川崎医会誌一般教，32号：39—51（2006）

非心ベキ正規分布のパラメータの推定

川崎医科大学　教材教具センター*，情報科学教室**

格和勝利*・近藤芳朗**

（平成18年11月20日受理）

On Estimation of Parameters in Power-Normal Distribution

Katsutoshi KAKUWA* and Yoshiro KONDO**

**Teaching Materials and Facilities Center, Kawasaki Medical School*
***Department of Information Sciences, Kawasaki Medical School,*
577 Matsushima, Kurashiki, Okayama, 701-0192, Japan
(Received on November, 20, 2006)

概　　要

本論文では、ベキ変換の一つであるBox-Cox変換をさらに一般化した「非心ベキ変換」を提唱する。この非心ベキ変換は、集団感染症の暴露日の推定には欠かせないもので、従来の単純なベキ変換はこの場合役に立たない。この研究では、非心ベキ変換を正規分布に適用する非心ベキ正規分布のパラメータについて論理的な推定のアルゴリズムを報告する。キーワード：非心ベキ変換、非心ベキ正規分布、パラメータ推定アルゴリズム、最尤法

Abstract

In this paper, we propose the term “non-central power transformation” for what is generally known as the Box-Cox transformation, which is one of the power transformations. This non-central power transformation is indispensable to making an estimate of the exposure point of an infectious disease. The conventional power transformation does not stand in this case. In this study, we present on a logical estimate algorithm for the parameters of non-central power normal distribution to apply non-central transformation in order to normal distribution. **Key words** : non-central power transformation, non-central power normal distribution, logical estimate algorithm, maximum likelihood estimation

Correspondence to Katsutoshi Kakuwa

1．はじめに

自然現象や社会現象では正規分布をする物理量が多い。正規分布でない場合でも、変換すると正規分布をする物理量もある。このような変換の一つにBox-Coxによる変換[1)]がある。この変換、つまりBox-Cox変換は$x>0$に対して

$$z=\begin{cases}\dfrac{x^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \ln(x) & \lambda=0\end{cases} \tag{1}$$

で定義されている。この変換の提唱者BoxとCoxは、これを用いて変数zは平均値μ、分散σ^2の正規分布$N(\mu,\sigma^2)$に従うものとし、もとの変数xの確率密度関数$g(x)$を

$$g(x)=\frac{x^{\lambda-1}}{\sqrt{2\pi}\sigma}\exp\{-\frac{(z-\mu)^2}{2\sigma^2}\} \tag{2}$$

とした。確率密度関数$g(x)$の形状は、λ,c,μ,σ^2などの大小関係によって表1のように表される。この表は、上坂、後藤[2)]の表1に基づいて作り直したものである。そして、母数μ,σ,λは上記の$g(x)$に基づく最尤法によって求められるとした。しかし、現実の多くの変数xは$x>0$で定義されているもので、それ故Box-Cox変換は意味を持つのであるが、この変換によるzの定義域は$\lambda>0$に対しては$[-1/\lambda,\infty]$、$\lambda<0$に対しては$[-\infty,-1/\lambda]$であって、正規分布で定義されている$[-\infty,\infty]$ではない。つまり、変数zは正規分布するとはいっても打ち切り正規分布をするのである。したがって、厳密にいえば(2)式は誤りである。実際、式(2)で与えられる$g(x)$に対しては

$$\int_0^{\infty} g(x)\,dx<1 \tag{3}$$

を満たし、規格化条件を満たさない。この欠点を改良したのがGoto et al.[3)]である。彼らは(2)の代わりに

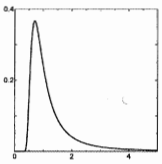
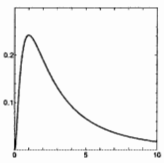
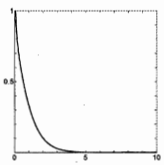
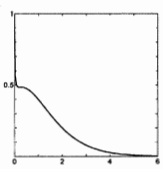
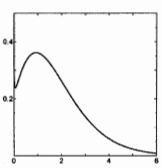
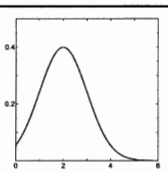
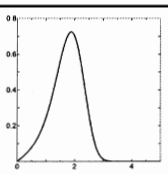
$$g(x)=\frac{x^{\lambda-1}}{\sqrt{2\pi}\sigma A}\exp\{-\frac{(z-\mu)^2}{2\sigma^2}\} \tag{4}$$

とおき、未知定数Aは規格化の条件

$$\int_0^{\infty} g(x)\,dx=1 \tag{5}$$

を満たすように定めた。その結果、$\Phi(x)$を標準正規分布関数として

表１　確率密度関数g(x)のパラメータと形状との関係

パラメータの制約		形　状	特　徴
$\lambda<0$			モーメントの次数 $<\lvert\lambda\rvert$
$\lambda=0$			対数正規分布
$0<\lambda<1$	$\sigma^2>c$		Twisted J-shaped
	$\sigma^2=c$		極大と極小が一致する
	$\sigma^2<c$		極大と極小が存在する
$\lambda=1$			打切られた正規分布
$\lambda>1$			歪度は負

$$A=\begin{cases}\Phi(\dfrac{\lambda\mu+1}{\lambda\sigma}) & \lambda>0\\ 1 & \lambda=0\\ \Phi(-\dfrac{\lambda\mu+1}{\lambda\sigma}) & \lambda<0\end{cases} \tag{6}$$

を得、彼らも母数μ,σ,λを推定するために最尤法を用いこれらを決定するための方程式を導出した。しかし、彼らが臨床検査データからμ,σ,λを推定するために用いたアルゴリズムは、ニュートン・ラフソン法を援用したものであるが必ずしも論理的に明確なものではない。

また、Box-Cox変換にはもう一つの問題点がある。それは、我々が取り扱うデータは必ずしも$x>0$が満たされているとは限らない。一般に$x>c$(cは定数)となっている場合もある。定数cが既知であれば$x-c$を更めてxと置くことにより式(1)が適用できるので、本質的な違いはない。しかしながら、cが未知の場合もある。たとえば、出血性大腸菌O157による集団食中毒の発症分布は、対数正規分布をするといわれているが、分布の原点$x=c$が未知である。この分布の原点$x=c$が暴露日であって、既知の場合もあるが一般的には未知である。そして、この暴露日$x=c$を定めることは疫学的に非常に重要な問題となっている[4),5)]。このような分布の原点$x=c$が未知のデータを取り扱うためには、Box-Cox変換を一般化して

$$z=\begin{cases}\dfrac{(x-c)^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \ln(x-c) & \lambda=0\end{cases} \tag{7}$$

としなければならない。これが筆者達の提唱する変換で「非心ベキ変換」と名づける。

この論文は、筆者達の計画している研究のなかでは理論編ともいうべきもので、最尤法によってμ,σ,λ,cを決定するための方程式を導出し、これらの方程式に基づいて、与えられたデータからμ,σ,λ,cを数値的に決定するための理詰めのアルゴリズムを紹介する。

2．尤度関数

非心正規分布をする確率変数xの確率密度関数は、

$$g(x)=\frac{1}{\sqrt{2\pi}\sigma A}(x-c)^{\lambda-1}\exp\{-\frac{(z-\mu)^2}{2\sigma^2}\} \tag{8}$$

と表される。観測データをx_1、x_2、・・・、x_nとすると、(7)式は

$$z_i=\begin{cases}\dfrac{(x_i-c)^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \ln(x_i-c) & \lambda=0\end{cases} \tag{9}$$

となり、尤度関数（Likelihood Function）Lは、

$$L=\prod_{i=1}^{n}g(x_i)$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma A}\right)^n\prod_{i=1}^{n}\left\{(x_i-c)^{\lambda-1}\exp\left[-\frac{(z_i-\mu)^2}{2\sigma^2}\right]\right\} \tag{10}$$

と表され、これをもとにした対数尤度関数は、

$$\ln L=-n\ln(\sqrt{2\pi}\sigma A)+(\lambda-1)\sum_{i=1}^{n}\ln(x_i-c)-\frac{1}{2\sigma^2 A}\sum_{i=1}^{n}(z_i-\mu)^2$$

$$=-\frac{1}{2\sigma^2 A}\sum_{i=1}^{n}(z_i-\mu)^2+(\lambda-1)\sum_{i=1}^{n}\ln(x_i-c)-\frac{n}{2}\ln\sigma^2-n\ln A-\ln\sqrt{2\pi} \tag{11}$$

となる。ここで

$$k=\frac{\lambda\mu+1}{\lambda\sigma} \tag{12}$$

とすると、Aは次式で与えられる。

$$A=\begin{cases}\Phi(k) & \lambda>0\\ 1 & \lambda=0\\ \Phi(-k) & \lambda<0\end{cases} \tag{13}$$

3．最尤推定値

対数尤度関数を最大にする最尤推定値$\hat{\mu},\hat{\sigma},\hat{\lambda},\hat{c}$は、

$$\frac{\partial}{\partial\mu}\ln L=0,\quad \frac{\partial}{\partial\sigma}\ln L=0,\quad \frac{\partial}{\partial\lambda}\ln L=0,\quad \frac{\partial}{\partial c}\ln L=0 \tag{14}$$

で与えられる４元の偏微分方程式全てを満たすことで求められる。

1）$\dfrac{\partial}{\partial\mu}\ln L=0$ より

$$\mu=\bar{z}-\frac{\sigma^2}{A}\frac{\partial A}{\partial\mu} \tag{15}$$

が得られる。ここに

$$\bar{z}=\frac{1}{n}\sum_{i=1}^{n}z_i \tag{16}$$

である。

2）$\dfrac{\partial}{\partial\sigma}\ln L=0$ より

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 - \frac{\sigma^3}{A}\frac{\partial A}{\partial \sigma} \tag{17}$$

3） $\frac{\partial}{\partial \lambda}\ln L = 0$ より

$$-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i-\mu)\frac{\partial A}{\partial \lambda} + \sum_{i=1}^{n}\ln(x_i-c) - n\frac{1}{A}\frac{\partial A}{\partial \lambda} = 0 \tag{18}$$

4） $\frac{\partial}{\partial c}\ln L = 0$ より

$$-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i-\mu)\frac{\partial z_i}{\partial c} - (\lambda-1)\sum_{i=1}^{n}\frac{1}{x_i-c} = 0 \tag{19}$$

などが与えられる。

4．最尤推定値の数値解法の基礎方程式

まず与えられたc,λのもとでμ,σを求めるアルゴリズムについて考える。

$$\frac{\partial A}{\partial \mu} = \begin{cases} \phi(k)\dfrac{\partial k}{\partial \mu} & \lambda > 0 \\ 0 & \lambda = 0 \\ -\phi(-k)\dfrac{\partial k}{\partial \mu} & \lambda < 0 \end{cases} \tag{20}$$

$$\frac{\partial k}{\partial \mu} = \frac{1}{\sigma},\quad \frac{\partial k}{\partial \sigma} = -\frac{k}{\sigma},\quad \frac{\partial k}{\partial \lambda} = -\frac{1}{\lambda^2\sigma} \tag{21}$$

ここに$\phi(x)$は、標準正規密度関数

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2) \tag{22}$$

である。ここで符号関数$\mathrm{sgn}(\lambda)$

$$\mathrm{sgn}(\lambda) = \begin{cases} 1 & \lambda > 0 \\ 0 & \lambda = 0 \\ -1 & \lambda < 0 \end{cases} \tag{23}$$

を用いると、

$$\frac{\partial A}{\partial \mu} = \frac{\mathrm{sgn}(\lambda)\phi(\mathrm{sgn}(\lambda)k)}{\sigma} \tag{24}$$

である。同様にして

$$\frac{\partial A}{\partial \sigma} = -\frac{\mathrm{sgn}(\lambda)\phi(\mathrm{sgn}(\lambda)k)\,k}{\sigma} \tag{25}$$

これによって(15)式は

$$\mu = \bar{z} - \frac{\sigma^2}{A}\frac{\mathrm{sgn}(\lambda)\phi(\mathrm{sgn}(\lambda)k)}{\sigma} \tag{26}$$

$$= \bar{z} - \sigma q \tag{27}$$

となる。ここに、

$$q = \frac{\mathrm{sgn}(\lambda)\phi(\mathrm{sgn}(\lambda)k)}{A} \tag{28}$$

である。

また、(17)式は、(27)式を用いると

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 - \frac{\sigma^3}{A}\left(-\frac{\mathrm{sgn}(\lambda)\phi(\mathrm{sgn}(\lambda)k)\,k}{\sigma}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 + k\sigma^2 q\\ &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 + \sigma^2 q^2 + k\sigma^2 q\\ &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 + \sigma^2 q(k+q)\end{aligned} \tag{29}$$

に帰着される。式(12)と式(27)から μ を消去すると

$$\sigma = \frac{\bar{z}+\dfrac{1}{\lambda}}{k+q} \tag{30}$$

が得られ、これを式(29)に代入すると、

$$\left(\frac{\bar{z}+\dfrac{1}{\lambda}}{k+q}\right)^2 = S^2 + \frac{q}{k+q}\left(\bar{z}+\frac{1}{\lambda}\right)^2 \tag{31}$$

が最終的に得られる。ここに

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 \tag{32}$$

である。方程式(31)の未知数は c,λ が既知であるとすると k だけである。というのは、式(13)から A は k の関数であり、式(28)から q も k の関数となり、$z_i,\bar{z}$ は c,λ が既知であるから k に対しては定数である、からである。したがって、方程式(31)から原理的に k が定まる。k が定まると式(30)から σ が定まり、式(27)から μ が定まる。こうして、c,λ が既知であると式(31)から k,σ,μ が決まるのである。

次に λ は式(18)から原理的に決定され、最後に c は式(19)から原理的に決定される。

5．推定値を求めるための具体的なアルゴリズム：2分割法（2分法）

最尤推定値を求めるための方程式は一般に連立の超越方程式となる。これを解く従来の方法は、ニュートン・ラフソン法である。この方法は初期値を解の近傍に選ばなければ収束しないし、また解が二つある場合などには適さない。我々の用いる2分割法は、探索する区間を2分割し、区間を拡大・縮小する方法であって、解がいくつあっても確実に求め得る方法である。その概要を説明する。

方程式(31)はkを決めるための方程式であるが、関数$F(c,\lambda,k)$（以下単に$F(k)$と略す）を

$$F(c,\lambda,k)=\left\{\frac{\bar{z}+1/\lambda}{k+q}\right\}^2-\left\{S^2+\frac{q}{k+q}\left(\bar{z}+\frac{1}{\lambda}\right)^2\right\} \tag{33}$$

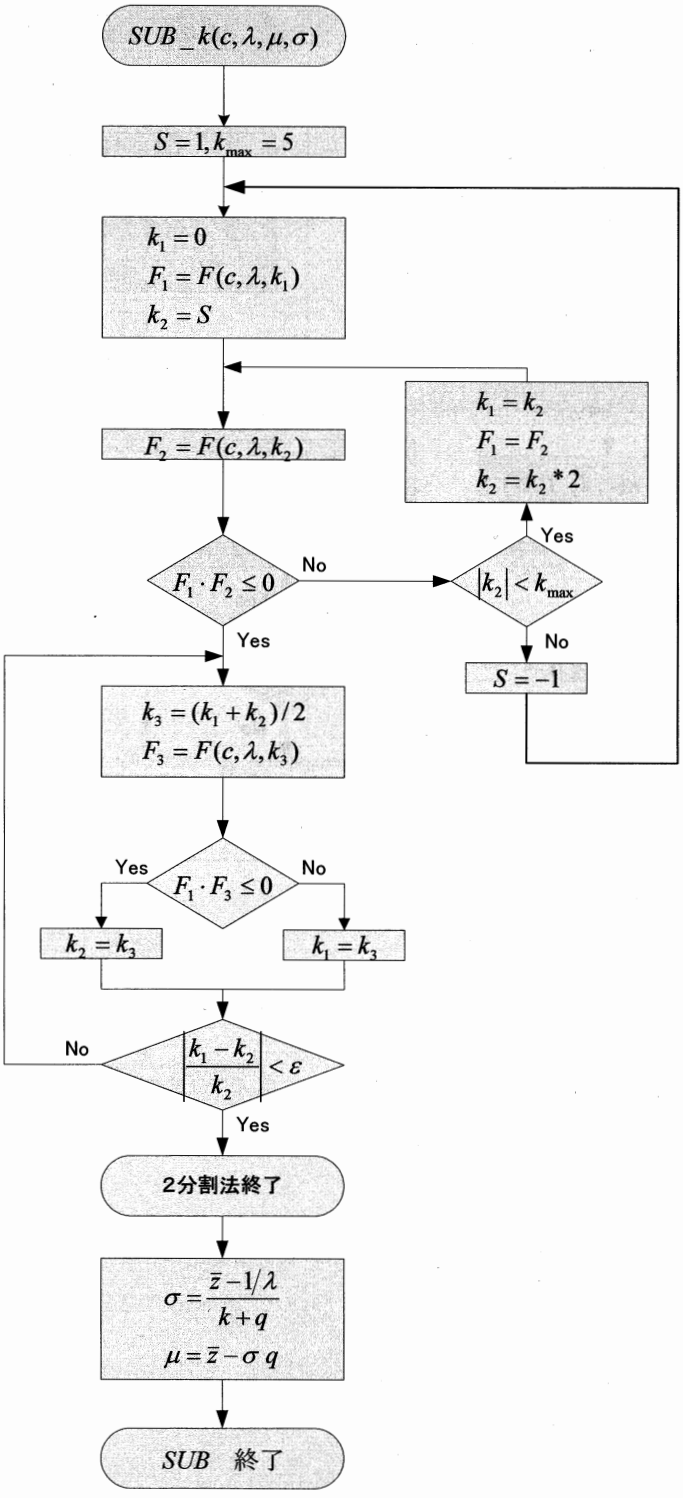
と定義すると、求めるkは$F(k)$の値を0とならしめるものである。いま、$\lambda\geq 0$として、kの決定法について説明する。まず、$k\geq 0$の範囲の解を探索する。$k=2^n$ $(n=-\infty,0,1,2\cdots)$として$k<k_{\max}$の範囲で$F(k)$の符号を調べる。$F(k)$の符号が変わるとき、すなわち$F(2^n)F(2^{n+1})\leq 0$が成り立つとき、$F(k)=0$の解は$2^n\leq k\leq 2^{n+1}$の範囲に存在することがわかる。次にこの範囲を2分し、どちらの区間に存在するかを調べていく。この2分法を求める精度まで繰り返していくと$F(k)=0$の解が定まる。これが2分割法である。もし、$k<k_{\max}$までに解が存在しない場合は、$-k_{\max}\leq k\leq 0$の範囲で2分割法によって解の存在を調べ求める。

フローチャート1を以下に示す。$\lambda<0$の場合も同様で$k<0$の範囲で解を探索する。このような2分割法をλに対しても、cに対しても行う。しかし、解が2つ以上ある可能性も考慮してcについては等間隔に探索区間を広げていき、解が見つかっても一定の範囲まで探索する。以上を総合したフローチャートをフローチャート2，3に示している。このフローチャートに現れる関数G,Hは次式で定義される。

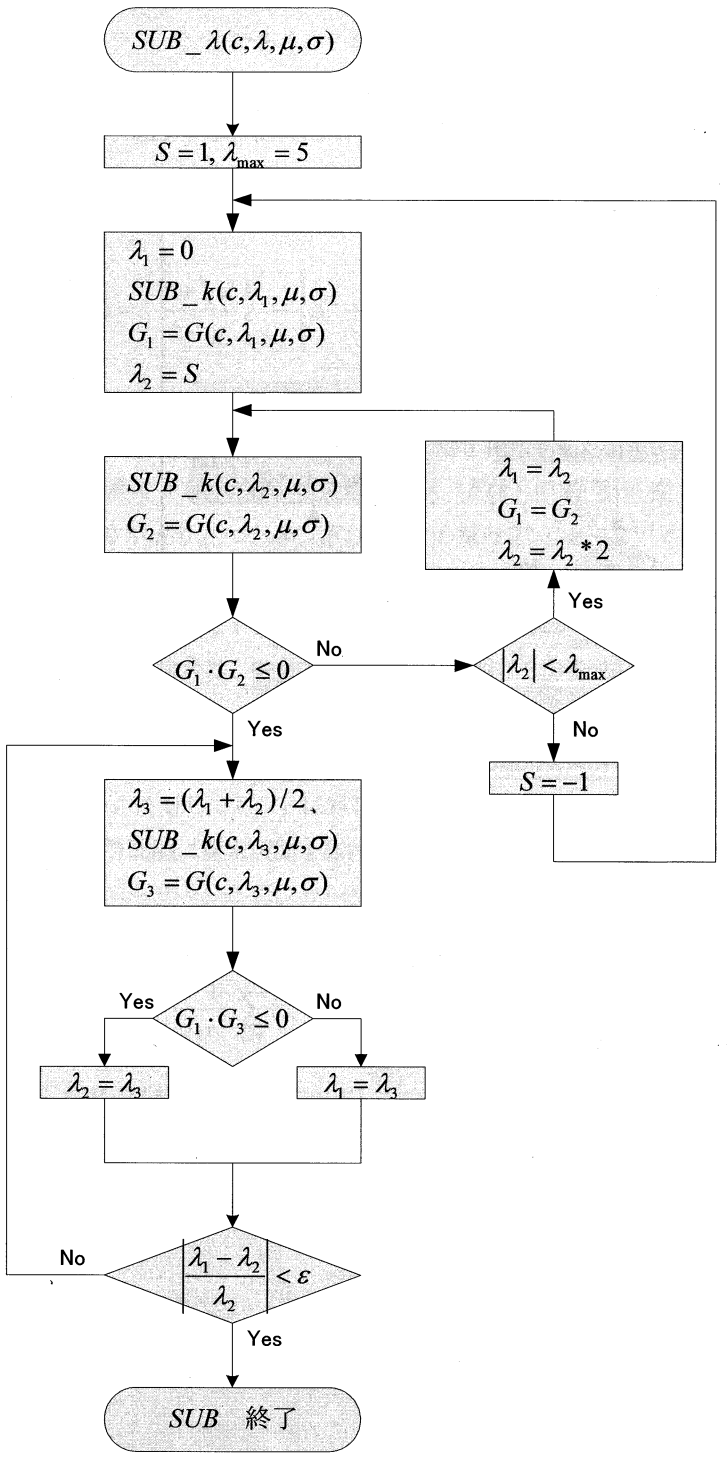
$$G(c,\lambda,\mu,\sigma)=-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i-\mu)\frac{\partial A}{\partial\lambda}+\sum_{i=1}^{n}\ln(x_i-c)-n\frac{1}{A}\frac{\partial A}{\partial\lambda} \tag{34}$$

$$H(c,\lambda,\mu,\sigma)=-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i-\mu)\frac{\partial z_i}{\partial c}-(\lambda-1)\sum_{i=1}^{n}\frac{1}{x_i-c} \tag{35}$$

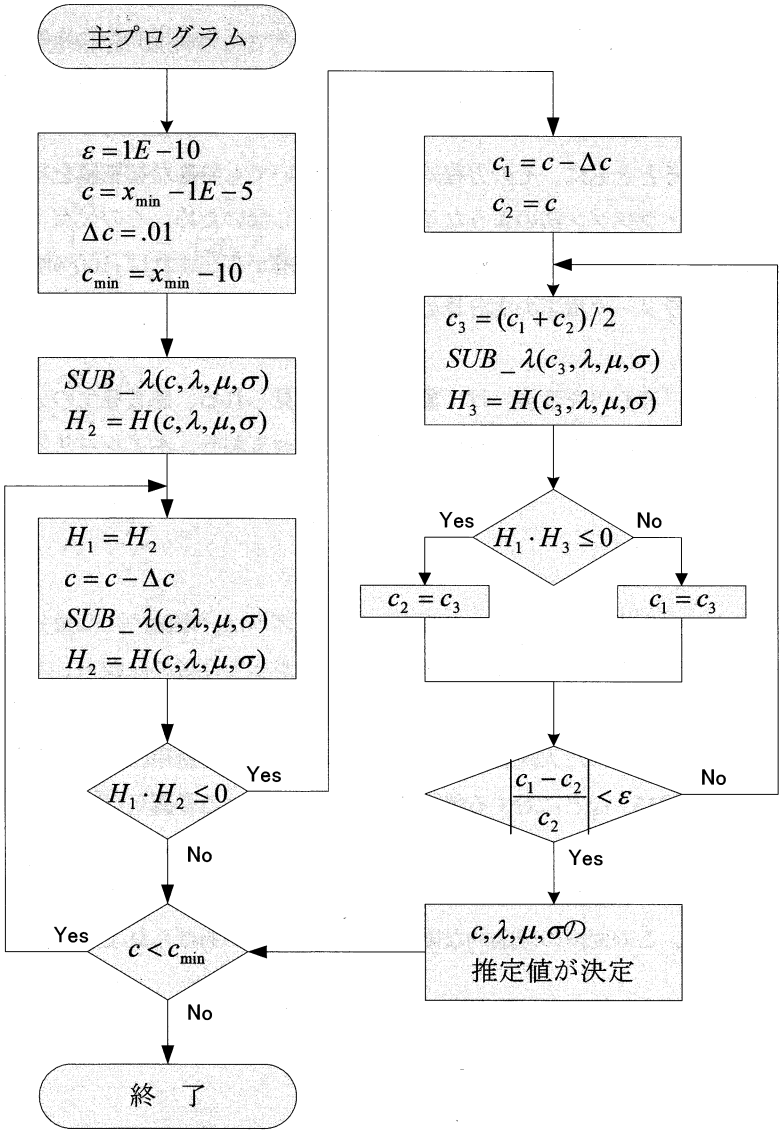
フローチャート1　2分割法



フローチャート2　λを求める副プログラム



フローチャート3　主プログラム



6．まとめと考察

本論文では、Box-Cox 変換を一般化した「非心ベキ変換」を提唱し、その変換パラメータ推定のアルゴリズムについて検討した。その結果、個々のパラメータを推定するための論理的な方程式の導出ができたことにより、フローチャート１～３に示すような、理詰めで簡潔なアルゴリズムを開発することができた。このアルゴリズムは、パラメータ変数に対して順を追って算出することが可能であるとともに、その方程式の解法においても特殊な初期値を与えて解を収束させる、ニュートン・ラフソン法のような手法も必要としないため、その欠点（初期値の与え方によっては解が収束しない）も存在しない。むしろ、解が存在すれば 100%解が得られる２分割法の採用は、パラメータ推定の実計算の部分においてアルゴリズムをより論理的にできたといえる。

また、我々の提唱する「非心ベキ変換」は、集団食中毒に見られる、患者発生の分布の原点と暴露日が異なっている場合にも対応できる形で一般化を図っており、本アルゴリズムは疫学的にも非常に有効であるといる。

7．今後の課題

今回報告した「非心ベキ変換」のパラメータ推定アルゴリズムは、理論的には微分方程式の展開や各パラメータの推定式の導出には成功している。したがって今後の課題としては、実データへの適用があげられる。臨床検査の生化学検査データへの適用による標準化への寄与や、これまで単純平均で検討されてきた入院患者の平均在院日数に対する適用。さらに疫学分野においては、出血性大腸菌 O157 などに対する集団食中毒事件における暴露日の推定への適用などがあげられる。

また数学的には、パラメータが既知の場合に対して人工的にデータを作成することによって変換精度を求めることと、この変換の理論的な限界を定めることがあげられよう。

参 考 文 献

1）Box, G.E.P. and Cox,D.R.:An analysis of transformations,J.Roy. Statist. Soc.,B26: 211-243,1964

2）上坂浩之,後藤昌司:ベキ変換に基づく臨床検査データの解析,応用統計学,9:23-33,1980

3）Goto,M.,Inoue,T. & Tsuchiya,Y.:On estimation of parameters ın power-normal-distribution,Bulletin of Informatics and Cybernetics,21:41-53,1984

4）格和勝利,近藤芳朗: 感染症の平均潜伏期の計算法について, 川崎医療福祉学会誌,6:381-387,1996

5）格和勝利,近藤芳朗: 感染症の平均潜伏期の計算法についてⅡ, 川崎医療福祉学会誌,7:199-203,1997

付　録

A.標準正規分布関数$\Phi(x)$の算出方法

$$erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \quad \text{とすると}$$

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_{-\infty}^x e^{-\frac{t^2}{2}}dt$$

$$=\frac{1}{2}+\frac{1}{2}\cdot\frac{2}{\sqrt{\pi}}\int_0^{\frac{x}{\sqrt{2}}} e^{-t^2}dt$$

$$=\frac{1}{2}+\frac{1}{2}\cdot erf\left(\frac{x}{\sqrt{2}}\right)$$

ここに$erf(x)$は、

$|x|\le 2$において

$$erf(x)=\sum_{n=0}^{\infty} a_n$$

$$a_0=\frac{2x}{\sqrt{\pi}}e^{-x^2}$$

$$a_n=\frac{x^2}{n+1/2}a_{n-1}$$

$|x|>2$において

$$erf(x)=1-\frac{e^{-x^2}}{\sqrt{\pi}}\left(\frac{1}{\lfloor x}+\frac{1/2}{\lfloor x}+\frac{1}{\lfloor x}+\frac{3/2}{\lfloor x}+\cdots\right)$$

を用いて計算する。