1.4 次序统计量及其分布

一、次序统计量

二、样本中位数和样本极差

一、次序统计量

1、定义1.12 次序统计量

设$(X_1, X_2, \cdots, X_n)^T$是从总体$X$中抽取的一个样本，$(x_1, x_2, \cdots, x_n)^T$是其一个观测值，将观测值按由小到大的次序重新排列为

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

当$(X_1, X_2, \cdots, X_n)^T$取值为$(x_1, x_2, \cdots x_n)^T$时，定义$X_{(k)}$取值为$x_{(k)}(k = 1, 2, \cdots n)$，由此得到

$$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})^T$$

称为样本$(X_1, X_2, \cdots, X_n)^T$的**次序统计量**.

对应的$(x_{(1)}, x_{(2)}, \cdots x_{(n)})$称为其观测值.

注 由于每个$X_{(k)}$都是样本$(X_1, X_2, \cdots, X_n)$的函数，所以$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$也都是随机变量。

？ $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$是否相互独立？

2、次序统计量的性质

次序统计量是充分统计量！

证明略

3、次序统计量的分布（针对连续总体讨论）

定理1.19 设总体X的分布密度为$f(x)$(或分布函数为$F(x)$), $X_1, X_2, \cdots, X_n$为来自总体X的样本，则第k个次序统计量$X_{(k)}$的分布密度为

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k} f(x),$$

$$k = 1, 2, \cdots, n.$$

证 设$v_n(x)$ $(-\infty < x < +\infty)$表示 $x_1, x_2, \cdots, x_n$中不超过 于 x的个数. $v_n(x) \sim B(n, F(x))$,因此

$$\{V_n(x) = k\} = \{X_{(k)} \le x < X_{(k+1)}\}, k = 1, 2, \cdots, n-1,$$

$$\{V_{(n)}(x) = n\} = \{X_{(n)} \le x\}$$

至少有k个样本值不超过x

则有$\sum_{i=k}^{n}\{V_n(x)=i\}=\sum_{i=k}^{n-1}\{X_{(i)}\le x<X_{(i+1)}\}+\{X_{(n)}\le x\}=\{X_{(k)}\le x\}$

$$=\{X_{(k)}\le x<X_{(k+1)}\}+\{X_{(k+1)}\le x<X_{(k+2)}\}+\cdots$$
$$+\underline{\{X_{(n-1)}\le x<X_{(n)}\}}+\underline{\{X_{(n)}\le x\}}$$

$$=\{X_{(k)}\le x<X_{(k+1)}\}+\{X_{(k+1)}\le x<X_{(k+2)}\}+\cdots$$
$$+\{X_{(n-2)}\le x<X_{(n-1)}\}+\{X_{(n-1)}\le x\}$$

$$=\{X_{(k)}\le x<X_{(k+1)}\}+\{X_{(k+1)}\le x<X_{(k+2)}\}+\cdots$$
$$+\{X_{(n-3)}\le x<X_{(n-2)}\}+\{X_{(n-2)}\le x\}$$

$$=\cdots=\{X_{(k)}\le x\}.$$

于是，$X_{(k)}$的分布函数为

$$F_{X_{(k)}}(x)=P\{X_{(k)}\le x\}=P\{\sum_{i=k}^{n}\{V_n(x)=i\}\}=\sum_{i=k}^{n}P\{V_n(x)=i\}$$

$$=\sum_{i=k}^{n}C_n^i[F(x)]^i[1-F(x)]^{n-i}$$

$$=\frac{n!}{(k-1)!(n-k)!}\int_0^{F(x)}t^{k-1}(1-t)^{n-k}\mathrm{d}t \quad (\text{利用分部积分})$$

后面给出详细推导

因此

$$f_{X_{(k)}}(x)=\frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

$$=kC_n^k[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

注：由

$$f_{X_{(k)}}(x) = kC_n^k[F(x)]^{k-1}[1-F(x)]^{n-k} f(x)$$

得

(1) 最大次序统计量$X_{(n)}$的分布密度为

$$f_{X_{(n)}}(x) = n[F(x)]^{n-1} f(x)$$

(2) 最小次序统计量$X_{(1)}$的分布密度为

$$f_{X_{(1)}}(x) = n[1-F(x)]^{n-1} f(x)$$

推导 $$\frac{n!}{(k-1)!(n-k)!}\int_0^{F(x)} t^{k-1}(1-t)^{n-k}\,\mathrm{d}t$$

$$=\sum_{i=k}^{n} C_n^i [F(x)]^i [1-F(x)]^{n-i}$$

$$=\frac{n!}{(k-1)!(n-k)!}\frac{1}{k}\int_0^{F(x)}(1-t)^{n-k}\,\mathrm{d}t^k$$

$$=\frac{n!}{(k-1)!(n-k)!}\frac{1}{k}\left[(1-t)^{n-k}t^k\Big|_0^{F(x)}\right.$$

$$\left.+(n-k)\int_0^{F(x)} t^k(1-t)^{n-k-1}\,\mathrm{d}t\right]$$

$$=\frac{n!}{k!(n-k)!}[F(x)]^k[1-F(x)]^{n-k}$$

$$+\frac{n!}{k!(n-k-1)!}\int_0^{F(x)} t^k(1-t)^{n-k-1}\,\mathrm{d}t$$

巧妙地逆推导

$$=\frac{n!}{k!(n-k)!}\left[F(x)\right]^{k}\left[1-F(x)\right]^{n-k}$$

$$+\frac{n!}{k!(n-k-1)!}\int_{0}^{F(x)}t^{k}(1-t)^{n-k-1}\mathrm{d}t$$

$$=C_{n}^{k}\left[F(x)\right]^{k}\left[1-F(x)\right]^{n-k}$$

$$+\frac{n!}{k!(n-k-1)!}\frac{1}{k+1}\left[(1-t)^{n-k-1}t^{k+1}\Big|_{0}^{F(x)}\right.$$

$$\left.+(n-k-1)\int_{0}^{F(x)}t^{k+1}(1-t)^{n-k-2}\mathrm{d}t\right]$$

$$=C_n^k\left[F(x)\right]^k\left[1-F(x)\right]^{n-k}$$

$$+C_n^{k+1}\left[F(x)\right]^{k+1}\left[1-F(x)\right]^{n-k-1}$$

$$+\frac{n!}{(k+1)!(n-k-2)!}\int_0^{F(x)}t^{k+1}(1-t)^{n-k-2}\mathrm{d}t$$

依次进行*n-k*次分部积分，可以证明上式等于

$$\sum_{i=k}^{n}C_n^i[F(x)]^i[1-F(x)]^{n-i}$$

例1(例1.18)设总体X服从区间[0,1]上的均匀分布，$(X_1, X_2, \cdots, X_n)$为总体X的样本，试求$X_{(k)}$的分布.

证 总体X的分布密度为

$$f(x)=\begin{cases}1, & 0 \le x \le 1 \\ 0, & \text{其他}\end{cases}$$

X的分布函数为

$$F(x)=\begin{cases}0, & x<0 \\ x, & 0 \le x \le 1 \\ 1, & x>1\end{cases}$$

$$f_{X_{(k)}}(x)=\frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}f(x)$$

$$=\frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k},\quad 0\le x\le 1.$$

定理1.20 设总体X的分布密度为$f(x)$(或分布函数为$F(x)$), $X_1, X_2, \cdots, X_n$为来自总体X的样本，则次序统计量$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})^T$的联合分布密度为

$$f(y_1, y_2, \cdots, y_n) = \begin{cases} n!\prod_{i=1}^{n} f(y_i), & y_1 < y_2 < \cdots < y_n \\ 0, & \text{其他}, \end{cases}$$

证明略

$$f_{X_{(k)}}(x) = kC_n^k [F(x)]^{k-1}[1-F(x)]^{n-k} f(x)$$

注：显然可以验证

$$f(x_1, x_2, \cdots, x_n) \neq f_{X_{(1)}}(x) f_{X_{(2)}}(x) \ldots f_{X_{(n)}}(x)$$

$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$不相互独立！！！

例2(例1.19) 设总体X服从区间$[0,\theta]$上的均匀分布,$(X_1, X_2, \cdots, X_n)$为总体X的样本,试求次序统计量$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})^{\mathrm{T}}$的联合分布密度.

证 总体X的分布密度为

$$f(x)=\begin{cases}\dfrac{1}{\theta}, & 0\le x\le\theta\\ 0, & \text{其他}\end{cases}$$

次序统计量的联合分布为

$$f(y_1, y_2, \cdots, y_n)=\begin{cases}\dfrac{n!}{\theta^n}, & 0<y_1<y_2<\cdots<y_n<\theta,\\ 0, & \text{其他},\end{cases}$$

定理1.21 设总体X的分布密度为$f(x)$(或分布函数为$F(x)$), $X_1, X_2, \cdots, X_n$为来自总体X的样本，则次序统计量$(X_{(1)}, X_{(n)})^T$的联合分布密度为

$$f_{(X_{(1)},X_{(n)})}(x,y)=\begin{cases} n(n-1)[F(y)-F(x)]^{n-2} f(x)f(y), & x<y, \\ 0, & x \ge y. \end{cases}$$

证 根据分布函数的定义可得，任给 $x, y \in R$

$$F_{(X_{(1)},X_{(n)})}(x,y)=P\{X_{(1)} \le x, X_{(n)} \le y\}$$

以下分两种情形讨论：

(1)当$x \ge y$时，

$$F_{(X_{(1)},X_{(n)})}(x,y)=P\{X_{(1)}\le x, X_{(n)}\le y\}=P\{X_{(n)}\le y\}$$

$$=P\{X_1\le y,\cdots,X_n\le y\}=\prod_{i=1}^{n}P\{X_i\le y\}=[F(y)]^n$$

(2)当$x<y$时，

$$F_{(X_{(1)},X_{(n)})}(x,y)=P\{X_{(1)}\le x, X_{(n)}\le y\}$$

又由于

$$\{X_{(n)}\le y\}=\{X_{(1)}\le x, X_{(n)}\le y\}+\{X_{(1)}> x, X_{(n)}\le y\}$$

$$=\{X_{(1)}\le x, X_{(n)}\le y\}+\{x<X_1\le y,\cdots,x<X_n\le y\}$$

因而 $[F(y)]^n=F_{(X_{(1)},X_{(n)})}(x,y)+[F(y)-F(x)]^n$

所以 $F_{(X_{(1)},X_{(n)})}(x,y)=[F(y)]^n-[F(y)-F(x)]^n$

此时只与y有关

于是可以得到其联合分布密度为

$$f_{(X_{(1)},X_{(n)})}(x,y)=\frac{\partial^2 F_{(X_{(1)},X_{(n)})}(x,y)}{\partial x\partial y}$$

$$=\begin{cases} n(n-1)[F(y)-F(x)]^{n-2}f(x)f(y), & x<y, \\ 0, & x\ge y. \end{cases}$$

二、样本中位数和样本极差

1、样本中位数

定义 设 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})^T$ 为样本 $(X_1, X_2, \cdots, X_n)^T$ 的次序统计量，样本的中位数定义为

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})}, & n\text{为奇数}, \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}], & n\text{为偶数}, \end{cases}$$

其观测值为

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & n\text{为奇数}, \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & n\text{为偶数}, \end{cases}$$

2、样本中位数的意义

样本中位数主要用来描述样本位置的特征，具有和样本均值类似的含义，但它不受样本异常值的影响，同时也容易计算，也可以作为总体均值的估计．缺点是分布不容易计算，因而在理论讨论时，带来一定困难.

3、样本极差

定义设$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})^T$为样本$(X_1, X_2, \cdots, X_n)^T$的次序统计量，样本的极差定义为

$$R = X_{(n)} - X_{(1)} = \max_{1 \le i \le n} X_i - \min_{1 \le i \le n} X_i$$

其观测值为$r = x_{(n)} - x_{(1)} = \max\limits_{1 \le i \le n} x_i - \min\limits_{1 \le i \le n} x_i$

4、样本极差的意义

样本极差主要用来描述样本变化幅度以及离散程度的特征，具有和样本方差类似的含义，也可以作为总体均方差的估计. 但它受样本异常值的影响较大。却由于容易计算，在实际中应用比较广泛.

例3(例1.20) 从总体中抽取容量为6的样本，测得样本值为 32，65，28，35，30，29

试求, 样本中位数、样本均值、样本极差、样本方差、以及样本标准差。

证 首先将样本观测值进行排序，可得

28，29，30，32，35，65，

则 样本中位数：$\tilde{x}=\dfrac{1}{2}(x_{(3)}+x_{(4)})=31$

样本均值：$\bar{x}=\dfrac{1}{6}\sum_{i=1}^{6}x_i=36.5$

样本极差：$r=\max\limits_{1\le i\le 6}x_i-\min\limits_{1\le i\le 6}x_i=65-28=37$

样本方差：$s_n^2 = \frac{1}{6}\sum_{i=1}^{6} x_i^2 - \overline{x}^2 = 167.583$

样本标准差：$s_n = \sqrt{\frac{1}{6}\sum_{i=1}^{6} x_i^2 - \overline{x}^2} = 12.954$

Thank You!