第五章

数理统计中的统计量

及其分布

随机样本和经验分布函数

统计量

三大抽样分布

正态总体下常用统计量的一些重要结论

数理统计——以概率论为基础，主要研究如何*收集*、*整理*和*分析*实际问题的数据（**有限的资源**），以便对所研究的问题作出有效的（**精确而可靠**）结论。

☆基础——概率论

☆功能——处理数据

☆目的——作出科学推断（就概率特征）

5.1 总体与随机样本

总体—— 作为研究对象的随机变量

记作 $X, Y, \cdots, \quad \xi, \eta, \cdots$

样本—— 对总体进行n次试验所得到的结果

记作 $(X_1, X_2, \cdots, \quad X_n), (Y_1, Y_2, \cdots, \quad Y_n), \cdots$

注意：$X_1, X_2, \cdots, \quad X_n, Y_1, Y_2, \cdots, \quad Y_n$都是随机变量

样本容量—— n

样本观测值——样本 $(X_1, X_2, \cdots, \quad X_n)$的一组具体数值，记作$(x_1, x_2, \cdots, \quad x_n)$

简单随机样本—— $X_1, X_2, \cdots, \quad X_n$独立同分布

结论：设$X_1, X_2, \cdots, X_n$为来自总体X的一组样本，则

(1)若总体 X是离散型随机变量，概　　率分布为

$P\{X = x\}$, 则$X_1, X_2, \cdots, X_n$的联合概率分布为

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

(2)若总体X是连续型随机变量，概率密度为$p(x)$,

则$X_1, X_2, \cdots, X_n$的联合概率密度为

$$p^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i)$$

(3)若总体X的分布函数为$F(x)$, 则$X_1, X_2, \cdots, X_n$

的联合分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

用样本估计总体的分布

——数理统计的主要任务之一。

经验分布函数

若总体X，样本观测值为 $x_1, x_2, \cdots, x_n$

将观测值从小到大排列：

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

则由大数定理，取值 $x_{(i)}$的概率$P\{X = x_{(i)}\}$可以用取值$x_{(i)}$的频率来估计。

并写出频率分布表：

观测值 $x_{(i)}$	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(l)}$
频数 m_i	m_1	m_2	$\cdots$	m_l
频率 $\omega_i = \dfrac{m_i}{n}$	ω_1	ω_2	$\cdots$	ω_l

其中，$x_{(1)} < x_{(2)} < \cdots < x_{(l)}$，$\sum_{i=1}^{l} m_i = n$，$\sum_{i=1}^{l} \omega_i = 1$

经验分布函数如下：

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)}; \\ \sum\limits_{x_{(i)} < x} \omega_i, & \text{当 } x_{(i)} \le x < x_{(i+1)}; \\ 1, & \text{当 } x \ge x_{(l)}. \end{cases}$$

★经验分布函数 $F_n(x)$ 的性质：

(1) $0 \le F_n(x) \le 1$

(2) $F_n(x)$ 是非减函数

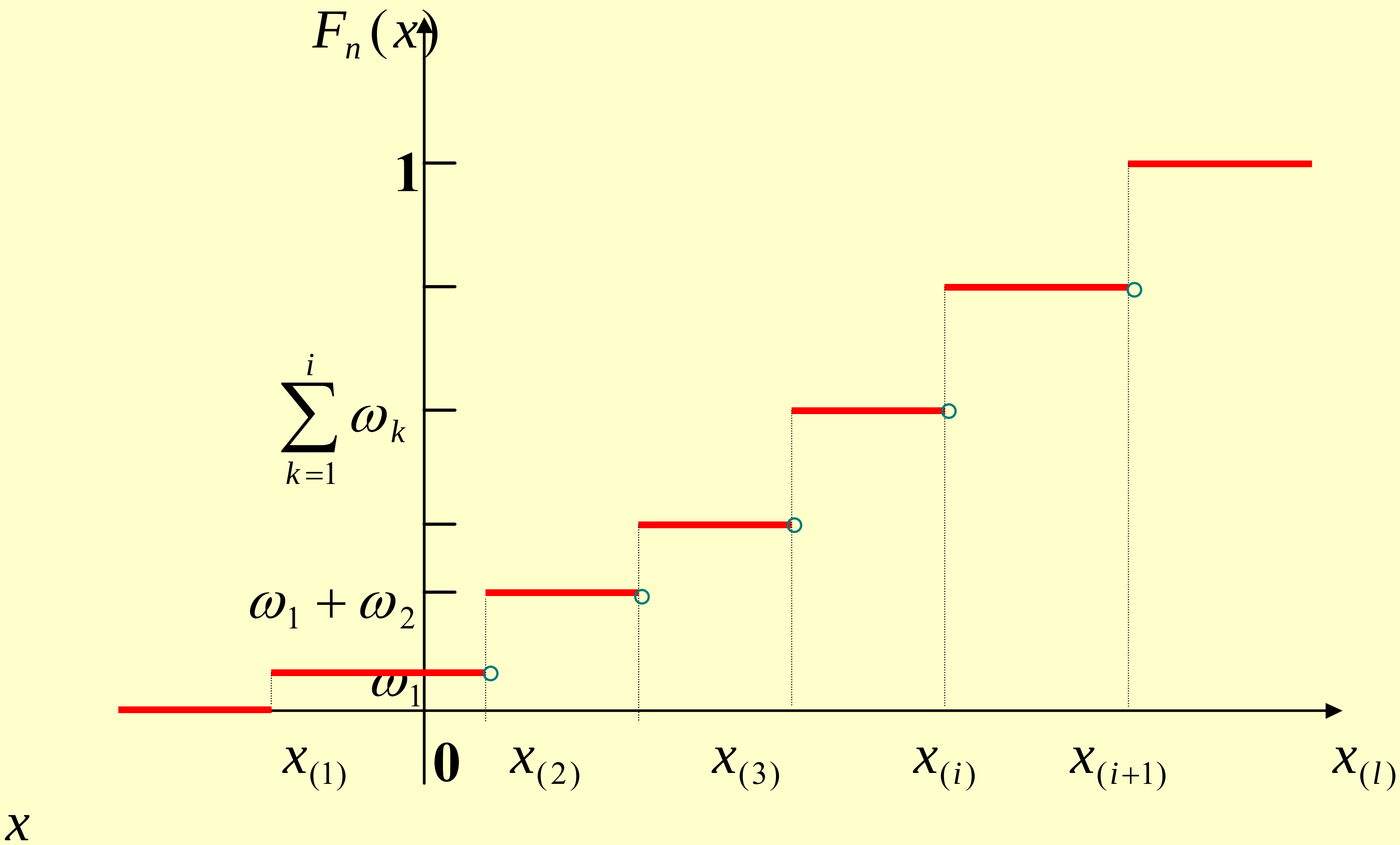
(3) $F_n(-\infty) = 0, F_n(+\infty) = 1$

(4) $F_n(x)$ 在每个观测点 $x_{(i)}$ 处是右连续的，点 $x_{(i)}$ 是 $F_n(x)$ 的跳跃间断点，$F_n(x)$ 在点 $x_{(i)}$ 处的跳跃度就等于频率 ω_i。

★经验分布函数 $F_n(x)$ 是事件 $\xi \le x$ 的频率；
总体分布函数 $F(x)$ 是事件 $\xi \le x$ 的概率。
则当 $n \to \infty$ 时，$\boldsymbol{P\{\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0\} = 1}$

！这是我们在数理统计中*用样本推断总体*的理论基础。

图形特点：右连续，台阶形



2. 统计量

定义．设$(X_1, X_2, \cdots, X_n)$为总体的样本，称T为统计量，若其满足两条要求：

(1)T为样本的函数，即$T = T(X_1, X_2, \cdots, X_n)$；

(2)T的表达式中不含有任何未知参数。

注意: 1.对于样本提供的信息要进行提炼。

2.样本的函数中不包含任何未知参数，是为了推断的可行性。

3.“统计量”与后面 的“枢轴量”不同。

统计量是随机变量.

常用的统计量

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

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

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

有的书中定义为： $$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

$$=\frac{n-1}{n}S_{n-1}^2$$

3. 样本标准差：

$$\sigma_{n-1} = S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

4. 样本k阶原点矩：

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

5. 次序统计量：

将样本的各个分量$X_1, X_2, \cdots, X_n$按从小到大的次序排列，得到$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$，常称$X_{(i)}$为样本的第$i$个“次序统计量”。特别地，$X_{(1)} = \min\limits_{1\le i\le n} X_i$称为最小次序统计量，$X_{(n)} = \max\limits_{1\le i\le n} X_i$称为最大次序统计量。

6.样本中位数:

$$M_e \overset{\Delta}{=} \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n\text{为奇数} \\ \frac{1}{2}\left[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right], & n\text{为偶数} \end{cases}$$

7.极差:

$$R \overset{\Delta}{=} X_{(n)} - X_{(1)} = \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i$$

与第二章中描述统计中不同的是这里强调了“随机性”。例如对应于（随机的）样本均值 $\overline{X}(\omega) \overset{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega)$ (强调概率分布)，第二章中所给出的 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 可以认为是样本观测值的平均(注重数值)。

（1）使用计算器计算统计量的值。

（2）使用**EXCEL**计算统计量的值。

样本均值（**AVERAGE**）

样本方差（**VAR**）

样本标准差（STDEV）

菜单中的“工具\数据分析\描述统计” 详见**P37**

例1. 已知样本观测值为**15.8**，**24.2**，**14.5**，**17.4**，**13.2**，**20.8**，**17.9**，**19.1**，**21.0**，**18.5**，**16.4**，**22.6**。计算样本平均值、样本方差。

3. 三大抽样分布

除正态分布外，最著名的就是“ χ^2 分布”、“ t 分布”与“ F 分布”，它们被称为数理统计中的“三大抽样分布”。

1. χ^2 分布（卡方分布）

构造定理：设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，并且都服从标准正态分布 $N(0,1)$，则随机变量

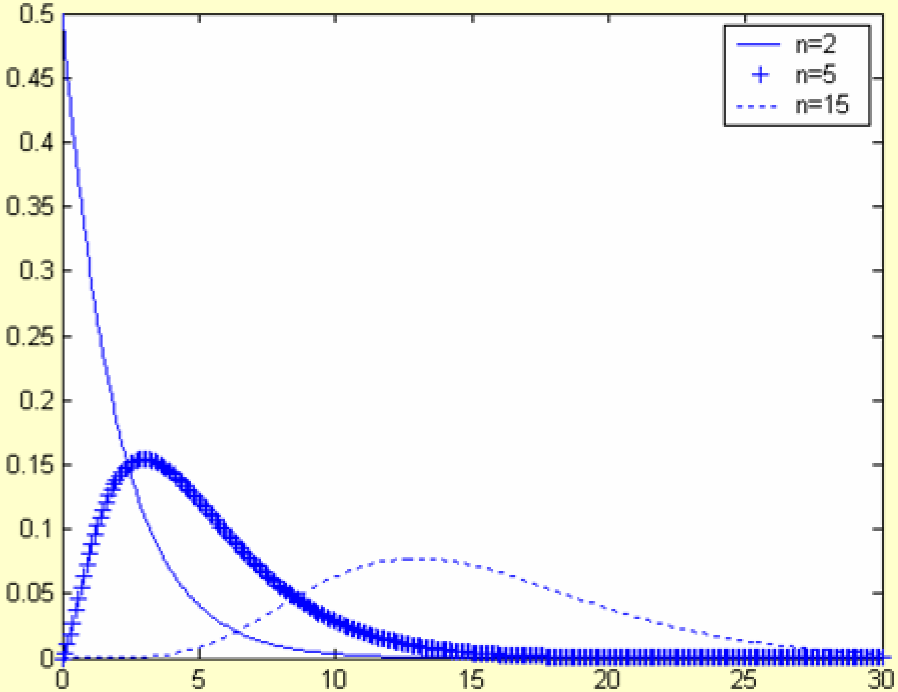
$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从自由度为 n 的 χ^2 分布，记为 $\chi^2(n)$。

自由度——独立随机变量的个数。

χ^2 分布的密度函数：

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



$p(x)$的性质：

(1) $x \to +\infty$时, $p(x) \to 0$

(2) $x = n-2$时, $p(x)$取到最大值。

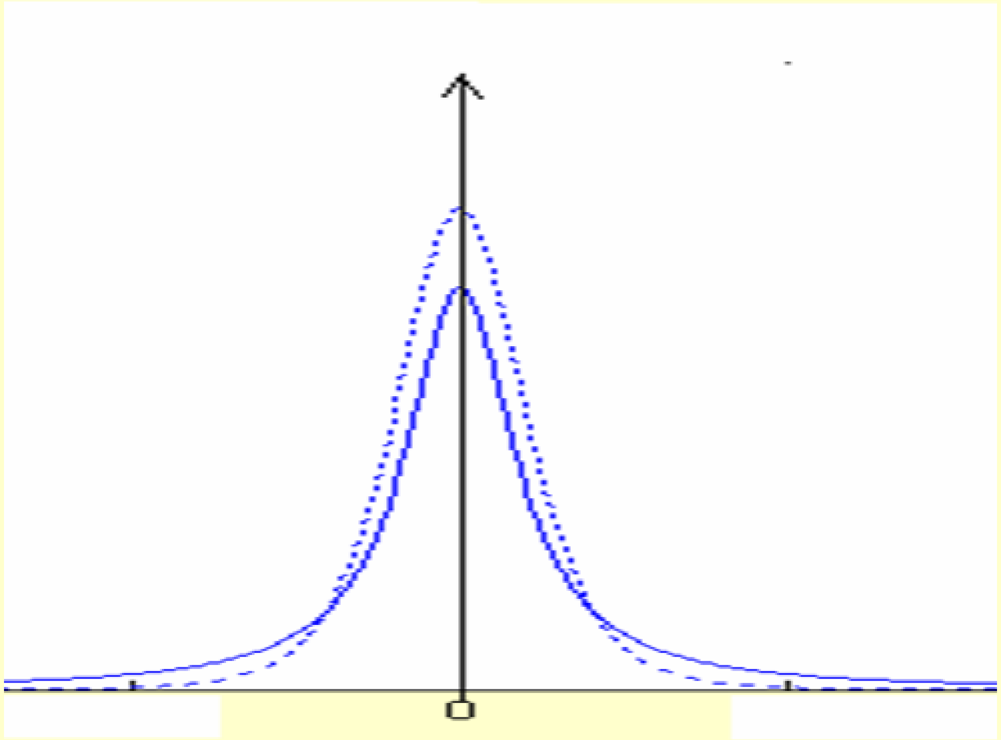
χ^2分布的性质：设$\xi \sim \chi^2(k_1)$，$\eta \sim \chi^2(k_2)$，ξ与η相互独立，则$\xi+\eta \sim \chi^2(k_1+k_2)$。

表7　χ^2分布的临界值（$P.265$）

2. t 分布（学生分布）

构造定理： 设随机变量 X 与 Y 相互独立，并且 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，则随机变量 $T = \dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ 服从自由度为 n 的 t 分布，记为 $t(n)$。

密度函数：$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$



$p(x)$的性质：

$(1)\, x \to \pm\infty$时，

$p(x) \to 0$

$(2)\, x = 0$时，$p(x)$

取到最大值。

$(3)\, p(x)$关于$x = 0$对称。

$(4)\, n \to \infty$时，$p(x) \to \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$(标准正态）

表6　t分布的临界值（P.264）

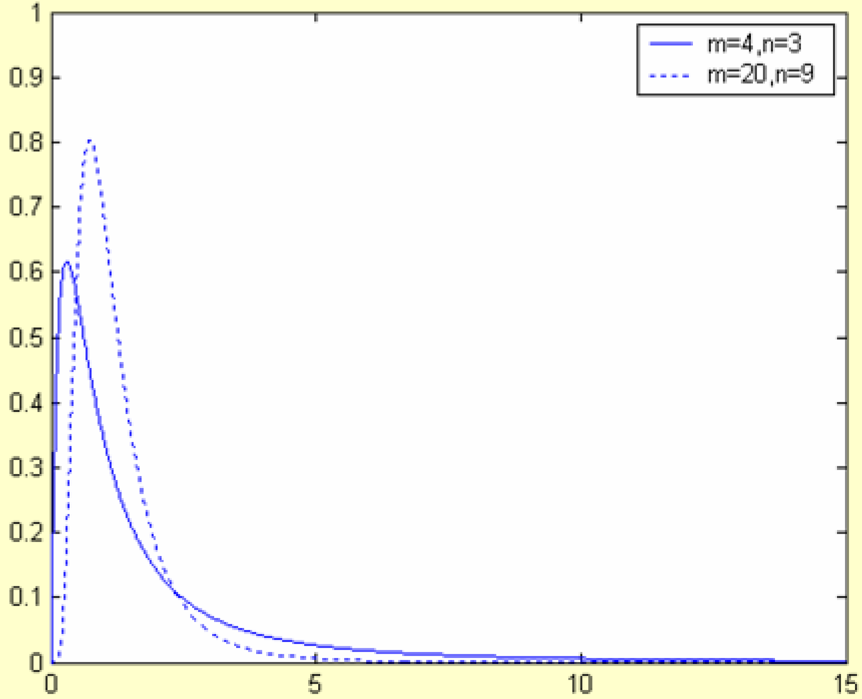
4. F 分布

构造定理： 设随机变量 X 与 Y 相互独立，并且 $X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，则随机变量 $F = \dfrac{X/m}{Y/n}$ 服从自由度为 (m,n) 的 F 分布，记为 $F(m,n)$。

$$\text{密度函数}: p(x) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, & x > 0\text{时} \\ 0, & x \le 0\text{时} \end{cases}$$

F **分布的性质：** $F_{1-\alpha}(m,n) = \dfrac{1}{F_{\alpha}(n,m)}$

表7　F分布的临界值（$P.311$）



F 分布的性质：$F_{1-\alpha}(m,n)=\dfrac{1}{F_{\alpha}(n,m)}$

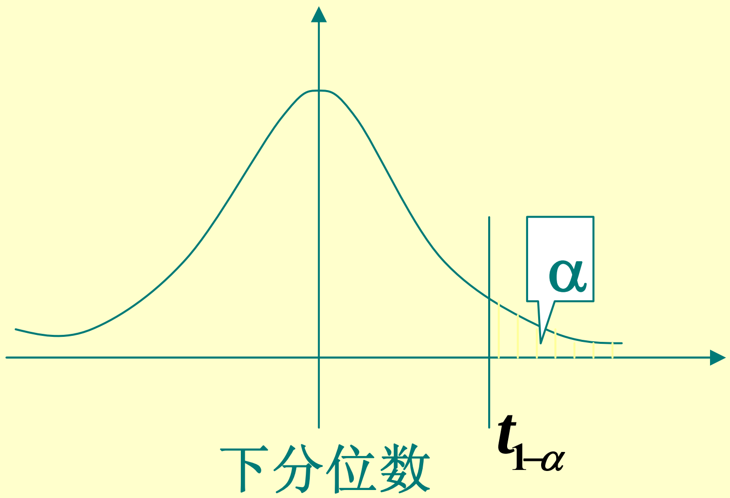
表8　*F*分布的临界值（*P*.266）

分位数

当随机变量 $T \sim t(n)$ 时，称满足关系式：

$$P(T \le t_{1-\alpha}(n)) = 1 - \alpha \text{ ,}$$

的数 $t_{1-\alpha}(n)$ 为自由度为 n 的 t 分布的“$1-\alpha$ 临界值”或“$1-\alpha$ 分位数”。通常可以通过查表找出临界值 $t_{1-\alpha}(n)$，例如当 $n=5, \alpha=0.05$ 时，查 t 分布表可得 $t_{1-0.05}(5) = 2.015$ 。



使用EXCEL计算分位数

t-分布（TINV）双侧分位数

χ^2 -分布（CHINV）上分位数

F-分布（FINV）上分位数

例如:

CHINV （0.05， 10）=18.30703 等价于 $\chi^2{}_{0.95}(10)$

$TINV(0.05, 10) = 2.2281$,等价于 $t_{0.975}(10)$

$FINV(0.05, 6, 3) = 8.9406$,等价于 $\boldsymbol{F_{0.95}(6, 3)}$

4. 正态总体下常用统计量的一些重要结论

假设总体服从正态分布。

定理1 设$(X_1, X_2, \cdots, X_n)$是取自正态总体$N(\mu, \sigma^2)$的样本，

其样本均值和样本方差分别为 $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ 和

$S_{n-1}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$,则有如下结论：

（1） $\overline{X}$ 与 S_{n-1}^2 相互独立；

(2） $\overline{X} \sim N(\mu, \sigma^2 / n)$；

（3） $\dfrac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2(n-1)$。

推论：在定理 5.4.1 的条件下，有 $\dfrac{\overline{X}-\mu}{S_{n-1}}\sqrt{n}\sim t(n-1)$

证：$\because \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1),\quad \dfrac{(n-1)S_{n-1}^2}{\sigma^2}\sim \chi^2(n-1)$

又 $\because \overline{X}$ 和 S_{n-1}^2 相互独立

$$\therefore \frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S_{n-1}^2}{\sigma^2}\Big/(n-1)}}=\frac{\overline{X}-\mu}{S_{n-1}}\sqrt{n}\sim t(n-1)$$

例 2. 设随机变量 X 和 Y 都服从标准正态分布，则

（A） $X+Y$ 服从正态分布　（B） X^2+Y^2 服从 χ^2 分布

（C） X^2 和 Y^2 都服从 χ^2 分布（D） X^2/Y^2 服从 F 分布

分析：如果加上 X，Y 相互独立的条件四个选项都对.

取 $Y=-X$ 可排除（A），（B），（D）。

答案：（C）

例 3. 设 X_1, X_2, X_3, X_4 是来自正态总体 $N(0,2^2)$ 的简单样本， $X = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$，则当 $a =$ ________ ，$b =$ ______ 时，统计量 X 服从 χ^2 分布，其自由度为 _____ 。

分析：由于 X 服从 χ^2 分布，可取 $\sqrt{a}(X_1 - 2X_2) \sim N(0,1)$，$\sqrt{b}(3X_3 - 4X_4) \sim N(0,1)$，

$$E\sqrt{a}(X_1 - 2X_2) = 0 \qquad E\sqrt{b}(3X_3 - 4X_4) = 0$$

$$D\sqrt{a}(X_1 - 2X_2) = a(4 + 4\times 4) = 20a$$

$$D\sqrt{b}(3X_3 - 4X_4) = b(9\times 4 + 16\times 4) = 100b$$

故有 $a = \dfrac{1}{20}, b = \dfrac{1}{100}, X \sim \chi^2(2)$ 。