1　基础知识

本章回顾一些相关数理统计的知识，如果读者学过这方面内容，可跳过，直接阅读第 2 章或感兴趣的章节。

1.1　数学基础

从严格意义上来说，所有自然科学理论都需要量化，都可归结为数学，但是，本节所指的数学是指最基础的数学理论，主要包括统计计算中常用的矩阵理论、Taylor 定理与极限理论。

1.1.1　数学记号与矩阵理论

为论述方便，本书不用黑体区分向量与矩阵，M 既可表示常数，也可表示向量、矩阵，以 M^{T} 或 M' 表示转置，以 I 表示单位矩阵，1或0也表示元素为 1 或 0 的向量，读者可根据上下文进行区分，也许初学者感觉有点混乱，但会慢慢发现这样表述的优点。

一个泛函就是一个函数空间中的实值函数，例如，如果 $T(f)=\int_{-\infty}^{+\infty} f(x)\mathrm{d}x$，则泛函 T 为可积函数到一维实数的映射。

示性函数 $I_{\{A\}}=1$，即如果 A 成立，等于 1，否则等于 0。一维实空间记为 $\mathbf{R}$，n 维实空间记为 $\mathbf{R}^n$。

如果对所有非零向量 x，$x^{\mathrm{T}}Ax>0$ 成立，则称对称方阵 A 正定。正定的等价条件是其所有特征根为正。如果对所有非零向量 x，$x^{\mathrm{T}}Ax\geqslant 0$，则称方阵 A 非负定或半正定。

函数 f 在点 x 的导数记作 $f'(x)$，如果 f 是 n 元实函数，则 f 在自变量 $x=(x_1,\cdots,x_n)^{\mathrm{T}}\in\mathbf{R}^n$ 处的梯度为 n 维列向量

$$\nabla f(x)=\left(\frac{\partial f(x)}{\partial x_1},\cdots,\frac{\partial f(x)}{\partial x_n}\right)^{\mathrm{T}}=\left(\frac{\mathrm{d}f(x)}{\mathrm{d}x_1},\cdots,\frac{\mathrm{d}f(x)}{\mathrm{d}x_n}\right)^{\mathrm{T}} \tag{1.1.1}$$

也称为函数 f 在 x 处的一阶导数。

函数 f 在 x 处的 **Hesse** 矩阵为 $n\times n$ 矩阵 $\nabla^2 f(x)$，第 i 行第 j 列元素为

$$(\nabla^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\mathrm{d}^2 f(x)}{\mathrm{d}x_i \mathrm{d}x_j}, \quad (1 \leqslant i, j \leqslant n) \tag{1.1.2}$$

也称为函数 f 在 x 处的二阶导数。

如果梯度 $\nabla f(x)$ 的每个分量在 x 处都连续，则称 f 在 x 一阶连续可微。如果 Hesse 矩阵 $\nabla^2 f(x)$ 的各个分量函数都连续，则称 f 在 x 二阶连续可微。如果 f 在开集 D 的每一点都连续可微，则称 f 在 D 上一阶连续可微。如果 f 在开集 D 的每一点都二阶连续可微，则称 f 在 D 上二阶连续可微。

若 f 在 x 二阶连续可微，则 $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$, $(i, j = 1, 2, \cdots, n)$，即 **Hesse** 矩阵 $\nabla^2 f(x)$ 是对称阵。负的 Hesse 矩阵在统计推断中具有重要的应用。

考虑向量值函数 $\boldsymbol{h}(x) = (h_1(x), \cdots, h_m(x))^{\mathrm{T}}$，其中每个分量 $h_i(x)$ 为 n 元实函数，假设对所有 i, j 偏导数 $\frac{\partial h_i(x)}{\partial x_j}$ 存在，则 $\boldsymbol{h}$ 在 x 的 **Jacobi** 矩阵为

$$\begin{pmatrix} \frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_1(x)}{\partial x_2} & \cdots & \frac{\partial h_1(x)}{\partial x_n} \\ \frac{\partial h_2(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_2} & \cdots & \frac{\partial h_2(x)}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial h_m(x)}{\partial x_1} & \frac{\partial h_m(x)}{\partial x_2} & \cdots & \frac{\partial h_m(x)}{\partial x_n} \end{pmatrix} \tag{1.1.3}$$

考虑到标量函数的梯度定义，有时也把向量函数 $h(x)$ 的 Jacobi 矩阵的转置称为 $\boldsymbol{h}$ 在 x 的梯度矩阵，记为 $\nabla h(x) = J_h(x)^{\mathrm{T}}$。

1.1.2 Taylor 定理

为了描述函数收敛的相对阶数，我们首先定义 O（同阶）和 o（高阶）。设 f, g 为定义在同一区间上函数，x_0 为此区间内或边界上一点，函数 $g(x) \neq 0$，其中在 x_0 的一个邻域内 $x \neq x_0$，如果存在一个常数 M 满足：当 $x \to x_0$ 时，$|f(x)| \leqslant M|g(x)|$，则称

$$f(x) = O(g(x)) \tag{1.1.4}$$

例如，当 $n \to \infty$ 时，$\frac{n+1}{n^2} = O\left(\frac{1}{n}\right)$。如果 $\lim\limits_{x \to x_0} \frac{f(x)}{g(x)} = 0$，则称

$$f(x) = o(g(x)) \tag{1.1.5}$$

Taylor 定理给出了函数 f 的一个多项式近似。设 f 在区间 (a,b) 上具有 $n+1$ 阶导数，在区间 $[a,b]$ 上有连续的 n 阶导数，则对于任意一个不同于 x 的 $x_0 \in [a,b]$，函数 f 在点 x_0 的 Taylor

级数展开为

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}f^{(k)}(x_0)(x-x_0)^k+R_n \tag{1.1.6}$$

其中，$f^{(k)}(x_0)$为函数 f 在点 x_0 的 k 阶导数，且

$$R_n=\frac{1}{(n+1)!}f^{(n+1)}(\xi)(x-x_0)^{n+1}\text{，}\quad \xi\in[x,x_0]$$

注意，当 $|x-x_0|\to 0$ 时，$R_n=O(|x-x_0|^{n+1})$。

多元 Taylor 定理与之类似，例如，多元函数 $f(x)$ 在 $\bar{x}$ 的二阶 Taylor 展开式为

$$f(x)=f(\bar{x})+\nabla f(\bar{x})^{\mathrm{T}}(x-\bar{x})+\frac{1}{2}(x-\bar{x})^{\mathrm{T}}\nabla^2 f(\bar{x})(x-\bar{x})+o(\|(x-\bar{x})\|^2)$$

其中，$o(\|(x-\bar{x})\|^2)$ 当 $\|(x-\bar{x})\|^2\to 0$ 时，关于 $\|(x-\bar{x})\|^2$ 是高阶无穷小。

Taylor 展开式在理论证明及算法设计中都发挥了很大的作用。

Euler-Maclaurin 公式在渐近分析中很有用。如果 f 在 $[0,1]$ 上具有 $2n$ 阶连续导数，则

$$\int_0^1 f(x)\mathrm{d}x=\frac{f(0)+f(1)}{2}-\sum_{i=0}^{n-1}\frac{b_{2i}(f^{(2i-1)}(1)-f^{(2i-1)}(0))}{(2i)!}-\frac{b_{2n}f^{(2n)}(\xi)}{(2n)!} \tag{1.1.7}$$

其中，$0\leqslant\xi\leqslant 1$，$b_i=B_i(0)$。由下列迭代关系确定

$$\sum_{i=0}^{m}\binom{m+1}{i}B_i(z)=(m+1)z^m$$

其初值 $B_0(z)=1$。

此结论由分步积分证得。

1.1.3 数学极限理论

下面我们介绍用有限差分来数值近似一个函数的导数。例如，函数 f 在点 x 处梯度的第 i 个分量为

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x_i}\approx\frac{f(x+\varepsilon_i e_i)-f(x-\varepsilon_i e_i)}{2\varepsilon_i} \tag{1.1.8}$$

其中，ε_i是任意的一个小数；e_i 是第 i 个梯度方向的单位向量。一般地，人们可从 $\varepsilon_i=0.01$ 或 $\varepsilon_i=0.001$ 开始，采用逐步减少 ε_i 的序列来近似所求导数，并且这种近似方法一般可逐步得到改善，直到当 ε_i 非常小时导致计算退化且计算完全由计算机四舍五入所控制。

有限差分法可用来近似函数 f 在 x 处的二阶导数，即

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x_i\mathrm{d}x_j}\approx\frac{f(x+\varepsilon_i e_i+\varepsilon_j e_j)-f(x+\varepsilon_i e_i-\varepsilon_j e_j)-f(x-\varepsilon_i e_i+\varepsilon_j e_j)+f(x-\varepsilon_i e_i-\varepsilon_j e_j)}{4\varepsilon_i\varepsilon_j}$$

它仍可以用类似 ε_i 序列来改进近似差分。

设已给出 3 个节点 $x_0, x_1=x_0+h, x_2=x_0+2h$ 上的函数值，则它们的导数分别是

$$f'(x_0)\approx\frac{1}{2h}[-3f(x_0)+4f(x_1)-f(x_2)] \quad \text{（端点的导数）}$$

$$f'(x_1)\approx\frac{1}{2h}[-f(x_0)+f(x_2)] \quad \text{（称中心差商，两端点除外的导数）}$$

$$f'(x_2)\approx\frac{1}{2h}[f(x_0)-4f(x_1)+3f(x_2)] \quad \text{（端点的导数）}$$

例 1.1.1　已知 $y=f(x)$ 的数值如表 1.1 所示，用上述公式计算 $x=2.5,2.7$ 处的函数一阶导数值。

表 1.1　$y=f(x)$ 的数值

x	2.5	2.6	2.7	2.8	2.9
y	12.182 5	13.463 7	14.879 7	16.444 6	18.174 1

R 程序如下：

```
zx=c(2.5,2.6,2.7,2.8,2.9);
y=c(12.1825,13.4637,14.8797,16.4446,18.1741);
n=length(x);a=x[2]-x[1];z=0;
for(i in 1:1:n){
    if(i==1) {z[i]=1/(2*a)*(-3*y[1]+4*y[2]-y[3]);}
    else if(i==n) {z[i]=1/(2*a)*(y[n-2]-4*y[n-1]+3*y[n]);}
    else {z[i]=1/(2*a)*(y[i+1]-y[i-1]);}
}
z
```

运行结果为：

```
[1] 12.1380 13.4860 14.9045 16.4720 18.1180
```

所以 $f'(2.5)=12.138\ 0$，$f'(2.7)=14.904\ 5$。

1.2　概率论

目前，中学阶段对学生概率统计知识已有一定要求，基本可满足随机模拟的初步学习。

1.2.1 概率的定义与计算

在一定条件下并不总是出现相同结果的现象称为随机现象，这类现象有两个特点：① 结果不止一个；② 哪一个结果出现，事先不能确定。

随机现象有大量和个别之分。在相同条件下可以（至少原则上可以）重复出现的随机现象，称为大量随机现象；带有偶然性但原则上不能在相同条件下重复出现的随机现象，称为个别随机现象，如一切带有偶然性特点的具体历史事件。

在相同的条件下可以重复的随机现象称为随机试验，即一个试验如果满足：

（1）可以在相同的条件下重复进行；

（2）其结果具有多种可能性；

（3）在每次试验前，不能预言将出现哪一个结果，但知道其所有可能出现的结果。

则称这样的试验为随机试验。

随机试验是对随机现象的一次观察或试验，通常用大写字母 E 表示，简称试验。随机试验的一切可能基本结果组成的集合称为样本空间（**sampling space**），用 Ω 表示；其每个元素称为样本点（**sample point**），又称为基本结果，用 ω 表示。样本点是抽样的基本单元，认识随机现象首先要列出它的样本空间。

在样本空间 Ω 中，具有某种性质的样本点构成的子集称为随机事件，简称事件（**event**），常用大写字母 A,B,C 等表示。在一般情况下，我们也称 Ω 的子集为随机事件。任何样本空间 Ω 都有两个特殊子集，即空集 $\varnothing$ 和 Ω 自身，其中空集 $\varnothing$ 称为不可能事件，指每次试验一定不会发生的事件； Ω 自身称为必然事件，指在每次试验中都必然发生的事件。事件之间的关系与集合之间的关系建立了一定的对应法则，因而事件之间的运算法则与 Borel（布尔）代数中集合的运算法则相同。

如果 F 是样本空间 Ω 中的某些子集的集合，它满足：

（1） $\Omega\in\mathrm{F}$ ；

（2）若 $A\in\mathrm{F}$ ，则 $\overline{A}\in\mathrm{F}$ ；

（3）若 $A_i\in\mathrm{F}, i=1,2,\cdots$ ，则 $\bigcup_{i=1}^{\infty}A_i\in\mathrm{F}$ 。

则称 F 为 σ 域，或 σ 代数、事件域。最简单的 σ 域为 $\{\Omega,\varnothing\}$ ，称为平凡 σ 域。

在概率论中， (Ω,F) 又称为可测空间，这里可测指的是 F 中的元素都具有概率，即都是可度量的。

设 $\Omega=\mathbf{R}$ ， $x\in\mathbf{R}$ ，由所有半无限闭区间 $(-\infty,x]$ 生成的最小 σ 域称为 $\mathbf{R}$ 上的 **Borel** σ 域，记为 $\mathrm{B}(\mathbf{R})$ ，其中的元素称为 **Borel** 集合。

定义 1.2.1　设 Ω 是一个样本空间， F 为 Ω 的某些子集组成的一个事件域，如果对 $\forall A\in\mathrm{F}$ ，定义在 F 上的一个集合函数 $P(A)$ 与之对应，它满足：

（1）非负性公理： $0\leqslant P(A)\leqslant 1$ ；

（2）正则性公理： $P(\Omega)=1$；

（3）可列可加性公理：设 $A_i \in \mathrm{F}, i=1,2,\cdots$，且 $A_i \cap A_j = \varnothing, i \neq j$，有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

则称 $P(A)$ 为事件 A 的概率；P 称为概率测度，简称为概率（**probability**）；三元总体 (Ω, F, P) 称为概率空间。

概率的公理化定义刻画了概率的本质，即概率是集合函数且满足上述三条公理。事件域的引进使我们的模型有了更大的灵活性，在实际问题中我们根据问题的性质选择合适的 F，一般选 Ω 一切子集为 F。事件域可以保证随机事件经过各种运算后仍是随机事件。

利用概率的公理化定义，可导出概率的一系列性质。

性质 1.2.1　$P(\varnothing)=0$

概率论中，将概率很小（通常认为小于 0.05）的事件称作小概率事件（**small probability event**）。小概率事件原理又称为实际推断原理：在原假设成立的条件下，小概率事件在一次试验中可以看成不可能事件，如果在一次试验中小概率事件发生了，则矛盾，即原假设不正确。

小概率事件原理是概率论的精髓，是统计学发展、存在的基础，它使得人们在面对大量数据而需要做出分析与判断时，能够依据具体情况的推理来做出决策，从而使统计推断具备严格的数学理论依据。

性质 1.2.2　有限可加性：设 $A_i \in \mathrm{F}, i=1,2,\cdots,n$，且 $A_i \cap A_j = \varnothing, i \neq j$，则

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

性质 1.2.3　加法公式：对任意两事件 A, B，有

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

加法公式还能推广到多个事件。设 A_1, A_2, A_3 为任意三个事件，则有

$$P(A_1 \cup A_2 \cup A_3) = \sum_{i=1}^{3} P(A_i) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3)$$

利用样本点在等式两端计算次数相等可直观证明这个公式。

一般，对于任意 n 个事件 $A_1, \cdots, A_n$，可以用数学归纳法得

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leqslant i < j \leqslant n} P(A_iA_j) + \sum_{1 \leqslant i < j < k \leqslant n} P(A_iA_jA_k) + \cdots + (-1)^{n-1} P(A_1A_2\cdots A_n)$$

称为容斥原理，也称为多去少补原理。

事件的概率通常是未知的，但公理化定义并没有告诉人们如何去确定概率。历史上在公理化定义出现之前，概率的统计定义、古典定义、几何定义和主观定义都在一定的场合下有着确定概率的方法，所以有了公理化定义后，它们可以作为概率的确定方法。

定义 1.2.2　设 A,B 是两个随机事件，且 $P(B)>0$，称

$$P(A\mid B)=\frac{P(AB)}{P(B)}$$

为在事件 B 发生条件下事件 A 发生的条件概率（**conditional probability**）。

由条件概率的定义立即可得乘法公式

$$P(A\mid B)=\frac{P(AB)}{P(B)}\Rightarrow P(AB)=P(B)P(A\mid B),\quad (P(B)>0)$$

$$P(B\mid A)=\frac{P(AB)}{P(A)}\Rightarrow P(AB)=P(A)P(B\mid A),\quad (P(A)>0)$$

乘法公式也称为联合概率，是指两个任意事件乘积的概率，或称之为交事件的概率，也可称为链式规则，它是一种把联合概率分解为条件概率的方法。对 $\forall n$ 个事件 $A_1,\cdots,A_n$，若 $P(A_1\cdots A_n)>0$，则

$$P(A_1\cdots A_n)=P(A_1)P(A_2\mid A_1)P(A_3\mid A_1A_2)\cdots P(A_n\mid A_1\cdots A_{n-1})$$

全概率公式是概率论中的一个重要公式，它提供了计算复杂事件概率的一条有效途径，使一个复杂事件的概率计算问题化繁为简。设 $A_1,A_2,\cdots,A_n$ 是 Ω 的一组事件，若 $\bigcup_{i=1}^{n}A_i=\Omega$ 且 $A_iA_j=\varnothing(i\neq j)$，则称 $A_1,A_2,\cdots,A_n$ 为 Ω 的一个完备事件组或一个分割。

定理 1.2.1　全概率公式：设 $A_1,\cdots,A_n$ 是 Ω 的一个分割，且 $P(A_i)>0$，$i=1,\cdots,n$，则对任一事件 B 有

$$P(B)=\sum_{i=1}^{n}P(A_i)P(B\mid A_i)$$

已知结果求原因在实际中更为常见，它所求的是条件概率，是已知某结果发生条件下，探求各原因发生可能性大小。

定理 1.2.2　贝叶斯公式：设 $A_1,A_2,\cdots,A_n$ 是 Ω 的一个分割，且 $P(A_i)>0$，$i=1,2,\cdots,n$，若对任一事件 B，$P(B)>0$，则有

$$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{\sum_{j=1}^{n}P(A_j)P(B\mid A_j)},\qquad (i=1,2,\cdots,n)$$

一般来说，$P(A\mid B)\neq P(A)$，$P(B)>0$，这表明事件 B 的发生提供了一些信息，影响了事

件 A 发生的概率。但在有些情况下，$P(A\mid B)=P(A)$，从这可以想象得到这必定是事件 B 的发生对 A 发生的概率不产生任何影响，或不提供任何信息，即事件 A 与 B 发生的概率是相互不影响的，这就是事件 A,B 相互独立。

定义 1.2.3　若两事件 A,B 满足 $P(AB)=P(A)P(B)$，则称 A 与 B 相互独立。

由于概率为 0 或 1 的事件之间具有非常复杂的关系，故请初学者注意：

（1）$\varnothing,\Omega$ 与任何事件相互独立；进一步有：概率为 0 或 1 的事件与任何事件相互独立。假如往线段 $[0,1]$ 上任意投一点，令事件 $A=$“点落在 0”，事件 $B=$“点落在 0 或 1”，则 $A\subset B$，但事件 A,B 相互独立。

（2）事件的独立是指事件发生的概率互不影响，但可同时发生，而互不相容只是说两个事件不能同时发生，故事件 A,B 互不相容 Á 事件 A,B 相互独立。

对于三个事件 A,B,C，若下列四个等式同时成立

$$\begin{cases}P(AB)=P(A)P(B)\\P(AC)=P(A)P(C)\\P(BC)=P(B)P(C)\\P(ABC)=P(A)P(B)P(C)\end{cases}\tag{1.2.1}$$

则称 A,B,C 相互独立（**independent**）。

若式 1.2.1 中只有前三个式子成立，则称 A,B,C 两两相互独立。

直观上说，当进行几个随机试验时，如果每个试验无论出现什么结果都不影响其他试验中各事件出现的概率，就称这些试验是独立的。

设有两个试验 E_1 和 E_2，假如试验 E_1 的任一结果与试验 E_2 的任一结果都是相互独立的事件，则称这两个试验是相互独立的。

1.2.2　随机变量

为全面研究随机试验的结果，揭示客观存在的统计规律性，我们必须将随机试验的结果与实数对应起来，即必须把随机试验的结果数量化，这就是引入随机变量的原因。随机变量的引入使得对随机现象的处理更简单与直接，也更统一而有力，更便于进行定量的数学处理。

定义 1.2.4　定义在样本空间 $\Omega=\{\omega\}$ 到实数集上的一个实值单值函数 $X(\omega)$ 称为随机变量（**random variables**），若对 $\forall x\in\mathbf{R}$，$\{\omega\mid X(\omega)\leqslant x\}$ 都是随机事件，即

$$\forall x\in\mathbf{R},\ \{\omega\mid X\leqslant x\}\in\mathrm{F}$$

进一步有：设 (Ω,F,P) 为一概率空间，$X(.)=(X_1,\cdots,X_n)$ 是定义在 $\Omega=\{\omega\}$ 上的 n 元实值函数，如果对 $\forall x=(x_1,\cdots,x_n)\in\mathbf{R}^n$，有

$$\{\omega\mid X_1(\omega)\leqslant x_1,\cdots,X_n(\omega)\leqslant x_n\}\in\mathrm{F}$$

则称 $X(.)=(X_1,\cdots,X_n)$ 为 n 维随机向量，也称为 n 维随机变量。

为掌握 X 的统计规律，只需掌握 X 取各值的概率，于是引入概率分布来描述随机变量的统计规律，但概率分布难以表示，我们可根据概率的累加特性，引入分布函数 F 来描述随机变量的统计规律。

设 X 为随机变量， x 为任意实数，称函数

$$F(x)=P\{X\leqslant x\}$$

为 X 的分布函数（**cumulative distribution function**），且称 X 服从 $F(x)$，记为 $X\sim F(x)$。

进一步有：

$$F(x)=F(x_1,\cdots,x_n)=P(X_1\leqslant x_1,\cdots,X_n\leqslant x_n)，\quad \forall x=(x_1,\cdots,x_n)\in\mathbf{R}^n$$

称为 $X=(X_1,\cdots,X_n)$ 的联合分布函数。

联合分布函数含有丰富的信息，我们的目的是将这些信息从联合分布中挖掘出来。下面，我们先讨论每个变量的分布，即边际分布。

二维随机向量 (X,Y) 具有联合分布 $F(x,y)$， X,Y 都是随机变量，各自有分布函数，将它们记为 $F_X(x)$， $F_Y(y)$，依次称为二维随机向量 (X,Y) 关于 X 和关于 Y 的边际分布函数。边际分布也称为边缘分布。

顾名思义，边际分布就是二维随机向量 (X,Y) 边的分布，即 X,Y 的分布函数，边际分布函数可以由联合分布函数确定。事实上，

$$F_X(x)=P(X\leqslant x)=P(X\leqslant x,Y\leqslant+\infty)=F(x,+\infty)，\text{即}$$

$$F_X(x)=F(x,+\infty)$$

同理可得

$$F_Y(y)=F(+\infty,y)$$

随机变量在概率与统计的研究中应用非常普遍，因此可以这样说，如果微积分是研究变量的数学，那么概率与统计是研究随机变量的数学。通常可将随机变量分为两类：

（1）离散型随机变量：所有取值可以一一列举。

若随机变量 X 只可能取有限或可列个值，则称 X 为离散型随机变量。如果离散型随机变量 X 的所有可能取值为 $x_1,\cdots,x_n,\cdots$，则称 X 取 x_i 的概率为

$$p_i=p(x_i)=P(X=x_i),\qquad (i=1,2,\cdots)$$

为 X 的概率分布列，简称分布列，记作 $X\sim\{p_i\}$。

分布列也可表示为 $\begin{pmatrix}x_1 & \cdots & x_n & \cdots\\ p_1 & \cdots & p_n & \cdots\end{pmatrix}$，或表示为概率分布表

X	x_1	x_2	$\cdots$	x_n	$\cdots$
P	p_1	p_2	$\cdots$	p_n	$\cdots$

如果二维随机向量 (X,Y) 只取有限个或可列个数对 (x_i,y_j)，则称 (X,Y) 为二维离散随机向量，称

$$p_{ij}=P(X=x_i,Y=y_j),\quad (i,j=1,2,\cdots)$$

为 (X,Y) 的联合分布列。

对于离散型随机变量 X，边际分布为

$$F_X(x)=F(x,+\infty)=\sum_{x_i\leqslant x}\sum_{j=1}^{+\infty}p_{ij}\ ,\quad F_Y(y)=F(+\infty,y)=\sum_{y_j\leqslant y}\sum_{i=1}^{+\infty}p_{ij}$$

离散随机变量 X 的分布列为

$$P(X=x_i)=P(X=x_i,Y\leqslant+\infty)=\sum_{j=1}^{+\infty}P(X=x_i,Y=y_j)=\sum_{j=1}^{+\infty}p_{ij}=p_{i\square},\ (i=1,2,\cdots)$$

同理可得离散随机变量 Y 的分布列为

$$P(Y=y_j)=\sum_{i=1}^{+\infty}p_{ij}=p_{\square j}\ ,\quad (j=1,2,\cdots)$$

分别称 $p_{i\square},i=1,2,\cdots$ 和 $p_{\square j}$，$j=1,2,\cdots$ 为 (X,Y) 关于 X，Y 的边际分布列。

对一切使得 $P(Y=y_j)=\sum_{i=1}^{+\infty}p_{ij}=p_{\square j}>0$ 的 y_j，称

$$p_{i|j}=P(X=x_i\mid Y=y_j)=\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)}=\frac{p_{ij}}{p_{\square j}}\ ,\quad (i=1,2,\cdots)$$

为在给定 $Y=y_j$ 条件下 X 的条件分布列。

同理，对一切使得 $P(X=x_i)=\sum_{j=1}^{+\infty}p_{ij}=p_{i\square}>0$ 的 x_i，称

$$p_{j|i}=P(Y=y_j\mid X=x_i)=\frac{P(X=x_i,Y=y_j)}{P(X=x_i)}=\frac{p_{ij}}{p_{i\square}}\ ,\quad (j=1,2,\cdots)$$

为在给定 $X=x_i$ 条件下 Y 的条件分布列。

有了条件分布列，我们就可以定义离散随机向量的条件分布。

在给定 $Y=y_j$ 条件下 X 的条件分布函数为

$$F(x\mid y_j)=P(X\leqslant x\mid Y=y_j)=\sum_{x_i\leqslant x}P(X=x_i\mid Y=y_j)=\sum_{x_i\leqslant x}p_{i|j}$$

在给定 $X=x_i$ 条件下 Y 的条件分布函数为

$$F(y\mid x_i)=\sum_{y_i\leqslant y}P(Y=y_j\mid X=x_i)=\sum_{y_i\leqslant y}p_{j|i}$$