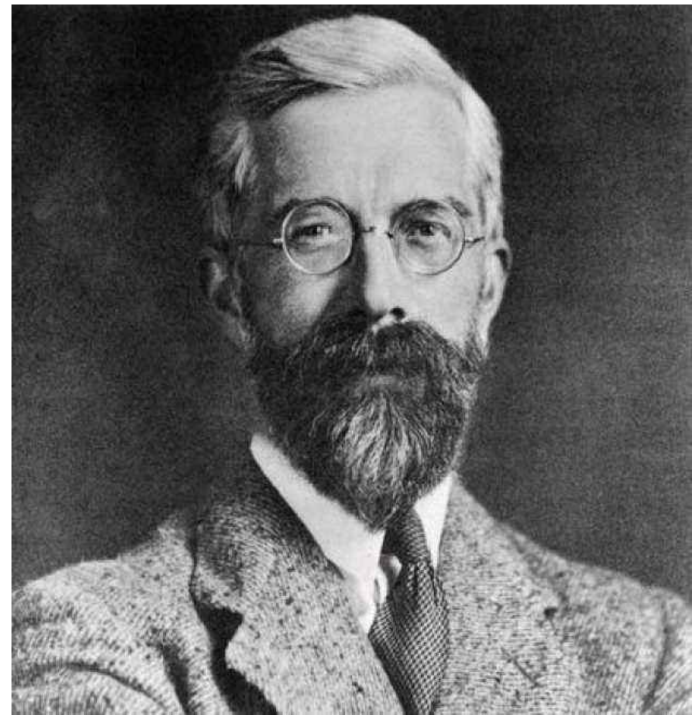
第六章 样本与统计量

- 现代统计学产生于19世纪末，20世纪初。奠基人为英国的Pearson与Fisher。尤其Fisher的贡献很大，被称为“统计之父”。到二战期间，统计的理论基础完成。二战以后，随着计算机的发展，统计进入快速发展阶段，其应用领域几乎扩展到所有学科。
- 数理统计：收集，分析带有随机影响的数据的学科。
- 注：统计的思想方法与数学不同，数学是**演绎**的思想；统计是**归纳**的思想。

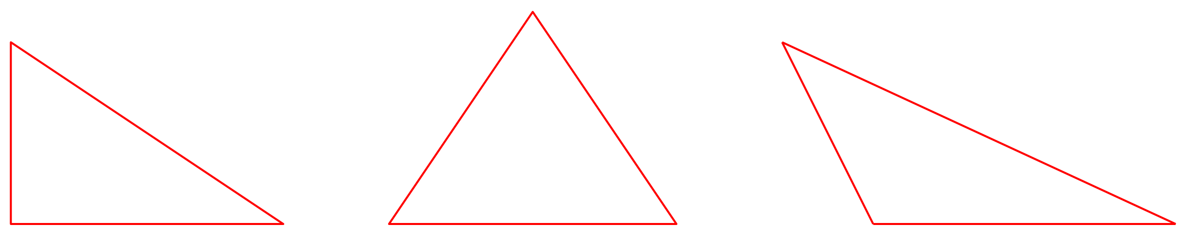
K.Pearson(1857-1936)



R.A.Fisher(1890-1972)

统计与数学思想的区别

- **命题**：三角形内角之和为180度。
- **数学家**：严格推导。
- **统计学家**：
- 1.取样；



- 2.得到数据；
- 3.由假设检验理论得出结果。

§1. 总体与样本

- **总体**：研究对象的全体。
- **个体**：总体的每个成员。
- **样本**：从总体中随机抽取的一些个体。

- **总体的数学表示**：用随机变量X来表示总体。
- **样本的数学表示**：用随机向量(X_1，X_2，…，X_n)来表示。

- **样本的二重性**：
- 1.(X_1，X_2，…，X_n)在一次具体的观测中，是一些已知的数。
- 2. 在不同的观测中,样本取值可能不同,应将其看做随机变量。

- **简单随机样本**：.(X_1，X_2，…，X_n)相互独立，且具有与总体相同的分布。

- 统计的任务：用样本推断总体的未知特征。

§ 2. 统计量与抽样分布

- **定义**：设X_1，X_2，…，X_n是抽自总体X的一个样本，$T(X_1, X_2, \ldots, X_n)$是X_1，X_2，…，X_n的不含未知参数的函数，则称$T=T(X_1, X_2, \ldots, X_n)$为一个**统计量**。

- 统计量是样本的函数，同样具有二重性。统计量的分布称为**抽样分布**.

几个常用的统计量

- 1.样本均值

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

- 2.样本方差

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

- 3.样本*k*阶原点矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, \cdots$$

- 4.样本*k*阶中心矩

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

- 总体（X, Y)，样本（X_1, Y_1)，…，（X_n, Y_n),记

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

$$S_X^{*2}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2,\qquad S_Y^{*2}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

- 5.样本协方差: $S_{XY}=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$

- 6.样本相关系数: $\rho=\dfrac{S_{XY}}{S_X^{*}\cdot S_Y^{*}}$

- **定理**：　设总体X的均值为μ,方差为 σ^2，$X_1,\ldots,X_n$为取自X的一个样本，则近似地有

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- 大样本理论：求统计量的极限分布。
- 小样本理论：求统计量的精确分布。

§3 正态总体

- **1.χ^2分布**
- 定义： 设X_i~N(0, 1)，i =1,2,…,n且相互独立，则称随机变量

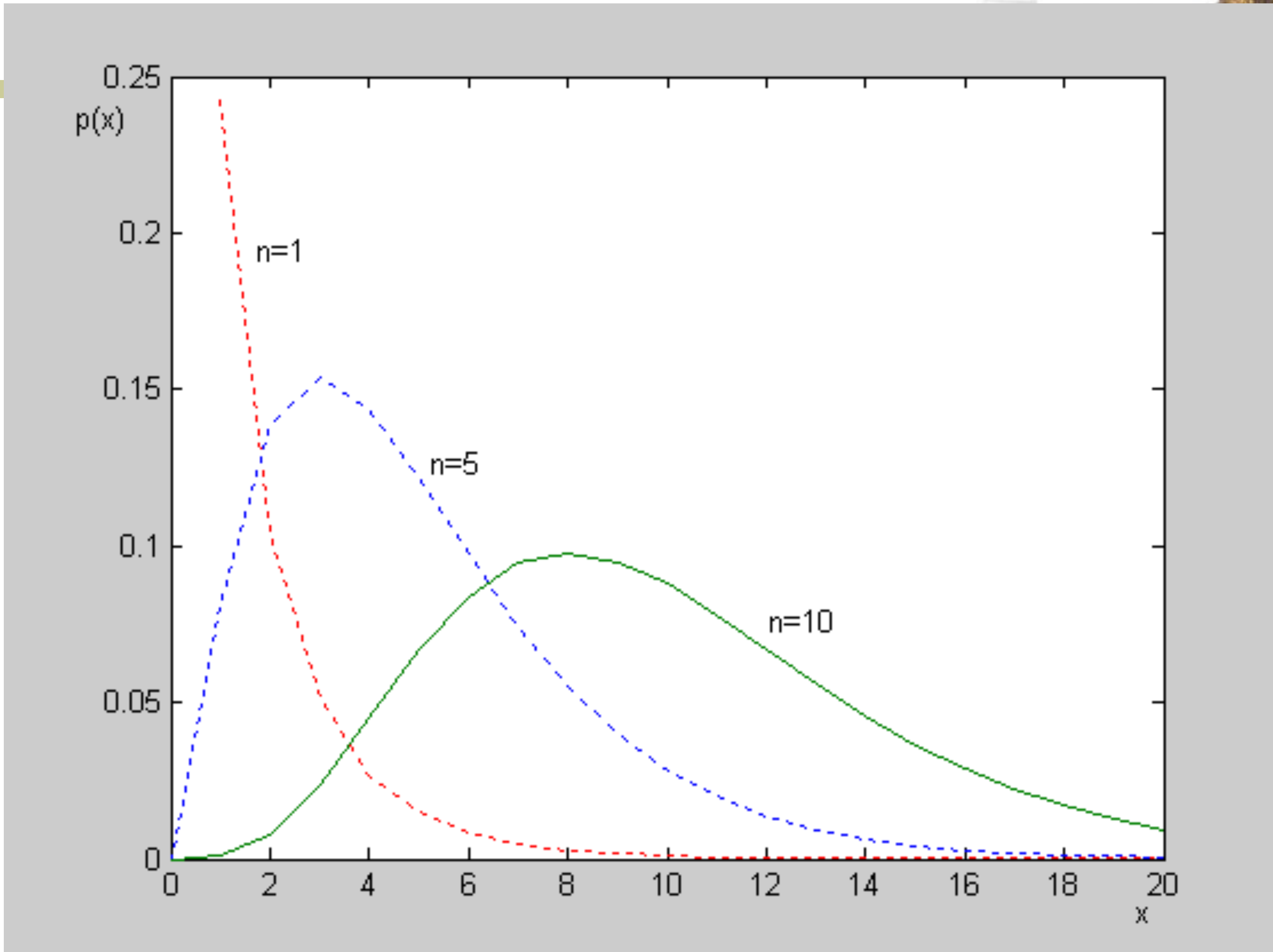
$$\chi^2=X_1^2+X_2^2+\dots+X_n^2$$

服从自由度（或参数）为n的χ^2分布，记作$\chi^2 \sim \chi^2(n)$。其密度函数为：

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

- (1) χ^2分布具有可加性。
  设$\chi_1^2 \sim \chi^2(n_1)$，$\chi_2^2 \sim \chi^2(n_2)$且相互独立，则
  $$\chi_1^2+\chi_2^2 \sim \chi^2(n_1+n_2)$$
- (2) χ^2分布的均值和方差。
  设$\chi^2 \sim \chi^2(n)$，则
  $$E(\chi^2)=n \quad D(\chi^2)=2n$$

χ^2分布的密度函数图：



n越大，图形越扁平，对称性越强

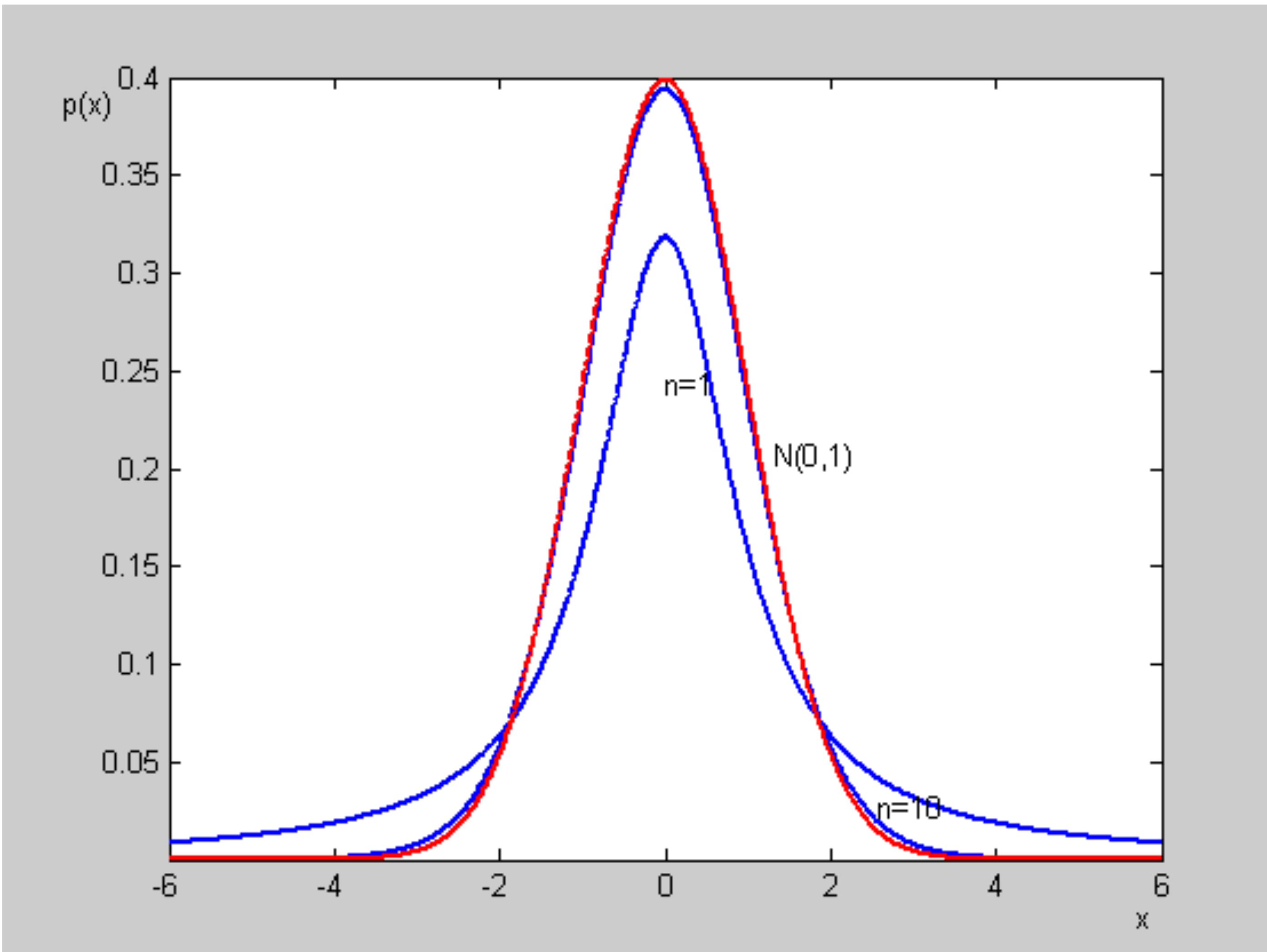
2. t分布

- 定义：设X~N(0,1),Y~χ^2(n), 且X与Y 独立，则称随机变量

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

服从自由度为n的t分布，记为t~t(n)。t分布又称学生（student）分布，是英国统计学家W.S.Gosset于1908年发表的，t分布开创了统计的小样本的理论。

其密度函数图形如下：



- t分布密度函数为：

$$p(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\ x\in R$$

- t分布的性质：
- (1). p(x)是偶函数，E(t)=0。
- (2). t分布近似N(0,1)。

- **3. F分布**
- **定义：** 设U~$\chi^2(n_1)$， V~$\chi^2(n_2)$，且U与V相互独立，则称随机变量

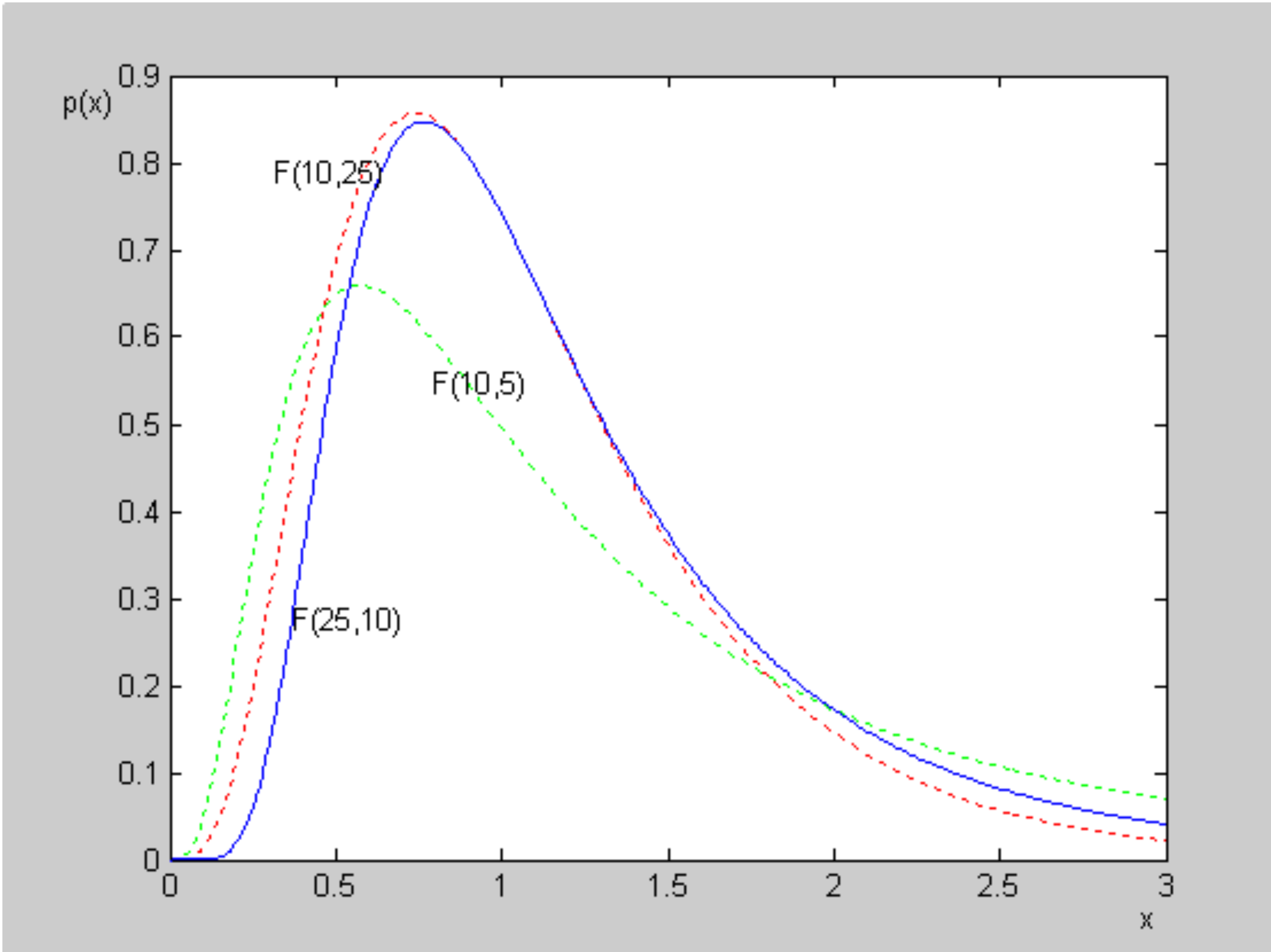
$$F=\frac{U/n_1}{V/n_2}$$

服从**自由度为n_1，n_2的F分布**，记作F~F (n_1, n_2)。

F分布密度函数为:

$$p(x)=\frac{\Gamma(\frac{n_1+n_2}{2})\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}x^{\frac{n_1}{2}-1}(1+\frac{n_1}{n_2}x)^{\frac{n_1+n_2}{2}}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma(\frac{n_2}{2})},\quad x>0$$

其密度函数图形如下：



F分布的性质

- (1)若F~F (n_1，n_2)，则

  1/ F~F (n_2，n_1)

- (2)t分布与F分布有如下关系，若T~t (n)，则

  T^2~ F (1，n)

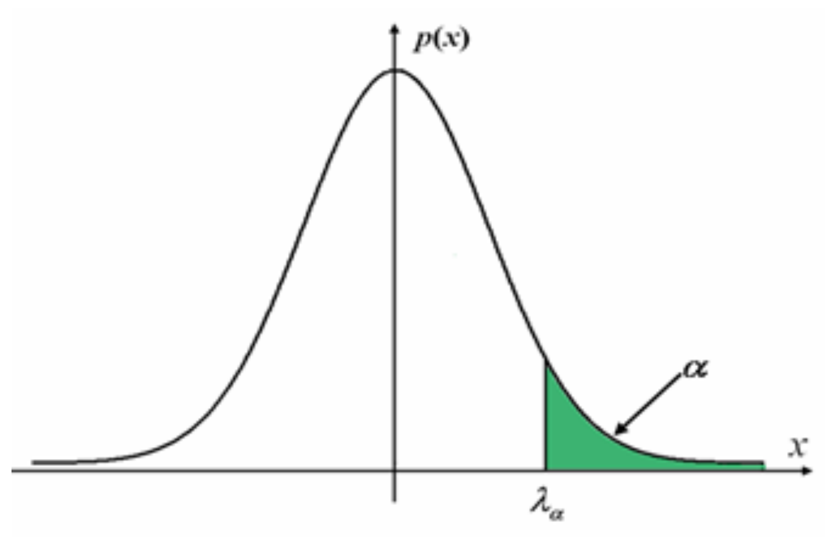
$$F_{1-\alpha}(n_1,n_2)=\frac{1}{F_{\alpha}(n_2,n_1)}$$

四.上α分位点

- 定义：设X是一个随机变量，对0<α<1，称满足条件

$$P(X > \lambda_\alpha) = \alpha$$

的实数λ_α为X的上α分位点。



- 注：上α分位点与分布函数的关系：$F_X(\lambda_\alpha)=1-\alpha$

- **正态分布的上α分位点：** $\boldsymbol{u_\alpha}$
- $\boldsymbol{u_{0.05}=}$ **，** $\boldsymbol{u_{0.95}=}$ **，** $\boldsymbol{u_{0.025}=}$ **，**

1.645, -1.645, 1.96

- 三大分布的上α分位点：

$\chi_\alpha^2(n):\chi^2(n)$分布的上$\alpha$分位点；

$t_\alpha(n):t(n)$分布的上α分位点；

$F_\alpha(n_1,n_2):F(n_1,n_2)$分布的上$\alpha$分位点。

- $t_{0.05}(8)=$， $t_{0.05}(130)=$，

1.86， 1.645

- $\chi^2_{0.05}(8)$, $\chi^2_{0.95}(8)$, $F_{0.05}(8，9)$ ， $F_{0.95}(8，9)$

15.51 , 2.733 , 3.23 ， 1/3.39

- 例：设总体X服从正态分布N(0,4)，总体Y服从正态分布N(1,1)，而$X_1, X_2, \ldots X_5$及Y_1, Y_2, Y_3分别来自总体X,Y的简单随机样本。求常数C,使得随机变量 $C\sum_{i=1}^{5}X_i^2 / \sum_{j=1}^{3}(Y_j-1)^2$ 为F分布。

- 解： C=3/20

$$\frac{\sum_{i=1}^{5}\left(\frac{X_i}{2}\right)^2/5}{\sum_{j=1}^{3}(Y_j-1)^2/3} = \frac{3}{20}F(5,3)$$

- **定理：** 设总体X ~ N(μ, σ^2)，X_1，X_2， …，X_n为取自X的一个样本，$\bar{X}, S^2$分别为样本均值和样本方差，则
- （1） $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
- （2） $$(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$$
- （3） $\bar{X}$与S^2相互独立

- 证明：1）由第三章中关于正态分布的线性组合结论可以得到。下证2），3）。取**A**是一个如下形式的正交阵

$$\mathbf{A}=\begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ * & * & * \end{pmatrix}$$

则利用**n**元正态分布的性质得：

$$X \sim N\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 I_n\right) \Rightarrow Y = A\left(X - \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}\right)/\sigma \sim N(0, I_n)$$

- 于是Y=(Y$_1$， Y$_2$， …， Y$_n$)^{T}中的Y$_i$，i=1,2,…,n为相互独立的标准正态分布。且

$$Y_1=\frac{1}{\sigma}\left(\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}}\right)(X_1-\mu,X_2-\mu,\cdots,X_n-\mu)^T=\frac{n\bar{X}-n\mu}{\sigma\sqrt{n}}=\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sigma}$$

$$\sum_{i=1}^{n}Y_i^2=\mathbf{Y'Y}=\left(\frac{1}{\sigma}\mathbf{A(X-\mu)}\right)'\left(\frac{1}{\sigma}\mathbf{A(X-\mu)}\right)=\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$$

$$=\frac{(n-1)S^2}{\sigma^2}+\frac{n\left(\bar{X}-\mu\right)^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2}+Y_1^2$$

- 所以 $\frac{(n-1)S^2}{\sigma^2}=\sum_{i=2}^{n}Y_i^2\sim\chi^2(n-1)$,且与$\bar{X}$相互独立.

- **推论1**： $T=\dfrac{\sqrt{n}(\bar{X}-\mu)}{S}\sim t(n-1)$
- **证明：**

由定理$\bar{X}\sim N(\mu,\dfrac{\sigma^2}{n})$，所以$\dfrac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\sim N(0,1).$

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

$$T=\frac{\sqrt{n}(\bar{X}-\mu)}{S}=\frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}}\sim t(n-1)$$

- **推论2：** 设总体X~N(μ_1, σ_1^2)，X_1，X_2，…，X_{n1}为取自X的样本，总体Y~N(μ_2, σ_2^2)，Y_1，Y_2，…，Y_{n2}为取自Y的样本,两样本相互独立，

  $\overline{X}, S_1^2$ 为总体X的样本均值和样本方差，

  $\overline{Y}, S_2^2$ 为总体Y的样本均值和样本方差，

则

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

- 证明：由定理

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

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

- **推论3**： 设总体X~N(μ_1, σ^2)，X_1，X_2， …， X_{n1} 为取自X的样本， 总体Y~N(μ_2, σ^2)，Y_1， Y_2， … ，Y_{n2}为取自Y的样本,两样本相互独立，

  $\bar{X}, S_1^2$ 为总体X的样本均值和样本方差，

  $\bar{Y}, S_2^2$ 为总体Y的样本均值和样本方差，

则

$$T=\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}}\sim t(n_1+n_2-2)$$

- 证明：由定理

$$\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2,(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2),$$

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

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

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma} / [\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)\sigma^2}}]$$

$$=\sqrt{\frac{n_1 n_2(n_1+n_2-2)}{n_1+n_2}}\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}} \sim t(n_1+n_2-2)$$