AKADEMIA EKONOMICZNA W KATOWICACH

JANUSZ WYWIAŁ

WIELOWYMIAROWE ASPEKTY METODY REPREZENTACYJNEJ

OSSOLINEUM

Janusz Wywiał

WIELOWYMIAROWE ASPEKTY METODY REPREZENTACYJNEJ

Matce i pamięci Ojca

AKADEMIA EKONOMICZNA W KATOWICACH

Janusz Wywiał

WIELOWYMIAROWE ASPEKTY METODY REPREZENTACYJNEJ

OSSOLINEUM

WROCŁAW-WARSZAWA- KRAKÓW
ZAKŁAD NARODOWY IMIENIA OSSOLIŃSKICH
WYDAWNICTWO 1995

Publikację opiniowali:
Prof. dr hab. Czesław Bracha
Prof. dr hab. Stanisław Maciej Kot

Wydanie pracy sfinansował Komitet Badań Naukowych

Wrocław 1995
Printed in Poland

ISBN

Skład i łamanie tekstu
Agencja Wydawnicza „TRIO"
40-881 Katowice ul.Chrobrego 26
tel. 150-47-27

SPIS TREŚCI

WSTĘP

W praktyce badań reprezentacyjnych zwykle mamy do czynienia z problemem wnioskowania o wielu parametrach analizowanych cech populacji. Rzadko celem takiego badania jest ocena wartości jednego parametru, chociaż temu właśnie przypadkowi jest głównie poświęcana większość prac metodologicznych z metody reprezentacyjnej. Bynajmniej nie oznacza to, iż te prace mijają się z praktycznymi potrzebami badań statystycznych, ponieważ otrzymywane wyniki dotyczące wnioskowania o pojedynczym parametrze jednowymiarowej cechy można w wielu zagadnieniach bezpośrednio uogólnić na przypadek wielowymiarowy. W tej dziedzinie są jednak problemy jednoczesnego wnioskowania o wielu parametrach, które wymagają szczególnego podejścia. Należą do nich problem sposobu oceny dokładności estymacji wektora parametrów oraz interpretacja używanych do tego celu wskaźników. Kluczowe znaczenie ma także usystematyzowanie podstawowych wiadomości pozwalających na porównywanie dokładności estymatorów wektorowych. Następna kwestia dotyczy optymalizacji badań próbkowych, a zwłaszcza optymalizacji rozmiarów prób złożonych, gdy występują ograniczone nakłady na badania reprezentacyjne oraz żądania spełnienia wymaganej dokładności oceny parametrów.

W ogólności właśnie wymienionym problemom poświęcona jest niniejsza praca. Prezentowano w niej głównie zagadnienia dotyczące jednoczesnej estymacji wielu parametrów cech w populacji. Nacisk położono na prezentację oryginalnych wyników otrzymanych przez autora. Treść niniejszej pracy jest wynikiem kontynuacji badań autora nad wielowymiarowymi problemami występującymi w metodzie reprezentacyjnej.

W pracy ograniczono się głównie do analizy problemu estymacji wektora wartości średnich w populacji. Otrzymane na tym polu wyniki można jednak łatwo przenieść na zagadnienie oceny innych ważnych z punktu widzenia praktyki parametrów, takich jak suma wartości cechy w populacji, ilość elementów z cechą wyróżnioną w populacji, częstość względna występowania określonego zjawiska w populacji.

W pierwszym rozdziale przedstawiono podstawowe definicje związane z rozkładami cech w populacji ustalonej, jak i w tzw. nadpopulacji. Szczegółowiej potraktowano problem interpretacji miar zróżnicowania wartości wielowymiarowej zmiennej. Oprócz podstawowych wiadomości o własnościach planów i schematów losowania przedstawiono szczegółowo specyficzne plany losowania zależne od cech pomocniczych. Uwzględniono również plany losowania prób z populacji przestrzennych zależne od położenia względem siebie elementów populacji.

Podstawowym wiadomościom o własnościach estymatorów wektorowych poświęcono drugi rozdział. Prezentowane tu są definicje i twierdzenia, które zwykle są bezpośrednimi uogólnieniami na przypadek wielowymiarowy odpowiednich określeń znanych z przypadku wnioskowania o jednowymiarowym parametrze. Szczególny nacisk położono na problem porównywania dokładności estymatorów wektorowych.

W trzecim rozdziale podstawiono podstawowe własności wektora średnich z próby prostej. Analizowano tu także problem wyznaczania niezbędnej liczebności próby. Korzystano m.in. z wprowadzonego tu uogólnienia nierówności Czebyszewa.

Podstawowe parametry rozkładu wektora znanych estymatorów Horvitza-Thompsona są prezentowane w czwartym rozdziale. Wyznaczano tu również w przybliżony sposób wariancje tego estymatora dla wybranych planów losowania zależnych od cech pomocniczych.

Piąty rozdział jest poświęcony wektorowym estymatorom regresyjnym, ilorazowym i iloczynowym. Prezentowane są ich macierze wariancji i kowariancji w przypadkach, gdy losowana jest próba prosta bądź podwójna. Gdy estymatory te są wyznaczane na podstawie obserwacji cech badanych i pomocniczych w próbie podwójnej, formułowano i rozwiązywano zadania optymalnego ustalania liczebności obu prób składowych próby podwójnej.

W szóstym rozdziale prezentowano podstawowe własności rozkładu wektora estymatorów z próby warstwowej. Konstruowano i rozwiązywano zadania optymalnej lokalizacji prób w warstwach. Zagadnienie optymalnego tworzenia warstw w populacji ograniczono do problemu wykorzystania formalnych metod grupowania na podstawie obserwowanych cech w populacji do wyodrębniania w niej warstw. Wskazano także na możliwość wykorzystania takich metod do warstwowania próby prostej wylosowanej z populacji. Z tak utworzonych warstw w następnym kroku są losowane próby proste, w których już obserwuje się cechy badane. Problem ten jest podobny do znanego zagadnienia warstwowania próby po jej wylosowaniu.

Podstawowe parametry rozkładu wektora średnich z próby grupowej prezentowano w rozdziale siódmym. Parametry te przedstawiono jako funkcje tzw. współczynników korelacji wewnątrzgrupowej. Podjęto problem optymalnego wyróżniania grup w populacji bądź w wylosowanej próbie prostej na podstawie obserwacji cech pomocniczych. Podobnie jak w uprzednim punkcie, do tego celu wykorzystano formalne metody grupowania zbiorów.

Rozdział ósmy dotyczy estymacji wektora średnich w populacji na podstawie wektora średnich z próby dwustopniowej. Oprócz podstawowych własności rozkładu wektora tych estymatorów zaprezentowano rozwiązania zadań optymalizacji rozmiarów prób składowych próby dwustopniowej.

Dziewiąty rozdział poświęcono wybranemu zagadnieniu estymacji wektora średnich cech w okresie bieżącym na podstawie estymatora będącego funkcją statystyk wyznaczanych z prób losowanych w okresach bieżącym i wyjściowym badania. Formułowano i także rozwiązywano zadania dotyczące optymalizacji liczebności losowanych prób.

Rezultaty otrzymane w pracy powinny przyczynić się do racjonalizacji badań reprezentacyjnych populacji. Prezentowane własności estymatorów wektorowych i planów losowania prób winny ułatwić wybór najodpowiedniejszych spośród nich w praktyce badań statystycznych. Celem optymalnego tworzenia warstw bądź grup w populacji jest w konsekwencji zwiększenie dokładności oceny parametrów na podstawie prób warstwowych bądź grupowych. Zastosowanie tych procedur przyczyni się również do oszczędności nakładów przeznaczanych na badania reprezentacyjne. W tej kwestii zwłaszcza mają bezpośrednie znaczenie analizowane w pracy zagadnienia optymalnego ustalania liczebności prób złożonych z uwzględnieniem kosztów badania i żądanej dokładności ocen parametrów.

1. PODSTAWY PRÓBKOWANIA

Przedmiotem badań metody reprezentacyjnej jest analiza własności charakterystyk zbioru obiektów zwanego populacją. W praktyce zwykle nie interesują nas wszystkie poziomy cech przypisane poszczególnym elementom populacji, lecz tylko pewne ich funkcje, nazywane funkcjami parametrycznymi, które mają syntetycznie charakteryzować własności populacji. Zaliczamy do nich np.: wartość średnią i wariancję cechy, współczynnik korelacji, parametry regresji. Wielkości te są również nazywane parametrami cechy[1].

Gdy parametr populacji jest ustalony, to mówi się o wnioskowaniu klasycznym prowadzonym na podstawie populacji ustalonej[2], a gdy jest losowy, mamy do czynienia z modelowym podejściem do wnioskowania o parametrach opisowych populacji. W drugim przypadku trzeba wprowadzić założenia o rozkładzie prawdopodobieństwa zmiennej losowej wielowymiarowej, której realizacją jest parametr populacji. Zbiór realizacji tej zmiennej jest nazywany nadpopulacją w stosunku do rozważanej podczas badania populacji obiektów. Podejście modelowe jest także nazywane wnioskowaniem na podstawie modelu nadpopulacji czy superpopulacji[3].

1.1. Populacja z parametrem ustalonym

Najprostszy model populacji jest określany przez Cassela i in. (1977), s. 4, następująco:

Definicja 1.1: Zbiór obiektów $\Omega \{\omega_1,....,\omega_N\}$, $N < \infty$ jest nazywany populacją skończoną N-elementową.

Dla wygody nazwę populacja skończona i ustalona skrócimy do słowa populacja.

---

[1] Por. np. pracę Hellwiga (1987), s. 99.

[2] Ang. design approach.

[3] Ang. model approache lub superpopulation approach.

Definicja 1.2: [Cassel i in. (1977)]: Mówimy, że elementy populacji Ω są identyfikowalne, jeśli mogą być jednoznacznie ponumerowane od 1 do N i gdy każdy element odpowiadający danemu numerowi jest obserwowalny.

Definicja 1.3: Wektor liczb rzeczywistych $\mathbf{y}^T = [y_1...y_N]$ jest parametrem populacji Ω, jeśli każdemu elementowi $i \in \Omega$ jest przyporządkowany i-ty element wektora **y**.

Symbolem $\mathcal{Y}$ oznaczmy przestrzeń parametru populacji. Zwykle jest ona N wymiarową podprzestrzenią przestrzeni kartezjańskiej, czyli $\mathcal{Y} \in R^N$.

Definicja 1.4 [Bracha (1987)]: Cechą (zmienną) w populacji nazywamy przyporządkowanie $y : \Omega \rightarrow R^1$ takie, że dla każdego i=1,...,N zachodzi, iż $y(\omega_1)=y_1$.

Zmienną m-wymiarową oznaczamy przez $y=[y_1...y_m]$, a jej wartosci są elementami macierzy $\mathbf{y}=[y_{ij}]$, i=1,...,N, j=1,...,m, przy czym y_{ij} jest wartością j-tej zmiennej przypisaną i-temu elementowi populacji.

Wnioskowanie o parametrach populacji można znacznie polepszyć, jeśli są o niej dostępne informacje dodatkowe. W szczególności są to z góry znane badaczowi obserwacje tzw. cech pomocniczych (wspomagających). Wektor tych cech oznaczamy przez $x=\{x_1,....,x_p\}$, a macierz ich wartości przez $\mathbf{x}=[x_{ij}]$, która ma wymiary N×p.

1.2. Podstawowe parametry opisowe populacji

W praktyce przedmiotem wnioskowania nie jest parametr populacji, lecz jego funkcje charakteryzujące własności zmiennych.

Definicja 1. 5: m - wymiarowym parametrem opisowym populacji lub określonej na niej cechy m wymiarowej *y* nazwiemy taką funkcję wektorową $\boldsymbol{\theta} = [\theta_1...\theta_m]$ definiowaną na przestrzeni $\mathcal{Y}$ parametru populacji, że $\boldsymbol{\theta} : \mathcal{Y} \rightarrow R^m$.

Do najczęściej używanych parametrów opisowych należy wektor wartości średnich zmiennych w populacji $\bar{\mathbf{y}} = [\bar{y}_1 \ldots \bar{y}_m]^T$, gdzie:

$$\bar{\mathbf{y}} = \mathbf{N}^{-1}\mathbf{y}^T\mathbf{J}_N \qquad (1.1)$$

Przez $\mathbf{J}_N$ oznaczono wektor jedynkowy o wymiarze N×1. Szczególnym przypadkiem wartości średniej danej zmiennej jest częstość względna występowania wyróżnionego poziomu cechy w populacji.

Wektor wartości globalnych zmiennych w populacji ma postać:

$$\tilde{\mathbf{y}} = N\bar{\mathbf{y}} \qquad (1.2.)$$

Jego szczególnym przypadkiem jest liczba elementów populacji z cechą wyróżnioną.

Przegląd parametrów zróżnicowania zmiennej wielowymiarowej rozpoczynamy od macierzy wariancji i kowariancji wektora cech $y=[y_1...y_m]$, którą oznaczamy przez: $\mathbf{C}_*(y)=[c_{*tj}]$ (t,j=1,...,m), gdzie:

$$c_{*tj}=c(y_t, y_j) = (N-1)^{-1}\sum_{i=1}^{N}(y_{it} - \overline{y}_t)(Y_{ij} - \overline{y}_j) \tag{1.3}$$

przy czym wariancję oznaczamy przez $v_*(y_j) = c_*(y_j, y_j)$.

Macierz współczynników korelacji wektora zmiennych *y* w populacji oznaczamy przez $\mathbf{R} = \mathbf{R}(y) = [r_{tj}]$ (t, j=1,...,m), gdzie:

$$r_{tj} = \frac{c(y_t, y_j)}{\sqrt{v_*(y_t)v_*(y_j)}} \tag{1.4}$$

Definicja 1.6: Średni promień rozkładu zmiennej m wymiarowej *y* określamy jako pierwiastek z sumy wariancji poszczególnych jej składowych, czyli:

$$q_*(y) = \sqrt{\mathrm{tr}\mathbf{C}_*(y)} \tag{1.5}$$

gdzie przez tr oznaczono ślad macierzy.

Parametr $q_*(y)$ jest pierwiastkiem ze średniej kwadratów odległości punktów o współrzędnych będących obserwacjami cechy *y* od punktu, którego współrzędnymi są wartości średnie tych cech.

Definicja 1.7 [Wilks (1932)]: Uogólnioną wariancją rozkładu m wymiarowej zmiennej *y* jest wyznacznik z jej macierzy wariancji i kowariancji[4], czyli

$$g(y) = \det(\mathbf{C}_*(y)) \tag{1.6}$$

Z geometrycznego punktu widzenia uogólnioną wariancję można interpretować na kilka sposobów. Niech $\mathbf{e}_{*j} = \mathbf{y}_{*j} - \mathbf{J}_N \overline{Y}_j$ będzie wektorem odchyleń obserwacji j-tej zmiennej od jej średniej. Anderson (1958), s. 167 wykazuje twierdzenie:

Twierdzenie 1.1: Uogólniona wariancja g(*y*) jest proporcjonalna do kwadratu objętości równoległościanu rozpiętego na wektorach zaczepionych wspólnie w początku układu współrzędnych $\mathbf{o}_N$ i końcach w punktach $\mathbf{e}_{*i},...,\mathbf{e}_{*m}$ w przestrzeni N wymiarowej.

Niech $m(\mathbf{y}_{*i_1},\ldots,\mathbf{y}_{*i_m},\mathbf{y}_{*i_{m+1}})$ będzie objętością (miarą) m wymiarową równoległościanu rozpiętego na wektorach wspólnie zaczepionych w punkcie $\mathbf{y}_{*i_{m+1}}$ o końcach w punktach $\mathbf{y}_{*i_1},\ldots,\mathbf{y}_{*i_m}$, którą określa wyrażenie[5] :

$$m(\mathbf{y}_{*i_1},..,\mathbf{y}_{*i_m},\mathbf{y}_{*i_{m+1}}) = \left|\det\begin{bmatrix}\mathbf{y}_{*i_1} - \mathbf{y}_{*i_{m+1}}\\ \ldots\ldots\ldots\ldots\\ \mathbf{y}_{*i_m} - \mathbf{y}_{*i_{m+1}}\end{bmatrix}\right| \tag{1.7}$$

Z kolei niech $m(\mathbf{y}_{*i_1},\ldots,\mathbf{y}_{*i_m},\overline{\mathbf{y}})$ będzie objętością m wymiarową równoległościanu rozpiętego na wektorach wspólnie zaczepionych w punkcie $\overline{\mathbf{y}}$ i końcach w punktach $\mathbf{y}_{*i_1},\ldots,\mathbf{y}_{*i_m}$. Anderson (1958), s. 168-170, wykazuje następującą własność:

---

[4] Definicję tę podajemy za Cramerem (1958) i Fiszem (1967).

[5] Por. np. pracę Borsuka (1976). Dodajmy także, że decyzja o wyborze, który punkt spośród danej kombinacji m+1 punktów jest punktem zaczepienia wektorów, a które punkty ich końcami, jest dowolna.

Twierdzenie 1.2: Uogólniona wariancja q(*y*) jest proporcjonalna do sumy kwadratów objętości równoległościanów rozpinanych na wektorach wspólnie zaczepionych w punkcie $\bar{\mathbf{y}}$ i końcach w punktach, których współrzędnymi są m elementowe kombinacje bez powtórzeń wierszy obserwacji y, co określa wyrażenie:

$$q_*(y) = N^{-m} \sum_{\{i_1,..,i_m\}} m^2\left(\mathbf{y}_{*i_1},...\mathbf{y}_{*i_m},\bar{\mathbf{y}}\right) \tag{1.8}$$

przy czym sumacja przebiega po wszystkich m elementowych kombinacjach bez powtórzeń numerów wierszy macierzy **y**.

Twierdzenie 1.3 [Wywiał (1989 b); (1992 a)]: Uogólniona wariancja q(*y*) jest proporcjonalna do sumy kwadratów objętości równoległościanów rozpinanych na kombinacjach (m+1) punktów o współrzędnych będących wierszami macierzy obserwacji cech **y**, co określa równanie:

$$q_*(y) = N^{-m-1} \sum_{\{i_1,..,i_{m+1}\}} m^2(\mathbf{y}_{*i_1},\ldots,\mathbf{y}_{*i_m},\mathbf{y}_{*i_{m+1}}) \tag{1.9}$$

przy czym sumacja przebiega po wszystkich (m+1) elementowych kombinacjach bez powtórzeń numerów wierszy macierzy **y**.

Uogólniona wariancja może więc służyć do oceny stopnia zróżnicowania obserwacji zmiennej wielowymiarowej traktowanych jako współrzędne punktów w przestrzeni cech lub obiektów. Jej wartość jest proporcjonalna do kwadratów objętości odpowiednio rozpinanych na tych punktach równoległościanów. Jeśli q(*y*) = 0, to obserwacje m wymiarowej zmiennej leżą na pewnej co najwyżej (m-1) wymiarowej hiperpłaszczyźnie, co wykazuje Anderson (1958).

Oznaczmy przez λ_j (j=1,...,m) pierwiastki charakterystyczne (wartości własne) macierzy wariancji i kowariancji $\mathbf{C}_*(y)$, przy czym niech $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$. Na podstawie znanych własności tzw. głównych składowych oraz warunków prostopadłości hiperpłaszczyzny i prostej wnioskujemy[6] , że jeśli $\det\mathbf{C}_*(y)>0$, to dla każdej wartości λ_j istnieje taka (m-1) wymiarowa hiperpłaszczyzna $H^{(j)}_{m-1}: \sum_{k=1}^{m} a_k^{(j)}\mathbf{x}_k = 0$, że średnia suma kwadratów odległości od niej punktów $\mathbf{e}_{i*} = \mathbf{y}_{i*} - \bar{\mathbf{y}}$ (i=1,...N) jest równa λ_j. Taką hiperpłaszczyznę Pearson (1901) nazywa hiperpłaszczyzną regresji ortogonalnej[7] . W szczególności gdy $\lambda_m = 0$, to wszystkie punkty $\mathbf{e}_{i*}$ leżą na hiperpłaszczyźnie $H^{(m)}_{m-1}$. W przypadku gdy $\lambda > 0$, dla j=1,..., m_0`-1 oraz $\lambda_j = 0$ dla j=$m_0$`, m_0`+1,...,m, to punkty $\mathbf{e}_{i*}$ (i=1,...,N) leżą na pewnej hiperpłaszczyźnie ($m_0$`-1) wymiarowej.

Parametr λ_j jest równy wariancji j-tej głównej składowej, której wartości są wyznaczane z równania: $\mathbf{u}_{ij} = \mathbf{a}^{(j)}\mathbf{e}_{i*}^T$ (i=1,...,N), gdzie: $\mathbf{a}^{(j)} = [a_1^{(j)}\ldots a_m^{(j)}]$ oraz $\mathbf{a}^{(j)}(\mathbf{C}_*(y)-\mathbf{I}_m\lambda_j)=\mathbf{0}_m$ czyli $\mathbf{a}^{(j)}$ jest wektorem własnym macierzy $\mathbf{C}_*$. Wektor wartości j-tej głównej składowej oznaczamy przez $\mathbf{u}_j = [u_{1j}...u_{Nj}]^T$. Z geometrycznego punktu widzenia przekształcenie przeprowadzające zmienne pierwotne w główne składowe jest takim obrotem układu współrzędnych, w

---

[6] Por. np. Anderson (1958), C.R.Rao (1982), Wilks (1962).

[7] Wiadomość tę podajemy za Cramerem (1958), s. 298.

wyniku którego otrzymane główne składowe są do siebie prostopadłe, czyli wektory $\overrightarrow{\mathbf{o}_N\mathbf{u}_j}$ i $\overrightarrow{\mathbf{o}_N\mathbf{u}_i}$ (j≠t=1,...,m) są do siebie prostopadłe. Wówczas wyznacznik $\det \mathbf{C}_*(y) = \prod_{j=1}^{m} \lambda_j$. Stąd wnioskujemy, że uogólniona wariancja jest równa kwadratowi m wymiarowej objętości prostopadłościanu rozpiętego na wektorach $\overrightarrow{\mathbf{o}_N\mathbf{u}_j}$ (j=1,...,m), przy czym długość j-tego wektora wynosi $\sqrt{\lambda_j}$.

Przeciętny poziom niezerowych wartości własnych macierzy wariancji i kowariancji określamy za pomocą średniej geometrycznej w następujący sposób:

$$\tilde{\lambda} = \sqrt[m']{\prod_{j=1}^{m'} \lambda_j} \tag{1.10}$$

W szczególności, gdy wszystkie pierwiastki charakterystyczne macierzy $\mathbf{C}_*(y)$ są dodatnie, to średnia $\tilde{\lambda}$ jest równa pierwiastkowi m-tego stopnia z uogólnionej wariancji $g_*(y)$.

Maksymalna wartość pierwiastka charakterystycznego macierzy kwadratowej jest nazywana[8] jej promieniem spektralnym, czyli promień spektralny macierzy $\mathbf{C}_*(y)$ wynosi λ_1.

Definicja 1.8: Parametr $\rho(y) = \sqrt{\lambda_1}$ będziemy nazywać promieniem spektralnym rozkładu zmiennej *y*.

Niech $\mathbf{z} = [z_1...z_N]^T$ będzie wektorem wartości zmiennej *z*, który jest kombinacją liniową kolumn macierzy obserwacji **y** zmiennych *y*, a przez $\mathbf{b} = [b_1...b_m]^T$ oznaczamy wektor współczynników tej kombinacji, przy czym zakładamy, iż $\mathbf{b}^T\mathbf{b} = 1$. Wtedy $\mathbf{z} = \mathbf{yb}$. Po odpowiednich operacjach algebraicznych otrzymujemy wariancję zmiennej *z*:

$$v(z) = \mathbf{b}^T\mathbf{C}_*(y)\mathbf{b}$$

Na podstawie znanych własności form kwadratowych[9] wnioskujemy, że jeśli $\mathbf{b} = \mathbf{b}_1$, gdzie $\mathbf{b}_1$ jest unormowanym wektorem własnym odpowiadającym maksymalnej wartości własnej λ_1 macierzy $\mathbf{C}_*(y)$, to

$$v(z) = \max_{\mathbf{b}^T\mathbf{b}=1} imum\left\{\mathbf{b}^T\mathbf{C}_*(y)\mathbf{b}\right\} = \lambda_1 \tag{1.11}$$

gdzie z_1 jest zmienną, której wartości wyznacza transformacja: $\mathbf{z}_1=\mathbf{yb}_1$. Promień spektralny rozkładu zmiennej *y* jest więc równy odchyleniu standardowemu $\sqrt{v(z_1)}$ zmiennej będącej kombinacją liniową wartości wyjściowych zmiennych. Współczynniki tej kombinacji są elementami wektora $\mathbf{b}_1$ tak wybranego, by parametr $v(z_1)$ osiągał wartość maksymalną.

Przykładowo niech $\mathbf{p} = [p_1...p_m]^T$ będzie wektorem cen m dóbr, które są miesięcznie kupowane przez i-te (i=1,...,N) gospodarstwo domowe w ilości y_{ij} (j=1,...,m). Przez **b** oznaczymy wektor cen unormowanych tak, że $\mathbf{b} = \alpha^{-1}\mathbf{p}$, gdzie $\alpha^2 = \mathbf{p}^T\mathbf{p}$. Wówczas wektor $\mathbf{b}_1$ jest

[8] Por. np. pracę Ralstona (1975).

[9] Por. np. pracę Rao (1982), s. 81 i 587.

najmniej korzystnym wektorem cen unormowanych w tym sensie, że wariancja miesięcznego wydatku $z_1 = y\mathbf{b}_1$ osiąga wartość nie mniejszą od wartości wariancji otrzymanej przy innym wektorze cen **b**. Ponadto:

$$v(z_1) = v(y\mathbf{b}_1) = \alpha_1^{-2} v(y\mathbf{p}_1) = \alpha_1^{-2} v(w_1)$$

gdzie $\alpha_1^2 = \mathbf{p}_1^T \mathbf{p}_1$, natomiast $w_1 = y\mathbf{p}_1$ określa rzeczywiste wydatki przy najmniej korzystnym układzie reprezentowanym wektorem cen $\mathbf{p}_1 = \alpha^2 \mathbf{b}_1$. Zatem odchylenie standardowe $\sqrt{v(z_1)}$ można w tym przypadku także interpretować jako specyficzny współczynnik zmienności rozkładu wydatków rzeczywistych wyliczany jako iloraz ich odchylenia standardowego $\sqrt{v(w_1)}$ przez długość α_1 wektora cen $\mathbf{p}_1$.

Definicja 1.9 [Frish (1929)]: Współczynnik rozrzutu (rozsiewu) wielowymiarowej zmiennej określa wzór[10] :

$$u(y) = \sqrt{\det \mathbf{R}(y)} \tag{1.12}$$

Współczynnik u(y) przyjmuje wartości z przedziału <0;1>. Jeśli u(y) = 0, to obserwacje zmiennej m wymiarowej y leżą na co najwyżej (m-1) wymiarowej hiperpłaszczyźnie, a zatem są zależne liniowo. Gdy u(y) = 1, to składowe wektora cech $y = [y_1...y_m]$, są nieskorelowane, czyli liniowo niezależne. Przyjmuje się, że zależność liniowa między składowymi wektora y jest tym silniejsza, im wartość współczynnika u(y) jest bliższa zeru.

Searle (1966) proponuje następującą dekompozycję uogólnionej wariancji przy założeniu, że jest ona dodatnia[11] :

$$g_*(y) = u^2(y)\tilde{v}_*^m(y) \tag{1.13}$$

gdzie:

$$\tilde{v}_*(y) = \sqrt[m]{\prod_{j=1}^{m} v_*^2(y_j)} \tag{1.14}$$

jest średnią geometryczną wariancji składowych zmiennej y. Uogólnioną wariancję $g_*(y)$ zapisano więc w postaci iloczynu czynnika wskazującego stopień zróżnicowania wartości poszczególnych cech i czynnika określającego siłę zależności liniowej między nimi.

Kowal (1971) krytycznie ustosunkowuje się do używania uogólnionej wariancji jako parametru określającego stopień zróżnicowania rozkładu wielowymiarowej zmiennej. Przede wszystkim dlatego, że parametr ten jest funkcją wariancji zmiennych, których wartości mają różne miana. Niepokoi go także dekompozycja (1.13). Uważa, że porównań zróżnicowania rozkładów zmiennych można dokonywać tylko wtedy, gdy mają te same miana i gdy u(y) jest równe lub bliskie jedności. Obawy te wydają się nieuzasadnione w przypadku wartości współczynnika rozsiewu w kontekście przyjętej jego interpretacji jako wskaźnika stopnia współliniowości obserwacji zmiennych. Wydaje się, że w ostateczności dwa rozkłady można porównać osobno z punktu widzenia stopnia skorelowania cech za pomocą współczynnika rozsiewu i z punktu widzenia średniej wariancji zmiennych wyliczanej na podstawie wzoru (1.14), lecz rzeczywiście tylko wtedy, gdy mają te same miana.

---

[10] Definicję tę podajemy za Cramerem (1958) i Fiszem (1967).

[11] Wiadomość tę podajemy za Kowalem (1971).

Wnioskowanie o względnym zróżnicowaniu rozkładu cechy wielowymiarowej można prowadzić na podstawie następującej macierzy współczynników zmienności:

$$\boldsymbol{\gamma}_*(y)=(\text{diag}\,\bar{\mathbf{y}})^{-1}\mathbf{C}_*(y)(\text{diag}\bar{\mathbf{y}})^{-1} \tag{1.15}$$

przy czym zakładamy, że każda składowa wektora średnich $\bar{\mathbf{y}}$ jest różna od zera. Przez diag $\bar{\mathbf{y}}$ oznaczono macierz diagonalną, ktorej elementy diagonalne tworzą kolejne składowe wektora $\bar{\mathbf{y}}$. Na podstawie macierzy $\boldsymbol{\gamma}_*(y)$ można definiować syntetyczne współczynniki zróżnicowania zmiennej podobne np. do wprowadzonej wyżej uogólnionej wariancji.

1.3. Modele nadpopulacji

Koncepcję wnioskowania na podstawie modelu nadpopulacji wprowadzali początkowo Cochran (1939; 1946), Deming i Stephan (1941) oraz Madow i Madow (1944), co podajemy za Casselem, Sarndalem i Wretmanem (1977). Do zwolenników tej metody należy m.in. Barnard (1971).

Przyjmijmy, że określony uprzednio definicją 1.3 parametr populacji **y** jest realizacją pewnej macierzy losowej $\mathbf{Y} = [Y_{ij}]$ (i=1,...,N; j=1,...,m). Wtedy przestrzeń parametru **y** staje się przestrzenią prób macierzy losowej **Y**. Oznaczmy przez F(**y**) dystrybuantę rozkładu prawdopodobieństwa macierzy **Y**. Cassel i in. (1977) podają następujące określenie modelu nadpopulacji:

Definicja 1.10: Modelem nadpopulacji jest nazywany zespół warunków lub parametrów definiujących klasę rozkładów prawdopodobieństwa, do której należy dystrybuanta F(**y**).

Dodajmy, że populacja ustalona jest szczególnym przypadkiem nadpopulacji, bowiem jej parametr **y** będący zarazem macierzą wartości cech występuje z prawdopodobieństwem jeden.

Model nadpopulacji określany dystrybuantą F(**y**) stanowi dodatkową wiedzę o populacji obok informacji niesionych przez tzw. zmienne pomocnicze, co stwierdza m.in. Basu (1971).

Wartość oczekiwaną, wariancję i kowariancję funkcji macierzy losowej **y** oznaczamy przez $\mathcal{E}(.),\mathcal{D}^2(.)$ i $\mathcal{C}ov(.)$, a w szczególności mamy:

$$\mathcal{E}(Y_{ij}) = \mu_{ij}, \quad \mathcal{D}^2(Y_{ij}) = \mathcal{E}(Y_{ij} - \mu_{ij})^2 = \sigma^2_{ij}$$

$$\mathcal{C}ov(Y_{ij},Y_{kt}) = \mathcal{E}(Y_{ij} - \mu_{ij})(Y_{kt} - \mu_{kt}) = \sigma_{ik,\,jt}$$

Niech $\mathbf{Y} = [\mathbf{Y}_{*1}...\mathbf{Y}_{*m}]$, gdzie $\mathbf{Y}_{*j}^{T} = [Y_{1j}...Y_{Nj}]$ (j=1,...,m). Realizacja wektora losowego $\mathbf{Y}_{*j}$ stanowi obserwację j-tej cechy w populacji. Cassel i in. (1977) przedstawiają model transformacyjny nadpopulacji dla jednowymiarowej cechy, który w naszym przypadku dla j-tej cechy określamy dystrybuantą $F_{*j}(\mathbf{y}_{*j}/\mathbf{a}_{*j},\mathbf{b}_{*j},\mu,\sigma,\rho)$ (j=1,...,m). Składowe wektorów $\mathbf{a}_{*j}=[a_{1j}...a_{Nj}]^T$ ($a_{ij}\neq 0$ dla każdej pary i,j) oraz $\mathbf{b}_{*j} = [b_{1j}...b_{Nj}]^T$ są parametrami transformacji:

$$U_{ij} = \frac{Y_{ij} - b_{ij}}{a_{ij}}, \; i = 1,...,N \tag{1.16}$$

Parametry składowych wektora $\mathbf{U}_{*j} = [U_{1j}...U_{Nj}]^T$ są następujące:

$$\mathcal{E}(Z_{ij}) = \mu, \;\; \mathcal{D}^2(U_{ij}) = \sigma^2, \;\; \mathcal{C}ov(U_{ij},U_{kj}) = \rho\sigma^2 \tag{1.17}$$

przy czym

$$-(N-1)^{-1} \leq \rho \leq 1 \tag{1.18}$$

Stąd i ze wzoru (1.16) wynika, że:

$$\begin{cases} \mathcal{E}(Y_{ij}) = a_{ij}\mu + b_{ij} \\ \mathcal{D}^2(Y_{ij}) = a_{ij}^2\sigma^2 \\ \mathcal{C}ov(Y_{ij},Y_{kj}) = a_{ij}a_{kj}\rho\sigma^2 \end{cases} \tag{1.19}$$

Opisany model oznaczamy symbolem G_T i uogólniamy na przypadek łącznego rozkładu m cech. Określającą go dystrybuantę oznaczamy przez $F(\mathbf{y} \mid \mathbf{a},\mathbf{b},\mu,\sigma,\rho)$, przy czym $\mathbf{a}=[\mathbf{a}_{*1}...\mathbf{a}_{*m}]$ i $\mathbf{b} = [\mathbf{b}_{*1}...\mathbf{b}_{*m}]$. Macierzy losowej $\mathbf{Y}$ odpowiada macierz $\mathbf{U} = [U_{ij}]$ (i=1,...,N; j=1,...,m), której składowe określa wzór (1.16). Parametry elementów macierzy $\mathbf{U}$ określają wzory (1.17) oraz dla j, t=1,...,m, $i \neq k = 1,...,N$ równania:

$$\begin{cases} \mathcal{C}ov(U_{ij},U_{it}) = \rho_{jt}\sigma^2 \\ \mathcal{C}ov(U_{ij},U_{kt}) = \rho\sigma^2 \quad dla \;\; i \neq k \;\; oraz \;\; j \neq t \end{cases} \tag{1.20}$$

Z kolei parametry rozkładu składowych macierzy $\mathbf{Y}$ określają równania (1.19) oraz dla j,t=1,...,m, i≠k=1,...,N wzory:

$$\begin{cases} \mathcal{C}ov(Y_{ij},Y_{it}) = a_{ij}a_{it}\rho_{jt}\sigma^2 \\ \mathcal{C}ov(Y_{ij},Y_{kt}) = a_{ij}a_{kt}\rho\sigma^2 \quad \text{dla} \;\; i \neq k, \;\; j \neq t \end{cases} \tag{1.21}$$

Macierz korelacji wektora $\mathbf{U}_{i*}=[U_{i1}...U_{im}]$ (i=1,...,N) oznaczmy przez $\mathbf{R}=[\rho_{ij}]$. Niech $\mathbf{P} = \rho\mathbf{J}_m\mathbf{J}_m^T$. Wprowadźmy wektor $\mathbf{U}_*=[U_{1*} \; ... \; U_{N*}]^T$. Wtedy na podstawie wyrażeń

(1.17) i (1.20) wnioskujemy, że macierz korelacji $\mathbf{R}_*$ wektora $\mathbf{U}_*$ jest stopnia Nm i ma następującą postać blokową:

$$\mathbf{R}_* = \begin{bmatrix} \mathbf{R} & \mathbf{P} & \dots & \mathbf{P} \\ \mathbf{P} & \mathbf{R} & \dots & \mathbf{P} \\ \dots & \dots & \dots & \dots \\ \mathbf{P} & \mathbf{P} & \dots & \mathbf{R} \end{bmatrix}$$

Twierdzenie 1.4 [Wywiał (1992)]: Jeśli $\det(\mathbf{R}_*)>0$, to współczynnik korelacji ρ spełnia nierówności:

$$-\frac{1}{(N-1)\mathbf{J}_m^T\mathbf{R}\mathbf{J}_m} \le \rho \le \frac{1}{\mathbf{J}_m^T\mathbf{R}\mathbf{J}_m} \tag{1.22}$$

a gdy m=1, to nierówność ta redukuje się do danej wzorem (1.18).

Szczególny w stosunku do opisanego wyżej jest model nadpopulacji warstwowej. Oznaczmy przez Ω_h wzajemnie rozłączne podzbiory elementów populacji Ω, przy czym $\Omega = \bigcup_{h=1}^{L} \Omega_h$ Ω_h. Podzbiór Ω_h jest nazywany warstwą, a jego liczebność jest oznaczana przez N_h. Przyjmijmy, że dla każdego $i \in \Omega_h$ rozkład wektora $\mathbf{Y}_{i*}$ ma taki sam wektor wartości oczekiwanych i tę samą macierz wariancji i kowariancji, co oznacza, że we wzorze (1.19) należy przyjąć, iż $a_{ij} = a_{kj}$ oraz $b_{ij} = b_{kj}$, dla każdego $i,k \in \Omega_h$.Wprowadzone założenia charakteryzują właśnie model nadpopulacji warstwowej.

Przyjmijmy, że dla każdego i=1,...,N oraz j=1,...,m $b_{ij} = 0$ oraz $a_{ij} = 1$, a także $\rho = 0$. Wtedy określony wyrażeniami (1.19) i (1.21) model nadpopulacji znacznie upraszcza się. Dodatkowo zakładając, że wektory $\mathbf{Y}_{1*},...,\mathbf{Y}_{*N}$ są parami niezależne, mamy do czynienia z modelem powszechnie rozważanym w klasycznej statystyce matematycznej, który jest równoważny definicji prostej próby statystycznej, por. np. prace Fisza (1967) i Pawłowskiego (1976). Przy dodatkowym założeniu, że każda ze zmiennych $\mathbf{Y}_{i*}$ ma rozkład normalny, otrzymany model będziemy nazywać prostym modelem normalnym nadpopulacji.

Załóżmy, że jest dostępna macierz $\mathbf{x} = [x_{it}]$ (i=1,...N: t=1,...p) obserwacji p zmiennych pomocniczych $x=[x_1...x_p]$ w populacji. Wprowadźmy także oznaczenia: $\mathbf{x}^T=\left[\mathbf{x}_{1*}^T \dots \mathbf{x}_{N*}^T\right]$, gdzie $\mathbf{x}_{i*} = [x_{i1}...x_{ip}]$ oraz $\mathbf{x} = [\mathbf{x}_{*1}...\mathbf{x}_{*p}]$, gdzie $\mathbf{x}_{*t}^T = [x_{1t} \dots x_{Nt}]$.Przyjmijmy, że występujący w uprzednio prezentowanym wzorami (1.16)-(1.19) modelu transformacyjnym parametr $\mu = 0$ oraz dla każdego i=1,...N; j=1,...,m.

$$E(Y_{ij}) = b_{ij} = \mathbf{x}_{i*}\boldsymbol{\beta}_j \tag{1.23}$$

gdzie $\boldsymbol{\beta}_j = [\beta_{1j}...\beta_{pj}]^T$ jest wektorem parametrów strukturalnych regresji liniowej j-tej cechy reprezentowanej wektorem $\mathbf{Y}_{*j}$ względem zmiennych pomocniczych. Rozważane równania syntetycznie da się zapisać następująco:

$$E(\mathbf{Y}) = \mathbf{b} = \mathbf{x}\boldsymbol{\beta} \tag{1.24}$$

gdzie $\boldsymbol{\beta} = [\beta_1 ... \beta_m]$.

Oznaczmy przez $\mathbf{U} = \mathbf{Y} - E(\mathbf{Y})$ tzw. macierz resztową. Wówczas macierz losową **Y**, reprezentującą model regresyjny, zapisujemy równaniem:

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{U} \tag{1.25}$$

Momenty centralne drugiego rzędu składowych macierzy **U** określają wzory (1.19) i (1.21). Dodajmy, że występujące w tych wzorach współczynniki proporcjonalności a_{ij} są zwykle określane także jako funkcje cech pomocniczych. Określony model w uproszczonej wersji dla m=1 i ρ=0 był m.in. studiowany przez Royalla (1970).

1.4. Plany i schematy losowania próby

1.4.1. Definicje i własności podstawowe

Wszystkie prezentowane niżej definicje pochodzą z pracy Cassela, Sarndala i Wretmana (1977).

Definicja 1.11: Ciąg $\underline{s} = \{k_1,...,k_n\}$, gdzie $k_i \in \Omega$, nazywamy próbę uporządkowaną, przy czym n jest liczebnością próby. Zbiór $\underline{\mathcal{S}} = \underline{\mathcal{S}}(\Omega)$ wszystkich prób typu $\underline{s}$ nazywamy przestrzenią tych prób.

Definicja 1.12: Liczba niepowtarzających się w próbie $\underline{s}$ elementów jest oznaczana przez $v \le n$ i nazywana efektywną liczebnością próby $\underline{s}$.

Pomijając w próbie $\underline{s}$ porządek elementów oraz informację o ich powtarzaniu się otrzymujemy zbiór:

$$s = \{k : k \in \underline{s}\}$$

Definicja 1.13: Niepusty podzbiór s populacji Ω nazywamy próbą nieuporządkowaną bądź po prostu próbą, której liczebność oznaczamy przez n. Symbolem $\mathcal{S}$ oznaczamy przestrzeń prób typu s.

Próba s o liczebności n jest n elementową kombinacją bez powtórzeń wybraną z populacji Ω, a zatem zbiór $\mathcal{S}$ składa się z $\binom{N}{n}$ prób.

Definicja 1.14: Planem losowania próby uporządkowanej $\underline{s}$ nazywamy taki rozkład prawdopodobieństwa $P(\underline{s})$ określony na zbiorze $\underline{\mathcal{S}}$, że dla każdej próby $\underline{s} \in \underline{\mathcal{S}}$ zachodzi, iż

$$P(\underline{s}) \ge 0 \quad \text{i} \quad \sum_{\underline{s} \in \underline{\mathcal{S}}} P(\underline{s}) = 1 \tag{1.26}$$

W szczególności bezzwrotnie losowaną próbę z populacji charakteryzuje plan $P(\underline{s})>0$ dla $\underline{s} \in \underline{\mathcal{S}}^* \subset \underline{\mathcal{S}}$ i $\sum_{\underline{s} \in \underline{\mathcal{S}}^*} P(\underline{s}) = 1$.

Definicja 1.15: Planem losowania próby nieuporządkowanej s nazywamy rozkład prawdopodobieństwa P(s) określony na przestrzeni $\mathcal{S}$, który dla każdej $s \in \mathcal{S}$ spełnia warunki:

$$P(s) \geq 0 \text{ i } \sum_{s \in \mathcal{S}} P(s) = 1$$

Zauważmy, że plan losowania P(s) można utworzyć z planu $P(\underline{s})$, gdzie $\underline{s} \in \underline{S}^*$ następująco

$$P(s) = \sum_{\underline{s} \in \underline{S}^*(s)} P(\underline{s}) \tag{1.27}$$

przy czym $\underline{S}^*(s)$ jest zbiorem wszystkich wariancji bez powtórzeń $\underline{s}$, jakie można utworzyć jednocześnie na podstawie wszystkich elementów zbioru s, czyli każde dwie próby wchodzące w skład $\underline{S}^*(s)$ są utworzone z elementów zbioru s, a różnią się porządkiem występowania elementów próby s.

Próbę uporządkowaną $\underline{s} = \{k_1...,k_n\}$ możemy traktować jako realizację zmiennej losowej n wymiarowej $\underline{S} = \{K_1,...,K_n\}$. Jej rozkład prawdopodobieństwa w uproszczonej notacji $P(\underline{s}) = P(\underline{S}) = \underline{s}$ określa definicja 1.15. Każda składowa K_i (i=1,...,n) zmiennej losowej $\underline{S}$ przyjmuje wartości spośród liczb naturalnych {1,...,N}, które są jednocześnie numerami (etykietami) elementów populacji. Podobnie próbę nieuporządkowaną s można traktować jako realizację zbioru losowego **S.**

Określamy zbiór:

$$A(k_1,...,k_r) = \{s: k_i \in s, \text{ dla } i=1,...,r\}$$

Definicja 1.16: Prawdopodobieństwo $\pi_{k_1...k_r}$ wyboru (inkluzji) rzędu r-tego, czyli prawdopodobieństwo wyboru do próby s elementów populacji $k_1,...,k_r$ wyliczamy na podstawie wzoru:

$$\pi_{k_1...k_r} = \sum_{s \in A(k_1....k_r)} P(s) \tag{1.28}$$

Definicja[12] 1.17: Jeśli do próby $\underline{s}$ (s) dobierane są indywidualnie elementy z populacji oraz $P(\underline{s}) = \text{const}$ (P(s) - const) dla każdej $\underline{s} \in \underline{\mathcal{S}}$ ($s \in \mathcal{S}$), to $\underline{s}$ (s) jest nazywana prostą próbą uporządkowaną (nieuporządkowaną).

Plan losowania prostej próby uporządkowanej ma postać:

[12] Por. Kish (1965), s. 38-40.

$$\bigwedge_{\underline{s}\in\underline{\mathcal{S}}} \; P_1(\underline{s}) = N^{-n} \tag{1.29}$$

Drugi plan losowania uporządkowanej próby prostej przyporządkowuje dodatnie prawdopodobieństwa tylko próbom charakteryzującym się stałą liczebnością efektywną n, czyli należącym do zbioru $\mathcal{S}$:

$$\bigwedge_{\underline{s}\in\underline{\mathcal{S}}_*} \; P_2(\underline{s}) = \frac{(N-n)!}{N!} \tag{1.30}$$

Przypomnijmy, że zbiór $\mathcal{S}_* \subset \mathcal{S}$ zawiera n elementowe wariancje bez powtórzeń elementów wybranych z populacji Ω. Plan P_2 nazywamy planem losowania próby uporządkowanej z ustaloną liczebnością efektywną n.

Plan losowania prostej próby nieuporządkowanej otrzymujemy poprzez zawężenie przestrzeni prób $\mathcal{S}_*$ do przestrzeni $\mathcal{S}$. Plan ten przyporządkowuje stałe wartości elementom zbioru $\mathcal{S}$ w następujący sposób:

$$\bigwedge_{s\in\mathcal{S}} P_3(s) = \frac{1}{\binom{N}{n}} \tag{1.31}$$

Prawdopodobieństwa wyboru każdego elementu populacji do próby w przypadku planów P_2 i P_3 są takie same i wynoszą:

$$\pi_i^{(2)} = \pi_i^{(3)} = \frac{n}{N}, \quad \text{dla } i=1,...,N \tag{1.32}$$

natomiast w przypadku planu losowania P_1 dla każdego i=1,...,N:

$$\pi_i^{(1)} = 1-(1-N^{-1})^n \tag{1.33}$$

Czerniak (1971) wyprowadził prawdopodobieństwa inkluzji drugiego rzędu dla planu P_1. Prawdopodobieństwa losowania elementów populacji o numerach $k_1,...,k_r$ do próby n elementowej (n > r) dla planów P_2 i P_3. wyprowadził Herzel (1986). Dodajmy, że Stam (1978 i 1986) bada różnice między planami P_1 i P_3 na gruncie teorii informacji.

W pracy będą także rozważane nieproste plany losowania., które charakteryzują się tym, że nie są równomiernymi rozkładami prawdopodobieństwa. Plany tego rodzaju konstruuje się wtedy, gdy istnieją racjonalne przesłanki różnicowania prawdopodobieństw wyboru poszczególnych prób, które powinny prowadzić do podniesienia dokładności wnioskowania o parametrach opisowych[13].W tym celu wykorzystuje się z góry znane informacje o populacji, co zwykle sprowadza się do obserwacji wszystkich wartości cech pomocniczych, silnie

[13] Obszerny przegląd planów losowania można znaleźć w pracach Brewera i Hanifa (1983), Koopa (1963) i Vosa (1980).

związanych z cechami badanymi. Zaznaczmy jednak, że do informacji pomocniczych nie będziemy zaliczać zdobytej podczas losowania próby wiedzy o parametrze **y** populacji. Plany losowania takich prób są nazywane nieinformatywnymi lub niesekwencyjnymi, które dokładnie określają Cassel i in. (1977) oraz Basu (1971), s. 207. Tego typu planów w niniejszej pracy nie będziemy analizować.

W praktyce czasem jest wygodniej dokładnie nie określać planu losowania próby, lecz poprzestać na ustaleniu prawdopodobieństw doboru poszczególnych elementów populacji do próby przynajmniej pierwszego rzędu. Pozwala to już na wykorzystanie odpowiednich metod estymacji parametrów opisowych, chociaż pojawia się tu następny problem wyznaczenia planu próby na podstawie założonych uprzednio prawdopodobieństw inkluzji. Zagadnieniem tym zajmowali się m.in. Brewer i Hanif (1983), Chaudhuri (1971), Chaudhuri i Vos (1988), Sinha (1973) oraz Gabler i Schweigkoffer (1990). Dotychczas określiliśmy plan losowania próby, czyli prawdopodobieństwa, z jakim powinny być wybierane zestawy elementów populacji nazywane próbami. Teraz należy ustalić taki mechanizm losowania, który umożliwi wybór próby z populacji według ustalonego planu próbkowania. Próba jest tworzona poprzez kolejne losowe wybory poszczególnych elementów populacji, a zatem jest rzeczą naturalną traktować ją jako ciąg uporządkowanych elementów $\underline{s}$ o liczebności n. Przez $p(k_1)$ oznaczamy prawdopodobieństwo wylosowania do próby k_i-tego elementu populacji za pierwszym razem, natomiast przez $p(k_i \mid k_{i-1},...,k_1)$ oznaczmy prawdopodobieństwo wyboru elementu k_i populacji w i-tym losowaniu (i=2,...,n) pod warunkiem, że uprzednio do próby już wylosowano elementy o numerach: $k_{i-1},...,k_1$, przy czym (k_i=1,...,N dla i=1,...,n). Zatem:

$$p(k_1)\prod_{i=2}^{n} p(k_i|k_{i-1},\ldots,k_1) = P(\underline{s}) \tag{1.34}$$

przy czym k_i=1,...,N oraz i=1,...,n.

Definicja 1.18 [Cassel i in. (1977)]: Mechanizm losowania realizujący plan wyboru elementów populacji do próby $\underline{s}$ według danych powyższym wzorem prawdopodobieństw warunkowych nazywamy schematem losowania próby lub schematem próbkowania.

Schemat losowania próby umożliwia więc techniczną realizację założonego planu losowania próby.

Twierdzenie 1.5 [T.V.H. Rao (1962)]: Dla każdego planu próbkowania $P(\underline{s})$ istnieje przynajmniej jeden schemat losowania próby realizujący tenże plan.

Schemat losowania próby prostej, której plan losowania określa wzór (1.29), polega na zwrotnym losowaniu elementów z tym samym prawdopodobieństwem wynoszącym N^{-1}. Opisany schemat losowania próby nazywamy wariantem zwrotnym schematu losowania indywidualnego i nieograniczonego ze stałymi prawdopodobieństwami wyboru elementów do próby lub schematem losowania niezależnej próby prostej[14]. Dalej będziemy schemat ten nazywać schematem zwrotnego losowania próby prostej.

Schemat losowania próby realizujący plany $P_2(\underline{s})$ lub $P_3(s)$ określone odpowiednio wzorami (1.30) i (1.31) polega na wyborze bezzwrotnym kolejnych elementów do próby, tak aby prawdopodobieństwo wylosowania elementu o numerze k_i w i-tym losowaniu pod warunkiem, że wcześniej już wybrano elementy $k_1,...,k_{i-1}$ wynosiło:

[14] Por. np. pracę Pawłowskiego (1972).

$$p(k_i|k_{i-1},\ldots,k_1) = \frac{1}{N-i+1} \tag{1.35}$$

Przedstawiony schemat jest nazywany wariantem bezzwrotnym schematu losowania indywidualnego i nieograniczonego ze stałymi prawdopodobieństwami wyboru elementów populacji do próby lub schematem losowania zależnej próby prostej. Dalej także będziemy go nazywać schematem bezzwrotnym losowania próby prostej.

Własności innych popularnych schematów losowania można znaleźć w każdym podręczniku z metody reprezentacyjnej. Tutaj ograniczymy się do prezentacji szczególnych schematów losowania tzw. prób przenikających się i planów losowania prób zależnych od cech pomocniczych.

1.4.2. Schematy losowania prostych prób wielokrotnych

Próbę wielokrotną nieuporządkowaną traktujemy jako zbiór prób pojedynczych także nieuporządkowanych i oznaczamy symbolem: $s^{(h)} = \{s_1,...,s_h\}$. Składowe s_i (i=1,..,h) próby $s^{(h)}$ mogą być wybierane bezpośrednio z całej populacji bądź dana próba może być losowana z uprzednio już wybranej próby s_{i-1}. W efekcie składowe próby $s^{(h)}$ mogą być zależne albo niezależne.

Załóżmy, że próba prosta $\underline{s}_1$ jest losowana bezzwrotnie z całej populacji Ω, a następna próba prosta $\underline{s}_2$ o liczebności $n_2 \le n_1$ jest losowana spośród elementów zbioru s_1 będącego nieuporządkowaną próbą utworzoną z uprzednio już dobranej próby uporządkowanej $\underline{s}_1$. Z kolei próbę $\underline{s}_3$ losujemy spośród elementów próby s_2 itd. W końcu otrzymujemy próbę wielokrotną określaną ciągiem[15]: $\underline{s}^{(h)} = \{\underline{s}_1,...,\underline{s}_h\}$. Plan losowania pierwszej próby ma postać:

$$\bigwedge_{\underline{s}_1 \in \underline{\mathcal{S}}_{n_1}(\Omega)} \quad P(\underline{s}_1) = \frac{(N-n_1)!}{N!}$$

Warunkowe plany losowania pozostałych prób są dla i=2,,...,h postaci:

$$\bigwedge_{\underline{s}_i \in \underline{\mathcal{S}}_{n_i}(s_{i-1})} \quad P(\underline{s}_i|\underline{s}_{i-1},...,\underline{s}_1) = \frac{(n_{i-1}-n_i)!}{n_{i-1}!} \tag{1.36}$$

Stąd wynika już plan losowania próby wielokrotnej:

$$\bigwedge_{\underline{s}^{(h)} \in \underline{\mathcal{S}}_{**}} \quad P_4\left(\underline{s}^{(h)}\right) = \prod_{i=1}^{h} P_2\left(\underline{s}_i|\underline{s}_{i-1},\ldots,\underline{s}_1\right) = \prod_{i=1}^{h} \frac{\left(n_{i-1}-n_i\right)!}{n_{i-1}!} \tag{1.37}$$

gdzie: $\underline{s}_0 = \Omega$, $n_0 = N$, a $\underline{\mathcal{S}}_{**}$ jest przestrzenią prób składającą się z ciągów $\underline{s}^{(h)} = \{\underline{s}_1,...,\underline{s}_h\}$,

[15] Zob. Wywiał (1988 b).

gdzie $\underline{s}_i$ jest n_1 elementową wariacją bez powtórzeń wybraną spośród elementów zbioru s_{i-1} o liczebności $n_{i-1} \geq n_i$.

Wywiał (1992) analizował rozkłady brzegowe poszczególnych prób (i ich par) składowych próby wielokrotnej. Pomijając porządek elementów populacji tworzących próby składowe wielokrotnej $\underline{s}^{(h)}= \{\underline{s}_1,...,\underline{s}_h\}$ otrzymujemy ciąg prób nieuporządkowanych w postaci: $s^{(h)} = \{s_1,...,s_h\}$. Stąd, że $P(s_i) = n_i!P(\underline{s}_i)$ oraz z wzoru (1.36) otrzymujemy dla i=1,...,h warunkowe plany losowania prób:

$$\bigwedge_{s_1 \in \mathcal{S}_{n_i}(s_i - 1)} P\left(\underline{s}_i \middle| \underline{s}_{i-1},\ldots,\underline{s}_1\right) = \binom{n_{i-1}}{n_i}^{-1} \tag{1.38}$$

oraz plan losowania próby wielokrotnej $s^{(h)}$:

$$\bigwedge_{s^{(h)} \in \mathcal{S}_{**}} P_5\left(s^{(h)}\right) = \prod_{i=1}^{h}\binom{n_{i-1}}{n_i}^{-1} = q^{-1} \tag{1.39}$$

gdzie: $\mathcal{S}_{**} = \{s^{(h)} : s_1 \subseteq s_2 \subseteq ...\subseteq s_h\}$, $n_o = N$

Z konstrukcji ciągu $\underline{s}^{(h)}$ lub $s^{(h)}$ wynika, że te próby można również nazywać ciągami prób zawierających się lub krótko próbami zawierającymi się [16].

Przedstawiamy teraz schemat losowania ciągu nieuporządkowanych prób prostych, który jest różny od opisanego wyżej, chociaż realizuje ten sam plan losowania określony wyrażeniem (1.39)[17].

Próby składowe ciągu $s^{(h)}=\{s_1,...,s_h\}$ losujemy niejako w odwrotnej kolejności do poprzednio określonej. Oznacza to, że najpierw z populacji Ω jest bezzwrotnie losowania próbka prosta s_h o liczebności n_h. Potem ze zbioru Ω-s_h losujemy także bezzwrotnie próbę prostą s'_{h-1} i dołączamy ją do próby s_h tworząc tą drogą próbę prostą $s_{h-1} = s'_{h-1} \cup s_h$ o liczebności $n_{h1}= n'_{h-1} + n_h$. Trzecią próbę otrzymujemy podobnie losując również bezzwrotnie próbkę prostą s'_{h-2} o liczebności n'_{h-2} spośród elementów zbioru $\Omega - s_h - s'_{h-1} = \Omega - s_{h-1}$ i potem dołączamy ją do zbioru s_{h-1} otrzymując próbę $s_{h-2}= s'_{h-2} \cup s_{h-1}$ o liczebności $n_{h-2}= n'_{h-2} + n_{h-1}$. Procedurę tę powtarzamy aż do uzyskania próby $s_1= s'_1 \cup s_2$ o liczebności $n_1= n'_1 + n_2$. Wtedy warunkowy plan losowania próby dla i=1,...h ma postać:

$$\bigwedge_{s_i \in S_{n_i}(\Omega - s_i + 1)} P(s'_i \cup s_{i+1} | s_{i+1},\ldots,s_h) = \frac{1}{w_i} \tag{1.40}$$

gdzie:

---

[16] Kończak (1995) nazywa je próbami zstępującymi.

[17] Zob. Wywiał (1988 b).

$$w_i = \binom{N - n_{i+1}}{n_i - n_{i+1}}, \qquad n_i - n_{i+1} = n_i' \geq 0, \; n_{h+1}=0, \quad s_i = s_i' \cup s_{i+1}$$

Stąd już otrzymujemy plan losowania próby $s^{(h)}$:

$$\bigwedge_{s^{(h)} \in \mathcal{S}_{**}} P\left(s^{(h)}\right) = \prod_{i=1}^{h} w_i^{-1} = w^{-1} \tag{1.41}$$

gdzie:

$$\mathcal{S} = \{s^{(h)} : s_1 \subseteq s_2 \subseteq ... \subseteq s_h\}, \; n_o = N$$

Wywiał (1992) wykazał, że określone wyżej schematy losowania prób $s^{(h)} = \{s_1,...,s_h\}$ realizują ten sam plan P_5 losowania określony wyrażeniem (1.39).

1.4.3. Schemat losowania dla planu proporcjonalnego do średniej z próby

Przyjmijmy, że wartości nieujemnej zmiennej pomocniczej będące elementami wektora $\mathbf{x}=[x_1...x_N]$ są jednoznacznie przyporządkowane elementom populacji $\Omega=\{1,...,N\}$. Przypomnijmy, że przez $\bar{x}_{\underline{s}} = \frac{1}{n}\sum_{k \in \underline{s}} x_k$ i $\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$ oznaczamy wartości średnie zmiennej pomocniczej *x* obserwowane odpowiednio w próbie i w populacji. Zakładamy, że w próbie nie występują powtórzenia elementów, czyli n jest liczebnością efektywną.

Przyjmijmy, że plan losowania próby jest proporcjonalny do sumy wartości zmiennej pomocniczej w próbie, czyli $P(\underline{s}) \; \alpha \; n\bar{x}_{\underline{s}}$. Dokładne wyrażenie określające plan losowania próby ma postać:

$$P_6(\underline{s}) = \frac{n\bar{x}_s}{n\sum\limits_{\underline{s} \in \underline{\mathcal{S}}} \bar{x}_{\underline{s}}}$$

Prawdopodobieństwa planu losowania dla każdej próby $\underline{s} \in \underline{\mathcal{S}}$ mają postać [por. Konijn (1973), s. 250]:

$$P_6(\underline{s}) = \frac{1}{N^{(n)}} \frac{\bar{x}_s}{\bar{x}} = \frac{1}{n(N-1)^{(n-1)}} \frac{n\bar{x}_s}{N\bar{x}} \tag{1.42}$$

Prawdopodobieństwo wylosowania próby $\underline{s}$ jest więc proporcjonalne do udziału sumy wartości cechy pomocniczej obserwowanej w próbie do sumy wartości tej cechy w populacji.

Wywiał (1991b i 1992) wyprowadził prawdopodobieństwa inkluzji dowolnego rzędu oraz prawdopodobieństwa warunkowe charakteryzujące schemat losowania próby, które po-

niżej przytaczamy.Prawdopodobieństwo wylosowania do próby s za pierwszym razem elementu k-tego populacji wynosi:

$$p(k) = \frac{N-n}{n(N-1)}\frac{x_k}{N\bar{x}} + \frac{(n-1)}{n(N-1)} \tag{1.43}$$

p(k) jest także prawdopodobieństwem wylosowania k-tego elementu populacji niekoniecznie za pierwszym razem, lecz w dowolnym i-tym losowaniu (i=1,...,n). Zatem prawdopodobieństwo wejścia do próby s k-tego elementu wynosi π_k=np(k), co po odpowiednich prostych przekształceniach zapisujemy następująco:

$$\pi_k = \frac{N-n}{N-1}\frac{x_k}{N\bar{x}} + \frac{n-1}{N-1} \text{ lub} \qquad \pi_k = \frac{N-n}{(N-1)N}\frac{x_k-\bar{x}}{\bar{x}} + \frac{n}{N} \tag{1.44}$$

Stąd wnioskujemy, że prawdopodobieństwo wylosowania k-tego elementu populacji do próby zależy od odchylenia k-tej obserwacji zmiennej pomocniczej od jej średniej. Gdy odchylenie to jest dodatnie (ujemne), to prawdopodobieństwo dostania się do próby k-tego elementu jest większe (mniejsze), niż by to było w przypadku losowania próby prostej realizującej plan P_2 dany wzorem (1.30).

Prawdopodobieństwa inkluzji rzędu drugiego są następujące:

$$\pi_{kl} = \frac{(n-1)(N-n)}{(N-2)(N-1)}\frac{x_k+x_l}{N\bar{x}} + \frac{(n-1)(n-2)}{(N-1)(N-2)} \tag{1.45}$$

lub

$$\pi_{kl} = \frac{(n-1)(N-n)}{(N-2)(N-1)N}\frac{x_k+x_l-2\bar{x}}{\bar{x}} + \frac{n(n-1)}{N(N-1)} \tag{1.46}$$

Prawdopodobieństwa warunkowe charakteryzujące schemat losowania próby:

$$p(k_r|k_{r-1},\ldots,k_1) = \frac{(N-n)\sum_{i=1}^{r} x_{k_i} + N(n-r)\bar{x}}{(N-n)\sum_{i=1}^{r-1} x_{k_i} + N(n-r+1)\bar{x}}\frac{1}{N-r} \tag{1.47}$$

przy czym $k_r \neq k_{r-1} \neq k_1 = 1,\ldots,N$ oraz $r = 2,\ldots,n$. Prawdopodobieństwo wylosowania danego elementu populacji do próby za pierwszym razem określa wzór (1.43).

Dodajmy, że Lahiri (1951) proponuje inny schemat losowania realizujący rozważany tutaj plan wyboru próby.

1.4.4. Schemat losowania dla planu proporcjonalnego do nieobserwowanej w próbie sumy wartości zmiennej

Wywiał (1992) analizuje własności planu losowania próby proporcjonalnego do sumy wartości cechy *x* nie obserwowanych w próbie, czyli $P_7(\underline{s}) \alpha N\bar{x} - n\bar{x}_{\underline{s}}$. Plan ten można przekształcić do postaci:

$$P_7(\underline{s}) = \frac{N\bar{x} - n\bar{x}_s}{N^{(n)}(N-n)\bar{x}} \tag{1.48}$$

Stąd i na podstawie wzoru (1.42) wnioskujemy, że:

$$P_7(\underline{s}) = \frac{1}{(N-1)^{(n)}} - \frac{n}{N-n} P_6(\underline{s})$$

$$P_7(\underline{s}) = \frac{1}{(N-1)^{(n)}} \left(1 - \frac{n\bar{x}_s}{N\bar{x}}\right) \tag{1.49}$$

Prawdopodobieństwa inkluzji do drugiego rzędu włącznie:

$$\pi_k = \frac{n}{N-n} - \frac{nx_k}{(N-1)N\bar{x}} - \frac{(n-1)n}{(N-1)(N-n)} \tag{1.50}$$

lub $\pi_k = \frac{n}{N-1}\left(1 - \frac{x_k}{N\bar{x}}\right)$ lub $\pi_k = \frac{n}{N}\left(1 - \frac{x_k - \bar{x}}{(N-1)\bar{x}}\right)$ (1.51)

$$\pi_{kl} = \frac{n(n-1)}{(N-1)(N-2)}\left(1 - \frac{x_k + x_l}{N\bar{x}}\right) \tag{1.52}$$

lub $$\pi_{kl} = \frac{n(n-1)}{N(N-1)}\left(1 - \frac{(x_k - \bar{x}) + (x_l - \bar{x})}{(N-2)\bar{x}}\right) \tag{1.53}$$

Prawdopodobieństwa warunkowe schematu losowania:

$$p(k_r | k_{r-1}, \ldots, k_1) = \frac{1}{N-r} \frac{N(n-r)\bar{x} - \left((N-n)\sum_{i=1}^{r} x_{k_i} + (n-r)N\bar{x}\right)}{N(n-r-1)\bar{x} - \left((N-n)\sum_{i=1}^{r-1} x_{k_i} + (n-r)N\bar{x}\right)} \tag{1.54}$$

przy czym $k_1 \neq k_2 \neq ... \neq k_r = k_r = 1,...,N$; $r > 1$. Prawdopodobieństwo wylosowania za pierwszym razem k-tego (k = 1,...,N) elementu populacji do próby wynosi: $p(k) = \frac{1}{n}\pi_k$.

1.4.5. Schematy losowania dla planów zależnych od obserwowanej w próbie sumy kwadratów odchyleń wartości zmiennej od jej średniej w populacji

Kwadraty odchyleń wartości zmiennej pomocniczej od jej średniej w populacji oznaczne przez $z_k = (x_k - \bar{x})^2$ można traktować jako wartości nowej zmiennej pomocniczej. Wówczas wariancje cechy pomocniczej są równe odpowiednim średnim cechy z:

$$v_{\#s} = \frac{1}{n}\sum_{k\in s}(x_k - \bar{x})^2 = \frac{1}{n}\sum_{k\in s} z_k = \bar{z}_s , \qquad v = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2 = \frac{1}{N}\sum_{k=1}^{N} z_k = \bar{z}$$

Załóżmy, że $v > 0$. Zadanie polega na wyznaczeniu planu losowania próby proporcjonalnego do stopnia zróżnicowania zmiennej pomocniczej mierzonego wariancją $v_{\#s}$, czyli $P_8(\underline{s}) \alpha v_{\#s} = \bar{z}_s$.

Równości $v_{\#s} = \bar{z}_s, v = \bar{z}$ pozwalają na natychmiastowe wnioskowanie o własnościach planu losowania $P_8(\underline{s})$. Na podstawie wzoru (1.42) mamy:

$$P_8(\underline{s}) = \frac{1}{N^{(n)}}\frac{v_{\#s}}{v} \tag{1.55}$$

Korzystając ze wzorów (1.28) i (1.44) wyliczamy prawdopodobieństwo inkluzji rzędu pierwszego dla k=1,...,N

$$\pi_k = \frac{N-n}{N-1}\frac{(x_k - \bar{x})^2}{Nv} + \frac{n-1}{N-1}$$

lub

$$\pi_k = \frac{N-n}{(N-1)N}\frac{(x_k - \bar{x})^2 - v}{v} + \frac{n}{N} \tag{1.56}$$

Stąd wynika, że prawdopodobieństwo doboru do próby k-tego elementu populacji może być mniejsze lub większe od wartości $\frac{n}{N}$, która jest prawdopodobieństwem inkluzji rzędu pierwszego dla próby prostej losowanej bezzwrotnie. Szanse wejścia k-tego elementu do próby są tym większe, im bardziej k-ta wartość zmiennej pomocniczej różni się od wartości średniej w populacji.

Prawdopodobieństwa inkluzji rzędu drugiego, jak i prawdopodobieństwa warunkowe schematu losowania próby wyliczamy na podstawie wzorów (1.45)-(1.47), zastępując w nich obserwacje cechy pomocniczej x przez kwadrat jej odchylenia od średniej w populacji.

Plan losowania P_8 daje większe szanse doboru prób, które charakteryzują się dużym zróżnicowaniem wartości zmiennej pomocniczej względem jej średniej w populacji.

Drugi plan losowania jest proporcjonalny do nie obserwowanej w próbie sumy kwadratów odchyleń wartości cechy pomocniczej od jej średniej w populacji, czyli $P_9(\underline{s}) \alpha Nv - nv_{\#\underline{s}}$. Wtedy:

$$P_9(\underline{s}) = \frac{1}{(N-1)^{(n)}}\left(1 - \frac{nv_{\#s}}{Nv}\right) \tag{1.57}$$

Podstawiając we wzorach (1.50)-(1.53) w miejsce x_k i $\bar{x}$ odpowiednio wielkości $(x_k - \bar{x})^2$ i v wyliczamy prawdopodobieństwa inkluzji do drugiego rzędu.

1.4.6. Schemat losowania dla planu proporcjonalnego do wariancji z próby

Wariancję z próby zmiennej pomocniczej oznaczamy przez

$$v_{\underline{s}} = v_s = \frac{1}{n}\sum_{k\in s}\left(x_k - \bar{x}\right)^2 \qquad \text{lub} \qquad v_{*\underline{s}} = v_{*s} = \frac{n}{n-1}v_s$$

Niech $e_k = x_k - \bar{x}$ dla k=1,...,N. Wtedy wariancję w populacji zmiennej pomocniczej można zapisać w postaci:

$$v = \frac{1}{N}\sum_{k=1}^{N} e_k^2 \qquad \text{lub} \qquad v_* = \frac{1}{N-1}\sum_{k=1}^{N} e_k^2 = \frac{N}{N-1}v$$

Singh i Srivastava (1980) zaproponowali plan losowania próby proporcjonalny do wariancji $v_{\underline{s}}$, czyli $P_{10}(\underline{s}) \alpha v_{\underline{s}}$. Jego dokładna postać jest następująca:

$$P_{10}(\underline{s}) = \frac{n(N-1)}{N(n-1)N^{(n)}}\frac{v_s}{v} = \frac{1}{N^{(n)}}\frac{v_{*s}}{v_*} \tag{1.58}$$

Odpowiadający rozkładowi prawdopodobieństwa $P_{10}(\underline{s})$ plan losowania próby nieuporządkowanej $P_{10}(s)$ ma postać: $P_{10}(s)=n!P_{10}(\underline{s})$, a zatem

$$P_{10}(s) = \frac{1}{\binom{N}{n}}\frac{v_{*s}}{v_*}$$

Singh i Srivastava (1980) określili dwa schematy losowania realizujące plan $P_{10}(s)$. Niżej prezentujemy inny schemat losowania próby realizujący obydwa plany $P_{10}(\underline{s})$ i $P_{10}(s)$ oraz prawdopodobieństwa inkluzji dowolnego rzędu wyprowadzone przez Wywiała (1992).

Prawdopodobieństwo doboru dowolnych elementów populacji $k_1 \neq k_2 \neq,...,\neq k_r = 1,...,N$ w pierwszych r losowaniach ma postać:

$$p(k_1,...,k_r) = \frac{(N-r-2)!}{n(n-1)N!}\left\{(N-n)\left((n-1)(N-r+1) - 2(r-1)\sum_{i=1}^{r}\left(x_{k_i} - \bar{x}\right)^2 + \right.\right.$$

$$\left.\left. -(N-n)(N-n-1)\sum_{i\neq j=1}^{r}\sum^{r}(x_{k_i} - \bar{x})(x_{k_j} - \bar{x}) + (n-r)(N-1)\left((n-1)(N-r) - r\right)v_*\right\} \qquad (1.59)$$

Prawdopodobieństwa inkluzji do drugiego rzędu są następujące:

$$\pi_k = \frac{(n-1)(N-1)-1}{(N-2)N} + \frac{(N-n)}{(N-1)(N-2)}\frac{(x_k - \bar{x})^2}{v_*}$$

lub

$$\pi_k = \frac{n}{N} + \frac{N-n}{N-2}\left\{\frac{(x_k - \bar{x})^2}{(N-1)v_*} - \frac{1}{N}\right\} \qquad (1.60)$$

Z kolei prawdopodobieństwa inkluzji rzędu drugiego dla elementów populacji $k \neq t = 1,....,N$ mają postać:

$$\pi_{kt} = \frac{(N-4)!}{N!v_*}\left\{(N-n)\left((n-1)(N-1)+2\right)\left((x_k - \bar{x})^2 + (x_t - \bar{x})^2\right) + \right.$$

$$\left. -2(N-n)(N-n-1)(x_k - \bar{x})(x_t - \bar{x}) + 2(n-2)(N-1)\left((n-1)(N-2)-1\right)v_*\right\} \qquad (1.61)$$

Prawodpodobieństwa warunkowe $P(k_r | k_{r-1},...k_1)$ schematu losowania realizującego plan P_{10} wyboru próby można łatwo obliczyć na podstawie wzoru (1.59).

1.4.7. Schemat losowania dla planu proporcjonalnego do funkcji wariancji z próby i populacji

Rozważmy plan losowania próby proporcjonalny do różnicy między sumą kwadratów odchyleń wszystkich obserwacji zmiennej x od jej średniej w populacji i sumą kwadratów odchyleń wartości zmiennej pomocniczej w próbie od jej średniej w tej próbie, czyli:

$$P_{11}(\underline{s}) \alpha\ Nv - nv_{\underline{s}} = a(\underline{s})$$

Niech $u=\Omega-s$, natomiast $\underline{u}$ niech będzie szczególną permutacją elementów zbioru u. Wykażemy, że $a(\underline{s}) \geq 0$:

$$a(\underline{s}) = a(s) = \sum_{k\in s}\left(\left(x_k - \bar{x}_s\right) + \left(\bar{x}_s - \bar{x}\right)\right)^2 + \sum_{k\in u}\left((x_k - \bar{x}_u) + (\bar{x}_u - \bar{x}\right)^2 - \sum_{k\in s}(x_k - \bar{x}_s)^2 =$$
$$= n(\bar{x}_s - \bar{x})^2 + \sum_{k\in u}(x_k - \bar{x})^2 + (N-n)(\bar{x}_u - \bar{x})^2 \geq 0$$

Stąd wnioskujemy, że:

$$\sum_{\underline{s}\in\underline{S}} a(\underline{s}) = (N-1)N^{(n)}v_* - (n-1)N^{(n)}v_* = (N-n)N^{(n)}v_*$$

Dalej już mamy:

$$P_{11}(\underline{s}) = \frac{N-1}{(N-n)N^{(n)}} - \frac{(n-1)v_{*s}}{(N-n)N^{(n)}v_*} \tag{1.62}$$

Prawdopodobieństwo doboru do próby w r ustalonych losowaniach elementów populacji o numerach $k_1,...,k_r$ określa wzór:

$$p(k_1,..,k_r) = \frac{N-1}{N-n}\left(\frac{(N-r)!}{N!} - \frac{n-1}{N-1}p_{20}(k_1,\ldots,k_r)\right) \tag{1.63}$$

Na podstawie otrzymanego wzoru można wyliczyć już prawdopodobieństwa warunkowe $p(k_r|k_{r-1},...,k_1)$ schematu losowania realizującego plan P_{11}.

Prawdopodobieństwa inkluzji rzędu pierwszego dla k=1,...,N określają wzory:

$$\pi_k = \frac{n}{N} + \frac{n-1}{N-2}\left(\frac{1}{N} - \frac{(x_k - \bar{x})^2}{(N-1)v_*}\right) \tag{1.64}$$

lub
$$\pi_k = \frac{n}{N} + \frac{n-1}{N(N-2)}\left(1 - \frac{(x_k - \bar{x})^2}{v}\right)$$

Prawdopodobieństwa inkluzji rzędu drugiego są postaci:

$$\pi_{kt} = \frac{n(n-1)}{N(N-n)} - \frac{n-1}{N-1}\pi_{kt}^{(10)} \tag{1.65}$$

przy czym $k \neq t = 1,...,N$, natomiast, przez $\pi_{kt}^{(10)}$ oznaczono prawą stronę równania (1.61).

1.4.8. Schemat losowania dla planu proporcjonalnego do kwadratu odchylenia między średnimi z próby i z populacji

Zgodnie z powyższym tytułem szukamy planu losowania próby $P_{12}(\underline{s})\alpha\left(\overline{x}_{\underline{s}} - \overline{x}\right)^2$. Wtedy mamy:

$$\sum_{\underline{s}\in\underline{\mathcal{S}}}\left(\overline{x}_{\underline{s}} - \overline{x}\right)^2 = n^{-2}\sum_{\underline{s}\in\underline{\mathcal{S}}}\left(\sum_{k\in\underline{s}} e_k\right)^2 = \frac{N-n}{nN}N^{(n)}v_* \tag{1.66}$$

Stąd wynika, że

$$P_{12}(\underline{s}) = \frac{Nn}{(N-n)N^{(n)}}\frac{(\overline{x}_s - \overline{x})^2}{v_*} \tag{1.67}$$

Wyprowadzimy teraz prawdopodobieństwo $p(k_1,...,k_r)$ dobrania do próby podczas r ustalonych losowań elementów populacji: $k_1,...,k_r$. Wyznaczamy je zakładając, co nie zmniejszy ogólności wyniku, że mamy do czynienia z pierwszymi r losowaniami, w wyniku których są dobierane do próby elementy populacji o kolejnych numerach: 1,..,r. Niech u = s - {1,...,r}. Wtedy definiując zbiory:

$$\underline{A}(1,...,r) = \{\underline{s} : i \in \underline{s}\ ,\quad \text{dla } i = 1,...,r\};$$

$$A(1,...,r) = \{s : i \in s,\quad \text{dla } i = 1,...,r\}$$

mamy:

$$p(1,\ldots,r) = \sum_{\underline{s}\in\underline{A}(1,...,r)} P_{12}(\underline{s}) = (n-r)!\sum_{s\in A(1,...,r)} P_{12}(s) = c\sum_{s\in A(1,...,r)}\left(\sum_{k\in s}(x_k - \overline{x})\right)^2 =$$

$$= c\sum_{s\in A(1,...,r)}\left(\sum_{k\in s} e_k\right)^2$$

przy czym: $$c = \frac{N(n-r)!}{n(N-n)N^{(n)}v_*}$$

$$p(1,\ldots,r) = c\sum_{s\in A(1,...,r)}\left(\sum_{k=1}^{r} e_k + \sum_{k\in u} e_k\right)^2 = c\sum_{s\in A(1,...,r)}\left(\left(\sum_{k=1}^{r} e_k\right)^2 + 2\sum_{k=1}^{r} e_k\sum_{k\in u} e_k + \sum_{k\in u} e_k^2 +\right.$$

$$\left. + \sum_{k,t\in u}\sum e_k e_t\right) = c\left\{\binom{N-r}{n-r}\left(\sum_{k=1}^{r} e_k\right)^2 + 2\binom{N-r-1}{n-r-1}\sum_{k=1}^{r} e_k\sum_{k=r+1}^{N} e_k + \binom{N-r-1}{n-r-k}\sum_{k=r+1}^{N} e_k^2 +\right.$$

$$+\binom{N-r-2}{n-r-2}\sum_{k\neq t=r+1}^{N}\sum^{N}e_k e_t\Big\}=c\Bigg\{\binom{N-r}{n-r}\left(\sum_{k=1}^{r}e_k\right)^2+2\binom{N-r-1}{n-r-1}\sum_{k=1}^{r}e_k\left(\sum_{k=1}^{N}e_k-\sum_{k=1}^{r}e_k\right)+$$

$$+\binom{N-r-1}{n-r-1}\left(\sum_{k=1}^{N}e_k^2-\sum_{k=1}^{r}e_k^2\right)+\binom{N-r-2}{n-r-2}\left[\left(\sum_{k=1}^{N}e_k-\sum_{k=1}^{r}e_k\right)^2-\sum_{k=r+1}^{N}e_k^2\right]\Bigg\}=$$

$$c\frac{(N-r-2)!(N-n)}{(N-n)!(n-r)!}\left\{(N-n-1)\left(\sum_{k=1}^{r}e_k\right)^2-(n-r)\sum_{k=1}^{r}e_k^2+(n-r)(N-1)v_*\right\}=$$

$$=\frac{(N-r-2)!}{n(N-1)!\,v_*}\left\{(N-n-1)\left(\sum_{k=1}^{r}\left(x_k-\bar{x}\right)\right)^2-(n-r)\sum_{k=1}^{r}\left(x_k-\bar{x}\right)^2+(n-r)(N-1)v_*\right\}$$

Uogólniając otrzymany wynik wnioskujemy, że prawdopodobieństwo wyboru elementów populacji $k_1 \neq k_2 \neq ... \neq k_r$ w dowolnych ustalonych r losowaniach wynosi:

$$p(k_1,\ldots,k_r)=\frac{(N-r-2)!}{n(N-1)!v_*}\Bigg((N-n-1)\left(\sum_{i=1}^{r}(x_{k_i}-\bar{x})\right)^2-(n-r)\sum_{i=1}^{r}(x_{k_i}-\bar{x})^2+$$

$$+(n-r)(N-1)v_*\Bigg) \tag{1.68}$$

Niech

$$\bar{x}(r)=\frac{1}{r}\sum_{i=1}^{r}x_{k_i}\,,\qquad v_*(r)=\frac{1}{r-1}\sum_{i=1}^{r}(x_{k_i}-\bar{x}(r))^2$$

przy czym $k_i \neq k_j = 1,...,N$ dla $i \neq j = 1,...,r \leq n$. Wówczas mamy:

$$\left(\sum_{i=1}^{r}\left(x_{k_i}-\bar{x}\right)\right)^2=\left(\sum_{i=1}^{r}\left(x_{k_i}-\bar{x}(r)\right)+\left(\bar{x}(r)-\bar{x}\right)\right)^2=\left(\sum_{i=1}^{r}\left(x_{k_i}-\bar{x}(r)\right)\right)^2+$$

$$+2r(\bar{x}(r)-\bar{x})\sum_{i=1}^{r}(x_{k_i}-\bar{x}(r))+r^2(\bar{x}(r)-\bar{x})^2$$

Stąd, że $\sum_{i=1}^{r}(x_{k_i}-\bar{x}(r))=0$ wynika, iż

$$\left(\sum_{i=1}^{r}\left(x_{k_i}-\bar{x}\right)\right)^2 = r^2\left(\bar{x}(r)-\bar{x}\right)^2 \tag{1.69}$$

Z kolei:

$$\sum_{i=1}^{r}(x_{k_i}-\bar{x})^2 = \sum_{i=1}^{r}\left(\left(x_{k_i}-\bar{x}(r)\right)+\left(\bar{x}(r)-\bar{x}\right)\right)^2 = \sum_{i=1}^{r}(x_{k_i}-\bar{x}(r))^2 -$$

$$-2(\bar{x}(r)-\bar{x})\sum_{i=1}^{r}(x_{k_i}-\bar{x}(r)) + r(\bar{x}(r)-\bar{x})^2$$

W końcu otrzymujemy:

$$\sum_{i=1}^{r}(x_{k_i}-\bar{x})^2 = (r-1)v_*(r) + r(\bar{x}(r)-\bar{x})^2 \tag{1.70}$$

Podstawiając wzory (1.69) i (1.70) do (1.68) po odpowiednich przekształceniach otrzymujemy:

$$p(k_1,\ldots,k_r) =$$

$$= \frac{(N-r-2)!}{n(N-1)!v_*}\left(r(Nr-nr-n)(\bar{x}(r)-\bar{x})^2 + (n-r)(r-1)v_*(r) + (n-r)(N-1)v_*\right) \tag{1.71}$$

Na podstawie otrzymanego wyrażenia bądź wzoru (1.68) już łatwo obliczamy prawdopodobieństwa warunkowe $p(k_r| k_{r-1},\ldots,k_1)$ schematu losowania realizującego plan P_{12} wyboru próby.

Prawdopodobieństwa inkluzji rzędu pierwszego są postaci:

$$\pi_k = \frac{N-2n}{(N-1)(N-2)}\frac{(x_k-\bar{x})^2}{v_*} + \frac{n-1}{N-2}$$

lub

$$\pi_k = \frac{n}{N} + \frac{N-2n}{N(N-2)}\left(\frac{\left(x_k-\bar{x}\right)^2}{v} - 1\right) \tag{1.72}$$

przy czym k = 1,...,N.

Prawdopodobieństwo doboru pary $k\neq t=1,\ldots,N$ elementów populacji do próby wynosi:

$$\pi_{kt} = \frac{(n-1)(N-n-1)(x_k-\bar{x}+x_t-\bar{x})^2}{(N-1)(N-2)(N-3)v_*} - \frac{(n-1)(n-2)(x_k-\bar{x})^2 + (x_t-\bar{x})^2}{(N-1)(N-2)(N-3)v_*} +$$

$$+\frac{(n-1)(n-2)}{(N-2)(N-3)} \tag{1.73}$$

1.4.9. Schemat losowania dla planu proporcjonalnego do malejącej funkcji kwadratu odchylenia między średnimi z próby i populacji

Studiowany tutaj plan losowania próby $P_{13}(\underline{s})$ jest proporcjonalny do różnicy między sumą kwadratów odchyleń wartości zmiennej pomocniczej od jej średniej w populacji i n-krotną wartością kwadratu odchylenia średniej z próby zmiennej pomocniczej od jej średniej w populacji, czyli

$$P_{13}(\underline{s})\alpha(N-1)v_* - n(\overline{x}_{\underline{s}} - \overline{x})^2$$

Nierówność $P_{13}(\underline{s}) \geq 0$ wykazujemy podobnie jak uprzednio nierówność $a(\underline{s})\geq 0$. Na podstawie wzoru (1.66) wyliczamy, że

$$\sum_{\underline{s}\in\underline{\mathcal{S}}}\left((N-1)v_* - n(\overline{x}_{\underline{s}} - \overline{x})^2\right) = N^{(n)}v_*\frac{N(N-2)+n}{N}$$

Stąd wynika, że

$$P_{13}(\underline{s}) = \frac{N(N-1)}{(N(N-2)+n)N^{(n)}} - \frac{nN(\overline{x}_s - \overline{x})^2}{N^{(n)}(N(N-2)+n)v_*} \tag{1.74}$$

lub

$$P_{13}(\underline{s}) = \frac{N(N-1)}{(N(N-2)+n)N^{(n)}} - \frac{N-n}{N(N-2)+n}P_{12}(\underline{s}) \tag{1.75}$$

przy czym plan $P_{12}(\underline{s})$ określa wzór (1.67).

Prawdopodobieństwa doboru do próby w r ustalonych losowaniach elementów populacji $k_1,...,k_r$ obliczamy za pomocą wzorów (1.67), (1.71), (1.74) i (1.75) następująco:

$$p(k_1,\ldots,k_r) = \sum_{\underline{s}\in\underline{A}(1,\ldots,r)}P_{13}(\underline{s}) = (n-r)!\sum_{s\in A(1,..,r)}P_{13}(s) =$$

$$= \frac{N(N-1)(N-r)^{(n-r)}}{(N(N-2)+n)N^{(n)}} - \frac{N-n}{N(N-2)+n}\sum_{\underline{s}\in\underline{A}(1,\ldots,r)}P_{12}(\underline{s})$$

$$p(k_1,\ldots,k_r) = \frac{(N-r)!}{(N(N-2)+n)(N-2)!} +$$

$$-\frac{(N-n)(N-r-2)!}{n(N(N-2)+n)(N-1)!v_*}\left\{(N-n-1)\left(\sum_{i=1}^{r}(x_{k_i}-\overline{x})\right)^2 +\right.$$

$$\left.(n-r)\sum_{i=1}^{r}(x_{k_i}-\overline{x})^2 + (n-r)(N-1)v_*\right\} \tag{1.76}$$

Stąd można już obliczyć potrzebne prawdopodobieństwa warunkowe schematu losowania realizującego plan losowania P_{13}.

Prawdopodobieństwa inkluzji są następujące:

$$\pi_k = \frac{n}{N} - \frac{(N-n)(N-2n)}{(N(N-2)+n)(N-2)N}\left(\frac{(x_k - \bar{x})^2}{v_*}\right), \qquad k = 1,...,N \tag{1.77}$$

$$\pi_{kt} = \frac{n(n-1) - (N-n)\pi_{kt}^{(12)}}{N(N-2)+n}, \qquad k \neq t = 1,...,N \tag{1.78}$$

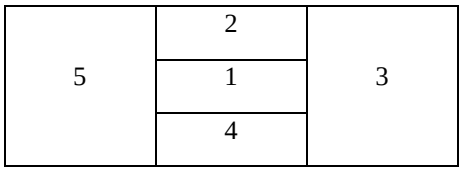
przy czym przez $\pi_{kt}^{(12)}$ oznaczono prawą stronę wzoru (1.73).

Plan losowania próby P_{13} w przeciwieństwie do planu P_{12} preferuje wybór próby, z której średnia cechy pomocniczej mało odchyla się od średniej tej zmiennej w populacji.

1.4.10. Plany losowania próby zależne od sąsiedztwa elementów w populacji przestrzennej

Załóżmy, że wzajemna pozycja elementów populacji w przestrzeni jest określona. Przykładowo, geograficzne ułożenie gmin względem siebie jest ustalone. Ułożenie elementów populacji można identyfikować za pomocą macierzy sąsiedztwa, którą oznaczamy symbolem $\mathbf{A}=[a_{ij}]$. Jeśli elementy populacji o numerach (i,j) sąsiadują (nie sąsiadują) ze sobą, to $a_{ij}=1$ ($a_{ij}=0$).

Rozważmy następującą populację:



Wtedy macierz sąsiedztwa ma postać:

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

Przestrzeń trójelementowych prób nieuporządkowanych ma postać:

S={(1,2,3); (1,2,4); (1,2,5); (2,3,4); (2,3,5); (3,4,5); (1,2,4); (2,4,5); (1,4,5)}

Macierz sąsiedztwa dla elementów tworzących próbę oznaczamy przez $A(s)=[a_{ij}(s)]$. Wtedy dla przestrzeni prób S mamy:

$$A(1,2,3)=\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix};\qquad A(1,2,4)=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\end{bmatrix};\qquad A(1,2,5)=\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}$$

$$A(2,3,4)=\begin{bmatrix}1&1&0\\1&1&1\\0&1&1\end{bmatrix};\qquad A(2,3,5)=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\end{bmatrix};\qquad A(3,4,5)=\begin{bmatrix}1&1&0\\1&1&1\\0&1&1\end{bmatrix}$$

$$A(1,3,5)=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\end{bmatrix};\qquad A(2,4,5)=\begin{bmatrix}1&0&1\\0&1&1\\1&1&1\end{bmatrix};\qquad A(1,4,5)=\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}$$

Zamierzamy tak losować próbę, by preferować wejście do niej elementów sąsiadujących ze sobą. Odpowiadający temu postulatowi plan losowania można sformułować następująco:

$$P_{14}(S)=\frac{\sum_{i>j} a_{ij}(S)}{\sum_{s\in S}\sum_{i>j} a_{ij}(S)} \tag{1.79}$$

Wtedy:

$$P_{14}(1,2,3)=P_{14}(1,2,5)=P_{14}(1,3,4)=P_{14}(1,4,5)=\frac{1}{2}$$

$$P_{14}(1,2,4)=P_{14}(2,3,4)=P_{14}(2,3,5)=P_{14}(3,4,5)=P_{14}(1,3,5)=P_{14}(2,4,5)=\frac{1}{12}$$

Prawdopodobieństwa inkluzji pierwszego i drugiego rzędu są następujące:

$$\pi_1=\frac{2}{3};\qquad \pi_2=\pi_3=\pi_4=\pi_5=\frac{7}{12}$$

$$\pi_{12}=\pi_{13}=\pi_{14}=\pi_{15}=\frac{1}{3};\qquad \pi_{23}=\pi_{25}=\pi_{34}=\pi_{45}=\frac{7}{24};\qquad \pi_{24}=\pi_{35}=\frac{1}{4}$$

W przypadku zwrotnego losowania prawdopodobieństwo wylosowania danego elementu populacji ustalamy proporcjonalnie do sumy elementów, z którymi graniczy analizowany element, a więc:

$$p_{15}(k) = \frac{\sum_{j=1}^{N} a_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}} \tag{1.80}$$

W przypadku rozważanej wyżej populacji mamy:

$$P_{15}(1) = \frac{5}{21}, \qquad P_{15}(2)=P_{15}(3)=P_{15}(4)=P_{15}(5)=\frac{4}{21}$$

Załóżmy teraz, że wybierana próba ma być tak losowana, aby było preferowane wejście do niej elementów nie sąsiadujących ze sobą. Plan losowania takiej próby określa więc następujące wyrażenie:

$$P_{16}(S)\alpha \frac{1}{2}\left(n^2 - n\right) + \alpha - \sum_{i>j} a_{tj}$$

Parametr α wprowadzono po to, by $P_{16}(s) > 0$. Zatem:

$$P_{16}(s) = \frac{\frac{1}{2}n(n-1) + \alpha - \sum_{i>j} a_{tj}}{\binom{N}{n}\left[\frac{1}{2}n(n-1) + \right] + \beta} \tag{1.81}$$

$$P_{16}(s) = \frac{\frac{1}{2}n(n-1) + \alpha - \beta P_1}{\binom{N}{n}\left[\frac{1}{2}n(n-1) + \alpha\right] + \beta} \tag{1.82}$$

gdzie:

$$\beta = \sum_{s \in S}\sum_{i>j} a_{tj}$$

Łatwo można wykazać, że wraz z nieograniczonym wzrostem α plan P_{16} zmierza do planu losowania próby prostej losowanej bezzwrotnie. Zatem

$$\lim_{\alpha \to \infty} P_{16} = \frac{1}{\binom{N}{n}} \tag{1.83}$$

W przypadku wcześniej rozważanej przykładowej populacji dla $\alpha = 0.1$ mamy:

$$P_{16}(1,2,3) = P_{16}(1,2,5) = P_{16}(1,3,4) = P_{16}(1,4,5) = \frac{1}{16}$$

$$P_{16}(1,2,4) = P_{16}(2,3,4) = P_{16}(2,3,5) = P_{16} =(3,4,5) = P_{16}(1,3,5) = P_{16}(2,4,5) = \frac{1}{8}$$

$$\pi_1 = \frac{13}{35}; \qquad \pi_2 = \pi_3 = \pi_4 = \pi_5 = \frac{23}{35}$$

$$\pi_{12} = \pi_{13} = \pi_{14} = \pi_{15} = \frac{13}{70}; \qquad \pi_{23} = \pi_{25} = \pi_{34} = \pi_{45}\frac{23}{70}; \qquad \pi_{24} = \pi_{35} = \frac{33}{70}$$

Gdy $\alpha = 0.5$, to

$$P_{16}(1,2,3) = P_{16}(1,2,5) = P_{16}(1,3,4) = P_{16}(1,4,5) = \frac{1}{22}$$

$$P_{16}(1,2,4) = P_{16}(2,3,4) = P_{16}(2,3,5) = P_{16}(3,4,5) = P_{16}(1,3,5) = P_{16}(2,4,5) = \frac{3}{22}$$

$$\pi_1 = \frac{1}{2}; \qquad \pi_2 = \pi_3 = \pi_4 = \pi_5 = \frac{5}{8}$$

$$\pi_{12} = \pi_{13} = \pi_{14} = \pi_{15} = \frac{5}{22}; \qquad \pi_{23} = \pi_{25} = \pi_{34} = \pi_{45} = \frac{7}{22}; \qquad \pi_{24} = \pi_{35} = \frac{9}{22}$$

W przypadku losowania zwrotnego prawdopodobieństwa wyboru do próby poszczególnych elementów populacji można określić tak, by większe szanse dostania się do próby miały elementy nie sąsiadujące ze sobą:

$$p_{17}(k) = \frac{N + 1 - \sum_{t=1}^{N} a_{kt}}{N(N+1) - \sum_{k=1}^{N}\sum_{t=1}^{N} a_{kt}} \tag{1.84}$$

W przypadku rozważanej populacji mamy:

$$p_{17}(1) = \frac{1}{9}; \qquad p_{17}(2) = p_{17}(3) = p_{17}(4) = p_{17}(5) = \frac{2}{9}$$

1.4.11. Wybrane plany losowania z populacji uporządkowanej.

Graficznie N-elementową populację uporządkowaną prezentuje rysunek:

1	2	3	...	N-1	N

Charakteryzuje ją następująca macierz sąsiedztwa:

$$A=\begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 1 & 1 & \dots & 0 & 0 \\ . & . & . & . & \dots & . & . \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \end{bmatrix}$$

Plan preferujący wejście do próby elementów sąsiadujących ze sobą konstruujemy następująco. Oznaczmy przez $s=(i_1,...,i_n)$ próbę, przy czym ustalmy, co nie umniejsza otrzymanych dalej wyników, że $i_1 < i_2 < ... < i_n$. Niech:

$$\delta_j(s)=\begin{cases} 1, & \text{gdy } i_j - i_{j+1}=1 \\ 0, & \text{gdy } i_j - i_{j+1}>1 \end{cases}$$

Wtedy postulowany plan określamy wyrażeniem:

$$P_{18}(s)\alpha\varepsilon+\sum_{j=1}^{n-1}\delta_j(s) \tag{1.85}$$

Zatem:

$$P_{18}(s)=\frac{\varepsilon+\sum_{j=1}^{n-1}\delta_j(s)}{\binom{N}{n}\varepsilon+\sum_{s\in S}\sum_{j=1}^{n-1}\delta_j(s)} \tag{1.86}$$

W szczególności niech n=2. Wtedy:

$$P_{18}(s)\alpha P_{18}(i_1,i_2)\alpha\theta_1(s)=\begin{cases} \varepsilon+1, & \text{gdy } i_i - i_2=1 \\ \varepsilon, & \text{gdy } i_1 - i_2>1 \end{cases}$$

Wtedy:

$$\sum_{s\in S}\theta_1(s)=\frac{(N-1)(2+N\varepsilon)}{2}$$

W końcu:

$$P_{18}(s)=\begin{cases}\dfrac{2(\varepsilon+1)}{(N-1)(2+N\varepsilon)}, & \text{gdy } i_1-i_2=1\\ \dfrac{2\varepsilon}{(N-1)(2+N\varepsilon)}, & \text{gdy } i_1-i_2>1\end{cases} \tag{1.87}$$

$$\pi_1=P_{18}(1,2)+\sum_{i=3}^{N}P_{18}(1,i)=\frac{2(N\varepsilon+1-\varepsilon)}{(N-1)(2+N\varepsilon)}, \qquad \pi_1=\pi_N \tag{1.88}$$

$$\pi_k = P_{18}(k,k+1)+P_{18}(k-1,k)+\sum_{t=1}^{k-2}P_{18}(i,k)+\sum_{i=k+2}^{N}P_{18}(k,i)=\frac{2(2+N\varepsilon-\varepsilon)}{(N-1)(2+N\varepsilon)} \tag{1.89}$$

przy czym k=1,...N-1.

Plan losowania preferujący wybór do próby elementów nie sąsiadujących ze sobą konstruujemy następująco:

$$P_{19}(s)\alpha\varepsilon+n-1-\sum_{j=1}^{n-1}\delta_j(s), \qquad \varepsilon>0 \tag{1.90}$$

Zatem:

$$P_{19}=\frac{\varepsilon+n-1\sum_{j=1}^{n-1}\delta_j(s)}{\binom{N}{n}(n+\varepsilon)-\sum_{s\in S}\sum_{j=1}^{n-1}\delta_j(s)} \tag{1.91}$$

W szczególności, gdy n=2, to

$$P_{19}(s)=\begin{cases}\dfrac{2(\varepsilon+1)}{(N-1)(N\varepsilon+N-2)}, & \text{gdy } i_i-i_2=1\\ \dfrac{2\varepsilon}{(N-1)(N\varepsilon+N-2)}, & \text{gdy } i_1-i_2>1\end{cases} \tag{1.92}$$

Prawdopodobieństwa inkluzji rzędu pierwszego są następujące:

$$\pi_1=\pi_N=\frac{2\varepsilon(N-1)+2(N-2)}{(N-1)(N\varepsilon+N-2)} \tag{1.93}$$

$$\pi_k=\frac{2\varepsilon(N-1)+2(N-3)}{(N-1)(N\varepsilon+N-2)} \tag{1.94}$$

1.4.12. Plany losowania zależne od sąsiedztwa elementów populacji przestrzennej i obserwacji cechy dodatkowej

Podobnie jak w dwóch ostatnich punktach zakładamy, że ustalone jest sąsiedztwo elementów populacji. Ponadto zakładamy, że dysponujemy obserwacjami cechy dodatkowej przypisanymi poszczególnym elementom populacji przestrzennej. W szczególności, gdy elementami populacji są grunty, to cechą dodatkową może być poziom sprzedaży nawozów sztucznych w poszczególnych gminach lub pwierzchnia gruntów ornych przeznaczonych pod uprawę zbóż itp. Wówczas można spodziewać się, że niektóre spośród analizowanych cech będą wykazywały autokorelację przestrzenną. Rozumiemy ją w tym sensie, że wartości cechy dla sąsiadujących ze sobą elementów populacji będą dodatnio (ujemnie) skorelowane.

Rozważmy następujący model nadpopulacji. Załóżmy, że i-temu (i = 1,...,μ) elementowi populacji jest przypisana zmienna losowa X_i oraz że

$$\bigwedge_{i=1,...N} \mathcal{E}(X_i) = \mu\,; \qquad \mathcal{D}^2(X_i) = \sigma^2 \tag{1.95}$$

Niech **A** = [a_{ij}] będzie macierzą sąsiedztwa określoną w punkcie 1.4.10. Dodatkowo załóżmy, że

$$\bigwedge_{i\neq j=1,...,N} \mathcal{C}ov(X_i, X_j) = \begin{cases} \rho & gdy \quad a_{ij} = 1 \\ 0 & gdy \quad a_{ij} = 0 \end{cases} \tag{1.96}$$

Niech dla każdego i = 1,...N:

$$Z_i = X_i - \overline{X} \tag{1.97}$$

gdzie:
$$\overline{X} = \frac{1}{N}\sum_{j=1}^{N} X_j$$

Załóżmy, że w przeprowadzanym badaniu statystycznym są obserwowane wartości wszystkich zmiennych losowych X_i (i=1,...,N). Wtedy dysponujemy ciągami wartości: $\{x_i\}=\{x_i,...,x_n\}$ i $\{z_i\} = \{z_i,...z_n\}$.

Pierwszy z analizowanych tutaj planów losowania określamy następująco:

$$P_{20}(s)\ \alpha\ b_1(s) = c + \sum_{k,h\in s} (x_k - x_h)^2 a_{kh} \tag{1.98}$$

Jeśli próba ma być losowana bezzwrotnie, to:

$$P_{20}(s) = \frac{c + \sum_{k,h\in s} (x_k - x_h)^2 a_{kh}}{c\binom{N}{n} + \sum_{s\in \mathcal{S}} \sum_{k,h\in s} (x_k - x_h)^2 a_{kh}} \tag{1.99}$$

przy czym c>0 jest stałą wprowadzoną po to, aby zapewnić, iż P(s)>0 dla każdej $s \in \mathcal{S}$.

Drugi plan ma postać

$$P_{21}(s)\alpha b_2(s) = c + \sum_{k,h \in s} (z_k + z_h)^2 a_{kh} \tag{1.100}$$

Wtedy przy bezzwrotnym losowaniu mamy:

$$P_{21}(s) = \frac{c + \sum_{k,h \in s} (z_k + z_h)^2 a_{kh}}{\binom{N}{n} c + \sum_{s \in \mathcal{S}} \sum_{k,h \in s} (z_k + z_h)^2 a_{kh}} \tag{1.101}$$

Zakładając zwrotne losowanie próby wystarczy określić prawdopodobieństwa wylosowania poszczególnych elementów z populacji w pojedynczym ciągnieniu. Wtedy można określić je następująco:

$$p_{22}(k) = \frac{c + \sum_{h=1}^{N} (x_k - x_h)^2 a_{kh}}{Nc + \sum_{k=1}^{N} \sum_{h=1}^{N} (x_k - x_h)^2 a_{kh}} \tag{1.102}$$

$$p_{23}(k) = \frac{c + \sum_{h=1}^{N} (z_k - z_h)^2 a_{kh}}{Nc + \sum_{k=1}^{N} \sum_{h=1}^{N} (z_k - z_h)^2 a_{kh}} \tag{1.103}$$

przy czym c > 0.

Zapisane wzory określające plany losowania można przedstawić w prostszej postaci wtedy, gdy zostanie ustalona konkretna postać macierzy sąsiedztwa.

Po to, aby ocenić wpływ przestrzennej autokorelacji na proponowane plany losowania, wyznaczamy ich oczekiwane wartości, przy określonych wzorami (1.95) i (1.96) założeniach modelu nadpopulacji

$$\mathcal{E}(B_1(s))\ c + n_{kh}\sigma^2 (1- \rho) \tag{1.104}$$

$$\mathcal{E}(B_2(s) = c + n_{kh}\sigma^2 (1 + \rho) \tag{1.105}$$

przy czym przez n_{kh} oznaczono ilość sąsiadujących ze sobą w próbie elementów. Stąd wynika, że w przypadku występowania ujemnej autokorelacji plany P_{20} będą preferowały wejście do próby sąsiadujących ze sobą elementów. Z kolei plany P_{21} i P_{23} będą sprzyjały wejściu do próby sąsiadujących ze sobą elementów, gdy współczynnik atokorelacji będzie dodatni.

1.5. Dane o parametrze populacji i statystyki

Wcześniej wprowadzono pojęcie parametru populacji jako macierzy rzeczywistej y o wymiarach N x m. Każdy wiersz tej macierzy jest jednoznacznie przyporządkowany jednemu elementowi populacji. Z kolei elementy każdej kolumny tej macierzy są traktowane jako wartości (realizacje) jednowymiarowej zmiennej (cechy).

W wyniku obserwacji elementów populacji o numerach $k \in \underline{s}$ otrzymujemy dane o realizacjach m wymiarowej cechy, które zapisujemy wektorami wierszowymi: $\mathbf{y}_{k*} = [y_{k1}...y_{km}]$. Oznaczmy przez $\mathbf{y}_{\underline{s}}$ macierz o wymiarach n x m składającą się z wierszy: $\mathbf{y}_{k_i*} = [y_{k_i 1} \cdots y_{k_i m}]$ i=1,...,n obserwowanych na kolejnych elementach tworzących próbę uporządkowaną $\underline{s}=\{k_1,...,k_n\}$, a zatem

$$\mathbf{y}_{\underline{s}} = \begin{bmatrix} \mathbf{y}_{k_1*} \\ .... \\ \mathbf{y}_{k_m*} \end{bmatrix}$$

Oznaczenia te wykorzystujemy do bezpośredniego uogólnienia wprowadzonego przez Cassela i in. (1977) pojęcia danych o wektorowym parametrze populacji (o jednowymiarowej zmiennej) na przypadek macierzowego parametru lub, innymi słowy, danych o wielowymiarowej zmiennej:

Definicja 1.19: Daną identyfikowalną o parametrze **y** populacji (o m wymiarowej zmiennej) obserwowaną w próbie uporządkowanej $\underline{s}$ nazywamy macierz $\underline{\mathbf{d}} = [\underline{s} : \mathbf{y}_{\underline{s}}]$.

Przypomnijmy, że przez s oznaczono próbę nieuporządkowaną, którą uzyskuje się z próby $\underline{s}$ poprzez pominięcie w niej powtarzających się elementów i zaniedbanie porządku, w jakim one były losowanie. Basu (1971) nazywa próbę s jądrem próby $\underline{s}$. Dane $\mathbf{d} = [s : \mathbf{y}_s]$ stanowią więc jądro danych $\underline{\mathbf{d}}$.

Dane można traktować jako realizacje pewnej macierzy losowej $\underline{\mathbf{D}}$ bądź **D**.

Definicja 1.20 [Cassel i in. (1977), s. 20]: Przestrzenią prób zmiennej losowej $\underline{\mathbf{D}}$ przyjmującej wartości $\underline{\mathbf{d}}$ jest zbiór:

$$\underline{\mathcal{D}} = \{\underline{d}: \underline{s} \in \underline{\mathcal{S}}(\Omega),\ \ y \in \mathcal{Y}\} \tag{1.106}$$

gdzie $\mathcal{Y} \in R^{nm}$ jest przestrzenią parametru populacji. Przestrzenią prób **Y** macierzy losowej **D** jest zbiór:

$$\mathcal{D} = \{\mathbf{d}: s \in \mathcal{S}(\Omega),\ \ \mathbf{y} \in \mathcal{Y}\} \tag{1.107}$$

Zbiory $\underline{\mathcal{D}}$ i $\mathcal{D}$ są więc przestrzeniami prób danych $\underline{\mathbf{d}}$ i **d**.

Macierz losową $\underline{\mathbf{D}}$ można zapisać w sposób blokowy jako $\underline{\mathbf{D}} = [\underline{S} : \mathbf{y}_{\underline{s}}]$, gdzie:

$$\underline{S} = \{K_1, \ldots, K_n\}, \quad \mathbf{y}_{\underline{S}} = \begin{bmatrix} \mathbf{y}_{K_1 *} \\ \ldots. \\ \mathbf{y}_{K_n *} \end{bmatrix}$$

a jej wartość $\underline{\mathbf{d}} = [\underline{s} : \mathbf{y}_{\underline{s}}]$, gdzie $\underline{s} = \{k_1,...,k_n\}$ jest realizacją próby losowej $\underline{S} = \{K_1,...,K_n\}$.

Statystyk często dysponuje obserwacjami cech, lecz bez informacji o numerach elementów populacji, którym są one przypisane. Zatem zaszła potrzeba wprowadzenia pojęcia danej nieidentyfikowalnej. Dana nieidentyfikowalna i nieuporządkowana $\mathbf{y}_s$ (uporządkowana $\mathbf{y}_{\underline{s}}$) jest realizacją zmiennej losowej $\mathbf{y}_S$ ($\mathbf{y}_{\underline{S}}$) otrzymanej z macierzy losowej **D** (**D**) po pominięciu w niej wielkości S ($\underline{S}$). Zatem zmienna losowa $\mathbf{y}_S$ ($\mathbf{y}_{\underline{S}}$) niesie mniej informacji ze sobą niż **D** (**D**), ponieważ nie są znane numery elementów populacji, którym są przyporządkowane poszczególne wartości zmiennych losowych $\mathbf{y}_S$ ($\mathbf{y}_{\underline{S}}$).

Cassel i in. (1977), s. 20 wprowadzili pojęcie statystyki, jako funkcji danych o zmiennej jednowymiarowej, które jest także ważne dla przypadku cechy wielowymiarowej.

Definicja 1.21: Statystyką jest nazywana każda funkcja rzeczywista $Z = u(\underline{\mathbf{D}})$ określona na przestrzeni $\mathcal{D}$ taka, że dla danej $\underline{s} \in \mathcal{S}$ wartość $u(\underline{\mathbf{d}})$ zależy od parametru **y** tylko poprzez te obserwacje $\mathbf{y}_{k*}$, dla których $k \in \underline{s}$.

Szczególnymi przypadkami statystyk są m.in. estymatory parametrów cech populacji, które analizowane są w dalszej części pracy.

2. PODSTAWY TEORII ESTYMACJI

2.1. Estymacja parametrów populacji ustalonej

2.1.1. Podstawowe własności strategii próbkowania

Celem estymacji jest określony definicją 1.5. wektor $\boldsymbol{\theta}=[\theta_1...\theta_m]$ funkcji parametrycznych (parametrów opisowych populacji) zależnych od parametru populacji **y**.

Estymację parametrów $\boldsymbol{\theta}$ prowadzi się na podstawie realizacji **d** macierzy losowej **D** danych o cechach populacji, którą określa definicja 1.19. Zakładamy, że obserwacje cech elementów populacji wchodzących do próby są bezbłędne. Rozkład prawdopodobieństwa macierzy **D** zależy od parametru populacji **y** oraz od planu losowania próby ustalonego mniej lub bardziej subiektywnie przez statystyka. Niech $\boldsymbol{\Theta}\in R^m$ będzie zbiorem możliwych wartości wektora parametrów opisowych $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{y})$, a zatem $\boldsymbol{\theta}$: $\mathcal{Y}\rightarrow\boldsymbol{\Theta}$. Bartoszewicz (1989, s. 145) wprowadza precyzyjną definicję estymatora parametru $\boldsymbol{\theta}$ na gruncie klasycznej teorii statystyki. Korzystając z jej uproszczenia używanego przez Zielińskiego (1990, s. 10) formułujemy ją następująco.

Definicja 2.1: Wektor statystyk $t_D=[t_{D1}...t_{Dm}]$ przyjmujący wartości ze zbioru $\boldsymbol{\Theta}$ nazywamy estymatorem wektora parametrów opisowych $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Podobnie określamy estymator $\mathbf{t}_{\underline{D}}$ jako funkcję danych $\underline{\mathbf{D}}$ z próby uporządkowanej $\underline{S}$. W niektórych pracach z dziedziny metody reprezentacyjnej używa się zamiast oznaczeń estymatorów $\mathbf{t}_D$, $\mathbf{t}_{\underline{D}}$ symboli odpowiednio $\mathbf{t}_S$, $\mathbf{t}_{\underline{S}}$. Te ostatnie również będziemy stosować w niniejszej pracy. Podkreślmy, że estymator $\mathbf{t}_S$ jest zmienną losową zależną od wyników obserwacji cech w próbie losowej S, natomiast $\mathbf{t}_s$ jest wartością statystyki $\mathbf{t}_S$ obliczoną na podstawie obserwacji pochodzących z konkretnej realizacji s próby S.

W celu otrzymania oceny parametrów $\boldsymbol{\theta}$ należy wybrać odpowiedni estymator i plan losowania próby. Te dwa czynniki wpływają na dokładność oceny wnioskowania. Zatem jest

wygodnie podczas porównywania dokładności różnych sposobów estymacji posługiwać się pojęciem strategii (próbkowania) wprowadzonej przez Cassela i in. (1977).

Definicja 2.2: Ustaloną parę uporządkowaną $\{\mathbf{t}_S, P(s)\}$ lub $\left\{\mathbf{t}_{\underline{S}}, P(\underline{s})\right\}$ nazywiemy strategią estymacji parametrów opisowych populacji $\boldsymbol{\theta}$.

W przypadku gdy szacowany parametr $\boldsymbol{\theta}$ jest skalarem, to błąd estymacji jest definiowany [18] jako różnica między estymatorem i parametrem θ. Ta definicja jest przenoszona bezpośrednio na przypadek wektorowej estymacji. Zatem błąd estymacji określa wzór:

$$B = B(t_s, \theta) = t_S - \theta \tag{2.1}$$

Definicja [19] 2.3: Estymator $\mathbf{t}_S$ daje nieobciążone oceny wektora θ wtedy i tylko wtedy, gdy:

$$\bigwedge_{y \in \mathcal{Y}} E(t_S) = \sum_{s \in \mathcal{S}} t_s P(s) = \boldsymbol{\theta} \tag{2.2}$$

gdzie przez $\mathcal{Y}$ i $\mathcal{S}$ oznaczono odpowiednio przestrzeń parametru populacji i przestrzeń prób.

Zaznaczmy, że estymator nieobciążony względem wybranego planu losowania próby może być jednak obciążony względem innego planu.

Definicja 2.4: Estymator $\mathbf{t}_S$ daje ε nieobciążone oceny wektorów $\boldsymbol{\theta}(\mathbf{y})$, jeśli dla każdego $\mathbf{y} \in \mathcal{Y}$ i dla ustalonego błędu dopuszczalnego estymacji $\varepsilon > 0$ zachodzi, że

$$\bigvee_{n_0 \le N} \bigwedge_{i=1,\ldots,m} \left|E(t_{iS}) - \theta_i\right| \le \varepsilon \tag{2.3}$$

Liczba ε (którą można ustalać na różnym poziomie dla każdej składowej wektora $\boldsymbol{\theta}$ jest interpretowana jako dopuszczalny poziom obciążenia wektora estymacji i ε może zależeć od liczebności próby. Zwykle ε jest określane za pomocą wielkości $0(n^{-r})$ $(r > 0)$, czyli wraz ze wzrostem liczebności próby n obciążenie ε maleje tak jak ciąg $\{n^{-r}\}$. Gdy obciążenie estymatora jest rzędu mniejszego niż $0(n^{-1})$, to taki estymator jest nazywany prawie nieobciążonym[20].

Przegląd innych definicji nieobciążoności estymatorów i definicji ich zgodności można znaleźć np. w pracy Wywiała (1992).

Niech macierz **A** ma wymiary m × m, natomiast **b** jest wektorem wierszowym o wymiarze 1 × m, natomiast $\mathbf{J}_N$ jest kolumną jedynkową o wymiarze N × 1. Przez **D** i **D'** oznaczono dane o populacjach charakteryzowanych odpowiednio parametrami **y** i **y'** , przy czym $\mathbf{y'}=\mathbf{yA}+\mathbf{J_N b}$.

Definicja 2.5: Estymator $\mathbf{t}_D$ nazywamy niezmienniczym liniowo wtedy i tylko wtedy, gdy dla każdej transformacji $\mathbf{y}` = \mathbf{yA} + \mathbf{J_n b}$ i każdej danej **D** zachodzi:

$$\mathbf{t_{D'}} = \mathbf{b} + \mathbf{t_D A} \tag{2.4}$$

---

[18] Por. np. Pawłowski (1976).

[19] Por. definicję nieobciążoności Basu (1971), s. 208.

[20] Por. H.P. Singh i in. (1985).

Gdy $\mathbf{A} = \mathbf{I}_m$ oraz $\mathbf{b} \neq \mathbf{o}_m$, gdzie $\mathbf{o}_m$ jest m elementowym wektorem zerowym o wymiarze 1 x m, to $\mathbf{t}_D$ jest nazywany estymatorem niezmienniczym względem przesunięcia parametru populacji. W przypadku gdy **A** jest macierzą diagonalną o dodatnich elementach jej głównej przekątnej i $\mathbf{b} = \mathbf{o}_m$, to $\mathbf{t}_D$ nazywamy niezmienniczym względem zamiany skali. W końcu gdy $\mathbf{b} \neq \mathbf{o}_m$ i **A** jest diagonalną o dodatnich elementach, to $\mathbf{t}_D$ nazywamy niezmienniczym względem przesunięcia i zamiany skali.

2.1.2. Mierniki rzędu dokładności estymacji

W przypadku jednowymiarowym do oceny dokładności estymacji wybranego parametru zwykle używa się tzw. błędu średniokwadratowego estymacji tego parametru[21], który jest równy drugiemu momentowi zwykłemu błędu estymacji. Sposób ten jest przenoszony na przypadek estymacji wektorowej. Niech $\mathbf{V}_{SR}(\mathbf{t}_D) = \mathbf{V}_{SR}(\mathbf{t}_S) = E(\mathbf{B}^T\mathbf{B})$, gdzie wektor błędów estymacji **B** o wymiarze $1 \times z$ określa wzór (2.1). Macierz $\mathbf{V}_{SR}(\mathbf{t}_S)$ drugich momentów mieszanych błędów estymacji nazywamy macierzą błędów średniokwadratowych oceny wektora parametrów $\boldsymbol{\theta}$ na podstawie estymatora $\mathbf{t}_S$. W szczególności jej element o numerze (i,k) (i,k=1,...,m) wyliczamy według wzoru:

$$E(B_iB_k) = \sum_{s \in \mathcal{S}} (t_{is} - \theta_i)(t_{ks} - \theta_k)P(s) \qquad (2.5)$$

W przypadku jednowymiarowym do oceny precyzji estymacji używa się wariancji estymatora [22]. W przypadku wielowymiarowym precyzję estymatora wektorowego $\mathbf{t}_S$ charakteryzujemy przy pomocy jego macierzy wariancji i kowariancji $\mathbf{V}(\mathbf{t}_S) = E(\mathbf{t}_S\text{-}E(\mathbf{t}_S))^T(\mathbf{t}_S\text{-}E(\mathbf{t}_S))$. Wówczas łatwo sprawdzamy, że

$$\mathbf{V}_{SR}(\mathbf{t}_S) = \mathbf{V}(\mathbf{t}_S) + E^T(\mathbf{B})E(\mathbf{B}) \qquad (2.6)$$

Stąd wnioskujemy, że jeśli estymator $\mathbf{t}_S$ jest nieobciążony dla $\boldsymbol{\theta}$, to $\mathbf{V}_{SR}(\mathbf{t}_S) = \mathbf{V}(\mathbf{t}_S)$.

Zwykle rząd dokładności estymacji parametrów $\boldsymbol{\theta}$ prowadzony na podstawie statystyki $\mathbf{t}_S$ jest charakteryzowany za pomocą błędów średniokwadratowych poszczególnych składowych wektora $\mathbf{t}_S$, czyli $[E(B_1)^2...E(B_z)^2]$, który redukuje się do wektora wariancji $[V(t_{1S})...V(t_{zS})]$, jeśli estymator $\mathbf{t}_S$ daje oceny nieobciążone parametrów $\boldsymbol{\theta}$.

Drugi sposób oceny dokładności polega na obliczeniu śladu macierzy błędów średniokwadratowych [23], który oznaczamy przez $q^2_{SR}(\mathbf{t}_S) = \mathrm{tr}\mathbf{V}_{SR}(\mathbf{t}_S)$ i nazywamy średniokwadratowym promieniem estymacji wektora $\boldsymbol{\theta}$ za pomocą estymatora $\mathbf{t}_S$. Parametr $q^2_{SR}(\mathbf{t}_S)$ wskazuje poziom średniego kwadratu odległości punktu, którego współrzędnymi są realizacje wektora $\mathbf{t}_S$, od punktu, którego współrzędnymi są elementy wektora parametrów $\boldsymbol{\theta}$.

---

[21] Ang. mean sguare error. Por. np. Cochran (1963), Kish (1965), Jessen (1978). Kordos (1987; 1988) używa do oceny dokładności estymacji pierwiastka z błędu średniokwadratowego estymacji, który Kisch (1965) nazywa błędem ogólnym (total error).

[22] Tamże.

[23] Por. np. Rao (1982).

Parametr $q(\mathbf{t}_S)=\sqrt{\mathrm{tr}\mathbf{V}(\mathbf{t}_S)}$ nazwiemy średnim promieniem estymatora wektorowego $\mathbf{t}_S$, i interpretujemy jako średnią odległość punktu, którego współrzędnymi są składowe wektora $E(\mathbf{t}_S)$, od punktu, którego współrzędnymi są elementy wektora $\mathbf{t}_S$. Parametr $q(\mathbf{t}_S)$ służy do oceny precyzji estymacji wektorowej.

Jeśli $E(\mathbf{t}_S) = \boldsymbol{\theta}$, to $q_{SR}(\mathbf{t}_S) = q(\mathbf{t}_S)$.

Kolejnymi parametrami służącymi do oceny rzędu dokładności i precyzji estymacji wektora $\boldsymbol{\theta}$ są odpowiednio wyznaczniki: $g_{SR}(\mathbf{t}_S) = \det\mathbf{V}_{SR}(\mathbf{t}_S)$ i $g(\mathbf{t}_S) = \det\mathbf{V}(\mathbf{t}_S)$, z których pierwszy nazwiemy uogólnionym błędem średniokwadratowym estymacji parametrów $\boldsymbol{\theta}$ na podstawie wektora $\mathbf{t}_S$, natomiast drugi Wilks (1932) nazwał[24] uogólnioną wariancją estymatora $\mathbf{t}_S$.

Korzystając z ogólnych wyników zamieszczonych w pracy Wilksa (1962), s. 546 wnioskujemy, że:

$$g_{SR}(\mathbf{t}_S) = g(\mathbf{t}_S)\,(1 + E(\mathbf{B})\mathbf{V}^{-1}(\mathbf{t}_S)E^T(\mathbf{B})) \geq g(\mathbf{t}_S) \tag{2.7}$$

Stąd wynika, że jeśli $\mathbf{t}_S$ daje nieobciążone oceny parametrów $\boldsymbol{\theta}$, czyli $E(\mathbf{B})=\mathbf{0}$, to $g_{SR}(\mathbf{t}_S)=g(\mathbf{t}_S)$.

Przypomnijmy [por. np. Ralston (1975)], że promień spektralny danej macierzy jest równy jej maksymalnej wartości własnej. Niech $\rho_{SR}(\mathbf{t}_S)$ i $\rho(\mathbf{t}_S)$ będą promieniami spektralnymi poprzednio zdefiniowanych odpowiednio macierzy błędów średniokwadratowych $\mathbf{V}_{SR}(\mathbf{t}_S)$ i macierzy wariancji i kowariancji $\mathbf{V}(\mathbf{t}_S)$. Niech $t_{oS} = \mathbf{t}_S\mathbf{w}^T$, gdzie $\mathbf{w} = [w_1...w_m]$ i $\mathbf{w}\mathbf{w}^T= 1$ oraz $\mathbf{w}\in R^z - \{\mathbf{o}_m\}$.

Statystyka t_{oS} estymuje kombinację liniową: $\theta_o = \boldsymbol{\theta}\mathbf{w}^T$ z błędem średniokwadratowym: $V_{SR}(t_{oS}) = E(t_{oS} - \theta_o)^2 = \mathbf{w}\mathbf{V}_{SR}(\mathbf{t}_S)\mathbf{w}^T$. Podobnie $D^2(t_{oS}) = \mathbf{w}\mathbf{V}(\mathbf{t}_S)\mathbf{w}^T$. Na podstawie znanych własności wartości własnych macierzy [por. np. C.R.Rao (1982)] wnioskujemy, że:

$$\rho_{SR}(\mathbf{t}_S) = \underset{\mathbf{w}\mathbf{w}^T = 1}{\mathrm{maximum}}\{\mathbf{w}\mathbf{V}_{SR}(\mathbf{t}_S)\mathbf{w}^T\}$$

$$\rho(\mathbf{t}_S) = \underset{\mathbf{w}\mathbf{w}^T = 1}{\mathrm{maximum}}\{\mathbf{w}\mathbf{V}(t_S)\mathbf{w}^T\} \tag{2.8}$$

Stąd wynika[25], że $\rho_{SR}(\mathbf{t}_S)$ jest błędem średniokwadratowym estymacji kombinacji liniowej $\theta_o=\boldsymbol{\theta}\mathbf{w}^T$ za pomocą statystyki $t_{oS} = \mathbf{t}_S\mathbf{w}^T$ przy najmniej korzystnym układzie wektora współczynników $\mathbf{w}_*$ w tym sensie, że $\rho_{SR}(\mathbf{t}_S)=E(\mathbf{t}_S w_*^T - \boldsymbol{\theta}w_*^T)^2 \geq E(\mathbf{t}_S\mathbf{w}^T-\boldsymbol{\theta}\mathbf{w}^T)^2$, dla których $\mathbf{w}\mathbf{w}^T= 1$, $\mathbf{w}_*\mathbf{w}_*^T = 1$ i $\mathbf{w},\mathbf{w}_*\in R^m-\{\mathbf{o}_m\}$. Podobnie interpretujemy parametr $\rho(\mathbf{t}_S)$, ponieważ $\rho(\mathbf{t}_S)=\rho_{SR}(\mathbf{t}_S)$, gdy $\mathbf{t}_S$ jest nieobciążonym estymatorem parametrem $\boldsymbol{\theta}$.

Parametr $\rho_{SR}(\mathbf{t}_S)$ nazywamy średniokwadratowym promieniem spektralnym estymacji parametrów $\boldsymbol{\theta}$ na podstawie wektora $\mathbf{t}_S$. Z kolei $\rho(\mathbf{t}_S)$ nazywamy promieniem spektralnym estymatora $\mathbf{t}_S$. Parametry $\rho_{SR}(\mathbf{t}_S)$ i $\rho(\mathbf{t}_S)$ służą do oceny odpowiednio dokładności i precyzji wektora parametrów $\boldsymbol{\theta}$ przy pomocy estymatora $\mathbf{t}_S$.

---

[24] Wiadomość tę podajemy za Cramerem (1958).

[25] Por. Rao (1982).

2.1.3. Porównywanie strategii

Wprowadzane dalej definicje pozwalające porządkować strategie z punktu widzenia dokładności i precyzji dawanych przez nie ocen parametrów sformułujemy na podstawie sposobu mierzenia zróżnicowania wartości wielowymiarowej zmiennej prezentowanego w pracy Borowkowa (1984), s. 99-103.

Niech $\mathbf{X}_1$, $\mathbf{X}_2$ będą dwoma $1 \times m$ wymiarowymi wektorami losowymi o różnych rozkładach prawdopodobieństwa. Przez $\mathbf{a}$, $\boldsymbol{\alpha}$ oznaczamy wektory rzeczywiste o wymiarze $1\times m$, natomiast przez $\mathbf{o}_m$ wektor zerowy o wymiarze $1\times m$. Stopień zróżnicowania rozkładu prawdopodobieństwa wektora $\mathbf{X}_j$ (i=1,2) jest określony parametrem $e_i^2(\mathbf{a},\boldsymbol{\alpha}) = E(\mathbf{X}_i\mathbf{a}^T - \boldsymbol{\alpha}\mathbf{a}^T)^2$, (i=1,2). Współczynnik $e_i^2(\mathbf{a},\boldsymbol{\alpha})$ mierzy przeciętny stopień zróżnicowania rozkładu realizacji wektora $\mathbf{x}_i$ wokół punktu o współrzędnych $\boldsymbol{\alpha}$.

Definicja 2.6. [Borowkow (1984), str. 99]: Mówimy, że średniokwadratowe zróżnicowanie rozkładu zmiennej losowej $\mathbf{X}_1$ wokół ustalonego punktu $\boldsymbol{\alpha} \in \mathbf{R}^m$ jest nie większe od zróżnicowania rozkładu wektora $\mathbf{X}_2$, jeśli dla każdego $\mathbf{a} \in \mathbf{R}^m - \{\mathbf{o}_m\}$.

$$e_1^2(\mathbf{a},\boldsymbol{\alpha}) = E(\mathbf{X}_1\mathbf{a}^T - \boldsymbol{\alpha}\mathbf{a}^T)^2 \le E(\mathbf{X}_2\mathbf{a}^T - \boldsymbol{\alpha}\mathbf{a}^T)^2 = e_2^2(\mathbf{a},\boldsymbol{\alpha}) \qquad (2.9)$$

Jeśli przynajmniej dla jednego wektora $\mathbf{a}\in R^m$ zachodzi, że $e_1^2(\mathbf{a},\boldsymbol{\alpha}) < e_2^2(\mathbf{a},\boldsymbol{\alpha})$, to mówimy, że rozkład wektora $\mathbf{X}_1$ jest mniej zróżnicowany wokół punktu $\boldsymbol{\alpha}$ niż rozkład wektora $\mathbf{X}_2$.

Niech $\mathbf{C}_i(\boldsymbol{\alpha}) = E(\mathbf{X}_i - \boldsymbol{\alpha})^T(\mathbf{X}_i - \boldsymbol{\alpha})$, i = 1,2.

Lemat 2.1. [Borowkow (1984), s. 100]: Średniokwadratowe zróżnicowanie rozkładu zmiennej losowej $\mathbf{X}_1$ wokół ustalonego punktu $\boldsymbol{\alpha}\in R^m$ jest nie większe od zróżnicowania wektora $\mathbf{X}_2$ wtedy i tylko wtedy, gdy różnica macierzy $(\mathbf{C}_2(\boldsymbol{\alpha})-\mathbf{C}_1(\boldsymbol{\alpha}))$ jest nieujemnie określona.

Wprowadźmy oznaczenie: $l_i(\mathbf{A},\boldsymbol{\alpha})=E(\mathbf{X}_i-\boldsymbol{\alpha})\mathbf{A}(\mathbf{X}_i-\boldsymbol{\alpha})^T$, i=1,2, przy czym macierz $\mathbf{A}$ jest kwadratową stopnia m.

Lemat 2.2 [Borowkow (1984), s. 101]: Macierz $\mathbf{C}_2(\boldsymbol{\alpha})-\mathbf{C}_1(\boldsymbol{\alpha})$ jest nieujemnie określona wtedy i tylko wtedy, gdy dla każdej nieujemnie określonej macierzy $\mathbf{A}$ (z wyjątkiem macierzy zerowej) zachodzi nierówność:

$$l_1(\mathbf{A},\boldsymbol{\alpha}) \le l_2(\mathbf{A},\boldsymbol{\alpha}) \qquad (2.10)$$

Cramer (1958) proponuje[26] porównywać rozrzut wartości wielowymiarowych zmiennych za pomocą elipsoidy koncentracji rozkładu wielowymiarowego. Załóżmy, że każdy wektor losowy $\mathbf{X}_i$ (i=1,2) o wymiarach $1 \times m$ ma nieosobliwy rozkład z wektorem wartości oczekiwanych $E(\mathbf{X}_i) = \boldsymbol{\mu}_i$ oraz macierzą wariancji i kowariancji $\mathbf{V}_i$.

Definicja 2.7: Elipsoidę koncentracji (rozrzutu) rozkładu wektora losowego $\mathbf{X}_i$ określa wyrażenie:

$$\mathcal{E}(\mathbf{X}_i) = \{\mathbf{a}: (a - \boldsymbol{\mu}_i)\ \mathbf{V}_i^{-1}(\mathbf{a} - \boldsymbol{\mu})^T \le z + 2,\ \mathbf{a} \in R^z\} \qquad (2.11)$$

gdzie $\mathbf{V}_i$ jest macierzą wariancji i kowariancji wektora $\mathbf{X}_i$.

---

[26] Por. Borowkow (1984), s. 101-102.

Elipsoida $\mathcal{E}(\mathbf{X}_i)$ służy do geometrycznej prezentacji rozrzutu (zróżnicowania) realizacji wektora $\mathbf{X}_i$ oraz ma następującą własność: Jeśli wektor losowy $\mathbf{U}_i$ ma jednostajny rozkład na elipsoidzie $\mathcal{E}(\mathbf{X}_i)$, to wektory $\mathbf{U}_i$, $\mathbf{X}_i$ mają wspólne wektor wartości oczekiwanych i macierz wariancji i kowariancji.

Lemat 2.3 [Borowkow (1984), s. 102]: Jeśli wektory losowe mają rozkłady nieosobliwe, to średniokwadratowe zróżnicowanie rozkładu wektora $\mathbf{X}_1$ wokół $E(\mathbf{X}_1)$ nie jest większe od zróżnicowania $\mathbf{X}_2$ wokół $E(\mathbf{X}_2)$ wtedy i tylko wtedy, gdy elipsoida koncentracji $\mathcal{E}(\mathbf{X}_1)$ zawiera się w elipsoidzie $\mathcal{E}(\mathbf{X}_2)$.

Strategię estymacji należy tak wybrać, aby dawała najlepsze oceny wektora parametrów $\boldsymbol{\theta}$. Zagadnienie to w przypadku szacowania skalarnego parametru opisowego populacji jest sprowadzone do wyboru strategii, dla której błąd średniokwadratowy oceny parametru jest najmniejszy. W przypadku estymacji wektora parametrów $\boldsymbol{\theta}$ kryterium wyboru strategii nie jest już tak oczywiste. W tym przypadku można żądać, by najlepsza strategia dawała np. możliwie mały uogólniony błąd średniokwadratowy bądź mały średniokwadratowy promień oceny parametrów. Okazuje się, że m.in. wymienione kryteria są szczególne w stosunku do kryterium pozwalającego porządkować wektorowe strategie estymacji $(\mathbf{t}_S, P(s))$ w pewnym zbiorze ϕ poprzez porównanie dokładności estymacji liniowej kombinacji $\boldsymbol{\theta}\mathbf{a}^T$ za pomocą $\mathbf{t}_S\mathbf{a}^T$, gdzie $\mathbf{a}=[a_1...a_m] \in R^z - \{\mathbf{o}_m\}$.

Niech $(\mathbf{t}_S^{(1)}, P^{(1)}(s)), (\mathbf{t}_S^{(2)}, P^{(2)}(s))$ będą dowolnymi strategiami estymacji wektora parametrów $\boldsymbol{\theta} = [\theta_1...\theta_m]$.

Definicja[27] 2.8: Mówimy, że strategia $(\mathbf{t}_S^{(1)}, P^{(1)}(s))$ wektora $\boldsymbol{\theta}(\mathbf{y})$ jest nie gorsza (lepsza) od strategii $(\mathbf{t}_S^{(2)}, P^{(2)}(s))$ wtedy i tylko wtedy, gdy zróżnicowanie średniokwadratowe w sensie definicji 2.6 rozkładu wektora losowego $\mathbf{t}_S^{(1)}$ wokół wektora $\boldsymbol{\theta}(\mathbf{y})$ nie jest większe (mniejsze) od zróżnicowania średniokwadratowego rozkładu wektora $\mathbf{t}_S^{(2)}$ dla każdego $\mathbf{y} \in \mathcal{Y}$.

Stąd i na podstawie lematu 2.1 natychmiast wynika własność:

Twierdzenie 2.1: Strategia $(\mathbf{t}_S^{(1)}, P^{(1)}(s))$ parametru $\boldsymbol{\theta}(\mathbf{y})$ jest nie gorsza (nie lepsza) od $(\mathbf{t}_S^{(2)}, P^{(2)}(s))$, wtedy i tylko wtedy, gdy dla każdego $\mathbf{y} \in \mathcal{Y}$ różnica macierzy: $\mathbf{R}=\mathbf{V}_{SR}(\mathbf{t}_S^{(2)}, P^{(2)}(s))-\mathbf{V}_{SR}(\mathbf{t}_S^{(1)}, P^{(1)}(s))$ jest nieujemnie (niedodatnio) określona. Gdy przynajmniej dla jednego wektora $\mathbf{y} \in \mathcal{Y}$ macierz $\mathbf{R}$ jest dodatnio (ujemnie) określona, to strategia $(\mathbf{t}_S^{(1)}, P^{(1)}(s))$ jest lepsza (gorsza) od $(\mathbf{t}_S^{(2)}, P^{(2)}(s))$.

Szczególną w stosunku do definicji 2.8 jest następująca:

Definicja 2.9: Dla ustalonego planu losowania próby statystyka $\mathbf{t}_S^{(1)}$ jest nie gorszym (nie lepszym) estymatorem parametrów $\boldsymbol{\theta}(\mathbf{y})$ od statystyki $\mathbf{t}_S^{(2)}$ wtedy i tylko wtedy, gdy zróż-

---

[27] Definicja ta jest adaptowaną, do potrzeb estymacji parametrów opisowych populacji ustalonej, formą odpowiedniego określenia Borowkowa (1984). Porównaj również przypadek jednowymiarowy rozważany przez Cassela i in. (1977), s. 39.

nicowanie rozkładu wektora losowego $\mathbf{t}_S^{(1)}$ wokół punktu $\boldsymbol{\theta}(\mathbf{y})$ jest nie większe (mniejsze) od wektora $\mathbf{t}_S^{(2)}$ dla każdego $\mathbf{y}\in\mathcal{Y}$.

Wprowadzony sposób porównywania strategii staje się bardzo zasadny w niektórych sytuacjach praktycznych. Przypuśćmy, że pewne decyzje są podejmowane na podstawie oceny funkcji $\boldsymbol{\theta}\mathbf{a}^T$, przy czym z góry nie jest znany wektor współczynników $\mathbf{a}$.Wówczas określone definicją 2.9 i twierdzeniem 2.1. kryterium pozwala wybrać spośród dwóch estymatorów wektorowych ten, który będzie szacował parametr $\boldsymbol{\theta}\mathbf{a}^T$ z mniejszym błędem, przy nie znanym z góry wektorze $\mathbf{a}$.

Przypuśćmy, że strategie $(\mathbf{t}_S^{(i)},P^{(i)}(s))$ (i=1,2) są nieobciążone oraz nieosobliwe, czyli $\det \mathbf{V}\left(\mathbf{t}_S^{(i)}\right)>0$. Wówczas na podstawie definicji 2.7, 2.8 i lematu 2.3 wnioskujemy, że jeśli strategia $(\mathbf{t}_S^{(1)},P^{(1)}(s))$ jest nie gorsza od $(\mathbf{t}_S^{(2)},P^{(2)}(s))$, to elipsoida koncentracji $\mathcal{K}(\mathbf{t}_S^{(1)})$ zawiera się w elipsoidzie $\mathcal{K}(\mathbf{t}_S^{(2)})$.

Związek między określonym definicją 2.8 sposobem porządkowania strategii a własnościami miar syntetycznych dokładności estymacji wektorowej określa twierdzenie:

Twierdzenie 2.2 [Rao (1982)]: Jeśli nieosobliwa strategia estymacji $(\mathbf{t}_S^{(1)},P^{(1)}(s))$ parametrów $\boldsymbol{\theta}(\mathbf{y})$ jest nie gorsza od strategii $(\mathbf{t}_S^{(2)},P^{(2)}(s))$ w sensie definicji 2.8, to:

$$q_{SR}\left(\mathbf{t}_S^{(1)},P^{(1)}(s)\right)\le q_{SR}\left(\mathbf{t}_S^{(2)},P^{(2)}(s)\right) \tag{2.12}$$

$$g_{SR}\left(\mathbf{t}_S^{(1)},P^{(1)}(s)\right)\le g_{SR}\left(\mathbf{t}_S^{(2)},P^{(2)}(s)\right) \tag{2.13}$$

$$\rho_{SR}\left(\mathbf{t}_S^{(1)},P^{(1)}(s)\right)\le \rho_{SR}\left(\mathbf{t}_S^{(2)},P^{(2)}(s)\right) \tag{2.14}$$

$$\bigwedge_{i=1,\ldots,m} v_{SR}\left(t_{Si}^{(1)},P^{(1)}(s)\right)\le v_{SR}\left(t_{Si}^{(2)},P^{(2)}(s)\right) \tag{2.15}$$

Oznaczmy przez $\mathcal{A}_t$ zbiór estymatorów $\mathbf{t}_S$ parametru $\boldsymbol{\theta}(\mathbf{y})$, gdzie $\mathbf{y}\in\mathcal{Y}$. Klasa $\mathcal{A}_t$ charakteryzuje się zwykle pewną własnością, jak np. nieobciążoność czy liniowość.

Definicja 2.10: Strategia $(\mathbf{t}_S^*,P^*)\in\mathcal{A}_y$ wektora parametrów $\boldsymbol{\theta}(\mathbf{y})$ jest nazywana strategią dopuszczalną w klasie $\mathcal{A}_y$ wtedy i tylko wtedy, gdy w klasie $\mathcal{A}_y$ nie ma od niej lepszej strategii estymacji w sensie definicji 2.8.

Wprowadzone pojęcie jest uogólnieniem na przypadek wielowymiarowy znanej definicji dopuszczalności strategii skalarnych. Własność dopuszczalności estymatora okazuje się niewystarczająca z punktu widzenia wyboru dobrej strategii, ponieważ w danej klasie strategii zwykle jest wiele strategii dopuszczalnych. Zatem to kryterium nie prowadzi do jednoznacznego wyboru strategii.

Określmy teraz najsilniejsze kryterium wyboru estymatora w klasie $\mathcal{A}_{to}$ estymatorów nieobciążonych wektora parametrów opisowych populacji $\boldsymbol{\theta}(\mathbf{y})$, takiego, że $\mathbf{y}\in\mathcal{Y}$. Powszech-

nie znaną definicję efektywności estymatora uogólniono na przypadek wektorowych estymatorów poprzez adaptację pojęcia efektywności estymatorów wektorowych wprowadoznej przez Borowkowa (1984), s. 109.

Definicja 2.11: Statystyka $\mathbf{t}_S^{@} \in \mathcal{A}_{to}$ jest efektywnym estymatorem wektora parametrów opisowych $\boldsymbol{\theta}(\mathbf{y})$ w klasie nieobciążonych estymatorów $\mathcal{A}_{to}$ przy ustalonym planie losowania próby wtedy i tylko wtedy, gdy $\mathbf{t}_S^{@}$ jest nie gorszy w sensie definicji 2.9 od każdego estymatora $\mathbf{t}_S \in \mathcal{A}_{to}$, dla każdego $\mathbf{y} \in \mathcal{Y}$.

Rzeczą ważną staje się określenie warunków, przy których istnieją estymatory efektywne. Godambe i Joshi (1965) wykazali, że dla prowadzącego do badania wyczerpującego planu losowania nie istnieje efektywny estymator w klasie $\mathcal{A}_{to}$ nieobciążonych estymatorów skalarnego parametru $\theta(\mathbf{y})$.

Z drugiej strony są określane zbiory estymatorów, w których jest możliwe wyodrębnienie estymatora efektywnego. Estymatory takie istnieją, gdy plan losowania próby uporządkowanej s spełnia następujące warunki podane przez Cassela i in. (1977), s. 71:

a) każda próba uporządkowana, jaką można wylosować, liczy stałą liczbę elementów populacji i elementy te nie powtarzają się w próbie,
b) n! prób uporządkowanych, z których każda jest permutacją ustalonej próby nieuporządkowanej n elementowej, ma takie samo prawdopodobieństwo wylosowania,
c) $\pi_k > 0$ dla każdego k=1,...,N.

Wymienione trzy postulaty znajdują się wśród założeń szczegółowych o istnieniu estymatorów efektywnych, które przytaczamy w następnym rozdziale.

2.2. Ocena parametru opisowego populacji jako zagadnienie predykcji

W niniejszym paragrafie uogólniono definicje i twierdzenia związane z problemami oceny wartości średniej w nadpopulacji na przypadek wektorowy.

Zadanie polega na ocenie wektora parametrów opisowych populacji $\boldsymbol{\theta}(\mathbf{y})$ o wymiarze 1×m, przy czym teraz macierz **y** o wymiarach N×m jest realizacją macierzy losowej **Y**, której rozkład prawdopodobieństwa jest określony założonym modelem nadpopulacji. Stąd wynika, że parametr $\boldsymbol{\theta}(\mathbf{y})$ jest realizacją zmiennej losowej $\boldsymbol{\Theta}=\boldsymbol{\theta}(\mathbf{y})$. Z omawianych wcześniej własności modeli nadpopulacji wynika, że rozkład macierzy **Y** może zależeć od wektora parametrów $\boldsymbol{\gamma}=[\gamma_1...\gamma_a]$. Zatem rozkład zmiennej losowej $\boldsymbol{\Theta}$ może pośrednio zależeć od parametrów $\boldsymbol{\gamma}$.

Przyjmijmy, że plan losowania daje próby z ustaloną liczebnością. W wylosowanej próbie $\underline{s}=\{k_1,...,k_n\}$ są obserwowane realizacje $[\mathbf{y}_{k_1}...\mathbf{y}_{k_n}]$ zmiennych losowych $[\mathbf{y}_{k_1}...\mathbf{y}_{k_n}]$ tworzących podmacierz macierzy **y**. W paragrafie 1.5 zdefiniowano pojęcie danych o parametrze populacji. Wprowadzono m.in. dane identyfikowalne z próby nieuporządkowanej $\mathbf{D}=\{(k, \mathbf{y}_k), k\in S\}$ i jej realizację $\mathbf{d}= (k, \mathbf{y}_k), k\in s\}$. Teraz dodatkowo wprowadzamy dane o nadpopulacji: $\mathcal{D}=\{(k,Y_k), k\in S\}$ z próby nieuporządkowanej S oraz dane o nadpopulacji $\mathbf{D}=\{(k,\mathbf{Y}_k), k\in s\}$ z realizacji próby s. Zauważmy, że dana **d** jest realizacją danej **D**.

Zakładamy, że plan losowania próby jest nieinformatywny (niesekwencyjny), co oznacza, że wybór elementów do próby s nie zależy od rozkładu prawdopodobieństwa obserwowanych w niej zmiennych losowych $\mathbf{Y}_{k_1} \ldots \mathbf{Y}_{k_n}$. Wtedy rozkład prawdopodobieństwa zmiennej losowej $\mathcal{D}$ jest zapisywany [por. Cassel i in. (1977)] następująco:

$$P(\mathcal{D} = \mathbf{d}) = P(S = s)\, P\,(\mathbf{Y}_k = \mathbf{y}_k, \text{ dla } k \in s) \qquad (2.16)$$

Wartość parametru opisowego $\boldsymbol{\theta}(\mathbf{y})$ jest oceniana poprzez predykcję realizacji zmiennej losowej $\boldsymbol{\Theta} = \boldsymbol{\theta}(\mathbf{Y})$ [28]. Predyktor wektora losowego $\boldsymbol{\Theta}$ jest statystyką $\mathbf{t}(\mathcal{D})$, którą będziemy także oznaczać przez $\mathbf{T}_S$. Dla ustalonej próby s predyktor jest funkcją danych **D**, czyli $\mathbf{T}_s=\mathbf{t}(\mathbf{D})$. Z kolei dla ustalonych obserwacji zmiennych w próbie otrzymujemy realizację danych **d**, której funkcją jest wartość predyktora $\mathbf{t}_s = \mathbf{t}(\mathbf{d})$.

Parę uporządkowaną $(\mathbf{T}_S,P)$, gdzie P=P(S) jest ustalonym planem losowania próby, nazywamy strategią predykcji parametru $\boldsymbol{\Theta}$.

Definicja 2.12 [Cassel i in. (1977), s. 92]: Strategię predykcji $(\mathbf{T}_S,P)$ nazywamy ξ nieobciążoną dla $\boldsymbol{\Theta}$ wtedy i tylko wtedy, gdy dla danego rozkładu prawdopodobieństwa ξ macierzy losowej **Y** zachodzi równanie:

$$\mathcal{E}(\mathbf{T}_S - \boldsymbol{\Theta}) = \mathbf{0} \qquad (2.17)$$

Strategia predykcji jest P-ξ nieobciążona dla $\boldsymbol{\Theta}$ wtedy i tylko wtedy, gdy:

$$\mathcal{E}\, E(\mathbf{T}_S - \boldsymbol{\Theta}) = \mathbf{0} \qquad (2.18)$$

Stąd oraz z założenia o nieinformatywności próby wynika, że:

$$E\, \mathcal{E}(\mathbf{T}_S - \boldsymbol{\Theta}) = \mathbf{0} \qquad (2.19)$$

Dokładność strategii predykcji jest charakteryzowana za pomocą wartości oczekiwanej macierzy mieszanych momentów zwykłych rzędu drugiego składowych wektora błędów predykcji: $\mathbf{B}_S = \mathbf{T}_S - \boldsymbol{\Theta}$, którą określa wzór:

$$\mathcal{E}[\mathbf{V}_{SR}(\mathbf{T}_S)] = \mathcal{E}\, E(\mathbf{T}_S - \boldsymbol{\Theta})^T(\mathbf{T}_S - \boldsymbol{\Theta}) \qquad (2.20)$$

Oznaczmy przez $\boldsymbol{\mathcal{V}}(\mathbf{T}_s) = \mathcal{E}(\mathbf{T}_s - \mathcal{E}(\mathbf{T}_s)]^T[(\mathbf{T}_s - \mathcal{E}(\mathbf{T}_s)]$ ξ macierz wariancji i kowariancji, a przez $\boldsymbol{\mathcal{B}}(\mathbf{T}_s) = \mathcal{E}[\mathbf{T}_s - \mathcal{E}(\mathbf{T}_s)]$ ξ obciążenie strategii predykcji $\mathbf{T}_s$ dla ustalonej próby s.

Twierdzenie [29] 2.3.: Jeśli $(\mathbf{T}_S, P)$ jest dowolną strategią predykcji wektora $\boldsymbol{\Theta}$, to dla każdego rozkładu prawdopodobieństwa ξ i każdego nieinformatywnego planu losowania próby P zachodzi:

$$\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_S) = E\boldsymbol{\mathcal{V}}(\mathbf{T}_S) + E\boldsymbol{\mathcal{B}}^2(\mathbf{T}_S) + \boldsymbol{\mathcal{V}}(\boldsymbol{\Theta}) - 2\mathcal{E}\{\boldsymbol{\Theta} - \mathcal{E}(\boldsymbol{\Theta})\}E\{T_S - \mathcal{E}(\boldsymbol{\Theta})\} \qquad (2.21)$$

---

[28] Wartością zmiennej losowej $\boldsymbol{\Theta}$ jest właśnie parametr $\boldsymbol{\theta}(\mathbf{y})$.

[29] Twierdzenie jest bezpośrednim uogólnieniem na przypadek predykcji wektorowej lematu o dekompozycji błędu średniokwadratowego predykcji, który podają Cassel i in. (1977), s. 94.

W szczególności:
a) jeśli $\mathbf{T}_S$ jest P nieobciążony, to:

$$\mathcal{E}\mathbf{V}(\mathbf{T}_S) = E\boldsymbol{\mathcal{V}}(\mathbf{T}_S) + E\boldsymbol{\mathcal{B}}^2(\mathbf{T}_S) - \boldsymbol{\mathcal{V}}(\Theta) \tag{2.22}$$

b) jeśli $\mathbf{T}_S$ jest P-ξ nieobciążony, to:

$$\mathcal{E}\mathbf{V}(\mathbf{T}_S) = E\boldsymbol{\mathcal{V}}(\mathbf{T}_S) - \boldsymbol{\mathcal{V}}(\Theta) \tag{2.23}$$

Podobnie jak to miało miejsce w przypadku studiowanego w rozdziale 2.1 zagadnienia estymacji parametrów populacji skończonej i ustalonej do oceny dokładności strategii predykcji wykorzystujemy miary syntetyczne takie, jak średni promień strategii $\{\mathbf{T}_S, P\}$ predykcji wektora Θ:

$$q_{SR}(\mathbf{T}_S)=\sqrt{\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_S)} \tag{2.24}$$

Uogólniony błąd średniokwadratowy strategii predykcji:

$$g_{SR}(\mathbf{T}_S)=\det \mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_S) \tag{2.25}$$

Średni promień spektralny strategii $\rho_{SR}(\mathbf{T}_S)$ jest równy maksymalnej wartości własnej macierzy $\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_S)$.

Przyjmijmy, że własności nadpopulacji opisują rozkłady prawdopodobieństwa ξ macierzy losowej $\mathbf{Y}$ zależne od wektora parametrów $\boldsymbol{\gamma} = [\gamma_1...\gamma_a]$. Klasę tych rozkładów oznaczmy przez $\boldsymbol{\mathcal{M}}$.

Zgodnie z definicją 2.6 wnioskujemy, że rozkład błędu predykcji $\mathbf{B}_{S1}=\mathbf{T}_{S1}-\Theta$ ma nie większe zróżnicowanie od błędu $\mathbf{B}_{S2}=\mathbf{T}_{S2}-\Theta$ wokół wektora zerowego wtedy i tylko wtedy, gdy:

$$\bigwedge_{\mathbf{e}\in R^z-\{\mathbf{o}_z\}} \mathcal{E}E(\mathbf{B}_{S1}\mathbf{e}^T)^2 \le \mathcal{E}E(\mathbf{B}_{S2}\mathbf{e}^T)^2 \tag{2.26}$$

Jeśli przynajmniej dla jednego wektora wierszowego $\mathbf{e} \in R^z$ zapisana wyżej nierówność staje się ostra, to mówimy, że rozkład $\mathbf{B}_{S1}$ jest mniej zróżnicowany od $\mathbf{B}_{S2}$ wokół wektora zerowego. Dodajmy, że po odpowiednich prostych przekształceniach nierówność (2.26) daje się sprowadzić do postaci:

$$\bigwedge_{\mathbf{e}\in R^z-\{\mathbf{o}_z\}} \mathbf{e}\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_{S1},P_1)\mathbf{e}^T \le \mathbf{e}\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_{S2}P_2)\mathbf{e}^T \tag{2.27}$$

Definicja [30] 2.13: Strategia predykcji $\{\mathbf{T}_1, P_1\}$ jest nie gorsza od $\{\mathbf{T}_2,P_2\}$, jeśli dla każdego $\xi\in\mathcal{M}$ zachodzi nierówność (2.26) lub (2.27), czyli gdy zróżnicowanie rozkładu błędu predykcji $\mathbf{B}_{S1}$ wokół wektora zerowego jest nie większe od takiego zróżnicowania rozkładu błędu $\mathbf{B}_{S2}$. Gdy istnieją takie $\mathbf{e}\in R^m$ i $\xi\in\mathcal{M}$, że nierówności (2.26) i (2.27) stają się ostre, to mówimy, iż strategia predykcji $\{\mathbf{T}_{S1}, P_1\}$ jest lepsza od $\{\mathbf{T}_{S2}, P_2\}$.

Z lematu 2.1 wynika następująca własność:

Twierdzenie 2.4: Jeśli dla każdego rozkładu $\xi\in\mathcal{M}$, macierz $\mathbf{L}=\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_{S2},P_2)-\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_{S1},P_1)$ jest nieujemnie określona, to strategia predykcji $\{\mathbf{T}_{S1},P_1\}$ jest nie gorsza od $\{\mathbf{T}_{S2},P_2\}$.

Strategię predykcji $\{\mathbf{T}_{S1},P\}$ nazwiemy nieosobliwą wtedy i tylko wtedy, gdy oczekiwana macierz błędu średniokwadratowego predykcji jest nieosobliwa, czyli: $\det\mathcal{E}\mathbf{V}_{SR}(\mathbf{T}_S,P)>0$.

Podobnie jak twierdzenie 2.2 można wykazać następujące:

Twierdzenie 2.5: Jeśli nieosobliwa strategia predykcji $\{\mathbf{T}_{S1},P_1\}$ daje nie gorsze prognozy realizacji wektora parametrów Θ od strategii predykcji $\{\mathbf{T}_{S2},P_2\}$ w sensie definicji 2.12 dla każdego $\xi\in\mathcal{M}$, to:

$$\mathcal{E}q_{SR}(\mathbf{T}_{S1},P_1) \le \mathcal{E}q_{SR}(\mathbf{T}_{S2},P_2) \tag{2.28}$$

$$\mathcal{E}g_{SR}(\mathbf{T}_{S1},P_1) \le \mathcal{E}g_{SR}(\mathbf{T}_{S2},P_2) \tag{2.29}$$

$$\mathcal{E}\rho_{SR}(\mathbf{T}_{S1},P_1) \le \mathcal{E}\rho_{SR}(\mathbf{T}_{S2},P_2) \tag{2.30}$$

2.3. Uwagi o budowie przedziałów i obszarów ufności

Estymacja przedziałowa jest w pewnym sensie uzupełniającym w stosunku do estymacji punktowej sposobem oceny parametrów opisowych populacji. Podstawy teorii estymacji przedziałowej wprowadził Neyman (1937), co podajemy za Wilksem (1962). W przypadku estymacji skalarnego parametru $\theta(\mathbf{y}) \in < \theta_1; \theta_2 >$ przedział ufności określamy następująco: Dla zadanego poziomu ufności $\gamma\in(0;1)$ wyznaczamy takie statystyki $T_-(\mathbf{D})$ i $T_+(\mathbf{D})$, że:

$$P\{T_-(\mathbf{D}) < \theta(\mathbf{y}) < T_+(\mathbf{D}) \mid \boldsymbol{\theta}\} = \gamma \tag{2.31}$$

Przedział $(T_-(\mathbf{D}); T_+(\mathbf{D}))$ nazywamy przedziałem ufności dla parametru opisowego populacji $\theta(\mathbf{y})$.

Do oceny precyzji estymacji obok współczynnika ufności służy długość przedziału ufności oznaczana przez $\Delta = T_+ - T_-$. Zmierza się do tego, by znaleźć przedział ufności o najkrótszej długości przy ustalonym poziomie ufności. Formalne metody wyznaczania takich przedziałów ufności są prezentowane np. w pracach Borowkowa (1984) i Wilksa (1962). Są

[30] Określenie to jest rozszerzeniem na przypadek predykcji wektorowej odpowiednich pojęć wprowadzonych przez Cassela i in. (1977), s. 93.

one jednak konstruowane w ramach klasycznej koncepcji wnioskowania i zwykle jest trudno je przenieść na grunt estymacji wykorzystującej próby losowane z ustalonych i skończonych populacji. Z tego powodu ograniczymy się do podania pewnej ogólnej formuły określającej przedział ufności dla parametru opisowego $\theta(\mathbf{y})$ populacji ustalonej. W tym celu niech $t(\mathbf{D})$ będzie nieobciążonym estymatorem punktowym parametru $\theta(\mathbf{y})$. Przez $v(t(\mathbf{D}))$ oznaczamy wariancję statystyki $t(\mathbf{D})$, a przez z_γ jej kwantyl rzędu $\frac{1}{2}(1+\gamma)=P\left(t(\mathbf{D})<z_\gamma\right)$. Wtedy przedział ufności wyznaczamy z wyrażenia:

$$P\left\{t(\mathbf{D})-z_\gamma\sqrt{v(t(\mathbf{D}))}<\theta(\mathbf{y})<z_\gamma\sqrt{v(t(\mathbf{D}))}+t(\mathbf{D})\right\}\geq\gamma \tag{2.32}$$

Gdy wariancja $v(t(\mathbf{D}))$ nie jest znana, to zastępujemy ją przez odpowiedni estymator. Określony przedział ufności jest często stosowany w metodzie reprezentacyjnej [por. np. Cochran (1963); Konijn (1973)], gdy liczebność próby jest bardzo duża i statystyka $t(\mathbf{D})$ jest przynajmniej asymptotycznie nieobciążonym estymatorem parametru $\theta(\mathbf{y})$ i jest znany jej rozkład graniczny, co pozwala obliczyć wartość z_γ.

Niech zbiór losowy $\mathbf{L}(\mathbf{D})$ zależny od danych $\mathbf{D}$ będzie określony układem nierówności: $G_i(\mathbf{D},\boldsymbol{\theta}(\mathbf{y})) < g_i$ $(i=1,...,h)$. Załóżmy, że każda funkcja G_i jest określona na iloczynie kartezjańskim przestrzeni $\mathcal{P}_{\boldsymbol{\theta}}=\{\boldsymbol{\theta}: \boldsymbol{\theta}=\boldsymbol{\theta}(\mathbf{y}), \mathbf{y}\in\mathcal{Y}\}$ wektorowych parametrów $\boldsymbol{\theta}$ i przestrzeni prób $\mathcal{D}$ macierzy losowej $\mathbf{D}$, za wyjątkiem podzbioru miary zero. Ponadto niech rozkład każdej z tych funkcji nie zależy od parametrów $\boldsymbol{\theta}$.Wówczas [por. np. Borowkow (1984) i Wilks (1962)] $\mathbf{L}(\mathbf{D})$ nazywamy zbiorem (obszarem) ufności dla wektora $\boldsymbol{\theta}$, gdy

$$P\{\boldsymbol{\theta}\in\mathbf{L}(\mathbf{D})\,|\,\boldsymbol{\theta}\}\geq\gamma \tag{2.33}$$

W niniejszej pracy rozważane są obszary ufności w kształcie prostokąta bądź elipsoidy. Niech $1\times m$ wymiarowy wektor $\mathbf{t}(\mathbf{D})$ będzie nieobciążonym estymatorem parametrów $\boldsymbol{\theta}$. Przez nieosobliwą macierz $\tilde{\mathbf{V}}(t(D))$ oznaczamy nieobciążony estymator macierzy wariancji i kowariancji $\mathbf{V}(\mathbf{t}(\mathbf{D}))$. Wtedy zbiór ufności w kształcie elipsoidy dla wektora $\boldsymbol{\theta}$ wyznacza wyrażenie:

$$P\{Q(\mathbf{D}) < q_\gamma\,|\,\boldsymbol{\theta}\}\geq\gamma \tag{2.34}$$

gdzie:

$$Q(\mathbf{D}) = (\mathbf{t}(\mathbf{D}) - \boldsymbol{\theta})\ \tilde{\mathbf{V}}^{-1}(\mathbf{t}(\mathbf{D}))(\mathbf{t}(\mathbf{D}) - \boldsymbol{\theta})^T \tag{2.35}$$

Precyzja estymacji jest określana za pomocą objętości elipsoidy [por. np. Cramer (1958)]:

$$H = cq_\gamma^{m/2}E\{\det^{m/2}\mathbf{V}[\mathbf{t}(\mathbf{D})]\} \tag{2.36}$$

gdzie:

$$c = \frac{\pi^{m/2}}{\Gamma(0.5m+1)} \tag{2.37}$$

natomiast $\mathbf{V}(\mathbf{t}(\mathbf{D}))$ jest macierzą wariancji i kowariancji wektora estymatorów $\mathbf{t}(\mathbf{D})$.

Przy spełnieniu pewnych ogólnych warunków dowodzi się, że statystyka Q(**D**) ma asymptotycznie rozkład χ^2_m z m stopniami swobody.

Prostokątny obszar ufności wyznacza wyrażenie:

$$P\left\{(t_i(\mathbf{D}) - z_\gamma \sqrt{v(t_i(\mathbf{D}))} < \theta_i(y) < z_\gamma \sqrt{v(t_i(\mathbf{D})} + t_i(\mathbf{D}))\right\} \geq \gamma \qquad (2.38)$$

przy czym m dla każdego i=1,...,m statystyka $t_i(\mathbf{D})$ jest nieobciążonym estymatorem parametru $\theta_i(\mathbf{y})$ składowego wektora $\boldsymbol{\theta}$. Niech $Z_i = (t_i(\mathbf{D}) - \theta_i)/\sqrt{v(t_i(\mathbf{D}))}$. Wtedy związek między kwantylem z_* oraz poziom ufności ma postać:

$$P\left\{\prod_{i=1}^{m} (|Z_i| < z_\gamma | \boldsymbol{\theta})\right\} \geq \gamma$$

Zwykle w praktyce możliwe jest jedynie wyznaczenie rozkładu asymptotycznego zmiennych losowych $Z_1,Z_2,...,Z_m$ przy dostatecznie dużej liczebności prób. Precyzja estymacji na podstawie prostokąta ufności jest mierzona jego objętością:

$$H_p = 2^m z_\gamma^m \prod_{i=1}^{m} \sqrt{v(t_i(\mathbf{D}))} \qquad (2.39)$$

Boki określonego prostokąta ufności są równoległe do odpowiednich osi układu współrzędnych. Dodajmy, że prostokąt ten można także konstruować tak, aby znalazł się w innej pozycji w stosunku do tych osi. Może to doprowadzić do zmniejszenia objętości prostokątnego obszaru ufności.

2.4. Nieobciążone estymatory uogólnionej wariancji

Mikhail i Mir (1981) zaproponowali nieobciążone estymatory uogólnionej wariancji z próby prostej losowanej zwrotnie. Otrzymane wcześniej wyniki pozwalają teraz na konstrukcję nieobciążonych estymatorów uogólnionej wariancji, lecz dla bezzwrotnego wariantu losowania próby prostej.

Niech $\mathbf{x}=[x_{ij}]$ będzie macierzą o wymiarach m×N wartości m-zmiennych objaśniających w populacji N-elementowej. Przez $\mathbf{x}_{*j}^T=[x_{1j}...x_{mj}]$ (j=1,...,N) oznaczamy obserwację m-wymiarowej zmiennej poczynioną na j-tym elemencie populacji. Wektor $\mathbf{x}_{i*}=[x_{i1}...x_{iN}]$ (i=1, ..., m) składa sie z obserwacji wartości i-tej zmiennej. Zatem macierz obserwacji **x** można określić następująco:

$$\mathbf{x} = [\mathbf{x}_{*1}\ \mathbf{x}_{*2}\ \ldots\ \mathbf{x}_{*N}] \text{ lub } \mathbf{x} = \begin{bmatrix} \mathbf{x}_{1*} \\ \cdots\cdots \\ \mathbf{x}_{m*} \end{bmatrix}$$

Obserwacją m-wymiarowej zmiennej w próbie s składającej się z n-elementów jest wektor $\mathbf{x}_s = [\mathbf{x}_{*j_1} \ldots \mathbf{x}_{*j_n}]$, przy czym $s=\{j_1, \ldots j_n\}$.

Niech $\mathbf{z} = [x_{ij} - \bar{x}_i]$, gdzie $\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$, (i=1,...,m) oraz $\mathbf{z}_s = [\mathbf{z}_{*j_1} \ldots \mathbf{z}_{*j_n}]$, gdzie $\mathbf{z}^T_{*j_k} = [(x_{1j_k} - \bar{x}_1) \mathrm{K}\ (x_{mj_k} - \bar{x}_m)]$, (k=1,...,n). Niech $\mathbf{u} = [x_{ij} - \bar{x}_i(s)]$, gdzie $\bar{x}_i(s) = \frac{1}{n}\sum_{j\in l} x_{ij}$. Ponadto: $\mathbf{u}_s = [\mathbf{u}_{*j_1} \ldots \mathbf{u}_{*j_n}]$, gdzie $\mathbf{u}^T_{*j_k} = [(x_{1j_k} - \bar{x}_1(s)) \ldots (x_{mj_k} - \bar{x}_m(s))]$, (k=1,…, n). Zatem $\mathbf{z}_s$ i $\mathbf{u}_s$ są podmacierzami odpowiednio macierzy **z** i **u**. Otrzymujemy je pozostawiając w macierzach **z** i **u** kolumny o numerach elementów populacji wchodzących do próby s. Uogólnioną wariancję w populacji definiuje wzór:

$$g = N^{-m} \left| \mathbf{z}\,\mathbf{z}^T \right| \tag{2.40}$$

a uogólnioną wariancję z próby s wzór:

$$g_s = N^{-m} \left| \mathbf{u}_s \mathbf{u}_s^T \right| \tag{2.41}$$

Niech s_m będzie m-elementowym podzbiorem próby nieuporządkowanej s składającej się z n-elementów. W pracach Andersona (1958) lub Mostowskiego i Starka (1977) znajdujemy dekompozycję wyznacznika macierzy, którą przy naszych oznaczeniach można zapisać następująco:

$$\left| \mathbf{z}_s \mathbf{z}_s^T \right| = \sum_{s_m \in s} \left| \mathbf{z}_{s_m} \right|^2 \tag{2.42}$$

gdzie sumacja odbywa się po wszystkich m-elementowych kombinacjach s_m wybieranych z n-elementowego zbioru s. Podobnie otrzymujemy, że:

$$\left| \mathbf{z}\mathbf{z}^T \right| = \sum_{s_m \in U} \left| \mathbf{z}_{s_m} \right|^2 \tag{2.43}$$

Niech $s_{m+1}=\{k_1, \ldots, k_{m+1}\}$ i $s_m=\{k_1, \ldots, k_m\}$ będą podzbiorami próby s, czyli $s_m \subset s_{m+1} \subseteq s$. Miarę (objętość) m-wymiarową równoległościanu rozpiętego na układzie wektorów wspólnie zaczepionych w punkcie o współrzędnych $\mathbf{x}_{k_{m+1}}$ i końcach w punktach o współrzędnych $\mathbf{x}_{k_1}, \ldots, \mathbf{x}_{k_m}$ definiuje wzór (por. np. Jefimow i Rozendorn (1974), s. 262):

$$q(s_{m+1}) = \left| \mathbf{x}_{s_m} - \mathbf{x}_{*k_{m+1}} \mathbf{J}_m^T \right|^2 \tag{2.44}$$

gdzie: $\mathbf{J}_m$ jest m-elementową kolumną jedynek.

Na podstawie otrzymanych przez Wywiała (1989b i 1992a) rezultatów dotyczących własności wyznaczników macierzy w szczególności mamy, że:

$$\left|\mathbf{u}_s\,\mathbf{u}_s^T\right| = \frac{1}{n}\sum_{s_{m+1}\in s} q(s_{m+1}) \tag{2.45}$$

$$N\left|\mathbf{z}\mathbf{z}^T\right| = \sum_{s_{m+1}\in U} q(s_{m+1}) \tag{2.46}$$

Korzystając ze wzoru (2.45) mamy:

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{u}_{\underline{s}}\mathbf{u}_{\underline{s}}^T\right| = n!\sum_{s\in S}\left|\mathbf{u}_s\mathbf{u}_s^T\right| = n!\sum_{s\in S}\frac{1}{n}\sum_{s_{m+1}\in s} q(s_{m+1}) = \quad (n-1)!\sum_{s\in S}\sum_{s_{m+1}\in s} q(s_{m+1})$$

Otrzymaną sumę rozdzielamy na powtarzające się składniki. Dla ustalonego zbioru s_{m+1} składnik $q(s_{m+1})$ powtarza się tyle razy, ile można dobrać (n-m-1) elementowych kombinacji bez powtórzeń tworzących zbiory s-s_{m+1}, spośród elementów zbioru U-s_{m+1}. Zatem:

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{u}_{\underline{s}}\mathbf{u}_{\underline{s}}^T\right| = (n-1)!\binom{N-m-1}{n-m-1}\sum_{s_{m+1}\in U} q(s_{m+1}) \tag{2.47}$$

Stąd i z wyrażenia (2.46) mamy:

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{u}_{\underline{s}}\mathbf{u}_{\underline{s}}^T\right| = (n-1)!\binom{N-m-1}{n-m-1} N\left|\mathbf{z}\,\mathbf{z}^T\right|$$

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{u}_{\underline{s}}\mathbf{u}_{\underline{s}}^T\right| = (n-1)!\binom{N-m-1}{n-m-1} N^{m+1} g \tag{2.48}$$

Korzystając ze wzoru (2.41) mamy:

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{z}_{\underline{s}}\,\mathbf{z}_{\underline{s}}^T\right| = n!\sum_{s\in S}\left|\mathbf{z}_s\,\mathbf{z}_s^T\right| = n!\sum_{s\in S}\sum_{s_m\in s}\left|\mathbf{z}_{s_m}\right|^2$$

W otrzymanej sumie dla ustalonego zbioru s_m składnik $|z_{s_m}|^2$ pojawia się tyle razy ile, można dobrać różniących się od siebie (n-m) elementowych kombinacji bez powtórzeń s-s_m wybranych spośród elementów zbioru U-s_m, a zatem:

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{z}_{\underline{s}}\,\mathbf{z}_{\underline{s}}^{T}\right| = n!\binom{N-m}{n-m}\sum_{s\in U}\left|\mathbf{z}_{s_m}\right|^2$$

Wtedy na podstawie wzoru (2.43) mamy:

$$\sum_{\underline{s}\in\underline{S}}\left|\mathbf{z}_{\underline{s}}\,\mathbf{z}_{\underline{s}}^{T}\right| = n!\binom{N-m}{n-m}\left|\mathbf{z}\,\mathbf{z}^{T}\right| \tag{2.49}$$

Przypomnijmy, że plany losowania próby prostej uporządkowanej i nieuporządkowanej określają odpowiednio wzory:

$$\bigwedge_{\underline{s}\in\underline{S}} P_1(\underline{s}) = \frac{(N-n)!}{N!}, \qquad \bigwedge_{s\in S} P_1(s) = \frac{1}{\binom{N}{n}} \tag{2.50}$$

Rozważmy następujące statystyki:

$$\bar{g}_s = \frac{\binom{N}{n}N^{-m}}{\binom{N-m}{n-m}}\left|z_s z_s^T\right| \tag{2.51}$$

$$\tilde{g}_s = \frac{n\binom{N}{n}N^{-m-1}}{\binom{N-m-1}{n-m-1}}\left|\mathbf{u}_s\mathbf{u}_s^T\right| \tag{2.52}$$

przy czym macierz $\mathbf{z}_S$ i $\mathbf{u}_S$ zdefiniowano w pierwszym punkcie tej pracy. Wtedy mamy:

Twierdzenie 2.6: Jeśli próba prosta jest losowana bezzwrotnie, to statystyki $\bar{g}_S$ i $\tilde{g}_S$ są nieobciążonymi estymatorami uogólnionej wariancji g, danej wzorem (2.40).

Dowód: Ze wzorów (2.49), (2.50) i (2.51) wynika, że $E(\bar{g}) = g$, ponieważ:

$$E(\bar{g}) = \frac{\binom{N}{n}N^{-m}}{\binom{N-m}{n-m}}E\left|\mathbf{z}_s\mathbf{z}_s^T\right| = \frac{\binom{N}{n}N^{-m}}{\binom{N-m}{n-m}}\sum_{s\in S}\left|\mathbf{z}_s\mathbf{z}_s^T\right|P_1(s) =$$

$$= \frac{\binom{N}{n} N^{-m}}{\binom{N-m}{n-m}} \sum_{\underline{s} \in \underline{S}} \left| z_{\underline{s}} z_{\underline{s}}^T \right| P_1(\underline{s}) = \frac{\binom{N}{n} N^{-m}}{\binom{N-m}{n-m}} n! \left| zz^T \right| \binom{N-m}{n-m} \frac{1}{n! \binom{N}{n}} = \left| zz^T \right| N^{-m} = g$$

Podobnie równość $E(\tilde{g}) = g$ wynika ze wzorów (2.50), (2.52) i (2.48):

$$E(\tilde{g}) = \frac{n\binom{N}{n} N^{-m}}{\binom{N-m-1}{n-m-1}} E\left| u_s u_s^T \right| = \frac{n\binom{N}{n} N^{-m}}{\binom{N-m-1}{n-m-1}} \sum_{\underline{s} \in \underline{S}} \left| u_{\underline{s}} u_{\underline{s}}^T \right| P_1(\underline{s}) =$$

$$= \frac{n\binom{N}{n} N^{-m-1}}{\binom{N-m-1}{n-m-1}} (n-1)! \binom{N-m-1}{n-m-1} N^{m+1} g \frac{1}{n! \binom{N}{n}} = g$$

cbdu.

Otrzymane tu nieobciążone estymatory uogólnionej wariancji można użyć do oceny objętości elipsoidalnego obszaru ufności wyznaczonego z próby prostej losowanej bezzwrotnie. Objętość ta jest traktowana jako wskaźnik precyzji estymacji na podstawie elipsoidalnego obszaru ufności, którego ogólną zasadę konstrukcji przedstawiono w uprzednim paragrafie.

3. WEKTOR ŚREDNICH Z PRÓBY PROSTEJ

3.1. Parametry rozkładu wektora średnich

Przypomnijmy, że przez $\bar{\mathbf{y}} = [\bar{y}_1 \ldots \bar{y}_m]$ oznaczyliśmy określony wzorem (1.1) wektor średnich w populacji. Z kolei przez $\bar{\mathbf{y}}_S = [\bar{y}_{1S} \ldots \bar{y}_{mS}]$ oznaczamy wektor średnich z próby prostej, przy czym:

$$\bar{y}_{iS} = \frac{1}{n} \sum_{k \in S} y_k \tag{3.1}$$

Wektor $\bar{\mathbf{y}}_S$ daje nieobciążone oceny parametrów $\bar{\mathbf{y}}$, gdy próba S jest prosta.

Gdy próba jest losowana bezzwrotnie zgodnie z danym wzorem (1.31) planem losowania, to macierz wariancji i kowariancji estymatorów $\bar{\mathbf{y}}_S$ ma postać:

$$\mathbf{V}\left(\bar{\mathbf{y}}_S, P_3\right) = \frac{N-n}{Nn} \mathbf{C}_* \tag{3.2}$$

gdzie $\mathbf{C}_*$ jest określoną wzorem (1.3) macierzą wariancji i kowariancji cechy y w populacji.

Gdy losowanie próby prostej jest zwrotne i zgodne z planem danym wzorem (1.29), to

$$\mathbf{V}\left(\bar{\mathbf{y}}_S, P_1\right) = \frac{N-1}{Nn} \mathbf{C}_* = \frac{1}{n} \mathbf{C} \tag{3.3}$$

Nieobciążone estymatory zapisanych macierzy wariancji i kowariancji otrzymujemy zastępując w nich elementy ich odpowiednimi wariancjami i kowariancjami z próby prostej.

3.2. Wyznaczanie niezbędnego rozmiaru próby przy ustalonych błędach średnich szacunku

Przypomnijmy, że przez $\bar{\mathbf{y}}=[\bar{y}_1 \ldots \bar{y}_m]$ i $\bar{\mathbf{y}}_S=[\bar{y}_{1S} \ldots \bar{y}_{mS}]$ oznaczaliśmy wektory średnich m cech odpowiednio w populacji i w próbie prostej. Niech d_i będzie dopuszczalnym poziomem błędu szacunku i-tej średniej, czyli $D(\bar{y}_{iS}) \le d_i$ (i=1,...,m). Rozwiązując te nierówności dla wariantu bezzwrotnego losowania próby otrzymujemy ciąg niezbędnych liczebności obserwacji:

$$n \ge \underline{n}_i = \left(d^2 v_*^{-2} + N^{-1}\right)^{-1}, \quad i=1,...,m \tag{3.4}$$

Zwykle liczebności $\underline{n}_i$ (i=1,...,m) nie są równe. W tej sytuacji jest wybierana najważniejsza z punktu widzenia celu badania zmienna i odpowiadającą jej liczebność obserwacji przyjmuje się także za rozmiar próby, jakkolwiek może on być niewystarczający do spełnienia żądanych poziomów dokładności oceny parametrów innych cech. Dlatego powinno się ustalać niezbędny rozmiar próby na poziomie największej liczebności obserwacji, czyli $\underline{n} = \max_{i=1,...,m}\{\underline{n}_i\}$.

W praktyce liczebności $\underline{n}_i$ (i=1,...,m) należy oceniać w podobny sposób, jak to już miało miejsce wcześniej, gdzie wyznaczano niezbędną liczebość próby przy estymacji jednej wartości średniej.

3.3. Wyznaczanie niezbędnego rozmiaru próby przy ustalonym poziomie ryzyka estymacji punktowej

Dla uproszczenia otrzymywanych wyników ograniczamy rozważania do przypadku estymacji na podstawie próby losowanej zwrotnie. Symbolem r_o oznaczamy ryzyko dopuszczalne.

Niech funkcja ryzyka ma postać $r=\sum_{i=1}^{m} a_i D^2(\bar{y}_{iS})$, gdzie a_i jest stratą jednostkową spowodowaną niedokładnością szacunku i-tej średniej. Przyjmując, że próba jest losowana zwrotnie mamy:

$$r(n)=\frac{1}{n}\sum_{i=1}^{m} k_i v_i \tag{3.5}$$

Wtedy na podstawie nierówności $r(n) \le r_o$ otrzymujemy:

$$n \geq \underline{n} = \frac{1}{r_o} \sum_{i=1}^{m} k_i v_i \tag{3.6}$$

Niech teraz funkcja ryzyka będzie równa uogólnionej wariancji estymatora wektorowego $\bar{\mathbf{y}}_S$, którą już oznaczono symbolem $g(\bar{\mathbf{y}}_S)$. Wówczas na podstawie nierówności $r_o \leq g(\bar{\mathbf{y}}_S) = n^{-m}\det(\mathbf{C})$ mamy:

$$n \geq \underline{n} = \sqrt[m]{r_o^{-1} \det(\mathbf{C})} \tag{3.7}$$

Uogólniona wariancja estymatora wektorowego jest funkcją objętości elipsoidalnego obszaru ufności wyznaczanego na podstawie tego estymatora (por. rozdział 2). Pozwala to w sposób bezpośredni wykorzystać otrzymane wyniki wyznaczania niezbędnej liczebności próby do otrzymania elipsoidalnego obszaru ufności o żądanej objętości. Uogólnioną wariancję $\det(\mathbf{C})$ można oceniać za pomocą nieobciążonych estymatorów analizowanych w czwartym paragrafie uprzedniego rozdziału.

Ryzyko estymacji można określić jako promień spektralny estymatora $\bar{\mathbf{y}}_S$, który w rozdziale 2 oznaczono symbolem $\rho(\bar{\mathbf{y}}_S)$. Parametr ten jest równy maksymalnej wartości własnej macierzy wariancji i kowariancji $v(\bar{\mathbf{y}}_S) = n^{-1}\mathbf{C}$. Zatem z nierówności $r_0 \geq \rho(\bar{\mathbf{y}}_S)$ wynika, że:

$$n \geq \underline{n} = r_o^{-1} \lambda_1 \tag{3.8}$$

gdzie przez λ_1 oznaczono maksymalną wartość własną macierzy $\mathbf{C}$.

Wyznaczana wyżej liczebność niezbędna $\underline{n}$ jest funkcją elementów macierzy wariancji i kowariancji $\mathbf{C}$. Zastępując wariancje i kowariancje ich nieobciążonymi ocenami wyznaczanymi z próby wstępnej losowanej zwrotnie otrzymujemy estymator $\underline{n}_S$ parametru $\underline{n}$. Statystyka $\underline{n}_S$ jest zgodnym estymatorem liczebności $\underline{n}$, co można wykazać na podstawie znanego twierdzenia Słuckiego o zbieżności stochastycznej rzeczywistych funkcji zmiennych losowych, które można znaleźć np. w pracach Fisza (1967) lub Rao (1982).

3.4. Minimalizacja ryzyka całkowitego estymacji wektora średnich

Zakładamy, że próba jest losowana zwrotnie. Funkcję ryzyka całkowitego wnioskowania formułujemy w następujący sposób:

$$r_c(n) = r(n) + nk_n \tag{3.9}$$

przy czym pierwszy składnik jest kwadratową funkcją ryzyka estymacji wektora $\bar{\mathbf{y}}$ określoną wzorem (3.5), natomiast k_n jest kosztem jednostkowym losowania elementu populacji i obserwacji zmiennych. Wyliczamy, że $r_C(n)$ osiąga minimum dla $n = \underline{n}$, gdzie:

$$\underline{n}=\sqrt{\frac{1}{k_n}\sum_{i=1}^{m}a_i v_i} \tag{3.10}$$

Zatem liczebność niezbędna jest malejącą funkcją kosztów jednostkowych losowania i obserwacji, a rosnącą względem poszczególnych współczynników straty.

Dodajmy, że funkcja ryzyka r(n) może być uogólnioną wariancją, promieniem spektralnym macierzy wariancji i kowariancji wektora estymatorów itp.

3.5. Wyznaczanie niezbędnej liczebności próby przy ustalonym prawdopodobieństwie przekroczenia błędu dopuszczalnego estymacji

Przez d_i oznaczmy dopuszczalny błąd estymacji i-tej średniej, czyli $|\bar{y}_{is}-\bar{y}_i|\le d_i$ (i=1,...,m). Postulat określający spełnienia żądanej dokładności estymacji wszystkich średnich formułujemy w postaci następującej alternatywy zdarzeń:

$$P\left\{\bigcup_{i=1}^{m}\left(|\bar{y}_{is}-\bar{y}_i|\ge d_i\right)\right\}\le\alpha \tag{3.11}$$

Zatem z prawdopodobieństwem nie większym od α może wystąpić przekroczenie dopuszczalnego błędu estymacji przynajmniej jednej spośród szacowanych średnich. Wywiał (1992) korzystając z podanego przez Godwina (1964) uogólnienia nierówności Czebyszewa otrzymał, że:

$$P\left\{\bigcup_{i=1}^{m}\left(|\bar{y}_{is}-\bar{y}_i|>d_i\right)\right\}\le\sum_{i=1}^{m}\frac{D^2(\bar{y}_s)}{d^2}\le\alpha \tag{3.12}$$

Przyjmując, że próba prosta jest losowana zwrotnie z populacji mamy:

$$n\ge\underline{n}=\frac{1}{\alpha}\sum_{i=1}^{m}\frac{v_i}{d_i^2} \tag{3.13}$$

Stąd więc m.in. wynika, że niezbędna liczebność próby jest odwrotnie proporcjonalna do prawdopodobieństwa α, z jakim może być przekraczany dopuszczalny błąd estymacji.

Przypuśćmy, że celem estymacji jest kombinacja liniowa $\bar{z}(\boldsymbol{\beta})=\bar{\mathbf{y}}\boldsymbol{\beta}^T$, gdzie $\boldsymbol{\beta}=[\beta_1...\beta_m]$ i $\boldsymbol{\beta}\boldsymbol{\beta}^T=1$. Do oceny parametru $\bar{z}(\boldsymbol{\beta})$ użyjemy statystyki $\bar{z}_s(\boldsymbol{\beta})=\bar{\mathbf{y}}_s\boldsymbol{\beta}^T$, przy czym $\bar{\mathbf{y}}_s$ jest wektorem nieobciążonych estymatorów wektora $\bar{\mathbf{y}}$. Z ogólnie znanej nierówności Czebyszewa mamy:

$$P\left\{\left|\bar{z}_s(\beta)-\bar{z}(\beta)\right|\geq d\right\}\leq\frac{D^2\left(\bar{z}_s(\beta)\right)}{d^2} \tag{3.14}$$

Nierówność ta jest równoważna następującej:

$$P\left\{\left|(\bar{y}_s-\bar{y})\beta\right|\geq d\right\}\leq\frac{\beta V(\bar{y}_s)\beta^T}{d^2} \tag{3.15}$$

Stąd i ze wzoru (2.8) wynika, że

$$\beta V(\bar{y}_s)\beta^T\leq\rho(\bar{y}_s)=\max_{\beta\beta^T=1}\text{imum}\left\{\beta V(\bar{y}_s)\beta^T\right\}$$

Stąd i ze wzoru (3.15) otrzymujemy następujące uogólnienie nierówności Czebyszewa:

$$P\left\{\left|(\bar{y}_s-\bar{y})\beta^T\right|\geq d\right\}\leq\frac{\rho(\bar{y}_s)}{d^2} \tag{3.16}$$

gdzie: przez $\rho(\bar{y}_s)$ oznaczono promień spektralny (czyli maksymalną wartość własną) macierzy wariancji i kowariancji $V(\bar{y}_s)$. W przypadku jednowymiarowym, gdy m = 1, to wyrażenie (3.16) redukuje się do nierówności Czebyszewa.

W podobny sposób otrzymujemy, że

$$P\left\{\left|(\bar{y}_s-\bar{y})\beta^T\right|<d\right\}>1-\frac{\rho(\bar{y}_s)}{d^2} \tag{3.17}$$

Załóżmy, że $\bar{y}_s$ jest wektorem średnich z próby prostej losowanej zwrotnie. Wtedy ze wzoru (3.16) mamy:

$$P\left\{\left|(\bar{y}_s-\bar{y})\beta^T\right|\geq d\right\}\leq\frac{\lambda_1}{nd^2} \tag{3.18}$$

gdzie: λ_1 jest maksymalną wartością własną macierzy wariancji i kowariancji **C**. Zatem ograniczenie prawdopodobieństwa tego, że błąd oceny kombinacji liniowej wektora średnich (przy nieznanych wartościach współczynników β kombinacji lecz spełniających równość $\beta\beta^T=1$) przekroczy dopuszczalny poziom, jest wprost proporcjonalne do promienia spektralnego macierzy wariancji i kowariancji, a odwrotnie proporcjonalne do liczebności próby.

Postulując, aby określone wyrażenieniem (3.18) prawdopodobieństwo było mniejsze od poziomu α, otrzymujemy:

$$n\geq\underline{n}=\frac{\lambda_1}{\alpha d^2} \tag{3.19}$$

Zatem niezbędna liczność próby jest malejącą funkcją postulowanego prawdopodobieństwa α oraz dopuszczalnego błędu d oceny kombinacji liniowej średnich.

Koszty obserwacji wartości poszczególnych cech w próbie mogą być bardzo zróżnicowane. W związku z tym mniej obserwacji można zebrać w próbie tych cech, których pomiar jest bardzo kosztowny, a więcej obserwacji tych zmiennych, których pomiar jest tańszy. Problemem optymalizacji liczb obserwacji poszczególnych cech w próbie prostej zajmował się Wywiał (1988b, 1992).

4. WEKTOR ESTYMATORÓW HORVITZA-THOMPSONA

W niniejszym rozdziale przedstawimy najważniejsze własności estymatora Horvitza i Thompsona (1952). Estymator ten jest wyznaczany na podstawie próby losowanej z różnymi prawdopodobieństwami doboru do niej elementów, które są zwykle funkcją wartości zmiennych pomocniczych obserwowanych w całej populacji. Dla wybranych schematów losowania próby wyprowadzono przybliżone wartości wariancji. Pozwala to na wybór planu losowania próby optymalnego z punktu widzenia własności populacji charakteryzowanej spodziewanym układem parametrów zmiennej badanej i cech pomocniczych. Wyznaczono także macierz wariancji i kowariancji tych estymatorów.

4.1. Własności podstawowe

Przyjmijmy, że z populacji jest losowana bezzwrotnie próba S o ustalonej liczebności n. Niech $a_k=1$, gdy k-ty element populacji należy do próby, czyli $k\in S$, natomiast a=0, gdy $k\notin S$. Wtedy mamy [por. np. pracę Cassela i in. (1977)]:

$$\pi_k = E(a_k), \qquad \pi_{ki} = E(a_k a_i), \qquad \sum_{k=1}^{N} \pi_k = n,$$

$$D^2(a_k) = \pi_k(1-\pi_k), \qquad Cov(a_k, a_i) = \pi_{ki} - \pi_k \pi_i.$$

Horvitz i Thompson (1952) zaproponowali estymator średniej $\bar{y}$ w populacji ustalonej, który ma postać:

$$t_{HTS} = \frac{1}{N} \sum_{k=1}^{N} \frac{a_k y_k}{\pi_k} \tag{4.1}$$

Statystyka t_{HTS} jest nieobciążonym estymatorem parametru $\bar{y}$. Jej wariancja jest postaci:

$$D^2\left(t_{HTS}\right)=\frac{1}{N^2}\sum_{k=1}^{N}\left(\frac{y_k}{\pi_k}\right)^2\pi_k\left(1-\pi_k\right)+\frac{1}{N^2}\sum_{k\neq i=1}^{N}\sum^{N}\frac{y_k y_i}{\pi_k\pi_i}\left(\pi_{ki}-\pi_i\pi_k\right) \tag{4.2}$$

Alternatywny sposób wyliczania wariancji podają Sen (1953) oraz Yates i Grundy (1953):

$$D^2\left(t_{HTS}\right)=\frac{1}{N^2}\sum_{k<i=1}^{N}\sum^{N}\left(\pi_k\pi_i-\pi_{ki}\right)\left(\frac{y_k}{\pi_k}-\frac{y_i}{\pi_i}\right)^2 \tag{4.3}$$

Do estymacji parameru $D^2\left(t_{HTS}\right)$ jest używana statystyka:

$$\overline{D}^2(t_{HTS})=\frac{1}{N^2}\sum_{k=1}^{N}a_k\left(\frac{y_k}{\pi_k}\right)^2\left(1-\pi_k\right)+\frac{1}{N^2}\sum_{k\neq i=1}^{N}\sum^{N}a_k a_i\,\frac{y_k y_i}{\pi_k\pi_i}\,\frac{\pi_{ki}-\pi_k\pi_i}{\pi_{ki}} \tag{4.4}$$

która jest nieobciążonym estymatorem wariancji $D^2(t_{HTS})$, chociaż w pewnych sytuacjach może przyjmować wartości ujemne z prawdopodobieństwem większym od zera.

Jeśli $\pi_k\,\pi_i - \pi_{ki} > 0$ dla każdego $k \neq i = 1,...,N$, to tzw. estymator Sena-Yatesa-Grundy'ego przyjmuje tylko wartości nieujemne i ma postać:

$$\tilde{D}^2(t_{HTS})=\frac{1}{N^2}\sum_{k>i i=1}^{N}\sum^{N}a_k a_i\,\frac{y_k y_i}{\pi_k\pi_i}\,\frac{\pi_k\pi_i-\pi_{ki}}{\pi_{ki}}\left(\frac{y_k}{\pi_k}-\frac{y_i}{\pi_i}\right)^2 \tag{4.5}$$

Załóżmy, że próba o ustalonej liczebności n jest teraz losowana zwrotnie. Niech p_k będzie prawdopodobieństwem doboru k-tego elementu populacji do próby. Przez $0\le n_{Sk}\le n$ oznaczmy krotność pojawienia się k-tego elementu w próbie S, przy czym $n=\sum_{k=1}^{N}n_{Sk}$. Wtedy $E(n_{Sk})=\pi_k = np_k$ [por. np. Konijn (1973), s. 240], natomiast estymator Horvitza-Thompsona przyjmuje postać:

$$\bar{t}_{HTS}=\frac{1}{nN}\sum_{k=1}^{N}\frac{y_k n_{Sk}}{p_k} \tag{4.6}$$

Statystyka $\bar{t}_{HTS}$ jest nieobciążonym estymatorem średniej $\bar{y}$, a jej wariancję określa wzór [por. Konijn (1973), s. 240]:

$$D^2(\bar{t}_{HTS})=\frac{1}{n}\sum_{k>i}^{N}\sum_{i=1}^{N}\left(\frac{y_k}{\pi_k}-\frac{y_i}{\pi_i}\right)^2 p_k p_i \tag{4.7}$$

Z kolei nieobciążonym estymatorem zapisanej wariancji jest statystyka [por. np. Konijn (1973), s. 241]:

$$\overline{D}^2(\bar{t}_{HTS)} = \frac{1}{n(n-1)} \sum_{k=1}^{N} \left(\frac{y_k}{\pi_k} - \bar{t}_{HTS} \right)^2 n_{Sk} \tag{4.8}$$

lub $$\overline{D}^2\left(\bar{t}_{HTS}\right) = \frac{1}{n(n-1)} \sum_{k>i}^{N} \sum_{i=1}^{N} \left(\frac{y_k}{\pi_k} - \frac{y_i}{\pi_i} \right)^2 n_{Si} n_{Sk} \tag{4.9}$$

Przeprowadzimy teraz analizę wpływu losowania próby z dowolnymi prawdopodobieństwami inkluzji pierwszego rzędu na wartość oczekiwaną zwykłej średniej z próby. Niech $\bar{y}_S$ będzie średnią z próby losowanej bezzwrotnie. Wtedy:

$$E(\bar{y}_S) = E\left(\frac{1}{n} \sum_{k \in S} y_k \right) = E\left(\frac{1}{n} \sum_{k=1}^{N} y_k a_k \right) = \frac{1}{n} \sum_{k=1}^{N} y_k \pi_k =$$

$$= \bar{y} - \bar{y}(1 - \frac{N}{n}\bar{\pi}) + \frac{N}{n} v(y)v(g)\rho(y,g) \tag{4.10}$$

gdzie: g to zmienna, której realizacjami są prawdopodobieństwa inkluzji pierwszego rzędu, a $\rho(y,g)$ jest współczynnikiem korelacji liniowej między zmiennymi *y* i *g*. W przypadku próby losowanej zwrotnie mamy:

$$E(\bar{y}_S) = E\left(\frac{1}{n} \sum_{k \in S} y_k n_{Sk} \right) = E\left(\frac{1}{n} \sum_{k=1}^{N} y_k n_k \right) = \frac{1}{n} \sum_{k=1}^{N} y_k p_k =$$

$$= \bar{y} + N v(y)v(p)\rho(y,p) \tag{4.11}$$

gdzie: *p* jest zmienną, której realizacjami są prawdopodobieństwa, z jakimi są losowane zwrotnie poszczególne elementy populacji do próby. Zatem $\rho(y,p)$ jest współczynnikiem korelacji tej zmiennej ze zmienną badaną *y*.

Ze wzorów (4.10) i (4.11) wynika, że zwykła średnia z próby może dać obciążone oceny przeciętnej w populacji, gdy próba jest losowana z różnymi prawdopodobieństwami inkluzji rzędu pierwszego.

Statystyka t_{HTS} jest szczególnym przypadkiem studiowanego przez Cassela i in. (1977) uogólnionego estymatora różnicowego.

Godambe i Joshi (1965) wykazali, co następuje:

Twierdzenie 4.1: Dla każdego planu losowania z $\pi_k > 0$ /k=1,..,N/ estymator t_{HTS} jest dopuszczalny w klasie $\mathcal{A}_u$ wszystkich nieobciążonych estymatorów średniej $\bar{y}$.

Z twierdzenia tego wynika, że w szczególności estymator t_{HTS} jest dopuszczalny w klasie $\mathcal{A}_{uo}$ wszystkich nieobciążonych i liniowych estymatorów średniej $\bar{y}$, co wcześniej udowodnili niezależnie od siebie Godambe (1960) oraz Roy i Chakravarti (1960).

Godambe i Joshi (1965) udowodnili, że dla każdego planu losowania próby z nie ustaloną efektywną liczebnością próby statystyka $\bar{t}_{HTS}$ jest niedopuszczalnym estymatorem średniej $\bar{y}$ w klasie $\mathcal{A}$ wszystkich estymatorów.

Z kolei Joshi (1965, 1966) wykazał następujące twierdzenie:

Twierdzenie 4.2: Dla każdego planu losowania próby z ustaloną efektywną liczebnością próby estymator t_{HTS} jest dopuszczalny w klasie $\mathcal{A}$ wszystkich estymatorów średniej $\bar{y}$.

Niech $\mathcal{K}_u$ będzie zbiorem strategii (t_S, P), gdzie P jest planem losowania z ustaloną na poziomie n oczekiwaną efektywną liczebnością próby, natomiast t_S jest nieobciążonym estymatorem $\bar{y}$.

Twierdzenie 4.3 [Ramakrishnan (1975)]: Każda strategia estymacji (t_{HTS}, P) jest dopuszczalna w $\mathcal{K}_u$ dla średniej $\bar{y}$.

Twierdzenie 4.4 [Sarndal (1976)][1]: Dla każdego planu losowania próby uporządkowanej spełniającego warunki: a/, b/, c/ określone na końcu punktu 2.1.3 estymator t_{HTS} ma najmniejszą wariancję w klasie nieobciążonych estymatorów średniej $\bar{y}$ zależnych od danych tylko poprzez nieuporządkowany zbiór wielkości:

$$\left\{ \frac{ny_k}{N\pi_k}, k \in S \right\}$$

Na zakończenie dodajmy, że Hartley i Rao (1968, 1969) podjęli studia nad warunkami, przy których statystyka Horvitza-Thompsona jest estymatorem metody największej wiarygodności.

4.2. Estymacja wskaźnika struktury

Częstym celem estymacji jest wskaźnik struktury $\alpha = \frac{M}{N}$ elementów populacji z cechą wyróżnioną. Niech $\delta_k = 1$, gdy k-ty element populacji posiada wyróżnioną cechę, natomiast $\delta_k = 0$, gdy jej nie posiada. Wtedy populację $\Omega = \{\omega_k\}$ można rozłożyć na takie dwa rozłączne podzbiory Ω_0 i Ω_1, że $\Omega_0 \cup \Omega_1 = \Omega$ i $\Omega_0 = \{\omega_k : \omega_k \in \Omega \text{ i } \delta_k = 0\}$ oraz $\Omega_1 = \{\omega_k : \omega_k \in \Omega \text{ i } \delta_k = 1\}$. Wtedy parametr α możemy zapisać następująco:

$$\alpha = \frac{1}{N} \sum_{k=1}^{N} \delta_k$$

[1] Szczególne twierdzenie w stosunku do cytowanego wykazał także Sarndal (1972).

Nieobciążonym estymatorem parametru α z próby losowanej bezzwrotnie jest statystyka będąca szczególnym przypadkiem estymatora Horvitza-Thompsona, danego wzorem (4.1). Ma ona postać:

$$\alpha_{HTS} = \frac{1}{N}\sum_{k\in\Omega_1}\frac{a_k}{\pi_k} \qquad (4.12)$$

Jej wariancję otrzymujemy ze wzoru (4.2):

$$D^2(\alpha_{HTS}) = N^{-2}\left(\sum_{k\in\Omega_1}\pi_k^{-1} + \sum_{k\neq l}\frac{\pi_{kl}}{\pi_k\pi_l} - M^2\right) \qquad (4.13)$$

Nieobciążony estymator zapisanej wariancji otrzymujemy ze wzoru (4.4) lub (4.5) zastępując w nich zmienną y_k przez δ_k (k=1,...,N).

W szczególności jeśli bezzwrotnie losowana próba S o liczebności n jest próbą prostą, to estymator α_{HTS} redukuje się do zwykłej częstości względnej z próby:

$$\alpha_S = \frac{m_S}{n} = \frac{1}{n}\sum_{k\in\Omega_1}a_k$$

(4.14)

gdzie m_S jest ilością elementów z cechą wyróżnioną w próbie S.

W przypadku zwrotnego losowania ze wzoru (4.6) mamy następujący estymator:

$$\overline{\alpha}_{HTS} = \frac{1}{nN}\sum_{k\in\Omega_1}\frac{n_{Sk}}{p_k}$$

(4.15)

gdzie n_{Sk} (k=1,...,N) jest zmienną losową, której wartości to możliwe ($0\le n_{sk}\le n$) krotności pojawienia się k-tego elementu w próbie S. Wariancję statystyki $\overline{\alpha}_{HTS}$ oraz jej estymator można wyliczyć ze wzorów (4.8) lub (4.9).

Jeśli S jest prostą próbą losowaną zwrotnie, to estymator $\overline{\alpha}_{HTS}$ redukuje się do znanej postaci:

$$\overline{\alpha}_S = \frac{1}{n}\sum_{k\in\Omega_1}n_{Sk} = \frac{\overline{m}_S}{n}$$

(4.16)

Podobnie, jak to miało miejsce wcześniej w przypadku estymacji wartości średniej zbadajmy wpływ losowania próby z różnymi prawdopodobieństwami na wartość oczekiwaną estymatorów α_S i $\overline{\alpha}_S$. Na podstawie wzorów (4.10) i (4.11) mamy:

$$E(\alpha_S)=\frac{1}{n}\sum_{k\in\Omega_1}\pi_k=\alpha-\alpha(1-\frac{N}{n}\bar{\pi}_1)$$

(4.17)

gdzie $\bar{\pi}_1=\frac{1}{M}\sum_{k\in\Omega_1}\pi_k$ jest średnią arytmetyczną prawdopodobieństw inkluzji (rzędu pierwszego) przypisanych elementom populacji z cechą wyróżnioną.

$$E(\bar{\alpha}_S)=\frac{1}{n}\sum_{k\in\Omega_1}p_k=\alpha-\alpha(1-N\bar{p}_1)$$

(4.18)

gdzie $\bar{p}_1=\frac{1}{M}\sum_{k\in\Omega_1}p_k$ jest średnią arytmetyczną prawdopodobieństw zwrotnego losowania elementów populacji z cechą wyróżnioną.

4.3. Parametry rozkładu wektora estymatorów

Załóżmy, że próba jest losowana bezzwrotnie z populacji ustalonej. Oznaczmy przez $\mathbf{t}_{HTS}=[t_{HT1S}...t_{HTmS}]$ wektor estymatorów średnich $\bar{y}=[\bar{y}_1\Lambda\ \bar{y}_m]$ m cech w populacji ustalonej, przy czym:

$$t_{HTiS}=\frac{1}{N}\sum_{k=1}^{m}\frac{a_k y_{ik}}{\pi_k}\quad,\quad i=1,K,m$$

Stąd, że $E(t_{HTiS})=\bar{y}_i$ dla każdego i=1,..,m, wynika, że wektor $\mathbf{t}_{HTiS}$ jest nieobciążonym estymatorem wektora $\bar{\mathbf{y}}$.

Kowariancje par składowych wektora $\mathbf{t}_{HTiS}$ wyprowadził Wywiał (1992):

$$Cov(t_{HTiS},t_{HTjS})=N^{-2}\sum_{k=1}^{N}\frac{y_{ik}y_{jk}}{\pi_k^2}D^2(a_k)+2N^{-2}\sum_{k>h=1}^{N}\sum^{N}\frac{y_{ik}y_{jh}}{\pi_k\pi_h}Cov(a_k,a_h)\qquad(4.19)$$

Stąd otrzymujemy:

$$Cov(t_{HTiS},t_{HTjS})=N^{-2}\sum_{k=1}^{N}\frac{y_{ik}y_{jk}}{\pi_k^2}\pi_k(1-\pi_k)+$$

$$+2N^{-2}\sum_{k>h=1}^{N}\sum^{N}\frac{y_{ik}y_{jh}}{\pi_k\pi_h}\left(\pi_{kh}-\pi_k\pi_h\right) \tag{4.20}$$

lub:

$$\mathrm{Cov}(t_{HTiS},t_{HTjS})=N^{-2}\sum_{k>h=1}^{N}\sum^{N}\left(\frac{y_{ik}}{\pi_k}-\frac{y_{ih}}{\pi_h}\right)\left(\frac{y_{jk}}{\pi_k}-\frac{y_{jh}}{\pi_h}\right)\left(\pi_k\pi_h-\pi_{kh}\right) \tag{4.21}$$

Po to, by w sposób syntetyczny zapisać macierz wariancji i kowariancji wektora estymatorów $\mathbf{t}_{HTS}$, wprowadzamy następujące oznaczenia. Przez $\mathbf{G}_\pi$ oznaczamy macierz diagonalną stopnia N z elementami głównej przekątnej równymi kolejno prawdopodobieństwom π_k (k=1,..,N). Niech $\mathbf{a}$ = $[a_1 \ldots a_N]$ i $\mathbf{\Pi}$ = $[\pi_1 \ldots \pi_N]$. Wtedy:

$$\mathbf{t}_{\mathrm{HTS}}-\mathbf{y}=\left(\frac{1}{N}\sum_{k=1}^{N}\frac{y_{ik}}{\pi_k}\left(a_k-\pi_k\right)\right)=\left(\mathbf{a}-\mathbf{\Pi}\right)\mathbf{G}_\pi^{-1}\mathbf{y}$$

Stąd już otrzymujemy:

$$\mathbf{V}(\mathbf{t}_{HTS})=N^{-2}\,\mathbf{y}^T\ \mathbf{G}_\pi^{-1}\mathbf{V}(\mathbf{a})\mathbf{G}_\pi^{-1}\mathbf{y} \tag{4.22}$$

przy czym $\mathbf{V(a)}=[E(a_i\ a_j)]$ jest macierzą wariancji i kowariancji elementów wektora $\mathbf{a}$.

Jesli dla każdego k≠h=1,..,N zachodzi, ze $\pi_{kh}>0$, to nieobciążonym estymatorem kowariancji składowej macierzy $\mathbf{V}(\mathbf{t}_{HTS})$ jest dla i,j=1,..,m statystyka [por. np. Konijn (1973), s. 235]:

$$\overline{\mathrm{C}}\mathrm{ov}(t_{\mathrm{HTiS}},t_{\mathrm{HTjS}})=N^{-2}\sum_{k=1}^{N}\frac{y_{ik}y_{jk}}{\pi_k^2}\left(1-\pi_k\right)a_k+$$

$$+2N^{-2}\sum_{k>h=1}^{N}\sum^{N}\frac{y_{ik}y_{jk}}{\pi_k\pi_h}\frac{\left(\pi_{kh}-\pi_k\pi_h\right)}{\pi_{kh}}a_ka_h \tag{4.23}$$

Drugi estymator jest rozszerzeniem na przypadek kowariancji określonego wzorem (4.5) estymatora wariancji Yatesa-Grundy'ego. Dla i,j=1,..,m ma on postać:

$$\overline{\mathrm{C}}\mathrm{ov}(t_{\mathrm{HTiS}},t_{\mathrm{HTjS}})=\sum_{k>h=1}^{N}\sum^{N}\frac{\left(\pi_{kh}-\pi_k\pi_h\right)}{\pi_{kh}}\left(\frac{y_{ik}}{\pi_k}-\frac{y_{ih}}{\pi_h}\right)\left(\frac{y_{jk}}{\pi_k}-\frac{y_{jh}}{\pi_h}\right)a_ka_h \tag{4.24}$$

Załóżmy teraz, że próba jest losowana bezzwrotnie. Przez $\bar{\mathbf{t}}_{\mathrm{HTS}}=[\bar{t}_{\mathrm{HT1S}}\ldots\bar{t}_{\mathrm{HTmS}}]$ oznaczamy wektor estymatorów średnich $\bar{\mathbf{y}}$, gdzie:

$$\bar{t}_{\mathrm{HTiS}}=\frac{1}{Nn}\sum_{k=1}^{N}\frac{y_{ik}n_{kS}}{p_k} \tag{4.25}$$

gdzie, przypomnijmy, n_{kS} jest częstością pojawienia się k-tego elementu populacji w próbie, a p_k prawdopodobieństwem wyboru za każdym razem do próby k-tego elementu populacji.

Na podstawie wyników z uprzedniego paragrafu wnioskujemy, że wektor statystyk $\bar{\mathbf{t}}_{HTS}$ jest nieobciążonym estymatorem wektora $\bar{\mathbf{y}}$. Macierz wariancji i kowariancji tego wektora otrzymujemy na podstawie wyrażenia (4.19), w którym kowariancje i wariancje składowych wektora $\mathbf{a}=[a_1 \ldots a_N]$ należy odpowiednio zastąpić przez kowariancje i wariancje elementów wektora częstości $\mathbf{n}_S=[n_{1S} \ldots n_{NS}]$, czyli $D^2(n_{kS})=np_k(1-p_k)$, $Cov(n_{kS},n_{hS})=-np_k p_h$ $k\neq h=1,\ldots,N$. Podobnie otrzymujemy macierz $\mathbf{V}(\bar{\mathbf{t}}_{HTS})$ zastępując we wzorze (4.22) macierz $\mathbf{V}(\mathbf{a})$ przez $\mathbf{V}(\mathbf{n}_S)=[Cov(n_{kS},n_{hS})]$.

Nieobciążonym estymatorem kowariancji dla $i\neq j=1,\ldots,m$ jest statystyka:

$$\overline{\mathrm{Cov}}\left(t_{HTiS}, t_{HTjS}\right)=\frac{1}{n(n-1)N^2}\sum_{k=1}^{N}\left(\frac{y_{ik}}{\pi_k}-\bar{t}_{HTiS}\right)\left(\frac{y_{jk}}{\pi_k}-\bar{t}_{HTS}\right)n_{kS} \tag{4.26}$$

która da się sprowadzić do postaci:

$$\overline{\mathrm{Cov}}\left(t_{HTiS}, t_{HTjS}\right)=\frac{1}{(Nn)^2(n-1)}\sum_{k>}^{N}\sum_{h=1}^{N}\left(\frac{y_{ik}}{\pi_k}-\frac{y_{ih}}{\pi_h}\right)\left(\frac{y_{jk}}{\pi_k}-\frac{y_{jh}}{\pi_h}\right)n_{kS}n_{hS} \tag{4.27}$$

Jeśli próba jest pobierana z nadpopulacji, wówczas macierz błędów średniokwadratowych predykcji wektora $\overline{\mathbf{Y}}$ prowadzonej przy pomocy wektorów $\mathbf{T}_{HTS}$, $\overline{\mathbf{T}}_{HTS}$ otrzymujemy wyznaczając wartość oczekiwaną odpowiednio macierzy $\mathbf{V}(\mathbf{T}_{HTS})$, $\mathbf{V}(\overline{\mathbf{T}}_{HTS})$ po rozkładzie prawdopodobieństwa, który określa przyjęty model nadpopulacji.

4.4. Parametry przybliżone rozkładu wektora estymatorów przy wybranych planach losowania próby

W niniejszym paragrafie wyznaczamy przybliżone wariancje i kowariancje określonego w uprzednim paragrafie wektora estymatorów t_{HTS} dla wybranych planów losowania próby zależnych od wartości zmiennych pomocniczych. Otrzymane tą drogą charakterystyki estymatorów zależą w różny sposób od parametrów populacji. W szczególności na podstawie wyprowadzanych wariancji estymatora Horvitza-Thompsona wnioskujemy, że dla danego planu losowania próby wariancja ta przyjmuje minimalną wartość, dla specyficznych własności populacji reprezentowanych momentami obserwowanych w niej zmiennych. W konsekwencji ułatwia to dla spodziewanego układu momentów zmiennych w populacji wybrać optymalny plan losowania próby.

4.4.1. Aproksymacja kowariancji estymatorów

Załóżmy, że próba jest losowana bezzwrotnie. Kowariancję daną wzorem (4.20) rozwijamy w szereg Taylora względem argumentów π_k i π_{kh} $(k\neq h=1,K,N)$ wokół punktów $\pi_k=\frac{n}{N}$ i $\pi_{kh}=\frac{n(n-1)}{N(N-1)}$. Wówczas pierwszy wyraz tego rozwinięcia wynosi $\frac{N-n}{Nn}c_{*ij}$, ponieważ przyjęte wartości prawdopodobieństw doboru elementów do próby odpowiadają schematowi bezzwrotnego losowania próby prostej. Po to, by wyznaczyć następny wyraz rozwinięcia, obliczamy pierwsze pochodne cząstkowe:

$$\frac{\partial \mathrm{Cov}\left(t_{HT\mathrm{iS}}\ ,\ t_{HT\mathrm{jS}}\right)}{\partial \pi_k}=\frac{1}{N^2}\frac{y_{ik}y_{jk}}{\pi_k^2}-\frac{1}{N^2}\sum_{k\neq h}^{N}\left(y_{ik}y_{jh}+y_{ih}y_{jk}\right)\frac{\pi_{kh}}{\pi_k^2\pi_h}$$

$$\frac{\partial \mathrm{Cov}\left(t_{HT\mathrm{iS}}\ ,\ t_{HT\mathrm{jS}}\right)}{\partial \pi_{kh}}=\frac{1}{N^2}\sum_{k\neq h=1}^{N}\sum^{N}\frac{y_{ik}y_{jh}}{\pi_k\pi_h}$$

Podczas wyliczania tych pochodnych było wygodnie kowariancję sprowadzić do postaci:

$$\mathrm{Cov}\left(t_{HT\mathrm{iS}},t_{HT\mathrm{jS}}\right)=\frac{1}{N^2}\left(\sum_{k=1}^{N}\frac{y_{ik}y_{jh}}{\pi_k}+\sum_{k\neq h=1}\sum y_{ik}y_{jh}\frac{\pi_{kh}}{\pi_k\pi_h}\right)-\bar{y}_i\bar{y}_j$$

Na podstawie wyliczonych pochodnych rozwijamy kowariancję w szereg Taylora z dokładnością do jego dwóch pierwszych składników:

$$\mathrm{Cov}\left(t_{HT\mathrm{iS}},t_{HT\mathrm{jS}}\right)=\frac{N-n}{Nn}c_{*ij}-\frac{1}{n^2}\sum_{k=1}^{N}y_{ik}y_{jk}\left(\pi_k-\frac{n}{N}\right)-\frac{2}{n^2}\sum_{k\neq h=1}\sum y_{ik}y_{jh}\left(\pi_k-\frac{n}{N}\right)+$$

$$+\frac{1}{n^2}\sum_{k\neq h=1}\sum y_{ik}y_{jh}\left(\pi_k-\frac{n}{N}\right)\left(\pi_{kh}-\frac{n(n-1)}{N(N-1)}\right) \qquad (4.28)$$

Załóżmy, że dla każdego $k\neq h=1,\ldots N$ $\pi_k=O(N^{-1})$ i $\pi_{kh}=O(N^{-2})$. Wtedy można wykazać, że dwa ostatnie składniki wzoru (4.28) są rzędu nie wyższego od $O(N^{-1})$. Przy dużej liczebności N populacji składniki te powinny mieć mały wpływ na wartość kowariancji i dlatego pomijamy je w dalszej analizie. Korzystając z równości $y_{ik}y_{jk}=\left[\left(y_{ik}-\bar{y}_i\right)+\bar{y}_i\right]\left[\left(y_{jk}-\bar{y}_j\right)+\bar{y}_j\right]$ tak przekształcamy drugi z kolei składnik wyrażenia (4.28), że

$$\mathrm{Cov}\left(t_{HT\mathrm{iS}}, t_{HT\mathrm{jS}}\right) = \frac{N-n}{Nn} c_{*\mathrm{ij}} - \frac{1}{n^2}\left(\sum_{k=1}^{N} \left(y_{\mathrm{ik}} - \bar{y}_{\mathrm{i}}\right)\left(y_{\mathrm{jk}} - \bar{y}_{\mathrm{j}}\right)\left(\pi_{\mathrm{k}} - \frac{n}{N}\right) + \bar{y}_{\mathrm{i}}\bar{y}_{\mathrm{j}} \sum_{k=1}^{N}\left(\pi_{\mathrm{k}} - \frac{n}{N}\right) + \right.$$

$$\left. + \bar{y}_{\mathrm{i}} \sum_{k=1}^{N}\left(y_{\mathrm{jk}} - \bar{y}_{\mathrm{j}}\right)\left(\pi_{\mathrm{k}} - \frac{n}{N}\right) + \bar{y}_{\mathrm{j}} \sum_{k=1}^{N}\left(y_{\mathrm{ik}} - \bar{y}_{\mathrm{i}}\right)\left(\pi_{\mathrm{k}} - \frac{n}{N}\right)\right) + O\left(N^{-1}\right) \qquad (4.29)$$

przy czym i,j=1,…,m. Stąd wnioskujemy, że w szczególności dla i=j=1,…,m zachodzi:

$$D^2\left(t_{HT\mathrm{iS}}\right) = \frac{N-n}{Nn} v_{*\mathrm{i}} - \frac{1}{n^2}\left(\sum_{k=1}^{N}\left(y_{\mathrm{ik}} - \bar{y}_{\mathrm{i}}\right)^2\left(\pi_{\mathrm{k}} - \frac{n}{N}\right) + \frac{1}{\bar{y}_{\mathrm{i}}^2} \sum_{k=1}^{N}\left(\pi_{\mathrm{k}} - \frac{n}{N}\right) + \right.$$

$$\left. + 2\bar{y}_{\mathrm{i}} \sum_{k=1}^{N}\left(y_{\mathrm{ik}} - \bar{y}_{\mathrm{i}}\right)\left(\pi_{\mathrm{k}} - \frac{n}{N}\right)\right) + O(N^{-1}) \qquad (4.30)$$

4.4.2. Plany losowania prób zależne od sumy wartości zmiennych pomocniczych

W pierwszym rozdziale opisano własności planu $P_6(\underline{S})$ losowania próby uporządkowanej. Prawdopodobieństwo wylosowania próby $\underline{S}$ jest proporcjonalne do obserwowanej w niej sumy wartości dodatniej zmiennej pomocniczej. Z kolei prawdopodobieństwo wylosowania pojedynczego elementu populacji bądź pary do próby określają wzory (1.44) i (1.46). Podstawiając je do wzoru (4.29) otrzymujemy:

$$\mathrm{Cov}(t_{HT\mathrm{Si}}, t_{HT\mathrm{Sj}}, P_6) = \frac{N-n}{Nn}\left\{ c_*\left(y_i, y_{\mathrm{j}}\right) - \frac{1}{n\bar{x}}\left(c_{*111}\left(y_i, y_j, x\right) + \right.\right.$$

$$\left.\left. + \bar{y}_{\mathrm{i}} c_*\left(y_j, x\right) + \bar{y}_{\mathrm{j}} c_*\left(y_i, x\right)\right)\right\} + 0(N^{-1}), \qquad (4.31)$$

przy czym i,j=1,..,m.

W szczególności wariancja estymatora t_{HTS} składowego wektora $\mathbf{t}_{\mathrm{HTS}}$ ma postać[32] :

$$D^2\left(t_{HT\mathrm{S}}, P_6\right) \approx \frac{N-n}{Nn}\left\{ v_*\left(y\right) - \frac{1}{n\bar{x}}\left(c_{*21}\left(y, x\right) + 2\bar{y} c_*\left(y, x\right)\right)\right\} \qquad (4.32)$$

która jest równoważna wyrażeniu:

[32] Por. Wywiał (1993).

$$D^2\left(t_{HTS}, P_6\right) \approx \frac{N-n}{Nn} v_*(y)\left\{1-\frac{1}{n}\gamma(x)\left(\eta_{21}(y,x)+2\gamma^{-1}(y)\rho(y,x)\right)\right\} \qquad (4.33)$$

przy czym $\gamma(.)$, jest współczynnikiem zmienności, $\eta(.)$ momentem centralnym zestandaryzowanym, a $\rho(.)$ współczynnikiem korelacji liniowej.

Na podstawie otrzymanego wyrażenia wnioskujemy, że wraz ze wzrostem liczebności próby wariancja strategii estymacji (t_{HTS}, P_6) zmierza do wariancji średniej z próby prostej losowanej bezzwrotnie. Wariancja strategii (t_{HTS}, P_6) maleje, jeśli współczynnik zmienności cechy pomocniczej rośnie przy ustalonych pozostałych parametrach i przy nie zmieniającej się liczebności próby. Podobnie wariancja ta maleje, gdy moduł wartości współczynnika korelacji $\rho(x,y)$ rośnie i iloczyn $\gamma(y)\rho(y,x)>0$. Interesującym jest wpływ trzeciego mieszanego momentu centralnego zestandaryzowanego $\eta_{21}(y,x)$ na wartość wariancji. Wywiał (1983) wykazał, że wraz ze wzrostem modułu z wartości tego momentu rośnie także ścisłość zależności w sensie regresji liniowej drugiego rodzaju między kwadratami odchyleń $(y_k-\bar{y})^2$ a różnicami $(x_k-\bar{x})$.

Wniosek 4.1: Strategia estymacji (t_{HTS}, P_6) średniej w populacji $\bar{y}$ jest lepsza od strategii $(\bar{y}_S, P_2)$, nazywanej krótko średnią z próby prostej losowanej bezzwrotnie, gdy drugi składnik znajdujący się w nawiasie wąsiastym wyrażenia (4.33) jest dodatni.

Przypomnijmy, ze określony wzorami (1.48) lub (1.49) plan P_7 preferuje podczas losowania te próby, w których obserwowana suma wartości zmiennej pomocniczej $n\bar{x}_S$ jest mała w stosunku do globalnej wartości tej cechy w populacji. Na podstawie wzorów (1.51) i (4.30) wyznaczamy przybliżoną wartość kowariancji między estymatorami t_{HTiS} oraz t_{HTjS} (i,j=1,..,m):

$$\mathrm{Cov}(t_{HTiS}, t_{HTjS}, P_7) \approx \frac{N-n}{Nn} c_*\left(y_i, y_j\right) +$$

$$+\frac{1}{nN}\gamma(x)\left(c_{*111}\left(y_i, y_j, x\right)+\bar{y}_i c_*\left(y_j, x\right)+\bar{y}_j c_*\left(y_i, x\right)\right) \qquad (4.34)$$

W szczególności wariancja ma postać:

$$D^2(t_{HTS}, P_7) = \frac{N-n}{Nn} v_*(y) + \frac{1}{nN} v_*(y)\gamma(x)\left(\eta_{21}(y,x)+2\gamma^{-1}(y)\rho(y,x)\right) \qquad (4.35)$$

Interpretacja otrzymanego wyniku jest podobna do wniosków wysnutych ze wzoru (4.33). Dodajmy tylko, że tym razem ujemna zależność liniowa między zmiennymi x i $(y-\bar{y})^2$ przyczynia się do spadku wariancji estymatora t_{HTiS}.

Załóżmy teraz, że próba pochodzi z nadpopulacji, w której cechy (x,y) mają dwuwymiarowy rozkład normalny. Korzystajac ze wzorów (4.33) i (4.35) wyznaczamy błędy średniokwadratowe strategii predykcji (T_{HTS}, P_6) i (T_{HTS}, P_7) dla wartości średniej μ w nadpopulacji. Korzystając z faktu, że dla dwuwymiarowego rozkładu normalnego $\mathcal{E}\eta_{21}(y,x)=0$, po odpowiednich obliczeniach otrzymujemy:

$$\mathcal{E}D^2(T_{HTS},P_6) \approx \frac{N-n}{Nn}\sigma^2(y)\left(1-\frac{2}{n}\gamma(x)\gamma^{-1}(y)\rho(y,x)\right) \tag{4.36}$$

$$\mathcal{E}D^2(T_{HTS},P_7) = \frac{N-n}{Nn}\sigma^2(y)+\frac{2}{nN}\sigma^2(y)\gamma(x)\gamma^{-1}(y)\rho(y,x) \tag{4.37}$$

Wniosek 4.2: Przyjmujemy, że $\gamma(x)>0$ i $\gamma(y)>0$. Wówczas, gdy $\rho(y,x)>0$, to efektywniejszą od strategii predykcji $(\overline{Y}_S, P_2)$ [będącej średnią z próby prostej losowanej bezzwrotnie] jest strategia (T_{HTS}, P_6). Z kolei gdy $\rho(y,x)<0$, to lepszą od $(\overline{Y}_S, P_2)$ jest strategia (T_{HTS}, P_7). Zatem obie strategie (T_{HTS}, P_6) i (T_{HTS}, P_7) uzupełniają się.

4.4.3. Plany losowania prób zależne od sumy kwadratów odchyleń cechy pomocniczej od jej średniej w populacji

W pierwszym rozdziale skonstruowano dwa plany losowania próby P_8 i P_9. Pierwszy z nich, dany wzorem (1.55), preferuje wybór tej próby, której odpowiada duża obserwowana w próbie suma kwadratów odchyleń zmiennej pomocniczej od jej średniej w populacji $\bar{x}$. W przeciwieństwie drugi plan P_9 losowania, określony wzorem (1.57), daje większe szanse wylosowania takiej próby, dla której ta suma kwadratów odchyleń jest mała w stosunku do sumy kwadratów tych odchyleń w populacji.

W przypadku gdy próba jest losowana zgodnie z planem P_8, na podstawie wzorów (1.56) i (4.30) mamy:

$$D^2\left(t_{HTS},P_8\right) \approx \frac{N-n}{Nn}v_*(y)\left(1-\frac{A}{n}\right) \tag{4.38}$$

gdzie:

$$A = \eta_{22}(y,x)-1+2\gamma^{-1}(y)\eta_{12}(y,x) \tag{4.39}$$

Gdy próba jest losowana zgodnie z określonym wzorem (1.57) planem P_9, to

$$D^2\left(t_{HTS},P_9\right) \approx v_*(y)\left(\frac{N-n}{Nn}+\frac{A}{nN}\right) \tag{4.40}$$

Gdy A>0, to $D^2\left(t_{HTS},P_8\right) < D^2\left(\overline{Y},P_2\right) = \frac{N-n}{Nn}v_*(y)$. Z kolei jeśli A<0, to $D^2\left(t_{HTS},P_9\right) < D^2\left(\overline{Y},P_2\right)$. Zatem strategie $\left(t_{HTS},P_8\right)$ i $\left(t_{HTS},P_9\right)$ uzupełniają się.

Załóżmy, że próba jest dobierana z nadpopulacji opisanej modelem normalnym. Wówczas można wykazać[33], że błędy średniokwadratowe predykcji średniej $\overline{Y}$ w populacji są postaci:

$$\mathcal{E}D^2\left(t_{HTS}, P_8\right) \approx \frac{N-n}{Nn}\sigma(y)\left(1-\frac{2}{n}\rho^2(y,x)\right) \tag{4.41}$$

$$\mathcal{E}D^2\left(t_{HTS}, P_9\right) \approx \sigma^2(y)\left(\frac{N-n}{Nn}+\frac{2}{nN}\rho^2(y,x)\right) \tag{4.42}$$

Gdy rozmiar populacji jest nieograniczony, czyli jeśli $N \to \infty$, to

$$\mathcal{E}D^2\left(t_{HTS}, P_8\right) \approx \frac{1}{n}\sigma(y)\left(1-\frac{2}{n}\rho^2(y,x)\right) \tag{4.43}$$

$$\mathcal{E}D^2\left(t_{HTS}, P_9\right) \approx \frac{1}{n}\sigma^2(y) = \mathcal{E}D^2\left(\overline{Y}_S, P_2\right) \tag{4.44}$$

Stąd wnioskujemy, że w przypadku predykcji wartości średniej w nadpopulacji opisanej prostym modelem normalnym strategia predykcji $\left(t_{HTS}, P_8\right)$ jest lepsza od klasycznej strategii $\left(\overline{Y}_S, P_2\right)$ gdy $\rho(y,x)\neq 0$. Z kolei średnia z próby $\left(\overline{Y}_S, P_2\right)$ dla $\rho(y,x)\neq 0$ jest dokładniejsza od strategii $\left(t_{HTS}, P_9\right)$ i dopiero w populacji o nieograniczonym rozmiarze obie te strategie są równie dokładne.

Wniosek 4.3: Spośród trzech strategii predykcji: $\left(\overline{Y}_S, P_2\right)$, $\left(t_{HTS}, P_9\right)$ i $\left(t_{HTS}, P_8\right)$ należy preferować ostatnią.

4.4.4. Plany losowania zależne od wariancji zmiennej pomocniczej w próbie

W rozdziale 1 konstruowano plany losowania prób P_{10} i P_{11}, które określają odpowiednio wzory (1.58) i (1.62). Pierwszy z nich jest proporcjonalny do wariancji cechy pomocniczej w próbie, a drugi preferuje próby o małej wariancji cechy wspomagającej w próbie w stosunku do wariancji tej cechy w populacji.

Na podstawie równań (1.60) i (4.30) otrzymujemy:

---

[33] Podczas dowodu wykorzystujemy znaną równość $\eta_{22}(y,x)=1+2\rho^2(y,x)$, która zachodzi dla przypadku dwuwymiarowego rozkładu normalnego, por. np. Kendall i Stuart (1966), s. 133.

$$D^2(t_{HTS}, P_{10}) \approx \frac{N-n}{Nn} v_*(y)\left(1-\frac{(N-1)A}{n(N-2)}\right) \tag{4.45}$$

przy czym wielkość A określa wyrażenie (4.39). Z kolei z równania (4.30) i wzoru (1.64) wynika:

$$D^2(t_{HTS}, P_{11}) = v_*(y)\left(\frac{N-n}{Nn}+\frac{(n-1)(N-1)}{n^2N(N-2)}A\right) \tag{4.46}$$

Otrzymane wyrażenia są zbliżone do wariancji odpowiednich strategii estymacji rozważanych w uprzednim podpunkcie. Podobne również wyciągamy wnioski.

Gdy próba jest losowana z nadpopulacji opisanej prostym modelem normalnym, to błędy średniokwadratowe predykcji przeciętnej $\overline{Y}$ są postaci:

$$\mathcal{E}D^2(T_{HTS}, P_{10}) \approx \frac{N-n}{Nn}\sigma^2(y)\left(1-\frac{2(N-1)}{n(N-2)}\rho^2(y,x)\right) \tag{4.47}$$

$$\mathcal{E}D^2(T_{HTS}, P_{11}) = \sigma^2(y)\left(\frac{N-n}{Nn}+\frac{2(n-1)(N-1)}{n^2(N-2)N}\rho^2(y,x)\right) \tag{4.48}$$

Przy $N\to\infty$ parametry $\mathcal{E}D^2(T_{HTS},P_{10})$ i $\mathcal{E}D^2(T_{HTS},P_{11})$ są równe prawym stronom odpowiednio równań (4.43) i (4.44).

Wniosek 4.4: W przypadku gdy próba jest losowana z nadpopulacji normalnej ze współczynnikiem korelacji $\rho(y,x)\neq 0$, to spośród planów losowania próby P_{10} , P_{11} i P_2 należy użyć pierwszego z wymienionych, ponieważ w przybliżeniu ma najmniejszą wartość błędu średniokwadratowego predykcji przeciętnej $\overline{Y}$.

4.4.5. Plany losowania zależne od kwadratu błędu szacunku średniej zmiennej pomocniczej w populacji

Rozważmy teraz plan losowania próby P_{12} określony wzorem (1.67). Podstawiając odpowiednio wyrażenie (1.72) do wzoru (4.30) otrzymujemy:

$$D^2(t_{HTS}, P_{12}) = v_*(y)\left(\frac{N-n}{Nn}-\frac{(N-2n)(N-1)}{n^2N(N-2)}A\right) \tag{4.49}$$

przy czym wielkość A jest wyjaśniana przez wzór (4.39). Stąd wnioskujemy, że strategia estymacji (t_{HTS}, P_{12}) jest lepsza od średniej z próby prostej losowanej bezzwrotnie $(\bar{y}_S, P_2)$, gdy $n<\frac{N}{2}$ i $A > 0$, bądź gdy $n>\frac{N}{2}$ i $A < 0$.

Gdy próba jest losowana z nadpopulacji o rozkładzie normalnym, to błąd średniokwadratowy predykcji przeciętnej $\overline{Y}$ prowadzonej na podstawie strategii (T_{HTS}, P_{12}) ma postać:

$$\mathcal{E}D^2(T_{HTS}, P_{12}) \approx \sigma^2(y)\left(\frac{N-n}{Nn} - \frac{2(N-2n)(N-1)}{n^2N(N-2)}\rho^2(x,y)\right) \tag{4.50}$$

a gdy $N \to \infty$, to przybliżoną wartość błędu średniokwadratowego określa wzór (4.43).

Na podstawie wzorów (4.30) i 1.77) wyliczymy przybliżony poziom wariancji strategii (T_{HTS}, P_{13}), gdzie plan P_{13} określa wzór (1.74). Po odpowiednich obliczeniach mamy:

$$D^2(t_{HTS}, P_{13}) \approx \frac{N-n}{Nn} v_*(y) + \frac{N-1}{Nn^2} v_*(y)(A+2) \tag{4.51}$$

gdzie A wyjaśnia wzór (4.39).

Gdy próba pochodzi z nadpopulacji o dwuwymiarowym rozkładzie normalnym, to błąd średniokwadratowy strategii (t_{HTS}, P_{13}) w przybliżeniu wynosi:

$$\mathcal{E}D^2(t_{HTS}, P_{13}) \approx \frac{N-n}{Nn}\sigma^2(y) + 2n^{-2}\left[1+\rho^2(x,y)\right]\sigma^2 \tag{4.52}$$

Stąd wynika następująca konkluzja:

Wniosek 4.5: Gdy próba jest losowana z nadpopulacji o dwuwymiarowym rozkładzie normalnym i gdy $\rho(x,y) \neq 0$, to strategia predykcji (t_{HTS}, P_{12}) jest lepsza od średniej z próby prostej $(\overline{Y}, P_2)$, a ta z kolei jest lepsza od strategii (t_{HTS}, P_{13}).

4.5. Predykcja średniej w modelu nadpopulacji z zależnymi zmiennymi

Przyjmijmy następujące założenia modelu nadpopulacji:

$Y_k=\mu+Z_k$ i $\mathcal{E}(Z_k)=0$ i $\mathcal{D}^2(Z_k)=\sigma^2$ oraz $\mathcal{E}(Z_kZ_l)=\sigma^2\rho_{kl}$ dla $k\neq l=1,...,N$

Przyjmując, że $\mathbf{Y}^T=[Y_1,...Y_N]$ i $\mathbf{Z}=[Z_1...Z_N]$, mamy:

$$\mathbf{Y}=\mu\mathbf{Z}, \quad \mathcal{E}(\mathbf{Z})=0, \quad \mathcal{D}^2(\mathbf{Z})=\sigma^2\mathbf{V}$$

Prawdopodobieństwa inkluzji rzędu pierwszego oznaczmy wektorem $\boldsymbol{\pi}^T=[\pi_1...\pi_N]$. Z kolei niech prawdopodobieństwa inkluzji rzędu drugiego π_{kl} będą elementami spoza głównej

przekątnej macierzy $\mathbf{\Gamma}=[\pi_{kl}]$, przy czym niech $\pi_{kk}=\pi_k$. Ponadto niech $\mathbf{a}^T=[a_1...a_N]$ oraz $E(a_k)=\pi_k$, $E(a_k a_l)=\pi_{kl}$.

Zadanie polega na predykcji średniej w populacji:

$$\overline{Y}=\frac{1}{N}\sum_{k=1}^{N}Y_k$$

Do tego celu adaptujemy znaną statystykę Horvitza-Thompsona:

$$T_{HTS}=\frac{1}{N}\sum_{k=1}^{N}\frac{Y_k}{\pi_k}a_k \tag{4.53}$$

którą w sposób macierzowy można zapisać następująco:

$$T_{HTS}=\frac{1}{N}\mathbf{Y}^T\mathrm{diag}^{-1}(\pi)\mathbf{a} \tag{4.54}$$

Można wykazać [por. np. Cassel i in. (1977)], że:

$$E(T_{THS})=\overline{Y} \tag{4.55}$$

$$E(T_{THS})=\frac{\mu}{N}\mathbf{J}_N^T\mathrm{diag}^{-1}(\pi)\mathbf{a} \tag{4.56}$$

$$EE(T_{HTS})=\mu \tag{4.57}$$

przy czym $\mathbf{J}_N$ jest N-elementową kolumną jedynkową.

Zatem statystyka T_{HTS} jest p-nieobciążonym oraz p-ξ nieobciążonym predyktorem parametrów odpowiednio $\overline{Y}$ i μ.

Wariancję statystyki T_{HTS} można zapisać następująco:

$$D^2(T_{HTS})=\frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\Delta_{kl}\frac{Y_kY_l}{\pi_k\pi_l} \tag{4.58}$$

gdzie:

$$\Delta_{kl}=\begin{cases}\pi_{kl}-\pi_k\pi_l & \text{dla} \quad k\neq l\\ \pi_k(1-\pi_k) & \text{dla} \quad k=l\end{cases} \tag{4.59}$$

Niech $\mathbf{\Delta}=[\Delta_{kl}]=\mathbf{\Gamma}-\boldsymbol{\pi}\boldsymbol{\pi}^T$ oraz

$$\mathbf{B}=\mathrm{diag}^{-1}(\boldsymbol{\pi})\mathbf{\Delta}\mathrm{diag}^{-1}(\boldsymbol{\pi}) \tag{4.60}$$

Wtedy w zapisie macierzowym:

$$D^2(T_{HTS})=N^{-2}\mathbf{Y}^T\mathbf{B}\mathbf{Y} \tag{4.61}$$

Pozwala to sprawniej wyprowadzić wartość oczekiwaną:

$$\mathcal{E}D^2(T_{HTS})=N^{-2}\mathcal{E}(\mathbf{Y}^T\mathbf{B}\mathbf{Y})=N^{-2}tr\mathbf{B}\mathcal{E}(\mathbf{Y}\mathbf{Y}^T)$$

$$\mathcal{E}(\mathbf{Y}\mathbf{Y}^T)=\mathcal{E}(\mu\mathbf{J}_N+\mathbf{Z})(\mu\mathbf{J}_N+\mathbf{Z})^T=\mu^2\mathbf{J}_N(\mathbf{J}_N)^T+\sigma^2\mathbf{V}$$

Zatem:

$$\mathcal{E}D^2(T_{HTS})=N^{-2}(\mu^2(\mathbf{J}_N)^T\mathbf{B}\mathbf{J}_N)+\sigma^2tr\mathbf{B}\mathbf{V} \tag{4.62}$$

gdzie:

$$\mathbf{J}_N^T\mathbf{B}\mathbf{J}_N=\sum_{k=l}^{N}\sum_{l=1}^{N}\left(\frac{1}{\pi_k}-\frac{1}{\pi_l}\right)^2\left(\pi_k\pi_l-\pi_{kl}\right) \tag{4.63}$$

Jeśli losujemy bezzwrotnie próbę prostą, to wiadomo, że

$$\pi_k=\frac{n}{N} \quad i \quad \pi_{kl}=\frac{n(n-1)}{N(N-1)}, \quad k\neq l=1,...,N$$

Wówczas wyliczamy, że elementy macierzy $\mathbf{B}=[b_{ij}]$ są postaci:

$$b_{ij}=\begin{cases}\dfrac{N-n}{n} & \text{dla} \quad i=j\\[2ex] -\dfrac{N-n}{n(N-1)} & \text{dla} \quad i\neq j\end{cases}$$

Ten wynik pozwala macierz **B** zapisać następująco:

$$\mathbf{B}=\frac{N(N-n)}{n(N-1)}\mathbf{M}, \qquad \mathbf{M}=\mathbf{I}_N-\frac{1}{N}\mathbf{J}_N\mathbf{J}_N^T \tag{4.64}$$

przy czym $\mathbf{I}_N$ jest macierzą jednostkową stopnia N. Wtedy można wykazać, że $(\mathbf{J}_N)^T\mathbf{B}\mathbf{J}_N=0$ oraz na podstawie wzoru (4.62), że

$$\mathcal{E}D^2(T_{HTS})=\frac{\sigma^2(N-n)}{Nn(n-1)}\left(tr\mathbf{V}-\frac{1}{N}\mathbf{J}_N^T\mathbf{V}\mathbf{J}_N\right) \tag{4.65}$$

W szczególności, jeśli $\mathbf{V}=\mathbf{I}_N$, to

$$\mathcal{E}D^2(T_{HTS}) = \frac{N-n}{Nn}\sigma^2 \tag{4.66}$$

Niech:

$$\mathbf{V}=\begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ . & . & . & . & . \\ \rho & \rho & \rho & \ldots & 1 \end{bmatrix} \tag{4.67}$$

Wtedy:

$$\mathcal{E}D^2(T_{HTS}) = \frac{N-n}{Nn}\sigma^2(1-\rho) \tag{4.68}$$

Stąd i z faktu, że $-(N-1)^{-1}<\rho<1$ [por. Rao (1982)], wynika, że wzrost wartości ρ przyczynia się do wzrostu precyzji predykcji.

Niech:

$$\mathbf{V}=\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \ldots & \rho^{N-2} & \rho^{N-1} \\ \rho & 1 & \rho & \rho^2 & \ldots & \rho^{N-3} & \rho^{N-2} \\ \rho^2 & \rho & 1 & \rho & \ldots & \rho^{N-4} & \rho^{N-3} \\ . & . & . & . & \ldots & . & . \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \rho^{N-4} & \ldots & 1 & \end{bmatrix}$$

Wtedy mamy:

$$\mathsf{E}D^2(T_{HTS}) = \frac{N-n}{Nn}\sigma^2\left(1-\frac{2}{N(N-1)}\sum_{k=1}^{N-1}(N-k)\rho^k\right) \tag{4.69}$$

Stad wynika, że jeśli współczynnik autokorelacji ρ jest dodatni i rośnie, to precyzja predykcji również zwiększa się.

5. WEKTOROWY ESTYMATOR REGRESYJNY

W niniejszym rozdziale uwagę koncentrujemy na własnościach wektora znanych estymatorów regresyjnych. Ten sposób oceny parametrów populacji wykorzystuje dostępne informacje o wartościach tzw. zmiennych dodatkowych. Przyczynia się to do podniesienia dokładności estymacji, jeśli cechy dodatkowe są dostatecznie silnie skorelowane ze zmiennymi badanymi.

W badaniach próbkowych oprócz estymatorów regresyjnych są polecane także i inne estymatory wykorzystujące własności cech dodatkowych. Chodzi tu o estymatory ilorazowe oraz iloczynowe, o których podstawowe informacje znajdziemy w każdym podręczniku z metody reprezentacyjnej i dlatego pomijamy tu ich prezentację.

Dodajmy, że ciekawe ich własności analizowali m.in. Herzel (1989), Murthy (1964), Rao i Mudholkar (1967), Sengupta (1981), Shah i Gupta (1987), Shukla (1976), Singh (1989), Srivastava (1983), Srivastava, Shukla i Bhatnagor (1981). Estymatorami tego typu zależnymi od wielu cech dodatkowych zajmowali się John (1969), Lynch (1978), Olkin (1958) i Tripathi (1976). Z kolei Wywiał (1992) przedstawił podstawowe własności wektora estymatorów ilorazowych i wektora estymatorów iloczynowych.

5.1. Wektorowy estymator różnicowy

Uogólnienie na przypadek wektorowy znanego estymatora różnicowego przebiega następująco. Niech **A** będzie macierzą rzeczywistą o wymiarze z×m, gdzie m jest liczbą szacowanych średnich tworzących elementy wektora $\bar{\mathbf{y}}=[\bar{y}_1 ... \bar{y}_m]$. Z kolei z oznacza ilość składowych wektora $\bar{\mathbf{x}}=[\bar{x}_1 ... \bar{x}_z]$ średnich cech pomocniczych. W końcu przez $\bar{y}_S$ i $\bar{\mathbf{x}}_S$ oznaczamy wektory średnich z próby prostej losowanej bezzwrotnie. Wtedy wektorowy estymator różnicowy ma postać:

$$\mathbf{t}_{AS} = \bar{\mathbf{y}}_S + (\bar{\mathbf{x}}_S - \bar{\mathbf{x}})\mathbf{A} \tag{5.1}$$

Estymator $\mathbf{t}_{AS}$ daje nieobciążone oceny wektora średnich $\bar{\mathbf{y}}$, a jego macierz wariancji i kowariancji ma postać:

$$\mathbf{V}(\mathbf{t}_{AS}) = \frac{N-n}{Nn}(\mathbf{C}_{*yy} + \mathbf{C}_{*yx}\mathbf{A} + \mathbf{A}^T\mathbf{C}_{*xy} + \mathbf{A}^T\mathbf{C}_{*xx}\mathbf{A}) \tag{5.2}$$

Stąd wynika, że bez dodatkowych założeń jest trudno porównywać precyzję estymatora $\mathbf{t}_{AS}$ z precyzją średniej z próby. Staje się to dopiero możliwe, gdy założymy, iż m=z=1. Wtedy estymator $\mathbf{t}_{AS}$ redukuje się do postaci:

$$t_{dS} = \bar{y}_S + b(\bar{x}_S - \bar{x}) \tag{5.3}$$

Jego wariancja jest następująca:

$$D^2(t_{dS}) = \frac{N-n}{Nn}\left(v_*(y) + 2b\sqrt{v_*(x)v_*(y)}r(x,y) - bv_*(x)\right) \tag{5.4}$$

Wtedy:

$$D^2(\bar{y}_S) - D^2(t_{dS}) = \frac{N-n}{Nn}\sqrt{v_*(x)v_*(y)}b\left(2r(x,y) - b\sqrt{\frac{v_*(x)}{v_*(y)}}\right)$$

Zatem przy planie P_3 losowania bezzwrotnego próby prostej estymator różnicowy t_{dS} jest lepszy od średniej $\bar{y}_S$, gdy $0<b<2r(x,y)\sqrt{\frac{v_*(y)}{v_*(x)}} = \frac{c(x,y)}{v^2(x)}$ i $r(x,y)>0$ bądź $2r(x,y)\sqrt{\frac{v_*(y)}{v_*(x)}} <b<0$ i $r(x,y)<0$

Analiza własności estymatora różnicowego t_{dS} uprości się, jeżeli założymy, że $|b| =1$. Wówczas po to, by oceniać wartość $\bar{y}$, trzeba dodatkowo znać tylko wartość średnią $\bar{x}$ zmiennej wspomagającej. Wtedy $D^2(\bar{y}_S) > D^2(t_{dS})$, gdy $r(x,y) < -\frac{1}{2}\sqrt{\frac{v_*(x)}{v_*(y)}}$ i b=1 bądź $r(x,y) > \frac{1}{2}\sqrt{\frac{v_*(x)}{v_*(y)}}$ i b=-1.

Wnioskujemy więc, że gdy |b|=1, estymator różnicowy t_{dS} należy stosować zamiast średniej $\bar{y}_S$ wtedy, gdy cechy *x* i *y* są silnie skorelowane oraz wariancja cechy pomocniczej jest mała w stosunku do wariancji zmiennej *y*.

W praktyce badań metodą reprezentacyjną nie zawsze można dysponować informacją o wartości średniej cechy pomocniczej w populacji. Wtedy do estymacji wektora $\bar{y}$ można użyć wektora estymatorów różnicowych z próby podwójnej, który ma postać[34] :

[34] Własności tego estymatora w przypadku jednowymiarowym studiowali m.in. Raj (1965) i Konijn (1973).

$$\bar{\mathbf{t}}_{AS} = \bar{\mathbf{y}}_{S_2} + \left(\bar{\mathbf{x}}_{S_2} - \bar{\mathbf{x}}_{S_1}\right)\mathbf{A} \tag{5.5}$$

przy czym $S = \{S_1, S_2\}$ jest próbą podwójną, której plan wyboru P_5 opisano w rozdziale 1. Przypomnijmy, że S_2 jest próbą prostą o liczebności n_2 losowaną bezzwrotnie z uprzednio już wylosowanej również bezzwrotnie próby prostej S_1 o rozmiarze $n_1 \geq n_2$. Wektor $\bar{t}_{AS}$ jest nieobciążonym estymatorem parametrów $\bar{\mathbf{y}}$, a jego macierz wariancji i kowariancji ma postać:

$$\mathbf{V}(\bar{\mathbf{t}}_{\mathbf{AS}}) = \mathrm{E}_{S_1}\left(\mathbf{V}_{S_2}\left(\bar{\mathbf{t}}_{\mathbf{AS}} | S_1 = s_1\right)\right) + \mathbf{V}_{S_1}\left(\mathrm{E}_{S_2}\left(\bar{\mathbf{t}}_{\mathbf{AS}} | S_1 = s_1\right)\right) =$$

$$= \frac{n_1 - n_2}{n_1 n_2}\left(\mathbf{C}_{*yy} + \mathbf{C}_{*yx}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{C}_{*xy} + \mathbf{A}^{\mathrm{T}}\mathbf{C}_{*xx}\mathbf{A}\right) + \frac{N - n_1}{N n_1}\mathbf{C}_{*yy}$$

Znaczenie praktyczne estymatora $\bar{t}_{AS}$ jest ograniczone, ponieważ decydując się na obserwacje cech w dostatecznie dużych próbach składowych próby podwójnej można znaleźć inne lepsze estymatory, które rozważamy dalej.

Na zakończenie sygnalizujemy, że Raj (1965) proponuje pewną modyfikację analizowanego tutaj estymatora różnicowego dla pojedynczej wartości średniej wyznaczanego z próby prostej pojedynczej bądź podwójnej. Proponowana przez niego statystyka jest kombinacją liniową estymatorów różnicowych średniej zmiennej *y* wyznaczanych względem kolejnych cech wspomagających. Zakładając, że współczynniki tej kombinacji sumują się do jedynki Raj wyznacza ich wartości optymalne. Ponadto wykazuje, że szczególnym przypadkiem jego estymatora jest wielowymiarowy estymator ilorazowy Olkina (1958).

5.2. Własności wektorowego estymatora regresyjnego

Zagadnieniem estymacji wartości średniej w populacji ustalonej za pomocą estymatorów regresyjnych zajmowali się m.in.: Bracha (1978, 1979, 1982, 1983), Cochran (1963), Greń (1969, 1970), Konijn (1962, 1973), Murthy (1977), Tripathi (1973) i Wywiał (1992).

Zastąpmy w wyrażeniu (5.2) macierz **A** przez

$$\mathbf{B} = -\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}, \qquad \det(\mathbf{C}_{xx}) > 0 \tag{5.6}$$

co prowadzi do wektora estytmatorów regresyjnych w postaci:

$$\mathbf{t}_{BS} = \bar{\mathbf{y}}_S + (\bar{\mathbf{x}}_S - \bar{\mathbf{x}})\mathbf{B} \tag{5.7}$$

Elementami wektora $\mathbf{t}_{BS}=[t_{B1S}...t_{BmS}]$ są znane w metodzie reprezentacyjnej estymatory regresyjne wyznaczane z próby prostej S losowanej bezzwrotnie. Wektor t_{BS} daje nieobciążone oceny wektora $\bar{\mathbf{y}}$, a jego macierz wariancji i kowariancji wynika ze wzorów (5.6) i (5.2):

$$\mathbf{V}(\mathbf{t}_{BS}) = \frac{N-n}{Nn}\left(\mathbf{C}_{*yy} - \mathbf{C}_{*yx}\mathbf{C}_{*xx}^{-1}\mathbf{C}_{*xy}\right) \tag{5.8}$$

Oznaczmy przez $\mathcal{A}$ zbiór wszystkich estymatorów różnicowych $\mathbf{t}_{AS}$ różniących się od siebie elementami macierzy **A**, której funkcją jest wektor estymatorów określony wzorem (5.1).

Twierdzenie 5.1 [Wywiał (1988)]: Jeśli próba prosta jest losowana bezzwrotnie, to estymator $\mathbf{t}_{BS}$ jest efektywny w klasie $\mathcal{A}$ estymatorów różnicowych.

Z twierdzenia 5.1 wynika, że indeks efektywności estymatora $\mathbf{t}_{BS}$ w klasie $\mathcal{A}$ spełnia nierówność: $e(\mathbf{t}_{BS}/\mathcal{A})\le 1$. W szczególności indeks efektywności względnej estymatora $\mathbf{t}_{BS}$ w stosunku do wektora średnich z próby prostej $\bar{\mathbf{y}}_S$ ma postać:

$$e_w\left(\mathbf{t}_{BS}/\bar{\mathbf{y}}_S\right)=\det\mathbf{V}(\mathbf{t}_{BS})\det\mathbf{V}^{-1}(\bar{\mathbf{y}}_S)=\frac{\det(\mathbf{C}_{yy}-\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy})}{\det(\mathbf{C}_{yy})}$$

Oznaczmy przez r_{qi} , i=1,...,m'=min(m,z), współczynniki korelacji kanonicznej mierzące zależność liniową między wektorami cech $x=[x_1...x_z]$ i $y=[y_1...y_m]$. Zależność między wektorami *y* i *x* zwiększa się wraz ze wzrostem wartości parametrów $r_{qi}^2 \le 1$ (i=1,...,m). Wielkość r_{qi}^2 jest otrzymywana jako i-ty pierwiastek charakterystyczny macierzy $\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}$ [por. np. Anderson (1958) i C.R.Rao (1982)]. Wtedy po odpowiednich przekształceniach algebraicznych otrzymujemy:

$$\det\mathbf{V}(\mathbf{t}_{BS}) = \det(\mathbf{C}_{*yy})\prod_{i=1}^{m'}(1-r_{qi}^2) \tag{5.9}$$

$$e_w(\mathbf{t}_{BS}/\bar{\mathbf{y}}_S) = \prod_{i=1}^{m'}(1-r_{qi}^2) \tag{5.10}$$

Wniosek 5.1: Dla próby prostej losowanej bezzwrotnie efektywność wektora $\mathbf{t}_{BS}$ względem wektora średnich $\bar{\mathbf{y}}_S$ rośnie wraz ze wzrostem zależności między wektorami zmiennych *y* i *x* mierzonej współczynnikami korelacji kanonicznej.

Dodajmy, że z twierdzeń 2.2 i 5.1 wynika, że $q(\mathbf{t}_{BS}) \le q(\bar{\mathbf{y}}_S)$ oraz $\rho(\mathbf{t}_{BS}) \le \rho(\bar{\mathbf{y}}_S)$.

Symbolem v_{*i} r_i^2 (i=1,..,m) oznaczmy diagonalne elementy macierzy $\mathbf{C}_{*yx}\mathbf{C}_{*xx}\mathbf{C}_{*xy}$, przy czym $0 \le r_i \le 1$ jest znanym w statystyce współczynnikiem korelacji wielorakiej mierzącej ścisłość zależności między cechą y_i oraz zmiennymi $x = [x_1...x_z]$ traktowanymi łącznie. Z kolei wielkość r_i^2 jest nazywana współczynnikiem determinacji, który określa, jaka część

zmienności zmiennej y_i została wyjaśniona przez cechy x za pośrednictwem funkcji regresji liniowej. Stąd na podstawie wzoru (5.8) i znanych własności śladu macierzy mamy:

$$q(\mathbf{t}_{BS}) = q(\bar{\mathbf{y}}_S)\sqrt{1-\bar{r}^2} \tag{5.11}$$

gdzie:

$$\bar{r}^2 = \sum_{i=1}^{m} g_i r_i^2 \tag{5.12}$$

$$g_i = v_i \left(\sum_{i=1}^{m} v_i \right)^{-1}$$

Przez $\bar{r}^2$ oznaczono średni ważony współczynnik determinacji, który spełnia nierówność $0 \le \bar{r}^2 \le 1$, ponieważ $0 \le q(\mathbf{t}_{BS}) \le q(\bar{\mathbf{y}}_S)$. Na podstawie wzoru (5.11) mamy:

Wniosek 5.2: Średni promień wektora $\mathbf{t}_{BS}$ maleje wraz ze wzrostem siły zależności między cechami y i x mierzonej średnim współczynnikiem determinacji.

Przypuśćmy teraz, że statystykowi są znane wartości składowych wektora średnich $\bar{\mathbf{x}}$ cech w populacji. Pozostaje jeszcze problem zdobycia informacji o macierzy parametrów **B**. Jej elementy szacujemy na podstawie obserwacji cech y i x w losowanej zwrotnie próbie prostej o liczebności n, za pomocą macierzy statystyk:

$$\tilde{\mathbf{B}} = -\tilde{\mathbf{C}}_{*xx}^{-1}\tilde{\mathbf{C}}_{*xy} \tag{5.13}$$

gdzie: $\tilde{\mathbf{C}}_{*xy}$ i $\tilde{\mathbf{C}}_{*xx}$ są macierzami wariancji i kowariancji z próby. Zastępując we wzorze (5.6) macierz **B** przez $\tilde{\mathbf{B}}$ mamy:

$$\tilde{\mathbf{t}}_{BS} = \bar{\mathbf{y}}_S + (\bar{\mathbf{x}}_S - \bar{\mathbf{x}})\tilde{\mathbf{B}} \tag{5.14}$$

Można wykazać następującą własność:

Wniosek 5.3: Estymator $\tilde{\mathbf{t}}_{BS}$ jest asymptotycznie nieobciążony, natomiast jego macierz wariancji i kowariancji w przybliżeniu określa wzór (5.8), którą oceniamy poprzez zastąpienie w tym wzorze momentów $c_*(x_i,y_j)$ $(i,j=1,...,m)$ przez ich estymatory nieobciążone, którymi są odpowiednie momenty $c_{*S}(x_i,y_j)$ z próby prostej losowanej bezzwrotnie.

Zakładamy teraz, że jest nieznana nie tylko macierz **B**, od której zależy wektor estymatorów regresyjnych, lecz także wektor $\bar{\mathbf{x}}$ średnich w populacji cech wspomagających. W tej sytuacji zachodzi potrzeba estymacji tych parametrów. W tym celu według okreslonego wzorem (1.39) planu P_5 losujemy bezzwrotnie prostą próbę podwójną $S=\{S_1, S_2\}$. W próbie prostej S_1 losowanej bezzwrotnie z całej populacji są obserwowane wartości cech pomocniczych, a w próbie prostej S_2 losowanej bezzwrotnie ze zbioru s_1 są jeszcze obserwowane wartości wektora zmiennych y. Następnie na podstawie tych obserwacji są wyznaczane nieobciążone estymatory $\bar{\mathbf{C}}_{*yx} = \left[\bar{c}_*\left(y_i, x_j\right)\right]$, $\bar{\mathbf{C}}_{*xx} = \left[\bar{c}_*\left(x_i, x_j\right)\right]$ odpowiednio macierzy $\mathbf{C}_{*yx}$, $\mathbf{C}_{*xx}$, przy czym:

$$\bar{c}_*\left(y_i, x_j\right) = \frac{1}{n_2 - 1} \sum_{k \in S_2} \left(y_{ik} - \bar{y}_{iS_2}\right)\left(x_{jk} - \bar{x}_{jS_2}\right)$$

$$\bar{c}_*\left(x_i, x_j\right) = \frac{1}{n_2 - 1} \sum_{k \in S_2} \left(x_{ik} - \bar{x}_{iS_2}\right)\left(x_{jk} - \bar{x}_{jS_2}\right)$$

Wtedy przy założeniu, że $\det\left(\bar{\mathbf{C}}_{*xx}\right) > 0$, określamy estymator danej wzorem (5.6) macierzy **B** za pomocą statystyki:

$$\bar{\mathbf{B}} = -\bar{\mathbf{C}}_{*xx}^{-1}\bar{\mathbf{C}}_{*xy}$$

Podstawiając ten wzór do wyrażenia (5.7), w którym także wektor $\bar{\mathbf{x}}$ zastępujemy przez $\bar{\mathbf{x}}_{S_1}$, otrzymujemy estymator wektora $\bar{\mathbf{y}}$:

$$\bar{\mathbf{t}}_{BS} = \bar{\mathbf{y}}_{S_2} + (\bar{\mathbf{x}}_{S_2} - \bar{\mathbf{x}}_{S_1})\bar{\mathbf{B}} \tag{5.15}$$

Parametry wektora $\bar{\mathbf{t}}_{BS}$ wyprowadził Wywiał (1988 i 1992):

$$E(\bar{\mathbf{t}}_{BS}) = \bar{\mathbf{y}} + 0(n_1^{-1}) + 0(n_2^{-1}) \tag{5.16}$$

$$\mathbf{V}(\bar{\mathbf{t}}_{BS}) = \frac{N - n_1}{Nn_1}\mathbf{C}_{*yy} + \frac{n_1 - n_2}{n_1 n_2}\left(\mathbf{C}_{*yy} - \mathbf{C}_{*yx}\mathbf{C}_{*xx}^{-1}\mathbf{C}_{*xy}\right) + 0(n_1^{-1}) + 0(n_2^{-2}) + 0(n_1^{-1}n_2^{-1}) \tag{5.17}$$

Korzystając z definicji średniego promienia estymatora oraz ze wzorów (5.17), (5.11) i (5.12) dochodzimy do równania:

$$q^2(\bar{\mathbf{t}}_{BS}) = \mathrm{tr}(\mathbf{C}_{*yy})\left(\frac{\bar{r}^2}{n_1} + \frac{1 - \bar{r}^2}{n_2} - \frac{1}{N}\right) \tag{5.18}$$

przy czym parametr $\bar{r}$ określa wzór (5.12).

5.3. Optymalizacja liczebności prób składowych próby podwójnej przy ustalonych kosztach obserwacji cech

Niech k_1 będzie kosztem jednostkowym obserwacji cech dodatkowych, a k_2 kosztem jednostkowym obserwacji zmiennych, których średnie są przedmiotem estymacji. Nakłady przeznaczane na obserwację wszystkich cech oznaczamy przez K. Liniową funkcję kosztów określa wzór:

$$k(n_1,n_2) = k_1n_1 + k_2n_2 \tag{5.19}$$

Poniżej formułujemy i rozwiązujemy zadania mające na celu wyznaczenie liczebności prób, które przy ustalonych kosztach dopuszczalnych obserwacji zmiennych minimalizują wybraną funkcję ryzyka estymacji. Większość konstruowanych tutaj zadań optymalizacji liczebności prób była rozwiązywana przez autora (1988).

Rozmiary prób są oczywiście liczbami naturalnymi. Jednakże podczas poszukiwania ich poziomów optymalnych traktujemy je jako liczby rzeczywiste, co pozwala stosować wygodny rachunek różniczkowy do poszukiwania ekstremów funkcji.

5.3.1. Minimalizacja kwadratowej funkcji ryzyka estymacji

Kwadratową funkcję ryzyka określamy jako sumę ważonych wariancji składowych wektorowej strategii $(\bar{\mathbf{t}}_{BS}, P_5)$, określonej wzorem (5.15), którą zapisujemy równaniem:

$$u_1(n_1,n_2) = \sum_{i=1}^{m} a_i D^2(\bar{\mathbf{t}}_{BiS}, P_5) \tag{5.20}$$

przy czym $\mathbf{a} = [a_1...a_m] \in R^m - \{\mathbf{o}_m\}$. Ze wzoru (5.18) wynika, że:

$$D^2(\bar{\mathbf{t}}_{BiS}, P_{15}) = v_*(y_i)\left(r_i^2 n_1^{-1} + \left(1 - r_i^2\right)n_2^{-1} - N^{-1}\right) \tag{5.21}$$

a zatem:

$$u_1(n_1,n_2) = q^2\left(\frac{\bar{r}^2}{n_1} + \frac{1-\bar{r}^2}{n_2} - \frac{1}{N}\right) \tag{5.22}$$

gdzie:

$$q^2 = \sum_{i=1}^{m} a_i v_*(y_i) \tag{5.23}$$

$$\bar{r}^2 = \sum_{i=1}^{m} r_i^2 w_i, \qquad w_i = q^{-2} a_i v_*(y_i) \tag{5.24}$$

Parametr $\bar{r}^2$,jest więc ważoną wartością średnią kwadratów współczynników korelacji wielorakiej r_i (i=1,...,m). Przypomnijmy, że r_i mierzy siłę zależności korelacyjnej między zmienną y_i i cechami pomocniczymi x oraz jest i-tym z kolei elementem głównej przekątnej macierzy $\mathbf{C}_{*yx} C_{*xx}^{-1} \mathbf{C}_{*xy}$. Stąd i z faktu, że dla każdego i=1,...,m zachodzą nierówności $0 \le r_i \le 1$, $0 \le w_i \le 1$, wynika, iż $0 \le \bar{r} \le 1$.Współczynnik $\bar{r}^2$ można również traktować jako ważoną średnią współczynników determinacji r_i^2 i(i=1,...,m).

Nasze zadanie polega na takim ustaleniu liczebności prób składowych próby podwójnej, aby określona funkcja ryzyka osiągnęła minimum przy ustalonych kosztach dopuszczalnych obserwacji wszystkich cech, co ściślej zapisujemy wyrażeniem:

$$\begin{cases} u_1(n_1,n_2) = \min\text{im} \\ k(n_1,n_2) \le K, \;\; 2 \le n_2 < n_1 \le N \end{cases} \tag{5.25}$$

przy czym funkcję kosztów określa wzór (5.19).

Wywiał (1992) wykazuje, że jeśli $3k_1+2k_2 \le K$ i $\bar{r} > r_* = \frac{k_1}{k_1+k_2}$, to funkcja $u_1(n_1,n_2)$ osiąga minimum dla pary liczebności $(\underline{n}_1,\underline{n}_2)$:

$$\begin{cases} \underline{n}_1 = \min\text{imum}\left\{N, n_1', n_1^*\right\} \\ \underline{n}_2 = \max\text{imum}\left\{n_2', 2, n_2^*\right\} \end{cases} \tag{5.26}$$

przy czym:

$$n_1' = \frac{K-2k_1}{k_1}, \qquad n_2' = \frac{K-Nk_1}{k_2}, \qquad n' = \frac{K}{k_1+k_2} \qquad n_1^* = g\frac{\bar{r}}{\sqrt{k_1}}, \qquad n_2^* = g\sqrt{\frac{1-\bar{r}^2}{k_2}}$$

gdzie:

$$g = \frac{K}{\bar{r}\sqrt{k_1} + \sqrt{\left(1-\bar{r}^2\right)k_2}}$$

Ponadto strategia $\left(\bar{\mathbf{t}}_{BS}, P_5\right)$ jest nie gorszą od $\left(\bar{\mathbf{y}}_S, P_3\right)$, w sensie kwadratowej funkcji ryzyka, jeśli jest spełniona nierówność:

$$u_1\left(\underline{n}_1,\underline{n}_2\right) \le u_2\left(\underline{n}\right) = k_2^{-1}K - N^{-1} \tag{5.27}$$

która zachodzi:

a) dla pary $\left(\underline{n}_1,\underline{n}_2\right) = \left(n_1^*, n_2^*\right)$, gdy

$$\bar{r} \ge r_o = 2\frac{\sqrt{k_1k_2}}{k_1+k_2} \ge r_*^2 \tag{5.28}$$

przy czym $r_* = \frac{k_1}{k_1+k_2}$

b) dla pary $\left(\underline{n}_1,\underline{n}_2\right) = \left(N, n_2'\right)$, jeśli:

$$\bar{r} \ge r_{**} = N\sqrt{\frac{k_1k_2}{K\left[N\left(k_1+k_2\right)-K\right]}} \ge r_o \tag{5.29}$$

c) dla pary $\left(\underline{n}_1,\underline{n}_2\right) = \left(n_1', 2\right)$, jeżeli:

$$\bar{r} \geq r_{***} = \frac{K - 2k_2}{\sqrt{K\left(K - 2k_1 - 2k_2\right)}} \geq r_o \tag{5.30}$$

Wniosek 5.4: Warunkami koniecznymi opłacalności stosowania wektorowej strategii regresyjnej $\left(\bar{\mathbf{t}}_{BS}, P_5\right)$ zamiast wektora średnich z próby prostej $\left(\bar{\mathbf{y}}_S, P_3\right)$ są aby: koszt jednostkowy k_2 obserwacji zmiennych badanych *y* był większy od kosztu jednostkowego k_1 obserwacji zmiennych wspomagających *x* oraz by wartość pierwiastka ze średniego współczynnika determinacji była większa od poziomu r_o, określonego wzorem (5.28).

Na zakończenie dodajmy, że jeśli wszystkie elementy wektora **a**=[a_1 ...a_m] parametrów funkcji ryzyka, danej wzorem (5.20), są równe liczbie jeden, to zadanie (5.25) staje się problemem minimalizacji warunkowej kwadratu średniego promienia strategii estymacji $\left(\bar{\mathbf{t}}_{BS}, P_5\right)$. Zaś w przypadku gdy $a_i = \bar{y}_i^{-2}$ $\left(i = 1, K, m\right)$, funkcja ryzyka staje się sumą kwadratów współczynników zmienności poszczególnych estymatorów składowych wektora $\bar{\mathbf{t}}_{BS}$.

5.3.2. Minimalizacja uogólnionej wariancji

W rozdziale 2 uzasadniano, że uogólniona wariancja estymatorów wektora parametrów jest monotoniczną funkcją miary objętości elipsoidalnego obszaru ufności dla tych parametrów. Dlatego uogólniona wariancja służy do oceny rzędu precyzji estymacji prowadzonej na podstawie takiego obszaru ufności. Należy więc tak wyznaczyć liczebności prób, by uogólniona wariancja osiągała minimum przy ustalonych kosztach dopuszczalnych obserwacji cech. Traktując uogólnioną wariancję jako funkcję liczebności prób składowych próby podwójnej oznaczamy ją symbolem $u_5(n_1,n_2)=\det\mathbf{V}\left(\bar{\mathbf{t}}_{BS}, P_5\right)$, gdzie estymator $\bar{\mathbf{t}}_{BS}$ i macierz $\mathbf{V}\left(\bar{\mathbf{t}}_{BS}, P_5\right)$ określają odpowiednio wzory (5.15) i (5.17). Zadanie optymalizacyjne formułujemy wyrażeniem:

$$\begin{cases} u_5\left(n_1, n_2\right) = \min \operatorname{im} \\ k\left(n_1, n_2\right) \leq K, \quad 2 \leq n_2 < n_1 \leq N \end{cases} \tag{5.31}$$

przy czym funkcję kosztów określa wzór (5.19).

Załóżmy, że macierz wariancji i kowariancji $\mathbf{C}_{*yy}$ jest dodatnio określona. Wówczas na podstawie znanego twierdzenia algebry liniowej [por. np. Rao (1982), s. 52] dotyczącego jednoczesnej diagonalizacji dwóch macierzy, z której przynajmniej jedna jest dodatnio określona, wnioskujemy, że istnieje nieosobliwa macierz kwadratowa **Q** stopnia m taka, iż $\mathbf{Q}^T\mathbf{Q}=\mathbf{C}_{*yy}$ i $\mathbf{Q}^T\mathbf{R}_Q\mathbf{Q}=\mathbf{C}_{*yx}\mathbf{C}_{*xx}^{-1}\mathbf{C}_{*xy}$, przy czym $\mathbf{R}_Q$ jest macierzą diagonalną, której elementy diagonalne oznaczamy przez r_{Qi}^2, $i = 1, K, m$. Wtedy wyznacznik $u_5(n_1,n_2)$ na podstawie wzoru (5.17) zapisujemy następująco:

$$u_5(n_1,n_2)=\det\left\{\mathbf{Q}^T\left(\frac{1}{n_1}\mathbf{R}_Q+\frac{1}{n_2}\left(\mathbf{I}_m-\mathbf{R}_Q\right)-\frac{1}{N}\mathbf{I}_m\right)\mathbf{Q}\right\}$$

Na podstawie znanych własności wyznacznika iloczynu macierzy mamy:

$$u_5(n_1,n_2)=\det\left(\mathbf{C}_{*yy}\prod_{i=1}^{m}f_{*i}\right) \tag{5.32}$$

gdzie:

$$f_{*i}=f_i-\frac{1}{N},\qquad f_i=\frac{r_{Qi}^2}{n_i}+\frac{1-r_{Qi}^2}{n_2}$$

Elementy diagonalne macierzy $\mathbf{R}_Q$ można również wyznaczyć jako pierwiastki następującego równania wyznacznikowego [por. np. pracę Rao (1982), s. 580]:

$$\det(\mathbf{C}_{*yx}\mathbf{C}_{*xx}^{-1}\mathbf{C}_{*xy}-r_Q^2\mathbf{C}_{*yy})=0$$

Element diagonalny $0\le r_{Qi}^{-1}\le 1$(i=1,..,m) jest kwadratem tzw. współczynnika korelacji kanonicznej, który mierzy ścisłość zależności między wektorem cech *y* i zmiennych wspomagających *x*.

Podczas wyznaczania rozwiązania będzie pożytecznym następujący lemat:

Lemat 5.1 [Wywiał (1992)]: Funkcja $u_5(n_1,n_2)$ jest ściśle wypukła w obszarze $\mathcal{D}=\{(n_1,n_2):0<n_2<n_1\le N\}$.

Wywiał (1992) wykazuje, że punkt optymalny $\underline{A}(\underline{n}_1,\underline{n}_2)$ znajduje się na odcinku o końcach A'(n'$_1$,n'$_2$) i A"(n",n"), przy czym:

$$\begin{cases} n_1'=\min\left\{\dfrac{K-2k_2}{k_1},N\right\} \\ n_2'=\max\left\{2,\dfrac{K-Nk_1}{k_2}\right\} \\ n''=\dfrac{K}{k_1+k_2} \end{cases} \tag{5.33}$$

Dodajmy, że odcinek $\overline{A'A''}$ zawiera się w prostej $k_1n_1+k_2n_2=K$.

Korzystając ze znanej metody czynników nieoznaczonych Lagrange'a można wykazać, że rozwiązanie $\left(n_1^*, n_2^*\right)$ układu równań:

$$n_1 = c\sqrt{\frac{1}{k_1}\sum_{i=1}^{m}\frac{r_{Qi}^2}{f_{*i}}}, \qquad n_2 = c\sqrt{\frac{1}{k_2}\sum_{i=1}^{m}\frac{1-r_{Qi}^2}{f_{*i}}} \tag{5.34}$$

gdzie:

$$\frac{1}{c} = \frac{1}{k}\left\{\sqrt{k_1\sum_{i=1}^{m}\frac{r_{Qi}^2}{f_{*i}}} + \sqrt{k_2\sum_{i=1}^{m}\frac{1-r_{Qi}^2}{f_{*i}}}\right\}$$

stanowią współrzędne punktu $A_*(n_1^*, n_2^*)$, w którym funkcja $u_5(n_1,n_2)$ osiąga minimum na odcinku $\overline{A'''A''}$, gdzie $A'''=A'''(K k_1^{-1},0)$. Jeśli $A_* \in \overline{A'''A''}$ i $A_* \neq A''$, to rozwiązanie zadania (5.31) określa wyrażenie:

$$\begin{cases} \underline{n}_1 = \min\left\{n_1', n_1^*\right\} \\ \underline{n}_2 = \max\left\{n_2', n_2^*\right\} \end{cases} \tag{5.35}$$

Stąd wynika, że po to, by otrzymać liczebności optymalne, należy najpierw wyznaczyć rozwiązanie (n_1^*, n_2^*) nieliniowego układu równań (5.34), które należy wyliczyć jedną z przybliżonych procedur, do których należą opisywane przez Demidowicza i Marona (1965) metody gradientowa, Newtona i iteracyjna. Zastosujemy ostatnią z wymienionych, przy czym dodatkowo będziemy zakładać, że liczebność populacji jest nieograniczona. To uproszczenie powoduje, że gdy $N \to \infty$, to $f_{*i} \to f_i$ dla każdego i=1,...,m, przy czym funkcje f_{*i} , f_i określono przy definiowaniu wyrażenia (5.32). Uwzględniając to dzielimy drugie z kolei równanie układu (5.34) przez pierwsze. Następnie po wprowadzeniu podstawienia $n_2 = wn_1$ otrzymujemy[35] :

$$w = p(w) \tag{5.36}$$

gdzie:

$$p(w) = \sqrt{\frac{k_1}{k_2}\frac{\sum_{i=1}^{m}\frac{1-r_{Qi}^2}{h_i}}{\sum_{i=1}^{m}\frac{r_{Qi}^2}{h_i}}} = \sqrt{\frac{k_1}{k_2}\left(\frac{\sum_{i=1}^{m}\frac{1}{h_i}}{\sum_{i=1}^{m}\frac{r_{Qi}^2}{h_i}} - 1\right)} \tag{5.37}$$

[35] Zabieg ten przy rozwiązywaniu innego zadania optymalizacyjnego zastosowali Greń i Koźniewska (1964).

$$h_i = wr_{Qi}^2 + 1 - r_{Qi}^2, \qquad i=1,...m \tag{5.38}$$

Theil (1979) proponje, aby siłę zależności liniowej między dwoma wektorami określać za pomocą wartości średniej współczynników korelacji kanonicznej, która ma postać:

$$\bar{r}_Q^2 = \frac{1}{m'}\sum_{i=1}^{m} r_{Qi}^2 = \frac{1}{m'}\mathrm{tr}(\mathbf{C}_{*yx}\mathbf{C}_{*xx}\mathbf{C}_{*xy}) \tag{5.39}$$

przy czym m' = min{m,z} gdzie, przypomnijmy, literą z oznaczono ilość zmiennych wspomagających. Zatem im bliższa jedności jest wartość średniej $\bar{r}_Q^2$, tym silniej są skorelowane ze sobą wektory zmiennych *y i x*.

Lemat 5.2 [Wywiał (1992)]: Jeśli $\frac{m'}{m}\bar{r}_Q^2 > \frac{k_1}{k_2}$, to $0<w<1$ i $n_1>n_2$.

Rozwiązanie równania (5.36) otrzymujemy z następującego procesu iteracyjnego:

$$w_{j+1} = p(w_j), \qquad j = 0,1,2,... \tag{5.40}$$

przy czym startowe rozwiązanie w_0 wybiera się tak, aby $0<w_0<1$. Wywiał (1992) wykazał, że proces iteracyjny (5.40) jest zbieżny z rozwiązaniem dokładnym w_* równania (5.36) niezależnie od przybliżenia początkowego $z_0 \in (0;1)$.

Demidowicz i Maron (1965), s. 117 podają sposób oceny dokładności rozwiązania w j-tej iteracji, co w naszym przypadku określa nierówność:

$$| w_* - w_j | \le 2^{1-j} | w_1 - w_0 | \tag{5.41}$$

Po obliczeniu $w_1=p(w_0)$ można więc łatwo wyznaczyć niezbędną ilość iteracji zapewniającą otrzymanie rozwiązania z żądaną dokładnością.

Po oszacowaniu przybliżonej wartości pierwiastka w_*, wreszcie oceniamy parę szukanych liczebności (n_1^*, n_2^*) na podstawie równań $n_2 = wn_1$ i $k_1n_1+k_2n_2=K$. Wtedy:

$$n_1^* = \frac{K}{k_1 + w_* k_2}, \qquad n_2^* = w_* n_1^* \tag{5.42}$$

Wywiał (1992) sugeruje, aby rozwiązanie w_* równania (5.36) wyznaczyć znaną metodą siecznej lub tzw. równego podziału, które opisują m.in. Demidowicz i Maron (1965) oraz Rice (1983).

Wniosek 5.5: Jeśli wektor wartości średnich w populacji jest oceniany na podstawie obszaru ufności konstruowanego za pomocą strategii estymacji regresyjnej $\left(\bar{\mathbf{t}}_{BS}, P_5\right)$ to do minimalizacji objętości tego obszaru ufności prowadzi rozwiązanie zadania (5.31). Optymalne liczebności prób składowych prostej próby podwójnej określają wyrażenia (5.35) i (5.33). Z lematu 5.2 wynika, że takie liczebności można wyznaczyć, jeśli są spełnione dwa warunki. Po pierwsze, gdy koszt jednostkowy k_1 obserwacji cech pomocniczych *x* jest mniejszy od kosztu jednostkowego k_1 obserwacji zmiennych *y*, których średnie w populacji są przedmiotem naszej

estymacji. Po drugie, siła zależności mierzona średnią kwadratów współczynników korelacji kanonicznej między zmiennymi *x* i *y* musi być na tyle duża, aby nierówność występująca w założeniu lematu 5.2 była spełniona. Należy więc tak dobierać cechy *x*, aby były one silnie skorelowane ze zmiennymi *y* oraz by koszt jednostkowy obserwacji zmiennych *x* był znacznie niższy od takiego kosztu k_2 dla cech *y*.

5.3.3. Optymalizacja celowa liczebności prób

Formułowane tutaj zadania należą do klasy tzw. celowych zadań programowania wielokryterialnego, których ogólne własności badano m.in. w pracach Konarzewskiej-Gubały (1980) oraz Galasa, Nykowskiego i Żółkiewskiego (1987). Wyróżnia się tutaj dwa rodzaje problemów. Pierwszy polega na znalezieniu optymalnych rozwiązań cząstkowych osobno ze względu na każdą funkcję kryterium. Następnie tak wyznacza się rozwiązanie, by było ono bliskie w określonym sensie otrzymanym wcześniej rozwiązaniom optymalnym. Drugi sposób polega na obliczeniu wartości optymalnych poszczególnych funkcji kryteriów, które są nazywane rozwiązaniem idealnym lub utopijnym, lecz w przestrzeni kryterialnej. Potem poszukuje się jednego rozwiązania, dla którego wartości funkcji celów będą wykazywać dużą zgodność z ich optymalnymi wartościami. Zatem określanie tego typu zagadnień mianem celowych znajduje uzasadnienie w dążeniu do uzyskania rozwiązania najbardziej zgodnego z docelowymi rozwiązaniami idealnymi, które zwykle nie mogą być jednocześnie osiągnięte. Dodajmy, że można na wiele sposobów formułować kryteria zgodności między osiąganym rozwiązaniem a rozwiązaniami cząstkowymi lub idealnymi.

Niech $(\underline{n}_1,\underline{n}_{2i})$, i=1,...,m, będą optymalnymi liczebnościami otrzymanymi jako rozwiązania zadania:

$$\begin{cases} u_{6i}(n_1,n_2) = \min \\ k_1 n_1 + k_{2i} n_2 \le K, \quad 2 \le n_2 < n_1 \le N \end{cases} \tag{5.43}$$

przy czym $u_{6i}(n_1,n_2)=D^2(\bar{t}_{BiS},P_5)$ określa wzór (5.21). Rozwiązanie optymalne postawionego zadania jest szczególnym przypadkiem rozwiązania $(\underline{n}_{1i}, \underline{n}_{2i})$ zadania (5.25) dla m=1, i $k_2=k_{2i}$, a określają je wzory (5.26), przy czym występujące w nich wielkości k_{2i} i $\bar{r}$ należy zastąpić przez odpowiednie parametry k_{2i} oraz r_i, gdzie ostatni z nich jest współczynnikiem korelacji wielorakiej między cechą y_i i zmiennymi pomocniczymi *x*.

Otrzymane tą drogą pary liczebności $(\underline{n}_{1i}, \underline{n}_{2i})$, i=1,...,m będą na ogół różniły się między sobą. W celu wyznaczenia rozwiązania kompromisowego wprowadzamy dodatkowe kryterium:

$$u_6(n_1,n_2) = \sum_{i=1}^{m}(\underline{n}_{1i}-n_1)^2 w_{1i} + \sum_{i=1}^{m}(\underline{n}_{2i}-n_2)^2 w_{2i} \tag{5.44}$$

przy czym $\sum_{i=1}^{m} w_{hi} = 1$ oraz w $w_{hi} > 0$ dla każdego i=1,..,m, h=1,2. Przyjmujemy je za funkcję celu zadania optymalizacyjnego postaci:

$$\begin{cases} u_6(n_1,n_2) = \text{minimum} \\ k_1 n_1 + k_2 n_2 \le K, \quad 2 \le n_2 < n_1 \le N \end{cases} \tag{5.45}$$

przy czym $k_2 = \sum_{i=1}^{m} k_{2i}$. Postulujemy więc znaleźć takie liczebności ($\underline{n}_1$,$\underline{n}_2$), które będą najbliższe rozwiązaniom cząstkowym ($\underline{n}_{1i}$, $\underline{n}_{2i}$), i=1,..,m w sensie ważonej sumy kwadratów odchyleń.

Wywiał (1992) podaje rozwiązanie zadania (5.45), gdy $n_1^* < n'$ i $n_2^* > n'$, gdzie $n' = \frac{K}{k_1 + k_2}$, natomiast

$$n_1^* = \underline{\bar{n}}_1 + ck_1, \qquad n_2^* = \underline{\bar{n}}_2 + ck_2 \tag{5.46}$$

gdzie:

$$c = \frac{K - k_1 \underline{\bar{n}}_1 - k_2 \underline{\bar{n}}_2}{k_1^2 + k_2^2}, \qquad \underline{\bar{n}}_k = \sum_{i=1}^{m} \underline{n}_{ki} w_{ki} \qquad \text{dla k=1,2}$$

Wtedy, można wykazać, że rozwiązanie zadania (5.45) określa wyrażenie (5.26), przy czym tym razem występujące w nim liczebności (n_1^*, n_2^*) są dane wzorem (5.46).

Niech $\underline{u}_{6i}$ ($\underline{n}_{1i}$,$\underline{n}_{2i}$)=$\underline{u}_6$ dla i=1,...,m będzie minimalną wartością funkcji celu zadania (5.43). Wtedy wektor $\underline{\mathbf{u}}_6$=[$\underline{u}_{61}$...$\underline{u}_{6m}$] jest rozwiązaniem idealnym w przestrzeni kryterialnej zadania jednoczesnej i warunkowej minimalizacji wektora wariancji estymatorów regresyjnych składowych wektora $\bar{\mathbf{t}}_{BS}$, które sformułowano w podpunkcie drugim. Współczynnik efektywności względnej określamy wzorem:

$$e_i = \frac{\underline{u}_{6i}}{u_{6i}(n_1,n_2)}, \qquad i=1,...,m \tag{5.47}$$

Sumę odwrotności tych współczynników przyjmujemy za funkcję celu zadania optymalizacyjnego w postaci:

$$\begin{cases} u_7(n_1,n_2) = \text{minimum} \\ k_1 n_1 + k_2 n_2 \le K, \quad 2 \le n_2 < n_1 \le N \end{cases} \tag{5.48}$$

gdzie:

$$u_7(n_1,n_2) = \sum_{i=1}^{m} \frac{1}{e_i} \tag{5.49}$$

Sformułowane zadanie jest szczególnym przypadkiem problemu optymalizacyjnego (5.25), ponieważ zastępując we wzorze (5.20) wektor **a** parametrów przez wektor odwrotności składowych minimalnych wartości wariancji, czyli przez $\mathbf{a}=\left[\frac{1}{\underline{u}_{61}} \dots \frac{1}{\underline{u}_{6m}}\right]$, otrzymujemy wyżej zapisaną funkcję u_7. Zatem rozwiązanie zadania (5.48) wynika ze wzorów (5.26).

5.4. Minimalizacja kosztów obserwacji cech przy ustalonym ryzyku estymacji

5.4.1. Ustalony poziom dopuszczalny kwadratowej funkcji ryzyka

Zakładamy, że została określona dopuszczalna wartość u_o funkcji ryzyka $u_1(n_1,n_2)$, danej wzorami (5.20)-(5.24), czyli $u_1(n_1,n_2) \le u_o$. Nierówność ta jest równoważna następującej

$$u_*(n_1,n_2) = \frac{\bar{r}^2}{n_1} + \frac{1-\bar{r}^2}{n_2} \le u_*$$

gdzie $\bar{r}$ określa wzór (5.24), natomiast $u_* = \frac{1}{N} + \frac{u_o}{q^2}$, przy czym q^2 określa wzór (5.23). Przy tak wprowadzonym warunku należy zminimalizować funkcję liniową kosztów, daną wzorem (5.19). Po uwzględnieniu ograniczeń wynikających ze sposobu losowania prostej próby podwójnej wybieranej bezzwrotnie mamy zadanie:

$$\begin{cases} k(n_1,n_2) = \min \\ u_{*1}(n_1,n_2) \le u_* \\ 2 \le n_2 < n_1 \le N \end{cases} \tag{5.50}$$

Wprowadzamy następujące oznaczenia:

$$n_1'' = \frac{2\bar{r}^2}{2u_* + \bar{r}^2 - 1}, \qquad n_2'' = \frac{N(1-\bar{r}^2)}{Nu_* - \bar{r}^2}, \qquad n'' = \frac{1}{u_*} \tag{5.51}$$

$$n_1^{**} = c\frac{\bar{r}}{\sqrt{k_1}}, \qquad n_2^{**} = c\sqrt{\frac{1-\bar{r}^2}{k_2}} \tag{5.52}$$

gdzie:

$$c = \frac{\bar{r}\sqrt{k_1} + \sqrt{(1-\bar{r}^2)k_2}}{u_*} \tag{5.53}$$

Wtedy np. w pracy Wywiała (1992) znajdujemy, że jeśli $q^2\left(\frac{1}{2} + \frac{1}{N}\right) > u_{o2}$ i $\bar{r} > r_* = \frac{k_1}{k_1 + k_2}$, to rozwiązanie naszego zadania ma postać:

$$\begin{cases} \underline{n}_1 = \min\left\{N, n_1'', n_1^{**}\right\} \\ \underline{n}_2 = \max\left\{n_2'', 2, n_2^{**}\right\} \end{cases} \tag{5.54}$$

Ponadto Wywiał (1992) wykazuje, że jeśli $k_1<k_2$, to koszt obserwacji cech *x* i *y* przy posługiwaniu się strategią estymacji $(\bar{\mathbf{t}}_{BS}, P_5)$ nie jest większy od kosztu obserwacji cechy *y* przy użyciu strategii $(\bar{y}_S, P_3)$, gdy liczebności $(\underline{n}_1, \underline{n}_2)$, dane wzorem (5.54), spełniają nierówność:

$$k(\underline{n}_1, \underline{n}_2) \le \frac{k_2}{u_*} \tag{5.55}$$

5.4.2. Ustalone błędy szacunku średnich poszczególnych cech

Zadanie polega na takim ustaleniu liczebności prób składowych próby podwójnej, by dana wzorem (5.19) funkcja kosztów $k(n_1, n_2)$ osiągnęła minimum przy założeniu, że wariancje estymatorów składowych wektora $\bar{\mathbf{t}}_{BS}$ nie przekroczą z góry ustalonych poziomów, czyli że $D^2(\bar{t}_{iBS}, P_5) \le d_i^2$, i=1,...,m. Te nierówności na podstawie wzoru (5.21) są równoważne następującym: $u_i(n_1,n_2) \le d_{*i}$, i=1,...,m, gdzie:

$$u_i(n_1,n_2) = \frac{r_i^2}{n_1} + \frac{1-r_i^2}{n_2}, d_{*i} = \frac{d_i}{v_*(y_i)} + \frac{1}{N}$$

Formułowane zadanie ma więc postać:

$$\begin{cases} k(n_1, n_2) = \min \\ u_i(n_1, n_2) \le d_{*i}, \quad i=1,...,m \\ 2 \le n_2 < n_1 \le N \end{cases} \tag{5.56}$$

Obszar $\mathcal{D}$ rozwiązań dopuszczalnych zadania jest iloczynem zbiorów $\mathcal{D}_i$, a więc $\mathcal{D} = \prod_{i=1}^{m} \mathcal{D}_i$. Każdy nie pusty zbiór $\mathcal{D}_i$ jest wspólną częścią simpleksu danego nierównościami $2 \le n_2 < n_1 \le N$ i obszaru wypukłego ograniczonego hiperbolą o równaniu $u_1(n_1,n_2) = d_{*i}$.

Obszar $\mathcal{D}$ jest wypukły, bo jest wspólną częścią zbiorów wypukłych $\mathcal{D}_i$ (i=1,..,m). Punkt $P_{ij}(n_{1ij},n_{2ij})$ przecięcia dwóch hiperbol $u_i(n_1,n_2)=d_{*i}$, $u_j(n_1,n_2)=d_{*j}$ (i≠j=1,...m) ma współrzędne:

$$n_{1ij} = \frac{r_i^2(r_j^2 d_{*i} - r_i^2 d_{*j})}{q_{*i}(r_j^2 d - r_i^2 d) - (1-r_i^2)\left[r_j^2(1-r_i^2) - r_i^2(1-r_j^2)\right]} \tag{5.57}$$

$$n_{2ij} = \frac{r_j^2 d_{*i} - r_i^2 d_{*j}}{r_j^2(1-r_i^2) - r_i^2(1-r_j^2)} \tag{5.58}$$

Załóżmy, że zbiór $\mathcal{D}$ nie jest pusty. Wtedy z wypukłości obszaru rozwiązań dopuszczalnych $\mathcal{D}$ i liniowości funkcji celu wynika, że rozwiązanie optymalne postawionego zadania znajduje się na brzegu zbioru $\mathcal{D}$. Rozwiązaniem są współrzędne jednego z wierzchołków obszaru $\mathcal{D}$ lub punkt styczności prostej $k_1n_1+k_2n_2=k_m$ z jego krawędzią będącą odcinkiem odpowiedniej krzywej $u_{1i}(n_1,n_2)=d_{*i}$ (i=1,..,m), gdzie k_m oznacza wartość funkcji celu w punkcie styczności. Spośród wymienionych potencjalnych punktów rozwiązaniem naszego zadania jest ten, dla którego funkcja celu osiągnie wartość minimalną.

Wniosek 5.6: Algorytm wyznaczania tego rozwiązania jest następujący: W pierwszym kroku obliczamy rozwiązania zadania (5.56) osobno dla kolejno ustalanych wskaźników i=1,...,m. W ten sposób otrzymujemy rozwiązania $\underline{P}_i(\underline{n}_{1i},\underline{n}_{2i})=\underline{P}i$ zadania w obszarach $\mathcal{D}_i$, i=1,...,m. Dodajmy, że przy obliczaniu współrzędnych punktów P_i zakładamy, że $2d_{*i}<1$ oraz $r_i>r_*=k_1/(k_1+k_2)$. Wówczas współrzędne punktów $\underline{P}_i$ otrzymujemy jako szczególne rozwiązanie problemu optymalizacyjnego (5.50) dla m=1 i a=1, które obliczamy na podstawie wzorów (5.51)-(5.54) zastępując w nich parametry $\bar{r}$, u_*, q^2 odpowiednio przez r_i, d_{*i} i $v_*(y_i)$. Oznaczmy przez $\mathcal{D}_a$ zbiór tych punktów wśród $\{\underline{P}_i, i=1,...,m\}$, które spełniają wszystkie ograniczenia zadania (5.56). Następnie obliczamy wszystkie punkty P_{ij} (i=1,..,m) przecięcia się hiperbol według wzorów (5.57), (5.58) i wybieramy spośród nich zbiór $\mathcal{D}_b$ tych punktów, które spełniają wszystkie warunki ograniczające zadania (5.,56). W końcu wyznaczamy parę $(\underline{n}_1,\underline{n}_2)$ liczebności optymalnych według wzoru:

$$k(\underline{n}_1,\underline{n}_2) = \min_{(n_1,n_2)\in\mathcal{D}_a\cup\mathcal{D}_b} \text{imum}\{k(n_1,n_2)\} \tag{5.59}$$

Otrzymaliśmy więc liczebności prób składowych próby podwójnej minimalizujące funkcję kosztów obserwacji przy ustalonych dopuszczalnych wariancjach estymatorów regresyjnych wektora średnich w populacji ustalonej.

5.4.3. Ustalony poziom dopuszczalny uogólnionej wariancji

Zakładamy, że jest ustalona objętość elipsoidy ufności, na podstawie której jest szacowany wektor wartości średnich w populacji. Jeśli elipsoida jest budowana na podstawie wektora estymatorów regresyjnych $\bar{\mathbf{t}}_{BS}$, to jej objętość jest proporcjonalna do uogólnionej wariancji tego wektora statystyk. Zatem poziom dopuszczalny objętości elipsoidy ufności można zastąpić warunkiem $\det\mathbf{V}(\bar{\mathbf{t}}_{BS},P_5)=u_5(n_1,n_2)\le d$, gdzie d jest poziomem dopuszczalnym uogólnionej wariancji strategii $(\bar{\mathbf{t}}_{BS},P_5)$. Zadanie polega na takim ustaleniu liczebności $(\underline{n}_1,\underline{n}_2)$ prób składowych próby podwójnej, by dana wzorem (5.19) funkcja kosztów $k(n_1,n_2)$ osiągnęła minimum, a więc:

$$\begin{cases} k(n_1, n_2) = \min \\ u_5(n_1, n_2) \le d, \quad 2 \le n_2 < n_1 \le N \end{cases} \tag{5.60}$$

przy czym funkcję u_5 w dogodnej postaci określa wzór (5.32).

Wywiał (1992) wykazuje, że rozwiązanie optymalne naszego zadania ma postać:

$$\begin{cases} \underline{n}_1 = \min\{N, n_1'', n_1^*\} \\ \underline{n}_2 = \max\{2, n_2'', n_2^*\} \end{cases} \tag{5.61}$$

Sposób wyliczania liczebności n_1'' i n_2'' wyjaśnia wyrażenie (5.51), natomiast przez (n_1^*, n_2^*) oznaczono rozwiązanie układu równań:

$$\begin{cases} \alpha(n_1, n_2) = -k_1 k_2^{-2} \\ u_5(n_1, n_2) = d \end{cases} \tag{5.62}$$

gdzie:

$$\alpha(n_1, n_2) = -\frac{F_1(n_1, n_2)}{F_2(n_1, n_2)} \tag{5.63}$$

Przez F_1 i F_2 oznaczono pierwsze pochodne cząstkowe funkcji uwikłanej $F(n_1,n_2)= u_5(n_1,n_2)-d=0$ wyznaczone odpowiednio względem argumentów n_1 i n_2 .

Rozwiązanie postawionego problemu istnieje, jeśli obszar rozwiązań dopuszczalnych nie jest pusty, a więc gdy $n_2'' < N$ i gdy $\underline{n}_1 > \underline{n}_2$.

Przy założeniu, że $N \to \infty$, rozwiązanie układu (5.62) nieco upraszcza się i przebiega podobnie jak szukanie pierwiastka układu równań (5.34). Wprowadzamy podstawienie $n_2 = w n_1$ oraz pierwsze równanie układu (5.62) przekształcamy do postaci danej wzorem (5.36), a drugie równanie do postaci:

$$\frac{1}{n_1^m} \det C_{*yy} \prod_{i=1}^{m} \left(r_{Qi}^2 + \frac{1 - r_{Qi}^2}{w} \right) = d \tag{5.64}$$

Wartość w_* będącą rozwiązaniem równania (5.36) można z zadaną dokładnością wyliczyć iteracyjnie na podstawie wzoru (5.40). Wtedy na podstawie wzoru (5.64) i równania $n_2 = w n_1$ otrzymujemy:

$$\begin{cases} n_1^* = \sqrt{\dfrac{1}{d} \det \mathbf{C}_{*yy} \displaystyle\prod_{i=1}^{m} \left(r_{Qi}^2 + \frac{1 - r_{Qi}^2}{w_*} \right)} \\ n_2^* = w_* n_1^* \end{cases} \tag{5.65}$$

Zaznaczmy, że warunki, przy których zachodzi nierówność $0<w<1$, określa lemat 5.2.

5.5. Minimalizacja ryzyka całkowitego estymacji

Niech kwadratowa funkcja ryzyka całkowitego w naszym przypadku ma postać:

$$u_8(n_1,n_2)=u_1(n_1,n_2) + k(n_1,n_2) \tag{5.66}$$

przy czym funkcje kosztów obserwacji cech $k(n_1,n_2)$ określa wzór (5.19), natomiast kwadratową funkcję ryzyka wzory (5.20)-(5.24). Rozważane tutaj zadanie optymalizacyjne ma postać:

$$\begin{cases} u_8(n_1,n_2)=\min \\ 2\le n_2 < n_1 \le N \end{cases} \tag{5.67}$$

Wprowadźmy oznaczenia:

$$n_1^* = \frac{q\bar{r}}{\sqrt{k_1}}, \qquad n_2^* = q\sqrt{\frac{1-\bar{r}^2}{k_2}} \tag{5.68}$$

Wywiał (1992) wykazuje, że obszarem dopuszczalnym $\mathcal{L}$ rozwiązań zadania (5.67) jest trójkąt o wierzchołkach $A_1(N,2)$, $A_2(N,N)$ i $A_3(2,2)$ bez odcinka $\overline{A_2A_3}$. Jeśli do tego zbioru należy punkt $A_*(n_1^*,n_2^*)$, to jego współrzędne stanowią szukane liczebności optymalne $\underline{A}(\underline{n}_1,\underline{n}_2)$. W przeciwnym razie jeśli punkt A_* nie należy do $\mathcal{L}$, to

a) $\underline{A} = (n_1^*,2)$, jeśli $2 \le n_1^* \le N$ i $n_2^* < 2$

b) $\underline{A} = (N,n_2^*)$, jeżeli $n_1^* > N$ i $2 \le n_2^* < N$

c) $\underline{A} = (N,2)$, gdy $n_1^* \ge N$ i $n_2^* \le 2$

d) $\underline{A}$ nie istnieje, jeżeli $n_1^* < 2$ i $n_2^* < 2$, bądź $n_1^* > N$ i $n_2^* > N$.

Wywiał (1992) również porównuje rozważane tu ryzyko z podobnie definiowanym ryzykiem całkowitym estymacji na podstawie średniej z próby prostej.

6. WEKTOR ŚREDNICH Z PRÓBY WARSTWOWEJ

6.1. Wiadomości wstępne

Zakładamy, że populacja Ω jest podzielona na niepuste podzbiory Ω_h (h=1,..,H) zwane warstwami oraz $\Omega = \bigcup_{h=1}^{H} \Omega_h$. Liczbę elementów populacji tworzących h-tą warstwę oznaczmy przez $0<N_h<N$, przy czym $N = \sum_{h=1}^{H} N_h$. Niech $w_h=N_h N^{-1}$. Przez S_h oznaczmy próbę prostą o liczebności $0<n_h \le N_h$ losowaną bezzwrotnie z h-tej warstwy. Plan losowania bezzwrotnego próby warstwowej [por. Wywiał (1992)] ma postać:

$$P_w(s) = \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1} \tag{6.1}$$

a przy losowaniu zwrotnym:

$$P'_w(s) = \prod_{h=1}^{H} N_h^{-n_h} \tag{6.2}$$

Wprowadźmy oznaczenia:

$$\bar{y}_{ih} = \frac{1}{N_h} \sum_{k=1}^{N_k} y_{ihk}, \qquad v_{*h}(y_i) = c_{*h}(y_i, y_i)$$

$$c_{*h}(y_i, y_j) = \frac{1}{N_{h=1}} \sum_{k=1}^{N_k} (y_{ihk} - \bar{y}_{ih})(y_{jhk} - \bar{y}_{jh})$$

$$\bar{\mathbf{y}}_h = [\bar{y}_{h1} \mathrm{L}\ \bar{y}_{hm}], \qquad \mathbf{C}_{*h} = [c_{*h}(y_i, y_j)]$$

Nieobciążonymi estymatorami parametrów $\bar{\mathbf{y}}_h$ i $\mathbf{C}_{*h}$ są odpowiednio statystyki: $\bar{\mathbf{y}}_S = [\bar{y}_{1S_h} \mathrm{L}\ \bar{y}_{mS_h}]$ i $\mathbf{C}_{*S_h} = [c_{*S_h}(y_i, y_j)]$, gdzie:

$$c_{*S_h}(y_i, y_j) = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_{ihk} - \bar{y}_{iS_h})(y_{jhk} - \bar{y}_{jS_h}); \qquad \bar{y}_{iS_h} = \frac{1}{n_h} \sum_{k \in S_h} y_{ihk} \tag{6.3}$$

Wektor średnich $\bar{\mathbf{y}}$ cech w populacji jest następującą funkcją wektorów średnich zmiennych w poszczególnych warstwach.

$$\bar{\mathbf{y}} = \sum_{h=1}^{H} w_h \bar{\mathbf{y}}_h$$

Znanym w metodzie reprezentacyjnej nieobciążonym estymatorem wektora $\bar{\mathbf{y}}$ wartości średnich cech w populacji jest statystyka:

$$\bar{\mathbf{y}}_{wS} = \sum_{h=1}^{H} w_h \bar{\mathbf{y}}_{hS} \tag{6.4}$$

Jej macierz wariancji i kowariancji ma postać [por. np. Greń (1963); Ghosh (1963a)]:

$$\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w) = \sum_{h=1}^{H} w_h^2 \mathbf{V}(\bar{\mathbf{y}}_{S_h}) = \sum_{h=1}^{H} w_h^2 \frac{N_h - n_h}{N_h n_h} \mathbf{C}_{*h} \tag{6.5}$$

Wariancję i-tego estymatora składowego wektora $\bar{\mathbf{y}}_{wS}$ określa wzór:

$$D^2(\bar{y}_{wiS}, P_w) = \sum_{h=1}^{H} w_h^2 \frac{N_h - n_h}{N_h n_h} v_{*h}(y_i) \tag{6.6}$$

W przypadku zwrotnego losowania prób:

$$\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w) = \sum_{h=1}^{H} \frac{1}{n_h} w_h^2 \mathbf{C}_h \tag{6.7}$$

gdzie: $$\mathbf{C}_h = \frac{N_h - 1}{N} \mathbf{C}_{*h}$$

Nieobciążony estymator macierzy wariancji i kowariancji $\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w)$ otrzymujemy zastępując we wzorze (6.5) macierze $\mathbf{C}_{*h}$ (h=1,..,H) odpowiednio macierzami $\mathbf{C}_{*S_h} = [c_{*S_h}(y_i, y_j)]$, których składowe określa wzór (6.3).

Problemem optymalnej lokalizacji prób w warstwach początkowo zajmował się Neyman (1934). Otrzymane przez niego wyniki dotyczą optymalizacji liczebności prób w przypadku estymacji średniej pojedynczej cechy w populacji. Zaproponował, by przy ograniczonej sumie liczebności wszystkich prób losowanych z warstw tak wyznaczyć liczebności tych prób, aby wariancja estymatora osiągnęła minimum. Tym sposobem obliczane rozmiary prób są nazywane lokalizacją Neymana[36], który sugeruje, że jego sposób otrzymywania liczebności jest także użyteczny w przypadku jednoczesnej estymacji średnich wielu cech. Proponuje, aby optymalizować liczebności prób na podstawie cechy najsilniej skorelowanej z pozostałymi. Uważa też, że ta cecha z punktu widzenia celu badania powinna być najważniejsza.

W niniejszym rozdziale koszty obserwacji zmiennych traktujemy jako funkcję liniową liczebności prób w poszczególnych warstwach postaci:

$$k(\mathbf{n}) = k(n_1,\ldots,n_H) = \sum_{h=1}^{H} k_h n_h \qquad (6.8)$$

gdzie, $\mathbf{n}=[n_1 \ldots n_h]$, natomiast k_h jest kosztem jednostkowym obserwacji cech w h-tej warstwie. Koszty stałe obserwacji pomijamy, ponieważ ich uwzględnienie nie ma istotnego wpływu na otrzymywane rozwiązania formułowanych dalej zadań optymalizacyjnych. Dopuszczalny koszt zmienny obserwacji cech oznaczamy przez K.

Dodajmy, że Beardwood, Halton i Hammersley (1959) proponowali nieliniową funkcję kosztów, która różni się od wyżej zapisanej tym, że zamiast liczebności $n_h (h=1,\ldots,H)$ bierze się ich pierwiastki. Współczynnik k_h interpretuje się jako koszt jednostkowy dojazdu do obiektów będących elementami h-tej warstwy.

6.2. Lokalizacja proporcjonalna prób w warstwach

Poniżej uogólniono na przypadek estymacji wektorowej znane własności podstawowe proporcjonalnego sposobu wyznaczania liczebności prób przy estymacji pojedynczej wartości średniej.

Przez n oznaczamy sumę liczebności prób losowanych ze wszystkich warstw, czyli $n=\sum_{h=1}^{H} n_h$. Wtedy liczebność $n_{\#h}$ elementów losowanych z h-tej warstwy jest proporcjonalna do frakcji ilości elementów populacji w h-tej warstwie, gdy:

$$n_{\#h} = n w_h, \quad h=1,\ldots,H \qquad (6.9)$$

[36] Cochran (1963), s. 97 sygnalizuje, że to rozwiązanie już wcześniej otrzymał Tschuprow (1923).

Plan losowania prób prostych z tak ustalanymi liczebnościami elementów losowanych z poszczególnych warstw oznaczamy przez P_p. Wtedy na podstawie wzoru (6.5) wnioskujemy, że macierz wariancji i kowariancji strategii estymacji $(\bar{\mathbf{y}}_{ws}, P_p)$ w przypadku bezzwrotnego losowania próby przyjmuje postać:

$$\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_p) = \frac{N-n}{Nn} \sum_{h=1}^{H} w_h \mathbf{C}_{*h} \tag{6.10}$$

W przypadku losowania zwrotnego prób prostych składowych ciągu S mamy:

$$\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_p') = \frac{1}{n} \sum_{h=1}^{H} w_h \mathbf{C}_h \tag{6.11}$$

Jeśli funkcja kosztów obserwacji, dana wzorem (6.8), jest ograniczona kosztem dopuszczalnym K, to na podstawie wzoru (6.9) wnioskujemy, że $n \sum_{h=1}^{H} k_h w_h \le K$. Stąd wynika, że $n \le K \left(\sum_{h=1}^{H} k_h w_h \right)^{-1}$.

Twierdzenie 6.1 [Wywiał (1992)]: Jeżeli liczebności prób prostych losowanych zwrotnie z warstw są ustalane w sposób proporcjonalny do odpowiednich frakcji elementów populacji w warstwach, to strategia estymacji $(\bar{\mathbf{y}}_{wS}, P_w')$ jest nie gorsza od strategii $(\bar{\mathbf{y}}_S, P_1)$.

Wniosek 6.1: Z twiedzeń 2.2 i 6.1 wynika, że miary precyzji estymacji, takie jak średni promień strategii estymacji wektorowej lub jej uogólniona wariancja, osiągają w przypadku strategii $(\bar{\mathbf{y}}_{wS}, P_w')$ wartości nie większe od wartości tych parametrów w przypadku $(\bar{\mathbf{y}}_S, P_1)$.

Twierdzenie 6.1 jest również prawdziwe dla wariantu bezzwrotnego losowania prób, lecz pod warunkiem, że liczebności warstw są dostatecznie duże.

6.3. Optymalizacja liczebności prób na podstawie funkcji kwadratowej ryzyka i funkcji kosztów obserwacji cech

6.3.1. Warunkowa minimalizacja funkcji kwadratowej ryzyka

Analizujemy tu następującą funkcję kwadratową ryzyka estymatora wektorowego $\bar{\mathbf{y}}_{wS}$, która ma postać:

$$f_1(\mathbf{n}) = \sum_{i=1}^{m} a_i D^2(t_{wiS}, P_5) \tag{6.12}$$

gdzie $a_i > 0$ dla każdego i=1,...,m, natomiast $D^2(\bar{y}_{wiS}, P_5)$ określa wzór (6.4). Po prostych przekształceniach otrzymujemy:

$$f_1(\mathbf{n}) = \sum_{h=1}^{H} \frac{w_h^2 b_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{H} w_h b_h \tag{6.13}$$

gdzie:

$$b_h^2 = \sum_{i=1}^{m} a_i v_{*h}(y_i) \tag{6.14}$$

Stąd wnioskujemy, że gdy $a_i = 1$ dla każdego i = 1,...,m, to $f_1(\mathbf{n})$ i jest równy kwadratowi średniego promienia strategii estymacji $(\bar{\mathbf{y}}_{wS}, P_w)$.

Nasze zadanie polega na takim ustaleniu wektora liczebności $\underline{\mathbf{n}} = [\underline{n}_1 ... \underline{n}_H]$, aby przy ustalonych kosztach zmiennych K funkcja ryzyka osiągała minimum, co syntetycznie zapisujemy wyrażeniem:

$$\begin{cases} f_1(\mathbf{n}) = \text{minium} \\ k(\mathbf{n}) \le K, \mathbf{o}_h < \mathbf{n} \le N\mathbf{w} \end{cases} \tag{6.15}$$

Przy czym $k(\mathbf{n})$ jest określoną wzorem (6.8) funkcją kosztów, $\mathbf{o}_H$ jest zerowym wektorem wierszowym o wymiarze $1 \times m$, natomiast $\mathbf{w}$ wektorem frakcji, czyli $N\mathbf{w} = [N_1 ... N_H]$.

Postawione zadanie w nieco ogólniejszej postaci jest rozwiązywane przez Hughesa i Rao (1979). Ich algorytm wyznaczania rozwiązania w naszym przypadku ma następującą postać. Najpierw obliczamy wyrazy:

$$A_h = \frac{b_h}{N\sqrt{k_h}}, h = 1,...,H$$

Przyjmujemy, co nie zmniejsza ogólności wyników, że ciąg $\{A_h\}$ jest nierosnący. Następnie wyznaczamy wielkości:

$$\begin{cases} G_1 = \dfrac{1}{A_1}\sum_{h=1}^{H} w_h b_h \sqrt{k_h} \\ G_{i+1} = \dfrac{1}{A_{i+1}}\left(\sum_{h=i+1}^{H} w_h b_h \sqrt{k_h} \right) + \sum_{h=1}^{i} k_h N_h \end{cases} \tag{6.16}$$

Wskaźniki te liczymy do momentu, gdy $K \in (G_z; G_{z+1})$, gdzie z=0,1,..,H-1 oraz $G_0=0$ i $G_h = \sum_{h=1}^{H} k_h N_h$. Wówczas rozwiązanie optymalne ma postać:

$$\begin{cases} \underline{n}_h = N_h \qquad \text{dla } h=1,...,z \\ \underline{n}_h = c_o \dfrac{w_h b_h}{\sqrt{k_h}} \quad \text{dla } h=z+1,...,H \end{cases} \tag{6.17}$$

gdzie:

$$c_o = \frac{K - \sum_{h=1}^{z} k_h N_h}{\sum_{h=z+1}^{H} w_h b_h \sqrt{k_h}}$$

Wniosek 6.2: Optymalna liczebność próby losowanej z h-tej warstwy jest równa liczebności tej warstwy bądź jest wprost proporcjonalna do iloczynu frakcji w_h i parametru b_h, a odwrotnie proporcjonalna do pierwiastka z kosztu jednostkowego obserwacji w niej zmiennych. Wartość minimalna funkcji kryterium zadania przyjmuje postać:

$$f_1(\underline{\mathbf{n}}) = \left(K - \sum_{h=1}^{z} k_h N_h \right)^{-1} \left(\sum_{h=z+1}^{H} w_h b_h \sqrt{k_h} \right)^2 - \frac{1}{N} \sum_{h=z+1}^{H} w_h b_h^2 \tag{6.18}$$

Każda z liczebności prób będzie mniejsza od ilości elementów warstwy, z której jest losowana (czyli $\underline{n}_h < N_h$ dla każdego h = 1,..,H), jeśli $N_h \to \infty$ dla każdego h = 1,...,H lub

$$N_h > \frac{K}{k_0} + 1 - \frac{1}{k_0} \sum_{h=1}^{H} k_h = n_0 \tag{6.19}$$

gdzie $k_o = \min\limits_{h=1,...,H} \{k_h\}$. Prawa strona nierówności określa maksymalną liczność próby, jaką można uzyskać, gdyby losowano tylko jedną próbę z jednej warstwy przy minimalnym koszcie obserwacji. Założenie (6.19) jest użyteczne, ponieważ znacznie upraszcza rozwiązywanie wielu dalej formułowanych zadań. Ponadto będzie ono w praktyce badań statystycznych spełnione, gdyż zwykle warstwy są bardzo liczne, a koszty dopuszczalne małe.

Jeśli liczba cech m=1 i wektor **a** redukuje się do liczby 1, to otrzymujemy wspomnianą wcześniej lokalizację Neymana (1934) prób w warstwach, czyli liczebności prób, które minimalizują wariancję estymatora wartości średniej zmiennej jednowymiarowej.

Szczególnym przypadkiem funkcji ryzyka $f_1(\mathbf{n})$ jest kwadrat średniego promienia $q^2(\bar{\mathbf{y}}_{wS}, P_w)$, który otrzymujemy przyjmując we wzorze (6.13), że $a_i=1$ dla każdego i=1,...,m. Dodajmy jednak, że użycie parametru $q(\bar{\mathbf{y}}_{wS}, P_w)$ za kryterium optymalizacyjne zadania (6.15) jest nierozsądne, gdy badane zmienne są mierzone w różnych skalach. Wówczas Greń (1964) proponuje zastąpić kwadrat średniego promienia wektora $\bar{\mathbf{y}}_{wS}$ sumą kwadratów współczynników zmienności jego składowych. Podejście to jest równoważne założeniu, że $a_i = \frac{1}{\bar{y}_i^2}$ dla każdego i=1,...,m.

Dodajmy, że J.K.Ghosh (1963) zagadnienie wyboru liczebności prób losowych z warstw traktuje jako grę statystyczną. Wywiał (1992) rozważa minimalizację innej funkcji ryzyka uwzględniającej prawdopodobieństwo przekroczenia błędów estymacji. Ponadto Wywiał (1992) analizuje problem efektywności lokalizacji próby optymalnej względem proporcjonalnej. Wykazuje, że optymalny wariant ustalania liczebności prób daje nie gorsze oceny wartości średnich cech (w sensie średniego promienia estymatora $\bar{\mathbf{y}}_{wS}$) od wariantu proporcjonalnego wyznaczania tych liczebności. W końcu zaznaczmy, że Mukerjee i Rao (1985) wyznaczają kres dolny ilorazu wariancji średniej dla optymalnej lokalizacji próby w warstwach przez wariancję tej średniej wyznaczoną dla dowolnej lokalizacji prób w warstwach. Ograniczenie to nie przekracza liczby jeden i jest tym mniejsze, im bardziej różnią się od siebie ilorazy liczebności prób dowolnie ustalanych przez odpowiadające im liczebności optymalne. Sugeruje się, że ten wynik jest przydatny podczas analizy kompromisowych lokalizacji próby w warstwach.

6.3.2. Minimalizacja kosztów obserwacji cech przy ustalonym ryzyku estymacji

Formułujemy problem minimalizacji kosztów obserwacji cech przy ustalonym poziomie dopuszczalnym funkcji kwadratowej ryzyka za pomocą wyrażenia:

$$\begin{cases} k(\mathbf{n}) = \text{minimum} \\ f_1(\mathbf{n}) \le f_0, \mathbf{o}_h < \mathbf{n} < N\mathbf{w} \end{cases} \tag{6.20}$$

Użyte tutaj symbole wyjaśniono wcześniej przy formułowaniu zadania (6.15). Rozwiązanie tego zadania jest szczególnym przypadkiem rozwiązania ogólniejszego zagadnienia optymalizacyjnego otrzymanego przez Hughesa i Rao (1979). Ograniczymy się więc do przedstawienia algorytmu dochodzenia do rozwiązania zadania (6.20). Najpierw obliczamy wielkości:

$$B_h = \frac{N\sqrt{k_h}}{b_h}, h=1,...H \tag{6.21}$$

przy czym parametr b_h określa wzór (6.14). Przyjmijmy, co nie zmniejsza ogólności wyników, że ciąg $\{B_h\}$ jest niemalejący. Następnie wyznaczmy parametry:

$$F_{i+1} = \frac{1}{B_{i+1}} \sum_{h=i+1}^{H} \sqrt{k_h}\, w_h b_h - \sum_{h=i+1}^{H} \frac{w_h^2 b_h^2}{N_h} \tag{6.22}$$

Liczenie tych wielkości kontynuujemy do momentu, gdy $f_0 \in (F_q ; F_{q+1}]$, q=0,1,...,H-1, przy czym $f_0 = 0$. Wtedy rozwiązanie optymalne ma postać:

$$\begin{cases} \underline{n}_h = N_h & \text{dla } h=1,...,q \\ \underline{n}_h = c_1 \dfrac{w_h b_h}{\sqrt{k_h}} & \text{dla } h=q+1,...,H \end{cases} \tag{6.23}$$

gdzie:

$$c_1 = \frac{\sum_{h=q+1}^{H} \sqrt{k_h}\, w_h b_h}{f_0 + \sum_{h=q+1}^{H} \frac{w_h b_h}{N_h}}$$

Wniosek 6.3: Gdy $f_0 \in (0,F_1]$, to q = 0 i wtedy liczebność próby losowanej z h-tej warstwy jest wprost proporcjonalna do iloczynu $w_h b_h$, a odwrotnie proporcjonalna do pierwiastka z kosztu jednostkowego obserwacji cech w tej warstwie.

6.3.3. Minimalizacja funkcji ryzyka całkowitego

Funkcja ryzyka całkowitego w naszym przypadku jest sumą funkcji kosztów k(**n**) danych wzorem (6.8) i funkcji kwadratowej ryzyka f_1(**n**), danej wzorem (6.13) i ma postać[37]:

$$f_3(\mathbf{n}) = \sum_{h=1}^{H} k_h n_h + \sum_{h=1}^{H} \frac{N_h - n_h}{N_h n_h} w_h^2 b_h^2 \tag{6.24}$$

przy czym parametr b_h^2 określa wzór (6.14).

Nasze zadanie polega na takim wyznaczeniu liczebności $\underline{\mathbf{n}}$, aby wartość funkcji $f_3(\underline{\mathbf{n}})$ była minimalna przy założeniu, że $\mathbf{o}_h < \mathbf{n} < N\mathbf{w}$. Problem ten rozwiązał Yates (1960) dla przypadku losowania zwrotnego prób.

---

[37] Kryterium to zaproponował Blythe (1945), co podajemy za Daleniusem (1957).

Wywiał (1992) otrzymuje liczebności optymalne:

$$\underline{n}_h = \text{minimum}\left\{\frac{w_h b_h}{\sqrt{k_h}}, N_h\right\} \tag{6.25}$$

Wniosek 6.4: Jeśli liczebność optymalna próby losowanej z h-tej warstwy nie jest równa rozmiarowi tej warstwy, to jest ona wprost proporcjonalna do stopnia zróżnicowania cech mierzonych współczynnikiem b_h , a odwrotnie proporcjonalna do pierwiastka z kosztów jednostkowych obserwacji cech.

6.4. Optymalna lokalizacja prób w warstwach przy ustalonej dokładności estymacji poszczególnych wartości średnich

Rozważany teraz problem polega na takim ustaleniu liczebności prób, aby funkcja kosztów obserwacji zmiennych osiągnęła minimum przy z góry ustalonych poziomach błędów dopuszczalnych szacunku średnich poszczególnych cech. Zagadnienie to sformułował Dalenius (1957) i rozwiązał dla przypadku losowania zwrotnego prób z dwóch warstw podczas estymacji średnich dwóch zmiennych. W przypadku ogólnym, lecz dla zwrotnego wariantu losowania prób problem ten rozwiązywali m.in.: Greń (1963) i (1966), Hartley (1965), Huddleston, Claypool i Hocking (1970), Jaganathan (1965) i (1965 a), Kokan (1963), Yates (1960), a w sposób wyczerpujący dla zwrotnego i bezzwrotnego wariantu losowania próby zadanie to rozwiązali Kokan i Khan (1967), które teraz streścimy. Zadanie optymalizacyjne ma postać:

$$\begin{cases} k(\mathbf{n}) = \text{minimum} \\ D^2(y_{wiS}) \le e_i, \quad i = 1,\ldots,m \\ 1 \le n_h \le N_h, \quad h,\ldots,H \end{cases} \tag{6.26}$$

przy czym funkcję kosztów określa wzór (6.8), natomiast wariancję $D^2(.)$ wyrażenie (6.6). Postawione zadanie jest transformowane poprzez wprowadzenie podstawienia $x_h = \frac{1}{n_h}$, h=1,...,H, do postaci:

$$\begin{cases} z(\mathbf{x}) = \text{minimum} \\ \mathbf{u}_i \mathbf{x}^T \le e_{*i}, i = 1,\ldots,m \\ \frac{1}{N_h} \le x_h \le 1, h = 1,\ldots,H \end{cases} \tag{6.27}$$

przy czym:

$$\mathbf{x} = [x_1 \ldots x_H], \qquad z(\mathbf{x}) = \sum_{h=1}^{H} \frac{k_h}{x_h}, \qquad \mathbf{u}_i = \left[w_1^2 v_{*1}(y_i) \ldots w_H^2 v_{*H}(y_i) \right]$$

$$e_{*i} = e_i + \frac{1}{N} \sum_{h=1}^{H} w_h^2 v_{*h}(y_i)$$

Zbiór rozwiązań dopuszczalnych otrzymanego zadania jest simpleksem. Zatem jego rozwiązanie stanowią współrzędne pewnego punktu $\underline{\mathbf{x}}$ wspólnego hiperboloidy $z(\mathbf{x}) = z(\underline{\mathbf{x}})$ ze ścianą, krawędzią bądź wierzchołkiem simpleksu. Kokan i Khan (1967) zaproponowali zbieżny algorytm dochodzenia do rozwiązania.

Zadaniem dualnym do (6.30) zajmował się Chatterjee (1968). Z kolei Bethel (1989) proponuje uproszczoną metodę rozwiązywania tego zadania oraz bada wrażliwość jego rozwiązań na małe zmiany parametrów e_i (i=1,...,m) określających postulowaną dokładność estymacji odpowiednich wartości średnich.

Na zakończenie sformułowany przez J.K. Ghosha (1963) problem optymalizacyjny w pewnym sensie szczególny w stosunku do zadania (6.26) zapisujemy wyrażeniem:

$$\begin{cases} k(\mathbf{n}) = \text{minimum} \\ D^2(\bar{y}_{wiS}) D^{-2}(\bar{y}_{w1S}) \le r_i, i=1,...,m \\ 1 \le n_h \le N_h, h=1,...,H \end{cases} \tag{6.28}$$

Pierwsze (m-1) nierówności określają dopuszczalny poziom wariancji i-tej zmiennej w stosunku do pierwszej. Zatem ograniczenia te określają względny stopień ważności cech. Jeśli pierwsze (m-1) nierówności sprowadzimy do równości oraz przyjmiemy, że $r_i = 1$ dla każdego i=2,..,m, to warunki ograniczające zadania odzwierciedlają żądanie estymacji wektora średnich ze stałą dokładnością. Zasygnalizujmy jeszcze, że jeśli H<m, to tak określone warunki ograniczające zadania mogą być ze sobą sprzeczne.

6.5. Optymalizacja na podstawie promienia spektralnego wektora estymatorów

Kwadrat promienia spektralnego strategii estymacji wektorowej $(\bar{\mathbf{y}}_{wS}, P_w)$ jest równy maksymalnej wartości własnej macierzy wariancji i kowariancji $\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w)$, danej wyrażeniem (6.5). Parametr ten będziemy dalej oznaczać symbolem f(**n**), jako funkcję liczebności $\mathbf{n}=[n_1 \ldots n_H]$. Problem nasz polega na znalezieniu rozwiązania zadania optymalizacyjnego w postaci:

$$\begin{cases} f(\mathbf{n}) = \text{minimum} \\ k(\mathbf{n}) \le K, \quad \mathbf{o}_H < \mathbf{n} \le N\mathbf{w} \end{cases} \tag{6.29}$$

gdzie: $\mathbf{w} = [w_1 \ldots w_H]$.

Funkcję celu możemy zapisać jako maksymalną wartość pierwiastka wielomianu charakterystycznego macierzy $\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w)$, który zapisujemy równaniem:

$$F(\mathbf{n}, f) = \sum_{i=0}^{m} g_i f^i \tag{6.30}$$

gdzie:

$$g_i = \sum_{j=1}^{\binom{m}{i}} g_{ij}, \qquad g_m = 1 \tag{6.31}$$

Współczynnik g_i jest [por. Demidowicz i Maron (1965)] równy sumie wszystkich minorów głównych stopnia (m-i) macierzy $\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w)$, które oznaczono przez g_{ij}.

Problem (6.29) można więc sprowadzić do zagadnienia szukania ekstremum warunkowego funkcji uwikłanej $F(\mathbf{n},f) = 0$, którego rozwiązanie wymaga stosowania jednak dość złożonych metod numerycznych. Dlatego przedstawimy uproszczenie tego zadania[38]. Promień spektralny $f(\mathbf{n})$ jest z góry ograniczony funkcją $r(\mathbf{n})$ w postaci[39]:

$$r(\mathbf{n}) = \sum_{h=1}^{H} w_h^2 \frac{N_h - n_h}{N_h n_h} \rho(\mathbf{C}_{*h}) \tag{6.32}$$

gdzie $\rho^2(\mathbf{C}_{*h})$ jest kwadratem promienia spektralnego macierzy $\mathbf{C}_{*h}$ wariancji i kowariancji cech w h-tej warstwie. Wówczas zadanie optymalizacyjne ma postać:

$$\begin{cases} r(\mathbf{n}) = \text{minimum} \\ k(\mathbf{n}) \le K, \mathbf{o}_h < \mathbf{n} \le N\mathbf{w} \end{cases} \tag{6.33}$$

Rozwiązanie tego zagadnienia otrzymujemy podobnie jak rozwiązanie zadania (6.15), a określa je wyrażenie (6.17), w którym wielkości b_h należy zastąpić odpowiednio przez $\rho(\mathbf{C}_{*h})$, k=1,...,H.

Zadanie polegające na minimalizacji kosztów obserwacji przy ustalonym poziomie dopuszczalnym wartości funkcji $r(\mathbf{n})$ ograniczającej promień spektralny wektora $\bar{\mathbf{y}}_{wS}$ oraz przy założeniu, że $\mathbf{o}_H < \mathbf{n} < \mathbf{N}_w$, rozwiązujemy podobnie, jak problem optymalizacyjny (6.20) zamieniając występujące w nim wielkości b_h na parametr $\rho(\mathbf{C}_{*h})$, h=1,...,H.

Wprowadźmy następującą formę kwadratową macierzy $\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w)$,

$$q(\boldsymbol{\alpha}, \mathbf{n}) = \boldsymbol{\alpha}\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_w)\boldsymbol{\alpha}^T \tag{6.34}$$

---

[38] Zob. Wywiał (1988a).

[39] Wynika to bezpośrednio z własności norm macierzy, którą jest również jej promień spektralny. Por. np. Ralston (1975), s. 427 i 428.

gdzie: $\boldsymbol{\alpha}=[\alpha_1 \ldots \alpha_m]$ jest dowolnym lecz niezerowym wektorem rzeczywistym. Ze znanych twierdzeń algebry liniowej wiadomo, że

$$f(\mathbf{n},\boldsymbol{\alpha})=\max_{\boldsymbol{\alpha}\boldsymbol{\alpha}^T=1}\{q(\boldsymbol{\alpha},\mathbf{n})\} \tag{6.35}$$

Parametr ten, jak już wspominano, jest wariancją kombinacji liniowej $\bar{\mathbf{y}}_{wS}\boldsymbol{\alpha}^T$, przy najmniej korzystnym układzie współczynników $\boldsymbol{\alpha}$, tzn. przy takim, że wariancja kombinacji jest maksymalna.

W celu znalezienia drugiego przybliżonego sposobu rozwiązania zadania (6.29) Wywiał i Kończak (1994) wykorzystują tzw. metodę iteracyjną [por. Demidowicz i Maron (1965)]. Korzystając z wyrażeń (6.34) i (6.35) formułują zadanie:

$$\begin{cases} \min\limits_{\mathbf{n}\in\mathbf{B}} \max\limits_{\boldsymbol{\alpha}\in\mathbf{A}} \{r(\boldsymbol{\alpha},\mathbf{n})\} \\ \mathbf{B}=\{\mathbf{n}:\mathbf{J}\mathbf{n}^T=\mathbf{n}_o>0\} \\ \mathbf{A}=\{\boldsymbol{\alpha}:\boldsymbol{\alpha}\boldsymbol{\alpha}=1\} \end{cases} \tag{6.36}$$

Rozwiązanie zadania przebiega następująco:

$$r(\tilde{\boldsymbol{\alpha}},\tilde{\mathbf{n}})=\max_{\boldsymbol{\alpha}\in\mathbf{A}}\min_{\mathbf{n}\in\mathbf{B}}\{r(\boldsymbol{\alpha},\mathbf{n})\} \tag{6.37}$$

Przy ustalonym wektorze $\boldsymbol{\alpha}$ funkcja $r(\boldsymbol{\alpha},\mathbf{n})$ osiaga minimum w punkcie $\mathbf{n}(\boldsymbol{\alpha})=[n_1(\boldsymbol{\alpha})\ldots n_H(\boldsymbol{\alpha})]$, gdzie:

$$n_h(\boldsymbol{\alpha})=\frac{n_0}{q(\boldsymbol{\alpha})}w_h\sqrt{\bar{q}_h(\boldsymbol{\alpha})},\quad h=1,2\ldots,H \tag{6.38}$$

gdzie:

$$q_h(\boldsymbol{\alpha})=\boldsymbol{\alpha}\mathbf{C}_h\boldsymbol{\alpha}^T,\quad q(\boldsymbol{\alpha})=\sum_{h=1}^{H}w_h\sqrt{q_h(\boldsymbol{\alpha})} \tag{6.39}$$

Wtedy

$$r(\boldsymbol{\alpha},\tilde{\mathbf{n}})=f(\boldsymbol{\alpha})=\frac{1}{n_0}q^2(\boldsymbol{\alpha})$$

Funkcja Lagrange'a ma postać :

$$F(\boldsymbol{\alpha},\boldsymbol{\lambda})=r(\boldsymbol{\alpha},\tilde{\mathbf{n}})-\lambda(\boldsymbol{\alpha}\boldsymbol{\alpha}^T-1) \tag{6.40}$$

Warunek konieczny istnienia ekstremum określają wzory:

$$\frac{\partial F}{\partial\boldsymbol{\alpha}}=\frac{2}{n_0}q(\boldsymbol{\alpha})\frac{\partial q}{\partial\boldsymbol{\alpha}}-\lambda\boldsymbol{\alpha}=0 \tag{6.41}$$

$$\frac{\partial \mathrm{q}}{\partial \boldsymbol{\alpha}} = \boldsymbol{\alpha}\frac{1}{2}\mathbf{G}(\boldsymbol{\alpha}) \tag{6.42}$$

gdzie:

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{h=1}^{H} \frac{w_h}{\sqrt{q_h(\boldsymbol{\alpha})}} \mathbf{C}_h \tag{6.43}$$

Mnożąc równanie (6.41) obustronnie przez $\boldsymbol{\alpha}^{\mathrm{T}}$ mamy:

$$\lambda = \frac{1}{\mathrm{n}_0} \mathrm{q}(\boldsymbol{\alpha})\boldsymbol{\alpha}\mathbf{G}(\mathbf{a})\boldsymbol{\alpha}^{\mathrm{T}} \tag{6.44}$$

Podstawiając ten wynik ponownie do (6.41) mamy:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}\mathbf{P}(\boldsymbol{\alpha}) \tag{6.45}$$

gdzie:

$$\mathbf{P}(\boldsymbol{\alpha}) = \frac{\mathbf{G}(\boldsymbol{\alpha})}{\boldsymbol{\alpha}\mathbf{G}(\boldsymbol{\alpha})\boldsymbol{\alpha}^{\mathrm{T}}} \tag{6.46}$$

Wynik ten prowadzi do następującego wzoru iteracyjnego:

$$\boldsymbol{\alpha}_{\mathrm{t+1}} = \boldsymbol{\alpha}_{\mathrm{t}}\mathbf{P}(\boldsymbol{\alpha}_{\mathrm{t}}), \quad \mathrm{t}=1,2,... \tag{6.47}$$

Otrzymane tą drogą przybliżenie wektora $\tilde{\boldsymbol{\alpha}}$ podstawiamy do wzorów (6.38) i (6.39), które dają oceny liczebności, oznaczane dalej symbolem $\mathrm{n_h}(\boldsymbol{\alpha}_{\mathrm{t}})$. Proces wyliczania wektora $\mathbf{n}_t = [n_1(\boldsymbol{\alpha}_{\mathrm{t}}) \ldots n_H(\boldsymbol{\alpha}_t)]$ zatrzymujemy, gdy odległość między $\mathbf{n}_{\mathrm{t}}$ i $\mathbf{n}_{\mathrm{t=1}}$ jest mniejsza od z góry przyjętej liczby lub gdy dopuszczalna liczba iteracji zostanie przekroczona.

Wywiał i Kończak (1994) porównują lokalizację próby w warstwach przy różnych postaciach macierzy wariancji i kowariancji w próbach. Liczebność prób, które mają być losowane z poszczególnych warstw, otrzymano jako: a) dokładne rozwiązanie poprzez wyliczanie wartości funkcji celu dla wszystkich rozwiązań dopuszczalnych, b) przybliżone rozwiązanie otrzymane ze wzoru (6.47) i c) przybliżone rozwiązanie pomocniczego zadania (6.33). Okazało się, że obie przybliżone metody rozwiązywania zadania dawały podobne lokalizacje prób w warstwach, które nie odbiegają znacznie od dokładnego rozwiązania zadania (6.29) optymalizacji liczebności prób losowanych z warstw.

6.6. Wyznaczanie liczebności prób wykorzystujące cząstkowe lokalizacje optymalne Neymana

Na podstawie wcześniejszych wyników można wyznaczyć optymalną lokalizację próby w warstwach proponowaną przez Neymana (1934) jako szczególny przypadek rozwiązania zadania (6.15). Optymalną lokalizację Neymana wyznaczaną ze względu na i-tą cechę określa wyrażenie (6.17), w którym wielkość b_h należy zastąpić przez odchylenie standardowe i-tej cechy w h-tej warstwie, czyli przez $\sqrt{v_{*h}(y_i)}$, h=1,..,H. Otrzymane w ten sposób wektory liczebności optymalnych oznaczamy przez $\underline{\mathbf{n}}^{(i)}$ (i=1,...,m). Dokonując również zamiany parametrów b_h , h=1,...,H, na $\sqrt{v_{*h}(y_i)}$ we wzorze (6.18) otrzymujemy wartość minimalną wariancji estymatora i-tej cechy (i=1,...,H) przy optymalnym dla niej wariancie lokalizacji próby $\underline{\mathbf{n}}^{(i)}$, którą oznaczamy przez $\underline{f}_{1i} = f_{1i}(\underline{n}^{(i)})$.

Poszukiwaniem rozwiązania kompromisowego problemu lokalizacji próby na podstawie optymalnych rozwiązań cząstkowych $\underline{\mathbf{n}}^{(i)}$ (i=1,...,m) zajmowali się m.in. Dalenius (1953 i 1957), Geary (1949), Greń (1963, 1963a, 1964), Kish (1961), Neyman (1934), Mahalanobis (1944), Srikantan (1963). Proponowane przez nich sposoby optymalizacji liczebności prób polegają na minimalizacji warunkowej określanej funkcji mierzącej stopień rozbieżności między szukanym rozwiązaniem optymalnym a rozwiązaniami cząstkowymi $\underline{\mathbf{n}}^{(i)}$, i=1,..,m.

Greń (1964) formułuje następujący problem[40] :

$$\begin{cases} \sum_{h=1}^{H} \sum_{i=1}^{m} (n_h - \underline{n}_h^{(i)})^2 k_h = \text{minimum} \\ k(\mathbf{n}) \le K, \quad \mathbf{o}_h < \mathbf{n} \le N\mathbf{w} \end{cases} \tag{6.48}$$

Przyjmujac, że $N \to \infty$, można wykazać, iż lokalizację kompromisową prób w warstwach określają wzory:

$$\underline{n}_h = \bar{\underline{n}}_h + \frac{K - \sum_{t=1}^{H} k_t \bar{\underline{n}}_t}{H\bar{k}} \tag{6.49}$$

gdzie:

$$\bar{\underline{n}}_h = \frac{1}{m} \sum_{i=1}^{m} n_h^{(i)}, \qquad \bar{k} = \frac{1}{H} \sum_{h=1}^{H} k_h$$

Prawie wszyscy z wymienionych autorów zajmowali się zagadnieniem określanym wyrażeniem:

---

[40] W oryginalnej postaci problem ten jest określony prościej dla przypadku ustalonej sumy liczebności prób losowanych z warstw.

$$\begin{cases} \sum_{i=1}^{m} \frac{1}{e_i} \\ k(\mathbf{n}) \le K, \quad \mathbf{J}_H^T < \mathbf{n} \le N\mathbf{w} \end{cases} \tag{6.50}$$

gdzie przez e_i oznaczono współczynnik efektywności względnej estymacji wartości średniej i-tej cechy przy dowolnej lokalizacji próby względem lokalizacji optymalnej $\underline{\mathbf{n}}^{(i)}$, który definiuje wyrażenie:

$$e_i = \frac{f_i(\underline{\mathbf{n}}^{(i)})}{D^2(\bar{y}_{wiS})} \tag{6.51}$$

przy czym wariancję $D^2(\bar{y}_{wiS}, P_w)$ określa wyrażenie (6.6), a $f_1(.)$ wzór (6.13), w którym parametry b_h zastępujemy przez $\sqrt{v_{*h}(y_i)}$. Szukaną lokalizację prób otrzymuje się prawie tak samo jak rozwiązanie zadania (6.15), a określa ją wyrażenie (6.17), w którym należy przyjąć, że dla każdego h=1,...,H

$$b_h = f_1^{-1}(\underline{n}^{(i)}) \sum_{i=1}^{m} v_{*h}(y_i) \tag{6.52}$$

Na zakończenie dodajmy, że można tworzyć również inne zadania optymalizacyjne prowadzące do lokalizacji bliskiej rozwiązaniom cząstkowym Neymana[41] . Do tego celu można wykorzystać pewne ogólne metody tworzenia tzw. zadań celowych lub kompromisowych, których własności są studiowane m.in. w pracach Konarzewskiej (1980) lub Galasa, Nykowskiego i Żółkiewskiego (1987).

6.7. Estymacja wektora średnich na podstawie elipsoidy ufności

Elipsoidalny obszar ufności konstruujemy zgodnie z ogólną zasadą wyjaśnioną w rozdziale 2. Przypomnijmy, że przez γ oznaczono poziom ufności, natomiast przez Q_S formę kwadratową postaci:

$$Q_s = (\bar{\mathbf{y}}_{wS} - \bar{\mathbf{y}})\mathbf{V}_S^{-1}(\bar{\mathbf{y}}_{wS})(\bar{\mathbf{y}}_{wS} - \bar{\mathbf{y}})^T \tag{6.53}$$

przy czym przez $\mathbf{V}_S(\bar{\mathbf{y}}_{wS}, P'_w)$ oznaczono macierz wariancji i kowariancji z próby warstwowej losowanej zwrotnie, której elementy wyjaśniają wzory (6.7). Elipsoidalny obszar ufności dla wektora $\bar{\mathbf{y}}$ określa więc wyrażenie:

$$P\{Q_S \le q_\gamma\} = \gamma \tag{6.54}$$

---

[41] Por. np. interesujące zadanie sformułowane i rozwiązywane przez Melaku (1987).

przy czym q_γ jest kwantylem rzędu γ rozkładu formy kwadratowej Q_S. Przy dostatecznie dużych liczebnościach prób prostych losowanych z warstw oraz liczebności tych warstw zmienna losowa Q_S ma rozkład x_m^2 z m stopniami swobody. Wynika to z uogólnionych na przypadek wielowymiarowy twierdzeń Lindeberga-Levy'ego [por. np. Fisz (1967)] i twierdzenia o zbieżności rozkładu prawdopodobieństwa ciągłych funkcji wektorów losowych [por. Rao (1982), s. 143].

6.7.1. Minimalizacja uogólnionej wariancji przy ustalonych kosztach obserwacji cech

Miarą precyzji estymacji na podstawie elipsoidalnego obszaru ufności jest jego objętość, która jest proporcjonalna do m-tej potęgi z pierwiastka uogólnionej wariancji wektora estymatorów $\bar{\mathbf{y}}_{wS}$. Zatem po to, by otrzymać obszar ufności o minimalnej objętości przy ustalonych kosztach dopuszczalnych obserwacji cech, wystarczy minimalizować uogólnioną wariancję, którą oznaczamy przez $f_4(n)=\det\mathbf{V}(\bar{\mathbf{y}}_{wS}, P_5)$. Tak sformułowane przez Daleniusa (1953) zadanie określamy precyzyjniej wyrażeniem:

$$\begin{cases} f_4(\mathbf{n}) = \text{minimum} \\ k(\mathbf{n}) \le K, \quad \mathbf{o}_h < \mathbf{n} \le N\mathbf{w} \end{cases} \tag{6.55}$$

przy czym funkcję kosztów określa wzór (6.8).

Rozwiązaniem postawionego problemu zajmował się S. P. Ghosh (1958), który proponował pewne iteracyjne postępowanie optymalizacyjne, przy czym warunek kosztowy był zastępowany prostszym, określającym dopuszczalną sumę liczebności prób losowanych z warstw. Dowód zbieżności tej metody podali Greń i Koźniewska (1964) lecz dla szczególnego przypadku, gdy h = m = 2 i $Cov(\bar{\mathbf{y}}_{w1S}, \bar{\mathbf{y}}_{w2S}, P_5) = 0$.

Metodę gradientową[42] do poszukiwania rozwiązania zadania stosowali Arwanitis i Afonia (1971) dla dowolnej liczby warstw, lecz przy jednoczesnej estymacji najwyżej trzech wartości średnich zmiennych. Zanim opiszemy algorytm poszukiwania rozwiązania uproszczonej wersji postawionego zadania dla dowolnej ilości cech, wykażemy najpierw twierdzenie.

Lemat 6.1 [Wywiał (1989) i (1992)]: Jeśli przynajmniej jedna z macierzy wariancji i kowariancji wewnątrzwarstwowej $\mathbf{C}_{*h}$ (h=1,..,H) jest dodatnio określona, to uogólniona wariancja $f_4(\mathbf{n})$ jest ściśle wypukła w obszarze $\mathcal{D}_b$ dla bezzwrotnego wariantu losowania prób prostych z warstw, natomiast w obszarze $\mathcal{D}_z$ dla przypadku losowania zwrotnego, przy czym:

$$\mathcal{D}_b = \{\mathbf{n}: \mathbf{o}_H < 2\mathbf{n} < N\mathbf{w}\} \tag{6.56}$$

---

[42] Różne odmiany tej metody można znaleźć w pracach opisujących metody numeryczne lub programowanie nieliniowe: Demidowicz i Maron (1965), Dryja i Jankowscy (1988), Findeisen, Szymanowski i Wierzbicki (1967), Grabowski (1982), Kręglewski, Rogowski, Ruszczyński i Szymanowski (1984), Martos (1983), Rice (1983), Wit (1986).

$$\mathcal{D}_z = \{\mathbf{n}:\ \mathbf{n} > \mathbf{o}_H\} \tag{6.57}$$

Wywiał (1992) wykazuje, że rozwiązaniem optymalnym analizowanego zadania są pierwiastki następującego układu równań:

$$\begin{cases} \dfrac{1}{k_h}\dfrac{\partial f_4(\mathbf{n})}{\partial n_h} = \dfrac{1}{k_h}\dfrac{\partial f_4(\mathbf{n})}{\partial n_H}, h=1,...,H-1 \\ k(\mathbf{n}) = K \end{cases} \tag{6.58}$$

Pierwsze pochodne cząstkowe funkcji f_4 można wyliczyć na podstawie wzoru ogólnego na pochodną wyznacznika macierzy zamieszczonego w pracy Kubika i Krupowicza (1982), s. 445-446, który w naszym przypadku ma postać[43]:

$$\frac{\partial f_4(n)}{\partial n_h} = -\frac{w_h^2}{n_h^2}\sum_{i=1}^{m}\det\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ . & . & \cdots & . \\ c_{hi1} & c_{hi2} & \cdots & c_{him} \\ . & . & \cdots & . \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{bmatrix} \tag{6.59}$$

przy czym przez v_{ij} oznaczono elementy macierzy $\mathbf{V}(\bar{\mathbf{y}}_{wS}, P'_w)$, natomiast przez $c_{hij}=c(y_i,y_j)$ elementy macierzy $\mathbf{C}_h$ wariancji i kowariancji cech w h-tej warstwie.

6.7.2. Minimalizacja kosztów obserwacji przy ustalonym poziomie dopuszczalnym uogólnionej wariancji

Przypuśćmy, że określono dopuszczalną objętość elipsoidalnego obszaru ufności, co determinuje poziom uogólnionej wariancji wektora estymatorów $\bar{\mathbf{y}}_{wS}$. Pojawia się więc teraz problem takiego ustalenia liczebności $\underline{\mathbf{n}}$ warstw, aby minimalizowały koszty obserwacji zmiennych w próbach przy ustalonym poziomie uogólnionej wariancji. Zagadnienie to dla przypadku losowania zwrotnego prób prostych bądź losowania bezzwrotnego takich prób z warstw o nieograniczonej liczbie elementów w sposób pełny określa wyrażenie:

$$\begin{cases} k(\mathbf{n}) = \text{minimum} \\ f_4(\mathbf{n}) \le f_o, \quad \mathbf{o}_h < \mathbf{n} \end{cases} \tag{6.60}$$

przy czym liniową funkcję kosztów określa wzór (6.6), natomiast $f_4(\mathbf{n})=\det \mathbf{V}(\bar{\mathbf{y}}_{wS})$, gdzie macierz $\mathbf{V}(\bar{\mathbf{y}}_{wS})$ jest dana wzorem (6.5).

---

[43] Inny sposób liczenia tych pochodnych proponuje Wywiał (1989) lub (1992).

Wywiał (1992) otrzymuje następujący układ równań:

$$\frac{1}{mf_o} f'_{4h}(\underline{\mathbf{n}}) = \frac{k_h}{k(\underline{\mathbf{n}})}, \qquad h = 1,\ldots,H \tag{6.61}$$

Rozwiązanie otrzymanego układu równań jest szukanym wektorem liczebności optymalnych $\underline{\mathbf{n}}$. Niestety układ ten jest nieliniowy, a więc do jego rozwiązania trzeba użyć odpowiednich metod numerycznych.

6.8. Warstwowanie populacji

W poprzednich paragrafach uzasadniono, że estymacja wartości średnich na podstawie próby losowanej z warst jest na ogół efektywniejsza niż np. ocena na podstawie próby prostej. Precyzja estymacji była znaczna zwłaszcza wtedy, gdy zróżnicowanie cech wewnątrz poszczególnych warstw było małe, a między nimi duże. Zatem warstwy w populacji należy tak wyróżniać aby rozproszenie cech było małe wewnątrz nich. Nie zawsze jest to możliwe, ponieważ w praktyce zazwyczaj mamy do czynienia z naturalnym podziałem populacji na warstwy lub podziałem wynikającym z dostępności operatu losowania. Najczęściej są tworzone operaty losowania, które są listami elementów warstwy. W tym przypadku odpada zmartwienie związane z poprawnym tworzeniem warstw, gdyż one już są dane.

W sytuacji, gdy istnieje możliwość tworzenia warstw, Kish (1965) podaje następujące uwagi praktyczne dotyczące ich konstruowania: Warstwy w populacji wyróżnia się zwykle na podstawie informacji dodatkowych o badanym rozkładzie cech, a są one bądź jakościowymi, bądź ilościowymi zmiennymi określonymi na elementach populacji. Z racji tego, że te cechy są wykorzystywane do faktycznego wyróżniania warstw, nazywane są zmiennymi warstwującymi. Postuluje się, aby była dostępna informacja o wzajemnie jednoznacznym przyporządkowaniu elementów populacji i wartości cech warstwujących. Jeśli zdarzy się, że w pewnym podzbiorze populacji nie jest dokładnie znane przyporządkowanie między wartościami zmiennych porządkujących i elementami populacji, to podzbiór ten należy potraktować jako osobną warstwę. Gdy operat losowania jest sumą rozłącznych list ze spisami elementów populacji, to zbiory odpowiadające tym listom można uznać za warstwy populacji.

Podczas wyboru zmiennych warstwujących lub przy tworzeniu warstw nie jest wymagana ani regularność, ani obiektywność. Zaleca się korzystać z opinii wielu niezależnych specjalistów zajmujących się badanym problemem. W związku z tym preferuje się mniej formalną procedurę warstwowania na podstawie wskazywanych przez ekspertów wielu zmiennych, zamiast wyczerpującej i niejednokrotnie finezyjnej z metodologicznego punktu widzenia analizy wykorzystującej tylko jedną zmienną. Proponuje się wybierać tylko jedną spośród dwóch silnie skorelowanych cech warstwujących. Preferuje się cechy warstwujące słabo zależne liniowo między sobą, lecz silnie związane ze zmiennymi, których parametry są przedmiotem estymacji. Obserwowane cechy pomocnicze mogą być także wykorzystane do podniesienia dokładności ocen parametrów poprzez użycie ich jako zmiennych wspomagających, np. w estymatorach regresyjnych i ilorazowych.

Oprócz wymaganego postulatu jednorodności rozkładu cech wewnątrz tworzonych warstw, należy dążyć do tego, by wartości średnie zmiennych między warstwami były możliwie silnie zróżnicowane, gdy jest stosowany wariant proporcjonalnego ustalania liczebności prób losowanych z warstw. W przypadku wariantu optymalnego wyliczania rozmiarów tych prób pożądane jest duże zróżnicowanie miar rozproszenia wewnątrzwarstwowego.

W przypadku silnie zróżnicowanego rozkładu cech w populacji zwykle zmierza się do tworzenia dużej liczby warstw. Gdy z każdej warstwy jest losowany tylko jeden element, to pojawia się kłopot z oceną wariancji estymatorów wartości średnich, który jest rozwiazywany przez łączenie równolicznych warstw, czym zajmuje się m.in. Cochran (1963).

Jeśli warstwowanie populacji może się okazać przedsięwzięciem zbyt kosztownym, to można zastosować znaną procedurę tzw. warstwowania po wylosowaniu próby. Polega to na na tym, że na podstawie wybranych zmiennych pomocniczych jest warstwowana realizacja próby, którą wylosowano z całej populacji. Z tak wyodrębnionych warstw w próbie pierwotnej w następnym etapie losowane są próbki, w których dopiero obserwuje się zmienne, których parametry są przedmiotem właściwego wnioskowania.

Dalenius (1957), s. 159-194 przywiązuje dużą wagę do problemu warstwowania populacji, ponieważ od własności rozkładu cech w warstwach w znacznej mierze zależy jakość dalszego wnioskowania statystycznego. Kwestię tę uznaje za tak ważną, że dobry podział populacji na warstwy utożsamia z prawidłowo postawionym problemem badawczym, zgadzając się z Robbinsonem (1952), że zwykle lepiej mieć rozsądne rozwiązanie dobrze postawionego problemu niż rozwiązanie optymalne źle sformułowanego problemu[44]. Dalenius podchodzi do zagadnienia warstwowania populacji w sposób bardziej formalny wyróżniając kwestie: a) wybór cech warstwujących, b) wybór liczby warstw, c) określenie sposobu warstwowania, d) wybór optymalny liczebności prób z uwzględnieniem wcześniejszych trzech problemów. W niniejszej pracy nie zajmujemy się problemem warstwowania przez podział obszaru zmienności cechy badanej. Tym zagadnieniem zajmowali się m.in. Bracha (1991), Cochran (1963), Dalenius (1957), Hess, Sethi i Balakrishnan (1966), Jonin, Jonina i Żurawlew (1978), Serfling (1968).

6.8.1. Warstwowanie po wylosowaniu próby

Problemem warstwowania po wylosowaniu próby najpierw zajmował się Neyman (1938)[45], następnie Cochran (1963). Niżej problem ten referujemy opierając się głównie na wynikach zaczerpniętych z artykułu J.N.K.Rao (1973).

Nich S' będzie próbą prostą o liczebności n' losowaną z populacji. W tak otrzymanej próbie są identyfikowane elementy populacji należące do poszczególnych warst poprzez obserwację odpowiednich cech warstwujących. Niech S'_h będzie zbiorem elementów h-tej warstwy populacji zidentyfikowanych w próbie S', czyli $s'_h \in \Omega_h$. Przez $o \leq n_{S'_h} \leq n'$

[44] W oryginale: "[...] it is often better to have reasonably good solution of the proper problems than optimum solution of the wrong problems".

[45] Wiadomość tę podajemy za Cochranem (1963), s. 328.

(h=1,...,H), $\sum_{h=1}^{H} n_{S'_h} = n'$ oznaczamy liczebności zbioru S'_h. Niech $w'_h = \frac{1}{n'} n_{S'_h}$, natomiast $w_h = \frac{N_h}{N}$ dla h=1,..,H. Z kolei przez S_h (h=1,..,H) oznaczamy próbę prostą o liczebności n_{S_h} losowaną bezzwrotnie spośród tych elementów zbioru s'_h o liczności $n'_h \geq 1$, które należą do h-tej warstwy. Niech $n_{S_h} = \nu_h n_{S'_h}$, gdzie $0 < \nu_h \leq 1$ dla h=1,..,H oraz $\sum_{h=1}^{H} n_{S_h} = n$. Określony plan losowania próby $S=\{S',S_1,...,S_2\}$ oznaczymy przez P'.

Wszystkie niżej prezentowane wyniki otrzymano przy założeniu, że rozmiar n' próby S' jest na tyle duży, że $P\left\{n_{S'_h}, > 0\right\} = 1$ dla każdego h=1,...,H.

Do estymacji wartości średniej $\bar{y} = \sum_{h=1}^{H} w_h \bar{y}_h$ w populacji jest używana statystyka[46]:

$$\overset{V}{y}_{wS} = \sum_{h=1}^{H} w'_h \bar{y}_{S_h} \tag{6.62}$$

Strategia estymacji $(\overset{V}{y}_{wS}, P')$ daje nieobciążone oceny średniej $\bar{y}$, natomiast jej wariancja ma postać:

$$D^2(\overset{V}{y}_{wS}, P') = \sum_{h=1}^{H} \frac{w_h \nu_{*h}}{n' \nu_h} - \sum_{h=1}^{H} \frac{E(w'_h)^2 \nu_{*h}}{N_h} + \frac{g' B}{n'} \tag{6.63}$$

gdzie:

$$B = \sum_{h=1}^{H} w_h (\bar{y}_h - \bar{y})^2, \qquad g' = \frac{N - n'}{N - 1}$$

lub

$$D^2(\overset{V}{y}_{wS}, P') = \left(\frac{1}{n'} - \frac{1}{N}\right) \nu_* + \sum_{h=1}^{H} \frac{w_h \nu_{*h}}{n'} \left(\frac{1}{\nu_h} - 1\right)$$

Jeśli plan P'_p daje proporcjonalną lokalizację prób $\{S_h\}$ w realizacji próby S', czyli gdy $\nu_h = \frac{n'}{n}$ dla każdego h=1,...,H, to w przybliżeniu [Cochran (1963), s. 330]:

$$D^2(\overset{V}{y}_{wS}; P'_p) \approx \frac{1}{n'} \sum_{h=1}^{H} \nu_{*h} + \frac{g' B}{n'} \tag{6.64}$$

---

[46] Bracha (1991) wykazuje, że statystyka $\overset{V}{y}_{wS}$ da się zapisać jako estymator regresyjny średniej $\bar{y}$ zależny od zero-jedynkowych cech dodatkowych identyfikujących warstwy w populacji skończonej.

Jeśli n' jest na tyle duże, że $P\left\{n_{S_h'} \geq 2\right\}=1$ dla każdego h=1,...,H, to nieobciążonym estymatorem wariancji $D^2(\overset{V}{y}_{wS}, P')$ jest statystyka:

$$d_S^2 = \frac{N-1}{N}\sum_{h=1}^{H}\left(\frac{n_{S_h'}-1}{n'-1}-\frac{n_h-1}{N-1}\right)\frac{w_h'}{n_h}v_{*S_h}+\frac{N-n'}{N(n'-1)}\sum_{h=1}^{H}w_h'(\bar{y}_{S_h}-\overset{V}{y}_{wS})^2$$

lub:

$$d_S^2 = \frac{1}{Nn'}\left\{\frac{N-1}{n'-1}\sum_{h=1}^{H}n_{S_h'}v_{*S_h}\left(\frac{1}{v_h}-1\right)+\frac{N-n'}{n'-1}\left(\sum_{h=1}^{H}\frac{1}{v_h}\sum_{k\in S_h}y_{hk}^2-n'\overset{V2}{y}_{wS}\right)\right\}$$

Przy założeniu, że z góry są znane wszystkie frakcje w_h elementów populacji w poszczególnych warstwach, Cochran (1963), s. 135 proponuje procedurę estymacji polegającą na pominięciu etapu wyboru podprób z powarstwowanej próby prostej $S'=\bigcup_{h=1}^{H}S_h'$ losowanej na początku z populacji. Używa do tego celu estymatora:

$$\overset{W}{y}_{wS}=\sum_{h=1}^{H}w_h\bar{y}_{S_h'} \tag{6.65}$$

Statystyka ta daje nieobciążone oceny średniej $\bar{y}$ a przybliżony wzór na jej wariancję ma postać:

$$D^2(\overset{W}{y}_{wS})=\frac{N-n'}{Nn'}\sum_{h=1}^{H}w_hv_{*h}+\frac{1}{(n')^2}\sum_{h=1}^{H}(1-w_h)v_{*h} \tag{6.66}$$

Rao (1973) również formułuje i rozwiązuje zadanie optymalizacyjne dotyczące ustalania rozmiarów prób składowych próby $S=\{S',S_1,...,S_H\}$.

6.8.2. Wykorzystanie metod grupowania danych do warstwowania populacji

Rozważmy następujący model regresyjny nadpopulacji:

$$Y_k=\alpha_0+\boldsymbol{\alpha}\mathbf{x}_k+U_k,\ k=1,...,N \tag{6.67}$$

przy czym:

$$\boldsymbol{\alpha}=[\alpha_1 K\ \alpha_z],\ \boldsymbol{\alpha}\boldsymbol{\alpha}^T=1,\ \mathbf{x}_k^T=[x_{1k}K\ x_{zk}]$$

$$\mathcal{E}\left(U_k|\mathbf{x}_k\right)=0,\ \mathcal{E}\left(Y_k|\mathbf{x}_k\right)=y_k,\ \mathcal{D}^2\left(U_k|\mathbf{x}_k\right)=\sigma^2,\ Cov\left(U_k,U_j|\mathbf{x}_k,\mathbf{x}_j\right)=0$$

gdzie $k\neq j=1,...,N$.

Średnią $\overline{Y}$ cechy w nadpopulacji oceniamy przy pomocy strategii predykcji $\left(\overline{Y}_{wS}, P_w\right)$, gdzie:

$$\overline{Y}_{wS} = \sum_{h=1}^{H} w_h \overline{Y}_{hS}, \qquad \overline{Y}_{hS} = \frac{1}{n_h} \sum_{k \in S} y_{hk} \tag{6.68}$$

Błąd średniokwadratowy wyliczono na podstawie wzoru (6.6):

$$\mathcal{E}D^2(\overline{Y}_{wS}, P_5) = \sum_{h=1}^{H} w_h^2 \frac{\mathcal{E}[v_{*h}(y)]}{n_h} - \frac{1}{N} \sum_{h=1}^{H} w_h \mathcal{E}[v_{*h}(y)] \tag{6.69}$$

Przyjmijmy, że ogólny rozmiar próby jest ograniczony i wynosi $n \le N_h$ dla każdego h=1,...,H. Wtedy jeśli liczebności prób są losowane z warstw w sposób proporcjonalny do ich liczebności, czyli $n_h = n w_h$, to:

$$\mathcal{E}D^2(\overline{Y}_{wS}, P_\rho) = \frac{N-n}{Nn} \sum_{h=1}^{H} w_h \mathcal{E}[v_{*h}(y)] \tag{6.70}$$

Po odpowiednich przekształceniach otrzymujemy [por. Wywiał (1992)]:

$$\mathcal{E}D^2(\overline{Y}_{wS}, P_\rho) = \frac{N-n}{Nn} [\rho(\mathbf{X},\boldsymbol{\alpha}) + \sigma^2] \tag{6.71}$$

gdzie:

$$\rho(\mathbf{X},\alpha) = \boldsymbol{\alpha} \sum_{h=1}^{H} w_h \mathbf{C}_{*h} \boldsymbol{\alpha}^T \tag{6.72}$$

przy czym $\mathbf{C}_{*h}$ jest macierzą wariancji i kowariancji cech dodatkowych w h-tej warstwie. Z kolei $\mathbf{X}=[\mathbf{X_1}...\mathbf{X_H}]$ jest macierzą blokową obserwacji cech wspomagających w poszczególnych warstwach. Podmacierz $\mathbf{X}_h$ o wymiarach z $\times N_h$ składa się z N_h kolumn obserwacji z-wymiarowej zmiennej wspomagającej w h-tej warstwie. Kolumny macierzy $\mathbf{X}_h$ są więc wzajemnie jednoznacznie przyporządkowane elementom populacji tworzącym h-tą warstwę. Zatem macierz blokowa $\mathbf{X}$ reprezentuje podział na warstwy populacji.

Ze wzoru (6.71) wynika, że błąd średniokwadratowy predyktora $\overline{Y}_{wS}$ przy proporcjonalnej lokalizacji prób prostych w warstwach jest najmniejszy, gdy funkcja $\rho(\mathbf{X},\mathbf{a})$ osiąga minimum. Wektor parametrów $\boldsymbol{\alpha}$ nie jest jednak znany. Najrozsądniej będzie więc przyjąć taką wartość wektora $\boldsymbol{\alpha}=\underline{\boldsymbol{\alpha}}$, dla którego funkcja $\rho(\mathbf{X},\underline{\mathbf{a}})$ = maximum, przy ustalonym podziale populacji na warstwy. Należy zatem szukać takiego podziału populacji reprezentowanego macierzą blokową $\mathbf{X}$ oraz wektora $\underline{\boldsymbol{\alpha}}$, aby był spełniony warunek:

$$\rho(\underline{\mathbf{X}},\underline{\boldsymbol{\alpha}}) = \underset{\mathbf{X} \in \mathcal{X}}{\text{minimum}} \; \underset{\alpha\alpha^T=1}{\text{maximum}} \{\rho(\mathbf{X},\boldsymbol{\alpha})\} \tag{6.73}$$

przy czym $\mathcal{X}$ jest zbiorem wszystkich dopuszczalnych podziałów populacji na warstwy. Na podstawie definicji 1.8 promienia spektralnego rozkładu zmiennej wielowymiarowej i

twierdzeń podanych przez Rao (1982), s. 81, wnioskujemy, że parametr $\rho(\mathbf{X},\underline{\boldsymbol{\alpha}})$ jest promieniem spektralnym macierzy wariancji i kowariancji wewnątrzwarstwowej, którą zapisujemy jako macierz $\mathbf{C}_w(\mathbf{X})=\sum_{h=1}^{H} w_h \mathbf{C}_{*h}$ formy kwadratowej (6.72). Zatem parametr $\rho(\mathbf{X},\underline{\boldsymbol{\alpha}})$ jest maksymalną wartością własną macierzy $\mathbf{C}_w(\mathbf{X})$.

Wniosek 6.5: Gdy populacja zostanie tak podzielona na warstwy, że promień spektralny macierzy wariancji i kowariancji wewnątrzwarstwowej cech pomocniczych osiągnie wartość minimalną, to błąd średniokwadratowy strategii $(\overline{Y}_{wS}, P_p)$ jest najmniejszy przy najmniej korzystnym układzie unormowanych parametrów $\boldsymbol{\alpha}=[\alpha_1...\alpha_z]$ funkcji regresji modelu nadpopulacji.

Procedurę grupowania populacji na warstwy według kryterium (6.73) można skonstruować poprzez adaptację algorytmu grupowania populacji minimalizującego uogólnioną wariancję wektora cech, którą proponował Friedman i Rubin (1967). Ich procedura polega na przemieszczaniu kolejno każdego elementu populacji z jednej warstwy do innej, aż do uzyskania pożądanej wartości funkcji kryterium. Załóżmy, że numery kolumn macierzy obserwacji **X,** będące jednocześnie numerami elementów populacji, nie zmieniaja się podczas przemieszczania ich z jednej warstwy do innej. Algorytm wyróżniania warstw rozpoczynamy od arbitralnego podziału populacji na warstwy, który oznaczamy przez $\mathbf{X}^{(0)}$. Każda następna iteracja algorytmu jest dzielona na N operacji. Przez $\mathbf{X}^{(e,k)}$ oznaczamy stan podziału populacji na warstwy otrzymany w wyniku k-tej operacji e-tej iteracji, podczas której należy najpierw znaleźć numer warstwy zawierającej k-ty element populacji. Przypuśćmy, że tym numerem jest indeks ν. Następnie k-ty element populacji reprezentowany obserwacją $\mathbf{x}_k$ przemieszczamy z ν-tej warstwy kolejno do pozostałych warstw reprezentowanych macierzami $\mathbf{X}_h$ (h=1,..,H), za każdym razem wyliczając wartość funkcji:

$$\rho(\mathbf{x}^{(e,k)}, \underline{\boldsymbol{\alpha}} | \mathbf{x}_k \in \mathbf{X}_h) = \underset{\boldsymbol{\alpha}\boldsymbol{\alpha}^T=1}{\text{maximum}} \left\{ \rho(\mathbf{X}^{(e,k)}, \boldsymbol{\alpha} | \mathbf{x}_k \in \mathbf{X}_h) \right\} \tag{6.74}$$

co sprowadza się do obliczenia maksymalnej wartości własnej macierzy $\mathbf{C}_w(\mathbf{X}^{(e,k)})$. Potem wyznaczamy tak wskaźnik t warstwy aby:

$$\rho(\mathbf{x}^{(e,k)}, \underline{\boldsymbol{\alpha}} | \mathbf{x}_k \in \mathbf{X}_\tau) = \min_{h=1,...,H} \text{imum} \left\{ \rho(\mathbf{X}^{(e,k)}, \boldsymbol{\alpha} | \mathbf{x}_k \in \mathbf{X}_h) \right\} \tag{6.75}$$

Jeśli $\tau=\nu$, to k-ty element populacji pozostaje w ν-tej warstwie, natomiast gdy $\tau\neq\nu$, to k-ty element populacji należy przemieścić z ν-tej wartwy do τ-tej. Na tym kończy się k-ta operacja i należy przejść do (k+1) operacji powtarzając wyżej opisane czynności. Gdy wskaźnik k osiągnie wartość N, to kończy się e-ta iteracja. Wtedy ewentualnie możemy rozpocząć (e+1) iterację ponownie, ustalając wskaźnik operacji na poziomie k=1. Dodajmy, że jeśli podczas danej operacji natrafimy na zbiór jednoelementowy, to nie przemieszczamy go do żadnej innej warstwy, lecz przechodzimy do następnej operacji. Ta decyzja nie dopuści do wystąpienia zbioru pustego podczas procesu grupowania.

Opisaną procedurę zatrzymujemy, jeśli napotkamy iterację, podczas której nie wystąpi żadne przemieszczenie elementu z jednej warstwy do innej. Druga reguła stopu algorytmu polega na arbitralnym ustaleniu dopuszczalnych liczby iteracji.

Promień spektralny macierzy wariancji i kowariancji wewnątrzgrupowej cech pomocniczych jest majoryzowany przez ślad tej macierzy. Zatem warstwowanie populacji minimalizujące ślad tej macierzy można traktować jako procedurę pokrewną względem wyżej opisanej. W tym przypadku do tworzenia warstw można wykorzystać ogólnie znany algorytm grupowania zwany metodą k-średnich. Ponadto można używać metody Warda (1963), najbliższego sąsiada itd[47].

6.8.3. Warstwowanie po wylosowaniu próby poprzez jej grupowanie

Korzystając z oznaczeń wprowadzonych w uprzednim punkcie załóżmy, że na elementach wstępnej próby prostej S' o liczebności n wylosowanej z populacji obserwujemy zespół cech pomocniczych w próbie. Ich związek ze zmienną badaną Y opisuje model regresyjny wprowadzony w uprzednim punkcie. Na podstawie tych cech grupujemy zbiór S' na H rozłącznych, wyczerpujących zbiór S' podzbiorów S'_h , h=1, ..., H o liczności n'_h. Do tego celu można wykorzystać wybrany algorytm grupowania. Przez S_h, h=1,...,H oznaczamy próby proste losowane bezzwrotnie z warstwy S'_h .

Do predykcji wartości średniej w populacji $\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k$ użyjemy statystyki:

$$\tilde{\overline{Y}}_{ws} = \sum_{h=1}^{H} w'_h \overline{Y}_{S_h} \tag{6.76}$$

gdzie:

$$w'_h = \frac{n'_h}{n'}, \qquad \overline{Y}_{S_h} = \frac{1}{n_h}\sum_{k\in S_h} Y_k$$

Wartość oczekiwaną predyktora $\tilde{\overline{Y}}_{wS}$ wyznaczamy nastepująco:

$$E(\tilde{\overline{Y}}_{wS}) = E_{S'}E_{S/S'}(\tilde{\overline{Y}}_{wS})$$

$$E_{S/S'}(\tilde{\overline{Y}}_{wS}) = \sum_{h=1}^{H}\frac{n'_h}{n'}E_{S/S'}(\overline{Y}_{S_h}) = \sum_{h=1}^{H}\frac{n'_h}{n'}\overline{Y}_{S'_h} = \overline{Y}_{S'}$$

$$E(\tilde{\overline{Y}}_{wS}) = E_{S'}(\overline{Y}_{S'}) = \overline{Y} \tag{6.77}$$

Zatem predyktor $\tilde{\overline{Y}}_{wS}$ jest p-nieobciążony.

Błąd średniokwadratowy predykcji wyznaczamy następująco:

---

47 Opis tych procedur można znaleźć m.in. w pracach Grabińskiego i in. (1989) oraz Kolonki (1980).

$$\mathsf{E}D^2(\tilde{Y}_{wS}) = \mathsf{E}E(\tilde{Y}_{wS} - \overline{Y})^2 = \mathsf{E}E_{S'}E_{S/S'}(\tilde{Y}_{wS} - \overline{Y})^2 \tag{6.78}$$

$$E_{S/S'}(\tilde{Y}_{wS} - \overline{Y})^2 = E_{S/S'}[(\tilde{Y}_{wS} - \overline{Y}_{S'}) + (\overline{Y}_{S'} - \overline{Y})]^2 =$$

$$= E_{S/S'}(\tilde{Y}_{wS} - \overline{Y}_{S'})^2 + 2(\overline{Y}_{S'} - \overline{Y})E_{S/S'}(\tilde{Y}_{wS} - \overline{Y}_{S'}) + (\overline{Y}_{S'} - \overline{Y})^2$$

$$E_{S/S'}(\tilde{Y}_{wS} - \overline{Y})^2 = D^2_{S/S'}(\tilde{Y}_{wS}|S' = s') + (\overline{Y}_{S'} - \overline{Y})^2 \tag{6.79}$$

gdzie z uwagi na to, że próby proste S_h, h=1,...,H są losowane z odpowiednich ustalonych zbiorów (warstw) s_h', mamy:

$$D^2_{S/S'}(\tilde{Y}_{wS}|S' = s') = \sum_{h=1}^{H}(w'_h)^2 \frac{n'_h - n_h}{n'_h n_h} v_{*S'_h}(Y|s') \tag{6.80}$$

Przyjmijmy, że liczebności n_h prób s_h, h=1,...,H są ustalone proporcjonalnie do frakcji $w'_h = \frac{n'_h}{n'}$, a zatem $n_h = w'_h n$. Wtedy:

$$D^2_{S/S'}(\tilde{Y}_{wS}|S' = s') = \frac{n' - n}{n' n}\sum_{h=1}^{H} w'_h v_{*S'_h}(Y|s') \tag{6.81}$$

Wariancje $v_{*S'_h}(Y|s')$ zależą od wylosowanej próby s' i otrzymanego w wyniku grupowania podziału jej na warstwy $s'_1,...,s'_H$. Zatem:

$$E_{S'}D^2_{S/S'}\left(\tilde{Y}_{wS}|S' = s'\right) = \frac{1}{\binom{N}{n'}}\frac{n' - n}{n' n}\sum_{s' \in \mathbf{S}}\sum_{h=1}^{H} w'_h v_{*S'_h}\left(Y|s'\right) \tag{6.82}$$

gdzie S jest przestrzenią prób typu s'. Stąd i ze wzoru (6.79) dla proporcjonalnego wariantu próby warstwowej mamy:

$$D^2_p\left(\tilde{Y}_{wS}\right) = \frac{1}{\binom{N}{n'}}\frac{n' - n}{n' n}\sum_{s' \in S}\sum_{h=1}^{H} w'_h v_{*S'_h}\left(Y|s'\right) + \frac{N - n'}{Nn'}v_*(Y) \tag{6.83}$$

Przyjmując założenia regresyjnego modelu nadpopulacji mamy:

$$E\left[v_{*S'_h}\left(Y|s'\right)\right] = \boldsymbol{\alpha}\mathbf{C}_{*h}(s')\boldsymbol{\alpha}^T + \sigma^{2'}, \qquad \mathcal{E}\left[v_*\left(Y|s'\right)\right] = \boldsymbol{\alpha}\mathbf{C}_*\boldsymbol{\alpha}^T + \sigma^2$$

Stąd i ze wzoru (6.83) mamy:

$$\mathrm{ED}_p^2(\tilde{\bar{Y}}_{wS}) = \frac{1}{\binom{N}{n'}} \frac{n'-n}{n'n} \sum_{s' \in S} \sum_{h=1}^{H} w'_h \boldsymbol{\alpha} \mathbf{C}_{*h}(s') \boldsymbol{\alpha}^T + \frac{N-n'}{Nn'} \boldsymbol{\alpha} \mathbf{C}_* \boldsymbol{\alpha}^T - \frac{N-n}{Nn} \sigma^2 \quad (6.84)$$

lub

$$\mathrm{ED}_p^2\left(\tilde{\bar{Y}}_{wS}\right) = \frac{1}{\binom{N}{n'}} \sum_{s' \in S} \psi(s') - \frac{N-n}{Nn} \sigma^2 \quad (6.85)$$

gdzie:

$$\psi(s') = \boldsymbol{\alpha} \left(\frac{n'-n}{n'n} \sum_{s' \in S} \sum_{h=1}^{H} w'_h \mathbf{C}_{*h}(s') + \frac{N-n'}{Nn'} \mathbf{C}_* \right) \boldsymbol{\alpha}^T \quad (6.86)$$

W szczególności żądając podziału zbioru s' na równoliczne warstwy mamy:

$$\psi(s') = \boldsymbol{\alpha} \left(\frac{n'-n}{n'nH} \sum_{s' \in S} \sum_{h=1}^{H} \mathbf{C}_{*h}(s') + \frac{N-n'}{Nn'} \mathbf{C}_* \right) \boldsymbol{\alpha}^T \quad (6.87)$$

Stąd wynika, że

$$\mathrm{ED}_p^2\left(\tilde{\bar{Y}}_{wS}\right) \le \frac{1}{\binom{N}{n'}} \sum_{s' \in S} \rho(s') - \frac{N-n}{Nn} \sigma^2 \quad (6.88)$$

gdzie $\rho(s')$ jest promieniem spektralnym macierzy formy kwadratowej $\psi(s')$ danej wzorem (6.87). Zatem funkcja $\rho(s')$ może być stosowana jako kryterium grupowania na równoliczne warstwy pierwotnej próby s' według cech dodatkowych.

Dodajmy, że algorytm grupowania może być adaptacją procedury tworzenia warstw opisanej w uprzednim paragrafie przy założeniu, że tworzone warstwy mają być równoliczne. Zatem w każdym kroku opisanego tam algorytmu przemieszczając dany element populacji np. z h-tej warstwy do l-tej warstwy trzeba jednocześnie także jeden element przenieść z l-tej wartwy do h-tej warstwy.

Na zakończenie porównajmy precyzję predyktora $\tilde{\bar{Y}}_{wS}$ ze średnią $\bar{Y}_S$ z próby prostej o liczebności n. Przyjmując we wzorze (6.83), że warstwy są równoliczne, mamy:

$$D_p^2\left(\tilde{\bar{Y}}_{wS}\right) = \frac{1}{\binom{N}{n'}} \frac{1}{H} \frac{n'-n}{n'n} \sum_{s' \in S} \sum_{h=1}^{H} v_{*S'_h}\left(Y|s'\right) + \frac{N-n'}{Nn'} v_*(Y) \quad (6.89)$$

Wtedy można sprawdzić, że

$$\left(n'-1\right) v_{*s'}(Y) = \left(\frac{n'}{H} - 1\right) \sum_{h=1}^{H} v_{*S'_h}(Y|S'=s') + \frac{n'}{H} \sum_{h=1}^{H} \left(\bar{Y}_{s'_h} - \bar{Y}_{s'}\right)^2 \quad (6.90)$$

Stąd mamy:

$$\frac{1}{H}\sum_{h=1}^{H} v_{*S'_h}(Y|S'=s') = \frac{n'-1}{n'-H} v_{*s'}(Y) - \frac{n'}{H(n'-H)} \sum_{h=1}^{H} \left(\overline{Y}_{s'_h} - \overline{Y}_{s'}\right)^2$$

Podstawiając ten wynik odpowiednio do wzoru (6.89) mamy:

$$D_p^2\left(\tilde{\overline{Y}}_{wS}\right) = \frac{1}{\binom{N}{n'}} \frac{n'-n}{nH(n'-H)} \sum_{s'\in S} \sum_{h=1}^{H} (\overline{Y}_{S'_h} - \overline{Y}_{S'})^2 - \left[\frac{H-1}{n'-H}\frac{n'-n}{n'n} + \frac{N-n}{Nn}\right] v_*(Y) \qquad (6.91)$$

Stąd wynika, że

$$D^2\left(\overline{Y}_S\right) - D_p^2\left(\tilde{\overline{Y}}_{wS}\right) = \left[\frac{H-1}{n'-H}\frac{n'-n}{n'n}\right] v_*(Y) - \frac{1}{\binom{N}{n'}} \frac{n'-n}{nH(n'-H)} \sum_{s'\in S} \sum_{h=1}^{H} (\overline{Y}_{s'_h} - \overline{Y}_{s'})^2 \qquad (6.92)$$

Zatem predyktor $\tilde{\overline{Y}}_{wS}$ może być efektywniejszy od średniej $\overline{Y}_S$, gdy średnie wewnątrzwarstwowe są silnie zróżnicowane lub ilość warstw jest mała.

W szczególności niech liczebności n' oraz $\frac{n'}{H}$ są duże.Wówczas na podstawie wzoru (6. 90) mamy:

$$n' v_{*s'}(Y) \approx \frac{n'}{H} \sum_{h=1}^{H} v_{*S'_h}(Y|S'=s') + \frac{n'}{H} \sum_{h=1}^{H} \left(\overline{Y}_{s'_h} - \overline{Y}_{s'}\right)^2$$

Stąd mamy:

$$\frac{1}{H}\sum_{h=1}^{H} v_{*S'_h}(Y|S'=s') \approx v_{*s'}(Y) - \frac{1}{H} \sum_{h=1}^{H} \left(\overline{Y}_{s'_h} - \overline{Y}_{s'}\right)^2$$

Wtedy podstawiając otrzymany wynik do wzoru (6.89) mamy:

$$D_p^2\left(\tilde{\overline{Y}}_{wS}\right) \approx \frac{N-n}{Nn} v_*(Y) - \frac{1}{\binom{N}{n'}} \frac{n'-n}{n'nH} \sum_{s'\in S} \sum_{h=1}^{H} (\overline{Y}_{s'_h} - \overline{Y}_{s'})^2 \qquad (6.93)$$

Stąd wynika, że

$$D^2\left(\overline{Y}_S\right) - D_p^2\left(\tilde{\overline{Y}}_{wS}\right) \approx \frac{1}{\binom{N}{n'}} \frac{n'-n}{nn'H} \sum_{s'\in S} \sum_{h=1}^{H} (\overline{Y}_{s'_h} - \overline{Y}_{s'})^2 \qquad (6.94)$$

Zatem gdy rozmiar n' próby S' jest dostatecznie duży i liczba warstw H mała, to należy spodziewać się, że predyktor $\tilde{\tilde{Y}}_{wS}$ jest precyzyjniejszy od $\overline{Y}_S$. Przeprowadzone porównanie precyzji predyktorów ma sens, gdy koszty losowania próby S' i obserwacji w niej cech są znikome.

Pozostańmy przy założeniach, że liczba warstw H jest mała oraz rozmiar próby S' jest duży. Oznaczmy wariancję daną wzorem (6.93) jako funkcję liczebności n i n'. Niech koszt jednostkowy losowania próby S' i obserwacji w niej cech pomocniczych wynosi k_x, a koszt jednostkowy obserwacji cechy badanej y wynosi k_y. Niech dopuszczalna suma nakładów na losowanie prób i obserwacji w nich cech wynosi K. Wtedy zadanie optymalizacji liczebności prób n i n' formułujemy nastepująco:

$$\begin{cases} f(n,n') = \min imum \\ 0 < n < n' \le N \\ k_x n' + k_y n \le K \end{cases} \tag{6.95}$$

Używając znanej metody czynnika nieoznaczonego Lagrange'a i przy dużym rozmiarze N populacji optymalne rozwiązanie postawionego problemu ma postać:

$$n'_* = c\sqrt{\frac{\eta}{k_x}}; \qquad n_* = c\sqrt{\frac{k_y}{v_*(Y)-\eta}} \tag{6.96}$$

gdzie:

$$c = \frac{K}{\sqrt{k_x \eta} + \sqrt{k_y (v_*(Y) - \eta)}}$$

$$\eta = \frac{1}{\binom{N}{n'}} \frac{1}{H} \sum_{s' \in \mathcal{S}} \sum_{h=1}^{H} (\overline{Y}_{s'_h} - \overline{Y}_{s'})^2$$

Wtedy minimalny poziom wariancji wynosi:

$$D^2_{pK}\left(\tilde{\tilde{Y}}_{wS}\right) = f(n_*, n'_*) = \frac{1}{K}\left(\sqrt{k_x \eta} + \sqrt{k_y (v_*(Y) - \eta)}\right)^2 - \frac{1}{N} v_*(Y) \tag{6.97}$$

Przy predykcji na podstawie średniej z próby prostej optymalna liczebność wynosi $m = \frac{K}{k_y}$, a wartość jej wariancji: $D^2_K\left(\overline{Y}_S\right) = \frac{k_y}{K} v_*(Y) - \frac{1}{N} v_*(Y)$. Wtedy predyktor $\tilde{\tilde{Y}}_{wS}$ jest precyzyjniejszy od średniej $\overline{Y}_S$, gdy

$$D_K^2\left(\overline{Y}_S\right) - D_{pK}^2\left(\widetilde{\overline{Y}}_{wS}\right) = \frac{1}{K}\left[k_y v_*(Y) - \left(\sqrt{k_x \eta} + \sqrt{k_y\left(v_*(Y) - \eta\right)}\right)^2 \right] > 0 \qquad (6.98)$$

Nierówność ta zachodzi, gdy

$$\frac{k_x}{k_y} < \frac{(v_*(Y) - \sqrt{v_*(Y) - \eta})^2}{\eta} \qquad (6.99)$$

Stąd wynika, że predyktor $\widetilde{\overline{Y}}_{wS}$ opłaca się stosować zamiast średniej $\overline{Y}_S$, gdy koszty obserwacji cech dodatkowych są bardzo małe w stosunku do kosztu obserwacji cechy badanej.

6.8.4. Warstwowanie populacji przed i po wylosowaniu próby

Załóżmy, że przed losowaniem próby jest znany wektor cech pomocniczych $x = [x_1 \ldots x_z]$ w populacji, natomiast w wylosowanej próbie jest obserwowana cecha *p*. Zakłada się, że związek liniowy (np. w sensie współczynnika korelacji wielorakiej) między zmiennymi *p* i cechami *x* jest wysoki. Szacowana jest wartość średnia zmiennej *y*. Przyjmijmy, że zamierza się warstwować populację według wartości cechy *p*, co nie jest jednak możliwe przed losowaniem próby. Wtedy Armstrong i Wu (1992) proponują utworzenie G-warstw na podstawie cech *x*. Następnie z każdej tak utworzonej warstwy jest bezzwrotnie losowana próba prosta S_{1g} o liczebności n_g. W tak dobranych próbach jest obserwowana zmienna pomocnicza *p*. Na jej podstawie wyróżnionych jest H-warstw w każdej już dobranej próbce S_{1g}. Oznaczmy je jako zbiory S'_{1gh} o liczebności, n'_{gh}, g = 1,...,G i h = 1,... H_g, w których bezzwrotnie losowane są próbki proste S'_{2gh} o liczebności n_{gh}. Niech $n_g = f_g N_g$, gdzie N_g=NW_g jest ilością elementów w g-tej warstwie otrzymanej przez warstwowanie populacji na podstawie cech pomocniczych *x*. Z kolei niech $n_{gh} = f_{gh} n'_{gh}$.

Armstrong i Wu (1992) proponują następny nieobciążony estymator średniej:

$$\widetilde{\overline{Y}}_{ws} = \sum_{g=1}^{G} \sum_{h=1}^{H} W_g \frac{n'_{gh}}{n_g} \overline{y}_{s_{2gh}} \qquad (6.100)$$

gdzie:

$$\overline{y}_{s_{2gh}} = \frac{1}{n_{gh}} \sum_{k \in S_{2gh}} y_k \qquad (6.101)$$

Wariancja tej statystyki ma postać:

$$D^2\left(\widetilde{Y}_{ws}\right)=\frac{1}{N^2}\sum_{h=1}^{H}\sum_{g=1}^{G}\left[\left(\frac{1}{f_g f_{gh}}-1\right)A_{gh}+\left(\frac{1}{f_g}-1\right)B_{gh}\right] \qquad (6.102)$$

gdzie:

$$A_{gh}=N_{gh}v_{*gh}, \qquad B_{gh}=\left(\frac{N_g-N_{gh}}{N_g-1}\right)\left(N_{gh}\bar{y}_{gh}^2-v_{*gh}\right)$$

przy czym N_{gh} jest liczbą elementów w warstwie otrzymanej w wyniku podziału populacji za pomocą cech *x* i *p*.

Autorzy tej metody estymacji analizują również szeroko zagadnienia optymalizacji liczebności wyżej opisanych prób dwustopniowych losowanych z warstw.

Proponowaną metodę można zmodyfikować na etapie tworzenia warstw w populacji na podstawie cech *x*. Można bowiem do tego celu wykorzystać jedną z metod grupowania danych, co doprowadzi do wyróżnienia warstw jednorodnych z punktu widzenia zmienności warstw cech *x*. W tak utworzonych warstwach również zróżnicowanie wartości zmiennej y też powinno być małe ze względu na oczekiwane silne związki między cechami *y*, *p i* x. Do grupowania można użyć wspomnianych już metod: k-średnich Warda oraz ich modyfikacji lub uogólnień, z których jedną rozważano w uprzednim punkcie.

Z praktycznego punktu widzenia byłoby wskazane, aby tworzone warstwy były spójne w tym sensie, że każdy element należący do warstwy sąsiaduje przynajmniej z jednym elementem tej warstwy. To winno przyczynić się w niektórych wypadkach do oszczędności nakładów ponoszonych na badania statystyczne. Przykładowo niech elementami populacji będą gminy. Wtedy sąsiadujące ze sobą gminy, tworzące spójną warstwę, winny przyczynić się do obniżki kosztów przejazdów między gminami.

Ponieważ wcześniej opisywany algorytm grupowania nie zapewnia spójności warstw, Wywiał (1994) wprowadził prostą modyfikację metod prowadzącą do podziału populacji na spójne grupy. Wykorzystuje ona następujące pojęcia: Grupa elementów składa się z sąsiadujących ze sobą elementów, gdy każdy wchodzący w jej skład element sasiaduje przynajmniej z jednym elementem tej grupy. W szczególności jednoelementową grupę również zaliczamy do tej klasy grup, a zatem każdy element sam sąsiaduje ze sobą. Po to, by otrzymać podział populacji na spójne warstwy, wystarczy więc w każdej metodzie grupowania założyć, że na każdym etapie tworzenia do istniejącej warstwy będą dołączane grupy elementów, z których przynajmniej jeden sąsiaduje z daną warstwą.

7. WEKTOR ŚREDNICH Z PRÓBY GRUPOWEJ

W niniejszym rozdziale rozważamy problem estymacji wektora wartości średnich na podstawie próby grupowej. Naszą uwagę skoncentrujemy na porównaniu dokładności wektora średnich z próby grupowej z wektorem średnich z próby prostej. Ponadto rozważania będą toczyły się wokół zagadnień optymalizacyjnych. Pierwsze dotyczy wyznaczania liczebności grupy, a drugie podziału populacji na grupy na podstawie cechy pomocniczej. Problemy te są dość złożone i dlatego, aby je przedstawić w możliwie jasny sposób, ograniczono się do przypadku estymacji średnich na podstawie grup równolicznych. Wnioskowanie na podstawie prób składających się z nierównolicznych grup zwykle jest prowadzone przy planach losowania próby z prawdopodobieństwami doboru grup proporcjonalnymi do ich rozmiarów.

Przypomnijmy jeszcze, że losowanie grupowe zwykle jest stosowane wtedy, gdy nie jest możliwe posiadanie dokładnego operatu losowania, a jedynie spisu grup (zespołów) elementów tworzących populację. Zespoły te zazwyczaj są zdeterminowane w sposób naturalny, jak np. zbiór gmin traktowanych jako zespoły gospodarstw domowych.

7.1. Podstawowe własności wektora średnich z próby grupowej

Przyjmujemy, że populacja Ω jest podzielona na G rozłącznych i niepustych zbiorów Ω_p (p=1,...,G) tak, że $\bigcup_{p=1}^{G}\Omega_p = \Omega$. Zbiory te są nazywane grupami lub zespołami elementów populacji. Zakładamy, że grupy są równoliczne i każda z nich składa się z M elementów, czyli rozmiar populacji N = GM.

Przyjmujemy, że g-elementowa próba S jest dobierana za pomocą znanego schematu losowania prostej próby grupowej realizującej plan losowania:

$$P_g(s) = \frac{1}{\binom{G}{g}}$$

Tak wybierana próba różni się od próby prostej tylko tym, że zamiast poszczególnych elementów populacji są losowane do niej zespoły tych elementów. Ponadto wprowadzamy oznaczenia: Przez y_{ik} oznaczamy k-tą obserwację i-tej zmiennej. Sumę wartości i-tej zmiennej w p-tej grupie określa wzór:

$$z_{ip} = \sum_{k \in \Omega_p} y_{ik}$$

a wartość średnią w p-tej grupie:

$$\bar{y}_{ip} = \frac{1}{M} z_{ip}$$

Średnia suma wartości i-tej zmiennej przypadająca na grupę:

$$\bar{z}_i = \frac{1}{G} \sum_{p=1}^{G} z_{ip}$$

Średnia wartość i-tej zmiennej w populacji:

$$\bar{y}_i = \frac{1}{N} \sum_{p=1}^{G} z_{ip}$$

Macierz wariancji i kowariancji cech: $C_* = [c_*(\mathbf{y}_i, \mathbf{y}_j)]$, gdzie:

$$c_*\left(y_i, y_j\right) = \frac{1}{N-1} \sum_{k=1}^{G} \sum_{k \in \Omega_p} \left(y_{ik} - \bar{y}_i\right)\left(y_{jk} - \bar{y}_j\right)$$

Macierz wariancji i kowariancji sum wartości cech w grupach: $\mathbf{C}_{z^*}=[c_*(\boldsymbol{z}_i, \boldsymbol{z}_j)]$, gdzie:

$$c_*\left(z_i, z_j\right) = \frac{1}{G-1} \sum_{p=1}^{G} \left(z_{ip} - \bar{z}_i\right)\left(z_{jp} - \bar{z}_j\right)$$

Wektor $\bar{\mathbf{y}} = \left[\bar{y}_1 K\ \bar{y}_m\right]$ jest oceniany za pomocą estymatorów $\bar{\mathbf{y}}_{gS} = \left[\bar{y}_{1gS} K\ \bar{y}_{mgS}\right]$, gdzie:

$$\bar{y}_{igS} = \frac{1}{gM} \sum_{p \in S} \sum_{k \in \Omega_p} y_{ik} = \frac{1}{gM} \sum_{p \in S} z_{ip} \qquad (7.1)$$

Wektor $\bar{\mathbf{y}}_{gS}$ daje nieobciążone oceny średnich $\bar{\mathbf{y}}$.

Podobnie jak wariancje składowych wektora $\bar{\mathbf{y}}_{gS}$ można wyprowadzić ich kowariancje[48] :

$$\mathrm{Cov}\left(\bar{y}_{igS}, \bar{y}_{jgS}\right) = \frac{G-g}{GgM^2} c_*\left(z_i, z_j\right) \tag{7.2}$$

Nieobciążony estymator tej kowariancji otrzymujemy zastępując parametr $c_*(\mathbf{z}_i, \mathbf{z}_j)$ przez jego nieobciążony estymator:

$$c_{*S}\left(z_i, z_j\right) = \frac{1}{g-1} \sum_{p \in S} \left(z_{ip} - \bar{z}_i\right)\left(z_{jp} - \bar{z}_j\right) \tag{7.3}$$

Określoną wcześniej kowariancję $c_*(\mathbf{z}_i, \mathbf{z}_j)$ można zdekomponować do postaci[49] :

$$c_{*S}\left(z_i, z_j\right) = \frac{N-1}{G-1} \sqrt{v_*\left(y_i\right) v_*\left(y_j\right)} \left[r_{ij} + (M-1) r_{ij}^{(w)}\right] \tag{7.4}$$

gdzie:

$$r_{ij} = \frac{c_*\left(y_i, y_j\right)}{\sqrt{v_*\left(y_i\right) v_*\left(y_j\right)}}$$

$$r_{ij}^{(w)} = \frac{c^{(w)}\left(y_i, y_j\right)}{\sqrt{v_*\left(y_i\right) v_*\left(y_j\right)}} \tag{7.5}$$

$$c^{(w)}\left(y_i, y_j\right) = \frac{2}{(N-1)(M-1)} \sum_{p=1}^{G} \sum_{k \neq l \in \Omega_p} \left(y_{ik} - \bar{y}_i\right)\left(y_{jl} - \bar{y}_j\right) \tag{7.6}$$

Parametr $r_{ii}^{(w)}$ nazywamy współczynnikiem korelacji wewnątrzgrupowej i-tej zmiennej. Przez podobieństwo parametr $r_{ij}^{(w)}$ nazwiemy współczynnikiem korelacji wewnątrzgrupowej pary zmiennych o numerach (i,j).

Współczynnik $r_{ii}^{(w)}$ przyjmuje wartości z przedziału[50] $\left\langle -\frac{1}{M-1}; 1 \right\rangle$.

Na podstawie wzorów (7.4) i (7.2) otrzymujemy:

$$\mathrm{Cov}\left(\bar{y}_{igS}, \bar{y}_{jgS}\right) = \frac{(N-1)(G-g)}{(G-1)M^2 Gg} \sqrt{v_*(y_i) v_*(y_j)} \left[r_{ij} + (M-1) r_{ij}^{(w)}\right] \tag{7.7}$$

---

[48] Por. np. Cochran (1963), Pawłowski (1972), Zasępa (1972) i Konijn (1973).

[49] Tamże.

[50] Tamże.

Niech $\mathbf{R}^{(w)} = \left[r_{ij}^{(w)} \right]$ będzie macierzą współczynników korelacji wewnątrzgrupowej, natomiast $\mathbf{R}^{(w)} = \left[r_{ij} \right]$ macierzą zwykłych współczynników korelacji liniowej i w końcu przez $\mathbf{D}_*$ oznaczamy macierz diagonalną, której elementami są wariancje kolejnych zmiennych, czyli $\mathbf{D}_*$=diag$\mathbf{C}_*$. Wtedy macierz wariancji i kowariancji wektora $\bar{\mathbf{y}}_{gS}$ zapisujemy na podstawie wzoru (7.7) w postaci syntetycznej:

$$\mathbf{V}\left(\bar{y}_{gS}, P_g\right) = \frac{(N-1)(G-g)}{(G-1)M^2 Gg} \mathbf{D}_*^{1/2} \left(\mathbf{R} + (M-1)\mathbf{R}^{(w)} \right) \mathbf{D}_*^{1/2} \tag{7.8}$$

Jeśli N i G są duże, to:

$$\mathbf{V}\left(\bar{y}_{gS}, P_g\right) \approx \frac{1}{Mg} \mathbf{D}^{1/2} \left(\mathbf{R} + (M-1)\mathbf{R}^{(w)} \right) \mathbf{D}^{1/2} \tag{7.9}$$

przy czym $\mathbf{D} = \frac{N-1}{N} \mathbf{D}_*$.

Niech standardową postać obserwacji y_{ik} określa wzór:

$$h_{ik} = \frac{y_{ik} - \bar{y}_i}{\sqrt{v_*(y)}} \tag{7.10}$$

Wtedy określony wzorami (7.5) i (7.6) współczynnik korelacji wewnątrzgrupowej można zapisać następująco:

$$r_{ij}^{(w)} = \frac{2}{(N-1)(M-1)} \sum_{p=1}^{G} \sum_{k \neq l \in \Omega_p} \sum h_{ik} h_{jl} \tag{7.11}$$

Oznaczmy przez **y** macierz o wymiarach N×m, której każdy wiersz zawiera obserwacje wszystkich m zmiennych. M pierwszych wierszy macierzy **y** zawiera wartości zmiennych tworzących pierwszą grupę populacji, M następnych wierszy to obserwacje zmiennych w drugiej grupie itd. Symbolem $\mathbf{J}_b$ oznaczamy wektor o wymiarach b×1 składający się z samych jedynek. Symbolem $\otimes$ oznaczmy iloczyn Kroneckera dwóch macierzy. Wtedy na podstawie wzorów (7.10) i (7.11) można wykazać, że:

$$\mathbf{R}^{(w)} = \frac{1}{(N-1)(M-1)} \mathbf{h}^T \left(\mathbf{B}\mathbf{B}^T - \mathbf{I}_N \right) \mathbf{h} \tag{7.12}$$

gdzie:

$$\mathbf{h} = \mathbf{A}\mathbf{y}\mathbf{D}^{-1/2}$$

$$\mathbf{A} = \mathbf{I}_N - \frac{1}{N}\mathbf{J}_N\mathbf{J}_N^T \tag{7.13}$$

$$\mathbf{B} = \mathbf{I}_N \otimes \mathbf{J}_M^T$$

Kolumny macierzy **h** są więc otrzymywane poprzez standaryzację obserwacji zmiennych tworzących odpowiednie kolumny macierzy **y**, por. wzór (7.10).

Niech **P** będzie taką macierzą ortogonalną stopnia m, że $\mathbf{P}^T\mathbf{R}^{(w)}\mathbf{P} = \mathbf{D}_R$, gdzie $\mathbf{D}_R$ jest macierzą diagonalną pierwiastków charakterystycznych macierzy $\mathbf{R}^{(w)}$, które oznaczamy przez d_{Ri} (i=1,...,m). Stąd i ze wzoru (7.12) wynika, że:

$$\mathbf{D}_R = \frac{1}{(N-1)(M-1)}\mathbf{u}^T\left(\mathbf{B}^T\mathbf{B} - \mathbf{I}_N\right)\mathbf{u} \tag{7.14}$$

gdzie **u** = **hP**. Wtedy ze wzoru (7.13) otrzymujemy:

$$\mathbf{u} = \mathbf{A}\mathbf{y}\mathbf{D}^{-1/2}\mathbf{P} \tag{7.15}$$

Element diagonalny d_{Ri} macierzy $\mathbf{D}_R$ jest więc współczynnikiem korelacji wewnątrzgrupowej i-tej zmiennej, której obserwacje tworzą i-tą kolumnę macierzy **u**. Tę zaś otrzymano w wyniku danej wzorem (7.15) transformacji macierzy **y** zawierającej obserwacje wyjściowych zmiennych. Ze znanych własności współczynnika korelacji wewnątrzgrupowej [por. np. Konijn (1973); Zasępa (1972)] wynika, że

$$-\frac{1}{M-1} \le d_{Ri} \le 1, \qquad i = 1,K,m \tag{7.16}$$

gdzie:

$$d_{Ri} = 1 - \frac{v_w(u_i)}{v(u_i)} \tag{7.17}$$

$$v(u_i) = \frac{1}{N}\sum_{p=1}^{G}\sum_{k\in\Omega_p}\left(u_{ik} - \bar{u}_i\right)^2, \qquad \bar{u}_i = \frac{1}{N}\sum_{p=1}^{G}\sum_{k\in\Omega_p} u_{ik} \tag{7.18}$$

$$v_w(u_i) = \frac{1}{G(M-1)}\sum_{p=1}^{G}\sum_{k\in\Omega_p}\left(u_{ik} - \bar{u}_{ip}\right)^2, \qquad \bar{u}_{ip} = \frac{1}{M}\sum_{k\in\Omega_p} u_{ik} \tag{7.19}$$

Parametr $v_w(u_i)$ jest wariancją wewnątrzgrupową, a zatem d_{Ri} mierzy względne odchylenie wariancji wewnątrzgrupowej od zwykłej wariancji.

Podobnie jak to ma miejsce w przypadku jednowymiarowym [por. np. Konijm (1973), s. 225-227; Zasępa (1972), s. 311-312] macierz współczynników korelacji można zapisać następująco:

$$\mathbf{R}^{(w)} = \mathbf{D}^{-½}\left(\mathbf{C}_m - \frac{1}{M}\mathbf{C}_w\right)\mathbf{D}^{-½} \tag{7.20}$$

lub:
$$\mathbf{R}^{(w)} = \mathbf{D}^{-½}\left(\mathbf{C} - \mathbf{C}_w\right)\mathbf{D}^{-½} \tag{7.21}$$

lub:
$$\mathbf{R}^{(w)} = \frac{1}{M-1}\mathbf{D}^{-½}\left(M\mathbf{C}_m - \mathbf{C}\right)\mathbf{D}^{-½} \tag{7.22}$$

przy czym $\mathbf{C} = \frac{N-1}{N}\mathbf{C}_*$, natomiast przez $\mathbf{C}_m = \left[c_m\left(y_i, y_j\right)\right]$ oznaczono macierz wariancji i kowariancji międzygrupowych gdzie:

$$c_m\left(y_i, y_j\right) = \frac{1}{G}\sum_{p=1}^{G}\left(\bar{y}_{ip} - \bar{y}_i\right)\left(\bar{y}_{jp} - \bar{y}_j\right) \tag{7.23}$$

Przez $\mathbf{C}_w = \left[c_w\left(y_i, y_j\right)\right]$ oznaczono macierz wariancji i kowariancji wewnątrzgrupowej, przy czym:

$$c_w\left(y_i, y_j\right) = \frac{1}{G(M-1)}\sum_{p=1}^{G}\sum_{k\in\Omega_p}\left(y_{ik} - \bar{y}_{ip}\right)\left(y_{jk} - \bar{y}_{jp}\right) \tag{7.24}$$

Ze znanych twierdzeń o ilorazie wyznaczników macierzy [patrz np. C.R. Rao (1982), s. 89] i na podstawie wzorów (7.21), (7.22) i (7.14) można wykazać co następuje:

Twierdzenie 7.1: Załóżmy, że macierze $\mathbf{C}$ i $\mathbf{C}_w$ są dodatnio określone. Wtedy a) jeśli macierz $\mathbf{R}^{(w)}$ jest nieujemnie określona, to $\det\mathbf{C} \geq \det\mathbf{C}_w$ i $d_{Ri} \geq 0$, dla i=1,...,m; b) jeżeli $\mathbf{R}^{(w)}$ jest niedodatnio określona, to $\det\mathbf{C} \leq \det\mathbf{C}_w$ oraz $d_{Ri} \leq 0$ dla i=1,...,m. Zapisane nierówności stają się ostrymi, gdy macierz $\mathbf{R}^{(w)}$ jest w przypadku a) dodatnio określona, natomiast w przypadku b) ujemnie określona.

Zróżnicowanie wewnątrzgrupowe obserwacji zmiennej wielowymiarowej w stosunku do ich rozrzutu w populacji jest małe, gdy macierz $\mathbf{R}^{(w)}$ jest dodatnio określona. Gdy $\mathbf{R}^{(w)}$ jest ujemnie określona, to zróżnicowanie wewnątrzgrupowe wartości zmiennych jest większe od ich zróżnicowania w populacji.

Dodajmy, że Wywiał (1992) wprowadza wskaźnik siły skorelowania wewnątrzgrupowego.

Własności wektora średnich z próby prostej losowanej bezzwrotnie studiowano w rozdziale 3. Przypomnijmy, że macierz wariancji i kowariancji tej strategii estymacji ma postać:

$$\mathbf{V}\left(\bar{\mathbf{y}}_S, P_3\right) = \frac{N-n}{Nn}\mathbf{C}_* = \frac{N-n}{Nn}\mathbf{D}_*^{½}\mathbf{R}\mathbf{D}_*^{½}$$

Przyjmując, że $G \to \infty$ i $n = gM$ na podstawie (7.9) mamy:

$$\mathbf{V}\left(\bar{\mathbf{y}}_{gS}, P_g\right) - \mathbf{V}\left(\bar{\mathbf{y}}_{S}, P_3\right) = \frac{M-1}{gM}\,\mathbf{D}_*^{1/2}\mathbf{R}^{(w)}\mathbf{D}_*^{1/2} \tag{7.25}$$

Twierdzenie 7.2 [Wywiał (1992)]: Jeśli grupy, na jakie jest podzielona populacja, są równoliczne i ich ilość $G \to \infty$ oraz macierz $\mathbf{R}^{(w)}$ jest niedodatnio (nieujemnie) określona, to strategia $\left(\bar{\mathbf{y}}_{gS}, P_g\right)$ jest nie gorszą (nie lepszą) strategią od $\left(\bar{\mathbf{y}}_{S}, P_3\right)$.

Stąd, na podstawie wzoru (7.25) i twierdzenia 2.3 wnioskujemy, że

$$e_1 = \frac{\det \mathbf{V}\left(\bar{\mathbf{y}}_{gS}, P_g\right)}{\det \mathbf{V}\left(\bar{\mathbf{y}}_{S}, P_3\right)} = \frac{\det\left(\mathbf{R} + (M-1)\mathbf{R}^{(w)}\right)}{\det \mathbf{R}} \tag{7.26}$$

$$e_2 = \frac{q^2\left(\bar{\mathbf{y}}_{gS}, P_g\right)}{q^2\left(\bar{\mathbf{y}}_{S}, P_3\right)} = \sum_{i=1}^{m} r_{ii}^{(w)} a_i = \tilde{r} \tag{7.27}$$

gdzie:

$$a_i = \frac{v(y_i)}{\sum_{i=a}^{m} v(y_i)}$$

Stąd i z twierdzeń 7.1 i 7.2 wynika, że przy dostatecznie dużej ilości równolicznych grup, wektor średnich z próby grupowej może być efektywniejszy od wektora średnich z próby prostej, gdy macierz współczynników korelacji wewnątrzgrupowej jest ujemnie określona. Zatem o ile jest to możliwe, należy tak wyróżniać rozłączne grupy w populacji, by jak najsilniej były zróżnicowane obserwacje cech wewnątrz tych grup.

7.2. Optymalizacja rozmiarów próby i grupy

W niektórych zagadnieniach praktycznych, np. dotyczących badań reprezentacyjnych w rolnictwie i środowisku naturalnym, pojawia się możliwość tworzenia grup. Wtedy najprościej tworzyć grupy równoliczne. Wobec ograniczonych kosztów przeznaczonych na badania trzeba zdecydować, jak dużo grup mamy losować i o jakiej liczebności. Ponadto pojawia się potrzeba określenia związku między liczebnością grupy i wariancją wewnątrzgrupową bądź współczynnikiem korelacji wewnątrzgrupowej. Tym problemem w przypadku estymacji cechy jednowymiarowej, zajmowali się początkowo Jessen (1942), Mahalanobis (1944), Hedricks (1944)[51] . Na podstawie badań empirycznych wywnioskowali, że wariancja wewnątrzgrupowa v_w , dana wzorem (7.24) dla i=j, jest niemalejącą funkcją li-

[51] Informację tę podajemy za Cochranem (1963) i Zasępą (1972).

czebności grupy. Zależność tę zaproponowali modelować za pomocą funkcji potęgowej, jak następuje:

$$v_w = \alpha M^{\beta}, \quad \alpha > 0 \; i \; \beta > 0 \tag{7.28}$$

Parametry α i β proponuje się szacować metodą najmniejszych kwadratów na podstawie próby wstępnej. Przy dużych liczebnościach G i N ze wzorów (7.9) i (7.22) dla m=1 wynika:

$$D^2\left(\bar{y}_{gS}, P_g\right) = \frac{1}{Mg}\left[v + (M-1)\left(v - v_w\right)\right] = \frac{v}{g} - \frac{m-1}{mg} v_w = \frac{v}{g} - \alpha \frac{(M-1)M^{\beta}}{Mg}$$

$$D^2\left(\bar{y}_{gS}, P_g\right) = \frac{1}{g}\left(v - \alpha(M-1)M^{\beta-1}\right) \tag{7.29}$$

Funkcję kosztów zmiennych gromadzenia obserwacji cech określają Hansen i in. (1953) równaniem:

$$k(g, M) = k_1\sqrt{g} + k_2 g + k_3 gM \tag{7.30}$$

gdzie $k_1\sqrt{g}$ mierzy koszty przejazdu między grupami, k_2 jest kosztem jednostkowym sporządzania operatu losowania, który w badaniach praktycznych zazwyczaj bywa pomijany. Przez k_3 oznaczono koszt jednostkowy obserwacji wartości zmiennej.

Formułowane są dwa zadania optymalizacyjne. Pierwsze polega na takim ustalaniu liczebności g i M, aby dana wzorem (7.29) wariancja osiągnęła minimum przy ustalonych kosztach dopuszczalnych badania statystycznego. Drugie zadanie ma na celu znalezienie takich liczebności g i M, które zminimalizują funkcję kosztów daną wzorem (7.30) przy ustalonej dokładności estymacji określonej dopuszczalnym poziomem wariancji, danej wzorem (7.29).

Postawione zadania nie mają prostych analitycznych rozwiązań i trzeba je wyznaczać odpowiednimi metodami numerycznymi. Przy założeniu, że $k_2 = 0$, Cochran (1963) badał problem istnienia rozwiązań postawionych zadań.

7.3. Metody tworzenia grup równolicznych

7.3.1. Grupowanie populacji na podstawie jednej cechy pomocniczej

Przyjmijmy, że celem wnioskowania jest wartość średnia cechy jednowymiarowej w nadpopulacji. Znane są wartości cechy wspomagającej w całej populacji. Przypuśćmy, że mamy do czynienia z modelem regresyjnym w najprostszej postaci:

$$Y_k = \alpha x_k + \beta + U_k, \quad k \in \Omega_p, \; p = 1,..,G \tag{7.31}$$

przy czym wszystkie rozłączne grupy Ω_p (p=1,...,G) zawierają po M elementów populacji i $\bigcup_{p=1}^{G}\Omega_p = \Omega$. Parametry zmiennych Y_k określają wyrażenia:

$$\mathcal{E}(Y_k) = \alpha x_k + \beta = \mu_k , \qquad \mathcal{E}(U_k) = 0$$

$$\mathcal{D}^2(Y_k) = \mathcal{D}^2(U_k) = \sigma^2 , \qquad \mathcal{C}ov(Y_k , Y_1) = \mathcal{C}ov(U_k , U_1) = 0$$

przy czym $k \neq 1 \in \Omega$. Sumę zmiennych w p-tej grupie określamy wzorem:

$$Z_p = \sum_{k \in \Omega_p} Y_k \tag{7.32}$$

Wtedy:

$$\mathcal{E}(Z_p) = M\beta + \alpha \sum_{k \in \Omega_p} x_k , \qquad \mathcal{D}^2(Z_k) = M\sigma^2 , \quad \mathcal{C}ov(Z_p , Z_h) = 0$$

Rozważmy zadanie polegające na ocenie wartości średniej $\overline{Y} = \sum_{p=1}^{G} \sum_{k \in \Omega_p} Y_k$ na podstawie próby składającej się z g grup. Próba jest dobierana zgodnie ze schematem losowania prostej próby grupowej, tzn. grupy są losowane ze stałym prawdopodobieństwem ich wyboru do próby. Do tego celu użyjemy predyktora w postaci:

$$\overline{Y}_{gS} = \frac{1}{Mg} \sum_{p=1}^{g} Z_p \tag{7.33}$$

Predyktor ten jest P nieobciążony oraz P-ξ nieobciążony. Jego błąd średniokwadratowy wyznaczył Wywiał (1992):

$$\mathcal{E}E\left(\overline{Y}_{gS} - \overline{Y}\right)^2 = \frac{G-g}{gN}\left(\frac{1}{M}\alpha^2 v_*(b) + \sigma^2\right) \tag{7.34}$$

gdzie $v_*(b)$ określa wzór (7.4), w którym zmienne z_i i z_j należy zastąpić przez rozważaną tutaj zmienną *b*. Jej wartości są sumami cechy pomocniczej w poszczególnych grupach. Dodajmy, że p-ξ nieobciążony estymator błędu średniokwadratowego otrzymujemy ze wzoru (7.34) zastępując w nim $v_*(b)$ przez wariancję z próby:

$$V_{*S}(b) = \frac{1}{g-1} \sum_{k \in S} \left(b_k - \overline{b}_S\right)^2 , \qquad \overline{b}_S = \frac{1}{g} \sum_{k \in S} b_k \tag{7.35}$$

Korzystając z rezultatów otrzymanych w paragrafie 7.1 mamy:

$$v_*(b) = \frac{N-1}{G-1} v_*(x)\left[1 + (M-1) r_x^{(w)}\right]$$

gdzie:

$$r_x^{(w)} = \frac{2}{(N-1)(M-1)} \sum_{p=1}^{g} \sum_{k \neq l \in \Omega_p} (x_k - \bar{x})(x_l - \bar{x})$$

jest współczynnikiem korelacji wewnątrzgrupowej zmiennej wspomagającej. Podstawiając ten wynik do wzoru (7.34) mamy:

$$\mathcal{E}E\left(\bar{Y}_{gS} - \bar{Y}\right)^2 = \frac{G-g}{gN}\left(\frac{1}{M}\alpha^2 \frac{N-1}{G-1} v_*(x)\left[1 + (M-1)r_x^{(w)}\right] + \sigma^2\right) \tag{7.36}$$

Stąd wynika, że błąd strategii predykcji $(\bar{y}_{gS}, P_g)$ maleje wraz ze spadkiem wartości współczynnika korelacji wewnątrzgrupowej cechy pomocniczej. Stąd wnioskujemy, że populację należy tak dzielić na równoliczne grupy, aby wewnątrzgrupowe zróżnicowanie zmiennej wspomagającej było jak największe. Ze wzoru (7.34) wynika, że wniosek ten jest równoważny następującemu. Błąd średniokwadratowy predykcji maleje wraz ze spadkiem wariancji sum cechy wspomagającej w grupach, czyli parametru $v_*(b)$. Zatem należy populację tak podzielić na równoliczne i rozłączne grupy, aby ta wariancja była jak najmniejsza.

Wariancję v_* (b) zapisujemy w postaci równania:

$$v(b) = \frac{1}{G-1}\left\{\sum_{p=1}^{G}\left(\sum_{k \in \Omega_p} x_k\right)^2 - MN\bar{x}^2\right\}$$

Stąd wynika, że minimalizacja wariancji $v_*(b)$ jest równoważna minimalizacji funkcji:

$$F(\Omega_1, ..., \Omega_G) = \sum_{p=1}^{G} f(\Omega_p) \tag{7.37}$$

gdzie:

$$f(\Omega_p) = \left(\sum_{k \in \Omega_p} x_k\right)^2 \tag{7.38}$$

Niech $\Psi = \{\Omega_1, ..., \Omega_G\}$ będzie podziałem populacji Ω na rozłączne grupy, z których każda liczy M elementów populacji i $\bigcup_{p=1}^{G} \Omega_p = \Omega$. Niech $\mathcal{L}(\Omega)$ będzie zbiorem wszystkich możliwych podziałów ψ. Na podstawie ogólnego wzoru podanego przez Lipskiego i Marka (1986), s. 48 wnioskujemy, że zbiór $\mathcal{L}(\Omega)$ liczy k_* elementów, gdzie

$$k_* = \frac{N!}{G!(M!)^G} \tag{7.39}$$

Zadanie grupowania populacji według cechy pomocniczej można więc precyzyjniej sformułować następująco: Należy znaleźć taki układ grup $\underline{\Psi}$, aby:

$$F(\underline{\Psi}) = \min_{\psi \in \mathcal{P}(\Omega)} \text{imum}\{F(\Psi)\} \tag{7.40}$$

Bezpośrednie wyszukiwanie układu $\underline{\Psi}$ wydaje się w praktyce niewykonalne, bowiem już dla stosunkowo małych N liczność k_* jest bardzo wysoka[52]. Dlatego niżej konstruujemy pewien algorytm iteracyjnego poszukiwania układu grup $\underline{\Psi}$.

Niech $\Psi_t = \{\Omega_1^{(t)},...,\Omega_G^{(t)}\}$ będzie rozmieszczeniem elementów populacji w grupach otrzymanym w wyniku t-tej iteracji. Zakładamy, że elementy populacji są w każdej iteracji identyfikowane wzajemnie jednoznacznie przez ciąg liczb naturalnych od jeden do N, czyli $\Omega=\{1,2,...,N\}$. Pojawiającą się w trakcie t-tej iteracji grupę zawierającą k-ty element populacji oznaczamy przez $\Omega^{(t)}(k)$. Podstawowa zmiana podziału Ψ_t populacji polega na przemieszczeniu elementu k-tego populacji z grupy $\Omega^{(t)}(k)$ do grupy $\Omega^{(t)}(a)$ oraz jednocześnie elementu a-tego z grupy $\Omega^{(t)}(a)$ do grupy $\Omega^{(t)}(k)$, dlatego by obie nowo tworzone grupy pozostawały równoliczne. Tak otrzymane grupy oznaczamy przez $\Omega^{(t+1)}(a)=\{\Omega^{(t)}(k)-k\cup a\}$ i $\Omega^{(t+1)}(k)=\{\Omega^{(t)}(a)-a\cup k\}$. W ten sposób otrzymujemy jeden z możliwych układów $\Psi_{(t+1)}$ podczas (t+1) iteracji algorytmu.

Różnica między wartością funkcji kryterium dla optymalnego w t-tej iteracji układu grup a tą funkcją dla układu aktualnego ma postać:

$$d_{t+1}(k,a)= F(\underline{\Psi}_t) - F(\Psi_{t+1}) \tag{7.41}$$

gdzie na podstawie wzoru (7.37) mamy:

$$d_{t+1}(k,a)= f[\Omega^{(t+1)}(k)] + f[\Omega^{(t+1)}(a)] - f[\Omega^{(t)}(k)] - f[\Omega^{(t)}(a)] \tag{7.42}$$

Następnie wyznaczamy tak parę $(\underline{k},\underline{a})$ elementów populacji, aby:

$$d_{t+1}(\underline{k},\underline{a}) = \max_{k=1,...,N;\ a\notin\Omega^{(t)}(k)} \left\{ d_{k+1}(k,a) \right\} \tag{7.43}$$

Wtedy zastępując w podziale populacji Ψ_t grupy $\underline{\Omega}^{(t)}(k)$ i $\underline{\Omega}^{(t)}(a)$ przez odpowiednio $\underline{\Omega}^{(t+1)}(k)=\{\underline{\Omega}^{(t)}(a)-\underline{a}\cup\underline{k}\}$ i $\underline{\Omega}^{(t+1)}(a)=\{\underline{\Omega}^{(t)}(k)-\underline{k}\cup\underline{a}\}$ otrzymujemy optymalny układ grup na koniec (t+1) iteracji algorytmu, ponieważ określone przemieszczenie elementów populacji między grupami zapewniło maksymalny spadek funkcji kryterium $F(\Psi)$.

Określony iteracyjny proces kontynuujemy aż do momentu, gdy $d(\underline{k},\underline{a})=0$, ponieważ wtedy żadne przemieszczenie elementów między grupami nie spowoduje spadku wartości minimalizowanej wartości funkcji kryterium $F(\Psi)$. Jeśli w praktyce ze względu na dużą

[52] Np. jeśli N = 15, G = 5 i M = 3, to k = 1 401 400.

liczebność populacji proces poszukiwania układu optymalnego grup miałby się znacznie wydłużyć, to można ograniczyć ilość dopuszczalną iteracji algorytmu w sposób arbitralny.

Proponowany algorytm grupowania można tak zmodyfikować, by dawał grupy spójne elementów. Oznacza to, iż wymaga się, aby każdy element należący do danej grupy sąsiadował przynajmniej z jednym innym elementem tej grupy. Aby to osiągnąć, trzeba łączyć tylko te elementy w grupy, które sąsiadują ze sobą. Tak zmodyfikowany algorytm będziemy nazywać warunkową procedurą tworzenia grup spójnych.

Na zakończenie dodajmy, że Konijn (1973) za Daleniusem (1957) i Ghoshem (1963a) rozważa problem estymacji na podstawie próby podwójnej losowanej z populacji podzielonej na grupy. Pierwsza próba prosta złożona z elementów populacji jest losowana bezzwrotnie. W niej obserwuje się cechy pomocnicze identyfikujące przynależność każdego elementu do poszczególnych grup. Do drugiej próby są już losowane grupy elementów wyróżnione wśród elementów pierwszej próby. Dopiero w tej drugiej próbie jest obserwowana zmienna, której średnia jest celem estymacji.

7.3.2 Grupowanie populacji na podstawie wielowymiarowej cechy pomocniczej

Określony wzorem (7.31) model nadpopulacji uogólniamy do postaci:

$$Y_k=\boldsymbol{\alpha}\mathbf{x}_k+\beta+U_k, \quad k\in\Omega_p,\ p=1,..,G \tag{7.44}$$

gdzie $\mathbf{x}_k=[x_{k1} \ldots x_{km}]$ jest wektorem m nielosowych cech dodatkowych obserwowanych na k-tym elemencie populacji, $\boldsymbol{\alpha}=[\alpha_1 \ldots \alpha_m]$ jest wektorem parametrów regresji. Zakładamy, że

$$\mathcal{E}(Y_k) = \boldsymbol{\alpha}\mathbf{x}_k^T + \beta = \mu_k, \quad \mathcal{E}(U_k) = 0$$

$$\mathcal{D}^2(Y_k) = \mathcal{D}^2(U_k) = \sigma^2, \qquad \mathcal{C}ov(Y_k, Y_1) = \mathcal{C}ov(U_k, U_1) = 0$$

przy czym $k \neq 1 \in \Omega$.

Sumę zmiennych w p-tej grupie określamy wzorem:

$$Z_p = \sum_{k\in\Omega_p} Y_k \tag{7.45}$$

Wtedy:

$$\mathcal{E}(Z_p) = M\beta + \boldsymbol{\alpha}\sum_{k\in\Omega_p}\mathbf{x}_k^T; \qquad \mathcal{D}^2(Z_k)=M\sigma^2; \qquad \mathcal{C}ov(Z_p,Z_k)=0$$

Predykcja wartości średniej $\overline{Y} = \frac{1}{MG}\sum_{p=1}^{G}\sum_{k\in\Omega_p} Y_k$ jest prowadzona, na podstawie próby składającej się z g-grup, za pomocą statystyki $\overline{Y}_{gS}$ określonej wzorem (7.33), która jest P-nieobciążona i ξ-P nieobciążona.

Niech $\mathbf{b}_p = \sum_{k \in \Omega_p} \mathbf{x}_k$ oraz niech $\mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ L \\ \mathbf{b}_G \end{bmatrix}$ będzie macierzą sum wartości cech dodatkowych w poszczególnych grupach. Przyjmując, że **J** jest jedynkową kolumną o G elementach, wektor średnich sum przypadających na grupę wartości zmiennych dodatkowych wyrażamy wzorem: $\bar{\mathbf{b}} = \frac{1}{G}\sum_{p=1}^{G} \mathbf{b}_p = \frac{1}{G}\mathbf{J}^T\mathbf{B}$. Wtedy macierz wariancji i kowariancji wektora cech *b* ma postać:

$$\mathbf{C}_*(b) = \mathbf{B}^T\left(\mathbf{I}_G - \frac{1}{G}\mathbf{J}\mathbf{J}^T\right)\mathbf{B}$$

gdzie $\mathbf{C}_*(b)=[c_*(b_i,b_j)]$.

Wywiał (1993a) otrzymuje następującą postać błędu średnio-kwadratowego predykcji:

$$\mathcal{E}\mathrm{E}\left(\overline{Y}_{gS} - \overline{Y}\right)^2 = \frac{G-g}{gN}\left(M^{-1}\boldsymbol{\alpha}\mathbf{C}_*(b)\boldsymbol{\alpha}^T + \sigma^2\right) \qquad (7.46)$$

Niech $\mathbf{e} = \frac{1}{\sqrt{\boldsymbol{\alpha}\boldsymbol{\alpha}^T}}$ będzie unormowanym współczynnikiem regresji. Wtedy, na podstawie wzoru (7.46), otrzymujemy:

$$\mathcal{E}\mathrm{E}\left(\overline{Y}_{gS} - \overline{Y}\right)^2 = \frac{G-g}{gN}\left(M^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{e}\mathbf{C}_*(b)\mathbf{e}^T + \sigma^2\right) \qquad (7.47)$$

Niech $\mathbf{e}_1$ będzie takim wektorem, że

$$\lambda_1 = \mathbf{e}_1\mathbf{C}_*(b)\mathbf{e}_1^T = \max_{\mathbf{e}\mathbf{e}^T=1} \operatorname{imum}\left\{\mathbf{e}\mathbf{C}_*(b)\mathbf{e}^T\right\} \qquad (7.48)$$

Zatem $\mathbf{e}_1$ jest wektorem własnym macierzy $\mathbf{C}_*(b)$ odpowiadającym jej maksymalnej wartości własnej λ_1 . Stąd i ze wzorów (7.47) i (7.48) wynika nierówność:

$$\mathcal{E}\mathrm{E}\left(\overline{Y}_{gS} - \overline{Y}\right)^2 \leq = \frac{G-g}{gN}\left(M^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\ \lambda_1 + \sigma^2\right) \qquad (7.49)$$

Błąd średniokwadratowy predykcji przyjmuje więc wartość maksymalną, gdy unormowany wektor współczynników regresji $\mathbf{e} = \mathbf{e}_1$. Wtedy układ unormowanych współczynników regresji **e** jest najmniej korzystny z punktu widzenia wartości błędu średniokwadratowego predykcji. Populację należy więc tak podzielić na równoliczne i rozłączne grupy, aby promień spektralny λ_1 macierzy $\mathbf{C}_*(b)$ był jak najmniejszy.

Przez $\mathbf{C}_*(b|\Psi)$ oznaczamy macierz wariancji i kowariancji sum wartości cech dodatkowych obserwowanych w poszczególnych grupach tworzących rozbicie Ψ populacji na grupy. Podobnie symbolem $\lambda_1(\Psi)$ oznaczamy promień spektralny macierzy $\mathbf{C}_*(b|\Psi)$.

Zadanie grupowania populacji według cechy dodatkowej można więc precyzyjniej sformułować następująco. Znaleźć taki układ grup $\underline{\Psi} = \{\underline{\Omega}_1,...,\underline{\Omega}_G\}$, aby

$$\lambda_1(\underline{\Psi}) = \min_{\psi \in \mathcal{J}(\Omega)} \operatorname{imum}\{\lambda_1(\Psi)\} \tag{7.50}$$

W celu optymalnego wyróżnienia grup można użyć iteracyjnego algorytmu grupowania opisywanego w uprzednim punkcie, przy czym występującą tam funkcję kryterium F należy zastąpić przez określoną tutaj funkcję λ_1. Iteracyjny proces wyszukiwania optymalnego układu równolicznych grup kontynuujemy aż do momentu, gdy żadne przemieszczenie elementów między grupami nie spowoduje spadku minimalizowanej wartości funkcji kryterium $\lambda_1(\Psi)$. Jeśli ze względu na dużą liczebność populacji proces poszukiwania optymalnego układu grup miałby się w praktyce znacznie wydłużyć, to można w sposób arbitralny ograniczyć dopuszczalną liczbę iteracji.

7.3.3. Grupowanie próby po jej wylosowaniu

Korzystając z oznaczeń wprowadzonych w uprzednim punkcie załóżmy, że na elementach wstępnej próby prostej S' o liczebności n wylosowanej z populacji obserwujemy zespół cech pomocniczych w próbie. Ich związek ze zmienną badaną Y opisuje model regresyjny wprowadzony w uprzednim punkcie. Na podstawie tych cech grupujemy zbiór **S**' na G rozłącznych i wyczerpujacych zbiór S' grup U_h, h=1,...,H, z których każda liczy m-elementów. Problem wyboru kryterium grupowania omówimy później. Następnie spośród tak utworzonych grup losujemy bezzwrotnie i ze stałymi prawdopodobieństwami wyboru g-grup do próby S. Przyjmujemy model nadpopulacji regresyjnej opisany w uprzednim paragrafie.

Do predykcji wartości średniej w populacji $\overline{Y} = \sum_{i=1}^{N} Y_i$ użyjemy statystyki [por. wzór (7.1)]:

$$\widetilde{Y}_{gS} = \frac{1}{gm} \sum_{p \in S} \sum_{k \in U_p} Y_k = \frac{1}{gm} \sum_{p \in S} Z_p \tag{7.51}$$

Teraz wyliczamy:

$$E(\widetilde{Y}_{gS}) = E_{S'} E_{S/S'}(\widetilde{Y}_{gS}) = E_{S'}(\overline{Y}_{S'}) = \overline{Y} \tag{7.52}$$

$$\mathcal{E} E(\widetilde{Y}_{gS}) = \mathcal{E}(\overline{Y}) = \mu \tag{7.53}$$

Zatem statystyka $\widetilde{Y}_{gS}$ jest p-nieobciążonym oraz p-ξ nieobciążonym predyktorem średniej $\overline{Y}$. Jej błąd średniokwadratowy otrzymujemy następująco:

$$\mathcal{E}\mathrm{E}(\widetilde{\mathrm{Y}}_{\mathrm{gS}} - \overline{\mathrm{Y}})^2 = \mathcal{E}\mathrm{E}_{\mathrm{S}'}\mathrm{E}_{\mathrm{S/S}'}[(\widetilde{\mathrm{Y}}_{\mathrm{gS}} - \overline{\mathrm{Y}}_{\mathrm{S}'}) + (\overline{\mathrm{Y}}_{\mathrm{S}'} - \overline{\mathrm{Y}})]^2 = \mathcal{E}\mathrm{E}_{\mathrm{S}'}\,\mathrm{D}^2_{\mathrm{S/S}'}\left(\widetilde{\mathrm{Y}}_{\mathrm{gS}} \middle| \mathrm{S}' = \mathrm{s}'\right) + \mathcal{E}\,\mathrm{D}^2_{\mathrm{S}'}\left(\overline{\mathrm{Y}}_{\mathrm{S}'}\right)$$

przy czym na podstawie wzoru (7.2) mamy:

$$\mathrm{E}_{\mathrm{S/S}'}[(\widetilde{\mathrm{Y}}_{\mathrm{gS}} - \overline{\mathrm{Y}}_{\mathrm{S}'}) + (\overline{\mathrm{Y}}_{\mathrm{S}'} - \overline{\mathrm{Y}})]^2 = \mathrm{D}^2_{\mathrm{S/S}'}\left(\widetilde{\mathrm{Y}}_{\mathrm{gS}} \middle| \mathrm{S}' = \mathrm{s}'\right) + (\overline{\mathrm{Y}}_{\mathrm{S}'} - \overline{\mathrm{Y}})^2 \tag{7.54}$$

$$\mathrm{D}^2_{\mathrm{S/S}'}\left(\widetilde{\mathrm{Y}}_{\mathrm{gS}} \middle| \mathrm{S}' = \mathrm{s}'\right) = \frac{\mathrm{G} - \mathrm{g}}{\mathrm{G}\,\mathrm{gm}^2}\,\mathrm{v}_*\left(\mathrm{Z} \middle| \mathrm{s}'\right) \tag{7.55}$$

$$\mathrm{E}_{\mathrm{S}'}\mathrm{D}^2_{\mathrm{S/S}'}\left(\widetilde{\mathrm{Y}}_{\mathrm{gS}} \middle| \mathrm{S}' = \mathrm{s}'\right) = \binom{\mathrm{N}}{\mathrm{n}}^{-1} \frac{\mathrm{G} - \mathrm{g}}{\mathrm{G}\,\mathrm{gm}^2} \sum_{\mathrm{s}' \in \mathcal{S}'} \mathrm{v}_*\left(\mathrm{Z} \middle| \mathrm{s}'\right) \tag{7.56}$$

gdzie $\mathcal{S}'$ jest przestrzenią prób s'.

$$\mathrm{D}^2_{\mathrm{S}'}\left(\overline{\mathrm{Y}}_{\mathrm{S}'}\right) = \frac{\mathrm{N} - \mathrm{n}'}{\mathrm{N}\mathrm{n}'}\,\mathrm{v}_*(\mathrm{Y}) \tag{7.57}$$

Korzystając z wyników otrzymanych przez Wywiała (1993a) mamy:

$$\mathcal{E}\left(\mathrm{v}_*(Z|\mathrm{s}')\right) = \frac{1}{\mathrm{m}}\boldsymbol{\alpha}\mathbf{C}_*(b|s')\boldsymbol{\alpha}^{\mathrm{T}} + \sigma^2 \tag{7.58}$$

$$\mathcal{E}\left(\mathrm{v}_*(Y)\right) = \boldsymbol{\alpha}\mathbf{C}_*(X)\boldsymbol{\alpha}^{\mathrm{T}} + \sigma^2 \tag{7.59}$$

Przez $\mathbf{C}_*(b|\mathrm{s}')$ oznaczono macierz wariancji i kowariancji sum cech dodatkowych w próbie s', którą określa wzór:

$$\mathbf{C}_*(b|s') = \frac{1}{\mathrm{g} - 1}\mathbf{B}^{\mathbf{T}}_{\mathbf{s}'}\left(\mathbf{I}_{\mathbf{g}} - \frac{1}{\mathrm{g}}\mathbf{J}_{\mathbf{g}}\mathbf{J}^{\mathrm{T}}_{\mathrm{g}}\right)\mathbf{B}_{\mathbf{s}'}$$

gdzie:

$$\mathbf{B}_{\mathbf{s}'} = \begin{bmatrix} \mathbf{b}_{\mathbf{i}_1} \\ \dots \\ \mathbf{b}_{\mathbf{i}_n} \end{bmatrix}, \qquad \mathrm{i}_j \in \mathrm{s}'$$

a przez $\mathbf{I}_g$ oznaczono macierz jednostkową stopnia g, natomiast przez $\mathbf{J}_g$ kolumnę składającą się z g jedynek. Przypomnijmy, że $\mathbf{C}_*(X)$ jest macierzą wariancji i kowariancji nielosowych cech dodatkowych w populacji.

Na podstawie wzorów (7.54)-(7.59) otrzymujemy błąd średniokwadratowy predykcji:

$$\mathcal{E}\mathrm{E}(\widetilde{\mathrm{Y}}_{\mathrm{gS}} - \overline{\mathrm{Y}})^2 = \binom{\mathrm{N}}{\mathrm{n}}^{-1} \sum_{\mathrm{s}' \in \mathcal{S}'} \boldsymbol{\alpha}\mathbf{A}(\mathrm{s}')\boldsymbol{\alpha}^{\mathrm{T}} + \left(\frac{\mathrm{G}-\mathrm{g}}{\mathrm{Ggm}^2} + \frac{\mathrm{N}-\mathrm{n}'}{\mathrm{Nn}'}\right)\sigma^2 \tag{7.60}$$

gdzie:

$$\mathbf{A}(\mathrm{s}') = \frac{\mathrm{G}-\mathrm{g}}{\mathrm{Ggm}^3}\mathbf{C}_*(b|\mathrm{s}') + \frac{\mathrm{N}-\mathrm{n}}{\mathrm{Nn}}\mathbf{C}_*(X) \tag{7.61}$$

Niech $\kappa_1(\mathrm{s}')$ będzie promieniem spektralnym macierzy $\mathbf{A}(\mathrm{s}')$. Wtedy:

$$\mathcal{E}\mathrm{E}(\widetilde{\mathrm{Y}}_{\mathrm{gS}} - \overline{\mathrm{Y}})^2 \leq \binom{\mathrm{N}}{\mathrm{n}}^{-1} \sum_{\mathrm{s}' \in \mathcal{S}'} \kappa_1(\mathrm{s}') + \left(\frac{\mathrm{G}-\mathrm{g}}{\mathrm{Ggm}} + \frac{\mathrm{N}-\mathrm{n}}{\mathrm{Nn}}\right)\sigma^2 \tag{7.62}$$

Wynik ten sugeruje, że elementy tworzące próbę s' należy tak grupować według cech dodatkowych, aby $\kappa(\mathrm{s}')$ osiągało minimalną wartość. Przy tak określonym kryterium konstrukcja algorytmu tworzenia grup nie odbiega znacznie od procedury opisanej w uprzednim paragrafie.

8. WEKTOR ŚREDNICH Z PRÓBY DWUSTOPNIOWEJ

8.1. Podstawowe własności wektora estymatorów

Utrzymujemy w mocy oznaczenia wprowadzone w rozdziale 7 w związku z wnioskowaniem na podstawie populacji podzielonej na grupy. Niech N_h będzie licznością h-tej grupy, natomiast $\overline{N}$ będzie średnią liczebnością grupy, czyli $\overline{N} = \frac{1}{G}\sum_{h=1}^{G} N_h$. Zakładamy, że próba S jest dobierana za pomocą znanego w metodzie reprezentacyjnej dwustopniowego schematu losowania. Pierwszy stopień losowania polega na bezzwrotnym losowaniu g grup, przy czym zakłada się, że każda z grup jest losowana z tym samym prawdopodobieństwem jej wyboru do próby. Następnie z każdej tak wybranej grupy Ω_h, gdzie $h \in S$, jest losowana bezzwrotnie próbka prosta S_h o liczebności n_h. Plan tak losowanej próby określa wzór [por. np. Wywiał (1992)]:

$$P_d(s) = \binom{G}{g}^{-1} \prod_{h=1}^{g} \binom{N_h}{n_h}^{-1}$$

Studiujemy tutaj własności strategii $(\tilde{\mathbf{y}}_{gS}, P_d)$, gdzie $\tilde{\mathbf{y}}_{gS} = [\tilde{y}_{giS} \ldots \tilde{y}_{gim}]$, przy czym:

$$\tilde{y}_{giS} = \frac{1}{g\overline{N}} \sum_{h \in S} N_h \bar{y}_{s_h} \tag{8.1}$$

gdzie:

$$\bar{y}_{S_h} = \frac{1}{n_h} \sum_{k \in S_h} y_k$$

Strategia $(\tilde{\mathbf{y}}_{gS}, P_d)$ daje nieobciążone oceny wektora $\bar{\mathbf{y}}$ [por. np. Zasępa (1972), s. 229]. Macierz wariancji i kowariancji strategii $(\tilde{\mathbf{y}}_{gS}, P_d)$ można wyprowadzić podobnie jak wariancję każdej jego składowej [por. Zasępa (1972), s. 338-346]. Macierz ta ma postać:

$$\mathbf{V}(\tilde{\mathbf{y}}_{gS}, P_d) = \frac{G-g}{Gg}\mathbf{C}_{*m} + \frac{1}{gG}\sum_{h=1}^{G} \frac{N_h(N_h - n_h)}{n_h}\mathbf{C}_{*h} \tag{8.2}$$

przy czym $\mathbf{C}_{*m} = \frac{G}{G-1}\mathbf{C}_m$, gdzie elementy macierzy $\mathbf{C}_m = [c_m(y_i, y_j)]$ określa wzór (7.23). Elementy macierzy wariancji i kowariancji $\mathbf{C}_{*h} = [c_{*h}(y_i, y_j)]$ wektora cech y w h-tej grupie określa wzór:

$$c_{*h}(y_i, y_j) = \frac{1}{N_h - 1}\sum_{k \in \Omega_h}(y_{ik} - \bar{y}_{ih})(y_{jk} - \bar{y}_{jh}) \tag{8.3}$$

Estymator macierzy $\mathbf{V}(\tilde{\bar{\mathbf{y}}}_{gS}, P_d)$ otrzymujemy zastępując elementy macierzy $\mathbf{C}_{*m}$ i $\mathbf{C}_{*h}$ odpowiednimi ich ocenami wyznaczonymi na podstawie próby. Estymatory elementów tych macierzy są bezpośrednimi uogólnieniami estymatorów wariancji składowych wektora $\bar{\mathbf{y}}_{gS}$, które można znaleźć w pracy Zasępy (1972), s. 230. Estymatorem i-tego oraz j-tego elementu macierzy $\mathbf{C}_{*m}$ jest kowariancja:

$$c_{*mS}(y_i, y_j) = \frac{1}{g-1}\sum_{h=1}^{G}(\bar{y}_{iS_h} - \bar{y}_{iS})(y_{jS_h} - \bar{y}_{jS})$$

gdzie:

$$\bar{y}_{iS} = \frac{1}{g}\sum_{h=1}^{g}\bar{y}_{iS_h}$$

natomiast estymatorami nieobciążonymi elementów macierzy $\mathbf{C}_{*h}$ są statystyki:

$$c_{*hS}(y_i, y_j) = \frac{1}{N_h - 1}\sum_{k \in S_h}(y_{ik} - \bar{y}_{iS_h})(y_{jk} - \bar{y}_{jS_h}) \tag{8.4}$$

Zakładamy, że próba S jest automatycznie wyważana, co oznacza [por. Zasępa (1972), s. 230], że z każdej wybranej do próby S grupy Ω_h jest losowana próbka S_h o liczebności równej ustalonej frakcji f elementów tej grupy, czyli dla każdego h=1,...,G mamy:

$$n_h = fN_h \tag{8.5}$$

Wtedy estymator i-tej średniej ma postać:

$$\tilde{\bar{y}}_{igS} = \frac{1}{g\bar{N}f}\sum_{h \in S}\sum_{k \in S_h} y_{ik}, \quad i = 1, \ldots, G \tag{8.6}$$

Oznaczając plan losowania próby dwustopniowej zrównoważonej przez P'_d na podstawie wzorów (8.2) i (8.5) wnioskujemy, że:

$$\mathbf{V}(\tilde{\bar{\mathbf{y}}}_{gS}, P'_d) = \frac{G-g}{Gg}\mathbf{C}_{*m} + \frac{1-f}{gf}\bar{N}\mathbf{C}_{*w} \tag{8.7}$$

gdzie:

$$\mathbf{C}_{*w} = \sum_{h=1}^{G} w_h \mathbf{C}_{*h}, \qquad w_h = \frac{N_h}{N} \tag{8.8}$$

Macierz $\mathbf{C}_{*w}$ nazywamy macierzą wariancji i kowariancji wewnątrzgrupową cech, natomiast określoną wcześniej macierz $\mathbf{C}_{*m}$ międzygrupową macierzą wariancji i kowariancji cech.

Po to, by zapisać macierz $\mathbf{V}(\tilde{\mathbf{y}}_{gS}, P'_d)$ w dogodny sposób przy analizie jej wyznacznika, wprowadzamy dodatkowo następujące oznaczenia: $\mathbf{y} = [\mathbf{y}_{1\#\#} \ldots \mathbf{y}_{m\#\#}]$ jest macierzą o wymiarach N × m obserwacji zmiennych w populacji, przy czym $\mathbf{y}_{i\#\#} = [\mathbf{y}_{i\#1}^T \mathrm{L}\ \mathbf{y}_{i\#G}^T]^T$ jest wektorem kolumnowym o wymiarze N×1 obserwacji i-tej zmiennej w populacji. Z kolei $\mathbf{y}_{i\#h} = [y_{il_h} \mathrm{L}\ y_{iN_h}]^T$ jest podwektorem kolumnowym o wymiarze N×1 obserwacji i-tej zmiennej w h-tej grupie. Niech $\mathbf{J}_a$ będzie wektorem jednostkowym o wymiarze a×1. Sumę wartości i-tej cechy w h-tej grupie oznaczamy przez $z_{hi} = \mathbf{y}_{i\#h}^T \mathbf{J}_{N_h}$. Niech $\mathbf{z}=[\mathbf{z}_{\#1} \ldots \mathbf{z}_{\#m}]$ będzie macierzą o wymiarze G×m, przy czym $\mathbf{z}_{\#i}^T = [z_{1i} \mathrm{L}\ z_{Gi}]$ wektorem sum wartości i-tej zmiennej w poszczególnych grupach. Przez $\mathbf{D}_J$ oznaczamy macierz o wymiarze G×N w postaci:

$$\mathbf{D}_{J=} \begin{bmatrix} \mathbf{J}_{N_1}^T & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{J}_{N_2}^T & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \mathbf{J}_{N_G}^T \end{bmatrix} \tag{8.9}$$

Z definicji macierzy **y**, **z**, **D** wynika:

$$\mathbf{z} = \mathbf{D}_J \mathbf{y} \tag{8.10}$$

Przez $\bar{\mathbf{z}} = [\bar{z}_1 \mathrm{L}\ \bar{z}_m]^T$ oznaczamy wektor średnich sum wartości poszczególnych zmiennych przypadających na grupę, który można zapisać jako funkcję macierzy **z**:

$$\bar{\mathbf{z}} = \frac{1}{G} \mathbf{J}_G^T \mathbf{z} \tag{8.11}$$

Wtedy macierz $\mathbf{C}_{*m}$ jest następującą funkcją wektora **z**:

$$C_{*m} = \frac{1}{G-1} (\mathbf{z} - \mathbf{J}_G \bar{\mathbf{z}})^T (\mathbf{z} - \mathbf{J}_G \bar{\mathbf{z}}) \tag{8.12}$$

Na podstawie wzorów (8.10) i (8.11) mamy:

$$\mathbf{z} - \mathbf{J}_G \bar{\mathbf{z}} = \mathbf{A}_G \mathbf{D}_J \mathbf{y} \tag{8.13}$$

gdzie:

$$\mathbf{A}_G = \mathbf{I}_G - \frac{1}{G}\mathbf{J}_G\mathbf{J}_G^T, \qquad \mathbf{A}_G^2 = \mathbf{A}_G \tag{8.14}$$

Z idempotentności macierzy $\mathbf{A}_G$ oraz ze wzorów (8.12) i (8.13) wynika, że:

$$\mathbf{C}_{*m} = \mathbf{y}^T\mathbf{A}\mathbf{y} \tag{8.15}$$

gdzie:

$$\mathbf{A} = \frac{1}{G-1}\mathbf{D}_J^T\mathbf{A}_G\mathbf{D}_J \tag{8.16}$$

Określone wzorem (8.3) kowariancje przekształcamy jak następuje:

$$c_{*h}\left(y_i, y_j\right) = \frac{1}{N_h - 1}\left(\mathbf{y}_{i\#h} - \mathbf{J}_{N_h}\bar{\mathbf{y}}_{ih}\right)^T\left(\mathbf{y}_{j\#h} - \mathbf{J}_{N_h}\bar{\mathbf{y}}_{jh}\right)$$

gdzie średnią i-tej zmiennej z h-tej grupy wyliczamy z wzoru:

$$\bar{y}_{ih} = \frac{1}{N_h}\mathbf{J}_{N_h}^T\mathbf{y}_{i\#h}$$

Stąd wynika, że

$$\mathbf{y}_{i\#h} - \mathbf{J}_{N_h}\bar{y}_{ih} = \mathbf{B}_h\mathbf{y}_{i\#h}$$

gdzie:

$$\mathbf{B}_h = \mathbf{I}_{Nh} - \frac{1}{N_h}\mathbf{J}_{N_h}\mathbf{J}_{N_h}^T, \qquad \mathbf{B}_h^2 = \mathbf{B}_h \tag{8.17}$$

Skutkiem tego wyprowadzenia jest uproszczenie wzoru na kowariancję, danej wzorem (8.3), który przyjmuje postać:

$$c_{*h}\left(y_i, y_j\right) = \frac{1}{N_h - 1}\mathbf{y}_{i\#h}^T\mathbf{B}_h\mathbf{y}_{j\#h} \tag{8.18}$$

Stąd wynika, że element macierzy kowariancji wewnątrzgrupowej określonej wzorem (8.8) można zapisać wzorem:

$$c_{*w}\left(y_i, y_j\right) = \frac{1}{N}\sum_{h=1}^{G}\frac{N_h}{N_h - 1}\mathbf{y}_{i\#h}^T\mathbf{B}_h\mathbf{y}_{j\#h} \tag{8.19}$$

Wprowadźmy macierz postaci:

$$\mathbf{B} = \frac{1}{N}\begin{bmatrix} q_1\mathbf{B}_1 & 0 & \dots & 0 \\ 0 & q_2\mathbf{B}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & q_G\mathbf{B}_G \end{bmatrix} \tag{8.20}$$

gdzie:

$$q_h = \frac{N_h}{N_h - 1}$$

Wyrażenie (8.19) można więc zapisać wzorem:

$$c_{*h}\left(y_i, y_j\right) = \mathbf{y}_{i\#\#}^{T}\mathbf{B}\mathbf{y}_{j\#\#}$$

Stąd już wnioskujemy, że dana wzorem (8.8) macierz $\mathbf{C}_{*w}$ wyraża się równaniem:

$$\mathbf{C}_{*w} = \mathbf{y}^{T}\mathbf{B}\mathbf{y} \tag{8.21}$$

Stąd i ze wzorów (8.15) i (8.7) już otrzymujemy:

$$\mathbf{V}(\$_{wS}, P_d') = a\mathbf{y}^{T}\mathbf{A}\mathbf{y} + b\mathbf{y}^{T}\mathbf{B}\mathbf{y} \tag{8.22}$$

gdzie:

$$a = \frac{G - g}{Gg}, \qquad b = \frac{1 - f}{gf}\overline{N} \tag{8.23}$$

Załóżmy, że macierz $\mathbf{C}_{*m}$ jest dodatnio określona. Oznaczmy przez $\mathbf{H}$ taką macierz ortogonalną stopnia m, tj. $\mathbf{H}^{T}\mathbf{H} = \mathbf{I}_m$, że

$$\mathbf{H}^{T}\mathbf{C}_{*m}\mathbf{H} = \mathbf{D}_m \tag{8.24}$$

gdzie $\mathbf{D}_m$ jest macierzą diagonalną stopnia m pierwiastków charakterystycznych d_{mi} (i=1,...,m) macierzy $\mathbf{C}_{*m}$. Z dodatniej określoności macierzy $\mathbf{C}_{*m}$ wynika, że elementy diagonalne macierzy $\mathbf{D}_m$ są dodatnie. Przez $\mathbf{D}_m^{-½}$ oznaczamy macierz diagonalną stopnia m, której każdy element diagonalny jest odwrotnością pierwiastka z odpowiedniego elementu diagonalnego macierzy $\mathbf{D}_m$. Przyjmijmy, że $\mathbf{F}$ jest taką macierzą ortogonalną stopnia m, że $\mathbf{F}^{T}\mathbf{F} = \mathbf{I}_m$ i:

$$\mathbf{F}^{T}\mathbf{D}_m^{-½}\mathbf{H}^{T}\mathbf{C}_{*w}\mathbf{H}\mathbf{D}_m^{-½}\mathbf{F} = \mathbf{D}_w \tag{8.25}$$

$\mathbf{D}_w = [d_i]$ jest macierzą diagonalną stopnia m pierwiastków charakterystycznych macierzy $\mathbf{D}_m^{-½}\mathbf{H}^{T}\mathbf{C}_{*w}\mathbf{H}\mathbf{D}_m^{-½}$. Oznaczamy:

$$\mathbf{G} = \mathbf{H}\mathbf{D}_m^{-\frac{1}{2}}\mathbf{F} \tag{8.26}$$

Stąd i z równań (8.24) i (8.25) wynika, że jednocześnie zachodzą dwa równania [por. np, Rao (1982), s. 58]:

$$\mathbf{G}^T\mathbf{C}_{*m}\mathbf{G} = \mathbf{I}_m, \qquad \mathbf{G}^T\mathbf{C}_{*w}\mathbf{G} = \mathbf{D}_w \tag{8.27}$$

Stąd mamy:

$$\mathbf{C}_{*m} = (\mathbf{G}^T)^{-1}\mathbf{G}^{-1}, \qquad \mathbf{C}_{*w} = (\mathbf{G}^T)^{-1}\mathbf{D}_w\mathbf{G}^{-1} \tag{8.28}$$

Wynik ten pozwala nam wyrażenie (8.7) zapisać w postaci:

$$\mathbf{V}(\tilde{\mathbf{y}}_{gS}, P_d') = (\mathbf{G}^T)^{-1}\left(a\mathbf{I}_m + b\mathbf{D}_w\right)\mathbf{G}^{-1} \tag{8.29}$$

gdzie a i b określa wzór (8.23).

Przekształćmy macierz obserwacji **y** na macierz **u** o wymiarach N×m za pomocą równania:

$$\mathbf{u} = \mathbf{y}\mathbf{G} \tag{8.30}$$

Obserwacje zmiennych zawarte w macierzy **y** są przekształcane w wartości nowych zmiennych, które tworzą kolumny macierzy **u** o wymiarach N×m. Stąd i ze wzorów (8.15), (8.21), (8.27) i (8.30) mamy, że $\mathbf{I}_m = \mathbf{u}^T\mathbf{A}\mathbf{u}$, $\mathbf{D}_w = \mathbf{u}^T\mathbf{B}\mathbf{u}$. Zatem wyrażenie (8.29) sprowadzamy do postaci:

$$\mathbf{V}(\tilde{\mathbf{y}}_{gS}, P_d') = (\mathbf{G}^T)^{-1}\left(a\mathbf{u}^T\mathbf{A}\mathbf{u} + b\mathbf{u}^T\mathbf{B}\mathbf{u}\right)\mathbf{G}^{-1} \tag{8.31}$$

Stąd wnioskujemy, że macierz jednostkowa $\mathbf{I}_m = \mathbf{u}^T\mathbf{A}\mathbf{u}$ jest macierzą wariancji i kowariancji międzygrupowej zmiennych, których obserwacje są elementami macierzy **u**. Z kolei macierz diagonalna $\mathbf{D}_w = \mathbf{u}^T\mathbf{B}\mathbf{u}$ jest macierzą wariancji i kowariancji wewnątrzgrupowej tych zmiennych. Zatem element diagonalny d_i macierzy $\mathbf{D}_w$ jest wariancją wewnątrzgrupową zmiennej, której obserwacjami są elementy i-tej kolumny macierzy **u**.

Przyjmując, że macierz $\mathbf{C}_{*w}$ jest nieosobliwa, można otrzymać podobny rozkład macierzy $\mathbf{V}(\tilde{\mathbf{y}}_{gS}, P_d')$, w którym macierzy $\mathbf{C}_m$ będzie odpowiadać macierz jednostkowa, a macierzy $\mathbf{C}_w$ macierz diagonalna.

Na podstawie znanych własności wyznacznika macierzy oraz wzorów (8.23), (8.28) i (8.29) otrzymujemy:

$$\det \mathbf{V}(\tilde{\mathbf{y}}_{gS}, P_d') = \det \mathbf{C}_{*m} \prod_{i=1}^{m}\left(a + bd_i\right)$$

$$\det \mathbf{V}(\tilde{\mathbf{y}}_{gS}, P_d') = \det \mathbf{C}_{*m} \prod_{i=1}^{m}\left(\frac{1 - \bar{N}d_i}{g} + \frac{\bar{N}d_i}{gf} - \frac{1}{G}\right) \tag{8.32}$$

W dalszych rozważaniach zakładamy, że elementy diagonalne macierzy $\mathbf{D}_w$ spełniają nierówności: $0 \le d_{wi} \le \frac{1}{\overline{N}}$. Przez $\mathbf{v}_{*m}(y_i)$ i $v_{*w}(y_i)$ oznaczmy wariancję odpowiednio między i wewnątrzgrupową, które są elementami diagonalnymi macierzy $\mathbf{C}_{*m}$ i C_{*w}. Wtedy na podstawie wzoru (8.7) mamy [por. przypadek jednowymiarowy studiowany przez Zasępę (1972)]:

$$q^2(\bar{\mathbf{y}}_{gS}, P_d') = \mathrm{tr}\mathbf{V}(\bar{\mathbf{y}}_{gS}, P_d') = \frac{G-g}{Gg}\sum_{h=1}^{m} v_{*m}(y_i) + \frac{1-f}{gf}\overline{N}\sum_{h=1}^{m} v_{*w}(y_i) \tag{8.33}$$

Wprowadźmy oznaczenie:

$$d_{\#i} = v_{*m}^{-1}(y_i)v_{*w}(y_i) \tag{8.34}$$

Wtedy zakładając, że $\mathbf{C}_{*m}$ jest dodatnio określona, mamy:

$$q^2(\bar{\mathbf{y}}_{gS}, P_d') = \frac{1}{2}\mathrm{tr}\mathbf{C}_{*m}\left(1 + \overline{N}\frac{1-f}{f}\bar{d}_{\#}\right) - \frac{1}{G}\mathrm{tr}\mathbf{C}_{*m} \tag{8.35}$$

gdzie:

$$\bar{d}_{\#} = \sum_{i=1}^{m} w_i d_{\#i}, \qquad w_i = \frac{v_{*m}(y_i)}{\mathrm{tr}\mathbf{C}_{*m}} \tag{8.36}$$

Uogólniając wnioski otrzymane przez Zasępę (1972), s. 236 stwierdzamy, że współczynnik $\bar{d}_{\#}$ ocenia średni stopień zróżnicowania wewnątrzgrupowego cechy m wymiarowej. Zróżnicowanie to jest tym większe, im jest wyższy poziom wartości wskaźnika $\bar{d}_{\#}$.

Wniosek 8.1: Kwadrat średniego promienia strategii $(\bar{\mathbf{y}}_{gS}, P_d')$ jest rosnącą funkcją wartości wskaźnika $\bar{d}_{\#}$.

8.2. Minimalizacja oczekiwanych kosztów badania przy ustalonej dokładności estymacji

Przez k_1 oznaczmy koszt jednostkowy przygotowania losowania dwustopniowego próby i jego realizacji, a przez k_2 koszt jednostkowy obserwacji wartości zmiennych w wylosowanej próbie. Wtedy [por. np. Konijn (1973), s. 322] oczekiwany koszt losowania i obserwacji w niej wartości zmiennych określa funkcja:

$$k(g, n_1, \ldots, n_G) = k_1 g + k_2 \frac{g}{G}\sum_{h=1}^{G} n_h \tag{8.37}$$

Stąd i ze wzoru (8.5) wynika, że w przypadku prób automatycznie wyważonych mamy:

$$k(g,f) = k_1 g + k_2 \overline{N} g f \tag{8.38}$$

przy czym używane tutaj symbole wyjaśniono w paragrafie 8.1. W kolejnych punktach formułujemy i rozwiązujemy zadania ustalania optymalnych liczebności g grup i frakcji f, które, podkreślmy, minimalizują oczekiwany koszt obserwacji, a nie koszt rzeczywisty. Realny koszt obserwacji cech jest możliwy do wyliczenia dopiero po wylosowaniu próby dwustopniowej, bowiem zależy on od liczebności grup, ktore znajdą się w próbie.

8.2.1. Ustalony poziom dopuszczalny funkcji ryzyka kwadratowego

Analizowana dalej funkcja ryzyka ma postać:

$$h(g,f) = \sum_{i=1}^{m} a_i D^2\left(\tilde{y}_{igS}\right) = \frac{G-g}{Gg} \sum_{i=1}^{m} a_i v_{*m}(y_i) + \frac{1-f}{gf} \overline{N} \sum_{i=1}^{m} a_i v_{*w}(y_i)$$

lub jeśli $v_{*m}(y_i) > 0$ dla każdego i=1,...,m, to:

$$h(g,f) = q_m \left\{ \frac{1}{g}\left(1 + \frac{1-f}{f} \overline{N} \tilde{d}\right) - \frac{1}{G} \right\} \tag{8.39}$$

gdzie:

$$q_m = \sum_{i=1}^{m} a_i v_{*m}(y_i) \tag{8.40}$$

$$\tilde{d} = \sum_{i=1}^{m} d_{\#i} w_i', \qquad w_i' = \frac{a_i v_{*m}(y_i)}{q_m} \tag{8.41}$$

przy czym $d_{\#i}$ wyjaśnia wzór (8.34).

Zadanie optymalizacyjne polega na takim ustaleniu liczebności g i frakcji f, aby dana wzorem (8.38) funkcja kosztów osiągnęła minimum przy ustalonym poziomie dopuszczalnym ryzyka estymacji, czyli:

$$\begin{cases} k(g,f) = \text{minimum} \\ h(g,f) \le h_d \\ 0 < f \le 1, \qquad 1 < g \le G \end{cases} \tag{8.42}$$

Podczas poszukiwania rozwiązania zadania wykorzystujemy następujące twierdzenia.

Lemat 8.1 [Wywiał (1992)]: Określona wzorem (8.38) funkcja k(g,f) ma dodatnią pochodną w kierunku każdego wektora zaczepionego w punkcie (g=1,f=0) i końcu w punkcie (g^o,f^o), gdzie $g^o > 1$ i $f^o > 0$.

Lemat 8.2 [Wywiał (1992)]: Jeśli $\tilde{d} < \overline{N}^{-1}$, to funkcja h(g,f) jest ściśle wypukła dla f >0 i g>0.

Wprowadźmy następujące oznaczenia:

$$g_* = \frac{q_m}{h_\#}\left\{1 - \overline{N}\tilde{d} + \overline{N}\sqrt{\frac{k_2\left(1-\tilde{d}\right)\tilde{d}}{k_1}}\right\} \tag{8.43}$$

$$g_{**} = q_m h_\#^{-1} \tag{8.44}$$

$$f_* = \sqrt{\frac{k_1}{k_2}\frac{\tilde{d}}{1-\overline{N}\tilde{d}}} \tag{8.45}$$

$$f_{**} = \frac{\overline{N}\tilde{d}}{Gg_{**}^{-1} + \overline{N}\tilde{d} - 1} \tag{8.46}$$

$$f_{***} = \frac{\overline{N}\tilde{d}}{g_{**}^{-1} + \overline{N}\tilde{d} - 1} \tag{8.47}$$

Oznaczając przez $(\underline{g},\underline{f})$ otrzymane przez Wywiała (1992) rozwiązanie zadania (8.42) zapisujemy w następującej formie:

Jeśli $\tilde{d} < \frac{1}{N}$ oraz

a) jeśli $g_* \le G$ i $f_* \le 1$, to para optymalna $(\underline{g},\underline{f}) = (g_*,f_*)$,

b) jeśli $f_*>1$ i $g<G$, to $(\underline{g},\underline{f}) = (g_{**},1)$,

c) jeśli $g_*>G$ i $f_*<1$, to $(\underline{g},\underline{f}) = (G,f_{**})$,

d) jeśli $g_* <1$ i $f_* <1$, to $(\underline{g},\underline{f}) = (1,f_{***})$.

Dodajmy, że Kokan (1963) i Kokan i Khan (1967) rozwiązywali zadanie (8.42) lecz dla przypadku, gdy podczas drugiego stopnia losowania są wybierane próbki równoliczne z wcześniej dobranych grup. Proponowany przez nich sposób rozwiązywania bezpośrednio rozciąga się na przypadek optymalizacji liczebności zrównoważonej próby dwustopniowej, który rozważał Wywiał (1992).

8.2.2. Ustalony poziom dopuszczalny uogólnionej wariancji

Zakładamy, że ilość grup G jest na tyle duża, iż występujący w wyrażeniu (8.32) składnik G^{-1} można pominąć. Wtedy oznaczając uogólnioną wariancję wektora $\bar{\mathbf{y}}_{gS}$ przez u(g,f) mamy:

$$u(g,f) = \frac{1}{g^m} \det \mathbf{C}_{*m} \prod_{i=1}^{m} \left(1 - \overline{N} d_i + \frac{1}{f} \overline{N} d_i\right) \tag{8.48}$$

przy czym przez d_i oznaczono element diagonalny macierzy $\mathbf{D}_w$ określonej wyrażeniami (8.25)-(8.27).

Nasze zadanie polega na znalezieniu takich argumentów g i f, aby oczekiwany koszt badania, dany wzorem (8.38), osiągnął minimum przy ustalonym poziomie dopuszczalnym uogólnionej wariancji. Rozwiązanie postawionego zadania nie zmieni się, jeśli uogólnioną wariancję zastąpimy jej pierwiastkiem m-tego stopnia, co oznaczamy przez:

$$u_{\#}(g,f) = \sqrt[m]{\det \mathbf{V}(\bar{\mathbf{y}}_{gS})} = \frac{1}{g} c_o \sqrt[m]{\prod_{i=1}^{m} \left(b_i + \frac{1}{f}\right)} \tag{8.49}$$

gdzie:

$$c_o = \sqrt[m]{\det \mathbf{C}_{*m}} \tag{8.50}$$

$$b_i = \frac{1}{\overline{N} d_i} - 1 \tag{8.51}$$

Zatem pierwotnie sformułowane zadanie jest równoważne problemowi[53]:

$$\begin{cases} k(g,f) = \text{minimum} \\ u_{\#}(g,f) \le u_d \\ g \ge 1, \qquad 0 < f \le 1 \end{cases} \tag{8.52}$$

Lemat 8.3 [Wywiał (1992)]: Określona wzorem (8.49) funkcja $u_{\#}$ jest ściśle wypukła dla g>0 i f>0.

Wprowadźmy oznaczenia: Niech f_* będzie pierwiastkiem równania:

$$\frac{1}{k_0 + f} - \frac{1}{mf} \sum_{i=1}^{m} \frac{1}{b_i f + 1} = 0 \tag{8.53}$$

gdzie:

[53] Por Wywiał (1987).

$$k_0 = \frac{k_1}{\bar{N} k_2} \tag{8.54}$$

$$g_* = \frac{c_o}{u_d} \sqrt[m]{\prod_{i=1}^{m} \left(b_i - \frac{1}{f_*} \right)} \tag{8.55}$$

Wartość f' jest pierwiastkiem równania $u_{\#}(1,f)=u_d$, natomiast g' pierwiastkiem równania $u_{\#}(g,1)=u_d$.

Korzystając z lematu 8.3 Wywiał (1992) otrzymuje następujące rozwiązanie ($\underline{g}$,$\underline{f}$) postawionego zadania (8.52):

Gdy jest spełniona nierówność

$$\bar{d} > \frac{k_2}{k_1 + \bar{N} k_2} \tag{8.56}$$

gdzie:

$$\bar{d} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} d_i} \tag{8.57}$$

oraz

a) jeśli $f_* \in (0,1>$ i $g_* > 1$, to ($\underline{g}$,$\underline{f}$)=(g_*,f_*),

b) jeżeli $f_* > 1$, to ($\underline{g}$,$\underline{f}$) = (g',1),

c) gdy $g_* < 1$, to ($\underline{g}$,$\underline{f}$) = (1,f').

Potrzebny pierwiastek f_* równania (8.53) należy wyliczyć za pomocą odpowiedniej metody rozwiązywania równań nieliniowych.

8.3. Maksymalizacja precyzji estymacji przy ustalonych kosztach badania

8.3.1. Minimalizacja funkcji kwadratowej ryzyka

Naszym zadaniem jest znalezienie takiej liczebności $\underline{g}$ grup i frakcji $\underline{f}$ potem losowanych z nich elementów, aby dana wzorem (8.39) funkcja ryzyka strategii estymacji osiągnęła minimum przy ustalonych oczekiwanych kosztach dopuszczalnych obserwacji zmiennych określonych wzorem (8.38). Problem ten zapisujemy syntetycznie następująco:

$$\begin{cases} h(g, f) = \text{minimum} \\ k(g, f) \le K \\ f_d \le f \le 1, \quad 1 \le g \le G \end{cases} \tag{8.58}$$

Formalnie, sposób rozwiązywania zadania jest taki sam jak dla m=1, który jest prezentowany przez Zasępę (1972). Poszukiwanie rozwiązania znacznie upraszcza się, gdy przyjmiemy, że:

$$x=gf, \qquad g=g \tag{8.59}$$

Wtedy zagadnienie (8.58) jest równoważne następującemu:

$$\begin{cases} h_{@}(g,x) = \text{minimum} \\ k_{@}(g,x) \le K \\ 0 < x \le g \le G, \quad g \ge 1 \end{cases} \tag{8.60}$$

przy czym:

$$h_{@}(g,x) = \frac{1}{g}\left(1-\overline{N}\widetilde{d}\right) + \frac{1}{x}\overline{N}\widetilde{d} - \frac{1}{G} \tag{8.61}$$

$$k_{@}(g,x) = k_1 g + k_2 \overline{N} x \tag{8.62}$$

Poprzez badanie określoności hesjanu funkcji $h_{@}$ wykazujemy prawie natychmiast następującą własność:

Lemat 8.4: Jeśli $\widetilde{d}\overline{N} < 1$, to funkcja $h_{@}$ jest ściśle wypukła dla x>0 i g>0.

Wprowadźmy oznaczenia:

$$g' = x' = \frac{f_d K}{k_1 + k_2 \overline{N} f_d} \tag{8.63}$$

$$g'' = \frac{K - k_2 \overline{N} f_d}{k_1} \tag{8.64}$$

$$x'' = \frac{K - k_1 G}{\overline{N} k_2} \tag{8.65}$$

$$g_* = e\sqrt{\frac{1-\overline{N}\widetilde{d}}{k_1}}, \qquad x_* = e\sqrt{\frac{\widetilde{d}}{k_2}} \tag{8.66}$$

gdzie:

$$e = K\left(\sqrt{\left(1-\overline{N}\widetilde{d}\right)k_1} + \overline{N}\sqrt{k_2\widetilde{d}}\right)^{-1}$$

Wywiał (1992) otrzymał następujące rozwiązanie (g,f) zadania (8.58): Jeżeli $\widetilde{d}\overline{N} < 1$ i $G > g'$ oraz

a) gdy $P_* \in \overline{AC}$, to $\left(\underline{g}, \underline{f}\right) = \left(g_*, \frac{x_*}{g_*}\right)$,

b) gdy $g_* < g'$, to $(\underline{g},\underline{f}) = (g',1)$,

c) gdy $G < g''$ i $g_* > G$, to $\left(\underline{g}, \underline{f}\right) = \left(G, \frac{x''}{G}\right)$.

8.3.2. Minimalizacja uogólnionej wariancji

Zakładamy, że liczba grup jest nieograniczona, czyli $G \to \infty$. Zadanie polega na takim wyznaczeniu liczebności $\underline{g}$ losowanych grup i frakcji $\underline{f}$ losowanych z nich elementów, aby uogólniona wariancja strategii $(\bar{y}_{gS}, P_d')$ osiągnęła minimum przy ustalonym dopuszczalnym koszcie przeciętnym obserwacji zmiennych. Tak sformułowane zadanie ma takie same rozwiązanie, jak problem minimalizacji pierwiastka m-tego stopnia z uogólnionej wariancji przy tym samym warunku kosztowym. Zagadnienie to ma postać[54] :

$$\begin{cases} u_{\#}\left(g, f\right) = \text{minimum} \\ k\left(g, f\right) \le K \\ g \ge 1, \quad f_d \le f \le 1 \end{cases} \tag{8.67}$$

przy czym funkcje $u_{\#}$, k określają odpowiednio wzory (8.49), (8.38). Wprowadźmy oznaczenia:

$$g_* = \frac{K}{k_1 + \bar{N} k_2 f_*} \tag{8.68}$$

Gdy $f_* > 1$, to $\underline{f}=1$ i z powyższego równania wyliczamy:

$$g' = \frac{K}{k_1 + \bar{N} k_2} \tag{8.69}$$

Wywiał (1992) otrzymuje, że jeśli $k_1 + k_2 \bar{N} < K$ i jest spełniona nierówność (8.56) oraz

a) gdy $f_* \le 1$, to $(\underline{g},\underline{f})=(g_*,f_*)$,

b) jeśli $f_* > 1$, to $(\underline{g},\underline{f})=(g',1)$.

Miejsca zerowego f_* określonej wzorem (8.53) pierwszej pochodnej funkcji $k_*(f)$ należy poszukiwać za pomocą odpowiednich metod numerycznych rozwiązywania równań nieliniowych.

---

[54] Por. Wywiał (1987).

9. WNIOSKOWANIE O WEKTORZE ŚREDNICH W BADANIACH DWUOKRESOWYCH POPULACJI

Badania reprezentacyjne cech populacji zwykle prowadzi się powtarzalnie co najmniej w dwu okresach czasu. Naturalną jest idea wykorzystania w bieżącym badaniu statystycznym informacji uzyskanych o kształtowaniu się rozkładu cech w poprzednich okresach czasu. Nie bez racji należy oczekiwać, że uwzględnienie przy estymacji wektora wartości średnich w bieżącym okresie informacji o parametrach tych zmiennych z przeszłości prowadzi do podniesienia dokładności ich oceny. Dla przypadku jednowymiarowego sygnalizowany probleme m.in. już analizowali Cochran (1963), Eckler (1955) oraz Tripathi, Mir i Chaturvedi (1989). Prezentowane tu rozważania sprowadzają się do uogólnienia znanych wyników dotyczących estymacji średniej jednej cechy na przypadek szacowania wektora wartości średnich.

Drugi z analizowanych tu problemów dotyczy testowania równości rozkładów prawdopodobieństwa obserwowanych w próbie prostej w dwóch kolejnych okresach czasu. Rozkłady te mogą być zależne, co ma zasadniczy wpływ na konstrukcję sprawdzianu testu dla weryfikacji postawionej hipotezy.

9.1. Podstawowe własności estymatorów

Symbolem $\bar{\mathbf{y}}_t = \{\bar{y}_{t1} \quad \bar{y}_{t2} \ldots \bar{y}_{ta}\}$ oznaczamy wektor średnich badanych zmiennych w okresie bieżącym (t=1) i podstawowym (t=0). Macierz wariancji i kowariancji badanych zmiennych w populacji oznaczamy następującą macierzą blokową

$$\mathbf{C} = [c_{ij}] = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{10} \\ \mathbf{C}_{01} & \mathbf{C}_{00} \end{bmatrix}$$

gdzie: $(N-1)c_{ltij} = \sum_{h=1}^{N}(y_{lih} - \bar{y}_{ti})(y_{tjh} - \bar{y}_{tj})$, l,t=0,1; i,j=1,...,r. Przez $\mathbf{C}_{00}$, $\mathbf{C}_{11}$ oznaczono macierze wariancji i kowariancji badanych zmiennych odpowiednio w okresie podstawowym (wyjściowym) i bieżącym, natomiast przez $\mathbf{C}_{10}=\mathbf{C}_{01}^{T}$ macierz kowariancji pomiędzy badanymi zmiennymi z okresu bieżącego, a zmiennymi z okresu podstawowego. Zakładamy, że $\mathbf{C}_{00}$ i $\mathbf{C}_{11}$ są nieosobliwe.

Populację oznaczamy symbolem U. W okresie wyjściowym jest bezzwrotnie losowana z całej populacji próba prosta S_0 składająca się z m-elementów. Następnie losujemy również bezzwrotnie n-elementową próbę prostą S_{10} spośród elementów już dobranej próby S_0. W bieżącym okresie jest analizowana próba S_1 o liczebności k-elementów losowana bezzwrotnie bezpośrednio ze zbioru $U_1 = U\text{-}S_o$.

Oznaczenia wierszowych wektorów średnich z prób są następujące:

$\overline{\mathbf{Y}}_{0S_0}$ - z próby S_0 w okresie t=0;

$\overline{\mathbf{Y}}_{0S_{10}}$ - z próby S_{10} dla t=0;

$\overline{\mathbf{Y}}_{1S_{10}}$ -z próby S_{10} dla t=1;

$\overline{\mathbf{Y}}_{1S_1}$ - z próby S_1 dla t=1;

$\overline{\mathbf{Y}}_{1U_1}$ - ze zbioru U_1 dla t=1.

Wektory średnich typu $\overline{\mathbf{Y}}_{0.}$ są nieobciążonymi estymatorami wektora średnich $\bar{\mathbf{y}}_0$ w okresie wyjściowym, natomiast $\overline{\mathbf{Y}}_{1.}$ są także nieobciążonymi estymatorami wektora średnich $\bar{\mathbf{y}}_1$ w okresie bieżącym.

Ogólna konstrukcja rozważanego dalej estymatora jest następującą kombinacją liniową statystyk:

$$\overline{\mathbf{Y}}_{\#} = b\overline{\mathbf{Y}}_{1S_1} + (1-b)\overline{\mathbf{Y}}_{*}, \qquad 0 \le b \le 1 \tag{9.1}$$

Symbolem $\overline{\mathbf{Y}}_{*}$ oznaczono nieobciążony estymator wektora $\bar{\mathbf{y}}_1$, który jest pewną funkcją średnich z próby i innych momentów badanych zmiennych w obu okresach obserwowanych w próbach S_0 i S_{10}.

Wartość oczekiwana wektora $\overline{\mathbf{Y}}_{\#}$ jest wyznaczana następująco:

$$E(\overline{\mathbf{Y}}_{\#}) = E_0 E_{1/0}(\overline{\mathbf{Y}}_{\#}) \tag{9.2}$$

gdzie przez E_0 oznaczono operator nadziei matematycznej wyznaczanej na podstawie rozkładu prawdopodobieństwa próby prostej S_0, natomiast $E_{1/0}$ jest operatorem nadziei matematycznej wyliczanej na podstawie warunkowego rozkładu prawdopodobieństwa prób prostych S_1', S_{10} przy ustalonym składzie wcześniej już wylosowanej próby S_0. Zatem:

$$E_{1/0}(\overline{\mathbf{Y}}_{\#}) = bE_{1/0}(\overline{\mathbf{Y}}_{1S_1}) + (1-b)E_{1/0}(\overline{\mathbf{Y}}_{*}) \tag{9.3}$$

Załóżmy, że $E_{1/0}(\overline{\mathbf{Y}}_{*}) = \bar{\mathbf{y}}_{S_0}$. Wtedy

$$E_{1/0}(\overline{\mathbf{Y}}_{\#}) = b\overline{\mathbf{y}}_{U_1} + (1-b)\overline{\mathbf{y}}_{S_0} \tag{9.4}$$

Z równania $\overline{\mathbf{y}} = \frac{N-m}{N}\overline{\mathbf{y}}_{U_1} + \frac{m}{N}\overline{\mathbf{y}}_{S_0}$ wyliczamy, że

$$\overline{\mathbf{y}}_{U_1} = \frac{1}{N-m}\left(N\overline{\mathbf{y}} - m\overline{\mathbf{y}}_{S_0}\right) \tag{9.5}$$

Stąd i ze wzoru (9.4) już wynika:

$$E(\overline{\mathbf{Y}}_{\#}) = bE_0(\overline{\mathbf{y}}_{U_1}) + (1-b)E_0(\overline{\mathbf{y}}_{S_0}) = \frac{b}{N-m}E_0\left(N\overline{\mathbf{y}} - m\overline{\mathbf{y}}_{S_0}\right) + (1-b)\overline{\mathbf{y}}_1 =$$

$$= \frac{b}{N-m}\left(N\overline{\mathbf{y}} - mE_0(\overline{\mathbf{y}}_{S_0})\right) + (1-b)\overline{\mathbf{y}}_1 = b\overline{\mathbf{y}}_1 + (1-b)\,\overline{\mathbf{y}}_1 = \overline{\mathbf{y}}_1$$

Zatem estymator $\overline{\mathbf{Y}}_{\#}$ daje nieobciążone oceny wektora $\overline{\mathbf{y}}_1$, jeśli $E_{1/0}(\overline{\mathbf{Y}}_*) = \overline{\mathbf{y}}_{S_0}$.

Macierz wariancji i kowariancji jest wyznaczana na podstawie następującego znanego wyrażenia:

$$\mathbf{V}\,(\overline{\mathbf{Y}}_{\#}) = E_0\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_{\#}) + \mathbf{V}_0(E_{1/0}(\overline{\mathbf{Y}}_{\#})) \tag{9.6}$$

gdzie przez $\mathbf{V}_0(.)$ oznaczono macierz wariancji i kowariancji wyznaczaną na podstawie rozkładu prawdopodobieństwa próby prostej S_0, natomiast $\mathbf{V}_{1/0}(.)$ jest macierzą wariancji i kowariancji wyliczaną na podstawie warunkowego rozkładu prawdopodobieństwa prób prostych S_1' i S_{10} przy ustalonym składzie wcześniej już wylosowanej próby S_0. Gdy skład próby S_0 jest ustalony, to próby S_{10} i S_1 są losowane z rozłącznych zbiorów S_0 i U_1. Zatem są one niezależne. Prowadzi to do następującego rezultatu:

$$\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_{\#}) = \mathbf{V}_{1/0}(b\overline{\mathbf{Y}}_{1S_1} + (1-b)\overline{\mathbf{Y}}_*) = b^2\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_{1S_1}) + (1-b)^2\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_*)$$

$$\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_{\#}) = b^2\frac{N-m-k}{(N-m)k}\mathbf{C}_{1U_1} + (1-b)^2\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_*) \tag{9.7}$$

Przez $\mathbf{C}_{1U_1} = [c_{1ijU_1}]$ oznaczono macierz wariancji i kowariancji cech badanych w okresie bieżącym w zbiorze U_1, gdzie

$$c_{1ijU_1} = \frac{1}{N-m-1}\sum_{h\in U_1}(y_{1ih} - \overline{y}_{1iU_1})(y_{1jh} - \overline{y}_{1jU_1}), \qquad \overline{y}_{1iU_1} = \frac{1}{N-m}\sum_{h\in U_1} y_{1ih}$$

Zbiór U_1 jest dopełnieniem próby S_0 do populacji U. Zatem rozkład prawdopodobieństwa pojawiania się zbioru U_1 jest taki sam jak próby S_0. Konsekwencją tego jest to, że znane

własności statystyki z próby S_0 przenoszą się także na statystyki ze zbioru (próby) U_1. Wtedy kowariancja c_{1ijU_1} jest nieobciążonym estymatorem odpowiedniego elementu macierzy $\mathbf{C}_{11}$ określonej w wyrażeniu (9.1). Stąd i ze wzoru (9.7) mamy więc:

$$E_0\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_{\#}) = b^2\frac{N-m-k}{(N-m)k}\mathbf{C}_{11} + (1-b)^2E_0\mathbf{V}_{1/0}(\overline{\mathbf{Y}}_{*}) \tag{9.8}$$

a na podstawie wzorów (9.4) i (9.5) mamy:

$$\mathbf{V}_0(E_{1/0}(\overline{\mathbf{Y}}_{\#})) = \mathbf{V}_0(b\overline{\mathbf{y}}_{U_1} + (1-b)\overline{\mathbf{y}}_{S_0}) = b^2\mathbf{V}_0(\overline{\mathbf{y}}_{U_1}) + 2b(1-b)\mathrm{Cov}_0(\overline{\mathbf{y}}_{U_1}, \overline{\mathbf{y}}_{S_0}) +$$

$$+(1-b)^2\mathbf{V}_0(\overline{\mathbf{y}}_{S_0}) = b^2\mathbf{V}_0\left(\frac{1}{N-m}(N\overline{\mathbf{y}} - m\overline{\mathbf{y}}_{S_0})\right) +$$

$$+2b(1-b)\mathrm{Cov}_0\left(\frac{1}{N-m}(N\overline{\mathbf{y}} - m\overline{\mathbf{y}}_{S_0}), \overline{\mathbf{y}}_{S_0}\right) + (1-b)^2\frac{N-m}{Nm}\mathbf{C}_{11} =$$

$$= \frac{b^2m^2}{(N-m)^2}\mathbf{V}_0(\overline{\mathbf{y}}_{S_0}) - \frac{2b(1-b)m}{N-m}\mathrm{Cov}_0(\overline{\mathbf{y}}_{S_0}, \overline{\mathbf{y}}_{S_0}) + (1-b)^2\frac{N-m}{Nm}\mathbf{C}_{11} =$$

$$= \frac{mb^2}{(N-m)N}\mathbf{C}_{11} - \frac{2b(1-b)m}{N-m}\mathbf{V}_0(\overline{\mathbf{y}}_{S_0}) + (1-b)^2\frac{N-m}{Nm}\mathbf{C}_{11}$$

$$\mathbf{V}_0(E_{1/0}(\overline{\mathbf{Y}}_{\#})) = \left(\frac{mb^2}{(N-m)N} - \frac{2b(1-b)}{N} + (1-b)^2\frac{N-m}{Nm}\right)\mathbf{C}_{11}$$

(9.9)

Podstawiajac otrzymany wynik oraz wzór (9.8) odpowiednio do wyrażenia (9.6) i przeprowadzając redukcję wyrazów otrzymujemy:

$$\mathbf{V}^2(\overline{\mathbf{Y}}_{\#}) = \left(b^2\left(\frac{1}{k} - \frac{2}{N-m} - \frac{1}{N}\right) - \frac{2b(1-b)}{N} + (1-b)^2\frac{N-m}{Nm}\right)\mathbf{C}_{11} + (1-b)^2E_0\mathbf{V}^2_{1/0}(\overline{\mathbf{Y}}_{*}) \tag{9.10}$$

Precyzję estymatora wektorowego ocenia się promieniem estymatora wektorowego. W naszym przypadku określa on średni kwadrat odległości punktu, którego współrzędnymi są wartości składowych wektora $\overline{\mathbf{y}}_{\#}$, od punktu, którego współrzędnymi są elementy wektora $E(\overline{\mathbf{y}}_{\#})$. Parametr ten oznaczamy symbolem $q^2(\overline{\mathbf{y}}_{\#})$.

Drugi sposób polega na wyznaczeniu promienia spektralnego macierzy wariancji i kowariancji wektora estymatorów, który oznaczamy przez $\gamma(\overline{\mathbf{Y}}_{\#})$. Można wykazać, że

$$\gamma(\overline{\mathbf{Y}}_{\#}) = \max_{\boldsymbol{\alpha} \in R^r - \{\mathbf{0}\}} \operatorname{imum} \left\{ \frac{D^2(\overline{\mathbf{Y}}_{\#}\boldsymbol{\alpha}^T)}{\boldsymbol{\alpha}\boldsymbol{\alpha}^T} \right\} \tag{9.11}$$

Stąd wynika, że promień spektralny $\gamma(\overline{\mathbf{Y}}_{\#})$ jest względnym parametrem rozproszenia rozkładu kombinacji liniowej $\overline{Y}_{\#}\boldsymbol{\alpha}^T$, przy najmniej korzystnym wektorze współczynników $\boldsymbol{\alpha}$. Określany jest ilorazem wariancji kombinacji liniowej $\overline{\mathbf{Y}}_{\#}\boldsymbol{\alpha}^T$ przez kwadrat odległości od zera wektora jej współczynników. Dodajmy, że kombinacja $\overline{\mathbf{Y}}_{\#}\boldsymbol{\alpha}^T$ jest estymatorem parametru $\bar{y}_l\boldsymbol{\alpha}^T$, który może być jednym z celów badania statystycznego, chociaż wektor współczynników $\boldsymbol{\alpha}$ nie jest z góry znany. Wyrażenie (9.11) można zapisać także w następującej postaci:

$$\gamma(\overline{\mathbf{Y}}_{\#}) = \max_{\boldsymbol{\alpha} \in R^a - \{\mathbf{0}\}} \operatorname{imum} \left\{ \frac{\boldsymbol{\alpha}\mathbf{V}(\overline{\mathbf{Y}}_{\#})\boldsymbol{\alpha}^T}{\boldsymbol{\alpha}\boldsymbol{\alpha}^T} \right\} \tag{9.12}$$

lub

$$\gamma(\overline{\mathbf{Y}}_{\#}) = \max_{\mathbf{v}\mathbf{v}^T = 1} \operatorname{imum} \left\{ \mathbf{v}\mathbf{V}(\overline{\mathbf{Y}}_{\#})\mathbf{v}^T \right\} \tag{9.13}$$

Stąd wynika, że parametr $\gamma(\overline{\mathbf{Y}}_{\#}) = \rho(\mathbf{V}(\overline{\mathbf{Y}}_{\#}))$, gdzie przez $\rho(\mathbf{V}(\overline{\mathbf{Y}}_{\#}))$ oznaczono maksymalną wartość własną (promień spektralny) macierzy wariancji i kowariancji wektora estymatorów $\overline{\mathbf{Y}}_{\#}$.

9.2. Wykorzystanie wektora estymatorów regresyjnych

Niech staystyka $\overline{\mathbf{Y}}_*$ będzie wektorem estymatorów regresyjnych analizowanych w rozdziale 5:

$$\tilde{\overline{\mathbf{Y}}}_R = \overline{\mathbf{Y}}_{1S_{10}} + (\overline{\mathbf{Y}}_{0S_0} - \overline{\mathbf{Y}}_{0S_{10}})\overline{\mathbf{B}} \tag{9.14}$$

gdzie:

$$\overline{\mathbf{B}} = \tilde{\overline{\mathbf{C}}}_{00}^{-1}\overline{\mathbf{C}}_{10}, \qquad \det \tilde{\overline{\mathbf{C}}}_{00} > 0 \tag{9.15}$$

Składowe macierzy $\overline{\mathbf{C}}_{10} = [\bar{c}_{ij}]$ i $\tilde{\overline{\mathbf{C}}}_{00} = [\tilde{\bar{c}}_{ij}]$ określają wzory:

$$\bar{c}_{ij} = \frac{1}{n-1} \sum_{h \in S_{10}} (y_{0ih} - \bar{y}_{0S_{10}i})(y_{1jh} - \bar{y}_{1S_{10}j}),$$

$$\tilde{c}_{ij} = \frac{1}{n-1}\sum_{h\in S_0}(y_{0ih} - \bar{y}_{0S_0 i})(y_{0jh} - \bar{y}_{0S_0 j}),$$

przy czym średnie $\overline{Y}_{0S_0 i}$, $\overline{Y}_{IS_{10}i}$ oraz $\overline{Y}_{0S_{10}i}$ są elementami odpowiednio wektorów $\overline{\mathbf{Y}}_{0S_0}$, $\overline{\mathbf{Y}}_{IS_{10}}$ oraz $\overline{\mathbf{Y}}_{0S_{10}}$.

Na podstawie własności wektora estymatorów regresyjnych wnioskujemy, że $E(\overline{\mathbf{Y}}_R) = \bar{\mathbf{y}}_1 + 0(n^{-1}) + 0(m^{-1})$, czyli $\tilde{\overline{\mathbf{Y}}}_R$ daje asymptotycznie nieobciążone oceny średniej $\bar{\mathbf{y}}_1$. Z dokładnością do wyrazów $O(n^{-1})$ i $O(m^{-1})$ wartość oczekiwana warunkowej macierzy wariancji i kowariancji ma postać:

$$E_0 D^2_{I/0}(\tilde{\overline{\mathbf{Y}}}_R) = \frac{m-n}{mn}\mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01} \tag{9.16}$$

Podstawiając odpowiednio wyrażenie (9.16) do wzoru (9.10) mamy:

$$\mathbf{V}(\tilde{\overline{\mathbf{Y}}}_{\#}) = \left[b^2\left(\frac{1}{k} - \frac{2}{N-m} - \frac{1}{N}\right) - \frac{2b(1-b)}{N} + (1-b)^2\frac{N-m}{Nm}\right]\mathbf{C}_{11} +$$

$$+(1-b)^2\frac{m-n}{mn}\left(\mathbf{C}_{11} - \mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01}\right) \tag{9.17}$$

Pomijając wyrazy rzędu niższego od $O(k^{-1})$, $O(n^{-1})$ i $O(m^{-1})$ otrzymujemy:

$$\mathbf{V}(\tilde{\overline{\mathbf{Y}}}_{\#}) = \frac{b^2}{k}\mathbf{C}_{11} + \frac{(1-b)^2}{n}\left(\mathbf{C}_{11} - \mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01}\right) + \frac{(1-b)^2}{m}\mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01} \tag{9.18}$$

Kwadrat średniego promienia estymatora $\tilde{\overline{\mathbf{Y}}}_{\#}$ wynosi:

$$q^2(\tilde{\overline{\mathbf{Y}}}_{\#}) = \frac{b^2}{k}g_1 + \frac{(1-b)^2}{n}g_3 + \frac{(1-b)^2}{m}(g_1 - g_3) \tag{9.19}$$

gdzie:

$$g_1 = \mathrm{tr}\mathbf{C}_{11}, \qquad g_2 = \mathrm{tr}\mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01}, \qquad g_3 = g_1 - g_2 \tag{9.20}$$

Dodajmy, że w miejsce estymatora $\overline{\mathbf{Y}}_*$ można również wstawić wektory estymatorów ilorazowych lub iloczynowych itp.

9.3. Minimalizacja warunkowa średniego promienia wektora estymatorów

Załóżmy, że ilość obserwacji badanych zmiennych jest taka sama w każdym z dwóch okresów czasu. Liczbę tych obserwacji oznaczamy przez m. W okresie wyjściowym liczebność próby S_0 wynosi więc m. Suma liczebności prób S_1 i S_{01} losowanych w okresie bieżącym wynosi także m=k+n. Załóżmy, że k=wm i n=(1-w)m, gdzie 0<w<1.

Rozważmy dwa zadania optymalizacji liczebności k i n prób S_1 i S_{10} przy ustalonym rozmiarze m próby pierwotnej S_0. Sprowadza się to do optymalizacji parametrów b i w, od których zależy macierz wariancji i kowariancji wektora estymatorów $\tilde{\mathbf{Y}}_{\#}$. Pierwsze z nich polega na warunkowej minimalizacji średniego promienia wektora $\tilde{\mathbf{Y}}_{\#}$ Zastępując we wzorze (9.19) argumenty k i n odpowiednio przez wm i (1-w)m zadanie optymalizacyjne zapisujemy w następującej postaci:

$$\begin{cases} f(b,w) = \text{minimum} \\ 0 \le w \le 1;\ b \in R \end{cases} \tag{9.21}$$

gdzie:

$$f(b,w) = \frac{1}{m}\left(\frac{b^2}{w} g_1 + \frac{(1-b)^2}{1-w} g_3 + (1-b)^2 (g_1 - g_3) \right) \tag{9.22}$$

Lemat 9.1: Funkcja f(b,w) jest ściśle wypukła dla 0<w<1 i $b \in R$.
Dowód: Obliczamy pochodne funkcji f:

$$\begin{cases} \dfrac{\partial f}{\partial b} = \dfrac{2b g_1}{mw} - \dfrac{2(1-b)}{m}\left(\dfrac{g_3}{1-w} + g_1 - g_3 \right) = 0 \\ \dfrac{\partial f}{\partial w} = -\dfrac{b^2 g_1}{m w^2} + \dfrac{(1-b)^2 g_3}{m(1-w)^2} = 0 \end{cases} \tag{9.23}$$

$$\begin{cases} \dfrac{\partial^2 f}{\partial b^2} = \dfrac{2g_1}{mw} + \dfrac{2}{m}\left(\dfrac{g_3}{1-w} + g_1 - g_3 \right) > 0 \\ \dfrac{\partial^2 f}{\partial w^2} = \dfrac{2b^2 g_1}{mw^3} + \dfrac{2(1-b)^2 g_3}{m(1-w)^3} > 0 \\ \dfrac{\partial^2 f}{\partial w \partial b} = -\dfrac{2b g_1}{mw^2} + \dfrac{2(1-b) g_3}{m(1-w)^2} > 0 \end{cases} \tag{9.24}$$

Forma kwadratowa, której macierzą jest hesjan funkcji f, ma postać:

$$Q(u_1,u_2) = \frac{2}{m}\left[\frac{g_1}{w}\left(\frac{b}{w}u_1-u_2\right)^2+\frac{g_3}{1-w}\left(\frac{1-b}{1-w}u_1-u_2\right)^2+(g_1-g_3)u_2^2\right] \tag{9.25}$$

Stąd wynika, że dla każdych u_1, $u_2 \in R-\{0\}$ oraz $w\in(0;1)$ i $b\in R$ forma kwadratowa $Q(u_1,u_2)\geq 0$. Zatem hesjan funkcji f jest nieujemnie określony. Oznacza to, że funkcja f jest wypukła dla $0<w<1$ i $b\in R$, cbdu.

Rozwiązując równanie $\frac{\partial f}{\partial w}=0$ względem w otrzymujemy:

$$w_1=\frac{1}{1-\left|\frac{1-b}{b}\right|\sqrt{e}} \tag{9.26}$$

$$w_2=\frac{1}{1+\left|\frac{1-b}{b}\right|\sqrt{e}} \tag{9.27}$$

gdzie:

$$e=\frac{g_3}{g_1} \tag{9.28}$$

jest współczynnikiem względnej efektywności estymatora

$$\overline{\mathbf{Y}}_R = \overline{\mathbf{Y}}_{S_0} + (\overline{\mathbf{x}} - \overline{\mathbf{X}}_{S_0})\mathbf{C}_{00}^{-1}\mathbf{C}_{01} \tag{9.29}$$

w stosunku do średniej $\overline{\mathbf{Y}}_{S_0}$. Rozwiązanie w_2 jest tylko dopuszczalne, bo $0<w_2<1$. Podstawiając je do równania $\frac{\partial f}{\partial b}=0$ po odpowiednim przekształceniu otrzymujemy:

$$b\left[2+\left(\left|\frac{1-b}{b}\right|+\left|\frac{1}{1-b}\right|\right)\sqrt{e}\right]-\left|\frac{1}{1-b}\right|-1=0$$

Przyjmując, że $0<b<1$ wyliczamy, że pierwiastkiem powyższego równania jest $b_0=\frac{1}{2}$. Stąd i z równania (9.27) wynika, że optymalnymi rozwiązaniami zadania jest para liczb (w_o,b_o), gdzie:

$$w_o=\frac{1}{1+\sqrt{e}}, \qquad b_o=\frac{1}{2} \tag{9.30}$$

Stąd, że 0<e<1 wynika, iż $\frac{1}{2}$<w_o<1 i w_o jest malejącą funkcją współczynnika e. Średni promień spektralny wektora estymatorów $\tilde{\mathbf{Y}}_{\#}$ dla (w_o,b_o) wynosi:

$$f(b_o,w_o) = \frac{g_1}{4m}\left(1+\sqrt{e}\right) \tag{9.31}$$

Precyzja oceny wektora średnich $\bar{y}_1$ za pomocą wektora estymatorów $\tilde{\mathbf{Y}}_{\#}$ rośnie wraz ze spadkiem współczynnika e. Zatem ta precyzja rośnie wraz ze wzrostem siły skorelowania między badanymi wektorami cech w obu okresach czasu.

9.4. Minimalizacja warunkowa promienia spektralnego wektora estymatorów

Drugie zadanie optymalizacyjne polega na minimalizacji promienia spektralnego estymatora $\tilde{\mathbf{Y}}_{\#}$, przy warunku, że k+n=m i ustalonym parametrze b=$\frac{1}{2}$. Prosty sposób rozwiązania postawionego zadania polega na wyszukaniu takiej kombinacji liczb k_o i n_o, że k_o+n_o=m i $\gamma(\mathbf{Y}_{\#})$=minimum.

W celu wyznaczenia rozwiązania analitycznego tego zadania wprowadzamy taką znienną w, że k=mw i n=(1-w)m. Niech z(w)= $\gamma(\tilde{\mathbf{Y}}_{\#})$ i Σ(w)= $\mathbf{V}(\tilde{\mathbf{Y}}_{\#})$ dla b=$\frac{1}{2}$. Wtedy zadanie optymalizacyjne określa wyrażenie:

$$\begin{cases} z(w) = \min imum \\ 0 < w < 1 \end{cases} \tag{9.32}$$

Na podstawie wzorów (9.13) i (9.18) mamy:

$$z(w) = \max_{\mathbf{v}\mathbf{v}^T=1} imum\{l(w,\mathbf{v})\} \tag{9.33}$$

gdzie:

$$1(w,\mathbf{v})=\mathbf{v}\boldsymbol{\Sigma}(w)\mathbf{v}^T \tag{9.34}$$

$$\boldsymbol{\Sigma}(w)=\frac{1}{4m}\left[\left(1+\frac{1}{w}\right)\mathbf{C}_{11}-\left(1-\frac{1}{1-w}\right)\left(\mathbf{C}_{11}-\mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01}\right)\right] \tag{9.35}$$

$$l(w,v)=\frac{1}{4m}\left[\left(1+\frac{1}{w}\right)q_1-\left(1-\frac{1}{1-w}\right)q_3\right] \tag{9.36}$$

$$q_1 = \mathbf{v}\mathbf{C}_{11}\mathbf{v}^T, \qquad q_2 = \mathbf{v}\mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{10}\mathbf{v}^T, \qquad q_3 = q_1 - q_2 \tag{9.37}$$

Teraz optymalizacyjne zagadnienie (9.32) można zastąpić zadaniem poszukiwania takiej pary $(w_0, \mathbf{v}_0)$, że

$$l(w_0, \mathbf{v}_0) = \min_{0<w<1} \operatorname{imum} \max_{\mathbf{v}\mathbf{v}^T=1} \operatorname{imum}\{l(w, \mathbf{v})\} \tag{9.38}$$

Będzie wygodniej analizować ten problem za pomocą funkcji Lagrange'a:

$$L(w,\mathbf{v}) = l(w,\mathbf{v}) - \lambda(\mathbf{v}\mathbf{v}^T - 1) \tag{9.39}$$

Wtedy zadanie polega na wyznaczeniu takiej trójki $(w_o\mathbf{v},\lambda)$, że

$$L(w_o,\mathbf{v}_o\lambda_1) = \underset{0<w<1}{\text{minimum}} \ \underset{\mathbf{v}\in R^a}{\text{maximum}} \{L(w,\mathbf{v},\lambda)\} \tag{9.40}$$

Pochodne funkcji l i L mają postać:

$$\frac{\partial l}{\partial w} = \frac{1}{4m}\left(-\frac{q_1}{w^2} + \frac{q_3}{(1-w)^2}\right) \tag{9.41}$$

$$\frac{\partial l}{\partial \mathbf{v}} = \mathbf{\Sigma}(w)\mathbf{v}^T \tag{9.42}$$

$$\frac{\partial^2 l}{\partial w^2} = \frac{1}{2m}\left(\frac{q_1}{w^3} + \frac{q_3}{(1-w)^3}\right) > 0 \tag{9.43}$$

$$\frac{\partial^2 l}{\partial \mathbf{v}^2} = \mathbf{\Sigma}(w) \tag{9.44}$$

$$\frac{\partial^2 l}{\partial \mathbf{v}\partial w} = \frac{1}{4m}\left(-\frac{1}{w^2}\mathbf{C}_{11} + \frac{1}{(1-w)^2}(\mathbf{C}_{11} - \mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01})\right)\mathbf{v}^T \tag{9.45}$$

$$\frac{\partial L}{\partial w} = \frac{\partial l}{\partial w}, \qquad \frac{\partial L}{\partial \mathbf{v}} = \frac{\partial l}{\partial \mathbf{v}} - \lambda\mathbf{v}$$

$$\frac{\partial^2 L}{\partial w^2} = \frac{\partial^2 l}{\partial w^2}, \qquad \frac{\partial^2 L}{\partial \mathbf{v}\partial w} = \frac{\partial^2 l}{\partial \mathbf{v}\partial w} \tag{9.46}$$

$$\frac{\partial^2 L}{\partial \mathbf{v}^2} = \mathbf{\Sigma}(w) - \lambda\mathbf{I}_a \tag{9.47}$$

gdzie $\mathbf{I}_a$ jest macierzą jednostkową stopnia a. Ze znanych własności macierzy wiadomo, że jeśli przynajmniej jedna z dwu symetrycznych macierzy rzeczywistych jest nieosobliwa, to istnieje taka nieosobliwa macierz **P**, że

$$\mathbf{P}\mathbf{P}^{\mathrm{T}}=\mathbf{C}_{11}, \qquad \mathbf{P}\mathbf{R}\mathbf{P}^{\mathrm{T}}=\mathbf{C}_{10}\mathbf{C}_{00}^{-1}\mathbf{C}_{01} \tag{9.48}$$

gdzie $\mathbf{R} = \left[r_i^2\right]$ jest macierzą diagonalną, której elementami są kwadraty współczynników korelacji kanonicznej. Wtedy hesjan określony wzorem (9.47) przyjmuje postać:

$$\frac{\partial^2 L}{\partial \mathbf{v}^2}=\frac{1}{4m}\mathbf{P}\left[\left(1+\frac{1}{w}\right)\mathbf{I}_a-\left(1-\frac{1}{1-w}\right)\left(\mathbf{I}_a-\mathbf{R}\right)\right]\mathbf{P}^{\mathrm{T}}-\lambda\mathbf{I}_a \tag{9.49}$$

Najpierw rozpatrzymy szczególny przypadek zakładając dodatkowo, że macierze $\mathbf{C}_{00}=\mathbf{C}_{11}=\mathbf{D}=[d_i]$ i $\mathbf{C}_{01}$ są diagonalne. Wtedy $\mathbf{P}=\mathbf{D}^{1/2}$, a określony wzorami (9.47) lub (9.49) hesjan funkcji L przyjmuje postać:

$$\frac{\partial^2 L}{\partial \mathbf{v}^2}=\frac{1}{4m}\mathbf{D}\left[\left(1+\frac{1}{w}\right)\mathbf{I}_a-\left(1-\frac{1}{1-w}\right)\left(\mathbf{I}_a-\mathbf{R}\right)\right]-\lambda\mathbf{I}_a$$

lub

$$\frac{\partial^2 L}{\partial \mathbf{v}^2}=\frac{1}{4m}\mathbf{D}(\mathbf{I}_a-w^2\mathbf{R})-\lambda\mathbf{I}_a \tag{9.50}$$

Otrzymany hesjan jest macierzą diagonalną. Wtedy jego maksymalną wartość własną wyznaczamy na podstawie równania charakterystycznego:

$$\prod_{i=1}^{a}\left(\frac{d_i}{4mw(1-w)}\left(1-w^2r_i^2\right)-\lambda\right)=0 \tag{9.51}$$

Wartości własne hesjanu są więc wyznaczane na podstawie funkcji:

$$\lambda_i(w)=\frac{d_i(1-w^2r_i^2)}{4mw(1-w)}, \quad i=1,...,a \tag{9.52}$$

Ich pochodne są postaci:

$$\frac{\partial\lambda_i}{\partial w}=-\frac{d_i r_i^2(w-w_{1i})(w-w_{2i})}{4mw^2(1-w)^2} \tag{9.53}$$

gdzie:

$$w_{1i}=\frac{1+\sqrt{1-r_i^2}}{r_i^2}>1 \tag{9.54}$$

$$w_{2i} = \frac{1-\sqrt{1-r_i^2}}{r_i^2} \tag{9.55}$$

Dopuszczalnym rozwiązaniem jest pierwiastek w_{2i}, bo $0<w_{2i}< 1$.Druga pochodna funkcji λ_i jest równa zeru dla $w=w_{2i}$. Jednak ze wzoru (9.52) wynika, że wraz ze wzrostem argumentu funkcji $\lambda_i(w)$ w przedziale (0;1) jej pierwsza pochodna zmienia znak z ujemnego na dodatni w punkcie $w=w_{2i}$. Stąd wynika, że λ_i jest wypukła w przedziale (0;1) i $\lambda_i(w_{2i})$=minimum, które wynosi:

$$\lambda_i(w_{2i}) = \frac{d_i r_i^2}{2m\left(1-\sqrt{1-r_i^2}\right)} \tag{9.56}$$

Z ciągłości i wypukłości funkcji $\lambda_i(w)$ w przedziale (0;1) wynika, że

$$z(w) = \max_{i=1,\ldots,a} imum\left\{\lambda_i(w)\right\} \tag{9.57}$$

gdzie przypomnijny, że z(w) jest promieniem spektralnym macierzy $\mathbf{\Sigma}(w)$. Możemy wyróżnić przynajmniej trzy przypadki. Pierwszy, najprostszy, gdy funkcje $\lambda_i(w)$ pokrywają się dla wszystkich i=1,...,a. Zachodzi ta sytuacja wtedy, gdy $r_i^2 = r^2$ oraz $d=d_i$ dla każdego i=1,...,a. Wtedy:

$$w_o = \frac{1-\sqrt{1-r^2}}{r^2}, \qquad z(w_o) = \frac{dr^2}{2m\left(1-\sqrt{1-r^2}\right)} \tag{9.58}$$

Drugi przypadek jest związany z sytuacją, gdy funkcje $\lambda_i(w)$, i=1,...,a nie mają punktów wspólnych w przedziale (0;1). Wtedy:

$$w_o = w_{2t}: \ \lambda_t(w_{2t}) = \max_{i=1,\ldots,a} imum\left\{\lambda_i(w_{2i})\right\} \tag{9.59}$$

Trzeci przypadek dotyczy sytuacji, gdy dla każdego $i\neq j=1,...,a$ funkcje $\lambda_i(w)$, $\lambda_j(w)$ mają jeden punkt wspólny w przedziale (0;1), który oznaczamy przez w_{ij}. Punkt ten jest rzeczywistym pierwiastkiem równania $\lambda_i(w) - \lambda_j(w) = 0$, które jest równoważne następującym:

$$d_i(1 - wr_i^2) = d_j(1 - w^2 r_j^2)$$

$$w^2 = \frac{d_i - d_j}{d_i r_i^2 - d_j r_j^2} = x_{ij}$$

Gdy $x_{ij}<0$, to $\lambda_i(w)>\lambda_j(w)$ bądź $\lambda_i(w)<\lambda_j(w)$ dla każdego $w\in(0;1)$. W przeciwnym przypadku szukany punkt wynosi:

$$w_{ij} = \sqrt{x_{ij}}, \qquad i \neq j \tag{9.60}$$

Wtedy maksymalna wartość własna macierzy $\Sigma(w)$ jest wyrażana w dwóch podprzedziałach przedziału (0;1) za pomocą wzoru:

$$z(w)=\begin{cases}\lambda_v(w)=\max\text{imum}\{\lambda_i(w)\} & \text{dla } 0<w<w_{vu}\\ \lambda_u(w)=\max\text{imum}\{\lambda_i(w)\} & \text{dla } w_{vu}\le w<1\end{cases} \tag{9.61}$$

Reasumując, w zależności od wartości parametrów r_i, d_i (i=1,...,a) promień spektralny z(w) może być określony wzorem (9.59) lub (9.61), a szukane rozwiązanie optymalne zadania (9.32) wyrażeniem:

$$w_o=\begin{cases}w_{2t}, & \text{gdy } z(w) \text{ określa wzór (57)}\\ w_{vu}, & \text{gdy } z(w) \text{ określa wzór (61)}\end{cases} \tag{9.62}$$

Rozważmy kilka szczególnych przypadków otrzymanego rozwiązania. Niech $r_i^2 = r^2$ dla każdego i=1,...,a. Wtedy frakcję optymalną w_o i wartość funkcji celu z(w) określa wzór:

$$w_o=\frac{1-\sqrt{1-r^2}}{r^2}, \qquad z(w_o)=\frac{d_o r^2}{2m\left(1-\sqrt{1-r^2}\right)} \tag{9.63}$$

gdzie:

$$d_o=\max\text{imum}\{d_i\}$$

Drugi przypadek dotyczy sytuacji, gdy $r_i^2>0$ i d_i=d dla każdego i=1,...,a. Wówczas ze wzoru (9.52) wynika, że dla każdego $w\in(0;1)$ zachodzi nierówność $\lambda_i(w)>\lambda_j(w)$ wtedy i tylko wtedy, gdy $r_i^2<r_j^2$. Stąd i ze wzorów (9.52), (9.53) i (9.55) wnioskujemy więc, że

$$w_o=\frac{1-\sqrt{1-r_o^2}}{r_o^2}, \qquad \text{gdzie:} \quad r_o^2=\min\text{imum}\left\{r_i^2\right\} \tag{9.64}$$

$$z(w_o)=\frac{d\ r_o^2}{2m\left(1-\sqrt{1-r_0^2}\right)} \tag{9.65}$$

Zakładając, że istnieje wskaźnik j (j=1,...,a), iż r_j =0 oraz d_i =d i dla każdego i=1,...,a, ze wzoru (9.52) wnioskujemy, że:

$$w_o = \frac{1}{2}, \qquad z\left(\frac{1}{2}\right) = \frac{d}{m} \tag{9.66}$$

Wróćmy do przypadku ogólnego, gdy macierz $\mathbf{\Sigma}(w)$ nie jest diagonalną. Wtedy optymalną frakcję w_o i odpowiadającą jej maksymalną wartość własną macierzy $\mathbf{\Sigma}(w)$ wyznaczamy na podstawie równania charakterystycznego:

$$\psi(\lambda,w) = \det(\mathbf{\Sigma}(w) - \lambda\mathbf{I}_a) = 0 \tag{9.67}$$

Równanie $\psi(\lambda,w)=0$ określa funkcję uwikłaną wartości własnej $\lambda(w)$ względem frakcji w. Zadanie polega na znalezieniu takiej wartości w, dla której $\lambda(w_o)$ osiąga maksimum absolutne. Na podstawie wzorów (9.47) i (9.49) równanie $\psi(\lambda,w)=0$ można zapisać następująco:

$$\psi(\lambda,w) = \det\left\{\frac{1}{4m}\left[\left(1+\frac{1}{w}\right)\mathbf{I}_a - \left(1-\frac{1}{1-w}\right)(\mathbf{I}_a - \mathbf{R})\right] - \lambda\mathbf{P}^{-1}\left(\mathbf{P}^T\right)^{-1}\right\} = 0$$

Niech $\mathbf{P}^{-1}(\mathbf{P}^T)^{-1}=\mathbf{H}$ i macierz diagonalna $\mathbf{U}(w)=[u_i(w)]$, gdzie:

$$u_i(w) = \frac{1-w^2 r_i^2}{4mw(1-w)}, \qquad i=1,...,a \tag{9.68}$$

Wtedy:

$$\psi(\lambda,w)=\det(\mathbf{U}(w)-\lambda\mathbf{H})=0 \tag{9.69}$$

Korzystając z dekompozycji wyznacznika sumy macierzy [por. Wywiał (1986)]mamy:

$$\psi(\lambda, w) = (-\lambda)^a \det\mathbf{H} + \sum_{v=1}^{a}(-\lambda)^{a-v} \sum_{\{i_1,...,i_v\}} \det\overline{\mathbf{H}}(i_1,\ldots,i_v)\prod_{j=1}^{v} u_{i_j}(w) \tag{9.70}$$

gdzie $i_1,..,i_v$ jest kombinacją numerów kolumn (wierszy) macierzy $\mathbf{H}$. Przez $\overline{\mathbf{H}}(i_1,..,i_v)$ oznaczono macierz kwadratową stopnia a-v otrzymaną z macierzy $\mathbf{H}$ po skreśleniu z niej kolumn i wierszy o numerach $i_1,..,i_v$. Zakłada się, że jeżeli v=a, to $\det\overline{\mathbf{H}}(1,..,a)=1$.

Wiadomo, że pochodną funkcji $\lambda(w)$ określa wzór:

$$\frac{\partial\lambda}{\partial w} = \frac{\partial\psi}{\partial w} : \frac{\partial\psi}{\partial\lambda} \tag{9.71}$$

Na podstawie wzoru (9.70) wyznaczamy pochodne cząstkowe:

$$\frac{\partial\psi}{\partial w} = \sum_{v=1}^{a}(-\lambda)^{a-v} \sum_{\{i_1,...,i_v\}} \det\overline{\mathbf{H}}(i_1,..,i_v)\sum_{j=1}^{a}\frac{\partial u_{i_j}}{\partial w}\prod_{t=1}^{v} u_{i_t}(w) \tag{9.72}$$

gdzie:

$$\frac{\partial u_i}{\partial w} = -\frac{r_i^2 (w - w_{1i})(w - w_{2i})}{4mw^2 (1-w)^2} \tag{9.73}$$

Pierwiastki w_{1i} i w_{2i} określają wzory (9.54) i (9.55).

$$\frac{\partial \psi}{\partial \lambda} = (-1)^a a\lambda^{a-1} \det \mathbf{H} + \sum_{v=1}^{a-1} (-1)^{a-v} (a-v-1)\lambda^{a-v-1} \sum_{\{i_1,..,i_v\}} \det \overline{\mathbf{H}}(i_1,..,i_v) \prod_{j=1}^{v} u_{i_j}(w) \tag{9.74}$$

Punkty osobliwe wyliczamy jako rozwiązania układu równań:

$$\begin{cases} \psi(\lambda, w) = 0 \\ \dfrac{\partial \psi}{\partial w} = 0 \end{cases} \tag{9.75}$$

Spośród pierwiastków $\{(\lambda_i, w_i)$ tego układu należy jako rozwiązanie zadania (9.32) wybrać tę parę (λ_o, w_o), dla której $\lambda_o = \max\limits_{i=1,...,a} \{\lambda_i\}$ oraz określona wzorami (9.71)-(9.74) pochodna zmienia znak z ujemnego na dodatni w otoczeniu punktu w_o.

Układ (9.75) składa się z dwóch równań nieliniowych. Możliwe jest jedynie przybliżone wyliczenie jego pierwiastków za pomocą odpowiedniej metody numerycznej rozwiązywania takich układów równań.

9.5. Predykcja wartości średniej w nadpopulacji regresyjnej

Zakładamy, że w okresie bieżącym rozkład cechy badanej jest opisany modelem regresyjnym nadpopulacji. Zmienną pomocniczą występującą jako niezależna w równaniu regresji jest cecha badana w okresie podstawowym. Szczegółowo prosty model nadpopulacji regresyjnej określamy następująco: Niech $\mathbf{Y}_1 = [Y_{11} ... Y_{1N}]$ będzie wektorem zmiennych losowych przyporządkowanych kolejnym elementom populacji. Z kolei $\mathbf{y}_0 = [y_{01} ... y_{0N}]$ jest wektorem ustalonych wartości cechy obserwowanej w okresie podstawowym. Przyjmujemy, że

$$Y_{1i} = a y_{0i} + U_i \tag{9.76}$$

$$\begin{cases} \mathcal{E}(U_i) = 0 \quad \text{i} \quad \mathcal{D}^2(U_i) = \sigma^2 \quad \text{dla} \quad i = 1,...,N \\ \mathcal{E}(U_i U_h) = 0 \quad \text{dla} \quad i \neq h = 1,...,N \end{cases} \tag{9.77}$$

Problem polega na predykcji wartości średniej $\overline{Y}_1 = \frac{1}{N}\sum_{i=1}^{N} Y_{1i}$ na podstawie danych obserwowanych w bezzwrotnie losowanych próbach prostych S_0, S_{10} i S_1 o liczebnościach odpowiednio m, n, k. Zbiory S_0 i S_{10} tworzą próbę podwójną, gdzie S_{10} jest losowana spośród elementów próby S_0, natomiast S_0 i S_1 są losowane z całej populacji. Niech

$$\overline{Y}_{1S_1} = \frac{1}{k}\sum_{i \in S_1} Y_{1i} \tag{9.78}$$

Niech statystyka $T_{\otimes}$ będzie funkcją obserwacji cechy badanej w okresie bieżącym w próbie S_{10} oraz cechy badanej w okresie wyjściowym w próbach S_{10} i S_0. Zakładamy, że $T_{\otimes}$ daje p-ξ nieobciążone oceny średniej $\overline{Y}_1$.

Do predykcji średniej $\overline{Y}_1$ użyjemy statystyki:

$$T = \alpha \overline{Y}_{1S_1} + (1-\alpha) T_{\otimes} \tag{9.79}$$

Statystyka T daje p-ξ nieobciążone oceny średniej $\overline{Y}_1$. Próba S_0 jest bezzwrotnie losowana z całej populacji. Podobnie, także z całej populacji jest losowana bezzwrotnie próba S_1. Zatem są one niezależne i również są niezależne rozkłady statystyk $\overline{Y}_{1S_1}$ i $T_{\otimes}$. Zatem błąd średniokwadratowy predykcji można zdekomponować w następujący sposób:

$$\mathcal{E}E(T-\overline{Y}_1)^2 = \alpha^2 \mathcal{E}E(\overline{Y}_{1S_1} - \overline{Y}_1)^2 + (1-\alpha)^2 \mathcal{E}E(T_{\otimes} - \overline{Y}_1)^2$$

Wtedy:

$$\mathcal{E}E(\overline{Y}_{1S_1} - \overline{Y}_1)^2 = \frac{N-k}{Nk}\mathcal{E}\left(v_*(\mathbf{Y}_1)\right) = \frac{N-k}{(N-1)Nk}\sum_{i=1}^{N}\mathcal{E}\left[a(y_{0i} - \bar{y}_0) - U_i + \overline{U}\right]^2 =$$

$$= \frac{N-k}{Nk}\left[a^2 v_*(y_0) + \frac{1}{N-1}\sum_{i=1}^{N}\mathcal{E}(U_i - \overline{U})^2\right] = \frac{N-k}{Nk}\left[a^2 v_*(\mathbf{y}_0) + \sigma^2\right]$$

Stąd już mamy:

$$\mathcal{E}E(T-\overline{Y}_1)^2 = \alpha^2 \frac{N-k}{Nk}\left[a^2 v_*(y_0) + \sigma^2\right] + (1-\alpha)^2 \mathcal{E}E(T_{\otimes} - \overline{Y}_1)^2 \tag{9.80}$$

9.5.1. Predyktor ilorazowy

Wywiał (1992) s. 211-213 analizuje własności predyktora ilorazowego o postaci:

$$\tilde{\bar{Y}}_I = \frac{\bar{Y}_{1S_{10}}}{\bar{y}_{0S_{10}}} \bar{y}_{0S_0} \tag{9.81}$$

Predyktor ten przy założonym modelu nadpopulacji daje p-ξ nieobciążone oceny średniej $\bar{Y}_1$. Z kolei jego błąd średniokwadratowy predykcji ma postać:

$$\mathbf{E}E(\tilde{\bar{Y}}_I - \bar{Y}_1)^2 = E\left\{a^2(\bar{y}_{0S_0} - \bar{y}_0)^2 + \left(\frac{(N-m)n}{Nm} + \frac{m-n}{m}\frac{\bar{y}_{0,\bar{s}_{10}}}{\bar{y}_{0,s_{10}}}\right)\frac{\sigma^2}{n} + \frac{N-n}{N^2}\sigma^2\right\}$$

lub

$$\mathbf{E}E(\tilde{\bar{Y}}_I - \bar{Y}_1)^2 = a^2\frac{N-m}{Nm}v_*(y_0) + \left(\frac{(N-m)n}{Nm} + \frac{m-n}{m}E\left(\frac{\bar{y}_{0,\bar{s}_{10}}}{\bar{y}_{0,s_{10}}}\right)\right)\frac{\sigma^2}{n} + \frac{N-n}{N^2}\sigma^2 \tag{9.82}$$

gdzie: $\bar{s}_{10} = s_0 - s_{10}$. Błąd ten przyjmuje wartość najmniejszą, gdy próba s_{01} jest dobierana celowo tak, aby otrzymać zbiór elementów s_{10}^* spełniający równanie:

$$\max_{s_{10}\in\mathcal{S}(s_{10})}\left\{\sum_{h\in s_{10}} y_{0h}\right\} = \sum_{h\in s_{10}^*} y_{0h} \tag{9.83}$$

Predyktor $\tilde{\bar{Y}}_I$ można podstawić we wzorze (9.79) w miejsce statystyki $T_\otimes$. Wtedy mamy:

$$T_I = \alpha\bar{Y}_{1S_1} + (1-\alpha)\tilde{\bar{Y}}_{I_I} \tag{9.84}$$

Otrzymany predyktor jest p-ξ nieobciążony, a jego błąd średniokwadratowy wynika ze wzorów (9.80) i (9.82):

$$\boldsymbol{\mathcal{E}}E(T_I - \bar{Y}_1)^2 = \alpha^2\frac{N-k}{Nk}\left(a^2 v_*(y_0) + \sigma^2\right) +$$

$$+(1-\alpha)^2\left\{a^2\frac{N-m}{Nm}v_*(y_0) + \left(\frac{(N-m)n}{Nm} + \frac{m-n}{m}E\left(\frac{\bar{y}_{0,\bar{s}_{10}}}{\bar{y}_{0,s_{10}}}\right)\right)\frac{\sigma^2}{n} + \frac{N-n}{N^2}\sigma^2\right\} \tag{9.85}$$

Stąd i ze wzoru (9.83) wynika, że błąd średniokwadratowy predyktora T_I będzie najmniejszy, jeśli próba s_{10} będzie dobierana celowo tak, aby otrzymać zbiór s_{10}^* spełniajacy równanie (9.83).

9.5.2. Predyktor regresyjny

Wywiał (1992), s. 213-214 analizuje predyktor regresyjny z próby podwójnej (S_0,S_{10}), postaci:

$$\tilde{\bar{Y}}_B = \bar{Y}_{S_{10}} - B_{S_{10}}(\bar{Y}_{S_{10}} - \bar{Y}_0) \tag{9.86}$$

gdzie:

$$B_{S_{10}} = \frac{\sum_{i \in S_{10}} \left(y_{0i} - \bar{y}_{0S_{10}}\right)\left(Y_{1i} - \bar{Y}_{1S_{10}}\right)}{\sum_{i \in S_{10}} \left(y_{0i} - \bar{y}_{0S_{10}}\right)^2}$$

Predyktor ten jest p-ξ nieobciążony, a jego błąd średniokwadratowy ma postać:

$$\mathbf{E}E(\tilde{\bar{Y}}_B - \bar{Y}_1)^2 = a^2 \frac{N-m}{Nm} v_*(y_0) + \frac{N-n}{Nn}\sigma^2 + \sigma^2 E(g(S_0, S_{10})) \tag{9.87}$$

gdzie:

$$g(S_0,S_{10}) = \frac{\left(\bar{y}_{0S_0} - \bar{y}_{0S_{10}}\right)^2}{\sum_{i \in S_{10}} \left(y_{0i} - \bar{y}_{0S_{10}}\right)^2} \tag{9.88}$$

Błąd średniokwadratowy predyktora $\tilde{\bar{Y}}_B$ osiąga minimum, gdy próba S_{10} jest dobierana celowo tak, aby otrzymać zbiór S_{10}^{**} spełniający wyrażenie:

$$\underset{s_{10} \in \mathcal{S}(s_{10})}{\text{minimum}} \left\{g(S_0,S_{10})\right\} = g(S_0,S_{10}^{**}) \tag{9.89}$$

Podstawiając we wzorze (9.79) w miejsce statystyki $T_\otimes$ predyktor $\tilde{\bar{Y}}_B$ otrzymujemy:

$$T_B = \alpha \bar{Y}_{1S_1} + (1-\alpha)\tilde{\bar{Y}}_B \tag{9.90}$$

Predyktor ten jest p-ξ nieobciążony, a jego błąd średniokwadratowy wyznaczany na podstawie wzorów (9.80) i (9.87) ma postać:

$$\mathcal{E}E(T_B - \bar{Y}_1)^2 = \alpha^2 \frac{N-k}{Nk}\left[a^2 v_*(\mathbf{y}_0) + \sigma^2\right] +$$

$$+(1-\alpha)^2\left(a^2 \frac{N-m}{Nm} v_*(\mathbf{y}_0) + \frac{N-n}{Nn}\sigma^2 + \sigma^2 E(g(S_0, S_{10}))\right) \tag{9.91}$$

Zatem błąd średniokwadratowy predyktora T_B jest minimalny, gdy próba S_{10} jest dobierana celowo tak, aby otrzymać zbiór S_{10}^{**} spełniający wyrażenie (9.89).

9.5.3. Średnia z próby powarstwowanej

W punkcie 6.8.3. wprowadzono predyktor będący średnią z próby warstwowej. W rozważanym teraz przypadku zakładamy, że próba prosta S_0 losowana w podstawowym okresie jest grupowana na H równolicznych warstw, które oznaczamy przez S_{0h}, h=1,...,H. Ilość warstw H należy tak ustalić, aby liczebność każdej z nich była liczbą całkowitą $n_0 = \frac{m}{H}$. Warstwy oznaczamy jako zbiory S_{0h}, h=1,...,H i są one wydzielane w sposób opisany wcześniej w paragrafie 6.8.3. za pomocą grupowania na podstawie cechy y_0 obserwowanej w wyjściowym okresie w całej próbie S_0. Z każdej warstwy S_{0h} jest losowana bezzwrotnie próbka prosta S_{1h} o liczebności n_0. Suma elementów populacji w próbie warstwowej wynosi więc $n=Hn_0$.

Średnią z próby warstwowej S_{1h} oznaczamy następująco:

$$\overline{Y}_{S_{1h}} = \frac{1}{n_0} \sum_{i \in S_{1h}} Y_{1i} \tag{9.92}$$

Przyjmując, że warstwy są równoliczne, średnią z próby warstwowej określamy wzorem:

$$\tilde{\overline{T}}_w = \frac{1}{H} \sum_{h=1}^{H} \overline{Y}_{S_{1h}} \tag{9.93}$$

Wtedy na podstawie wzoru (6.89) otrzymujemy błąd średniokwadratowy predyktora $\tilde{\overline{T}}_w$ średniej $\overline{Y}_1$ w postaci:

$$\mathbf{E}E(\tilde{\overline{T}}_w - \overline{Y}_1)^2 = \frac{1}{\binom{N}{m}} \frac{1}{H} \frac{m-n}{mn} \sum_{s_0 \in S_0} \sum_{h=1}^{H} \mathbf{E}\left(v_{*S_{0h}}(Y_1 | s_0) \right) + \frac{N-m}{Nm} v_*(Y_1) \tag{9.94}$$

Podobnie jak wzóry (6.85) i (6.87) otrzymujemy:

$$E\left(\tilde{T}_w - \bar{Y}_1\right)^2 = \frac{1}{\binom{N}{n}} \frac{a^2}{H} \frac{n-m}{mn} \sum_{s_0 \in S} \sum_{h=1}^{H} v_*(y_0|s_0) +$$

$$+a^2 \frac{N-m}{Nm} v_*(y_0) - \frac{N-n}{Nn}\sigma^2 \tag{9.95}$$

Przez $\mathcal{S}_0$ oznaczono przestrzeń prób typu s_0. Z kolei symbolem $v_{*S_{0h}}(y_0|s_0)$ oznaczono wariancję teraz już cechy pomocniczej y_0 obserwowanej w h-tej warstwie s_{0h} wydzielonej z próby s_o. Wariancję tej cechy w całej populacji oznaczono przez $v_*(y_0)$.

Podstawiając do wzoru (9.81) predyktor $\tilde{T}_w$ w miejsce statystyki $T_\otimes$ mamy:

$$T_w = \alpha \bar{Y}_{1S_1} + (1-\alpha)\tilde{T}_w \tag{9.96}$$

Błąd średniokwadratowy predyktora T_w wynika ze wzorów (9.80) i (9.95):

$$\mathcal{E}E(T_w - \bar{Y}_1)^2 = \alpha^2 \frac{N-k}{Nk}\left[a^2 v_*(y_0) + \sigma^2\right] +$$

$$+(1-\alpha)^2 \left(\frac{1}{\binom{N}{m}} \frac{a^2}{H} \frac{m-n}{mn} \sum_{s_0 \in \mathcal{S}_0} \sum_{h=1}^{H} v_{*S_{0h}}(y_0|s_0) + \right.$$

$$\left. +a^2 \frac{N-m}{Nm} v_*(y_0) - \frac{N-n}{Nn}\sigma^2 \right) \tag{9.97}$$

Predyktor T_w zależy m.in. od średniej $\tilde{T}_w$ z próby warstwowej, która, jak wiadomo, ma zwykle dużą precyzję oceny średniej w populacji. W naszym przypadku warstwy, z których są losowane podpróbki, tworzono na podstawie wartości cechy y_0 badanej lecz obserwowanej w okresie podstawowym bezpośrednio poprzedzającym okres bieżący. W związku z tym należy się spodziewać dużej autokorelacji pomiędzy badanymi cechami y_0 i y_1 w obu okresach. To z kolei pozwala przypuszczać, że warstwy utworzone na podstawie cechy y_0 będą bliskie optymalnemu układowi warstw, który otrzymalibyśmy na podstawie cechy y_1. Wtedy należy się spodziewać dużej dokładności predyktora $\tilde{T}_w$, a co za tym idzie także predyktora T_w. Precyzyjnych odpowiedzi na pytanie o efektywności predyktora T_w względem innych można uzyskać z badań symulacyjnych prowadzonych na podstawie rzeczywistych cech charakteryzujących populacje empiryczne lub zmiennych sztucznie generowanych z wykorzystaniem techniki komputerowej.

9.5.4. Śrdnia z próby pogrupowanej

W punkcie 7.3.3. analizowano określoną wzorem (7.51) średnią z próby grupowej, przy czym grupy losowano bezzwrotnie ze stałymi prawdopodobieństwami ich wyboru spośród grup utworzonych w próbie wstępnej na podstawie cechy pomocniczych. Rezultaty, które tam otrzymano, adaptujemy tutaj zakładając, że w wylosowanej w okresie podstawowym próbie prostej S_0 wydzielamy G rozłącznych i wyczerpujących zbiór S_0 grup równolicznych. Każdą z nich oznaczamy symbolem U_{Sh}, h=1,...,G, przy czym liczność każdej z nich wynosi M=m/G elementów populacji. Grupowanie odbywa się na zasadzie opisanej w punkcie 7.3.3., lecz na podstawie cechy y_0 obserwowanej w próbie S_0 w okresie podstawowym. Następnie ze zbioru tak utworzonych grup losujemy bezzwrotnie próbę prostą składającą się z g-grup. Wtedy na podstawie wzoru (7.51) konstruujemy następujący predyktor:

$$\tilde{\bar{T}}_g = \frac{1}{gM}\sum_{p=1}^{g} Z_p \tag{9.98}$$

gdzie:

$$Z_p = \sum_{i \in U_{Sp}} Y_{1i} \tag{9.99}$$

Na podstawie wzorów (7.60) i (7.61) otrzymujemy błąd średniokwadratowy predyktora $\tilde{\bar{T}}_g$:

$$\mathbf{E}\mathrm{E}(\tilde{\bar{T}}_g - \bar{Y}_1)^2 = \frac{a^2}{\binom{N}{m}} \frac{G-g}{GgM^3} \sum_{s_0 \in S_0} v_*(b_s | s_0) + a^2 \frac{N-m}{Nm} v_*(y_0) +$$

$$+\left(\frac{G-g}{GgM^2} - \frac{N-m}{Nm}\right)\sigma^2 \tag{9.100}$$

gdzie wartości zmiennej b_s są sumami obserwacji cechy y_o w poszczególnych grupach utworzonych w próbie s_o. Zatem

$$b_{Sh} = \sum_{i \in U_{Sh}} y_{oi}$$

Zastępując we wzorze (9.79) statystykę $T_{\otimes}$ przez $\tilde{\bar{T}}_g$ mamy następujący predyktor:

$$T_g = \alpha \bar{Y}_{1S_1} + (1-\alpha)\tilde{\bar{T}}_g \tag{9.101}$$

Na podstawie wzorów (9.80) i (9.100) wyznaczamy błąd średniokwadratowy predyktora T_g:

$$\mathcal{E}E(T_g-\overline{Y}_1)^2=\alpha^2\frac{N-k}{Nk}\left[a^2v_*(y_0)+\sigma^2\right]+(1-\alpha)^2\left(\frac{a^2}{\binom{N}{m}}\frac{G-g}{GgM^3}\sum_{s_0\in\mathcal{S}_0}v_*(b_s|s_0)+\right.$$

$$\left.+a^2\frac{N-m}{Nm}v_*(y_0)+\left(\frac{G-g}{GgM^2}-\frac{N-m}{Nm}\right)\sigma^2\right) \qquad (9.102)$$

Podobnie jak w uprzednim punkcie porównanie dokładności określonego predyktora z innymi wymaga osobnych i dość złożonych analiz, przy czym wydaje się, że badania symulacyjne z wykorzystaniem elektronicznej techniki komputerowej winny przyczynić się do rozwiązania tego problemu.

9.6. Weryfikacja hipotezy o równości dwóch zależnych rozkładów skokowych

Funkcję prawdopodobieństwa dwuwymiarowej zmiennej (X,Y) oznaczamy przez $P(X = x_i, Y = y_j) = p_{ij}$, dla i,j = l ..., m. W szczególności wartości zmiennej *X* mogą być obserwowane w okresie podstawowym, a wartości cechy Y w okresie bieżącym. Należy dodać, że cechy *X i Y* mogą być jakościowymi. Wtedy ich poziomy, są zdarzeniami losowymi.

Wprowadzimy oznaczenia:

$$p_{i.}=\sum_j p_{ij},\quad p_{.j}=\sum_i p_{ij}$$

$$\mathbf{p}_x=\begin{bmatrix}p_{1.}\\p_{2.}\\\dots\\p_{m.}\end{bmatrix},\qquad \mathbf{p}_y=\begin{bmatrix}p_{.1}\\p_{.2}\\\dots\\p_{.m}\end{bmatrix}$$

$$\mathbf{P}_{xy}=\left[p_{ij}\right];\qquad \mathbf{P}_{yx}=\mathbf{P}_{xy}^{T}$$

Określone prawdopodobieństwa szacujemy na podstawie n-elementowej próby prostej. Nieobciążonymi estymatorami wyżej zapisanych prawdopodobieństw są następujące statystyki:

$$\tilde{p}_{ij}=\frac{K_{ij}}{n},\qquad \tilde{p}_{i.}=\sum_j\tilde{p}_{ij},\qquad \tilde{p}_{.j}=\sum_i\tilde{p}_{ij} \qquad (9.103)$$

$$K_{i.} = \sum_j K_{ij,} \qquad K_{.j} = \sum_i K_{ij,}$$

gdzie przez K_{ij} oznaczono zmienną losową, której wartości są obserwowanymi w próbie liczbami obserwacji par wartości (x_i,y_j) zmiennych losowych (X,Y). Z kolei wartości zmiennej losowej $K_{i.}$ ($K_{.j}$) są liczbami wartości x_i (y_i) obserwowanymi w próbie.

$$\tilde{p}_{i.} = \frac{K_{i.}}{n}, \qquad \tilde{p}_{.j} = \frac{K_{.j}}{n}, \tag{9.104}$$

$$\tilde{\mathbf{P}}_{xy} = \left[\tilde{p}_{ij}\right], \qquad \tilde{\mathbf{P}}_{xy} = \tilde{\mathbf{P}}_{yx}$$

$$\tilde{\mathbf{p}}_x = \begin{bmatrix} \tilde{p}_{1.} \\ \tilde{p}_{2.} \\ \cdots \\ \tilde{p}_m \end{bmatrix}, \qquad \tilde{\mathbf{p}}_y = \begin{bmatrix} \tilde{p}_{.1} \\ \tilde{p}_{.2.} \\ \cdots \\ \tilde{p}_{.m} \end{bmatrix}$$

Różnicę prawdopodobieństw rozkładów brzegowych oznaczamy symbolem:

$$\tilde{\mathbf{d}} = \tilde{\mathbf{p}}_x - \tilde{\mathbf{p}}_y \tag{9.105}$$

Zakładając, że próba prosta jest losowana zwrotnie, macierz wariancji i kowariancji jest postaci:

$$\mathbf{V}\left(\tilde{\mathbf{d}}\right) = \mathbf{V}\left(\tilde{\mathbf{p}}_x\right) + \mathbf{V}\left(\tilde{\mathbf{p}}_y\right) - \mathbf{V}\left(\tilde{\mathbf{p}}_x, \tilde{\mathbf{p}}_y\right) - \mathbf{V}\left(\tilde{\mathbf{p}}_y, \tilde{\mathbf{p}}_x\right)$$

$$\mathbf{V}(\tilde{\mathbf{d}}) = \frac{1}{n}\left(\mathbf{D}_x + \mathbf{D}_y - \mathbf{P}_{xy} - \mathbf{P}_{yx} - \mathbf{d}\mathbf{d}^{\mathbf{T}}\right), \tag{9.106}$$

gdzie:

$$\mathbf{D}_x = \text{diag}\left(\mathbf{p}_x\right), \mathbf{D}_y = \text{diag}\left(\mathbf{p}_y\right).$$

Problem polega na weryfikacji hipotezy H_0: $\mathbf{p_x} = \mathbf{p}_y$ przy alternatywnej H_1: $\mathbf{p_x} \# \# \# \mathbf{p_y}$. Macierz wariancji i kowariancji przy prawdziwości H_0 ma postać:

$$\mathbf{V}\left(\tilde{\mathbf{d}} \middle| H_0\right) = \frac{1}{n}(\mathbf{D} - \mathbf{P}) = \frac{1}{n}\mathbf{A} \tag{9.107}$$

gdzie

$$\mathbf{D} = \mathbf{D_x} + \mathbf{D_y} \ , \quad \mathbf{P} = \mathbf{P_{xy}} + \mathbf{P_{yx}} \ .$$

Elementy macierzy $\mathbf{A} = \left[a_{ij} \right]$ można zapisać następująco:

$$a_{ij} = \begin{cases} p_{i.} + p_{.i} - 2 & \text{gdy} \quad i = j \\ -p_{ij} - p_{ji} & \text{gdy} \quad i \neq j \end{cases}$$

lub

$$a_{ij} = \begin{cases} \sum_{h=1}^{m} p_{ih} + \sum_{h=1}^{m} p_{hi} - 2\,p_{ii}, & \text{dla} \quad i = j \\ -p_{ij} - p_{ji} & \text{dla} \quad i \neq j \end{cases}$$

Zastąpmy prawdopodobieństwa, których funkcją są elementy macierzy **A** przez odpowiednie częstości względne określone wcześniej. Wtedy otrzymujemy macierz $\tilde{\mathbf{A}} = \left[\tilde{a}_{ij} \right]$, gdzie:

$$\tilde{a}_{ij} = \begin{cases} \sum_{h=1}^{m} \tilde{p}_{ih} + \sum_{h=1}^{m} \tilde{p}_{hi} - 2\,\tilde{p}_{ii}, & \text{dla} \quad i = j \\ -\tilde{p}_{ij} - \tilde{p}_{ji} & \text{dla} \quad i \neq j \end{cases} \tag{9.108}$$

Niech $\mathbf{A}^{-}$ i $\tilde{\mathbf{A}}^{-}$ będą g-inwersjami odpowiednich macierzy **A** i $\tilde{\mathbf{A}}$, a zatem $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$ oraz $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{-}\tilde{\mathbf{A}} = \tilde{\mathbf{A}}$.

Do testowania hipotezy H_0 użyjemy nastepującego sprawdzianu:

$$\tilde{Q} = \tilde{\mathbf{d}}^{T}\tilde{\mathbf{A}}^{-}\tilde{\mathbf{d}} \tag{9.109}$$

Duże wartości statystyki $\tilde{Q}$ świadczą przeciwko hipotezie H_0; Wartości krytyczne testu przy dużej liczebności próby wyznaczamy na podstawie twierdzenia:

Twierdzenie 9.1: Jeżeli rozkład dwuwymiarowej cechy (X,Y) jest symetryczny i liczebność prostej próby statystycznej $n \to \infty$, to $\tilde{Q}$ ### χ^2_r , gdzie liczba stopni swobody r jest równa rzędowi macierzy **A** formy kwadratowej $\tilde{Q}$.

Dowód: Można wykazać, że jeśli rozkład dwuwymiarowej cechy (X,Y) jest symetryczny, to hipoteza H_0, jest prawdziwa. Wtedy, przy $n \to \infty$ rozkład wektora $\tilde{\mathbf{d}}$ jest niezależny od rozkładu macierzy $\tilde{\mathbf{A}}$ tej formy kwadratowej.

Na podstawie znanego słabego prawa wielkich liczb Bernoulliego wnioskujemy, że każdy element macierzy $\tilde{\mathbf{A}}$ jest stochastycznie zbieżny do odpowiedniego elementu macierzy **A**.

Niech $\mathbf{A}^-$ i $\tilde{\mathbf{A}}^-$ będą g-inwersjami odpowiednich macierzy $\mathbf{A}$ i $\tilde{\mathbf{A}}$. Elementy macierzy $\tilde{\mathbf{A}}^-$ są stochastycznie zbieżne do odpowiednich elementów macierzy $\mathbf{A}^-$, ponieważ elementy macierzy $\tilde{\mathbf{A}}^-$ i $\mathbf{A}^-$ są wymiernymi funkcjami elementów macierzy odpowiednio $\tilde{\mathbf{A}}$ i $\mathbf{A}$ [por.np.Fisz (1967), s.251].

Na podstawie wielowymiarowej wersji znanego centralnego twierdzenia Lindeberga-Levy'ego wnioskujemy, że wektor $\tilde{\mathbf{d}}$ zmierza do rozkładu zmiennej **d**, która ma m-wymiarowy rozkład normalny z wektorem wartości oczekiwanych E(**d**)=### i macierzą wariancji i kowariancji **A**.

Stąd i z faktu stochastycznej zbieżności $\tilde{\mathbf{A}}^- \to \mathbf{A}^-$ wynika na podstawie znanego twierdzenia o zbieżności rozkładu statystyk [por.np. Rao (1982), s. 395], że rozkład statystyki $\tilde{Q}$ zmierza do rozkładu zmiennej losowej $Q = \mathbf{d}^T \mathbf{A}^- \mathbf{d}$. Z kolei na podstawie ogólnego twierdzenia demonstrowanego w pracy Rao i Mitry (1971), s.173 wnioskujemy, że Q### χ^2_r, gdzie r jest rzędem macierzy $\mathbf{A}^-$, co należało wykazać.

Wiele proponowanych testów dla weryfikacji hipotezy o równości rozkładów cechy w dwóch okresach czasu wymaga spełnienia założenia o niezależności tych rozkładów. Trudno oczekiwać, by to założenie było spełnione w praktyce. Przedstawiony tutaj test nie wymaga spełnienia tego założenia. Może być on użyty np. do testowania równości wydatków rodzin na określone dobra w dwóch okresach czasu.

BIBLIOGRAFIA

Anderson T.W. (1958): An introduction to multivariate statistical analysis. John Wiley & Sons, New York.

Armstrong J.B., Wu C.F.J. (1992): A sample allocation method for two phase survey design. "Survey Methodology", vol. 18, nr 2, s. 253-262.

Arwanitis L.G., Afonia B. (1971): Use of generalized variance and the gradient projection method in multivariate stratified sampling. "Biometrics", vol. 27, s. 119-127.

Barnard G.A. (1971): Discussion of paper by V.P.Godambe and M.E.Thompson. "Journal of the Royal Statistical Society", vol. B 33, s. 376-378.

Bartoszewicz J. (1989): Wykłady ze statystyki matematycznej. PWN, Warszawa.

Basu D. (1969): Role of the sufficiency and likelihood principles in sample survey theory. "Sankhya", vol. A 31, s. 441-454.

Basu D. (1971): An essay on the logical foundations of survey sampling. In: Fundations of Statistical Inference. Edited by V.P.Godambe and D.A.Sprott. Holt, Rinehart and Winston of Canada Ltd. Toronto-Montreal.

Beardwood J., Halton J.H., Hammersley J.M. (1959): The shortest path through many points. "Proceedings of Cambridge Phil. Soc.", vol. 55.

Bethel J. (1989): Sample allocation in multivariate surveys. "Survey Methodology", vol. 15, nr 1, s. 47-57.

Biggeri L. (1977): Some application of cluster analysis in sample design. "Bulletin of the International Statistical Institute", vol. XLVII, book 4.

Blythe J.R.H. (1945): The economics of sample size applied to the scaling of sowlongs. "Biometrics Bulletin", vol. 1, s. 67-70.

Borowkow A.A. (1984): Matematiczeskaja statistika. Ocenka parametrow. Prowierka gipotez. Nauka, Moskwa.

Borsuk K. (1976): Geometria analityczna wielowymiarowa. PWN, Warszawa.

Bracha Cz. (1978): Szacowanie parametrów liniowej funkcji regresji na podstawie próby losowanej bez zwracania z populacji skończonej. "Przegląd Statystyczny", vol. 25.

Bracha Cz. (1979): Szacowanie liniowej funkcji regresji na podstawie prób nieprostych. SGPiS, Warszawa (nie publikowana).

Bracha Cz. (1982): Szacowanie parametrów liniowej funkcji regresji na podstawie próby losowanej dwustopniowo. "Przegląd Statystyczny", vol. 29.

Bracha Cz. (1983): Regresja liniowa w badaniach reprezentacyjnych. Prace i Materiały Instytutu Cybernetyki i Zarządzania. SGPiS, Warszawa.

Bracha Cz. (1987): Wykorzystanie informacji o cechach dodatkowych w badaniach reprezentacyjnych. ZBS-GUS i PAN, Warszawa.

Bracha Cz. (1987 a): Wykorzystanie metody podprób do szacowania błędów średniokwadratowych estymatorów złożonych. "Wiadomości Statystyczne" nr 32/2, s. 9-13.

Bracha Cz. (1987 b): Wykorzystanie metody podprób do redukcji obciążenia estymatorów. Z Prac Zakładu Badań Statystyczno Ekonomicznych - zeszyt nr 166 na temat: Zastosowanie Metody Reprezentacyjnej w Badaniach Statystycznych GUS (1981-1986). GUS, Warszawa.

Bracha Cz. (1991): Wybrane problemy losowania warstwowego. Prace i Materiały Instytutu Cybernetyki i Zarządzania. Tom 20. SGPiS, Warszawa.

Brewer K.R.W. (1963): A model of systematic sampling with unequal probabilities. "Australian Journal of Statistics", vol. 5, s. 5-13.

Brewer K.R.W.,Hanif M. (1983): Sampling with unequal probabilities. Springer Verlag, New York-Heidelberg-Berlin 1983.

Cassel C.M., Sarndal C.E., Wretman J.H. (1977): Foundation of inference in survey sampling. John Wiley & Sons, New York-London-Sydney-Toronto.

Chatterjee S. (1968): Multivariate stratified surveys. "Journal of the American Statistical Association", vol. 63, s. 530-534.

Chaudhuri A. (1971): Some sampling shemes to use Horvitz-Thompson estimator in estimation a finite population total. "Bulletin of the Calcutta Statistical Association", vol. 20, s. 37-66.

Chaudhuri A., Vos J.W.E. (1988): Unified theory of survey sampling. North Holland, Amsterdam-New York-Oxford-Tokyo.

Cochran W.G. (1939): The use of analysis of variance in enumeration by sampling. "Journal of the American Statistical Association", vol. 34, s. 492-510.

Cochran W.G. (1946): Relative accuracy of systematic and stratified random samples for a certain class of population. "Annals of Mathematical Statistics", vol. 17, s. 164-177.

Cochran W.G. (1961): Comparison of methods for determining stratum boundaries. "Bulletin of International Statistical Institute", vol.38, nr 2, s. 345-358.

Cochran W.G. (1963): Sampling techniques. John Wiley & Sons, New York.

Cramer H. (1958): Metody matematyczne w statystyce. PWN, Warszawa.

Czerniak W. (1971): O losowaniu niezależnym ze zwracaniem. "Biblioteka Wiadomości Statystycznych". Tom 15, s. 40-86.

Dalenius T. (1953): The multivariate sampling problem. "Scandinavisk Aktuarietidskrift", vol. 36, s. 92-102.

Dalenius T. (1953 a): The economic of one stage stratified sampling. "Sankhya", vol. 16, s. 351-356.

Dalenius T. (1957): Sampling in Sweden. Contribution to methods and theories of sample survey practice. Almqvist & Wiksells, Stockholm.

Dalenius T., Gurney M. (1951): The problem of optimum stratification II. "Scandinavisk Aktuarietidskrift", vol. 34, s. 133-148.

Dalenius T., Hodges J.L.Jr. (1959): Minimum variance stratification. "Journal of the American Statistical Association", vol. 54, s. 88-101.

Dayal S. (1985): Allocation of sample using values of auxiliary characteristic. "Journal of Statistical Planning and Inference", vol. 11, s. 321-328.

Demidowicz B.P., Maron I.A. (1965): Metody numeryczne, cz. I. PWN, Warszawa.

Deming W.E., Stephan F. (1941): On the interpretation of censuses as samples. "Journal of the American Statistical Association", vol. 36, s. 45-49.

Dryja M., Jankowscy J. M. (1988): Przegląd metod i algorytmów numerycznych. WNT, Warszawa.

Eckler A.R. (1955): Rotating sampling. "Annals of Mathematical Statistics", vol. 26, s. 664-685.
Fellegi I.P. (1963): Sampling with varying probabilities without replacement. Rotating and non rotating samples. "Journal of the American Statistical Association", vol. 58, s. 183-201.
Findeisen W., Szymanowski J., Wierzbicki A. (1967): Teoria i metody obliczeniowe optymalizacji. PWN, Warszawa.
Fisher R.A. (1922): On the mathematical fundation of theoretical statistics. "Phil. Trans. Roy. Soc.", A 222, s. 309-368.
Fisz M. (1967): Wstęp do rachunku prawdopodobieństwa i statystyki matematycznej. PWN, Warszawa.
Folks J.L., Antle Ch.E. (1965): Optimum allocation of sampling units to strata when there are R responses of interest. "Journal of the American Statistical Association", vol. 60.
Friedman H.P., Rubin J. (1967): On some invariant criteria for groupping data. "Journal of the American Statistical Association", vol. 62, s. 1159-1178.
Frish R. (1929): Correlation and scatter in statistical variables. "Nord. Statist. Tidskr.", vol. 8, s. 36.
Gabler S., Schweigkoffer R. (1990): The existance of sampling designs with preassigned inclusion probabilities. "Metrika", vol. 37, s. 87-96.
Galas Z., Nykowski I., Żółkiewski Z. (1987): Programowanie wielokryterialne. PWN, Warszawa.
Geary R.C. (1949): Sampling methods applied to Irish agricultural statistics. Technical Series.
Ghosh J.K. (1963): A game theory approach to the problem of optimum allocation in stratified sampling with multiple characters. "Calcutta Statistical Association Bulletin", vol. 12, s. 4-12.
Ghosh J.K. (1963a.): "Annals of Mathematical Statistics", vol. 34, s. 587-597.
Ghosh S.P. (1958): A note on stratified random sampling with multiple characters. "Calcutta Statistical Association Bulletin", vol. 8, s. 81-90.
Godambe V.P. (1955): A unified theory of sampling from finite populations. "Journal of the Royal Statistical Society", vol. B 17, s. 269-278.
Godambe V.P. (1960): An admissible estimate for any sampling design. "Sankhya", vol. 22, s. 285-288.
Godambe V.P. (1966): A new approach to sampling from finite population I,II. "Journal of the Royal Statistical Society", vol. B 28, s. 310-328.
Godambe V.P., Joshi V.M. (1965): Admissibility and Bayes estimation in sampling finite populations I. "Annals of Mathematical Statistics", vol. 36, s. 1707-1722.
Godwin H.J. (1964): Inequalities on distribution functions. Charles Griffin and Company Limited, London.
Grabiński T., Wydymus S., Zeliaś A. (1989): Metody taksnomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych. PWN, Warszawa.
Grabowski W. (1982): Programowanie matematyczne. PWE, Warszawa.
Greń J. (1963): Lokalizacja próby w wieloparametrowym losowaniu warstwowym. "Przegląd Statystyczny", vol. 10, s. 291-302.
Greń J. (1963 a): Zagadnienie lokalizacji próby w wieloparametrowym losowaniu warstwowym. SGPiS, Warszawa (nie publikowana).
Greń J. (1964): O pewnych metodach wyznaczania lokalizacji próby w losowaniu warstwowym wieloparametrowym. "Przegląd Statystyczny", vol. 11, s. 361-369.
Greń J. (1966): O pewnym zastosowaniu programowania nieliniowego do metody reprezentacyjnej. "Przegląd Statystyczny", vol. 13, s. 203-217.

Greń J. (1969): Wielowymiarowy estymator regresyjny średniej. "Przegląd Statystyczny", vol. 16.

Greń J. (1970): Wielowymiarowy estymator regresyjny średniej dla skończonej populacji. "Przegląd Statystyczny", vol. 17, s. 73-78.

Greń J., Koźniewska I. (1964): Rozwiązywanie pewnego równania rekurencyjnego związanego z dwuparametrowym losowaniem warstwowym. "Przegląd Statystyczny", vol. 11, s. 169-176.

Hansen M.H., Hurwitz W.N., Madow W.G. (1953): Sampling survey methods and theory. Tom I i II. John Wiley & Sons, New York.

Hartley H.O. (1965): Multiple purpose optimum allocation in stratifed sampling. Proceedings of the American Statistical Association, Social Statistics Section, s. 258-261.

Hartley H.O., Rao J.N.K. (1968): A new estimation theory for sample surveys. "Biometrika", vol. 55, s. 547-557.

Hartley H.O., Rao J.N.K. (1969): A new estimation theory for sample surveys II. In: New developments in survey sampling. Edited by N.L. Johnson and H. Smith. Wiley-Interscience, New York, s. 147-169.

Hartley H.O., Ross A. (1954): Unbiased ratio estimates. "Nature", vol. 174, s. 270-271.

Hedricks W.A. (1944): The relative efficiences of groups of farms as sampling units. "Journal of the American Statistical Association", vol. 39, s. 367-376.

Hellwig Z. (1987): Elementy rachunku prawdopodobieństwa i statystyki matematycznej. PWN, Warszawa.

Herzel A. (1986): Sampling without replacement with unequal probabilities: sample designs with preassigned joint inclusion probabilities of any order. "Metron", vol. 44, nr. 1-4, s. 49-68.

Herzel A. (1989): On product estimation in simple random sampling. "Statistica", vol. XLIX, s. 3-20.

Hess I., Sethi V.K., Balakrishnan T.R. (1966): Stratification - a practical investigation. "Journal of the American Statistical Association", vol. 61.

Horvitz D.G., Thompson D.J. (1952): A generalization of sampling without replacement from finite universe. "Journal of the American Statistical Association", vol. 47, s. 663-685.

Huddleston H.F., Claypool P.L., Hocking R.R. (1970): Optimal sample allocation to strata using convex programming. "Applied Statistics", vol. 19, s. 273-278.

Hughes E., Rao J.N.K. (1979): Some problems of optimal allocation in sample surveys involving inequality constrains. "Communcation in Statistics", vol. A8, s. 1551-1571.

Jaganathan R. (1965): The programming approach in multiple characters studies. "Econometrica", vol. 33, s. 236-237.

Jaganathan R. (1965 a): A method for solving a nonlinear programming problem in sample surveys. "Econometrica", vol. 33, s. 841-846.

Jefimow N.W., Rozendorn E.R. (1974): Algebra liniowa z geometrią wielowymiarową. PWN, Warszawa.

Jessen R.J. (1942): Statistical investigation of a sample survey for obtaining farm facts. "Iowa Agr. Exp. Sta. Res. Bull.", vol.304.

Jessen R.J. (1978): Statistical survey techniques. John Wiley & Sons, New York-Chichester-Brisbane-Toronto-Singapore.

John S. (1969): On multivariate ratio and product estimators. "Biometrika", vol. 56, s. 533-537.

Jonin B.G., Jonina N.P., Żurawlew N.M. (1978): Ispolzowanie procedur optimizacji pri kłasifikacji objektow i faktorow. W: Ekonomika i statisticzeskije modieli w prognozirowani płanirowani promyszlennowo proizbodstwa. Nauka, Moskwa.

Joshi V.M. (1965): Admissibility and Bayes estimation in sampling finite populations II i III. "Annals of Mathematical Statistics", vol. 36, s. 1658-1670.
Joshi V.M. (1966): Admissibility and Bayes estimation in sampling finite populations IV. "Annals of Mathematical Statistics", vol. 37, s. 1658-1670.
Kendall M.G., Stuart A. (1966): Teoria raspriedielienij. Nauka, Moskwa.
Kish L. (1961): Efficient allocation of multipurpose sample. "Econometrica", vol. 29, s. 363-385.
Kish L. (1965): Survey sampling. John Wiley & Sons, Inc. New York- London-Sydney.
Kish L. (1979): Population for survey sampling. "Survey Statistician", vol. 1, s. 14-15.
Kokan A.R. (1963): Optimum allocation in multivariate surveys. "Journal of the Royal Statistical Society", vol. A 126, s. 557-565.
Kokan A.R., Khan S. (1967): Optimum allocation in multivariate surveys: an analytical solution. "Journal of the Royal Statistical Society", vol. B 29, s. 115-125.
Kolonko J. (1980): Analiza dyskryminacyjna i jej zastosowania w ekonomii. PWN, Warszawa.
Konarzewska-Gubała E. (1980): Programowanie przy wielorakości celów. PWN, Warszawa.
Konijn H.S. (1962): Regression analysis in sample surveys. "Journal of the American Statistical Association", vol. 57.
Konijn H.S. (1973): Statistical theory of sample survey and analysis. North-Holland Publishing Company, Inc., Amsterdam-London, American Elsevier Publishing Company, Inc., New York.
Kończak G. (1995): Metody oceny dokładności szacowania indeksów ekonomicznych. Niepublikowana praca doktorska, Akademia Ekonomiczna w Katowicach.
Koop J.C. (1963): On the axioms of sample formation and their bearing on the construction of linear estimators in sampling theory for finite universes. "Metrika", vol. 7, s. 165-204.
Kordos J. (1987): Dokładność danych w badaniach społecznych. Biblioteka Wiadomości Statystycznych, GUS, Warszawa.
Kordos J. (1988): Jakość danych statystycznych. PWE, Warszawa.
Kostrykin A.I. (1984): Wstęp do algebry. PWN, Warszawa.
Kowal R.R. (1971): Disadvantages of the generalized variance as a measure of variability. "Biometrics", vol. 27, s. 213-216.
Kręglewski T., Rogowski T., Ruszczyński A., Szymanowski J. (1984): Metody optymalizacyjne w języku FORTRAN. PWN, Warszawa.
Kubik L.T., Krupowicz A. (1982): Wprowadzenie do rachunku prawdopodobieństwa i jego zastosowań. PWN, Warszawa.
Lahiri G.W. (1951): A method for sample selection providing unbiased ratio estimator. "Bulletin of the International Statistical Institute", vol. 33, s. 133-140.
Lipski W., Marek W. (1986): Analiza kombinatoryczna. PWN, Warszawa.
Lynch G.W. (1978): The choise of auxiliary variables in multivariate ratio and regression estimators. Proceedings of the Section on Survey Research Methods. American Statistical Association.
Madow W.G., Madow L.H. (1944): On the theory of systematic sampling. "Annals of Mathematical Statistics", vol. 15, s. 1-24.
Magnus J.R., Neudecker H. (1988): Matrix differential calculs with application in statistics and econometrics. John Wiley & Sons, Chichester-New York-Birsbane-Toronto-Singapore.
Mahalanobis P.C. (1944): On large scale sample surveys. "Phil. Trans. Roy. Soc." London, vol. B 231, s. 329-451.

Marczyńska K. (1984): Podział populacji na warstwy przeprowadzony przy pomocy zmiennej pomocniczej. W: Modele, prognozy i optymalne programy działania. Prace Naukowe Akademii Ekonomicznej w Katowicach, s. 137-145.
Martos B. (1983): Programowanie nieliniowe. Teoria i metody. PWN, Warszawa.
Melaku A. S. (1987): L d-dnorm and other methods for sample 1 allocation in multivariate stratified surveys. "Computational Statistics & Data Analysis", vol. 5, s. 415-423.
Mikhail N.N., Mir. M.A. (1981): Unbiased estimates of the generalized variance for finite population. "Journal of the Indian Statistical Association", vol. 19, s. 85-92.
Mostowski A., Stark M. (1977): Elementy algebry wyższej. PWN, Warszawa.
Mukerjee R., Rao T.J. (1985): On a problem of allocation of sample size in stratified random sampling. "Biometrical Journal", vol. 27, nr 3, s. 327-331.
Murthy M.N. (1964): Product method of estimation. "Sankhya", vol. Ad26, s. 69-74.
Murthy M.N. (1977): Sampling theory and practice. Statistical Publishing Society, Calcutta.
Neyman J. (1934): On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. "Journal of the Royal Statistical Society", vol. 97, s. 558-606.
Neyman J. (1937): Outline of theory of statistical estimation based on the classical theory of probability. "Philos. Trans. Roy. Soc.", vol. A 236, s. 333-380.
Neyman J. (1938): Contribution to the theory of sampling human population. "Journal of the American Statistical Association", vol. 33, s. 101-116.
Olkin I. (1958): Multivariate ratio estimation for finite populations. "Biometrika", vol 45, s. 154-165.
Patterson H.D. (1950): Sampling on successive occasions with partial replacement of units. "Journal of the Royal Statistical Society", vol. B12, s. 241-255.
Pawłowski Z. (1972): Wstęp do statystycznej metody reprezentacyjnej. PWN, Warszawa.
Pawłowski Z. (1976): Statystyka matematyczna. PWN, Warszawa.
Peaerson K. (1901): On lines planes of closest fit to systems of points in space. "Phil. Mag.", vol. 6, s. 559.
Prakasa Rao B.L.S. (1987): Asymptotic theory of statistical inference. John Wiley & Sons, New York-Chihester-Brisbane-Toronto-Singapore.
Quenouille M.H. (1956): Notes on bias in estimation. "Biometrika", vol. 43, s. 353-360.
Raj D. (1965): On a method of using multi-auxiliary information in sample surveys. "Journal of the American Statistical Association", vol. 60, s. 270-277.
Ralston A. (1975): Wstęp do analizy numerycznej. PWN, Warszawa.
Ramakrishnan M. K. (1975): Choise of an optimum sampling strategy I. "Annals of Statistics", vol. 3, s. 669-679.
Rao C.R. (1982): Modele liniowe statystyki matematycznej. PWN, Warszawa.
Rao C. R., Mitra S. K. (1971): Generalized inverse of matrices and its applications. John Wiley & Sons, New York-London-Sydney-Toronto.
Rao J.N.K. (1973): On double sampling for stratification and analytical surveys. "Biometrika", vol. 60, nr 1, s. 125-133.
Rao P.S.R.S., Mudholkar G.S. (1967): Generalized multivariate estimator for the mean of finite populations. "Journal of the American Statistical Association", vol. 62, s. 1009-1012.
Rao T.V.H. (1962): An existence theorem in sampling theory. "Sankhya", vol. A 24, s. 327-330.
Reddy V.N. (1974): A transformed ratio method of estimator. "Sankhya", vol. C 36, s. 59-70.
Rice J.R. (1983): Numerical methods, software and analysis. McGraw-Hill Book Company, New York.

Robbinson H. (1952): Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, s. 527-535.
Robson D.S. (1957): Application of multivariate polykage of the theory of unbiased ratio-type estimator. "Journal of the American Statistical Association", vol. 52, s. 511-522.
Roy J., Chakravarti J.M. (1960): Estimating the mean of a finite population. "Annals of Mathematical Statistics", vol. 31, s. 392-398.
Royall R.M. (1968): An old approach to finite population sampling theory. "Journal of the American Statistical Association", vol. 63, s. 1269-1279.
Royall R.M. (1970): On finite population sampling theory under certain linear regression models. "Biometrika", vol. 57, s. 377-387.
Sarndal C.E. (1972): Sample survey theory vs. general statistical theory: Estimation of the population mean. "Revue of the International Statistical Institute", vol. 40, s. 1-12.
Sarndal C.E. (1976): On uniformly minimum variance estimation in finite populations. "Annals of Statistics", vol. 4, s. 993-997.
Sahoo J. (1994): Some estimation problems in finite population sampling using auxiliary information. Nie publikowana praca doktorska, Utkal University.
Searle S.R. (1966): Matrix algebra for the biological sciences. John Wiley & Sons, New York.
Sengupta S. (1981): Jack-knifing the ratio and the product estimators in double sampling. "Metrika", vol. 28, s. 245-256.
Serfling R.J. (1968): Approximately optimum stratification. "Journal of the American Statistical Association", vol. 63.
Shah D.N., Gupta M.R. (1987): An efficiency comparison of dual ratio and product estimators. "Communication in Statistics-Theory and Methods", vol. 16, s. 693-703.
Shukla N.D. (1976): Almost unbiased product type estimator. "Metrika", vol. 23, s. 127-133.
Sikorski R. (1972): Rachunek różniczkowy i całkowy. Funkcje wielu zmiennych. PWN, Warszawa.
Singh H.P. (1989): A class of unbiased estimators of product of population means. "Journal of the Indian Society of Agricultural Statistics", vol. 41, s. 113-118.
Singh H.P., Iachan R., Upadhyaya L.N. (1985): Almost unbiased ratio and product estimators based on interpenetring subsamples. "Communication in Statistics-Theory and Methods", vol. 14, s. 963-978.
Singh P., Srivastava A.K. (1980): Sampling schemes providing unbiased regresion estimators. "Biometrika", vol. 67, s. 205-209.
Sinha B.K. (1973): On sampling schemes to realize preassigned sets of inclusion probabilities of selection. "Biometrika", vol. 54, s. 499-513.
Srikantan (1963): Problems in optimum allocation. "Operational Research", vol. 11, s. 265-273.
Srivastava S.K. (1983): Predictive estimation of finite population mean using product estimator. "Metrika", vol. 30, s. 93-99.
Srivastava V.K., Shukla N.D., Bhatnagor S. (1981): Unbiased product estimator. "Metrika", vol. 28, s. 191-196.
Srivenkataramana T. (1980): A dual to ratio estimator in sample surveys. "Biometrika", vol. 67, s. 199-204.
Stam A.J. (1978): Distance between sampling with and without replacement. "Statistica Neerlandica", vol. 32, s. 81-91.
Stam A.J. (1986): A note on sampling with and without replacement. "Statistica Neerlandica", vol. 40, s. 35-38.

Steczkowski J. (1988): Zastosowanie metody reprezentacyjnej w badaniach społeczno-ekonomicznych. PWN, Warszawa.

Theil H. (1979): Zasady ekonometrii. PWN, Warszawa.

Tripathi T.P. (1973): Double sampling for indusion probabilities and regression method of estimation. "Jourlnal of the Indian Statistical Association", vol. 10, s.33-46.

Tripathi T.P. (1976): On double sampling for multivariate ratio and difference methods at estimation. "Jourlnal of the Indian Society of Agricultural Statistics". vol. 28, nr 1, s.33-54.

Tripathi T.P., Mir A.H., Chaturvedi D.K. (1989): Estimation of mean on second occasion using PPS sampling and multivariate information. "Aligarh Jourlnal of Statistics", vol. 9, s.16-27.

Tschuprow A.A. (1923): On the mathematical expectation of the moments of frequency distribution in the case of correlated observations. "Metron", vol. 2, s. 461-493 i s. 646-683.

Vos J.W.E. (1980): Mixing of direct, ratio, and product method estimators. "Statistica Neerlandica", vol. 34, s. 209-219.

Ward J.H. (1963): Hierarchical groupping to optimise an objective function. "Journal of the American Statistical Association", vol. 58.

Wilks S.S. (1932): Certain generalization in the analysis of variance. "Biometrika", vol. 24, s. 471-494.

Wilks S.S. (1960): Multidimensional statistical scatter. W: Contribution to probability and statistics. Editor I. Olkin. Stanford University Press.

Wilks S.S. (1962): Mathematical statistics. John Wiley & Sons, Inc., New York-London.

Wit R. (1986): Metody programowania nieliniowego. WNT, Warszawa.

Wywiał J. (1983): Badanie zmienności parametrów rozkładów warunkowych. "Przeglad Statystyczny", vol. 30, s.213-222.

Wywiał J. (1985): Estymacja wektora wartości średnich w skończonych populacjach statystycznych. Akademia Ekonomiczna w Katowicach. Nie publikowana praca doktorska.

Wywiał J. (1986): O wyznaczniku sumy macierzy. "Zeszyty Naukowe Akademii Ekonomicznej w Katowicach", nr 4/104, s. 95-118.

Wywiał J. (1987): Optymalizacja liczebności prób otrzymywanych z dwustopniowego schematu losowania przy estymacji wektora średnich w skończonej populacji. W: Problemy budowy i zastosowania modeli ekonometrycznych. SGPiS, Warszawa, s. 379-395.

Wywiał J. (1988): Estymacja wektora wartości średnich za pomocą estymatorów różnicowych i regresyjnych. "Przegląd Statystyczny", vol. 35, s. 19-35.

Wywiał J. (1988 a): Lokalizacja próby w warstwach minimalizująca promień spektralny macierzy wariancji i kowariancji wektora średnich z próby. "Prace Naukowe Akademii Ekonomicznej we Wrocławiu", nr 404, s. 195-200.

Wywiał J. (1988 b): Estymacja wartości średnich wielowymiarowej zmiennej w skończonej populacji przy pomocy wektora wartości średnich z prób przenikających się. W: Metody statystyczne, ekonometryczne i porogramowania matematycznego. Prace Naukowe Akademii Ekonomicznej w Katowicach, s. 223-235.

Wywiał J. (1989): O minimalizacji uogólnionej wariancji wektora średnich z próby losowanej z warstw. "Wiadomości Statystyczne", nr 11, s. 23-24.

Wywiał J. (1989 a): Wpływ sposobu grupowania populacji na dokładność estymacji wartości przeciętnych prowadzonej przy pomocy wektora średnich z próby grupowej. W: Informatyka w dydaktyce i badaniach naukowych szkół ekonomicznych. Prace Naukowe Akademii Ekonomicznej w Katowicach, s. 105-114.

Wywiał J. (1989 b): Zastosowanie dekompozycji uogólnionej wariancji do badania zmian struktury w przestrzeni obiektów i cech. W: Ekonometryczna analiza zmian struktury.

Praca zbiorowa pod red. K.Zadory, w ramach problemu CPBP 10.09. Akademia Ekonomiczna w Katowicach.

Wywiał J. (1989 c): Współczynniki rzetelności danych wielowymiarowych. Biblioteka Wiadomości Statystycznych. Tom 36. Warszawa, s. 120-126.

Wywiał J. (1990): Estymacja średnich wartości zmiennych o niejednorodnej strukturze rozkładu w populacji. W: Ekonometryczna analiza zmian struktury. Praca zbiorowa pod red. K.Zadory, w ramach problemu CPBP 10.09. Akademia Ekonomiczna w Katowicach.

Wywiał J. (1990 a): Weryfikacja wybranych hipotez o błędach specyfikacji modelu regresji liniowej. "Zeszyty Naukowe Akademii Ekonomicznej w Katowicach", nr 115, s. 132-144.

Wywiał J. (1990 b): O optymalnej lokalizacji próby w losowaniu warstwowym przy estymacji wektora wartości średnich w skończonych zbiorowościach. "Zeszyty Naukowe Akademii Ekonomicznej w Katowicach", nr 117, s. 19-32.

Wywiał J. (1991): Badanie mocy wybranych testów na występowanie wartości oddalonych. W: Ekonometria. Materiały XXV Konferencji Ekonometrycznej i VII Seminarium im Zbigniewa Pawłowskiego. Akademia Ekonomiczna w Katowicach, s. 35-49.

Wywiał J. (1991 a): Własności średniej z próby zespołowej losowanej z populacji pogrupowanej przy pomocy cechy pomocniczej. "Prace Naukowe Akademii Ekonomicznej we Wrocławiu", nr 559, s. 107-112.

Wywiał J. (1991 b): O planie losowania próby proporcjonalnym do średniej z próby cechy pomocniczej. "Wiadomości Statystyczne", s. 21-22.

Wywiał J. (1992): Statystyczna metoda reprezentacyjna w badaniach ekonomicznych (Optymalizacja badań próbkowych). Prace Naukowe Akademii Ekonomicznej w Katowicach.

Wywiał J. (1992 a): O pewnych współczynnikach zróżnicowania wielowymiarowej zmiennej i uogólnieniu metody Friedmana-Rubina grupowania populacji skończonej. "Zeszyty Naukowe Akademii Ekonomicznej w Katowicach", nr 120, 129-149.

Wywiał J. (1993): Variances of the Horvitz-Thompson Estimator for Two Sampling Designs Dependent on an Auxiliary Variable. "Statistics in Transition", vol. 1., no 1, 79-87.

Wywiał J.(1993a): Predykcja średniej w populacji pogrupowanej według cech dodatkowych. "Przegląd Statystyczny", vol. XL, zeszyt 3-4, s. 303-308.

Wywiał J. (1994): O metodzie Warda grupowania zbiorów. Prace Naukowe Akademii Ekonomicznej weWrocławiu, nr 667, s. 119-122.

Wywiał J., Kończak G. (1994): O lokalizacji próby w warstwach minimalizującej promień spektralny wewnątrzwarstwowej macierzy wariancji i kowariancji. W: XI Seminarium Ekonometryczne im. Profesora Zbigniewa Pawłowskiego. Trzemieśnia 24-26 III. Akademia Ekonomiczna w Krakowie, s. 85-92.

Yates F., Grundy P.M. (1953): Selection without replacement from within strata with probability proporcional to size. "Journal of the Royal Statistical Society", vol. B15, s. 235-261.

Yates F. (1960): Sampling methods for censuses and surveys. Griffin & Company Lth., London.

Zacks S. (1969): Bayes sequential designs for sampling finite populations. "Journal of the American Statistical Association", vol. 64, s. 1342-1349.

Zarkovich S.S. (1966): Quality of statistical data. FAO, Roma.

Zasępa R. (1972): Metoda reprezentacyjna. PWE, Warszawa.

Zieliński R. (1990): Siedem wykładów wprowadzających do statystyki matematycznej. PWN, Warszawa.

Multidimensional Aspects of Survey Sampling

Summary

Statistical research is usually connected with simultaneous estimation of at least one parameter. The theory of survey sampling is well developed in the case connected with the estimation of a singular parameter. This book is treated as a contribution to the development of the method of vector estimation in a finite population. Almost all procedures presented in this book have been proposed by the author or they are generalizations or modifications of solutions to the problems well known in a one-dimensional case. In the following paragraphs the author enumerates the results of his research to be found in the book.

The contributions to interpretations of the following measures of accuracy of vector estimators: the generalized variance, the mean radius and spectral radius defined as a determinant, the trace and the maximal eigenvalue of the variance-covariance matrix, respectivelly.

Extentions of some definitions and theorems, known in a one-dimensional case, on the vector estimation case. They let compare the accuracy of vector estimators. Some generalization of Tchebysheff inequality is introduced, too.

The definitions of new sampling designs and sampling schemes dependent on the following parameters of auxilliary variables: the sample variance, the squared difference between the sample mean and the population mean, the adjacent population elements.

The approximate expressions of the variance of the Horovitz-Thompson estimator of the mean value are derived for the above sampling designs.

The unbiased estimators of the generalized variance are found in the cases when the simple sample is drawn with as well as without replacement.

The basic properties of the vector of the regression estimators are derived. It is proved that the vector of regression estimators is efficient in the class of the vector of the difference estimators in the case of a simple sample.

Let the double sample consist of the following two samples: the first one is a simple sample drawn without replacement from a population. The other one is also a simple sample but selected from the first sample. Several problems concerning the optimisation of determining the size of the above two samples are formulated and solved. The square risk function (or the generalized variance of the vector of the regression estimators) is minimized under the fixed total cost of observation of variables in the double sample. Next the cost function is minimized under the fixed variances of the regression estimators of the mean of particular variables either under the fixed value of the generalized variance of the vector estimators or under the fixed square risk function.

The similar optimization problems are formulated and solved to determine the sizes of a stratified sample or a two-stage one or some samples drawn on two occassions. Moreover, the sizes of the samples drawn from given strata are determined through minimization of the spectral radius of the vector of the sample mean under the fixed total cost of observations of variable values in the sample. This problem is considered in the case of estimation of the mean vector on the basis of a sample drawn on two occassions. It is shown that in the case of the proportional allocation of the sample in the strata the mean vector of the stratified sample is not less accurate than the mean vector of the simple sample.

The multidimensional auxiliary variables are used to stratify a population. Strata can be selected on the basis of the well known cluster method by Ward or the k-mean method. In the case of the regressiv superpopulation the strata are selected through minimization of the spectral radius of the variance-covariance matrix of auxiliary variables.

The properties of the mean vector from cluster sample are studied. Its variance-covariance matrix is expressed as a function of the introduced matrix of the coefficients of the intra-cluster correlation. It is proved that the vector of the cluster means is not a less accurate estimator of the vector of population averages than the vector of the simple sample means when the matrix of the coefficients of the intra-cluster correlation is defined as non-positive. The new method of dividing a fixed and finite population into groups of the same size on the basis of a multidimensional auxiliary variable is proposed. This method maximizes the intra-cluster scatter of the observations of a multidimensional auxiliary variable.

At the end the hypothesis concerning the equality of distributions of two dependent discrete variables is considered. The probability distribution of the test statistic is approximately of the chi-square distribution when the simple sample size is sufficiently large.

ISBN 83-04-04298-3