İstatistik, Journal of the Turkish Statistical Association
2000, Vol. 3, No 3, pp. 58-82

PREDICTION INTERVALS FOR A DISCRETE EXPONENTIAL FAMILY OF DISTRIBUTIONS AND ITS APPLICATIONS

Masafumi Akahira and Eisuke Hida
Institute of Mathematics
University of Tsukuba
Ibaraki 305-8571, Japan

Abstract

Let $\mathbf{X}$ be an observable random vector and Y a random variable to be observed in future. Assume that the joint distribution of $\mathbf{X}$ and Y depends on an unknown parameter. In this paper we consider a way of the construction of a prediction interval for Y based on $\mathbf{X}$ for a discrete exponential family of distributions. In particular we asymptotically construct the prediction interval in the binomial and Poisson cases, and give practical applications to the prediction of the number of wins of the Japanese professional baseball teams and that of home runs of the players in the major league of the United States.

Key Words: (Similar) prediction region; prediction intervals; confidence coefficient; sufficient statistics; Cornish–Fisher expansion; binomial case; Poisson case.

1. Introduction

In a statistical inference, we may consider a predictive procedure for an unobserved random variable based on an observable random vector (see, e.g. Guttman(1970), Lauritzen(1974), Takeuchi(1975), Hinkley(1979), Butler(1986), Akahira(1990), Bjϕrnstad(1990), Geisser(1993), Takada(1996), Barndorff-Nielsen and Cox(1996)).

Suppose that $\mathbf{X} = (X_1, \cdots, X_m)$ is an observable random vector, Y is a random variable to be observed in future, and the joint distribution of $(\mathbf{X}, Y)$ depends on an unknown parameter θ in Θ, where Θ is a parameter space. Let $\mathcal{Y}$ be a space representing the possible outcomes of Y. If for any α $(0 < \alpha < 1)$ there exists a subset $S_{\mathbf{X}}$ (of $\mathcal{Y}$) based on $\mathbf{X}$ such that

$$P_\theta\{Y \in S_{\mathbf{X}}\} \geq 1 - \alpha, \qquad \text{for all } \theta \in \Theta, \tag{1}$$

then $S_{\mathbf{X}}$ is called a prediction region of Y at confidence coefficient $1 - \alpha$. If $\mathcal{Y}$ is a subset of R^1 and $S_{\mathbf{X}}$ is an interval $[a(\mathbf{X}), b(\mathbf{X})]$, then $S_{\mathbf{X}}$ is called a prediction interval of Y at confidence coefficient $1 - \alpha$ (see Figure 1). If $\mathbf{X}$ takes a realized value $\boldsymbol{x} = (x_1, \cdots, x_m)$,

then the interval $[a(\boldsymbol{x}), b(\boldsymbol{x})]$ is called a prediction interval of Y at confidence coefficient $100(1-\alpha)\%$. If, in particular, the equality in (1) holds, then the prediction region $S_{\mathbf{X}}$ is said to be similar.

In this paper we consider the case when the joint distribution of $(\mathbf{X}, Y)$ belongs to a discrete exponential family of distributions with an unknown one-dimensional parameter θ. Since there exists a complete and sufficient statistic T, using a conditional distribution of Y given T we obtain the conditional mean, variance and third cumulant, and give a way to construct a prediction interval of Y based on $\mathbf{X}$, by the Cornish–Fisher expansion. Indeed, for the binomial and Poisson cases, we asymptotically obtain the prediction intervals and curves for Y, and give practical applications to the prediction of the number of wins of the Japanese professional baseball teams and that of home runs of the players in the major league of the United States.

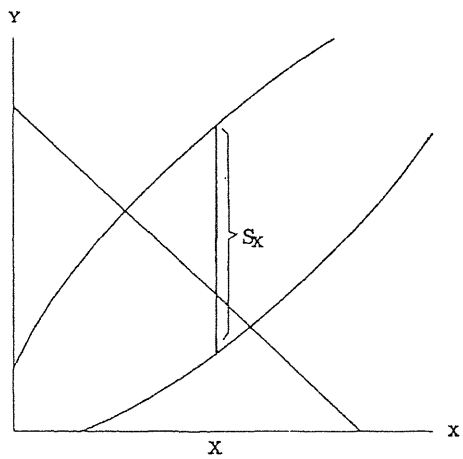

Figure 1: Prediction interval S_X of Y based on X

2. Prediction intervals for a discrete exponential family of distributions

Suppose that $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are independent and identically distributed random variables according to a one-parameter exponential type distribution with a probability mass function (or p.m.f. for short)

$$f(x;\theta) = c(\theta)h(x)\exp\{\eta(\theta)t(x)\}$$

for $x = 0, 1, 2, \cdots, \theta \in \Theta = \mathrm{R}^1$, where $c(\theta)$ and $h(x)$ are nonnegative real-valued functions of θ and x, respectively, and $\eta(\theta)$ and $t(x)$ are real-valued functions of θ and x, respectively. Then the joint p.m.f. of $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ is given

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

Letting $T := \sum_{i=1}^{m} t(X_i) + \sum_{j=1}^{n} t(Y_j)$, T is a complete and sufficient statistic for θ, hence the conditional p.m.f. of $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ given T is independent of θ. So, using the conditional distribution of $Y := \sum_{j=1}^{n} t(Y_j)$ given the sufficient statistic T, we can construct a prediction interval which is independent of unknown parameter θ. Actually, we construct a prediction interval of Y according to the following procedures (i) to (iii).

(i) Let $f_{Y|T}(\cdot|t)$ be a conditional p.m.f. of Y given $T = t$. Since T is sufficient for θ, it follows that $f_{Y|T}(\cdot|t)$ is independent of θ. Using $f_{Y|T}(\cdot|t)$, we obtain the conditional mean $\mu_t := E[Y|T = t]$, the conditional variance $\sigma_t^2 := \mathrm{Var}(Y|T = t)$ and the conditional third cumulant $\kappa_{3,t} := \kappa_3(Y|T = t) = E[(Y - \mu_t)^3|T = t]$ of Y given $T = t$.

(ii) Using the Cornish–Fisher expansion with μ_t, σ_t^2 and $\kappa_{3,t}$ in (i), we asymptotically get $\underline{y}(t)$, $\bar{y}(t)$ such that

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

for any α $(0 < \alpha < 1)$ and any $t \in \mathbf{R}^1$.

(iii) From (2), we have for any $\theta \in \Theta$

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

Since $T := \sum_{i=1}^{m} t(X_i) + \sum_{j=1}^{n} t(Y_j) = \sum_{i=1}^{m} t(X_i) + Y$ is complete and sufficient, we asymptotically obtain $a(\,\cdot\,)$, $b(\,\cdot\,)$ such that

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

Then the interval $[a(\mathbf{X}),\ b(\mathbf{X})]$ is a prediction interval of Y at confidence coefficient $1-\alpha$.

2.1. Binomial case

Suppose that X is an observable random variable, Y is a random variable to be observed in future, and X and Y are independent. Further, assume that X is distributed according to the binomial distribution $B(m, p)$ whose p.m.f.

$$f_X(x;p) = \binom{m}{x} p^x q^{m-x} \qquad (x = 0, 1, \cdots, m;\ \ 0 < p < 1 \ \text{ and } \ q = 1 - p),$$

and Y is distributed according to the binomial distribution $B(n,p)$, where m and n are known natural numbers, and p is unknown. Then we construct a prediction interval of Y based on X at confidence coefficient $1-\alpha$. Since the joint p.m.f. of (X,Y) is given by

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

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

it follows that the statistic $T := X+Y$ is sufficient for p, and T is distributed according to the binomial distribution $B(m+n,p)$. Then the conditional p.m.f. of Y given $T=t$ is

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

which is independent of p. This means that the prediction interval of Y based on the sufficient statistic T is constructed independently of p. The distribution with the above p.m.f. $f_{Y|T}(y|t)$ is called the hypergeometric distribution $H(t,n,m+n)$. When $T=t$ is given, the conditional mean μ_t, the conditional variance σ_t^2 and the conditional third cumulant $\kappa_{3,t}$ of Y are given by

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

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

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

respectively.

When m and n are large, using the Cornish–Fisher expansion we asymptotically obtain the upper $100(\alpha/2)$ percentile $y_{\alpha/2}(t)$ of the hypergeometric distribution $H(t,n,m+n)$ such that

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

First, by the Cornish–Fisher expansion we have

$$\frac{y_{\alpha/2}(t) - \mu_t + \frac{1}{2}}{\sigma_t} = u_{\alpha/2} + \frac{\kappa_{3,t}}{6\sigma_t^3}u_{\alpha/2}^2 + \cdots,$$

that is,

$$y_{\alpha/2}(t) = \mu_t - \frac{1}{2} + \sigma_t u_{\alpha/2} + \frac{\kappa_{3,t}}{6\sigma_t^2} u_{\alpha/2}^2 + \cdots$$

$$= \frac{tn}{m+n} - \frac{1}{2} + u_{\alpha/2}\sqrt{t\left(1-\frac{t}{m+n}\right)\frac{mn}{(m+n)(m+n-1)}}$$

$$+\frac{m-n}{6(m+n-2)}\left(1-\frac{2t}{m+n}\right)u_{\alpha/2}^2 + \cdots, \tag{4}$$

where $u_{\alpha/2}$ is the upper $100(\alpha/2)$ percentile of the standard normal distribution $N(0,1)$. Letting $y := y_{\alpha/2}(t)$, $a := n/(m+n)$, $b := mn/\{(m+n)(m+n-1)\}$, $c := (m-n)/(m+n-2)$, $u = u_{\alpha/2}$ and $t := x+y$, then we obtain from (4)

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

which implies that

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

Hence we have

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

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

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

$$-b(x+y)u^2 + \frac{b(x+y)^2}{m+n}u^2 - \frac{c}{6}u^2 = 0,$$

which implies that

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

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

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

$$+\left[\left\{\frac{c}{3}u^2-1\right\}\left\{a-\frac{cu^2}{3(m+n)}\right\}-bu^2\right]x+\frac{c^2}{36}u^4+\frac{1}{4}-\frac{c}{6}u^2=0. \tag{6}$$

Putting

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

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

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

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

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

$$F:=\frac{c^2}{36}u^4+\frac{1}{4}-\frac{c}{6}u^2,$$

we have from (6)

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

whose solution is given by

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

From (3), we asymptotically get a prediction interval $[a(X), b(X)]$ of Y at confidence coefficient $1-\alpha$ such that

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

for $0 < p < 1$. Then $a(X)$ and $b(X)$ are given by

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

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

Drawing the curves $Y = a(X)$ and $Y = b(X)$, i.e. the prediction curves of Y, we can get the prediction interval of Y in Figures 2 and 3.

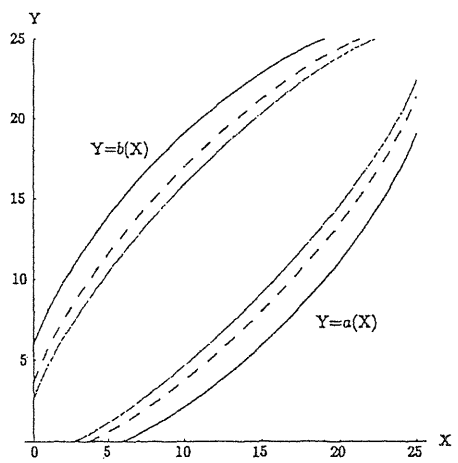

Figure 2: Prediction curves $Y = a(X)$ and $Y = b(X)$ for Y at confidence coefficient $1-\alpha$ for $m = n = 25$

$1-\alpha$: ———— 99% , — — — — 95% , ———— 90%

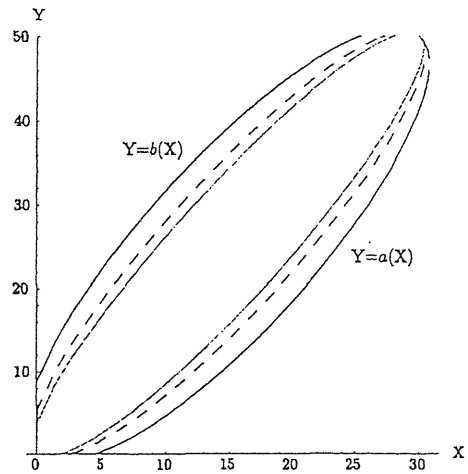

Figure 3: Prediction curves $Y = a(X)$ and $Y = b(X)$ for Y at confidence coefficient $1-\alpha$ for $m = 30$ and $n = 50$

$1-\alpha$: ———— 99% , — — — — 95% , ———— 90%

2.2. Poisson case

Suppose X is an observable random variable, Y is a random variable to be unobserved, and X and Y are independent. Further, we assume that X is distributed according to the Poisson distribution $Po(m\lambda)$ whose p.m.f.

$$f_X(x) = \frac{e^{-m\lambda}(m\lambda)^x}{x!} \qquad (x = 0, 1, 2, \cdots; \quad \lambda > 0),$$

and Y is distributed according to the Poisson $Po(n\lambda)$, when m and n are known natural numbers, and λ is unknown. Then we construct a prediction interval of Y based on X. Since the joint p.m.f. of (X, Y) is given by

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

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

it follows that the statistic $T := X + Y$ is sufficient for λ, and T is distributed according to the Poisson distribution $Po((m+n)\lambda)$. Then the conditional p.m.f. of Y given $T = t$ is the binomial distribution $B(t, n/(m+n))$ which is independent of λ. This means that the prediction interval of Y based on the sufficient statistic T is constructed independently of unknown parameter λ. When $T = t$ is given, the conditional mean μ_t, the conditional variance σ_t^2 and the conditional third cumulant $\kappa_{3,t}$ of Y are given by

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

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

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

respectively. When m and n are very large, in a similar way to (3) using the Cornish–Fisher expansion we asymptotically obtain the upper $100(\alpha/2)$ percentile $y_{\alpha/2}(t)$ of the binomial distribution $B(t, n/(m+n))$ such that

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

By the Cornish–Fisher expansion we have

$$\frac{y_{\alpha/2}(t) - \mu_t + \frac{1}{2}}{\sigma_t} = u_{\alpha/2} + \frac{\kappa_{3,t}}{6\sigma_t^3}u_{\alpha/2}^2 + \cdots,$$

that is,

$$\begin{aligned} y_{\alpha/2}(t) &= \mu_t - \frac{1}{2} + \sigma_t u_{\alpha/2} + \frac{\kappa_{3,t}}{6\sigma_t^2}u_{\alpha/2}^2 + \cdots \\ &= \frac{nt}{m+n} - \frac{1}{2} + u_{\alpha/2}\sqrt{\frac{mnt}{(m+n)^2}} + \frac{m-n}{6(m+n)}u_{\alpha/2}^2 + \cdots, \end{aligned} \tag{8}$$

where $u_{\alpha/2}$ is the upper $100(\alpha/2)$ percentile of the standard normal distribution $N(0,1)$. Letting $y := y_{\alpha/2}(t)$, $a := n/(m+n)$, $b := mn/(m+n)^2$, $c := (m-n)/\{6(m+n)\}$, $u = u_{\alpha/2}$ and $t := x+y$, then we obtain from (8)

$$y \fallingdotseq a(x+y) - \frac{1}{2} + u\sqrt{b(x+y)} + cu^2, \tag{9}$$

which implies that

$$\left\{y - a(x+y) - cu^2 + \frac{1}{2}\right\}^2 \fallingdotseq b(x+y)u^2.$$

Hence we have

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

Putting $A := (1-a)^2$, $B := a-a^2$, $C := a^2$, $D := -\{acu^2 - cu^2 - (bu^2/2) - 1/2 + a/2\}$, $E := acu^2 - (bu^2/2) - a/2$, $F := c^2u^4 + 1/4 - cu^2$, we have from (10)

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

whose solution is given by

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

From (7), we asymptotically get a prediction interval $[a(X), b(X)]$ of Y at confidence coefficient $1-\alpha$ such that

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

for $\lambda > 0$. Then $a(X)$ and $b(X)$ are given by

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

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

Drawing the curves $Y = a(X)$ and $Y = b(X)$, i.e. the prediction curves for Y, we can get the prediction interval of Y in Figures 4 and 5.

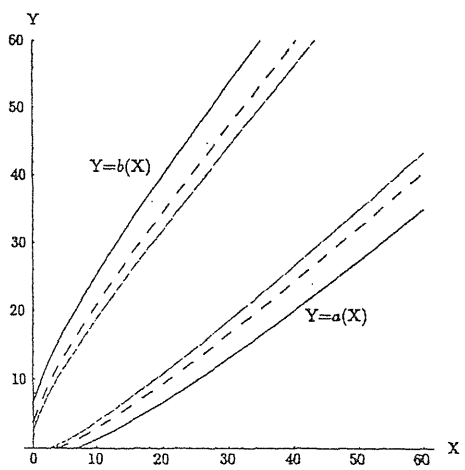

Figure 4: Prediction curves $Y = a(X)$ and $Y = b(X)$ for Y at confidence coefficient $1-\alpha$ for $m = n = 25$

$1-\alpha$: ———— 99% , — — — — 95% , ———— 90%

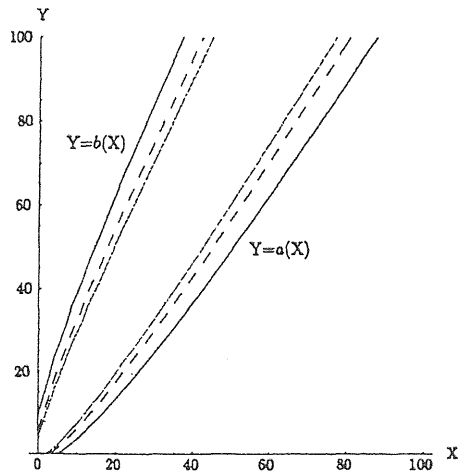

Figure 5: Prediction curves $Y = a(X)$ and $Y = b(X)$ for Y at confidence coefficient $1 - \alpha$ for $m = 30$ and $n = 50$

$1-\alpha$: ———— 99% , — — — — 95% , ———— 90%

2.3. Randomized prediction function

In the previous sections, we consider a non-randomized prediction interval, but we need to take a randomized prediction interval to attain the confidence coefficient $1 - \alpha$ (Takeuchi, 1975).

If for any $\alpha(0 < \alpha < 1)$ there exists an interval $[a(\mathrm{X}), b(\mathrm{X})]$ such that

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

for all $\theta \in \Theta$, then the interval is called a prediction interval of Y at confidence coefficient $1 - \alpha$. We also define a randomized prediction function ϕ at confidence coefficient $1 - \alpha$ as

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

where $a(\boldsymbol{x}, y)$ and $b(\boldsymbol{x}, y)$ are functions satisfying

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

for all $\theta \in \Theta$. Let $\phi(\boldsymbol{x}, y)$ be a randomized prediction function at confidence level $1 - \alpha$, and $\boldsymbol{x}$ be any fixed. Then there exists $y^*(\boldsymbol{x})$ such that $\phi(\boldsymbol{x}, y)$ is monotone increasing in y for $0 \leq y \leq y^*(\boldsymbol{x})$, monotone decreasing in y for $y^*(\boldsymbol{x}) \leq y$. Then the set $\{y|\phi(\boldsymbol{x}, y) \geq u\}$ also becomes an interval $[c(\boldsymbol{x}, u), d(\boldsymbol{x}, u)]$ for all $u(0 \leq u \leq 1)$ when $\boldsymbol{x}$ is arbitrarily fixed. So, letting U be a uniformly distributed random variable over the interval $[0, 1]$, then

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

for all $\theta \in \Theta$ and, if we take ϕ such that

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

then we obtain a similar randomized prediction function ϕ at confidence coefficient $1-\alpha$. We also get a randomized prediction interval

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

at confidence coefficient $1-\alpha$, based to $\mathbf{X}$. Since, in a discrete exponential family of distributions with a parameter θ, a complete and sufficient statistic $T = T(\mathbf{X})$ for θ exists, hence a necessary and sufficient condition for (13) to hold is

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

Now, we consider the binomial case in Section 2.1 as a concrete example. Suppose that X is an observable random variable, Y is a random variable to be observed in future, and X and Y are independent. Further, assume that X is distributed according to the binomial distribution $B(m, p)$ and Y is distributed according to the binomial distribution $B(n, p)$, where m and n are known natural numbers, and p is unknown. The statistic $T := X + Y$ is sufficient for p, and T is distributed according to the binomial distribution $B(m+n, p)$. For each $t = 0, 1, \cdots, m+n$ we take a randomized prediction function $\phi_t(y)$ such that

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

where integers $y_0(t)$, $y_1(t)$ $(0 \leq y_0(t) \leq y_1(t) \leq n)$ and $\gamma_0(t)$, $\gamma_1(t)$ $(0 \leq \gamma_0(t) < 1, 0 < \gamma_1(t) \leq 1)$ are determined by (14). But, the way of the construction of a randomized prediction function $\phi_t(y)$ is not unique. Here, we choose $y_0(t)$, $y_1(t)$, $\gamma_0(t)$ and $\gamma_1(t)$ such that

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

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

Indeed, we consider the case when $\alpha = 0.05$, 0.10 for $m = n = 20$. Since, in the case, the conditional joint distribution of Y given $T = t$ is symmetric with respect to m and

n, x and $2v - x$, y and $2v - y$, it is enough to consider only the case $v \leq t \leq 2v$. In the case, $\gamma_0(t) \equiv \gamma_1(t)$ and the values of $y_0(t)$, $y_1(t)$, $\gamma_0(t)$ are given by Tables 1 and 2. From Tables 1 and 2, using a uniformly distributed random number over the interval $[0, 1]$, we obtain a randomized prediction interval

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

at confidence coefficient $1 - \alpha$. As a result, the difference between the non-randomized prediction interval and the randomized one seems to be small (see Figures 6 and 7). It is also easier to construct a non-randomized prediction interval (curve) in a way in Section 2.1 than to do a randomized prediction one.

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

Table 1: The values of $y_0(t)$, $y_1(t)$, $\gamma_0(t)$ in the randomized prediction function $\phi_t(y)$ for $\alpha = 0.05$

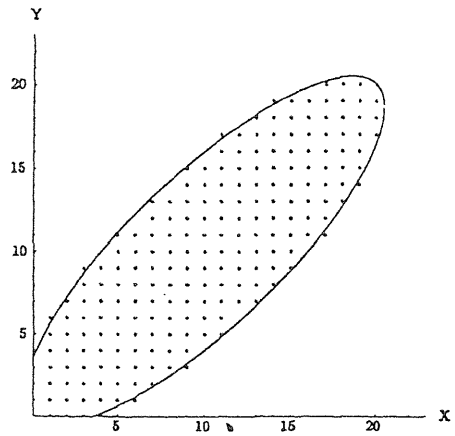

Figure 6: The dots representing the randomized prediction interval of Y based on the randomized prediction function ϕ_t at the confidence coefficient (c.c.) 0.95 and the non-randomized prediction curves at the c.c. 0.95 given in Section 2.1

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

Table 2: The values of $y_0(t)$, $y_1(t)$, $\gamma_0(t)$ in the randomized prediction function $\phi_t(y)$ for $\alpha = 0.10$

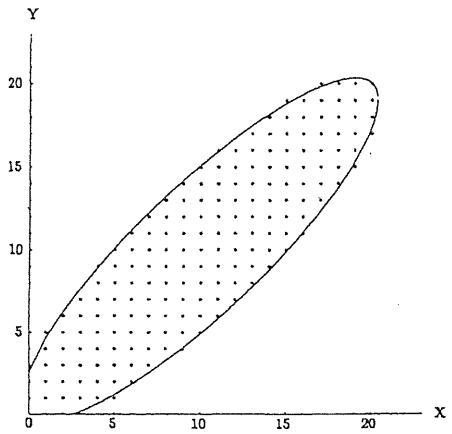

Figure 7: The dots representing the randomized prediction interval of Y based on the randomized prediction function ϕ_t at the confidence coefficient (c.c.) 0.9 and the non-randomized prediction curves at the c.c. 0.9 given in Section 2.1

3. Applications of the prediction interval

First, when some professional baseball team had m games and X wins in them, we consider a prediction interval for the number Y of wins in n residual games, applying to the binomial case. Second, some professional baseball player hited X home runs until certain time, we consider a prediction interval for the number Y of home runs in the rest of games based on X, applying to the Poisson case.

Example 1 (Prediction of the number of wins of the Japanese professional baseball teams). The day, September 10, 1998 was near to the end of the professional baseball season in Japan. In the Central League consisting of six teams, the team "Giants" had the third place but six successive wins up to the day, hence the fans were interested in the final result of the season. So, for the three teams "Bay Stars", "Dragons" and "Giants", we obtain a prediction interval for the number of wins in the rest of games. When each team had m games and X wins in them, we obtain a prediction interval of the number Y of wins in the n games of the rest for the team, applying to the binomial case. Indeed, we get the prediction intervals of Y and prediction curves for Y at confidence coefficient $100(1-\alpha)\%$ including the randomized confidence intervals (see Tables 3 and 4 and Figures 8 to 13).

Team's name	Nos. of finished games	Nos. of wins	Nos. of defeats	No. of draw	Nos. of the rest of games
Bay Stars	110(109)	65	44	1	26
Dragons	115(114)	63	51	1	21
Giants	119	64	55	0	16

Table 3: The result of the three teams in September 10, 1998.

In the above table, (·) means the number of finished games except for the draw. Here, the numbers of the rest of games include those of the draw games, since it is ruled that the draw games are played again in the Central League.

Then we have prediction intervals of the number of wins in the rest of games as follows.

Confidence coefficient(%)	Bay Stars	Dragons	Giants
99	[7.435, 21.748]	[4.483, 17.256]	[2.447, 13.442]
95	[9.227, 20.199]	[6.034, 15.824]	[3.769, 12.197]
90	[10.146, 19.381]	[6.833, 15.074]	[4.452, 11.546]
80	[11.203, 18.419]	[7.757, 14.197]	[5.243, 10.786]
70	[11.914, 17.759]	[8.381, 13.597]	[5.778, 10.267]
60	[12.478, 17.229]	[8.878, 13.117]	[6.203, 9.852]
50	[12.960, 16.770]	[9.303, 12.703]	[6.568, 9.494]
The real numbers of wins in (·) games of the rest	14 (26)	12 (21)	9 (16)

Table 4: The prediction intervals of the number of wins in the rest of games for the three teams "Bay Stars", "Dragons" and "Giants"

Next, at the end of the time of the first half of the season in 1998, that is, in July 21, 1998, the rest of games of the upper three teams was following.

Team's name	Nos. of finished games	Nos. of wins	Nos. of defeats	No. of draw	Nos. of the rest of games
Bay Stars	74	45	28	1	62
Dragons	77	42	34	1	59
Giants	79	41	38	0	56

Table 5: The result of the three teams in July 21, 1998.

Then we obtain prediction intervals of the number of wins at confidence coefficient $100(1-\alpha)\%$ in the latter half of the season in 1998 (see Table 6).

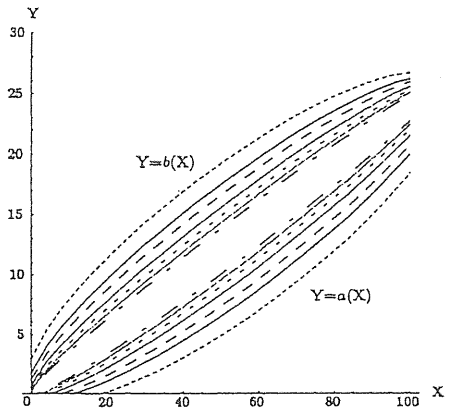

Figure 8: The prediction curves $Y = a(X)$ and $Y = b(X)$ for "Bay Stars"

Confidence coefficient: ——— 99% ; ——— 95%; - — — — — 90%
——— 80% ; - - - - - - - - 70%; ——— 60%
—— - —— - 50%

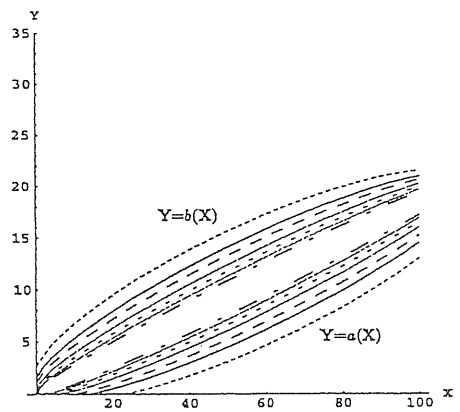

Figure 9: The prediction curves $Y = a(X)$ and $Y = b(X)$ for "Dragons"

Confidence coefficient: ——— 99% ; ——— 95%; - — — — — 90%
——— 80% ; - - - - - - - - 70%; ——— 60%
—— - —— - 50%

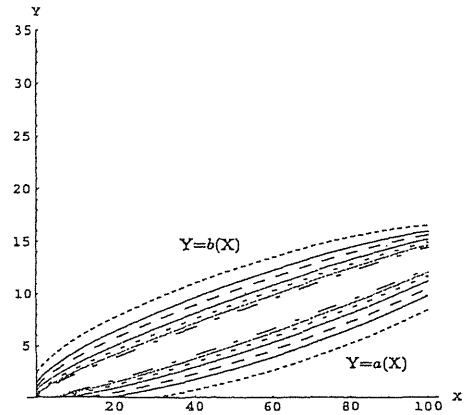

Figure 10: The prediction curves $Y = a(X)$ and $Y = b(X)$ for "Giants"

Confidence coefficient: ——— 99% ; ——— 95%; - — — — — 90%
——— 80% ; - - - - - - - - 70%; ——— 60%
— - — - 50%

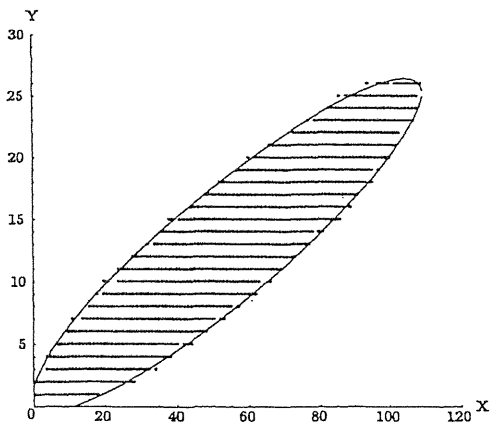

Figure 11: The dots representing the randomized prediction interval for "Bay Stars" based on the randomized prediction function at the confidence coefficient (c.c.) 0.95 and the non-randomized prediction curves at the c.c. 0.95 given Section 2.1

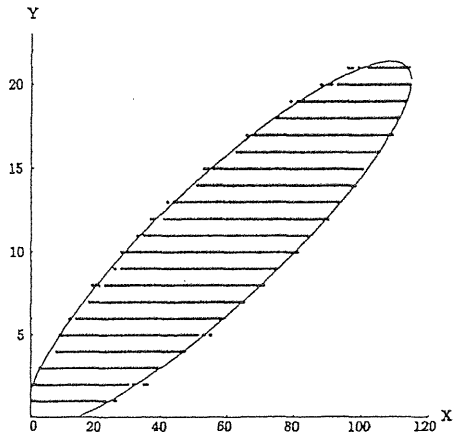

Figure 12: The dots representing the randomized prediction interval for "Dragons" based on the randomized prediction function at the confidence coefficient (c.c.) 0.95 and the non-randomized prediction curves at the c.c. 0.95 given Section 2.1

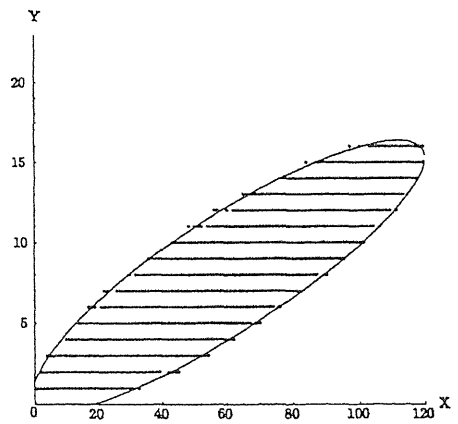

Figure 13: The dots representing the randomized prediction interval for "Giants" based on the randomized prediction function at the confidence coefficient (c.c.) 0.95 and the non-randomized prediction curves at the c.c. 0.95 given Section 2.1

Confidence coefficient(%)	Bay Stars	Dragons	Giants
99	[23.4401, 49.8896]	[18.5622, 44.3707]	[15.7539, 40.5331]
95	[26.7683, 47.0776]	[21.6591, 41.4794]	[18.6568, 37.6838]
90	[28.4767, 45.5848]	[23.2636, 39.9608]	[20.1675, 36.1955]
80	[30.4449, 43.8215]	[25.1249, 38.1810]	[21.9261, 34.4582]
70	[31.7693, 42.6079]	[26.3854, 36.9646]	[23.1207, 33.2751]
60	[32.8184, 41.6306]	[27.3885, 35.99]	[24.0737, 32.3295]
50	[33.7154, 40.7836]	[28.2495, 35.1489]	[24.8932, 31.5152]
The real numbers of wins in the latter half	34	33	32

Table 6: The prediction intervals of the number of wins for the three teams "Bay Stars", "Dragons" and "Giants" in the latter half

We also get the prediction curves of wins of the three teams at confidence coefficient $100(1-\alpha)\%$ in the latter half (see Figures 14 to 16). From the above, we see that the way of construction of a prediction interval in the binomial case in Section 2.1 seems to be reasonable.

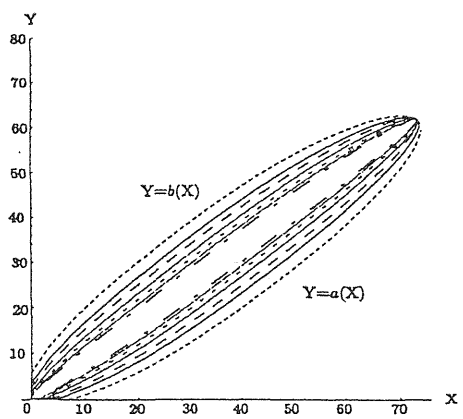

Figure 14: The prediction curves $Y = a(X)$ and $Y = b(X)$ for "Bay Stars"

Confidence coefficient: ——— 99%; ——— 95%; - - - - - 90%
——— 80%; - - - - - - - - 70%; ——— 60%
—— - —— - 50%

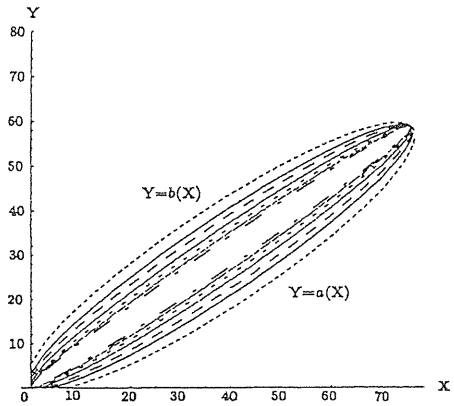

Figure 15: The prediction curves $Y = a(X)$ and $Y = b(X)$ for "Dragons"

Confidence coefficient: ——— 99% ; ——— 95%; - — - — - — 90%
——— 80% ; - - - - - - - - 70%; ——— 60%
—— - —— - 50%

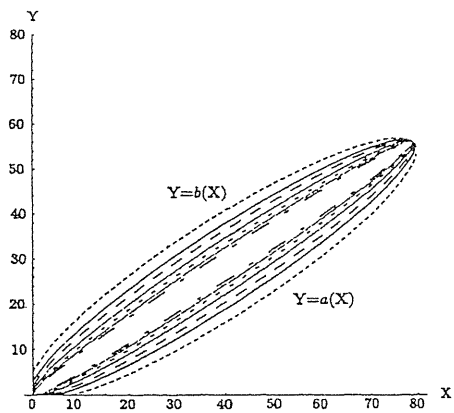

Figure 16: The prediction curves $Y = a(X)$ and $Y = b(X)$ for "Giants"

Confidence coefficient: ——— 99% ; ——— 95%; - — - — - — 90%
——— 80% ; - - - - - - - - 70%; ——— 60%
—— - —— - 50%

Example 2 (Prediction of the number of home runs in the major league). In the major league in the United States, Mark McGwire and Sammy Sosa hited 61 and 58 home runs in September 8, 1998, respectively. When a player hited X home runs in the finished games, we obtain a prediction interval of the number Y of home runs in the rest of games, applying the Poisson case. Indeed, we get the prediction intervals and the prediction curves for Y at confidence coefficient $100(1-\alpha)\%$ including the randomized confidence intervals (see Table 7 and Figures 17 and 18).

Confidence coefficient(%)	McGwire	Sosa
99	[1.071, 16.690]	[0.868, 16.101]
95	[2.371, 14.213]	[2.122, 13.669]
90	[3.091, 13.014]	[2.819, 12.495]
80	[3.968, 11.689]	[3.669, 11.197]
70	[4.589, 10.829]	[4.271, 10.355]
60	[5.099, 10.164]	[4.767, 9.705]
50	[5.549, 9.607]	[5.205, 9.161]
The real number of home runs in 19 games of the rest	9	8

Table 7: The prediction intervals of the number of home runs of McGwire and Sosa in 19 games of the rest

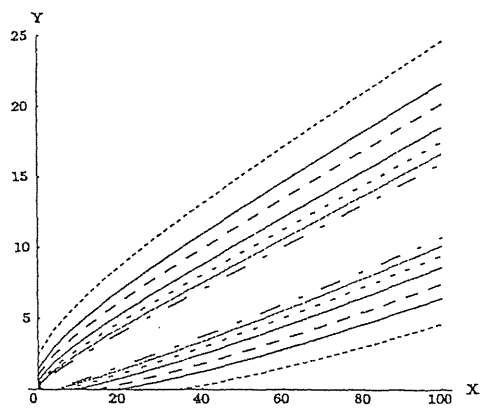

Figure 17: The prediction curves of the number Y of home runs of McGwire and Sosa in 19 games of the rest

Confidence coefficient: 99%; 95%; 90%; 80%; 70%; 60%; 50%

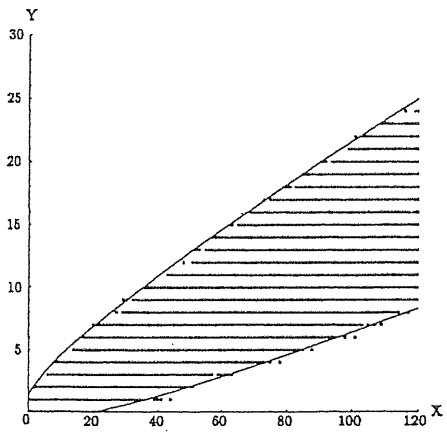

Figure 18: The dots representing the randomized prediction interval for McGwire and Sosa based on the randomized prediction function at the confidence coefficient (c.c.) 0.95 and the non-randomized predictien curves at the c.c. 0.95 given in Section 2.2

Next, at the time when McGwire played 116 games, he hited 46 home runs and the number of his rest of games was 47. On the other hand, at the time when Sosa played 118 games, he hited 44 home runs and the number of his rest of games was 45. Then we get prediction intervals and prediction curves of Y at confidence coefficient $100(1-\alpha)\%$ (see Table 8, Figures 15 to 16).

Confidence coefficient(%)	McGwire	Sosa
99	[6.69146, 33.1830]	[5.57338, 30.5031]
95	[9.05895, 29.1299]	[7.7759 , 26.6592]
90	[10.3445, 27.1589]	[8.97485, 24.7928]
80	[11.8914, 24.9703]	[10.4201, 22.7229]
70	[12.9755, 23.5434]	[11.4344, 21.3747]
60	[13.8607, 22.4374]	[12.2635, 20.3306]
The real number of home runs in (·) games of the rest	24 (47)	22 (45)

Table 8: The prediction intervals of the number of home runs of McGwire and Sosa in games of the rest

From the above, we see that the way of construction of a prediction interval in the Poisson case in Section 2.2 seems to be reasonable.

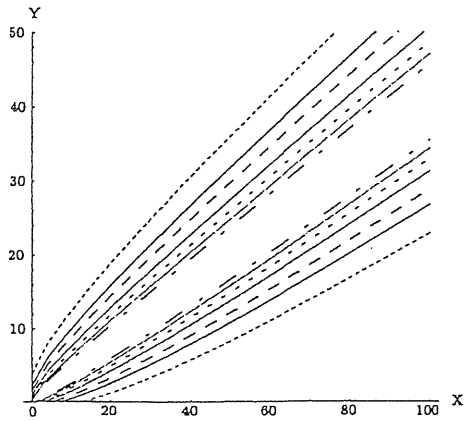

Figure 19: The prediction curves of the number of home runs of McGwire in games of the rest

Confidence coefficient: 99%; 95%; 90%; 80%; 70%; 60%; 50%

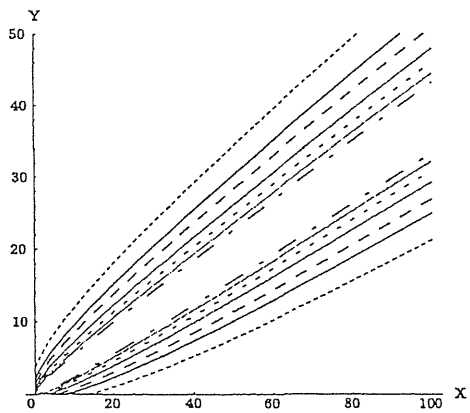

Figure 20: The prediction curves of the number of home runs of Sosa in games of the rest

Confidence coefficient: 99%; 95%; 90%; 80%; 70%; 60%; 50%

REFERENCES

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

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

[3] Bjϕrnstad, J. F. (1990). Predictive likelihood: A review (with discussion). *Statist. Sci.* **5**, 242-265

[4] Butler, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. R. Statist. Soc.*, **B**, **41**,279-312

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

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

[7] Hinkley, D. V. (1979). Predictive likelihood. *Ann Statist.* **7**, 718-728

[8] Lauritzen, S. L. (1974). Sufficiency, prediction, and extreme models. *Scand. J. Statist.* **6**. 128-134

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

[10] Takeuchi, K. (1975). *Statistical Prediction Theory.* (In Japanese), Baifukan, Tokyo.

ÖZET

X gözlenen rasgele vectör ve Y gelecekde gözlenecek rasgele değişken olsun. **X** ve Y 'nin ortak dağılımının bilinmeyen parametreden bağomlı olduğunu varsayalım. Bu makalede biz kesikli üstel dağılımlar ailesi için Y 'nin **X**'e dayalı öngörü güven aralığını kurmağa çalışıyoruz. Özel halde binomial ve Poisson dağılımları durumunda öngöru güven aralıkları kuruluyor ve pratik problemler üzerinde uygulamalar yapılıyor.