第七章 数理统计基础知识

§7.1 总体与样本

顾名思义，总体就是研究对象的全体，总体中每个成员就称为个体。但一般来说,我们并不研究总体的一切属性，而只研究其某一项数量指标，因此，我们把总体以及个体的定义改写为：总体即研究对象某项数量指标的全体；个体即总体中的每个元素。

例如：某工厂生产的全体灯泡的寿命是一个总体（而非全体灯泡），每一个灯泡的寿命是一个个体；某学校全体男生的身高是一个总体（而不是全体男生），每个男生的身高是一个个体。

总体（按数量分）：有限总体、无限总体。若有限总体中个体数量很大，也可近似地认为是无限总体。

那么，怎么才能了解总体的性质呢？最好的莫过于对每个个体都进行观测，试验，但这往往是行不通的，特别是无限总体，我们要面临这些问题：能无穷次地抽取下去吗？所以一般采用抽样调查的方法：从总体中抽出一些个体，并对这些个体进行观测，试验，用得到的的数据去推测总体并对总体作出判断。

从总体中抽出个体，在抽到某个个体前，这个个体的数量指标事先并不确知，因而是随机变量，用X

表示。一般地，我们将总体及其所对应的随 机变量不加区别，都记为X。

为了解总体的性质，我们从总体中抽出了n 个个体 $X_1, X_2, \ldots, X_n$，称为是来自于总体的容量为 n 的样本，对该样本进行观测、试验，就可以得到相应的一组数值 $x_1, x_2, \ldots, x_n$，称为样本值（或观测值）.

那么，如何抽取样本呢？首先，为使样本能充分反映总体的状况，每个个体被抽到的机会应相等，即满足随机性；其次，每次抽样应该独立进行，其结果不受其它抽样结果的影响，也不影响其它的结果，即独立性。满足这样两条性质的抽样方法称为简单随机抽样，

其样本称为**简单随机样本**。以后我们所谈 到的抽样（或样本），均指简单随机抽样（或简单随机样本）。

设$X_1, X_2, \ldots, X_n$为来自总体的一个样本，则 $X_1, X_2, \ldots, X_n$ 是n个相互独立且与总体X同分布的随机变量，于是（$X_1, X_2, \ldots, X_n$）构成一个n 维随机变量，若样本值为$x_1, x_2, \ldots, x_n$，那么（$x_1, x_2, \ldots, x_n$）为（$X_1, X_2, \ldots, X_n$）的取值。此时，我们认为 n个事件：$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ 已经发生了。于是

（1）若总体X是连续型的，密度为f(x)，则随机变量$(X_1, X_2, \cdots, X_n)$的密度为 $f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i);$

（2）若总体X离散，分布律为 $P(X=x)=p(x)$ ，则随机变量 $(X_1,X_2,\cdots X_n)$ 的联合分布律为

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

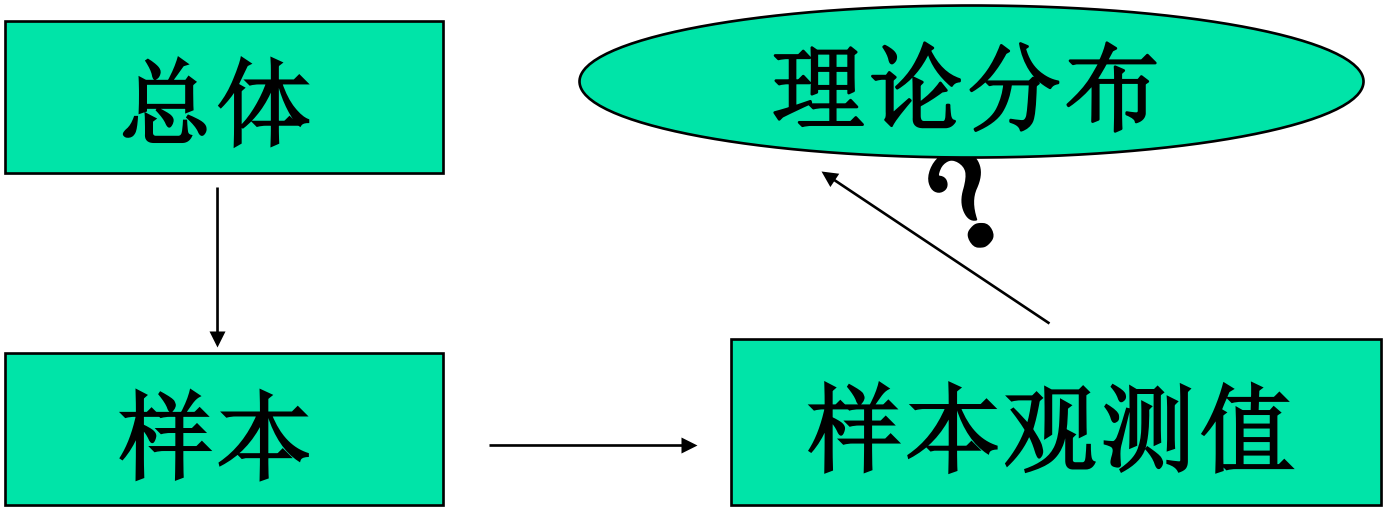
为了方便，我们把这个联合分布律也写为 $f(x_1,x_2,\cdots,x_n)$，称之为**联合概率函数**。从而，离散型和连续型有了统一的表达式。

例：若 $X\sim B(1,p)$，其分布律为 $P(X=x)=p(x)=p^xq^{1-x}$（x=0 或 1），则样本 $X_1,X_2,\cdots,X_n$ 的联合概率函数为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p(x_i)=\prod_{i=1}^{n}p^{x_i}q^{1-x_i}=p^{\sum_{i=1}^{n}x_i}q^{n-\sum_{i=1}^{n}x_i},$$

$x_1,x_2,\cdots,x_n=0,1;\quad p=1-q\in[0,1].$

总体、样本、样本观测值的关系



统计是从手中已有的资料---样本观测值,去推断总体的情况---总体分布.样本是联系两者的桥梁.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观测值的规律,因而可以用样本观测值去推断总体.

§7.2.1 χ^2 分布

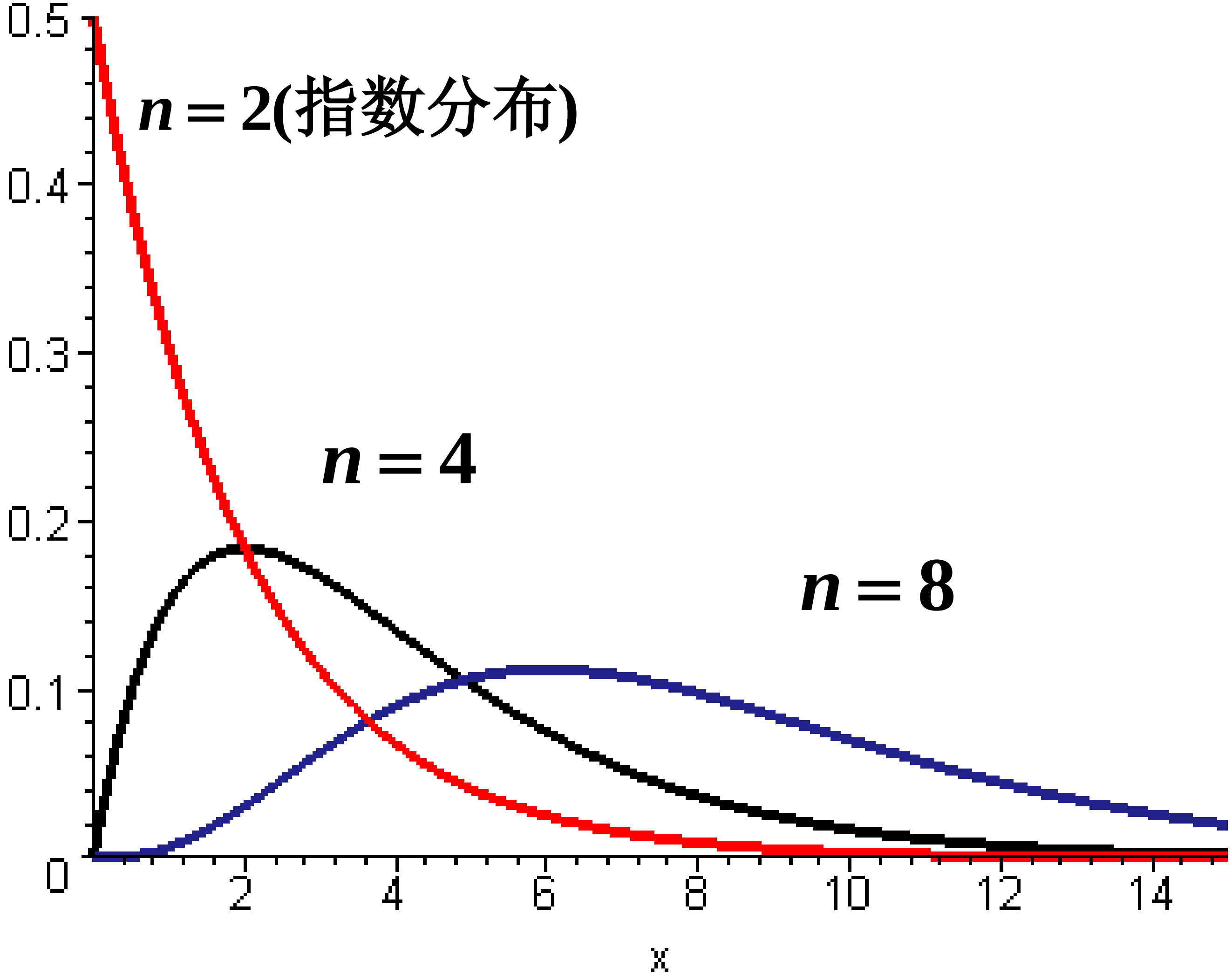
定义7.2 设$X_1, X_2, \ldots, X_n$是来自标准正态总体$N(0,1)$的样本,称随机变量$\chi^2=\sum_{i=1}^{n}X_i^2$所服从的分布为自由度为$n$的 χ^2 分布，记为

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

定理7.1 $\chi^2(n)$ 分布也为$\Gamma(\frac{n}{2},\frac{1}{2})$,即 $\chi^2(n)$ 有密度函数

$$f(x,n)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}},x>0$$

χ^2 分布的密度曲线



χ^2分布的可加性

定理7.2 设$\chi_1^2 \sim \chi^2(n), \chi_2^2 \sim \chi^2(m)$, 且 χ_1^2与χ_2^2相互独立, 则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n+m)$

证： 因为 $\chi_1^2 \sim \Gamma(\frac{n}{2}, \frac{1}{2}), \chi_2^2 \sim \Gamma(\frac{m}{2}, \frac{1}{2})$

且 χ_1^2与χ_2^2 相互独立,由 Γ 分布可加性，有

$$\chi_1^2 + \chi_2^2 \sim \Gamma(\frac{n+m}{2}, \frac{1}{2}) \Leftrightarrow \chi^2(n+m)$$

χ^2分布的期望与方差

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

例: $X \sim N(\mu,\sigma^2)$ (X_1,X_2,X_3)为X的一个样本

求 $\left(\frac{X_1-\mu}{\sigma}\right)^2+\left(\frac{X_2-\mu}{\sigma}\right)^2+\left(\frac{X_3-\mu}{\sigma}\right)^2$ 的分布.

§7.2.2 t分布

定义7.3 设随机变量 $X \sim N(0,1), Y \sim \chi^2(n)$

且X与Y相互独立，称随机变量

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

所服从的分布为自由度为n的t分布，

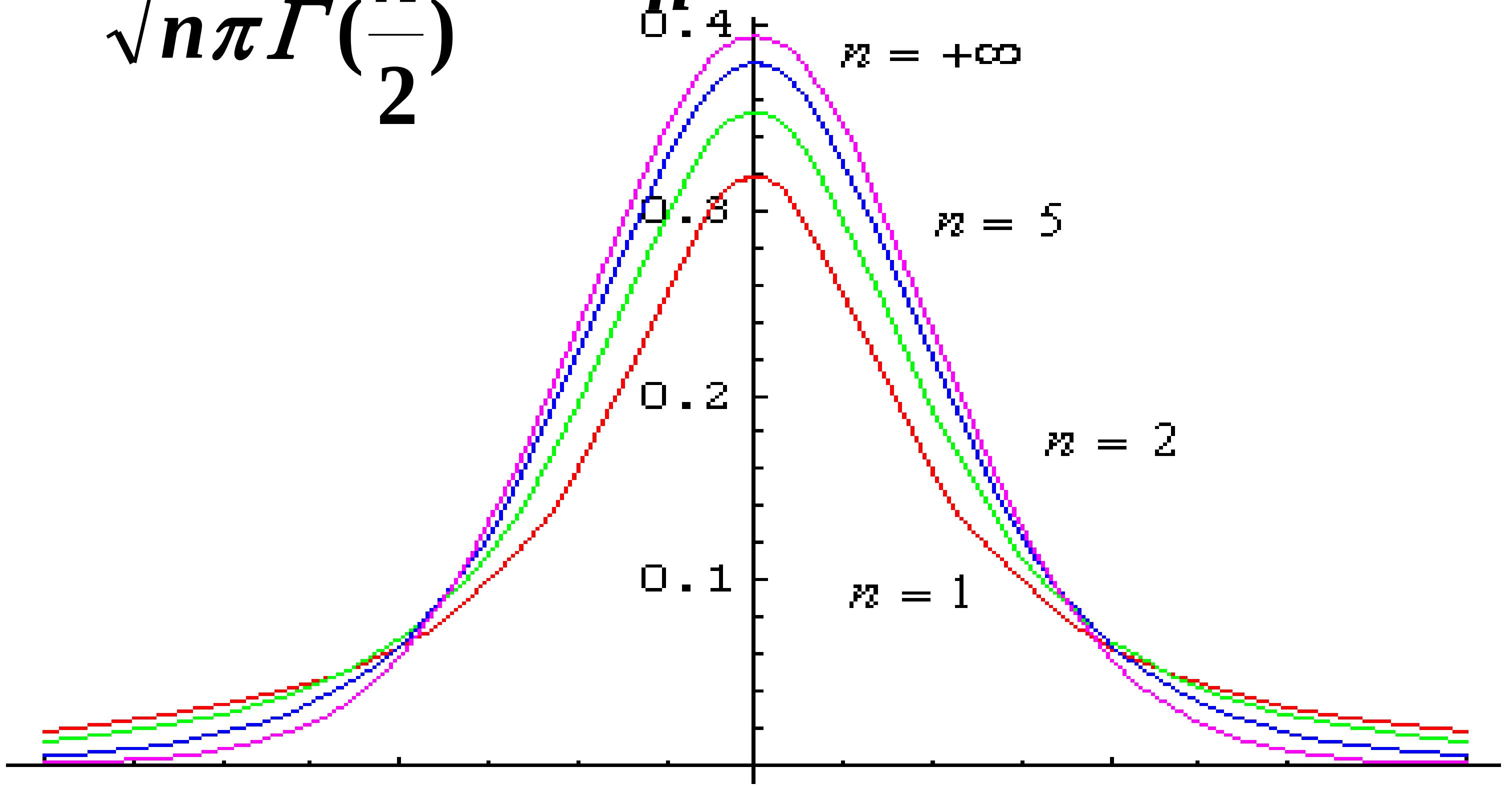
记为 $t \sim t(n)$

t分布的密度曲线关于纵坐标对称,可以证明

当n充分大时,t分布具有渐近正态性.

t分布的密度曲线

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



基本性质

(1) $t(x,n)$关于$x=0$(纵轴)对称

(2) $t(n)$的极限为$N(0,1)$的密度函数，即

$$\lim_{n\to\infty} t(x,n)=\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},-\infty<x<\infty$$

例

$X \sim N(\mu, \sigma^2)$ (X_1, X_2, X_3)为X的一个样本,求

$$\frac{\sqrt{2}(X_1-\mu)}{\sqrt{(X_2-\mu)^2+(X_3-\mu)^2}}$$ 的分布.

解 $\dfrac{X_1-\mu}{\sigma} \sim N(0,1)$ $\left(\dfrac{X_2-\mu}{\sigma}\right)^2+\left(\dfrac{X_3-\mu}{\sigma}\right)^2 \sim \chi^2(2)$

$$\frac{\dfrac{X_1-\mu}{\sigma}}{\sqrt{\dfrac{\left(\dfrac{X_2-\mu}{\sigma}\right)^2+\left(\dfrac{X_3-\mu}{\sigma}\right)^2}{2}}} \sim t(2)$$

§7.2.3 F分布

定义7.4 设随机变量 $X \sim \chi^2(n), Y \sim \chi^2(m)$

且X与Y相互独立，称随机变量

$$F = \frac{X/n}{Y/m}$$

所服从的分布为自由度为(n,m)的F分布，

其中n称为第一自由度，m称为第二自由度.

由F分布的定义，易见当

$$F \sim F(n,m)\text{时}, \frac{1}{F} \sim F(m,n)$$

例7.4 设 $X_1, X_2, \dots, X_n$ 为来自正态总体 $N(0,\sigma^2)$ 的样本，证明：1) $\sum_{i=1}^{8} X_i \Big/ \sqrt{\sum_{i=9}^{16} X_i^2} \sim t(8)$

$$2)\ \sum_{i=1}^{8} X_i^2 \Big/ \sum_{i=9}^{16} X_i^2 \sim F(8,8)$$

证： 1) $\because \sum_{i=1}^{8} X_i \sim N(0,8\sigma^2)$ $\quad \therefore \sum_{i=1}^{8} X_i \Big/ \sqrt{8}\sigma \sim N(0,1)$

又 $\sum_{i=9}^{16} X_i^2 / \sigma^2 \sim \chi^2(8)$ 由t分布定义即得

§7.2.4 分布的分位点

定义7.5 设X是随机变量，$0<p<1$，若实数 a_p 满足

$$F(a_p)=P(X\le a_p)=p$$

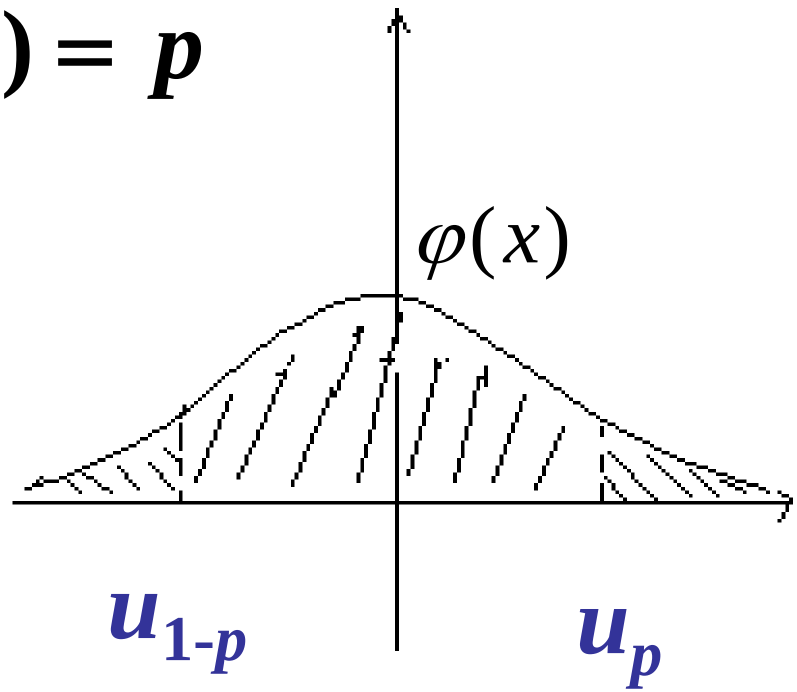
则称 a_p 为X的 p 分位点，当 $p=\dfrac{1}{2}$ 时，

$a_{1/2}$ 称为中位数.

$N(0,1)$分布的分位点u_p

$$\Phi(u_p) = P(X \le u_p) = p$$

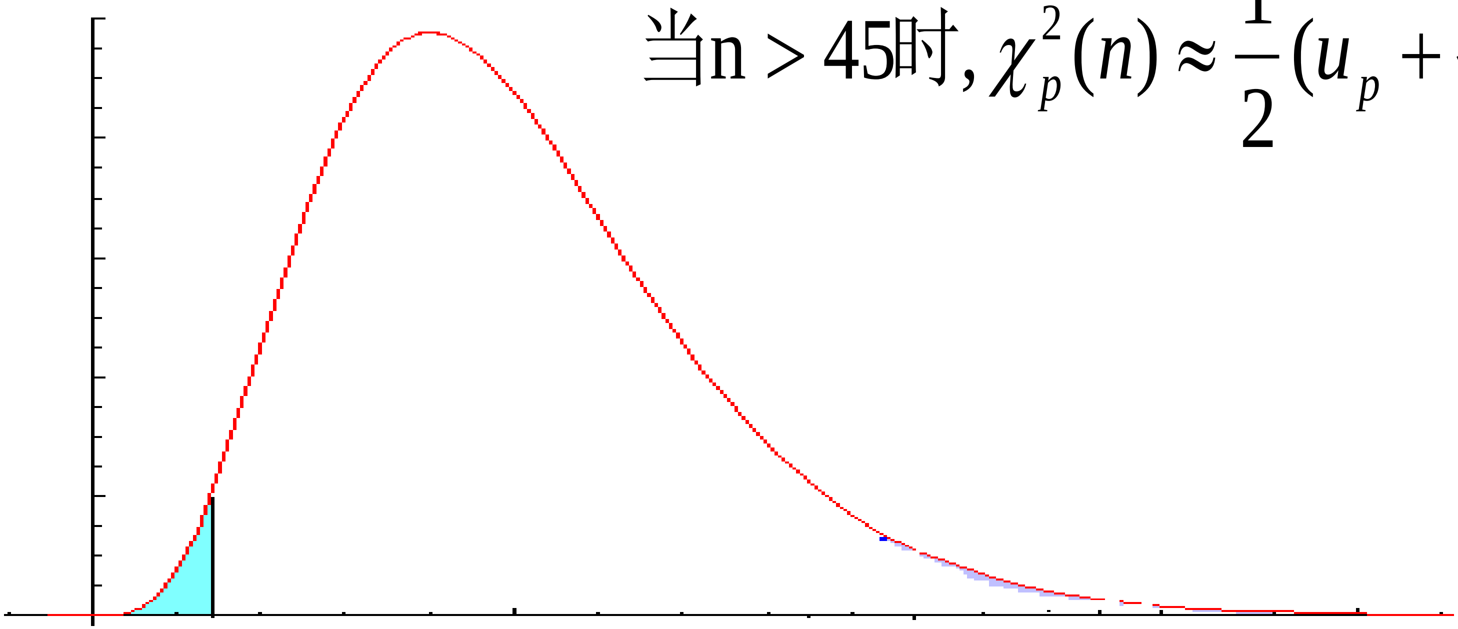
$$\Rightarrow u_{1-p} = -u_p$$



χ^2分布的分位点$\chi_p{}^2(n)$

$$P(X \le \chi_p^2(n)) = p$$

当n > 45时, $\chi_p^2(n) \approx \frac{1}{2}(u_p + \sqrt{2n-1})^2$



例

总体$X \sim N(1,4)$,抽取样本$(X_1,\ldots,X_n)$,

$$Y = \sum_{i=1}^{n}(X_i - 1)^2$$ 要$P(Y \le 100) \ge 0.95$,

n最大可以取多少?

解: $$\frac{X_i - 1}{2} \sim N(0,1) \Rightarrow \underbrace{\sum_{i=1}^{n}\left(\frac{X_i - 1}{2}\right)^2}_{Y/4} \sim \chi^2(n)$$

$$P\left(\frac{Y}{4} \le 25\right) = P(Y \le 100) \ge 0.95$$ 即是要$P(\chi^2(n) \le 25) \ge 0.95$

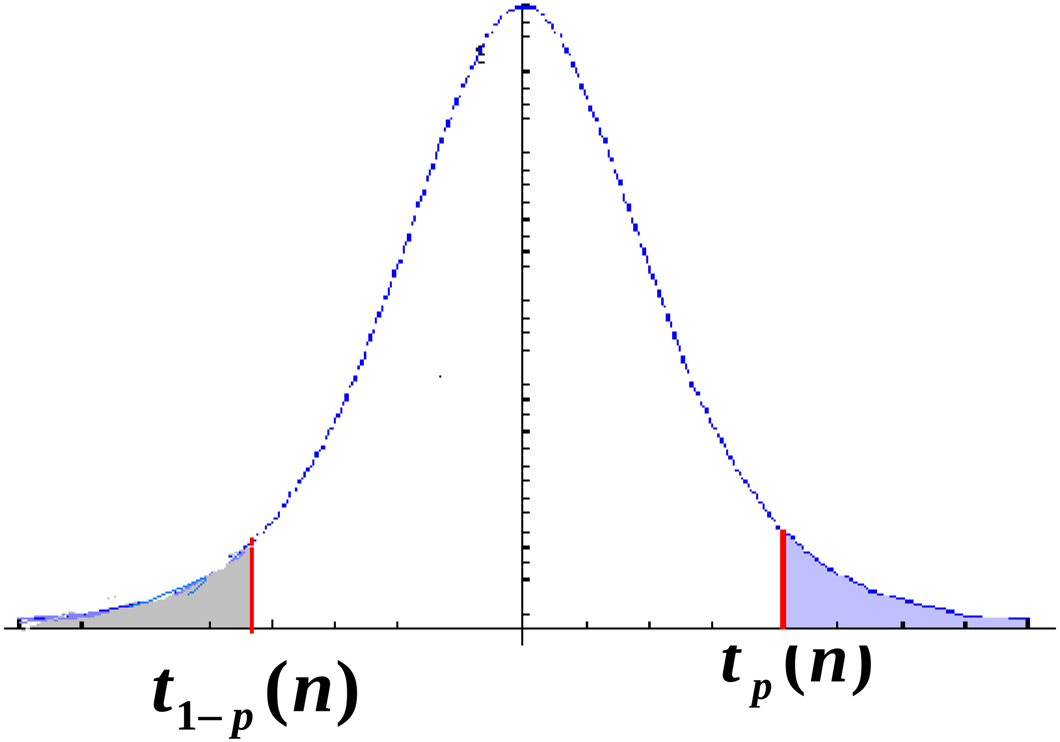
查χ^2分布表,使得$\chi^2_{0.95}(n) \approx 25$的$n = ?$ 故$n \le 15$

$t(n)$分布的分位点$t_p(n)$

$$P(t \le t_p(n)) = p$$

$$\Rightarrow t_{1-p}(n) = -t_p(n)$$

n>45时，用极限分布（正态）近似计算，即 $t_p(n) \approx u_p$

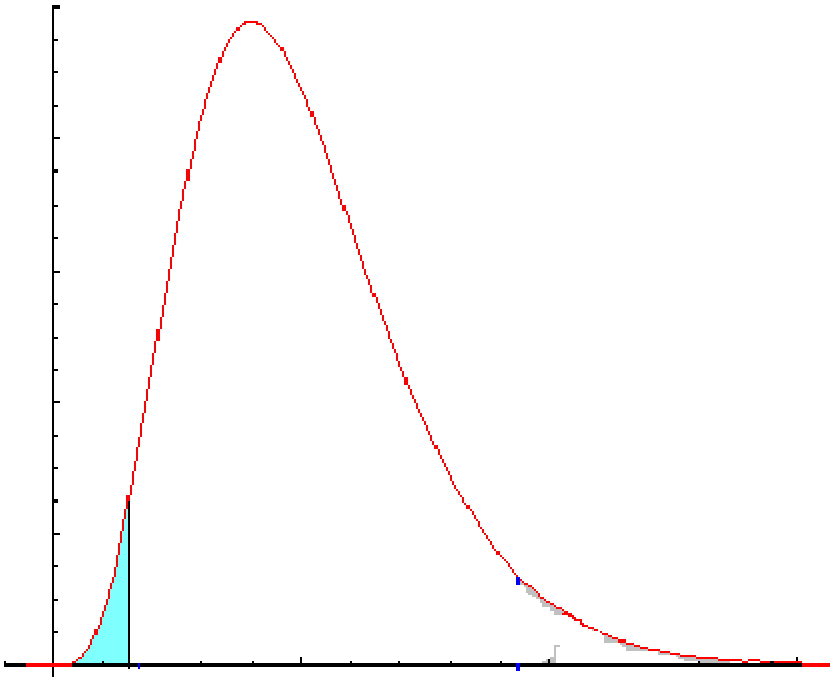


$F(n,m)$分布的分位点$F_p(n,m)$

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

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

利用它来计用$p<\frac{1}{2}$时的值



$F_p(n,m)$

§7.3.1 统计量

当人们观测样本时就可以得到一组样本值，但这些样本值是杂乱无章的。为了研究总体的规律，我们须对这些数据作进一步处理，进行“加工”和“提炼”,将分散于样本中的信息集中起来,通常是构造一相应的函数，这样的函数称为统计量。

定义7.6

样本 $X_1, X_2, ..., X_n$ 的一个连续函数 $g(X_1, X_2, \cdots, X_n)$ 称为一个**样本函数,**

若 $g(X_1, X_2, \cdots, X_n)$ 不含任何未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 为一个**统计量**，而代入样本观测值后 $g(x_1, x_2, \cdots, x_n)$ 叫 **统计量的观测值**.

构造统计量的目的是用它来推断总体.

例

$X \sim N(\mu,\sigma^2)$ μ,σ^2 未知，

$(X_1,X_2,\ldots,X_n)$为X的一个样本

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$， $\sum_{i=1}^{n} X_i^2$ 均为统计量

$\overline{X} - \mu$， $\frac{1}{\sigma^2}\sum X_i^2$ 不是统计量

常用的统计量

统计量名称	统计量	统计量观测值
样本均值	$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$
样本方差	$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}[\sum_{i=1}^{n}X_i^{\,2}-n\bar{X}^2]$	$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\,2}-n\bar{x}^2]$
样本标准差	$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$	$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

统计量名称	统计量	统计量观测值
样本k阶原点矩	$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^{\ k}$ $k=1,2,\cdots$	$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^{\ k}$ $k=1,2\cdots$
样本k阶中心矩	$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$ $k=1,2,\cdots$	$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$ $k=1,2\cdots$