第六册

大青花鱼

目录

第一章　向量

第五册中，我们学习了用三角函数解三角形。三角函数是定量研究平面形的利器。不过，三角函数本身并不是简单的函数。我们目前只能通过查表的方式得到函数值。这让我们思考，能不能打造一种更方便定量研究的体系呢?

回顾我们对平面形的研究，我们从几条公理出发，得出点、直线、三角形、圆等形状之间的定性关系。公理体系的缺陷在于没有与数紧密结合。比如，“两点之间直线最短”，除了定性的“最短”，没有提供别的信息。我们需要一种根本上和数量结合的体系，来理解各种平面形状。

此外，公理体系中并没有强调运动的概念。我们说点运动形成了线，旋转形成角度和圆，但并没有相关的工具来描述具体的运动。我们需要一种根本上和运动结合的体系，来理解形状之间的关系。

1.1　点、向量和直线

学习有理数的时候，我们使用数轴上的点表示。每个点代表一个实数。两点重合，当且仅当它们代表同一个数。这种表示方法把数和直线上的点牢牢绑在一起。我们可以用数的关系表示直线上点的关系。数轴使我们可以定量理解直线。

至于平面中的点，我们用相互垂直的数轴定义了点的坐标。每个点代表一个有序数对。两个数按顺序排列，对应平面中一点。

能不能像数轴一样，用一个量代表平面中一点呢？数轴之所以能用一个数代表一个点，是因为直线只有两个方向，使用正负号就可以代表方向。平面中不止两个方向，我们无法用正负来表示方向了。为此，我们引入一个新的量来代表平面中的点：**向量**。

自然数、有理数、实数都有自己的运算法则。向量作为代表点的量，需要满足怎样的运算法则呢？我们从运动出发，给出以下的法则：

- 向量的加法就是平移：两个向量相加得到另一个向量。向量的加法满足结合律和交换律。
- 零向量表示静止不变：存在这样一个向量，任何向量与它相加，仍然是自己。这个向量叫做零向量。零向量不定义方向，也可以说它与任何向量同向或反向。它对应的点称为**原点**。
- 非零向量放缩，引出一根数轴：任何实数乘以向量，得到方向相同或相反的向量。这个运算称为**数乘运算**。数乘运算对应图形的放缩。
- 放缩和四则运算相容：数轴上可以做数的运算。
- 平移和放缩相容：先平移再放缩，和先放缩再平移，结果一样。

按照定义，**向量就是点**，所以可以用大写字母来表记。比如零向量就是原点，记为 O。此外，**向量就是平移**。点 A 就是把 O 对应到 A 的平移，也是 O 平移的结果，记为 $\overrightarrow{OA}$。

让我们用数学语言把这些法则具体写出来。我们把平面看作集合，记为 $\mathbb{V}$，其中的元素称为向量或点，用粗体小写字母表示，以便和代表数的量区分：

- 加法结合律：$\forall\ \mathbf{a},\mathbf{b},\mathbf{c}\in\mathbb{V}$，$\mathbf{a}+(\mathbf{b}+\mathbf{c})=(\mathbf{a}+\mathbf{b})+\mathbf{c}$。
- 加法交换律：$\forall\ \mathbf{a},\mathbf{b}\in\mathbb{V}$，$\mathbf{a}+\mathbf{b}=\mathbf{b}+\mathbf{a}$。
- 存在零向量：$\forall\ \mathbf{a}\in\mathbb{V}$，$\mathbf{a}+\mathbf{0}=\mathbf{a}$。

- 放缩和四则运算相容：$\forall\ \mathbf{a} \in \mathbb{V}$，$1 \cdot \mathbf{a} = \mathbf{a}$。$\forall s, t \in \mathbb{R}$，$(s+t) \cdot \mathbf{a} = (s \cdot \mathbf{a}) + (t \cdot \mathbf{a})$，$(s \cdot t) \cdot \mathbf{a} = s \cdot (t \cdot \mathbf{a})$。
- 放缩和平移相容：$\forall\ \mathbf{a}, \mathbf{b} \in \mathbb{V}$，$\forall\ t \in \mathbb{R}$，$t \cdot (\mathbf{a} + \mathbf{b}) = t \cdot \mathbf{a} + t \cdot \mathbf{b}$。

从以上法则出发，我们可以定义直线：

定义 1.1.1. 过原点的直线是非零向量放缩得到的集合。不过原点的直线是过原点的直线按一点平移得到的集合。

给定非零向量 $A = \mathbf{a}$，$\{t\mathbf{a} \,|\, t \in \mathbb{R}\}$ 是一条过原点 O 和 A 的直线 OA，称为 A **引出**的直线，记为 $\mathbb{R}\mathbf{a}$。给定向量 $B = \mathbf{b}$，$\{t\mathbf{a} + \mathbf{b} \,|\, t \in \mathbb{R}\}$ 是一条过 B 的直线，记为 $\mathbb{R}\mathbf{a} + \mathbf{b}$，其中 $\mathbb{R}\mathbf{a}$ 称为它的直部；而 $\{t\mathbf{a} + (1-t)\mathbf{b} \,|\, t \in \mathbb{R}\}$ 就是直线 AB。

给定非零向量 $\mathbf{a}$，如果向量 $\mathbf{b}$ 可以通过 $\mathbf{a}$ 放缩得到，或者说 $\mathbf{b} \in \mathbb{R}\mathbf{a}$，就称两者**共轴**，表示两者在共同的数轴上。

类比可以定义线段和射线：给定非零向量 $A = \mathbf{a}$ 和向量 $B = \mathbf{b}$，$\{(1-t)\mathbf{a} + t\mathbf{b} \,|\, t \in [0;1]\}$ 是线段 AB，$\{(1-t)\mathbf{a} + t\mathbf{b} \,|\, t \geqslant 0\}$ 是射线 AB。

这样定义的线段和射线，也具备了数轴的性质。比如，在线段 $\{(1-t)A + tB \,|\, t \in [0;1]\}$ 中，t 的不同值就对应了不同的点：$t = 0$ 对应点 A，$t = 1$ 对应点 B。对一般的 $t \in (0;1)$，$(1-t)A + tB$ 对应的点 P 在 AB 上，可以看作数轴 AB 上的点。因此满足：

$$\begin{aligned} |AP| &= |P - A| = |tB - tA| && = t|AB|, \\ |PB| &= |B - P| = |(1-t)B - (1-t)A| && = (1-t)|AB|. \end{aligned}$$

也就是说，P 是线段 AB 上使得 $\frac{|AP|}{|PB|} = \frac{t}{1-t}$ 的点。

反过来，设 $\frac{|AP|}{|PB|}$ 等于定值 $k > 0$，对应的点 P 是什么点呢？这个问题实际上是求方程：

$$\frac{t}{1-t} = k$$

的解。容易解出这个方程的唯一解：$t = \frac{k}{k+1}$。因此我们得到结论：

定理 1.1.1. 定比分点定理 线段 AB 上到两端距离之比 $\frac{|AP|}{|PB|}$ 为定值 k 的点 P 恰有一个，称为它的 k **分点**。

正数 k 越小，k 分点距离 A 越近，k 越大，k 分点离 B 越近；$k=1$ 时，我们就得到线段的中点。

例题 1.1.1. 三角形 ABC 的边 AB、BC 上分别有点 P、Q，使得：

$$\frac{|AP|}{|PB|}=\frac{|BQ|}{|QC|}=2,$$

定义点 X：

$$X=\frac{A+2B+4C}{7}.$$

验证：X 是直线 AQ 和 PC 的交点。

解答. 证明：设 $P=(1-t)A+tB$。P 是线段 AB 的 2 分点，所以

$$t=\frac{2}{2+1}=\frac{2}{3}.$$

即：

$$P=\frac{1}{3}A+\frac{2}{3}B=\frac{A+2B}{3}.$$

同理，设 $Q=(1-t)B+tC$，则 $t=\frac{2}{3}$，即：

$$Q=\frac{1}{3}B+\frac{2}{3}C=\frac{B+2C}{3}.$$

由于线段的定比分点在直线上，只要找出合适的参数 t，使得 $X=(1-t)A+tQ$，就说明 X 在 AQ 上。

$$X=(1-t)A+tQ=(1-t)A+\frac{t}{3}B+\frac{2t}{3}C,$$

因此，令 $t=\frac{6}{7}$，就有 $1-t=\frac{1}{7}$、$\frac{t}{3}=\frac{2}{7}$、$\frac{2t}{3}=\frac{4}{7}$。即 $(1-t)A+\frac{t}{3}B+\frac{2t}{3}C=\frac{A+2B+4C}{7}$。这说明：

$$X=(1-\frac{6}{7})A+\frac{6}{7}Q.$$

也就是说，X 在直线 AQ 上，是线段 AQ 的 $\frac{6}{7}$ 分点。
同理可以验证，

$$(1-\frac{4}{7})\,P+\frac{4}{7}\,C=\frac{3}{7}\cdot\frac{B+2C}{3}+\frac{4}{7}\,C=\frac{A+2B+4C}{7}.$$

这说明：

$$X=(1-\frac{4}{7})\,P+\frac{4}{7}\,C.$$

也就是说，X 在直线 PC 上，是线段 PC 的 $\frac{4}{7}$ 分点。
综上所述，X 既在直线 AQ 上也在直线 PC 上，是直线 AQ 和 PC 的交点。

以上我们讨论了 $k>0$ 的情况，显然，$k=0$ 对应 $P=A$。对于负数 k，有没有对应的点呢？我们用平移的思想考虑这个问题，从 A 到 P 经历的平移是 $\overrightarrow{AP}=t\overrightarrow{AB}$，从 P 到 B 经历的平移是 $\overrightarrow{PB}=(1-t)\overrightarrow{AB}$。它们的系数之比就是 $\frac{t}{1-t}$。于是，我们可以对一般的 k 定义定比分点：如果 k 能使得方程

$$\frac{t}{1-t}=k$$

有唯一解，那么我们就把对应的点 P 称为 AB 的 k 分点。

如果 $k<-1$，那么 k 分点对应的 $t=\frac{k}{k+1}>1$，也就是说，P 在线段 AB 沿 B 的延长线上。如果 $-1<k<0$，那么 k 分点对应的 $t=\frac{k}{k+1}<0$，也就是说，P 在线段 AB 沿 A 的延长线上。如果 $k=-1$，以上方程无解，这说明 -1 分点不存在。

共轴的向量，通过数轴，可以方便地讨论相互的位置关系。不共轴的向量之间，如何讨论位置关系呢？为此，我们要引入**平面的根本性质**：

- 给定任何非零向量 A，平面中总有另一个向量 B，不在直线 OA 上。我们说两者**不共轴**。
- 从不共轴的向量 A,B 出发，经过放缩、平移，可以得到平面中任何向量。具体来说，任何向量都可以表示成 $sA+tB$ 的形式，集合 $\{sA+tB\,|\,s,t,\in\mathbb{R}\}$ 就是整个平面。这样的 A,B 称为平面的一组**基**或**基底**。

举例来说，在直角坐标系中，我们选择了原点重合、互相垂直的两条数轴，以每条数轴上数 1 对应的点（记为 $\mathbf{e}_x, \mathbf{e}_y$）出发，通过放缩和平移，就得到平面所有的点。平面中任一点可以写成 $x\mathbf{e}_x + y\mathbf{e}_y$，其中 x, y 就是点的坐标。直角坐标系其实是一种用向量描述平面的方法。$\mathbf{e}_x, \mathbf{e}_y$ 就是一组基底。给定基底后，平面中的点用两个实数构成的有序数对（也就是坐标）表示，所以平面也记为 $\mathbb{R}^2$。

我们知道，平面中的点的坐标是唯一的。从向量的角度来说，可以表达成：用基底 A, B 表示向量的方法只有一种。如果向量 P 既能写成 $P = s_1A + t_1B$，又能写成 $P = s_2A + t_2B$，那么 $s_1 = s_2$、$t_1 = t_2$。我们说，作为基底的向量 A, B 是**直无关**的：

定义 1.1.2. 考虑一组向量 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ 和同样数量的系数 $t_1, t_2, \cdots, t_n$。以下的向量：

$$t_1\mathbf{a}_1 + t_2\mathbf{a}_2 + \cdots + t_n\mathbf{a}_n$$

称为这组向量（关于系数 $t_1, t_2, \cdots, t_n$）的**直组合**。$t_1, t_2, \cdots, t_n$ 称为直组合的系数。能表示为 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ 的直组合的向量，称为这组向量的**生成向量**，所有生成向量的集合称为这组向量的**生成集**。

如果一组向量的生成向量只能用唯一的方式表示为它们的直组合，就说这组向量是**直无关**的，否则就说这组向量是**直约束**的或**直相关**的。

从这些定义出发，我们容易想到一些基本的性质。比如，一组向量的生成集总会包括这些向量自身，也总包括零向量。

从平面的根本性质，可以得到：

定理 1.1.2. 平面基本定理 平面中不共轴的两个向量总是直无关的，生成集是整个平面。多于两个向量如果不共轴，总是直约束的。

证明： 从平面的根本性质第二条可以直接得出：平面中不共轴的两个向量，生成集是整个平面，称为平面的基底。下面证明平面的基底总是直无关的。

使用反证法来证明。反设某组基底 A, B 是直约束的，即有某个向量 P 可以用两种形式表示成 A, B 的直组合：

$$P = s_1A + t_1B = s_2A + t_2B.$$

且有序数对 $(s_1,\ t_1)$ 和 $(s_2,\ t_2)$ 不完全相等。不妨设 $s_1 \neq s_2$。从上式可以得到零向量的表示：

$$\mathbf{0} = (s_1 - s_2)A + (t_1 - t_2)B.$$

于是

$$A = -\frac{t_1 - t_2}{s_1 - s_2}B.$$

A, B 共轴，矛盾！因此原命题成立：平面的基底总是直无关的。

多于两个向量 $P_1, P_2, \cdots, P_n$，如果不共轴，那么至少有两个向量不共轴。记为 P_1, P_2，则 P_1, P_2 构成了平面的基底。其他向量都可以表示成 P_1, P_2 的直组合。

考虑用这组向量生成零向量。首先，令所有系数为 0。零向量总能表示为关于系数 $(0, 0, \cdots, 0)$ 的直组合。其次，在其他向量里面选一个，比如 P_k，P_k 可以表示为 P_1, P_2 的直组合：$P_k = sP_1 + tP_2$。于是令 P_1, P_2 的系数分别为 s, t，P_k 的系数为 -1，其他向量系数为 0。零向量表示为：

$$\mathbf{0} = s \cdot P_1 + t \cdot P_2 + (-1) \cdot P_k + \sum_{i \neq 1,2,k} 0 \cdot P_i$$

零向量表示方式不唯一，说明这组向量是直约束的。 □

从证明中可以看出，零向量的表示方式是判别向量是否直相关的方便方法。这是因为生成集里面某向量表示不唯一，等于说零向量表示不唯一。而我们总知道零向量的一种表示方式：所有系数为 0。这样，我们只需要讨论另一种表示方式是否存在。我们也可以把这个思考方法总结成：

定理 1.1.3. 直无关向量组的判定 直无关的向量组只能用系数全为零的方式表示零向量。只能用系数全为零的方式表示零向量的向量组是直无关的。

换句话说，给定 n 个向量 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$，如果存在不全为零的实数 $t_1, t_2, \cdots, t_n$，使得直组合 $t_1\mathbf{a}_1 + t_2\mathbf{a}_2 + \cdots + t_1\mathbf{a}_n$ 是零向量，那么这 n 个向量直相关。如果不存在这样一组实数，那么这 n 个向量直无关。

平面基底是直无关的，也就是说，任何向量表示成基底的直组合，方法是唯一的。使用恰当的基底，可以从更好的角度看问题。

例题 1.1.2. 三角形 ABC 的边 AB、BC 上分别有点 P、Q，使得:

$$\frac{|AP|}{|PB|} = \frac{|BQ|}{|QC|} = 2,$$

试用 A, B, C 对应的向量，表示直线 AQ 和 PC 的交点。

解答. 前面我们验证了这个交点存在。这里我们来思考：如果一开始不知道交点，如何把它求出来。我们已经知道如何用 A, B, C 表示 P 和 Q:

$$P = \frac{A+2B}{3}, \quad Q = \frac{B+2C}{3}.$$

记交点为 X。下面我们用 $\overrightarrow{BA}$ 和 $\overrightarrow{BC}$ 来表示 $\overrightarrow{BP}$ 和 $\overrightarrow{BQ}$，然后分别用 $\overrightarrow{BP}$ 和 $\overrightarrow{BC}$、$\overrightarrow{BA}$ 和 $\overrightarrow{BQ}$ 来表示交点 X。$\overrightarrow{BA}$ 和 $\overrightarrow{BC}$ 不共轴，所以构成平面的基底。根据平面基本定理，这两种表示方式都可以归于 $\overrightarrow{BA}$ 和 $\overrightarrow{BC}$ 的直组合，即归根结底是同一个方式。这样，我们就能建立方程组，推出 X 的具体表示方式。

首先考虑 $\overrightarrow{BP}$，

$$\overrightarrow{BP} = P - B = \frac{A+2B}{3} - B = \frac{A-B}{3} = \frac{1}{3}\overrightarrow{BA},$$

同理，

$$\overrightarrow{BQ} = \frac{2}{3}\overrightarrow{BC}.$$

设 $\overrightarrow{BX} = (1-t)\overrightarrow{BP} + t\overrightarrow{BC} = (1-s)\overrightarrow{BA} + s\overrightarrow{BQ}$，则:

$$\begin{aligned}\frac{1-t}{3}\overrightarrow{BA} + t\overrightarrow{BC} &= (1-t)\overrightarrow{BP} + t\overrightarrow{BC} \\ &= (1-s)\overrightarrow{BA} + s\overrightarrow{BQ} = (1-s)\overrightarrow{BA} + \frac{2s}{3}\overrightarrow{BC}.\end{aligned}$$

两种表达方式一致，所以对应系数相同。因此有以下方程组：

$$\begin{cases} \frac{1-t}{3} &= 1-s \\ t &= \frac{2s}{3} \end{cases}$$

解得 $t=\frac{4}{7}$，$s=\frac{6}{7}$。代入上面 $\overrightarrow{BX}$ 的表达式，可以求得：

$$\overrightarrow{BX}=(1-t)\overrightarrow{BP}+t\overrightarrow{BC}=\frac{3}{7}\overrightarrow{BP}+\frac{4}{7}\overrightarrow{BC}=\frac{1}{7}\overrightarrow{BA}+\frac{4}{7}\overrightarrow{BC}.$$

即：

$$X-B=\frac{A-B}{7}+\frac{4C-4B}{7}.$$

于是：

$$X=\frac{A+2B+4C}{7}.$$

这也是我们前面验证过的交点。

思考 1.1.1.

1. 设平面上有两点 A,B，以 OA,OB 为邻边作平行四边形 $AOBC$。向量 $\overrightarrow{OA}$ 和 $\overrightarrow{BC}$ 是什么关系？

2. 设平面上有两点 A,B，三角形 OAB 中，连接边 OA,OB 的中点 M,N。向量 $\overrightarrow{AB}$ 和 $\overrightarrow{MN}$ 是什么关系？

3. 例题 *1.1.2* 的解答里，如果关于 s,t 的方程组无解，代表什么？是否存在无解的情况？

习题 1.1.1.

1. 证明：零向量只有一个，任何向量乘 0 得到零向量。

2. 证明：零向量乘任何数得到零向量。

3. 证明：任何向量 $\mathbf{a}$ 都有唯一的反向量 $\mathbf{b}$，满足 $\mathbf{a}+\mathbf{b}=\mathbf{0}$。

4. 设 $\mathbf{a},\mathbf{b}$ 不共轴，如果 $s\mathbf{a}+t\mathbf{b}=\mathbf{0}$，证明：$s=t=0$。

直角坐标系 xOy 中，设 $\mathbf{a}=4\mathbf{e}_x+\mathbf{e}_y$，$\mathbf{b}=\mathbf{e}_x-2\mathbf{e}_y$ ，$\mathbf{b}=-\mathbf{e}_x+2\mathbf{e}_y$。

5. 在坐标轴上标出 $\mathbf{a}$，$\mathbf{b}$ 和 $\mathbf{c}$。

6. 用 $\mathbf{a}$ 和 $\mathbf{b}$ 表示 $\mathbf{e}_x$、$\mathbf{e}_y$ 和点 $(3,0)$。

7. 用 $\mathbf{a}$ 和 $\mathbf{b}$ 表示它们的中点、3 分点、-0.5 分点、-3 分点。写出这些点的坐标和直线的方程。

8. 证明：向量组的生成集总包含向量组中的各个向量，以及零向量。

9. 证明：共轴的向量总是直约束的。

10. 用 $\mathbf{a}, \mathbf{b}, \mathbf{c}$ 表示顶点为 $\mathbf{a}, \mathbf{b}, \mathbf{c}$ 的三角形三边和重心。

1.2　角度与长度

根据平面的根本性质，任何向量都可以用两个不共轴向量表示。如何讨论它们的位置关系呢？下面我们定义一种关系，把长度、距离和角度统一起来。

给定平面基底 $\mathbf{e}_1, \mathbf{e}_2$，我们给出这样一个二元映射 f：

$$\forall\ \mathbf{a} = x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ \mathbf{b} = x_B\mathbf{e}_1 + y_B\mathbf{e}_2\ \in \mathbb{R}, \quad f(\mathbf{a},\ \mathbf{b}) = x_Ax_B + y_Ay_B.$$

f 把两个向量对应到一个实数。它满足以下五个性质：

- 向量的顺序不影响关系大小：

$$\begin{aligned} & f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ x_B\mathbf{e}_1 + y_B\mathbf{e}_2) \\ = {} & x_Ax_B + y_Ay_B \\ = {} & x_Bx_A + y_By_A \\ = {} & f(x_B\mathbf{e}_1 + y_B\mathbf{e}_2,\ x_A\mathbf{e}_1 + y_A\mathbf{e}_2). \end{aligned}$$

- 零向量和任意向量关系为 0：

$$\begin{aligned} f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ \mathbf{0}) &= f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ 0\mathbf{e}_1 + 0\mathbf{e}_2) \\ &= x_A \cdot 0 + y_A \cdot 0 \\ &= 0. \end{aligned}$$

- 非零向量与自身的关系总是正的：x_A, y_A 不全为零时，

$$f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ x_A\mathbf{e}_1 + y_A\mathbf{e}_2) = x_A^2 + y_A^2 > 0.$$

- 和向量的放缩相容：

$$\begin{aligned}
&f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ t\cdot(x_B\mathbf{e}_1 + y_B\mathbf{e}_2))\\
=\ &x_A\cdot t\cdot x_B + y_A\cdot t\cdot y_B\\
=\ &t\cdot(x_Ax_B + y_Ay_B)\\
=\ &t\cdot f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ x_B\mathbf{e}_1 + y_B\mathbf{e}_2).
\end{aligned}$$

- 和向量的平移相容：

$$\begin{aligned}
&f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ (x_B\mathbf{e}_1 + y_B\mathbf{e}_2) + (x_C\mathbf{e}_1 + y_C\mathbf{e}_2))\\
=\ &x_A(x_B + x_C) + y_A(y_B + y_C)\\
=\ &(x_Ax_B + y_Ay_B) + (x_Ax_C + y_Ay_C)\\
=\ &f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ x_B\mathbf{e}_1 + y_B\mathbf{e}_2) + f(x_A\mathbf{e}_1 + y_A\mathbf{e}_2,\ x_C\mathbf{e}_1 + y_C\mathbf{e}_2).
\end{aligned}$$

满足以上五个条件的映射 f 称为平面向量的**内积**。从第四个性质可知，向量与自身的内积总是正数。我们把这个数的平方根叫做**向量的长度**，记为：

$$\forall\ \mathbf{a}\in\mathbb{V},\quad \|\mathbf{a}\| = \sqrt{f(\mathbf{a},\mathbf{a})}.$$

两个向量之差的长度，称为向量之间的距离。

$$\forall\ \mathbf{a},\mathbf{b}\in\mathbb{V},\quad \|\mathbf{a}-\mathbf{b}\| = \sqrt{f(\mathbf{a}-\mathbf{b},\mathbf{a}-\mathbf{b})}.$$

如果基底 $\mathbf{e}_1, \mathbf{e}_2$ 是直角坐标系的基，那么

$$\begin{aligned}
&\forall\ \mathbf{a} = x_A\mathbf{e}_x + y_A\mathbf{e}_y,\\
&\qquad \|\mathbf{a}\| = \sqrt{x_A^2 + y_A^2},\\
&\forall\ \mathbf{a} = x_A\mathbf{e}_x + y_A\mathbf{e}_y,\ \ \mathbf{b} = x_B\mathbf{e}_x + y_B\mathbf{e}_y,\\
&\qquad \|\mathbf{a}-\mathbf{b}\| = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}.
\end{aligned}$$

给定向量 $A=\mathbf{a}$、$B=\mathbf{b}$，$\|\mathbf{a}\|$ 就是 $|OA|$，$\|\mathbf{a}-\mathbf{b}\|$ 就是 $|AB|$。也就是说，通过内积定义的映射 f，分别与直观经验中长度和距离的概念相符合。

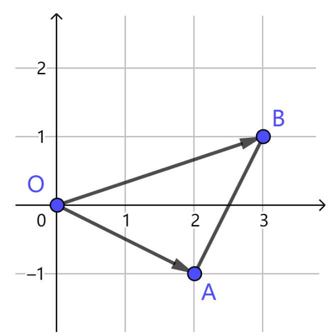
那么，f 本身有什么含义呢？我们来计算 $\frac{|OA|^2+|OB|^2-|AB|^2}{2}$.

$$\begin{aligned}\frac{|OA|^2+|OB|^2-|AB|^2}{2}&=\frac{x_A^2+y_A^2+x_B^2+y_B^2-(x_A-x_B)^2-(y_A-y_B)^2}{2}\\&=x_Ax_B+y_Ay_B=f(\mathbf{a},\mathbf{b}).\end{aligned}$$

另一方面，余弦定理告诉我们，

$$\frac{|OA|^2+|OB|^2-|AB|^2}{2}=|OA||OB|\cos\angle AOB.$$

也就是说，$f(\mathbf{a},\mathbf{b})=\|\mathbf{a}\|\|\mathbf{b}\|\cos\angle AOB$。内积 f 的本质是向量夹角的余弦与向量长度的乘积。通过内积，我们把角度和长度统一起来了。



例题 1.2.1. 给定向量 $A(2,\ -1)$ 和 $B(3,\ 1)$，求它们的内积。

解答. 向量 A 和 B 的内积为 $2\cdot3+(-1)\cdot1=5$。如右图，如果计算线段 OA 和 OB 长度，可得：$|OA|=\sqrt{5}$，$|OB|=\sqrt{10}$。于是

$$\cos\angle AOB=\frac{f(\overrightarrow{OA},\ \overrightarrow{OB})}{|OA|\cdot|OB|}=\frac{5}{\sqrt{5}\cdot\sqrt{10}}=\frac{\sqrt{2}}{2}.$$

即 $\angle AOB=45^\circ$。

向量夹角的余弦值总在 -1 和 1 之间，所以向量的内积的绝对值不大于向量长度的乘积：

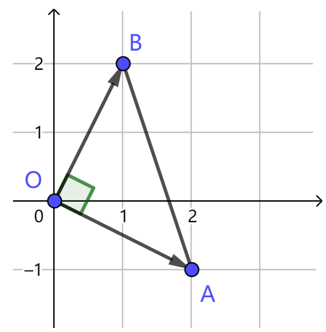
$$|x_Ax_B+y_Ay_B|\leqslant\sqrt{x_A^2+y_A^2}\sqrt{x_B^2+y_B^2}.$$

可以验证这个关系对任意 x_A,y_A,x_B,y_B 成立。从这个关系出发，可以得到：

$$|AB|=\sqrt{(x_A-x_B)^2+(y_A-y_B)^2}\leqslant\sqrt{x_A^2+y_A^2}+\sqrt{x_B^2+y_B^2}=|OA|+|OB|.$$

这符合直观经验中“三角形两边之和大于第三边”或“两点之间线段距离最短”的性质。

内积为 0，就表示向量夹角的余弦为 0，即两个向量垂直。比如，令 $\mathbf{a} = 2\mathbf{e}_x - \mathbf{e}_y$，$\mathbf{b} = \mathbf{e}_x + 2\mathbf{e}_y$，那么 $f(\mathbf{a},\ \mathbf{b}) = 2 \cdot 1 - 1 \cdot 2 = 0$。在平面上画出对应的点 A, B，可以验证 $\angle AOB = 90°$。



内积映射并不是唯一的，我们来看另一个内积映射 f_2：

$$\forall\ x_A, y_A, x_B, y_B \in \mathbb{R}, \quad f_2(x_A\mathbf{e}_1 + y_A\mathbf{e}_2, x_B\mathbf{e}_1 + y_B\mathbf{e}_2) = 2x_Ax_B + y_Ay_B.$$

可以验证，f_2 也满足 f 满足的五个性质。从 f_2 出发，我们也可以定义距离和长度：

$$\forall\ \mathbf{a} = x_A\mathbf{e}_x + y_A\mathbf{e}_y, \quad \|\mathbf{a}\|_2 = f_2(\mathbf{a}, \mathbf{a}) = 2x_A^2 + y_A^2.$$

这样定义的距离和长度和我们直观经验中有些不一样，不过，我们可以验证，这样定义的距离也满足“两点之间线段最短”的性质。

$$|2x_Ax_B + y_Ay_B| \leqslant \sqrt{2x_A^2 + y_A^2}\sqrt{2x_B^2 + y_B^2}.$$

因此，f_2 也是内积。

我们把符合直观经验的内积 f 称为**经典内积**，一般称内积都默认指经典内积；把对应的长度称为向量的**模**。我们把向量 $\mathbf{a}, \mathbf{b}$ 的（经典）内积记为 $(\mathbf{a} \,|\, \mathbf{b})$，不至于混淆时，也常称为**点积**，记为 $\mathbf{a} \cdot \mathbf{b}$；把它们的模记为 $|\mathbf{a}|$、$|\mathbf{b}|$。

既然有余弦，自然有正弦。记 $\alpha = \angle AOB$，则 $(\mathbf{a} \,|\, \mathbf{b}) = |\mathbf{a}||\mathbf{b}| \cos\alpha$，于是，

$$|\mathbf{a}|^2|\mathbf{b}|^2 \sin^2\alpha = |\mathbf{a}|^2|\mathbf{b}|^2 - (\mathbf{a} \,|\, \mathbf{b})^2$$

记 $\mathbf{a} = x_A\mathbf{e}_x + y_A\mathbf{e}_y$，$\mathbf{b} = x_B\mathbf{e}_x + y_B\mathbf{e}_y$，则

$$\begin{aligned}
(x_A^2 + y_A^2)(x_B^2 + y_B^2)\sin^2\alpha &= (x_A^2 + y_A^2)(x_B^2 + y_B^2) - (x_Ax_B + y_Ay_B)^2 \\
&= (x_Ay_B - x_By_A)^2 \\
|\sin\alpha| &= \frac{|x_Ay_B - x_By_A|}{\sqrt{x_A^2 + y_A^2}\sqrt{x_B^2 + y_B^2}}
\end{aligned}$$

我们得出了夹角 $\angle AOB$ 正弦的绝对值。

观察向量夹角的正弦和余弦，我们注意到，它们的表达式与和差角公式有相似之处。$x_Ax_B+y_Ay_B$ 与差角余弦公式形式相似，$x_Ay_B-x_By_A$ 与差角正弦公式形式相似。

让我们在直角坐标系中找几个例子，看看直观结果。设有点 $A(1,\ 0)$、$B(\frac{1}{2}\ \frac{\sqrt{3}}{2})$。不难得出 $\angle AOB=60°$。我们用以上公式计算 $\angle AOB$ 的正弦和余弦：

$$\frac{x_Ay_B-x_By_A}{\sqrt{x_A^2+y_A^2}\sqrt{x_B^2+y_B^2}}=\frac{\sqrt{3}}{2},\quad \frac{x_Ax_B+y_Ay_B}{\sqrt{x_A^2+y_A^2}\sqrt{x_B^2+y_B^2}}=\frac{1}{2}.$$

把 P 的坐标换成 $(0,\ 1)$、$(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$、$(\frac{\sqrt{3}}{2},-\frac{1}{2})$ 等，我们发现，通过以上两个公式得到的值，分别是 $\angle AOB$ 的正弦、余弦值。记 $\angle AOB=\alpha$，那么：

$$\sin\alpha=\frac{x_Ay_B-x_By_A}{\sqrt{x_A^2+y_A^2}\sqrt{x_B^2+y_B^2}},\quad \cos\alpha=\frac{x_Ax_B+y_Ay_B}{\sqrt{x_A^2+y_A^2}\sqrt{x_B^2+y_B^2}}.$$

我们是通过面积定义正弦的。比如，邻边为 OA 和 OB 的平行四边形，面积是 $|OA||OB|\sin\angle AOB$。对照上面正弦的表达式，可以发现这个面积等于 $x_Ay_B-x_By_A$。于是，我们把对应的映射

$$(\mathbf{a},\ \mathbf{b})\mapsto x_Ay_B-x_By_A.$$

称为向量 A,B 的**面积**，记为 $\mathbf{a}\wedge\mathbf{b}$。

向量的面积和内积，分别对应正弦和余弦。两向量面积为零，当且仅当它们共轴；两向量内积为零，当且仅当它们互相垂直。

习题 1.2.1.

1. 直角坐标系中，已知两向量，计算它们的内积和面积，讨论它们的关系。

1.1. $A(0,\ 2)$, $B(1,\ 1)$

1.2. $A(2,\ 1)$, $B(0.5,\ -1)$

1.3. $A(1.6,\ 0.2)$, $B(-0.9,\ -3)$

1.4. $A(1,\ -0.28)$, $B(-0.45,\ -0.6)$

2. 直角坐标系中，已知向量 B 的模为 2，根据以下条件，求向量 A, B 的内积:

2.1 $A = (-4,\ 2)$, $\angle AOB = 60°$

2.2 $A = (0,\ 5)$, $\angle AOB = 135°$

2.3.$A = (3,\ -2.5)$, $\angle AOB = 45°$

3. 直角坐标系中，已知点 $P(2,\ 1)$，求使得 P, Q 内积为 4 的点 Q。

4. 直角坐标系中，已知点 $P(2,\ 1)$，求使得 P, Q 面积为 4 的点 Q。

思考 1.2.1. 如果直角坐标系 xOy 的 x 轴和 y 轴沿逆时针排布，而角度的正方向为顺时针方向，角 α 的正弦和余弦是否还能写成

$$\sin\alpha = \frac{x_A y_B - x_B y_A}{\sqrt{x_A^2 + y_A^2}\sqrt{x_B^2 + y_B^2}}, \quad \cos\alpha = \frac{x_A x_B + y_A y_B}{\sqrt{x_A^2 + y_A^2}\sqrt{x_B^2 + y_B^2}}$$

的形式?

1.3 直线的方程

我们已经讨论过直线和方程的关系。直角坐标系中，任意给定二元一次方程 $ax + by + c = 0$（a、b 不全为 0），它的解集都对应平面中一条直线。我们把这个二元一次方程称为直线的**一般式**方程。比如，$3x - 2y + 1 = 0$ 就是一条直线的一般式方程。不过，从这个式子，我们不容易看出相应的直线有什么性质。已知直线的一些直观性质，我们希望将这些直观性质和直线的方程联系起来，以便通过方程的方式，来分析直线以及直线和其他平面形的关系。

比如，已知直线过点 $A(x_A,\ y_A)$，斜率为 k。直线斜率为 k，说明直线是某个一次函数 $x \mapsto kx + b$ 的图像。对比可知，直线方程为:

$$y - y_A = k(x - x_A).$$

我们把这个方程称为直线的**点斜式**方程。已知直线上一点和直线的斜率，可以写出直线的点斜式方程。比如，过 $(1,\ 2)$，斜率为 2 的直线方程为：$y-2=2(x-1)$，即 $y-2x=0$。

下面我们用向量的语言，给出有不同直观性质的直线的方程。

点向式：已知直线过点 $A(x_A,\ y_A)$，方向向量为 $\mathbf{b}=(x_B,\ y_B)$，那么直线为 $A+\mathbb{R}\mathbf{b}$。考虑直线上一点 $P(x,\ y)$，$\overrightarrow{AP}$ 和 $\mathbf{b}$ 共轴，所以面积为 0。于是 $(x,\ y)$ 满足方程：

$$(x-x_A)y_B-(y-y_A)x_B=0.$$

我们把这个二元一次方程称为直线的点向式方程。已知直线上一点和直线的方向，可以写出直线的点向式方程。比如，过 $(1,\ 2)$，方向为 $(-1,\ 1)$ 的直线方程为：$1\cdot(x-1)-(-1)\cdot(y-2)=0$，即 $x+y=3$。

两点式：已知直线过点 $A(x_A,\ y_A)$ 和点 $B(x_B,\ y_B)$，那么直线为 $A+\mathbb{R}\overrightarrow{AB}$。考虑直线上一点 $P(x,\ y)$，则 $\overrightarrow{AP}$ 和 $\overrightarrow{AB}$ 共轴。于是 $(x,\ y)$ 满足方程：

$$(x-x_A)(y_B-y_A)-(y-y_A)(x_B-x_A)=0.$$

我们把以上方程称为直线的两点式方程。已知直线上不同的两点，可以写出直线的两点式方程。比如，过 $(1,\ 2)$、$(-2,\ 1)$ 的直线方程为：$(x-1)(1-2)-(y-2)(-2-1)=0$，即 $-x+3y=5$。

点法式：已知直线过点 $A(x_A,\ y_A)=\mathbf{a}$，并且和 $\mathbf{b}=(x_B,\ y_B)$ 垂直。考虑直线上一点 $P(x,\ y)=\mathbf{p}$，$\mathbf{p}-\mathbf{a}$ 和 $\mathbf{b}$ 垂直，所以内积为 0。于是 $(x,\ y)$ 满足方程：

$$(x-x_A)x_B+(y-y_A)y_B=0.$$

我们把 $\mathbf{b}$ 称为直线的**法向量**，把以上方程称为直线的点法式方程。已知直线上一点和法向量，可以写出直线的点法式方程。比如，过 $(1,\ 2)$，法向量为 $(3,\ -1)$ 的直线方程为：$(x-1)\cdot 3+(y-2)\cdot(-1)=0$，即 $3x-y=1$。

等高式：已知点 $P(x,\ y)=\mathbf{p}$ 与 $B(x_B,\ y_B)=\mathbf{b}$ 的面积为 S，则 $(x,\ y)$ 满足方程：

$$xy_B-yx_B=S.$$

我们把这个二元一次方程称为直线的等高式方程。从直观上看，它表示所有以 OB 为底，面积相等（从而高相等）的三角形 OBP 的顶点 P 的集合，即一条平行于 OB 的直线。比如，与 $(1,\ -1)$ 的面积为 3 的点构成直线，方程为：$-x+y=3$。

$S=0$ 时，我们就得到直线 OB。也就是说，已知 B 的坐标 $(x_B,\ y_B)$，直线 OB 的方程是 $xy_B-yx_B=0$。

等垂式：已知点 $P(x,\ y)=\mathbf{p}$ 与 $B(x_B,\ y_B)=\mathbf{b}$ 的内积为 T，则 $(x,\ y)$ 满足方程：

$$xx_B+yy_B=T.$$

我们把这个二元一次方程称为直线的等垂式方程。

从直观上看，作 P 到直线 OB 的垂线，垂足为 H。设 $\overrightarrow{OH}=t\overrightarrow{OB}$，

$$T=\overrightarrow{OP}\cdot\overrightarrow{OB}=\overrightarrow{OH}\cdot\overrightarrow{OB}+\overrightarrow{HP}\cdot\overrightarrow{OB}.$$

$HP\perp OB$，所以 $\overrightarrow{HP}\cdot\overrightarrow{OB}=0$。于是 $T=t\overrightarrow{OB}\cdot\overrightarrow{OB}=t|OB|^2$，算得 $t=\frac{T}{|OB|^2}$，$\overrightarrow{OH}=\frac{T}{|OB|^2}\overrightarrow{OB}$。这说明 H 是定点。

反之，过 H 点作垂直 OB 的直线。直线上任一点到 OB 的垂足是 H，因此 $\overrightarrow{OP}\cdot\overrightarrow{OB}=T$。

这说明直线 $xx_B+yy_B=T$ 是一条垂直于 OB 的直线，是所有到 OB 的垂足为定点 H 的点 P 的集合。比如，与点 $B(1,\ -1)$ 的内积为 3 的点构成直线，方程为：$x-y=3$。它垂直于直线 OB：$x+y=0$。

使用等高式和等垂式，我们可以讨论直线平行和垂直的关系。我们来看两个典型问题：

例子 1.3.1. 求经过点 $P(x_P,\ y_P)$ 并平行于直线 l： $ax+by+c=0$ 的直线（如果存在）的方程。

解答. 首先把直线方程转为等高式：$ax-(-b)y=-c$，它表示平行于向量 $(a,-b)$，与它的面积为 $-c$ 的直线。如果 P 在直线上，那么 P 满足 $ax_P-(-b)y_P=-c$。这时符合要求的平行线不存在。如果 P 不在直线上，我们可以构造直线 l'： $ax-(-b)y=ax_P-(-b)y_P$。l' 过 P，且与向量 $(a,-b)$ 平行，因此平行于 l。

另一方面，如果有直线过点 P 并平行于 l，那么它与向量 $(a,-b)$ 平行。设它的方程为 $ax-(-b)y=S$。由于 P 在直线 $ax-(-b)y=S$ 上，P 满足 $ax_P-(-b)y_P=S$，即平行线方程为 $ax-(-b)y=ax_P-(-b)y_P$。显然，$ax_P-(-b)y_P=-c$ 时，两直线重合，不符合平行线定义。

因此，$ax_P-(-b)y_P=-c$（P 在直线上）时无平行线；其他情况下，恰有一条平行线：$ax-(-b)y=ax_P-(-b)y_P$。化简后的方程为：

$$ax+by=ax_P+y_P.$$

另解: 首先把直线方程转为等垂式：$ax+by=-c$，它表示与点 $M(a,b)$ 的内积为 $-c$ 的直线，也就是垂直于 OM，垂足 Q 满足 $\overrightarrow{OQ}=\frac{-c}{|OM|^2}\overrightarrow{OM}$ 的直线。如果有直线过点 P 并平行于 l，那么它也和 OM 垂直。把它写成等垂式：$ax+by=T$。由于 P 在 l' 上，所以 $T=ax_P+y_P$。于是 l' 的方程为 $ax+by=ax_P+y_P$。显然，$ax_P+by_P=-c$ 时，两直线重合，不符合平行线定义。

另一方面，考察 l'： $ax+by=ax_P+y_P$，它垂直于 OM，所以与 l 平行或重合。容易验证，两线重合当且仅当 $ax_P+by_P=-c$。

因此，$ax_P+by_P=-c$（P 在直线上）时无平行线；其他情况下，恰有一条平行线：

$$ax+by=ax_P+y_P.$$

我们发现，与 $ax+by+c=0$ 平行的直线，总可以写成 $ax+by+c'=0$ 的形式。我们把 $ax+by$ 称为**直线的头**。头相同的直线相互平行或重合。

例题 1.3.1. 求经过点 $P(x_P,\ y_P)$ 并垂直于直线 l： $ax+by+c=0$ 的直线（如果存在）的方程。

解答. 把直线方程写成点法式：$(x-x_Q)a+(y-y_Q)b=0$，其中 $Q(x_Q,y_Q)$ 是 l 上一点。这说明 (a,b) 是 l 的法向量。因此，以 (a,b) 为方向的直线：$(x-x_P)b-(y-y_P)a=0$ 就是过 P 且垂直于 l 的直线。

我们发现，与 $ax+by+c=0$ 垂直的直线，总可以写成 $bx-ay+c'=0$ 的形式。

例题 1.3.2. 给定一点 $P(x_P,\ y_P)$ 和一条直线 l： $ax+by+c=0$，如何计算 P 到直线 l 的距离 d?

解答. 首先把直线方程转为等高式: $ax-(-b)y=-c$, 它表示与点 $M(a,-b)$ 的面积为 $-c$ 的直线。P 与点 M 的面积为 $ax_P-(-b)y_P$。如果 P 到 l 的垂足为 Q，P 到 OM 的垂足为 H，那么 P 到 l 的距离 $d=|PQ|$ 是 $|HQ|$ 与 $|PH|$ 的差。平行四边形的面积等于底乘以高，所以，考虑 P、Q 分别与 O,M 生成的平行四边形，它们的面积之差等于高的差乘以底。而两者对应的高分别是 $|HQ|$ 与 $|PH|$。也就是说，P 到 l 的距离 d 是两个面积的差除以底的商:

$$d=\frac{|-c-(ax_P-(-b)y_P)|}{\sqrt{a^2+(-b)^2}}=\frac{|ax_P+by_P+c|}{\sqrt{a^2+b^2}}.$$

另解: 首先把直线方程转为等垂式: $ax+by=-c$, 它表示与点 $M(a,b)$ 的内积为 $-c$ 的直线, 也就是垂直于 OM，垂足 Q 满足 $\overrightarrow{OQ}=\frac{-c}{|OM|^2}\overrightarrow{OM}$ 的直线。作 P 到 OM 的垂线 l', 记垂足为 H, 则 P 到 l 的距离就是 $|HQ|$, 也就是 $|OQ|$ 和 $|OH|$ 的差。l 与 l' 垂直于同一条直线 OM，因此平行或重合。于是 l' 的等垂式为 $ax+by=ax_P+by_P$，$\overrightarrow{OH}=\frac{T}{|OM|^2}\overrightarrow{OM}=\frac{ax_P+by_P}{|OM|^2}\overrightarrow{OM}$。于是 P 到 l 的距离为:

$$d=|\overrightarrow{HQ}|=|\overrightarrow{OQ}-\overrightarrow{OH}|=\frac{|-c-(ax_P+by_P)|}{|OM|^2}|OM|=\frac{|ax_P+by_P+c|}{\sqrt{a^2+b^2}}.$$

习题 1.3.1.

1. 根据已知条件，写出直线的方程:

1.1. 过点 $(1,\ -3)$，与 $(0.5,\ 2.1)$ 共轴。

1.2. 过点 $(2,\ -0.8)$、$(-2,\ 2.5)$。

1.3. 过点 $(-1,\ 1)$，与 $(-0.5,\ 1.5)$ 垂直。

1.4. 过点 $(-2.25,\ -6)$，斜率为 -1.7。

1.5. 与 $(4.5,\ -5)$ 内积为 -1.2。

1.6. 与 $(5.6,\ 1)$ 面积为 -8。

1.7. 与直线 $2x-y+4$ 平行，且过点 $(1,\ 1)$。

2. 求平行直线 $x-3y+1=0$ 和 $x-3y-4=0$ 的距离。

3. 两直线的方程分别为 $y=k_1x+b_1$、$y=k_2x+b_2$，证明：两直线垂直，当且仅当 $k_1k_2+1=0$。

思考 1.3.1.

1. 直线 l 按向量 P 平移得到直线 l'，l 和 l' 之间有什么关系?

2. 一次函数的直部与直线方程的头有什么关系?

1.4 圆的方程

圆是到一点距离相同的点的集合。用向量的语言来说，设有以 $W(x_W,\ y_W)$ 为圆心、以正数 r 为半径的圆，那么圆上的点 $P(x,\ y)$ 是方程:

$$|P-W|=r$$

的解集。直角坐标系中，方程可以写为:

$$\sqrt{(x-x_W)^2+(y-y_W)^2}=r$$

根号中的值总大于等于零，所以这个方程的解集就是方程

$$(x-x_W)^2+(y-y_W)^2=r^2$$

的解集。我们把这个方程称为圆的**径心式**方程，它的解集就是以 W 为圆心、r 为半径的圆。比如，

$$x^2+y^2=4$$

表示圆心为 $(0,\ 0)$、半径为 2 的圆。

我们还可以通过其他直观性质来得到圆的方程。

点心式：已知 $W(x_W,\ y_W)$ 为圆心，$A(x_A,\ y_A)$ 为圆上一点，那么半径为 $\sqrt{(x_A-x_W)^2+(y_A-y_W)^2}$。于是圆的方程为：

$$(x-x_W)^2+(y-y_W)^2=(x_A-x_W)^2+(y_A-y_W)^2.$$

我们把它称为圆的点心式方程。

直径式：已知 $A(x_A,\ y_A)$、$B(x_B,\ y_B)$ 连成的线段为圆的直径，那么对圆上任一点 P，$AP\perp BP$，因此两向量内积为 0。于是可得到方程：

$$(x-x_A)(x-x_B)+(y-y_A)(y-y_B)=0.$$

我们把这个方程称为圆的直径式方程。比如，以两点 $(0,\ 0)$、$(1,\ 2)$ 为直径的圆，方程为：$(x-0)(x-1)+(y-0)(y-2)=0$，即 $x^2+y^2-x-2y=0$。

例题 1.4.1. 求过点 $(1,\ 0)$、$(0,\ 2)$、$(-1,\ -1)$ 的圆的方程。

解答. 设圆的方程为

$$(x-x_W)^2+(y-y_W)^2=r^2$$

三点都在圆上，因此都满足圆的方程，于是我们得到方程组：

$$\begin{cases} (1-x_W)^2+(0-y_W)^2 &= r^2 \qquad (1)\\ (0-x_W)^2+(2-y_W)^2 &= r^2 \qquad (2)\\ (1+x_W)^2+(1+y_W)^2 &= r^2 \qquad (3)\end{cases}$$

分别计算 $(1)-(2)$、$(1)-(3)$，得到另一个方程组：

$$\begin{cases} 1-2x_W &= 4-4y_W\\ 1-2x_W &= 1+2x_W+1+2y_W\end{cases}$$

解这个方程组，得到

$$\begin{cases} x_W &= & -0.5 \\ y_W &= & 0.5 \end{cases}$$

代入 (1)，得到 $r^2 = 2.5$。于是圆的方程为:

$$(x+0.5)^2 + (y-0.5)^2 = 2.5$$

例题 1.4.2. 过点 $P(1,\ 2)$ 的直线与圆 $(x+1)^2+(y-1)^2 = 3$ 相切，求直线方程。

解答. 设直线 l 过点 P，且与圆 $(x+1)^2+(y-1)^2 = 3$ 相切，那么圆心 $W(-1,1)$ 到 l 的距离平方为 3。将直线 l 写成点斜式: $y-2 = k(x-1)$，其中斜率 k 是待定的未知数。将它写成等垂式:

$$kx + (-1)y = k-2.$$

可以算出，W 到它的距离为 $\frac{|-k-1-k+2|}{\sqrt{k^2+1}}$。于是可以列出方程:

$$\frac{(-k-1-k+2)^2}{k^2+1} = 3.$$

方程两边乘以 k^2+1，并将所有项整理到等式左边，得到:

$$k^2 - 4k - 2 = 0.$$

这是一个一元二次方程。解得斜率 $k = 2+\sqrt{6}$ 和 $k = 2-\sqrt{6}$，分别对应方程: $y+(-2+\sqrt{6})x-\sqrt{6} = 0$、$y+(-2-\sqrt{6})x+\sqrt{6} = 0$。经验证，$W$ 到这两条直线的距离都是 $\sqrt{3}$。因此，过点 $P(1,\ 2)$ 且与圆 $(x+1)^2+(y-1)^2 = 3$ 相切的直线有两条，方程为: $y+(-2+\sqrt{6})x-\sqrt{6} = 0$、$y+(-2-\sqrt{6})x+\sqrt{6} = 0$。

习题 1.4.1.

1. 写出以 $(-3,\ 2)$ 为圆心，半径为 5 的圆的方程。

2. 写出圆心为 $(-3,\ 2)$，过 $(1,\ 1.3)$ 的圆的方程。

3. 写出以 $(1,\ -1)$ 为圆心，过点 $(0,\ 4)$ 的圆的方程。

4. 直线过点 $(2,\ 5)$，且和点 $(0,\ 1)$ 的距离是 2.3，求直线的方程。
5. 直线 l 过点 $(4,\ 2)$，且和圆 $(x+1)^2+(y-1.5)^2=4$ 相切。求直线 l 的方程和对应切点的坐标。

思考 1.4.1.
1. 给定实数 θ，点 $(1+3\cos\theta,\ -2+3\sin\theta)$ 到 $(1,\ -2)$ 的距离是多少?
2. 如果点 $P(x_P,\ y_P)$ 到 $(1,\ -2)$ 的距离位 3，是否总存在实数 θ，使得 $x_P=1+3\cos\theta$、$y_P=-2+3\sin\theta$?

1.5 平面形的变换

我们已经学习过一些平面形的变换。用向量的语言，我们可以把这些变换做一次总结。

首先，定义这样一类映射。它们把平面的点（向量）映射到平面的点。我们称之为平面的**点映射**。

平面形的变换，就可以看作点映射的效果。

具体来说，平面形作为平面的子集，经过点映射，变成平面的另一个子集，这就是平面形的变换。因此，我们也把平面的点映射称为平面上的**变换**。

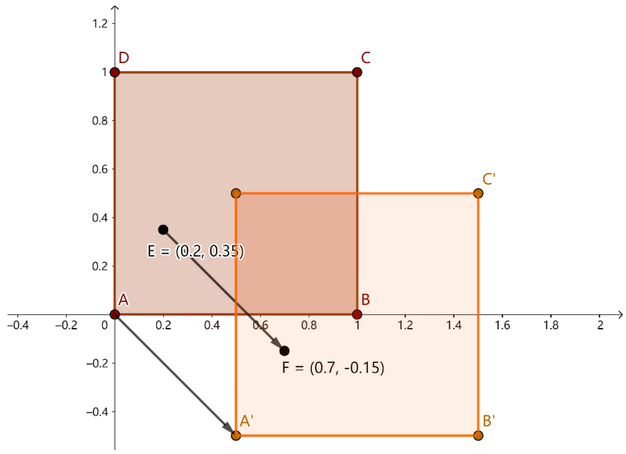
举例来说，设有平面 γ，那么以下映射就是一个点映射：

$$\begin{aligned} \gamma &\to \gamma \\ P &\mapsto P+U \end{aligned}$$

它将平面上每一点 P 按向量 U（或者记作 $\mathbf{u}$）平移。

设 S 是以 $(0,\ 0)$、$(1,\ 0)$、$(1,\ 1)$、$(0,\ 1)$ 为顶点的正方形（下面称为标准单位正方形）。S 可以写作：

$$S=\{(x,\ y)\,|\,0\leqslant x,y\leqslant 1\}$$



设 $U=(0.5,\ -0.5)$，那么点映射：

$$P \mapsto P+U$$

就把 S 映射成以 $(0.5,\ -0.5)$、$(1.5,\ -0.5)$、$(1.5,\ 0.5)$、$(0.5,\ 0.5)$ 为顶点的正方形。我们知道，这样的点映射叫做**平移变换**。

从例子中可以看到，按向量 U 的平移变换可以表示为：

$$\begin{aligned}\mathrm{Py}_U:\ \gamma\ &\to\ \gamma\\ P &\mapsto\ P+U\end{aligned}$$

平移变换不改变图形的大小和形状，只改变图形的位置。大部分的变换对图形的大小、形状和位置都有改变。比如给定一个长度为 1 的向量 U，考虑以下变换：

$$T:\ \ (x,\ y) \mapsto (x,\ 0)$$

对平面中一点 $P(x_P,\ y_P)$，T 把它映射到 $(x_P,\ 0)$，也就是它到 x 轴的垂足。直观上，如果把 x 轴看作“地面”，想象光线从正上方照射，那么 T 把点映射到它在“地面”上的投影。我们把这样的变换称为投影变换。

更一般地来说，可以这样定义投影变换：

定义 1.5.1. 给定平面基底 $\mathbf{u},\mathbf{v}$，平面中任意向量 P 都能唯一表示成 $\mathbf{u},\mathbf{v}$ 的直组合：

$$P = a\mathbf{u} + b\mathbf{v}.$$

其中 a,b 是相应的分量。那么，把 P 映射到 $a\mathbf{u}$ 的点变换

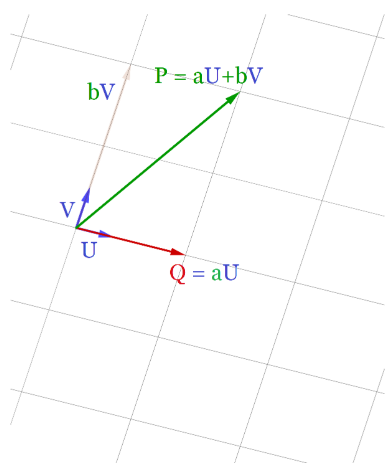
$$\begin{aligned} 投_{\mathbf{u},\mathbf{v}}:\ \gamma &\to \gamma \\ P &\mapsto a\mathbf{u} \end{aligned}$$

就称为沿 $\mathbf{v}$ 到 $\mathbf{u}$ 的**投影变换**。它把任何向量映射到直线 $\mathbb{R}\mathbf{u}$ 上。$\mathbf{u}$ 称为 投$_{\mathbf{u},\mathbf{v}}$ 的**方向向量**，$\mathbf{v}$ 称为**投影向量**。

直角坐标系中，记基向量为 $\mathbf{e}_x,\mathbf{e}_y$。上面的例子中，$T$ 就是沿 $\mathbf{e}_y$ 到 $\mathbf{e}_x$ 的投影变换。

设 S 是标准单位正方形，S 中任一点 $(x,\ y)$ 变换后为 $(x,\ \ 0)$。因此，S 变成从原点到 $(1,\ 0)$ 的线段。这就是标准单位正方形在 $(1,\ 0)$ 方向的投影。

T 的投影向量垂直于方向向量，就好像正午的阳光一样。我们把这样的投影变换称为**正投影**。



沿 V 到 U 的投影变换

以上定义中，没有说明怎样得到 a,b。如果平面上定义了内积，我们可以用内积求出 a,b 的值。已知 P 在基底 $\mathbf{u},\mathbf{v}$ 下的分解：

$$P = a\mathbf{u} + b\mathbf{v}.$$

等式两边与 $\mathbf{u},\mathbf{v}$ 做内积，得到：

$$\begin{cases} (\mathbf{u}\,|\,P) &= a(\mathbf{u}\,|\,\mathbf{u}) + b(\mathbf{u}\,|\,\mathbf{v}) \\ (P\,|\,\mathbf{v}) &= a(\mathbf{u}\,|\,\mathbf{v}) + b(\mathbf{v}\,|\,\mathbf{v}) \end{cases}$$

解这个关于 a, b 的二元一次方程组，得到:

$$\begin{cases} a &= \frac{(\mathbf{v}\,|\,\mathbf{v})(\mathbf{u}\,|\,P)-(\mathbf{u}\,|\,\mathbf{v})(\mathbf{P}\,|\,\mathbf{v})}{(\mathbf{u}\,|\,\mathbf{u})(\mathbf{v}\,|\,\mathbf{v})-(\mathbf{u}\,|\,\mathbf{v})^2} \\ b &= \frac{(\mathbf{u}\,|\,\mathbf{v})(\mathbf{u}\,|\,P)-(\mathbf{u}\,|\,\mathbf{u})(\mathbf{P}\,|\,\mathbf{v})}{(\mathbf{u}\,|\,\mathbf{v})^2-(\mathbf{u}\,|\,\mathbf{u})(\mathbf{v}\,|\,\mathbf{v})} \end{cases}$$

也就是说，投影变换可以写为:

$$\text{投}_{\mathbf{u},\mathbf{v}}:\quad P \mapsto \frac{(\mathbf{v}\,|\,\mathbf{v})(\mathbf{u}\,|\,P)-(\mathbf{u}\,|\,\mathbf{v})(\mathbf{P}\,|\,\mathbf{v})}{(\mathbf{u}\,|\,\mathbf{u})(\mathbf{v}\,|\,\mathbf{v})-(\mathbf{u}\,|\,\mathbf{v})^2}\mathbf{u}.$$

如果 $\mathbf{u}, \mathbf{v}$ 是标准直角坐标系的基底，那么

$$\begin{aligned} (\mathbf{u}\,|\,\mathbf{v}) &= 0 \\ (\mathbf{u}\,|\,\mathbf{u}) = (\mathbf{v}\,|\,\mathbf{v}) &= 1 \end{aligned}$$

于是

$$\begin{cases} a &= (\mathbf{u}\,|\,P) \\ b &= (P\,|\,\mathbf{v}) \end{cases}$$

投影变换写为:

$$\text{投}_{\mathbf{u}}:\quad P \mapsto (\mathbf{u}\,|\,P)\mathbf{u}.$$

同样考虑正投影。如果方向向量换成 $\mathbf{u} = \left(\frac{1}{2},\ \frac{\sqrt{3}}{2}\right)$，投影向量换成 $\mathbf{v} = \left(-\frac{\sqrt{3}}{2},\ \frac{1}{2}\right)$。那么两者构成标准直角坐标系的基底。平面中的点 $P(x,\ y)$ 经过投影变换，就变为

$$\begin{aligned} (\mathbf{u}\,|\,P)\mathbf{u} &= \left(\frac{1}{2}x + \frac{\sqrt{3}}{2}y\right)\mathbf{u} \\ &= \left(\frac{x+\sqrt{3}y}{4},\ \frac{\sqrt{3}x+3y}{4}\right) \end{aligned}$$

也就是说，S 变为从原点到 $\left(\frac{\sqrt{3}+1}{4},\ \frac{3+\sqrt{3}}{4}\right)$ 的线段。这就是标准单位正方形在 $\left(\frac{1}{2},\ \frac{\sqrt{3}}{2}\right)$ 方向的正投影。

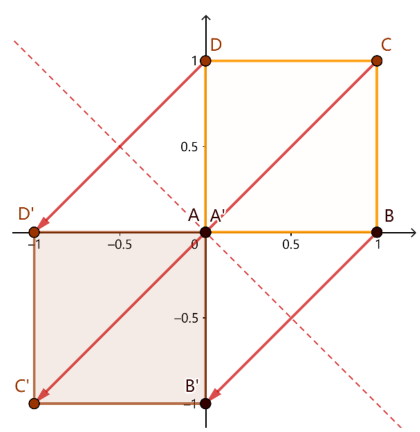
我们学习过点和图形的轴对称。把图形变成它的轴对称图形，这样的变换叫做**反射变换**。

投影变换和反射变换有密切关系。给定长度为 1 的方向向量 $\mathbf{u}$[①]，平面中一点 P 以及它关于 $\mathbb{R}\mathbf{u}$ 的对称点 Q，到 $\mathbf{u}$ 的正投影是重合的。设 P 到 $\mathbf{u}$ 的正投影为点 H，则：

$$\overrightarrow{OH} = \text{投}_{\mathbf{u}}(P) = (\mathbf{u}\,|\,P)\mathbf{u},$$

H 是 P 到 $\mathbb{R}\mathbf{u}$ 的垂足，而 $\overrightarrow{HP}$ 和 $\overrightarrow{HQ}$ 大小相等、方向相反。因此：

$$\begin{aligned}\overrightarrow{OQ} &= \overrightarrow{OH} + \overrightarrow{HQ}\\ &= \overrightarrow{OH} - \overrightarrow{HP}\\ &= \overrightarrow{OH} + \overrightarrow{OH} - \overrightarrow{OP}\\ &= (\mathbf{u}\,|\,P)\mathbf{u} + (\mathbf{u}\,|\,P)\mathbf{u} - P\\ &= 2(\mathbf{u}\,|\,P)\mathbf{u} - P\end{aligned}$$



也就是说，给定单位方向向量 $\mathbf{u}$，关于直线 $\mathbb{R}\mathbf{u}$ 的反射变换可以表示为：

$$\begin{aligned}\text{反}_{\mathbf{u}}:\ \gamma &\to \gamma\\ P &\mapsto 2(\mathbf{u}\,|\,P)\mathbf{u} - P\end{aligned}$$

其中 $\mathbf{u}$ 是对称轴的单位方向向量。

举例来说，设方向向量为 $\mathbf{u} = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$，标准单位正方形经过关于

[①]也叫做**单位方向向量**。

$\mathbb{R}\mathbf{u}$ 的轴对称变换，点 $P(x,\ y)$ 变为

$$\begin{aligned}2(\mathbf{u}\,|\,P)\mathbf{u}-P&=2\cdot\left(-\frac{\sqrt{2}}{2}x+\frac{\sqrt{2}}{2}y\right)\mathbf{u}-P\\&=(x-y,\ y-x)-(x,\ y)\\&=(-y,\ -x)\end{aligned}$$

即它关于直线 $x+y=0$ 的对称点。于是 $反_{\mathbf{u}}$ 将标准单位正方形变换为它关于 $x+y=0$ 的对称图形，即以 $(0,\ 0)$、$(0,\ -1)$、$(-1,\ -1)$、$(-1,\ 0)$ 为顶点的正方形。

如果对称轴不过原点，可以通过平移变换将它移到过原点的位置。比如，设对称轴为 $\mathbb{R}\mathbf{u}+\mathbf{v}$，点 P 经过反射变换变为 Q，则 $P-\mathbf{v}$ 和 $Q-\mathbf{v}$ 关于 $\mathbb{R}\mathbf{u}$ 对称。因此：

$$Q-\mathbf{v}=2(\mathbf{u}\,|\,P-\mathbf{v})\mathbf{u}-P+\mathbf{v}.$$

也就是说，一般的反射变换可以写成：

$$\begin{aligned}反_{\mathbf{u},\mathbf{v}}:\ \gamma\ &\to\ \gamma\\P\ &\mapsto\ 2(\mathbf{u}\,|\,P-\mathbf{v})\mathbf{u}-P+2\mathbf{v}.\end{aligned}$$

更一般地来说，可以这样定义反射变换：

定义 1.5.2. 给定平面基底 $\mathbf{u},\mathbf{v}$，平面中任意向量 P 都能唯一表示成 $\mathbf{u},\mathbf{v}$ 的直组合：

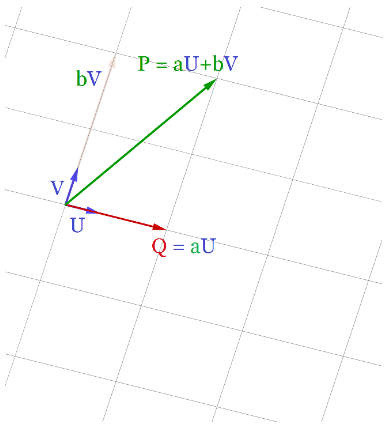
$$P=a\mathbf{u}+b\mathbf{v}.$$

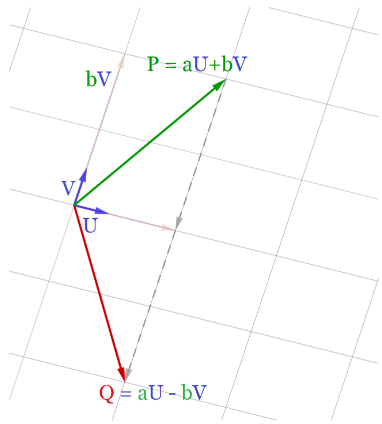
其中 a,b 是相应的分量。那么，把 P 映射到 $a\mathbf{u}$ 的点变换

$$\begin{aligned}投_{\mathbf{u}}:\ \gamma\ &\to\ \gamma\\P\ &\mapsto\ a\mathbf{u}-b\mathbf{v}\end{aligned}$$

就称为沿 $\mathbf{v}$ 到 $\mathbf{u}$ 的**反射变换**。

直观上，沿 $\mathbf{v}$ 到 $\mathbf{u}$ 的反射变换和投影变换分别是：





投影变换（左）和反射变换（右）的关系

虽然这个定义已经脱离了轴对称，直观上看，仍然与投影变换有密切联系。如果 $\mathbf{u}, \mathbf{v}$ 是标准直角坐标系的基底（对应正投影），以上定义的反射变换就是关于坐标轴的轴对称变换。

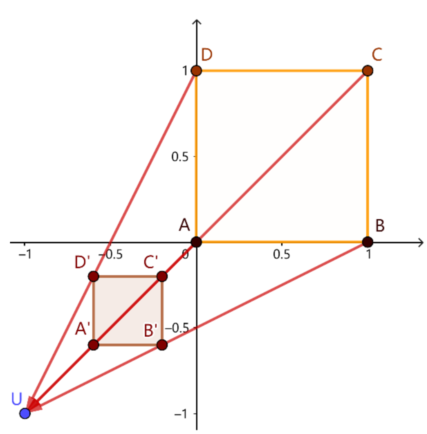
另一种简单的变换是**位似变换**。**位似**是一种特殊的相似关系。举例来说，设 $\triangle ABC$ 相似于 $\triangle A'B'C$。如果直线 AA'、BB'、CC' 交于一点 U，就说两个三角形位似，U 是**位似中心**。

位似变换可以表示为：

$$\begin{aligned} \mathrm{Ws}_{U,k}: \gamma &\rightarrow \gamma \\ P &\mapsto kP + (1-k)U \end{aligned}$$

其中 U 是位似中心，k 是变换前后的相似比。我们说平面形以 U 为中心放缩，变成原来 k 倍大小。

举例来说。设 S 是标准单位正方形，四个顶点分别是 $A(0,\ 0)$、$B(1,\ 0)$、$C(1,\ 1)$、$D(0,\ 1)$。指定位似中心 $U(-1,\ -1)$，相似比 $k = 0.4$。于是位似变换中，A 变为



$A'(-0.6,\ -0.6)$，B 变为 $B'(-0.2,\ -0.6)$。线段 AB 上一点 $(x,\ 0)$ 变为 $(0.4x - 0.6,\ -0.6)$。于是线段 AB 变成线段 $A'B'$。同理线段 BC、CD、DA 变成线段 $B'C'$、$C'D'$、$D'A'$，其中 $C' = (-0.2,\ -0.2)$，$D' = (-0.6,\ -0.2)$。标准单位正方形 S 变成边长为 0.4 的正方形 $A'B'C'D'$。

最后来看**旋转变换**。旋转变换指的是以给定角的顶点为中心，把某个图形按该角旋转。

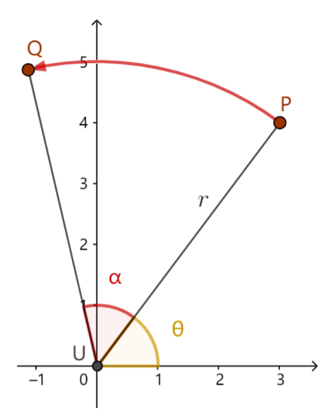
首先来看一点的旋转。给定角 α，顶点为 U。平面中一点 P 关于 α 旋转的结果，是唯一使得 $\angle PUQ = \alpha$ 且 $|UP| = |UQ|$ 的点 Q[①]。以 U 为圆心，以 $|UP|$ 为半径作圆，则 P 沿圆弧运动到 Q。顶点 U 称为**旋转中心**，α 称为**旋转角**。

最简单的情况是 U 为原点。这时，设点 P 的坐标为 $(x_P,\ y_P)$，$|UP| = r = \sqrt{x_P^2 + y_P^2}$。取 x 轴上的点 $S(1,\ 0)$ 为基准，设 $\angle SUP = \theta$，容易验证，$|x_P| = r|\cos\theta|$，$|y_P| = r|\sin\theta|$。对比各个象限中 x_P、y_P 和 $\cos\theta$、$\sin\theta$ 的正负号，可以发现：任何情况下，x_P 总与 $\cos\theta$ 同正负，y_P 总与 $\sin\theta$ 同正负。

定理 1.5.1. 设 $P(x_P,\ y_P)$ 是平面上原点外一点，设以 x 轴正半轴为始边、射线 OP 为终边的角为 θ，则

$$x_P = r\cos\theta, \quad y_P = r\sin\theta.$$

其中 $r = \sqrt{x_P^2 + y_P^2}$。



这个定理将点的坐标和角度联系起来，是我们讨论平面角度和旋转问题的关键。

如果将 P 按角 α 旋转，那么与 x 轴正半轴张成的角应该从 θ 变为 $\theta + \alpha$。它的长度仍

[①]如果 $P = U$，则旋转后仍为 P。

然是 r，而角度是 $\theta+\alpha$，因此，它的坐标是
$(r\cos(\theta+\alpha),\ r\sin(\theta+\alpha))$。使用和角公式，可
以将它转化为:

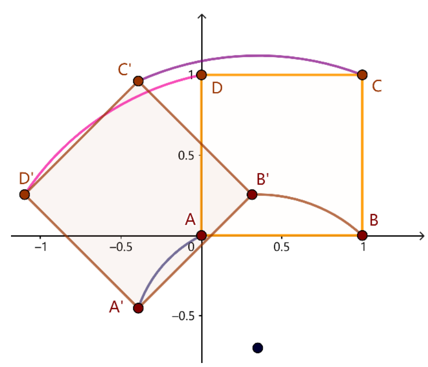
$$(r\cos\theta\cos(\alpha)-r\sin\theta\sin(\alpha),\ r\cos\theta\sin(\alpha)+r\sin\theta\cos(\alpha)).$$

用 x,y 表示，旋转变换就是:

$$(x,\ y)\mapsto\ (x\cos(\alpha)-y\sin(\alpha),\ x\sin(\alpha)+y\cos(\alpha)).$$

如果 U 不是原点，坐标是 $(x_U,\ y_U)$，那么 $\overrightarrow{UP}=(x_P-x_U,\ y_P-y_U)$。$\overrightarrow{UP}$ 旋转变为 $\overrightarrow{UQ}$ 的运动和前面类似。因此通过 $Q=U+\overrightarrow{UQ}$ 就可得到 Q。于是旋转变换可以表示为:

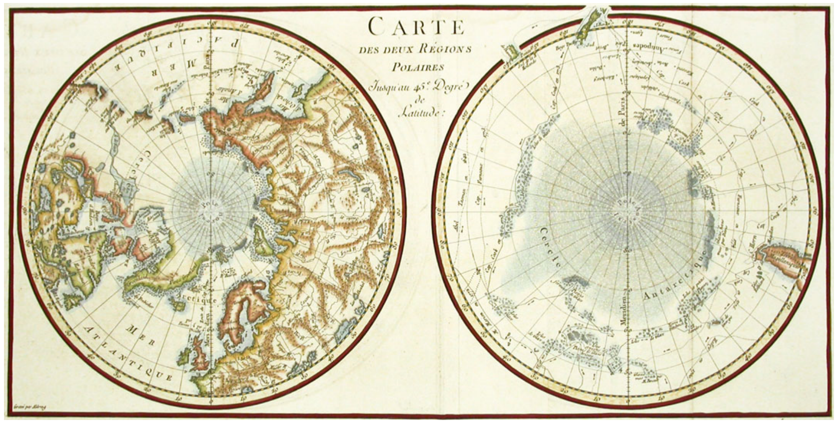
$$\begin{aligned}\text{旋}_{(U,\,\alpha)}:\ \gamma\ &\to\ \gamma\\(x,\ y)&\mapsto\ (x_U,\ y_U)+\big((x-x_U)\cos(\alpha)-(y-y_U)\sin(\alpha),\\&\qquad (x-x_U)\sin(\alpha)+(y-y_U)\cos(\alpha)\big).\end{aligned}$$



可以看到，旋转变换的表示形式繁琐。这是向量和直角坐标系的性质决定的。

直角坐标系与向量的共轴性质和内积相关。共轴性质、内积的基本性质都只和向量的加法、数乘运算相关。而向量定义中的加法和数乘运算，分别对应平移和放缩变换。向量定义中并没有与旋转对应的运算，共轴性、内积的性质也不包括旋转，因此，要通过向量在直角坐标系下的坐标表示旋转，自然不是简单的事情。

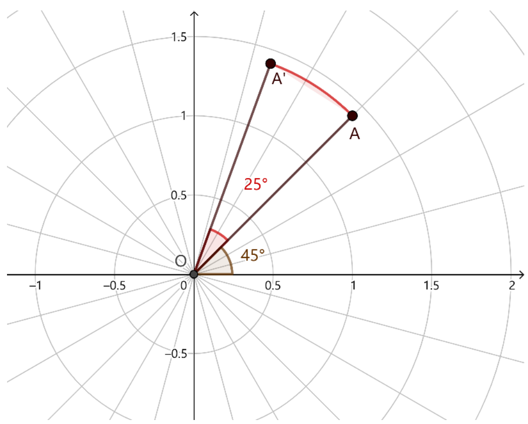
一种折中的方法是使用**极坐标系**。极坐标系是科学家研究南北极地时发明的。在南北极圈中，各条经线汇于极点。人们发现，使用到极点的距离和经度描述位置，比使用经纬度更精确、更方便。



使用极坐标系的极地地图

如下图所示，极坐标系由一条射线构成。这条射线上配备了单位向量，形成数轴的正半轴。为了方便起见，我们同样称它为 x 轴。一点 A 在极坐标系中的坐标称为极坐标，它由**模长**和**辐角**组成。模长就是点 A 到原点 O 的距离，辐角是 x 轴正半轴到 $\overrightarrow{OA}$ 的角。

比如，直角坐标系中的点 $(1,\ 1)$，在极坐标系中可以表示为 $\left(\sqrt{2},\ 45^\circ\right)$。



使用极坐标系，可以比较方便地描述以原点为中心的旋转运动。比如，直角坐标系中，点 $(1,\ 1)$ 以原点为中心逆时针旋转 25°，可以用极坐标系描述为：

$$\left(\sqrt{2},\ 45^\circ\right) \mapsto \left(\sqrt{2},\ 70^\circ\right)$$

然而，用极坐标系描述平移、反射等变换就比较复杂。用极坐标描述中心非原点的旋转变换，仍然繁琐。

思考 1.5.1.

1. 如何描述图形关于某点的中心对称变换？

2. 如何描述一般的相似变换？

3. 给定向量 $\mathbf{u}, \mathbf{v}$，能否用 $反_{\mathbf{u}}$ 和 $\mathrm{Ty}_{\mathbf{u}}$ 表示 $反_{\mathbf{u},\mathbf{v}}$？

4. 如果平面中任两点之间的距离经过某个变换后都不变，就说这个变换是**保距变换**。本节提到的变换中，哪些变换是保距变换？

5. 如果平面中任意角经过某个变换后角度都不变，就说这个变换是**保角变换**。本节提到的变换中，哪些变换是保角变换？

习题 1.5.1.

1. 将标准单位正方形以原点为中心，按 45° 顺时针旋转，最终得到的图形是什么？

2. 设点 A、B、C 的坐标分别是 $(0,\ 4)$、$(-1,\ 0)$、$(1,\ 1)$。

2.1. 以原点为中心，把 $\triangle ABC$ 按 90° 顺时针旋转，然后以原点为中心，将其放大为原来的 3 倍，再按 $(-5,\ 0)$ 平移。最终得到的图形是什么？

2.2. 把 $\triangle ABC$ 按 $\mathbb{R}\left(\frac{1}{2},\ \frac{\sqrt{3}}{2}\right)$ 作反射变换，再按 $(0,\ 6)$ 平移。最终得到的图形是什么？

2.3. 以原点为中心，把 $\triangle ABC$ 分别按 30°、120° 顺时针旋转，然后投影到 $\left(\frac{2\sqrt{5}}{5},\ -\frac{\sqrt{5}}{5}\right)$ 方向上。最终分别得到什么图形？

3. 某工厂的机械臂要完成一个平面动作，可以简单描述为把端点为 $A(1,\ 2)$、$B(-2,\ 1)$ 的线段，变换为端点为 $A'(-4,\ 4)$、$B'(-4,\ -6)$ 的线段。机械臂可以进行 3 种操作：以原点为中心的旋转变换、按任一向量的平移变换、以线段一端为中心的位似变换。

3.1. 请设计一种操作组合，把 AB 移动到 $A'B'$。

3.2. 请设计一种操作组合，把 AB 移动到 $A'B'$，并确保 AB 的中点经过点 $(-2,\ -2)$。

3.3. 已知机械臂的每种操作成本。旋转变换的成本与线段中点经过的距离成正比，每单位长度 0.65 元；平移变换的成本和平移距离成正比，每

单位长度 0.4 元；位似变换的成本和变换前后线段长度的差值成正比，每单位长度 0.8 元。请设计一种操作方式，使成本最低。

3.4. 同上，如果要确保 AB 的中点经过点 $(-2,\ -2)$，那么怎样操作成本最低?

第二章　平面中的运动

在前面的学习中，我们已经接触过运动的概念。比如，我们说点运动形成了线。这一章我们来初步讨论各种和运动相关的问题。

2.1　轨迹

我们知道，点运动形成线，而直线和线段是最简单的“线”。因此，我们从直线和线段开始定义运动。

给定数轴上的线段 XY 以及线段上的点 P，定义线段上的运动为一种二元关系（两个元素之间的关系）：$A \triangleright B$，表示点 P 运动时经过的先后，称为**运动的先后关系**。如果点 P 运动时先经过点 A 再经过点 B，就记为 $A \triangleright B$。P 运动时经过的点的集合，称为它的**轨迹**。运动的先后关系 $\triangleright$ 有以下性质：

- $A \triangleright B$ 和 $B \triangleright A$ 至少有一个成立。
- 如果 $A \triangleright A$，那么必有 $B \neq A$，使得 $A \triangleright B$ 且 $B \triangleright A$。
- 如果 $A \triangleright B$、$B \triangleright C$，则 $A \triangleright C$。

如果轨迹中某点 S 使得对轨迹中任何点 A 都有 $S \triangleright A$，就说 S 是**运动的起点**。如果轨迹中某点 T 使得对轨迹中任何点 A 都有 $A \triangleright T$，就说 T 是**运动的终点**。

如果用数轴上的数来代表这些点，那么不难验证，常见的“小于”（$<$）和“大于”（$>$）关系都满足这些要求。当 $\triangleright$ 是 $<$ 时，我们把这个运动称为“从左到右沿 XY 运动”；当 $\triangleright$ 是 $>$ 时，我们把这个运动称为“从右到左沿 XY 运动”。

比如，令 $X=(0,\ 0)$、$Y=(1,\ 0)$，那么从左到右沿 XY 运动，就是实数 a 从 0 不断增大到 1 时，点 $(a,\ 0)$ 的轨迹。如果实数 $a<b$，点 P 就先经过 $(a,\ 0)$，后经过 $(b,\ 0)$。

从右往左沿 XY 运动则恰好相反，是实数 a 从 1 不断减小到 0 时，点 $(a,\ 0)$ 的轨迹。如果实数 $a>b$，点 P 就先经过 $(a,\ 0)$，后经过 $(b,\ 0)$。

这样定义的运动，不仅可以定义在线段上，也可以定义在直线上。比如，考虑数轴所在直线 l，那么“从左到右沿 l 运动”就对应“小于”关系，“从右到左沿 l 运动”就对应“大于”关系。直线上定义的运动和线段上定义的运动，区别在于前者没有起点和终点，后者有起点和终点。

对一般的运动来说，要说明先后关系 $\triangleright$ 并不容易。我们采用映射的方法来描述。如果某运动有起点 S 和终点 T，轨迹为 $\mathcal{C}$，我们可以把它和沿数轴上的线段 $[0;1]$ 从左到右的运动对应起来。定义映射 f:

$$
\begin{aligned}
f:[0;1] &\to \mathbb{R}^2\\
t &\mapsto f(t)
\end{aligned}
$$

其中 t 表示沿 $[0;1]$ 从左到右的运动中经过的点，而 $f(t)$ 就是它对应在该运动中的点。$f(0)$ 就是起点 S，$f(1)$ 是终点 T。$\mathcal{C}=\{f(t)\,|\,t\in[0;1]\}$，一般记作 $f([0;1])$。

举例来说，给定起点 $(-2,\ 1)$ 和终点 $(1,\ 3)$，如果我们要描述一个点沿以它们为端点的线段运动，则可以用以下映射表示：

$$
\begin{aligned}
g:[0;1] &\to \mathbb{R}^2\\
t &\mapsto (-2+3t,\ 1+2t)
\end{aligned}
$$

可以验证，$f(0) = (-2,\ 1)$，$f(1) = (1,\ 3)$，且 $[0;1]$ 中的数对应的点都在线段上。

如果运动没有起点和终点，那么可以把它和沿数轴直线左到右的运动对应起来：

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R}^2 \\ t &\mapsto f(t) \end{aligned}$$

要注意的是：运动的起点和终点相同，轨迹并不一定相同。比如，t 沿 $[0;1]$ 从左到右运动时，运动

$$h:\ t\ \mapsto (-2 + 2t + t^2,\ 1 + 2t)$$

的起点和终点也分别是 $(-2,\ 1)$ 和 $(1,\ 3)$，但 g 和 h 对应的轨迹并不一样。$h(0.5) = (0.25,\ 2)$，而 g 不经过 $(0.25,\ 2)$。

思考 2.1.1.

1. 运动的先后关系满足的性质中，第二条应该如何理解？能否去掉？
2. 如何定义只有起点没有终点的运动？只有终点没有起点的运动呢？
3. 比较两个运动：$\forall t \geqslant 0$，$f_1:\ t\ \mapsto (1 - \frac{1}{t+1},\ 0)$，和 $\forall t \in [0;1)$，$f_2:\ t\ \mapsto (t,\ 0)$。这两个运动有什么相同和不同之处？

习题 2.1.1.

1. 给定点 $A(1,\ 0)$、$B(-1,\ 2)$，描述沿线段 AB 的运动。
2. 描述一个从 $A(1,\ 0)$ 到 $B(-1,\ 2)$，并经过 $(0,\ 0)$ 的运动。
3. 描述一个起点和终点都是 $(0,\ 1)$，且经过 $(-1,\ -2)$ 的运动。
4. 描述一个没有起点和终点，经过 $(2,\ 1)$ 和 $(1,\ 4)$ 的运动。
5. t 沿 $[0;1]$ 从左到右运动时，$f:\ t\ \mapsto (t^2 - t + 1,\ t + 1)$ 是什么运动？

2.2 变与不变

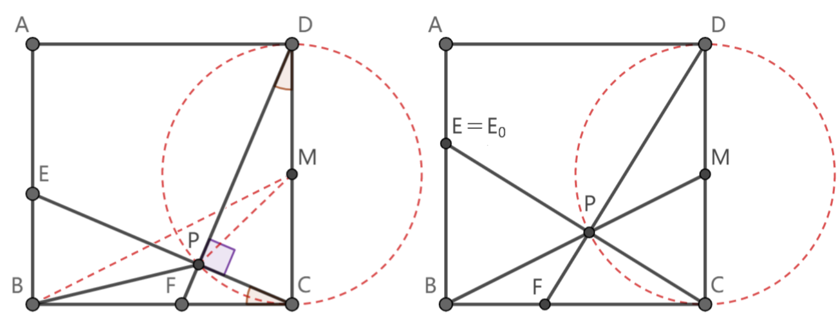
运动是对变化的描述。以下运动中，哪些量变了，哪些没变？

- $t \mapsto (3t+1,\ 4)$，$t \in [0;1]$。
- $t \mapsto (\cos t,\ \sin t)$，$t \in [0;180°]$。
- $t \mapsto (1-2t,\ 2-t)$，$t \in [0;180°]$。

第一个运动中，随着 t 变化，横坐标在变，纵坐标不变。第二个运动中，随着 t 变化，纵横坐标在变，但两者的平方和总是 1，即点到 $(0,\ 0)$ 的距离不变。第三个运动中，随着 t 变化，纵横坐标在变，但纵坐标的两倍与横坐标的差总是 3，即“纵坐标的两倍与横坐标的差”不变。

运动中保持不变的量，称为运动的**不变量**。在变化中保持不变的量，很可能对我们理解事物的本质有帮助。

例题 2.2.1. 如图，正方形 $ABCD$ 的边长为 8。点 E 沿边 AB 从 A 往 B 运动。在 BC 有一点 F，使得 $|AE| = |BF|$。CE、DF 交于点 P。求 $|BP|$ 的最小值。

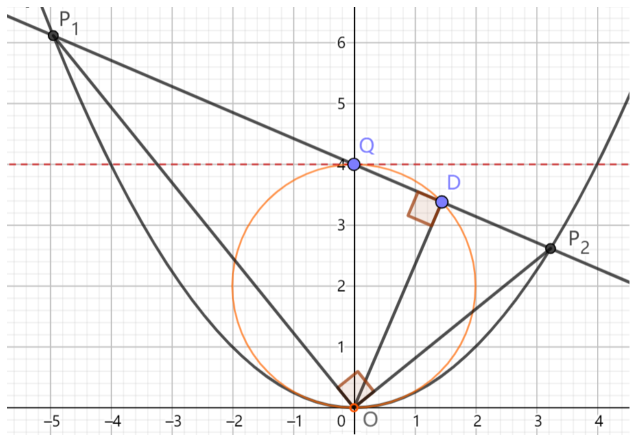


解答. 任何时刻，总有 $\triangle EBC \simeq \triangle FCD$。因此 $\angle PCB = \angle PDC$，$\angle CPD$ 为直角。这说明 P 总在以 CD 为直径的圆上，即 E 运动时，P 到 CD 中点 M 的距离总为 4。也就是说，$\triangle BPM$ 中，$|PM| = 4$、$|BM| = 4\sqrt{5}$ 都是定值。于是

$$|BP| \geqslant |BM| - |PM| = 4\sqrt{5} - 4.$$

这个最小值是可以达到的。作以 CD 为直径的圆，和线段 BM 交于点 Q。延长 CQ 与 AB 交于 E_0，则 E 运动到 E_0 时，$P = Q$，这时 B, P, M 三点共轴，因此 $|BP| = |BM| - |PM|$，达到最小值。

例题 2.2.2. 如图，点 P_1 在抛物线 $y=\frac{x^2}{4}$ 上从左到右运动。直线 OP_2 垂直于 OP_1，并且和抛物线交于点 P_2。过原点 O 作 P_1P_2 的垂线，垂足为 D。当 P_1 运动时，求点 D 的轨迹。



解答. 选几个不同的 P_1 点，作对应的直线 P_1P_2。通过观察，可以发现这些点和 y 轴似乎交于同一点。猜测这是一个不变量。给定 P_1，连接 P_1P_2，交 y 轴于点 Q。设直线 OP_1 的斜率为 a，则它的方程为

$$y=ax.$$

和抛物线方程联立求解，得到 P_1 坐标为 $(a,\ \frac{a^2}{4})$。设直线 OP_2 斜率为 b，则它的方程为

$$y=bx.$$

同理解得 P_2 坐标为 $(b,\ \frac{b^2}{4})$。于是 P_1P_2 的两点式方程为:

$$(x-a)\left(\frac{b^2}{4}-\frac{a^2}{4}\right)-\left(y-\frac{a^2}{4}\right)(b-a)=0$$

令 $x=0$，就得到直线 P_1P_2 和 y 轴交点的纵坐标: $y_Q=-\frac{ab}{4}$。由于 $OP_1\perp OP_2$，两者的方向向量内积为 0。分别选方向向量 $(1,\ \frac{a}{4})$、$(1,\ \frac{b}{4})$，就得到 $1+\frac{ab}{16}=0$，即 $ab=-16$。代入 Q 的纵坐标，就得到 $y_Q=4$。即 P_1P_2 与 y 轴的交点总是 $Q(0,\ 4)$。

过 O 作 P_1P_2 的垂线，垂足为 D，那么 $\triangle ODQ$ 是直角三角形，$\angle ODQ$ 是直角。因此 D 总在以 OQ 为直径的圆：$x^2 + y(y-4) = 0$ 上。

另一方面，给定以 OQ 为直径的圆上一点 $D \neq Q$，连接 DQ，交抛物线于两点：P_1, P_2。DQ 的两点式方程为：

$$(x - x_D)(4 - y_D) - (y - y_D)(0 - x_D) = 0.$$

将 $y = \frac{x^2}{4}$ 代入，得到关于 x 的一元二次方程：

$$\frac{x_D}{4}x^2 + (4 - y_D)\,x - 4x_D = 0.$$

只要 $x_D \neq 0$，二次方程的判别式：

$$(4 - y_D)^2 - 4 \cdot \frac{x_D}{4} \cdot (-4x_D) = (4 - y_D)^2 + 4x_D^2 > 0.$$

于是方程总有两个解。它的两个解就是 P_1, P_2 的横坐标 a, b。因此根据根与系数的关系，P_1, P_2 的横坐标乘积为

$$ab = -4x_D \div \frac{x_D}{4} = -16.$$

而 P_1, P_2 的纵坐标分别是 $\frac{a^2}{4}$、$\frac{b^2}{4}$，于是两直线方向向量的内积为：

$$ab + \frac{a^2}{4} \cdot \frac{b^2}{4} = ab(1 + \frac{ab}{16}) = 0.$$

即 $OP_1 \perp OP_2$。

如果 $D = Q$，则 OD 垂直于抛物线上两点：$P_1(-4,\ 4)$ 和 $P_2(4,\ 4)$ 的连线。可以验证 $OP_1 \perp OP_2$。于是点 Q 也在轨迹上。

综上可知，一方面 D 的轨迹总在以 OQ 为直径的圆上。另一方面，除了原点 O，以 OQ 为直径的圆都在 D 的轨迹上。因此，D 的轨迹是圆：

$$\{(x,\ y)\,|\,x^2 + y(y-4) = 0,\ (x,\ y) \neq (0,\ 0)\}.$$

习题 2.2.1.

1. 设 c 为正实数，k 是大于 1 的实数。考虑平面上的点 $C_+ = (c,\ 0)$ 和直线 $l_+ : x = k^2c$。考虑集合：

$$T = \left\{ P \,\middle|\, \frac{P\text{到}l_+\text{的距离}}{|PC_+|} = k \right\}.$$

1.1. 证明：T 包含点 $(ck,\ 0)$、$(-ck,\ 0)$。

1.2. 记 $C_- = (-c,\ 0)$。$P = (x,\ y)$ 是 T 中一点。证明：

$$|PC_-|^2 - |PC_+|^2 = 4cx.$$

1.3. 证明：P 到直线 $l_- : x = -k^2c$ 的距离是 P 到 C_- 距离的 k 倍。

1.4. 考虑集合：

$$S = \{P \,|\, |PC_+| + |PC_-| = 2ck\}.$$

证明：$T = S$。

2.3 运动的速度

生活中，我们常用快慢来描述运动。比如，从商店回家，骑自行车比走路更快；从某城市到另一个城市，坐飞机比开汽车更快。等等。我们用长度来刻画事物的长短。对于运动的快慢，我们也希望用某种方法来具体描述。

定义 2.3.1. 设 P 是平面 $\mathbb{R}^2$ 中某个点的运动：

$$\begin{aligned} P : \mathbb{R} &\to \mathbb{R}^2 \\ t &\mapsto P(t) \end{aligned}$$

给定两数 $t_1 < t_2$，则 P 在区间 $[t_1; t_2]$ 上的**速度**是：

$$\frac{P(t_2) - P(t_1)}{t_2 - t_1}.$$

显然，这里的自变量是日常生活中时间概念的抽象，我们一般认为自变量属于某个数轴，称为**时间轴**，称具体的 t 为**时刻**，轴上的区间为**时间段**。而以上定义的速度是某个时间段 $[t_1;t_2]$ 上的速度，也称为这段时间的**平均速度**。它是两个量的商。作为被除数的 $P(t_2)-P(t_1)$ 是平面上的向量，从 t_1 时刻所在的点指向 t_2 时刻所在的点，表示点在这段时间的位置的改变量，或者说移动情况，因此称为点（在该时间段的）**位移**。作为除数的 t_2-t_1 是时长。

换句话说，某段时间里，**点运动的（平均）速度等于位移除以时长**。

类似的定义还出现在很多地方。比如，材料科学中，会考虑温度的改变对材料电阻的影响。实践中，考虑电阻率关于温度的函数：

$$\begin{aligned} R : \mathbb{R} &\to \mathbb{R} \\ t &\mapsto R(t) \end{aligned}$$

这里的自变量是温度，而应变量是电阻率 R。可以认为温度作为变量也属于某个数轴，如果我们试图描述温度改变时电阻率 R 改变的快慢，那么可以尝试用类似的方式描述：

$$\frac{R(t_2)-R(t_1)}{t_2-t_1}.$$

其中 $t_1<t_2$ 表示起始和最终温度，$R(t_2)-R(t_1)$ 就是温度从 t_1 增加到 t_2 时电阻率的改变量。

要注意的是，以上两个例子中，时间作为自变量，增长是不可逆的，而温度是可上升也可下降的。因此，两者并不完全等同。但我们在讨论变化的快慢时，用到了同样的方法，即用映射的应变量的改变量除以自变量的改变量。我们把这个比率称为实变映射（在某个区间上的）**变率**。

定义 2.3.2. 实变映射的变率 设 f 是定义在实数集上的映射，给定实数 $t_1<t_2$，则 f 在 $[t_1;t_2]$ 上的**变率**是：

$$\frac{f(t_2)-f(t_1)}{t_2-t_1}.$$

速度可以用来描述一段时间内运动的状态。最容易想到的是这样的运动: 任意时间段的速度都是一样的。我们把这样的运动称为**匀速运动**。显然，如果时间轴是整个数轴，匀速运动的轨迹就是直线。直线的方向向量就是速度向量。记匀速运动的速度为 $\mathbf{v}$，则时间段 $[t_1;t_2]$ 里，位移就是 $(t_2-t_1)\mathbf{v}$。设点在初始时刻 t_0 的位置是 P_0，则它在 t 时刻的位置是:

$$P(t)=P_0+(t-t_0)\mathbf{v}.$$

对于上面电阻率的例子来说，电阻率 R 随温度的变化也可以是匀速的。如果经实验证明，在某个温度范围 I 内，给定任意温度 $t_1,t_2\in I$，变率

$$\frac{R(t_2)-R(t_1)}{t_2-t_1}$$

都等于某个常数 q，就说相关材料的电阻率是匀速变化的。如果已知物体在某个温度 $t_0\in I$ 下的电阻率为 R_0，那么，对任意 $t\in I$，

$$R(t)=R_0+q(t-t_0).$$

再来看另一类简单的运动：给定圆 $\odot(O,r)$，点从 $(r,\ 0)$ 出发，逆时针方向沿圆周运动。这时，我们发现，任意时间段里的位移方向一直在改变。为此，我们改为研究点经过的**路程**。

路程是物体某段时间内运动轨迹的长度。它是一个数，而不是向量。显然，同样的位移，可以通过走不同的路来达到，而对应轨迹的长度是不一样的。也就是说，同样的位移可以对应不同的路程。

路程，作为时间的函数，可以定义其变率。我们把路程关于时间的变率叫做**速率**，和速度相区别。它们的主要区别和路程与位移的区别是一样的。速率是数，而速度是向量。

对于圆周运动，我们可以讨论它的速率，就是运动的路程关于时间的变率。最简单的圆周运动，任意时间段的速率是定值。我们把这样的运动叫做**匀速圆周运动**。要注意的是，这里的“速”指“速率”。

在上面的例子中,我们可以假设运动的速率总是 a。t_0 时刻,点从 $(r,\ 0)$ 出发，于是点在 t 时刻经历的路程为 $a(t-t_0)$。也就是说，物体沿圆周走过的弧长为 $a(t-t_0)$。于是，$t=\frac{2\pi r}{a}+t_0$ 时，点走过的路程为 $2\pi r$，即转动了一周。

我们可以把这个路程换算成点对应的向量和 t_0 时的向量的夹角。这样，我们就得到角度关于时间的函数 θ:

$$\theta:\ \ t\mapsto \frac{a(t-t_0)\cdot 360^\circ}{2\pi r}=\frac{a(t-t_0)\cdot 180^\circ}{\pi r}.$$

注意到角度关于时间的变率也是常数:

$$\frac{\theta(t_2)-\theta(t_1)}{t_2-t_1}=\frac{a\cdot 180^\circ}{\pi r}.$$

我们把这个变率称为圆周运动的**角速率**，把弧长对应的速率 a 称为**线速率**。圆周运动的线速率和角速率成正比。角速率不变的圆周运动是匀速圆周运动。

角速率在描述多个点的圆周运动时比较方便。比如，几个点绕着同一个圆心作圆周运动，但半径不相同。这时，考虑各个点的角速率，就容易理解点之间的关系。

思考 2.3.1.

1. 除了上文给出的例子，你还能说出哪些有意义的变率?

2. 我们常说“某个时刻的速度”，这种说法和以上定义的速度是否冲突? 你觉得可以怎样定义“某个时刻的速度”?

习题 2.3.1.

1. 点 P 沿单位圆逆时针运动，$t=0$ 时从 $(1,\ 0)$ 出发，$t=3$ 时到达 $\left(-\frac{1}{2},\ \frac{\sqrt{3}}{2}\right)$。这段时间里，它的位移是多少? 经过的路程是多少? 平均速度、线速率和角速率分别是多少?

2. 给定平面上的极坐标系。$t=0$ 时，点 P 从 $(1,\ 0)$ 出发。它的模长

和辐角是关于时间的函数：

$$\begin{aligned} r:&\ \ t \mapsto 1 + 2t^2 - 2.4t, \\ \theta:&\ \ t \mapsto 120^\circ t. \end{aligned}$$

2.1. $[0;3]$ 这段时间里，点 P 的位移是多少？平均速度、平均角速率分别是多少？

2.2. 圆 C 以原点为圆心，半径为 4.2。P 的轨迹与圆 C 有几个交点？在什么时刻相交？求交点的坐标。

第三章　从平面到立体

我们已经初步了解了简单平面图形的性质。下面我们来认识立体形状。

我们生活的世界是立体空间。人类自身和自然万物，都是立体的。立体形状是我们最常接触的形状。不过，人类的眼睛和大脑并不能直接处理立体形状，只能感知立体事物的平面图像，在大脑中还原事物的形状。因此，人类总是通过平面图像来了解立体事物，通过平面图像还原、表现现实。

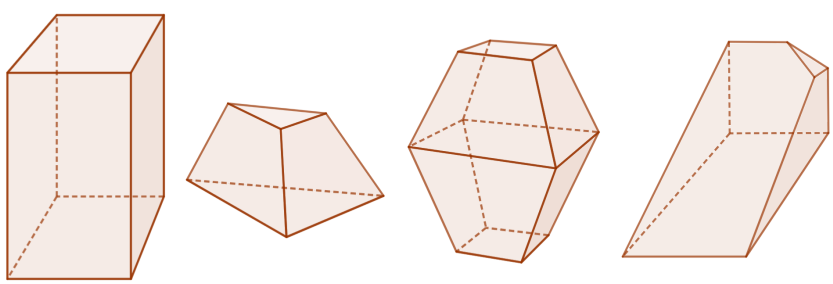
平面图像与现实

3.1 多面体

观察以下的物品，它们有什么相似之处？

如果立体形状是几个平面围成的，就说它是**多面体**。多面体的每个面都是多边形。

我们按照多面体的面数来称呼它。比如，由四个面围成的多面体叫四面体，由九个面围成的多面体叫九面体。



多面体相邻两面相交的部分总是线段，这些线段叫做**多面体的棱**。三条或以上的棱，交会在共同的端点上，这些点也是多面体相邻的几个面的共同点，称为**多面体的顶点**。

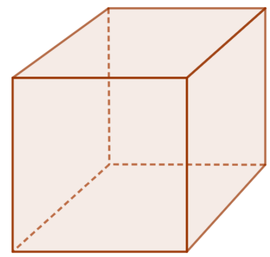
数一数，上图中的多面体各有几个面？有几条棱？有几个顶点？

思考 3.1.1.

1. 日常生活中，你还见过哪些多面体？

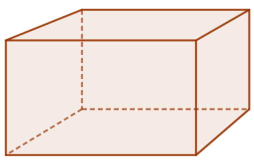
2. 多面体的顶点有没有可能只是两条棱的共同端点?

3. 多面体相邻两面的交点是否一定是线段? 多面体的棱有没有可能是三个或以上的相邻面的共同部分?



下面来认识几种特殊的多面体。

右图中的多面体由六个正方形组成。它叫做**正方体**或**正六面体**。观察正方体，你能总结出什么特征?



把正方体按某条棱的方向延长或压缩，得到的多面体叫做**长方体**。正方体就是一种特殊的长方体。显然，长方体也是六面体。观察长方体，你能总结出什么特征?

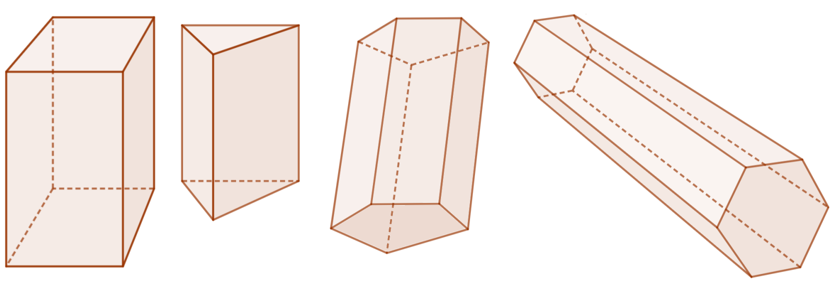
思考 3.1.2.

1. 正方体有几条棱? 棱与棱之间有什么关系?

2. 正方体有几个顶点? 顶点与顶点之间有什么关系?

3. 正方体是否是对称形状? 你能说出正方体的哪些对称性质? 长方体呢?

观察下图中的多面体，它们有什么共同特征?



把一个多边形朝指定方向平移，可以得到多面体。这样的多面体叫做**棱柱**。平移的多边形叫做**棱柱的底面**。其它面叫做**棱柱的侧面**。

思考 3.1.3.

1. 棱柱的侧面有什么特征?

2. 正方体和长方体是棱柱吗?

3. 你能说出生活中见过的棱柱吗?

4.《九章算术》刘徽注:“邪(斜)解立方得两堑堵,邪解堑堵,其一为阳马,一为鳖臑,阳马居二,鳖臑居一,不易之率也。合两鳖臑成一阳马,合三阳马而成一立方,故三而一。”其中的“立方“就是正方体。已知“堑堵”和“阳马”是五面体、“鳖臑”是四面体,自行探究:刘徽说的“堑堵”、“阳马”、“鳖臑”分别是什么?如何理解刘徽说的话?

3.2 旋转体和球

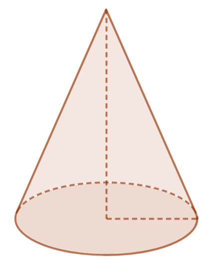
观察以下的物品,它们有什么相似之处?

如果立体形状是某个平面图形绕着定轴旋转一周而成的,就说它是**旋转体**。

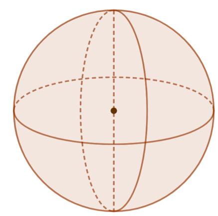
多面体可以是旋转体吗?

任何点绕定轴旋转一周,必然形成一个圆。考虑多面体表面的点,它们旋转之后是否能留在所属的平面上?

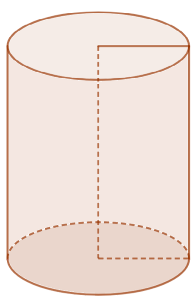
观察右图中的旋转体，它有几个面？它的面和多面体的面有什么不同？



直角三角形绕着直角边所在的直线旋转一周，形成的立体形状叫做**圆锥**。圆锥有两个面，一个面是平的，一个面是曲面。我们把平的面叫做**圆锥的底面**，把曲面叫做**圆锥的侧面**。侧面的顶端叫做**圆锥的锥顶**。



矩形绕着它某条边所在的直线旋转一周，形成的立体形状叫做**圆柱**。圆柱有三个面，其中两个面是平的，一个面是曲面。我们把平的面叫做**圆柱的底面**，把曲面叫做**圆柱的侧面**。



圆形绕着过圆心的直线旋转一周，形成的立体形状叫做**球**或**球体**。球只有一个面，叫做**球面**。

思考 3.2.1.

1. 日常生活中，你还见过哪些旋转体？

2. 圆锥、圆柱的底面是什么形状？为什么？

3. 圆柱和棱柱有什么相似之处？能不能用别的方式定义圆柱？

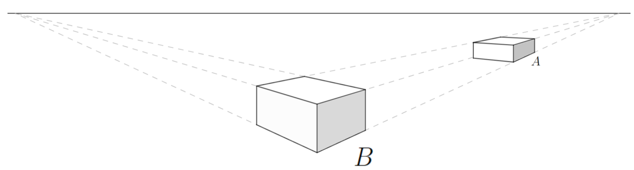
4. 圆锥、圆柱和球是否是对称形状？你能说出它们的哪些对称性质？

3.3 在平面上表示立体形状

让我们在平面上还原我们看到的立体事物。为什么下图中的 A 显得远，B 显得近？

大脑还原事物的形状时，遵循“近大远小”的规律。

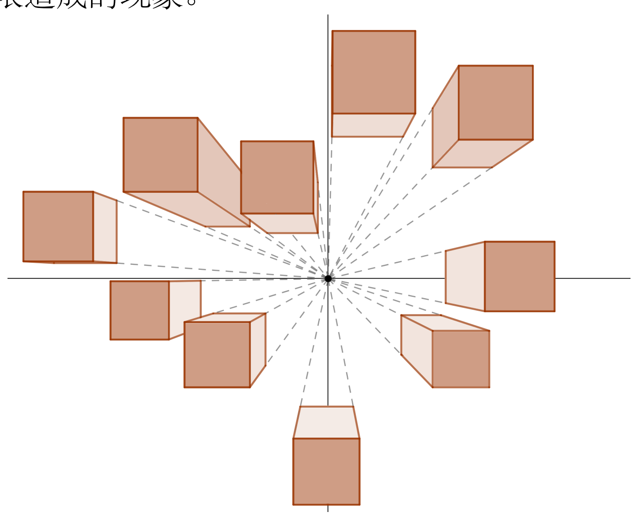
同一个物体，离眼睛越远，就显得越小；离眼睛越近，就显得越大。



近大远小

在平面中，可以使用“近大远小”的方法，表现立体事物的远近。这种表现方法称为**透视法**。在平面中表现立体形状的技法，叫做**作图法**，因为传统上，我们用画图的方法来表示现实世界中的立体形状。大部分写实绘画的作图法中都会用到透视法，让平面中表现的立体景象显得更真实。

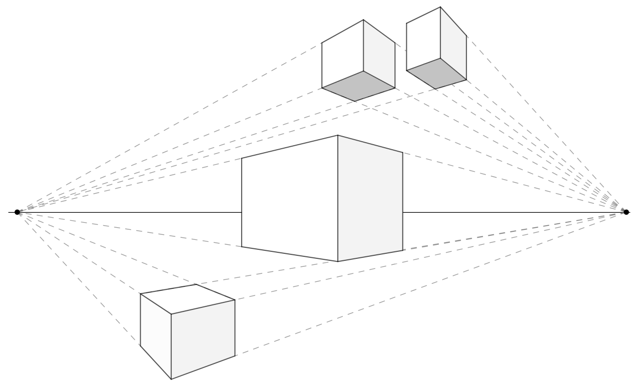
透视法有两个基本要诀，一个是“近大远小”，另一个是“向前缩短”。“近大远小”我们已经介绍过了，“向前缩短”指的是往远方的线并不会真的无限延伸，因此会显得比实际长度短一些。“向前缩短”可以说是“近大远小”和人的视力有限造成的现象。



单点透视

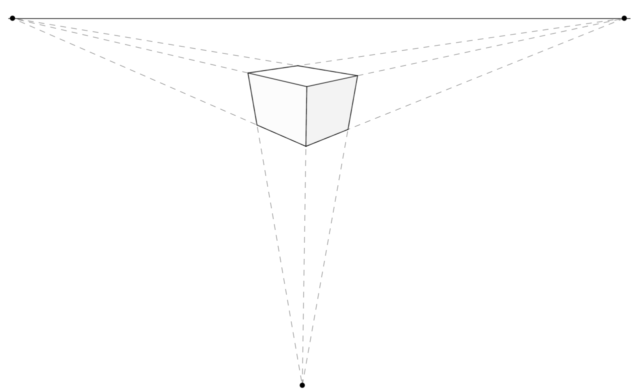
最简单的透视法是**直线透视法**。直线透视法认为，现实世界中的平行线，在平面作图时会交于同一点，即便在图中没有相交，其延长线也一定交于同一点。这个点代表着我们的视力能看到的最远极限，叫做**灭点**或**消失**

点。灭点可以是一个、两个或三个。对应的透视法称为**单点透视**、**两点透视**和**三点透视**。



两点透视

观察单点透视、两点透视和三点透视的图画，你觉得图中哪些线是现实世界中的平行线？相比现实，它们有什么变化？



三点透视

思考 3.3.1.

1. 如何理解“近大远小”现象？请试着将学过的知识推广到立体空间中，来解释“近大远小”现象。

2. 如何理解单点透视、两点透视、三点透视之间的区别和联系？为什么它们都可以反映真实的立体景象？你觉得哪种透视更接近真实？

3. 你还知道哪些透视法？自行探究并比较各种透视法，它们和以上介绍的透视法有什么不同？

第四章　同余

例子 4.0.1. 7^{65} 的个位数是多少?

解答. 从 $7^0, 7^1, 7^2, 7^3\cdots$ 开始找规律。$7^0 = 1$，$7^1 = 7$，$7^2 = 49$，$7^3 = 343$，$7^4 = 2401$，$7^5 = 16807$。7^4 和 7^0 的个位数都是 1，7^5 和 7^1 的个位数都是 7。我们可以总结出这样的规律：个位数是 1 的，乘以 7 得到 7；个位数是 7 的，乘以 7 得到 9；个位数是 9 的，乘以 7 得到 3；个位数是 3 的，乘以 7 得到 1。

也就是说，如果把 $7^0, 7^1, 7^2, 7^3\cdots$ 的个位数写成一列，应该是这个样子的：

$$1, 7, 9, 3, 1, 7, 9, 3, 1, 7, \cdots$$

用归纳法不难证明，这列数字以 4 为周期不断重复。所以，要求 7^{65} 的个位数，可以看 65 在相关的周期里处于哪个位置。换句话说，只要看 65 除以 4 的余数。$65 = 16 \times 4 + 1$，所以 7^{65} 的个位数和 7^1 的个位数一样，都是 7。

从这个例子可以看出，两个整数除以同一个数得到相同的余数，是一个重要的性质。我们把这种性质称为**同余**。比如，65 和 1 除以 4 余数都是 1，我们就说 65 和 1 模 4 同余，记为：

$$65 \overset{4}{\equiv} 1.$$

7^{65} 和 7^1 除以 10 余数都是 7，我们说 7^{65} 和 7^1 模 10 同余，记为：

$$7^{65} \overset{10}{\equiv} 7^1.$$

4.1　同余类

整数除以 3，余数有 $0,1,2$ 三种可能。整数除以 10，余数有 $0,1,\cdots,9$ 十种可能。一般来说，给定正整数 n，整数除以 n，余数有 $0,1,\cdots,n-1$ 这 n 种可能。因此，按除以 n 的余数，可以把整数集分成 n 类。同属一类的数，模 n 同余，所以这 n 类数叫作模 n **同余类**。所有模 n 同余类的集合，叫作模 n **同余系**。

每个模 n 同余类，可以写成 $\{kn+a\,|\,k\in\mathbb{Z}\}$ 的形式。也就是说，可以看成某个数 a 不断加上或减去 n 得到的所有数的集合。这个集合是无穷的。

不同的模 n 同余类，交集是空集，并集是 $\mathbb{Z}$。也就是说，它们是 $\mathbb{Z}$ 的分划。

为了方便，我们从每个模 n 同余类中选一个元素，代表这个同余类。一般来说，可以选 $0,1,\cdots,n-1$ 个数。我们给它们加个上划线，以便和作为整数的 $0,1,\cdots,n-1$ 区分：

$$\overline{0},\overline{1},\cdots,\overline{n-1}$$

如果要强调 n，可以把 n 加在右上角：

$$\overline{0}^n,\overline{1}^n,\cdots,\overline{n-1}^n$$

举例来说，当 $n=3$ 时，我们可以把全体整数分为三个模 3 同余类，分别记作 $\overline{0}$、$\overline{1}$、$\overline{2}$，它们分别表示除以 3 余数是 0、1、2 的三类整数。比如说，$\overline{1}$ 就表示集合 $\{3k+1\,|\,k\in\mathbb{Z}\}$，也就是 $1,4,7,\cdots$ 这样的数的集合。而这三个模 3 同余类的集合：$\left\{\overline{0},\overline{1},\overline{2}\right\}$，就叫做模 3 同余系。

给定整数 m，我们可以把它对应到某个模 n 同余类，称为对 n **取模**。比如 $n=5$ 时，$24\overset{5}{\equiv}4$，我们把 24 对应到 $\overline{4}^5$，或者说，24 对 5 取模，得 $\overline{4}^5$。

同余关系和相等关系很像，它们是否有一样的性质呢？我们可以验证，同余关系满足以下的性质：

- $\forall\, a \in \mathbb{Z},\quad a \overset{n}{\equiv} a$;
- $\forall\, a, b \in \mathbb{Z}$，如果 $a \overset{n}{\equiv} b$，那么 $b \overset{n}{\equiv} a$;
- $\forall\, a, b \in \mathbb{Z}$，如果 $a \overset{n}{\equiv} b$，$b \overset{n}{\equiv} c$，那么 $a \overset{n}{\equiv} c$。

满足以上三个性质的二元关系称为**等价关系**。数与数的相等关系是一种等价关系，数与数的同余关系也是一种等价关系。我们可以把数与数的同余关系看作同余类之间的相等关系。

定义 4.1.1. 给定正整数 n，如果模 n 同余类 $\bar{a}$ 中任意元素与同余类 $\bar{b}$ 中任意元素同余，就说模 n 同余类 $\bar{a}$ 和 $\bar{b}$ 相等。

整数之间有四则运算，模 n 同余类之间，也可以进行运算。以 $n = 5$ 为例子。我们分别计算 24 和 37 除以 5 的余数，以及它们的和 61 除以 5 的余数：

$$24 \overset{5}{\equiv} 4,\ \ 37 \overset{5}{\equiv} 2,\ \ 61 \overset{5}{\equiv} 1$$

可以发现：$4 + 2 \overset{5}{\equiv} 1$，也就是说，取模和加法可以交换顺序。可以验证，两个同余类中各取一个元素相加，和所在的同余类，就是两者模 n 余数的和所在的同余类。用集合的语言，可以写成：

$$\{kn + a + ln + b \,|\, k \in \mathbb{Z},\, l \in \mathbb{Z}\} = \{kn + a + b \,|\, k \in \mathbb{Z}\}$$

所以，可以定义同余类的加法：

$$\bar{a} + \bar{b} = \overline{a + b}$$

其中的 $\overline{a + b}$ 指的是 $a + b$ 所在的同余类。为了方便，我们用 $a + b$ 作为代表。

可以验证，同余类的加法也满足结合律和交换律。这里我们只证明同余类的加法满足结合律，交换律的证明留做习题：

证明： 由上可知 $\overline{a}+\overline{b}=\overline{a+b}$，所以

$$(\overline{a}+\overline{b})+\overline{c}=\overline{a+b}+\overline{c}=\overline{a+b+c}.$$

类似可得:

$$\overline{a}+(\overline{b}+\overline{c})=\overline{a}+\overline{b+c}=\overline{a+b+c}.$$

于是

$$(\overline{a}+\overline{b})+\overline{c}=\overline{a+b+c}=\overline{a}+(\overline{b}+\overline{c}).$$

□

类似可以定义同余类的减法和乘法:

$$\overline{a}-\overline{b}=\overline{a-b},\ \ \overline{a}\cdot\overline{b}=\overline{a\cdot b}$$

可以验证，同余类的减法性质和整数减法一样，同余类的乘法也满足结合律、交换律和分配律。

能否定义同余类的除法呢？我们来看一个例子。设 $n=6$，考虑等式 $12\div 4=3$。12、4 和 3 分别对 6 取模，得到 0、4 和 3。考虑等式 $60\div 10=6$。60、10 和 6 分别对 6 取模，得到 0、4 和 0。也就是说，两个模 6 同余类中各取元素相除，商所在的同余类不是唯一的。所以，我们没法定义模 6 同余类的除法。

再看另一个例子。设 $n=5$，考虑以下的“乘法表”:

$\times$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{1}$	$\overline{3}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{1}$	$\overline{4}$	$\overline{2}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

可以看出，任何模 5 同余类乘以 $\overline{0}$ 都得到 $\overline{0}$，非 $\overline{0}$ 同余类乘以不同的同余类，结果也不同。这说明每个同余类除以另一个同余类（非 $\overline{0}$），都必然有唯一的结果。这样我们就定义了模 5 同余系里的除法。

习题 4.1.1.
动手做一做:

1. 证明同余关系满足等价关系所要求的三个性质。

2. 证明同余类的加法满足交换律。

3. 证明同余类的减法是加法的逆运算。

4. 证明同余类的乘法满足结合律和交换律。

5. 证明同余类的乘法满足分配律。

6. 证明: 如果某模 n 同余类的代表与 n 的最大公因数是 d，则其中所有元素与 n 的最大公因数都是 d。

7. 分别画出模 3 同余系和模 4 同余系的“乘法表”。它们和模 5 同余系的“乘法表”哪些地方相同，哪些地方不同?

4.2 完全同余系和简化同余系

上一节我们提到模 6 同余系无法定义除法，而模 5 同余系可以定义除法。两者有什么不同呢? 我们画出模 6 同余系的“乘法表”:

$\times$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{0}$	$\overline{2}$	$\overline{4}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{2}$	$\overline{0}$	$\overline{4}$	$\overline{2}$
$\overline{5}$	$\overline{0}$	$\overline{5}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

可以看到，这个“乘法表”和模 5 同余系的大有不同。同一行或同一列常有重复。这说明不同的同余类乘同一个同余类得到同一个结果。比如

$$\overline{2} \times \overline{4} = \overline{5} \times \overline{4} = \overline{2}.$$

这就使我们没法定义除法。

如果我们把上面的等式稍作变化，会得到：

$$\overline{0} = (\overline{5} - \overline{2}) \times \overline{4} = \overline{3} \times \overline{4}.$$

也就是说，有非 $\overline{0}$ 的同余类相乘等于 $\overline{0}$。同余类乘法的这个性质和整数乘法完全不同。我们把这种非 $\overline{0}$ 同余类叫做**零因子**。整数中没有零因子：非 0 的整数相乘必然不是 0。而只要有这种零因子存在，同余系中就会发生“不同的同余类乘同一个同余类得到同一个结果”的现象，从而无法定义除法。

有什么办法在模 6 同余系中定义除法呢？我们可以选一部分同余类，在其中定义除法。如果同余类 $\overline{a}$ 的代表 a 与 6 不互素，设最大公因数是 b，那么

$$\frac{a}{b} \times 6 = a \times \frac{6}{b}$$

于是有 $\overline{a} \times \overline{\frac{6}{b}} = \overline{0}$，出现零因子。因此，为了避免零因子问题，我们只选和 6 互素的数所在的同余类，也就是 $\overline{1}$ 和 $\overline{5}$。我们发现 $\{\overline{1}, \overline{5}\}$ 中可以定义乘法和除法（但不再满足加减法）。

$\times$	$\overline{1}$	$\overline{5}$
$\overline{1}$	$\overline{1}$	$\overline{5}$
$\overline{5}$	$\overline{5}$	$\overline{1}$

我们把模 6 同余系称为模 6 的**完全同余系**，把 $\{\overline{1}, \overline{5}\}$ 称为模 6 的**简化同余系**。

一般来说，我们把模 n 同余系称为模 n 的完全同余系，在其中可以定

义加减法和乘法；把其中所有和 n 互素的同余类的集合称为模 n 的简化同余系 [1]。

定理 4.2.1. 给定正整数 n，在模 n 的简化同余系中可以定义乘法和除法。

证明： 模 n 同余类的乘法已经定义好了。我们只需要说明：简化同余系中的同余类相乘，仍然在简化同余系中。这是因为与 n 互素的整数相乘，结果还是与 n 互素。

接下来定义除法。除法是乘法的逆运算。比照数的除法：$a \div b = a \times \frac{1}{b}$。因此，只要将简化同余系中每个同余类都对应一个“倒数”，就可以用“乘以倒数”来定义除法。

我们把模 n 简化同余系中的同余类用小于 n 且与 n 互素的正整数来代表，记为

$$1 = b_1 < b_2 < \cdots < b_{\varphi(n)} = n - 1.$$

其中 $\varphi(n)$ 是模 n 简化同余系的元素个数。考虑任一元素 b_i，我们接下来会证明：$b_ib_1, b_ib_2, \cdots, b_ib_{\varphi(n)}$ 模 n 两两不同余。于是，它们中恰有一个模 n 余 1。设 $b_ib_j \overset{n}{\equiv} 1$，那么 b_j 就是 b_i 的“倒数”。

最后用反证法证明命题：$b_ib_1, b_ib_2, \cdots, b_ib_{\varphi(n)}$ 模 n 两两不同余。

反设命题不成立，即存在 b_j, b_k 使得 $b_ib_j \overset{n}{\equiv} b_ib_k$。这说明 $n|b_i(b_j - b_k)$。由于 b_i 和 n 互素，根据倍和析因定理，存在整数 p, q，使得：

$$b_ip + nq = 1.$$

两边乘以 $b_j - b_k$，就得到：

$$b_i(b_j - b_k)p + nq(b_j - b_k) = b_j - b_k.$$

等式左边是 n 的倍数，因此 b_j 和 b_k 模 n 同余，这与它们的定义矛盾。

因此命题的否定为假，原命题为真。 □

[1] 通常不把 $\overline{0}$ 计入简化剩余系，以省去讨论除以 $\overline{0}$ 的问题。

简化同余系的除法和整数不同，任何同余类都能整除另一个同余类，不需要余数、带余除法的概念。每个同余类都有自己的“倒数”，比如在模 6 简化同余系中，$\overline{5}\times\overline{5}=\overline{1}$。我们把同余类的“倒数”称为它的（乘法）**逆**。

习题 4.2.1.

1. 写出模 12 的简化同余系。写出 $\overline{7}^{12}$ 的逆。

2. 比较模 12 简化同余系中的乘除法和模 4 完全同余系中的加减法，它们有何异同?

3. 写出模 10 的简化同余系。写出 $\overline{7}^{10}$ 的逆。

4. 比较模 10 简化同余系中的乘除法和模 4 完全同余系中的加减法，它们有何异同?

5. 给定素数 n，写出模 n 简化同余系。

4.3 方余定理

与模 n 简化同余系密切相关的一个定理是方余定理[①]。

定理 4.3.1. 方余定理 设 a 是模 n 简化同余系中某个同余类中的元素，则:

$$a^{\varphi(n)} \overset{n}{\equiv} 1$$

其中 $\varphi(n)$ 是模 n 简化同余系中同余类的个数。

比如，模 10 简化同余系有 4 个元素：$\overline{1},\overline{3},\overline{7},\overline{9}$。7 属于同余类 $\overline{7}$，则 $7^4 \overset{10}{\equiv} 1$。

证明： 我们把模 n 简化同余系中的同余类用小于 n 且与 n 互素的正整数来代表，记为

$$1=b_1<b_2<\cdots<b_{\varphi(n)}=n-1.$$

[①]这个定理也称为欧拉定理。但以欧拉命名的定理太多了。为了避免混淆，这里不采用。

它们两两不同余。把它们各自乘以 a, 得到 $\varphi(n)$ 个整数: $ab_1, ab_2, \cdots, ab_{\varphi(n)}$。
前面我们已经证明了，它们仍然两两不同余。
这说明这 $\varphi(n)$ 个整数也分别代表模 n 简化同余系中的各个同余类。
考虑乘积：$b_1b_2\cdots b_{\varphi(n)}$。$(ab_1)(ab_2)\cdots(ab_{\varphi(n)})$ 和它同余。也就是说：

$$b_1b_2\cdots b_{\varphi(n)} \overset{n}{\equiv} (ab_1)(ab_2)\cdots(ab_{\varphi(n)}) \overset{n}{\equiv} a^{\varphi(n)}b_1b_2\cdots b_{\varphi(n)}.$$

由于 $b_1b_2\cdots b_{\varphi(n)}$ 也与 n 互素，我们把等式两边除以 $b_1b_2\cdots b_{\varphi(n)}$，就得到:

$$a^{\varphi(n)} \overset{n}{\equiv} 1.$$

□

如果 n 是素数，那么 $1, 2, \cdots, n-1$ 都和它互素，于是模 n 的简化同余系就是 $\{\overline{1}, \overline{2}, \cdots, \overline{n-1}\}$，$\varphi(n) = n-1$。根据方余定理，只要 a 不是 n 的倍数，就有:

$$a^{n-1} \overset{n}{\equiv} 1.$$

这个结论也叫做费马小定理。

习题 4.3.1.

1. 给定素数 n，证明:

1.1. 除了 $\overline{1}$ 和 $\overline{n-1}$，其它同余类的逆都不是自己。

1.2. $(n-1)! \overset{n}{\equiv} -1$.

2. 设 a 与 n 互素，称使得 $a^m \overset{n}{\equiv} 1$ 的最小正整数 m 为 a 模 n 的**阶**。

2.1. 证明 a 的阶整除 $\varphi(n)$。

2.2. 如果 a 的阶等于 $\varphi(n)$，就说 a 是模 n 的**原根**。证明：如果 a 是模 n 的原根，那么模 n 简化同余系可以写成：$\{\overline{a^0}, \overline{a^1}, \cdots, \overline{a^{\varphi(n)-1}}\}$。

2.3. 找出所有模 7 的原根。

第五章　用数据说话

数据是客观事物的定量记录。比如，以下列表中有某班级学生的身高数据。

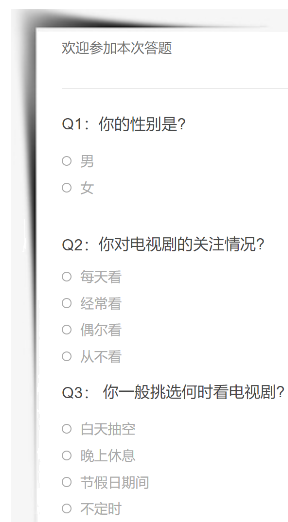
姓名	学号	身高（厘米）
张三	214092	166.4
李四	214033	171.5
王五	210819	164.8
⋮	⋮	⋮

生产生活中，我们常常以数量等形式记录客观事物的特征和属性。以合理的方式组织、呈现的数据，能帮助我们了解事物的本质，成为我们讨论、判断、决策的依据。

数据一般由**义**和**值**两部分构成。比如，我国 2020 年国民生产总值为 101 兆 5986 亿元人民币。这个数据的义是“我国 2020 年国民生产总值”，值是“101 兆 5986 亿元人民币”。数据的义和值缺一不可。没有义，就无法理解数据代表什么、与什么事物有关；没有值，就无法使用数据来了解相关事物。

5.1 样本和特征

要了解事物的某种性质，我们需要收集数据。比如，要了解学生的身体素质，我们可以通过体检收集学生的身高、体重、肺活量、血压等相关数据。如果我们要了解学生的身高，就要从每个学生的身高数据出发进行研究。这里每个学生的身高称为一份**样本**。对学生的各项数据来说，和某种性质相关的某一项，称为一个**特征**。比如，我们收集了 100 名学生的身高、体重、肺活量和血压数据。从总体来说，每个学生的数据是一个样本，共有 100 份样本。每个学生的数据都包括身高、体重、肺活量、血压。因此，这 100 份样本涉及 4 个特征。如果我们只研究学生的身高状况，那么可以把每个学生的身高数据看作一份样本。



实际生活中，收集数据样本的方式多种多样。一种常见的方式是通过调查问卷获得数据。右图是一份关于电视剧的调查问卷（部分）。

利用调查问卷，可以收集人们的主观想法、喜好。如果要收集事物的客观性质，更多是通过科技手段。

直接收集到的数据，也许无法直接反映事物的本质，难以让我们总结事物的规律。为此，我们对收集到的数据进行整理加工，得到新的数据。为了区别直接收集到的数据和经过人为整理加工的数据，我们一般把前者称为**原始数据**，把后者称为**生成数据**。

通常来说，整理加工数据的手段主要有以下几种：提高数据质量、转化数据形态、变换数据结构、新造特征。

为了提高数据质量而加工数据，称为数据清洗。我们收集到的数据，往往并不是我们理想中的样子。比如，我们收集 100 名学生的身高、体重、肺活量、血压数据，结果中可能有 3 名学生的血压数据缺失，6 名学生的肺活

量明显错误，4 名学生的数据重复记录，7 名学生的身高与上次体检结果严重不一致等情况。这些问题会影响我们研究数据，作出结论。因此，需要清洗数据，提高数据质量，以便接下来进行研究。

检查数据质量，通常从以下几个方向入手：**完整性**、**唯一性**、**一致性**、**正确性**。

- 完整性检查，就是检查数据是否有缺失的。比如，收集 20 个城市的就业率做研究，结果缺了某城市的数据，则数据不完整。
- 唯一性检查，就是检查不同途径、时间收集的数据是否有重复的。比如，收集学生对新游泳馆的评价，出现同一名学生的两份样本，说明数据重复了。
- 一致性检查，就是检查有关系的数据是否一致、是否合理。比如，收集市场的交易数据，各账户买入和卖出的数量分别相加，得到的总买入量应该等于总卖出量。如果总买入量和总卖出量不相等，则数据不一致。
- 正确性检查，就是核实数据是否正确无误。比如，收集学生的身高数据，出现某学生身高 173 米，说明数据有误。

不同的研究中，出于实际需要和综合考虑，会使用不同的手段清洗数据，处理数据的完整性、一致性、唯一性、正确性问题。

转化数据形态，就是把数据的值用另一种形式呈现。转化数据形态往往是为了从另一个角度看待数据的值。比如，收集 100 名学生的肺活量：

姓名	肺活量（毫升）
张三	2800
李四	3200
王五	4100
⋮	⋮
钱六	3850

我们可以把数据转为:

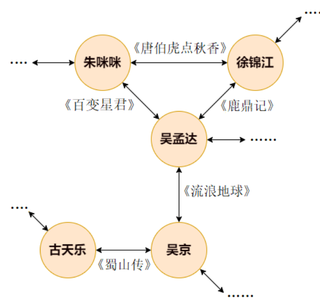
姓名	$[0;1)$	$[1;2)$	$[2;3)$	$[3;4)$	$[4;5)$	>5
张三	0	0	1	0	0	0
李四	0	0	0	1	0	0
王五	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
钱六	0	0	0	1	0	0

我们以升为单位，把肺活量的范围分划为几个区间，每个人的肺活量的值落在哪个区间，就在对应的列记 1，其他列记 0。这让我们用另一种角度来观察肺活量数据。

变换数据结构，是指用另一种方式组织数据。原始数据不容易理解，可能是因为数据组织的方式不合其义。组织数据的方式越符合其义，我们就越容易通过数据理解事物的本质。比如，我们收集某些影视演员的出演影视作品的数据，研究演员之间的关系。原始数据是各个演员的作品列表:

姓名	作品
周比利	《少年陈真》、《精武英雄》、《鼠胆龙威》、……
斯琴高娃	《北京爱情故事》、《骆驼祥子》、《香魂女》、……
朱咪咪	《百变星君》、《审死官》、《唐伯虎点秋香》、……
⋮	⋮
李绮虹	《西贡的童话》、《爱未央》、《圣诞玫瑰》、……
张涵予	《风声》、《智取威虎山》、《红海行动》、……

我们可以根据共演关系，把数据结构改为下图。每个圆点代表一个演员。曾经共演的演员之间连线，否则不连线。



从图中我们能够看出不同演员之间的合作关系。我们也可以把共演关系用数表的方式呈现。曾经共演的演员，相应的格子中填入共演作品数目，否则填 0：

	朱咪咪	吴孟达	古天乐	吴京	⋯
朱咪咪	-	3	0	0	⋯
吴孟达	3	-	3	1	⋯
古天乐	0	3	-	4	⋯
吴京	0	1	4	-	⋯
⋮	⋮	⋮	⋮	⋮	

新造特征，就是从原始数据的某些特征出发，构造新的数据。比如，从身高和体重数据出发，用体重除以身高的平方，就得到新特征：体质指数。

5.2 描述和分析

上一节的例子中，我们收集了 100 名学生的身高、体重、肺活量、血压数据。如何通过这些数据，了解这些学生的身体素质情况呢？我们可以查看每个学生的数据。然而，人的认知和理解能力是有限的。我们无法直接把这 400 个数值直接反映在头脑中。为此，我们希望用更少的信息来描述这些数据，让我们对这些数据有大概的认识。这种手段叫作**描述统计**。描述统

计中最基本的方式，是从数据出发算得一个数量，来描述数据。我们把这样的数量称为**统计量**。以下是一些常见的统计量。

极值：某特征所有样本值中最大的，称为该特征的**极大值**；所有样本值中最小的，称为该特征的**极小值**。比如，我们可以查看所有学生的体重，找出其中最大的值和最小的值。

平均数：某特征所有样本值之和，除以样本的个数，称为该特征的平均数或均值。比如，我们把所有学生的身高值加起来，除以 100，就得到学生平均身高。

中位数：如果某个数值比某特征一半样本的值大，比另一半样本的值小，就说它是该特征的中位数。显然，这个定义是不严谨的，实际计算时需要严谨的定义。最简单也是最常用的定义是：将该特征所有样本的值从小到大排列。如果样本个数 N 是奇数：$N=2n+1$，那么取第 $n+1$ 大的数为中位数；如果样本个数是偶数：$N=2n$，那么取第 n 大和第 $n+1$ 大的数的均值为中位数。比如，把 100 名学生的肺活量值从小到大排列，第 50 大的值和第 51 大的值之和除以 2，就是中位数。

根据极值、平均数和中位数，我们可以大致描述数据的总体情况。比如，知道了学生的身高极小值是 144 厘米，极大值是 191 厘米，平均值是 165.7 厘米，中位数是 163.8 厘米，我们就对这 100 名学生的身高有了基本的认识。

要更进一步分析数据，可以通过**加权平均数**、**分位数**和**直方图**。

加权平均数是平均数的推广，指对每个样本的值乘以一个系数（称为**权重**）来描述它的“重要程度”，然后求和，最后除以权重之和。比如，我们想知道男生身高的平均数时，可以把每个男生的身高值乘以权重 1，把女生的身高值乘以权重 0，求和之后除以权重之和（也就是男生的人数），就得到加权平均数。

分位数是中位数概念的推广。可以想象把所有样本的值标记在数轴上，

用手指从左边第一个值起，往右划过去，数过一个又一个值。数过一半样本时的值，大致就是中位数。类似地，设有 $0 < q < 1$，数过的样本数量和总数量之比为 q 的时候，指向的值就可以称为样本的 q **分位数**。q 称为分位，一般用百分数表示。中位数就是 50% 分位数。

显然，从右往左数也可以定义一种分位数。我们把从左往右数得到的称为**左** q **分位数**或**下** q **分位数**，把从右往左数得到的称为**右** q **分位数**或**上** q **分位数**。

分位数和中位数一样，是一种模糊的说法。实际计算中，我们要采用更严谨的定义。

给定 N 个样本值 $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_N$ 和某个值 a，定义 n^- 为小于等于 a 的样本个数。数轴上，a 从 $x_1 - 1$ 向 $x_N + 1$ 运动时，n^- 从 0 增大到 N。a 从某个 x_i 离开，运动到 x_{i+1} 期间，n^- 是不变的。a 到达 x_{i+1} 时，n^- 增大。

分位数的自然定义：如果 a 到达某个 x_i 前，$n^- < qN$，而到达 x_i 时 $n^- \geqslant qN$，就定义 x_i 为左 q 分位数。如果在某段 x_i 到 x_{i+1} 期间 $n^- = qN$，就定义 $(1-q)x_i + qx_{i+1}$ 为左 q 分位数。

此外还可以定义顺分位数。区别在于，如果在某段 x_i 到 x_{i+1} 期间 $n^- = qN$，顺分位数仍定义左端 x_i 为左 q 分位数。

自然定义的分位数和之前中位数的定义相洽，顺分位数则更方便计算。

分位数按照样本的数量来划分样本，划分越细，越有利于我们了解样本的分布情况。比如，根据 100 个学生的身高，求得分位数如下：

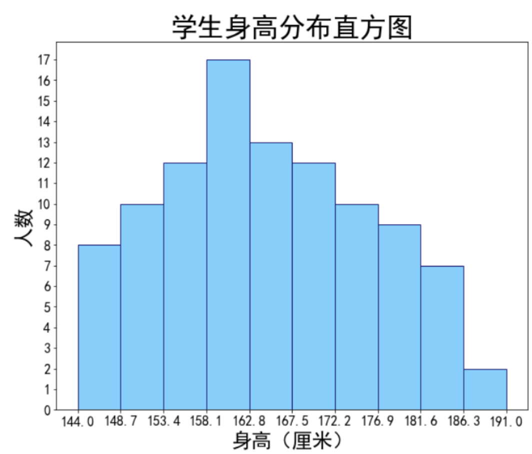
q	10%	20%	30%	40%	50%	60%	70%	80%	90%
分位数 (cm)	149.7	154.3	158.0	161.2	163.8	167.5	171.7	175.6	180.2

可以看出，学生身高在 40% 到 50% 阶段较为集中，在 80% 到 90% 阶段较为稀疏。

直方图是另一种展示样本分布情况的方法。我们取定某个包含了所有样本值的区间，然后将其均匀分划为若干份。比如，我们取学生身高的极小值和极大值构建区间 $[144; 191]$，将其均匀分划分成 10 份，每份长为 $\frac{191-144}{10} = 4.7$。数轴上，左起第一个区间是 $[144; 148.7)$，第二个区间是 $[148.7; 153.4)$，等等。最右边的区间是 $[186.3; 191]$。把这 10 个区间从左到右编号，每个样本的身高值恰好属于其中一个。我们可以把这 10 个区间想象成 10 个“抽屉”，把 100 个学生的身高值“放进”这些“盒子”里，然后统计每个“盒子”里的样本数。最终我们可以得到一个以下样子的表格：

区间序号	1	2	3	4	5	6	7	8	9	10
样本数	8	10	12	17	13	12	10	9	7	2

我们还可以把这个结果用图表形式画出来：



直方图比分位数表格更直观、更容易理解。我们可以直接看出，学生身高在 $[158.1; 162.8)$ 区间最为集中，共有 17 人的身高落在这个区间里；在 $[186.3; 191]$ 区间最为稀疏，只有 2 人的身高落在这个区间里。

比直方图更简单的描述方法是**平均数和方差**。平均数是所有样本值之和除以样本数的商。如果所有样本值相差不多，那么平均数可以告诉我们样本值大致分布在哪里。比如，算得 100 个学生的身高值之和为 16426.5 厘米，则他们身高的平均值为 164.265 厘米。如果所有学生身高相差不多，就

可以说他们的身高大致是 164 厘米。

当然，我们从直方图可以看出，学生的身高差距并非"不多"。因此，我们引入**方差**的概念，描述样本值的集中程度。方差和平均数一起，可以粗略呈现样本分布的情况。样本的方差定义为样本值与均值之差的平方的均值。给定 N 个样本值：$x_1, x_2, \cdots, x_N$，记它们的均值为 $\mathbf{u}$，方差为 v，则：

$$v = \frac{1}{N}\sum_{i=1}^{N}(x_i - u)^2.$$

由于 $u = \frac{1}{N}\sum_{i=1}^{N} x_i$，可以推出：

$$v = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - u^2.$$

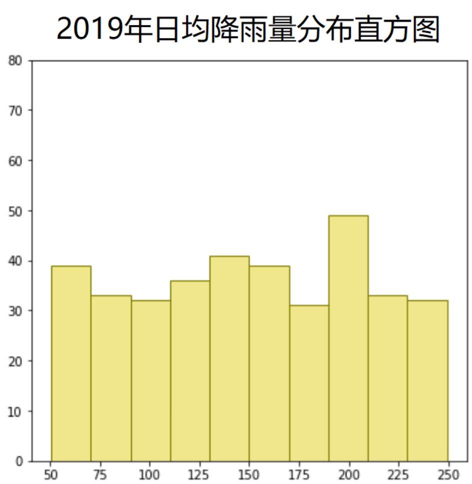
比如，$1, 2, 3, 4, 5$ 的平均数为：

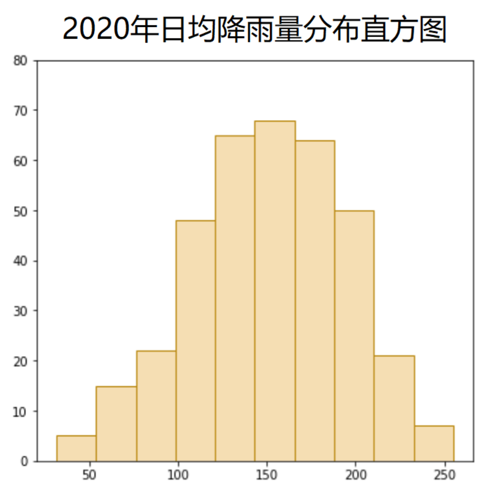
$$\frac{1+2+3+4+5}{5} = 3,$$

方差为：

$$\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} = 2.$$

方差的概念在单批样本中比较难理解，但如果用于比较两批样本，就容易理解了。举例来说，考察某地连续两年的日均降雨量，绘制直方图：





可以看出，虽然两年平均降水量差不多，但相比 2019 年，2020 年的日均降雨量分布较为集中。计算得到的方差也说明这一点：2019 年日均降雨量的方差为 3147.87，而 2020 年日均降雨量的方差只有 1872.21。

思考 5.2.1.

1. 能否把分位数表格的信息用图表形式展示？

2. 绘制直方图时，是否有别的方式划分区间？

3. 分位数表格和直方图有什么联系？

4. 直方图划分的区间个数是否越多越好？你认为应该如何确定划分的区间数？

5. 除了方差，你还能想到什么方法来描述数据的集中程度？它们和方差相比有什么好处？

习题 5.2.1.

1. 写出右分位数的自然定义。

2. 写出顺右分位数的定义。

5.3 数据的结构

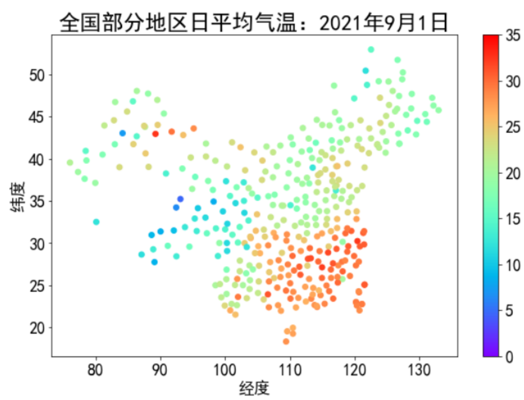
上一节中我们描述、分析数据的时候，并未考虑数据本身的结构。什么是数据的结构？数据结构是不同数据样本之间因数据自身含义产生的联系的总和。用适合的方式展现数据的结构，有助于我们理解数据包含的信息，描述并分析数据。

一种常见的数据结构是时空结构：不同的数据样本代表了不同时间不同地点的信息。用合适的方式展现这种时空结构，对我们研究相关事物大有帮助。

举例来说，下表中有我国部分地区某日的日均气温数据。

地区编号	日期	经度	纬度	气温（°C）
391027	2021–09–01	110.530128	34.754986	15.9
391028	2021–09–01	111.105681	35.516971	18.9
391029	2021–09–01	111.183294	34.207944	19.3
⋮	⋮	⋮	⋮	⋮

从列表中难以看出数据之间的联系。我们可以把相关的数据展示在对应的地理位置上，形成下图：



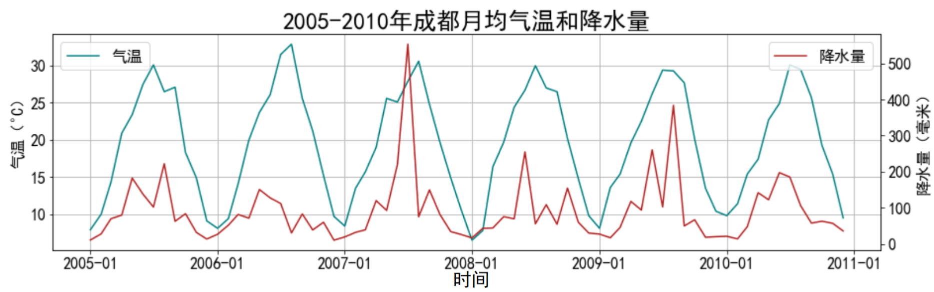
这种图叫**热力图**，能直观表现各地气温之间的关系。

以上是空间地域结构的展示。再来看时间结构的展示。下表是 2005 年至 2010 年成都市月均气温和降水量数据。

月份	气温（°C）	降水量（毫米）
2005–03	14.3	69.3
2005–01	7.9	10.3
2005–02	10.0	27.5
⋮	⋮	⋮
2010–12	9.5	35.5

从列表中难以看出数据之间的联系。可以把数据按时间先后顺序排列，

这样得到的数据称为**时间序列**。以 x 轴表示时间，以 y 轴表示气温和降水量，按照时间先后顺序，把每个月的气温和降水量对应的点用线段连起来，就形成下图：

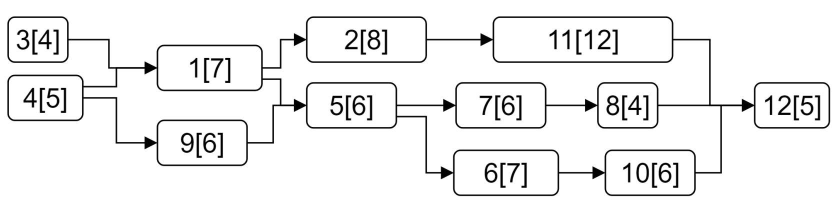


从图中，我们容易看出当地降水量和气温数据随时间变化的方式。这种呈现数据的方式称为**折线图**。

另一个例子涉及更复杂的关系。下表中有某工厂完成一个项目所需的任务、各项任务的前置任务（完成前置任务才能开始本任务）和预计用时：

任务编号	预计用时	前置任务
1	7	3，4
2	8	1
3	4	-
4	5	-
⋮	⋮	⋮

从列表中难以判断应该如何组织工作。我们可以把各项任务的关系用箭头连接起来，用图表来表示：



这个展示数据的方法称为**流程图**。图中将任务用圆角方块表示，方块长度和任务预计用时成正比。方块中标明了任务编号和预计用时（方括号内）。用箭头把每个任务指向它的后继任务（必须完成本任务才能开始后继任务）。据此，我们就能更好地理解完成任务的必要次序，更合理地安排工作。

第六章　数据和概率

之前，我们学习了用概率来讨论不确定的事件。如何判断某件事情将来会不会发生？我们需要通过已有的知识和经验来分析。数据就是我们对已有知识和经验的总结、提炼。通过分析数据，我们可以讨论不确定的事件发生的概率。

6.1　实验和样本

为了研究我们关心的问题中某些随机事件发生的规律，我们会特意设计**实验**，布置相关条件，并观察发生的事件，记录结果。每次实验，我们得到的结果就是一个数据样本。比如，为了研究投骰子各面朝上的规律，我们设定骰子的大小、材料和形制，设计投掷的高度，并观察投掷结果。这就是一个实验。

实际中，很多问题并不是人类有能力设计、布置必要条件的，但我们仍然关心其中随机事件发生的规律，比如台风、地震、火山爆发等等。因此，我们假设大自然设计了相关的实验，框定一些假设条件，并观察发生的事件，记录结果。我们称之为广义的实验。比如，为了研究我们所在地区的降雨情况，我们观察每天是否降雨，记录降雨量，就是一次广义的实验。为了研究方便，我们会假设记录地点代表了所在地区（比如全市）的平均降雨情况，或者设定每天降雨量为早上 8 点到次日早上 8 点之间的记录量等等，

但我们无法完全控制实验的条件。

一次大实验可以包含几个小实验。比如，我们可以把连续记录一周的日降雨量称为一次实验，也可以把每天的记录称为一次实验。为了好说话，我们会区分**单一实验**和**复合实验**。如何区分，取决于我们关心的问题的性质和我们的研究方法。比如，体检中，我们可以把测量每个学生身高、体重、血压和肺活量合称为一个单一实验，这样，测量 100 个学生就是复合实验。如果我们关心的是这些学生的整体状况，我们可以认为这是 1 次复合实验，其中每个学生的测量称为单一实验。如果我们关心的是某个学生的状况，那么我们可以重复测量这个学生 100 次，称为复合实验，其中每次测量称为单一实验。

实验中，我们观察某些事件发生的情况，把我们关心的部分用数据记录下来，这样就得到一个数据样本。这样记录的叫做实验的**终态**或**结果**。所有可能的终态构成的集合称为终集。终态是终集的元素。

6.2 随机变量

一般来说，我们把实验的每个结果关联到某个方便理解的数值，作为数据记录。这样的映射，称为**随机映射**或**随机变量**。比如，设计抛掷硬币的实验，每次投掷如果正面向上，就规定结果是 1，否则规定结果是 0。那么我们可以把投掷结果用随机变量表示：

$$
\begin{aligned}
h \;:\; \{\text{正面},\text{反面}\} &\rightarrow \{0,\ 1\} \\
h(\text{正面}) &= 1 \\
h(\text{反面}) &= 0
\end{aligned}
$$

随机变量 h 把终态“正面”映射到 1，把终态“反面”映射到 0。

又比如，为了研究投骰子各面朝上的规律，我们设计投掷骰子的实验，每次投掷的结果是朝上的数字，我们约定，每次投掷骰子，得到的收益是这

个数值的 10 倍。如果数字 2 朝上，就获得 20 元；数字 5 朝上，就获得 50 元，等等。则我们可以把每次投掷骰子的收益用随机变量表示：

$$\begin{aligned} f \quad : \quad & \{1,2,3,4,5,6\} \to \{10,20,30,40,50,60\} \\ & i \to 10 \times i \end{aligned}$$

随机变量 f 把终态 1 映射到 10，把终态 2 映射到 20，把终态 3 映射到 30，等等。

一般来说，设试验的终集为 S，如果某个映射把终集里的每个终态映射到另一个集合 X 里，就说这个映射是 S 上的 X–随机映射或随机变量。如果集合 X 是有限集合或有理数集的子集，就说随机变量是**离散**的；如果集合 X 是实数集或区间，就说随机变量是**连续**的。

我们把随机变量背后随机事件的概率分布，称为**随机变量的分布**或随机变量遵循的分布。可以看出，随机变量的性质与其分布有关。随机变量的分布如何具体影响随机变量呢？我们来看两个例子。

例子 6.2.1. 考虑这样的映射 $\mathbb{E}$：它将离散随机变量 f 映射为

$$\mathbb{E}(f) = \sum_{s \in S} f(s)\mathbb{P}(\{s\}).$$

这个映射称为**随机变量的期望**。比如，对以上掷骰子试验的随机变量 f，它的期望为：

$$\mathbb{E}(f) = \sum_{i=1}^{6} 10 \cdot i\mathbb{P}(\{i\}).$$

如果投骰子的分布是均匀分布，那么对任何 i，$\mathbb{P}(\{i\})$ 总是 $\frac{1}{6}$，于是：

$$\mathbb{E}(f) = \sum_{i=1}^{6} \frac{10 \cdot i}{6} = 35.$$

如果 1 朝上的概率是 50%，其他数字朝上的概率都是 10%，那么：

$$\mathbb{E}(f) = 10 \cdot 1 \cdot 50\% + \sum_{i=2}^{6} \frac{10 \cdot i}{10} = 25.$$

随机变量分布改变，它的期望也可能改变。

例子 6.2.2. 考虑这样的映射 Var：它将离散随机变量 f 映射为

$$\mathrm{Var}(f) = \sum_{s\in S} (f(s) - \mathbb{E}(f))^2\, \mathbb{P}(\{s\}).$$

这个映射称为**随机变量的方差**。比如，对以上掷骰子试验的随机变量 f，它的期望为：

$$\mathrm{Var}(f) = \sum_{i=1}^{6} (10i - \mathbb{E}(f))^2\, \mathbb{P}(\{i\}).$$

如果投骰子的分布是均匀分布，那么对任何 i，$\mathbb{P}(\{i\})$ 总是 $\frac{1}{6}$，于是 $\mathbb{E}(f) = 35$，因此：

$$\mathrm{Var}(f) = \sum_{i=1}^{6} \frac{(10i - 35)^2}{6} = \frac{875}{3} \approx 291.67$$

如果 1 朝上的概率是 50%，其他数字朝上的概率都是 10%，那么 $\mathbb{E}(f) = 25$，因此：

$$\mathbb{E}(f) = (10\cdot 1 - 25)^2 \cdot 50\% + \sum_{i=2}^{6} \frac{(10\cdot i - 25)^2}{10} = 325$$

随机变量分布改变，它的方差也可能改变。

另一种表示随机变量分布的方法，是**概率密度函数**。概率密度函数从随机变量的值出发描述它遵循的分布，把 X 中元素值映射到“随机变量等于该值”这一事件的概率。比如，投掷骰子的随机变量 f 的分布，可以用以下的概率密度函数表示：

$$\begin{aligned} p_f \;:\; \{10,20,30,40,50,60\} &\to [0;1] \\ x &\mapsto \mathbb{P}(\{s \in \{1,2,3,4,5,6\} \mid f(s) = x\}) \end{aligned}$$

概率分布不同，随机变量 f 的概率密度函数也不同。例如，投骰子的分布是均匀分布时，$p_f(10) = \frac{1}{6}$；如果 1 朝上的概率是 50%，那么 $p_f(10) = \frac{1}{2}$。

概率密度函数是另一种描述随机变量的方法，可以看作如何描述随机变量的值的分布，因此，我们也把概率密度函数叫做随机变量的值遵循的分布，或者随机变量值的分布。

使用概率密度函数，也可以表示随机变量的期望和方差。对 X 中每个元素 x，考虑使 $f(s)=x$ 的 s。我们发现:

$$\begin{aligned}\sum_{f(s)=x} f(s)\mathbb{P}(\{s\}) &= \sum_{f(s)=x} x\mathbb{P}(\{s\})\\ &= x\sum_{f(s)=x}\mathbb{P}(\{s\})\\ &= x\mathbb{P}(\{s\in S\mid f(s)=x\})\\ &= xp_f(x)\end{aligned}$$

因此，随机变量 f 的期望

$$\mathbb{E}(f)=\sum_{s\in S} f(s)\mathbb{P}(\{s\})=\sum_{x\in X} xp_f(x).$$

这是因为当 x 遍历 X 时，使 $f(s)=x$ 的 s 也遍历了 S，没有重复也没有遗漏[①]。

同理可以推出，随机变量 f 的方差可以写成:

$$\mathrm{Var}(f)=\sum_{x\in X}(x-\mathbb{E}(f))^2 p_f(x).$$

如何理解随机变量的期望和方差呢?

首先来看期望。随机变量的期望可以看作一种对实验结果预期的描述。举例来说，考虑以上掷硬币的实验的随机变量 h。假设硬币正面朝上的概率是 p，那么 h 的期望是:

$$\mathbb{E}(h)=1\cdot p+0\cdot(1-p)=p.$$

[①]不至于混淆时，$p_f(x)$ 也写作 $\mathbb{P}(f=x)$。

于是，期望越大，表示 p 越大，说明掷硬币越可能正面朝上；反之说明掷硬币更可能反面朝上。又比如，考虑以上掷骰子的实验的随机变量 f。假设各面朝上的概率分别是 $p_1, p_2, \cdots, p_6$，那么 f 的期望是：

$$\mathbb{E}(f) = 10p_1 + 20p_2 + \cdots + 60p_6.$$

如果 p_1 接近于 1，其它概率接近 0，那么 $\mathbb{E}(f)$ 接近于 10，即预期收益和 1 点向上的收益相近。如果 p_1 接近于 5，其它概率接近 0，那么 $\mathbb{E}(f)$ 接近于 50，即预期收益和 5 点向上的收益相近。

一般来说，随机变量的期望是一种方便的描述预期结果的方法，和平均数很像。和平均数的区别在于，随机变量的期望和各个终态的概率有关。概率越大的终态，对期望的影响越大，反之亦然。如果把各个终态看作已知数据，那么这种描述数据的方法就是**加权平均数**。加权平均数是指对每个数据值乘以一个系数（称为权重）来描述它的“重要程度”，然后求和，最后除以权重之和。要注意区分的是：随机变量的期望和数据的加权平均数的区别在于，随机变量的期望是一种预估方法，根据终态的概率估计预期结果，而数据的加权平均是对已知数据的描述方法。虽然两者的计算方式相同，但它们的含义不同。

再来看方差的作用。在投骰子的例子里，考虑以下两种概率分布：

- $p_1 = p_6 = 40\%$，$p_2 = p_3 = p_4 = p_5 = 5\%$。
- $p_3 = p_4 = 40\%$，$p_1 = p_2 = p_5 = p_6 = 5\%$。

两种情况下，随机变量的期望相同，都是 35，但显然这两种分布有很大区别。而方差可以描述这两种分布的区别。计算两种情况的方差，分别得到 525 和 85。可以观察到，期望相同的情况下，分布的方差越大，结果远离期望值的概率越大，方差越小，结果远离期望值的概率越小。随机变量的方差和数据的方差计算方法相同，但它们的含义不同。随机变量的方差是对预期结果的描述方法，而数据的方差是对已知数据的描述方法。

通过期望和方差，我们可以对概率分布有粗略的了解。

思考 6.2.1.

1. 概率映射和随机变量有什么不同?

2. 比较随机变量的期望和样本数据的平均数。它们有什么相似之处?有什么不同?

3. 比较随机变量的方差和样本数据的方差。它们有什么相似之处?有什么不同?

4. 随机变量的分布改变，它的期望一定改变吗?

5. 随机变量的分布改变，它的方差一定改变吗?

6. 随机变量的期望和方差的不同写法有哪些要注意的地方?你觉得每种写法好在哪里?

习题 6.2.1.

1. 小李投掷一枚 1 元硬币，约定正面向上则小王给小李 10 元钱，否则小王给小李 1 元钱。

1.1. 假设硬币正面向上的概率是 47%，试将小王给小李的钱数用相关的随机变量表示。

1.2. 该随机变量的分布是什么分布?

1.3. 写出该随机变量的概率密度函数。

1.4. 写出该随机变量的期望和方差。

2. 投掷一枚骰子，假设 1 点向上和 5 点向上的概率分别是 40%，其余点数向上的概率是 5%，如果以点数的 3 倍为随机变量，求它的期望和方差。

3. X 是 $\mathbb{Q}$ 的子集。证明 S 上的 X-随机变量 f 的方差也可以写成:

$$\mathrm{Var}(f) = \mathbb{E}(f^2) - \mathbb{E}(f)^2.$$

其中 f^2 表示把每个终态 s 映射到 $f(s)^2$ 的随机变量。

6.3 条件概率和独立事件

讨论随机事件，总要知道前提条件。只有明确了前提条件，才能讨论事件的概率。为了研究不同的前提条件对事件的影响，我们引入**条件概率**的概念。

仍以“明天杭州是否下雨”为例。我们讨论台风“凤凰”动向对下雨概率的影响，可以说：“如果台风靠近，那么下雨的概率有 90%”、“如果台风转向，那么不下雨的概率为 40%”。这里我们把台风靠近与否作为明天杭州下雨的前提条件。以上两句话可以记作：

$$\mathbb{P}(\{\text{下雨}\}\,|\,\{\text{台风靠近}\}) = 0.9, \quad \mathbb{P}(\{\text{不下雨}\}\,|\,\{\text{台风转向}\}) = 0.4$$

这个记法称为条件概率。其中分隔符“|”左边是结果对应的事件，右边是前提条件对应的事件，合起来表示在前提条件下，结果发生的概率。

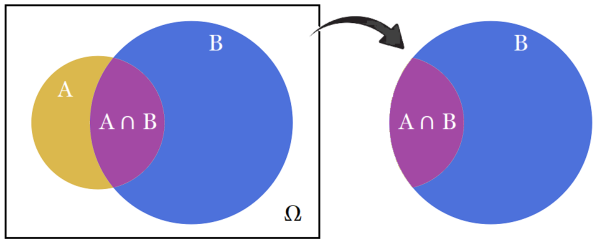
如果台风靠近，那么不下雨的概率是 10%；如果台风转向，那么下雨的概率是 60%。我们也可以把这些概率用表格的形式表示：

条件	下雨	不下雨
台风靠近	90%	10%
台风转向	60%	40%

可以看到，台风靠近与否，会影响明天杭州下雨的概率。如果把台风“凤凰”的动向视为一个实验，把“明天杭州是否下雨”视为另一个实验，我们也可以说，前一个实验影响了后一个实验的结果。

条件概率和概率之间的联系是怎样的呢？观察下图，可以发现，在假定事件 B 发生的条件下，发生事件 A 的条件概率，就是 A, B 同时发生的概率，除以事件 B 的概率。用公式表达出来，就是：

$$\mathbb{P}(A\,|\,B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$



把“B 发生”作为前提条件，A 发生的概率，就是 A, B 同时发生的概率，除以 B 发生的概率

我们称 $\mathbb{P}(A \mid B)$ 为“B 时 A 的概率”。要注意的是，这个公式没有考虑 $\mathbb{P}(B) = 0$ 的情况。如果 B 的概率是 0，那么讨论 B 发生时 A 的概率比较困难。这时我们暂时不定义条件概率 $\mathbb{P}(A \mid B)$。

从映射的角度来看，从事件 A 到 B 时 A 的概率形成这样一个映射：

$$A \mapsto \mathbb{P}(A \mid B)$$

可以验证，这个映射也是一个概率映射，叫作关于 B 的条件概率映射。我们一般把这个映射记为 $\mathbb{P}_B$ 或 $\mathbb{P}(\cdot \mid B)$（其中的 $\cdot$ 表示自变量的位置）。

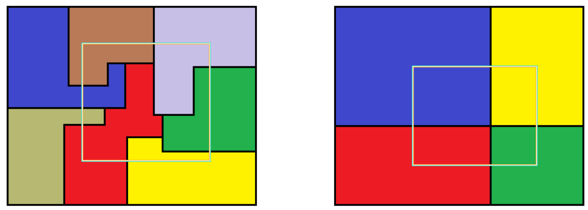
再来看另一个例子：有面值为 5 角和 1 元的硬币各一枚。首先抛出 1 元硬币，再抛出 5 角硬币，把 1 元硬币抛掷结果看作 5 角硬币抛掷结果的前提条件。假设 1 元硬币正面朝上时，5 角硬币正面朝上的概率是 60%，反面朝上的概率是 40%；1 元硬币反面朝上时，5 角硬币正面朝上的概率是 60%，反面朝上的概率是 40%。可以用表格表示条件概率：

条件	5 角正面朝上	5 角反面朝上
1 元正面朝上	60%	40%
1 元反面朝上	60%	40%

可以看到，1 元硬币抛掷结果，不影响 5 角硬币各面朝上的概率。“抛

掷 1 元硬币”这个实验的结果，不影响“抛掷 5 角硬币”实验的结果。

如果两个实验的结果互不影响，就说两者**相互独立**。一般来说，当我们讨论实验中发生的事件时，也说两个事件相互独立。



前提条件可能会改变不同事件发生的概率，也可能不会改变

如果事件 A, B 相互独立，那么 B 时 A 的概率应该和 A 的概率一样。也就是说，

$$\mathbb{P}(A) = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

即：

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

同样地，

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B).$$

多个事件之间也可能相互独立。给定 n 个事件 $A_1, A_2, \cdots, A_n$，如果它们的交集的概率等于各自概率的乘积：

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\cdots\mathbb{P}(A_n),$$

就说它们相互独立。

例题 6.3.1. 一共有 1000 个灯泡，其中有 k 个次品。从中随机抽取 10 个检验，发现有 3 个次品。k 最有可能是多少？

解答. 记事件“10 个抽检灯泡里有 3 个次品”为 Y。把不同的 k 作为条件，求出条件下发生事件 Y 的概率，然后比较找出使得条件概率最大的 k。

设 1000 个灯泡中有 k 个次品，那么事件 Y 的条件概率是:

$$\mathbb{P}(Y\,|\,k)=\frac{\mathrm{C}_k^3\cdot\mathrm{C}_{1000-k}^7}{\mathrm{C}_{1000}^{10}}=\frac{\frac{k!}{3!(k-3)!}\cdot\frac{(1000-k)!}{7!(993-k)!}}{\frac{1000!}{990!\cdot 10!}}=\frac{10!\cdot 990!}{3!\cdot 7!\cdot 1000!}\cdot\frac{k!(1000-k)!}{(k-3)!(993-k)!}$$

比较 $\mathbb{P}(Y\,|\,k)$ 和 $\mathbb{P}(Y\,|\,k+1)$ 的大小，可以发现两者比值为:

$$\frac{\mathbb{P}(Y\,|\,k)}{\mathbb{P}(Y\,|\,k+1)}=\frac{\frac{k!(1000-k)!}{(k-3)!(993-k)!}}{\frac{(k+1)!(999-k)!}{(k-2)!(992-k)!}}=\frac{(k-2)(1000-k)}{(k+1)(993-k)}$$

可以看到, k 很小 (比如 $k=3$) 的时候, 比值小于 1; k 很大 (比如 $k=992$) 的时候，比值大于 1。实际上，

$$\frac{(k-2)(1000-k)}{(k+1)(993-k)}-1=\frac{10k-2993}{(k+1)(993-k)}$$

因此 $k<300$ 时比值小于 1, 而 $k\geqslant 300$ 时比值大于 1。这说明概率 $\mathbb{P}(Y\,|\,k)$ 从 k 很小的时候开始不断增大，至 $k=300$ 时达到最大值，而从 $k=301$ 起逐渐减小。

因此，k 最有可能是 300。

习题 6.3.1.

1. 证明: $\mathbb{P}(A\,|\,B)+\mathbb{P}(A^{\complement}\,|\,B)=1$。

2. 证明: 如果事件 A_1,A_2 互斥，那么 $\mathbb{P}(A_1\cup A_2\,|\,B)=\mathbb{P}(A_1\,|\,B)+\mathbb{P}(A_2\,|\,B)$。

3. 证明: $\mathbb{P}(A)=\mathbb{P}(A\,|\,B)\mathbb{P}(B)+\mathbb{P}(A\,|\,B^{\complement})\mathbb{P}(B^{\complement})$。

4. 证明全概率公式: 如果 $B_1,B_2,\cdots,B_k$ 两两互斥, 且并集为全集, 那么 $\mathbb{P}(A)=\mathbb{P}(A\,|\,B_1)\mathbb{P}(B_1)+\mathbb{P}(A\,|\,B_2)\mathbb{P}(B_2)+\cdots+\mathbb{P}(A\,|\,B_k)\mathbb{P}(B_k)$。

5. 证明: $\mathbb{P}(B\,|\,A)=\dfrac{\mathbb{P}(A\,|\,B)\mathbb{P}(B)}{\mathbb{P}(A\,|\,B)\mathbb{P}(B)+\mathbb{P}(A\,|\,B^{\complement})\mathbb{P}(B^{\complement})}$。

6. 证明反条件概率公式: 如果 $B_1,B_2,\cdots,B_k$ 两两互斥，且并集为全集，那么

$$\forall i\in[1..k],\ \ \mathbb{P}(B_i\,|\,A)=\frac{\mathbb{P}(A\,|\,B_i)\mathbb{P}(B_i)}{\mathbb{P}(A\,|\,B_1)\mathbb{P}(B_1)+\mathbb{P}(A\,|\,B_2)\mathbb{P}(B_2)+\cdots+\mathbb{P}(A\,|\,B_k)\mathbb{P}(B_k)}.$$

7. 抛掷两枚硬币，假设两次抛掷互不影响，每枚硬币抛出正反的概率相同。定义 A 为事件“第一枚抛出正面”，B 为“第二枚抛出正面”，C 为“两枚硬币抛掷结果相同”。

7.1. 求 A、B、C 的概率。

7.2. 求 $A\cap B$、$A\cap C$、$B\cap C$ 的概率。

7.3. A,B 是否相互独立？B,C 是否相互独立？$A.C$ 是否相互独立？A,B,C 是否相互独立？

6.4 大数定律

我们做这样一个复合实验：取同一个骰子，投掷 100 次，每次为一次单一实验。记录每次朝上的点数作，按其 10 倍给出收益，作为结果。我们假定投掷结果是两两相互独立的事件，即前面投掷的结果不影响后面投掷的结果，反之亦然。我们还假定每次投掷的实验条件都是相同的，忽略投掷姿势、高度等差别。在这样的假定下，我们说，这 100 次投掷是 100 次**独立重复实验**。

独立重复实验是寻找随机事件规律的常用手段。比如，我们对投掷这枚骰子时各面朝上的概率一无所知，只有一份 100 次独立重复的投掷实验的结果。这时，如果要讨论下一次投掷的结果，我们如何给出合理的预测呢？

对于这 100 次独立重复实验，我们可以定义随机变量 $f_1, f_2, \cdots, f_{100}$ 来表示各次实验。由于我们假设这些随机变量相互独立，而且遵循相同的分布，我们说这 100 个随机变量是**独立同分布**的。

对于独立同分布的一组随机变量，我们有以下的定理：

定理 6.4.1. 大数定律设一组随机变量 $f_1, f_2, \cdots, f_N$ 相互独立，且都遵循同

一个分布，它们的期望同为实数 u，那么 N 越大，它们的平均数:

$$\frac{1}{N}\left(f_1 + f_2 + \cdots + f_N\right)$$

就几乎总会越来越接近 u。

大数定律是长久以来人们观察得出的一系列结论的总称，表示事情重复发生的数量足够多的时候，就会呈现一定的规律。后来人们找到了合适的数学语言来描述这些定律，发现大数定律是一类数学定理。大数定律的证明较为复杂，这里我们不作证明，只给出其中一个结论。

运用大数定律，我们可以为独立重复实验的结果给出什么结论呢?

大数定律建立了沟通实验结果平均数和随机变量期望的桥梁。已知独立重复实验的结果，那么对结果求平均数，就能估计出随机变量的期望。而随机变量的期望，可以用来估计下一次实验的结果。

以上面投掷骰子的实验为例。设 100 次投掷实验的结果为 $Y_1, Y_2, \cdots, Y_{100}$。它们分别是 100 个独立同分布的随机变量 $f_1, f_2, \cdots, f_{100}$ 的值。设下一次投掷实验的收益用独立同分布的随机变量 f 表示。根据大数定律，平均数

$$\overline{Y}_{100} = \frac{1}{100}(Y_1 + Y_2 + \cdots + Y_{100})$$

接近 f 的期望。因此，我们可以用 $\overline{Y}_{100}$ 代表 f 的期望。比如，如果算得 $\overline{Y}_{100} = 34.4$，那么可以估计下一次投掷实验的预期收益是 34.4。

再来看掷硬币实验，它对应的随机变量 h 将正面朝上映射到 1，反面朝上映射到 0。已知正反面朝上的概率 $p_{\text{正}}$ 和 $p_{\text{反}}$，可以算得 h 的期望为:

$$\mathbb{E}(h) = 1 \cdot p_{\text{正}} + 0 \cdot p_{\text{反}} = p_{\text{正}}.$$

所以，h 的期望就是 $p_{\text{正}}$。因此，如果投掷 1000 次硬币，其中 487 次正面朝上，那么 h 的期望就接近于 $\frac{487}{1000} = 0.487$。于是我们可以估计 $p_{\text{正}} = 0.487$。再掷一次硬币，正面朝上的概率是 48.7%。

这个实验中，我们用随机变量 h“揭示”了概率 $p_{正}$ 的性质，我们把这样设计的随机变量称为（$p_{正}$ 的）**示性随机变量**。“示性”这个词，我们将在不少地方碰到，指用来揭示某些对象性质的数学工具。

从掷硬币实验的例子，还可以得到估计某事件概率的方法。如果能够作独立重复实验来观察事件 A 是否发生，那么我们把 n 次独立重复实验中事件 A 发生的次数与试验次数 n 的比值称为该实验中 A 的**频率**，记为 q_A：

$$q_A = \frac{\text{事件}A\text{发生的次数}}{\text{总试验次数}}.$$

考虑把“事件 A 发生”映射到 1，把“A 不发生”映射到 0 的示性随机变量 f_A，那么频率 q_A 就是 n 个与 f_A 同分布的独立随机变量的平均数。根据大数定律，n 很大的时候，q_A 就接近 f_A 的期望，也就是 $\mathbb{P}(A)$。于是，通过记录多次实验中事件的频率，我们就可以估计事件发生的概率。

思考 6.4.1.

1. 投掷一枚骰子，如果朝上的点数为 i，则将 $(2i-7)^2$ 作为结果。如果某骰子 1 向上的概率为 50%，其余点数朝上概率为 10%；另一枚骰子 6 向上的概率为 50%，其余点数朝上概率为 10%。设第一枚骰子的结果为随机变量 g_1，第二枚为 g_2，这两个随机变量是否同分布?

2. 大数定律说明，独立重复实验次数足够多时，随机变量的期望可以用结果的平均数来估计。那么，独立重复实验次数足够多时，随机变量的方差能否用结果来估计呢? 你有什么想法?

习题 6.4.1.

1. 将一枚硬币重复抛掷 200 次，其中 104 次正面向上。如何估计下一次抛掷时正面向上的概率?

2. 将一枚骰子重复抛掷 1000 次，统计每次正面朝上的点数为结果。统计发现有 169 次 1 点、142 次 2 点、182 次 3 点、163 次 4 点、170 次 5 点、174 次 6 点。

2.1. 试估计下一次抛掷时得到 5 点的概率是多少?

2.2. 试估计下一次抛掷时得到偶数点的概率是多少?

2.3. 试估计下一次抛掷时点数不超过 3 的概率是多少?

2.4. 试估计单次掷骰子结果的分布。