Anna B a r t k o w i a k o w a (Wrocław)

SCHEMATY LOSOWANIA NIEOGRANICZONEGO, WARSTWOWEGO I GRUPOWEGO

1. Ogólne podstawy metod reprezentacyjnych

Przypuśćmy, że interesuje nas populacja charakteryzująca się pewnymi cechami, które chcielibyśmy bliżej poznać. Z pewnych względów nie możemy jednak zbadać całej populacji, ale musimy ograniczyć się do zbadania tylko pewnej części populacji, tak zwanej próbki.

Metody reprezentacyjne zajmują się dwoma zasadniczymi zagadnieniami: 1° W jaki sposób pobierać próbkę z populacji? 2° W jaki sposób wnioskować o odpowiednich własnościach populacji na podstawie informacji zebranych z próbki?

W dalszym ciągu będziemy rozważać te zagadnienia przy założeniu, że liczebność rozważanej populacji jest skończona. W populacji tej będą nas interesować takie wskaźniki statystyczne jak średnia arytmetyczna wybranej cechy ilościowej lub frakcja elementów populacji posiadających interesującą nas własność w przypadku gdy badana cecha jest cechą jakościową.

Omówimy trzy dość powszechnie stosowane schematy pobierania próbki: 1° losowanie nieograniczone z całej populacji, kiedy to wybiera się do próbki indywidua losując je w pewien określony sposób bezpośrednio z całej populacji; 2° losowanie warstwowe, w którym przed pobraniem próbki rozdziela się populację na tak zwane warstwy i następnie pobiera się próbkę losując indywidua z tych warstw; 3° losowanie grupowe, w którym korzysta się z naturalnego ugrupowania populacji w pewne zespoły (np. młodzież szkolna jest ugrupowana w zespoły - klasy, młodzież studencka w grupy studenckie). Przy losowaniu grupowym losuje się pewną liczbę zespołów i te bada się w całości.

Dla porównania poszczególnych metod należy zdefiniować jakieś kryteria dobroci, które dawałyby wskazówki co do tego, kiedy można się spodziewać, że jakiś schemat jest korzystniejszy od innych. Zagadnienie to można rozpatrywać w następujący sposób:

Przypuśćmy, że na podstawie próbki chcemy oszacować pewien parametr Q charakteryzujący rozkład badanej cechy w interesującej nas populacji. Parametr ten chcemy oszacować za pomocą funkcji $t = t(x_1, \ldots, x_n)$. Podstawiając na miejsce argumentów tej funkcji wartości $x_1, x_2, \ldots, x_n$ zaobserwowane dla wylosowanych elementów otrzymujemy wartość t odpowiadającą wylosowanej próbce.

Ponieważ wielkości $x_1, \ldots, x_n$ obserwowane w próbce są zmiennymi losowymi, więc i estymator $t = t(x_1, \ldots, x_n)$ rozpatrywany jako funkcja zmiennych losowych jest również zmienną losową i wobec tego ma swój rozkład prawdopodobieństwa. Jeśli tak, to możemy obliczać jego wartość oczekiwaną i wariancję.

Planując jakieś badanie statystyczne należy przede wszystkim określić estymatory, przy pomocy których zamierzamy szacować nieznane parametry populacji. Jest zrozumiałe, że jedne estymatory mogą się nadawać do tego lepiej, inne gorzej.

Następnym zagadnieniem jest ustalenie schematu losowania. Można się zgodzić, że schemat losowania jest tym lepszy, im mniejsza jest wariancja estymatora obliczonego przy tym schemacie losowania.

W praktyce jednak trzeba wziąć pod uwagę również rozmaite inne czynniki. W zależności od posiadanych informacji o badanej populacji, od warunków organizacyjno-technicznych, od kosztów badania eliminujemy z góry pewne schematy i z pozostałych wybieramy najlepszy.

Przy różnych schematach stosuje się w rozmaity sposób "losowanie" elementów z populacji. "Losowanie" takie przeprowadza się za pomocą liczb losowych. Tablice takich liczb można znaleźć np. w [8] lub [9]. Przy schemacie losowania nieograniczonego, kiedy to losujemy elementy do próbki bezpośrednio z całej populacji, najbardziej zalecany sposób polega na tym, że wszystkie elementy badanej populacji numerujemy, a następnie za pomocą tablic liczb losowych losujemy próbkę n-elementową, wybierając z populacji elementy o numerach wskazanych przez tablice liczb losowych. Przy losowaniu warstwowym, gdy populacja jest podzielona na warstwy, stosujemy to samo postępowanie do każdej warstwy.

Czasem rozgraniczenie populacji na warstwy jest trudne, np.

gdy badaną populację stanowią wyroby jakiejś fabryki w dłuższym okresie czasu, przy czym skądinąd wiadomo, że jakość tych wyrobów zmieniała się z czasem. Wtenczas można stosować liczby złote lub żelazne wprowadzone przez H. Steinhausa (patrz: [6], [7]). Łączą one postulat losowego wyboru z postulatem równomiernego rozstawienia wylosowanych elementów w populacji.

2. Losowanie nieograniczone

Przez losowanie nieograniczone rozumiemy następujące postępowanie: Jeśli badana populacja zawiera N elementów, to elementy te numerujemy kolejnymi liczbami naturalnymi od 1 aż do N. Ustalamy liczebność próbki n. Następnie za pomocą odpowiednio dobranej tablicy liczb losowych wyznaczamy do próbki elementy populacji odpowiadające kolejnym odczytywanym liczbom losowym tak długo, aż uzyskamy żądaną liczebność próbki n.

§ 2.1. Losowanie niezależne i zależne.

Jeśli losowanie nieograniczone wykonujemy przy użyciu tak zwanych tablic liczb losowych (a nie tablic liczb przetasowanych, złotych lub żelaznych), to może się zdarzyć, że niektóre elementy populacji wejdą do próbki więcej niż jeden raz. W związku z tym powstają dwie możliwości: 1° pozostawiamy próbkę taką, jaką wylosowaliśmy, czyli dopuszczamy możliwość powtarzania się niektórych elementów; 2° powtarzające się elementy próbki wykreślamy, a na ich miejsce dołączamy elementy odpowiadające następnym liczbom losowym tak długo, aż liczba różnych elementów w próbce będzie równa żądanej liczebności n.

Opiszemy teraz modele probabilistyczne odpowiadające wymienionym możliwościom 1° i 2°.

Ad 1°. W urnie mamy N kul (populacja). Z urny tej wyciągamy n razy po jednej kuli. Po każdym ciągnieniu wylosowaną kulę wrzucamy z powrotem do urny i mieszamy kule w urnie tak, aby prawdopodobieństwo wyciągnięcia dowolnej kuli w następnym losowaniu było dla każdej kuli takie samo jak na początku losowania. Przy takim sposobie losowania wylosowane w kolejnych ciągnieniach kule przedstawiają zmienne losowe niezależne, a odpowiadający mu schemat losowania nazywa się schematem losowania niezależnego lub losowania ze zwracaniem.

Ad. 2°. Z urny, zawierającej N kul, wyciągamy n razy po jed-

nej kuli nie wkładając ich z powrotem po wylosowaniu. Przy takim postępowaniu mamy gwarancję, że wyciągniemy różne kule, ale prawdopodobieństwa wylosowania dla każdej kuli będą różne w różnych ciągnieniach i będą zależeć od wyników poprzednich ciągnień. Wynik losowania można przedstawić jako ciąg zmiennych losowych zależnych, a odpowiadający mu schemat losowania nazywa się schematem losowania zależnego lub losowania bez zwracania.

Na podstawie zaobserwowanych w próbce wielkości $x_1, \ldots, x_n$ obliczamy wartość estymatora interesującego nas parametru Q populacji. Przy obliczaniu wartości oczekiwanej i wariancji różnych estymatorów rachunki upraszczają się na ogół znacznie, gdy mamy do czynienia ze zmiennymi losowymi niezależnymi. Natomiast funkcje zmiennych losowych zależnych mają bardzo skomplikowane rozkłady probabilistyczne i stąd wynikają pewne trudności w ostatecznym przedstawieniu wyników dotyczących np. przedziałów ufności dla parametru Q na podstawie próbki uzyskanej przy pomocy schematu losowania zależnego. Tak więc ze względu na opracowanie probabilistyczne preferuje się próbkę uzyskaną na podstawie schematu losowania niezależnego. Dla praktyka natomiast schemat losowania zależnego jest bardziej rozsądny, ponieważ nie ma tam kilkakrotnego powtarzania tego samego, przypadkowo wybranego wyniku.

§ 2.2. Średnia i wariancja w próbce. Oznaczenia.

Każdemu elementowi rozważanej populacji przypisujemy określoną wartość pewnej cechy ilościowej X. Oznaczmy przez X_k wartość cechy X dla elementu o numerze k (k=1, ..., N). Jako cel naszych badań stawiamy sobie oszacowanie średniej arytmetycznej tej cechy w badanej populacji. Oznaczmy tę średnią symbolem $\bar{X}$. Tak więc z definicji

$$\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k \tag{1}$$

Wariancję cechy X w badanej populacji określamy następująco:

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} (X_k - \bar{X})^2 \tag{2}$$

Z pewnych względów, które wyjaśnią się później, wygodnie jest posługiwać się czasem tak zwaną wariancją zmodyfikowaną, definiowaną następującym wzorem:

$$(3) \qquad S^2 = \frac{1}{N-1} \sum_{k=1}^{N} (X_k - \bar{X})^2.$$

Porównując wzory (2) i (3) widzimy, że między wariancją σ^2 a wariancją zmodyfikowaną S^2 zachodzi następujący związek:

$$(4) \qquad S^2 = \frac{N}{N-1} \sigma^2$$

Badaną cechą może być również tak zwana cecha zerojedynkowa przyjmująca tylko dwie wartości 0 i 1. Zwykle przyjmuje się, że X=0, gdy badany element nie posiada uwzględnianej charakterystyki oraz, że X=1, gdy badany element uwzględnianą charakterystykę posiada, czyli gdy jest on tak zwanym elementem wyróżnionym. W takim przypadku średnia arytmetyczna badanej cechy jest po prostu frakcją elementów wyróżnionych danej populacji:

$$(5) \qquad \bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k = \frac{N^+}{N} = P$$

gdzie N^+ oznacza liczbę elementów populacji posiadających daną charakterystykę.

Ponieważ dla cechy zerojedynkowej $X_k^2 = X_k$, więc dla wariancji takiej cechy otrzymujemy:

$$(6) \qquad \sigma^2 = \frac{1}{N} \sum_{k=1}^{N} (X_k - \bar{X})^2 = \frac{1}{N} \sum_{k=1}^{N} X_k^2 - (\bar{X})^2 = P - P^2 = P(1-P).$$

Odpowiednie wartości cechy zaobserwowane w próbce oznaczamy małymi literami. Zaobserwowaną wartość cechy X dla i-tego elementu próbki będziemy oznaczać przez x_i. Niech n będzie liczebnością próbki. Wtenczas średnia arytmetyczna z próbki obliczana jest z wzoru:

$$(7) \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Wariancję próbkową s^2 definiujemy następująco:

$$(8) \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Dla cechy zerojedynkowej wzory (7) i (8) przyjmują następującą postać:

$$(9) \qquad \bar{x} = p = \frac{n^+}{n}$$

$$(10) \qquad s^2 = \frac{n}{n-1} p(1-p)$$

§ 2.3. Twierdzenia o średniej arytmetycznej i wariancji z próbki.

Wyprowadzimy teraz wzory na wartości oczekiwane średniej arytmetycznej, wariancję średniej arytmetycznej oraz oczekiwaną wartość wariancji z próbki, jeśli próbka ta została pobrana z populacji według schematu losowania zależnego lub niezależnego. Wzory te będą potrzebne przy budowie przedziałów ufności dla średniej. Czytelnik, który nie interesuje się wywodami teoretycznymi, ale chciałby poznać jedynie końcowe rezultaty, może znaleźć je w tabeli I na str. 10.

Twierdzenie 1. Wartość oczekiwana średniej arytmetycznej z próbki pobranej według schematu losowania niezależnego równa się średniej arytmetycznej $\bar{X}$ populacji, z której ta próbka została wylosowana.

Wynik tego twierdzenia można wyrazić wzorem

$$E(\bar{x}) = \bar{X} . \tag{11}$$

Dowód. W dowodzie skorzystamy z następującego podstawowego twierdzenia z rachunku prawdopodobieństwa: Wartość oczekiwana dowolnej kombinacji liniowej zmiennych losowych równa się kombinacji wartości oczekiwanych tych zmiennych. Wobec tego

$$E\left[\frac{1}{n}(x_1 + \ldots + x_n)\right] = \frac{1}{n}\left[E(x_1) + \ldots + E(x_n)\right].$$

Obliczymy teraz wartość oczekiwaną $E(x_i)$ dla i-tego elementu próbki. Z definicji

$$E(x_i) = \sum_{k=1}^{N} X_k \cdot Pr\left\{x_i = X_k\right\}$$

Przy losowaniu niezależnym jako i-ty z kolei element próbki można otrzymać równie dobrze każdy z N elementów populacji generalnej, przy czym każdy z nich z prawdopodobieństwem $\frac{1}{N}$ może zostać wybrany do próbki. Stąd

$$E(x_i) = \sum_{k=1}^{N} X_k \cdot \frac{1}{N} = \bar{X}.$$

Wobec tego

$$E(\bar{x}) = \frac{1}{n}\left[\bar{X} + \ldots + \bar{X}\right] = \bar{X}$$ cbdo.

Twierdzenie 2. Wartość oczekiwana średniej arytmetycznej z próbki pobranej według schematu losowania zależnego równa się śred-

niej arytmetycznej $\bar{X}$ populacji, z której ta próbka została wylosowana.

Wynik tego twierdzenia można znów wyrazić wzorem (11).

Dowód: Podobnie w tw. 1

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i),$$

przy czym

$$E(x_i) = \sum_{i=1}^{n} X_k \cdot \Pr\left\{x_i = X_k\right\}.$$

Przy losowaniu zależnym prawdopodobieństwo wylosowania za i-tym razem elementu populacji generalnej o numerze k, czyli prawdopodobieństwo zdarzenia $x_i = X_k$, jest równe prawdopodobieństwu, że element oznaczony numerem k w populacji nie zostanie wylosowany w pierwszych i-1 losowaniach oraz że zostanie on wylosowany w i-tym losowaniu. Prawdopodobieństwo to wyraża się następującym wzorem:

$$\Pr\left\{x_i = X_k\right\} = (1 - \frac{1}{N})(1 - \frac{1}{N-1})\ldots(1 - \frac{1}{N-i+2})(\frac{1}{N-i+1}) =$$

$$= \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdots \frac{N-i+1}{N-i+2} \cdot \frac{1}{N-i+1} = \frac{1}{N}$$

Tak więc podobnie jak w tw. 1 otrzymujemy

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{N} X_k \cdot \frac{1}{N} = \frac{1}{n} \cdot n \cdot \bar{X} = \bar{X} \quad \text{cbdo.}$$

Wariancję estymatora t definiujemy następująco:

$$D^2(t) = E\left[t - E(t)\right]^2 = \sum_{j}\left[t_j - E(t)\right]^2 \cdot \Pr\left\{t = t_j\right\}$$

Wobec tego wariancję średniej arytmetycznej z próbki należy obliczać według następującego wzoru:

$$D^2(\bar{x}) = E(\bar{x} - \bar{X})^2.$$

Twierdzenie 3. Wariancja średniej arytmetycznej z próbki pobranej według schematu losowania niezależnego wyraża się wzorem

$$D^2(\bar{x}) = \frac{\sigma^2}{n} \tag{12}$$

Dowód. W dowodzie posługujemy się następującym twierdzeniem rachunku prawdopodobieństwa: Jeśli t jest kombinacją liniową niezależnych zmiennych losowych $t_1, \ldots, t_n$, czyli $t = \sum_{i=1}^{n} a_i t_i$, to

zachodzi następująca równość:

$$D^2(t) = D^2\left(\sum_{i=1}^{n} a_i t_i\right) = \sum_{i=1}^{n} a_i^2 D^2(t_i)$$

Posługując się tym wzorem mamy

$$D^2(\bar{x}) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(x_i)$$

ale

$$D^2(x_i) = E(x_i-\bar{X})^2 = \sum_{k=1}^{N}(X_k-\bar{X})^2 \cdot \Pr\{x_i=X_k\} = \frac{1}{N}\sum_{k=1}^{N}(X_k-\bar{X})^2 = \sigma^2 .$$

Wobec tego

$$D^2(\bar{x}) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(x_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n} \qquad \text{cbdo.}$$

Twierdzenie 4. Wariancja średniej arytmetycznej z próbki pobranej według schematu losowania zależnego wyraża się wzorem

$$D^2(\bar{x}) = \frac{N-n}{N-1}\,\frac{\sigma^2}{n} \tag{13}$$

Dowód tego twierdzenia jest dłuższy i bardziej uciążliwy, ponieważ możliwe wyniki losowania są zmiennymi losowymi zależnymi. Wobec tego dowód ten pominiemy. Można go znaleźć np. w [3] lub [5].

Twierdzenie 5. Oczekiwana wartość wariancji próbkowej (zdefiniowanej wzorem (8)) przy losowaniu niezależnym wyraża się następującym wzorem:

$$E(s^2) = \sigma^2. \tag{14}$$

Dowód. W dowodzie będziemy korzystać z następujących wzorów na oczekiwaną wartość iloczynu zaobserwowanych w próbce niezależnych zmiennych losowych x_i oraz x_j:

$$E(x_i x_j) = \begin{cases} \overline{X^2} & \text{dla } i = j \\ [E(X)]^2 & \text{dla } i \neq j \end{cases}$$

Część pierwsza powyższego wzoru wynika z tego, że gdy i=j to

$$E(x_i x_j) = E(x_i^2) = \sum_{k=1}^{N} X_k^2 \cdot \frac{1}{N} = \frac{1}{N}\sum_{k=1}^{N} X_k^2 = \overline{X^2}$$

Uwzględniając fakt, że zmienne losowe x_i, x_j są niezależne, otrzymujemy drugą część powyższego wzoru w sposób następujący:

$$E(x_i x_j) = (Ex_i)(Ex_j) = [E(X)]^2$$

Korzystając z wypisanego na str. 8 wyrażenia $E(x_i x_j)$ i przy użyciu znanych przekształceń otrzymujemy kolejno

$$E(s^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n(\bar{x})^2\right]\right\} =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}E(x_i^2) - \frac{1}{n-1}E\left[n(\bar{x})^2\right] =$$

$$= \frac{1}{n-1}\,n.\overline{X^2} - \frac{1}{n-1}E\left[\frac{1}{n}\sum_{i=1}^{n}x_i^2 + \frac{1}{n}\sum_{i\neq j}x_i x_j\right] =$$

$$= \frac{n}{n-1}\overline{X^2} - \frac{1}{n-1}\overline{X^2} - \frac{n(n-1)}{(n-1)n}\left[E(X)\right]^2 = \overline{X^2} - \left[E(X)\right]^2 = \sigma^2 \text{ cbdo.}$$

Twierdzenie 6. Oczekiwana wartość wariancji próbkowej przy losowaniu zależnym wyraża się następującym wzorem:

$$E(s^2) = \frac{N}{N-1}\sigma^2 = S^2. \tag{15}$$

Szkic dowodu. Będziemy korzystać z następujących wzorów na oczekiwaną wartość iloczynu zaobserwowanych w próbce zmiennych losowych x_i oraz x_j otrzymanych przy schemacie losowania zależnego:

$$E(x_i x_j) = \begin{cases} \overline{X^2} & \text{dla } i=j \\ \dfrac{\sigma^2}{N-1} + \left[E(X)\right]^2 & \text{dla } i\neq j \end{cases}$$

Pierwszą część powyższego wzoru można otrzymać z rozważań takich, jak w dowodzie tw. 5, druga część jest trudniejsza do wykazania i można ją znaleźć np. w [3]. Znając powyższy wzór w sposób całkowicie podobny jak w dowodzie tw. 5 dochodzimy do szukanego rezultatu.

Tabela I

Oznaczenia i definicje	
Elementy populacji: $X_1, X_2, \ldots, X_N$	
Średnia populacyjna $\overline{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$	
Wariancja populacyjna $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(X_k-\overline{X})^2$	
Elementy próbki: $x_1, x_2, \ldots, x_n$	
Średnia próbkowa $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	
Wariancja próbkowa $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$	
Wyniki twierdzeń 1-6	
losowanie niezależne	**losowanie zależne**
$E(\overline{x}) = \overline{X}$ $D^2(\overline{x}) = \frac{\sigma^2}{n}$ $E(s^2) = \sigma^2$	$E(\overline{x}) = \overline{X}$ $D^2(\overline{x}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$ $E(s^2) = \frac{N}{N-1}\sigma^2$

§ 2.4. Porównanie wyników dla schematu losowania zależnego i schematu losowania niezależnego.

Porównując wzory na wariancję średniej arytmetycznej w obu schematach losowania otrzymujemy następującą nierówność:

$$\frac{\sigma^2}{n} > \frac{N-n}{N-1}\frac{\sigma^2}{n} \quad \text{dla } n>1.$$

Wypływa stąd następujący wniosek: Wariancja średniej arytmetycznej przy schemacie losowania zależnego jest mniejsza niż analogiczna wariancja przy schemacie losowania niezależnego. Znaczy to, że ten ostatni schemat daje oszacowanie nieznanej wartości średniej $\overline{X}$ badanej populacji bardziej rozproszone dookoła tejże wartości. Róż-

nice są tym większe, im większa jest liczebność próbki w stosunku do liczebności populacji. Wniosek ten jest zgodny z intuicją. Przy losowaniu niezależnym zdarza się, że niektóre elementy populacji mierzymy częściej niż raz i tym samym do próbki wchodzi na ogół mniej niż n różnych elementów populacji. Wobec tego z próbki tej otrzymujemy na ogół mniej informacji o badanej populacji niż gdybyśmy przebadali próbkę składającą się z n różnych elementów, co mamy zagwarantowane przy losowaniu zależnym.

Wariancja średniej arytmetycznej z próbki zależy od wariancji populacyjnej σ^2. Jeśli ta ostatnia jest nieznana i zamiast niej chcemy posłużyć się jej oszacowaniem z próbki czyli wielkością s^2, a oszacowanie to ma być nieobciążone, to przy losowaniu niezależnym na miejsce σ^2 można podstawić s^2, natomiast przy losowaniu zależnym lepiej posłużyć się wynikiem twierdzenia 6 i na miejsce σ^2 podstawić wielkość $\frac{N-1}{N} s^2$. Podstawione w ten sposób wielkości będą nieobciążonymi estymatorami nieznanej wielkości σ^2.

W poszczególnych przypadkach wariancja z próbki może znacznie odbiegać od wariancji w populacji.

Kiedy zamiast wariancji populacyjnej można podstawić wariancję z próbki i nie pomylić się za bardzo? Zagadnienie to jest szczegółowo rozpatrywane np. w [10], gdzie dochodzi się do następującej konkluzji: Jeśli cecha X ma w populacji rozkład w przybliżeniu normalny, to dla oszacowania nieznanej wariancji w populacji z dokładnością do połowy jej prawdziwej wartości wystarczy pobrać próbkę liczącą 50 elementów.

§ 2.5. Przedziały ufności dla średniej arytmetycznej.

Rozkład średniej z próbki zależny jest od rozkładu badanej cechy w populacji. Jeśli rozkład cechy w populacji jest normalny, a losowanie jest niezależne, to rozkład średniej z próbki jest również normalny. Jeśli tak, to możemy zbudować przedział ufności dla nieznanej średniej populacyjnej, tj. taki przedział, aby z góry zadanym prawdopodobieństwem $1-\alpha$ nieznana średnia $\overline{X}$ mieściła się w tym przedziale. Przedział ten jest wyznaczony następującą nierównością:

$$(16) \qquad \overline{x} - t\sqrt{D^2(\overline{x})} < \overline{X} < \overline{x} + t\sqrt{D^2(\overline{x})}.$$

W powyższej nierówności $\overline{x}$ jest średnią arytmetyczną obliczoną na podstawie próbki, $D^2(\overline{x})$ - wariancją tejże średniej. Liczba t jest wielkością odczytaną z tablic rozkładu normalnego w zależności od

przyjętego poziomu ufności $1-\alpha$, to znaczy tak, aby

$$Pr\left\{\bar{x} - t\sqrt{D^2(\bar{x})} < \bar{X} < \bar{x} + t\sqrt{D^2(\bar{x})}\right\} = 1-\alpha.$$

I tak przykładowo dla $1-\alpha = 0{,}95$ należy przyjąć $t=1{,}96$; dla $1-\alpha = 0{,}99$ przyjmujemy $t = 2{,}58$.

Przedział ufności wyznaczony nierównościami (16) można inaczej zapisać w postaci

$$\bar{x} - d < \bar{X} < \bar{x} + d$$

gdzie d nosi nazwę dokładności oszacowania średniej populacyjnej i jest zdefiniowane równością

$$d = t\sqrt{D^2(\bar{x})}$$

W praktyce stosuje się przy konstrukcji przedziałów ufności wzór (16) również wtedy, gdy rozkład cechy X w populacji nie jest normalny. Z twierdzeń statystyki matematycznej wynika bowiem, że nawet jeśli rozkład cechy X w badanej populacji znacznie odbiega od rozkładu normalnego, to rozkład średniej arytmetycznej wylosowanych wartości przy zwiększającej się liczebności próbki jest szybko zbieżny do rozkładu normalnego.

Jak wynika z wzoru (16), końce przedziału ufności są funkcjami $D^2(\bar{x})$, która to wielkość z kolei zależy od wariancji populacyjnej σ^2. Jeśli ta ostatnia jest nieznana, a próbka jest duża ($n > 50$), to na miejsce σ^2 można podstawić wartość σ^2 obliczoną z próbki. Jeśli liczebność próbki jest stosunkowo mała, rozkład badanej cechy X w populacji jest normalny z nieznaną średnią i wariancją populacyjną, a obserwacje w próbce są niezależne, to możemy skonstruować przedział ufności dla średniej populacyjnej $\bar{X}$ w oparciu o tablice rozkładu t Studenta. W takim przypadku przedział ufności wyraża się również wzorem (16) z tą różnicą, że wielkości t są odczytywane z tablic rozkładu t-Studenta i zależą od liczebności próbki n.

Przy losowaniu zależnym konstruuje się przedział ufności dla nieznanej średniej populacyjnej również w oparciu o wzór (16), podstawiając na miejsce $D^2(\bar{x})$ wartość obliczoną według wzoru (13). Jeśli wariancja populacyjna σ^2 jest nieznana, to podstawiamy na jej miejsce jej nieobciążony estymator $\frac{N-1}{N} s^2$ tak, że ostatecznie wariancja średniej arytmetycznej wyraża się wzorem

$$D^2(\bar{x}) = \frac{N-n}{N} \cdot \frac{s^2}{n} . \qquad (17)$$

Ostatecznie więc otrzymujemy następujące przedziały ufności dla nieznanej średniej populacyjnej $\bar{X}$:
1° przy losowaniu niezależnym:

$$\bar{x} - t\frac{s}{\sqrt{n}} < \bar{X} < \bar{x} + t\frac{s}{\sqrt{n}}, \tag{18}$$

2° przy losowaniu zależnym:

$$\bar{x} - t\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} < \bar{X} < \bar{x} + t\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \tag{19}$$

Szczegółową dyskusję na temat wpływu nienormalności rozkładu na otrzymane przedziały ufności można znaleźć w monografii [2]. Niektórzy autorzy (np. Hansen, Hurwitz, Madow [4]) nie dyskutują tego zagadnienia, lecz zabezpieczają się przed skutkami niespełnienia założeń przy których otrzymano przedział ufności (16) w ten sposób, że konstruują przedział ufności w oparciu o tak zwane prawo 3σ. Uzasadnia się je następująco: Jeśli próbka pochodzi z rozkładu normalnego, to szanse na to, że różnica między średnią próbkową a średnią populacyjną przekroczy trzy dyspersje średniej są znikome; wynoszą one 1 na 1000. Jeśli rozkład w populacji nie jest normalny, to szanse te wzrastają, lecz pozostają w dalszym ciągu znikome. Tak jest dla ogółu spotykanych w praktyce rozkładów, szczególnie przy stosunkowo dużych próbkach. Wobec tego w przypadku, gdy rozkład populacji nie jest dokładnie znany, nie można skonstruowanemu przedziałowi ufności przypisać dokładnego poziomu ufności, można jednak wyrazić przekonanie, że jesteśmy "praktycznie pewni", iż nieznany parametr badanej populacji znajduje się w podanym przedziale.

Przykład 1. Spośród N=8000 dziewczynek 11-letnich danego miasta wylosowano próbkę liczącą n = 324 różnych dziewczynek. Średnia arytmetyczna wzrostu dziewczynek z próbki wynosiła $\bar{x}$=145,0 cm, a dyspersja s = 5,4 cm. W jakich granicach znajduje się średnia arytmetyczna wzrostu $\bar{X}$ ogółu dziewczynek 11-letnich danego miasta? Wiadomo, że rozkład wzrostu w badanej populacji jest normalny, a próbka została pobrana przy użyciu schematu losowania zależnego.

Korzystając z ostatnich informacji należy skonstruować przedział ufności na podstawie wzoru (19). Poziomowi ufności $1-\alpha$ = = 0,95 odpowiada odczytana z tablic rozkładu normalnego wartość t = 1,96. Po podstawieniu odpowiednich danych do wzoru (19) otrzy-

mujemy

$$145 - 1{,}96\cdot\frac{5{,}4}{18}\cdot\sqrt{1-\frac{324}{8000}} < \bar{X} < 145 + 1{,}96\cdot\frac{5{,}4}{18}\sqrt{1-\frac{324}{8000}}$$

$$145 - 0{,}588\cdot 0{,}9795 < \bar{X} < 145 + 0{,}588\cdot 0{,}9795$$

$$145 - 0{,}5759 < \bar{X} < 145 + 0{,}5759$$

$$144{,}42 < \bar{X} < 145{,}58$$

Cochran [2] podaje następujący przykład konstrukcji przedziału ufności w oparciu o wzór (16) dla danych odbiegających znacznie od rozkładu normalnego:

Przykład 2. Przy pewnej petycji zebrano podpisy na 676 arkuszach. Na każdym arkuszu było miejsce na 42 podpisy, ale na niektórych arkuszach podpisów było mniej. Wylosowano 50 arkuszy i policzono na każdym z nich liczbę podpisów. Z danych tych oszacowano łączną liczbę podpisów na wszystkich arkuszach. Badaną cechą X jest tu liczba podpisów na poszczególnych arkuszach. Z zaobserwowanych w próbce wartości cechy X ułożono szereg rozdzielczy, który przedstawiał się następująco:

x_j	42	41	36	32	29	27	23	19	16	15	14	11	10	9	7	6	5	4	3
n_j	23	4	1	1	1	2	1	1	2	2	1	1	1	1	1	3	2	1	1

Wskaźnik j przyjmuje tu wartości j=3, 4, ..., 42; $\sum_{j=3}^{42} n_j = 50$.

Z danych tych obliczono

$$\bar{x} = \frac{1}{50}\sum_{j=3}^{42} n_j x_j = \frac{1471}{50} = 29{,}42$$

Z próbki mamy więc, że średnio na jednym arkuszu znajduje się 29,42 podpisów. Stąd otrzymujemy szacunkową liczbę podpisów na wszystkich 676 arkuszach

$$N\cdot\bar{x} = 676\cdot 29{,}42 = 19\ 888.$$

Następnie obliczamy wariancję próbkową dla zaobserwowanej liczby podpisów na 50 arkuszach:

$$s^2 = \frac{1}{n-1}\sum_{j=3}^{42} n_j(x_j-\bar{x})^2 = \frac{1}{n-1}\left[\sum_{j=3}^{42} n_j x_j^2 - \frac{\left(\sum_{j=3}^{42} n_j x_j\right)^2}{\sum_{j=3}^{42} n_j}\right] =$$

$$= \frac{1}{49}\left[54\ 497 - \frac{1471^2}{50}\right] = 229{,}0\ .$$

Konstruujemy teraz przedział ufności dla $N\bar{X}$ według wzoru (16) podstawiając $D^2(N\bar{x}) = N^2 \cdot \frac{s^2}{n} \cdot (1-\frac{n}{N})$. Chcąc zbudować przedział na poziomie ufności $1-\alpha = 0{,}80$ przyjmujemy $t = 1{,}28$. Podstawiając odpowiednie dane do wzoru (16) otrzymujemy następujące granice dla przedziału ufności:

$$19\,888 \pm \frac{tNs}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = 19\,888 \pm \frac{1{,}28 \cdot 676 \cdot 15{,}13 \cdot \sqrt{1-0{,}0740}}{\sqrt{50}}$$

$$= 19\,888 \pm 1781 .$$

Stąd otrzymujemy, że łączna liczba podpisów z prawdopodobieństwem 0,80 zawiera się w przedziale 18 107 do 21 669.

Po wykonaniu tego oszacowania porachowano prawdziwą liczbę podpisów na zebranych 676 arkuszach. Okazało się, że było ich 21 045!

§ 2.6. Twierdzenia o wartości oczekiwanej i wariancji dla frakcji elementów wyróżnionych w próbce.

Jak już stwierdziliśmy, frakcję elementów wyróżnionych możemy uważać za średnią arytmetyczną cechy, która przyjmuje tylko dwie wartości: wartość 1, gdy rozważany element posiada badaną charakterystykę czyli wyróżnia się, i wartość 0, gdy rozważany element badanej charakterystyki nie posiada. Twierdzenia 1-6 pozostają również słuszne dla takiej cechy. Jeśli skorzystamy z zależności $\sum_{i=1}^{n} x_i = np$ i $\sum_{k=1}^{N} X_k = NP$, gdzie p i P są to wcześniej już wprowadzone frakcje elementów wyróżnionych w próbce i populacji, to twierdzenia te można przedstawić jako twierdzenia o wartości oczekiwanej i wariancji zmiennej losowej p. W tabeli II podano odpowiedniki wyrażeń tabeli I przedstawione jako funkcje p lub P.

Tabela II

Oznaczenia i definicje
Elementy populacji: $X_k = \begin{cases} 1 & \text{gdy k-ty element wyróżnia się} \\ 0 & \text{" " " nie wyróżnia się} \end{cases}$
$k = 1, \ldots, N$
Średnia populacja $\bar{X} = P = \frac{1}{N}\sum_{k=1}^{n} X_k$
Wariancja populacyjna $\sigma^2 = P(1-P) = \frac{1}{N}\sum_{k=1}^{N}(X_k - \bar{X})^2$
Elementy próbki: $x_i = \begin{cases} 1 & \text{gdy i-ty element wyróżnia się} \\ 0 & \text{gdy i-ty element nie wyróżnia się} \end{cases}$
$i = 1, \ldots, n$
Średnia próbkowa $\bar{x} = p = \frac{1}{n}\sum_{i=1}^{n} x_i$
Wariancja próbkowa $s^2 = \frac{n}{n-1}\cdot p\cdot(1-p) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Wyniki twierdzeń 1-6 dla cechy zerojedynkowej	
losowanie niezależne	losowanie zależne
$E(p) = P$	$E(p) = P$
$D^2(p) = \frac{P(1-P)}{n}$	$D^2(p) = \frac{N-n}{N-1}\frac{P(1-P)}{n}$
$E\left[\frac{n}{n-1}\cdot p\cdot(1-p)\right] = P(1-P)$	$E\left[\frac{n}{n-1}\cdot\frac{N-1}{N}\cdot p(1-p)\right] = P(1-P)$

Przykład 3 (zaczerpnięty z monografii Cochrana [2]). Z listy obejmującej 3042 nazwiska i adresy pobrano próbkę nieograniczoną zależną obejmującą 200 nazwisk, po czym po sprawdzeniu adresów okazało się, że 38 było fałszywych. Oszacować liczbę adresów fałszywych na tej liście i znaleźć błąd tego oszacowania (błędem oszacowania nazywa się pierwiastek z wariancji estymatora). W naszym wypadku estymatorem nieznanej liczby fałszywych adresów jest wielkość Np, a więc mamy znaleźć $\sqrt{D^2(Np)}$.

Stosując oznaczenia według tabeli II mamy N = 3042, n = 200, p = 38/200 = 0,19. Mamy oszacować NP. Nieobciążonym estymatorem

tej wielkości jest Np, ponieważ zgodnie z tw. 2 mamy $E(Np) = N \cdot E(p) = NP$. Wariancja tego estymatora zgodnie z tw. 4 wynosi

$$D^2(Np) = N^2 \cdot D^2(p) = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{P(1-P)}{n} .$$

Podstawiając na miejsce iloczynu P(1-P) jego nieobciążony estymator zgodnie z tw. 6 otrzymujemy

$$D^2(Np) = \frac{N(N-n)}{n-1} \cdot p(1-p).$$

Podstawiając odpowiednie dane liczbowe otrzymujemy, że estymator oczekiwanej liczby fałszywych adresów równa się

$$Np = 3042 \cdot 0,19 = 578.$$

Wariancja tego estymatora równa się odpowiednio

$$D^2(Np) = \frac{3042 \cdot 2842 \cdot 0,19 \cdot 0,81}{199} = 6685.$$

Błąd estymatora czyli pierwiastek z jego wariancji równa się wobec tego

$$\sqrt{D^2(Np)} = \sqrt{6685} = 81,8.$$

Gdybyśmy obliczyli błąd według wzorów dla losowania niezależnego, to otrzymalibyśmy

$$\sqrt{D^2(Np)} = N \cdot \sqrt{\frac{p(1-p)}{n}} = 3042 \cdot \sqrt{\frac{0,19 \cdot 0,81}{200}} = 84,4 .$$

Wynik ten potwierdza nasze spostrzeżenie z § 2.4., że wariancja średniej arytmetycznej z próbki przy losowaniu niezależnym jest większa niż przy losowaniu zależnym.

§ 2.7. Przedziały ufności dla nieznanej frakcji elementów wyróżnionych w populacji.

Aby zbudować przedział ufności dla parametru badanej populacji należy znać rozkład estymatora tego parametru. Jeśli chodzi o frakcję elementów wyróżnionych w próbce, to wiadomo, że ma ona rozkład hipergeometryczny. Rozkład ten ze wzrostem liczebności próbki jest zbieżny do rozkładu normalnego. Szybkość tej zbieżności zależy od liczebności próbki oraz od wielkości P czyli frakcji elementów wyróżnionych w populacji. Szczegółową dyskusję tego zagadnienia można znaleźć np. w [2] lub [10]. Według Cochrana [2] rozkład frakcji p można przyjmować za normalny, jeśli są spełnione warunki podane w tabelce zamieszczonej na str. 18.

P	Mniejsza z liczb nP oraz n(1-P)	Liczebność próbki n conajmniej
0,5	15	30
0,4 lub 0,6	20	50
0,3 lub 0,7	24	80
0,2 lub 0,8	40	200
0,1 lub 0,9	60	600
0,05 lub 0,95	70	1400

Jeśli są podstawy do stwierdzenia, że rozkład frakcji elementów wyróżnionych w próbce można przybliżyć rozkładem normalnym, to możemy skonstruować przedziały ufności dla P w oparciu o wzór (16) podstawiając odpowiednie wyrażenia na $D^2(p)$ według tabeli II. Otrzymamy wtedy następujące wzory:

1° przy losowaniu niezależnym:

$$(20) \qquad p - t\sqrt{\frac{p(1-p)}{n-1}} < P < p + t\sqrt{\frac{p(1-p)}{n-1}},$$

2° przy losowaniu zależnym:

$$(21) \qquad p - t\sqrt{(1-\frac{n}{N})\frac{p(1-p)}{n-1}} < P < p + t\sqrt{(1-\frac{n}{N})\frac{p\cdot(1-p)}{n-1}}$$

We wzorach tych t oznacza wielkość odczytaną z tablic rozkładu normalnego.

Porównując wzory (20) i (21) można zauważyć, że wzory dla schematu losowania niezależnego otrzymuje się ze wzoru dla schematu losowania zależnego przez podstawienie $N = \infty$. Jeśli populacja składa się z nieskończenie wielu elementów, to prawdopodobieństwo, że jakiś element zostanie wylosowany więcej niż jeden raz wynosi zero i schemat losowania zależnego staje się schematem losowania niezależnego.

Przykład 4. Na n = 300 zbadanych dzieci szkolnych $n^+ = 48$ okazało się być nosicielami dyfterii. Przyjmując, że zbadane dzieci stanowią próbkę ogółu dzieci szkolnych danej miejscowości wyznaczyć przedział ufności dla frakcji P nosicieli dyfterii wśród dzieci tej miejscowości. Przyjąć poziom ufności $1-\alpha = 0,95$.

Populację, z której zbadane dzieci zostały wybrane, uważamy za nieskończenie wielką. Wobec tego zgodnie z wzorem (20) mamy

$$\frac{48}{300} - 1,96\sqrt{\frac{1}{299}\cdot\frac{48}{300}\left(1-\frac{48}{300}\right)} < P < \frac{48}{300} + 1,96\sqrt{\frac{1}{299}\cdot\frac{48}{300}\left(1-\frac{48}{300}\right)}$$

co po wykonaniu obliczeń daje, że z prawdopodobieństwem 0,95 frakcja nosicieli dyfterii wśród dzieci szkolnych tej miejscowości zawiera się w granicach

$$0{,}139 < P < 0{,}181,$$

które mieliśmy obliczyć.

W nowszych podręcznikach statystyki (por. [1],[5]) konstruuje się przedziały ufności niekoniecznie symetryczne względem wartości oczekiwanej. Można założyć prawdopodobieństwa α_1 i α_2, a następnie szukać oszacowań nieznanego parametru populacyjnego P z dołu i z góry przez dwie liczby $\underline{P}$ i $\overline{P}$ takie, że

$$\Pr\left\{x \geqslant x_0 \mid P = \underline{P}\right\} = 1-\alpha_2 \tag{22}$$

$$\Pr\left\{x \leqslant x_0 \mid P = \overline{P}\right\} = \alpha_1 \tag{23}$$

przy czym x_0 oznacza zaobserwowaną liczbę elementów wyróżnionych w próbce, x oznacza zmienną losową równą potencjalnej liczbie elementów wyróżnionych w próbce. Metoda ta pochodzi od wybitnego polskiego statystyka J. Spławy-Neymana. Konstruując w ten sposób ograniczenia dolne i górne nieznanego parametru P otrzymujemy przedziały, które z prawdopodobieństwem $\alpha_2-\alpha_1$ pokryją nieznaną wartość parametru P.

Jeśli liczebność próbki jest duża, a P niezbyt małe lub niezbyt wielkie, to możemy prawdopodobieństwa (22) i (23) przybliżyć rozkładem normalnym. Oznaczmy przez t_1 i t_2 wartości odczytane z tablic rozkładu normalnego odpowiednio dla α_1 i α_2. Korzystając z tzw. poprawki na nieciągłość otrzymujemy następujące równania na $\underline{P}$ i $\overline{P}$:

$$\frac{x_0 - 1/2 - n\underline{P}}{\sqrt{n\underline{P}(1-\underline{P})}} = t_2$$

$$\frac{x_0 + 1/2 - n\overline{P}}{\sqrt{n\overline{P}(1-\overline{P})}} = t_1$$

Rozwiązując te równania ze względu na $\underline{P}$ i $\overline{P}$ otrzymujemy

$$\underline{P} = \frac{1}{n+t_2^2}\left[x_0 - \frac{1}{2} + \frac{t_2^2}{2} - t_2\sqrt{\frac{(x_0-1/2)(n-x_0+1/2)}{n} + \frac{t_2^2}{4}}\right]. \tag{24}$$

$$\overline{P} = \frac{1}{n+t_1^2}\left[x_0 + \frac{1}{2} + \frac{t_1^2}{2} - t_1\sqrt{\frac{(x_0+1/2)(n-x_0-1/2)}{n} + \frac{t_1^2}{4}}\right]. \tag{25}$$

Jeśli nie chcemy lub nie możemy korzystać z przybliżenia prawdopodobieństw (22) i (23) rozkładem normalnym, to możemy je przybliżyć rozkładem Bernoulliego, a następnie wykorzystać związek między rozkładem Bernoulliego a rozkładem F Snedecora i otrzymać dokładniejsze oszacowanie $\overline{P}$ ze wzoru

$$\overline{P} = \frac{(x_0+1)\cdot F}{n-x_0+(x_0+1)\cdot F} \tag{26}$$

przy czym F oznacza wielkość odczytaną z tablic F Snedecora dla $\alpha = 1-\alpha_1$ i liczby stopni swobody $n_1 = 2(x_0+1)$, $n_2 = 2(n-x_0)$. Podobnie dokładniejsze oszacowanie $\underline{P}$ otrzymujemy ze wzoru

$$\underline{P} = \frac{x_0}{x_0+(n-x_0+1)\cdot F}, \tag{27}$$

przy czym F należy tym razem odczytać dla $\alpha = \alpha_2$ i liczby stopni swobody $n_1 = 2(n-x_0+1)$, $n_2 = 2x_0$.

Przykład 5. (według [1]). Na 30 samochodów wybranych losowo z produkcji pewnej fabryki stwierdzono 8 samochodów z usterkami. W jakich granicach zawiera się frakcja samochodów z usterkami charakteryzująca produkcję tej fabryki?

Przyjmijmy $\alpha_1 = 0{,}025$, $\alpha_2 = 0{,}975$. Wtedy $t_1 = -1{,}96$, a $t_2 = +1{,}96$. Podstawiając do wzorów (24) i (25) $x_0=8$ i $n=30$ otrzymujemy $\underline{P} = 0{,}130$, $\overline{P} = 0{,}462$.

Jeśli chcemy skorzystać z dokładniejszych wzorów (26) i (27), to najpierw znajdujemy odpowiednie liczby stopni swobody dla odczytania wartości F. I tak dla wzoru (26) mamy $n = 2(8+1) = 18$, $n_2 = 2(30-8) = 44$. Wartość F odczytana dla tych stopni swobody i dla $\alpha = 1-0{,}025 = 0{,}975$ równa się $F = 2{,}07$.
Stąd otrzymujemy

$$\overline{P} = \frac{(8+1)\cdot 2{,}07}{30-8+(8+1)\cdot F} = 0{,}459.$$

Dla wyznaczenia $\underline{P}$ z wzoru (27) obliczamy $n_1 = 2\cdot(30-8+1) = 46$, $n_2 = 2\cdot 8 = 16$. Wartość F odczytana dla tych stopni swobody przy $\alpha = 0{,}975$ wynosi $F = 2{,}49$. Podstawiając te dane do wzoru (27) otrzymujemy

$$\underline{P} = \frac{8}{8+(30-8+1)\cdot 2{,}49} = 0{,}123.$$

Porównanie tych granic z wartościami otrzymanymi metodą przybliżoną według wzorów (24) i (25) wykazuje niewielkie różnice pomimo tego, że liczebność próbki $n = 30$ nie była zbyt duża.

§ 2.8. Dokładność oszacowania średniej arytmetycznej.

Dokładność oszacowania d określiliśmy na stronicy 12 jako maksymalną przy danym poziomie ufności różnicę między zaobserwowaną z próbki średnią arytmetyczną $\bar{x}$, a prawdziwą średnią populacyjną $\bar{X}$. Dokładność ta w myśl wzoru $d = t\sqrt{D^2(\bar{x})}$ zależy od przyjętego poziomu ufności, co wyraża się w odpowiedniej wielkości t. Dokładność d zależy również od liczebności próbki n, ponieważ $D^2(\bar{x})$ jest funkcją n. Często w praktyce stawiamy następujące pytanie: Jak liczną powinna być próbka, aby na jej podstawie oszacować nieznaną średnią populacyjną z zadaną z góry dokładnością? Aby odpowiedzieć na to pytanie należy rozwiązać następujące równania:

$$d = t \cdot \sqrt{\frac{\sigma^2}{n}} \qquad \text{dla losowania niezależnego,}$$

$$d = t \frac{S}{\sqrt{n}} \cdot \sqrt{1-\frac{n}{N}} \qquad \text{dla losowania zależnego.}$$

Otrzymujemy stąd

(27) $$n = \frac{\sigma^2 t^2}{d^2} \qquad \text{dla losowania niezależnego,}$$

(28) $$n = \frac{N}{1 + \frac{Nd^2}{t^2 S^2}} \qquad \text{dla losowania zależnego .}$$

Jeśli wariancje populacyjne σ^2 lub S^2 są nieznane, to w powyższych wzorach możemy podstawić ich nieobciążone estymatory czyli wariancję próbkową $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ uzyskaną z wstępnej próbki.

Dla cechy zerojedynkowej, gdy słuszne jest przybliżenie rozkładu p przez rozkład normalny, mamy

$$d = t \cdot \sqrt{\frac{P(1-P)}{n}} \qquad \text{dla losowania niezależnego,}$$

$$d = t \sqrt{\frac{N-n}{N-1} \frac{P(1-P)}{n}} \qquad \text{dla losowania zależnego.}$$

Otrzymujemy stąd

(29) $$n = \frac{P(1-P) \cdot t^2}{d^2} \qquad \text{dla losowania niezależnego,}$$

(30) $$n = \frac{N}{1 + \frac{(N-1)d^2}{t^2 P(1-P)}} \qquad \text{dla losowania zależnego.}$$

Jeśli na miejsce iloczynu P(1-P) podstawimy jego nieobciążony estymator obliczony na podstawie wstępnej próbki zgodnie z tw. 5 lub tw. 6 (wyniki w tabeli II), to zakładając, że próbka ta była wystarczająco duża, aby w przybliżeniu przyjąć $\frac{n}{n-1} \approx 1$, otrzymamy:

(29 a) $$n = \frac{p(1-p)\cdot t^2}{d^2}$$ dla losowania niezależnego,

(30 a) $$n = \frac{N}{1 + \frac{Nd^2}{t^2 p(1-p)}}$$ dla losowania zależnego.

Przykład 6 (zaczerpnięty z [10]). Należy zbadać średnią wagę noworodka w pewnej populacji składającej się z N =12000 noworodków. Wydaje się uzasadnione przypuszczenie, że waga noworodka ma rozkład zbliżony do normalnego. Stawiamy warunek, żeby przy oszacowaniu średniej wagi noworodka nie pomylić się więcej niż o d = = 50 g. Z poprzednich badań wynika, że wariancję wagi noworodka S^2 można przyjąć jako równą 490 000 g^2. Jak wielka powinna być próbka, aby zagwarantować sobie wymienioną dokładność d = 50 g na poziomie ufności 0,95?

Podstawiając N = 12 000, d = 50, t = 1,96, S^2 = 490 000, otrzymujemy

$$\frac{Nd^2}{t^2 S^2} = \frac{12\,000\cdot 2500}{3{,}8416\cdot 490\,000} = 15{,}9.$$

Stąd z wzoru (28) mamy

$$n = \frac{12\,000}{1+15{,}9} = 710\ .$$

Należy więc z populacji generalnej wylosować 710 noworodków.

Przykład 7 (zaczerpnięty z [10]). W mieście liczącym N = 12000 mieszkańców należy zbadać odsetek dzieci w wieku 7-13 lat, przy czym dokładność oszacowania tego odsetka ma wynosić d = 0,02. Na podstawie wyników wstępnej próbki przypuszczamy, że $p \approx 0{,}2$. Przyjmując t=3 i podstawiając odpowiednie wartości do wzoru (30 a) otrzymujemy

$$n = \frac{12000}{1+\frac{12000\cdot 0{,}0004}{9\cdot 0{,}16\cdot 0{,}84}} = \frac{12000}{1+\frac{48000}{9\cdot 16\cdot 84}} = \frac{12000\cdot 756}{3756} \approx 2415.$$

Tak więc, aby nie pomylić się w ocenie szukanego odsetka dzieci więcej niż o 0,02 należy zbadać próbkę liczącą n = 2415 osób.

3. Losowanie warstwowe

§ 3.1. Oznaczenia i definicje.

Dotychczas omawialiśmy sytuacje, w których losowanie do próbki odbywało się z całej populacji generalnej. Było to tak zwane losowanie nieograniczone. Jeśli przed przystąpieniem do losowania rozkładamy populację generalną na części i losowanie odbywa się z każdej części oddzielnie, mówimy o losowaniu warstwowym. W takiej sytuacji jesteśmy na przykład, jeśli chcemy znaleźć średnią jakiejś cechy antropologicznej populacji polskiej i całą populację rozkładamy na dwie warstwy: ludność wiejską i miejską. Podobnie, wybierając próbkę reprezentacyjną młodzieży akademickiej możemy rozłożyć całą populację młodzieży akademickiej na warstwy według uczelni, do których ta młodzież należy, a następnie z każdej warstwy-uczelni wylosować odpowiednią liczbę osób według schematu losowania niezależnego lub zależnego.

Wprowadzimy teraz oznaczenia odpowiadające takiemu podziałowi populacji. Niech populacja składająca się z N elementów dzieli się na k warstw liczących odpowiednio $N_1, N_2, \ldots, N_k$ elementów tak, że $N_1+\ldots+N_k = N$. Każdy element populacji generalnej oznaczamy dwoma wskaźnikami: pierwszy niech oznacza numer warstwy, do której ów element przynależy, drugi wskaźnik niech oznacza numer kolejny elementu w warstwie. Jeśli badaną cechą jest X, to X_{ij} oznacza wartość cechy j-tego elementu w i-tej warstwie. Wartość średnią cechy X w i-tej warstwie oznaczamy przez $\overline{X}_i$. Mamy więc

$$\overline{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij} . \tag{31}$$

Średnią ogólną rozważanej populacji wyznaczamy ze wzoru

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} X_{ij} . \tag{32}$$

Przy rozważaniu próbki n-elementowej składającej się odpowiednio z $n_1, \ldots, n_k$ elementów poszczególnych warstw otrzymujemy średnie próbkowe dla warstw i średnią próbkową ogólną zamieniając we wzorach (31) i (32) duże litery na małe.

Wariancję cechy X w populacji czyli średnie kwadratowe odchy-

lenie od średniej ogólnej będziemy nazywali wariancją całkowitą i oznaczali przez σ_c^2. Mamy więc

$$\sigma_c^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}})^2. \tag{33}$$

Wariancję wewnątrz i-tej warstwy oznaczamy przez σ_i^2. Mamy więc

$$\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2. \tag{34}$$

Wariancję międzywarstwową definiujemy wzorem

$$\sigma_m^2 = \frac{1}{N} \sum_{i=1}^{k} N_i (\bar{X}_i - \bar{\bar{X}})^2. \tag{35}$$

Między tymi wariancjami zachodzi następujący związek:

$$\sigma_c^2 = \frac{1}{N} \sum_{i=1}^{k} N_i \sigma_i^2 + \sigma_m^2 . \tag{36}$$

Pobieranie próbki według schematu losowania warstwowego odbywa się w ten sposób, że pobieramy z i-tej warstwy jak gdyby z odrębnej populacji próbkę o liczności n_i (według jakich zasad ustalać liczebności n_i powiemy później). Suma liczebności n_i z poszczególnych warstw daje całkowitą liczebność próbki n. Próbki $n_1, \ldots, n_k$ z poszczególnych warstw są pobierane według ustalonego przedtem schematu losowania (niezależnego lub zależnego), tego samego dla wszystkich warstw. Wobec tego w dalszym ciągu będziemy mówić o schemacie losowania warstwowego niezależnego i schemacie losowania warstwowego zależnego.

Jeśli już wylosowaliśmy z poszczególnych warstw próbkę o łącznej liczebności n, to znając N_i (i=1, ..., k) czyli liczebności poszczególnych warstw możemy obliczyć następującą średnią ważoną poszczególnych średnich warstwowych próbkowych:

$$\bar{\bar{x}} = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{x}_i. \tag{37}$$

Posługując się wynikami tw. 1 i tw. 2 można bardzo łatwo wykazać, że wartością oczekiwaną takiej średniej ważonej zarówno przy schemacie losowania warstwowego niezależnego jak i warstwowego zależnego jest średnia arytmetyczna populacji generalnej $\bar{\bar{X}}$ zdefiniowana wzorem (32). Oznacza to, że średnia próbkowa zdefiniowana wzo-

rem (37) jest nieobciążonym estymatorem średniej populacyjnej (32). Można to zapisać przy pomocy następującej równości:

$$(38) \qquad E(\bar{x}) = E\left(\frac{1}{N}\sum_{i=1}^{k}\bar{x}_i N_i\right) = \frac{1}{N}\sum_{i=1}^{k}\bar{X}_i N_i = \bar{X}.$$

§ 3.2. Porównanie losowania nieograniczonego i warstwowego.

Jak już wspomnieliśmy na stronicy 2, schemat losowania jest tym lepszy, im mniejsza jest wariancja estymatora wyznaczanego parametru przy tym schemacie. Jeśli celem naszego badania jest uzyskanie informacji o wielkości $\bar{X}$ czyli średniej generalnej cechy X w rozważanej populacji, to w myśl wzoru (38) za nieobciążony estymator tej wielkości można przyjąć przy losowaniu warstwowym średnią ważoną (37). Aby ocenić dobroć tego estymatora, należy znaleźć jego wariancję czyli $D^2(\bar{x})$. W zależności od schematu losowania warstwowego niezależnego i zależnego rozróżniamy $D^2(\bar{x})_{\text{los. warstw. niez.}}$ i $D^2(\bar{x})_{\text{los. warstw. zal.}}$. W obu przypadkach wariancje te sprowadzają się do znalezienia odpowiednich kombinacji liniowych wariancji średnich próbkowych warstwowych:

$$D^2(\bar{x}) = D^2\left(\frac{1}{N}\sum_{i=1}^{k}N_i\bar{x}_i\right) = \frac{1}{N^2}\sum_{i=1}^{k}N_i^2 D^2(\bar{x}_i).$$

Jeśli każdą warstwę traktować jako odrębną populację, to posługując się odpowiednimi wynikami tw. 3 i tw. 4 otrzymujemy, że

1° przy stosowaniu schematu losowania warstwowego niezależnego

$$(39) \qquad D^2(\bar{x})_{\text{los. warstw. niez.}} = \frac{1}{N^2}\sum_{i=1}^{k}N_i^2\cdot\frac{\sigma_i^2}{n_i};$$

2° przy stosowaniu schematu losowania warstwowego zależnego

$$(39\text{ a}) \qquad D^2(\bar{x})_{\text{los. warstw. zal.}} = \frac{1}{N^2}\sum_{i=1}^{k}N_i^2\cdot\frac{N_i-n_i}{N_i-1}\cdot\frac{\sigma_i^2}{n_i}.$$

Porównamy teraz schemat losowania nieograniczonego zależnego ze schematem losowania warstwowego zależnego.

Przypuśćmy, że z populacji liczącej N elementów pobraliśmy według schematu losowania zależnego bez rozkładania tej populacji na warstwy próbkę liczącą n elementów. Z próbki tej obliczamy średnią arytmetyczną $\bar{x}$, która w myśl tw. 2 jest nieobciążonym estymatorem średniej populacji $\bar{X}$. Zgodnie z wzorem (13) jest $D^2(\bar{x}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$, gdzie σ^2 jest to wariancja całkowita cechy X w badanej

populacji. Wariancja ta została zdefiniowana wzorem (2). Gdyby w rozważanej populacji wyróżnić warstwy o licznościach $N_1,\ldots,N_k$, to wariancję całkowitą cechy X w badanej populacji można przedstawić zgodnie z wzorem (36) jako $\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i\sigma_i^2 + \sigma_m^2$. Wobec tego wariacja estymatora średniej arytmetycznej $\bar{X}$ przy losowaniu nieograniczonym zależnym wyraża się następującym wzorem:

$$(40)\qquad D^2(\bar{x})_{\text{los. nieogr. zal.}} = \frac{N-n}{(N-1)\cdot n}\left[\frac{1}{N}\sum_{i=1}^{k} N_i\sigma_i^2 + \sigma_m^2\right].$$

Wariancję tę porównujemy z wariancją estymatora średniej populacyjnej $\bar{X}$ przy schemacie losowania warstwowego zależnego czyli z wyrażeniem (39 a), obliczając ich różnicę. Jeśli liczebności warstw są tak duże, że można przyjąć przybliżenie $N_i - 1 \approx N_i$ oraz $N-1 \approx N$, to różnicę tę można przedstawić ostatecznie w następującej postaci:

$$(41)\qquad D^2(\bar{x})_{\text{los. nieogr. zal.}} - D^2(\bar{x})_{\text{los. warstw. zal.}} =$$

$$= \frac{1}{N^2}\sum_{i=1}^{k} N_i\cdot\sigma_i^2\left(\frac{N}{n} - \frac{N_i}{n_i}\right) + \frac{N-n}{nN}\sigma_m^2 .$$

Drugi składnik w powyższym wyrażeniu jest nieujemny. Jest on równy zero, gdy wariancja między warstwami równa się zero, a dzieje się to wtedy, gdy wartości średnie poszczególnych warstw są równe wartości średniej całej populacji. Dalej stwierdzamy, że pierwszy składnik w wyrażeniu (41) może być zarówno dodatni jak i ujemny. Jeśli liczebności próbek z poszczególnych warstw wybierzemy tak, aby

$$(42)\qquad n_i = n\cdot\frac{N_i}{N},$$

to mówimy o losowaniu warstwowym proporcjonalnym. W tym przypadku pierwszy składnik w wyrażeniu (41) równa się zero, pozostaje więc tylko drugi składnik zależny od wariancji międzywarstwowej σ_m^2. Wynika stąd, że jeśli stosujemy tzw. losowanie proporcjonalne, to różnica (41) jest tym większa, im większa jest wariancja rozważanej cechy między poszczególnymi warstwami. W tym przypadku losowanie warstwowe jest co najmniej tak efektywne jak losowanie nieograniczone. Wariancja estymatora średniej arytmetycznej przy losowaniu warstwowym proporcjonalnym wyraża się następującym wzorem:

$$(43)\qquad D^2(\bar{x})_{\text{los. warstw. prop.}} = \frac{1}{N^2}\left[\left(\frac{N}{n}-1\right)\sum_{i=1}^{k}\frac{N_i^2\sigma_i^2}{N_i-1}\right].$$

Jeśli liczebności n_i z poszczególnych warstw będziemy wybierać inaczej, to może się zdarzyć, że pierwszy składnik występujący po prawej stronie wzoru (41) będzie nie tylko ujemny, ale nawet przeważy co do bezwzględnej wartości drugi składnik tak, że cała różnica będzie ujemna. Wtedy losowanie warstwowe jest gorsze od losowania nieograniczonego bez warstwowania.

§ 3.3. Optymalny schemat losowania warstwowego.

Wariancja (39a) jest funkcją liczebności próbek z poszczególnych warstw. Przypuśćmy, że mamy z góry ustaloną liczebność próbki n. Zapytajmy się, dla jakiego układu liczb $n_1, \ldots, n_k$ spełniających warunek uboczny $n_1 + \ldots + n_k = n$ wariancja $D^2(\bar{x})_{\text{los. warst. zal.}}$ osiąga minimum. Rozwiązując to zadanie znanymi metodami analizy matematycznej otrzymujemy, że

$$(44)\qquad n_i^2 = Q^2 \cdot \frac{N_i^3\sigma_i^2}{N_i-1},$$

przy czym stałą Q należy dobrać tak, aby $\sum_{i=1}^{n} = n$.

Jeśli liczebności warstw są dostatecznie duże aby liczby N_i-1 można zastąpić przez N_i, to wyrażenie (44) upraszcza się do wzoru

$$(45)\qquad n_i = Q \cdot N_i\sigma_i, \quad \text{gdzie} \quad Q = \frac{n}{\sum_{i=1}^{k} N_i\sigma_i}.$$

Optymalne liczby n_i są więc proporcjonalne do N_i czyli do liczebności warstw i do σ_i czyli do dyspersji wewnątrzwarstwowych badanej cechy.

Z powyższych rozważań wynika praktyczny wniosek, że jeśli znamy liczebności poszczególnych warstw oraz dyspersje wewnątrzwarstwowe badanej cechy lub przynajmniej ich oszacowania na podstawie wstępnej próbki i mamy zadecydować, ile elementów z poszczególnych warstw ma wejść do próbki, to aby otrzymać minimalną wariancję estymatora średniej populacyjnej należy z każdej warstwy pobrać próbkę o liczebności n_i zgodnie z wzorami (44) lub (45).

Jeśli liczebności próbki warstwowej n_i spełniają warunek (44) lub (45), to schemat losowania nosi nazwę optymalnego schematu warstwowego Neymana. Wariancja średniej próbkowej przy takim sche-

macie losowania wynosi wtedy

$$D^2(\bar{x})_{\text{los. warstw. zal. opt.}} = \frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i-n_i}{N_i-1}\cdot\frac{N_i^2Q_i^2}{n_i} =$$

$$= \frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i-n_i}{N_i-1}N_i^2\cdot\frac{N_i-1}{N_i}\cdot\frac{S_i^2}{n_i} = \frac{1}{N^2}\sum_{i=1}^{k}\left(\frac{N_i^2S_i^2}{n_i} - N_iS_i^2\right) =$$

$$= \frac{1}{N^2}\sum_{i=1}^{k}\left(\frac{N_i^2S_i^2}{QN_iS_i} - N_iS_i^2\right) = \frac{1}{N^2}\left(\frac{\sum_{i=1}^{k}N_iS_i}{Q} - \sum_{i=1}^{k}N_iS_i^2\right),$$

i ostatecznie

$$(46)\quad D^2(\bar{x})_{\text{los. warstw. zal. opt.}} = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{i=1}^{k}N_iS_i\right)^2 - \sum_{i=1}^{k}N_iS_i^2\right].$$

§ 3.4. Przykłady.

Przykład 8. Celem badania jest wyznaczenie średniej ilości punktów uzyskiwanych przez dzieci pewnego miasta w wyniku przeprowadzenia określonego testu. Dzieci dzielimy na warstwy w zależności od klasy, do której dziecko uczęszcza. Z wstępnych badań mamy pewne informacje o dyspersji (odchyleniu standardowym) badanej cechy wewnątrz poszczególnych warstw. Dane podano w kolumnach 1-4 poniższej tablicy:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Nr warstwy i	Klasy	Liczebność warstwy N_i	Dyspersja wewnątrz warstwy s_i	N_is_i	QN_is_i	n_i	$N_is_i\cdot s_i$
1	1-2	250	0,3	75	4,34	4	22,5
2	3-4	580	0,7	406	23,45	23	284,2
3	5-6	620	0,8	496	28,65	29	396,8
4	7-8	450	1,0	450	26,00	26	450,0
5	9-10	210	1,2	252	14,56	15	302,4
6	11	90	2,5	225	13,00	13	562,5
Razem		N = 2200		1904	110,00	110	2018,4

Możemy zbadać jedynie n=110 dzieci. Ile dzieci wybrać z każdej warstwy do próbki, aby otrzymać najlepsze oszacowanie nieznanej średniej ogólnej?

Ponieważ mamy oszacowania dyspersji wewnątrzwarstwowych, wy-

bierzemy liczebności n_i według optymalnego schematu Neymana, tj. według wzoru (45). W tym celu obliczamy najpierw iloczyny $N_i s_i$, i obliczamy ich sumę (patrz kolumna 5 tablicy). Następnie obliczamy stałą Q:

$$Q = \frac{n}{\sum_{i=1}^{6} N_i s_i} = \frac{110}{1904} = 0{,}0578.$$

Mnożąc otrzymaną w ten sposób wartość Q = 0,0578 przez odpowiedni iloczyn $N_i s_i$ otrzymujemy szukane liczebności n_i (patrz kolumna 6 tablicy). Liczebności te zaokrąglamy do liczb całkowitych (patrz kolumna 7).

Dla obliczenia wariancji średniej otrzymanej na podstawie tak zlokalizowanej w warstwach próbki należy jeszcze obliczyć $\sum_{i=1}^{6} N_i s_i^2$ (patrz kolumna 8). Podstawiając odpowiednie wyrażenia do wzoru (46) otrzymujemy

$$D^2(\bar{x}) = \frac{1}{2200^2}\left[\frac{1}{110} \cdot 1904^2 - 2018{,}4\right] =$$

$$= \frac{1}{484 \cdot 10^4}\left[32956{,}5 - 2018{,}4\right] = 0{,}0064.$$

Dokładność d tak wyznaczonej średniej wynosi (przyjmując t=2)

$$d = 2\sqrt{D^2(\bar{x})} = 0{,}16$$

Zapytajmy się teraz, jak liczną próbkę należałoby pobrać z rozważanej populacji nie rozkładając jej na warstwy, aby uzyskać tę samą dokładność w oszacowaniu nieznanej średniej tej populacji. Przypuśćmy, że na podstawie wstępnych badań wiadomo, że $s_c^2 = 1{,}4$.

Dla wyznaczenia n użyjemy wzoru (28), podstawiając w nim N = = 2200, $S^2 = s_c^2 = 1{,}4$, d = 0,16, t = 2. Otrzymujemy

$$n = \frac{2200}{1 + \frac{2200 \cdot 0{,}16^2}{2^2 \cdot 1{,}4}} = \frac{2200}{11} = 200.$$

Widzimy więc, że aby otrzymać estymator średniej o dokładności d = 0,16 na poziomie ufności 0,95 należy przy nieograniczonym zależnym schemacie losowania pobrać do próbki n=200 dzieci, natomiast przy losowaniu warstwowym dla otrzymania tej samej dokładności wystarczy próbka złożona z n= 110 dzieci. Zysk z warstwowa-

nia jest tu ogromny. Dzieje się to dlatego, że badana cecha jest znacznie zróżnicowana pomiędzy warstwami.

Przykład 9 (według [10]). Badana populacja (liczby umowne) została podzielona na 8 warstw, których liczebności podano w poniższej tablicy. Biorąc pod uwagę wartości dyspersji s_i z warstw otrzymane z wstępnej próbki należy ustalić optymalną lokalizację próbki przy założeniu, że próbka powinna obejmować n=1250 elementów. Szacujemy $\bar{X}$.

Zgodnie z założoną w § 3.3 teorią optymalnego schematu losowania warstwowego dzielimy n=1250 przez $\sum_{i=1}^{8} N_i s_i = 45000$ i otrzymujemy współczynnik 0,02778, przez który należy pomnożyć liczby z kolumny 4, aby uzyskać liczby losowań z poszczególnych warstw oznaczone przez n_i' i umieszczone w kolumnie 5 tablicy. Wobec tego, że

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Nr warstwy i	N_i	s_i	$N_i s_i$	n_i'	n_i	$N_i s_i^2$
1	4000	2	8000	222	227	16000
2	6000	1	6000	167	170	6000
3	3000	4	12000	333	341	48000
4	200	25	5000	139	142	125000
5	2700	3	8100	225	230	24300
6	5000	0,2	1000	28	28	200
7	100	45	4500	125	100	-
8	4000	0,1	400	11	12	40
Razem	25000		45000	1250	1250	219540

z warstwy nr 7 należałoby wylosować 125 elementów, podczas gdy warstwa ta obejmuje $N_7 = 100$ elementów, trzeba obliczyć nowe liczby losowań z pozostałych siedmiu warstw. Dzieląc łączną liczebność pozostałych siedmiu warstw przez $\sum_{\substack{i=1 \\ i \neq 7}}^{8} N_i s_i$ otrzymujemy

$$Q = \frac{1150}{40500} = 0{,}0284.$$

Mnożąc przez ten współczynnik liczby z kolumny 4 - oczywiście za wyjątkiem wiersza 7 - otrzymujemy szukane liczebności n_i.

Oszacujmy jeszcze wariancję $D^2(\bar{x})$. Przypatrując się wzorowi (39a) spostrzegamy, że składnik siódmej warstwy równa się zero, bo $n_i = N_i$. Jest to intuicyjnie zrozumiałe: badając całą warstwę wyznaczamy średnią arytmetyczną tej warstwy dokładnie, bez żadnego błędu. Wobec tego na wariancję średniej arytmetycznej całej populacji składają się jedynie te warstwy, dla których zostały pomierzone tylko pewne reprezentacje. Wobec tego zgodnie z wzorem (46) wariancja estymatora średniej przy zaproponowanym w tym przykładzie sposobie pobrania próbki warstwowej równa się

$$D^2(\bar{x}) = \frac{1}{25000^2}\left[\frac{1}{1150}(40500)^2 - 219540\right] = 0{,}00193.$$

Przykład 10. Badana populacja składa się z 800 mężczyzn i 1200 kobiet. Celem badania jest wyznaczenie średniego poziomu enzymu X w tej populacji. Z wstępnych badań wiadomo, że wariancja cechy X wynosi w grupie mężczyzn $s_1^2=16{,}00\ j^2$, w grupie kobiet $s_2^2=20{,}25\ j^2$, natomiast w całej badanej populacji wariancja całkowita wynosi $s_c^2 = 25{,}00\ j^2$. Możemy przebadać 100 osób. Który z następujących schematów losowania daje najlepsze oszacowanie średniej populacyjnej $\bar{X}$: 1° losowanie proporcjonalne do liczebności w warstwach; 2° losowanie według optymalnego schematu Neymana; 3° losowanie nieograniczone w całej populacji bez rozbijania jej na warstwy?

Aby odpowiedzieć na to pytanie, porównujemy wariancje estymatorów średnich obliczonych przy tych trzech schematach losowania.

Ad 1°. Dla losowania proporcjonalnego zgodnie ze wzorem (43) mamy

$$D^2(\bar{x}) = \frac{1}{N^2}\left[\left(\frac{N}{n}-1\right)\sum_{i=1}^{2} N_i s_i^2\right] = \frac{1}{4\cdot 10^6}\left[37100\cdot\left(\frac{2000}{100}-1\right)\right] =$$

$$= \frac{7049}{4\cdot 10^4} = 0{,}1762.$$

Ad 2°. Dla losowania według optymalnego schematu Neymana zgodnie z wzorem (46) mamy

$$D^2(\bar{x}) = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{i=1}^{2} n_i s_i\right)^2 - \sum_{i=1}^{2} N_i s_i^2\right] = \frac{1}{4\cdot 10^6}\left[\frac{73960000}{100} - 37100\right] =$$

$$= 702500:(4\cdot 10^6) = 0{,}1756.$$

Ad 3°. Dla losowania nieograniczonego zgodnie z wzorem (17) mamy

$$D^2(\bar{x}) = \frac{s_c^2}{n}\left(1 - \frac{n}{N}\right) = \frac{25}{100}\left(1 - \frac{100}{2000}\right) = 0{,}25\cdot 0{,}95 = 0{,}2375.$$

Widzimy, że najlepszy jest schemat optymalny Neymana, chociaż prawie równie dobry jest schemat losowania proporcjonalnego.

4. Losowanie grupowe (x)

Często populacja generalna, którą badamy, jest podzielona na pewne grupy, niekoniecznie jednakowo liczne. I tak młodzież szkolna jest ugrupowana w klasy, chorzy danego szpitala według sal itp. Przy pobieraniu próby na ogół dość trudno dotrzeć do poszczególnych grup, natomiast jeśli się już do nich dotarło, to stosunkowo łatwo jest zbadać całą grupę. Na przykład jeśli badamy rozkład próchnicy wśród młodzieży szkolnej i do jakiejś szkoły przybędzie ekipa stomatologiczna, to stosunkowo łatwo zbadać przynajmniej jedną klasę.

Przypuśćmy więc, że populacja licząca M elementów jest podzielona na N grup liczących odpowiednio $M_1, M_2, \ldots, M_N$ elementów. Oznaczmy średnią ilość elementów w grupie literą $\bar{M}$. Z definicji mamy

$$\bar{M} = \frac{M}{N} . \tag{47}$$

Podobnie jak przy podziale na warstwy możemy określić dla każdej grupy średnią i wariancję, możemy również zdefiniować wariancję całkowitą S^2 oraz wariancję międzygrupową, a ponadto możemy określić tak zwany współczynnik korelacji wewnątrzgrupowej według następującego wzoru:

$$r_w = \frac{2 \sum_{k=1}^{N} \sum_{l_1=1}^{M_k-1} \sum_{l_2=l_1+1}^{M_k} (X_{kl_1} - \bar{X})(X_{kl_2} - \bar{X})}{N \cdot \bar{M}(\bar{M}-1) \cdot S^2} , \tag{48}$$

gdzie $S^2 = \frac{1}{M-1} \sum_{k=1}^{N} \sum_{l=1}^{M_k} (X_{kl} - \bar{X})^2$

Tak zdefiniowany współczynnik może być zarówno dodatni jak i ujemny. r_w jest dodatnie, gdy wartości badanej cechy obserwowane dla poszczególnych elementów danej grupy są podobne, natomiast

(x) Termin ten przyjęłam za Oktabą według Słownika polsko-rosyjsko-angielskiego statystyki matematycznej i teorii doświadczenia. Zasępa w swej monografii używa terminu "losowanie zespołowe". Perkal używał wyrażenia "losowanie gronowe". Termin angielski: cluster sampling (por. np. [2] lub [4]).

same grupy różnią się znacznie między sobą. r_w jest ujemne, gdy elementy wewnątrz grupy są zróżnicowane. Wyjaśnimy to bliżej na przykładach.

Wyobraźmy sobie, że na podstawie egzaminu wstępnego zbadaliśmy zdolności i wiedzę ubiegających się o przyjęcie kandydatów, a następnie dokonujemy podziału na grupy według ilości otrzymanych punktów: do grupy A przydzielamy najlepszych, do grupy B trochę gorszych itd. aż do ostatniej grupy, w której znajdą się osoby z końcowych miejsc naszej listy. Przy takim doborze osób do grup można się spodziewać, że poziom w poszczególnych grupach będzie dość wyrównany, zaznaczą się natomiast w sposób mniej lub bardziej wyraźny różnice między poszczególnymi grupami. Współczynnik korelacji wewnątrzgrupowej w takiej sytuacji będzie dodatni.

Wyobraźmy sobie teraz inną możliwość: na podstawie wstępnych wyników dzielimy uczniów na klasy według zasady, żeby w każdej grupie znaleźli się najlepsi i najgorsi, inaczej mówiąc, żeby poziom w klasach był taki sam, a poszczególne klasy nie wykazywały specjalnych różnic. W takiej sytuacji współczynnik korelacji wewnątrzgrupowej będzie ujemny.

W losowaniu grupowym jednostkami losowanymi są całe grupy. Po dokonaniu losowania bada się wszystkie elementy wchodzące w skład wylosowanych grup. Przypuśćmy że wylosowaliśmy do próbki n grup o liczebnościach $M_{i_1}, M_{i_2}, \ldots, M_{i_n}$. Nazwijmy te liczebności $m_1, m_2, \ldots, m_n$. Za estymator średniej populacyjnej można przyjąć następujące wyrażenie:

$$\bar{x} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} x_{ij}}{\sum_{i=1}^{n} m_i} . \tag{49}$$

Można pokazać, że dla jednakowo licznych grup wyrażenie określone wzorem (49) jest nieobciążonym estymatorem średniej populacyjnej $\bar{X}$. Jeśli liczebności poszczególnych grup są różne, to wyrażenie (49) jest obciążonym estymatorem średniej populacyjnej $\bar{X}$. Obciążenie to zależy od liczebności poszczególnych grup. Wariancja tego estymatora zależy od wariancji średnich poszczególnych grup.

Losując n grup z populacji liczącej ogółem N zespołów otrzymujemy jako próbkę $n\bar{M}$ elementów, jeśli grupy są jednakowo liczne, i około $n\bar{M}$ elementów, jeśli grupy są o różnej liczebności.

Przypuśćmy teraz, że zamiast losowania grupowego stosujemy losowanie nieograniczone, to znaczy z populacji liczącej M elementów losujemy próbkę składającą się z $n\bar{M}$ elementów. Który z tych sposobów jest efektywniejszy, to jest daje estymator o mniejszej wariancji?

Jeśli liczba grup w badanej populacji jest dostatecznie duża, to można otrzymać wyraźną odpowiedź na to pytanie. Można wykazać, że

$$D^2(\bar{x})_{\text{dla losowania grupowego}} \gtreqless D^2(\bar{x})_{\text{dla losowania nieograniczonego}}$$

w zależności od tego, czy współczynnik korelacji wewnątrzgrupowej

$$r_w \gtreqless 0.$$

Wynika stąd następujący wniosek: jeśli $r_w > 0$, to losowanie grupowe jest mniej efektywne od losowania nieograniczonego indywidualnego. Jeśli $r_w = 0$, to obydwa schematy losowania są tak samo efektywne. Jeśli $r_w < 0$, to losowanie grupowe jest efektywniejsze od nieograniczonego.

Wynik ten nie powinien nas dziwić. r_w dodatnie oznacza, że elementy wewnątrz zespołów są podobne. Jeśli wybierzemy do próbki n elementów podobnych, to mamy gorszą reprezentację badanej populacji niż gdybyśmy wybrali próbkę w sposób losowy z całej populacji. r_w ujemne oznacza, że elementy wewnątrz grup są zróżnicowane. Wobec tego jeśli do próbki weźmiemy elementy wchodzące do grup z założenia zróżnicowanych, to zapewnimy sobie tym samym lepszą reprezentację badanej populacji niż gdybyśmy wybierali do próbki elementy na chybił-trafił.

W przypadku, gdyby podział badanej populacji na grupy odbył się w sposób losowy według cechy nieskorelowanej z badaną cechą, to losowanie grupowe byłoby średnio tak samo efektywne jak losowanie nieograniczone.

Jednak nawet gdy współczynnik korelacji wewnątrzgrupowej jest równy zero, to może się opłacać przeprowadzenie losowania grupowego. Będzie tak w przypadku, gdy koszt losowania nieograniczonego będzie tak wysoki, że przekroczy on wraz z kosztem zebrania informacji statystycznych koszt wylosowania i zebrania materiału w przypadku losowania grupowego tak samo licznej próbki.

Literatura cytowana

[1] Brownlee, K. A., Statistical theory and methodology in science and engineering, New York, 1960.
[2] Cochran, W. G., Sampling techniques, New York 1953.
[3] Fisz, M., Rachunek prawdopodobieństwa i statystyka matematyczna, Warszawa 1958.
[4] Hansen, M. H., Hurwitz, W. N., Madow, W. G., Sample survey methods and theory, vol. I-II, New York 1953.
[5] Sadowski, W., Statystyka matematyczna, Warszawa 1965.
[6] Steinhaus, H., Tablice liczb przetasowanych czterocyfrowych, Rozprawy matematyczne 4 (1954).
[7] Steinhaus, H., Liczby złote i żelazne, Zastosowania matematyki 3 (1956), str. 51-65.
[8] Tablice statystyczne, pod red. W. Sadowskiego, Warszawa 1957.
[9] Vielrose, E., Tablice liczb losowych, Warszawa 1951.
[10] Zasępa, R., Badania statystyczne metodą reprezentacyjną, Warszawa 1962.