Spis treści

1 Wstęp

Metody Reprezentacyjne

Przedmiot

Metoda reprezentacyjna (lub reprezentatywna) jest częściowym badaniem statystycznym opartym na próbie pobranej ze zbiorowości generalnej w sposób losowy. Z teoretycznego i praktycznego punktu widzenia metoda ta jest najbardziej prawidłową formą badania częściowego. W metodzie reprezentacyjnej dokonuje się wyboru próby na dwa sposoby. Może to być wybór przez losowanie, albo przez celową selekcję.

1.1 Literatura

Literatura

Literatura

- Ravindra Singh, Naurang Singh Mangat (1996), *Elements of Survey Sampling*, Originally published by Kluwer Academic Publishers in 1996, Springer Science+Business Media Dordrecht.
- Thompson M.E. (1997), *Theory of Sample Surveys*, Originally published by Chapman & Hall in 1997, Springer-Science+Business Media, B.Y.

1.2 Podstawowe pojęcia

Podstawowe pojęcia

Podstawowe pojęcia

- **Element** jest jednostką badania.
- **Populacja** lub zbiorowość statystyczna jest zbiorem elementów.
- **Jednostkami losowania** są rozłączne podzbiory elementów populacji.
- Wykaz wszystkich elementów populacji nazywamy **operatem losowania** lub krótko **operatem**.
- Podzbiór populacji nazywamy **próbą**.

Podstawowe pojęcia

Podstawowe pojęcia

- Zbiór informacji o każdej jednostce populacji nazywamy **spisem**.
- Liczba jednostek (niekoniecznie rozłącznych), z których składa się próba nazywamy **rozmiarem tej próby** i najczęściej oznaczana jest przez n, natomiast liczbę jednostek populacji nazywamy **rozmiarem populacji** i oznaczamy ją przez N. Iloraz n/N jest nazywany **frakcją losowania**.
- Metodę według której losujemy próbę z populacji, nazywamy **schematem losowania**.

Podstawowe pojęcia

Podstawowe pojęcia

- Jeżeli jednostki w próbie są wybrane zgodnie z pewnym mechanizmem losowym, taki mechanizm nazywam **wyborem probabilistycznym.**
- Procedura wyboru próby bez zastosowania mechanizmu losowego nazywamy **wyborem nieprobabilistycznym**.

Podstawowe pojęcia

Podstawowe pojęcia

- W **losowaniu prostym ze zwracaniem (lpzz)**, jednostki są losowane kolejno z populacji, przy czym po każdym pojedynczym losowaniu jednostka jest zwracana do populacji.
- W **losowaniu prostym bez zwracania (lpbz)**, jednostki są losowane kolejno z populacji, przy czym żadna z wylosowanych jednostek nie jest zwracana do populacji przed następnym losowaniem.

Podstawowe pojęcia

Zadanie 1

W tabeli zamieszczono wagi czterech noworodków (w funtach):

Dziecko	A	B	C	D
Waga	5.5	8.0	6.5	7.0

1 Wypisz wszystkie możliwe próby rozmiaru 2 w schemacie **lpzz**. Przy każdej próbie zapisz wagi odpowiednich jednostek.

2 Wypisz wszystkie możliwe próby rozmiaru 2 w schemacie **lpbz**. Przy każdej próbie zapisz wagi odpowiednich jednostek.

Podstawowe pojęcia

Zadanie 2

Powierzchnie (w hektarach) sześciu wsi przedstawia tabela:

Wieś	A	B	C	D	E	F
Powierzchnia	760	343	657	550	480	935

1 Wypisz wszystkie możliwe próby rozmiaru 3 w schemacie **lpzz**. Przy każdej próbie zapisz powierzchnię odpowiednich jednostek.

2 Wypisz wszystkie możliwe próby rozmiaru 3 w schemacie **lpbz**. Przy każdej próbie zapisz powierzchnię odpowiednich jednostek.

2 Estymator

Populacja

Podstawowe pojęcia

- Populacja $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$
- Cecha $Y : \mathcal{U} \to R$
- Parametr populacji $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$
- Funkcja parametryczna $T : \mathcal{Y} \to R$

Populacja

Funkcja parametryczna (przykłady)

- Wartość globalna $Y = \sum_{i=1}^{N} Y_i$
- Średnia $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$
- Wariancja $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$
- Wartość największa $\max\{Y_i : i =, \ldots, N\}$

Parametry populacji

Parametry

- Każda funkcja rzeczywista określona na jednostkach populacji jest **parametrem populacji** lub krótko **parametrem**.

Niech $Y_1, Y_2, \ldots, Y_N$ będą wartościami cechy Y dla N jednostek populacji.

$$\text{Średnia populacyjna: } \mu = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$\text{Wariancja populacyjna: } \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Statystyka

Estymacja/Oszacowanie

- Każda funkcja rzeczywista określona na jednostkach próby jest **statystyką**. Stosuje się ją do oszacowania parametru populacji i nazywa **estymatorem**.

  Niech $y_1, y_2, \ldots, y_n$ będą wartościami cechy Y w $n-$elementowej próbie.

$$\text{średnia próbkowa: } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\text{wariancja próbkowa: } \hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Plan losowania

Plan losowania

Rozważmy wszystkie próby s będące podzbiorami $\mathcal{U} = \{1, \ldots, N\}$. Oznaczmy przez $\mathcal{S}$ kolekcję wszystkich podzbiorów s zbioru $\mathcal{U}$. **Planem losowania** nazywamy rozkład prawdopodobieństwa na $\mathcal{S}$. Z każdą próbą s związane jest prawdopodobieństwo $p(s)$ wylosowania tej próby. Każde $p(s)$ is wartością z przedziału $[0, 1]$ oraz

$$\sum_{s\in\mathcal{S}} p(s) = 1.$$

Plan losowania

Przykład 1

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

Plan losowania

Prawdopodobieństwo inkluzji

Prawdopodobieństwem pierwszego rzędu dla elementu j nazywamy prawdopodobieństwo wylosowania z populacji elementu j:

$$\pi_j = \sum_{s:\ j \in s} p(s)$$

Plan losowania

Przykład 1 *cd.*

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

π_A	0.5000	0.2427	0.4748	0.5237	0.6046
π_B	0.5000	0.6129	0.3389	0.5958	0.2848
π_C	0.5000	0.4655	0.5925	0.6036	0.4587
π_D	0.5000	0.6788	0.5938	0.2769	0.6518

Plan losowania

Prawdopodobieństwo inkluzji

Prawdopodobieństwem drugiego rzędu dla elementu j oraz k jest prawdopodobieństwo wylosowania elementu j oraz k:

$$\pi_{jk} = \sum_{s:\ j,k \in s} p(s)$$

Plan losowania

Plan losowania

$z_1, \ldots, z_N$: wartości pewnej cechy Z

E_p: wartość oczekiwana według planu losowania p

Dla próby s sumą próbkową Z jest $\sum_{j \in s} z_j$

$$\begin{aligned} E_p\left(\sum_{j \in s} z_j\right) &= E_p\left(\sum_{j=1}^{N} z_j \mathbf{1}(j \in s)\right) \\ &= \sum_{j=1}^{N} z_j E_p \mathbf{1}(j \in s) = \sum_{j=1}^{N} z_j \pi_j \end{aligned}$$

Plan losowania

Plan losowania

- Jeżeli $z_j \equiv 1$ to $\sum_{j \in s} z_j = n(s)$ jest próbą rozmiaru $n(s)$ oraz

$$E_p(n(s)) = \sum_{j=1}^{N} \pi_j$$

- Dla planu losowania o stałym rozmiarze próby n, prawdopodobieństwa pierwszego rzędu sumują się do n

$$E_p n = n = \sum_{j=1}^{N} \pi_j$$

Plan losowania

Przykład 1 *cd.*

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

π_A	0.5000	0.2427	0.4748	0.5237	0.6046
π_B	0.5000	0.6129	0.3389	0.5958	0.2848
π_C	0.5000	0.4655	0.5925	0.6036	0.4587
π_D	0.5000	0.6788	0.5938	0.2769	0.6518

Statystyka

Rozkład prawdopodobieństwa

- Dla danej populacji oraz danego schematu losowania zbiór możliwych wartości estymatora łącznie z odpowiednimi prawdopodobieństwami realizacji tych wartości nazywamy **rozkładem prawdopodobieństwa** tego estymatora.

Plan losowania

Przykład 1 *cd.*

(A,B)	0.1667	0.1195	0.0844	0.2474	0.1355
(A,C)	0.1667	0.0066	0.1718	0.1933	0.1865
(A,D)	0.1667	0.1166	0.2186	0.0830	0.2826
(B,C)	0.1667	0.1951	0.1500	0.2824	0.0261
(B,D)	0.1667	0.2983	0.1046	0.0660	0.1231
(C,D)	0.1667	0.2639	0.2707	0.1280	0.2461

Plan losowania

Przykład 1 *cd.*

średnia próbkowa					
6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6.00	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.50	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461

średnia populacyjna: 6.75

Statystyka

Rozkład prawdopodobieństwa

- Różnicę pomiędzy uzyskaną z próby wartością estymatora a wartością parametru populacji nazywamy **błędem losowym**. Jeżeli θ jest parametrem populacji oraz $\hat{\theta}$ jest jego oszacowaniem, to $\hat{\theta} - \theta$ jest błędem losowym.

Plan losowania

Przykład 1 *cd.*

błąd					
0.00	0.1667	0.1195	0.0844	0.2474	0.1355
-0.75	0.1667	0.0066	0.1718	0.1933	0.1865
-0.50	0.1667	0.1166	0.2186	0.0830	0.2826
0.50	0.1667	0.1951	0.1500	0.2824	0.0261
0.75	0.1667	0.2983	0.1046	0.0660	0.1231
0.00	0.1667	0.2639	0.2707	0.1280	0.2461

Statystyka

Własności

- Estymator $\hat{\theta}$ jest **nieobciążony** dla parametru θ, jeżeli

$$E_p\hat{\theta} = \theta$$

- Jeżeli $E_p(\hat{\theta}) \neq \theta$, to estymator $\hat{\theta}$ jest **obciążony**. Obciążeniem parametru θ nazywamy różnicę

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta$$

Plan losowania

Przykład 1 *cd.*

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461
E_p	6.75	7.008	6.6653	6.7543	6.5743
B_p	0	0.258	-0.0847	0.0043	-0.1757

Statystyka

Podstawowe własności

- **Wariancja** estymatora $\hat{\theta}$ względem planu losowania p:

$$Var_p(\hat{\theta}) = E_p(\hat{\theta} - E\hat{\theta})^2$$

- **Błąd średniokwadratowy** (MSE) mierzy rozbieżność wartości estymatora od rzeczywistej wartości parametru:

$$MSE_p(\hat{\theta}) = E_p(\hat{\theta} - \theta)^2$$

Plan losowania

Przykład 1 *cd.*

6.75	0.1667	0.1195	0.0844	0.2474	0.1355
6	0.1667	0.0066	0.1718	0.1933	0.1865
6.25	0.1667	0.1166	0.2186	0.0830	0.2826
7.25	0.1667	0.1951	0.1500	0.2824	0.0261
7.5	0.1667	0.2983	0.1046	0.0660	0.1231
6.75	0.1667	0.2639	0.2707	0.1280	0.2461
Var_p	0.2708	0.2495	0.2476	0.2372	0.2514
MSE_p	0.2708	0.3161	0.2548	0.2372	0.2823

Statystyka

Podstawowe własności

- Niech $\hat{\theta}_1$ oraz $\hat{\theta}_2$ będą estymatorami parametru θ. **Względną efektywność** estymatora $\hat{\theta}_2$ w stosunku do estymatora $\hat{\theta}_1$ definiujemy następująco

$$RE\left(\hat{\theta}_2|\hat{\theta}_1\right) = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

Plan losowania

Plan losowania

The **Estymator Horvitza-Thompsona** (*estymator HT*) średniej populacyjnej:

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j}$$

Estymator HT jest nieobciążony dla $\mu = \frac{1}{N} \sum_{j=1}^{N} Y_j$

Pokazaliśmy, że $E_p(\sum_{j \in s} z_j) = \sum_{j=1}^{N} z_j \pi_j$

Wystarczy podstawić $z_j = y_j / \pi_j$

Plan losowania

Plan losowania

Plan losowania jest **zrównoważony** jeżeli *prawdopodobieństwa pierwszego rzędu* są sobie równe.

Dla planu losowania, który jest *zrównoważony* oraz *rozmiar próby jest stały* **średnia z próby** jest nieobciążonym estymatorem średniej populacyjnej.

Plan losowania

Plan losowania

Dla próby s: średnia z próby $\bar{y}_s = \frac{1}{n} \sum_{j \in s} y_j$

Ponieważ $n = \sum_{j=1}^{N} \pi_j$ oraz wszystkie π_j są równe: $\pi_j = \frac{n}{N}$

Estymator HT jest *średnią z próby*:

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{j \in s} \frac{y_j}{\pi_j} = \frac{1}{n} \sum_{j \in s} y_j = \bar{y}$$

Statystyka

Zadanie 3

Cztery krowy A, B, C oraz D w gospodarstwie rolnym dawały odpowiednio 5.00, 5.50, 6.00 oraz 6.50 kg mleka dziennie. Wyznaczyć rozkład prawdopodobieństwa średniej $\bar{y}$ liczonej z próby dwóch ($n = 2$) krów, wylosowanych zgodnie ze schematem **lpbz** oraz **lpzz**.

1 Wyznaczyć $P(\bar{y} > 5.9)$.

2 Wyznaczyć wariancję estymatora $\bar{y}$.

3 Wyznaczyć średnią populacyjną μ oraz błąd średniokwadratowy $MSE(\bar{y}) = E(\bar{y} - \mu)^2$.

4 Naszkicować dystrybuantę estymatora $\bar{y}$.

Statystyka

Zadanie 3 *cd.*

Cztery krowy A, B, C oraz D w gospodarstwie rolnym dawały odpowiednio 5.00, 5.50, 6.00 oraz 6.50 kg mleka dziennie. Wyznaczyć rozkład prawdopodobieństwa średniej $\bar{y}$ liczonej z próby dwóch ($n = 2$) krów, wylosowanych zgodnie ze schematem **lpbz** oraz **lpzz**.

5 Sprawdzić czy estymator $\bar{y}$ średniej mleczności krów jest nieobciążony ($E\bar{y} \stackrel{?}{=} \mu$).

6 Niech $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ będzie estymatorem σ^2. Wyznaczyć obciążenie tego estymatora.

Statystyka

Zadanie 4

Oszacowanie błędu średniokwadratowego estymatora $\hat{\theta}_1$ oraz $\hat{\theta}_2$ wynosi odpowiednio 4861.79 oraz 5258.62. Wyznaczyć względną efektywność estymatora $\hat{\theta}_2$ w stosunku do estymatora $\hat{\theta}_1$. Który estymator jest efektywniejszy?

2.1 CTG i przedział ufności

Przybliżony przedział ufności

Centralne twierdzenie graniczne

Niech $\hat{\theta}$ będzie nieobciążonym estymatorem θ o wariancji $Var(\hat{\theta})$

Dla dużych N oraz dużych n

$$\hat{\theta}\text{ma w przybliżeniu rozkład } N\left(\theta, Var(\hat{\theta})\right)$$

równoważnie

$$\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \text{ ma w przybliżeniu rozkład } N(0,1)$$

Przybliżony przedział ufności

Przykład 2

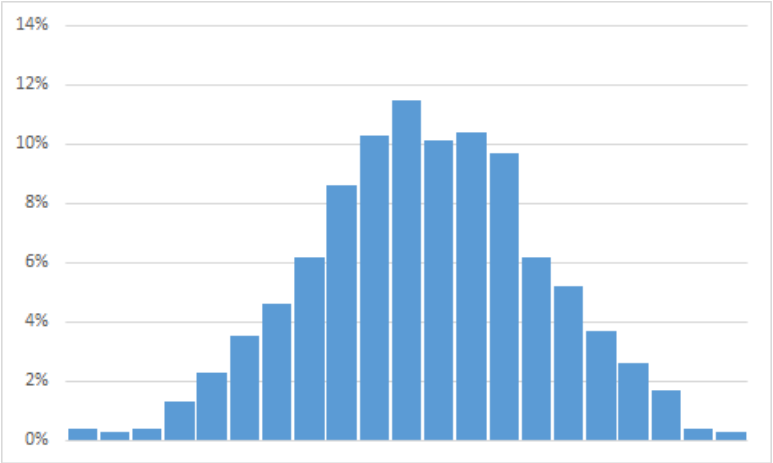
Zpopulacji rozmiaru $N = 10000$ wylosowano 1000 próbek rozmiaru $n = 500$.

Dla każdej próbki wyznaczono wartość estymatora HT.

Z wyznaczonych wartości skonstruowano histogram.

Przybliżony przedział ufności

Przykład 2 *cd.*



Przybliżony przedział ufności

Definicja

Niech $1-\alpha \in (0,1)$ będzie zadanym poziomem ufności.

Wtedy

$$P\left\{\frac{|\hat{\theta}-\theta|}{\sqrt{Var(\hat{\theta})}} \leq u_{1-\frac{\alpha}{2}}\right\} = 1-\alpha$$

gdzie $u_{1-\frac{\alpha}{2}}$ jest $(1-\frac{\alpha}{2})$-kwantylem standardowego rozkładu normalnego.

Przybliżony przedział ufności

Definicja

Przedział ufności dla θ na poziomie ufności $1-\alpha$

$$\left(\hat{\theta} - u_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})}, \hat{\theta} + u_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})}\right)$$

Przybliżony przedział ufności

Definicja

Podstawiając za $Var(\hat{\theta})$ estymator $v(\hat{\theta})$ otrzymujemy

$$\left(\hat{\theta} - u_{1-\frac{\alpha}{2}}\sqrt{v(\hat{\theta})}, \hat{\theta} + u_{1-\frac{\alpha}{2}}\sqrt{v(\hat{\theta})}\right)$$

W szczególności za $1-\alpha$ przyjmujemy 0.95. Wtedy $\alpha = 0.05$ oraz

$$u_{1-\frac{0.05}{2}} = u_{0.975} = 1.96$$

3 Losowanie proste

Losowanie proste

Definicja

Plan eksperymentu nazwiemy **losowaniem prostym** jeżeli wszystkie prawdopodobieństwa pierwszego rzędu są sobie równe:

$$\pi_j = \pi \text{ for all } j \in \{1, 2, \ldots, N\}$$

3.1 Ze zwracaniem

Losowanie proste ze zwracaniem

Średnia próbkowa

Estymator nieobciążony średniej populacyjnej μ

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Losowanie proste ze zwracaniem

Dowód nieobciążoności średniej próbkowej (szkic)

Niech $\mathcal{U} = \{1, \ldots, N\}$ oraz niech $s = (i_1, \ldots, i_n)$ (n jest ustalonym rozmiarem próby). W schemacie **lpzz** $p(s) = 1/N^n$ oraz dla danej próby s mamy

$$\bar{y} = \frac{1}{n}(Y_{i_1} + \cdots + Y_{i_n}) = \frac{1}{n}\sum_{i=1}^{N} Y_i \mathbf{1}(i \in s).$$

Losowanie proste ze zwracaniem

Dowód nieobciążoności średniej próbkowej (szkic)

$$\begin{aligned} nE(\bar{y}) &= \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) = \\ &= \sum_{i=1}^{N} Y_i \sum_{s:\ i\in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i\in s} \frac{1}{N^n} = \\ &= \sum_{i=1}^{N} Y_i n \cdot N^{n-1} \cdot \frac{1}{N^n} = \frac{n}{N}\sum_{i=1}^{N} Y_i = n\bar{Y} \end{aligned}$$

Losowanie proste ze zwracaniem

Wariancja estymatora

Wariancja estymatora $\bar{y}$

$$Var(\bar{y}) = \frac{1}{n}\sigma^2$$

Estymator nieobciążony wariancji estymatora

$$v(\bar{y}) = \frac{1}{n} s^2$$

3.2 Bez zwracania

Losowanie proste bez zwracania

Średnia populacyjna

Estymator nieobciążony średniej populacyjnej μ

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Losowanie proste bez zwracania

Dowód nieobciążoności średniej próbkowej (szkic)

Niech $\mathcal{U} = \{1, \ldots, N\}$ oraz niech $s = (i_1, \ldots, i_n)$ (n jest ustalonym rozmiarem próby). W schemacie **lpbz** $p(s) = 1/\binom{N}{n}$ oraz dla danej próby s mamy

$$\bar{y} = \frac{1}{n}(Y_{i_1} + \cdots + Y_{i_n}) = \frac{1}{n} \sum_{i=1}^{N} Y_i \mathbf{1}(i \in s).$$

Losowanie proste bez zwracania

Dowód nieobciążoności średniej próbkowej (szkic)

$$\begin{aligned} nE(\bar{y}) &= \sum_{i=1}^{N} Y_i E\mathbf{1}(i \in s) = \sum_{i=1}^{N} Y_i P(\{s : i \in s\}) = \\ &= \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} p(s) = \sum_{i=1}^{N} Y_i \sum_{s:\ i \in s} \frac{1}{\binom{N}{n}} = \\ &= \sum_{i=1}^{N} Y_i \cdot \binom{N-1}{n-1} \cdot \frac{1}{\binom{N}{n}} = \frac{n}{N} \sum_{i=1}^{N} Y_i = n\bar{Y} \end{aligned}$$

Losowanie proste bez zwracania

Wariancja estymatora

Wariancja estymatora $\bar{y}$

$$Var(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

Estymator nieobciążony wariancji estymatora

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2$$

Losowanie proste bez zwracania

Wariancja estymatora $\bar{y}$ (wyprowadzenie formuły)

$$Var(\bar{y}) = E\left(\bar{y} - \bar{Y}\right)^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)\right]^2$$
$$= \frac{1}{n^2}\left\{E\left[\sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2\right] + E\left[\sum_{i\neq j}\left(y_i - \bar{Y}\right)\left(y_j - \bar{Y}\right)\right]\right\}$$
$$= \frac{1}{n^2}\left\{\frac{n}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 + \frac{n(n-1)}{N(N-1)}\sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)\right\}$$

Losowanie proste bez zwracania

Wariancja estymatora $\bar{y}$ (wyprowadzenie formuły)

$$\left[\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\right]^2 = \sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 + \sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)$$
$$0 = (N-1)\sigma^2 + \sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)$$
$$\sum_{i\neq j}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right) = -(N-1)\sigma^2$$

Losowanie proste bez zwracania

Wariancja estymatora $\bar{y}$ (wyprowadzenie formuły)

$$\begin{aligned} Var(\bar{y}) &= \frac{1}{Nn}\left\{\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2+\frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)\right\} \\ &= \frac{1}{Nn}\left\{(N-1)\sigma^2+\frac{n-1}{N-1}\sum_{i\neq j}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)\right\} \\ &= \frac{1}{Nn}\left\{(N-1)-(n-1)\right\}\sigma^2 \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2 \end{aligned}$$

Losowanie proste bez zwracania

Estymacja średniej za pomocą niepotarzających się elementów

Jednostki, które są powtórzone w losowaniu, nie wnoszą dodatkowej informacji o szacowanym parametrze. Informacja zawarta w niepowtarzających się elementach jest wystarczająca do oszacowania średniej populacyjnej. Niech $y_1, \ldots, y_d$ odpowiadają d rozłącznym elementom.

Losowanie proste

Estymacja średniej za pomocą niepowtarzających się elementów

Estymator średniej populacyjnej μ: $\bar{y}_d = \frac{1}{d}\sum_{i=1}^{d} y_i$

Wariancja estymatora $\bar{y}_d$: $Var(\bar{y}_d) = \left(E(\frac{1}{d}) - \frac{1}{N}\right)\sigma^2$

Estymator wariancji: $v(\bar{y}_d) = (\frac{1}{d} - \frac{1}{N})s_d^2$,
gdzie $d \geq 2$ oraz $s_d^2 = \frac{1}{d-1}(y_i - \bar{y}_d)^2$

Losowanie proste

Zadanie 5

Z populacji 69 wsi w losowaniu prostym ze zwracaniem pobrano następującą próbę :

Nr wsi	23	28	54	52	49	6	44	30	10	6	53	66	53	56	6
Liczba traktorów	7	21	11	8	38	21	29	59	10	21	12	20	12	8	21

Oszacować przeciętną liczbę traktorów przypadającą na jedną wieś za pomocą estymatora $\bar{y}_d$. Wyznaczyć przedział ufności dla przeciętnej liczby traktorów oraz ich łącznej liczby. Zauważyć, że wieś o numerze 6 została wybrana 3 razy.

3.3 Optymalny rozmiar próby

Optymalny rozmiar próby

Próba wstępna

Niech n_1 będzie rozmiarem próby wstępnej wybranej według schematu **lpbz**. Niech $\bar{y} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ oraz $s^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i - \bar{y})^2$ będą parametrami tej próby wstępnej.

Optymalny rozmiar próby

$$n^\star = \frac{N u^2_{1-\frac{\alpha}{2}} s^2}{N d^2 + u^2_{1-\frac{\alpha}{2}} s^2} = \frac{N u^2_{1-\frac{\alpha}{2}} v^2}{N \delta^2 + u^2_{1-\frac{\alpha}{2}} v^2},$$

gdzie $v = s/\bar{y}$, d jest zadanym błędem dopuszczalnym oraz $\delta = d/\bar{y}$.

Optymalny rozmiar próby

Rozmiar próby $n^\star$ (szkic dowodu)

$$P(|\bar{y} - \mu| < d) = P(\bar{y} - d < \mu < \bar{y} + d) = 1 - \alpha$$

$$P(\bar{y} - u_{1-\frac{\alpha}{2}}\sqrt{Var(\bar{y})} < \mu < \bar{y} + u_{1-\frac{\alpha}{2}}\sqrt{Var(\bar{y})}) \approx 1 - \alpha$$

gdzie $Var(\bar{y}) = (1/n - 1/N)\sigma^2 \approx (1/n - 1/N)s^2$

$$d = u_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)s^2} \;\Rightarrow\; n^\star = \frac{Nu_{1-\frac{\alpha}{2}}^2 s^2}{Nd^2 + u_{1-\frac{\alpha}{2}}^2 s^2}$$

Optymalny rozmiar próby

$N = 1000000$; **rozmiar próby?**

d/S	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
0.01	26343	36994	62221
0.05	1082	1535	2647
0.10	271	384	664
0.20	68	97	166

Zadana precyzja

Rozwiązanie

$$\frac{d}{S} = u_{1-\alpha/2}\sqrt{\frac{1}{n} - \frac{1}{N}}$$

$N = 1000000$; **precyzja estymacji?**

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
100	164.49%	196.00%	257.58%
500	73.56%	87.65%	115.19%
1000	52.01%	61.98%	81.45%
1500	42.47%	50.61%	66.51%

4 Losowanie warstwowe

4.1 Bez powtórzeń

Losowanie warstwowe

Losowanie warstwowe bez powtórzeń

Niech $\mathcal{U}$ będzie sumą rozłącznych warstw $\mathcal{U}_1, \ldots, \mathcal{U}_H$.

Rozmiary $N_1, \ldots, N_H$ warstw są znane: $\sum_{h=1}^{H} N_h = N$.

Dla każdego h losujemy **bez zwracania** próbę rozmiaru n_h z warstwy $\mathcal{U}_h$.

Losowanie warstwowe

Losowanie warstwowe bez powtórzeń

Niebciążony estymator średniej populacyjnej μ:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$$

Losowanie warstwowe

Losowanie warstwowe bez powtórzeń

Wariancja estymatora $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2$$

Nieobciążony estymator wariancji $Var(\bar{y}_{st})$:

$$v(y_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_h^2$$

Losowanie warstwowe

Proporcjonalna alokacja próby w losowaniu warstwowym bez powtórzeń

Niech n będzie całkowitym rozmiarem próby.

Proporcjonalna alokacja próby:

$$n_h = n \cdot W_h = n \cdot \frac{N_h}{N}$$

dla $h = 1, \ldots, H$

Losowanie warstwowe

Proporcjonalna alokacja próby w losowaniu warstwowym bez powtórzeń

Nieobciążony estymator średniej populacyjnej μ:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$$

Losowanie warstwowe

Proporcjonalna alokacja próby w losowaniu warstwowym bez powtórzeń

Wariancja estymatora $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h \sigma_h^2$$

Nieobciążony estymator wariancji $Var(\bar{y}_{st})$:

$$v(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h s_h^2$$

4.2 Optymalna alokacja próby

Losowanie warstwowe

Optymalna alokacja próby w losowaniu warstwowym bez powtórzeń z uwzględnieniem kosztów

Niech budżet eksperymentu wynosi C.

Niech c_h będzie jednostkowym kosztem zaobserwowania Y w $h-$tej warstwie, $h = 1, \ldots, H$

Zagadnienie: znaleźć $n_1, \ldots, n_h$ minimalizujące wariancję

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2$$

przy ograniczeniu

$$\sum_{h=1}^{H} n_h c_h \leq C$$

Losowanie warstwowe

Optymalna alokacja próby w losowaniu warstwowym bez powtórzeń z uwzględnieniem kosztów

Mnożniki Lagrange'a:

$$\mathcal{L}(n_1, \ldots, n_h, \lambda) =$$
$$\sum_{h=1}^{H} \frac{W_h^2 \sigma_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{H} W_h \sigma_h^2 - \lambda \left(C - \sum_{h=1}^{H} n_h c_h \right)$$

$$\frac{\partial \mathcal{L}}{\partial n_h} = -\frac{W_h^2 \sigma_h^2}{n_h^2} + \lambda c_h, \ h = 1, \ldots, H$$

Losowanie warstwowe

Optymalna alokacja próby w losowaniu warstwowym bez powtórzeń z uwzględnieniem kosztów

Układ równań:

$$\begin{cases} -\frac{W_h^2\sigma_h^2}{n_h^2} + \lambda c_h = 0, \ h = 1, \ldots, H \\ C - \sum_{h=1}^{H} n_h c_h = 0 \end{cases}$$

$$n_h^* = \frac{W_h \sigma_h}{\sqrt{\lambda c_h}}$$

Losowanie warstwowe

Optymalna alokacja próby w losowaniu warstwowym bez powtórzeń z uwzględnieniem kosztów

$$\frac{1}{\sqrt{\lambda}} = C \left(\sum_{h=1}^{M} \sqrt{c_h} W_h \sigma_h \right)^{-1}$$

$$n_h = \frac{W_h \sigma_h C}{\sqrt{c_h} \sum_{h=1}^{M} \sqrt{c_h} W_h \sigma_h}, \ h = 1, \ldots, H$$

Losowanie warstwowe

Optymalna alokacja próby w losowaniu warstwowym bez powtórzeń z uwzględnieniem kosztów

Jeżeli $c_h = c$ dla każdego h, to

$$n_h = \frac{C}{c} \cdot \frac{W_h \sigma_h}{\sum_{k=1}^{H} W_k \sigma_k}.$$

Alokacja Neymana

Losowanie warstwowe

Alokacja o zadanej precyzji w losowaniu warstwowym bez powtórzeń

Niech V_0 oznacza zadaną precyzję estymacji

Let c_h będzie jednostkowym kosztem zaobserwowania Y w $h-$tej warstwie, $h = 1, \ldots, H$

Zagadnienie: wyznaczyć $n_1, \ldots, n_h$ minimalizujące koszt

$$\sum_{h=1}^{H} n_h c_h$$

przy ograniczeniu

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) W_h^2 \sigma_h^2 \leq V_0$$

Losowanie warstwowe

Alokacja o zadanej precyzji w losowaniu warstwowym bez powtórzeń

Mnożniki Lagrange'a:

$$\mathcal{L}(n_1, \ldots, n_h, \lambda) =$$
$$\sum_{h=1}^{H} n_h c_h + \lambda \left(\sum_{h=1}^{H} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) W_h^2 \sigma_h^2 - V_0 \right)$$
$$\frac{\partial \mathcal{L}}{\partial n_h} = c_h - \lambda \frac{W_h^2 \sigma_h^2}{n_h^2}, \; h = 1, \ldots, H$$

Losowanie warstwowe

Alokacja o zadanej precyzji w losowaniu warstwowym bez powtórzeń

Układ równań:

$$\begin{cases} c_h - \lambda \frac{W_h^2 \sigma_h^2}{n_h^2} = 0, \ h = 1, \ldots, H \\ \sum_{h=1}^{H} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) W_h^2 \sigma_h^2 - V_0 = 0 \end{cases}$$

$$\frac{1}{n_h^*} = \frac{\sqrt{c_h}}{\sqrt{\lambda} W_h \sigma_h}$$

Losowanie warstwowe

Alokacja o zadanej precyzji w losowaniu warstwowym bez powtórzeń

$$\frac{1}{\sqrt{\lambda}} = \frac{\sum_{h=1}^{H} \frac{1}{N_h} W_h^2 \sigma_h^2 + V_0}{\sum_{h=1}^{H} \sqrt{c_h} W_h \sigma_h}$$

$$n_h^* = \frac{W_h \sigma_h}{\sqrt{c_h}} \frac{\sum_{k=1}^{H} \sqrt{c_k} W_k \sigma_k}{V_0 + \sum_{k=1}^{H} \frac{1}{N_k} W_k^2 \sigma_k^2}, \ h = 1, \ldots, H$$

Losowanie warstwowe

Alokacja o zadanej precyzji w losowaniu warstwowym bez powtórzeń

Jeżeli $c_h = c$ dla dowolnego h, to

$$n_h = W_h S_h \frac{\sum_{k=1}^{H} W_k \sigma_k}{V_0 + \frac{1}{N} \sum_{k=1}^{H} W_k \sigma_k^2}$$

Alokacja Neyman'a

5 Losowanie dwustopniowe

Losowanie dwustopniowe

Wstęp

$$\mathcal{U} = \bigcup_{i=1}^{H} \mathcal{U}_h, \qquad \mathcal{U}_g \cap \mathcal{U}_h = \emptyset \text{ for } g \neq h$$

$$\mathcal{U}_h = \{u_{h1}, \ldots, u_{hN_h}\}$$

$$N_1 + \cdots + N_H = N$$

Losowanie dwustopniowe

Oznaczenia

$$\bar{Y}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} Y_{hj} \qquad \bar{Y} = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{N_h} Y_{hj} = \sum_{h=1}^{H} W_h \bar{Y}_h$$

$$\sigma_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (Y_{hj} - \bar{Y}_h)^2$$

$$\sigma^2 = \frac{1}{N - 1} \sum_{h=1}^{H} \sum_{j=1}^{N_h} (Y_{hj} - \bar{Y})^2$$

Losowanie dwustopniowe

Schemat lpbz+ lpbz

- Krok 1: wylosuj $2 < m < H$ warstw według schematu **lpbz**
- Krok 2: Z każdej warstwy wylosuj próbę według schematu **lpbz**
- Całkowity rozmiar próby

$$n = \sum_{h=1}^{m} n_{(h)}$$

Losowanie dwustopniowe

Schemat lpbz+ lpbz: estymator dwustopniowy

$$\bar{y}_{(2)} = \frac{1}{N}\left[\frac{H}{m}\sum_{h=1}^{m} W_{(h)}\bar{y}_{(h)}\right]$$

Własności

- $E\bar{y}_{(2)} = \bar{Y}$
- $Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{m}\left[\left(1-\frac{m}{H}\right)\sigma_1^2 + \frac{1}{H}\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}\right]$
- $\sigma_1^2 = \frac{1}{H-1}\sum_{h=1}^{H}\left(W_h\bar{Y}_h - \frac{\bar{Y}}{H}\right)^2 = \frac{1}{N^2}\left[\frac{1}{H-1}\sum_{h=1}^{H}\left(Y_h - \frac{Y}{H}\right)^2\right]$

Losowanie dwustopniowe

lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami

Niech C oznacza wielkość ustalonego z góry budżetu.

Niech c_h będzie jednostkowym kosztem obserwacji Y w h-tej warstwie, $h = 1, \ldots, H$

Zagdnienie: wyznaczyć $1 < m < H$ oraz $n_1, \ldots, n_H$ minimalizujące wariancję

$$Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{m}\left[\left(1-\frac{m}{H}\right)\sigma_1^2 + \frac{1}{H}\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}\right]$$

przy ograniczeniu

$$E\left[\sum_{h=1}^{m} n_{(h)}c_{(h)}\right] \leq C$$

Losowanie dwustopniowe

lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

- Założżenie: dla wszystkich warstw $\frac{n_h}{N_h} = f_2 = const$
- $Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{m}\left[\left(1 - \frac{m}{H}\right)\sigma_1^2 + \frac{1-f_2}{f_2}\bar{N}\sigma_2^2\right]$
- $\sigma_2^2 = \sum_{h=1}^{H} W_h\sigma_h^2; \quad \bar{N} = \frac{N}{H}$

Losowanie dwustopniowe

lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

- Jednostkowy koszt wylosowania warstwy: c_1 (constant)
- Jednostkowy koszt obserwacji Y w warstwie: c_2 (constant)
- Koszt doświadczenia $\hat{C} = mc_1 + c_2\sum_{h=1}^{m} n_{(h)}$
- $E(\hat{C}) = m(c_1 + f_2\bar{N}c_2)$
- Ograniczenie: $E(\hat{C}) \leq C$

Losowanie dwustopniowe

lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

Zagadnienie: wyznaczyć m oraz f_2 minimalizujące wariancję

$$Var(\bar{y}_{(2)}) = \frac{1}{N^2}\frac{H^2}{m}\left[\left(1 - \frac{m}{H}\right)\sigma_1^2 + \frac{1-f_2}{f_2}\bar{N}\sigma_2^2\right]$$

przy ograniczeniu

$$m(c_1 + f_2\bar{N}c_2) \leq C$$

Losowanie dwustopniowe

Schemat lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

$$f_2 = \sqrt{\frac{c_1\sigma_2^2}{c_2(\sigma_1^2 - \bar{N}\sigma_2^2)}}$$

$$m = \frac{C}{c_1 + c_2\bar{N}f_2}$$

Zastosować regułę mnożników Lagrange'a

Losowanie dwustopniowe

Schemat lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

$$\frac{\partial\mathcal{L}}{\partial m} = -\frac{H^2}{N^2}\left[\frac{\sigma_1^2}{m^2} + \frac{\bar{N}(1-f_2)}{f_2 m^2}\sigma_2^2\right] - \lambda(c_1 + f_2\bar{N}c_2)$$

$$\frac{\partial\mathcal{L}}{\partial f_2} = -\frac{H^2}{N^2}\frac{\bar{N}\sigma_2^2}{m f_2^2} - \lambda m\bar{N}c_2$$

$$\frac{\partial\mathcal{L}}{\partial m} = 0 \quad \frac{\partial\mathcal{L}}{\partial f_2} = 0$$

Losowanie dwustopniowe

Schemat lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

$$\lambda m\bar{N}c_2 = -\frac{H^2}{N^2}\frac{\bar{N}\sigma_2^2}{m f_2^2}$$

$$\lambda(c_1 + f_2\bar{N}c_2) = -\frac{H^2}{N^2m^2}\left[\sigma_1^2 + \bar{N}\frac{(1-f_2)}{f_2}\sigma_2^2\right]$$

$$\lambda = -\frac{H^2}{N^2}\frac{\sigma_2^2}{c_2m^2f_2^2}$$

Losowanie dwustopniowe

Schemat lpbz+ lpbz: optymalna alokacja próby z ustalonymi kosztami (uproszczenie)

$$f_2 = \sqrt{\frac{c_1\sigma_2^2}{c_2(\sigma_1^2 - \bar{N}\sigma_2^2)}}$$

$$m = \frac{C}{c_1 + c_2\bar{N}f_2}$$

Losowanie dwustopniowe

Schemat lpbz+ lpbz: dla każdej warstwy $f_2 = N_h/N = const$

Estymator nieobciążony σ_2^2**:**

$$\hat{\sigma}_2^2 = \frac{H}{m}\sum_{h=1}^{m} W_h s_h^2$$

Estymator nieobciążony σ_1^2**:**

$$\hat{\sigma}_1^2 = \frac{1}{m-1}\sum_{r=1}^{m}(t_r - \bar{t}_m)^2 - \frac{1-f_2}{f_2}\bar{N}\hat{\sigma}_2^2,$$

gdzie

$$t_r = N_r\bar{y}_r,\ \bar{t}_m = \frac{1}{m}\sum_{r=1}^{m} t_r$$

Losowanie dwustopniowe

Losowanie dwustopniowe

W pierwszym kroku losujemy próbę $\mathcal{L}$ jednostek pierwszego stopnia. Następnie niezależnie dla każdego $r \in \mathcal{L}$ próba s_r rozmiaru $n(s_r)$ jednostek drugiego stopnia pobierana jest z $\mathcal{U}_r$ zgodnie z zadanym schematem losowania. Zgodnie z oznaczeniami całkowity rozmiar próby wynosi $n(s) = \sum_{r\in\mathcal{L}} n(s_r)$, gdzie

$$s = \bigcup_{r\in\mathcal{L}} s_r.$$

Losowanie dwustopniowe

Losowanie dwustopniowe

Prawdopodobieństwa pierwszego rzędu losowania jednostek pierwszego stopnia określa się następująco

$$\Pi_r = P(r \in \mathcal{L})$$

Warunkowe prawdopodobieństwo pierwszego rzędu, t.j. prawdopodobieńnstwo, że j jest w s_r pod warunkiem, że r jest w L określa się przez

$$\pi_{j|r} \text{ dla } r \in \mathcal{U}_r.$$

Warunkowe prawdopodobieństwo pierwszego rzędu π_j może być wyznaczone przez

$$\pi_j = \Pi_r \cdot \pi_{j|r}.$$

Losowanie dwustopniowe

Losowanie dwustopniowe

Estymator HT dla μ jest postaci

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{r\in\mathcal{L}} \frac{\hat{T}_r}{\Pi_r},$$

gdzie $\hat{T}_r = \sum_{j\in s_r} \frac{y_j}{\pi_{j|r}}$

Losowanie dwustopniowe

Losowanie dwustopniowe

- Niech M będzie ustaloną liczbą wylosowanych jednostek pierwszego stopnia. Wtedy
$$\sum_{r=1}^{N} \Pi_r = M$$
- Jeżeli prawdopodobieństwo pierwszego rzędu Π_r jest proporcjonalne do rozmiaru warstwy U_r to
$$\Pi_r = M \cdot \frac{N_r}{N}$$

Losowanie dwustopniowe

Losowanie dwustopniowe

- W pierwszym kroku wybieramy ustaloną liczbę M jednostek losowania pierwszego stopnia ($n(\mathcal{L}) = \mathcal{M}$ dla każdej próby $\mathcal{L}$). Z warstwy U_r losujemy próbę rozmiaru n_r według schematu **lpbz** (dla $r \in \mathcal{L}$). Wtedy

$$Var(\hat{\mu}_{HT}) = \frac{1}{N^2} \sum_{r=1}^{H} \frac{N_r^2}{\Pi_r n_r} \left(1 - \frac{n_r}{N_r}\right) \sigma_r^2 + \\ + \frac{1}{2N^2} \sum_{j \neq k} (\Pi_j \Pi_k - \Pi_{jk}) \left(\frac{T_j}{\Pi_j} - \frac{T_k}{\Pi_k}\right)^2, \quad (1)$$

gdzie $T_j = \sum\limits_{r \in U_j} Y_r$, $\sigma_r^2 = \frac{1}{1-N_r} \sum\limits_{j \in U_r} (Y_j - \bar{Y}_r)^2$, $\bar{Y}_r = T_r / N_r$.

Losowanie dwustopniowe

Losowanie dwustopniowe

- Jeżeli prawdopodobieństwo pierwszego rzędu Π_r jest proporcjonalne do rozmiaru warstwy U_r oraz dla $r \in \mathcal{L}$, próba s_r jest wybrana według losowania prostego bez zwracania, przy n_r losowań z U_r, to

$$\pi_{j|r} = \frac{n_r}{N_r} \text{ and } \pi_j = M \cdot \frac{N_r}{N} \cdot \frac{n_r}{N_r} = M \cdot \frac{n_r}{N}.$$

Losowanie dwustopniowe

Zadanie 6

Pokazać, że w tym przypadku **estymator HT** średniej μ jest postaci $\hat{\mu}_{HT} = \frac{1}{M} \sum_{r \in \mathcal{L}} \bar{y}_r$.

Losowanie dwustopniowe

Losowanie dwustopniowe

W pierwszym kroku wybieramy M jednostek pierwszego stopnia przy założeniu, że prawdopodobieństwo pierwszego rzędu Π_r jest proporcjonalne do rozmiaru warstwy U_r. W drugim kroku z warstwy U_r pobieramy próbę rozmiaru n_r według schematu **lpbz**. Można pokazać

$$Var(\hat{\mu}_{HT}) = \frac{1}{M^2} \sum_{r=1}^{H} \frac{\Pi_r}{n_r} \left(1 - \frac{n_r}{N_r}\right) \sigma_r^2 + \\ + \frac{1}{2N^2} \sum_{j \neq k} (\Pi_j \Pi_k - \Pi_{jk}) \left(\frac{T_j}{\Pi_j} - \frac{T_k}{\Pi_k}\right)^2, \quad (2)$$

gdzie $\Pi_r = \frac{MN_r}{N}$

Losowanie dwustopniowe

W pierwszym kroku wybieramy M jednostek pierwszego stopnia według schematu **lpbz**. W drugim kroku z warstwy U_r pobieramy próbę rozmiaru n_r według schematu **lpbz**. Wtedy

$$Var(\hat{\mu}_{HT}) = \frac{H}{MN^2} \left[(H - M)\sigma_{(1)}^2 + \sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r}\right], \quad (3)$$

gdzie

$$\sigma^2_{(1)} = \frac{1}{2H(H-1)} \sum_{j \neq k} (T_j - T_k)^2 = \frac{1}{H-1} \sum_{j=1}^{H} (T_j - \mu_T)^2,$$

$$\mu_T = \frac{1}{H} \sum_{r=1}^{H} T_r.$$

Losowanie dwustopniowe lpbz+lpbz

Estymator nieobciążony dla $\sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r}$:

$$\frac{H}{M} \sum_{r=1}^{M} N_r(N_r - n_r) \frac{s_r^2}{n_r} \tag{4}$$

Estymator nieobciążony dla $\sigma^2_{(1)}$:

$$\frac{1}{M-1} \sum_{r=1}^{M} (t_r - \bar{t}_M)^2 - \frac{1}{M} \sum_{r=1}^{M} N_r(N_r - n_r) \frac{s_r^2}{n_r}, \tag{5}$$

$$\text{gdzie } t_r = N_r \bar{y}_h, \ \bar{t}_M = \frac{1}{M} \sum_{r=1}^{M} t_r$$

Dowód równania (1).

$$\begin{aligned}
Var(\hat{\mu}_{HT}|\mathcal{L}) &= \frac{1}{N^2} \sum_{r \in \mathcal{L}} \frac{1}{\Pi_r^2} Var(\hat{T}_r|\mathcal{L}) \\
&= \frac{1}{N^2} \sum_{r \in \mathcal{L}} \frac{1}{\Pi_r^2} N_r^2 \left(\frac{1}{n_r} - \frac{1}{N_r} \right) \sigma_r^2 \\
E(Var(\hat{\mu}_{HT}|\mathcal{L})) &= \frac{1}{N^2} E \left(\sum_{r=1}^{H} \frac{1}{\Pi_r^2} N_r^2 \left(\frac{1}{n_r} - \frac{1}{N_r} \right) \sigma_r^2 \mathbb{1}(r \in \mathcal{L}) \right) \\
&= \frac{1}{N^2} \sum_{r=1}^{H} \frac{1}{\Pi_r^2} N_r^2 \left(\frac{1}{n_r} - \frac{1}{N_r} \right) \sigma_r^2 \Pi_r \\
&= \frac{1}{N^2} \sum_{r=1}^{H} \frac{N_r^2}{\Pi_r n_r} \left(1 - \frac{n_r}{N_r} \right) \sigma_r^2
\end{aligned}$$

$$
\begin{aligned}
Var(E(\hat{\mu}_{HT}|\mathcal{L})) &= Var\left(E\left(\frac{1}{N}\sum_{r\in\mathcal{L}}\frac{\hat{T}_r}{\Pi_r}\middle|\mathcal{L}\right)\right) \\
&= Var\left(\frac{1}{N}\sum_{r\in\mathcal{L}}\frac{1}{\Pi_r}E(\hat{T}_r|\mathcal{L})\right) \\
&= Var\left(\frac{1}{N}\sum_{r\in\mathcal{L}}\frac{1}{\Pi_r}T_r\right) \\
&= \frac{M^2}{N^2}Var\left(\frac{1}{M}\sum_{r\in\mathcal{L}}\frac{1}{\Pi_r}T_r\right) \\
&= \frac{M^2}{N^2}\cdot\frac{1}{2M^2}\sum_{j\neq k}(\Pi_j\Pi_k-\Pi_{jk})\left(\frac{T_j}{\Pi_j}-\frac{T_k}{\Pi_k}\right)^2
\end{aligned}
$$

Wiadomo, że

$$Var(\hat{\mu}_{HT}) = E(Var(\hat{\mu}_{HT}|\mathcal{L})) + Var(E(\hat{\mu}_{HT}|\mathcal{L})),$$

□

Dowód równania (2). Wystarczy podstawić $\Pi_r = \frac{MN_r}{N}$ do pierwszej części równania (1) □

Dowód równania (3). Wystarczy podstawić $\Pi_j = \frac{M}{H}$ oraz $\Pi_{jk} = \frac{M(M-1)}{H(H-1)}$ dla każdego j, k, do równania (1) □

Dowód (4) oraz (5)

Niech $s = (s_1, s_2, \ldots, s_H)$.

$$
\begin{aligned}
E\left(\sum_{r\in\mathcal{L}} N_r(N_r - n_r)\frac{s_r^2}{n_r}\right) &= E_s\left(E_{\mathcal{L}}\left(\sum_{r\in\mathcal{L}} N_r(N_r - n_r)\frac{s_r^2}{n_r}\middle| s\right)\right) \\
&= E_s\left(\frac{M}{H}\sum_{r=1}^{H} N_r(N_r - n_r)\frac{s_r^2}{n_r}\right) \\
&= \frac{M}{H}\left(\sum_{r=1}^{H} N_r(N_r - n_r)\frac{E_s s_r^2}{n_r}\right) \\
&= \frac{M}{H}\left(\sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r}\right)
\end{aligned}
$$

$$
\begin{aligned}
E_s t_r^2 = Es\left(\sum_{i\in s_r} N_r\bar{y}_i\right)^2 &= \frac{N_r^2}{n_r^2}E_s\left(\sum_{i\in s_r} y_i\right)^2 \\
&= \frac{N_r^2}{n_r^2}E_s\left(\sum_{i\in s_r} y_i^2 + \sum_{\substack{j\neq k \\ j,k\in s_r}} y_i y_j\right) \\
&= \frac{N_r^2}{n_r^2}\left(\sum_{i=1}^{N_r} y_i^2\pi_i + \sum_{\substack{r \\ j\ \neq\ k}}^{N}\sum_{r}^{N} y_i y_j \pi_{jk}\right)
\end{aligned}
$$

$$= \frac{N_r^2}{n_r^2}\left(\sum_{i=1}^{N_r} y_i^2 \frac{n_r}{N_r} + \sum_{\substack{r \\ j \neq k}}^{N}\sum_{r}^{N} y_i y_j \frac{n_r(n_r-1)}{N_r(N_r-1)}\right)$$

$$= \frac{N_r}{n_r}\left(\sum_{i=1}^{N_r} y_i^2 + (n_r-1)\frac{1}{N_r-1}\sum_{\substack{r \\ j \neq k}}^{N}\sum_{r}^{N} y_i y_j\right)$$

$$= \frac{N_r}{n_r}\left(\sum_{i=1}^{N_r} y_i^2 + (n_r-1)\left(\sum_{i=1}^{N_r} y_i^2 - N_r\sigma_r^2\right)\right)$$

$$= \frac{N_r}{n_r}\left(n_r\sum_{i=1}^{N_r} y_i^2 - (n_r-1)N_r\sigma_r^2\right)$$

$$= \frac{N_r}{n_r}\left(n_r\left((N_r-1)\sigma_r^2 + N_r\mu_r^2\right) - (n_r-1)N_r\sigma_r^2\right)$$

$$= N_r(N_r-n_r)\frac{\sigma_r^2}{n_r} + (N_r\mu_r)^2$$

$$= N_r(N_r-n_r)\frac{\sigma_r^2}{n_r} + T_r^2$$

$E\left(\frac{1}{M-1}\sum_{r\in\mathcal{L}}(t_r-\bar{t})^2\right) =$

$$= E_s\left(E_{\mathcal{L}}\left(\frac{1}{M-1}\sum_{r=1}^{M}(t_r-\bar{t}_M)^2\middle| s\right)\right)$$

$$= E_s\left(\frac{1}{H-1}\sum_{r=1}^{H}(t_r-\bar{t}_H)^2\right)$$

$$= E_s\left(\frac{1}{H}\sum_{r=1}^{H} t_r^2 - \frac{1}{H(H-1)}\sum_{\substack{j \\ j \neq k}}^{H}\sum_{k}^{H} t_j t_k\right)$$

$$
\begin{aligned}
&= \frac{1}{H}\sum_{r=1}^{H} E_s t_r^2 - \frac{1}{H(H-1)}\sum_{j \neq k}^{H}\sum^{H} E_s t_j E_s t_k \\
&= \frac{1}{H}\sum_{r=1}^{H} E_s t_r^2 - \frac{1}{H(H-1)}\sum_{j \neq k}^{H}\sum^{H} T_j T_k \\
&= \frac{1}{H}\sum_{r=1}^{H} \left(N_r(N_r - n_r)\frac{\sigma_r^2}{n_r} + T_r^2 \right) - \frac{1}{H(H-1)}\sum_{j \neq k}^{H}\sum^{H} T_j T_k \\
&= \frac{1}{H}\left\{ \sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r} + \sum_{r=1}^{H} T_r^2 - \frac{1}{H-1}\sum_{j \neq k}^{H}\sum^{H} T_j T_k \right\} \\
&= \frac{1}{H}\left\{ \sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r} + H\sigma_{(1)}^2 \right\} \\
&= \frac{1}{H}\sum_{r=1}^{H} N_r(N_r - n_r)\frac{\sigma_r^2}{n_r} + \sigma_{(1)}^2
\end{aligned}
$$

□

6 Estymacja na podpopulacjach

Przedmiot

Często nie dysponujemy takim operatem losowania, który by zawierał jedynie te jednostki w populacji, które nas interesują. Na przykład jesteśmy zainteresowani pobraniem próby gospodarstw domowych, w których pracują zarówno mąż, jak i żona, albo chcemy chcemy pobrać próbę gospodarstw domowych, których członkami są osoby dorosłe w wieku powyżej 50 lat. Niestety najlepszym dostępnym operatem w obu przypadkach jest lista wszystkich gospodarstw domowych. W takim przypadku przed każdą próbką jednostka jest obserwowana bez możliwości poznania czy jakaś konkretna wybrana jednostka jest członkiem rozważanej pod-populacji czy nie.

Estymacja parametrów podpopulacji

Oznaczenia

- N: rozmiar populacji
- N_1: rozmiar podpopulacji
- n: rozmiar próby według schematu **lpbz**popbranej z populacji
- n_1: liczba jednostek należących do podpopulacji w pobranej próbie
- Y_{si}: wartość badanej cechy Y dla i-tej jednostki należącej do podpopulacji
- y_{si}: wartość badanej cechy Y dla i-tej jednostki należącej do podpopulacji w pobranej próbie

Estymacja parametrów podpopulacji

Oznaczenia

Średnia, wartość globalna oraz Średni błąd kwadratowy dla podpopulacji:

- $\bar{Y}_s = \frac{1}{N_1} \sum_{i=1}^{N_1} Y_{si}$
- $Y_s = \sum_{i=1}^{N_1} Y_{si}$
- $S_s^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (Y_{si} - \bar{Y}_s)^2$

Estymacja parametrów podpopulacji

Estymacja średniej

Let $s_s^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{si} - \bar{y}_s)^2$

- Nieobciążony estyimator średniej podpopulacyjnej $\bar{Y}_s$ kiedy N_1 jest znane:

$$\bar{y}_s = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{si}$$

- Wariancja estymatora $\bar{y}_s$: $Var(\bar{y}_s) = [E(1/n_1) - 1/N_1]S_s^2$
- Estymator wariancji $Var(\bar{y}_s)$: $v(\bar{y}_s) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right) s_s^2$

Uwaga

W przypadku nieznanego N_1 można go zastąpić przez $n_1 N/n$.

Estymacja średniej podpopulacyjnej

Zadanie 7

Departament Planowania Rodziny zamierza przeprowadzić ankietę wśród rodzin, które mieszkają w akademikach i które posiadają dwójkę dzieci. Celem badań jest oszacowanie średniego czasu (w miesiącach) między urodzeniem jednego a drugiego dziecka. Dostępna lista obejmuje 800 rodzin. Ponieważ wcześniejsza identyfikacja rodzin z dwójką dzieci nie była możliwa, postanowiono wybrać losową próbkę 80 rodzin. W pobranej próbce zidentyfikowano 32 rodziny posiadające dwoje dzieci. Na podstawie przeprowadzonych wywiadów zebrano potrzebne informacje, które przedstawia tabela:

Estymacja średniej podpopulacyjnej

Zadanie 7 *cd.*

Rodzina	Różnica	Rodzina	Różnica	Rodzina	Różnica	Rodzina	Różnica
1	24	9	64	17	57	25	42
2	30	10	32	18	65	26	16
3	50	11	58	19	26	27	37
4	41	12	48	20	35	28	61
5	27	13	51	21	31	29	34
6	47	14	22	22	17	30	29
7	47	15	69	23	28	31	19
8	39	16	54	24	55	32	57

Estymacja średniej podpopulacyjnej

Zadanie 7 *cd.*

Oszacować średni czasu między urodzeniem jednego a drugiego dziecka **Uwaga:** $n_1 = 32, n = 80,$ **oraz** $N = 800$