数理解析研究所講究録
1101 巻 1999 年 43-68

離散指数型分布族における区間予測とその応用

筑波大・理工　飛田 英祐 (Eisuke Hida)
筑波大・数学　赤平 昌文 (Masafumi Akahira)

1. はじめに

統計的推測理論では未知の母数をもつ母集団分布からの標本に基づいて，その母数の推測方式の最適性などについて論じる．それに対して統計的予測論では未観測の確率変数を観測データに基づいて予測する方式を考える ([G70], [Take75], [A90], [Taka96], [BC96]). その際，観測データが従う分布は未知の母数をもつから，そのことも考慮しなければならない.

本論では推測理論の区間推定に対応する区間予測について考察し，現実の問題に適用して数値的に検討し，本論の区間予測の妥当性を確認する.

2. 問題設定

観測データを確率ベクトル $\mathbf{X}=(X_1,\cdots,X_m)$ とし，未観測確率変数を Y，同時分布 P_θ は未知母数 θ に依存するとし，Y のとり得る値全体の空間を $\mathcal{Y}$ とする．ただし，θ は母数空間 Θ の元とする．このとき，任意の α $(0<\alpha<1)$ に対して $\mathbf{X}$ に基づく集合 $S_{\mathbf{X}}(\subset\mathcal{Y})$ をとって

$$P_\theta\{Y\in S_{\mathbf{X}}\}\geq 1-\alpha, \qquad {}^\forall\theta\in\Theta \tag{1}$$

となるとき，$S_{\mathbf{X}}$ を Y の信頼係数 $1-\alpha$ の予測域といい，$\mathcal{Y}\subset\mathbf{R}^1$ で $S_{\mathbf{X}}$ が閉区間 $[a(\mathbf{X}),b(\mathbf{X})]$ になるとき，$S_{\mathbf{X}}$ を Y の信頼度 $1-\alpha$ の予測区間という (図 1 参照)．また，$\mathbf{X}$ が実現値 $\boldsymbol{x}=(x_1,\cdots,x_m)$ をとるとき，区間 $[a(\boldsymbol{x}),b(\boldsymbol{x})]$ を Y の信頼係数 $100(1-\alpha)\%$ 予測区間という．特に，(1) において等号が成り立つとき，予測域 $S_{\mathbf{X}}$ は相似 (similar) であるという.

3. 離散指数型分布族における区間予測

いま，$X_1,\cdots,X_m,Y_1,\cdots,Y_n$ を互いに独立にいずれも確率関数

$$f(x;\theta)=c(\theta)h(x)\exp\{\eta(\theta)t(x)\} \qquad (x=0,1,2,\cdots)$$

をもつ 1 母数離散指数型分布に従う確率変数とする．ただし $\theta\in\Theta=\mathbf{R}^1$ で $c(\theta)$，$h(x)$ は非負値関数，$\eta(\theta)$，$t(x)$ は実数値関数とする．このとき，$X_1,\cdots,X_m,Y_1,\cdots,Y_n$ の同時確率関数は

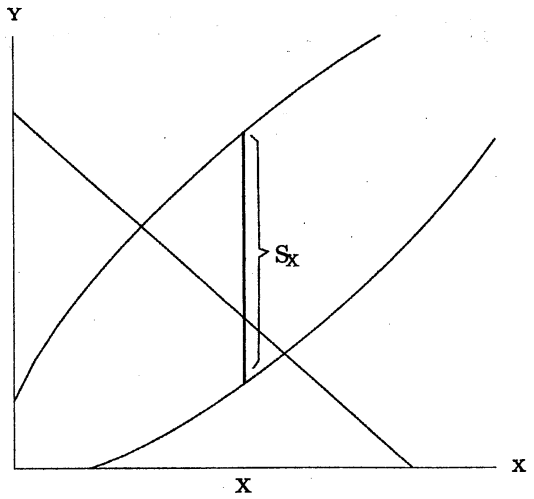

図 1: X に基づく Y の予測区間 S_X

$$f_{X_1,\cdots,X_m,Y_1,\cdots,Y_n}(x_1,\cdots,x_m,y_1,\cdots,y_n;\theta) = c^{m+n}(\theta)\prod_{i=1}^{m}h(x_i)\prod_{j=1}^{n}h(y_j) \cdot\exp\left\{\eta(\theta)\left(\sum_{i=1}^{m}t(x_i)+\sum_{j=1}^{n}t(y_j)\right)\right\}$$

になるから，$T:=\sum_{i=1}^{m}t(X_i)+\sum_{j=1}^{n}t(Y_j)$ は θ に対する完備十分統計量となる．ここで，T が十分統計量であるとは，$T=t$ を与えたときの $(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$ の条件付確率関数が θ に無関係になることをいう．従って，十分統計量 T を与えたときの $Y:=\sum_{j=1}^{n}t(Y_j)$ の条件付分布を利用して，予測区間を未知の母数 θ に無関係に構成できる．実際には，次の手順 (i) ~ (iii) によって Y の予測区間を構成する．

(i) $T=t$ を与えたときの Y の条件付確率関数を $f_{Y|T}(\cdot|t)$ とすれば，これは θ に無関係になり，この確率関数から $T=t$ を与えたときの Y の条件付平均 $\mu_t:=E[Y|T=t]$，条件付分散 $\sigma_t^2:=\mathrm{Var}(Y|T=t)$，条件付 3 次のキュムラント $\kappa_{3,t}:=\kappa_3(Y|T=t)=E[(Y-\mu_t)^3|T=t]$ を求める．

(ii) 任意の α $(0<\alpha<1)$ に対して，各 t について

$$P\{\underline{y}(t)\le Y\le\bar{y}(t)|T=t\}=1-\alpha \tag{2}$$

となる $\underline{y}(t)$，$\bar{y}(t)$ を，(i) で求めた μ_t，σ_t^2，$\kappa_{3,t}$ を用いて (漸近的に) 求める．

(iii) (2) から，任意の $\theta\in\Theta$ について

$$P_\theta\{\underline{y}(T)\le Y\le\bar{y}(T)\}=1-\alpha$$

になり，また完備十分統計量Tが$\sum_{i=1}^{m} t(X_i) + \sum_{j=1}^{n} t(Y_j) = \sum_{i=1}^{m} t(X_i) + Y$であることから

$$P_\theta\{a(\mathbf{X}) \leq Y \leq b(\mathbf{X})\} = 1 - \alpha$$

となる$a(\cdot)$，$b(\cdot)$を(漸近的に)求める．このとき，区間$[a(\mathbf{X}),\ b(\mathbf{X})]$は，$Y$の信頼係数$1-\alpha$の(相似な)予測区間になる．

3.1. 2項分布の場合

観測されるデータを確率変数X，未観測の確率変数をYとし，XとYは互いに独立に，Xは2項分布$B(m,p)$，Yは2項分布$B(n,p)$に従うとする．ただし，m,nは，自然数で既知とし，pは$0<p<1$で未知とする．このとき，Xに基づいてYの区間予測を行なう．まず，X,Yの同時確率関数は

$$f_{X,Y}(x,y;p) = \binom{m}{x}\binom{n}{y} p^{x+y} q^{m+n-(x+y)}$$

$$(x = 0,1,\ldots,m;\ y = 0,1,\ldots,n;\ 0<p<1,\ q = 1-p)$$

となることから，統計量$T := X+Y$はpに対する十分統計量であり，Tは$B(m+n,p)$に従う．このとき，$T=t$を与えたときのYの条件付確率関数は

$$f_{Y|T}(y|t) = \frac{\binom{n}{y}\binom{m}{t-y}}{\binom{m+n}{t}} \qquad (\max(0,t-m) \leq y \leq \min(t,n))$$

になり，これはpに無関係である．このことは，十分統計量Tに基づくYの予測区間が未知の母数pに無関係に構成することができることを意味している．また，この分布は超幾何分布$H(t,n,m+n)$と呼ばれている．また，$T=t$を与えたときのYの条件付平均μ_t，条件付分散σ_t^2，条件付3次キュムラント$\kappa_{3,t}$は次のようになる．

$$\mu_t := E[Y|T=t] = \frac{tn}{m+n},$$

$$\sigma_t^2 := \mathrm{Var}(Y|T=t) = \frac{tmn(m+n-t)}{(m+n)^2(m+n-1)},$$

$$\kappa_{3,t} := \kappa_3(Y|T=t) = \frac{tmn(m-n)(m+n-t)(m+n-2t)}{(m+n)^3(m+n-1)(m+n-2)}.$$

そこで，m,n が十分大きいとき

$$P\{\min(t,n)-y_{\alpha/2}(t)\leq Y\leq y_{\alpha/2}(t)|T=t\}=1-\alpha \tag{3}$$

となるような超幾何分布 $H(t,n,m+n)$ の上側 $100(\alpha/2)\%$ 点 $y_{\alpha/2}(t)$ を漸近的に求める．

まず，Cornish–Fisher 展開によって

$$\frac{y_{\alpha/2}(t)-\mu_t}{\sigma_t}=u_{\alpha/2}+\frac{\kappa_{3,t}}{6\sigma_t^3}u_{\alpha/2}^2+\cdots$$

より

$$\begin{aligned} y_{\alpha/2}(t) &= \mu_t+\sigma_t u_{\alpha/2}+\frac{\kappa_{3,t}}{6\sigma_t^2}u_{\alpha/2}^2+\cdots \\ &= \frac{tn}{m+n}+u_{\alpha/2}\sqrt{t\left(1-\frac{t}{m+n}\right)\frac{mn}{(m+n)(m+n-1)}} \\ &\quad +\frac{m-n}{6(m+n-2)}\left(1-\frac{2t}{m+n}\right)u_{\alpha/2}^2+\cdots \end{aligned} \tag{4}$$

になる．ただし，$u_{\alpha/2}$ は正規分布 $N(0,1)$ の上側 $100(\alpha/2)\%$ 点とする．ここで，(4) において，$y=y_{\alpha/2}(t)$ とおき，$a:=n/(m+n)$，$b:=mn/\{(m+n)(m+n-1)\}$，$c:=(m-n)/(m+n-2)$，$u=u_{\alpha/2}$ とし，$t=x+y$ に注意すれば，(4) から

$$y\fallingdotseq a(x+y)+u\sqrt{(x+y)\left(1-\frac{x+y}{m+n}\right)b}+\frac{c}{6}\left(1-\frac{2(x+y)}{m+n}\right)u^2 \tag{5}$$

になる．そこで，(5) の辺々を 2 乗すると

$$\left[\left\{1-a+\frac{cu^2}{3(m+n)}\right\}y-\left\{a-\frac{cu^2}{3(m+n)}\right\}x-\frac{c}{6}u^2\right]^2\fallingdotseq b(x+y)\left(1-\frac{x+y}{m+n}\right)u^2$$

となる．従って

$$\left\{1-a+\frac{cu^2}{3(m+n)}\right\}^2y^2+\left\{a-\frac{cu^2}{3(m+n)}\right\}^2x^2+\frac{c^2}{36}u^4$$

$$-2\left\{1-a+\frac{cu^2}{3(m+n)}\right\}\left\{a-\frac{cu^2}{3(m+n)}\right\}xy$$

$$+\frac{c}{3}u^2\left\{a-\frac{cu^2}{3(m+n)}\right\}x-\frac{c}{3}u^2\left\{1-a+\frac{cu^2}{3(m+n)}\right\}y$$

$$-b(x+y)u^2+\frac{b(x+y)^2}{m+n}u^2=0$$

となり，これをまとめると

$$\left[\left\{1-a+\frac{cu^2}{3(m+n)}\right\}^2+\frac{bu^2}{m+n}\right]y^2$$

$$-2\left[\left\{1-a+\frac{cu^2}{3(m+n)}\right\}\left\{a-\frac{cu^2}{3(m+n)}\right\}-\frac{bu^2}{m+n}\right]xy$$

$$+\left[\left\{a-\frac{cu^2}{3(m+n)}\right\}^2+\frac{bu^2}{m+n}\right]x^2-\left[\frac{c}{3}u^2\left\{1-a+\frac{cu^2}{3(m+n)}\right\}+bu^2\right]y$$

$$+\left[\frac{c}{3}u^2\left\{a-\frac{cu^2}{3(m+n)}\right\}-bu^2\right]x+\frac{c^2}{36}u^4=0 \qquad (6)$$

ここで，

$$A:=\left\{1-a+\frac{cu^2}{3(m+n)}\right\}^2+\frac{bu^2}{m+n}\ ,$$

$$B:=\left\{1-a+\frac{cu^2}{3(m+n)}\right\}\left\{a-\frac{cu^2}{3(m+n)}\right\}-\frac{bu^2}{m+n}\ ,$$

$$C:=\left\{a-\frac{cu^2}{3(m+n)}\right\}^2+\frac{bu^2}{m+n}\ ,$$

$$2D:=\frac{c}{3}u^2\left\{1-a+\frac{cu^2}{3(m+n)}\right\}+bu^2,$$

$$2E:=\frac{c}{3}u^2\left\{a-\frac{cu^2}{3(m+n)}\right\}-bu^2,$$

$$F := \frac{c^2}{36}u^4$$

とすると，(6) より

$$Ay^2 - 2(Bx + D)y + Cx^2 + 2Ex + F = 0$$

になり，これを y について解けば

$$y = \frac{1}{A}\left\{Bx + D \pm \sqrt{(Bx + D)^2 - A(Cx^2 + 2Ex + F)}\right\}$$

となる．従って，これを用いて (3) から，任意の p $(0 < p < 1)$ について

$$P_p\{a(X) \leq Y \leq b(X)\} \fallingdotseq 1 - \alpha$$

となる Y の予測区間 $[a(X), b(X)]$ を漸近的に得る．ただし，

$$a(X) = \frac{1}{A}\left\{Bx + D - \sqrt{(Bx + D)^2 - A(Cx^2 + 2Ex + F)}\right\},$$

$$b(X) = \frac{1}{A}\left\{Bx + D + \sqrt{(Bx + D)^2 - A(Cx^2 + 2Ex + F)}\right\}$$

とする．また，Y の予測区間を得るための曲線 (Y の予測曲線) $Y = a(X)$，$Y = b(X)$ を図 2, 3 において示す．

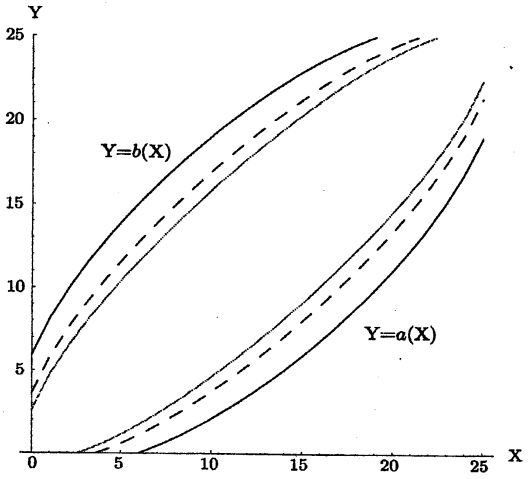

図 2: $m = n = 25$ のときの Y の予測曲線 $Y = a(X), Y = b(X)$

———— 信頼係数99%; − − − − 信頼係数95%; ———— 信頼係数90%

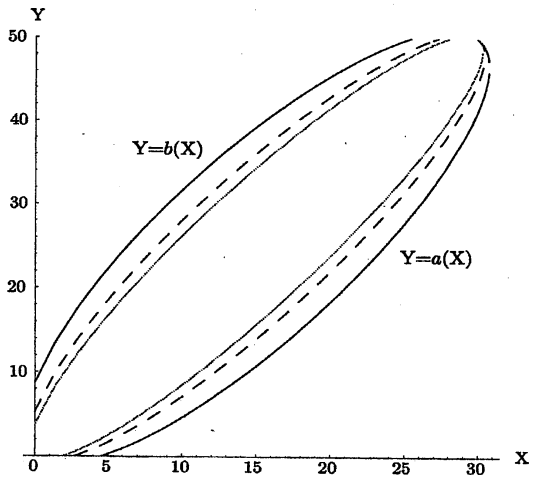

図 3: $m = 30, n = 50$ のときの Y の予測曲線 $Y = a(X), Y = b(X)$

———— 信頼係数99%; − − − − 信頼係数95%; ———— 信頼係数90%

3.2. ポアソン分布の場合

観測されるデータを確率変数 X，未観測の確率変数を Y とし，X と Y は互いに独立に，X はポアソン分布 $Po(m\lambda)$，Y はポアソン分布 $Po(n\lambda)$ に従うとする．ただし，m,n は自然数で既知，λ は正で未知とする．このとき，X に基づいて Y の区間予測を行なう．まず，X,Y の同時確率関数は

$$f_{X,Y}(x,y;\lambda)=\frac{e^{-(m+n)\lambda}m^x n^y \lambda^{x+y}}{x!y!}$$

$$(x=0,1,2,\ldots;\ y=0,1,2,\ldots;\ m,n=1,2,\ldots;\ \lambda>0)$$

となるから，統計量 $T:=X+Y$ は λ に対する十分統計量であり，T の分布は $Po((m+n)\lambda)$ に従う．このとき，$T=t$ を与えたときの Y の条件付分布は2項分布 $B(t,n/(m+n))$ に従い，これは λ に無関係になる．このことは，十分統計量 T に基づく Y の予測区間が未知の母数 λ に無関係に構成することができることを意味している．また，$T=t$ を与えたときの Y の条件付平均 μ_t，条件付分散 σ_t^2，条件付3次キュムラント $\kappa_{3,t}$ は次のようになる．

$$\begin{aligned}
\mu_t &:= E[Y|T=t]=\frac{tn}{m+n},\\
\sigma_t^2 &:= \mathrm{Var}(Y|T=t)=\frac{tmn}{(m+n)^2},\\
\kappa_{3,t} &:= \kappa_3(Y|T=t)=\frac{tmn(m-n)}{(m+n)^3}.
\end{aligned}$$

そこで，m,n が十分大きいとき，(3) と同様に

$$P\{t-y_{\alpha/2}(t)\le Y\le y_{\alpha/2}(t)|T=t\}=1-\alpha \tag{7}$$

が成り立つような2項分布 $B(t,n/(m+n))$ の上側 $100(\alpha/2)\%$ 点 $y_{\alpha/2}(t)$ を漸近的に求める．あとは前節の場合と同様にして，Cornish–Fisher 展開によって

$$\frac{y_{\alpha/2}(t)-\mu_t}{\sigma_t}=u_{\alpha/2}+\frac{\kappa_{3,t}}{6\sigma_t^3}u_{\alpha/2}^2+\cdots$$

より

$$\begin{aligned}
y_{\alpha/2}(t) &= \mu_t+\sigma_t u_{\alpha/2}+\frac{\kappa_{3,t}}{6\sigma_t^2}u_{\alpha/2}^2+\cdots\\
&= \frac{nt}{m+n}+u_{\alpha/2}\sqrt{\frac{mnt}{(m+n)^2}}+\frac{m-n}{6(m+n)}u_{\alpha/2}^2+\cdots
\end{aligned} \tag{8}$$

になる．ただし，$u_{\alpha/2}$ は正規分布 $N(0,1)$ の上側 $100(\alpha/2)\%$ 点とする．よって，(8) において，$y = y_{\alpha/2}(t)$ とおき，$a := n/(m+n)$，$b := mn/(m+n)^2$，$c := (m-n)/\{6(m+n)\}$，$u = u_{\alpha/2}$ とし，$t = x+y$ に注意すれば，(8) から

$$y \fallingdotseq a(x+y) + u\sqrt{b(x+y)} + cu^2 \tag{9}$$

を得る．そこで，(9) の辺々を 2 乗すると

$$\{y - a(x+y) - cu^2\}^2 \fallingdotseq b(x+y)u^2$$

となり，

$$\begin{aligned}(1-a)^2y^2 \;+\; & 2\left\{(a^2-a)x + acu^2 - cu^2 - \frac{1}{2}bu^2\right\}y \\ & + \; a^2x^2 + 2\left(acu^2 - \frac{1}{2}bu^2\right)x + c^2u^4 = 0\end{aligned} \tag{10}$$

となる．ここで，$A := (1-a)^2$，$B := a - a^2$，$C := a^2$，$D := -\{acu^2 - cu^2 - (bu^2/2)\}$，$E := acu^2 - (bu^2/2)$，$F := c^2u^4$ とおくと，(10) より

$$Ay^2 - 2(Bx+D)y + Cx^2 + 2Ex + F = 0$$

になり，これを y について解くと

$$y = \frac{1}{A}\left\{Bx + D \pm \sqrt{(Bx+D)^2 - A(Cx^2+2Ex+F)}\right\}$$

になる．従って，これを用いて (7) から，任意の $\lambda > 0$ について

$$P_\lambda\{a(X) \le Y \le b(X)\} \fallingdotseq 1 - \alpha$$

となる Y の予測区間 $[a(X), b(X)]$ を漸近的に得る．ただし，

$$a(X) = \frac{1}{A}\left\{Bx + D - \sqrt{(Bx+D)^2 - A(Cx^2+2Ex+F)}\right\},$$

$$b(X) = \frac{1}{A}\left\{Bx + D + \sqrt{(Bx+D)^2 - A(Cx^2+2Ex+F)}\right\}$$

とする．また，Y の予測曲線 $Y = a(X)$，$Y = b(X)$ を図 4, 5 において示す．

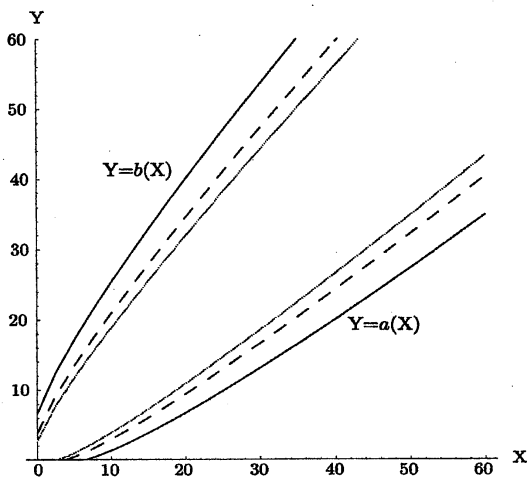

図 4: $m = n = 25$ のときの Y の予測曲線 $Y = a(X), Y = b(X)$

———— 信頼係数99%; − − − − 信頼係数95%; ———— 信頼係数90%

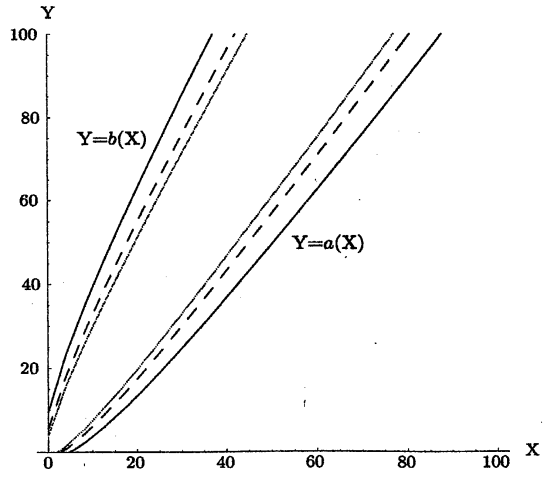

図 5: $m = 30, n = 50$ のときの Y の予測曲線 $Y = a(X), Y = b(X)$

———— 信頼係数99%; − − − − 信頼係数95%; ———— 信頼係数90%

3.3. ランダム予測関数

前節までに論じた予測区間は非ランダム予測区間であるが，信頼係数$1-\alpha$を達成する予測区間を考えるためにはランダム予測区間を導入する必要がある ([Take75]).

前出の母数θをもつ離散指数型分布族についての区間予測において，任意のθについて

$$P_\theta\{a(\mathbf{X}) \leq Y \leq b(\mathbf{X})\} \geq 1-\alpha \tag{11}$$

となる$a(\cdot)$，$b(\cdot)$を求める方法について述べ，区間$[a(\mathbf{X}), b(\mathbf{X})]$を$Y$の信頼係数$1-\alpha$の予測区間といった．そこで，非ランダム予測関数$\phi$を

$$\phi(\boldsymbol{x}, y) = \begin{cases} 1 & (a(\boldsymbol{x}, y) \leq y \leq b(\boldsymbol{x}, y)), \\ 0 & (y < a(\boldsymbol{x}, y),\ \ y > b(\boldsymbol{x}, y)) \end{cases}$$

によって定義すると，(11)から，任意のθについて

$$E_\theta[\phi(\mathbf{X}, Y)] \geq 1-\alpha \tag{12}$$

になる．

次に，一般に，任意の$\boldsymbol{x}$，yについて$0 \leq \phi(\boldsymbol{x}, y) \leq 1$で，任意の$\theta$について(12)を満たす$\phi$を信頼係数$1-\alpha$の$Y$のランダム予測関数という．そこで，$\phi$をランダム予測関数とし，任意に$\boldsymbol{x}$を固定するとき，$y^*(\boldsymbol{x})$が存在して，$0 \leq y \leq y^*(\boldsymbol{x})$において$\phi(\boldsymbol{x}, y)$は$y$に関して単調増加であり，$y^*(\boldsymbol{x}) \leq y$において$\phi(\boldsymbol{x}, y)$は$y$に関して単調減少であるとする．このとき，任意に$\boldsymbol{x}$を固定するとき，任意の$u(0 \leq u \leq 1)$について，集合$\{y|\phi(\boldsymbol{x}, y) \geq u\}$は区間$[c(\boldsymbol{x}, u), d(\boldsymbol{x}, u)]$になる．従って，$U$を区間$[0, 1]$上の一様分布に従う確率変数とすれば，任意の$\theta$について

$$P_\theta\{c(\mathbf{X}, U) \leq Y \leq d(\mathbf{X}, U)\} = E_\theta[\phi(\mathbf{X}, Y)]$$

になり，

$$E_\theta[\phi(\mathbf{X}, Y)] \equiv 1-\alpha \tag{13}$$

となるϕをとれば，信頼係数$1-\alpha$の相似なランダム予測関数を得て，そこから$\mathbf{X}$に基づいて

$$\{Y|\phi(\mathbf{X}, Y) \geq U\} = [c(\mathbf{X}, U), d(\mathbf{X}, U)]$$

というランダム予測区間を得る．なお，母数θをもつ離散指数型分布族の場合にはθに対する完備十分統計量Tが存在するから，(13)であるための必要十分条件は

$$E[\phi(\mathbf{X},Y)|T]=1-\alpha \tag{14}$$

になる．

そこで，具体的な場合として第2.1節の2項分布の場合を考える．観測されるデータを確率変数X，未観測確率変数をYとし，XとYはたがいに独立にXは2項分布$B(m,p)$，Yは2項分布$B(n,p)$に従うとする．ただしm,nは自然数で既知とし，pは$0<p<1$で未知とする．このとき統計量$T:=X+Y$はpに対する十分統計量であり，Tは$B(m+n,p)$に従う．いま，各$t=0,1,\cdots,m+n$に対して，整数$y_0(t)$，$y_1(t)$ $(0\le y_0(t)\le y_1(t)\le n)$と$0\le\gamma_0(t)<1$，$0<\gamma_1(t)\le 1$となる$\gamma_0(t)$，$\gamma_1(t)$を適当に定めて

$$\phi_t(y)=\begin{cases}0 & (y<y_0(t),\ y>y_1(t)),\\ \gamma_0(t) & (y=y_0(t)),\\ \gamma_1(t) & (y=y_1(t)),\\ 1 & (y_0(t)<y<y_1(t))\end{cases}$$

となるランダム予測関数$\phi_t(y)$をつくり，(14)の条件を満たすようにする．しかし，このランダム予測関数$\phi_t(y)$のつくり方は一意的ではない．ここでは

$$P\{Y<y_0(t)|T=t\}+(1-\gamma_0(t))P\{Y=y_0(t)|T=t\}=\frac{\alpha}{2},$$

$$P\{Y>y_1(t)|T=t\}+(1-\gamma_1(t))P\{Y=y_1(t)|T=t\}=\frac{\alpha}{2}$$

となるように$y_0(t)$，$y_1(t)$，$\gamma_0(t)$，$\gamma_1(t)$を定めることにする．

実際，$m=n=20$の場合に$\alpha=0.05,0.10$とする．この場合，$T=t$を与えたときのYの条件付確率関数はmとn，xと$20-x$，yと$20-y$に関して対称になるから，$0\le t\le 20$の範囲について考えれば十分である．このとき，$\gamma_0(t)\equiv\gamma_1(t)$であり，$y_0(t)$，$y_1(t)$，$\gamma_0(t)$の値は表1, 2のようになる．そして，実際に表1, 2から得られるランダム予測関数から，区間$[0,1]$上の一様乱数Uを用いて，Xに基づくランダム予測区間

$$\{Y|\phi_{X+Y}(Y)\ge U\}=[c(X,U),d(X,U)]$$

を得ることができる．

t	$y_0(t)$	$y_1(t)$	$\gamma_0(t)$
0	0	0	0.975
1	0	1	0.95
2	0	2	0.8974
3	0	3	0.7833
4	0	4	0.5284
5	1	4	0.9902
6	1	5	0.8155
7	1	6	0.4988
8	2	6	0.9666
9	2	7	0.7183
10	2	8	0.2627
11	3	8	0.8467
12	3	9	0.4721
13	4	9	0.9316
14	4	10	0.5943
15	5	10	0.9886
16	5	11	0.6679
17	5	12	0.0807
18	6	12	0.7079
19	6	13	0.1346
20	7	13	0.7207

表 1: $\alpha = 0.05$ の場合のランダム予測関数 $\phi_t(y)$ の $y_0(t)$, $y_1(t)$, $\gamma_0(t)$ の値

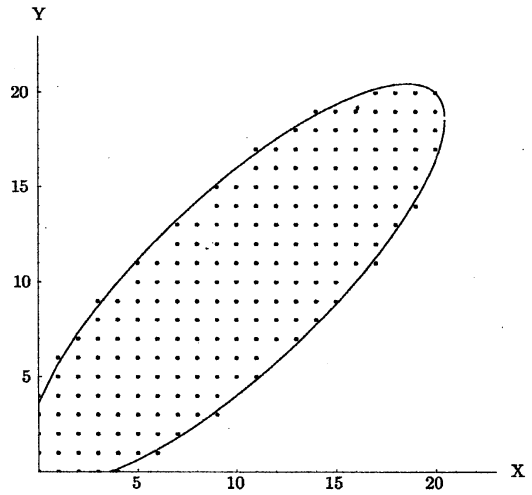

図 6: ランダム予測関数 ϕ_t に基づく Y の 95% ランダム予測区間を表示する点と
95% 非ランダム予測曲線

t	$y_0(t)$	$y_1(t)$	$\gamma_0(t)$
0	0	0	0.95
1	0	1	0.9
2	0	2	0.7947
3	0	3	0.5667
4	0	4	0.0569
5	1	4	0.8206
6	1	5	0.5061
7	2	5	0.9730
8	2	6	0.7055
9	2	7	0.2542
10	3	7	0.8313
11	3	8	0.4442
12	4	8	0.9193
13	4	9	0.5619
14	5	9	0.9815
15	5	10	0.6375
16	5	11	0.0644
17	6	11	0.6835
18	6	12	0.1274
19	7	12	0.7053
20	7	13	0.1472

表 2: $\alpha = 0.10$ の場合のランダム予測関数 $\phi_t(y)$ の $y_0(t)$, $y_1(t)$, $\gamma_0(t)$ の値

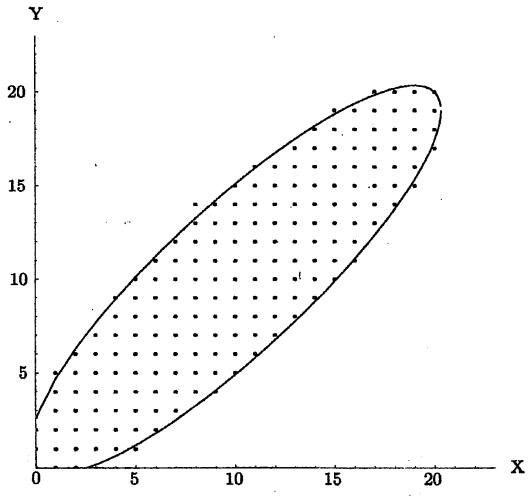

図 7: ランダム予測関数 ϕ_t に基づく Y の 90% ランダム予測区間を表示する点と 90% 非ランダム予測曲線

4. 区間予測の応用

まず，プロ野球で，あるチームがm試合消化した段階でX勝しているとき，残りn試合での勝数Yを区間予測する問題を２項分布の場合に適用する．また，プロ野球である選手がある時点でそれまでに打ったホームラン数Xに基づいて残り試合におけるホームラン数Yを区間予測する問題をポアソン分布の場合に適用する．

例１ (日本のプロ野球チームの勝数の予測). 日本のプロ野球もいよいよ大詰めを迎えた(1998年9月10日)現在，セ・リーグにおいて巨人は３位であるが６連勝した．果たしてミラクルは起こるのか? そこで，横浜，中日も含めて残り試合での勝数の区間予測を行なう．各チームがm試合消化した段階でX勝しているとき，残りn試合での勝数Yを各チームについて区間予測を２項分布の場合の方法で行なうと，Yの信頼係数$100(1-\alpha)\%$の予測区間と予測曲線を得る(表３～４, 図８～１３参照).

チーム	試合数	勝数	負数	引分	残り試合数
横浜	110(109)	65	44	1	26
中日	115(114)	63	51	1	21
巨人	119	64	55	0	16

表3: 1998年9月10日現在の３チームの成績

このとき，残り試合での勝数の信頼係数$100(1-\alpha)\%$の予測区間は次のようになる．

信頼係数(%)	横浜	中日	巨人
99	[8.123, 22.268]	[5.150, 17.747]	[3.103, 13.897]
95	[9.902, 20.745]	[6.685, 16.341]	[4.407, 12.680]
90	[10.813, 19.940]	[7.476, 15.604]	[5.080, 12.044]
80	[11.861, 18.992]	[8.390, 14.741]	[5.858, 11.299]
70	[12.565, 18.341]	[9.007, 14.151]	[6.384, 10.791]
60	[13.123, 17.818]	[9.497, 13.679]	[6.803, 10.385]
50	[13.600, 17.365]	[9.918, 13.271]	[7.161, 10.034]
実際に残り試合数(・)での勝数	14 (26)	12 (21)	9 (16)

表4: 残り試合での各チームの勝数の予測区間

上記のことから，第3.1節の２項分布の場合の区間予測の方法は妥当に思われる．

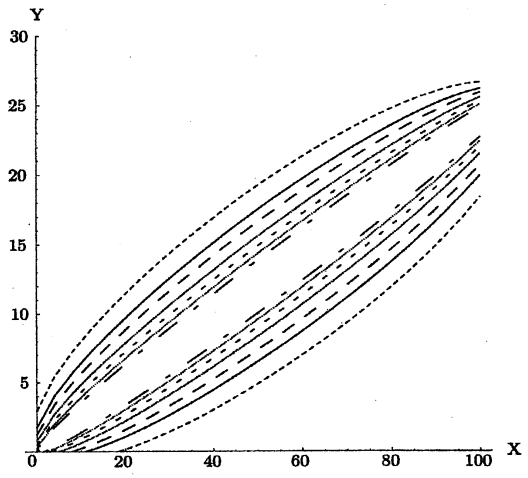

図 8: 横浜の勝数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

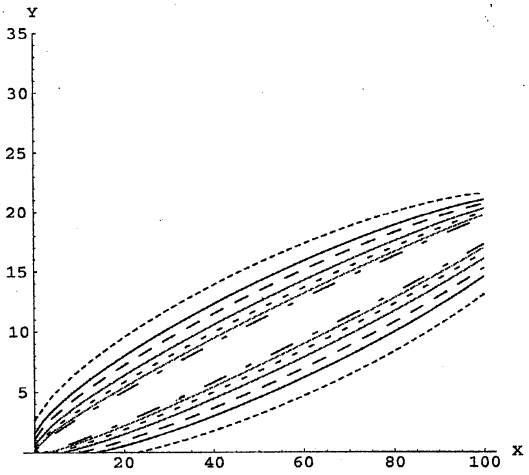

図 9: 中日の勝数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

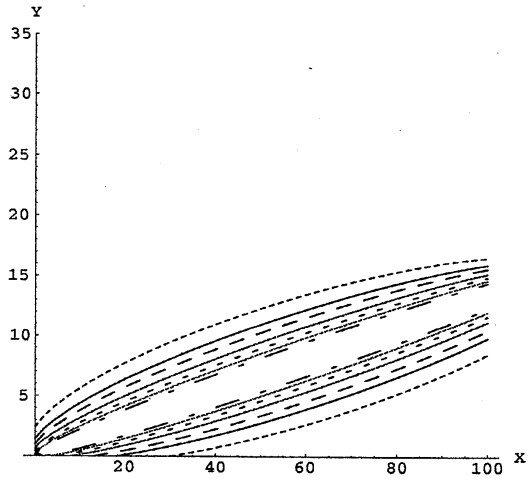

図 10: 巨人の勝数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

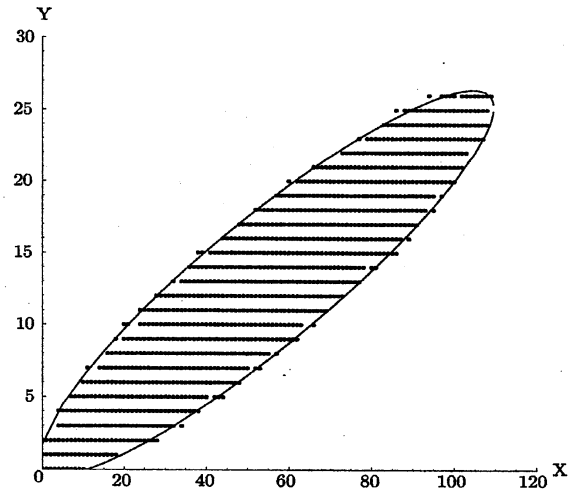

図 11: ランダム予測関数に基づく横浜の勝数 Y の 95% ランダム予測区間を表示する点と 95% 非ランダム予測曲線

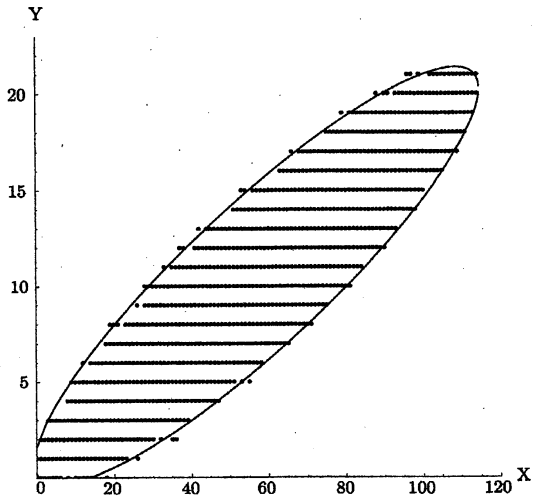

図 12: ランダム予測関数に基づく中日の勝数 Y の 95% ランダム予測区間を表示する点と 95% 非ランダム予測曲線

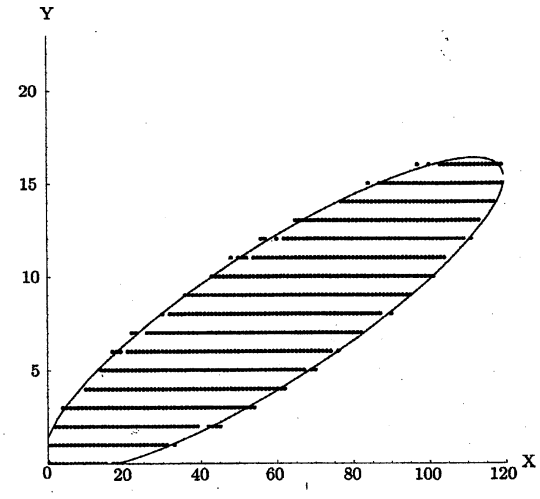

図 13: ランダム予測関数に基づく巨人の勝数 Y の 95% ランダム予測区間を表示する点と 95% 非ランダム予測曲線

また，前半が終了した1998年7月21日現在のセ・リーグの上位3チームの結果は次の表のようであった．

チーム	試合数	勝数	負数	引分	残り試合数
横浜	74	45	28	1	62
中日	77	42	34	1	59
巨人	79	41	38	0	56

表5: 1998年7月21日現在の3チームの成績

このとき，各チームの後半での勝数の信頼係数 $100(1-\alpha)\%$ の予測区間は次のようになる．

信頼係数 (%)	横浜	中日	巨人
99	[24.3777, 50.7226]	[19.4472, 45.1903]	[16.602 , 41.3285]
95	[27.7015, 47.9382]	[22.5515, 42.3211]	[19.5123, 38.4987]
90	[29.4195, 46.4584]	[24.1582, 40.8125]	[21.0254, 37.0192]
80	[31.3868, 44.7088]	[26.0207, 39.0432]	[22.7856, 35.2909]
70	[32.7095, 43.5038]	[27.2812, 37.8331]	[23.9805, 34.1131]
60	[33.7568, 42.5328]	[28.2838, 36.8631]	[24.9333, 33.1715]
50	[34.6517, 41.6909]	[29.144 , 36.0255]	[25.7524, 32.3602]
実際の後半の試合での勝数	34	33	32

表6: 後半戦における3チームの勝数の予測区間

また，3チームの後半での勝数の予測曲線も得る (図14 ～ 16参照)．上記のことから，第3.1節の2項分布の場合の区間予測の方法は妥当に思われる．

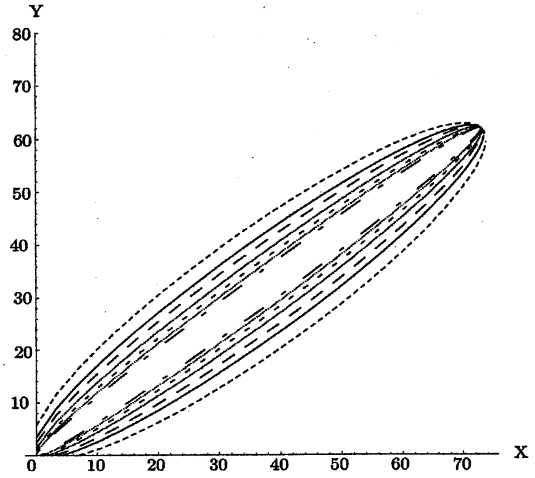

図 14: 横浜の勝数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

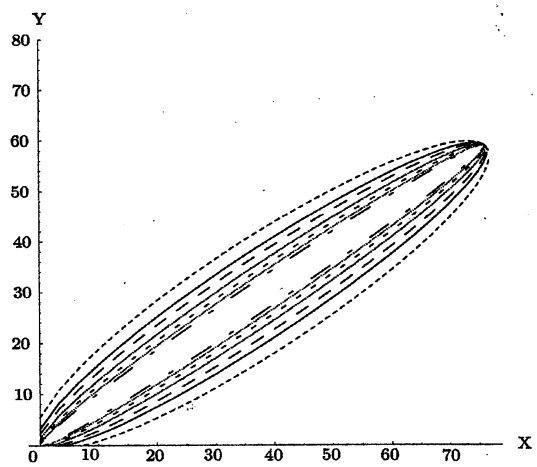

図 15: 中日の勝数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

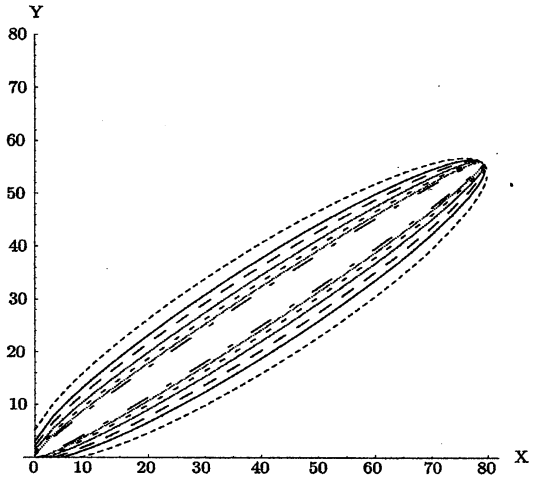

図 16: 巨人の勝数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

例 2 (米国の大リーグ選手のホームラン数の予測). 米国の大リーグの上記の両選手は，1998 年 9 月 8 日現在，144 試合消化した時点でマグワイア選手は 61 本，ソーサ選手は 58 本のホームランを打っている．一般に，その時点での各選手のホームラン数を X とする．残り試合は両選手とも 19 試合である．このとき残り試合でのホームラン数 Y の区間予測をポアソン分布の場合の方法で行なうと，その信頼係数 $100(1-\alpha)\%$ の予測区間と予測曲線を得る (表 7，図 17 ～ 18 参照)．

信頼係数 (%)	マグワイア	ソーサ
99	[1.605, 17.287]	[1.401, 16.699]
95	[2.913, 14.803]	[2.663, 14.261]
90	[3.637, 13.601]	[3.364, 13.082]
80	[4.518, 12.271]	[4.218, 11.779]
70	[5.142, 11.408]	[4.824, 10.934]
60	[5.655, 10.741]	[5.323, 10.281]
50	[6.107, 10.182]	[5.762, 9.735]
実際に残り 19 試合で打ったホームラン数	9	8

表 7: マグワイア，ソーサ両選手の残り 19 試合でのホームラン数の予測区間

上記のことから，第 3.2 節のポアソン分布の場合の区間予測の方法は妥当に思われる．

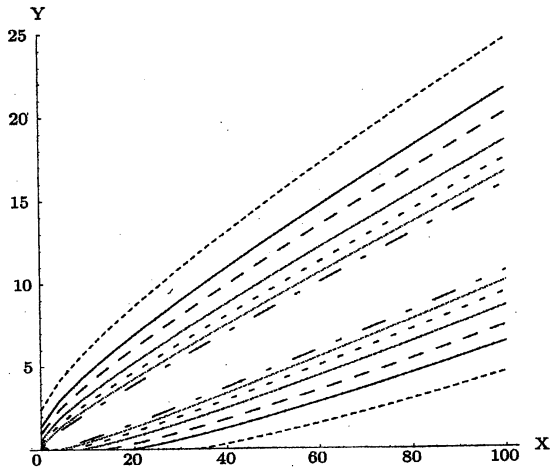

図 17: マグワイア，ソーサのホームラン数 Y の予測曲線

信頼係数99%； 信頼係数95%； 信頼係数90%
信頼係数80%； 信頼係数70%； 信頼係数60%
信頼係数50%

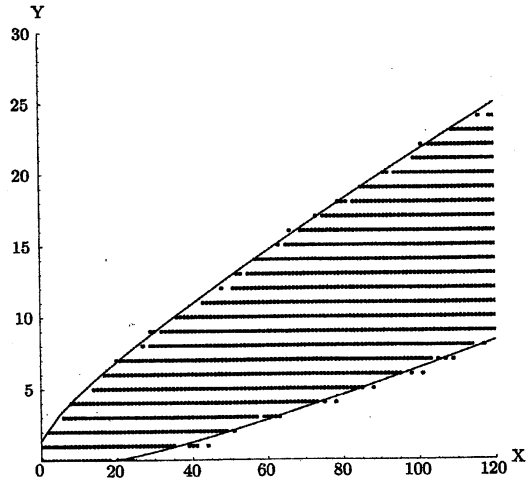

図 18: ランダム予測関数に基づく両選手のホームラン数 Y の 95% ランダム予測区間を表示する点と 95% 非ランダム予測曲線

また，マグワイア選手が116試合消化した時点で46本のホームランを打ち，残り試合数は47であり，ソーサ選手は118試合消化した時点で44本のホームランを打ち，残り試合数は45であった．このとき残り試合でのホームラン数の信頼係数$100(1-\alpha)\%$の予測区間と予測曲線を得る(表8，図19 ~ 22参照)．

信頼係数(%)	マグワイア	ソーサ
99	[7.32294, 33.9567]	[6.19415, 31.2636]
95	[9.70724, 29.8867]	[8.41319, 27.4032]
90	[11.0014, 27.9071]	[9.62060, 25.5284]
80	[12.5584, 25.7086]	[11.0757, 23.4486]
70	[13.6492, 24.2749]	[12.0967, 22.0938]
60	[14.5399, 23.1634]	[12.9312, 21.0444]
実際に残り試合数(・)で打ったホームラン数	24 (47)	22 (45)

表8: マグワイア，ソーサ両選手の残り試合におけるホームラン数の予測区間

上記のことから，第3.2節のポアソン分布の場合の区間予測の方法は妥当に思われる．

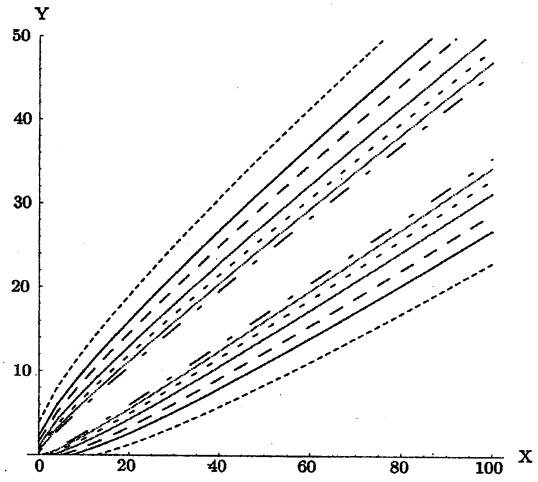

図19: マグワイアのホームラン数Yの予測曲線

信頼係数99%； 信頼係数95%； 信頼係数90%
信頼係数80%； 信頼係数70%； 信頼係数60%
信頼係数50%

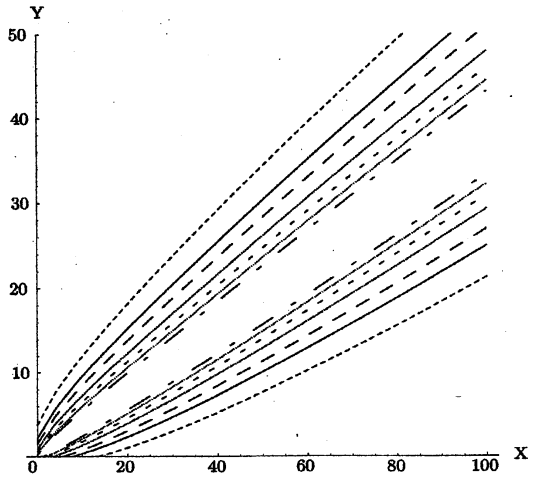

図 20: ソーサのホームラン数 Y の予測曲線

信頼係数99%; 信頼係数95%; 信頼係数90%
信頼係数80%; 信頼係数70%; 信頼係数60%
信頼係数50%

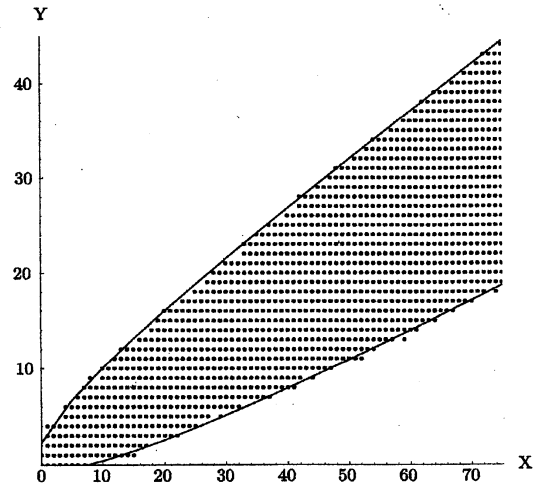

図 21: ランダム予測関数に基づくマグワイアのホームラン数 Y の 95% ランダム予測区間を 表示する点と 95% 非ランダム予測曲線

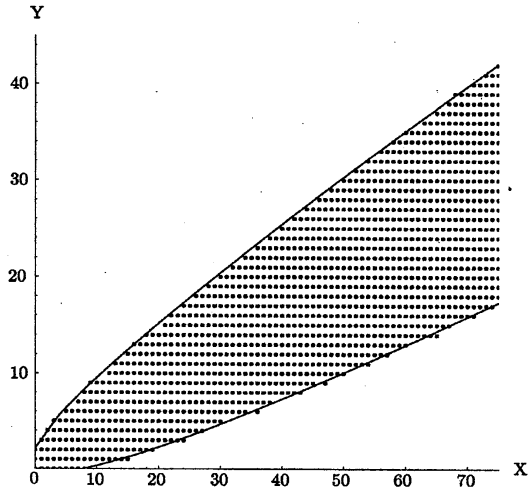

図 22: ランダム予測関数に基づくソーサのホームラン数 Y の 95% ランダム予測区間を表示する点と 95% 非ランダム予測曲線

5. おわりに

本論において，離散指数型分布族における未観測確率変数の区間予測法を十分統計量を通して論じた．具体的には 2 項分布とポアソン分布の場合には十分統計量を与えたときの未観測確率変数の条件付分布はそれぞれ超幾何分布，2 項分布になることを利用して，予測区間を漸近的に構成できることを示した．またそれぞれの場合に，現実の問題への応用として，1998 年の日本のプロ野球の 3 チームの残り試合での勝数の区間予測，米国の大リーグのホームラン新記録を作ったマグワイア，ソーサ両選手の残り試合でのホームラン数の区間予測について数値的に検討した．その結果，本論の区間予測が妥当なものであることが確かめられた．さらに，この区間予測法は現実の他の問題へも適用可能であると考えられる．

参考文献

[A90] Akahira, M. (1990). *Theory of Statistical Prediction.* Lecture Note at the Middle East Technical University, Ankara.

[BC96] Barndorff-Nielsen, O. E. and Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2**, 319-340.

[G93] Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman & Hall, New York.

[G70] Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian.* Griffin, London.

[Taka96] Takada, Y. (1996). Statistical properties of prediction intervals. *Sugaku Expositions* **9**, 153-168.

[Take75] 竹内 啓 (1975). 統計的予測論. 培風館.