$n_i/n \to p_i =$ 常数，

$$\frac{1}{n}\ln W \to -\sum_{i=1}^{m} p_i \ln p_i = H(p_1, \cdots, p_m), \tag{11.29}$$

因此，此游戏最有可能产生的概率分配就是在给定信息 I 时具有最大熵的分配.

你可能会反对说这个游戏仍然不完全“公平”，因为我们得到第一个可以接受的结果后就停止了，看不到其他可以接受的结果. 为了消除这一反对意见，可以考虑所有可能的可接受分布并选择它们的平均 $\overline{p_i}$. 这里“大数定律”能起到作用. 我们以练习的方式留给读者证明在大 n 的极限中，这个游戏中可以产生的所有可接受的概率分配中的绝大多数会无限接近于最大熵分布. ①

从概念上讲，沃利斯推导非常有吸引力. 它完全独立于香农的函数方程 (11.8)，不需要关于概率和频率之间存在联系的任何假设，也不假设不同可能性 $\{1, \cdots, m\}$ 本身是可重复随机试验的结果. 此外，它自动导致了 H 的最大化——而不是以其他方式处理——而无须用诸如“不确定性程度”这样的模糊概念对 H 进行准哲学解释. 因此，任何接受所提出的游戏作为分配不由先验信息确定的概率的公平方法的人都不可避免地导致最大熵原理.

让我们强调这一点. 试图将太多的哲学意义赋予导致 (11.23) 的定理是一个很大的错误. 特别地，回顾起来，用“熵”来表达信息似乎是很不幸的，因为它始终会给很多人带来错误的暗示. 香农本人对他的工作可能导致的反应有着先见之明，在提出定理之后，他马上指出该理论没有必要遵循，试图淡化它. 他的意思是说：H 满足的不等式已经足以证明使用的合理性，其实并不需要定理的进一步支持. 而该定理是从直观表示“不确定性程度”性质的函数方程中推导出来的.

尽管这是完全正确的，但我们现在想表明：如果确实接受熵作为由概率分布表示的“不确定性程度”的正确表达式，这将导致总体上更加统一的概率论. 这将使我们把无差别原则、概率与频率的诸多联系看作单一原理的特殊情况. 而统计力学、通信理论和大量其他应用是单一推理方法的应用实例.

11.5 一个示例

让我们来看看该原理在以上讨论的示例中如何工作，以检验该原理. 在该示例中，m 仅可以采用的值 $1, 2, 3, \overline{m}$ 是给定的. 我们可以再次使用拉格朗日乘子方法来解决此问题，与 (11.2) 中一样，

$$\delta\left[H - \lambda\sum_{m=1}^{3} m p_m - \mu\sum_{m=1}^{3} p_m\right] = \sum_{m=1}^{3}\left[\frac{\partial H}{\partial p_m} - \lambda m - \mu\right]\delta p_m = 0. \tag{11.30}$$

① 这一结果将通过后面要给出的熵集中定理更完整地形式化.

现在我们有

$$\frac{\partial H}{\partial p_m} = -\ln p_m - 1, \tag{11.31}$$

所以我们的解是

$$p_m = \mathrm{e}^{-\lambda_0 - \lambda_m}, \tag{11.32}$$

其中 $\lambda_0 \equiv \mu + 1$.

因此，在给定平均值的条件下，最大熵分布将呈指数形式. 我们需要拟合常数 λ_0 和 λ 来满足 p 的总和为 1 且期望值等于指定平均值 $\overline{m}$ 的约束. 可以通过定义函数

$$Z(\lambda) \equiv \sum_{m=1}^{3} \mathrm{e}^{-\lambda m}, \tag{11.33}$$

来轻松完成此任务. 我们在第 9 章中称其为*分拆函数*. 固定我们的拉格朗日乘子的 (11.3) 和 (11.4) 使用形式

$$\lambda_0 = \ln Z(\lambda), \tag{11.34}$$

$$\overline{m} = -\frac{\partial \ln Z(\lambda)}{\partial \lambda}. \tag{11.35}$$

我们发现参数形式的 $p_1(\overline{m}), p_2(\overline{m}), p_3(\overline{m})$ 值为

$$p_k = \frac{\mathrm{e}^{-k\lambda}}{\mathrm{e}^{-\lambda} + \mathrm{e}^{-2\lambda} + \mathrm{e}^{-3\lambda}} = \frac{\mathrm{e}^{(3-k)\lambda}}{\mathrm{e}^{2\lambda} + \mathrm{e}^{\lambda} + 1}, \qquad k = 1, 2, 3, \tag{11.36}$$

$$\overline{m} = \frac{\mathrm{e}^{2\lambda} + 2\mathrm{e}^{\lambda} + 3}{\mathrm{e}^{2\lambda} + \mathrm{e}^{\lambda} + 1}. \tag{11.37}$$

在更复杂的问题中，我们需要将其保留为参数形式，但是在这种特殊情况下，可以消去参数 λ，从而得到显式解

$$\begin{aligned} p_1 &= \frac{3 - \overline{m} - p_2}{2}, \\ p_2 &= \frac{1}{3}\left[\sqrt{4 - 3(\overline{m} - 2)^2} - 1\right], \\ p_3 &= \frac{\overline{m} - 1 - p_2}{2}. \end{aligned} \tag{11.38}$$

作为 $\overline{m}$ 的函数，p_2 是椭圆的弧，在端点有单位斜率. p_1 和 p_3 也是椭圆的弧，但以两种不同的方式倾斜.

我们终于有了一个满足前面两个标准的解. 最大熵分布 (11.36) 自然具有 $p_k \geqslant 0$ 的性质，因为对数的奇点为 0，这是我们永远无法越过的. 此外，它还具有以下特征：不允许机器人将零概率分配给可能假设，除非有证据表明其概率为 0.[①] 概率为 0 的情况是 $\overline{m}$ 恰好为 1 或 3 的极限情况. 当然，在这种极限情况

① 戴维·布莱克韦尔强调了此性质，他认为这是分配概率的合理程序的最基本要求.

下，无论我们使用什么原理，根据演绎推理，某些概率的确都必须为 0.

11.6 推广：更严格的证明

最大熵解可以以多种方式推广. 假设变量 x 可以有 n 个不同的离散值 $(x_1, \cdots, x_n)$，它们对应于 n 个不同的命题 $(A_1, \cdots, A_n)$，x 有 m 个不同的函数

$$f_k(x), \quad 1 \leqslant k \leqslant m < n, \tag{11.39}$$

并且我们希望它们有期望值

$$\langle f_k(x) \rangle = F_k, \quad 1 \leqslant k \leqslant m, \tag{11.40}$$

其中 $\{F_k\}$ 是问题陈述中给我们的值. 机器人会分配什么概率 $(p_1, \cdots, p_n)$ 给 $(x_1, \cdots, x_n)$ 呢？我们有

$$F_k = \langle f_k(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i), \tag{11.41}$$

为了找到具有满足所有这些约束条件的最大熵的 p_i 集合，我们引入多个拉格朗日算子作为约束：

$$\begin{aligned} 0 &= \delta \left[H - (\lambda_0 - 1) \sum_i p_i - \sum_{j=1}^{m} \lambda_j \sum_i p_i f_j(x_i) \right] \\ &= \sum_i \left[\frac{\partial H}{\partial p_i} - (\lambda_0 - 1) - \sum_{j=1}^{m} \lambda_j f_j(x_i) \right] \delta p_i. \end{aligned} \tag{11.42}$$

根据 (11.23)，我们的解为

$$p_i = \exp\left\{ -\lambda_0 - \sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}, \tag{11.43}$$

像往常一样，它是约束的指数函数. 所有概率的和必须为 1，因此

$$1 = \sum_i p_i = \exp\{-\lambda_0\} \sum_i \exp\left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}. \tag{11.44}$$

如果我们现在定义分拆函数

$$Z(\lambda_1, \cdots, \lambda_m) \equiv \sum_{i=1}^{m} \exp\left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}, \tag{11.45}$$

那么 (11.44) 化为

$$\lambda_0 = \ln Z(\lambda_1, \cdots, \lambda_m). \tag{11.46}$$

在概率范围内，平均值 F_k 必须等于 $f_x(x)$ 的期望值

$$F_k = \exp\{-\lambda_0\} \sum_i f_k(x_i) \exp\left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}, \tag{11.47}$$

或者

$$F_k = -\frac{\partial \ln Z(\lambda_1, \cdots, \lambda_m)}{\partial \lambda_k}. \tag{11.48}$$

熵的最大值是

$$H_{\max} = \left[-\sum_{i=1}^{m} p_i \ln p_i \right]_{\max}, \tag{11.49}$$

根据 (11.43)，我们得到

$$H_{\max} = \lambda_0 + \sum_{j=1}^{m} \lambda_j F_j. \tag{11.50}$$

现在，这些结果有许多新的应用，因此有尽可能严格的证明非常重要. 但是，我们刚才通过变分法解决最大化问题并不是 100% 严格的. 我们的拉格朗日乘数乘子方法具有能立即给出答案的优点；然而它也有一个不好的方面，那就是我们完成后不确定答案是否正确. 假设我们要定位一个函数的最大值，该函数的全局最大值恰好发生在尖点（斜率不连续的点）而不是圆点. 变分法将找到一些圆滑的局部极大值，但找不到全局最大尖点. 即使我们已经证明达到了通过变分法所能达到的最大值，但函数仍然有可能在某些点处具有更大的值，而这是我们无法通过变分法找到的. 如果我们只使用变分法，总会有一点儿疑问.

因此，现在我们给出一个完全不同的推导方法. 该方法可以弥补变分法的不足之处. 为此，我们需要一个引理. 令 p_i 为可能概率分布的任何数值集合，即

$$\sum_{i=1}^{n} p_i = 1, \quad p_i \geqslant 0. \tag{11.51}$$

令 u_i 为另一个可能的概率分布，

$$\sum_{i=1}^{n} u_i = 1, \quad u_i \geqslant 0. \tag{11.52}$$

现在，我们有

$$\ln x \leqslant x - 1, \quad 0 \leqslant x < +\infty, \tag{11.53}$$

当且仅当 $x = 1$ 时等号成立. 因此

$$\sum_{i=1}^{n} p_i \ln \frac{u_i}{p_i} \leqslant \sum_{i=1}^{n} p_i \left(\frac{u_i}{p_i} - 1 \right) = 0, \tag{11.54}$$

或者

$$H(p_1, \cdots, p_n) \leqslant \sum_{i=1}^{n} p_i \ln \frac{1}{u_i}, \tag{11.55}$$

当且仅当 $p_i = u_i$（$i = 1, \cdots, n$）时等号成立. 这是我们需要的引理.

现在我们简单、随意地定义分布 u_i 如下：

$$u_i \equiv \frac{1}{Z(\lambda_1, \cdots, \lambda_m)} \exp\left\{ -\sum_{j=1}^{m} \lambda_j f_j(x_i) \right\}, \tag{11.56}$$

其中 $Z(\lambda_1, \cdots, \lambda_m)$ 由 (11.45) 定义. 为什么要以这种特定方式选择 u_i 呢? 一会儿我们就会明白为什么. 现在可以将不等式 (11.55) 写成

$$H \leqslant \sum_{i=1}^{n} p_i \left[\ln Z(\lambda_1, \cdots, \lambda_m) + \sum_{j=1}^{m} \lambda_j f_j(x_i) \right], \tag{11.57}$$

或者

$$H \leqslant \ln Z(\lambda_1, \cdots, \lambda_m) + \sum_{j=1}^{m} \lambda_j \langle f_j(x) \rangle. \tag{11.58}$$

现在让 p_i 在满足约束 (11.41) 的所有可能概率分布上变化，(11.58) 的右侧保持不变. 我们的引理表明，当且仅当选择 p_i 为规范分布 (11.56) 时，H 才能达到其绝对最大值 $H_{\max}$，从而使 (11.58) 中的等号成立.

这是严格的证明，不会有我们尝试把问题当作变分问题求解时可能发生的问题. 正如我们所看到的，该论证在变分法很弱时是强的，而在变分法很强时比较弱，因为我们在得到分布 (11.56) 时需要将解从中抽出来. 我们必须先知道解才能证明它. 如果同时拥有两种论证，那么整个故事就完整了.

11.7 最大熵分布的形式性质

现在我们要列出规范分布 (11.56) 的形式性质. 从某种意义上讲，这不是一种好方法，因为它听起来很抽象，我们看不到它与实际问题的联系. 另外，如果首先了解理论中的所有形式性质，我们就能更快地理解这一理论. 然后，当讨论特定的物理问题时，我们就会发现这些形式关系中的每一种对不同的问题有不同的意义.

固定平均值时所能达到的最大 H 值当然依赖于我们指定的平均值，

$$H_{\max} = S(F_1, \cdots, F_m) = \ln Z(\lambda_1, \cdots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k. \tag{11.59}$$

我们可以将 H 视为任何概率分布中“不确定性程度”的度量. 最大化后，它成为问题中确定数据 $\{F_i\}$ 的函数，因此我们将其称为最大 $S(F_1, \cdots, F_m)$，以期在物理学中得到初始应用. 它仍然是“不确定性”的量度，但是它是当我们仅拥有这些数字信息时的不确定性. 从某种意义上说，它完全是“客观的”，因为它仅取决于问题的给定数据，而不取决于任何人的性格或意愿.

如果 S 仅是 $(F_1,\cdots,F_m)$ 的函数，则在 (11.59) 中 $Z(\lambda_1,\cdots,\lambda_m)$ 也必须被认为是 $(F_1,\cdots,F_m)$ 的函数. 最初这些 λ 只是未确定的拉格朗日乘子，但最终我们想确定它们. 如果选择不同的 λ_i，就是在选择不同的概率分布 (11.56). 我们在 (11.48) 中看到，如果

$$F_k = \langle fk\rangle = -\frac{\partial \ln Z(\lambda_1,\cdots,\lambda_m)}{\partial \lambda_k}, \quad k=1,2,\cdots,m, \tag{11.60}$$

这些分布的平均值与给定的平均值 F_k 相符. (11.60) 是 m 个联立的非线性方程组，必须根据 F_k 对 λ 求解. 通常，在非平凡的问题中，显式地求解 λ 是不切实际的 [尽管下面有一个简单的形式解 (11.62)]. 我们将保留 λ_k，以参数形式表示所需要的东西. 实际上，这并不是悲剧，因为 λ 通常具有重要的物理意义，因此我们很高兴将其作为自变量. 但是，如果我们可以显式计算函数 $S(F_1,\cdots,F_m)$，则可以将 λ 作为 $\{F_k\}$ 的显式函数给出如下.

假设我们对 F_k 之一进行小扰动，这将如何改变最大可达到的 H 呢? 根据 (11.59) 可以得到

$$\frac{\partial S(F_1,\cdots,F_m)}{\partial F_k} = \sum_{j=1}^{m}\left[\frac{\partial \ln Z(\lambda_1,\cdots,\lambda_m)}{\partial \lambda_j}\right]\left[\frac{\partial \lambda_j}{\partial F_k}\right] + \sum_{j=1}^{m}\frac{\partial \lambda_j}{\partial F_k}F_k + \lambda_k, \tag{11.61}$$

鉴于 (11.60)，这简化为

$$\lambda_k = \frac{\partial S(F_1,\cdots,F_m)}{\partial F_k}, \tag{11.62}$$

其中明确给出了 λ_k.

将该式与 (11.60) 比较：一个根据 λ_k 明确给出 F_k，另一个根据 F_k 明确给出 λ_k. 这表明指定 $\ln Z(\lambda_1,\cdots,\lambda_m)$ 或 $S(F_1,\cdots,F_m)$ 是等效的，因为每个都给出了有关概率分布的完整信息. 实际上 (11.59) 只是从一种表征函数转化为另一表征函数的勒让德变换.

通过对 (11.60) 或 (11.62) 进行微分，我们可以得出一些更有趣的定律. 因为 $\ln Z(\lambda_1,\cdots,\lambda_m)$ 的二阶交叉导数在 j 和 k 中是对称的，如果我们将 (11.60) 对 λ_j 进行微分，则可以得到

$$\frac{\partial F_k}{\partial \lambda_j} = \frac{\partial^2 \ln Z(\lambda_1,\cdots,\lambda_m)}{\partial \lambda_j \partial \lambda_k} = \frac{\partial F_j}{\partial \lambda_k}. \tag{11.63}$$

这是一个对通过熵最大化来解决的任何问题都成立的通用互反定律. 同样，如果对 (11.62) 再次进行微分，可以得到

$$\frac{\partial \lambda_k}{\partial F_j} = \frac{\partial^2 S}{\partial F_j \partial F_k} = \frac{\partial \lambda_j}{\partial F_k}, \tag{11.64}$$

这是另一个互反定律，但它并不独立于 (11.63)，因为如果我们通过 $A_{jk}=\partial\lambda_j/\partial F_k$ 和 $B_{jk}=\partial F_j/\partial\lambda_k$ 定义矩阵，容易明白它们互为逆矩阵：$\boldsymbol{A}=\boldsymbol{B}^{-1}$，$\boldsymbol{B}=\boldsymbol{A}^{-1}$. 这些互逆定律很容易得到，可能显得微不足道. 但是当我们研究实际应用时，会发现它们具有非凡和并不显而易见的物理含义. 过去，其中一些定律是通过烦琐的方式得到的，使得它们显得神秘而晦涩.

现在，我们考虑函数 $f_k(x)$ 之一包含可变参数 α 的可能性. 如果要考虑应用，可以说 $f_k(x_i;\alpha)$ 代表某个系统的第 i 个能级，α 代表该系统的体积，能级取决于体积. 或者，如果它是一个磁共振系统，我们可以说 $f_k(x_i)$ 代表自旋系统的第 i 个稳态的能量，α 代表施加于其上的磁场. 通常，我们想要预测随着 α 的变化某些量会如何变化. 我们可能想要计算压力或磁化率. 根据最小均方误差准则，导数的最优估计将是概率分布的均值

$$\left\langle\frac{\partial f_k}{\partial\alpha}\right\rangle=\frac{1}{Z}\sum_i\exp\{-\lambda_1 f_1(x_i)-\cdots-\lambda_k f_k(x_i;\alpha)\\ -\cdots-\lambda_m f_m(x_i)\}\frac{\partial f_k(x_i;\alpha)}{\partial\alpha}, \tag{11.65}$$

这可以简化为

$$\left\langle\frac{\partial f_k}{\partial\alpha}\right\rangle=-\frac{1}{\lambda_k}\frac{\partial\ln Z(\lambda_1,\cdots,\lambda_m;\alpha)}{\partial\alpha}. \tag{11.66}$$

在这个推导中，我们假设 α 只出现在一个函数 f_k 中. 如果相同的参数出现在几个不同的 f_k 中，容易验证结论可以推广为

$$\sum_{k=1}^m\lambda_k\left\langle\frac{\partial f_k}{\partial\alpha}\right\rangle=-\frac{\partial\ln Z(\lambda_1,\cdots,\lambda_m;\alpha)}{\partial\alpha}. \tag{11.67}$$

该一般规则包含任何热力学系统的状态方程等.

当我们将 α 添加到问题中时，$Z(\lambda_1,\cdots,\lambda_m;\alpha)$ 和 $S(F_1,\cdots,F_k;\alpha)$ 都成为 α 的函数. 如果对 $\ln Z(\lambda_1,\cdots,\lambda_m;\alpha)$ 或 $S(F_1,\cdots,F_k;\alpha)$ 求导，将得到相同的结果：

$$\frac{\partial S(F_1,\cdots,F_k;\alpha)}{\partial\alpha}=-\sum_{k=1}^m\lambda_k\left\langle\frac{\partial f_k}{\partial\alpha}\right\rangle=\frac{\partial\ln Z(\lambda_1,\cdots,\lambda_m;\alpha)}{\partial\alpha}, \tag{11.68}$$

复杂之处是：我们必须理解，在 (11.68) 中，对于导数 $\partial S(F_1,\cdots,F_m;\alpha)/\partial\alpha$，我们保持 F_k 固定；对于导数 $\partial\ln Z(\lambda_1,\cdots,\lambda_m;\alpha)/\partial\alpha$，我们保持 λ_k 固定. 然后根据勒让德变换 (11.59) 得出这两个导数的相等性. 显然，如果在这个问题中有几个不同的参数 $\{\alpha_1,\alpha_2,\cdots,\alpha_r\}$，对于它们中的每一个，形如 (11.68) 的关系都成立.

现在，让我们得出一些一般的“波动定律”或矩定理. 首先对符号做一些说明：我们使用 F_k 和 $\langle f_k\rangle$ 代表相同的数. 它们是相等的，因为我们指定期望值 $\{\langle f_1\rangle,\cdots,\langle f_m\rangle\}$ 等于给定数据 $\{F_1,\cdots,F_m\}$. 当我们要强调这些数是规范分

布 (11.56) 上的期望值时，将使用符号 $\langle f_k\rangle$；当我们想强调它们是给定的数据时，将其称为 F_k. 现在我们想强调前者，所以互反定律 (11.63) 可以写成

$$\frac{\partial\langle f_k\rangle}{\partial\lambda_j}=\frac{\partial\langle f_j\rangle}{\partial\lambda_k}=\frac{\partial^2\ln Z(\lambda_1,\cdots,\lambda_m)}{\partial\lambda_j\partial\lambda_k}. \tag{11.69}$$

在改变 λ 时，我们从规范分布 (11.56) 变为一种略有不同的分布，其中 $\langle f_k\rangle$ 略有不同. 由于对应于 $(\lambda_k+d\lambda_k)$ 的新分布仍然是规范形式，它是对应于略有不同的数据 (F_k+dF_k) 的最大熵分布. 因此，我们正在比较两个略有不同的最大熵问题. 为了以后的物理应用，在解释互反定律 (11.69) 时很重要的是要认识到这一点.

现在我们要证明 (11.69) 中的量对于单个最大熵问题也具有重要意义. 在规范分布 (11.56) 中，不同量 $f_k(x)$ 如何相互关联？更具体地说，与平均值 $\langle f_k\rangle$ 的偏离如何关联？该度量是分布的协方差或第二中心矩：

$$\begin{aligned}\big\langle(f_j-\langle f_j\rangle)(f_k-\langle f_k\rangle)\big\rangle&=\big\langle f_jf_k-f_j\langle f_k\rangle-\langle f_j\rangle f_k+\langle f_j\rangle\langle f_k\rangle\big\rangle\\&=\langle f_jf_k\rangle-\langle f_j\rangle\langle f_k\rangle.\end{aligned} \tag{11.70}$$

如果大于平均值 $\langle f_k\rangle$ 的 f_k 值可能伴随有大于其平均值 $\langle f_j\rangle$ 的 f_j 值，则协方差为正；如果它们倾向于在相反的方向波动，则协方差为负；如果它们的变化不相关，则协方差为 0. 如果 $j=k$，这就变成*方差*：

$$\big\langle(f_k-\langle f_k\rangle)^2\big\rangle=\langle f_k^2\rangle-\langle f_k\rangle^2\geqslant 0. \tag{11.71}$$

要直接从规范分布 (11.56) 计算这些量，我们可以首先计算

$$\begin{aligned}\langle f_jf_k\rangle&=\frac{1}{Z(\lambda_1,\cdots,\lambda_m)}\sum_{i=1}^{n}f_j(x_i)f_k(x_i)\exp\left\{-\sum_{j=1}^{m}\lambda_jf_j(x_i)\right\}\\&=\frac{1}{Z(\lambda_1,\cdots,\lambda_m)}\sum_{i=1}^{n}\frac{\partial^2}{\partial\lambda_j\partial\lambda_k}\exp\left\{-\sum_{j=1}^{m}\lambda_jf_j(x_i)\right\}\\&=\frac{1}{Z(\lambda_1,\cdots,\lambda_m)}\frac{\partial^2Z(\lambda_1,\cdots,\lambda_m)}{\partial\lambda_j\partial\lambda_k},\end{aligned} \tag{11.72}$$

然后应用 (11.60)，协方差变为

$$\langle f_jf_k\rangle-\langle f_j\rangle\langle f_k\rangle=\frac{1}{Z}\frac{\partial^2Z}{\partial\lambda_j\partial\lambda_k}-\frac{1}{Z^2}\frac{\partial Z}{\partial\lambda_j}\frac{\partial Z}{\partial\lambda_k}=\frac{\partial^2\ln Z}{\partial\lambda_j\partial\lambda_k}. \tag{11.73}$$

但是这只是量 (11.69)，因此互反定律有更大的意义：

$$\langle f_jf_k\rangle-\langle f_j\rangle\langle f_k\rangle=-\frac{\partial\langle f_j\rangle}{\partial\lambda_k}=-\frac{\partial\langle f_k\rangle}{\partial\lambda_j}. \tag{11.74}$$

为我们提供了互反定律的 $\ln Z(\lambda_1,\cdots,\lambda_m)$ 的二阶导数也给出了我们分布中的 f_j 和 f_k 的协方差.

注意，(11.74) 仅是更一般规则的特例. 令 $q(x)$ 为任意函数，容易验证 $q(x)$ 与 $f_k(x)$ 的协方差为

$$\langle qf_k\rangle - \langle q\rangle\langle f_k\rangle = -\frac{\partial\langle q\rangle}{\partial\lambda_k}. \tag{11.75}$$

练习 11.3 通过比较 (11.60) (11.69) 和 (11.74)，我们可以期望 $\ln Z(\lambda_1,\cdots,\lambda_m)$ 的更高阶导数对应于分布 (11.56) 的更高阶中心矩. 通过计算 $\ln Z(\lambda_1,\cdots,\lambda_m)$ 的第三和第四中心矩来验证这一猜想是否成立.
提示：有关累积量的理论见附录 C.

对于非中心矩，习惯上定义矩母函数

$$\Phi(\beta_1,\cdots,\beta_m) \equiv \left\langle \exp\left\{\sum_{j=1}^{m}\beta_j f_j\right\}\right\rangle, \tag{11.76}$$

它显然具有性质

$$\left\langle f_i^{m_i} f_j^{m_j}\cdots\right\rangle = \left(\frac{\partial^{m_i}}{\partial\beta_i^{m_i}}\frac{\partial^{m_j}}{\partial\beta_j^{m_j}}\cdots\right)\Phi(\beta_1,\cdots,\beta_m)\Big|_{\beta_k=0}. \tag{11.77}$$

由 (11.76) 可以得到

$$\Phi(\beta_1,\cdots,\beta_m) = \frac{Z([\lambda_1-\beta_1],\cdots,[\lambda_m-\beta_m])}{Z(\lambda_1,\cdots,\lambda_m)}, \tag{11.78}$$

因此，分拆函数 $Z(\lambda_1,\cdots,\lambda_m)$ 可以达到此目的. 不同于 (11.77)，我们可以得到

$$\left\langle f_i^{m_i} f_j^{m_j}\cdots\right\rangle = \frac{1}{Z(\lambda_1,\cdots,\lambda_m)}\left(\frac{\partial^{m_i}}{\partial\beta_i^{m_i}}\frac{\partial^{m_j}}{\partial\beta_j^{m_j}}\cdots\right)Z(\lambda_1,\cdots,\lambda_m), \tag{11.79}$$

这是 (11.72) 的推广.

现在，我们可能会问：f_k 的导数相对于参数 α 的协方差是多少？定义

$$g_k \equiv \frac{\partial f_k}{\partial\alpha}. \tag{11.80}$$

如果 f_k 是能量，α 是体积，则 $-g_k$ 是压力. 我们可以轻松地验证另一个互反关系：

$$\frac{\partial\langle g_j\rangle}{\partial\lambda_k} = -\left[\langle g_j f_k\rangle - \langle g_j\rangle\langle g_k\rangle\right] = \frac{\partial\langle g_k\rangle}{\partial\lambda_j}, \tag{11.81}$$

这类似于 (11.74). 通过类似的推导可以得到等式

$$\sum_{j=1}^{m}\lambda_j\left[\langle g_j g_k\rangle - \langle g_j\rangle\langle g_k\rangle\right] = \left\langle\frac{\partial g_k}{\partial\alpha}\right\rangle - \frac{\partial\langle g_k\rangle}{\partial\alpha}. \tag{11.82}$$

在意识到其通用性之前，我们已经发现并使用了一些特殊情况.

$\ln Z(\lambda_1,\cdots,\lambda_m)$ 的其他导数与 f_k 及其相对于 α 的导数的各阶矩有关. 比如，与 (11.82) 密切相关的是

$$\frac{\partial^2 \ln Z(\lambda_1,\cdots,\lambda_m)}{\partial\alpha^2} = \sum_{jk}\lambda_j\lambda_k\big[\langle g_j g_k\rangle - \langle g_j\rangle\langle g_k\rangle\big] - \sum_k \lambda_k\left\langle\frac{\partial g_k}{\partial\alpha}\right\rangle. \tag{11.83}$$

二阶交叉导数是一个简单而有用的关系，

$$\frac{\partial^2 \ln Z(\lambda_1,\cdots,\lambda_m)}{\partial\alpha\partial\lambda_k} = -\frac{\partial\langle f_k\rangle}{\partial\alpha} = \sum_j \lambda_j\big[\langle f_k g_j\rangle - \langle f_k\rangle\langle g_j\rangle\big] - \langle g_k\rangle, \tag{11.84}$$

这也可以由 (11.69) 和 (11.75) 得到. 通过进一步求导，可以获得类似的无限层次的矩关系. 正如我们将在后面看到的那样，上述定理在特殊情况下具有我们熟悉的关系，例如关于黑体辐射和气体或液体密度的爱因斯坦波动定律、奈奎斯特电压波动定律或可逆电池产生的“噪声”定律，等等.

显然，如果不同参数 $\{\alpha_1,\cdots,\alpha_r\}$ 存在，以上关系将对它们每一个都成立. 新的关系，比如

$$\begin{aligned}\frac{\partial^2 \ln Z(\lambda_1,\cdots,\lambda_m)}{\partial\alpha_1\partial\alpha_2} &= \sum_k \lambda_k\left\langle\frac{\partial^2 f_k}{\partial\alpha_1\partial\alpha_2}\right\rangle \\ &\quad - \sum_{kj}\lambda_j\lambda_k\left[\left\langle\frac{\partial f_k}{\partial\alpha_1}\frac{\partial f_j}{\partial\alpha_2}\right\rangle - \left\langle\frac{\partial f_k}{\partial\alpha_1}\right\rangle\left\langle\frac{\partial f_j}{\partial\alpha_2}\right\rangle\right]\end{aligned} \tag{11.85}$$

也会出现.

$\ln Z(\lambda_1,\cdots,\lambda_m;\alpha_1,\cdots,\alpha_r)$ 与 $S(\langle f_1\rangle,\cdots,\langle f_m\rangle;\alpha_1,\cdots,\alpha_r)$ 的关系表明它们也都可以用 S 的导数（即变分性质）表示，见 (11.59). 但是对于 S 还有更一般的重要变分性质.

在 (11.62) 中，我们假设函数 $f_k(x)$ 的定义是固定的，而 $\langle f_k\rangle$ 的变化仅仅是由 p_i 的变化引起的. 现在我们将导出一个更一般的变分陈述，其中这两个量均发生变化. 针对 k 和 i 独立地随意指定 $\delta f_k(x_i)$，独立于 $\delta f_k(x_i)$ 指定 $\delta\langle f_k\rangle$，并考虑从最大熵分布 p_i 到一个稍微不同的分布 $p_i' = p_i + \delta p_i$ 的变化，通过上述方程，δp_i 和 $\delta\lambda_k$ 的变化将根据 $\delta f_k(x_i)$ 和 $\delta\langle f_k\rangle$ 确定地变化. 换句话说，我们现在正在考虑两个略有不同的最大熵问题，其中问题的所有条件（包括基础函数 $f_k(x)$ 的定义）都可以随意变化. $\ln Z(\lambda_1,\cdots,\lambda_m)$ 的变化为

$$\begin{aligned}\delta\ln Z(\lambda_1,\cdots,\lambda_m) &= \frac{1}{Z}\sum_{i=1}^n\left[\sum_{k=1}^m\big[-\lambda_k\delta f_k(x_i) - \delta\lambda_k f_k(x_i)\big]\exp\left\{-\sum_{j=1}^m\lambda_j f_j(x_i)\right\}\right] \\ &= -\sum_{k=1}^m\big[\lambda_k\langle\delta f_k\rangle + \delta\lambda_k\langle f_k\rangle\big],\end{aligned} \tag{11.86}$$

根据勒让德变换 (11.59)，

$$\delta S = -\sum_k \lambda_k \big[\delta\langle f_k\rangle - \langle \delta f_k\rangle\big] \quad \text{或者} \quad \delta S = \sum_k \lambda_k \delta Q_k, \tag{11.87}$$

其中

$$\delta Q_k \equiv \delta\langle f_k\rangle - \langle \delta f_k\rangle = \sum_{i=1}^{n} f_k(x_i)\delta p_i. \tag{11.88}$$

这一结果推广了 (11.62)，它表明熵 S 不仅在导致规范分布 (11.56) 最大化的意义上是稳定的，而且如果 p_i 保持固定，则熵对于函数 $f_k(x_i)$ 的微小变化也保持不变.

作为 (11.87) 的特例，假设函数 f_k 像 (11.85) 一样包含参数 $\{\alpha_1, \cdots, \alpha_r\}$，它们通过

$$\delta f_k(x_i, \alpha_j) = \sum_{j=1}^{r} \frac{\partial f_k(x_i, \alpha)}{\partial \alpha_j}\delta\alpha_j \tag{11.89}$$

生成 $\delta f_k(x_i)$. 虽然 δQ_k 通常不是任何函数 $Q_k(\langle f_i\rangle; \alpha_j)$ 的精确微分，但 (11.87) 表明 λ_k 是一个积分因子，使得 $\sum \lambda_k \delta Q_k$ 是“状态函数” $S(\langle f_i\rangle; \alpha_j)$ 的精确微分. 这一点在那些研究热力学的人来说看起来似乎很熟悉. 最后，我们留给读者根据 (11.87) 证明

$$\sum_{k=1}^{m} \langle f_k\rangle \frac{\partial \lambda_k}{\partial \alpha} = 0, \tag{11.90}$$

其中 $\langle f_1\rangle, \cdots, \langle f_r\rangle$ 在微分中保持不变.

显然，现在有一大类新问题可以让机器人来解决，它可以批量地解决这些问题. 它首先计算分拆函数 Z，或者最好是计算 $\ln Z$. 然后，通过以各种可能的方式对其所有参数对 $\ln Z$ 求微分，就可以得到最大熵分布的均值形式的各种预测. 这是一个非常简洁的数学过程，当然，大家会明白我们在这里所做的事情. 这些关系只是吉布斯带给我们的统计力学的标准方程，但是其中所有的物理学内容都被删除了，只留下数学形式.

实际上，几乎所有已知的热力学定律现在都被视为最大熵理论的简单数学恒等式的特例. 这些定律是一个多世纪以来通过多样化、复杂的物理实验和推理得到的. 这清楚地表明，这些关系实际上独立于任何特定的物理假设，是扩展逻辑的一般性质. 这使得我们对热力学的关系为何独立于任何特定物质的性质有了新的认识. 吉布斯的统计力学在历史上是最大熵原理的最早应用，并且至今仍然是使用得最多的（尽管它的许多应用者仍然不知道它的一般性）.

最大熵的数学形式在物理学之外还有大量其他应用. 在第 14 章中，我们将通过此方法为库存控制的非平凡问题提供完整的数值解；在第 22 章中，我们将

给出通信理论中最优编码问题的非平凡的解析解. 从某种意义上说，一旦我们理解了本章所述的最大熵原理，那么概率论的大多数应用能被视为是在使用它来分配初始概率——无论在技术上称为先验概率还是抽样概率. 每当我们分配均匀的先验概率时，我们都可以说在应用最大熵原理（尽管在这种情况下，结果是如此简单直观，因此我们不需要上述任何数学形式）. 正如我们在第 7 章中所看到的，每当分配高斯抽样分布时，这与给定第一和第二阶矩应用最大熵原理相同. 我们在第 9 章中看到，在分配二项抽样分布时，这在数学上等价于在更深的假设空间上分配均匀的最大熵分布.

11.8　概念问题-频率对应

最大熵原理相当简单明了. 正如我们刚刚看到的，在给定信息是平均值的情况下，如果可以计算函数 $Z(\lambda_1,\cdots,\lambda_m;\alpha_1,\cdots,\alpha_r)$，则一切都可以得到，所以它会产生非常简洁的数学形式，然而，这似乎会产生概念上的严重困难，特别是对于那些被训练成只会在频率意义上考虑概率的人来说. 因此，在转向应用之前，我们将研究并希望解决其中的一些困难. 以下是对最大熵原理的一些异议.

(A) 如果使用规范分布 (11.56) 的唯一理由是“最大化不确定性”，那就是消极的做法，不可能导致任何有用的预测，单单出于无知无法获得可靠的结果.

(B) 通过最大熵获得的概率与物理预测无关，因为它们与频率无关——绝对没有理由假设实验观察到的分布与通过最大熵发现的分布相符.

(C) 最大熵原理仅限于约束条件为平均值的情况，如果给定数据 $\{F_1,\cdots,F_n\}$ 几乎都不是任何事物的平均值，它们则是确定的测量值. 当你将它们设置为等于平均值 $F_k=\langle f_k\rangle$ 时，你就犯了逻辑矛盾，因为给定的数据说 f_k 的值为 F_k，但是你立即写下了一个概率分布，该概率分布将非零概率分配给 $f_k\neq F_k$ 的值.

(D) 因为不同的人有不同的信息，所以该原理不可能导致任何确定的物理结果，而是将导致不同的分布——结果基本上是任意的.

异议 (A) 当然只是语言游戏. “不确定性”一直存在. 我们最大化熵并不会产生任何“无知”或“不确定性”；相反，它能定量确定已经存在的不确定性的范围. 如果不这样做——意味着使用的分布信息比我们实际拥有的要多——将得出不可靠的结论.

当然，作为对我们最大熵分布的约束而放入理论中的信息可能太少了——分

布受极弱的无信息均匀分布的约束——以至于无法从中做出可靠的预测. 但是在这种情况下，正如我们稍后将看到的，该理论会自动告诉我们：如果某一个量（例如压力、磁化强度、电流密度、扩散速率等）呈现很宽的概率分布，这就是机器人告诉我们的："你没有给我足够的信息来做任何确定的预测."但是，如果我们得到一个非常尖锐的分布［例如——这也是许多实际问题中所发生的典型现象——理论上说 θ 在区间 θ_0（1 ± 10^{-6}）中的几率大于 $10^{10}:1$］，则给定的信息足以做出非常明确的预测.

在两种极端情况以及其他中间情况下，θ 的分布总是根据方程中的输入信息告诉我们关于 θ 能得出哪些结论. 如果有人有其他可靠信息，但没有将其纳入计算，那么结果不能说明最大熵方法的失败，而只是其被误用了.

为了回击异议 (B)，我们想说情况远比那要复杂得多. 最大熵原理基本上与任何可重复的"随机试验"无关. 一些最重要的应用是在分布 (11.56) 中的概率 p_i 与频率没有关联的情况——x_i 只是在单一场景下列举所有可能情况，例如渡轮问题中的汽车数.

然而，没有什么能阻止我们将最大熵原理应用于重复实验生成 x_i 的情况. 在这种情况下，可以对最大熵概率 $p(x_i)$ 与观测到 x_i 的频率之间的关系进行数学分析. 我们证明：(1) 在这种情况下，最大熵概率确实与频率有确定的关联；(2) 然而在大多数实际问题中，这种关联对于该方法的使用是不必要的；(3) 实际上，在观察到的频率与最大熵概率不一致时，最大熵原理才对我们最有用.

现在假设 x 的值是由一些随机试验确定的，在每次实验中，最终结果都是值 x_i（$i = 1, 2, \cdots, n$）. 在掷骰子问题中 $n = 6$. 但是现在，我们不问概率 p_i 是多少，而是问一个完全不同的问题：根据已知信息，关于各种 x_i 发生的相对频率 f_i 我们能说什么?

假设实验由 N 次试验组成（我们对 $N \to +\infty$ 的极限特别感兴趣，因为这是通常的频率概率理论所考虑的情况），然后对结果的每种可能序列进行分析. 每个试验都可以独立给出结果 $\{x_1, \cdots, x_n\}$，因此在整个实验中有 n^N 种可能的结果，但是其中许多种将与给定的信息不符.［我们再次假设它由几个函数的平均值 $f_k(x)$（$k = 1, 2, \cdots, m$）组成. 很明显，最终结论与采用这种形式或其他形式无关.］当然，我们将假定实验结果与该信息相符——如果不相符，则给定的信息是错误的，这是在解决错误的问题. 在整个实验中，结果 x_1 将获得 n_1 次，x_2 将获得 n_2 次，依此类推. 我们当然有

$$\sum_{i=1}^{n} n_i = N, \tag{11.91}$$

如果在实际实验中观察到给定平均值 F_k，则我们有附加关系

$$\sum_{i=1}^{n} n_i f_k(x_i) = NF_k, \qquad k = 1, 2, \cdots, m. \tag{11.92}$$

如果 $m < n-1$，则约束 (11.91) 和 (11.92) 不足以确定相对频率 $f_i = n_i/N$，但是，我们确实有理由偏爱 f_i 的某些选择. 例如，在最初可能出现的 n^N 种结果中，有多少会导致样本数 $\{n_1, n_2, \cdots, n_n\}$ 的集合？答案当然是多项式系数

$$W = \frac{N!}{n_1! n_2! \cdots n_n!} = \frac{N!}{(Nf_1)!(Nf_2)! \cdots (Nf_n)!}. \tag{11.93}$$

因此，能以最大数量实现的一组频率 $\{f_1, \cdots, f_n\}$ 是在约束 (11.91) 和 (11.92) 下使得 W 最大化的一组频率. 当然，我们同样可以最大化 W 的任何单调递增函数，尤其是 $N^{-1} \ln W$，但是当 $N \to +\infty$ 时，正如我们在 (11.29) 中所看到的，

$$\frac{1}{N} \ln W \to -\sum_{i=1}^{n} f_i \ln f_i = H_f. \tag{11.94}$$

因此，可以看到，在 (11.91) (11.92) 和 (11.94) 中，我们形式化了与最大熵原理推导过程中完全相同的数学问题，因此这两个问题将具有相同的解. 该论证在数学上使人联想到 11.4 节中给出的沃利斯推导. 通过直接应用贝叶斯定理，在所有 n^N 种可能的结果中分配均匀的先验概率，并且取极限 $N \to +\infty$，也可以得到相同的结果.

作为对异议 (C) 的部分回击，我们看到，无论约束条件是否采用平均值形式，数学等式都会成立. 如果给定的信息确实包含平均值，那么数学上就特别简洁，会导致分拆函数，如此而已. 但是，对于对问题施加了任何确定约束的给定信息，我们会得出相同的结论：使熵最大化的概率分布在数值上与可以与频率分布中最可能的频率相同.

此外，W 的最大值非常尖锐. 为了显示这一点，令 $\{f_1, \cdots, f_n\}$ 是最大化 W 并具有熵 H_f 的频率集合，$\{f'_1, \cdots, f'_n\}$ 是任何其他可能的频率集合（即满足约束 (11.91) 和 (11.92) 并具有熵 $H'_f < H_f$ 的集合）. 根据 (11.94)，比率（f_i 可以实现的方式数量）/（f'_i 可以实现的方式数量）渐近地是

$$\frac{W}{W'} \to \exp\{N(H_f - H_{f'})\}, \tag{11.95}$$

当 $N \to +\infty$ 会很快超过任何界限值. 因此在实验中，通过最大熵原理预测的频率分布会比任何其他满足相同约束的频率以占绝对优势多的方式实现.

这里，我们在概率和频率之间建立了另一种精确而一般的联系. 它与概率的定义无关，而是作为扩展逻辑的概率论的数学结果. 第 12 章将介绍概率与频率之间的另一种联系，其精确的数学陈述形式与这里不同，但具有相同的实际结果.

关于异议 (C)，施加约束的目的是将某些信息融入我们的概率分布中. 现在说概率分布“包含”某些信息意味着什么? 我们认为这意味着可以使用通常的规则估计期望值来从中提取信息. 通常数据 F_k 的准确性未知，因此仅使用数据约束 $\langle F_k\rangle$ 就是诚实的过程，而 $f_k(x)$ 的分布宽度将由可能数据 x_i 的阈值和密度确定. 但是如果我们确实有关于 F_1 准确性的信息，则可以通过在 $\langle f_1(x_i)^2\rangle$ 上添加一个新约束来融入这一信息，数学形式允许这样做. 但这很少会对最终结论产生任何实质性的影响，因为与合理的均方实验误差相比，$f_1(x)$ 的最大熵分布的方差通常很小.

现在我们来谈谈异议 (D)，并仔细分析情况，因为这可能是所有异议中最常见的一种. 概率与频率之间的上述联系是否证明了我们的预测，即最大熵分布实际上将在实验中作为频率分布被观察到? 从演绎的角度来看，显然不是这样的，因为正如异议 (D) 指出的那样，我们不得不承认不同的人可能拥有不同的信息，这将导致他们写出不同的分布，从而对可观察的事实做出不同的预测，而他们不可能全都是对的. 但这错失了要点，让我们仔细分析一下.

考虑一种特定情况：A 对平均值 $\langle f_1(x)\rangle$ 和 $\langle f_2(x)\rangle$ 施加约束，以使其与数据 F_1 和 F_2 一致. B 有更多信息，另外对 $\langle f_3(x)\rangle$ 施加了约束，以与其额外数据 F_3 一致. 两个人都根据自己的信息求最大熵分布. 由于 B 的熵是在受到更多约束时最大化，因此我们有

$$S_B \leqslant S_A. \tag{11.96}$$

假设 B 的额外信息是多余的，从某种意义上说，这只是 A 从他的分布中可以预测到的信息. 这时，A 针对概率分布的所有变量的变化最大化了熵，这些变化使 $\langle f_1\rangle$ 和 $\langle f_2\rangle$ 固定为指定值 F_1 和 F_2. 因此，相对于较少的变量变化，不用说他也会获得最大值，这也将 $\langle f_3\rangle$ 固定为最终获得的值. 因此在这种情况下，A 的分布也解决了 B 的问题，$\lambda_3 = 0$，并且 A 和 B 具有相同的概率分布. 只有在这种情况下，(11.96) 中的等号才成立.

从中我们学到了两件事. (1) 具有不同信息的两个人未必会得出不同的最大熵分布；只有当 B 的额外信息对 A 是“意外的”时，情况才如此. (2) 在定义最大熵问题时，没有必要保证所使用的不同信息独立：任何冗余信息都不会被计数两次，而是会自动退出方程. 确实，这不仅符合我们的基本合情条件，即布尔代数中 $AA = A$，而且对于任何变分原理也是这样的（如果旧的解已经满足该约束，则施加新的约束不能改变解）.

现在假设相反的情况：B 的额外信息在逻辑上与 A 所知道的相矛盾. 例如，可能 $f_3(x) = f_1(x) + 2f_2(x)$，但是 B 的数据不满足 $F_3 = F_1 + 2F_2$. 显然，没有

符合 B 数据的概率分布. 我们的机器人将如何告诉我们这一点呢? 数学上，你将发现方程

$$F_k = -\frac{\partial \ln Z(\lambda_1, \lambda_2, \lambda_3)}{\partial \lambda_k} \tag{11.97}$$

没有兼容的实数解 λ_k. 在上述例子中，

$$\begin{aligned} Z(\lambda_1, \lambda_2, \lambda_3) &= \sum_{i=1}^{n} \exp\{-\lambda_1 f_1(x_i) - \lambda_2 f_2(x_i) - \lambda_3 f_3(x_i)\} \\ &= \sum_{i=1}^{n} \exp\{-(\lambda_1 + \lambda_3) f_1(x_i) - (\lambda_2 + 2\lambda_3) f_2(x_i)\} \end{aligned} \tag{11.98}$$

以及

$$\frac{\partial Z(\lambda_1, \lambda_2, \lambda_3)}{\partial \lambda_3} = \frac{\partial Z(\lambda_1, \lambda_2, \lambda_3)}{\partial \lambda_1} + 2\frac{\partial Z(\lambda_1, \lambda_2, \lambda_3)}{\partial \lambda_2}, \tag{11.99}$$

因此方程 (11.97) 对于 $\lambda_1, \lambda_2, \lambda_3$ 没有解，除非 $F_3 = F_1 + 2F_2$. 因此当一条新的信息在逻辑上与先前的信息矛盾时，最大熵原理就应该失效，从而拒绝提供任何分布.

最有趣的是，B 的额外信息既非多余的，也不是矛盾的中间情况. 这样，他发现与 A 具有不同的最大熵分布，且 (11.96) 中的不等号成立，这表明 B 的额外信息是“有用的”，进一步缩小了 A 信息所允许的可能性范围. 该范围的度量正是 W. 根据 (11.95)，我们渐近地有

$$\frac{W_A}{W_B} \sim \exp\{N(S_A - S_B)\}. \tag{11.100}$$

对于大的 N，即使熵稍有减小，也会导致可能性数量的极大减少.

现在假设我们在 A 和 B 的观察下开始进行实验. 由于 A 预测的平均值 $\langle f_3 \rangle$ 与 B 已知的正确平均值不同，因此很明显，实验分布无法在所有方面都与 A 的预测一致. 我们也无法事先确定它是否也将与 B 的预测一致，因为可能有 B 不知道的其他约束 $f_4(x), f_5(x), \cdots$ 在影响实验.

以上展示的性质证明了以下较弱的概率与频率对应关系：如果融入最大熵分析中的信息包括随机试验中实际的所有约束，那么最大熵方法预测的分布将最有可能被观测到. 实际上，自然界中观察到的大多数频率分布都是最大熵分布，因为它们的实现方式比其他任何分布要多得多.

相反，如果实验没有证实最大熵预测，并且在重复实验时这种差别持续存在，由于根据假设数据 F_i 是不完整的，那么我们可以得出结论，实验的物理机制中一定包含一些附加约束，而这些约束在最大熵计算中并未考虑在内. 然后，观察到的偏差为新约束的性质提供了线索. 这样，A 可以根据经验发现他的信息不完整.

总而言之，最大熵原理不是能告诉我们哪些预测一定正确的金科玉律. 它是一种归纳推理的规则，会告诉我们当前信息最强烈支持的是什么预测.

11.9 评注

11.8 节中的情景很好地描述了吉布斯当时所面对的情况，这是历史上最重要的统计分析应用之一. 众所周知，到 1901 年，在经典统计力学中，规范系综（吉布斯基于指定的能量平均值对于经典状态空间，或相空间推导出的最大熵分布）的使用无法正确预测某些热力学性质（热容量、状态方程等）. 数据分析表明实际物理系统的熵总是小于预测值. 因此，当时吉布斯处于 A 的境况. 结论是，物理学的微观定律中一定包含古典力学定律中未包含的一些附加约束.

吉布斯于 1903 年去世，其他人继续寻找这种约束的性质：首先是普朗克对黑体辐射，然后是爱因斯坦和德拜对固体，最后是玻尔对原子，他们都找到了相应约束. 约束是可能的能量值的离散性，此后称为能级. 1927 年，海森伯和薛定谔提出了可以根据第一原理进行计算的数学理论.

因此，一个历史事实是，通过最大熵原理的“不成功”应用，发现了需要量子理论，及表明新理论某些必要特征的第一线索. 我们可能会期望这种情况将来会再次发生. 这是以下观点的基础：最大熵原理仅在无法正确预测实验事实的情况下对我们最有用. 这也说明了归纳推理在科学中的真正性质、作用和价值. 杰弗里斯也强调了这一点（见 1957 年版的 Jeffreys，1931）.

吉布斯（1902）用形式

$$w(q_1,\cdots,q_n;p_1,\cdots,p_n)=\exp\{\eta(q_1,\cdots,q_n)\}, \tag{11.101}$$

在相空间中写出概率密度，并将函数 η 称为“相位概率指标”. 他分别对平均能量、平均能量和粒子数的约束得出了他的正则系综和大正则系综［见吉布斯的著作（Gibbs，1902，第 143 页）］，并指出这是“在不违反该条件的情况下，相位分布给出了相位概率指标平均值 $\bar{\eta}$ 的最小值……”，当然，这就是我们今天描述的最大熵约束条件.

不幸的是，由于健康状况不佳，吉布斯并没有完成工作. 他没有给出任何明确的解释，我们只能猜测他是否有相对于其他函数，为什么要最大化这一特定函数的解释. 因此，他的程序对许多人来说似乎很随意. 60 年来，人们对于吉布斯方法的合理性一直感到困惑并存在争议. 它们被一些统计力学的学者完全拒绝，而被其他人以最大程度谨慎对待. 只有在香农的著作（Shannon，1948）中，才能从根本上看到这种新思想. 这些历史问题在杰恩斯的著作（Jaynes，1967 和 Jaynes，1992b）中有更详细的讨论.

第 12 章　无知先验和变换群

无知胜于错误，认为自己不知道的人比有着错误信念的人更接近真理.

——托马斯·杰斐逊（Thomas Jefferson，1781）

将先验信息唯一地转化为先验概率分配的问题是概率论中的另一个重要问题. 对于此问题，前一章的最大熵原理提供了一个重要工具. 但是这一问题尚未解决，因为几十年来它一直被那些无法将概率分布视为代表信息的人所拒绝. 正是由于长期以来的忽视，许多当前的科学、工程、经济和环境问题亟待此问题的答案，因为不解决此问题则许多重要的应用将无法继续进行.

12.1　我们要做什么?

令人感到奇怪的是，即使不同的人在应该计算什么的问题上意见完全一致，他们对实际在做什么以及为什么这样做的看法也可能截然不同. 例如，有一个庞大的贝叶斯社区，其成员自称为“主观贝叶斯主义者”，他们的态度处于“正统”统计学和我们的理论之间. 他们大部分接受了标准的正统统计训练，但是由于随后看到了其中的荒谬而叛离正统统计哲学，但同时保留了使用正统统计学术语和符号的习惯.

这些表达习惯使得主观贝叶斯主义者遇到了严重的障碍. 尽管他们看到概率不能仅仅表示频率，但是他们仍然将抽样概率视为“随机变量”的频率. 对他们来说，先验概率和后验概率仅代表个人意见，这些观点根据德菲内蒂的连贯性原则进行更新. 幸运的是，这将得出贝叶斯定理，因此与我们的计算方式相同.

在分配先验概率时，主观贝叶斯主义者在问题的初始阶段面临着模糊不清的尴尬情况. 如果先验概率仅代表先验个人意见，那么它们基本上是随意而未定义的. 似乎只有通过内省才能分配先验概率，并且不同的人会做出不同的分配. 然而，大多数主观贝叶斯主义者使用的语言中暗示着在实际问题中存在某种未知的“真实”先验概率分布. 我们认为，在认识到以下三个核心点之前，推断问题是没有良好定义的.

(A) 先验概率代表我们的先验信息，其值不是通过自省，而是通过对该信息的逻辑分析来确定的.

(B) 由于最终结论同时取决于先验信息和数据，在提出问题时，必须指定要使用的先验信息，就像指定数据一样.

(C) 我们的目标是，在具有相同先验信息的两个人必须分配相同的先验概率的意义上，推断应该是完全“客观的”.

如果没有指定先验信息，那么推断问题就如同没有指定数据一样没有得到良好的定义. 确实，自从拉普拉斯时代以来，概率论的应用就因先验信息的处理困难而受到阻碍. 在实际推断问题中，典型的情况是我们有与问题高度相关的强有力的先验信息，不考虑这些信息就会进行明显的不一致性推理，从而可能导致荒谬或危险的误导性结果.

在指定先验信息之后，我们便面临着将该信息转化为特定的先验概率的问题. 这种形式转化过程占据概率论的整整一半篇幅，因为这是实际应用所需要的. 然而，这方面在正统统计学中完全缺失，在主观贝叶斯理论中也只能被隐约地感觉到.

就像 0 是数列相加的自然起点一样，许多先验信息转化的自然起点是完全无知的状态. 在上一章中我们看到，对于离散概率，最大熵原理表明——与我们显而易见的直觉相一致——完全无知是由均匀先验概率分配表示的. 对于连续概率，这一问题要困难得多，因为直觉不能告诉我们结果，我们必须诉诸正式的必备条件和原则. 本章中，我们将为此探讨变换群数学工具的使用.

有些人反对尝试表征完全无知，理由是完全无知的状态并不“存在”. 对此，我们会回答说：完美的三角形也不存在，但是不了解完美三角形性质的测量者将是不称职的. 对我们来说，完全无知先验是真实先验信息的理想极限情况，正如完美三角形是测量者测量的真实三角形的理想极限情况一样. 如果没有学会如何处理完全无知先验，那么我们几乎不可能解决真正的问题.

到目前为止研究的这些相对简单的问题可以通过合理的常识——几乎总是能看出先验应该是什么——来解决. 当我们处理更复杂的问题时，如何找到无知先验的形式理论变得越来越必要. 在很多情况下，只要有最大熵原理就足够了，但是我们的工具箱中也应该提供诸如变换群、边缘化理论和编码理论之类的原理. 本章将研究变换群方法. 在开始研究之前，我们将介绍性地讨论连续分布的最大熵原理，并说明这如何自然地导致表征完全无知分布的思想.

12.2 无知先验

到目前为止，我们仅仅考虑了离散情况下的最大熵原理，并且发现：如果所寻求的分布可以视为由随机试验产生的，那么概率和频率之间存在对应关系，并且

结果与其他概率论原理一致. 但是在数学上，并没有要求实际执行或构想任何随机试验，因此我们从最广泛的意义上解释该原理，从而赋予它最广泛的适用性，即无论是否涉及随机试验，最大熵分布都代表着对我们的知识状态最“诚实”描述.

在这样的应用中，最大熵原理非常易于应用，并且会得出我们期望的结果. 例如，我在一篇文章（Jaynes，1963a）中，分析了在不确定性条件下决策的一系列问题（本质上是库存控制问题）. 这类问题在实践中经常出现. 在这里，自然状态不是任何随机试验的结果，没有样本分布也没有样本，因此从切尔诺夫和摩西（Chernoff & Moses，1959）的角度来看，这可能被认为是一个“无数据”决策问题. 但是在各个阶段中，可以获得越来越多的先验信息. 通过最大熵方法来吸收它们，可以得到一系列先验分布，其中的可能性区间逐渐变小. 它们会导致一系列决策，每一种决策都是基于该阶段可用信息的理性决策，这对应于早期阶段根据直觉就能够看到答案的直观常识性判断. 很难想象，如果不使用最大熵原理或者与之等效的工具，这一问题该如何解决.

在将最大熵原理应用于物理与工程学问题的多年实践中，我们尚未发现在离散先验时，它无法产生有用且合理结果的情况. 据我所知，还没有人提出分配离散先验的其他通用方法. 看来，最大熵原理可能是分配离散先验问题的最终解.

12.3　连续分布

但是，在应用最大熵原理分配连续先验分布时需要更深入的分析，因为乍一看，结果似乎依赖于参数的选择. 我们在这里并不是指以下众所周知的事实，即量

$$H' = -\int \mathrm{d}x\, p(x|I) \ln[p(x|I)] \tag{12.1}$$

在参数 $x \to y(x)$ 的变换下缺乏不变性. 因为 (12.1) 并不是任何推导的结果，事实证明它也不是连续分布的正确信息度量. 香农定理使用 (11.23) 作为信息度量仅适用于离散分布. 为了在连续分布的情况下找到对应的表达式，我们可以对离散分布取极限. 以下论证可以根据我们的要求变得更严格，但是会大大地牺牲明晰性.

在离散熵表达式中

$$H_I^{\mathrm{d}} = -\sum_{i=1}^{n} p_i \ln p_i. \tag{12.2}$$

我们假设离散点 x_i（$i = 1, 2, \cdots, n$）变得越来越多，以至于当 $n \to +\infty$ 时，

$$\lim_{n\to+\infty} \frac{1}{n}(a < x < b \text{ 的点数}) = \int_a^b \mathrm{d}x\, m(x). \tag{12.3}$$

如果达到极限的行为足够好，那么在 x 的任何特定值附近的差值 $(x_{i+1} - x_i)$ 也

会趋于 0，即

$$\lim_{n\to+\infty}\big[n(x_{i+1}-x_i)\big]=\big[m(x_i)\big]^{-1}. \tag{12.4}$$

离散概率分布 p_i 将通过如下极限形式变为连续概率 $p(x|I)$：

$$p_i=p(x_i|I)(x_{i+1}-x_i); \tag{12.5}$$

或者，根据 (12.4)，

$$p_i\to p(x_i|I)\big[nm(x_i)\big]^{-1}. \tag{12.6}$$

离散熵 (12.2) 将变成积分

$$H_I^{\mathrm{d}}\to\int \mathrm{d}x\,p(x|I)\ln\left[\frac{p(x|I)}{nm(x)}\right]. \tag{12.7}$$

在求极限时，它包含一个无限大项 $\ln n$. 如果减去此项，则差值将接近确定的极限，我们将其作为连续分布信息的度量：

$$H_I^{\mathrm{c}}\equiv\lim_{n\to+\infty}\big[I_I^{\mathrm{d}}-\ln n\big]=-\int \mathrm{d}x\,p(x|I)\ln\left[\frac{p(x|I)}{m(x)}\right]. \tag{12.8}$$

“不变测度”函数 $m(x)$ 与离散点的极限密度成正比. 在迄今为止研究的所有应用中 $m(x)$ 都是行为良好的连续函数，因此我们继续使用黎曼积分的概念. 我们将 $m(x)$ 称为“测度”只是为了适当的概括，如果实际问题需要，可以随时提供. 由于 $p(x|I)$ 和 $m(x)$ 在变量变换时以相同的方式变化，H_I^{c} 是不变的.

我们寻求归一化的概率密度 $p(x|I)$：

$$\int \mathrm{d}x\,p(x|I)=1 \tag{12.9}$$

（我们知道积分范围是整个参数空间），它受 m 个不同函数 $f_k(x)$ 的平均值的信息约束，即

$$F_k=\int \mathrm{d}x\,p(x|I)f_k(x),\qquad k=1,2,\cdots,m, \tag{12.10}$$

其中 F_f 是给定的数值. 在这些约束条件下，我们要最大化度量 (12.8). 解还是基础的：

$$p(x|I)=Z^{-1}m(x)\exp\{\lambda_1 f_1(x)+\cdots+\lambda_m f_m(x)\}, \tag{12.11}$$

其中分拆函数是

$$Z(\lambda_1,\cdots,\lambda_m)\equiv\int \mathrm{d}x\,m(x)\exp\{\lambda_1 f_1(x)+\cdots+\lambda_m f_m(x)\}, \tag{12.12}$$

拉格朗日乘子 λ_k 由以下方程组确定，

$$F_k=-\frac{\partial\ln Z(\lambda_1,\cdots,\lambda_m)}{\partial\lambda_k},\qquad k=1,\cdots,m. \tag{12.13}$$

那么我们对其他任何量 $q(x)$ 的“最优”估计（根据平方损失函数）是

$$\langle q\rangle = \int \mathrm{d}x\, q(x)p(x|I). \tag{12.14}$$

从这些方程可以明显看出，当我们使用 (12.8) 而不是 (12.1) 作为信息量度时，不仅最终结论 (12.14) 而且分拆函数和拉格朗日乘子在参数变换 $x \to y(x)$ 时都不变. 这些量在应用中具有确定的物理意义.

但是仍然存在一个实际的难题：如果参数空间不是任何极限过程的结果，那么是什么决定了 $m(x)$ 的适当度量？结论显然依赖于我们采取的度量. 这是迄今为止最大熵原理的不足. 我们必须对此不足进行弥补才能将最大熵原理视为对先验概率问题的全面解决方案. 让我们看看该度量的直观含义. 考虑一维情况，假设我们已知 $a < x < b$，但是没有其他先验信息，那么没有拉格朗日乘子 λ_k，并且 (12.11) 可以简化为

$$p(x|I) = \left[\int_a^b \mathrm{d}x\, m(x)\right]^{-1} m(x), \qquad a < x < b. \tag{12.15}$$

除了相差一个常数因子外，度量 $m(x)$ 也是一种描述 x 的“完全无知”的先验分布. 因此，模糊性只是一直困扰着贝叶斯统计的古老问题：我们如何找到表征“完全无知”的先验？一旦解决了这一问题，最大熵原理将导致一种确定、与参数无关、根据可检验的先验信息建立先验分布的方法. 由于关于这一问题已经有 200 多年的讨论和争议，我们希望以一种建设性的态度来说明它.

像某些人所做的那样，以完全无知的状态不“存在”为由拒绝这个问题，与以不存在物理点为由拒绝欧几里得几何一样荒谬. 在归纳推理的研究中，完全无知的概念像算术中的 0 的概念一样自然而然地出现在理论中.

如果有人以完全无知的概念模糊不清且未良好定义为由拒绝考虑它，那么回应是，在任何完整的推理理论中都不能逃避该概念. 因此，如果仍未良好定义，那么主要且紧迫的目标就是找到一个符合直觉并且在数学理论中具有实际用途的精确定义.

带着这种认识，让我们对于前人有关该问题的一些思想做一下回顾. 贝叶斯建议，在某种特殊情况下，通过分配均匀先验概率密度来表达完全无知. 该规则有其应用范围，因为拉普拉斯将其应用于分析天文数据而获得了一些天体力学中最重要的发现，但是贝叶斯规则有一个明显的困难，那就是它不具有参数变换不变性，并且似乎没有任何标准可以告诉我们应该使用哪种参数.（顺便指出，无偏估计量、有效估计量和最小置信区间的概念有同样模糊不清的问题，并且后果同样严重，因此正统统计学不能声称比贝叶斯理论更好地解决了这个问题.）

杰弗里斯（Jeffreys，1931；1939，1957）建议我们分配先验 $\mathrm{d}\sigma/\sigma$ 给已知为正的连续参数 σ，理由是无论使用参数 σ 还是 σ^m，意义是一样的. 这肯定是朝着正确方向迈出了一步，但是它不能向更一般的参数变换扩展. 我们不希望（并且显然不能）在所有参数变换下保持先验形式不变. 我们想要的是内容的不变性，而概率论法则已经决定先验在参数变换下必须如何变换.

因此必须以不同的方式陈述真正的问题. 我们建议恰当的问题是："给定形式（例如贝叶斯或杰弗里斯形式）的先验适用于哪种参数选择?" 我们的参数空间似乎具有类似软体动物的性质，使我们无法对此做出回答，除非能找到一种新原则，能赋予它们"刚性".

根据这种陈述方式，我们认识到这类问题已经在其他数学分支中出现并得到了解决. 在黎曼几何和广义相对论中，我们允许任意连续的坐标变换. 然而刚性是通过不变线元的概念来体现的，这使我们能够独立于坐标选择而做出几何和物理意义上的明确陈述. 在连续群理论中，在哈尔（Haar，1933）、庞特里亚金（Pontryagin，1946）和维格纳（Wigner，1959）引入不变群测度之前，群参数空间也具有类似软体动物的性质. 我们试图对统计的参数空间做类似的事情.

庞加莱（Poincaré，1912）以及最近的哈蒂根（Hartigan，1964）、斯通（Stone，1965）和弗雷泽（Fraser，1966）讨论了在与此相关的问题中利用变换群的想法. 在以下各节中，我们将给出四个以不同的群论方法进行推理的例子. 这些方法主要是由维格纳（Wigner，1959）和外尔（Weyl，1961）发展的，它们在物理问题上取得了巨大的成功，并且似乎特别适合解决我们现在的问题.

12.4 变换群

最好通过一些简单的例子来说明这种推理方法，其中第一个例子在实践中也是最重要的.

12.4.1 位置和比例参数

我们从具有两个参数的连续分布

$$p(x|\nu\sigma) = \phi(x, \nu, \sigma)\mathrm{d}x \tag{12.16}$$

中抽样，并且考虑以下问题 A.

问题 A

给定样本 $\{x_1, \cdots, x_n\}$，估计参数 ν 和 σ. 除非引入如下确定的先验分布，此问题在数学和概念上都是不确定的.

$$p(\nu\sigma|I)\mathrm{d}\nu\mathrm{d}\sigma = f(\nu, \sigma)\mathrm{d}\nu\mathrm{d}\sigma, \tag{12.17}$$

但是如果我们只是说“对先验分布完全无知”，则不能确定要使用什么函数 $f(\nu,\sigma)$.

假设我们通过式子

$$\begin{aligned}\nu' &= \nu + b,\\ \sigma' &= a\sigma,\\ x' - \nu' &= a(x-\nu)\end{aligned} \tag{12.18}$$

将旧变量变换为新变量 $\{x',\nu',\sigma'\}$，其中 $0 < a < +\infty, -\infty < b < +\infty$. 分布 (12.16) 用新变量表示为

$$p(x'|\nu'\sigma') = \psi(x',\nu',\sigma') = \phi(x,\nu,\sigma)\mathrm{d}x, \tag{12.19}$$

或者根据 (12.18) 是

$$\psi(x',\nu',\sigma') = a^{-1}\phi(x,\nu,\sigma). \tag{12.20}$$

类似地，先验分布变为 $g(\nu',\sigma')$，其中，根据变换 (12.18) 的雅可比变换，

$$g(\nu',\sigma') = a^{-1}f(\nu,\sigma). \tag{12.21}$$

以上关系对于任何分布 $\phi(x,\nu,\sigma)$, $f(\nu,\sigma)$ 都成立.

现在假设分布 (12.16) 在变换群 (12.18) 下是不变的，因此无论 a 和 b 的值是什么，ψ 和 ϕ 都是相同的函数：

$$\psi(x,\nu,\sigma) = \phi(x,\nu,\sigma). \tag{12.22}$$

这种不变性的条件是 $\phi(x,\nu,\sigma)$ 必须满足函数方程

$$\phi(x,\nu,\sigma) = a\phi(ax - a\nu + \nu + b, \nu + b, a\sigma). \tag{12.23}$$

将此方程对 a 和 b 求微分并求解所得的微分方程，可以得到函数方程 (12.23) 的一般解是

$$\phi(x,\nu,\sigma) = \frac{1}{\sigma}h\left(\frac{x-\nu}{\sigma}\right), \tag{12.24}$$

其中 $h(q)$ 是任意函数. 因此，位置参数 μ 和比例参数 σ 的通常定义等价于指定在变换群 (12.18) 下分布将是不变的.

除了知道 ν 是位置参数且 σ 是比例参数外，我们对 ν 和 σ “完全无知”，这种说法的意思是什么？为了回答这一问题，我们可以做如下推理：如果缩放比例可以使问题变得不同，那么我们就不是完全无知的，我们一定掌握了有关该问题的绝对数值范围的某种信息；同样，如果位置移动会使问题变得不同，那么我们一定已经有一些位置相关的先验信息. 换句话说，位置参数和比例参数的“完全无知”是一种知识状态，通过缩放比例和位置平移不会改变这种知识状态. 我们目前必须更仔细地说明这一点，但是首先来看看这种知识状态的后果吧. 考虑问题 B.

问题 B

给定样本 $\{x'_1,\cdots,x'_n\}$，估计 ν' 和 σ'. 如果我们在以上描述的意义上“完全无知”，那么必须将问题 A 和问题 B 视为完全等价的，它们具有相同的抽样分布，并且我们对问题 B 中的 ν' 和 σ' 的先验知识与问题 A 中的 ν 和 σ 完全相同.

现在，因为我们提出了两个具有相同先验信息的问题，基本合情条件获得了非平凡的内容. 一致性要求我们分配相同的先验概率分布. 因此，无论 (a,b) 的值是什么，f 和 g 必须是相同的函数：

$$f(\nu,\sigma)=g(\nu,\sigma). \tag{12.25}$$

这样，先验分布的形式现在是唯一确定的，因为联合 (12.18) (12.21) 和 (12.25)，我们看到 $f(\nu,\sigma)$ 必须满足函数方程

$$f(\nu,\sigma)=af(\nu+b,a\sigma), \tag{12.26}$$

其一般解是

$$f(\nu,\sigma)=\frac{\text{常数}}{\sigma}. \tag{12.27}$$

这正是杰弗里斯的规则!

我们不能就此跳到先验分布 (12.27) 是由总体的分布 (12.24) 形式确定的结论. 确实，如果先验分布的形式仅由我们进行抽样的总体决定，那将非常令人不安，任何导致这种结论的原则都是可疑的. 检查上述推理过程表明：结果 (12.27) 是由变换群 (12.18) 唯一确定的，而不是由分布 (12.24) 的形式确定.

为了阐明这一点，注意在分布 (12.24) 不变的情况下有多个变换群. 在变换 (12.18) 中，我们通过因子 a 改变比例并通过因子 b 平移. 用符号 (a,b) 表示此操作，我们可以进行变换 (a_1,b_1)，然后进行变换 (a_2,b_2)，根据 (12.18) 可以获得群元素的组成定律：

$$(a_2,b_2)(a_1,b_1)=(a_2a_1,b_2+b_1). \tag{12.28}$$

因此，群 (12.18) 是阿贝尔群，这是两个单参数群的直积. 它有矩阵形式的表示

$$\begin{pmatrix} a & 0\\ 0 & \mathrm{e}^b\end{pmatrix}. \tag{12.29}$$

现在考虑先进行比例变换 a，然后进行平移 b 的变换群. 该群为

$$\begin{aligned}\nu'&=a\nu+b,\\ \sigma'&=a\sigma,\\ x'&=ax+b.\end{aligned} \tag{12.30}$$

该群具有组成定律

$$(a_2,b_2)(a_1,b_2)=(a_2a_1,a_2b_1+b_2), \tag{12.31}$$

因此群 (12.30) 是非阿贝尔群，它有不能简化为对角形式的矩阵形式的表示

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}, \tag{12.32}$$

因此，群 (12.18) 和群 (12.30) 是完全不同的群.

如果指定变换群 (12.30) 而不是 (12.18)，则 (12.21) 和 (12.23) 会变为

$$g(\nu', \sigma') = a^{-2} f(\nu, \sigma), \tag{12.33}$$

和

$$\phi(x, \nu, \sigma) = a\phi(ax + b, a\nu + b, a\sigma). \tag{12.34}$$

但是我们发现 (12.34) 的一般解也是 (12.24)，因此两个群都很好地定义了位置参数和比例参数，但是它们对先验的影响是不同的，因为函数方程 (12.26) 变为

$$f(\nu, \sigma) = a^2 f(a\nu + b, a\sigma), \tag{12.35}$$

其一般解是

$$f(\nu, \sigma) = \frac{\text{常数}}{\sigma^2}. \tag{12.36}$$

因此在变换群 (12.18) 和 (12.30) 下不变的知识状态是不同的，我们看到“完全无知”的概念有了新的含义. 为了明确地定义它，仅仅说“缩放比例和位置平移不会改变这种知识状态”是不够的. 我们必须指定执行这些操作的确切方式，即我们必须指定确定的变换群.

因此我们面临一个问题：群 (12.18) 和 (12.30) 中的哪个真正描述了完全无知的先验信息？群 (12.30) 的困难在于方程组 $x' = ax + b$, $\nu' = a\nu + b$，因此缩放操作是在 $x = 0, \nu = 0$ 的两个点上进行的. 但是如果我们对于位置“完全无知”，那么 $x = 0$ 并没有特殊意义. 是什么决定了要对某一固定点进行比例缩放呢？

在我能想到的所有问题中，都是对应着杰弗里斯先验概率规则的群 (12.18) 更适合. 在这里比例缩放仅涉及差值 $\{x - \nu\}$，因此它是针对一个本身任意的点执行的，群 (12.18) 没有定义任何“固定点”. 但是，很有意思的问题是，是否有人能找到点 $x = 0$ 具有特殊含义的例子，从而证明这时更强的先验 (12.36) 更合理.

总结如下：如果仅仅指定“初始状态完全无知”，我们并不能获得任何明确的先验分布，因为这样的陈述太模糊，无法定义任何数学上的适定问题. 如果可以指定一组将问题转化为等效问题的操作，那么我们将更加明确地定义这种知识状态. 在找到这样一组操作之后，一致性的基本合情条件就对先验概率的形式施加了不小的限制.

12.4.2 泊松率

再举一个数学上差别不大但描述起来差别较大的例子. 考虑泊松过程，在时间间隔 t 内恰好会发生 n 个事件的概率为

$$p(n|\lambda t)=\exp\left\{-\frac{(\lambda t)^n}{n!}\right\}. \tag{12.37}$$

通过观察单位时间事件发生的数量，我们希望估计速率常数 λ. 最初，我们对 λ 完全无知，除了知道它是物理量纲为（秒）$^{-1}$ 的速率常数，即我们完全不知道过程的绝对时间尺度.

然后假设两个观测者 X 和 X′ 的手表以不同的速率运行，使得他们对给定间隔的测量值以 $t=qt'$ 关联进行实验. 由于他们在观察相同的物理实验，因此他们的速率常数必须与 $\lambda't'=\lambda t$ 或 $\lambda'=q\lambda$ 相关联. 他们分配的先验概率为

$$p(\mathrm{d}\lambda|X)=f(\lambda)\mathrm{d}\lambda, \tag{12.38}$$

$$p(\mathrm{d}\lambda'|X')=g(\lambda')\mathrm{d}\lambda'. \tag{12.39}$$

如果两式相互一致（即它们具有相同的物理内容），则必须有 $f(\lambda)\mathrm{d}\lambda=g(\lambda')\mathrm{d}\lambda'$ 或者 $f(\lambda)=qg(\lambda')$. 但是 X 和 X′ 都完全无知，并且处于相同的知识状态，因此 f 和 g 必须是相同的函数：$f(\lambda)=g(\lambda)$. 结合这些关系可得出函数方程 $f(\lambda)=qf(q\lambda)$ 或者

$$p(\mathrm{d}\lambda|X)\sim\lambda^{-1}\mathrm{d}\lambda. \tag{12.40}$$

使用除此以外的任何其他先验将会使得时间尺度的变化导致先验形式的变化，这意味着不同的先验知识的状态. 但是如果我们对于时间尺度完全无知，那么所有时间尺度都应该是等价的.

12.4.3 未知成功概率

第三个例子不那么普通——直觉无法预知其结果，我们将考察成功概率未知的伯努利试验. 这里成功的概率本身是要估计的参数 θ. 给定 θ，我们在 n 次试验中观测到 r 次成功的概率是

$$p(r|n\theta)=\binom{n}{r}\theta^r(1-\theta)^{n-r}. \tag{12.41}$$

问题同样是：什么样的先验分布 $f(\theta)\mathrm{d}\theta$ 能描述 θ 完全无知的状态?

在讨论这个问题时，拉普拉斯承袭了贝叶斯的做法，并用以下名言回答了这一问题："当一个简单事件的概率未知时，我们可以假设 0 和 1 之间的所有值都具有相同的可能性." 换句话说，贝叶斯和拉普拉斯使用的是均匀分布 $f_B(\theta)=1$. 但是，杰弗里斯（Jeffreys，1939）和卡尔纳普（Carnap，1952）指出，由此产生的拉

普拉斯连续法则似乎与我们直观进行的归纳推理不太吻合. 杰弗里斯提出，如果理论要解释科学家做出的那种推断，则 $f(\theta)$ 应该赋予端点 $\theta=(0,1)$ 更大的权重.

例如，在化学实验室中，我们发现一个罐子中装有未知且未做标记的化合物. 我们开始完全不知道该化合物是否会溶于水. 但是在观察到一部分样品确实溶于水后，我们立即推断出该化合物的所有样品都是水溶性的. 尽管这一结论并没有演绎证明的同等效力，但我们认为这一推断是非常合理的. 然而贝叶斯-拉普拉斯规则认为这种情况成立的可能性很小，并且预测检验的下一部分样品溶于水的概率为 2/3.

现在让我们从变换群的角度来研究这个问题. 由于 $f(\theta)\mathrm{d}\theta$ 是“概率的概率”，这里有一个概念上的困难，但是这种困难可以通过将分裂人格的概念带到极致来消除. 我们假设 $f(\theta)$ 描述的不是任何人的知识状态，而是假设有对成功概率持不同信念的大量不同个体，而 $f(\theta)$ 描述了他们的信念的分布. 是否有可能尽管每个人都有明确的意见，但总体上对于 θ 完全无知呢? $f(\theta)$ 的何种分布描述了对该问题处于完全迷糊状态的总体?

由于我们关注的作为扩展逻辑的概率论必须具有一致性，必须假定每个个体都根据概率论的数学法则（贝叶斯定理等）进行推理. 因此，他们持有不同信念的原因是，他们获得了不同且相互矛盾的信息: 比如一个人读了《圣路易斯邮报》，另一个人读了《洛杉矶时报》，一个人读了《工人日报》，另一个人读了《国家评论》，等等. 概率论中的任何内容都不会使人们怀疑他在问题陈述中被告知的事实.

现在假设在进行实验之前，对所有这些人同时给出一个确定的证据 E. 每个人都会根据贝叶斯定理改变自己的信念状态，X 先生曾认为成功概率为

$$\theta=p(S|X), \tag{12.42}$$

随后将变为

$$\theta'=p(S|EX)=\frac{p(S|X)p(E|SX)}{p(E|SX)p(S|X)+p(E|FX)p(F|X')}, \tag{12.43}$$

其中 $p(F|X)=1-p(S|X)$ 是他对失败概率的先验信念. 因此该新证据生成了参数空间 $0\leqslant\theta\leqslant1$ 到其自身的映射，根据 (12.43) 是

$$\theta'=\frac{a\theta}{1-\theta+a\theta}, \tag{12.44}$$

其中

$$a=\frac{p(E|SX)}{p(E|FX)}. \tag{12.45}$$

如果作为总体的人群不能从这一新证据中学到任何东西，那么合理的说法是，由于相互矛盾的宣传，人们对该问题处于完全迷惑的状态. 因此我们通过以下条件

来定义“完全迷惑”或“完全无知”的状态：经过 (12.44) 的变换，拥有信念在任何给定范围 $\theta_1 < \theta < \theta_2$ 内的个体数量与以前相同.

数学问题同样很简单. 信念的原始分布 $f(\theta)$ 通过变换 (12.44) 变为新的分布 $g(\theta')$，其中

$$f(\theta)\mathrm{d}\theta = g(\theta')\mathrm{d}\theta'. \tag{12.46}$$

如果人群总体上没有学到什么，那么 f 和 g 必须是相同的函数：

$$f(\theta) = g(\theta). \tag{12.47}$$

结合 (12.44) (12.46) 和 (12.47)，我们发现 $f(\theta)$ 必须满足函数方程

$$af\left(\frac{a\theta}{1-\theta-a\theta}\right) = (1-\theta+a\theta)^2 f(\theta). \tag{12.48}$$

这可以通过消去 (12.44) 和 (12.48) 之间的 a 值直接求解，或者以更常见的方式通过对 a 进行微分并使得 $a=1$ 来求解. 这将导致微分方程

$$\theta(1-\theta)f'(\theta) = (2\theta-1)f(\theta), \tag{12.49}$$

其解是

$$f(\theta) = \frac{\text{常数}}{\theta(1-\theta)}, \tag{12.50}$$

它具有杰弗里斯预期的定性特征. 这样，假想的个体集合已经达到了揭示问题的变换群 (12.44) 的目的，让它们再次合并为一个单一思想（希望估计 θ 的统计学家的思想）. 让我们研究一下使用 (12.50) 作为先验分布的后果.

如果我们在 n 次试验中观察到 r 次成功，则根据 (12.41) 和 (12.50)，θ 的后验分布（假定 $r \geqslant 1, n-r \geqslant 1$）是

$$p(\mathrm{d}\theta|rn) = \frac{(n-1)!}{(r-1)!(n-r-1)!}\theta^{r-1}(1-\theta)^{n-r-1}\mathrm{d}\theta. \tag{12.51}$$

该分布具有期望值和方差

$$\langle\theta\rangle = \frac{r}{n} = f, \tag{12.52}$$

$$\sigma^2 = \frac{f(1-f)}{n+1}. \tag{12.53}$$

因此，根据平方损失函数的准则，成功概率的“最优”估计等于观察到的成功频率 f. 这也等于下一次试验成功的概率，与研究过伯努利试验的每个人的直觉一致. 另外，贝叶斯-拉普拉斯均匀先验将根据连续法则得到均值 $\langle\theta\rangle_B = (r+1)/(n+2)$，这一结果总让人觉得有些奇怪.

对于区间估计，数值分析表明，根据分布 (12.51) 得出的结论在实际应用上都与基于置信区间的结论相同（即 θ 的最小 90% 置信区间几乎等于最小的 90%

根据分布 (12.51) 确定的后验概率区间). 如果 $r \gg 1$ 并且 $(n-r) \gg 1$，那么对分布 (12.51) 的正态逼近将是有效的，并且 $100P\%$ 后验概率区间为 $(f \pm q\sigma)$，其中 q 是正态分布的 $(1+P)/2$ 百分位数；对于 90%、95% 和 99% 的置信水平，q 分别为 1.645、1.960 和 2.576. 在正态逼近有效的条件下，该结果与确切的置信区间之间的差异通常小于根据不同逼近方法计算得出的各种已公开发表的置信区间之间的差异.

如果 $r = (n-r) = 1$，那么分布 (12.51) 简化为 $p(\mathrm{d}\theta|r,n) = \mathrm{d}\theta$，正是贝叶斯和拉普拉斯作为先验的均匀分布. 因此，我们可以将贝叶斯-拉普拉斯先验解释为不是在描述一种完全无知的状态，而是在描述已经有一次成功和一次失败的知识状态. 因此，如果先验信息向我们保证实验在物理上有可能成功或失败，那么贝叶斯-拉普拉斯的选择将是正常先验，但是完全无知的分布 (12.50) 则描述了我们甚至不确定这一点的“前先验”知识状态.

如果 $r = 0$ 或 $r = n$，那么分布 (12.51) 的推导过程将无效，并且后验分布仍然无法归一化，分别与 $\theta^{-1}(1-\theta)^{n-1}$ 或 $\theta^{n-1}(1-\theta)^{-1}$ 成正比. 权重全部集中在值 $\theta = 0$ 或 $\theta = 1$ 上，因此先验分布 (12.50) 解释了我们在化学药品问题中注意到的归纳推理，这也是我们可以凭直觉得出的结论. 但是，我们一旦看到至少一次成功和一次失败，就知道该实验是一个真正的物理上的二元实验. 从那时起，所有后验分布 (12.51) 都可归一化，允许对 θ 进行确定的推断.

因此变换群方法得到的先验似乎也满足针对拉普拉斯连续法则提出的一致反对，但是我们也能看到分布 (12.50) 或贝叶斯-拉普拉斯先验是否恰当地依赖于具体的先验信息.

12.4.4　贝特朗问题

最后，我们给出一个例子，其中可以使用变换群发现更多的有信息先验. 贝特朗问题（Bertrand，1889）最初描述为“随机”画一条与圆相交的直线. 更具体地考虑这一问题有助于我们理解，假设我们不违背问题作者的原意（比如，它仍然是“随机”的），比如我们是向圆上扔稻草，但是没有指定如何扔. 因此，我们将问题表述如下.

将一根长长的稻草随机扔到一个圆上，假定它落下时与圆相交，那么相交形成的弦的长度大于该圆内接正三角形的边长的概率是多少?自从贝特朗在 1889 年提出该问题以来，它已经被讲给一代代学生，以证明拉普拉斯的“无差别原则”包含逻辑上的矛盾. 因为似乎有很多种不同的方法来定义“同等可能”情况，它们会导致不同的结果. 有三种方法是指定均匀概率密度给 (A) 弦中心点与圆心之间

的线段，(B) 相交弦在圆周上的相交角，(C) 弦中心在圆内部区域的位置. 这三种分配方式分别导致结果为 $p_A = 1/2$、$p_B = 1/3$ 和 $p_C = 1/4$.

哪个答案是正确的?在 10 位作者（Bertrand 1889, Borel 1909, Poincare 1912, Uspensky 1937, Northrop 1944, von Mises 1957, Gnedenko 1962, Kendell & Moran 1963, Mosteller 1965）中，只有博雷尔（Borel）愿意表达明确的偏向，尽管他没有提供任何证明. 冯 · 米泽斯（Von Mises）则采取相反的态度，宣称此类问题（包括类似的布丰投针问题）根本不属于概率论的范畴. 包括贝特朗本人在内的其他人则持中间态度，只是说问题没有明确的解，因为问题是不适定的，“随机”一词未定义.

在概率论著作中，这种情况几乎被普遍地解释为，这表明必须完全拒绝无差别原则. 通常会得出进一步的结论：分配概率的唯一有效依据是某一随机试验的频率. 这样看来，回答贝特朗问题的唯一方法似乎就是进行实验.

但是我们真的相信通过“纯粹思想”不能预测如此简单的实验结果吗? 解决此问题的意义远不只是解决一个几何难题，因为正如本章结论中讨论的那样，概率论在物理实验中的应用通常会导致此类问题，它们开始看起来似乎是不确定的，有许多可能的解让我们不知如何选择. 例如，给定气体的平均粒子密度和总能量，预测其黏性. 问题的答案显然依赖于分子的确切空间和速度分布（实际上，它非常依赖于位置-速度的相关性），并且在给定数据中似乎没有任何东西告诉我们要假定哪种分布. 然而，物理学家在无差别原则的引导下做出了确定的选择，他们使我们对黏性及许多其他物理现象做出了正确且重要的预测.

因此，尽管在某些问题上，无差别原则使我们陷入悖论，但在另一些问题上，它却产生了概率论最重要与成功的应用. 在没有任何更好选择的情况下拒绝该原则将导致不可接受的后果. 许多年来，即使是那些自以为最忠实的概率频率定义的遵从者，也会设法忽略这些逻辑上的困难，以保留某些非常有用的解.

显然，我们应该更仔细地研究诸如贝特朗问题中存在的表面上的悖论. 将概率论应用于实际物理问题中有重要的一点是需要了解的.

显然，如果圆足够大并且抛掷者足够熟练，那么可以随意获得各种结果. 但是在必须用一个比圆大的“不确定区域”来描述抛掷者技能的极限情况下，弦长度的分布必定是一种可通过“纯粹思想”确定的唯一函数. 若认为概率论既不能告诉我们如何根据第一原理计算出此函数，又不能否认这样做的可能性，则将对概率论的应用范围造成很大的限制——对于一名物理学家来说，这是无法容忍的.

庞加莱（Poincaré, 1912）将不变性论证应用于此类问题. 这最近为肯德尔与莫兰（Kendall & Moran, 1963）所引用. 我们考虑在 xy 平面上“随机”画的

线，每一条线通过直线方程为 $ux+vy=1$ 的两个参数 (u,v) 来确定. 这样我们可以问：哪种概率密度 $p(u,v)\mathrm{d}u\mathrm{d}v$ 在欧几里得（旋转和平移）变换群下具有形式不变性？这是一个容易解决的问题，答案为 $p(u,v)=(u^2+v^2)^{-3/2}$（Kendall & Moran，1963）.

但是这似乎没有说服力. 后来的作者都忽略了庞加莱的不变性论证，并坚持贝特朗最初持有的该问题没有确定答案的判断. 由于问题陈述中并未说明直线的分布具有这种不变性，并且我们没有充分的理由认为真实实验中扔的稻草会具有这种不变性，这是可以理解的. 这种假设似乎只是一种直觉的判断，并没有比以上三种答案更坚实的基础. 所有这些都可以归结为通过直觉指定“相等可能”事件来猜测随机的稻草雨应该具有什么性质. 结果仍然是，不同的直觉判断会导致不同的结果.

以上观点是迄今为止最普遍的，显然代表了一种解释问题的有效方法. 如果我们可以找到另一种观点，问题根据这一观点确实具有确定解，并且可以定义这些解能通过实验验证的条件，那么虽然说这种新观点原则上比传统观点更“正确”可能是夸大其词，但它肯定是更实用的.

现在，我们提出这样的观点，并且从一开始就理解，我们现阶段并不关心各种事件发生的频率. 相反，我们会问：当唯一的可用信息是上述问题陈述中给出的信息时，哪种概率分布描述了我们的知识状态？这一分布必须符合第 1 章中所描述的一致性合情条件：在两个具有相同知识状态的问题中，我们必须分配相同的概率. 关键是：如果我们假设，尽管有许多未确定的方面，贝特朗问题仍有一个确定的解，那么问题的陈述会自动暗示某些不变性，而这绝不依赖于我们的直觉判断. 在找到解后，无论它与频率是否有任何对应关系，都可以将其作为贝叶斯推断的先验. 任何可能出现的频率都将被视为额外的奖励，这也证明了将其直接用于物理预测的合理性.

贝特朗问题具有明显的旋转对称性，这一点在所有提出的解中都被认识到. 但是，这种对称性与弦长的分布无关. 此外，还有两个高度相关的“对称性”：无论是贝特朗的原始陈述，还是我们通过稻草进行的重新陈述，都没有指定圆的确切大小或位置. 因此，如果问题要有一个确定的解，那么它必须不依赖于这些条件，即对圆的大小或位置进行小的改变时，解必须保持不变. 我们将会看到，这一看似微不足道的陈述完全决定了解.

通过定义一个四参数变换群，可以同时考虑所有这些不变性需求，于是完整的解将像变魔术一样突然出现. 但是，对这些不变性的作用进行单独分析，看看每一方面如何影响解的形式，将更具启发性.

旋转不变性

令圆的半径为 R，弦的位置由弦中点的极坐标 (r,θ) 表示. 我们试图回答一个比贝特朗问题更具体的问题：应该给圆内部分配怎样的概率密度 $f(r,\theta)\mathrm{d}A = f(r,\theta)r\mathrm{d}r\mathrm{d}\theta$? 实际上，对 θ 的依赖与贝特朗问题无关，因为弦长度的分布仅取决于径向分布

$$g(r) = \int_0^{2\pi} \mathrm{d}\theta f(r,\theta). \tag{12.54}$$

但是，直觉表明 $f(r,\theta)$ 应该与 θ 无关，正式的变换群理论对于旋转对称性的处理方法如下.

出发点是注意到问题陈述中并没有指明观察者的朝向，因此如果有确定解，那么它一定不依赖于观察者视线的方向，所以假设有两个不同的观察者 X 和 Y 正在观看此实验. 他们从不同的方向观看实验，视线之间的角度为 α. 每个人都使用沿其视线方向的坐标系. X 在系统坐标系 S 上分配概率密度 $f(r,\theta)$，Y 在系统坐标系 S_α 上分配概率密度 $g(r,\theta)$. 显然，如果它们描述的是相同的情况，则必须有

$$f(r,\theta) = g(r,\theta-\alpha), \tag{12.55}$$

它表示简单的变量变换，将固定分布 f 转换到新的坐标系. 无论问题是否具有旋转对称性，这一关系都将成立.

但是我们意识到，由于旋转对称性，问题在 X 的坐标系中看起来与 Y 在自己的坐标系中完全一样. 由于他们处于相同的知识状态，我们的一致性合情条件要求他们分配相同的概率分布. 因此 f 和 g 必须是相同的函数：

$$f(r,\theta) = g(r,\theta). \tag{12.56}$$

这一关系对于 $0 \leqslant \alpha \leqslant 2\pi$ 中的所有 α 都成立，因此唯一的可能性是 $f(r,\theta) = f(r)$.

与我们明显的直觉灵光相比，这种正式的论证可能显得笨拙. 当然，当将其应用于这样一个简单问题时，事实确实如此. 然而正如维格纳（Wigner，1931）和外尔（Weyl，1946）在其他物理问题中所表明的那样，正是这种笨拙的论证可以立即推广到一些直觉无法得到结论的不普通情形. 它总是由两个步骤组成：首先找到一个像 (12.55) 的变换方程，该方程显示两个问题如何相互关联，而与对称无关；然后类似 (12.56) 的对称关系表明我们已经提出了两个等价的问题. 将两个步骤结合在大多数情况下会得到一个对其分布形式有所限制的函数方程.

尺度不变性

根据旋转对称性，问题已经简化为确定函数 $f(r)$. 根据归一化条件有

$$\int_0^{2\pi} \mathrm{d}\theta \int_0^R r\mathrm{d}r f(r) = 1. \tag{12.57}$$

我们再考虑两个不同的问题. 考虑与半径为 R 的圆同心、半径为 aR（$0 < a \leqslant 1$）的圆. 在较小的圆中，有概率密度 $h(r)r\mathrm{d}r\mathrm{d}\theta$，它回答了这一问题：假定一根稻草与较小的圆相交，其弦中点位于 $\mathrm{d}A = r\mathrm{d}r\mathrm{d}\theta$ 区域内的概率是多少?

任何与小圆相交的稻草也会定义一条在较大圆上的弦，因此在小圆内 $f(r)$ 一定与 $h(r)$ 成比例. 当然，此比例是由条件概率的标准公式给出的，在这种情况下其形式为

$$f(r) = 2\pi h(r) \int_0^{\alpha R} r\mathrm{d}r f(r), \quad 0 < \alpha \leqslant 1, \quad 0 \leqslant r \leqslant aR. \tag{12.58}$$

这一变换方程无论问题是否具有尺度不变性都将成立.

但是我们现在应用尺度不变性：对于眼球大小不同的两位观察者，大圆和小圆的问题将显得完全相同. 如果存在独立于圆大小的唯一解，则 $f(r)$ 和 $h(r)$ 之间一定存在另一种关系，它表明一个问题仅仅是另一个问题的按比例缩小版本. 面积 $r\mathrm{d}r\mathrm{d}\theta$ 和 $(ar)\mathrm{d}(ar)\mathrm{d}\theta$ 分别以相同的方式与大小圆相关，所以必须分别通过分布 $f(r)$ 和 $h(r)$ 给它们分配相同的概率：

$$h(ar)(ar)\mathrm{d}(ar)\mathrm{d}\theta = f(r)r\mathrm{d}r\mathrm{d}\theta, \tag{12.59}$$

或者

$$a^2 h(ar) = f(r), \tag{12.60}$$

这就是对称方程. 结合 (12.58) 和 (12.60)，我们看到尺度不变性要求概率密度满足函数方程

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} u\mathrm{d}u f(u), \quad 0 < a \leqslant 1, \quad 0 \leqslant r \leqslant R. \tag{12.61}$$

将以上方程对 a 微分，并令 $a = 1$，求解所得的微分方程，我们发现满足归一化条件 (12.57) 的方程 (12.61) 的最一般解为

$$f(r) = \frac{qr^{q-2}}{2\pi R^q}, \tag{12.62}$$

其中 q 是满足 $0 < q < +\infty$ 的常数，不再由尺度不变性确定.

我们注意到引言中提出的解 B 已经被排除了，因为它对应于选择函数 $f(r) \sim 1/\sqrt{R^2 - r^2}$，其形式与 (12.62) 不一致. 这意味着，如果圆周上弦的相交部分以角度均匀且独立地分布在一个圆上，则对于居于其中的较小的圆就不成立，也就

是说解 B 的概率分配最多对于仅仅一个大小的圆成立. 但是解 A 和解 C 仍然与尺度不变性兼容，分别对应于选择 $q=1$ 和 $q=2$.

平移不变性

现在，我们研究给定稻草 S 可以与两个具有相同半径 R 但相对位移为 b 的圆 C 和 C' 相交这一事实的后果. 参照图 12-1，相对于圆 C 的弦中点是点 P，坐标为 (r,θ)，而同一稻草定义的相对于圆 C' 的弦中点 P' 的坐标为 (r',θ'). 根据图 12-1，$(r,\theta)\to(r',\theta')$ 的坐标变换为

$$r'=|r-b\cos\theta|, \tag{12.63}$$

$$\theta'=\begin{cases}\theta, & r>b\cos\theta,\\ \theta+\pi, & r<b\cos\theta.\end{cases} \tag{12.64}$$

当 P 在区域 Γ 内变化时，P' 在 Γ' 内变化，反之亦然. 因此稻草定义了 Γ 到 Γ' 的一一映射.

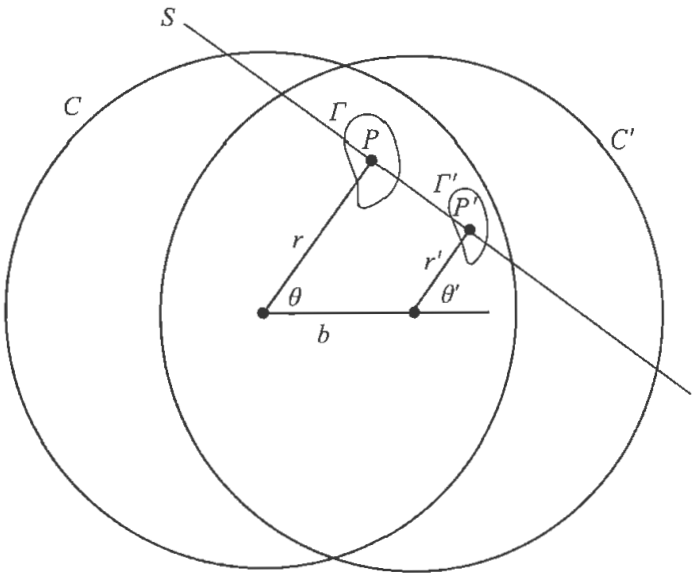

图 12-1 稻草 S 与两个略微移位的圆 C 和 C' 相交

现在考虑平移不变性. 由于问题陈述中没有给定圆位置的信息，C 和 C' 的问题对于两个稍有移位的观察员 O 和 O' 是相同的，因此我们的一致性合情条件要求它们对于 C 和 C' 必须分配相同的如 (12.62) 的形式和相同的 q 值的概率密度.

进一步要求两个观察者对于区域 Γ 和 Γ' 分配相同的概率分布，由于 (a) 它们是同一事件的概率，(b) 一根稻草与一个圆相交也将与另一个圆相交，因此建立起这种对应关系的概率在两个问题中是相同的. 让我们看看这两个需求能否兼容.

弦与 C 相交的中点在 Γ 中的概率为

$$\int_{\Gamma} r\mathrm{d}r\mathrm{d}\theta f(r)=\frac{q}{2\pi R^q}\int_{\Gamma}\mathrm{d}r\mathrm{d}\theta r^{q-1}. \tag{12.65}$$

弦与 C' 相交的中点在 Γ' 中的概率为

$$\frac{q}{2\pi R^q}\int_{\Gamma}\mathrm{d}r'\mathrm{d}\theta'(r')^{q-1}=\frac{q}{2\pi R^q}\int_{\Gamma}\mathrm{d}r\mathrm{d}\theta\,|r-b\cos\theta|^{q-1}, \tag{12.66}$$

这里我们已经应用 (12.63) 和 (12.64) 将积分转换回变量 (r,θ) 上，注意雅可比行列式是 1. 显然，对于任意 Γ，当且仅当 $q=1$ 时，(12.65) 和 (12.66) 相等，因此分布 $f(r)$ 现在是唯一确定的.

引言中提出的解 C 由于缺乏平移不变性也被排除了. 对于一个圆具有假定性质的一堆稻草，对于稍微移位的圆不具有相同的性质.

我们发现不变性要求决定了概率密度为

$$f(r,\theta)=\frac{1}{2\pi Rr},\quad 0\leqslant r\leqslant R,\quad 0\leqslant\theta\leqslant 2\pi, \tag{12.67}$$

对应于引言中的解 A. 有趣的是，它在中心有一个奇点，对此的理解如下. 中点 (r,θ) 落在小区域 Δ 的条件对弦的可能方向施加了限制，但是随着 Δ 向内移动，一旦它包含了圆心，就会突然允许所有的角度，因此“概率流形”会无限快速地变化.

进一步的分析（对于图 12-1 进行思考几乎可以明显看出）表明，平移不变性的要求如此严格，以至于它已经唯一地确定了结果 (12.67)，因此引言中提出的解 B 与尺度不变性和平移不变性都不兼容. 为了得到解 (12.67)，实际上没有必要考虑尺度不变性. 但是无论如何，都必须对解 (12.67) 进行尺度不变性测试，如果解未能通过该测试，我们将得出结论，提出的问题没有解. 也就是说，尽管乍看之下这一问题似乎是欠定的，但从变换群的角度来看，它应该被认为是超定的. 幸运的是，这些要求是兼容的，所以这个问题有唯一解.

根据解 (12.67) 立即可以得到弦长的分布. 弦中点在 (r,θ) 处的长度为 $L=2\sqrt{R^2-r^2}$. 根据归一化的弦长 $x\equiv L/2R$，我们得到普遍的分布定律

$$p(x)\mathrm{d}x=\frac{x\mathrm{d}x}{\sqrt{1-x^2}},\quad 0\leqslant x\leqslant 1, \tag{12.68}$$

这与博雷尔（Borel，1909）的猜想一致.

频率对应

从其推导过程来看，分布 (12.68) 似乎只具有主观意义. 尽管贝特朗问题的陈述中有很多方面没有说明，这一分布描述了与唯一解相对应的唯一可能的知识状态，但是我们还没有理由认为它与实际实验中观察到的频率有任何关系. 当然，我们一般也不能这样断言. 仅凭我们的知识状态没有理由更倾向于两个事件中的

某一个这一事实，并不足以保证两件事实际上会以同样的可能发生！事实上，显然任何“纯思想”论证——无论是基于变换群还是任何其他原则——都不能确定地预测在实际实验中一定会发生什么. 我们很容易想象出一种精密机器，该机器能以指定的任意弦长分布抛掷稻草.

但是，我们有权对结果 (12.68) 宣称某种明确的频率对应. 因为以上推导过程已经证明了一个“客观事实”：任何与分布 (12.68) 不一致的稻草雨一定会在不同的圆上产生不同的分布.

为了肯定地预测分布 (12.68) 将在“不确定区域”比圆大的任何实验中被观察到，这是我们所需要的所有结论. 因为我们如果缺乏肯定能与圆相交的扔稻草的技巧，那么肯定也缺乏在此不确定区域内在不同圆上一致地扔出不同分布的技巧.

正是出于这一原因，通过变换群方法预测的分布最终具有频率对应. 严格来说，该结果仅在“完全无技巧”的极限情况下成立. 但是，稍微想一下就能明白，产生与分布 (12.68) 存在任何显著偏差所需的技巧是如此之高，以至于在实践中即使使用机器也很难实现.

这些结论似乎与冯 · 米泽斯（von Mises，1957）的结论直接矛盾. 冯 · 米泽斯完全否认此类问题属于概率论领域. 在我们看来，如果严格且一致地采用冯 · 米泽斯的概率论哲学，那么概率论的合法物理应用范围几近为零. 由于我们已经做出了明确的预测，该问题现在已经从哲学领域转移至可验证的科学领域. 通过实验，可以很容易地验证变换群方法应用于此问题以及其他问题的预测能力.

实际上，我和查尔斯 · 泰勒博士做过贝特朗实验，将稻草从垂直位置扔到位于地板上直径 5 英寸的圆上. 结果将弦长分为 10 类，128 次成功抛掷以极低的卡方值确认了 (12.68) 的正确性. 但是如果这一实验结果是由其他人报告的，无疑将更令人信服.

12.5 评注

贝特朗问题比初看起来更重要，因为它是从一开始就弥漫在概率论领域中的一个更深层悖论的简单缩影. 在“实际”物理应用中，每当我们尝试使用概率术语来表达感兴趣的问题时，几乎总是会发现类似贝特朗的陈述. 因为显然有重要的事情没有说明，它似乎太模糊了，无法得到任何明确的解.

我们将详细说明 12.4.4 节引言中提到的例子：给定体积为 V 的 N 个分子组成的气体，其分子间力的总能量 E，预测其分子速度分布、压力、压力起伏分布、黏度、导热系数和扩散常数. 再一次，大多数概率论作者表达的观点是，因为该问题是不适定的，所有没有确定解，所指定的宏观状态不足以确定微观状态上的

任何唯一概率分布. 如果我们拒绝无差别原则，并且坚持认为在随机试验中分配概率的唯一有效基础是频率，那么确定这些量的唯一方法似乎就是进行实验.

但是，一个多世纪以前的历史记录表明，在没有任何有关分子位置与速度的频率数据的情况下，麦克斯韦通过“纯思想”的概率分析正确地预测所有这些量的值，这相当于认识到某种“等可能”情况. 对于黏度，所预测的与密度的依赖关系最初似乎与常识相矛盾，这使得人们对麦克斯韦的分析产生了怀疑. 但是进行实验后，人们证实了麦克斯韦的预言，导致了动力学理论的第一次伟大胜利. 这些是实实在在的成就，如果麦克斯韦不使用无差别原则，就不可能取得这些成就.

同样，我们计算扑克牌中获得各种牌型的概率，并且对结果非常有信心，因此我们愿意冒险对计算表明的、对我们有利的情形下注. 然而这些计算的基础是对所有牌的分布等可能的直觉判断. 如果使用不同的判断，我们将得出不同的计算结果. 我们再次通过“纯思想”论证来预测确定、可验证的事实，这完全基于存在“等可能”情况的认识. 而目前的统计学，无论是正统统计还是主观贝叶斯统计，都否认无差别原则是分配概率的有效依据!

这里的两难困境显而易见: 一方面，通过指出诸如贝特朗悖论之类的东西，人们不能否认使用无差别原则的模糊性与危险性; 另一方面，同样不可否认的是，人们使用这一原则一遍遍地得出了正确、重要且有用的预测结果. 因此看起来，尽管我们不能完全接受无差别原则，但也不能完全拒绝它，因为这样做会将一些最重要和成功的应用排除在概率论之外.

变换群方法源于作者对无差别原则在过去一直受到不公平对待的信念. 我们需要的不是一味地非难这一原则，而是认识到使用它的正确方法. 与大多数其他概率论学者一样，我们认为在事件层次使用无差别原则是危险的，因为正如贝特朗悖论所表明的那样，我们的直觉在这类事情上的引导是非常不可靠的.

我们认为无差别原则可以合理地应用于更抽象的问题层次，因为那是已经由问题陈述确定的，与我们的直觉无关. 在问题陈述中未指定的每种情况都定义了一种不变性，如果有确定解，那么这一解必须具有这种不变性. 变换群数学地表达了这些不变性，对解的形式施加了确定的限制，并且在许多情况下完全确定了它.

当然，并非所有的不变性都是有用的. 例如，贝特朗问题的陈述中未指定扔稻草的时间、圆的颜色、猎户座 α 星的亮度或切萨皮克海湾的牡蛎数量. 如果所描述的问题有唯一解，就可以正确地推断出它一定不取决于这些情况. 但是除非我们之前认为这些事情是相关的，否则这些信息对我们没有帮助.

对多个示例的研究表明，上面提到的两难困境现在可以通过如下方式解决. 我们认为，过去成功应用无差别原则的情况是，问题的解可以重新描述，使得实

际计算的过程是在问题而不是事件之间应用无差别原则.

杰弗里斯先验的变换群推导过程使我们能够以新的视角看待这一先验. 一直以来, 也许很明显的是, 杰弗里斯规则的真正依据不能仅仅在于参数为正的事实. 举一个简单的例子, 假设已知 μ 是一个位置参数, 那么直觉和前面的分析都认为, 均匀先验密度是表达 μ 完全无知的正确方法. 关系 $\mu = \theta - \theta^{-1}$ 定义了从 $-\infty < \mu < +\infty$ 到 $0 < \theta < +\infty$ 的一一映射, 但是杰弗里斯规则不能应用于参数 θ, 一致性要求其先验密度与 $\mathrm{d}\mu = (1 + \theta^{-2})\mathrm{d}\theta$ 成正比. 看来杰弗里斯规则的基本依据不仅是参数为正, 而且它必须是一个*比例参数*.

由变换群发现的表征完全无知的分布不能归一化的事实可以通过以下两种方式解释. 一方面可以说, 这仅仅是因为我们对完全无知的表述是一种理想化, 并不严格适用于任何实际问题. 从圣路易斯到仙女座星云的位置变化, 或者从原子大小到银河系大小的尺度变化, 都不会将任何世俗关心的问题转化为完全等价的问题. 在实际问题中, 我们总是具有关于位置和比例的某些先验知识, 因此群参数 (a, b) 不能在真正无限范围内变化. 因此, 严格来说, 变换 (12.50) 不会形成一个群. 但是, 在确实表达了我们先验无知的范围内, 上述论证仍然适用. 在此范围内, 函数方程式和先验的结果形式一定仍然成立.

另一方面, 我们对最大熵方法的讨论显示了看待这一问题的一种更具建设性的方法. 找到完全无知的分布只是找到任何现实问题的先验的第一步. 虽然变换群产生的先验分布并不严格表示任何现实的知识状态, 但是它确实定义了我们参数空间的不变度量. 没有该方法, 通过最大熵找到现实先验的问题在数学上是不确定的.

第 13 章　决策论：历史背景

“从结果来看，你的行为是不明智的.”我大声说.
他庄严地注视着我，然后说道：
“在选择行动路线时，并没有结果在引导我.”
——安布罗斯·比耶尔斯（Ambrose Bierce）

在前面的几处讨论中，我们插入了括号以表示“此处仍然缺少一个要点，当我们应用决策论时将提供此要点”. 虽然将这一主题推迟到现在讨论，但是我们并没有剥夺读者所需的技术工具，因为在我们看来，决策问题的求解是如此直接和直观，甚至不需要诉诸任何基础的正式理论.

13.1　推断与决策

从我们将概率论应用于第一个问题开始，推断与决策的评估问题就出现了. 当我们在第 4 章通过序列检测说明贝叶斯定理的应用时，就注意到了概率论本身没有什么可以告诉机器人以改变其决定——无论是接受、拒绝这批样品，还是继续进行检测——的临界水平值. 这些临界水平值的确定，除了依赖于概率之外，显然在某种程度上还取决于价值判断：做出错误决策的后果是什么，继续进行检测的代价是什么?

在第 6 章中，当机器人进行参数估计时，也面临相同的状况. 概率论只决定机器人对于参数的知识状态，并不能告诉机器人实际上应该做出怎样的估计. 我们当时注意到，取后验 PDF 的均值与最小化期望均方误差的决策是等价的. 但是我们也指出，在某些情况下，我们可能更偏好中位数.

定性和直观地讲，这些考虑已经很清楚了. 但是在我们宣称已经对机器人进行真正完整的设计之前，必须澄清这里的逻辑，并证明我们的流程并不仅仅是基于直觉的特定工具——按照某一明确定义的准则，它是最优的. 沃尔德的决策论旨在实现这一目标.

目前所考虑的所有问题都有一个共同特征：概率论只能解决推断问题，也就是说它只能为我们提供一种概率分布，该分布代表了考虑所有的先验信息和数据的机器人的最终知识状态. 但是实际上，工作还没有结束. 在我们的机器人设计中仍然缺少的一个核心要素是，将最终的概率分配转换为确定的行动过程的规则.

但是对我们来说，正式的决策论只是合理化——而不是改变——我们的直觉已经告诉我们要做的事情.

决策论对我们而言具有不同的意义，因为它使我们充分理解持续了几百年的关于概率论基础的争议. 关于什么是概率论，有两种截然相反的观点，决策论可以从这两者之中很好地推导出来，因此在它们之间构建了一座桥梁，并且暗示了决策论可能有助于解决其中的争议. 我们将在这里详细介绍决策问题的两种解决方案的历史背景以及它们之间的联系.

13.2 丹尼尔·伯努利的建议

正如人们根据概率论最基本的应用中出现的情况所能预期的那样，两种决策论方法之间的关系并不是一个新问题. 丹尼尔·伯努利（Daniel Bernoulli，1738）就清楚地认识到这一点，并为某一类问题提供了明确的解决方案. 相同原理的粗略形式甚至可以追溯到更早的时间，在概率论几乎只涉及赌博问题时甚至更早. 尽管今天我们似乎很难理解，但历史记录清楚地表明，“期望收益”的概念对于概率论的早期工作者来说非常直观，甚至比概率更直观.

考虑 n 种可能性，$i=1,2,\cdots,n$，给它们分配概率 p_i 以及表示第 i 种可能性为真时我们将获得的收益 M_i. 那么按照我们的标准表示法，收益的期望是

$$E(M)=\langle M\rangle=\sum_{i=1}^{n}p_iM_i. \tag{13.1}$$

17 世纪生意兴隆的阿姆斯特丹商人们就像有形商品一样买卖期望. 在许多人看来，一个完全基于个人利益行事的人应该始终以获得最大化期望收益的方式行事. 但是这导致了一些悖论（尤其是著名的圣彼得堡问题），这使得伯努利认识到简单的最大化期望收益并不总是明智的行动标准.

例如，假设你拥有的信息使得你为某个有偏硬币正面朝上分配了 0.51 的概率. 现在，你有两种选择：(1) 用你所有的钱押注下一次抛硬币时正面朝上；(2) 根本不下注. 根据最大化期望收益的标准，当面临以上选择时你应该始终选择下注. 如果不下注，你的期望收益是 0；但是如果你下注，期望收益是

$$\langle M\rangle=0.51M_0+0.49(-M_0)=0.02M_0>0, \tag{13.2}$$

其中 M_0 是你现在有的钱. 然而对于伯努利来说，正如对于普通读者一样，显而易见的是，没有一名头脑清醒的人会做出第一种选择. 这意味着在某些情况下，我们的常识会拒绝最大化预期收益的标准.

假设你有以下机会：你可以下注任意金额，并将以概率 $1-10^{-6}$ 输钱，以 10^{-6} 的概率赢得 1 000 001 倍下注金额. 这次，最大化预期收益的标准仍然表明，

你应该拿所有的钱下注. 常识会更加强烈地拒绝这种选择.

丹尼尔 · 伯努利提出了解决这些悖论的方法. 他认识到，一个人获取一定数量金钱的真正价值并不能简单地通过所得到的金钱数来衡量，还取决于他已经有多少钱. 伯努利还换过一种方式说，我们应该认识到收益的数学期望与“道德期望”不同. 一名现代经济学家在谈到“货币的边际效用递减”时也是在表达同样的思想.

在圣彼得堡博弈中，我们抛掷一枚无偏的硬币，直到它第一次出现正面，游戏就此结束. 如果在第 n 次抛掷时出现正面，玩家将获得 2^n 美元. 问题是：为了获得玩这个游戏的权利，玩家要支付的“公平”入场费是多少? 如果我们使用公平游戏的标准是入场费等于期望收益，你会看到将发生什么. 期望是无限的：

$$\sum_{k=1}^{\infty}(2^{-k})(2^k)=\sum_{k=1}^{\infty}1=+\infty. \tag{13.3}$$

然而还是很清楚的是，除只用很少的本钱，没有一名理智的人会愿意冒险玩这个游戏. 在这一点上，我们引用拉普拉斯（Laplace，1814，1819）的话：

> 确实，很明显，1 法郎对一个只拥有 100 法郎的人比对一个百万富翁来说价值要高得多. 这样，我们应该将期望收益的绝对价值与其相对价值区分开. 后者取决于使之成为理想的动机，而前者则与之无关. 我们不能给出分配该相对价值的一般性原则，但丹尼尔 · 伯努利给出了一个原则，该原则在许多情况下可以使用：无限小额的金钱的相对价值等于其绝对值除以相关人的总财富.

换句话说，伯努利认为金钱数 M 的“道德价值”或现代经济学家所称的“效用”与 $\ln M$ 成正比. 拉普拉斯在讨论圣彼得堡问题和这一标准时，没有给出计算过程就报告了以下结果：一个总财富为 200 法郎的人在这场游戏中的赌注不应超过 9 法郎. 180 年后，让我们检查一下拉普拉斯的计算.

对于初始“财富”为 m 法郎的人，公平入场费 $f(m)$ 是通过将其当前效用与如果他付费玩游戏的期望效用相等来确定的，即 $f(m)$ 是方程

$$\ln m=\sum_{n=1}^{\infty}\frac{1}{2^n}\ln(m-f+2^n) \tag{13.4}$$

的根. 通过计算机计算得出 $f(200)=8.7204$. 没有计算机的拉普拉斯的计算结果非常好. 我们同样可以得到，$f(10^3)=10.95$, $f(10^4)=14.24$, $f(10^6)=20.87$. 即使是百万富翁，在此可疑游戏中愿意承担的风险也不应该超过 21 法郎.

在我们看来，这个数值结果是完全合理的. 然而，效用的对数分配，无论是

在（正如拉普拉斯所指出的）极少财富的情况下，还是在极大财富的情况下，都不能简单加以接受，如以下萨维奇（Savage，1954）的例子所示.

假设你当前的财富为 1 000 000 美元，如果你的金钱效用与金额的对数成正比，那么相对于不接受，你应该同样有可能接受以下赌博方式：你有一半可能性只剩下 1000 美元，而另一半可能性将拥有 1 000 000 000 美元. 大多数人会认为这样的赌注对拥有前面提到的初始财富的人显然是不利的. 这表明我们直觉上金钱的“效用”的增长速度对于非常大的数值而言应该比对数增长速度更快. 切尔诺夫和摩西（Chernoff & Moses，1959）声称它是有界的，从理论上看来这似乎是合理的，但是在现实世界中并没有真正得到证明.

因此，丹尼尔·伯努利建议的要点是，在面对不确定性的决策问题中，人们应该按照最大化期望价值来行动. 期望价值未必就是收益，而是收益的某一函数，他称之为“道德价值”. 使用更现代的术语，乐观主义者称其为“最大化期望效用”，而悲观主义者则称其为“最小化期望损失”，“损失函数”等于效用函数的相反数.

13.3 保险的理论依据

让我们以保险为例简要说明上述一些观点，这在某些方面类似于圣彼得堡博弈. 以下场景显然过于简化，但是可以从中得到一些有效且重要的结论. 保险费总是要设置得足够高，以保证保险公司对于合同涉及的所有意外费用都有正的期望收益. 公司赚的每一分钱都是客户支付的，那为什么还有人愿意买保险呢?

关键在于，个人客户对于金钱的效用函数在 1000 美元左右可能会急剧变化，但是保险公司的资产规模是如此之大，以至于其效用函数在数百万美元的范围内仍呈线性关系. 因此，令 P 为某一保险合同的保费，令 $i=1,\cdots,n$ 列举所涵盖的所有意外且第 i 项具有概率 w_i，如果出险则保险公司将产生费用 L_i. 假定潜在客户对于金钱具有丹尼尔·伯努利的对数效用函数并具有初始金额 M. 当然，我们应该将 M 理解为他具有的“资产净值”，而不仅仅是他手头上的现金量. 那么保险公司和个人客户（无论是否购买保险）的期望效用如表 13-1 所示. 因此如果 $\langle L\rangle < P$，则公司希望出售该保险，而如果 $\langle \ln(M-L)\rangle < \ln(M-P)$，则客户希望购买该保险. 如果保费在以下范围内：

$$\langle L\rangle < P < M - \exp\langle \ln(M-L)\rangle, \tag{13.5}$$

将对双方都有利，双方都愿意做这笔生意.

我们将其作为练习让读者根据 (13.5) 证明：穷人应该购买保险，但是富人除非对期望损失 $\langle L\rangle$ 的估计比保险公司的估计大得多，否则不应该购买保险. 的确，

表 13-1　期望效用

	买	不买
公司	$P-\sum w_iL_i$	0
客户	$\ln(M-P)$	$\sum w_i\ln(M-L_i)$

如果你现在拥有的财富远大于任何可能的损失，那么你的金钱效用在所关心的区间内就几乎与保险公司一样呈线性，还不如自己成为自己的保险公司.

注意到如果 $M \gg \langle L\rangle$，我们可以展开 M^{-1} 的幂：

$$M-\exp\langle\ln(M-L)\rangle=\langle L\rangle+\frac{\mathrm{var}(L)}{2M}+\cdots, \tag{13.6}$$

其中 $\mathrm{var}(L)=\langle L^2\rangle-\langle L\rangle^2$，这样可以进一步了解富人的心理. 因此，即使保费略高于他的期望损失，中等富裕的人也可能愿意购买保险，因为这样可以消除他将不得不承受的实际损失的不确定性 $\mathrm{var}(L)$. 我们不仅对风险，而且对不确定性有厌恶感.

将 (13.5) 的右侧写为形式

$$M-\exp\langle\ln(M-L)\rangle=M-\prod_i\exp\{w_i\ln(M-L_i)\} \tag{13.7}$$

可以进一步了解穷人的心理. 排列 L_i 使得 $L_1\geqslant L_2\geqslant L_3\geqslant\cdots$，那么除非 $M>L_1$，否则该表达式无意义. 但可以假定不可能使得 $M<L_1$，因为一个人的损失不能超过他的所有. 但是如果 M 接近 L_1，则最后一项将接近于 $\mathrm{e}^{-\infty}$ 并可以忽略. 这样 (13.5) 可以简化为 $\langle L\rangle<P<M$. 看起来这个不幸的人始终应该购买保险，即使这会使他像发生了最严重的意外事故一样变得贫穷!

当然，这仅仅说明对数效用对于很小的数量是不现实的. 实际上，效用显然也局限于该区间. 只有一分钱的人不会认为失去它是一场灾难. 我们可以通过将 $\ln M$ 替换为 $\ln(M+b)$ 来修正此问题，其中 b 的值很小，以至于我们认为它实际上毫无价值. 这在某种程度上修正了我们根据 (13.7) 得出的结论. 我们将具体结论留给读者得出，也许可以为 b 提供一个合理值.

13.4　熵与效用

对数效用分配对于许多情况是合理的，只要不把它推向极端. 顺便提一下，它也与熵的概念紧密相关，如贝尔曼和卡拉巴（Bellman & Kalaba，1956，1957）所示. 一名在游戏中预先获得部分可靠小窍门的赌博者会采取行动（即决定在赌哪一边及下注多少）来最大化期望对数财富. 贝尔曼和卡拉巴指出：(1) 遵循这一策略永远不会破产，这与最大化期望收益的策略形成了鲜明对比，在后一种策略

下，很容易看到破产将最终以 1 的概率发生（经典的“赌徒破产”情况）；(2) 一个人在任何一局游戏中可以期望赚取的金额显然与他的初始金额 M_0 成正比，因此在 n 局游戏之后，他可能预期得到的金额为 $M = M_0 \mathrm{e}^{\alpha n}$. 显然，使用对数效用函数的作用是使 α 的期望最大化.

练习 13.1 证明可达到的 $\langle \alpha \rangle$ 最大值只是 $H_0 - H$，其中 H 是描述赌博者对所获小窍门真实有效性的不确定性的熵，而 H_0 是所获小窍门完全没有信息时的最大可能熵.

类似的结果也会在后面导出. 这表明随着概率论的进一步发展，熵可能在指导商人或股票市场投资者的策略中具有重要的地位.

这些考虑有着更微妙的用途：不仅有可能最大化我们自己的效用，而且有可能通过巧妙地利用他人对于效用的考虑因素，诱使他们按照我们的意愿行事. 能干的管理者本能地（但只是定性地）知道如何进行奖励和惩罚，以保持其组织的平稳运行. 下面是一个大大简化但是定量的例子.

13.5 诚实的天气预报员

假设天气预报员根据先验信息和数据得出明天下雨的概率为 $p = P(\text{降雨}|\text{数据}, I)$，那么他在晚间电视天气预报节目中公开宣告的概率 q 是多少? 这取决于他的效用函数. 我们怀疑天气预报员会系统性地夸大恶劣天气的概率，即宣布的值 $q > p$，因为他们认为如果没有预测暴风雨的到来，将会招致更多的批评. [①]

然而我们更希望被告知由当数据前预示的 p 的实际值. 的确，如果我们确定被告知这一点，并且我们是理性的，就不能批评天气预报员的预测错误. 是否可以给天气预报员创造一个环境，使他总是愿意说实话呢?

假设我们在与天气预报员签订的雇用合同上规定，他永远不会因为做出太多的错误预测而被解雇. 但是每一天，当他宣布下雨的概率 q 时，如果第二天实际下雨，则该天的工资为 $B\ln(2q)$，否则为 $B\ln(2[1-q])$，其中 B 是基本工资常数，对于我们目前的考虑并不重要，只要足以使他想要这份工作即可. 那么，如果天气预报员宣布概率 q，则今天的预期工资为

$$B[p\ln(2q) + (1-p)\ln(2[1-q])] = B[\ln 2 + p\ln q + (1-p)\ln(1-q)]. \tag{13.8}$$

取一阶和二阶导数，我们发现当 $q = p$ 时会得到最大值.

① 这方面的证据是：在圣路易斯，我们几乎每隔一周就会遇到一次天气预报说有暴风雨，但是实际上没有. 但是没有预测到的暴风雨非常罕见，会成为重大新闻.

由于任何连续的效用函数的一小部分都为线性，因此如果天气预报员认为一天的工资足够少，以至于其效用是线性的，那么说实话永远对他有利. 事实上，存在着奖励和效用函数的组合，使得诚实是最好的策略.

更一般地，假如存在 n 个可能的事件 $(A_1, \cdots, A_n)$，根据先验信息和数据预测的概率为 $(p_1, \cdots, p_n)$，但是预测者选择改为宣告概率为 $(q_1, \cdots, q_n)$. 如果随后发生事件 A_i，则让他获得 $B\ln(nq_i)$ 的报酬，即他因对真实事件赋予很高的概率而受到奖励. 那么他的薪水期望值是

$$B[\ln n - I(q;p)], \tag{13.9}$$

其中 $I(q;p) \equiv \sum p_i \ln q_i$ 本质上是分布的相对熵，今天通常称为库尔贝克-莱布勒信息（Kullback & Leibler，1951），尽管其基本性质已经被吉布斯（Gibbs，1902，第 11 章）证明和利用. 那么，宣告 $q_i = p_i$ 对天气预报员总是有利的，而他的最大期望报酬是

$$B[\ln n - H(p_1, \cdots, p_n)], \tag{13.10}$$

其中 $H(p_i) = -\sum p_i \ln p_i$ 是衡量他对 A_i 不确定性的熵. 不仅说实话对他有利，获得最大可能的信息量以减少该熵也对他有利.

举一个非常真实的具体例子，考虑一家只有有限研发资源的制药公司. 它有两种潜在的新药：药物 A 能治疗每年折磨 10^6 人的疾病，而药物 B 每年仅可帮助 1000 人. 假设初步证据表明两种药物具有同样的有效性和安全性，该公司自然会更愿意将研发资源投入到药物 A 上. 我们可以肯定地预测这个决定会受到一些愤世嫉俗者的攻击，他们会指责制药公司只对自己的利益感兴趣. 然而，如果深入思考，他可能会意识到，这一策略虽然无可否认会给公司带来好处，但是也会给社会上的更多人带来好处.

13.6　对丹尼尔·伯努利和拉普拉斯的反应

得到这些结果使用的数学知识很基础，但显然很重要. 这可能使人们认为，一旦丹尼尔·伯努利和拉普拉斯开始了这种思维方式，这种事情一定不仅会被许多人觉察到，而且会立即得到很好的利用. 确实，回顾历史，令人感到惊讶的是，在吉布斯之前的 100 年中竟然没有人在这一过程中发现熵的概念.

实际的历史进程截然不同. 在 20 世纪的大部分时间里，“频率主义”学派要么忽略上述推理方式，要么谴责其为形而上学的胡说. 在关于概率论的一本名著（Feller，1950，第 199 页）中，费勒甚至没有给出描述就否定丹尼尔·伯努利对圣彼得堡悖论的解决方式，仅仅向读者说明伯努利“徒劳地试图通过道德期望的概念解决它”. 沃伦·赫希在对该书的评论中将其放大如下：

> 过去人们对这一悖论进行了各种神秘的“解释”，包括道德期望的概念. 这些解释对于现代的概率论学生来说是难以理解的. 费勒给出了一个简单的数学论证，可得到确定有限的入场费，其中圣彼得堡博弈具有公平博弈的所有属性.

我们刚刚已经看到了丹尼尔·伯努利的方法是多么“徒劳”和“难以理解”. 在阅读费勒的书时，我们发现他只是通过定义和分析另一种博弈来“解决”这一悖论的. 他试图以同样的方式解释保险的理论依据；但是由于他拒绝了丹尼尔·伯努利的曲线效用函数的概念，因此得出结论，保险对被保险人来说必然是“不公平”的. 这些解释是现代经济学家难以理解的.

20 世纪三四十年代，奈曼和皮尔逊阐述了另一种决策规则的形式，作为假设检验的附属品. 它在电气工程师（Middleton，1960）和经济学家（Simon，1977）中曾经享有一定的知名度，但由于缺少两个现今被认为必不可少的基本性质，现在已经过时了. 在第 14 章中，我们将给出奈曼 - 皮尔逊方法的一个简单示例，以说明这一方法与其他方法之间的关系. 20 世纪 50 年代，亚伯拉罕·沃尔德（Abraham Wald）提出了一种在更根本的层次上起作用的表述方式. 尽管这一方式为丹尼尔·伯努利的直觉想法提供了相当基本的辩护，看起来更具有恒久的有效性，但是并未得到各方赞赏. 莫里斯·肯德尔（Maurice Kendall，1963）写道：

> 在美国，有一种认为推断是决策论的一个分支的浪潮. 费希尔会（在我看来是正确地）认为科学推断不是决策问题，而且在任何情况下，都不存在基于一种或另一种收益的决策选择标准. 从广义上讲，这是英国人与美国人之间的态度分歧……我以为，在认为行动比思想重要的国家和认为思想比行动重要的国家之间，这种态度差异是不可避免的.

关于费希尔对决策论的态度，我们不必依赖二手资料. 如第 16 章所述，费希尔从不会在任何事情上放弃表达自己观点的机会. 在讨论显著性检验时，他（Fisher，1956 年，第 77 页）写道：

> ……最近……有相当多的学说试图根据完全不同的接受流程来解释或重新解释这些检验. 对于我来说，这两者之间的差别似乎很多，而且我认为，这种重新解释的作者如果对自然科学的工作有真正的了解，或者意识到可以增进科学理解的观察记录的这些特征，就不可能忽略它们.

然后，他将奈曼和沃尔德视为批评的对象.

显然，肯德尔诉诸学者们通常会拒绝的动机，认为决策论是美国人（而不是英国人）性格缺陷（尽管奈曼和沃尔德都不是在美国出生或接受教育的——他们是从欧洲逃到了美国）的体现. 费希尔认为这是不精通自然科学的一种思维偏差所致（尽管这一流程最初起源于丹尼尔 · 伯努利和拉普拉斯，他们作为自然科学家的地位很容易被拿来与费希尔进行比较）.

我们同意肯德尔的观点，沃尔德的做法给人的印象确实是，推断只是决策的一种特殊情况. 和他一样，我们对此深表遗憾. 但是我们观察到，在原始的伯努利-拉普拉斯公式（以及我们的公式）中，这两种功能有着应有的明显区别. 尽管我们理解推断与决策之间的这种必要区别，但我们也意识到，没有决策的推断在很大程度上是无用的. 除非为了某种目的，否则没有真正的自然科学家愿意进行推断的工作.

这些引语反映了丹尼尔 · 伯努利和拉普拉斯的思想要想得到广泛的接受所必须克服的障碍. 这些思想本来是完全自然且极其有用的. 200 年后，任何提出此类建议的人仍然受到顽固“正统”统计阵营不遗余力的攻击，而且攻击的方式没有反映出攻击者的信誉. 现在让我们检查一下沃尔德的理论.

13.7　沃尔德的决策论

沃尔德的表述在其最初阶段与概率论没有明显的联系. 我们首先设想（即列举）一组可能的“自然状态”$\{\theta_1, \theta_2, \cdots, \theta_N\}$，数量 N 在实际情况中总是有限的，尽管将其视为无限大甚至形成连续体可能是一个有用的极限近似. 在第 4 章的质量控制例子中，“自然状态”是批次中未知数量的坏部件数.

这里可能产生一些错误的观念. 让我们澄清其中的一种：在列举不同的自然状态时，我们并未描述自然的任何真实（可验证）性质，因为其中只有一种是真实的. 列举仅是描述有关可能性范围的知识状态的一种手段. 具有不同先验信息的两个人或机器人可能会以不同的方式列举 θ_j，这之间没有谁对谁错，也没有任何矛盾. 人们只能努力利用自己拥有的信息尽力而为，而我们希望拥有更好信息的人自然会做出更好的决定. 这不是悖论，而是老生常谈.

我们理论的下一步是对可能的决策 $\{D_1, D_2, \cdots, D_k\}$ 进行类似的列举. 在质量控制例子中，每个阶段都有三种可能的决定：

$$D_1 \equiv \text{接受该批次};$$

$$D_2 \equiv \text{拒绝该批次};$$

$$D_3 \equiv \text{再次检测}.$$

在第 6 章 B 先生的粒子计数器问题中，我们要估计在第 1 秒通过计数器的粒子

数 n_1，其中有无数种可能的决定：

$$D_i \equiv n_1 \text{ 估计等于 } 0, 1, 2, \cdots$$

如果我们要估计源强度，那么就有太多可能的估计值，以致我们认为它们构成了可能决策的连续体，尽管实际上只能写下有限数量的数字.

除非我们真正“做出决策”，即“决定按照仿佛此决策是正确决策的观念去行动”，否则该理论显然毫无用处. 除非我们准备在 $n_1 = 150$ 的假设下采取行动，否则机器人“决定” $n_1 = 150$ 是最优估计值是徒劳无益的. 因此，给予机器人 D_i 是一种描述我们知道哪些行动可行的知识的一种方法. 考虑任何我们事先知道不可能采取行动相对应的决策是无用的，并且浪费计算资源.

还有一种原因可能会排除特定的决策. 即使 D_1 易于执行，我们也可能事先知道它会导致无法承受的后果. 汽车驾驶员可以随时急转弯，但是常识通常告诉他不要这么做. 这里我们看到另外两点：(1) 存在连续的渐变——行动后果可能很严重，而并非绝对不能容忍；(2) 行动后果通常取决于自然的真实状态——突然的急转弯并不总是会导致灾难，实际上可能会避免灾难.

这令人想到第三个必要概念——损失函数 $L(D_i, \theta_j)$，它是一组数字，代表我们对于 θ_j 是自然状态时做出决策 D_i 导致的“损失”的判断. 如果 D_i 和 θ_j 都是离散的，损失函数则为损失矩阵 (L_{ij}).

仅用 θ_j, D_i, L_{ij} 就能完成很多工作，并且已经有大量文献只使用这几个概念讨论决策标准. 卢斯和雷法的著作（Luce & Raiffa，1989）以一种非常可读和有趣的形式对该理论的早期成果做了总结. 同样，前面提到的切尔诺夫和摩西的基础教科书（Chernoff & Moses，1959）也对此做了总结，这本书今天仍然值得一读. 这最终导致了雷法和施莱弗更高级的著作（Raiffa & Schlaifer，1961），由于其中有大量有用的数学材料，它仍然是标准的参考书.

伯杰的著作（James Berger，1985）中有比我们在此给出的更为详尽的关于哲学和数学的现代论述. 这本书是以与我们几乎相同的贝叶斯视角写的. 它用了很长的篇幅来说明很多对于推断来说很重要的技术详情，但在我们看来，这并不是决策论的真正组成部分.

最小最大准则是指：对于每个 D_i，找出最大可能损失 $M_i = \max_j(L_{ij})$；然后选择 M_i 最小的 D_i. 如果我们将自然视为一名聪明的对手，他会预见到我们的决定并故意选择能使我们有最大挫败感的自然状态，那么这将是一个合理的策略. 在某些博弈理论中，这不是一种完全不现实的情况，因此最小化最大策略在博弈论中具有根本的重要性（von Neumann & Morgenstern，1953）.

在科学家、工程师或经济学家面对的大多数决策问题中，我们并没有聪明的

对手，而最小最大准则是愁眉苦脸的悲观主义者的准则. 他全神贯注于可能发生的最糟糕的事情，从而错失了有利的机会.

在我们看来，满眼繁星闪烁的乐观主义者的态度也是不理智的. 他认为自然界总是刻意设法帮助他，因此拥护最小化最小准则：对于 D_i 找到可能的最小损失 $m_i = \min_j(L_{ij})$ 并选择了使 m_i 最小的 D_i.

显然，对于科学家、工程师或经济学家而言，合理的决策标准在某种意义上介于最小最大与最小最小之间，表示我们相信自然对于我们的目标是保持中立的. 人们已经提出了许多其他准则，例如最大最小效用（Wald）、α-乐观-悲观（Hurwicz）、最小最大后悔（Savage），等等. 如卢斯和雷法所详细描述的那样，通常的流程是分析所提议的准则，以了解它是否满足十几个定性的常识条件，如下所示.

(1) 传递性：如果 D_1 优先于 D_2 且 D_2 优先于 D_3，则 D_1 应优先于 D_3.

(2) 强主导：如果对于所有自然状态 θ_j 有 $L_{ij} < L_{kj}$，则 D_i 始终优先于 D_k.

这种分析虽然简单明了，但也可能变得很乏味. 我们在此不再赘述，因为最终结果是只有一类决策准则能通过所有测试，并且通过不同的推理方式可以更轻松地获得这一类决策准则.

当然，完整的决策论不仅仅与 θ_j, D_i, L_{ij} 有关系. 在典型问题中，我们还拥有其他证据 E 与决策有关. 我们必须学习如何将 E 纳入理论之中. 在第 4 章的质量控制例子中，E 包含先前检测的结果.

在这一点上，沃尔德的决策论需要经过冗长、困难且不必要的迂回途径才能达到目标. 定义“策略” S 为一系列如下形式的规则：“如果我收到新证据 E_i，那么我将做出决策 D_k.” 原则上，首先列举所有可能的策略（然而即使在非常简单的问题中，其数量也是天文数字），然后排除根据以下准则所认为不允许的策略. 定义

$$p(D_k|\theta_j S) = \sum_i p(D_k|E_i\theta_j S)p(E_i|\theta_j) \tag{13.11}$$

为如果 θ_j 是真实自然状态，则策略 S 会导致我们做出决策 D_k 的抽样概率，并且将 θ_j 为真时使用策略 S 的风险定义为该分布上的期望损失：

$$R_j(S) = \langle L\rangle_j = \sum_k p(D_k|\theta_j S)L_{kj}. \tag{13.12}$$

那么，如果不存在其他策略 S' 使得

$$\text{对于所有 } j \text{ 有} \quad R_j(S') \leqslant R_j(S), \tag{13.13}$$

则将策略 S 称为**可容许的**. 如果存在 S'，对于至少一个 θ_j 以上不等式严格成立，则 S 称为**不可容许的**. 风险和可容许性的概念显然是抽样论而不是贝叶斯理论的