są obdarzeni wrażliwością w sposób nam nieznany i dlatego traktujemy ją jako wielkość losową o wartościach rzeczywistych. Czyli przyjmujemy, że można ją skalować.

Przez pojęcie reakcji na zło społeczne rozumiany jest tu nie bunt, strajk itp., lecz jedynie wyrażenie opinii, ale opinii bardzo zdecydowanej, w kategoriach „tak" lub „nie", kodowanych liczbowo w postaci jedynki i zera.

Reakcja na j-ty aspekt zła społecznego zależy od wrażliwości (traktowanej jako wielkość losowa), więc sama jest wielkością losową.

Zależność reakcji od wrażliwości określana jest za pomocą rozkładu warunkowego $g(y|z)$ zmiennej losowej (wektorowej) Y pod warunkiem zmiennej losowej Z, której rozkład oznaczony jest jako $h(z)$. Wrażliwość nie jest bezpośrednio obserwowalna, obserwujemy jedynie reakcję na stresory. Rozkład zmiennej losowej charakteryzujący tę reakcję, czyli odpowiedzi na pytania, oznaczany jest jako $f(y)$. Wyrażenie określające ten rozkład nazywa się modelem statystycznym. Model taki został sformułowany na podstawie wymienianych hipotez. Jeśli hipotezy nie są prawdziwe, to model jest bezwartościowy. Oceny modelu dokonuje się więc przez weryfikację hipotez. Aby to zrobić, należy przede wszystkim „dostroić" ogólną postać modelu do zaobserwowanych danych. W języku statystycznym nazywa się to estymacją modelu. Przed omówieniem tego zagadnienia dokonajmy pewnych upraszczających przekształceń funkcji. Zauważmy, że dwa zapisy:

$$P(Y_j = 1|z) = \frac{\exp(z-\alpha_j)}{1+\exp(z-\alpha_j)} \text{ oraz } P(Y_j = 0|z) = \frac{1}{1+\exp(z-\alpha_j)}$$

można zapisać w postaci jednego wyrażenia:

$$P(Y_j = y_j|z) = \frac{\exp(y_j(z-\alpha_j))}{1+\exp(z-\alpha_j)}.$$

Wówczas wzór (4) można zapisać następująco:

$$P(Y_1 = y_1, ..., Y_p = y_p|z) = \prod_{j=1}^{p} P(Y_j = y_j|z) = \prod_{j=1}^{p} \frac{\exp(y_j(z-\alpha_j))}{1+\exp(z-\alpha_j)} =$$

$$= \frac{\exp(y_1 \cdot z + y_2 \cdot z + ... + y_p \cdot z - y_1\alpha_1 - y_2\alpha_2 - ... - y_p\alpha_p)}{\prod_{j=1}^{p}(1+\exp(z-\alpha_j))}.$$

Wprowadźmy oznaczenie $t = y_1 + y_2 + ... + y_p$. Możemy wówczas wzór powyższy przedstawić w postaci:

$$\prod_{j=1}^{p} P(Y_j = y_j|z) = \frac{\exp\left(t \cdot z - \sum_{j=1}^{p} y_j \alpha_j\right)}{\prod_{j=1}^{p} (1 + \exp(z - \alpha_j))}. \tag{11}$$

Czyli mamy następującą postać modelu określonego wzorem (10) [2; 3]):

$$f(y;\alpha,\mu,\sigma^2) = \int \frac{\exp\left(tz - \sum_{j=1}^{p} y_j \alpha_j\right)}{\prod_{j=1}^{p} (1 + \exp(z - \alpha_j))} h(z;\alpha,\mu,\sigma^2) dz. \tag{12}$$

Wielkość t określa liczbę odpowiedzi „tak" na pytania $Y_1, Y_2, ..., Y_p$, tzn. liczbę jedynek wśród składowych wektora $(y_1, y_2, ..., y_p)$. Wielkość t jest oczywiście realizacją zmiennej losowej $T = Y_1 + Y_2 + ... + Y_n$.

Rozkład zmiennej T zależy od wielkości z, dlatego też oznaczmy go jako $\upsilon(t|z)$. Rozkład tej zmiennej określić można bez większego trudu. A mianowicie

$$P(T = t|z) = P(T = y_1 + y_2 + ... + y_p|z) = \sum_{=t} P(Y_1 = y_1|z) \cdot P(Y_2 = y_2|z) ... P(Y_p = y_{p1}|z) =$$

$$= \sum_{=t} \frac{\exp\left(tz - \sum_{j=1}^{p} y_j \alpha_j\right)}{\prod_{j=1}^{p} (1 + \exp(z - \alpha_j))}. \tag{13}$$

Znak $\sum_{=t}$ oznacza sumowanie po wszystkich wektorach $(y_1, y_2, ..., y_p)$ takich, że $t = y_1 + y_2 + ... + y_p$.

Rozkład ten można też przedstawić następująco:

$$\upsilon(t|z) = P(T = t|z) = \exp(tz) \frac{\sum_{=t} \exp\left(-\sum_{j=1}^{p} y_j \alpha_j\right)}{\prod_{j=1}^{p} (1 + \exp(z - \alpha_j))}. \tag{14}$$

Dzieląc rozkład $f(y;\alpha,\mu,\sigma^2)$ przez rozkład $\upsilon(t|z)$, uzyskamy warunkowy rozkład odpowiedzi na pytania $Y_1, Y_2, ..., Y_p$, wiedząc, że było t odpowiedzi twierdzących. Rozkład ten, oznaczony jako $\varphi(y|t)$, jest następujący (por. [5]):

$$\varphi(y|t)=\varphi(y_1, y_2, ..., y_p|t)=\frac{\exp\left(-\sum_{j=1}^{p} y_j\alpha_j\right)}{\gamma(t;\alpha)} \tag{15}$$

gdzie

$$\gamma(t;\alpha)=\sum_{=t}\exp\left(-\sum_{j=1}^{p} y_j\alpha_j\right) \tag{16}$$

przy czym sumowanie rozciąga się na takie wektory, dla których zachodzi warunek $t = y_1 + ... + y_n$. Wykorzystując ten rozkład warunkowy, sformułowany model można teraz przedstawić następująco:

$$f(y;\alpha,\mu,\sigma^2)=\int\varphi(y|t)\,\upsilon(t|z)\,h(z)\,dz=\varphi(y|t)\int\upsilon(t|z)\,h(z)\,dz. \tag{17}$$

Taka postać modelu jest szczególnie wygodna do estymacji parametrów. Parametry $\alpha_1, \alpha_2, ..., \alpha_p$ można bowiem estymować niezależnie od parametrów μ oraz σ^2.

4. Estymacja

Załóżmy, że dane są obserwacje

$$y_{i1}, y_{i2}, ..., y_{ip}\ , i = 1, 2, ..., n.$$

Funkcja wiarygodności jest następująca:

$$L(\alpha,\mu,\sigma^2)=\prod_{i=1}^{n} P(Y_{i1}=y_{i1}, ..., Y_{ip}=y_{ip})=\prod_{i=1}^{n} f(y_{i1}, y_{i2}, ..., y_{ip}),$$

gdzie $f(y)$ określone jest wzorem (17).
Przedstawmy ją w następującej postaci:

$$\begin{aligned} L(\alpha,\mu,\sigma^2)&=\prod_{i=1}^{n}\int\varphi(y_{i1}, y_{i2}, ..., y_{ip}|t_i)\,\upsilon(t_i|z)\,h(z;\mu,\sigma^2)\,dz= \\ &=\prod_{i=1}^{n}\varphi(y_{i1}, y_{i2}, ..., y_{ip}|t_i)\prod_{i=1}^{n}\int\upsilon(t_i|z)\,h(z;\mu,\sigma^2)\,dz \end{aligned} \tag{18}$$

gdzie $\upsilon(t_i|z) = P(T_i = t_i|z)$, T_i zaś określone jest wzorem: $T_i = \sum_{j=1}^{p} Y_{ij}$.

Tak więc wielkość $\upsilon(t_i|z)$ oznacza prawdopodobieństwo, że i-ta osoba charakteryzująca się wrażliwością z odpowie „tak" na t_i spośród p pytań.

Wprowadzając oznaczenia

$$L_c(\alpha_1, \alpha_2, ..., \alpha_p) = \prod_{i=1}^{n} \varphi(y_{i1}, ..., y_{ip}|t_i),$$

$$L_p(\alpha_1, \alpha_2, ..., \alpha_p, \mu, \sigma^2) = \prod_{i=1}^{p} \int \upsilon(t_i|z) h(z; \mu, \sigma^2) dz,$$

funkcję wiarygodności (17) przedstawimy w postaci iloczynu

$$L(\alpha, \mu, \sigma^2) = L_c(\alpha) \; L_p(\alpha, \mu, \sigma^2). \tag{19}$$

Pierwszy czynnik iloczynu nazywa się warunkową funkcją wiarygodności, zależy ona wyłącznie od parametrów $\alpha_1, \alpha_2, ..., \alpha_p$. Funkcję tę wykorzystuje się więc do estymacji parametrów α_j. Drugi czynnik iloczynu określa funkcję wiarygodności służącą do estymacji parametrów μ oraz σ^2 charakteryzujących rozkład wrażliwości w populacji.

Wykorzystując określenie funkcji $\varphi(y|t)$, funkcję warunkowej wiarygodności można przedstawić następująco:

$$L_c(\alpha) = \prod_{i=1}^{n} \varphi(y|t_i) = \frac{\exp\left(-\sum_{j=1}^{p} y_j \alpha_j\right)}{\prod_{i=1}^{n} \gamma(t_i; \alpha)}. \tag{20}$$

Zauważmy, że każda zmienna losowa T_i może przybierać tylko następujące wartości: 0, 1, 2, ..., p, oznacza ona bowiem liczbę „punktów", jakie uzyskała i-ta osoba. Przez liczbę punktów rozumiemy liczbę odpowiedzi aprobujących. Oznaczmy symbolami $n_0, n_1, ..., n_p$ liczbę tych osób, które uzyskały odpowiednio 0, 1, ..., p punktów. Oczywiście $n_o + n_1 + ... + n_p = n$.

Funkcję wiarygodności L_c można wówczas zapisać następująco:

$$L_c(\alpha) = \exp\left(-\sum_{j=1}^{p} y_j \alpha_j\right) \left[\prod_{k=0}^{p} \gamma(k; \alpha)^{n_k}\right]^{-1}. \tag{21}$$

Jeśli liczba analizowanych stresorów p jest znacznie mniejsza od liczby respondentów n, to powyższa postać funkcji wiarygodności stanowi znaczne uproszczenie w porównaniu z pierwotną postacią. W celu określenia estymatorów $\alpha_1, \alpha_2, ..., \alpha_p$ należy rozwiązać następujący układ równań:

$$\frac{\partial \ln L_c(\alpha)}{\partial \alpha_i} = 0,\ i = 1, 2, ..., p. \tag{22}$$

Rozwiązanie tego układu równań nie jest proste. Numeryczny sposób jego rozwiązania jest podany np. w pracy [2].

Zauważmy, że do określenia tych estymatorów nie była potrzebna znajomość rozkładu wrażliwości w populacji, tzn. rozkładu zmiennej losowej Z.

Podstawę estymacji parametrów charakteryzujących tę zmienną, czyli wrażliwość społeczną na różne przejawy zła społecznego, stanowi rozkład zmiennej losowej T_i. Rozkład ten uzyskamy z rozkładu warunkowego (14) w sposób następujący:

$$v(t) = \int v(t|z) h(z;\mu,\sigma^2)\, dz = \int \exp(tz) \frac{\sum_{=t} \exp\left(-\sum_{j=1}^{p} y_j \alpha_j\right)}{\prod_{j=1}^{p} \left(1 + \exp(z - \alpha_j)\right)} h(z;\mu,\sigma^2)\, dz = \tag{23}$$

$$= \gamma(t;\alpha) \int \frac{\exp(tz)}{\prod_{j=1}^{p} \left(1 + \exp(z - \alpha_j)\right)} h(z;\mu,\sigma^2)\, dz$$

gdzie $\gamma(t;\alpha)$ jest określone wzorem (16).

Ponieważ przyjęliśmy, że zmienna losowa Z ma rozkład normalny, czyli

$$h(z;\mu,\sigma^2) = \left(\sqrt{2\pi\sigma}\right)^{-1} \exp\left(-\frac{1}{2}(z-\mu)^2/\sigma^2\right),$$

to wzór (23) przybiera następującą postać:

$$v(t) = \gamma(t;\alpha) \int \frac{\exp(tz)}{\prod_{j=1}^{p} \left(1 + \exp(z - \alpha_j)\right)} \left(\sqrt{2\pi\sigma}\right)^{-1} \exp\left(-\frac{1}{2}(z-\mu)^2/\sigma^2\right) dz. \tag{24}$$

Funkcja wiarygodności L_p ma następującą postać:

$$L_p(\alpha,\mu,\sigma_2) = \prod_{i=1}^{n} v(t_i).$$

Oznaczmy, tak jak poprzednio, symbolami $n_0, n_1, ..., n_p$ liczbę tych osób, które uzyskały odpowiednio $0, 1, ..., p$ punktów. Funkcję wiarygodności L_p można wówczas przedstawić następująco:

$$L_p(\alpha,\mu,\sigma^2)=\prod_{k=0}^{p}v(t)^{n_k}.$$

Podstawiając zamiast $v(t_i)$ wyrażenie (24) określające tę funkcję, otrzymujemy następującą postać funkcji wiarygodności:

$$L_p(\alpha,\pi,\sigma^2)=\prod_{k=0}^{p}\left[\gamma(t;\alpha)\int\frac{\exp(tz)}{K}\left(\sqrt{2\pi\sigma}\right)^{-1}\exp\left(\frac{(z-\mu)^2}{2\sigma^2}\right)dz\right]^{n_k}, \tag{25}$$

gdzie

$$K=\prod_{j=1}^{p}\left(1+\exp(z-\alpha_j)\right). \tag{26}$$

Logarytmując, otrzymujemy:

$$\ln L_p(\alpha,\pi,\sigma^2)=\sum_{k=0}^{p}n_k\ln\gamma(t;\alpha)+\sum_{k=0}^{p}n_k\ln\left(\int\frac{\exp(tz)}{K}h(z)dz\right). \tag{27}$$

Ponieważ chcemy estymować parametry μ oraz σ^2, a pierwszy składnik powyższej sumy tych parametrów nie zawiera, to przedstawmy logarytm funkcji wiarygodności następująco:

$$\ln L_p(\mu,\sigma^2)=C+\sum_{k=0}^{p}n_k\ln\left(\int\frac{\exp(tz)}{K}h(z;\mu,\sigma^2)dz\right).$$

W celu estymacji parametrów μ oraz σ^2 należy rozwiązać następujące równania:

$$\frac{\partial\ln L_p(\mu,\sigma^2)}{\partial\mu}=0,$$

$$\frac{\partial\ln L_p(\mu,\sigma^2)}{\partial\sigma^2}=0.$$

Rozpatrzmy nieco bardziej szczegółowo pierwsze z nich. Obliczmy pochodną logarytmu funkcji L_p względem parametru μ:

$$\frac{\partial \ln L_p(\mu,\sigma^2)}{\partial \mu} = 0 + \sum_{k=0}^{p} n_k \cdot \frac{1}{\int \frac{\exp(tz)}{K} h(z;\mu,\sigma^2)\, dz} \cdot \int \frac{\exp(tz)}{K} \frac{\partial h(z)}{\partial \mu}\, dz.$$

Ponieważ

$$\frac{\partial h(z;\mu,\sigma^2)}{\partial \mu} = \frac{\partial \left(\sqrt{2\pi}\,\sigma\right)^{-1} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)}{\partial \mu} = h(z;\mu,\sigma^2) \cdot \frac{z-\mu}{\sigma^2},$$

więc

$$\frac{\partial \ln L_p(\mu,\sigma^2)}{\partial \mu} = \sum_{k=0}^{p} n_k \frac{\int \frac{\exp(tz)}{K} h(z) \cdot \frac{z-\mu}{\sigma^2}\, dz}{\int \frac{\exp(tz)}{K} h(z)\, dz}.$$

W podobny sposób otrzymujemy lewą stronę drugiego równania:

$$\frac{\partial \ln L_p(\mu,\sigma^2)}{\partial \sigma^2} = \sum_{k=0}^{p} n_k \frac{A}{B},$$

gdzie A oraz B są określone następująco:

$$A = \int \frac{\exp(tz)}{K} h(z) \cdot \frac{1}{2\sigma^2} \left(\frac{(z-\mu)^2}{\sigma^2} - 1 \right) dz,$$

$$B = \int \frac{\exp(tz)}{K} h(z)\, dz.$$

Zastępując $\alpha_1, \alpha_2, ..., \alpha_p$ przez ich estymatory $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_p$, uzyskane z rozwiązania równań (22), otrzymuje się następujący układ równań:

$$\frac{\partial \ln L_p(\hat{\alpha},\mu,\sigma^2)}{\partial \mu} = 0,$$

$$\frac{\partial \ln L_p(\hat{\alpha},\mu,\sigma^2)}{\partial \sigma^2} = 0,$$

który trzeba rozwiązać względem niewiadomych μ oraz σ.

Oczywiście rozwiązanie takiego układu równań możliwe jest tylko przy zastosowaniu procedury rekurencyjnej.

5. Weryfikacja

Weryfikacja uzyskanego modelu stanowi kolejny ważny etap w procesie modelowania statystycznego.

Istotę weryfikacji można przedstawić bardzo prosto. A mianowicie trzeba „uruchomić" skonstruowaną maszynę statystyczną i zobaczyć, czy „wyprodukowane" przez nią dane są dostatecznie bliskie danym zebranym w wyniku przeprowadzonego sondażu.

Dane zebrane w wyniku sondażu stanowią następującą macierz:

$$\begin{bmatrix} y_{11}, y_{12}, \ldots, y_{1p} \\ y_{21}, y_{22}, \ldots, y_{2p} \\ \vdots \\ y_{n1}, y_{n2}, \ldots, y_{np} \end{bmatrix}$$

W celu ilustracji załóżmy, że do 7 respondentów skierowano 3 pytania, na które uzyskano następujące odpowiedzi:

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Przedstawmy je w następującej tabeli

Odpowiedzi na pytania	Liczba respondentów	Względna częstość odpowiedzi
0 0 0	1	1/7
1 0 0	2	2/7
1 1 0	1	1/7
0 1 1	2	2/7
1 1 1	1	1/7

Powiemy, że skonstruowany model dobrze pasuje do zaobserwowanych danych, jeśli według jego wskazań podane typy odpowiedzi powinny się pojawiać z mniej więcej takimi prawdopodobieństwami jak zaobserwowane względne częstości.

Prawdopodobieństwa te oblicza się z uzyskanego modelu. W przypadku tego ilustracyjnego przykładu prawdopodobieństwa te oblicza się z następujących wzorów:

$$P(Y_1 = 0, Y_2 = 0, Y_3 = 0) =$$

$$\frac{1}{\sqrt{2\pi}\hat{\sigma}} \int_{-\infty}^{+\infty} \frac{1}{1+\exp(z-\hat{\alpha}_1)} \cdot \frac{1}{1+\exp(z-\hat{\alpha}_2)} \cdot \frac{1}{1+\exp(z-\hat{\alpha}_3)} e^{-\frac{(z-\hat{\mu})^2}{2\hat{\sigma}^2}} dz,$$

$$P(Y_1 = 1, Y_2 = 0, Y_3 = 0) =$$

$$\frac{1}{\sqrt{2\pi}\hat{\sigma}} \int \frac{\exp(z-\hat{\alpha}_1)}{1+\exp(z-\hat{\alpha}_1)} \cdot \frac{1}{1+\exp(z-\hat{\alpha}_2)} \cdot \frac{1}{1+\exp(z-\hat{\alpha}_3)} e^{-\frac{(z-\hat{\mu})^2}{2\hat{\sigma}^2}} dz.$$

Po obliczeniu wszystkich prawdopodobieństw tego typu należy je porównać z odpowiednimi częstościami. Prostszy sposób polega na wykorzystaniu statystyki, która nazwana była liczbą punktów uzyskanych przez respondenta.

Przedstawmy zaobserwowane liczby punktów w postaci wektora.

$$[n_0, n_1, ..., n_p].$$

W przypadku przykładu ilustracyjnego mamy [1 2 3 1]. Zinterpretujmy poszczególne składowe:

$n_0 = 1$ oznacza, że jedna osoba uzyskała 0 punktów,

$n_1 = 2$ oznacza, że 2 osoby uzyskały po 1 punkcie,

$n_2 = 3$ oznacza, że 3 osoby uzyskały po 2 punkty (są to odpowiedzi (1,1,0), (0,1,1) oraz (0,1,1)),

$n_3 = 1$ oznacza, że 1 osoba uzyskała 3 punkty, czyli odpowiedziała twierdząco na każde z trzech pytań.

Aby określić liczbę n_k wynikającą z modelu, trzeba obliczyć prawdopodobieństwo $\upsilon(k) = P(T = k)$ oraz pomnożyć przez liczbę n, tzn. liczbę wszystkich respondentów. Oznaczmy symbolami $\hat{n}_0, \hat{n}_1, ..., \hat{n}_p$ liczby produkowane przez skonstruowany mechanizm, czyli

$$\hat{n}_k = n \cdot \upsilon(k), \quad k = 0, 1, ..., p.$$

Miarę bliskości dwóch wektorów:

$$[n_0, n_1, ..., n_p]$$

oraz

$$[\hat{n}_0, \hat{n}_1, ..., \hat{n}_p],$$

określa się za pomocą następującego wzoru:

$$q = \sum_{k=0}^{p} \frac{[n_k - n \cdot \upsilon(k)]^2}{n \cdot \upsilon(k)}. \tag{28}$$

Z teorii statystyki wiadomo, że wartość q jest to zaobserwowana wartość zmiennej losowej *chi-kwadrat* z p-2 stopniami swobody. Oznaczmy tę zmienną losową jako $X^2_{p=i}$. Jeśli okaże się, że $P(X^2_{p-2} > q)$ jest duże, np. wynosi więcej niż 0,05, to można uważać, że model został dobrze skonstruowany. Używając technicznego języka statystycznego, mówi się, że nie ma istotnych powodów, aby twierdzić, że skonstruowany model źle „odtwarza" zaobserwowane dane.

6. Skalowanie wrażliwości

Skalowanie jest to szczególny sposób pomiaru cech, a więc jest to pewien sposób przyporządkowywania liczb obiektom ze względu na określaną cechę. W rozpatrywanym problemie cechą jest wrażliwość społeczna. Przyjęliśmy, że jest to cecha jednowymiarowa. Jeśli model skonstruowany na podstawie tego założenia został zweryfikowany, to można przystąpić do skalowania.

W konstrukcji modelu przyjęliśmy, że wrażliwość jest wielkością losową o rozkładzie *a priori* $h(z)$. Od wielkości tej, czyli od stopnia wrażliwości, zależą odpowiedzi na pytania zawarte w ankiecie. Zależność ta została określona w postaci rozkładu warunkowego $g(y|z)$. Na podstawie tych dwóch rozkładów określony został rozkład wyników odpowiedzi:

$$f(y) = \int g(y|z)\, h(z)\, dz. \tag{29}$$

Określmy teraz rozkład *a posteriori* wartości określających stopień wrażliwości. Korzystając z twierdzenia Bayesa, rozkład ten określa się ze wzoru [9]:

$$h(z|y) = \frac{g(y|z)\, h(z)}{f(y)}. \tag{30}$$

Rozkład ten stanowi podstawę do skalowania wrażliwości. A mianowicie jako wielkość wrażliwości można przyjąć dowolny parametr położenia uzyskanego rozkładu *a posteriori*. W szczególności może to być wartość modalna. Jeśli odpowiedzi i-tego respondenta oznaczymy jako $y_i = (y_{i1}, y_{i2}, ..., y_{ip})$, to „miejsce" tego respondenta na skali wrażliwości, oznaczane jako w_i, określimy następująco:

$$w_i = \arg\max_z h(z|y_i).$$

7. Model alternatywny

Skonstruowany w poprzednich punktach model statystyczny jest tylko jednym z możliwych. Rozpatrzmy dla przykładu nieco inne podejście do analizy tego samego problemu, a mianowicie analizy stresorów społecznych. Przyjmijmy tym

razem, że grupa n respondentów jest *ustalona*. Zakładamy przy tym, że każdy respondent może różnie reagować na ten sam stresor w różnych okolicznościach.

Wysłanie polskich żołnierzy na wojnę poza granicami kraju raz może np. wywołać silny sprzeciw, innym razem ten sam respondent może się do tego problemu odnieść zupełnie obojętnie.

Poprzednio przyjmowaliśmy, że każdy respondent jest konsekwentny i za każdym razem odpowiada jednakowo na ten sam aspekt zła społecznego.

Losowość zdarzenia „$Y = 1$" wynikała z losowego wyboru respondentów. Tym razem zakładamy więc, że losowość „zawarta" jest w każdym respondencie. Zapis $P(Y = 1)$ oznacza więc prawdopodobieństwo, że ustalony respondent odpowie twierdząco na pytanie Y. Dlatego też w tym przypadku przyjmujemy, że i-ty respondent charakteryzuje się losową wrażliwością zależną od pewnego *parametru* liczbowego, który oznaczymy symbolem z_i, $i = 1, 2, ..., n$. Zauważmy, że w poprzednim modelu wielkość z była traktowana jako zmienna losowa. Natomiast każdy aspekt zła w tym modelu, tak jak poprzednio, scharakteryzujemy za pomocą parametru α_j, $j = 1, ..., p$.

Prawdopodobieństwo udzielenia odpowiedzi aprobującej stwierdzenie j-tego aspektu zła przez i-tego respondenta określmy za pomocą wzoru:

$$\pi_{ij} = \pi_j(z_i) = P\left(Y_{ij} = 1 \middle| z_i, \alpha_j\right) = \frac{\exp\left(z_i - \alpha_j\right)}{1 + \exp\left(z_i - \alpha_j\right)}. \tag{31}$$

Czyli w ogólnym przypadku mamy następujący model:

$$P\left(Y_{ij} = y_{ij}\right) = \pi_{ij}^{y_{ij}}\left(1 - \pi_{ij}\right)^{1-y_{ij}}, \quad y_{ij} \in \{0,1\}. \tag{32}$$

Tym razem trzeba estymować następujące parametry:

$$\alpha_1, \alpha_2, ..., \alpha_p, z_1, z_2, ..., z_n.$$

Można to uczynić, korzystając z metody największej wiarygodności.

Funkcja wiarygodności przybierze w tym przypadku następującą postać:

$$\begin{aligned} L\left(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p\right) &= \prod_{i=1}^{n}\prod_{j=1}^{p}\pi_i(z_i)^{y_{ij}}\left[1 - \pi_j(z_i)\right]^{1-x_{ij}} = \\ &= \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{p} z_i y_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{p}\alpha_i y_{ij}\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{p}\left(1 + e^{z_i - \alpha_j}\right)\right)^{-1} = \\ &= \exp\left(\sum_{i=1}^{n} z_i t_i - \sum_{j=1}^{p}\alpha_j q_j\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{p}\left(1 + \exp\left(z_i - \alpha_j\right)\right)\right)^{-1}, \end{aligned} \tag{33}$$

gdzie

$$t_i = \sum_{j=1}^{p} y_{ij} \quad q_j = \sum_{i=1}^{n} y_{ij}. \tag{34}$$

Estymatory parametrów $\alpha_1, \alpha_2, ..., \alpha_p$ oraz $z_1, z_2, ..., z_n$ uzyskujemy rozwiązując następujący układ równań:

$$\frac{\partial \ln L(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p)}{\partial z_i} = 0,$$

$$\frac{\partial \ln L(z_1, z_2, ..., z_n, \alpha_1, \alpha_2, ..., \alpha_p)}{\partial \alpha_j} = 0.$$

W celu zapewnienia jednoznaczności rozwiązania należy przyjąć dodatkowe ograniczenie postaci

$$\sum_{i=1}^{n} z_i = 0$$

lub też postać:

$$\sum_{j=1}^{p} \alpha_j = 0.$$

Po dokonaniu przekształceń algebraicznych uzyskuje się następujący układ do rozwiązania:

$$t_i = \sum_{j=1}^{p} \frac{\exp(z_i - \alpha_j)}{1 + \exp(z_i - \alpha_j)}, \quad i = 1, 2, ..., n,$$

$$q_j = \sum_{i=1}^{n} \frac{\exp(z_i - \alpha_j)}{1 + \exp(z_i - \alpha_j)}, \quad j = 1, 2, ..., p,$$

gdzie t_i oraz q_j określone są wzorem (34).

Do rozwiązania tych równań należy zastosować metodę kolejnych przybliżeń.

8. Zakończenie

W artykule niniejszym przedstawiono ogólną ideę statystycznego modelowania dychotomicznych odpowiedzi na pytania ankietowe. Tę ogólną idę omówiono następnie na podstawie dwóch konkretnych modeli statystycznych, które były dotychczas stosowane głównie w badaniach psychometrycznych. Modele te zostały

zaadaptowane do badań społecznych jako statystyczne modele stresorów społecznych (smss). Spośród czterech głównych etapów modelowania: specyfikacji modelu, jego identyfikacji, estymacji i weryfikacji, najwięcej uwagi poświęcono specyfikacji, najmniej zaś weryfikacji. Wynikało to z tego, że specyfikacji omówionych tu modeli poświęca się bardzo mało uwagi, natomiast problem weryfikacji nie został jeszcze w pełni rozwiązany. Dlatego też w artykule przedstawiono tylko ogólną ideę weryfikacji i jedną z najprostszych metod wykorzystującą test chi kwadrat.

Oprócz popularyzacji prezentowanych modeli w środowisku ekonomiczno-społecznym, ubocznym celem artykułu było zaproponowanie i wyrażenie sprzeciwu przeciwko stosowaniu dziwolągów językowych w pracach dotyczących badań sondażowych (nie zaś surwejowych).

Literatura

[1] Amemiya T., *Qualitative Response Models: A Survey*, „Journal of Economic Literature" 1981, 19, 1483-1536.

[2] Andersen E.B., *Discrete Statistical Models with Social Science Applications*, North-Holland, Amsterdam 1980.

[3] Andersen E.B., *Latent Trait Models*, „Journal of Econometric" 1983, 22, 215-227.

[4] Andrich D., *A Rating Formulation for Ordered Response Categories*, „Psychometrika" 1978, 43, 561-573.

[5] Andrich D., *Rasch Models for Measurement*, Sage University Papers No 68, London 1988.

[6] Bartholomew D.J., Steele F., Moustaki I., Galbraith J.I., *The Analysis and Interpretation of Multivariate Data for Social Scientists*, Chapman and Hall, London 2002.

[7] Bartholomew D.J., *Latent Variable Models for Ordered Categorical Data*, „Journal of Econometrics" 1983, 22, 229-243.

[8] Bartholomew D.J., *Scaling Binary Data Using a Factor Model*, „Journal of Royal Statistical Society" 1984, B46, 120-123.

[9] Bartholomew D.J., *Factor Analysis for Categorical Data*, „Journal of the Royal Statistical Society" 1980, B42, 293-321.

[10] Brzeziński J. (red.), *Metodologia badań psychologicznych. Wybór tekstów*, PWN, Warszawa 2004.

[11] Collett D., *Modelling Binary Data*, Chapman and Hall, London 2003.

[12] Everitt B.S., *An Introduction to Latent Variable Models*, Chapman and Hall, London 1984.

[13] Fischer G.H., Molenaar I.W. (red.), *Rasch Models, Foundations, Recent Developments and Applications*, Spinger-Verlag, Berlin 1994.

[14] Gleick J., *Szybciej*, Zysk i S-ka, Poznań 2003.

[15] Gruszczyński L.A., *Elementy metod i technik badań socjologicznych*, Wyższa Szkoła Zarządzania i Nauk Społecznych w Tychach, Tychy 2002.

[16] Haberman S., *Maximum Likelihood Estimates in Exponential Response Models*, „Annals of Statistics" 1977, 5, 815-841.

[17] Jonnson V.E., Albert J.H., *Ordinal Data Modeling*, Springer, New York 1999.

[18] Krauth J., *Itestkonstruktion und Testtheorie*, BELTZ Psychologie Verlags Union, Weinheim 1995.

[19] Lord F.M., *A Strong True-Score Theory with Applications*, „Psychometrika" 1965, 30, 239-270.

[20] Lutyńska K., *Surveye w Polsce. Spojrzenie socjologiczno-antropologiczne*, IFIS PAN, Warszawa 1993.

[21] Maddala G.S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge 1983.
[22] Ostasiewicz W., *Istota pomiaru statystycznego*, [w:] W. Ostasiewicz (red.), *Pomiar statystyczny*, AE, Wrocław 2003, s. 11-45.
[23] Ostasiewicz W., *Kwantyfikacja zjawisk społeczno-gospodarczych*, Prace Naukowe Akademii Ekonomicznej nr 1097, AE, Wrocław 2005, 17-36.
[24] Sanathanan L., Blumenthal S., *The Logistic Model and Estimation of Latent Structure*, „Journal of the American Statistical Association" 1978, vol. 73, 794-799.

STATISTICAL MODELS OF SURVEY RESEARCH

Summary

The paper contains the basic definitions of survey research in social sciences. The whole process of statistical modelling of social stressors is discussed in detail. Two different approaches to this kind of modelling are presented.