

第 3 章 数学回顾

3.1 微积分

1. 导数

对于一元函数 $y = f(x)$, 记其一阶导数(first derivative)为 $\frac{dy}{dx}$ 或 $f'(x)$, 其定义为

$$\frac{dy}{dx} \equiv f'(x) \equiv \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} \equiv \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad (3.1)$$

“ \equiv ”表示“定义”。几何上，(一阶)导数就是函数 $y = f(x)$ 在 x 处的切线斜率，参见图 3.1。

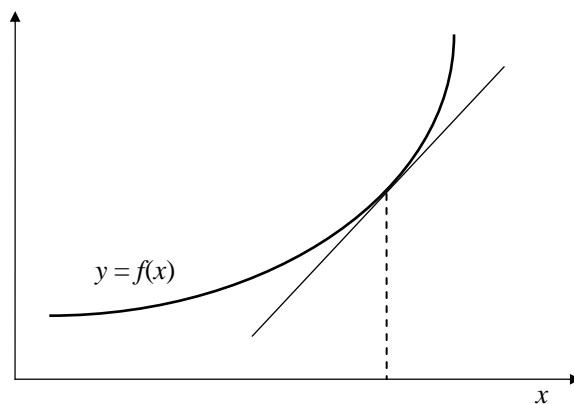


图 3.1 导数的示意图

一阶导数 $f'(x)$ 仍是 x 的函数，可定义 $f'(x)$ 的导数，即二阶导数 (second derivative):

$$\frac{d^2 y}{dx^2} \equiv f''(x) \equiv \frac{d\left(\frac{dy}{dx}\right)}{dx} \equiv [f'(x)]' \quad (3.2)$$

直观上，二阶导数表示切线斜率的变化速度，即曲线 $f(x)$ 的弯曲程度，也称“曲率” (curvature)。

2.一元最优化

计量中常见的两种估计方法为最小二乘法与最大似然估计。二者都是最优化问题 (optimization)，前者为最小化问题 (minimization)，后者为最大化问题 (maximization)。

考虑无约束的一元最大化问题(参见图 3.2),

$$\max_x f(x) \quad (3.3)$$

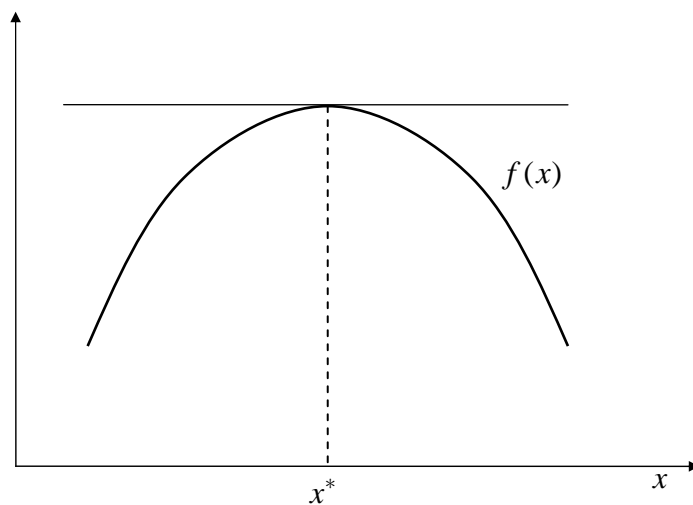


图 3.2 最大化的示意图

函数 $f(x)$ 在山峰顶端 x^* 处达到最大值。在山顶 x^* 处, $f(x)$ 的切线恰好为水平线, 故切线斜率为 0。

故一元最大化问题的必要条件为

$$f'(x^*) = 0 \quad (3.4)$$

称为一阶条件(first order condition)。考虑无约束一元最小化问题(参见图 3.3),

$$\min_x f(x) \quad (3.5)$$

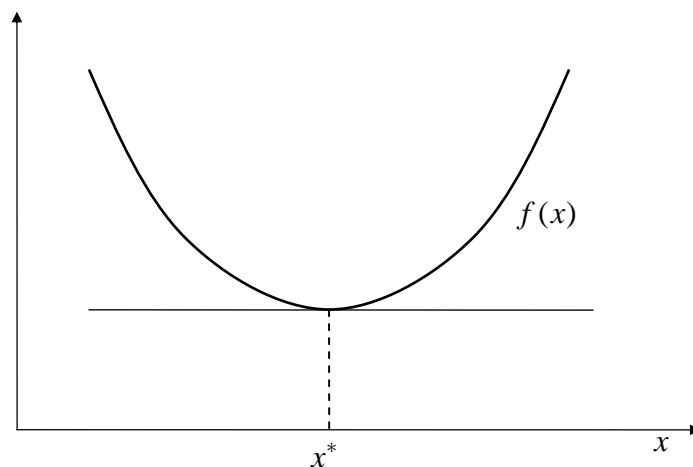


图 3.3 最小化的示意图

最小化问题的一阶条件与最大化问题相同，都要求在最优值 x^* 处的切线斜率为 0，即 $f'(x^*) = 0$ 。

二者的区别仅在于二阶条件(second order condition)，即最大化要求二阶导数 $f''(x^*) \leq 0$ ，而最小化要求 $f''(x^*) \geq 0$ 。

一般假设二阶条件满足，主要关注一阶条件。

3.偏导数

对于多元函数 $y = f(x_1, x_2, \dots, x_n)$ ，定义 y 对于 x_1 的偏导数(partial derivative)为

$$\frac{\partial y}{\partial x_1} \equiv \frac{\partial f(x_1, x_2, \dots, x_n)}{\partial x_1} \equiv \lim_{\Delta x_1 \rightarrow 0} \frac{f(x_1 + \Delta x_1, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{\Delta x_1} \quad (3.6)$$

在计算 y 对 x_1 的一阶偏导数时，首先给定 x_2, \dots, x_n 为常数(视为参数)，则 $y = f(x_1, x_2, \dots, x_n)$ 可看成 x_1 的一元函数 $y = f(x_1, \cdot)$ 。

$\frac{\partial y}{\partial x_1}$ 便是此“一元函数” $y = f(x_1, \cdot)$ 的导数。

类似地，可定义 y 对 x_i ($i = 2, \dots, n$) 的偏导数 $\frac{\partial y}{\partial x_i}$ 。

在经济学中，如果 $y = f(x_1, x_2, \dots, x_n)$ 为效用函数，则 $\frac{\partial y}{\partial x_1}$ 表示商品 x_1 所能带来的边际效用(marginal utility)。

如果 $y = f(x_1, x_2, \dots, x_n)$ 为生产函数，则 $\frac{\partial y}{\partial x_1}$ 表示生产要素 x_1 所能带来的边际产出(marginal output)。

4.多元最优化

考虑无约束的多元最大化问题，

$$\max_{\mathbf{x}} f(\mathbf{x}) \equiv f(x_1, x_2, \dots, x_n) \quad (3.7)$$

其中， $\mathbf{x} \equiv (x_1, x_2, \dots, x_n)$ 。

一阶条件要求在最优值 \mathbf{x}^* 处，所有偏导数均为 0:

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_1} = \frac{\partial f(\mathbf{x}^*)}{\partial x_2} = \dots = \frac{\partial f(\mathbf{x}^*)}{\partial x_n} = 0 \quad (3.8)$$

多元最小化的一阶条件与此相同。

此一阶条件要求在最优值 \mathbf{x}^* 处，曲面 $f(\mathbf{x})$ 在各个方向的切线斜率都为 0。

5.积分

考虑计算连续函数 $y = f(x)$ 在区间 $[a, b]$ 上的面积，参见图 3.4。

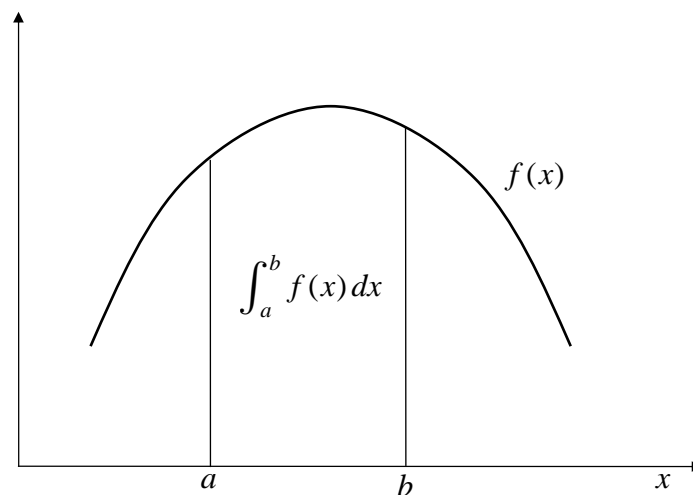


图 3.4 定积分的示意图

将区间 $[a, b]$ 划分为 n 等分，即 $[a, x_1], (x_1, x_2], \dots, (x_{n-1}, b]$ ，从每个

区间 $(x_{i-1}, x_i]$ ($i = 1, \dots, n$) 中任取一点 ξ_i (记 a 为 x_0 , 而 b 为 x_n)。

每个区间的长度为 $\Delta x \equiv \frac{b-a}{n}$, 此面积近似等于 $\sum_{i=1}^n f(\xi_i) \Delta x$ 。

不断细分这些区间, 让 $n \rightarrow \infty$, 可得此面积的精确值, 即函数 $f(x)$ 在区间 $[a, b]$ 上的定积分(definite integral):

$$\int_a^b f(x) dx \equiv \lim_{n \rightarrow \infty} \sum_{i=1}^n f(\xi_i) \Delta x \quad (3.9)$$

在极限处, 将 Δx 记为 dx , 将求和符号 Σ (英文 Summation) 记为 \int , 由大写字母 S 向上拉长而成。

定积分的实质就是求和(只不过是无穷多项之和)。

3.2 线性代数

1. 矩阵

将 $m \times n$ 个实数排列成如下矩状的阵形,

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad (3.10)$$

称 \mathbf{A} 为 $m \times n$ 级矩阵(matrix), m 为矩阵 \mathbf{A} 的行数(row dimension), n 为矩阵 \mathbf{A} 的列数(column dimension)。 \mathbf{A} 中元素 a_{ij} 表示矩阵 \mathbf{A} 的第 i 行、第 j 列元素。

矩阵 \mathbf{A} 有时也记为 $\mathbf{A}_{m \times n}$ ，以强调矩阵的维度。

如果 \mathbf{A} 中所有元素都为 0，则称为零矩阵(zero matrix)，记为 $\mathbf{0}$ 。

零矩阵在矩阵运算中的作用，相当于 0 在数的运算中的作用。

2. 方阵

如果 $m = n$ ，则称 \mathbf{A} 为 n 级方阵(square matrix)，即

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \quad (3.11)$$

称 $a_{11}, a_{22}, \dots, a_{nn}$ 为主对角线上的元素(diagonal elements), 而 \mathbf{A} 中的其他元素为非主对角线元素(off-diagonal elements)。

如果方阵 \mathbf{A} 中的元素满足 $a_{ij} = a_{ji}$ (任意 $i, j = 1, \dots, n$), 则称矩阵 \mathbf{A} 为对称矩阵(symmetric matrix)。

如果方阵 \mathbf{A} 的非主对角线元素全部为 0, 则称为对角矩阵(diagonal matrix):

$$\mathbf{A} \equiv \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \quad (3.12)$$

如果一个 n 级对角矩阵的主对角线元素都为 1，则称为 n 级单位矩阵(identity matrix)，记为 \mathbf{I} 或 \mathbf{I}_n ：

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n \times n} \quad (3.13)$$

单位矩阵在矩阵运算中的作用，相当于 1 在数的运算中的作用。

3. 矩阵的转置

如果将矩阵 $\mathbf{A} = (a_{ij})_{m \times n}$ 的第 1 行变为第 1 列，第 2 行变为第 2 列，……，第 m 行变为第 m 列，可得其转置矩阵(transpose)，记为 \mathbf{A}' (英文读为 \mathbf{A} prime)，其维度为 $n \times m$ 。

矩阵 \mathbf{A}' 的 (i, j) 元素 $(\mathbf{A}')_{ij}$ 正好是矩阵 \mathbf{A} 的 (j, i) 元素 $(\mathbf{A})_{ji}$ ，即

$$(\mathbf{A}')_{ij} \equiv (\mathbf{A})_{ji} \quad (3.14)$$

如果 \mathbf{A} 为对称矩阵，则 \mathbf{A} 的转置还是它本身，即 $\mathbf{A}' = \mathbf{A}$ 。

矩阵转置的转置仍是它本身，即 $(\mathbf{A}')' = \mathbf{A}$ 。

4. 向量

如果 $m = 1$ ，则矩阵 $A_{1 \times n}$ 为 n 维行向量(row vector)。

如果 $n = 1$ ，则矩阵 $A_{m \times 1}$ 为 m 维列向量(column vector)。

向量是矩阵的特例。

考察 n 维列向量 $\mathbf{a} = (a_1 \ a_2 \ \cdots \ a_n)'$ 与 $\mathbf{b} = (b_1 \ b_2 \ \cdots \ b_n)'$ 。

向量 \mathbf{a} 与 \mathbf{b} 的内积(inner product)或点乘(dot product)可定义为

$$\mathbf{a}'\mathbf{b} \equiv (a_1 \quad a_2 \quad \cdots \quad a_n) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \equiv a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^n a_i b_i \quad (3.15)$$

如果 $\mathbf{a}'\mathbf{b} = 0$ ，则称向量 \mathbf{a} 与 \mathbf{b} 正交(orthogonal)，意味着两个向量在 n 维向量空间中相互垂直(夹角为90度)，参见图3.5。

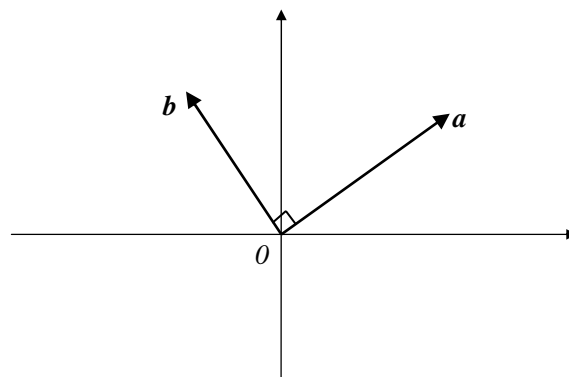


图 3.5 正交的向量

任何形如 $\sum_{i=1}^n a_i b_i$ 的乘积求和，都可写为向量内积 $\mathbf{a}'\mathbf{b}$ 的形式。

平方和 $\sum_{i=1}^n a_i^2$ 可写为 $\mathbf{a}'\mathbf{a}$ ：

$$\mathbf{a}'\mathbf{a} \equiv (a_1 \quad a_2 \quad \cdots \quad a_n) \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \equiv a_1^2 + a_2^2 + \cdots + a_n^2 = \sum_{i=1}^n a_i^2 \quad (3.16)$$

5. 矩阵的加法

如果两个矩阵的维度相同，则可相加。

对于 $m \times n$ 级矩阵 $\mathbf{A} = (a_{ij})_{m \times n}$ ， $\mathbf{B} = (b_{ij})_{m \times n}$ ，矩阵 \mathbf{A} 与 \mathbf{B} 之和定义为两个矩阵相应元素之和，即

$$\mathbf{A} + \mathbf{B} \equiv (a_{ij})_{m \times n} + (b_{ij})_{m \times n} \equiv (a_{ij} + b_{ij})_{m \times n} \quad (3.17)$$

矩阵加法满足以下规则：

- (1) $\mathbf{A} + \mathbf{0} = \mathbf{A}$ (加上零矩阵不改变矩阵)
- (2) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ (加法交换律)
- (3) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$ (加法结合律)
- (4) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$ (转置为线性运算)

6. 矩阵的数乘

矩阵 $\mathbf{A} = (a_{ij})_{m \times n}$ 与实数 k 的数乘(scalar multiplication)定义为此实数 k 与矩阵 $\mathbf{A} = (a_{ij})_{m \times n}$ 每个元素的乘积:

$$k\mathbf{A} \equiv k(a_{ij})_{m \times n} \equiv (ka_{ij})_{m \times n} \quad (3.18)$$

7. 矩阵的乘法

如果矩阵 \mathbf{A} 的列数与矩阵 \mathbf{B} 的行数相同, 则可以定义矩阵乘积(matrix multiplication) $\mathbf{A} \times \mathbf{B}$, 简记 \mathbf{AB} 。

假设矩阵 $\mathbf{A} = (a_{ij})_{m \times n}$ ，矩阵 $\mathbf{B} = (b_{ij})_{n \times q}$ ，则矩阵乘积 \mathbf{AB} 的 (i, j) 元素即为矩阵 \mathbf{A} 第 i 行与矩阵 \mathbf{B} 的第 j 列的内积：

$$(\mathbf{AB})_{ij} \equiv (a_{i1} \quad a_{i2} \quad \cdots \quad a_{in}) \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{pmatrix} = \sum_{k=1}^n a_{ik} b_{kj} \quad (3.19)$$

矩阵乘法不满足交换律，即一般来说， $\mathbf{AB} \neq \mathbf{BA}$ 。

只有矩阵 \mathbf{B} 的列数 q 等于矩阵 \mathbf{A} 的行数 m ， $\mathbf{B}_{n \times q} \mathbf{A}_{m \times n}$ 才有定义。

在做矩阵乘法时，需区分左乘 (premultiplication) 与右乘 (postmultiplication)。A 左乘 B 为 AB ，而 A 右乘 B 为 BA 。

矩阵的乘法满足以下规则：

$$(1) \mathbf{IA} = \mathbf{A}, \quad \mathbf{AI} = \mathbf{A} \quad (\text{乘以单位矩阵不改变矩阵})$$

$$(2) (\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \quad (\text{乘法结合律})$$

$$(3) \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \quad (\text{乘法分配律})$$

$$(4) (\mathbf{AB})' = \mathbf{B}'\mathbf{A}', \quad (\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}' \quad (\text{转置与乘积的混合运算})$$

8. 线性方程组

考虑由 n 个方程, n 个未知数构成的线性方程组:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \quad (3.20)$$

$(x_1 \ x_2 \ \cdots \ x_n)$ 为未知数。根据矩阵乘法定义, 可将上式写为

$$\underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_A \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_x = \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}}_b \quad (3.21)$$

记上式中的相应矩阵分别为 \mathbf{A} ， \mathbf{x} 与 \mathbf{b} ，可得

$$\mathbf{Ax} = \mathbf{b} \quad (3.22)$$

如将此方程左边的方阵 \mathbf{A} “除”到右边去，可得 \mathbf{x} 的解。为此，引入逆矩阵的概念。

9. 逆矩阵

对于 n 级方阵 \mathbf{A} ，如果存在 n 级方阵 \mathbf{B} ，使得 $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$ ，则称 \mathbf{A} 为可逆矩阵(invertible matrix)或非退化矩阵(nonsingular matrix)，而 \mathbf{B} 为 \mathbf{A} 的逆矩阵(inverse matrix)，记为 \mathbf{A}^{-1} 。

逆矩阵的逆矩阵还是矩阵本身，即 $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$ 。

方阵 \mathbf{A} 可逆的充分必要条件为其行列式 $|\mathbf{A}| \neq 0$ 。

如果 \mathbf{A} 可逆，则其逆矩阵 \mathbf{A}^{-1} 是唯一的。

假设方程(3.22)中的矩阵 \mathbf{A} 可逆，则在该方程两边同时左乘其逆矩阵 \mathbf{A}^{-1} 可得：

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \Rightarrow \mathbf{I}\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \Rightarrow \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \quad (3.23)$$

矩阵求逆满足以下规则：

$$(1)(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1} \quad (\text{求逆与转置可交换次序})$$

$$(2)(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad (\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$$

(求逆与乘积的混合运算)

10. 矩阵的秩

考虑两个 n 维列向量 \mathbf{a}_1 与 \mathbf{a}_2 。

如果 \mathbf{a}_1 正好是 \mathbf{a}_2 的固定倍数，则在向量组 $\{\mathbf{a}_1, \mathbf{a}_2\}$ 中，真正含有信息的只是其中的一个向量。

更一般地，考虑由 K 个 n 维向量构成的向量组 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ ，如果存在 c_1, c_2, \dots, c_K 不全为零，使得

$$c_1 \mathbf{a}_1 + c_2 \mathbf{a}_2 + \dots + c_K \mathbf{a}_K = \mathbf{0} \quad (3.24)$$

则称向量组 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ 线性相关(linearly dependent)。

如果 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ 线性相关，则其中至少有一个向量可写为其他向量的线性组合(linear combination)，也称线性表出。

反之，如果方程(3.24)必然意味着 $c_1 = c_2 = \dots = c_K = 0$ ，则称 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ 线性无关(linearly independent)。

如果 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ 线性相关，但从中去掉一个向量后，就变得线性无关，则 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ 中正好有 $(K-1)$ 个向量真正含有信息，称

$(K-1)$ 为此向量组的秩。

向量组 $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ 的极大线性无关部分组所包含的向量个数，称为该向量组的秩(rank)。

对于 $m \times n$ 级矩阵 \mathbf{A} ，可将其 n 个列向量看成一个向量组，称此列向量组的秩为矩阵 \mathbf{A} 的**列秩**(column rank)。

如果矩阵 $\mathbf{A}_{m \times n}$ 的列秩正好等于 n ，则称矩阵 \mathbf{A} **满列秩**(full column rank)。

将矩阵 $\mathbf{A}_{m \times n}$ 的 m 个行向量看成一个向量组，称此行向量组的秩

为矩阵 \mathbf{A} 的行秩(row rank)。

任何矩阵的行秩与列秩一定相等，称为矩阵的秩(matrix rank)。

11. 二次型

对于 n 维列向量 $\mathbf{x} = (x_1 \ x_2 \ \cdots \ x_n)'$ ，如何度量它与零向量 $\mathbf{0}$ 的距离？最简单的方法为欧几里得距离 (Euclidean distance)，即内积

$$x_1^2 + x_2^2 + \cdots + x_n^2 = (x_1 \ x_2 \ \cdots \ x_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{x}'\mathbf{x} \quad (3.25)$$

上式可写为

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}_{\mathbf{I}_n} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{x}' \mathbf{I}_n \mathbf{x} \quad (3.26)$$

单位矩阵 \mathbf{I}_n 相当于给予此内积的每一项相同的权重。

如果允许不同的权重，则可使用任意对称矩阵 \mathbf{A} ，构成如下二次型(quadratic form):

$$f(x_1, x_2, \dots, x_n) = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j \quad (3.27)$$

对称矩阵 \mathbf{A} 称为二次型的矩阵。所谓二次型，就是 x_1, x_2, \dots, x_n 的二次齐次多项式函数：

$$\begin{aligned} f(x_1, x_2, \dots, x_n) = & a_{11}x_1^2 + 2a_{12}x_1x_2 + \cdots + 2a_{1n}x_1x_n \\ & + a_{22}x_2^2 + \cdots + 2a_{2n}x_2x_n \\ & + \cdots \cdots \cdots \cdots \cdots \\ & + a_{nn}x_n^2 \end{aligned} \quad (3.28)$$

任意二次型(3.28)，都可写为 $\mathbf{x}'\mathbf{A}\mathbf{x}$ 的形式，其中 \mathbf{A} 为对称矩阵。

例如，考虑一般的二维二次型：

$$f(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \quad (3.29)$$

此二次型可写为：

$$f(x_1, x_2) = (x_1 \ x_2) \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (3.30)$$

其中， $a_{21} = a_{12}$ 。

如果 $\mathbf{x} = \mathbf{0}$ ，则二次型 $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ 。

当 $\mathbf{x} \neq \mathbf{0}$ 时，二次型 $\mathbf{x}'\mathbf{A}\mathbf{x}$ 如何取值？

首先，考虑一维二次型：

$$f(x_1) = a_{11}x_1^2 = x_1' a_{11} x_1 \quad (3.31)$$

此二次型的矩阵就是常数 a_{11} 。如果 $a_{11} > 0$ ，则只要 $x_1 \neq 0$ ，就有 $f(x_1) = a_{11}x_1^2 > 0$ 。此时，称此二次型为“正定”(positive definite)，其图形为开口向上的抛物线(参见图 3.6)：

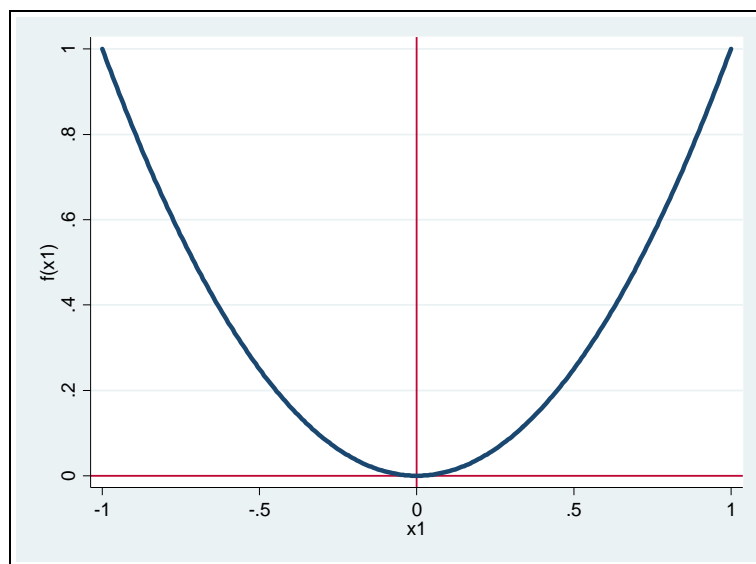


图 3.6 正定的一维二次型

反之，如果 $a_{11} < 0$ ，则只要 $x_1 \neq 0$ ，就有 $f(x_1) = a_{11}x_1^2 < 0$ 。此时，称此二次型为“负定” (negative definite)，其图形为开口向下的抛物线(参见图 3.7)：

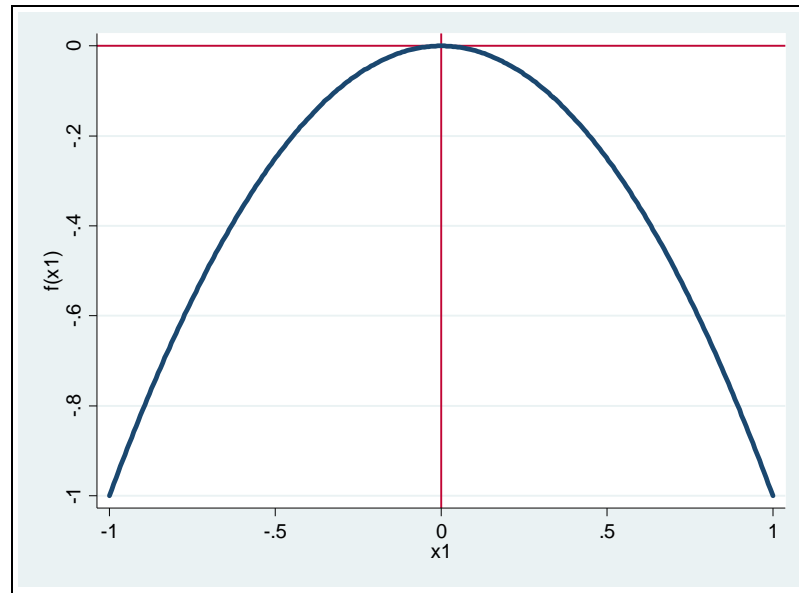


图 3.7 负定的一维二次型

对于二维的二次型，其取值的确定性更为复杂。

例如, 对于 x_1, x_2 不全为 0, 二次型 $(x_1^2 + x_2^2)$ 一定为正, 故为正定; 二次型 $(-x_1^2 - x_2^2)$ 一定为负, 故为负定; 而二次型 $(x_1^2 - x_2^2)$ 则可正可负, 称为“不定” (indefinite)。

这依然没有穷尽所有情形。考虑以下二次型:

$$f(x_1, x_2) = x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2 \quad (3.32)$$

二次型 $(x_1 + x_2)^2 \geq 0$ (必然非负); 但即使 x_1, x_2 不全为 0, 也可能出现 $(x_1 + x_2)^2 = 0$, 只要 $x_1 = -x_2$ 。此时, 称此二次型为“半正定” (positive semidefinite)。

另一方面，二次型 $-(x_1 + x_2)^2 \leq 0$ (必然非正)；但即使 x_1, x_2 不全为0，也可能出现 $(x_1 + x_2)^2 = 0$ ，只要 $x_1 = -x_2$ 。此时，称此二次型为“半负定”(negative semidefinite)。

在一般的 n 维情况下，给定对称矩阵 \mathbf{A} ，针对二次型 $\mathbf{x}'\mathbf{A}\mathbf{x}$ 的取值确定性，可引入以下定义。

(1) 对于任意非零列向量 \mathbf{x} ，都有 $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ ，则对称矩阵 \mathbf{A} 为正定矩阵(positive definite)。

(2) 对于任意非零列向量 \mathbf{x} ，都有 $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ ，则对称矩阵 \mathbf{A} 为半正定矩阵(positive semidefinite)。

(3) 对于任意非零列向量 \mathbf{x} ，都有 $\mathbf{x}'\mathbf{A}\mathbf{x} < 0$ ，则对称矩阵 \mathbf{A} 为负定矩阵(negative definite)。

(4) 对于任意非零列向量 \mathbf{x} ，都有 $\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$ ，则对称矩阵 \mathbf{A} 为半负定矩阵(negative semidefinite)。

正定矩阵一定半正定，而负定矩阵也一定半负定。

如果对称矩阵 \mathbf{A} 为正定矩阵，则该矩阵可通过线性变换转换为一个主对角线元素全为正数的对角矩阵；而这些主对角线元素正好是矩阵 \mathbf{A} 的特征值，故正定矩阵 \mathbf{A} 一定可逆，即逆矩阵 \mathbf{A}^{-1} 存在。

线性变换后的正定二次型可写为

$$f(x_1, x_2, \dots, x_n) = \alpha_{11}x_1^2 + \alpha_{22}x_2^2 + \dots + \alpha_{nn}x_n^2 \quad (3.33)$$

其中， $\alpha_{11}, \dots, \alpha_{nn}$ 全部为正数，故当 x_1, \dots, x_n 不全为 0 时， $f(x_1, x_2, \dots, x_n)$ 必然大于 0。

如果 $\alpha_{11}, \dots, \alpha_{nn}$ 全部为正数或 0，则 **A** 为半正定矩阵。

如果 $\alpha_{11}, \dots, \alpha_{nn}$ 全部为负数，则 **A** 为负定矩阵。

如果 $\alpha_{11}, \dots, \alpha_{nn}$ 全部为负数或 0，则 **A** 为半负定矩阵。

反之，如果 $\alpha_{11}, \dots, \alpha_{nn}$ 有正有负，则 \mathbf{A} 为不定的(indefinite)，其二次型 $\mathbf{x}'\mathbf{A}\mathbf{x}$ 的取值可正可负。

在计量中，常使用形如 $\mathbf{x}'[\text{Var}(\mathbf{x})]^{-1}\mathbf{x}$ 的二次型。

其中， \mathbf{x} 为 n 维随机向量，而 $[\text{Var}(\mathbf{x})]^{-1}$ 为其协方差矩阵 $\text{Var}(\mathbf{x})$ 的逆矩阵。

二次型 $\mathbf{x}'[\text{Var}(\mathbf{x})]^{-1}\mathbf{x}$ 的直观含义是，以 $[\text{Var}(\mathbf{x})]^{-1}$ 为权重，将 \mathbf{x} 到零向量 $\mathbf{0}$ (原点)的距离标准化(避免受到 \mathbf{x} 度量单位的影响)。

在一维情况下，此二次型可写为

$$\mathbf{x}'[\text{Var}(\mathbf{x})]^{-1}\mathbf{x} = \frac{x^2}{\text{Var}(x)} = \left(\frac{x}{\sqrt{\text{Var}(x)}} \right)^2 \quad (3.34)$$

此一维二次型度量的是， x 离原点 0 有几个标准差的距离。

比如， $\frac{x}{\sqrt{\text{Var}(x)}} = 2$ ，则 x 离原点 0 有两个标准差的距离(以标准差 $\sqrt{\text{Var}(x)}$ 为度量单位)。

3.3 概率与条件概率

1. 概率

概率为在大量重复实验下，事件发生的频率趋向的某个稳定值。

记事件“下雨”为 A ，其发生的“概率”(probability)为 $P(A)$ 。

2. 条件概率

例 已知明天会出太阳，则下雨的概率有多大？

记事件“出太阳”为 B ，则在出太阳的前提条件下，降雨的**条件概率**(conditional probability)为

$$P(A|B) \equiv \frac{P(AB)}{P(B)} \quad (3.35)$$

AB 表示事件 A 与 B 同时发生(即交集, 也记为 $A \cap B$), 故 $P(AB)$ 为“太阳雨”的概率, 参见图 3.8。条件概率是计量的重要概念之一。

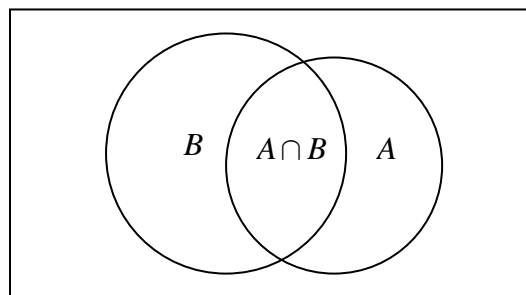


图 3.8 条件概率的示意图

例 股市崩盘的可能性为无条件概率；在已知经济已陷入严重衰退的情况下，股市崩盘的可能性则为条件概率。

3. 独立事件

如果条件概率等于无条件概率, $P(A|B) = P(A)$, 即 B 是否发生不影响 A 的发生, 则称 A, B 为相互独立的随机事件。

此时, $P(A|B) \equiv \frac{P(AB)}{P(B)} = P(A)$, 故

$$P(AB) = P(A)P(B) \quad (3.36)$$

也可将此式作为独立事件的定义。

4. 全概率公式

如果事件组 $\{B_1, B_2, \dots, B_n\}$ ($n \geq 2$)两两互不相容, 但必有一件事发生, 且每件事的发生概率均为正数, 则对任何事件 A (无论 A 与 $\{B_1, B_2, \dots, B_n\}$ 是否有任何关系), 都有

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i) \quad (3.37)$$

全概率公式把世界分成了 n 个可能的情形 $\{B_1, B_2, \dots, B_n\}$, 再把每种情况下的条件概率 $P(A|B_i)$ “加权平均”而汇总成无条件概率(权重为每种情形发生的概率 $P(B_i)$)。

该式有助于理解后面的迭代期望定律。

3.4 分布与条件分布

1. 离散型概率分布

假设随机变量 X 的可能取值为 $\{x_1, x_2, \dots, x_k, \dots\}$, 其对应概率为 $\{p_1, p_2, \dots, p_k, \dots\}$, 即 $p_k \equiv P(X = x_k)$, 则称 X 为离散型随机变量, 其分布律可表示为

$$\begin{array}{cccccc} X & x_1 & x_2 & \cdots & x_k & \cdots \\ p & p_1 & p_2 & \cdots & p_k & \cdots \end{array} \quad (3.38)$$

其中, $p_k \geq 0$, $\sum_k p_k = 1$ 。常见的离散分布有两点分布(Bernoulli)、二项分布(Binomial)、泊松分布(Poisson)等。

2. 连续型概率分布

连续型随机变量 X 可以取任意实数, 其概率密度函数(probability density function, 简记 pdf) $f(x)$ 满足,

$$(1) f(x) \geq 0, \forall x;$$

$$(2) \int_{-\infty}^{+\infty} f(x) dx = 1;$$

$$(3) X \text{ 落入区间 } [a, b] \text{ 的概率为 } P(a \leq X \leq b) = \int_a^b f(x) dx。$$

概率密度函数的示意图参见图 3.9。

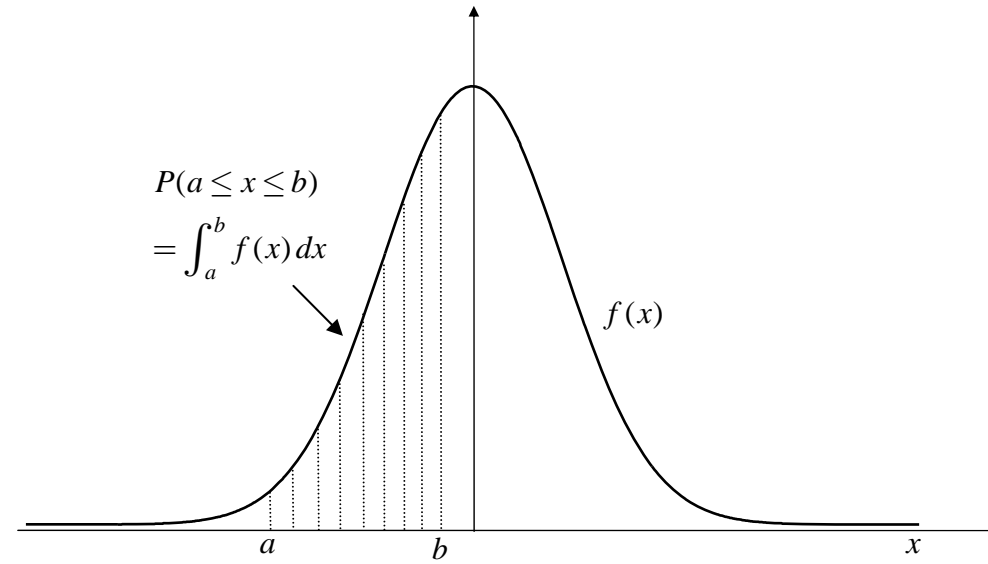


图 3.9 概率密度函数的示意图

定义累积分布函数(cumulative distribution function, 简记 cdf):

$$F(x) \equiv P(-\infty < X \leq x) = \int_{-\infty}^x f(t) dt \quad (3.39)$$

其中, t 为积分变量。 $F(x)$ 度量的是, 从 $-\infty$ 至 x 为止, 概率密度函数 $f(t)$ 曲线下的面积, 参见图 3.10。

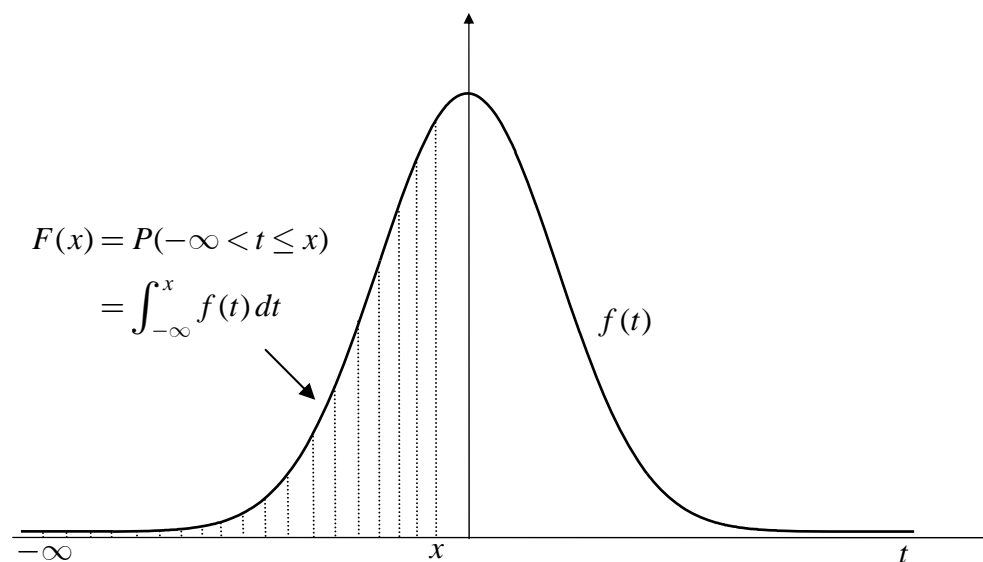


图 3.10 累积分布函数的示意图

3. 多维随机向量的概率分布

为研究变量间关系，常同时考虑两个或多个随机变量，即随机向量(random vector)。二维连续型随机向量 (X, Y) 的联合密度函数(joint pdf) $f(x, y)$ 满足：

(i) $f(x, y) \geq 0, \forall x, y;$

(ii) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1;$

(iii) (X, Y) 落入平面某区域 D 的概率为

$$P\{(X, Y) \in D\} = \iint_D f(x, y) dx dy。$$

二维随机向量的联合密度函数就像倒扣的草帽。落入平面某区域 D 的概率就是此草帽下在区域 D 之上的体积，参见图 3.11。

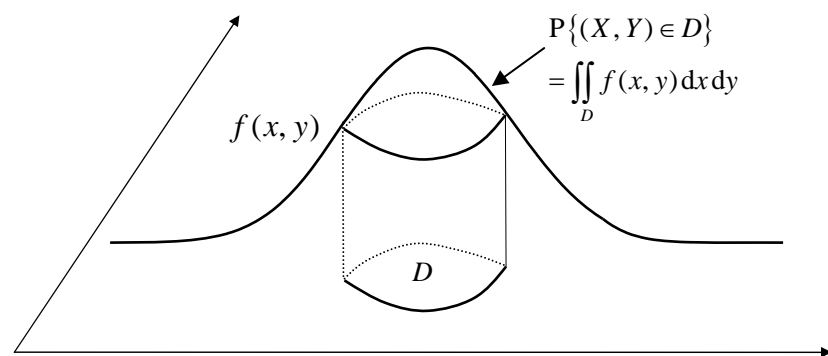


图 3.11 二维联合密度函数的示意图

n 维连续型随机向量 (X_1, X_2, \dots, X_n) 可由联合密度函数 $f(x_1, x_2, \dots, x_n)$ 来描述。

从二维联合密度 $f(x, y)$ ，可计算 X 的(一维)边缘密度函数 (marginal pdf):

$$f_x(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad (3.40)$$

即给定 $X = x$ ，把所有 Y 取值的可能性都“加总”起来(积分的本质就是加总)。

类似地，可以计算 Y 的(一维)边缘密度函数:

$$f_y(y) = \int_{-\infty}^{+\infty} f(x, y) dx \quad (3.41)$$

即给定 $Y = y$ ，把所有 X 取值的可能性都“加总”起来。

定义二维随机向量 (X, Y) 的累积分布函数为:

$$F(x, y) \equiv P(-\infty < X \leq x; -\infty < Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(t, s) dt ds \quad (3.42)$$

4. 条件分布

条件分布(conditional distribution)的概念对于计量至关重要。

考虑在 $X = x$ 条件下 Y 的条件分布, 记为 $Y|X = x$ 或 $Y|x$ 。

对于连续型分布, 此条件分布相当于在“草帽”(联合密度函数)上 $X = x$ 的位置垂直地切一刀所得的截面。

由于 X 为连续型随机变量, 事件 $\{X = x\}$ 发生的概率为0。

如何计算 $Y|X = x$ 的条件概率密度(conditional pdf)?

考虑 x 附近的小邻域 $[x - \varepsilon, x + \varepsilon]$ 。

计算在 $X \in [x - \varepsilon, x + \varepsilon]$ 条件下 Y 的累积分布函数，即 $P\{Y \leq y | X \in [x - \varepsilon, x + \varepsilon]\}$ (参见图 3.12)，然后让 $\varepsilon \rightarrow 0^+$ ，则可证明条件密度函数为，

$$f(y|x) = \frac{f(x,y)}{f_x(x)} \quad (3.43)$$

直观上，此公式与条件概率公式(3.35)类似。

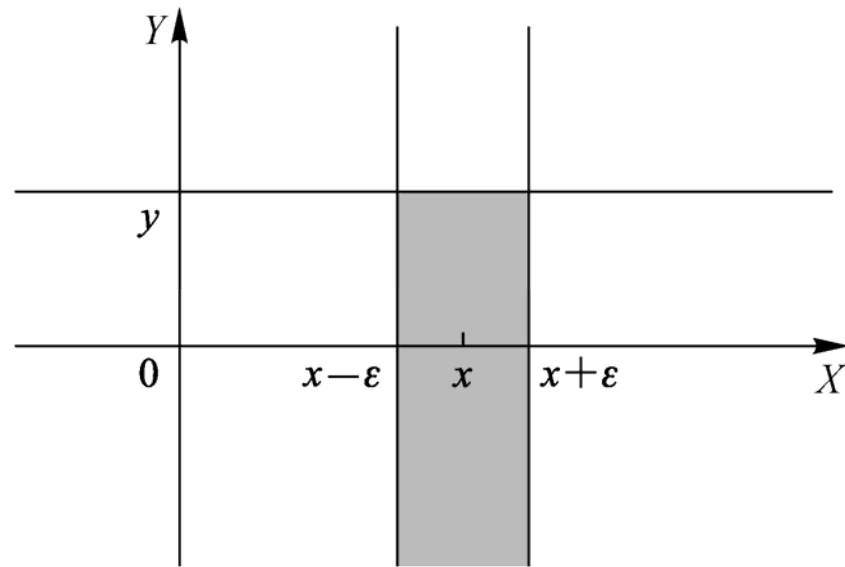


图 3.12 条件密度函数的计算

3.5 随机变量的数字特征

定义 对于分布律为 $p_k \equiv P(X = x_k)$ 的离散型随机变量 X ，其期望(expectation)为

$$E(X) \equiv \mu \equiv \sum_{k=1}^{\infty} x_k p_k \quad (3.44)$$

期望的直观含义就是对 x_k 进行加权平均，而权重为概率 p_k 。

定义 对于概率密度函数为 $f(x)$ 的连续型随机变量 X ，其期望为

$$E(X) \equiv \mu \equiv \int_{-\infty}^{+\infty} x f(x) dx \quad (3.45)$$

上式也是对 x 进行加权平均，权重为概率密度 $f(x)$ 。

有时称求期望这种运算为**期望算子**(expectation operator)。

期望算子满足**线性性**(linearity)，即对于任意常数 k 都有

$$E(X + Y) = E(X) + E(Y), \quad E(kX) = k E(X) \quad (3.46)$$

定义 随机变量 X 的**方差**(variance)为

$$\text{Var}(X) \equiv \sigma^2 \equiv E[X - E(X)]^2 \quad (3.47)$$

方差越大，则随机变量取值的波动幅度越大。

称方差的平方根为**标准差**(standard deviation)，记为 σ 。

在计算方差时，常利用以下简便公式：

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \quad (3.48)$$

常需要考虑两个变量之间的相关性，即一个随机变量的取值会对另一随机变量的取值有多大影响。

定义 随机变量 X 与 Y 的协方差(covariance)为

$$\text{Cov}(X, Y) \equiv \sigma_{XY} \equiv E[(X - E(X))(Y - E(Y))] \quad (3.49)$$

如果当随机变量 X 的取值大于(小于)其期望 $E(X)$ 时，随机变量 Y 的取值也倾向于大于(小于)其期望值 $E(Y)$ ，则 $\text{Cov}(X, Y) > 0$ ，二者存在正相关；

反之，如果当随机变量 X 的取值大于(小于)其期望 $E(X)$ 时，随机变量 Y 的取值反而倾向于小于(大于)其期望值 $E(Y)$ ，则 $\text{Cov}(X, Y) < 0$ ，二者存在**负相关**。

如果 $\text{Cov}(X, Y) = 0$ ，则说明二者**线性不相关**(uncorrelated)，但不一定**相互独立**(independent)，因为还可能存在着非线性的相关关系。

在计算协方差时，常使用以下简便公式：

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) \quad (3.50)$$

协方差的运算也满足线性性，可以证明：

$$\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z) \quad (3.51)$$

协方差的缺点是，它受 X 与 Y 计量单位的影响。为将其标准化，引入相关系数。

定义 随机变量 X 与 Y 的相关系数(correlation)为

$$\rho \equiv \text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (3.52)$$

相关系数一定介于 -1 与 1 之间，即 $-1 \leq \rho \leq 1$ 。

如果以上各定义式中的积分不收敛，则随机变量的数字特征可能不存在。

比如，自由度为 1 的 t 分布变量，其期望与方差都不存在。

更一般地，对于随机变量 X ，可以定义一系列的数字特征，即各阶矩(moment)的概念。

定义 一阶原点矩为 $E(X)$ (即期望)，二阶原点矩为 $E(X^2)$ ，三阶原点矩为 $E(X^3)$ ，四阶原点矩为 $E(X^4)$ ，等等。

定义 二阶中心矩为 $E[X - E(X)]^2$ (即方差)，三阶中心矩为 $E[X - E(X)]^3$ ，四阶中心矩为 $E[X - E(X)]^4$ ，等等。

一阶原点矩(期望)表示随机变量的平均值。

二阶中心矩(方差)表示随机变量的波动程度。

三阶中心矩表示随机变量密度函数的不对称性(偏度)。

四阶中心矩表示随机变量密度函数的最高处(山峰)有多“尖”及尾部有多“厚”(峰度)。

三、四阶中心矩还取决于变量的单位。为此，首先将变量“标准化”(即减去期望 μ ，再除以标准差 σ)，并引入以下定义。

定义 随机变量 X 的偏度(skewness)为 $E[(X - \mu)/\sigma]^3$ 。

如果随机变量为对称分布(比如，正态分布)，则其偏度为 0；因为根据微积分，奇函数在关于原点对称的区间上积分为 0。

定义 随机变量 X 的峰度(kurtosis)为 $E[(X - \mu)/\sigma]^4$ 。

对于正态分布，其峰度为 3。如果随机变量 X 的峰度大于 3(比如 t 分布)，则其密度函数的最高处(山峰)比正态分布更“尖”，而两侧尾部则更“厚”，称为“厚尾”(fat tails)。存在厚尾的概率分布更容易在尾部取值，称为**极端值(outlier)**，参见图 3.13。

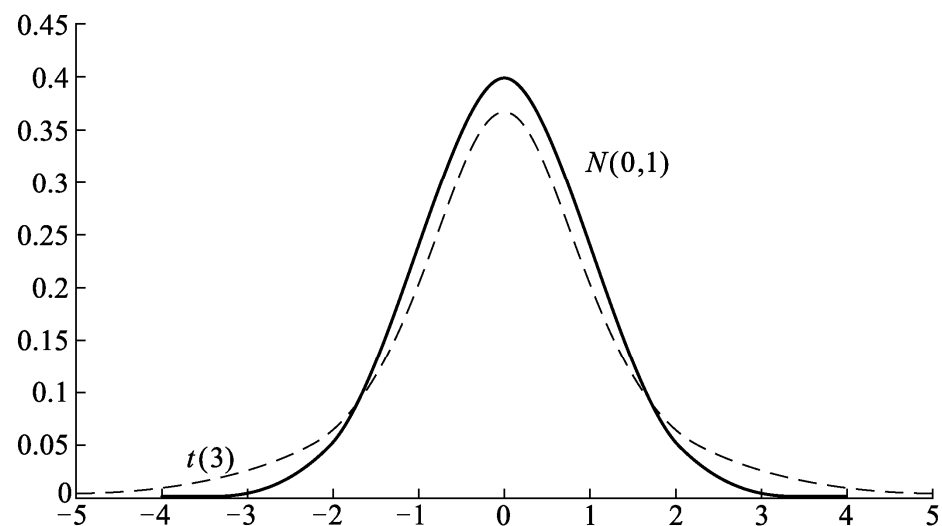


图 3.13 $N(0,1)$ 与 $t(3)$ 的概率密度

定义 随机变量 X 的超额峰度 (excess kurtosis) 为 $E[(X - \mu)/\sigma]^4 - 3$ 。

由于正态分布的偏度为 0，峰度为 3，故可使用正态分布的偏度与峰度性质来检验某个分布是否为正态分布。

更一般地，对于随机变量 X 与任意函数 $g(\cdot)$ ，称随机变量函数 $g(X)$ 的期望 $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$ 为矩(moment)。

定义 条件期望(conditional expectation)就是条件分布 $Y|x$ 的期望, 即

$$E(Y | X = x) \equiv E(Y | x) = \int_{-\infty}^{+\infty} yf(y|x)dy \quad (3.53)$$

由于 y 已被积分积掉, 故 $E(Y|x)$ 只是 x 的函数, 参见图 3.14。

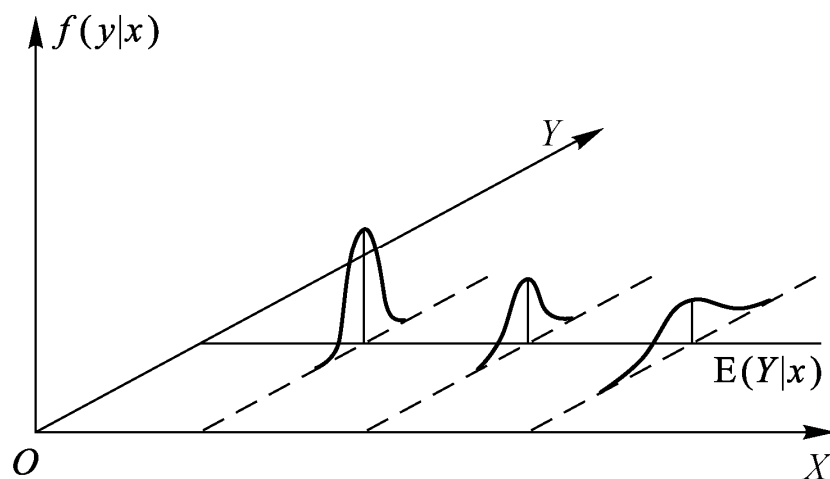


图 3.14 条件期望与条件方差示意图

定义 条件方差(conditional variance)就是条件分布 $Y | x$ 的方差

$$\text{Var}(Y | X = x) \equiv \text{Var}(Y | x) = \int_{-\infty}^{+\infty} [y - E(Y | x)]^2 f(y | x) dy \quad (3.54)$$

由于 y 已被积分积掉, 故 $\text{Var}(Y | x)$ 也只是 x 的函数, 参见图 3.14。

定义 设 $\mathbf{X} = (X_1 \ X_2 \ \cdots \ X_n)'$ 为 n 维随机向量, 则其协方差矩阵(covariance matrix)为 $n \times n$ 的对称矩阵:

$$\begin{aligned}
\text{Var}(\mathbf{X}) &\equiv \text{E} \left[(\mathbf{X} - \text{E}(\mathbf{X}))(\mathbf{X} - \text{E}(\mathbf{X}))' \right] \\
&= \text{E} \left[\begin{pmatrix} X_1 - \text{E}(X_1) \\ \vdots \\ X_n - \text{E}(X_n) \end{pmatrix} \begin{pmatrix} X_1 - \text{E}(X_1) & \cdots & X_n - \text{E}(X_n) \end{pmatrix} \right] \\
&= \text{E} \left(\begin{pmatrix} [X_1 - \text{E}(X_1)]^2 & \cdots & [X_1 - \text{E}(X_1)][X_n - \text{E}(X_n)] \\ \vdots & \ddots & \vdots \\ [X_1 - \text{E}(X_1)][X_n - \text{E}(X_n)] & \cdots & [X_n - \text{E}(X_n)]^2 \end{pmatrix} \right) \\
&= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}
\end{aligned}
\tag{3.55}$$

主对角线元素 $\sigma_{ii} \equiv \text{Var}(X_i)$ ，非主对角线元素 $\sigma_{ij} \equiv \text{Cov}(X_i, X_j)$ 。

协方差矩阵必然为半正定矩阵(positive semidefinite)。在一维情况下，这意味着随机变量的方差必然为非负。

对于随机向量 \mathbf{X} 的期望与协方差矩阵的运算，有如下法则。假设 \mathbf{A} 为 $m \times n$ 常数矩阵(不含随机变量)，可以证明：

$$(1) E(\mathbf{A}\mathbf{X}) = \mathbf{A} E(\mathbf{X}) \quad (\text{期望算子的线性性})$$

$$(2) \text{Var}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})[E(\mathbf{X})]'$$
 (一维公式的推广)

$$(3) \text{Var}(\mathbf{A}\mathbf{X}) = \mathbf{A} \text{Var}(\mathbf{X}) \mathbf{A}' \quad (\text{夹心估计量})$$

如果 \mathbf{A} 为对称矩阵，则 $\text{Var}(\mathbf{A}\mathbf{X}) = \mathbf{A} \text{Var}(\mathbf{X}) \mathbf{A}$ ，称为**夹心估计量** (sandwich estimator)，其中两边的 \mathbf{A} 为“面包”，而夹在中间的 $\text{Var}(\mathbf{X})$ 为“菜”，在形式上类似于三明治。

以上随机变量的数字特征都可视为“总体矩” (population moments)。

在抽取随机样本后，可用样本数据计算相应的“样本矩” (sample moments)，作为相应总体矩的估计值。

即以“求样本平均值运算” $\left(\frac{1}{n} \sum_{i=1}^n (\cdot) \right)$ 来替代总体矩表达式的期望算子 $E(\cdot)$ 。

比如, 可用样本均值(sample mean) $\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$ 来估计总体均值(population mean)或期望 $E(X)$ 。

以数据集 `grilic.dta` 为例。

该数据集截取自 Griliches(1976)对教育投资回报率的经典研究, 由 Blackburn and Neumark(1992)更新数据, 包括 758 名美国年轻男子的数据。

首先，打开此数据集，看它所包含的变量。

```
. use grilic.dta,clear  
. describe
```

Contains data from D:\desktop\±¼¿Æ¼Æ¿\2014\¼Æ¿¼-¼ÃÑ§¼°StataÓ ÓÃ\Data\grilic.dta				
obs:	758			
vars:	11		15 Sep 2014 07:21	
size:	13,644			
variable name	storage type	display format	value label	variable label
rns	byte	%8.0g		south = 1
mrt	byte	%8.0g		married = 1
smsa	byte	%8.0g		big cities =1
med	byte	%8.0g		mother's education
iq	int	%8.0g		IQ
kww	byte	%8.0g		KWW
age	byte	%8.0g		age
s	byte	%8.0g		schooling
expr	float	%9.0g		experience
tenure	byte	%8.0g		tenure
lnw	float	%9.0g		ln(wage)
Sorted by:				

此数据集的样本容量为 758。

被解释变量为 `lnw` (工资对数)。

解释变量：`s` (教育年限)、`expr` (工龄)、`tenure` (在现单位工作年限)、`age` (年龄)、`iq` (智商)、`kww` (在 KWW(Knowledge of the World of Work)测试的成绩)、`med` (母亲的教育年限)、`mrt` (婚否)、`smsa` (是否住在大城市)以及 `rns` (是否住在美国南方)。

看各变量的基本统计指标。

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rns	758	.2691293	.4438001	0	1
mrt	758	.5145119	.5001194	0	1
smsa	758	.7044855	.456575	0	1
med	758	10.91029	2.74112	0	18
iq	758	103.8562	13.61867	54	145
kww	758	36.57388	7.302247	12	56
age	758	21.83509	2.981756	16	30
s	758	13.40501	2.231828	9	18
expr	758	1.735429	2.105542	0	11.444
tenure	758	1.831135	1.67363	0	10
lnw	758	5.686739	.4289494	4.605	7.051

如想看 lnw 的更多统计指标，比如偏度、峰度，可加选择项“detail”：

```
. sum lnw,detail
```

ln(wage)				
	Percentiles	Smallest		
1%	4.804	4.605		
5%	5.011	4.605		
10%	5.165	4.654	Obs	758
25%	5.38	4.718	Sum of Wgt.	758
50%	5.684		Mean	5.686739
		Largest	Std. Dev.	.4289494
75%	5.991	6.786		
90%	6.252	6.844	Variance	.1839976
95%	6.399	6.869	Skewness	.1744968
99%	6.706	7.051	Kurtosis	2.73237

通过画 lnw 的直方图来看其(无条件)分布，结果参见图 3.15。

```
. hist lnw,width(0.1)
(bin=25, start=4.605, width=.1)
```

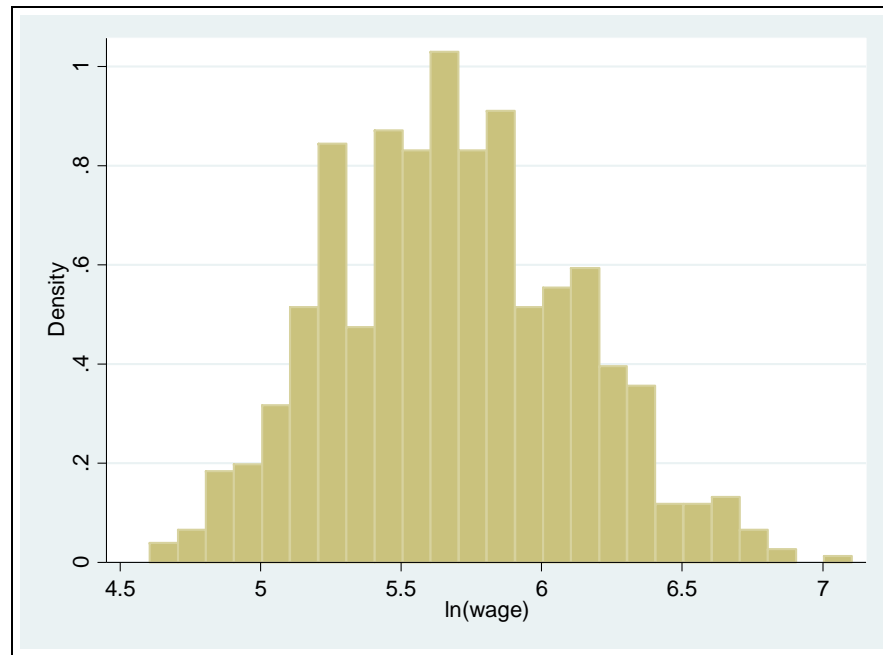


图 3.15 工资对数的直方图

但直方图不连续。如想得到概率密度函数的连续估计，可输入命令(参见图 3.16)

```
. kdensity lnw,normal normop(lpattern(dash))
```

“kdensity”表示核密度估计(kernel density estimation)。

选择项“normal”表示画正态分布的密度函数作为对比。

选择项“normop(lpattern(dash))”则指示将正态密度用虚线(dash)来画(其中，normop表示 normal options；而 lpattern 表示 line pattern)。

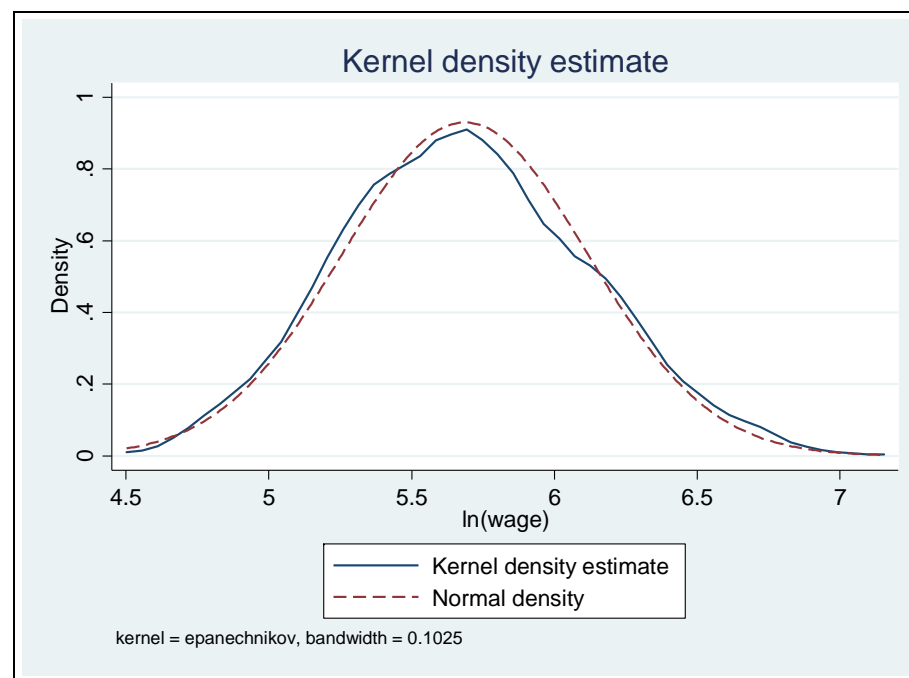


图 3.16 工资对数的核密度估计

工资对数的分布接近于正态分布，也基本对称。

作为对比，考察工资水平本身的分布，参见图 3.17。

```
. gen wage=exp(lnw)  
. kdensity wage
```

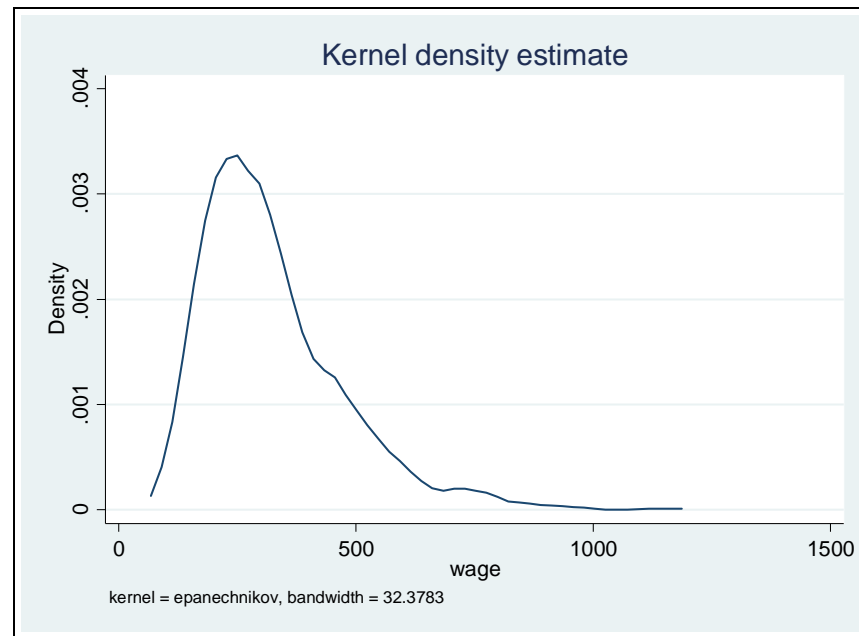


图 3.17 工资的核密度估计

工资水平的分布相去正态分布甚远，为非对称分布，在右边存在很长的尾巴，称为“向右偏”。

而工资对数的分布则很接近正态，这是使用工资对数作为被解释变量的原因之一。

对于取值为正的非对称分布，有时可通过取对数使其变得更加对称，也更接近于正态分布。

以上考察的均为无条件分布以及无条件期望等。

下面考察给定教育年限情况下的条件分布。

比如，给定教育年限为 16 年(大学毕业)，工资对数的条件密度(参见图 3.18)。

```
. kdensity lnw if s==16
```

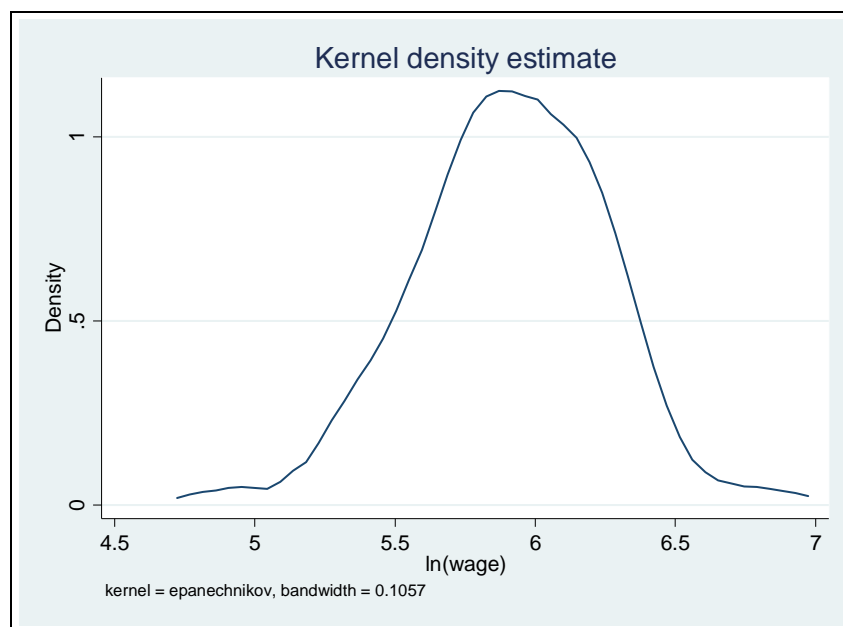


图 3.18 给定 $s=16$ 的工资对数条件密度

将 $\ln w$ 的无条件密度与条件密度画在一起(参见图 3.19):

```
. twoway kdensity lnw || kdensity lnw if  
s==16,lpattern(dash)
```

“||”为分隔符(separator)。

分隔符“||”的作用，也可以通过两个括号“() ()”来等价地实现，比如：

```
. twoway (kdensity lnw) (kdensity lnw if  
s==16,lpattern(dash))
```

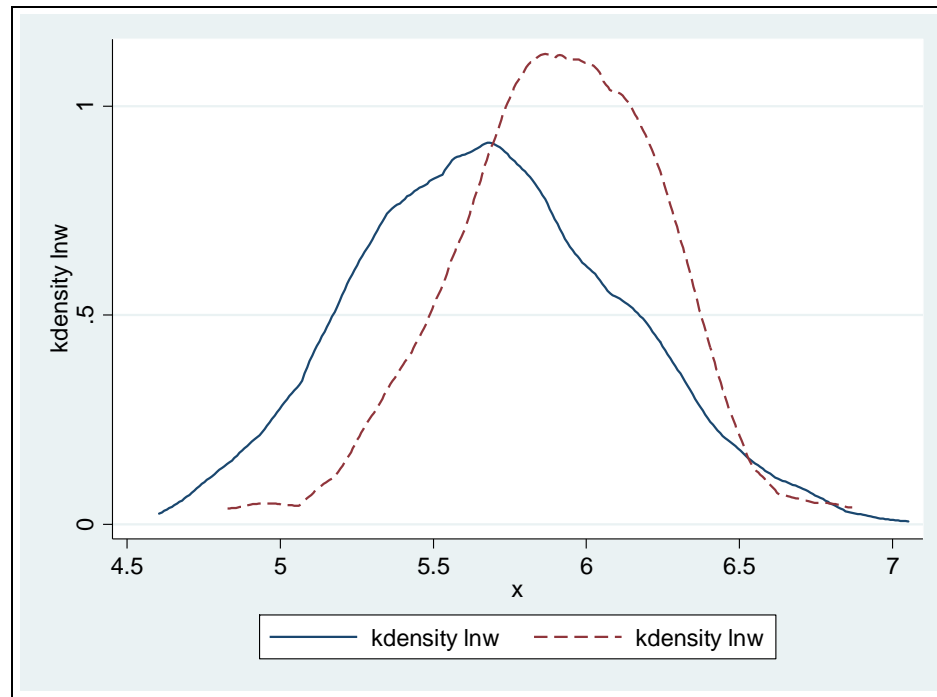


图 3.19 给定 $s=18$ 的工资对数条件密度

给定 $s=16$ 的工资对数条件密度(虚线), 明显比工资对数的无条件密度向右移, 故条件期望增大, 而条件方差似乎也变小。

比较 `lnw` 的无条件期望、方差与条件期望、条件方差。

```
. sum lnw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnw	758	5.686739	.4289494	4.605	7.051

```
. sum lnw if s==16
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnw	151	5.907338	.3396442	4.828	6.869

条件期望为 5.91，大于无条件期望 5.69。

条件标准差为 0.34，小于无条件标准差 0.43。

比较在 $s=12$ (中学毕业)与 $s=16$ (大学毕业)情况下, $\ln w$ 的条件密度(参见图 3.20)。

```
. twoway (kdensity lnw if s==12) (kdensity lnw  
if s==16,lpattern(dash))
```

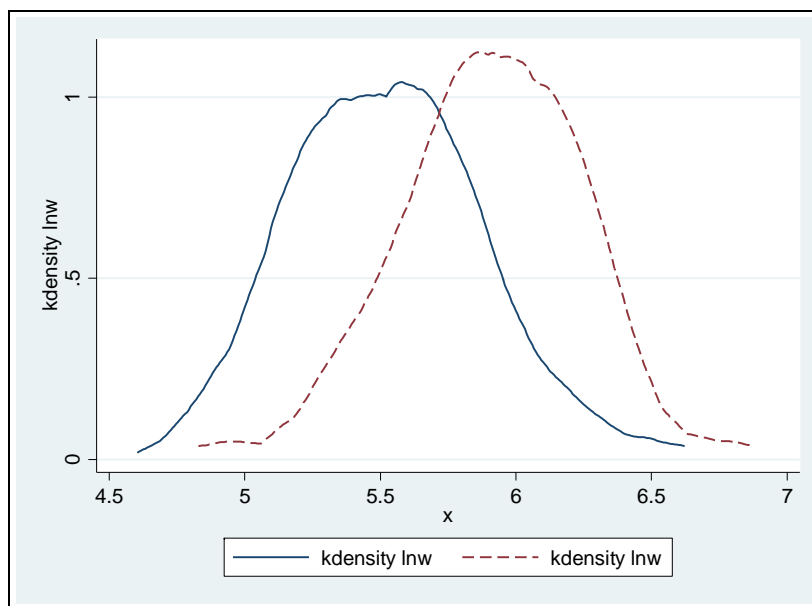


图 3.20 给定 $s=12$ (实线)与 $s=16$ (虚线)的工资对数条件密度

大学毕业生($s=16$)的工资对数条件分布相对于中学毕业生($s=12$)向右移, 故大学毕业生工资对数的条件期望高于中学毕业生。

3.6 迭代期望定律

定理 对于条件期望的运算, 有以下重要的迭代期望定律(Law of iterated expectation),

$$E(Y) = E_X [E(Y | x)] \quad (3.56)$$

无条件期望 $E(Y)$ 等于, 给定 $X = x$ 情况下 Y 的条件期望 $E(Y | x)$ (仍为 x 的函数), 再对 X 求期望。

如果 X 为离散随机变量，则根据期望定义，上式可写为：

$$E(Y) = \sum_i P(X = x_i) E(Y | x_i) \quad (3.57)$$

无条件期望等于条件期望之加权平均，而权重为条件“ $X = x$ ”的概率(取值可能性)。

以数据集 `grilic.dta` 为例，来验证迭代期望定律，即

$$E(\ln w) = E_{rns} [E(\ln w | rns)] \quad (3.58)$$

其中，`rns` 为美国南方居民的虚拟变量，取值为 0 或 1。

首先，计算 $rns = 0$ 情况下， $\ln w$ 的条件期望：

```
. use grilic.dta, clear  
. sum lnw if rns==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnw	554	5.725644	.4129207	4.605	7.051

北方居民有 554 位，其条件期望 $E(\ln w | rns = 0) = 5.725644$ 。

其次，计算 $rns = 1$ 情况下， $\ln w$ 的条件期望：

```
. sum lnw if rns==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnw	204	5.581083	.4542189	4.718	6.844

南方居民有 204 位，其条件期望 $E(\ln w | rns = 1) = 5.581083$ ，故美国南方居民的工资略低于北方。

以北方与南方居民所占比重作为权重，将北方与南方居民的平均工资对数进行加权平均。

```
. dis 5.725644*(554/(554+204))+5.581083*(204/(554+204))  
5.6867384
```

最后，用命令 `summarize` 直接计算无条件期望。

```
. sum lnw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnw	758	5.686739	.4289494	4.605	7.051

二者的结果完全相等，验证等式(3.58)成立。

以离散型变量为例，证明等式(3.57)。假设 X 的可能取值为 $x_1, x_2, \dots, x_i, \dots$ ，而 Y 的可能取值为 $y_1, y_2, \dots, y_j, \dots$ 。记 $p_i \equiv P(X = x_i)$ ， $q_j \equiv P(Y = y_j)$ ，而 $p_{ij} \equiv P(X = x_i, Y = y_j)$ 。

证明：从等式(3.57)的右边开始证明。

$$E_X[E(Y | x)] = \sum_i P(X = x_i) E(Y | x_i) \quad (\text{期望的定义式})$$

$$= \sum_i P(X = x_i) \left[\sum_j P(Y = y_j | x_i) \cdot y_j \right] \quad (\text{条件期望的定义式})$$

$$= \sum_i \cancel{P(X = x_i)} \left[\sum_j \frac{P(X = x_i, Y = y_j)}{\cancel{P(X = x_i)}} \cdot y_j \right] \quad (\text{条件概率的定义式})$$

$$= \sum_i \left[\sum_j P(X = x_i, Y = y_j) \cdot y_j \right] \quad (\text{消去 } P(X = x_i))$$

$$= \sum_j \left[\sum_i P(X = x_i, Y = y_j) \cdot y_j \right] \quad (\text{交换加总的次序})$$

$$= \sum_j \left[y_j \sum_i P(X = x_i, Y = y_j) \right] \quad (y_j \text{ 与 } i \text{ 无关, 可提出})$$

$$= \sum_j y_j P(Y = y_j) \quad (\text{边缘概率与联合概率的关系})$$

$$= E(Y) \quad (\text{期望的定义式})$$

迭代期望定律很像全概率公式：无条件期望等于条件期望之加权平均，权重为条件概率密度。

将迭代期望定律(3.56)推而广之，对于任意函数 $g(\cdot)$ ，可得

$$E[g(Y)] = E_x E[g(Y)|x] \quad (3.59)$$

有时期望算子 E_x 的下标被省去，需注意对什么变量求期望。

3.7 随机变量无关的三个层次概念

定义 对于连续型随机变量 X 与 Y ，如果其联合密度等于边缘密度的乘积，即 $f(x, y) = f_x(x)f_y(y)$ ，则称 X 与 Y 相互独立 (independent)。

如果 X 与 Y 相互独立, 则 X 与 Y 没有任何关系, 故 X 的取值不对 Y 的取值产生任何影响, 反之亦然。

“相互独立”是有关随机变量“无关”的最强概念。

线性不相关的概念则更弱, 仅要求协方差为 0, 即 $\text{Cov}(X, Y) = 0$ 。

“相互独立”意味着“线性不相关”, 但反之不然。

在二者之间还有一个中间层次的无关概念, 即“均值独立”(mean-independence), 在计量中很有用。

定义 假设条件期望 $E(Y | x)$ 存在。如果 $E(Y | x)$ 不依赖于 X , 则称 Y 均值独立于 X (Y is mean-independent of X)。

均值独立不是一种对称的关系，即“ Y 均值独立于 X ”并不意味着“ X 均值独立于 Y ”。

命题 Y 均值独立于 X , 当且仅当 $E(Y | x) = E(Y)$ (条件期望等于无条件期望)。

证明: (1) 假设 Y 均值独立于 X , 则 $E(Y | x)$ 不依赖于 X , 故 $E_X[E(Y | x)] = E(Y | x)$ 。根据迭代期望定律, $E(Y) = E_X[E(Y | x)] = E(Y | x)$ 。

(2) 假设 $E(Y | x) = E(Y)$, 则显然 $E(Y | x)$ 不依赖于 X , 故 Y 均值独立于 X 。

命题 如果 X 与 Y 相互独立, 则 Y 均值独立于 X , 且 X 均值独立于 Y 。

定理 如果 Y 均值独立于 X , 或 X 均值独立于 Y , 则 $\text{Cov}(X, Y) = 0$ 。

证明: $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$ (协方差的定义)

$$= E_X E_Y [(X - E(X))(Y - E(Y)) | x] \quad (\text{迭代期望定律})$$

$$= E_X [(X - E(X)) E_Y (Y - E(Y) | x)] \quad ((X - E(X)) \text{ 视为常数提出})$$

$$= E_X [(X - E(X))(E(Y | x) - E(Y))] \quad (\text{期望算子的线性性})$$

$$= E_X [(X - E(X)) \cdot 0] = 0 \quad (\text{均值独立的性质})$$

“相互独立” \Rightarrow “均值独立” \Rightarrow “线性不相关”; 反之不然。

3.8 常用连续型统计分布

1. 正态分布：如果随机变量 X 的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad (3.60)$$

则称 X 服从正态分布(normal distribution), 记为 $X \sim N(\mu, \sigma^2)$, 其中 μ 为期望, 而 σ^2 为方差。

将 X 进行标准化, 定义 $Z \equiv \frac{X-\mu}{\sigma}$, 则 Z 服从标准正态分布 (standard normal distribution), 记为 $Z \sim N(0,1)$, 其概率密度函数为

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} \quad (3.61)$$

标准正态分布的概率密度以原点为对称，呈钟形(参见图 3.11)，通常记为 $\phi(x)$ ；其累积分布函数记为 $\Phi(x)$ 。

在 Stata 中，使用函数 `normalden(x)` 与 `normal(x)` 分别表示标准正态的密度函数 $\phi(x)$ 与累积分布函数 $\Phi(x)$ 。

比如，计算标准正态变量小于 1.96 的概率：

```
. dis normal(1.96)  
.9750021
```

如画标准正态的密度函数，可输入命令(参见图 3.21)：

```
. twoway function y=normalden(x),range(-5 5)  
xline(0) ytitle(概率密度)
```

选择项“`range(-5 5)`”表示在横轴区间 $(-5, 5)$ 上画此图；默认为“`range(0 1)`”，即在 $(0, 1)$ 区间画图。“`xline(0)`”表示在横轴 $x = 0$ 处画一条直线。“`yttitle(概率密度)`”表示将纵轴标签设为“概率密度”。

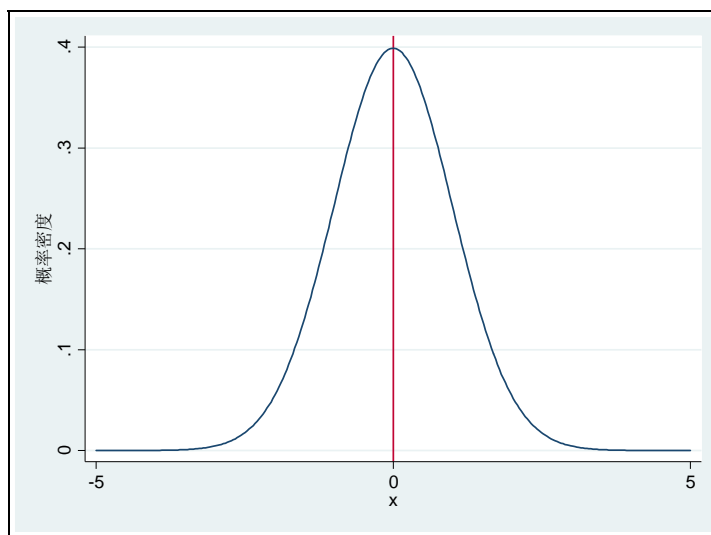


图 3.21 标准正态的概率密度

正态分布 $N(m, s^2)$ 的密度函数可用 `normalden(x, m, s)` 来表示，其中 m 与 s 分别为期望与标准差。

将 $N(0,1)$ 与 $N(1,4)$ 的密度函数画在一起(参见图 3.22):

```
. twoway function y=normalden(x), range(-5 10) ||  
function      z=normalden(x,1,2), range(-5      10)  
lpattern(dash) ytitle(概率密度)
```

选择项 “`lpattern(dash)`” 表示使用虚线画图。

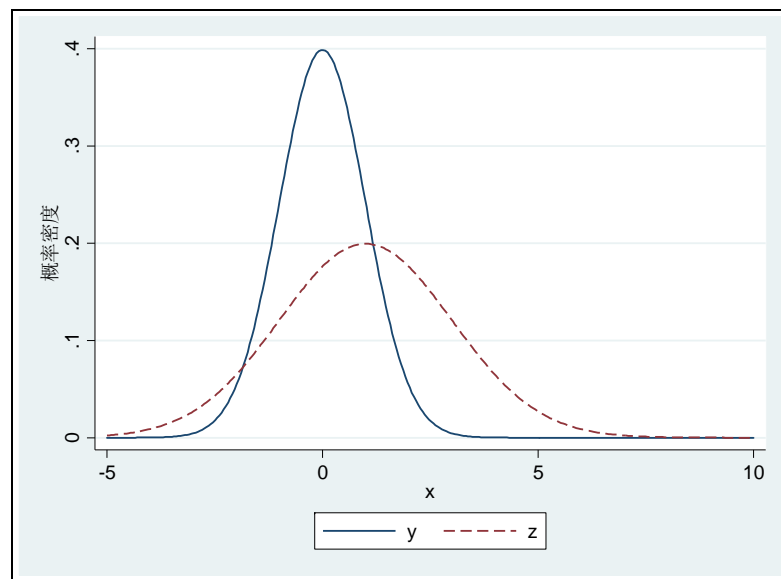


图 3.22 $N(0,1)$ 与 $N(1,4)$ 的密度函数

多维正态分布：如果 n 维随机向量 $\mathbf{X} = (X_1 \ X_2 \ \cdots \ X_n)'$ 的联合密度函数为

$$f(x_1, \cdots, x_n) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\} \quad (3.62)$$

则称 \mathbf{X} 服从期望为 $\boldsymbol{\mu}$ 、协方差矩阵为 $\boldsymbol{\Sigma}$ 的 n 维正态分布，记为 $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 。

其中， $(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$ 为 $(\mathbf{X} - \boldsymbol{\mu})$ 的二次型，其二次型矩阵为协方差矩阵的逆矩阵 $\boldsymbol{\Sigma}^{-1}$ ； $|\boldsymbol{\Sigma}|$ 为协方差矩阵 $\boldsymbol{\Sigma}$ 的行列式。

多维正态分布具有良好的性质。比如，多维正态的每个分量都是正态，其分量之任意线性组合仍然是正态。反之，每个分量均为一维正态并不足以保证其联合分布也是多维正态的。

如果 (X_1, X_2, \dots, X_n) 服从 n 维正态分布，则“ X_1, X_2, \dots, X_n 相互独立”与“ X_1, X_2, \dots, X_n 两两不相关”是等价的。利用此性质，有时容易证明正态变量的独立性。

2. χ^2 分布(卡方分布, **Chi-square**)

如果 $Z \sim N(0,1)$, 则 $Z^2 \sim \chi^2(1)$, 即自由度为 1 的 χ^2 分布。

如果 $\{Z_1, \dots, Z_k\}$ 为独立同分布的标准正态, 则其平方和服从自由度为 k 的卡方分布, 记为

$$\sum_{i=1}^k Z_i^2 \sim \chi^2(k) \quad (3.63)$$

参数 k 为自由度(degree of freedom), 因为 $\sum_{i=1}^k Z_i^2$ 由 k 个相互独立(自由)的随机变量所构成。

χ^2 分布来自标准正态的平方和, 故取值为正。可以证明, $\chi^2(k)$ 分布的期望为 k , 而方差为 $2k$ 。

在 Stata 中，使用函数 `chi2den(k,x)` 与 `chi2(k,x)` 分别表示自由度为 k 的卡方分布的概率密度与累积分布函数。比如，输入命令将 $\chi^2(3)$ 与 $\chi^2(5)$ 的密度函数画在一起(参见图 3.23)：

```
. twoway function chi3=chi2den(3,x),range(0 20) || function chi5=chi2den(5,x),range(0 20)  
    lpattern(dash) ytitle(概率密度)
```

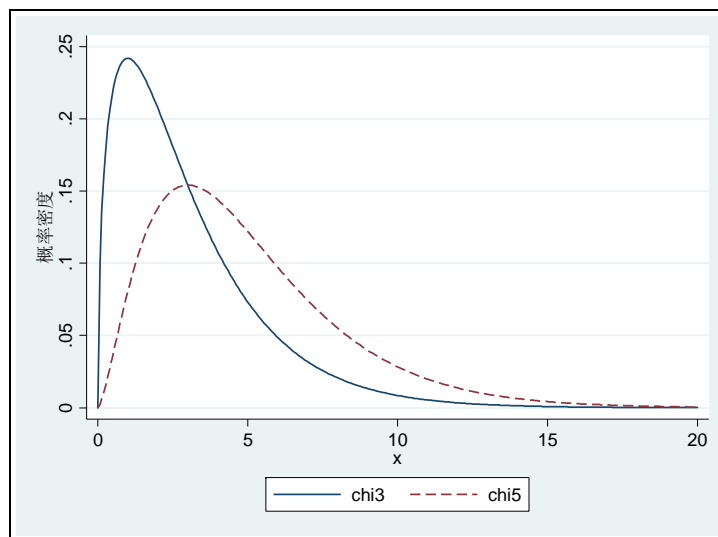


图 3.23 $\chi^2(3)$ 与 $\chi^2(5)$ 的概率密度

3. t 分布

假设 $Z \sim N(0,1)$, $Y \sim \chi^2(k)$, 且 Z 与 Y 相互独立, 则 $\frac{Z}{\sqrt{Y/k}}$ 服从自由度为 k 的 t 分布, 记为

$$\frac{Z}{\sqrt{Y/k}} \sim t(k) \quad (3.64)$$

k 为自由度。如果表达式(3.64)中的分子与分母不相互独立, 则一般不服从 t 分布。

t 分布也以原点为对称, 但与标准正态分布相比, 中间的“山峰”更低(但更尖), 而两侧有“厚尾”(fat tails), 参见图 3.13。

当自由度 $k \rightarrow \infty$ 时, t 分布收敛于标准正态分布。

在 Stata 中, 使用函数 `tden(k,t)` 与 `t(k,t)` 分别表示自由度为 k 的 t 分布的概率密度与累积分布函数。

比如, 使用命令将 $t(1)$ 与 $t(5)$ 的密度函数画在一起(参见图 3.24):

```
. twoway function t1=tden(1,x),range(-5 5) ||  
function t5=tden(5,x),range(-5 5) lpattern(dash)  
ytitle(概率密度)
```

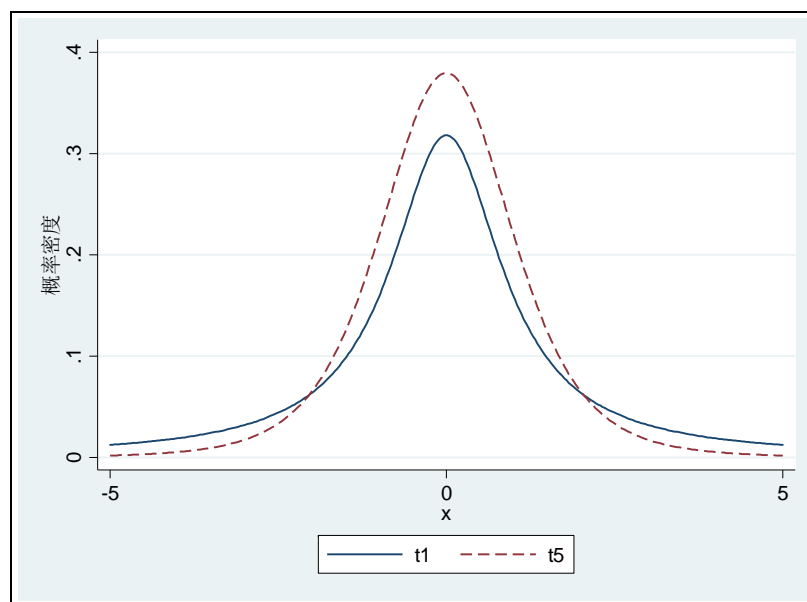


图 3.24 $t(1)$ 与 $t(5)$ 的密度函数

Stata 以函数 `ttail(k, t)` 表示自由度为 k 的 t 分布的右侧尾部概率，即 $P(T > t)$ ，正好是反向的累积分布函数。

4. F 分布

假设 $Y_1 \sim \chi^2(k_1)$, $Y_2 \sim \chi^2(k_2)$, 且 Y_1, Y_2 相互独立, 则 $\frac{Y_1/k_1}{Y_2/k_2}$ 服从自由度为 k_1, k_2 的 F 分布, 记为

$$\frac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2) \quad (3.65)$$

其中, k_1, k_2 为自由度。

F 分布的取值也只能为正数, 其概率密度形状与 χ^2 分布相似。

如表达式(3.65)中的分子与分母不独立, 则一般不服从 F 分布。

Stata 使用函数 $\text{Fden}(k_1, k_2, x)$ 与 $\text{F}(k_1, k_2, x)$ 分别表示自由度为 (k_1, k_2) 的 F 分布的概率密度与累积分布函数。比如，输入命令将 $F(10, 20)$ 与 $F(10, 5)$ 的密度函数画在一起(参见图 3.25):

```
. twoway function F20=Fden(10,20,x),range(0 5)
|| function F5=Fden(10,5,x),range(0 5)
lpattern(dash) ytitle(概率密度)
```

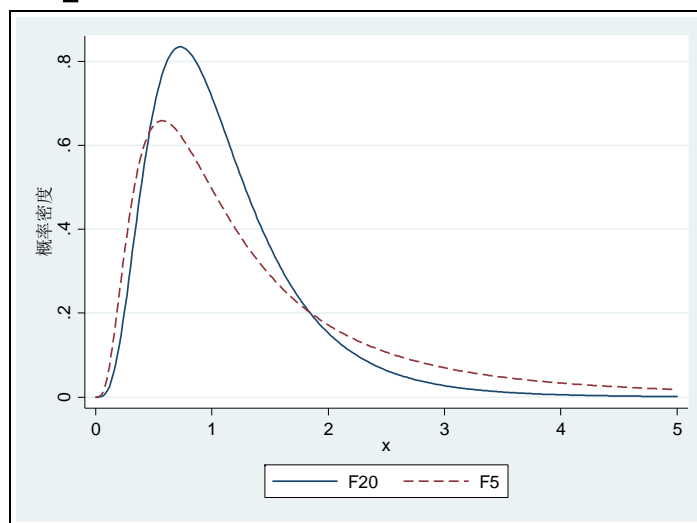


图 3.25 $F(10, 20)$ 与 $F(10, 5)$ 的概率密度

如想对图 3.25 作进一步的编辑,比如将变量标签改为“ $F(10,20)$ ”与“ $F(10,5)$ ”,可在图像上点菜单“File”→“Start Graph Editor”,启动 Stata 的图像编辑器,参见图 3.26。

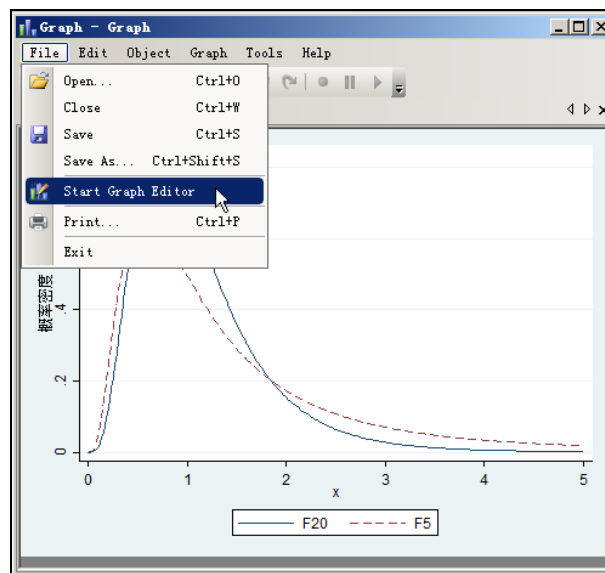


图 3.26 启动 Stata 的图像编辑器

直接点击原来的变量标签“F20”与“F5”即可进行编辑，将标签分别改为“F(10, 20)”与“F(10, 5)”，参见图 3.27。

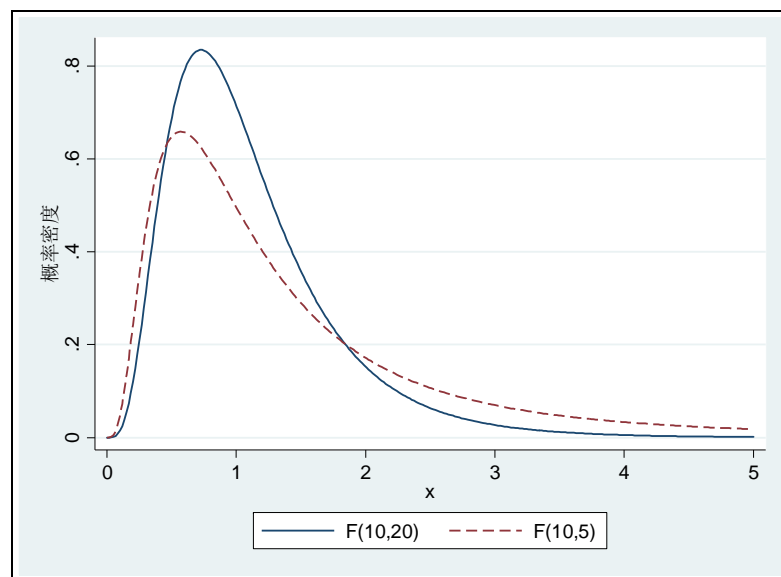


图 3.27 编辑变量标签的结果

F 分布与 t 分布存在着密切关系，因为 t 分布的平方就是 F 分布。

命题 如果 $X \sim t(k)$, 则 $X^2 \sim F(1, k)$ 。

证明 由于 $X \sim t(k)$, 故根据 t 分布的定义, 可将 X 写为 $X = \frac{Z}{\sqrt{Y/k}} \sim t(k)$, 其中 $Z \sim N(0, 1)$, $Y \sim \chi^2(k)$, 且 Z 与 Y 相互独立。因此,

$$X^2 = \left(\frac{Z}{\sqrt{Y/k}} \right)^2 = \frac{Z^2/1}{Y/k} \sim F(1, k) \quad (3.66)$$

其中, 由于 $Z \sim N(0, 1)$, 故 $Z^2 \sim \chi^2(1)$; 而且, 由于 Z 与 Y 相互独立, 故 Z^2 也与 Y 相互独立。根据 F 分布的定义, X^2 服从自由度为 $(1, k)$ 的 F 分布。

更多有关概率分布的 Stata 函数, 参见 “help density function”。

3.9 统计推断的思想

计量经济学的主要方法是数理统计的统计推断(statistical inference)。

称我们感兴趣的研究对象全体为总体(population)，其中的每个研究对象称为个体(individual)。

由于总体包含的个体可能很多，普查成本较高，故常从总体抽取部分个体，称为样本(sample)，参见图 3.28。

样本所包含的个体数目称为样本容量(sample size)。

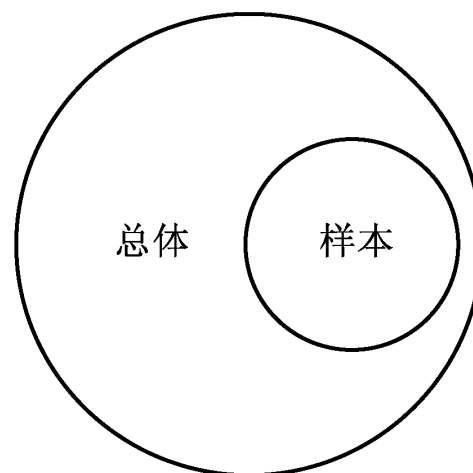


图 3.28 总体与样本

通常希望样本为**随机样本**(random sample)，即总体中的每位个体都有相同的概率被抽中，且被抽中的概率相互独立，称为**独立同分布**(independently identically distributed，简记 iid)。

由于样本来自总体，必然带有总体的信息。

统计推断就是根据样本数据，对总体性质进行推断的科学。

统计推断的主要形式有参数估计(点估计、区间估计)、假设检验及预测等，其中点估计为统计推断的基础。

假设随机变量 X 的概率密度为 $f(x; \theta)$ ，其中 θ 为待估参数。

为估计总体参数 θ ，从总体中抽取了样本容量为 n 的样本数据 $\{x_1, x_2, \dots, x_n\}$ 。

我们希望根据此样本数据来设计一个性质良好的统计量 $\hat{\theta}(x_1, x_2, \dots, x_n)$ ，以此估计 θ 。

统计量 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为样本数据 $\{x_1, x_2, \dots, x_n\}$ 的函数，故仍为随机变量，且随着样本不同而变化。

由于使用 $\hat{\theta}$ (英文读为“theta hat”)来估计 θ ，故称 $\hat{\theta}$ 为 θ 的估计量(estimator)。

相应地，给定 $\{x_1, x_2, \dots, x_n\}$ 后，可得到估计量 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 的具体取值，称为估计值(estimate)。

比如， θ 为总体均值，即 $E(X) = \theta$ ，则一般使用样本均值来估计 θ ，即估计量 $\hat{\theta} = \bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ 。

θ 的其他可能估计量包括第一个观测值 x_1 , 中位数 $median(x_1, x_2, \dots, x_n)$, 最大值 $max(x_1, x_2, \dots, x_n)$, 最小值 $min(x_1, x_2, \dots, x_n)$ 等。

由于潜在的估计量很多(所有样本数据的函数都可视为估计量), 需要有评判估计量优劣的标准。

首先, 希望估计量没有系统性偏差(systematic error), 即 $\hat{\theta}$ 不会系统地高估或低估 θ 。

定义 以估计量 $\hat{\theta}$ 来估计参数 θ , 则其偏差为 $Bias(\hat{\theta}) \equiv E(\hat{\theta}) - \theta$ 。

定义 如果偏差 $Bias(\hat{\theta}) = 0$, 则称 $\hat{\theta}$ 为无偏估计量 (unbiased estimator); 反之, 则称为有偏估计 (biased estimator), 参见图 3.29。

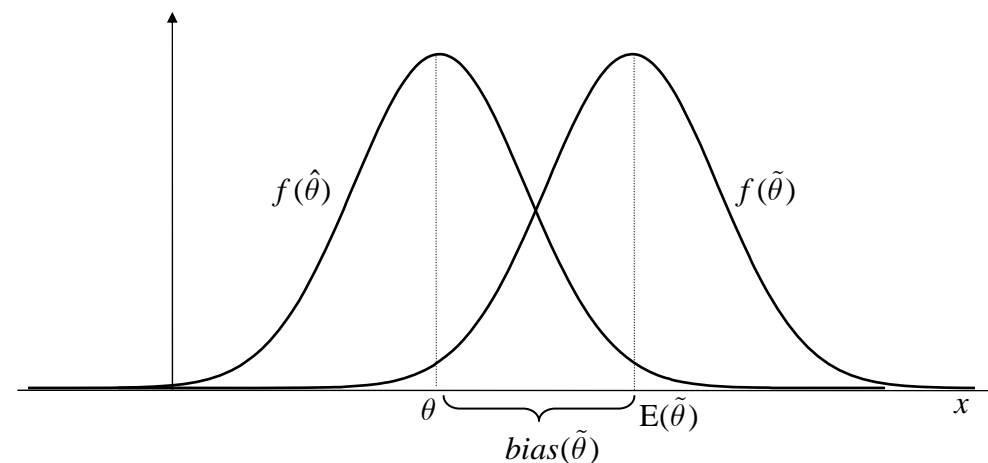


图 3.29 无偏估计量 $\hat{\theta}$ 与有偏估计量 $\tilde{\theta}$ 的概率分布

其次，希望抽样误差(sampling error) $(\hat{\theta} - \theta)$ 尽量地小，即 $\hat{\theta}$ 离真实参数 θ 越近越好。

由于 $(\hat{\theta} - \theta)$ 可正可负，故考虑误差平方(squared error) $(\hat{\theta} - \theta)^2$ 。

但 $\hat{\theta}$ 是随机变量，故引入“均方误差”的概念。

定义 以估计量 $\hat{\theta}$ 来估计参数 θ ，则其均方误差(Mean Squared Error, 简记 MSE)为 $\text{MSE}(\hat{\theta}) \equiv E[(\hat{\theta} - \theta)^2]$ 。

最优的估计量应在所有估计量中具有最小的均方误差。

估计量 $\hat{\theta}$ 的均方误差来源于 $\hat{\theta}$ 的方差与偏差。

命题 均方误差可以分解为方差与偏差平方之和，即

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2 \quad (3.67)$$

证明： $\text{MSE}(\hat{\theta}) \equiv E[(\hat{\theta} - \theta)^2] = E\left\{ \left[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta \right]^2 \right\}$

$$\begin{aligned}
&= E\left[\hat{\theta} - E(\hat{\theta})\right]^2 + 2E\left\{\left[\hat{\theta} - E(\hat{\theta})\right]\left[E(\hat{\theta}) - \theta\right]\right\} + E\left[E(\hat{\theta}) - \theta\right]^2 \\
&= \text{Var}(\hat{\theta}) + 2E\left\{\left[\hat{\theta} - E(\hat{\theta})\right]\left[E(\hat{\theta}) - \theta\right]\right\} + \left[\text{Bias}(\hat{\theta})\right]^2
\end{aligned}$$

只需证明上式的交叉项为 0 即可：

$$E\left\{\left[\hat{\theta} - E(\hat{\theta})\right]\left[E(\hat{\theta}) - \theta\right]\right\} = \left[E(\hat{\theta}) - \theta\right]E\left[\hat{\theta} - E(\hat{\theta})\right] = \left[E(\hat{\theta}) - \theta\right] \cdot 0 = 0$$

因此，均方误差最小化，可视为在“估计量方差”与“偏差”之间进行权衡(trade-off)。

比如，一个无偏估计量 $\hat{\theta}$ ，如果方差很大，可能不如一个有偏但方差很小的估计量 $\tilde{\theta}$ (英文读为 theta tilde)，参见图 3.30。

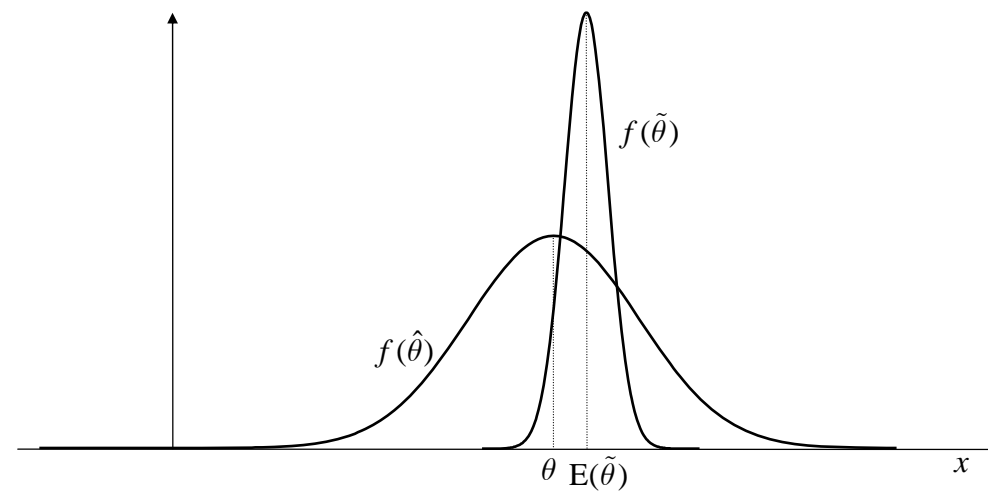


图 3.30 无偏估计量 $\hat{\theta}$ 与有偏估计量 $\tilde{\theta}$ 之间的权衡