文章编号: 1004-4353(2016)03-0192-04

次序统计量概率密度函数的新的推导方法

孙婷婷， 洪义成*， 沈京虎， 金哲植
(延边大学理学院 数学系，吉林 延吉 133002)

摘要: 提出一种利用二项分布和微积分中的分部积分法来推导次序统计量概率密度函数的新方法,该方法不仅可以使学生在学习次序统计量的过程中对二项分布有更进一步地了解,而且也体现了数学思维的逻辑性、转换性、承接性和多样性.

关键词: 二项分布；次序统计量；概率密度函数
中图分类号: O221.3　　　**文献标识码:** A

A new derivation method for probability density function of order statistics

SUN Tingting， HONG Yicheng*， SHEN Jinhu， JIN Zhezhi
(*Department of Mathematics*, *College of Science*, *Yanbian University*, *Yanji* 133002, *China*)

Abstract: In this paper, we present a new derivation method for probability density function of order statistics by using binomial distribution and partial integral method in calculus. This derivation is not only can make the students have further knowledge of the binomial distribution in the process of learning order statistic, but also can reflect the logic, transformation, undertake and diversity of the mathematical thinking.

Keywords: binomial distribution; order statistics; probability density function

众所周知,次序统计量在数理统计上有着十分广泛的应用.然而,在讨论次序统计量的概率密度函数时,有关教程中都采用了分区间法或类似分区间法的概率元法来导出其概率密度函数[1-5].虽然文献[6]认为这种方法不够严谨,并对这种方法进行了相应的修改,但总的证明思路仍没有太大的变化.鉴于此,本文根据二项分布的定义,结合微积分中的分部积分法来推导次序统计量的概率密度函数,该方法不仅可以使学生在教学过程中可以回顾更多以往学过的概率分布知识和内容,还能达到启发学生独立思考和培养学生数学思维逻辑的目的.

1　传统的次序统计量推导方法

1.1　分区间法

设总体 X 的累积分布函数与概率密度函数分别为 $F(x)$ 和 $f(x)$, $X_1, \cdots, X_n$ 是来自该总体的简单随机样本, $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为其次序统计量,第 k 个次序统计量 $X_{(k)}$ 的累计分布函数与概率密度函数

收稿日期: 2016 - 09 - 01
* **通信作者:** 洪义成(1980—),男,理学博士,副教授,研究方向为金融统计学、保险精算学.

分别记作 $F_k(x)$ 与 $f_k(x)$. 对任意的实数 x，考虑次序统计量 $X_{(k)}$ 取值落在小区间$(x, x+\Delta x]$内这一事件，它等价于“样本容量为 n”的样本中有一个观测值落在$(x, x+\Delta x]$之间，而有 $k-1$ 个落入$(-\infty, x]$内，其余 $n-k$ 个落入$(x+\Delta x, \infty)$内.

样本中观测值落于$(-\infty, x]$内的概率为 $F(x)$，落入区间$(x, x+\Delta x]$内的概率为 $F(x, x+\Delta x)-F(x)$，落入区间$(x+\Delta x, \infty)$内的概率为 $1-F(x, x+\Delta x)$. 将 n 个观测值分成这样的 3 组，则其总的分法共有$\frac{n!}{(k-1)!\ 1!\ (n-k)!}$种. 于是，由多项分布可得：

$$F_k(x+\Delta x)-F_k(x) \approx \frac{n!}{(k-1)!\ (n-k)!} F(x)^{k-1}(F(x+\Delta x)-F(x))(1-F(x+\Delta x))^{n-k}. \tag{1}$$

对式(1)两边除以 Δx，并令 $\Delta x \to 0$，即有：

$$f_k(x)=\lim_{\Delta x} \frac{F_k(x+\Delta x)-F_k(x)}{\Delta x}=\frac{n!}{(k-1)!\ (n-k)!} F(x)^{k-1} f(x)(1-F(x))^{n-k}. \tag{2}$$

1.2　概率元法

如果连续随机变量落在很小的区间$(x, x+\mathrm{d}x]$内的概率为：

$$P(x<X \leqslant x+\mathrm{d}x)=f(x)\mathrm{d}x+o(\mathrm{d}x), \tag{3}$$

其中 $o(\mathrm{d}x)$ 是比 $\mathrm{d}x$ 高阶的无穷小量，可知左端概率的主要部分就是 $f(x)\mathrm{d}x$，$f(x)\mathrm{d}x$ 称为是 X 的概率元. 同样，若存在函数 $f(x)$ 使式(3)成立，则 $f(x)$ 就是 X 的密度函数. 这种寻求密度函数的方法称为概率元法. 下面利用概率元法求单个的次序统计量.

设总体 X 的累积分布函数与概率密度函数分别为 $F(x)$ 和 $f(x)$，$X_1, \cdots, X_n$ 是来自该总体的简单随机样本，$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为其次序统计量，它们的观测值依次记为 $y_1 \leqslant \cdots \leqslant y_n$. 第 k 个次序统计量 $X_{(k)}$ 的累计分布函数与概率密度函数分别记作 $F(y_k)$ 与 $f(y_k)$，它的观测值为 y_k. 以 y_k 为基础把实数轴分为 3 个区间：$(-\infty, y_k)$，$[y_k, y_k+\mathrm{d}y_k)$，$[y_k+\mathrm{d}y_k, \infty)$，其中第 2 个区间的长度 $\mathrm{d}y_k$ 很小，使得样本观测值中只有一个落入该区间，而两个或更多个观测值落入该区间的概率为零或为 $o(\mathrm{d}y_k)$，这里只要 $\mathrm{d}y_k$ 充分小就可以. 这样，要使 $X_{(k)}$ 的观测值落入$[y_k, y_k+\mathrm{d}y_k)$内，就要使样本的 n 个观测值中有 $k-1$ 个落入$(-\infty, y_k)$内，其余 $n-k$ 个落入$[y_k+\mathrm{d}y_k, \infty)$内. 根据多项分布，可得 $X_{(k)}$ 的概率元为：

$$f(y_k)\mathrm{d}y_k=\frac{n!}{(k-1)!\ 1!\ (n-k)!}[F(y_k)]^{k-1}[f(y_k)\mathrm{d}y_k][1-F(y_k+\mathrm{d}y_k)]^{n-k}+o(\mathrm{d}y_k). \tag{4}$$

式(4)两边约去 $\mathrm{d}y_k$ 后，再让 $\mathrm{d}y_k \to 0$，即得 $X_{(k)}$ 的概率密度函数：

$$f(y_k)=\frac{n!}{(k-1)!\ (n-k)!}[F(y_k)]^{k-1}[1-F(y_k)]^{n-k} f(y_k). \tag{5}$$

2　次序统计量密度函数的新的推导方法

设总体 X 的累积分布函数与概率密度函数分别为 $F(x)$ 和 $f(x)$，$X_1, \cdots, X_n$ 是来自该总体的简单随机样本，$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为其次序统计量. 第 k 个次序统计量 $X_{(k)}$ 的累计分布函数与概率密度函数分别记作 $F_k(x)$ 与 $f_k(x)$，则由次序统计量的定义可知，对 $\forall x \in \mathbf{R}$ 有：

$$F_k(x)=P(X_{(k)} \leqslant x)=P(x_1, \cdots, x_n \text{ 当中至少有 } k \text{ 个小于等于 } x)=$$

$$\sum_{j=k}^{n} P(x_1, \cdots, x_n \text{ 当中正好有 } j \text{ 个小于等于 } x). \tag{6}$$

其中 $x_1, \cdots, x_n$ 是 n 个随机样本 $X_1, \cdots, X_n$ 的观测值.

因事件$\{x_1, \cdots, x_n$ 当中正好有 j 个小于等于 $x\}$ 发生的概率可看作 n 重贝努利试验中成功 j 次的概

率,故式(6)可改写成

$$F_k(x)=\sum_{j=k}^{n}P(x_1,\cdots,x_n\text{ 当中正好有 }j\text{ 个小于等于 }x)=\sum_{j=k}^{n}C_n^j[F(x)]^j[1-F(x)]^{n-j}=$$
$$C_n^k[F(x)]^k[1-F(x)]^{n-k}+\cdots+C_n^{n-1}[F(x)]^{n-1}[1-F(x)]+[F(x)]^n,\tag{7}$$

而式(7)中最后两项可写成

$$C_n^{n-1}[F(x)]^{n-1}[1-F(x)]+[F(x)]^n=\frac{n!}{(n-1)!\ 1!}[F(x)]^{n-1}[1-F(x)]+$$
$$\frac{n!}{n(n-1)!\ 1!}[F(x)]^n=\frac{n!}{(n-1)!\ 1!}\left\{[F(x)]^{n-1}[1-F(x)]+\int_0^{F(x)}t^{n-1}(1-t)^0\mathrm{d}t\right\}=$$
$$\frac{n!}{(n-1)!\ 1!}\left\{[F(x)]^{n-1}[1-F(x)]-\int_0^{F(x)}t^{n-1}\mathrm{d}(1-t)\right\},\tag{8}$$

根据分部积分法,式(8)可写成

$$C_n^{n-1}[F(x)]^{n-1}[1-F(x)]+[F(x)]^n=\frac{n!}{(n-2)!\ 1!}\int_0^{F(x)}t^{n-2}(1-t)\mathrm{d}t=$$
$$-\frac{n!}{(n-2)!\ 2!}\int_0^{F(x)}t^{n-2}\mathrm{d}(1-t)^2,\tag{9}$$

再根据分部积分法,式(9)与式(7)中的倒数第 3 项之和可改写成

$$C_n^{n-2}[F(x)]^{n-2}[1-F(x)]^2-\frac{n!}{(n-2)!\ 2!}\int_0^{F(x)}t^{n-2}\mathrm{d}(1-t)^2=$$
$$\frac{n!}{(n-3)!\ 2!}\int_0^{F(x)}t^{n-3}(1-t)^2\mathrm{d}t=-\frac{n!}{(n-3)!\ 3!}\int_0^{F(x)}t^{n-3}\mathrm{d}(1-t)^3.$$

继续这一过程,直到式(7)写成一个如下积分式子:

$$F_k(x)=C_n^k[F(x)]^j[1-F(x)]^{n-j}+\cdots+C_n^{n-1}[F(x)]^{n-1}[1-F(x)]+[F(x)]^n=$$
$$C_n^k[F(x)]^j[1-F(x)]^{n-j}-\frac{n!}{k!\ (n-k)!}\int_0^{F(x)}t^k\mathrm{d}(1-t)^{n-k}=$$
$$\frac{n!}{(k-1)!\ (n-k)!}\int_0^{F(x)}t^{k-1}(1-t)^{n-k}\mathrm{d}t,\tag{10}$$

式(10)两边对 x 求导,可得

$$\frac{\mathrm{d}}{\mathrm{d}x}F_k(x)=\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{n!}{(k-1)!\ (n-k)!}\int_0^{F(x)}t^{k-1}(1-t)^{n-k}\mathrm{d}t\right)=$$
$$\frac{n!}{(k-1)!\ (n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x),$$

即

$$f_k(x)=\frac{n!}{(k-1)!\ (n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x).\tag{11}$$

这一结果与文献[1-5]中关于单个次序统计量概率密度函数的结果完全一致.

3 学习效果的比较及结论

在教学过程中,使用分区间和概率元方法进行推导虽然简洁,但是得到概率密度函数的过程不够严谨,并且也没有把连续随机变量的分布函数与概率密度函数紧密地联系起来,不能起到温故知新的作

用.而使用本文给出的推导方法，可以通过次序统计量的定义导出式(6)，这样既可加深学习者对次序统计量的理解，同时也可以回顾事件的表示法和概率的可列可加性性质.进一步地，通过把事件$\{x_1,\cdots,x_n$ 当中正好有 j 个小于等于 $x\}$ 与贝努利概型进行比较，导出式(7)，这一过程不仅可以启发学习者如何应用好贝努利概型，也可以使学习者对二项分布有进一步的了解.

本文利用微积分中的分部积分法来导出式(8)和式(9)，然后再利用迭推关系导出式(10)，这正是本文方法区别于分区间法和概率元法的关键.另外，在推导过程中，本文还得出如下有用的恒等式：

$$\sum_{k=m}^{n} C_n^k p^k (1-p)^{n-k} = mC_n^m \int_0^p t^{m-1}(1-t)^{n-m}\,\mathrm{d}t. \tag{12}$$

由式(12)发现，其左侧为 n 重贝努利试验中试验成功的次数不小于 m 次的概率，而其右侧为不完全贝塔函数.

最后再对式(10)求导得出式(11)，由此可使学习者对分布函数与概率密度函数间的关系有更进一步的认识.

综上所述，本文方法在整个推导过程中不仅没有过于复杂的数学逻辑，而且也不失数学上严密的思维逻辑性.这种推导方式不仅可以给攻读数学类或统计专业的学生非常直观的印象，而且还可以教会学生数学逻辑的思维转换性和思维多样性.

参考文献：

[1]　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，2007：259-264.

[2]　魏宗舒.概率论与数理统计教程[M].2 版.北京：高等教育出版社，2013：247-256.

[3]　茆诗松，王静龙，濮晓龙.高等数理统计[M].北京：高等教育出版社，2004：34-39.

[4]　盛骤，谢式千，潘承毅.概率论与数理统计[M].4 版.北京：高等教育出版社，2008：34-39.

[5]　茆诗松，王静龙.数理统计[M].上海：华东师范大学出版社，1986：20-25.

[6]　李秀兰.对教材中有关次序统计量分布证明的一点看法[J].雁北师范学院学报，2000，16(6)：26-27.

[7]　刘玉胜，田丽娜，王志林，等.利用概率元法推证次序统计量的密度函数[J].甘肃高师学报，2015，20(5)：12-14.

[8]　Sun Zhongfeng，Qin Huizeng，Li Aijuan. Extension of the partial derivatives of the incomplete beta function for complex values[J]. Applied Mathematics and Computation，2016，275：63-71.