第2章 抽样分布及若干预备知识

§2.1 引言

统计基本目的: 从样本出发推断总体分布.——统计推断(statistical inference)

$$
\text{Statistical Inference}\begin{cases}\text{Sampling distribution}\\ \text{Estimation}\begin{cases}\text{Point Estimation}\\ \text{Interval Estimation}\end{cases}\\ \text{Hypothesis Testing}\end{cases}
$$

Ronald Aylmer Fisher

Born: 17 Feb 1890 in London, England

Died: 29 July 1962 in Adelaide, Australia

抽样分布

统计量是对样本的加工,将样本中分散的信息浓缩集中起来.
统计量浓缩集中样本中关于总体分布的信息能力是通过它的分布来体现的.

统计量的分布通常称为抽样分布(sampling distribution), 或称诱导分布.

当总体X的分布类型已知时, 样本$(X_1, X_2, \cdots, X_n)$的分布类型也已知了. 理论上可以推导出统计量
$T = T(X_1, X_2, \cdots, X_n)$ 的分布. 但实际上推导出T的分布的明显表达式常常不容易.

如果对每个n, 都能导出统计量$T = T(X_1, X_2, \cdots, X_n)$的分布的明显表达式, 此时统计量的分布为精确的抽样分布.

目前的精确分布大多是在正态总体条件下得到的, 主要涉及"统计三大分布"——χ^2分布, t分布, F分布.

§2.2 正态总体样本均值和样本方差的分布

这里,我们首先给出样本均值与样本方差的分布.

已知结论

总体$X \sim F$, $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$为来自该总体的简单随机样本, $\overline{X}$和S^2为其样本均值与样本方差. 若总体的方差存在, 并记$\mathsf{E}X=\mu$, $\mathsf{Var}X=\sigma^2$时, 则有

$$\mathsf{E}\overline{X}=\mu, \quad \mathsf{Var}\overline{X}=\sigma^2/n;$$

$$\mathsf{E}(S^2)=\sigma^2, \quad \mathsf{E}(S_n^2)=\mathsf{E}(\frac{n-1}{n}S^2)=\frac{n-1}{n}\sigma^2.$$

Lemma

设在两个随机变量$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)'$ 和$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)'$ 间有一个线性变换

$$\boldsymbol{Y} = A\boldsymbol{X},$$

其中$A = (a_{ij})$为$n \times n$阶方阵, 则

$$\mathsf{E}\boldsymbol{Y} = A(\mathsf{E}\boldsymbol{X}),$$

$$\mathsf{Var}\boldsymbol{Y} = A(\mathsf{Var}\boldsymbol{X})A'.$$

证明: 线性变换$\boldsymbol{Y} = A\boldsymbol{X}$即为

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j, \quad i = 1, 2, \cdots, n.$$

于是

$$\mathsf{E}Y_i = \sum_{j=1}^{n} a_{ij} \mathsf{E}X_j, \quad i = 1, 2, \cdots, n.$$

证明: 线性变换$\boldsymbol{Y}=A\boldsymbol{X}$即为

$$Y_i=\sum_{j=1}^{n}a_{ij}X_j,\quad i=1,2,\cdots,n.$$

于是

$$\mathsf{E}Y_i=\sum_{j=1}^{n}a_{ij}\mathsf{E}X_j,\quad i=1,2,\cdots,n.$$

这就是

$$\mathsf{E}\boldsymbol{Y}=A(\mathsf{E}\boldsymbol{X}).$$

上式说明求数学期望可和线性变换交换次序.

至于第二个等式，我们有

$$\begin{aligned}
\mathrm{Var}\boldsymbol{Y} &= \mathsf{E}\left[(\boldsymbol{Y}-\mathsf{E}\boldsymbol{Y})(\boldsymbol{Y}-\mathsf{E}\boldsymbol{Y})'\right] \\
&= \mathsf{E}\left[(A\boldsymbol{X}-A\mathsf{E}\boldsymbol{X})(A\boldsymbol{X}-A\mathsf{E}\boldsymbol{X})'\right] \\
&= \mathsf{E}\left[A(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})'A'\right] \\
&= A\mathsf{E}\left[(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})'A'\right] \\
&= A\left(\mathsf{E}\left[(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})(\boldsymbol{X}-\mathsf{E}\boldsymbol{X})'\right]\right)A' \\
&= A(\mathrm{Var}\boldsymbol{X})A'.
\end{aligned}$$

Theorem

定理2.2.3 设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$的一组简单随机样本. $\overline{X}$和S^2为其样本均值与样本方差, 则

(1) $\overline{X} \sim N(\mu, \sigma^2/n)$;

(2) $\frac{(n-1)S^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(n-1)$;

(3) $\overline{X}$与S^2独立.

在证明定理2.2.3之前,回忆一下几个概率论的结果:

- χ^2分布的定义: 设$X_1, ..., X_n$ i.i.d.$\sim N(0,1)$，则

$$K = \sum_{i=1}^{n} X_i^2$$

  的分布定义为自由度是n的χ^2分布，记为$K \sim \chi^2(n)$.

- 如果n维随机变量$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)'$服从$n$维正态分布, 那么它线性变换$\boldsymbol{Y} = A\boldsymbol{X}$也服从正态分布.

- 如果$(X_1, X_2, \cdots, X_n)'$服从n维正态分布, 那么$X_1, X_2, \cdots, X_n$相互独立等价于它们两两不相关.

定理2.2.3的证明: 记$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)'$. 则$\boldsymbol{X}$服从$n$维正态分布. 构造一个正交矩阵$A = (a_{ij})$使得第一行的元素都为$1/\sqrt{n}$:

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\cdot 1}} & \frac{-1}{\sqrt{2\cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & \frac{-2}{\sqrt{3\cdot 2}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}$$

作线性变换

$$\boldsymbol{Y} = A\boldsymbol{X}.$$

则$\boldsymbol{Y}$也服从n维正态分布,

且$Y_1 = (X_1 + X_2 + \cdots + X_n)/\sqrt{n} = \sqrt{n}\overline{X}$.

作线性变换

$$\boldsymbol{Y} = A\boldsymbol{X}.$$

则$\boldsymbol{Y}$也服从n维正态分布,

且$Y_1 = (X_1 + X_2 + \cdots + X_n)/\sqrt{n} = \sqrt{n}\overline{X}$. 由前面的定理,可得

$$\mathsf{E}\boldsymbol{Y} = A\mathsf{E}\boldsymbol{X} = A(\mu, \mu, \cdots, \mu)' = (\sqrt{n}\mu, 0, \cdots, 0)',$$

$$\mathsf{Var}\boldsymbol{Y} = A(\mathsf{Var}\boldsymbol{X})A' = A(\sigma^2 I)A' = \sigma^2 I.$$

作线性变换

$$\boldsymbol{Y} = A\boldsymbol{X}.$$

则$\boldsymbol{Y}$也服从n维正态分布,

且$Y_1 = (X_1 + X_2 + \cdots + X_n)/\sqrt{n} = \sqrt{n}\overline{X}$. 由前面的定理,可得

$$\mathsf{E}\boldsymbol{Y} = A\mathsf{E}\boldsymbol{X} = A(\mu, \mu, \cdots, \mu)' = (\sqrt{n}\mu, 0, \cdots, 0)',$$

$$\mathsf{Var}\boldsymbol{Y} = A(\mathsf{Var}\boldsymbol{X})A' = A(\sigma^2 I)A' = \sigma^2 I.$$

因此, $Y_1, Y_2, \cdots, Y_n$相互独立, $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$,
$Y_k \sim N(0, \sigma^2)$, $k = 2, \cdots, n$.

另一方面,

$$
\begin{aligned}
\sum_{i=1}^{n} Y_i^2 =& \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'A'A\boldsymbol{X} \\
=& \boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n} X_i^2 \\
=& \sum_{i=1}^{n}(X_i - \overline{X})^2 + n\overline{X}^2 \\
=& (n-1)S^2 + Y_1^2.
\end{aligned}
$$

另一方面,

$$\begin{aligned}\sum_{i=1}^n Y_i^2 =& \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'A'A\boldsymbol{X}\\ =&\boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^n X_i^2\\ =&\sum_{i=1}^n (X_i-\overline{X})^2 + n\overline{X}^2\\ =&(n-1)S^2 + Y_1^2.\end{aligned}$$

所以

$$(n-1)S^2 = \sum_{i=2}^n Y_i^2 \text{ 与 } \overline{X} = Y_1/\sqrt{n} \text{ 独立}$$

并且

$$(n-1)S^2/\sigma^2 = \sum_{i=2}^{n}(Y_i/\sigma)^2 \sim \chi^2(n-1),$$

$$\overline{X} = Y_1/\sqrt{n} \sim N(\mu, \sigma^2/n),$$

定理得证.

二、渐近分布(Asymptotic distribution)

在大多数场合，精确抽样分布不易求出，或精确抽样分布过于复杂而难于应用, 这时常常求统计量

$$T = T(X_1, X_2, \cdots, X_n)$$

当$n \to \infty$时的极限分布, 这种极限分布称为渐近分布.
当n较大时, 用渐近分布当作抽样分布的近似.

Example

例 设$X_1, X_2, \cdots, X_n$取自正态总体$X \sim N(0, \sigma^2)$. 则

$$\sqrt{n}\overline{X}/\sigma \xrightarrow{D} N(0,1) \text{ 和 } S_n^2 \xrightarrow{P} \sigma^2.$$

所以

$$T = \sqrt{n}\overline{X}/S_n = \frac{\sqrt{n}\overline{X}/\sigma}{S_n/\sigma} \xrightarrow{D} N(0,1), \quad \text{当} n \to \infty,$$

即

$$T \sim AN(0,1).$$

Example

例 设$X_1, X_2, \cdots, X_n$取自总体$X \sim F(x)$,且$\mu_{2k} = EX^{2k}$存在. 则

$$a_{n,k} \sim AN(\mu_k, \frac{\mu_{2k} - (\mu_k)^2}{n}).$$

三、用随机模拟法(simulation)求统计量的近似分布

设想有一个统计量$T = T(X_1, X_2, \cdots, X_n)$, 为获得$T$的分布函数$F^{(n)}(t)$, 我们可以重复地做类似的试验(譬如说$N$次). 每次从总体中随机抽取样本容量为$n$的样本, 计算对应的统计量的值. 这样我们可以得到统计量T的N个观察值: $t_1, t_2, \cdots, t_N$. 根据这N个观察值可以写出经验分布函数$F_N^{(n)}(t)$. 由格里汶科定理知, 当N充分大时, $F_N^{(n)}(t)$是$F^{(n)}(t)$的一个很好近似.

三、用随机模拟法(simulation)求统计量的近似分布

设想有一个统计量$T = T(X_1, X_2, \cdots, X_n)$, 为获得$T$的分布函数$F^{(n)}(t)$, 我们可以重复地做类似的试验(譬如说$N$次). 每次从总体中随机抽取样本容量为$n$的样本, 计算对应的统计量的值. 这样我们可以得到统计量T的N个观察值: $t_1, t_2, \cdots, t_N$. 根据这N个观察值可以写出经验分布函数$F_N^{(n)}(t)$. 由格里汶科定理知, 当N充分大时, $F_N^{(n)}(t)$是$F^{(n)}(t)$的一个很好近似.

在实际中, 我们做重复试验的方法求统计的分布一般难以实现(很多时候成本也太高). 通常这一过程可由计算机实现.

Example

例 设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$的一个样本,其样本峰度为

$$b_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2} - 3.$$

b_k的精确分布难以求得. 可以证明, $\sqrt{n}b_k$有渐近分布$N(0, 24)$,但其收敛速度甚"慢". 下面介绍如何用随机模拟的方法得到其近似分布.

如何获得样本峰度b_k的大量观察值?

正态总体$N(\mu, \sigma^2)$中的μ和σ未知, 我们不能从这个总体中获得样本.

如何获得样本峰度b_k的大量观察值?

正态总体$N(\mu,\sigma^2)$中的μ和σ未知, 我们不能从这个总体中获得样本.

作变换$X_i^*=(X_i-\mu)/\sigma$. 则$X_1^*,X_2^*,\cdots,X_n^*$是来至$N(0,1)$的样本. 并且易知

$$b_k=\frac{\frac{1}{n}\sum_{i=1}^n(X_i^*-\overline{X^*})^4}{\left[\frac{1}{n}\sum_{i=1}^n(X_i^*-\overline{X^*})^2\right]^2}-3.$$

这样我们只要从正态$N(0,1)$中抽取样本计算b_k的样本观察值即可.

模拟步骤如下:

a. 由计算机产生n个标准正态$N(0,1)$随机数. 许多软件可以产生样的随机数. 在Matlab只要用normrnd(0,1,n,1)命令, 在R中只要rnorm(n)即可(一般正态的随机数可用rnorm(n,mean,sd)).

模拟步骤如下:

a. 由计算机产生n个标准正态$N(0,1)$随机数. 许多软件可以产生样的随机数. 在Matlab只要用normrnd(0,1,n,1)命令, 在R中只要rnorm(n)即可(一般正态的随机数可用rnorm(n,mean,sd)).

b. 计算此样本的样本峰度值b_k.

模拟步骤如下:

a. 由计算机产生n个标准正态$N(0,1)$随机数. 许多软件可以产生样的随机数. 在Matlab只要用normrnd(0,1,n,1)命令, 在R中只要rnorm(n)即可(一般正态的随机数可用rnorm(n,mean,sd)).

b. 计算此样本的样本峰度值b_k.

c. 重复(a到b步)N次(比如N取1000, 5000或10000), 可得b_k的N次观察值.

模拟步骤如下:

a. 由计算机产生n个标准正态$N(0,1)$随机数. 许多软件可以产生样的随机数. 在Matlab只要用normrnd(0,1,n,1)命令, 在R中只要rnorm(n)即可(一般正态的随机数可用rnorm(n,mean,sd)).

b. 计算此样本的样本峰度值b_k.

c. 重复(a到b步)N次(比如N取1000, 5000或10000), 可得b_k的N次观察值.

d. 把b_k的N次观察值从小到大排序, 从中可以找到b_k分布的0.01,0.05,0.10,...等的分位数. (在统计中, 我们常常需要的是这样的分位数,并不需要知道分布函数的每一个值)

模拟步骤如下:

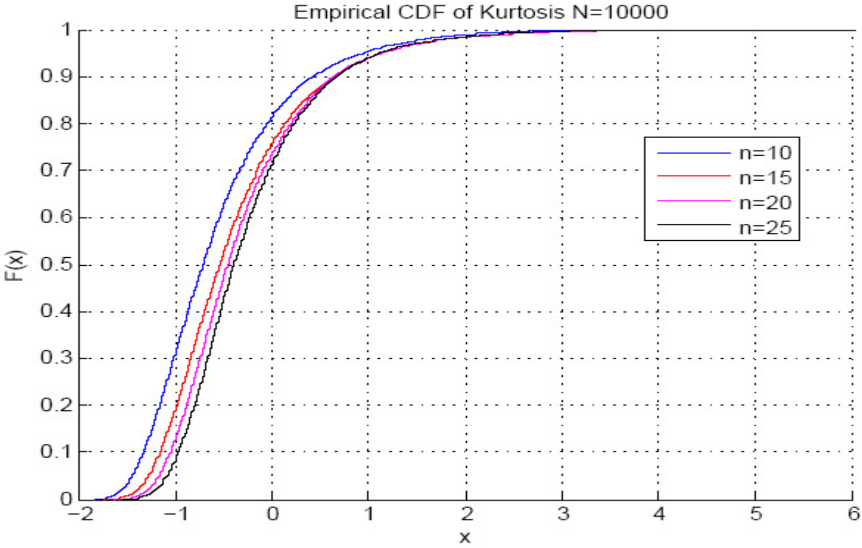
a. 由计算机产生n个标准正态$N(0,1)$随机数. 许多软件可以产生样的随机数. 在Matlab只要用normrnd(0,1,n,1)命令, 在R中只要rnorm(n)即可(一般正态的随机数可用rnorm(n,mean,sd)).

b. 计算此样本的样本峰度值b_k.

c. 重复(a到b步)N次(比如N取1000, 5000或10000), 可得b_k的N次观察值.

d. 把b_k的N次观察值从小到大排序, 从中可以找到b_k分布的0.01,0.05,0.10,...等的分位数. (在统计中, 我们常常需要的是这样的分位数,并不需要知道分布函数的每一个值)

e. 改变样本容量n的值,重复a到d步, 又可在另一个n处得到b_k分布的各种分位数.


Empirical CDF of Kurtosis N=10000
F(x)
x
n=10
n=15
n=20
n=25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
−2
−1
0
1
2
3
4
5
6

N=10000

	0.0100	0.0500	0.1000	0.9000	0.9500	0.9900
10	-1.5936	-1.4339	-1.3197	0.4775	0.9733	1.9213
15	-1.4744	-1.2807	-1.1598	0.6311	1.1435	2.3744
20	-1.3595	-1.1697	-1.0535	0.6435	1.0982	2.2568
25	-1.2838	-1.0912	-0.9671	0.6393	1.1057	2.3223

N=100000

	0.0100	0.0500	0.1000	0.9000	0.9500	0.9900
10	-1.6179	-1.4441	-1.3270	0.4354	0.9489	1.9900
15	-1.4567	-1.2737	-1.1586	0.6663	1.1418	2.3706
20	-1.3585	-1.1644	-1.0470	0.6366	1.1423	2.4021
25	-1.2736	-1.0875	-0.9669	0.6870	1.1206	2.2455

N=1000000

	0.0100	0.0500	0.1000	0.9000	0.9500	0.9900
10	-1.6138	-1.4364	-1.3203	0.4637	0.9458	1.9916
15	-1.4619	-1.2787	-1.1588	0.6170	1.1180	2.3194
20	-1.3553	-1.1691	-1.0493	0.6679	1.1497	2.3487
25	-1.2730	-1.0873	-0.9679	0.6805	1.1398	2.3057

§2.3 次序统计量及其分布

一、次序统计量

Definition

定义 样本$X_1, X_2, \cdots, X_n$按从小到大的顺序重排为

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)},$$

则$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$称为次序统计量(order statistics). 特别地, $X_{(1)}$称为最小次序统计量, $X_{(n)}$称为最大次序统计量. 其观察值$(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$也简称为次序统计量.
有时为了强调样本容量n, 次序统计量也记为$(X_{n1}, X_{n2}, \ldots, X_{nn})$.

Definition

定义' $X_{(i)}$称为样本$X_1, X_2, ..., X_n$的第i个次序统计量,是指它是样本的满足如下条件的函数: 每当样本的一组观测值为$x_1, x_2, ..., x_n$,将它们按从小到大的顺序重排为$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$,其中的第$i$个值$x_{(i)}$便是$X_{(i)}$的观测值.

Definition

定义′ $X_{(i)}$称为样本$X_1, X_2, ..., X_n$的第i个次序统计量,是指它是样本的满足如下条件的函数: 每当样本的一组观测值为$x_1, x_2, ..., x_n$,将它们按从小到大的顺序重排为$x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$,其中的第$i$个值$x_{(i)}$便是$X_{(i)}$的观测值.

不难看出次序统计量与经验分布函数是相互唯一确定的.

例如，假定有一组容量为6的样本，其观测值如下：

x_1	x_2	x_3	x_4	x_5	x_6
5	3	8	6	2	4

则次序统计量的观测值为

x_5	x_2	x_6	x_1	x_4	x_3
2	3	4	5	6	8

在这个例子中，$n = 6$,

$$x_{(1)} = x_5 \qquad x_{(2)} = x_2 \qquad x_{(3)} = x_6$$

$$x_{(4)} = x_1 \qquad x_{(5)} = x_4 \qquad x_{(6)} = x_3.$$

Example

总体X来自两点分布族$\mathcal{F} = \{B(1,p) : 0 < p < 1\}$, 从中抽取了容量为3的一组简单随机样本$(X_1, X_2, X_3)$.

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0				

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0	0	0	0	

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0	0	0	0	$(1-p)^3$

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0	0	0	0	$(1-p)^3$
②	1	0	0				

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0	0	0	0	$(1-p)^3$
②	1	0	0	0	0	1	

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0	0	0	0	$(1-p)^3$
②	1	0	0	0	0	1	$(1-p)^2p$

则样本的可能取值, 对应的次序统计量的值及发生的概率分别如下:

	X_1	X_2	X_3	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	P
①	0	0	0	0	0	0	$(1-p)^3$
②	1	0	0	0	0	1	$(1-p)^2p$
③	0	1	0	0	0	1	$(1-p)^2p$
④	0	0	1	0	0	1	$(1-p)^2p$
⑤	1	1	0	0	1	1	$(1-p)p^2$
⑥	1	0	1	0	1	1	$(1-p)p^2$
⑦	0	1	1	0	1	1	$(1-p)p^2$
⑧	1	1	1	1	1	1	p^3

那么次序统计量的分布为:

$$P(X_{(1)}=1)=p^3, \quad P(X_{(1)}=0)=1-p^3;$$
$$P(X_{(2)}=1)=3(1-p)p^2+p^3, \quad P(X_{(2)}=0)=1-3p^2+2p^3;$$
$$P(X_{(3)}=1)=1-(1-p)^3, \quad P(X_{(3)}=0)=(1-p)^3.$$

那么次序统计量的分布为:

$$P(X_{(1)}=1)=p^3,\quad P(X_{(1)}=0)=1-p^3;$$
$$P(X_{(2)}=1)=3(1-p)p^2+p^3,\quad P(X_{(2)}=0)=1-3p^2+2p^3;$$
$$P(X_{(3)}=1)=1-(1-p)^3,\quad P(X_{(3)}=0)=(1-p)^3.$$

简单随机样本$\{X_i, i=1,2,\cdots,n\}$是独立同分布的, 但次序统计量$\{X_{(i)}, i=1,2,\cdots,n\}$不一定是独立同分布的.

二、次序统计量的分布

设总体为连续分布, 分布函数为$F(x)$, 概率密度函数为$p(x)$,则

1. $(X_{(1)}, X_{(2)}, ..., X_{(n)})$的联合密度函数为

$$g(y_1, y_2, \cdots, y_n) = n!p(y_1)\cdots p(y_n), \quad y_1 \leq \cdots \leq y_n.$$

2. 最大次序统计量$X_{(n)}$的分布函数为$F^n(y)$, 密度为$np(y)F^{n-1}(y)$;
3. 最小次序统计量$X_{(1)}$的分布函数为$1-(1-F(y))^n$, 密度为$np(y)(1-F(y))^{n-1}$;
4. $X_{(k)}$的密度函数为

$$\frac{n!}{(k-1)!(n-k)!}p(y)F^{k-1}(y)(1-F(y))^{n-k};$$

5. $(X_{(i)}, X_{(j)})$ $i < j$的联合密度函数为

$$g(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} p(y_i)p(y_j) \times F^{i-1}(y_i)(F(y_j) - F(y_i))^{j-i-1}(1 - F(y_j))^{n-j}, \quad y_i \le y_j.$$

6. 特别地, $(X_{(1)}, X_{(n)})$的联合密度函数为

$$g_{1n}(y_1, y_n) = n(n-1)(F(y_n) - F(y_1))^{n-2} p(y_1)p(y_n) \quad y_1 \le y_n.$$

(4)的证明: 记$A_i = \{X_i < y\}$,则$X_{(k)}$的分布函数为

$$
\begin{aligned}
F_k(y) =& P(X_{(k)} < y) = P(\text{事件}A_1, \cdots, A_n\text{中至少有}k\text{个发生}) \\
&= \sum_{i=k}^{n} P(\text{事件}A_1, \cdots, A_n\text{中恰好有}i\text{个发生}) \\
&= \sum_{i=k}^{n} \binom{n}{i} F^i(y)(1 - F(y))^{n-i}.
\end{aligned}
$$

(4)的证明: 记$A_i = \{X_i < y\}$,则$X_{(k)}$的分布函数为

$$F_k(y) = P(X_{(k)} < y) = P(\text{事件}A_1, \cdots, A_n\text{中至少有}k\text{个发生})$$
$$= \sum_{i=k}^{n} P(\text{事件}A_1, \cdots, A_n\text{中恰好有}i\text{个发生})$$
$$= \sum_{i=k}^{n} \binom{n}{i} F^i(y)(1 - F(y))^{n-i}.$$

利用恒等式

$$\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} = k\binom{n}{k} \int_0^p t^{k-1}(1-t)^{n-k} dt,$$

得

$$F_k(y) = \sum_{i=k}^{n} \binom{n}{i} F^i(y)(1-F(y))^{n-i}$$
$$= k\binom{n}{k} \int_0^{F(y)} t^{k-1}(1-t)^{n-k}dt.$$

得

$$
\begin{aligned}
F_k(y) =& \sum_{i=k}^{n} \binom{n}{i} F^i(y)(1-F(y))^{n-i} \\
=& k\binom{n}{k}\int_0^{F(y)} t^{k-1}(1-t)^{n-k}dt.
\end{aligned}
$$

密度函数为

$$
g_k(y) = k\binom{n}{k} F^{k-1}(y)(1-F(y))^{n-k}p(y).
$$

(5)的证明: 不妨设$y_i < y_j$. 由于

$$P(y_i \le X_{(i)} < y_i+dy_i, y_j \le X_{(j)} < y_j+dy_j) \approx p(y_i, y_j)dy_idy_j.$$

右边是左边概率的主要部分,即概率微元.

(5)的证明: 不妨设$y_i < y_j$. 由于

$$P(y_i \le X_{(i)} < y_i + dy_i, y_j \le X_{(j)} < y_j + dy_j) \approx p(y_i, y_j) dy_i dy_j.$$

右边是左边概率的主要部分,即概率微元. 对充分小的微元$dy_i dy_j$, 事件$\{y_i \le X_{(i)} < y_i + dy_i, y_j \le X_{(j)} < y_j + dy_j\}$意味着:

样本$X_1, X_2, \cdots, X_n$ 的观察值中有

- $i-1$个观察值落在$(-\infty, y_i)$内, (每个观察值落在这个区间的概率为$F(y_i)$);
- 1个观察值落在$[y_i, y_i+dy_i)$内, (每个观察值落在这个区间的概率为$F(y_i+dy_i)-F(y_i)\approx p(y_i)dy_i$);
- $j-i-1$个观察值落在$[y_i+dy_i, y_j)$内, (每个观察值落在这个区间的概率为$F(y_j)-F(y_i+dy_i)\approx F(y_j)-F(y_i)$);
- 1个观察值落在$[y_j, y_j+dy_j)$内,(每个观察值落在这个区间的概率为$F(y_j+dy_j)-F(y_j)\approx p(y_j)dy_j$);
- $n-j$个观察值落在$[y_i+dy_i, +\infty)$内,(每个观察值落在这个区间的概率为$1-F(y_j+dy_j)\approx 1-F(y_j)$).

因此

$$
\begin{aligned}
& P(y_i \le X_{(i)} < y_i + dy_i, y_j \le X_{(j)} < y_j + dy_j) \\
\approx \; & \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times F^{i-1}(y_i) \times p(y_i)dy_i \\
& \times (F(y_j) - F(y_i))^{j-i-1} \times p(y_j)dy_j \times (1 - F(y_j))^{n-j}.
\end{aligned}
$$

因此

$$
\begin{aligned}
& P(y_i \le X_{(i)} < y_i + dy_i, y_j \le X_{(j)} < y_j + dy_j) \\
\approx\ & \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times F^{i-1}(y_i) \times p(y_i)dy_i \\
& \times (F(y_j) - F(y_i))^{j-i-1} \times p(y_j)dy_j \times (1 - F(y_j))^{n-j}.
\end{aligned}
$$

即

$$
\begin{aligned}
& p(y_i, y_j)dy_idy_j \\
=\ & \frac{n!}{(i-1)!(j-i-1)!(n-j)!} p(y_i)p(y_j)F^{i-1}(y_i) \\
& \times (F(y_j) - F(y_i))^{j-i-1}(1 - F(y_j))^{n-j}dy_idy_j.
\end{aligned}
$$

结论得证.

Theorem

定理 $X_{(j_1)}, X_{(j_2)} \cdots, X_{(j_r)}$ (其中$1 \le j_1 < j_2 < \cdots < j_r$) 的联合密度函数为

$$
\begin{aligned}
& g(y_{j_1}, y_{j_2}, \cdots, y_{j_r}) \\
= & \frac{n!}{(j_1-1)!(j_2-j_1-1)!\cdots(j_r-j_{r-1}-1)!(n-j_r)!} \\
& \times p_1^{j_1-1} p_2^{j_2-j_1-1} \cdots p_r^{j_r-j_{r-1}-1} p_{r+1}^{n-j_r} \\
& \times p(y_{j_1}) p(y_{j_2}) \cdots p(y_{j_r}),
\end{aligned}
$$

其中,

$$a = y_{j_0} \le y_{j_1} \le y_{j_2} \le \cdots \le y_{j_r} \le y_{j_{r_1}} = b,$$

$$p_k = \int_{y_{j_{k-1}}}^{y_k} p(x)dx = F(y_{j_k}) - F(y_{j_{k-1}}), \quad k = 1, 2, \cdots, r+1.$$

三、极差、样本中位数与分位数

由次序统计量出发可以构造许多有用的统计量，例如

样本极差(sample range)：

$$R_n = X_{(n)} - X_{(1)}.$$

若总体$X \sim F(x)$,其密度函数为$p(x)$,则样本极差R_n的密度函数为

$$p_{R_n}(r) = \int_{-\infty}^{\infty} n(n-1)(F(r+z)-F(z))^{n-2}p(r+z)p(z)dz,\ \ r > 0.$$

三、极差、样本中位数与分位数

由次序统计量出发可以构造许多有用的统计量，例如

样本极差(sample range)：

$$R_n = X_{(n)} - X_{(1)}.$$

若总体$X \sim F(x)$,其密度函数为$p(x)$,则样本极差R_n的密度函数为

$$p_{R_n}(r) = \int_{-\infty}^{\infty} n(n-1)(F(r+z)-F(z))^{n-2}p(r+z)p(z)dz, \ \ r > 0.$$

极差反映了总体标准差的的信息.

样本中位数(sample median):

$$m_e = \begin{cases} X_{(\frac{n+1}{2})}, & \text{当}n\text{为奇数}, \\ \left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right)/2, & \text{当}n\text{为偶数}. \end{cases}$$

样本中位数反映了总体中位数的信息.

总体分位数定义如下: 对$0 < p < 1$, 若

$$F(\xi_p) = p$$

或者

$$F(\xi_p) < p \text{ 但 } F(\xi_p + 0) \geq p,$$

则称ξ_p为总体X(或分布函数$F(x)$)的(下侧)p分位数, $1/2$分位数称为中位数.

样本分位数: $X_{(r)}$ 定义为样本$r/(n+1)$分位数. 对一般的$0 < p < 1$, 由线性插入法,定义p分位数.

Definition

定义 设$X_1, X_2, \cdots, X_n$是取自总体$F(x)$的一个样本.
对$0 < p < 1$, 称

$$m_p = X_{([np])} + (n+1)\left(p - \frac{[np]}{n+1}\right)(X_{([np]+1)} - X_{([np])})$$

为该样本的(下侧)p分位数.

一些教材也直接定义$X_{([np])}$为样本(下侧)p分位数, 而称$X_{([n(1-p)])}$为样本上侧p分位数.

一般地, 分布F的p分位数ξ_p是指满足$F(\xi_p)=p$的一个数. 由于样本的经验分布函数不是严格单调也不是连续的,因此$F_n(x)=p$的解可能不存在, 即使存在也不一定唯一. 对于观测值$x_1,\ldots,x_n$, 样本p分位数m_p应使得

$$\#\{x_i<m_p\}\approx np, \qquad \#\{x_i>m_p\}\approx n(1-p).$$

在实用中, 样本分位数有多种规定, 在样本容量n较大时它们的数值时相近的.

$$m_p = \begin{cases} (1-g)X_{(j)} + gX_{(j+1)}, & j = [(n-1)p]+1, \\ & g = (n-1)p - j + 1, \\ & \text{Excel, R, Splus}, \\ (1-g)X_{(j)} + gX_{(j+1)}, & j = [np], g = np - j, \\ & \text{SAS 公式1}, \\ (1-g)X_{(j)} + gX_{(j+1)}, & j = [(n+1)p], g = (n+1)p - j, \\ & \text{JMP, SPSS}, \\ X_{[np]}, & \text{SAS 公式3}. \end{cases}$$

$$m_p = \begin{cases} \frac{1}{2}(X_{(j)} + X_{(j+1)}), & g = 0 \\ X_{(j+1)}, & g > 0, \end{cases} \qquad j = [np], g = np - [np],$$

SAS 公式5 (缺省情况).

样本四分位数(quartile)和四分位距:

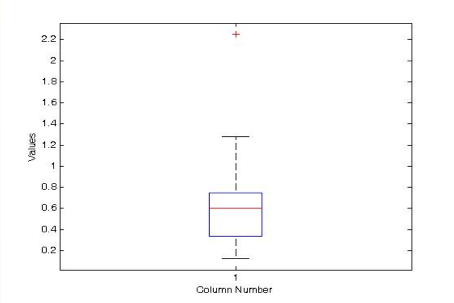
0.25分位数$m_{0.25}$—下四分位数

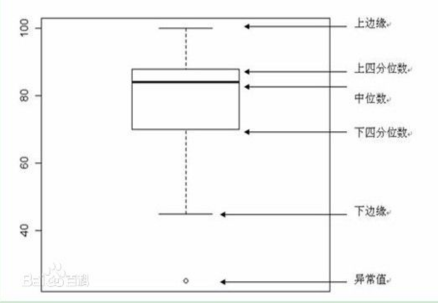
0.75分位数$m_{0.75}$—上四分位数

$IQR = m_{0.75} - m_{0.25}$—四分位距(interquartile range).

箱线图(盒型图)(Boxplot):

+—1.5 IQR 以外的点, outliers in the data


2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
Values
1
Column Number


100
80
60
40
上边缘
上四分位数
中位数
下四分位数
下边缘
异常值

样本分位数的精确分布比较复杂,但在一定的条件下可以求得渐近分布.

Theorem

定理 设总体的p分位数为ξ_p, 又设总体的密度函数$f(x)$在ξ_p处连续且不为零, 那么当$n \to \infty$时, 有

$$\sqrt{n}(m_p - \xi_p) \xrightarrow{D} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right).$$

即

$$m_p \sim AN\left(\xi_p, \frac{p(1-p)}{nf^2(\xi_p)}\right).$$

特别地

$$m_e \sim AN\left(\xi_{0.5}, \frac{1}{4nf^2(\xi_{0.5})}\right).$$