On the asymptotic construction of prediction intervals for a discrete exponential family of distributions

Eisuke Hida
Institute of Mathematics, University of Tsukuba
Ibaraki 305-8571, Japan
eisuke@math.tsukuba.ac.jp

Masafumi Akahira
Institute of Mathematics, University of Tsukuba
Ibaraki 305-8571, Japan
akahira@math.tsukuba.ac.jp

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. Akahira (1990), Geisser (1993), Takeuchi (1975)). 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. Now 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 asymptotically a prediction interval of Y based on $\mathbf{X}$, by the Cornish–Fisher expansion (Akahira and Hida (2000)). 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.

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)

$$f(x;\theta) = c(\theta)h(x)\exp\{\eta(\theta)t(x)\} \qquad \text{for } x = 0, 1, 2, \cdots;\ \theta \in \Theta = \mathbf{R}^1,$$

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**), for any α $(0 < \alpha < 1)$ we asymptotically get $\underline{y}(t)$, $\bar{y}(t)$ such that

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

for any $t \in \mathbf{R}^1$.

(**iii**) From (**ii**), 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 $1 - \alpha$.

Further, we give practical applications to the prediction, and make sure that the way of construction of a prediction interval is reasonable.

REFERENCES

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

Akahira, M. and Hida, E. (2000). Prediction intervals for a discrete exponential family of distributions and its applications. In press in *Istatistik* **3**.

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

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