統計数理 第34巻 第2号（1986）

順序統計量とベイズ型平滑化条件に基づく条件付き分布の百分位数の推定

統計数理研究所　柏　木　宣　久
江戸川区医師会　伊　谷　昭　幸

（1986年8月　受付）

1. はじめに

連続な分布関数 $F(x, y)$ 及び密度関数 $f(x, y)$ の存在を仮定しても不都合でないような集団から，大きさ n のサンプル $(X_i, Y_i), 1 \leqq i \leqq n$ が得られているとする．このサンプルから，条件付き分布 $F(y \mid x)$ の特定の複数の百分位数を，同時に推定する問題について考える．但し，百分位数は x に関して滑らかに変化する，という事前情報が与えられているとする．

現在，筆者達は小児の発育と肥満という問題に取り組んでいる．小児の発育と肥満の実態を把握する作業の一環として，身長毎にみた体重の百分位数を使った形態発育評価を行う必要が生じ，上記の問題を考えるに至った．第4節で形態発育評価に対する適用例について述べる．

さて，百分位数の推定問題であるが，この問題が密度関数あるいは分布関数の推定によっても解決されてしまうのはいうまでもない．密度関数あるいは分布関数の推定法には，パラメトリックな方法とノンパラメトリックな方法の二種類があるが，形態発育評価といった記述統計に近い問題では，実行可能性が問題にならないかぎり，ノンパラメトリックな方法が好まれる．ノンパラメトリックな方法については Wertz and Schneider (1979), Prakasa Rao (1983) 等が詳しい．従来，ヒストグラムの区間の幅や kernel の window の幅といったノンパラメトリックモデルのパラメータを，主観によって決めていたために，ノンパラメトリックな方法は，特に小サンプルの場合，便宜的な方法にすぎないと思われてきた．しかし，最近，ノンパラメトリックモデルのパラメータをデータに基づき選択しようとする Burman (1985) 等の試みがあり，ノンパラメトリックな方法が必ずしも便宜的とはいえない情勢となりつつある．

しかしながら，任意の百分位数を求めるためにならともかく，特定の百分位数にのみ興味がある場合に，百分位数の推定問題を密度関数あるいは分布関数の推定によって解決しようとするのは，必ずしも適切な方法のようには思われない．密度関数あるいは分布関数の推定では，美的感覚を満足するために，しばしば関数の形状を滑らかにしようとする．関数が定義されている全区間において関数の形状を滑らかにしようとするのは，特定の百分位数にのみ興味がある場合には，不必要な努力である．また，従来のノンパラメトリックな密度関数あるいは分布関数の推定法に，百分位数の位置に関する事前情報を組み込んで，実行容易な方法を構成できるとも思われない．

本稿では，百分位数は x に関して滑らかに変化するという事前情報が与えられている場合に，特定の複数の百分位数を，順序統計量の漸近的な性質とベイズ型平滑化条件を使って，同時に推定する方法を提案する．本稿の方法の特徴は，以下の通りである．① データそのものの分布は仮定していない．その意味ではノンパラメトリックな方法に近い．② モデルのパラメー

タをデータに基づいて選択している．③ 求めるべき百分位数だけをモデリングの対象にしている．④ 階差制約を使って平滑化している．従って，多項式による平滑化の際に問題になる，関数の癖の影響がほとんどない．⑤ 百分位数の位置に関する事前情報を組み込んでいるにもかかわらず，実行が容易である．必要な算法は，ハウスホルダー変換程度の基本的算法だけである．

まず次節では，推定に使用するモデルを提案する．モデルは二つのサブモデルから成っている．一つはデータに関するモデルであり，もう一つは事前情報に関するモデルである．データに関するモデルとしては，Mosteller (1946) の順序統計量に関する漸近的な結果を参考に，容易に同定可能な実用的なモデルを仮定する．そして，百分位数の滑らかな変化という事前情報を，2階差モデルにより定式化する．第3節では，これらのモデルに対応する尤度関数を記述し，百分位数をベイズ推定するための事後尤度と，ベイズ型モデルを選択するためのABICを示し，簡単に推定手続きについて述べる．そして第4節では，本稿の方法を開発する動機を与えた適用例について述べる．

2. モ デ ル

本節では，推定に使用するモデルについて述べる．

何の仮定も設定せずに，有限の大きさのサンプルから，連続な x の各点における条件付き分布の百分位数を推定するのは困難である．そこで，x をある幅を持つ区間に分割し，各区間内の条件付き分布は一様であると仮定し，この仮定の下で百分位数を推定する問題を考える．実際，観測値は有限桁の数値で与えられるのが普通であるから，この考え方は自然であろう．そこで今，例えば整数に注目し，サンプルを次のように分割したとする．

$$C_t=\{Y_i \mid t-0.5\leqq X_i<t+0.5,\ 1\leqq i\leqq n\},\ t\in Z.$$

但し，Z は整数を表わす．サンプルの大きさは有限であるから，実際に t の動く範囲は標本最小値から標本最大値に対応する範囲でよい．それ故，一般性を失うことなく，$1\leqq t\leqq T$ とできる．また，n_t をサブサンプル C_t の大きさとしよう．もちろん，

$$\sum_{t=1}^{T} n_t=n,$$

である．こうする時，各区間 $[t-0.5,\ t+0.5)$ 内での条件付き分布は一様であると仮定し，$1\leqq t\leqq T$ なる t について，サブサンプル C_t から，対応する条件付き分布の百分位数を推定する．

さて，$Y_{t(j)},\ 1\leqq j\leqq n_t$ を，$C_t,\ 1\leqq t\leqq T$ の要素についての順序統計量としよう．$(X_i,\ Y_i),\ 1\leqq i\leqq n$ がランダムサンプルである場合の位数 $Q_{tk}=Y_{t(n_{tk})},\ 1\leqq k\leqq m$ の同時分布について，Mosteller (1946) により次のような漸近的な結果が与えられている．但し，ここに $n_{tk}=[n_t\cdot\lambda_k]+1,\ 0<\lambda_1<\lambda_2<\cdots<\lambda_m<1$ である．

定理 条件付き密度関数 $f_t(y)=f(y \mid t-0.5\leqq x<t+0.5)$ が，λ_k に対応する母集団位数 ξ_{tk} の近傍で微分可能であり，$f_t(\xi_{tk})\neq 0,\ 1\leqq k\leqq m$ であるならば，$Q_{t1},\ Q_{t2},\ \cdots,\ Q_{tm}$ の同時分布は，漸近的に，平均 $\xi_{t1},\ \xi_{t2},\ \cdots,\ \xi_{tm}$，共分散，

$$\text{(1)}\qquad \operatorname{cov}(Q_{tk},\ Q_{tk'})=\lambda_k(1-\lambda_{k'})/\{n_t\cdot f_t(\xi_{tk})\cdot f_t(\xi_{tk'})\},\ k\leqq k',$$

の m 変量正規分布に従う．

本稿では，位数 Q_{tk} について，次のようなモデルを仮定する．

$$Q_{tk}=\xi_{tk}+U_{tk},\ 1\leqq t\leqq T,\ 1\leqq k\leqq m. \tag{2}$$

ここに，U_{tk} は平均 0 で，t に関しては独立に，そして k に関してはある共分散構造をもって変動している正規変量とする．Mosteller の結果を最大限利用しようとするならば，U_{tk} の共分散としては，(1) の構造を仮定すべきであろう．しかしながら，一般に $f_t(y)$ は未知である．そこで本稿では，U_{tk} の共分散を次のようにモデリングする．

$$\begin{aligned}&\mathrm{cov}\,(U_{tk},\ U_{tk'})=\alpha\cdot\lambda_k(1-\lambda_{k'})/\{n_t\cdot\phi(\zeta_k)\cdot\phi(\zeta_{k'})\},\\&\zeta_k=\Phi^{-1}(\lambda_k),\ 1\leqq t\leqq T,\ 1\leqq k,\ k'\leqq m,\ k\leqq k'.\end{aligned} \tag{3}$$

但し，α は未知乗数であり，$\phi(\cdot)$ はある分布の確率密度関数であり，$\Phi^{-1}(\cdot)$ はその分布関数の逆関数である．未知乗数 α は，U_{tk} の分布のスケールを，データに基づき調整するために導入した．$\phi(\cdot)$ としては，先験情報に基づき，適当と思われる分布の確率密度関数を仮定する．いくつかの分布の確率密度関数を仮定し，その内から当てはまりのよい分布を選択しようとする場合には，次節で述べる ABIC を基準に判断する．

Mosteller の結果を最大限利用するために，密度関数の推定と百分位数の推定を組合わせる方法が考えられる．そうした方法は，本来の条件付き密度関数と，仮定した密度関数が極端に異なる場合，特に有効であろう．しかしながら，本稿ではそこまでの精密化は試みない．

サブサンプルの大きさ n_t が，特定の t において，極端に小さくなってしまう場合が想定される．そうした場合の処置として，4 つの方法が考えられる．第 1 が，特に対処せず，(3) 式中にある n_t の効果に期待する方法，第 2 が，そうしたサブサンプルを捨ててしまう方法，第 3 が，区間の分割をやり直すか，あるいは極端に小さくなってしまうのが $t=1$，T といった場合には，隣り合う区間のサブサンプルと統合してしまう方法，そして第 4 が，モデルを変更してしまう方法である．第 1 から第 3 の方法は，実行が容易である．試した後に，どの方法を採用するかを判断するのがよいであろう．第 4 の方法については，今後その必要が生じた段階で検討する．

さて，今まで x を固定して考えてきたわけであるが，データを実際に取り扱う現場からの要請として，位数は x 方向について滑らかであってほしいという希望，あるいは，そうあるべきであるという信念が表明されることがある．そうしたデータのひとつが第 4 節で紹介される．本稿では，そうした要請を，次のような 2 階差モデルにより定式化する．

$$\xi_{tk}=2\xi_{t-1,k}-\xi_{t-2,k}+V_{tk},\ 3\leqq t\leqq T,\ 1\leqq k\leqq m. \tag{4}$$

ここに，V_{tk} は，互いに独立に，平均 0，分散 β の正規分布に従う確率変量である．

モデル (4) の残差項の分布として，平均 0 の正規分布を仮定したのは，母集団位数の 2 階差が 0 の近辺で変動するよう制約するためと，計算の便利のためであって，母集団位数が $\pm\infty$ の範囲で自由に変動すると積極的に表明するためではない．とはいえ，厳密にみた場合，

$$\xi_{t1}\leqq\xi_{t2}\leqq\cdots\leqq\xi_{tm}, \tag{5}$$

という制約条件が満足されない確率は 0 ではなくなり，一見不合理のようにも思われる．しかしながら，この程度のことは，近似的接近法により引き起こされた小さな不合理として，一般に許容されているのではないだろうか．例えば，$\pm\infty$ の身長の人間がいないという理由だけで，身長の分布に正規分布を当てはめる仕方に，異議を唱えたりはしないだろう．また，変量としての標本位数に対し，その順序に制約を設けることなく，標本位数の同時分布が漸近的に多変量正規分布に従うという Mosteller の結果は，広く一般に受け入れられているように思われる．

このような状況から，母集団位数のベイズ推定値を求める目的にとって，(5) という制約条件が満足されない確率が 0 でないという分布論的な不整合は，重要でないと判断した．ただし，ベイズ推定値が (5) の制約を侵す場合には，(5) を明示して設定する必要があるが，今回の解析では，そうした事態は生じなかったので，モデル (4) を仮定した．

3. 尤度関数と ABIC

データに関するモデルのパラメータに対し，その事前分布を階差モデルにより与え，事前分布のパラメータをベイズ型情報量基準(ABIC)により選択しようとする方法は，Akaike (1980) 以降，多くの適用例を生み出した．本稿の方法も，そのひとつと言える．本節では，前節で仮定したモデルに対応する尤度関数と ABIC を示し，簡単に推定手続きを説明する．

先ず，データに関するモデル (2) に対応し，尤度関数，

$$g(\boldsymbol{q}\mid\boldsymbol{\xi},\ \alpha)\propto\alpha^{-Tm/2}\mid\Omega\mid^{-1/2}\exp\left\{-\frac{1}{2\alpha}(\boldsymbol{q}-\boldsymbol{\xi})'\Omega^{-1}(\boldsymbol{q}-\boldsymbol{\xi})\right\},$$

が求まる．ここに，$\boldsymbol{q}$, $\boldsymbol{\xi}$ は，各々 q_{tk}, ξ_{tk}, $1\leqq t\leqq T$, $1\leqq k\leqq m$ を要素とするベクトルであり，Ω は，その要素が (3) 式で表わされる分散共分散行列を，α で除した行列を表わす．一方，母集団位数に関するモデル (4) に対応して，尤度関数，

$$w(\boldsymbol{\xi}_p\mid\boldsymbol{\xi}_i,\ \beta)\propto\beta^{-(T-2)m/2}\exp\left\{-\frac{1}{2\beta}(D\boldsymbol{\xi})'(D\boldsymbol{\xi})\right\},$$

$$D=\begin{bmatrix} I_m & -2I_m & I_m & & \\ & I_m & -2I_m & I_m & \\ & & \cdots\cdots & & \\ & & I_m & -2I_m & I_m \end{bmatrix},$$

が求まる．ここに，$\boldsymbol{\xi}_i$ は ξ_{tk}, $1\leqq t\leqq 2$, $1\leqq k\leqq m$ を要素とし，$\boldsymbol{\xi}_p$ は ξ_{tk}, $3\leqq t\leqq T$, $1\leqq k\leqq m$ を要素とするベクトルで，$\boldsymbol{\xi}'=(\boldsymbol{\xi}_i',\ \boldsymbol{\xi}_p')$ であり，I_m は m 階の単位行列である．以上から，ABIC は以下のように求まる．

$$\begin{aligned}\mathrm{ABIC}&=-2\log\int g(\boldsymbol{q}\mid\boldsymbol{\xi},\ \alpha)w(\boldsymbol{\xi}_p\mid\boldsymbol{\xi}_i,\ \beta)d\boldsymbol{\xi}_p\\&=Tm\log\alpha-(T-2)m\log\gamma+\log\mid\Omega\mid+\log\mid\Omega_p^{-1}+\gamma D_p'D_p\mid\\&\quad+\alpha^{-1}\{(\boldsymbol{q}-\tilde{\boldsymbol{\xi}})'\Omega^{-1}(\boldsymbol{q}-\tilde{\boldsymbol{\xi}})+\gamma(D\tilde{\boldsymbol{\xi}})'(D\tilde{\boldsymbol{\xi}})\}+\mathrm{constant}.\end{aligned}$$

ここに，$\gamma=\alpha/\beta$ であり，Ω_p, D_p は，各々 $t\geqq 3$ の母集団位数に対応する Ω, D の部分行列であり，そして $\tilde{\boldsymbol{\xi}}$ は，$\boldsymbol{\xi}$ の要素の内 $\boldsymbol{\xi}_p$ を，次に述べるベイズ推定量 $\hat{\boldsymbol{\xi}}_p$ に置き換えたベクトルを表わす．

推定手続きは，$\boldsymbol{\xi}_p$ の事後尤度に比例した量，

$$h(\boldsymbol{\xi}_p\mid\boldsymbol{q},\ \boldsymbol{\xi}_i,\ \alpha,\ \beta)=g(\boldsymbol{q}\mid\boldsymbol{\xi},\ \alpha)\cdot w(\boldsymbol{\xi}_p\mid\boldsymbol{\xi}_i,\ \beta),$$

を最大にするように $\boldsymbol{\xi}_p$ の推定量 $\hat{\boldsymbol{\xi}}_p$ を求めるベイズ推定と，ABIC を最小にするように $\boldsymbol{\xi}_i$, α, γ の推定量 $\hat{\boldsymbol{\xi}}_i$, $\hat{\alpha}$, $\hat{\gamma}$ を求める最尤推定の，二つの推定手続きから成っている．実際の計算では，$\hat{\boldsymbol{\xi}}_i$, $\hat{\boldsymbol{\xi}}_p$, $\hat{\alpha}$ は，簡単な計算によって，以下のように陽的に表現できるので，この関係を使って値を決めればよい．

$$\hat{\xi}=(\hat{\xi}_i',\ \hat{\xi}_p')'=(\Omega^{-1}+\gamma D'D)^{-1}\Omega^{-1}q,$$
$$\hat{\alpha}=\frac{1}{Tm}\{(q-\hat{\xi})'\Omega^{-1}(q-\hat{\xi})+\gamma(D\hat{\xi})'(D\hat{\xi})\}.$$

しかし，$\hat{\gamma}$ の陽的な表現はできないので，γ については，数値的な方法により最終的な値を決める．具体的には，いくつかの値を $\hat{\gamma}$ の候補として用意し，これらの値に対応する ABIC を計算し，そして，その内で ABIC を最小にする値を選択する．値の候補としては，$2^{k/2}$, $k=0,1,\cdots,20$ を考えておけば十分であろう．また，$\phi(\cdot)$ としていくつかの分布の確率密度関数を仮定する場合には，各分布毎に上記の手続きを実行し，その中で ABIC を最小にする分布を選択する．

4. 適　用　例

本節では，本稿の方法を開発する動機を与えた適用例について述べる．

小児保健の分野に，小児の形態発育評価という問題がある．これは，容易に観測可能な身長，体重といった量から，小児の発育の具合を評価しようという問題である．形態発育評価法としては，きわめて多くの方法が，それぞれの必然性に基づき提案されている．そうした方法の内，日本で最も広く利用されているのは，高石によるパーセンタイル図であろう．（まとまった参考文献としては，高石，樋口，小島（1981）が一般的である．）

一般によく利用されるパーセンタイル図は，母子手帳に掲載されている，年齢で条件付けした身長あるいは体重のパーセンタイル図である．これらの図は，当初は，食糧不足による発育不良を改善するために作成されたと理解している．ところが，現在は飽食の時代と呼ばれるように，単純な食糧不足による発育不良の問題は希薄となり，代わりに，偏食による栄養不良，あるいは過食による栄養過多といった問題が生じ，身長と体重のバランスが問題にされるようになってきた．こうした問題に対しては，身長で条件付けした体重のパーセンタイル図が，ある程度の目安を与えてくれると考えられている．但し，小児の場合，同性同一身長であっても，年齢により体重の分布が違うといった現象があり，男女別各年齢毎の，そうしたパーセンタイル

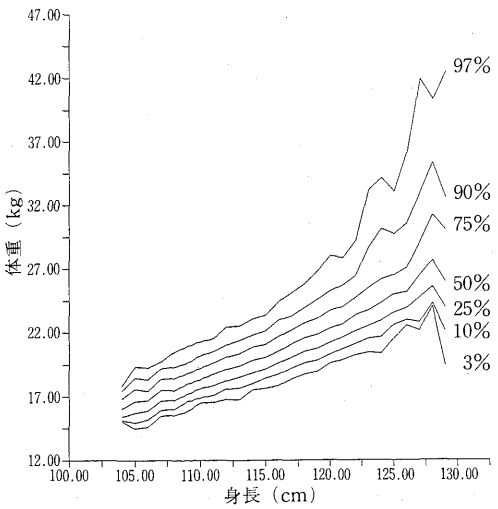

図1．　男児6歳のデータによるパーセンタイル図

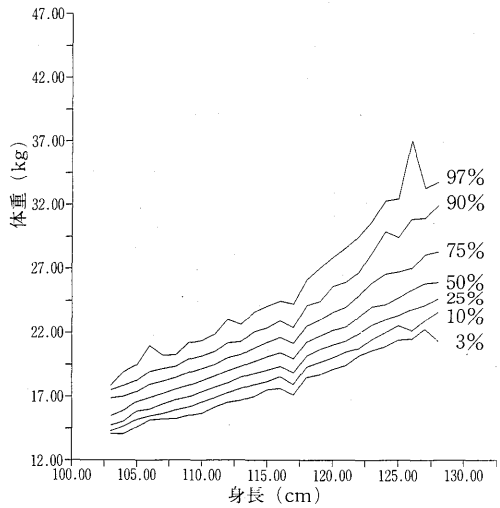

図2．　女児6歳のデータによるパーセンタイル図

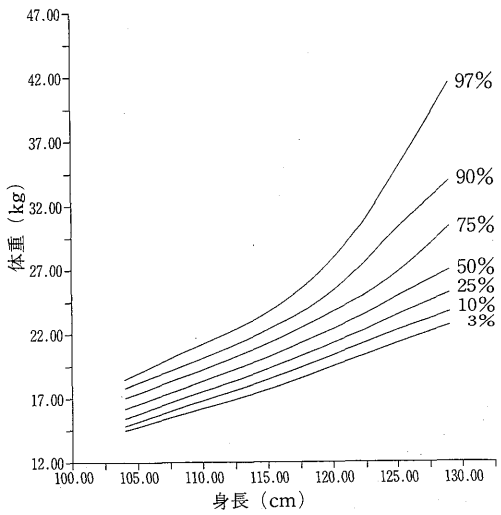

図3. 男児6歳の平滑化されたパーセンタイル図

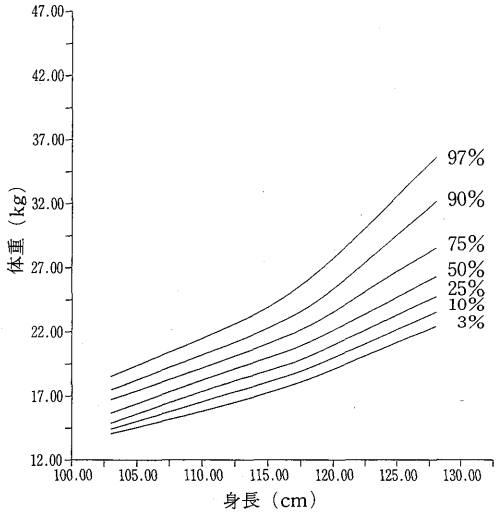

図4. 女児6歳の平滑化されたパーセンタイル図

図が必要となる.

図1および2は,昭和57年度に実施された学校保健統計調査の集計結果から計算された標本位数をもとに作成した,男女6歳児の身長で条件付けした体重のパーセンタイル図である.一般に,位数は身長に関して滑らかに変化するのが望ましいとされているが,この図は,そうした希望を満足していない.細かな変動のみならず,系統的とみられる変動までをも見てとれる.滑らかな位数を求めるために,平滑化の必要が生じた.

図3および4は,前節までに述べた方法を,図1および2に示したデータに適用した結果得られたパーセンタイル図である.望むべく平滑化されている様子を見てとれるであろう.ちなみに,昭和57年度学校保健統計調査報告書には,0.01%単位の比率に変換された集計結果しか掲載されていないため,各サブサンプルの実際の大きさは分からなかったが,この影響は全てαで吸収できるので,本稿の方法を適用するうえで問題は生じなかった.尚,今回の計算では,$\phi(\cdot)$として,① 分散1の正規分布,② 対数変換後に分散1の正規分布になる対数正規分布,③ 対数変換後に分散4の正規分布になる対数正規分布,④ スケール1のコーシー分布,という4種類の分布の密度関数を仮定した.モデルの性質上,ロケーションは任意でよい.正規分布とコーシー分布のスケールはαで表現できるので,これらの分布については,代表的な分布を一つだけ考えた.結局,最終的に選択された分布は,分散1の正規分布であった.また,小さいサブサンプル(0.2%未満)は,隣り合う区間に統合した.これは,そうしたサブサンプルの出現が,身長が極めて低い,あるいは高い場合に限られていたのと,今後の解析の都合上,できるだけデータを捨てたくない,という特別の事情に基づく決定である.

【参考】 学校保健統計調査は,文部省が主体となって実施している指定統計の一つで,調査実施校に指定された学校の長には,申告の義務が果せられている.対象は,全国の幼稚園,小学校,中学校,高校に所属する,5歳以上の児童,生徒で,標本調査形式により,毎年1回,4月～6月頃実施されている.標本の大きさは,昭和57年度の場合,幼稚園児が72,380人,小学生が250,980人,中学生が214,320人,そして高校生が124,080人であった.標本抽出の方法は,学校の抽出については生徒数に基づく確率比例抽出が,また児童,生徒の抽出については

系統抽出が使用されている.

5. おわりに

本稿で述べた百分位数の推定法の応用として，2変数 X, Y の同時確率密度関数 $f(x, y)$ の推定法が考えられる.具体的には,順序統計量の端点に関する漸近的な結果と本稿の方法を使って，条件付き分布 $F(y|x)$ の百分位数を密に求めれば，それはある意味で条件付き確率密度関数 $f(y|x)$ の近似的な推定であるから，後は x の周辺密度関数 $f(x)$ を推定すれば，$f(x, y)=f(y|x)f(x)$ の関係により，同時確率密度関数 $f(x, y)$ の近似的な推定が得られる．この問題については，別稿で詳しく述べる.

本稿では，身長で条件付けした体重のパーセンタイル図を適用例として提示したが，一般によく利用される年齢で条件付けした身長あるいは体重のパーセンタイル図を，本稿の方法によって求められるのはいうまでもない．こうして求めたパーセンタイル図と，他の検査データを使って，更なる解析が著者達によって試みられている.

謝　辞

示唆に富む助言を与えてくださった，安芸重雄博士(統計数理研究所)に感謝いたします．また，計算機プログラムの開発に当って御助力いただきました，坂本光子女史に感謝いたします．査読者の方々の有益な助言に感謝いたします.

参考文献

Akaike, H. (1980). Likelihood and the Bayes procedure, *Trab. Estadist.*, **31**, 143-166.

Burman, P. (1985). A data dependent approach to density estimation, *Zeit. Wahrscheinlichkeitsth.*, **69**, 609-628.

David, H.A. (1970). *Order Statistics,* John Wiley and Sons.

Kashiwagi, N. and Kishino, H. (1984). A Bayesian method for smoothing time-series data collected by sample surveys, *Research Memo,* No. **282**, The Institute of Statistical Mathematics, Tokyo.

文部省 (1982). 昭和57年度学校保健統計調査報告書.

Mosteller, F. (1946). On some useful "inefficient" statistics, *Ann. Math. Statist.*, **17**, 377-408.

Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation,* Academic press.

高石昌弘，樋口　満，小島武次 (1981). からだの発達，大修館書店.

Wertz, W. and Schneider, B. (1979). Statistical density estimation : a bibliography, *International Statist. Review,* **47**, 155-175.

A Bayesian Method for Estimating Percentiles of a Conditional Distribution Using Order Statistics and a Smoothness Prior

Nobuhisa Kashiwagi

(The Institute of Statistical Mathematics)

Teruyuki Itani

(Edogawa-ku Medical Association)

A Bayesian method for estimating percentiles of a conditional distribution $F(y \mid x)$ when a sample (X_i, Y_i), $1 \leqq i \leqq n$ from a continuous distribution with cdf $F(x, y)$ is obtained, is proposed here. The asymptotic result about the joint distribution of quantiles given by Mosteller (1946, *Ann. Math. Statist.*, **17**, 377-408) and the Bayes procedure developed by Akaike (1980, *Trab. Estadist.*, **31**, 143-166) play important roles in the proposed method. An example shows that the proposed method can provide a percentile trace useful for assessing growth of the child.