

# 第6章 数理统计基本概念 抽样分布 (Sampling distribution)

1. 总体和随机样本.(Population and Random Samples)
2. 统计量 Statistic,
3. 抽样分布 Sampling distribution.
4. 单个正态总体统计量的分布 (The distribution of the statistics of single normal population)

# 1. 总体和随机样本 Population and Random Sample.

考虑一个随机试验:

令  $X$  为该随机试验对应的随机变量, 通常称为**总体**。其元素通常称为**个体** (individuals)

重复该随机试验  $n$  次, 令  $X_k$  为第  $k$  次试验对应的随机变量.

这些随机变量的集合  $\{X_1, X_2, \dots, X_n\}$  就称为总体  $X$  的一个容量为  $n$  的**随机样本**。简记为  $X_1, X_2, \dots, X_n$

随机样本  $X_1, X_2, \dots, X_n$  是独立同分布的 (independent and identically distributed): 如果  $F(x)$  是总体  $X$  的分布函数, 那么  $F(x)$  也是  $X_i$  的分布函数

每一个  $X_i$  通常称为总体  $X$  的一次**观察** (observation)

观察的结果记为  $x_1, x_2, \dots, x_n$ , 称为  $X_1, X_2, \dots, X_n$  的样本值

在获得样本值以前, 由于随机试验具有不确定性, 因此在观察结果出来以前, 每一次观察通常被视为随机变量

总体可以是有限的、或者无限的

If the CDF of  $X$  is  $F(x)$ , then the joint CDF of  $X_1, X_2, \dots, X_n$  is

$$F(x_1, x_2, \dots, x_n) = F(x_1)F(x_2) \cdots F(x_n)$$

If a pdf of  $X$  is  $f(x)$ , then a joint pdf of  $X_1, X_2, \dots, X_n$  is

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n)$$

## 2. 统计量 (Statistic) .

设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的随机样本.

$g(X_1, X_2, \dots, X_n)$  是  $X_1, X_2, \dots, X_n$  的一个函数

如果  $g(X_1, X_2, \dots, X_n)$  中不含有未知的参数, 则称为  $X_1, X_2, \dots, X_n$  的一个统计量 (statistic) .

$g(x_1, x_2, \dots, x_n)$  称为统计量  $g(X_1, X_2, \dots, X_n)$  的观测值, 如果  $x_1, x_2, \dots, x_n$  是  $X_1, X_2, \dots, X_n$  的样本值

i.e.  $X_1 + 2X_2X_3 - X_n$ ,  $X_1X_2X_3 \cdots X_n$ ,  $X_1 + X_2 + \cdots + X_n$

# 一些重要的统计量

观测值

(1) 样本均值 (The sample mean)

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

(2) 样本方差 The sample variance

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

$$\hat{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

### (3) 样本的标准差

观测值

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{S} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{s} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

(4)  $k$  阶样本原点矩 The  $k$ th moment of the sample about origin

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

$$a_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

(5)  $k$  阶样本中心矩 The  $k$ th moment of the sample about center

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

### 3. 抽样分布(Sampling distribution)

(1) 卡方分布,  $\chi^2$  distribution, Chi-square distribution /Kai/

$X_1, X_2, \dots, X_n$  是来自具有标准正态分布的总体  $X \sim N(0,1)$  随机样本

那么, 统计量  $\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$  服从自由度为  $n$

的卡方分布. ( $\chi^2$  distribution) 表示为  $\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2 \sim \chi^2(n)$

$$X_i \sim N(0,1), \quad E(X_i) = 0, \quad D(X_i) = 1;$$

$$\text{由 } D(X_i) = E(X_i^2) - E^2(X_i) \quad \text{得 } E(X_i^2) = 1$$

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

$$E(\chi^2) = n \quad \text{Var}(\chi^2) = 2n$$

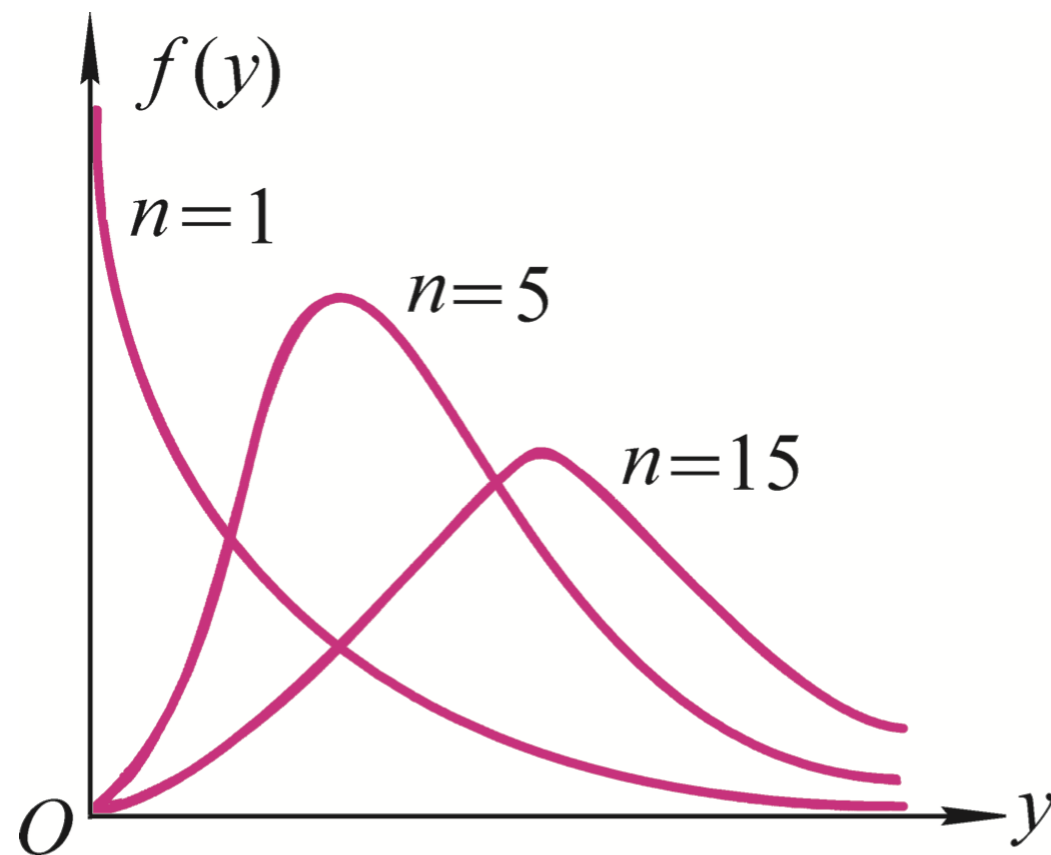


**定理** 假设  $\chi_1^2 \sim \chi^2(n_1)$ ,  $\chi_2^2 \sim \chi^2(n_2)$ , 如果  $\chi_1^2$  和  $\chi_2^2$  相互独立  
则  $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$

$\chi^2$  是连续型随机变量, 其密度函数为

$$f(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, & y > 0 \\ 0, & \text{其他} \end{cases}$$

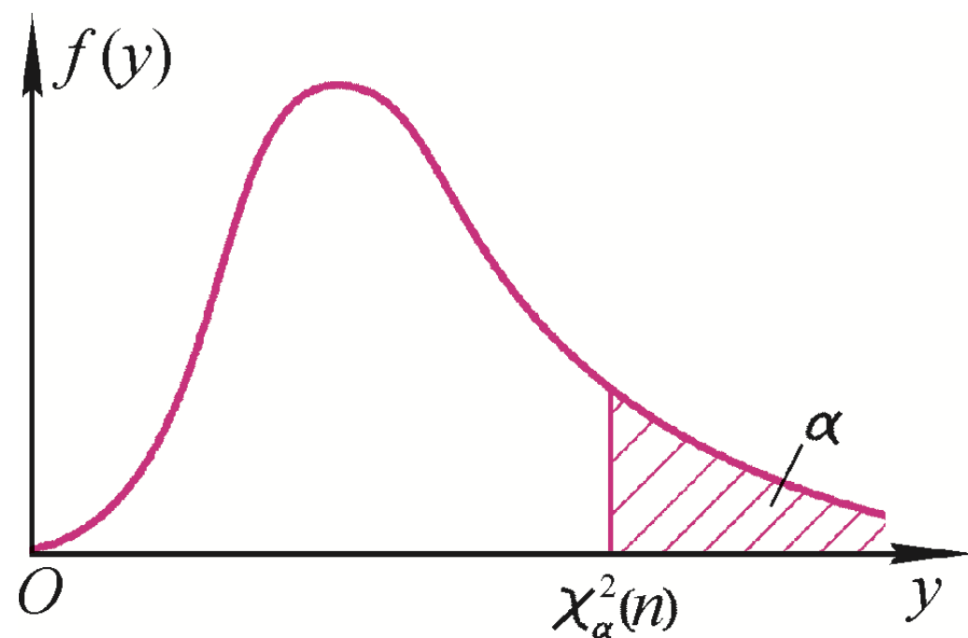
$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt, x > 0$$



定义:  $\alpha$  分位点  $\chi^2_{\alpha}(n)$ , 密度曲线位于该点右侧的面积为  $\alpha$

Upper  $\alpha$  quantile

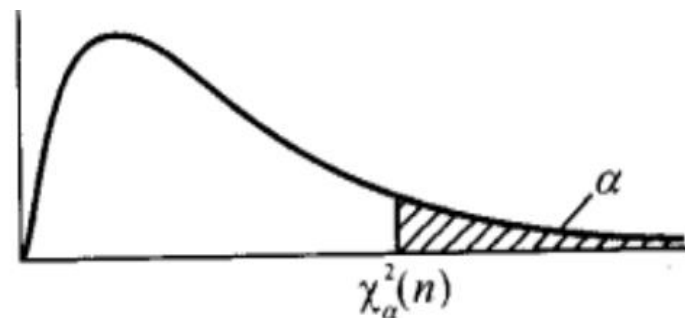
$$P(\chi^2 > \chi^2_{\alpha}(n)) = \int_{\chi^2_{\alpha}(n)}^{+\infty} f(y) dy = \alpha$$



$$P(\chi^2 \leq \chi^2_{\alpha}(n)) = 1 - \alpha$$

$$\chi^2_{0.05}(9)$$

$$= 16.919$$



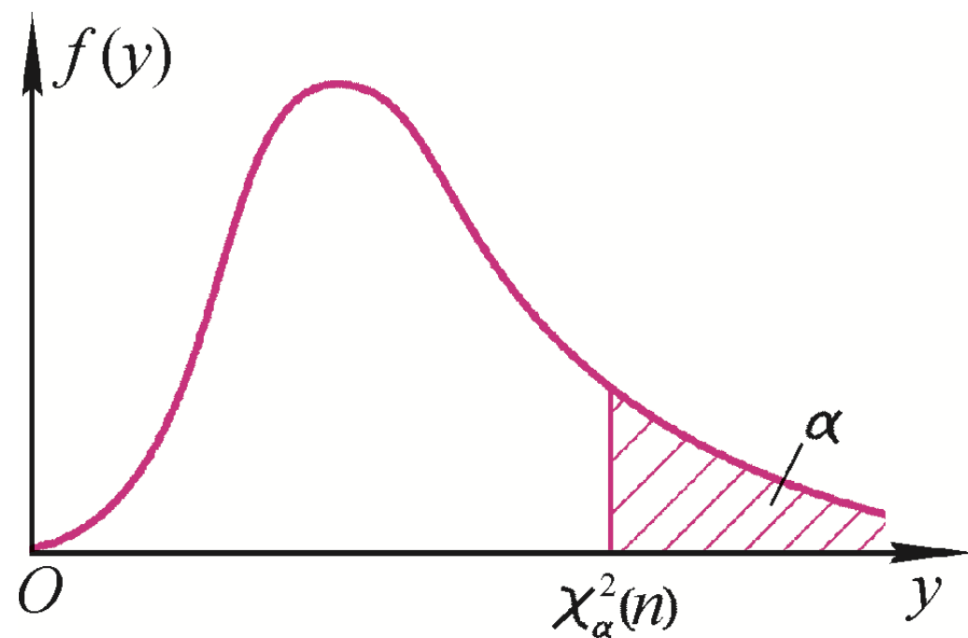
$$P\{\chi^2(n) > \chi^2_\alpha(n)\} = \alpha$$

$\alpha \backslash n$	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799

when  $n > 45$ ,

$$\chi^2_{\alpha}(n) \approx \frac{1}{2} \left( z_{\alpha} + \sqrt{2n-1} \right)^2$$

$z_{\alpha}$  是标准正态分布的  $\alpha$  分位点



$$\chi^2_{0.05}(50) \approx \frac{1}{2} \left( 1.645 + \sqrt{99} \right)^2 = 67.221$$

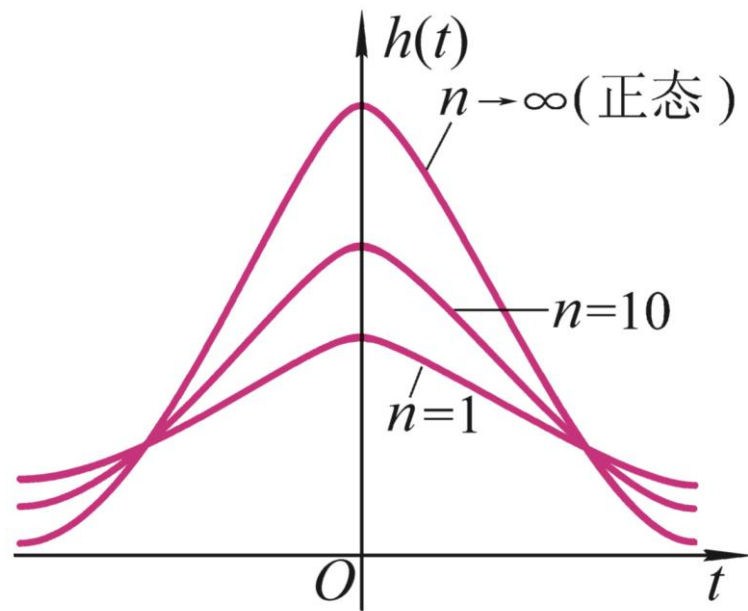
(2)  $t$ 分布, 或  $t$ -distribution, or student distribution

William Sealy Gosset, 1908年发表

$X \sim N(0,1), Y \sim \chi^2(n)$ ,  $X$  and  $Y$  are independent

Then

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



obeys a  $t$ -distribution with a degree of freedom  $n$ , denoted by a  $t \sim t(n)$ ,  $t$  is a continuous random variable. A pdf of  $t$

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

当  $n$  趋于无穷的时候,  $t$ -分布的极限分布就是标准正态分布  
The limit distribution of the  $t$ -distribution as  $n \rightarrow \infty$  is the standard normal distribution

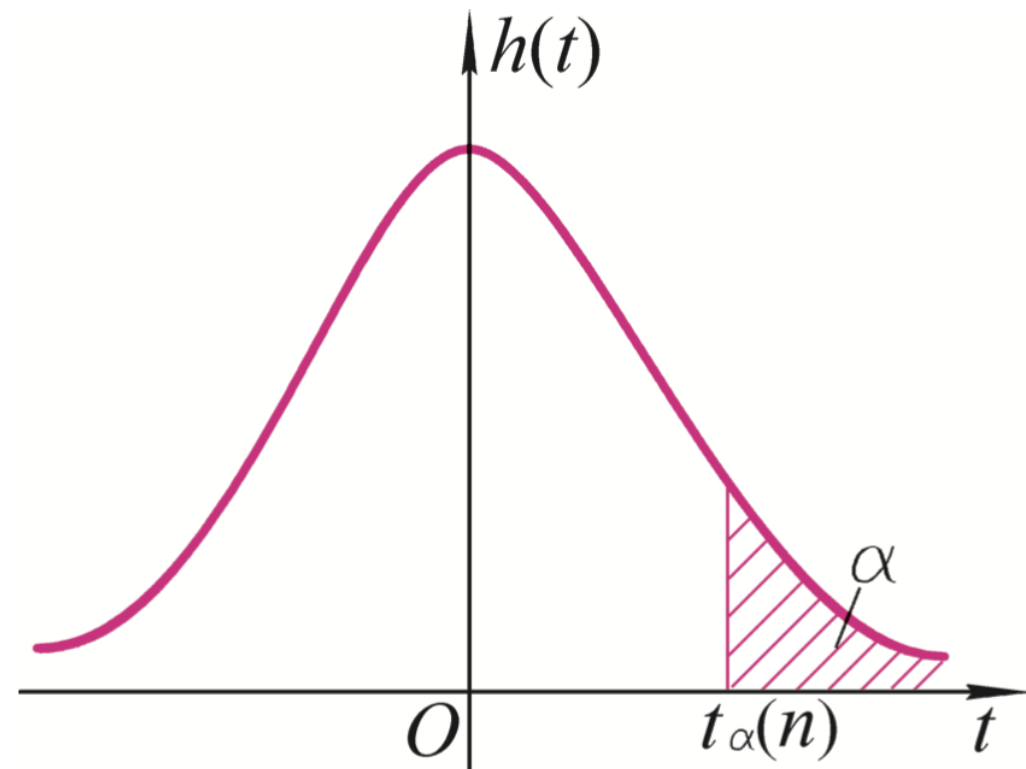
定义:  $\alpha$  分位点 ( $0 < \alpha < 1$ )

$$P(t > t_{\alpha}(n)) = \int_{t_{\alpha}(n)}^{+\infty} h(t)dt = \alpha$$

$h(t)$  的对称性表明

$$t_{1-\alpha}(n) = -t_{\alpha}(n)$$

当  $n > 45$ ,  $t_{\alpha}(n) \approx z_{\alpha}$



### (3) $F$ 分布, $F$ -distribution

$U \sim \chi^2(n_1), V \sim \chi^2(n_2),$   $U$  和  $V$  相互独立

则  $F = \frac{U/n_1}{V/n_2}$  服从自由度为  $n_1, n_2$  的  $F$ -分布,

记为  $F \sim F(n_1, n_2),$

$F$  是一个连续型随机变量. A possible pdf is

$$\psi(y) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} y^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left[1+\frac{n_1 y}{n_2}\right]^{\frac{n_1+n_2}{2}}}, y > 0; \psi(y) = 0, y \leq 0$$

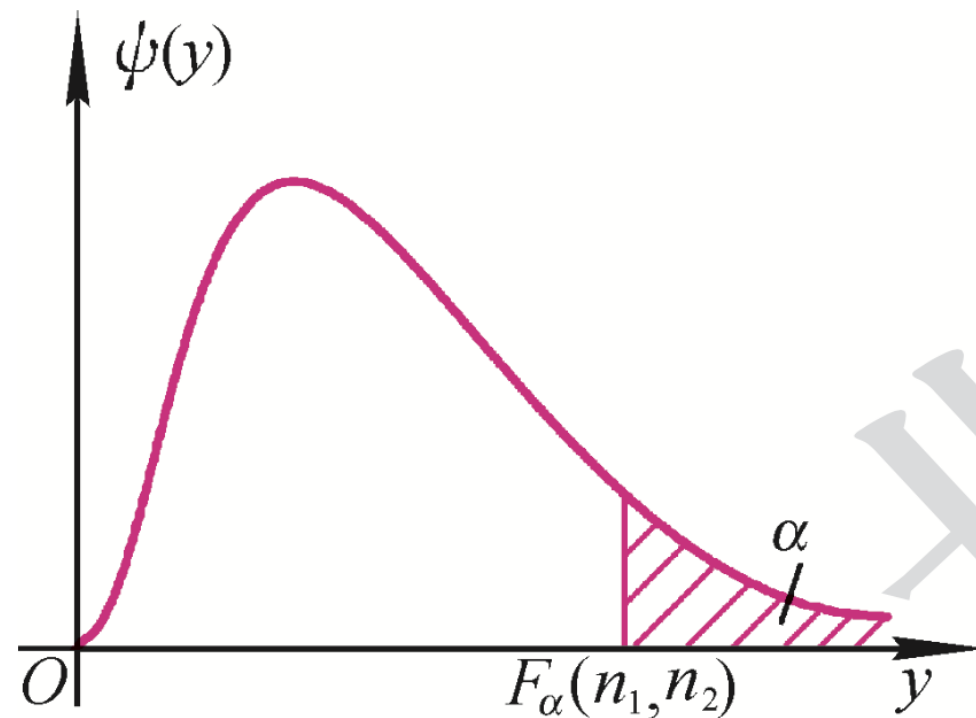
定义:  $\alpha$  分位点

$$P(F > F_{\alpha}(n_1, n_2)) = \int_{F_{\alpha}(n_1, n_2)}^{+\infty} \psi(y) dy = \alpha$$

$$F = \frac{U/n_1}{V/n_2} \sim F(n_1, n_2)$$

$$\frac{1}{F} = \frac{V/n_2}{U/n_1} \sim F(n_2, n_1)$$

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_{\alpha}(n_2, n_1)}$$



$$F_{0.95}(12, 9) = \frac{1}{F_{0.05}(9, 12)} = \frac{1}{2.80} \approx 0.357$$



## 4. 单个正态总体样本均值与样本方差的分布

$X_1, X_2, \dots, X_n$  是来自均值  $\mu$ , 方差  $\sigma^2$  的正态总体的简单随机样本

$$(1) \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ 或 } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$(2) \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$(3) \bar{X} \text{ 和 } \frac{(n-1)S^2}{\sigma^2} \text{ 相互独立}$$

$$(4) \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

$$(1) \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ 或 } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \quad \text{由于} \quad X_i \sim N(\mu, \sigma^2)$$

$$E(\bar{X}) = E(X_1) + E(X_2) + \cdots + E(X_n) = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2} (Var(X_1) + Var(X_2) + \cdots + Var(X_n)) = \frac{\sigma^2}{n}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{单位化} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$(2) \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \quad (3) \bar{X} \text{ 和 } \frac{(n-1)S^2}{\sigma^2} \text{ 相互独立. 证明略}$$

$$(4) \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} / (n-1)}} \sim t(n-1) \quad \text{化简得 } \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

**例 6.2**

设总体  $X$  服从正态分布  $N(62, 100)$ , 为使样本均值大于 60 的概率不小于 0.95, 问样本容量  $n$  至少应取多大?

**解** 设需要样本容量为  $n$ , 则

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$P(\bar{X} > 60) = P\left(\frac{\bar{X} - 62}{\frac{10}{\sqrt{n}}} > \frac{60 - 62}{\frac{10}{\sqrt{n}}}\right) = 1 - \Phi\left(-\frac{\sqrt{n}}{5}\right) = \Phi\left(\frac{\sqrt{n}}{5}\right)$$

$$\text{已知 } \Phi(1.64) \approx 0.95 \quad \text{故 } \frac{\sqrt{n}}{5} \geq 1.64$$

$$\text{得 } \sqrt{n} \geq 8.2, \quad n \geq 67.24$$

样本的容量至少应该取 68

