第七章 数理统计的基本概念

数理统计的分类

描述统计学

对随机现象进行观测、试验，以取得有代表性的观测值

推断统计学

对已取得的观测值进行整理、分析,作出推断、决策,从而找出所研究的对象的规律性

- 推断统计学
  - 参数估计
  - 假设检验
  - 方差分析
  - 回归分析

§ 7.1 基本概念

总体和样本

总体 —— 研究对象的全体组成的集合

一般地，研究对象全体的某个(或某些)数量指标,是一个随机变量(或多维随机变量).

若为一个随机变量,可记为X. 例如, 某钢铁厂生产的钢锭的强度.

X 的分布函数和数字特征称为总体的分布函数和数字特征.

个体 —— 组成总体的每一个元素(单元)

总体中每个元素的数量指标,可以看作随机变量 X 的某个取值.用x_i表示.

抽样 —— 从总体中抽取个体，做随机试验并记录其结果

样本 ——从总体中抽取的部分个体.

假设抽取n个个体，则每一个体的数量指标分别为一个随机变量，记为：$X_1, X_2,\ldots ,X_n$，则$(X_1, X_2,\ldots ,X_n)$为样本变量，简称样本.

n 称为样本容量.

样本值 ——
从总体中抽取的部分个体.其数量指标的观察值$(x_1, x_2, \ldots, x_n)$，称为总体 X 的一个容量为n的**样本值**.

或称为样本$(X_1, X_2, \ldots, X_n)$的一个实现.

样本空间 —— 样本$(X_1, X_2, \ldots, X_n)$所有可能取值的集合.

简单随机样本

设$(X_1, X_2,\ldots, X_n)$是来自总体 X 的一个样本,它满足:

(1) 同分布：$X_1, X_2,\ldots, X_n$都与X 有相同的分布

(2) 独立性: $X_1, X_2,\ldots, X_n$相互独立

则称$(X_1, X_2,\ldots, X_n)$为简单随机样本.

一般地,对有限总体,采用放回抽样所得到的样本为简单随机样本,但使用不方便,常用不放回抽样代替.当总体中个体的数目N 与样本容量 n 之比$N/n \geq 10$时,可将不放回抽样近似地看作放回抽样.

设总体 X 的分布函数为$F(x)$, $(X_1, X_2, \ldots, X_n)$为总体 X 的简单随机样本,

则$(X_1, X_2, \ldots, X_n)$的联合分布函数为

$$F_{总}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

若总体X的概率密度函数为$f(x)$,

则$(X_1, X_2, \ldots, X_n)$的联合概率密度函数为

$$f_{总}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

统计量

定义

设$(X_1, X_2,\ldots ,X_n)$是取自总体X的一个样本,

$$g(r_1, r_2, \cdots, r_n)$$

为一实值连续函数,且不含有未知参数,

则称随机变量$g(X_1, X_2,\ldots ,X_n)$为**统计量**.

若$(x_1, x_2,\ldots ,x_n)$是一个样本值,

则称 $g(x_1, x_2, \cdots, x_n)$

为统计量$g(X_1, X_2,\ldots ,X_n)$的一个样本值

例 $X \sim N(\mu,\sigma^2), \mu,\sigma^2$ 是未知参数,

$(X_1, X_2, \cdots, X_n)$ 是一样本,则

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

是统计量, 其中 $X_i \sim N(\mu,\sigma^2)$

但 $\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$ 不是统计量.

若 μ ,σ 已知,则为统计量

常用的统计量

设$(X_1, X_2,\dots ,X_n)$是来自总体 X 的容量为 n 的样本,

称统计量

(1) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 为**样本均值**

(2) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ 为**样本方差**

$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$ 为**样本标准差**

(3) $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 为样本的 k 阶**原点矩**

(4) $B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$ 为样本的 k 阶**中心矩**

例如

$$A_1 = \overline{X}$$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

而样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

注 样本方差 S^2 与样本二阶中心矩 S_n^2 不同:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

而 $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

1）关系式 $S^2 = \dfrac{n}{n-1} S_n^2$

2） 设 $E(X)=\mu\ ,D(X)=\sigma^2$ 则

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

$$\sum_{i=1}^{n}X_i=n\overline{X}$$

推导如下：

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}(X_i^2-2X_i\overline{X}+\overline{X}^2)$$

$$=\frac{1}{n}\left(\sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\overline{X}^2\right)=\frac{1}{n}\left(\sum_{i=1}^{n}X_i^2-2n\overline{X}^2+n\overline{X}^2\right)$$

$$=\frac{1}{n}\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right)=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2$$

则

$$E\left(\overline{X}\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\mu \qquad D\left(\overline{X}\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n}\sigma^2$$

$$E(S_n^2)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-E\left(\overline{X}^2\right)$$

$$=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-\left[D\left(\overline{X}\right)+E^2\left(\overline{X}\right)\right]$$

$$EX_i^2$$
$$=E^2(X_i)+DX_i$$
$$=\sigma^2+\mu^2$$

$$=\sigma^2+\mu^2-\left(\frac{1}{n}\sigma^2+\mu^2\right)=\frac{n-1}{n}\sigma^2$$

$$E(S^2)=E\left[\frac{n}{n-1}S_n^2\right]=\frac{n}{n-1}ES_n^2=\sigma^2$$

(5) **顺序统计量与极差**

设$(X_1, X_2, \cdots, X_n)$为样本,

$(x_1, x_2, \cdots, x_n)$为样本值,且 $x_1^* \le x_2^* \le \cdots \le x_n^*$

当$(X_1, X_2, \cdots, X_n)$取值为$(x_1, x_2, \cdots, x_n)$时

定义随机变量 $X_{(k)} = x_k^*, k = 1, 2, \cdots, n$

则称统计量$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$为**顺序统计量**.

其中, $X_{(1)} = \min\limits_{1 \le k \le n}\{X_k\}$, $X_{(n)} = \max\limits_{1 \le k \le n}\{X_k\}$

称$D_n = X_{(n)} - X_{(1)}$为**极差**

例 在总体 $N(52,\ 6.3^2)$ 中,随机地抽取一个容量为36的样本,求样本均值 $\overline{X}$ 落在50.8到53.8之间的概率

解 $$\overline{X} \sim N(52,\ \frac{6.3^2}{36})$$

故 $$P(50.8 < \overline{X} < 53.8) = F_{\overline{X}}(53.8) - F_{\overline{X}}(50.8)$$

$$= \Phi\left(\frac{53.8-52}{\frac{6.3}{6}}\right) - \Phi\left(\frac{50.8-52}{\frac{6.3}{6}}\right)$$

$$= \Phi(1.7143) - \Phi(-1.1429)$$

$$= 0.8239$$

§ 7.3 统计量的分布

确定统计量的分布—— 抽样分布, 是数理统计的基本问题之一. 采用求随机向量的函数的分布的方法可得到抽样分布. 由于样本容量一般不止 2 或 3 (甚至还可能是随机的), 故计算往往很复杂, 有时还需要特殊技巧或特殊工具.

由于正态总体是最常见的总体, 故本节介绍的几个抽样分布均对正态总体而言.

统计中常用分布

(1) 正态分布

若 $X_1, X_2, \cdots, X_n \sim N(\mu_i, \sigma_i^2)$

则 $$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

independent and identically distributed

特别地,

若 $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$

则 $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(2) $\chi^2(n)$ 分布(n为自由度)

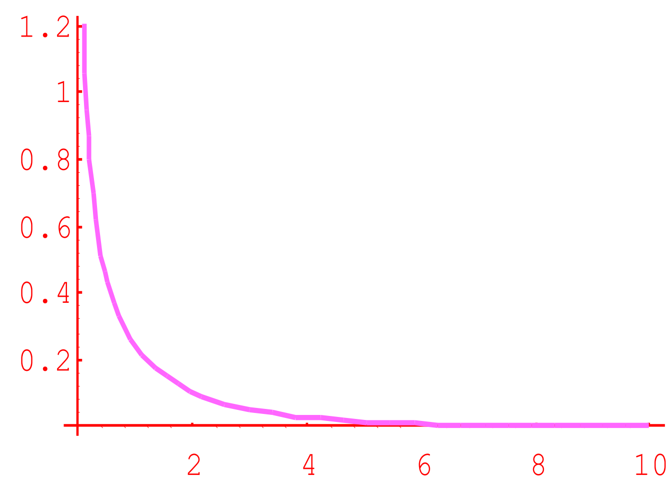
定义 设$(X_1, X_2,\dots ,X_n)$相互独立,

且都服从标准正态分布$N(0,1)$,则称

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

$n = 1$ 时,其密度函数为

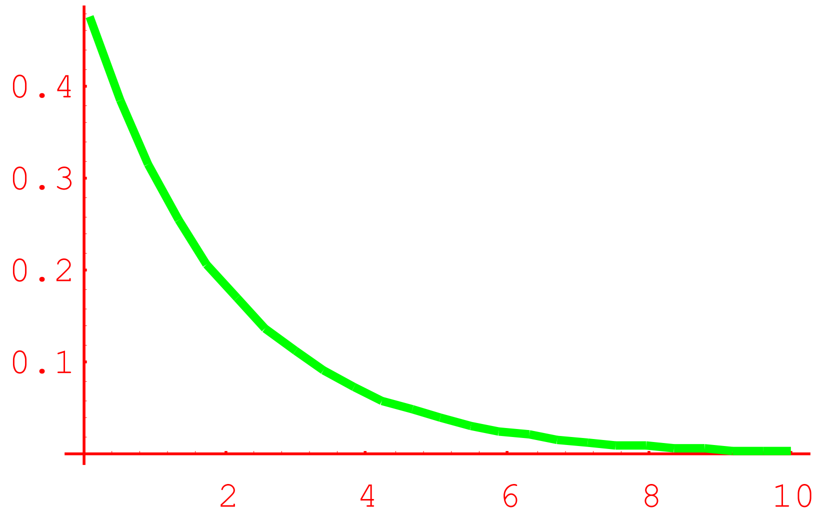
$$f(x)=\begin{cases} \dfrac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}, & x>0 \\ 0, & x \le 0 \end{cases}$$



$n=2$ 时,其密度函数为

$$f(x)=\begin{cases}\dfrac{1}{2}e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$$

为参数为1/2的指数分布.



一般地, 自由度为 n 的$\chi^2(n)$的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}, & x>0\\ 0, & x\le 0\end{cases}$$

其中，$\Gamma(x)=\int_0^{+\infty}t^{x-1}e^{-t}dt$

在$x>0$时收敛，称为Γ函数.

$\chi^2(n)$ 分布的性质

1、$E\left(\chi^2(n)\right)=n, D\left(\chi^2(n)\right)=2n$

2、若 $X_1=\chi^2(n_1), X_2=\chi^2(n_2)$,

X_1, X_2相互独立，则

$X_1+X_2\sim\chi^2(n_1+n_2)$

3、 $\chi^2(n)$的α分位数有表可查。

n>45时 $\chi_\alpha^2(n)\approx\frac{1}{2}(z_\alpha+\sqrt{2n-1})^2$

证 1° 设 $\chi^2(n)=\sum_{i=1}^{n}X_i^2$ $X_i\sim N(0,1)$ $i=1,2,\cdots,n$

$X_1,X_2,\cdots,X_n$ 相互独立,

则 $E(X_i)=0,\ D(X_i)=1,\ E(X_i^2)=1$

$$E\left(\chi^2(n)\right)=E\left(\sum_{i=1}^{n}X_i^2\right)=n$$

$$E(X_i^4)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x^4e^{-\frac{x^2}{2}}\mathrm{d}x=3$$

$$D(X_i^2)=E(X_i^4)-E^2(X_i^2)=2$$

$$D\left(\chi^2(n)\right)=D\left(\sum_{i=1}^{n}X_i^2\right)=2n$$

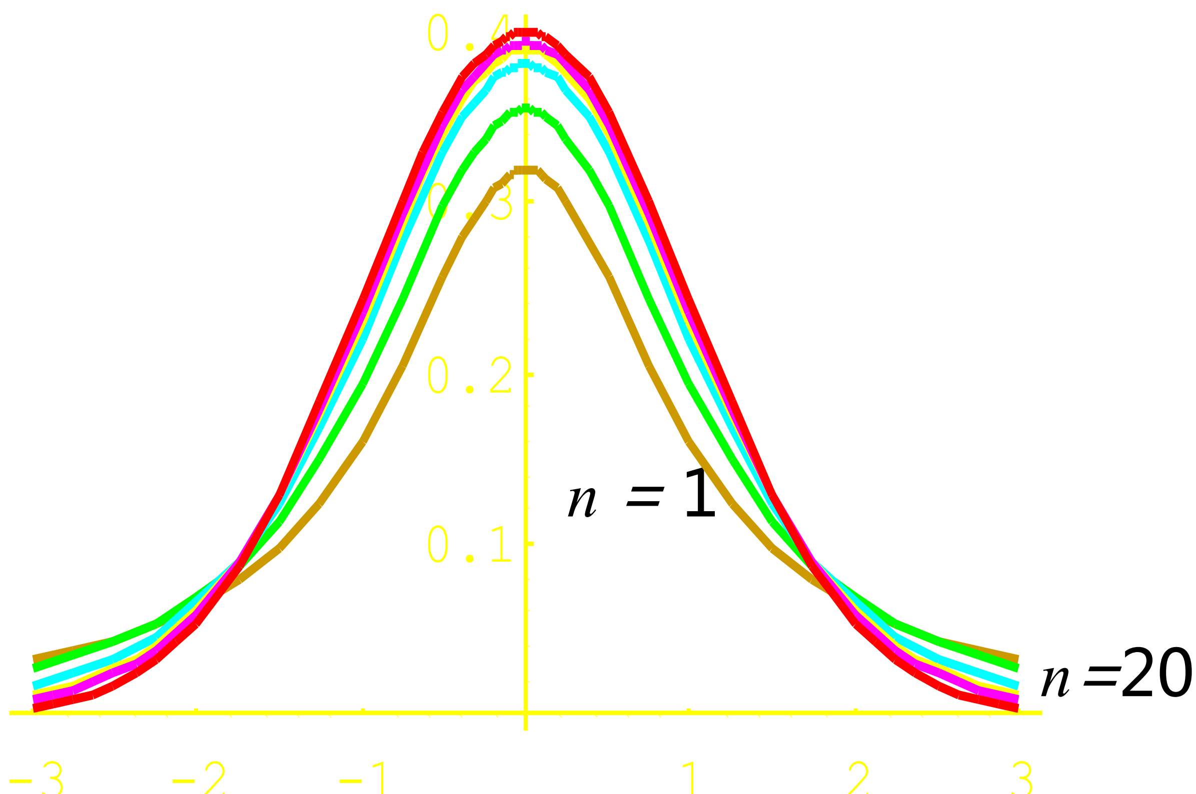
(3) t 分布 (Student 分布)

定义 设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$, X, Y 相互独立,

$$T = \frac{X}{\sqrt{Y/n}}$$

则T所服从的分布称为自由度为 n 的T分布
其密度函数为

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad -\infty < t < \infty$$



t 分布的图形(红色的是标准正态分布)

t 分布的性质

$1°\ f_n(t)$是偶函数,

$$n \to \infty, f_n(t) \to \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

$2°\ T$分布的α分位数t_α有表可查

(4) F 分布

定义 设 $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, X ，Y 相互独立，

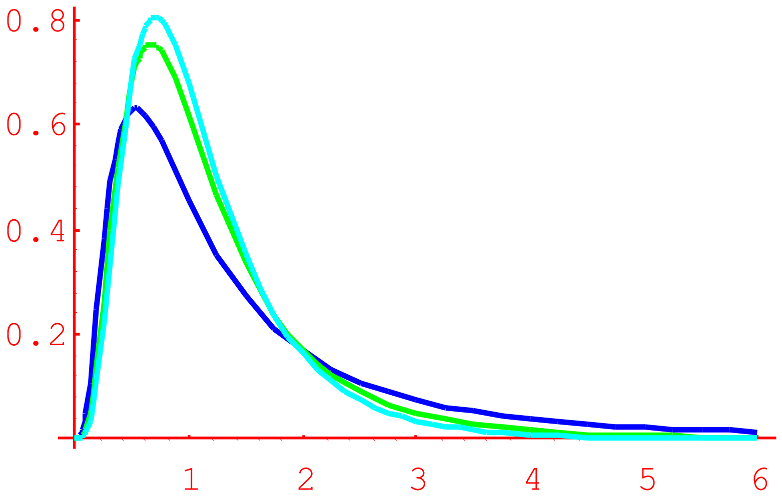
令 $$F = \frac{X/n}{Y/m}$$

则F 所服从的分布称为**第一自由度为n ，第二自由度为 m 的F 分布**

记为 $F \sim F(n,m)$

其密度函数为

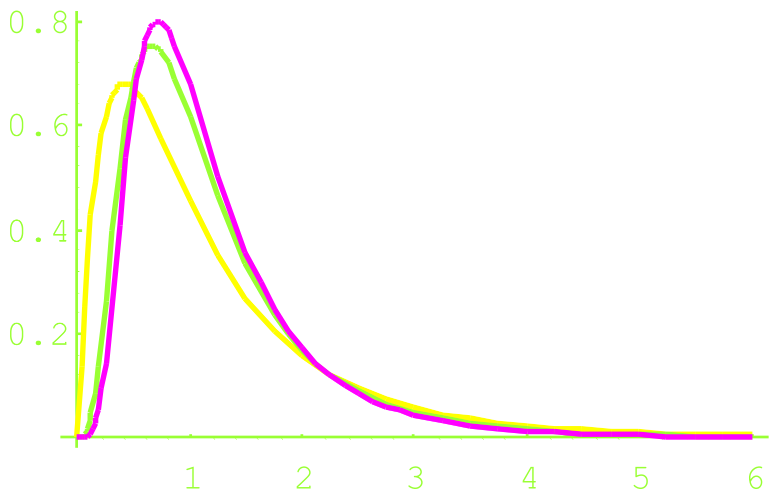
$$f(t,n,m)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n+m}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)\Gamma\left(\dfrac{m}{2}\right)}\left(\dfrac{n}{m}\right)^{\frac{n}{2}}t^{\frac{n}{2}-1}\left(1+\dfrac{n}{m}t\right)^{-\frac{n+m}{2}} & t>0\\ 0, & t\le 0\end{cases}$$



$m = 10, n = 4$

$m = 10, n = 10$

$m = 10, n = 15$



$m = 4, n = 10$

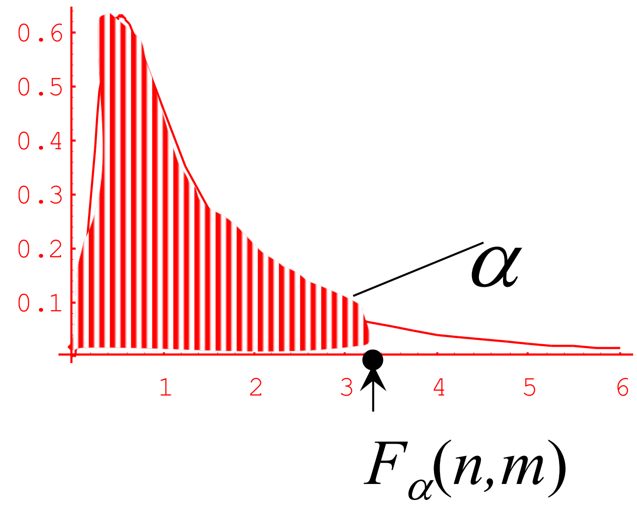
$m = 10, n = 10$

$m = 15, n = 10$

F 分布的性质

1、$F(n,m)$的α分位数$F_\alpha(n,m)$有表可查:

$$P(F \le F_\alpha(n,m)) = \alpha$$



2、若$F\sim F(n, m)$,则 $\frac{1}{F} \sim F(m,n)$

事实上，若 $F \sim F(n,m)$

则可设 $F = \frac{X/n}{Y/m}$

故有 $\frac{1}{F} = \frac{Y/m}{X/n} \sim F(m,n)$

从而，在查表时可能用到结论：$F_{1-\alpha}(n,m)=\dfrac{1}{F_{\alpha}(m,n)}$

$$1-\alpha=P(F\le F_{1-\alpha}(n,m))=P\left(\frac{1}{F}>\frac{1}{F_{1-\alpha}(n,m)}\right)$$

$$=1-P\left(\frac{1}{F}\le\frac{1}{F_{1-\alpha}(n,m)}\right)$$

即 $P\left(\dfrac{1}{F}\le\dfrac{1}{F_{1-\alpha}(n,m)}\right)=\alpha$ 而 $\dfrac{1}{F}\sim F(m,n)$

即 $F_{\alpha}(m,n)=\dfrac{1}{F_{1-\alpha}(n,m)}$ $F_{1-\alpha}(n,m)=\dfrac{1}{F_{\alpha}(m,n)}$

例如 $F_{0.95}(5,4)=6.26$

故 $F_{0.05}(4,5)=\dfrac{1}{F_{0.95}(5,4)}=0.159$