Order Statistics and Data Augmentation

次序統計量和資料擴增

Chiahua Ling[1]
凌嘉華
實踐大學會計學系副教授

Abstract

In this study, we use the method of order statistic technique. In traditional statistical method one may assume a conjugate prior distribution. Our approach of data augmentation is supposed to help for small samples or cases where a very few observations are available. We use prior distributions, centered at the given observations (including order statistics), to generate a larger artificial dataset which may be termed as second generation dataset. This larger second generation dataset is then used to draw inferences. The method is dependent on computational resources, and may be useful in applied problems.

Keywords：Order statistics、prior distribution、conjugate distribution

摘　要

本研究我們使用次序統計量技術的方法,傳統的統計方法是假設有先驗分配(可能是共軛分配)。我們用小樣本中非常少的觀察值作資料增加擴大，用以觀察值(包含次序統計量)為中心的先驗分配來產生一個較大的資料集(稱作第二產生資料集),這個第二產生資料集可以用來做統計推論，這方法需要用到電腦資源而且是對應用問題有幫助的。

關鍵詞：次序統計量、先驗分配、共軛分配

[1] Associate Professor, Department of Accounting, Shih Chien University
E-mail: ling@mail.usc.edu.tw

I、Introduction

The increasing availability of computers and statistical software packages has enlarged the role of statistics as a tool for empirical research. Computational resources and availability of affordable computational facilities have changed the face of research in mathematical and statistical sciences. The techniques of Gibbs sampling (see Casella and George (1992)) the EM algorithm (see Dempster, Laird and Rubin (1997)), bootstrap (see Efron (1982)), etc. are very useful, but can not be implemented without the computational resources. The purpose of this note is also to propose a simple technique with the help of computational resources (see SAS) which can enable to understand a small dataset.

The idea of our proposed technique is very simple, and borrows the idea of standard Bayesian method. It has a flavor of parametric bootstrap since the technique is built up on a parametric model for the idea, and the dataset has been used twice. The proposed technique is described as follows. Assume we have iid observations $X_1, X_2, \ldots, X_n$ from a distribution $f(x \mid \theta)$. When n is large, efficient inferences on θ is possible with the help of classical methods. But when n is small (very small) then one may seek the help of other methods notably the Bayesian one. In such a case one assumes a suitable prior distribution $\theta \sim \pi(\theta)$, often a conjugate one, to draw inferences on θ. What we are proposing here is that given the original dataset $(X_1, X_2, \ldots, X_n)$, call it first stage observations, generate the augmented dataset or the second stage dataset X_{ij} (j = 1, 2, …, m_i; i = 1, 2, … , n) as X_{ij} iid $f(\cdot \mid x_i)$, j = 1, 2, …, m_i. Once this is done, then draw the inferences on θ based on $X_{ij}(1 \le \mathrm{j} \le \mathrm{m_i}; 1 \le i \le n)$ with a new sample size $m = m_1 + m_2 + \ldots + m_n$ which can be made much larger than the original sample size n.

In the following two sections we describe the above mentioned method for the Gamma as well as the exponential distributions. This is still an ongoing work and we hope to expand it for other distributions as well.

II、Gamma Distribution

Now let us consider the gamma model where $X_1, X_2, \ldots, X_n$ are iid Gamma(α) with pdf

f(x|α) = $\dfrac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$, x $\ge$ 0, α> 0.

After observing x_i's , the second stage observations are generated as

$$X_{11}, X_{12}, \ldots, X_{1m_1} \; iid \; Gamma(X_1);$$
$$\vdots$$
$$X_{n1}, X_{n2}, \ldots, X_{nm_n} \; iid \; Gamma(X_n).$$

As a demonstration (see Faraway (1992)) (Andrews and Pregibon (1978)), we generate n = 15 iid observations from Gamma(1) distribution (taking α= 1). The following figure gives the pdf curve superimposed on the relative frequency histogram of the original sample.

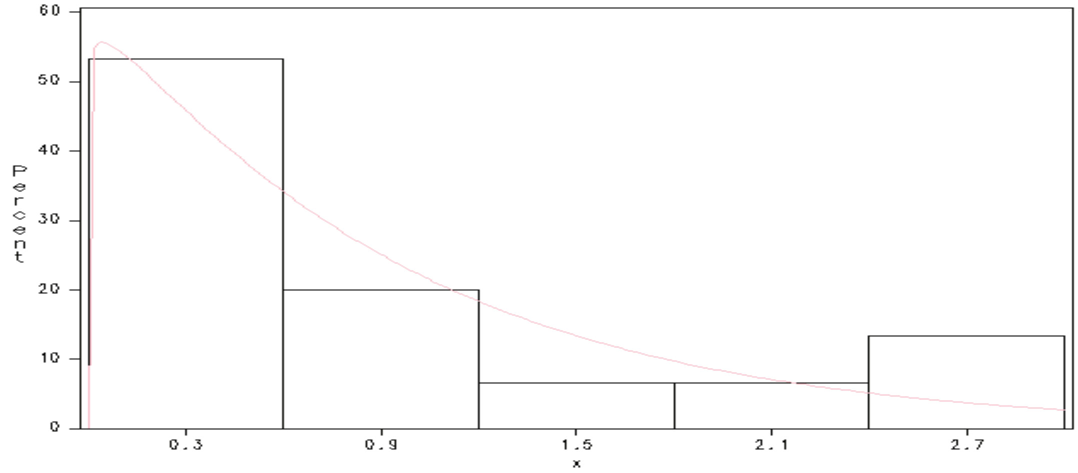


Figure 2. 1. Histogram of the gamma data with n = 15

Using $m_1 = m_2 = \ldots = m_{15} = 10 = m_0$ (say as a demonstration) = 10, we now generate $m = nm_0 = 150$ second stage observations from the above first stage observations. The following diagram shows the relative histogram

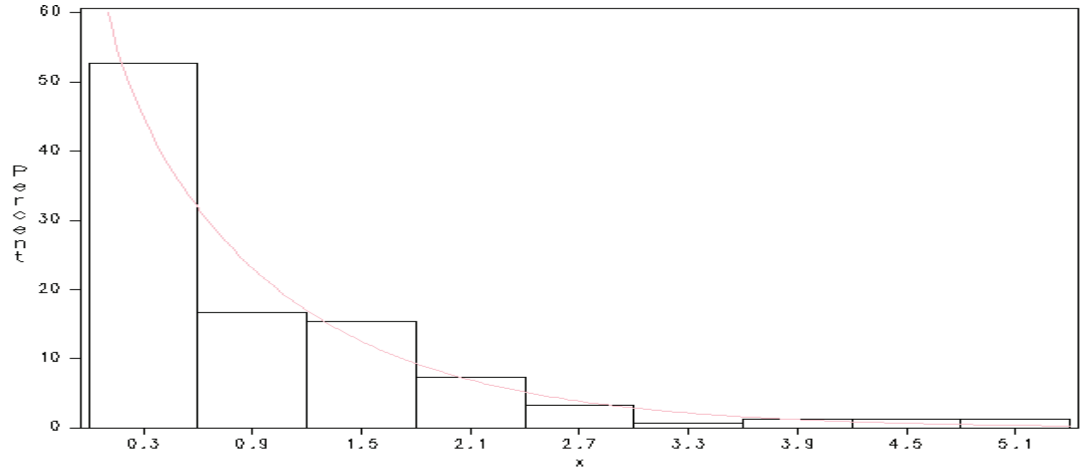


Figure 2. 2. Histograms of the gamma data with m=150.

A modified sampling scheme can be used where m_i's can be chosen suitably. Since the model is positively skewed, we'll give more emphasis on the smaller observations which have higher probabilities to appear.

III、Exponential Distribution

Let us consider the exponential model where $X_1, X_2, \ldots, X_n$ are iid Exp(θ) with pdf f(x|θ) = $\frac{1}{\theta}\exp(-\frac{x}{\theta})$, x > 0, θ > 0.

After observing x_i's, the second stage observations are generated as

$$\begin{aligned} &X_{11}, X_{12}, \ldots, X_{1m_1}\ iid\ Exp(X_1); \\ &\qquad\qquad\vdots \\ &X_{n1}, X_{n2}, \ldots, X_{nm_n}\ iid\ Exp(X_n). \end{aligned}$$

As a demonstration, we take and generate n = 15 iid observations and draw a random sample of size 15 from Exp(1) distribution (taking θ = 1). The following figure gives the pdf curve superimposed on the relative frequency histogram of the original sample.

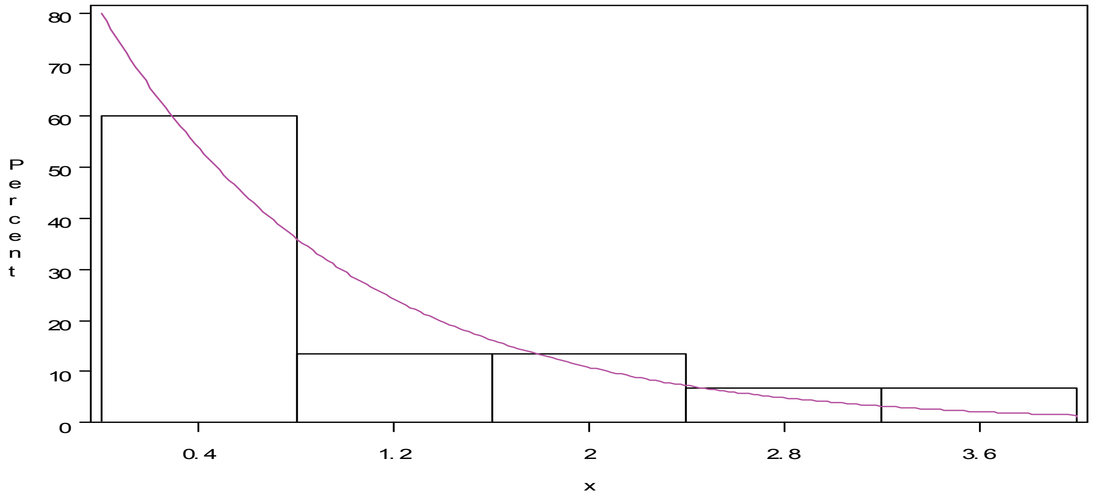
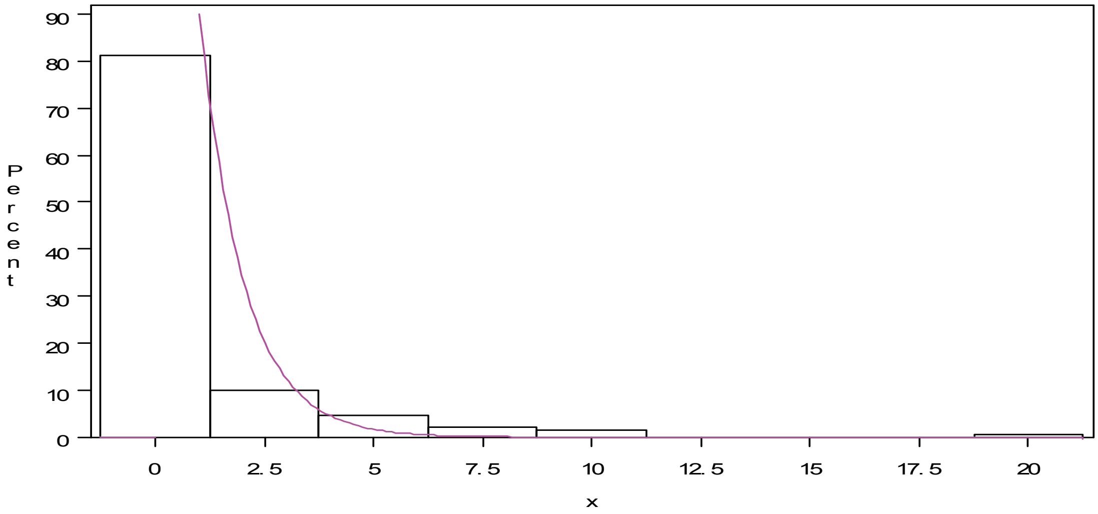


Figure 3. 1. Histogram of the exponential data with n = 15

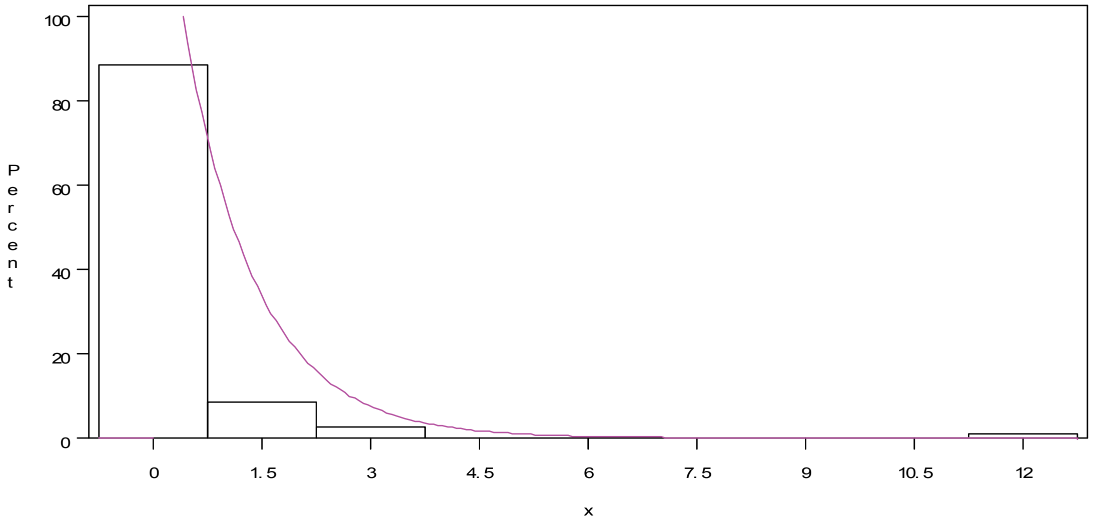
Using $m_1 = m_2 = \ldots = m_{15} = 10 = m_0$ (say as a demonstration) = 10, we now generate $m = nm_0 = 150$ second stage observations from the above first stage observations. The following diagram shows the relative histogram



Fi

gure 3. 2. Histograms of the exponential data with m=150.

So, first we use order statistics (see David and Nagaraja (2003)) to order the observations $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. Then take m_i proportional to (n + 1 - i). For instance, take (n + 1 - i) ; i.e. , m_1= 15, m_2= 14, … , m_{15}= 1, and m = 120. So, generate $X_{i1}, X_{i2}, \cdots, X_{im_i}$ iid $\text{Exp}(X_{(i)})$. Thus the second stage sample size is m = 120 and the following figure shows the relative histograms.



Fi

gure 3. 3. Histograms of the exponential data with m=120.

Note that with the above modified sampling scheme. $X_{i1},\cdots,X_{im_i}$ are iid $\mathrm{Exp}(X_{(i)})$. Therefore, given $x_{(i)}$, $X_{i\cdot}=\sum_{j=1}^{m_i}x_{ij}$ follows Gamma $(x_{(i)},m_i)$ with pdf $(x_{(i)}{}^{m_i}/\Gamma(m_i))\exp(-x_{(i)}x_{i\cdot})x_{i\cdot}{}^{m_i-1}$. So, from the first stage data, we have the second stage sufficient statistic $(X_{1\cdot},X_{2\cdot},\cdots,X_{n\cdot})$. $X_i\sim$ $\mathrm{Exp}(\theta)$, $x_{i\cdot}\,|\,x_i\sim$ Gamma (x_i,m_i). So, the joint density of $(x_i,x_{i\cdot})$ is given by $(\theta x_i{}^{m_i}/\Gamma(m_i))\exp\{-x_i(\theta+x_{i\cdot})\}x_{i\cdot}{}^{m_i-1}$. Therefore, the marginal density of $x_{i\cdot}$ is given by $g_\theta(x_{i\cdot})=m_i x_{i\cdot}{}^{m_{i-1}}\theta(\theta+x_{i\cdot})^{-(m_i+}$. Using $g(x_{i\cdot})$ as likelihood for each $x_{i\cdot}$, one can get the joint likelihood as $L_A=\prod_{i=1}^{n}g_\theta(x_{i\cdot})$, called the likelihood of the augmented data. The value of θ which maximizes L_A (or $\ln L_A$) is denoted by $\widehat{\theta}_{(2)}$ which satisfies the equation

$$\text{n} = \widehat{\theta}_{(2)}\sum_{i=1}^{n}(m_i+1)/(\widehat{\theta}_{(2)}+x_{i\cdot}). \qquad \ldots(\,3.\,1\,)$$

In the following we observe the bias and SE of the above $\widehat{\theta}_{(2)}$ along with those of $\widehat{\theta}_{(1)}$ here $\widehat{\theta}_{(1)}=(1/\bar{x})=$ the maximum likelihood estimator (MLE) of θ based on the original sample. Using n = 10^4, we compute the bias and standard error (SE)

$$\mathrm{Bias}(\widehat{\theta}_{(1)}) \approx \sum_{j=1}^{N}\widehat{\theta}_{(1)}^{\,j}/n-1 \;; \qquad \mathrm{Bias}(\widehat{\theta}_{(2)}) \approx \sum_{j=1}^{N}\widehat{\theta}_{(2)}^{\,j}/n-1;$$

$$\mathrm{SE}(\widehat{\theta}_{(1)}) \approx \sqrt{\sum_{j=1}^{N}(\widehat{\theta}_{(1)}^{(j)}-\widehat{\theta}_{(1)}^{(\cdot)})^2/n} \;; \quad \mathrm{SE}(\widehat{\theta}_{(2)}) \approx \sqrt{\sum_{j=1}^{N}(\widehat{\theta}_{(2)}^{(j)}-\widehat{\theta}_{(2)}^{(\cdot)})^2/n} \;;$$

The computed values are

Bias($\widehat{\theta}_{(1)}$) = 0.001005436 ; Bias($\widehat{\theta}_{(2)}$) = -0.72291 ;

SE($\widehat{\theta}_{(1)}$) = 0.25738 ; SE($\widehat{\theta}_{(2)}$) = 0.180156 .

IV、Conclusion and Further Applications

In this study we follow the method of Bayesian technique with a difference. It has a flavor of parametric bootstrap since the technique is built up on a parametric model for the idea, and the dataset has been used twice. Our approach of data augmentation is supposed to help for small samples or cases where a very few observations are available. We use some computational resources, and may be useful in applied problems. In the future, we can use the proposed method for chi-squared and logistic distributions.

References

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. J. R. Statist. Soc. B 40, (pp. 85-93).

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. The American Statistican, 46, (pp. 167-174).

David, Herbert A. and Nagaraja, Haikady N., Order Statistics, 3rd ed. (2003). Wiley Series in Probability and Statistics. John Wiley & Sons.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, (pp. 1-38).

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, Pa., USA.

Faraway, J. J. (1992). On the Cost of Data Analysis. Journal of Computational and Graphical Statistics 1, (pp. 213-29).

Pelosi, M. K. and Sandifer, T. M. (2002). Airsapace Data Set, Doing Statistics For Business: Data, Inference, and Decision Making. 2nd ed. John Wiley & Sons, Inc., New York.

SAS User's Guide: Basics, 5th ed. (1985). Sas Institute Inc., Cary, North Carolina.

SAS User's Guide: Statistics, 5th ed. (1985). Sas Institute Inc., Cary, North Carolina.