第二节　抽样分布

一、基本概念

二、常见分布

三、小结

一、基本概念

1. 统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, $g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数, 若 g 中不含未知参数, 则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量.

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值, 则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

实例1 设 X_1,X_2,X_3 是来自总体 $N(\mu,\sigma^2)$ 的一个样本, 其中 μ 为已知, σ^2 为未知, 判断下列各式哪些是统计量, 哪些不是?

$$T_1 = X_1, \qquad T_2 = X_1 + X_2 e^{X_3},$$

$$T_3 = \frac{1}{3}(X_1 + X_2 + X_3),$$

$$T_4 = \max(X_1, X_2, X_3), \qquad T_5 = X_1 + X_2 - 2\mu,$$

是

$$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2).$$

不是

2. 几个常用统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值.

(1)样本平均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;

其观察值 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2)样本方差

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

其观察值

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right).$$

(3)样本标准差

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

其观察值 $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\ .$

(4) 样本 k 阶(原点)矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, $k = 1, 2, \cdots$;

其观察值 $\alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$, $k = 1, 2, \cdots$.

(5)样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, \ k = 2, 3, \cdots;$$

其观察值 $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$, $k = 2, 3, \cdots$.

由以上定义得下述结论:

若总体 X 的 k 阶矩 $E(X^k)$ 记成 μ_k 存在,

则当 $n \to \infty$ 时, $A_k \xrightarrow{P} \mu_k$, $k = 1, 2, \cdots$.

证明　因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布,

所以 $X_1^k, X_2^k, \cdots, X_n^k$ 独立且与 X^k 同分布,

故有　$E(X_1^k) = E(X_2^k) = \cdots = E(X_n^k) = \mu_k$.

再根据第五章辛钦定理知

辛钦定理

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k\ , \quad k = 1, 2, \cdots;$$

由第五章关于依概率收敛的序列的性质知

$$\boxed{g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k),}$$

其中 g 是连续函数.

以上结论是下一章所要介绍的矩估计法的理论根据.

3. 经验分布函数

总体分布函数 $F(x)$ 相应的统计量称为经验分布函数.

经验分布函数的做法如下:

设 $X_1, X_2, \cdots, X_n$ 是总体 F 的一个样本,

用 $S(x)\ (-\infty < x < +\infty)$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量的个数,

定义经验分布函数 $F_n(x)$ 为

$$F_n(x) = \frac{1}{n} S(x), \qquad (-\infty < x < +\infty)$$

对于一个样本值, $F_n(x)$ 的观察值容易求得.

($F_n(x)$ 的观察值仍以 $F_n(x)$ 表示.)

实例2 设总体 F 具有一个样本值 1, 2, 3,

则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{1}{3}, & 1\le x<2,\\ \dfrac{2}{3}, & 2\le x<3,\\ 1, & x\ge 3.\end{cases}$$

实例3 设总体 F 具有一个样本值 1, 1, 2,

则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{2}{3}, & 1\le x<2,\\ 1, & x\ge 2.\end{cases}$$

一般地,

设 $x_1, x_2, \cdots, x_n$ 是总体 F 的一个容量为 n 样本值,

先将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列,

并重新编号, $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$,

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, \\ 1, & x \ge x_{(n)}. \end{cases}$$

格里汶科定理

格里汶科

对于任一实数 x, 当 $n \to \infty$ 时, $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$, 即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} \left|F_n(x) - F(x)\right| = 0\right\} = 1.$$

对于任一实数 x 当 n 充分大时, 经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别, 从而在实际上可当作 $F(x)$ 来使用.

二、常见分布

统计量的分布称为抽样分布.

1. χ^2分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0,1)$ 的样本，则称统计量 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 服从自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$.

自由度：

指 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 中右端包含独立变量的个数.

随机数演示

分布函数与密度函数演示

$\chi^2(n)$分布的概率密度为

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

证明　因为 $\chi^2(1)$ 分布即为 $\Gamma\left(\dfrac{1}{2},2\right)$ 分布,

又因为 $X_i\sim N(0,1)$,　由定义 $X_i^2\sim\chi^2(1)$,

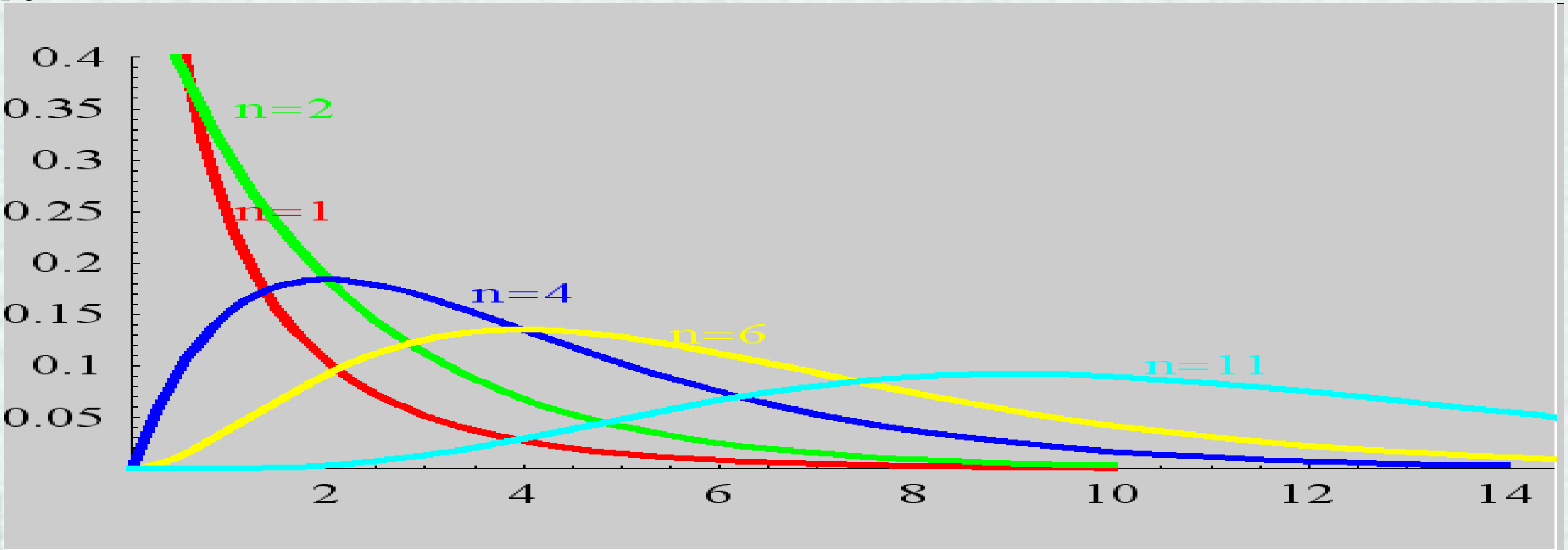
即 $X_i^2\sim\Gamma\left(\dfrac{1}{2},2\right)$,　$i=1,2,\cdots,n.$

因为 $X_1, X_2, \cdots, X_n$ 相互独立,

所以 $X_1^2, X_2^2, \cdots, X_n^2$ 也相互独立,

根据 Γ 分布的可加性知 $\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \Gamma\left(\frac{n}{2}, 2\right)$.

$\chi^2(n)$ 分布的概率密度曲线如图.



χ^2 分布的性质

性质1 (χ^2 分布的可加性)

设 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$, 并且 χ_1^2, χ_2^2 独立, 则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$.

(此性质可以推广到多个随机变量的情形.)

设 $\chi_i^2 \sim \chi^2(n_i)$, 并且 χ_i^2 $(i = 1, 2, \cdots, m)$ 相互独立, 则 $\sum_{i=1}^{m} \chi_i^2 \sim \chi^2(n_1 + n_2 + \cdots + n_m)$.

性质2 (χ^2分布的数学期望和方差)

若 $\chi^2 \sim \chi^2(n)$, 则 $E(\chi^2) = n, \quad D(\chi^2) = 2n.$

证明 因为 $X_i \sim N(0,1)$, 所以 $E(X_i^2) = D(X_i) = 1,$

$$D(X_i^2) = E(X_i^4) - [E(X_i^2)]^2 = 3 - 1 = 2, \; i = 1, 2, \cdots, n.$$

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

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

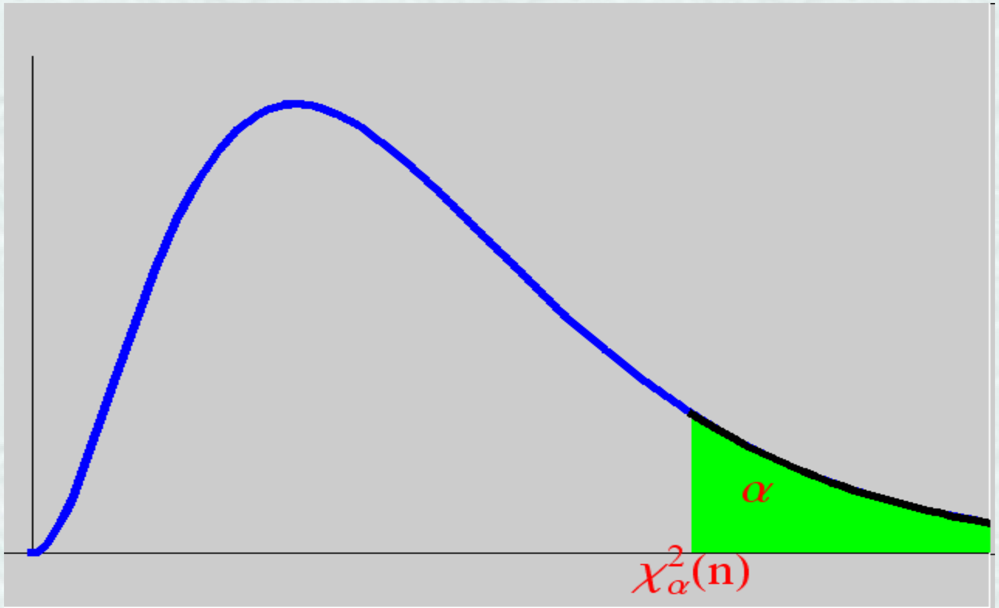
χ^2 分布的分位点

对于给定的正数α, $0<\alpha<1$, 称满足条件

$$P\{\chi^2>\chi_\alpha^2(n)\}=\int_{\chi_\alpha^2(n)}^{\infty}f(y)\mathrm{d}y=\alpha$$

的点$\chi_\alpha^2(n)$为$\chi^2(n)$分布的上α分位点.

对于不同的α, n,
可以通过查表求
得上α分位点的值.

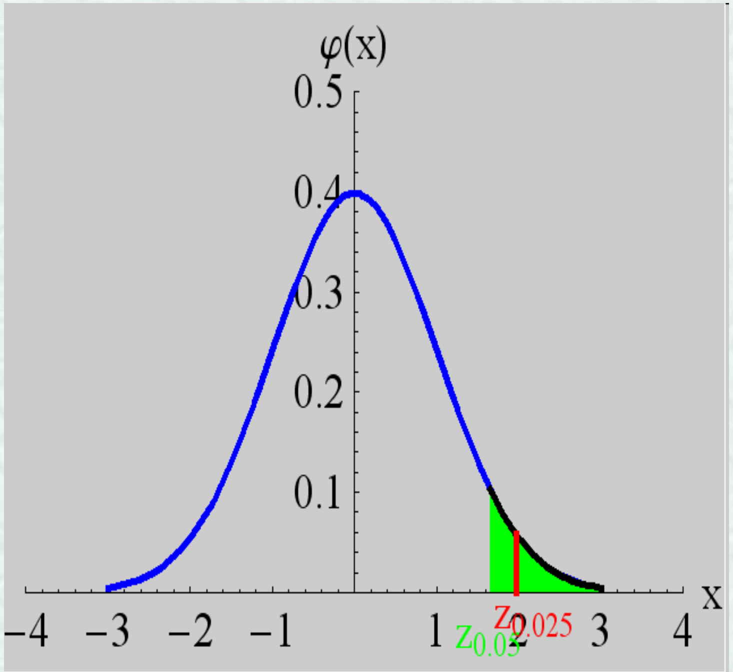


例1　设 X 服从标准正态分布 $N(0,1)$, $N(0,1)$ 的上 α 分位点 z_α 满足 $P\{X > z_\alpha\} = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{z_\alpha}^{+\infty} e^{-\frac{x^2}{2}}\,dx = \alpha$,

求 z_α 的值, 可通过查表完成.

$z_{0.05} = 1.645$,　　附表2-1

$z_{0.025} = 1.96$,　　附表2-2



根据正态分布的对称性知

$$z_{1-\alpha} = -z_\alpha.$$

例2 设 $Z \sim \chi^2(n)$, $\chi^2(n)$ 的上 α 分位点满足

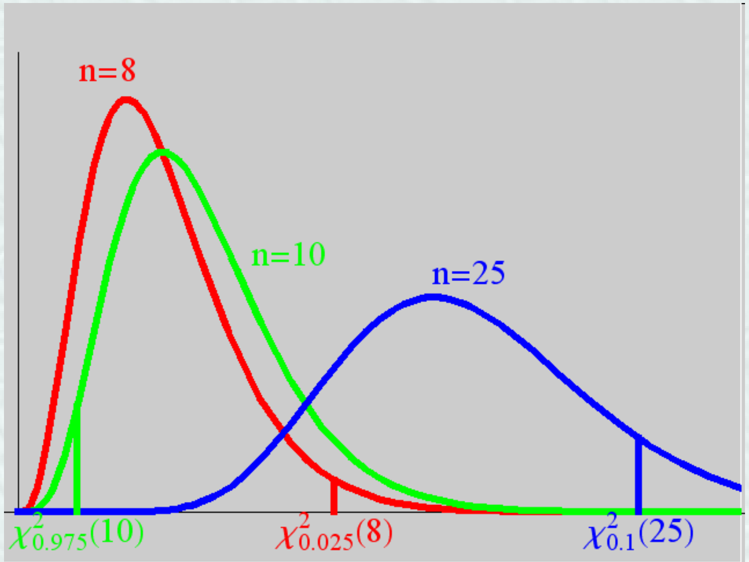
$$P\{Z > \chi^2_\alpha(n)\} = \int_{\chi^2_\alpha(n)}^{+\infty} \chi^2(y;n)\mathrm{d}y = \alpha,$$

求 $\chi^2_\alpha(n)$ 的值, 可通过查表完成.

$\chi^2_{0.025}(8) = 17.535,$ 附表4-1

$\chi^2_{0.975}(10) = 3.247,$ 附表4-2

$\chi^2_{0.1}(25) = 34.382.$ 附表4-3



附表4只详列到 n=45 为止.

在Matlab中求解

费舍尔(R.A.Fisher)证明:

费舍尔资料

当 n 充分大时, $\chi_{\alpha}^{2}(n) \approx \frac{1}{2}(z_{\alpha}+\sqrt{2n-1})^{2}$.

其中 z_{α} 是标准正态分布的上 α 分位点.

利用上面公式,

可以求得 $n>45$ 时, 上 α 分位点的近似值.

例如 $\chi_{0.05}^{2}(50) \approx \frac{1}{2}(1.645+\sqrt{99})^{2}=67.221.$

而查详表可得 $\chi_{0.05}^{2}(50)=67.505$.

2. t 分布

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

则称随机变量 $t = \dfrac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t

分布, 记为 $t \sim t(n)$.

学生氏资料

t 分布又称学生氏(Student)分布.

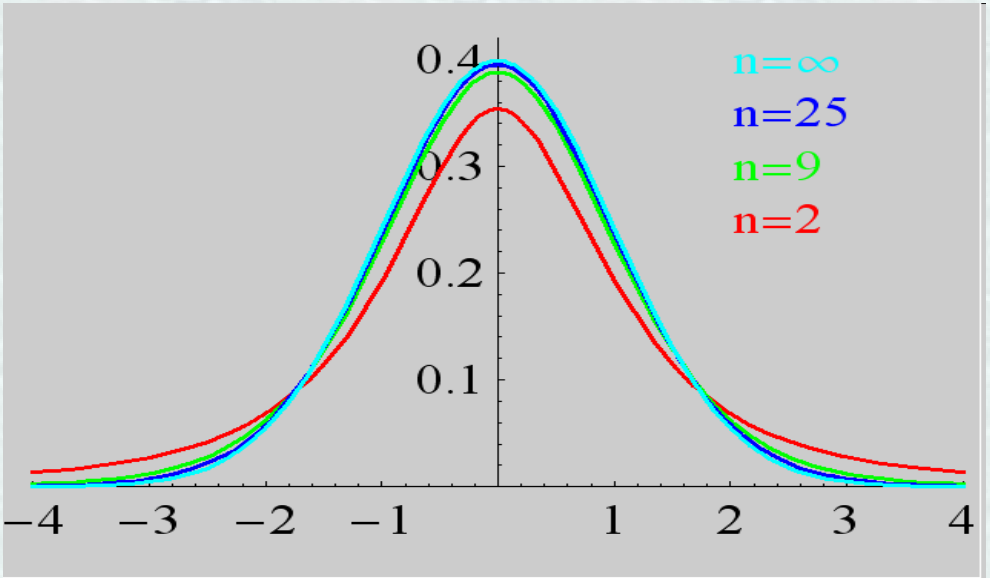
随机数演示

$t(n)$ 分布的概率密度函数为

分布函数与密度函数演示

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

t 分布的概率密度曲线如图



显然图形是关于 $t=0$ 对称的.

当 n 充分大时, 其图形类似于标准正态变量概率密度的图形.

因为 $\lim\limits_{n\to\infty} h(t) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$,

所以当 n 足够大时 t 分布近似于 $N(0,1)$ 分布,

但对于较小的 n, t 分布与 $N(0,1)$ 分布相差很大.

t 分布的分位点

对于给定的 α, $0<\alpha<1$, 称满足条件

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

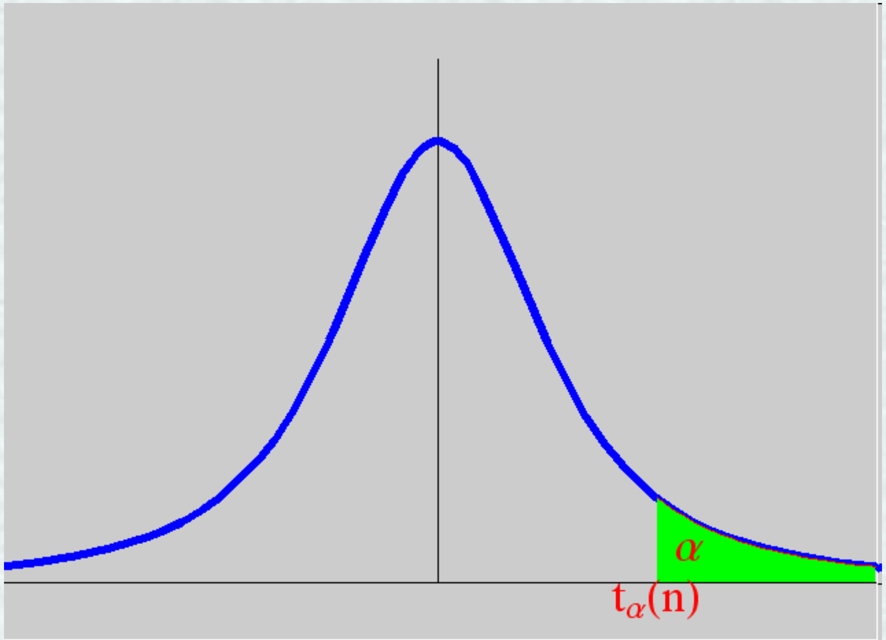
的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位点.

可以通过查表求得上α分位点的值.

由分布的对称性知

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

当 $n>45$ 时, $t_\alpha(n)\approx z_\alpha$.



例3 设 $T \sim t(n)$, $t(n)$ 的上 α 分位点满足

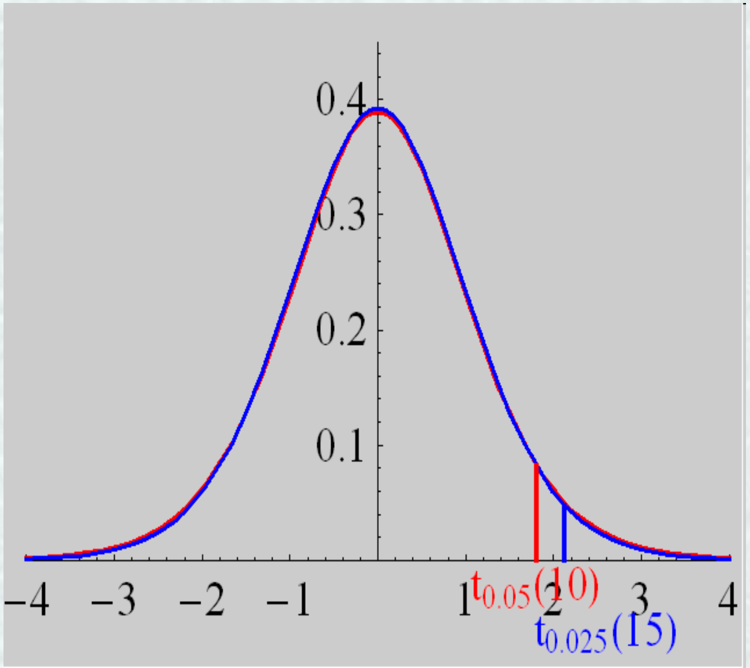
$$P\{T > t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} t(y;n)\mathrm{d}y = \alpha,$$

求 $t_\alpha(n)$ 的值, 可通过查表完成.

$t_{0.05}(10) = 1.8125,$ 附表3-1

$t_{0.025}(15) = 2.1315.$ 附表3-2

在Matlab中求解



3. F分布

设$U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U, V独立, 则称随机变量 $F = \dfrac{U/n_1}{V/n_2}$服从自由度为(n_1, n_2)的F分布, 记为$F \sim F(n_1, n_2)$.

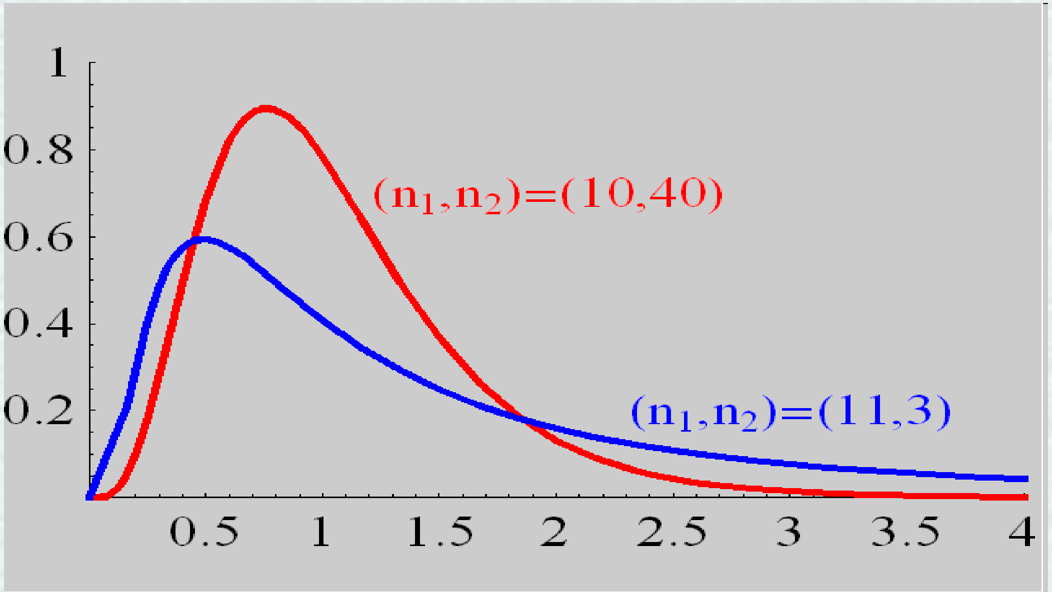
随机数演示

分布函数与密度函数演示

$F(n_1, n_2)$分布的概率密度为

$$\psi(y)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} y^{\frac{n_1}{2}-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left[1+\left(\dfrac{n_1 y}{n_2}\right)\right]^{\frac{n_1+n_2}{2}}}, & y>0, \\ 0, & \text{其他}.\end{cases}$$

F分布的概率密度曲线如图



根据定义可知,

若$F \sim F(n_1, n_2)$,

则$\dfrac{1}{F} \sim F(n_2, n_1)$.

F 分布的分位点

对于给定的α, $0 < \alpha < 1$, 称满足条件

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

的点$F_\alpha(n_1, n_2)$为$F(n_1, n_2)$分布的上α分位点.

例4 设 $F(n_1, n_2)$分布的上α分位点满足

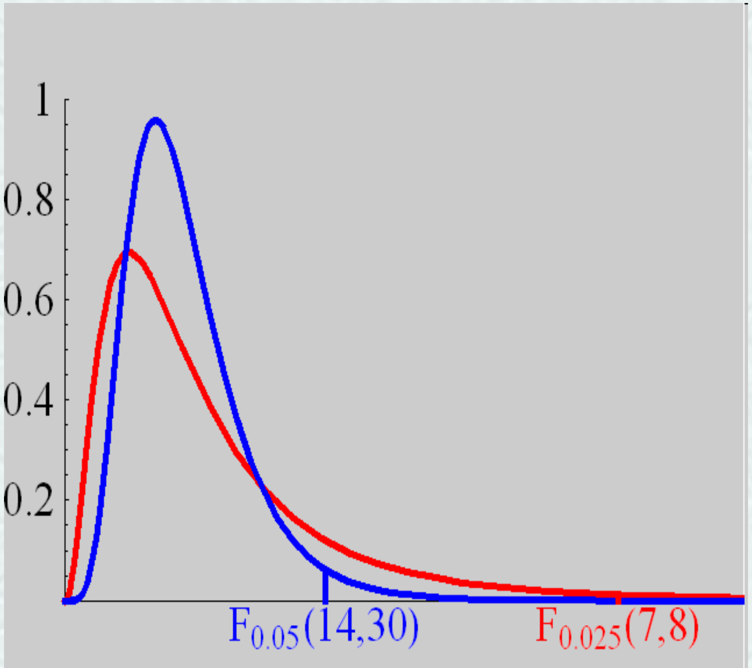
$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1, n_2)}^{+\infty} \psi(y)\mathrm{d}y = \alpha,$$

求 $F_\alpha(n_1, n_2)$的值, 可通过查表完成.

$F_{0.025}(7,8) = 4.90,$ 附表5-1

$F_{0.05}(14,30) = 2.31.$ 附表5-2

在Matlab中求解



F 分布的上 α 分位点具有如下性质：

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

证明　因为$F \sim F(n_1, n_2)$,

所以 $1-\alpha = P\{F > F_{1-\alpha}(n_1, n_2)\}$

$$= P\left\{\frac{1}{F} < \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = 1 - P\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\}$$

$$= 1 - P\left\{\frac{1}{F} > \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\},$$

故 $P\left\{\frac{1}{F} > \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = \alpha,$

因为 $\dfrac{1}{F} \sim F(n_2, n_1)$, 所以 $P\left\{\dfrac{1}{F} > F_\alpha(n_2, n_1)\right\} = \alpha,$

比较后得 $\dfrac{1}{F_{1-\alpha}(n_1, n_2)} = F_\alpha(n_2, n_1),$

即 $F_{1-\alpha}(n_1, n_2) = \dfrac{1}{F_\alpha(n_2, n_1)}.$

用来求分布表中未列出的一些上α分位点.

例 $F_{0.95}(12,9) = \dfrac{1}{F_{0.05}(9, 12)} = \dfrac{1}{0.28} = 0.357.$

4. 正态总体的样本均值与样本方差的分布

定理一

设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}$ 是样本均值, 则有 $\overline{X} \sim N(\mu, \sigma^2/n)$.

正态总体 $N(\mu,\sigma^2)$ 的样本均值和样本方差有以下两个重要定理.

定理二

设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}, S^2$ 分别是样本均值和样本方差,则有

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

(2) $\overline{X}$ 与 S^2 独立.

定理三 设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}, S^2$ 分别是样本均值和样本方差, 则有

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

证明 因为 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$,

且两者独立, 由 t 分布的定义知

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$$

定理四 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是具有相同方差的两正态总体 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ 的样本, 且这两个样本互相独立, 设 $\overline{X} = \dfrac{1}{n_1}\sum_{i=1}^{n_1} X_i$,

$\overline{Y} = \dfrac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的均值,

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

分别是这两个样本的方差, 则有

(1) $$\frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1);$$

(2) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

其中 $S_w^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$, $S_w = \sqrt{S_w^2}$.

证明 (1) 由定理二

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1),$$

由假设 S_1^2, S_2^2 独立, 则由 F 分布的定义知

$$\frac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2} \Bigg/ \frac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2} \sim F(n_1-1, n_2-1),$$

即 $$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

(2) 因为 $\overline{X}-\overline{Y} \sim N\left(\mu_1-\mu_2, \dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}\right)$

所以 $U=\dfrac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim N(0,1)$,

由 $\dfrac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1)$, $\quad \dfrac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1)$,

且它们相互独立, 故由 χ^2 分布的可加性知

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2),$$

由于 U 与 V 相互独立, 按 t 分布的定义.

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}$$

$$=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2).$$

三、小结

两个最重要的统计量:

样本均值 $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$

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

三个来自正态分布的抽样分布:

χ^2 分布, t 分布, F 分布.