第六章 数理统计的基本概念

关键词：

- 总 体，个 体，样 本， 统 计 量
- χ^2-分布 $t-$分布 $F-$分布

- 数理统计学 是一门以数据为基础的科学, 可以定义为收集数据, 分析数据和由数据得出结论的一组概念、原则和方法。

- 例如：若规定灯泡寿命低于1000小时者为次品，如何确定次品率？由于灯泡寿命试验是破坏性试验，不可能把整批灯泡逐一检测，只能抽取一部分灯泡作为样本进行检验，以样本的信息来推断总体的信息，这是数理统计学研究的问题之一。

§6.1随机样本与统计量

- 总体：研究对象的全体；
- 个体：总体中的成员；
- 总体的容量：总体中包含的个体数；
- 有限总体：容量有限的总体；
- 无限总体：容量无限的总体，通常将容量非常大的总体也按无限总体处理。

- **例：** 现要研究某一个公司员工工资水平及其影响工资水平的因素. 这个公司的每个员工就是"个体", 而所有的员工构成一个"总体". 由于公司的员工总数是有限的, 因此, 是一个有限总体. 每个员工都附着有年龄, 性别, 工种, 工资, 受教育程度等指标(变量).

- 总体的某个指标X, 对于不同的个体来说有不同的取值, 这些取值可以构成一个分布, 因此X可以看成一个**随机变量**. 有时候就把X称为总体. 假设X的分布函数为**F(x),** 也称F(x)为总体.

- 数理统计主要任务是从总体中抽取一部分个体，根据这部分个体的数据对总体分布给出推断。被抽取的部分个体叫做总体的一个 样本.

- **随机样本：** 从总体中随机地取n个个体, 称为一个随机样本。
- **简单随机样本：** 满足以下两个条件的随机样本$(X_1, X_2, \cdots, X_n)$称为容量是n的简单随机样本。

1. 每个X_i与X同分布；
2. $X_1, X_2, \cdots, X_n$是相互独立的随机变量。

[说明]：后面提到的样本均指简单随机样本。

由概率论知，若总体X具有概率密度f(x)，则样本（$X_1, X_2, \cdots, X_n$）具有联合密度函数：

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

[注意]：一个容量为n的样本 $X_1, X_2, \cdots X_n$

是指n个独立与总体分布相同的随机变量。

一旦对样本进行观察，得到实际数值

$$x_1, x_2, \cdots x_n$$

称为样本观察值（或样本值）。

两次观察，样本值可能是不同的。

- 如何取得的样本才称是简单随机样本?

对于有限总体, 采用放回抽样就能得到简单随机样本.

但当总体容量很大的时候,放回抽样有时候很不方便, 因此在实际中当总体容量比较大时,通常将不放回抽样所得到的样本近似当作简单随机样本来处理.

对于无限总体, 一般采取不放回抽样.

- **统计量：** 样本的不含任何未知参数的函数。
- **常用统计量：** 设（$X_1, X_2, \cdots, X_n$）为取自总体X的样本。常用的统计量如下：

1. 样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

2. 样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$，$S$为样本标准差

3. 样本矩　　k阶矩：$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \quad (k=1,2,\cdots)$

　　k阶中心矩：$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k \quad (k=1,2,\cdots)$

[思考题]：

对于总体X, $X_1,\ldots,X_n$是来自总体的样本，设下列数字特征存在，

$E(X)=\mu, D(X)=\sigma^2$，$E(X^k)$，$E[(X-\mu)^k]$，

问：（1）$\bar{X}$与μ，（2）S^2与σ^2，

（3）A_k与$E(X^k)$，（4）B_k与$E[(X-\mu)^k]$

是一回事吗？

答：不是。前者是随机变量，观察两次得到的统计值可能不一样；后者是数，可能已知也可能未知。

当总体数字特征未知时

- 一般, 用样本均值 $\overline{X}$ 作为总体均值 $\mu = E(X)$ 的估计;
- 用样本方差 S^2 作为总体方差 $\sigma^2 = E(X-\mu)^2$ 的估计;
- 用样本的原点矩 A_k 作为总体原点矩 $\mu_k = E(X^k)$ 的估计;
- 用样本的中心矩 B_k 作为总体中心矩 $\nu_k = E(X-\mu)^k$ 的估计.

（假设总体各阶矩存在）

■ 总体方差的估计可以用S^2也可以B_2， 主要的区别是S^2作为总体方差估计是无偏估计，但B_2作为总体方差的估计是有偏的(关于估计的无偏性将在下一节讨论).

§ 6.2 χ^2-分布 $t-$分布 $F-$分布

- 统计量的分布称为抽样分布.
- 在数理统计中, 最重要的三个分布分别为:

$$\chi^2-\text{分布} \quad t-\text{分布} \quad F-\text{分布}$$

χ^2分布

定义：设随机变量$X_1, X_2, \cdots X_n$相互独立，

$X_i \sim N(0,1)\ \ (i=1,2,\cdots,n)$

则称 $\chi_n^2 = \sum_{i=1}^{n} X_i^2 \qquad (1)$

服从自由度为n的χ^2分布，记为$\chi^2 \sim \chi^2(n)$

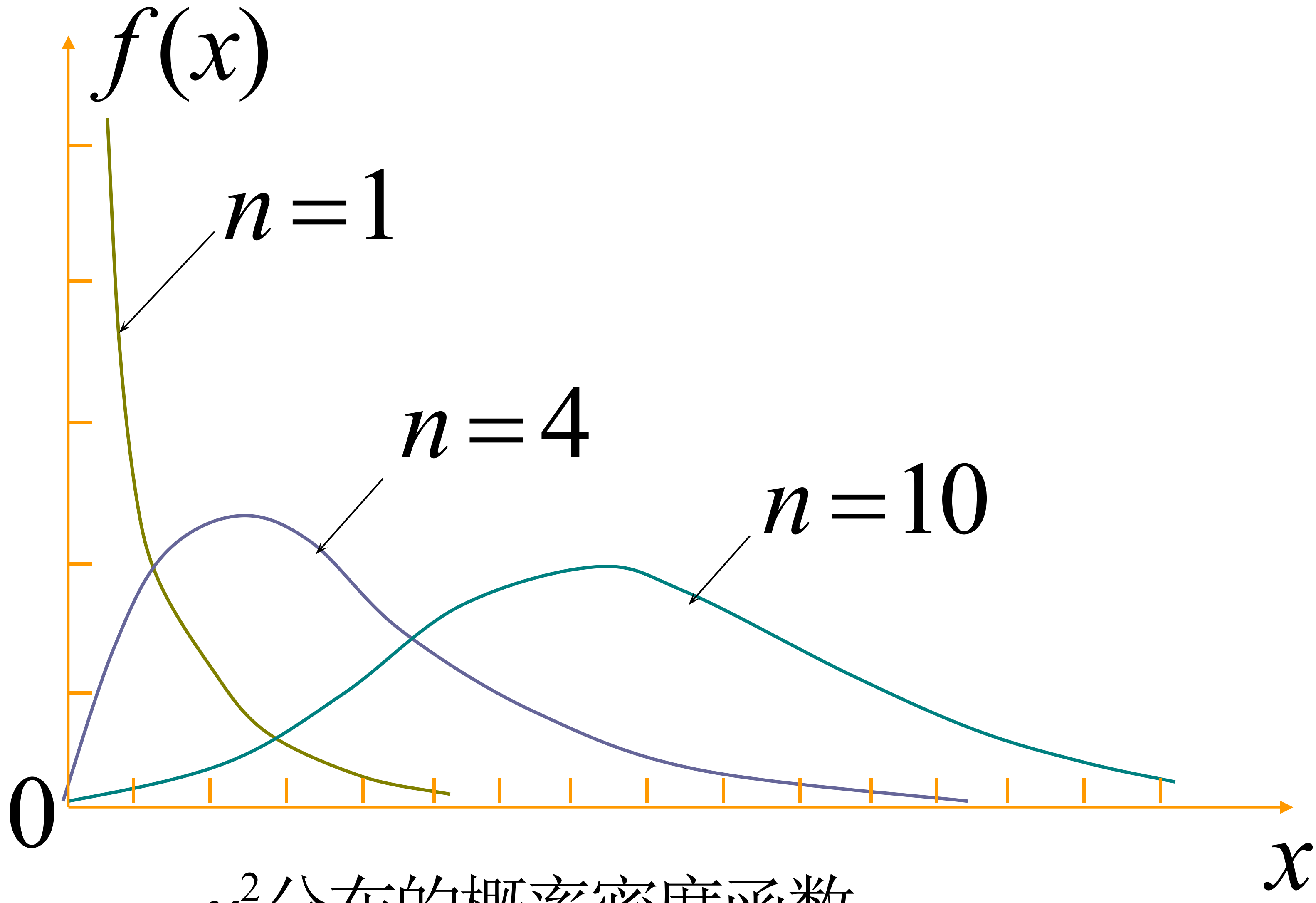
自由度指(1)式右端包含的独立变量的个数.

χ^2分布

$\chi^2(n)$分布的概率密度为:

$$f_n(y)=\begin{cases}\dfrac{1}{2\Gamma(n/2)}\left(\dfrac{y}{2}\right)^{\frac{n}{2}-1}e^{-\frac{y}{2}} & y>0\\ 0 & y\le 0\end{cases}$$

其中$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}dx$



χ^2分布的概率密度函数

χ^2分布的一些重要性质：

1. 设$\chi^2 \sim \chi^2(n)$,则有$E(\chi^2)=n, D(\chi^2)=2n$

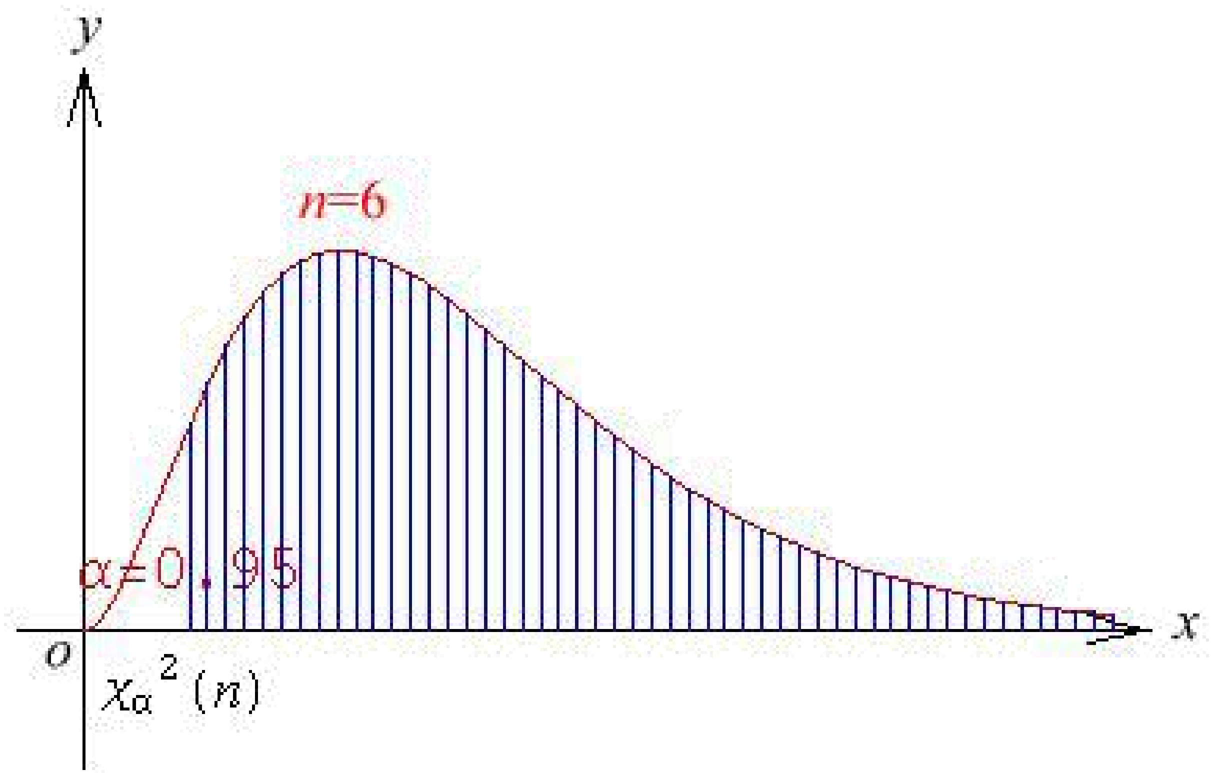
2. 设$Y_1 \sim \chi^2(n_1), Y_2 \sim \chi^2(n_2)$,且$Y_1, Y_2$相互独立，则有$Y_1+Y_2 \sim \chi^2(n_1+n_2)$

性质2称为χ^2分布的可加性，可推广到有限个的情形：

$Y_i \sim \chi^2(n_i), i=1,2,\cdots m$,并假设$Y_1, Y_2, \cdots Y_m$相互独立，

则$\sum_{i=1}^{m} Y_i \sim \chi^2\left(\sum_{i=1}^{m} n_i\right)$

对给定的概率$\alpha, 0<\alpha<1$,称满足条件$\int_{\chi_{\alpha}^{2}(n)}^{\infty} f_n(y)\,dy=\alpha$的点$\chi_{\alpha}^{2}(n)$为$\chi^{2}(n)$分布的上$\alpha$分位数,上$\alpha$分位数$\chi_{\alpha}^{2}(n)$的值可查$\chi^{2}$分布表



例：求 $\chi^2_{0.1}(25)$

- **通过Excel给出.**

（1）具体如下: 在Excel表单的任一单元格输入“=”；

（2）在主菜单中点击“插入”，点击“函数(F)”；

（3）在选择类别的下拉式菜单中选择“统计”选择“CHIINV”点击“确定”在函数参数表单中输入Probability=0.1.

（4） Deg_freedom=25, 点击"确定" 即在单元格中出现 "34.382".

例：设总体$X \sim N\left(\mu, \sigma^2\right), \mu, \sigma^2$ 已知。

$\left(X_1, X_2, \cdots, X_n\right)$是取自总体$X$的样本

求(1)统计量 $\chi^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$ 的分布；

（2）设$n = 5$, 若$a(X_1 - X_2)^2 + b(2X_3 - X_4 - X_5)^2 \sim \chi^2(k)$,

则a, b, k各为多少？

解：(1) 作变换 $Y_i = \dfrac{X_i - \mu}{\sigma} \quad i = 1, 2, \cdots, n$

显然 $Y_1, Y_2, \cdots, Y_n$ 相互独立，且 $Y_i \sim N(0,1) \quad i = 1, 2, \cdots, n$

于是 $\chi^2 = \sum_{i=1}^{n} (\dfrac{X_i - \mu}{\sigma})^2 = \sum_{i=1}^{n} Y_i^2 \sim \chi^2(n)$

(2) $X_1-X_2\sim N(0,2\sigma^2)$, $\dfrac{X_1-X_2}{\sqrt{2}\sigma}\sim N(0,1)$

$2X_3-X_4-X_5\sim N(0,6\sigma^2)$, $\dfrac{2X_3-X_4-X_5}{\sqrt{6}\sigma}\sim N(0,1)$

$\dfrac{X_1-X_2}{\sqrt{2}\sigma}$与$\dfrac{2X_3-X_4-X_5}{\sqrt{6}\sigma}$相互独立,

故$\dfrac{(X_1-X_2)^2}{2\sigma^2}+\dfrac{(2X_3-X_4-X_5)^2}{6\sigma^2}\sim\chi^2(2)$

$a=\dfrac{1}{2\sigma^2}$,

$b=\dfrac{1}{6\sigma^2}$,

$k=2.$

$t-$分布

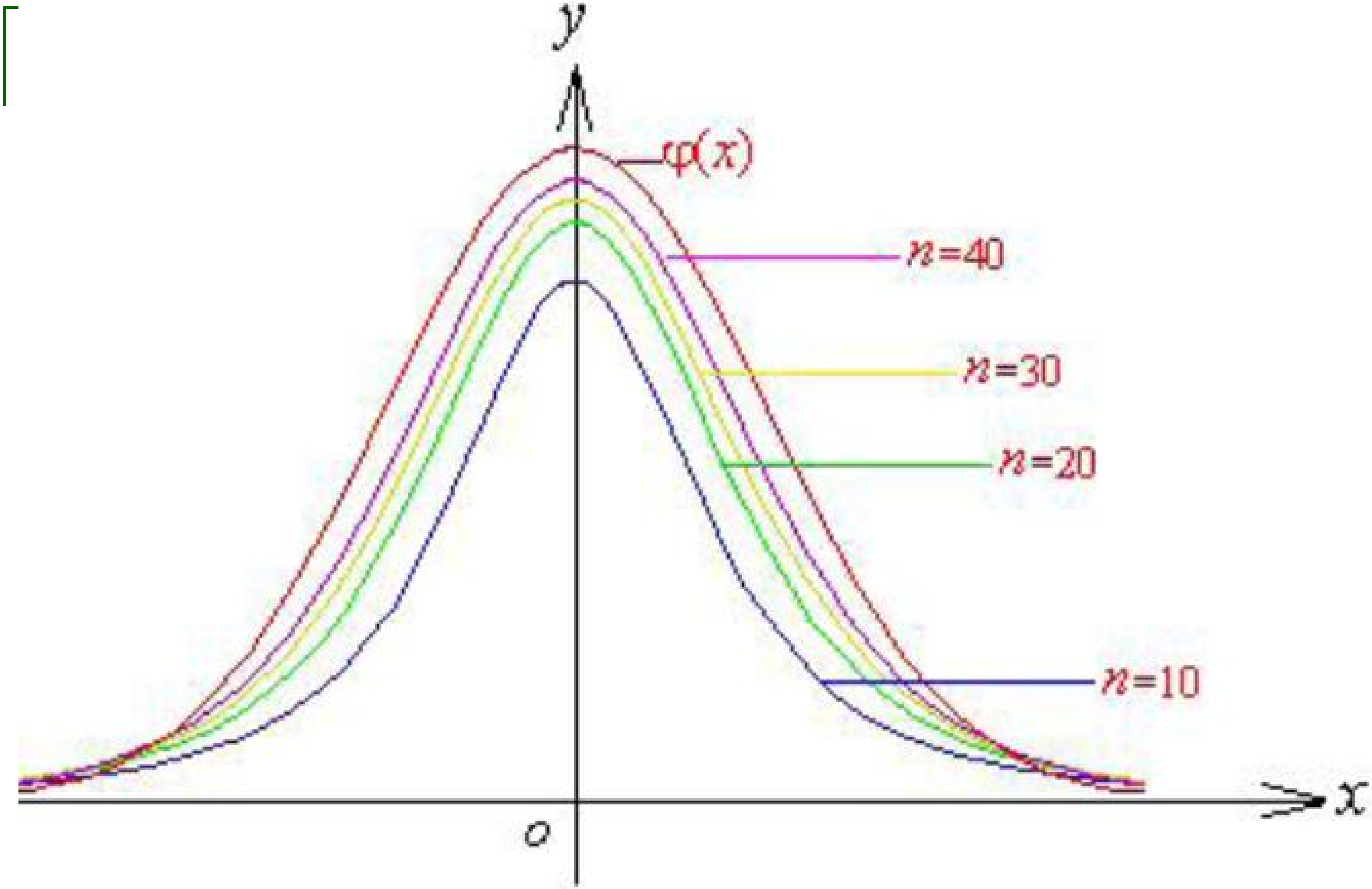
设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，并且假设 X,Y 相互独立，

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 $T \sim t(n)$

$t(n)$ 分布的概率密度为：

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

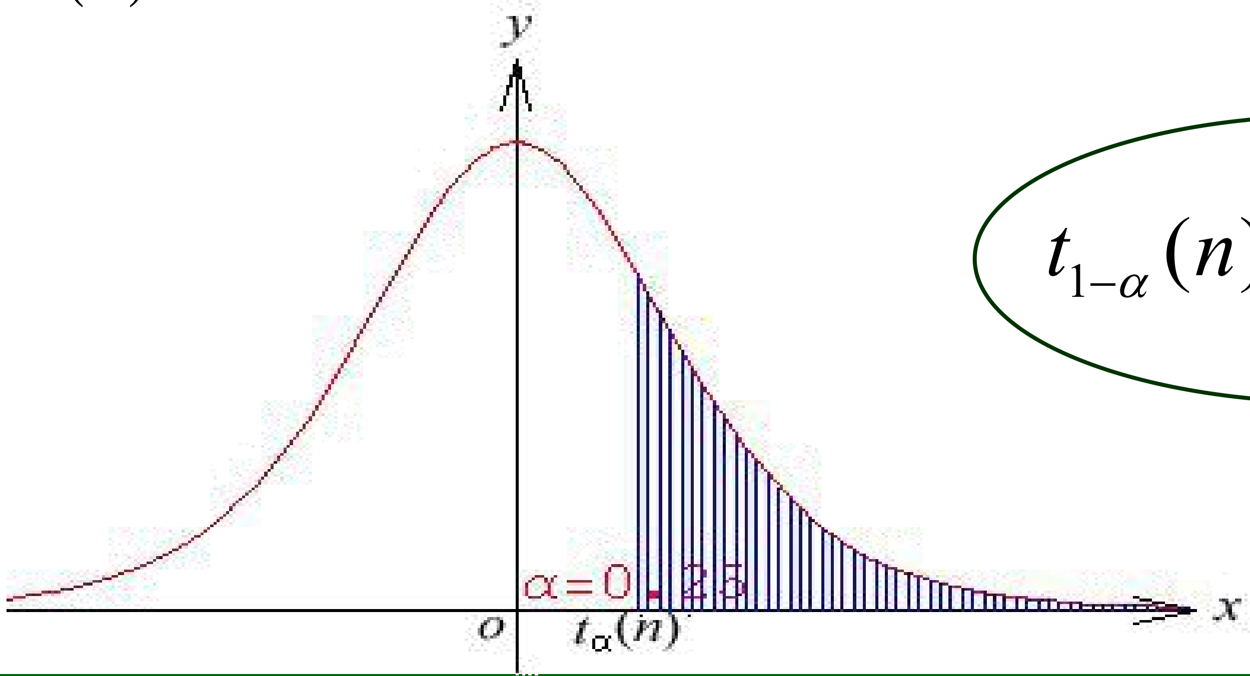


$t(n)$ 分布概率密度函数

$t-$分布

对给定的α, $0<\alpha<1$, 称满足条件$\int_{t_\alpha(n)}^{\infty} f(t,n)\,dt=\alpha$的点$t_\alpha(n)$

为$t(n)$分布的上α分位数。t分布的上α分位数可查t分布表



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

例：求 $t_{0.05}(25)$

- **通过Excel给出.**

（1）具体如下: 在Excel表单的任一单元格输入 “=”；

（2）在主菜单中点击“插入”， 点击“函数(F)”；

（3）在选择类别的下拉式菜单中选择“统计” 选择 “TINV” 点击“确定“在函数参数表单中输入 Probability=0.95.

（4） Deg_freedom=25, 点击“确定” 即在单元格中出现 “1.708” .

F分布

定义：设$X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$，且$X, Y$独立，则

称随机变量$F = \dfrac{X/n_1}{Y/n_2}$服从自由度(n_1, n_2)的F分布，

记为$F \sim F(n_1, n_2)$

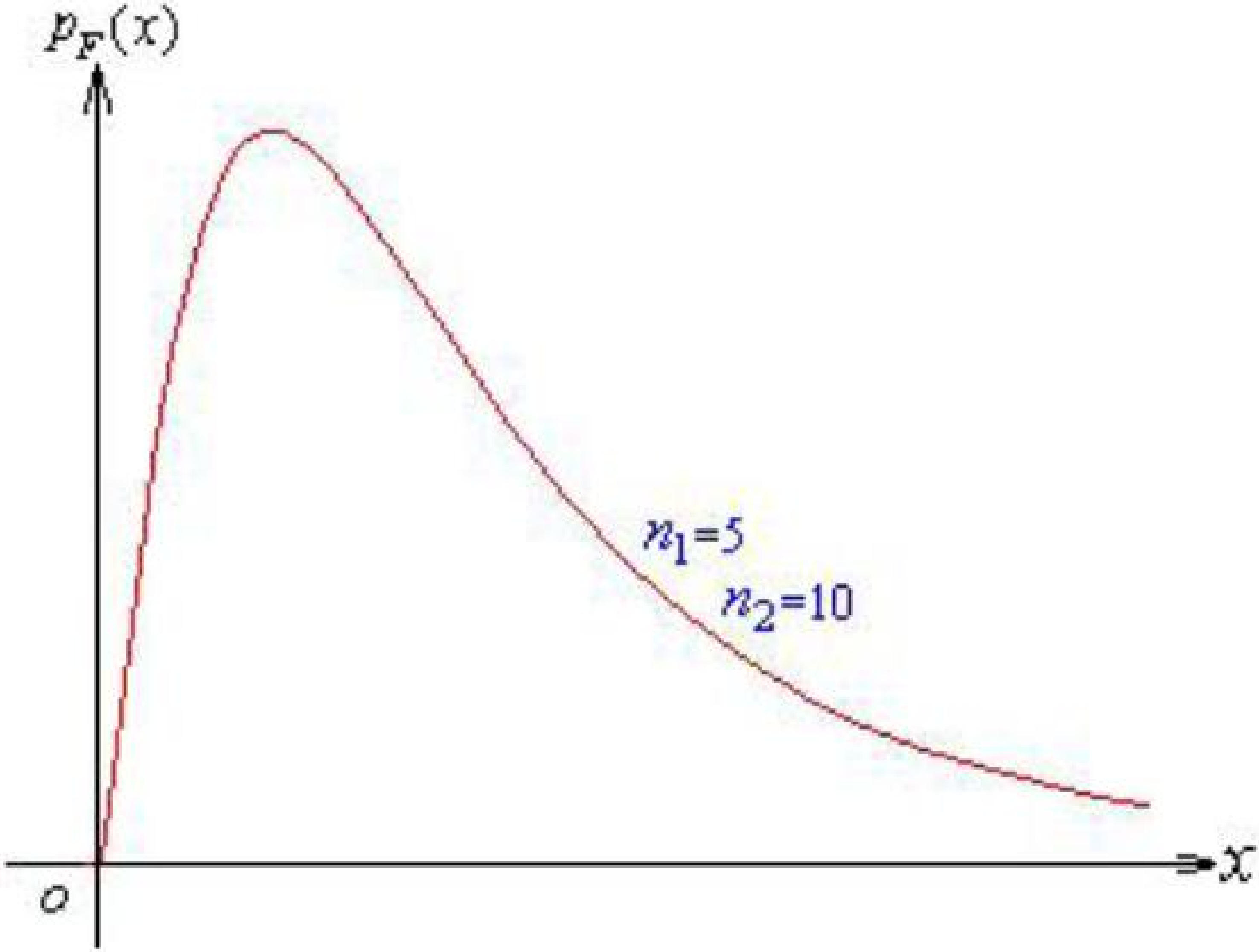
其中n_1称为第一自由度，n_2称为第二自由度

性质：$F \sim F(n_1, n_2)$，则$F^{-1} \sim F(n_2, n_1)$

$F\left(n_1,n_2\right)$分布的概率密度为：

$$f\left(x;n_1,n_2\right)=\begin{cases}\dfrac{1}{B\left(n_1/2,n_2/2\right)}n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}x^{\frac{n_1}{2}-1}\left(n_2+n_1x\right)^{-\frac{n_1+n_2}{2}} & x>0\\ 0 & x\le 0\end{cases}$$

其中$B\left(a,b\right)=\int_0^1 x^{\alpha-1}\left(1-x\right)^{b-1}dx=\dfrac{\Gamma\left(a\right)\Gamma\left(b\right)}{\Gamma\left(a+b\right)}$



F（5,10）分 布 概 率 密 度 函 数

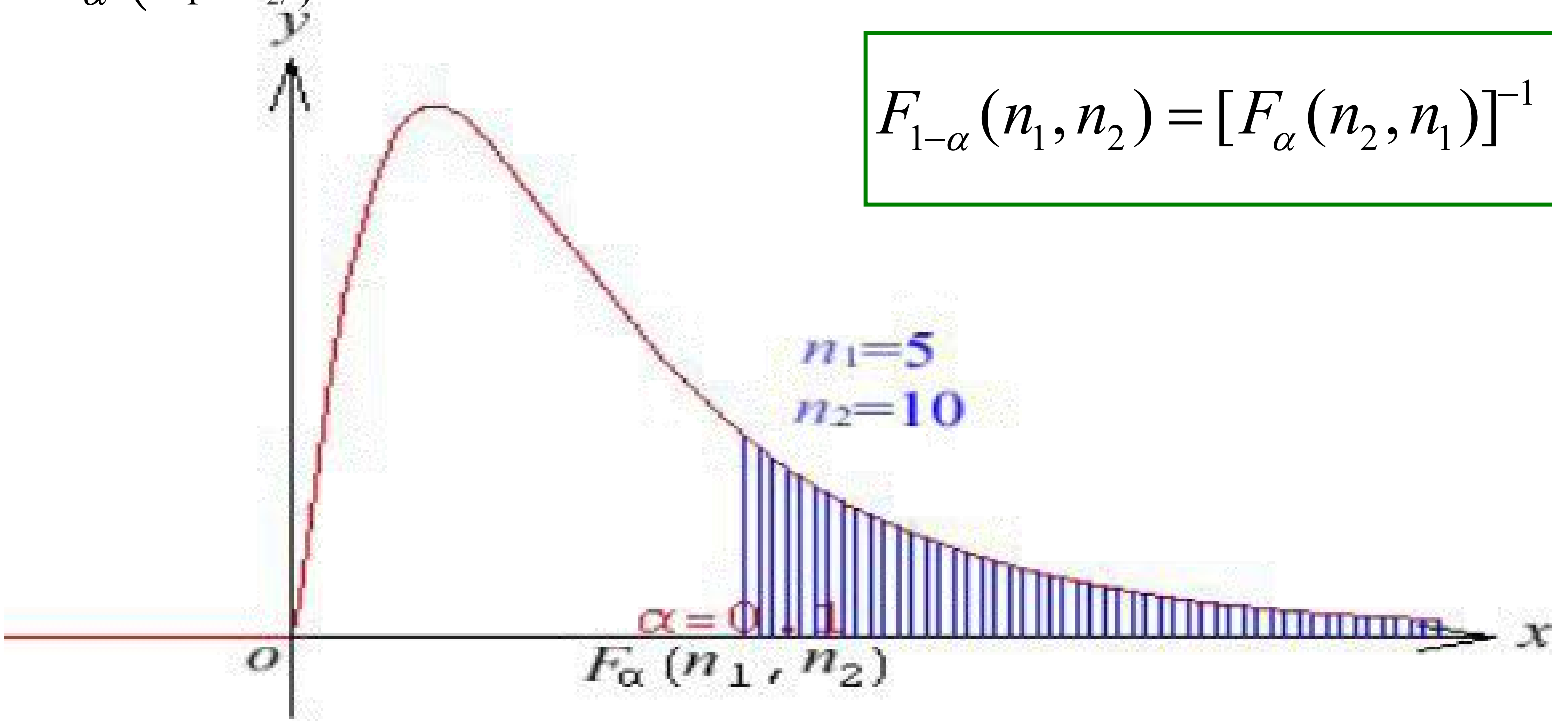
对于给定的α, $0<\alpha<1$,称满足条件

$$\int_{F_\alpha(n_1,n_2)}^{\infty} f(x;n_1,n_2)\,dx=\alpha$$

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

$F_\alpha(n_1,n_2)$的值可查F分布表

$$\boxed{F_{1-\alpha}(n_1,n_2)=[F_\alpha(n_2,n_1)]^{-1}}$$



例：求 $F_{0.1}(9,10)$

- 通过Excel给出.

（1）具体如下: 在Excel表单的任一单元格输入“=”；

（2）在主菜单中点击“插入”，点击“函数(F)”；

（3）在选择类别的下拉式菜单中选择“统计” 选择“FINV” 点击“确定”在函数参数表单中输入 Probability=0.9.

（4） Deg_freedom1=9, Deg_freedom2=10点击“确定” 即在单元格中出现 “2.347”.

§6.3正态总体下的抽样分布

定理 6.3.1 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体

$N(\mu, \sigma^2)$ 的简单随机样本，$\overline{X}$ 是样本均值，

S^2 是样本方差，则有：

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

定理 6.3.2 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu,\sigma^2)$ 的简单随机样本，$\overline{X}$ 是样本均值，S^2 是样本方差，则有：

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

(2) $\overline{X}$ 与 S^2 相互独立.

[思考题]：

设$X_1,X_2,\cdots,X_n$是来自正态总体$N(\mu,\sigma^2)$的简单随机样本，$\bar{X}$和S^2分别是样本均值和样本方差。问：

(1) $\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$服从什么分布？

(2) $\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$服从什么分布？

答：（1）$\chi^2(n-1)$，（2）$\chi^2(n)$.

定理 6.3.3 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的简单随机样本，$\bar{X}$ 是样本均值，S^2 是样本方差，则有:

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

定理6. 3. 4

设样本$\left(X_1,\cdots,X_{n_1}\right)$和$\left(Y_1,\cdots,Y_{n_2}\right)$分别来自总体

$N\left(\mu_1,\sigma_1{}^2\right)$和$N\left(\mu_2,\sigma_2{}^2\right)$ 并且它们相互独立，

其样本方差分别为S_1^2,S_2^2,

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

(2) $$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

其中$\overline{X},\overline{Y}$分别为样本均值.

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

例：设总体X的均值μ，方差σ^2存在，

$\left(X_1,\cdots,X_n\right)$是取自总体$X$的样本，

$\bar{X},S^2$为样本均值和样本方差；

求（1）$E(\bar{X}),D(\bar{X}),E(S^2)$;

(2)X_1与$\bar{X}$的相关系数；

(3)若$X\sim N\left(\mu,\sigma^2\right)$，求$D(S^2)$）.

解：（1）$E(\bar{X})=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu,$

$$D(\bar{X})=D(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{\sigma^2}{n},$$

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

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

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

$$(2)\ Cov(X_1, \bar{X}) = Cov(X_1, \frac{1}{n}\sum_{i=1}^{n} X_i)$$

$$= \frac{1}{n} D(X_1) = \frac{\sigma^2}{n},$$

$$\rho_{X_1\bar{X}} = \frac{Cov(X_1, \bar{X})}{\sqrt{D(X_1)D(\bar{X})}}$$

$$= \frac{\sigma^2}{n} \Big/ \sqrt{\sigma^2 \times \frac{\sigma^2}{n}} = \frac{1}{\sqrt{n}},$$

$$(3) X \sim N\left(\mu, \sigma^2\right), \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

$$\Rightarrow D\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$$

$$\Rightarrow D(S^2) = \frac{2\sigma^4}{n-1}.$$

例：设总体$X \sim N\left(\mu, \sigma^2\right), \left(X_1, \cdots, X_4\right)$

与$\left(Y_1, \cdots, Y_9\right)$是取自总体$X$的两个独立样本，

$\overline{X}, S_1^2$和$\overline{Y}, S_2^2$分别为样本均值和样本方差；

求（1）若$a\dfrac{\overline{X}-\overline{Y}}{S_1} \sim t(k)$，则$a, k$各为多少？

(2) $\sum_{i=1}^{4}(X_i-\mu)^2 \Big/ 4S_2^2$服从什么分布？

(1) $\because \bar{X} \sim N(\mu, \frac{\sigma^2}{4}), \bar{Y} \sim N(\mu, \frac{\sigma^2}{9})$，且$\bar{X}$与$\bar{Y}$相互独立，

$$\therefore \bar{X} - \bar{Y} \sim N(0, \frac{13\sigma^2}{36}), \quad \frac{6(\bar{X} - \bar{Y})}{\sqrt{13}\sigma} \sim N(0,1)$$

又$\frac{3S_1^2}{\sigma^2} \sim \chi^2(3)$，且$\bar{X} - \bar{Y}$与$S_1^2$相互独立，

由t分布定义，$\frac{6}{\sqrt{13}} \frac{\bar{X} - \bar{Y}}{\sigma} \Big/ \sqrt{\frac{3S_1^2}{3\sigma^2}} = \frac{6\sqrt{13}}{13} \frac{\bar{X} - \bar{Y}}{S_1} \sim t(3)$

$$\Rightarrow a = \frac{6\sqrt{13}}{13}, k = 3.$$

$$(2)\ \frac{1}{\sigma^2}\sum_{i=1}^{4}(X_i-\mu)^2 \sim \chi^2(4),\quad \frac{8S_2^2}{\sigma^2} \sim \chi^2(8),$$

且$\sum_{i=1}^{4}(X_i-\mu)^2$与$S_2^2$独立，

由F分布定义知，

$$\frac{1}{4\sigma^2}\sum_{i=1}^{4}(X_i-\mu)^2 \Bigg/ \frac{8S_2^2}{8\sigma^2} = \sum_{i=1}^{4}(X_i-\mu)^2 \Bigg/ 4S_2^2 \sim F(4,8).$$