標本データの抽出、標本平均

1　有限母集団と無限母集団

母集団の構成要素、これから調べようとする対象の個数が有限個か無限個によって、有限母集団、無限母集団とよぶ。ただし実際には無限母集団というときには、標本の抽出を繰り返して何度でもできる場合が無限母集団の議論で、繰り返しによって、調査の対象の状況が変わってくる場合が有限母集団である。しかし理論的ではなく、近似的に大きな数、たとえば　N が大きく $1/N$ がゼロと近いとみなせる場合には、標本の抽出の前後でによって変化しないと考えられるから、有限母集団と無限母集団は同じと考えられる。これに対して、抽出の前後で母集団が変わってくるとき、復元抽出と非復元抽出の区別をする必要がでてくる。

いま母集団には 2 種類の属性 (♡と♣) がそれぞれ M,N 個あるものとしてこれから標本を取り出す。番号をつけて区別するとして、起こりえるあらゆる場合を表にまとめてみる。表の行は最初の取り出しで、横の列は 2 回目の取り出しの結果を表すとしよう。まず母集団は $\overbrace{\clubsuit 1\ \clubsuit 2\ \cdots \clubsuit M}^{M}\overbrace{\heartsuit 1\ \heartsuit 2\ \cdots \heartsuit N}^{N}$　　$\clubsuit k, k = 1, 2, \cdots M$: $\heartsuit j, j = 1, 2, \cdots N$ であり、2 個の標本をとりだしたときから考える。

[復元抽出]

	$\clubsuit 1$	$\clubsuit 2$	$\cdots$	$\clubsuit M$	$\heartsuit 1$	$\cdots$	$\heartsuit N$
$\clubsuit 1$	$\clubsuit 1\clubsuit 1$	$\clubsuit 1\clubsuit 2$		$\clubsuit 1\clubsuit M$	$\clubsuit 1\heartsuit 1$		$\clubsuit 1\heartsuit N$
$\clubsuit 2$	$\clubsuit 2\clubsuit 1$	$\clubsuit 2\clubsuit 2$		$\clubsuit 2\clubsuit M$	$\clubsuit 2\heartsuit 1$		$\clubsuit 2\heartsuit N$
$\vdots$							
$\clubsuit M$	$\clubsuit M\clubsuit 1$	$\clubsuit M\clubsuit 2$		$\clubsuit M\clubsuit M$	$\clubsuit M\heartsuit 1$		$\clubsuit M\heartsuit N$
$\heartsuit 1$	$\heartsuit 1\clubsuit 1$	$\heartsuit 1\clubsuit 2$		$\heartsuit 1\clubsuit M$	$\heartsuit 1\heartsuit 1$		$\heartsuit 1\heartsuit N$
$\vdots$							
$\heartsuit N$	$\heartsuit N\clubsuit 1$	$\heartsuit N\clubsuit 2$		$\heartsuit N\clubsuit M$	$\heartsuit N\heartsuit 1$		$\heartsuit N\heartsuit N$

要素の個数

1 回目 ＼ 2 回目	♣ 1	♣ 2	$\cdots$	♣ M	♡ 1	$\cdots$	♡ N
♣ 1 ♣ 2 $\vdots$ ♣ M	M^2				MN		
♡ 1 $\vdots$ ♡ N	MN				N^2		

2 個の結果を属性でまとめて合計してみると

$$\left.\begin{array}{ll} \clubsuit\clubsuit & : M^2 \\ \clubsuit\heartsuit + \heartsuit\clubsuit & : 2MN \\ \heartsuit\heartsuit & : N^2 \end{array}\right\} = (M+N)^2 = (M+N)(M+N)$$

つぎに非復元では 1 回目と 2 回目に同じものはないから、つまり非復元の場合には対角線の部分を除く。
[非復元抽出] 表中の × は起こらないことを意味している

	$\clubsuit 1$	$\clubsuit 2$	$\cdots$	$\clubsuit M$	$\heartsuit 1$	$\cdots$	$\heartsuit N$
$\clubsuit 1$	×	$\clubsuit 1\clubsuit 2$		$\clubsuit 1\clubsuit M$	$\clubsuit 1\heartsuit 1$		$\clubsuit 1\heartsuit N$
$\clubsuit 2$	$\clubsuit 2\clubsuit 1$	×		$\clubsuit 2\clubsuit M$	$\clubsuit 2\heartsuit 1$		$\clubsuit 2\heartsuit N$
$\vdots$							
$\clubsuit M$	$\clubsuit M\clubsuit 1$	$\clubsuit M\clubsuit 2$		×	$\clubsuit M\heartsuit 1$		$\clubsuit M\heartsuit N$
$\heartsuit 1$	$\heartsuit 1\clubsuit 1$	$\heartsuit 1\clubsuit 2$		$\heartsuit 1\clubsuit M$	×		$\heartsuit 1\heartsuit N$
$\vdots$							
$\heartsuit N$	$\heartsuit N\clubsuit 1$	$\heartsuit N\clubsuit 2$		$\heartsuit N\clubsuit M$	$\heartsuit N\heartsuit 1$		×

したがって要素の個数は

	$\clubsuit 1$ $\clubsuit 2$ $\cdots$ $\clubsuit M$	$\heartsuit 1$ $\cdots$ $\heartsuit N$
$\clubsuit 1$ $\clubsuit 2$ $\vdots$ $\clubsuit M$	$(M)_2$	MN
$\heartsuit 1$ $\vdots$ $\heartsuit N$	MN	$(N)_2$

合計は

$$\left.\begin{array}{ll} \clubsuit\clubsuit & :(M)_2 \\ \clubsuit\heartsuit+\heartsuit\clubsuit & :2MN \\ \heartsuit\heartsuit & :(N)_2 \end{array}\right\} = (M+N)_2 = (M+N)(M+N-1)$$

さらに 3 回くり返しを考えると、つぎの表の行には 2 回目までの属性による合計結果を表し、横の列は 3 回目の結果で分けてみる。

[復元抽出] での要素の数とその合計は

2 回目まで ＼ 3 回目	$\clubsuit$	$\heartsuit$
$\clubsuit\clubsuit : M^2$	M^3	M^2N
$\clubsuit\heartsuit : 2MN$	$2M^2N$	$2MN^2$
$\heartsuit\heartsuit : N^2$	MN^2	N^3

$$\left.\begin{array}{ll} \clubsuit+\clubsuit+\clubsuit & :M^3 \\ \clubsuit+\clubsuit+\heartsuit & :3M^2N \\ \clubsuit+\heartsuit+\heartsuit & :3MN^2 \\ \heartsuit+\heartsuit+\heartsuit & :N^3 \end{array}\right\} = (M+N)^3$$

[非復元抽出] での要素の数とその合計は

2 回目まで ＼ 3 回目	$\clubsuit$	$\heartsuit$
$\clubsuit\clubsuit : (M)_2$	$(M)_3$	$(M)_2N$
$\clubsuit\heartsuit : 2MN$	$2(M)_2N$	$2M(N)_2$
$\heartsuit\heartsuit : (N)_2$	$M(N)_2$	$(N)_3$

$$\left.\begin{array}{ll} \clubsuit+\clubsuit+\clubsuit & :(M)_3 \\ \clubsuit+\clubsuit+\heartsuit & :3(M)_2N \\ \clubsuit+\heartsuit+\heartsuit & :3M(N)_2 \\ \heartsuit+\heartsuit+\heartsuit & :(N)_3 \end{array}\right\} = (M+N)_3$$

以上の計算から、一般に n 回くり返した場合には
「復元抽出」

$$\left.\begin{array}{ll}
\overbrace{\clubsuit+\clubsuit+\cdots+\clubsuit}^{n} & :\binom{n}{0}M^n \\
\overbrace{\clubsuit+\cdots+\clubsuit}^{n-1}+\heartsuit & :\binom{n}{1}M^{n-1}N \\
\overbrace{\clubsuit+\cdots+\clubsuit}^{n-k}+\overbrace{\heartsuit+\cdots+\heartsuit}^{k} & :\binom{n}{r}M^{n-k}N^k \\
\overbrace{\heartsuit+\heartsuit+\cdots+\heartsuit}^{n} & :\binom{n}{n}N^n
\end{array}\right\}=(M+N)^n$$

「非復元抽出」

$$\left.\begin{array}{ll}
\overbrace{\clubsuit+\clubsuit+\cdots+\clubsuit}^{n} & :\binom{n}{0}(M)_n \\
\overbrace{\clubsuit+\cdots+\clubsuit}^{n-1}+\heartsuit & :\binom{n}{1}(M)_{n-1}N \\
\overbrace{\clubsuit+\cdots+\clubsuit}^{n-k}+\overbrace{\heartsuit+\cdots+\heartsuit}^{k} & :\binom{n}{k}(M)_{n-k}(N)_k \\
\overbrace{\heartsuit+\heartsuit+\cdots+\heartsuit}^{n} & :\binom{n}{n}(N)_n
\end{array}\right\}=(M+N)_n$$

確率変数 X を抽出における標本での n 個のうち ♣ の個数を表すとすると $k=0,1,2\cdots,n$ に対して

$$\begin{aligned}
P(X=k) &= P(\overbrace{\clubsuit+\cdots+\clubsuit}^{k}+\overbrace{\heartsuit+\cdots+\heartsuit}^{n-k}) \\
&= \begin{cases} \binom{n}{k}\dfrac{M^kN^{n-k}}{(M+N)^n} & :\text{復元抽出} \\ \binom{n}{k}\dfrac{(M)_k(N)_{n-k}}{(M+N)_n} & :\text{非復元抽出} \end{cases}
\end{aligned}$$

復元抽出はパラメータ $n, p=\dfrac{M}{M+N}$ の 2 項分布 $Binom\left(n,\dfrac{M}{M+N}\right)$ であり、

$$f(k)=\binom{n}{k}p^k(1-p)^{n-k}, k=0,1,2,\cdots,n$$

また非復元抽出の場合には超幾何分布 $Hyper(M+N,M,n)$ と呼ばれ、

$$f(k)=\binom{n}{k}\frac{(M)_k(N)_{n-k}}{(M+N)_n}=\frac{\binom{M}{k}\binom{N}{n-k}}{\binom{M+N}{n}}, k=0,1,2,\cdots,n$$

と表されることが多い。

これを比較すればわかるよう、べき乗の数と組合せの数が対応している。復元抽出では $M^n=\overbrace{M\cdot M\cdots M}^{n}$ に対して非復元抽出では $(M)_n=\overbrace{M(M-1)\cdots(M-n+1)}^{n}$ となっていることに注意する。

2　標本平均の平均と分散

標本データから求められる基本的な統計量としては、標本平均がある。単純に算術平均を求めることである。いま母集団には 2 種類の属性があるときに、標本平均に対する平均（期待値）と分散を求める。属性が 2 種類しかないから、標本平均というより、標本比率あるいは標本割合ということが多い。つまり母集団の属性が ♣, ♡ として、i 番目 $(i=1,2,\cdots,n)$ に抽出された標本に対して $X_i=\begin{cases} 1 & \text{if} \quad \clubsuit\text{が抽出されたとき} \\ 0 & \text{if} \quad \heartsuit\text{が抽出されたとき} \end{cases}$

とおくと、これらの和 $X = X_1 + X_2 + \cdots + X_n$ の分布は復元抽出ならば、2 項分布 $Binom\left(n, \dfrac{M}{M+N}\right)$ であって、もし非復元抽出ならば、超幾何分布 $Hyper(M+N, M, n)$ にしたがう。したがってこれを利用すれば、標本平均（標本比率）

$$\overline{X}_n = \frac{X}{n} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$

の平均と分散が求められる。

$$\text{復元抽出:}\quad E(\overline{X}_n) = \frac{M}{M+N}, \qquad V(\overline{X}_n) = \frac{1}{n}\frac{M}{M+N}\frac{N}{M+N}$$

$$\text{非復元抽出:}\quad E(\overline{X}_n) = \frac{M}{M+N}, \quad V(\overline{X}_n) = \frac{1}{n}\frac{M}{M+N}\frac{N}{M+N}\frac{M+N-n}{M+N-1}$$

非復元抽出の分散で $\dfrac{M+N-n}{M+N-1}$ は有限補正項とよばれるが、標本数 n に比べて母集団サイズが大きく $M+N$ が大とすると、この値は 1 に近いとみなせるから、復元と非復元の違いはなくなってくることが分かる。

非復元抽出で平均 $E(X_i) = p$, 分散 $V(X_j) = p(1-p)$ ただし $p = M/(M+N)$ を示し、また共分散 $Cov(X_i, X_j) = P(X_j = 1)P(X_j = 1|X_i = 1) - p^2$ を計算せよ。