4．Ewensのサンプリング公式、ポアソン・ディリクレ分布とポリアの壺様モデル

4．1 Ewensのサンプリング公式

生物集団においては一つの種内においても多くの遺伝的多様性が存在する。古くはエンドウ豆の色、ショウジョウバエの羽の変異、染色体の逆位、血液型など表現形として明確に識別できるものに限られていたが、近年、分子レベルの解析の進歩とともに、タンパク質のアミノ酸配列の多型、遺伝子のDNA塩基配列の個人差まで識別可能となってきた。生物集団内の遺伝的多様性の推定は集団遺伝学の主要な問題の一つである。遺伝的多様性の源は種々の突然変異であり、DNA塩基配列がヒトでは３０億塩基対ということを考えると、生じる突然変異はほとんど常に新しいタイプと考えられる。このように突然変異は常に集団中にこれまで存在しなかった新しいタイプと仮定するモデルを無限対立遺伝子モデルという。Ewens(1972)は中立な無限対立遺伝子モデルの下で、集団からランダムに取り出したn個の遺伝子サンプルの中に存在する異なる対立遺伝子（アレル）の数、およびサンプル内アレル構成の分布について「Ewensのサンプリング公式」と呼ばれる分布を求めた。対立遺伝子（アレル）は自然選択に対して中立と仮定する。個々のアレルタイプに個性はないので、我々は集団から取り出したn個のサンプルのアレルタイプによる分割構造に興味がある。そこでサンプル中にi個含まれているアレルタイプの数を$a_i\,(i=1,2,...,n)$としよう。例えば１０個の遺伝子をサンプルしたとき、アレルAが３個、Bが１個、Cが１個、Dが２個、Eが３個含まれているとすると$a_1=2,\ a_2=1,\ a_3=2$、他は全て$a_i=0$である。$a=(a_1,a_2,...,a_n)$は遺伝子サンプル中のアレル構成による分割構造を表現している。全てが同じアレルタイプのときは$a_n=1$、$a_1=...=a_{n-1}=0$である。$\sum_{i=1}^{n} ia_i=n$（サンプル数）であり、$|a|=\sum_{i=1}^{n} a_i$はサンプル中の異なるアレルタイプ数を表す。

定理４．１

集団からランダムに取り出した遺伝子サンプルのアレル分割構造が$a=(a_1,a_2,...,a_n)$である確率は次の式で与えられる。

$$P(a_1,a_2,..,a_n)=\frac{n!}{(\vartheta)_n}\prod_{j=1}^{n}\frac{\vartheta^{a_j}}{j^{a_j}a_j!} \qquad \text{(Ewensのサンプリング公式)} \qquad (4.1)$$

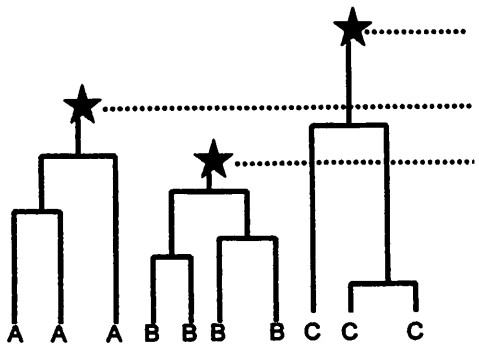
またアレルタイプの数については $P(|a|=k)=\dfrac{\vartheta^k S(n,k)}{\vartheta(\vartheta+1)....(\vartheta+n-1)}$ (4.2)

ここで、$S(n,k)$は第１種スターリング数（付録(C.1)参照）である。

(証明)

これらの結果はEwens(1972),Karlin&MacGregor(1972),Watterson(1984), Kingman (1982)などにより求められているが、ここでは第３章で紹介した突然変異存在下での遺伝子系図を用いて求めてみよう。まずp種のアレルを考え、集団からn個のサンプルを取り出

し、サンプルには1番からn番までナンバーをつける。このサンプル中に$A_1, A_2, ..., A_p$のp種のアレルを順に、それぞれk_1個、k_2個、・・・k_p個づつ取り出す確率を求める（ただし$k_1 + ... + k_p = n$）。そのためには、n個のサンプルは$A_1, A_2, ..., A_p$のp種のアレルに対応してk_1個、k_2個、・・・k_p個のp個のクラスに分けられ、各々のクラスは同じタイプのアレルで一つの共通なmutantを祖先とする一つの血族同値類に属すことになる。



上記の条件を満たす系図を描いたとき、i番目のクラスのk_i個の遺伝子の系図を遡ると$k_i - 1$回のcoalesceと最後の突然変異の合計k_i回の事象、従って全部で$k_1 + ... + k_p = n$回の事象が存在するはずである。系図上で祖先の数jのとき、生じる事象がその時点でk個の祖先を持つアレルクラスで起こる合祖である確率は$\frac{k(k-1)/2}{\{j\vartheta + j(j-1)\}/2}$であり、突然変異である確率は$\frac{\vartheta/2}{\{j\vartheta + j(j-1)\}/2}$である。さらに、条件を満たす可能な系図は各$i$番目のクラスの遺伝子について合祖(coalesce)と突然変異(mutation)合わせてk_i個のeventがあり、全部で合計n回の事象を順に並べる方法の数だけあるので$\frac{n!}{\prod_{i=1}^{p} k_i!}$通りとなる。よって1から$n$までナンバーのついた遺伝子中に$A_1, A_2, ..., A_p$の$p$種のアレルをそれぞれ$k_1$個、$k_2$個、・・・$k_p$個ずつ含まれる確率は次式で与えられる。

$$P(k_1, k_2, ..., k_p) = [\prod_{j=n}^{1} \frac{1}{\{j\vartheta + j(j-1)\}/2}]\{\prod_{i=1}^{p}(\prod_{j=k_i}^{2} \frac{j(j-1)}{2})\}(\frac{\vartheta}{2})^p \times \frac{n!}{\prod_{i=1}^{p} k_i!}$$

$$= \frac{\vartheta^p}{(\vartheta)_n} \prod_{i=1}^{p} (k_i - 1)! \qquad \text{ただし } (\vartheta)_n = \vartheta(\vartheta+1)(\vartheta+2)...(\vartheta+n-1) \qquad (4.3)$$

中立な遺伝子の場合、その個々の遺伝子のアレルタイプに興味は無く、サンプル中のアレルタイプ数とその構成に興味がある。そこでa_i $(1 \le i \le n)$をi個の遺伝子がサンプル中に存在するアレルタイプの数とすると、$a = (a_1, a_2, ..., a_n)$ ， $\sum_{i=1}^{n} i a_i = n$, $\sum_{i=1}^{n} a_i = p$を満たす。例えば、$(k_1, k_2, k_3, k_4, k_5) = (5,5,3,3,1)$と$(k_1, k_2, k_3, k_4, k_5) = (3,5,3,5,1)$は同じアレル分割$a = (a_i; 1 \le i \le 17, a_1 = 1, a_3 = 2, a_5 = 2, a_i = 0 (i \ne 1,3,5))$を持つ。n個のサンプル遺伝子に対してその中に、分割構成$(a_1, .., a_n)$を持つような対立遺伝子を含んでいる確率$P(a_1, a_2, \cdots, a_n)$を求めてみよう。ある一つの分割構造$(a_1, .., a_n)$に対して、$\sum_{i=1}^{n} a_i = p$のとき、i個のサンプルを含むアレルタイプがa_i種あるので、ラベルの与え方は$\prod_{i=1}^{p} a_i!$通りある。その各々のアレルタイプ内でi個の遺伝子について順番の付け方は$i!$通り。n個のサンプル遺伝子について、アレルタイプと、そのタイプ内での順番によって$\dfrac{n!}{\prod_{j=1}^{n} \{(j!)^{a_j} a_j!\}}$通りのラベリングの方法がある。この因子を(4.3)式に掛けて次のサンプリング公式(4.1)を得る。

$P(a_1, a_2, .., a_n) = \dfrac{n!}{(\vartheta)_n} \prod_{j=1}^{n} \dfrac{\vartheta^{a_j}}{j^{a_j} a_j!}$。また付録(C.1)の定理C2、及び$\sum_{k=1}^{n} P(|a| = k) = 1$より

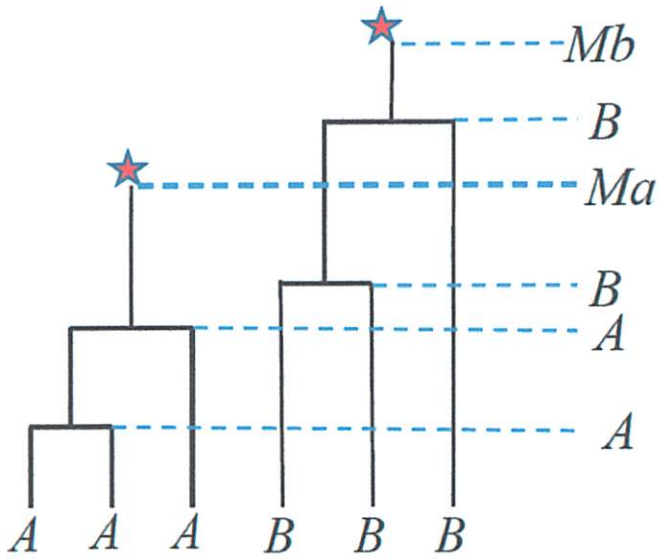
$$P(|a| = k) = \sum_{\substack{a \\ |a|=k}} P(a_1, a_2, \cdots, a_n) = \frac{\vartheta^k}{(\vartheta)_n} \sum_{\substack{a \\ |a|=k}} \frac{n!}{j^{a_j} a_j!} = \frac{\vartheta^k S(n,k)}{\vartheta(\vartheta+1)....(\vartheta+n-1)}$$ を得る。

4．2　アレルの年齢

集団からランダムに取り出した遺伝子の系図を遡るとそのアレルタイプの共通祖先、すなわちそのタイプの起源となる突然変異を起こした祖先遺伝子に到達する。サンプル中の各アレルタイプの起源に到達するまでの時間をそのアレルタイプの年齢という。アレルの年齢については古くは Kimura&Ohta(1973),Maruyama(1974)など多くの研究がある。ここでは遺伝子系図の結果を使ってアレルの年齢を求める。n個のサンプルを取り出したとき、その中に$A_1, A_2, ..., A_p$のp種のアレルがそれぞれk_1個、k_2個、・・・k_p個ずつ含まれ、かつA_1アレルが最も若く、次にA_2アレル、・・・、最も古いのがA_pアレルとなる確率$P(A_1 \rhd A_2 \rhd ... \rhd A_p; (k_1, k_2, \cdots, k_p))$を求めよう。ここで記号$A_1 \rhd A_2$　はアレルA_1がA_2

よりも若いことを意味する。そのための条件は起源突然変異 $A_1, A_2, .., A_p$ がその年齢の順に

系図上で生じることである。たとえば2種のアレル A,B が n 個のサンプル中にそれぞれ $n-k$ 個と k 個含まれ、かつ $A \triangleright B$ となる確率を考えてみよう。



上図は $n=6, k=3$ の場合の図である。アレル A での合祖を A、アレル B での合祖を B A,B の起源となる突然変異を Ma,Mb で表し、上の系図を古い事象から順に表すと（Mb,B,Ma,B,A,A）と表示される。一般に A,B が n 個のサンプル中にそれぞれ $n-k$ 個と k 個含まれるとき、$n-k-1$ 個の A と $k-1$ 個の B，および Ma, Mb の合計 n 個の記号を左端が Mb で始まり、B の位置は残り $n-1$ 個の場所から $k-1$ 個選ぶ方法だけある。Ma は残り $n-k$ 個の場所の最も左の位置に置き、他は全て A となる。条件を満たす可能な系図の総数は $\frac{(n-1)!}{k!(n-1-k)!}$ 通りとなる。同様にして、$A_1, A_2, ..., A_p$ の p 種のアレルがそれぞれ k_1 個、k_2 個、・・・k_p 個づつ含まれる場合には、合祖と突然変異を合わせた計 n 回の事象に対応して n 個の席を横一列に準備し、左から右へ事象の古い順に並べる。最も古いアレル A_p からその起源となる突然変異の事象は左端の位置に置かれ、アレル A_p で生じた k_p-1 回の合祖事象を残り $n-1$ 個の場所に入れるので $\binom{n-1}{k_p-1}$ 通りある。次のアレル A_{p-1} について、起源の突然変異事象は残りの中の最も左端の位置に置かれ、$k_{p-1}-1$ 回と合祖事象は残り $n-k_p-1$ 個の位置から選ぶ。これを繰り返して次式を得る。

$$P(A_1 \rhd A_2 \rhd \ldots \rhd A_p;(k_1,k_2,\cdots,k_p))$$

$$=\frac{\vartheta^p}{(\vartheta)_n n!}[\prod_{i=1}^{p}\{k_i!(k_i-1)!\}]\times\binom{n-1}{k_p-1}\binom{n-k_p-1}{k_{p-1}-1}\cdots\binom{k_1-1}{k_1-1}$$

$$=\frac{\vartheta^p}{(\vartheta)_n n!}\{\prod_{i=1}^{p}k_i!\prod_{i=1}^{p}(k_i-1)!\}\frac{(n-1)!}{(k_p-1)!(n-k_p)!}\times\frac{(n-k_p-1)!}{(k_{p-1}-1)!(n-k_p-k_{p-1})!}\times\ldots.$$

$$=\frac{\vartheta^p}{(\vartheta)_n}\times\frac{\prod_{i=1}^{p}k_i!}{n(n-k_p)(n-k_p-k_{p-1})\ldots.k_1}=\frac{\vartheta^p}{(\vartheta)_n}\times\frac{\prod_{i=1}^{p}k_i!}{k_1(k_1+k_2)\ldots(k_1+k_2+\ldots+k_p)}$$

(Donnelly&Tavare(1986)の(4.1)式)　(4.4)

n個のサンプル遺伝子において、どの遺伝子がどのクラス（アレル）に属するか区別を無視するために$\frac{n!}{\prod_{i=1}^{p}k_i!}$を掛けると、単にp種のアレルがそれぞれ$k_1$個、$k_2$個、・・・$k_p$個づつサンプルされているとき、$k_1$個のアレルが最も若く、・・・、$k_p$個のアレルが最も古い確率$P(k_1 \rhd k_2 \rhd \ldots \rhd k_p)$について次の定理を得る。

定理４．２

$$P(k_1 \rhd k_2 \rhd \ldots \rhd k_p)=\frac{\vartheta^p\, n!}{(\vartheta)_n k_1(k_1+k_2)\ldots.(k_1+k_2+\ldots+k_p)} \qquad (4.5)$$

これはDonnelly&Tavare(1986)の式(4.2)と一致する。

最も古いアレルA_pの年齢分布は指数分布の畳み込み$\overset{n}{\underset{j=1}{*}}Exp(\frac{2}{j\vartheta+j(j-1)})$に従い、期待値$=\sum_{j=1}^{n}\frac{2}{j\vartheta+j(j-1)}$である。

４．３　固有変異の多型（Unique Event Polymorphism）

ヒトではDNA塩基の数が10の９乗のオーダーということを考えると生物のゲノム上に生じる各突然変異は進化の過程でただ１回起こる事象と考えられる。このように各突然変異をユニークな事象と見るとき、サンプル中に見られる多型の起源となる固有な突然変異の年齢を考える。その例としてn個のサンプル中に、あるUEPのサンプルがk個$(1 \le k \le n-1)$含まれているとき、そのUEPの年齢を推定しよう。

T_Δを突然変異が起きた時刻、$\frac{\vartheta}{2}$を突然変異率、サンプル遺伝この系図過程をα_tで表し、

$|\alpha_t| = i$の状態の時に突然変異が生じたと仮定しよう（$2 \le i \le n-k+1$）。

突然変異型：$t \le T_\Delta$では$k-1$回の合祖と最後に突然変異が生じる。$t > T_\Delta$では突然変異タイプの祖先も含めたi個の野生型祖先が$i-1$回合祖して最終的なMRCAに到達する。
野生型：$t \le T_\Delta$では$n-k$個の系統が$i-1$個の系統になるまで合祖する。
$t > T_\Delta$では突然変異型と同じ。

事象U：n個のサンプル中に$(k, n-k)$の割合で固有変異の多型が観察される事象

事象U_i：事象Uの中で$|\alpha_t| = i$の状態の時に突然変異が生じた事象

$$P_\vartheta(U_i) = \left(\frac{\vartheta/2}{\{\vartheta i + i(i-1)\}/2}\right)\prod_{j=2}^{i}\left(\frac{j(j-1)/2}{\{\vartheta j + j(j-1)\}/2}\right)\times\left\{\prod_{j=i+1}^{n}\frac{1}{\{\vartheta j + j(j-1)\}/2}\right\}$$
$$\times\left(\prod_{j=i}^{n-k}\frac{j(j-1)}{2}\right)\left(\prod_{j=2}^{k}\frac{j(j-1)}{2}\right)\times {}_{n-i}C_{k-1}$$

$$P_\vartheta(U) = \sum_{i=2}^{n-k+1} P(U_i)$$

事象U_iの条件下で突然変異の年齢(T_Δ)の分布$P_\vartheta(T_\Delta = t|U_i)$は

$$P_\vartheta(T_\Delta = t|U_i) = \mathop{*}_{j=n}^{i} Exp\left(\frac{2}{j(\vartheta + j - 1)}\right)$$　　（指数分布の畳み込み）、

事象Uの条件下で突然変異の年齢(T_Δ)の分布は

$$P_\vartheta(T_\Delta = t|U) = \frac{1}{P_\vartheta(U)}\sum_{i=2}^{n-k+1} P_\vartheta(T_\Delta = t, U_i) = \sum_{i=2}^{n-k+1}\frac{P_\vartheta(U_i)}{P_\vartheta(U)} P_\vartheta(T_\Delta = t|U_i)$$

固有変異は突然変異率が非常に小さい事象と考えられるので$\vartheta \to 0$の極限を考える。

$$\lim_{\vartheta\to 0}\frac{P_\vartheta(U_i)}{P_\vartheta(U)} = \frac{{}_{n-i}C_{k-1}}{\sum_{i=2}^{n-k+1} {}_{n-i}C_{k-1}} = \frac{{}_{n-i}C_{k-1}}{{}_{n-1}C_k}$$、　$$\lim_{\vartheta\to 0} P_\vartheta(T_\Delta = t|U_i) = \mathop{*}_{j=n}^{i} Exp\left(\frac{2}{j(j-1)}\right)$$

定理４．３　固有変異の年齢の分布

$$P(T_\Delta = t|U) = \sum_{i=2}^{n-k+1}\frac{{}_{n-i}C_{k-1}}{{}_{n-1}C_k}\left\{\mathop{*}_{j=n}^{i} Exp\left(\frac{2}{j(j-1)}\right)\right\}$$

平均は$E[T_\Delta|U] = \sum_{i=2}^{n-k+1}\frac{{}_{n-i}C_{k-1}}{{}_{n-1}C_k}\times\frac{2(n-i+1)}{n(i-1)}$となる。この結果はTavare(2004)(p116,p117)

と一致する。特に$\frac{k}{n}=x$（一定）として、サンプル数nを無限に大きくすると

$$\lim_{n\to\infty}E[T_\Delta|U]=\lim_{n\to\infty}\sum_{i=2}^{n-k+1}\frac{{}_{n-i}C_{k-1}}{{}_{n-1}C_k}\times\frac{2(n-i+1)}{n(i-1)}=\frac{2x}{1-x}\sum_{i=2}^{\infty}\frac{(1-x)^i}{i-1}=-\frac{2x}{1-x}\log x$$

これは拡散過程を用いたKimura&Ohta(1973)の結果と一致する。

(cf. Wiuf and Donnelly(1999))

○固有変異多型（UEP）としてのSNP（単一塩基多型）の分布

第2節の分離サイトの数の分布で使った$\vartheta=4Nu$は全ゲノム上に発生するＳＮＰに関する全突然変異率である。しかし特定の領域あるいはサイトに着目するとそこで発生するＳＮＰは非常に稀にしか起きないＵＥＰと考えられる。現実のサンプルで検出されるＳＮＰはある領域（例えばある遺伝子のイントロン領域）でのSNPでありこのデータから系図を考察するのはＵＥＰとして扱うのが適切である。第2節の分離サイトの結果から$\vartheta\to 0$の極限を考える。定理３．６より以下の結果を得る。

$$\lim_{\vartheta\to 0}P(T=t|S=k)=\sum_{|r|=k}\frac{\prod_{j=2}^{n}(\frac{1}{j-1})^{r_j}}{P(S=k)}\{\overset{n}{\underset{j=2}{*}}Gamma(r_j+1,\frac{2}{j(j-1)})\}$$

$$\text{ただし}\quad P(S=k)=\sum_{|r|=k}\prod_{j=2}^{n}(\frac{1}{j-1})^{r_j}$$

$$E[T|S=k]=\sum_{|r|=k}A(r)\{\sum_{j=2}^{n}\frac{2(r_j+1)}{j(j-1)}\}\qquad\text{ただし}\quad A(r)=\prod_{j=2}^{n}(\frac{1}{j-1})^{r_j}/P(S=k)$$

４．４ ディリクレ分布とポリアの壺モデル

４．４．１ ベータ分布、ガンマ分布、ディリクレ分布

相互に関連の深い次の三つの確率分布の説明から始めることにしよう。

（１） ベータ分布 $Beta(\alpha,\beta)$

$$f(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\quad(0<x<1)$$

ここで$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ ； $\Gamma(\cdot)$はガンマ関数。

（２）ガンマ分布 $\Gamma(\alpha,\lambda)$

$$g(x;\alpha,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}\quad(0<x<\infty),\ \alpha>0,\ \lambda>0$$

α を shape parameter, λ を scale parameter という。

特性関数は以下の様に求められる。

$$\phi(t)=\int_0^\infty e^{itx}g(x;\alpha,\lambda)dx=\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}\exp[(it-\lambda)x]dx=\left(1-\frac{it}{\lambda}\right)^{-\alpha}$$

これより、確率変数 $X_1,\cdots,X_n$ が独立で、各 X_i が $\Gamma(\alpha_i,\lambda)$ に従うとき（λ は共通)、和 $X_1+\cdots+X_n$ はガンマ分布 $\Gamma(\alpha_1+\cdots+\alpha_n,\lambda)$ に従う。

（3） ディリクレ分布　$D(\alpha_1,\alpha_2,\cdots,\alpha_{k+1})$（ただし $\alpha_i>0,\ i=1,\ldots,k+1$)

領域 $\Delta=\{\vec{x}=(x_1,\cdots,x_k);\ 0\le x_i\le 1,\ \sum_{i=1}^{k}x_i\le 1\}$ 上の分布

$h(\vec{x};\vec{\alpha})=\dfrac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}\prod_{i=1}^{k+1}x_i^{\alpha_i-1}$ 、ただし $\vec{\alpha}=(\alpha_1,\cdots,\alpha_{k+1}),\ x_{k+1}=1-\sum_{i=1}^{k}x_i$ 。

$k=1$ のときはベータ分布となり、ベータ分布の多次元版と考えられる。

補題４．４

ディリクレ分布 $D(\alpha_1,\alpha_2,\cdots,\alpha_{k+1})$ を変数 x_k について $0\le x_k\le 1-\sum_{i=1}^{k-1}x_i$ の区間で積分した $(x_1,\cdots,x_{k-1})$ の周辺分布密度はディリクレ分布 $D(\alpha_1,\cdots,\alpha_{k-1},\alpha_k+\alpha_{k+1})$ となる。

(証明)

$$\frac{\Gamma(\alpha_1+\cdots\alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}\prod_{i=1}^{k-1}x_i^{\alpha_i-1}\int_0^t x_k^{\alpha_k-1}\left\{(1-\sum_{i=1}^{k-1}x_i)-x_k\right\}^{\alpha_{k+1}-1}dx_k \qquad (\text{ただし} t=1-\sum_{i=1}^{k-1}x_i)$$

$x_k=(1-\sum_{i=1}^{k-1}x_i)y=ty$ と変数変換すると

$$=\frac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}\prod_{j=1}^{k-1}x_j^{\alpha_j-1}\left(1-\sum_{i=1}^{k-1}x_i\right)^{\alpha_k+\alpha_{k+1}-1}\int_0^1 y^{\alpha_k-1}(1-y)^{\alpha_{k+1}-1}dy$$

$$=\frac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}\times B(\alpha_k,\alpha_{k+1})\left(\prod_{j=1}^{k-1}x_j^{\alpha_j-1}\right)\left(1-\sum_{i=1}^{k-1}x_i\right)^{\alpha_k+\alpha_{k+1}-1}$$

$$=\frac{\Gamma(\alpha_1+\cdots\alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k-1})\Gamma(\alpha_k+\alpha_{k+1})}\left(\prod_{j=1}^{k-1}x_j^{\alpha_j-1}\right)x_k^{\alpha_k+\alpha_{k+1}-1} \qquad \text{ただし} x_k=1-\sum_{i=1}^{k-1}x_i$$

次にガンマ分布とディリクレ分布の関係について次の補題を証明する。

補題４．５(Mosimann(1962))

$Y_j\,(j=1,2,\cdots,k)$ をそれぞれガンマ分布$\Gamma(\alpha_j,\lambda)$に従う独立な確率変数とする。

このとき、$P_j=\dfrac{Y_j}{\sum_{i=1}^{k}Y_i},(j=1,2,\cdots,k)$ とすると、$(P_1,P_2,\cdots P_k)$はディリクレ分布
$D(\alpha_1,\cdots,\alpha_k)$に従う。

（証明）

$(Y_1,\cdots,Y_k)$の同時分布密度を$g(y_1,\cdots,y_k)$とすると

$$g(y_1,\cdots,y_k)=\left(\prod_{j=1}^{k}\frac{\lambda^{\alpha_j}}{\Gamma(\alpha_j)}y_j^{\alpha_j-1}\right)\exp\left[-\lambda\left(\sum_{i=1}^{k}y_i\right)\right] \tag{4.6}$$

$P_j=Y_j\Big/\left(\sum_{i=1}^{k}Y_i\right),\ j=1,2,\cdots,k$、　$W=\sum_{i=1}^{k}Y_i$ とする。これより、

$Y_j=P_jW\ \ (j=1,2,\cdots,k-1),\ Y_k=\left(1-\sum_{i=1}^{k-1}Y_i\right)W$。

$(Y_1,\cdots,Y_k)$の分布から$(P_1,\cdots,P_{k-1},W)$の分布を求める。変数変換のヤコビアンは

$$\frac{\partial(y_1,y_2,\cdots,y_k)}{\partial(p_1,\cdots,p_{k-1},w)}=\begin{vmatrix}\dfrac{\partial y_1}{\partial p_1} & \dfrac{\partial y_1}{\partial p_2} & \cdots & \dfrac{\partial y_1}{\partial p_{k-1}} & \dfrac{\partial y_1}{\partial w}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \dfrac{\partial y_k}{\partial p_1} & \dfrac{\partial y_k}{\partial p_2} & \cdots & \dfrac{\partial y_k}{\partial p_{k-1}} & \dfrac{\partial y_k}{\partial w}\end{vmatrix}=\begin{vmatrix}w & 0 & \cdots & p_1\\ 0 & w & \cdots & p_2\\ \vdots & \vdots & \vdots & \vdots\\ -w & -w & \cdots & 1-\sum_{j=1}^{k-1}p_j\end{vmatrix}$$

第１行から第$k-1$行までの和を第k行に加えると

$$=\begin{vmatrix}w & 0 & \cdots & 0 & p_1\\ 0 & w & \cdots & 0 & p_2\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & w & p_{k-1}\\ 0 & 0 & \cdots & 0 & 1\end{vmatrix}=w^{k-1}。$$

これより$(P_1,P_2,\cdots P_k)$の分布密度関数は(4.6)を変数変換し、wについて積分すると

$$f(p_1,p_2,\cdots,p_{k-1})=\left\{\prod_{j=1}^{k}\frac{\lambda^{\alpha_j}}{\Gamma(\alpha_j)}{p_j}^{\alpha_j-1}\right\}\int_0^{\infty}w^{\alpha-1}e^{-\lambda w}dw\quad,\ \alpha=\sum_{i=1}^{k}\alpha_i$$

$$=\frac{\Gamma(\alpha)}{\prod_{j=1}^{k}\Gamma(\alpha_j)}\prod_{j=1}^{k}p_j^{\alpha_j-1}\quad。$$

補題４．６(Lukacs(1955))

X,Y を互いに独立で退化しない正値の確率変数とする。このとき $P=\dfrac{Y}{X+Y}$ と $W=X+Y$ が独立となるための必要十分条件は X,Y が同じスケールパラメーターのガンマ分布に従うことである。

（証明）まず必要性を証明しよう。

X,Y,P,W の分布関数を $F(x),G(y),H_1(p),H_2(w)$ とする。X,Y の特性関数を

$$f(v)=\int_0^\infty e^{ivx}dF(x),\quad g(v)=\int_0^\infty e^{ivy}dG(y)\qquad (v=s+it) \tag{4.7}$$

とする。ラプラス変換の性質より $t>0$ で $f(v),g(v)$ は解析的である。さらに、次の連続性を仮定する。

$$\lim_{t\downarrow 0} f(s+it)=f(s),\ \lim_{t\downarrow 0} g(s+it)=g(s) \tag{4.8}$$

v について微分すると、

$$f'(v)=i\int_0^\infty xe^{ivx}dF(x),\quad f''(x)=-\int_0^\infty x^2e^{ivx}dF(x) \tag{4.9a}$$

$$g'(v)=i\int_0^\infty ye^{ivy}dG(y),\quad g''(x)=-\int_0^\infty y^2e^{ivy}dG(y) \tag{4.9b}$$

$0\le P=\dfrac{Y}{X+Y}\le 1$ より確率変数 P のモーメントは常に存在する。

$$\theta_1=E[P]=E\left[\frac{Y}{X+Y}\right],\ \theta_2=E[P^2]=E\left[\left(\frac{Y}{X+Y}\right)^2\right] \tag{4.10}$$

とすると、明らかに $0<\theta_1{}^2\le\theta_2<1$。$W$ と P は独立なので、(W,P) の特性関数は

$E[\exp(izP+ivW)]=E[\exp(izP)]E[\exp(ivW)]$、すなわち

$$E\left[\exp\left\{iz\left(\frac{Y}{X+Y}\right)+ivW\right\}\right]=E\left[\exp\left\{iz\left(\frac{Y}{X+Y}\right)\right\}\right]E[\exp\{iv(X+Y)\}] \tag{4.11}$$

両辺を v および z で微分した後、$z=0$ と置くと

$E[Y\exp\{iv(X+Y)\}]=\theta_1E[(X+Y)\exp\{iv(X+Y)\}]$、ただし(4.10)を用いた。

X,Y の独立性より

$E[Y\exp(ivY)]E[\exp(ivX)]=\theta_1\{E[X\exp(ivX)]E[\exp(ivY)]+E[\exp(ivX)]E[Y\exp(ivY)]\}$

(4.9a,b)より　$(-i)g'(v)f(v)=\theta_1\{(-i)f'(v)g(v)+(-i)g'(v)g(v)\}$、これより

$(1-\theta_1)g'(v)f(v)=\theta_1f'(v)g(v)$　ただし $\mathrm{Im}(v)>0$ (4.12)

また$f(0)=g(0)=1$, と$f(\nu),g(\nu)$の連続性より原点の近傍（ただし$\mathrm{Im}(\nu)>0$）で、$f(\nu)\neq 0,\ g(\nu)\neq 0$が成り立つ。特性関数の定義より明らかに$f(it)>0, g(it)>0$ ただし$t>0$の実数。$f(\nu),g(\nu)$の連続性とハイネ＝ボレルの被覆定理より$0<\mathrm{Im}(\nu)\leq 1$を含むある単連結領域で$f(\nu)\neq 0, g(\nu)\neq 0$が成り立つ。以後この領域をDとする。
$\nu\in D$のとき、(4.12)の両辺を$f(\nu)g(\nu)$で割ると

$$(1-\theta_1)\frac{g'(\nu)}{g(\nu)}=\theta_1\frac{f'(\nu)}{f(\nu)} \tag{4.13}$$

初期条件$f(0)=g(0)=1$より、積分して次の関係式を得る。

$$\{g(\nu)\}^{1-\theta_1}=\{f(\nu)\}^{\theta_1} \qquad \text{ただし} \quad \mathrm{Im}(\nu)>0 \tag{4.14}$$

次に、(4.11)の両辺をν,zでそれぞれ2回微分し$z=0$と置くと、

$$E\left[Y^2\exp\{i(X+Y)\nu\}\right]=E\left[\left(\frac{Y}{X+Y}\right)^2\right]E\left[(X+Y)^2\exp\{i\nu(X+Y)\}\right]$$

X,Yの独立性に注意し、(4.7),(4.9a,b)及び(4.10)を使うと
$g''(\nu)f(\nu)=\theta_2\{f''(\nu)g(\nu)+2f'(\nu)g'(\nu)+g''(\nu)f(\nu)\}$、$\nu\in D$のとき、両辺を$f(\nu)g(\nu)$で割ると

$$\frac{g''(\nu)}{g(\nu)}=\theta_2\left\{\frac{f''(\nu)}{f(\nu)}+2\frac{f'(\nu)}{f(\nu)}\times\frac{g'(\nu)}{g(\nu)}+\frac{g''(\nu)}{g(\nu)}\right\} \tag{4.15}$$

$\phi(\nu)=\log f(\nu),\ \psi(\nu)=\log g(\nu)$とすると

$$\frac{f'(\nu)}{f(\nu)}=\phi'(\nu),\quad \frac{f''(\nu)}{f\nu)}=\phi''(\nu)+\{\phi'(\nu)\}^2 \tag{4.16}$$

$\psi(\nu)$についても同様の式が成り立つ。(4.13)を$\phi(\nu),\psi(\nu)$で表すと
$(1-\theta_1)\psi'(\nu)=\theta_1\phi'(\nu)$　これより　$(1-\theta_1)\psi''(\nu)=\theta_1\phi''(\nu)$ (4.17)
(4.15)を(4.16)(4.17)を用いて$\phi(\nu)$のみで表すと

$$(1-\theta_1)(\theta_1-\theta_2)\phi''(\nu)=(\theta_2-\theta_1^2)\{\phi'(\nu)\}^2 \tag{4.18}$$

（1）$0<\theta_1^2<\theta_2<\theta_1<1$（等号が付かないとき）
$\rho=\dfrac{(1-\theta_1)(\theta_1-\theta_2)}{(\theta_2-\theta_1^2)}$と置くと$\rho>0$である。

(4.18)より$\nu\in D,\ \mathrm{Im}(\nu)>0$のとき　$$\frac{\phi''(\nu)}{\{\phi'(\nu)\}^2}=\frac{1}{\rho} \tag{4.19}$$

$k_1 = E[\exp(-X)]$, $k_2 = E[X\exp(-X)]$, $\lambda = (k_1\rho - k_2)/k_2$ とする。 (4.20)

初期条件 $\phi'(i) = \dfrac{f'(i)}{f(i)} = \dfrac{ik_2}{k_1}$, $(i = \sqrt{-1})$ より $C = \dfrac{\lambda}{i\rho}$。故に $\phi'(v) = \dfrac{i\rho}{\lambda - iv}$。積分して

$f(v) = A\left(1 - \dfrac{iv}{\lambda}\right)^{-\rho}$,(A は定数)。連続性の条件より $\lim_{t\downarrow 0} f(it) = f(0) = 1$、よって A=1。

(4.14)より $f(v) = \left(1 - \dfrac{iv}{\lambda}\right)^{-\rho}$, $g(v) = \left(1 - \dfrac{iv}{\lambda}\right)^{-\sigma}$ (ただし $\sigma = \dfrac{\theta_1\rho}{1-\theta_1}$) (4.21)

(4.21)は領域 D で成り立つが、$f(v)$ および $\left(1 - \dfrac{iv}{\lambda}\right)^{-\rho}$ は複素平面の上半部 $\mathrm{Im}(v) > 0$ で解析的でかつ領域 D で一致する。故に解析関数の一致の定理より $\mathrm{Im}(v) > 0$ で(4.21)が成り立つ。$g(v)$ についても同様である。連続性の仮定(4.8)より実数 v についても成り立つ。

X, Y の特性関数は(4.21)で与えられ、ともに同じスケールパラメーター λ のガンマ分布に従う。

（2） $\theta_1 = \theta_2$ または $\theta_1^2 = \theta_2$ のとき

$\theta_1 = \theta_2$ のとき、$\theta_2 = \theta_1 > \theta_1^2$ なので(4.18)より $\phi'(v) = 0$、よって $\phi(v) = C$。また $\theta_1^{\,2} = \theta_2$ ならば $\theta_1 > \theta_1^2 = \theta_2$ なので(4.18)より $\phi''(v) = 0$ すなわち $\phi(v) = Av + C$。いずれにしても $f(v) = \exp(Av + C)$ となりラプラス逆変換により X は退化した分布になり定理の条件を満たさない。Y についても同様である。故に P と W が独立ならば X,Y は同じスケールパラメーターの Γ 分布に従う。

十分性の証明：X,Y を独立でそれぞれガンマ分布 $\Gamma(\alpha,\lambda), \Gamma(\beta,\lambda)$ に従う確率変数とする。

このとき $W = X + Y$ と $P = \dfrac{Y}{X+Y}$ の結合分布の特性関数は

$$E\left[\exp\left\{iv(X+Y) + iz\left(\frac{Y}{X+Y}\right)\right\}\right]$$

$$= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^\infty\int_0^\infty \exp\left[-(\lambda - iv)(x+y) + iz\left(\frac{y}{x+y}\right)\right] \times x^{\alpha-1}y^{\beta-1}dxdy$$

$\xi = (\lambda - iv)x$, $\eta = (\lambda - iv)y$ とし (x,y) から (ξ,η) に変数変換すると

$$= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \left\{ \frac{1}{(\lambda - iv)^{\alpha+\beta}} \right\} \int_0^\infty \int_0^\infty \exp\left[-(\xi+\eta) + iz\left(\frac{\eta}{\xi+\eta} \right) \right] \times \xi^{\alpha-1} \eta^{\beta-1} d\xi d\eta$$

すなわち特性関数が v の関数と z の関数の積として表されているので $W = X + Y$ と $P = \dfrac{Y}{X+Y}$ は独立である。　　（証明終わり）

証明は省略するが、補題４．６を一般化した次の補題が知られている。

補題４．７

（１）(Mosiman(1962)) $\{Y_i;\ i = 1,2,\cdots,k\}$ を k 個の独立で非退化、正値の確率変数とする。$k-1$ 個の確率変数 $P_j = Y_j \Big/ \left(\sum_{i=1}^{k} Y_i \right),\ (j = 1,2,\cdots,k-1)$ が $W = \sum_{i=1}^{k} Y_i$ と独立となるのは、全ての $Y_j\ (j = 1,\cdots,k)$ が同じスケールパラメーター λ の Γ 分布に従う時に限る。

（２）$\{Y_i;\ i = 1,2,\cdots,k\}$ を k 個の独立で非退化、正値の確率変数とする。$P_j = Y_j \Big/ \left(\sum_{i=1}^{k} Y_i \right),$ $(j = 1,2,\cdots,k-1)$ とすると、各 P_j が和 $W = \sum_{i=1}^{k} Y_i$ と独立ならば $(P_1,\cdots,P_{k-1})$ はディリクレ分布に従う。

４．４．２　完全中立性とディリクレ分布の一般化

$P_1, P_2, \cdots, P_k$ を非負で $\sum_{i=1}^{k} P_i = 1$ を満たす確率変数とする。また $S_0 = 0,\ j \geq 1$ のとき

$$S_j = \sum_{i=1}^{j} P_i,\quad Z_j = \frac{P_j}{1 - S_{j-1}} = \frac{P_j}{P_j + P_{j+1} + \cdots + P_k}$$ によって $\{S_j;\ 0 \leq j \leq k\},\ \{Z_j;\ 1 \leq j \leq k\}$

を定義する。明らかに $Z_1 = P_1,\ Z_k = 1$ である。これより、逆に $P_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i)$ と表わされる。Z_j は全体から $P_1, P_2, \cdots, P_{j-1}$ を取り除いた残りの中での P_j の相対的な割合 $P_j / (1 - S_{j-1})$ を表わしている。確率ベクトル $\vec{P} = (P_1,\cdots,P_k)$ に対して $\vec{P}_j^{\,1} = (P_1,\cdots,P_j)$、$\vec{P}_j^{\,2} = (P_{j+1},\cdots,P_k);\ (j < k)$ とし、$\vec{W}_j = \dfrac{\vec{P}_j^{\,2}}{1 - S_j} = \dfrac{(P_{j+1}, P_{j+2}, \cdots, P_k)}{P_{j+1} + \cdots + P_k}$ と定義する。

この小節では Connor and Mosimann(1969)に従って、確率ベクトルの完全中立性(completely neutral)と一般化されたディリクレ分布について紹介する。

定義：上で定義した確率ベクトル$\vec{P}=(P_1,\cdots,P_k)$に対して、P_1が確率ベクトル

$\vec{W}_1=\left(\frac{P_2}{1-P_1},\frac{P_3}{1-P_1},\cdots,\frac{P_k}{1-P_1}\right)$と独立のとき$P_1$は中立(neutral)であると言う。

定理４．８

確率ベクトル$\vec{P}=(P_1,\cdots,P_k)$に対してP_1が中立ならば、P_1は$\vec{W}_j$ $(j=1,2,\cdots,k-1)$と独立である。

(証明) $X_2=\frac{P_2}{1-P_1},X_3=\frac{P_3}{1-P_1},\cdots$とすると条件より$P_1$は$(X_2,X_3,\cdots,X_k)$と独立である。

$P_2=(1-P_1)X_2,\ P_3=(1-P_1)X_3,\cdots,P_k=(1-P_1)X_k$より

$\frac{P_3}{1-P_1-P_2}=\frac{(1-P_1)X_3}{1-P_1-(1-P_1)X_2}=\frac{X_3}{1-X_2}$,同様に$\frac{P_j}{1-P_1-P_2}=\frac{X_j}{1-X_2};(3\le j\le k)$。

故にP_1は$(X_2,\cdots,X_k)$と独立なので$(X_3,X_4,\cdots,X_k)/(1-X_2)=\vec{W}_2$と独立である。

同様にしてP_1は$\vec{W}_j$ $(j=1,2,\cdots,k-1)$と独立である。

中立性の概念をベクトルまで拡張しよう。

定義：確率ベクトル$\vec{P}=(\vec{P}_j^{\,1},\vec{P}_j^{\,2}),\ j=1,2,\cdots,k-1$に対して、$\vec{P}_j^{\,1}$が$\vec{W}_j$と独立であるとき確率ベクトル$\vec{P}_j^{\,1}$は中立であるという。また全ての$j$について$\vec{P}_j^{\,1}$が中立のとき、確率ベクトル$\vec{P}=(P_1,\cdots,P_k)$は完全中立という。

すなわち完全中立とは任意のjに対して最初のj個の要素の確率$P_1,\cdots,P_j$を与えたときそれを除いた残りの確率を他の$k-j$個の要素で分割する方法$\vec{W}_j$が$P_1,\cdots,P_j$と独立であることを意味する。ただしベクトル$\vec{P}$の成分の順序に依存することを注意しなければならない。すなわち(P_1,P_2,P_3,P_4)のP_1は中立であるが、順序を入れ替えて(P_2,P_1,P_3,P_4)とする

とP_2は中立でないこともある。確率ベクトルの中立性と確率変数系$\{Z_j; j=1,2,\cdots,k\}$の独立性に関して次の定理が成り立つ。

定理４．９

$j=1,2,\cdots,r$について、$\vec{P}_j^{\,1}$が中立ならば$Z_1,Z_2,\cdots,Z_r$は互いに独立である。

（証明）

$Z_1,Z_2,\cdots,Z_r$の同時分布$F_r(z_1,\cdots,z_r)$が各Z_jの分布$f_j(z_j)$の積として

$F_r(z_1,\cdots,z_r)=\prod_{j=1}^{r} f_j(z_j)$となることを示せばよい。まず、$P_1$が中立であることより、$Z_1=P_1$と$Z_2=P_2/(1-P_1)$が独立。故に、$F_2(z_1,z_2)=f_1(z_1)f_2(z_2)$が成り立つ。次に$\vec{P}_2^{\,1}=(P_1,P_2)$について$\vec{P}_2^{\,1}$が中立より$(P_1,P_2)$は$\left(\frac{P_3}{1-S_2},\frac{P_4}{1-S_2},\cdots\right)=\left(Z_3,\frac{P_4}{1-S_2},\cdots\right)$と独立である。故に$(Z_1,Z_2)$は$(P_1,P_2)$の関数なので$(Z_1,Z_2)$は$Z_3$と独立である。これより$F_3(z_1,z_2,z_3)=F_2(z_1,z_2)f_3(z_3)=f_1(z_1)f_2(z_2)f_3(z_3)$となり、$z_1,z_2,z_3$は独立である。以下同様にして$Z_1,Z_2,\cdots,Z_r$が独立であることが示される。

この定理より$\vec{P}=(P_1,\cdots,P_k)$が完全中立ならば$\{Z_1,Z_2,\cdots,Z_{k-1}\}$は互いに独立であるが、この条件が必要十分であることが示される。

定理４．１０

$\vec{P}=(P_1,\cdots,P_k)$が完全中立であるための必要十分条件は$\{Z_1,Z_2,\cdots,Z_{k-1}\}$が互いに独立であることである。ただし、$Z_k=1$とする。

（証明）

必要性は定理 4.9 で証明されたので十分性を示す。$\{Z_1,Z_2,\cdots,Z_{k-1}\}$が互いに独立のとき$\vec{P}_j^{\,1}$が$\vec{W}_j$ $(j=1,2,\cdots,k-1)$と独立であることを示せば十分である。Z_jの定義より$\prod_{i=1}^{j}(1-Z_i)=1-S_j$, $(1\le j\le k-1)$。これより$P_j=(1-S_j)Z_j=Z_j\prod_{i=1}^{j-1}(1-Z_i)$、また

$$\frac{P_{j+r}}{1-S_j}=\frac{Z_{j+r}\prod_{i=1}^{j+r-1}(1-Z_i)}{\prod_{i=1}^{j}(1-Z_i)}=Z_{j+r}\prod_{i=j+1}^{j+r-1}(1-Z_i)\ ,\ (1\le j\le k-j)$$。すなわち$\vec{P}_j^{\,1}$は$\{Z_1,\cdots,Z_j\}$で表され、$\vec{W}_j$ $(j=1,2,\cdots,k-1)$は$(Z_{j+1},\cdots,Z_r)$のみに依存する。従って$\vec{P}_j^{\,1}$と$\vec{W}_j$は独立である。この二つの定理より$1\le j\le k-2$に対して$\vec{P}_j^{\,1}$がZ_{j+1}と独立ならば$(Z_1,\cdots,Z_j)$はZ_{j+1}と独立になるので$\{Z_1,Z_2,\cdots,Z_{k-1}\}$は独立である。故に、定理４．１０より$\vec{P}=(P_1,\cdots,P_k)$は完全中立である。すなわち次の定理が成り立つ。

定理４．１１

$\vec{P}=(P_1,\cdots,P_k)$が完全中立であるための必要十分条件は$\vec{P}_j^{\,1}$とZ_{j+1}が全て独立となることである。

この完全中立性を利用して一般化されたディリクレ分布を定義し、ディリクレ分布と完全中立性との関係を調べる。$\{Z_1,Z_2,\cdots,Z_{k-1}\}$を独立な確率変数系で各$Z_j$はベータ分布$Beta(\alpha_j,\beta_j)$に従うものとする。すなわち$Z_j$の分布密度は

$$f_j(z_j)=\frac{1}{B(\alpha_j,\beta_j)}z_j^{\alpha_j-1}(1-z_j)^{\beta_j-1}\ ,\ (0<z_j<1)$$。

また$P_j=Z_j\left\{\prod_{i=1}^{j-1}(1-Z_i)\right\}$によって$\vec{P}=(P_1,\cdots,P_k)$を定義する。

$Z_j=\frac{P_j}{1-S_{j-1}}=\frac{P_j}{1-(P_1+\cdots+P_{j-1})}$より、$j>i$のとき$\frac{\partial Z_i}{\partial P_j}=0$, $\frac{\partial Z_j}{\partial P_j}=\frac{1}{1-S_{j-1}}$,

$j<i$のとき$\frac{\partial Z_i}{\partial P_j}=\frac{P_iP_j}{(1-S_{j-1})^2}$となる。これより、変数変換のヤコビアンは

$$J=\frac{\partial(Z_1,\cdots,Z_{k-1})}{\partial(P_1,\cdots,P_{k-1})}=\begin{vmatrix}\frac{\partial Z_1}{\partial P_1} & \cdots & \frac{\partial Z_{k-1}}{\partial P_{k-1}}\\ \vdots & \vdots & \vdots\\ \frac{\partial Z_{k-1}}{\partial P_1} & \cdots & \frac{\partial Z_{k-1}}{\partial P_{k-1}}\end{vmatrix}=\frac{1}{\prod_{i=1}^{k-1}(1-S_{i-1})}。 \tag{4.22}$$

この変数変換により$\vec{Z}=(Z_1,\cdots,Z_{k-1})$から$\vec{P}=(P_1,\cdots,P_k)$の分布を求める。

$$\prod_{j=1}^{k-1}\left\{B(\alpha_j,\beta_j)\right)^{-1}z_j{}^{\alpha_j-1}(1-z_j)^{\beta_j-1}\Big\}$$

$$=\left\{\prod_{j=1}^{k-1}B(\alpha_j,\beta_j)\right\}^{-1}\left\{\prod_{j=1}^{k-1}\left(\frac{P_j}{1-S_{j-1}}\right)^{\alpha_j-1}\right\}\left\{\prod_{j=1}^{k-1}\left(1-\frac{P_j}{1-S_{j-1}}\right)^{\beta_j-1}\right\}\times J$$

$$=\left\{\prod_{j=1}^{k-1}B(\alpha_j,\beta_j)\right\}^{-1}\left\{\prod_{j=1}^{k-1}\left(\frac{P_j}{1-S_{j-1}}\right)^{\alpha_j-1}\right\}\left\{\prod_{j=1}^{k-1}\left(\frac{1-S_j}{1-S_{j-1}}\right)^{\beta_j-1}\right\}\times\frac{1}{\prod_{j=1}^{k-1}(1-S_{j-1})}$$

$$=\left\{\prod_{j=1}^{k-1}B(\alpha_j,\beta_j)\right\}^{-1}\left\{\prod_{j=1}^{k-1}P_j{}^{\alpha_j-1}\right\}\left\{\prod_{j=1}^{k-1}(1-S_{j-1})^{\beta_{j-1}-(\alpha_j+\beta_j)}\right\}(1-S_{k-1})^{\beta_{k-1}-1}$$

$1-S_{j-1}=\sum_{i=j}^{k}P_i$　（ただし$\sum_{j=1}^{k}P_j=1$）および$1-S_{k-1}=P_k$より

$$=\left\{\prod_{j=1}^{k-1}B(\alpha_j,\beta_j)\right\}^{-1}P_k{}^{\beta_{k-1}-1}\left\{\prod_{j=1}^{k-1}\left(P_j{}^{\alpha_j-1}\left(\sum_{i=j}^{k}P_i\right)^{\beta_{j-1}-(\alpha_j+\beta_j)}\right)\right\} \tag{4.23}$$

ここでβ_0は任意の実数としてよい。また$\beta_{i-1}=\alpha_i+\beta_i\ (i=2,3,\cdots,k-1)$のとき(4.23)は

$\left\{\prod_{j=1}^{k-1}B(\alpha_j,\beta_j)\right\}^{-1}\left\{\prod_{j=1}^{k-1}P_j{}^{\alpha_j-1}\right\}P_k{}^{\beta_{k-1}-1}$　すなわちディリクレ分布$D(\alpha_1,\cdots,\alpha_{k-1},\beta_{k-1})$となる。従って、ディリクレ分布に従う確率変数$(P_1,P_2,\cdots,P_k)$は完全中立であり、ディリクレ分布は任意の変数の置換をしてもディリクレ分布なので$\vec{P}=(P_1,P_2,\cdots,P_k)$は任意の置換をしても完全中立な確率ベクトルである。それに対して(4.23)で定義される一般化されたディリクレ分布では一般に$(P_1,P_2,\cdots,P_k)$の順序において完全中立となる。
特に対称ディリクレ分布$D(\alpha,\cdots,\alpha)$となるのは(4.23)において$\beta_{i-1}=\alpha_i+\beta_i$かつ$\alpha_1=\alpha_2=\cdots=\alpha_{k-1}=\beta_{k-1}=\alpha$のときなので、$\alpha_j=\alpha$, $\beta_j=(k-j)\alpha$　$(1\le j\le k-1)$よりZ_jはベータ分布$Beta(\alpha,(k-j)\alpha)$に従うことを注意しておこう。

$(P_1,P_2,\cdots,P_k),\ \sum_{i=1}^{k}P_i=1$ がパラメーター α の対称ディリクレ分布に従うとき単に $D(\alpha;k)$ と書くことにする。完全中立性と関連して対称ディリクレ分布の size-biased permutation について触れておこう。

$\vec{P}=(P_1,P_2,\cdots,P_k),\ P_i\geq 0,\sum_{i=1}^{k}P_i=1$ を $\{1,2,\cdots,k\}$ 上のランダム確率測度とする。$\vec{P}$ の size-biased permutation $\vec{P}^S=(P_1^S,P_2^S,\cdots,P_k^S)$ とは次のように定義される。

定義：Size-biased permutation

任意の $\{1,2,\cdots,k\}$ の置換 σ に対して、$\vec{P}^S=(P_1^S,P_2^S,\cdots,P_k^S)=(P_{\sigma(1)},P_{\sigma(2)},\cdots,P_{\sigma(k)})$ となる確率が次式で与えられるとき、$\vec{P}^S$ は $\vec{P}$ の size-biased permutation という。

$$P\left(P_j^S=P_{\sigma(j)};1\leq j\leq i\middle|\vec{P}\right)=P_{\sigma(1)}\prod_{j=2}^{i}\left(\frac{P_{\sigma(j)}}{1-\sum_{r=1}^{j-1}P_{\sigma(r)}}\right),\ 1\leq i\leq k \qquad (4.24)$$

すなわち $\{1,2,\cdots,k\}$ の任意の置換 σ および $\vec{P}$ に対して、集団からランダムに一つの個体をサンプルし、その個体のタイプの頻度を P_1^S とする。その確率は $P(P_1^S=P_{\sigma(1)})=P_{\sigma(1)}$ である。この個体と同じタイプの個体を集団からすべて除き、残りの集団からまたランダムに１個体サンプルする。２番目にサンプルしたタイプの元の全集団での遺伝子頻度を P_2^S とする。その確率は $P\left(P_2^S=P_{\sigma(2)}\middle|P_1^S=P_{\sigma(1)}\right)=\frac{P_{\sigma(2)}}{1-P_{\sigma(1)}}$。この操作を全てのタイプが無くなるまで続け、順に $P_1^S,P_2^S,\cdots,P_k^S$ とするとき $\vec{P}^S=(P_1^S,P_2^S,\cdots,P_k^S)=(P_{\sigma(1)},P_{\sigma(2)},\cdots,P_{\sigma(k)})$ となる確率を(4.24)式右辺は示している。

集団の遺伝子頻度 $\vec{P}=(P_1,P_2,\cdots,P_k)$ が対称ディリクレ分布 $D(\alpha;k)$ に従うとき、その Size-biased permutation $\vec{P}^S$ の分布を求めてみよう。

$$P\left(P^S=(x_1,\cdots,x_k)\right)=\sum_{\sigma\in\mathfrak{R}(k)}P\left(P_j^S=x_{\sigma(j)};1\le j\le k\middle|\vec{P}=(x_1,\cdots,x_k)\right)\times P(\vec{P}=(x_1,..,x_k))$$

$$=\sum_{\sigma\in\mathfrak{R}(k)}\left\{x_{\sigma(1)}\prod_{j=2}^{k}\left(\frac{x_{\sigma(j)}}{1-\sum_{l=1}^{j-1}x_{\sigma(l)}}\right)\right\}\times\frac{\Gamma(n\alpha)}{\{\Gamma(\alpha)\}^n}\prod_{j=1}^{k}x_j^{\alpha-1}$$

$$=\sum_{\sigma^{-1}\in\mathfrak{R}(k)}\left\{x_1\prod_{j=2}^{k}\left(\frac{x_j}{1-\sum_{l=1}^{j-1}x_l}\right)\right\}\frac{\Gamma(n\alpha)}{\{\Gamma(\alpha)\}^n}\left(\prod_{j=1}^{k}x_{\sigma^{-1}(j)}\right)^{\alpha-1}$$

$$=\frac{k!\Gamma(k\alpha)}{\{\Gamma(\alpha)\}^k}\left\{\prod_{j=1}^{k}x_j\right\}^{\alpha-1}\left\{x_1\prod_{j=2}^{k}\frac{x_j}{1-\sum_{l=1}^{j-1}x_l}\right\}$$

最後の式は$B(\alpha,\beta)=\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$に注意して整理すると

$$\left\{\prod_{j=1}^{k-1}B(1+\alpha,(k-i)\alpha)\right\}^{-1}x_k^{\alpha-1}\left\{\prod_{j=1}^{k-1}\left(x_j^{\alpha}\left(\sum_{l=j}^{k}x_l\right)^{-1}\right)\right\}\ \text{となる。}\qquad(4.25)$$

これは(4.23)で$\alpha_i=1+\alpha,\ \beta_i=(k-i)\alpha$と置いた一般化されたディリクレ分布の形である。

すなわち、対称ディリクレ分布$D(\alpha;k)$のsize-biased permutationは$Z_j\ (1\le j\le k-1)$

がベータ分布$Beta(1+\alpha,(k-j)\alpha)$に従う一般化されたディリクレ分布である。そして

$\vec{P}^S=(P_1^S,P_2^S,\cdots,P_k^S),\ P_1^S=Z_1,P_2^S=(1-Z_1)Z_2,P_3^S=(1-Z_1)(1-Z_2)Z_3,\cdots$、

$P_k^S=(1-Z_1)\cdots(1-Z_{k-1})Z_k$　（ただし$Z_k=1$）と表される。

4．4．3　ポリアの壺とディリクレ過程

ポリアの壺とディリクレ分布の間の興味ある関係について紹介する。

ポリアの壺とは、壺の中に例えばb個の黒球とr個の赤球が入っているとする。まず1個の球をランダムに壺の中から取り出し、この球と同色の球を1個加えて壺に戻す。この操作を繰り返すとき、壺の中の2色の球の割合を問う確率論の問題を、発案者の名よりポリアの壺と呼んでいる。n回の試行の後、壺の中の黒球の割合をY_nとする。$Y_0=\dfrac{b}{b+r}$であ

る。Y_nは試行の度に確率的に変化するので確率過程である。n回の試行後の２色の球の個数をb_n, r_nとすると$Y_n = \frac{b_n}{b_n + r_n}$, $b_n + r_n = b + r + n$である。このとき、

$E[Y_{n+1}|Y_n] = \frac{b_n}{b_n + r_n} \times \frac{b_n + 1}{b_n + r_n + 1} + \frac{r_n}{b_n + r_n} \times \frac{b_n}{b_n + r_n + 1} = \frac{b_n}{b_n + r_n} = Y_n$、よって$\{Y_n\}$は有界なマルチンゲールである。マルチンゲールの収束定理より、極限$\lim_{n\to\infty} Y_n = Y$が存在する。

最初のn回が全て黒球である確率をμ_nとすると$\mu_n = \frac{b(b+1)\cdots(b+n-1)}{(b+r)(b+r+1)\cdots(b+r+n-1)}$、確率変数$X_n$を$n$回目に取り出した球が黒球のとき$X_n = 1$、赤球のとき$X_n = 0$で定義すると、$(X_1, X_2, \cdots, X_n)$は簡単な計算により交換可能である。すなわち$n$次元同時分布が変数の順序を変えても同じ分布に従う。$\mu_n = P(X_1 = 1, X_2 = 1, \cdots, X_n = 1)$に注意すると、ド・フィネッティの定理より$\mu_n = \int_0^1 x^n F(dx)$となる$[0, 1]$上の確率速度$F(x)$が存在する(参考：Feller「確率論とその応用Ⅱ」Ⅶ章. ４節)。Γ関数の性質を利用すると、

$\mu_n = \frac{\Gamma(b+n)/\Gamma(b)}{\Gamma(b+r+n)/\Gamma(b+r)} = \frac{\Gamma(b+n)\Gamma(b+r)}{\Gamma(b+r+n)\Gamma(b)}$と表せる。これはパラメーターが$(b, r)$のベータ分布$f_{(b,r)}(x) = \frac{\Gamma(b+r)}{\Gamma(b)\Gamma(r)} x^{b-1}(1-x)^{r-1}$の$n$次のモーメントに等しい。すなわち、$F(x) = f_{(b,r)}(x)$となる。$S_n = \sum_{i=1}^{n} X_i$とすると、$P(S_n = k) = \binom{n}{k} \int_0^1 x^k (1-x)^{n-k} f_{(b,r)}(x) dx$

すなわち、S_nの分布は二項分布のランダム化で表される。Yのラプラス変換を考えると

$$E[\exp(-\lambda Y)] = \lim_{n\to\infty} E[\exp(-\lambda Y_n)] = \lim_{n\to\infty} E[\exp(-\lambda S_n / n)] = \lim_{n\to\infty} \int_0^1 (xe^{-\lambda/n} + 1 - x)^n f_{(b,r)}(x) dx$$
$$= \int_0^1 e^{-\lambda x} f_{(b,r)}(x) dx$$

これより極限$\lim_{n\to\infty} Y_n = Y$はベータ分布$f_{(b,r)}(x)$に従うことが分かる。

さらに一般化して、壷の中にk種の色の球がそれぞれ、α_1個、α_2個、・・、α_k個ずつ入っているとする。壷から１個ランダムに球を取り出し、同じ色の球を１個付け加えて壷に戻す。このとき、n回の試行の後、各色の球の割合を$Y_n = (Y_n^{(1)}, \cdots, Y_n^{(k)})$とすると、極限$\lim_{n\to\infty} Y_n = Y$が存在し、$Y$はディリクレ分布$D(\alpha_1, \cdots, \alpha_k)$に従うことが示される。

Blackwell&MacQueen(1973)はポリアの壷とディリクレ分布の関係をより一般化した定理を示した。

定義：ディリクレ過程

完備で可分な距離空間 X 上の有界な正値測度 μ とランダム確率測度 μ^* があり、X の任意の可測な有限分割 $(B_1, B_2, \cdots, B_r)$ に対して、r 次元ベクトル $(\mu^*(B_1), \cdots, \mu^*(B_r))$ がパラメーター $(\mu(B_1), \cdots, \mu(B_r))$ のディリクレ分布に従うとき μ^* はパラメーター μ のディリクレ過程と呼ばれる。ただし、$\mu(B_i) = 0$ ならば $\mu^*(B_i) \equiv 0$ とする。詳しくは Ferguson(1973) を参照されたい。

定義：ポリア系列

X 値の確率変数系 $\{X_1, X_2, X_3, \cdots\}$ が次の条件を満たすとき、パラメーター μ のポリア系列という。任意の可測部分集合 $B \subset X$ に対して

（1） $P(X_1 \in B) = \dfrac{\mu(B)}{\mu(X)} \quad , j = 1,2,\cdots,r$

（2） $P(X_{n+1} \in B | X_1, X_2, \cdots, X_n) = \dfrac{\mu_n(B)}{\mu_n(X)}$ ， ただし $\mu_n(X) = \mu(X) + n$

$$\mu_n(B) = \mu(B) + \sum_{i=1}^{n} \delta_{X_i}(B) \quad , \ \delta_{X_i}(B) = \begin{cases} 1 & if \ X_i \in B \\ 0 & if \ X_i \notin B \end{cases}.$$

これは X が有限集合 $X = \{1,2,\cdots,r\}$ のとき、最初 r 色の球がそれぞれ $\mu_1, \mu_2, \cdots, \mu_r$ 個ずつ入っている壺から 1 個球を取り出し、その球と同色の球を 1 個付け加えて壺に戻すという試行を繰り返す、ポリアの壺を一般化したものである。この時、次の定理が成り立つ。

定理４．１２　Blackwell&MacQueen(1973)

$\{X_n ; n = 1,2,3,\cdots\}$ をパラメーター μ のポリア系列とする。このとき、次が成り立つ。

（1） $m_n = \dfrac{\mu_n}{\mu_n(X)}$ は $n \to \infty$ のとき、X 上のランダム確率測度 μ^* に確率 1 で収束する。

（2） μ^* はパラメーター μ の Ferguson 分布を持つ。

（3） μ^* を 1 つ与えたとき（μ^*_ω とする）、$X_1, X_2, X_3, \cdots$ は分布 μ^*_ω に従う独立な確率変数系である。

（証明）

ここでは X が有限集合 $X = \{1,2,\cdots,r\}$ の場合のみ示す。一般の場合は Blackwell and MacQueen(1973)を参照されたい。

μ^* を X 上のパラメーター μ のディリクレ過程とする。すなわち、$(\mu^*(1), \mu^*(2), \cdots, \mu^*(r))$

はパラメーター$(\mu(1),\cdots,\mu(r))$のディリクレ分布に従い、$\sum_{j=1}^{r}\mu^*(j)=1$。またμ_ω^*を与えたとき、$X_1,X_2,X_3,\cdots$を分布μ_ω^*に従う独立な確率変数系とする。π_nを$X_1,X_2,\cdots,X_n$の経験分布、$\pi_n(j)=\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(j)$とする。大数の強法則より$n\to\infty$のとき$\pi_n$は$\mu_\omega^*$に確率1で収束する。$m_n=\frac{\mu+n\pi_n}{\mu(X)+n}$より、$n\to\infty$のとき、確率1で$\mu^*$に収束する。従って最後に$\{X_1,X_2,\cdots\}$がパラメーター$\mu$のポリア系列であることを証明すればよい。その為には$A=\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$, $x_i\in\{1,2,\cdots,r\}$とするとき、その確率が

$P(A)=\frac{\prod_{j=1}^{r}(\mu(j))_{(n(j))}}{\mu(X)_{(n)}}$、ただし$n(j)=^{\#}\{i\ ;x_i=j\}$, $a_{(n)}=a(a+1)\cdots(a+n-1)$で与えられることを示せば十分である。$\mu_\omega^*$を与えたとき、$X_1,X_2,X_3,\cdots$は独立で$\mu_\omega^*$に従うので

$P(A|\mu_\omega^*)=\prod_{i=1}^{n}P(X_i=x_i|\mu_\omega^*)=\prod_{j=1}^{r}(\mu_\omega^*(j))^{n(j)}$、$\mu_\omega^*$の分布で平均し

$$P(A)=E\left[\prod_{j=1}^{r}(\mu^*(j))^{n(j)}\right]\qquad\text{とすると}\tag{4.26}$$

$(\mu^*(1),\mu^*(2),\cdots,\mu^*(r))$はパラメーター$(\mu(1),\cdots,\mu(r))$のディリクレ分布に従うので、$\mu(i)=\alpha_i$, $n(j)=n_j$と書くと(4.26)より

$$\begin{aligned}P(A)&=\frac{\Gamma(\alpha_1+\cdots+\alpha_r)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_r)}\int\cdots\int\left(\prod_{j=1}^{r}x_j^{\alpha_j+n_j-1}\right)dx_1\cdots dx_{r-1}\\&=\frac{\Gamma(\alpha_1+\cdots+\alpha_r)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_r)}\times\frac{\Gamma(\alpha_1+n_1)\Gamma(\alpha_2+n_2)\cdots\Gamma(\alpha_r+n_r)}{\Gamma(\alpha_1+\cdots+\alpha_r+n_1+\cdots+n_r)}\\&=\frac{\Gamma(\alpha_1+\cdots+\alpha_r)}{\Gamma(n+\alpha_1+\cdots+\alpha_r)}\times\frac{\Gamma(\alpha_1+n_1)}{\Gamma(\alpha_1)}\times\frac{\Gamma(\alpha_2+n_2)}{\Gamma(\alpha_2)}\times\cdots\times\frac{\Gamma(\alpha_r+n_r)}{\Gamma(\alpha_r)}\\&=\frac{\alpha_{1(n_1)}\alpha_{2(n_2)}\cdots\alpha_{r(n_r)}}{(\alpha_1+\cdots+\alpha_r)_{(n)}}=\frac{\prod_{j=1}^{r}(\mu(j))_{(n(j))}}{(\mu(X))_{(n)}}\end{aligned}$$

よって$\{X_1,X_2,X_3,\cdots\}$はポリア系列である。　　（証明終わり）

４．５　ポアソン・ディリクレ分布

４．５．１　ポアソン・ディリクレ分布

領域$\Delta_n=\{\vec{x}=(x_1,\cdots,x_n);\ x_i\geq 0,\ \sum_{i=1}^{n}x_i\leq 1\}$上の対称ディリクレ分布$D(\alpha,n+1)$の分布密度は　$f_n(\vec{x};\alpha)=\dfrac{\Gamma((n+1)\alpha)}{\{\Gamma(\alpha)\}^n}\left(\prod_{i=1}^{n+1}x_i\right)^{\alpha-1}$　ただし$x_{n+1}=1-\sum_{i=1}^{n}x_i$　　(4.27)

明らかに、$E[x_i]=\dfrac{1}{n+1}$で$\lim_{n\to\infty}E[x_i]=0$より$n\to\infty$における$\vec{x}=(x_1,\cdots,x_n)$の極限分布をうまく構成できない。Kingman(1975)は$(x_1,x_2,\cdots,x_n)$の順序統計量$x_{(1)}\geq x_{(2)}\geq\cdots\geq x_{(n)}$の分布を考え、$n\to\infty$で極限分布（ポアソン・ディリクレ分布）が存在することを示した。

$\nabla_\infty=\left\{\vec{x}=(x_1,x_2,x_3,\cdots;\ x_1\geq x_2\geq x_3\geq\cdots,\ \sum_{i=1}^{\infty}x_i=1\right\}$とする。

定理４．１３（Kingman(1975)）

$0<\theta<\infty$に対して、次の性質を持つ∇_∞上の測度P_θが存在する。

$(x_1,x_2,\cdots,x_n)$が対称ディリクレ分布に従うとき、任意のｋに対してその順序統計量$x_{(1)}\geq x_{(2)}\geq\cdots\geq x_{(k)}$の分布は$n\alpha\to\theta,n\to\infty$の極限で$P_\theta$のｋ次元周辺分布に収束する。

（証明）

$\{\xi(t);\ t\geq 0\}$をΓ過程とする。すなわち、$\xi(0)=0,\ \xi(t_2)-\xi(t_1)$は shape parameter $\alpha=t_2-t_1$、scale parameter $\lambda=1$のΓ分布に従う。Γ過程は独立増分を持つ加法過程でジャンプのみで増加する。各nについて

$$q_j=q_j(n)=\{\xi(j\alpha)-\xi((j-1)\alpha)\}/\xi(n\alpha)\qquad(4.27)$$

とすると、定理４．５より$q=(q_1,q_2,\cdots,q_n)$は対称ディリクレ分布$D(\alpha;n+1)$に従う。

$q=(q_1,q_2,\cdots,q_n)$の順序統計量を$(q_{(1)},q_{(2)},\cdots,q_{(n)})$とする。(4.27)より$n\alpha\to\theta,n\to\infty$の極限を取ると、後述するように極限が存在し

$$\lim_{n\to\infty}q_{(j)}=\frac{\delta\xi_{(j)}}{\xi(\theta)}\qquad(4.28)$$

ここで$\delta\xi_{(j)}$はΓ過程$\xi(t)$の区間$(0,\theta)$におけるj番目に大きなジャンプを表す。

$\xi(\theta)=\sum_{j=1}^{\infty}\delta\xi_{(j)}$より$\left(\dfrac{\delta\xi_{(1)}}{\xi(\theta)},\dfrac{\delta\xi_{(2)}}{\xi(\theta)},\cdots\right)\in\nabla_\infty$であり、極限分布$P_\theta$の存在が示される。

そこで最後に(4.28)を証明する。Γ過程の一つのサンプルに対してN_0を十分大きく取り、$N \geq N_0$のとき大きな方からk個のジャンプ$\{\delta\xi_{(j)}; j=1,2,..,k\}$が$(0, N\alpha_N)$をN等分した異なる区間$((i-1)\alpha_N, i\alpha_N)$に生じているようにしておく。(4.27)より

$$\xi(N\alpha_N)q_{(j)} \geq \delta\xi_{(j)} \quad , j=1,2,\cdots,k \ (N \geq N_0)、故に \quad \liminf_{N\to\infty} q_{(j)} \geq \frac{\delta\xi_{(j)}}{\xi(\theta)} \qquad (4.29)$$

kは任意なので全ての$j=1,2,3,\cdots$について(4.29)は成り立つ。

$$また \limsup_{N\to\infty} q_{(j)} = \limsup_{N\to\infty}\left(1-\sum_{i\neq j} q_{(i)}\right) \leq 1-\sum_{i\neq j}\liminf_{N\to\infty} q_{(j)} \leq 1-\sum_{i\neq j}\frac{\delta\xi_{(i)}}{\xi(\theta)} = \frac{\delta\xi_{(j)}}{\xi(\theta)} \quad (4.30)$$

(4.29)(4.30)より(4.28)が成り立つ。　　(証明終わり)

より一般に非対称ディリクレ分布$D(\alpha_1, \alpha_2, \cdots, \alpha_{n+1})$の場合はΓ過程$\xi(t)$に対して、

$$q_{(j)}(n) = \frac{\xi(\alpha_1+\cdots+\alpha_j)-\xi(\alpha_1+\cdots+\alpha_{j-1})}{\xi(\alpha_1+\cdots+\alpha_{n+1})} \qquad (4.31)$$

と置き、$n\to\infty$, $\alpha_1+\cdots+\alpha_{n+1}\to\theta$かつ$\max(\alpha_1,\cdots,\alpha_{n+1})\to 0$となる様に極限をとると$q_{(j)}(n)$が収束することが同様に示される。この極限分布をポアソン・ディリクレ分布と呼び、$PD(\theta)$で表す。$PD(\theta)$の具体的な形は扱い易い形ではない。Watterson(1976)は対称ディリクレ分布の順序統計量$(x_{(1)}, x_{(2)}, \cdots, x_{(k)})$の分布から求めているので、結果だけを紹介しよう。$PD(\theta)$の$k$次元周辺分布を$f_\theta(x_1, x_2, \cdots, x_k)$とすると

$$f_\theta(x_1,x_2,\cdots,x_k) = \theta^k\Gamma(\theta)e^{\gamma\theta}g\left(\frac{1-x_1-\cdots-x_k}{x_k}\right)\left(\prod_{i=1}^{k-1}x_i\right)^{-1}x_k^{\theta-2} \qquad (4.32)$$

ここで$1\geq x_1\geq x_2\geq\cdots\geq x_k\geq 0$, $\sum_{i=1}^{k}x_i \leq 1$　　$k=1,2,3,\cdots$

$g(\cdot)$はラプラス変換が次の式で与えられる無限分解可能な分布密度関数である。

$$E[\exp(-tz)] = \int_0^\infty e^{-tz}g(z)dz = \exp(-\gamma\theta)t^{-\theta}\exp[-\theta E_1(t)]$$

ただし　$E_1(t) = \int_t^\infty e^{-x}x^{-1}dx = -\gamma - \log t - \sum_{k=1}^{\infty}\frac{(-t)^k}{k\cdot k!}$、$\gamma = 0.57721$（オイラーの定数)。

4．5．2　ポアソン・ディリクレ分布のsize-biased permutationとEwensのサンプリング公式

前節で述べたようにディレクレ分布から順序統計量の分布の極限としてえられたポアソン・ディレクレ分布は扱い易い形とは言い難い。しかし、以下で示すように4．4．2

節で導入したsize-viased permutationの分布を考えると非常に扱いやすい表現が得られる。すなわち、対称ディレクレ分布$D(\alpha;n)$のsize-biased permutationは$Z_j\,(1\le j\le n-1)$がベータ分布$Beta(1+\alpha,(n-j)\alpha)$に従う一般化されたディレクレ分布であり、

$\vec{P}^S=(P_1^S,P_2^S,\cdots,P_k^S)$, $P_1^S=Z_1, P_2^S=(1-Z_1)Z_2, P_3^S=(1-Z_1)(1-Z_2)Z_3,\cdots$、

$P_n^S=(1-Z_1)\cdots(1-Z_{n-1})Z_n$　（ただし$Z_n=1$）と表される。これより、$n\alpha\to\theta$, $n\to\infty$の極限をとると、$Z_j\,(j=1,2,3,\cdots)$はベータ分布$Beta(1,\theta)$に従う独立な確率変数であり、$P_1^S=Z_1, P_2^S=Z_1(1-Z_2),\cdots,P_n^S=Z_1(1-Z_2)\cdots(1-Z_{n-1})Z_n,\cdots$と表される。

この$P^S=(P_1^S,P_2^S,P_3^S,\cdots)$がポアソン・ディリクレ分布のsize-biased permutationである。ベータ分布$Beta(1,\theta)$に従う独立な確率変数系$\{Z_j\,(j=1,2,3,\cdots)\}$に対して上で定義した確率変数系$P_1^S\ P_2^S,\cdots,P_n^S,\cdots$はG.E.M.(Griffiths-Engen-McCloskey)分布と呼ばれている。

４．６　Ewensのサンプリング公式とポリアの壷様モデル

定理４．１でEwensのサンプリング公式を遺伝子系図を用いて導いたが、ポリアの壷モデルとの関係がHoppe(1984)によって示された。このポリアの壷モデルを用いたサンプリング公式に関連した種々の性質の証明と遺伝子系図との関係などについてHoppe(1984, 1987)およびDonnelly(1986)に従って紹介する。

４．６．１　ポリアの壷様モデル

１個の黒球と他の種々の色の球が入った壷を考える。黒球の重さをϑ、他の色の球は全て重さ１とする。壷からそれぞれの球の重さに比例した確率でランダムに１個の球を取り出す。取り出した球が黒以外の球であれば、その球と同じ色の球を１個付け加えて壷に戻す。取り出した球が黒球のときは壷の中にない新しい色の球を付け加えて黒球と一緒に壷に戻す。最初($t=0$)、壷の中には黒球１個のみ入っているとする。時刻$t=1$に黒球を取り出し、最初に黒球と一緒に壷に入れる球のラベル（色）を１とする。以後は新たに壷に付加する色を順に2,3,…と自然数でラベルする。n回目の試行で壷に戻す球のラベルを確率変数X_nで現わす。例えば$X_1=1,X_2=1,X_3=2,X_4=1,X_5=3$などとなる。この５回の試行の結果、壷の中にはラベル１の球が３個とラベル２，３の球がそれぞれ１個、黒球が１個の計６個の球が入っている。n回の試行の後、壷の中の球のラベルの種類を確率変数Kで表す。この壷モデルは黒球を入れることにより新しい色の球が付加されて行くのでポリ

アの壷様モデルと呼ぶことにする。ラベル i$(1 \le i \le K)$ の球の数を n_i とすると $n_1 + n_2 + \cdots n_K = n$ であり、この占有数 $\{n_1, n_2, \cdots, n_K\}$ によって次の様に n の分割(partition)が決まる。$1 \le i \le n$ に対して、i 個の球を含むラベルの種類を a_i とする。すなわち、$a_i = {}^{\#}\{j; n_j = i\}$, $\vec{a} = (a_1, \cdots, a_n)$ とする。明らかに $\sum_{i=1}^{n} i a_i = n$ である。n 個の球が全て同じラベル（同色）ならば $\vec{a} = (0, \cdots, 0, 1)$ であり、全て異なるならば $\vec{a} = (n, 0, \cdots, 0)$ である。$\vec{a} = (a_1, a_2, \cdots, a_n)$ を n のアレル分割(allelic partition)と呼ぶ。アレル分割と呼ぶ理由は n 回の試行によって得られる確率変数系 $\{X_1, X_2, \cdots, X_n\}$ によって決まる n のアレル分割を Π_n とするとき、Π_n の分布が Ewens のサンプリング公式（定理４．１）と同じ分布になることによる。すなわち次の定理が成り立つ。

定理４．１４(Hoppe(1984))

Π_n は次の分布に従うマルコフ過程である。

$$P(\Pi_n = \vec{a}) = \frac{n!}{\vartheta_{(n)}} \prod_{i=1}^{n} \frac{\vartheta^{a_i}}{i^{a_i} a_i!} \quad 、\ \vartheta_{(n)} = \vartheta(\vartheta+1)\cdots(\vartheta+n-1),\ \vec{a} = (a_1, \cdots, a_n),\ \sum_{i=1}^{n} i a_i = n$$

（証明）

ポリアの壷様モデル $\{X_1, X_2, \cdots, X_n\}$ から決まる占有数を $\{n_1, n_2, \cdots, n_K\}$、さらにこの占有数から得られるアレル分割を $\vec{a} = (a_1, a_2, \cdots, a_n)$ とする。占有数が $\{n_1, n_2, \cdots, n_K\}$ となる一つのサンプルを $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ とすると、その確率は次式で与えられる。

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \frac{\vartheta^K}{\vartheta_{(n)}} \prod_{i=1}^{K} (n_i - 1)! \qquad (4.33)$$

なぜなら１回球を取り出すごとに壷の中の球の数が一つずつ増えることより、分母は $\vartheta_{(n)}$ となる。分子の ϑ^K は K 種類のラベルの球があることより、黒球が K 回取り出され、そのたびに新たなラベルの球が付加されたことに対応する。さらにその各々に対してラベル j $(1 \le j \le K)$ の球が１個ずつ $n_j - 1$ 回付加されるので、その場合の数が $\prod_{i=1}^{K} (n_i - 1)!$ となる。以上より(4.33)が得られる。従って、アレル分割 $\vec{a} = (a_1, a_2, \cdots, a_n)$ となる異なるサンプル

$\{X_1, X_2, \cdots, X_n\}$ が $\dfrac{n!}{\left(\prod_{i=1}^{n} i^{a_i} a_i!\right)\prod_{j=1}^{K} (n_j - 1)!} = \dfrac{n!}{\left(\prod_{i=1}^{n} a_i!\right)\left(\prod_{j=1}^{K} n_j!\right)}$ 通りあることを示せば、この因子を(4.33)に掛けて証明が終わる。i 個のサンプルを含むアレルタイプの数が a_i 種あり、

それらのラベルの与え方はアレル分割$\vec{a}=(a_1,a_2,\cdots,a_n)$に対して$\prod_{i=1}^{n}a_i!$通りある。

また、n_j個は同じアレルタイプなので、その並べ方は全部で$\prod_{j=1}^{K}n_j!$通りある。ポリアの壷様モデルでは出現した順にラベルを自然数で与えるので、全て異なるときの順列$n!$をアレルのラベルと同じアレルの順列から求めた上記の場合の数の積$\prod_{i=1}^{n}a_i!\prod_{j=1}^{K}n_j!$で割るとアレル分割$\vec{a}=(a_1,a_2,\cdots,a_n)$に対応するポリアの壷様モデルのサンプル$\{X_1,X_2,\cdots,X_n\}$の場合の数となる。

定理４．１２でポリアの壷モデルとディリクレ分布の関係について述べた。この定理の応用として、修正されたポリアの壷モデル（ポリアの壷様モデル）とポアソン・ディリクレ分布について関係が導かれる。$i=1,2,\cdots,K$に対して、ラベルiで重さ１の球がα_i個ずつ入っている壷を考える。この壷から毎時刻１個の球をランダムに取り出し、その球と同じラベルの球を付け加えて戻すポリアの壷モデルを考える。ただし、α_iが非整数の場合はα_iに比例した確率でラベル i の球は取り出されるとする。n回目の試行で壷に付加した球のラベルをX_n, $\sum_{i=1}^{K}\alpha_i=\tilde{\vartheta}$とする。

このとき、明らかに$P\left(X_{n+1}=i\middle|\ X_1,X_2,\cdots,X_n\right)=\frac{\alpha_i+\nu_i(n)}{\tilde{\vartheta}+n}$が成り立つ。ここで、$\nu_i(n)$は$n$回の試行の後、壷の中のラベルｉの球の数を表す。定理４．１２より

$\lim_{n\to\infty}\frac{\nu_i(n)}{n}=p_i$, $\sum_{i=1}^{K}p_i=1$ $(a.s.)$であり、$P=(p_1,p_2,\cdots,p_K)$はディリクレ分布$D(\alpha_1,\cdots,\alpha_K)$に従う。以上を準備として次の定理を証明する。

定理４．１５　(Hoppe(1987))

ポリアの壷様モデルのサンプル$\{X_1,X_2,\cdots,X_n\}$においてｎ回の試行の後の壷の中のラベルｉの球の数を$S_i(n)$とすると　$\lim_{n\to\infty}\frac{S_i(n)}{n}=P_i$, $\sum_{i=1}^{\infty}P_i=1$ $(a.s.)$が成り立つ。

（証明）

$\{\lambda_i;i=1,2,\cdots\}$を分布$P(\lambda_i=1)=\frac{\vartheta}{\vartheta+i-1}$, $P(\lambda_i=0)=\frac{i-1}{\vartheta+i-1}$に従う独立な確率変数系とする。すなわち、$\lambda_i=1$のとき黒球が取り出され、新しいラベルの球が壷に入れられる。マルコフ時刻τ_i $(i=1,2,\cdots)$を$\tau_1=1$、$i\geq 2$に対して$\tau_i=\min\{j>\tau_{i-1}\ ;\ \lambda_j=1\}$で定義する。$\tau_i$はラベルｉの球が初めて壷に入れられた時刻を表す。整数$k(\geq 1)$を固定し、$n=1,2,\cdots$

に対して、球が初めて壷に入れられた時刻によって球の再ラベリングをしたマルコフ連鎖 $\{Y_n^{(k)};n=1,2,3\cdots\}$ を次のように定義する。

$$Y_n^{(k)}=\begin{cases} j & (1\le j\le k,\ \lambda_j=1,\ X_{n+k}=X_j\text{のとき}) \\ 0 & (\text{その他}) \end{cases}。$$

最初のポリアの壷様モデル $\{X_1,X_2,\cdots,X_n\}$ では出現する順に１，２，３、・・・と球にラベリングしたが、$\{Y_n^{(k)};n=1,2,\cdots\}$ はｋ回の試行 $\{X_1,X_2,\cdots,X_k\}$ の条件下での黒球を含めて最大 $k+1$ 色のポリアの壷モデルである。そして初期条件としてのｋ回の試行 $\{X_1,X_2,\cdots,X_k\}$ の中でその色の球が最初に付加された時刻をその色のラベルとし、黒球はラベルを０とした再ラベリングである。そして、$\alpha_i^{(k)}=\begin{cases}\lambda_i S_{X_i}(k) & (1\le i\le k)\\ \vartheta & (i=0)\end{cases}$ によって $(\alpha_0^{(k)},\alpha_1^{(k)},\cdots,\alpha_k^{(k)})$ を定義する。例えば $k=7$ で、サンプルが

$(X_1,X_2,X_3,X_4,X_5,X_6,X_7)=(1,1,2,1,3,2,3)$ とすると、

$(\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5,\lambda_6,\lambda_7)=(1,0,1,0,1,0,0)$ であり、$\alpha_1^{(7)}=\lambda_1 S_{X_1}(7)=S_1(7)=3,$

$\alpha_2^{(7)}=\lambda_2 S_{X_2}(7)=0,\quad \alpha_3^{(7)}=\lambda_3 S_{X_3}(7)=S_2(7)=2,\quad \alpha_4^{(7)}=\lambda_4 S_{X_4}(7)=0,$
$\alpha_5^{(7)}=\lambda_5 S_{X_5}(7)=S_3(7)=2,\quad \alpha_6^{(7)}=\lambda_6 S_{X_6}(7)=0,\quad \alpha_7^{(7)}=\lambda_7 S_{X_7}(7)=0$

すなわち、$(\alpha_0^{(7)},\alpha_1^{(7)},\alpha_2^{(7)},\alpha_3^{(7)},\alpha_4^{(7)},\alpha_5^{(7)},\alpha_6^{(7)},\alpha_7^{(7)})=(\vartheta,3,0,2,0,2,0,0)$ となる。

$\{Y_n^{(k)};n=1,2,\cdots\}$ は初期条件 $(\alpha_0^{(k)},\alpha_1^{(k)},\cdots,\alpha_k^{(k)})$ のポリアの壷モデルになる。ポリアの壷モデル $\{Y_n^{(k)};n=1,2,3\cdots\}$ においてｎ回の試行でラベルｉ、すなわち時刻ｉに最初に壷に付加された球が $k+1$ 回目以降、ｎ回の試行で壷に付加された回数を

$S_i(n,k)=^{\#}\{j;Y_j^{(k)}=i,1\le j\le n\}\quad (i=0,1,2,\cdots,k)$ とする。定理４．１２より

$$P\left(\lim_{n\to\infty}\frac{S_i(n,k)+\alpha_i^{(k)}}{n+k+\theta}=\gamma_i^{(k)}\,\middle|\,X_1,X_2,\cdots,X_k\right)=1,\quad \sum_{i=0}^{K}\gamma_i^{(k)}=1。\ \gamma_i^{(k)}\text{は}(\alpha_0^{(k)},\cdots,\alpha_k^{(k)})$$

に依存して決まる確率変数である。全てのサンプル $\{X_1,X_2,\cdots,X_k\}$ について、期待値を取ると、確率１で極限 $\lim_{n\to\infty}\frac{S_i(n,k)+\alpha_i^{(k)}}{n+k+\vartheta}$ が存在することが分かる。他方、$S_i(n,k)$ の定義よ

り $S_i(n,k)+\alpha_i^{(k)}=\alpha_i^{(n+k)}$ なので、$\lim_{n\to\infty}\frac{S_i(n,k)+\alpha_i^{(k)}}{n+k+\vartheta}=\lim_{n\to\infty}\frac{\alpha_i^{(n+k)}}{n+k+\vartheta}=\gamma_i^{(k)}$ であり極限値はｋによらない。この極限を $\lim_{n\to\infty}\frac{\alpha_i^{(n+k)}}{n+k+\vartheta}=\lim_{n\to\infty}\frac{\alpha_i^{(n)}}{n}=\gamma_i\ \ (1\le i<\infty)$　(a.s)とする。

条件 $\{X_1,X_2,\cdots,X_k\}$ の下で、定理４．１２より $(\gamma_0^{(k)},\gamma_1^{(k)},\cdots,\gamma_k^{(k)})$ はディリクレ分布 $D(\alpha_0^{(k)},\alpha_1^{(k)},\cdots,\alpha_k^{(k)})$（ただし $\alpha_0^{(k)}=\theta$）に従う。その周辺分布として $\gamma_0^{(k)}$ はベータ分布 $Beta(\theta,\sum_{i=1}^{k}\alpha_i^{(k)})=Beta(\vartheta,k)$ に従うので、$E\left[\gamma_0^{(k)}\middle|X_1,\cdots,X_k\right]=\frac{\vartheta}{\vartheta+k}$ となる。

全ての初期条件 $\{X_1,X_2,\cdots,X_k\}$ に関して期待値を取ると $E[\gamma_0]=E\left[1-\sum_{i=1}^{k}\gamma_i\right]=\frac{\vartheta}{\vartheta+k}$ 。

これより、$E\left[1-\sum_{i=1}^{\infty}\gamma_i\right]=E\left[1-\lim_{k\to\infty}\sum_{i=1}^{k}\gamma_i\right]=\lim_{k\to\infty}E[1-\sum_{i=1}^{k}\gamma_i]=\lim_{k\to\infty}\frac{\vartheta}{\vartheta+k}=0$　。

$1-\sum_{i=1}^{\infty}\gamma_i\ge 0$ (a.s.) なので、$\sum_{i=1}^{\infty}\gamma_i=1$ である。最初のポリアの壺様モデル $\{X_1,X_2,\cdots,X_n\}$ に戻って、ラベルｉの球が最初に付加される時刻を τ_i とすると、$n\ge\tau_i$ のとき、$S_i(n)=\alpha_{\tau_i}^{(n)}$ である。これより $\lim_{n\to\infty}\frac{S_i(n)}{n}=\gamma_{\tau_i}\equiv P_i$　(a.s.) と置くことにする。$(\gamma_1(\omega),\gamma_2(\omega),\gamma_3(\omega)\cdots)$ の中で正の値は $(\gamma_{\tau_1}(\omega),\gamma_{\tau_2}(\omega),\gamma_{\tau_3}(\omega),\cdots)$ であり、ポリアの壺様モデルにおいて有限個の色（ラベル）しか現れない確率は

$P(\text{有限個の } i \text{でのみ} \tau_i<\infty)=\lim_{n\to\infty}P(\lambda_i=0\ ;\text{全ての } i\ge n)=\lim_{n\to\infty}\prod_{i=n}^{\infty}\frac{i-1}{\vartheta+i-1}=0$。

よって、確率１で無限個のラベルが現れるので $\sum_{i=1}^{\infty}P_i=\sum_{i=1}^{\infty}\gamma_{\tau_i}=1$ である。

定理４．１６

$P_1,P_2,\cdots P_n\cdots$ を定理４．１５で定義される確率変数とするとき、

（１）$Z_n=\frac{P_n}{\sum_{i=n}^{\infty}P_i}$，$(n=1,2,3,\cdots)$ とすると、$\{Z_1,Z_2,Z_3\cdots\}$ はベータ分布 $Beta(1,\vartheta)$ に従う独立な確率変数系である。

（２）$P_1=Z_1$，$P_2=Z_2(1-Z_1)$，$\cdots$、一般に $P_n=Z_n\prod_{i=1}^{n-1}(1-Z_i)$　$(n\ge 2)$。

すなわち $\{P_1,P_2,\cdots P_n\cdots\}$ はポアソン・ディリクレ分布の size-biased permutation で

ある。

（3） $\lambda_j = 1$のとき、γ_jは$Beta(1, \vartheta + j - 1)$に従う。

(証明)

（1） 第４．４．２節の式(4.23)で述べたように、$Y_1, Y_2, \cdots, Y_{n+1}$がディリクレ分布

$D(\alpha_1, \alpha_2, \cdots, \alpha_{n+1})$に従うとき、$\vec{Y} = (Y_1, \cdots, Y_n)$は完全中立な確率ベクトルであり、

$U_j = Y_j \Big/ (1 - \sum_{i=1}^{j-1} Y_j) = Y_j \Big/ (\sum_{i=j}^{n+1} Y_j)$ $(j = 1,2,\cdots,n)$は独立でベータ分布$Beta(\alpha_j, \sum_{i=j+1}^{n+1} \alpha_i)$に従う。ポリアの壷様モデル$\{X_i; i \geq 1\}$において、ラベル n の球が最初に壷に入れられた時刻をτ_nとする。時刻τ_nでのラベルk $(1 \leq k \leq n)$の球の数を$\alpha_k = S_k(\tau_n)$とする。明らかに$\alpha_n = 1$であり、壷の中には黒球が１個あるので$\alpha_{n+1} = \vartheta$とする。初期条件$(\alpha_1, \cdots, \alpha_n, \alpha_{n+1})$の下で、ポリアの壷モデルを$\{Y_k^{(n)}; k = 1,2,3,\cdots\}$とする。ここでラベル$n+1$は黒色であり、黒色の球が取り出されたときは、黒球を１個付加して壷に戻す。条件$(X_1, X_2, \cdots, X_{\tau_n})$の下で$(P_1, P_2, \cdots, P_{n+1})$はディレクレ分布$D(\alpha_1, \alpha_2, \cdots, \alpha_{n+1})$に従う。故に、$(X_1, X_2, \cdots, X_{\tau_n})$の下で$Z_i = P_i \Big/ \sum_{j=i}^{n+1} P_j$ $(1 \leq i \leq n)$はベータ分布$Beta(\alpha_i, \sum_{j=i+1}^{n+1} \alpha_j)$に従う独立な確率変数である。特に$\alpha_n = 1, \alpha_{n+1} = \theta$より条件$(\alpha_1, \cdots, \alpha_n, \alpha_{n+1})$の下で$Z_1, Z_2, \cdots, Z_n$の同時分布は

$$\left[\prod_{i=1}^{n-1} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} z_i^{\alpha_i - 1} (1 - z_i)^{\beta_i - 1} \right] \times \vartheta (1 - z_n)^{\vartheta - 1} \qquad \text{ただし} \beta_i = \sum_{j=i+1}^{n+1} \alpha_j \text{。}$$

初期条件$(\alpha_1, \cdots, \alpha_n, \alpha_{n+1})$の下で、$(Z_1, \cdots, Z_{n-1})$と$Z_n$は独立である。
初期条件$(\alpha_1, \cdots, \alpha_{n-1}, \alpha_n, \alpha_{n+1}) = (\alpha_1, \cdots, \alpha_{n-1}, 1, \vartheta)$の確率分布は$(\alpha_1, \cdots, \alpha_{n-1})$に関する確率分布なのでこれを上式に掛けると、$Z_1, Z_2, \cdots, Z_n$の同時分布は適当な関数$\Phi$を用いて$\Phi(z_1, \cdots, z_{n-1}) \vartheta (1 - z_n)^{\vartheta - 1}$と書ける。これより$(Z_1, \cdots, Z_{n-1})$と$Z_n$は独立で$Z_n$の分布は$\vartheta (1 - z_n)^{\vartheta - 1}$となることが分かる。nは任意なので、全ての自然数について成り立つ。

(2) $P_n = Z_n \left(\sum_{i=n}^{\infty} P_i \right) = Z_n (1 - P_1 - \cdots - P_{n-1}) = Z_n \left(\frac{1 - P_1 - \cdots - P_{n-1}}{1 - P_1 - \cdots - P_{n-2}} \right) (1 - P_1 - \cdots - P_{n-2})$

$$= Z_n\left(1-\frac{P_{n-1}}{1-P_1-\cdots-P_{n-2}}\right)(1-P_1-\cdots-P_{n-2}) = Z_n(1-Z_{n-1})(1-P_1-\cdots-P_{n-2})$$

$$= Z_n\prod_{i=1}^{n-1}(1-Z_i) \quad (n\geq 2)$$

（3）$\lambda_j = 1$のとき、この球のラベルをkとする。条件$(X_1, X_2, \cdots, X_j)$の下で

$(P_1, P_2, \cdots, P_k, P_{k+1})$はディレクレ分布$D(\alpha_1, \alpha_2, \cdots, \alpha_{k+1})$に従う、ただし$\alpha_k = 1$、$\alpha_{k+1} = \vartheta$

$\sum_{i=1}^{k+1}\alpha_i = j+\vartheta$である。これより、$P_k$の分布はディリクレ分布の周辺分布として、

補題４．４よりベータ分布$Beta(1, \sum_{i\neq k}\alpha_i) = Beta(1, \vartheta + j - 1)$に従う。

４．６．２　ポアソン・ディリクレ集団からのサンプルとポリアの壺様モデル

４．５．２節で述べたようにポアソン・ディレクレ分布からのサンプルのアレル分割分布はEwensのサンプリング公式に従う。Watterson(1976)はK個のアレルを含む定常な集団モデルからのサンプルの分布から極限操作によりEwensのサンプリング公式が導かれることを示した。K個のアレルを$1,2,\cdots,K$でラベルし、その集団内頻度を$X_1, X_2, \cdots, X_K$とする。集団からランダムにn個の遺伝子をサンプルしたとき、各々のアレルがn_1個、n_2個$\cdots, n_K$個含まれる確率は

$$P_n(n_1,\cdots,n_K|X) = \frac{n!}{\prod_{i=1}^{K} n_i!}\left(\prod_{j=1}^{K} X_j^{n_j}\right),\ X = (X_1,\cdots,X_K),\ (n_1+\cdots+n_K = n)。$$

$X = (X_1, \cdots, X_K)$がディリクレ分布$D(\varepsilon, K)$に従うとし、上式の期待値を取ると

$$P_n(n_1,\cdots,n_K) = \frac{n!}{\prod_{i=1}^{K} n_i!}\iint_{\Delta_K}\prod_{j=1}^{K} x_j^{n_j} \times \left\{\frac{\Gamma(K\varepsilon)}{(\Gamma(\varepsilon))^K}\left(\prod_{j=1}^{K} x_j^{\varepsilon-1}\right)\right\}dx_1\cdots dx_{K-1}$$

ただし、$\Delta_K = \{x = (x_1,\cdots,x_{K-1});\ x_j \geq 0,\ \sum_{j=1}^{K-1} x_j \leq 1\}$。ディリクレ分布の積分より

$$P_n(n_1,\cdots,n_K) = \frac{n!}{\prod_{i=1}^{K} n_i!} \times \frac{\Gamma(K\varepsilon)\prod_{j=1}^{K}\Gamma(\varepsilon+n_j)}{(\Gamma(\varepsilon))^K\Gamma(K\varepsilon+n)} \quad となる。$$

$n_1, n_2, \cdots, n_K$の中で$n_j \geq 1$であるアレルの数をkとし、大きさの順に$n_{(1)} \geq n_{(2)} \geq \cdots \geq n_{(k)}$

と並べる。kと$n_{(1)} \geq n_{(2)} \geq \cdots \geq n_{(k)}$を与えたとき、確率$P(n_{(1)}, n_{(2)}, \cdots, n_{(k)}; k)$は

$$P(n_{(1)},n_{(2)},\cdots,n_{(k)};k)=\frac{n!}{\prod_{l=1}^{k}n_{(l)}!}\times\frac{\Gamma(K\varepsilon)\prod_{l=1}^{k}\Gamma(\varepsilon+n_{(l)})}{(\Gamma(\varepsilon))^{k}\Gamma(K\varepsilon+n)}\times M$$

ここで、Mは異なるK種のアレルを$n_{(1)}\geq n_{(2)}\geq\cdots\geq n_{(k)}$となるように割り当てる場合の数であり、$n_{(1)},n_{(2)},\cdots,n_{(k)},0,\cdots,0$のK個の数を異なるK個の部屋に割り当てる場合の数に等しい。$\alpha_j=^{\#}\{i;n_{(i)}=j,1\leq i\leq k\}$とすると、同じものを含む順列になるので

$M=\dfrac{K!}{\alpha_1!\alpha_2!\cdots\alpha_n!(K-k)!}$である。ここで、アレルの数を無限に増やし$K\to\infty,\varepsilon\to 0,$
そして$K\varepsilon\to\vartheta$となるように極限を取ると

$$\lim_{K\to\infty}\frac{M}{(\Gamma(\varepsilon))^{k}}=\lim_{K\to\infty}\left\{\frac{1}{\alpha_1!\cdots\alpha_n!}\times\frac{K!}{(\Gamma(\varepsilon))^{k}(K-k)!}\right\}$$

$$=\frac{1}{\alpha_1!\alpha_2!\cdots\alpha_n!}\lim_{K\to\infty}\left\{\frac{K}{\Gamma(\varepsilon)}\right\}^{k}\left(1-\frac{1}{K}\right)\left(1-\frac{2}{K}\right)\cdots\left(1-\frac{k-1}{K}\right)$$

ここで、$\lim_{\varepsilon\to 0}\varepsilon\Gamma(\varepsilon)=\lim_{\varepsilon\to 0}\int_0^\infty\varepsilon t^{\varepsilon-1}e^{-t}dt=\lim_{\varepsilon\to 0}\Gamma(\varepsilon+1)=\Gamma(1)=1$より

$\lim_{\substack{K\to\infty\\K\varepsilon\to\theta}}\dfrac{K}{\Gamma(\varepsilon)}=\lim_{\varepsilon\to 0}\dfrac{\theta}{\varepsilon\Gamma(\varepsilon)}=\theta$なので$\lim_{K\to\infty}\dfrac{M}{(\Gamma(\varepsilon))^{k}}=\dfrac{\theta^{k}}{\alpha_1!\alpha_2!\cdots\alpha_n!}$。故に$K\to\infty$, $K\varepsilon\to\vartheta$の極限を取ると次のEwensのサンプリング公式を得る。

$$\lim_{\substack{K\to\infty\\K\varepsilon\to\theta}}P(n_{(1)},n_{(2)},\cdots,n_{(k)};k)=\frac{n!\vartheta^{k}\Gamma(\vartheta)}{n_{(1)}n_{(2)}\cdots n_{(k)}\alpha_1!\alpha_2!\cdots\alpha_n!\Gamma(\vartheta+n)}。$$

N個の半数体個体から成る有限集団をQ^Nとする。可算無限個のアレルタイプを仮定する。集団Q^Nから非復元抽出によりn個の遺伝子サンプルを取り出したとき、そのサンプル構成が$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n),\ \sum_{i=1}^{n}i\alpha_i=n$となる確率を$P_N(\alpha)$とする。$N\to\infty$のとき、この分布がEwensのサンプリング公式(4.1)に収束するとき、この集団の系列Q^NはEwensのサンプリング性を持つと言うことにしよう。Kingman(1977)は有限集団の系列Q^NがEwensのサンプリング性を持つための必要十分条件はこの集団がポアソン・ディリクレ極限を持つことであることを示した。

４．６．３　ポリアの壺様モデルおよび逆進プロセスの推移確率

ポリアの壺様モデルによるアレル分割Π_nのプロセスは壺の中に黒球１個の初期状態（これを$\Pi_0=\phi$で表す）から出発するマルコフ連鎖である。より一般に任意の初期状態Π_0から出発するアレル分割Π_nのマルコフプロセスの１ステップ推移確率を

$P(\vec{a},\vec{b})=P(\Pi_{n+1}=\vec{b}\,|\,\Pi_n=\vec{a}),\ \vec{a}=(a_1,\cdots,a_n),\ \vec{b}=(b_1,\cdots,b_{n+1})$とすると、明らかに

$$P(\vec{a},\vec{b})=\begin{cases}\dfrac{\vartheta}{\vartheta+n} & if\ \vec{b}=(a_1+1,a_2,\cdots,a_n,0)\\ \dfrac{ia_i}{\vartheta+n} & if\ \vec{b}=(a_1,\cdots,a_i-1,a_{i+1}+1,\cdots,a_n,0)\\ \dfrac{na_n}{\vartheta+n} & if\ \vec{b}=(a_1,\cdots,a_n-1,1)\end{cases}\qquad(4.34)$$

これより時間を逆向きにした逆進プロセスの推移確率を考えると

$P(\Pi_n=\vec{a}\,|\,\Pi_{n+1}=\vec{b})=\dfrac{P(\Pi_{n+1}=\vec{b}\,|\,\Pi_n=\vec{a})P(\Pi_n=\vec{a})}{P(\Pi_{n+1}=\vec{b})}$。これは初期分布$\Pi_0$に依存するのでポリアの壺様モデルの初期条件$\Pi_0=\phi$とすると、$P(\Pi_n=\vec{a})$は定理４．１４で与えられる。よって

$$P(\Pi_n=\vec{a}\,|\,\Pi_{n+1}=\vec{b})=\begin{cases}\dfrac{a_1+1}{n+1} & if\ \vec{b}=(a_1+1,a_2,\cdots,a_n,0)\\ \dfrac{(i+1)(a_{i+1}+1)}{n+1} & if\ \vec{b}=(a_1,\cdots,a_i-1,a_{i+1}+1,\cdots,a_n,0)\\ 1 & if\ \vec{b}=(a_1,\cdots,a_n-1,1)\end{cases}\qquad(4.35)$$

この逆進プロセスの推移確率行列を$M_{n,n+1}$で現わす。

$P_n(\vec{a})=P(\Pi_n=\vec{a})=P(\Pi_n=\vec{a}\,|\,\Pi_0=\phi)$とすると

$P_n(\vec{a})=\sum_{\vec{b}}P(\Pi_n=\vec{a}\,|\,\Pi_{n+1}=\vec{b})P(\Pi_{n+1}=\vec{b})$より次の漸化式を得る。

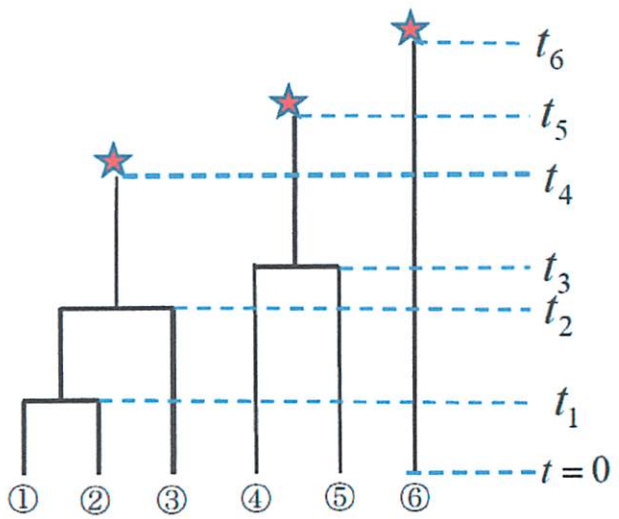
$$P_n(a_1,\cdots,a_n)=\frac{a_1+1}{n+1}P_{n+1}(a_1+1,a_2,\cdots,a_n,0)+\sum_{r=2}^{n}\frac{r(a_r+1)}{n+1}P_{n+1}(a_1,\cdots,a_{r-1}-1,a_r+1,\cdots,a_n,0)+P_{n+1}(a_1,\cdots,a_n-1,1)\qquad(4.36)$$

(4.36)は Kingman(1978)によっても Ewens のサンプリングの分割構造(partition structure)が満たす式として導出されている。同様にして任意の自然数$k < m < n$に対して行列M_{mn}をその成分が$M_{mn}(\vec{a},\vec{b}) = P(\Pi_m = \vec{a} | \Pi_n = \vec{b})$で定義される行列、$P_m, P_n$をその成分が$P_m(\vec{a})$, $P_n(\vec{b})$のベクトルとすると、線形変換$P_m = M_{mn}P_n$および$M_{kn} = M_{km}M_{mn}$が成り立つ。これは逆進プロセスの Chapman-Kolmogorov 方程式である。

４．６．４　合祖モデルとポリアの壺様モデル

第３章では(3.32)式で Ewens のサンプリング公式がモランモデルを用いて遺伝子系図過程から導かれることを示した。また、第４章の定理４．１では推移確率が(3.8)式で与えられる突然変異を含む遺伝子系図過程を用いて同様に Ewens のサンプリング公式を導いた。この節では突然変異を含む遺伝子系図、サンプリング公式およびポリアの様壺モデルの三者の間に深い関連があることを示す。

推移確率(3.8)の連続時間マルコフ連鎖を$\mathfrak{R}_t$とする。$\mathfrak{R}_t$は合祖と突然変異によって１つずつ減少する死滅過程であるが、$\mathfrak{R}_t$（ただし$\mathfrak{R}_0 = n$）のサンプルパスとして１つの系図が与えられたとき、次のような同値関係を定義する。$\mathfrak{R}_t = k$のとき k 個のメンバーの２つの遺伝子についてその２つの遺伝子が突然変異を経ることなく、共通祖先に到達するとき、この２つの遺伝子は同値とする。すなわちこの２つの遺伝子はある突然変異を祖先として共有していることを意味する。下の図で①と②は同値であるが②と⑤は同値ではない。



また、合祖をしたとき、その合祖によって現れた共通祖先には２つの遺伝子の若い番号をラベルとして与えることにする。上図で①と②が合祖した共通祖先は①というラベルを持つ。$\mathfrak{R}_t$の同値関係のプロセスを$\tilde{\mathfrak{R}}_t$とする。上の系図では$\tilde{\mathfrak{R}}_0 = \{(1,2,3),(4,5),(6)\}$、

$\tilde{\Re}_{t_1} = \{(1,3),(4,5),(6)\}$, $\tilde{\Re}_{t_2} = \{(1),(4,5),(6)\}$, $\tilde{\Re}_{t_3} = \{(1)(4)(6)\}$, $\tilde{\Re}_{t_4} = \{(4),(6)\}$

$\tilde{\Re}_{t_5} = \{(6)\}$, $\tilde{\Re}_{t_6} = \phi$ となる。突然変異を起こしていない個体について、その同値関係を時間を遡って追跡したプロセスである。同値関係のプロセス $\tilde{\Re}_t$ はまた、次のように分割(partition)のプロセス Ω_t を定義する。上の例では初期時刻にサンプル数がそれぞれ１，２，３の３つの部分に分割されているので $\Omega_0 = (a_1, a_2, a_3) = (1,1,1)$、以下同様にして

$\Omega_{t_1} = (1,2,0)$, $\Omega_{t_2} = (2,1,0)$, $\Omega_{t_3} = (3,0,0)$, $\Omega_{t_4} = (2,0,0)$, $\Omega_{t_5} = (1,0,0)$, $\Omega_{t_6} = (0,0,0)$ となる。Ω_t のジャンププロセスを $\Omega_k^*(k = n, n-1, .., 0)$ とする。上の例では $\Omega_6^* = (1,1,1)$、

$\Omega_5^* = (1,2,0)$, $\Omega_4^* = (2,1,0)$, $\Omega_3^* = (3,0,0)$, $\Omega_2^* = (2,0,0)$, $\Omega_1^* = (1,0,0)$, $\Omega_1^* = (0,0,0)$ となる。ジャンププロセス $\Omega_k^*(k = n, n-1, .., 0)$ の逆進過程を $\Omega_0^*, \Omega_1^*, \Omega_2^*, \cdots, \Omega_n^*$ とすると、次の定理が成り立つ。

定理４．１７

逆進プロセス $\Omega_0^*, \Omega_1^*, \Omega_2^*, \cdots, \Omega_n^*$ はポリアの壷様モデルから定義されるアレル分割のプロセス $\Pi_0, \Pi_1, \Pi_2, \cdots \Pi_n$ と確率論的に同値である。

（証明）

$\Omega_k^*(k = 0,1,2,\cdots,n)$ と $\Pi_k(k = 0,1,2,\cdots,n)$ の推移確率が一致することを示す。分割 $\vec{a} = (a_1, a_2, \cdots, a_n), \sum_{i=1}^{n} i a_i = n-1$; $\vec{b} = (b_1, b_2, \cdots, b_n), \sum_{i=1}^{n} i b_i = n$ に対して、推移確率を $P\left(\Omega_n^* = \vec{b} \middle| \Omega_{n-1}^* = \vec{a}\right)$ とする。

（１） $a_{r-1} = b_{r-1} + 1$, $a_r = b_r - 1$、かつ $j \neq r, r-1$ $(j \leq n-1)$ のとき $a_j = b_j$ の場合

時刻 T_{n-1} に分割 $\vec{b}$ の中の同値な r 個の遺伝子を含む同値類クラスにおいて合祖が起き、そのクラスはサイズが $r-1$ のクラスとして a_{r-1} 個のクラスの１つとなることを意味する。そこで、結合確率 $P(\Omega_n^* = \vec{b}, \Omega_{n-1}^* = \vec{a}) = P(\Omega_n^* = \vec{b} | \Omega_{n-1}^* = \vec{a}) P(\Omega_{n-1}^* = \vec{a})$ を求めよう。

$P(B)=P$(時刻T_{n-1}に起きた事象が合祖である$\left|\Omega_{n-1}^{*}=\vec{a}\right.$)

分割$\vec{a}$のサイズ$r-1$のクラスを$C_1,C_2,\cdots,C_m;(m=a_{r-1})$とする。

$P(C_i)=P\left(\text{合祖によってクラス}C_i\text{を生じる}\middle|\text{時刻}T_{n-1}\text{で合祖、}\Omega_{n-1}^{*}=\vec{a}\right)$とすると

$P(\Omega_n^{*}=\vec{b}\left|\Omega_{n-1}^{*}=\vec{a}\right.)=\sum_{i=1}^{m}P(B)P(C_i)$ ただし$m=a_{r-1}$。ポアソン過程の性質より、

$P(B)=\dfrac{n(n-1)}{2}\Big/\dfrac{n(n-1+\theta)}{2}=\dfrac{n-1}{n-1+\vartheta}$。

合祖が起きるという条件の下で、時間を逆向きにして分割$\vec{a}$のa_{r-1}個のサイズ$r-1$のあるクラスC_iが合祖の結果生じる確率は$n-1$個体間で等確率なので、$P(C_i)=\dfrac{r-1}{n-1}$となる。

よって、$P(\Omega_n^{*}=\vec{b}\left|\Omega_{n-1}^{*}=\vec{a}\right.)=\sum_{i=1}^{m}P(B)P(C_i)=\dfrac{n-1}{n-1+\vartheta}\times\dfrac{r-1}{n-1}\times m=\dfrac{(r-1)a_{r-1}}{n-1+\vartheta}$。

（2） $a_1=b_1-1,\ a_j=b_j\ (j\le n-1)$の場合

時刻T_{n-1}に分割$\vec{b}$の中の１個の遺伝子を含む同値類クラスにおいて突然変異が起き、除かれることを意味する。従って

$P(\Omega_n^{*}=\vec{b}\left|\Omega_{n-1}^{*}=\vec{a}\right.)=P\left(\text{時刻}T_{n-1}\text{に突然変異}\middle|\Omega_{n-1}^{*}=\vec{a}\right)=\dfrac{\vartheta}{n-1+\vartheta}$

以上をまとめると
$$P(\Omega_n^{*}=\vec{b}\left|\Omega_{n-1}^{*}=\vec{a}\right.)=\begin{cases}\dfrac{(r-1)a_{r-1}}{n-1+\vartheta} & (1)\\ \dfrac{\vartheta}{n-1+\vartheta} & (2)\end{cases}\qquad(4.37)$$

これはポリアの壷様モデルの推移確率(4.34)と同じである。初期条件はΩ_1^{*},Π_1いずれも１個の状態から始まるので、２つのプロセスは確率的に同値である。

系図過程$\Re_t$のジャンプが起きる時刻を$T_i=\inf\{t:\Re_t=i\},T_n=0$とする。$T_{i-1}-T_i$は状態iへの滞在時間で指数分布$P(T_{i-1}-T_i>t)=\exp\left(-\dfrac{i(i-1+\vartheta)}{2}t\right)$に従う。

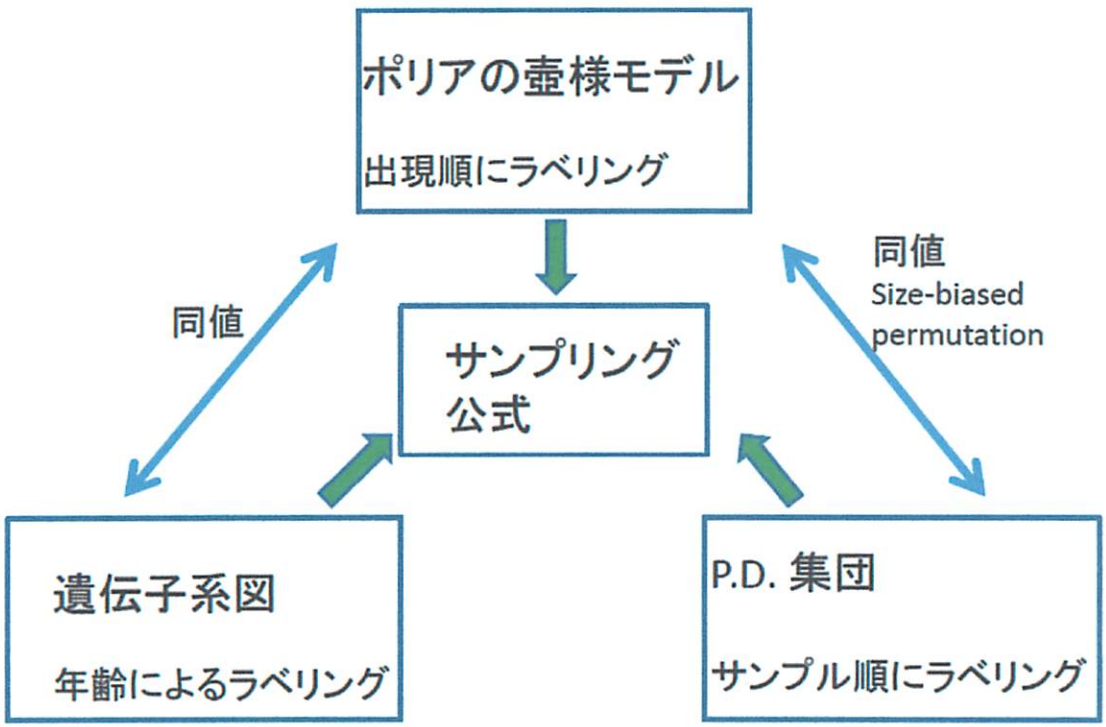
この時刻$T_n=0<T_{n-1}<T_{n-2}<\cdots<T_2<T_1<T_0$に突然変異または合祖が起こっている。

従って、逆進過程$\Omega_k^{*}(k=0,1,2,\cdots,n)$は突然変異によって新しい遺伝子タイプが生じ、合祖においては、時間を逆さに見ると、分枝によって同じタイプが追加されるプロセスである。逆進プロセスにおいて古い突然変異から順にラベルをつけ、合祖では同じラベルを与

え、ラベルを生じた順に$Y_1, Y_2, \cdots, Y_n$とすると、定理４．１７よりポリアの壺のラベルのプロセス$X_1, X_2, \cdots X_n$と同値であることが分かる。以上のことから、ポリアの壺様モデルにおいて現れるボールのラベリングは突然変異を含む合祖過程を時間について古い方から現れる順にラベリングしたものに一致する。すなわち、サンプル中に現れる順に遺伝子にラベリングすることは系図上で古い方から順にラベリングすることに確率論的に同値である。

系４．１８

年齢によるラベリングされたサンプル中のアレル数の分布はサンプル中に現れた順によるラベリングされたアレル数の分布と確率的に等しい。



以上をまとめると、上図のようになる。すなわち、ポリアの壺様モデルのサンプルのラベリング、ポアソン・ディリクレ(P.D.)集団からのサンプルの出現順のラベリングおよび遺伝子系図で古いアレルから年齢によるラベリング、この３者が確率論的に同値であるということである。これらの事実より、集団遺伝学における種々の結果について、この３つの視点から統一的な解釈を与えることができる。次の節で紹介しよう。

４．６．５　アレルの年齢

ポリアの壺様モデルからアレルの年齢に関する種々の公式を導くことができる。
$\{X_1, X_2, X_3, \cdots\}$をポアソン・ディリクレ集団$\underline{P}$の観測系列とし
$\{X_1, X_2, \cdots, X_n\}$　を観測サンプル、　$\{X_{n+1}, X_{n+2}, \cdots\}$で集団を現わす。

ε を n 個サンプルからランダムに取り出した一つのアレルすなわち $P(\varepsilon = X_j) = \frac{1}{n}$、

$S \equiv^{\#} \{1 \le j \le n; X_j = \varepsilon\}$ を ε と同じタイプのアレルの数とする。さらに、サンプル中で

i番目のサンプルと同じタイプのアレルの数を $S_i =^{\#} \{1 \le j \le n; X_j = X_i\}$ で表す。

1．(Watterson and Guess(1977))：

(a)集団内で頻度がYのアレルが集団内で最も古いアレルである確率はYである。

(b)もし或るアレルの頻度がYであるならば、それが標本中で最初にサンプルされる確率はYである。

(証明)頻度Yのアレルが最初にサンプルされる確率はYであり、それは同時に最も古いアレルである確率でもある。

2．(a)Watterson and Guess(1977), Kelly(1977)：サイズnのサンプル中に或るアレルがi個含まれているとき、そのアレルが集団中で最も古いアレルである確率は $\frac{i}{\vartheta + n}$。

(b) $P(X_{n+1} = \varepsilon | S = i) = \frac{i}{\vartheta + n}$

(証明) $P(X_{n+1} = X_j | S_j = i) = \frac{i}{\vartheta + n}$ より、

$$P(X_{n+1} = \varepsilon | S = i) = \frac{P(X_{n+1} = \varepsilon, S = i)}{P(S = i)} = \frac{1}{P(S = i)} \sum_{j=1}^{n} P(\varepsilon = X_j, X_{n+1} = X_j, S_j = i)$$

$$= \sum_{j=1}^{n} P(\varepsilon = X_j) \frac{P(X_{n+1} = X_j, S_j = i)}{P(S = i)} = \sum_{j=1}^{n} \frac{1}{n} P(X_{n+1} = X_j | S_j = i) \frac{P(S_j = i)}{P(S = i)} = \frac{i}{\vartheta + n}$$

3．(a)Watterson and Guess(1977), Kelly(1977)：サイズnのサンプル中に或るアレルがi個含まれているとき、それがサンプル中で最も古いアレルである確率は $\frac{i}{n}$ である。

(b) $P(X_1 = \varepsilon | S = i) = \frac{i}{n}$

(証明) (b)n個のサンプル中にあるアレルがi個含まれるとき、それが最初のサンプルである確率、すなわちサンプル中で最も古い確率は $\frac{i}{n}$ である。

4．(a)Kelly(1977)：サイズnのサンプル中に、最も古いアレルがi個含まれる確率は

$$\frac{\vartheta}{n}\binom{n}{i}\Big/\binom{\vartheta+n-1}{i}。$$

(b) $P(S_1=i)=\dfrac{\vartheta}{n}\binom{n}{i}\Big/\binom{\vartheta+n-1}{i}$。

（証明）ポリアの壺様モデルより、最初にサンプルするアレルがラベル１となる。そのアレルがその後のｎ－１個のサンプルに何個含まれるかは二色の球を含むポリアの壺モデルで表現される。ラベル１のアレル頻度はベータ分布 $Beta(1,\vartheta)$ に従い、ラベル１の頻度が x のとき、サンプル中に含まれるラベル１の個数は二項分布 $B(n-1,x)$ に従う。よって、$P(S_1=i)$ は次の式で表される。

$$P(S_1=i)=\int_0^1\binom{n-1}{i-1}x^{i-1}(1-x)^{n-i}\times\vartheta(1-x)^{\theta-1}dx=\vartheta\binom{n-1}{i-1}\frac{\Gamma(i)\Gamma(n+\vartheta-i)}{\Gamma(n+\vartheta)}$$
$$=\frac{\vartheta}{n}\binom{n}{i}\Big/\binom{\vartheta+n-1}{i}$$

５．(a) Saunders, Tavare and Watterson(1984)：サイズｎのサンプル中で最も古いアレルより古い集団中のアレルの数を N_n とすると

$$P(N_n=k)=\frac{n}{\vartheta+n}\left(\frac{\vartheta}{\vartheta+n}\right)^k,\ k=0,1,2,\cdots$$

(b) サンプル $\{X_1,X_2,\cdots,X_n\}$ の中の最も古いアレルよりも古いアレルが集団 $\{X_{n+1},X_{n+2},\cdots\}$ において見出されるアレルタイプの数を ν_n とすると

$$P(\nu_n=k)=\frac{n}{n+\vartheta}\left(\frac{\vartheta}{\vartheta+n}\right)^k,k=0,1,2,\cdots$$

（証明）事象 $\{\nu_n=k\}$ は集団で最も古いアレル、二番目に古いアレル、...、ｋ番目に古いアレルがサンプル $\{X_1,X_2,\cdots,X_n\}$ の中に現れず、k+1 番目に古いアレルはサンプル中に現れる事象である。また最も古いアレルが ｎ 個のサンプル中に現れない確率はポアソン・ディリクレ分布の size-biased permutation より

$$E[(1-Z_1)^n]=\int_0^1(1-z)^n\vartheta(1-z)^{\vartheta-1}dz=\vartheta\int_0^1(1-z)^{n+\vartheta-1}dz=\frac{\vartheta}{n+\vartheta}$$

以下同様にして、$Z_1,Z_2,Z_3,\cdots$ はベータ分布 $Beta(1,\theta)$ に従う独立な確率変数なので

$$P(\nu_n=k)=E[(1-Z_1)^n(1-Z_2)^n\cdots(1-Z_k)^n\{1-(1-Z_{k+1})^n\}]$$
$$=\prod_{i=1}^{k}E[(1-Z_i)^n]\times E[\{1-(1-Z_{k+1})^n\}]=\left(\frac{\vartheta}{n+\vartheta}\right)^k\frac{n}{n+\vartheta}$$

6．Watterson(1974)：サイズnのサンプル中にi個含まれるアレルの種類数をa_iとする

と$E[a_i]=\frac{\vartheta}{i}\binom{n}{i}\Big/\binom{\vartheta+n-1}{i}$

（証明）$P(S_1=i|a_1,\cdots,a_n)=\frac{ia_i}{n}$、より、$(a_1,\cdots,a_n)$の全てについて期待値を取ると4.

の計算より　$E[a_i]=\frac{n}{i}P(S_1=i)=\frac{n}{i}\times\frac{\vartheta}{n}\binom{n}{i}\Big/\binom{\vartheta+n-1}{i}=\frac{\vartheta}{i}\binom{n}{i}\Big/\binom{\vartheta+n-1}{i}$。

これは、サンプル中で最も古いアレルがサンプル中にi個含まれる確率$P(S_1=i)$が

$P(S_1=i)=\frac{i}{n}E[a_i]$であることを意味する。

7．Ewens(1972)：集団からランダムに取り出した二つの遺伝子が同じアレルタイプである確率は$E[\sum P_i^2]=\frac{1}{1+\vartheta}$。

（証明）ポリアの壺様モデルで2番目の球が1番目の球である確率$P(X_1=X_2)=\frac{1}{1+\vartheta}$。

8．Watterson and Guess(1977)：最大の頻度を$P_{(1)}=\underset{i}{Max}P_i$とする、このとき

$0.5\le x\le 1$に対して、$P_{(1)}$の分布密度は$\vartheta x^{-1}(1-x)^{\vartheta-1}$である。

（証明）λ_jを定理4．15で導入した確率変数とする、

$$P(P_{(1)}>x)=\sum_{j=1}^{\infty}P(P_{(1)}=P_j,\ P_j>x)=\sum_{j=1}^{\infty}P(\gamma_j>x\ \ and\ P_{(1)}=\gamma_j)$$

$$=\sum_{j=1}^{\infty}P(\gamma_j>x)=\sum_{j=1}^{\infty}P(\gamma_j>x|\lambda_j=1)P(\lambda_j=1)$$

定理4．16（3）より$\lambda_j=1$のとき、γ_jは$Beta(1,\vartheta+j-1)$に従うので、

$P(\lambda_j=1)=\frac{\vartheta}{\vartheta+j-1}$より

$$P(P_{(1)}>x)=\sum_{j=1}^{\infty}\int_x^1\frac{\Gamma(\vartheta+j)}{\Gamma(\vartheta+j-1)}(1-t)^{\vartheta+j-2}dt\times\frac{\vartheta}{\vartheta+j-1}$$

$$=\vartheta\int_x^1\sum_{j=1}^{\infty}(1-t)^{\vartheta+j-2}dt=\vartheta\int_x^1 t^{-1}(1-t)^{\vartheta-1}dt$$

9．Ewens(1972)：サイズnのサンプル中での異なるアレルタイプ数をK_nとする。

$P(K_n=k)=\dfrac{\vartheta^k S(n,k)}{\vartheta_{(n)}}$，ただし$S(n,k)$は第1種スターリング数（付録(C.1)参照)。

（証明）$K_n=\lambda_1+\lambda_2+\cdots+\lambda_n$なので、$K_n$の母関数は$\lambda_1,\lambda_2,\cdots\lambda_n$が独立より

$$E\left[s^{K_n}\right]=E\left[s^{\lambda_1+\lambda_2+\cdots+\lambda_n}\right]=\prod_{i=1}^{n}E[s^{\lambda_i}]=\prod\left\{1\times\frac{i-1}{\vartheta+i-1}+s\times\frac{\vartheta}{\vartheta+i-1}\right\}=\prod_{i=1}^{n}\left(\frac{i-1+\vartheta s}{i+\vartheta-1}\right)$$

s^kの係数に注意すると第1種スターリング数及び$P(K_n=k)$が得られる。

10．Ewens(1972)：ポアソン・ディリクレ集団$\underline{P}$の頻度スペクトルを$\phi(x)$とすると
$\phi(x)=\vartheta x^{-1}(1-x)^{\vartheta-1}\quad(1<x<1)$。

（証明）　$f(x)$を[0, 1]上の有界可測関数とする。$Q_1=P_1^S$をポアソン・ディリクレ$\underline{P}$の size-biased Permutationの最初の頻度とすると、$\sum_i f(P_i)P_i=E\left[f(Q_1)\middle|\underline{P}\right]$。　Q_1は
$Beta(1,\theta)$に従うので、上式の期待値を取ると

$$E\left[\sum_i f(P_i)P_i\right]=E[f(Q_1)]=\int_0^1 f(x)\vartheta(1-x)^{\vartheta-1}dx。$$

ここで$f(x)=\begin{cases}1/x & (x>t\text{のとき})\\ 0 & (x\le t\text{のとき})\end{cases}$とすると

$$E\left[\sum_i f(P_i)P_i\right]=E\left[\sum_i 1(P_i>t)\right]=\int_t^1\vartheta x^{-1}(1-x)^{\vartheta-1}dx、$$

ただし$1(P_i>t)=\begin{cases}1 & (P_i>t\text{のとき})\\ 0 & (P_i\le t\text{のとき})\end{cases}$とする。これより頻度スペクトル$\phi(x)$は

$\phi(x)=\vartheta x^{-1}(1-x)^{\vartheta-1}\quad(1<x<1)$。　また$g(x)=xf(x)$とすると

$$E\left[\sum_i g(P_i)\right]=E\left[\sum_i\frac{g(P_i)}{P_i}P_i\right]=E\left[\frac{g(Q_1)}{Q_1}\right]=\int_0^1\frac{g(x)}{x}\vartheta(1-x)^{\vartheta-1}dx=\int_0^1 g(x)\phi(x)dx。$$