# 大样本回归

大样本理论研究的是当样本容趋于无穷大时统计量的性质。

## 随机收敛

依概率 a 收敛：在随机序列中，任意给定$\forall\epsilon > 0$，当样本数量 n 越来越大时，随机变量$x_n$落在区间$(a-\epsilon, a+\epsilon)$之外的概率收敛于0。

依均方收敛：$\lim_{n\to\infty} \mathrm{E}\left(x_n\right) = a$，$\lim_{n\to\infty} \mathrm{Var}\left(x_n\right) = 0$。依均方收敛是依概率收敛的充分条件。

依分布收敛：$\forall c$，$\lim_{n\to\infty} F_n\left(c\right) = F\left(c\right)$，即随着样本量增大，两个随机变量的密度函数越来越像。

## 大数定律和中心极限定理

### 大数定律

样本均值会随着 n 的不断增大，依概率收敛，即足够大的样本的均值能近似反映总体的均值，能用频率近似代替概率，用样本均值近似代替总体均值。

> 如果考试没有要求，实际应用中大数定律不需要会推导或者深入理解，大数定律的作用是保证抽样调查、蒙特卡洛模拟等方式有效性的理论基础。

### 中心极限定理

极简定义：**当样本量足够大时，样本均值的分布慢慢变成正态分布。**

描述性定义（参考3Blue1Brown）：对于任意分布的随机变量 $X$，样本均值$\mu$和方差$\sigma^2$存在，当正整数N趋近于无穷大时，那么$M=\frac{\left(\sum_{i=1}^{N} X_n\right)-N\cdot\mu}{\sigma\sqrt{N}}$这个值落在$a,b(b>a)$区间的概率为标准正态概率密度函数在 a 与 b 区间的定积分。

$$\lim_{N\to\infty} P\left(a<M<b\right)=\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx \tag{1}$$

书本定义：若$\{x_n\}_{n=1}^{\infty}$为独立同分布的随机序列，且$\mathrm{E}\left(x_1\right)=\mu$，$\mathrm{Var}\left(x_1\right)=\sigma^2$存在，则 $\sqrt{n}\left(\bar{x}_n-\mu\right)\to N\left(0,\sigma^2\right)$

## 统计量的大样本性质

均方误差（Mean Squred Error）：
$\mathrm{MSE}\left(\hat{\beta}\right)=E\left[\left(\hat{\beta}-\beta\right)^2\right]=\mathrm{Var}\left(\hat{\beta}\right)+\left[E\left(\hat{\beta}\right)-\beta\right]^2$

其中$\hat{\beta}$是一维参数$\beta$的估计量，$\left[E\left(\hat{\beta}\right)-\beta\right]$为以估计量 估计参数的误差。

一致估计量：如果$\hat{\boldsymbol{\beta}}$依概率收敛于$\boldsymbol{\beta}$，则估计量$\hat{\boldsymbol{\beta}}$是参数$\boldsymbol{\beta}$的一致估计量。

## 随机过程的性质

现在，将$\{x_n\}_{n=1}^{\infty}$称为随机过程（stochastic process），如果下标为时间，则记录为 $\{x_t\}_{t=1}^{\infty}$，即时间序列（time series）。

## 严格平稳过程

若将$\{x_{t_1}, x_{t_2},\cdots,x_{t_m}\}$这个实时间序列的时间下标全部向前或向后移动 k 期，不改变其分布，则称$\{x_t\}_{t=1}^{\infty}$为严格平稳过程。

如果$E(x_t)$不依赖于 t，且 $\mathrm{Cov}(x_t,x_{t+k})$只依赖于k不依赖于 t，则成随机过程为弱平稳过程或协方差平稳过程。

如果一个协方差平稳过程对于任意 t 都有 $E(x_t)=0$，又被称为白噪声过程。

> 平稳的时间序列的性质不随观测时间的变化而变化。因此具有趋势或季节性的时间序列不是平稳时间序列——趋势和季节性使得时间序列在不同时段呈现不同性质。与它们相反，白噪声则是平稳的——不管观测的时间如何变化，它看起来都应该是一样的。一般而言，一个平稳的时间序列从长期来看不存在可预测的特征。
>
> 来源：《预测：方法与实践》

## 渐进独立性

随机过程没有长记忆，即如果给予足够的时间，则系统的演化将忘记自己是从什么初始条件起步的，这就是渐进独立。

假设$\{x_t\}_{t=1}^{\infty}$时渐进独立的严格平稳过程，且$E\left(x_i\right)=\mu$，则样本均值$\bar{x}_n$是总体均值的一致性估计。

## 大样本 OLS 的假定

无需假定严格外生性何正态随机扰动项。

- 线性假定
- 渐进独立的平稳过程
- 前定解释变量：在某个时间点或情况中，解释变量与误差项的当前和未来值不相关。这与严格外生性不同，严格外生性要求解释变量在所有时候和误差项都不相关。
- 秩条件假定：确保 OLS 的最小二乘解是唯一的，这意味着每个参数估计有一个明确的数值，避免多重共线性。

## 假设检验

$H_0: \beta_k=\bar{\beta}_k$

$t_k \equiv \frac{b_k-\bar{\beta}_k}{\mathrm{SE}^*\left(b_k\right)}$

$SE^*(b_k)$被称为稳健标准误。为了应对异方差性问题，统计学家引入了稳健标准误（robust standard errors），在误差项存在异方差性时依然提供有效的标准误估计，从而使得假设检验和置信区间仍然是可靠的。

统计量$t_k$被称为稳健t比值，服从标准正态分布而不是 t 分布。$|t_k|$越大，越倾向于拒绝原假设。

## Stata 命令

```
reg lntc lnq lnpl lnpk lnpf, robust
```

使用稳健标准误进行回归，得到的回归系数完全相同，但假设检验相关项会因此变动。

```
testnl _b[lnpl] = _b[lnq]^2
```

非线性假设检验，检验lnpl的系数是lnq系数的平方，`_b[变量名]` 用于引用特定变量的估计系数。