第五章 抽样分布

5.1 总体与样本

（一）总体与个体

（二）样本

（一） 总体与个体

1. 总体

试验的全部可能的观察值称为总体.

2. 个体

总体中的每个可能观察值称为个体.

实例1 在研究2000名学生的年龄时，这些学生的年龄的全体就构成一个总体，每个学生的年龄就是个体.

3. 有限总体和无限总体

实例2 　某工厂10月份生产的灯泡寿命所组成的总体中，个体的总数就是10月份生产的灯泡数，这是一个有限总体;而该工厂生产的所有灯泡寿命所组成的总体是一个无限总体，它包括以往生产和今后生产的灯泡寿命.

有限总体包含的个体的总数很大时，可近似地将它看成是无限总体.

1. 样本的定义

设 X 是具有分布函数 F 的随机变量，若 $X_1, X_2, \cdots, X_n$ 是具有同一分布函数 F、相互独立的随机变量，则称 $X_1, X_2, \cdots, X_n$ 为从分布函数 F (或总体 F、或总体 X) 得到的容量为 n 的简单随机样本，简称样本.

它们的观察值 $x_1, x_2, \cdots, x_n$ 称为样本值，又称为 X 的 n 个独立的观察值.

2. 样本容量

样本中所含的样品数称为样本容量

5.2 样本函数与样本分布函数

（一）样本函数

定义5.1 样本函数

设$X_1,X_2,\ldots,X_n$是总体X的一个样本，$\phi=\phi(X_1,X_2,\ldots,X_n)$为一个连续的实值函数——样本函数.如果$\phi$中不包含任何未知参数,则称$\phi(X_1,X_2,\ldots,X_n)$为统计量.

设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本,可定义样本函数:

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

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

3. 样本均方差或样本标准差 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

4. 样本 k 阶原点矩 $\hat{A}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

5. 样本 k 阶中心矩 $\hat{B}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$.

（一） 样本函数

定义5.2 设 $X_1,X_2,\cdots,X_n$ 是取自总体 X 的样本,记 $x_1,x_2,\cdots,x_n$ 是样本的任一组观测值,将它们按由小到大的顺序重新排列为 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$.若 $X_{(k)}=x_{(k)}(k=1,2,\cdots,n)$,则称 $X_{(1)},X_{(2)},\cdots,X_{(n)}$为样本 $X_1,X_2,\cdots,X_n$ 的次序统计量,称 $X_{(i)}$为第 i 个次序统计量.特别地,$X_{(1)}=\min\{X_1,X_2,\cdots,X_n\}$和 $X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$分别称为最小次序统计量和最大次序统计量,它们的值分别称为极小值和极大值.

由于次序统计量 $X_{(k)}(k=1,2,\cdots,n)$是样本的函数,所以 $X_{(1)},X_{(2)},\cdots,X_{(n)}$一般不是相互独立的.

(二) 样本分布函数

1. 直方图与概率密度函数

作直方图的步骤

⑴从 n 个原始数据中,确定最大值和最小值,取 a 略小于最小值,b 略大于最大值.

⑵对数据进行整理分组.分组的个数,可根据实际的经验来确定,例如:数据在 50 个以下,可分为 5~6 个组;50~100 个数据,可分为 6~10 个组;100~250 个数据,可分成 7~12 个组;

⑶求出组距和组限.设分组的个数为 K,那么每个组的组距 $d=\frac{\square-\square}{\square}$,第一组应包含数据的最小值,最后一组应包含数据的最大值.

（二） 样本分布函数

1. 直方图与概率密度函数

⑷计算频数、频率、频率密度.每组包含的数据的个数,称为频数,记为 f_i,$i=1,2,\cdots,K$,频数除以数据的总数 n,即$\frac{\square\square}{\square}$,称为频率.把每一组至第 i组的频率累加,称为第 i组的累积频率,频率和组距 d 的比,称为频率密度.

⑸列出有关组距、频数、频率、频率密度等的统计表.

(二) 样本分布函数

1. 直方图与概率密度函数

⑹制作直方图.

建立平面直角坐标系 xOy,横坐标表示随机变量的取值,即数据的范围;纵坐标表示频率密度,即频率与组距的比值.这样,以每一组的组距 d 为底,以相应于这个小区间的频率密度为高,就得到一排竖直的小长方形,这样作出的图形,称为频率直方图,简称直方图.

1. 直方图与概率密度函数

例 1 某企业生产某种电子元件，因受到各种偶然因素的影响，其长度是有差异的，将其长度 X 看成随机变量，用直方图法分析 X 服从什么分布，抽样取得的 100 个数据如下：

（二）样本分布函数

1. 直方图与概率密度函数

1.36	1.49	1.43	1.41	1.37	1.40	1.32	1.42	1.47	1.39
1.41	1.36	1.40	1.34	1.42	1.42	1.45	1.35	1.42	1.39
1.44	1.42	1.39	1.42	1.42	1.30	1.34	1.42	1.37	1.36
1.37	1.34	1.37	1.37	1.44	1.45	1.32	1.48	1.40	1.45
1.39	1.46	1.39	1.53	1.36	1.48	1.40	1.39	1.38	1.40
1.36	1.45	1.50	1.43	1.38	1.43	1.41	1.48	1.39	1.45
1.37	1.37	1.39	1.45	1.31	1.41	1.44	1.44	1.42	1.47
1.35	1.36	1.39	1.40	1.38	1.35	1.38	1.43	1.42	1.42
1.42	1.40	1.41	1.37	1.46	1.36	1.37	1.27	1.37	1.38
1.42	1.34	1.43	1.42	1.41	1.41	1.44	1.48	1.55	1.37

（二） 样本分布函数

1. 直方图与概率密度函数

解 (1)先找出数据中的最大值 1.55,最小值 1.27,并取 a=1.265,b=1.565. a 略小于最小值,b 略大于最大值,使最大值、最小值在组内.

(2)对数据进行分组,可分成 10 组,即 K=10.

(3)找出组距:$\frac{□-□}{□}=\frac{1.565-1.265}{10}$=0.03.显然,第一组为[1.265,1.295],⋯,第十组为[1.535,1.565].

(4)计算频数、频率、频率密度等.

(5)列出有关组距、频数、频率、频率密度的分布表:

（二） 样本分布函数

1. 直方图与概率密度函数

组序号	分组组距	唱票统计	频数	频率	频率密度
1	1.265~1.295	一	1	0.01	0.33
2	1.295~1.325	止	4	0.04	1.33
3	1.325~1.355	正丅	7	0.07	2.33
4	1.355~1.385	正正正正丅	22	0.22	7.33
5	1.385~1.415	正正正正止	24	0.24	8.00
6	1.415~1.445	正正正正止	24	0.24	8.00
7	1.445~1.475	正正	10	0.10	3.33
8	1.475~1.505	正一	6	0.06	2.00
9	1.505~1.535	一	1	0.01	0.33
10	1.535~1.565	一	1	0.01	0.33
合计			100	1.00	

(二) 样本分布函数

1. 直方图与概率密度函数

(6)作直方图(见图 5—1).

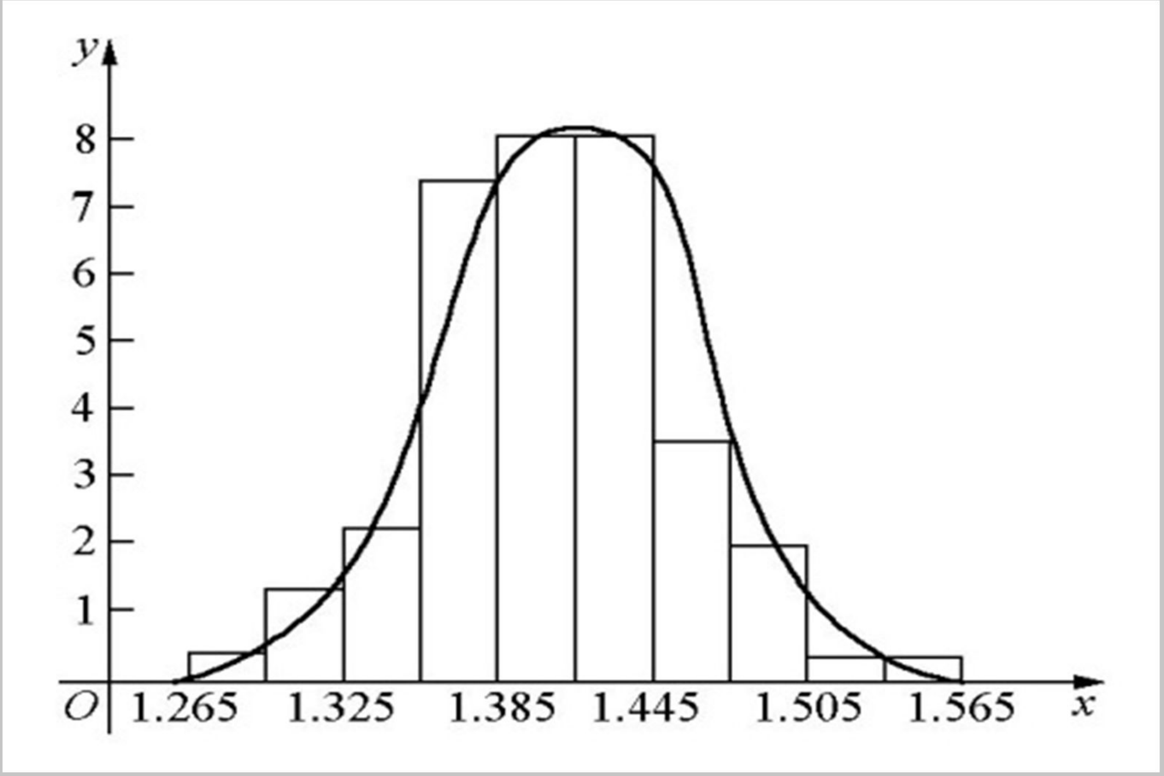


图 5—1

2. 样本分布函数

设总体 X 的 n 个样本值可以按大小次序排列成：

$$x_1 \le x_2 \le \cdots \le x_n,$$

如果 $x_k \leq x < x_{k+1}$，则不大于 x 的样本值的频率为 $\frac{k}{n}$. 因而函数

$$F_n(x) = \begin{cases} 0, & x < x_1, \\ \cdots \\ \frac{k}{n}, & x_k \le x < x_{k+1}, \quad (k = 1, 2, \cdots, n-1) \\ \cdots \\ 1, & x \ge x_n, \end{cases}$$

与事件 $\{X \leq x\}$ 在 n 次重复独立试验中的频率是相同的. 我们称 $F_n(x)$ 为**样本分布函数**或经验分布函数.

(二) 样本分布函数

2. 样本分布函数

设总体 X 的 n 个样本值可以按大小次序排列成:

$$\square_1 \le \square_2 \le \cdots \le \square_{\square},$$

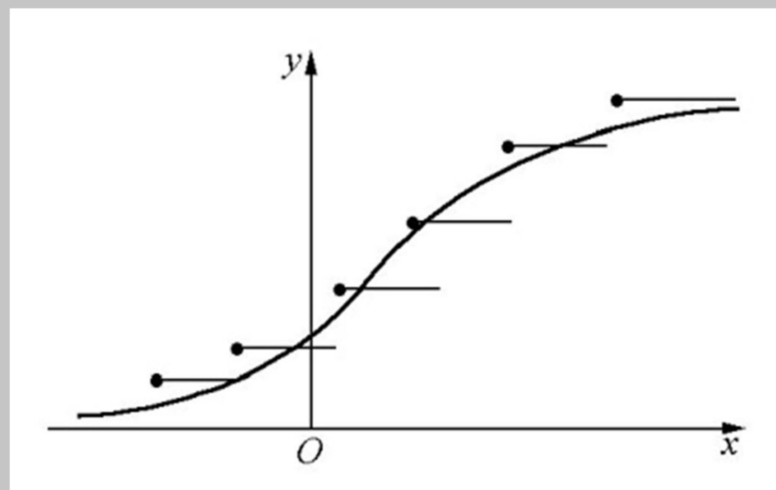


图 5—2

在图 5—2 中,阶梯形的曲线是经验分布函数,而光滑曲线是总体 X 的分布函数 F(x)的近似曲线.

（二） 样本分布函数

2. 样本分布函数

例 2 已知样本值:6.60,4.60,5.40,5.80,5.40,试构造出它们的经验分布函数.

解 先将这一组数据从小到大重新排列:4.60,5.40,5.40,5.80,6.60,则经验分布函数为:

$$F_5(x)=\begin{cases}0, & x<4.60,\\ \dfrac{1}{5}, & 4.60\le x<5.40,\\ \dfrac{3}{5}, & 5.40\le x<5.80,\\ \dfrac{4}{5}, & 5.80\le x<6.60,\\ 1, & 6.60\le x.\end{cases}$$

从上面经验分布函数的概念和例题,我们看到样本量 n 越大,样本取值的分布就越能近似地反映总体 X 的分布.

5.3 抽样分布

（一）几个常用的分布

（二）抽样分布的分位点

（三）正态总体的抽样分布

统计量的分布称为抽样分布.

1. χ^2 分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0,1)$ 的样本，则称统计量 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 服从自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$.

自由度：

指 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 中右端包含独立变量的个数.

(一) 几个常用的分布

$\chi^2(n)$ 分布的概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}y^{\frac{n}{2}-1}\mathrm{e}^{-\frac{y}{2}}, & y>0\\ 0 & \text{其他.}\end{cases}$$

证明　因为 $\chi^2(1)$ 分布即为 $\Gamma\left(\dfrac{1}{2},2\right)$ 分布,

又因为 $X_i\sim N(0,1)$, 由定义 $X_i^2\sim\chi^2(1)$,

即 $X_i^2\sim\Gamma\left(\dfrac{1}{2},2\right)$, $i=1,2,\cdots,n.$

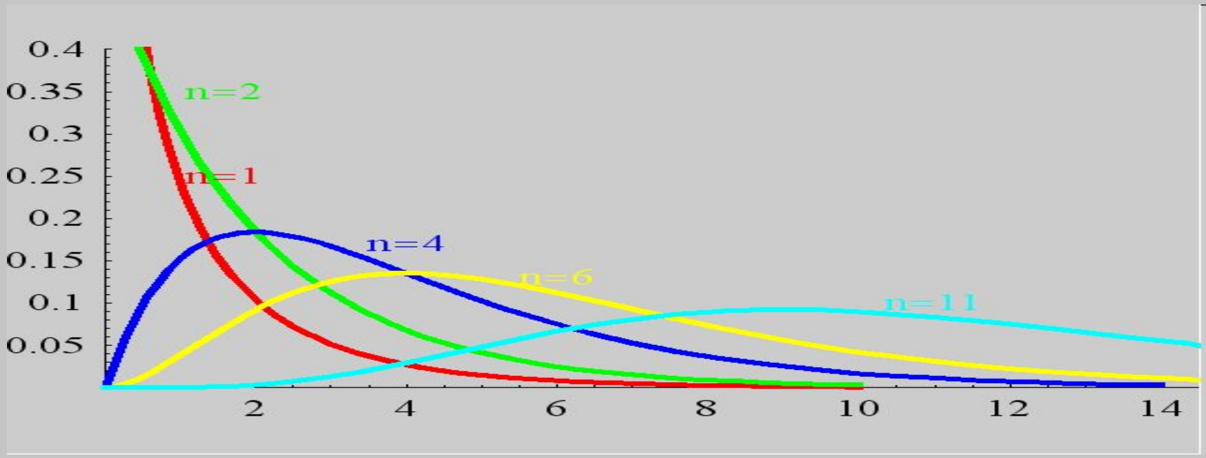
(一) 几个常用的分布

因为 $X_1, X_2, \cdots, X_n$ 相互独立，

所以 $X_1^2, X_2^2, \cdots, X_n^2$ 也相互独立，

根据Γ分布的可加性知 $\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \Gamma\left(\frac{n}{2}, 2\right).$

$\chi^2(n)$分布的概率密度曲线如图.



χ^2 分布的性质

性质1 （χ^2 分布的可加性）

设 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$, 并且 χ_1^2, χ_2^2 独立, 则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$.

（此性质可以推广到多个随机变量的情形.）

设 $\chi_i^2 \sim \chi^2(n_i)$, 并且 χ_i^2 $(i = 1, 2, \cdots, m)$ 相互独立, 则 $\sum_{i=1}^{m} \chi_i^2 \sim \chi^2(n_1 + n_2 + \cdots + n_m)$.

性质2 (χ^2分布的数学期望和方差)

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

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

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

$$= 3 - 2 = 1, \qquad i = 1, 2, \cdots, n.$$

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

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

（一） 几个常用的分布

例 1 设随机变量 $Y_1\sim\chi^2(m)$,$Y_2\sim\chi^2(n)$,且 Y_1 与 Y_2 相互独立,则
$Y_1+Y_2\sim\chi^2(m+n)$.

证明 根据 χ^2 分布的定义及 Y_1 与 Y_2 相互独立,我们可以把 Y_1 和 Y_2 分别表示为

$$Y_1 = X_1^2 + X_2^2 + \dots + X_m^2,$$

$$Y_2 = X_{m+1}^2 + X_{m+2}^2 + \dots + X_{m+n}^2,$$

其中 $X_1,X_2,\cdots,X_m,X_{m+1},\cdots,X_{m+n}$ 相互独立且均服从分布 $N(0,1)$,于是

$$Y_1 + Y_2 = X_1^2 + X_2^2 + \dots + X_{m+n}^2.$$

再由 χ^2 分布的定义知,$Y_1+Y_2\sim\chi^2(m+n)$.

（一） 几个常用的分布

例 2 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $X\sim N(\mu,\sigma^2)$的样本.求随机变量 $Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 的概率分布.

解 因为 $X_1,X_2,\cdots,X_n$ 相互独立,且 $X_i\sim N(\mu,\sigma^2)(i=1,2,\cdots,n)$,故令

$$Y_i=\frac{X_i-\mu}{\sigma},\quad i=1,2,\ldots,n,$$

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

根据定义可知, $Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}Y_i^2$ 服从自由度为 n 的 χ^2 分布.

2. t 分布

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

$t(n)$ 分布的概率密度函数为

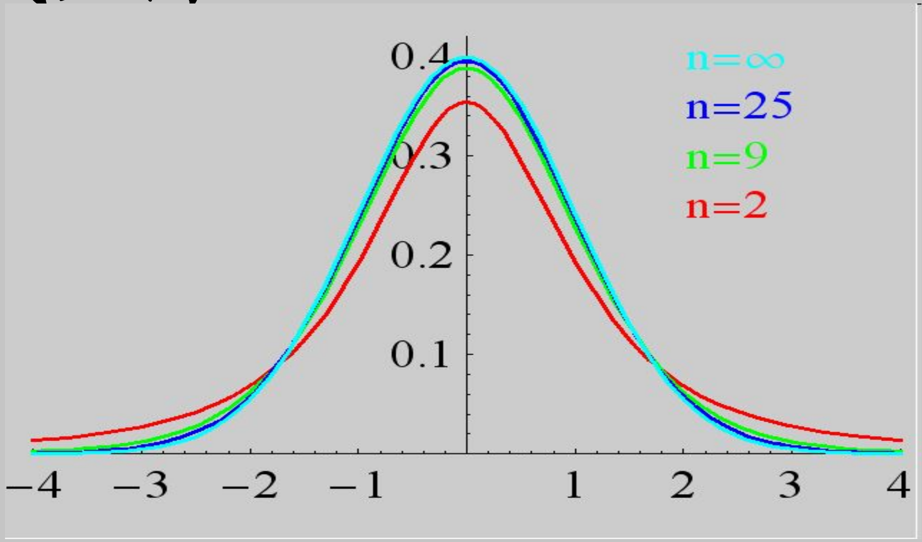
$$h(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < +\infty$$

（一） 几个常用的分布

t 分布的概率密度曲线如图

显然图形是关于

$t = 0$对称的.



当 n 充分大时，其图形类似于标准正态变量概率密度的图形.

因为 $\lim\limits_{n\to\infty} h(t) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$,

所以当 n 足够大时 t 分布近似于 $N(0,1)$ 分布,

但对于较小的n，t分布与$N(0,1)$分布相差很大.

（一） 几个常用的分布

例 3 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $X \sim N(0,4)$ 的样本. 试问:统计量 $\dfrac{\sqrt{n-1}\,X_1}{\sqrt{\sum_{i=2}^{n} X_i^2}}$ 服从什么分布?

解 因为 $X_i \sim N(0,4)\ (i=1,2,\cdots,n)$，所以 $\dfrac{\sum_{i=2}^{n} X_i^2}{4} \sim \chi^2(n-1)$，$\dfrac{X_1}{2} \sim N(0,1)$ 且它们相互独立. 根据 t 分布的定义，有

$$\frac{\sqrt{n-1}\,X_1}{\sqrt{\sum_{i=2}^{n} X_i^2}} = \frac{X_1/2}{\sqrt{\dfrac{\sum_{i=2}^{n} X_i^2}{4}\Big/(n-1)}} \sim t(n-1).$$

3. F 分布

设 $U \sim \chi^2(n_1)$，$V \sim \chi^2(n_2)$，且 U, V 独立，则称随机变量 $F = \dfrac{U/n_1}{V/n_2}$ 服从自由度为 (n_1, n_2) 的 F 分布，记为 $F \sim F(n_1, n_2)$.

$F(n_1, n_2)$分布的概率密度为

$$\psi(y)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} y^{\frac{n_1}{2}-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left[1+\left(\dfrac{n_1 y}{n_2}\right)\right]^{\frac{n_1+n_2}{2}}}, & y>0, \\ 0, & \text{其他}.\end{cases}$$

（一） 几个常用的分布

例 4 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $X\sim N(0,1)$ 的样本.试问:统计量 $\dfrac{(n-3)\sum_{i=1}^{3}X_i^2}{3\sum_{i=4}^{n}X_i^2}$ 服从什么分布?

解 因为 $\sum_{i=1}^{3}X_i^2\sim\chi^2(3)$, $\sum_{i=4}^{n}X_i^2\sim\chi^2(n-3)$,且二者相互独立,所以

$$\frac{(n-3)\sum_{i=1}^{3}X_i^2}{3\sum_{i=4}^{n}X_i^2}=\frac{\sum_{i=1}^{3}X_i^2/3}{\sum_{i=4}^{n}X_i^2/(n-3)}\sim F(3,n-3).$$

（二） 抽样分布的分位点

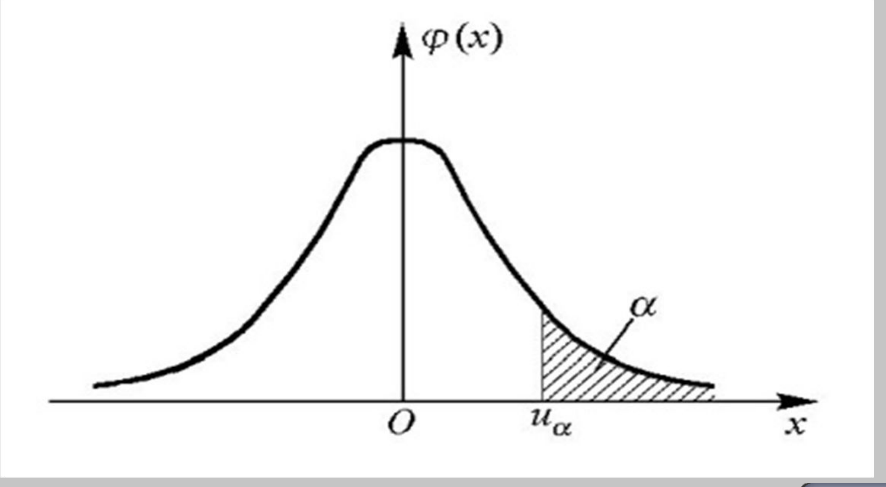
定义4.4 四种常见抽样分布的分位点概念

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

$$P(X > u_\alpha) = \int_{u_\alpha}^{+\infty} \varphi(x)\,dx = \alpha,$$

即 $\Phi(u_\alpha) = 1 - P(X > u_\alpha) = 1-\alpha$，

则称 u_α 为 X 关于 α 的上侧分位点（见图 5—6）.



例 5 公共汽车车门的高度是按男子与车门碰头的概率在 0.01 以下来设计的,设男子身高(单位:cm)$X \sim N(170,36)$,问:车门高度应如何确定?

解 设车门高度为 h cm 时,男子与车门碰头概率是 0.01,则

$$P(X > h) = 1 - \Phi\left(\frac{h-170}{6}\right) = 0.01,$$

化简为 $\Phi\left(\frac{h-170}{6}\right)=1-0.01=0.99$,

这里 $\frac{h-170}{6}$ 也就是标准正态随机变量的上侧分位点 $u_{0.01}$.

查附表 2 得 $\frac{h-170}{6}=u_{0.01}=2.33$,所以 $h=170+13.98\approx184$.因此要使男子与车门碰头几率在 0.01 以下,车门高度至少为 184cm .

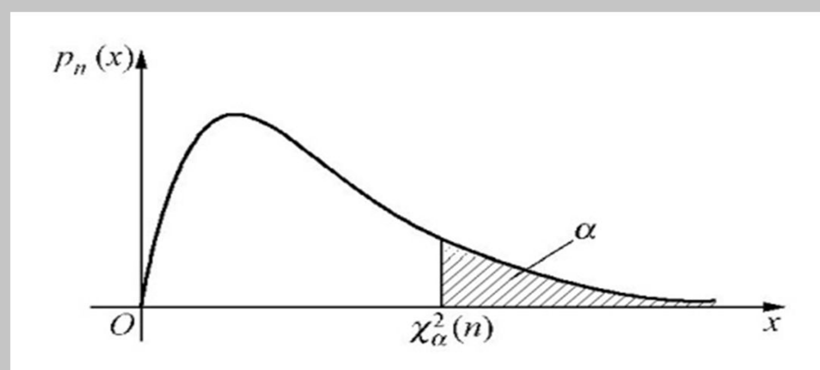
（二） 抽样分布的分位点

定义4.4 **四种常见抽样分布的分位点概念**

(2) 设 $X \sim \chi^2(n)$，对于给定的 $0<\alpha<1$，存在 $\chi^2_\alpha(n)$ 满足

$$P(\chi^2 > \chi^2_\alpha(n)) = \int_{\chi^2_\alpha(n)}^{+\infty} p_n(x)\,dx = \alpha,$$

其中，$p_n(x)$ 为 χ^2 的概率密度，则称点 $\chi^2_\alpha(n)$ 为 χ^2 分布关于 α 的上侧分位点，如图 5—7 所示.

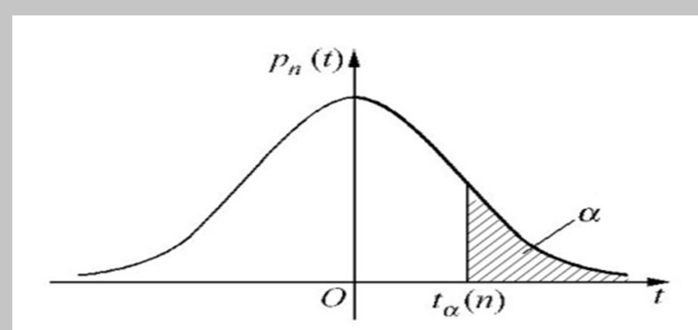


定义4.4 四种常见抽样分布的分位点概念

(3)设 $X \sim t(n)$,对于给定的 $0<\alpha<1$,存在 $t_\alpha(n)$满足

$$P(T > t_\alpha(n)) = \int_{t_\alpha(n)}^{+\infty} p_n(t)\,dt = \alpha,$$

其中,$p_n(t)$为 T 的概率密度,则称点 $t_\alpha(n)$为 t分布关于 α的上侧分位点,如图所示.



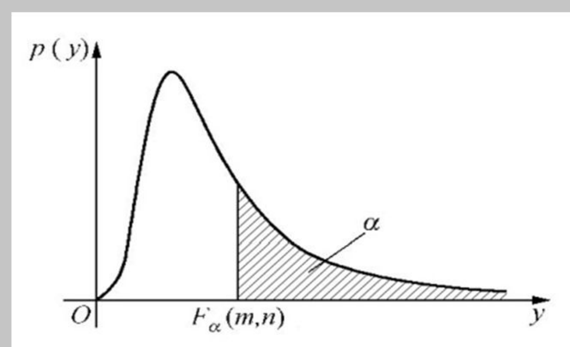
(二) 抽样分布的分位点

定义4.4 **四种常见抽样分布的分位点概念**

(4)设 $X \sim F(m,n)$,对于给定的 $0<\alpha<1$,存在 $F_{\alpha}(m,n)$满足

$$P(F > F_{\alpha}(m,n)) = \int_{F_{\alpha}(m,n)}^{+\infty} p(y)\,dy = \alpha,$$

其中,$p(y)$为 F 的概率密度,则称点 $F_{\alpha}(m,n)$为 F 分布关于 α的上侧分位点,如图所示.



（二）抽样分布的分位点

正态分布，χ^2分布、t分布及F分布的分位点有如下一些性质：

(1) $u_{1-\alpha}=-u_{\alpha}$;

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

(3) $F_{\alpha}(m,n)=\dfrac{1}{F_{1-\alpha}(n,m)}$;

(4)当 n 较大(n>45)时,有

$$\chi_{\alpha}^{2}(n)\approx\frac{1}{2}\left(u_{\alpha}+\sqrt{2n-1}\right)^{2},\quad t_{\alpha}(n)\approx u_{\alpha},$$

这里 u_{α}是标准正态分布 N (0,1)关于 α的上侧分位点.

定理5.1 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的样本，则

(1) $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$;

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

(3) $\bar{X}$ 与 S^2 相互独立;

(4) $\dfrac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1)$.

这里 $\bar{X}$ 为样本均值，S^2 为样本方差，即

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

（三） 正态总体的抽样分布

定理5.2 设 $X_1, X_2, \cdots, X_m$ 是来自正态总体 $X \sim N(\mu_1, \sigma^2)$ 的样本，$Y_1, Y_2, \cdots, Y_n$ 是来自正态总体 $Y \sim N(\mu_2, \sigma^2)$ 的样本，且 X 与 Y 相互独立，则

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_\omega\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} \sim t(m+n-2),$$

其中 $\bar{X}=\frac{1}{m}\sum_{i=1}^{m} X_i$，$S_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2$，$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n} Y_i$，

$$S_2^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2,\quad S_\omega^2=\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}=\frac{m-1}{m+n-2}S_1^2+\frac{n-1}{m+n-2}S_2^2$$

（三） 正态总体的抽样分布

定理5.3 设 $X_1,X_2,\cdots,X_m$ 是来自正态总体 $X\sim N(\mu_1,\sigma_1^2)$ 的样本，$Y_1,Y_2,\cdots,Y_n$ 是来自正态总体 $Y\sim N(\mu_2,\sigma_2^2)$ 的样本，且 X 与 Y 相互独立，则

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(m-1,n-1),$$

其中

$$S_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,\quad S_2^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2.$$