DOI :10.3969/J. ISSN. 1672-6693. 2009. 04. 015

二维混合关联及算法*

邹辛程, 刘 坤
(常州工学院 理学院, 江苏 常州 213022)

摘要 :为了研究两个不同类型随机变量的关联及算法,提出了二维混合关联和二维混合单关联的概念,以及二维混合关联的两个基本模型。建构了二维混合单关联的两个基本模型。建构了二维混合型随机变量,给出了二维混合密度的精确定义,推出了求解关联变量自身分布的算法,对于关联模型 $Y(X)$,Y 的边缘密度为 $p_Y(y) = \sum_k p(x_k, y)$, $y \in \mathbf{R}$,对于关联模型 $X(Y)$,X 的边缘分布列为 $p_X(x_k) = \int_{\mathbf{R}} p(x_k, y)\,\mathrm{d}y$,$k = 1, 2, \cdots$。再利用重期望公式,推出求解了关联变量自身分布的数学期望的简化算式。

关键词 :自由变量 ;关联变量 ;二维混合单关联 ;二维混合型随机变量 ;混合密度
中图分类号 :O212.1 文献标识码 :A 文章编号 :1672-6693(2009)04-0074-04

一个离散型随机变量和一个连续型随机变量相关联是一个常见的现象。如对某一学生群体,可用离散型随机变量 X 描述学生的性别,用连续型随机变量 Y 描述学生的身高,显然 X 和 Y 是相关联的。本文称两个不同类型随机变量的关联为二维混合关联,下面将讨论二维混合关联的基本模型、研究目标、基本算法和数字特征。

1 基本模型

定义 1[1-8] 设 $X = X(e)$, $Y = Y(e)$ 是定义在同一个样本空间 $S = \{e\}$ 上两个随机变量,不妨设 X 为一个为离散型随机变量,Y 为一个为连续型随机变量。如果一个随机变量的取值不受另一个随机变量取值的影响,则称这个随机变量为自由变量,否则称为关联变量。当 X, Y 其中一个是自由变量,另一个是关联变量时,称这种二维混合关联为二维混合单关联。

关联模型约定为 $Y(X)$ 表示 X 是自由变量,Y 是关联变量 ;$X(Y)$ 表示 Y 是自由变量,X 是关联变量。

二维混合单关联是二维混合关联的基本形式。事实上,当 X, Y 都是关联变量时,在没有第三方信息的条件下,即为循环式关联,此时是无法获得 X 与 Y 之间关联的确切算法的。

2 研究目标

本文的主要研究目标是,对于二维混合单关联,在已知自由变量的分布和关联变量的条件分布情况下,如何使关联变量解除与自由变量的关联,而获得其自身分布的基本算法。

例 1 用 X 描述某一学生群体学生的性别,定义 $X = 1$ 表示男生,$X = 0$ 表示女生,则 X 服从 $B(1, p)$。用 Y 描述该学生群体学生的身高,定义男生身高和女生身高的概率密度分别为 $p_1(y)$, $p_2(y)$,试求 Y 的概率分布密度。

分析 X 为离散型自由变量,由已知身高 Y 的取值(概率密度)依赖性别 X 的取值,则 Y 为连续型关联变量,其关联模型为 $Y(X)$。

事实上,概率密度 $p_1(y)$ 是在“$X = 1$”的条件下 Y 的条件密度,即 $p_1(y) = p_Y(y \mid X = 1)$;概率密度 $p_2(y)$ 是在“$X = 0$”的条件下 Y 的条件密度,即 $p_2(y) = p_Y(y \mid X = 0)$。所以,例 1 是在已知自由变量 X 和关联变量 Y 的条件分布下,求解 Y 自身分布的问题。

* 收稿日期 :2009-03-28 修回日期 :2009-04-28
作者简介 :邹辛程,男,副教授,研究方向为概率统计。

3 基本算法

求解关联变量的自身分布,采用通常的数学方法,将自由变量和关联变量合成一个二维随机变量 (X,Y) 来加以考虑,这里称此 (X,Y) 为二维混合型随机变量[1],然后借鉴一般二维随机变量关于边缘分布的算法,达到求解关联变量自身分布的目标。

3.1 关联变量的条件分布

对于关联模型 $Y(X)$,记自由变量 X 的分布列为 $p_X(x_k)=P(X_k=x_k),k=1,2,\cdots$,且当 $p_X(x_k)>0$ 时,记关联变量 Y 的条件分布密度为 $p_Y(y|x_k)=p_Y(y|X=x_k),y\in\mathbf{R}$。

对于关联模型 $X(Y)$,记自由变量 Y 的分布密度为 $p_Y(y),y\in\mathbf{R}$,且当 $p_Y(y)>0$ 时,记关联变量 X 的条件分布列为 $p_X(x_k|y)=p(X=x_k|Y=y),k=1,2,\cdots$。

3.2 二维混合密度

为了简便起见,对二维混合型随机变量 (X,Y) 的概率密度采用以下定义方式给出。

定义2 对于关联模型 $Y(X)$,若关联变量 Y 的条件分布密度 $p_Y(y|x_k)$ 存在,则称

$$p(x_k,y)=p_X(x_k)p_Y(y|x_k),k=1,2,\cdots,y\in\mathbf{R}$$

为二维混合型随机变量 (X,Y) 的混合密度。

对于关联模型 $X(Y)$,若关联变量 X 的条件分布列 $p_X(x_k|y),k=1,2,\cdots$ 存在,则称

$$p(x_k,y)=p_Y(y)p_X(x_k|y),k=1,2,\cdots,y\in\mathbf{R}$$

为二维混合型随机变量 (X,Y) 的混合密度。

可以验证,无论是 $Y(X)$ 或 $X(Y)$,(X,Y) 的混合密度都有

1) 非负性:$p(x_k,y)\geqslant 0$;

2) 规范性:$\sum\limits_k\int\limits_{\mathbf{R}}p(x_k,y)\,\mathrm{d}y=1$ 或 $\int\limits_{\mathbf{R}}\sum\limits_k p(x_k,y)\,\mathrm{d}y=1$。

3.3 基本算法

显然,关联变量的自身分布,就是二维混合型随机变量 (X,Y) 关于关联变量的边缘分布。

定理1 对于关联模型 $Y(X)$,Y 的边缘密度为 $p_Y(y)=\sum\limits_k p(x_k,y),y\in\mathbf{R}$。对于关联模型 $X(Y)$,X 的边缘分布列为 $p_X(x_k)=\int\limits_{\mathbf{R}}p(x_k,y)\,\mathrm{d}y,k=1,2,\cdots$。

证明 对于 $Y(X)$ 有 $F_Y(y)=P(Y\leqslant y)=P(-\infty<X<+\infty,Y\leqslant y)=\int_{-\infty}^{y}\sum\limits_k p(x_k,t)\,\mathrm{d}t$,两边对 y 求导,得 $p_Y(y)=\sum\limits_k p(x_k,y),y\in\mathbf{R}$。

对于 $X(Y)$ 有 $p_X(x_k)=P(X=x_k,-\infty<Y<+\infty)=\int\limits_{\mathbf{R}}p(x_k,y)\,\mathrm{d}y,k=1,2,\cdots$。 证毕

例2 试求例1中 Y 的概率分布密度。

解 由定义二得关联模型 $Y(X)$,(X,Y) 的混合密度为

$$p(k,y)=p_X(k)p_Y(y|k)=\begin{cases}(1-p)\cdot p_2(y),k=0,y\in\mathbf{R}\\ p\cdot p_1(y),k=1,y\in\mathbf{R}\end{cases}$$

由定理1,有 $p_Y(y)=\sum\limits_{k=0}^{1}p_X(k)p_Y(y|k)=(1-p)\cdot p_2(y)+p\cdot p_1(y),y\in\mathbf{R}$。

例3 设 $Y\sim U(0,1)$,$X\sim \mathrm{B}(n,Y)$。试求1) (X,Y) 的混合密度;2) 关于 X 的边缘分布列。

解 关联变量 $X\sim \mathrm{B}(n,Y)$ 是指条件分布 $X|Y\sim \mathrm{B}(n,Y)$。

1) 由定义二的关联模型 $X(Y)$,(X,Y) 的混合密度为

$$p(k,y)=p_Y(y)p_X(k|y)=\begin{cases}\mathrm{C}_n^k y^k(1-y)^{n-k},0<y<1,k=0,1,\cdots,n\\ 0,y\leqslant 0\text{ 或 }y\geqslant 1\end{cases}$$

2) 由定理 1,X 的边缘分布列为

$$p_X(k) = \int_{-\infty}^{+\infty} p(k,y)\,\mathrm{d}y = \int_0^1 \mathrm{C}_n^k y^k (1-y)^{n-k}\mathrm{d}y = \mathrm{C}_n^k \int_0^1 y^k (1-y)^{n-k}\mathrm{d}y =$$

$$\mathrm{C}_n^k \mathrm{B}(k+1,n-k+1) = \frac{1}{n+1}, k = 0,1,\cdots,n$$

其中贝塔函数 $\mathrm{B}(k+1,n-k+1) = \dfrac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \dfrac{1}{(n+1)\mathrm{C}_n^k}$。

例 4　设随机变量 Y 服从参数为 θ 的指数分布 $E(\theta)$,随机变量 X 服从参数为 Y 的泊松分布 $P(Y)$。试求关于 X 的边缘密度。

解　由定义二得关联模型 $X(Y)$ 及定理 1,有

$$p_X(k) = \int_{-\infty}^{+\infty} p_Y(y)p(k|y)\,\mathrm{d}y = \int_0^{+\infty} \theta \mathrm{e}^{-\theta y}\frac{y^k}{k!}\mathrm{e}^{-y}\mathrm{d}y = \frac{\theta}{k!}\int_0^{+\infty} y^k \mathrm{e}^{-(\theta+1)y}\mathrm{d}y = \frac{\theta}{(\theta+1)^{k+1}}, k = 0,1,\cdots$$

上式运用 $\Gamma(k+1,\theta+1)$ 分布的规范性 $\int_0^{+\infty} \dfrac{(\theta+1)^k}{\Gamma(k+1)} y^k \mathrm{e}^{-(\theta+1)y}\mathrm{d}y = 1$ 所得。

以上各例表明,关联变量的条件分布与它自身的分布是类型相同的不同分布。

4 数字特征

求解关联变量自身分布的数字特征,可以类比二维随机变量的算法进行,这里从略。

一种形式简明、应用较多的混合单关联,即将自由变量作为关联变量分布中的常见参数使用(事实上,把分布中的未知参数,当作一个随机变量去看待,是贝叶斯学派的一个重要观点[2]),本文称这种关联为混合参数单关联。在混合参数单关联的情况下,定理 2 给出了求解关联变量自身数学期望的一个算法,将大大简化求解过程。

定理 2　1) 对于关联模型 $Y(X)$,若 $E(Y|X) = X$,则关联变量 Y 自身的数学期为 $E(Y) = E(X)$。

2) 对于关联模型 $X(Y)$,若 $E(X|Y) = Y$,则关联变量 X 自身的数学期望为 $E(X) = E(Y)$。

证明　由重期望公式[2],有 $E[E(Y|X)] = E(Y)$,$E[E(X|Y)] = E(X)$。再结合已知条件 $E(Y|X) = X$ 或 $E(X|Y) = Y$,即得 1)、2) 的结果。　证毕

例 5　根据以下混合单关联,求解关联变量自身的数学期望。1) 已知 $X \sim P(\lambda)$(泊松分布),$Y \sim N(X,\sigma^2)$。2) 已知 $Y \sim U(a,b)$(均匀分布),$(b > a > 0)$,$X \sim P(Y)$。

解　1) 由 $Y \sim N(X,\sigma^2)$,则 $E(Y|X) = X$,且 $E(X) = \lambda$。由定理 2 得 $E(Y) = E(X) = \lambda$。

2) 由 $X \sim P(Y)$,则 $E(X|Y) = Y$,且 $E(Y) = \dfrac{a+b}{2}$。由定理 2 得 $E(X) = E(Y) = \dfrac{a+b}{2}$。

5 结语

类似于两个同类型随机变量的关联,二维混合单关联有着广泛而且更灵活的应用。

例如,某厂有 n 种型号的大批电子产品,每种型号的寿命标记为 $x_k, k = 1,2,\cdots,n$,且知其分布列为 $P(X = x_k) = p_k, k = 1,2,\cdots,n$。又知寿命标记 x_k 的电子产品的实际使用寿命服从指数分布 $E(x_k), k = 1,2,\cdots,n$。今从中任取一种型号的电子元件,试求该电子元件的实际使用寿命 Y 的分布。显然这是一个二维混合单关联的应用问题,不难算得 Y 的分布密度 $p_Y(y) = \sum_{k=1}^{n} p_k x_k \mathrm{e}^{-x_k y}, y > 0$。假如又把它们看成随机生成过程中的一个“链结”,那么多个这样的有序“链结”,就会产生一个特殊的随机过程。

从例 3 中还看出,当自由变量 $Y \sim U(0,1)$ 作了二项分布 $X \sim \mathrm{B}(n,Y)$ 的第二个参数时,使得 X 的性质发生了变化,即关联变量 X 的自身分布成为 $n+1$ 个点上的等可能分布了。因此,有目的地使分布参数随机化,可以改变关联变量的自身分布,实现对随机分布的生成与控制。

参考文献：

[1] 刘坤,邹辛程. 概率论与数理统计[M]. 南京:南京大学出版社,2009.

[2] 茆诗松,程依明,濮晓龙. 概率论与数理统计教程[M]. 北京:高等教育出版社,2004.

[3] 缪铨生,赵跃生,邹辛程. 概率与统计[M]. 南京:南京大学出版社,2000.

[4] 唐万宏,杨继龙,邹辛程,等. 概率统计[M]. 南京:南京师范大学出版社,2001.

[5] 盛骤,谢式千,潘承毅. 概率论与数理统计(第三版)[M]. 北京:高等教育出版社,2001.

[6] 孙激流,沈大庆. 概率论与数理统计[M]. 北京:首都经济贸易大学出版社,2005.

[7] 周概容. 概率论与数理统计[M]. 北京:高等教育出版社,1984.

[8] 邓永录. 应用概率及其理论基础[M]. 北京:清华大学出版社,2005.

Two Dimension Mixed Conjunctions and Their Arithmetic

ZOU Xin-cheng, LIU Kun

(School of Science, Changzhou Institute of Technology, Changzhou Jiangsu 213022, China)

Abstract: In order to research the two different types of random variants, and their arithmetic, this article raises the concept of two-dimensional mixed conjunctions, two-dimensional mixed single conjunctions, and two basic models of two-dimensional mixed single conjunctions. On the basis of determined research target, the author builds random variants of two-dimensional mixed conjunction, provides a precise definition on the density of two-dimensional mix, and works out the arithmetic and theorem of solving the distribution of the conjunction variants themselves. For the conjunction model $Y(X)$, the marginal density of Y $p_Y(y) = \sum_k p(x_k, y)$, $y \in \mathbf{R}$, and for the conjunction model $X(Y)$, the marginal distribution the row X is $p_X(x_k) = \int_{\mathbf{R}} p(x_k, y)\,\mathrm{d}y, k = 1, 2, \cdots$ The author then moves on to utilize repeatable formulas and works out simplified arithmetic for mathematic expectancy of solving the conjunction variants´distributions. The article exemplifies the use of various kinds of arithmetric and finally summarizes the applicaion of two dimensional mixed conjunctions by demonstrating an application example.

Key words: freeness variable; conjunction variable; two dimension mixed single conjunction; two dimension mixed random variable; mixed density

(责任编辑 黄 颖)