数理统计

—以概率论为基础，研究如何收集，整理和分析受随机因素影响的数据，以便得出合理有效推断的数学分支

☆基础——概率论

☆功能——处理数据

☆目的——作出科学推断（就概率特征）

- 数理统计
  - 统计推断
    - 估计
    - 假设检验
  - 试验设计

第二章　　抽样与抽样分布

– 总体与样本
– 总体分布的估计
– 统计量
– 数理统计中的三大抽样分布
– 正态总体的抽样分布

总体与样本

1.总体—— 研究对象的全体称为总体

如果研究对象有有限个元素(个体)称之为有限总体

如果研究对象有无穷多个元素则称之为无限总体

说明: 1) 当总体元素个数有限但非常多时,可视为无限总体

2) 作为研究对象的总体可能有很多特性,我们关心的是其某个数量指标.因此,总体从数量上看就是一堆数

3) 总体从数量上看就是一堆数,这些数有的可能出现次数较多,有的出现次数较少.它有一个自己的分布.因此,总体可视为一个随机变量, 记作 $X, Y, \cdots,\ \xi, \eta, \cdots$

比如:要考察某宿舍四名同学的英语学习情况,他们的英语成绩分别为：68,75,82,75;这四名同学的英语成绩就是总体

这些成绩的频率分布表为:

成绩	68	75	82
频率	1/4	1/2	1/4

设X为任取该宿舍一名同学的英语成绩,则X可视为一个随机变量,X的分布列为:

X	68	75	82
概率	1/4	1/2	1/4

X与总体具有相同的分布,因此,可把总体视为随机变量X

又比如：要考察某厂生产的灯管的寿命．则该厂生产的所有灯管（的寿命）构成总体．

总体（该厂所有灯管寿命）的分布与任取该厂一个灯管的寿命X的分布相同．因此，可以把总体用随机变量X来表示．

等价定义：总体—— 作为研究对象的随机变量

如何来考察总体性质和特征呢?

答案是做试验．但全面试验有时因为成本太大，或者试验具有破坏性是不可取的．此时，可通过"抽样试验"的方式，即从总体中抽取一部分个体进行试验，由部分个体的试验结果来推断总体的性质和特征

2.样本

从总体中抽取若干个个体进行试验, 这若干个个体的试验结果叫样本.

即: 样本就是对总体进行的若干次试验（考察）所得到的结果.

抽样— 就是从总体中抽取若干个个体的过程.

抽样的方法包括: 随机抽样, 机械抽样, 整群抽样, 分层抽样 等.

数理统计中涉及的是"随机抽样"

----它表示:总体中的每个个体都有相同的可能性被抽到.

$$\text{随机抽样}\begin{cases}\text{有返回抽样}\\\text{不返回抽样}\end{cases}$$

二者的区别是: 有返回抽样满足独立性

当总体是无限总体时, 有返回抽样和不返回抽样差别不大,

可以把不返回抽样视为有返回抽样来处理.

样本是对总体进行的若干次试验考察的结果.

但是,在试验考察之前, 我们并不知道各次试验的结果是多少

因此,各次试验的结果也都是随机变量

因此,样本可记作 $(X_1, X_2, \cdots, X_n), (Y_1, Y_2, \cdots, Y_n), \cdots$

其中X_i是对总体进行的第i次考察的结果, 是随机变量

n是抽样个数(对总体考察的次数), 称为样本的容量

一旦抽样试验结束, 就得到了样本 $(X_1, X_2, \cdots, X_n)$的一组具体取值(一组数)，这组数称为样本观测值,

记作$(x_1, x_2, \cdots, x_n)$

3.简单随机样本

如果样本 $(X_1, X_2, \cdots, X_n)$ 满足:

1) 代表性: X_i与总体服从相同的分布

2) 独立性: $X_1, X_2, \cdots, X_n$相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为一个简单随机样本(子样),

简称: 样本.

注: 样本的代表性要求抽样方法是随机抽样

独立性要求抽样方法是有返回的抽样

4.样本的联合分布

设$X_1, X_2, \cdots, X_n$为来自总体X的一组样本，

(1)若总体X是离散型随机变量，则$X_1, X_2, \cdots, X_n$ 的联合概率分布为

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} P\{X_i = x_i\}$$

(2)若总体X是连续型随机变量，密率密度为$p(x)$，

则$X_1, X_2, \cdots, X_n$的联合概率密度为

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

(3)若总体X的分布函数为$F(x)$，则$X_1, X_2, \cdots, X_n$的联合分布函数为

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

例1 设总体ξ~P(λ),($X_1,X_2,...,X_n$)是ξ的样本，求($X_1,X_2,...,X_n$)的联合概率分布.

解：因为ξ~P(λ),ξ的概率分布为$P(\xi=k)=\frac{\lambda^k}{k!}e^{-\lambda},k=0,1,2,...$

所以($X_1,X_2,...,X_n$)的联合概率分布为

$$\prod_{i=1}^{n}p\{\xi=x_i\}=\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}=\frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}e^{-n\lambda},x_i=0,1,2,...(i=1,2,...,n)$$

例2 设总体ξ~E(λ),($X_1,X_2,\dots,X_n$)是样本，

求($X_1,X_2,\dots,X_n$)得联合概率密度

解：因为ξ~E(λ),概率密度为$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} & x_i > 0 (i=1,2,\dots,n) \\ 0 & 其他 \end{cases}$

$$= \begin{cases} \lambda^n e^{-\lambda x} & \min\limits_i x_i > 0 \\ 0 & 其他 \end{cases}$$

5.总体分布估计

——如何根据样本数据估计总体的分布

1) 总体分布函数的估计

总体的分布函数 $F(x) = P\{X \le x\}$,根据大数定律,

它可由n次试验中事件 "$X \le x$" 发生的频率来估计

即: $F(x) = P\{X \le x\} \approx$ 事件 "$X \le x$" 发生的频率

$$= \frac{\text{n个观测值中不超过x的个数}}{\text{试验总次数n}}$$

经验分布函数

若总体X,样本观测值为$x_1, x_2, \cdots, \ x_n$

将观测值从小到大排列：

$$x_{(1)} < x_{(2)} < \cdots < x_{(l)} \qquad (l \le n),$$

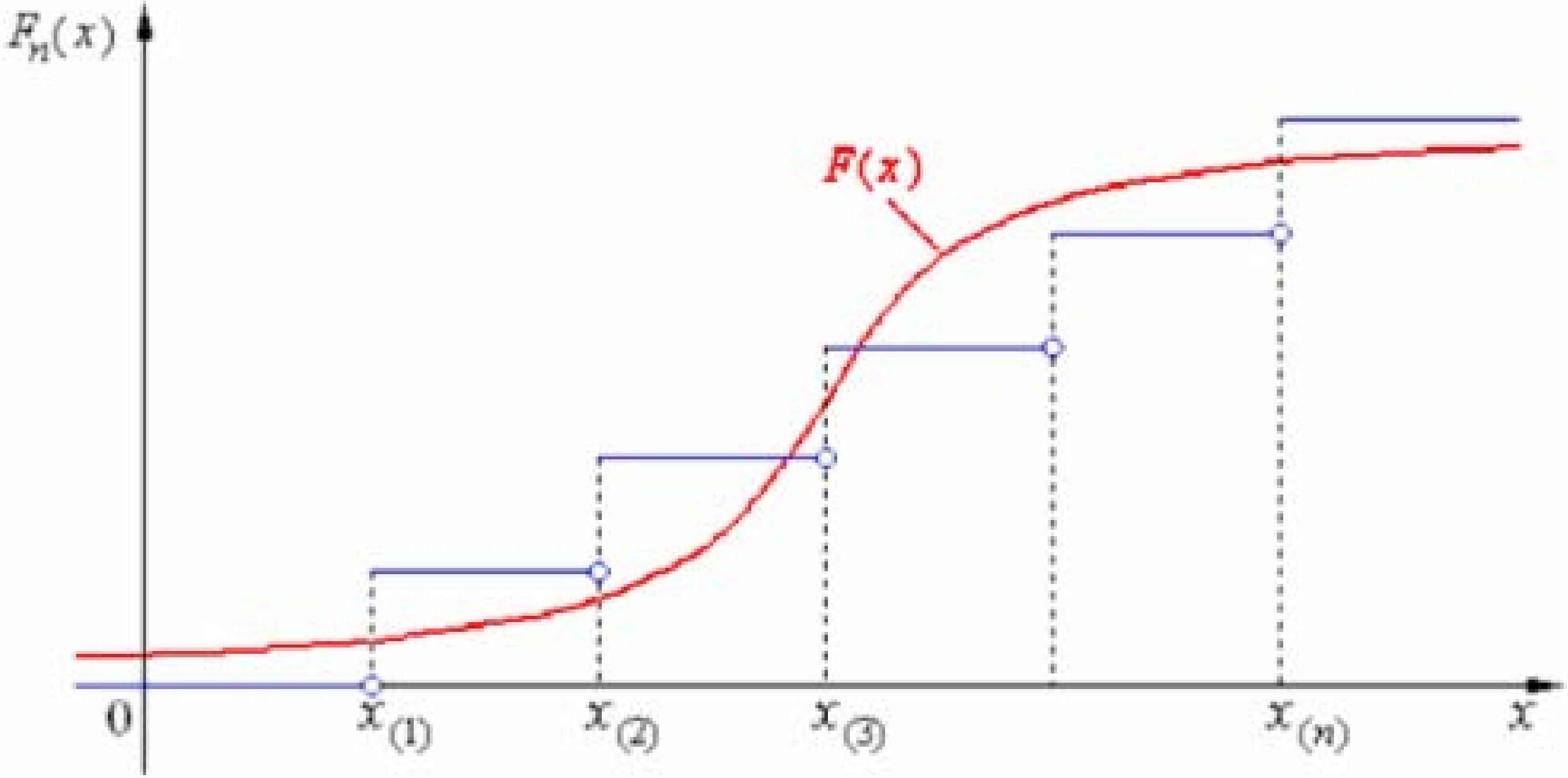
写出频率分布表：

观测值 $x_{(i)}$	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(l)}$
频数 m_i	m_1	m_2	$\cdots$	m_l
频率 $\omega_i = \dfrac{m_i}{n}$	ω_1	ω_2	$\cdots$	ω_l

经验分布函数如下：

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)}; \\ \sum\limits_{x_{(i)} \le x} \omega_i, & \text{当 } x_{(i)} \le x < x_{(i+1)}; \\ 1, & \text{当 } x \ge x_{(l)}. \end{cases}$$

总体分布函数与样本经验分布函数图示



★经验分布函数$F_n(x)$的性质：

(1) $0 \le F_n(x) \le 1$

(2) $F_n(x)$是非减函数

(3) $F_n(-\infty) = 0, F_n(+\infty) = 1$

(4) $F_n(x)$在每个观测点$x_{(i)}$处是右连续的，点$x_{(i)}$是$F_n(x)$的跳跃间断点，$F_n(x)$在点$x_{(i)}$处的跳跃度就等于频率ω_i。

★经验分布函数$F_n(x)$是事件$\xi \le x$的频率；
总体分布函数$F(x)$是事件$\xi \le x$的概率。
则当$n\to\infty$时，$P\{\sup\limits_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0\} = 1$

注：这是我们在数理统计中*用样本推断总体*的理论基础。

2) 离散型总体分布列的估计

总体X的样本观测值为$x_1, x_2, \cdots$，x_n将观测值从小到大排列

$$x_{(1)} < x_{(2)} < \cdots < x_{(l)} \qquad (l \le n),$$

写出频率分布表：

观测值 $x_{(i)}$	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(l)}$
频数 m_i	m_1	m_2	$\cdots$	m_l
频率 $\omega_i = \frac{m_i}{n}$	ω_1	ω_2	$\cdots$	ω_l

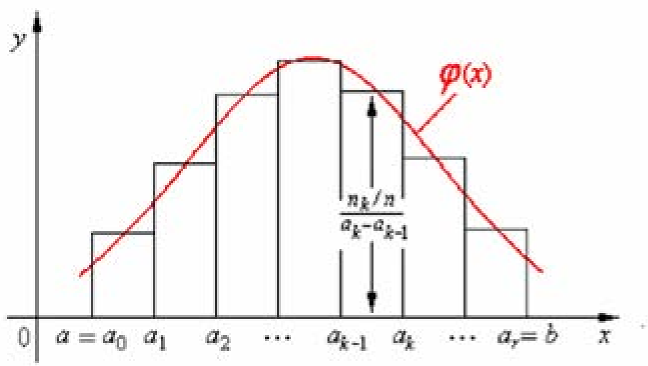
样本频率分布表--------→总体分布列

3) 连续型总体密度函数的估计

如果总体ξ是一个连续型随机变量，

我们可以用下列方法来估计它的概率密度φ(x)

作分点$a=a_0<a_1<a_2<\ldots a_r=b$,将$\xi$的样本取值范围（a, b）分成r个区间。设共进行了n次试验，落在$(a_{k-1}, a_k]$中的样本观测值的个数为n_k, n_k 称为频数，n_k/n称为频率. 在每一个区间$(a_{k-1}, a_k]$上，$\dfrac{n_k/n}{a_k-a_{k-1}}$为高度。这样得到的一排长方形，称为频率直方图（见下图）.



6统计量

定义：样本的不含未知参数的函数叫统计量

注意: 1.统计量(函数)是对样本信息的提炼。

2.样本的函数中不包含任何未知参数，是为了推断的可行性。

3. 统计量是随机变量.

统计量的分布称为抽样分布

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

3. 修正样本方差： $$S^{*2} = \frac{1}{n\text{-}1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

有的书直接定义修正样本方差为样本方差,二者关系：

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

4. 样本标准差： $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

5. 修正样本标准差： $S^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

6. 样本k阶原点矩： $\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

7. 次序统计量： 将样本的各个分量 $X_1, X_2, \cdots, X_n$ 按从小到大的次序排列，得到 $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$，常称 $X_{(i)}$ 为样本的第 i 个“次序统计量”。特别地， $X_{(1)} = \min\limits_{1 \le i \le n} X_i$ 称为最小次序统计量， $X_{(n)} = \max\limits_{1 \le i \le n} X_i$ 称为最大次序统计量。

8.样本中位数:

$$M_e \overset{\Delta}{=} \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n\text{为奇数} \\ \dfrac{1}{2}\left[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right], & n\text{为偶数} \end{cases}$$

9.极差:

$$R \overset{\Delta}{=} X_{(n)} - X_{(1)} = \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i$$

例3 设有样本观测值0.7,0.1,0.8,0.4，求样本均值 $\bar{X}$、样本2阶矩 $\bar{X}^2$、样本方差S^2、样本标准差S、修正样本方差S^{*2}、修正样本标准差S^*的值

$$\bar{X} = \frac{0.7+0.1+0.8+0.4}{4} = \frac{2}{4} = 0.5 \quad ;$$

$$\overline{X^2} = \frac{0.7^2+0.1^2+0.8^2+0.4^2}{4} = \frac{1.3}{4} = 0.325 \quad ;$$

$$S^2 = \overline{X^2} - \bar{X}^2 = 0.325 - 0.5^2 = 0.325 - 0.25 = 0.075 \quad ;$$

$$S = \sqrt{S^2} = \sqrt{0.075} = 0.27386128 \quad ;$$

$$S^{*2} = \frac{n}{n-1} S^2 = \frac{4}{4-1} \times 0.075 = 0.1 \quad ;$$

$$S^* = \sqrt{S^{*2}} = \sqrt{0.1} = 0.31622777 \quad 。$$

注意：用带统计功能的计算器计算常用统计量的观测值时，要注意其方差的定义（是除以n，还是除以n-1的）

可以用EXCEL的统计分析功能来求常用统计量的观测值时

定理2.1 设总体ξ的数学期望$E\xi$和方差$D\xi$都存在，$(X_1, X_2, ...X_n)$是ξ的样本，$\bar{X}$是样本均值，S^2是样本方差，S^{*2}是修正样本方差，则有

$$(1)E\bar{X}=E\xi;(2)D\bar{X}=\frac{D\xi}{n};(3)E(S^2)=\frac{n-1}{n}D\xi;(4)E(S^{*2})=D\xi$$

证 （1）$E\overline{X}=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{1}{n}\sum_{i=1}^{n}E\xi=E\xi$ ；

（2）$D\overline{X}=D(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}\sum_{i=1}^{n}DX_i=\frac{1}{n^2}\sum_{i=1}^{n}D\xi=\frac{D\xi}{n}$ ；

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

$=[D\xi+(E\xi)^2]-[D\overline{X}+(E\overline{X})^2]=[D\xi+(E\xi)^2]-[\frac{D\xi}{n}+(E\xi)^2]=\frac{n-1}{n}D\xi$ ；

（4）$E(S^{*2})=E(\frac{n}{n-1}S^2)=\frac{n}{n-1}E(S^2)=\frac{n}{n-1}\frac{n-1}{n}D\xi=D\xi$。

7.三大抽样分布

除正态分布外,常用的 χ^2 分布, t 分布, F 分布被称为数理统计中的“三大抽样分布”

χ^2 分布（卡方分布）

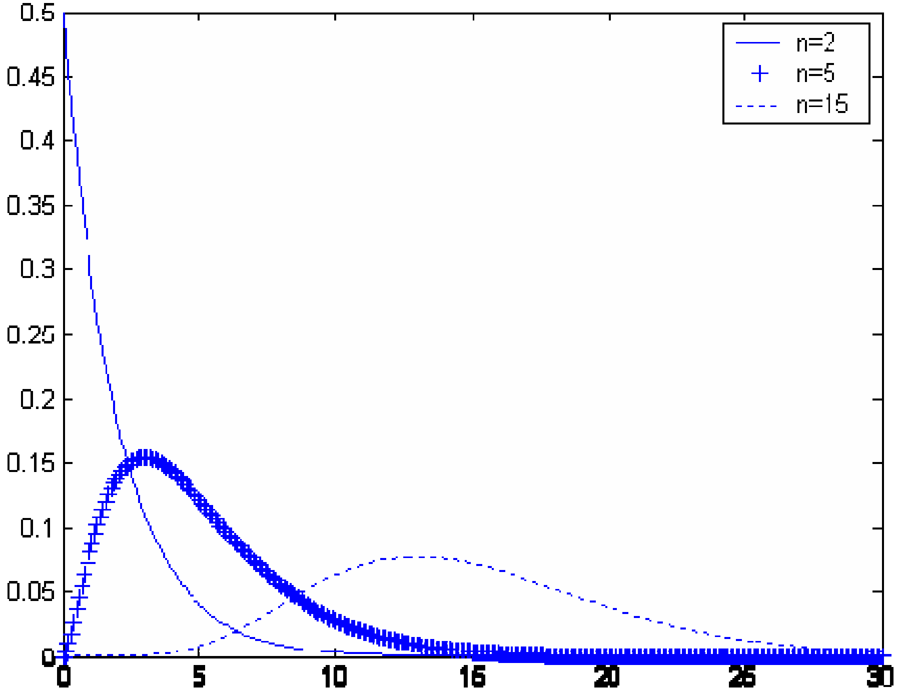
构造性定义：设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，并且都服从标准正态分布 $N(0,1)$，则随机变量

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

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

χ^2分布的密度函数：

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



$p(x)$的性质

1) 当 $x = n - 2$时，$p(x)$取到最大值。

2) 当 n=2 时, p(x) 与 E(1/2) 的密度函数相同

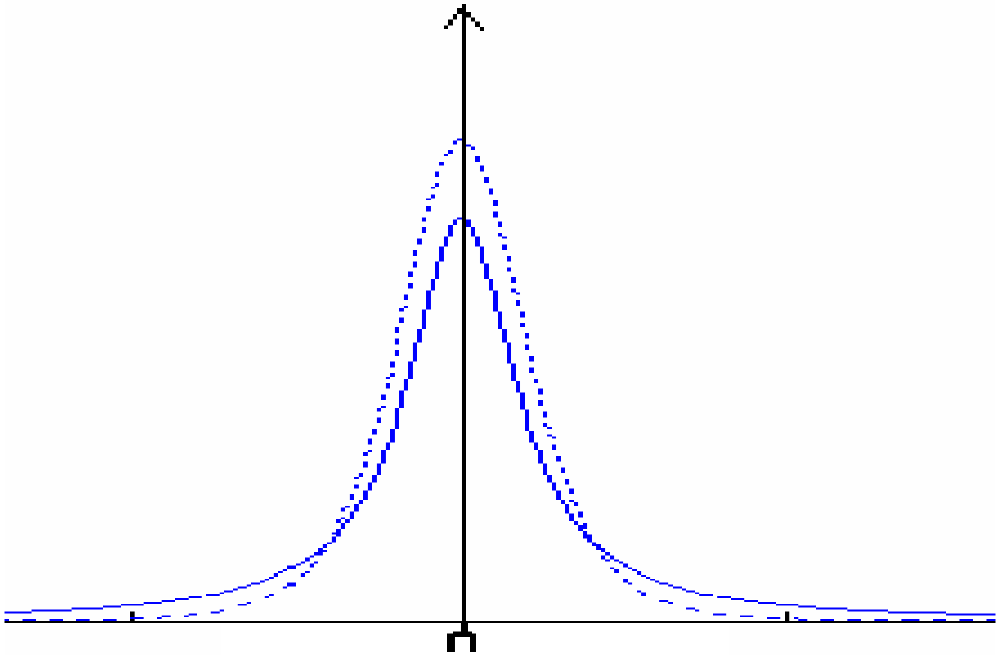
3) $E\chi^2$=n； $D\chi^2$=2n

χ^2分布的性质：设$\xi \sim \chi^2(k_1)$， $\eta \sim \chi^2(k_2)$， ξ与η相互独立，则$\xi+\eta\sim\chi^2(k_1+k_2)$。

t 分布（学生分布）

构造性定义：设随机变量 X 与 Y 相互独立，并且 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，则随机变量 $T = \dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ 服从自由度为 n 的 t 分布，记为 $T \sim t(n)$

密度函数：$p(x) = \dfrac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}(1 + \dfrac{x^2}{n})^{-\frac{n+1}{2}}$



$p(x)$的性质：

(1) $x \to \pm\infty$时，

$p(x) \to 0$

(2) $x = 0$时，$p(x)$取到最大值。

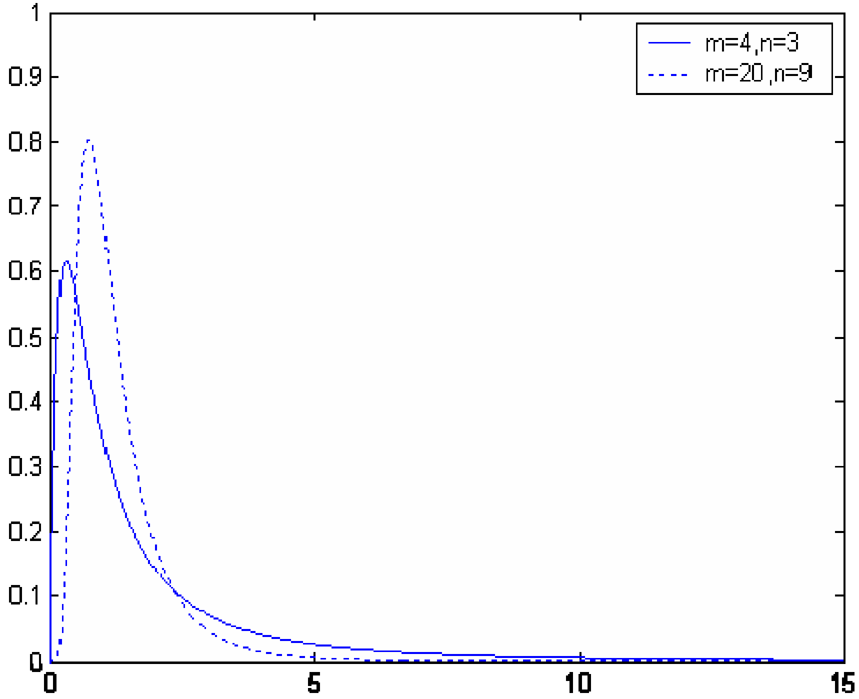
(3) $p(x)$关于$x = 0$对称。

(4) $n \to \infty$时，$p(x) \to \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$（标准正态）

F 分布

构造性定义： 设随机变量 X 与 Y 相互独立，并且 $X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，则随机变量 $F = \dfrac{X/m}{Y/n}$ 服从自由度为 (m,n) 的 F 分布，记为 $F \sim F(m,n)$。

$$\text{密度函数}: p(x) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, & x > 0\text{时} \\ 0, & x \le 0\text{时} \end{cases}$$



F 分布的性质： $F \sim F(m,n) \Rightarrow \frac{1}{F} \sim F(n,m)$

分位数（临界点）: 对给定的概率p及随机变量X,若有常数c,使得 $P\{X \le c\} = p$,则称c为X所服从的分布的(左侧) p分位数

当 $X \sim N(0,1)$ 时, c 记为 u_p; 即 $\mathrm{P}\{X \le u_p\} = p$;

当 $X \sim t(n)$ 时, c 记为 $\mathrm{t}_p(\mathrm{n})$; 即 $\mathrm{P}\{X \le t_p(n)\} = p$;

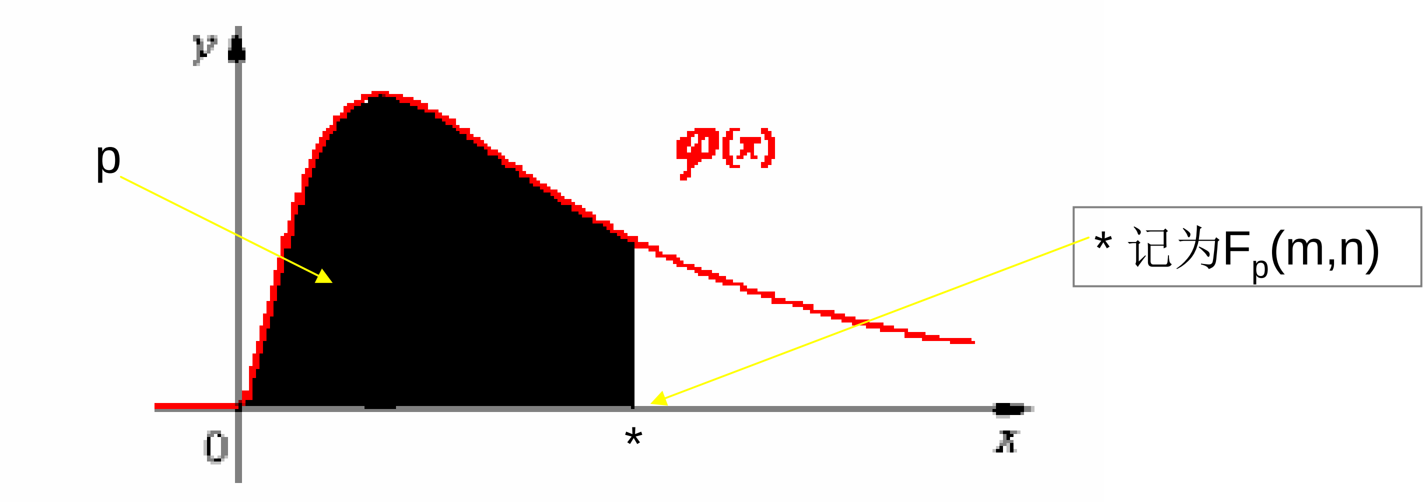
当 $X \sim \chi^2(n)$ 时, c 记为 $\chi_p^2(\mathrm{n})$; 即 $\mathrm{P}\{X \le \chi_p^2(\mathrm{n})\} = p$;

当 $X \sim F(m,n)$ 时, c 记为 $\mathrm{F}_p(\mathrm{m,n})$; 即 $\mathrm{P}\{X \le F_p(m,n)\} = p$

注: 不同教材分位数的含义和表示可能是不同的

正态分布及三大抽样分布的分位数可查表

分位数（临界点）图示及结论



定理：

$$\boxed{\begin{aligned} & u_{1-\alpha} = -u_\alpha \\ & t_{1-\alpha}(n) = -t_\alpha(n) \\ & F_\alpha(m,n) = \frac{1}{F_{1-\alpha}(n,m)} \end{aligned}}$$

证明:

标准正态的密度函数

如图;阴影部分面积为α

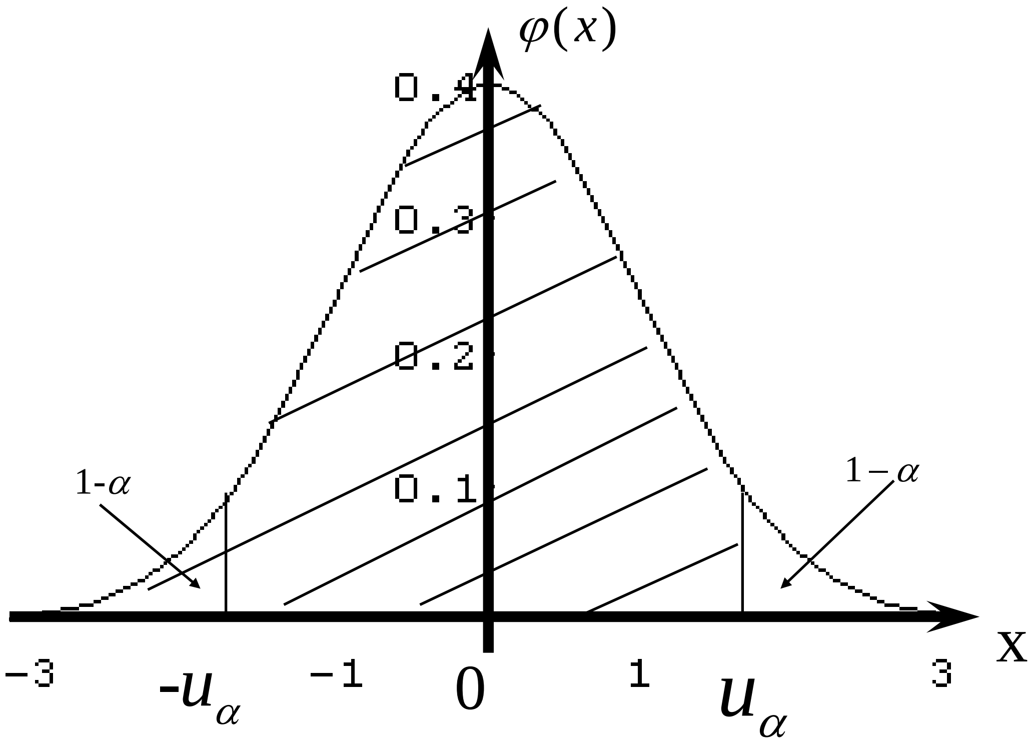
两侧面积均为 1-α

分位数标示如图

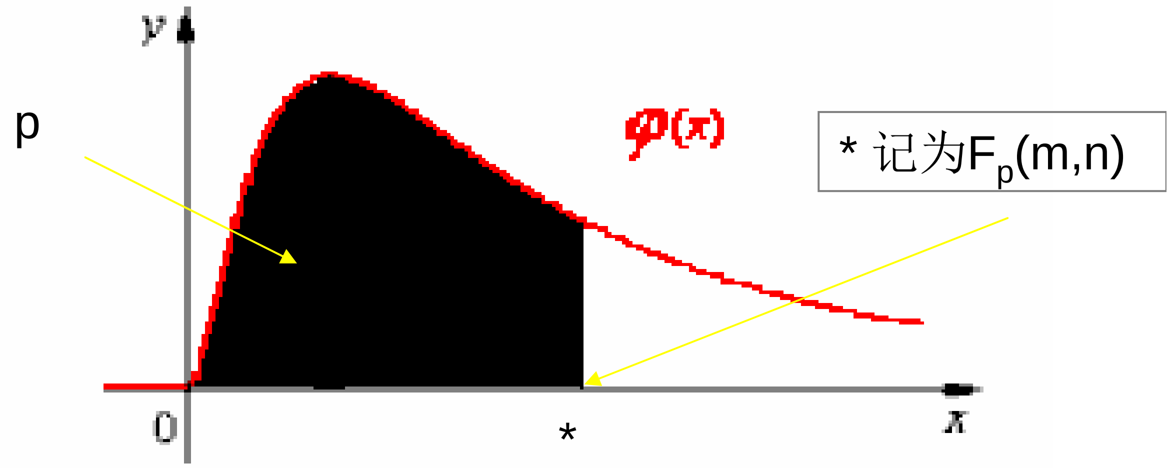
因此: $u_{1-\alpha} = -u_\alpha$

同理可证:

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



证明: $F_p(m,n) = \dfrac{1}{F_{1-p}(n,m)}$



设 $F \sim F(m,n)$

根据分位数定义,

$$P\{F \le F_p(m,n)\} = p$$

于是:

$$P\{\frac{1}{F} \ge \frac{1}{F_p(m,n)}\} = p$$

即 $P\{\dfrac{1}{F} \le \dfrac{1}{F_p(m,n)}\} = 1-p$

因 $F \sim F(m,n)$, 有 $\dfrac{1}{F} \sim F(n,m)$

把 $\dfrac{1}{F_p(m,n)}$ 用分布 $F(n,m)$ 的分位数表示即是 $F_{1-p}(n,m)$

例3 设X_1,X_2,X_3,X_4是来自正态总体$N(0,2^2)$的简单样本，$X=a(X_1-2X_2)^2+b(3X_3-4X_4)^2$，则当$a=$__，$b=$___时，统计量$X$服从$\chi^2$分布，其自由度为____。

分析：由于X服从χ^2分布，可令

$$\sqrt{a}(X_1-2X_2)\sim N(0,1)\text{，}\quad \sqrt{b}(3X_3-4X_4)\sim N(0,1)\text{，}$$

于是,由: $E\sqrt{a}(X_1-2X_2)=0\quad \boldsymbol{E\sqrt{b}(3X_3-4X_4)=0}$

$$D\sqrt{a}(X_1-2X_2)=a(4+4\times4)=20a=1$$

$$D\sqrt{b}(3X_3-4X_4)=b(9\times4+16\times4)=100b=1$$

得： $a=\dfrac{1}{20},b=\dfrac{1}{100},X\sim\chi^2(2)$。