

遺伝子解析とポアソン過程

京大・数理研

山崎泰郎

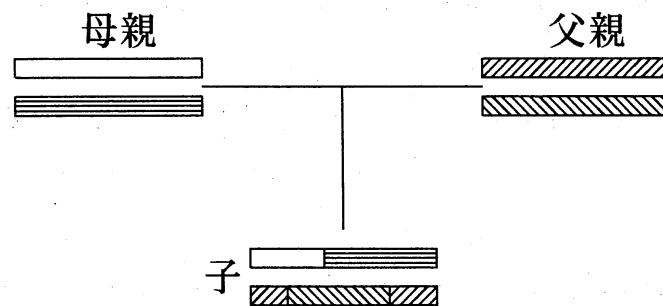
京都産大・理

山崎京子

本稿は山崎京子の提起した問題に、両者で確率論的考察を加えたものである。数学の理論とその適用される実際の現場とで、問題意識も大いにずれており、一方で明らかと思われることが他方ではそうとみなされず議論の対象となったり、或いは逆に問題のある点が無視されて全く考察の対象とされていないなかったり、と言うことはよくある現象である。このようなことは数学と物理学の間でしばしば見られることであるが、数学が広く諸科学に適用されるにつれて、より離れた分科との間により大きなギャップをもって現れるであろう。別の学問体系であるからそのようなことがあっても不都合はない、と言う考え方もあり得るだろうが、学問が個別化、専門化して共通の理念を失いつつあることの表れであると思えば、放置できないことである。今回は遺伝学から題材を取って、一つの考察を試みる。誤解を避けるために一言すれば、筆者は遺伝学の研究には十分な敬意を持っており、数学的に欠陥があって数学の方が優位に立っているなどと主張するつもりは決してない。

今回取り上げるのは、E. Feingold 氏の論文 “ Markov processes for modeling and analyzing a new genetic mapping method ” (文献 [1]) であり、ここに現れている数学的内容は、もっと簡単かつ正確に得られるものであることを注意する。

まず数学以前の状況の説明から始める。人間の遺伝情報は、23 対ある染色体 (これは長い帯状のものである) のそれぞれの上に分布している遺伝子によって伝えられる。生殖に際して、染色体の各対において減数分裂を通じて、父親からの一本と母親からの一本が子の染色体として引き継がれる。ただし単純に二本のうちの一本と言うわけではなく、二本の染色体のつぎはぎされたもの (cross over という) が子に写される。たとえば次の図の通り。



この図の例では、母親からの染色体には一回 cross over が、父親からの染色体には二回 cross over が起きている。

したがって、子の染色体（の一方）の遺伝子は必ず母親の染色体の遺伝子から来ている。しかし孫には子の染色体がさらに cross over を起こして伝えられるので、孫の染色体（の一方）には祖父から来た遺伝子と祖母から来た遺伝子のつぎはぎされたものが伝えられている。

今、祖母と孫が同じ病気にかかったとして、これが遺伝病ではないかと疑ったとする。そして遺伝病だったとした場合、その責任遺伝子を同定したい。そのような祖母と孫の pair を例えば 100 組取り、孫の遺伝子と祖母の遺伝子と同じかどうか、ある染色体のある遺伝子について 100 組のうちの何組が祖母と孫とで合致しているかを調べる。もし合致している割合が異常に高い遺伝子があれば、それが責任遺伝子ではないかと疑われるわけである。

ここで、あらかじめ責任遺伝子の予想がついている場合と、何の予想もなくいわば当てずっぽうに探している場合とで検定の仕方が違う。前者の場合は、ある特定の遺伝子が責任遺伝子であるかどうかを検定することになり、問題としては比較的単純であり、本稿の考察の対象とはしない。後者の場合は、23 対の染色体のどこかに責任遺伝子がないかを検定することになる。すなわち、「合致している割合が異常に高い」と言えるかどうかを考察するわけである。言い換えれば、全く at random に cross over が起きているとして、たくさんある遺伝子のどこかでは合致する割合が、偶然高くなる可能性があるのではないか、あるモデルのもとに「偶然でない」と言い切れるためにはどうしたらよいか、ということである。

そのモデルとしては、次のように考えるのが最も自然であろう。まず、母から子に伝えられる染色体は、二本のうちどちらであるかは全く等確率であること、次に cross over は染色体の位置によらずに、どこでも同じ確率で起こること、また一回 cross over が起こったとして、二回目に cross over が起こるかどうか、どこで起こるかは一回目と確率的に独立であること、である。このモデルに従って、一番簡単な祖母と孫の場合について、上記の検定の

問題を考えてみよう。

サンプルの数を n とする。(先の例では $n = 100$)。また染色体の長さを l 、単位長さ当りの cross over の起こる確率を λ とする。

cross over の起こる回数の期待値は $l\lambda$ であり、 n 個のサンプル全体を通じて cross over の回数の期待値は $nl\lambda$ となる。ところで染色体上の位置は、染色体の一方の端からの距離 x によって表すことができる。この x は実数であり、同じく実数である時間と同じようにみなすことにより、ポアソン過程に帰着させることができる。すなわち x を時間とみなすことにより、前述のモデルは「cross over がいつ起こるかは時刻によらずにいつでも同じ確率で起こり、また一回 cross over が起こったとして、二回目に cross over が起こるかどう、いつ起こるかは一回目と確率的に独立である」と言う風に述べられる。このモデルの従う確率過程がポアソン過程であり、 $0 \leq x \leq l$ における cross over の回数はポアソン分布に従う。回数の期待値が $nl\lambda$ であるから、cross over が k 回起こる確率はポアソン分布：

$$(1) \quad p_k = e^{-nl\lambda} \frac{(nl\lambda)^k}{k!}$$

に従う。

そして cross over が起こるたびに、祖母と孫の遺伝子の合致の個数は一個ふえ、または一個減る。問題は $0 < b \leq n$ をみたす b を一つ fix し、合致の個数がどこかで b 組以上になる確率を計算することである。もしこの確率が例えば $\frac{1}{100}$ であると、実際にあるところで b 組以上の合致が観測されれば、これは「偶然とは考えられない」と主張できるであろう。(危険率 $\frac{1}{100}$ の検定)。

今の問題は、cross over が起こるかどうかがポアソン過程と言う確率論的な現象である上に、cross over が起こったときに遺伝子の合致の個数がどう変わるかを解析せねばならない。これは一見複雑なことのように入る。そこで Feingold 氏はモンテ・カルロ法により、実際に乱数を発生させて計算機上で 10 万回の試行を行ない、その結果として上記の確率を求めると言う方法を取った。

一方 Aldous (文献 [2]) は、上記の確率を求める簡単な近似式を与えている。これは今の祖母と孫の場合には

$$(2) \quad \alpha = 1 - \exp \left\{ -l \binom{n}{b} \frac{1}{2^n} (2b - n) \lambda \right\}$$

と書ける。(この近似式の導出については本稿では省略する)。

しかしこれは近似式であるから、Feingold はもっと正確な結果を得ようとして、上記のように 10 万回の simulation を行ない、その結果を Aldous の近似式と比較している。ところが後述のように、上記の確率 α の計算は簡単に厳密に（すなわち一切の近似なしで）できるのであり、数値的にもポケコンかパソコン程度で計算できるのである。しかもこの厳密解は Aldous 近似式とよく一致していて、むしろ Feingold の simulation の結果の方が誤差が大きい。言い換えれば、Feingold はより大規模な大変な計算をして精度の悪い結果を得たのであり、それを Aldous の近似式と比較しても全く意味をなさないことになる。以下厳密解の計算について説明する。

条件付き確率の概念を用いると、問題は次のように単純化される。cross over が k 回起こったとして、どこかで合致の個数が b 以上になる条件付き確率を α_k とすると、求める確率 α は

$$(3) \quad \alpha = \sum_{k=0}^{\infty} \alpha_k p_k$$

によって定まる。そこで α_k を定める式を見つけさえすればよいが、ここで問題は完全に discrete なものとなり、単に α_0 および数列 $\{\alpha_k\}$ のみたす漸化式を見出すことに帰着される。

cross over が k 回起きたとしたときの、どこでも合致の個数が b 組以上にならず、かつ最後の k 回目の cross over の結果合致の個数が i 組 ($0 \leq i < b$) となっている条件付き確率を $a_{k,i}$ とする。明らかに

$$(4) \quad \forall k, \sum_{i=0}^{b-1} a_{k,i} + \alpha_k = 1$$

が成り立つ。

$k = 0$ すなわち cross over が全く起こっていないときは、二本の染色体（祖母から来た染色体と祖父から来た染色体）のどちらが選ばれるかは等確率だから、祖母の遺伝子と孫の遺伝子の合致の個数は確率 $\frac{1}{2}$ の二項分布に従う。

従って

$$(5) \quad \begin{cases} a_{0,i} = \frac{1}{2^n} \binom{n}{i} & 0 \leq i < b \\ \alpha_0 = 1 - \sum_{i=0}^{b-1} a_{0,i} \end{cases}$$

となる。

次に $\{\alpha_k\}$ の漸化式を作る。 k 回 cross over が起こっているところに $k+1$ 回目の cross over が起こると、 n 組のうち 1 組について合致、非合致が逆になり、従って合致の個数は一つ増えるか一つ減る。一つ増えるのは、今まで非合致だった場合であり、その確率は (n 組のうちのどの一組であるかは等確率だから) $\frac{n-i}{n}$ であり、一つ減るのは今まで合致だった場合だから、その確率は $\frac{i}{n}$ である。従って次の漸化式が生ずる。

$$(6) \quad \begin{cases} a_{k+1,0} = \frac{1}{n} a_{k,1} \\ a_{k+1,i} = \frac{n-(i-1)}{n} a_{k,i-1} + \frac{i+1}{n} a_{k,i+1} \quad (1 \leq i < b) \\ \alpha_{k+1} = \alpha_k + \frac{n-(b-1)}{n} a_{k,b-1} \end{cases}$$

ただし $a_{k,b} = 0$ とみなす。

以上初期値 (5) と漸化式 (6) により、 $\{\alpha_k\}$ は帰納的に数値計算され、これを代入して (3) により α が計算される。(3) 式は無限級数であるが、有効数字が何桁欲しいか指定すると、何項目までの和を計算すればよいか、残余項の和の大きさ (誤差) を評価するのは容易である。従ってここに説明した計算方法により、任意の精度で求める確率 α を exact に計算することができる。またその計算は小さなパソコンでできる程度の規模のものである。

これまで祖母と孫の pair について議論して来たことは、他の親族の pair についても多少の修正を施せば通用する。以下で Feingold の取り上げた四つの場合について考察する。

(1) 半兄弟の場合。すなわち母親は同じで父親が異なる兄弟の場合である。このとき共通の母親から来る染色体は各一本であり、その上の遺伝子の合致、非合致を調べることになる。このときは状況は祖母と孫の pair の場合と同じであるが、祖母と孫の場合は祖母の遺伝子は動かなくて孫の遺伝子がそれと合致しているかどうかを調べた。今度の場合は、半兄弟のそれぞれについて cross over が起こり、どちらに cross over が起こっても合致、非合致の個数は変る。そこで cross over の起こる確率 λ を 2λ に置き換えることだけで、見かけ上祖母と孫の場合の結果と同じ結果が生ずる。すなわち見かけ上 λ の値を二倍にすればよいだけのことである。

(2) 兄弟の場合。この場合、各一本の染色体については母親からきた遺伝子の合致、非合致が問題になり、残り一本の染色体については父親から来た遺伝子の合致、非合致が問題になる。ここで、どちらかの染色体で合致が起こっている個数を調べることにする。(実験上、これだけが観測可能だそうである)。

$k = 0$ すなわち cross over が全く起こっていないとき、父親から来た染色体が合致している確率は $\frac{1}{2}$ 、母親から来た染色体が合致している確率も $\frac{1}{2}$ 、従ってどちらかが合致している確率は $\frac{3}{4}$ である。よって (5) 式は確率 $\frac{3}{4}$ の二項分布に置き換える。

$$(5)' \quad \begin{cases} a_{0,i} = \binom{n}{i} \left(\frac{3}{4}\right)^i \left(\frac{1}{4}\right)^{n-i} & 0 \leq i < b \\ \alpha_0 = 1 - \sum_{i=0}^{b-1} a_{0,i} \end{cases}$$

もう少し詳しく分類して、両方の染色体について遺伝子が合致しているのが i_2 組、片方の染色体について遺伝子が合致しているのが i_1 組、となる確率 a_{0,i_1,i_2} は三項分布に従い

$$(7) \quad a_{0,i_1,i_2} = \frac{n!}{i_1!i_2!(n-i_1-i_2)!} \left(\frac{1}{2}\right)^{i_1} \left(\frac{1}{4}\right)^{i_2} \left(\frac{1}{4}\right)^{n-i_1-i_2}$$

となる。そして $a_{0,i} = \sum_{i_1+i_2=i} a_{0,i_1,i_2}$ である。また $i_1+i_2=i$ としたときの i_1 の分布は、確率 $\frac{2}{3}$ の二項分布に従い、従って i_1 の期待値は $\frac{2}{3}i$ である。

次に漸化式について考える。(6) 式と同様にして a_{k,i_1,i_2} について次の漸化式が成り立つ。
 $0 \leq i_1 + i_2 < b$ として

$$(8) \quad \begin{aligned} a_{k+1,i_1,i_2} = & \frac{n-(i_1+i_2-1)}{n} a_{k,i_1-1,i_2} + \frac{i_1+1}{2n} a_{k,i_1+1,i_2} \\ & + \frac{i_1+1}{2n} a_{k,i_1+1,i_2-1} + \frac{i_2+1}{n} a_{k,i_1-1,i_2+1} \end{aligned}$$

(第二項、第三項の分母が $2n$ になっているのは、片方の染色体のみ合致している場合、合致している方の染色体に cross over が起こるか、合致していない方の染色体に cross over が起こるかは等確率で $\frac{1}{2}$ だから)。ただし $a_{k,-1,i_2} = a_{k,i_1,-1} = a_{k,b-i_2,i_2} = 0$ と考える。そして $a_{k,i} = \sum_{i_1+i_2=i} a_{k,i_1,i_2}$ であり、

$$(9) \quad \alpha_{k+1} = \alpha_k + \frac{n-(b-1)}{n} a_{k,b-1}$$

となる。従って漸化式は前より複雑になるが、この場合もパソコンにより帰納的に α_k は数値計算され、これより $\alpha = \sum_{k=0}^{\infty} \alpha_k p_k$ も数値計算される。

なお、cross over は二人の兄弟それぞれの二本の染色体のどれで起こってもよいので、一本の染色体について単位長さ当たり cross over の起こる確率を λ とすると、見かけ上 cross

over が 4λ の確率で起こるように考えて数値計算すればよい。すなわち λ の値を見かけ上四倍して数値計算することになる。

(3) 伯母と姪の場合（性別無関係に兄弟の子との比較）。

伯母と母が姉妹とする。姪の染色体のうち、母の方から来た一本だけが問題になる。

姪の遺伝子と伯母の遺伝子が合致するのは、姪の遺伝子が祖父方から来ていてかつ伯母と母の祖父方の遺伝子が合致している場合、および姪の遺伝子が祖母方から来ていてかつ伯母と母の祖母方の遺伝子が合致している場合である。

姪の遺伝子が祖父方から来ているか祖母方から来ているかは、祖母-孫の場合と同じで、 $k = 0$ のとき確率 $\frac{1}{2}$ の二項分布、そして後は確率 λ で cross over を起こす。伯母と母の祖父方の遺伝子が合致しているかどうか（祖母方の遺伝子についても同じ）は (1) の半兄弟の場合と同じで、 $k = 0$ のとき確率 $\frac{1}{2}$ の二項分布、そして後は確率 2λ で cross over を起こす。そしてこの三種類の cross over は互いに独立である。

さて姪の遺伝子が祖父方から来ているときは、伯母と母の祖父方の染色体の cross over が起これば合致の数は一つ減りまたは増え、伯母と母の祖母方の染色体の cross over が起こっても合致の数は変らない。姪の遺伝子が祖母方から来ている場合は、上のことを逆に考えて考えればよい。

次に姪の（母方の）染色体の cross over を起こして遺伝子が祖父方から祖母方に変った場合（またはその逆の場合）、cross over 後、伯母の祖母方の遺伝子と合致するかどうかは確率 $\frac{1}{2}$ である。すなわち確率 $\frac{1}{2}$ で合致の個数が一つ減りまたは増え、確率 $\frac{1}{2}$ で合致の個数は変わらない。

いずれにしても合致の個数が一つ減りまたは増える確率は単位長さ当たり $2\lambda + \frac{\lambda}{2} = \frac{5}{2}\lambda$ である。 $k = 0$ のとき合致の個数は確率 $\frac{1}{2}$ の二項分布に従う。従ってこれらを総合して、伯母と姪の場合は、祖母と孫の場合で単に λ を見かけ上 $\frac{5}{2}\lambda$ に置き換えればよい。（すなわち λ の値を $\frac{5}{2}$ 倍するだけで祖母-孫の場合の数値計算がそのまま通用する）。

(4) いとこの場合。双方の母親同士が姉妹とする。母親同士について祖父方の染色体（祖母方の染色体についても同じ）の合致、非合致は (1) の半兄弟の場合と同じ。いとこ同士の母方の染色体について、祖父方から来ているか祖母方から来ているかの合致、非合致は、祖母-孫の場合において λ を見かけ上 2λ に置き換えればよい。

いとこの母方の染色体の遺伝子が合致するのは、ともに祖父方の遺伝子でかつ母親同士の祖父方の遺伝子が合致しているか、またはともに祖母方の遺伝子でかつ母親同士の祖母方

の遺伝子が合致する場合である。

$k = 0$ のときいとこの遺伝子が合致する確率は $\frac{1}{4} \times \frac{1}{2} \times 2 = \frac{1}{4}$ だから、 $a_{0,i}$ は確率 $\frac{1}{4}$ の二項分布に従う。

もう少し細かく分類して、いとこ同士の遺伝子がともに祖父方またはともに祖母方であるが合致はしていないものが j 組あったとする。結合分布 $a_{0,i,j}$ は確率 $\frac{1}{4}, \frac{1}{4}$ の（従って残りの確率 $\frac{1}{2}$ の）三項分布に従う。すなわち

$$(10) \quad a_{0,i,j} = \frac{n!}{i!j!(n-i-j)!} \left(\frac{1}{4}\right)^i \left(\frac{1}{4}\right)^j \left(\frac{1}{2}\right)^{n-i-j}$$

である。もちろん $a_{0,i} = \sum_{j=0}^{n-i} a_{0,i,j}$ である。

次に漸化式を求めよう。いとこ同士の遺伝子がともに祖父方であったとする。母親同士の祖父方の染色体に cross over が起これば、合致の個数は一つ減りまたは増えるが、祖母方の染色体に cross over が起こっても合致の個数は変らない。いとこ同士の遺伝子がともに祖母方でも同様である。

いとこ同士の染色体の祖父方、祖母方が食い違っているとき、片方のいとこの染色体に cross over が起こって、例えばともに祖父方にそろったとする。このとき cross over 後に遺伝子が合致するかどうかの確率は $\frac{1}{2}$ である。従ってこの cross over によって確率 $\frac{1}{2}$ で i が一つ増え、確率 $\frac{1}{2}$ で j が一つ増える。また逆に cross over の結果、祖父方、祖母方がそろっていたのが食い違うことになった場合は、 i が一つ減りまたは j が一つ減る。

以上 cross over は祖父方（または祖母方）の染色体、祖父方、祖母方の入れ替わり、の双方がそれぞれ単位長さ当たり 2λ の確率で起こる。よってどちらかの cross over が起こる確率は 4λ であり、それがどちらであるかの確率は各 $\frac{1}{2}$ である。

従って、 $\frac{i}{2n}$ の確率で i は一つ減って j が一つ増え、 $\frac{j}{2n}$ の確率で j が一つ減って i が一つ増え、 $\frac{n-(i+j)}{4n}$ の確率で i が一つ増え、同じ確率で j が一つ増え、 $\frac{i}{2n}$ の確率で i が一つ減り、 $\frac{j}{2n}$ の確率で j が一つ減る。残りすなわち $\frac{n-(i+j)}{2n}$ の確率で、 i, j は変らない。従って次の漸化式が成り立つ。

$$(11) \quad \begin{aligned} a_{k+1,i,j} = & \frac{i+1}{2n} a_{k,i+1,j-1} + \frac{j+1}{2n} a_{k,i-1,j+1} \\ & + \frac{n-(i+j-1)}{4n} a_{k,i-1,j} + \frac{n-(i+j-1)}{4n} a_{k,i,j-1} \end{aligned}$$

$$+\frac{i+1}{2n}a_{k,i+1,j}+\frac{j+1}{2n}a_{k,i,j+1}+\frac{n-(i+j)}{2n}a_{k,i,j}$$

従って、単位長さ当たり 4λ の確率で起こるポアソン過程を考え、一回ごとに (11) で定まる漸化式によって $a_{k,i,j}$ が変化することとなる。もちろん $a_{k,i} = \sum_{j=0}^{n-i} a_{k,i,j}$ であり、 $\alpha_k = 1 - \sum_{i=0}^{b-1} a_{k,i}$ となる。((11) の右辺において、 $a_{k,i,j}$ は $0 \leq i < b, 0 \leq i+j \leq n$ をみたさないものは 0 と考える)。

(11) から α_k についての漸化式として

$$(12) \quad \alpha_{k+1} = \alpha_k + \frac{n-(b-1)}{4n}a_{k,b-1} + \sum_{j=0}^{n-b+1} \frac{j}{4n}a_{k,b-1,j}$$

も得ることができる。いずれにしても、これらの漸化式により、いとこの場合も $\alpha = \sum_{k=0}^{\infty} \alpha_k p_k$ を数値計算することができる。

以上の場合を通じて、単純な祖母-孫の場合に比べて二つの理由で複雑になっている。一つは適切な漸化式を作る際、合致の個数 i だけでなく、補助的な変数（兄弟の場合二重に合致している個数 i_2 、いとこの場合祖父方か祖母方かは合致している個数 j 、など）を導入して、この補助変数を用いて初めて漸化式が記述できることである。対応する数値計算には、二つ（またはそれ以上）の変数の組の結合分布 $a_{k,i_1,i_2,\dots}$ のメモリーが必要であり、補助変数の個数が多くなればメモリ不足になったり計算時間が長くなったりして、小型のコンピュータでは計算しにくくなる心配がある。上で取り上げた場合には、補助変数の個数はせいぜい一個であり、何ら問題にはならないが、問題を無制限に複雑化、一般化することには、上述のような制約がある。

もう一つの問題の複雑化の要因は、合致、非合致にいくつもの染色体の cross over が関与し、従って何段階かの cross over の総合結果として合致、非合致が定まることである。これは何種類ものポアソン過程が同時に考えられ、合致、非合致の変化はどのポアソン分布からも起因し得ると言うことである。このことは一見複雑なように思えるが、実は以下に説明するように、「一つのポアソン過程」の場合に容易に帰着できるのである。従って、関係するポアソン過程の複数化は、数値計算の大規模化をもたらさず、何の障害にもならない。なお、前述の (1)~(4) で取り上げた場合では、何種類かのポアソン過程についてパラメータ λ の値は簡単な整数比であったが、以下の議論でこのことは全く必要でないことが注意される。

簡単のため、補助変数を用いずに合致の個数 i だけを用いて漸化式が書けているとする。 i の変化には n 個の異なる独立なポアソン過程が関与し、それぞれの単位時間（長さ）当たり起こる確率を $\lambda_j (1 \leq j \leq n)$ とする。また j 番目のポアソン過程に対し、これが起こったとき i が一つ増える確率を $r_j(i)$, i が一つ減る確率を $s_j(i)$, i が変化しない確率を $t_j(i)$ とする。 $\forall i, j, r_j(i) + s_j(i) + t_j(i) = 1$ と考える。

いま $\lambda \geq \sum_{j=1}^n \lambda_j$ をみたす λ を任意に一つ fix する。そして単位時間（長さ）当たり λ の確率で起こる仮想的なポアソン過程を考える。この仮想的ポアソン過程が起こったとき、確率 λ_j/λ で j 番目のポアソン過程と同じような i の変化が生じ、確率 $1 - \sum_{j=1}^n \lambda_j/\lambda$ で i には何の変化も起こらないとする。結果として生ずる i の変化は、 n 個のポアソン過程全体によって生ずる i の変化と同じである。従ってこれら n 個のポアソン過程の結果は、ただ一つの仮想的ポアソン過程の結果に（漸化式を書き換えることによって）帰着されることになる。

仮想的ポアソン過程に伴う漸化式は次のようになる。

$$(13) \quad a_{k+1,i} = r(i-1)a_{k,i-1} + s(i+1)a_{k,i+1} + t(i)a_{k,i} \quad (0 \leq i < b)$$

$$r(i) = \sum_{j=1}^n r_j(i) \frac{\lambda_j}{\lambda}, s(i) = \sum_{j=1}^n s_j(i) \frac{\lambda_j}{\lambda}, t(i) = 1 - \sum_{j=1}^n \frac{\lambda_j}{\lambda} + \sum_{j=1}^n t_j(i) \frac{\lambda_j}{\lambda}$$

この議論により、関係するポアソン過程が複数あることから来る困難は解消し、一つの単独なポアソン過程の場合の計算だけを考えればよいことになる。

最後に数値計算の結果をあげる。 $n = 100, l = 100, \lambda = 0.01$ として、祖母-孫、兄弟、伯母-姪、いとこの各場合について、どこかの遺伝子で合致の個数 $\geq b$ となる確率 α を数値計算して数表にまとめる。祖母-孫の場合のみ $n = 20$ の場合の結果もあげておく。これらは Feingold の元の論文で計算されたのと同じ場合であり、比較する目的で同じ場合について exact な計算を行った。ただし、Feingold は 5% の帰無検定を念頭に置き $\alpha_{23} \approx 0.05$ となるような b についてだけ計算をしているが、我々は $b \geq (i \text{ の期待値}) = (\text{祖母-孫、伯母-姪のとき } 50, \text{ 兄弟のとき } 75, \text{ いとこのとき } 25)$ のすべての b について計算した。（同じプログラムで b の値を変えるだけであり、 b を変えることは何の負担にもならない）。また、 α は一対の染色体についての値であり、23 対の染色体のすべてを通じてどこかで合致の個数が b 以上となる確率を α_{23} とすると

$$(14) \quad \alpha_{23} = 1 - (1 - \alpha)^{23}$$

により計算される。実際の検定には、 α_{23} の値の方が意味のあるものである。

我々の exact な数値計算と、Feingold の simulation および Aldous 近似の結果を合わせて数表にした。1 chromosome と書いてある欄が α の値、23 chromosomes と書いてある欄が α_{23} の値である。また数表における λ の値は、一本の染色体について単位長さ当たりの cross over の確率ではなく、問題の pair について合致、非合致の変化を起こさせるような幾つかの cross over の起こる全体としての確率である。従って一本の染色体について cross over の確率を 0.01 とするとき、上の意味の λ の値は祖母-孫のとき 0.01, 兄弟のとき 0.04, 伯母-姪のとき 0.025, いとこのとき 0.04 となる。

祖母-孫の場合以外の Aldous 近似の式は次のようである。(祖母-孫のときは (2) 式)。

兄弟のとき

$$(15) \quad \alpha = 1 - \exp \left\{ -l \binom{n}{b} \left(\frac{3}{4} \right)^b \left(\frac{1}{4} \right)^{n-b} \left(\frac{4}{3}b - n \right) \lambda \right\}$$

伯母-姪のときは (2) と同じ。

いとこのとき

$$(15)' \quad \alpha = 1 - \exp \left\{ -l \binom{n}{b} \left(\frac{1}{4} \right)^b \left(\frac{3}{4} \right)^{n-b} \left(\frac{4}{3}b - \frac{1}{3}n \right) \lambda \right\}$$

ただし λ は上記の意味の見かけの値（例えばいとこの場合 $\lambda = 0.04$ ）である。

数値計算の結果を見ると、Aldous 近似の結果は exact な計算結果とよく合っている。特に漸化式に補助変数の入って来ない祖母-孫の場合は有効数字の最後まで合っている。これと比べて simulation (10 万回の乱数実験) による結果の、exact な計算結果からのずれは同程度またはそれ以上である。Feingold の計算の意図が、simulation によって Aldous 近似の精度を問おうと言うことであつたとすれば、この計算は全く無意味であつたことになる。精度のより悪いものと比較しても何の結論も出て来ないし、計算自体の規模もかなり大掛かりである。我々の注意したように、exact な数値計算が小型のパソコンで十分実行可能であり、それで Aldous 近似が非常によい近似であることが完全に確かめられる。

最後に 5% の帰無検定の結果について一言する。 $\alpha_{23} \leq 0.05$ となるような b の最小値は $n = 100$ のとき、祖母-孫について 69, 兄弟について 92, 伯母-姪のとき 70, いとこのとき 44 である。 $n = 20$ とすると、祖母-孫について $b \geq 19$ (すなわち $b = 19$ または 20) である。

$n = 100$ の場合はさておき、 $n = 20$ のときは、19 組の祖母-孫の pair について (23 対の染色体上のどこかの遺伝子で) 合致が起こって初めて、5% の危険度をもって「それは偶

然の結果でない」と言い切れる。言い換えれば合致が18組以下ならば、とくに責任遺伝子がなくとも偶然に合致する可能性があるということである。

帰無検定の場合、サンプルの数を大きくすれば問題無いわけだが、実際にはしばしばそうも行かない。例えば珍しい病気であれば、入手できるあらゆる症例を集めても、サンプルはせいぜい数十程度にしかないだろう。従って現実には、少ないサンプルから何らかの有意の結論を得ることが必要なわけである。そこで20組のうち15組も合致が起こっていれば、その統計的意味付けなどを考慮することもなく「責任遺伝子の存在」を主張したくなるのではなかろうか。このような統計的裏付けの乏しい「研究成果」にまどわされることはないだろうか。

本稿では数学の立場から見て、「遺伝統計」で議論されていることは実は殆ど自明なことではないかと言う例をあげ、分野によって理解の種類にギャップがあることを見てきた。けれどその「統計」と医療の現場との間には、またギャップがあって、統計学上あまり意味のないことが堂々と通用しているケースもあるようである。

科学が発達するにつれて、すべてに通じて総合的に見渡せることは不可能に近くなりつつあるが、細かい知識はさておき、なるべく総合的にとらえたいと言う「目」なり「感覚」なりは重要ではなかろうか。今の例では $n = 20$ で合致が17組くらいであった場合、(大サンプルが入手できない状況であれば)これでポジティブな結論を主張することはやむを得ない場合もあるが、少なくとも「統計的な裏付けはない」ことを理解した上で主張することが望まれる。またこれを聞く側も、このことを理解した上で批判的に聞くことが望まれる。

参考文献

- [1] E. Feingold: *Markov processes for modeling and analyzing a new genetic mapping method*. J. Appl. Prob. **30** (1993) 766-779
- [2] D. Aldous: *Probability Approximations via the Poisson Clumping Heuristic*. Springer-Verlag. New York. (1989)

注1 数表の後に計算プログラムを載せておく。順に祖母-孫、兄弟、いとこの場合の計算プログラム (UBASIC) である。

注2 Simulation による計算結果の有効数字が一桁しかないことは次のようにして分かる。

N 回の simulation を行ったとき、一对の染色体についてどこかで合致の個数が b 以上となる回数は確率 α の二項分布に従う。従ってそのような回数の期待値は $N\alpha$ 、標準偏差は $\sqrt{N\alpha(1-\alpha)} \approx \sqrt{N\alpha}$ である。

$\alpha_{23} \approx 0.05$ のとき $\alpha \approx 0.00223$. $N = 10^5$ で

$$\sqrt{N\alpha(1-\alpha)} \approx 14.9, \sqrt{N\alpha(1-\alpha)}/N \approx 0.00015$$

$N = 2 \times 10^4$ で

$$\sqrt{N\alpha(1-\alpha)} \approx 6.7, \sqrt{N\alpha(1-\alpha)}/N \approx 0.00033$$

となる。これを $\alpha \approx 0.00223$ と比べると simulation によって得た α の値は有効数字一桁しかない。simulation による α の値の計算の有効数字を一桁増やすためには標準偏差を前の $\frac{1}{10}$ にする必要があるので、回数は 100 倍にしなければならない。そこで α を二桁求めるためには約 10^7 回の simulation が必要になる。

Table 1 Grandparent–Grandchild pair case

 $l = 100$ $n = 100$ $\lambda = 0.0100000000$

b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
50	0.9570057006	1.0000000000	76	0.0000032681	0.0000751640
51	0.9283564916	1.0000000000	77	0.0000010580	0.0000243337
52	0.8871863815	1.0000000000	78	0.0000003236	0.0000074423
53	0.8317421153	1.0000000000	79	0.0000000933	0.0000021468
54	0.7616991802	1.0000000000	80	0.0000000253	0.0000005830
55	0.6786120185	1.0000000000	81	0.0000000065	0.0000001488
56	0.5859727030	0.9999999984	82	0.0000000015	0.0000000356
57	0.4887922263	0.9999998015	83	0.0000000003	0.0000000080
58	0.3927826807	0.9999896028	84	0.0000000001	0.0000000017
59	0.3033641792	0.9997550184	85	0.0000000000	0.0000000003
60	0.2247797124	0.9971372859	86	0.0000000000	0.0000000001
61	0.1595494016	0.9816439012	87	0.0000000000	0.0000000000
62	0.1083628986	0.9284965460	88	0.0000000000	0.0000000000
63	0.0703596090	0.8132568639	89	0.0000000000	0.0000000000
64	0.0436428965	0.6416874461	90	0.0000000000	0.0000000000
65	0.0258461556	0.4524375954	91	0.0000000000	0.0000000000
66	0.0146065829	0.2871106889	92	0.0000000000	0.0000000000
67	0.0078734374	0.1662366928	93	0.0000000000	0.0000000000
68	0.0040460899	0.0890332153	94	0.0000000000	0.0000000000
69	0.0019812607	0.0445895096	95	0.0000000000	0.0000000000
70	0.0009239311	0.0210358329	96	0.0000000000	0.0000000000
71	0.0004100681	0.0093891441	97	0.0000000000	0.0000000000
72	0.0001730934	0.0039735778	98	0.0000000000	0.0000000000
73	0.0000694323	0.0015957235	99	0.0000000000	0.0000000000
74	0.0000264424	0.0006079973	100	0.0000000000	0.0000000000
75	0.0000095510	0.0002196490			

Simulation (10^5 repetitions)			Aldous approximation		
b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
67	0.00827	0.174	67	0.00787	0.166
68	0.00425	0.093	68	0.00405	0.089
69	0.00201	0.045	69	0.00199	0.045
70	0.00099	0.023	70	0.00093	0.021

Table 2 Grandparent-Grandchild pair case

 $l = 100$ $n = 20$ $\lambda = 0.0100000000$

b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
10	0.9578621601	1.0000000000	16	0.0553500394	0.7300825588
11	0.8785293625	1.0000000000	17	0.0151357065	0.2958632321
12	0.7305827773	1.0000000000	18	0.0028840944	0.0642715967
13	0.5257925384	0.9999999647	19	0.0003419865	0.0078361711
14	0.3139826738	0.9998279293	20	0.0000190214	0.0004374012
15	0.1499969865	0.9761948619			

Simulation (10^5 repetitions)			Aldous approximation		
b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
17	0.01520	0.297	17	0.01511	0.295
18	0.00281	0.063	18	0.00289	0.065
19	0.00027	0.006	19	0.00034	0.008

Table 3 Sibling case (Brothers and Sisters)

 $l = 100$ $n = 100$ $\lambda = 0.0400000000$

b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
75	0.9981121636	1.0000000000	88	0.0388710613	0.5982295108
76	0.9943411389	1.0000000000	89	0.0173278003	0.3310411474
77	0.9849605281	1.0000000000	90	0.0069273603	0.1477571125
78	0.9645011331	1.0000000000	91	0.0024684344	0.0552587309
79	0.9253938516	1.0000000000	92	0.0007775410	0.0177313162
80	0.8598896199	1.0000000000	93	0.0002141424	0.0049136903
81	0.7637205895	1.0000000000	94	0.0000508164	0.0011681252
82	0.6398814752	0.9999999999	95	0.0000101870	0.0002342754
83	0.4998390792	0.9999987999	96	0.0000016784	0.0000386022
84	0.3605264505	0.9999658084	97	0.0000002183	0.0000050208
85	0.2383442836	0.9980925629	98	0.0000000210	0.0000004835
86	0.1436306653	0.9717378321	99	0.0000000013	0.0000000307
87	0.0785678424	0.8477149074	100	0.0000000000	0.0000000010
Simulation (10^5 repetitions)			Aldous approximation		
b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
89	0.01723	0.330	89	0.01896	0.356
90	0.00691	0.147	90	0.00749	0.159
91	0.00248	0.056	91	0.00264	0.059
92	0.00081	0.018	92	0.00082	0.019

Table 4 Aunt–Niece pair case

 $l = 100$ $n = 100$ $\lambda = 0.0250000000$

b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
50	0.9978659257	1.0000000000	76	0.0000079785	0.0001834898
51	0.9943992572	1.0000000000	77	0.0000025878	0.0000595173
52	0.9867274989	1.0000000000	78	0.0000007928	0.0000182332
53	0.9714906308	1.0000000000	79	0.0000002290	0.0000052674
54	0.9442608505	1.0000000000	80	0.0000000623	0.0000014324
55	0.9003536966	1.0000000000	81	0.0000000159	0.0000003659
56	0.8362858806	1.0000000000	82	0.0000000038	0.0000000876
57	0.7514240666	1.0000000000	83	0.0000000009	0.0000000196
58	0.6490489397	1.0000000000	84	0.0000000002	0.0000000041
59	0.5361694846	0.9999999788	85	0.0000000000	0.0000000008
60	0.4219920286	0.9999966544	86	0.0000000000	0.0000000001
61	0.3156369525	0.9998372239	87	0.0000000000	0.0000000000
62	0.2240467382	0.9970743798	88	0.0000000000	0.0000000000
63	0.1508402825	0.9767321728	89	0.0000000000	0.0000000000
64	0.0963264837	0.9026657340	90	0.0000000000	0.0000000000
65	0.0583714105	0.7492553778	91	0.0000000000	0.0000000000
66	0.0335832155	0.5441912337	92	0.0000000000	0.0000000000
67	0.0183548052	0.3469378168	93	0.0000000000	0.0000000000
68	0.0095338810	0.1977469884	94	0.0000000000	0.0000000000
69	0.0047075045	0.1028464522	95	0.0000000000	0.0000000000
70	0.0022096871	0.0496063748	96	0.0000000000	0.0000000000
71	0.0009858602	0.0224305766	97	0.0000000000	0.0000000000
72	0.0004179086	0.0095678406	98	0.0000000000	0.0000000000
73	0.0001682222	0.0038619587	99	0.0000000000	0.0000000000
74	0.0000642538	0.0014767939	100	0.0000000000	0.0000000000
75	0.0000232666	0.0005349960			

Simulation (2×10^4 repetitions)			Aldous approximation		
b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
67	0.01840	0.348	67	0.01957	0.365
68	0.00865	0.181	68	0.01010	0.208
69	0.00415	0.091	69	0.00496	0.108
70	0.00215	0.048	70	0.00231	0.052
71	0.00090	0.020	71	0.00103	0.023

Table 5 Cousins pair case

 $l = 100$ $n = 100$ $\lambda = 0.0400000000$

b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
25	0.9982426859	1.0000000000	63	0.0000000000	0.0000000000
26	0.9944207609	1.0000000000	64	0.0000000000	0.0000000000
27	0.9848143727	1.0000000000	65	0.0000000000	0.0000000000
28	0.9641652757	1.0000000000	66	0.0000000000	0.0000000000
29	0.9258621816	1.0000000000	67	0.0000000000	0.0000000000
30	0.8640081562	1.0000000000	68	0.0000000000	0.0000000000
31	0.7763037478	1.0000000000	69	0.0000000000	0.0000000000
32	0.6661907253	1.0000000000	70	0.0000000000	0.0000000000
33	0.5427675447	0.9999999848	71	0.0000000000	0.0000000000
34	0.4182593863	0.9999961205	72	0.0000000000	0.0000000000
35	0.3043268050	0.9997626872	73	0.0000000000	0.0000000000
36	0.2090423467	0.9954550998	74	0.0000000000	0.0000000000
37	0.1356891637	0.9650530697	75	0.0000000000	0.0000000000
38	0.0833638747	0.8649388977	76	0.0000000000	0.0000000000
39	0.0485701294	0.6818250461	77	0.0000000000	0.0000000000
40	0.0268889901	0.4657618671	78	0.0000000000	0.0000000000
41	0.0141707978	0.2798240581	79	0.0000000000	0.0000000000
42	0.0071209170	0.1515694181	80	0.0000000000	0.0000000000
43	0.0034166409	0.0756988041	81	0.0000000000	0.0000000000
44	0.0015670388	0.0354273852	82	0.0000000000	0.0000000000
45	0.0006876682	0.0156973015	83	0.0000000000	0.0000000000
46	0.0002889478	0.0066247192	84	0.0000000000	0.0000000000
47	0.0001163211	0.0026719659	85	0.0000000000	0.0000000000
48	0.0000448850	0.0010318453	86	0.0000000000	0.0000000000
49	0.0000166076	0.0003819060	87	0.0000000000	0.0000000000
50	0.0000058939	0.0001355513	88	0.0000000000	0.0000000000
51	0.0000020067	0.0000461532	89	0.0000000000	0.0000000000
52	0.0000006556	0.0000150778	90	0.0000000000	0.0000000000
53	0.0000002055	0.0000047268	91	0.0000000000	0.0000000000
54	0.0000000618	0.0000014220	92	0.0000000000	0.0000000000
55	0.0000000178	0.0000004105	93	0.0000000000	0.0000000000
56	0.0000000049	0.0000001137	94	0.0000000000	0.0000000000

Table 5 (continued)

b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
57	0.0000000013	0.0000000302	95	0.0000000000	0.0000000000
58	0.0000000003	0.0000000077	96	0.0000000000	0.0000000000
59	0.0000000001	0.0000000019	97	0.0000000000	0.0000000000
60	0.0000000000	0.0000000004	98	0.0000000000	0.0000000000
61	0.0000000000	0.0000000001	99	0.0000000000	0.0000000000
62	0.0000000000	0.0000000000	100	0.0000000000	0.0000000000
Simulation (2×10^4 repetitions)			Aldous approximation		
b	1 chromosome	23 chromosomes	b	1 chromosome	23 chromosomes
41	0.01350	0.268	41	0.01498	0.293
42	0.00655	0.140	42	0.00748	0.159
43	0.00305	0.068	43	0.00357	0.079
44	0.00135	0.031	44	0.00163	0.037
45	0.00060	0.014	45	0.00071	0.016

```

10  point -8:word -9
20  input "l=";L
30  if or{type(L)<>1,L<1} then end
40  input "n=";N
50  if type(N)<>1 then end
60  input "lambda=";Lambda
70  if or{type(Lambda)>3,Lambda<0} then end
80  print "Table 1 Grandparent-Grandchild pair case":print
90  print "l=",L:print "n=",N:print "lambda=",:print using(,10),Lambda:print
100 print " b"," 1 chromosome"," 23 chromosomes"
110 dim A(N),A1(N)
120 for B=N\2 to N
130  X=N*L*Lambda:Y#=1:E#=exp(X)
140  clr S,P1,P2,P
150  for I=0 to B-1
160    A(I)=combi(N,I)/2^N
170    P2+=A(I)
180  next I
190  P1=1-P2
200  P=P1*Y#/E#
210  repeat
220    inc S
230    Y#=Y#*X/S
240    A1(0)=A(1)/N
250    for I=1 to B-1
260      A1(I)=A(I-1)*(N-(I-1))/N+A(I+1)*(I+1)/N
270    next I
280    P1=A(B-1)*(N-(B-1))/N:P2-=P1
290    block A(0..(B-1))=block A1(0..(B-1))
300    P+=(1-P2)*Y#/E#
310  until Y#=0
320  print B,:print using(,10),P,1-(1-P)^23
330  next B

```

```

10 point -8:word -9:emaword -9
20 input "l=";L
30 if or{type(L)<>1,L<1} then end
40 input "n=";N
50 if type(N)<>1 then end
60 input "lambda=";Lambda
70 if or{type(Lambda)>3,Lambda<0} then end
80 print "Table 3 Sibling case (Brothers and Sisters)":print
90 print "l=",L:print "n=",N:print "lambda=",Lambda:print using(,10),Lambda:print
100 print " b"," 1 chromosome"," 23 chromosomes"
110 dim ema(N,N),ema(1,N,N)
120 for B=3*N\4 to N
130 X=N*L*Lambda:Y#=1:E#=exp(X)
140 clr S,P1,P2,P
150 for I1=0 to B-1
160 for I2=0 to B-1-I1
170 ema(I1,I2)=combi(N,I1+I2)*combi(I1+I2,I1)/(2^I1*4^I2*4^(N-I1-I2))
180 P2+=ema(I1,I2)
190 next I2
200 next I1
210 P1=1-P2
220 P=P1*Y#/E#
230 repeat
240 inc S
250 Y#=Y#*X/S
260 ema(1;0,0)=ema(1,0)/(2*N)
270 for I=1 to B-1
280 ema(1;0,I)=ema(1,I)/(2*N)+ema(1,I-1)/(2*N)
290 for I1=1 to I-1
300 I2=I-I1
310 ema(1;I1,I2)=ema(I1-1,I2)*(N-(I1+I2-1))/N+ema(I1+1,I2)*(I1+1)/(2*N)+ema(
I1+1,I2-1)*(I1+1)/(2*N)+ema(I1-1,I2+1)*(I2+1)/N
320 next I1
330 ema(1;I,0)=ema(I-1,0)*(N-(I-1))/N+ema(I+1,0)*(I+1)/(2*N)+ema(I-1,1)/N
340 next I
350 clr P1
360 for I=0 to B-1
370 P1+=ema(I,B-1-I)
380 next I
390 P1*=(N-(B-1))/N:P2-=P1
400 block ema(0..(B-1),0..(B-1))=block ema(1;0..(B-1),0..(B-1))
410 P+=(1-P2)*Y#/E#
420 until Y#=0
430 print B,:print using(,10),P,1-(1-P)^23
440 next B

```

```

10 point -8:word -9:emaword -9
20 input "l=";L
30 if or{type(L)<>1,L<1} then end
40 input "n=";N
50 if type(N)<>1 then end
60 input "lambda=";Lambda
70 if or{type(Lambda)>3,Lambda<0} then end
80 print "Table 5 Cousins pair case ":print
90 print "l=",L:print "n=",N:print "lambda=",,:print using(,10),Lambda:print
100 print " b"," 1 chromosome"," 23 chromosomes"
110 dim ema(N,N),ema(1;N,N)
120 for B=N\4 to N
130 X=N*L*Lambda:Y#=1:E#=exp(X)
140 clr S,P1,P2,P
150 for I=0 to B-1
160 for J=0 to N-I
170 ema(I,J)=combi(N,I+J)*combi(I+J,I)/(4^(I+J)*2^(N-I-J))
180 P2+=ema(I,J)
190 next J
200 next I
210 P1=1-P2
220 P=P1*Y#/E#
230 repeat
240 inc S
250 Y#=Y#*X/S
260 ema(1;0,0)=ema(1,0)/(2*N)+ema(0,1)/(2*N)+ema(0,0)/2
270 for J=1 to N-1
280 ema(1;0,J)=ema(1,J-1)/(2*N)+ema(0,J-1)*(N-J+1)/(4*N)+ema(1,J)/(2*N)+ema(
0,J+1)*(J+1)/(2*N)+ema(0,J)*(N-J)/(2*N)
290 next J
300 ema(1;0,N)=ema(1,N-1)/(2*N)+ema(0,N-1)/(4*N)+ema(1,N)/(2*N)
310 for I=1 to B-1
320 ema(1;I,0)=ema(I-1,1)/(2*N)+ema(I-1,0)*(N-I+1)/(4*N)+ema(I+1,0)*(I+1)/(2
*N)+ema(I,1)/(2*N)+ema(I,0)*(N-I)/(2*N)
330 for J=1 to N-I
340 ema(1;I,J)=ema(I+1,J-1)*(I+1)/(2*N)+ema(I-1,J+1)*(J+1)/(2*N)+ema(I-1,J)*
(N-(I+J-1))/(4*N)+ema(I,J-1)*(N-(I+J-1))/(4*N)+ema(I+1,J)*(I+1)/(2*N)+ema(I,J+1)
*(J+1)/(2*N)+ema(I,J)*(N-(I+J))/(2*N)
350 next J
360 next I
370 clr P1
380 for J=0 to N-B+1
390 P1+=ema(B-1,J)*(N+J-B+1)/(4*N)
400 next J
410 P2-=P1
420 block ema(0...(B-1),0..N)=block ema(1;0...(B-1),0..N)
430 P+=(1-P2)*Y#/E#
440 until Y#=0
450 print B,:print using(,10),P,1-(1-P)^23
460 next B

```