Gibbs 抽样在产犊难易中的应用

陈琦,张立岭*,赵静,马月辉

(1.内蒙古农业大学动物科学与医学学院,内蒙古呼和浩特010018;2.中国农业科学院北京畜牧兽医研究所,北京100094)

摘要　介绍了Gibbs 抽样的原理,并用WinBUGS 1.4 软件,通过模拟数据,利用犊牛出生重/母牛产犊前体重(CBW CCW)数据(先验信息),确定产犊难易4 个分类对应的正态分布的参数μ_i 和σ_i^2,进一步说明Gibbs 抽样在产犊难易中的应用。对 μ_i 和 σ_i^2($i=1,2,3,4$) 进行抽样的结果为:$\mu_1=0.079\ 66$,$\sigma_1^2=1.954$;$\mu_2=0.136\ 6$,$\sigma_2^2=0.756$;$\mu_3=0.293\ 2$,$\sigma_3^2=0.834$;$\mu_4=0.440\ 3$,$\sigma_4^2=0.786$。根据混合分布模型理论计算,得出混合以后的参数 $\mu=0.197\ 4$,$\sigma^2=0.908$。

关键词　Gibbs 抽样;产犊难易;应用

中图分类号　S811　　文献标识码　A　　文章编号　0517-6611(2009)03-01111-02

The Application of Gibbs Sampling on Calving Ease

CHEN Qi et al　(Department of Animal science and medicine,Inner Mongolia Agricultural University,Huhhot,Inner Mongolia 010018)

Abstract　This study aimed to introduce the theory of Gibbs sampling and parameters of normal distribution corresponding four classification of Calving ease,were defined by WinBUGS 1.4 software and analogical data of calf birth weight/precalving cow weight(CBW CCW) in order to explain the application of Gibbs sampling in calving ease.The results of sampling μ_i and σ_i^2($i=1,2,3,4$) is $\mu_1=0.079\ 66$, $\sigma_1^2=1.954$; $\mu_2=0.136\ 6$, $\sigma_2^2=0.756$; $\mu_3=0.293\ 2$, $\sigma_3^2=0.834$; $\mu_4=0.440\ 3$, $\sigma_4^2=0.786$.Then,mixture parameters were calculated according to the theory of mixture distribution model and the result is $u=0.197\ 4$, $\sigma_2^2=0.908$.

Key words　Gibbs sampling;Calving ease;Application

产犊难易(Calving ease)是肉牛育种研究的一个重要性状,也是肉牛育种的一个目标。影响Calving ease 的因素至少有20 种,对于Calving ease 选择,犊牛出生重/母牛产犊前体重(CBW CCW)可能是比单一出生重指标更重要的参数。因为出生体重相同的犊牛,体况大的母牛产犊比体况中小的母牛容易。因此,不考虑CBW CCW,Calving ease 选择的育种值估计或方差分析以及遗传参数的估计是无效的。

Calving ease 受遗传影响,并且是犊牛和母牛联合作用的结果。以前一般采用一个被几个未知阈截断的标准正态分布的阈模型估计肉牛的Calving ease,但是存在一些参数估计的难题没有克服,例如分布、阈值、每个类型的均值和方差估计等。笔者将Gibbs 抽样应用于Calving ease,估计每个分类的未知参数(均值和方差)。

1　研究方法

1.1　Gibbs 抽样原理[1]　2 个参数的贝叶斯模型:$y_i|\mu,\sigma^2\sim N(\mu,\sigma^2)$,$i=1,\cdots,n$;$\mu\sim N(0,1)$,$\sigma^2|S,v\sim S\chi_v^{-2}$。

假设 μ 的先验分布,$p(\mu)$ 是正态分布,均值0,方差1,方差的先验分布 $p(\sigma^2)$ 为尺度反卡方分布。S 和v 是它的超参数,假设它们是已知的。

假设数据是条件独立的,已知 μ 和 σ^2,则:

$$p(y|\mu,\sigma^2)=\prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right)$$

所有参数的后验分布由下式可得:

$$p(\mu,\sigma^2|y)\propto p(\mu)p(\sigma^2)p(y|\mu,\sigma^2)\qquad(1)$$

反卡方分布与下式成比例:

$$p(\chi^2)\propto(\chi^2)^{-(\frac{v}{2}+1)}\exp\left(-\frac{1}{2\chi^{-2}}\right),\ \chi^2>0$$

σ^2 的尺度反卡方分布为:

$$p(\sigma^2|S,v)\propto(\sigma^2)^{-(\frac{v}{2}+1)}\exp\left(-\frac{vS}{2\sigma^2}\right),\ \sigma^2>0$$

因此,后验分布式(1)变为:

$$p(\mu,\sigma^2|y)\propto\exp\left(-\frac{\mu^2}{2}\right)(\sigma^2)^{-(\frac{v}{2}+1)}\exp\left(-\frac{vS}{2\sigma^2}\right)(\sigma^2)^{\frac{n}{2}}\exp\left(-\frac{\sum_i(y_i-\mu)^2}{2\sigma^2}\right)\qquad(2)$$

定义 $\sigma^2>0$(否则 $p(\mu,\sigma^2|y)=0$)。

从全条件分布中获得 $p(\mu|\sigma^2,y)$,从式(2)中挑选只包括 μ 的项。由此,可得:

$$p(\mu,\sigma^2|y)\propto\exp\left(-\frac{\sum_i(y_i-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{\mu^2}{2}\right)$$

$$=\exp\left(-\frac{\sum_i[(y_i-\bar{\mu})+(\bar{\mu}-\mu)]^2}{2\sigma^2}\right)\exp\left(-\frac{\mu^2}{2}\right)$$

其中,$\bar{\mu}=n^{-1}\sum_{i=1}^{n}y_i$。展开第一项的平方,注意:

$\sum_i[(y_i-\bar{\mu})+(\bar{\mu}-u)]^2=(\bar{\mu}-u)\sum_i(y_i-\bar{\mu})=0$,最后导出:

$$\mu|\sigma^2,y\sim N\left(\frac{n\bar{\mu}}{\sigma^2+n},\frac{\sigma^2}{\sigma^2+n}\right)\qquad(3)$$

为获得 σ^2 的全条件后验分布,从式(2)中挑选含有 σ^2 的项:

$$p(\sigma^2|\mu,y)\propto(\sigma^2)^{-(\frac{v}{2}+1)}\exp\left(-\frac{vS}{2\sigma^2}\right)(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{\sum_i(y_i-\mu)^2}{2\sigma^2}\right)$$

$$=(\sigma^2)^{-(\frac{n+v}{2}+1)}\exp\left(-\frac{vS+\sum_i(y_i-\mu)^2}{2\sigma^2}\right)$$

$$=(\sigma^2)^{-(\frac{v}{2}+1)}\exp\left(-\frac{vS}{2\sigma^2}\right)\qquad(4)$$

其中,$S=\frac{vS+\sum_i(y_i-\mu)^2}{v}$,$v=n+v$。式(4)为尺度反卡方分

基金项目　国家科技支撑计划"畜禽基因资源发掘与种质创新利用研究"(2006BAD13B08)。

作者简介　陈琦(1980-),女,山东蓬莱人,博士研究生,研究方向:动物遗传育种。* 通讯作者,博士,教授,博士生导师。

收稿日期　2008-11-10

布,自由度 v,尺度参数 $vS + \sum_i (y_i - \mu)^2$。要从式(4)中抽取一个样本,先从自由度 v 的卡方分布中抽样,对这个数求倒数,乘以 $vS = vS + \sum_i (y_i - \mu)^2$。

执行这个模型的 Gibbs 抽样,从式(3)中抽 μ,在式(4)中更新 μ 并抽取 σ^2,再返回式(3),更新 σ^2 并抽取 μ。如此循环,收敛时,样本 μ_i($i = 1, \cdots$,Gibbs 抽样数)看作来自边缘后验分布 $p(\mu | y)$,样本 $(\sigma^2)_i$ 看作来自边缘后验分布 $p(\sigma^2 | y)$ 的样本。

1.2 应用 模拟100 个母牛产犊数据,犊牛出生重28 ~40 kg,母牛产犊前体重450 ~650 kg,CBW CCW 比值0.043 ~0.089。将 Calving ease 分为4 类:正常的占11.4 %,难产不需要辅助的占58.99 %,难产需要辅助的占15.38 %,剖腹产的占14.23 %。每一类对应不同的 CBW/CCW 数据(先验信息),分别服从不同的正态分布,均值和方差分别为 μ_i 和 σ_i^2($i = 1, 2, 3, 4$)。通过 Gibbs 抽样,利用每个分类的 CBW CCW 数据,确定 Calving ease 4 个分类的正态分布的 μ_i 和 σ_i^2。再将4 个正态分布按照混合模型混合成一个正态分布,确定其参数。

利用 Excel 及 WinBUGS 1.4,对 μ_i 和 σ_i^2 进行抽样。

2 结果与分析

Gibbs 抽样结果如表1 和图1 所示。

表1 每个分类抽样所得的各参数

Table 1 The parameters in dassification sampling

参数名称 Parameter name	均值 Mean value	标准差 Standard deviation	MC 误差 MC error	中值 Median	迭代初始数 Iterative initial number	抽样数 Sampling number
μ_1	0.079 66	0.410 3	0.004 026	0.079 37	1001	10 000
S_1	1.398 00	0.372 7	0.003 577	1.334 00	1001	10 000
μ_2	0.136 6	0.278 0	0.002 724	0.136 90	1001	10 000
S_2	0.886 6	0.240 0	0.002 353	0.843 00	1001	10 000
μ_3	0.293 2	0.285 7	0.002 799	0.294 60	1001	10 000
S_3	0.913 3	0.247 4	0.002 421	0.868 40	1001	10 000
μ_4	0.440 3	0.292 4	0.002 865	0.442 20	1001	10 000
S_4	0.935 5	0.253 7	0.002 481	0.888 90	1001	10 000

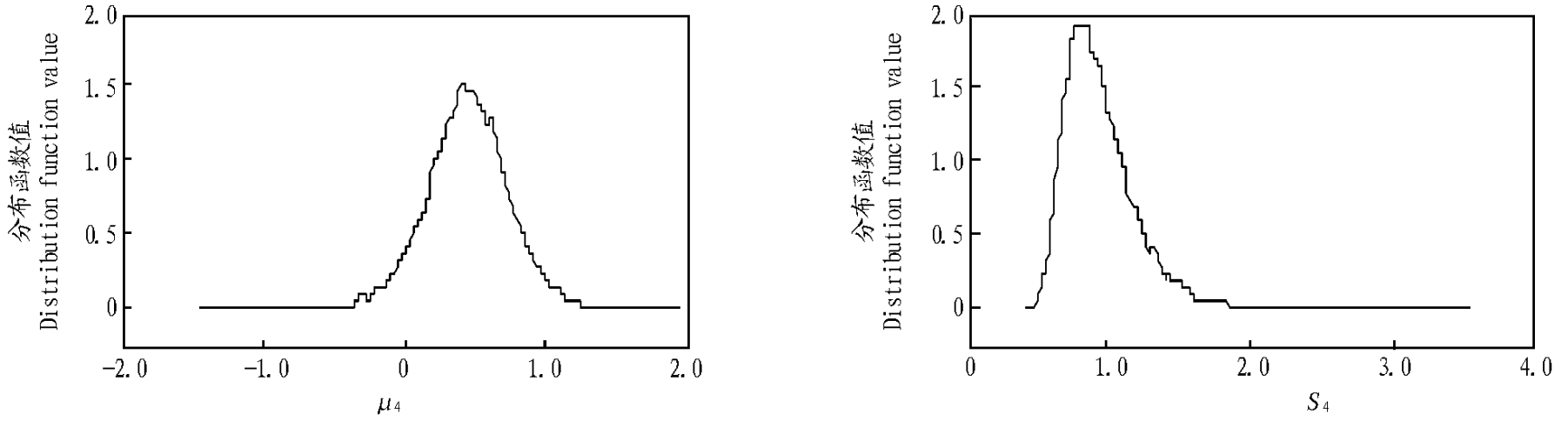

图1 第四类 Gibbs 抽样所得图表

Fig.1 The graph from Gibbs sampling of the fourth dassification

经过计算得出:$\mu_1 = 0.079\ 66$,$\sigma_1^2 = 1.954$;$\mu_2 = 0.136\ 6$,$\sigma_2^2 = 0.756$;$\mu_3 = 0.293\ 2$,$\sigma_3^2 = 0.834$;$\mu_4 = 0.440\ 3$,$\sigma_4^2 = 0.786$。根据混合分布模型理论计算,得出混合以后的参数 μ = 0.197 4,$\sigma^2 = 0.908$。

3 小结

该研究将 Gibbs 抽样用于 Calving ease 的研究,解决了阈模型在 Calving ease 研究中存在的某些问题。而且 Gibbs 抽样的优点是不依赖于初始数据,通过 Gibbs 抽样迭代过程,只要迭代过程足够大,最终结果收敛于真值[2]。这对于解决数据缺失及数据不准确的问题有重要意义。

参考文献

[1] SORENSEN D. Gibbs sampling in quantitative genetics[J]. Internal Report, 1996,82:119 - 122.

[2] CASELLA G, GEORGE E I. Explaining the Gibbs sampler[J]. The American Statistician, 1992,46(2):167 - 174.