

ELEMENTY TEORII RANDOMIZACJI
I. PROBA ZRANDOMIZOWANA

RADOŚLAW KALA

Katedra Metod Matematycznych i Statystycznych
Akademii Rolniczej w Poznaniu

Streszczenie

Praca rozpoczyna cykl trzech prac poświęconych roli randomizacji w przeprowadzaniu eksperymentów badawczych. Szczególny nacisk położony jest na sformułowanie matematycznego modelu obserwacji eksperymentalnych. W artykule wprowadzone są niezbędne pojęcia podstawowe. Zbudowane są modele próby zrandomizowanej oraz próby prostej. Zbadane są pewne własności tych modeli.

1. WSTĘP

Wnioskowanie na podstawie eksperymentu jest zagadnieniem złożonym bo obejmującym trudne do pogodzenia, jeśli nie wręcz przeciwstawne, tendencje, z których główną jest chęć matematycznego, a więc abstrakcyjnego, opisania tego co realizuje się w świecie rzeczywistym. Złożoność tę pogłębia szereg uwarunkowań i ograniczeń wynikających bądź to z przedmiotu eksperymentu i jego zakresu, bądź to z indukcyjnej natury samego wnioskowania statystycznego.

Wśród wielu czynników mających wpływ na przebieg wnioskowania centralną rolę odgrywają procesy randomizacyjne, które, jak dowiadujemy się z pracy Stiglera (1978) (patrz także Folks, 1984), pojawiły się w praktyce eksperymentalnej już bardzo wcześnie, bo w roku 1884 w związku z doświadczeniem przeprowadzonym przez Peirce'a i Jastrowa, w którym badano próg percepcji sensorycznej. Wprowadzenie zasady randomizacji do teorii statystyki matematycznej nastąpiło znacznie później i było wynikiem prac prowadzonych w latach dwudziestych naszego stulecia przez pionierów doświadczalnictwa - R.A. Fishera i J. Neymana. W latach późniejszych

Słowa kluczowe: jednostki eksperymentalne, randomizacja jednostek, jednorodność próby

zagadnienia te były rozwijane przez wielu statystyków, wśród których wymienić można nazwiska Kempthorna, Wilka, Tukeya, Zyskinda, Scheffego, Yatesa, Neldera, Bailey. Mimo tak znacznego zainteresowania procesami randomizacyjnymi oraz prawie powszechnego ich stosowania we współczesnej praktyce eksperymentalnej, rola randomizacji w doświadczalnictwie ciągle jeszcze wzbudza żywe kontrowersje (patrz Thornett, 1982, a także Greenberg, 1951, Harville, 1975, Kempthorne, 1977).

W pracy niniejszej, rozpoczynającej cykl trzech prac poświęconych teorii randomizacji, usystematyzowano pojęcia podstawowe oraz przedstawiono własności prób zrandomizowanych. Podejście tu zaprezentowane, choć nawiązuje do dobrze znanych idei, szeroko dyskutowanych w literaturze statystycznej, odbiega jednak pod pewnymi względami od podejść tam prezentowanych. Dostarczyło ono systematycznego wyjaśnienia sytuacji, w których prowadzenie randomizacji jest uzasadnione, a w których zbędne.

2. JEDNOSTKI EKSPERYMENTALNE. POPULACJA PODSTAWOWA

Podstawą materialną dla przeprowadzenia eksperymentu są zwykle jednostki eksperymentalne. Mogą być nimi: maszyny, urządzenia, ich produkty, rośliny, poletka, zwierzęta, próbki materiałów itd. Każdą jednostkę charakteryzuje wiele cech, z których część można zmierzyć bezpośrednio, a określenie wartości innych wymaga zwykle mniej lub bardziej wyspecjalizowanej aparatury laboratoryjnej. Pomijając trudności z określeniem wartości poszczególnych cech, każdą jednostkę doświadczalną można w sensie matematycznym utożsamiać z punktem w przestrzeni euklidesowej odpowiednio dobranego wymiaru, przy czym kolejne cechy, opisujące ustaloną jednostkę, odpowiadają kolejnym współrzędnym punktu.

Wszystkie potencjalne jednostki mogące wziąć udział w ustalonym eksperymencie tworzą pewien zbiór, który określać będziemy mianem populacji podstawowej i oznaczać symbolem \mathcal{P} . Określenie tej zbiorowości jest w każdym eksperymencie sprawą bardzo istotną, gdyż stanowi ona podstawę odniesienia dla wszystkich wyprowadzanych z doświadczenia konkluzji, przy czym każde wzbogacenie tej zbiorowości powoduje rozszerzenie ogólności wyciąganych wniosków (porównaj Kempthorne, 1977, s.5).

Z uwagi na zróżnicowanie jednostek, wynikające z naturalnego zróżnicowania wartości cech je charakteryzujących, populacja podstawowa na ogół nie redukuje się w przedstawieniu matematyczno-geometrycznym do jednego punktu, lecz stanowi zbiór rozproszony. Rozproszenie to jest zazwyczaj znaczne w przypadku populacji przyrodniczych, gdzie zmienność jednostek regulowana zjawiskami naturalnymi jest duża, a bywa niewielkie w przypadku populacji technicznych, gdzie niekiedy można doprowadzić do pełnej unifikacji jednostek (porównaj Kempthorne, 1977, s.5).

Charakteryzując populację podstawową jednostek należy także zwrócić uwagę na jej liczebność. Nie ma ona większego znaczenia w przypadku, gdy wszystkie jej jednostki można uznać za identyczne. Na ogół jednak takie założenie nie może być przyjęte i wtedy w procesie modelowania eksperymentu liczebność populacji podstawowej musi być brana pod uwagę. Wyróżnia się tutaj dwie kategorie populacji. Pierwsza to tzw. populacje skończone, zawierające z góry określoną liczbę N jednostek. Populacje te oznaczamy będziemy symbolem \mathcal{P}_N . Druga kategoria, to populacje o nieskończonej liczbie jednostek lub wręcz populacje nieprzeliczalne. Do tej kategorii zaliczymy również populacje skończone lecz o tak dużej liczbie jednostek, że potraktowanie ich jako populacji nieskończonych nie będzie budziło istotnych zastrzeżeń. W tej kategorii znajdują się zatem populacje skończone, ale bardzo liczne, a także te, w których liczba jednostek nie jest dokładnie znana, a jedynie oszacowana od dołu znaczną liczbą naturalną, i wreszcie takie populacje, w których nie jest możliwe ustalenie zasady numerowania jednostek. Populacje te oznaczamy będziemy symbolem \mathcal{P}_∞ , traktując je równocześnie jako przypadek graniczny, przy $N \rightarrow \infty$, populacji \mathcal{P}_N .

Chcąc dokonać pełniejszej formalizacji poczynionych do tej pory uwag oznaczmy przez u_1, u_2, \dots jednostki populacji, a przez x q -wymiarowy wektor, którego kolejne składowe reprezentują kolejne wyróżnione cechy jednostek, w ogólnej liczbie q . Napiszemy ponadto $x_j = x(u_j)$ w celu utożsamienia wybranej, powiedzmy j -tej jednostki z punktem w q -wymiarowej przestrzeni euklidesowej R^q .

Przy tych oznaczeniach, populacji skończonej $\mathcal{P}_N = \{u_1, u_2, \dots, u_N\}$ odpowiada skończony zbiór punktów $\{x_j \in R^q, j = 1, 2, \dots, N\}$. W wypadku populacji nieskończonej \mathcal{P}_∞ zbiór odpowiadających jej punktów jest nieskończony lub wręcz nieprzeliczalny. Tym samym opis populacji \mathcal{P}_∞ nie może być tak prosty jak w wypadku populacji \mathcal{P}_N . Przy jego sformułowaniu używamy najczęściej środków probabilistycznych, które jednakże ujmują poszczególne jednostki raczej przeciętnie niż indywidualnie.

3. OBIEKTY DOŚWIADCZALNE

W odróżnieniu od badań, w których jedynym celem jest dostarczenie informacji opisowych o określonych populacjach naturalnych, w badaniach ściśle eksperymentalnych pojawiają się tzw. obiekty (zabiegi) doświadczalne, którymi mogą być porównywane technologie, procesy, warianty nawożeń, odmiany roślin uprawnych, terapie, diety pokarmowe, bądź kombinacje poziomów określonych czynników eksperymentalnych takich jak na przykład temperatura, ciśnienie, czas, wilgotność, dawki nawożeniowe itp. Tak ogólnie rozumiane obiekty charakteryzują się możliwością ich stosowania na jednostkach doświadczalnych. Zakłada się przy tym, że na każdej jednostce może być zastosowany dokładnie jeden

obiekt (porównaj Bailey, 1981, s.215) oraz że ten proces stosowania podlega pełnej kontroli ze strony eksperymentatora. W rezultacie działania zastosowanych obiektów doświadczalnych wartości poszczególnych cech opisujących jednostki ulegają określonym zmianom, które mogą okazać się zmianami pożądanymi lub niekorzystnymi. Modyfikacje te będziemy ogólnie określać mianem efektów obiektowych.

Wszystkie obiekty biorące udział w danym doświadczeniu, lub szerzej wszystkie te obiekty, o których eksperymentator zamierza się wypowiedzieć po przeprowadzeniu analizy wyników doświadczenia, tworzą pewien zbiór J . Najczęściej zbiór ten jest skończony, ale nic nie stoi na przeszkodzie aby, podobnie jak w przypadku populacji podstawowej P , dopuścić również sytuację, gdy jest on nieskończony.

Liczba jednostek eksperymentalnych przeznaczonych dla kolejnych obiektów biorących udział w doświadczeniu wynika z tzw. planu doświadczenia, na którego wybór decyduje się eksperymentator mając na względzie z jednej strony możliwość weryfikacji postawionej hipotezy badawczej oraz precyzję i efektywność doświadczenia, a z drugiej strony ograniczenia techniczne i ekonomiczne związane z prowadzeniem eksperymentu.

4. STRATEGIA EKSPERYMENTALNA

Ogólnie rzecz ujmując, zadaniem eksperymentatora jest wyprowadzenie wniosków o wzajemnych relacjach w zbiorze obiektów J odniesionych do ustalonej populacji podstawowej P . Zgodnie z zasadami wnioskowania indukcyjnego cel ten może być osiągnięty przez odpowiednie zaplanowanie i przeprowadzenie doświadczenia z ustalonymi obiektami ze zbioru J zastosowanymi na wybranym z populacji P skończonym zbiorze jednostek eksperymentalnych. Uzyskane z doświadczenia obserwacje, po poddaniu ich odpowiedniej analizie, stanowią podstawę dla sformułowania wniosków końcowych wyrażonych poprzez określenie wpływu różnych obiektów na zmiany cech jednostek populacji podstawowej. Elementem wiążącym wszystkie te etapy postępowania eksperymentalnego jest model probabilistyczny opisujący zachowanie się potencjalnych obserwacji. Model ten winien powstać w stadium planowania eksperymentu, przy czym jego struktura powinna być związana ściśle ze sposobem przeprowadzenia doświadczenia, tak aby mógł on stanowić teoretyczną podstawę dla analizy wyników uzyskanych z eksperymentu. Doświadczenie i związany z nim model obserwacji muszą stwarzać wystarczające podstawy dla uchwycenia i oddzielenia zmienności jednostek wynikającej z zastosowanych na nich różnych obiektów doświadczalnych od zmienności naturalnej samych jednostek. Interesująca zmienność związana z działaniem różnych obiektów dotyczy zatem jednostek, a dokładniej cech je charakteryzujących, i byłaby bez wątpienia łatwa do zaobserwowania, gdyby jednostki doświadczalne przed

przeprowadzeniem eksperymentu były identyczne. W tej sytuacji stwierdzenie jakiegokolwiek zmienności wśród jednostek, na których zastosowano różne obiekty, można byłoby w całości tłumaczyć różnicowaniem wpływem tych obiektów. To powszechne przekonanie całkowicie uzasadnia ogólną strategię eksperymentalną sprowadzającą się do dwóch kolejnych przeciwstawnych działań. Pierwsze z nich polega na początkowym wyrównaniu jednostek, na przykład na drodze odpowiedniej ich selekcji, natomiast drugie związane jest z celowym różnicowaniem jednostek przez poddanie ich działaniu różnych obiektów. Postępowanie takie, choć możliwe w doświadczeniach ściśle kontrolowanych, prowadzonych w skali laboratoryjnej, ma jednak ograniczone zastosowanie w powszechnej praktyce eksperymentalnej. Ograniczenie to wiąże się bądź z trudnościami technicznymi polegającymi na niemożliwości utworzenia dostatecznie licznej zbiorowości jednostek początkowo identycznych, bądź z założeniem metodologicznym związanym z chęcią odniesienia się we wnioskach końcowych do populacji podstawowej, w której jednostki z natury rzeczy są różnicowane (porównaj Kempthorne, 1977, s.5-6).

Problem początkowej unifikacji jednostek, odgrywający tak ważną rolę w strategii eksperymentalnej, może być jednak rozwiązany innymi środkami. Rozwiązanie to polega na wprowadzeniu takich mechanizmów losowania, które umożliwiają traktowanie jednostek populacji podstawowej \mathcal{P} tak, jakby były one w pewnym sensie identyczne.

W wypadku populacji nieskończonej \mathcal{P}_∞ , kiedy jest rzeczą oczywistą, że tylko część jednostek populacji może wziąć udział w eksperymencie, unifikację jednostek mamy w istocie zawartą w przyjmowanym a priori opisie probabilistycznym samej populacji. Mianowicie, każdej jednostce u pobranej z populacji \mathcal{P}_∞ w sposób losowy, niezależny i za pomocą jednakowego mechanizmu losowania przypisuje się identyczną pod względem rozkładu prawdopodobieństwa q -wymiarową zmienną losową $x = x(u)$, której kolejne składowe związane są z kolejnymi cechami wyróżnionymi na jednostkach. W dalszych rozważaniach przyjmujemy, że dla wektora losowego x istnieje wartość oczekiwana,

$$E(x) = \mu,$$

oraz macierz dyspersji,

$$D(x) = \Sigma,$$

przy czym ani q -wymiarowy wektor μ ani $q \times q$ -wymiarowa, dodatnio określona macierz Σ nie muszą być znane. Takie przedstawienie jednostek nazwiemy modelem, który zapiszemy w postaci uporządkowanej trójki $\{x, \mu, \Sigma\}$.

W przypadku populacji skończonej \mathcal{P}_N unifikację jednostek można uzyskać również na drodze probabilistycznej, która teraz sprowadza się do losowego ich indeksowania. Ponieważ postępowanie to nawiązuje do zasady randomizacji Fishera (1935), lecz dotyczy jednostek rozpatrywanych w oderwaniu od obiektów doświadczalnych, więc będziemy nazywać je randomizacją jednostek (patrz także Preece i in., 1978, s.114, Bailey, 1981, s.216 oraz Thornett, 1982, s.139).

5. RANDOMIZACJA JEDNOSTEK

Populacja skończona P_N obejmuje N indywidualnych jednostek u_1, u_2, \dots, u_N uporządkowanych w pewien początkowy sposób, wynikający np. z ich rozmieszczenia w przestrzeni lub czasie. Jak już to zostało powiedziane, jednostki te można utożsamiać z wektorami $x_1, x_2, \dots, x_N \in \mathbb{R}^q$ określającymi wartości ustalonego zespołu q cech dla kolejnych jednostek. Randomizacja jednostek polega na przyporządkowaniu im nowych indeksów w sposób przypadkowy i niezależny od aktualnych wartości opisujących je cech. Proces ten przedstawimy wprowadzając za Kempthornem (1952, s.137) zmienne losowe

$$\delta_{rj} = \delta_{rj}(l_1, l_2, \dots, l_N) = \begin{cases} 1, & \text{gdy } l_r = j, \\ 0, & \text{gdy } l_r \neq j, \end{cases} \quad \text{dla } r = 1, 2, \dots, N,$$

określone na zbiorze wszystkich permutacji (l_1, l_2, \dots, l_N) indeksów $(1, 2, \dots, N)$. Równość $\delta_{rj} = 1$ oznacza zatem, że w wyniku randomizacji indeks r jednostki, nadany jej w początkowym uporządkowaniu jednostek, zostaje zastąpiony indeksem j

Rozkład prawdopodobieństwa zmiennych losowych δ_{rj} jest uzależniony od miary prawdopodobieństwa P wprowadzonej na zbiorze permutacji. Przyjmując, że numeracja jednostek nie powinna prowadzić do jakiegokolwiek ich preferencji, przypiszemy każdej permutacji to samo prawdopodobieństwo $1/N!$. W rezultacie każda z N zmiennych losowych $\delta_{1j}, \delta_{2j}, \dots, \delta_{Nj}$ ma rozkład zdeterminowany równościami

$$P(\delta_{rj}=1) = \frac{(N-1)!}{N!} = \frac{1}{N}, \quad P(\delta_{rj}=0) = 1 - \frac{1}{N}. \quad (5.1)$$

Ponadto łatwo pokazać, że

$$P(\delta_{rj}=1, \delta_{r'j'}=1) = \begin{cases} \frac{1}{N(N-1)}, & \text{gdy } r \neq r', j \neq j', \\ 0, & \text{gdy } r \neq r', j = j' \text{ lub } r = r', j \neq j', \end{cases} \quad (5.2)$$

• także stwierdzić, że dla $j = 1, 2, \dots, N$

$$\sum_{r=1}^N \delta_{rj} = 1.$$

Niech teraz $\delta_j \in \mathbb{R}^N$ będzie wektorem losowym zdefiniowanym równością

$$\delta_j = (\delta_{1j}, \delta_{2j}, \dots, \delta_{Nj})'.$$

W świetle (5.1) oraz (5.2)

$$E(\delta_j) = \frac{1}{N} \mathbf{1}_N, \quad E(\delta_j \delta_j') = \begin{cases} \frac{1}{N} \mathbf{I}_N, & j=j', \\ \frac{-1}{N(N-1)} (\mathbf{I}_N - \mathbf{J}_N), & j \neq j', \end{cases} \quad (5.3)$$

gdzie $\mathbf{1}_N$ oznacza N -wymiarowy wektor o każdej składowej równej 1, \mathbf{I}_N oznacza macierz jednostkową stopnia N , a $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}'_N$. Stąd macierz dyspersji wektora δ_j i macierz kowariancji wektorów δ_j i $\delta_{j'}$ wyrażają się odpowiednio wzorami

$$D(\delta_j) = \frac{1}{N} (\mathbf{I}_N - \frac{1}{N} \mathbf{J}_N) \quad (5.4)$$

oraz

$$C(\delta_j, \delta_{j'}) = \frac{-1}{N(N-1)} (\mathbf{I}_N - \frac{1}{N} \mathbf{J}_N), \quad j \neq j'. \quad (5.5)$$

Wprowadzone zmienne losowe odniesiemy teraz do jednostek populacji \mathcal{P}_N reprezentowanych przed randomizacją wektorami $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$. Mianowicie, jednostką j -tą uzyskaną w wyniku procesu randomizacji jest

$$u_{(j)} = \sum_{r=1}^N \delta_{rj} u_r,$$

lub, przy wektorowej reprezentacji jednostek, wektor

$$\mathbf{x}_{(j)} = \mathbf{X}' \delta_j = \sum_{r=1}^N \delta_{rj} \mathbf{x}_r, \quad (5.6)$$

gdzie $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)'$ jest $N \times q$ -wymiarową macierzą, której wiersze utworzone są z wektorów charakteryzujących poszczególne jednostki populacji \mathcal{P}_N .

Wobec obecności w równości (5.6) zmiennych losowych δ_{rj} , wektory $\mathbf{x}_{(j)}$, $j = 1, 2, \dots, N$, są teraz q -wymiarowymi zmiennymi losowymi. W świetle (5.3) i (5.4), każda z nich ma wartość oczekiwana

$$E(\mathbf{x}_{(j)}) = \bar{\mathbf{x}}, \quad (5.7)$$

gdzie

$$\bar{\mathbf{x}} = \frac{1}{N} \mathbf{X}' \mathbf{1}_N = \frac{1}{N} \sum_{r=1}^N \mathbf{x}_r \quad (5.8)$$

jest wektorem średnich arytmetycznych, oraz macierz dyspersji

$$D(\mathbf{x}_{(j)}) = \mathbf{S}, \quad (5.9)$$

gdzie

$$N\mathbf{S} = \mathbf{X}' (\mathbf{I}_N - \frac{1}{N} \mathbf{J}_N) \mathbf{X} = \sum_{r=1}^N \mathbf{x}_r \mathbf{x}'_r - N \bar{\mathbf{x}} \bar{\mathbf{x}}', \quad (5.10)$$

jest tzw. poprawioną macierzą sum kwadratów i iloczynów. Zauważmy jeszcze, że na mocy (5.5) macierz kowariancji dwóch jednostek zrandomizowanych ma postać

$$C(\mathbf{x}_{(j)}, \mathbf{x}_{(j')}) = \frac{-1}{N-1} \mathbf{S}, \quad j \neq j'. \quad (5.11)$$

W rezultacie jednostki populacji skończonej \mathcal{P}_N po randomizacji mogą być przedstawione analogicznie jak jednostki populacji nieskończonej \mathcal{P}_∞ , tj. jako zmienne losowe o identycznych rozkładach z parametrami (ściślej - momentami rzędu pierwszego i drugiego) określonymi w (5.7) i

(5.9). Odpowiedni model i -tej zrandomizowanej jednostki z populacji \mathcal{P}_N możemy zapisać w postaci $\{x_{(j)}, \bar{x}, S\}$. Od modelu jednostki pochodzącej z populacji nieskończonej \mathcal{P}_∞ różni się on parametrami, a także, wyrażonym za pomocą wzoru (5.11) skorelowaniem zmiennych losowych $x_{(j)}$, $j = 1, 2, \dots, N$.

Pełniejsza zgodność przedstawionych modeli ma miejsce, gdy liczebność N populacji \mathcal{P}_N można uznać za wielkość wzrastającą nieograniczenie. Jeśli wtedy przyjąć dodatkowe założenie, że wektor średnich \bar{x} , określony wzorem (5.8), oraz macierz S , określona wzorem (5.10), mają, przy $N \rightarrow \infty$, skończone granice, wynoszące odpowiednio μ oraz Σ , to granicą macierzy kowariancji (5.11) jest macierz zerowa, zmienne $x_{(j)}$, $j = 1, 2, \dots, N$, stają się nieskorelowane i tym samym model jednostki zrandomizowanej pochodzącej z \mathcal{P}_N utożsamia się z modelem jednostki wylosowanej z \mathcal{P}_∞ . W tym sensie model jednostek z populacji \mathcal{P}_∞ można traktować jako model graniczny dla jednostek z populacji \mathcal{P}_N .

W przedstawionych opisach warto jeszcze umiejscowić wypadek, gdy populacja składa się z jednostek identycznych. Łatwo mianowicie spostrzec, że populacje takie będą scharakteryzowane odpowiednio równościami $S = O$, lub $\Sigma = O$, co w obu wypadkach prowadzi do degeneracji odpowiedniej zmiennej losowej, która traci atrybut losowości, stając się wielkością stałą.

6. POBIERANIE PRÓBY. MODEL PRÓBY

Kolejnym elementem, na który należy zwrócić uwagę, jest zagadnienie związane z indukcyjnym charakterem wnioskowania statystycznego, a polegające na wyborze z populacji podstawowej \mathcal{P} skończonego podzbioru n jednostek ją reprezentujących. Podzbiór ten oznaczymy symbolem \mathcal{P}_n i nazwiemy próbą.

Ograniczenie się w konkretnym doświadczeniu do próby jest oczywiste w wypadku populacji nieskończonych \mathcal{P}_∞ , natomiast w wypadku populacji skończonych, szczególnie niezbyt licznych, może okazać się zbyteczne. Pozostaniemy jednak przy takim podejściu jako bardziej uniwersalnym, obejmującym w szczególności sytuację, gdy $n = N$.

W celu zapewnienia reprezentatywności próby pochodzącej z populacji nieskończonej \mathcal{P}_∞ jej wybór winien odbyć się zgodnie z zasadą losowości. W rezultacie próba \mathcal{P}_n w opisie probabilistycznym jest przedstawiona przez układ n niezależnych zmiennych losowych x_1, x_2, \dots, x_n , każda o tym samym rozkładzie prawdopodobieństwa z wartością oczekiwaną μ i macierzą dyspersji Σ . W notacji wektorowej próbę \mathcal{P}_n z populacji \mathcal{P}_∞ można przedstawić za pomocą nxq -wymiarowej macierzy losowej

$$X = (x_1, x_2, \dots, x_n)' \quad (6.1)$$

o wartości oczekiwanej

$$E(X) = \mathbf{1}_n \mu' \quad (6.2)$$

oraz macierzy dyspersji

$$D(X) = D(csX) = \Sigma \otimes I_n, \quad (6.3)$$

gdzie csX jest nq -wymiarowym wektorem utworzonym przez ułożenie kolejnych kolumn macierzy X jedna pod drugą, a symbol \otimes oznacza iloczyn Kroneckera macierzy. Model próby \mathcal{P}_n pobranej z populacji nieskończonej \mathcal{P}_∞ możemy w rezultacie zapisać w postaci

$$\{X, \mathbf{1}_n \mu', \Sigma \otimes I_n\}. \quad (6.4)$$

W wypadku populacji skończonej \mathcal{P}_n , reprezentatywną próbę można uzyskać kierując się również zasadą losowości, która w zastosowaniu do populacji \mathcal{P}_n , poddanej uprzednio procesowi randomizacji, sprowadza się do wyboru dowolnych n jednostek zrandomizowanych, pod warunkiem jednakże, że wybór ten jest niezależny od uporządkowania jednostek przed randomizacją. Praktycznie oznacza to, że po wykonaniu randomizacji każdy niezależny wybór n jednostek, a więc również wybór najwygodniejszy pierwszych n jednostek według numeracji ustalonej w procesie randomizacji, prowadzi do próby reprezentatywnej \mathcal{P}_n . Komplet jednostek takiej próby można przedstawić przy użyciu nxq -wymiarowej macierzy losowej

$$X_{(,)} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})', \quad (6.5)$$

o wartości oczekiwanej

$$E(X_{(,)}) = \mathbf{1}_n \bar{x}', \quad (6.6)$$

oraz o macierzy dyspersji

$$D(X_{(,)}) = S \otimes \frac{N}{N-1} (I_n - \frac{1}{N} J_n). \quad (6.7)$$

Model próby \mathcal{P}_n pobranej z populacji skończonej \mathcal{P}_n można zatem zapisać w postaci

$$\{X_{(,)}, \mathbf{1}_n \bar{x}', S \otimes \frac{N}{N-1} (I_n - \frac{1}{N} J_n)\}. \quad (6.8)$$

W dalszym ciągu zbiory \mathcal{P}_n uzyskane opisanymi wyżej sposobami nazywać będziemy odpowiednio próbą prostą lub próbą zrandomizowaną.

Warto na koniec podkreślić, że w opisanych sposobach tworzenia zbioru \mathcal{P}_n ma miejsce pełna swoboda, jeśli chodzi o ustalenie jego liczebności. W wypadku pobierania próby z populacji \mathcal{P}_∞ wystarczy założenie, że kolejna jednostka będzie wybrana z zachowaniem zasady losowości, a więc będzie mogła być reprezentowana za pomocą zmiennej losowej niezależnej od zmiennych losowych reprezentujących jednostki już wybrane, a ponadto jej rozkład prawdopodobieństwa będzie identyczny z rozkładami tych

wcześniejszych zmiennych. Spełnienie opisanych warunków będzie miało miejsce, jeśli kolejno pobrana jednostka nie naruszy rozkładu prawdopodobieństwa w populacji podstawowej oraz gdy jej losowanie będzie dokonane niezależnie od rezultatów wcześniejszych losowań.

Zwiększanie liczebności próby \mathcal{P}_n pochodzącej z populacji skończonej \mathcal{P}_N jest znacznie prostsze. Kolejną jednostką może tu być mianowicie jednostka z kolejnym numerem ustalonym w procesie randomizacji jednostek populacji \mathcal{P}_N .

Warto wreszcie zaznaczyć, że modele prób o różnych liczebnościach pozostaną niezmiennione z dokładnością do wymiaru n odpowiednich wektorów i macierzy występujących w modelu.

7. JEDNORODNOŚĆ PRÓBY

W przedstawieniu modelowym jednostki eksperymentalne stanowiące próbę prostą lub próbę zrandomizowaną można uznać za probabilistycznie identyczne, gdyż są one opisane zmiennymi losowymi o identycznych rozkładach prawdopodobieństwa. Tę własność będziemy nazywać jednorodnością lub symetrią próby. Własność ta nawiązuje do tzw. wymiennalności zmiennych losowych (porównaj Lindley, 1977, s.68), które to określenie celnie odzwierciedla możliwość wymiennego traktowania jednostek próby, widzianych w pryzmacie modelu. Precyzyjne określenie wymiennalności związane jest z symetrią łącznego rozkładu prawdopodobieństwa zmiennych losowych stanowiących próbę, przy czym symetria rozumiana jest jako niezmienniczość rozkładu łącznego względem grupy wszystkich permutacji. Ponieważ rozważane przez nas modele specyfikują jedynie momenty pierwszego i drugiego rzędu, więc w dalszym ciągu mówiąc o symetrii rozkładu będziemy odnosić ją do wartości oczekiwanej i macierzy dyspersji. Własność tę charakteryzuje następujące twierdzenie będące wielowymiarowym uogólnieniem spostrzeżeń Davida (1977, s.292).

Twierdzenie 1. Niech $n \times q$ -wymiarowa macierz losowa Z ma wartość oczekiwaną $E(Z)$ i macierz dyspersji $D(Z)$. Jeśli rozkład macierzy Z jest niezmienniczy względem grupy wszystkich permutacji jej wierszy, to istnieje q -wymiarowy wektor \mathbf{m} oraz macierze \mathbf{V} i \mathbf{W} stopnia q takie, że

$$E(Z) = \mathbf{1}_n \mathbf{m}', \quad (7.1)$$

$$D(Z) = \mathbf{W} \otimes \mathbf{I}_n + \mathbf{V} \otimes \mathbf{J}_n. \quad (7.2)$$

Odwrotnie, jeżeli spełnione są równości (7.1) i (7.2), to wartość oczekiwana i macierz dyspersji macierzy Z są niezmiennicze względem grupy wszystkich permutacji.

Dowód. Niech $Z = (z_1, z_2, \dots, z_n)'$ oraz niech $E(z_i) = m_i$, $D(z_i) = M_i$, $C(z_i, z_{i'}) = V_{ii'}$ dla $i, i' = 1, 2, \dots, n$, $i \neq i'$. Wtedy wobec niezmienniczości

niczości rozkładu mamy

$$m_i = E(z_i) = E(z_j) = m_j,$$

$$M_i = D(z_i) = D(z_j) = M_j,$$

dla $i = 1, 2, \dots, n$ oraz

$$V_{i_1 i_2} = C(z_i, z_{i_1}) = C(z_j, z_{i_1}) = V_{j_1 i_2},$$

dla dowolnego $i \cdot = 1, 2, \dots, n$, $i \cdot \neq i$, $i \cdot \neq j$. Wobec dowolności wskaźników i , $i \cdot$ oraz j mamy zatem

$$E(z_i) = m, \quad D(z_i) = M \quad \text{dla } i = 1, 2, \dots, n$$

oraz

$$C(z_i, z_{i \cdot}) = V \quad \text{dla } i, i \cdot = 1, 2, \dots, n, i \neq i \cdot.$$

Stąd równość (7.1) wynika bezpośrednio, a równość (7.2) po położeniu zamiast różnicy $M - V$ macierzy W . Dla dowodu pozostałej części twierdzenia wystarczy zauważyć, że każda macierz P realizująca permutację wierszy macierzy Z spełnia relację $P1_n = 1_n$ oraz $PP' = I_n$. \square

Z przedstawionego twierdzenia w szczególności wynika, że model próby prostej oraz model próby zrandomizowanej są niezmiennicze względem grupy wszystkich permutacji jednostek, co pozwala w rezultacie traktować jednostki tych prób równoważnie. Wniosek taki łatwo ustalić zauważając, że model (6.4), próby prostej, i model (6.8), próby zrandomizowanej, należą do klasy modeli wyznaczonych równościami (7.1) i (7.2).

Warto tu przypomnieć, że jednorodność próby prostej jest implikowana przyjętymi a priori założeniami o populacji P_∞ , natomiast jednorodność próby zrandomizowanej jest wymuszona procesem randomizacji opartym o jednostajny rozkład prawdopodobieństwa nad zbiorem wszystkich permutacji. Fakt ten jest w pełni zgodny z ogólniejszym rezultatem ustalonym przez Thorntona (1982, s.141) w jego twierdzeniu 4, z którego wynika, że jeśli tylko randomizacja jest prowadzona niezależnie od rozkładu zmiennej losowej $x = x(u)$ opisującej jednostki eksperymentalne, to łączny rozkład jednostek zrandomizowanych jest symetryczny bez względu na początkową postać rozkładu zmiennej x .

8. WIELOKROTNE POBIERANIE PRÓBY

Udowodnimy tutaj dwa twierdzenia pozwalające określić postać modelu próby uzyskanej w wyniku wielokrotnego jej pobierania.

Twierdzenie 2. Jeżeli P_n jest n elementową q -wymiarową próbą prostą z populacji nieskończonej P_∞ , której jednostki podlegają modelowi $\{x, \mu, \Sigma\}$, to dowolna niezależnie pobrana m elementowa próba zrandomizowana ze zbioru

\mathcal{P}_n , traktowanego jako n elementowa populacja skończona, ma model postaci

$$\{X_{()}, 1_n \mu', \Sigma \otimes I_n\}. \quad (8.1)$$

Dowód. Zgodnie z założeniem, model dla zbioru \mathcal{P}_n , traktowanego jako próba prosta z populacji \mathcal{P}_∞ , możemy zapisać w postaci

$$\{X_n, 1_n \mu', \Sigma \otimes I_n\}, \quad (8.2)$$

gdzie $X_n = (x_1, x_2, \dots, x_n)'$ jest macierzą losową reprezentującą jednostki próby prostej. Z kolei macierz $X_{()}$, opisująca jednostki próby zrandomizowanej \mathcal{P}_n pobranej z \mathcal{P}_n , może być przedstawiona w postaci

$$X_{()} = \Delta' X_n, \quad (8.3)$$

gdzie $\Delta = (\delta_1, \delta_2, \dots, \delta_m)$ jest $n \times m$ -wymiarową macierzą losową, której elementy są zmiennymi losowymi opisującymi proces randomizacji zgodnie z uwagami paragrafu 5. Zauważmy ponadto, że macierz losowa Δ jest niezależna od macierzy losowej X_n oraz że z własności wektorów losowych $\delta_1, \delta_2, \dots, \delta_m$ wynikają dwa związki niestochastyczne

$$\Delta' 1_n = 1_m \quad \text{oraz} \quad \Delta' \Delta = I_m. \quad (8.4)$$

Biorąc teraz pod uwagę równość (8.3), model (8.2), pierwszą z równości (5.3) oraz korzystając z niezależności macierzy losowych Δ i X_n , mamy

$$E(X_{()}) = E(\Delta') E(X_n) = \frac{1}{n} 1_m 1_n' 1_n \mu' = 1_m \mu'. \quad (8.5)$$

W celu wyznaczenia macierzy dyspersji $D(X_{()}) = D(csX_{()})$ określmy na początek wartość oczekiwaną

$$E\{csX_{()}(csX_{()})'\} = E\{(I_m \otimes \Delta') csX_n (csX_n)' (I_m \otimes \Delta)\},$$

która za pośrednictwem warunkowych wartości oczekiwanych oraz równości (8.4), może być przekształcona do postaci

$$\begin{aligned} E\{csX_{()}(csX_{()})'\} &= E_{\Delta}\{(I_m \otimes \Delta') E_X[csX_n (csX_n)'] (I_m \otimes \Delta)\} \\ &= E_{\Delta}\{(I_m \otimes \Delta') [D(X_n) + csEX_n (csEX_n)'] (I_m \otimes \Delta)\} \\ &= E_{\Delta}\{(I_m \otimes \Delta') [(\Sigma \otimes I_n) + (\mu \otimes 1_n)(\mu' \otimes 1_n')] (I_m \otimes \Delta)\} \\ &= E_{\Delta}\{(\Sigma \otimes \Delta' \Delta) + (\mu \mu' \otimes \Delta' 1_n 1_n' \Delta)\} \\ &= \Sigma \otimes I_m + \mu \mu' \otimes 1_n 1_n'. \end{aligned}$$

Stąd oraz w oparciu o relację

$$D(X_{()}) = E\{csX_{()}(csX_{()})'\} - (csEX_{()})(csEX_{()})'$$

i równość (8.5) mamy ostatecznie $D(X_{()}) = \Sigma \otimes I_m$.

Twierdzenie 3. Jeżeli \mathcal{P}_n jest n elementową q -wymiarową próbą zrandomizowaną pochodzącą z populacji skończonej \mathcal{P}_M , to dowolna, niezależnie pobrana, m elementowa próba zrandomizowana \mathcal{P}_m ze zbioru \mathcal{P}_n , traktowanego jako n elementowa populacja skończona, ma model postaci

$$\{X_{(c)}, \mathbf{1}_m \bar{x}_M', S_M \otimes \frac{M}{M-1} (I_m - \frac{1}{M} J_m)\},$$

gdzie

$$\bar{x}_M = \frac{1}{M} \sum_{i=1}^M x_i$$

oraz

$$S_M = \frac{1}{M} (\sum_{i=1}^M x_i x_i' - M \bar{x}_M \bar{x}_M').$$

Dowód. Uzasadnienie tego twierdzenia przebiega analogicznie do dowodu twierdzenia 2. Zauważmy na początek, że zgodnie z (6.8) model n elementowej q -wymiarowej próby zrandomizowanej z populacji \mathcal{P}_M ma postać

$$\{X_{(n)}, \mathbf{1}_n \bar{x}_M', S_M \otimes \frac{M}{M-1} (I_n - \frac{1}{M} J_n)\},$$

gdzie $X_{(n)} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})'$ jest macierzą losową reprezentującą jednostki wybrane z populacji \mathcal{P}_M do próby \mathcal{P}_n . Ponadto macierz reprezentująca jednostki z próby zrandomizowanej \mathcal{P}_m pobranej z ze zbioru \mathcal{P}_n można wyrazić podobnie jak w (8.3), tj. w postaci

$$X_{(c)} = \Delta' X_{(n)}.$$

Korzystając teraz z analogicznych argumentów jak te przytoczone w dowodzie twierdzenia 2, mamy

$$E(X_{(c)}) = E(\Delta') E(X_{(n)}) = \frac{1}{n} \mathbf{1}_m \mathbf{1}_n' \mathbf{1}_n \bar{x}_M' = \mathbf{1}_m \bar{x}_M'$$

oraz

$$\begin{aligned} E\{csX_{(c)}(csX_{(c)})'\} &= E_{\Delta}\{(I_m \otimes \Delta') E_{X_{(n)}}[csX_{(n)}(csX_{(n)})'] (I_m \otimes \Delta)\} \\ &= E_{\Delta}\{(I_m \otimes \Delta') [S_M \otimes \frac{M}{M-1} (I_n - \frac{1}{M} J_n) + (\bar{x}_M \bar{x}_M' \otimes J_n)] (I_m \otimes \Delta)\} \\ &= S_M \otimes \frac{M}{M-1} (I_m - \frac{1}{M} J_m) + (\bar{x}_M \bar{x}_M' \otimes J_m), \end{aligned}$$

co kończy dowód. □

Z przedstawionych twierdzeń wynikają praktyczne wnioski dla eksperymentatora. Pierwszy z nich można sformułować w oparciu o twierdzenie 2, które w istocie orzeka, że prowadzenie randomizacji na jednostkach stanowiących próbę prostą z populacji nieskończonej \mathcal{P}_{∞} jest niecelowe, gdyż nie zmienia w żadnym stopniu własności jednorodności czy wymienialności początkowo zawartej w modelu populacji \mathcal{P}_{∞} . Jednocześnie można stwierdzić, że działanie takie jest dopuszczalne, bo wspomniana własność zachowuje, co jest łatwo widoczne z porównania modelu (8.2),

dotyczącego próby prostej \mathcal{P}_n , z modelem (8.1), próby zrandomizowanej \mathcal{P}_m pochodzącej z \mathcal{P}_n .

Umiejscawiając z kolei wymienione tu modele w klasie modeli przedstawionej w twierdzeniu 1, łatwo zauważyć dodatkową ich własność. Wyraża się ona brakiem skorelowania poszczególnych jednostek eksperymentalnych. W odniesieniu do modelu (8.2) własność ta wynika bezpośrednio z przyjętego założenia o niezależności zmiennych losowych opisujących te jednostki. Natomiast fakt przeniesienia się tej własności z próby prostej \mathcal{P}_n na próbę zrandomizowaną \mathcal{P}_m wynika z twierdzenia 2. Rezultat ten można również odczytać z wyniku Thornetta (1982, s.139) sformułowanego w jego twierdzeniu 1, które orzeka w szczególności, że jeżeli łączny rozkład wektorowej zmiennej losowej x jest niezmienniczy względem grupy wszystkich permutacji i jeżeli randomizacja składowych wektora x jest niezależna od niego samego, to wektor x po zrandomizowaniu jego składowych ma taki sam rozkład jak przed randomizacją.

Drugi wniosek wypływa z twierdzenia 3, które praktycznie oznacza, że jednokrotne pobranie próby zrandomizowanej generuje model jednostek eksperymentalnych, który, przy powtórzeniu procesu pobrania próby zrandomizowanej, nie ulega już dalszym zmianom, być może z wyjątkiem zmienionej liczby jednostek. W rezultacie prowadzenie wielokrotnej randomizacji w celu sformułowania próby można uznać za zbyteczne.

Również i ten rezultat można wyprowadzić ze wspomnianego twierdzenia Thornetta (1982). W tym celu wystarczy przypomnieć, że próbę zrandomizowaną \mathcal{P}_n cechuje jednorodność, a więc ponowne zastosowanie niezależnej randomizacji pozostawi tę cechę rozkładu łącznego bez zmiany.

Opisane tu własności nawiązują także do ogólniejszej konkluzji Thornetta (1982), w której autor stwierdza, że rola randomizacji jest dość ograniczona, bo jedynie zachowuje własność niezmienniczości początkowo zawartą w modelu populacji. Ta uwaga, uszczuplająca znaczenie randomizacji, stosuje się w pełni przy pobieraniu próby prostej z populacji nieskończonej (twierdzenie 2), lub próby zrandomizowanej z populacji skończonej, w której jednakże wstępnie dokonano randomizacji (twierdzenie 3). W odniesieniu natomiast do niezrandomizowanej populacji skończonej \mathcal{P}_n , a więc populacji traktowanej jako kolekcja jednostek niewymienialnych, indywidualnie zróżnicowanych, utożsamianych z układem ustalonych wektorów x_1, x_2, \dots, x_n , należy pamiętać o kreatywnej roli randomizacji, za pomocą której pożądaną własność niezmienniczości rozkładu względem grupy permutacji można do modelu wprowadzić.

LITERATURA

- Bailey, R.A. (1981). A unified approach to design of experiments. *J. R. Statist. Soc. A*, 144, 214-223.
- David, A.P. (1977). Invariant distributions and analysis of variance models. *Biometrika*, 64, 291-297.

- Folks, J.L. (1984). Use of randomization in experimental research. In: *Experimental Design, Statistical Models, and Genetic Statistics. Essays in Honor of O. Kempthorne* (Ed. K. Hinkelmann). Marcel Dekker, INC, New York.
- Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- Greenberg, B.G. (1951). Why randomize? *Biometrics* 7, 309-322.
- Harville, D.A. (1975). Experimental randomization: Who needs it? *The Amer. Statist.* 29, 27-31.
- Kempthorne, O. (1977). Why randomize? *J. Statist. Plan. and Infer.* 1, 1-25.
- Kempthorne, O. (1952). *The Design and Analysis of Experiments*. J. Wiley, New York.
- Lindley, D.V. (1977). Contribution to the discussion on J. A. Nelder's paper "A reformulation of linear models". *J. R. Statist. Soc A*, 140, 68-69.
- Preece, D.A., Bailey, R.A., Patterson, H.D. (1978). A randomization problem in forming designs with superimposed treatments. *Australian J. Statist.* 20, 111-125.
- Thornett, M.L. (1982). The role of randomization in model-based inference. *Australian J. Statist.* 24, 137-145.
- Stigler, S.M. (1978). Mathematical statistics in the early states. *Ann. Statist.* 6, 239-265.

Praca wpłynęła 1 sierpnia 1988:
w wersji ostatecznej 10 marca 1989

ELEMENTS OF THE RANDOMIZATION THEORY.
I. A RANDOMIZED SAMPLE

Summary

The paper begins a series of three papers discussing the role of randomization in conducting research experiments. The special emphasis is laid on formulation of the mathematical models for experimental observations. In this paper relevant basic notions are introduced, the models of a randomized sample and a simple sample are constructed and their properties are examined.