第五章 数理统计的基础知识

§5.1 数理统计的基本概念

定义1 统计学中称随机变量(或向量)X为**总体**, 并把随机变量(或向量)的分布称为**总体分布**.

注: ①有时个体的特性的直接描述并非是数量指标, 但是总可以将其数量化, 如检验某学校全体学生的血型, 试验的结果有O型,A型,B型,AB型4种, 若分别以1,2,3,4依次记这4种血型, 则试验的结果就可以用数量表示了.

总体的分布一般来说是未知的, 有时即使知道其分布的类型(如正态分布, 二项分布等), 但不知这些分布中所含的参数等(如μ,σ^2,p等). 数理统计的任务就是要通过对总体X进行一系列的独立的同环境下的随机试验, 获得一系列的试验值, 用这些试验值来对X的分布或参数进行统计推断.

二, 样本与样本分布

要研究总体X, 就要对X进行一系列的独立试验, 通常进行n次试验, 这样,将n次试验组合成一个试验, 又可以试验多次, 因此, n次试验相当于根据总体X产生了和总体的分布完全一样的, 相互独立的n个随机变量$X_1,X_2,\ldots,X_n$, 称它们是总体X的**样本**; 而每n次试验的一个结果产生n个数$x_1,x_2,\ldots,x_n$称为样本观察值或者**样本值**, n被称为**样本容量**(或**样本大小**).

三, 统计推断问题简述(略)
四, 分组数据统计表和频率直方图(略)
五, 经验分布函数(略)

六, 统计量

为了由样本推断总体, 需构造一些合适的统计量, 再由这些统计量来推断未知总体. 这里, 样本的统计量即为样本的函数. 广义地讲, 统计量可以是样本的任一函数, 但由于构造统计量的目的是为推断未知总体的分布, 故在构造统计量时, 就不应包括总体的未知参数.

定义3 设$X_1,X_2,\ldots,X_n$为总体X的一个样本，称此样本的任一不含总体分布未知参数的函数为该样本的统计量.

注: 当样本$X_1,X_2,\ldots,X_n$未取一组具体样本值时, 统计量用大写字母表示; 当样本$X_1,X_2,\ldots,X_n$取一组具体的样本值$x_1,x_2,\ldots,x_n$时, 统计量用小写字母表示.

七, 样本的数字特征

以下设 $X_1, X_2, \cdots, X_n$ 为总体 X 的一个样本.

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

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

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

注: 称$Q=\sum_{i=1}^{n}(X_i-\bar{X})^2$为样本的偏差平方和.

由$Q=\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X})=\sum_{i=1}^{n}X_i^2-n\bar{X}^2$

得
$$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right) \qquad (1.9)$$

3. 样本标准差

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad (1.10)$$

4. 样本(k 阶)原点矩

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\ \ k=1,2,\cdots \quad (1.11)$$

5. 样本(k 阶)中心矩

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k,\quad k=2,3,\cdots \quad (1.12)$$

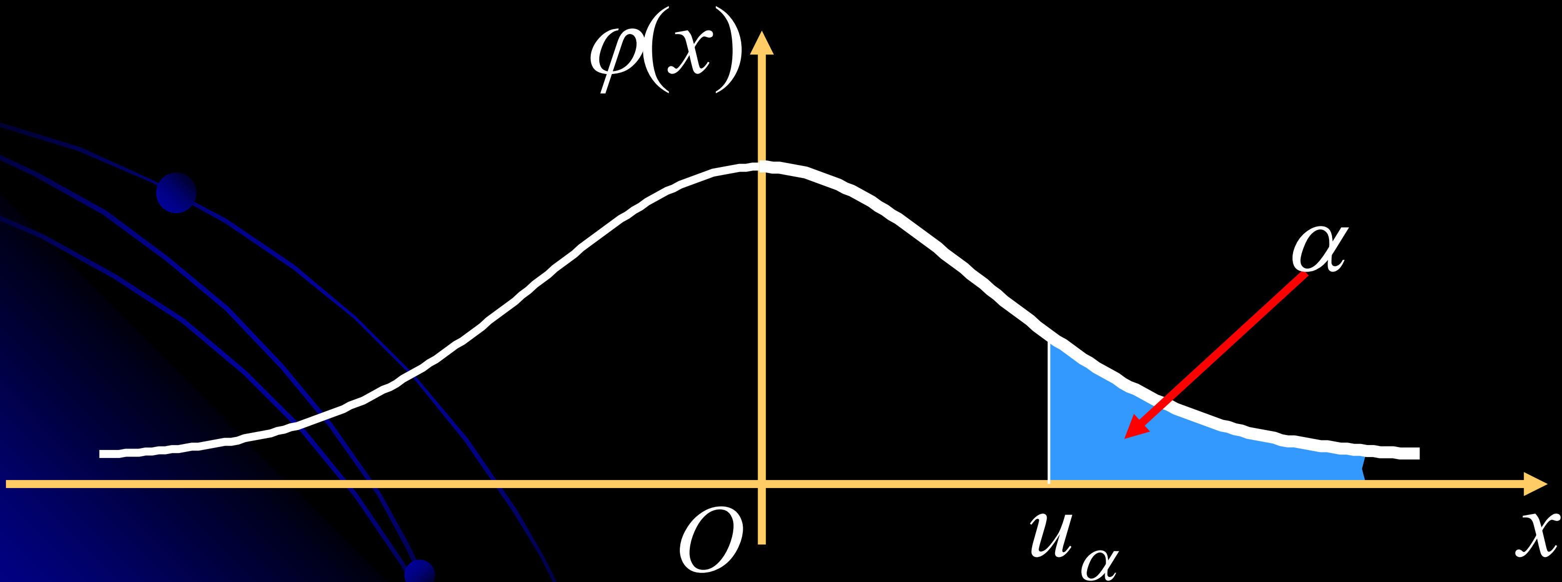
§5.2 常用统计分布

一, 分位数

设随机变量X的分布函数为$F(x)$, 对给定的实数$\alpha(0<\alpha<1)$, 若实数F_α满足不等式

$$P\{X>F_\alpha\}=\alpha,$$

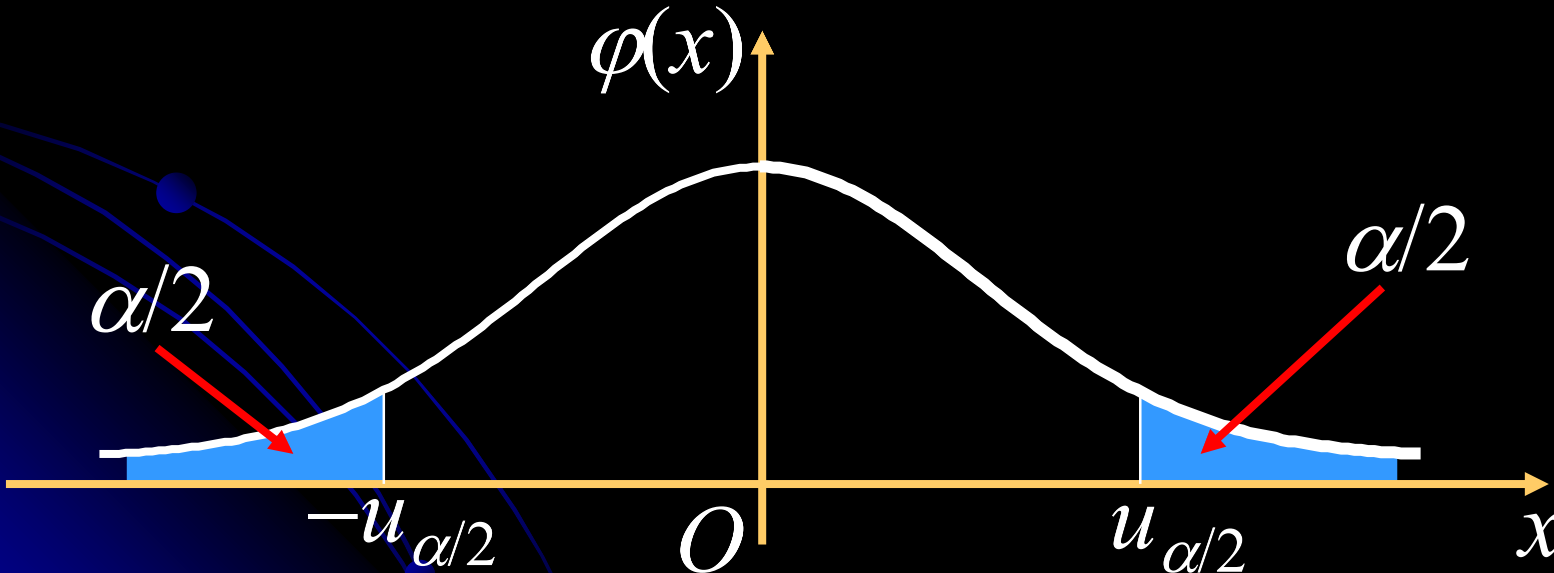
则称F_α为X分布的水平α的**上侧分位数.**



若实数T_α满足不等式

$$P\{|X|>T_\alpha\}=\alpha,$$

则称T_α为X分布的水平α的**双侧分位数.**



例1 设α=0.05, 求标准正态分布的水平0.05的上侧分位数和双侧分位数.

解: 由于$\Phi(u_{0.05})=1-0.05=0.95$,

查标准正态分布函数表可得$u_{0.05}=1.645$.

而水平0.05的双侧分位数为$u_{0.025}$, 它满足

$$\Phi(u_{0.025})=1-0.025=0.975,$$

查表得 $$u_{0.025}=1.96.$$

注: 今后, 分别记u_α与$u_{\alpha/2}$为标准正态分布的上侧分位数与双侧分位数.

二, χ^2分布

定义1 设$X_1,X_2,\ldots,X_n$是相互独立服从标准正态分布的随机变量, 则称统计量

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \qquad (2.1)$$

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

这里, 自由度是指(2.1)式右端所包含的独立变量的个数.

$\chi^2(n)$分布的概率密度:

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

其中Γ(x)为gamma函数, 定义为

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

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

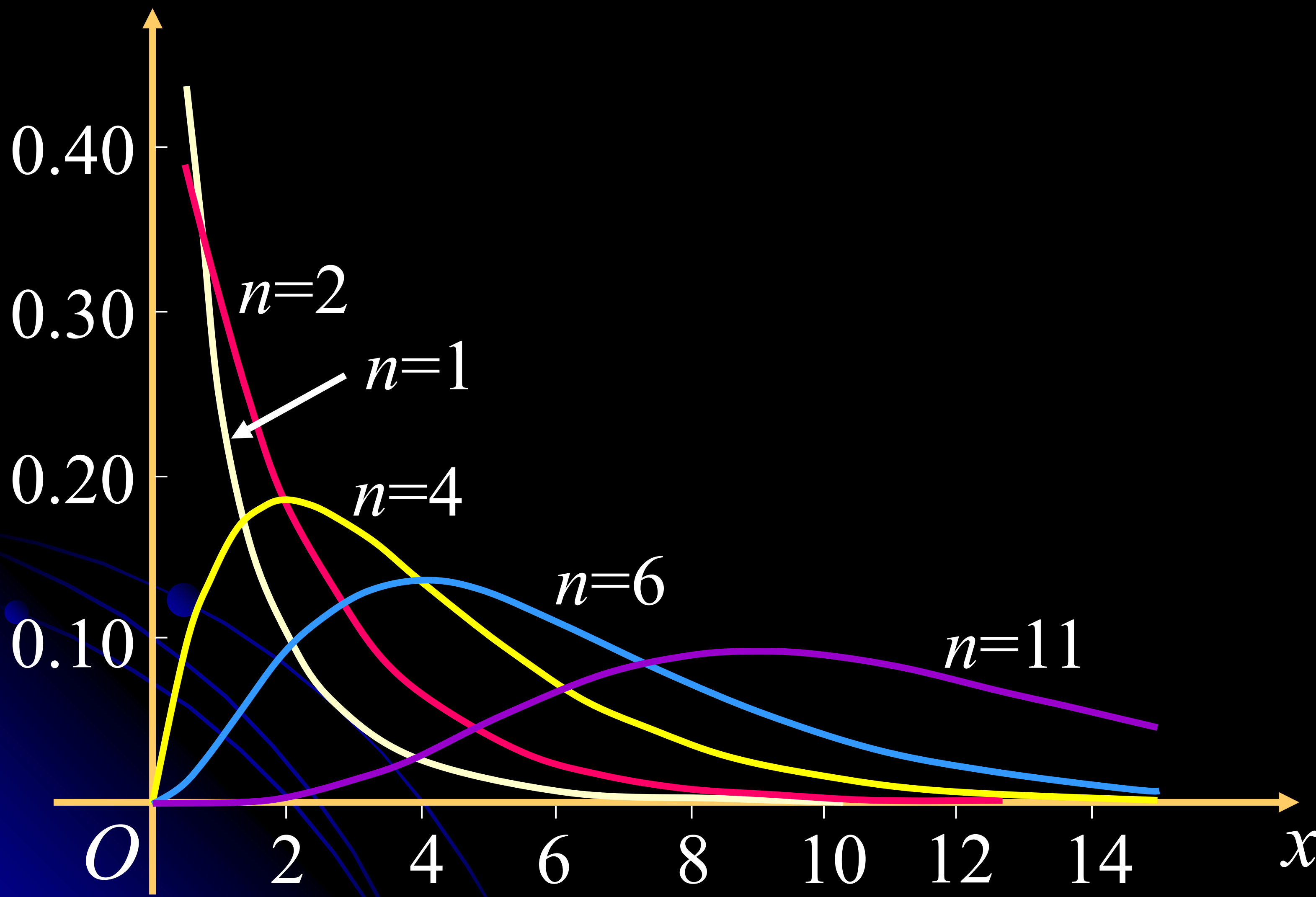
Gamma 数的性质:

(1) $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$;

(2) $\Gamma(n)=(\mathrm{n}\text{-}1)!$, n 为正整数;

(3) $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$

$f(x)$的图形:



χ^2分布的性质:

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

证明: 由$X_i \sim N(0,1)$, 故

$$E(X_i^2)=D(X_i)=1,$$

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

$$=\frac{-x^3}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Bigg|_{-\infty}^{+\infty}+3\int_{-\infty}^{+\infty}x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,\mathrm{d}x=3$$

$$E(X_i^2)=1, E(X_i^4)=3$$

$$D(X_i^2)=E(X_i^4)-[E(X_i^2)]^2=3-1=2,$$

$$i=1,2,\cdots,n$$

由$X_1, X_2, \ldots, X_n$的独立性, 于是

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

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

(2)若$\chi_1^2 \sim \chi^2(m), \chi_2^2 \sim \chi^2(n)$,且$\chi_1^2, \chi_2^2$相互独立，则

$$\chi_1^2 + \chi_2^2 \sim \chi^2(m+n).$$

证明 由χ^2分布的定义，可设

$$\chi_1^2 = X_1^2 + X_2^2 + \cdots + X_m^2, \chi_2^2 = X_{m+1}^2 + X_{m+2}^2 + \cdots + X_{m+n}^2$$

其中 $X_1, X_2, \cdots, X_m, X_{m+1}, X_{m+2}, \cdots, X_{m+n}$ 均服从 $N(0,1)$，且相互独立，于是由χ^2分布的定义

$$\chi_1^2 + \chi_2^2 = X_1^2 + \cdots + X_m^2 + X_{m+1}^2 + \cdots + X_{m+n}^2$$

服从$\chi^2(m+n)$.

(3) χ^2 分布的分位数:

设χ^2~$\chi^2(n)$, 对给定的实数$\alpha(0<\alpha<1)$, 称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2}^{+\infty} f(x)\,\mathrm{d}\,x = \alpha \quad (2.2)$$

的数$\chi_\alpha^2(n)$为**$\chi^2(n)$分布的水平α的上侧分位数**, 简称为上侧α分位数. 对不同的α与 n, 分位数的值已经编制成表供查用(参见附表).

例如, 查表得

$$\chi^2_{0.05}(10) = 18.307$$

$$\chi^2_{0.1}(25) = 34.382$$

	0.25	0.1	0.05	0.025	0.01	0.005
………………						
6	7.841	10.645	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.300
13	15.984	19.812	22.362	24.736	27.688	29.819
………………						
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.195	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993

例2 设$X_1,\ldots,X_6$是来自总体$N(0,1)$的样本,

又设

$$Y=(X_1+X_2+X_3)^2+(X_4+X_5+X_6)^2$$

试求常数C, 使CY服从χ^2分布.

解 因为 $X_1+X_2+X_3 \sim N(0,3)$

$$X_4+X_5+X_6 \sim N(0,3)$$

所以 $\dfrac{X_1+X_2+X_3}{\sqrt{3}} \sim N(0,1), \dfrac{X_4+X_5+X_6}{\sqrt{3}} \sim N(0,1),$

且相互独立.

于是

$$\left(\frac{X_1+X_2+X_3}{\sqrt{3}}\right)^2 + \left(\frac{X_4+X_5+X_6}{\sqrt{3}}\right)^2 \sim \chi^2(2)$$

故应取 $C=\dfrac{1}{3}$,则有 $\dfrac{1}{3}Y \sim \chi^2(2).$

三 t 分布

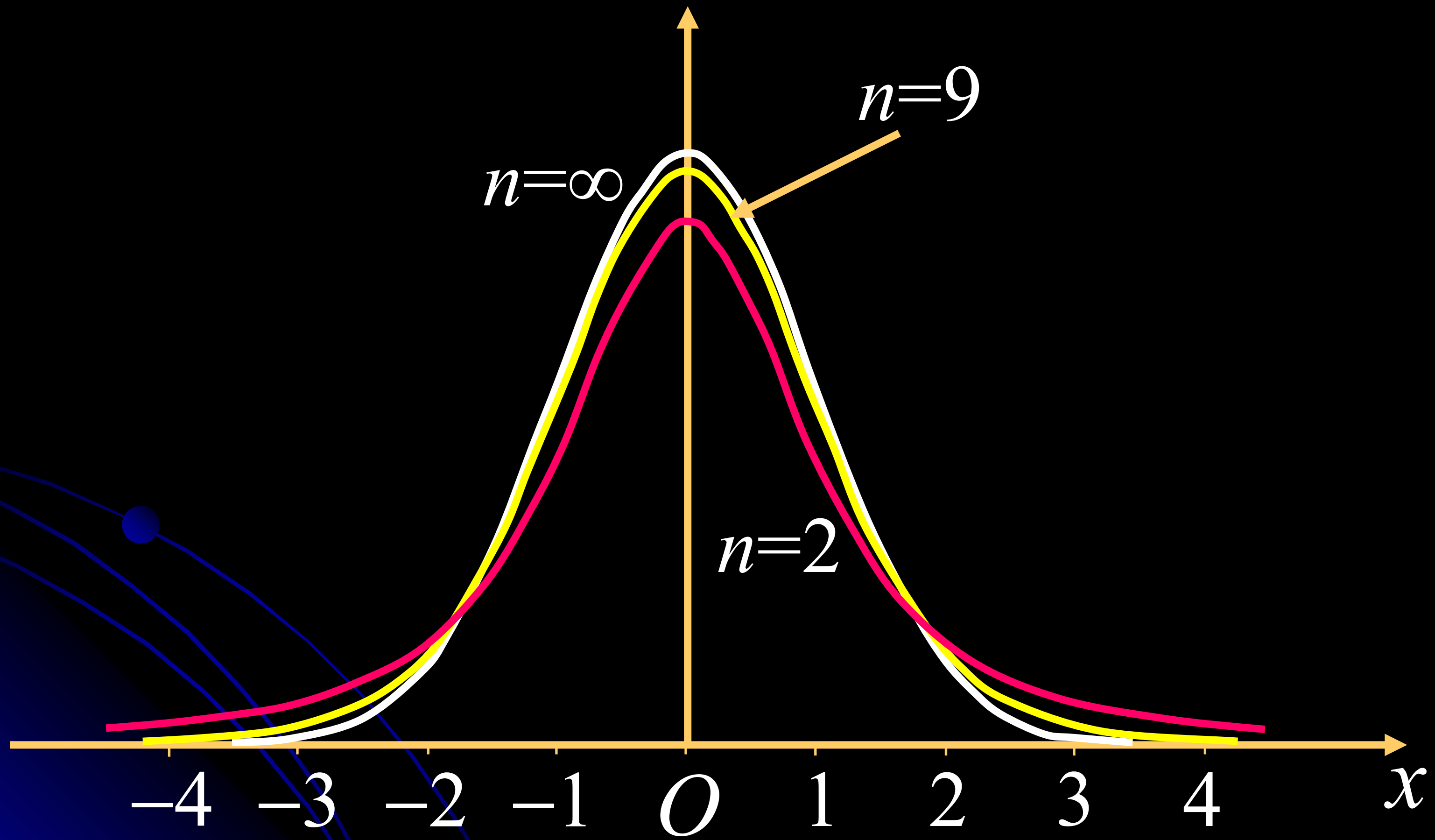
定义2 设$X\sim N(0,1)$, $Y\sim\chi^2(n)$, 且X与Y相互独立, 则称

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

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:

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

t 分布概率密度的图形:



t分布具有如下性质:

(1) $f(x)$的图形关于y轴对称, 且

$$\lim_{x\to\infty} f(x) = 0;$$

(2) 当n充分大时, t分布近似于正态分布;

$$\lim_{n\to\infty} f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

但当n较小时, t分布与标准正态分布仍相差较大.

(3) t分布的分位数:

设$T\sim t(n)$, 对给定的实数$\alpha(0<\alpha<1)$, 称满足条件

$$P\{T>t_\alpha(n)\}=\int_{t_\alpha(n)}^{+\infty}f(x)\,\mathrm{d}x=\alpha \qquad (2.5)$$

的数$t_\alpha(n)$为**$t(n)$分布的水平α的上侧分位数**. 由密度函数的对称性, 可得

$$t_{1-\alpha}(n)=-t_\alpha(n) \qquad (2.6)$$

类似地, 我们可以给出t分布的双侧分位数

$$P\{|T|>t_{\alpha/2}(n)\}=$$

$$\int_{-\infty}^{-t_{\alpha/2}} f(x)\,\mathrm{d}x+\int_{t_{\alpha/2}}^{+\infty} f(x)\,\mathrm{d}x=\alpha,$$

显然有

$$P\{T>t_{\alpha/2}(n)\}=\alpha/2;\ P\{T<-t_{\alpha/2}(n)\}=\alpha/2.$$

对不同的α与n, t分布的上侧分位数可从附表查得.

例如, 设$t\sim t(8)$, 对水平α=0.05, 查表得到

$t_{\alpha}(8)=t_{0.05}(8)=1.8595$, $t_{\alpha/2}(8)=t_{0.025}(8)=2.3060$

	0. 25	0. 1	0. 05	0. 025	0. 01	0. 005
1	1. 0000	3. 0777	6. 3138	12. 7062	31. 8205	63. 6567
2	0. 8165	1. 8856	2. 9200	4. 3027	6. 9646	9. 9248
3	0. 7649	1. 6377	2. 3534	3. 1824	4. 5407	5. 8409
4	0. 7407	1. 5332	2. 1318	2. 7764	3. 7469	4. 6041
5	0. 7267	1. 4759	2. 0150	2. 5706	3. 3649	4. 0321
6	0. 7176	1. 4398	1. 9432	2. 4469	3. 1427	3. 7074
7	0. 7111	1. 4149	1. 8946	2. 3646	2. 9980	3. 4995
8	0. 7064	1. 3968	1. 8595	2. 3060	2. 8965	3. 3554
9	0. 7027	1. 3830	1. 8331	2. 2622	2. 8214	3. 2498
10	0. 6998	1. 3722	1. 8125	2. 2281	2. 7638	3. 1693

例如, 设$T \sim t(8)$, 对水平$\alpha=0.05$, 查表得到

$t_{\alpha}(8)=t_{0.05}(8)=1.8595$, $t_{\alpha/2}(8)=t_{0.025}(8)=2.3060$

故有

$$P\{T>1.8595\}=P\{T<-1.8595\}$$

$$=P\{|T|>2.3060\}=0.05.$$

注: ①当自由度n充分大时, t分布近似于正态分布, 故有

$$t_{\alpha}(n) \approx u_{\alpha}; \quad t_{\alpha/2}(n) \approx u_{\alpha/2}.$$

从而当$n>45$时, t分布的分位数可用正态分布的分位数近似.

②设$t_{\alpha}(n)$为$t(n)$的上侧α分位数, 则

$$P\{T<t_{\alpha}(n)\}=1-\alpha, P\{T<-t_{\alpha}(n)\}=\alpha,$$

$$P\{|T|>t_{\alpha}(n)\}=2\alpha.$$

例3 设随机变量$X \sim N(2,1)$, 随机变量Y_1, Y_2, Y_3, Y_4均服从$N(0,4)$, 且$X, Y_i (i=1,2,3,4)$都相互独立, 令

$$T = \frac{4(X-2)}{\sqrt{\sum_{i=1}^{4} Y_i^2}},$$

试求T的分布, 并确定t_0的值, 使$P\{|T|>t_0\}=0.01$.

解 由于$X-2 \sim N(0,1)$, $Y_i/2 \sim N(0,1)$, $i=1,2,3,4$, 故由t分布的定义知

$$T=\frac{4(X-2)}{\sqrt{\sum_{i=1}^{4} Y_i^2}}=\frac{X-2}{\sqrt{\sum_{i=1}^{4}\left(\frac{Y_i}{2}\right)^2 \Big/ 4}} \sim t(4)$$

由$P\{|T|>t_0\}=0.01$, 对于$n=4$, $\alpha=0.01$查附表$t_0=t_{\alpha/2}(4)=t_{0.005}(4)=4.6041$.

四, F分布(用以方差分析, 协方差分析和回归分析)

定义3 设$X\sim\chi^2(m)$, $Y\sim\chi^2(n)$, 且X与Y相互独立, 则称

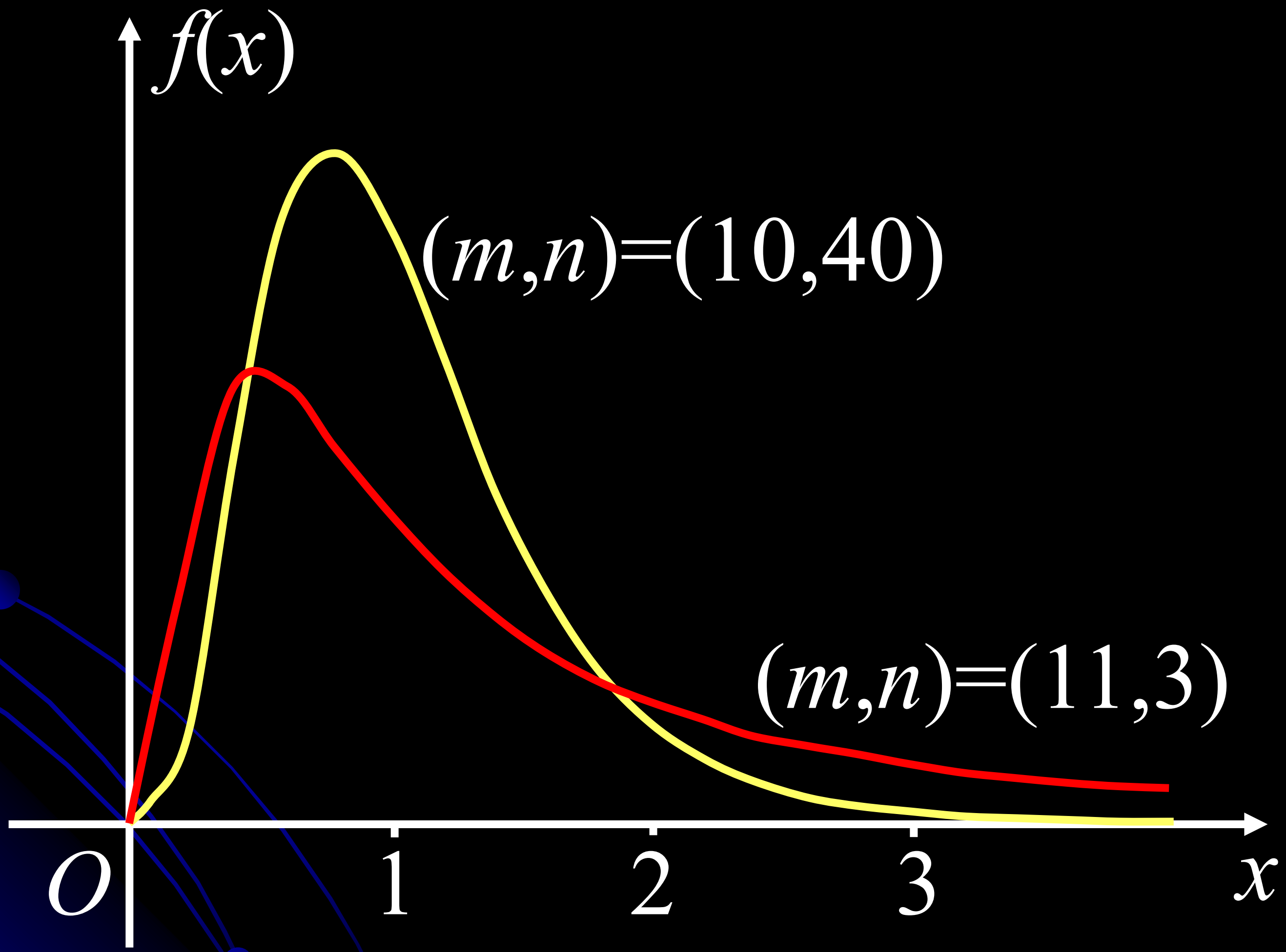
$$F=\frac{X/m}{Y/n}=\frac{nX}{mY} \qquad (2.7)$$

服从自由度为(m,n)的F分布, 记为$F\sim F(m,n)$

$F(m,n)$分布的概率密度:

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

F分布概率密度的图形:



F分布具有如下性质:

(1) 若$X \sim t(n)$, 则$X^2 \sim F(1,n)$;

(2) 若$F \sim F(m,n)$, 则

$$\frac{1}{F} \sim F(n,m).$$

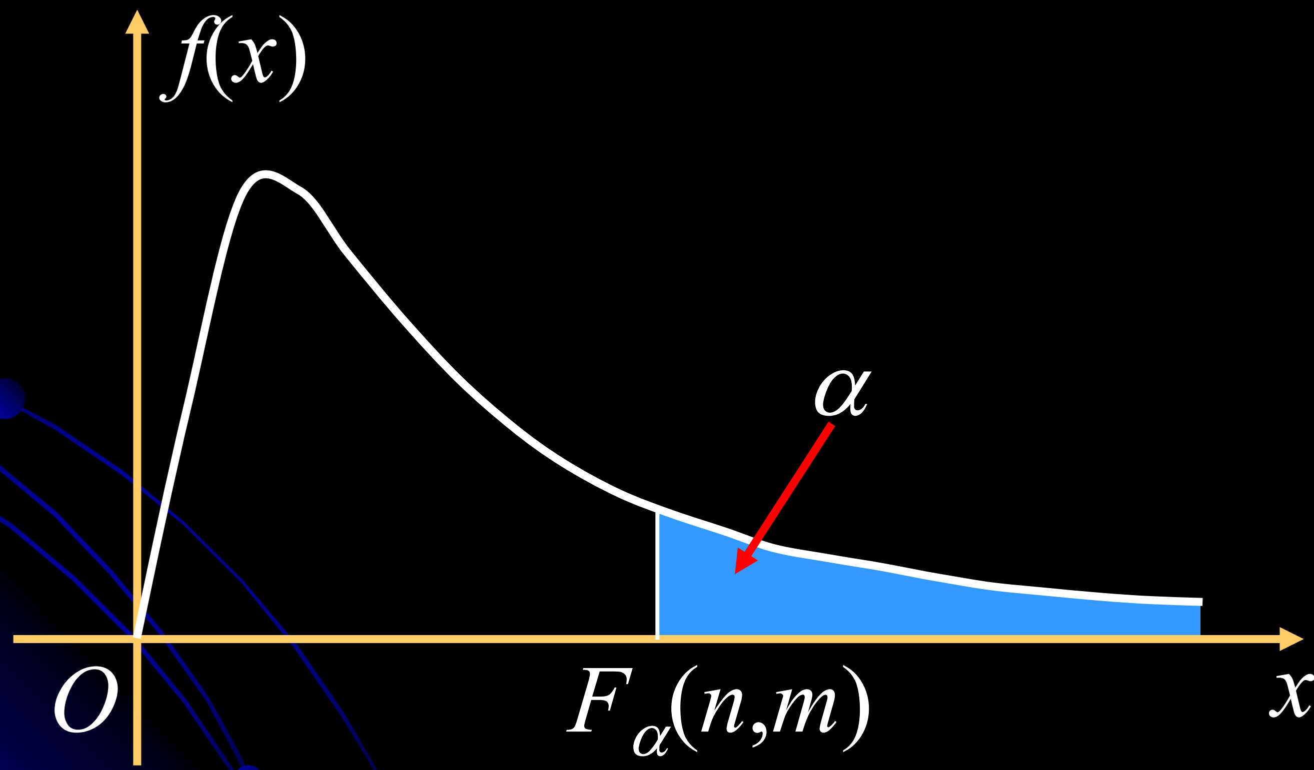
(3) *F*分布的分位数:

设$F\sim F(n,m)$, 对给定的实数$\alpha(0<\alpha<1)$, 称满足条件

$$P\{F > F_{\alpha}(n,m)\} = \int_{F_{\alpha}(n,m)}^{+\infty} f(x)\,\mathrm{d}\,x = \alpha$$

的数$F_{\alpha}(n,m)$为**$F(n,m)$分布的水平α的上侧分位数**. F分布的上侧分位数可自附表查得.

$F(n,m)$分布的水平α的上侧分位数的示意图:



(4) F分布的分位数的一个重要性质:

$$F_{\alpha}(m,n)=\frac{1}{F_{1-\alpha}(n,m)}.$$

此式常用来求F分布表中没有列出的某些上侧分位数, 可用上式求之

例如, $F_{0.05}(10,5)=4.74$, $F_{0.025}(5,10)=4.24$.

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

例4 设总体X服从标准正态分布，$X_1, X_2, \ldots, X_n$是来自X的一个简单随机样本，试问统计量

$$Y = \left(\frac{n}{5} - 1\right) \sum_{i=1}^{5} X_i^2 \Big/ \sum_{i=6}^{n} X_i^2, n > 5$$

服从何种分布？

解 因为 $X_i \sim N(0,1)$,

$\sum_{i=1}^{5} X_i^2 \sim \chi^2(5), \sum_{i=6}^{n} X_i^2 \sim \chi^2(n-5)$, 且 $\sum_{i=1}^{5} X_i^2$ 与 $\sum_{i=6}^{n} X_i^2$ 相互独立, 所以

$$\frac{\sum_{i=1}^{5} X_i^2 \Big/ 5}{\sum_{i=6}^{n} X_i^2 \Big/ (n-5)} \sim F(5, n-5),$$

$$\frac{\sum_{i=1}^{5} X_i^2 \Big/ 5}{\sum_{i=6}^{n} X_i^2 \Big/ (n-5)} \sim F(5, n-5),$$

由于$Y = \left(\frac{n}{5} - 1\right) \sum_{i=1}^{5} X_i^2 \Big/ \sum_{i=6}^{n} X_i^2 = \frac{\sum_{i=1}^{5} X_i^2 \Big/ 5}{\sum_{i=6}^{n} X_i^2 \Big/ (n-5)}$

故 $Y \sim F(5, n-5)$

课堂练习

1. 设 X_1,X_2,X_3,X_4,X_5 是来自正态总体 $N(0,2^2)$ 的样本.

(1) 求 C 使统计量 $Y_1=\dfrac{C(X_1+X_2)}{\sqrt{X_3^2+X_4^2+X_5^2}}$ 服从 t 分布.

(2) 求 $Y_2=\dfrac{(X_1+X_2)^2}{(X_4-X_3)^2}$ 所服从的分布.

作业 习题5-2 第138页开始

第4,5,6,9题