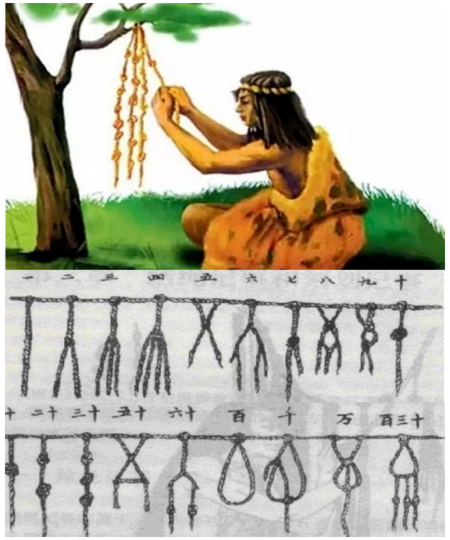
第六章　数理统计的基本知识

1. 总体与样本
2. 直方图与样本函数
3. 统计量及其分布
4. 常用统计量的分布

数理统计(Statistics)溯源久远



古代的统计行为
“结绳计事”

《二十四史》等历史典籍中有许多关于
钱粮、户口、地震、水灾等等的记载

到了十九世纪末二十世纪初,随着近代数学和概率论的发展,才真正诞生了数理统计学这门学科.

数理统计

以概率论为理论基础,根据试验或观测到的数据,研究如何利用有效的方法对这些已知数据进行整理、分析和推断, 从而对研究对象的性质和统计规律作出合理和科学的估计与判断.

计算机的诞生与发展,为数据处理提供了强有力的技术支持.国内外著名的统计软件包： SAS,SPSS,STAT等.

凡是有数据出现的地方就会有数理统计.

§1 总体和样本

1. 总体

2. 样本

总体

总体 研究对象的全体

个体 总体中每个元素

总体容量 总体中所包含的个体数目 $\begin{cases} \text{有限总体} \\ \text{无限总体} \end{cases}$

若有限总体所包含的个体相当多,则可以把它作为无限总体来处理.

以随机变量X表示总体(某一数量指标),而X的分布$F(x)$即为总体的分布.

样本

为研究总体,需观测个体,要付出人力,物力和财力,且某些试验具有破坏性.

样本　从总体中抽取的待测个体的集合.

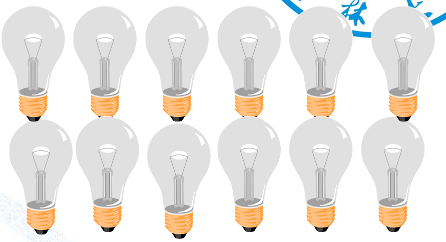
样本容量　样本所包含的个体数目.

样本观测值　样本的确切数值.

从总体X中抽取容量为n的样本常记为$(X_1, X_2, \ldots, X_n)$,

样本值记为$(x_1, x_2, \ldots, x_n)$.

例 某工厂为了检测一批出厂的十万只灯泡的寿命,出厂时随机抽取了1000只灯泡进行检测.



记灯泡的寿命为X,总体可用X或其分布函数$F(x)$来表示.

例 为了统计全国中学生的营养状况,规定每个地区随机抽取百分之一的中学生进行统计调查.

若关心的数量指标是身高和体重,用X和Y分别表示身高和体重,总体可用二维随机变量(X,Y)或其联合分布函数$F(x,y)$来表示.

简单随机抽样

简单随机样本 由简单随机抽样得到的样本$(X_1,X_2,\cdots,X_n)$,满足

(1) $X_1,X_2,\cdots,X_n$ 与X有相同的分布　　代表性

(2) $X_1,X_2,\cdots,X_n$ 相互独立　　独立性

可以用与总体同分布的n个相互独立的随机变量表示.

样本$(X_1,X_2,\ldots,X_n)$的联合分布函数

设总体X的分布函数为$F(x)$, 则样本$(X_1,X_2,\ldots,X_n)$的联合分布函数为

$$F(x_1,x_2,\cdots,x_n)=\prod_{k=1}^{n}F(x_k).$$

1) 若连续总体X具有密度函数$f(x)$, 则样本的联合密度函数为

$$f(x_1,x_2,\cdots,x_n)=\prod_{k=1}^{n}f(x_k).$$

2) 若离散总体X具有分布律$P\{X=x_k\}=p_k, k=1,2,\cdots$ 则样本的联合分布律为

$$P\{X=x_1,X=x_2,\cdots,X=x_k\}=\prod_{k=1}^{n}p_k.$$

例 设总体 X 服从参数为 $1/\lambda\ (\lambda>0)$ 的指数分布, $(X_1,X_2,\cdots,X_n)$ 是来自总体的样本, 求样本 $(X_1,X_2,\cdots,X_n)$ 的概率密度.

解 总体 X 的概率密度为 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$

由 $X_1,X_2,\cdots,X_n$ 相互独立, 且与 X 有相同的分布, 则 $(X_1,X_2,\cdots,X_n)$ 的概率密度为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\begin{cases}\lambda^n e^{-\lambda\sum\limits_{i=1}^{n}x_i}, & x_i>0\\ 0, & \text{其它}\end{cases}$$

例　设总体 $X \sim \pi(\lambda)$, $(X_1, X_2, \cdots, X_n)$是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的分布律.

解　因为$X_1, X_2, \cdots, X_n$ 相互独立, 且与 X有相同的分布, 故 $(X_1, X_2, \cdots, X_n)$的分布律为

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \qquad x_i = 0,1,2,\cdots, i = 1,2,\cdots n.$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

§2 直方图与样本分布函数

1. 直方图

2. 样本分布函数

频率直方图

样本的频率直方图是样本对总体概率密度函数的反映.

作频率直方图的一般步骤如下:

设总体X中抽取到样本观测值$x_1, x_2,\dots, x_n,$

(1) 找出$x_1,x_2,\cdots,x_n$中的最小值 $x_{(1)}$ 和最大值 $x_{(n)}$. 选取略小于 $x_{(1)}$ 的数 a 和略大于$x_{(n)}$ 的数 b .

(2) 根据样本容量确定组数 k , 如果样本容量小, 则组数少些. 如果样本容量大, 则组数多些. 一般来说, 组数取为8～16.

(3) 选取分点 $a=t_0<t_1<\cdots<t_{i-1}<t_i<\cdots<t_k=b$. 把区间 (a,b) 分为 k 个子区间 $(a,t_1],(t_1,t_2],\cdots,(t_{i-1},t_i],\cdots,(t_{k-1},b)$, 第 i 个子区间 $(t_{i-1},t_i]$ 的长度为

$$\Delta t_i=t_i-t_{i-1},\quad i=1,2,\cdots,k.$$

若取各子区间长度相等, 则有

$$\Delta t_i=\frac{b-a}{k},\quad i=1,2\cdots,k.$$

记 $\Delta t=\dfrac{b-a}{k}$. 称 Δt 叫做组距. 此时分点 $t_i=a+i\Delta t,\quad i=1,2,\cdots,k$

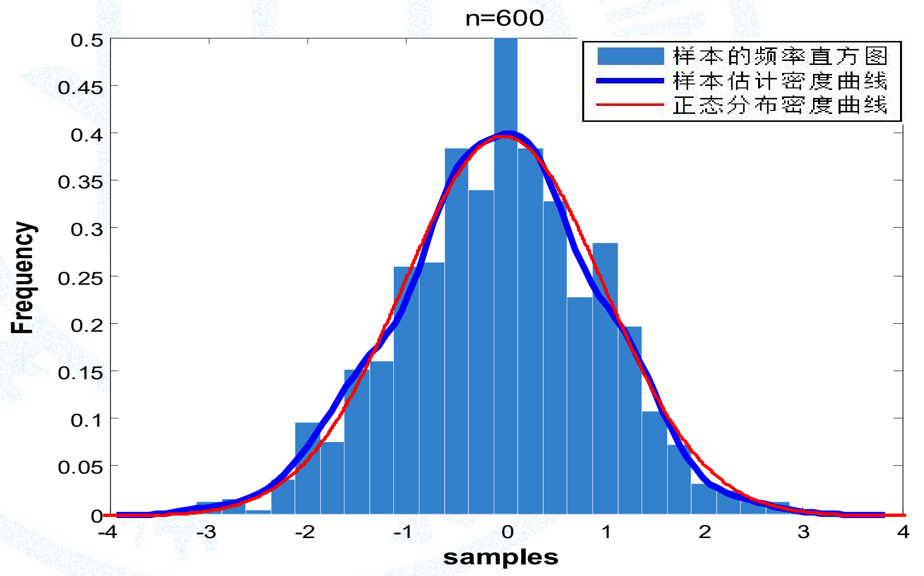
注 分点 t_i 应比样本观测值 x_i 多取一位有效数字.

(4) 数出 $x_1, x_2, \cdots, x_n$落在每个子区间 $(t_{i-1}, t_i]$内的频数n_i，再算出频率

$$f_i = \frac{n_i}{n}, \quad i = 1, 2, \cdots, k.$$

(5) 在 Ox 轴上画出各个分点$t_i\left(i = 0, 1, 2, \cdots, k\right)$，并以各子区间 $(t_{i-1}, t_i]$为底，以 $y_i = \dfrac{f_i}{\Delta t_i}$为高做小矩形.

这样做出的所有小矩形构成了直方图.



频率直方图作用

所有小矩形的面积之和等于1.

第 $i(i=1,2,\cdots,k)$ 个小矩形的面积等于样本观测值落在该子区间内的频率.

当样本容量n充分大时, 随机变量X落在第 $i(i=1,2,\cdots,k)$ 个小区间内的频率近似等于其概率.

直方图大致反映总体X的概率分布.

例 某门课程有120人参加考试，考试成绩 X 如下，

86	83	77	81	81	80	79	82	82	81
83	65	64	78	75	82	80	80	77	81
81	87	82	78	80	81	87	81	77	78
77	78	77	77	77	71	95	78	81	79
80	77	76	82	80	82	84	79	90	82
79	82	79	86	76	78	83	75	82	78
73	83	81	81	83	89	81	86	82	82
78	84	84	84	81	81	74	78	78	80
74	78	75	79	85	75	74	71	88	82
76	85	73	78	81	79	77	78	81	87
75	83	90	80	85	81	77	78	82	84
85	84	82	85	84	82	85	84	78	78

试根据这些数据作出直方图，并根据直方图估计 X 的分布.

解 从n=120个数据中找出最小值$x_{(1)}=64$及最大值$x_{(120)}=95$.

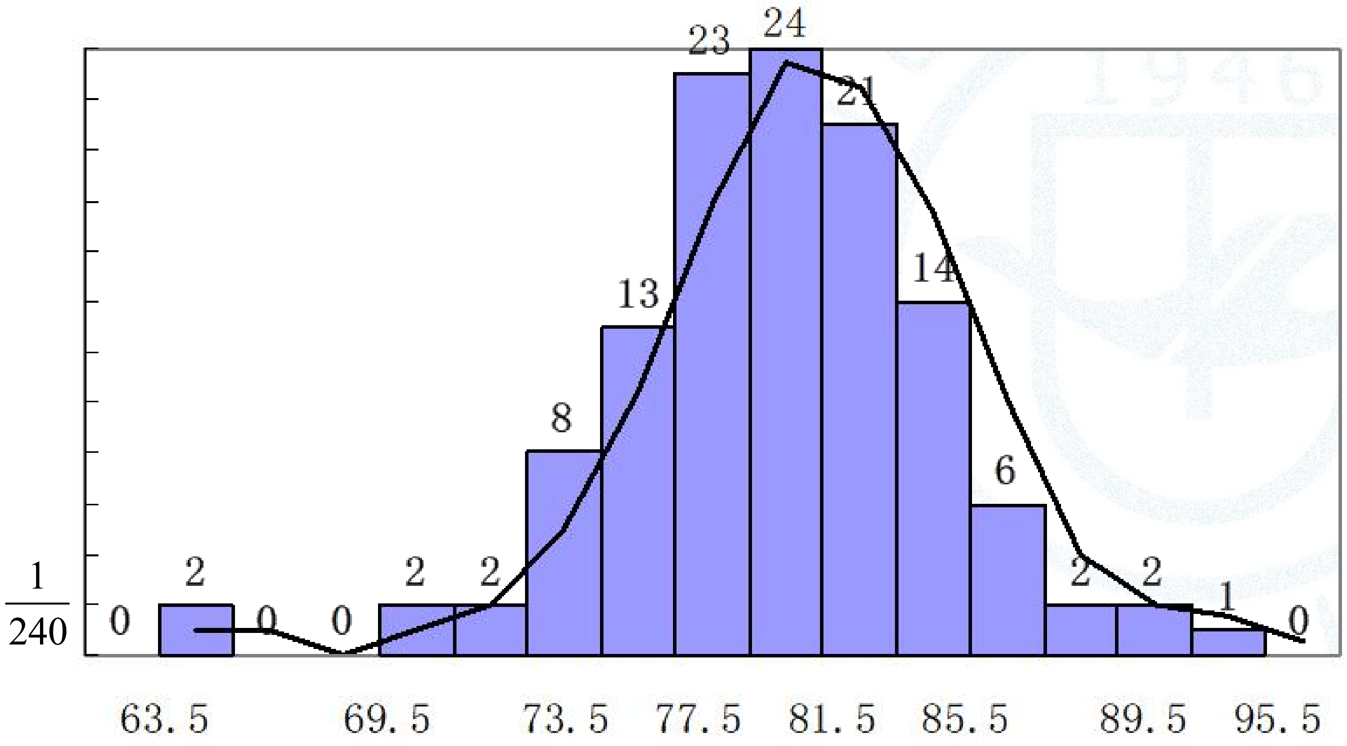
取 $a=63.5$, $b=95.5$, 分$k=16$组, 组距$\Delta t=\dfrac{95.5-63.5}{16}=2.$

分组$(t_{i-1},t_i]$	频数
63.5~65.5	2
65.5~67.5	0
67.5~69.5	0
69.5~71.5	2
71.5~73.5	2
73.5~75.5	8
75.5~77.5	13
77.5~79.5	23

分组$(t_{i-1},t_i]$	频数
79.5~81.5	24
81.5~83.5	21
83.5~85.5	14
85.5~87.5	6
87.5~89.5	2
89.5~91.5	2
91.5~93.5	0
93.5~95.5	1

以横轴 x 轴表示成绩, $a= t_0= 63.5$, $t_1=65.5$, …, $t_{15}= 93.5$, $b= t_{16} = 95.5$,

$\Delta t = 2$, 在$(t_{i-1},t_i]$上, 作高为$y_i = \dfrac{f_i}{\Delta t} = \dfrac{n_i}{n}\cdot\dfrac{1}{\Delta t} = \dfrac{n_i}{240}$的矩形.



样本分布函数

定义 设总体X的分布函数$F(x)$, 从总体X中抽取容量为n的样本, 样本观测值为 $x_1, x_2, \cdots, x_n$(相同的观测值可重复出现), 假设n个观测值中有k个各不相同的值, 按由小到大顺序依次记为

$$x_{(1)} < x_{(2)} < \cdots < x_{(k)} \quad (k \le n)$$

设 $x_{(i)}$ 出现的频数为 n_i , 则 $x_{(i)}$ 出现的频率为 $f_i = \dfrac{n_i}{n}, i = 1, 2, \cdots k.$

显然有 $\sum_{i=1}^{k} n_i = n, \sum_{i=1}^{k} f_i = 1.$ 则称

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}; \\ \sum_{j=1}^{i} f_j, & x_{(i)} \le x < x_{(i+1)}, \quad i = 1, 2, \cdots, k-1; \\ 1, & x \ge x_{(k)}. \end{cases}$$

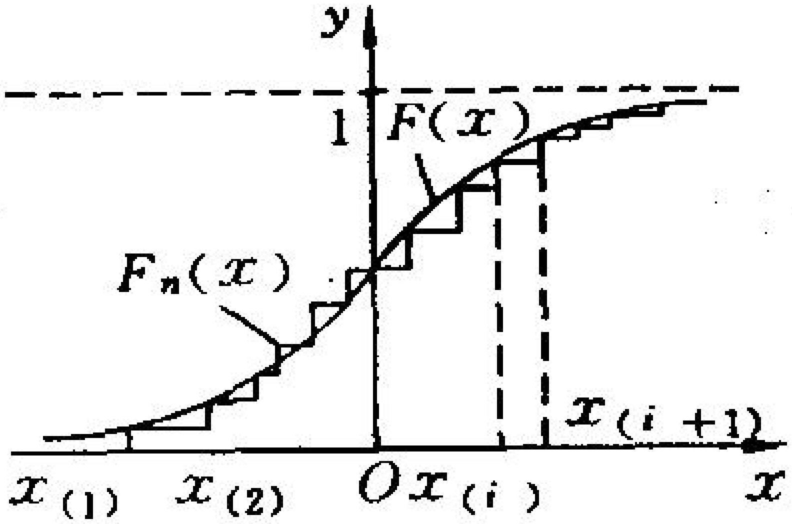
为总体X的样本分布函数.

样本分布函数$F_n(x)$的性质

1° $0 \le F_n(x) \le 1$;

2° $F_n(x)$是单调不减函数;

3° $F_n(x)$是处处右连续的,点$x_{(i)}$是$F_n(x)$跳跃间断点,跳跃度为频率f_i.



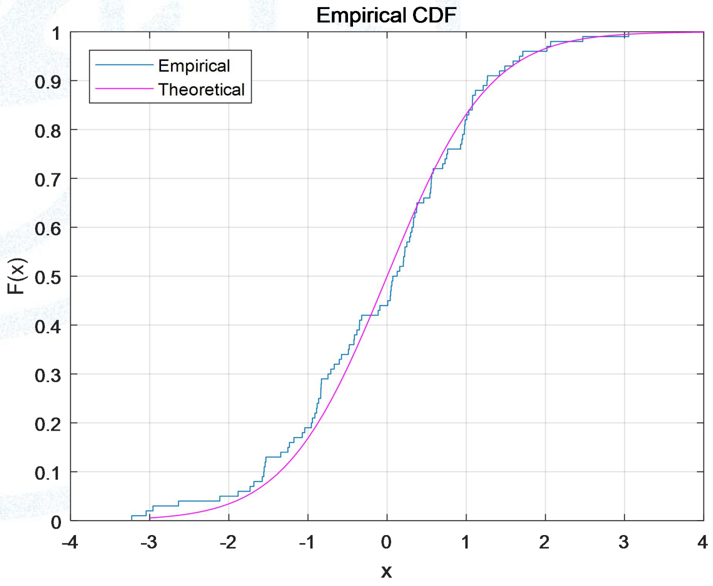
样本分布函数$F_n(x)$的图形呈跳跃上升的阶梯状,可以反映总体X的理论分布函数$F(x)$的图形.

定理 格利文科(W. Glivenko)定理

对于任一实数 x, 当 $n\to\infty$ 时, 样本分布函数$F_n(x)$ 以概率1 关于x 均匀收敛于总体分布函数 $F(x)$, 即

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

数理统计中依据样本推断总体特征的理论依据.



统计的一般步骤

总体

(理论分布)

选择个体

样本

观测样本

样本值 (数据)

数据处理

统计量

样本有关结论

推断

总体分布决定了样本取值的概率规律,也就是样本取到样本值的规律,因而可以由样本值去推断总体.样本是联系二者的桥梁.

§3 统计量及其分布

统计量

设$X_1,X_2,\dots,X_n$是来自总体 X 的样本,$x_1, x_2, \dots, x_n$ 是样本观测值,如果 $g(x_1, x_2,\dots, x_n)$ 为已知n元函数, 则称 $g(X_1, X_2, \dots, X_n)$为样本函数,它是一个随机变量, 称 $g(x_1, x_2,\dots, x_n)$为样本函数观测值.

称不含有任何未知参数的样本函数为统计量.

例 设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$的样本,其中μ, σ^2是未知参数,则

[1] $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ [2] $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ [3] $\min\{X_1, X_2, \cdots, X_n\}$,

是统计量

[4] $\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$ [5] $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ [6] $2\mu X_1, X_2, \cdots, X_n$.

不是统计量

几种常见统计量

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

反映了总体均值的信息

观察值为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

注 1.若$E(X)=\mu, D(X)=\sigma^2$,则$E(\bar{X})=\mu,\ D(\bar{X})=\frac{\sigma^2}{n}$.

2.由辛钦大数定律,$\bar{X}\xrightarrow{P}\mu$.

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

反映了总体方差和标准差的信息

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

观察值分别为 $$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

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

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

$E(S^2)=\sigma^2.$

样本k阶原点矩 $$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$$

反映了总体k阶矩的信息

样本k阶中心矩 $$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$$

k=1,2,…

反映了总体k阶中心矩的信息

观察值分别为

$$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\qquad b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k$$

注 1. $A_1=\overline{X}$,

2. $B_1=0$,

3. $B_2=\dfrac{n-1}{n}S^2, E(B_2)=\dfrac{n-1}{n}\sigma^2$,

4. 如果总体k阶矩$E(X^k)=\mu_k$存在,则$E(X_i^k)=\mu_k, i=1,2,\cdots,n.$

由辛钦定理,样本矩依概率收敛于总体同阶矩, 即

$$\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{P} \mu_k .$$

例　从一批机器零件毛坯中随机地抽取10件，测得其重量(公斤)：

230,　243,　185,　240,　215,

228,　196,　235,　200,　199

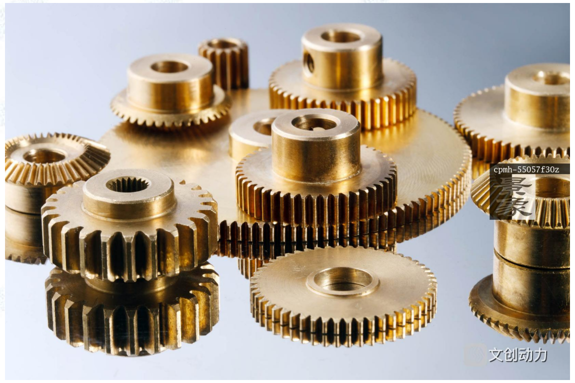
求这组样本值的均值、方差、二阶原点矩与二阶中心矩.

解 $$\overline{x}=\frac{1}{10}(230+243+185+240+215+228+196+235+200+199)=217.1$$

$$s^2=\frac{1}{9}\sum_{i=1}^{10}(x_i-\overline{x})^2=433.43$$

$$B_2=\frac{9}{10}s^2=\frac{1}{10}\sum_{i=1}^{10}(x_i-\overline{x})^2=390.0$$

$$A_2=\frac{1}{10}\sum_{i=1}^{10}x_i^2=47522.5$$



样本最大值 $X_{(n)}=\max(X_1,X_2,\cdots,X_n)$

样本最小值 $X_{(1)}=\min(X_1,X_2,\cdots,X_n)$

设总体X的分布函数为$F(x)$,则

$$F_{\max}(x)=P\{\max(X_1,X_2,\cdots,X_n)\le x\}=P\{X_1\le x,X_2\le x,\cdots,X_n\le x\}$$

$$=P\{X_1\le x\}P\{X_2\le x\}\cdots P\{X_n\le x\}=[F(x)]^n,$$

$$F_{\min}(x)=P\{\min(X_1,X_2,\cdots,X_n)\le x\}=1-P\{\min(X_1,X_2,\cdots,X_n)>x\}$$

$$=1-P\{X_1\ge x\}P\{X_2\ge x\}\cdots P\{X_n\ge x\}=1-[1-F(x)]^n.$$

例 设总体ξ服从区间$[0,\theta]$上的均匀分布,$(\xi_1,\xi_2,\cdots,\xi_n)$为总体$\xi$的样本,试求$\xi_{(1)}$和$\xi_{(n)}$的分布.

解 总体ξ的分布密度为

$$f(x)=\begin{cases}1/\theta, & 0\le x\le\theta\\ 0, & \text{其他}\end{cases}$$

ξ的分布函数为

$$F(x)=\begin{cases}0, & x<0\\ x/\theta, & 0\le x\le\theta\\ 1, & x>\theta\end{cases}$$

$$f_{\xi_{(n)}}(x)=\begin{cases}\dfrac{n}{\theta^n}x^{n-1}, & 0\le x\le\theta,\\ 0, & \text{其他.}\end{cases}$$

$$f_{\xi_{(1)}}(x)=\begin{cases}\dfrac{n}{\theta}(1-\dfrac{x}{\theta})^{n-1}, & 0\le x\le\theta,\\ 0, & \text{其他.}\end{cases}$$

例 设总体X的概率密度为

$$f(x)=\begin{cases}2\cos 2x, & 0\le x\le \pi/4\\ 0, & \text{其他}\end{cases}$$

从总体X的中抽取样本 $X_1,X_2,\cdots,X_n$，求样本容量n，使得

$$P\left\{\max(X_1,X_2,\cdots,X_n)>\frac{\pi}{12}\right\}\ge\frac{15}{16}$$

解 X的分布函数为$F(x)=\begin{cases}0, & x<0\\ \sin 2x, & 0\le x\le \pi/4\\ 1, & x>\pi/4\end{cases}$

$$\frac{15}{16}\le P\left\{\max(X_1,X_2,\cdots,X_n)>\frac{\pi}{12}\right\}=1-P\left\{\max(X_1,X_2,\cdots,X_n)\le\frac{\pi}{12}\right\}=1-\left[F\left(\frac{\pi}{12}\right)\right]^n=1-\left(\frac{1}{2}\right)^n$$

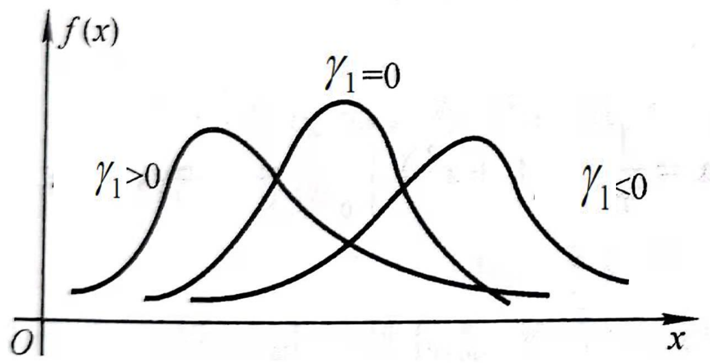
得$n\ge 4$.

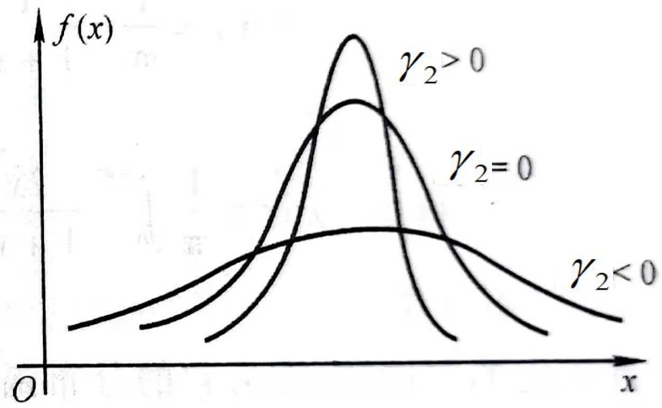
样本偏度 $$\gamma_1=\frac{b_3}{b_2^{\frac{3}{2}}}$$

反映了总体分布密度曲线的对称性信息

样本峰度 $$\gamma_2=\frac{b_4}{b_2^2}-3$$

反映了总体分布密度曲线在其峰值附近的陡峭程度





设$(X_1,Y_1),(X_2,Y_2),\cdots(X_n,Y_n)$是来自二维总体的样本,

样本协方差

$$S_{XY}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}).$$

样本相关系数

$$\rho_{XY}=\frac{S_{XY}^2}{S_X S_Y},$$

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

正态总体结论

(1) 单个正态总体

若$X \sim N(\mu,\sigma^2), X_1, X_2,\cdots, X_n$是来自总体的样本，

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

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

(2) 两个正态总体

若$X\sim N(\mu_1,\sigma_1^2)$, $Y\sim N(\mu_2,\sigma_2^2)$, 且$X$与$Y$相互独立,

$X_1,X_2,\cdots,X_{n_1}$, $Y_1,Y_2,\cdots,Y_{n_2}$分别为来自总体X和Y的样本,

则$\bar{X}\sim N(\mu_1,\dfrac{\sigma_1^2}{n_1})$, $\bar{Y}\sim N(\mu_2,\dfrac{\sigma_2^2}{n_2})$, $\bar{X}$与$\bar{Y}$相互独立,

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

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

例 在总体$N(52,\ 6.3^2)$中随机抽取一个容量为36的样本, 求样本均值$\overline{X}$落在50. 8到53. 8之间的概率.

解 $\overline{X} \sim N(52,\ 6.3^2/36)$

$$P\left\{50.8 < \overline{X} < 53.8\right\} = P\left\{\frac{50.8-52}{6.3/6} < \frac{\overline{X}-52}{6.3/6} < \frac{53.8-52}{6.3/6}\right\}$$

$$= \Phi\left(\frac{53.8-52}{6.3/6}\right) - \Phi\left(\frac{50.8-52}{6.3/6}\right)$$

$$= \Phi(1.7143) - \Phi(-1.1429) = 0.8239$$

例 某厂检验保温瓶的保暖性能, 在瓶中灌沸水, 24小时后测定其保温温度为T, 若已知 $T \sim N(62, 5^2)$, 求

(1) 随机抽取20只进行测定, 其样本均值低于60° 的概率;

(2) 若独立进行两次抽样测试, 各次分别抽取20只和12只, 那么两个样本平均值差的绝对值大于1° 的概率.

解 (1) $\overline{T} \sim N(62, 5^2/20)$, 即 $\overline{T} \sim N(62, 1.25)$

$$P\{\overline{T} < 60\} = P\left\{\frac{\overline{T} - 62}{1.12} < \frac{60 - 62}{1.12}\right\}$$

$$= \Phi(-1.79) = 0.0367$$

(2) $\overline{T}_1 \sim N(62, 5^2/20), \overline{T}_2 \sim N(62, 5^2/12)$,

则$\overline{T}_1 - \overline{T}_2 \sim N(62-62, \dfrac{5^2}{20} + \dfrac{5^2}{12})$,

即$\dfrac{\overline{T}_1 - \overline{T}_2}{\sqrt{25/20 + 25/12}} \sim N(0,1)$

$$P\left\{\left|\overline{T}_1 - \overline{T}_2\right| > 1\right\} = 1 - P\left\{-1 < \overline{T}_1 - \overline{T}_2 < 1\right\}$$

$$= 1 - P\left\{\frac{-1}{\sqrt{25/20+25/12}} < \frac{\overline{T}_1 - \overline{T}_2}{\sqrt{25/20+25/12}} < \frac{1}{\sqrt{25/20+25/12}}\right\}$$

$$= 2 - 2\Phi\left(\sqrt{3/10}\right) = 2 - 2\Phi\left(0.548\right) = 0.5824$$

例 设总体 $X \sim N(30,16)$, 从总体中抽取容量为n的样本, 要使 $P\{|\overline{X}-30|<1\} \geq 0.95$, 样本容量$n$至少应取多大.

解 $\dfrac{\overline{X}-30}{4/\sqrt{n}} \sim N(0,1)$

$$P\{|\overline{X}-30|<1\} = P\left\{\left|\frac{\overline{X}-30}{4/\sqrt{n}}\right| < \frac{1}{4/\sqrt{n}}\right\} = 2\Phi\left(0.25\sqrt{n}\right)-1$$

要使$P\{|\overline{X}-30|<1\} \geq 0.95$，则 $2\Phi\left(0.25\sqrt{n}\right)-1 \geq 0.95$

$$\Phi\left(0.25\sqrt{n}\right) \geq 0.975 = \Phi\left(1.96\right)$$

$0.25\sqrt{n} \geq 1.96$,即$n \geq 61.4656$,样本容量至少应取62 .

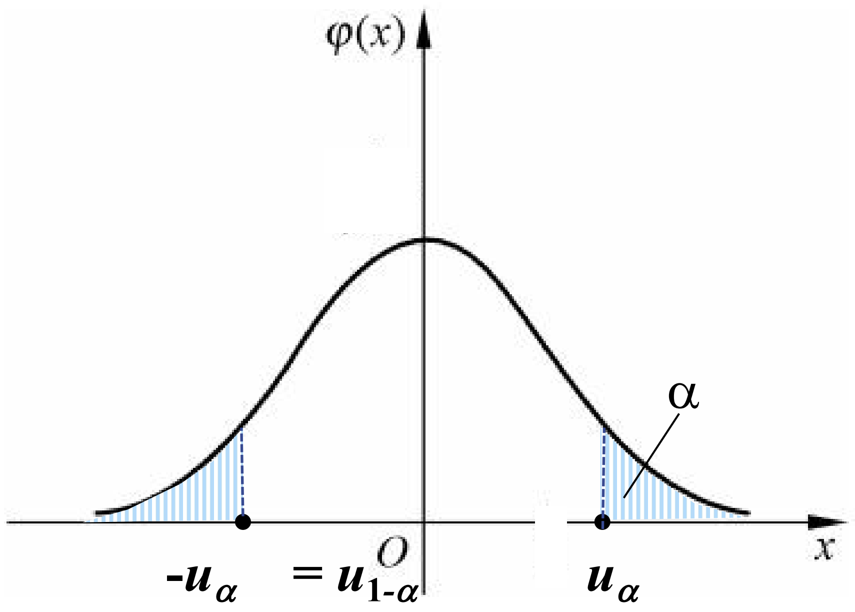
知识回顾 标准正态分布的上α分位点

设 $X \sim N(0,1)$, 对于给定的 $\alpha(0<\alpha<1)$. 若点 u_α 满足

$$P\{X \ge u_\alpha\} = \frac{1}{\sqrt{2\pi}} \int_{u_\alpha}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x = \alpha$$

则称点 u_α 为标准正态分布的上α分位点.

注 $\Phi(u_\alpha) = 1-\alpha, \quad -u_\alpha = u_{1-\alpha}.$



常用数据

$\Phi(1.96) = 0.975$

$u_{0.025} = 1.96$

$u_{0.05} = 1.645$

§4　常用统计量的分布

1.　χ^2分布

2.　t分布

3.　F分布

χ^2 分布

定义 设$X \sim N(0,1)$, $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, 称统计量

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

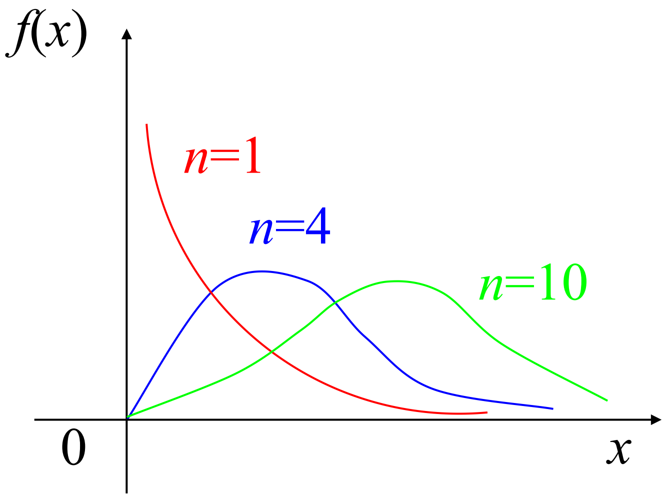
为**服从自由度为 n 的 χ^2分布**, 记作 $\chi^2 \sim \chi^2(n)$.

概率密度

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

其中$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} dx \ (t > 0)$

图像



χ^2分布的性质

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

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

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

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

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

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

性质2 设 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$, 且$X$与$Y$相互独立, 则

$$X+Y \sim \chi^2(n_1+n_2)$$ 可加性

性质3 设 $\chi^2 \sim \chi^2(n)$, 则对任意实数x,有

$$\lim_{n\to\infty} P\left\{\frac{\chi^2-n}{\sqrt{2n}} \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

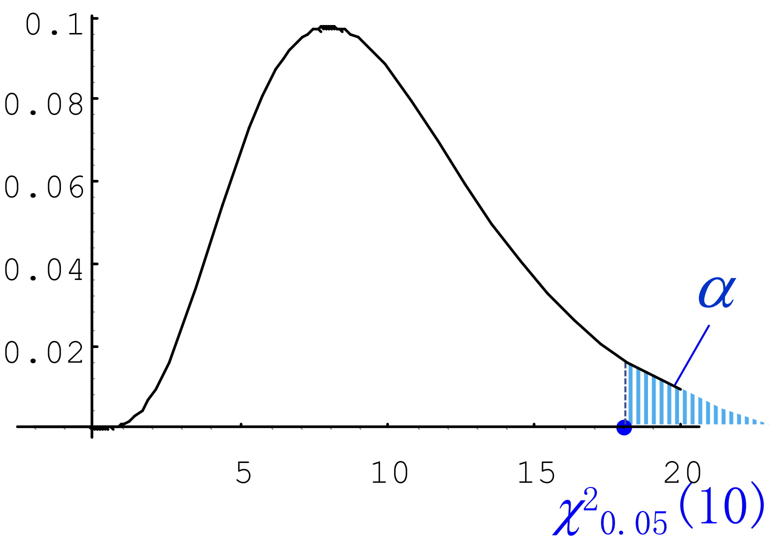
即 $\dfrac{\chi^2-n}{\sqrt{2n}} \xrightarrow{L} U \sim N(0,1)$

即n充分大时, 自由度为n 的χ^2分布近似服从$N(n,\ 2n)$.

定义 设 $\chi^2 \sim \chi^2(n)$，对于给定的正数 $\alpha(0<\alpha<1)$，称满足条件

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

的点 $\chi^2_\alpha(n)$ 为 $\chi^2(n)$ 分布的**上 α 分位点(数)**.



例如 取 $\alpha = 0.05, n = 7,$

查附表 $\chi^2_{0.05}(7) = 14.067$

$$P\{\chi^2(7) > 14.067\} = 0.05$$

定理 设$X \sim N(\mu,\sigma^2)$, $X_1, X_2, \cdots, X_n$为X的样本, 则

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

证明 $\dfrac{X_i-\mu}{\sigma} \sim N(0,1)$, 且相互独立,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$$

定理 设 $X\sim N(\mu,\sigma^2), X_1,X_2,\cdots,X_n$ 为X的样本, 样本均值和样本方差分别为$\overline{X}$ 和 S^2 ，则

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

即$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1);$$

(2) $\overline{X}$和S^2 相互独立.

注 $$D(S^2)=\frac{2\sigma^4}{n-1}$$

例 设$X_1,X_2,\ldots,X_{10}$ 为总体$N(0, 0.09)$的一个样本, 求

$$P\left\{\sum_{i=1}^{10} X_i^2 > 1.44\right\}.$$

解 $\dfrac{X_i}{0.3} \sim N(0,1),\quad i=1,2,\cdots,10.$

$$\chi^2 = \frac{1}{0.3^2}\sum_{i=1}^{10} X_i^2 \sim \chi^2(10)$$

$$P\left\{\sum_{i=1}^{10} X_i^2 > 1.44\right\} = P\left\{\frac{1}{0.3^2}\sum_{i=1}^{10} X_i^2 > \frac{1.44}{0.3^2}\right\} = P\{\chi^2 > 16\} = 0.10$$

例 设总体X服从$N\,(u,9)$,其中u是未知参数. 抽取容量为16的样本, 求样本方差小于16. 5的概率.

解 $$\chi^2=\frac{15}{9}S^2=\frac{5}{3}S^2\sim\chi^2(15)$$

$$P\left\{S^2<16.5\right\}=P\left\{\frac{5}{3}S^2<\frac{5}{3}\times 16.5\right\}=P\left\{\chi^2<27.5\right\}$$

由$\chi^2_{0.025}(15)=27.488$,

即$P\{\chi^2\geq 27.5\}\approx 0.025$,

$$P\left\{S^2<16.5\right\}\approx 1-0.025=0.975$$

例 设总体 $X \sim N(0,\sigma^2)$，样本$X_1, X_2, \ldots, X_6$, 设

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

求C,使CY服从χ^2分布, 并求自由度.

解 由 $X_i \sim N(0,\sigma^2)$, 且$X_1, X_2, \ldots, X_6$相互独立,

$$X_1+X_2+X_3 \sim N(0, 3\sigma^2) \qquad X_4+X_5+X_6 \sim N(0, 3\sigma^2)$$

$$Y_1 = \frac{X_1+X_2+X_3}{\sqrt{3}\sigma} \sim N(0,1) \qquad Y_2 = \frac{X_4+X_5+X_6}{\sqrt{3}\sigma} \sim N(0,1)$$

由独立性有 $Y_1^2+Y_2^2 \sim \chi^2(2)$，因此取$C = \dfrac{1}{3\sigma^2}$,

$$CY = Y_1^2+Y_2^2 \sim \chi^2(2)$$

自由度为2.

例 设总体 $X \sim N(0,4)$，样本 X_1, X_2, X_3, X_4，

则 $Y = \frac{1}{20}(X_1 - 2X_2)^2 + \frac{1}{100}(3X_3 - 4X_4)^2 \sim$ __$\chi^2(2)$__

t 分布

定义 设$X \sim N(0,1)$，$Y \sim \chi^2(n)$，且 X 与 Y 独立，则称随机变量

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

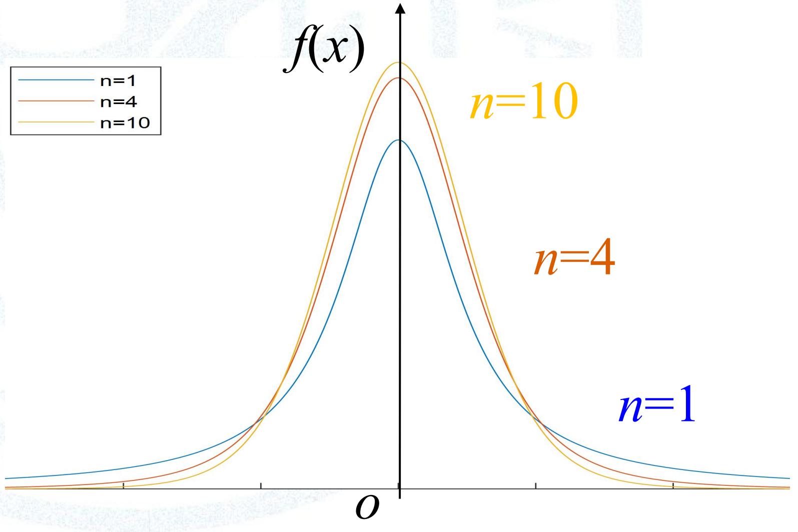
服从自由度为 n 的 **t 分布**，记作$t \sim t(n)$.

$t(n)$分布的概率密度

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

$$-\infty < x < +\infty$$

图像



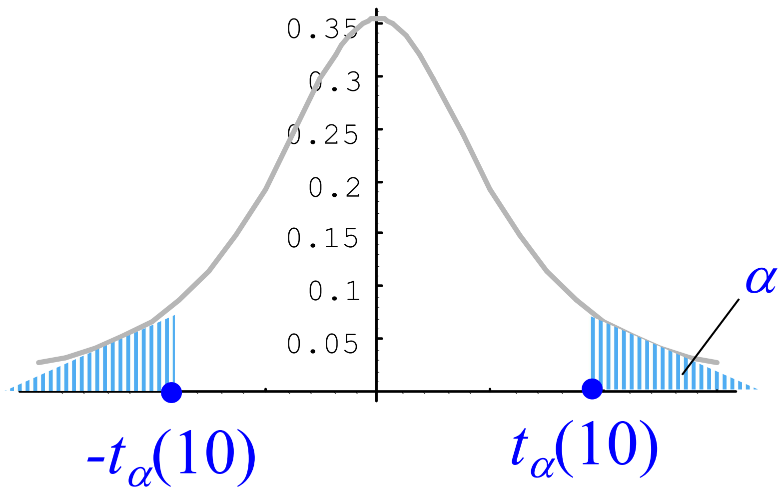
性质 $t(n)$分布的概率密度关于 y 轴对称, 且

$$\lim_{n\to\infty} f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < +\infty, \text{近似于} N(0,1).$$

定义 设 $t \sim t(n)$,对于给定正数 $\alpha\,(0<\alpha<1)$称满足条件

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

的点$t_\alpha(n)$为$t(n)$分布的上α分位点.



注 $-t_\alpha(n) = t_{1-\alpha}(n)$

n很大时 $t_\alpha(n) \approx u_\alpha$

例如 $t_{0.05}(10) = 1.8125$

$t_{0.95}(10) = -1.8125$

定理　设总体 $X \sim N(\mu, \sigma^2)$,$X_1, X_2, \cdots, X_n$为***X***的样本, 样本均值和样本方差分别为$\overline{X}$和 S^2 , 则随机变量

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

定理 正态总体 $X \sim N(\mu_1, \boxed{\sigma^2})$ 与 $Y \sim N(\mu_2, \boxed{\sigma^2})$ 相互独立,

样本容量依次为 n_1 和 n_2, 样本均值依次为 $\overline{X}$ 和 $\overline{Y}$, 样本方差依次为

S_1^2 和 S_2^2, 记

$$S_W = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

则随机变量

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

例 设样本$X_1, X_2, \ldots, X_n, X_{n+1}$来自总体$X \sim N(\mu, \sigma^2)$,

记$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, 证明

$$\sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\bar{X}}{S_n} \sim t(n-1)$$

证明 由$X_{n+1} - \bar{X} \sim N(0, \frac{n+1}{n}\sigma^2)$,$U = \dfrac{X_{n+1}-\bar{X}}{\sqrt{\frac{n+1}{n}\sigma^2}} \sim N(0,1)$,

$\chi^2 = \dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$ 又U与χ^2相互独立,

故 $\dfrac{U}{\sqrt{\chi^2/(n-1)}} \sim t(n-1)$.

F 分布

定义 设 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且 X 与 Y 独立, 则称随机变量

$$F = \frac{X / n_1}{Y / n_2}$$

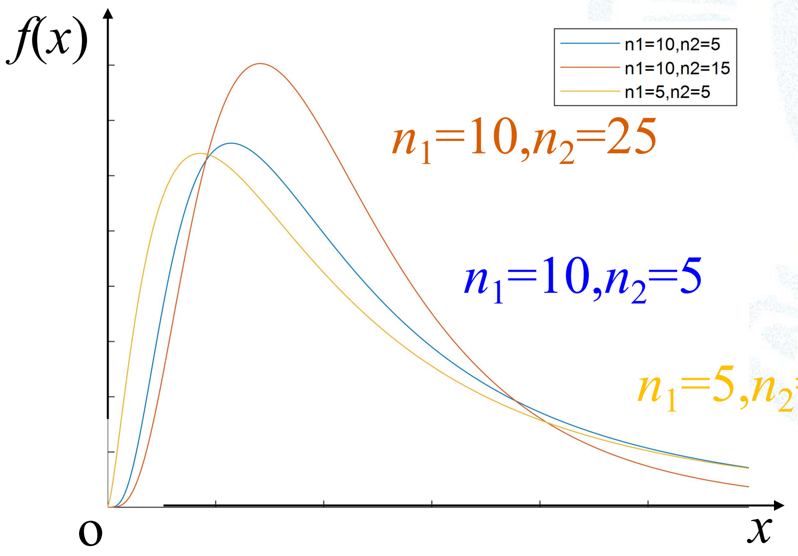
服从第一自由度n_1, 第二自由度n_2的F分布， 记作$F \sim F(n_1, n_2)$.

注 若$F \sim F(n_1, n_2)$, $\frac{1}{F} \sim F(n_2, n_1)$

$F(n_1,n_2)$分布的概率密度为

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

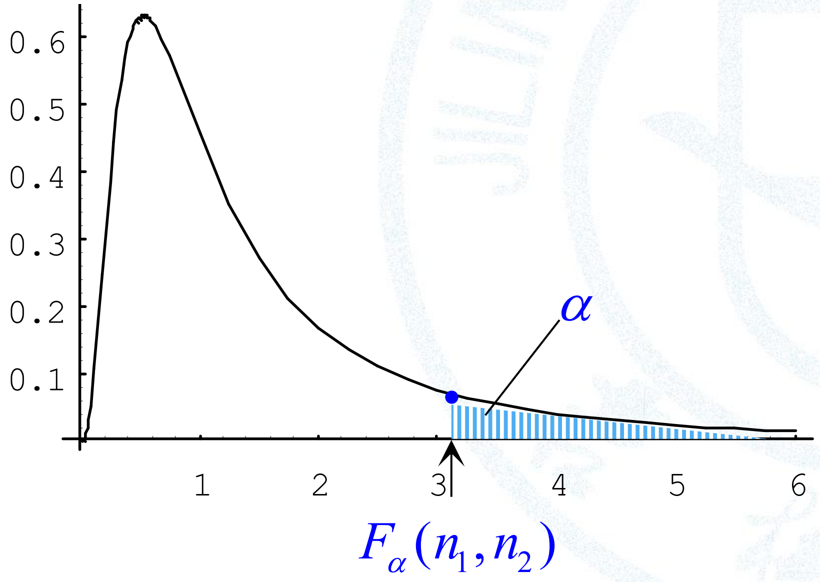
图像



定义 设$F \sim F(n_1, n_2)$,对于给定正数$\alpha(0<\alpha<1)$, 称满足条件

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

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



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

证明

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

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

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

如 $F_{0.025}(10,6)=5.46$ $F_{0.975}(6,10)=\dfrac{1}{F_{0.025}(10,6)}=0.183$

定理 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是来自正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$ 的相互独立的简单随机样本, 则

$$F = \frac{n_2}{n_1} \cdot \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum_{i=1}^{n_1} \left(X_i - \mu_1\right)^2}{\sum_{i=1}^{n_2} \left(Y_j - \mu_2\right)^2} \sim F\left(n_1, n_2\right)$$

定理 正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$相互独立, 样本容量依次为 n_1和 n_2, 样本均值依次为$\overline{X}$ 和 $\overline{Y}$, 样本方差依次为S_1^2和 S_2^2, 则随机变量

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

例 设$t \sim t(n)$, 证明$t^2 \sim F(1,n)$.

证明 设$t = \dfrac{U}{\sqrt{\chi^2 / n}}$,

$U \sim N(0,1)$, $\chi^2 \sim \chi^2(n-1)$, 且U与χ^2相互独立,

故 $t^2 = \dfrac{U^2}{\chi^2 / (n-1)} \sim F(1,n)$.

例 设二维随机变量$(X,Y)\sim N(0,1,4,9,0)$, 证明$F=\dfrac{9X^2}{4(Y-1)^2}\sim F(1,1)$.

证明 由$X\sim N(0,4), Y\sim N(1,9)$, X与Y独立,

$$\text{故}\ \frac{\left(\dfrac{X}{2}\right)^2}{\left(\dfrac{Y-1}{3}\right)^2}\sim F(1,1).$$

三大分布

χ^2 分布　设$X \sim N(0,1)$，　$X_1, X_2, \cdots, X_n$是X的样本，

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$$

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

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

F分布　设$X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$,且X与Y独立，

$$F = \frac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$$

$$1.\ u=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

$$2.\ t=\frac{\bar{X}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

$$3.\ \chi^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\sim\chi^2(n)$$

$$4.\ \frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

$$5.\ u=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^{\ 2}}{n_1}+\frac{\sigma_2^{\ 2}}{n_2}}}\sim N(0,1)$$

$$6.\ t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2).$$

$$其中\ S_W=\sqrt{\frac{(n_1-1)S_1^{\ 2}+(n_2-1)S_2^{\ 2}}{n_1+n_2-2}}$$

$$(\sigma_1^{\ 2}=\sigma_2^{\ 2}=\sigma^{\ 2})$$

$$7.\ F=\frac{n_2}{n_1}\cdot\frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_j-\mu_2)^2}\sim F(n_1,n_2)$$

$$8.\ F=\frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$$