第 6 章　大样本 OLS

6.1　为何需要大样本理论

大样本理论(large sample theory)，也称**渐近理论**(asymptotic theory)，研究当样本容量n趋向无穷大时统计量的性质。

大样本理论已成为当代计量经济学的主流方法，原因如下。

(1) 小样本理论的假设过强。首先，小样本理论的严格外生性假设要求解释变量与所有的扰动项均正交(不相关)。在时间序列模型中，这意味着解释变量与扰动项的过去、现在与未来值全部正交！

例　一阶自回归模型(first order autoregression，简记 AR(1))：

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (t = 2, \cdots, T) \qquad (6.1)$$

解释变量y_{t-1}为被解释变量y_t的一阶滞后；且$\mathrm{Cov}(y_{t-1}, \varepsilon_t) = 0$。

严格外生性要求，解释变量y_{t-1}与所有$\{\varepsilon_2, \cdots, \varepsilon_T\}$均不相关。

这意味着，y_t也不与ε_t相关。

根据自回归方程(6.1)，ε_t是y_t的一部分，故二者一定相关，因为

$$\mathrm{Cov}(y_t, \varepsilon_t) = \mathrm{Cov}\left[(\rho y_{t-1} + \varepsilon_t), \varepsilon_t\right] = \rho \underbrace{\mathrm{Cov}(y_{t-1}, \varepsilon_t)}_{=0} + \mathrm{Var}(\varepsilon_t) = \mathrm{Var}(\varepsilon_t) >$$

(6.2)

因此，以被解释变量滞后值为解释变量的自回归模型，必然违背严格外生性的假定。

大样本理论只要求解释变量与同期(同方程)的扰动项不相关。

其次，小样本理论假定扰动项为正态分布，而大样本理论无此限制。

在很多情况下，我们并无把握经济变量是否服从正态分布。比如，正态分布为对称分布，但许多经济变量的分布并不对称，例如工资收入。

即使考虑比较对称的工资对数，由于正态变量的取值范围为$(-\infty,+\infty)$，而工资对数一般为正数(假设工资大于 1)，故也不相符。

作为示例，下面将数据集 grilic.dta 的工资与工资对数的核密度图画在一起，结果参见图 6.1。

```
. use grilic.dta,clear
. gen wage=exp(lnw)
. twoway kdensity wage,xaxis(1) yaxis(1) xvarlab(wage) || kdensity lnw,xaxis(2) yaxis(2) xvarlab(ln(wage)) lp(dash)
```

选择项“xaxis(1) yaxis(1)”与“xaxis(2) yaxis(2)”指定对于变量 *wage* 与 *lnw* 分别使用不同的x轴与y轴，因为这两个变量的取值范围与概率密度均很不相同；

选择项“xvarlab(wage)”与“xvarlab(ln(wage))”将变量 *wage* 与 *lnw* 核密度图的横轴标签分别指定为“wage”与“ln(wage)”。

从图 6.1 可知，工资的分布与正态分布相去甚远；即使工资对数，在取值范围为$(-\infty, +\infty)$这一点上，也与正态分布不符。

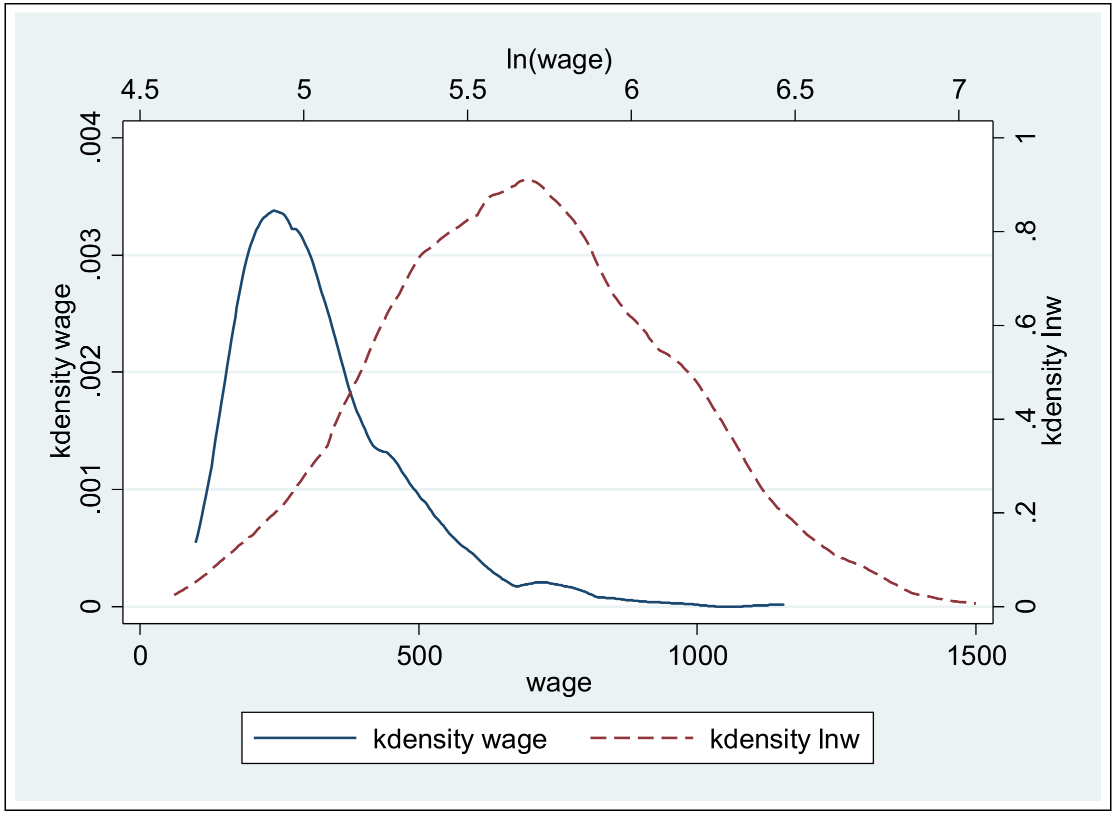

图 6.1 工资与工资对数的分布

被解释变量的分布可能为各种形状；有时即使取对数也不能使其接近正态分布。

继续以数据集 grilic.dta 为例，将教育年限(s)与其对数(lns)的核密度图画在一起，结果参见图 6.2。

```
. gen lns=log(s)
```

```
. twoway kdensity s,xaxis(1) yaxis(1) xvarlab(s) || kdensity lns,xaxis(2) yaxis(2) xvarlab(lns) lp(dash)
```

教育年限的分布呈现“双峰”形状，多数人为中学或大学毕业。

这种双峰形状，即使取对数后，也难以改变。

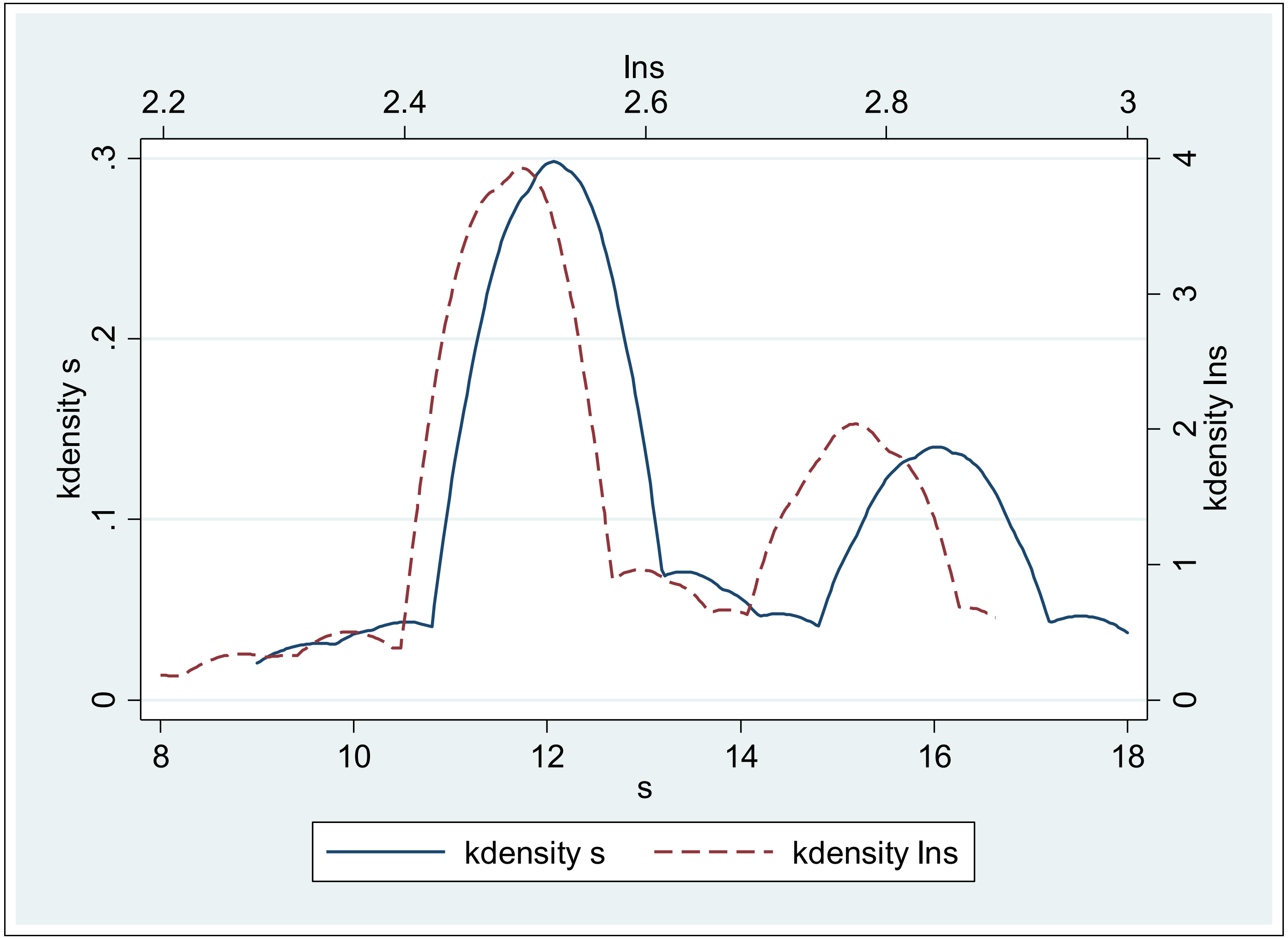

图 6.2　教育年限 *s* 与其对数 *lns* 的分布

无论教育年限还是其对数的分布，都与“单峰”的正态分布相去甚远。

这也说明，通过取对数使得变量的分布接近于正态并非万能。

对于小样本理论来说，为了进行统计推断(比如，推导t统计量与F统计量的有限样本分布)，必须假设扰动项服从正态分布(故被解释变量也服从正态分布)。

由于现实中的被解释变量可能服从各种分布(比如，变量婚否 mrt 为离散的两点分布)，故基于正态假设的小样本理论的适用范围受到很大限制。

(2) 在小样本理论的框架下，必须研究统计量的精确分布(exact distribution)，但常难以推导(即使在正态分布的假设之下)。

根据大样本理论，只要研究统计量的大样本分布，即当$n \to \infty$时的渐近分布，相对比较容易推导(可使用大数定律与中心极限定理)。

(3) 使用大样本理论的代价是要求样本容量较大，以便大数定律与中心极限定理可以起作用。

大样本理论对于样本容量的要求，一般认为至少$n \geq 30$，最好在 100 以上。

现代的数据集越来越大，经常成百上千(对于大数据，样本容量可能过亿)，使得渐近理论成为对统计量真实分布的很好近似。

6.2 随机收敛

1. 确定性序列的收敛

定义 确定性序列$\{a_n\}_{n=1}^{\infty}=\{a_1,a_2,a_3,\cdots\}$收敛(converge)于常数$a$，记为$\lim_{n\to\infty}a_n=a$或$a_n\to a$，如果对于任意小的正数$\varepsilon>0$，都存在$N>0$，只要$n>N$，就有$|a_n-a|<\varepsilon$，即在$a_N$以后的序列$\{a_{N+1},a_{N+2},\cdots\}$均落入区间$(a-\varepsilon,a+\varepsilon)$内，参见图 6.3。

此定义意味着，无论区间$(a-\varepsilon,a+\varepsilon)$多么小，如果去看序列$\{a_n\}_{n=1}^{\infty}$ “足够后面” (比如$n>N$)的项，则就会都落入此区间内。

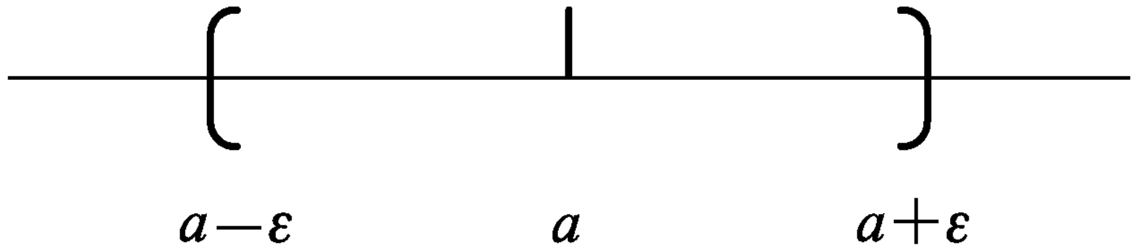

图 6.3　确定性序列的收敛

例 假设$a_n = 5 + \frac{1}{n}$，则$\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty}(5 + \frac{1}{n}) = 5 + \lim\limits_{n\to\infty}\frac{1}{n} = 5 + 0 = 5$。

2．随机序列的收敛

考虑随机序列$\left\{x_n\right\}_{n=1}^{\infty} = \left\{x_1, x_2, x_3, \cdots\right\}$，即由随机变量构成的序列，其中每个元素$x_n$都是随机变量，而下标$n$通常表示样本容量。

定义　随机序列$\{x_n\}_{n=1}^{\infty}$**依概率收敛**(converge in probability)于常数 a，记为$\operatorname*{plim}_{n\to\infty} x_n = a$，或$x_n \xrightarrow{p} a$，如果对于任意$\varepsilon > 0$，当$n \to \infty$时，都有$\lim_{n\to\infty} \mathrm{P}\left(\left|x_n - a\right| > \varepsilon\right) = 0$。

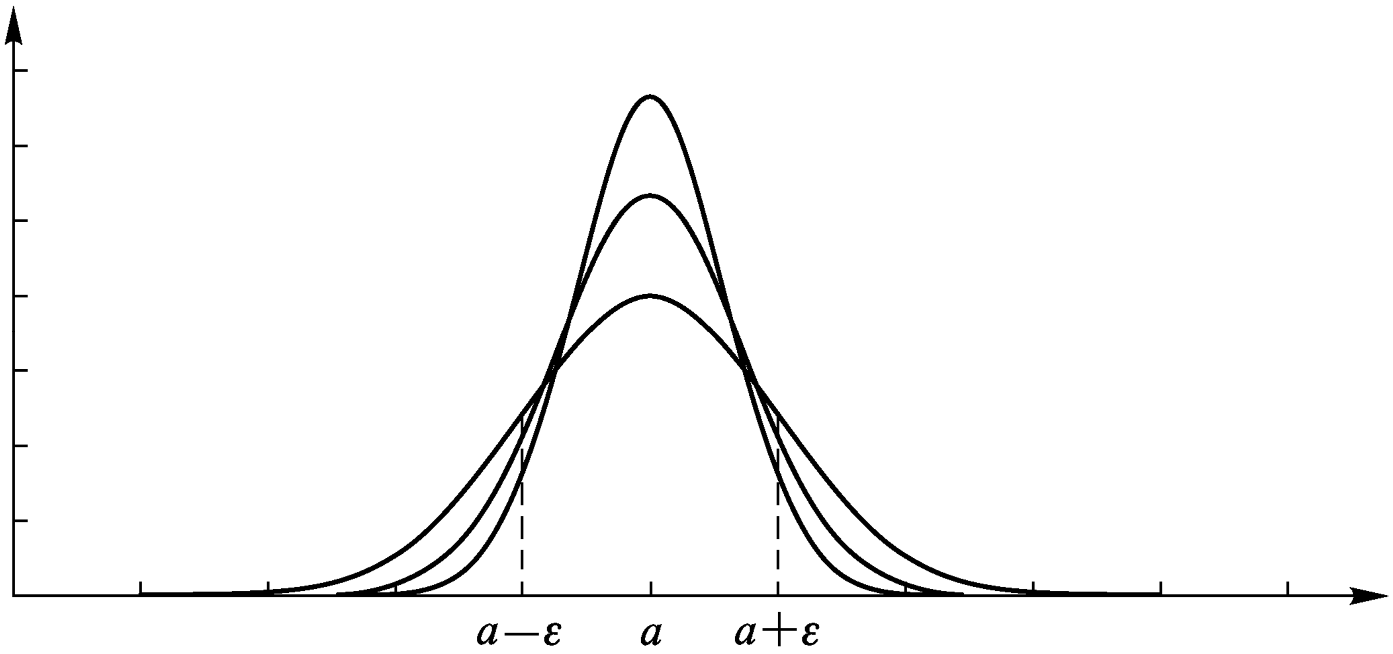

图 6.4　随机序列的收敛

这意味着，任意给定很小的正数$\varepsilon > 0$，当n越来越大时，随机变量x_n落在区间$(a-\varepsilon, a+\varepsilon)$之外的概率收敛于 0，参见图 6.4。

当n变大时，x_n远离常数 a 的可能性越来越小，变得几乎不可能。

虽然x_n为随机变量，各种取值都有可能，但当$n \to \infty$时，x_n的取值可能性越来越集中于 a 附近。

由于已将随机事件$\left(\left|x_n - a\right| > \varepsilon\right)$取概率，故$\mathrm{P}\left(\left|x_n - a\right| > \varepsilon\right)$其实是确定性序列。

$\lim\limits_{n\to\infty} \mathrm{P}\left(\left|x_n - a\right| > \varepsilon\right)$只是普通的微积分极限。

例 假设x_n服从如下两点分布：

$$x_n=\begin{cases}0 & \text{取值概率 } 1-1/n \\ n & \text{取值概率 } 1/n\end{cases} \qquad (6.3)$$

随着$n\to\infty$，x_n的分布越来越集中于 0，而取值为n的可能性越来越小(尽管n离 0 越来越远)。

故根据定义，$\operatorname*{plim}_{n\to\infty} x_n = 0$。

利用随机变量依概率收敛于常数的概念，可定义随机变量之间的随机收敛，只要随机变量之差别依概率收敛于 0。

定义 随机序列$\left\{x_n\right\}_{n=1}^{\infty}$**依概率收敛**于随机变量$x$，记为$x_n \xrightarrow{p} x$，如果随机序列$\left\{x_n - x\right\}_{n=1}^{\infty}$依概率收敛于 0。

概率收敛($\operatorname*{plim}_{n\to\infty}$)的运算规则类似于微积分中极限($\lim_{n\to\infty}$)的运算。比如，假设$g(\cdot)$为连续函数，则

$$\operatorname*{plim}_{n\to\infty} g(x_n) = g\left(\operatorname*{plim}_{n\to\infty} x_n\right) \qquad (6.4)$$

此结论称为**连续映射定理**(Continuous Mapping Theorem)。

概率极限$\operatorname*{plim}_{n\to\infty}$与连续函数$g(\cdot)$可交换运算次序。

无论先用函数$g(\cdot)$去作用x_n，再取概率极限；还是先对x_n取概率极限，再用函数$g(\cdot)$去作用，二者的效果是一样的。

当x_n的分布越来越集中于$x^* \equiv \operatorname*{plim}_{n\to\infty} x_n$附近时，$g(x_n)$的分布自然也就越来越集中于$g(x^*)$附近。

但期望算子$\mathrm{E}(\cdot)$无此性质。

一般来说，$\mathrm{E}(x^2) \neq \left[\mathrm{E}(x)\right]^2$。

这正是大样本理论的方便之处。

例　如果$\operatorname*{plim}_{n\to\infty} s^2=\sigma^2$(样本方差依概率收敛于总体方差)，则样本标准差$s$也依概率收敛于总体标准差$\sigma$，因为

$$\operatorname*{plim}_{n\to\infty} s=\operatorname*{plim}_{n\to\infty}\sqrt{s^2}=\sqrt{\operatorname*{plim}_{n\to\infty} s^2}=\sqrt{\sigma^2}=\sigma \qquad (6.5)$$

其中，“开根号” ($\sqrt{\ }$)是连续函数，故可与求概率极限的运算交换次序。

对于随机向量序列(即序列中每个元素都是随机向量)，也可类似地定义依概率收敛，只要定义其每个分量都依概率收敛即可。

比如，随机向量序列$\left\{\mathbf{x}_n\right\}_{n=1}^{\infty}=\left\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\cdots\right\}$依概率收敛于随机向量$\mathbf{x}$，意味着$\mathbf{x}_n$的每个分量都依概率收敛至$\boldsymbol{\beta}$的相应分量，记为$\underset{n\to\infty}{\operatorname{plim}}\mathbf{x}_n=\mathbf{x}$。

3．依均方收敛

定义　如果随机序列$\left\{x_n\right\}_{n=1}^{\infty}$的期望收敛于$a$，即$\lim\limits_{n\to\infty}\mathrm{E}(x_n)=a$；而方差收敛于 0，即$\lim\limits_{n\to\infty}\mathrm{Var}(x_n)=0$，则称$\left\{x_n\right\}_{n=1}^{\infty}$**依均方收敛**(converge in mean square)于常数a，记为$x_n\xrightarrow{ms}a$。

通过切比雪夫不等式，可以证明(参见附录)，依均方收敛意味着依概率收敛。

当x_n的均值越来越趋于 a，而方差越来越小并趋于 0 时，就有$\operatorname*{plim}_{n\to\infty} x_n = a$，即在极限处$x_n$退化为常数 a。

证明均方收敛通常比证明概率收敛更容易，故可通过证明前者来证明后者，这也是依均方收敛概念的主要用途之一。

反之，依概率收敛并不意味着均方收敛。

例 回到$\{x_n\}$服从两点分布的例子，即x_n取值为 0 的概率为 1-1/n，而取值为 n 的概率为1/n。虽然x_n依概率收敛到 0，但x_n并不依均方收敛到 0，因为此序列的期望恒等于 1：

$$\lim_{n\to\infty} \mathrm{E}(x_n) = \lim_{n\to\infty}\left[0\cdot\left(1-\frac{1}{n}\right)+n\cdot\frac{1}{n}\right]=1\neq 0 \qquad (6.6)$$

随着$n \to \infty$，随机序列x_n取值大于 0 的概率越来越小(为$1/n$)，但一旦取值为正数，则很大(等于n)，故此序列的期望始终为 1。

此序列的方差发散，即$\lim\limits_{n\to\infty} \mathrm{Var}(x_n) = \infty$(参见习题)。

4．依分布收敛

定义　记随机序列$\{x_n\}_{n=1}^{\infty}$与随机变量x的累积分布函数分别为$F_n(x)$与$F(x)$。如果对于任意给定x，都有$\lim\limits_{n\to\infty} F_n(x) = F(x)$，则称随机序列$\{x_n\}_{n=1}^{\infty}$**依分布收敛**(converge in distribution)于随机变量x，记为$x_n \xrightarrow{d} x$，并称x的分布为x_n的**渐近分布**(asymptotic distribution)或**极限分布**(limiting distribution)。

这意味着，当$n \to \infty$时，x_n的分布函数(概率密度函数)越来越像x的分布函数(概率密度函数)。

与依概率收敛或依均方收敛不同，依分布收敛并非随机变量序列本身的收敛，而本质上是(分布)函数序列(sequence of functions)的收敛，已无不确定性。

例 当t分布的自由度越来越大时，t分布依分布收敛于标准正态分布；即当$k \to \infty$时，$t(k) \xrightarrow{d} N(0,1)$。

为了显示依分布收敛的过程，在 Stata 中画 $N(0, 1)$，$t(1)$与 $t(5)$的累积分布函数，结果参见图 6.5。

```
. twoway function N=normal(x) ,range(-5 5) ||
function  t1=t(1,x),range(-5  5)   lp(dash)  ||
function  t5=t(5,x),range(-5  5)   lp(shortdash)
ytitle("累积分布函数")
```

其中，选择项“lp(shortdash) ”表示以短横来画线。

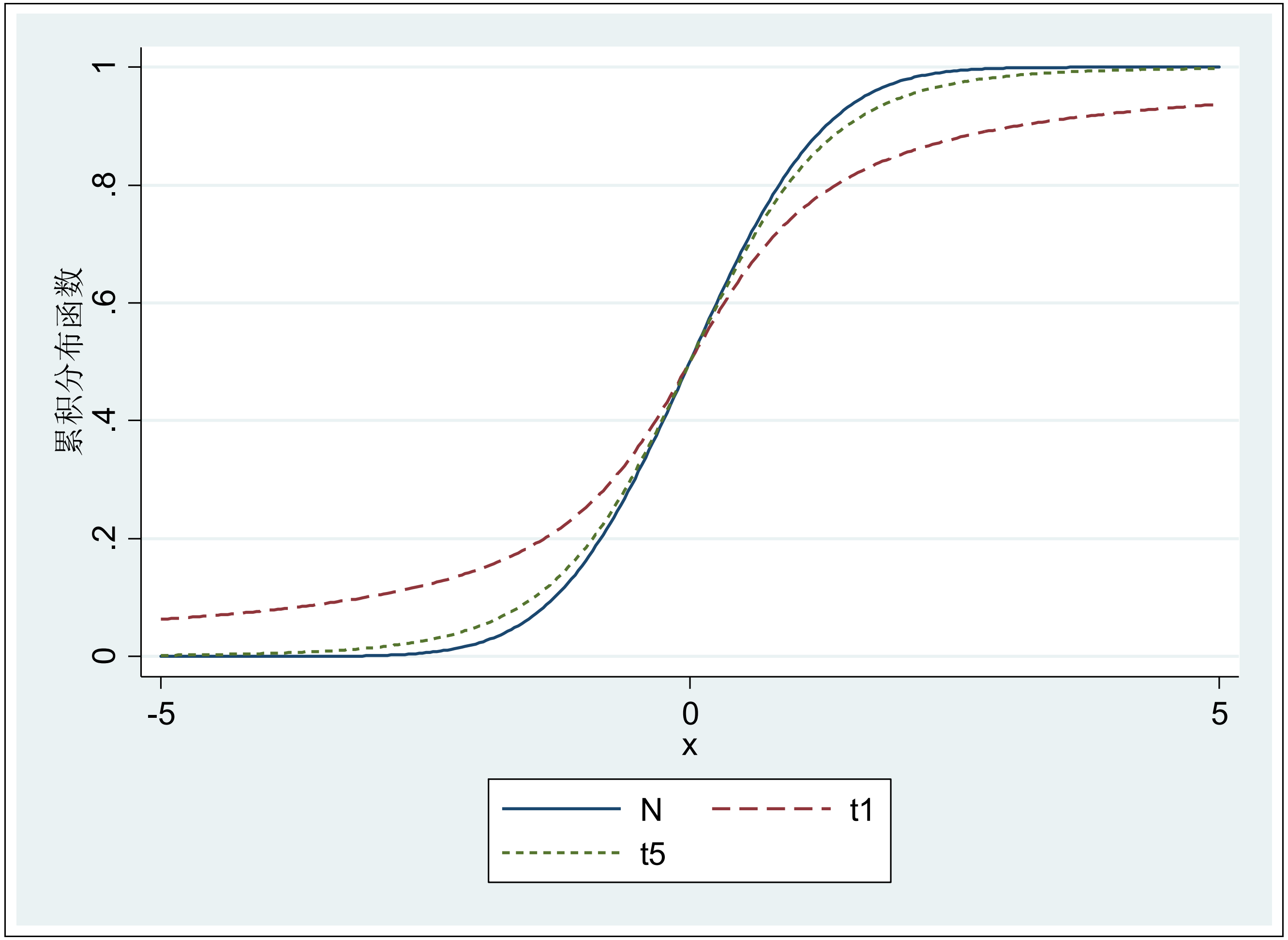

图 6.5　依分布收敛(累积分布函数)

通过概率密度函数，考察*t*分布依分布收敛于标准正态的过程，结果参见图 6.6。

```
. twoway function N=normalden(x) ,range(-5 5) || function  t1=tden(1,x),range(-5  5)  lp(dash)  || function t5=tden(5,x),range(-5 5)  lp(shortdash) ytitle(“概率密度”)
```

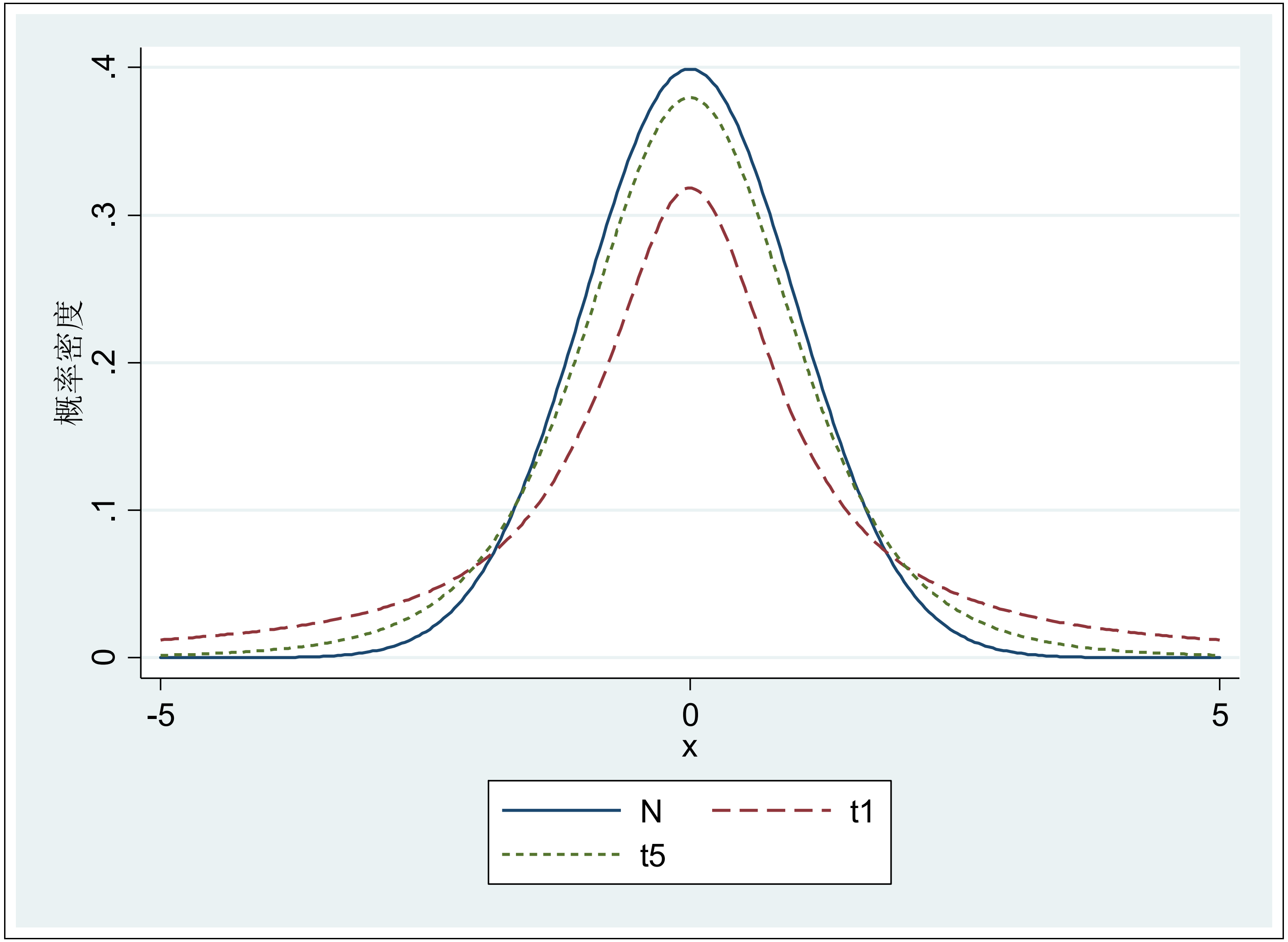

图 6.6　依分布收敛(概率密度函数)

在计量经济学中，许多统计量的大样本分布均为正态分布，故引入如下概念。

定义 如果$x_n \xrightarrow{d} x$，且x服从正态分布，则称x_n为**渐近正态**(asymptotically normal)，即当$n \to \infty$时，x_n的分布越来越像正态分布。

依分布收敛的运算也很方便。

比如，假设$x_n \xrightarrow{d} x$，而$g(\cdot)$为连续函数，则$g(x_n)$的渐近分布就是$g(x)$，即$g(x_n) \xrightarrow{d} g(x)$。

当x_n的分布越来越像x的分布时，$g(x_n)$的分布自然也越来越像$g(x)$的分布。

此结论也称为**连续映射定理**(Continuous Mapping Theorem)。

连续映射定理既适用于依概率收敛，也适用于依分布收敛。

例 假设$x_n \xrightarrow{d} z$，其中$z \sim N(0,1)$，则$x_n^2 \xrightarrow{d} z^2$，其中$z^2 \sim \chi^2(1)$，即$x_n^2 \xrightarrow{d} \chi^2(1)$，因为平方是连续函数。这意味着，渐近标准正态的平方服从渐近$\chi^2(1)$的分布。

“依概率收敛”比“依分布收敛”更强，前者是后者的充分条件；但反之，则不然。

首先，如果$x_n \xrightarrow{p} x$，则意味着$(x_n - x) \xrightarrow{p} 0$，即在极限处$x_n$与$x$的具体取值并无区别，故二者的概率分布也必然相同，所以$x_n \xrightarrow{d} x$。

其次，如果$x_n \xrightarrow{d} x$，这只说明在极限处x_n与x的分布函数相同，但x_n与x的实际取值仍可以很不相同(比如，x_n与x相互独立)。

依分布收敛只是分布函数的收敛(随机变量之间可以毫无关系)，而依概率收敛才是随机变量本身的收敛。

例 假设x与y都为标准正态，且相互独立。考虑随机序列$\left\{x_n = x + (1/n)\right\}_{n=1}^{\infty}$。

由于$1/n \to 0$，故x_n的渐近分布为标准正态，因此$x_n \xrightarrow{d} y$(y也是标准正态)。

但x_n却与y相互独立，x_n的具体取值也与y毫无关系，故x_n并不依概率收敛于y。

随机收敛的三个概念之间的强弱关系为：

$$\text{依均方收敛} \Rightarrow \text{依概率收敛} \Rightarrow \text{依分布收敛}$$

反之，此箭头的相反方向则不成立。

6.3 大数定律与中心极限定理

1. 大数定律(Law of Large Numbers)

假定 $\{x_n\}_{n=1}^{\infty}$ 为独立同分布的随机序列，且 $\mathrm{E}(x_1)=\mu$，$\mathrm{Var}(x_1)=\sigma^2$存在，则样本均值$\bar{x}_n \equiv \dfrac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mu$。

证明： 首先，$\mathrm{E}(\overline{x}_n)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_i)=\frac{1}{n}\cdot n\mu=\mu$，故样本均值$\overline{x}_n$的期望仍为$\mu$。

其次，$\mathrm{Var}(\overline{x}_n)=\mathrm{Var}\left(\frac{x_1+\cdots+x_n}{n}\right)=\frac{1}{n^2}(n\sigma^2)=\frac{\sigma^2}{n}\to 0$，样本均值$\overline{x}_n$的方差收敛到 0。

因此，$\overline{x}_n$ 依均方收敛于μ。

由于“依均方收敛”是“依概率收敛”的充分条件，故$\overline{x}_n \xrightarrow{p} \mu$。

当样本容量n很大时，样本均值趋于总体均值，故名“大数定律”。

2. 中心极限定理(Central Limit Theorem)

根据大数定律，当$n \to \infty$时，样本均值$\overline{x}_n$依概率收敛到总体均值μ。

在一般情况下，$\overline{x}_n$的具体分布则很难推导。

中心极限定理告诉我们，无论原序列$\{x_n\}_{n=1}^{\infty}$服从什么分布，当$n \to \infty$时，样本均值$\overline{x}_n$的渐近分布都为正态分布。

只要样本容量n足够大，则$\overline{x}_n$的真实分布将很接近于正态分布。

故可用正态分布来很好地近似$\overline{x}_n$的真实分布(此真实分布通常无法求解)，并以此渐近分布作为统计推断的基础。

中心极限定理 假定$\{x_n\}_{n=1}^{\infty}$为独立同分布的随机序列，且$\mathrm{E}(x_1)=\mu$，$\mathrm{Var}(x_1)=\sigma^2$存在，则

$$\frac{\bar{x}_n-\mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1) \qquad (6.7)$$

标准化之后的样本均值(即减去期望，除以标准差)的渐近分布为标准正态。

直观上，可视为$\bar{x}_n \xrightarrow{d} N\left(\mu,\sigma^2/n\right)$；但不严格，因为$\bar{x}_n$的方差$\sigma^2/n \to 0$。

根据连续映射定理，将表达式(6.7)两边同乘σ，并将分母的$\sqrt{1/n}$放到分子上，可得中心极限定理的等价表达式：

$$\sqrt{n}\,(\overline{x}_n-\mu)\xrightarrow{d}N(0,\sigma^2) \qquad (6.8)$$

显然，$\sqrt{n}\to\infty$；而根据大数定律，$(\overline{x}_n-\mu)\xrightarrow{p}0$，故上式用$\sqrt{n}\cdot(\overline{x}_n-\mu)$(即“$\infty\cdot 0$”型)得到非退化的渐近正态分布。

表达式(6.8)的好处是，容易推广到多维的情形。

多维的中心极限定理：假定$\{\mathbf{x}_n\}_{n=1}^{\infty}$为独立同分布的随机向量序列，且$\mathrm{E}(\mathbf{x}_1)=\boldsymbol{\mu}$，$\mathrm{Var}(\mathbf{x}_1)=\boldsymbol{\Sigma}$存在，则$\sqrt{n}\,(\overline{\mathbf{x}}_n-\boldsymbol{\mu})\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma})$。

6.4 使用蒙特卡罗法模拟中心极限定理

使用蒙特卡罗法来模拟中心极限定理。

假设x服从在$(0,1)$上的均匀分布，从此分布随机抽取观测值，

样本容量为 30，希望用蒙特卡罗法直观地“看到”样本均值$\overline{x}_{30}$的分布，并与正态分布相比较。

从$(0,1)$上的均匀分布抽取 10000 个样本容量为 30 的随机样本，得到 10000 个$\overline{x}_{30}$的观测值，然后画其直方图。

可使用如下 Stata 程序：

首先，用命令 `program` 定义一个叫“`onesample`”的程序，从均匀分布抽取一个样本容量为 30 的随机样本，并计算$\overline{x}_{30}$；

其次,用命令 `simulate` 重复此程序 10000 次,得到 10000 个$\overline{x}_{30}$的观测值；

最后，用命令 `histogram` 画$\overline{x}_{30}$的直方图。

可在 Stata 命令窗口依次输入如下命令：

```
. program onesample,rclass
```
(定义程序 `onesample`，并以 `r()`形式储存结果)

```
  drop _all
```
(删去内存中已有数据)

```
  set obs 30
```
(确定随机抽样的样本容量为 30)

```
  gen x=runiform()      (得到在(0,1)上均匀分布的随机样本)
  sum x                 (使用命令 sum 计算样本均值)
  return scalar mean_sample=r(mean)  (将样本均值记为 mean_sample)
  end                   (程序 onesample 结束)

. set more off          (指定 Stata 输出结果连续翻页)

. simulate xbar=r(mean_sample),seed(101) reps(10000) nodots: onesample
```

选择项“reps(10000)”表示，命令 simulate 将运行“onesample”程序 10000 遍，并生成变量 xbar 来记录这 10000 个样本均值。

选择项“seed(101)”用来确定随机数的初始值，以便再次模拟或别人运行此程序时，也能得到完全一样的结果。

选择项“nodots”表示不显示表示模拟过程的点点(默认以一个点表示抽取一个样本)。

```
. hist xbar,normal
```

选择项“normal”表示画出相应的正态分布，结果参见图 6.7。

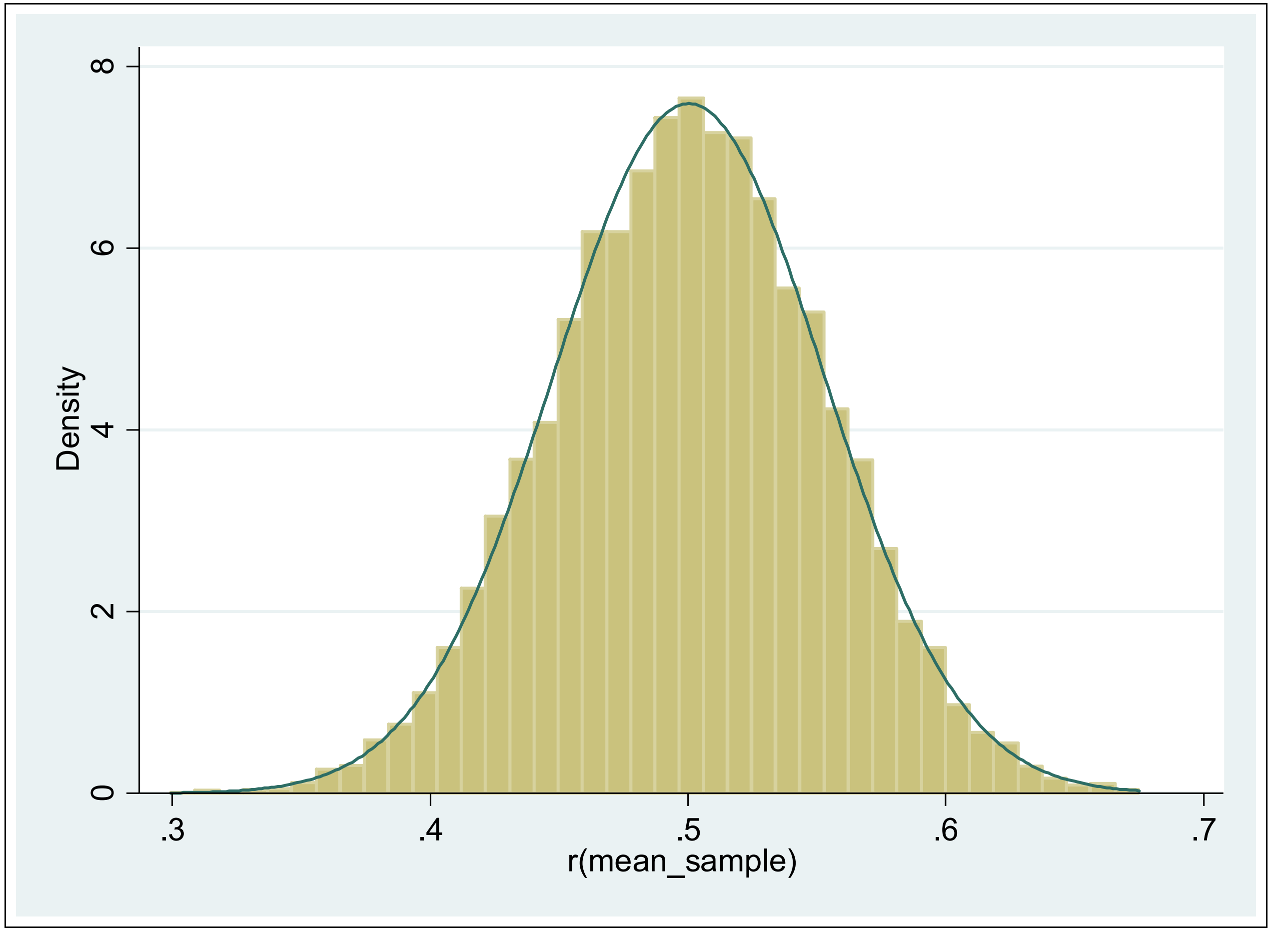

图 6.7　模拟中心极限定理

虽然样本容量仅为 30，但$\overline{x}_{30}$的分布已经很接近于正态分布。

作为练习，可从$\chi^2(10)$中抽取随机样本，重复上面的蒙特卡罗模拟。

只要将上面程序中的语句“`gen x=runiform()`”改为“`gen x=rchi2(10)`”即可。

6.5　统计量的大样本性质

在大样本理论下，我们关心当样本容量$n \to \infty$时，统计量是否具有良好的大样本性质。

1. 一致估计量

定义 考虑参数β的估计量$\hat{\beta}_n$，其中下标n为样本容量(强调$\hat{\beta}_n$对样本容量n依赖)。如果$\operatorname*{plim}_{n\to\infty} \hat{\beta}_n = \beta$，则称$\hat{\beta}_n$是参数$\beta$的**一致估计量**(consistent estimator)。

一致性(consistency)意味着，当样本容量足够大时，$\hat{\beta}_n$依概率收敛到真实参数β，参见图 6.8。

这是对估计量最基本，也是最重要的要求。

如果估计方法不一致，则意味着研究没有太大意义；因为无论样本容量多大，估计量也不会收敛到真实值。

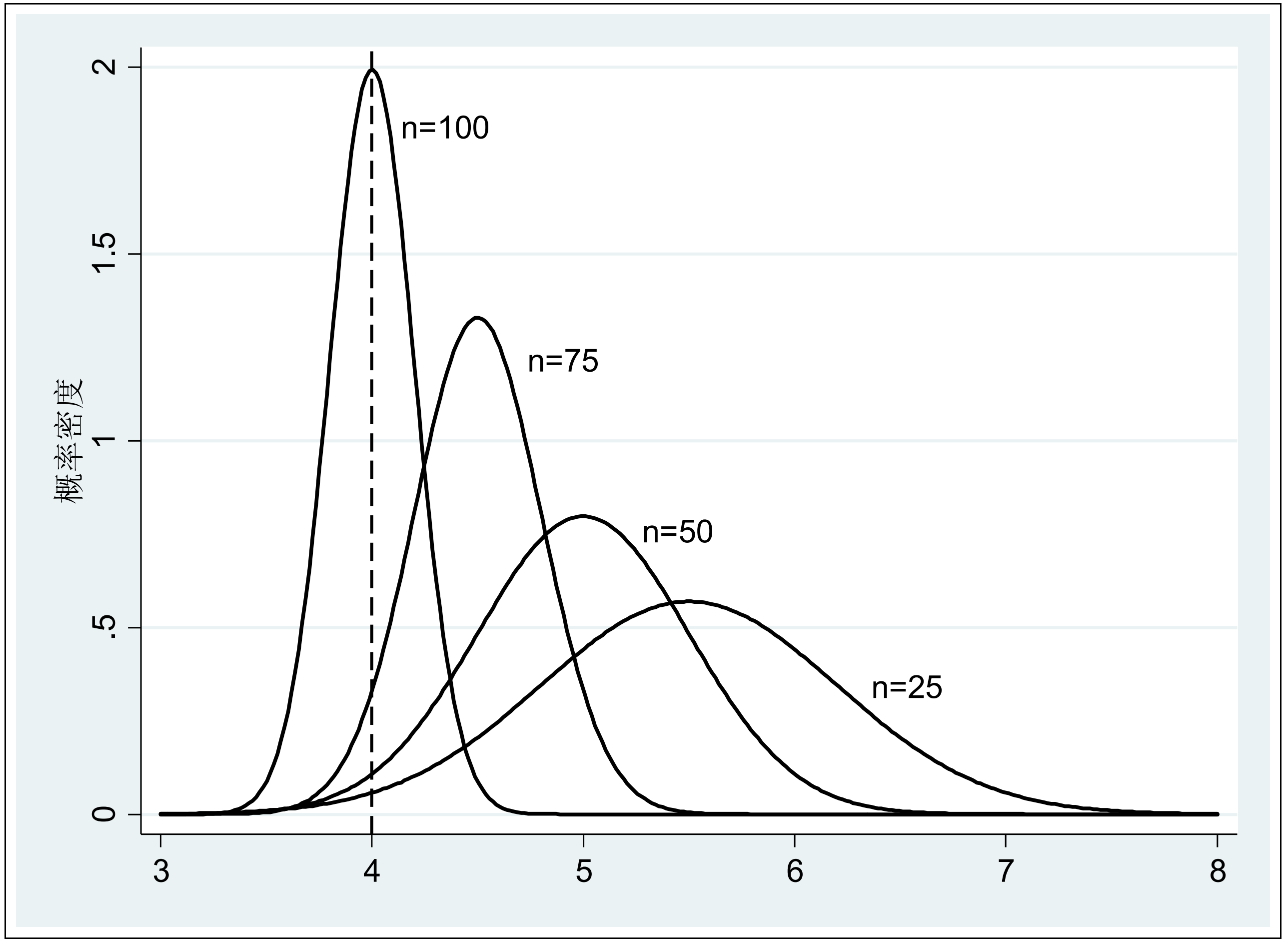

图 6.8　一致估计量示意图

在多维情况下，称估计量$\hat{\boldsymbol{\beta}}_n$是参数$\boldsymbol{\beta}$的一致估计量，如果$\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}$，即$\hat{\boldsymbol{\beta}}_n$的各分量都是$\boldsymbol{\beta}$相应分量的一致估计。

2．渐近正态分布与渐近方差

定义　如果$\sqrt{n}\,(\hat{\beta}_n - \beta) \xrightarrow{d} N(0,\sigma^2)$，则称$\hat{\beta}_n$为**渐近正态**(asymptotically normal)，称σ^2为其**渐近方差**(asymptotic variance)，记为$\text{Avar}(\hat{\beta}_n)$。

可近似认为$\hat{\beta}_n \xrightarrow{d} N\left(\beta, \sigma^2/n\right)$, 但不严格(因为$\sigma^2/n$趋于 0)。

在多维情况下，如果$\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma})$，其中$\boldsymbol{\Sigma}$为半正定矩阵，则称$\hat{\boldsymbol{\beta}}_n$为渐近正态分布，而称$\boldsymbol{\Sigma}$为$\hat{\boldsymbol{\beta}}_n$的渐近协方差矩阵，记为$\mathrm{Avar}(\hat{\boldsymbol{\beta}}_n)$。

3. 渐近有效

假设$\hat{\beta}_n$与$\tilde{\beta}_n$都是β的渐近正态估计量。如果$\mathrm{Avar}(\hat{\beta}_n)\le \mathrm{Avar}(\tilde{\beta}_n)$，则称$\hat{\beta}_n$比$\tilde{\beta}_n$**更为渐近有效**(asymptotically more efficient)。

这意味着，在大样本下，$\hat{\beta}_n$的方差小于$\tilde{\beta}_n$的方差(尽管在小样本下未必如此)。

在多维情况下，假设$\hat{\boldsymbol{\beta}}_n$与$\tilde{\boldsymbol{\beta}}_n$都是$\boldsymbol{\beta}$的渐近正态估计量。如果$[\text{Avar}(\tilde{\boldsymbol{\beta}}_n)-\text{Avar}(\hat{\boldsymbol{\beta}}_n)]$为半正定矩阵，则称$\hat{\boldsymbol{\beta}}_n$比$\tilde{\boldsymbol{\beta}}_n$**更为渐近有效**。

6.6 随机过程的性质

大数定律与中心极限定理假设随机序列为独立同分布(iid)，但对于大多数经济变量而言，此假定可能太强了。

比如，今年的通货膨胀率通常依赖于去年的通货膨胀率，二者并非相互独立。

我们需要研究随机序列的性质，并将常规的大数定律与中心极限定理进行推广。

随机序列$\{x_n\}_{n=1}^{\infty}$也称为**随机过程**(stochastic process)。如果下标为时间，则记为$\{x_t\}_{t=1}^{\infty}$，也称**时间序列**(time series)。

1. 严格平稳过程

数据集 price_retail.dta 包含中国 1978-2021 年的通货膨胀率数据(商品零售价格指数，上年=100)，可记为$\{\pi_{1978}, \pi_{1979}, \cdots, \pi_{2021}\}$。

打开此数据集，并画其时间趋势图：

```
. use price_retail.dta,clear
. graph twoway connect price year,
yline(100,lp(dash))
```

选择项“`yline(100,lp(dash))`”指定在纵轴等于 100 的位置画一条水平的虚线，结果参见图 6.9。

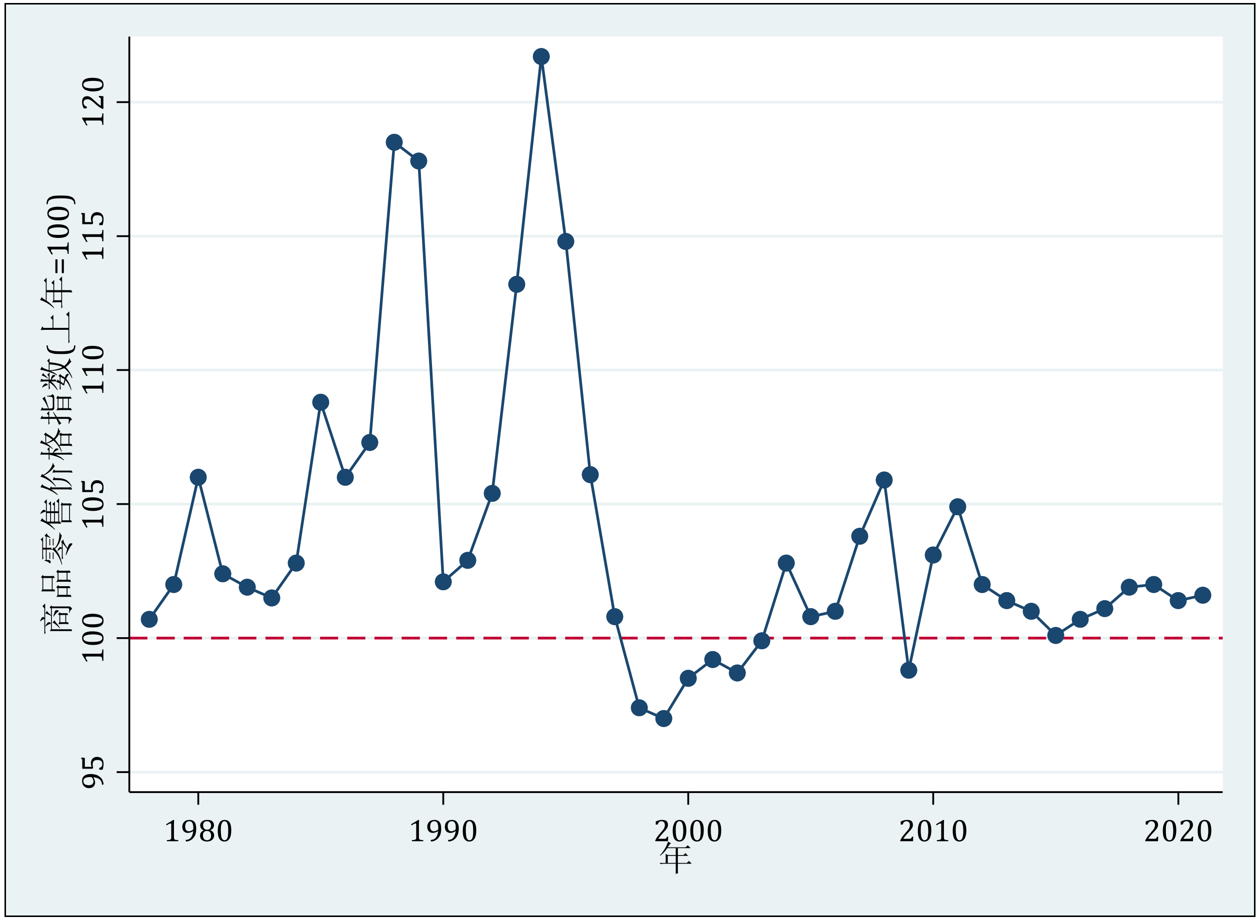

图 6.9　中国商品零售价格指数(上年=100)

数据来源：国家统计局网站(http://data.stats.gov.cn)

假如每年的通货膨胀率作为一个随机变量都有自己不同的分布，如何估计$\mathrm{E}(\pi_{1978})$与$\mathrm{Var}(\pi_{1978})$？

每年通货膨胀率的样本容量仅为 1，且历史不能重演(也无法穿越)！

如果这 44 年的通货膨胀率分布都不变，则可将$\bar{\pi} \equiv \frac{1}{44}\sum_{t=1978}^{2021} \pi_t$作为$\mathrm{E}(\pi_t)$的估计量。

通常要求随机过程$\{x_t\}_{t=1}^{\infty}$的概率分布不随时间推移而改变。

无论过去、现在还是未来去看此随机过程，它的概率分布性质都一样。

这种随机过程称为“严格平稳过程”，它要求随机过程的有限维分布不随时间推移而改变。

比如，x_t的分布与x_s的分布相同$(\forall t, s)$。

(x_1, x_4)的分布与(x_2, x_5)相同(二者均相隔 3 期)。

(x_1, x_2, x_3)的分布与(x_5, x_6, x_7)相同(二者均为连续 3 期)。

定义 随机过程$\{x_t\}_{t=1}^{\infty}$是**严格平稳过程**(strictly stationary process)，简称平稳过程，如果对任意 m 个时期的时间集合$\{t_1, t_2, \cdots, t_m\}$，随机向量$\left\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\right\}$的联合分布等于随机向量$\left\{x_{t_1+k}, x_{t_2+k}, \cdots, x_{t_m+k}\right\}$的联合分布，其中$k$为任意整数。

这意味着，将$\left\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\right\}$中每个变量的时间下标全部前移或后移$k$期，不会改变其分布。

$\left\{x_{t_1}, x_{t_2}, \cdots, x_{t_m}\right\}$的联合分布仅取决于$\left\{t_1, t_2, \cdots, t_m\right\}$各个时期之间的相对距离，而不依赖于其绝对位置。

例　如果随机过程$\left\{x_t\right\}_{t=1}^{\infty}$为 iid，则$\left\{x_t\right\}_{t=1}^{\infty}$是平稳过程，且不存在序列相关。

例　如果随机过程$\left\{x_t\right\}_{t=1}^{\infty}=\left\{x_1, x_1, x_1, \cdots\right\}$(即$x_t \equiv x_1$)，则$\left\{x_t\right\}_{t=1}^{\infty}$是平稳过程，且存在最强的序列相关。

例 考虑以下一阶自回归过程(AR(1))，

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (t = 1, \cdots, T) \qquad (6.9)$$

其中，$\{\varepsilon_t\}$为独立同分布，且$\mathrm{Cov}(y_{t-1}, \varepsilon_t) = 0$。

命题 如果$\rho = 1$，则$\{y_t\}$不是平稳过程。如果$|\rho| < 1$，则$\{y_t\}$为平稳过程。

证明： 如果 $\rho = 1$，则 $y_t = y_{t-1} + \varepsilon_t$。因此，$y_1 = y_0 + \varepsilon_1$，而 $y_2 = y_1 + \varepsilon_2 = y_0 + \varepsilon_1 + \varepsilon_2$，以此类推可知

$$y_t = y_0 + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t \qquad (6.10)$$

给定初始值 y_0，当 $t \to \infty$ 时，$\mathrm{Var}(y_t) = t\sigma_\varepsilon^2 \to \infty$，其中 $\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon_t)$，即方差越来越大，以至无穷。

因此，$\{y_t\}$不是平稳过程(平稳过程要求同分布，故方差不变)。

由于y_t只是在y_{t-1}的基础上，加上一个随机扰动项ε_t，故当$\rho = 1$时，称$\{y_t\}$为“随机游走”(random walk)。

如果$|\rho| < 1$，则$\mathrm{Var}(y_t)$会收敛到常数。对方程(6.9)两边同时取方差，可得

$$\mathrm{Var}(y_t) = \rho^2 \mathrm{Var}(y_{t-1}) + \sigma_\varepsilon^2 \qquad (6.11)$$

记$z_t \equiv \mathrm{Var}(y_t)$，$z_{t-1} = \mathrm{Var}(y_{t-1})$，则上式可写为

$$z_t = \rho^2 z_{t-1} + \sigma_\varepsilon^2 \tag{6.12}$$

这是确定性的一阶线性差分方程，因为$z_t \equiv \text{Var}(y_t)$为非随机。

由于$\rho^2 < 1$，故$\text{Var}(y_t)$将收敛到一个稳定值，参见图 6.10。

在方程(6.12)中，令$z_t = z_{t-1}$，可求解此收敛的稳定值z^*：

$$z^* = \rho^2 z^* + \sigma_\varepsilon^2 \tag{6.13}$$

整理后可得，$z^* = \dfrac{\sigma_\varepsilon^2}{1-\rho^2}$。如果忽略序列$\{y_t\}$的前面几项，则可将$\{y_t\}$的方差视为常数。进一步可证明，$\{y_t\}_{t=0}^{\infty}$是严格平稳过程。

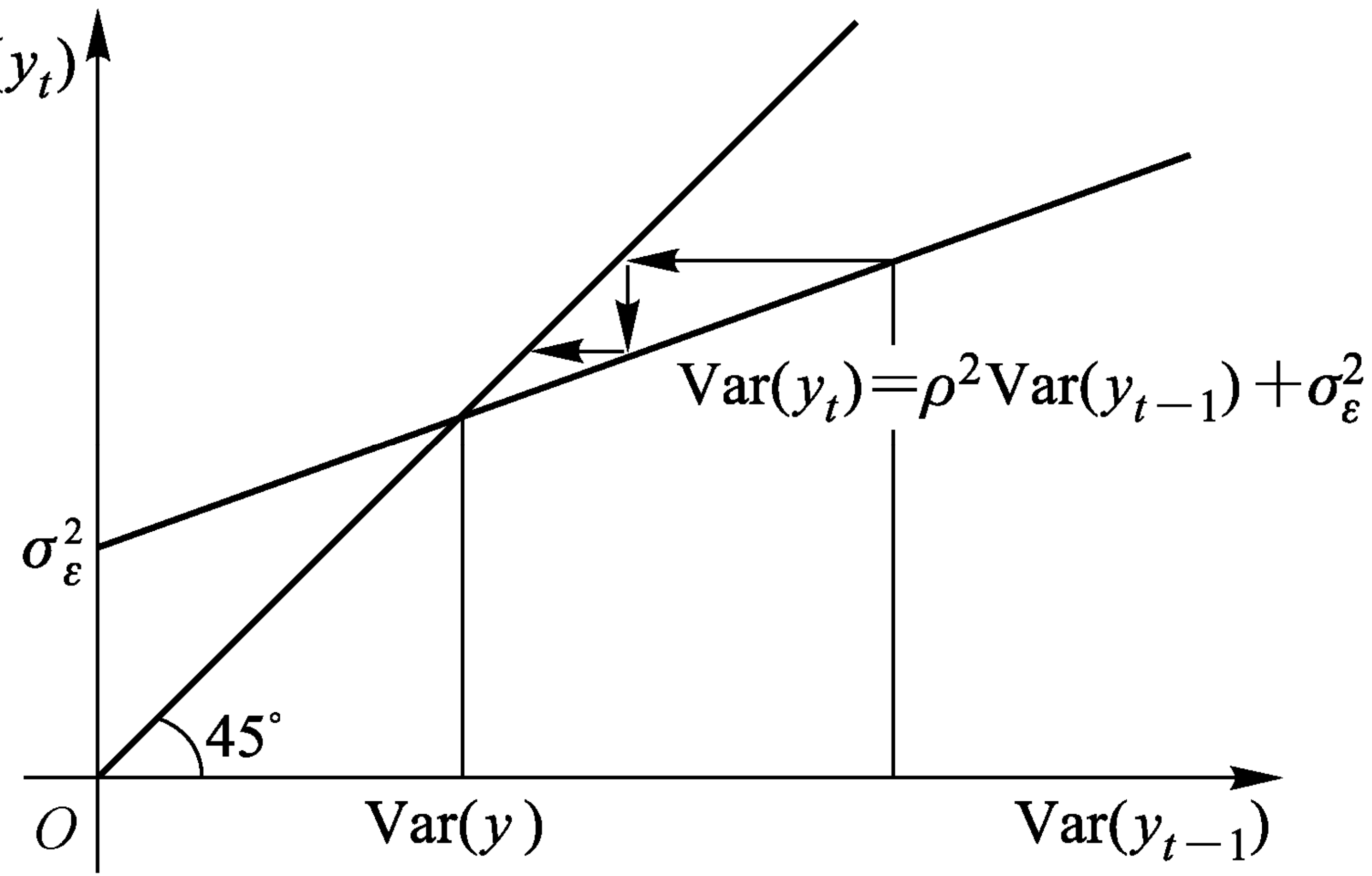

图 6.10　平稳一阶自回归过程的方差收敛

有时我们仅仅关心随机过程的期望、方差及协方差是否稳定，而不要求整个分布都稳定，故引入以下“弱平稳过程”的概念。

定义　随机过程$\{x_t\}_{t=1}^{\infty}$是**弱平稳过程**(weakly stationary process)或**协方差平稳过程**(covariance stationary process)，如果$\mathrm{E}(x_t)$不依赖于t，而且$\mathrm{Cov}(x_t, x_{t+k})$仅依赖于$k$(即$x_t$与$x_{t+k}$在时间上的相对距离)而不依赖于其绝对位置$t$。

对于弱平稳过程，由于$\mathrm{E}(x_t)$不依赖于t，故其期望为常数。

由于$\mathrm{Cov}(x_t, x_{t+k})$仅依赖于$k$，如果令$k=0$，则$\mathrm{Cov}(x_t, x_t)=\mathrm{Var}(x_t)$也不依赖于$t$，故弱平稳过程的方差也是常数。

严格平稳过程是弱平稳过程的充分条件；但反之则不然，因为弱平稳过程只要求二阶矩平稳(即期望、方差、协方差等不随时间而变)，而概率分布还可能依赖于更高阶的矩。

在实践中较常用的弱平稳过程是期望为 0，且不存在序列相关的白噪声过程。

定义 对于弱平稳过程$\{x_t\}_{t=1}^{\infty}$，如果对于$\forall\ t$，都有$\mathrm{E}(x_t)=0$，而且$\mathrm{Cov}(x_t, x_{t+k})=0 \ \ (\forall\ k\neq 0)$，则称为**白噪声过程**(white noise process)。

白噪声过程不一定独立同分布，也不一定是严格平稳过程。

“白噪声”是性质比较好的“噪声”，即该噪声的期望值为 0，而不同期之间的噪声互不相关。

对于随机向量过程$\left\{\mathbf{x}_t\right\}_{t=1}^{\infty}$, 可以类似地定义平稳过程或弱平稳过程(只要将上述定义中的x置换为$\mathbf{x}$即可)。

如果$\left\{\mathbf{x}_t\right\}_{t=1}^{\infty}$为(弱)平稳过程，则其每个分量都是(弱)平稳过程；反之，则不然。

2． 渐近独立性

“严格平稳过程”(相当于“同分布”假定)还不足以应用大数定律或中心极限定理，因为它们都要求独立同分布(iid)。

但“相互独立”的假定对于大多数经济变量而言过强了。

比如，今年的通胀率显然与去年的通胀率相关。

但今年的通胀率与 100 年前的通胀率或许可近似地视为相互独立，称为**渐近独立**(ergodic，也称**遍历性**)，或**弱相依**(weakly dependent)。

渐近独立意味着，只要两个随机变量相距足够远，可近似认为它们相互独立。

例　相互独立的随机序列是渐近独立的。

例 AR(1)是否渐近独立？考虑以下一阶自回归模型：

$$y_t = \rho y_{t-1} + \varepsilon_t \tag{6.14}$$

其中，$|\rho| < 1$，而ε_t为白噪声。

为考察其渐近独立性，分别计算其各阶“自协方差”(autocovariance)。当时间间隔为 1 期时，一阶自协方差为

$$\text{Cov}(y_t, y_{t-1}) = \text{Cov}(\rho y_{t-1} + \varepsilon_t, y_{t-1}) = \rho\sigma_y^2 + \underbrace{\text{Cov}(\varepsilon_t, y_{t-1})}_{=0} = \rho\sigma_y^2 \tag{6.15}$$

其中，σ_y^2为 y 的方差，而

$\mathrm{Cov}(\varepsilon_t, y_{t-1}) = \mathrm{Cov}(\varepsilon_t, \rho y_{t-2} + \varepsilon_{t-1}) = \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1}) = 0$，因为$\varepsilon_t$为白噪声(无序列相关)。

当时间间隔为 2 期时，原方程(6.14)可写为

$$y_t = \rho y_{t-1} + \varepsilon_t = \rho(\rho y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \rho^2 y_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t \qquad (6.16)$$

因此，二阶自协方差为

$$\mathrm{Cov}(y_t, y_{t-2}) = \mathrm{Cov}(\rho^2 y_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t, y_{t-2}) = \rho^2 \sigma_y^2 \qquad (6.17)$$

以此类推，当时间间隔为 j 期时，

$$\mathrm{Cov}(y_t, y_{t-j}) = \rho^j \sigma_y^2 \qquad (6.18)$$

由于$|\rho|<1$，故当上式$j\to\infty$时，$\mathrm{Cov}(y_t,y_{t-j})\to 0$。

相距越远，则序列$\{y_t\}$的自协方差越小，且在极限处变为 0(不相关)，故此 AR(1)模型为渐近独立的过程。

有了严格平稳过程与渐近独立的概念后，可将大数定律作以下重要推广。

渐近独立定理(Ergodic Theorem)　假设$\{x_i\}_{i=1}^{\infty}$为渐近独立的严格平稳过程，且$\mathrm{E}(x_i)=\mu$存在，则$\bar{x}_n \equiv \dfrac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mu$，即样本均值$\bar{x}_n$是总体均值$\mathrm{E}(x_i)$的一致估计。

渐近独立定理是对大数定律的重要推广，更适用于经济数据。

大数定律要求每个x_i相互独立，而渐近独立定理允许$\left\{x_i\right\}_{i=1}^{\infty}$存在“序列相关”(serial correlation)，只要此相关关系在极限处消失即可。

大数定律要求每个x_i的分布相同，而渐近独立定理要求$\left\{x_i\right\}_{i=1}^{\infty}$为严格平稳过程，故也是同分布的。

类似地，可将中心极限定理作相应的推广；即在一定条件下，中心极限定理也适用于渐近独立的平稳过程。

命题 如果$\{x_i\}_{i=1}^{\infty}$为渐近独立的严格平稳过程，则对于任何连续函数$f(\cdot)$，$\{y_i \equiv f(x_i)\}_{i=1}^{\infty}$也是渐近独立的严格平稳过程。

根据此命题，则渐近独立定理意味着，只要$f(\cdot)$为连续函数，则渐近独立平稳过程$\{x_i\}_{i=1}^{\infty}$的任何**总体矩**(population moment) $\mathrm{E}[f(x_i)]$，都可以由其对应的**样本矩**(sample moment)$\frac{1}{n}\sum_{i=1}^{n} f(x_i)$来一致地估计。

例　对于渐近独立的平稳过程 $\begin{pmatrix} x_i \\ y_i \end{pmatrix}_{i=1}^{\infty}$，样本方差 $s^2 \equiv \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ 是总体方差 $\mathrm{Var}(x) \equiv \mathrm{E}[x - \mathrm{E}(x)]^2$ 的一致估计；而样本协方差 $s_{xy} \equiv \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ 为总体协方差 $\mathrm{Cov}(x, y) \equiv \mathrm{E}[(x - \mathrm{E}(x))(y - \mathrm{E}(y))]$ 的一致估计。

6.7　大样本 OLS 的假定

假定 6.1　线性假定

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{6.19}$$

假定 6.2 $(K+1)$维随机过程$\{y_i, x_{i1}, \cdots, x_{iK}\}$为渐近独立的平稳过程(ergodic stationarity)，故适用大数定律与中心极限定理。

例 如果样本为随机样本，则$\{y_i, x_{i1}, \cdots, x_{iK}\}$独立同分布，故是渐近独立的平稳过程。

假定 6.3 前定解释变量(predetermined regressors)

所有解释变量均为**前定**(predetermined)，也称**同期外生**(contemporaneously exogenous)，即它们与同期(同方程)的扰动项正交，即$\mathrm{E}(x_{ik}\varepsilon_i)=0$，$\forall i, k$。

由于$\mathrm{E}(x_{ik}\varepsilon_i)=0$，故$x_{ik}$与$\varepsilon_i$不相关，仿佛在$\varepsilon_i$产生之前，$x_{ik}$已经确定，故名"前定解释变量"。

此假定比严格外生性假定更弱，因为后者要求扰动项与过去、现在及未来的解释变量都不相关(对于时间序列数据而言)，而前定变量仅要求与同期的扰动项不相关。

假定 6.4　秩条件(rank condition)

数据矩阵 **X** 满列秩，即 **X** 中没有多余(可由其他变量线性表出)的解释变量。

大样本理论的假定 6.1 与 6.4 与小样本理论相同，而假定 6.2 与 6.3 则比小样本理论更为放松。

大样本 OLS 无须假设“严格外生性”与“正态随机扰动项”，故具有更大的适用性与稳健性。

6.8 OLS 的大样本性质

在假定 6.1-6.4 之下，可以证明 OLS 估计量$\hat{\boldsymbol{\beta}}$具有以下良好的大样本性质。

(1) $\hat{\boldsymbol{\beta}}$为一致估计量，即$\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$；

以一元回归为例进行说明。考虑以下模型：

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \cdots, n) \qquad (6.20)$$

其中，β的 OLS 估计量为(参见第 4 章)

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (6.21)$$

此模型的离差形式为(参见习题)

$$y_i - \overline{y} = \beta(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}) \qquad (6.22)$$

其中，$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$，$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，而$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$。

将方程(6.22)代入原方程(6.20)可得

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\left[\beta(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})\right]}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \frac{\beta\sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \beta + \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \beta + \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \xrightarrow{p} \beta + \underbrace{\frac{\mathrm{Cov}(x_i, \varepsilon_i)}{\mathrm{Var}(x_i)}}_{=0} = \beta$$

$$(6.23)$$

其中，根据假定 6.3，$\mathrm{Cov}(x_i, \varepsilon_i) = 0$。

前定解释变量，或扰动项与解释变量同期不相关，是保证 OLS 一致的最重要条件。

反之，如果$\mathrm{Cov}(x_i, \varepsilon_i) \neq 0$，则$\operatorname*{plim}_{n\to\infty} \hat{\beta} = \beta + \dfrac{\mathrm{Cov}(x_i, \varepsilon_i)}{\mathrm{Var}(x_i)} \neq \beta$。

如果$\mathrm{Cov}(x_i, \varepsilon_i) > 0$，则$\operatorname*{plim}_{n\to\infty} \hat{\beta} > \beta$。比如，考察教育投资的回报率，$x_i$为教育年限，而$\varepsilon_i$为被遗漏的个人能力。

x_i与ε_i正相关(能力高者通常上学更久)，故 OLS 估计量将高估教育投资的回报率。

如果$\mathrm{Cov}(x_i, \varepsilon_i) < 0$，则$\operatorname*{plim}_{n \to \infty} \hat{\beta} < \beta$。比如，考察上医院对健康的作用，$x_i$为是否上医院，而$\varepsilon_i$为个人原来的健康状况(被遗漏)。

x_i与ε_i负相关(通常只有健康不佳者才上医院)，故 OLS 估计量将低估上医院对健康的正面作用(去医院者的健康往往不如未去医院者)。

可通过图示来考察$\mathrm{Cov}(x_i, \varepsilon_i) \neq 0$的后果，参见图 6.11。

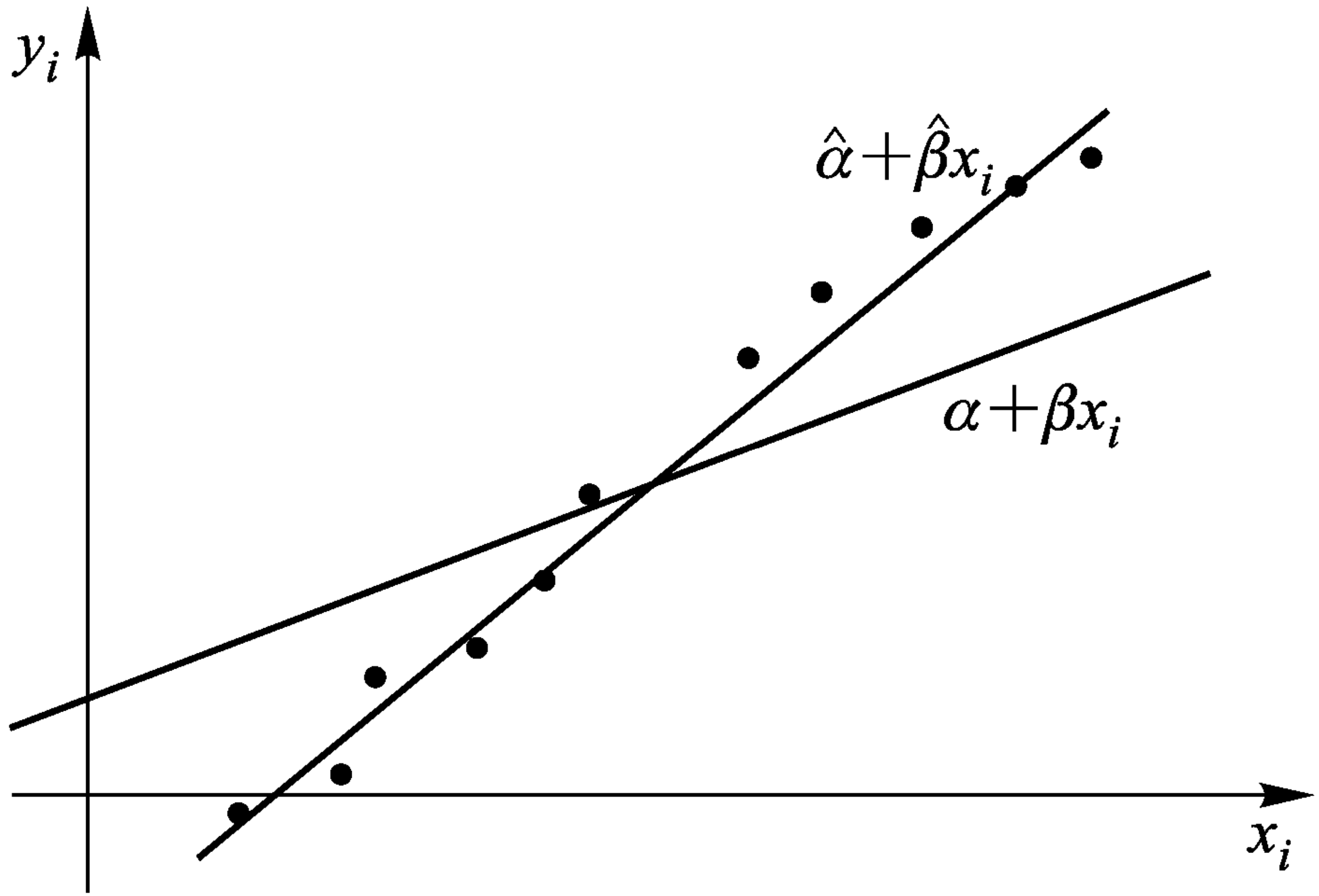

图 6.11　扰动项与解释变量相关导致不一致估计

如果解释变量与扰动项相关，即$\text{Cov}(x_i, \varepsilon_i) \neq 0$，则称此解释变量为**内生解释变量**(endogenous regressor)，简称“内生变量”；反之，则为**外生变量**(exogenous variable)。

由于内生变量的存在，致使 OLS 回归出现偏差，统称为**内生性偏差**(endogeneity bias)，简称**内生性**。

在什么情况下可能出现内生性偏差？

如果存在遗漏变量、双向因果关系、或解释变量测量误差(measurement errors)，则常会出现解释变量与扰动项同期相关的情形，导致 OLS 不一致。

(2) $\hat{\boldsymbol{\beta}}$服从渐近正态分布，即 $\sqrt{n}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}, \mathrm{Avar}(\hat{\boldsymbol{\beta}})\right)$，其中$\mathrm{Avar}(\hat{\boldsymbol{\beta}})$为$\hat{\boldsymbol{\beta}}$的渐近协方差矩阵。

$\hat{\boldsymbol{\beta}}$之所以服从渐近正态，是因为在一定条件下，中心极限定理适用于渐近独立的平稳过程。

(3) 由于大样本理论一般不假设球形扰动项，故渐近协方差矩阵$\mathrm{Avar}(\hat{\boldsymbol{\beta}})$的表达式更为复杂。

根据第 5 章公式(5.46)，OLS 估计量$\hat{\boldsymbol{\beta}}$的协方差矩阵可写为夹心估计量的形式：

$$\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad (6.24)$$

其中，$\mathrm{Var}(\boldsymbol{\varepsilon}\,|\,\mathbf{X})$为扰动项的协方差矩阵。

如果存在球形扰动项(同方差、无自相关)，则$\mathrm{Var}(\boldsymbol{\varepsilon}\,|\,\mathbf{X}) = \sigma^2\mathbf{I}_n$，上式可简化为

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}\,|\,\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad (6.25)$$

对于横截面数据，经常存在异方差，但无自相关(比如，各截面单位之间相互独立)，扰动项的协方差矩阵可写为

$$\mathrm{Var}(\boldsymbol{\varepsilon}\,|\,\mathbf{X}) = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix} \qquad (6.26)$$

其中，$\sigma_1^2, \cdots, \sigma_n^2$不全相等。

如何估计上式的$\left\{\sigma_1^2, \cdots, \sigma_n^2\right\}$？

由于$\left\{\sigma_1^2, \cdots, \sigma_n^2\right\}$为各扰动项的方差，且$\sigma_i^2 = \mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) - \underbrace{[\mathrm{E}(\varepsilon_i)]^2}_{=0} = \mathrm{E}(\varepsilon_i^2)$，故一个自然想法是，以 OLS 残差平方$\left\{e_1^2, \cdots, e_n^2\right\}$替代上式的$\left\{\sigma_1^2, \cdots, \sigma_n^2\right\}$，得到扰动项协方差矩阵的估计量：

$$\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})} = \frac{n}{n-K}\begin{pmatrix} e_1^2 & & 0 \\ & \ddots & \\ 0 & & e_n^2 \end{pmatrix} \tag{6.27}$$

其中，$\dfrac{n}{n-K}$为自由度的调整(在大样本下无差别)。

将表达式(6.27)代入方程(6.24)，可得如下方差估计量

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}\,|\,\mathbf{X})}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathrm{Var}(\boldsymbol{\varepsilon}\,|\,\mathbf{X})}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad (6.28)$$

上式只在小样本下才成立。

当$n\to\infty$时，对参数$\hat{\boldsymbol{\beta}}$的估计变得无限准确，故$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}\,|\,\mathbf{X})}\to 0$。

考虑$\sqrt{n}\,\hat{\boldsymbol{\beta}}$的方差估计量，即$\hat{\boldsymbol{\beta}}$的渐近方差估计量：

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}} \mid \mathbf{X})} = n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{6.29}$$

上式为$\hat{\boldsymbol{\beta}}$渐近协方差矩阵的一致估计量，即

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}} \mid \mathbf{X})} \xrightarrow{p} \mathrm{Avar}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) \tag{6.30}$$

由于在推导过程中并未假设“条件同方差”，故它提供了在“条件异方差”情况下也成立的标准误，称为**异方差稳健的标准误** (heteroskedasticity-consistent standard errors)，简称**稳健标准误** (robust standard errors)。

在形式上，稳健标准误也是夹心估计量。

稳健标准误的思想最早由 Eicker (1967)与 Huber (1967)提出，并由 White (1980)严格证明，故也称 White's standard errors，Huber-White standard errors，或 Eicker-Huber-White standard errors。

通过使用迭代期望定律可以证明，在条件同方差的假定下，稳健标准误还原为普通(非稳健)标准误。

特别地，考虑同方差的一种极端情形，即$e_1^2 = e_2^2 = \cdots = e_n^2$(所有残差的绝对值都相等，但符号可以相反)，则

$$\widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})} = \frac{n}{n-K} \begin{pmatrix} e_1^2 & & 0 \\ & \ddots & \\ 0 & & e_n^2 \end{pmatrix} = \frac{ne_i^2}{n-K} \mathbf{I}_n = \frac{\sum_{i=1}^{n} e_i^2}{\underbrace{n-K}_{=s^2}} \mathbf{I}_n = s^2 \mathbf{I}_n$$

(6.31)

此时，稳健的协方差矩阵可简化为同方差情况下的普通(非稳健)协方差矩阵：

$$\begin{aligned} \widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X})} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \widehat{\mathrm{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (s^2 \mathbf{I}_n) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} = s^2 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{6.32}$$

6.9 大样本统计推断

对于渐近独立的平稳过程，如果样本容量足够大，则 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的渐近正态分布是对其真实分布的较好近似，故可使用其渐近分布进行大样本假设检验与区间估计。

1．检验单个系数：$H_0: \beta_k = c$

考虑检验 $H_0: \beta_k = c$，其中 c 为已知常数。

根据大样本理论，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 服从渐近正态分布，即 $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}, \mathrm{Avar}(\hat{\boldsymbol{\beta}})\right)$，其中 $\mathrm{Avar}(\hat{\boldsymbol{\beta}})$ 为渐近协方差矩阵。

具体到$\hat{\boldsymbol{\beta}}$的第 k 个元素$\hat{\beta}_k$，则有

$$\sqrt{n}\,(\hat{\beta}_k - \beta_k) \xrightarrow{d} N\left(0,\ \mathrm{Avar}(\hat{\beta}_k)\right) \qquad (6.33)$$

其中，$\mathrm{Avar}(\hat{\beta}_k)$为$\hat{\beta}_k$的渐近方差，即渐近方差矩阵$\mathrm{Avar}(\hat{\boldsymbol{\beta}})$主对角线上的第$k$个元素。

在原假设H_0成立的情况下，$\beta_k = c$，故表达式(6.33)可写为

$$\sqrt{n}\,(\hat{\beta}_k - c) \xrightarrow{d} N\left(0,\ \mathrm{Avar}(\hat{\beta}_k)\right) \qquad (6.34)$$

记$\widehat{\text{Avar}(\hat{\beta}_k)}$为渐近方差矩阵估计量$\widehat{\text{Avar}(\hat{\boldsymbol{\beta}})}$主对角线上的第$k$个元素，则$\widehat{\text{Avar}(\hat{\beta}_k)}$是$\text{Avar}(\hat{\beta}_k)$的一致估计量。

定义t统计量为

$$t_k \equiv \frac{\sqrt{n}\,(\hat{\beta}_k - c)}{\sqrt{\widehat{\text{Avar}(\hat{\beta}_k)}}} = \frac{\hat{\beta}_k - c}{\sqrt{\frac{1}{n}\widehat{\text{Avar}(\hat{\beta}_k)}}} \equiv \frac{\hat{\beta}_k - c}{\text{SE}^*(\hat{\beta}_k)} \xrightarrow{d} N(0,1) \quad (6.35)$$

$\text{SE}^*(\hat{\beta}_k) \equiv \sqrt{\frac{1}{n}\widehat{\text{Avar}(\hat{\beta}_k)}}$即为异方差稳健的标准误。

由于在推导过程中并未用到“条件同方差”的假定，故在“条件异方差”的情况下也适用。

统计量t_k称为**稳健t比值**(robust t ratio)，服从渐近标准正态分布，而不是t分布。

对于双边检验(即$H_1 : \beta_k \neq c$)，则$|t_k|$越大，越倾向于拒绝H_0。

比如，对于 5%的显著性水平，如果$|t_k|$大于临界值 1.96，则可拒绝H_0。

也可通过p值进行检验,其方法与小样本理论相同(参见第 5 章)。

2．检验线性假设：$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$

在大样本理论下，对于多个线性假设的联合检验，与小样本理论下的F检验类似。考虑检验m个线性假设是否同时成立：

$$H_0 : \underbrace{\mathbf{R}}_{m\times K} \underbrace{\boldsymbol{\beta}}_{K\times 1} = \underbrace{\mathbf{r}}_{m\times 1}$$

其中，$\mathbf{r}$ 为 m 维列向量（$m < K$），$\mathbf{R}$ 为 $m\times K$ 矩阵，且 $\text{rank}(\mathbf{R}) = m$，即 $\mathbf{R}$ 满行秩，没有多余或自相矛盾的行或方程。

对于原假设 $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$，根据沃尔德检验原理，可考察 $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$ 的大小，譬如其二次型 $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$。

在 H_0 成立的情况下，可以证明统计量

$$W \equiv n(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\mathbf{R}\widehat{\text{Avar}(\hat{\boldsymbol{\beta}})}\mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \xrightarrow{d} \chi^2(m) \qquad (6.36)$$

其中，$\widehat{\mathbf{R}\mathrm{Avar}(\hat{\boldsymbol{\beta}})\mathbf{R}'}$为$(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})$的渐近方差矩阵(使用夹心估计量公式)。

如果统计量W大于$\chi^2(m)$的临界值，则拒绝原假设。

在表达式(6.36)中，虽然统计量W服从χ^2分布，而非小样本理论下的F分布，但χ^2分布与F分布在大样本情况下是等价的。

即使在大样本下使用稳健标准误进行假设检验，Stata 也依然汇报F统计量及其p值。

命题 假设统计量 $F \sim F(m,n)$ 分布，则当 $n \to \infty$ 时，$mF \xrightarrow{d} \chi^2(m)$。

证明：因为$F \sim F(m,n)$，故可写为$F = \dfrac{\chi^2(m)/m}{\chi^2(n)/n}$，其中分子与分母相互独立。

根据χ^2分布的性质，χ^2分布的期望等于自由度，而方差等于自由度的两倍；即$\mathrm{E}[\chi^2(n)] = n$，且$\mathrm{Var}[\chi^2(n)] = 2n$。

考察此F统计量的分母$\chi^2(n)/n$。

其期望为 $\mathrm{E}[\chi^2(n)/n]=n/n=1$，而方差为 $\mathrm{Var}[\chi^2(n)/n]=2n/n^2=2/n\to 0$ (当 $n\to\infty$ 时)。

因此，此 F 统计量的分母依均方收敛于 1，故依概率收敛于 1(前者是后者的充分条件)，即 $\chi^2(n)/n\xrightarrow{p}1$。

在大样本下，$\chi^2(n)/n$ 退化为 1，此 F 统计量的性质仅由分子 $\chi^2(m)/m$ 决定，故 $F\xrightarrow{d}\chi^2(m)/m$。

因此，在大样本下，$mF\xrightarrow{d}\chi^2(m)$。

6.10 大样本 OLS 的 Stata 实例

在 Stata 中，可方便地得到 OLS 估计的稳健标准误，其命令为

```
. reg y x1 x2 x3,robust
```

选择项“robust”表示稳健标准误。

以数据集 nerlove.dta 为例。

取自 Nerlove (1963)对电力行业规模报酬的经典研究。

此数据集包括 1955 年美国 145 家电力企业的横截面数据。

主要变量为 *tc* (total cost，总成本)，*q* (total output，总产量)，*pl* (price of labor，小时工资率)，*pk* (user cost of capital，资本的使用成本) 与 *pf* (price of fuel，燃料价格)，以及相应的对数值 *lntc*，*lnq*，*lnpl*，*lnpk*，与 *lnpf*。

Nerlove (1963)假设企业i的生产函数为 Cobb−Douglas 函数：

$$Q_i = A_i L_i^{\alpha_1} K_i^{\alpha_2} F_i^{\alpha_3} \qquad (6.37)$$

A, L, K, F 分别为生产率、劳动力、资本与燃料。

记$r \equiv \alpha_1 + \alpha_2 + \alpha_3$为规模效应(degree of returns to scale)。

如果$r = 1$，则规模报酬不变。

如果$r>1$，则规模报酬递增。

如果$r<1$，则规模报酬递减。

Nerlove (1963)的主要目的是确定美国电力行业的规模经济。

假设企业追求成本最小化，可以证明成本函数也为 Cobb−Douglas 函数：

$$TC_i = \delta_i Q_i^{1/r} (P_L)_i^{\alpha_1/r} (P_K)_i^{\alpha_2/r} (P_F)_i^{\alpha_3/r} \qquad (6.38)$$

δ_i是$A_i, \alpha_1, \alpha_2, \alpha_3$的函数。取对数后得到如下模型，

$$\ln TC_i = \beta_1 + \frac{1}{r}\ln Q_i + \frac{\alpha_1}{r}\ln P_{L,i} + \frac{\alpha_2}{r}\ln P_{K,i} + \frac{\alpha_3}{r}\ln P_{F,i} + \varepsilon_i \quad (6.39)$$

首先，打开数据集 nerlove.dta，并使用普通标准误对方程(6.39)进行 OLS 估计：

```
. use nerlove.dta,clear
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS
Model	269.524728	4	67.3811819
Residual	21.5420958	140	.153872113
Total	291.066823	144	2.02129738

Number of obs	=	145
F(4, 140)	=	437.90
Prob > F	=	0.0000
R-squared	=	0.9260
Adj R-squared	=	0.9239
Root MSE	=	.39227

lntc	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

R^2 =0.9260，$\bar{R}^2$ =0.9239，检验整个方程显著性的 F 统计量高达 437.9，其相应 p 值(`Prob > F`)为 0.0000，表明此回归方程高度显著。

lnpl 与 *lnpk* 这两个变量均不显著，其 p 值(`P>|t|`)分别为 0.131 与 0.528。

变量 *lnpk* 的系数(`Coef.`)符号为负，与经济理论的预测相反。Nerlove(1963)认为，这是由于“资本使用成本”的数据不太可靠。

由于 *lnq* 的系数为$1/r$(即规模报酬的倒数)，可估计规模报酬为

```
. display 1/_b[lnq]
1.387129
```

其中，“`_b[lnq]`”表示“*lnq*”的 OLS 系数估计值。

由于$\hat{r} = 1.387129 > 1$，故认为可能存在规模报酬递增。

为此，检验规模报酬不变的原假设“$H_0: r = 1$”，输入命令

```
. test lnq=1
```

此命令检验的原假设为，变量 *lnq* 的系数等于 1。

```
( 1)  lnq = 1

      F(  1,   140) =  256.27
           Prob > F =    0.0000
```

由于p值为 0.0000，故可强烈拒绝原假设，认为存在规模报酬递增。

其次，使用稳健标准误重新进行回归。

```
. reg lntc lnq lnpl lnpk lnpf,r
```

```
Linear regression                                Number of obs     =        145
                                                 F(4, 140)         =     177.19
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.9260
                                                 Root MSE          =     .39227

------------------------------------------------------------------------------
             |               Robust
        lntc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lnq |   .7209135   .0325376    22.16   0.000      .656585     .785242
        lnpl |   .4559645    .260326     1.75   0.082    -.0587139    .9706429
        lnpk |  -.2151476   .3233711    -0.67   0.507    -.8544698    .4241745
        lnpf |   .4258137   .0740741     5.75   0.000     .2793653    .5722622
       _cons |  -3.566513   1.718304    -2.08   0.040    -6.963693   -.1693331
------------------------------------------------------------------------------
```

对比以上两个回归的结果可知，使用选择项“robust”所得到的 OLS 回归系数完全相同，只是所得到的稳健标准误(Robust Std. Err.)与普通标准误(Std. Err.)不同。

对于变量 *lnq* 的系数，其稳健标准误(0.033)几乎是普通标准误(0.017)的两倍。

其他变量系数的稳健标准误反而比普通标准误有所下降。

如果认为存在异方差，则应使用稳健标准误。

在异方差的情况下，如果使用普通标准误，将大大低估变量 *lnq* 系数的真实标准误，从而导致不正确的统计推断。

在 Stata 中使用稳健标准误，即可进行大样本检验。

对单个变量系数显著性的检验，可使用上表中的稳健 t 统计量(服从渐近正态分布)来进行。

可直接看表中所列的 p 值(P>|t|)。

对于更一般的线性假设，仍可使用命令 `test` 来检验。

比如，检验变量 *lnq* 的系数是否为 1：

```
. test lnq=1
```

```
 ( 1)  lnq = 1

       F(  1,   140) =   73.57
            Prob > F =    0.0000
```

由于 p 值为 0.0000，即使使用稳健标准误，仍强烈拒绝原假设。

在使用稳健标准误的情况下，Stata 仍然汇报 F 统计量(服从 F 分布)，即依然使用小样本理论中的 F 统计量公式，但将协方差矩阵换成“稳健的协方差矩阵”。

F 分布与 χ^2 分布在大样本下是等价的，参见本章 6.9 节。

6.11 大样本理论的蒙特卡罗模拟

考虑以下数据生成过程(DGP)：

$$y=\alpha+\beta x+\varepsilon,\quad x\sim\chi^{2}(1),\quad \varepsilon\sim\chi^{2}(10)-10 \tag{6.40}$$

其中，$\alpha=1$，$\beta=2$，解释变量x服从$\chi^{2}(1)$分布；而扰动项ε服从经过位移后的$\chi^{2}(10)$分布，以保证其期望为零(卡方分布的期望为其自由度)；而且x与ε相互独立。

由于小样本理论要求扰动项服从正态分布，这个模型不满足小样本理论的假定，但符合大样本理论的要求。

首先，考虑样本容量为 20 的情形，看 OLS 估计量$\hat{\beta}$与真实值$\beta=2$的差距，以及$\hat{\beta}$的分布能否收敛到正态分布。

抽取 10000 个样本容量为 20 的随机样本，进行回归，得到 10000 个 $\hat{\beta}$。

先用命令 `program` 定义一个名为“`chi2data_20`”的程序进行一次抽样；然后，用命令 `simulate` 来重复此程序 10000 次：

```
. program chi2data_20,rclass
```
(定义程序 chi2data_20，以 r()形式储存结果)

```
  drop _all
```
(删去内存中已有数据)

```
  set obs 20
```
(确定随机抽样的样本容量为 20)

```
  gen x = rchi2(1)
```
(生成服从 $\chi^2(1)$ 分布的解释变量)

```
  gen y = 1 + 2*x + rchi2(10)-10
```
(生成被解释变量)

```
reg y x                          (线性回归)
return scalar b=_b[x]            (存储β̂的估计值)
end                              (程序 chi2data 结束)
```

```
. set more off                   (指定 Stata 输出结果连续翻页)
```

```
. simulate bhat=r(b),reps(10000) seed(10101) nodots:chi2data_20
```

其中，选择项“`reps(10000)`”表示通过命令 `simulate` 将程序“`chi2data_20`” 模拟 10000 次。

得到 10000 个 $\hat{\beta}$ 后，可计算其均值与标准差：

```
. sum bhat
```

Variable	Obs	Mean	Std. dev.	Min	Max
bhat	10,000	2.007048	.9918413	-6.570165	8.955813

$\hat{\beta}$的样本均值为 2.0071，很接近真实值 2，验证了$\hat{\beta}$为β的无偏估计。

但标准(误)差为 0.992，接近于 1，故估计误差较大(因为样本容量仅为 20)。

通过直方图来看这 10000 个$\hat{\beta}$的分布，结果参见图 6.12。

```
. hist bhat,normal
```

其中，选择项“`normal`”表示同时画相应的正态分布密度图。

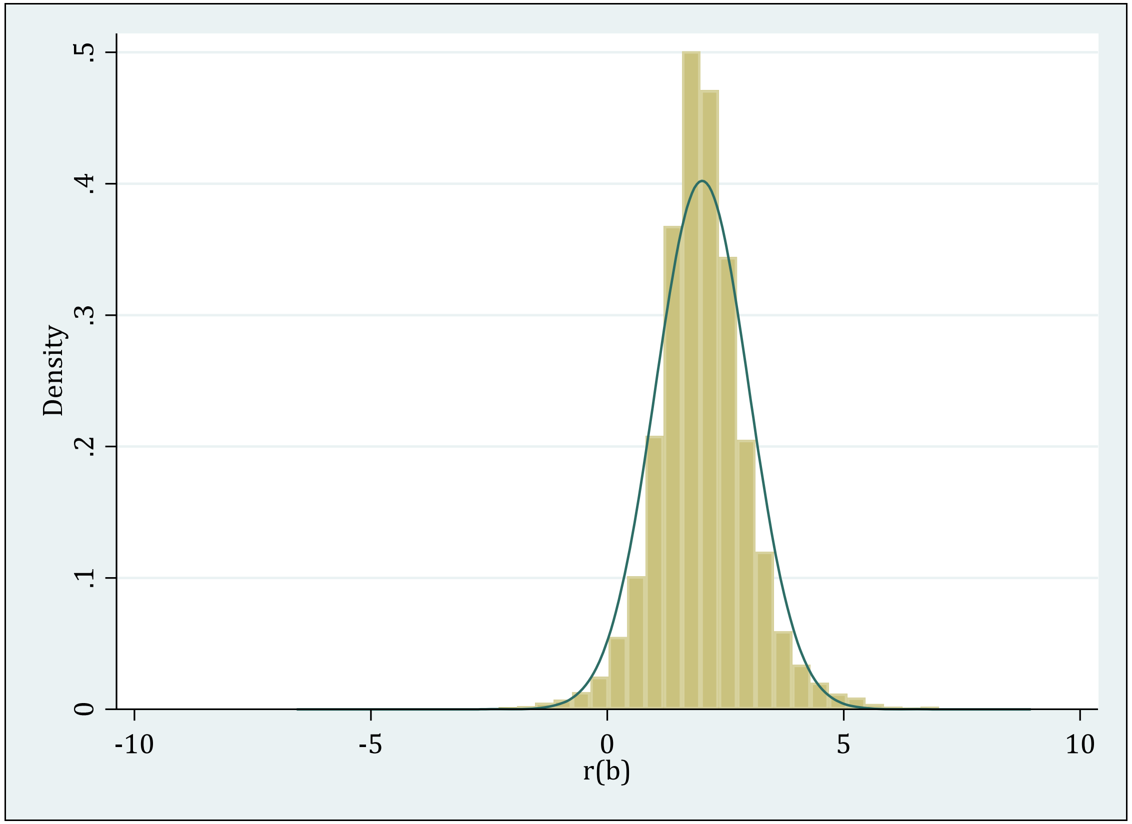

图 6.12　$\hat{\beta}$的分布(样本容量为 20)

当样本容量为 20 时，$\hat{\beta}$的真实分布与正态分布仍有一定差距。

其次，用命令 program 定义一个名为“chi2data_100”的程序，将样本容量增加至 100(将命令“set obs 20”改为“set obs 100”)，仍然抽取 10000 个随机样本，即在上述程序中，再次得到 10000 个$\hat{\beta}$；然后看$\hat{\beta}$的统计特征。

```
. program chi2data_100,rclass
  drop _all
  set obs 100
  gen x = rchi2(1)
  gen y = 1 + 2*x + rchi2(10)-10
  reg y x
  return scalar b=_b[x]
  end
. simulate bhat=r(b),reps(10000) seed(10101)
nodots:chi2data_100
```

```
. sum bhat
```

Variable	Obs	Mean	Std. dev.	Min	Max
bhat	10,000	2.00086	.3378564	.7553486	3.522173

$\hat{\beta}$的样本均值为 2.00086，更加接近真实值 2。

与当样本容量从 20 增加到 100 后，$\hat{\beta}$的标准(误)差从 0.992 下降到 0.338。

画$\hat{\beta}$的直方图，并与正态分布比较，结果参见图 6.13。

```
. hist bhat,normal
```

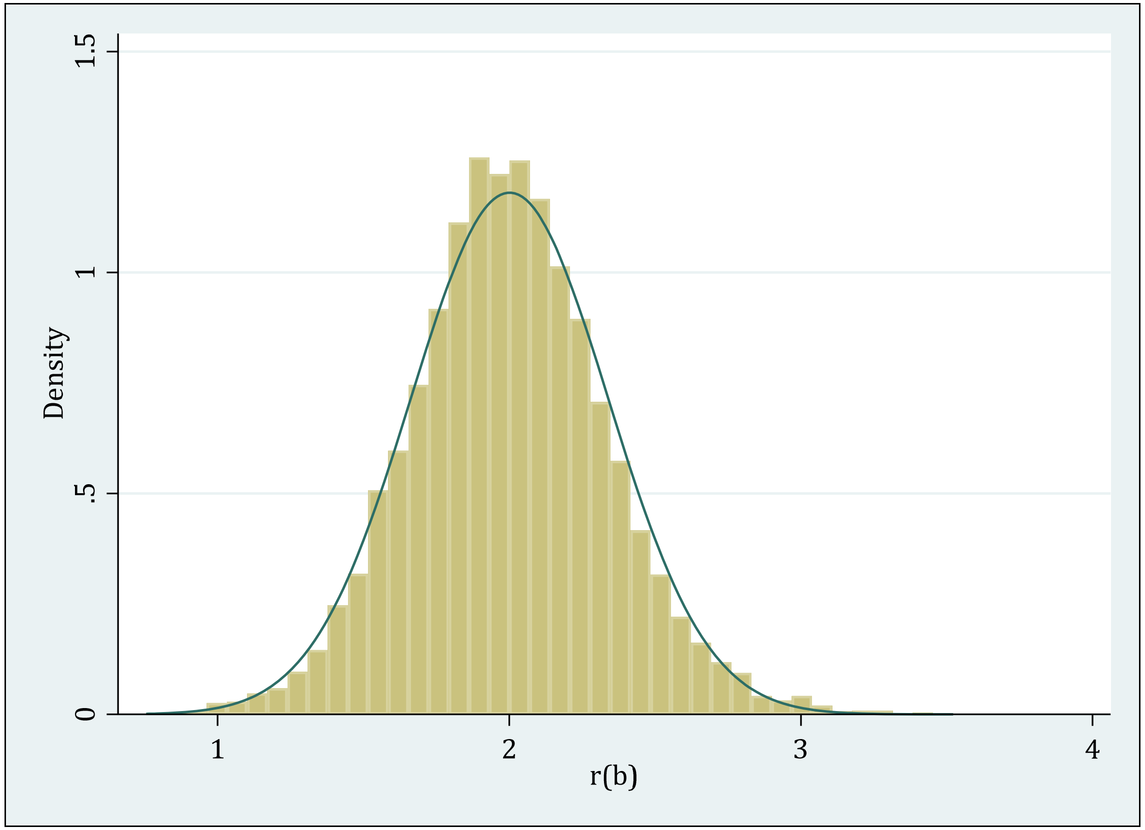

图 6.13　$\hat{\beta}$的分布(样本容量为 100)

当样本容量为 100 时，$\hat{\beta}$的真实分布与正态分布已较为接近。

为了进一步验证 OLS 估计量$\hat{\beta}$的一致性与渐近正态性，用命令 program 定义一个名为“chi2data_1000”的程序，将样本容量增加为 1000，得到 10000 个$\hat{\beta}$，再看其统计特征。

```
. program chi2data_1000,rclass
  drop _all
  set obs 1000
  gen x = rchi2(1)
  gen y = 1 + 2*x + rchi2(10)-10
  reg y x
  return scalar b=_b[x]
  end
. simulate bhat=r(b),reps(10000) seed(10101)
nodots:chi2data_1000
. sum bhat
```

Variable	Obs	Mean	Std. dev.	Min	Max
bhat	10,000	2.000912	.1001662	1.620001	2.440964

当样本容量增加为 1000 时，$\hat{\beta}$的样本均值为 2.000912；而$\hat{\beta}$的标准(误)差则下降为 0.1002。

这验证了$\hat{\beta}$依均方收敛于β，故$\operatorname*{plim}_{n\to\infty}\hat{\beta}=\beta$，即$\hat{\beta}$为一致估计量。

通过直方图看$\hat{\beta}$的真实分布，参见图 6.14。

```
. hist bhat,normal
```

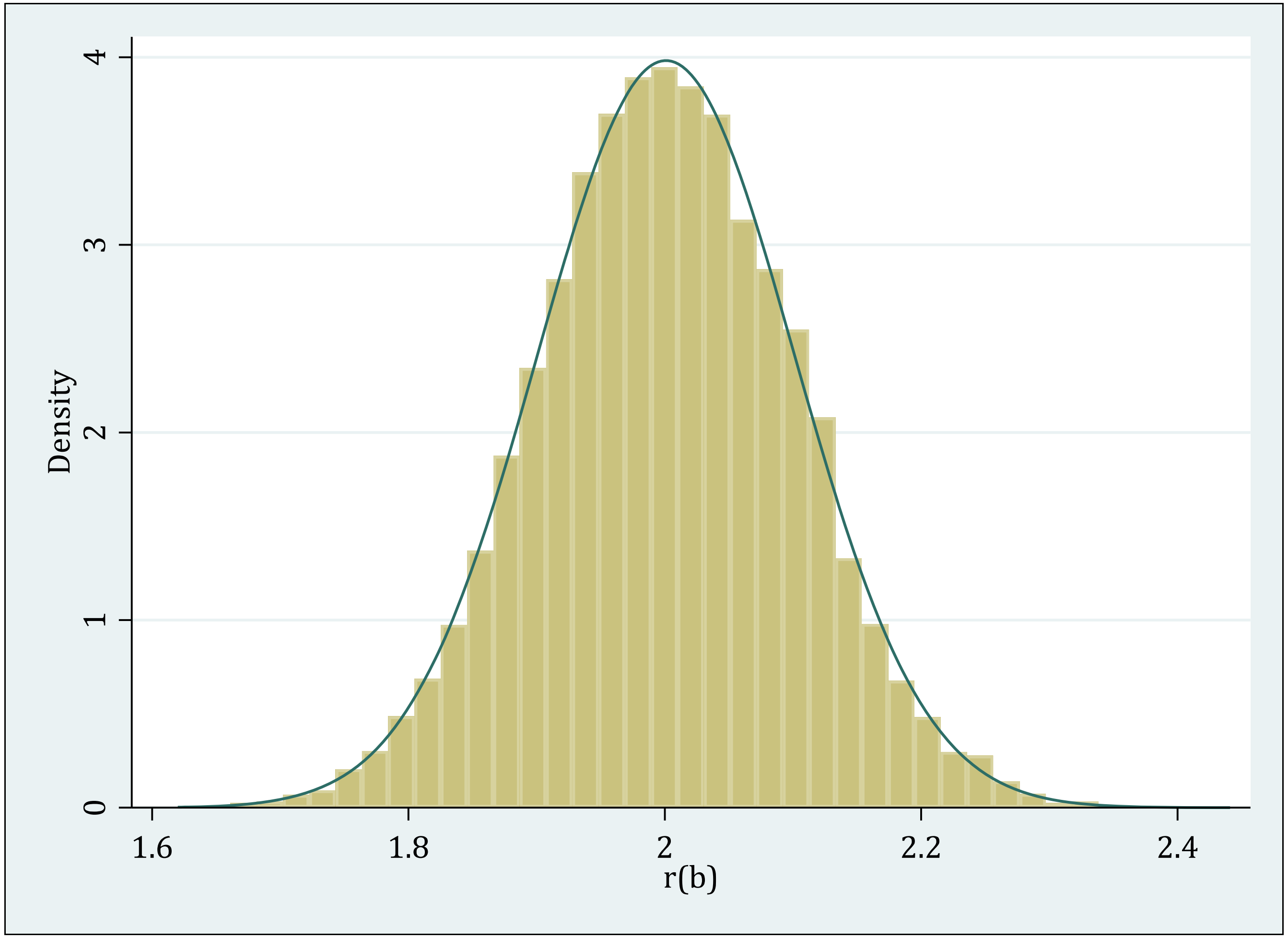

图 6.14　$\hat{\beta}$的分布(样本容量为 1000)

当样本容量增加为 1000 时，$\hat{\beta}$的真实分布已非常接近于正态分布，可以很放心地使用大样本理论进行统计推断。

上述蒙特卡罗模拟验证了 OLS 估计量的一致性与渐近正态性；这些性质即使在扰动项不服从正态分布的情况下也成立，使得大样本理论具有很大的适用性与稳健性。