数学地质丛书

# 概率统计

程光华　蒋耀淞　张一球　编

地质出版社

数学地质丛书

# 概率统计

程光华　蒋耀淞　张一球　编

地质出版社

数学地质丛书

# 概率统计

程光华　蒋耀淞　张一球　编

•

地质部书刊编辑室编辑

责任编辑：高书平

地质出版社出版

（北京西四）

地质印刷厂印刷

（北京海淀区学院路29号）

新华书店北京发行所发行·各地新华书店经售

•

开本：850×1168 1/32 印张：9 5/4 字数：255,000

1982年2月北京第一版·1982年2月北京第一次印刷

印数1—10,280册·定价1.80元

统一书号：15038·新679

# 前　言

近二十年来，随着电子计算机在地质学领域中日益广泛的应用，逐渐形成了一门地质与数学互相渗透、紧密结合的边缘学科——数学地质。

“数学地质”包括的内容相当广泛，研究随机现象的学科在其中占有特别重要的地位。地质现象是在漫长的地质年代中形成的，它的形成受着多种错综复杂因素的控制和影响，因而常表现出“随机性”的特点。地质多元统计、地质统计学以及地质学中的随机过程等都是数学地质中研究随机现象的分支学科。

本书是作为对“随机现象”的数量规律性的一般的、初步的研究而列入这套丛书的。对于丛书中的其它各册，它起着介绍必要的基础理论知识的作用。

为了使广大地质工作人员在阅读时亲切易懂些，在编写过程中，我们力图用大家所熟悉的问题和例子来阐明一些基本的概念和方法。但是，有时为了避免概念的不确切和阐述的过份冗长，还是不得不放弃这种企图而采用一些“概率统计”教科书中常见的典型例子。

在编写本书时，我们主要参考了武汉地质学院数学教研室蒋耀淞和王仁锋等同志历次为学生和短训班学员所编写的讲义。

这里我们特别要感谢长春地质学院数学地质研究室夏立显、矫希国两位同志，他们曾详细地审阅了初稿，并提出了很多宝贵的意见和建议。

本书插图由武汉地质学院绘图室王润斋同志绘制，我们表示感谢。

由于我们的水平有限，本书肯定会有错误和不妥之处，恳请读者批评指正。

编　者

1980年10月

# 目　录

带有*号的章节对部分读者来说可略去。

# 第一章 数据整理

在地质工作中经常会接触到许多数据资料，如岩石的化学分析资料，孔隙度的测定值，一个矿区的磁异常值，等等。这些数据资料在未经整理之前往往是比较杂乱的，不容易看出蕴含于其中的统计规律性，必须进行整理，才能从中提取有价值的信息。在实际工作中，还常常将整理的结果制成醒目的图表。这一章就来介绍数据的初步分析整理的方法。

## §1.1 频率分布直方图与累积频率多角形

先来看一个例子，我们将结合这个例子来说明数据整理的方法。

**例1.1** 对某地闪长岩体的60个样品进行了铜（Cu）的含量分析，得到结果如表1.1。

**1. 分组**

为了概括数据信息，通常将观测值分组。经验表明，一般说来可以分成10至20组，当观测值个数较少时（如少于50个），也可以分成10组以下（5～7组）。一般都用等间距分组，组的间隔长度叫做组距 $l$。组距 $l$ 可由数据的上界和下界以及分组数 $n$ 求得（上界和下界可取比数据的最大值、最小值再向外扩展一些的数）

$$\text{组距 } l = \frac{\text{上界} - \text{下界}}{n}$$

有时将最靠两端，即包含最极端值的组合并是比较方便的。

在地质工作中，对于大多数微量元素常常先求出含量的常用对数值，再按其对数值来整理图表。下面我们对表1.1中的数据

表1.1 某地闪长岩中的比色分析结果（附常用对数值）

| 样品号 | Cu含量(γ/g) | 对数值 | 样品号 | Cu含量(γ/g) | 对数值 | 样品号 | Cu含量(γ/g) | 对数值 |
|---|---|---|---|---|---|---|---|---|
| 1 | 28 | 1.4472 | 21 | 14 | 1.1451 | 41 | 22 | 1.3424 |
| 2 | 20 | 1.3010 | 22 | 10 | 1.0000 | 42 | 45 | 1.6532 |
| 3 | 4 | 0.6021 | 23 | 48 | 1.6812 | 43 | 13 | 1.1139 |
| 4 | 20 | 1.3010 | 24 | 40 | 1.6021 | 44 | 13 | 1.1139 |
| 5 | 16 | 1.2041 | 25 | 8 | 0.9031 | 45 | 5 | 0.6990 |
| 6 | 32 | 1.5051 | 26 | 96 | 1.9823 | 46 | 15 | 1.1761 |
| 7 | 10 | 1.0000 | 27 | 60 | 1.0000 | 47 | 18 | 1.2553 |
| 8 | 20 | 1.3010 | 28 | 20 | 1.3010 | 48 | 11 | 1.0414 |
| 9 | 16 | 1.2041 | 29 | 6 | 0.7782 | 49 | 17 | 1.2304 |
| 10 | 20 | 1.3010 | 30 | 20 | 1.3010 | 50 | 20 | 1.3010 |
| 11 | 8 | 0.9031 | 31 | 32 | 1.5051 | 51 | 13 | 1.1139 |
| 12 | 10 | 1.0000 | 32 | 10 | 1.0000 | 52 | 14 | 1.1461 |
| 13 | 12 | 1.0792 | 33 | 20 | 1.3010 | 53 | 8 | 0.9031 |
| 14 | 10 | 1.0000 | 34 | 11 | 1.0414 | 54 | 13 | 1.1139 |
| 15 | 16 | 1.2041 | 35 | 28 | 1.4472 | 55 | 24 | 1.3802 |
| 19 | 6 | 0.7782 | 36 | 8 | 0.9031 | 56 | 25 | 1.3979 |
| 17 | 16 | 1.2041 | 37 | 13 | 1.1139 | 57 | 8 | 0.9031 |
| 18 | 12 | 1.0792 | 38 | 13 | 1.1139 | 58 | 18 | 1.2553 |
| 19 | 16 | 1.2041 | 39 | 10 | 1.0000 | 59 | 6 | 0.7782 |
| 20 | 16 | 1.2041 | 40 | 20 | 1.3001 | 60 | 8 | 0.9031 |

就是先取对数再来整理。可以看到，数据的最大值是1.9823，最小值是0.6021。可取上界为2.08，下界为0.48，并分成8组，因而组距是

$$l=\frac{1.6}{8}=0.2$$

2．列表

按组距 $l=0.2$ 把 0.48～2.08 划分为8个间隔。间隔的中点值称为组中值，记为 $x_i(i=1,\cdots,8)$，落入各组中的数据个数称为频数，记作 $f_i^*$。计算频率 $f_i(\%)$，$f_i=\frac{f_i^*}{N}$（这里 $N=60$，是数据总个数，$i$ 由1到8）。将前面各组的频率依次相加，就得到累

积频率值$F_i(\%)$，$F_i=\sum_{k=1}^{i} f_k$，$i=1,\cdots,m$；$F_N=\sum_{k=1}^{m} f_k=1$。

于是，就可以列成表1.2。

**表 1.2 闪长岩中Cu含量对数值的频数、频率和累积频率分布表**

| 间　　隔 | 组 中 值 $x_i$ | 频 数 $f_i$* | 频 率 $f_i$(%) | 累积频率$F_i$(%) |
|---|---|---|---|---|
| 0.48～0.68 | 0.58 | 1 | 1.7 | 1.7 |
| 0.68～0.88 | 0.78 | 4 | 6.6 | 8.3 |
| 0.88～1.08 | 0.98 | 17 | 28.4 | 36.7 |
| 1.08～1.28 | 1.18 | 18 | 30.0 | 66.7 |
| 1.28～1.48 | 1.38 | 14 | 23.3 | 90.0 |
| 1.48～1.68 | 1.58 | 4 | 6.6 | 96.6 |
| 1.68～1.88 | 1.78 | 1 | 1.7 | 98.3 |
| 1.88～2.08 | 1.98 | 1 | 1.7 | 100.0 |

### 3. 制图

图形比表格更为直观。有了表 1.2 就不难进一步绘制成频率分布直方图和累积频率多角形。

直方图是由一系列相邻的长方形作成的。在频率分布直方图（图 1.1）中，各小长方形的底边的长即各间隔的宽度，而其上的小长方形的面积等于该组的频率值，从而第 $i$ 个长方形的高是

$$y_i=\frac{\text{长方形面积}}{\text{长方形底边长}}=\frac{\text{频率}\,f_i}{\text{组距}\,l}$$

$y_i$ 表示在 $x$ 轴的单位长度上平均分布了多大的频率，故 $y_i$ 的意义可以说是频率分布密度。容易看出，全部长方形的面积的总和等于 1（因为$\sum_{i=1}^{m} f_i=1$）。

如果在各组上限处作纵坐标等于对应累积频率值 $f_i(\%)$ 的点，并依次用线段连接各点，就得到累积频率多角形（图1.2）。因为最后一组的累积频率为 100％＝1，所以该组上限处的线段的长恒为 1。

为了更加直观、方便，实践中还常采用图表结合的方法，如

图 1.3 所示。

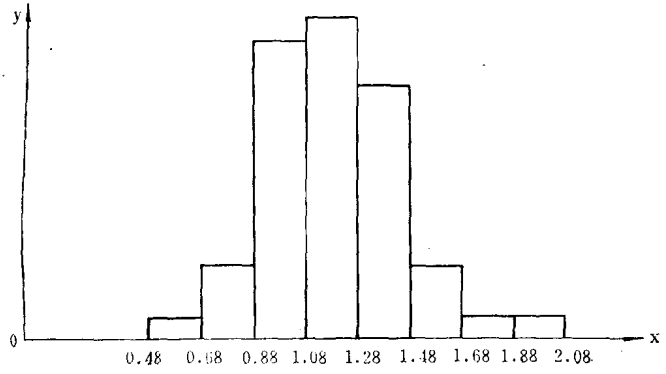



图 1.1 闪长岩中Cu含量对数值的频率分布直方图

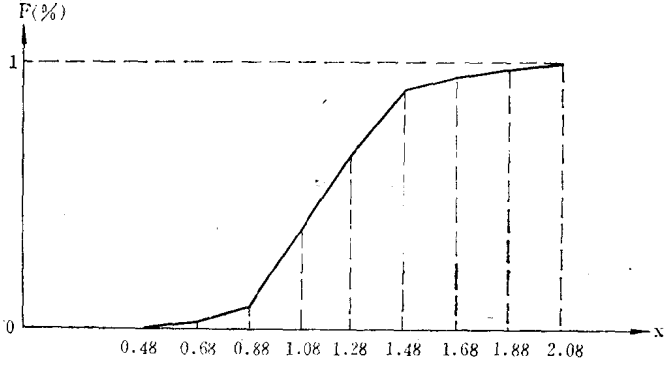
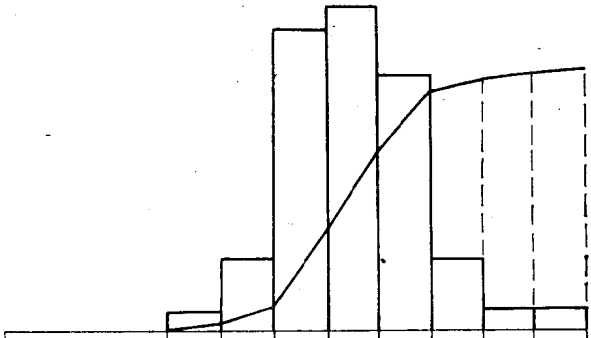



图 1.2 闪长岩中Cu含量对数值的累积频率多角形

## §1.2 两类重要的特征数

经过分组、列表、制图，我们对这一批数据的频率分布就有了一个直观的了解。例如，从上面的图表就立即可以看出，约有80%以上的样品的铜含量对数值落在 0.88～1.48 的范围之内，等等。但还可以引进几个能反映频率分布的重要统计特征的参数，以进一步简化上述表示。这些参数称为统计特征数或简称特征



| 组中值 $x_i$ | 0.58 | 0.78 | 0.98 | 1.18 | 1.38 | 1.58 | 1.78 | 1.98 |
|---|---|---|---|---|---|---|---|---|
| 频数 $f_i^*$ | 1 | 4 | 17 | 18 | 14 | 4 | 1 | 1 |
| 频率 $f_i$(%) | 1.7 | 6.6 | 28.4 | 30.0 | 23.3 | 6.6 | 1.7 | 1.7 |
| 累积频率 $F_i$(%) | 1.7 | 8.3 | 36.7 | 66.7 | 90.0 | 96.6 | 98.3 | 100.0 |

图 1.3 图表结合

数。

常用的特征数可分为两大类：一类是反映数据分布的集中情况或者说中心趋势的，因而它们可以用来作为这批数据的典型代表，如算术平均数、几何平均数、众数、中位数等；另一类是反映数据分布的离散程度的，如标准差、极差、变化系数等。

**1. 算术平均数（简称平均数）**

**定义1.1** $N$个数据$x_1$，$x_2$，…，$x_N$的**算术平均数**是

$$\bar{x}=\frac{1}{N}(x_1+x_2+\cdots+x_N)=\frac{1}{N}\sum_{i=1}^{N}x_i \qquad (1.1)$$

当$N$较大，且$x_1$，…，$x_N$分成为$n$组时，若以$y_1$，…，$y_n$记各组的组中值，以$f_1^*$，…，$f_n^*$记相应的频数，$f_1$，…，$f_n$记各频率，则平均数可用下式来定义

$$\bar{x}=\frac{1}{N}(f_1^*y_1+f_2^*y_2+\cdots+f_n^*y_n)=\frac{1}{N}\sum_{i=1}^{n}f_i^*y_i$$

$$=\sum_{i=1}^{n}\frac{f_i^*}{N}y_i=\sum_{i=1}^{n}f_iy_i \qquad (1.1)'$$

公式 (1.1)′ 中出现的平均数也称为是 $y_1$，…，$y_n$ 的“加权”平均数，在这里以频率 $f_i$ 作为权。

对例 1.1 中 60 个样品的 Cu 含量的对数值，按公式 (1.1) 可算得

$$\bar{x}=\frac{1}{60}(1.4472+1.3010+\cdots+0.9031)=1.1677$$

而按公式 (1.1)′，则可求得其平均数是

$$\bar{x}=\frac{1}{60}(1\times 0.58+4\times 0.78+17\times 0.98+18\times 1.18+14\times 1.38$$

$$+4\times 1.58+1\times 1.78+1\times 1.98)=\frac{71}{60}=1.18$$

**2. 几何平均数**

**定义1.2** $N$个数据 $x_1$，$x_2$，…，$x_N$ (均$>0$) 的**几何平均数**为

$$\bar{x}_g=\sqrt[N]{x_1x_2\cdots x_N}=\left(\prod_{i=1}^{N}x_i\right)^{\frac{1}{N}} \tag{1.2}$$

如果这$N$个数已分成了$n$组，组中值分别为$y_1$，…，$y_n$，仍以$f_1^*$，…，$f_n^*$记对应频数，$f_1$，…，$f_n$记对应频率，则几何平均数还可定义如下

$$\bar{x}_g=\sqrt[N]{y_1^{f_1^*}\cdot y_2^{f_2^*}\cdot\cdots\cdot y_n^{f_n^*}}=\left(\prod_{i=1}^{n}y_i^{f_i^*}\right)^{\frac{1}{N}}=\prod_{i=1}^{n}y_i^{f_i} \tag{1.2}'$$

为了方便，往往利用对数来计算，对上面 (1.2) 和 (1.2)′ 式取对数，分别有

$$\log\bar{x}_g=\frac{1}{N}(\log x_1+\cdots+\log x_N)=\frac{1}{N}\sum_{i=1}^{N}\log x_i \tag{1.3}$$

和

$$\log\bar{x}_g=\frac{1}{N}\sum_{i=1}^{n}f_i^*\log y_i=\sum_{i=1}^{n}f_i\log y_i \tag{1.3}'$$

下面我们给出一个简单的例子。

**例 1.2** 求八个数：6.5，8.5，4.7，9.4，11.3，8.5，9.7，

9.4 的算术平均数和几何平均数。

解

$$\bar{x}=\frac{1}{8}(6.5+8.5+4.7+9.4+11.3+8.5+9.7+9.4)=8.5$$

$$\bar{x}_g=\sqrt[8]{6.5\times8.5^2\times4.7\times9.4^2\times11.3\times9.7}=8.25$$

如先求其对数

| $x_i$ | 6.5 | 8.5 | 4.7 | 9.4 | 11.3 | 9.7 |
|---|---|---|---|---|---|---|
| $\log x_i$ | 0.8129 | 0.9294 | 0.6721 | 0.9731 | 1.0531 | 0.9868 |

按公式（1.3）计算

$$\log \bar{x}_g=\frac{1}{8}(0.8129+2\times0.9294+0.6721+2\times0.9731$$
$$+1.0531+0.9868)=0.9162$$

同样得到

$$\bar{x}_g=8.25$$

**3. 中位数与众数**

反映数据分布的中心趋势的特征数中最常用的是算术平均数，它具有易于计算，计算时考虑到了全部观测值，并且在不同的抽样中，其值也比较稳定等优点，因而在统计中被广泛地应用着。除了算术平均值与几何平均值以外，有时还用到“中位数”和“众数”。简略地说，中位数是把总频数分成相等两半的变量值，而众数是对应最大频数分布的变量值，或者说是最容易出现的变量值，因而可以定义如下：

**定义 1.3** 对于按大小次序排列的$N$个数$x_1,x_2,\cdots,x_N$，则中位数（记作$M_d$）是

$$M_d=x_{\frac{N+1}{2}} \qquad \text{当}N\text{是奇数}$$

$$M_d=\frac{1}{2}(x_{\frac{N}{2}}+x_{\frac{N}{2}+1}) \qquad \text{当}N\text{是偶数}$$

**定义 1.4** 对应于最大频率的组的组中值称为众数。当有多于一个组具有最大频率时，则有多个众数。

对于有同样中心（例如同样的算术平均值）的数据，仍能呈现着很不相同的情况，我们来看下面的图1.4。

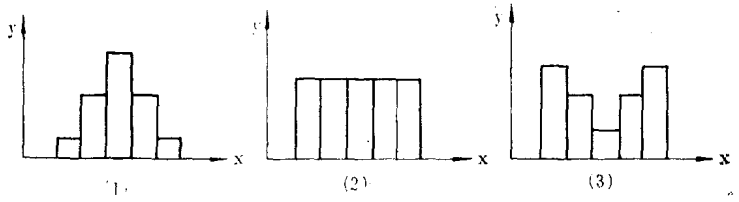



图 1.4

在图 1.4 的三种情形中，中心是相同的，但在（1）中，数据较多地围绕在中心的附近，在（2）中数据均匀地散布着，而在（3）中则较多的是偏离中心的极端的数据。因而我们必须来考虑描述离散情况的特征数。

**4. 极差**

**定义 1.5** 观测值中最大值与最小值之差称为**极差**。若一批观测数据为 $x_1, x_2, \cdots, x_N$，则极差 $R$ 是

$$R=\max\{x_1, \cdots, x_N\}-\min\{x_1, \cdots, x_N\} \qquad (1.4)$$

其中 $\max\{x_1, \cdots, x_N\}$ 和 $\min\{x_1, \cdots, x_N\}$ 分别表示 $x_1, x_2, \cdots, x_N$ 中的最大值和最小值。

极差的优点是计算简便、迅速，但由于它只利用了两个极端值，没有充分利用全部数据所提供的信息，因而本身极不稳定，反映实际情况的精确度较差。

**5. 标准差（或称均方差）**

描述数据离散情况的特征数中最重要的是标准差。

**定义1.6** $N$个观测值 $x_1, x_2, \cdots, x_N$ 的**方差**是

$$s^2=\frac{1}{N}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2]$$

$$=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 \qquad (1.5)$$

而称

$$s=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} \tag{1.6}$$

为**标准差**（或称**均方差**），其中$\bar{x}$是算术平均数。

标准差与观测值有相同的单位。在统计中广泛地应用方差或标准差来衡量观测值对其平均数的离散程度。$s$越大，数据越分散；$S$越小，数据越集中在$\bar{x}$附近。

当$N$个数据$x_1$，…，$x_N$被分成为$n$组时，以$y_1$，$y_2$，…，$y_n$记组中值，仍以$f_1^*$，$f_2^*$，…，$f_n^*$与$f_1$，$f_2$，…，$f_n$分别记各组频数与频率，且以$\bar{y}$记各$y_i$（以$f_i^*$为权）的加权平均数，则方差与标准差可定义如下

$$s^2=\frac{1}{N}[f_1^*(y_1-\bar{y})^2+f_2^*(y_2-\bar{y})^2+\cdots+f_n^*(y_n-\bar{y})^2]$$

$$=\frac{1}{N}\sum_{i=1}^{n}f_i^*(y_i-\bar{y})^2=\sum_{i=1}^{n}f_i(y_i-\bar{y})^2 \tag{1.5}'$$

$$s=\sqrt{\frac{1}{N}\sum_{i=1}^{n}f_i^*(y_i-\bar{y})^2}=\sqrt{\sum_{i=1}^{n}f_i(y_i-\bar{y})^2} \tag{1.6}'$$

对于例 1.1 中的数据，我们已算得$\bar{y}=1.18$，再由公式(1.5)′与 (1.6)′，可算得

$$s^2=\frac{1}{60}[1\times(0.58-1.18)^2+4\times(0.78-1.18)^2$$
$$+17\times(0.98-1.18)^2+18\times(1.18-1.18)^2$$
$$+14\times(1.38-1.18)^2+4\times(1.58-1.18)^2$$
$$+1\times(1.78-1.18)^2+1\times(1.98-1.18)^2]=0.0647$$

$$s=\sqrt{0.0647}=0.254$$

**6. 变化系数（或称变异系数）**

我们知道，在实际计算中，平均值不等的两批数据，如果它们的标准差相等，则平均值较大的那批数据相对地波动小些，而平均值较小的那批数据相对波动大些。因此有时需要反映数据相对离散程度的一个指标，于是就引进了变化系数。

**定义1.7** 标准差与平均数之比称为**变化系数**（通常以百分比表示），记作$c_v$，即

$$c_v=\frac{s}{\bar{x}}\times 100\%$$

它是一个无量纲的数。

一批观测数据在进行了上述的数据整理、编制图表、计算特征数之后，对于数据使用者来说，比起一开始时面对杂乱无章的一堆数据的状况，确实是前进了一步，它们所代表的含义清楚得多了。但在实际工作中，数据的整理不能到此为止。例如在例1.1中，对某地闪长岩体采取的60个样品进行铜含量分析的目的不仅仅是想描述一下这60个样品中铜含量的分布情况，求出这60个样品的平均含量，更主要的是想通过样品的分析研究而对整个闪长岩体中铜的含量进行某种合理的推断，以进一步指导找矿工作。也就是说，在直方图、特征数所提供的概括之外，我们还要求再前进一步，而这就必须先转向统计规律方面的理论的探讨，然后再回到解决这些实际问题上来。

# 第二章 预备知识
# （集合、排列组合）

## §2.1 集合及其运算

为了较好地理解概率论中的一些基本概念，如事件及其运算等，我们在这一章里把集合论的一些基本知识作一简单的介绍。

**1. 集合的概念**

集合这个概念是数学中最基本的概念之一。它是如此的基本，以至于找不到比它更原始的概念来给它下定义，而只能采取描述性的办法来加以说明。

**集合**（简称为集）便是具有某种共同特征的具体的或抽象的事物的总体，其中每个事物，就称为这个集合的元素。

**例2.1** 小于10的全部奇数构成一个集。它包含五个元素：1，3，5，7，9。

**例2.2** 全部自然数构成一个集。它包含可列无穷多个元素：1，2，3，…。

**例2.3** 全部有理数构成一个集合。

**例2.4** 数轴上全部点构成一个点集。

**例2.5** 圆周 $x^2+y^2=1$ 上全部的点构成一个集。

**例2.6** 地质博物馆中全部岩浆岩标本构成一个集合。

我们通常用大写字母 $A$，$B$，$C$，…表示集合，用小写字母 $a$，$b$，$c$，…表示集合的元素。如果能够明确写出集 $A$ 的所有元素，也可以都列举在大括号里面。例如，小于10的全部奇数构成的集，就是{1，3，5，7，9}。

取任何事物 $a$，对于某个集 $A$ 来说，或者 $a$ 是 $A$ 的元素——这时，我们说 $a$ 属于 $A$，记为 $a\in A$；或者 $a$ 不是 $A$ 的元素——即