第二节　抽样分布

一、基本概念

二、常见分布

三、小结

一、基本概念

1. 统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量.

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值，则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

【例1】 设X_1, X_2, X_3是来自总体$N(\mu, \sigma^2)$的一个样本, 其中μ为已知, σ^2为未知, 判断下列各式哪些是统计量, 哪些不是?

$T_1 = X_1$, $T_2 = X_1 + X_2 e^{X_3}$,

$T_3 = \frac{1}{3}(X_1 + X_2 + X_3)$,

$T_4 = \max(X_1, X_2, X_3)$, $T_5 = X_1 + X_2 - 2\mu$,

是

$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2)$.

不是

2. 几个常用统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值.

(1)样本平均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;

其观察值 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2)样本方差

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

其观察值

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

(3)样本标准差

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

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

(4) 样本 k 阶(原点)矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \ , k = 1, 2, \cdots;$

其观察值 $\alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \ , k = 1, 2, \cdots .$

(5)样本 k 阶中心矩

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

其观察值 $b_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k \ , k = 2, 3, \cdots .$

(6) 顺序统计量:

设$(X_1, X_2, \cdots, X_n)$是来自总体 X 的一个样本，每当样本得到一组观察值 $x_1, x_2, \cdots, x_n$，将其按从小到大的次序排列为$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

第k个值$x_{(k)}$作为$X_{(k)}$的观察值,

则$X_{(k)}$ (k=1,2,…,n)均为统计量, 统称为顺序统计量

$X_{(1)}$ 为最小项统计量; $X_{(n)}$为最大项统计量

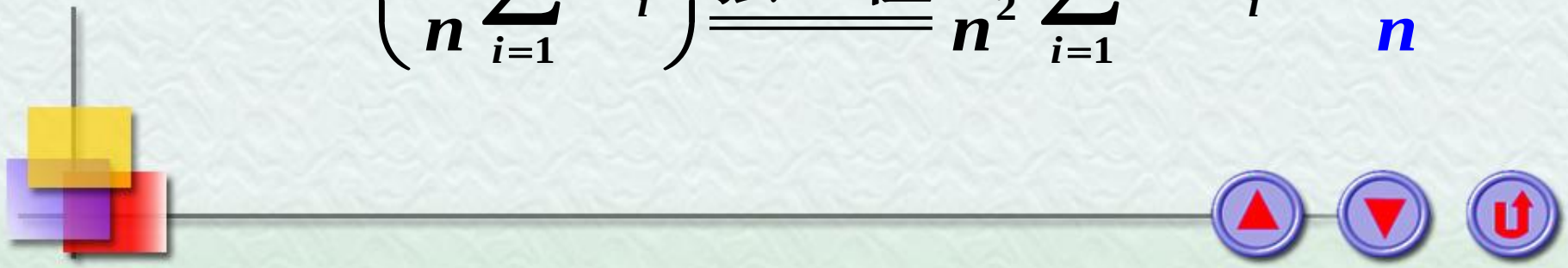
3. $\bar{X}$ 和 S_0^2 的数字特征

设总体X的期望$EX=\mu$,方差$DX=\sigma^2$

则 1)

$$E\bar{X}=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}EX_i=\mu$$

$$D\bar{X}=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\underline{\underline{\text{独立性}}}\frac{1}{n^2}\sum_{i=1}^{n}DX_i=\frac{\sigma^2}{n}$$



上式表明，样本均值 $\bar{X}$ 有这样的性质：其观测值以总体期望 μ 为中心，波动方差 σ^2/n。也就是说，样本容量越大，**$\bar{X}$ 的方差就越小，就越向总体期望 μ 集中。所以，我们用 $\bar{X}$ 的观测值来估计总体期望 μ** 是合理的，$\bar{X}$ 把样本中关于 μ 的信息提取出来了。那么，还会不会有比它更好的统计量能更有效地提取 μ 的信息呢？或者说，$\bar{X}$ 是否已经充分地提取了样本中关于 μ 的信息呢？我们在后面会讨论这个问题。

2) $$ES_0^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{n-1}{n}\sigma^2$$

令 $$S^2 = \frac{n}{n-1}S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

称为修正的样本方差 $(\because ES^2 = \sigma^2)$

$S = \sqrt{S^2}$ 称为修正的样本标准差

证明：$ES_0^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$

$$\because \quad S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) =$$

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

$$\therefore \quad ES_0^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2\right) = EX_i^2 - E\bar{X}^2 =$$

$$DX_i + (EX_i)^2 - \left(D\bar{X} + (E\bar{X})^2\right)$$

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

【例2】设X_1，X_2，…，X_n是来自总体X的样本，$X\sim N(\mu,\ \sigma^2)$，其中μ、σ^2为未知参数，则

X_1，　$\frac{1}{2}X_1+\frac{1}{3}X_2$，　$\min\{X_1,\ X_2,\ \ldots,\ X_n\}$

均为统计量，

但诸如　$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$，　$\frac{X_1}{\sigma}$

等均不是统计量，因它含有未知参数μ或σ.

【例3】 从一批机器零件毛坯中随机地抽取10件，测得其重量为(单位: 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 令 $(x_1, x_2, \cdots, x_{10})$

$$= (230, 243, 185, 240, 215, 228, 196, 235, 200, 199)$$

则 $$\bar{x} = \frac{1}{10}(230 + 243 + 185 + 240 + 215$$

$$+228 + 196 + 235 + 200 + 199)$$

$$= 217.10$$

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

$$a_2 = \frac{1}{10}\sum_{i=1}^{10} x_i^{\,2} = 47522.5$$

$$b_2 = \frac{9}{10}s^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = 390.0$$

4. 经验分布函数

如前所述，数理统计所研究的实际问题（总体）的分布一般来说是未知的，需要通过样本来推断．但如果对总体一无所知，那么，做出推断的可信度一般也极为有限．在很多情况下，我们往往可以通过具体的应用背景或以往的经验，再通过观察样本观测值的分布情况，对总体的分布形式有个大致了解．观察样本观测值的分布规律，了解总体X的概率密度和分布函数，常用直方图和经验分布函数．

总体分布函数 $F(x)$ 相应的统计量称为经验分布函数.

经验分布函数的做法如下:

设 $X_1, X_2, \cdots, X_n$ 是总体 F 的一个样本,

用 $S(x)(-\infty < x < +\infty)$ 表示 $X_1, X_2, \cdots, X_n$ 中不大于 x 的随机变量的个数,

定义经验分布函数 $F_n(x)$ 为

$$F_n(x) = \frac{1}{n} S(x), \quad (-\infty < x < +\infty)$$

为了解总体X的分布形式，根据样本观测值x_1，x_2，…，x_n构造一个函数$F_n(x)$来近似总体X的分布函数，函数$F_n(x)$称为经验分布函数．它的构造方法是这样的，将样本观测值x_1，x_2，…，x_n按从小到大可排成$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，定义

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

$$F_n(x)=\frac{\text{样本中小于 } x \text{ 的观测值的个数}}{n}, \quad \forall x \in R.$$

$F_n(x)$ 只在 $x = x_{(k)}$，（$k = 1, 2, \ldots, n$）处有跃度为 $1/n$ 的间断点，若有 l 个观测值相同，则$F_n(x)$在此观测值处的跃度为 l/n. 对于固定的 x，$F_n(x)$即表示事件$\{X \leq x\}$在n次试验中出现的频率，即 $F_n(x) = \frac{k}{n}$，其中k为落在$(-\infty, x)$中x_i的个数.

对于一个样本值，$F_n(x)$的观察值容易求得．

($F_n(x)$的观察值仍以 $F_n(x)$ 表示.)

【例4】 设总体 F 具有一个样本值 1, 2, 3,

则经验分布函数 $F_3(x)$的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \frac{1}{3}, & 1\le x<2,\\ \frac{2}{3}, & 2\le x<3,\\ 1, & x\ge 3.\end{cases}$$

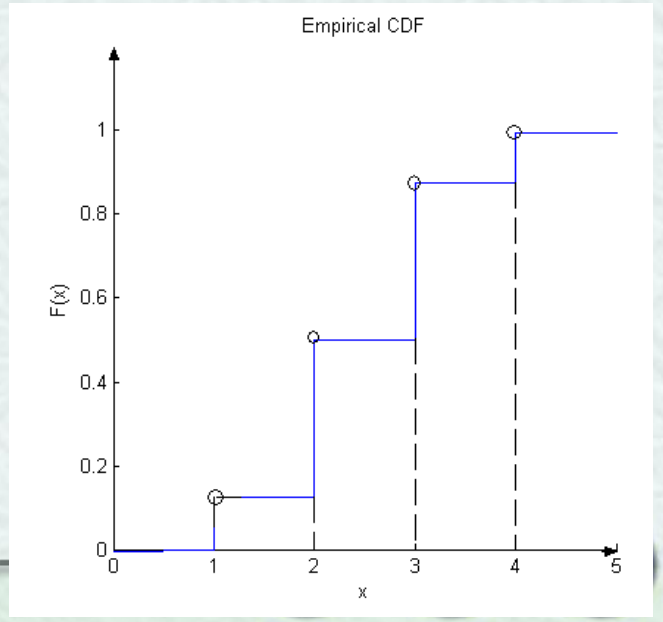
【例5】 设总体 F 具有一个样本值 $1, 1, 2,$

则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{2}{3}, & 1\le x<2,\\ 1, & x\ge 2.\end{cases}$$

【例6】 总体X，样本观察值1，2，2，2，3，3，3，4，则经验分布函数为

$$F_8(x)=\begin{cases}0, & x\le 1;\\ 1/8, & 1<x\le 2;\\ 4/8, & 2<x\le 3;\\ 7/8, & 3<x\le 4;\\ 1, & x>4;\end{cases}$$



关于经验分布函数，我们要注意以下几点：

经验分布函数是利用样本得到的，而样本是随机向量，所以经验分布函数也是随机的。同一个总体，即使是在相同的样本容量下，不同的样本也会给出不同的经验分布函数；

关于经验分布函数，我们要注意一下几点：

对于给定的 x，$F_n(x)$ 是一个随机变量，是事件 $\{X<x\}$ 在 n 重贝努里试验中发生的频率；

给定样本值后，经验分布函数就成为一个普通的跳跃函数，而且恰好是一个离散型随机变量的分布函数，该离散型随机变量的分布列为

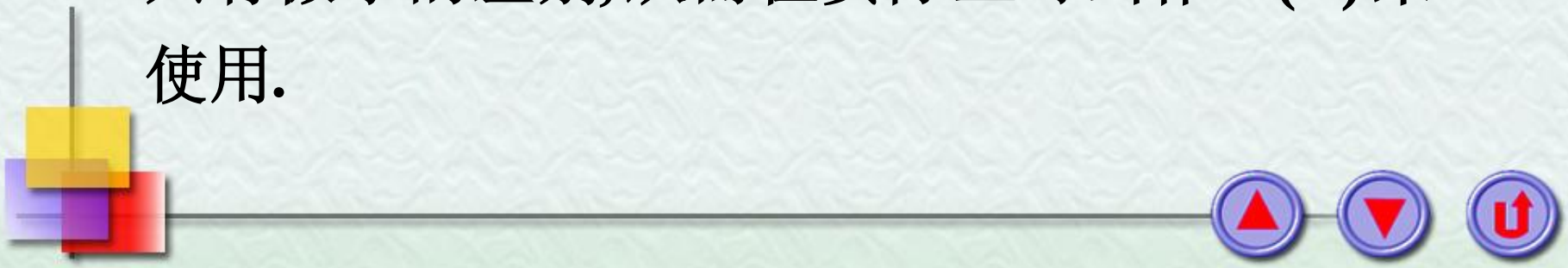
$$P_i = 1/n\ ,\ i = 1\ ,\ 2\ ,\ \ldots\ ,\ n$$

格里汶科定理

对于任一实数 x,当 $n \to \infty$ 时,$F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$,即

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

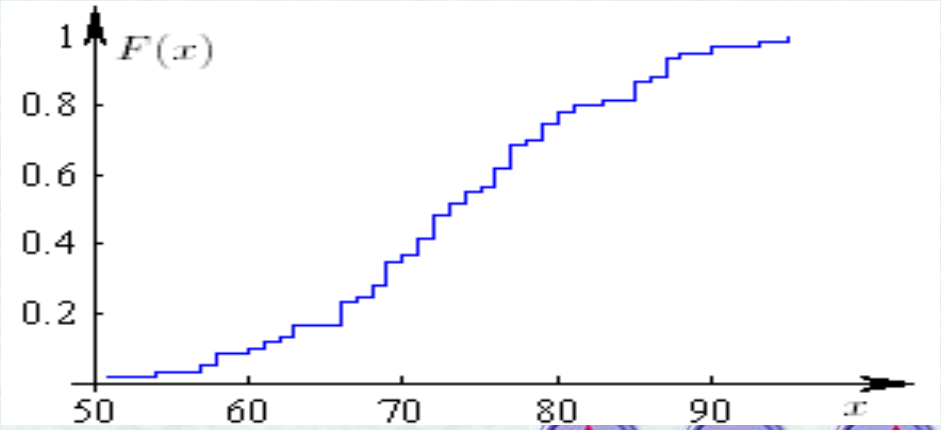
对于任一实数 x 当 n 充分大时,经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别,从而在实际上可当作 $F(x)$ 来使用.



由伯努利大数定理知$F_n(x)$依概率收敛于$F(x)$. 实际上，$F_n(x)$还一致地收敛于$F(x)$，所谓的格里文科定理指出了这一更深刻的结论，即

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

所以，当n充分大时经验分布函数$F_n(x)$是总体分布函数$F(x)$的一个良好的近似.



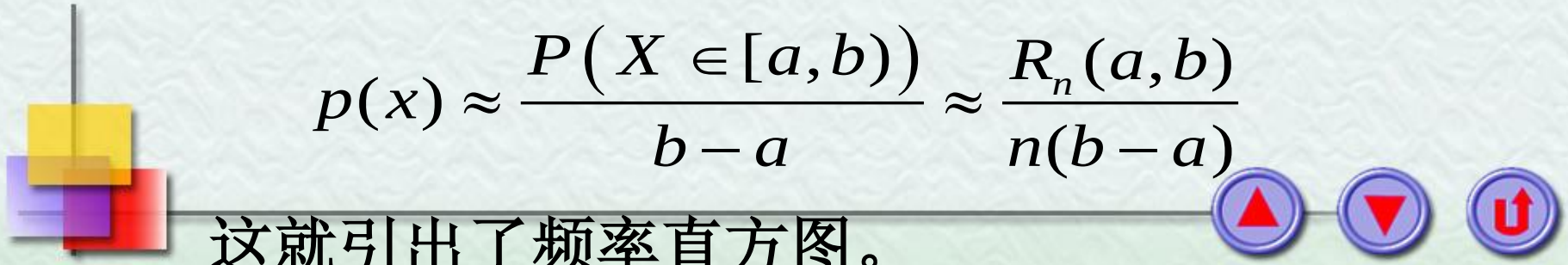
如果总体X的分布函数$F(x)$有密度函数$p(x)$，怎样利用样本$(X_1,X_2,\ldots,X_n)$来刻画这个密度函数？任意给定$x\in[a,b)$，则当$[a,b)$区间比较短，而且$p(u)$在$[a,b)$区间变化不大时，有

$$P(X\in[a,b))=\int_a^b p(u)du\approx p(x)(b-a)$$

再次利用频率近似概率的思想，用$R_n(a,b)$表示样本$(X_1,X_2,\ldots,X_n)$中落在$[a,b)$的个数，那么

$$p(x)\approx\frac{P(X\in[a,b))}{b-a}\approx\frac{R_n(a,b)}{n(b-a)}$$

这就引出了频率直方图。



依次给定$m+1$个实数$t_0<t_1<\ldots<t_m$，

其中$t_1-t_0=t_2-t_1=t_3-t_2=\ldots=t_m-t_{m-1}=h>0$. 令

$$f_n(x)=\begin{cases}\sum_{i=0}^{m-1}\dfrac{R_n(t_i,t_{i+1})}{nh}I_{[t_i,t_{i+1})}(x) & x\in[t_0,t_m)\\ 0 & \text{其它}\end{cases}$$

用$f_n(x)$作为密度函数$p(x)$的估计，这就是频率直方图法。

设(x_1，… ，x_n)是得到的样本观测值，在实际使用时，我们往往用以下步骤具体给出频率直方图

（1）找出$x_{(1)}$,$x_{(n)}$，选择适当的$a<x_{(1)}$,$b>x_{(n)}$.（例如将$x_{(1)}$缩小半个刻度作为a，将$x_{(n)}$放大半个刻度作为b）取m-1个分点$a=t_0<\ldots<t_m=b$，得到m个等分区间$[t_{i-1},t_i)(i=1,\ldots,m)$。$t_{i-1}$称为第$i$组的下组界；$t_i$称为第$i$组的上组界；$h=t_i-t_{i-1}$称为组距或步长。则每个数据都落在其中的一个小区间上。

（2）统计落在每一组上的频数$n_i=R_n(t_i,t_{i+1})$；计算 $f_i= n_i/nh$，$i=0,1,\ldots,m-1$.

（3）以$[t_{i-1},t_i)(i=1,\ldots,m)$为底，$f_i$为高作矩形，即频率直方图。

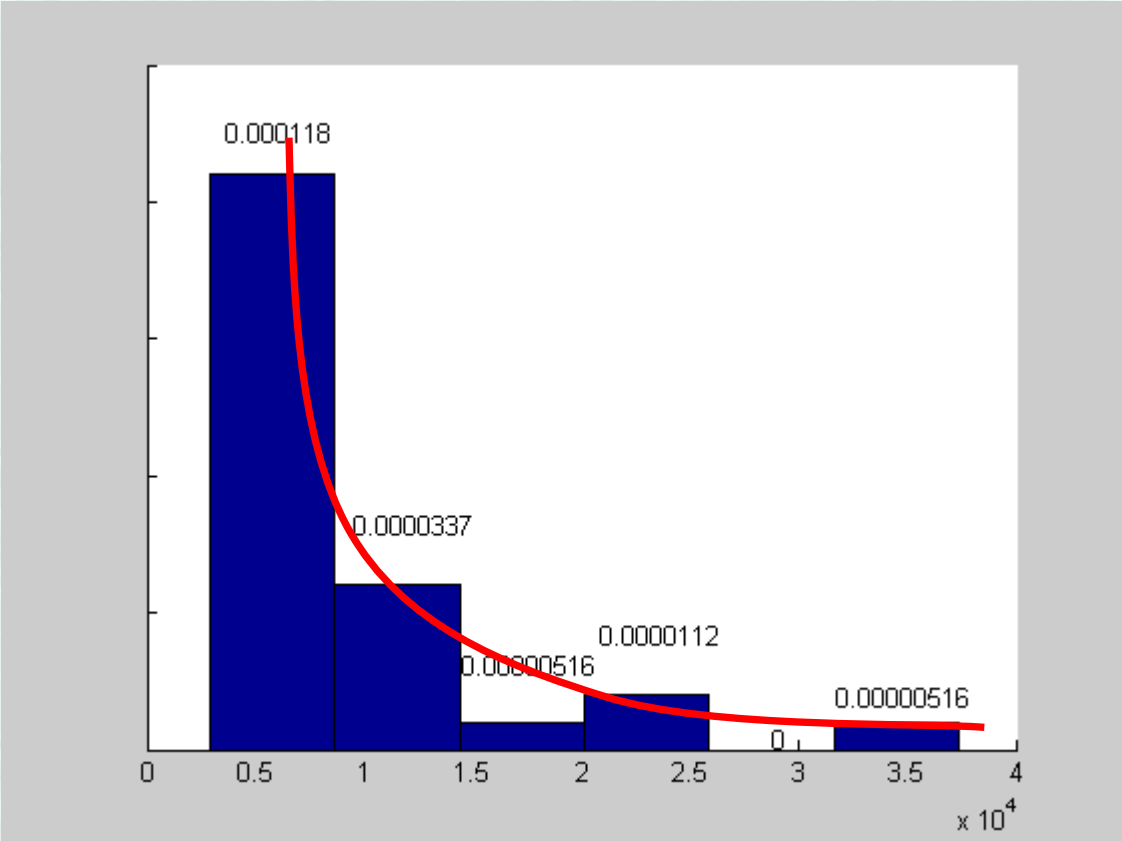
【例7】 下表为我国大陆各省、直辖市2001年人均国民生产总值（万元），试做出频率直方图，从中判断数据大概是来自什么样的总体？

北京 25523	天津 20154	河北 8362	山西 5460	内蒙古 6463	辽宁 12041	吉林 7640	黑龙江 9349
上海 37382	江苏 12922	浙江 14655	安徽 5221	福建 12362	江西 5221	山东 10465	河南 5924
湖北 7813	湖南 6054	广东 13730	广西 4668	海南 7135	重庆 5654	四川 5250	贵州 2895
云南 4866	西藏 5307	陕西 5024	甘肃 4163	青海 5735	宁夏 5340	新疆 7913	

解： 最小值为2895，最大值为37382，取a=2894.5，b=37382.5，将区间分成6等分，统计频数计算频率，得下表：

分组	频数	频率	频率/组距
[2894.5,8642.5)	21	0.677419	0.000118
[8642.5,14390.5)	6	0.193548	0.0000337
[14390.5,20138.5)	1	0.032258	0.00000516
[20138.5,25886.5)	2	0.064516	0.0000112
[25886.5,31634.5)	0	0	0
[31634.5,37382.5)	1	0.032258	0.00000516

频率直方图如下图所示：



初步判断数据是来自什么样的总体？

指数分布

这个例子中数据量相对来说比较少，一般情况下数据量最好大于100，分组的个数根据数据量来确定，一般介于[n/10,n/5]之间，最多不能超过20组。

直方图是对一组数据x_1，x_2，…，x_n的分布情况的图形描述.

将数据的取值范围分成若干区间（一般是等间隔的），在等间隔的情况，每个区间的长度称为组距．考察这些数据落入每一个小区间的频数和频率，在每一个区间上画一个矩形，它的宽度是组距，高度可以是频数、频率或频率/组距，所得直方图分别称为频数直方图、频率直方图和密度直方图.

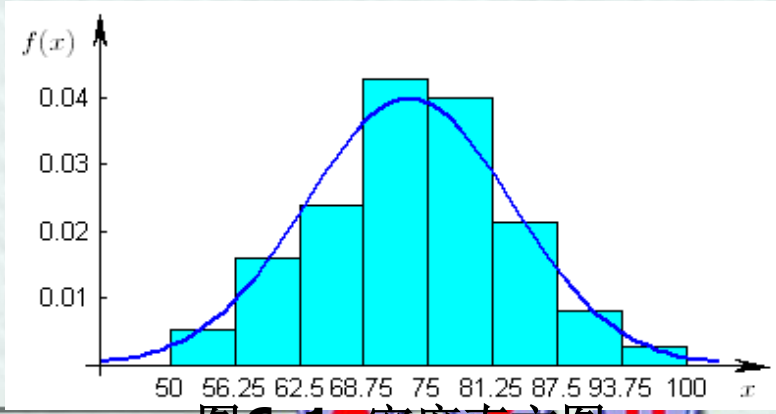
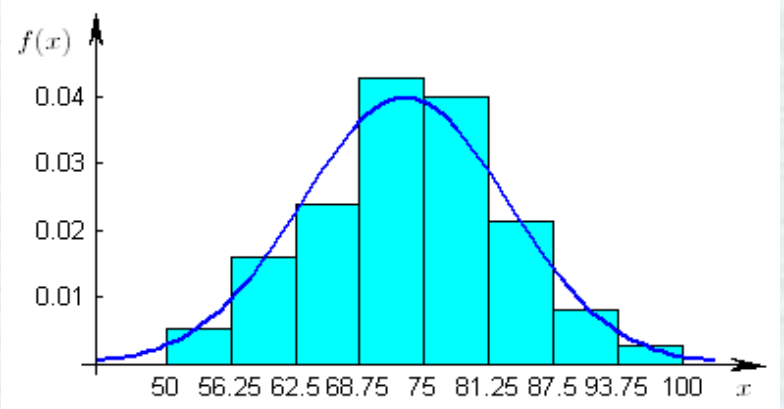
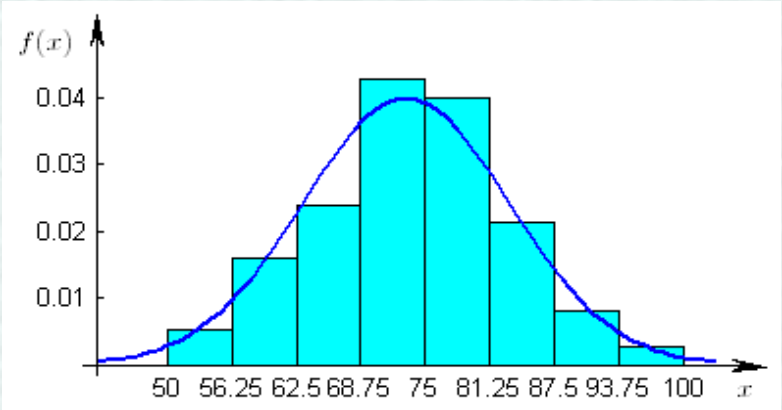


图6-1 密度直方图

如果数据x_1，x_2，…，x_n是来自连续总体X的样本观测值，其密度直方图中，每一个矩形的面积恰好是观测数据落入对应区间的频率，这种密度直方图可以用来估计总体的概率密度（用密度直方图的顶部折线估计X的概率密度曲线）．组距对直方图的形态有很大的影响，组距太小或太大，直方图反映概率密度的形态就不够准确.



一个合适的分组是希望密度直方图的形态接近总体的概率密度函数的形态．手工计算常取组数等于$\sqrt{n}$左右，一些统计软件会根据样本容量和样本的取值范围自动确定一个合适的分组方式，画出各种漂亮的直方图.



分布函数是随机变量的一个重要特征，既然总体可以用随机变量来表示，而样本又可对总体的信息进行提取。因此，怎样用样本$(X_1,\ldots,X_n)$估计总体X的分布函数$F(x)$?

任意给定自变量x，则

$$F(x)=P(X<x).$$

用事件$\{X<x\}$发生的频率作为其估计即可。这就引出了下面所谓经验分布函数的概念。

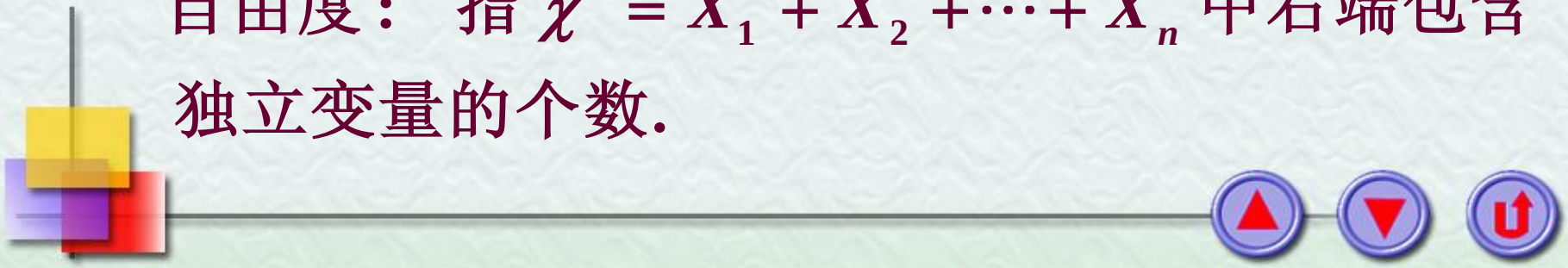
二、常见分布

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.



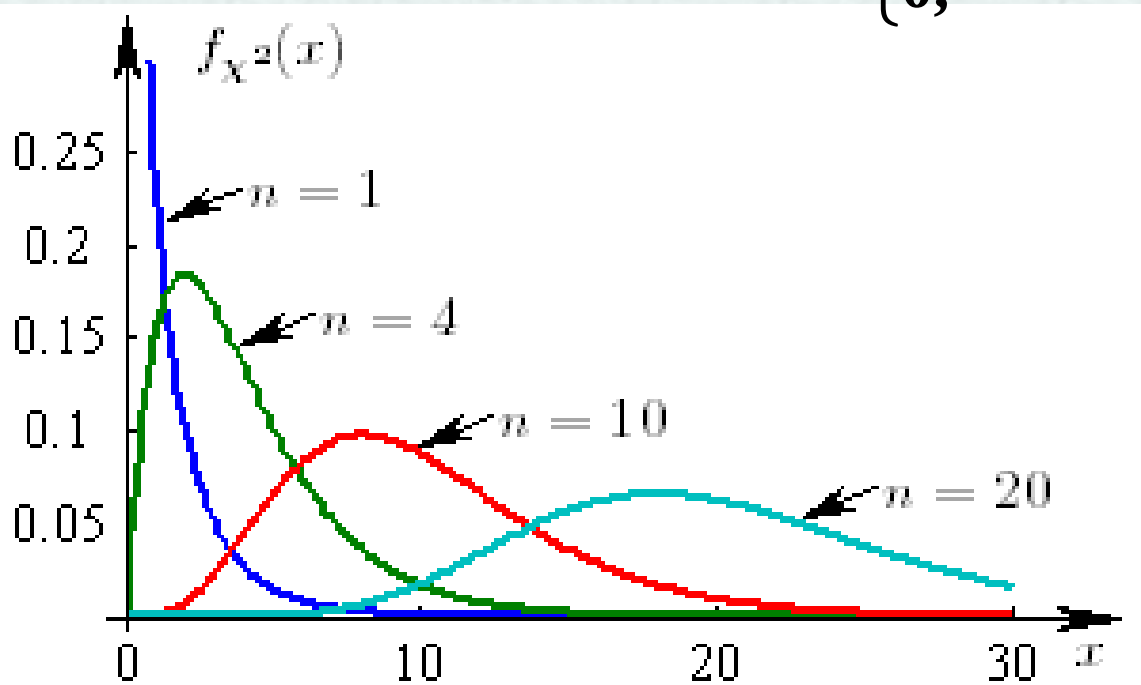
$\chi^2(n)$分布的概率密度为

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

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

$\Gamma(\alpha)$称为伽马函数

χ^2分布概率密度 $f_{\chi^2}(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$



可以看出，随着n的增大，的图形趋于“平缓”，其图形下区域的重心亦逐渐往右下移动.

$\chi^2(n)$分布的概率密度曲线

n越大，图形越扁平，对称性越强

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

$D(X_i^2)=E(X_i^4)-[E(X_i^2)]^2=3-2=1,\ i=1,2,\cdots,n.$

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

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

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

$$=\frac{-1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^3e^{-\frac{x^2}{2}}d\frac{-x^2}{2}=\frac{-1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^3de^{-\frac{x^2}{2}}$$

$$=\frac{-1}{\sqrt{2\pi}}\left[x^3e^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty}-\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}dx^3\right]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}dx^3$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}3x^2dx=\frac{3}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}\frac{1}{2}xdx^2$$

$$=\frac{-3}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}xd\frac{-x^2}{2}=\frac{-3}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}xde^{-\frac{x^2}{2}}$$

$$=\frac{-3}{\sqrt{2\pi}}\left[xe^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty}-\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}dx\right]=\frac{3}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}dx$$

$$=3\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=3\times1=3$$

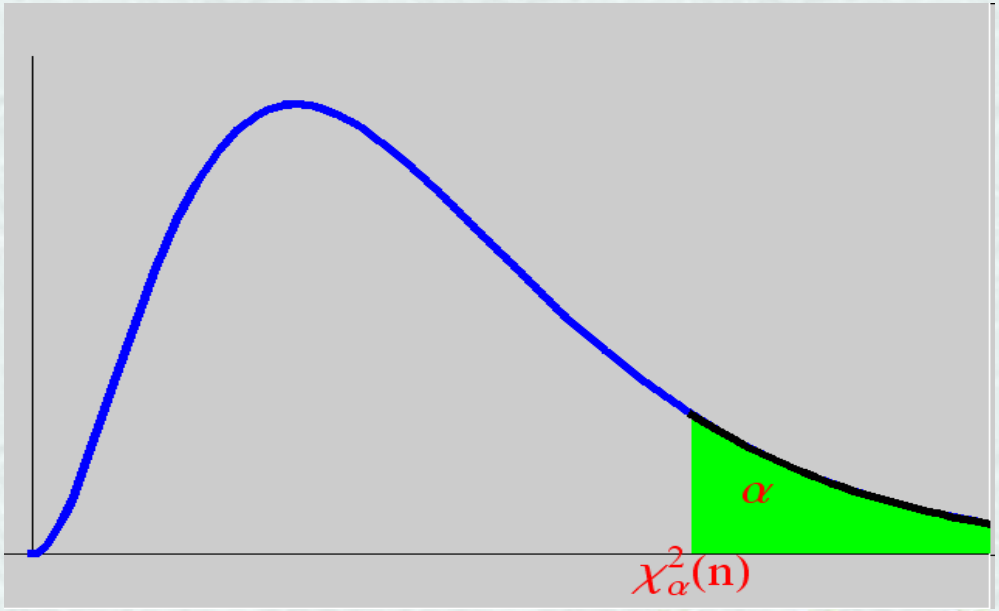
χ^2 分布的分位点

对于给定的正数α, $0<\alpha<1$, 称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(y)\mathrm{d}y = \alpha$$

的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点.

对于不同的$\alpha, n,$
可以通过查表求
得上α 分位点的值.



【例8】设 X_1, X_2， X_3，X_4来自正态总体，$N(0,2^2)$，

的简单随机样本，$X=\dfrac{(X_1-2X_2)^2}{a}+\dfrac{(3X_3-4X_4)^2}{b}$

则当a=_____，b=_____ 时，统计量X服从 χ^2

分布，其自由度为 _____ .

解： $X_1-2X_2\sim N(0,20)$

$$E(X_1-2X_2)=E(X_1)-2E(X_2)=0-2\times 0=0$$

$$D(X_1-2X)=D(X_1)+4D(X_2)=4+4\times 4=20$$

$$3X_3-4X_4\sim N(0,100)$$

$$E(3X_3-4X_4)=3E(X_3)-4E(X_4)=0$$

$$D(3X_3-4X_4)$$

$$=D(3X_3)+D(-4X_4)$$

$$=3^2D(X)+4^2D(X)$$

$$=9\times4+16\times4=36+64=100$$

$$X_1-2X_2\sim N(0,20)\qquad 3X_3-4X_4\sim N(0,100)$$

$$\left(\frac{X_1-2X_2-0}{\sqrt{20}}\right)^2+\left(\frac{3X_3-4X_4-0}{\sqrt{100}}\right)^2\sim\chi^2(2)$$

$$\Rightarrow\frac{(X_1-2X_2)^2}{20}+\frac{(3X_3-4X_4)^2}{100}\sim\chi^2(2)$$

$$\therefore a=20\ ,\ b=100\ ,\ n=2$$

【例9】设 X 服从标准正态分布 $N(0,1)$, $N(0,1)$ 的上 α 分位点 z_α 满足 $P\{X > z_\alpha\} = \dfrac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{+\infty} e^{-\frac{x^2}{2}} dx = \alpha$,

求 z_α 的值, 可通过查表完成.

$z_{0.05} = 1.645$,

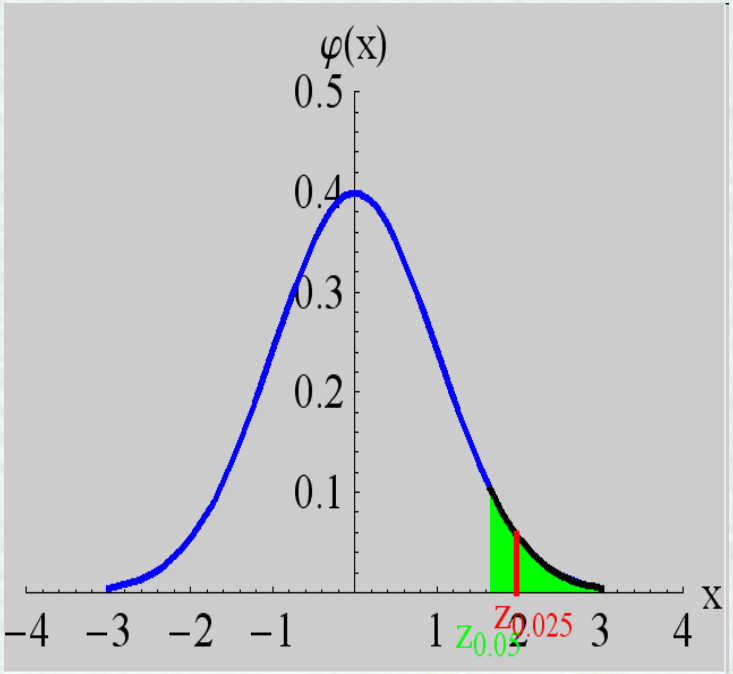
附表2-1

$z_{0.025} = 1.96$,

附表2-2

根据正态分布的对称性知

$$z_{1-\alpha} = -z_\alpha .$$



【例10】设 $Z \sim \chi^2(n)$, $\chi^2(n)$ 的上 α 分位点满足

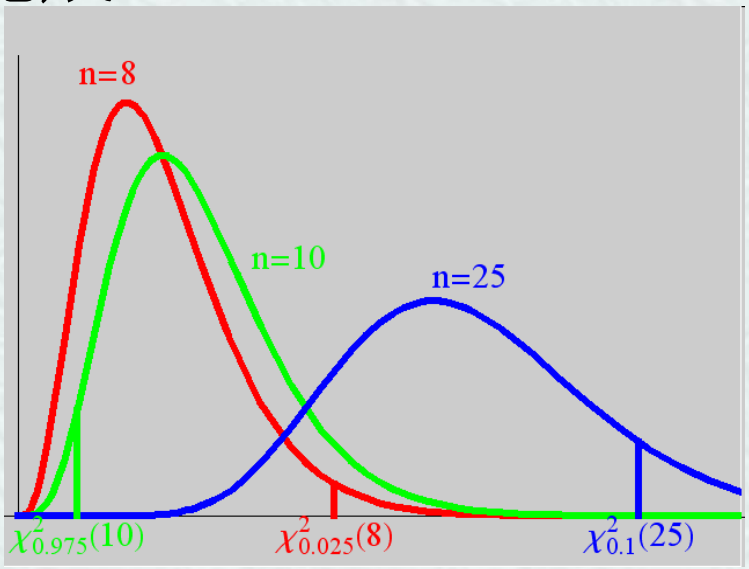
$$P\{Z > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{+\infty} \chi^2(y;n)\mathrm{d}y = \alpha,$$

求 $\chi_\alpha^2(n)$ 的值, 可通过查表完成.

$\chi_{0.025}^2(8) = 17.535,$ 附表4-1

$\chi_{0.975}^2(10) = 3.247,$ 附表4-2

$\chi_{0.1}^2(25) = 34.382.$ 附表4-3



附表4只详列到 n=45 为止.

费舍尔(R.A.Fisher)证明:

当 n 充分大时, $\chi_\alpha^2(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$.

其中 z_α 是标准正态分布的上 α 分位点.

利用上面公式,

可以求得 $n > 45$ 时, 上 α 分位点的近似值.

例如 $\chi_{0.05}^2(50) \approx \frac{1}{2}(1.645 + \sqrt{99})^2 = 67.221.$

而查详表可得 $\chi_{0.05}^2(50) = 67.505$.

【例11】X~N(0, σ^2)，计算 $E\left(X^2\right)$和 σ^2 。

$$EX^2 = DX + \left(EX\right)^2 = \sigma^2 + 0 = \sigma^2$$

$$DX^2 = EX^4 + \left(EX^2\right)^2 = ?$$

经常并不好算

$$\frac{X-0}{\sigma} \sim N(0,1) \qquad \therefore \left(\frac{X-0}{\sigma}\right)^2 \sim \chi^2(1)$$

$$D\left(\frac{X^2}{\sigma^2}\right) = 2, \ \therefore D\left(\frac{X^2}{\sigma^2}\right) = \frac{1}{\sigma^4} D\left(X^2\right) = 2 \Rightarrow D\left(X^2\right) = 2\sigma^4$$

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 $t \sim t(n)$.

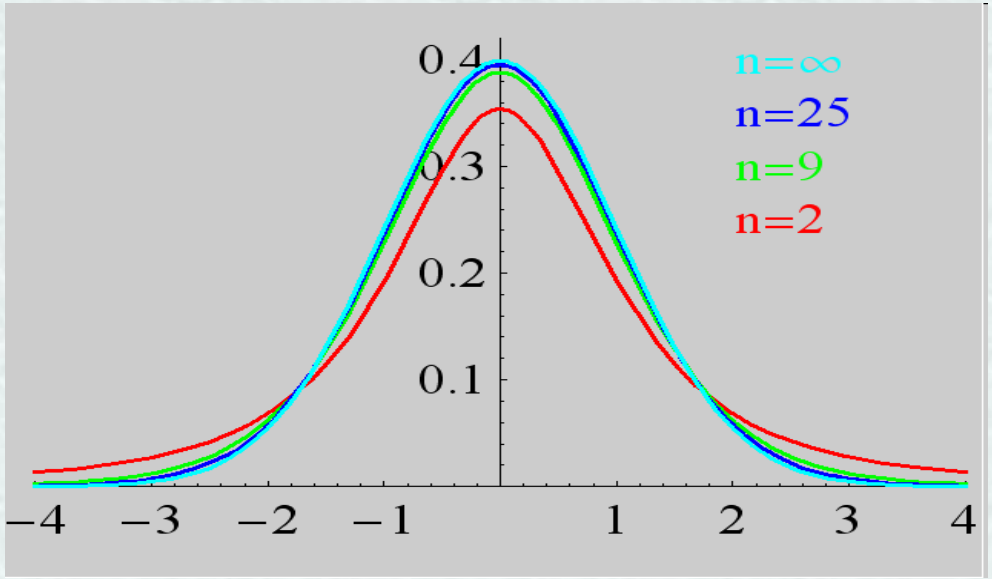
t 分布又称学生氏(Student)分布.

$t(n)$ 分布的概率密度函数为

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

t分布又叫学生氏分布，是英国统计学家W.S.Gosset于1908年凭经验发现的，以笔名Student发表。

t 分布的概率密度曲线如图



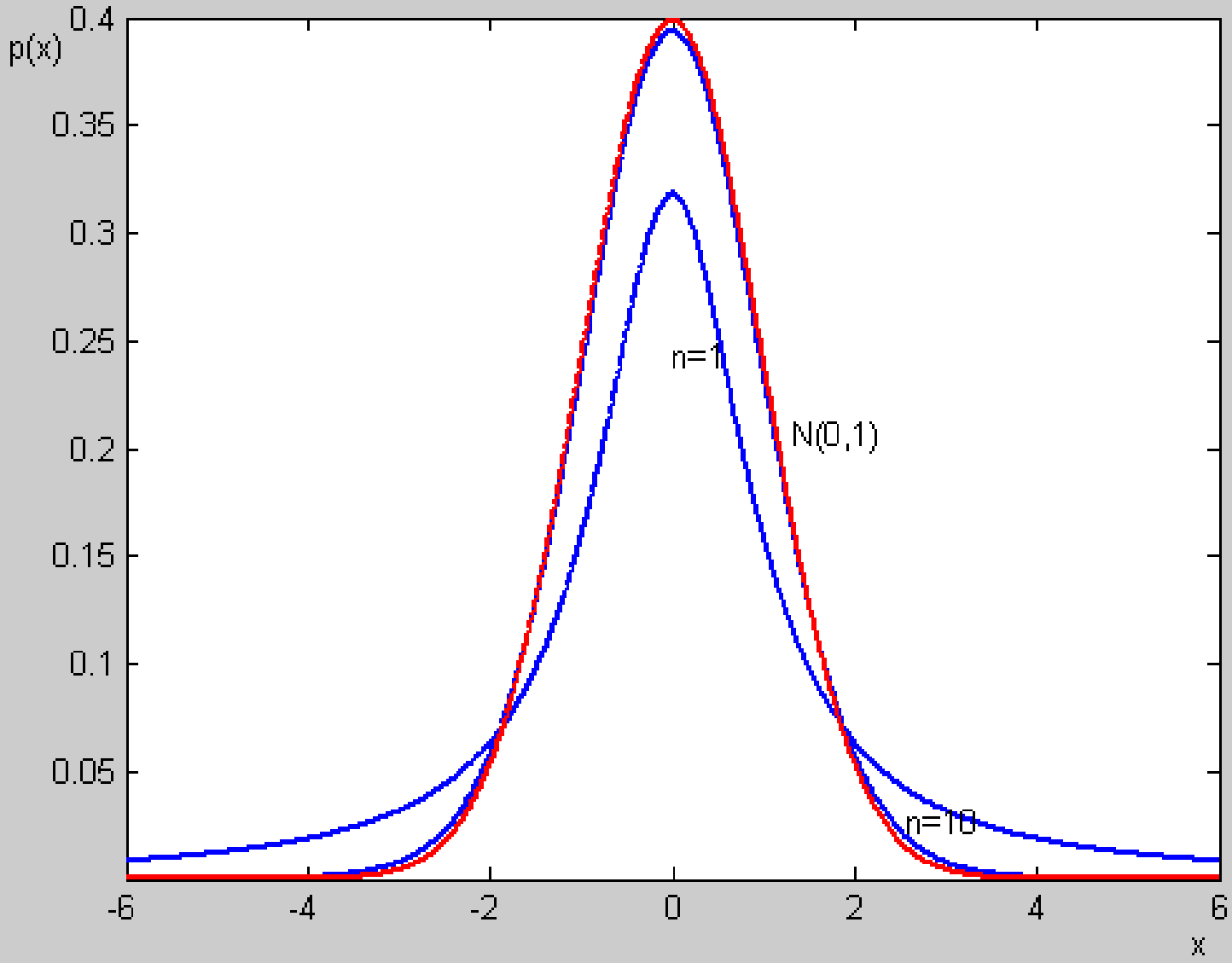
显然图形是关于 $t=0$ 对称的.

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.


p(x)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
-6
-4
-2
0
2
4
6
x
n=1
N(0,1)
n=10

t 分布的分位点

对于给定的 α，$0<\alpha<1$，称满足条件

$$P\{t>t_\alpha(n)\}=\int_{t_\alpha(n)}^{\infty}h(t)\mathrm{d}t=\alpha$$

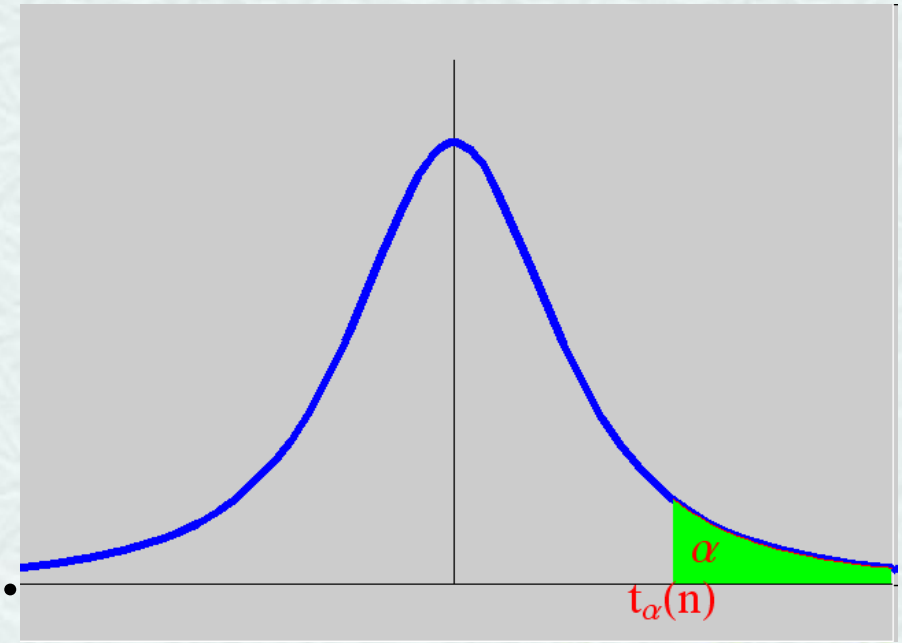
的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位点.

可以通过查表求得上α分位点的值.

由分布的对称性知

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

当 $n>45$ 时，$t_\alpha(n)\approx z_\alpha$.



【例12】设 $T \sim t(n)$, $t(n)$的上 α 分位点满足

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

求 $t_\alpha(n)$ 的值, 可通过查表完成.

$t_{0.05}(10) = 1.8125$, 附表3-1

$t_{0.025}(15) = 2.1315$. 附表3-2

【例13】设随机变量X和Y相互独立都服从$N(0,3^2)$，而 $X_1,\ldots\ldots,X_9$ 和 $Y_1,\ldots\ldots,Y_9$ 分别是来自总体X和Y的简单随机样本，则统计量 $\dfrac{X_1+\ldots\ldots+X_9}{\sqrt{Y_1^2+\ldots+Y_9^2}}$ 服从 _____ 分布，参数为

解：$$X_1+\ldots\ldots+X_9\sim N(0,9\times 9)$$

$$E(X_1+\ldots\ldots+X_9)=\sum_{i=1}^{9}EX_i$$

$$D(X_1+\ldots\ldots+X_9)=\sum_{i=1}^{9}DX_i=9\times 9$$

标准化 $\dfrac{X_1+\cdots\cdots+X_9-0}{9}\sim N(0,1)$

$$Y_i\sim\left(0,3^2\right),\quad \frac{Y_i-0}{3}\sim N(0,1)$$

$$\frac{Y_1^2}{3^2}+\frac{Y_2^2}{3^2}+\cdots+\frac{Y_9^2}{3^2}\sim\chi^2(9)\Rightarrow\frac{Y_1^2+\cdots+Y_9^2}{9}\sim\chi^2(9)$$

$$\Rightarrow\frac{\dfrac{X_1+\cdots+X_9}{9}}{\sqrt{\dfrac{Y_1^2+\cdots+Y_9^2}{9}/9}}\sim t(9)$$

$$\Rightarrow\frac{X_1+\cdots X_9}{\sqrt{Y_1^2+\cdots+Y_9^2}}\sim t(9)$$

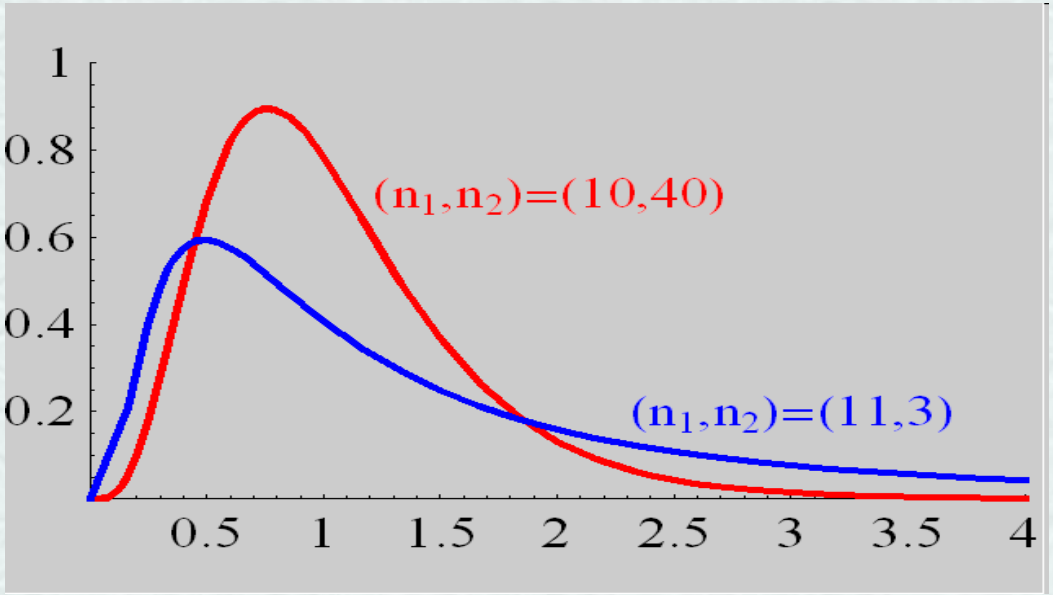
3. F分布

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

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

F分布的概率密度曲线如图



根据定义可知,

若$F \sim F(n_1, n_2)$,

则$\dfrac{1}{F} \sim F(n_2, n_1)$.

F 分布的分位点

对于给定的α, $0<\alpha<1$, 称满足条件

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

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

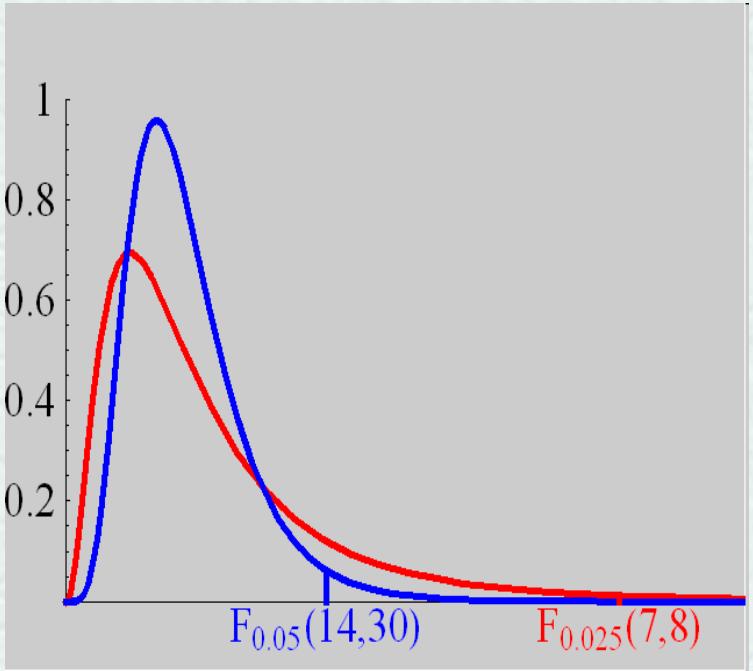
【例14】设 $F(n_1, n_2)$分布的上α分位点满足

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

求 $F_\alpha(n_1, n_2)$的值, 可通过查表完成.

$F_{0.025}(7,8) = 4.90,$ 附表5-1

$F_{0.05}(14,30) = 2.31.$ 附表5-2



F 分布的上 α 分位点具有如下性质：

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

证明 因为 $F \sim F(n_1, n_2)$,

所以 $1-\alpha = P\{F > F_{1-\alpha}(n_1, n_2)\}$

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

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

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

因为 $\frac{1}{F} \sim F(n_2, n_1)$，所以 $P\left\{\frac{1}{F} > F_\alpha(n_2, n_1)\right\} = \alpha$,

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

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

用来求分布表中未列出的一些上 α 分位点.

例 $F_{0.95}(12,9) = \frac{1}{F_{0.05}(9,12)} = \frac{1}{0.28} = 0.357$.

【例15】总体X服从 $N(0,2)$ 正态分布，而 $X_1+\cdots+X_{15}$ 是来自总体X的简单随机样本，则随机变量 $Y=\dfrac{X_1^2+\cdots+X_{10}^2}{2\left(X_{11}^2+\cdots+X_{15}^2\right)}$ 服从_____分布，参数为_____。

解：$X_i\sim N\left(0,2^2\right)$ $\therefore$ $\dfrac{X_i-0}{2}\sim N(0,1)$

$\left(\dfrac{X_1}{2}\right)^2+\cdots+\left(\dfrac{X_{10}}{2}\right)^2\sim\chi^2(10)$ 即 $\dfrac{X_1^2+\cdots+X_{10}^2}{4}\sim\chi^2(10)$

$$\left(\frac{X_{11}}{2}\right)^2+\cdots+\left(\frac{X_{15}}{2}\right)^2\sim\chi^2(5)\ \text{即}\ \frac{X_{11}^2+\cdots+X_{15}^2}{4}\sim\chi^2(5)$$

$$\frac{\dfrac{X_1^2+\cdots+X_{10}^2}{4}/10}{\dfrac{X_{11}^2+\cdots+X_{15}^2}{4}/5}=\frac{X_1^2+\cdots+X_{10}^2}{2\left(X_{11}^2+\cdots+X_{15}^2\right)}\sim F(10,5)$$

【例16】设随机变量 $X \sim t(n),(n>1)$，$Y=\dfrac{1}{X^2}$，

则（ c ）.

A. $Y \sim \chi^2(n)$ **B.** $Y \sim \chi^2(n-1)$

C. $Y \sim F(n,1)$ **D.** $Y \sim F(1,n)$

解:

$$X_1 \sim N(0,1) \quad , X_2 \sim \chi^2(n)$$

$$X=\frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t(n)$$

$$X^2=\frac{X_1^2}{\frac{X_2}{n}}=\frac{\frac{X_1^2}{1}}{\frac{X_2}{n}} \quad , \quad Y=\frac{1}{X^2}=\frac{\frac{X_2}{n}}{\frac{X_1^2}{1}} \sim F(n,1)$$

4. 正态总体的样本均值与样本方差的分布

定理一

设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}$ 是样本均值, 则有 $\overline{X} \sim N(\mu, \sigma^2/n)$.

正态总体 $N(\mu,\sigma^2)$ 的样本均值和样本方差有以下两个重要定理.

定理二

设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，$\overline{X}, S^2$ 分别是样本均值和样本方差，则有

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

(2) $\overline{X}$ 与 S^2 独立.

定理三 设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}, S^2$ 分别是样本均值和样本方差, 则有

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

证明 因为 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$,

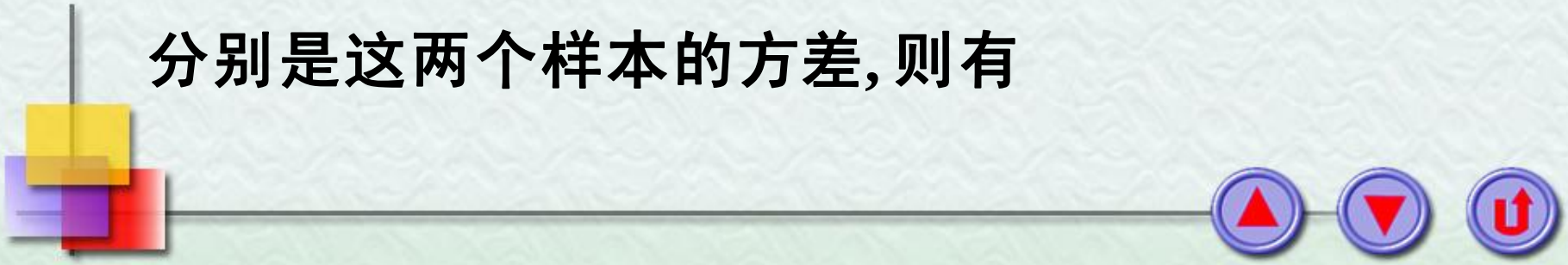
且两者独立, 由 t 分布的定义知

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$$

定理四 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是具有相同方差的两正态总体 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ 的样本, 且这两个样本互相独立, 设 $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的均值,

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

分别是这两个样本的方差, 则有



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

(2) 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时,

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

其中 $S_w^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, $S_w=\sqrt{S_w^2}$.

证明 (1) 由定理二

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

由假设 S_1^2, S_2^2 独立, 则由 F 分布的定义知

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

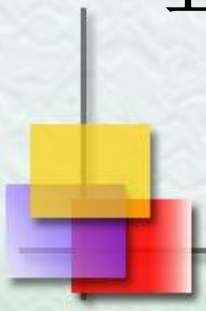
即 $$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1).$$

(2) 因为 $\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}\right)$

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

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

且它们相互独立, 故由 χ^2 分布的可加性知



$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2),$$

由于 U 与 V 相互独立, 按 t 分布的定义.

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}$$

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

三、小结

两个最重要的统计量:

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

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

三个来自正态分布的抽样分布:

χ^2 分布, t 分布, F 分布.