# 第一章： 次序统计量

王 成
http://math.sjtu.edu.cn/faculty/chengwang/

上海交通大学数学系

2015-10-8

## 1 学习目标

1. 了解次序统计量的定义以及极大、极小、中位数、极差等几个关键次序统计量；
2. 学会计算次序统计量的分布以及次序统计量线性函数分布的计算；
3. 学习次序统计量相关分布的大样本性质，包括极值分布，中位数分布等；
4. 可以用次序统计量构造一些统计应用并解释原理。

## 2 背景介绍

### 2.1 定义

给定一组样本 $X_1, \ldots, X_n$，将其按照大小排序得到的新的样本$X_{(1)} \leq \cdots \leq X_{(n)}$ 称为样本 $X_1, \ldots, X_n$的次序统计量。

**思考：** 次序统计量是随机变量嘛？ 给定一组样本，如何计算某一个次序统计量 $X_{(k)}$？简单了解数据排序算法。

### 2.2 重要的次序统计量

极大： $X_{(n)} = max(X_1, \ldots, X_n)$
极小： $X_{(n)} = min(X_1, \ldots, X_n)$
中位数： $n$为奇数时$X_{((n+1)/2)}$， 偶数时： $(X_{(n/2)} + X_{(n/2+1)})/2$.
极差： $X_{(n)} - X_{(1)}$.
其中，极值，以及奇数情形下的中位数都是单个的次序统计量，而偶数情形的中位数和极差是组合形式的次序统计量。

### 2.3 常见应用

次序统计量在社会经济中经常以各种不同的形式出现，例如”百年不遇的……”“前十大……” “世界首富”等等。下面我们以一个NBA球队工资为例来看下次序统计量的应用：

$$2297, 1969, 1641, 964, 899, 677, 500, 495, 210, 150, 150, 128, 118, 115, 98, 98, 84.5, 84.5.$$

样本平均： 593. 极大：2297；极小：84.5；中位数： 180 极差： 2212.5.

## 3 基本分布

假定$X_1, \ldots, X_n$是从一个分布为$F$的总体中抽取的iid样本，本节我们考虑次序统计量的相关分布。

### 3.1 $X_{(k)}$ 的分布

记$X_{(k)}$的分布函数为$F_k$，我们有

$$\begin{aligned} F_k(x) = P(X_{(k)} \le x) &= P(X_1, \ldots, X_n \text{中至少有k个不大于x}) \\ &= \sum_{r=k}^{n} P(X_1, \ldots, X_n \text{中r个不大于x, n-r个大于x}) \\ &= \sum_{r=k}^{n} C_n^r (P(X_1 \le x))^r (P(X_1 > x))^{n-r} \\ &= \sum_{r=k}^{n} C_n^r F^r(x)(1-F(x))^{n-r}. \end{aligned}$$

利用恒等式：

$$\sum_{j=r}^{n} C_n^r p^j (1-p)^{n-j} = \frac{n!}{(r-1)!(n-r)!} \int_0^p t^{r-1}(1-t)^{n-r} dt, \ \ r = 1, \ldots, n, \ \ 0 \le p \le 1,$$

我们有

$$F_k(x) = P(X_{(k)} \le x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{r-1}(1-t)^{n-r} dt.$$

如果$F(x)$的密度函数存在，记为$f(x)$，对应的我们有$X_{(k)}$的密度函数：

$$f_k(x) = F_k'(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x). \qquad (1)$$

特别的，对于两个极值，我们有

$$\begin{aligned} &X_{(1)} : F_1(x) = 1 - (1-F(x))^n, \ \ f_1(x) = n(1-F(x))^{n-1} f(x); \\ &X_{(n)} : F_n(x) = F^n(x), \ \ f_n(x) = nF^{n-1}(x) f(x). \end{aligned}$$

Remark 3.1 如果$X_1$的期望存在，即$\int |x| dF(x) < \infty$，那么$X_{(k)}$的期望也一定存在。可以从其分布函数出发推导，也可以直接由下面的不等式得到：

$$|X_{(k)}| \leq |X_1| + \cdots + |X_n|.$$

当然，我们也可以像一般的随机变量一样，研究次序统计量的期望，方差，特征函数，生成函数等等，稍后我们用均匀分布的例子来看看次序统计量的统计特征。

## 3.2 联合分布

对于任意的$k < j$，我们研究 $(X_{(k)}, X_{(j)})$的联合分布。对于任意的$x \leq y$,

$$\begin{aligned}
F_{kj}(x,y) &= P(X_{(k)} \leq x, X_{(j)} \leq y) \\
&= P(X_1, \ldots, X_n \text{中至少有k个不大于x, j个不大于y}) \\
&= \sum_{r=j}^{n} P(X_1, \ldots, X_n \text{中r个不大于y, n-r个大于y, 至少k个不大于x}) \\
&= \sum_{r=j}^{n} \sum_{i=k}^{r} P(X_1, \ldots, X_n \text{中i个不大于x, r-i个大于x不大于y，n-r个大于y}) \\
&= \sum_{r=j}^{n} \sum_{i=k}^{r} C_n^r C_r^i F(x)^i (F(y) - F(x))^{r-i} (1 - F(y))^{n-r} \\
&= \sum_{r=j}^{n} \sum_{i=k}^{r} \frac{n!}{i!(r-i)!(n-r)!} F(x)^i (F(y) - F(x))^{r-i} (1 - F(y))^{n-r}.
\end{aligned}$$

如果$F(x)$的密度函数$f(x)$存在，我们有$(X_{(k)}, X_{(j)})$密度函数：

$$f_{kj}(x,y) = \frac{n!}{(k-1)!(j-k-1)!(n-j)!} F(x)^{k-1} (F(y) - F(x))^{j-k-1} (1 - F(y))^{n-j} f(x) f(y).$$

对于$x > y$的情形，$P(X_{(k)} \leq x, X_{(j)} \leq y) = P(X_{(j)} \leq y)$ 退化到单个的情形，对应的密度函数为0.

Remark 3.2 从这里我们可以看出，即使对于iid的$X_1, \ldots, X_n$, $X_{(k)}$与$X_{(j)}$也是不独立的.

Example 3.1 对于极值$(X_{(1)}, X_{(n)})$的联合分布，我们有

$$F_{1n}(x,y) = \sum_{i=1}^{n} \frac{n!}{i!(n-i)!} F(x)^i (F(y) - F(x))^{n-i} = F^n(y) - (F(y) - F(x))^n.$$

如果$F(x)$的密度函数$f(x)$存在，我们有$(X_{(1)}, X_{(n)})$密度函数：

$$f_{1n}(x,y) = n(n-1)(F(y) - F(x))^{n-2} f(x) f(y).$$

Remark 3.3 全体次序统计量$(X_{(1)},\ldots,X_{(n)})$的联合密度函数为：

$$f_{1\ldots n}(x_1,\ldots,x_n)=n!f(x_1)\ldots f(x_n),\ \ x_1\le x_2\le\cdots\le x_n. \tag{2}$$

上述提及的所有单个的或者联合的本质上都可以通过对全体的联合密度函数积分得到。

### 3.3 极差和样本分位数的分布

### 3.4 条件分布

## 4 均匀分布$U(0,1)$的情形

对于任意的随机变量$X$，记其分布函数为$F$，我们首先研究随机变量$Y:=F(X)$的分布，对于任意的$0\le y\le 1$，

$$P(Y\le y)=P(F(X)\le y)=P(X\le F^{-1}(y))=F(F^{-1}(y))=y. \tag{3}$$

如果$F(x)$是严格增函数，上述推导没有问题。但是我们知道并不是所有分布函数都是严格增函数的，例如离散分布等。

Example 4.1 $X$是Bernoulli分布，我们有

$$F(x)=\begin{cases}0 & \text{if } x<0;\\ \frac{1}{2} & \text{if } 0\le x<1;\\ 1 & \text{if } x\ge 1.\end{cases}$$

和

$$F^{-1}(y)=\inf_x\{x:F(x)\ge y\}=\begin{cases}-\infty & \text{if } y<0;\\ 0 & \text{if } 0\le y<\frac{1}{2};\\ 1 & \text{if } \frac{1}{2}\le y<1;\\ \infty & \text{if } y\ge 1.\end{cases}$$

和

$$F(F^{-1}(y))=\begin{cases}0 & \text{if } y<0;\\ \frac{1}{2} & \text{if } 0\le y<\frac{1}{2};\\ 1 & \text{if } \frac{1}{2}\le y<1;\\ 1 & \text{if } y\ge 1.\end{cases}$$

而实际上，$Y=F(X)$的分布为一个$(\frac{1}{2},1)$的两点分布。

Theorem 4.1 设随机变量$X$的分布函数$F$处处连续，则

$$Y:=F(X)\sim U(0,1). \tag{4}$$

Proof: 记 $F^{-1}(y)=\inf\{x:F(x)\geq y\}$. 对于 $0<y<1$,

$$\begin{aligned}P(Y\leq y)=P(F(X)\leq y)&=P(X\in\{x:F(x)\leq y\})\\&=P(X\leq F^{-1}(y))\ \text{这里用到}F(x)\text{的连续性质}\\&=F(F^{-1}(y))=y.\end{aligned}$$

由此定理，对于处处连续的总体分布函数，研究次序统计量 $X_{(1)},\ldots,X_{(n)}$，可以等价的研究

$$(F(X_{(1)}),\ldots,F(X_{(n)}))\overset{d}{=}(U_{(1)},\ldots,U_{(n)}),\tag{5}$$

这里， $U_{(1)},\ldots,U_{(n)}$是来自$U(0,1)$的iid样本$U_1,\ldots,U_n$的次序统计量。 基于之前的结果，我们可以写出$U_{(1)},\ldots,U_{(n)}$的相关分布如下：

1. $U(k)$的分布函数：

$$F_k=P(U_{(k)}\leq x)=\frac{n!}{(k-1)!(n-k)!}\int_0^x t^{r-1}(1-t)^{n-r}dt,\ 0\leq x\leq 1.$$

密度函数：

$$f_k(x)=\frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}I_{(0,1)}(x).\tag{6}$$

2. 联合分布函数

$$F_{kj}(x,y)=\sum_{r=j}^{n}\sum_{i=k}^{r}\frac{n!}{i!(r-i)!(n-r)!}x^i(y-x)^{r-i}(1-y)^{n-r}.$$

联合密度函数：

$$f_{kj}(x,y)=\frac{n!}{(k-1)!(j-k-1)!(n-j)!}x^{k-1}(y-x)^{j-k-1}(1-y)^{n-j},\ 0<x<y<1.$$

3. 全体次序统计量的密度函数

$$f_{1\ldots n}(x_1,\ldots,x_n)=n!,\ 0<x_1\leq x_2\leq\cdots\leq x_n<1.\tag{7}$$