15 等間隔抽出法について

青山博次郎

大きさNなる母集団からn個のサンプルを抽出するとき等間隔抽出法を利用することが実際上よく用いられている。これに関して実用上の注意をのべることにしよう。

§1 $N \neq kn$ なるときの誤差について

通常 $N=kn$ の場合のみが論じられている。そうして

$N=kn-r, \quad 0<r<k$

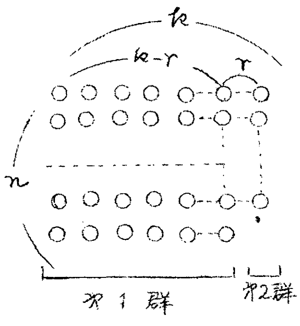
なるときは $k>50$ なるとき

近似的に $\quad E(\bar{x}) = \bar{X}$

但し $\bar{X}$ は母平均とす

が成立するとして議論が進められている。

この理由は容易に分るが念のために計算してみよう。



$1 \leqq i \leqq k$ として，i がランダム・スタートとしてえらばれるとき，サンプルの平均は

$$\bar{x}_i = \frac{x_i + x_{i+k} + \cdots + x_{i+(n'-1)k}}{n'} \qquad (1)$$

但し $\quad n'=n, \quad i \leqq k-r$

$\qquad n'=n-1, \quad i > k-r$

とする。

このとき

$$E(\bar{x}_i) = \frac{1}{nk}\sum_{j=1}^{k-r}(X_j + X_{j+k} + \cdots\cdots + X_{j+(n-1)k}) + \frac{1}{(n-1)k}\sum_{j=1}^{r}(X_{k-r+j}$$

$$+X_{k-r+1+k}+\cdots\cdots+X_{k-r+j+(n-2)k}\Big)$$

$$=\left(1-\frac{r}{k}\right)\overline{X}_1+\frac{r}{k}\overline{X}_2 \qquad (2)$$

ここで $\overline{X}_1$ は $n(k-r)$ 個のもの（第1群と名付ける）の平均，$\overline{X}_2$ は $(n-1)r$ 個の半端もの（第2群と名付ける）の平均とする。

一方母平均は

$$\overline{X}=\frac{\overline{X}_1 n(k-r)+\overline{X}_2 r(n-1)}{nk-r}=\overline{X}_1\frac{k-r}{k-\frac{r}{n}}+\overline{X}_2\frac{r\left(1-\frac{1}{n}\right)}{k-\frac{r}{n}} \qquad (3)$$

それ故 $n>50$ 程度であれば

$$E(\bar{x}_i)\fallingdotseq\overline{X} \qquad (4)$$

となる。

$\bar{x}_i$ の分散については

$$V(\bar{x}_i)=\left(1-\frac{r}{k}\right)\left(\overline{X}_1^2+\sigma_1^2\right)+\frac{r}{k}\left(\overline{X}_2^2+\sigma_2^2\right)-\frac{1}{k}\sum_{i=1}^{k}\overset{*}{\sigma}_i^2-\left\{\overline{X}_1\left(1-\frac{r}{k}\right)+\overline{X}_2\frac{r}{k}\right\}^2 \qquad (5)$$

但し σ_1^2, σ_2^2 はそれぞれ第1, 2群の母分散とし，$\overset{*}{\sigma}_i^2$ は第 i 組の母分散である。

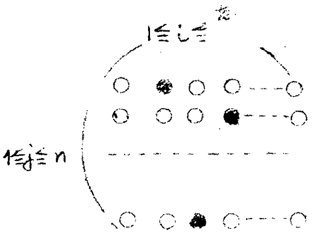
従って Mean Square Error は全体の母分散を σ^2 とするとき

$$M.S.E.(\bar{x}_i)\fallingdotseq\sigma^2-\frac{1}{k}\sum_{i=1}^{k}\overset{*}{\sigma}_i^2 \qquad (6)$$

§2 N=kn の場合で対象外のサンプルを含む場合

我々が適当サンプルを抽出するとき台帳として [illegible] を用いることが多い。例えば選挙人名簿がそれであるが，時々刻々変化しないまでも，調製時から現在までに名簿には訂正が加えられているのが普通である。このとき正確に人数が算えられておれば兎に角，通常の場合は大きさで抽出するために，抽出しやすい5とか10の倍数を用いて等間隔抽出を行い，もし対象外（転出，死亡等）の

の者が当ればこれを省いて，行く方法がとられている。この方法は bias をもっていることは明らかであるが，近似的に等間隔抽出法として取扱っている。これについて考えてみる。



●は対象外

左図の如く k 組に分割されたとして j 行 i 列の単位に対し

$$f(i,j) = 1,\quad \text{対象内のとき}$$
$$= 0,\quad \text{対象外のとき}$$

とおく，また第 i 列内の対象外を除いた単位の数を N_i' とし，$\sum_{i=1}^{k} N_i' = N'$ とおく。

ランダム・スタートとして i がえらばれたとき標本平均は

$$\bar{x}_i' = \frac{x_i f(i,1) + x_{i+k} f(i,2) + \cdots + x_{i+(n-1)k} f(i,n)}{\sum_{j=1}^{n} f(i,j)} \tag{7}$$

このとき

$$E(\bar{x}_i') = \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i' \tag{8}$$

となることは容易に分る。

母平均は

$$\bar{X}' = \frac{N_1'\bar{X}_1' + N_2'\bar{X}_2' + \cdots + N_k'\bar{X}_k'}{N'} \tag{9}$$

であるから，一般には

$$E(\bar{x}_i') \neq \bar{X}'$$

となり unbiased とはならない。

このとき

$$V(\bar{x}_i') = E\Big(\bar{x}_i' - \frac{1}{k}\sum_{j=1}^{k} \bar{X}_j'\Big)^2$$
$$= \frac{1}{k}\sum_{i=1}^{k} (\bar{X}_i' - \bar{X}')^2 + \Big(\bar{X}' - \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i'\Big)^2 \tag{10}$$

従って

$$M.S.E.(\bar{x}_i') = \frac{1}{k}\sum_{i=1}^{k}(\bar{X}_i' - \bar{X}')^2 \tag{11}$$

さて $\bar{x}_i'$ の bias の order を求めるために次のように考えてみる。即ちN個の単位がk組の $N_1', N_2', \cdots, N_k'$ に random に分割されると考えると

$$\frac{1}{k}\sum_{i}\bar{X}_i' = \frac{1}{k}\Big\{\frac{1}{N_1'}(X_{11} + \cdots + X_{1N_1'}) + \frac{1}{N_2'}(X_{21} + \cdots + X_{2N_2'}) + \cdots + \frac{1}{N_k'}(X_{k1} + \cdots + X_{kN_k'})\Big\} \tag{12}$$

であるから、この分母もまた確率変数である。

そこで各項毎に期待値をとると

$$E\Big\{\frac{1}{N_i'}(X_{i1} + \cdots + X_{iN_i'})\Big\} = \frac{E(X_{i1} + \cdots + X_{iN_i'})}{E(N_i')} + E(R_i) = \bar{X}' + E(R_i) \tag{13}$$

ここに R_i は誤差項を示す

従って

$$E\Big(\frac{1}{k}\sum_{i=1}^{k}\bar{X}_i'\Big) = \bar{X}' + E\Big(\frac{1}{k}\sum_{i=1}^{k}R_i\Big) \tag{14}$$

次節でのべる如く

$$E(N_i') = \frac{N'}{k} = \bar{N}' \tag{15}$$

となるから

$$E(R_i) \doteqdot \frac{E(X_{i1} + \cdots + X_{iN_i'})\,V(N_i')}{\{E(N_i')\}^2} = \frac{k^2\bar{X}'}{N'^2}\tau^2 = \frac{\bar{X}'}{\bar{N}'^2}\tau^2$$

但し　$V(N_i') = \tau^2$ (16)

$$\therefore\quad E\Big(\frac{1}{k}\sum_{i=1}^{k}R_i\Big) = \frac{\bar{X}'}{\bar{N}'^2}\tau^2 \tag{17}$$

故に以上の如き仮定の下において $E(\bar{x}_i')$ の平均は

$$\overline{E}(\bar{x}_i') = \bar{X}' + \frac{\bar{X}}{\bar{N}'^2}\tau^2 \tag{18}$$

また

$$V\left(\frac{1}{N_i'}(X_{i1}+\cdots+X_{iN_i'})\right) = \frac{V(X_{i1}+\cdots+X_{iN_i'})}{\{E(N_i')\}^2} + \frac{\bar{X}'^2\tau^2}{\{E(N_i')\}^2}$$

$$- 2\rho_i\tau\bar{X}'\frac{\sqrt{V(X_{i1}+\cdots+X_{iN_i'})}}{\{E(N_i')\}^2}$$

ここで ρ_i は N_i' と $X_i' = \sum_{j=1}^{N_i'} X_{ij}$ の相関係数とする。

然るに若干の計算の後，σ^2 を母分散として

$$V(X_{i1}+\cdots+X_{iN_i'}) = E\left(\sum_{j=1}^{N_i'} X_{ij} - \bar{N}'\bar{X}'\right)^2$$

$$= \frac{\bar{N}'^2(k-1)-\tau^2}{\bar{N}'-1}\cdot\sigma^2 + \tau^2\bar{X}'^2$$

となるから

$$V\left(\frac{1}{N_i'}\sum_{j=1}^{N_i'} X_{ij}\right) = \frac{\bar{N}'^2(k-1)-\tau^2}{\bar{N}'-1}\frac{\sigma^2}{\bar{N}'^2} + 2\tau^2\frac{\bar{X}'^2}{\bar{N}'^2} - 2\rho_i\tau\frac{\bar{X}'}{\bar{N}'^2}\sqrt{\frac{\bar{N}'^2(k-1)-\tau^2\sigma^2}{\bar{N}'-1} + \tau^2\bar{X}'^2} \tag{19}$$

従って $V(\bar{x}_i')$ の平均は

$$\bar{V}(\bar{x}_i') = E(V(\bar{x}_i')) = \frac{k-1}{k}\left(\frac{\bar{N}'^2(k-1)-\tau^2}{\bar{N}'-1}\frac{\sigma^2}{\bar{N}'^2} + 2\tau^2\frac{\bar{X}'^2}{\bar{N}'^2}\right)$$

$$- \frac{2\tau(k-1)\bar{X}'}{k^2\bar{N}'^2}\sqrt{\frac{\bar{N}'^2(k-1)-\tau^2}{\bar{N}'-1}\sigma^2 + \tau^2\bar{X}'^2}\sum_{i=1}^{k}\rho_i$$

$$+ (\bar{x}_i',\ \bar{x}_j' \text{の相関係数を含む項}) \tag{20}$$

近似的に $\tau \fallingdotseq 0$，$\rho_i \fallingdotseq 0$ 等とおけるならば

$$\overline{V}(\bar{x}_i') \fallingdotseq \frac{(k-1)^2}{k} \frac{\sigma^2}{N'-1} \fallingdotseq \frac{\sigma^2}{N'} \qquad (20)'$$

以上の所論を簡単な例によって示してみよう。次の例は昭和26年度の東京都知事選挙に際して参照した豊島区池袋4丁目の基本選挙人名簿からとったものである。実際上の端数は切捨てて，丁度 $N=nk$ （nはサンプル数，kは抽出間隔である）とすると

$$N=2580,\quad N'=2476,\quad k=60,\quad n=43,\quad \overline{N'} \fallingdotseq 41.3$$

このとき対象外の者の数は次のように現われた。

1, 4, 4, 2, 4, 6, 9, 11, 8, 0, 1, 2, 1, 1, 0,
2, 0, 1, 1, 2, 2, 1, 1, 0, 1, 3, 1, 1, 1, 2,
0, 1, 0, 0, 0, 4, 0, 1, 0, 2, 1, 0, 0, 2, 0,
1, 2, 1, 0, 4, 1, 1, 1, 2, 1, 2, 0, 3, 0, 1,

従ってこの数列から $\overline{V}(N_i')=\tau^2$ の推定値として 4.773 が得られる。これを (18) に代入すると $\overline{E}(\bar{x}_i')$ の偏りは $\overline{X}'$ の

$$\frac{\tau^2}{\overline{N'}^2} \times 100\% = 0.284\%$$

程度であることが分る。

§3 自然数の分割数について

前節で対象外の者が含まれたN人の者がk組に分けられる方法が等確率で起ると仮定して $\overline{E}(\bar{x}_i)$ を求めた。その時に用いた性質を以下証明してみよう。記号を簡単にするためしばらくダッシュを除いて書く。

今N個のものをk組（$N \geqq k$とする）に分ける方法は幾通りあるかというと，0個の組も許せば

$P_k(N+k)$

個である。ここで $P_k(m)$ は自然数 m を丁度 k 組の自然数に分ける方法の数を示し

$$\sum_{m,k=0}^{\infty} P_k(m)x^m z^k = \frac{1}{(1-xz)(1-x^2z)\cdots\cdots} \tag{21}$$

$$P_k(m) = \sum_{\ell=1}^{k} P_\ell(m-k) \tag{22}$$

なる性質があることが知られている。(注)

例えば $N=5$, $k=3$ ならば $P_3(8)=5$ となり

$$\begin{aligned} N &= 0+0+5 \\ &= 0+1+4 \\ &= 0+2+3 \\ &= 1+1+3 \\ &= 1+2+2 \end{aligned}$$

なる5通りの分割法がある。しかも我々は各通りの中の1つをサンプルとして抽出したと考えられるので、結局

$$N = N_1 + N_2 + \cdots\cdots + N_k$$

となった時の $E(N_i)$, $V(N_i)$ は $kP_k(N+k)$ 個の N_i の平均や、分散を求めればよいこととなるのである。

そこで 0 の数を $F_{N,k}(0)$, 1 の数を $F_{N,k}(1)$, ……… として表わせば

$$\sum_{x=0}^{N} F_{N,k}(x) = kP_k(N+k) \tag{23}$$

(注) 伏見康治:確率論及統計論, 河出書房, 昭和17年

また容易に分る如く

$$F_{N,k}(0)=(k-1)P_1(N)+(k-2)P_2(N)+\cdots\cdots+P_{k-1}(N) \tag{24}$$

数学的帰納法により

$$\begin{aligned}F_{N,k}(x)= & P_1(N-x)+P_2(N-x)+\cdots\cdots+P_{k-1}(N-x)\\ & +P_1(N-2x)+P_2(N-2x)+\cdots\cdots+P_{k-2}(N-2x)\\ & +\cdots\cdots\cdots\cdots\\ & \cdots\cdots\cdots\cdots\\ & +P_1(N-(k-1)x)+\delta_{N=kx}\\ & \fallingdotseq P_{k-1}(N-k-(x+1))+P_{k-2}(N+k-2(x+1))+\cdots\cdots\\ & +P_1(N+k-(k-1)(x+1))+\delta_{N=kx}\end{aligned} \tag{26}$$

なることが証明できる。ここで勿論 $m<k$ のとき

$P_k(m)=0$ とし、

$$\left.\begin{aligned}\delta_{N=kx}&=1, & N&=kx\\ &=0, & N&\neq kx\end{aligned}\right\}$$

とおく。

これらの関係式から

$$\sum_{x=0}^{N} x\,F_{N,k}(x)=N\,P_k(N+k) \tag{27}$$

なることが証明できる。それは（N, $k+1$）及び（N+1, k）について（27）が成立つとき

$$\sum_{x}^{N+1} x\,F_{N+1,k+1}(x)=\sum_{x}^{N+1} x\left(F_{N+1,k}(x)+F_{N-k,k+1}(x-1)\right)$$

$$= (N+1)P_k(N+k+1) + (N+1)P_{k+1}(N+1)$$

$$= (N+1)P_{k+1}(N+k+2)$$

となることより，二重帰納法を用いて証明できる。

故に分割数の組成数たる $kP_k(N+k)$ 個のものの平均は

$$\frac{\sum_{x}^{N} x F_{N,k}(x)}{kP_k(N+k)} = \frac{N}{k} \tag{28}$$

と成る。

分散については一般式が得られないので評価式を用いることにする。このため以下再び N' の記号を用いる。

先ず我々の問題の性質上，N' 個のものを k 組に分けるのであるが一組の個数は n（但し $N=nk$）より大きくはなり得ない。従って（28）式は変更を要する訳であるが，このとき $n+1$ 個以上のものを含む場合が A 通り出来るとすると，容易に分る如く

$$\text{平均} = \frac{N'P_k(N'+k) - N'A}{kP_k(N'+k) - kA} = \frac{N'}{k} \tag{28$'$}$$

となって平均は変らない。

分散については上述の A 通りを除いたものの F を F^* でおきかえると

$$\sum_{x=0}^{n} x^2 F^*_{N,k}(x) < n^2 \sum_{x=0}^{n} F^*_{N,k}(x)$$

$$\therefore\ \tau^2 = \frac{\sum x^2 F^*_{N,k}(x)}{\sum F^*_{N,k}(x)} - \frac{N'^2}{k^2} < n^2 - \frac{N'^2}{k^2} = \frac{N'^2}{k^2}\left(\frac{N^2}{N'^2} - 1\right) \tag{29}$$

故に前節の $\bar{E}(\bar{x}'_i)$ の bias の $\bar{X}'$ に対する相対割合は

$$\frac{\tau^2}{\bar{N}'^2} < \frac{N^2}{N'^2} - 1 \qquad (30)$$

となる。

池袋4丁目の例にあてはめると

$$N' = 2476 = 0.96N$$

であるから

$$\frac{\tau^2}{\bar{N}'^2} < 0.085$$

§4. サンプル数 n を確保する場合

§2で対象外を除いたまま平均を確定した。しかし問題によってはサンプル数 n を確保しなければならない場合がある。このとき欠けたものを残りのものから random にえらぶ場合を考えよう。即ち

第 i 列がえらばれたとき，残りの $N' - N'_i$ 個のものからサンプルを $n - N'_i$ 個えらぶのである。このときの推定平均を $\bar{x}''_i$ とすると

$$\bar{x}''_i = \frac{1}{n}\left(x_i f(i,1) + x_{i+k} f(i,2) + \cdots\cdots + x_{i+(n-1)k} f(i,n)\right)$$

$$+ \frac{1}{n}\left(x_{j1} + x_{j2} + \cdots + x_{j,n-N'_i}\right) \qquad (31)$$

このとき

$$E(\bar{x}_i'')=\frac{1}{kn}\left(\sum_{i=1}^{k}N_i'\bar{X}_i'+\sum_{i=1}^{k}\frac{n-N_i'}{N'-N_i'}\sum{}'N_j'\bar{X}_j'\right)$$

$$=\frac{1}{kn}\left\{N'\bar{X}'+\sum_i\frac{n-N_i'}{N'-N_i'}(N'\bar{X}'-N_i'\bar{X}_i')\right\} \tag{32}$$

ここで $\sum{}'$ は i を除いて加えることを示す。

当然ながら $\bar{X}=\bar{X}_i'$ ならば

$$E(\bar{x}_i'')=\bar{X}' \tag{33}$$

また $N_i'=\bar{N}'$ であっても (33) は成立する。

一般には (32) は $\bar{X}'$ に等しくないから $\bar{x}_i''$ は bias をもっている。その値を B_2 とし，§2 で得た $\bar{x}_i'$ の bias を B_1 とする。

このとき

$$B_1=\frac{1}{k}\sum_i\bar{X}_i'-\bar{X}'=\sum_i\bar{X}_i'\left(\frac{1}{k}-\frac{N_i'}{N'}\right)$$

$$B_2=\frac{N'}{N}\bar{X}'\left(1+\sum_i\frac{n-N_i'}{N'-N_i'}\right)-\frac{1}{N}\sum_i\frac{n-N_i'}{N'-N_i'}N_i'\bar{X}_i'-\bar{X}'$$

もし $N_i'\fallingdotseq\frac{N'}{k}$ とおけるならば

$$B_2\fallingdotseq\frac{N-N'}{N(k-1)}\sum_i\bar{X}_i'\left(\frac{N_i'}{N'}-\frac{1}{k}\right)$$

となるから $|B_2|\leqq|B_1|$

となる．

簡単な例をあげてみるとその様子が分る。

$$
\begin{array}{c|ccccc}
 & & & i & & \\
\hline
 & 1 & \times & 1 & 6 & 4 \\
 & 8 & 2 & \times & 8 & 1 \\
j & 0 & 5 & 2 & 9 & 1 \\
 & 9 & 4 & 5 & \times & 7 \\
 & 6 & 0 & 9 & 4 & 1
\end{array}
$$

（但し　$\times$ は対象外）

$N = 25,\quad N' = 22,\quad n = 5,\quad k = 5$

$\overline{X}' = 4.2272$

$E(\overline{x}_i') = 4.27,$

$E(\overline{x}_i'') = 4.2177,$

$|B_1| = 0.0428$

$|B_2| = 0.0095$

$\overline{x}_2''$ の分散は $N_i'\overline{x}_i'$,　$(n-N_i')\overline{x}_{n-N_i'}$ の相関係数を ρ とおくとき

$$
\begin{aligned}
V(\overline{x}_i'') = & \frac{1}{nN}\sum_i N_i'^2\overline{X}_i'^2 - \frac{N'^2}{N^2}\overline{X}'^2 + \frac{1}{n^2}\left\{\frac{1}{k}\sum_i\left(\frac{n-N_i'}{N'-N_i'}(N'\overline{X}'-N_i'\overline{X}_i')\right)^2\right. \\
& \left. -\left(\frac{1}{k}\sum_i\frac{n-N_i'}{N'-N_i'}(N'\overline{X}'-N_i'\overline{X}_i')\right)^2\right\} + \frac{N'-n}{nN}\sum_i\frac{n-N_i'}{N'-N_i'}\overline{\sigma}_i^2 \\
& + 2\rho\sqrt{\frac{1}{nN}\sum N_i'^2\overline{X}_i'^2 - \frac{N'^2}{N^2}\overline{X}'^2}\sqrt{\frac{N'-n}{nN}\sum_i\frac{n-N_i'}{N'-N_i'}\overline{\sigma}_i^2} \\
= & \frac{1}{n^2}\left\{V(N_i'\overline{X}_i') + V\left(\frac{n-N_i'}{N'-N_i'}(N'\overline{X}'-N_i'\overline{X}_i')\right) + \frac{N'-n}{k}\sum_i\frac{n-N_i'}{N'-N_i'}\overline{\sigma}_i^2\right. \\
& \left. + 2\rho\sqrt{V(N_i'\overline{X}_i')}\sqrt{\frac{N'-n}{k}\sum_i\frac{n-N_i'}{N'-N_i'}\overline{\sigma}_i^2}\right\} \qquad (34)
\end{aligned}
$$

となる

但し $\overline{\sigma}^2$ は　第 i 組を除いたものの分散を示す。

このときも $\frac{N_i'}{N} \fallingdotseq \frac{1}{k}$ とおければ

$$V(\bar{x}_i'') = \frac{N^2(k-1)^2 - (N-N')^2}{N^2(k-1)^2} V(\bar{x}_i')$$

$$+ \frac{(N-N')(N'-n)}{nNN'(k-1)} \sum \bar{\sigma}_i^2 + (\rho \text{ を含む項})$$

となり第2，3項が第1項に比べて無視できる場合は

$$V(\bar{x}_i'') < V(\bar{x}_i')$$

となるから M.S.E. は $\bar{x}_i''$ の方が小さくなることが分る。

前の数値例では

$$V(\bar{x}_i') = 2.1806, \quad V(\bar{x}_i'') = 1.6797$$

となり，

$$\text{M.S.E.}(\bar{x}_i') \fallingdotseq 2.1824$$

$$\text{M.S.E.}(\bar{x}_i'') \fallingdotseq 1.6798$$

となって $\bar{x}_i''$ の方が良い推定値を与える。