然，如果你除了数值 n 和 N 之外还有其他信息，则应该将这些信息考虑在内. 这时你正在考虑一个不同的问题，连续法则不再适用，你可能得到完全不同的答案. 概率论根据所输入的信息给出一致的合情推理结果.

我们必须承认，拉普拉斯提到日出例子本身就做了一个非常不幸的选择——因为出于他所指出的原因，连续法则并不真正适用于日出概率的估计. 此后，这一选择对拉普拉斯的声誉造成了灾难性的影响. 当读者像拉普拉斯一样将“概率”解释为表示部分知识状态的一种手段时，他的陈述才有意义. 但是对于那些认为概率是一种真实物理现象，独立于人类知识而存在的人们来说，拉普拉斯的立场是相当难以理解的. 所以他们会得出结论说，拉普拉斯犯了一个荒唐的错误，甚至不愿阅读他的完整说明.

以下是一些可以在文献中找到的关于连续法则的反对意见的著名例子.

(1) 假设完成了一次氢的凝固. 根据连续法则，如果重复实验，它能再次凝固的概率为 2/3. 这至少不代表任何科学家的信念状态.

(2) 今天有一个男孩 10 岁，根据连续法则，他有 11/12 的概率再活一年. 这个男孩的祖父是 70 岁，根据连续法则，他有 71/72 的概率再活一年. 连续法则显然违反常识!

(3) 考虑 $N = n = 0$ 的情况，那么任何未经验证的猜想为真的概率都是 1/2. 因此，火星上正好有 137 头大象的概率为 1/2. 此外，火星上有 138 头大象的概率也为 1/2. 因此，可以肯定火星上至少有 137 头大象. 但是连续法则还说，火星上没有大象的概率也为 1/2. 因此，该规则在逻辑上是自相矛盾的!

鉴于我们前面的评论，例 (1) 和例 (2) 的问题是显而易见的. 在每一种情况下，我们大家都知道的高度相关的先验信息被简单地忽略了，从而公然滥用了连续法则. 但是让我们更仔细看看例 (3). 连续法则在这里应用地不正确吗? 我们当然不能声称我们拥有关于火星上是否存在大象的先验信息，而这些信息被忽略了. 显然，如果连续法则要适用于例 (3)，那么有一些我们需要强调的关于应用概率论的基本要点.

当我们说一个主张没有“证据”时，这意味着什么? 问题不是我们所说的口头意思是什么. 问题是：这对机器人意味着什么? 就概率论而言，这意味着什么?

我们用于推导连续法则的先验信息是，机器人被告知只有两种可能性：A 为真或假. 它的整个“话语世界”只包含这两个命题. 在 $N = 0$ 的情况下，我们也可以通过直接应用无差别原则来解决问题，这当然会与我们根据连续法则得出的

答案相同：$P(A|X)=1/2$. 只要注意到这一点，我们就会发现问题出在哪里. 只要承认存在三个不同命题之一为真的可能性，而不是两个，我们就已经确定与用于推导连续法则的信息不同的先验信息. ①

如果告诉机器人考虑 137 种不同的 A 可能为假、只有一种可能为真的方式，并且没有其他任何信息，那么它对 A 为真的先验概率为 $1/138$，而不是 $1/2$. 因此，我们看到火星大象的例子还是对连续法则的严重滥用.

教训

像任何其他数学理论一样，除非我们提出确定的问题，概率论不能给出确定的答案. 我们应该始终从明确地列举出问题中要考虑的所有不同命题的“假设空间”开始定义一个问题. 这是必须要指定的“边界条件”的一部分，然后才有一个定义良好的数学问题. 如果我们说“我不知道可能的命题是什么”，这在数学上就相当于说“我不知道要解决什么问题”. 这时，机器人给出的答案只能是：“回去，等你知道时再来问我.”

18.7　杰弗里斯的异议

正如人们所预期的那样，杰弗里斯（Jeffreys，1939，第 107 页）使用的例子更加微妙. 他写道：

> 我可能在英格兰见过千分之一的“有羽毛的动物”. 根据拉普拉斯的理论，“所有有羽毛的动物有喙”这一命题成立的概率约为 $1/1000$. 这与我或任何其他人的信念不符.

尽管我们同意杰弗里斯所说的一切，但是必须指出，他未能补充两个重要的事实. 首先，根据拉普拉斯连续法则，的确有 $P(\text{全部都有喙})\approx 1/1000$. 但是也有 $P(\text{除一个以外都有喙})\approx 1/1000$，$P(\text{除两个以外都有喙})\approx 1/1000$，等等. 更具体地说，如果有 N 个有羽毛动物，我们已经看到 r 个（都有喙），那么用这种记号法重写 (18.24)，我们看到 $P(\text{全部都有喙})=P_0=(r+1)/(N+1)\approx 1/1000$，而 $P(\text{除 } n \text{ 个以外都有喙})$ 为

$$P_n=P_0\frac{(N-r)!(N-n)!}{N!(N-n-r)!}, \tag{18.26}$$

有 n_0 个或更多没有喙的概率为

$$\sum_{n=n_0}^{N}P_n=\frac{(N-r)!(N-n_0+1)!}{(N+1)!(N-n_0-r)!}\approx \mathrm{e}^{-rn_0/N}. \tag{18.27}$$

① 我们在这里看到显而易见之处：我们根据数据得出的结论可能取决于假设空间的大小. 我们在 (15.92) 之后对边缘化悖论的研究中看到了非常相似的事情，其中我们发现参数空间的大小会影响我们的推论. 也就是说，即使我们不知道其数值，引入一个新参数也可能改变我们的结论.

因此，如果有 100 万只有羽毛的动物，其中我们看到 1000 只（全部有喙），那么这等于是一个赌注，即至少有 $1000\ln 2 = 693$ 只没有喙. 当然，甚至可以打赌要少于这个数值. 如果仅有的信息是上述观察结果，我们认为这正是适当与合理的推断.

此外，拉普拉斯连续法则并不适用于该问题，因为我们有其他未考虑的先验信息：物种遗传的稳定性. 有羽毛而无喙的动物如果存在，一定会引起人们极大的好奇心. 即使我们没有看到它，也应该听说过它（就像鸭嘴兽一样）. 为了公平而详细地了解拉普拉斯连续法则 (18.24) 说的是什么，我们需要考虑与一个我们的先验信息与推导过程的假设更相应的问题.

18.8 鲈鱼还是鲤鱼?

一份真实而权威的指南向我们说明，某个湖泊中仅有两种鱼：鲈鱼和鲤鱼. 我们从中抓到 10 条，发现全部都是鲤鱼——那么我们对鲈鱼所占比例的信念状态是怎样的？常识告诉我们，如果鱼群中鲈鱼的比例高于 10%，那么在 10 次捕获中，我们就有相当大的概率找到 1 条，因此我们的信念状态在 10% 以上会迅速下降. 另外，这些数据 D 没有提供证据反对鲈鱼比例为 0 的假设. 因此，不做任何计算的常识将使我们得出结论：鲈鱼数量很可能在例如 $(0\%, 15\%)$ 的区间内，但是直觉并不能定量地告诉我们这种可能性有多大.

那么，拉普拉斯连续法则会得出什么结论呢? 用 f 表示鲈鱼的比例，其后验累积 PDF 为 $P(f < f_0|DX) = 1-(1-f_0)^{11}$. 因此，我们有 $1-(1-0.15)^{11} \approx 0.833$ 的概率或 5 : 1 的几率表明鲈鱼比例确实低于 15%. 同样可以得出该湖中包含小于 9.5% 的鲈鱼的概率为 2/3 或 2 : 1 的几率，而小于 19.6% 的几率为 10 : 1，后验中值是

$$f_{1/2} = 1 - \left(\frac{1}{2}\right)^{1/11} \approx 0.061, \tag{18.28}$$

即 6.1%，甚至有人愿意打赌鲈鱼的比例比这还少. 它的四分位间距为 $(f_{1/4}, f_{3/4}) = (2.6\%, 11.8\%)$，在此区间内外有相同的可能性. 按照最小均方误差准则对 f 进行的“最优”估计是拉普拉斯后验均值 (18.25) : $\langle f\rangle = 1/12$，即 8.3%.

现在假设第 11 次抓到的是鲈鱼. 这将如何改变我们的信念状态呢? 显然，我们将向上调整对 f 的估计，因为现在的数据确实提供了证据反对 f 很小的假设. 的确，如果鲈鱼比例少于 5%，那么我们就不太可能在 11 次中抓到 1 条，因此我们的信念状态在 5% 以下迅速降低，但下降的速度不及之前在 10% 以上.

拉普拉斯连续法则认为，现在最优均方估计为 $\langle f\rangle = 2/13$，即 15.4%，后验

密度为 $P(\mathrm{d}f|DX) = 132f(1-f)^{10}\mathrm{d}f$. 这得到 13.6% 的中值，由于新数据有效地消除了鲈鱼数量低于 3% 的可能性，而这恰恰是以前最可能的区域，因此中值大大提高了. 四分位间距现在为 (8.3%, 20.9%).

在我们看来，所有这些数值都与我们的常识判断非常一致. 那么，这就是拉普拉斯连续法则实际适用的问题. 也就是说，每次试验只有两种可能性，而我们的先验知识除了向我们保证只有这两种可能性外不提供任何其他信息. 每当根据拉普拉斯连续法则得出的结论与我们的直觉不一致时，我们就认为原因是我们的常识正在利用关于现实世界的其他先验信息，而该信息并未用于推导连续法则.

18.9 连续法则什么时候有用?

从数学上讲，连续法则是对由先验概率和数据定义的某一类推断问题的解. 200 年悬而未决的问题是：均匀先验概率 (18.3) 描述的是什么先验信息? 拉普拉斯对此也不是太清楚——他的讨论似乎援引了“概率的概率”的概念，在人们有内外层机器人的概念之前，这似乎都是形而上学的胡说——他的批评者们，不是建设性地试图更清楚地定义概念问题，而是抓住这一点来否定拉普拉斯关于概率论的整个研究方法.

在拉普拉斯的批评者中，只有杰弗里斯（Jeffreys，1939）和费希尔（Fisher，1956）似乎更为深思熟虑，意识到先验信息未清晰定义是问题的根源. 其他人只是——跟随维恩（Venn，1866）的例子——举出常识与拉普拉斯规则相矛盾的例子，并且在不试图理解其原因的情况下，在任何情况下都拒绝接受该规则. 正如我们在第 16 章中提到的那样，维恩的批评是如此不公平，以至于费希尔（Fisher，1956）都被迫在此问题上为拉普拉斯辩护.

在这方面，我们必须记住，概率论从来不能解决实践中的问题，因为所有这些问题都是无限复杂的. 我们只能解决经过理想化的实际问题，并且在理想化很好的一定范围内，解会很有用. 在氢凝固的示例中，我们的常识使用的先验信息看似非常简单，实际上却非常复杂，以至于没人知道如何将其转化为先验概率. 毫无疑问，概率论原则上有能力解决此类问题，但是我们还没有学会在不过分简化的前提下如何将它们翻译成数学语言.

总而言之，拉普拉斯连续法则为确定的实际问题提供了确定有用的解. 由于它不是某些不同问题的解，每个人都谴责它不过是胡说八道. 可以合理理想化的问题是只需考虑两个假设，有恒定的“因果机制”，并且没有其他先验信息的情况. 这是连续法则唯一适用的情况，但是我们当然可以将其推广为针对任意数量的假设的情况，如下所示.

18.10 推广

我们将详细给出整个推导过程，以介绍在许多其他问题中有用的拉普拉斯的数学技巧. 有 K 个不同的假设 $\{A_1, A_2, \cdots, A_K\}$，有“因果机制”恒定的信念，没有其他先验信息. 我们进行随机试验 N 次，并观察到 A_1 为真的情况 n_1 次，A_2 为真的情况 n_2 次，等等. 当然，我们有 $\sum_i n_i = N$. 根据这一证据，在后面的 $M = \sum_i m_i$ 次重复试验中，A_i 会出现 m_i 次为真的概率是多少？要得到所求的答案，即概率 $P(m_1 \cdots m_K | n_1, \cdots, n_K)$，我们通过 K 维均匀先验 A_p 密度

$$(A_{p_1} \cdots A_{p_K} | X) = C\delta(p_1 + \cdots + p_K - 1), \quad p_i \geqslant 0 \tag{18.29}$$

定义先验知识. 为了找到归一化常数 C，令

$$\int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K \, (A_{p_1} \cdots A_{p_K} | X) = 1 = CI(1), \tag{18.30}$$

其中

$$I(r) \equiv \int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K \delta(p_1 + \cdots + p_K - r). \tag{18.31}$$

直接计算很麻烦，因为第一项之后的所有积分都将在需要求出的极限之间. 因此，让我们使用以下技巧. 首先，对 (18.31) 进行拉普拉斯变换：

$$\int_0^{+\infty} \mathrm{d}r \mathrm{e}^{-\alpha r} I(r) = \int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K \exp\{-\alpha(p_1 + \cdots + p_K)\} = \frac{1}{\alpha^K}. \tag{18.32}$$

然后，根据柯西定理反转拉普拉斯变换

$$I(r) = \frac{1}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \mathrm{d}\alpha \frac{\mathrm{e}^{\alpha r}}{\alpha^K} = \frac{1}{(K-1)!} \frac{\mathrm{d}^{K-1}}{\mathrm{d}\alpha^{K-1}} \mathrm{e}^{\alpha r} \bigg|_{\alpha=0} = \frac{r^{K-1}}{(K-1)!}, \tag{18.33}$$

其中，根据拉普拉斯变换的标准理论，此处的积分路径通过原点的右侧，并且在左侧半平面上被一个无限半圆封闭，该半圆的积分为零. 因此有

$$C = \frac{1}{I(1)} = (K-1)!. \tag{18.34}$$

通过这一技巧，我们避免了不得不考虑不同 p_i 的不同积分范围的复杂细节. 如果尝试直接计算 (18.31)，就会出现这些细节. 使用相同的技巧，先验 $P(n_1 \cdots n_K | X)$ 是

$$\begin{aligned} P(n_1 \cdots n_K | X) &= \frac{N!}{n_1! \cdots n_K!} \int_0^{+\infty} \mathrm{d}p_1 \cdots \int_0^{+\infty} \mathrm{d}p_K \, p_1^{n_1} \cdots p_K^{n_K} (A_{p_1} \cdots A_{p_K} | X) \\ &= \frac{N!(K-1)!}{n_1! \cdots n_K!} J(1), \end{aligned} \tag{18.35}$$

其中

$$J(r) \equiv \int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K \, p_1^{n_1} \cdots p_K^{n_K} \delta(p_1 + \cdots + p_K - r), \tag{18.36}$$

这将通过做拉普拉斯变换

$$\int_0^{+\infty} \mathrm{d}r\, \mathrm{e}^{-\alpha r} J(r) = \int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K\, p_1^{n_1} \cdots p_K^{n_K} \exp\{-\alpha(p_1 + \cdots + p_K)\}$$
$$= \prod_{i=1}^{K} \frac{n_i!}{\alpha^{n_i+1}} \tag{18.37}$$

进行计算. 因此，如 (18.33) 所示，我们有

$$J(r) = \frac{n_1! \cdots n_K!}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \mathrm{d}\alpha\, \frac{\mathrm{e}^{\alpha r}}{\alpha^{N+K}} = \frac{n_1! \cdots n_K!}{(N+K-1)!}\, r^{N+K-1}, \tag{18.38}$$

以及

$$p(n_1 \cdots n_K|X) = \frac{N!(K-1)!}{(N+K-1)!}, \quad n_i \geqslant 0, \quad n_1 + \cdots + n_K = N. \tag{18.39}$$

因此，根据贝叶斯定理有

$$(A_{p_1} \cdots A_{p_K}|n_1 \cdots n_K) = (A_{p_1} \cdots A_{p_K}|X) \frac{P(n_1 \cdots n_K|A_{p_1} \cdots A_{p_K})}{P(n_1 \cdots n_K|X)}$$
$$= \frac{(N+K-1)!}{n_1! \cdots n_K!}\, p_1^{n_1} \cdots p_K^{n_K} \delta(p_1 + \cdots + p_K - 1), \tag{18.40}$$

最后得到

$$P(m_1 \cdots m_K|n_1 \cdots n_K)$$
$$= \int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K\, P(m_1 \cdots m_K|A_{p_1} \cdots A_{p_K})(A_{p_1} \cdots A_{p_K}|n_1 \cdots n_K)$$
$$= \frac{M!}{m_1! \cdots m_K!} \frac{(N+K-1)!}{n_1! \cdots n_K!} \int_0^{+\infty} \mathrm{d}p_1 \cdots \mathrm{d}p_K\, p_1^{n_1+m_1} \cdots p_K^{n_K+m_K}$$
$$\times \delta(p_1 + \cdots + p_K - 1). \tag{18.41}$$

只要做替换 $n_i \to n_i + m_i$，这个积分就与 $J(1)$ 相同. 所以，根据 (18.38)，

$$P(m_1 \cdots m_K|n_1 \cdots n_K) = \frac{M!}{m_1! \cdots m_K!} \frac{(N+K-1)!}{n_1! \cdots n_K!} \frac{(n_1+m_1)! \cdots (n_K+m_K)!}{(N+M+K-1)!}, \tag{18.42}$$

或者重新组织为二项式系数，则 (18.24) 的推广为

$$P(m_1 \cdots m_K|n_1 \cdots n_K) = \frac{\binom{n_1+m_1}{n_1} \cdots \binom{n_K+m_K}{n_K}}{\binom{N+M+K-1}{M}}. \tag{18.43}$$

在只需要得到下一次试验中 A_1 为真的概率时，我们需要 $M = m_1 = 1$ 且所有其他 $m_i = 0$ 的公式. 结果是广义连续法则

$$P(A_1|n_1 NK) = \frac{n_1 + 1}{N + K}. \tag{18.44}$$

我们看到，在 $N = n_1 = 0$ 的情况下，这可以简化为根据无差别原则所得到的答案，因此它就是其中的一种特殊情况. 如果 K 是 2 的幂，这与卡尔纳普（Carnap，1942）提出的归纳推理方法相同——在其归纳方法连续统中表示为 $c^*(h, e)$.

当然，在 N 很小的情况下使用连续法则是很愚蠢的. 未必是错的，只是愚蠢. 因为，我们没有 A 的先验证据，并且进行的观察数量如此之少，实际上并没有什么证据. 这并不是进行合情推理的有希望的起点. 我们不能指望从中得到任何有用的信息. 当然，我们确实得到了确定的概率值，但是这些值非常“软”，即非常不稳定. 因为对于小的 N，A_p 分布仍然非常宽. 常识告诉我们，对于小的 N，证据 N_n 没有为进一步的预测提供可靠的依据. 我们将看到，这一结论也是我们在此发展的理论的自然结果.

引入连续法则的真正原因在于我们确实从实验中获得了大量信息，即 N 很大的情况. 幸运的是，在这种情况下，我们几乎可以忽略与先验证据有关的具体细节. 出于与第 6 章的粒子计数器问题相同的原因，特定的初始赋值 $(A_p|X)$ 不会对结果产生太大影响. 对于导致 (18.43) 的推广情况，这仍然适用. 从 (18.44) 中可以看到，一旦观察的次数 N 与假设的数量 K 相比很大，那么分配给任何特定假设的概率从实用的角度而言仅取决于我们所观察到的数据，而不是有多少个先验假设. 如果对此思考 10 秒钟，常识将会告诉你 $N \gg K$ 是正确的条件.

从维恩（Venn，1866）开始，那些发表文章对拉普拉斯连续法则表示强烈反对的人简直令人难以置信. 一个人如何可能既拒绝拉普拉斯连续法则，同时又提倡概率的频率定义呢? 在多次试验中，给事件指定概率等于频率的任何人，都在按照拉普拉斯的规则行事！当观察次数与命题的数量相差不大时，推广的规则 (18.44) 提供了显然需要的改进，即一个小的修正项.

18.11 证实和证据的权重

我们关于 A_p 的计算提供了一些新思想——或者更确切地说，与熟悉的旧思想的一些联系. 尽管我们不会对其做任何特别应用，但是值得指出来. 我们看到，面对新证据时概率分配的稳定性基本上由 A_p 分布的宽度决定. 如果 E 是先验证据，F 是新证据，则

$$P(A|EF) = \int_0^1 \mathrm{d}p\, p\,(A_p|EF) = \frac{\int_0^1 \mathrm{d}p\, p\,(A_p|F)(A_p|E)}{\int_0^1 \mathrm{d}p\,(A_p|F)(A_p|E)}. \tag{18.45}$$

如果新证据 F 不会使 A 的概率发生任何显著的变化，我们可以说 F 就 A 而言与 E **兼容**：

$$P(A|EF) = P(A|E). \tag{18.46}$$

新证据可能在不改变一阶矩的情况下极大地改变 A_p 的分布. 它可能会大大地让它变尖锐或变宽. 我们可能对 A 变得更加确定或更不确定，但是如果 F 不改变 A_p 分布的重心，我们最终仍然会将相同的概率分配给 A.

现在，它具有更强的性质：如果新证据 F 与之兼容，并且使我们对其更有信心，则 F 可以证实先前的概率分配. 换句话说，我们排除一些可能性，并且有了新的证据 F，A_p 分布变窄了. 假设 F 是做一些随机试验并观察 A 成立的频率. 在这种情况下，$F = N_n$，我们先前的结果 (18.22) 给出

$$(A_p|N_n) = \frac{(N+1)!}{n!(N-n)!}p^n(1-p)^{N-1} \approx (\text{常数}) \cdot \exp\left\{-\frac{(p-f)^2}{2\sigma^2}\right\}, \tag{18.47}$$

其中

$$\sigma^2 = \frac{f(1-f)}{n}, \tag{18.48}$$

并且 $f = n/N$ 是观测到的 A 的频率. 近似是通过将 $\ln(A_p|N_p)$ 对其峰值做泰勒级数展开得到的，并且在 $n \gg 1$ 且 $N - n \gg 1$ 时有效. 如果满足这些条件，则 $(A_p|N_n)$ 关于其峰值非常接近对称. 那么，如果观测到的频率 f 接近先验概率 $P(A|E)$，则新证据 N_n 不会影响 A_p 分布的一阶矩，但会使其变尖，这按定义将构成一次证实.

这显示了概率和频率之间的另一种联系. 我们以与众不同的思想来定义概率分配的“证实”. 我们以与证实之前的认知状态直觉相一致的方式定义它. 但是相同的实验证据将构成对频率理论或我们理论的证实.

现在，我们从中可以看到另一个有用的概念，称为证据权重. 考虑 A_p，给定两种不同的证据 E 和 F，

$$(A_p|EF) = (\text{常数}) \times (A_p|E)(A_p|F). \tag{18.49}$$

如果分布 $(A_p|F)$ 比分布 $(A_p|E)$ 尖锐得多，那么两者的乘积实际上仍会在 F 确定的峰值处有一个峰值. 在这种情况下，我们凭直觉说证据 F 比证据 E 具有更大的“权重”. 如果我们有证据 F，那么是否考虑证据 E 并没有多大关系. 另外，如果我们没有证据 F，那么 E 所代表的任何证据都将是非常重要的，因为它将代表我们能做的最好推断. 因此，就所有实际目的而言，获取到一个权重很大的证据可以使得我们不必继续跟踪权重小的其他证据.

当然，这正是我们人类思维的方式. 当得到非常重要的证据时，我们将不再对其他不明确的证据给予太多关注. 这样一来，我们并不会自相矛盾，因为这无论如何也不会带来太大的变化. 因此，我们直观的证据权重概念与 A_p 分布的尖锐程度紧密相关. 我们认为非常重要的关于 A 的证据未必是使得 A 的概率发生很大变化的证据，它是使 A_p 的密度分布发生了很大变化的证据. 明白这一点，我们就可以对无差别原则有更多的了解，并且可以使得我们的理论与卡尔纳普的归纳推理方法联系起来.

无差别是基于知识还是无知?

在使用无差别原则分配概率值时，必须满足两个不同的条件：(1) 必须能够将情况分解为互斥穷尽的不同可能性；(2) 完成此操作后，必须找到可用的信息使得我们没有理由偏爱任何一种可能性. 实际上，除非问题中存在明显的对称性，否则这些条件很难满足. 但是，可能有两种完全不同的方式来满足条件 (2)：无知，或者对情况了解. 为了说明这一点，让我们假设一个已知非常不诚实的人将抛一枚硬币，并且有两个人 A 和 B 在看着他. A 可以检查硬币. 他拥有国家标准局的所有设备. 他用秤和卡尺、磁力计和显微镜、X 射线和中子束等对硬币进行了数百次实验. 最后，他坚信硬币是完全无偏的. B 则不能这样做. 他所知道的只是一个小人正在抛硬币. 他怀疑硬币有偏，但不知道偏向哪一面. 条件 (2) 对于他们两人都满足. 每个人都会开始为每面分配 $1/2$ 的概率. 相同的概率分配可以描述完全无知或充分了解的条件. 有很长一段时间，这似乎被认为是自相矛盾的. 为什么 A 的额外知识没有什么作用? 当然，它确实在起作用，并且是在起非常重要的作用. 但是直到我们开始实验时，这种作用才会显示出来. 差别不在 A 的概率，而是 A_p 的密度.

假设第一次抛掷的结果是正面. 对于 B 来说，这构成该硬币有偏且偏向正面的证据. 因此，在接下来的抛掷中，他将考虑以上证据分配新的概率. 但是对 A 来说，硬币无偏的证据远比一次抛掷结果的证据强烈，他将继续分配 $1/2$ 的概率.

你会看到将要发生的事情. 对于 B 来说，每一次抛硬币都代表了硬币有偏的新证据. 每次抛掷后，他都会修改下一次抛掷的概率分配. 但是经过多次抛掷后，他的概率分配将变得越来越稳定，并且在 $n \to +\infty$ 时，趋于观察到的正面朝上的频率. 对于 A 来说，对称性的先验证据比几乎任何次抛掷结果的证据具有更大的权重，他会坚持分配概率 $1/2$. 他们每个人都根据自己掌握的信息进行了一致的合情推理，我们的理论解释了每个人的行为.

如果你假设 A 具有完全的对称性知识，则可能得出结论，他的 A_p 分布是德尔塔函数. 在这种情况下，他的想法永远不会因任何新数据而改变. 当然，这是在实践中从来不会存在的极限例子. 甚至连标准局也不能给我们提供如此好的证据.

18.12 卡尔纳普的归纳法

哲学家鲁道夫 · 卡尔纳普（Carnap，1952）提出了一系列可能的“归纳法”，通过这些方法，人们可以将先验信息和频率数据转化为概率分配和对未来频率的

估计. 他的特定原则（即根据直觉而不是概率论规则找到的原则）是：最终的概率分配 $P(A|N_nX)$ 应该是先验概率 $P(A|N)$ 和观察到的频率 $f=n/N$ 的加权平均值. 将权重 N 分配给“经验因子” f，将任意权重 λ 分配给“逻辑因子” $P(A|X)$，将得到卡尔纳普用 $c_\lambda(h,e)$ 表示的方法. 引入 A_p 分布对此有更详细的解释. 这里的理论包括卡尔纳普的所有方法，它们是对应于不同先验密度 $(A_p|X)$ 的特例，并使我们将 λ 重新解释为先验证据的权重. 因此，在两种假设的情况下，卡尔纳普方法是你可以根据先验密度 $(A_p|X)=(\text{常数})\cdot[p(1-p)]^r$（其中 $2r=\lambda-2$）来计算的方法. 结果是

$$P(A|N_nX)=\frac{2n+\lambda}{2N+2\lambda}=\frac{(n+r)+1}{(N+2r)+2}. \tag{18.50}$$

更大的 λ 对应着更尖锐的峰值 $(A_p|X)$ 密度.

在抛硬币的例子中，来自标准局的 A 根据卡尔纳普方法推断 λ 约为数千. 而 B 由于对硬币的先验知识较少，可能会使用 5 或 6 的 λ.（$\lambda=2$ 的情况给出了拉普拉斯连续法则，它太宽以至于对于抛硬币不现实. B 肯定知道硬币的重心不可能偏离几何中心超过其厚度的一半. 实际上，正如我们在第 10 章中看到的那样，这种分析出于物理定理的原因并不总是适用于抛掷真实硬币的情况.）

根据第二种方式写出 (18.50)，我们看到卡尔纳普 λ 方法对应于先验证据的权重将被给予 $\lambda-2$ 次试验，其中恰好有一半的 A 被观察为真. 我们能否理解为什么先验证据的权重为 $\lambda=$（先验试验的次数 $+2$），而新证据的权重 N_p 只是（新试验的次数）$=N$ 呢？可以这样来看.（$+2$）的出现只是机器人告诉我们以下信息的方式：A 可能为真或假的先验知识等同于 A 至少一次为真、一次为假. 这几乎不是什么推导，但有一定的道理.

让我们继续探讨这一推理. 我们从陈述 X 开始：在任何一次试验中 A 可能为真也可能为假. 但这仍然是一个含糊的陈述. 假设我们将它解释为这意味着 A 已被观察到恰好一次为真、一次为假. 如果我们认为通过拉普拉斯方法分配 $(A_p|X)=1$ 正确描述了这种知识状态，那么*在获得数据 X 之前*，“*前先验*”*知识状态 X_0 是什么*？要回答这个问题，我们只需要反向应用贝叶斯定理，就像我们在第 5 章中使用虚构结果的方法和第 6 章从坛子中抽取球的问题中所做的那样. 结果是，我们的“前先验” A_p 分布必须是

$$(A_p|X_0)\mathrm{d}p=(\text{常数})\frac{\mathrm{d}p}{p(1-p)}. \tag{18.51}$$

这只是表示对参数空间的“完全无知”或“基本度量”的类分布. 这是我们在第 12 章的变换群方法中发现的，也是霍尔丹（Haldane, 1932）很久以前建议的. 因此，这是可能导致我们采取这一措施的另一种思路. 基于同样的思路，我们已经在第 6 章

给出的 (6.29) 是 (18.51) 的离散形式.

那么看起来，如果我们有明确的先验证据表明在任何一次试验中 A 都可能为真或假，那么使用拉普拉斯规则 $(A_p|X)=1$ 是合适的. 但是，如果开始我们非常不肯定，甚至不能确定在某些试验中 A 是否可能为真或假，那么我们应该使用先验 (18.51).

前先验分配 (18.51) 导致我们的数值结果有何不同? 根据此前先验分配重复对 (18.22) 的推导，我们发现，如果 n 不为 0 或 N，

$$(A_p|N_nX_0)=\frac{(N-1)!}{(n-1)!(N-n-1)!}p^{n-1}(1-p)^{N-n-1}, \tag{18.52}$$

得到的不是拉普拉斯连续法则，而是 p 的均值估计

$$P(A|N_nX_0)=\int_0^1 \mathrm{d}p\, p\,(A_p|N_n)=\frac{n}{N}, \tag{18.53}$$

它等于观测到的频率，并且与 p 的最大似然估计相同. 同样，假设 $0<n<N$，我们得到的不是 (18.24) 而是

$$P(M_m|N_nX_0)=\frac{\binom{m+n-1}{m}\binom{M-m+N-n-1}{M-m}}{\binom{N+M-1}{M}}. \tag{18.54}$$

所有这些结果都对应于观察到少一次的真和少一次的假.

18.13 可交换序列中的概率与频率

现在，我们可以对概率和频率之间的联系做更多的介绍. 主要有两种类型的联系：(a) 给定随机试验中的观察频率，将该信息转化为概率分配；(b) 给定概率分配，预测某种条件下的频率. 在第 11 章和第 12 章中，我们已经看到最大熵和变换群原理是如何实现概率分配的. 如果我们感兴趣的量是某个随机试验的结果，该概率分配将自动对应于预测频率，从而在某些情况下解决问题 (b).

连续法则在非常广泛的一类问题中为我们提供了问题 (a) 的解决方案. 如果我们在大量试验中观察到 A 是否为真，而且我们对 A 的唯一了解是该随机试验的结果以及背后“因果机制”的一致性，那么连续法则表明我们在下一次试验中分配给 A 的概率实际上应该等于观察到的频率. 实际上，这正是根据频率来定义概率的人们所做的：假设存在一个未知的“绝对”概率，该概率值可以通过做随机试验来找到. 当然，你必须进行大量的试验. 然后，将观察到的 A 的频率作为概率的估计. 如本章前面所述，当“频率主义者”通过取置信区间的中心来改善自己的方法时，拉普拉斯公式中的 $+1$ 和 $+2$ 也会出现. 因此，我不明白最狂热的概率频率理论的倡导者如何能在不谴责自己的流程的情况下谴责连续法则. 在

所有的争论不休背后，仍然存在一个简单事实，那就是他在自己的流程中，也在按照拉普拉斯连续法则行事. 实际上，用频率来定义概率等同于说连续法则是唯一可用于将观测数据转化为概率分配的规则.

18.14 频率预测

现在让我们考虑在以下情况下的问题 (b)：根据概率推理频率. 这只是一个参数估计问题，原则上跟其他参数估计问题没有什么不同. 假设我们不是问在下一次试验中得到 A 为真的概率，而是希望根据证据 N_n 来推断无限多次试验中 A 的相对频率. 我们必须取当 $M \to +\infty, m \to +\infty$ 时 (18.24) 的极限，使得 $m/M \to f$. 引入命题

$$A_f = \text{无限次的试验中 } A \text{ 为真的频率 } f, \tag{18.55}$$

我们发现在给定 N_n 的情况下，A_f 的概率密度为

$$P(A_f|N_n) = \frac{(N+1)!}{N!(N-n)!} f^n (1-f)^{N-n}, \tag{18.56}$$

这与 (18.22) 中的 $(A_p|N_n)$ 相同，其中 f 在数值上等于 p. 根据 (18.55)，最概然频率等于 n/N，即过去观察到的频率. 但是我们之前已经注意到在参数估计中（如果你反对我称 f 是“参数”，那么可以称其为“预测”），最概然值估计在小样本情况下通常比均值差，而且两者可能会明显不同. 频率的均值估计为

$$\overline{f} = \int_0^1 \mathrm{d}f\, f\, P(A_f|N_n) = \frac{n+1}{N+2}, \tag{18.57}$$

刚好与拉普拉斯连续法则所给出的 $P(A|N_n)$ 值 (18.25) 相同. 因此，我们可以用任何一种方式来解释. 拉普拉斯理论在一次试验中分配给 A 的概率在数值上等于最小化均方误差的对频率的估计. 你能看到这与最大熵和变换群论证中发现的概率和频率之间的关系相当吻合.

还请注意，对于小的 N 而言，分布 $P(A_f|N_n)$ 相当宽，这证实了我们在这种情况下不可能进行可靠预测的预期. 一个数值例子是，如果在 2 次试验中 1 次观测到 A 为真，则 $\overline{f} = P(A|N_n) = 1/2$，但是根据 (18.55)，真实频率 f 仍然有一半概率在范围 $0.326 < f < 0.674$ 之外. 在没有证据的情况下（$N = n = 0$），有一半概率 f 在范围 $0.25 < f < 0.75$ 之外. 更一般地，(18.55) 的方差为

$$\operatorname{var} P(A_f|N_n) = \overline{f^2} - \overline{f}^2 = \frac{\overline{f}(1-\overline{f})}{N+3}, \tag{18.58}$$

因此，估计值 (18.56) 的期望误差以 $1/\sqrt{N}$ 减小. 就实用而言，我们可以从 (18.56) 得出的关于预测可靠性的更详细的结论，都与统计学家通过置信区间方法得出的结论相同.

所有这些结果也适用于推广的连续法则. 当 $M \to +\infty, m_i/M \to f_i$ 时取 (18.43) 的极限，我们得到频率为 f_i 的 A_i 的联合概率密度函数为

$$P(f_1 \cdots f_k | n_1 \cdots n_k) = \frac{(N+K)!}{n_1! \cdots n_k!}(f_1^{n_1} \cdots f_k^{n_k})\delta(f_1 + \cdots + f_k - 1). \tag{18.59}$$

通过对满足 $f_1 \geqslant 0$, $f_2 + \cdots + f_k = 1 - f_1$ 的所有 $f_2, \cdots, f_k$ 取 (18.59) 的积分，可以得到频率 f_1 在 $\mathrm{d}f_1$ 范围内的概率. 这可以通过以众所周知的方式应用拉普拉斯变换，结果是

$$P(f_1 | n_1 \cdots n_k) = \frac{(N+K-1)!}{n_1!(N-n_1+K-2)!} f_1^{n_1}(1-f_1)^{N-n_1+K-2}, \tag{18.60}$$

我们从中发现最概然值是

$$(\hat{f}_1) = \frac{n_1}{N+K-2}, \tag{18.61}$$

均值是

$$\overline{f}_1 = \frac{n_1+1}{N+K}, \tag{18.62}$$

这正是拉普拉斯连续法则 (18.44).

通过在 (18.18) 中令 $M \to +\infty, m/M \to f$ 取 $P(M_m|A_p)$ 的极限，可以得到另外一个有趣的结果

$$P(M_m | A_p) = \delta(f-p). \tag{18.63}$$

同样，在 (18.22) 中令 $N \to +\infty$ 取 $(A_p|N_n)$ 的极限，我们得到

$$(A_p | A_f) = \delta(f-p). \tag{18.64}$$

这也可以从 (18.63) 通过应用贝叶斯定理得到. 因此，如果 B 为任意命题，根据我们的标准论证，

$$\begin{aligned} P(B|A_f) &= \int_0^1 \mathrm{d}p\,(BA_p|A_f) = \int_0^1 \mathrm{d}p\,P(B|A_pA_f)(A_p|A_f) \\ &= \int_0^1 \mathrm{d}p\,P(B|A_p)\delta(p-f). \end{aligned} \tag{18.65}$$

在最后一步中，我们应用了性质 (18.1)：A_p 自动中和关于 A 的任何其他陈述. 因此，如果 f 与 p 相等，则 $P(B|A_f) = P(B|A_p)$，A_p 和 A_f 在合情推理中是等价的陈述.

为了在某一种情况下验证这种等效性，请注意，在极限 $N \to +\infty, n/N \to f$ 时，(18.24) 中的 $P(M_m|N_n)$ 简化为 (18.18) 给出的二项分布 $P(M_m|A_p)$. 推广的 (18.43) 在相应的极限中变为多项分布，

$$P(m_1 \cdots m_k | f_1 \cdots f_k) = \frac{m!}{m_1! \cdots m_k!} f_1^{m_1} \cdots f_k^{m_k}. \tag{18.66}$$

这种等效性说明了概率与频率的概念为什么如此容易混淆，以及为什么在许多问题中这种混淆不会造成损害. 只要可以获得的信息包括大样本中观察到的频率以及具有恒定的“因果机制”，拉普拉斯理论就在数学上等同于频率理论. 大多数“经典”的统计问题（如人寿保险等）属于这种类型. 如果人们只解决这样的问题，就一切都很好. 但当我们考虑更普遍的问题时，这就会造成损害.

如今，物理和工程有许多重要的概率论应用，其中证据的绝对重要部分不能用频率来描述，或者我们需要进行合情推理的量与频率无关. 如果坚持使用（概率）≡（频率）作为公理，将使我们无法在这些问题中应用概率论.

18.15　一维中子倍增

到目前为止，我们的讨论很抽象，也许过于抽象了. 为了纠正这一现象，我想展示这些方程适用的特定物理问题. 贝尔曼、卡拉巴和温（Bellman, Kalaba & Wing，1957）首先描述了这个问题，然后，它在温的一本更新的著作（Wing，1962）中得到了进一步发展. 中子在可裂变材料中移动，我们希望估计由于一个入射中子而触发的新中子数量. 为了得到一个可解的数学问题，我们做如下简化假设.

(a) 中子仅以恒定速度在 $\pm x$ 方向上移动.

(b) 每次向右或向左移动的中子引发裂变反应，其结果正好是两个中子，一个向右移动，一个向左移动. 因此，最终结果是任何中子都将不时触发沿相反方向移动的子代中子.

(c) 子代中子同样可以立即触发更多的子代.

我们从左侧将单个触发中子发射到厚度为 x 的可裂变材料中，问题是要预测在整个过程中从左侧和从右侧出现的中子数量. 这至少就是我们想计算的. 当然，产生中子的数量不能*完全确定*，因此我们能做的最好的事情就是计算 n 个中子透射或反射的*概率*. 我们想通过应用于该问题对拉普拉斯理论和频率概率理论进行详细的比较. 我们主要关注将概率论与物理模型联系起来的基本原理.

许多频率理论的支持者基于纯粹哲学的理由谴责拉普拉斯理论，无论它在应用中成功与否. 有些人持更合乎情理的立场，他们认识到在目前的状况中没有自以为是的理由，而有更多需要保持谨慎的理由. 尽管他们认为目前频率理论是更优的，但是他们也说，就像一位通信者对我所说的那样：“如果有任何能使我得到更好理解及更有效形式的理论，我就会很乐意放弃频率理论.”问题在于，目前的统计文献使我们没有机会看到实际使用拉普拉斯理论的效果，因此无法进行有效

的比较. 这就是我们要在这里纠正的情况.

18.15.1 频率解

首先，让我们用频率理论来表述这个问题. “频率主义者”的推理方式如下.

实验者为我们测量了这种材料在很小厚度下的相对裂变频率 $p = a\Delta$. 这意味着他们向厚度为 Δ 的板中发射 N 个触发中子，并且观察到 n 例裂变. 由于 N 有限，因此我们无法从中得到 p 的确切值，但它近似等于观测频率 n/N. 更准确地说，我们可以得到 p 的置信区间. 在类似的情况下，我们可以预期大约有 $k\%$ 的时间，极限区间

$$\frac{N}{N+\lambda^2}\left[\frac{2n+\lambda^2}{2N} \pm \lambda\sqrt{\frac{n(N-n)}{N^3}+\frac{\lambda^2}{4N^2}}\right] \tag{18.67}$$

将包含 p 的真实值，其中 λ 是正常偏差的 $(100-k)\%$ 值（Cramér，1946，第 515 页）. 例如，当 $\lambda=\sqrt{2}$ 时，区间

$$\frac{n+1}{N+2} \pm \frac{N}{N+2}\sqrt{\frac{2n(N-n)}{N^3}+\frac{1}{N^2}} = \frac{n+1}{N+2} \pm \sqrt{\frac{2n(N-n)}{N^3}} \tag{18.68}$$

在类似情况下的大约 84% 将包含正确的 p.（同样，也存在拉普拉斯的连续法则中的 $+1$ 和 $+2$!）一般来说，λ 与 k 之间的关系为

$$\frac{1}{\sqrt{2\pi}}\int_{-\lambda}^{\lambda} \mathrm{d}x\, \exp\left\{-\frac{x^2}{2}\right\} = \frac{k}{100}. \tag{18.69}$$

当 n 和 $N-n$ 足够大时，(18.67) 近似有效. 确切的置信区间很难通过解析式表示，对于较小的 N，应参考皮尔逊和克洛珀（Pearson & Clopper，1934）的图. 当然，数 p 是可裂变材料的确定但不完全已知的物理常数特征.

现在，为了计算从厚度为 x 的材料中反射 n 个中子的相对频率，我们必须做一些附加的假设. 我们假设，每个中子在单位长度裂变概率总是相同的，与它的历史无关. 由于背后运作原因的复杂性，这种假设似乎是合理的. 但是，只有通过将我们的最终计算结果与实验结果进行比较，才能真正验证该假设是否成立. 该假设意味着，连续的厚度为 Δ 的板的裂变概率是相互独立的. 例如，入射中子在厚度为 Δ 的第二个板中发生裂变而不在第一个板中发生裂变的概率为 $p(1-p)$.

现在我们转向其数学，并通过几种可能的技术中的任何一种来解决问题，这些技术出现在以相对频率 $p_n(x)$ 和 $q_n(x)$ 反射或透射 n 个中子的问题中.［实际上，尚未找到解析解，但是温（Wing，1962）给出了数值积分的结果，这对于我们的目标而言已经足够好了.］

我们现在将这些预测与实验结果进行比较. 当第一个触发中子被发射到厚度为 x 的板中时，我们观察到反射的 r_1 中子和透射的 t_1 个中子. 这些数据不会影

响到 $p_n(x)$ 和 $q_n(x)$ 的分配，因为后者对于单个实验没有任何意义，而只是无限多次数实验的极限频率的预测. 因此，我们必须重复实验多次，并记录每次实验的结果数值 r_i 和 t_i. 如果发现 $r_i = n$ 的情况发生的频率足够接近 $p_n(x)$ [“足够接近”是由某一种显著性检验（例如卡方检验）确定的]，那么我们可以得出结论该理论是令人满意的，或者至少没有被数据拒绝. 但是，如果观察到的频率与 $p_n(x)$ 的偏差很大，那么就能知道我们的初始假设有问题.

当然，理论可能对也可能错. 如果理论错了，那么原则上整个理论都将被推翻，我们必须重新开始尝试找到正确的理论. 在实践中，可能只需要更改该理论的一个小方面即可，因此大多数旧的计算将仍然对新理论有用.

18.15.2 拉普拉斯解

现在我们用拉普拉斯理论来陈述同样的问题. 我们只是将其视为合情推理的一个例子，在其中对单次或有限数量的实验的结果进行最优猜测. 我们不关心极限频率的预测及存在与否，因为任何关于不可能实验结果的断言显然都没有意义，与任何应用都没有关系. 我们的推理如下.

实验者通过向厚度为 Δ 的板上发射 N 个中子，并且观察到 n 次裂变，为我们提供了 N_n 的证据. 由于根据假设，唯一的先验知识是中子将发生或不发生裂变，因此拉普拉斯连续法则适用，第 $N+1$ 个中子在厚度为 Δ 的板上发生裂变的概率是

$$p \equiv P(F_{N+1}|N_n) = \frac{n+1}{N+2}, \tag{18.70}$$

其中

$$F_n \equiv \text{第 } n \text{ 个中子将发生裂变.} \tag{18.71}$$

不管 N 的大小如何，概率“准确性”问题都不会出现——这根据定义是精确的. 当然，我们将希望具有尽可能大的 N 值，因为这会增加证据 N_n 的权重，并使概率 p 不仅更准确，而且更稳定. 概率 p 显然不是可裂变材料的物理性质，而是仅基于证据 N_n 来描述我们对其知识状态的一种手段. 如果初步实验得出了不同的结果 N'_n，那么我们当然会分配不同的概率 p'，但可裂变材料的性质将保持不变.

现在我们向厚度为 $x = M_\delta$ 的板发射一个中子，并且定义命题：

$F^n \equiv$ 中子会在第 n 个厚度为 Δ 的板中引起裂变；

$f^n \equiv$ 中子不会在第 n 个板中引起裂变.

在第 1 个板中发生裂变的概率为

$$p \equiv P(F_1|N_n) = \frac{n+1}{N+2}. \tag{18.72}$$

但是现在，裂变将在第 2 个板中发生而在第 1 个板中不发生的概率不像在频率解中那样是 $p(1-p)$. 在这一点上，我们看到了两种理论的根本差异之一. 根据乘法规则，我们有

$$\begin{aligned} P(F^2F^1|N_n) &= P(F^2|F^1N_n)P(F^1|N_n) \\ &= \frac{n+1}{N+2}\left[1-\frac{n+1}{N+2}\right] \\ &= \frac{(n+1)(N-n+1)}{(N+2)(N+3)}. \end{aligned} \tag{18.73}$$

不同之处在于，在计算概率 $P(F^2|F^1N_n)$ 时，我们必须考虑证据 F^1，即中子已经通过了更多一个厚度为 Δ 的板而没有裂变. 这算是导致 N_n 的实验之外的另一次实验. 证据 F^1 完全与 N_n 一样有说服力，因此只考虑其中一个显然是不一致的. 以这种方式继续下去，我们发现入射中子在穿过厚度为 $M\Delta$ 的板时将精确地发射 m 个第一子代的概率正好是

$$P(M_m|N_n) = \binom{M}{m}\frac{(n+m)!(N+1)!(N+M-n-m)!}{n!(N-n)!(N+M+1)!}, \tag{18.74}$$

这正是我们之前得出的 (18.24). 现在，如果 N 不是非常大，则可能与通过频率方法获得的以下值明显不同：

$$P(M_m|A_p) = \binom{M}{m}p^m(1-p)^{M-m}, \tag{18.75}$$

但是，再次注意，随着证据权重 N_n 的增加，我们发现在极限 $N\to+\infty, n/N\to p$ 时 $(A_{p'}|N_n)\to\delta(p'-n/N)$ 且

$$P(M_m|N_n)\to P(M_m|A_p). \tag{18.76}$$

每当 $N\gg M$ 时，即当证据权重 N_n 大大超过 M_m 时，两个结果的差异可忽略不计. 现在，让我们更仔细地研究 (18.74) 和 (18.75) 之间的差异. 根据 (18.74)，我们有对 m 的均值估计，根据拉普拉斯理论是

$$\overline{m} = M\frac{n+1}{N+2}. \tag{18.77}$$

要说明此估计的准确性，我们可以计算分布 (18.74) 的方差. 使用表示形式 (18.23)，最容易做到这一点：

$$\begin{aligned} \overline{m^2} &= \sum_{m=0}^{M} m^2\int_0^1 \mathrm{d}p\, P(M_m|A_p)(A_p|N_n) \\ &= \frac{(N+1)!}{n!(N-n)!}\int_0^1 \mathrm{d}p\left[M_p+M(M-1)p^2\right]p^n(1-p)^{N-n} \\ &= M\frac{n+1}{N+2}+M(M-1)\frac{(n+1)(n+2)}{(N+2)(N+3)}, \end{aligned} \tag{18.78}$$

这将给出方差

$$V \equiv \overline{M^2} - \overline{m}^2 = M\left[\frac{N+M+2}{N+3}\right]\left[\frac{n+1}{N+2}\right]\left[n - \frac{n+1}{N+2}\right]. \tag{18.79}$$

根据 (18.75)，频率理论给出

$$\overline{m_0} = M_p, \tag{18.80}$$

$$V_0 \equiv \left[\overline{m^2} - \overline{m}^2\right]_0 = M_p(1-p). \tag{18.81}$$

如果频率主义者将置信区间 (18.68) 的中心作为他对 p 的“最优”估计，那么他将在这些等式中取 $p = (n+1)/(N+2)$. 因此，我们都获得了相同的概率估计，但方差 (18.79) 将会大

$$V - V_0 = \frac{M-1}{N+3} M_p(1-p). \tag{18.82}$$

为什么会有这种差别？为什么拉普拉斯理论似乎不如频率理论那样更精确地确定 m 的值？这里的表象是骗人的. 事实是，拉普拉斯理论比频率理论*更精确地*确定了 m 的值. 方差 (18.81) 并不是像频率理论那样是 m 的不确定性的全部度量，因为 p 的“真实”值仍然存在不确定性. 根据 (18.81)，p 的不确定性约为 $\pm\sqrt{2p(1-p)/N}$，因此 (18.80) 的不确定性除了 (18.81) 表示的不确定性外还有

$$\pm\sqrt{\frac{2p(1-p)}{N}}. \tag{18.83}$$

如果我们假设不确定性 (18.81) 和 (18.83) 是独立的，则关于频率理论上 m 值的总均方不确定性将是 (18.81) 与

$$M^2\frac{2p(1-p)}{N} \tag{18.84}$$

的和，这足够消除差别 (18.82). (18.84) 中的因子 2 当然会因采用不同的置信度而有所改变，但是没有合理的选择可以对其进行很大的改变.

在频率理论中，两种不确定性 (18.81) 和 (18.84) 表现为完全独立的效应. 这是通过应用两种不同的原理确定的：一个是常规概率理论，另一个是置信区间理论. 在拉普拉斯理论中，不存在这种区别. 两者都是通过一次计算自动给出的. 在第 6 章中，当将机器人的流程与正统统计学家的流程进行比较时，我们在粒子计数器问题中发现了完全相同的情况.

拉普拉斯理论能够做到这一点的原因非常有趣. 这只是由于已经注意到的区别. 在 (18.74) 的推导中，我们一直在考虑新实验中积累的其他证据，例如 (18.73) 中的 F^1. 在频率理论中，由于给定 N_n 的初始实验仅提供了有限数量的数据，因此出现 p 的不确定性 (18.83). 正是出于这个原因，诸如 F^1 之类的新证据仍然有意义. 在对*所有*这些证据进行一致处理时，拉普拉斯理论自动考虑了初始数据有

限性的影响，而频率理论只有通过引入置信区间才可以粗略地做到这一点. 在拉普拉斯理论中，没有必要确定任何随意的“置信度”，因为当概率论一致地应用于整个问题时，已经告诉我们应该对初始数据 N_n 赋予多大权重. 我们得到的不仅是更统一的处理方式，拉普拉斯理论会得到 m 的更小的总不确定性，这表明频率理论的两种不确定性来源 (18.81) 和 (18.84) 不是独立的：它们具有小的负相关性，因此倾向于相互补偿. 这就是拉普拉斯理论得到较小的可能误差的原因. 如果仔细思考一下，你就能够凭直觉明白为什么存在这种负相关性——我不想剥夺你亲自解决这一问题的乐趣. 所有这些微妙之处在频率理论中都完全消失了.

“但是”，有人会反对说，“你忽略了一个非常实际的考虑因素，这是引入置信区间的最初原因. 虽然我认为原则上最好用一种计算方式来处理整个问题，但实际上通常必须将其分解为两个不同的问题. 毕竟，初始数据 N_n 是由一群人获得的，他们必须将其结果传达给另一群人，这群人使用这些数据进行二次计算. 实际上必要的是，第一群人必须能够如实地陈述他们的发现以及结论的可靠性. 他们的数据还可以用于二次计算之外的其他用途，因此置信区间的引入满足了不同人之间沟通的重要实际需求.”

当然，如果理解了到目前为止的所有内容，那么你就会知道对于这种异议的回答. 记忆存储问题是我们最初的出发点，而刚才讨论的问题只是更抽象的 (18.16) 的一个具体示例. 根据 (18.23) 以及我们对 (18.79) 的推导，你就会明白，分析整个问题所需的初始数据的唯一性质是初始实验得到的 A_p 分布 $(A_p|N_n)$. 引入置信区间的原理是为了满足非常实际的需求. 但是没有必要为此目的引入任何新的原理，它已经包含在概率论中. 概率论表明，传达你所学到知识的确切方式不是通过指定置信区间，而是通过指定最终的 A_p 分布.

作为进一步的比较，请注意在拉普拉斯理论中，没有必要在连续的厚度为 Δ 的板中引入任何关于事件独立性的“统计假设”. 实际上，该理论告诉我们（如 (18.73) 中所述），当只有有限数量的初始数据时，这些概率并不是独立的. 正是这一事实使得拉普拉斯理论能够考虑频率理论通过置信区间描述的不确定性.

这就得到了关于概率论的一个非常基本的点，它是频率理论未能认识到的，但是正如我们将在后面说明的那样对于传播理论和统计力学的应用都必不可少. 我们说两个事件“相互独立”究竟是什么意思?

在频率理论中，唯一认识到的独立性是因果独立性，即一个事件发生的事实本身并不会对另一事件产生任何物理作用. 因此，在第 6 章讨论的抛硬币示例中，硬币在一次抛掷中正面朝上的事实当然不会在物理上影响下一次抛掷的结果，因此在频率理论中，人们称抛硬币实验是典型的“独立重复随机试验”，两次抛掷结

果的概率一定是各次概率的乘积. 但是这样我们就无法描述 A 与 B 的推理之间的区别了!

在拉普拉斯理论中, "独立性" 意味着完全不同的东西. 我们根据乘法规则 $P(AB|C) = P(B|C)P(A|BC)$ 一目了然地知道, 独立是指 $P(A|BC) = P(A|C)$, 即知道 B 为真并不影响我们分配给 A 的概率. 因此, 独立性不只是意味着因果独立性, 而是逻辑独立性. 即使第一次抛掷并不会在物理上影响下一次抛掷正面朝上的倾向, 但我们所知道的第一次抛掷结果的知识可能会对我们对下一次抛掷正面朝上的预测产生很大的影响.

这一点的重要性在于, 各种极限定理 (将在后面详细介绍) 在推导中需要独立性. 因此, 即使可能存在严格的因果独立性, 但如果没有逻辑独立性, 则这些极限定理也不成立. 频率派思想的作者否认概率论与归纳推理有关, 只认识到因果联系的存在. 因此, 他们长期以来一直将这些极限定理应用于物理和通信过程. 我们声称这是错误和完全误导人的. 凯恩斯 (Keynes, 1921) 很早就注意到了这一点, 他也强调了同样的观点.

我认为这些比较让我们很清楚: 至少在这种问题上, 拉普拉斯理论确实提供了我的同事所追求的 "更好的理解及更有效的形式".

18.16 德菲内蒂定理

到目前为止, 我们考虑了 A_p 分布的概念, 并且在所有试验均服从相同 A_p 分布的限制条件下从中导出了一类的概率分布. 直观上, 这意味着我们假设背后的 "机制" 是不变而未知的. 显然, 这是一个限制性很强的假设, 并且自然会产生一个问题: 以这种方式可以获得的概率函数类别的一般性如何? 为了清楚地说明问题, 让我们定义

$$x_n \equiv \begin{cases} 1, & \text{如果在第 } n \text{ 次试验中 } A \text{ 为真}, \\ 0, & \text{如果在第 } n \text{ 次试验中 } A \text{ 为假}. \end{cases} \tag{18.85}$$

那么, N 次试验的知识状况最一般地是通过概率函数 $P(x_1 \cdots x_N|N)$ 描述的. 原则上, 这可以在 2^N 个点上任意定义 (除了必须满足归一化条件外).

现在我们问: 要从 A_p 分布中导出 $P(x_1 \cdots x_N|N)$ 的充分必要条件是什么? 我们可以对给定的分布 $P(x_1 \cdots x_N|N)$ 进行何种检验, 以判断该分布是否包含在上面给出的理论中? 根据前面的方程式可以清楚地看出一个必要条件: 从 A_p 分布获得的任何分布都必须具有以下性质: 在 n 次指定试验中 A 为真、在其余 $N-n$ 次为假的概率仅仅依赖于数 n 和 N, 即不能依赖于指定的具体试验. 如果是这样,

我们说 $P(x_1\cdots x_N|N)$ 定义了一个可交换序列.

德菲内蒂（de Finetti，1937）的一个重要定理断言反之亦然：任何可交换概率函数 $P(x_1\cdots x_N|N)$ 都可以通过一个 A_p 分布生成. 因此，存在函数 $(A_p|X)=g(p)$ 使得 $g(p)\geqslant 0,\int_0^1 \mathrm{d}p\, g(p)=1$，并且在 N 次试验中 A 在 n 次指定试验中为真、在其余 $N-n$ 次试验中为假的概率为

$$P(n|N)=\int_0^1 \mathrm{d}p\, p^n(1-p)^{N-n}g(p). \tag{18.86}$$

这可以证明如下. 注意，$p^n(1-p)^{N-n}$ 是 N 次多项式：

$$p^n(1-p)^{N-n}=p^n\sum_{m=0}^{N-n}\binom{N-m}{m}(-p)^m=\sum_{k=0}^{N}\alpha_k(N,n)p^k, \tag{18.87}$$

它定义了 $\alpha_k(N,n)$. 因此，如果 (18.86) 成立，我们将有

$$P(n|N)=\sum_{k=0}^{N}\alpha(N,n)\beta_k, \tag{18.88}$$

其中

$$\beta_k=\int_0^1 \mathrm{d}p\, p^n g(p) \tag{18.89}$$

是 $g(p)$ 的 n 阶矩. 因此，指定 $\beta_0,\cdots,\beta_N$ 等同于对于 $n=0,\cdots,N$ 指定所有 $P(n|N)$. 反之，对于给定的 N，指定 $P(n|N)$（$0\leqslant n\leqslant N$）等价于指定 $\{\beta_0,\cdots,\beta_N\}$. 实际上，$\beta_N$ 是 $x_1=x_2=\cdots=x_N=1$ 的概率，而不管以后试验的结果，并且可以无须引入任何函数 $g(p)$ 直接确定其与 $P(n|N)$ 的关系.

因此，问题可以简化为：如果指定 $\beta_0,\cdots,\beta_N$，在什么条件下存在函数 $g(p)\geqslant 0$ 使得 (18.89) 成立？这正好是众所周知的豪斯道夫矩问题. 它的解可以在很多地方找到，例如威德的著作（Widder，1941，第 3 章）. 转化为我们的符号，主要定理如下. 满足条件 (18.89)（因此也满足条件 (18.86)）的函数 $g(p)\geqslant 0$ 存在的充分必要条件是存在数 B 使得

$$\sum_{n=0}^{N}\binom{N}{n}P(n|N)\leqslant B,\qquad N=0,1,\cdots \tag{18.90}$$

但是，将 $P(n|N)$ 解释为概率，我们看到在 $B=1$ 时等号对于 (18.90) 始终成立，证毕.

还有另外一种看待这一定理的方法. 我们可能需要更多的工作来证明它，但是也许可以更清楚地揭示德菲内蒂定理的直观原因，并且在指定 $P(n|N)$ 时可以

立即表明我们对 $g(p)$ 说了什么. 想象 $g(p)$ 可以展开为

$$g(p)=\sum_{n=0}^{\infty}a_n\phi_n(p), \tag{18.91}$$

其中 $\phi_n(p)$ 是 $0\leqslant p\leqslant 1$ 的多项式的完全正交集合，本质上是勒让德函数

$$\phi_n(p)=\frac{\sqrt{2n+1}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}p^n}\big[p(1-p)\big]^n=(-1)^n\sqrt{2n+1}P_n(2p-1), \tag{18.92}$$

其中 $\phi_n(p)$ 是 n 次多项式，满足条件

$$\int_0^1\mathrm{d}p\,\phi_m(p)\phi_n(p)=\delta_{mn}. \tag{18.93}$$

如果将 (18.93) 代入 (18.86)，那么由于 $\phi_k(p)$ 与 $N<k$ 次多项式正交，只有有限项不为 0. 这样，容易看出对于给定的 N，指定 $P(n|N)$（$0\leqslant n\leqslant N$）的值等价于指定前面 $n+1$ 个展开系数 $\{a_0,\cdots,a_N\}$. 因此，当 $N\to+\infty$ 时，由 (18.91) 定义的函数 $g(p)$ 是唯一确定的，与傅里叶级数“几乎处处”唯一确定其生成函数一样. 这一论证的主要问题是根据 (18.91) 很难确定条件 $g(p)\geqslant 0$.

18.17 评注

德菲内蒂定理对于我们非常重要，因为它表明我们在本章中发现的概率与频率之间的联系对于相当广泛的一类概率函数 $P(x_1,\cdots,x_N|N)$ 成立，即对于所有可互换序列类成立. 当然，这些结果可以立即推广到每次试验有两个以上可能结果的情况.

然而，可能更重要的是，德菲内蒂定理使得概率论中最古老的争议之一——拉普拉斯对连续法则的最初推导——更容易被理解. 毫无疑问，A_p 分布的思想并不是我们的发明. 我们在这里引入它只是尝试将拉普拉斯的以下著名段落翻译为现代语言. 他说：“当一个简单事件的概率未知时，我们可以假设此概率介于 0 和 1 之间的所有值的可能性相同.”我们将此解释为，在没有任何先验证据的情况下 $(A_p|X)=$ 常数. 本世纪的几乎所有概率论学者都认为这一陈述没有意义. 当然，就概率的频率定义而言，拉普拉斯的陈述根本无法合理解释. 但是对于任何理论，这在概念上都是困难的，因为它似乎涉及“概率的概率”的概念. 并且，自从拉普拉斯时代以来，人们一直避免在计算中使用 A_p 分布.

德菲内蒂定理为这些方法奠定了坚实的基础. 与所有概念问题无关，这是一个数学定理. 每当你谈论某个情况时，某个结果序列的概率仅仅取决于成功的次数，而不取决于某次特定试验是否成功，那么你的所有概率分布可以由单个函数 $g(p)$ 生成，就像我们在这里所做的那样. 而且，生成函数的使用在数学上是一种

非常强大的技术. 如果你尝试在不使用 A_p 分布的情况下重复上述某些推导过程(例如 (18.24)),很快就会发现这一点. 因此,我们如何从概念上看待 A_p 分布并不重要. 它作为处理可交换序列的数学工具的有效性已被证实,无关乎哲学上的反对意见.

第 19 章　物理测量

我们在第 7 章中看到，连伟大的数学家欧拉也无法解决根据木星和土星位置的 75 个不同观测值估计 8 个轨道参数的问题. 他从演绎逻辑的角度思考，甚至无法想象解决此类问题的原理. 但是，在 38 年后，拉普拉斯从作为逻辑的概率论的角度出发，掌握了解决木星与土星之间巨大不平衡问题的正确原理. 在本章中，我们将通过考虑根据 3 个观测值估计 2 个参数的问题来展示今天的解法. 我们的一般解（以矩阵符号表示）将自动包含拉普拉斯的解决方案.

19.1　条件方程的简化

假设我们想确定电子的电荷 e 和质量 m. 密立根油滴实验可以直接测量 e. 通过测量电子束在已知电磁场中的偏转可以测量比率 e/m. 通过测量由于镜像电荷的吸引而产生的电子向金属板的偏移可以测量 e^2/m.

根据这三个实验中的任意两个结果，我们可以计算 e 和 m 的值. 但是所有实验的测量值都有误差，根据不同实验获得的 e 和 m 值将不一致. 不过每个实验结果中的确都包含一些与我们的问题相关的其他实验不包含的信息. 我们如何处理数据，利用所有可用信息获得 e 和 m 的最优估计呢？误差有多大？通过另一个给定精度的实验，情况会有多大改善？概率论给这些问题提供了简单而优雅的答案.

更具体地说，假设我们有以下实验结果：

(1) 以 $\pm 2\%$ 的精度测量 e；

(2) 以 $\pm 1\%$ 的精度测量 e/m；

(3) 以 $\pm 5\%$ 的精度测量 e^2/m.

假设预先已知的 e 和 m 的值大约为 $e \approx e_0, m \approx m_0$，那么测量结果就是校正项的线性函数. 将 e 和 m 的未知真值写成

$$\begin{aligned} e &= e_0(1 + x_1), \\ m &= m_0(1 + x_2), \end{aligned} \tag{19.1}$$

那么 x_1 和 x_2 是小于 1 的无量纲校正项，我们的问题就是要找到 x_1 和 x_2 的最

优估计. 这三个测量结果是三个数 M_1, M_2, M_3，我们将其写为

$$\begin{aligned} M_1 &= e_0(1+y_1), \\ M_2 &= \frac{e_0}{m_0}(1+y_2), \\ M_3 &= \frac{e_0^2}{m_0}(1+y_3), \end{aligned} \tag{19.2}$$

其中 y_i 是由 (19.2) 定义的小的无量纲数，可以根据旧估计 e_0, m_0 和新测量值 M_1, M_2, M_3 确定. 另外，$e, e/m, e^2/m$ 的真实值可以用 x_j 表示为

$$\begin{aligned} e &= e_0(1+x_1), \\ \frac{e}{m} &= \frac{e_0(1+x_1)}{m_0(1+x_2)} = \frac{e_0}{m_0}(1+x_1-x_2+\cdots), \\ \frac{e^2}{m} &= \frac{e_0^2(1+x_1)^2}{m_0(1+x_2)} = \frac{e_0^2}{m_0}(1+2x_1-x_2+\cdots), \end{aligned} \tag{19.3}$$

其中高阶项可以忽略不计. 比较 (19.2) 和 (19.3)，我们看到，如果测量准确，将有

$$\begin{aligned} y_1 &= x_1, \\ y_2 &= x_1 - x_2, \\ y_3 &= 2x_1 - x_2. \end{aligned} \tag{19.4}$$

但是，考虑到误差，已知的 y_i 与未知的 x_j 有关：

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \delta_1, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \delta_2, \\ y_3 &= a_{31}x_1 + a_{32}x_2 + \delta_3, \end{aligned} \tag{19.5}$$

其中系数 a_{ij} 形成 3×2 矩阵：

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 2 & -1 \end{pmatrix}, \tag{19.6}$$

δ_i 是这三个测量值的未知误差. 例如，$\delta_2 = -0.01$ 表示第二次测量得出的结果小了 1%.

更一般地，我们要根据 N 个不完美的观测 $\{y_1, \cdots, y_N\}$ 以及 N 个条件方程

$$y_i = \sum_{j=1}^{n} a_{ij}x_j + \delta_i, \qquad i = 1, 2, \cdots, N, \tag{19.7}$$

或者以矩阵符号表示

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\delta}, \tag{19.8}$$

其中 $\boldsymbol{A}$ 是 $N \times n$ 矩阵，来估计 n 个未知量 $\{x_1, \cdots, x_n\}$. 在当前讨论中，我们假设问题是“超定的”，即 $N > n$. 这种情况难倒了欧拉（Euler，1749），他面对的是 $N = 75, n = 8$ 的情况. 但是我们要记住，$N = n$（表面上看似定义良好）和 $N < n$（欠定）的情况也可能在实际问题中出现. 看看概率论会对这两种情况有什么结论吧，这会很有趣.

在 19 世纪初期，人们普遍会做如下推理. 似乎合情的是，每个 x_j 的最优估计是所有 y_i 的某种线性组合，但是如果 $N > n$，我们不能简单求解方程 (19.8) 得出 x，因为 $\boldsymbol{A}$ 不是方阵并且没有逆. 但是，如果我们有 n 个线性组合作为条件方程，就可以得到一个能求解出 x 的方程组. 也就是说，如果我们用某个 $n \times N$ 矩阵 $\boldsymbol{B}$ 去左乘方程 (19.8)，那么乘积 $\boldsymbol{BA}$ 存在，并且是一个 $n \times n$ 方阵. 选择 $\boldsymbol{B}$ 使得 $(\boldsymbol{BA})^{-1}$ 存在，那么线性组合是 n 行的

$$\boldsymbol{By} = \boldsymbol{BAx} + \boldsymbol{B\delta}, \tag{19.9}$$

它有唯一解

$$\boldsymbol{x} = (\boldsymbol{BA})^{-1}\boldsymbol{B}(\boldsymbol{y} - \boldsymbol{\delta}). \tag{19.10}$$

如果各个分误差 δ_i 的概率是对称的：$p(\delta_i) = p(-\delta_i)$ 使得 $\langle \delta_i \rangle = 0$，那么对应于任何给定的矩阵 $\boldsymbol{B}$，对于几乎任一合理的损失函数，x_j 的“最优”估计将是

$$\hat{\boldsymbol{x}} = (\boldsymbol{BA})^{-1}\boldsymbol{By} \tag{19.11}$$

的第 j 行，但是选择不同的 $\boldsymbol{B}$（即采用不同的线性组合的条件方程），我们会得到不同的估计值. 在欧拉问题中，存在数十亿种可能的选择，那么 $\boldsymbol{B}$ 的最优选择是什么?

在上文中，我们只是用现代的记号（但还是用旧的语言）重述了拉普拉斯《哲学评论》（Laplace，1812）中“条件方程的简化”问题. 解的流行准则是最小二乘原理：找到使得 $\hat{x}_j$ 误差平方和或加权和最小的矩阵 $\boldsymbol{B}$. 这个问题可以直接解决，下面我们将通过不同的方式得到相同的解.

19.2　重表述为决策问题

我们在第 13 章中实际上已经解决了这个问题. 在那里我们看到，在任何损失函数的标准下，任何参数的最优估计一般都可以通过应用贝叶斯定理找到以数据为条件的参数位于不同区间的概率，然后做出使后验概率的期望损失最小的估计来得到.

在上面给出的问题的原始表述中，x_j 的最优估计是如 (19.11) 所示的 y_i 的线性组合，这只是一个合情的猜测. 在第 13 章中，我们展示了一种更好的表述问题的方法，其中不必依赖于猜测. 与其在不知道采用哪种组合的情况下尝试线性组合，

不如将贝叶斯定理直接应用于条件方程. 这样，如果最优估计的确是 (19.11) 的线性形式，那么贝叶斯定理不仅应该能告诉我们这一点，还可以自动给出矩阵 $\boldsymbol{B}$ 的最优选择，并告诉我们最小二乘法不能给出的估计的精度.

让我们在给各种测量的误差 δ_i 分配高斯概率的情况下进行计算. 根据第 7 章的讨论，这几乎总是我们可以根据所拥有的信息分配概率的最优误差定律. 但是在正统文献中，人们不这样看. 人们会认为：在大多数物理测量中，总误差是许多因果独立的微小误差之和，利用中心极限定理将得到误差的高斯频率分布.[①] 这种观点也没有什么错，只是它误导了几代概率工作者，使得他们得出结论：如果误差的频率分布实际上不是高斯分布，那么分配高斯分布就是“假设”某种不正确的东西是对的. 这将导致我们的最终结论存在某种可怕的错误.

关于高斯误差分布的说明

第 7 章的讨论使我们确信，这种危险被严重夸大了. 关键是，在作为逻辑的概率论中，高斯概率分配不是关于误差频率的假设，而是我们对误差了解状态的描述. 关于误差，我们除了大致幅度之外几乎没有任何先验知识，这种知识可以合理地解释为指定误差分布的前二阶矩. 根据最大熵原理，在与该信息相符但没有其他假设时，这将导致独立的高斯概率分配. 合理可能噪声向量 $(\delta_1, \cdots, \delta_N)$ 域 Ω 或合理可能数据向量域 $\{\boldsymbol{Ax} + \boldsymbol{\delta}\}$ 在满足二阶矩约束时尽可能大. 在看到数据之前，误差的频率分布几乎总是未知的. 但是即使频率分布与高斯分布相差很大，高斯概率分配仍将使我们从已知信息中得出最优推断.

高斯分布的特殊地位在于一个更微妙的事实：如果新信息能根据旧信息预测到，那么获取到的新信息就不会影响我们的推断. 因此，如果我们分配的是高斯概率分布，然后获取到误差的真实频率分布确实是具有指定方差的高斯分布的新信息，那么这对我们没有什么帮助，因为这只是我们能预测到的. 但是，如果我们有关于误差频率偏离高斯分布的其他特定先验信息，那就是将可能的误差向量限制在较小域 $(\Omega_1 \subset \Omega)$ 的强有力的新信息. 这将使我们能够在以下参数估计的基础上进行改进，因为补集 $\Omega - \Omega_1$ 中的数据向量以前被认为是噪声，现在已经被识别为真实的“信号”. 贝叶斯定理能自动为我们完成所有这些工作.

因此，我们与大自然的契约比正统统计教义中所设想的要有利得多. 因为，在给定二阶矩时，非高斯频率分布不会使我们的推断变得更糟，但是对非高斯分布的了解可以使我们得到比下面更好的结果.

---

① 如第 14 章所述，这里有一个重要的限定条件：通常，高斯近似仅适用于总误差 δ 的值可以通过各种小误差的组合以多种方式产生的情况. 对于异常大的误差，我们不会预期，也几乎不会观察到高斯频率.

受以上消息的鼓舞，我们将误差 $\{\delta_1,\cdots,\delta_N\}$ 分别位于区间 $\{\mathrm{d}\delta_1,\cdots,\mathrm{d}\delta_N\}$ 中的概率分配为

$$p(\delta_1\cdots\delta_N)\mathrm{d}\delta_1\cdots\mathrm{d}\delta_N=(\text{常数})\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}w_i\delta_i^2\right\}\mathrm{d}\delta_1\cdots\mathrm{d}\delta_N,\tag{19.12}$$

其中“权重” w_i 是第 i 个测量的误差的方差倒数. 例如，第 1 次测量结果具有 $\pm2\%$ 精度的粗略描述现在变成更精确的描述，结果具有权重

$$w_1=\frac{1}{\langle\delta_1^2\rangle}=\frac{1}{0.02^2}=2500.\tag{19.13}$$

目前，我们假设这些权重是已知的. 这正是在天文和其他物理数据中的情况. 根据 (19.7) 和 (19.12)，给定真实值 $\{x_1,\cdots,x_n\}$，我们立即可以获得测量值 $\{y_1,\cdots,y_N\}$ 的抽样概率密度：

$$p(y_1\cdots y_N|x_1\cdots x_n)=C_1\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}w_i\left[y_i-\sum_{j=1}^{n}a_{ij}x_j\right]^2\right\},\tag{19.14}$$

其中 C_1 不依赖 y_i. 根据贝叶斯定理，如果我们给 x_j 分配均匀先验概率，则在给定实际值 y_i 的情况下，x_j 的后验概率密度为

$$p(x_1\cdots x_n|y_1\cdots y_N)=C_2\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}w_i\left[y_i-\sum_{j=1}^{n}a_{ij}x_j\right]^2\right\},\tag{19.15}$$

现在 C_2 不依赖 x_j. 接下来，就像在几乎所有的高斯计算中一样，我们需要重新将其组织为二次型形式，以分离出对 x_i 的依赖. 通过展开它，我们得到

$$\begin{aligned}\sum_{i=1}^{N}w_i\left(y_i-\sum_{j=1}^{n}a_{ij}x_j\right)^2&=\sum_{i=1}^{N}w_i\left(y_i^2-2y_i\sum_{j=1}^{n}a_{ij}x_j+\sum_{j,k=1}^{n}a_{ij}a_{ik}x_jx_k\right)\\&=\sum_{j,k=1}^{n}K_{jk}x_ix_k-2\sum_{j=1}^{n}L_jx_j+\sum_{i=1}^{N}w_iy_i^2,\end{aligned}\tag{19.16}$$

其中

$$K_{jk}=\sum_{i=1}^{N}w_ia_{ij}a_{ik},\qquad L_j=\sum_{i=1}^{N}w_iy_ia_{ij}.\tag{19.17}$$

或者，可以定义对角线“权重”矩阵 $\boldsymbol{W}$，其中 $W_{ij}=w_i\delta_{ij}$，我们得到矩阵 $\boldsymbol{K}$ 和向量 $\boldsymbol{L}$：

$$\boldsymbol{K}=\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A},\qquad\boldsymbol{L}=\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{y},\tag{19.18}$$

其中 $\tilde{\boldsymbol{A}}$ 是转置矩阵. 我们要将 (19.15) 写成

$$p(x_1\cdots x_n|y_1\cdots y_N)=C_3\exp\left\{-\frac{1}{2}\sum_{j,k=1}^{n}K_{jk}(x_j-\hat{x}_j)(x_k-\hat{x}_k)\right\},\tag{19.19}$$

其中 $\hat{x}_j$ 将是所需的均值估计值. 比较 (19.16) 和 (19.19)，我们看到

$$\sum_{k=1}^{n} K_{jk}\hat{x}_k = L_j, \tag{19.20}$$

所以如果 $\boldsymbol{K}$ 是非奇异矩阵，我们可以唯一地求解 $\hat{\boldsymbol{x}}$.

19.3 欠定情形：$\boldsymbol{K}$ 奇异

如果观测值的个数少于参数，即 $N < n$，那么根据 (19.17)，$\boldsymbol{K}$ 仍然是 $n \times n$ 矩阵，但是秩最多为 N，因此可能是奇异的. 因此问题不是方程 (19.20) 没有解，而是可能有无数个解. 最大似然不是在某个点上，而是在 $n-N$ 维线性流形上达到. 当然，尽管 $(\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A})^{-1}$ 不存在，但是 $(\boldsymbol{A}\tilde{\boldsymbol{A}})^{-1}$ 存在，从这一事实可以看出，最大似然解仍然存在，所以参数估计

$$\boldsymbol{x}^* = \tilde{\boldsymbol{A}}(\boldsymbol{A}\tilde{\boldsymbol{A}})^{-1}\boldsymbol{y} \tag{19.21}$$

使 (19.15) 中的二次型消失：$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^*$，达到了最大可能似然. 这称为规范逆解，可以通过最大熵原理进行计算. 但是规范逆绝不是唯一的，因为从 (19.8) 中可以看出，如果将齐次方程 $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$ 的任一解 $\boldsymbol{z}$ 加到估计 (19.21) 中，将得到相同似然的另一估计 $\boldsymbol{x}^* + \boldsymbol{z}$，且有维数为 $n-N$ 的此类向量 $\boldsymbol{x}^* + \boldsymbol{z}$ 的线性流形 Δ.

练习 19.1 证明规范逆解 (19.21) 也是最小二乘解，使得流形 Δ 的 $\sum(x_i^*)^2$ 最小. 不幸的是，并没有令人信服的理由使估计向量的长度最小.

很长时间以来，人们没有找到解决此类问题的满意方法. 但是我们并非完全无助，因为数据确实将参数 $\{x_i\}$ 的可能值限制在满足 (19.20) 的“可能集合” Δ 中. 单靠数据是无法挑选出该集合中的任何唯一点的. 尽管如此，如果数据加上先验信息，我们仍然有可能做出有用的选择. 这是“广义逆”问题，在许多应用（例如图像重建）中具有重要意义. 实际上，在现实世界中，广义逆问题可能占了绝大多数，因为在现实世界中，我们很少有提出定义良好问题所需要的所有信息. 然而，在许多情况下，可以通过最大熵找到有用的解，该解以第 11 章和第 20 章中所述的方式根据几种不同的准则是“最优”的.

19.4 超定情形：$\boldsymbol{K}$ 非奇异

根据 (19.17) 的定义，$\boldsymbol{K}$ 是 $n \times n$ 矩阵，并且对于所有实数 $\{q_1, \cdots, q_n\}$ 使得 $\sum q_i^2 > 0$，

$$\sum_{j,k=1}^{n} K_{jk}q_jq_k = \sum_{i=1}^{N} w_i \left(\sum_{j=1}^{n} a_{ij}q_j\right)^2 \geqslant 0, \tag{19.22}$$

因此，如果 $\boldsymbol{K}$ 的秩为 n，则它不仅是非奇异的，而且是正定的. 如果 $N \geqslant n$，就将是这种情况，除非我们在定义问题上做一些愚蠢的事情——包括无用的观察或不相关的参数，等等.

首先，我们假设所有权重 w_i 都为正：因为如果任何观测值 y_i 有权重 $w_i = 0$，那么它在我们的问题中是没有用的，即它无法传达有关参数的任何信息，我们根本不应该将其包含在数据集中，这时可以将 N 减 1.

其次，如果有一个非零向量 $\boldsymbol{q}$ 使得 $\sum_j a_{ij} q_j$ 对于所有 i 都是 0，那么在 (19.7) 中，对于所有 c，参数集 $\{x_j\}$ 和 $\{x_j + cq_j\}$ 将导致等价的数据而难以区分. 换句话说，问题中有一个与数据无关的不相关参数，这时我们将 n 减 1. 数学上，这意味着矩阵的列并不线性独立. 这样，如果 $\boldsymbol{q} \neq 0$，我们可以删除参数 x_k 和 $\boldsymbol{A}$ 的第 k 列而不对问题产生任何影响（即对我们能获取的信息没有改变）.

必要时删除不相关的观测值和参数，最后，观测值的数量至少与相关参数的个数一样，那么 $\boldsymbol{K}$ 将为正定矩阵，并且方程 (19.20) 具有唯一解

$$\hat{\boldsymbol{x}}_k = \sum_{j=1}^{n} \left(\boldsymbol{K}^{-1}\right)_{kj} \boldsymbol{L}_j. \tag{19.23}$$

根据 (19.18)，我们可以将结果写为

$$\hat{\boldsymbol{x}} = (\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A})^{-1}\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{y}, \tag{19.24}$$

与 (19.11) 相比，我们发现，在均匀先验概率的高斯情况下，最优估计的确是 (19.11) 形式的测量值的线性组合，而矩阵 $\boldsymbol{B}$ 的最优选择是

$$\boldsymbol{B} = \tilde{\boldsymbol{A}}\boldsymbol{W}, \tag{19.25}$$

这个结果也许是高斯首先发现的，并在拉普拉斯的《哲学评论》中重复出现. 让我们对我们的简单问题计算这个解.

19.5　结果的数值计算

将解 (19.24) 应用于估计 e 和 m 的问题，e, e/m, e^2/m 的测量精度分别为 2%, 1%, 5%，因此

$$w_2 = \frac{1}{0.01^2} = 10\,000, \qquad w_3 = \frac{1}{0.05^2} = 400, \tag{19.26}$$

我们之前发现 $w_1 = 2500$，因此

$$\boldsymbol{B} = \tilde{\boldsymbol{A}}\boldsymbol{W} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix} \begin{pmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_3 \end{pmatrix} = \begin{pmatrix} w_1 & w_2 & 2w_3 \\ 0 & -w_2 & -w_3 \end{pmatrix}, \tag{19.27}$$

$$\boldsymbol{K}=\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A}=\begin{pmatrix}[w_1+w_2+4w_3] & -[w_2+2w_3]\\ -[w_2+2w_3] & [w_2+w_3]\end{pmatrix}, \tag{19.28}$$

$$\boldsymbol{K}^{-1}=(\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A})^{-1}=\frac{1}{|\boldsymbol{K}|}\begin{pmatrix}[w_2+w_3] & [w_2+2w_3]\\ [w_2+2w_3] & [w_1+w_2+4w_3]\end{pmatrix}, \tag{19.29}$$

其中

$$|\boldsymbol{K}|=\det(\boldsymbol{K})=w_1w_2+w_2w_3+w_3w_1. \tag{19.30}$$

因此最终结果是

$$(\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A})^{-1}\tilde{\boldsymbol{A}}\boldsymbol{W}=\frac{1}{|\boldsymbol{K}|}\begin{pmatrix}w_1[w_2+w_3] & -w_2w_3 & w_2w_3\\ w_1[w_2+2w_3] & -w_2[w_1+2w_3] & w_3[w_2-w_1]\end{pmatrix}, \tag{19.31}$$

x_1 和 x_2 的最优点估计是

$$\begin{aligned}\hat{x}_1&=\frac{w_1(w_2+w_3)y_1+w_2w_3(y_3-y_2)}{w_1w_2+w_2w_3+w_3w_1},\\ \hat{x}_2&=\frac{w_1w_2(y_1-y_2)+w_2w_3(y_3-2y_2)+w_3w_1(2y_1-y_3)}{w_1w_2+w_2w_3+w_3w_1}.\end{aligned} \tag{19.32}$$

代入 w_1,w_2,w_3 的数值，我们得到

$$\begin{aligned}\hat{x}_1&=\frac{13}{15}y_1+\frac{2}{15}(y_3-y_2),\\ \hat{x}_2&=\frac{5}{6}(y_1-y_2)+\frac{2}{15}(y_3-2y_2)+\frac{1}{30}(2y_1-y_3),\end{aligned} \tag{19.33}$$

结果显示，最优估计值作为可能实验对中估计的加权平均值. 因此，y_1 是在第 1 个实验（直接测量 e 的值）中获得的 x_1 的估计值. 第 2 个和第 3 个实验相结合得出了一个由 $(e^2/m)(e/m)^{-1}$ 形式给出的 e 的估计. 由于

$$\frac{\frac{e_0^2}{m_0}(1+y_3)}{\frac{e_0}{m_0}(1+y_2)}\approx e_0(1+y_3-y_2), \tag{19.34}$$

y_3-y_2 是第 2 个和第 3 个实验给出的 x_1 的估计值. (19.33) 表示，x_1 的这两个独立估计值应该以权重 13/15 和 2/15 组合. 同样，$\hat{x}_2$ 是 x_2 的 3 个不同（尽管不是独立的）估计的加权平均值.

19.6 估计的精度

根据 (19.19) 我们发现 $p(x_1\cdots x_n|y_1\cdots y_N)$ 的第二中心矩为

$$\langle(x_j-\hat{x}_j)(x_k-\hat{x}_k)\rangle=\langle x_jx_k\rangle-\langle x_j\rangle\langle x_k\rangle=(\boldsymbol{K}^{-1})_{jk}. \tag{19.35}$$

因此，根据我们在 $\hat{x}_j$ 的计算中已经确定的 $n\times n$ 逆矩阵

$$\boldsymbol{K}^{-1}=(\tilde{\boldsymbol{A}}\boldsymbol{W}\boldsymbol{A})^{-1}, \tag{19.36}$$

我们还可以得到可能误差或者标准差. 根据 (19.29), 我们可以用 (均值)±(标准差) 形式将结果表示为

$$(x_j)_{\text{est}} = \hat{x}_j \pm \sqrt{\left(\boldsymbol{K}^{-1}\right)_{jj}}. \tag{19.37}$$

(19.24) 和 (19.37) 是欧拉想要的问题的一般解. 在当前情况下是

$$\begin{aligned}(x_1)_{\text{est}} &= \hat{x}_1 \pm \sqrt{\frac{w_2+w_3}{w_1w_2+w_2w_3+w_3w_1}},\\ (x_2)_{\text{est}} &= \hat{x}_2 \pm \sqrt{\frac{w_1+w_2+4w_3}{w_1w_2+w_2w_3+w_3w_1}},\end{aligned} \tag{19.38}$$

数值是

$$\begin{aligned}x_1 &= \hat{x}_1 \pm 0.0186,\\ x_2 &= \hat{x}_2 \pm 0.0216,\end{aligned} \tag{19.39}$$

因此, 根据这三个测量值, 我们得出 e 的精度为 $\pm 1.86\%$, m 的精度为 $\pm 2.16\%$.

e^2/m 的比较差的测量结果 (仅 $\pm 5\%$ 的精度) 对我们有多大帮助呢? 要回答这个问题, 注意, 如果没有该实验, 我们将在极限 $w_3 \to 0$ 的情况下得出 (19.28)、(19.29) 和 (19.32). 结论也很容易从问题的陈述中得出, 是

$$\begin{aligned}\hat{x}_1 &= y_1,\\ \hat{x}_2 &= y_1 - y_2,\end{aligned} \tag{19.40}$$

$$\boldsymbol{K}^{-1} = \frac{1}{w_1w_2}\begin{pmatrix} w_2 & w_2 \\ w_2 & [w_1+w_2] \end{pmatrix}, \tag{19.41}$$

或者, (均值) ± (标准差) 为

$$\begin{aligned}x_1 &= y_1 \pm \frac{1}{w_1} = y_1 \pm 0.020,\\ x_2 &= y_1 - y_2 \pm \sqrt{\frac{w_1+w_2}{w_1w_2}} = y_1 - y_2 \pm 0.024.\end{aligned} \tag{19.42}$$

正如基于常识所能预料的那样: 低精度测量对准确测量的结果影响很小, 如果 e^2/m 的测量精度远差于 $\pm 5\%$, 那么几乎不值得将其包含在计算中. 但是, 假设一种改进的技术为我们提供了 $\pm 2\%$ 精度的 e^2/m 测量值, 那将有多大帮助呢? 答案仍由我们前面的公式给出, 其中 $w_1 = w_3 = 2500, w_2 = 10\,000$. 我们发现均值估计对于使用 e^2/m 测量值的估计会给出更大权重:

$$\begin{aligned}\hat{x}_1 &= 0.556y_1 + 0.444(y_3 - y_2),\\ \hat{x}_2 &= 0.444(y_1 - y_2) + 0.444(y_3 - 2y_2) + 0.112(2y_1 - y_3),\end{aligned} \tag{19.43}$$

这可以与 (19.33) 比较. 标准差为

$$\begin{aligned} x_1 &= \hat{x}_1 \pm 0.0149, \\ x_2 &= \hat{x}_2 \pm 0.020. \end{aligned} \tag{19.44}$$

因为改进的测量涉及 e^2，但仅涉及 m 的一次幂，所以 $e(x_1)$ 的精度的改进大约是 $m(x_2)$ 的两倍.

练习 19.2 针对满足 $N \geqslant n$ 的一般的 N 和 n 问题编写一个计算机程序，并在刚才解决的问题上进行测试. 估计编译后的程序解决欧拉问题所需要的时间.

在上述情况下，我们假定权重 w_i 根据先验信息已知. 如果不是这样，关于它们将会有许多种可能的部分先验信息，这将导致许多不同的先验概率分配 $p(w_1 \cdots w_n|I)$. 这也将导致一些细节上的小的定量改变，但是没有原则上的问题，只需要遵循贝叶斯原则就可以直接进行数学推广.

19.7 评注

悖论

通过研究这个问题，我们可以学到更多东西. 例如，让我们注意一些一开始可能令人感到惊讶的东西. 如果你研究根据 3 个测量值得出 m 的最优估计 (19.32)，将看到 y_3（e^2/m 的测量结果）以不同于 y_1 和 y_2 的方式进入到公式中. 它一次具有正系数，一次具有负系数. 如果 $w_1 = w_2$，则这两个系数相等，(19.32) 简化为

$$\hat{x}_2 = y_1 - y_2. \tag{19.45}$$

现在充分理解一下这一点的意义：它是说我们在估计 m 时使用 e^2/m 测量值的唯一原因是 e 和 e/m 的测量值具有不同的精度. 无论我们如何准确地知道 e^2/m，如果 e 和 e/m 的测量碰巧具有相同的精度（不管其精度多差），那么就应该忽略高精度的测量 e^2/m，而仅根据 e 和 e/m 估计 m.

我们认为，在开始听到这个结论时，你的直觉会反对它，第一反应可能是 (19.32) 中肯定有错误. 因此，请在有时间时仔细检查这一推导过程. 这是概率论几乎不费吹灰之力就能得出结果，但是我们仅凭无辅助的常识思考多年可能都不会得到同样结果的完美例子. 我们不会剥夺你自己解决这个“悖论”，并向朋友解释一致归纳推理如何要求你抛弃原初的最优测量值的乐趣.

在第 17 章中，我们曾经抱怨正统统计学家有时会丢弃相关数据，以使问题适应他们预想的“独立随机误差”模型. 我们现在是否也犯有同样的错误呢？毫

无疑问，情况看起来非常像是这样. 但是我们其实是无辜的：如果我们具有相同精度的 e 和 e/m 的测量值，e^2/m 的值实际上与 m 的推断无关. 为了明白这一点，假设我们从一开始就精确地知道 e^2/m. 在这个问题上，你将如何利用这一信息？如果你尝试一下，就会知道为什么 e^2/m 不相关. 但是要解决问题，请尝试完成以下练习.

练习 19.3　考虑一种特定情况：$w_1 = w_2 = 1, w_3 = 100$，第三个测量的精度是前两个的 10 倍. 但是，如果问题的情况是当我们尝试使用 (19.22) 中所有三个测量值时，第三个测量值会消去，那么我们使用准确的第三个测量值的唯一方法似乎是丢弃第一个或第二个测量值. 请证明：尽管如此，在这种情况下，(19.32) 仅使用第一个和第二个测量值所做的估计，比使用第一个和第三个，或者第二个和第三个测量值所做的估计更准确. 直观地解释为什么会是这样，这里面没有悖论.

作为另一个例子，重要的是，我们要理解为什么结论依赖于误差函数 δ_i 的损失函数和概率分布的选择. 如果我们使用的不是高斯分布 (19.12)，而是更宽尾的分布，例如柯西分布 $p(\delta) \propto (1 + w\delta^2/2)^{-1}$，那么后验分布 $p(x_1x_2|y_1y_2y_3)$ 在 (x_1, x_2) 平面上可能有多个峰. 于是，二次损失函数，或者更一般地，任何凹损失函数（即误差翻倍将使损失增加一倍以上的函数）将使得人们对 x_1 和 x_2 的估计位于非常不可能的两个峰值之间. 如果我们有凸损失函数，就会出现不同的“悖论”：构造最优估计量的基本方程 (19.26) 可能有多个解，但是没有任何信息告诉我们要使用哪个解.

这些情况的出现是机器人告诉我们如下结论的方式：我们对 x_1 和 x_2 的了解过于复杂，无法通过给出最优估计和可能误差来充分描述. 描述我们所知的唯一诚实的方法是给出实际分布 $p(x_1x_2|y_1y_2y_3)$. 这是决策论的局限之一，我们需要了解才能正确使用它.

第 20 章　模型比较

如无必要，勿增实体.

——奥卡姆的威廉（William of Ockham，约 1330）

我们已经比较详细地了解了如何在预先指定模型（表示被观测现象的一些假设）的条件下进行推断——检验假设、估计参数、预测未来观测结果等. 但是科学家一定还关心一个更重要的问题：当两个模型似乎都能解释事实时，如何在它们之间做出选择. 事实上，科学的进步需要对可能的不同模型进行比较，任何数量的新数据都无法消除模型中内置的从未被质疑的错误前提.

大致说来，这个问题并不新鲜. 大约 650 年前，方济会修士奥卡姆的威廉就察觉到思维投射谬误的逻辑错误.[①] 这使他教导说：宗教中的一些问题可以通过理性解决，而其他问题只能诉诸信仰. 他将后者置于自己的讨论范围之外，并专注于可以应用理性的问题——正如贝叶斯主义者今天在抛弃正统的思维投射谬误（例如在从未进行过的实验中存在极限频率的断言）时尝试做的，专注于现实世界中有意义的事情. 他所说的“只能诉诸信仰”的命题，大致对应于非亚里士多德命题. 本章开头引用的他的名言，通常称为“奥卡姆剃刀”，代表了他想要的一个推理原则. 这个原则我们今天仍然需要，但它也非常微妙，只能通过现代贝叶斯分析才能够很好理解.

当然，从当前的角度来看，这与我们在第 4 章已经考虑过的复合假设检验明显是同一个问题. 这里，我们只需要对原来的处理进行推广并得出更多细节. 这样，我们可以看到，传统的显著性检验只是在指定的备择假设类中选择最优假设的模型比较问题.

但还需要注意另外一个方面. 只要我们在单个模型中工作，归一化常数往往会相互抵消，因此在大多数情况下根本不需要引入. 但是当两个不同模型出现在同一个方程中时，归一化常数通常不会抵消，所有概率都必须正确归一化.

---

① 奥卡姆的观点，用他那个时代的语言表述是“实在只存在于个体中，共相只是抽象符号”；翻译成 20 世纪的语言是“心灵的抽象创造不是外部世界的实在”. 对他来说不幸的是，那个时代被正统神学视为珍宝的“实在”正是他否认为实在的东西，所以这让他在当权派中陷入了麻烦. 显然，奥卡姆是现代贝叶斯学派的先驱，这一切对他来说非常熟悉.

20.1　问题表述

要了解归一化常数为什么不再抵消，让我们首先回顾一下贝叶斯定理告诉我们的关于参数估计的内容. 模型 M 包含由 θ 统一表示的各种参数. 给定数据 D 和先验信息 I，为了估计其参数，我们首先应用贝叶斯定理：

$$p(\theta|DMI) = p(\theta|MI)\frac{p(D|\theta MI)}{p(D|MI)}, \tag{20.1}$$

其中右侧 M 的存在表示我们假设模型 M 的正确性. 分母作为归一化常数：

$$p(D|MI) = \int \mathrm{d}\theta\, p(D\theta|MI) = \int \mathrm{d}\theta\, p(D|\theta MI)p(\theta|MI), \tag{20.2}$$

我们看到这是似然 $L(\theta) = p(D|\theta MI)$ 的先验期望，即它是对参数的先验概率分布 $p(\theta|MI)$ 的期望.

现在我们问一个更高层次的问题：根据先验信息和数据，判断一组不同模型 $\{M_1, \cdots, M_r\}$ 中哪个最有可能是正确的. 贝叶斯定理给出第 j 个模型的后验概率为

$$p(M_j|DI) = p(M_j|I)\frac{p(D|M_jI)}{p(D|I)}, \quad 1 \leqslant j \leqslant r. \tag{20.3}$$

但是我们可以像第 4 章那样通过计算优势比来消除分母 $p(D|I)$. 模型 M_j 对 M_k 的后验优势比是

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)}\frac{p(D|M_jI)}{p(D|M_kI)}, \tag{20.4}$$

我们看到，概率 $p(D|M_jI)$ 在单模型参数估计问题 (20.1) 中仅作为归一化常数出现，现在作为确定模型 M_j 相对其他模型状态的基本量.[①] 数据告诉我们的这方面的准确度量总是它的似然函数对于任何参数 θ_j 在这个模型中的先验概率 $p(\theta_j|M_jI)$（它们对于不同模型通常是不同的）的期望. 在这里概率必须正确归一化，否则就违背了我们的基本法则，(20.4) 中的优势比就是任意的.

直观地说，数据偏爱的是给观察数据分配最高概率，因而最好地“解释数据”的模型. 在更高层次上，这只是在模型中做参数估计的似然原理.

但是奥卡姆原则如何能从中产生呢？第一个困难是奥卡姆原则从来没有用定义明确的术语来表述. 后来的作者们几乎普遍地将我们的开篇引语解释为：选择

① 这种逻辑结构甚至比贝叶斯形式体系更一般化，它存在于纯粹的最大熵形式体系中. 在统计力学中，两种不同相（例如液态和固态）的相对概率 P_j/P_k 是它们的分拆函数 Z_j/Z_k 的比. 每个分拆函数是每一种相中预测子问题的归一化常数. 在贝叶斯分析中，当两个模型的归一化常数相等时，数据对于两个模型是没有区分能力的；在统计力学中，相变温度是两个分拆函数变得相等时的温度. 在贝叶斯分析中，我们通常更喜欢用对数几率形式表达 (20.4)；在化学热力学中，一个世纪以来一直习惯于将相位无差别条件表述为“自由能” $F_j \propto \ln(Z_j)$ 相等. 这说明了贝叶斯和最大熵推理之间的根本统一性，尽管它们由于所处理的信息种类不同而存在表面上的差异.

竞争模型的标准是模型的“简单性”，尽管尚不清楚奥卡姆本人是否使用过这一表达方式. 也许可以将开篇引语重新描述为“如果无助于推断效果不要引入细节”，但是几个世纪以来的哲学讨论并没有对“简单”的意义做出明显的澄清.[①] 我们认为，将注意力集中在未定义术语上将阻碍我们对核心要点的理解. 这其实只是因为具有未指定参数的模型是复合假设，而不是简单假设，需要像第 4 章中那样对复合假设进行分析. 这样就出现了一些新特征，这些特征源于所考虑模型的参数空间的不同内部结构.

20.2 公正的法官与残酷的现实主义者

现在考虑在什么情况下需要进行模型比较的问题. 可能有以下两种态度. (1) 我们可以采取严格公正的法官的态度，坚持公平地比较模型，要求每个模型都尽可能地提供最优性能. 这可以通过为每个模型提供其参数的最优先验概率来实现（类似地，在奥运会中我们会认为，当两名运动员中的其中一名生病或受伤时，以表现来评判他们是不公平的. 公平的裁判员希望在两人都发挥最好时进行比较）. (2) 我们可能认为有必要成为残酷的现实主义者，根据实际拥有的先验信息来评判每个模型. 也就是说，如果没有关于模型参数的最优先验信息，我们就会惩罚该模型，尽管这并不是模型本身的错误.

奥卡姆剃刀原则体现的是残酷现实主义者的态度，它只是将模型本身严格公平的比较——不管我们目前所能给出的先验概率如何——转化为现实的比较，坚持只考虑现在实际可能发生的事情. 虽然一名生病或受伤的运动员值得同情，但我们不能在明天的重要比赛中启用他. 同样，如果我们的先验信息会将其参数置于远离最大似然值的位置，那么潜在优越的模型可能无法使用. 当真正的结果至关重要时，我们不得不成为残酷的现实主义者.

20.2.1 参数预先已知

要明白这一点，首先假设没有内在参数空间，模型的参数预先已知（$\theta = \theta'$）. 这样模型实际上变成了一个简单假设而非复合假设，贝叶斯定理的简单形式适用. 这相当于分配了先验 $p(\theta_j|M_jI) = \delta(\theta_j - \theta'_j)$，因此 (20.2) 简化为

$$p(D|M_jI) = p(D|\theta'_jM_jI) = L_j(\theta'_j), \tag{20.5}$$

这正是第 j 个模型中 θ'_j 的似然. 显然，如果 θ'_j 恰好等于模型和数据的最大似然估计 $\hat{\theta}_j$，严格公正的法官将注意到 (20.5) 是最大值. 这样，后验优势比 (20.4) 将

① 有一段时间，简单的概念因为似乎无法定义而被放弃了. 罗森克兰茨 (Rosenkrantz, 1977) 描述了这些冗长的细节.

变为

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)(L_j)_{\max}}{p(M_k|I)(L_k)_{\max}}. \tag{20.6}$$

这种极端情况虽然在上述意义上是公平的，但是可能非常不现实. 参数通常是未知的，并且在可能进行有用推断的“讲推理”的问题中，我们关于参数的先验信息必须足够好以进行有用的推断.

我们在前面的章节中已经看到，如果有一定数量的数据，大多数模型会给出尖锐的似然函数，以至于先验对于参数的推断相对不重要. 但是先验对于模型的推断仍然很重要，因此由先验定义的奥卡姆因子在模型比较中仍然很重要. 费希尔研究的简单生物学问题通常属于这种类型.

当先验信息对于参数的推断也很重要时——无论是由于宽泛模型还是稀疏数据——奥卡姆因子在我们的模型比较中有着至关重要的作用. 在杰弗里斯研究的问题以及现代科学家和经济学家所面临的更复杂的问题中，忽略这些因子将会带来危险.

20.2.2　参数未知

假设模型 M 具有参数 $\theta \equiv \{\theta_1, \cdots, \theta_m\}$. 比较 (20.4) 和 (20.6)，我们写下

$$p(D|MI) = L_{\max} W, \tag{20.7}$$

这定义了奥卡姆因子 W，它只是模型 M 受到非最优先验信息的惩罚量. 明确写出来是

$$W \equiv \int \mathrm{d}\theta \, \frac{L(\theta)}{L_{\max}} p(\theta|MI). \tag{20.8}$$

如果像在费希尔问题中那样，数据比先验信息更能提供有关 θ 的信息，那么似然函数就是锐峰的，我们可以定义一个“高似然区域” Ω' 作为整个参数空间 Ω 中包含指定似然积分值（例如 95%）的最小子区域. 那么大部分积分 (20.8) 的贡献将来自区域 Ω'. 更好的方法是，可以通过条件积分似然刚好是

$$\int \mathrm{d}\theta \, L(\theta) = L_{\max} V(\Omega') \tag{20.9}$$

来定义体积 $V(\Omega')$ 以消除任意数字 0.95. 那么 Ω' 可以定义为包含最大积分似然的体积 $V(\Omega')$ 区域. 也就是说，在 Ω' 内部似然处处大于 Ω' 边界上的某个阈值 L_0.

如果先验密度 $p(\theta|MI)$ 非常宽，以至于它在最大似然点周围的高似然区域基本上是常数，那么 (20.8) 可以简化为

$$W \simeq V(\Omega') p(\hat{\theta}|MI), \tag{20.10}$$

所以在这种情况下，奥卡姆因子从本质上说只是数据挑选出的高似然区域 Ω' 中包含的先验概率量.

无论在哪种情况下，我们的基本模型比较规则 (20.4) 都将变为

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)} \frac{(L_j)_{\max}}{(L_k)_{\max}} \frac{W_j}{W_k}, \tag{20.11}$$

通过与 (20.6) 比较，我们看到，在模型的内部参数空间中产生了额外的奥卡姆因子 W_j/W_k. 在 (20.11) 中，似然因子只取决于数据和模型. 如果两个不同的模型有相同的似然 $(L_j)_{\max}$，那么它们能够同样好地解释数据，在正统理论中，我们似乎无法在它们之间做出选择. 然而，贝叶斯定理告诉我们，还有另一种因素需要考虑：正统理论所忽略的先验信息，仍然可以为模型的优劣提供强有力的判断依据. 事实上，(20.11) 中的奥卡姆因子可能差别非常大，以至于会逆转 (20.6) 中的似然判断.

20.3 简单性概念何在?

关系 (20.11) 有着简单直觉无法（或者至少没有）看到的含义. 如果数据与先验信息相比信息量很大，那么两个模型的相对优劣由两个因素决定：

(1) 在它们各自的参数空间 Ω_j, Ω_k 上可以达到多高的似然；

(2) 有多少先验概率集中在它们各自的高似然区域 Ω'_j, Ω'_k 中.

但这两个因素似乎都与简单性的直观概念无关（对我们大多数人来说，简单性似乎是指在定义模型时不同假设的数量——例如，引入的不同参数的数量）.

为了理解这一点，让我们问："如何凭直觉决定如何选择模型?"在观察到一些事实后，我们更倾向于其中一种解释的真正原因是什么? 假设两个解释 A 和 B 都可以很好地解释某些已证实的历史事实. 但是解释 A 做了 4 种假设，每种假设都非常合理；而解释 B 只做了 2 种假设，但它们似乎都很牵强，极不可能是真的. 在这种情况下，每位历史学家都会毫不犹豫地选择解释 A，尽管解释 B 在直觉上更简单. 因此，我们的直觉从根本上问的不是假设有多简单，而是它们有多合情.

当然，合情性和简单性之间也存在着松散的联系，因为可能的假设集合越复杂，某个特定假设的替代假设的流形就越大，因此集合中任一特定假设的先验概率一定越小.

现在我们明白了为什么"简单性"永远无法给出令人满意的定义（即一个以令人满意的方式解释推断过程的定义）. 这是一个选择不当的词，它将人们的注意力从推断考虑的重要因素中转移开. 但是，人们几个世纪以来毫无批判地接受

了“简单性”的概念，它变得不可动摇，以至于一些人即使在应用贝叶斯定理之后，仍然顽固地试图通过简单性解释贝叶斯定理.[①]

一代代作者模糊地认为“简单假设更加合情”而没有给出任何合乎逻辑的理由. 我们建议纠正这种观念：应该说“合情的假设往往更简单”. 更简单的假设是具有更少同样合理替代者的假设.

在正统统计理论的范围内，这一切都无法理解. 其理念不允许存在模型或未知固定参数概率的观念，因为它们不是“随机变量”. 正统统计试图完全根据抽样分布来比较模型，而不考虑模型的简单性或先验信息！但是连这一点甚至都无法做到，因为模型中的所有参数都变成了冗余参数，而同样的理念拒绝任何处理这些参数的方法.[②] 因此，正统统计在这一问题上完全失效——它甚至没有提供可以描述问题的词汇——这在 20 世纪的大部分时间里阻碍着这一领域的进一步发展.

值得注意的是，尽管这个问题在数学上微不足道，但这种观念未曾使得几代有数学才能的工作者明白. 一旦明白这一点，直觉上显而易见的是：单纯的“简单性”无法成为评判模型的准则. 这只会再次提醒我们：人类大脑是一种不完美的推理装置，虽然它很擅长得出合理的结论，但往往无法给出令人信服的理由. 为此，我们确实需要作为逻辑的概率论的帮助.

当然，贝叶斯定理确实承认简单性是推断的组成部分. 但是这是通过什么机制起作用的呢？尽管贝叶斯定理对于任何问题总是能给出正确答案，但它通常是以非常有效的方式做到这一点的，以至于我们常常因不太了解这如何发生而感到困惑. 当前问题就是一个很好的例子，所以让我们试着从直觉上更好地理解这种情况.

用 M_n 表示模型，其中 $\theta = \{\theta_1, \cdots, \theta_n\}$ 是定义在参数空间 Ω_n 的 n 维参数. 现在通过添加新参数 θ_{n+1} 并转到新参数空间 Ω_{n+1} 来引入新模型 M_{n+1}，这样 $\theta_{n+1} = 0$ 表示旧模型 M_n. 我们将对这种情况进行计算，但首先来一般地考虑这个问题.

在子空间 Ω_n 上，模型的这种变化不会改变似然：$p(D|\theta M_{n+1} I) = p(D|\theta M_n I)$. 但是先验概率 $p(\theta|M_{n+1}I)$ 现在分布在比以前更大的参数空间上，并且一般来说相对于旧模型会对 Ω_n 的邻域点 Ω' 分配一个较低的概率.

对于一个有着合理信息量的实验，我们预计似然会集中在小的子区域 $\Omega'_n \in \Omega_n$ 和 $\Omega'_{n+1} \in \Omega_{n+1}$ 中. 因此，如果对于 M_{n+1}，最大似然点出现在或接近 $\theta_{n+1} = 0$ 处，那么 Ω'_{n+1} 将被分配比模型 M_n 的 Ω'_n 更小的先验概率，我们将有 $p(D|M_n I) >$

① 确实，对一位作者而言，奥卡姆剃刀根据定义与简单性有关，因为贝叶斯分析没有展示简单性而拒绝它！

② 普拉特（Pratt，1961）很久以前就对正统假设检验理论提出了批评.

$p(D|M_{n+1}I)$，数据生成的似然比将倾向于选择 M_n. 这就是奥卡姆现象.

因此，如果旧模型已经足够灵活，能很好地解释数据，那么贝叶斯定理一般将像奥卡姆原则一样更倾向于旧模型. 如果我们所说的“更简单”是指模型的参数空间更小，从而将我们限制在更小的可能抽样分布范围内，这在直觉上就是更简单的. 通常，只有当最大似然点远离 $\theta_{n+1}=0$（即显著性检验表明需要新参数）时，由于似然在 Ω_n' 比在 Ω_{n+1}' 上要小得多，会补偿后者较小的先验概率，不等号的方向才会改变. 如前所述，奥卡姆不会不同意这一点.

但是直觉根本没有定量告诉我们，这种似然的差异必须多大才能达到模型之间无差别的点. 此外，在明白这一原理后，由于贝叶斯定理会考虑到奥卡姆原则中无法想象的更多情况，很容易构造贝叶斯定理和奥卡姆原则相矛盾的情况（例如，新参数的引入伴随着旧子空间 S_n 上先验概率的重新分布）. 所以我们需要具体的计算来使这些东西量化.

20.4 示例：线性响应模型

现在我们做简单分析以说明上述结论，并计算似然和奥卡姆因子的确定值. 我们的场景是：数据集 $D\equiv\{(x_1,y_1),\cdots,(x_n,y_n)\}$ 由 (x,y) 的 n 个测量值组成. 尽管并非必要，我们还是可以将 x 视为“原因”，将 y 视为“结果”. 对于以下一般关系，“自变量” x_i 不需要指标 i 均匀分布或者单调增加. 根据这些数据和先验信息，我们需要在两种可能的生成数据的模型之间做出选择. 对于模型 M_1，除了不规则测量误差 e_i 外，响应对于原因是线性的：

$$M_1:\ y_i=\alpha x_i+e_i,\quad 1\leqslant i\leqslant n. \tag{20.12}$$

而对于模型 M_2，还有一个二次项：

$$M_2:\ y_i=\alpha x_i+\beta x_i^2+e_i, \tag{20.13}$$

如果 β 为负，这表示一种初始饱和或稳定性效应（如果 β 为正，则表示初始不稳定性效应）. 为了具体化，我们可以认为 x_i 是给第 i 名患者某一药物的剂量，y_i 是由此导致的血压升高值. 然后我们试图确定对这种药物剂量的响应是线性的还是二次的. 但是这个数学模型同样适用于许多不同的场景.[①] 只要模型是正确的，我们假设 x_i 的测量误差可以忽略不计，但是 y_i 的测量误差对于任何模型都是相

① 例如，x_i 可能是第 i 年空气中的臭氧含量，y_i 是该年的平均温度. 或者，x_i 可能是第 i 只加拿大鼠摄入的某种食物添加剂的量，y_i 是它身体内的癌组织的量. 或者，x_i 可能是第 i 年德国北部的酸雨量，y_i 是那一年死亡的松树数量；等等. 换句话说，我们现在处于前言中提到的所谓“线性响应模型”的领域，这些计算的结果直接关系到许多当前有争议的健康与环境问题的答案. 当然，大多数实际问题需要更复杂的模型，但是在明白这里的简单计算之后，我们将清楚如何以不同的方式进行推广.

同的，因此我们给它们分配联合抽样分布：

$$p(e_1,\cdots,e_n|I)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{e_i^2}{2\sigma^2}\right\}=\left(\frac{w}{2\pi}\right)^{n/2}\exp\left\{-\frac{w}{2}\sum_i e_i^2\right\},\quad(20.14)$$

其中 $w\equiv 1/\sigma^2$ 是“权重”参数，比 σ^2 更方便计算. 这个简单场景的优点是所有计算都可以手动完成，因此最终结果在任意极端条件下都是正确的，我们也可以看到哪些极限操作表现良好，哪些不好.

离题：又一次说明

如第 7 章中所讨论的那样，我们想重申其含义. 在正统统计中，抽样分布总是被表述为似乎代表一种“客观”事实，即误差的频率分布. 但是我们怀疑是否有人在实际问题中有此类频率分布或极限频率分布存在的先验知识. 如何获得有关从未进行过的长期实验结果的信息呢？这是我们所抛弃的思维投射谬误的一部分.

我们认识到，抽样分布只是描述关于测量误差先验知识状态的一种手段. 参数 σ 表示我们预期误差的大小. 例如，先验信息 I 可能是在过去此类数据中观察到的可变性. 在物理实验中，它可能不是任何观察的结果，而是从统计力学原理中得出的设备在已知温度下的奈奎斯特噪声强度.

特别是，(20.14) 中相关性的缺失并不是断言真实数据中不存在相关性. 这只是承认我们没有这种相关性存在的先验知识，因此假设无论是正还是负的相关性都可能会损害或有助于推理的准确性. 从某种意义上说，我们只是坦诚地承认自己的无知. 但是从另一种意义上说，我们走的是最安全、最保守的途径：*无论相关性是否实际存在*，使用这种抽样分布都会产生合理结果. 但是如果我们知道任何此类相关性存在，就将能够通过包含相关性的抽样分布做出更好的推断（尽管未必会好很多）.

这样做的原因是，抽样分布中的相关性会告诉机器人样本向量空间中的某些区域比其他区域更有可能，尽管它们具有相同的均方误差 $\overline{e^2}$. 这样，数据中一般被视为噪声的某些细节可以被识别为模型中存在系统效应的进一步证据.

让我们回到问题本身. 模型 M_1 的抽样分布为

$$M_1:\ p(D|\alpha M_1)=\left(\frac{w}{2\pi}\right)^{n/2}\exp\left\{-\frac{nw}{2}Q_1(\alpha)\right\},\quad(20.15)$$

其二次型为

$$Q_1(\alpha)\equiv\frac{1}{n}\sum_{i=1}^{n}(y_i-\alpha x_i)^2=\overline{y^2}-2\alpha\overline{xy}+\alpha^2\overline{x^2},\quad(20.16)$$

其中上划线表示平均值. α 的最大似然估计可以根据 $\partial Q_1/\partial\alpha=0$ 得到，或者

$$\alpha=\hat{\alpha}\equiv\frac{\overline{xy}}{\overline{x}}.\quad(20.17)$$

在这种情况下，它也称为“普通最小二乘”估计. 假设权重 w 已知，模型 M_1 的似然 (20.15) 是

$$L_1(\alpha)=\left(\frac{w}{2\pi}\right)^{n/2}\exp\left\{-\frac{nw}{2}\left[\overline{y^2}+\overline{x^2}(\alpha-\hat{\alpha})^2-\overline{x^2}\hat{\alpha}^2\right]\right\}, \tag{20.18}$$

其中可以丢弃任何与 α 无关的因子，但它将在 (20.23) 中自行消失. 如果我们只是从数据中估计 α，结果将是

$$(\alpha)_{\text{est}}=\hat{\alpha}\pm\frac{1}{\sqrt{nw\overline{x^2}}}=\hat{\alpha}\pm\frac{1}{\sqrt{n}}\frac{\sigma}{x_{\text{rms}}}=\hat{\alpha}\pm\delta\alpha, \tag{20.19}$$

其中 $x_{\text{rms}}=\sqrt{\overline{x^2}}$ 是 x_i 的均方根值. 因此，可以粗略认为高似然区域 Ω' 的体积（在这种情况下是宽度）是 $V(\Omega')=2(\delta\alpha)$.

现在，使用 (20.17) 可得 (20.3) 中模型 M_1 的“全局”抽样分布包含两个因子：

$$p(D|M_iI)=\int \mathrm{d}\alpha\, p(D|\alpha M_1)p(\alpha|M_1I)=L_{\max}(M_1)W_1, \tag{20.20}$$

其中

$$L_{\max}(M_1)=L_1(\hat{\alpha}), \tag{20.21}$$

因此模型 M_1 的奥卡姆因子是

$$W_1=\int \mathrm{d}\alpha\,\frac{L_1(\alpha)}{L_1(\hat{\alpha})}p(\alpha|M_1I). \tag{20.22}$$

我们发现似然比是

$$\frac{L_1(\alpha)}{L_1(\hat{\alpha})}=\exp\left[-\frac{nw\overline{x^2}}{2}(\alpha-\hat{\alpha})^2\right]. \tag{20.23}$$

因为似然比不能超过 1，而先验是已经归一化的，显然 $W_1\leqslant 1$.

现在必须为 α 分配先验. 我们通常会有一些理由（例如根据以前对此类问题的经验）来猜测某一数量级的值 α_0. 我们对这种猜测的准确性毫无信心，除了认为 $|\alpha-\alpha_0|$ 不能非常大之外（否则我们就不可能关心这个问题），但是我们很少有更多具体的先验信息. 我们可以通过分配归一化的先验密度

$$p(\alpha|M_1I)=\sqrt{\frac{w_0}{2\pi}}\exp\left\{-\frac{w_0}{2}(\alpha-\alpha_0)^2\right\} \tag{20.24}$$

来表明这一点，这表明我们认为 $|\alpha-\alpha_0|$ 不太可能远大于 $\sigma_0=1/\sqrt{w_0}$. 根据第 7 章讨论的中心极限定理以及第 11 章讨论的最大熵原理，这种高斯函数形式的先验原则上比所有其他形式都更受欢迎，因为它代表了我们在几乎所有实际问题中的实际知识状态. 然后幸运的是，这种形式也让我们可以对 (20.22) 准确积分，结果是

$$W_1=\sqrt{\frac{w_0}{nw\overline{x^2}+w_0}}\exp\left\{-\frac{nw\overline{x^2}w_0}{2(nw\overline{x^2}+w_0)}(\hat{\alpha}-\alpha_0)^2\right\}. \tag{20.25}$$

通过高似然区域的半宽 $\delta\alpha = 1/\sqrt{nw\overline{x^2}}$ 和 α 先验的半宽 $\sigma_0 = 1/\sqrt{w_0}$ 重写此式，它变成

$$W_1 = \frac{1}{\sqrt{1+(\sigma_0/\delta\alpha)^2}} \exp\left\{-\frac{(\hat{\alpha}-\alpha_0)^2}{2\sigma_0^2}\right\}. \tag{20.26}$$

这有几种极限形式. 如果先验估计 α_0 恰好等于普通最小二乘估计 $\hat{\alpha}$，它简化为

$$W_1 = \frac{1}{\sqrt{1+(\sigma_0/\delta\alpha)^2}}. \tag{20.27}$$

这样，如果 $\sigma_0 \gg \delta\alpha$，我们有

$$W_1 \simeq \frac{\delta\alpha}{\sigma_0}, \tag{20.28}$$

这实际上只是高似然区域中包含的先验概率量. 在这种情况下，奥卡姆因子是参数空间被数据信息收缩的比率，它表示我们的先验信息的模糊性在多大程度上可以通过将先验概率置于高似然区域之外而使模型 M_1 的性能变差. 如果先验估计 α_0 与普通最小二乘估计 $\hat{\alpha}$ 的差异小于 σ_0，这个结论也大致正确.

如果在 (20.27) 中 $\sigma_0 \to 0$，我们就有 $W_1 \to 1$，趋于最大可能值. 如果先验信息已经准确告诉了我们根据数据得出的普通最小二乘估计值，且没有误差，那么模型 W_1 根本不会受到惩罚. 但在所有其他情况下，都会有一些惩罚. 举个例子，如果 $|\alpha_0 - \hat{\alpha}| \gg \sigma_0$，则数据与先验信息强烈矛盾，模型将受到严重惩罚.

对于模型 M_2，抽样分布仍由 (20.15) 给出，但现在具有二次型

$$Q_2(\alpha,\beta) \equiv \frac{1}{n}\sum\left(y_i - \alpha x_i - \beta x_i^2\right)^2 = \overline{y^2} + \alpha^2\overline{x^2} + \beta^2\overline{x^4} - 2\alpha\overline{xy} - 2\beta\overline{x^2y} + 2\alpha\overline{x^3}, \tag{20.29}$$

最大似然估计 $(\hat{\alpha},\hat{\beta})$ 现在是方程组 $\partial Q_2/\partial\alpha = 0$, $\partial Q_2/\partial\beta = 0$ 或者

$$\begin{aligned} \overline{x^2}\hat{\alpha} + \overline{x^3}\hat{\beta} &= \overline{xy}, \\ \overline{x^3}\hat{\alpha} + \overline{x^4}\hat{\beta} &= \overline{x^2y}, \end{aligned} \tag{20.30}$$

的根，其解是

$$\hat{\alpha} = \frac{(\overline{x^4})(\overline{xy}) - (\overline{x^3})(\overline{x^2y})}{(\overline{x^2})(\overline{x^4}) - (\overline{x^3})^2}, \qquad \hat{\beta} = \frac{(\overline{x^2})(\overline{x^2y}) - (\overline{x^3})(\overline{xy})}{(\overline{x^2})(\overline{x^4}) - (\overline{x^3})^2}, \tag{20.31}$$

我们注意到，随着 $\overline{x^3} \to 0$，这变成估计

$$\hat{\alpha} \to \frac{\overline{xy}}{\overline{x^2}}, \qquad \hat{\beta} \to \frac{\overline{x^2y}}{\overline{x^4}}, \tag{20.32}$$

其中 $\hat{\alpha}$ 是使用模型 M_1 (20.17) 找到的普通最小二乘估计. 现在，如同 (20.22)，模型 M_2 的奥卡姆因子是

$$W_2 = \int \mathrm{d}\alpha \int \mathrm{d}\beta \, \frac{L_2(\alpha,\beta)}{L_2(\hat{\alpha},\hat{\beta})} p(\alpha\beta|M_2 I), \tag{20.33}$$

经过一些相当冗长的代数运算，我们发现似然比只是一个熟悉的二次型：

$$\frac{L_2(\alpha,\beta)}{L_2(\hat{\alpha},\hat{\beta})}=\exp\left\{-\frac{nw}{2}Q(\alpha,\beta)\right\}, \tag{20.34}$$

其中

$$\begin{aligned}Q(\alpha,\beta)&\equiv Q_2(\alpha,\beta)-Q_2(\hat{\alpha},\hat{\beta})\\&=\overline{x^2}(\alpha-\hat{\alpha})^2+2\overline{x^3}(\alpha-\hat{\alpha})(\beta-\hat{\beta})+\overline{x^4}(\beta-\hat{\beta}).\end{aligned} \tag{20.35}$$

现在我们赋予联合先验

$$p(\alpha\beta|M_2I)=\sqrt{\frac{w_0}{2\pi}}\exp\left\{-\frac{w_0}{2}(\alpha-\alpha_0)^2\right\}\sqrt{\frac{w_1}{2\pi}}\exp\left\{-\frac{w_1}{2}(\beta-\beta_0)^2\right\}, \tag{20.36}$$

其中 w_0,α_0 与 (20.24) 中的相同，因此两个模型中 α 的边缘先验是相同的（否则在从 M_1 到 M_2 时将改变两种而不是一种条件，这将使结果很难解释）：

$$p(\alpha|M_1I)=p(\alpha|M_2I). \tag{20.37}$$

模型 M_2 的奥卡姆因子为

$$W_2=\frac{\sqrt{w_0w_1}}{2\pi}\int \mathrm{d}\alpha\int \mathrm{d}\beta\,\exp\left\{-\frac{1}{2}\left[nwQ(\alpha,\beta)+w_0(\alpha-\alpha_0)^2+w_1(\beta-\beta_0)^2\right]\right\}, \tag{20.38}$$

这一积分同样可以准确算出，结果是

$$W_2=\sqrt{\frac{w_0w_1}{(w_0+nw\overline{x^2})(w_1+nw\overline{x^4})}}\,\mathrm{e}^x. \tag{20.39}$$

新编练习 20.1 正如上面所写的，只有使用条件 $\overline{x^3}\to 0$ 时，(20.39) 中的分母才是正确的. 使用这一简化假设，推导出 W_2 并定义 x.

根据 (20.27) 和 (20.39) 计算倾向于 M_1 而非 M_2 的奥卡姆因子比是

$$\frac{W_1}{W_2}=\frac{1\Big/\sqrt{1+(\sigma_0/\delta\alpha)^2}}{\sqrt{w_0w_1/(w_0+nw\overline{x^2})(w_1+nw\overline{x^4})}\,\mathrm{e}^x}. \tag{20.40}$$

新编练习 20.2 根据半宽 $\delta\alpha=1\Big/\sqrt{nw\overline{x^2}}$, $\sigma_0=1\Big/\sqrt{w_0}$, $\delta\beta=1\Big/\sqrt{nw\overline{x^4}}$, $\sigma_1=1\Big/\sqrt{w_1}$ 重写 (20.40). 在什么情况下，模型 M_2 会比模型 M_1 更受青睐?

20.5 评注

实际的科学实践并不真正遵循奥卡姆剃刀原则，无论是在前面的“简单性”还是我们修订后的“合情性”形式中都是如此. 正如许多人感到痛惜的那样，迷人的

新假设或模型如此简洁、可信地解释了事实，以至于你想立刻相信它，但官方机构通常以一些单调乏味、复杂无趣的理由嗤之以鼻，或者甚至完全不提供其他选择. 科学的进步主要是由少数基本持不同意见的创新者推动的，例如哥白尼、伽利略、牛顿、拉普拉斯、达尔文、孟德尔、巴斯德、玻尔兹曼、爱因斯坦、魏格纳、杰弗里斯——他们都不得不经历这种最初的拒绝和攻击. 在伽利略、拉普拉斯和达尔文的事例中，这些攻击在他们死后持续了一个多世纪. 这并不是因为他们的新假设是错误的，而是因为这是科学社会学的一部分（实际上也是所有学术的一部分）. 在任何领域，当权派都很少追求真理，因为当权集团就是由那些真诚地相信自己已经拥有真理的人组成的.

此外，这也延缓了进步. 那些没有听从奥卡姆的威廉关于区分诉诸理性与诉诸信仰问题的教训的学者，注定过去而且现在仍然会一生都在胡说. 我们记录下了这种胡说过去最常见的形式.

终极原因

每一次关于科学推断的讨论似乎都迟早要面临对终极原因相信还是不相信的问题. 表现形式从雅克 · 莫诺（Jacques Monod，1970）禁止我们提及宇宙的目的，到宗教原教旨主义者坚持认为不相信该目的是邪恶的. 我们惊讶于那些宣扬相反观点的人表现出的教条与情绪化的强烈程度，而且他们没有丝毫提及支持其立场的事实证据.

但是几乎所有讨论过这个问题的人都认为，所谓“终极原因”意味着某种超自然力量可以中止自然法则并接管事件的控制权（即以与运动方程不同的方式改变分子位置与速度），以确保达到所需的某个最终条件. 在我们看来，几乎所有过去的讨论都存在缺陷，因为未能认识到终极原因的运行并不意味着需要控制分子细节.

当一本教科书的作者说“我写本书的目的是……”时，他就在表明有一个真正的“终极原因”支配着作者、笔、秘书、文字处理器等的许多活动，这通常会持续数年. 当一名化学家对他的实验系统强加条件，迫使它具有一定的体积和温度时，他就像终极原因的真正持有者一样，决定他希望具有的最终热力学状态. 瓦匠和厨师同样从事为特定目的援引终极原因的艺术. 但是几乎总是被忽略的是，这些终极原因是宏观的，它们不确定任何特定的“分子”细节. 在所有情况下，如果这些细节在数十亿种方式中的任何一种里存在不同，终极原因也会得到满足.

终极原因可以说具有熵，表示可以实现其目的的微观方式的数量. 熵越大，可能实现的概率就越大. 因此，最大熵原理也适用于此.

换句话说，虽然微观终极原因的想法与科学家的本能背道而驰，但是宏观终极原因是一种非常常见与真实的现象，我们每天都在援引它. 当我们所做的几乎所有事情都具有某个明确的目的时，我们几乎无法否认宇宙也存在目的. 的确，如果一个人在生活中不追求某种明确的长期目标，他的同事就会将其斥为游手好闲者. 这显然只是一个熟悉的事实，没有任何宗教含义——也没有反宗教含义. 每位科学家都相信宏观的终极原因，不相信超自然、违反物理定律的事情. 终极原因的持有者不是暂停物理定律，只是在选择某个系统根据物理定律演化的哈密顿量. 看不到这一点就会产生最不可思议、最神秘的胡说.

第 21 章　离群值与稳健性

每一名参与实际测量的人都很可能发现自己面临以下处境. 你尝试测量某个量 θ（可能是天狼星的赤经，π 介子的质量，100 公里深处的地震波速度，一种新有机化合物的熔点，消费者对苹果的需求弹性等），但是测量仪器或者数据获取流程总是不完善的. 因此，对 θ 进行 n 次独立测量后，你会得到 n 个不同的结果 $(x_1, \cdots, x_n)$. 你将如何报告对 θ 的了解呢？更具体地说，你应该报告怎样的“最优”估计，其准确性是多少？

如果这 n 个数据值紧密聚集在一起形成相当平滑的单峰直方图，你将接受前几章给出的解，可能会觉得即使没有概率论，从好数据中得出结论也不是很困难. 但是如果数据没有很好地聚集在一起：有一个值 x_j 与其他 $n-1$ 个值形成的良好聚类相距甚远，你将如何处理这个离群值[①]呢？它对你有理由得出的关于 θ 的结论有什么影响？

我们在第 4 章和第 5 章中已经看到，出乎意料的惊人数据可能如何导致“死假设复活”，似乎，类似这种东西可能在这里起作用. 事实上，任何出人意料的十分难看的数据都可能引发这个问题. 这里只考虑离群值的特殊情况，而将为其他类型的意外结构构建相应的理论留给读者作为练习.

21.1　实验者的困境

自 18 世纪以来，围绕数据中的离群值问题一直有热烈的讨论. 这一问题出现在天文学、大地测量学、量热学和许多其他测量中. 让我们将“仪器”广义地解释为获取数据的任何方法. 哲学上，关于离群值有两种截然相反的观点.

(I) 仪器一定发生了问题，离群值不是好数据的一部分，我们必须将其丢弃以免得出错误结论.

(II) 不能如此！仅仅因为数据离群就丢弃它是不诚实的. 这个离群值很可能是你拥有的最重要的数据，在你的数据分析中必须考虑到它，否则就是在随意“篡改”数据，不再具有科学客观性.

---

① 英文为 outlier，常见的翻译有“离群值”和“异常值”. “离群值”偏中性一些，只是说这个数据与众不同，不跟其他数据聚类在一起，但是未必就是坏的或异常的数据. 如果确定是坏的数据，那么翻译为“异常值”更合适些. 本章主要是对 outlier 做贝叶斯分析，主要说明 outlier 根据先验信息不同可能是正常（好）值或异常（坏）值. 为了更切合原意，一般翻译为离群值，而在确定指坏数据时翻译为异常值. ——译者注

从以上陈述中我们可以理解为什么这一问题会引起争议并且很难解决，其中不仅潜藏着一种正义的道德热情因素，同样清楚的是，这两种立场中都包含真理的成分. 这两者之间能调和吗?

在实用角度上，人们已经发明了几种随意的特定方法（例如一个世纪前天文学教科书中的肖维内准则）来决定何时拒绝离群值. 奇怪的是，这种随意的拒绝准则（如两个标准差等）似乎没有注意到以下方面. 我们认为它对于这一问题的任何理性解决方案都是至关重要的.

思考一下上面两个陈述，我们看到它们反映了关于测量仪器的不同先验信息. 这是在上述所有准则中都被忽略了的关键因素. 为了将此考虑在内，需要的不是更多的特定方法，而是直接的概率分析.

如果我们知道收集数据的方式不可靠，并且确实很可能在没有任何警告的情况下出错并给出错误数据，那么立场 (I) 似乎是合理的. 如果我们已经预料到这一点，那么离群值的出现似乎更有可能是由于“仪器错误”，而非真实效应. ①

另外，立场 (II) 对于一个对自己的仪器有绝对信心的人是合理的. 他确信他的电压表总是提供误差为 $\pm 0.5\%$ 的可靠读数，不可能出现 5% 的误差；或者相信他的望远镜在记录方向时的误差在 10 弧秒内，不可能差 1 度. 那么，离群值的出现无论多么出乎意料，都必须被视为重大事件，忽略它可能会错过一个重大发现.

但是 (I) 和 (II) 是极端的立场，真正的实验者几乎总是处于某种中间情况. 一方面，如果知道仪器非常不可靠，人们想必根本不会用它获取数据；但在生物学或经济学等领域，人们可能不得不使用大自然提供的任何“仪器”. 另一方面，很少有科学家——即使是在国家标准局最好的实验室里的科学家——会对他们的仪器如此有信心，以至于会武断地断言它绝对不会出错.

人们希望以明确的形式看到估计结果，例如 $(\theta)_{\text{est}} = A \pm B$，其中 A 和 B 是两个确定值，想必是数据 $D \equiv \{x_1, \cdots, x_n\}$ 的两个函数. 但是它们是哪两个函数呢? 当数据紧密聚集在一起时，将样本均值 $A = \overline{x} \equiv n^{-1} \sum x_i$ 作为估计值肯定是合情的猜测. 观察到的数据值 x_i 的离散度表明了测量的可重复性，人们可以认为这也表明了它们的准确性. 如果是这样，计算均值的均方偏差或样本方差似乎是合理的: $s^2 \equiv n^{-1} \sum (x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2$ 并选择 $B = s$，即样本标准差. 一个受过教育并且熟悉概率论基本结果的人可以凭直觉通过取 $B = s/\sqrt{n}$ 来改进这一结果，即使没有任何明确规定的准则表明它是最优的，结论

$$(\theta)_{\text{est}} = \overline{x} \pm \frac{s}{\sqrt{n}} \tag{21.1}$$

① 我们在 (6.97) 后面的讨论中看到过这种现象的另一个例子. 概率论告诉我们，如果能预期计数率的大波动是仪器的产物，那么观察到的波动对于估计光束强度的变化就变得不那么有说服力了.

不会在位置或准确性估计上被批评为非常不合理.

练习 21.1　我们在第 7 章中已经看到，在相当一般的条件下，高斯抽样分布 $p(x|\theta,\sigma) \propto \exp\{-(x-\theta)^2/2\sigma^2\}$ 将导致我们将数据均值 $\bar{x}$ 作为 θ 的点估计. 证明，任何具有圆顶的抽样分布（即 $p(x|\theta) = a_0 - a_1(x-\theta)^2 + \cdots$）都将在数据紧密聚集时的极限情况下产生相同的均值估计.

如果数据没有紧密聚集，上述讨论似乎只考虑了两种可能的操作：保留离群值并完全信任它，或者完全抛弃离群值. 有没有更合乎情理的中间立场?

21.2　稳健性

对于此类问题，最近出现了以胡贝尔（Huber，1981）为代表的另一种观点，这在第 6 章中已经做了简要说明. 它仍然试图通过基于直觉的特定流程来处理问题，不明确关注先验信息或概率论，但是确实在寻找一种中间立场. 人们寻求稳健的数据分析方法，这意味着对误差的具体抽样分布不敏感，或者通常描述为对模型不敏感，或者抗异的数据分析方法，也意味着一小部分数据中大的误差不会对结论有很大的影响.

笼统地说，一般思想是前几章使用的理论上的“最优性”，在实践中并不总是一个好的准则. 通常，我们不确定哪个模型是正确的，那么一种对各种不同模型有用的方法，尽管对任何模型都不是最优的，但是可能比完全适合一个特定模型却不适合其他模型的方法更可取.

显然，这种观点可能具有一些优点. 但是，以图基和莫斯特勒（例如 Tukey，1977）为代表的“稳健性/探索性”思想流派将其推向了反对所有最优性考虑的地步. 然而，试图不那么模糊地定义这一立场变得很麻烦. 给定数据 D 以及某个参数的任意两个估计量 $f(D)$ 和 $g(D)$，对于术语“稳健”或“抗异”是否有任何明确定义，使得一个估计量比另一个“更稳健”或“更抗异”有意义呢? 如果有，那么在给定的可能估计量集合 S 中，必然有“最优稳健”的估计量 $a(D)$ 和“最优抗异”的估计量 $b(D)$，它们未必是同一个. [①]

这里要指出的要点是，如果任何直观的性质（例如稳健性）被认为是需要的，那么一旦足够精确定义这一性质允许比较之后，就会遵循最优性原则. 因此不可能提出任何定义明确的推断性质而又拒绝最优性原则，这将缺乏一致性.

① “稳健/抗异估计量”一词是图基发明的. 我曾经向他说过，从字面上看，这意味着“一个拒绝稳健的估计量”，但他否认了这一点.