step 2　各k_iについて　$y_i=\sqrt{k_i}-\frac{1}{\sqrt{k_i}}\quad(i=1,\quad,n)$を計算する。

step 3　各y_iの値を階級値と考えて度数分布表を作る。階級の数は　$nk(\leq n)$　である。出来上がった度数分布表の各階級値を　kk_iとし，各階級の度数を　kd_i　とする。ここで，

$kk_i<kk_j\ \ (1\leq i<j\leq nk)$

step 4　$\hat{\Phi}_i=0.5\left(1+\frac{\sum_{j=1}^{i}kd_j}{n+1}\right),(1\leq i\leq nk)$　を計算する。

step 5　$\left(kk_i,\Phi^{-1}\left(\hat{\Phi}_i\right)\right)\ (1\leq i\leq nk)$を打点する。

ここて　$\Phi^{-1}(x)$　は，標準正規分布の分布関数の逆関数を表す。

以上の５つの step が本質的な逆ガウス型確率紙の構成法である。確率紙の解釈として，$\left(kk_i,\Phi^{-1}\left(\hat{\Phi}_i\right)\right)$ を打点した図において直線性が見られる場合には，実用上はデータは２母数逆ガウス型分布に従うとする。

確率紙作成において，表現上の見かけをよくするためには，縦軸を等間隔の座標系である　$\Phi^{-1}(x)$　から　$\Phi(x)$　にした目盛りや，横軸に関しても等間隔の座標系である y と kの目盛りを併記することを推奨する。

２　３　２　　確率紙の理論的な課題と実用上の対応

この確率紙の利用に関しては，以下の４の課題が理論的には考えられる。しかし，これらは実用上の大きな問題でない。

（１）　提案されている確率紙においては，μの推定値か求められない。その結果としてμと　$\bar{x}_A$　との間に大きな開きがある場合，$P_r\left(\frac{1}{k}\leq\frac{X_i}{\mu}\leq k\right)$　と確率推定値との間に大きな差が生じ，直線性に影響が現れる可能性がある。

（2） 母集団のレベルでも，(k,Φ_k) が直線上に並べば，$X \sim IG(\mu,c^2)$ であるという逆命題は必ずしも成立するとは保証されていない。

（3） $(x,F(x))$を打点するのではなく，$P_r\left(\frac{1}{k} \le \frac{X_i}{\mu} \le k\right)$の推定値に対応する$(k,\hat{\Phi}_k)$を打点しなければならない。

（4） この確率紙は$\bar{x}_A$を求めることを前提にしており，打ち切りデータの解析には利用できない。

以上の課題に対する実用上の対応としては，以下のように処理する。

（1） 母集団が逆ガウス型分布にしたがっているにもかかわらず，異常値などの存在によって確率紙上で直線性が認められないことがあったとしても，μ と $\bar{x}_A$ との間の大きな開きのために，偶然的に確率紙上で直線性が認められたという場合はほとんど起こらないとして取り扱う。

（2） 逆ガウス型分布に従うとする仮説は否定できないのみならず，機能的代替物としての代替的分布の存在は今のところないので，逆ガウス型分布に従うとして取り扱ってもよいであろう。したがって，実用上$(k,\hat{\Phi}_k)$がほぼ直線上に並んだならば，$X \sim IG(\mu,c^2)$ とする。

（3） 表現方法に関することは，実用上は問題がない。

（4） 打ち切りデータには，この確率紙を適用しない。

２　４　３母数逆ガウス型分布とその未知パラメータの推定

発症日のデータに基づいてその原因物質の摂取日を推定する問題の定式化は，発症日に関するパターンを記述する曲線として逆ガウス型分布の確率密度関数を仮定しているので，原点が未知の３母数逆ガウス型分布の未知母数推定問題として捉えることが可能である。しかし，２母数の寿命分布の推定と異なる点がある。３母数寿命分布の未知母数の推定では最尤推定量の性質として，特に，未知原点に関して推定の困難さが伴う。未知原点の推定値として標本最小値を考えると，尤度関数が無限大になる寿命分布がある。その一例が対数正規分布である。逆ガウス型分布はこのような取り扱いにくい性質を持っていない。

２　４　１　　３母数逆ガウス型分布の定義

２母数逆ガウス型分布の定義領域は，正の領域 $R^+=(0,\infty)$ である。ここで，定義領域の下限を未知とするために１つの新たな母数を導入し３母数逆ガウス型文を導入する。

３母数逆ガウス型分布 $IG_\lambda(\mu,\sigma^2)$ は，既に定義している２母数逆ガウス型分布を用いて以下のように定義される。

$$1+\lambda\frac{X-\mu}{\sigma}\sim IG\left(1,|\lambda|^2\right)$$

ここて，定義領域は，$sgn(\lambda)\ X>sgn(\lambda)\left(\mu-\frac{\sigma}{\lambda}\right)$ であり，

$-\infty<\mu<\infty,0<\sigma<\infty,-\infty<\lambda<\infty,\lambda\neq 0$ てある。また，$sgn(\lambda)=\begin{cases}-1 & \lambda<0\\ 0 & \lambda=0\\ 1 & \lambda>0\end{cases}$ である。

以上の定義に基づいて，以下のモーメントが計算できる。

$$E[X]=\mu,$$
$$E\left[(X-\mu)^2\right]=\sigma^2,$$
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]=3\lambda$$

また，２母数逆ガウス型分布と３母数逆ガウス型分布の間には以下のような関係かある。

$$\begin{array}{lccl} & a\ IG\left(\mu,\frac{\sigma}{\mu}\right) & \Leftrightarrow & IG\left(a\mu,\frac{\sigma}{\mu}\right) & \\ \{a>0,b=0,\lambda=\sigma/\mu\} & \uparrow & & \uparrow & \{a>0,b=0,\lambda=\sigma/\mu\} \\ & a\ IG_\lambda\left(\mu,\sigma^2\right)+b & \Leftrightarrow & IG_\lambda\left(a\mu+b,(a\sigma)^2\right) & \\ \{\lambda\to 0\} & \downarrow & & \downarrow & \{\lambda\to 0\} \\ & a\ N\left(\mu,\sigma^2\right)+b & \Leftrightarrow & N\left(a\mu+b,(a\sigma)^2\right) & \end{array}$$

２　４　２　　未知母数の推定

Iwase and Kanefuji(1992)は前節で定義した３母数逆ガウス型分布に対する未知母数の最尤推定法を提案している。ここでは最尤推定方程式として得られた結果のみを要約して述へると，以下のような最尤推定量と推定方程式を求められる。また，推定値については，以下の推定方程式をニュートン法などにより数値的に求めることが可能である。

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\sigma_0 = \tilde{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2\right]^{\frac{1}{2}},$$

$$\lambda_0 = \tilde{\lambda} = \frac{1}{3n}\sum_{i=1}^{n}\left(\frac{x_i - \hat{\mu}}{\tilde{\sigma}}\right)^3,$$

$$f(\lambda_k) = \frac{1}{n}\sum_{i=1}^{n}\left(1+\lambda_k\frac{x_i-\hat{\mu}}{\sigma_{k-1}}\right)^{-2} - 1$$
$$+3\left[1-\frac{1}{n}\sum_{i=1}^{n}\left(1+\lambda_k\frac{x_i-\hat{\mu}}{\sigma_{k-1}}\right)^{-1}\right]\frac{1}{n}\sum_{i=1}^{n}\left(1+\lambda_k\frac{x_i-\hat{\mu}}{\sigma_{k-1}}\right)^{-1},$$

$$g(\sigma_k) = \frac{1}{n}\sum_{i=1}^{n}\left(1+\lambda_k\frac{x_i-\hat{\mu}}{\sigma_k}\right)^{-1} - 1 - \lambda_k^2,$$

$$\varepsilon = 10^{-3}, \qquad k = 1$$

３　　２つの母集団の比較

集団食中毒性の特徴に，各地で発生している散発事例が、実は同一の汚染を原因としている事例である場合がある。この一例として，1998 年６月に富山、神奈川、東京等で発生した同一の加工業者が出荷したイクラを汚染原因とする事件があり，これは厚生省によって報告されている。一般に，このような食中毒の発生を散在的集団発生(diffuse outbreak)と呼んでいる。このケースでは，各地で分離された菌のＤＮＡ型が同一であったため原因食材および加工地が同一の業者であることが判明し、当該業者のイクラを回収することにより、それ以上の被害者の発生が防がれた。このことは画期的なことであったが，現実にこれらのことを行うには，候補となる集団をいかに選択するか，そして候補集団の菌のＤＮＡ型鑑定等を短時間で行うことが必要である。そこで，探索的な統計解析を行って候補集団を決定し，その後精細なＤＮＡ型の判定を行うことは資源の有効活用となる。

そこで，本節では以下のような統計的問題を検討し，これによって候補集団の選択を行う。

３　１　２つの逆ガウス型分布に関する検定

ここての仮定としては，発症日のパターンは逆ガウス型分布で記述可能であり，同一の原因物質を摂取した集団の発症パターンにおける各未知母数は，必然的に同じ値を取ることを仮定している。また，標本集団では構成が異なり（年齢分布）必然的に，発生分布の推定した未知パラメータが異なることは自然である。

二つの２母数逆ガウス型分布の母数に関する推定検定を行うために以下のことを仮定し，３つの検定方式を導入する。

仮定

$$\mathrm{X}_i \sim IG\left(\mu_x, c_x^2\right)$$
$$Y_i \sim IG\left(\mu_y, c_y^2\right)$$

且つ，X_i, Y_j は互いに独立とする。

基本定理

完備十分統計量は，それぞれ$\left(\sum \mathrm{X}_i, \sum \mathrm{X}_i^{-1}\right), \left(\sum Y_i, \sum Y_i^{-1}\right)$である。

$$\bar{\mathrm{X}}_A \sim IG\left(\mu_X, \frac{c_X^2}{n_X}\right),$$
$$\frac{n_X}{c_X^2}\frac{\mu_x}{\bar{\mathrm{X}}_A}\left\{\frac{\bar{\mathrm{X}}_A}{\bar{\mathrm{X}}_H} - 1\right\} \sim \chi^2\left(n_X - 1\right)$$

ここで，$\bar{\mathrm{X}}_A$ と $\bar{\mathrm{X}}_H$ はそれぞれ，算術平均と調和平均を表わす。且つ、これらは互いに独立てある。

初めに，c が既知のときの平均に関する比の２つの検定を構成する。

定理１

$c_X = c_Y = c$ のときは，c か未知であっても

$$\frac{\sum_{i=1}^{n_X}\left(\frac{\mu_X}{X_i}-\frac{\mu_X}{\overline{X}_A}\right)}{\sum_{i=1}^{n_Y}\left(\frac{\mu_Y}{Y_i}-\frac{\mu_Y}{\overline{Y}_A}\right)}$$

は分子の自由度n_X-1，分母の自由度n_Y-1のＦ分布に従う。従って，帰無仮説H_0 $\mu_X=\mu_Y$の下では検定統計量

$$\frac{\sum_{i=1}^{n_X}\left(\frac{1}{\mathrm{X}_i}-\frac{1}{\overline{\mathrm{X}}_A}\right)}{\sum_{i=1}^{n_Y}\left(\frac{1}{Y_i}-\frac{1}{\overline{Y}_A}\right)}$$

は分子の自由度n_X-1，分母の自由度n_Y-1のＦ分布に従う。また、対立仮説は両側でも片側でも検定可能である。

定理２

両方のcは異なるけれどもわかっている場合，帰無仮説H_0 $\mu_X=\mu_Y$の下で

$$\frac{c_Y^2\sum_{i=1}^{n_X}\left(\frac{1}{\mathrm{X}_i}-\frac{1}{\overline{\mathrm{X}}_A}\right)}{c_X^2\sum_{i=1}^{n_Y}\left(\frac{1}{Y_i}-\frac{1}{\overline{Y}_A}\right)}$$

は分子の自由度n_X-1，分母の自由度n_Y-1のＦ分布に従う。同時に、対立仮説は両側でも片側でも検定可能である。

次に，cに関する比の検定を構成する。

定理３．

μ_X,μ_Yは共に既知である場合，帰無仮説$H_0.c_X^2/c_Y^2=c_0^2$の検定統計量

$$\frac{\sum_{i=1}^{n_X}\left(\frac{\mu_X}{\mathrm{X}_i}-\frac{\mu_X}{\overline{\mathrm{X}}_A}\right)}{c_0^2\sum_{i=1}^{n_Y}\left(\frac{\mu_Y}{Y_i}-\frac{\mu_Y}{\overline{Y}_A}\right)}$$

は分子の自由度$n_\lambda - 1$，分母の自由度$n_\gamma - 1$のＦ分布に従う。また、対立仮説は両側ても片側でも検定可能てある。

ここて、以上の定理をまとめると、以下のようになる。

（１）$c_X^2 = c_\gamma^2$（未知）のとき，帰無仮説H_0 $\mu_X = \mu_Y$の両側片側検定が可能である。また、帰無仮説H_0 $\mu_X / \mu_Y = \mu_0$の両側片側検定まで拡張可能である。

（２）c_X^2, c_Y^2は共に既知のとき，帰無仮説H_0 $\mu_X = \mu_Y$の両側片側検定か可能である。同様に、帰無仮説H_0 $\mu_X / \mu_Y = \mu_0$の両側片側検定まで拡張可能である。

（３）μ_X, μ_Yは共に既知であるとき，帰無仮説H_0 $c_X^2 / c_Y^2 = c_0^2$の両側片側検定か可能である。

検定結果の解釈についてまとめると、以下のようになる。

（１） 異種の病原菌であるとすると平均発症時間に差があるかもしれない。逆に平均の比の検定を行うことによって，もしも有意てあったならば異種の病原菌であることを疑わせる根拠の一つになる。
（２） もしもｃか異なるとすると，同し平均発病時間を与える病原菌でも異なるメカニズムての発現をもたらす病原菌であることを疑わせる。

本検定方式の課題は、感染した時刻と発現した時刻とかわかっていなければならない。しかし通常は感染した時刻はわからないと考えるのか自然てある。仮にn人の人間か感染の結果，症状か発現したとすれば，これらの時刻（時間ではない）のデータから感染した時刻の推定が望まれる。病原菌とは何ら関係ない（即ち病原菌にとっては知ったことではない）適当な時刻の起点（例えば正月元日，このような起点は人間だけの都合て勝手に決められたものてあり病原菌にとっては何ら関係ない起点てある。ただし，季節などによる気温などの変化は関係あると考えるのか自然である。）から何日目に症状が発現したとの情報しかないときに，その適当な時刻の起点から何日目が感染日であると推定することか必要である。

確率変数Xを感染してから発現するまでの時間とする。適当な時刻の起点からの時間をδとすれば，適当な起点からの時刻$\delta + X$（定数とする）の実現値が観測値となる。無論δは未知である。従って，Xが逆ガウス型分布に従っていると仮定しても，適当な起

点からの発症の時刻である D= δ +Xは逆ガウス型分布には従わない。ここで原点未知の３母数逆ガウス型分布の利用が必要である。

４　　予測手法の構築

本節では，同じ原因食材をほぼ同時に摂取した集団においてある時点までの食中毒患者の発症データに基づいて，それ以降の対象となる集団の規模予測を行うための統計手法の検討を行った。

原因食材がある程度，確定されれば当然それを食べた集団の人数もある程度把握可能であるが，これまでに述べたようにＯ－１５７を原因物質とする集団食中毒の場合，発症までの時間が平均すると数日必要であり，その結果原因食材に関する確定が難しく，それに伴ってどの程度の人が原因物質を含む原因食材を摂取しているかは，到底早期に確定出来ない。

このような状況では，ある時点までの発症者の初症状の発生時刻に関するデータのみに基づいてそれ以降の発生状況を予測し，行政としての対策を施す必要がある。

ここで，注意しなければならない点は現状の「感染症の予防及び感染症の患者に対する医療に関する法律」に基づくデータ収集においても，reporting delay あるいは batch reporting により報告のタイムラグが生じる。実際にこれらの現象の影響がなくなるには１ヶ月程度必要になる。このような状況からもわかるように現実の対応には，集中的な人的・物的資源の投入による疫学調査が必要であり，それによってデータ解析の精度が高まる。

当然，このような状況を反映して考察した統計解析の手法が必要であるが，現状ではそれを補う統計手法は見当たらない。また，本研究でもそこまでは踏み込んで考察していないし，そのためにはさらなる研究を進めなければならない。

次に，データが reporting delay あるいは batch reporting により報告のタイムラグが生じないという非常に強い条件を仮定して予測に関する統計的手法の構築を進める。これは，ある時点においてそれまでの発症日のデータが正確に収集されている場合，それ以降の発症状況を予測する。

４　１　　微分方程式による予測式の構築

初めに，以下のような微分方程式を考える。

微分方程式

$$-\frac{1}{y}\frac{dy}{dx}=A+\frac{B}{x}+\frac{C}{x^2}$$

これは，yを時刻xでの発症者数とし，そのときの濃度当たりの増加率がいわば時間xの逆数に関して二次式で表されるモデルである。時間が経てば経つほど相対的な増加率か減少することを意味している。

この微分方程式の解は，

$$-\log y = k + Ax + B\log x - \frac{C}{x}$$

である。但しkは積分常数である。ここでは，対数関数は自然対数を採用する。

このとき上式をyについて解けば，

$$y = constant \quad x^{-B} exp\left(-Ax + \frac{C}{x}\right)$$

となる。

また，領域$(0, \infty)$の上ての積分か1になるように定数部分を決定すれば，この解曲線は2母数逆ガウス型分布の確率密度関数になる。これは，この微分方程式の初期条件てはなくて，このような制約条件て定数項が決定される。

先の微分方程式は

$$-\frac{x^2}{y}\frac{dy}{dx} = Ax^2 + Bx + C$$

と変形される。これに相当する差分方程式として

$$-\frac{x_i^2}{y_i}\frac{y_{i+1} - y_i}{h} = Ax_i^2 + Bx_i + C$$

を考え，上式左辺をz_iと置く。ここて，$h = x_{i+1} - x_i (i = 1, \ .n-1)$てある。

$$z_i = Ax_i^2 + Bx_i + C \quad (i = 1, \quad , n-1)$$

である。

$$\Delta z_i = z_{i+1} - z_i = 2Ahx_i + Ah^2 + Bh \quad (i = 1, \ ., n-2)$$
$$\Delta^2 z_i = 2Ah^2 \quad (i = 1, . \ , n-3)$$

よりn組の(x_i, y_i)からn-3個の$\Delta^2 z_i = 2Ah^2$が得られ，これらの算術平均からAの推定値が得られる。次に，n組の(x_i, y_i)からn-2個の$\Delta z_i = z_{i+1} - z_i = 2Ahx_i + Ah^2 + Bh$が得られ，これらの算術平均より$B$の推定値が得られる。同様に，$n$組の$(x_i, y_i)$から$n$-1

個の $z_i = Ax_i^2 + Bx_i + C$ か得られ，これらの算術平均より C の推定値が得られる。以下に各ステップでの途中結果を示している。

$$
\begin{array}{llllll}
x_1 & y_1 & & & & \\
x_2 = x_1 + h & y_2 & z_1 = -\dfrac{x_1^2}{y_1}\dfrac{y_2 - y_1}{h} & & & \\
x_3 = x_2 + h & y_3 & z_2 = -\dfrac{x_2^2}{y_2}\dfrac{y_3 - y_2}{h} & \Delta z_1 = z_2 - z_1 & & \\
x_4 = x_3 + h & y_4 & z_3 = -\dfrac{x_3^2}{y_3}\dfrac{y_4 - y_3}{h} & \Delta z_2 = z_3 - z_2 & \Delta^2 z_1 = \Delta z_2 - \Delta z_1 & \\
 & & & \cdot & & \\
x_i = x_{i-1} + h & y_i & z_{i-1} = -\dfrac{x_{i-1}^2}{y_{i-1}}\dfrac{y_i - y_{i-1}}{h} & \Delta z_{i-2} = z_{i-1} - z_{i-2} & \Delta^2 z_{i-3} = \Delta z_{i-2} - \Delta z_{i-3} & \\
\cdot & \cdot & & & \vdots & \\
x_{n-2} = x_{n-3} + h & y_{n-2} & z_{n-3} = -\dfrac{x_{n-3}^2}{y_{n-3}}\dfrac{y_{n-2} - y_{n-3}}{h} & \Delta z_{n-4} = z_{n-3} - z_{n-4} & \Delta^2 z_{n-5} = \Delta z_{n-4} - \Delta z_{n-5} & \\
x_{n-1} = x_{n-2} + h & y_{n-1} & z_{n-2} = -\dfrac{x_{n-2}^2}{y_{n-2}}\dfrac{y_{n-1} - y_{n-2}}{h} & \Delta z_{n-3} = z_{n-2} - z_{n-3} & \Delta^2 z_{n-4} = \Delta z_{n-3} - \Delta z_{n-4} & \\
x_n = x_{n-1} + h & y_n & z_{n-1} = -\dfrac{x_{n-1}^2}{y_{n-1}}\dfrac{y_n - y_{n-1}}{h} & \Delta z_{n-2} = z_{n-1} - z_{n-2} & \Delta^2 z_{n-3} = \Delta z_{n-2} - \Delta z_{n-3} &
\end{array}
$$

ここで

$$z_i = Ax_i^2 + Bx_i + C$$

である。故に，以下の結果が得られる。

$$
\begin{aligned}
\Delta z_{i+1} &= z_{i+2} - z_{i+1} \\
&= \left(Ax_{i+2}^2 + Bx_{i+2} + C\right) - \left(Ax_{i+1}^2 + Bx_{i+1} + C\right) \\
&= \left(A\left(x_{i+1} + h\right)^2 + B\left(x_{i+1} + h\right) + C\right) - \left(Ax_{i+1}^2 + Bx_{i+1} + C\right) \\
&= \left(Ax_{i+1}^2 + 2Ahx_{i+1} + Ah^2 + Bx_{i+1} + Bh + C\right) \\
&\quad - \left(Ax_{i+1}^2 + Bx_{i+1} + C\right) \\
&= 2Ahx_{i+1} + Ah^2 + Bh
\end{aligned}
$$

$$
\begin{aligned}
\Delta^2 z_i &= \Delta z_{i+1} - \Delta z_i \\
&= \left(2Ahx_{i+1} + Ah^2 + Bh\right) \\
&\quad -\left(2Ahx_i + Ah^2 + Bh\right) \\
&= 2Ah\left(x_{i+1} - x_i\right) \\
&= 2Ah^2
\end{aligned}
$$

４　２　発生強度関数

混乱をきたさないために言葉の定義について述べると，ここて求める曲線は分布関数でも確率密度関数でもない。一つの反応曲線とでも言うべきものてある。患者の発生の分布と言う表現は混乱を招くものである。降雨強度（単位時間当たりの降水量）の表現を真似れば発生強度とでも表現すべきものである。

また，先の方法で求めたＢが3/2に近いものであれば，発生強度は逆ガウス型分布の確率密度関数で記述できることを意味する。もしも一斉に各個人が菌に曝露されて，症状が発現するまでの時間か個体差で(又は固体環境差)攪乱されると仮定すれば，また，菌に曝露されてある一定の濃度に達すると発現すると仮定すれば，ウィーナー確率過程での初期通過時間分布としての逆ガウス型分布か妥当である。但し，この場合の分布とは曝露されてから発現するまでの時間のデータについての分布てある。

当然，半日または一日の幅で度数表が作成されていると解釈すれば，発生強度関数と初期通過時間分布とが同し関数になることは合理的であると考えられる。

４　３　提案した手法の課題とその対応

ここで提案した方法における課題を３つ述べると以下のように要約てきる。

（１）　予測可能になるまでに最低４周期の時間か必要になる。
（２）　曝露時の時点が未知てあるケースには提供できない。
（３）　データがばらついているものにも適用が難しい。

実用上においては，以下のような対策を施すことにより以上の課題か解決できると考えている。

（１）　なるべく階級の幅を小さくして，正確に初期発症時の度数分布表を作成する。出来れば，６時間間隔か１２時間間隔か望ましい。
（２）　実用上は，付加情報から探索的に曝露時の予測値を決めて適用する。
（３）　迅速な疫学調査を行い reporting delay や batch reporting によるデータ

のタイムラグを減らす。また，より重要なことはこの方法は差分を利用しているので，度数分布の変動に関して非常に敏感に反応し推定値が適当でない値を求めるケースも生じる。そこで，統計的な性質は保証できないが，度数分布を平滑化し各時点の予測値を用いることを薦める。

5　まとめ

各節において構築した手法の関係をまとめると図２．１になる。

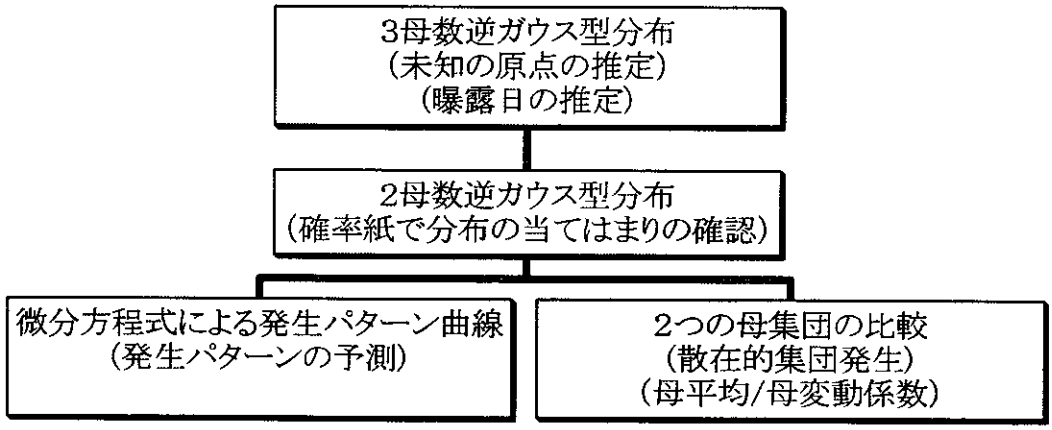

図２．１　統計手法の関連

Ｏ－１５７データの解析のうち曝露時点の推定に関して，対数正規分布を用いた研究として丹後(1998)がある。しかし，Ｏ－１５７データの発症パターンに関して対数正規分布を用いる明確な記述はない。

また，一般に３母数対数正規分布，３母数ワイブル分布，３母数ガンマ分布を用いて，未知母数の最尤推定を行う場合，未知原点の推定値は標本の最小値なることが知られており，その問題を克服する様々な手法が提案されている。丹後(1998)の方法も，３母数対数正規分布における最尤推定方程式を求める場合において，非線型方程式の解の収束性に関する課題を克服するための一つの方法である。

３母数逆ガウス型分布を仮定して推定を行う場合には以上に述べた問題はなく，特に実用上有効な性質の一つとしてほとんどのケースにおいて適切な解が求まる点が上げられる。

その他の予測手法としては経験的ベイズ手法が有効であると考えている。しかし，この手法は実際の適用に関して，現状では以下の問題点があると考えている。

この手法を利用するためには，ある程度原因食材や集団の特性が一致する集団に対する過去の発生データに関する事前情報を確定（推定）する必要がある。しかし，現実に

は,集団としての全データの正確性(発症日の記録がすべてに個体にあるとは限らない)か無い。

参考文献

岩瀬晃盛, 久保田洋志, 中村信人, 平木秀作, 畝正二　1991, 逆ガウス型確率紙の試作と検討,品質 vol 21, vol 3, pp 215-225

K Iwase and K Kanefuji 1992, Estimation for 3-Parameter Inverse Gaussian Distribution with Unknown Origin, *Technical Report of Hiroshima University, No 92-D4*

丹後俊郎　1998 潜伏期間に対数正規分布を仮定した集団食中毒の曝露時点の最尤推定法, 日本公衛誌, 45, 129-141

第３章　統計解析システムの実装に関する検討

１　　システム

健康危機関連の統計情報を扱う場合に求められることは，迅速なデータの処理とその解析結果に基づく各種疾患への迅速な対処法を構築する基礎となる情報提供である。これらの一連の処理を有効に行うためには，中心的となる研究機関でのシステムとしての運用を考えなければならない。そこで，本研究ではデータ解析手法の構築のみならず，構築した理論の実装についての検討を行い，実験として解析システムの構築を試みている。

システムとしての運用を検討する手始めに，システムの利用者と利用環境を想定する。本研究の目的はＯ－１５７等の集団食中毒の発生時に，行政が発生情報とそれに関する様々な付属情報を公開可能とするための基礎的研究である。そこで，国立感染症研究所のような健康危機に関するデータが全国の保健所から集まる中心的な研究機関において，その研究機関内（イントラネット）での利用に限定した。研究機関内の各部局においては，集約された情報を様々な観点から解析することによって有効な情報が得られ，それを様々な観点から検討し合うことによって精度の高い２次情報として利用できる。また，ここで得られた情報は現存の公開システム（厚生省のホームページ，各研究機関のホームページ等）を用いて，広く情報提供可能である。

現実の問題として，ここでの制約（情報が集約される研究機関内でのイントラネットに限る）は，本章で述べた統計解析理論の構築に関する研究の実装のみを考える場合においては，必ずしも必要なものではない。しかし，この制約は第４章で検討している可視化システムとの連携を考えた場合，現状の遅いインターネットの環境での遠隔地からの運用を検討した結果において必要である。

２　　健康危機関連情報処理システムの構成

初めに第２章で示した統計手法をシステムとして実装するために用いたハードウェアとソフトウェアについてまとめる。

ハードウェアとしては，特にＣＰＵには高性能な仕様は必要としない。今回の実験で用いたサーバは，Ｐｅｎｔｉｕｍ ＩＩプロセッサーでクロックが４００ＭＨｚであった。ただし，メモリーはソフトウェアの制約で２５６Ｍｂｙｔｅ以上必要であり，今回は，５１２Ｍｂｙｔｅを搭載した。また，ハードディスクは９Ｇｂｙｔｅであり，イン

トラネット等のネットワーク上での運用を前提としているので，イーサネットカートが必要てある。

表 3　1　システム構成

ハードウェア	
CPU	Pentium II（400MHz）
メインメモリー	512MB
ハートティスク	9Gbyte
ソフトウェア	
OS	Windows NT Server 4 0
コンポーネント	MathSoft社製 StatServer 2000 + S-Plus2000

以上の仕様は，現状のパーソナルコンピュータでは，普及しているものに相当する。特に，ハードウェアに関しては，現時点では安価なものになっている。

ソフトウェアに関しては，統計解析システム構築用にMathSoft社のStat Server2000 + SとPlus2000を用いた。その理由は統計解析プロクラム言語としてS言語の自由度が高く，グラフィック環境も整備されている点がある。また，インターネット上ての公開を考えた場合，以上の組み合わせがシステムの導入において比較的簡単である点かあげられる。基本OSとしては，Windows NTを採用した。その理由は，StatServerが現時点ではWindows NT上でしか動作しないためである。表3　1に以上の仕様をまとめている。

当然，CGI等をよく理解して場合は，以上のような商用ソフトウェアを用いてシステムを作り上げなくても，もっと安価な方法て同様のシステムか作成可能である。例えば，基本OSをLinuxにし，その上に Web サーハを立ち上げS-Plusの代わりにRを利用し，CGIを駆使すれば，ほとんどソフトに関する経費はかからない。

その他の関連するソフトウェアとして Microsoft 社の Excel か一部必要である。これは，システムとして必要なものてはなく。取り扱うファイル形式として利用している。また，その他の統計解析プログラム（SPSS,SAS 等）て取り扱われているファイル形式も取り扱うことも可能である。

図3　1はネットワーク（イントラネット）としてのシステムの概念図を示したものである。

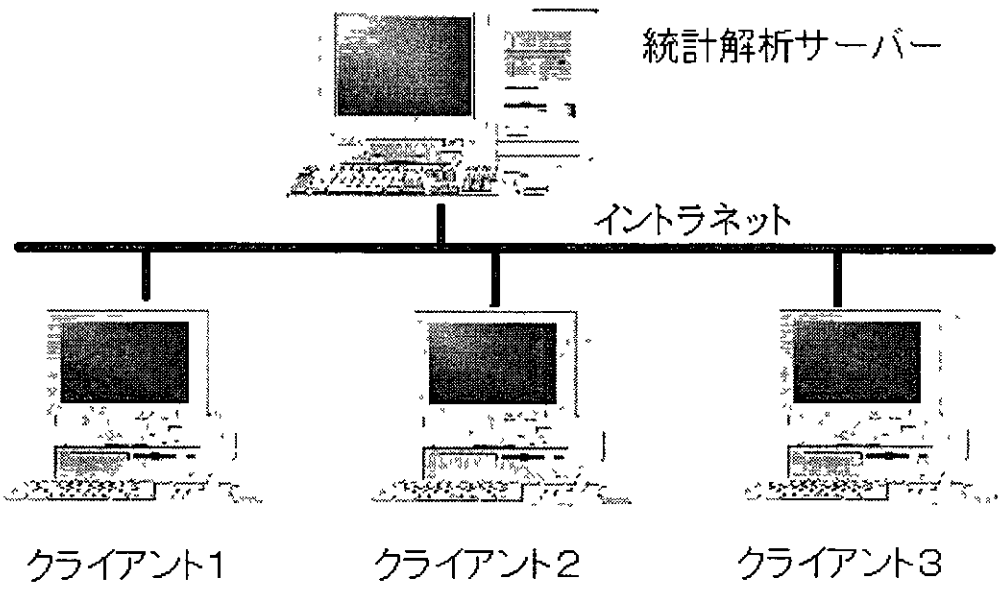

図３．１．ネットワーク

今回の実験システムでは，現時点の汎用性および実際にこのようなシステムを構築し運用する場合を想定したサポート体制の重要性を考え，以上のような組み合わせを用いた。

３　　システムに実装する解析手法

実験システムで実装を検討した統計解析手法は，第２章で述べた以下の３つの手法である。詳細についてはそれを参考にしていただきたい。

（１）　曝露日の予測プログラム
（２）　散在的集団発症に関する検定プログラム
（３）　患者の発症予測プログラム

３　１　　データの仕様

この節では，システムで利用可能なデータの仕様を述べる。ここでの制約は特別なものではなく，現在一般的に利用されているソフトウェアの利用を前提にしている。利用データの仕様は，Microsoft 社の Excel のワークシートやカンマ区切りのテキストファイル形式とする。ここの解析プログラムで許可している３つの入力項目の形式を以下の表３　２，表３　３，表３．４にまとめた。

表３．２　入力データの形式A（発症日）

YYYY／MM／DD
１９９９／１０／２１
２０００／３／１２

表３．３　入力データの形式B（単位区間当たりの発症個数）

YYYY／MM／DD	度数
１９９９／１０／２１	２０
１９９９／１０／２２	３０
１９９９／１０／２３	１０

表３．４　入力データの形式C（発症日）

YYYY／MM／DD	YYYY／MM／DD
１９９９／１０／２１	２０００／３／１２
１９９９／１０／２１	２０００／３／１３
１９９９／１０／２３	２０００／３／１４
１９９９／１０／２３	２０００／３／１４
１９９９／１０／２１	

ここで，形式（A，B）は曝露日の予測プログラムおよび患者の発症予測プログラムにおいて有効であり，形式Cは，散在的集団発症に関する検定プログラムで用いる。

４　　システムの動作実例

図３　２は実験として立ち上げた解析サーバのトップページである。ここには，

（１）研究の目的

（２）システムの目的

（３）システムの機能

（４）システムの実装

に関するサブページがあり，各ページには対応する内容と解析システムとしてのプログラムへのリンクが張られている。

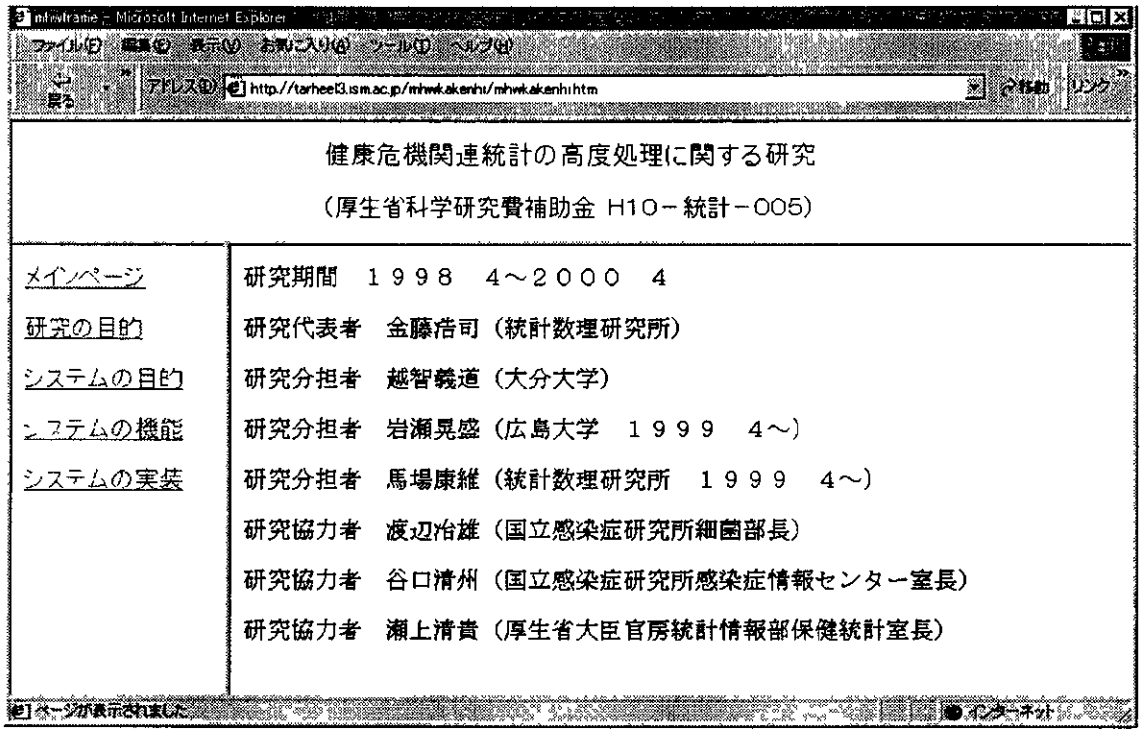

図3．2：トップページ

図3　3は研究の目的を述べているページを表示している。図3．4はシステムの機能について説明したページを表示している。各ページでは，本研究において中心的な役割を果たす統計用語に関する説明にリンクが張られ，各項目をクリックすることによってそれぞれの説明項目のページに移る構造になっている。実験システムでは，2母数の逆ガウス型分布，患者発症パターン等にその説明が加えられている。

このような構造は，利用者（この場合はある程度の統計的知識を有することを仮定している。）がシステムを利用する場合，解析結果についての判断の手助けとなることを期待して作成している。

このような機能が充実することによって，得られたデータをシステムで解析する場合，データが解析の仮定を満足しているかどうかといった問題を利用者が判断する材料となる点も重要である。

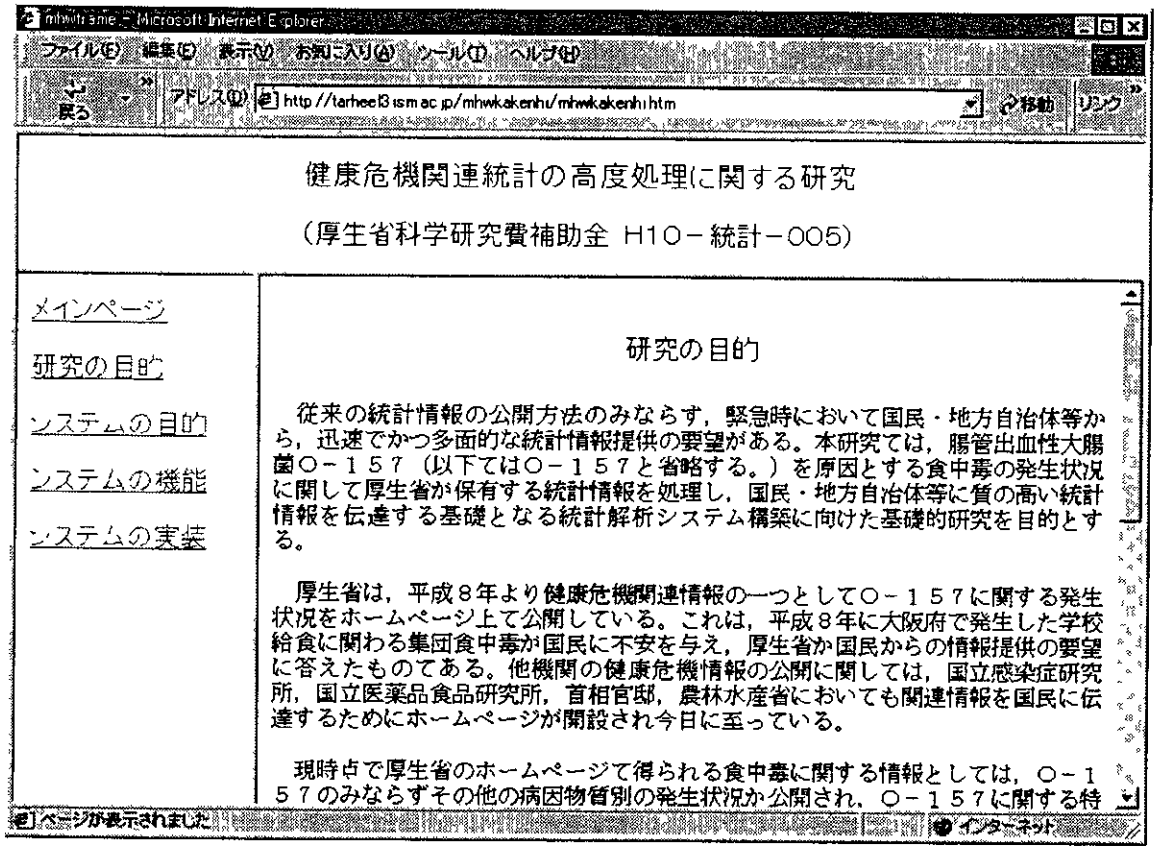

図３．３・　研究の目的

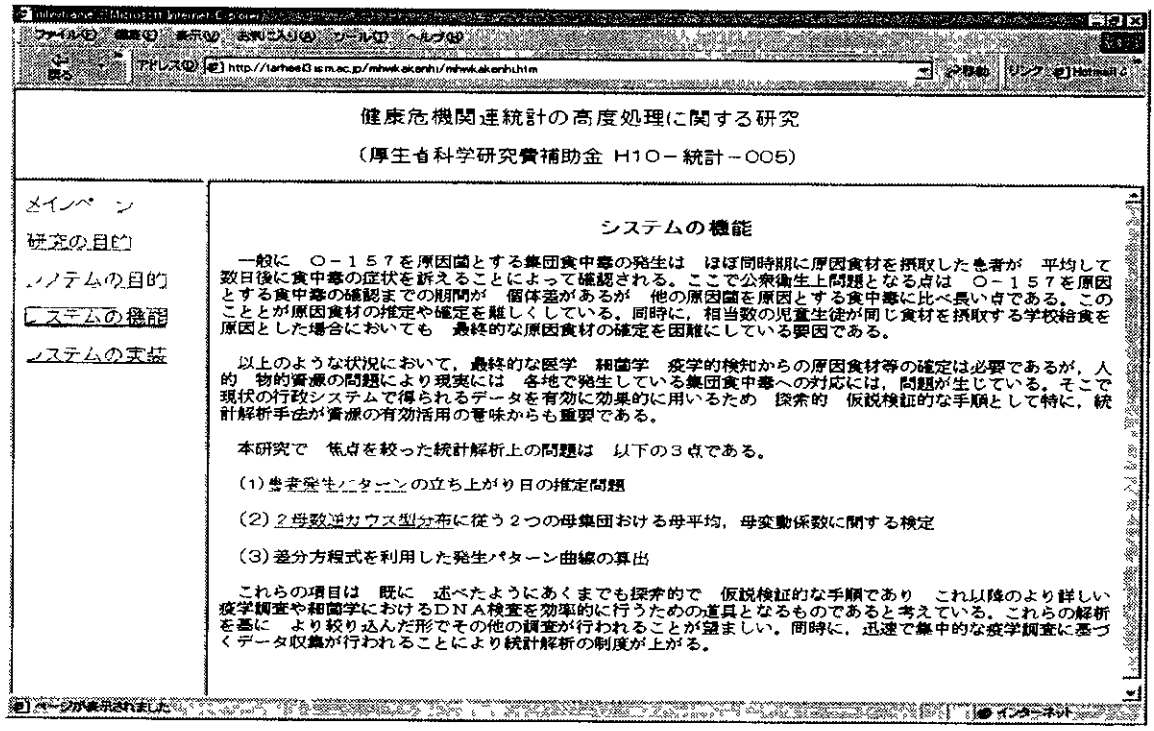

図３．４：システムの機能説明のページ

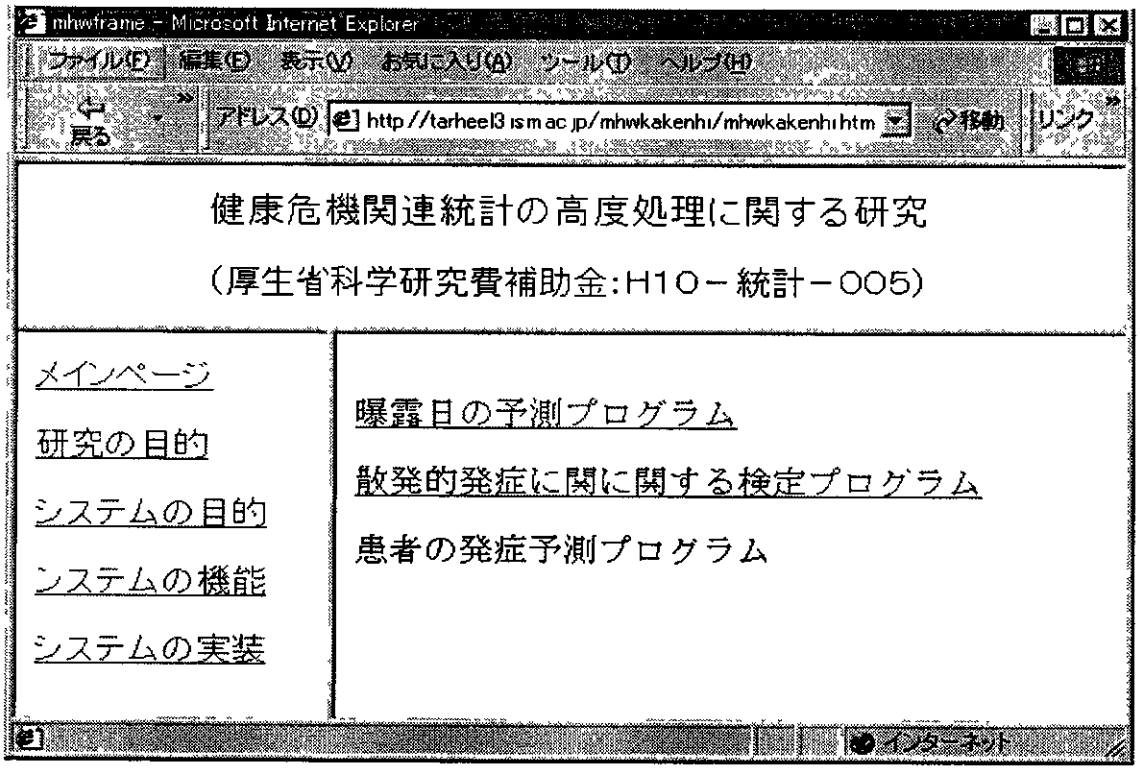

図３．５：解析プログラムのページ

図３　５はシステムの実装を選択した場合に現れる画面である。ここでは，各プログラムを選択することによって解析サーバ上で動作中のＳｔａｔＳｅｒｖｅｒで処理される解析プログラムにリンクが張られている。

これ以降，各プログラムを選択した場合に表れる画面とテストデータを用いたその解析結果の例を示す。

ここで，確認のために解析するファイルについて述べると，原則的には解析対象となるファイルは，既に示した３つの形式（A,B,C）の中で一つの形式として，解析者の手元にあるものとする。原則的にと述べた理由は，データがある場所（ネットワークでつながったデータベースサーバ等に構築されている）にある場合も利用可能である。

解析結果は，テキストのみの形式と，テキストとグラフが表示される場合がある。グラフに関しては，GIF,JPEG,WMF の３つの形式が選択できる。

図３．６は，ある原因食材を摂取したと思われる集団の発症日のデータを，３母数逆ガウス型分布にあてはめ，原因食材摂取日の推定を行った例である。推定される項目は，３つの未知母数とそれから計算される未知原点の推定値である。この未知原点の推定値が原因食材の摂取日に対応する。