第六章 数理统计的基本概念

1. 数理统计的涵义

1) 数理统计研究的数据必须带有随机性的影响，才能成为数理统计学的研究对象；

2) 数据随机性的来源：一是所研究对象为数很多；二是试验的随机误差；

3) 用有效的方法收集数据：一是建立一个好的模型；二是包含尽可能多的信息.

4) 有效地使用数据：集中和提取信息.

§6.1 总体与样本

在统计研究中，人们往往关心每个个体的一项(或几项)数量指标和该数量指标的分布情况。这时每个个体在该数量指标的所有取值就是总体。

该批灯泡寿命的所有可能的取值就是总体

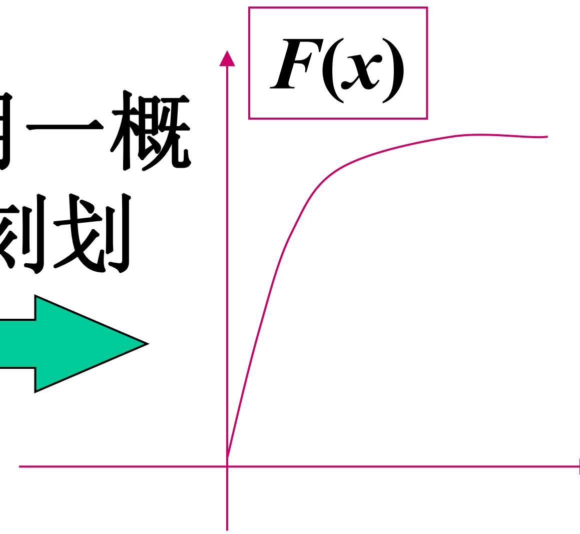
所有国产轿车每公里耗油量的所有可能的取值就是总体

由于每个个体的出现带有随机性，即相应的数量指标值的出现带有随机性。从而可把此种数量指标看作随机变量，我们用一个随机变量或其分布来描述总体。

通常，我们用随机变量X , Y , Z,…, 等表示总体。当我们说到总体，就是指一个具有某种确定概率分布的随机变量。

如:研究某批灯泡的寿命时，我们关心的数量指标就是寿命，那么，此总体就可以用随机变量X表示，或用其分布函数$F(x)$表示.

寿命X可用一概率分布来刻划



因此，在统计学中, 总体这个概念是：

总体就是一个概率分布。

二、随机样本的定义

1. 样本的定义

为推断总体的分布及各种特征，按一定的规则从总体中抽取若干个体进行观察试验，以获得有关总体的信息。这一抽取过程称为“抽样”。

所抽取的部分个体称为样本，通常记为

$$(X_1, X_2, \cdots, X_n)$$

样本中所包含的个体数目n称为样本容量。

容量为n的样本可以看作n维随机变量。当取定一组样本，得到的是n个具体的数，$(x_1, x_2, \cdots, x_n)$称此为样本的一次观察值，简称样本值。

2. 简单随机样本

抽取样本的目的是为了利用样本对总体进行统计推断, 这就要求样本能很好的反映总体的特性且便于处理. 为此, 需对抽样提出一些要求, 通常有两条:

1). 代表性： $X_1, X_2, \ldots, X_n$中每一个与所考察的总体X有相同的分布.

2). 独立性： $X_1, X_2, \ldots, X_n$是相互独立的随机变量.

满足上述两条性质的样本称为简单随机样本.

3.样本的分布

设$(X_1,X_2,\cdots,X_n)$为来自总体X的样本.

(1)若总体X的分布函数为$F(x)$,则样本$(X_1,X_2,\cdots,X_n)$

的联合分布函数为$\prod_{i=1}^{n}F(x_i)$.

(2)若总体X的密度函数为$f(x)$,则样本$(X_1,X_2,\cdots,X_n)$

的联合密度函数为$\prod_{i=1}^{n}f(x_i)$.

(3)若总体X的分布律为$P\{X=x_i\}=p(x_i)(i=1,2,\cdots)$,

则样本$(X_1,X_2,\cdots,X_n)$的联合分布律为$\prod_{i=1}^{n}p(x_i)$.

例1 设总体 X 服从参数为 $\lambda\ (\lambda>0)$ 的指数分布, $(X_1, X_2, \cdots, X_n)$ 是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数。

解 总体 X 的密度函数为: $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$

因为 $X_1, X_2, \cdots, X_n$ 相互独立, 且与 X 有相同的分布,

所以 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数为:

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

例2 设总体 X 服从两点分布 $B(1, p)$, 其中 $0 < p < 1$, $(X_1, X_2, \cdots, X_n)$是来自总体的样本，求样本 $(X_1, X_2, \cdots, X_n)$ 的分布律.

解 总体 X 的分布律为

$$P\{X = i\} = p^i (1 - p)^{1-i} \qquad (i = 0, 1)$$

因为 $X_1, X_2, \cdots, X_n$ 相互独立,

且与 X 有相同的分布,

所以 $(X_1, X_2, \cdots, X_n)$ 的分布律为

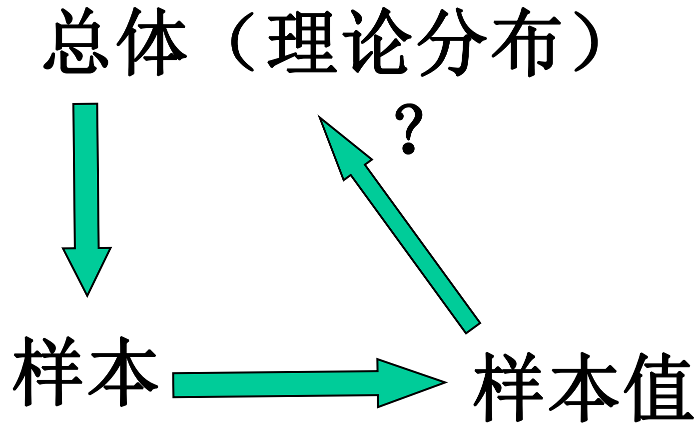
$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$$

$$= P\{X_1 = x_1\}P\{X_2 = x_2\}\cdots P\{X_n = x_n\}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \qquad x_i = 0,1,\ i = 1,2,\cdots,n$$

3. 总体、样本、样本值的关系

事实上我们抽样后得到的资料都是具体的、确定的值. 如我们从某班大学生中抽取10人测量身高，得到10个数，它们是样本取到的值而不是样本. 我们只能观察到随机变量取的值而见不到随机变量.



统计是从手中已有的资料--样本值，去推断总体的情况---总体分布$F(x)$的性质.

样本是联系二者的桥梁

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.

小结

基本概念：个体　总体　随机样本

说明1　一个总体对应一个随机变量X, 我们将不区分总体和相应的随机变量, 统称为总体 X.

说明2　在实际中遇到的总体往往是有限总体, 它对应一个离散型随机变量; 当总体中包含的个体的个数很大时, 在理论上可认为它是一个无限总体.

§6.2　统计量

由样本推断总体特征，需要对样本值进行“加工”，“提炼”。这就需要构造一些样本的函数，它把样本中所含的信息集中起来。

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.

2. 几个常用样本矩统计量的定义

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

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

其观察值

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

(3)样本标准差

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

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

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

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

(5)样本 k 阶中心矩

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

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

显然： $\overline{X} = M_1$

$$M_2^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$$

样本均值和样本方差的性质：

1） $\sum_{i=1}^{n}(X_i-\overline{X})=0$ ；

2）若总体的均值、方差存在，且$EX=\mu$，$DX=\sigma^2$，则：

$$E\overline{X}=\mu, D\overline{X}=\frac{\sigma^2}{n}$$

3）当$n\to\infty$时，$\overline{X}\xrightarrow{p}\mu$

4）如果DX存在，则 $ES^2=DX$

5）对任意实数x，有$\sum_{i=1}^{n}(X_i-\overline{X})^2\le\sum_{i=1}^{n}(X_i-x)^2$

例：设总体X~$\Gamma(1,3)$,B(100,0.1),P(3)，$X_1,X_2,\ldots,X_n$为X的样本。求：$E\bar{X},D\bar{X},ES^2,EM_2^*$

3.顺序统计量

定义 设$(X_1, X_2, \cdots, X_n)$是从总体X中抽取的一个样本，$(x_1, x_2, \cdots, x_n)$是其一个观测值，将观测值按由小到大的次序重新排列为

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

当$(X_1, X_2, \cdots, X_n)$取值为$(x_1, x_2, \cdots x_n)$时，定义$X_{(k)}$取值为$x_{(k)}(k = 1, 2, \cdots n)$，由此得到

$$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$$

称其为样本$(X_1, X_2, \cdots, X_n)$的次序统计量.对应的$(x_{(1)}, x_{(2)}, \cdots x_{(n)})$称为其观测值.

特别的

$$X_{(1)}=\min_{1\le i\le n} X_i \text{ 称为最小顺序统计量 .}$$

$$X_{(n)}=\max_{1\le i\le n} X_i \text{ 称为最大顺序统计量 .}$$

说明

由于每个$X_{(k)}$都是样本$(X_1,X_2,\cdots,X_n)$的函数,所以,$X_{(1)},X_{(2)},\cdots,X_{(n)}$也都是随机变量,并且它们一般不相互独立.

称：$R = X_{(n)} - X_{(1)}$ 为样本极差；

称：$\widetilde{X} = \begin{cases} X_{(\frac{n+1}{2})}, & \text{当n为奇数时} \\ \frac{1}{2}\left[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right], & \text{当n为偶数时} \end{cases}$ 为中位数。

设总体X的密度函数为$f(x)$(或分布函数为$F(x)$)，$(\mathbf{X}_{(1)},\mathbf{X}_{(2)},\cdots,\mathbf{X}_{(n)})$为总体$X$的样本$(X_1,X_2,\cdots,X_n)$的顺序统计量.则有：

(1) 最小顺序统计量$X_{(1)}$的密度函数为：

$$f_{(1)}(x)=n[1-F(x)]^{n-1}f(x)$$

(2) 最大顺序统计量$X_{(n)}$的分布密度为

$$f_{(n)}(x)=n[F(x)]^{n-1}f(x)$$

例：设总体X~$\Gamma(\lambda)$，$X_1,X_2,\ldots,X_n$为X的样本。

求：$f_{(1)}(x)$，$f_{(n)}(x)$。

重庆大学
CHONGQING UNIVERSITY

4. 经验分布函数(样本分布函数)

定义 设 $X_1, X_2, \cdots, X_n$ 是总体 X 的一个样本，$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 为总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 的次序统计量.

$(x_{(1)}, x_{(2)}, \cdots x_{(n)})$ 为其观测值，设 x 是任一实数，

称函数

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

为总体 X 的经验分布函数。其中，k 表示小于等于 x 的观察值个数。

例1 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

例2 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

经验分布函数性质:

$F_n(x)$满足分布函数的特征，是一个分布函数：

(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) $F_n(x)$在每个观察值 $x_{(i)}$处是右连续的.

3.格里汶科定理

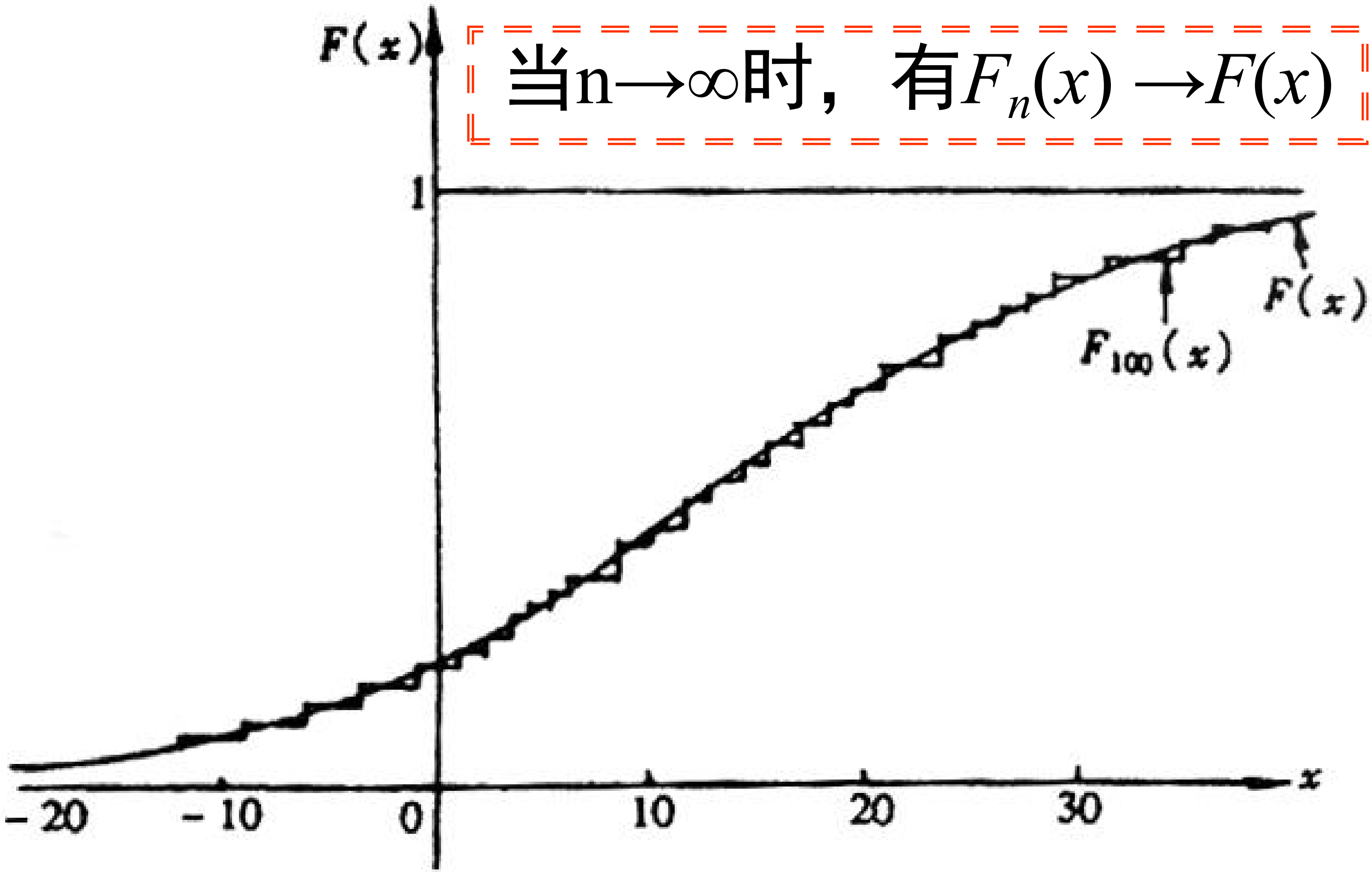
对于任一实数 x,当$n \to \infty$时,$F_n(x)$以概率1一致收敛于分布函数 $F(x)$,即

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

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.

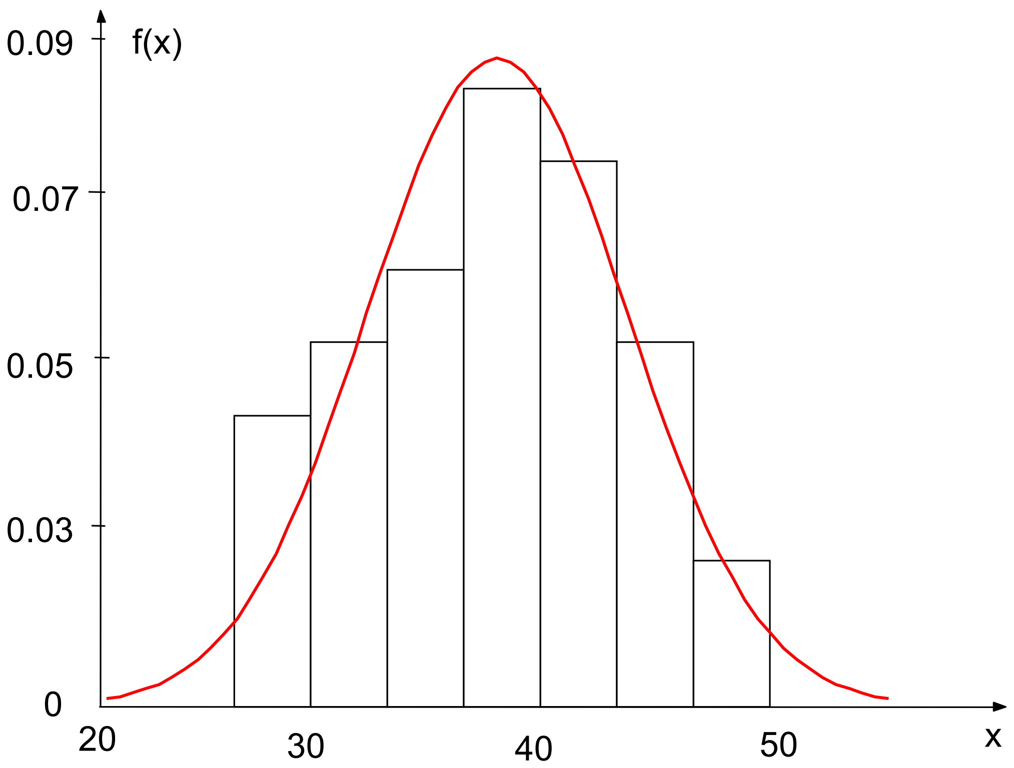
$F_n(x)$与$F(x)$之间的关系

当n→∞时，有$F_n(x) \to F(x)$



直方图

总体X是连续型的，有密度函数$f(x)$,则利用统计数据，可以画直方图逼近函数$f(x)$。



6.5 抽样分布

既然统计量是依赖于样本的, 而后者又是随机变量, 故统计量也是随机变量, 因而就有一定的分布. 称这个分布为“抽样分布”. 也即抽样分布就是统计量的分布

常见正态总体的抽样分布讨论。

§ 6.4 抽样分布

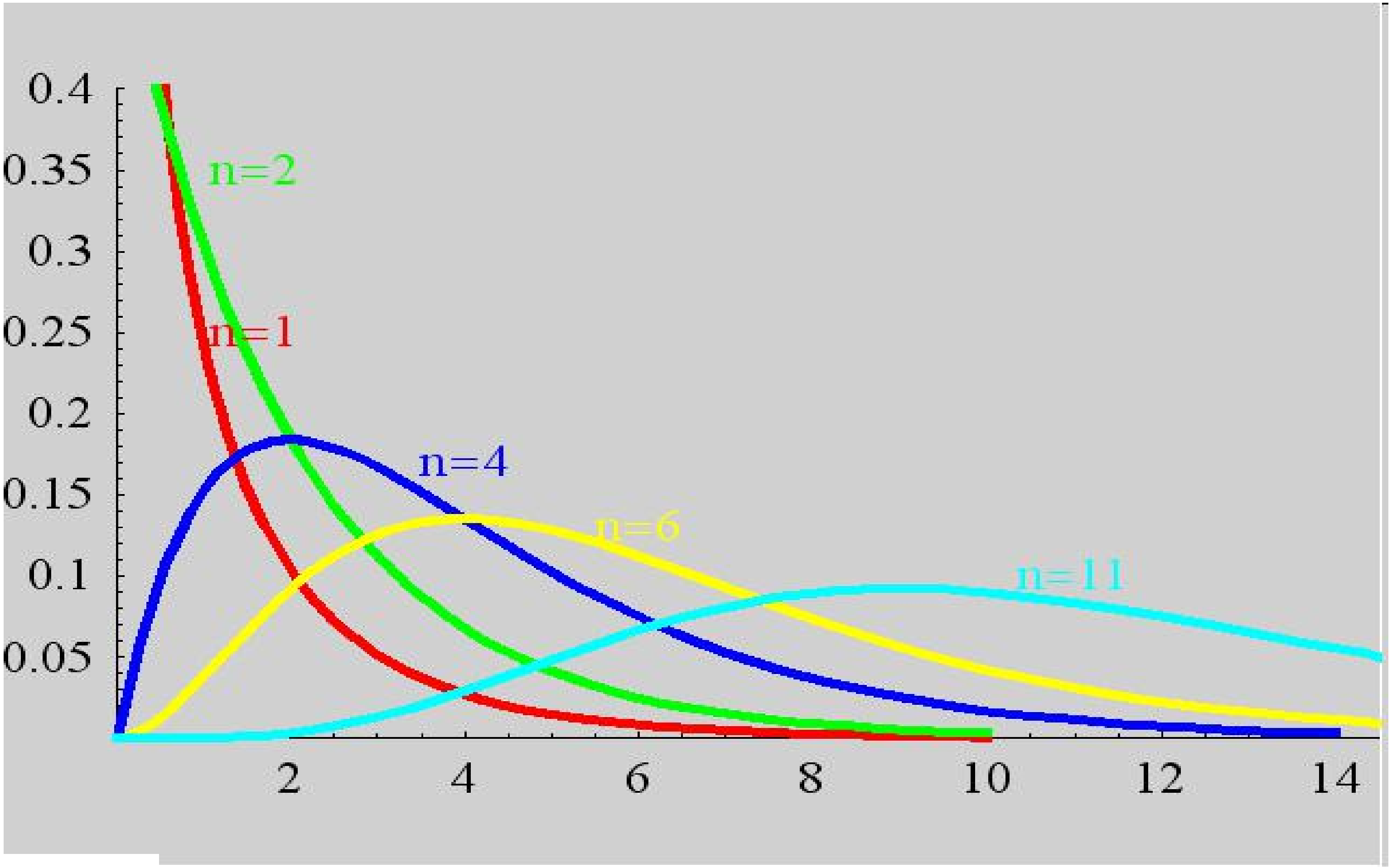
1. χ^2分布

定理1、设 $X_1, X_2, \cdots, X_n$ 相互独立，同服从 $N(0, 1)$ 分布,则随机变量 $\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 服从自由度为 n 的 χ^2分布, 记为 $\chi_n^2 \sim \chi^2(n)$.

其概率密度为

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

$\chi^2(n)$分布的概率密度曲线如图.



重庆大学
CHONGQING UNIVERSITY

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 $E(\chi^2) = n$, $D(\chi^2) = 2n$。

重庆大学 CHONGQING UNIVERSITY

2. t 分布（学生(Student)分布）

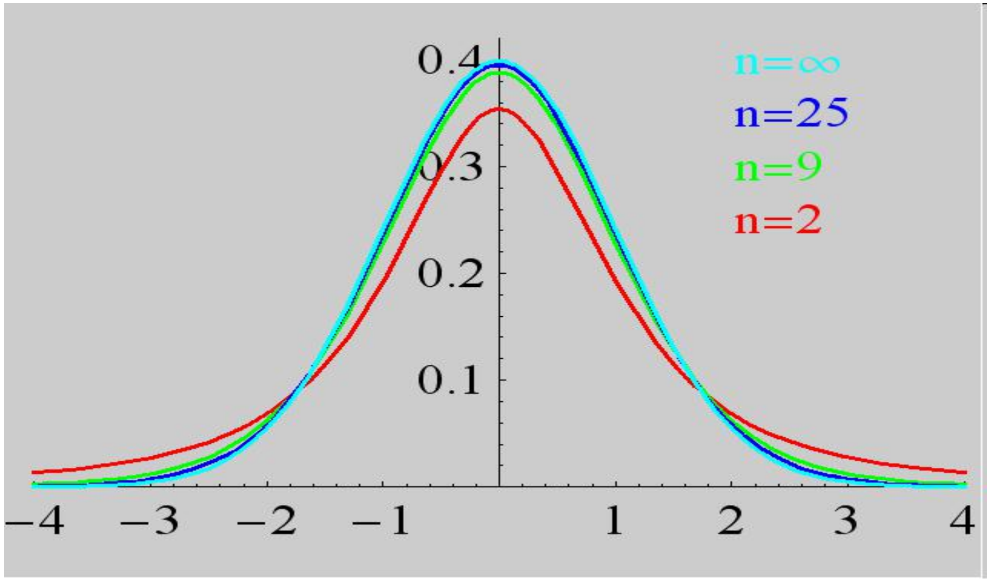
定理2： 设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 X, Y 独立,则随机变量 $T = \dfrac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布, 记为 $T \sim t(n)$.

t 分布的概率密度函数为

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

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

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 .

t 分布的性质

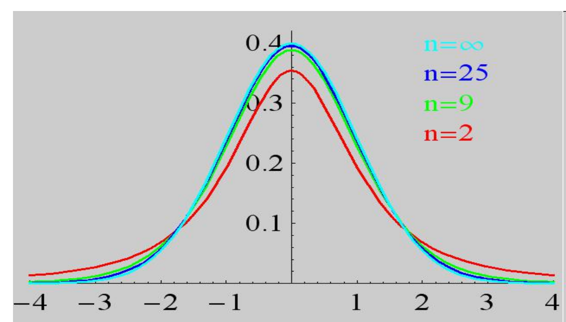
1）当n=1时，T的密度函数为*Cauchy*分布

$$f(x)=\frac{1}{\pi}\cdot\frac{1}{1+x^2},\quad x\in R$$

2）当n>1时，$E(T)$=0，$f(x)$的曲线关于y轴$(x=0)$对称。

3）当n>2时，$D(T)=\frac{n}{n-2}$

4）当$n\to\infty$时，$f(x)\to\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$



3. F分布

定理3：设$X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且X, Y独立,

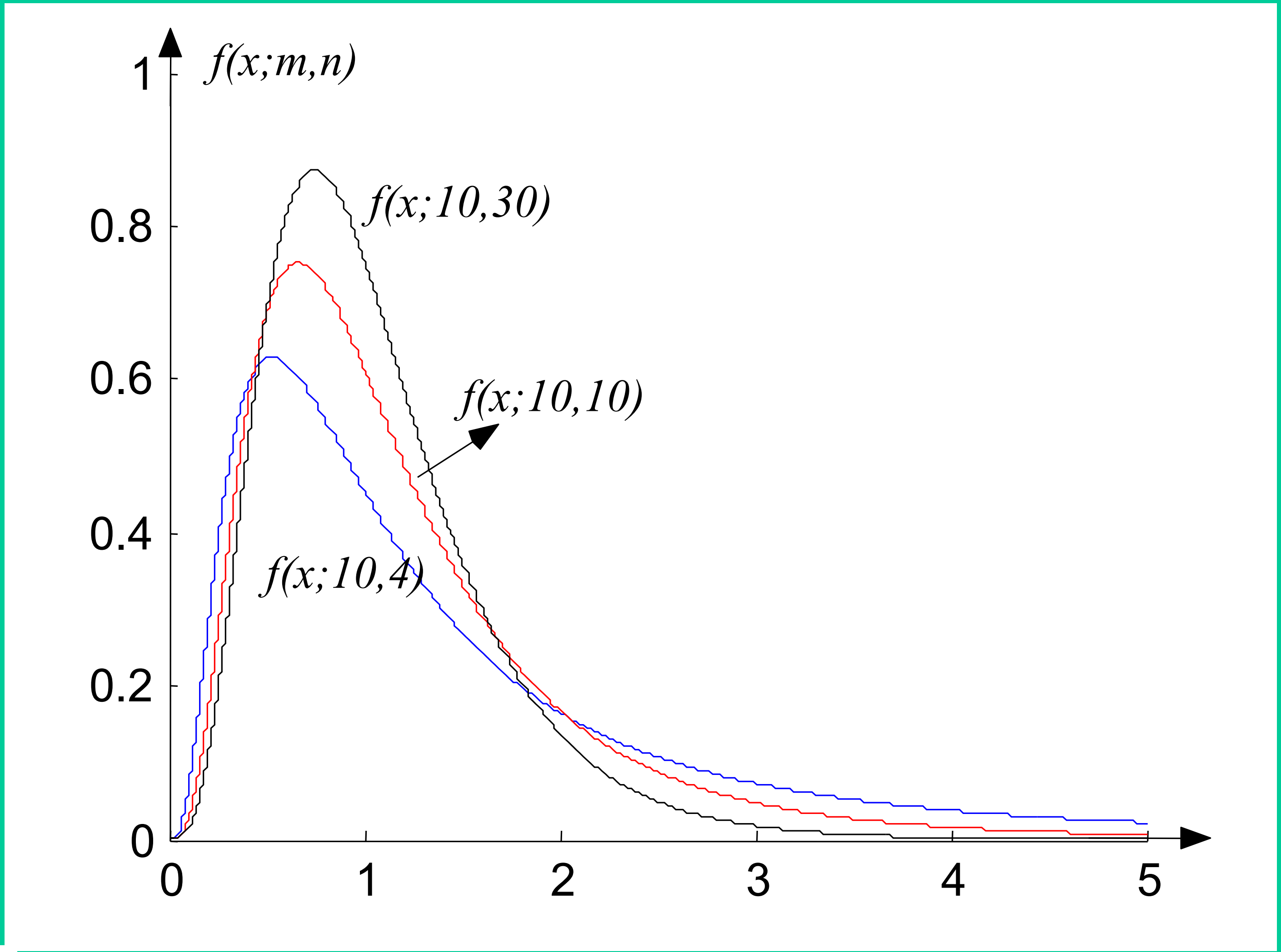
则称随机变量$F = \dfrac{X/n_1}{Y/n_2}$服从自由度为(n_1, n_2)

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

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

F 分布的图形特征



F分布有以下性质

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

(2) 当$T \sim t(n)$时，则$T^2 \sim F(1,n)$。

例4、设X_1, X_2, X_3, X_4独立同分 $N(0,2^2)$，令

$$Y_1 = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$$

$$Y_2 = c\frac{X_1 - X_2}{\sqrt{X_3^2 + X_4^2}} \quad Y_3 = d\frac{(X_1 - 2X_2)^2}{X_3^2 + X_4^2}$$

求：

1） 参数a,b，使 Y_1 服从 χ^2 分布，并求其自由度；

2） 参数c，使 Y_2 服从t 分布，并求其自由度；

3） 参数d，使得 Y_3 服从F分布，并求其自由度.

抽样分布定理

定理　设$(X_1,X_2,\ldots,X_n)$是来自总体X的样本，

设$X\sim N(\mu,\sigma^2)$，则

1. $\bar{X}\sim N(\mu,\dfrac{\sigma^2}{n})$, 或 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

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

3. $\bar{X}$与S^2相互独立。

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

5. $E(S^2)=\sigma^2$, $D(S^2)=\dfrac{2\sigma^4}{n-1}$

定理2 设$(X_1,X_2,\ldots,X_m)$是来自总体$X\sim N(\mu_1,\sigma_1^2)$的样本，$(Y_1,Y_2,\ldots,Y_n)$是来自总体$Y\sim N(\mu_2,\sigma_2^2)$的样本，两样本相互独立。且：

$$\bar{X}=\frac{1}{m}\sum_{i=1}^{m}X_i \qquad \bar{Y}=\frac{1}{n}\sum_{j=1}^{n}Y_j$$

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

则：1）$F=\dfrac{\sigma_2^2S_1^2}{\sigma_1^2S_2^2}\sim F(m-1,n-1)$

2）当$\sigma_1^2=\sigma_2^2=\sigma^2$时，

$$T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(m-1)S_1^2+(n-1)S_2^2}}\sqrt{\frac{mn(m+n-2)}{m+n}}\sim t(m+n-2)$$

分位数

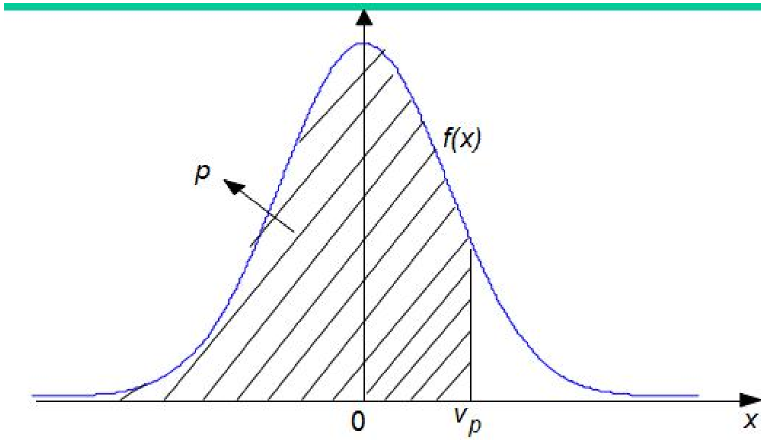
分位数又叫分位点或临界值，是随机变量的一类数字特征，它在统计推断中起着很重要的作用。

定义 设X为一个随机变量，对于给定的常数p

$0<p<1$，称满足下式

$$P\{X \le v_p\} = p$$

的实数v_p为X的p分位数。



常见分位数及性质

符号表示

正态分位数—— u_p

$$X \sim N(0,1), P\{X \le u_p\} = p$$

t 分位数 —— $t_p(n)$

$$X \sim t(n), P\{X \le t_p(n)\} = p$$

卡方分位数—— $\chi_p^2(n)$

$$X \sim \chi^2(n), P\{X \le \chi_p^2(n)\} = p$$

F 分位数 —— $F_p(m,n)$

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

查表计算：

$u_{0.992} = 2.41$

$u_{0.95} = 1.645$

$t_{0.95}(8) = 1.86$

$t_{0.90}(16) = 1.337$

$\chi_{0.95}^2(7) = 14.07$

$\chi_{0.05}^2(17) = 8.67$

$F_{0.95}(8,6) = 4.15$

$F_{0.99}(11,4) = 14.45$

分位数的性质

1） $u_p = -u_{1-p}$

2） $t_p(n) = -t_{1-p}(n)$

3）当n>45时，$t_p(n) \approx u_p$,

$$\chi_p^2(n) \approx \frac{1}{2}(u_p + \sqrt{2n-1})^2$$

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

查表计算：

$$u_{0.05} = -u_{1-0.05}$$

$$= -u_{0.95}$$ **=-1.645**

$$t_{0.05}(8) = -t_{0.95}(8)$$

=-1.86

$$F_{0.05}(6,8) = \frac{1}{F_{0.95}(8,6)}$$

$$= \frac{1}{4.15}$$