



统计数据分析

(from *Statistical Data Analysis*, by Glen Cowan)

YZW 译

北京大学高能物理研究中心

(内部教学使用)

目录

第一章 基本概念	1
1.1 概率与随机变量	1
1.2 概率的解释	3
1.2.1 概率的相对频率解释	4
1.2.2 主观概率	4
1.3 概率密度函数	5
1.4 随机变量的函数	10
1.5 期望值	12
1.6 误差传递	15
1.7 随机变量的正交变换	16
第二章 常用概率函数	19
2.1 二项分布和多项分布	19
2.2 泊松分布	21
2.3 均匀分布	22
2.4 指数分布	22
2.5 高斯分布	23
2.6 对数正态分布	25
2.7 卡方分布	26
2.8 柯西(布莱特-魏格纳)分布	27
2.9 朗道分布	27
第三章 蒙特卡罗方法	31
3.1 均匀分布的随机数	31
3.2 变量变换法	32
3.3 舍选法	32
3.4 蒙特卡罗方法的应用	34
第四章 统计检验	37
4.1 假设、检验统计量、显著性水平和功效	37
4.2 粒子选择的统计检验	38
4.3 用奈曼-皮尔逊引理选择拒绝域	40
4.4 构造检验统计量	40

4.4.1	线性检验统计量、费舍尔甄别函数	40
4.4.2	非线性检验统计量、神经网络	42
4.4.3	输入变量的选择	44
4.5	拟合优度检验	44
4.6	观测信号的显著性	46
4.7	皮尔逊卡方检验	47
第五章	参数估计的一般概念	51
5.1	样本、估计量、偏倚	51
5.2	均值、方差和协方差的估计量	52
第六章	极大似然法	55
6.1	极大似然估计量	55
6.2	极大似然估计量：指数分布	56
6.3	极大似然估计量：高斯分布的均值和方差	58
6.4	极大似然估计量的方差：解析方法	58
6.5	极大似然估计量的方差：蒙特卡罗方法	59
6.6	极大似然估计量的方差：RCF 边界方法	60
6.7	极大似然估计量的方差：图解法	61
6.8	双参数的极大似然估计	62
6.9	扩展的极大似然估计	64
6.10	分区间数据的极大似然估计	67
6.11	极大似然法的拟合优度检验	69
6.12	用极大似然法合并实验测量	70
6.13	极大似然与贝叶斯估计量的关系	71
第七章	最小二乘法	73
7.1	与极大似然的联系	73
7.2	线性最小二乘拟合	74
7.3	多项式的最小二乘拟合	75
7.4	分区间数据的最小二乘拟合	77
7.5	拟合优度最小二乘检验	79
7.6	用最小二乘法合并实验测量	81
7.6.1	存在相关性的测量数据的平均	83
7.6.2	求解协方差矩阵	85
第八章	矩方法	87
第九章	统计不确定度、置信区间和极限	91
9.1	标准差作为统计不确定度	91
9.2	经典置信区间 (严格方法)	91
9.3	高斯分布估计量的置信区间	95
9.4	泊松分布均值的置信区间	96

9.5	相关系数、参数变换的置信区间	98
9.6	用似然函数或卡方求置信区间	100
9.7	多维置信区域	101
9.8	物理边界的极限	103
9.9	含本底的泊松变量均值的上限	106
第十章	特征函数	109
10.1	特征函数的定义与性质	109
10.2	特征函数的应用	110
10.3	中心极限定理	111
10.4	用特征函数求估计量的概率密度函数	113
10.4.1	平均寿命和衰变常数的期望值	114
10.4.2	指数随机变量均值的置信区间	114
第十一章	解谱法	117
11.1	解谱问题的数学描述	117
11.2	响应矩阵求逆	121
11.3	修正因子法	124
11.4	正规化解谱法的一般策略	125
11.5	正规化函数	126
11.5.1	Tikhonov 正规化	126
11.5.2	基于熵的正规化函数	128
11.5.3	利用熵的贝叶斯动机	129
11.5.4	基于交叉熵的正规化函数	130
11.6	估计量的方差和偏倚	130
11.7	正规化参数的选择	133
11.8	解谱法举例	135
11.9	解谱法的数值实现	139
参考文献		143
索引		148

DRAFT ZV

第一章 基本概念

1.1 概率与随机变量

本书的主要目的是介绍统计数据分析中最重要的概念与方法。统计数据分析的核心概念是不确定性，不确定性的表现形式多种多样。例如，在重复实验中，我们经常会面临测量结果不确定且不可预言的情形。这种不确定的行为可能来自于与测量仪器有关的误差，也可能是所研究系统更深层次不可预言性的后果，例如量子力学的不可预言性。不确定性可能根源于各种未确定的因素，这些因素往往原则上可以确定但实际却无法做到。如果无法预言或者不能完全确定地预言某个系统，则称系统的这种特征为**随机性(random)**。

随机性的大小可以用**概率(probability)**的概念定量描述。概率的数学理论至少可以追溯到 17 世纪，并且已发展出多种不同的定义。我们将采用集合论语言给出的定义，即柯尔莫哥洛夫 (Kolmogorov) 于 1933 年首次提出的表述 [1]。考虑某集合 S (称为**样本空间(sample space)**) 包含一定数量的元素，并且暂不考虑这些元素的具体涵义。对 S 的任意子集 A ，可以指定一个称为概率的实数 $P(A)$ ，概率由以下三个公理定义：¹

1. 对于 S 的任意子集 A ， $P(A) \geq 0$ 。
2. 对于任意两个不相交的子集 A 和 B (即互斥， $A \cap B = \emptyset$)， A 与 B 并集的概率为两个概率之和，即 $P(A \cup B) = P(A) + P(B)$ 。
3. 整个样本空间的概率为 1，即 $P(S) = 1$ 。

利用这三个公理，可以推导出概率函数的更多性质，例如

$$\begin{aligned} P(\bar{A}) &= 1 - P(A), \text{ 其中 } \bar{A} \text{ 为 } A \text{ 的补集} \\ P(A \cup \bar{A}) &= 1 \\ 0 &\leq P(A) \leq 1 \\ P(\emptyset) &= 0 \\ \text{如果 } A &\subset B, \text{ 则 } P(A) \leq P(B) \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B). \end{aligned} \tag{1.1}$$

关于这些性质的证明以及其它性质，可参见文献 [2–4]。

如果一个变量对于集合 S 的每一个元素，都取相应的特定值，则称该变量为**随机变量(random variable)**。有时，可能每一个元素都需要用多个量来描述，这种情况下随机变量为多维矢量。

¹相比于某些更严格的教材 (例如 [2])，这里的公理有所简化，但对我们而言足矣。若要更准确地定义，被赋予概率的子集的集合必须构成所谓的 σ 域或事件域。

假设某样本空间 S 包含子集 A 和 B , 并且 $P(B) \neq 0$, 则可定义**条件概率(conditional probability)** $P(A|B)$ (读作给定 B 时 A 的概率 P),

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1.2)$$

图1.1给出了集合 A 、 B 和 S 的相互关系。容易证明, 条件概率本身也满足概率的三个公理。注意, 通常的概率 $P(A)$ 可以看作给定 S 时 A 的条件概率: $P(A) = P(A|S)$ 。

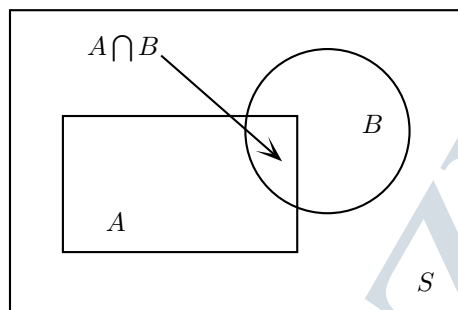


图 1.1: 条件概率定义中集合 A , B 和 S 的相互关系。

我们称两个子集 A 和 B 相互**独立(independent)**, 如果它们满足

$$P(A \cap B) = P(A)P(B). \quad (1.3)$$

对于相互独立的子集 A 和 B , 根据条件概率的定义, 它们满足 $P(A|B) = P(A)$ 以及 $P(B|A) = P(B)$ 。(不要把1.3式独立的子集与互斥的子集 $A \cap B = \emptyset$ 相混淆。如果两个子集互斥, 则它们必定不相互独立。)

从条件概率的定义, 我们还可以得到给定 A 时 B 的概率(假设 $P(A) \neq 0$),

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \quad (1.4)$$

由于 $A \cap B$ 与 $B \cap A$ 相同, 联立 (1.2) 与 (1.4) 式可以得到

$$P(B \cap A) = P(A|B)P(B) = P(B|A)P(A), \quad (1.5)$$

或者

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (1.6)$$

式 (1.6) 称为**贝叶斯定理(Bayes theorem)** [5]。

假设样本空间 S 可以划分成若干两两互斥的子集 A_i , 即 $S = \cup_i A_i$ 并且当 $i \neq j$ 时 $A_i \cap A_j = \emptyset$ 。进一步假设对所有 i 满足 $P(A_i) \neq 0$ 。显然, S 的任意子集 B 可以表示为 $B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$ 。由于 B 的这些子集 $B \cap A_i$ 两两互斥, 所以 B 的概率等于这些子集的概率相加。即,

$$\begin{aligned} P(B) &= P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) \\ &= \sum_i P(B|A_i)P(A_i). \end{aligned} \quad (1.7)$$

最后一行用到了 (1.4) 式条件概率的定义, 即对所有 i , 取 $A = A_i$, 利用条件概率计算概率 $P(B \cap A_i)$ 。式 (1.7) 称为**全概率定律(law of total probability)**。全概率定律有时很有用, 例如, 假设样本空间可以划分为

互斥的子集 A_i ，并且 A_i 的概率相对容易计算，那么结合全概率定律与贝叶斯定理(1.6)，在已知条件概率 $P(B|A)$ 的情况下，可以计算条件概率 $P(A|B)$ 。即

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}. \quad (1.8)$$

其中 A 可以是 S 的任意子集，例如某个 A_i 。

作为贝叶斯定理的例子，考虑已知在人群中患病率为 0.1% 的某种疾病，即患病或非患病的**验前概率**(也称先验概率)为

$$\begin{aligned} P(\text{患病}) &= 0.001, \\ P(\text{非患病}) &= 0.999. \end{aligned}$$

已知对于患病者，检验结果为阳性的概率为 98%，即

$$\begin{aligned} P(+|\text{患病}) &= 0.98, \\ P(-|\text{患病}) &= 0.02. \end{aligned}$$

然而，如果对非患病者进行相同的检验，也有 3% 的概率检验结果为阳性，即

$$\begin{aligned} P(+|\text{非患病}) &= 0.03, \\ P(-|\text{非患病}) &= 0.97. \end{aligned}$$

那么，如果某人检验结果为阳性，其患病的概率为多少？根据贝叶斯定理 (利用 (1.8) 式)，该条件概率为

$$\begin{aligned} P(\text{患病}|+) &= \frac{P(+|\text{患病})P(\text{患病})}{P(+|\text{患病})P(\text{患病}) + P(+|\text{非患病})P(\text{非患病})} \\ &= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \\ &= 0.032. \end{aligned}$$

可见，即使检验结果为阳性，其患病的概率仍然只有 3.2%。这看起来可能有些令人吃惊，因为对患病者检验结果错误的概率只有 0.2%，对非患病者检验错误的概率为 3%。但是，由于患病的验前概率非常低，仅为 0.1%，从而导致**验后概率**只有 3.2%。这里，我们尚未提及一个至关重要的问题：当我们说 $P(\text{患病}|+) = 0.032$ 时，到底指的是什么意思？换言之，应该如何准确解释这个概率。我们将在下一节仔细考察这个问题。

1.2 概率的解释

根据定义，任何满足上一节所列三个公理的函数都可称为概率函数。尽管如此，我们仍需明确如何解释样本空间中的元素，以及如何为之赋予概率值并加以解释。在数据分析中，常用的解释有两种，其中最重要的一种解释是**相对频率(relative frequency)**。一般来说，除了确定测量的统计不确定度，其它场合往往都使用相对频率的解释。另外一种常用的解释是**主观概率(subjective propability)**，例如，经常用主观概率定量给出系统不确定度。下面将详细介绍概率的这两种不同解释。

1.2.1 概率的相对频率解释

在数据分析中, 概率通常解释为**相对频率极限(limiting relative frequency)**。在这种解释下, 集合 S 的元素对应于测量的各种可能结果, 并且认为测量可重复 (至少可在假想中重复)。 S 的子集 A 对应于子集中任何一个结果的发生。这样的子集称为一个**事件(event)**, 如果测量结果在该子集中, 则称该事件发生。

如果 S 的子集只包含一个元素, 表示单一的**元输出(elementary outcome)**。元输出 A 的概率指定为: 当重复实验次数 n 趋于无穷时, A 发生的次数的比率, 即

$$P(A) = \lim_{n \rightarrow \infty} \frac{n \text{ 次测量中 } A \text{ 发生的次数}}{n}. \quad (1.9)$$

如果 A 不是元子集, 包含多个元输出, 那么, 这几个输出中任意一个发生的总概率就是 A 发生的概率, 可以根据概率公理的相加定律由单个输出的概率确定。这也符合事件发生的相对频率。

相对频率的解释与概率公理是一致的: 第一, 事件发生的比率总是大于等于零; 第二, 互斥子集的输出中任何一个发生的频率等于单个输出的频率之和; 第三, 测量必定会得到某种输出 (即 $P(S) = 1$)。根据相对频率解释, 条件概率 $P(A|B)$ 为 A 和 B 都发生的次数除以无论 A 是否发生而 B 发生的次数。即, $P(A|B)$ 是把子集 B 当作样本空间时 A 发生的频率。

显然, 基于这种模型的概率永远不可能从实验上精确确定。**经典统计(classical statistics)** 的基本任务是由有限的实验数据估计概率 (假设具有确定的值但该值未知), 并检验可以预言这一概率的特定模型或理论与观测数据的符合程度。

在研究物理规律时, 相对频率解释直接明了, 因为我们总是假设重复实验中物理规律具有相同的行为。指定的概率值的合理性可以通过重复实验予以检验, 这一观点在很多领域是合适的, 例如粒子物理。在粒子物理中, 重复的粒子碰撞就构成了重复实验。相对频率的概念在处理唯一现象时存在很大困难, 例如宇宙大爆炸。此时, 为了挽救频率解释, 我们可以想象存在大量类似的宇宙, 某种事件可以在一部分宇宙中发生。然而, 由于这种做法甚至原则上都无法实现, 这里的频率只能看作为表达对我们所在的单一宇宙的信心程度而进行的精神构想。

标准的概率统计教材中通常采用概率的频率解释, 例如 Fisher [6]、Stuart 和 Ord [7] 以及 Cramér [8] 的教材。von Mises [9, 10] 在他的书中对概率的频率解释作了一些哲学讨论。

1.2.2 主观概率

概率的另一种解释为**主观概率**, 也称作**贝叶斯概率(Bayesian probability)**。在这种解释下, 样本空间的元素对应于**假设(hypotheses)** 或者**命题 (propositions)**, 即要么正确要么错误的陈述。(在使用主观概率时, 样本空间经常被称为假设空间。) 人们把与假设关联的概率解释为可信性的量度, 即

$$P(A) = \text{假设 } A \text{ 为真的信心程度}. \quad (1.10)$$

在主观概率解释中, 构造的样本空间里的元假设必须互斥, 即只有一个假设为真。如果某个子集包含多个假设, 则当子集中任意一个假设为真时该子集就为真。也就是说, 集合的并集对应于布尔算符“或”, 集合的交集对应于布尔算符“与”。样本空间中必定存在某个假设为真, 即 $P(S) = 1$ 。

对于给定的测量结果其发生的频率是一定的, 我们可以把这一陈述看作一个假设。所以, 主观概率的框架包含了相对频率解释。然而, 主观概率还可以与未知常数的值联系起来, 它反映了我们对常数的数值处于某固定区间内的信心程度。未知常数的概率在频率解释中是没有意义的, 因为, 如果我们的实验依赖于准确值未知的物理参数 (例如电子质量), 在重复进行实验时, 参数值要么始终落在给定的固定区间, 要么永远处于区间之外, 相应的概率要么永远为 1, 要么永远为 0, 但我们不知道到底是哪种情况。然而, 在

主观概率中，我们可以给出“电子质量处于给定区间的概率为 95%”的论述，这个概率实际上反映了我们对问题的认识程度。

主观概率的应用与贝叶斯定理关系密切，它构成了贝叶斯统计(与经典统计相对)的基础。贝叶斯定理((1.6)式)中的子集 A 可以解释为某个理论为真的假设，子集 B 可以解释为实验产生某特定结果(即实验数据)的假设。于是贝叶斯定理可以表示成这样的形式

$$P(\text{理论}|\text{数据}) \propto P(\text{数据}|\text{理论}) \cdot P(\text{理论}).$$

其中 $P(\text{理论})$ 表示理论为真的验前概率(**prior probability**)，条件概率 $P(\text{数据}|\text{理论})$ 也称为**似然(likelihood)**，表示在给定的理论假设下观测到实际数据的概率。在得到实验的观测结果后理论为真的**验后概率(posterior probability)**由条件概率 $P(\text{理论}|\text{数据})$ 给出。这里没有明显给出数据的验前概率 $P(\text{数据})$ ，所以方程用正比关系表示。如何为理论设定验前概率？贝叶斯统计没有提供这样的基本原则。但是，一旦设定了这个验前概率，贝叶斯统计就可以告诉我们信心程度如何根据实验数据而变化。

让我们重新考虑在检验结果为阳性的条件下被检人员的患病概率。从研究者的角度，研究者研究大量该疾病的潜在携带者，这个问题中的概率可以解释为相对频率。验前概率 $P(\text{患病})$ 为患病人员的总比例，验后概率 $P(\text{患病}|\text{+})$ 为检验结果为阳性的人群中患病人员的比例。经典统计的中心问题是通过研究有限样本，即总体的子样本，估计可以描述总体的概率。

然而，对特定的个人而言，他可能对自己在检验结果为阳性时患病的主观概率更感兴趣。如果没有其它信息，人们一般将验前概率 $P(\text{患病})$ 取为疾病携带者的总比例，即与相对频率解释相同。不过，此时这个概率表示在进行检验之前某人患病的信心程度。如果能够获得其它信息，则可以根据这些信息设定不同的验前概率，正如主观概率解释这一名称本身所暗示，贝叶斯统计在这方面必然是主观的。然而，一旦设定了 $P(\text{患病})$ ，贝叶斯定理就可以告诉我们患病的概率(即在该假设条件下的信心程度)如何根据阳性检验结果变化。

我们将在6.13节、9.8节和11.5.3小节进一步讨论主观概率的应用。有很多关于主观概率的文献，比如 Jeffreys [11]、Savage [12]、de Finetti [13] 等人的书，以及 Cox [14] 的文章。贝叶斯方法的应用在 Lindley [15]、O'hagan [16]、Lee [17] 和 Sivia [18] 的书中均有所讨论。

1.3 概率密度函数

考虑某实验，其结果可以用一个连续变量 x 描述。样本空间对应于 x 可能取值的集合，我们可以求观测值在无限小区间 $[x, x + dx]$ 内的概率。² 这个概率由概率密度函数(**probability density function, p.d.f.**) $f(x)$ 给出：

$$\text{观测到 } x \text{ 处于区间 } [x, x + dx] \text{ 的概率} = f(x)dx. \quad (1.11)$$

在相对频率的解释中， $f(x)dx$ 给出观测次数无穷大极限下，观测到 x 处在区间 $[x, x + dx]$ 内的次数的比率。概率密度函数 $f(x)$ 是归一化的，以保证总概率为 1，

$$\int_S f(x)dx = 1, \quad (1.12)$$

其中积分区域 S 表示 x 可取值的总区间，即整个样本空间。

²这里的符号使用可能会引起混淆，因为 x 既表示随机变量本身，也表示随机变量的可能取值。很多作者使用大写字母表示随机变量，用小写字母表示随机变量的可能取值，即随机变量 X 在区间 $[x + dx]$ 中取值。简要起见，这里不采用这种符号形式，随机变量和它们的取值应该很容易从上下文中文中区分出来。

尽管我们将在第五章更全面地讨论有限数据样本的处理，这里仍有必要指出概率密度函数 $f(x)$ 与 x 的一组 (n 个) 观测值 x_1, \dots, x_n 的联系。这样一组观测值可以用**直方图(histogram)**表示，如图1.2所示。直方图的 x 轴分成了 m 个子区或**区间(bins)**，每个区间的宽度为 Δx_i ， $i = 1, \dots, m$ ，通常采用等区间宽度，有时也采用非等区间宽度的方式。纵轴表示 x 在第 i 个子区域的发生次数 n_i ，即该区间的频数。直方图下方的面积等于总事例数 n 乘以 Δx (对于不等宽区间，面积 $= \sum_{i=1}^m n_i \cdot \Delta x_i$)。所以，可以通过将每个频数 n_i 除以对应的区间宽度 Δx_i 和总事例数 n ，使直方图的面积归一化为 1。在区间宽度趋于零且总事例数无穷大的极限下， x 的归一化直方图就是概率密度函数 $f(x)$ ，如图1.2(d) 所示。

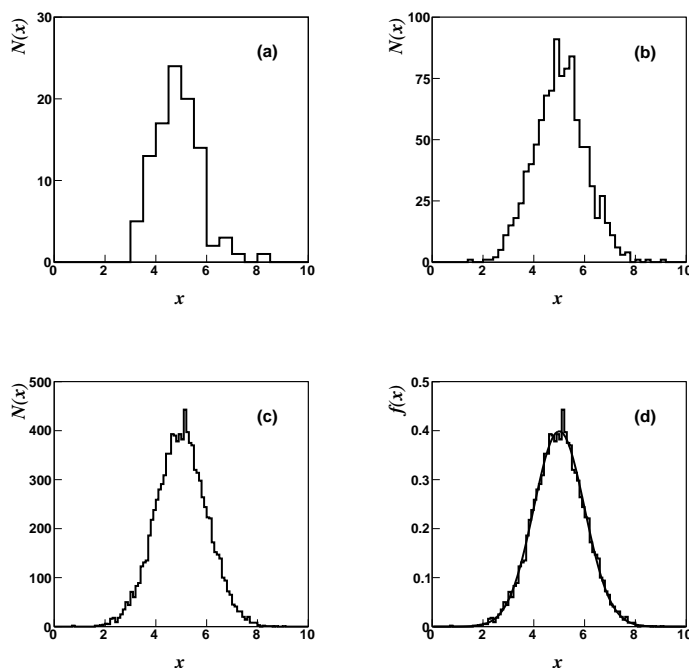


图 1.2: 概率密度函数为 $f(x)$ 的随机变量 x ，在给定不同观测次数 n 时的直方图。(a) 观测次数 $n = 100$ ，区间宽度 $\Delta x = 0.5$ 。(b) 观测次数 $n = 1000$ ， $\Delta x = 0.2$ 。(c) 观测次数 $n = 10000$ ， $\Delta x = 0.1$ 。(d) 直方图与 (c) 相同，但面积归一化为 1，同时画出对应的概率密度函数 $f(x)$ 。(a-c) 的纵轴 $N(x)$ 为 x 处在相应区间的频数。(d) 的纵轴为 $f(x) = N(x)/(n\Delta x)$ 。

对于离散情形，即变量 x 只能取离散值 x_i ，其中 $i = 1, \dots, N$ ， N 可以为无穷大。此时，对应的概率可以表示为

$$\text{观测值为 } x_i \text{ 的概率} = P(x_i) = f_i, \quad (1.13)$$

其中 $i = 1, \dots, N$ ，归一化条件为

$$\sum_{i=1}^N f_i = 1. \quad (1.14)$$

尽管后面的例子多数是连续变量，但是从连续变量到离散变量的变换非常直接，把对应的积分变成求和即可。

概率密度函数 $f(x)$ 的**累积分布(cumulative distribution)** $F(x)$ 为

$$F(x) = \int_{-\infty}^x f(x') dx', \quad (1.15)$$

即, $F(x)$ 表示随机变量取值小于等于 x 的概率。³ 实际上, $F(x)$ 通常定义为测量结果小于等于 x 的概率, 而概率密度函数 $f(x)$ 则定义为 $\partial F/\partial x$ 。数据分析中遇到的分布通常性质很好 (即 $F(x)$ 处处可微), 这种情况下两种定义是等价的。图1.3说明了概率密度 $f(x)$ 与累积分布 $F(x)$ 的联系。

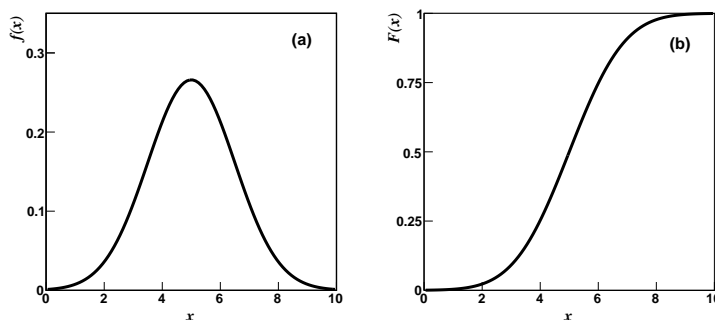


图 1.3: (a) 概率密度函数 $f(x)$ 。(b) 对应的累积分布函数 $F(x)$ 。

对于离散随机变量, 取值 x_i 的概率为 $P(x_i)$, 累积分布定义为观测值小于等于 x 的概率

$$F(x) = \sum_{x_i \leq x} P(x_i). \quad (1.16)$$

有一个很有用的概念与累积分布紧密相关, 就是所谓的 α 阶分位数(quantile of order α), 也称作 α -点(α -point)。分位数 x_α 定义为满足 $F(x_\alpha) = \alpha$ 的随机变量 x 的值, 其中 $0 \leq \alpha \leq 1$ 。也就是说, 分位数就是累积分布的反函数

$$x_\alpha = F^{-1}(\alpha). \quad (1.17)$$

常用的一种特殊情形为 $x_{1/2}$, 称为 x 的中位数(median)。中位数经常用于度量随机数的典型“位置”, 即观测到 x 大于或小于 $x_{1/2}$ 的概率相等。

另一个常用的位置度量为众数(mode), 定义为概率密度函数取最大值时随机变量的值。当然, 概率密度函数可能存在局域最大值。到目前为止, 最常用的位置参数是随机变量的期望值, 这将在第1.5节中介绍。

下面考虑一种情形, 测量结果不是用一个而是多个量一起描述, 这几个量可以看作多维随机变量, 或者说随机矢量。例如, 如果要研究某个人群, 对每个人可能要测量身高、体重、年龄等信息。假设某个测量可以用两个连续随机变量 x 和 y 描述, 令事件 A 表示 x 处在 $[x, x + dx]$ 之间而 y 处在任意位置, 令事件 B 表示 y 处在 $[y, y + dy]$ 之间而 x 处在任意位置, 如图1.4所示。

联合概率密度函数(joint p.d.f.) $f(x, y)$ 定义为

$$\begin{aligned} P(A \cap B) &= x \text{ 处在 } [x, x + dx] \text{ 且 } y \text{ 处在 } [y, y + dy] \text{ 的概率} \\ &= f(x, y) dx dy. \end{aligned} \quad (1.18)$$

因此, 联合概率密度函数 $f(x, y)$ 对应于散点图中事例数趋于无穷时散点的密度。因为 x 和 y 必定要取某个值, 归一化条件可以表示为

$$\iint_S f(x, y) dx dy = 1. \quad (1.19)$$

³数学家称 $F(x)$ 为“分布”函数, 而物理学家经常用分布一词表示概率密度函数。为了避免混淆, 我们使用累积分布和概率密度 (或概率密度函数) 的术语。

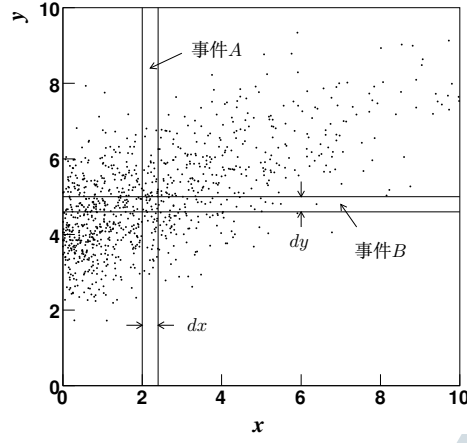


图 1.4: 二维随机变量 x 和 y 的散点图, 共 1000 个事例。观测到某一点处于两条区域交叉的方格中 (事件 $A \cap B$) 的概率由联合概率密度乘以面积元给出, 即 $f(x, y)dx dy$ 。

假设已知联合概率密度函数 $f(x, y)$, 希望得到 x 的概率密度函数, 不论 y 取任何值。这对应于图 1.4 中的事件 A 。如果将“事件 A ”看作包含面积元 $dx dy$ 的竖列, 每个面积元用指标 i 标记, 则 A 的概率直接就是对应的各个方格的概率之和,

$$P(A) = \sum_i f(x, y_i) dy dx = f_x(x) dx. \quad (1.20)$$

因此, 对应的概率密度由函数 $f_x(x)$ 给出, 称为 x 的**边缘概率密度函数(marginal p.d.f.)**。在 dy 无穷小的极限下, 求和变成积分, 所以边缘概率密度函数与联合概率密度函数的关系为

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy. \quad (1.21)$$

类似地, $f(x, y)$ 对 x 积分可以得到边缘概率密度函数 $f_y(y)$,

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx. \quad (1.22)$$

边缘概率密度函数 $f_x(x)$ 和 $f_y(y)$ 对应于, 将 x 和 y 的散点图投影到相应的坐标轴上并归一化后得到的直方图。图 1.5 说明了边缘概率密度函数与联合概率密度函数的关系。

从条件概率(1.2)的定义可知, 如果给定 x 处于区间 $[x, x + dx]$ 而 y 为任意值 (事件 A), 则 y 处于区间 $[y, y + dy]$ 而 x 为任意值 (事件 B) 的概率为

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x, y) dx dy}{f_x(x) dx}. \quad (1.23)$$

因此, 给定 x 时 y 的**条件概率密度函数(conditional p.d.f.)** $h(y|x)$ 定义为

$$h(y|x) = \frac{f(x, y)}{f_x(x)} = \frac{f(x, y)}{\int f(x, y') dy'}. \quad (1.24)$$

这是单个随机变量 y 的概率密度函数, 而 x 被当作常量参数。从 $f(x, y)$ 出发, 我们可以简单地认为 x 为常数, 然后将函数重新归一化, 使其单独对 y 积分等于 1。

条件概率密度函数 $h(y|x)$ 对应于, 将 (x, y) 散点图上 x 处的窄条 (即宽度 dx 无限小) 投影到 y 轴得到的归一化后的直方图。图 1.6 对此进行了说明, 对两个不同的 x 值, 得到了不同的条件概率密度函数 $h(y|x_1)$

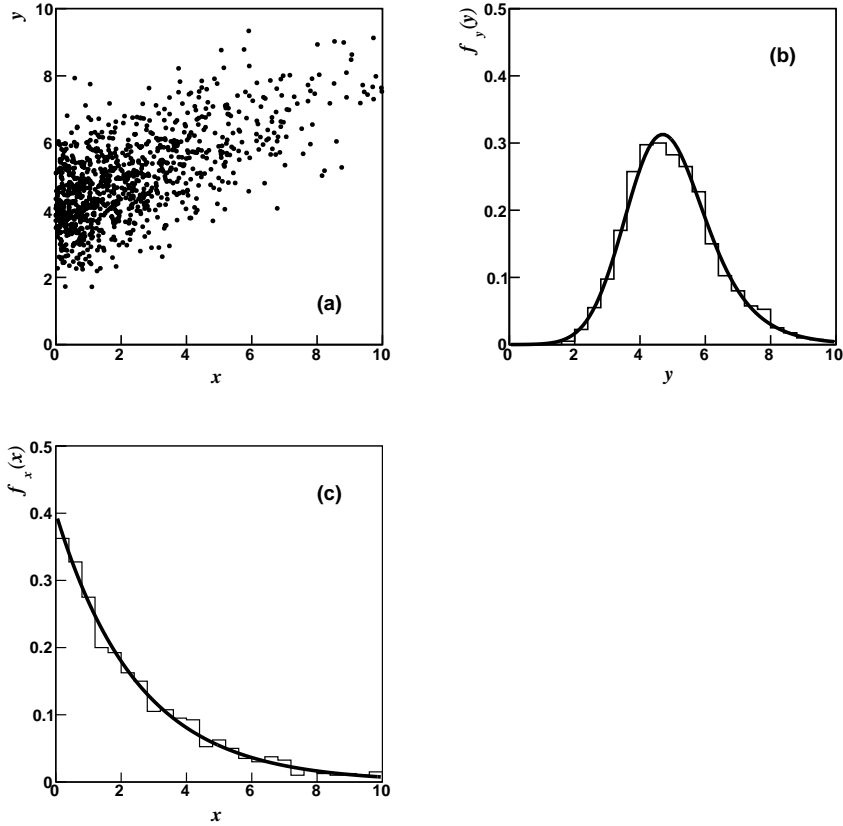


图 1.5: (a) 散点图上点的密度由联合概率密度函数 $f(x, y)$ 给出。 (b) 将散点图上的点投影到 y 轴并归一化后得到的直方图, 以及相应的边缘概率密度函数 $f_y(y)$ 。 (c) 将散点图投影到 x 轴得到 $f_x(x)$ 。

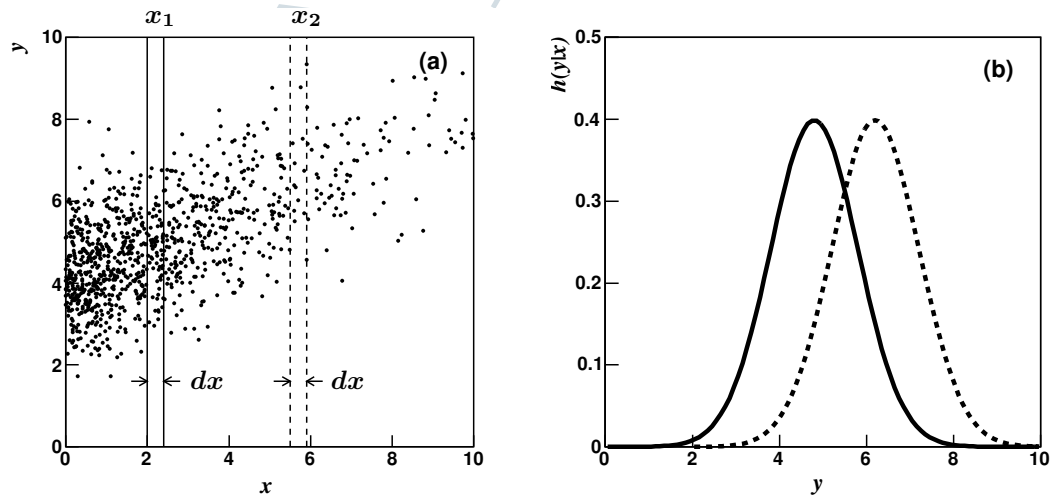


图 1.6: (a) 随机变量 x 和 y 的散点图, 图中标出了变量 x 在 x_1 (实线带) 和 x_2 (虚线带) 处两个极小宽度 (dx) 的长带。 (b) 条件概率 $h(y|x_1)$ 和 $h(y|x_2)$ 分别对应于两条长带在 y 轴上的投影。

和 $h(y|x_2)$ 。需要注意的是, 图1.6中的 $h(y|x_1)$ 和 $h(y|x_2)$ 都按照概率密度函数定义的要求归一化到了单位面积。

类似地, 给定 y 时, x 的条件概率为

$$g(x|y) = \frac{f(x, y)}{f_y(y)} = \frac{f(x, y)}{\int f(x', y) dx'}. \quad (1.25)$$

联立式 (1.24) 和 (1.25), 可以给出 $g(x|y)$ 与 $h(y|x)$ 的关系

$$g(x|y) = \frac{h(y|x)f_x(x)}{f_y(y)}, \quad (1.26)$$

这就是连续随机变量的贝叶斯定理(参见 (1.6) 式)。

利用 $f(x, y) = h(y|x)f_x(x) = g(x|y)f_y(y)$, 可以将边缘概率密度函数表示为

$$f_x(x) = \int_{-\infty}^{\infty} g(x|y)f_y(y)dy, \quad (1.27)$$

$$f_y(y) = \int_{-\infty}^{\infty} h(y|x)f_x(x)dx. \quad (1.28)$$

这对应于 (1.7) 式的全概率定律推广到连续随机变量的情形。

如果 “ x 处于 $[x, x + dx]$ 区间内而 y 任意”(事件 A) 和 “ y 处于 $[y, y + dy]$ 区间内而 x 任意”(事件 B) 相互独立, 即 $P(A \cap B) = P(A)P(B)$, 则 x 和 y 对应的联合概率密度函数可以因子化为

$$f(x, y) = f_x(x)f_y(y). \quad (1.29)$$

从式 (1.24) 和 (1.25) 可以看出, 对于相互独立的随机变量 x 和 y , 条件概率密度函数 $g(x|y)$ 与 y 无关, 即对所有 y , $g(x|y)$ 都相同。同样可得, $h(y|x)$ 与 x 无关。换言之, 了解其中一个变量的信息并不影响另一个变量的概率。然而, 图1.6中给出的两个随机变量 x 和 y 不相互独立, 因为 $h(y|x)$ 依赖于 x 的取值。

1.4 随机变量的函数

随机变量的函数本身也是随机变量。假设 $a(x)$ 为某连续随机变量 x 的连续函数, 其中 x 的分布服从概率密度函数 $f(x)$, 那么, 描述 a 的分布的概率密度函数是什么呢? 显然, x 发生在 x 到 $x + dx$ 区间内的概率等于 a 发生在 a 到 $a + da$ 区间内的概率, 可以根据这一要求确定 a 的概率密度函数, 即

$$g(a')da' = \int_{dS} f(x)dx, \quad (1.30)$$

其中无限小积分区间 dS 由 x -空间中 $a(x) = a'$ 和 $a(x) = a' + da'$ 之间的区域确定, 如图1.7所示。如果可求出函数 $a(x)$ 的逆函数 $x(a)$, 则 (1.30) 式给出

$$g(a)da = \left| \int_{x(a)}^{x(a+da)} f(x')dx' \right| = \int_{x(a)}^{x(a) + \left| \frac{dx}{da} \right| da} f(x')dx', \quad (1.31)$$

或者

$$g(a) = f(x(a)) \left| \frac{dx}{da} \right|. \quad (1.32)$$

积分限中对 dx/da 取绝对值是为了保证积分为正。如果函数 $a(x)$ 的逆不唯一, 对 dS 区域的积分必须包含 x -空间中 $a(x) = a'$ 和 $a(x) = a' + da'$ 之间的所有区间, 如图1.7(b) 所示。

如果 n 个随机变量 x_1, \dots, x_n 的联合概率密度函数为 $f(x_1, \dots, x_n)$, $a(x_1, \dots, x_n)$ 是这 n 个随机变量的函数, 则其概率密度函数 $g(a)$ 由下式确定:

$$g(a')da' = \int \cdots \int_{dS} f(x_1, \dots, x_n)dx_1 \dots dx_n, \quad (1.33)$$

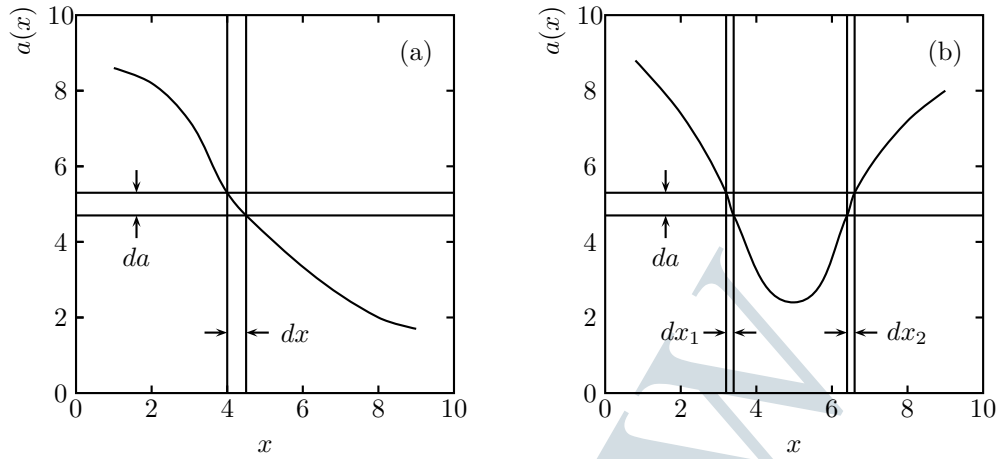


图 1.7: 随机变量的函数 $a(x)$ 的变换。(a) 函数 $a(x)$ 的逆函数 $x(a)$ 为单值函数；(b) 函数 $a(x)$ 的区间 da 对应于 x 的两个区间 dx_1 和 dx_2 。

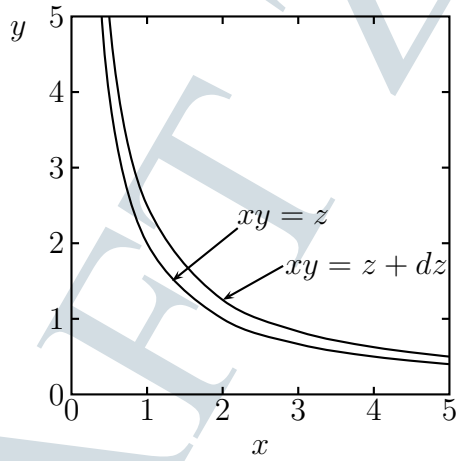


图 1.8: 积分区域 dS 为两条曲线 $xy = z$ 与 $xy = z + dz$ 之间的区域。 (x, y) 发生在两条曲线之间，对应于 z 的值处在 $[z, z + dz]$ 区间。

其中无限小体积元 dS 为 x_1, \dots, x_n -空间中 $a(x_1, \dots, x_n) = a'$ 和 $a(x_1, \dots, x_n) = a' + da'$ 这两个(超)平面之间的区域。

作为例子，考虑两个独立的随机变量 x 和 y ，其概率密度函数分别为 $g(x)$ 和 $h(y)$ 。我们希望得到乘积 $z = xy$ 的概率密度函数。由于 x 和 y 独立，其联合概率密度函数为 $g(x)h(y)$ 。于是，利用 (1.33) 式可以得到随机变量 z 的概率密度函数 $f(z)$

$$f(z)dz = \iint_{dS} g(x)h(y)dx dy = \int_{-\infty}^{\infty} g(x)dx \int_{z/|x|}^{(z+dz)/|x|} h(y)dy, \quad (1.34)$$

其中 dS 为 $xy = z$ 和 $xy = z + dz$ 之间的区域，如图1.8所示。容易得到

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(x)h(z/x) \frac{dx}{|x|} \\ &= \int_{-\infty}^{\infty} g(z/y)h(y) \frac{dy}{|y|}, \end{aligned} \quad (1.35)$$

其中第二个式子与第一个等价，可以直接通过变换积分次序得到。(1.35) 式经常记为 $f = g \otimes h$ ，函数 f 称为 g 和 h 的**梅林卷积(Mellin convolution)**。

类似地，可以得到 $z = x + y$ 的概率密度函数 $f(z)$

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(x)h(z-x)dx \\ &= \int_{-\infty}^{\infty} g(z-y)h(y)dy. \end{aligned} \quad (1.36)$$

(1.36) 式也经常记为 $f = g \otimes h$ ，函数 f 称为 g 和 h 的**傅立叶卷积(Fourier convolution)**。文献中提到这两种卷积时经常省略“梅林 (Mellin)”或“傅立叶 (Fourier)”，需要根据上下文判断到底是哪一类型的卷积。

假设有 n 个随机变量 $\mathbf{x} = (x_1, \dots, x_n)$ ，可以用下面的方法确定 n 个线性独立的函数 $a_i(\mathbf{x})$ 的联合概率密度函数。假设函数 a_1, \dots, a_n 可以反解出其反函数 $x_i(a_1, \dots, a_n)$ (其中 $i = 1, \dots, n$)，则联合概率密度函数为

$$g(a_1, \dots, a_n) = f(x_1, \dots, x_n)|J|, \quad (1.37)$$

其中 $|J|$ 为该变换的雅可比 (Jacobian) 行列式的绝对值，

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial a_1} & \frac{\partial x_1}{\partial a_2} & \dots & \frac{\partial x_1}{\partial a_n} \\ \frac{\partial x_2}{\partial a_1} & \frac{\partial x_2}{\partial a_2} & \dots & \frac{\partial x_2}{\partial a_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & \frac{\partial x_n}{\partial a_n} \end{vmatrix} \quad (1.38)$$

要确定其中某个函数 (例如 $g_1(a_1)$) 的边缘概率密度函数，需要对联合概率密度函数 $g(a_1, \dots, a_n)$ 中所有其它 a_i 进行积分。

很多情况下，上面的方法很难解析求解。例如，如果对 n 个随机变量的某个函数感兴趣，并且 n 很大，而且很可能本身也是一个随机数。从实际操作的角度，很难做到补充 $n-1$ 个附加函数，利用上面的方法求解联合概率密度函数，然后对 $n-1$ 个不需要的量进行积分。此时，通常采用第三章介绍的蒙特卡罗 (Monte-Carlo) 方法进行数值求解。如果只关心函数的均值和方差，可以利用1.6节介绍的方法，即所谓的“误差传递”过程进行求解。

某些情形下，随机变量函数的概率密度函数可以利用积分变换的技术求解，例如用傅立叶变换求解随机变量之和的概率密度函数，用梅林变换求解随机变量乘积的概率密度函数，基本思想是分别取 (1.35) 式或 (1.36) 式的梅林变换或傅立叶变换。于是，方程 $f = g \otimes h$ 转化为变换后的密度函数的乘积 $\tilde{f} = \tilde{g} \cdot \tilde{h}$ ，概率密度函数 f 可以通过 \tilde{f} 的逆变换得到。对这些方法的完整介绍超出了本书的范围，可以参见文献 [19]。本书在第十章给出了几个例子，利用傅立叶变换 (特征函数) 求随机变量之和的概率密度函数。

1.5 期望值

假设随机变量 x 服从概率密度函数为 $f(x)$ ，则 x 的**期望值(expectation value)** $E[x]$ 定义为

$$E[x] = \int_{-\infty}^{\infty} xf(x)dx = \mu. \quad (1.39)$$

x 的期望值经常用 μ 表示 (期望值也称为 x 的**总体平均(population mean)**，或者简称为均值)。注意 $E[x]$ 不是 x 的函数，而是依赖于概率密度函数 $f(x)$ 的形式。如果概率密度函数主要集中在某个区域，则 $E[x]$ 可以度量 x 很可能在哪个区域被观测到。然而，也可能 $f(x)$ 包含两个相距很远的峰，这样 $E[x]$ 处在两个峰的中间，然而 x 很少会在这个区域被观测到。

对于随机变量的函数 $a(x)$ ，其期望值为

$$E[a] = \int_{-\infty}^{\infty} ag(a)da = \int_{-\infty}^{\infty} a(x)f(x)dx, \quad (1.40)$$

其中 $g(a)$ 表示 a 的概率密度函数， $f(x)$ 表示 x 的概率密度函数。如果对 (1.30) 式等号两边同乘以 a ，然后对全空间积分，很容易看出这两个积分是等价的。

我们经常对一些特殊的期望值很感兴趣，例如

$$E[x^n] = \int_{-\infty}^{\infty} x^n f(x)dx = \mu'_n, \quad (1.41)$$

称为 x 的 n 阶代数矩。当 $n = 1$ 时，一阶代数矩就是随机变量的期望值，即 $\mu = \mu_1$ 。而期望值

$$E[(x - E[x])^n] = \int_{-\infty}^{\infty} (x - \mu)^n f(x)dx = \mu_n, \quad (1.42)$$

称为 x 的 n 阶中心矩。对二阶中心矩，

$$E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \sigma^2 = V[x], \quad (1.43)$$

称为 x 的**总体方差(population variance)**(或者简称为方差)，通常记作 σ^2 或 $V[x]$ 。注意， $E[(x - E[x])^2] = E[x^2] - \mu^2$ 。方差是随机变量 x 以均值为中心散开程度的一种度量。方差的平方根 σ 称为 x 的**标准差(standard deviation)**，实际应用中经常使用这个概念，因为它与 x 的具有相同的量纲。

如果 a 是多个随机变量 $\mathbf{x} = (x_1, \dots, x_n)$ 的函数，其期望值为

$$\begin{aligned} E[a(\mathbf{x})] &= \int_{-\infty}^{\infty} ag(a)da \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a(\mathbf{x})f(\mathbf{x})dx_1 \dots dx_n = \mu_a, \end{aligned} \quad (1.44)$$

其中 $g(a)$ 表示 a 的概率密度函数， $f(\mathbf{x})$ 表示 x_i 的联合概率密度函数。后面的章节中，我们经常使用 $\mu_a = E[a]$ 的表示形式。与单变量情形一样，(1.44) 式的两个积分是等价的，对 (1.33) 式等号两边同乘以 a ，然后对整个空间积分，很容易就可以看出这一点。 a 的方差为

$$\begin{aligned} V[a] &= E[(a - \mu_a)^2] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (a(\mathbf{x}) - \mu_a)^2 f(\mathbf{x})dx_1 \dots dx_n = \sigma_a^2, \end{aligned} \quad (1.45)$$

通常用 σ_a^2 或 $V[a]$ 表示。两个随机变量 x 和 y 的**协方差(covariance)** 定义为

$$\begin{aligned} V_{xy} &= E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x \mu_y \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dx dy - \mu_x \mu_y, \end{aligned} \quad (1.46)$$

其中 $\mu_x = E[x]$ ， $\mu_y = E[y]$ 。协方差矩阵 V_{xy} 也称误差矩阵，有时用 $\text{cov}[x, y]$ 表示。更一般的情况，如果 a 和 b 均是 n 个随机变量 $\mathbf{x} = (x_1, \dots, x_n)$ 的函数，其协方差 $\text{cov}[a, b]$ 为

$$\begin{aligned} \text{cov}[a, b] &= E[(a - \mu_a)(b - \mu_b)] \\ &= E[ab] - \mu_a \mu_b \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} abg(a, b)dadb - \mu_a \mu_b \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a(\mathbf{x})b(\mathbf{x})f(\mathbf{x})dx_1 \dots dx_n - \mu_a \mu_b, \end{aligned} \quad (1.47)$$

其中 $g(a, b)$ 表示 a 和 b 的联合概率密度函数, $f(\mathbf{x})$ 表示 x_i 的联合概率密度函数。与 (1.44) 式一样, 上式 V_{ab} 中的两个积分是等价的。需要注意的是, 协方差矩阵 V_{ab} 的构造本身关于 a 和 b 是对称的, 并且对角元素 $V_{aa} = \sigma_a^2$ (即方差) 为正。

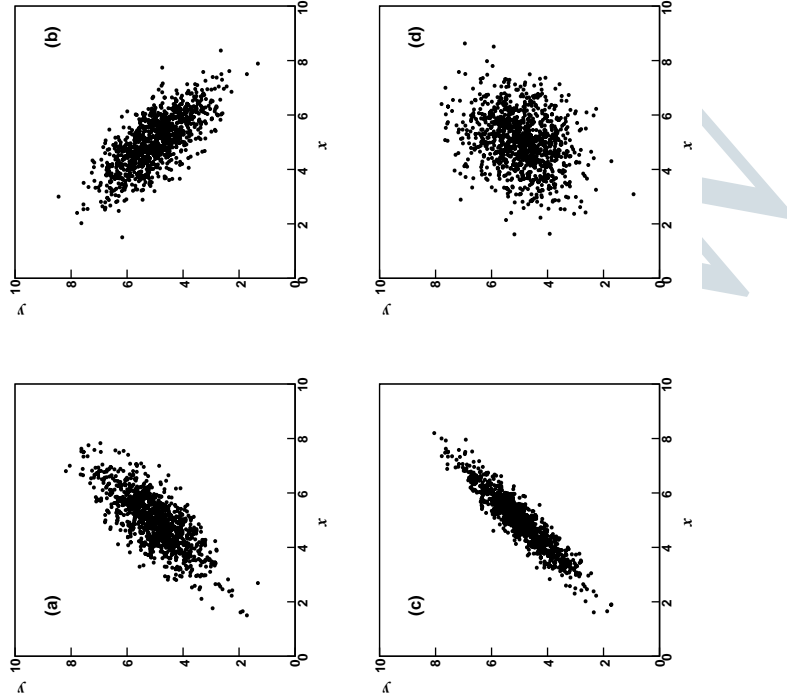


图 1.9: 随机变量 x 和 y 的散点分布。(a) 正相关, $\rho = 0.75$; (b) 负相关, $\rho = -0.75$; (c) $\rho = 0.95$; (d) $\rho = 0.25$ 。4 个图中, x 和 y 的标准差均为 $\sigma_x = \sigma_y = 1$ 。

有时需要给出两个随机变量 x 和 y 相关程度的无量纲度量, 这就需要使用**相关系数(correlation coefficient)**的概念, 定义为

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}. \quad (1.48)$$

容易证明 (参见 [3,20]), 相关系数的范围为 $-1 \leq \rho_{xy} \leq 1$ 。

两个随机变量 x 和 y 的协方差的意义大致可以这样理解: $(x - \mu_x)(y - \mu_y)$ 表示 x 和 y 与各自均值 μ_x 和 μ_y 的偏差的乘积, V_{xy} 表示这个乘积的期望值。假设 x 大于 μ_x 的条件可以提高 y 大于 μ_y 的概率, x 小于 μ_x 的条件可以提高 y 小于 μ_y 的概率, 那么 V_{xy} 大于零, 称这两个变量正相关。正相关的例子如图 1.9(a), (c) 和 (d) 的情形, 其相关系数 ρ_{xy} 分别为 0.75, 0.95 和 0.25。同样, $V_{xy} < 0$ 称为负相关, 即 $x > \mu_x$ 的条件提高 $y < \mu_y$ 的概率。负相关如图 1.9(b) 所示, 其中 $\rho_{xy} = -0.75$ 。

从式 (1.29) 和 (1.44) 可以看出, 对独立的随机变量 x 和 y ,

$$E[xy] = E[x]E[y] = \mu_x \mu_y, \quad (1.49)$$

式 (1.46) 给出 $V_{xy} = 0$, 所以相关系数为零。然而, 反过来并不一定成立, 即不能由相关系数为零推出两个随机变量相互独立。例如, 图 1.10 给出的概率密度函数的二维散点图, $V_{xy} = 0$, 但 x 和 y 不独立。也就是说, $f(x, y)$ 不能象 (1.29) 式那样因子化, 一个变量的已知条件将影响另外一个变量的条件概率。然而, 由于 $f(x, y)$ 在 x 轴上在均值 μ_x 两边是对称的, 所以 $V_{xy} = 0$ 。

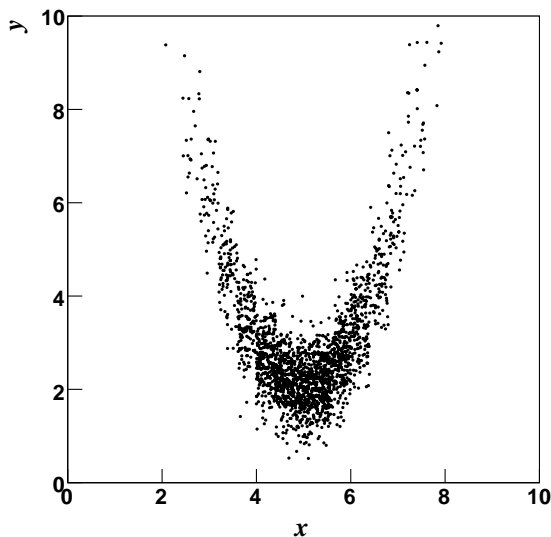


图 1.10: 随机变量 x 和 y 的散点分布。 x 和 y 不独立 (即 $f(x, y) \neq f_x(x)f_y(y)$), 但因为分布的特殊对称性, $V_{xy} = 0$ 。

1.6 误差传递

假设有 n 个随机变量 $\mathbf{x} = (x_1, \dots, x_n)$, 服从某联合概率密度函数 $f(\mathbf{x})$ 。假设我们对联合概率密度函数的了解不充分, 但 x_i 的均值 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ 以及协方差矩阵 V_{ij} 已知或者至少可以估计。(估计方法将在第五章介绍。)

考虑这 n 个变量的函数 $y(\mathbf{x})$ 。要确定 y 的概率密度函数, 原则上需要按照 1.4 节介绍的程序处理 (例如, 利用 (1.33) 或 (1.37) 式)。然而, 我们已经假设无法完全确定 $f(\mathbf{x})$, 而仅知道均值 $\boldsymbol{\mu}$ 和协方差矩阵 V_{ij} , 因此按照这些程序确定 y 的概率密度函数并不可行。不过, 我们可以通过将函数 $y(\mathbf{x})$ 围绕 x_i 的均值泰勒展开到一阶项, 得到 y 的期望值的近似值, 即

$$y(\mathbf{x}) \approx y(\boldsymbol{\mu}) + \sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_i - \mu_i). \quad (1.50)$$

由于 $E[x_i - \mu_i] = 0$, 所以近似到一阶项, y 的期望值为

$$E[y(\mathbf{x})] \approx y(\boldsymbol{\mu}). \quad (1.51)$$

y^2 的期望值为

$$\begin{aligned} E[y^2(\mathbf{x})] &= y^2(\boldsymbol{\mu}) + 2y(\boldsymbol{\mu}) \cdot \sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} E[x_i - \mu_i] \\ &\quad + E \left[\left(\sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_i - \mu_i) \right) \left(\sum_{j=1}^n \left[\frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_j - \mu_j) \right) \right] \\ &= y^2(\boldsymbol{\mu}) + \sum_{i,j=1}^n \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} V_{ij}, \end{aligned} \quad (1.52)$$

所以方差 $\sigma_y^2 = E[y^2] - (E[y])^2$ 为

$$\sigma_y^2 \approx \sum_{i,j=1}^n \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} V_{ij}. \quad (1.53)$$

类似地，对于 m 个函数 $y_1(\mathbf{x}), \dots, y_m(\mathbf{x})$ ，可以得到它们的协方差矩阵

$$U_{kl} = \text{cov}[y_k, y_l] \approx \sum_{i,j=1}^n \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} V_{ij}. \quad (1.54)$$

用矩阵形式可以表示为

$$U = A V A^T, \quad (1.55)$$

其中导数矩阵 A 为

$$A_{ij} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} \quad (1.56)$$

A^T 为矩阵 A 的转置。式 (1.53)-(1.56) 构成**误差传递(error propagation)**的基础 (即用来衡量统计不确定性的方差从 x_i 传递到 y_1, y_2, \dots)。 (“误差”这个词经常用来表示测量的不确定性，多数情况下由对应的随机变量的标准差给出。)

如果 x_i 之间不相关，即 $V_{ii} = \sigma_i^2$ ，则当 $i \neq j$ 时 $V_{ij} = 0$ ，式 (1.53) 和 (1.54) 变为

$$\sigma_y^2 \approx \sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}}^2 \sigma_i^2 \quad (1.57)$$

$$U_{kl} \approx \sum_{i=1}^n \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}} \sigma_i^2. \quad (1.58)$$

利用 (1.53) 式可以得到以下两种特殊情形。如果 $y = x_1 + x_2$ ，则 y 的方差为

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}. \quad (1.59)$$

对于 $y = x_1 x_2$ ，则得到

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2 \frac{V_{12}}{x_1 x_2}. \quad (1.60)$$

如果变量 x_1 和 x_2 不相关 ($V_{12} = 0$)，上面的结果表明：对于独立随机变量的和 $y = x_1 + x_2$ ，误差 (标准差) 平方相加；对于对立随机变量的积 $y = x_1 x_2$ ，相对误差平方和相加。

在推导误差传递公式时，我们假设已知原始随机变量 x_1, \dots, x_n 的均值和方差 (或者至少可以估计)，并假设待求的随机变量的函数可以在均值 μ_1, \dots, μ_n 附近作一阶泰勒展开。后一个假设显然只有在线性函数条件下才严格成立。如果在 x_i 的标准差 $\sigma_1, \dots, \sigma_n$ 的尺度上，函数 $y(\mathbf{x})$ (或 $\mathbf{y}(\mathbf{x})$) 在均值 $\boldsymbol{\mu}$ 附近非线性行为显著，这个近似就不成立了。在处理类似于 $y = 1/x$ 这样的函数时，如果 $E[x] = \mu$ 小于标准差或者与之相当，那么使用误差传递就需要特别小心。对于这种情形，利用第三章介绍的蒙特卡洛方法处理更好，或者可以利用9.2节介绍的置信区间来处理。

1.7 随机变量的正交变换

假设有 n 个随机变量 x_1, \dots, x_n ，已知协方差矩阵 $V_{ij} = \text{cov}[x_i, x_j]$ ，矩阵的非对角元不一定为零。很多情况下，重新定义 n 个不相关的变量 y_1, \dots, y_n ，可以简化问题的处理。由于 y_1, \dots, y_n 互不相关，新的协方差矩阵 $U_{ij} = \text{cov}[y_i, y_j]$ 为对角矩阵。下面将证明通过线性变换我们总是可以做到这一点，即

$$y_i = \sum_{j=1}^n A_{ij} x_j. \quad (1.61)$$

假设存在这样的变换，新变量的协方差矩阵为

$$\begin{aligned}
 U_{ij} &= \text{cov}[y_i, y_j] = \text{cov} \left[\sum_{k=1}^n A_{ik} x_k, \sum_{l=1}^n A_{jl} x_l \right] \\
 &= \sum_{k,l=1}^n A_{ik} A_{jl} \text{cov}[x_k, x_l] \\
 &= \sum_{k,l=1}^n A_{ik} V_{kl} A_{lj}^T.
 \end{aligned} \tag{1.62}$$

这实际上就是误差传递公式(1.54)的特殊形式。由于变换 (1.61) 是线性的，所以该式严格成立。

于是问题变为，寻找矩阵 A 使得 $U = AVA^T$ 为对角矩阵。这正是将实对称矩阵对角化的问题，是线性代数中的常见问题 (参见 [21])。求解这个问题需要先确定协方差矩阵 V 的本征矢量 \mathbf{r}^i , $i = 1, \dots, n$ 。即，必须求解方程

$$V\mathbf{r}^i = \lambda_i \mathbf{r}^i, \tag{1.63}$$

在矩阵形式下矢量 \mathbf{r}^i 应当理解为列矢量。本征矢量只能确定到相差一个相乘因子，可以选择该因子使得本征矢量的长度都是 1。而且，由于协方差矩阵是对称的，很容易证明本征矢量相互正交，即

$$\mathbf{r}^i \cdot \mathbf{r}^j = \sum_{k=1}^n r_k^i r_k^j = \delta_{ij}. \tag{1.64}$$

如果两个或多个本征值 $\lambda_i, \lambda_j, \dots$ 相等，则对应的本征矢量 $\mathbf{r}^i, \mathbf{r}^j, \dots$ 的方向不能唯一确定，但总可以选择其方向使得这些本征矢量相互正交。

变换矩阵 A 的 n 行由 n 个本征矢量 \mathbf{r}^i 给出 (任何顺序即可)，即 $A_{ij} = r_j^i$ ；转置矩阵的列则为本征矢量，即 $A_{ij}^T = r_i^j$ 。将 A 代入 (1.62) 式，很容易证明矩阵 A 满足所要求的性质，即

$$\begin{aligned}
 U_{ij} &= \sum_{k,l=1}^n A_{ik} V_{kl} A_{lj}^T = \sum_{k,l=1}^n r_k^i V_{kl} r_l^j \\
 &= \sum_{k=1}^n r_k^i \lambda_j r_k^j \\
 &= \lambda_j \mathbf{r}^i \cdot \mathbf{r}^j \\
 &= \lambda_j \delta_{ij}.
 \end{aligned} \tag{1.65}$$

因此，变换后的变量 y_1, \dots, y_n 的方差为原始协方差矩阵 V 的本征值， U 的所有非对角元等于零。由于本征矢量是正交归一的 (见 (1.64) 式)，所以有如下性质

$$\sum_{j=1}^n A_{ij} A_{jk}^T = \sum_{j=1}^n r_j^i r_j^k = \mathbf{r}^i \cdot \mathbf{r}^k = \delta_{ik}, \tag{1.66}$$

或者用矩阵形式表示为 $AA^T = 1$ ，因此 $A^T = A^{-1}$ 。这种变换称为正交变换，即该变换将矢量 \mathbf{x} 旋转为 \mathbf{y} ，而保持矢量的模不变，因为 $|\mathbf{y}|^2 = \mathbf{y}^T \mathbf{y} = \mathbf{x}^T A^T A \mathbf{x} = |\mathbf{x}|^2$ 。

可以利用线性代数中的标准方法求 V 的本征矢量 (例如参见 [21])。当变量超过三个时，解析求解实际上相当困难，此时需要用数值方法求解，例如**奇异值分解(singular value decomposition)**的方法 (参见 [3, 22])。

对于二维情形，随机变量 $\mathbf{x} = (x_1, x_2)$ 的协方差矩阵可以表示为

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{1.67}$$

通过要求系数矩阵的行列式为零，可以求解本征方程 $(V - I\lambda)\mathbf{r} = 0$ (其中 I 为 2×2 的单位矩阵)，即

$$\det(V - I\lambda) = 0. \quad (1.68)$$

计算可得两个本征值为

$$\lambda_{\pm} = \frac{1}{2} \left[\sigma_1^2 + \sigma_2^2 \pm \sqrt{(\sigma_1^2 + \sigma_2^2)^2 - 4(1 - \rho^2)\sigma_1^2\sigma_2^2} \right]. \quad (1.69)$$

两个正交的本征矢量 \mathbf{r}_{\pm} 可以用角度 θ 来参数化，

$$\mathbf{r}_+ = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix} \quad \mathbf{r}_- = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}. \quad (1.70)$$

将本征值 (1.69) 代回本征值方程可以确定出角度 θ ，

$$\theta = \frac{1}{2} \tan^{-1} \left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2} \right). \quad (1.71)$$

将本征矢量作为矩阵的行，即可得到所需的变换矩阵

$$A = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}. \quad (1.72)$$

这实际上就是将矢量 (x_1, x_2) 旋转角度 θ 。图1.11为随机变量正交变换的例子，其中原始的两个变量的标准差分别为 $\sigma_1 = 1.5$ 和 $\sigma_2 = 1.0$ ，相关系数为 $\rho = 0.7$ 。

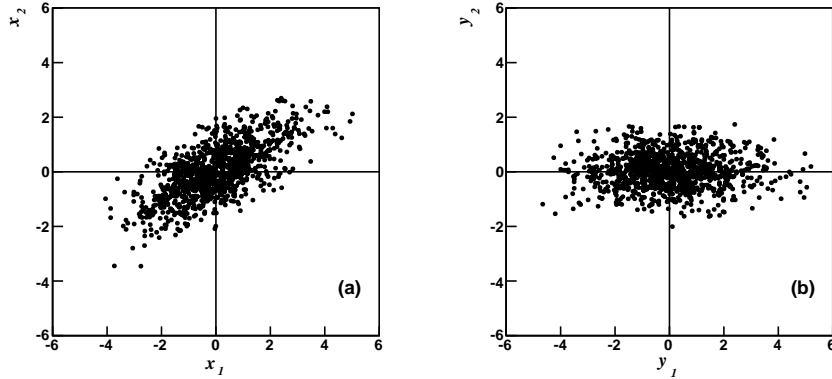


图 1.11: (a) 两个相关的随机变量 (x_1, x_2) 的散点分布，标准差分别为 $\sigma_1 = 1.5$ 和 $\sigma_2 = 1.0$ ，相关系数 $\rho = 0.7$ 。(b) 变换后的变量 (y_1, y_2) 的散点图，其协方差矩阵为对角矩阵。

尽管不相关的随机变量通常容易处理，但变换后的变量可能不象原始变量一样有直接的物理解释。正交变换的程序在第六章到第八章的参数估计中可能会用到，进行参数估计时，参数的估计量往往是相关的。

第二章 常用概率函数

本章将介绍一些常用的概率分布和概率密度函数。概率密度函数的均值、方差等性质将直接给出，一般不予以证明。概率密度函数的各阶矩可以利用第十章介绍的特征函数求出。其它更多的概率密度函数可以在参考文献 [20] 的第 4 章，[23] 的第 4 章以及 [3] 的第 5 章找到。

2.1 二项分布和多项分布

考虑 N 个独立的试验 (或观测)，每次试验都有两种可能的输出，称为“成功”或“失败”。成功的概率为某个常数 p 。这 N 个试验可以看作一次实验，用离散随机变量 n 描述， n 定义为成功的总数目，即样本空间为 N 次观测中成功的可能次数 n 。如果重复进行实验，每次实验都包含 N 个试验，得到的结果 n 将按照所谓的**二项分布(binomial distribution)**的相对频率发生。

二项分布的形式可以用下面的方法得到。假设单次观测成功的概率为 p ，失败的概率为 $1 - p$ 。由于每个观测相互独立，按照特定顺序给出的成功和失败序列的概率为每个事件概率的乘积。例如，五次试验中，按照“成功、成功、失败、成功、失败”顺序发生的概率为 $p \cdot p \cdot (1 - p) \cdot p \cdot (1 - p) = p^3(1 - p)^2$ 。进行 N 个试验的普遍情况下，按照特定顺序发生 n 次成功、 $N - n$ 次失败的概率为 $p^n(1 - p)^{N-n}$ 。然而，我们并不关心成功和失败序列的顺序，只关心最终成功的次数 n 。由于 N 个试验中，发生 n 次成功事件的序列数目为

$$\frac{N!}{n!(N - n)!}, \quad (2.1)$$

所以， N 次试验中发生 n 次成功事件的总概率为

$$f(n; N, p) = \frac{N!}{n!(N - n)!} p^n (1 - p)^{N-n}, \quad (2.2)$$

其中 $n = 0, \dots, N$ 。注意， $f(n; N, p)$ 本身为概率而不是概率密度。这里的符号约定为，把随机变量作为概率函数 (或概率密度函数) 的变量放到分号的左边，而其它所有参数 (这里为 N 和 p) 放到分号右边。 n 的期望值为

$$E[n] = \sum_{n=0}^N n \frac{N!}{n!(N - n)!} p^n (1 - p)^{N-n} = Np, \quad (2.3)$$

方差为

$$\begin{aligned} V[n] &= E[n^2] - (E[n])^2 \\ &= Np(1 - p). \end{aligned} \quad (2.4)$$

利用二项分布的特征函数 (参见第十章) 很容易就可以计算出均值和方差。

需要注意的是，期望值不是随机变量的函数，而依赖于概率函数的参数。对于二项分布，所依赖的参数即 p 和 N 。图2.1和2.2给出了 p 和 N 取不同参数值时的二项分布。

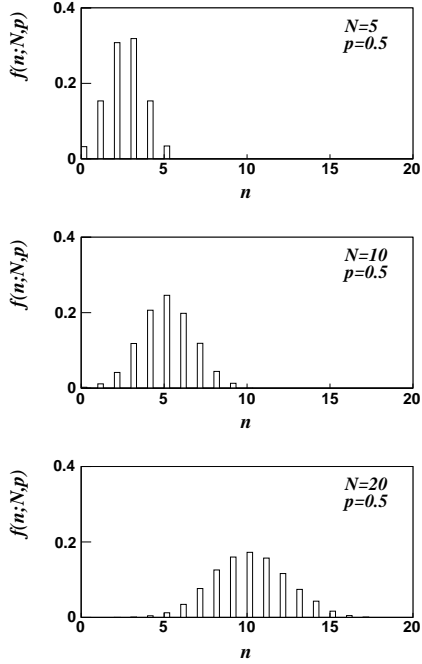


图 2.1: 二项分布, $p = 0.5$, $N = 5, 10, 20$ 。

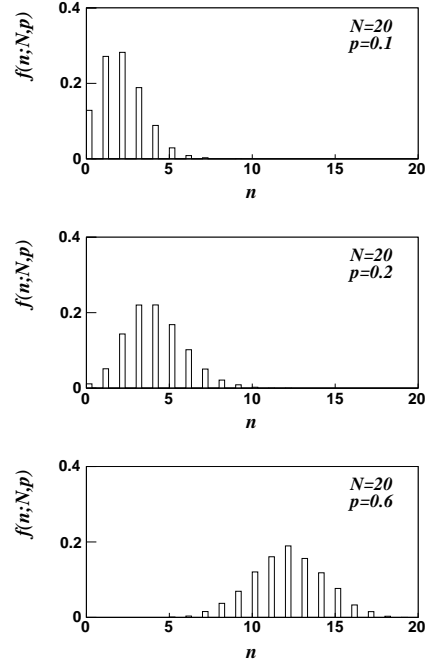


图 2.2: 二项分布, $N = 20$, $p = 0.1, 0.2, 0.6$ 。

多项分布(multinomial distribution) 是二项分布的推广, 即每次试验的输出结果不止两种 (“成功” 和 “失败”), 而存在 m 种不同的结果。对于某次试验, 第 i 种输出的概率为 p_i 。由于试验必定会有某种结果发生, 所以要求 $\sum_{i=1}^m p_i = 1$ 。

现在考虑试验次数为 N 的某个测量, 每次试验都产生 m 种可能输出中的一种。所有输出的一个特定序列, 例如第一次试验产生第 i 种结果, 第二次试验产生第 j 种结果, 等等, 其概率为 N 个对应的概率的乘积 $p_i p_j \cdots p_k$ 。第一种输出、第二种输出、……、第 m 种输出分别出现 n_1, n_2, \dots, n_m 次, 这样的序列的总数为

$$\frac{N!}{n_1! n_2! \cdots n_m!} \quad (2.5)$$

如果我们不关心各个输出出现的次序, 而只关心每种输出发生的总次数, 则第一种输出发生 n_1 次, 第二种输出发生 n_2 次, ……、第 m 种输出发生 n_m 次, 这个联合概率由多项分布给出

$$f(n_1, \dots, n_m; N, p_1, \dots, p_m) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m} \quad (2.6)$$

假设我们把 m 种输出分成两类: 输出 i (“成功”) 和不输出 i (“失败”)。这与前面讨论的二项分布的过程完全相同, 输出 i 的次数 n_i 显然服从二项分布。而且, 这一结论对所有的 i 都成立。由式 (2.3) 和 (2.4), 可以得到 n_i 的期望值为 $E[n_i] = N p_i$, 方差为 $V[n_i] = N p_i (1 - p_i)$ 。

如果考虑有三种可能的输出: i, j 以及所有其他输出。则第 i 种输出的次数为 n_i , 第 j 种输出的次数为 n_j , 其它输出的次数为 $N - n_i - n_j$ 的概率为

$$f(n_i, n_j; N, p_i, p_j) = \frac{N!}{n_i! n_j! (N - n_i - n_j)!} p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{N - n_i - n_j}, \quad (2.7)$$

因此, 协方差 $V_{ij} = \text{cov}[n_i, n_j]$ 在 $i \neq j$ 时为

$$\begin{aligned} V_{ij} &= E[(n_i - E[n_i])(n_j - E[n_j])] \\ &= -N p_i p_j, \end{aligned} \quad (2.8)$$

当 i 和 j 相等时, $V_{ii} = \sigma_i^2 = Np_i(1 - p_i)$ 。

作为多项分布的一个例子, 可以考虑对随机变量进行 N 次独立测量得到某个特定的直方图的概率, 即直方图分 m 个区间, 当总事例数为 N 时, 有 n_1 个事件在第一个区间, n_2 个事件在第二个区间, 等等。这个概率分布服从多项分布。从 (2.8) 式可以看出, 任意两个区间的事例数都是负相关的。换言之, 在 N 次试验中, 如果第 i 个区间包含的事例数超过平均事例数 ($n_i > Np_i$), 则另外一个区间 j 包含的事例数小于平均值的概率将增大。

2.2 泊松分布

考虑2.1节二项分布的极限行为。在 N 非常大的极限下, p 则变得非常小, 但要求乘积 Np (即成功事例数的期望值) 保持为某个有限的常数 ν 。可以证明 (2.2) 式在这个极限下变为 (见10.2节)

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}, \quad (2.9)$$

这个概率称为**泊松分布(Poisson distribution)**, 其中随机变量 $n = 0, 1, \dots, \infty$ 为整数。泊松分布的概率密度函数只有一个参数 ν 。图2.3给出了不同参数值 $\nu = 2, 5, 10$ 的泊松分布。

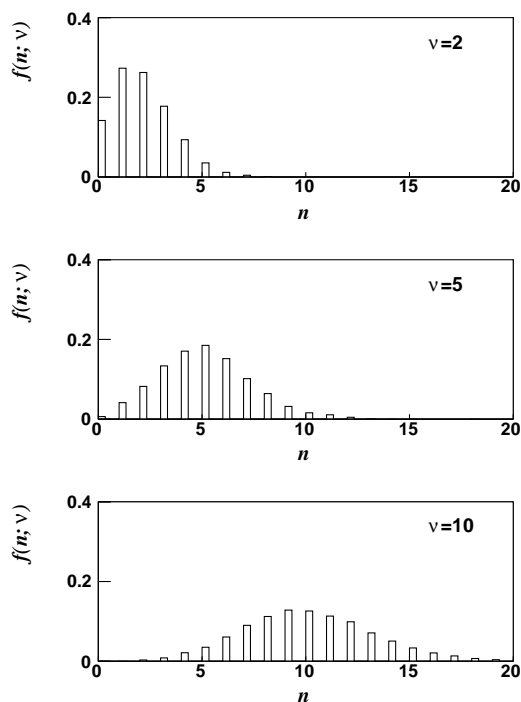


图 2.3: 参数 ν 取不同值时, 泊松分布的概率函数。

泊松随机变量 n 的期望值为

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n}{n!} e^{-\nu} = \nu, \quad (2.10)$$

方差为

$$V[n] = \sum_{n=0}^{\infty} (n - \nu)^2 \frac{\nu^n}{n!} e^{-\nu} = \nu. \quad (2.11)$$

尽管泊松变量是离散的，它可以被当作连续变量 x 处理，只要积分区间 Δx 明显大于 1。在第十章，我们将证明如果泊松分布的均值 ν 很大，则可以把它当成服从高斯分布 (参见2.5节) 的连续变量处理。

作为泊松随机变量的例子，可以考虑在固定时间间隔内，一定数量的放射性物质发生衰变的次数，要求的条件是可能衰变的总数 (即放射性原子的总数目) 非常大，并且给定时间间隔内单个衰变的概率很小。另外一个例子是，给定积分亮度 L 时，粒子散射实验中观测到某一类型事例的次数。事例数的期望值为

$$\nu = \sigma L \epsilon, \quad (2.12)$$

其中 σ 为事件的截面， ϵ 为效率，即事例被探测器探测到的概率。

2.3 均匀分布

连续变量 $x (-\infty < x < \infty)$ 的**均匀 (uniform)** 分布定义为

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \text{其它,} \end{cases} \quad (2.13)$$

即 x 在 α 和 β 之间的概率处处相同，而在其它地方的概率为零。 x 的均值和方差分别为

$$E[x] = \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx = \frac{1}{2}(\alpha + \beta), \quad (2.14)$$

$$V[x] = \int_{\alpha}^{\beta} [x - \frac{1}{2}(\alpha + \beta)]^2 \frac{1}{\beta - \alpha} dx = \frac{1}{12}(\beta - \alpha)^2. \quad (2.15)$$

均匀分布的一个重要特点是，对任何概率密度函数为 $f(x)$ 、累积分布函数为 $F(x)$ 的连续随机变量，都可以很容易地变换到新的随机变量 y ，使之服从 0 到 1 之间的均匀分布。变换后的随机变量 y 为

$$y = F(x), \quad (2.16)$$

也就是说，满足 0 到 1 之间均匀分布的新随机变量就是变量 x 的累积分布。对任意累积分布函数 $y = F(x)$ ，可以得到

$$\frac{dy}{dx} = \frac{d}{dx} \int_{-\infty}^x f(x') dx' = f(x), \quad (2.17)$$

所以，由式 (1.32)，我们可以得到 y 的概率密度函数为

$$g(y) = f(x) \left| \frac{dx}{dy} \right| = f(x) \left| \frac{dy}{dx} \right|^{-1} = 1, \quad (0 \leq y \leq 1). \quad (2.18)$$

第三章介绍蒙特卡洛方法时将用到均匀分布的这一性质。

作为均匀分布的例子，可以考虑中性 π 介子衰变 $\pi^0 \rightarrow \gamma\gamma$ 过程中，光子 (γ) 的能量分布。在实验室系中，任何光子的能量 E_{γ} 都均匀分布于 $E_{\min} = \frac{1}{2}E_{\pi}(1 - \beta)$ 和 $E_{\max} = \frac{1}{2}E_{\pi}(1 + \beta)$ 之间，其中 E_{π} 为 π 介子的能量， $\beta = v/c$ 为 π 介子的速度除以光速。

2.4 指数分布

连续随机变量 $x (0 \leq x < \infty)$ 的**指数 (exponential)** 概率密度函数为

$$f(x; \xi) = \frac{1}{\xi} e^{-x/\xi}. \quad (2.19)$$

指数分布的概率密度函数由单个参数 ξ 描述。 x 的期望值为

$$E[x] = \frac{1}{\xi} \int_0^{\infty} x e^{-x/\xi} dx = \xi, \quad (2.20)$$

x 的方差为

$$V[x] = \frac{1}{\xi} \int_0^{\infty} (x - \xi)^2 e^{-x/\xi} dx = \xi^2. \quad (2.21)$$

指数分布的一个例子是，质心系中测量不稳定粒子的衰变时间。参数 ξ 则表示平均寿命，通常用 τ 表示。参数 ξ 取不同值的指数分布如图2.4所示。

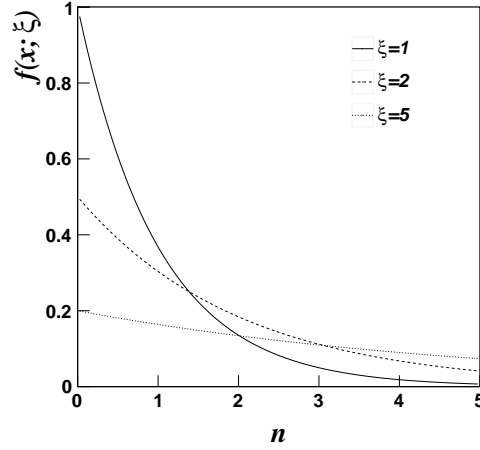


图 2.4: 参数 ξ 不同取值时指数分布的概率密度函数。

2.5 高斯分布

连续随机变量 $x(-\infty \leq x \leq \infty)$ 的高斯 (Gaussian)(又称正态 (normal)) 概率密度函数为

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad (2.22)$$

其中包含两个参数 μ 和 σ^2 。两个参数的名称直接来自 x 的均值和方差，计算可得

$$E[x] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx = \mu, \quad (2.23)$$

$$V[x] = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx = \sigma^2. \quad (2.24)$$

需要注意的是，在 (1.39) 和 (1.43) 式的定义中，我们用 μ 和 σ^2 表示所有概率密度函数的均值和方差，而不仅仅局限于高斯分布。同时还要注意，既可以用 σ 也可以用 σ^2 作为参数，二者是等价的。参数 μ 和 σ^2 不同取值组合时的高斯分布如图2.5所示。

高斯分布的一个特例非常重要，有必要另行表示。即，当 $\mu = 0$, $\sigma = 1$ 时，定义**标准高斯 (standard Gaussian)** 分布的概率密度函数 $\varphi(x)$ 为

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \quad (2.25)$$

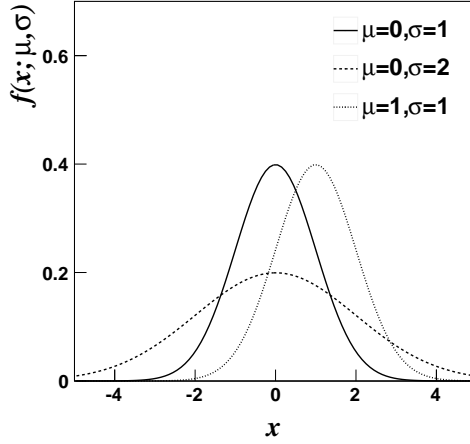


图 2.5: 参数 μ 和 σ 在不同取值组合时高斯分布的概率密度函数。

对应的累积分布 $\Phi(x)$ 为

$$\Phi(x) = \int_{-\infty}^x \varphi(x') dx'. \quad (2.26)$$

容易证明, 如果 y 服从均值为 μ 方差为 σ^2 的高斯分布, 则变量

$$x = \frac{y - \mu}{\sigma} \quad (2.27)$$

服从标准高斯分布 $\varphi(x)$, 累积分布的关系为 $F(y) = \Phi(x)$ 。累积分布 $\Phi(x)$ 没有解析表达式, 只能数值计算。 $\Phi(x)$ 的值以及分位数 $x_\alpha = \Phi^{-1}(\alpha)$ 在很多参考书中都可以查表得到 (例如 [3, 20, 24]), 也可从计算机程序库中计算得到, 例如 [25] 中的 **FREQ** 和 **GAUSIN** 子程序, 或者 ROOT 程序包 [26] 名字空间 **ROOT::Math** 中的函数 **normal_cdf()** 和 **gaussian_quantile()**。

高斯分布的重要性根源于**中心极限定理(central limit theorem)**。该定理表明, n 个独立的连续随机变量 x_i , 均值和方差分别为 μ_i 和 σ_i^2 , 在 n 趋于无穷大的极限下, 其和服从高斯分布, 均值和方差分别为 $\mu = \sum_{i=1}^n \mu_i$ 和 $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ 。该定理在相当普遍的条件下成立, 不管单个 x_i 的分布形式如何。正是因为中心极限定理, 我们可以将测量误差当作高斯分布来处理, 只要总误差为大量小误差之和 (即没有哪个误差占总误差的绝对贡献)。中心极限定理可以用特征函数证明, 如10.3节所述。

高斯分布的 N 维推广定义为

$$f(\mathbf{x}; \boldsymbol{\mu}, V) = \frac{1}{(2\pi)^{N/2} |V|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right], \quad (2.28)$$

其中 \mathbf{x} 和 $\boldsymbol{\mu}$ 为列矢量, 矢量元素分别为 x_1, \dots, x_N 和 μ_1, \dots, μ_N , \mathbf{x}^T 和 $\boldsymbol{\mu}^T$ 为对应的行矢量, $|V|$ 为 $N \times N$ 对称矩阵 V 的行列式, 矩阵 V 包含 $N(N+1)/2$ 个自由参数。尽管如一维情况一样, V 的命名来自协方差矩阵, 目前我们只把 V 看作高斯分布参数的一个记号。计算可得期望值、方差和协方差分别为

$$\begin{aligned} E[x_i] &= \mu_i \\ V[x_i] &= V_{ii} \\ \text{cov}[x_i, x_j] &= V_{ij}. \end{aligned} \quad (2.29)$$

对二维情形，概率密度函数为

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right] \right\}, \quad (2.30)$$

其中 $\rho = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ 为相关系数。

2.6 对数正态分布

如果连续变量 y 服从均值为 μ 方差为 σ^2 的高斯分布，则 $x = e^y$ 服从**对数正态 (log-normal)** 分布，概率密度函数如下

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp \left(-\frac{(\log x - \mu)^2}{2\sigma^2} \right). \quad (2.31)$$

期望值和方差用参数 μ 和 σ^2 表示为

$$E[x] = \exp(\mu + \frac{1}{2}\sigma^2), \quad (2.32)$$

$$V[x] = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]. \quad (2.33)$$

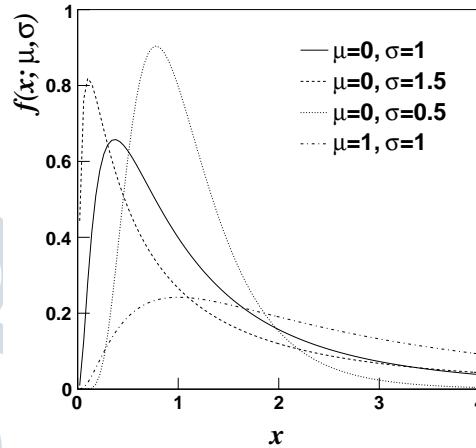


图 2.6: 参数 μ 和 σ 在不同取值时对数正态分布的概率密度函数。

与高斯概率密度函数的情形一样，既可以选择 σ^2 为参数，也可以选择 σ 为参数。然而，需要注意的是，此时 μ 和 σ^2 不是 x 的均值和方差，而是 $\log x$ 所对应的高斯分布的参数。参数 μ 和 σ 的不同取值组合下的对数正态分布如图2.6所示。

回顾上节提到的中心极限定理，如果随机变量 y 为很多小贡献的变量之和，则 y 服从高斯分布。由这个定理可以得到，如果变量 x 为很多变量因子的乘积，则 x 服从对数正态分布。因此，如果随机误差通过相乘因子的方式改变结果，则可以用对数正态分布来处理。

2.7 卡方分布

连续随机变量 $z(0 \leq z < \infty)$ 的卡方 (χ^2) 分布定义为

$$f(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}, \quad n = 1, 2, \dots, \quad (2.34)$$

其中参数 n 称为**自由度的数目 (number of degrees of freedom)**，伽马函数 $\Gamma(x)$ 定义为

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt. \quad (2.35)$$

要计算 χ^2 分布，只需了解伽马函数的以下性质： $\Gamma(n) = (n-1)!$ (其中 n 为整数)， $\Gamma(x+1) = x\Gamma(x)$ ， $\Gamma(1/2) = \sqrt{\pi}$ 。 z 的均值和方差为

$$E[z] = \int_0^\infty z \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} dz = n, \quad (2.36)$$

$$V[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} dz = 2n. \quad (2.37)$$

参数 n 取不同值时 χ^2 分布如图2.7所示。

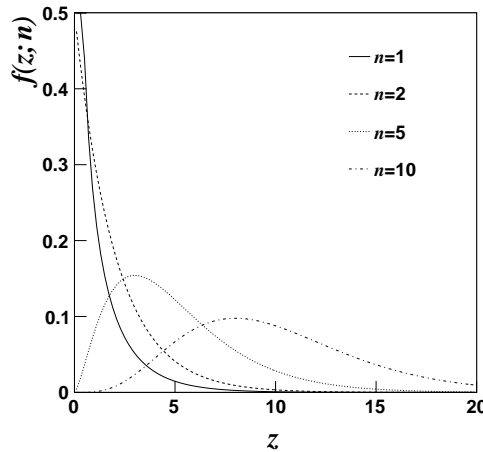


图 2.7: 参数 n 在不同取值时卡方分布的概率密度函数。

卡方分布的重要性在于，该分布与高斯分布随机变量的平方和关系密切。给定 N 个独立的高斯随机变量 x_i ，均值和方差分别为 μ_i 和 σ_i^2 ，则变量

$$z = \sum_{i=1}^N \frac{(x_i - \mu_i)^2}{\sigma_i^2} \quad (2.38)$$

服从自由度为 N 的卡方分布 (见10.2节)。更普遍地，如果 x_i 不独立但服从 N 维高斯分布((2.28) 式)，则变量

$$z = (\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (2.39)$$

服从自由度为 N 的卡方分布。服从卡方分布的随机变量在拟合优度检验中发挥重要作用 (4.5节和4.7节)，尤其是与最小二乘法(7.5)联合使用时更是如此。

2.8 柯西 (布莱特-魏格纳) 分布

连续随机变量 $x (-\infty < x < \infty)$ 的柯西 (Cauchy) 或布莱特-魏格纳 (Breit-Wigner) 概率密度函数定义为

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}. \quad (2.40)$$

这是粒子物理中经常出现的布莱特-魏格纳分布的特殊形式。一般形式为

$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}, \quad (2.41)$$

其中参数 x_0 和 Γ 分别表示共振粒子的质量和宽度。不同参数组合下的柯西分布 (布莱特-魏格纳分布) 如图2.8所示。

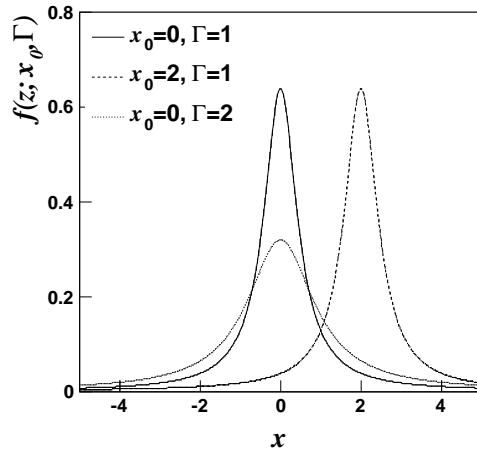


图 2.8: 不同参数 x_0 和 Γ 时, 柯西 (布莱特-魏格纳) 分布的概率密度函数。

柯西分布的期望值没有定义, 因为尽管概率密度函数本身关于零点 (或者 (2.41) 式中的 x_0) 对称, 但积分 $\int_{-\infty}^0 x f(x) dx$ 和 $\int_0^{\infty} x f(x) dx$ 都是发散的。方差和更高阶矩也都发散。不过, 参数 x_0 和 Γ 仍然可以用来给出概率密度函数的位置和宽度信息, 如图2.8所示。 x_0 为峰值的位置, Γ 为峰值处的半高宽。¹

2.9 朗道分布

在核与粒子物理中, 经常会遇到概率密度函数 $f(\Delta; \beta)$, 用来描述带电粒子穿过给定厚度的物质时能量损失 (简称能损) Δ 的分布。这个概率密度函数最早是由朗道 [27] 得到的, 定义为

$$f(\Delta; \beta) = \frac{1}{\xi} \phi(\lambda), \quad 0 \leq x < \infty, \quad (2.42)$$

其中参数 ξ 与物质性质和粒子速度 $\beta = v/c$ (以光速 c 为单位) 有关, $\phi(\lambda)$ 为无量纲随机变量 λ 的概率密度函数。变量 λ 与物质性质、粒子速度 β 以及能损 Δ 有关。这些量的定义为

$$\xi = \frac{2\pi N_A e^4 z^2 \rho \sum Z}{m_e c^2 \sum A} \frac{d}{\beta^2}, \quad (2.43)$$

¹这里采用的是高能物理的标准定义, 其中 Γ 解释为粒子的衰变率。在某些文献中, 例如 [20, 23], 参数 Γ 定义为半高处的半宽度, 相应的概率密度函数要将 (2.41) 式中的 Γ 替换为 2Γ 。

$$\lambda = \frac{1}{\xi} \left[\Delta - \xi \left(\log \frac{\xi}{\epsilon'} + 1 - \gamma_E \right) \right], \quad (2.44)$$

$$\epsilon' = \frac{I^2 \exp(\beta^2)}{2m_e c^2 \beta^2 \gamma^2}, \quad (2.45)$$

其中 N_A 为阿伏伽德罗常数, m_e 和 e 分别为电子的质量和电荷, z 为入射粒子的电荷数 (以电子电荷为单

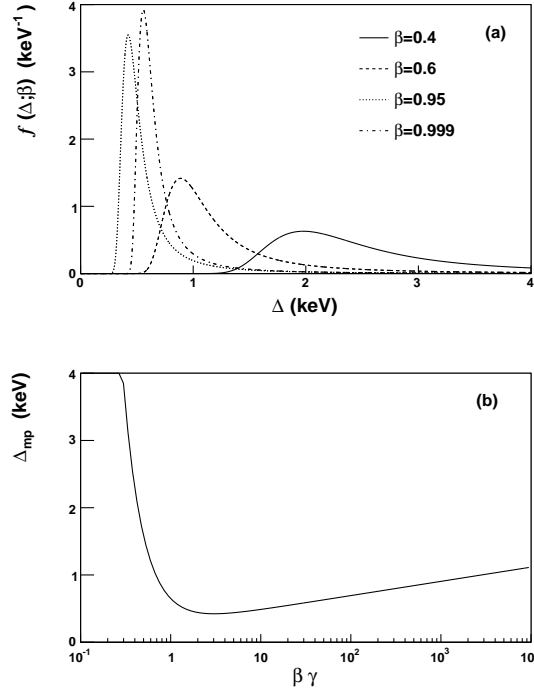


图 2.9: (a) 不同速度 β 的带电粒子穿过 4 mm 厚的氩气层时, 能损 Δ 的朗道分布, 氩气处于正常温度和标准大气压下。(b) 根据 (2.48) 式, (a) 中分布的峰值 (众数) 作为 $\beta\gamma$ 的函数。

位), $\sum Z$ 和 $\sum A$ 分别表示分子物质中原子数和原子量之和, ρ 为物质密度, d 为物质层的厚度, $I = I_0 Z$ 为物质的特征电离能, $I_0 \approx 13.5 \text{ eV}$, 洛伦兹因子 $\gamma = 1/\sqrt{1-\beta^2}$, 欧拉常数 $\gamma_E = 0.5772 \dots$ 。函数 $\phi(\lambda)$ 可以表示为

$$\phi(\lambda) = \frac{1}{2\pi i} \int_{\epsilon-i\infty}^{\epsilon+i\infty} \exp(u \log u + \lambda u) du, \quad (2.46)$$

其中 ϵ 为正无穷小量。或者, 经过变量变换后 $\phi(\lambda)$ 等价于

$$\phi(\lambda) = \frac{1}{\pi} \int_0^\infty \exp(-u \log u - \lambda u) \sin \pi u du. \quad (2.47)$$

该积分只能数值求解, 比如, 利用 **LANDAU** 子程序 [25,28], 或 **ROOT** 程序包 [26] **ROOT::Math** 名字空间中的函数 **landau_pdf()**。图 2.9(a) 给出了不同速度 $\beta = v/c$ 时的能损分布。由于朗道分布在能损 Δ 较高时尾端很长, 所以其均值和高阶矩都不存在。也就是说, $n \geq 1$ 时积分 $\int_0^\infty \Delta^n f(\Delta) d\Delta$ 发散。然而, 从图中可以看出, 能损的最可几值 Δ_{mp} 对粒子速度非常敏感。文献 [29] 数值计算得到最可几值为

$$\Delta_{mp} = \xi [\log(\xi/\epsilon') + 0.198], \quad (2.48)$$

如图 2.9(b) 所示。²

²式 (2.48)(Bethe-Bloch 公式) 是通过测量电离能损鉴别带电粒子的基础, 参见 [30]。

尽管布莱特-魏格纳分布和朗道分布的均值和高阶矩不存在，但实际描述物理过程的概率密度函数必须具有有限矩。例如，如果要多次测量某个特殊系统的能量损失，最终的平均结果将收敛于某个值，因为 Δ 不可能超过入射粒子的能量。类似地，共振态粒子的质量不可能小于质心系中衰变产物的静止质量之和，也不可能超过产生该共振态粒子的反应中质心系总能量。出现发散问题的原因在于，柯西分布和朗道分布都是物理系统的近似模型。在分布的尾端，模型已不再成立，而正是概率密度函数的尾端导致了均值和高阶矩的发散。

DRAFT ZV

第三章 蒙特卡罗方法

蒙特卡罗 (Monte Carlo) 方法是利用随机数序列计算概率及相关量的数值方法。对于单个随机变量的情形, 计算过程可以分成以下步骤。首先, 产生 $(0, 1)$ 区间内均匀分布的随机数序列 r_1, r_2, \dots , 即概率密度函数 $g(r)$ 为

$$g(r) = \begin{cases} 1 & 0 < r < 1, \\ 0 & \text{其它}. \end{cases} \quad (3.1)$$

其次, 利用随机数序列 r_1, r_2, \dots , 确定另外一个随机数序列 x_1, x_2, \dots , 使得 x 服从我们感兴趣的概率密度函数 $f(x)$ 。于是, x 的值可当作模拟测量值处理, 并由此估计 x 处于某个区域的概率。在这个过程中, 我们实际上计算了 $f(x)$ 的积分。这个例子看起来可能平庸之极, 因为我们已经知道了 $f(x)$ 的具体形式, 完全可以直接在关心的区间内进行积分。实际上, 蒙特卡罗方法的优势在多维问题中才能够得到真正的体现。在多维问题中, 在复杂区域内对联合概率密度函数 $f(x, y, z, \dots)$ 进行积分, 几乎没有任何其它办法真正可行, 除了蒙特卡罗积分。

3.1 均匀分布的随机数

为了产生服从均匀分布的随机数序列, 原则上可以利用随机的物理过程, 比如重复投掷硬币。然而, 实际上, 我们都是利用称作**随机数产生子(random number generator)**的计算机算法实现的。很多程序可以实现这样的算法, 用户只需要调用子程序即可 (例如子程序 **RANMAR** [31]、**RANLUX** [32, 33], 这些都是 [25] 的子程序; 或者, 利用 **ROOT** 程序包 [26] 中产生随机数的类 (如 **TRandom3**) 的成员函数 **Rndm()**、**Uniform()** 等)。关于随机数产生子的细节问题超出了本书的范畴, 感兴趣的读者可以参阅 [3, 34], 其中有更完整的讨论。我们在这里只介绍一种简单而有效的算法, 用以阐明产生子的概念。

常用的一种随机数产生子基于**乘法线性同余 (Multiplicative linear congruential)** 算法。从初始的某个整数 n_0 (称为**种子(seed)**) 出发, 根据下面的关系式产生一个整数序列 n_1, n_2, \dots

$$n_{i+1} = (an_i) \bmod m. \quad (3.2)$$

其中**乘子 (multiplier)** a 和**模数 (modulus)** m 均为整型常数, 模算符 (**mod**) 表示取 an_i 除以 m 的余数。 n_i 的值为 $[1, m-1]$ 区间内的一个周期序列。为了得到 $(0, 1)$ 之间服从均匀分布的值, 需要作以下变换

$$r_i = n_i / m. \quad (3.3)$$

注意, 0 和 1 在这里被排除在序列之外, 而在某些其它算法中, 边界值 0 和 1 可以包含在内。初值 n_0 和两个常数 a 和 m 决定了整个序列, 这当然不是真正的随机, 而是严格确定的。更严格地讲, 这样产生的结果应该称为**伪随机数(pseudorandom)**。实际上, 几乎所有的应用中伪随机数与真随机数可以等价处理, 唯一的区别在于伪随机数是可重复的, 例如, 如果用相同的随机数种子, 则伪随机数序列相同。

如何选取 m 和 a 的值呢? 其要求是, 产生的随机数序列在各种随机性检验下都有良好的随机性。在这些要求中, 最重要的是序列重复前的周期要长, 因为一旦随机数序列开始重复, 实际上已经视之为随机数了。此外, 我们还希望得到的随机数中两个随机数可能存在的关联最小。例如, 对于 32-位的整数实现, 已经证明取 $m = 2147483399$ 和 $a = 40692$ 时的结果非常好, 最大周期为 $m - 1 \approx 2 \times 10^9$ [35]。更复杂的算法可以提供更长的周期, 例如 **RANMAR** 产生子 [25,31] 的周期大约为 10^{43} , **ROOT** 程序包 [26] 中类 **TRandom3** 提供的随机数产生子周期可达 10^{6000} 左右。

3.2 变量变换法

假设已经给定在 $[0, 1]$ 区间均匀分布的随机数序列 r_1, r_2, \dots , 下一步就要确定随机数序列 x_1, x_2, \dots , 使其服从我们感兴趣的概率密度函数 $f(x)$ 。在变量变换法中, 这一步是这样实现的: 寻找合适的函数 $x(r)$, 使得均匀分布的 r 代入函数得到的数值 $x(r)$ 服从所需要的分布。这显然是1.4节讨论的变量变换问题。在1.4节中, 给定随机变量 x 的概率密度函数 $f(x)$ 和函数 $a(x)$, 可以求出函数 a 的概率密度函数 $g(a)$ 。这里的任务是, 给定 r 服从 0 到 1 之间的均匀分布, 如何求出函数 $x(r)$, 使其服从指定的概率密度函数 $f(x)$ 。

r 处于区间 $[r, r + dr]$ 的概率为 $g(r)dr$, 这应当等于 x 处于对应区间 $[x(r), x(r) + dx(r)]$ 的概率, 而后的概率显然为 $f(x)dx$ 。为了确定函数 $x(r)$ 使其满足这两个概率相等的条件, 可以要求 r 小于某个值 r' 的概率等于 x 小于 $x(r')$ 的概率。(从后面给出的例子, 我们将看到这个要求不唯一。) 也就是说, 我们需要找到一个函数 $x(r)$, 使其满足 $F(x(r)) = G(r)$, 其中 F 和 G 分别为概率密度函数 f 和 g 对应的累积分布。由于均匀分布概率密度函数的累积分布为 $G(r) = r$, 其中 $0 \leq r \leq 1$, 可以得到

$$F(x(r)) = \int_{-\infty}^{x(r)} f(x') dx' = \int_{-\infty}^r g(r') dr' = r. \quad (3.4)$$

(3.4) 式等效于说, 如果把累积分布函数 $F(x)$ 当作随机变量, 则它服从 0 到 1 之间的均匀分布 (参见 (2.18) 式)。

取决于 $f(x)$ 的具体形式, 方程 (3.4) 有可能可以解析求解出 $x(r)$, 也可能无法解析求解。考虑2.4节介绍的指数分布, (3.4) 式变为

$$\int_0^{x(r)} \frac{1}{\xi} e^{-x'/\xi} dx' = r. \quad (3.5)$$

积分并求解出 x 为

$$x(r) = -\xi \log(1 - r). \quad (3.6)$$

如果变量 r 服从 0 到 1 之间的均匀分布, 则 $r' = 1 - r$ 显然也服从 0 到 1 之间的均匀分布, 即从分布的角度说, r 与 $1 - r$ 等价。所以函数

$$x(r) = -\xi \log(r) \quad (3.7)$$

也满足我们的要求。换言之, 如果 r 服从 0 到 1 之间的均匀分布, 则 $x(r) = -\xi \log r$ 服从均值为 ξ 的指数分布。

3.3 舍选法

实际上, 在很多实际应用中, 要通过 (3.4) 式解析求解出 $x(r)$ 非常困难, 此时需要其他方法产生我们期望的随机分布。一个非常有用的方法是冯·诺依曼 (von Neumann) 的舍选法 (acceptance-rejection method) [36]。考虑如图3.1所示的概率密度函数 $f(x)$, $f(x)$ 可以完全包围在由 x_{\min} 、 x_{\max} 和高度 f_{\max} 确定的方框中。可以通过下面的算法产生服从 $f(x)$ 分布的随机数序列:

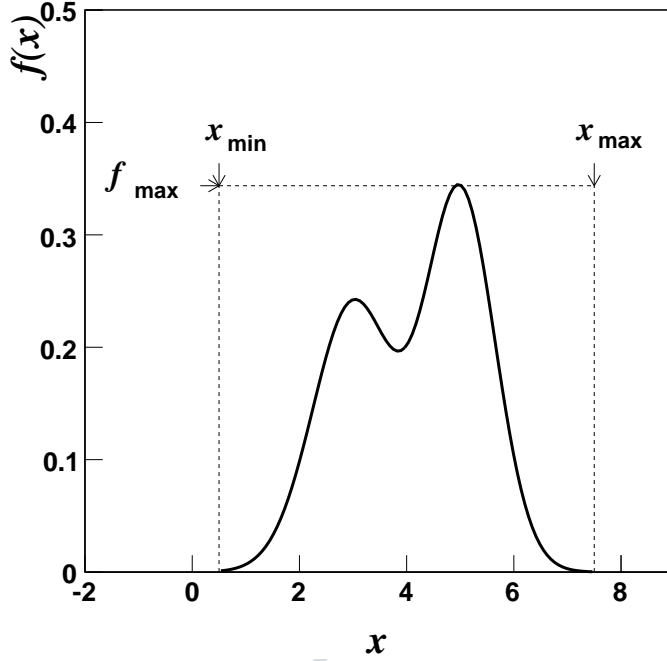


图 3.1: 概率密度函数包围在方框之内, 利用舍选法产生服从该概率密度函数的随机数。

1. 产生随机数 x , x 服从 x_{\min} 到 x_{\max} 之间的均匀分布, 即 $x = x_{\min} + r_1(x_{\max} - x_{\min})$, 其中 r_1 服从 0 到 1 之间的均匀分布。
2. 产生第二个独立的随机数 u , u 服从 0 到 f_{\max} 之间的均匀分布, 即 $u = r_2 f_{\max}$, 其中 r_2 为 0 到 1 之间的均匀分布。
3. 如果 $u < f(x)$, 则接受 x ; 否则, 则拒绝 x 。然后重复上述过程。

被接受的 x 服从概率密度函数 $f(x)$ 的分布, 因为对算法中第 (1) 步得到的每个 x , 其被接受的概率均正比于 $f(x)$ 。

作为例子, 考虑概率密度函数¹

$$f(x) = \frac{3}{8}(1 + x^2), \quad -1 \leq x \leq 1. \quad (3.8)$$

当 $x = \pm 1$ 时, 概率密度函数取最大值 $f_{\max} = 3/4$ 。图 3.2(a) 给出了按照舍选法产生的随机数 u 和 x 的散点分布图。处于曲线下方的 x 值被接受。图 3.2(b) 给出的是被接受的这些 x 值的归一化直方图。

该算法的效率 (即 x 被接受的比率) 等于概率密度函数包围的面积 (该面积为 1) 与方框包围的面积 $f_{\max} \cdot (x_{\max} - x_{\min})$ 之比。如果概率密度函数有很陡的尖峰结构, 则效率可能会非常低, 该算法在实际应用中将变得很慢。如果出现这种情形, 可以对该算法进行改进以提高效率。我们可以用另外一个函数曲线 $g(x)$ 将概率密度函数 $f(x)$ 包围起来, 服从 $g(x)$ 分布的随机数可以根据 $g(x) / \int g(x') dx'$ 产生, 比如采用变量变换法产生随机分布。

于是, 更一般的算法为:

1. 产生随机数 x , 服从概率密度函数 $g(x) / \int g(x') dx'$ 。

¹(3.8) 式表示的是 $e^+e^- \rightarrow \mu^+\mu^-$ 反应中散射角 θ 的分布, 取 $x = \cos \theta$ (参见 [37])。

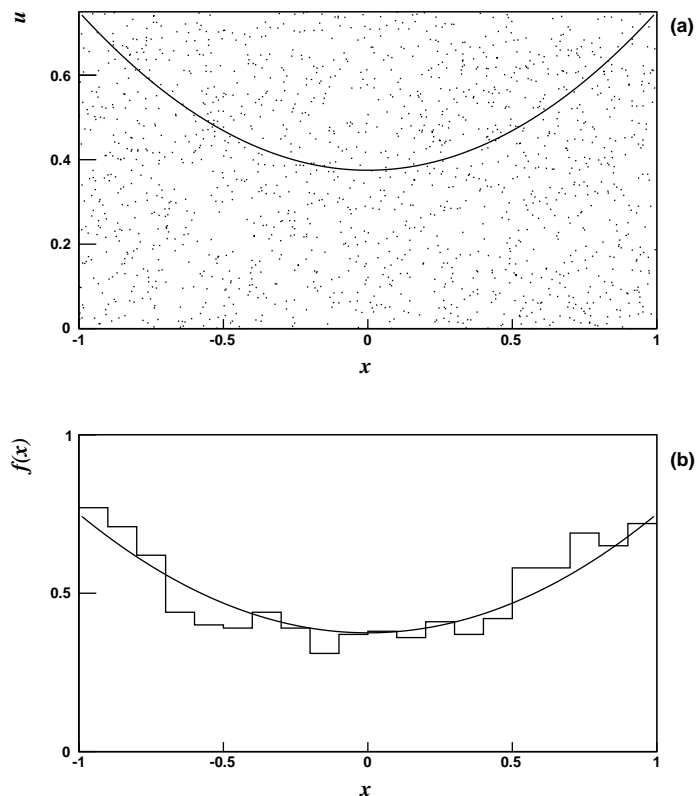


图 3.2: (a) 随机数对 (u, x) 的散点分布, 其中 x 服从 $-1 \leq x \leq 1$ 之间的均匀分布, u 服从 $0 \leq u \leq f_{\max}$ 之间的均匀分布。曲线下方的 x 值被接受。(b) 被接受的 x 的直方图 (已归一化) 和对应的概率密度函数。

2. 产生第二个独立的随机数 u , u 服从 0 到 $g(x)$ 之间的均匀分布。
3. 如果 $u < f(x)$, 则接受 x ; 否则, 则拒绝 x 。然后重复上述过程。

这里, 第 (1) 步产生 x 的概率正比于 $g(x)$, 第 (3) 步之后该值被保留的概率等于 $f(x)/g(x)$, 所以得到 x 的总概率正比于 $f(x)$, 这正是我们需要的。

3.4 蒙特卡罗方法的应用

只要问题的解可以与概率分布的某个参数联系起来, 这个问题就可以采用蒙特卡罗方法解决。这种联系既可以是概率密度函数中明显的参数, 也可以是概率分布在某个区域内的积分。蒙特卡罗产生的随机数序列可以象真实数据一样, 用来计算参数的估计量(或积分)。构造估计量的方法将在第五到第八章讨论。

恰当的估计量必须具备的一个重要特点是, 估计量的统计精度随着数据样本 (利用蒙特卡罗方法或其它手段获得) 容量 n 的增大而提高。可以证明, 在相当普遍的条件下, 估计量的标准差反比于 \sqrt{n} (见 6.6 节)。因此, 用蒙特卡罗方法进行数值积分, 精度随 $1/\sqrt{n}$ 提高。

可以将这个标准与利用梯形法计算积分的要求进行比较。梯形法中, 精度随 $1/n^2$ 提高, 提高速度远远快于蒙特卡罗方法。然而, 梯形法的精度提高依赖于积分的维度, 对于 d 维积分, 梯形法的精度随 $1/n^{2/d}$ 提高。而蒙特卡罗方法不论维度多少, 都随 $1/\sqrt{n}$ 提高。所以, 当 $d > 4$ 时, 蒙特卡罗方法的精度对 n 的

依赖更好。对于其它积分方法，比如高斯积分法，收敛性会比梯形法稍好。然而，如果维度非常高，蒙特卡罗方法总是最优的。细致的讨论详见文献 [38]。

蒙特卡罗方法提供了确定随机变量概率密度函数的一种方法。例如，假设有 n 个独立的随机变量 x_1, \dots, x_n ，服从的概率密度函数 $f_1(x_1), \dots, f_n(x_n)$ 已知，我们想计算某个函数 $a(x_1, \dots, x_n)$ (函数形式可能非常复杂) 服从的概率密度函数 $g(a)$ 。1.4节介绍的方法通常只能用于相对简单的函数，并且变量数目比较少。用蒙特卡罗方法，每个 x_i 的值都根据相应的概率密度函数 $f_i(x_i)$ 产生，然后计算 $a(x_1, \dots, x_n)$ 的值并记录下来 (例如记录到直方图中)。重复该过程，直到获得足够数量的 a ，按照要求的统计精度估计概率密度函数 $g(a)$ 的性质 (例如均值、方差等)。后面的章节有很多应用实例。

蒙特卡罗方法经常用于模拟实验数据。例如，在粒子物理中，模拟通常分成两个阶段：事例产生和探测器模拟。考虑某个实验，例如入射粒子 (电子) 经过打靶散射，被探测器探测到。假设某个理论可以预言事例发生的概率对散射角的依赖关系 (即微分散射截面)。首先，我们要构造蒙特卡罗程序，产生不同散射角的末态粒子 (这实际上确定了动量)。这样的程序称为**事例产生子(event generator)**。在高能物理中，事例产生子可以描述各种不同的粒子反应。

事例产生子的输出，即所产生的粒子的动量，则用来作为**探测器模拟程序(detector simulation program)**的输入。由于散射粒子穿过探测器的响应也包含很多随机过程，例如离子化过程、多重库仑散射等等，探测器响应程序也用蒙特卡罗方法实现。例如 GEANT [25] 程序包可以用来描述复杂的探测器几何，实验合作组通常花费相当的精力尽可能完整地实现探测器的模型。在投入大量时间和金钱建造真正的仪器设备之前，为了针对特定的物理过程而进行探测器优化，蒙特卡罗模拟尤其重要。

蒙特卡罗方法用于模拟实验数据时，可以简单地把这个过程看作利用计算机实现真正的随机过程。概率可以自然地解释为可重复实验输出的相对频率，而实验只是简单地在计算机上多次重复。然而，蒙特卡罗方法也可以看作对任何与概率有关的问题提供数值求解，求解的结果显然与概率解释无关。例如，用蒙特卡罗方法来实现变量变换，或者只是计算函数的积分。

DRAFT ZV

第四章 统计检验

本章将介绍统计检验理论的一些基本概念。统计检验包含的内容非常广泛，我们先作一般性介绍，之后再针对与粒子物理密切相关的几个方面予以讨论。例如，我们可能会对相互作用产生的粒子(事例)感兴趣，或者对事例中的某个特定粒子感兴趣。统计检验在粒子物理中的一个直接应用是候选粒子或事例的筛选，筛选后的粒子或事例将用于后续分析。这里我们关心的是，如何区分感兴趣的事例(信号)和其它类型的事例(本底)。这些问题将在4.2-4.4节讨论。统计检验另一个重要之处与拟合优度(goodness-of-fit)有关，我们将在4.5-4.7节予以讨论。

4.1 假设、检验统计量、显著性水平和功效

统计检验的目的是表述观测数据与预期概率(即假设(hypothesis))之间的符合程度。待考察的假设通常称为**原假设**或**零假设 (null hypothesis)** H_0 ，它可以确定随机变量 x 的概率密度函数 $f(x)$ 。如果假设可以唯一确定 $f(x)$ ，则称为**简单假设(simple hypothesis)**；如果假设可以确定概率密度函数的形式，但无法完全确定概率密度函数的参数 θ ，则称 $f(x; \theta)$ 为**复合假设(composite hypothesis)**。对于复合假设，可以采用第五章到第八章介绍的方法，利用数据估计未知参数的值。而在本章中，只考虑简单假设。

要表述原假设 H_0 的合理性，通常需要与**备择假设(alternative hypothesis)** H_1, H_2, \dots 进行比较。假定有一组包含 n 个测量值 $\mathbf{x} = x_1, \dots, x_n$ 的实验数据，以及一组假设 H_0, H_1, \dots ，每个假设都能确定一个联合概率密度函数，分别记为 $f(\mathbf{x}|H_0), f(\mathbf{x}|H_1), \dots$ ¹。概率密度函数的值可以表示对同一个随机变量的 n 次重复观测，也可以表示对 n 维随机变量的单次观测。为了研究观测数据与给定假设的符合程度，我们构造一个观测变量的函数 $t(\mathbf{x})$ ，称为**检验统计量(test statistic)**。给定某个假设的条件下，就会得到检验统计量 t 相应的概率密度函数，即 $g(t|H_0), g(t|H_1)$ 等等。

检验统计量 t 可以是多维矢量，即 $\mathbf{t} = (t_1, \dots, t_m)$ 。实际上，我们可以直接把原始的数据矢量 $\mathbf{x} = (x_1, \dots, x_n)$ 当作检验统计量。但是，我们更关心的是低维(即 $m < n$)检验统计量，构造低维检验统计量的关键在于，约化数据量的同时保证不丢失区分各个假设的能力。我们暂时先假设选定的检验统计量为标量函数 $t(\mathbf{x})$ ，并且如果 H_0 为真则其概率密度函数为 $g(t|H_0)$ ，如果 H_1 为真则概率密度函数为 $g(t|H_1)$ ，如图4.1所示。

通常，根据接受或拒绝原假设 H_0 的决定，描述数据与不同假设的匹配性。具体做法是定义 t 的**拒绝域(critical region)**，等效地，可以定义拒绝域的补，称为**接受域(acceptance region)**。如果实际得到的 t 值处于拒绝域，则拒绝 H_0 假设；否则，就接受 H_0 。拒绝域的选择，要求在 H_0 假设下，观测到 t 处于拒绝域的概率等于某个值 α ，其中 α 称为检验的**显著性水平(significance level)**。例如，可以选择拒绝域为 t 大

¹在经典统计中，如果假设 H 指的是可重复实验的输出，那么 H 仅仅被当作一个随机变量。尽管如此，对于给定假设 H 时随机变量 x 的概率密度函数，我们仍然采用条件概率密度函数的符号表示 $f(\mathbf{x}|H)$ (见1.3节)。对贝叶斯统计， \mathbf{x} 和 H 都是随机变量，所以这种符号表示在任何时候都没有问题。

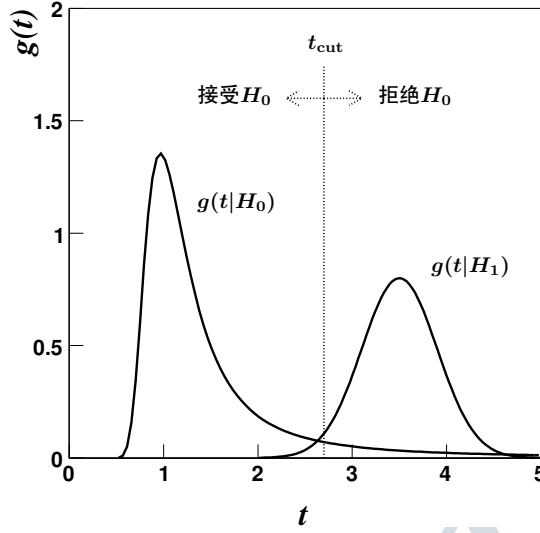


图 4.1: 检验统计量 t 在 H_0 和 H_1 假设下的概率密度。如果观测到 t 处于拒绝域 (这里为 $t > t_{\text{cut}}$)，则拒绝 H_0 。

于某个值 t_{cut} 的区域，如图 4.1 所示，其中 t_{cut} 称为**判选条件(cut)**。于是，显著性水平为

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0) dt. \quad (4.1)$$

因此，如果 t 的观测值小于 t_{cut} ，则接受 (或者严格地说的不拒绝) H_0 假设。显然，如果 H_0 为真，其被拒绝的概率为 α ，这称为**第一类误差(error of the first kind)**。如果 H_0 假设被接受 (即 t 的观测值小于 t_{cut})，但真假设不是 H_0 而是某个备择假设 H_1 ，这称为**第二类误差(error of the second kind)**。第二类误差的概率为

$$\beta = \int_{-\infty}^{t_{\text{cut}}} g(t|H_1) dt. \quad (4.2)$$

其中 $1 - \beta$ 称为区分备择假设 H_1 的**检验功效(power of the test)**。

4.2 粒子选择的统计检验

作为例子，考虑已知动量的带电粒子穿过探测器时发生电离现象，检验统计量 t 可以表示测量到的电离值。电离的大小随每个粒子而涨落，并依赖于粒子的质量 (在动量固定的情况下)。因此，图 4.1 中的概率密度函数 $g(t|H_0)$ 可以表示粒子为电子的假设，而 $g(t|H_1)$ 表示粒子为 π 介子的假设，即 $H_0 = e$ ， $H_1 = \pi$ 。

假设已知要研究的粒子要么是电子，要么是 π 介子，并且我们希望根据判选条件 $t \leq t_{\text{cut}}$ 选出电子样本。(电子被当作“信号”， π 介子被当作“本底”。) 接受某给定类型粒子的概率，即**选择效率(selection efficiencies)** ϵ_e 和 ϵ_π 分别为

$$\epsilon_e = \int_{-\infty}^{t_{\text{cut}}} g(t|e) dt = 1 - \alpha, \quad (4.3)$$

$$\epsilon_\pi = \int_{-\infty}^{t_{\text{cut}}} g(t|\pi) dt = \beta. \quad (4.4)$$

对于每个效率，我们都可以通过选择合适的拒绝域，使选择条件变得更松或更严格，从而导致效率任意接近于 0 或 1。然而，这是有代价的，提高信号的选择效率将增加样本中杂质的数量，即电子样本的纯度由于某些 π 介子进入选择域而降低。

如果 π 介子和电子的相对比例未知，这个问题就变成了参数估计问题 (第五到八章)。也就是说，检验统计量 t 服从的分布为

$$f(t; a_e) = a_e g(t|e) + (1 - a_e) g(t|\pi), \quad (4.5)$$

其中 a_e 和 $a_\pi = 1 - a_e$ 分别为电子和 π 介子的比率。估计出 a_e 则可以得到原始样本 N_{tot} 个粒子中电子的数目 N_e ，即 $N_e = a_e N_{\text{tot}}$ 。

我们也可能希望通过要求 $t \leq t_{\text{cut}}$ 选择出电子候选者的子样本，使得 N_{tot} 个粒子中的 n_{acc} 个被接受。此时，我们经常感兴趣的是，在施加 t 的判选之前一共有多少个电子。接受粒子的数目为²

$$\begin{aligned} N_{\text{acc}} &= \epsilon_e N_e + \epsilon_\pi N_\pi \\ &= \epsilon_e N_e + \epsilon_\pi (N_{\text{tot}} - N_e), \end{aligned} \quad (4.6)$$

从而得到

$$N_e = \frac{N_{\text{acc}} - \epsilon_\pi N_{\text{tot}}}{\epsilon_e - \epsilon_\pi}. \quad (4.7)$$

从 (4.7) 式可以看出，只有效率 ϵ_e 和 ϵ_π 不同时，才可以用 N_{acc} 确定电子数目 N_e 。如果 ϵ_e 和 ϵ_π 存在不确定度，则该不确定度将通过1.6节介绍的误差传递导致 N_e 的不确定度。实际应用中，我们往往可能希望以这样的方式确定拒绝域 (即确定电离值的判选值)，即令 N_e 的总误差最小。

观测值为 t 时，粒子为电子或 π 介子的概率 $h(e|t)$ 和 $h(\pi|t)$ ，可以利用贝叶斯定理(1.8) 以及概率密度函数 $g(t|e)$ 和 $g(t|\pi)$ 得到，

$$h(e|t) = \frac{a_e g(t|e)}{a_e g(t|e) + a_\pi g(t|\pi)}, \quad (4.8)$$

$$h(\pi|t) = \frac{a_\pi g(t|\pi)}{a_e g(t|e) + a_\pi g(t|\pi)}, \quad (4.9)$$

其中 a_e 和 $a_\pi = 1 - a_e$ 为假设 e 和假设 π 的验前概率。因此，为了给出被选择的粒子为电子的概率，我们需要知道所有可能假设的验前概率以及检验统计量 t 在这些假设下的概率密度函数。

尽管从本质上讲这是贝叶斯方法处理的问题，(4.8) 和 (4.9) 式在经典统计的框架下也是有意义的。如果我们处理的粒子样本非常大，则假设 $H = e$ 和 $H = \pi$ 涉及到随粒子的不同而随机变化的某种特性。这种情况下，用相对频率解释， $h(e|t)$ 是 t 给定时粒子为电子的比率。在贝叶斯统计中，利用主观概率，我们说 $h(e|t)$ 表示测量值为 t 时我们对给定粒子是电子的信心程度。

相比于给定某个粒子为电子的概率，我们可能对**纯度(purity)** p_e 更感兴趣，即根据 $t < t_{\text{cut}}$ 选择出来的电子候选者样本中电子的比率。纯度由下式给出

$$\begin{aligned} p_e &= \frac{t < t_{\text{cut}} \text{ 的电子数}}{t < t_{\text{cut}} \text{ 的所有粒子数}} \\ &= \frac{\int_{-\infty}^{t_{\text{cut}}} a_e g(t|e) dt}{\int_{-\infty}^{t_{\text{cut}}} (a_e g(t|e) + (1 - a_e) g(t|\pi)) dt} \\ &= \frac{a_e \epsilon_e N_{\text{tot}}}{N_{\text{acc}}}. \end{aligned} \quad (4.10)$$

利用 (4.8) 式可以得到，这正是电子概率 $h(e|t)$ 在 $(-\infty, t_{\text{cut}}]$ 区间内的平均值。即，

$$p_e = \frac{\int_{-\infty}^{t_{\text{cut}}} h(e|t) f(t) dt}{\int_{-\infty}^{t_{\text{cut}}} f(t) dt}. \quad (4.11)$$

²严格而论，(4.6) 和 (4.7) 式给出的是粒子数的期望值，而不是某次实验得到的有限数据样本一定会选择出的粒子数目。个中差异将在第五章详细讨论。

4.3 用奈曼-皮尔逊引理选择拒绝域

到目前为止，我们还没有谈到拒绝域的具体选择，即 t_{cut} 的值到底应该取多大。拒绝域的选择依赖于进一步分析对所选粒子(事例)的效率和纯度有何要求。判选条件的一种优化选择是，要求在给定效率的条件下，使得所选样本的纯度最高。(仍然没有确定需要的效率是多少。)对于上节一维检验统计量的例子，这个优化自动满足。在这个例子中，只需要一个判选条件 t_{cut} ，这个判选条件同时确定了效率和纯度。

然而，如果选择的检验统计量是多维的，即 $\mathbf{t} = (t_1, \dots, t_m)$ ，则拒绝域或接受域的定义就不那么明显了。假定我们想检验某个简单原假设 H_0 ，例如，选择某种给定类型的事例。假定备择假设 H_1 也是简单假设，即除了感兴趣的事例(信号)，还存在本底事例。因此，选出的样本中信号纯度通常达不到 100%。

奈曼-皮尔逊引理(Neyman-Pearson lemma) 声称，在给定显著性水平 α (或者给定选择效率 $\epsilon = 1 - \alpha$) 的条件下，使功效最高(等价于信号纯度最高)的接受域为 \mathbf{t} -空间中满足下面条件的区域：

$$\frac{g(\mathbf{t}|H_0)}{g(\mathbf{t}|H_1)} > c. \quad (4.12)$$

其中常数 c 由期待的效率确定。该引理的证明详见文献 [3]。需要注意的是，对于矢量形式的检验统计量 \mathbf{t} ，利用奈曼-皮尔逊引理确定接受域，实际上等价于由 (4.12) 式等号左边的比值给出的一维检验统计量：

$$r = \frac{g(\mathbf{t}|H_0)}{g(\mathbf{t}|H_1)}. \quad (4.13)$$

这个比值称为简单假设 H_0 和 H_1 的似然比(likelihood ratio)。显然，这个一维检验统计量相应的接受域由 $r > c$ 给出。

4.4 构造检验统计量

假设已知数据矢量 $\mathbf{x} = (x_1, \dots, x_n)$ ，并且我们希望以此构造一维检验统计量 $t(\mathbf{x})$ ，用来区分简单假设 H_0 和 H_1 。在上一节我们看到，在给定显著性水平(或选择效率)的条件下，最大功效(等价于最大信号纯度)意义上的最佳检验统计量由似然比给出：

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)}. \quad (4.14)$$

然而，为了构造这个检验统计量，我们需要知道 $f(\mathbf{x}|H_0)$ 和 $f(\mathbf{x}|H_1)$ 。通常，这必须通过对两种类型事例进行蒙特卡罗模拟得到，相应的概率密度用多维直方图表示。如果对 \mathbf{x} 的每个分量(共 n 个)都分成 M 个区间，则总的区间数为 M^n 。这意味着需要从蒙特卡罗数据确定 M^n 个参数。当 n 很大时，这个方法实际上完全不可行，因为需要的数据量极为庞大。

即使无法用 n 维直方图确定 $f(\mathbf{x}|H_0)$ 和 $f(\mathbf{x}|H_1)$ ，我们仍然可以对 $t(\mathbf{x})$ 的函数形式做一个简单的拟设(Ansatz)，然后根据某种判据选择一个具有拟设函数形式的最佳函数。我们将分别考虑 x_i 的线性和非线性函数形式，同时阐明一些与输入变量的选择有关的实际考虑。

4.4.1 线性检验统计量、费舍尔甄别函数

检验统计量 $t(\mathbf{x})$ 最简单的形式是线性函数，

$$t(\mathbf{x}) = \sum_{i=1}^n a_i x_i = \mathbf{a}^T \mathbf{x}, \quad (4.15)$$

其中 \mathbf{a}^T 为系数的转置矢量 (即行矢量)。我们的目标是选择合适的系数 a_i ，使之可以让两个概率密度函数 $g(t|H_0)$ 和 $g(t|H_1)$ 尽可能分开。对分离程度的不同定义将导致不同的确定系数的规则。费舍尔 (Fisher) [47] 基于以下考虑最早发展出一种方法。数据 \mathbf{x} 的均值和协方差矩阵为

$$\begin{aligned}(\mu_k)_i &= \int x_i f(\mathbf{x}|H_k) d\mathbf{x}_1 \dots d\mathbf{x}_n, \\(V_k)_{ij} &= \int (x_i - \mu_k)_i (x_j - \mu_k)_j f(\mathbf{x}|H_k) d\mathbf{x}_1 \dots d\mathbf{x}_n,\end{aligned}\tag{4.16}$$

其中指标 $i, j = 1, \dots, n$ 用来表示矢量 \mathbf{x} 的分量, $k = 0, 1$ 分别代表假设 H_0 或 H_1 。与此相对应, 每个假设也可以由 t 的均值和方差表征,

$$\begin{aligned}\tau_k &= \int t g(t|H_k) dt = \mathbf{a}^T \boldsymbol{\mu}_k, \\ \Sigma_k^2 &= \int (t - \tau_k)^2 g(t|H_k) dt = \mathbf{a}^T V_k \mathbf{a}.\end{aligned}\tag{4.17}$$

显然, 为了增大分离度, 我们需要尝试最大化 $|\tau_0 - \tau_1|$ 。此外, 我们还希望 $g(t|H_0)$ 和 $g(t|H_1)$ 尽量集中在各自的均值 τ_0 和 τ_1 附近, 这由方差 Σ_0^2 和 Σ_1^2 决定。同时考虑这两点的分离度可以由下式给出

$$J(\mathbf{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}.\tag{4.18}$$

将分子表示为 a_i 的函数, 可以得到

$$\begin{aligned}(\tau_0 - \tau_1)^2 &= \left(\sum_{i,j=1}^n a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j \right)^2 \\ &= \sum_{i,j=1}^n a_i a_j B_{ij} = \mathbf{a}^T \mathbf{B} \mathbf{a},\end{aligned}\tag{4.19}$$

其中矩阵 B 定义为

$$B_{ij} = (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j \quad i, j = 1, \dots, n,\tag{4.20}$$

表示与 H_0 和 H_1 对应的两个类别 ‘之间’(between) 的分离。用类似方法, (4.18) 式的分母可以化为

$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^n a_i a_j (V_0 + V_1)_{ij} = \mathbf{a}^T \mathbf{W} \mathbf{a},\tag{4.21}$$

其中 $W_{ij} = (V_0 + V_1)_{ij}$ 表示两个类别 ‘内’(within) 的协方差之和。因此, (4.18) 式的分离度变为

$$J(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}}.\tag{4.22}$$

令 $J(\mathbf{a})$ 对 a_i 的导数等于零, 可以求出, 使分离度最大的系数 \mathbf{a} 满足

$$\mathbf{a} \propto W^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1).\tag{4.23}$$

这些系数只确定到相差一个任意的标度因子。基于 (4.15) 和 (4.23) 式的检验统计量称为**费舍尔线性判别函数(Fisher linear discriminant function)**。为了确定系数 a_i , 需要知道矩阵 W 和期望值 $\boldsymbol{\mu}_{0,1}$, 这通常会利用一组训练数据 (training data) 进行估计, 比如利用蒙特卡罗模拟来估计。这个方法的关键点在于, 我们不需要完全确定联合分布函数 $f(\mathbf{x}|H_0)$ 和 $f(\mathbf{x}|H_1)$ (n 维直方图), 而只要求出 (4.16) 式的均值和协方差即可。也就是说, 训练数据只用来确定 $n(n+1)/2$ 个参数即可。5.2节将介绍如何从数据样本估计均值和协方差。

可以通过对 \mathbf{a} 简单地乘以一个常数，改变变量 t 的标度。我们可以将 t 的定义推广为下面的形式

$$t(\mathbf{x}) = a_0 + \sum_{i=1}^n a_i x_i, \quad (4.24)$$

从而可以利用偏移量 a_0 和标度的任意性，将期望值 τ_0 和 τ_1 固定为任意需要的值。于是，在固定 τ_0 和 τ_1 的条件下，最大化 (4.18) 式等价于最小化方差之和

$$\Sigma_0^2 + \Sigma_1^2 = E_0[(t - \tau_0)^2] + E_1[(t - \tau_1)^2], \quad (4.25)$$

其中 E_0 和 E_1 分别表示 t 在 H_0 假设和 H_1 假设下的期望值。因此，用来确定系数 \mathbf{a} 的判选条件，实际上类似于第七章在参数估计中将遇到的最小化原则，称为最小二乘原则。

值得特别关注的是特殊情形下费舍尔甄别函数的一些性质，即概率密度函数 $f(\mathbf{x}|H_0)$ 和 $f(\mathbf{x}|H_1)$ 都是多维高斯分布，并且具有相同的协方差矩阵 $V_0 = V_1 = V$ ，

$$f(\mathbf{x}|H_k) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T V^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) \right], \quad k = 0, 1. \quad (4.26)$$

此时，(4.24) 式包含偏移量的费舍尔甄别函数可以取为

$$t(\mathbf{x}) = a_0 + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T V^{-1} \mathbf{x}. \quad (4.27)$$

于是，(4.14) 式的似然比变为

$$\begin{aligned} r &= \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^T V^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T V^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \right] \\ &= \exp \left[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T V^{-1} \mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_0^T V^{-1} \boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T V^{-1} \boldsymbol{\mu}_1 \right] \\ &\propto e^t. \end{aligned} \quad (4.28)$$

也就是说， $t \propto \log r + \text{常数}$ ，即 (4.27) 式的检验统计量为似然比 r 的单调函数。因此，在这种情况下，费舍尔甄别函数与似然比具有相同的效果。而且，多维高斯和 $V_0 = V_1 = V$ 的条件会使得这些假设的验后概率形式非常简单。例如，对于假设 H_0 ，利用贝叶斯定理，在给定数据 \mathbf{x} 的条件下，其验后概率为

$$P(H_0|\mathbf{x}) = \frac{f(\mathbf{x}|H_0)\pi_0}{f(\mathbf{x}|H_0)\pi_0 + f(\mathbf{x}|H_1)\pi_1} = \frac{1}{1 + \frac{\pi_1}{\pi_0 r}}, \quad (4.29)$$

其中 π_0 和 π_1 分别为假设 H_0 和 H_1 的验前概率。将该式与 (4.28) 式中 r 的表达式结合起来，可以得到

$$P(H_0|\mathbf{x}) = \frac{1}{1 + e^{-t}} \equiv s(t), \quad (4.30)$$

这要求相应的偏移量 a_0 取为

$$a_0 = -\frac{1}{2}\boldsymbol{\mu}_0^T V^{-1} \boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T V^{-1} \boldsymbol{\mu}_1 + \log \frac{\pi_0}{\pi_1}. \quad (4.31)$$

(4.30) 式的函数 $s(t)$ 是 **S 型函数 (sigmoid function)** 的一个特例，其取值范围为 $(0, 1)$ ，因此也称为**逻辑 S 型 (logistic sigmoid)**。

4.4.2 非线性检验统计量、神经网络

如果联合概率密度函数 $f(\mathbf{x}|H_0)$ 和 $f(\mathbf{x}|H_1)$ 不是高斯分布，或者它们具有不同的协方差矩阵，那么，上节得到的费舍尔甄别函数的这些最佳性质不复存在。此时，可以尝试检验统计量 $t(\mathbf{x})$ 更一般的参数化形

式。这里我们将介绍的函数形式是**人工神经网络(artificial neural network)**的一个特例。近年来神经网络领域的文献非常多，尤其是与模式识别问题有关的文献。限于篇幅，我们将只介绍神经网络的一些主要思想，感兴趣的读者请参阅 [39–41] 等文献，这些文献对神经网络的介绍更完整。如果对神经网络在粒子物理中的应用感兴趣，可以参考 [42–44] 等文献。

假设我们取 $t(\mathbf{x})$ 为

$$t(\mathbf{x}) = s \left(a_0 + \sum_{i=1}^n a_i x_i \right). \quad (4.32)$$

函数 $s(\cdot)$ 通常称为**激活函数(activation function)**，这里我们将取为逻辑 S 型。激活函数可以取不同的形式，例如阶跃函数。激活函数的自变量为输入变量的线性函数，包括偏移量 a_0 (通常称为阈值)。这种形式的检验统计量通常称为**单层感知器(single-layer perceptron)**。

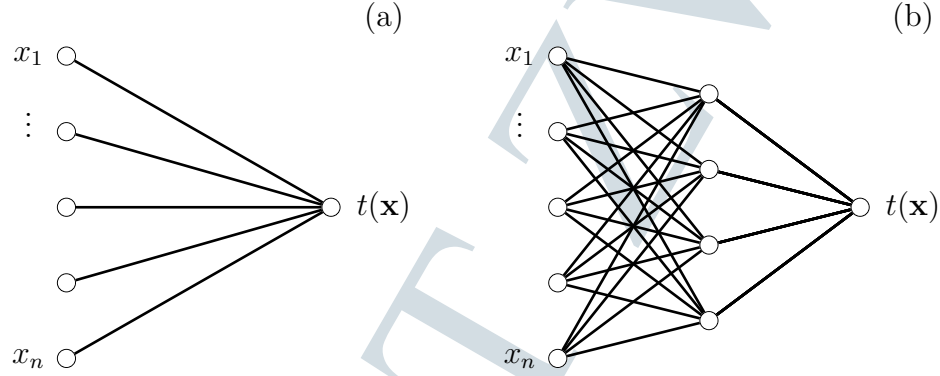


图 4.2: 两个前馈神经网络的结构示意图: (a) 单层感知器的神经网络; (b) 双层感知器的神经网络。

单层感知器的结构如图 4.2(a) 所示。输入值 \mathbf{x} 用一组节点表示，它们一起构成所谓的**输入层(input layer)**。最终的检验统计量 $t(\mathbf{x})$ 由**输出节点(output node)** 给出。一般来说，可以考虑多个输出节点，这对应于矢量形式的检验统计量。与上节一样，我们在这里只考虑单输出节点。

由于 S 型函数的单调性，单层感知器等价于线性检验统计量。然而，我们可以将其推广到**双层感知器(two-layer perceptron)**，如图 4.2(b) 所示。在双层感知器中，除了输入层和输出层，还存在**隐藏层(hidden layer)**，隐藏层包含 m 个节点 (h_1, \dots, h_m) 。在双层感知器中， h_i 担负着单层感知器中输入变量的角色，因此， t 变为

$$t(\mathbf{x}) = s \left(a_0 + \sum_{i=1}^m a_i h_i(\mathbf{x}) \right). \quad (4.33)$$

而 h_i 本身是上层节点 (对双层感知器而言上层节点即输入层) 的函数，

$$h_i(\mathbf{x}) = s \left(\omega_{i0} + \sum_{j=1}^n \omega_{ij} x_j \right). \quad (4.34)$$

很容易将双层感知器推广到包含任意多的隐藏层 (即多层感知器)。如图 4.2 所示，我们通常限制节点之间的链接，使得给定节点的值仅依赖于上层节点，这称为**前馈网络(feed-forward network)**。如果输入变量为 n 个，后面的隐藏层依次包含 m_1, m_2, m_3, \dots 个节点，则自由参数的数目为 $(n+1)m_1 + (m_1+1)m_2 + (m_2+1)m_3 + \dots$ 。(注意，这些参数已经包含了 a_0, ω_{i0} 等。) 参数 a_i, ω_{ij} 等称为**权重(weights)** 或**链接强度(connection strengths)**。

如果给定 (4.14) 式的似然比，或者等价地给定似然比的单调函数，通过引入更多的参数，我们可以更好地近似最优检验统计量。于是问题变成，如何选取合适的参数使得 $t(\mathbf{x})$ 可以对不同假设实现优化分离。

但是，这时参数的选择已经不像线性情形那样直接，它们可以与 \mathbf{x} 的均值和协方差有关。典型的做法是，按照最小化**误差函数(error function)**的原则来参数优化，例如

$$\epsilon = E_0[(t - t^{(0)})^2] + E_1[(t - t^{(1)})^2], \quad (4.35)$$

这类似于费舍尔甄别函数中对方差之和 (4.25) 进行最小化。然而，这里的 $t^{(0)}$ 和 $t^{(1)}$ 分别表示 H_0 和 H_1 预设的**目标值(target values)**。对于最后一层的逻辑 S 型激活函数，目标值分别取为 0 和 1。

为了确定使 ϵ 最小化的参数值，我们必须采用迭代的数值方法，称为**网络训练(network training)**或**网络学习(network learning)**。在实际应用中，参数的调整涉及到将 (4.35) 式的期望值替换为用训练数据 (例如通过蒙特卡罗模拟得到) 样本计算所得的样本均值。学习算法通常先随机为各个权重参数选定初值，然后利用部分或全部训练数据评估其性能。然后就可以利用各种不同的方法调整权重参数以最小化 ϵ 。比较流行的一种方法是**误差逆传播(error back-propagation)**。对这些方法的详细介绍超出了本书的范畴，感兴趣的读者请参阅 [39–42, 61]。

确定感知器的层数和每一层的节点数 (即网络结构) 需要根据具体问题和可用的训练数据而定。层数越多，则参数越多，从而可以达到更好地区分不同假设的效果。然而，在给定训练数据的情况下，参数越多则参数的优化越困难。可以证明，三层感知器足以对任意函数实现任意好的参数化 [39]。

4.4.3 输入变量的选择

到目前为止，我们用完整的数据矢量 $\mathbf{x} = (x_1, \dots, x_n)$ 作为输入变量构造了检验统计量 $t(\mathbf{x})$ 。其中系数 $\mathbf{a} = (a_1, \dots, a_n)$ (在神经网络中称为权重) 通常用蒙特卡罗模拟的训练数据确定，因此，统计精度必然是有限的。实际应用中，我们更倾向于采用 \mathbf{x} 一部分分量来构造检验统计量，只选择携带所研究假设最重要信息的那些分量。有可能 \mathbf{x} 的某些分量区分不同假设的能力很小，甚至完全没有区分能力，我们可以放心地扔掉这些分量。也可能两个或多个分量高度相关，同时使用这些分量并不提供更多信息，此时选择其中之一即可。一般来说，在给定有限训练数据的情况下，选用较少的输入变量可以更好地确定权重参数。

还需要注意的是，训练数据可能与实际数据存在系统上的差异，例如，蒙特卡罗模拟不可避免地存在一定的近似和不完善性。对区分能力贡献有限的一些变量，可能实际上没有得到很好的模拟。

实际操作中如何从原始输入数据 \mathbf{x} 选择一个子集来构造检验统计量呢？一般来说，可以先选择能最好地区分 H_0 和 H_1 的那个分量构造 t ，对于费舍尔甄别函数，这由 (4.22) 式 $J(\mathbf{a})$ 的值来判断；对于神经网络，由 (4.35) 式的误差函数 ϵ 判断。之后，我们可以将其他分量逐个考虑进来，保证每增加一个分量都可以最大程度提高检验统计量的区分能力。另外一种方法是，开始的时候把所有分量都考虑进来，然后逐个去除某些分量，保证每次去除分量之后检验统计量区分能力的降低最小。任何一种方法都无法保证训练结果最优，文献 [39, 40, 45] 更为详细地介绍了在实际操作中如何考虑输入变量的选择。

选择输入变量时，很重要的一点是，要考虑到统计检验的原始目的，我们往往是想选择属于给定类型的对象 (事例、粒子，等等)，用于进一步的分析研究。这实际上暗含一个事实，即我们对这些对象的性质的理解还不充分，否则就没有必要测量了。在确定输入变量时，必须避免所选择的量与后续分析中要研究的量存在相关性。这经常是一个很重要的限制，因为我们可能还没有很好地理解这些相关性。

4.5 拟合优度检验

有时候我们希望度量，在没有任何特别的备择假设时，给定的原假设 H_0 与观测数据之间的一致性到底有多好。这称为**拟合优度(goodness-of-fit)**检验，通常的做法是构造一个检验统计量，使其值本身即可反

映观测数据与 H_0 的预言结果之间的符合程度。在4.7节、6.11节和7.5节，我们将具体讨论如何构造合适的检验统计量。本节仅利用一个小例子阐明拟合优度检验的主要思想。

假设投掷硬币 N 次，其中 n_h 次正面朝上， $n_t = N - n_h$ 次背面朝上。那么，实验结果 n_h 和 n_t 与硬币‘公平’(即正面和背面的概率相同) 这一假设的符合程度如何呢？我们可以简单地选取正面的次数 n_h 作为检验统计量，对于公平的硬币而言， n_h 应当服从参数 $p = 0.5$ 的二项分布 ((2.2) 式)。换句话说，观测到 n_h 次正面的概率为

$$f(n_h; N) = \frac{N!}{n_h!(N - n_h)!} \left(\frac{1}{2}\right)^{n_h} \left(\frac{1}{2}\right)^{N - n_h}. \quad (4.36)$$

假设投掷次数 $N = 20$ ，并且观测到正面朝上的次数为 $n_h = 17$ 。由于 n_h 的期望值 (2.3式) 为 $E[n_h] = Np = 10$ ，所以期望值与实际观测值之间存在颇大的差异。为了定量描述这种差异的显著程度，我们可以将实际观测值与理论假设之间的差异作为一个参考点，计算重复试验中观测到的差异等于或大于这个参考点的概率。对本例而言，这个概率即 $n_h = 0, 1, 2, 3, 17, 18, 19, 20$ 的概率之和。利用 (4.36) 式，可以得到概率为 $P = 0.0026$ 。

因此，拟合优度检验的结果可以由所谓的 **P -值(P -value)** 给出，即在所研究的 H_0 假设条件下，与当前观测值跟 H_0 的符合程度相比，重复试验得到相同或更差符合程度的概率 P 。 P -值有时也称作检验的**观测显著性水平(observed significance level)** 或**置信水平(confidence level)**³。换句话说，如果我们已经为检验统计量预先指定了拒绝域，让显著性水平 α 正好等于得到的 P -值，那么检验统计量的值将位于拒绝域的边界上。然而，在拟合优度检验中， P -值是一个随机变量。这与4.1节的情形不同，在4.1节中，显著性水平 α 是一个常数，这个常数在实施检验之前就已经指定好了。

在经典统计中，讨论到此为止，我们并不尝试给出 H_0 假设为真的概率，因为经典统计中假设不被当作随机变量处理。(然而，显著性水平或 P -值经常被不恰当地解释成这样的概率。) 在贝叶斯统计中，我们将利用 (1.6) 式的贝叶斯定理给出 H_0 为真的概率，但是这需要给定验前概率，即在看到实验结果之前硬币公平这一假设的概率是多少。有些情况下，这个方法可以实际操作，然而，有些情况下不可操作。我们目前将局限于经典统计的框架内，仅仅给出 P -值，而不讨论 H_0 假设为真的概率。

因此， P -值表示，如果在相似的环境下多次重复实验 (即投掷硬币 20 次)，得到的结果与假设 H_0 不一致程度不好于实际观测结果的相对频数。所谓‘相似的环境’指的是每次实验都投掷 20 次，或者更一般地说是每次实验进行相同次数的观测。假设我们设计这样一个实验：不断投掷硬币，直到至少出现 3 次正面和 3 次背面，然后终止实验。再假设在实际实验中，我们投掷了 20 次硬币才终止了实验。在这种情形下，可以证明，投掷 20 次或 20 次以上才终止实验 (即观测结果对 H_0 假设的支持不好于当前观测) 的概率不是 0.26% 而是 0.072%。于是在验证 H_0 假设是否正确时好像出现了明显不同的结论。也存在一种可能，我们根本不知道实验者如何设计投掷硬币的实验，而只是得到了实验的结果。为了避免这种所谓**可选终止(optional stopping)** 问题，可以采用的一种方法是，将‘相似实验’解释为观测次数相同的实验。尽管这是一个任意选择的惯例，但它可以保证报告的显著性水平具有唯一解释。关于这一问题的深入讨论参见文献 [16, 46]。

在投掷硬币的例子中，检验统计量 $t = n_h$ 是合理的，因为从问题的对称性，很容易判断在哪些区域 t 与 H_0 假设的不一致程度不好于当前观测值。实际上，在这个例子中，各种可能的备择假设的集合就是所有参数 p 不等于 0.5 的集合，这些备择假设都会导致正面和背面次数的不对称。

³这与置信区间的置信水平有联系，但并不相同，参见9.2节。

4.6 观测信号的显著性

常用的一种拟合优度检验是判断数据与期望值之间的差异是否足够明显，以至于可以声明新发现。考虑某种特殊类型信号事例，信号数目 n_s 可以看作均值为 ν_s 的泊松变量。然而，除了信号之外，一般还存在一定数目的本底事例 n_b 。假设本底数目也可看作泊松变量，其均值为 ν_b ，并假定 ν_b 已知而且没有不确定度。因此，观测到的总事例数 $n = n_s + n_b$ 也是一个泊松变量，均值为 $\nu = \nu_s + \nu_b$ 。于是，观测到 n 个事例的概率为

$$f(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}. \quad (4.37)$$

假设在实验中观测到 n_{obs} 个事例。为了定量描述发现新效应 (即 $\nu_s \neq 0$) 的信心程度，我们可以在假设只有本底的情况下计算观测事例数不少于 n_{obs} 的概率。这个概率为

$$\begin{aligned} P(n \geq n_{\text{obs}}) &= \sum_{n=n_{\text{obs}}}^{\infty} f(n; \nu_s = 0, \nu_b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} f(n; \nu_s = 0, \nu_b) \\ &= 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}. \end{aligned} \quad (4.38)$$

例如，如果我们期望本底事例的均值为 $\nu_b = 0.5$ ，但是观测到的事例数为 $n_{\text{obs}} = 5$ ，那么从 (4.38) 式可以计算出 P -值为 1.7×10^{-4} 。需要强调的是，这并非 $\nu_s = 0$ 这一假设 (hypothesis) 的概率。准确地讲，这是在假设 $\nu_s = 0$ 的条件下观测到不少于当前观测事例数的概率。尽管解释有些微妙，在判断是否发现新效应时， P -值是一个值得考虑的有用的量。

测量结果经常表示为参数的估计值加减一个标准差 (在第九章我们会继续讨论如何报告测量不确定度的问题)。由于泊松变量的标准差等于其均值的平方根，我们可以将 ν 估计为 $5 \pm \sqrt{5}$ ，或者在扣除本底之后将 ν_s 估计为 4.5 ± 2.2 。然而，这个表示存在一定的误导性，因为这个结果与零事例相比只有两倍标准差，所以给人一种印象，即实验结果没有明显表现出与 $\nu_s = 0$ 的不一致性。但是，我们从 P -值已经看到，事实并非如此，实验结果与 $\nu_s = 0$ 明显不一致。这里，我们需要求出均值为 ν_b 的泊松变量涨落到 n_{obs} 或更高的概率，而不是均值为 n_{obs} 的变量涨落到 ν_b 或更少事例的概率。不幸的是，在用测量值的平方根表示泊松变量测量结果的误差棒时，容易导致不正确的解释。

还有一个危险之处是，我们已经假设 ν_b 已知并且没有不确定度。在上面的例子中，如果取 $\nu_b = 0.8$ 而不是 0.5， P -值几乎提高一个数量级，变成 1.4×10^{-3} 。因此，在评估新效应的显著性时，定量估计本底的系统不确定度非常重要。通常的做法是，根据 ν_b 的合理取值范围，给出 P -值的取值范围。

假设除了事例计数，我们还测量了每个事例的变量 x 。为了初步看一下数据的分布状况，通常先将数据画成直方图，如图4.3所示。为了对比，可以对理论期待值 (虚线直方图) 进行归一化，使得每个区间的值表示该区间理论预期的事例数。如果所有 x 值都相互独立，则每个区间的事例数都是泊松随机变量。给出图4.3这样的直方图后，一个很自然的问题是，图中的峰状结构是否对应新的发现？

原则上，我们可以将前面的分析过程简单地应用到任何给定区间的频数上，或者应用到几个联合的区间上。图4.3事例数最多的两个区间里，观测到的总事例数为 $n_{\text{obs}} = 11$ ，对应的本底期待值为 $\nu_b = 3.2$ 。对于均值为 $\nu_b = 3.2$ 的泊松变量，观测到不少于 11 个事例的概率为 5.0×10^{-4} 。

然而，通常情况下，我们无法先验地知道峰的位置。如果利用数据构造了某个包含很多区间的直方图，由于统计涨落，有些区间自然会表现出与理论的差异。因此，为了更好地表明这个峰的出现出乎意料，我们可以给出在任何区间观测到类似差异 (即与图中的峰状结构类似) 的概率。然而，我们还可以争论，认为给出这样的概率还不够，因为可能我们可以已经看过很多不同的直方图。但是不管怎样，相比于报告直方图某特定区间的显著性水平，这种做法都更有意义。

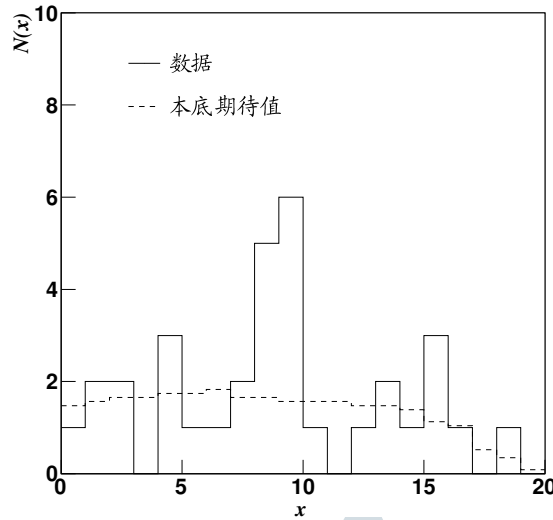


图 4.3: 变量 x 的观测值与期望值的直方图。数据给出相对显著的峰。

如果我们希望确定是否存在勉强有些显著的效应，很重要的一点是，一旦已经开始对数据进行考察，那么事例选择和分析过程都应当保持不变。如果为了增强信号显著性而修改事例选择或分析过程（通常没有办法说清楚这是否是刻意为之），那么前面关于显著性水平的解释将不再成立。然而，我们可以修改分析过程并应用于新的数据，然后利用新数据针对前一组数据样本峰值出现的区间计算 P -值。这当然仅适用于有额外数据可用的情况。

然而，这种方法仍存在缺点，因为图4.3峰值以外的区域也或多或少存在与理论期待值之间的偏离（有些地方高一些，有些地方低一些），在计算整体符合程度的时候需要以某种方式考虑进来。而且，如果选取不同的区间宽度，峰值所在区间的事例数也会有所变化，前面介绍的检验统计量的值将有所不同。确定直方图区间数目的典型做法是，峰值区域的宽度至少要定义为变量 x 的分辨率的几倍。

4.7 皮尔逊卡方检验

本节将讨论的拟合优度检验可以应用于变量 x 的分布。如图4.3所示，将观测到的 x 值填充到分 N 个区间的直方图。假设第 i 个区间的事例数为 n_i ，相应的期望值为 ν_i 。我们希望构造一个统计量，能够反映观测直方图与理论期待的直方图之间的符合程度。毫无疑问，基于皮尔逊卡方统计量(Pearson χ^2 statistic)是最常用的拟合优度检验：

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \nu_i)^2}{\nu_i}. \quad (4.39)$$

如果数据 $\mathbf{n} = (n_1, \dots, n_N)$ 服从均值为 $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$ 的泊松分布，并假设每个区间的事例数不太少（例如 $n_i \geq 5$ ），可以证明 (4.39) 式的统计量服从自由度为 N 的卡方分布，即 (2.34) 式。不论 x 服从何种分布，这一结论都成立。因此，我们经常说卡方检验与分布无关 (distribution free)。对每个区间事例数的限制（即 $n_i \geq 5$ ），等价于要求 n_i 可以近似为高斯分布。

由于均值为 ν_i 的泊松变量的标准差等于 $\sqrt{\nu_i}$ ，所以卡方检验统计量实际上等于，以相应的标准差为单位，观测值与理论期待值之差的平方和。 χ^2 值越大，说明数据与假设之间的差别越大。因此， P -值或显著性水平可以用卡方分布从观测到的 χ^2 值到无穷的积分表示，

$$P = \int_{\chi^2}^{\infty} f(z; n_d) dz, \quad (4.40)$$

其中自由度的数目为 $n_d = N$ 。⁴ 注意到卡方分布的期望值等于自由度的数目，所以经常用比值 χ^2/n_d 来衡量数据与假设的符合程度。然而，这个量传递的信息不如同时给出 χ^2 值和 n_d 。例如，对于 $\chi^2 = 15$ ， $n_d = 10$ ，相应的 P -值为 0.13；然而，对于 $\chi^2 = 150$ ， $n_d = 100$ 的情形，相应的 P -值为 9.0×10^{-4} 。

对于 (4.39) 式，我们实际上假定了总事例数 $n_{\text{tot}} = \sum_{i=1}^N n_i$ 本身也是泊松变量，理论预期的均值为 $\nu_{\text{tot}} = \sum_{i=1}^N \nu_i$ 。然而，我们也可以认为 n_{tot} 是固定的，因此数据 n_i 服从概率为 $p_i = \nu_i/n_{\text{tot}}$ 的多项分布。我们在这里并不想检验总事例数的理论期待值与实际观测值之间的符合程度，而只检验变量 x 的分布情况。于是，可以构造 χ^2 统计量为

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}} \quad (4.41)$$

可以证明，在每个区间的事例数都很大的极限下，(4.41) 式的检验统计量服从自由度数目为 $N - 1$ 的卡方分布。这里，我们已经假定概率 p_i 已知。一般而言，如果要从数据中估计 m 个参数，则自由度的数目减少 m 。我们将在第七章回到卡方检验上来，讨论它与最小二乘法的联系。

有时可用的数据量很小，无法满足每个区间至少 5 个事例的要求。此时，我们仍然可以构造卡方统计量，只要所有的 ν_i 都大于零。然而，这个统计量将不再服从卡方分布，其分布将依赖于变量 x 的分布（也就是说，这时的卡方检验不再是分布无关的检验）。以图 4.3 为例，我们可以得到 $\chi^2 = 29.8$ ，自由度为 $n_d = 20$ 。然而，多数区间不足 5 个事例，因此在计算 P -值时不能视其为卡方分布。

要得到正确的 P -值，需要利用蒙特卡罗方法确定检验统计量的分布。采用的做法是，对每个区间，根据均值 ν_i 产生相应的泊松变量的值，然后计算并记录相应 χ^2 值。（可以用 [25] 中的 **RNPSSN** 子程序产生服从泊松分布的随机数，也可以利用 ROOT [26] 中相应的随机数产生子，例如通过指针 **gRandom** 调用函数 **Poisson()**。）利用蒙特卡罗模拟进行大量实验得到的分布如图 4.4 所示，作为对比，图上还给出了 (2.34) 式定义的卡方分布。如果按照卡方分布处理，得到的 P -值为 0.073。然而，蒙特卡罗方法得到的分布表明， χ^2 较大的区域的概率实际上比卡方分布更大一些，相应的 P -值为 $P = 0.11$ 。需要注意的是，这种情形下，卡方检验对峰值的存在不太敏感，无法为拒绝只有本底没有额外信号的假设提供显著的证据。

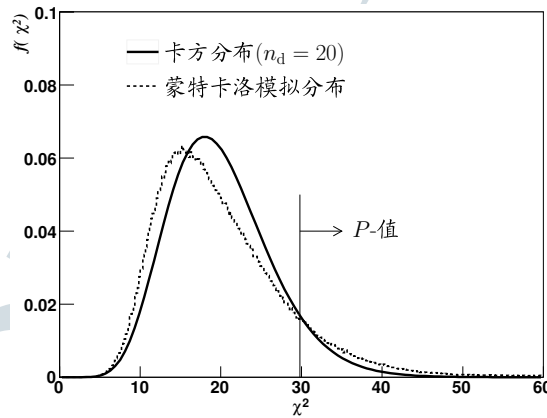


图 4.4: 利用蒙特卡罗模拟从图 4.3 的例子得到的卡方检验统计量的分布，以及自由度数目 $n_d = 20$ 的卡方分布的曲线。

卡方检验的一个不确定之处在于，我们必须选择特定的分区间方式。对于小数据样本，不同的分区间方式通常会导致检验得到的 P -值不同。一方面，区间宽度需要足够大，保证每个区间包含至少 5 个事例，从而可以用卡方分布计算显著性水平。另一方面，区间宽度太大会丢掉部分信息，例如同一个区间内每个 x 都丢失了具体值的信息。如果可以通过蒙特卡罗确定卡方检验统计量的分布，则可以放松对每个区间最

⁴卡方分布的累积分布，即 1 减去 (4.40) 式的积分，可以用 [25] 中的 **PROB** 子程序计算，也可以用 ROOT [26] 中的 **TMath::Prob()** 函数计算。

小事例数的要求，从而可以减小区间宽度。

适用于小数据样本的其他统计检验都是基于 x 的每一个具体值，即不分区间处理。柯尔莫哥洛夫-斯米尔诺夫 (Kolmogorov-Smirnov) 检验和斯米尔诺夫-米塞斯 (Smirnov-Cramér-von Mises) 检验就是这样的例子。关于这些检验的讨论在文献 [20, 23] 中可以找到。

DRAFT

DRAFT ZV

第五章 参数估计的一般概念

本章要介绍的是参数估计的一般概念，这些概念将应用于第六到八讨论的所有方法。此外，本章还给出了估计概率密度函数的均值和方差等性质的一般规则。

5.1 样本、估计量、偏倚

考虑服从概率密度函数 $f(x)$ 的随机变量 x 。样本空间为 x 的所有可能取值的集合。对变量 x 进行 n 次独立观测，所得观测值构成的集合称为容量为 n 的**样本(sample)**。可以定义新的样本空间为 n 维矢量 $\mathbf{x} = (x_1, \dots, x_n)$ 所有可能取值的集合，也就是说，把包含 n 次测量的整个实验看成一次随机测量，该测量用 n 个量 x_1, \dots, x_n 来描述。由于所有测量都可认为是相互独立的，每个 x_i 都服从相同的概率密度函数 $f(x)$ ，样本的联合概率密度函数 $f_{\text{sample}}(x_1, \dots, x_n)$ 可以表示为

$$f_{\text{sample}}(x_1, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n). \quad (5.1)$$

尽管实际应用中随机矢量的维数(即测量次数)可能非常大，但是，由于样本的联合概率密度函数为 n 个形式完全相同的概率密度函数的乘积，问题实际上被大大简化。

考虑另外一种情形，我们对随机变量 x 进行了 n 次测量，但 x 服从的概率密度函数 $f(x)$ 未知。统计的核心问题是根据观测值 x_1, \dots, x_n 推断概率密度函数 $f(x)$ 的性质。更明确地说，就是要构造 x_i 的函数，用来估计概率密度函数 $f(x)$ 的各种性质。通常，我们会假设概率密度函数为 $f(x; \theta)$ ，即它依赖于未知参数 θ (或多维参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$)，而我们的目标则是构造观测值 x_i 的函数，以估计未知参数的值。

观测值 x_1, \dots, x_n 的函数(不包含未知参数)称为**统计量(statistic)**。特别地，用来估计概率密度函数的性质(例如均值、方差或其它参数)的统计量称为**估计量(estimator)**。某个量 θ 的估计量通常用 $\hat{\theta}$ 表示，即在这个量的符号上加上“尖帽”，以区别于其真值 θ 。这个真值是未知的，而且很可能永远也不可能知道其大小。

如果 $\hat{\theta}$ 在大 n 极限下收敛于 θ ，则称该估计量为**相合(consistent)**估计量。这里所说的收敛意指概率意义上的收敛，即对于任意 $\epsilon > 0$ ，满足

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0. \quad (5.2)$$

相合性通常是有用的估计量的最基本要求。后面我们经常用“大样本”或“渐进”极限等字样表示大 n 极限。如果有必要明确区分，我们会用估计量表示样本的函数(即其函数形式)，而用**估计值(estimate)**表示用特定样本计算得到的估计量的值。利用给定数据样本 x_1, \dots, x_n 估计参数值的过程称为**参数拟合(parameter fitting)**。

由于估计量 $\hat{\theta}(x_1, \dots, x_n)$ 为观测值的函数，所以估计量本身也是随机变量。换句话说，如果多次重复整个实验，每次得到一个样本容量为 n 的新的样本 $\mathbf{x} = (x_1, \dots, x_n)$ ，估计量 $\hat{\theta}(\mathbf{x})$ 将按照一定的概率密度函

数 $g(\hat{\theta}; \theta)$ 取不同值, 这个概率密度函数原则上依赖于 θ 的真值。统计量的概率分布称为**抽样分布 (sampling distribution)**。后面几章的大多数问题都涉及到抽样分布及其性质, 尤其是期望值和方差等性质。

如果估计量 $\hat{\theta}$ 服从样本概率密度函数 $g(\hat{\theta}; \theta)$, 其期望值为

$$\begin{aligned} E[\hat{\theta}(\mathbf{x})] &= \int \hat{\theta} g(\hat{\theta}; \theta) d\hat{\theta} \\ &= \int \cdots \int \hat{\theta}(\mathbf{x}) f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n, \end{aligned} \quad (5.3)$$

其中用到了 (5.1) 式的联合概率密度函数。需要注意的是, $\hat{\theta}$ 的这个期望值是通过无限次相似实验获得的, 每次实验的样本容量都是 n 。定义估计量 $\hat{\theta}$ 的**偏倚 (bias)**为

$$b = E[\hat{\theta}] - \theta. \quad (5.4)$$

注意, 偏倚不依赖于样本测量值, 而依赖于样本容量、估计量的函数形式以及概率密度函数 $f(x)$ 的真实属性 (通常是未知的), 包括参数 θ 的真值。如果不论样本容量多大, 参数估计量的偏倚均为零, 则称该估计量为无偏估计量。如果当 $n \rightarrow \infty$ 时偏倚变为零, 则称估计量为渐进无偏估计量。需要注意的是, 即使估计量 $\hat{\theta}$ 是相合估计量, 它也可以是有偏估计量。也就是说, 当单次实验进行无限次测量时, $\hat{\theta}$ 收敛于真值 θ , 即便如此, 我们也无法得出这样的结论: 进行无限次实验, 每次实验包含有限次测量, 则 $\hat{\theta}$ 的平均值收敛于真值。因此, 如果我们希望把不同实验的结果结合起来, 估计量的无偏性特别有用。很多实际情况中, 偏倚相比于统计误差 (即标准差) 非常小。此时, 如果估计量拥有其他某些好的品质 (例如简单性), 我们通常也会考虑使用这样的估计量。

评价估计量品质好坏的另一个量度是**均方差 (mean squared error, MSE)**, 定义为

$$\begin{aligned} \text{MSE} &= E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] - \theta)^2 \\ &= V[\hat{\theta}] + b^2. \end{aligned} \quad (5.5)$$

即, 均方差等于方差与偏倚的平方和, 因此可以解释为统计不确定度与系统不确定度的平方和。

需要强调的是, 经典统计中构造估计量的方法并不唯一。然而, 只要给定某个估计量, 我们就可以指出该估计量在什么程度下具有理想的品质, 例如偏倚是否很小或者为零, 方差是否很小, 均方差是否很小, 等等。在第六章我们将看到, 偏倚与方差之间存在一定的权衡。如果给定估计量的偏倚, 则估计量的方差下限会更小。通常, 如果估计量的偏倚为零并且方差最小, 我们就认为这个估计量是“最优的”, 尽管有时也会考虑其他量评判估计量是否可取, 例如均方差 (参见11.7)。第六章到第八章介绍的方法, 在大多数实际情况下可以构造最优或接近最优的估计量。在逆问题 (解谱法) 中, 有偏估计量非常重要, 我们将在第十一章予以讨论。

5.2 均值、方差和协方差的估计量

假设有随机变量 x 的一个样本, 样本容量为 n : x_1, \dots, x_n 。假设 x 服从的概率密度函数 $f(x)$ 未知, 其函数的参数形式也未知。我们希望构造 x_i 的函数作为 x 的期望值 μ 的估计量。一个自然的想法是求 x_i 的代数平均值, 定义为

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (5.6)$$

样本元素的代数平均值称为**样本均值 (sample mean)**, 通常表示为随机变量的符号上方加一条横线, 例如 \bar{x} 。注意不要将样本均值与 x 的期望值 (总体均值) 混淆, x 的期望值用 μ 或 $E[x]$ 表示, 而样本均值 \bar{x} 是一个估计量。

样本均值的第一个重要性质由所谓的大数弱定律 (**weak law of large numbers**) 给出。该定律指出, 如果 x 的方差存在, 则 \bar{x} 为总体均值 μ 的相合估计量。换言之, 当 $n \rightarrow \infty$ 时, \bar{x} 在概率的意义上收敛于 μ , 参见 (5.2) 式。在文献 [3] 中可以找到该定理的证明。方差存在这一条件表明, 如果 x 服从柯西分布 (2.40) 则该定理不成立。此时, 我们实际上可以证明, 无论样本容量多大, \bar{x} 的概率密度函数与 x 的概率密度函数完全相同。然而, 在实际问题中, 代表物理量的随机变量的方差总是有限的 (参见 2.9 节), 因此大数弱定律总是成立的。

样本均值的期望值 $E[\bar{x}]$ 为 (见 (5.3) 式)

$$E[\bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu, \quad (5.7)$$

因为下式对所有 i 都成立:

$$E[x_i] = \int \cdots \int x_i f(x_1) \cdots f(x_n) dx_1 \cdots dx_n = \mu. \quad (5.8)$$

从 (5.7) 式可以看到, 样本均值 \bar{x} 是总体均值 μ 的无偏估计量。

样本方差 (sample variance) s^2 定义为

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} (\overline{x^2} - \bar{x}^2). \quad (5.9)$$

与计算样本均值 \bar{x} 的期望值类似, 我们可以计算样本方差 s^2 的期望值。 s^2 的定义中引入了因子 $1/(n-1)$, 以保证期望值等于 σ^2 , 即保证 s^2 为总体方差的无偏估计量。如果均值 μ 已知, 则按照下式定义的统计量 S^2 为方差 σ^2 的无偏估计量。

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 = \overline{x^2} - \mu^2, \quad (5.10)$$

用类似的方法可以证明, 两个随机变量 x 和 y 的协方差 V_{xy} 的无偏估计量为

$$\hat{V}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1} (\overline{xy} - \bar{x} \cdot \bar{y}). \quad (5.11)$$

可以将 \hat{V}_{xy} 除以变量 x 和 y 的样本方差的平方根, 得到相关系数 ρ 的估计量 r (见 (1.48) 式, 后面我们经常省略下标 xy , 即 $r = r_{xy}$):

$$\begin{aligned} r &= \frac{\hat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{j=1}^n (x_j - \bar{x})^2 \cdot \sum_{k=1}^n (y_k - \bar{y})^2\right)^{1/2}} \\ &= \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}. \end{aligned} \quad (5.12)$$

给定估计量 $\hat{\theta}$, 可以计算方差 $V[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$ 。注意到方差 $V[\hat{\theta}]$ (或者等价地, 方差的平方根 $\sigma_{\hat{\theta}}$) 是 $\hat{\theta}$ 围绕均值 (进行大量重复实验, 每次实验的样本容量为 n) 变化幅度的度量, 因此经常称其为 $\hat{\theta}$ 的统计不确定度。例如, 样本均值 \bar{x} 的方差为

$$\begin{aligned} V[\bar{x}] &= E[\bar{x}^2] - (E[\bar{x}])^2 = E\left[\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \left(\frac{1}{n} \sum_{j=1}^n x_j\right)\right] - \mu^2 \\ &= \frac{1}{n^2} \sum_{i,j=1}^n E[x_i x_j] - \mu^2 \\ &= \frac{1}{n^2} [(n^2 - n)\mu^2 + n(\mu^2 + \sigma^2)] - \mu^2 = \frac{\sigma^2}{n}, \end{aligned} \quad (5.13)$$

其中 σ^2 为 $f(x)$ 的方差，其中已经用到了 $E[x_i x_j] = \mu^2 (i \neq j \text{ 时})$ 和 $E[x_i^2] = \mu^2 + \sigma^2$ 。这个公式给出的结果非常重要。它说明，对 x 进行 n 次测量，其平均值的标准差等于 $f(x)$ 本身的标准差除以 \sqrt{n} 。

用同样的方法，可以计算出 (5.9) 式估计量 s^2 的方差为

$$V[s^2] = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \mu_2^2 \right), \quad (5.14)$$

其中 μ_k 为 (1.42) 式中的 k 阶中心矩，例如 $\mu_2 = \sigma^2$ 。对 (5.9) 式作一个简单的推广，可以将 μ_k 估计为

$$m_k = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^k. \quad (5.15)$$

相关系数的估计量 r 的期望值和方差，依赖于联合概率密度函数 $f(x, y)$ 的高阶矩。对于 (2.30) 式的二维高斯概率密度函数， r 的均值和方差为 (见 [49] 及其中的参考文献)

$$E[r] = \rho - \frac{\rho(1-\rho^2)}{2n} + \mathcal{O}(n^{-2}), \quad (5.16)$$

$$V[r] = \frac{1}{n} (1-\rho^2)^2 + \mathcal{O}(n^{-2}). \quad (5.17)$$

虽然 (5.12) 式给出的估计量 r 只是渐进无偏估计量，但是因其简单性仍被广泛使用。需要注意的是，尽管 \hat{V}_{xy} ， s_x^2 和 s_y^2 分别是 V_{xy} ， σ_x^2 和 σ_y^2 的无偏估计量，非线性函数 $\hat{V}_{xy}/(s_x s_y)$ 并非 $V_{xy}/(\sigma_x \sigma_y)$ 的无偏估计量 (参见 6.2 节)。因此，在应用 (5.17) 式评估观测到的相关性的显著性时，需要格外小心 (见 9.5 节)。

第六章 极大似然法

6.1 极大似然估计量

考虑服从概率密度函数 $f(x; \theta)$ 的随机变量 x 。假设已知 $f(x; \theta)$ 的函数形式，但其中至少一个参数 θ (或 $\theta = (\theta_1, \dots, \theta_m)$) 的取值未知。也就是说， $f(x; \theta)$ 表示概率密度函数的一个复合假设 (参见4.1节)。**极大似然法 (maximum likelihood)** 是给定有限数据样本的条件下进行参数估计的一种方法。假设对随机变量 x 进行了 n 次测量，测量值为 x_1, \dots, x_n 。这里， x 也可以表示多维随机矢量，即单次观测的结果需要用几个量一起表示。

如果给定假设 $f(x; \theta)$ 以及参数值 θ ，第一次测量结果处于区间 $[x_1, x_1 + dx_1]$ 的概率为 $f(x_1; \theta)dx_1$ 。由于假定所有测量都是独立的，所以第一次测量处于 $[x_1, x_1 + dx_1]$ ，第二次测量处于 $[x_2, x_2 + dx_2]$ 等等，这个总概率为

$$x_i \text{ 处于 } [x_i, x_i + dx_i] \text{ (对所有 } i \text{) 的概率} = \prod_{i=1}^n f(x_i; \theta) dx_i. \quad (6.1)$$

如果假设的概率密度函数形式和参数值都正确，对于实际测量得到的数据，我们期望这个概率比较高；如果假设的参数值远离真实值，这将导致测量结果的概率比较低。由于 dx_i 与参数无关，这个推理显然适用于下面的函数 L ：

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta), \quad (6.2)$$

这个函数称为**似然函数 (likelihood function)**。需要注意的是， $L(\theta)$ 实际上就是 x_i 的联合概率密度函数，尽管这里将其看作参数 θ 的函数。另一方面， x_i 被视为固定值 (即实验已经结束)。

根据这一动机，我们定义参数的极大似然 (ML) 估计量为使似然函数取极大值的参数值。只要似然函数对参数 $\theta_1, \dots, \theta_m$ 可导，并且最大值不在参数区间的边界，就可以由下式给出估计量

$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1, \dots, m. \quad (6.3)$$

如果存在多个局域极大值，则选择其中最大的一个。与其它类型的估计量一样，极大似然估计量通常加上尖帽号， $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$ ，用来区分估计量与准确值未知的真实参数 θ_i 。

极大似然法的基本思想如图6.1所示。根据均值 $\mu = 0.2$ 、标准差 $\sigma = 0.1$ 的高斯分布，用蒙特卡罗方法产生了包含 50 次测量的数据样本 (在横轴上用刻度符号 (tick mark) 表示)。图6.1(a) 的实线是根据参数 $\hat{\mu} = 0.204$ 和 $\hat{\sigma} = 0.106$ 计算而得，这两个参数值使似然函数取极大值 (因此似然函数的对数也取最大值)。虚线是根据高斯分布的真实参数值 (即 $\mu = 0.2, \sigma = 0.1$) 所画。由于随机涨落，估计值 $\hat{\mu}$ 和 $\hat{\sigma}$ 与相应的真值 μ 和 σ 不严格相等。估计量 $\hat{\mu}$ 和 $\hat{\sigma}$ 以及二者的方差 (反映了统计不确定度的大小) 将在6.3节讨论。图6.1(b) 是利用参数值远离真值的概率密度函数所得，对应的似然函数值也比较小。

前面给出的极大似然原则的动机，不一定能够保证得到的估计量具有优化性质。然而实际证明，极大似然法有很多优点，包括使用方便、不需要对数据分区间等。下面，我们将依据几个标准探讨极大似然估

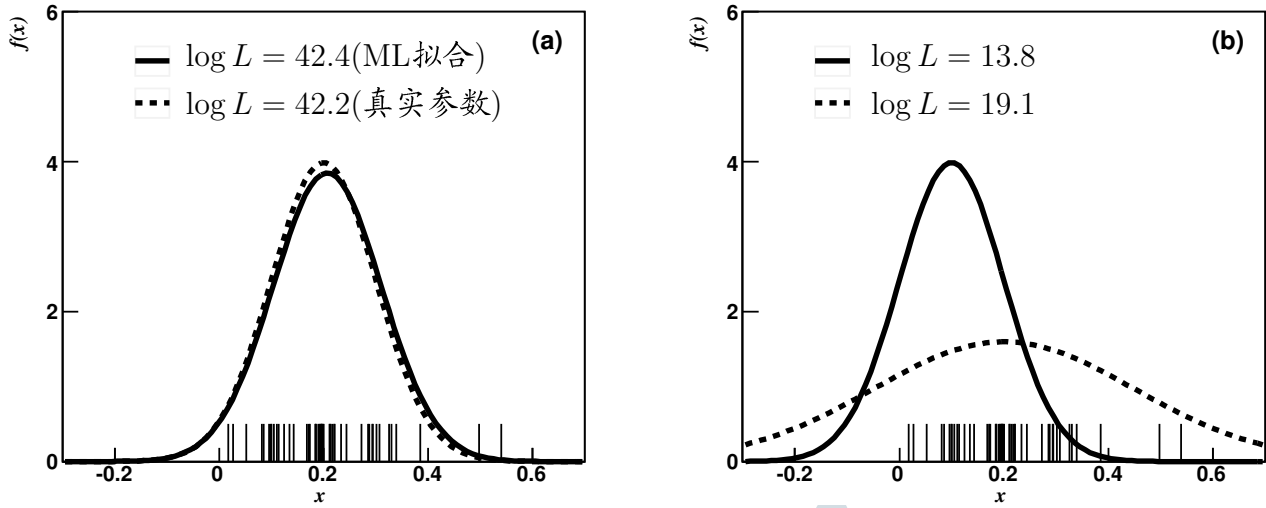


图 6.1: 对高斯随机变量(均值 $\mu = 0.2$, 标准差 $\sigma = 0.1$) 进行 50 次观测得到的样本。(a) 利用使似然函数取最大值的参数和真实参数计算得到的概率密度函数。(b) 利用远离真值的参数计算得到的概率密度函数, 相应的似然函数值比较小。

计量的可取性, 这些标准主要包括估计量的方差和偏倚。

6.2 极大似然估计量：指数分布

假设我们测量某种不稳定粒子的固有衰变时间, 观测到 n 次衰变, 测量值为 t_1, \dots, t_n 。并且, 我们选取的理论假设是 t 服从均值为 τ 的指数分布:

$$f(t; \tau) = \frac{1}{\tau} e^{-t/\tau}. \quad (6.4)$$

现在的任务是利用观测数据估计参数 τ 的值。与 (6.2) 式定义的似然函数相比, 使用似然函数的对数通常更为方便。由于对数是单调增函数, 所以使似然函数 L 最大的参数值必然使 $\log L$ 最大。对数的优势在于, L 中的乘积转换成了求和, 并且 f 中的指数转化成简单的因子。因此, **对数似然函数 (log-likelihood function)** 为

$$\log L(\tau) = \sum_{i=1}^n \log f(t_i; \tau) = \sum_{i=1}^n \left(\log \frac{1}{\tau} - \frac{t_i}{\tau} \right). \quad (6.5)$$

令 $\log L$ 关于 τ 取最大值, 可以给出极大似然估计量 $\hat{\tau}$,

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i. \quad (6.6)$$

可以看到, 在这个例子中, 极大似然估计量 $\hat{\tau}$ 就是时间测量值的样本均值。 $\hat{\tau}$ 的期望值为

$$\begin{aligned} E[\hat{\tau}(t_1, \dots, t_n)] &= \int \cdots \int \hat{\tau}(t_1, \dots, t_n) f_{\text{joint}}(t_1, \dots, t_n; \tau) dt_1 \cdots dt_n \\ &= \int \cdots \int \left(\frac{1}{n} \sum_{i=1}^n t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \cdots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \cdots dt_n \\ &= \frac{1}{n} \sum_{i=1}^n \left(\int t_i \frac{1}{\tau} e^{-t_i/\tau} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-t_j/\tau} dt_j \right) \end{aligned}$$

$$= \frac{1}{n} \sum_{i=1}^n \tau = \tau, \quad (6.7)$$

因此, $\hat{\tau}$ 是 τ 的无偏估计量。从2.4节和5.2节的结果立刻就可以得到这个结论, 因为在那里我们已经证明指数分布的期望值为 τ , 并且样本均值是任意概率密度函数期望值的无偏估计量。(参见10.4节对 $\hat{\tau}$ 的概率密度函数的推导。)

作为例子, 考虑用蒙特卡罗方法按指数分布生成的衰变时间 t 的样本, 样本容量 $n = 50$, 如图6.2所示。生成数据样本时选取的真实寿命为 $\tau = 1.0$ 。(6.6) 式给出极大似然估计值为 $\hat{\tau} = 1.062$ 。图中的曲线是利用极大似然估计得到的指数概率密度函数。

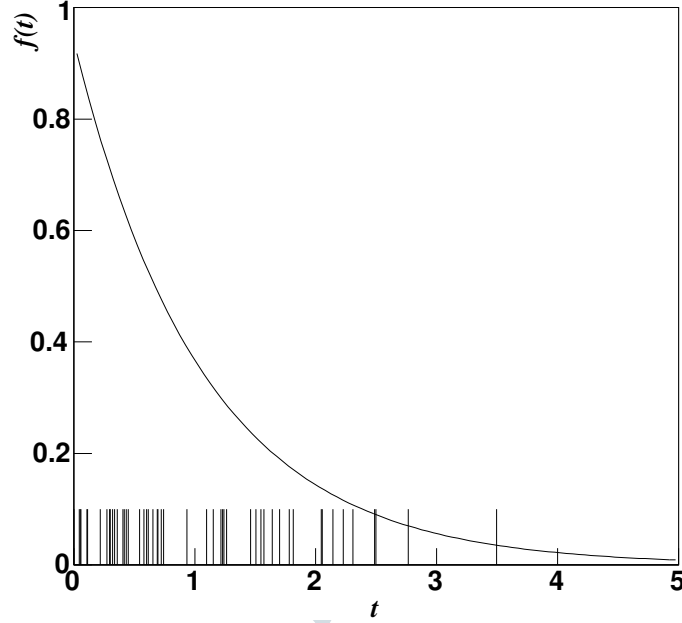


图 6.2: 利用蒙特卡罗模拟, 对均值 $\tau = 1.0$ 的指数随机变量 t 进行 50 次观测得到的样本。图中的曲线是极大似然拟合的结果, $\hat{\tau} = 1.062$ 。

假如我们感兴趣的不是平均寿命而是衰变常数 $\lambda = 1/\tau$, 那么应当如何估计 λ 呢? 一般来说, 给定某个参数 θ 的函数 $a(\theta)$, 可以得到

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial a} \frac{\partial a}{\partial \theta} = 0. \quad (6.8)$$

因此, $\partial L / \partial \theta = 0$ 意味着, 当 $a = a(\theta)$ 时 $\partial L / \partial a = 0$, 除非 $\partial a / \partial \theta = 0$ 。只要 $\partial a / \partial \theta \neq 0$, 我们就可以利用原始的极大似然估计量, 直接根据函数形式计算得到该函数的极大似然估计量, 即 $\hat{a} = a(\hat{\theta})$ 。因此, 衰变常数的极大似然估计量为 $\hat{\lambda} = 1/\hat{\tau} = n / \sum_{i=1}^n t_i$ 。这种变换不变性是极大似然法非常方便的一个性质, 但无偏估计量在变换后不一定能保证还是无偏的。在10.4节我们将会得到, $\hat{\lambda}$ 的期望值为

$$E[\hat{\lambda}] = \lambda \frac{n}{n-1} = \frac{1}{\tau} \frac{n}{n-1}, \quad (6.9)$$

因此, 只有在大 n 极限下, $\hat{\lambda} = 1/\hat{\tau}$ 才是 $1/\tau$ 的无偏估计量, 尽管对于任意 n , $\hat{\tau}$ 都是 τ 的无偏估计量。总之, 参数 θ 的函数 a 的极大似然估计量就是 $\hat{a} = a(\hat{\theta})$ 。但是, 如果 $\hat{\theta}$ 为 θ 的无偏估计量 ($E[\hat{\theta}] = \theta$), 并不能保证 $a(\hat{\theta})$ 也是 $a(\theta)$ 的无偏估计量。然而, 可以证明, 几乎对所有实际应用, 极大似然估计量的偏倚在大样本极限下都趋向于零。(一个例外的情形是, 随机变量允许的边界依赖于参数本身, 参见 [23] 第 8.3.3 节。)

6.3 极大似然估计量：高斯分布的均值和方差

假设随机变量 x 服从高斯分布，均值 μ 和方差 σ^2 未知，并且对 x 进行了 n 次测量。相应的对数似然函数为

$$\log L(\mu, \sigma^2) = \sum_{i=1}^n \log f(x_i; \mu, \sigma^2) = \sum_{i=1}^n \left(\log \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \log \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right). \quad (6.10)$$

令 $\log L$ 对 μ 的导数等于零，可以求解得到 μ 的极大似然估计量

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (6.11)$$

按照与 (6.7) 式相同的计算，可以得到估计量的期望值为 $E[\hat{\mu}] = \mu$ ，即 $\hat{\mu}$ 是 μ 的无偏估计量。（与平均寿命的估计量 $\hat{\tau}$ 一样， $\hat{\mu}$ 也正好等于样本均值，因此，从 2.5 节和 5.2 节已经知道 $\hat{\mu}$ 为 μ 的无偏估计量。）对 σ^2 重复同样的计算过程，并利用 $\hat{\mu}$ 的结果，可以得到

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2. \quad (6.12)$$

然而，计算 $\hat{\sigma}^2$ 的期望值得到，

$$E[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2. \quad (6.13)$$

因此，极大似然估计量 $\hat{\sigma}^2$ 是有偏估计量，但是在样本极大下偏倚变为零。

然而，在 5.1 节已经给出，样本方差 s^2 对任何概率密度分布都是方差的无偏估计量，因此

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (6.14)$$

为高斯分布参数 σ^2 的无偏估计量。总之，(6.12) 式给出了参数 σ^2 的极大似然估计量，并且其偏倚在 n 趋于无穷时趋向于零。(6.14) 式给出的统计量 s^2 是无偏的（好估计量），但这不是极大似然估计量。

6.4 极大似然估计量的方差：解析方法

给定随机变量 x 的 n 次观测，并假设其服从概率密度函数 $f(x; \theta)$ ，我们已经知道如何估计其中的未知参数。下一步需要做的是，如何评估这个估计的统计不确定度。换言之，如果多次重复整个实验（每次实验都进行 n 次观测），每次实验都将给出不同的参数估计值，我们希望知道参数估计值的分布有多宽？这个问题可以用估计量的方差（或标准差）来描述。

某些情形下，我们可以解析计算极大似然估计量的方差。例如，对于均值为 τ 的指数分布， τ 的极大似然估计量为 $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$ ，可以得到

$$\begin{aligned} V[\hat{\tau}] &= E[\hat{\tau}^2] - (E[\hat{\tau}])^2 \\ &= \int \cdots \int \left(\frac{1}{n} \sum_{i=1}^n t_i \right)^2 \frac{1}{\tau} e^{-t_1/\tau} \cdots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \cdots dt_n \\ &\quad - \left(\int \cdots \int \left(\frac{1}{n} \sum_{i=1}^n t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \cdots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \cdots dt_n \right)^2 \\ &= \frac{\tau^2}{n}. \end{aligned} \quad (6.15)$$

这是意料之中的结果，因为在5.2节我们已经看到，样本均值的方差等于 $1/n$ 乘以 t （对时间的单次测量）的概率密度函数的方差，在本例中，这个方差为 τ^2 (见2.4节)，而极大似然估计量 $\hat{\tau}$ 恰好等于样本均值。

需要注意的是，(6.15) 式计算得到的 $\hat{\tau}$ 的方差是真实参数 (未知) τ 的函数。那么问题来了：我们应当如何报道实验的统计不确定度？由于极大似然估计量的变换不变性 ((6.8) 式)，我们可以简单地将 τ 替换为其自身的估计量 $\hat{\tau}$ ，得到方差 $\sigma_{\hat{\tau}}^2 = \tau^2/n$ 的极大似然估计，即 $\widehat{\sigma_{\hat{\tau}}^2} = \hat{\tau}^2/n$ ，或者表示为标准差 $\hat{\sigma}_{\hat{\tau}} = \hat{\tau}/\sqrt{n}$ 。

如果实验者报道测量结果为 $\hat{\tau} = 7.82 \pm 0.43$ ，这表示参数 τ 的 (似然) 估计值为 7.82，并且如果多次重复实验，每次实验包含相同数目的观测，那么我们期待估计值的分布的标准差为 0.43。这是拟合参数的‘统计不确定度’的一种可能解释，与估计量的严格分布无关。然而，如果多次重复实验得到的不是高斯分布，这种解释不是标准解释。此时，我们通常给出所谓的 68.3% 的置信区间，这是第九章要讨论的问题。对于高斯分布的情形，68.3% 的置信区间等价于正负一倍标准差， $\pm\sigma$ 。可以证明 (例如参见 [7] 的 18.5 节)，在大样本极限下，极大似然估计实际上服从高斯分布。因此，在大样本极限下，两种方法给出的结果相同。

6.5 极大似然估计量的方差：蒙特卡罗方法

当解析求解很困难时，极大似然估计量的分布可以通过蒙特卡罗方法来研究。为此，我们必须对实验进行大量模拟，对每次模拟实验都计算极大似然估计值，考察得到的估计值的分布。对于蒙特卡罗程序中的‘真实’参数，可以采用真实实验得到的估计值。如上节所述，(5.9) 式定义的 s^2 是概率密度函数的方差的无偏估计量。因此，我们可以利用模拟实验得到的极大似然估计值计算 s ，并以此作为从真实测量估计出的参数值的统计不确定度。

作为例子，再次考虑指数分布的平均寿命测量 (见6.2节)。取真实寿命为 $\tau = 1.0$ ，观测次数 $n = 50$ 。假设利用测量得到的样本计算出参数的极大似然估计值为 $\hat{\tau} = 1.062$ (见图6.2)。将第一次蒙特卡罗实验当作‘真实’实验，然后再进行 1000 次模拟，每次模拟都包含 50 次观测。后面的 1000 次模拟中，参数真值取为 $\tau = 1.062$ ，即第一次 (模拟) 实验的极大似然估计值。

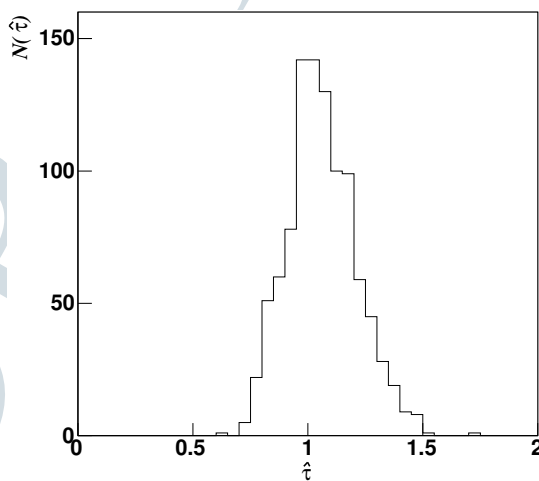


图 6.3: 1000 次蒙特卡罗实验得到的极大似然估计值 $\hat{\tau}$ 的直方图，每次实验包含 50 次观测。取图6.2的极大似然估计结果作为蒙特卡罗‘真实’参数 τ 。样本标准差为 $s = 0.151$ 。

图6.3给出了 1000 次模拟得到的极大似然估计值的直方图。这些估计的样本均值为 $\bar{\hat{\tau}} = 1.059$ ，与输入值非常接近，这是意料之中的事情，因为极大似然估计量 $\hat{\tau}$ 是无偏的。从这 1000 次实验得到的样本标准差为 $s = 0.151$ ，基本上跟 (6.15) 式计算得到的不确定度完全相同， $\hat{\sigma}_{\hat{\tau}} = \hat{\tau}/\sqrt{n} = 1.062/\sqrt{50} = 0.150$ 。因

此, 对于真实测量, 我们将报道测量结果为 $\hat{\tau} = 1.06 \pm 0.15$ (无论用哪种方式估计不确定度)。注意到, 估计量的分布近似为高斯分布。这是大样本极限下极大似然估计量的一般性质, 称为**渐进正态性(asymptotic normality)**。进一步的讨论参见文献 [23] 的第 7 章。

6.6 极大似然估计量的方差: RCF 边界方法

在很多应用中, 解析求解估计量的方差非常困难, 而蒙特卡罗方法的工作量也可能很大。此时, 通常采用所谓的 **RCF 不等式 (Rao-Cramér-Frechet inequality)** 给出估计量方差的最小边界值。RCF 不等式也称为**信息不等式 (information inequality)**, 适用于任何估计量, 不限于极大似然法构造的估计量。对于只有一个参数 θ 的情形, 该不等式给出

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / E \left[-\frac{\partial^2 \log L}{\partial \theta^2} \right], \quad (6.16)$$

其中 b 为 (5.4) 式定义的偏倚, L 为似然函数。该不等式的证明可以参见文献 [3]。实际上 (6.16) 式不是 RCF 不等式的最普遍形式, 但几乎所有实际应用都满足这种形式所要求的条件 (参见文献 [23] 的 7.4.5 节)。如果等号成立 (即方差最小), 估计量称为**有效 (efficient)** 估计量。可以证明, 如果要研究的问题存在有效估计量, 那么极大似然法一定可以找到该估计量。进一步可以证明, 在大样本极限下极大似然估计量总是有效的, 除非样本空间依赖于待估计的参数。在实际应用中, 我们经常假定估计量是有效的和零偏倚。如果对此有疑问, 可以用蒙特卡罗研究验证这些结论。关于有效估计量的一般条件, 在文献 [23] 的 7.4.5 节和文献 [7] 的第 18 章有所讨论。

对于均值为 τ 的指数分布的例子, 从 (6.5) 式可以得到

$$\frac{\partial^2 \log L}{\partial \tau^2} = \frac{n}{\tau^2} \left(1 - \frac{2}{\tau} \frac{1}{n} \sum_{i=1}^n t_i \right) = \frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right) \quad (6.17)$$

以及 $\partial b / \partial \tau = 0$, 因为偏倚 $b = 0$ (见 (6.7) 式)。因此, $\hat{\tau}$ 的方差的 RCF 边界 (也称为最小方差边界(minimum variance bound, MVB)) 为

$$V[\hat{\tau}] \geq \frac{1}{E \left[-\frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right) \right]} = \frac{1}{-\frac{n}{\tau^2} \left(1 - \frac{2E[\hat{\tau}]}{\tau} \right)} = \frac{\tau^2}{n}, \quad (6.18)$$

其中用到了 (6.7) 式 $E[\hat{\tau}]$ 的结果。由于 τ^2/n 就是严格计算得到的方差 (见 (6.15) 式), 所以 RCF 不等式中的等号成立, 而 $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$ 为参数 τ 的有效估计量。

如果参数不止一个, 以矢量形式表示为 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$, 估计量的协方差矩阵 $V_{ij} \equiv \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$ 的逆可以表示为 (假设估计量无偏且有效)

$$(V^{-1})_{ij} = E \left[-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right]. \quad (6.19)$$

(6.19) 式也可以写作

$$\begin{aligned} (V^{-1})_{ij} &= \int \cdots \int -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \left(\sum_{k=1}^n \log f(x_k; \boldsymbol{\theta}) \right) \prod_{l=1}^n f(x_l; \boldsymbol{\theta}) dx_l \\ &= n \cdot \int -f(x; \boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x; \boldsymbol{\theta}) dx, \end{aligned} \quad (6.20)$$

其中 $f(x; \boldsymbol{\theta})$ 为随机变量 x 的概率密度函数, 并且观测次数为 n 。也就是说, 协方差矩阵的 RCF 边界的逆正比于样本容量 n 。协方差矩阵的 RCF 边界也称为**费舍尔信息矩阵(Fisher information matrix)**, 参见文

献 [23] 的第 5.2 节和文献 [3]。由于 $V^{-1} \propto n$ ，所以 $V \propto 1/n$ 。因此，(6.20) 式给出了一个常用的结论：统计不确定度（即标准差）按照 $1/\sqrt{n}$ 的比例减小（至少对于有效估计量成立）。

实际上，很多时候要解析计算 RCF 边界几乎不可行，因为需要计算对数似然函数二阶导数的期望值（即需要对变量 x 积分）。在数据样本足够大的情况下，我们可以用观测数据和极大似然估计值 $\hat{\theta}$ 计算二阶导数，以此估计 V^{-1} ：

$$(V^{-1})_{ij} = - \left. \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right|_{\theta=\hat{\theta}}. \quad (6.21)$$

如果只有一个参数 θ ，上式简化为

$$\hat{\sigma}_{\hat{\theta}}^2 = \left(-1 / \left. \frac{\partial^2 \log L}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \right). \quad (6.22)$$

用数值方法求解似然函数极大值时，这是估计协方差矩阵的常用方法。¹

6.7 极大似然估计量的方差：图解法

前面讨论的利用 RCF 边界求估计量方差的方法，可以简单推广得到图解法求极大似然估计量方差的方法。考虑只有一个参数 θ 的情况，将对数似然函数围绕极大似然估计量 $\hat{\theta}$ 作泰勒展开：

$$\log L(\theta) = \log L(\hat{\theta}) + \left[\frac{\partial \log L}{\partial \theta} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!} \left[\frac{\partial^2 \log L}{\partial \theta^2} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots \quad (6.23)$$

根据 $\hat{\theta}$ 的定义，我们知道 $\log L(\hat{\theta}) = \log L_{\max}$ ，并且展开式的第二项等于零。利用 (6.22) 式，并忽略高阶项，可以得到

$$\log L(\theta) = \log L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}_{\hat{\theta}}^2}, \quad (6.24)$$

或者

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{\max} - \frac{1}{2}. \quad (6.25)$$

也就是说，如果参数 θ 偏离极大似然估计值一个标准差，对应的对数似然函数的值比最大值小 1/2。

可以证明，在大样本极限下，对数似然函数变成抛物线（即，似然函数变成高斯曲线）。即使 $\log L$ 不是抛物型，我们仍然可以采用 (6.25) 式作为统计不确定度的定义。这种不确定度的解释将在第九章作深入讨论。

作为图解法确定估计量方差的应用，让我们再回到 6.2 节和 6.5 节关于指数分布的例子。图 6.4 给出了包含 50 次观测的蒙特卡罗实验中，对数似然函数 $\log L(\tau)$ 作为参数 τ 的函数。 $\hat{\tau}$ 的标准差通过改变 τ 直到 $\log L(\tau)$ 比极大值减小 1/2 来确定，结果为 $\Delta\hat{\tau}_- = 0.137$ ， $\Delta\hat{\tau}_+ = 0.165$ 。在这个例子中， $\log L(\tau)$ 可以近似看成抛物线，所以近似有 $\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$ 。这个结果与利用严格的标准差 τ/\sqrt{n} 计算的结果基本相同，计算时取 $\tau = \hat{\tau}$ 。在第九章，我们将重新解释区间 $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ ，它近似为 68.3% **中心置信区间(central confidence interval)**（参见 9.6 节）。

¹例如，程序包 MINUIT [25, 48] 中的子程序 MIGRAD 和 HESSE，利用有限差分的方法，数值确定 $\log L$ 二阶导数的矩阵，计算其在极大似然估计处的值，然后通过求逆得到协方差矩阵。

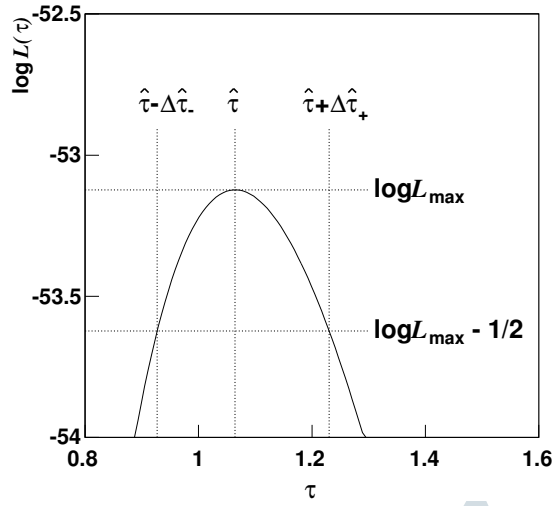


图 6.4: 对数似然函数 $\log L(\tau)$ 。在大样本极限下, 区间 $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau}]$ 和 $[\hat{\tau}, \hat{\tau} + \Delta\hat{\tau}_+]$ 的宽度对应于一个标准差 $\hat{\sigma}_{\hat{\tau}}$ 。

6.8 双参数的极大似然估计

作为双参数极大似然法的例子, 让我们考虑粒子散射的反应, 每个散射事例都由某个散射角 θ 来表征 (或者等价地用 $x = \cos \theta$ 表示)。假设某给定理论对角分布的预言为

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}. \quad (6.26)$$

(例如, $\alpha = 0$ 且 $\beta = 1$ 对应于量子电动力学中 $epem \rightarrow \mu^+ \mu^-$ 反应的最低阶效应 [37]。)注意到, 分母 $2 + 2\beta/3$ 是为了使 $f(x; \alpha, \beta)$ 在区间 $-1 \leq x \leq 1$ 内归一化。

为了使问题更复杂一些 (也更符合实际情况), 假设测量只能在限定的区域进行, 比如 $x_{\min} \leq x \leq x_{\max}$ 的区域。这个限制条件要求重新计算归一化因子, 容易给出

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{(x_{\max} - x_{\min}) + \frac{\alpha}{2}(x_{\max}^2 - x_{\min}^2) + \frac{\beta}{3}(x_{\max}^3 - x_{\min}^3)}. \quad (6.27)$$

图6.5给出了包含 2000 个事例的直方图, 这些事例通过蒙特卡罗实验产生, 模拟所用的参数为 $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$ 。用数值方法求对数似然函数的最大值, 可以得到

$$\begin{aligned} \hat{\alpha} &= 0.508 \pm 0.052, \\ \hat{\beta} &= 0.47 \pm 0.11, \end{aligned} \quad (6.28)$$

其中统计不确定度为方差的平方根。在6.6节, 我们已经对方差进行了估计, 先 (数值) 计算对数似然函数对参数的二阶导数矩阵, 然后求逆得到协方差矩阵。用同样的方法, 我们可以得到协方差的估计值 $\text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$, 或者等价地用相关系数表示 $r = 0.46$ 。可以看到, 估计量 $\hat{\alpha}$ 和 $\hat{\beta}$ 存在正相关。需要注意的是, 整个过程并没有用到直方图本身, 计算似然函数时直接使用了每个 x 的具体值。

为了更形象地理解这些结果, 可以进行 500 次类似的蒙特卡罗实验, 每次实验都包含 $\alpha = 0.5$ 和 $\beta = 0.5$ 的 2000 个事例。图6.6(a) 给出了极大似然估计量 $\hat{\alpha}$ 和 $\hat{\beta}$ 的散点分布, 其中散点的密度对应于 $\hat{\alpha}$ 和 $\hat{\beta}$ 的联合概率密度函数。图6.6(b) 和 (c) 分别给出了 $\hat{\alpha}$ 和 $\hat{\beta}$ 的投影直方图, 对应于边缘概率密度函数, 即对所有 $\hat{\beta}$ 积分后 $\hat{\alpha}$ 的分布, 以及对所有 $\hat{\alpha}$ 积分后 $\hat{\beta}$ 的分布。可以看到, $\hat{\alpha}$ 和 $\hat{\beta}$ 的边缘概率密度函数都近似服从高斯分布。

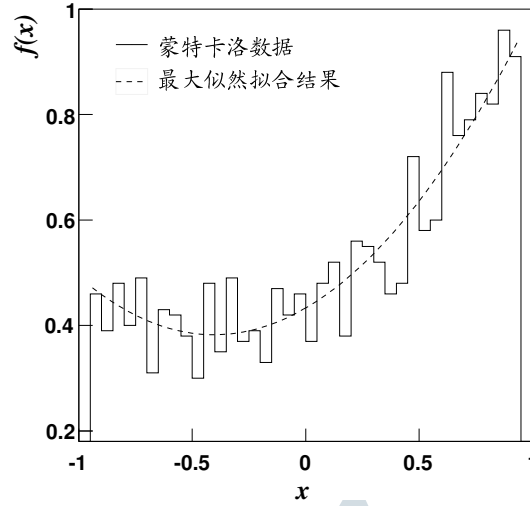


图 6.5: 根据 (6.27) 式的分布进行 2000 次蒙特卡罗模拟得到的直方图, 其中 $\alpha = 0.5, \beta = 0.5$ 。虚线为极大似然拟合的结果, $\hat{\alpha} = 0.508 \pm 0.052$, $\hat{\beta} = 0.466 \pm 0.108$ 。不确定度是利用 (6.21) 式数值计算得到。

从蒙特卡罗实验得到的样本均值、样本标准差、样本协方差以及样本相关系数 (见 5.2 节) 为

$$\begin{aligned} \bar{\hat{\alpha}} &= 0.499 & \bar{\hat{\beta}} &= 0.498 \\ s_{\hat{\alpha}} &= 0.051 & s_{\hat{\beta}} &= 0.111 \\ \text{cov}[\hat{\alpha}, \hat{\beta}] &= 0.0024 & r &= 0.42. \end{aligned} \quad (6.29)$$

可以看到, $\bar{\hat{\alpha}}$ 和 $\bar{\hat{\beta}}$ 与蒙特卡罗输入的 ‘真值’ ($\alpha = 0.5$ 和 $\beta = 0.5$) 符合得很好, 样本方差和样本协方差也与前面用 RCF 边界法估计的结果非常接近。

$\hat{\alpha}$ 和 $\hat{\beta}$ 两个参数存在相关性, 这一点可以从 $\hat{\alpha}$ 和 $\hat{\beta}$ 的散点分布图直接看出: 散点图中散点的分布带是倾斜的。也就是说, 如果我们要求 $\hat{\alpha} > \alpha$, 则会导致 $\hat{\beta} > \beta$ 的概率增大。换句话说, 相比于 $\hat{\alpha}$ 的边缘概率密度函数, 给定 $\hat{\beta} > \beta$ 条件下 $\hat{\alpha}$ 的条件概率密度函数的均值更高, 方差更小。

图 6.7 给出了参数空间中极大似然估计值的位置, 以及对应于 $\log L = \log L_{\max} - 1/2$ 的等值线。

在大样本极限下, 对数似然函数的形式为

$$\begin{aligned} \log L(\alpha, \beta) &= \log L_{\max} \\ &- \frac{1}{2(1-\rho^2)} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right], \end{aligned} \quad (6.30)$$

其中 $\rho = \text{cov}[\hat{\alpha}, \hat{\beta}] / (\sigma_{\hat{\alpha}} \sigma_{\hat{\beta}})$ 为 $\hat{\alpha}$ 和 $\hat{\beta}$ 的相关系数。因此, 等值线 $\log L = \log L_{\max} - 1/2$ 由下面方程给出

$$\frac{1}{(1-\rho^2)} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right] = 1. \quad (6.31)$$

这是一个椭圆方程, 椭圆的中心为极大似然估计值 $(\hat{\alpha}, \hat{\beta})$, 并且椭圆相对于 α 轴转动了角度 ϕ , ϕ 满足关系式

$$\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}. \quad (6.32)$$

需要特别注意的是, 椭圆垂直和水平方向的切线分别对应于 $\alpha = \hat{\alpha} \pm \sigma_{\hat{\alpha}}$ 和 $\beta = \hat{\beta} \pm \sigma_{\hat{\beta}}$ (见图 6.7)。如果两个参数的估计量是相关的, 则将其中一个参数改变一倍标准差, 对数似然值的减小通常会超过 1/2。如果

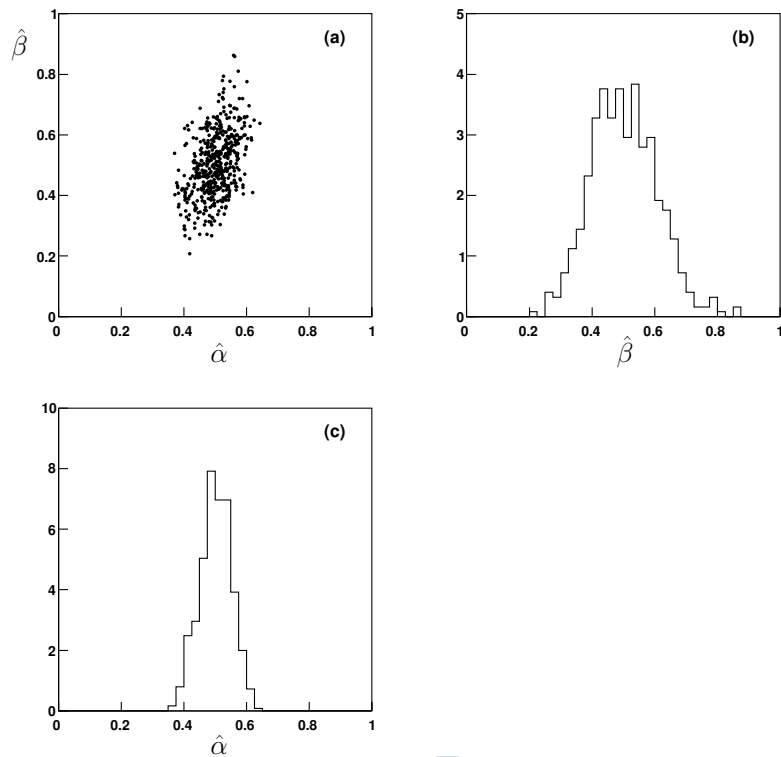


图 6.6: 对 500 个蒙特卡罗数据样本进行极大似然拟合得到结果。(a) $\hat{\alpha}$ 和 $\hat{\beta}$ 的拟合值。(b) $\hat{\beta}$ 的边缘分布。(c) $\hat{\alpha}$ 的边缘分布。

已知其中某个参数，例如 β ，则 $\hat{\alpha}$ 的标准差将会小一些，因为这个标准差将根据 $\log L(\alpha)$ 减小 $1/2$ 确定。类似地，如果拟合中还包含其它参数 (γ, δ, \dots) ，并且它们的估计量与 $\hat{\alpha}$ 相关，那么这将导致 $\hat{\alpha}$ 的标准差增大。

6.9 扩展的极大似然估计

考虑随机变量 x ，其概率密度函数为 $f(x; \theta)$ ，参数 $\theta = (\theta_1, \dots, \theta_m)$ 未知。假设数据样本为 x_1, \dots, x_n 。经常会遇到一种情形，样本中观测次数 n 本身是均值为 ν 的泊松随机变量。实验的结果可以用观测次数 n 以及 n 个数据 x_1, \dots, x_n 表示。于是，似然函数可以表示为两部分的乘积，一部分是观测次数为 n 的泊松概率（即 (2.9) 式），另一部分是正常的 x 的 n 个观测值的似然函数。即

$$L(\nu, \theta) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^n f(x_i; \theta) = \frac{e^{-\nu}}{n!} \prod_{i=1}^n \nu f(x_i; \theta). \quad (6.33)$$

上式称为**扩展的似然函数(extended likelihood function)**。实际上，这与正常的似然函数没什么差别，只是样本容量 n 变成了实验结果的一部分。我们对两种情形很感兴趣，一种是泊松参数 ν 为 θ 的函数，另一种是泊松参数 ν 被当作独立参数。

首先假设 ν 给定为 θ 的函数。这种情况下，扩展的对数似然函数为

$$\begin{aligned} \log L(\theta) &= n \log \nu(\theta) - \nu(\theta) + \sum_{i=1}^n \log f(x_i; \theta) \\ &= -\nu(\theta) + \sum_{i=1}^n \log(\nu(\theta) f(x_i; \theta)), \end{aligned} \quad (6.34)$$

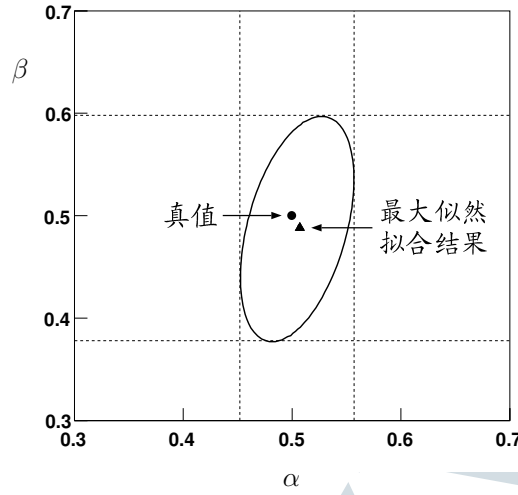


图 6.7: 对数似然函数 $\log L = \log L_{\max} - 1/2$ 的等值线, 以及参数 (α, β) 的真值和极大似然拟合结果 $(\hat{\alpha}, \hat{\beta})$ 。在大样本极限下, 曲线的切线对应于 $\hat{\alpha} \pm \hat{\sigma}_{\hat{\alpha}}$ 和 $\hat{\beta} \pm \hat{\sigma}_{\hat{\beta}}$ 。

其中已经忽略了与参数无关的附加项。(因为估计量只依赖于 $\log L$ 的导数, 所以与参数无关的附加项对估计量没有贡献。)通过考虑泊松项, 得到的估计量 $\hat{\theta}$ 既包含 n 的信息, 也包含变量 x 的信息。通常情况下, 这会导致 $\hat{\theta}$ 的方差比只利用 x 的信息时更小。

例如, 在粒子散射的反应中, 总反应截面以及用来描述反应事例性质的某个变量的分布 (比如出射粒子的角分布) 可能依赖于粒子质量和耦合常数等参数。一般来说, 如果同时考虑反应截面和角分布的信息, 则参数估计值的统计不确定度会相对小一些。总反应截面与泊松参数 ν 的关系为 $\nu = \sigma L \varepsilon$, 其中 L 为数据的积分亮度, ε 为事例被探测到的概率 (即探测效率)。估计量的标准差对应于大量重复实验中估计值涨落的大小。重复实验中, 要求每次实验的积分亮度相同, 但不要求事例数相同。

我们还关心一种情形, ν 和 θ 之间没有对应的函数关系。对 (6.33) 式取对数, 并令其对 ν 的导数为零, 不难猜到相应的估计量为

$$\hat{\nu} = n. \quad (6.35)$$

令 $\log L(\nu, \theta)$ 对 θ_i 的导数等于零, 得到的估计量 $\hat{\theta}_i$ 与正常的极大似然法相同。因此, 这种情形本质上与前面讨论的情形完全相同, 只是在这种情况下, 如果一个量同时依赖于 n 和 $\hat{\theta}$, 它将有额外的统计涨落来源, 因为 n 也被视为随机变量。

然而, 这种情况下的某些问题, 使用扩展的似然函数仍然会有意义。随机变量 x 的概率密度函数通常是几个分量的叠加,

$$f(x; \theta) = \sum_{i=1}^m \theta_i f_i(x), \quad (6.36)$$

而我们的目标是估计 θ_i , 即每个分量相对贡献的大小。假设概率密度函数 $f_i(x)$ 全部已知。这里, 参数 θ_i 不相互独立, 约束条件是所有 θ_i 之和为 1。通常情况下, 不采用扩展的似然函数, 可以将某个系数替换一下, 例如用 $1 - \sum_{i=1}^{m-1} \theta_i$ 替换 θ_m , 于是概率密度函数只包含 $m-1$ 个参数。然后, 我们可以构造似然函数, 并以此求出 θ_i 的估计量。

然而, 如果采用 (6.33) 式扩展的似然函数, 可以等价地但以更加对称的方式处理这个问题。对 (6.33) 式取对数, 并忽略与参数无关的项, 可以得到

$$\log L(\nu, \theta) = -\nu + \sum_{i=1}^n \log \left(\sum_{j=1}^m \nu \theta_j f_j(x_i) \right). \quad (6.37)$$

定义 $\mu_i = \theta_i \nu$ 为第 i 类事例的期待事例数，对数似然函数可以表示为 m 个参数 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)$ 的函数，

$$\log L(\boldsymbol{\mu}) = - \sum_{j=1}^m \mu_j + \sum_{i=1}^n \log \left(\sum_{j=1}^m \mu_j \theta_j f_j(x_i) \right). \quad (6.38)$$

这种情况下，参数 $\boldsymbol{\mu}$ 不再象 $\boldsymbol{\theta}$ 的分量那样受到约束。总事例数 n 被看作均值为 μ_i 的独立泊松变量之和。估计量 $\hat{\mu}_i$ 直接给出对不同类别事例平均事例数的估计值。这当然等价于利用各分量的极大似然估计量 $\hat{\theta}_i$ 以及泊松参数 ν 的估计量 $\hat{\nu} = n$ 。但是，在这种方法中，所有参数的处理是对称的。而且，参数 μ_i 往往与期待的最终结果联系更为紧密，例如，第 i 类事例的产生截面。

如果 (6.36) 式中不同的项表示对样本有贡献的不同事例类型，那么我们可以假设所有 θ_i 均大于或等于零。也就是说，第 i 类事例可以对样本有贡献，但不能导致 x 分布的某个区域内事例系统减少。我们也可以认为 (6.36) 式中某些 θ_i 可以是负的，例如，某些 $f_i(x)$ 表示量子力学干涉效应。不过在这里，我们只考虑所有 θ_i 都不小于零的情况。即使这样，由于数据的统计涨落，可能导致似然函数在某些 θ_i 小于零时取最大值，因此对应的 μ_i 为负。这种情况下，我们必须明确如何报告估计结果。

作为例子，考虑包含两种事例（例如信号和本底）的数据样本，其中每种事例都由连续随机变量 x 描述。假设信号事例的概率密度函数 $f_s(x)$ 为高斯分布，而本底事例的概率密度函数 $f_b(x)$ 为指数分布。假设信号事例数 n_s 服从均值为 μ_s 的泊松分布，本底事例数 n_b 服从均值为 μ_b 的泊松分布。

因此， x 的概率密度函数可以写为

$$f(x) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x), \quad (6.39)$$

其中假定已知概率密度函数 $f_s(x)$ 和 $f_b(x)$ 的形式，并且在固定区间内（在本例子中这个固定区间为 $0 \leq x \leq 2$ ）归一化到单位面积。假设给定 $n = n_s + n_b$ 个 x 的值，我们要以此估计 μ_s 。本底的期望值 μ_b 可能已知，但在某些情况下， μ_b 也需要用数据确定。在这个例子中，我们同时拟合 μ_s 和 μ_b 。图6.8给出了蒙特卡罗产生的两个数据样本（取 $\mu_s = 6$ 和 $\mu_b = 60$ ），以及扩展的极大似然拟合的结果。在图6.8(a)中，估计的信号数为 $\hat{\mu}_s = 8.7$ ，并且通过求对数似然函数二阶导数估计得到的 $\hat{\mu}_s$ 的标准差为 5.5。由于标准差与信号的估计值大小相当，所以可以想象这样的实验拟合可能得到负的信号估计值。实际上，图6.8(b) 就是这种结果，信号估计值 $\hat{\mu}_s = -1.8$ 。

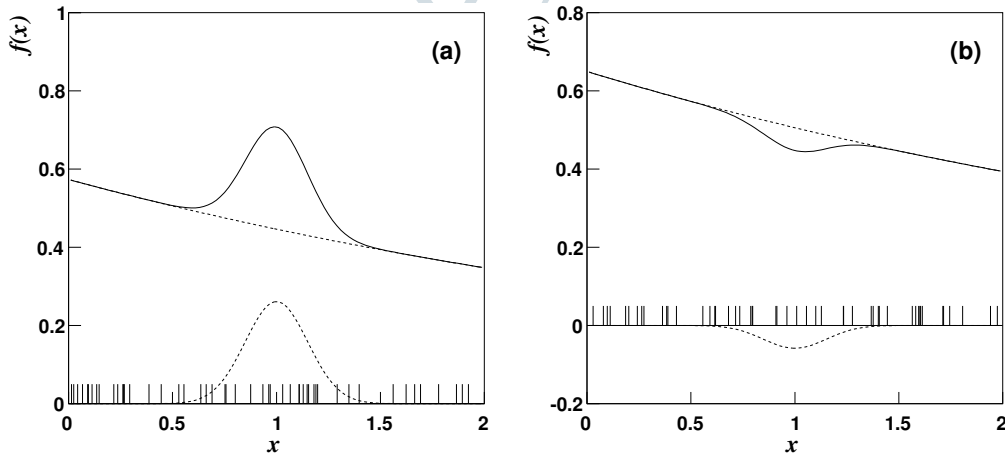


图 6.8: 对两个数据样本进行扩展的极大似然拟合，数据的分布是高斯概率密度与指数概率密度的叠加。由于数据的统计涨落，对高斯分量大小的估计可以为正 (a)，也可以为负 (b)。

在负估计值没有物理意义的情况下，我们可能会这样选择估计量：如果估计值为正，则采用拟合值；否则，就报告零结果。但是，这种估计量的问题在于它是有偏估计量。如果重复实验很多次，有些极大似

然估计为负，有些估计为正，但是平均结果将收敛于真值。(原则上，极大似然估计量仍然可能有一定的偏倚，但是如果每次实验的数据样本足够大，那么这个偏倚就会很小，并且在任何情况下都将远小于对所有负估计值取零带来的偏倚。)

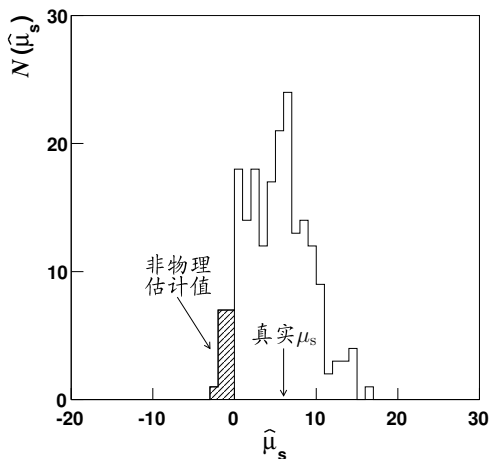


图 6.9: 200 次蒙特卡罗实验得到的 $\hat{\mu}_s$ 的估计值的直方图，蒙特卡罗模拟所用的真值为 $\mu_s = 6$ 。估计值的平均为 6.1 ± 0.4 ，大约 10% 的估计值为负。

例如，图6.9给出了 200 次上述蒙特卡罗实验得到的估计值 $\hat{\mu}_s$ 的直方图。 $\hat{\mu}_s$ 的平均值为 6.1，与真值 6 很接近。200 次实验的样本标准差为 5.3，与前面利用对数似然函数估计的标准差接近。因此， $\hat{\mu}_s$ 的平均值的标准差为 $5.3/\sqrt{200} = 0.37$ 。如果将所有负的估计值都取为零，则 $\hat{\mu}_s$ 的平均值变为 6.4。在本例中，这种做法导致的偏倚 $6.4 - 6 = 0.4$ 比标准差 5.3 要小很多，但是如果与多次实验的平均结果的标准差相比，这个偏倚可能会变得显著。

因此，如果想要将实验结果与其它实验的结果平均起来，对实验结果进行无偏报道非常重要，即使实验结果在非物理区域。此外，我们也可能希望给出参数的上限，第9.8节将讨论参数上限的估计方法。

6.10 分区间数据的极大似然估计

假设随机变量 x 的分布服从概率密度函数 $f(x; \theta)$ ，考虑对 x 的 n_{tot} 次观测，我们希望以此估计未知参数 $\theta = (\theta_1, \dots, \theta_m)$ 。如果数据样本非常大，对数似然函数的计算将非常困难，因为必须对所有 x_i 的 $\log f(x_i, \theta)$ 求和。这种情况下，通常不记录每次测量的具体值，而是将数据做成直方图，得到 N 个区间的频数 $\mathbf{n} = (n_1, \dots, n_N)$ 。频数的期望值 $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$ 为

$$\nu_i(\theta) = n_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \theta) dx, \quad (6.40)$$

其中 x_i^{\min} 和 x_i^{\max} 分别表示区间的下限和上限。可以将这个直方图看作对 N 维随机矢量的单次测量，这个 N 维矢量的联合概率密度函数由 (2.6) 式的多项分布给出，

$$f_{\text{joint}}(\mathbf{n}; \boldsymbol{\nu}) = \frac{n_{\text{tot}}!}{n_1! \cdots n_N!} \left(\frac{\nu_1}{\nu_{\text{tot}}} \right)^{n_1} \cdots \left(\frac{\nu_N}{\nu_{\text{tot}}} \right)^{n_N}. \quad (6.41)$$

测量值处于区间 i 的概率等于期望值 ν_i 除以总频数 n_{tot} 。取联合概率密度函数的对数得到对数似然函数，

$$\log L(\theta) = \sum_{i=1}^N n_i \log \nu_i(\theta), \quad (6.42)$$

其中已经省略了与参数无关的相加项。估计量 $\hat{\theta}$ 可以通过使 $\log L$ 最大化求得，不论使用哪种最大化方法都可以，例如数值求解。在区间宽度非常小的极限下（即 N 非常大），这个似然函数与非分区的极大似然法完全相同（(6.2) 式）。因此，即使某些区间只有很少事例甚至没有事例，分区间的极大似然法也没有任何问题。这与7.5节讨论的最小二乘法截然相反。

作为例子，重新考虑6.2节讨论的数据样本，即对粒子衰变时间的 50 次测量，见图6.2。图6.10给出了同一个样本的直方图表示，直方图的区间宽度为 $\Delta t = 0.5$ 。图6.10还给出了极大似然拟合的结果（似然函数基于 (6.42) 式）。拟合结果为 $\hat{\tau} = 1.067$ ，与非分区的结果 $\hat{\tau} = 1.062$ 符合得非常好。利用对数似然函数在最大值处的曲率对标准差进行估计（(6.22) 式），得到 $\hat{\sigma}_{\hat{\tau}} = 0.171$ ，这一结果比非分区极大似然法稍大。

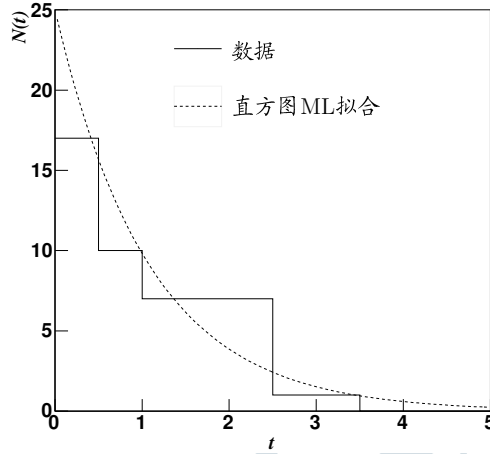


图 6.10: 6.2 节粒子衰变时间数据样本（观测到 50 次）的直方图，虚线为极大似然拟合结果。

在6.9节讨论过，很多问题中，我们想把总事例数 n_{tot} 当作均值为 ν_{tot} 的泊松随机变量。也就是说，测量包含两部分，首先根据泊松分布确定 n_{tot} ，然后将 x 的 n_{tot} 次测量做成区间数为 N 的直方图，得到 $\mathbf{n} = (n_1, \dots, n_N)$ 。 n_{tot} 和 n_1, \dots, n_N 的联合概率密度函数为泊松分布与多项分布的乘积，

$$f_{\text{joint}}(\mathbf{n}; \nu) = \frac{\nu_{\text{tot}}^{n_{\text{tot}}} e^{-\nu_{\text{tot}}}}{n_{\text{tot}}!} \frac{n_{\text{tot}}!}{n_1! \cdots n_N!} \left(\frac{\nu_1}{n_{\text{tot}}} \right)^{n_1} \cdots \left(\frac{\nu_N}{n_{\text{tot}}} \right)^{n_N}, \quad (6.43)$$

其中 $\nu_{\text{tot}} = \sum_{i=1}^N \nu_i$ ， $n_{\text{tot}} = \sum_{i=1}^N n_i$ 。将这两个关系式代入 (6.43) 式得到

$$f_{\text{joint}}(\mathbf{n}; \nu) = \prod_{i=1}^N \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}, \quad (6.44)$$

现在，每个区间事例数的期望值 ν_i 依赖于参数 θ 和 ν_{tot} ，

$$\nu_i(\nu_{\text{tot}}, \theta) = \nu_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \theta) dx. \quad (6.45)$$

从联合概率密度函数 (6.44) 可以看出，这个问题等价于将每个区间的事例数看作均值为 ν_i 的独立泊松随机变量 n_i 。对联合概率密度函数取对数，忽略参数无关的项，可以得到

$$\log L(\nu_{\text{tot}}, \theta) = -\nu_{\text{tot}} + \sum_{i=1}^N n_i \log \nu_i(\nu_{\text{tot}}, \theta). \quad (6.46)$$

这就是对应于分区间数据的扩展的对数似然函数，参见 (6.33) 式和 (6.37) 式。

之前讨论的关于 ν_{tot} 与其它参数 θ 之间依赖性的考虑，在这里同样使用。也就是说，如果 ν_{tot} 与 θ 之间不存在函数关系，则可以得到 $\hat{\nu}_{\text{tot}} = n_{\text{tot}}$ ，估计量 $\hat{\theta}$ 与不考虑 n_{tot} 的泊松项的情形完全相同。如果 ν_{tot} 为 θ 的函数，则估计量 $\hat{\theta}$ 的方差通常会因为引入了 n_{tot} 的信息而变小。

6.11 极大似然法的拟合优度检验

尽管极大似然原则提供了参数估计的方法，它本身并不直接给出拟合优度检验的方法。一种可能性是，利用似然函数的最大值 L_{\max} 作为拟合优度的统计量。然而，这种方法并不简单，因为我们无法先验地知道，假定概率密度函数的形式正确的前提下 L_{\max} 如何分布。

L_{\max} 的概率密度函数可以通过蒙特卡罗方法确定。用于产生数据的‘真实’蒙特卡罗参数，可以采用真实实验的极大似然估计结果。以6.8节讨论的散射实验为例，图6.11给出了蒙特卡罗模拟得到的 $\log L_{\max}$ 的分布。最初的例子(即图6.5的数据样本)给出 $\log L_{\max} = 2436.4$ 。以此，我们可以象4.5节介绍的那样，计算观测的显著水平 (P -值)，将其作为拟合优度的量度。对于本例，可以得到 $P = 0.63$ ，这个结果表明没有任何迹象反对拟合所用的概率密度函数形式。

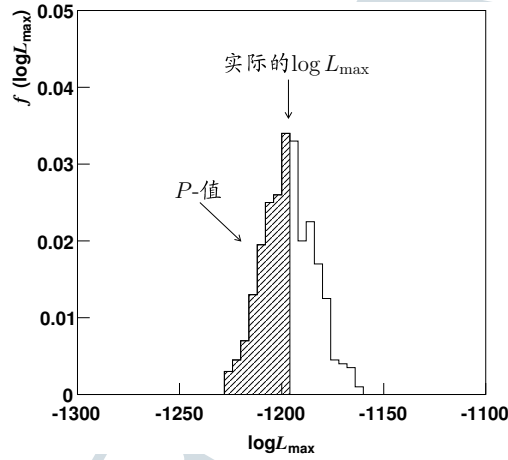


图 6.11: 5000 次蒙特卡罗实验得到的极大似然函数值 $\log L_{\max}$ 的直方图(已归一化)。竖直线标明了用图6.5中的数据得到的 $\log L_{\max}$ 值。

另一种方法是利用 n_{tot} 个测量值构造分 N 个区间的直方图 $\mathbf{n} = (n_1, \dots, n_N)$ 。均值 $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$ 的极大似然拟合估计为

$$\hat{\nu}_i = n_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \hat{\boldsymbol{\theta}}) dx, \quad (6.47)$$

其中参数 $\boldsymbol{\theta}$ 用它们的极大似然估计值计算。当然，对于非分区的极大似然拟合，我们也可以这么计算。直方图的优点是可以图形显示，并且初步给出数据与拟合结果的视觉比较。更量化的层面上，可以用数据 \mathbf{n} 和估计量 $\hat{\boldsymbol{\nu}}$ (或其它假设的 $\boldsymbol{\nu}$ 值) 构造拟合优度检验的统计量。

仍以6.10节中分区间的极大似然拟合为例，用似然函数构造这样的检验统计量。对于固定样本容量 n_{tot} 的情况，这个检验统计量即 (6.41) 式的多项分布概率密度函数；如果 n_{tot} 是均值为 ν_{tot} 的泊松变量，则检验统计量为 (6.44) 式泊松概率的乘积。首先考虑似然函数的比率

$$\lambda = \frac{L(\mathbf{n}|\boldsymbol{\nu})}{L(\mathbf{n}|\mathbf{n})} = \frac{f_{\text{joint}}(\mathbf{n}; \boldsymbol{\nu})}{f_{\text{joint}}(\mathbf{n}; \mathbf{n})}, \quad (6.48)$$

其中明显写出似然函数 $L(\mathbf{n}|\boldsymbol{\nu}) = f_{\text{joint}}(\mathbf{n}; \boldsymbol{\nu})$ 的目的是，强调似然函数同时依赖于数据 \mathbf{n} 和参数 $\boldsymbol{\nu}$ 。也就是说，(6.48) 式分母中直接令使似然函数中的 ν_i 等于 n_i 。对于多项分布的数据，这个比率变成

$$\lambda_M = \prod_{i=1}^N \left(\frac{\nu_i}{n_i} \right)^{n_i}, \quad (6.49)$$

而对于泊松分布的数据，可以得到

$$\lambda_P = e^{n_{\text{tot}} - \nu_{\text{tot}}} \prod_{i=1}^N \left(\frac{\nu_i}{n_i} \right)^{n_i}, \quad (6.50)$$

如果从数据中估计 m 个参数，则可以用 $\hat{\nu}$ 替换 ν 。如果假设是正确的，则在大样本极限下，多项分布的数据的检验统计量

$$\chi_M^2 = -2 \log \lambda_M = 2 \sum_{i=1}^N n_i \log \frac{n_i}{\hat{\nu}_i} \quad (6.51)$$

服从自由度为 $N - m - 1$ 的卡方分布(参见 [23,50] 及其中引用的其它参考文献)。对于泊松分布的数据，检验统计量

$$\chi_P^2 = -2 \log \lambda_P = 2 \sum_{i=1}^N \left(n_i \log \frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i \right) \quad (6.52)$$

服从自由度为 $N - m$ 的卡方分布。如果任何一个 n_i 等于零，则这两个量的定义都不再成立，但是这种情况下 λ 中的因子 $n_i^{n_i}$ 取为 1，对应的项对 (6.51) 式和 (6.52) 式没有贡献。

$\lambda(\nu) = L(\mathbf{n}|\nu)/L(\mathbf{n}|\mathbf{n})$ 这个量只与似然函数相差一个因子 $L(\mathbf{n}|\mathbf{n})$ ，并且这个因子与参数无关。因此，使 $\lambda(\nu)$ 取最大值的参数等于极大似然估计量。于是， $\lambda(\nu)$ 既可以用来构造估计量，也可以用来进行拟合优度检验。

我们也可以选择4.7节的某个卡方检验统计量。如果 n_{tot} 为泊松变量，则从 (4.39) 式可以得到

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \hat{\nu}_i)^2}{\hat{\nu}_i}, \quad (6.53)$$

或者，如果 n_{tot} 为固定值，则从 (4.41) 式可以得到

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \hat{p}_i n_{\text{tot}})^2}{\hat{p}_i n_{\text{tot}}}. \quad (6.54)$$

这里， $\hat{p}_i = \hat{\nu}_i / \hat{\nu}_{\text{tot}}$ 为某次测量处于区间 i 的概率的估计量。在大样本极限下，(6.53) 式的检验统计量服从自由度为 $N - m$ 的卡方分布，(6.54) 式的检验统计量服从自由度为 $N - m - 1$ 的卡方分布。

对于有限数据样本，上面的检验统计量都不严格服从卡方分布。如果直方图包含 $n_i < 5$ 的区间，则可以通过蒙特卡罗研究确定检验统计量的真实概率密度分布，并以此计算 P -值。

6.12 用极大似然法合并实验测量

考虑某个实验，我们对随机变量 x 进行了 n 次测量， x 的概率密度函数 $f_x(x; \theta)$ 依赖于未知参数 θ 。假设在另外的实验中，我们对另一个随机变量 y 进行了 m 次测量， y 的概率密度函数 $f_y(y; \theta)$ 依赖于同一个参数 θ 。例如， x 可以是质子-反质子对撞中产生的正负电子对的不变质量，而 y 可以是正负缪子对的不变质量。这两个分布都在 Z^0 玻色子的质量 M_Z 附近有个质量峰，因此两个分布都包含参数 M_Z 。于是，为了更好地估计 M_Z ，我们希望对两个实验测量进行合并。

这两个实验一起，可以解释为对一个矢量的单次测量，这个矢量包含 n 个 x 值和 m 个 y 值。因此，似然函数为

$$L(\theta) = \prod_{i=1}^n f_x(x_i; \theta) \cdot \prod_{j=1}^m f_y(y_j; \theta) = L_x(\theta) \cdot L_y(\theta), \quad (6.55)$$

或者，等价地用对数似然函数表示， $\log L(\theta) = \log L_x(\theta) + \log L_y(\theta)$ 。

因此，只要可以得到每个实验的似然函数，就可以构造完整的似然函数，从而基于两个实验确定 θ 的极大似然估计量。当然，这一技术也适用于 x 和 y 是同一个随机变量的特殊情形，这种情况下两个样本 x_1, \dots, x_n 和 y_1, \dots, y_m 仅仅表示数据的不同子样本。

更常见的情况是，我们不报告似然函数本身，而只是对参数进行估计。假设测量 x 和 y 的两个实验给出参数 θ 的估计量分别为 $\hat{\theta}_x$ 和 $\hat{\theta}_y$ ，而这两个估计量本身也是随机变量，对应的概率密度函数分别为 $g_x(\hat{\theta}_x; \theta)$ 和 $g_y(\hat{\theta}_y; \theta)$ 。可以把这两个估计量看作单次实验的输出，结果用二维矢量 $(\hat{\theta}_x, \hat{\theta}_y)$ 表示。只要 $\hat{\theta}_x$ 和 $\hat{\theta}_y$ 相互独立，对数似然函数就是二者的对数似然函数之和，即

$$\log L(\theta) = \log g_x(\hat{\theta}_x; \theta) + \log g_y(\hat{\theta}_y; \theta). \quad (6.56)$$

对于大数据样本，概率密度函数 g_x 和 g_y 可以看作高斯分布，我们将估计出的标准差 $\hat{\sigma}_{\hat{\theta}_x}$ 和 $\hat{\sigma}_{\hat{\theta}_y}$ 分别报告为 $\hat{\theta}_x$ 和 $\hat{\theta}_y$ 的误差。在第七章将看到，在大数据样本的情况下，这个问题等价于最小二乘法，合并数据对 θ 的估计由加权平均给出

$$\hat{\theta} = \frac{\hat{\theta}_x / \hat{\sigma}_{\hat{\theta}_x}^2 + \hat{\theta}_y / \hat{\sigma}_{\hat{\theta}_y}^2}{1 / \hat{\sigma}_{\hat{\theta}_x}^2 + 1 / \hat{\sigma}_{\hat{\theta}_y}^2}, \quad (6.57)$$

方差的估计量为

$$\hat{V}[\hat{\theta}] = \frac{1}{1 / \hat{\sigma}_{\hat{\theta}_x}^2 + 1 / \hat{\sigma}_{\hat{\theta}_y}^2}. \quad (6.58)$$

这一方法显然可以推广到合并任意多个实验测量结果。

6.13 极大似然与贝叶斯估计量的关系

将极大似然法与贝叶斯统计的参数估计进行对比很有指导意义。在贝叶斯统计中，不确定度用主观概率定量表述(参见1.2节)。对于前面的例子， x 的测量结果和参数 θ 都被当作随机变量。我们对 θ 的认知用概率密度函数描述，这个概率密度函数表达了我们对参数取某个给定值的信心有多大。

重新考虑对随机变量 x 的 n 次测量，假设 x 的概率密度函数为 $f(x; \theta)$ ，依赖于未知参数 θ 。(贝叶斯方法很容易推广到多个参数 $\theta = (\theta_1, \dots, \theta_m)$ 。简单起见，这里只考虑单参数的情况。)注意到，对于给定某个参数值 θ ，似然函数是数据 $\mathbf{x} = (x_1, \dots, x_n)$ 的联合概率密度函数，可以写为

$$L(\mathbf{x}|\theta) = f_{\text{joint}}(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i; \theta). \quad (6.59)$$

我们希望得到的是，给定数据的情况下 θ 的条件概率密度函数 $p(\theta|\mathbf{x})$ 。利用 (1.26) 式的贝叶斯定理，可以从似然函数得到这个概率密度函数，

$$p(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)\pi(\theta)}{\int L(\mathbf{x}|\theta')\pi(\theta')d\theta'}. \quad (6.60)$$

其中 $\pi(\theta)$ 为 θ 的**验前概率密度(prior probability density)**，反映了在考虑数据之前对 θ 的认识程度， $p(\theta|\mathbf{x})$ 称为在给定数据 \mathbf{x} 的情况下 θ 的**验后概率密度(posterior probability density)**。

在贝叶斯统计中，我们对 θ 的所有认识都包含在验后概率密度 $p(\theta|\mathbf{x})$ 中。由于报告整个概率密度很难实现，尤其在 θ 是多维参数的时候，所以必须寻找一种合适的方法总结验后概率密度的信息。第一步是选取一个估计量，通常取 $p(\theta|\mathbf{x})$ 最大时的 θ 值(即验后概率密度的众数)。如果验前概率密度 $\pi(\theta)$ 取为常数，则 $p(\theta|\mathbf{x})$ 正比于似然函数 $L(\mathbf{x}|\theta)$ ，此时贝叶斯估计量和极大似然估计量是一致的。于是，极大似然估计量可以看作贝叶斯估计量取均匀验前概率密度的特殊形式。因此，我们需要进一步讨论均匀分布的验前概率密度 $\pi(\theta)$ 有什么深层意义。

(6.60) 式表示的贝叶斯方法给出，如何根据新获取的数据更新我们的认知程度。然而，需要指出的是，所谓认知程度是进行测量之前的认知程度。如果之前我们一无所知，可以先验地假设 θ 取任何值都是等几率的。这个假设有时称为**贝叶斯假设(Bayes' postulate)**，表示为 $\pi(\theta) = \text{常数}$ 。如果 θ 的取值范围无穷大，则常数的 $\pi(\theta)$ 无法归一化，称其为**非正当验前概率(improper prior)**。实际上，这通常不是什么问题，因为 $\pi(\theta)$ 的出现总伴随着乘以相应的似然函数，最终导致验后概率密度可归一化。然而，对于某些非正当验前概率密度，并不一定满足这种情况，参见第九章的 (9.45) 式。

在进行参数变换时，常数验前概率密度会导致更难解决的问题。例如，考虑在 $[0, 10]$ 区间内定义的连续参数 θ 。利用贝叶斯假设，(6.60) 式中的验前概率密度取为 $\pi(\theta) = 0.1$ ，以此计算验后概率密度 $p(\theta|\mathbf{x})$ 。然而，另一个实验者可能认为以某个非线性函数 $a(\theta)$ 作为参数更合适。通过变量变换，可以得到对应的概率密度 $p_a(a|\mathbf{x}) = p(\theta|\mathbf{x})|d\theta/da|$ 。与此对应，可以将似然函数直接表示为 a 的函数，并假设验前概率密度 $\pi_a(a)$ 为常数。例如，如果 $a = \theta^2$ ，则在区间 $[0, 100]$ 取 $\pi_a(a) = 0.01$ 。然而，将这个验前概率密度代入 (6.60) 式，得到的验后概率密度通常不同于用变量变换得到的 $p_a(a|\mathbf{x})$ 。也就是说，如果取 θ 的验前概率密度为常数 ($\pi_\theta(\theta) = \text{常数}$)，那么 θ 的非线性函数的验前概率密度就不是均匀分布 ($\pi_a(a) \neq \text{常数}$)。

如果 $\pi_a(a)$ 不是常数，则验后概率密度 $p_a(a|\mathbf{x})$ 的众数将不同于似然函数 $L_a(\mathbf{x}|a) = f_{\text{joint}}(\mathbf{x}|a)$ 取最大值的位置。即，贝叶斯估计量在参数变化下通常不能保持不变。然而，如6.2节所述，极大似然估计量在参数变化下是不变的。也就是说，对于极大似然估计量，如果 $\hat{\theta}$ 使 $L_\theta(\mathbf{x}|\theta)$ 取最大值，则使 $L_a(\mathbf{x}|a)$ 取最大值的 a 就等于 $a(\hat{\theta})$ 。

第七章 最小二乘法

7.1 与极大似然的联系

很多情形，测量值 y 可以看作以真值 λ 为中心值的高斯随机变量。从中心极限定理可以得到这一结论，只要总误差（即与真值的偏差）为大量小误差的贡献之和。

考虑 N 个独立的高斯随机变量 y_i , $i = 1, \dots, N$ ，每个 y_i 都对应于另一个假定已知并且无误差的变量 x_i 。例如，我们可能在不同位置 x_i 对温度进行了 N 次测量 $T(x_i)$ 。假设不同 y_i 的均值 λ_i 不同并且未知，其方差 σ_i^2 也不同但是大小已知。 N 次测量 y_i 可以等价地看作对 N 维随机矢量的单次测量，该矢量的联合概率密度函数为 N 个高斯分布的乘积，

$$g(y_1, \dots, y_N; \lambda_1, \dots, \lambda_N, \sigma_1^2, \dots, \sigma_N^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(y_i - \lambda_i)^2}{2\sigma_i^2}\right). \quad (7.1)$$

进一步假设真值 λ 为 x 的函数，即 $\lambda = \lambda(x; \theta)$ ，该函数依赖于未知参数 $\theta = (\theta_1, \dots, \theta_m)$ 。最小二乘法的目标是对参数 θ 进行估计。此外，最小二乘法很容易给出假设函数 $\lambda(x; \theta)$ 的拟合优度。图7.1说明了最小二乘法的基本要素。

对联合概率密度函数取对数并忽略参数无关的相加项，可以得到对数似然函数，

$$\log L(\theta) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2}. \quad (7.2)$$

这个对数似然函数的最大化，等价于求使下面这个量最小化的参数 θ 的值，

$$\chi^2(\theta) = \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2}, \quad (7.3)$$

即，以方差的倒数为权重，测量值与假设值之差的平方和。这是**最小二乘法(method of least squares, LS)**的基础，即使每个 y_i 不服从高斯分布，但只要它们之间相互独立，也可以这样定义。

如果测量不相互独立，但测量值可以用 N 维高斯概率密度函数描述，并且已知该分布的协方差矩阵 V ，而分布的均值未知，此时，相应的对数似然函数由 (2.28) 式的联合概率密度函数的对数给出，

$$\log L(\theta) = -\frac{1}{2} \sum_{i,j=1}^N (y_i - \lambda(x_i; \theta))(V^{-1})_{ij}(y_j - \lambda(x_j; \theta)), \quad (7.4)$$

其中已经省略了参数无关的相加项。该对数似然函数的最大化等价于求下面这个量的最小值，

$$\chi^2(\theta) = \sum_{i,j=1}^N (y_i - \lambda(x_i; \theta))(V^{-1})_{ij}(y_j - \lambda(x_j; \theta)), \quad (7.5)$$

如果协方差矩阵是对角的（则矩阵的逆也是对角的），上式退化为 (7.3) 式。

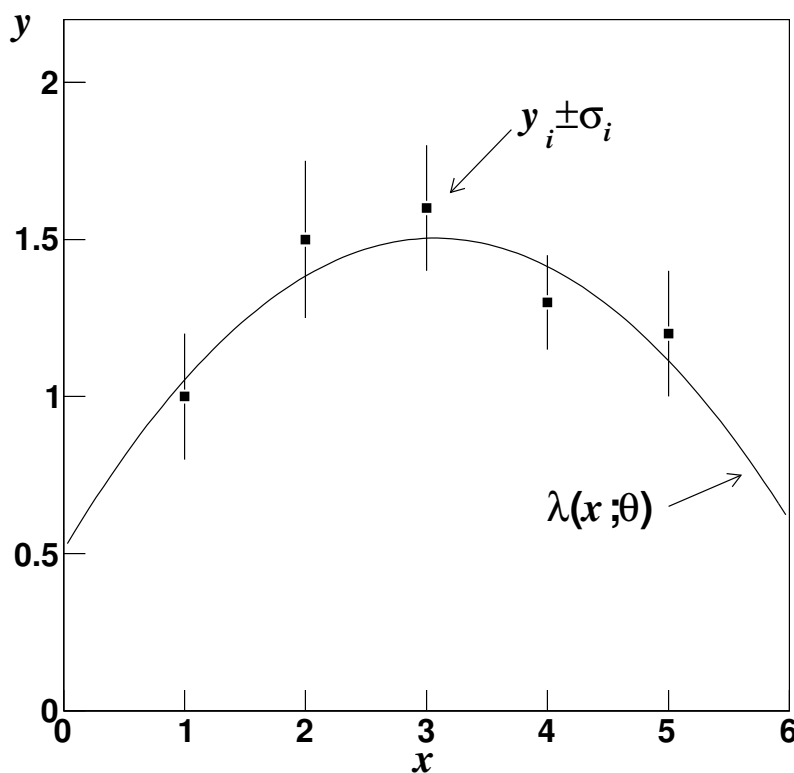


图 7.1: 最小二乘问题的几个要素: 在给定 x 值 x_1, \dots, x_N (无误差) 处测量到 N 个值 y_1, \dots, y_N , 对应的误差为 $\sigma_1, \dots, \sigma_N$ 。假定 y_i 的真值由函数 $\lambda_i = \lambda(x_i; \theta)$ 给出。调整 θ 的值使 (7.3) 式的 χ^2 值最小。

使 χ^2 取最小值的参数 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 称为最小二乘估计量。在 7.5 节将讨论, 某些条件下得到的最小 χ^2 , 服从 2.7 节定义的卡方分布。由于这个原因, (7.3) 式和 (7.5) 式定义的量通常称为卡方 (χ^2), 即使在更一般的条件下其最小值并不服从卡方分布。

7.2 线性最小二乘拟合

尽管对任意函数 $\lambda(x; \theta)$ 都可以应用最小二乘法, 但是当 $\lambda(x; \theta)$ 为参数 θ 的线性函数时, 得到的 χ^2 值和最小二乘估计量具有特别理想的性质, 即

$$\lambda(x; \theta) = \sum_{j=1}^m a_j(x) \theta_j, \quad (7.6)$$

其中 $a_j(x)$ 为 x 的任意线性独立函数。(要求 λ 是参数 θ_j 的线性函数, 而 $a_j(x)$ 一般来说不要求是 x 的线性函数, 只要它们之间相互独立即可, 即其中任意一个不能表示为其他函数的线性组合。) 在这种情况下, 可以解析求解最小二乘估计量及其方差, 尽管很多时候仍然会选择计算机数值求解 χ^2 的最小值。另外, 线性最小二乘估计量的偏倚为零, 并且方差最小。这个结论可以利用高斯-马可夫 (Gauss-Markov) 定理得到 (参见文献 [7]), 并且不论测量次数 N 多大, 不论单次测量的概率密度函数的形式, 这个结论都成立。

函数 $\lambda(x; \theta)$ 在 x_i 处的值可以写为

$$\lambda(x_i; \theta) = \sum_{j=1}^m a_j(x_i) \theta_j = \sum_{j=1}^m A_{ij} \theta_j \quad (7.7)$$

其中 $A_{ij} = a_j(x_i)$ 。于是, (7.5) 式 χ^2 的一般表达式可以写成矩阵形式,

$$\begin{aligned}\chi^2 &= (\mathbf{y} - \boldsymbol{\lambda})^T V^{-1} (\mathbf{y} - \boldsymbol{\lambda}) \\ &= (\mathbf{y} - A\boldsymbol{\theta})^T V^{-1} (\mathbf{y} - A\boldsymbol{\theta}),\end{aligned}\quad (7.8)$$

其中 $\mathbf{y} = (y_1, \dots, y_N)$ 为测量值的矢量, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N)$ 包含期望值 $\lambda_i = \lambda(x_i; \boldsymbol{\theta})$ 。在矩阵方程中, \mathbf{y} 和 $\boldsymbol{\lambda}$ 都应当理解为列矢量, 而上标 T 表示矢量的转置 (即行矢量)。

为了求解 χ^2 的最小值, 我们令其对参数 θ_i 的导数等于零,

$$\nabla \chi^2 = -2(A^T V^{-1} \mathbf{y} - A^T V^{-1} A \boldsymbol{\theta}) = 0. \quad (7.9)$$

假设 $A^T V^{-1} A$ 为非奇异矩阵, 则由上式可以求解出估计量 $\hat{\boldsymbol{\theta}}$,

$$\hat{\boldsymbol{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \mathbf{y} \equiv B \mathbf{y}. \quad (7.10)$$

也就是说, 求解得到的估计量 $\hat{\boldsymbol{\theta}}$ 是原始测量值 \mathbf{y} 的线性函数。利用误差传递可以求出估计量的协方差矩阵 $U_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$, 即

$$U = B V B^T = (A^T V^{-1} A)^{-1}. \quad (7.11)$$

等价地, 协方差矩阵的逆 U^{-1} 可以表示为

$$(U^{-1})_{ij} = \frac{1}{2} \left[\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}. \quad (7.12)$$

需要注意的是, 当 y_i 服从高斯分布时, $\log L = -\chi^2/2$, 此时 (7.12) 式协方差矩阵的逆与 RCF 边界一致, 参见6.6节和7.1节。

对于 $\lambda(x; \boldsymbol{\theta})$ 为参数 $\boldsymbol{\theta}$ 的线性函数的情形, χ^2 为 $\boldsymbol{\theta}$ 的二次型函数:

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{i,j=1}^m \left[\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j). \quad (7.13)$$

将该式与 (7.12) 式的方差表达式结合起来, 可以得到参数空间的等值线, 等值线的切线位于 $\hat{\theta}_i \pm \hat{\sigma}_i$, 对应于离开最小二乘估计值一个标准差:

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + 1 = \chi_{\min}^2 + 1. \quad (7.14)$$

这直接对应于6.11节极大似然问题中常数似然的等值线。如果函数 $\lambda(x; \boldsymbol{\theta})$ 不是参数的线性函数, 则 (7.14) 式定义的等值线一般来说不是椭圆, 无法从切线得到标准差。然而, 这在参数空间定义了一个区域, 可以解释为**置信区域(confidence region)**, 区域的大小反映了拟合参数的统计不确定度。在第九章, 将对置信区域的概念进行更精确的定义。需要注意的是, (7.14) 式定义的区域置信水平依赖于拟合参数的数目: 对一个参数, 置信水平为 6.83%; 对两个参数, 为 39.4%, 对三个参数, 为 19.9%; 等等 (参见9.7节)。

7.3 多项式的最小二乘拟合

作为最小二乘法的例子, 考虑图7.2给出的数据, 图中包含在不同 x 处对 y 的 5 个测量值, 误差为 Δy 。假设每个测量值 y_i 服从中心值为 λ_i (未知) 标准差为 $\sigma_i = \Delta y_i$ (假定已知) 的高斯分布。作为 $\lambda(x; \boldsymbol{\theta})$ 的一个假设, 可能尝试 m 次多项式 (即 $m+1$ 个参数),

$$\lambda(x; \theta_0, \dots, \theta_m) = \sum_{j=0}^m x^j \theta_j. \quad (7.15)$$

这是7.2节介绍的线性最小二乘的特殊形式，系数 $a_j(x)$ 为 x 的幂次方。图7.2给出了利用零次、一次和四次多项式进行最小二乘法拟合的结果。零次多项式就是测量值的平均，每个点的权重为误差平方的倒数。这个假设给出的结果为 $\hat{\theta}_0 = 2.66 \pm 0.13$ ， $\chi^2 = 45.5$ ，自由度的数目为 4(5 个点减去 1 个参数)。这些数据用直线拟合(一次多项式)可以更好地描述，结果为 $\hat{\theta}_0 = 0.93 \pm 0.30$ ， $\hat{\theta}_1 = 0.68 \pm 0.10$ ， $\chi^2 = 3.99$ ，自由度的数目为 3。由于只有 5 个数据点，4 次多项式(5 个参数)完全通过所有数据点，相应的 χ^2 为零。在7.5节将讨论如何利用 χ^2 值评价拟合优度。

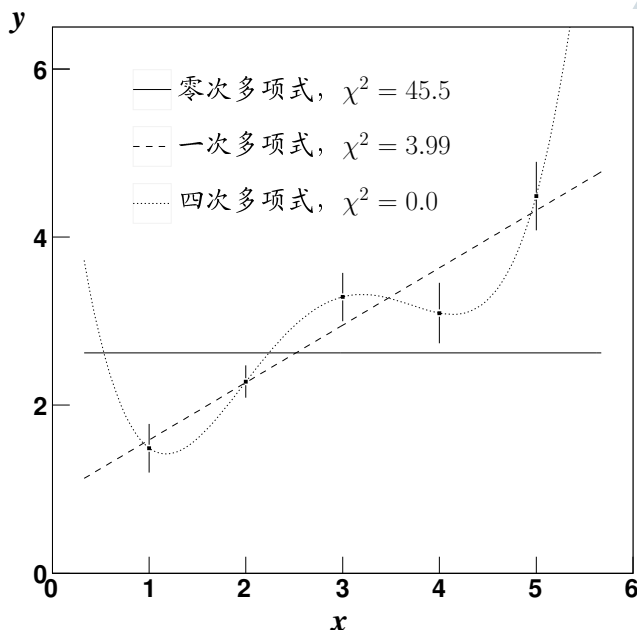


图 7.2: 用 0 次、1 次和 4 次多项式分别对 5 个测量数据进行最小二乘拟合。

与极大似然法一样，最小二乘估计量的统计误差和协方差也可以通过多种方式进行估计。当参数离开使 χ^2 取最小值的位置时， χ^2 的取值会有所变化，这些估计方法都与这个变化有关。图7.3(a)给出了用零次多项式拟合时 χ^2 作为 θ_0 的函数。 χ^2 的曲线为抛物线，因为假设的拟合函数为参数 θ_0 的线性函数(见(7.13)式)。最小二乘估计量 $\hat{\theta}_0$ 的方差可以用7.2节讨论的任何一种方法进行估计：让参数的变化使最小 χ^2 值增加 1；求抛物线在最低点的曲率(二阶导数)；或者，直接利用(7.11)式求解。

图7.3(b)给出了一次多项式拟合时 $\chi^2 = \chi^2_{\min} + 1$ 的等值线。从椭圆的倾斜度和宽度，可以看到估计量 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 存在负相关。(7.11)式给出

$$\begin{aligned}\hat{\sigma}_{\hat{\theta}_0} &= \sqrt{\hat{U}_{00}} = 0.30 \\ \hat{\sigma}_{\hat{\theta}_1} &= \sqrt{\hat{U}_{11}} = 0.10 \\ \text{cov}[\hat{\theta}_0, \hat{\theta}_1] &= \hat{U}_{01} = -0.028,\end{aligned}$$

对应的相关系数为 $r = -0.90$ 。与极大似然法一样，这里的标准差对应于椭圆的切线，相关系数对应于椭圆的宽度和倾斜角(参见6.31式和6.32式)。

由于两个参数 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 之间存在强负相关，在报告拟合结果时，必须同时报告参数之间的协方差或相关系数。在1.7节曾经介绍过，总可以通过正交变换的方法，用原来的 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 定义两个新的变量 $\hat{\eta}_0$ 和 $\hat{\eta}_1$ ，使新变量满足 $\text{cov}[\hat{\eta}_0, \hat{\eta}_1] = 0$ 。然而，尽管处理非相关的量更加简单，但是，变换后的量可能不如原来的量有直接的物理解释相对应。

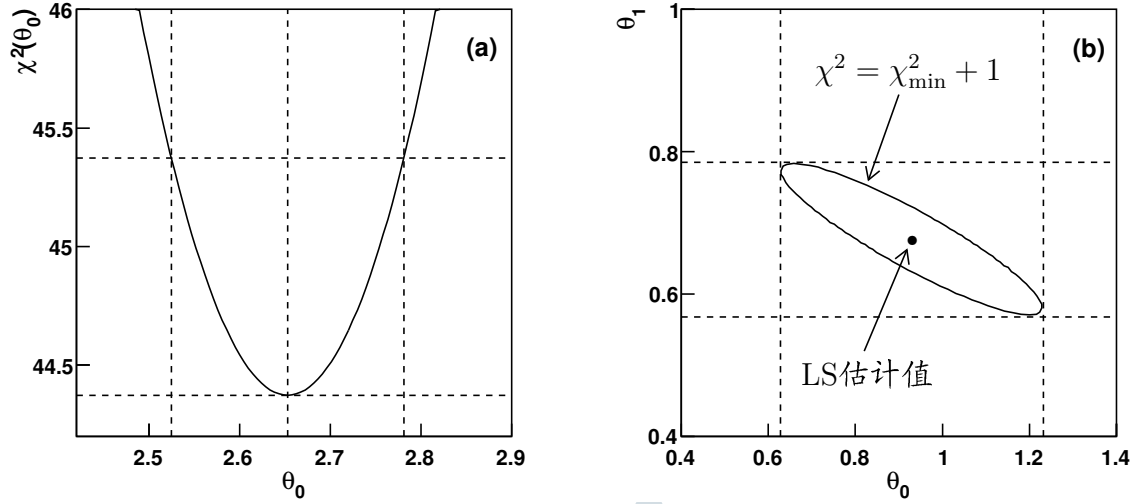


图 7.3: (a) 图 7.2 零次多项式拟合中, χ^2 作为 θ_0 的函数。水平线分别表示 χ^2_{\min} 和 $\chi^2_{\min} + 1$ 。对应的 θ_0 的值 (竖线) 为最小二乘估计值 $\hat{\theta}_0$ 和 $\hat{\theta}_0 \pm \hat{\sigma}_{\hat{\theta}_0}$ 。(b) 图 7.2 一次多项式拟合的最小二乘估计值 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 。等值线 $\chi^2(\hat{\theta}_0, \hat{\theta}_1) = \chi^2_{\min} + 1$ 的切线对应于 $\hat{\theta}_0 \pm \hat{\sigma}_{\hat{\theta}_0}$ 和 $\hat{\theta}_1 \pm \hat{\sigma}_{\hat{\theta}_1}$ 。

7.4 分区间数据的最小二乘拟合

在前一个例子中, 联系 ‘真值’ λ 和变量 x 的函数不一定是 x 的概率密度函数, 而可以是任意函数。然而, 这个函数可以取为概率密度函数, 或者正比于概率密度函数。例如, 假设我们对随机变量 x 进行次测量, 并做成分 N 个区间的直方图。令 y_i 为区间 i 的频数, $f(x; \theta)$ 为假设的概率密度函数, 用以估计参数 $\theta = (\theta_1, \dots, \theta_m)$ 。于是, 区间 i 的预期事例数 $\lambda_i = E[y_i]$ 为

$$\lambda_i(\theta) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \theta) dx = np_i(\theta), \quad (7.16)$$

其中 x_i^{\min} 和 x_i^{\max} 分别为区间的下限和上限, $p_i(\theta)$ 为事例处于区间 i 的概率。参数 θ 可以通过最小化下面这个量求得

$$\chi^2(\theta) = \sum_{i=1}^N \frac{(y_i - \lambda_i(\theta))^2}{\sigma_i^2}, \quad (7.17)$$

其中 σ_i^2 为区间 i 的事例数的方差。这里需要注意的是, 函数 $f(x; \theta)$ 归一化为 1, 因为它是概率密度函数, 而用来拟合直方图的函数为 $\lambda_i(\theta)$ 。

如果每个区间的平均事例数相比总事例数都很小, 则每个区间的事例数近似服从泊松分布, 因此其方差等于平均值 (见 (2.11) 式)。于是, (7.17) 式变成

$$\chi^2(\theta) = \sum_{i=1}^N \frac{(y_i - \lambda_i(\theta))^2}{\lambda_i(\theta)} = \sum_{i=1}^N \frac{(y_i - np_i(\theta))^2}{np_i(\theta)}. \quad (7.18)$$

另一种方法是, 用区间 i 的实际观测事例数近似相应的方差, 而不用预期值 $\lambda_i(\theta)$ 。这就是所谓的**修正的最小二乘法(modified least-squares method, MLS)**。修正的最小二乘法要求对下式求最小值

$$\chi^2(\theta) = \sum_{i=1}^N \frac{(y_i - \lambda_i(\theta))^2}{y_i} = \sum_{i=1}^N \frac{(y_i - np_i(\theta))^2}{y_i}. \quad (7.19)$$

这个方法容易进行计算, 但存在一个缺点是, 如果某些区间事例数很少或者没有事例, 则误差估计可能很差 (甚至导致 χ^2 无定义)。

在使用最小二乘法拟合直方图时，需要注意以下潜在的问题。通常我们不使用实际观测的总事例数 n 从 (7.16) 式求 λ_i ，而是引入额外的可调参数 ν 作为归一化因子。于是，区间 i 的预期事例数 $\lambda_i(\boldsymbol{\theta}, \nu) = E[y_i]$ 变为

$$\lambda_i(\boldsymbol{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \boldsymbol{\theta}) dx = \nu p_i(\boldsymbol{\theta}). \quad (7.20)$$

这一步看起来避免了计算总事例数 n 。然而很容易证明，引入可调归一化参数会导致总事例数的错误估计。考虑方差取为预期事例数 ($\sigma_i^2 = \lambda_i$) 的最小二乘法。 λ_i 取为 (7.20) 式的形式，并令 χ^2 对 ν 的导数等于零，可以得到估计量

$$\hat{\nu}_{\text{LS}} = n + \frac{\chi^2}{2}. \quad (7.21)$$

对于修正的最小二乘法 ($\sigma_i^2 = y_i$)，可以得到

$$\hat{\nu}_{\text{MLS}} = n - \chi^2. \quad (7.22)$$

由于每个区间对 χ^2 的贡献一般来说大约为 1，所以，对于最小二乘法，总事例数的相对误差大约偏高 $N/2n$ ；对于修正的最小二乘法，总事例数的相对误差大约偏低 N/n 。如果我们根据经验取每个区间至少包含 5 个事例，在归一化中将带来 10 – 20% 的额外误差。

尽管引入的偏倚可能小于对应的统计误差，但是，对很多这种拟合进行平均得到的结果，引入的错误很容易超过平均值的统计误差。因此，我们应当直接从总事例数确定归一化因子。如果用总事例数作归一化因子无法实现 (例如因为计算机程序的限制)，我们至少应当清楚存在潜在的问题，区间大小的选择应当保证引入的偏倚在可接受范围之内。

可以对分区间数据的最小二乘法与 6.10 节的极大似然法相比较。在极大似然法中，各区间事例数 y_i 的联合概率密度函数取为多项分布，或者也可以认为每个 y_i 是泊松随机变量。在后一种情形中，期待的总事例数 ν 被当作可调参数，可以自动得到正确的估计量 $\hat{\nu} = n$ (见 6.9 节和 6.10 节)。而且，文献 [23] (见 8.4.5 节及其中给出的参考文献) 指出，极大似然估计量的方差比最小二乘法或修正的最小二乘估计量更快地收敛于最小方差边界。这个结论也建议我们在直方图拟合中采用极大似然法。

作为例子，考虑图 7.4 给出的直方图，直方图分 20 个区间，包含 400 个事例。数据是用蒙特卡罗方法在区间 $[0, 2]$ 内产生，所用的概率密度函数为高斯分布和指数分布的线性组合，即

$$f(x; \theta, \sigma, \xi) = \theta \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\int_0^2 \exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right) dx'} + (1-\theta) \frac{e^{-x/\xi}}{\xi(1-e^{-2/\xi})}, \quad (7.23)$$

其中 $\theta = 0.5$ ， $\mu = 1$ ， $\xi = 4$ 及 $\sigma = 0.35$ 。假设 μ ， σ 和 ξ 均已知，而我们希望确定总事例数对高斯分量和指数分量的贡献，即 $\nu_{\text{Gauss}} = \nu\theta$ 和 $\nu_{\text{exp}} = \nu(1-\theta)$ 。

图 7.4(a) 给出了将 ν 当作自由参数的拟合结果。对于 $\sigma_i^2 = \lambda_i$ (即最小二乘法，LS)，正如 (7.21) 式所期待，拟合结果为 $\chi^2 = 17.1$ ， $\hat{\nu}_{\text{LS}} = 408.5$ ，并且 $\hat{\theta} = 0.498 \pm 0.056$ 。对于 $\sigma_i^2 = y_i$ (即修正的最小二乘法，MLS)，拟合结果为 $\chi^2 = 17.8$ ， $\hat{\nu}_{\text{MLS}} = 382.2 \pm 19.5$ ，与 (7.22) 式的结果一致，相应地， $\hat{\theta} = 0.551 \pm 0.062$ 。从图中很容易看出两个拟合曲线下的面积不同，从而导致对高斯和指数两种成分的事例数的不同估计。

图 7.4(b) 给出的拟合结果中，对 ν 的处理更为正确。对于最小二乘法 (LS) 的曲线，方差取为 $\sigma_i^2 = \lambda_i$ ，总事例数固定为真实事例数 $\nu = n = 400$ 。相应的拟合结果为 $\chi^2 = 17.3$ ， $\hat{\theta} = 0.496 \pm 0.055$ 。图 7.4(b) 也给出了极大似然拟合的结果， $\hat{\theta} = 0.514 \pm 0.057$ ，其中似然函数取为泊松分布的乘积。如 6.10 节所示，极大似然估计量自动给出正确的事例数。极大似然拟合中的拟合优度可以用 (6.52) 式的检验统计量进行评估。对于本例，这个结果为 $\chi_p^2 = 17.6$ ，与最小二乘拟合的 χ^2 接近。尽管所有这些拟合方法中， $\hat{\theta}$ 的标准差都很接近，但推荐使用图 7.4(b) 中的拟合，因为这两种拟合中总事例数是正确的。

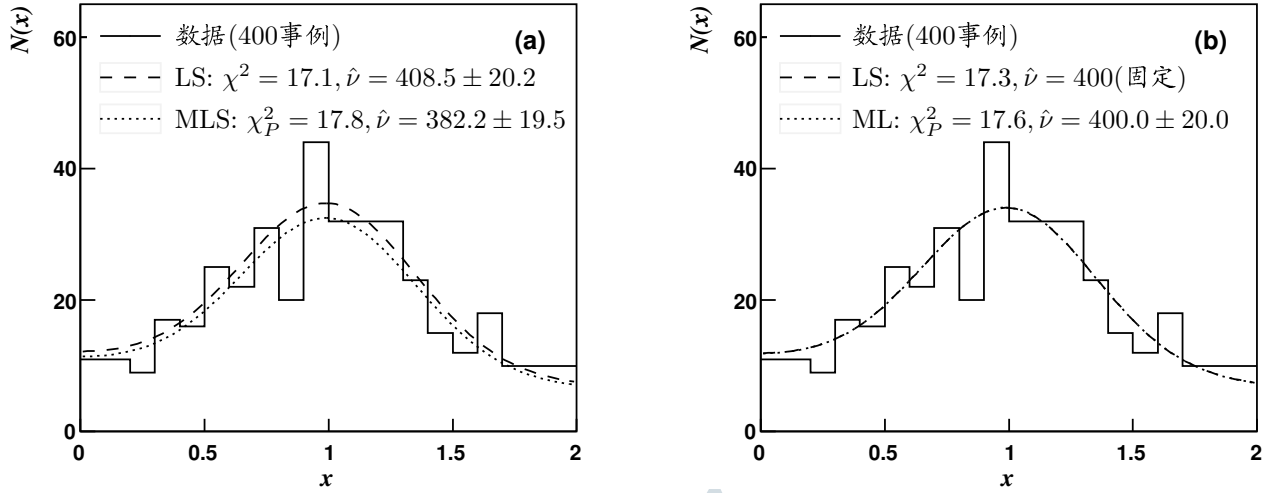


图 7.4: (a) 按照 (7.23) 式产生的蒙特卡罗数据的拟合，其中总事例数 ν 被当作可调参数。(b) 总事例数固定为真实事例数的最小二乘拟合结果，以及极大似然拟合结果。

7.5 拟合优度最小二乘检验

如 7.1 节所示，如果测量值 y_i 服从高斯分布，则最小二乘估计量与极大似然估计量一致。并且， χ^2 值可以用来检验假设为真的条件下得到观测数据的可能性有多大。

$(y_i - \lambda(x_i; \theta))/\sigma_i$ 这个量可以衡量第 i 个测量值 y_i 与函数 $\lambda(x_i; \theta)$ 之间的偏差，因此， χ^2 衡量观测数据与假设之间的总体符合程度。可以证明，如果

1. $y_i (i = 1, \dots, N)$ 为方差 σ_i^2 已知相互独立的高斯随机变量 (或者 y_i 服从 N 维高斯分布，协方差矩阵 V 已知)，
2. $\lambda(x; \theta_1, \dots, \theta_m)$ 这个假设是参数 θ_i 的线性函数，
3. 假设的函数形式是正确的，

那么 (7.3) 式 (或者对于相关的 y_i ，则为 (7.5) 式) 定义的最小 χ^2 值服从自由度为 $N - m$ 的卡方分布，见 (2.7) 节中 (2.34) 的定义。在 4.7 节，我们已经遇到了一个特殊的例子，这个例子不需要从数据中估计参数。

由 2.7 节可以看到，服从卡方分布的随机变量 z 的期望值等于自由度的数目。因此，经常用 χ^2 除以自由度的数目 n_d (数据点的数目减去独立参数的数目) 来衡量拟合优度。如果这个比值接近于 1，则一切都符合预期。如果比值远小于 1，则在给定的测量误差条件下，拟合好于预期。从能否提供否定假设的证据的角度来说，这不是一件坏事，但这时候通常需要仔细检查误差 σ_i 是否被高估，或者是否存在相关性。

如果 χ^2/n_d 远大于 1，则有理由怀疑假设的正确性。如 4.5 节所讨论，经常为给定的 χ^2 提供一个显著水平 (P -值)，用以表示当前假设将导致更差 (即更大) χ^2 值的概率 (相比于实际数据)。也就是说，

$$P = \int_{\chi^2}^{\infty} f(z; n_d) dz, \quad (7.24)$$

其中 $f(z; n_d)$ 为自由度数等于 n_d 的卡方分布。这个积分值可以通过数值计算得到 (例如用 [25] 中的子程序 **PROB**，或用 [26] 中的 **TMath** 空间下的函数 **Prob(double chi2, int ndf)**)，或者通过标准图表得到 (例如，参考文献 [3, 51])。 P -值多大时可以决定拒绝某个假设有一定的主观性，但需要注意的是，低估误差 σ_i 会使正确的假设给出很差的 χ^2 值。

对于7.3节考虑的多项式拟合，用直线拟合得到 $\chi^2 = 3.99$ ，自由度的数目为 3(即 5 个数据点减去 2 个自由参数)。利用 (7.24) 式计算显著水平得到 $P = 0.263$ 。也就是说，如果真实函数 $\lambda(x)$ 为直线，并且进行很多次重复实验，每次实验可以得到 $\hat{\theta}_0$ 、 $\hat{\theta}_1$ 和 χ^2 的一组值。我们预期这个 χ^2 值有 26.3% 的可能性比实际得到的 $\chi^2 = 3.99$ 要更差(更大)。通过大量的蒙特卡罗实验可以对此进行验证，其中‘真实’参数 θ_0 和 θ_1 取为实际实验的结果，而每个数据点的‘测量’值用高斯分布产生，高斯分布的宽度 σ 等于相应数据点的误差。图7.5给出了 1000 次模拟实验得到的归一化直方图，同时给出了自由度数目等于 3 的卡方分布(预期分布)以作比较。

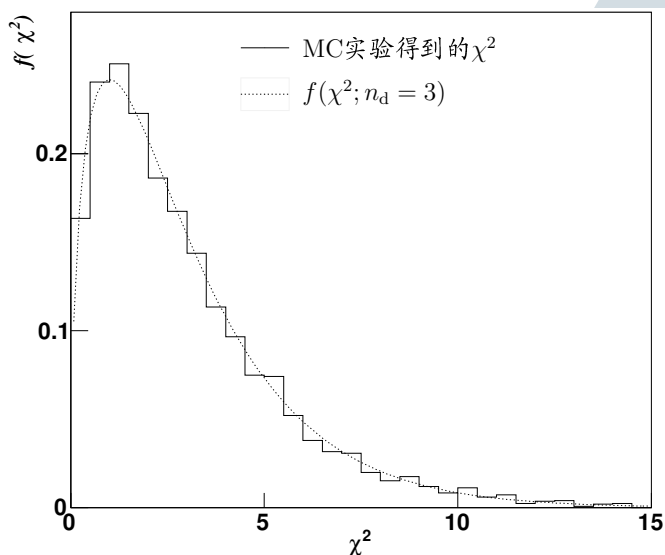


图 7.5: 1000 次蒙特卡罗实验得到的 χ^2 值的直方图(已归一化)，以及自由度数目为 3 的卡方概率密度函数。

用水平直线拟合给出 $\chi^2 = 45.5$ ，自由度数目等于 4。对应的显著水平为 $P = 3.1 \times 10^{-9}$ 。如果水平直线的假设是正确的，则我们预期十亿次实验中有 3 次会得到更大的 χ^2 值，因此这个假设显然应当被否定掉。在计算 P -值时，假定已知标准差 σ_i (或者对于存在相关性的测量来说是协方差矩阵 V)。需要注意的是，低估测量误差 σ_i 或者不正确处理测量的相关性会使正确的假设导致很大的 χ^2 值。

应当谨记小统计误差和好的 χ^2 值之间的区别。统计误差反映了参数离开拟合值时 χ^2 的变化，而与 χ^2 本身的绝对值无关。从 (7.11) 式可以看到，估计量的协方差矩阵 U 只依赖于系数函数 $a_j(x)$ (即复合假设 $\lambda(x; \theta)$) 和原始测量的协方差矩阵 V ，而与测量值 y_i 无关。

估计量 $\hat{\theta}$ 的标准差 $\sigma_{\hat{\theta}}$ ，反映了重复多次实验时估计量的分布有多宽。然而，如果假设的函数形式不正确，估计值 $\hat{\theta}$ 仍然可能明显不同于 θ 的真值，这个真值指的是在真实复合假设中的定义。也就是说，如果假设的形式不正确，则标准差小(即统计误差小)不足以保证参数估计的不确定性也小。

为了证明这一点，考虑7.3节的水平直线拟合，其结果为 $\hat{\theta}_0 = 2.66 \pm 0.13$ ， $\chi^2 = 45.5$ ，对应的自由度数目为 4。图7.6给出了相同 x 值处并且数据误差 Δy 也相同的另外 5 个数据点，但是 y 的值不同。对这组数据进行水平直线拟合得到 $\hat{\theta}_0 = 2.84 \pm 0.13$ ， $\chi^2 = 4.48$ 。 $\hat{\theta}_0$ 的误差与7.3节的结果相同，但现在的 χ^2 值表明水平直线的假设可以很好地描述数据。两种情况下的 χ^2 - θ_0 曲线具有相同的曲率，但是其中一个相对于另一个在垂直方向上有一个常数偏移量。

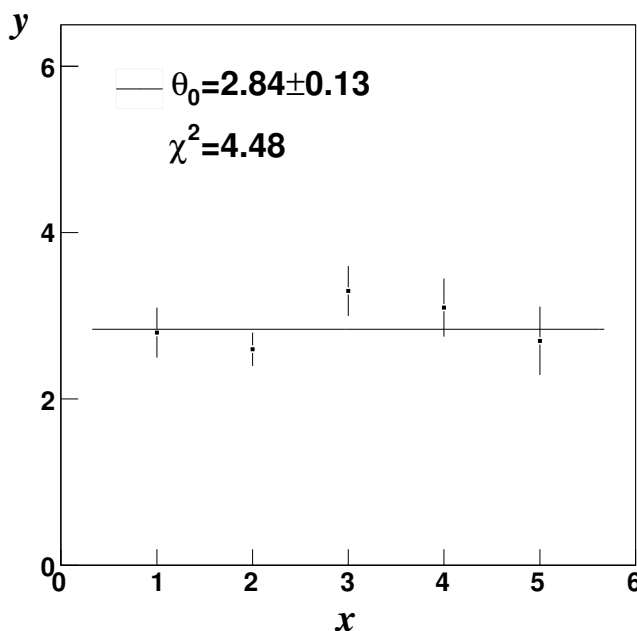


图 7.6: 用零次多项式对数据进行最小二乘拟合, 所用数据的 x 和误差与图 7.2 相同, 但 y 值不同。尽管得到的 χ^2 值远小于前一个例子, 但是 $\hat{\theta}_0$ 的误差保持不变。

7.6 用最小二乘法合并实验测量

最小二乘法的一个特殊应用是合并对同一个量的很多不同测量。假设对某个真值 λ 未知的量进行了 N 次测量 (例如, 进行了 N 次不同的实验), 得到了相互独立的测量值 y_i 以及估计的误差 (标准差) $\sigma_i (i = 1, \dots, N)$ 。由于我们假定对所有测量来说真值是相同的, λ 的值是个常数, 即函数 $\lambda(x)$ 不依赖于 x , 所以, 实际上变量 x 不会在这个问题中出现。(7.3) 式变为

$$\chi^2(\lambda) = \sum_{i=1}^N \frac{(y_i - \lambda)^2}{\sigma_i^2}, \quad (7.25)$$

其中 λ 相当于参数 θ 。令 χ^2 对 λ 的导数等于零, 并求解 λ 可以得到最小二乘估计量 $\hat{\lambda}$,

$$\hat{\lambda} = \frac{\sum_{i=1}^N y_i / \sigma_i^2}{\sum_{j=1}^N 1 / \sigma_j^2}, \quad (7.26)$$

这就是常见的加权平均公式。利用 χ^2 的二次导数, 可以得到 $\hat{\lambda}$ 的方差 (见 (7.12) 式),

$$V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^N 1 / \sigma_i^2}. \quad (7.27)$$

从 (7.27) 式可以看出, 加权平均的方差小于任何单次测量的方差。而且, 如果某个测量值 y_i 的方差远远小于其他测量的方差, 则该测量决定了加权平均的值和方差。

这一过程很容易推广到测量值 y_i 不相互独立的情形。例如, 如果测量值部分使用相同数据, 则测量值就不相互独立。假设 N 次测量的协方差矩阵 V 已知, (7.25) 式可以用 χ^2 更一般的定义 ((7.5) 式) 表示, 即

$$\chi^2(\lambda) = \sum_{i,j=1}^N (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda). \quad (7.28)$$

λ 的最小二乘估计量可以通过令 $\chi^2(\lambda)$ 对 λ 的导数等于零得到。与非相关测量的情形类似，得到的估计量为测量值 y_i 的线性组合，

$$\hat{\lambda} = \sum_{i=1}^N \omega_i y_i, \quad (7.29)$$

其中权重 ω_i 为

$$\omega_i = \frac{\sum_{j=1}^N (V^{-1})_{ij}}{\sum_{k,l=1}^N (V^{-1})_{kl}}. \quad (7.30)$$

如果测量没有相关性，即 $(V^{-1})_{ij} = \delta_{ij}/\sigma_i^2$ ，上式显然退化为 (7.26) 式。

注意到，权重 ω_i 之和等于 1，

$$\sum_{i=1}^N \omega_i = \frac{\sum_{i,j=1}^N (V^{-1})_{ij}}{\sum_{k,l=1}^N (V^{-1})_{kl}} = 1. \quad (7.31)$$

假定单次测量 y_i 都是 λ 的无偏估计，则估计量 $\hat{\lambda}$ 也是无偏的，

$$E[\hat{\lambda}] = \sum_{i=1}^N \omega_i E[y_i] = \lambda \sum_{i=1}^N \omega_i = \lambda. \quad (7.32)$$

只要保证 $\lambda \sum_{i=1}^N \omega_i = 1$ ，这个结论对权重的任意选择都成立。可以证明，最小二乘法得到的特殊权重 (7.30) 使得无偏估计量 $\hat{\lambda}$ 的可能方差最小。一方面，这可以由所有最小二乘估计量都满足的高斯-马可夫定理得到。等价地，可以简单地假定 (7.29) 式的加权平均形式，并要求 $\sum_{i=1}^N \omega_i = 1$ ，通过要求 $\hat{\lambda}$ 的方差最小来确定各个权重。利用 (1.53) 式的误差传递，可以得到方差为

$$V[\hat{\lambda}] = \sum_{i,j=1}^N \omega_i V_{ij} \omega_j, \quad (7.33)$$

或者，写成矩阵形式， $V[\hat{\lambda}] = \mathbf{w}^T V \mathbf{w}$ ，其中 \mathbf{w} 为包含 N 个权重的列矢量， \mathbf{w}^T 为对应的转置矢量（即行矢量）。通过用 $1 - \sum_{i=1}^{N-1} \omega_i$ 替换 ω_N ，并令 (7.33) 式对前 $N-1$ 个 ω_i 的导数等于零，就可以得到 (7.30) 式定义的权重。（另一种办法是，利用拉格朗日乘子处理约束条件 $\sum_{i=1}^N \omega_i = 1$ 。）

作为例子，考虑两个测量 y_1 和 y_2 ，其协方差矩阵为

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad (7.34)$$

其中 $\rho = V_{12}/(\sigma_1\sigma_2)$ 为相关系数。于是，协方差矩阵的逆为

$$V^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}. \quad (7.35)$$

将这个结果代入 (7.29) 式和 (7.30) 式，得到加权平均为

$$\hat{\lambda} = \omega y_1 + (1 - \omega) y_2, \quad (7.36)$$

其中

$$\omega = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \quad (7.37)$$

从 (7.33) 式可以求得 $\hat{\lambda}$ 的方差为

$$V[\hat{\lambda}] = \frac{(1 - \rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sigma^2, \quad (7.38)$$

或者，等价地写为

$$\frac{1}{\sigma^2} = \frac{1}{1-\rho^2} \left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2} \right]. \quad (7.39)$$

相关性的存在会导致一些有趣的结果，例如，参见文献 [52,53] 等。从 (7.39) 式可以看到，由于第二次测量 y_2 导致方差的倒数的变化为

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1-\rho^2} \left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2} \right)^2. \quad (7.40)$$

这个结果显然恒大于或等于零，也就是说，第二次测量总是有助于减小总方差 σ^2 ，或者至少不会使 σ^2 变坏。如果 $\rho = \sigma_1/\sigma_2$ ，则方差的变化为零。一个平庸的例子是 $\rho = 1$ ， $\sigma_1 = \sigma_2$ ，也就是说，同一个测量被考虑了两次。

如果 $\rho > \sigma_1/\sigma_2$ ，权重 ω 变成负值，这表示加权平均不处于 y_1 和 y_2 之间。如果 y_1 和 y_2 之间存在强正相关，两个测量值可能都位于真值 λ 的同一侧，就可能发生 ω 为负的情况。这个结果相当令人惊讶，很值得更仔细地考察下面的例子。

7.6.1 存在相关性的测量数据的平均

考虑用两个不同的尺子测量物体的长度，两个尺子用不同的材料制作，因此热膨胀系数不同。假设对两个尺子都进行了刻度，在温度为 T_0 时给出准确的测量结果，但在其它温度，真实长度 λ (未知) 的修正估计 y 需要通过下式得到

$$y_i = L_i + c_i(T - T_0). \quad (7.41)$$

其中指标 i 表示尺子 1 或 2， L_i 为未修正的测量值， c_i 为膨胀系数，而 T 为温度，需要测量得到。将温度的测量值当作标准差为 σ_T 的高斯随机变量，并假定两次测量所处的温度 T 相同，即两次测量是同时进行的。将未修正的测量值 L_i 当作标准差为 σ_{L_i} 的随机变量，并假定 c_i ， σ_{L_i} 和 σ_T 都已知。

为了得到 y_1 和 y_2 的加权平均，需要它们的协方差矩阵。修正的测量结果的方差 $V[y_i] = \sigma_i^2$ 为

$$\sigma_i^2 = \sigma_{L_i}^2 + c_i^2 \sigma_T^2. \quad (7.42)$$

假定测量是无偏的，即 $E[y_i] = \lambda$ ， $E[T] = T_{\text{true}}$ ，其中 T_{true} 为真实温度(未知)。于是可以得到期望值 $E[L_i] = \lambda - c_i(T_{\text{true}} - T_0)$ ， $E[T^2] = \sigma_T^2 + T_{\text{true}}^2$ ， $E[L_i, L_j] = \delta_{ij} \sigma_{L_i}^2$ 。由这些期望值可以得到协方差 $V_{12} = \text{cov}[y_1, y_2]$ 为

$$\begin{aligned} V_{12} &= E[y_1 y_2] - \lambda^2 \\ &= E[(L_1 + c_1(T - T_0))(L_2 + c_2(T - T_0))] - \lambda^2 \\ &= c_1 c_2 \sigma_T^2, \end{aligned} \quad (7.43)$$

相关系数为

$$\rho = \frac{V_{12}}{\sigma_1 \sigma_2} = \frac{c_1 c_2 \sigma_T^2}{\sqrt{(\sigma_{L_1}^2 + c_1^2 \sigma_T^2)(\sigma_{L_2}^2 + c_2^2 \sigma_T^2)}}. \quad (7.44)$$

因此，(7.26) 的加权平均为

$$\hat{\lambda} = \frac{(\sigma_{L_2}^2 + (c_2^2 - c_1 c_2) \sigma_T^2) y_1 + (\sigma_{L_1}^2 + (c_1^2 - c_1 c_2) \sigma_T^2) y_2}{\sigma_{L_1}^2 + \sigma_{L_2}^2 + (c_1 - c_2)^2 \sigma_T^2}. \quad (7.45)$$

如果温度的误差 σ_T 很小，可以忽略不计，则 ρ 趋向于零， $\hat{\lambda}$ 将位于 y_1 和 y_2 之间。然而，如果标准差 σ_{L_i} 非常小，而温度的误差 σ_T 很大，则 (7.44) 的相关系数趋向于 1，(7.37) 式的权重变成负值。在 $\rho = 1$ 的极端情况下，(7.45) 式变为

$$\hat{\lambda} = \frac{-c_1}{c_1 - c_2} y_1 + \frac{c_2}{c_1 - c_2} y_2, \quad (7.46)$$

其方差为零，参见 (7.38) 式。

图7.7对这个结果的原因进行了说明。两个倾斜的阴影带表示，在给定测量长度 L_i 的条件下，修正长度 y_i 的可能值作为温度的函数。假设已经非常精确地知道 L_i 的值，如果 y_1 和 y_2 仍然相差很大，则唯一的解释是，真实温度一定与测量值 T 不同。因此，加权平均 $\hat{\lambda}$ 被拖向两条阴影带 $y_1(T)$ 和 $y_2(T)$ 的交叉点。

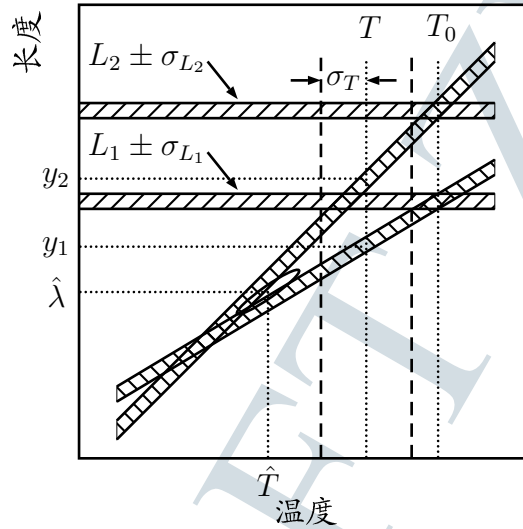


图 7.7: 水平条状区间表示用不同尺子对物体长度的测量值 L_i ，斜对角的条状区间表示修正长度作为温度的函数。温度的测量值 T 给出修正的长度 y_1 和 y_2 。误差椭圆对应于 $\chi^2 = \chi_{\min}^2 + 1$ ，中心位置由长度估计值 $\hat{\lambda}$ 和温度估计值 \hat{T} 给出。

这个问题等价于，给定非相关测量 L_1 、 L_2 和温度 T ，对长度 λ 和真实温度 T_{true} 进行最小二乘拟合。温度的最小二乘估计量为

$$\hat{T} = T - \frac{(c_1 - c_2)(y_1 - y_2)\sigma_T^2}{\sigma_{L_1}^2 + \sigma_{L_2}^2 + (c_1 - c_2)^2\sigma_T^2}. \quad (7.47)$$

认为两次测量所处温度相同，以及假定温度 T 的误差至少是 y_1 和 y_2 存在差异的部分原因，我们以这两点为前提寻求 λ 和 T_{true} 的估计值。

需要问的一个问题是，在实际应用中我们在多大程度上可以信赖这两个前提。两次测量实际上部分基于共同的测量，即对温度的测量（温度本身也是一个随机变量），这是相关性的根源所在。在对高度相关的量作平均时，我们需要倍加小心，因为每次测量之间的协方差矩阵中很小的不确定度都会导致平均值 $\hat{\lambda}$ 出现很大的不确定度，并且其方差估计也不再正确。当不确定度的相关性被高估时，尤其如此。

在上面的例子中，如果对温度的修正 $\Delta T = \hat{T} - T$ 比 σ_T 还大，则意味着我们对测量的前提假设很可能不正确。这将反映在得到的最小 χ^2 值很大，从而对应的 P -值很小。图7.7的例子正是如此，其中给出 $\chi^2 = 10.7$ ，自由度的数目为 1。这时候必须修改模型，很可能要重新评估 $\sigma_{L_1}^2$ ， $\sigma_{L_2}^2$ 和 σ_T^2 ，或者修改测量长度与修正长度之间的函数关系。这个例子的正确结论为，由于 χ^2 值很大，两次测量 y_1 和 y_2 不匹配，因此其平均结果不可靠。

例如，可以假设，大 χ^2 值的原因是 L_i 的标准差被低估。如果这两个标准差取为原来的 4 倍，则将得到 $\chi^2 = 1.78$ ， $\hat{\lambda}$ 将位于 y_1 和 y_2 之间。然而，如果假设将 σ_T 取为原来的 4 倍，会导致 χ^2 下降为 0.97，因此没有任何证据否定系统的模型。 $\hat{\lambda}$ 和 \hat{T} 的误差椭圆（即 $\chi^2 = \chi_{\min}^2 + 1$ 的等值线）将向两个阴影带交叉的点靠近，但椭圆大小基本保持不变。如果我们知道 σ_T 应当取更大的值，并且测量仪器的模型也是可信的，那么这个结果完全可以接受。然而，如果 σ_T 只是因为过于保守而被高估，那么我们将被引向错误的结果。

7.6.2 求解协方差矩阵

为了用最小二乘法对结果进行平均，需要知道协方差矩阵 V 。对于简单的情形，可以直接计算协方差矩阵，例如，如果两次测量仅仅部分而非全部使用相同数据。比如，考虑对某个量 x 的独立观测（‘事例’），其中 x 本身的方差为 σ^2 。假设第一次测量 y_1 是由 n 个独立事例的样本均值构造而成，即

$$y_1 = \frac{1}{n} \sum_{i=1}^n x_i, \quad (7.48)$$

第二次测量 y_2 使用了 m 个 x 值，

$$y_2 = \frac{1}{m} \sum_{i=1}^m x_i. \quad (7.49)$$

进一步假设两组数据中有 c 个事例是共用的，即它们表示相同的观测。利用 $\text{cov}[x_i, x_j] = \delta_{ij}\sigma^2$ ，可以得到协方差矩阵 $\text{cov}[y_1, y_2]$ 为

$$\text{cov}[y_1, y_2] = \frac{c\sigma^2}{nm}, \quad (7.50)$$

或者等价地， $\rho = c/\sqrt{nm}$ 。也就是说，如果 $c = 0$ ，两组数据没有重叠，则相关系数 $\rho = 0$ ；如果 $c \equiv n \equiv m$ ，则两次测量使用的数据几乎完全相同，因此高度相关。

在这个例子中，单次测量的方差为 $\sigma_1 = \sigma/\sqrt{n}$ ， $\sigma_2 = \sigma/\sqrt{m}$ 。可以得到 $\sigma_1/\sigma_2 = \sqrt{m/n}$ ，因此

$$\rho = \frac{c}{\sqrt{nm}} = \frac{\sigma_1}{\sigma_2} \frac{c}{m} \leq \frac{\sigma_1}{\sigma_2}. \quad (7.51)$$

所以，在这种情况下，不会出现负权重，平均值总是处于 y_1 和 y_2 之间。

通常情况下，两个估计量之间的相关性更难确定。即使两个数据样本完全重叠，不同的估计量 y_1, \dots, y_N 可能对数据的依赖方式不同。因此，确定其协方差矩阵非常困难，可能需要用蒙特卡罗方法进行大量模拟实验，对每次实验都确定出 y_1, \dots, y_N ，然后利用 (5.11) 式估计两两实验之间的协方差。文献 [52] 给出了用这种方法处理粒子平均寿命的例子。

有些情况下，即使一次模拟实验也可能需要很长时间的蒙特卡罗计算，实际应用中这个方法可能太慢。然而，我们仍然可以进行模拟实验，只是每次模拟实验包含的事例数可以比实际事例数少一些。取决于估计量的不同，有可能可以从只包含一个事例的实验构造估计量。于是，我们可以以此来确定 N 个估计量 y_1, \dots, y_N 的相关系数矩阵 ρ_{ij} 。对于有效估计量（大样本极限下，极大似然估计量和最小二乘估计量都是有效估计量），协方差反比于样本容量（参见 (6.20) 式），因此， ρ_{ij} 与每个子样本中的事例数无关。于是，可以使用关系式 $V_{ij} = \rho_{ij}\sigma_i\sigma_j$ ，其中 σ_i 和 σ_j 可以从数据进行估计，也可以从蒙特卡罗进行估计。如果有足够的数据，我们可以不必求助于蒙特卡罗计算，而是将真实数据样本分成大量的子样本，对每个子样本确定估计量 y_1, \dots, y_N ，并以此估计相关系数矩阵。

即使不精确已知协方差矩阵，用这个技术对测量进行平均仍然是正确的，而且这可能是最佳的实际解。注意到，可以任意选择 (7.29) 式的权重 ω_i ，只要保证其和为 1，得到的估计量 $\hat{\lambda}$ 就是无偏的。如果协方差矩阵只是近似已知，这一点仍然成立。但是这种情况下，将无法得到最小可能的方差，平均值的方差的估计将不再正确。

DRAFT ZV

第八章 矩方法

尽管极大似然法和最小二乘法得到的估计量一般来说拥有最优或近似最优的性质，有时这两种方法很难具体实现。还有一种可供选择的参数估计方法，称为**矩方法(method of moments, MM)**。尽管矩方法估计量的方差通常不如极大似然估计量那样小，但是从计算的角度来说，矩方法通常更简单。

假设有随机变量 x 的 n 个观测值 x_1, \dots, x_n 以及某个假设，该假设预言概率密度函数的形式为 $f(x; \boldsymbol{\theta})$ ，其中包含 m 个未知参数 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ 。矩方法的思想是，首先构造 m 个线性独立的函数 $a_i(x)$, $i = 1, \dots, m$ 。 $a_i(x)$ 本身也是随机变量，其数学期望值 $e_i = E[a_i(x)]$ 是真实参数的函数，

$$E[a_i(x)] = \int a_i(x) f(x; \boldsymbol{\theta}) dx \equiv e_i(\boldsymbol{\theta}). \quad (8.1)$$

函数 $a_i(x)$ 的选择要求可以计算 (8.1) 式的数学期望值，从而确定函数 $e_i(\boldsymbol{\theta})$ 。

在5.2节已经看到，样本均值为随机变量数学期望值的无偏估计量。所以，可以用 x 的观测值计算函数 $a_i(x)$ ，然后用函数 $a_i(x)$ 的算术平均估计数学期望值 $e_i = E[a_i(x)]$ ，

$$\hat{e}_i = \bar{a}_i = \frac{1}{n} \sum_{j=1}^n a_i(x_j). \quad (8.2)$$

参数 $\boldsymbol{\theta}$ 的矩方法估计量按如下方法定义：令期望值 $e_i(\boldsymbol{\theta})$ 等于对应的估计量 \hat{e}_i ，并对参数进行求解。即，求解下面关于 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 的 m 个方程：

$$\begin{aligned} e_1(\hat{\boldsymbol{\theta}}) &= \frac{1}{n} \sum_{i=1}^n a_1(x_i) \\ &\vdots \\ e_m(\hat{\boldsymbol{\theta}}) &= \frac{1}{n} \sum_{i=1}^n a_m(x_i) \end{aligned} \quad (8.3)$$

函数 $a_i(x)$ 的一种可能选择是变量 x 的整数幂次方： x^1, \dots, x^m ，所以数学期望值 $E[a_i(x)] = E[x^i]$ 为 x 的 i 阶矩 (这就是“矩方法”名称的来源)。然而，也可以选择其它 m 个线性独立的函数，只要能够计算其数学期望值，并得到关于参数的 m 个线性独立的函数即可。

我们可能还希望评估估计量 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 的协方差矩阵。为此，首先利用 (5.11) 式估计协方差 $\text{cov}[a_i(x), a_j(x)]$ ，

$$\widehat{\text{cov}}[a_i(x), a_j(x)] = \frac{1}{n-1} \sum_{k=1}^n (a_i(x_k) - \bar{a}_i)(a_j(x_k) - \bar{a}_j). \quad (8.4)$$

该协方差可以与函数的算数平均的协方差联系起来,

$$\begin{aligned}\text{cov}[\bar{a}_i, \bar{a}_j] &= \text{cov}\left[\frac{1}{n} \sum_{k=1}^n a_i(x_k), \frac{1}{n} \sum_{l=1}^n a_j(x_l)\right] \\ &= \frac{1}{n^2} \sum_{k,l=1}^n \text{cov}[a_i(x_k), a_j(x_l)] \\ &= \frac{1}{n} \text{cov}[a_i, a_j].\end{aligned}\quad (8.5)$$

最后一行用到了对 k 和 l 的求和只有 $k = l$ 的 n 项有贡献, 并且每项贡献为 $\text{cov}[a_i, a_j]$ 。其余 $n^2 - n$ 项中 $k \neq l$, 由于每个 x 都是独立的, 所以协方差 $\text{cov}[a_i(x_k), a_j(x_l)]$ 等于零。于是, 期望值 $\hat{e}_i = \bar{a}_i$ 的估计量的协方差矩阵 $\text{cov}[\hat{e}_i, \hat{e}_j]$ 可以估计为

$$\widehat{\text{cov}}[\hat{e}_i, \hat{e}_j] = \frac{1}{n(n-1)} \sum_{k=1}^n (a_i(x_k) - \bar{a}_i)(a_j(x_k) - \bar{a}_j). \quad (8.6)$$

为了得到参数估计量本身的协方差矩阵 $\text{cov}[\hat{\theta}_i, \hat{\theta}_j]$, 可以对 (8.6) 式使用误差传递公式 (1.54),

$$\text{cov}[\hat{\theta}_i, \hat{\theta}_j] = \sum_{k,l} \frac{\partial \hat{\theta}_i}{\partial \hat{e}_k} \frac{\partial \hat{\theta}_j}{\partial \hat{e}_l} \text{cov}[\hat{e}_k, \hat{e}_l]. \quad (8.7)$$

需要注意的是, 即使利用了所有观测值 x_i (即没有数据的分区间现象), 通常也无法获得概率密度函数形式的所有信息。例如, 取 $a_i(x) = x^i$, $i = 1, \dots, m$, 只用到了 x 的前 m 阶矩的信息, 但是某些参数可能对更高阶矩敏感。由于这个原因, 矩方法的方差通常比极大似然法或最小二乘法得到的方差更大 (参见文献 [23] 的 8.2.2 节, [20] 的第 11 和 12 章)。然而, 由于矩方法的简单性, 如果参数的估计过程需要重复很多次, 使用矩方法将特别有效。

作为例子, 考虑连续变量 x 的概率密度函数

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{d_1 + \alpha d_2 + \beta d_3}, \quad (8.8)$$

其中 $x_{\min} \leq x \leq x_{\max}$, 并且

$$d_k = \frac{1}{k} (x_{\max}^k - x_{\min}^k). \quad (8.9)$$

这个概率密度函数在 6.8 节已经碰到过, 当时利用极大似然法对参数 α 和 β 进行了估计。这里, 为了进行比较, 我们采用矩方法。为此, 需要 x 的两个线性独立函数, 选择的要求是函数期望值的计算简单。一个显然的选择是

$$\begin{aligned}a_1 &= x, \\ a_2 &= x^2.\end{aligned}\quad (8.10)$$

可以计算得到期望值 $e_1 = E[a_1]$ 和 $e_2 = E[a_2]$ 为

$$\begin{aligned}e_1 &= \frac{d_2 + \alpha d_3 + \beta d_4}{d_1 + \alpha d_2 + \beta d_3}, \\ e_2 &= \frac{d_3 + \alpha d_4 + \beta d_5}{d_1 + \alpha d_2 + \beta d_3}.\end{aligned}\quad (8.11)$$

其中 d_n 由 (8.9) 式给出。用 \hat{e}_1 和 \hat{e}_2 代替 e_1 和 e_2 , 并对这两个方程求解 α 和 β , 即得到矩方法的估计量

$$\begin{aligned}\hat{\alpha} &= \frac{(\hat{e}_1 d_3 - d_4)(\hat{e}_2 d_1 - d_3) - (\hat{e}_1 d_1 - d_2)(\hat{e}_2 d_3 - d_5)}{(\hat{e}_1 d_2 - d_3)(\hat{e}_2 d_3 - d_5) - (\hat{e}_1 d_3 - d_4)(\hat{e}_2 d_2 - d_4)}, \\ \hat{\beta} &= \frac{(\hat{e}_1 d_1 - d_2)(\hat{e}_2 d_2 - d_4) - (\hat{e}_1 d_2 - d_3)(\hat{e}_2 d_1 - d_3)}{(\hat{e}_1 d_2 - d_3)(\hat{e}_2 d_3 - d_5) - (\hat{e}_1 d_3 - d_4)(\hat{e}_2 d_2 - d_4)}.\end{aligned}\quad (8.12)$$

从6.8节的例子，我们得到包含 2000 个 x 值的数据样本，数据产生的条件为 $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$ 。利用同样的数据，用矩方法得到

$$\hat{\alpha} = 0.493 \pm 0.051,$$

$$\hat{\beta} = 0.410 \pm 0.106.$$

其中的统计误差是利用 \hat{e}_1 和 \hat{e}_2 的协方差，通过误差传递方法得到的，而 \hat{e}_1 和 \hat{e}_2 的协方差根据 (8.6) 式估计得到。类似地，可以得到相关系数 $r = 0.42$ 。

这些结果与极大似然法得到的结果类似，估计的标准差实际上略小一些。然而，标准差略小的原因只是估计方差时统计涨落的结果。矩方法估计量的方差通常大于等于极大似然估计量的方差。在这个特定的例子中，二者的方差几乎相同。然而，矩方法的优点是，不要求似然函数的最大值即可得到估计量。在这个例子中，求似然函数的最大值需要更复杂的数值计算。

DRAFT ZV

第九章 统计不确定度、置信区间和极限

从第五章到第八章，我们讨论了几种估计概率密度函数性质(矩以及其他参数)的方法，并介绍了求估计量方差的技术。到目前为止，对于‘误差分析’(亦称不确定度分析)来说，我们仅局限于报道估计量的方差(及协方差)，或者等价地，报道标准差和相关系数。在某些情况下，仅报道这些是不充分的，必须寻找其它方法交流统计不确定度。

9.1节将简单介绍，以标准差作为统计不确定度的估计有何意义。9.2节将介绍**置信区间(confidence interval)**，置信区间让我们可以定量表述，进行大量重复实验时参数的真值位于这个区间内的百分比是多少。从9.3节到9.6节处理了很多重要情形的置信区间，并在9.7节推广到多维情形。在9.8节和9.9节，同时用贝叶斯置信区间和经典置信区间对物理排除区域的参数极限进行了估计。

9.1 标准差作为统计不确定度

假设实验的结果是对某个参数的估计。估计量的方差(或者方差的平方根，即标准差)反映了，如果进行多次重复实验并且每次实验包含相同数目的观测，估计值分布的宽度有多大。同样，标准差 σ 经常用来报道测量的统计不确定度，也称为**统计误差(standard error)**。

例如，假定对随机变量 x 进行了 n 次观测，并且对概率密度函数的假设为 $f(x; \theta)$ ，其中包含一个未知参数 θ 。从数据样本 x_1, \dots, x_n ，可以构造函数 $\hat{\theta}(x_1, \dots, x_n)$ (例如用极大似然法)作为参数 θ 的估计量。利用第五-八章讨论的方法(例如解析法，RCF边界，蒙特卡罗方法，图解法)，可以估计出 $\hat{\theta}$ 的标准差。令 $\hat{\theta}_{\text{obs}}$ 为实际观测的估计量的值， $\hat{\sigma}_{\hat{\theta}}$ 为估计量的标准差的估计值。 θ 的测量结果报道为 $\hat{\theta}_{\text{obs}} \pm \hat{\sigma}_{\hat{\theta}}$ ，其意思是，如果对 θ 进行多次估计，每次估计都基于 x 的 n 次测量，那么这些估计值将按照概率密度函数 $g(\hat{\theta})$ 分布， $g(\hat{\theta})$ 的中心值为真值 θ ，宽度为真实的标准差 $\sigma_{\hat{\theta}}$ ，它们的估计结果分别为 $\hat{\theta}_{\text{obs}}$ 和 $\hat{\sigma}_{\hat{\theta}}$ 。

对于大多数实际的估计量，在大样本极限下，样本概率密度函数 $g(\hat{\theta})$ 近似为高斯分布。如果超过一个参数需要估计，则概率密度函数变为协方差矩阵为 V 的多维高斯分布。因此，通过估计标准差，或者对多参数情形估计协方差矩阵，我们等效地总结了估计值分布的所有可获得的信息。利用1.6节介绍的误差传递技术，协方差矩阵等价地给出(至少近似给出)估计量函数的信息。

尽管原则上不论估计量的概率密度函数 $g(\hat{\theta})$ 的形式如何，都可以使用统计误差棒的‘标准差’定义，然而，如果 $g(\hat{\theta})$ 不是高斯分布，这实际上不是常规的定义。这种情况下，通常使用下一节将介绍的置信区间，并且一般来说会导致不对称的误差棒。9.3节将证明，如果 $g(\hat{\theta})$ 为高斯分布，则所谓的68.3%置信区间与 $\hat{\theta}_{\text{obs}} \pm \hat{\sigma}_{\hat{\theta}}$ 覆盖的区间完全相同。

9.2 经典置信区间(严格方法)

报道测量的统计不确定度的另一种方法(通常是等价的方法)是尼曼(Neyman) [54] 首先提出的置信区间。如上一节，假定对随机变量 x 进行了 n 次观测，以此计算参数 θ 的估计量 $\hat{\theta}(x_1, \dots, x_n)$ ，并且得到估

计值为 $\hat{\theta}_{\text{obs}}$ 。进一步假设, 利用某种方法, 例如解析法或蒙特卡罗方法, 得到估计量 $\hat{\theta}$ 的概率密度函数为 $g(\hat{\theta}; \theta)$, 以真值 θ 为参数。也就是说, 真值 θ 未知, 但只要给定 θ 值, 就知道 $\hat{\theta}$ 的概率密度函数。

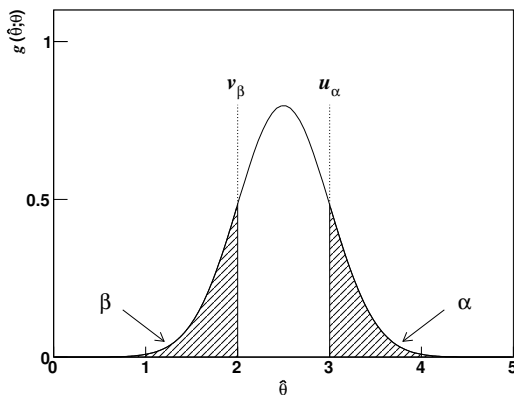


图 9.1: 给定真实参数 θ 为某个值, 估计量 $\hat{\theta}$ 的概率密度函数 $g(\hat{\theta}; \theta)$ 。两个阴影区域分别表示 $\hat{\theta} \leq v_\beta$ 的值 (概率为 β) 和 $\hat{\theta} \geq u_\alpha$ 的值 (概率为 α)

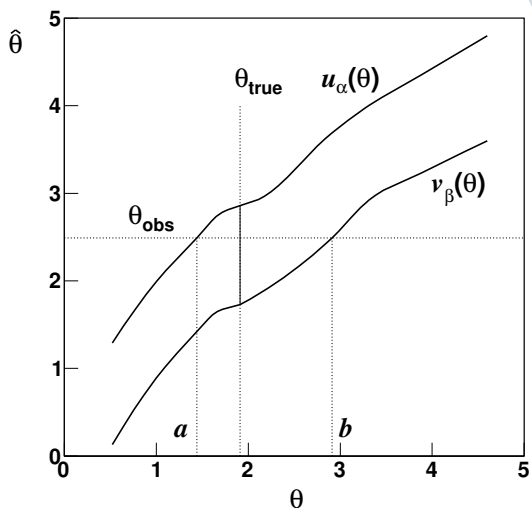


图 9.2: 给定参数 θ 的估计量 $\hat{\theta}$ 的某个观测值 $\hat{\theta}_{\text{obs}}$, 构造置信区间 $[a, b]$ 。

图9.1给出了参数 θ 取某特定值时估计量 $\hat{\theta}$ 的概率密度。利用 $g(\hat{\theta}; \theta)$, 可以确定 u_α 的值, 使得观测到 $\hat{\theta} \geq u_\alpha$ 的概率为固定值 α 。类似地, 可以确定出 v_β , 使得观测到 $\hat{\theta} \leq v_\beta$ 的概率为 β 。 u_α 和 v_β 的值依赖于 θ 的真值, 可以由下面两个式子确定, 即

$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}; \theta) d\hat{\theta} = 1 - G(u_\alpha(\theta); \theta), \quad (9.1)$$

和

$$\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}; \theta) d\hat{\theta} = G(v_\beta(\theta); \theta), \quad (9.2)$$

其中 G 为概率密度函数 $g(\hat{\theta}; \theta)$ 对应的累积分布。图9.2给出了 $u_\alpha(\theta)$ 和 $v_\beta(\theta)$ 作为真值 θ 的函数的示意图。两条曲线之间的区域称为**置信带(confidence belt)**。不管 θ 取什么值, 估计量位于置信带内的概率为

$$P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta. \quad (9.3)$$

只要 $u_\alpha(\theta)$ 和 $v_\beta(\theta)$ 为 θ 的单增函数 (如果 $\hat{\theta}$ 为 θ 的好估计量, 通常满足这一条件), 就可以给出反函数

$$\begin{aligned} a(\hat{\theta}) &\equiv u_\alpha^{-1}(\hat{\theta}), \\ b(\hat{\theta}) &\equiv v_\alpha^{-1}(\hat{\theta}). \end{aligned} \quad (9.4)$$

于是, 不等式

$$\begin{aligned} \hat{\theta} &\geq u_\alpha(\theta), \\ \hat{\theta} &\leq v_\alpha(\theta), \end{aligned} \quad (9.5)$$

分别对应于

$$\begin{aligned} a(\hat{\theta}) &\geq \theta, \\ b(\hat{\theta}) &\leq \theta. \end{aligned} \quad (9.6)$$

因此, (9.1) 式和 (9.2) 式变成

$$\begin{aligned} P(a(\hat{\theta}) \geq \theta) &= \alpha, \\ P(b(\hat{\theta}) \leq \theta) &= \beta, \end{aligned} \quad (9.7)$$

或者, 写到一起,

$$P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta. \quad (9.8)$$

如果函数 $a(\hat{\theta})$ 和 $b(\hat{\theta})$ 用实验中实际得到的估计值 $\hat{\theta}_{\text{obs}}$ 计算, 则会得到两个值 a 和 b , 如图9.2所示。区间 $[a, b]$ 称作**置信水平(confidence level)**或**范围概率(coverage probability)**为 $1 - \alpha - \beta$ 时的**置信区间(confidence interval)**。这样构造置信区间的原因是, 不管未知的真值 θ 取何值, (9.7) 式和 (9.8) 式的范围概率始终成立。需要强调的是, a 和 b 都是随机变量, 因为它们依赖于估计量 $\hat{\theta}$, 而 $\hat{\theta}$ 本身是数据的函数。如果多次重复实验, 区间 $[a, b]$ 包含参数 θ 真值的实验占总实验次数的比率将是 $1 - \alpha - \beta$ 。

从图9.2中, 通过考虑假设的真值 θ_{true} , 我们可以理解区间 $[a, b]$ 和范围概率 $1 - \alpha - \beta$ 的关系。如果 θ_{true} 为参数 θ 的真值, 则 $\hat{\theta}_{\text{obs}}$ 与 $u_\alpha(\theta_{\text{true}})$ 和 $v_\beta(\theta_{\text{true}})$ 之间的垂直线段相交的概率为 $1 - \alpha - \beta$ 。从图中可以看出, 如果 $\hat{\theta}_{\text{obs}}$ 与这个线段相交, 则区间 $[a, b]$ 包含真值 θ_{true} , 反之, 则不包含真值。

在某些情况下, 我们可以只关心**单侧置信区间(one-sided confidence interval)**或**极限(limit)**。也就是说, a 表示参数 θ 的下限, 满足 $a \leq \theta$ 的概率为 $1 - \alpha$ 。类似地, b 表示参数 θ 的上限, 满足 $P(\theta \leq b) = 1 - \beta$ 。

双侧区间 (即同时指定 a 和 b) 无法由置信水平 $1 - \alpha - \beta$ 唯一确定。例如, 我们经常选择 $\alpha = \beta = \gamma/2$, 给出概率为 $1 - \gamma$ 的所谓**中心置信区间(central confidence interval)**。需要注意的是, 中心置信区间不一定表示 a 和 b 离估计值 $\hat{\theta}$ 的距离相等, 仅仅表示概率 α 和 β 相等。

根据置信区间的构造, 假设真实参数取值等于 a , 那么对 $\hat{\theta}$ 的重复估计中将有 α 的比率超过实际得到的估计值 $\hat{\theta}_{\text{obs}}$, 如图9.3所示。类似地, 如果假设参数 θ 的真值为 b , 则有 β 的比率小于实际观测值 $\hat{\theta}_{\text{obs}}$ 。也就是说, 取 $\hat{\theta}_{\text{obs}} = u_\alpha(a) = v_\beta(b)$, (9.1) 式和 (9.2) 式变为

$$\begin{aligned} \alpha &= \int_{\hat{\theta}_{\text{obs}}}^{\infty} g(\hat{\theta}; a) d\hat{\theta} = 1 - G(\hat{\theta}_{\text{obs}}; a), \\ \beta &= \int_{-\infty}^{\hat{\theta}_{\text{obs}}} g(\hat{\theta}; b) d\hat{\theta} = G(\hat{\theta}_{\text{obs}}; b). \end{aligned} \quad (9.9)$$

于是, 前面介绍的确定置信区间的过程等价于对 (9.9) 式求解 a 和 b , 该方程一般需要数值求解。

图9.3还阐明了置信区间与拟合优度检验之间的联系, 参见4.5节。例如, 可以用 $\hat{\theta}$ 作为检验统计量来检验 $\theta = a$ 的假设。如果将 $\hat{\theta} \geq \hat{\theta}_{\text{obs}}$ 定义为与假设的符合程度不好于实际结果的区域 (单侧检验), 那么该

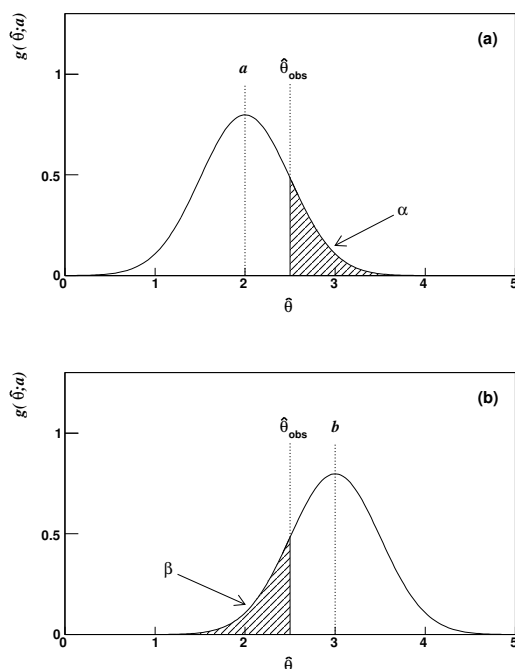


图 9.3: (a) 概率密度函数 $g(\hat{\theta}; a)$, 其中 a 为置信区间的下限。如果参数 θ 的真值为 a , 将有 α 的概率得到 $\hat{\theta}$ 的估计值大于实际观测值 $\hat{\theta}_{\text{obs}}$ 。(b) 概率密度函数 $g(\hat{\theta}; b)$, 其中 b 为置信区间的上限。如果参数 θ 的真值为 b , 将有 β 的概率得到 $\hat{\theta}$ 的估计值大于实际观测值 $\hat{\theta}_{\text{obs}}$ 。

检验的 P -值为 α 。不同的是, 对于置信区间, 概率 α 是预先指定的, 而 a 是依赖于数据的随机量; 对于拟合优度检验, 要预先给定假设 (这个例子中假设为 $\theta = a$), 而 P -值是个随机变量。

需要注意的是, 有时候我们称 P -值 (这里等于 α) 为检验的 ‘置信水平’, 然而单侧置信区间 $\theta \geq a$ 的置信水平为 $1 - \alpha$ 。也就是说, 对于统计检验来说, α 小表示假设 $\theta = a$ 的置信水平低。对于置信区间, 小 α 表示 $\theta \geq a$ 区间包含真实参数具有高置信水平。为了避免混淆, 我们对拟合优度检验使用术语 P -值或 (观测的) 显著水平, 而将置信水平留给置信区间的范围概率。

在报道测量结果时, 置信区间 $[a, b]$ 经常表示成 $\hat{\theta}^{+d}_{-c}$, 其中 $\hat{\theta}$ 为估计值, 而 $c = \hat{\theta} - a$ 和 $d = b - \hat{\theta}$ 通常显示为**误差棒(error bars)**。很多情况下, 概率密度函数 $g(\hat{\theta}; \theta)$ 近似为高斯分布, 所以, 围绕测量值正负一个标准差的区间对应于 $1 - \gamma = 0.683$ 的中心置信区间 (见 9.3 节)。68.3% 的中心置信区间通常为误差棒定义的惯例, 即使估计量的概率密度函数不服从高斯分布。

例如, 如果实验结果报道为 $\hat{\theta}^{+d}_{-c} = 5.79^{+0.32}_{-0.25}$, 这个结果表示, 如果根据前面描述的方法, 对每次实验包含相同测量次数的大量重复实验构造区间 $[\hat{\theta} - c, \hat{\theta} + d]$, 那么将有 $1 - \alpha - \beta$ 的比率构造的区间包含真值。这并不表示真值 θ 位于固定区间 $[5.54, 6.11]$ 内的概率 (相对频率意义的概率) 为 $1 - \alpha - \beta$ 。在相对频率解释中, 真实参数 θ 不是随机变量, 假定不随实验的不同而涨落。这种意义下, θ 位于 $[5.54, 6.11]$ 的概率要么是 0 要么是 1, 但我们不知道到底是 0 还是 1。然而, 构造的区间本身是涨落的, 因为它由数据构造而成。

构造置信区间的困难是, 估计量 $g(\hat{\theta}; \theta)$ 的概率密度函数或累积分布 $G(\hat{\theta}; \theta)$ 必须是已知的。10.4 给出一个例子, 即如何获得指数分布均值 ξ 的估计量的概率密度函数, 并以此确定 ξ 的置信区间。在很多实际应用中, 估计量服从高斯分布 (至少近似服从)。这种情况下, 置信区间很容易确定, 下一节将详细介绍如何处理这种问题。然而, 即使对于非高斯估计量, 也可以利用极大似然函数进行近似处理, 9.6 节将对此进行介绍。

9.3 高斯分布估计量的置信区间

置信区间的简单而重要的应用是当 $\hat{\theta}$ 服从均值为 θ 标准差为 $\sigma_{\hat{\theta}}$ 的高斯分布。也就是说， $\hat{\theta}$ 的累积分布为

$$G(\hat{\theta}; \theta, \sigma_{\hat{\theta}}) = \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(-\frac{(\hat{\theta}' - \theta)^2}{2\sigma_{\hat{\theta}}^2}\right) d\hat{\theta}'. \quad (9.10)$$

这种情况经常出现，因为根据中心极限定理，任何估计量，如果它是随机变量之和的线性函数，那么这个估计量在大样本极限下服从高斯分布。在这种情况下我们将看到，上一节相对复杂的过程变成确定置信区间的简单诀窍。

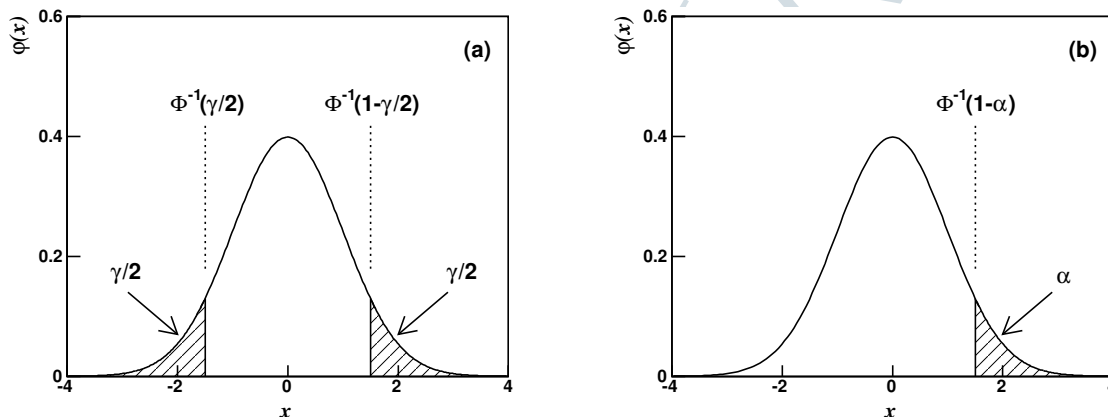


图 9.4: 用标准高斯概率密度函数 $\varphi(x)$ 展示分位数 Φ^{-1} 与置信水平的关系，其中 (a) 中为中心置信区间，(b) 中为单侧置信区间。

假设标准差 $\sigma_{\hat{\theta}}$ 已知，实验中得到参数估计值为 $\hat{\theta}_{\text{obs}}$ 。根据 (9.9) 式，置信区间 $[a, b]$ 可以通过对下面的方程求解 a, b 来确定，即

$$\begin{aligned} \alpha &= 1 - G(\hat{\theta}_{\text{obs}}; a, \sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{\text{obs}} - a}{\sigma_{\hat{\theta}}}\right), \\ \beta &= G(\hat{\theta}_{\text{obs}}; b, \sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{\text{obs}} - b}{\sigma_{\hat{\theta}}}\right), \end{aligned} \quad (9.11)$$

其中 G 已经用标准高斯分布函数的累积分布 Φ 表示出来，见 (2.26) 式和 (2.27) 式。于是可以得到

$$\begin{aligned} a &= \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}} \Phi^{-1}(1 - \alpha), \\ b &= \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}} \Phi^{-1}(1 - \beta). \end{aligned} \quad (9.12)$$

其中 Φ^{-1} 为 Φ 的反函数，即标准高斯分布的分位数。并且，为了使两个方程形式上对称，已经用到了 $\Phi^{-1}(\beta) = -\Phi^{-1}(1 - \beta)$ 。

分位数 $\Phi^{-1}(1 - \alpha)$ 和 $\Phi^{-1}(1 - \beta)$ 表示，以标准差 $\sigma_{\hat{\theta}}$ 为单位，区间下限 a 和上限 b 相对于估计值 $\hat{\theta}_{\text{obs}}$ 的距离。标准高斯分布的分位数与置信水平之间的关系在图 9.4 进行了说明，其中图 9.4(a) 对应于中心置信区间，图 9.4(b) 对应于单侧置信区间。

考虑 $\alpha = \beta = \gamma/2$ 的中心置信区间。经常选择置信水平 $1 - \gamma$ 使分位数为某个小整数，例如 $\Phi^{-1}(1 - \gamma/2) = 1, 2, 3, \dots$ 。类似地，对于单侧置信区间(极限)，我们也经常选择 $\Phi^{-1}(1 - \alpha)$ 等于某个小整数。表 9.1 给出了中心置信区间和单侧置信区间的常用值。或者，我们可以选择置信水平而不是分位数为一个约整数。

表9.2给出了常用值。其他可能的值可以从文献 [3, 20, 24] 得到，或者利用计算程序计算得到 (例如 [25] 中的子程序 **GAUSIN**，或 [26] 中高斯分布的分位数函数，即 **TMath::NormQuantile(double p)**)。

$\Phi^{-1}(1 - \gamma/2)$	$1 - \gamma$	$\Phi^{-1}(1 - \alpha)$	$1 - \alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987

表 9.1: 标准高斯分布不同分位数 Φ^{-1} 对应的置信水平：左侧为中心置信区间的分位数 $\Phi^{-1}(1 - \gamma/2)$ 和置信水平 $1 - \gamma$ ；右侧为单侧置信区间的分位数 $\Phi^{-1}(1 - \alpha)$ 和置信水平 $1 - \alpha$ 。

$1 - \gamma$	$\Phi^{-1}(1 - \gamma/2)$	$1 - \alpha$	$\Phi^{-1}(1 - \alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

表 9.2: 标准高斯分布不同置信水平对应的分位数 Φ^{-1} ：左侧为中心置信区间的置信水平 $1 - \gamma$ 和分位数 $\Phi^{-1}(1 - \gamma/2)$ ；右侧为单侧置信区间的置信水平 $1 - \alpha$ 和分位数 $\Phi^{-1}(1 - \alpha)$ 。

对于惯用的 68.3% 置信区间， $\alpha = \beta = \gamma/2$ ，相应地 $\Phi^{-1}(1 - \gamma/2) = 1$ ，即，‘一倍标准差误差棒’。这给出非常简单的结果

$$[a, b] = [\hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}, \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}]. \quad (9.13)$$

因此，对于服从高斯分布的估计量，68.3% 的中心置信区间由估计值加减一个标准差给出。于是， θ 的测量结果可以简单报道为 $\hat{\theta}_{\text{obs}} \pm \sigma_{\hat{\theta}}$ 。

如果标准差 $\sigma_{\hat{\theta}}$ 无法先验地知道，而需要从数据中估计得到，那么原则上情况会更复杂一些。例如，如果使用估计的标准差 $\hat{\sigma}_{\hat{\theta}}$ 而不是 $\sigma_{\hat{\theta}}$ ，则累计分布 $G(\hat{\theta}; \theta, \hat{\sigma}_{\hat{\theta}})$ 与标准高斯分布的累积分布 Φ 之间的关系就不那么简单，因为 $\hat{\sigma}_{\hat{\theta}}$ 一般来说依赖于 $\hat{\theta}$ 。然而，在实际使用中，即使使用估计值 $\hat{\sigma}_{\hat{\theta}}$ 而非 $\sigma_{\hat{\theta}}$ ，仍然可以应用前面的方法，只要 $\hat{\sigma}_{\hat{\theta}}$ 对真实标准差的近似足够好，例如当数据样本足够多时。对于小数据样本的情形，其中 $\hat{\theta}$ 表示标准差未知的 n 个高斯随机变量的均值，可以通过将累积分布 $G(\hat{\theta}; \theta, \hat{\sigma}_{\hat{\theta}})$ 与学生分布 $t(\text{Student's } t \text{ distribution})$ 联系起来 (参见文献 [20], [24] 的 10.2 节)，确定置信区间。

如果估计量的概率密度函数 $g(\hat{\theta}; \theta)$ 不服从高斯分布，则置信区间的准确确定将会更加困难，或者更糟糕的情况是，这个概率密度函数没有解析形式。对于非高斯分布的概率密度函数，有时可能通过进行参数变换 $\theta \rightarrow \eta(\theta)$ ，使得估计量 $\hat{\eta}$ 的概率密度函数近似为高斯分布。之后，可以将变换参数 η 的置信区间转换回 θ 的区间。9.5 节给出了这种方法的一个例子。

9.4 泊松分布均值的置信区间

除了高斯分布的估计量，另一种常见的情况是测量结果为泊松变量 $n(n = 0, 1, 2, \dots)$ 。由 (2.9) 式知道，观测到 n 的概率为

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}, \quad (9.14)$$

并且参数 ν 等于期望值 $E[n]$ 。很容易得到 ν 的极大似然估计量为 $\hat{\nu} = n$ 。假设单次测量得到的估计值为 $\hat{\nu}_{\text{obs}} = n_{\text{obs}}$ ，并且我们希望以此构造均值 ν 的置信区间。

对于离散变量的情形，不能直接应用9.2节介绍的确定置信区间的过程，原因是，用来确定置信带的函数 $u_\alpha(\theta)$ 和 $v_\beta(\theta)$ 不一定对参数 θ 的所有值都存在。例如，对于泊松分布来说，我们需要对参数 ν 的所有值求出 $u_\alpha(\nu)$ 和 $v_\beta(\nu)$ ，使其满足 $P(\hat{\nu} \geq u_\alpha(\nu)) = \alpha$ ， $P(\hat{\nu} \leq v_\beta(\nu)) = \beta$ 。但是，如果 α 和 β 固定，由于 $\hat{\nu}$ 只能取离散值，这两个方程通常来说只对 ν 的特殊值才成立。

然而，我们仍然可以利用 (9.9) 式确定置信区间 $[a, b]$ 。对于离散随机变量并且参数为 ν 的情况，这两个方程变为

$$\begin{aligned}\alpha &= P(\hat{\nu} \geq \hat{\nu}_{\text{obs}}; a), \\ \beta &= P(\hat{\nu} \leq \hat{\nu}_{\text{obs}}; b),\end{aligned}\tag{9.15}$$

对于泊松变量的特殊情况，可以得到

$$\begin{aligned}\alpha &= \sum_{n=n_{\text{obs}}}^{\infty} f(n; a) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} f(n; a) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{a^n}{n!} e^{-a}, \\ \beta &= \sum_{n=0}^{n_{\text{obs}}} f(n; b) = \sum_{n=0}^{n_{\text{obs}}} \frac{b^n}{n!} e^{-b}.\end{aligned}\tag{9.16}$$

对于估计值 $\hat{\nu} = n_{\text{obs}}$ 以及给定的概率 α 和 β ，可以数值求解出 a 和 b 。这里，可以利用泊松分布与卡方分布之间的关系，

$$\begin{aligned}\sum_{n=0}^{n_{\text{obs}}} \frac{\nu^n}{n!} e^{-\nu} &= \int_{2\nu}^{\infty} f_{\chi^2}(z; n_d = 2(n_{\text{obs}} + 1)) dz \\ &= 1 - F_{\chi^2}(2\nu; n_d = 2(n_{\text{obs}} + 1)),\end{aligned}\tag{9.17}$$

其中 f_{χ^2} 是自由度数目为 n_d 的卡方概率密度函数， F_{χ^2} 是对应的累积分布。于是可以得到

$$\begin{aligned}a &= \frac{1}{2} F_{\chi^2}^{-1}(\alpha; n_d = 2n_{\text{obs}}), \\ b &= \frac{1}{2} F_{\chi^2}^{-1}(1 - \beta; n_d = 2(n_{\text{obs}} + 1)).\end{aligned}\tag{9.18}$$

卡方分布的分位数 $F_{\chi^2}^{-1}$ 可以查表得到 (例如文献 [3])，或者利用计算机程序得到，如 [25] 中的 **CHISIN** 子程序，或 [26] 中的函数 **TMath::ChisquareQuantile(double p, double ndf)**。表9.3给出了 $n_{\text{obs}} = 0, \dots, 10$ 对应的一些值。

需要注意的是，如果 $n_{\text{obs}} = 0$ ，则无法确定下限 a 。(9.15) 式表示，如果 $\nu = a(\nu = b)$ ，则观测到不小于 (不大于) 实际观测值的概率为 $\alpha(\beta)$ 。由于 (9.15) 中的不等式包含了等号，即 $\hat{\nu} = \hat{\nu}_{\text{obs}}$ ，我们保守地得到了一个大的置信区间，即

$$\begin{aligned}P(\nu \geq a) &\geq 1 - \alpha, \\ P(\nu \leq b) &\geq 1 - \beta, \\ P(a \leq \nu \leq b) &\geq 1 - \alpha - \beta.\end{aligned}\tag{9.19}$$

一个很重要的特殊情形是观测数目 n_{obs} 为零的情况，此时，我们对上限 b 感兴趣。(9.15) 式变为

$$\beta = \sum_{n=0}^0 \frac{b^n e^{-b}}{n!} = e^{-b},\tag{9.20}$$

n_{obs}	下限 a			上限 b		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.01$
0	-	-	-	2.30	3.00	4.61
1	0.105	0.051	0.010	3.89	4.74	6.64
2	0.532	0.355	0.149	5.32	6.30	8.41
3	1.10	0.818	0.436	6.68	7.75	10.04
4	1.74	1.37	0.823	7.99	9.15	11.60
5	2.43	1.97	1.28	9.27	10.51	13.11
6	3.15	2.61	1.79	10.53	11.84	14.57
7	3.89	3.29	2.33	11.77	13.15	16.00
8	4.66	3.98	2.91	12.99	14.43	17.40
9	5.43	4.70	3.51	14.21	15.71	18.78
10	6.22	5.43	4.13	15.41	16.96	20.14

表 9.3: 观测事例数为 n_{obs} 时, 泊松分布置信区间的下限和上限。

或者说 $b = -\log \beta$ 。对于置信水平 $1 - \beta = 95\%$ 的上限, 可以得到 $b = -\log(0.05) = 2.996 \approx 3$ 。所以, 如果某种稀有事例发生的次数可以当做均值为 ν 的泊松变量, 并且我们在寻找这种事例的实验中观测到的事例数为零, 那么 95% 置信水平的均值上限为 3。也就是说, 如果平均值实际上就是 $\nu = 3$, 那么观测到零事例的概率为 5%。

9.5 相关系数、参数变换的置信区间

很多情况下, 可以假定估计量的概率密度函数为高斯分布, 于是可以利用9.3节的结果得到置信区间。但有时这个条件并不满足。作为例子, 考虑两个连续随机变量 x 和 y 的相关系数, 这两个变量服从二维高斯概率密度函数 $f(x, y)$ (见 (2.30) 式)。假设样本包含 x 和 y 的 n 次独立观测, 并且根据估计量 r 确定相关系数 ρ 的置信区间 (参考 (5.12) 式),

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{j=1}^n (x_j - \bar{x})^2 \cdot \sum_{k=1}^n (y_k - \bar{y})^2 \right)^{1/2}}. \quad (9.21)$$

概率密度函数 $g(r; \rho, n)$ 的形式相当复杂, 文献 [49] 第 151 页给出了具体形式。图9.5给出了不同真实相关系数 ρ 下样本容量 $n = 20$ 时该函数的形状。可以看到, $g(r; \rho, n)$ 为非对称函数, 并且不对称度与 ρ 有关。可以证明, 在大样本极限下, $g(r; \rho, n)$ 趋近于高斯分布, 但是要使这个近似成立, 需要的样本数目相当大。(根据文献 [3], 至少需要 $n \geq 500$ 。) 对于图9.5中的小样本来说, $g(r; \rho, n)$ 的高斯近似是不可靠的, 因此不能使用 (9.12) 式确定置信区间。

原则上, 我们必须回到9.2节的过程, 而对于本例来说, 这个过程很难计算。然而, 存在一种简单的方法, 可以近似确定 ρ 的置信区间。费舍尔曾经证明, 相比于 r , 统计量

$$z = \tanh^{-1} r = \frac{1}{2} \log \frac{1+r}{1-r} \quad (9.22)$$

的概率密度函数随样本容量 n 趋向于高斯极限的速度要快很多 (见文献 [6] 及其中提到的其它文献)。可以

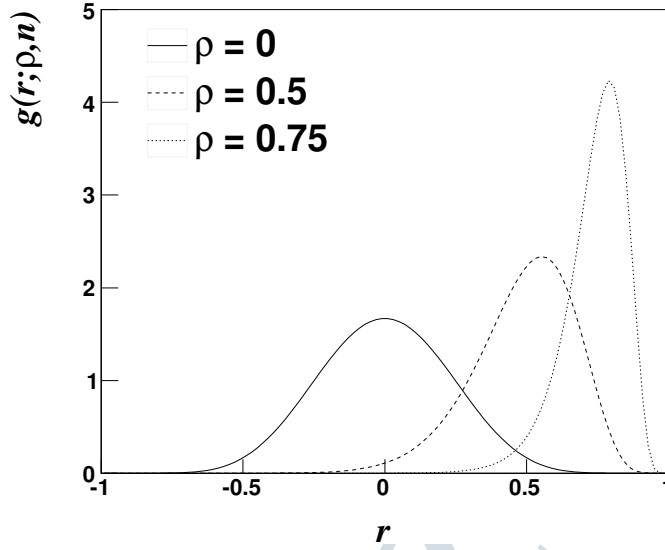


图 9.5: 样本容量 $n = 20$ 时, 相关系数 ρ 的估计量 r 的概率密度函数 $f(r; \rho, n)$, 不同曲线选取不同的 ρ 值。

以此作为 ζ 的估计量, ζ 定义为

$$\zeta = \tanh^{-1} \rho = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}. \quad (9.23)$$

可以证明, z 的期望值近似为

$$E[z] = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} + \frac{\rho}{2(n - 1)}. \quad (9.24)$$

其方差为

$$V[z] = \frac{1}{n - 3}. \quad (9.25)$$

我们将假定样本足够大, 可以保证 z 的概率密度函数为高斯分布, 并且 (9.24) 式中的偏倚项 $\rho/2(n - 1)$ 可以忽略不计。给定对 x 和 y 进行 n 测量的样本, z 可以由 (9.22) 式确定, 其标准差 $\hat{\sigma}_z$ 可以利用 (9.25) 式的方差进行估计。利用这些结果, 可以确定区间 $[z - \hat{\sigma}_z, z + \hat{\sigma}_z]$ 或者 (9.12) 式给出的一般性区间 $[a, b]$, 从而给出了置信水平为 $1 - \alpha$ 时 ζ 的下限 a , 以及置信水平为 $1 - \beta$ 时的上限 b 。于是, 可以利用 (9.22) 式的反函数, 将 $\zeta = \tanh^{-1} \rho$ 的置信区间 $[a, b]$ 变换回 ρ 的置信区间 $[A, B]$, 即 $A = \tanh a$, $B = \tanh b$ 。

例如, 考虑样本容量 $n = 20$, 并且得到估计值 $r = 0.5$ 。利用 (5.17) 式, r 的标准差可以估计为 $\hat{\sigma}_r = (1 - r^2)/\sqrt{n} = 0.168$ 。如果采取不正确的近似, 在这样的小样本情况下认为 r 服从高斯分布, 可以得到 ρ 的 68.3% 中心置信区间为 $[0.332, 0.668]$, 或者 99% 置信水平下的置信区间为 $[0.067, 0.933]$ 。由于样本相关系数 r 几乎是标准差 $\hat{\sigma}_r$ 的 3 倍, 我们可能会得到不正确的结论, 认为存在明显证据表明相关系数 ρ 不为零, 即所谓 ‘ 3σ 效应’。然而, 利用 z -变换, 可以得到 $z = 0.549$, $\hat{\sigma}_z = 0.243$ 。这对应于 ζ 的 99% 中心置信区间 $[-0.075, 1.174]$, 变换到 ρ 为 $[-0.075, 0.826]$ 。因此, 99% 中心置信区间包含了零相关系数的情况。

我们知道, 假设重复实验中 r 比实际观测值更大的概率为 α , 满足这个条件的真实参数的假设值就等于置信区间的下限。例如, 我们可以问, 下限为零的置信水平是多少。如果已经假定 $g(r; \rho, n)$ 为高斯分布, 对应的概率将是 0.14%。然而, 利用 z -变换, 下限为零的置信水平为 2.3%, 也就是说, 如果 ρ 为零, 得到 r 不小于实际观测值 $r = 0.5$ 的概率为 2.3%。因此, 相对于假定 r 为高斯分布利用标准差 $\hat{\sigma}_r$ 得到的结论, 相关系数不为零的实际证据并不那么强烈。

9.6 用似然函数或卡方求置信区间

即使对于非高斯分布的估计量，也可以用简单的近似方法确定置信区间，这个近似方法利用了似然函数或者卡方函数（二者等价，存在关系式 $L = \exp(-\chi^2/2)$ ）。首先考虑大样本极限下参数 θ 的极大似然估计量 $\hat{\theta}$ 。在大样本极限下，可以证明（见 [7] 的第 18 章）概率密度函数 $g(\hat{\theta}; \theta)$ 变成高斯分布，

$$g(\hat{\theta}; \theta) = \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(-\frac{(\hat{\theta}' - \theta)^2}{2\sigma_{\hat{\theta}}^2}\right), \quad (9.26)$$

中心值为参数真值 θ ，标准差为 $\sigma_{\hat{\theta}}$ 。

在大样本极限下，还可以证明，似然函数本身也是高斯形式，中心值为极大似然估计值 $\hat{\theta}$ ，即

$$L(\theta) = L_{\max} \exp\left(-\frac{(\hat{\theta}' - \theta)^2}{2\sigma_{\hat{\theta}}^2}\right). \quad (9.27)$$

利用 (6.16) 式的 RCF 不等式，该不等式对大样本极限的极大似然估计量变成等式，可以得到，(9.27) 式极大似然函数中的 $\sigma_{\hat{\theta}}$ 与 (9.26) 式中的 $\sigma_{\hat{\theta}}$ 相同。这在 6.7 节的 (6.24) 式已经遇到过，当时用似然函数对估计量 $\hat{\theta}$ 的方差进行估计。这给出估计 $\sigma_{\hat{\theta}}$ 的一个简单方法，因为通过将参数 θ 改变 N 个标准差，对数似然函数从其最大值减小 $N^2/2$ ，

$$\log L(\hat{\theta} \pm N\sigma_{\hat{\theta}}) = \log L_{\max} - \frac{N^2}{2}. \quad (9.28)$$

然而，从上一节得到的结果可以知道，对于高斯分布的估计量 $\hat{\theta}$ ，68.3% 的中心置信区间可以用估计量及其标准差的估计值 $\hat{\sigma}_{\hat{\theta}}$ 进行构造，即 $[a, b] = [\hat{\theta} - \hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \hat{\sigma}_{\hat{\theta}}]$ （或者更一般地根据 (9.12) 式构造置信水平为 $1 - \gamma$ 的置信区间）。因此，68.3% 的置信区间由使似然函数从最大值减小 1/2 的 θ 值给出。（当然，这需要假设 $\hat{\theta}$ 为极大似然估计量，因此对应于似然函数的最大值。）

实际上，可以证明，即使似然函数不是参数的高斯函数，中心置信区间 $[a, b] = [\hat{\theta} - c, \hat{\theta} + d]$ 仍然可以用下式进行近似，即

$$\log L(\hat{\theta}_{-c}^{+d}) = \log L_{\max} - \frac{N^2}{2}. \quad (9.29)$$

其中 $N = \Phi^{-1}(1 - \gamma/2)$ 为标准高斯分布对应于 $1 - \gamma$ 置信水平的分位数。（例如， $N = 1$ 对应于 68.3% 的中心置信区间，参见表 9.1。）在对高斯不确定度进行最小二乘拟合的情况下，即 $\log L = -\chi^2/2$ ，上式变为

$$\chi^2(\hat{\theta}_{-c}^{+d}) = \chi_{\max}^2 + N^2. \quad (9.30)$$

(9.29) 式和 (9.30) 式可以近似 9.2 节的经典置信区间，文献 [20, 23] 对此给出了探索性的证明。(9.29) 式和 (9.30) 式给出了估计统计不确定性最常用的方法之一。然而，需要记住的是，只有在大样本极限下，与 9.2 节方法的对应才是严格的。有些作者推荐使用‘似然区间’的术语表示从似然函数得到的区间 [20, 55]。不管用什么名称，都需要牢记的一点是，这里将这个区间解释为经典置信区间的近似，也就是说，这是构造的一个随机区间，其包含参数真值的概率为某个给定值。

作为例子，考虑指数分布参数 τ 的估计量 $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$ ，这个例子与 6.2 节的例子一样（也参见 6.7 节）。在 6.2 节，利用给定样本，即对指数分布随机变量 t 进行 $n = 50$ 次测量，采用极大似然法对参数 τ 进行了估计。这个样本足够大，标准差 $\sigma_{\hat{\tau}}$ 近似为使对数似然函数从最大值减小 1/2 的 τ 值（见图 6.4）。得到的结果为 $\hat{\tau} = 1.06$ ， $\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$ 。

图 9.6 给出了对数似然函数 $\log L(\tau)$ 作为 τ 的函数，其中样本是对指数分布随机变量仅有 $n = 5$ 次测量，数据是通过蒙特卡罗方法产生的，参数真值为 $\tau = 1$ 。由于样本容量很小，相比之前的例子 ($n = 50$)，现在的对数似然函数不太像抛物线。

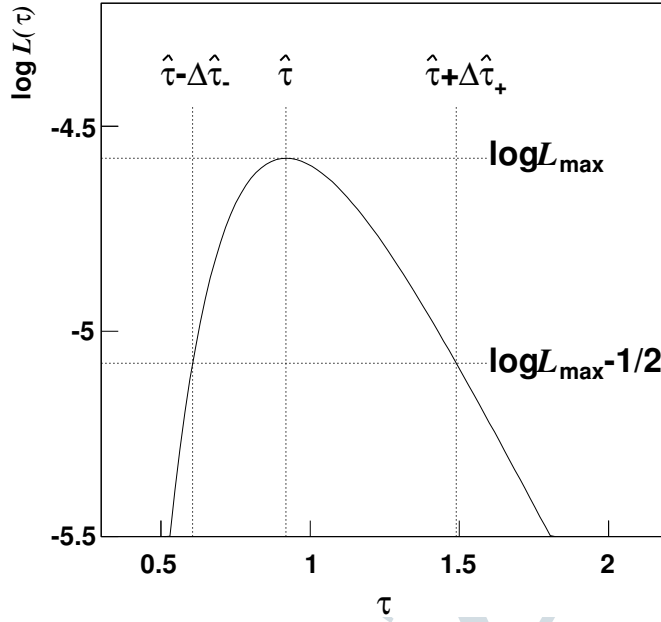


图 9.6: 样本容量 $n = 5$ 时, 对数似然函数 $\log L(\tau)$ 作为 τ 的函数。 $\log L(\tau) = \log L_{\max} - 1/2$ 确定的区间 $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ 可以用来近似 68.3% 的中心置信区间。

我们仍然可以用 $\log L_{\max} - 1/2$ 确定的区间的半宽度近似标准差 $\sigma_{\hat{\tau}}$, 但这并不是我们真正希望的。用置信区间可以更好地表达统计不确定度, 因为根据置信区间可以立刻知道这个区间覆盖参数真值的概率。而且, 通过给出中心置信区间 (因此给出不对称不确定度, $\Delta\hat{\tau}_- \neq \Delta\hat{\tau}_+$), 真实参数高于或低于区间上下限的概率相同。如图 9.6 中所说明, 中心置信区间可以近似为 $\log L(\tau) = \log L_{\max} - 1/2$ 的 τ 值, 结果为 $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+] = [0.55, 1.37]$, 或者 $\hat{\tau} = 0.85^{+0.52}_{-0.30}$ 。

实际上, 在 6.7 节同样可以这样处理, 给出的结果为 $\hat{\tau} = 1.062^{+0.165}_{-0.137}$ 。究竟应该选择这种方法, 还是简单地报道平均的对称不确定度 (即 $\hat{\tau} = 1.06 \pm 0.15$), 取决于需要给出什么精度的统计不确定度。对于图 9.6 中 $n = 5$ 的例子, 误差棒的不对称度非常大, 很可能需要使用 68.3% 的中心置信区间, 给出 $\hat{\tau} = 0.85^{+0.52}_{-0.30}$ 的结果。

9.7 多维置信区域

在 9.2 节, 我们构造了置信区间 $[a, b]$, 使其包含参数 θ 的概率为 $1 - \gamma$ 。为了推广到 n 个参数 $\theta = (\theta_1, \dots, \theta_n)$ 的情况, 可能会尝试构造 n 维置信区间 $[\mathbf{a}, \mathbf{b}]$, 使得同时满足 $a_i < \theta_i < b_i (i = 1, \dots, n)$ 的概率为给定值。实际证明这种构造很难计算, 一般很少使用。

然而, 在参数空间构造一个**置信区域(confidence region)**, 使得真实参数 θ 包含于这个区间的概率为给定值 (至少近似成立), 这种做法相对来说简单得多。这个区域不存在 $a_i < \theta_i < b_i, i = 1, \dots, n$ 的形式, 表达式非常复杂, 在大样本极限下趋近于 n 维超椭球体。

与单参数情形一样, 大样本极限下, 估计量 $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ 的联合概率密度函数和似然函数都变成高斯函数。也就是说, $\hat{\theta}$ 的联合概率密度函数变成

$$g(\hat{\theta}|\theta) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp \left[-\frac{1}{2} Q(\hat{\theta}|\theta) \right], \quad (9.31)$$

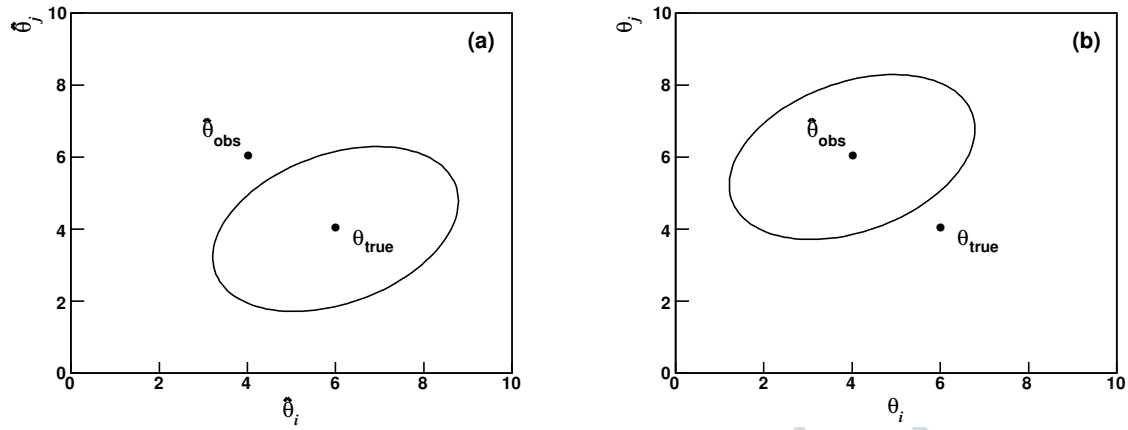


图 9.7: (a) $\hat{\theta}$ -空间中一个等值线 $g(\hat{\theta}; \theta_{\text{true}})$ (即 $Q(\hat{\theta}, \theta_{\text{true}})$ 为常数)。 (b) $\hat{\theta}$ -空间中对应于 $Q(\hat{\theta}_{\text{obs}}, \theta)$ 为常数的等值线 $L(\theta)$ 。 θ_{true} 和 $\hat{\theta}_{\text{obs}}$ 分别表示 θ 和 $\hat{\theta}$ 的特殊值。

其中 Q 定义为

$$Q(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^T V^{-1} (\hat{\theta} - \theta). \quad (9.32)$$

这里 V^{-1} 为协方差矩阵的逆, 上标 T 表示转置矢量 (即行矢量)。 $g(\hat{\theta}|\theta)$ 为常数的等值线对应于 $Q(\hat{\theta}, \theta)$ 为常数。这是 $\hat{\theta}$ -空间中的椭圆 (或者, 对于超过二维的情形来说是超椭球体), 以真实参数 θ 为中心。图 9.7(a) 给出了 $Q(\hat{\theta})$ 的等值线, 其中 θ_{true} 表示 θ 的一个特殊值。

与一维情形类似, 也可以证明似然函数 $L(\theta)$ 为高斯形式, 中心值为极大似然估计量 $\hat{\theta}$,

$$L(\theta) = L_{\max} \exp \left[-\frac{1}{2} (\hat{\theta} - \theta)^T V^{-1} (\hat{\theta} - \theta) \right] = L_{\max} \exp \left[-\frac{1}{2} Q(\hat{\theta}, \theta) \right]. \quad (9.33)$$

协方差矩阵的逆 V^{-1} 与 (9.31) 式中一样, 这可以从 (6.19) 式的 RCF 不等式看出, 其中用到了大样本极限下极大似然估计量可以取 RCF 边界。 Q 这个量在这里被看作参数 θ 的函数, 在估计值 $\hat{\theta}$ 处取最大值。图 9.7(b) 说明了这一点, 其中 $\hat{\theta}$ 等于某个特殊值 $\hat{\theta}_{\text{obs}}$ 。由于 (9.32) 式的定义关于 θ 和 $\hat{\theta}$ 对称, 在 (9.31) 式的概率密度函数和 (9.33) 式的似然函数中, Q 这个量的值相同, 即 $Q(\hat{\theta}, \theta) = Q(\theta, \hat{\theta})$ 。

如 7.5 节所讨论, 可以证明, 如果 $\hat{\theta}$ 服从 n 维高斯概率密度函数 $g(\hat{\theta}, \theta)$, 则 $Q(\hat{\theta}, \theta)$ 服从自由度数目为 n 的卡方分布。 $Q(\hat{\theta}, \theta)$ 小于某个值 Q_γ (即估计值处于真值 θ 的一定范围之内) 意味着 $Q(\theta, \hat{\theta}) < Q_\gamma$ (即真值 θ 处于估计值相同的范围之内)。因此, 这两个事例具有相同的概率,

$$P(Q(\hat{\theta}, \theta) \leq Q_\gamma) = \int_0^{Q_\gamma} f(z; n) dz, \quad (9.34)$$

其中 $f(z; n)$ 是自由度数目为 n 的卡方分布 (式 (2.34))。 Q_γ 值的选择对应于给定的概率,

$$\int_0^{Q_\gamma} f(z; n) dz = 1 - \gamma. \quad (9.35)$$

也就是说,

$$Q_\gamma = F^{-1}(1 - \gamma; n) \quad (9.36)$$

是卡方分布对应于 $1 - \gamma$ 的分位数。 θ -空间中 $Q(\theta, \hat{\theta}) \leq Q_\gamma$ 定义的区域称为置信水平为 $1 - \gamma$ 的**置信区域(confidence region)**。对于高斯形式的似然函数, 形如 (9.33) 式, 可以通过求对数似然函数从最大值减小 $Q_\gamma/2$ 的 θ 值构造置信区域, 即通过求解下式进行构造,

$$\log L(\theta) = \log L_{\max} - \frac{Q_\gamma}{2}. \quad (9.37)$$

与单参数情形一样，我们仍然可以使用 (9.37) 式给出的公式，即使似然函数不是高斯形式，因而 (9.34) 式的概率只是近似成立。随着参数数目的增加，趋向于高斯极限的速度作为样本容量的函数将不断变慢，而且，当样本足够大时，(9.34) 式的应用很难定量处理。如果需要，可以利用蒙特卡罗计算，确定根据 (9.37) 式构造的区域包含真实参数的概率。

表9.4给出了几个不同置信水平 $1 - \gamma$ 下卡方分布的分位数 $Q_\gamma = F^{-1}(1 - \gamma; n)$ ，其中参数数目分别取 $n = 1, 2, 3, 4, 5$ 。不同分位数 Q_γ 值对应的置信水平由表9.5给出。

$1 - \gamma$	Q_γ				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

表 9.4: 不同置信水平 $1 - \gamma$ 值，拟合参数 $n = 1, 2, 3, 4, 5$ 时，对应的 Q_γ 值。

Q_γ	$1 - \gamma$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

表 9.5: 不同 Q_γ 值，拟合参数 $n = 1, 2, 3, 4, 5$ 时，对应的置信水平 $1 - \gamma$ 的值。

对于 $n = 1$ ，可以证明 (9.36) 式 Q_γ 的表达式满足

$$\sqrt{Q_\gamma} = \Phi^{-1}(1 - \gamma/2), \quad (9.38)$$

其中 Φ^{-1} 为标准正态分布 (累积分布) 的反函数。因此，这里的过程简化为9.6节的单参数问题，其中 $N = \sqrt{Q_\gamma}$ 为标准差区间的半宽度 (见 (9.28) 式和 (9.29) 式)。因此，表9.4和9.5中 $n = 1$ 的值通过 (9.38) 式与表9.1和表9.2中的值联系起来。

随着 n 的增大，给定 Q_γ 的置信水平不断下降。例如，对于单参数情形， Q_γ 等于 1 对应于 $1 - \gamma = 0.683$ 。对于 $n = 2$ ， $Q_\gamma = 1$ 给出的置信水平只有 0.393，为了得到 $1 - \gamma = 0.683$ ，需要 $Q_\gamma = 2.30$ 。

需要强调的是，与单参数情形一样，置信区域 $Q(\theta, \hat{\theta}) \leq Q_\gamma$ 是 θ -空间的随机区域。置信区域随实验的重复而变化，因为 $\hat{\theta}$ 是随机变量。相反，真实参数则是未知常数。

9.8 物理边界的极限

实验的目的经常是寻找新的效应，新效应的存在意味着某个参数不等于零。例如，我们可以尝试测量中微子的质量，中微子在标准模型理论中的质量为零。如果数据得出的参数值明显不是零，则说明发现了新效应，结果用参数值和反映其不确定度的置信区间给出。相反，如果数据得出参数的拟合值与零一致，

则实验结果以给出参数上限的方式报道。(也存在这样的类似情况,即如果没有新现象则对应的参数非常大或者无穷,这时候要给出参数下限。为简单起见,这里只考虑上限的情况。)

当估计量可以取排除区域的值时,就会出现困难。如果参数 θ 的估计量 $\hat{\theta}$ 形如 $\hat{\theta} = x - y$, 其中 x 和 y 都是随机变量(即它们都有随机测量不确定度),就会出现这种情况。例如,粒子质量的平方可以通过独立测量粒子能量 E 和动量 p , 并利用 $\hat{m}^2 = E^2 - p^2$ 进行估计。尽管质量平方应当为正, E^2 和 p^2 的测量不确定度可能导致 \hat{m}^2 的值为负。那么,问题是如何给出 m^2 的上限,或者更一般地,当估计值位于或接近排除区域时,如何给出参数 θ 的上限。

进一步考虑估计量 $\hat{\theta} = x - y$ 的例子,其中 x 和 y 为高斯随机变量,均值分别为 μ_x 和 μ_y , 方差分别为 σ_x^2 和 σ_y^2 。可以证明,两个随机变量的差 $\hat{\theta} = x - y$ 也是高斯变量,均值 $\theta = \mu_x - \mu_y$, 方差为 $\sigma_{\hat{\theta}}^2 = \sigma_x^2 + \sigma_y^2$ 。(可以用第十章介绍的特征函数证明。)

假设先验已知 θ 为非负(例如,质量平方),并假设实验得到估计量 $\hat{\theta}$ 的估计值为 $\hat{\theta}_{\text{obs}}$ 。根据(9.12)式,置信水平 $1 - \beta$ 的上限 θ_{up} 为

$$\theta_{\text{up}} = \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}} \Phi^{-1}(1 - \beta). \quad (9.39)$$

对于常用的 95% 置信水平,从表9.2可以到分位数 $\Phi^{-1}(0.95) = 1.645$ 。

区间 $(-\infty, \theta_{\text{up}}]$ 是这样构造的,不论 θ 的真实值是多少,这个区间包含 θ 真值的概率为 95%。假设标准差为 $\sigma_{\hat{\theta}} = 1$, 实验结果为 $\hat{\theta}_{\text{obs}} = -2.0$ 。利用(9.39)可以得到 95% 置信水平下 $\theta_{\text{up}} = -0.355$ 。不仅 $\hat{\theta}_{\text{obs}}$ 位于禁止区域(因为如果 θ 的确等于零,半数估计值应当处于禁止区域),而且置信区域的上限也小于零。这并不奇怪,如果 θ 的真值的确为零,这种情况发生的概率大约为 5%。

就置信区间的定义而言,没有出现根本的问题。置信区间定义为,在重复实验中,有给定比率的区间覆盖 θ 的真值,而显然我们遇到的是 θ 不在这个区间的某次实验。但是这个结果不令人满意,因为即使不做任何实验之前,我们已经知道 θ 大于零(当然也大于 $\theta_{\text{up}} = -0.355$)。

不管上限如何,报道实际估计值及其标准差 $\hat{\theta}_{\text{obs}} \pm \sigma_{\hat{\theta}}$ 都很重要,即使估计值在物理排除区域。只有这样,很多实验的平均结果才会收敛于正确的值(见7.6节),当然前提是估计量是无偏的。如果 $\hat{\theta}$ 的概率密度函数明显不是高斯分布,需要给出完整的似然函数 $L(\theta)$, 从而可以与其他实验的似然函数合并起来,如6.12节所讨论。

然而,多数实验者希望报道某种上限,在前面讨论的情形中,提出了很多方法(参见[56,57])。不幸的是,到目前为止还没有形成惯例,告诉我们到底应该如何处理,因此,我们需要声明采用了什么方法。

为了处理非物理区域上限问题的困难,可以简单地尝试提高置信水平,直到上限进入允许区域。在前面的例子中,如果取置信水平 $1 - \beta = 0.99$, 则从表9.2可以得到 $\Phi^{-1}(0.99) = 2.326$, 从而有 $\theta_{\text{up}} = 0.326$ 。这个结果使我们在 99% 的高置信水平下,给出比实验内在分辨率($\sigma_{\hat{\theta}} = 1$)还小的上限,这显然具有误导性。当然,更糟糕的是,通过调整置信水平给出任意小的上限,例如,在 97.725% 置信水平下, $\Phi^{-1}(0.97725) = 2.00001$, 得到 $\theta_{\text{up}} = 10^{-5}$ 。

为了避免这类困难,常用的方法是在应用(9.39)式之前,将负估计值简单地平移为零,即

$$\theta_{\text{up}} = \max(\hat{\theta}_{\text{obs}}, 0) + \sigma_{\hat{\theta}} \Phi^{-1}(1 - \beta). \quad (9.40)$$

这样,上限至少始终保持与实验分辨率的量级相同。如果 $\hat{\theta}_{\text{obs}}$ 为正,得到的上限与经典过程一致。这种方法从直觉上很有吸引力,所以被经常采用,但是,这时无法象前面一样,将上限解释为按照 $1 - \beta$ 的概率包含参数真值的区间。在这种方法中,范围概率显然大于 $1 - \beta$, 因为(9.40)式平移后的上限总是大于或等于(9.39)式的值。

另一种可供选择的方法是,基于贝叶斯验后概率密度函数 $p(\theta|\mathbf{x})$ 报道参数区间。如6.13节所述,这个

验后概率可以由贝叶斯定理得到,

$$p(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)\pi(\theta)}{\int L(\mathbf{x}|\theta')\pi(\theta')d\theta'}, \quad (9.41)$$

其中 \mathbf{x} 表示观测数据, $L(\mathbf{x}|\theta)$ 为似然函数, $\pi(\theta)$ 为 θ 的验前概率密度函数。在6.13节, $p(\theta|\mathbf{x})$ 的众数被用作 θ 的估计量, 并且已经证明, 如果验前概率密度函数 $\pi(\theta)$ 为均匀分布, 那么这个估计量与极大似然估计量一致。这里, 我们可以用 $p(\theta|\mathbf{x})$ 确定出区间 $[a, b]$, 使得对于给定概率 α 和 β , 满足

$$\begin{aligned} \alpha &= \int_{-\infty}^a p(\theta|\mathbf{x})d\theta, \\ \beta &= \int_b^{\infty} p(\theta|\mathbf{x})d\theta. \end{aligned} \quad (9.42)$$

选择 $\alpha = \beta$ 则给出中心置信区间, 例如可以选择 $1 - \alpha - \beta = 68.3\%$ 。另一种可能性是, 选择 α 和 β , 使得 $[a, b]$ 区间内所有 $p(\theta|\mathbf{x})$ 的值都大于区间外的值, 这意味着 $p(a|\mathbf{x}) = p(b|\mathbf{x})$ 。可以证明, 这种方法给出最短可能区间。

贝叶斯区间的优点在于, 先验知识 (例如 $\theta > 0$) 可以很容易地考虑进来, 只要令验前概率密度函数 $\pi(\theta)$ 在排除区域等于零即可。于是, 贝叶斯定理给出的验后概率 $p(\theta|\mathbf{x})$ 满足, 当 $\theta = 0$ 时, $p(\theta|\mathbf{x}) = 0$ 。因此, 上限可以通过下式确定

$$1 - \beta = \int_{-\infty}^{\theta_{\text{up}}} p(\theta|\mathbf{x})d\theta = \frac{\int_{-\infty}^{\theta_{\text{up}}} L(\mathbf{x}|\theta)\pi(\theta)d\theta}{\int_{-\infty}^{\infty} L(\mathbf{x}|\theta)\pi(\theta)d\theta}. \quad (9.43)$$

这里的困难在6.13节已经提到, 即验前概率密度函数 $\pi(\theta)$ 无法唯一指定。通常的选择是

$$\pi(\theta) = \begin{cases} 0 & \theta < 0, \\ 1 & \theta \geq 0. \end{cases} \quad (9.44)$$

具体的方法实际上是这样的: 在物理区域将似然函数归一化为单位面积, 然后积分至 θ_{up} , 使覆盖的面积比率为 $1 - \beta$ 。这个方法尽管简单, 但存在一些概念上的问题。对于已知 $\theta \geq 0$ 的情形 (例如中微子质量), 我们并不真的相信 $0 < \theta < 1$ 的验前概率与 $10^{40} < \theta < 10^{40} + 1$ 的概率相同。而且, 从 $\pi(\theta) = \text{常数}$ 得到的上限在参数的非线性变换下不能保持不变。

文献 [11] 中讨论过, 如果仅仅先验地知道 $\theta \geq 0$, 而没有其它先验信息, 那么应当采用

$$\pi(\theta) = \begin{cases} 0 & \theta \leq 0, \\ \frac{1}{\theta} & \theta > 0. \end{cases} \quad (9.45)$$

其好处是, 上限在参数任何幂次变换下都是不变的。

这等价于为 $\log \theta$ 选取形如 (9.44) 式的均匀验前概率密度。然而, 要想使用这个验前概率密度函数, 要求似然函数在 $\theta \rightarrow 0$ 和 $\theta \rightarrow \infty$ 时趋向于零, 或者要求 (9.43) 式的积分发散。所以, 这个方法无法应用于很多感兴趣的实际例子, 包括本节讨论的例子。因此, 尽管存在概念上的困难, 在求参数上限时, 最常用选择仍然是均匀分布的验前概率密度函数。

图9.8给出了 95% 置信水平下, 根据经典方法、平移方法和贝叶斯方法得到的参数上限作为 $\hat{\theta}_{\text{obs}} = x - y$ 的函数, 其中 $\sigma_{\hat{\theta}} = 1$ 。对于贝叶斯上限, 采用的验前概率为 $\pi(\theta) = \text{常数}$ 。在 $\hat{\theta}_{\text{obs}}$ 的区域, 平移方法和经典方法完全相同。贝叶斯上限总是正的, 并且始终大于经典上限。当 $\hat{\theta}_{\text{obs}}$ 超过实验精度 $\sigma_{\hat{\theta}}$ 时, 贝叶斯上限与经典上限很快相互趋近。

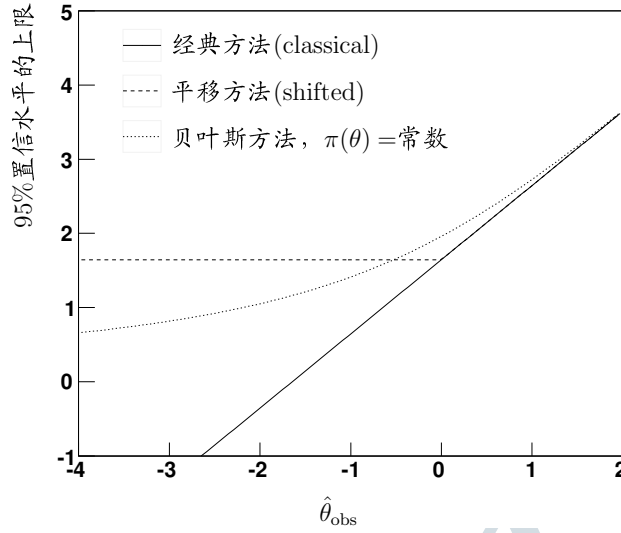


图 9.8: 9.8节的例子 95% 置信水平的上限, 分别对应经典方法、平移方法和贝叶斯方法。 $\hat{\theta}_{\text{obs}} \geq 0$ 时, 平移方法和经典方法的结果相同。

9.9 含本底的泊松变量均值的上限

作为最后一个例子, 重新考虑9.4节为均值等于 ν 的泊松变量 n 设置上限。经常碰到的情况会更复杂一些, 即 n 的观测值为期望的信号事例 n_s 与本底事例 n_b 之和,

$$n = n_s + n_b, \quad (9.46)$$

其中 n_s 和 n_b 均被视为泊松变量, 均值分别为 ν_s 和 ν_b 。暂时先假设本底的均值 ν_b 已知, 并且没有不确定度。对于 ν_s , 仅先验地知道 $\nu_s \geq 0$ 。我们的目标是, 在给定测量值 n 的情况下, 构造出信号参数 ν_s 的上限。

由于 n 是两个泊松变量之和, 可以证明 n 本身也是泊松变量, 概率密度函数为

$$f(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}. \quad (9.47)$$

ν_s 的极大似然估计量为

$$\hat{\nu}_s = n - \nu_b, \quad (9.48)$$

该估计量是无偏的, 因为 $E[n] = \nu_s + \nu_b$ 。用于确定置信区间的 (9.15) 式变为

$$\begin{aligned} \alpha &= P(\hat{\nu}_s \geq \hat{\nu}_s^{\text{obs}}; \nu_s^{\text{lo}}) = \sum_{n \geq n_{\text{obs}}} \frac{(\nu_s^{\text{lo}} + \nu_b)^n e^{-(\nu_s^{\text{lo}} + \nu_b)}}{n!}, \\ \beta &= P(\hat{\nu}_s \leq \hat{\nu}_s^{\text{obs}}; \nu_s^{\text{up}}) = \sum_{n \leq n_{\text{obs}}} \frac{(\nu_s^{\text{up}} + \nu_b)^n e^{-(\nu_s^{\text{up}} + \nu_b)}}{n!}. \end{aligned} \quad (9.49)$$

通过数值求解可以得到下限 ν_s^{lo} 和上限 ν_s^{up} 。与 $\nu_b = 0$ 的情形对比, 可以看到 (9.49) 式得到的上下限与没有本底情形的关系为

$$\begin{aligned} \nu_s^{\text{lo}} &= \nu_s^{\text{lo}}(\text{无本底}) - \nu_b, \\ \nu_s^{\text{up}} &= \nu_s^{\text{up}}(\text{无本底}) - \nu_b. \end{aligned} \quad (9.50)$$

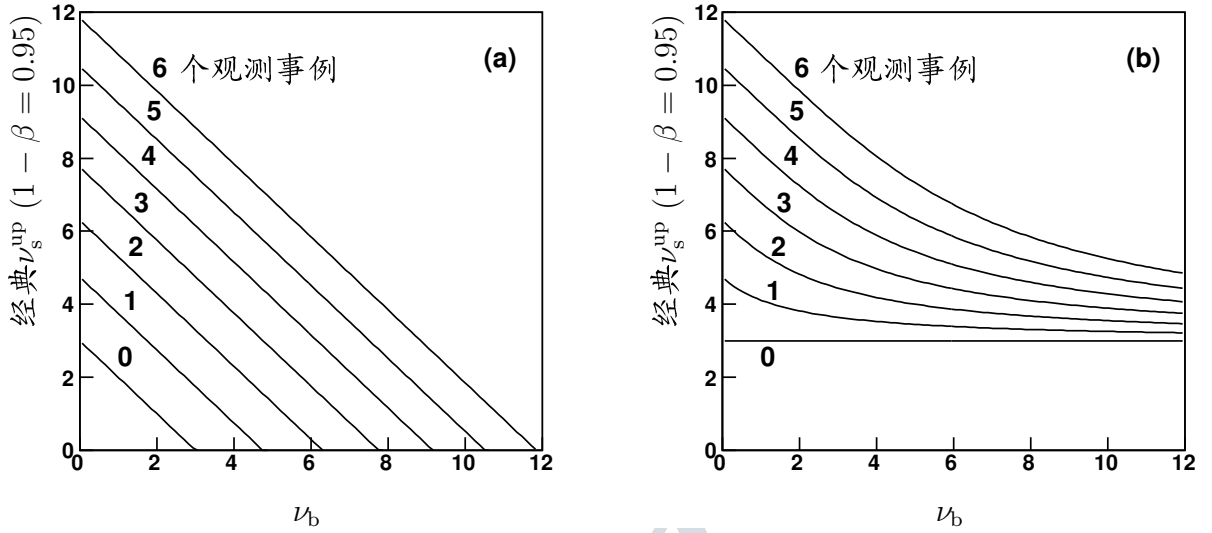


图 9.9: 对于不同观测事例数 n_{obs} , 置信水平 $1 - \beta = 0.95$ 时上限 ν_s^{up} 作为本地事例期望值 ν_b 的函数。(a) 经典极限。(b) 贝叶斯极限 (ν_s 的验前概率密度取为均匀分布)。

这里的困难与上一个例子遇到的困难类似。如果观测到的总事例数 n_{obs} 小于本底事例的期望值 ν_b , 就会出现困难。图 9.9(a) 给出了 $1 - \beta$ 置信水平下 ν_s^{up} 作为本底事例期望值 ν_b 的函数。如果 n_{obs} 很小, 而本底 ν_b 很大, ν_s^{up} 不存在非负解。由于 n_s 和 n_b 的涨落, 这种情况显然可以发生。

由于这些困难, 这种情况下不推荐使用经典极限的方法。如前面提到的一样, 我们总是应当报道 $\hat{\nu}_s$ 及其方差, 即使得到的 $\hat{\nu}_s$ 为负。只有这样, 很多实验的平均才会收敛于正确的值。另外, 如果我们希望报道 ν_s 的上限, 可以使用贝叶斯方法, 验前概率可以采用均匀分布 [58]。似然函数由 (9.47) 式的概率给出, 只是现在被当作 ν_s 的函数,

$$L(n_{\text{obs}}|\nu_s) = \frac{(\nu_s + \nu_b)^{n_{\text{obs}}}}{n_{\text{obs}}!} e^{-(\nu_s + \nu_b)}. \quad (9.51)$$

ν_s 的验后概率密度函数由贝叶斯定理给出,

$$p(\nu_s|n_{\text{obs}}) = \frac{L(n_{\text{obs}}|\nu_s)\pi(\nu_s)}{\int_{-\infty}^{\infty} L(n_{\text{obs}}|\nu'_s)\pi(\nu'_s)d\nu'_s}. \quad (9.52)$$

当 $\nu_s \geq 0$ 时, $\pi(\nu_s)$ 取为常数, 而当 $\nu_s < 0$ 时 $\pi(\nu_s)$ 取为零。置信水平为 $1 - \beta$ 的上限 ν_s^{up} 由下式给出

$$\begin{aligned} 1 - \beta &= \frac{\int_0^{\nu_s^{\text{up}}} L(n_{\text{obs}}|\nu_s)d\nu_s}{\int_0^{\infty} L(n_{\text{obs}}|\nu_s)d\nu_s} \\ &= \frac{\int_0^{\nu_s^{\text{up}}} (\nu_s + \nu_b)^{n_{\text{obs}}} e^{-(\nu_s + \nu_b)} d\nu_s}{\int_0^{\infty} (\nu_s + \nu_b)^{n_{\text{obs}}} e^{-(\nu_s + \nu_b)} d\nu_s}. \end{aligned} \quad (9.53)$$

其中的积分可以与不完全伽马函数联系起来 (参见 [21]), 或者, 由于 n_{obs} 为正整数, 可以通过变量替换 $x = \nu_s + \nu_b$, 并作 n_{obs} 次分部积分。于是, (9.53) 式变为

$$\beta = \frac{e^{-(\nu_s^{\text{up}} + \nu_b)} \sum_{n=0}^{n_{\text{obs}}} \frac{(\nu_s^{\text{up}} + \nu_b)^n}{n!}}{e^{-\nu_b} \sum_{n=0}^{n_{\text{obs}}} \frac{\nu_b^n}{n!}}. \quad (9.54)$$

通过数值求解可以得到上限 ν_s^{up} 。图 9.9(b) 给出了 n_{obs} 取不同值时上限作为 ν_b 的函数。对于没有本底的情形, 令 $\nu_b = 0$ 得到

$$\beta = e^{-\nu_s^{\text{up}}} \sum_{n=0}^{n_{\text{obs}}} \frac{(\nu_s^{\text{up}})^n}{n!}, \quad (9.55)$$

与9.16式的经典上限完全相同。这可以通过比较图9.9(a) 和 (b) 看出。贝叶斯极限总是大于等于对应的经典极限，只有当 $\nu_b = 0$ 时，二者才完全重合。

然而，必须将没有本底时二者的符合看作一种偶然，因为贝叶斯极限依赖于验前概率 $\pi(\nu_s)$ 的特殊选择，即选择常数验前概率。然而，这种吻合使我们可以避免选择经典极限还是贝叶斯极限的争论，这可能也是这种情况下均匀验前概率被广泛接受的原因之一。

实验结果往往不仅仅是观测到的事例数 n_{obs} ，而且还包含事例某些性质的测量值 x_1, \dots, x_n 。假设 x 的概率密度函数为

$$f(x; \nu_s, \nu_b) = \frac{\nu_s f_s(x) + \nu_b f_b(x)}{\nu_s + \nu_b}, \quad (9.56)$$

其中 $f_s(x)$ 为信号分量， $f_b(x)$ 为本底分量，二者均假设已知。如果这些概率密度函数具有不同的形式，则 x 的值包含区分观测事例为信号还是本底的额外信息。通过扩展的似然函数，可以象6.9节定义的那样，将这些信息结合到 ν_s 的极限中，

$$\begin{aligned} L(\nu_s) &= \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)} \prod_{i=1}^n \frac{\nu_s f_s(x_i) + \nu_b f_b(x_i)}{\nu_s + \nu_b} \\ &= \frac{e^{-(\nu_s + \nu_b)}}{n!} \prod_{i=1}^n [\nu_s f_s(x_i) + \nu_b f_b(x_i)], \end{aligned} \quad (9.57)$$

或者，也可以利用6.10节讨论分区间数据的对应公式。

在经典情形，利用极大似然函数求估计量 $\hat{\nu}_s$ 。然而，为了求经典上限，需要知道 $\hat{\nu}_s$ 的概率密度函数。这个概率密度函数的求解不再象之前只考虑事例数那样简单，通常必须数值求解。例如，可以进行蒙特卡罗实验，利用给定的 ν_s 值 (以及已知的 ν_b 值)，按照泊松分布产生 n_s 和 n_b 。通过调整 ν_s ，可以求出一个值，满足 $\hat{\nu}_s \leq \hat{\nu}_s^{\text{obs}}$ 的概率等于 β 。这里，仍然必须处理上限可能为负值的问题。

在贝叶斯方法中，象以前一样， $L(\nu_s)$ 可以直接应用于贝叶斯定理。从 (9.53) 式求解 ν_s^{up} 一般来说也必须用数值方法。这个方法的优点是不需要估计量 $\hat{\nu}_s$ 的样本概率密度函数，另外一个优点前面已经提到过，即该方法自动将先验知识 $\nu_s \geq 0$ 结合到上限中。

贝叶斯极限与经典极限对比的深入讨论可以参见文献 [56, 57, 59]。将系统不确定度结合到极限中的方法可以参见文献 [60]。

第十章 特征函数

10.1 特征函数的定义与性质

假设随机变量 x 服从的概率密度函数为 $f(x)$ ，其特征函数(characteristic function) $\phi_x(k)$ 定义为 e^{ikx} 的期望值，即

$$\phi_x(k) = E[e^{ikx}] = \int_{-\infty}^{\infty} e^{ikx} f(x) dx. \quad (10.1)$$

这实际上就是概率密度函数的傅立叶变换。特征函数在一些重要定理的证明中非常有用，尤其是与随机变量之和有关的一些证明中。可以证明，特征函数与概率密度函数之间存在一一对应的关系，所以知道了特征函数的信息就知道了概率密度函数的信息，反之亦然。表10.1列出了一些重要概率密度函数的特征函数。

表 10.1: 一些常用概率密度函数的特征函数。

分布	概率密度函数	特征函数 $\phi_x(k)$
二项分布	$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$	$[p(e^{ik} - 1) + 1]^N$
泊松分布	$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}$	$\exp[\nu(e^{ik} - 1)]$
均匀分布	$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \leq x \leq \beta \\ 0 & \text{其它} \end{cases}$	$\frac{e^{i\beta k} - e^{i\alpha k}}{(\beta-\alpha)ik}$
指数分布	$f(x; \xi) = \frac{1}{\xi} e^{-x/\xi}$	$\frac{1}{1-ik\xi}$
高斯分布	$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\exp(i\mu k - \frac{1}{2}\sigma^2 k^2)$
卡方分布	$f(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$	$(1-2ik)^{-n/2}$
柯西分布	$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$	$e^{- k }$

假设有 n 个独立的随机变量 x_1, \dots, x_n ，分别服从概率密度函数 $f_1(x_1), \dots, f_n(x_n)$ ，对应的特征函数分别为 $\phi_1(k), \dots, \phi_n(k)$ 。考虑这 n 个随机变量之和 $z = \sum_i x_i$ ， z 的特征函数 $\phi_z(k)$ 与 x_i 的特征函数的关

系为

$$\begin{aligned}
 \phi_z(k) &= \int \cdots \int \exp\left(ik \sum_{i=1}^n x_i\right) f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n \\
 &= \int e^{ikx_1} f_1(x_1) dx_1 \cdots \int e^{ikx_n} f_n(x_n) dx_n \\
 &= \phi_1(k) \cdots \phi_n(k).
 \end{aligned} \tag{10.2}$$

也就是说，独立随机变量之和的特征函数等于每一个特征函数的乘积。

概率密度函数 $f(z)$ 可以通过反傅立叶变换得到

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_z(k) e^{-ikz} dk. \tag{10.3}$$

即使有时无法通过反傅里叶变换求得随机变量 z 的概率密度函数 $f(z)$ ，利用特征函数可以很容易确定 z 的各阶矩。对特征函数作 m 次微分得到

$$\left. \frac{d^m}{dk^m} \phi_z(k) \right|_{k=0} = \left. \frac{d^m}{dk^m} \int e^{ikz} f(z) dz \right|_{k=0} = i^m \int z^m f(z) dz = i^m \mu'_m \tag{10.4}$$

其中 $\mu'_m = E[z^m]$ 为 z 的 m 阶代数矩。

10.2 特征函数的应用

本节将通过证明一些结论说明特征函数的用处，这些结论在前面的章节中曾经提到过，但没有进行证明。作为第一个例子，可以用 (10.4) 式求第二章介绍的各种分布的均值、方差以及高阶矩。例如，高斯分布的均值和方差为

$$\begin{aligned}
 E[x] &= -i \left. \frac{d}{dk} [i\mu k - \frac{1}{2}\sigma^2 k^2] \right|_{k=0}, \\
 V[x] &= E[x^2] - (E[x])^2 = - \left. \frac{d^2}{dk^2} [\exp(i\mu k - \frac{1}{2}\sigma^2 k^2)] \right|_{k=0} - \mu^2 = \sigma^2.
 \end{aligned} \tag{10.5}$$

采用同样方法，可以求出表10.1中其它分布的各阶矩，除了柯西分布。柯西分布的特征函数 $\phi(k) = e^{-|k|}$ 在 $k=0$ 处不可导，所以均值和高阶矩都不存在，这在2.8节已经提到过。

利用特征函数，可以很容易研究概率分布在参数取极限时的行为。例如，对于二项分布，特征函数为

$$\phi(k) = [p(e^{ik} - 1) + 1]^N. \tag{10.6}$$

取极限 $p \rightarrow 0$, $N \rightarrow \infty$ ，并且要求 $\nu = pN$ 为常数，则特征函数变为

$$\phi(k) = \left(\frac{\nu}{N} (e^{ik} - 1) + 1 \right)^N \rightarrow \exp[\nu(e^{ik} - 1)], \tag{10.7}$$

这正是泊松分布的特征函数。

类似地，可以证明均值为 ν 的泊松变量 n ，在 $\nu \rightarrow \infty$ 的极限下变成高斯变量。尽管泊松变量是离散的，当 n 很大时，可以当作连续变量 x 处理，只要积分区间大于 1。注意到，泊松变量的方差等于均值 ν ，为了方便起见，作以下变量变换

$$x = \frac{n - \nu}{\sqrt{\nu}}, \tag{10.8}$$

显然 x 的均值为零，方差为 1，并且其特征函数为

$$\phi_x(k) = E[e^{ikx}] = E[e^{ikn/\sqrt{\nu}} e^{-ik\sqrt{\nu}}] = \phi_n\left(\frac{k}{\sqrt{\nu}}\right) e^{-ik\sqrt{\nu}}, \quad (10.9)$$

其中 ϕ_n 为泊松分布的特征函数。将表10.1中泊松变量的特征函数代入上式，对其中的指数部分进行展开，并取极限 $\nu \rightarrow \infty$ ，(10.9) 式变为

$$\phi_x(k) = \exp\left[\nu\left(e^{ik/\sqrt{\nu}} - 1\right) - ik\sqrt{\nu}\right] \rightarrow \exp\left(-\frac{1}{2}k^2\right). \quad (10.10)$$

这正是标准高斯分布的特征函数。再作变量变换，变换回原来的泊松变量 n ，可以得到，当 ν 很大时， n 服从均值和方差都为 ν 的高斯分布。

(10.2) 式的相加定理让我们可以证明很多重要的结论。例如，考虑两个高斯变量 x 和 y 的和，其均值分别为 μ_x 和 μ_y ，方差分别为 σ_x^2 和 σ_y^2 。根据 (10.2) 式， z 的特征函数与 x 和 y 的特征函数的关系为

$$\begin{aligned} \phi_z(k) &= \phi_x(k)\phi_y(k) = \exp(i\mu_x k - \frac{1}{2}\sigma_x^2 k^2) \cdot \exp(i\mu_y k - \frac{1}{2}\sigma_y^2 k^2) \\ &= \exp[i(\mu_x + \mu_y)k - \frac{1}{2}(\sigma_x^2 + \sigma_y^2)k^2]. \end{aligned} \quad (10.11)$$

这表明 z 是高斯随机变量，均值为 $\mu_z = \mu_x + \mu_y$ ，方差为 $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$ 。在9.8节中，我们利用了两个高斯随机变量之差的相应性质，而这里讨论的是两个高斯变量之和的性质。用相同的方法，可以证明，均值为 ν_i 的泊松变量之和是均值为 $\sum_i \nu_i$ 的泊松变量。

同样利用 (10.2) 式，可以证明，均值为 μ_i 方差为 σ_i^2 的 n 个独立的高斯随机变量，其平方和

$$z = \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} \quad (10.12)$$

服从自由度为 n 的卡方分布。要证明这个结论，首先考虑 $n = 1$ 的情形。通过变量变换，可以证明

$$y = \frac{x_i - \mu_i}{\sigma_i} \quad (10.13)$$

服从标准高斯概率密度函数 $\varphi(y) = (1/\sqrt{2\pi})e^{-y^2/2}$ ，这对所有的 i 都成立。因此 $z = y^2$ 服从

$$f(z; n=1) = 2\varphi(y) \left| \frac{dy}{dz} \right| = \frac{1}{\sqrt{2\pi z}} e^{-z/2}, \quad (10.14)$$

其中因子 2 是因为 y 为正和负都有贡献。与 (2.34) 式比较，可以发现这是 $n = 1$ 的 χ^2 分布。从表10.1知道， z 的特征函数为

$$\phi(k) = (1 - 2ik)^{-1/2}. \quad (10.15)$$

如果有 n 项求和，即 $z = \sum_{i=1}^n y^2$ ，则 z 的特征函数为 n 个 (10.15) 式的因子相乘，直接得到自由度数目为 n 的卡方分布的特征函数。

与此类似，可以证明中心极限定理，这个证明非常重要，值得在单独一节中讨论。下一节即给出中心极限定理的证明。

10.3 中心极限定理

假设有 n 个独立的随机变量 x_j ，均值和方差分别为 μ_j 和 σ_j^2 。它们的均值和方差不一定相等。中心极限定理表明，当 n 非常大时，随机变量之和 $\sum_j x_j$ 服从高斯分布，均值为 $\sum_j \mu_j$ ，方差为 $\sum_j \sigma_j^2$ 。这个定理成立的条件相当普遍，不论单个 x_j 服从何种分布。

为了方便起见，将变量减去 μ_j ，使变量的均值变为零。另外，先把 n 当作固定值，并定义

$$y_j = \frac{x_j - \mu_j}{\sqrt{n}}, \quad (10.16)$$

所以， $E[y_j] = 0$ ， $E[y_j^2] = \sigma_j^2/n$ 。考虑所有方差都相等的情况， $\sigma_j^2 = \sigma^2$ 。对 y_j 的特征函数 $\phi_j(k)$ 进行泰勒展开，

$$\phi_j(k) = \sum_{m=0}^{\infty} \frac{d^m \phi}{dk^m} \Big|_{k=0} \frac{k^m}{m!} = \sum_{m=0}^{\infty} \frac{(ik)^m}{m!} E[y_j^m] = 1 - \frac{k^2}{2n} \sigma^2 - \frac{ik^3}{3!} \frac{E[(x_j - \mu_j)^3]}{n^{3/2}} + \dots, \quad (10.17)$$

其中已经用到了 (10.4) 式将微分 $d^m \phi_j / dk^m$ 与矩 $E[y_j^m]$ 联系起来。利用 (10.2) 式，和式 $z = \sum_j y_j$ 的特征函数 $\phi_z(k)$ 为

$$\phi_z(k) = \prod_{j=1}^n \left(1 - \frac{k^2}{2n} \sigma^2 - \frac{ik^3}{3!} \frac{E[(x_j - \mu_j)^3]}{n^{3/2}} + \dots \right). \quad (10.18)$$

如果在 n 很大时可以忽略三阶和更高阶矩，则可以得到

$$\phi_z(k) \approx \left(1 - \frac{k^2}{2n} \sigma^2 \right)^n \rightarrow \exp \left(-\frac{1}{2} \sigma^2 k^2 \right). \quad (10.19)$$

这是均值为零，方差为 σ^2 的高斯分布的特征函数。变换回 $\sum_j x_j$ ，可以得到均值为 $\sum_j \mu_j$ ，方差为 $n\sigma^2$ 的高斯分布。当 σ_j 不等时，在某种更严格的条件下，该定理也成立，变量之和的方差为 $\sum_j \sigma_j^2$ 。

相比于指明什么条件下中心极限定理在 $n \rightarrow \infty$ 时严格成立 (参见文献 [2])，在实际的数据分析中，更重要的是了解 n 有限的情况下高斯近似的合理程度。这很难定量说明，但大致可以说，只要求和是由大量的小贡献构成，高斯近似基本成立。如果单项的分布具有很长的尾端，则有可能偶然取很大值，使其占求和的很大比重，这种情况下就会产生偏差。这些大数的贡献导致求和中出现“非高斯”的尾端，明显改变观测到远离均值事件的概率。

物理量的测量值服从高斯分布，中心值为真值，这个假设就是中心极限定理的一个常见的隐性应用。在构造置信区间时，经常用到这个假设，参见第九章。如果分布存在非高斯的尾端，置信区间可能被明显低估。特别是，置信水平和置信区间的大小将与高斯规则有明显差别，例如在 9.12 式中，68.3% 对应“ 1σ ”的区间，95.4% 对应“ 2σ ”的区间，等等。有时可以通过对求和中每个变量进行细致的蒙特卡罗模拟，获得对非高斯尾端的更好理解。

有时中心极限定理不成立，例如带电粒子穿过物质层时产生电子-离子对的总数目。对于给定厚度的物质层，产生的电子-离子对数目服从朗道分布，见 2.9 节。图 2.9 显示了粒子穿过 4 毫米厚的氩气层产生的电子-离子对数目分布。如果穿过 1 米长的气体容量，电离的总数目则是 25 个这样的变量之和。一般来说，如果将总容量看作分割成很多薄层，则总电离数为每一层的贡献之和，我们期望可以应用中心极限定理。然而，朗道分布具有很长的尾端，有相当的概率取比较大的值，所以，相对极少的强电离碰撞可以在总电离中占有非常显著的份额。这时候，在实际的系统中，高斯近似通常不再成立 (参见文献 [30])。

高斯近似失效的另一个例子是，带电粒子穿过物质层时的散射角。粒子穿过物质时，与其中的原子核发生碰撞，每次碰撞都会被散射一个小角度，这些角度之和被认为是总的散射角 (多重散射)。由于粒子穿过时会发生很多次碰撞，所以我们期望总的散射角服从高斯分布。然而，单次碰撞的散射角分布在大角度处有很长的尾端。对于有限的厚度，少数大角度的碰撞在总散射角中占有显著的比例，这导致最终的散射角不服从高斯分布。

10.4 用特征函数求估计量的概率密度函数

考虑对随机变量 x 的 n 次独立观测, x 服从指数分布 $f(x; \xi) = (1/\xi) \exp(-x/\xi)$ 。在6.2节, 已经看到参数 ξ 的极大似然估计量 $\hat{\xi}$ 为观测值 x_i 的样本均值:

$$\hat{\xi} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (10.20)$$

如果实验重复多次, 可以得到很多 $\hat{\xi}$ 值, 服从概率密度函数 $g(\hat{\xi}; n, \xi)$, 这个概率密度函数依赖于每次实验观测的次数 n 以及参数 ξ 的真值。

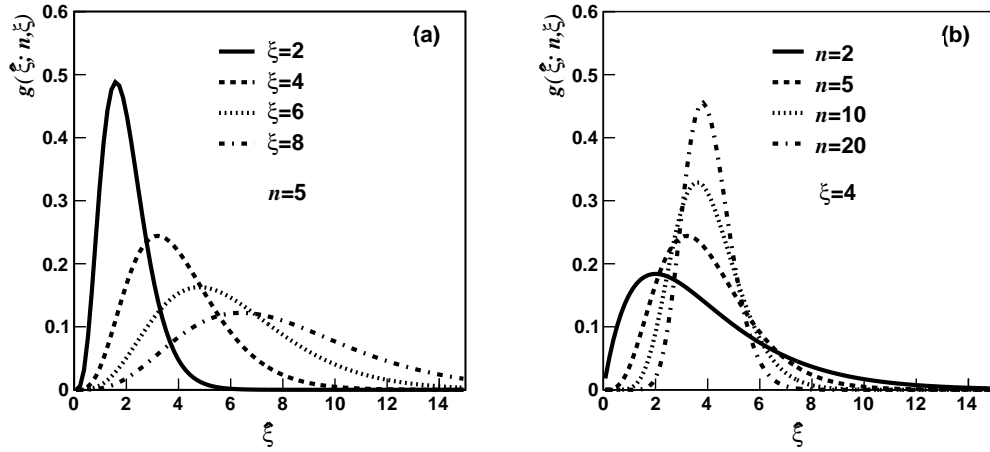


图 10.1: 估计量 $\hat{\xi}$ 的样本概率密度函数 $g(\hat{\xi}; n, \xi)$ 。(a) 每次实验测量次数 $n=5$, 取不同的参数真值 ξ 。(b) 参数真值为 $\xi=4$, 每次实验的测量次数 n 不同。

假设我们想求概率密度函数 $g(\hat{\xi}; n, \xi)$ 。随机变量 x 的特征函数为

$$\phi_x(k) = \int_{-\infty}^{\infty} e^{ikx} f(x) dx = \int_0^{\infty} e^{ikx} \frac{1}{\xi} e^{-x/\xi} dx = \frac{1}{1 - ik\xi}. \quad (10.21)$$

将 (10.2) 式应用于求和 $z = \sum_{i=1}^n x_i = n\hat{\xi}$, 可以得到

$$\phi_z(k) = \frac{1}{(1 - ik\xi)^n}. \quad (10.22)$$

随机变量 z 的概率密度函数 $g_z(z)$ 可以通过计算 $\phi_z(k)$ 的反傅立叶变换得到

$$g_z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ikz}}{(1 - ik\xi)^n} dk. \quad (10.23)$$

被积函数在 k 复平面的 $-i/\xi$ 处为 n 阶奇点。选取下半平面的围道积分, 利用留数定理可以得到

$$g_z(z) = \frac{1}{(n-1)!} \frac{z^{n-1}}{\xi^n} e^{-z/\xi}. \quad (10.24)$$

通过变量变换, 可以得到估计量 $\hat{\xi} = z/n$ 的概率密度函数为

$$g(\hat{\xi}; n, \xi) = g_z(z) \left| \frac{dz}{d\hat{\xi}} \right| = n g_z(n\hat{\xi}) = \frac{n^n}{(n-1)!} \frac{\hat{\xi}^{n-1}}{\xi^n} e^{-n\hat{\xi}/\xi}, \quad (10.25)$$

这是伽马分布(参见文献 [23] 第 4 章)的特殊形式。图10.1给出了不同参数组合下 $g(\hat{\xi}; n, \xi)$ 的形状。每次实验的测量次数为 $n = 5$ 时, 概率密度函数大致以真值 ξ 为中心, 但在 $\hat{\xi}$ 比较大的地方有长尾端。从图10.1(b) 可以看出, 随着测量次数 n 的增加, 概率密度函数如中心极限定理所要求, 变得接近高斯分布。

现在, 可以利用已知估计量概率密度函数的优势, 解决更多的问题。例如, 在10.4.1小节, 我们用 (10.25) 式的概率密度函数计算期望值, 在10.4.2小节中, 用这个概率密度函数构造置信区间。

10.4.1 平均寿命和衰变常数的期望值

先转换到粒子寿命的惯用符号表示, 利用对衰变时间的 n 次测量结果构造估计量 $\hat{\tau} = (1/n) \sum_{i=1}^n t_i$, 用来估计粒子的平均寿命 τ , (10.25) 式给出该估计量的概率密度函数。在6.2节中, 我们已经利用下面的公式计算估计量 $\hat{\tau}$ 的期望值

$$E[\hat{\tau}(t_1, \dots, t_n)] = \int \cdots \int_0^\infty \left(\frac{1}{n} \sum_{i=1}^n t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \cdots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \cdots dt_n = \tau. \quad (10.26)$$

这个结果也可以直接用 $\hat{\tau}$ 的概率密度函数 (见 (10.25) 式) 得到,

$$E[\hat{\tau}] = \int_0^\infty \hat{\tau} g(\hat{\tau}; n, \tau) d\hat{\tau} = \int_0^\infty \hat{\tau} \frac{n^n}{(n-1)!} \frac{\hat{\tau}^{n-1}}{\tau^n} e^{-n\hat{\tau}/\tau} d\hat{\tau} = \tau. \quad (10.27)$$

6.2节还证明了, 参数的函数的极大似然估计量等于同样函数形式下该参数极大似然估计量的函数。例如, 衰变常数 $\lambda = 1/\tau$ 的极大似然估计量为 $\hat{\lambda} = 1/\hat{\tau}$ 。利用 $g(\hat{\tau}; n, \tau)$, 可以计算出概率密度函数 $h(\hat{\lambda})$,

$$h(\hat{\lambda}; n, \lambda) = g(\hat{\tau}; n, \tau) \left| d\hat{\tau}/d\hat{\lambda} \right| = \frac{n^n}{(n-1)!} \frac{\lambda^n}{\hat{\lambda}^{n+1}} e^{-n\hat{\lambda}/\hat{\lambda}}. \quad (10.28)$$

估计量 $\hat{\lambda}$ 的期望值为

$$E[\hat{\lambda}] = \int_0^\infty \hat{\lambda} h(\hat{\lambda}; n, \lambda) d\hat{\lambda} = \int_0^\infty \frac{n^n}{(n-1)!} \frac{\lambda^n}{\hat{\lambda}^n} e^{-n\lambda/\hat{\lambda}} d\hat{\lambda} = \frac{n}{n-1} \lambda. \quad (10.29)$$

可以看到, 尽管 τ 的极大似然估计量 $\hat{\tau} = (1/n) \sum_{i=1}^n t_i$ 为无偏估计量, 但 $\hat{\lambda} = 1/\hat{\tau}$ 不是 $\lambda = 1/\tau$ 的无偏估计量。然而, 当 n 趋于无穷时, 偏倚变为零。

10.4.2 指数随机变量均值的置信区间

可以用 (10.25) 式的概率密度函数 $g(\hat{\xi}; n, \xi)$, 根据9.2节的程序确定置信区间。假设对指数随机变量 x 进行 n 次观测, 用以计算参数 ξ 的估计量 $\hat{\xi}$, 并且得到的估计值为 $\hat{\xi}_{\text{obs}}$ 。我们的目标是, 用给定数据样本 x_1, \dots, x_n 确定区间 $[a, b]$, 使得不论真值 ξ 为多少, 只要给定 α 和 β , 概率 $P[a < \xi] = \alpha$ 和 $P[\xi < b] = \beta$ 成立。

置信区间可以利用 (9.9) 式求解 a 和 b 得到, 即

$$\begin{aligned} \alpha &= \int_{\hat{\xi}_{\text{obs}}}^\infty g(\hat{\xi}; a) d\hat{\xi}, \\ \beta &= \int_{\hat{\xi}_{\text{obs}}}^\infty g(\hat{\xi}; b) d\hat{\xi}. \end{aligned} \quad (10.30)$$

图10.2给出了置信区间随 n 的变化关系, 假设观测值为 $\hat{\xi}_{\text{obs}} = 1$ 。图中还给出了观测值加减一个标准差估计值得到的区间。 n 很大时, 概率密度函数 $g(\hat{\xi}; n, \xi)$ 变成高斯分布 (中心极限定理的要求), 此时 68.3% 的中心置信区间趋向于 $[\hat{\xi}_{\text{obs}} - \hat{\sigma}_{\hat{\xi}}, \hat{\xi}_{\text{obs}} + \hat{\sigma}_{\hat{\xi}}]$ 。

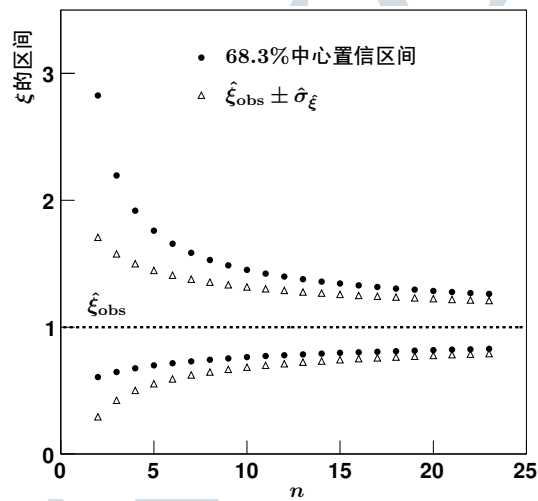


图 10.2: 假设观测值 $\hat{\xi}_{\text{obs}} = 1$, 观测次数 n 取不同值时, 指数分布参数 ξ 的经典置信区间 (实心圆点之间), 以及区间 $[\hat{\xi}_{\text{obs}} - \hat{\sigma}_{\xi}, \hat{\xi}_{\text{obs}} + \hat{\sigma}_{\xi}]$ (空心三角之间)。

DRAFT ZV

第十一章 解谱法

到目前为止，我们已经考虑了很多随机变量，例如粒子能量、衰变时间等等，并且通常假设这些量的测量值没有误差。本章将考虑测量仪器有限分辨率的影响。由于有限的分辨率，测量值将会有附加的随机涨落，从而使分布变形。对这些变形进行修正的程序称为**解谱 (unfolding)**。解谱与更普遍的题目**逆问题 (inverse problem)**在数学上是相同的，这一类的问题也称为反卷积 (deconvolution) 或去模糊化 (unsmearing)。尽管这里的介绍基本局限于粒子物理领域，但这些概念在很多领域中得到发展和应用，例如光学图像重建、射线天文学、晶体成像学和医学成像等。

这里介绍的方法基本都是经典统计的概念，在很多文献中都有所介绍 [62–66]。然而，也有一些方法采用贝叶斯统计，参见文献 [18, 67–69]。

在11.1节，将给出解谱问题的数学描述以及符号定义。在11.2节，将讨论通过对响应矩阵求逆进行解谱。响应矩阵求逆的解谱法很少会在实际中应用，但这是寻求更优解的起点。11.3节将介绍基于修正因子的简单方法。而本章的主要内容正规化解谱将在11.4节到11.7节进行详细讨论，包括求解的策略，几种不同正规化函数的概要介绍，以及对结果的方差和偏倚进行估计的方法。11.8节中，通过不同的例子对这些要点进行了说明。11.9节给出了如何数值实现这些方法。

需要强调的是，很多问题并不需要对测量得到的分布进行解谱，尤其当我们的目的只是将结果与已知理论预言进行比较的时候。这种情况下，只需要对理论预言进行修正，将探测器对分布造成的变形考虑进去，然后用得到的新预言直接与测量结果进行比较。相比于对测量值进行解谱并与原始 (未修正的) 理论分布进行比较，这个过程通常更为简单。

然而，如果不进行解谱，测量结果将无法与其它实验进行比较，因为不同实验的分辨率效应通常不同。也存在一种情况，即实验测量之后多年又发展出新的理论，此时已经无法找到用来修正理论的分辨率效应，即响应函数或响应矩阵 (见后面内容)。如果特别重要的测量分布想要保持其价值，则需要同时保存测量结果和响应矩阵的信息。然而不幸的是，这通常不可行，也很少这么做。

通过对分布进行解谱，提供的结果可以直接与其它实验结果进行比较，也可以与各种理论预言进行比较。在某些应用中，必须进行解谱处理，例如图像重建。在图像重建的问题中，如果不对分布进行修正，图像的某些特征可能无法识别。本章，我们假设已经考虑过这些问题，并决定进行解谱。

11.1 解谱问题的数学描述

考虑某个随机变量 x ，我们希望确定其概率密度函数。本章中，我们将允许对 x 的测量具有有限分辨率，探测器的效率也可能小于 100%，并且可能存在本底过程。作为例子，考虑放射性核素 β 衰变产生的电子能量分布，即随机变量 x 指的是发射电子的能量。

所谓“有限分辨率”指的是，由于测量误差， x 的测量值可能与实际产生时的真值有所差异，并且这种差异是随机的。例如，某次 β 衰变产生的电子能量是确定的，但由于测量仪器的分辨率影响，记录下来的测量值通常会稍有不同。因此，每次测量都由两个量描述：真值 y (可能是未知的) 以及测量值 x 。

一般来说，还必须考虑这种情况：真值存在，但却没有得到任何测量值。在 β 衰变的例子中，这种情况对应于发射电子完全没有被探测到就跑掉了，因为探测器可能没有覆盖包围放射源的整个立体角，或者电子能量低于某个阈值而无法产生可被探测的足够强的信号。事例被测量到的概率称为探测器的**效率 (efficiency)**¹ ϵ_y 。一般来说，效率依赖于事例的真值 y 。

假设真值服从的概率密度函数为 $f_{\text{true}}(y)$ 。为了构造 $f_{\text{true}}(y)$ 可用的估计量，该估计量的参数数目必须有限。如果无法预先知道 $f_{\text{true}}(y)$ 的函数形式，仍然可以将其表示为分 M 个区间的归一化直方图。在第 j 个区间发现 y 的概率就是对该区间的积分，

$$p_j = \int_{\text{区间 } j} f_{\text{true}}(y) dy. \quad (11.1)$$

假设我们实施一次实验，实验中发生的事例总数为 m_{tot} ，观测到的事例数通常与 m_{tot} 不同。可以将 m_{tot} 当作固定值，也可以当作随机变量。不管哪种情况，我们都称总事例数的期望值为 $\mu_{\text{tot}} = E[m_{\text{tot}}]$ ，所以，区间 j 里事例数的期望值为

$$\mu_j = \mu_{\text{tot}} p_j. \quad (11.2)$$

我们将称矢量 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ 为“真实直方图”。需要注意的是，这不是各个区间内实际的事例数，而是对应的期望值，即 μ_i 一般来说不是整数。可以将区间 i 里的真实事例数当作均值为 μ_i 的随机变量 m_i 。然而，由于有限分辨率和效率的影响， m_i 无法直接观测，它甚至不会出现在解谱问题的数学公式中。相反，我们将直接构造参数 μ_i 的估计量。

为方便起见，通常也构造一个观测值的直方图。假设从 x 的观测样本出发，并且观测值被填充至分成 N 个区间的直方图，用矢量表示为 $\mathbf{n} = (n_1, \dots, n_N)$ 。这些值也可以是样本矩、傅里叶系数等等。实际上，变量 x 也可以是多维的，不但包含真正感兴趣的物理量 y 的测量值，还包含可以为 y 提供附加信息的相关量。

一般来说，区间数目 N 大于、小于或等于真实直方图区间数 M 都有可能。假设观测直方图第 i 个区间包含 n_i 个事例，并且总事例数为 $\sum_i n_i = n_{\text{tot}}$ 。通常可以将变量 n_i 当作期望值为 ν_i 的独立泊松变量。也就是说，在这个假设下，在区间 i 观测到 n_i 个事例的概率为

$$P(n_i; \nu_i) = \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}. \quad (11.3)$$

由于泊松变量之和仍然是泊松变量 (参见10.4节)，所以 n_{tot} 服从期望值为 $\nu_{\text{tot}} = \sum_i \nu_i$ 的泊松变量。也可能考虑 n_{tot} 为固定参数的情形，则 n_i 服从多项分布。不管是哪种分布，我们都说期望值为

$$\mu_i = E[n_i]. \quad (11.4)$$

要构造似然函数，需要知道数据 $\mathbf{n} = (n_1, \dots, n_N)$ 的概率分布 (泊松分布、多项分布等等)，基于极大似然法的解谱法需要用到这个构造似然函数。另一种可能是，给出的是协方差矩阵

$$V_{ij} = \text{cov}[n_i, n_j], \quad (11.5)$$

基于最小二乘法的解谱法需要用到协方差矩阵。假设要么已知测量数据的概率分布形式，要么已知其协方差矩阵。

¹如果事例没有被探测到的原因与几何有关，例如探测器的立体角有限，这时的效率通常称为接受度。有时，效率一词用来表示已知事例在探测器灵敏区域内该事例被探测到的概率。这里，我们采用的更广义上的效率，表示事例被探测到的总概率。

利用全概率定律(1.27) 式，可以将期望值 $\nu_i = E[n_i]$ 表示为

$$\begin{aligned}\nu_i &= \mu_{\text{tot}} P(\text{事例在区间} i \text{被观测到}) \\ &= \mu_{\text{tot}} \int dy P(\text{在区间} i \text{被观测到} | \text{真值为} y \text{且被探测到}) \epsilon(y) f_{\text{true}}(y) \\ &= \mu_{\text{tot}} \int_{\text{区间} i} dx \int dy s(x|y) \epsilon(y) f_{\text{true}}(y)\end{aligned}\quad (11.6)$$

其中 $s(x|y)$ 为给定真值为 y 且事例在某处被观测到的前提下，观测值 x 的条件概率密度函数，即 $s(x|y)$ 是归一化的， $\int s(x|y) dx = 1$ 。 s 被称为**分辨率函数(resolution function)**，在成像应用领域中也被称为**点扩散函数(point spread function)**。也可以定义**响应函数(response function)**，

$$r(x|y) = s(x|y) \epsilon(y). \quad (11.7)$$

响应函数给出已知真值为 y 的条件下，观测值为 x 的概率，其中考虑了有限效率的影响。需要注意的是，响应函数没有按照 x 的条件概率归一化。我们称真实分布被响应函数**包藏 (folded)** 起来，因此称估计 f_{true} 的任务为**解谱(unfolding)**。

将 (11.6) 式中对 y 的积分分成对所有区间内积分的求和，并将分子与分母同乘以 μ_j ，区间 i 中观测事例数的期望值为

$$\begin{aligned}\nu_i &= \sum_{j=1}^M \frac{\int_{\text{区间} i} dx \int_{\text{区间} j} dy s(x|y) \epsilon(y) f_{\text{true}}(y)}{(\mu_j / \mu_{\text{tot}})} \mu_j \\ &= \sum_{j=1}^M R_{ij} \mu_j.\end{aligned}\quad (11.8)$$

其中**响应矩阵(response matrix)** R 为

$$\begin{aligned}R_{ij} &= \frac{\int_{\text{区间} i} dx \int_{\text{区间} j} dy s(x|y) \epsilon(y) f_{\text{true}}(y)}{\int_{\text{区间} j} dy f_{\text{true}}(y)} \\ &= \frac{P(\text{真值在区间} j \text{观测值在区间} i)}{P(\text{真值在区间} j)} \\ &= P(\text{观测值在区间} i | \text{真值在区间} j).\end{aligned}\quad (11.9)$$

因此，响应矩阵的元素 R_{ij} 为已知真值 y 在区间 j 但测量值 x 在区间 i 的条件概率。响应矩阵中的非对角元素将所有精细结构模糊化。真实直方图中主要集中在一个区间的峰，在观测直方图中将分布在几个区间内。真实直方图中相距只有几个区间的两个峰，将被合并成一个宽峰。

从 (11.9) 式的第一行可以看出，响应矩阵依赖于概率密度函数 $f_{\text{true}}(y)$ 。然而，这个概率密度函数先验未知，因为我们的目标就是确定 $f_{\text{true}}(y)$ 。如果被解谱的直方图的区间宽度非常小，以至于在同一个区间内 $s(x|y)$ 和 $\epsilon(y)$ 都可以近似看成常数，与 y 无关，则 R_{ij} 对 $f_{\text{true}}(y)$ 的直接依赖被抵消。在后面的讨论中，我们将假定这个近似成立，并且响应矩阵中由于 $f_{\text{true}}(y)$ 的不确定性导致的误差可以忽略。实际上，在进行实验之前，我们将利用可以很好地近似 $f_{\text{true}}(y)$ 的任何可用分布确定响应矩阵。

尽管 $s(x|y)$ 和 $\epsilon(y)$ 的构造要求与给定值 y 发生的概率无关 (即，与 $f_{\text{true}}(y)$ 无关)，但通常来说不能做到完全模型无关。变量 y 可能不是唯一影响测量值 x 概率的量。在 β 衰变的例子中， y 和 x 分别表示发射电子能量的真值和测量值， $s(x|y)$ 和 $\epsilon(y)$ 一般来说与电子的角分布有关 (探测器的某些部位可能比其它地方分辨率更好)，而不同的 β 衰变模型可能会得到不同的角分布预言。

后面的讨论中，我们将忽略这种模型依赖，并假定分辨率函数 $s(x|y)$ 和效率 $\epsilon(y)$ 只依赖于测量仪器，因此，响应矩阵 R_{ij} 也只与测量仪器有关。实际上，我们将假设响应矩阵的确定可以忽略不确定性，无论

从模型依赖的角度，还是从探测器响应的角度。实际运用中， R 可以通过真值 y 先验已知的刻度实验确定，也可以通过蒙特卡罗模拟确定，前提是对探测器中发生的各种物理过程的正确理解。在实际问题中，模型依赖可能不能忽略，并且对探测器的理解本身不可能完美，这些都必须当作系统误差的可能来源进行处理。

注意到，响应矩阵 R_{ij} 一般来说不是对角的（甚至不一定是方阵），第一个指标 $i = 1, \dots, N$ 表示观测直方图的区间，第二个指标 $j = 1, \dots, M$ 表示真实直方图的区间。对第一个指标求和，并利用 $\int s(x|y)dx = 1$ ，可以得到

$$\begin{aligned}\sum_{i=1}^N R_{ij} &= \sum_{i=1}^N \frac{\int_{\text{区间 } i} dx \int_{\text{区间 } j} dy s(x|y) \epsilon(y) f_{\text{true}}(y)}{(\mu_j / \mu_{\text{tot}})} \\ &= \frac{\int_{\text{区间 } i} dy \epsilon(y) f_{\text{true}}(y)}{\int_{\text{区间 } j} f_{\text{true}}(y) dy} \\ &\equiv \epsilon_j,\end{aligned}\tag{11.10}$$

即，得到的是区间 j 中的效率平均值。

除了有限分辨率和效率，还必须考虑一种可能性，即所研究类型的真实事例没有发生但测量仪器得到了某个测量值，也就是说，测量值是由某种**本底 (background)**过程产生的。在 β 衰变的例子中，本底可能来源于探测器中的乱真信号，样本中存在其它类型的放射性核素，宇宙线等外部粒子与探测器的相互作用，等等。假设本底过程导致在区间 i 可以观测到的事例数的期望值为 β_i ，则 (11.8) 式修改为

$$\nu_i = \sum_{j=1}^M R_{ij} \mu_j + \beta_i.\tag{11.11}$$

注意， β_i 已经包含了探测器的有限分辨率和效率等效应。分辨率和效率通常通过刻度实验或蒙特卡罗模拟确定，模拟既包括本底过程也包括探测器响应。我们将假定 β_i 的值已知，尽管实际上只能在一定程度上确定本底的值。因此，本底的不确定性是解谱结果的系统误差来源之一。

总结一下，我们介绍了以下一些量（也统称为直方图或分布）：

1. 真实直方图（每个区间真实事例数的期望值）， $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ ，
2. 归一化的真实直方图（概率）， $\mathbf{p} = (p_1, \dots, p_M) = \boldsymbol{\mu} / \mu_{\text{tot}}$ ，
3. 观测事例数的期望值， $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$ ，
4. 实际观测的事例数（数据）， $\mathbf{n} = (n_1, \dots, n_N)$ ，
5. 效率 $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ ，
6. 本底期望值 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)$ 。

假设要么已知数据 \mathbf{n} 的概率分布形式，要么已知协方差矩阵 $V_{ij} = \text{cov}[n_i, n_j]$ 。概率密度分布的形式用于构造似然函数，协方差矩阵用来构造 χ^2 函数。另外，响应矩阵为 R_{ij} ，其中 $i = 1, \dots, N$ 表示观测直方图的区间， $j = 1, \dots, M$ 表示真实直方图的区间。假定 R 和 $\boldsymbol{\beta}$ 都已知。矢量 $\boldsymbol{\mu}$ 、 $\boldsymbol{\nu}$ 、 $\boldsymbol{\beta}$ 以及响应矩阵 R 的关系为

$$\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta},\tag{11.12}$$

其中 $\boldsymbol{\mu}$ 、 $\boldsymbol{\nu}$ 和 $\boldsymbol{\beta}$ 在矩阵方程中应理解为列矢量。我们的目标是构造真实直方图的估计量 $\hat{\boldsymbol{\mu}}$ ，或者概率的估计量 $\hat{\mathbf{p}}$ 。

11.2 响应矩阵求逆

本节将考察一种构造真实直方图 μ 的估计量的方法。这种方法看起来很自然，但得出的结果通常不可接受。考虑真实直方图与观测直方图区间数相等的情况，即 $M = N$ 。现在，假设可以对矩阵关系 $\nu = R\mu + \beta$ 求逆，得到

$$\mu = R^{-1}(\nu - \beta). \quad (11.13)$$

ν 的估计量的一种自然选择是对应的数据，即

$$\hat{\nu} = \mathbf{n} \quad (11.14)$$

则 μ 的估计量为

$$\hat{\mu} = R^{-1}(\mathbf{n} - \beta). \quad (11.15)$$

实际上很容易证明，这正是最大化对数似然函数的结果，

$$\log L(\mu) = \sum_{i=1}^N \log P(n_i; \nu_i), \quad (11.16)$$

其中 $P(n_i; \nu_i)$ 为泊松分布或二项分布。这也是最小二乘解，即最小化

$$\chi^2(\mu) = \sum_{i,j=1}^N (\nu_i - n_i)(V^{-1})_{ij}(\nu_j - n_j) \quad (11.17)$$

的结果。注意到，由于关系式 $\nu = R\mu + \beta$ 始终成立，所以 $\log L(\mu)$ 和 $\chi^2(\mu)$ 可以写成 μ 或 ν 的函数。也就是说，在 (11.16) 式或 (11.17) 式对 μ_i 求微分时，用到了 $\partial \nu_i / \partial \mu_j = R_{ij}$ 。

在证明这样构造的估计量为什么失败之前，有必要先研究一下其偏倚和方差。 $\hat{\mu}_j$ 的期望值为

$$\begin{aligned} E[\hat{\mu}_j] &= \sum_{i=1}^N (R^{-1})_{ji} E[n_i - \beta_i] = \sum_{i=1}^N (R^{-1})_{jk} (\nu_i - \beta_i) \\ &= \mu_j, \end{aligned} \quad (11.18)$$

所以估计量 $\hat{\mu}_j$ 是无偏估计量，因为根据假设 $\hat{\nu}_i = n_i$ 是无偏的。对于估计量 $\hat{\mu}$ 的协方差矩阵，可以得到

$$\begin{aligned} \text{cov}[\hat{\mu}_i, \hat{\mu}_j] &= \sum_{k,l=1}^N (R^{-1})_{ik} (R^{-1})_{jl} \text{cov}[n_k, n_l] \\ &= \sum_{k=1}^N (R^{-1})_{ik} (R^{-1})_{jk} \nu_k, \end{aligned} \quad (11.19)$$

其中最后一行用到了独立泊松变量的协方差矩阵 $\text{cov}[n_k, n_l] = \delta_{kl} \nu_k$ 。

后面我们将采用符号 $V_{ij} = \text{cov}[n_i, n_j]$ 表示数据的协方差矩阵，用 $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$ 表示真实分布估计量的协方差矩阵。于是，(11.19) 式可以写成矩阵形式

$$U = R^{-1} V (R^{-1})^T. \quad (11.20)$$

考虑图11.1中的例子。原始的真实分布如图11.1(a)所示，观测分布的期望值 ν 如图11.1(b)的直方图所示。

直方图 ν 是通过 $\nu = R\mu$ 计算得到的，即本底 β 取为零。响应矩阵 R 基于高斯分辨率函数和效率 ϵ_i ，其中高斯分辨率函数的标准差等于 1.5 倍的区间宽度，效率 ϵ_i 处处为 1。这个响应矩阵造成的结果是，事

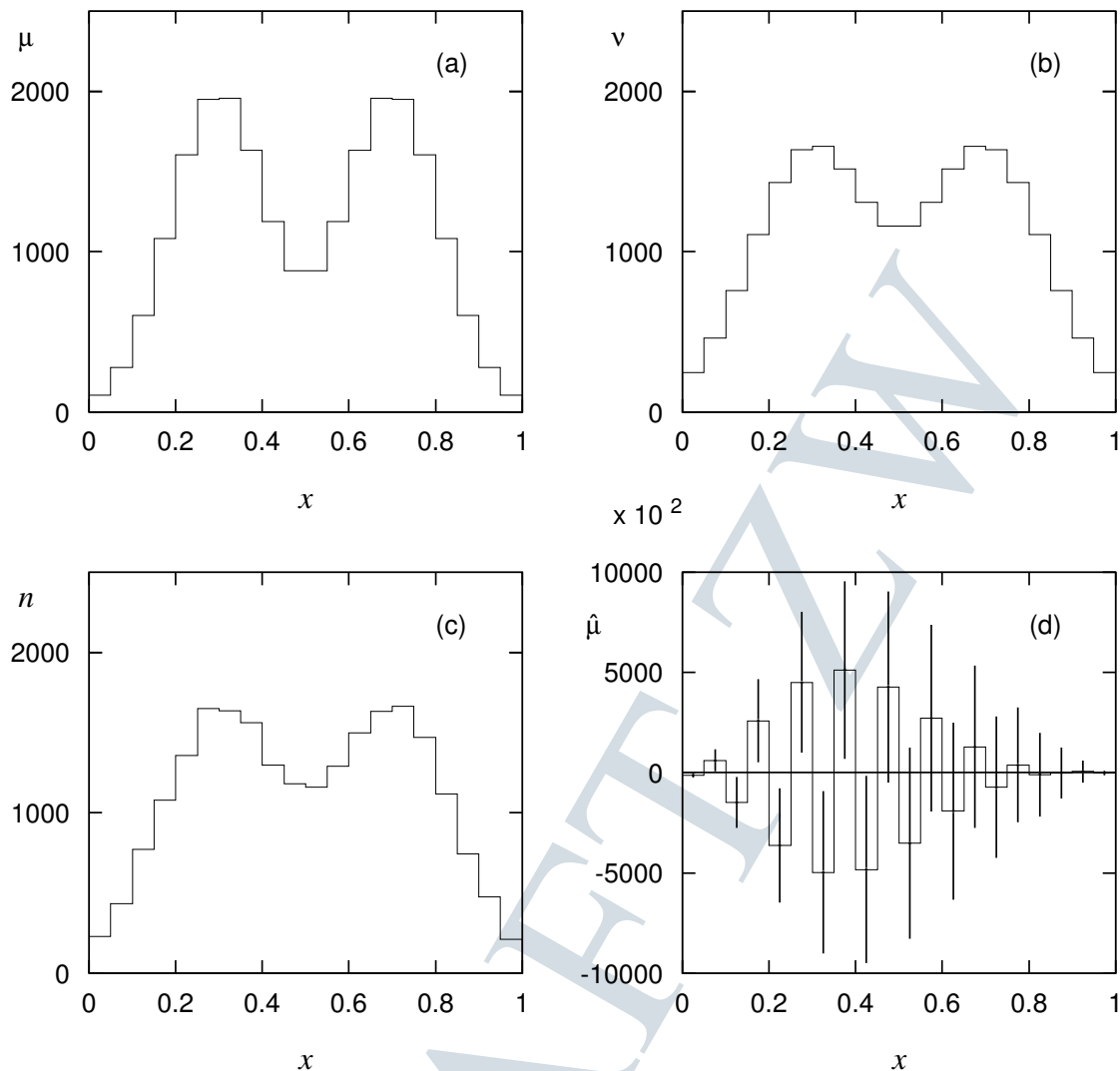


图 11.1: (a) 假定的真实直方图 μ ; (b) 期望值 $\nu = R\mu$ 的直方图; (c) 观测数据 \mathbf{n} 的直方图; (d) 利用响应矩阵的逆求得的估计量 $\hat{\mu}$ 。

例保留在产生的区间内的概率是 26%，42% 的概率在左右相邻的区间被观测到，32% 的概率出现在其它区间内。

图 11.1(c) 给出数据 $\mathbf{n} = (n_1, \dots, n_N)$ 。该数据是根据泊松分布利用蒙特卡罗方法产生，泊松分布的均值 ν_i 为图 11.1(b) 各个区间的值。由于每个区间事例数的范围大约为 10^2 到 10^3 ， n_i 的相对误差(标准差与均值的比率)在 3 – 10% 的范围内。

图 11.1(d) 给出利用 (11.15) 式矩阵求逆得到的估计值 $\hat{\mu}$ 。误差棒表示每个区间的标准差。 $\hat{\nu}_j$ 的精度比 n_i 达到的精度 3 – 10% 差很多， $\hat{\nu}_j$ 的值随区间涨落非常厉害，误差棒的大小与估计值本身差不多大。(要注意图 (d) 的纵坐标标度比其它几个图要大。) 相邻区间的相关系数接近 -1 ，即几乎完全负相关。

出现这种灾难性失败的原因是，我们不知道期望值 ν 。如果 ν 已知，则可以直接计算 $\mu = R^{-1}\nu$ 。然而，我们只有数据 \mathbf{n} ，而 \mathbf{n} 为随机变量，存在随机涨落。之前曾经提到过，响应矩阵将所有的精细结构都模糊化。如果 μ 中的峰非常接近，那么在 ν 中尽管它们将被合并到一起，其中仍然会保留精细结构的一些残留信息。将 R^{-1} 作用到 ν 上，残留的信息将把原来的精细结构完全恢复。数据 \mathbf{n} 实际上在不同区间都存在统计涨落，这种涨落导致的结果定性地讲与 ν 中精细结构的残留是一样的。也就是说，解谱后的结果

给出了大量精细结构，如图11.1(d) 显示。

有趣的是，将 (11.19) 式的协方差矩阵 U 与 RCF 不等式 (参见6.6节) 给出的方差进行比较。RCF 不等式给出所有可能估计量的最小方差。为此，我们将 n_i 看作均值为 ν_i 的泊松变量。因此对数似然函数为

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^N \log P(n_i; \nu_i) = \sum_{i=1}^N \log \left(\frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!} \right). \quad (11.21)$$

省略掉与 $\boldsymbol{\mu}$ 无关的相加项，可以得到

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^N (n_i \log \nu_i - \nu_i). \quad (11.22)$$

令 $\log L$ 对 $\boldsymbol{\mu}$ 的各个分量的导数都为零，

$$\frac{\partial \log L}{\partial \mu_k} = \sum_{i=1}^N \frac{\partial \log L}{\partial \nu_i} \frac{\partial \nu_i}{\partial \mu_k} = \sum_{i=1}^N \left(\frac{n_i}{\nu_i} - 1 \right) R_{ik} = 0, \quad (11.23)$$

实际上得到 $\boldsymbol{\nu}$ 的估计量就是前面用到的 $\hat{\boldsymbol{\nu}} = \mathbf{n}$ 。对 $\log L$ 再求一次导数得到，

$$\frac{\partial^2 \log L}{\partial \mu_k \partial \mu_l} = - \sum_{i=1}^N \frac{n_i R_{ik} R_{il}}{\nu_i^2}. \quad (11.24)$$

在零偏倚的情况下，RCF 边界给出的协方差矩阵的逆 (见 (6.19) 式) 则为

$$\begin{aligned} (U^{-1})_{kl} &= -E \left[\frac{\partial^2 \log L}{\partial \mu_k \partial \mu_l} \right] \\ &= \sum_{i=1}^N \frac{E[n_i] R_{ik} R_{il}}{\nu_i^2} \\ &= \sum_{i=1}^N \frac{R_{ik} R_{il}}{\nu_i}. \end{aligned} \quad (11.25)$$

对该方程两边同乘 U 一次，同乘 R^{-1} 两次，然后对适当的指标求和，可以从 (11.25) 式解出协方差矩阵的 RCF 边界，

$$U_{ij} = \sum_{k=1}^N (R^{-1})_{ik} (R^{-1})_{jk} \nu_k. \quad (11.26)$$

这与 (11.19) 中的严格结果完全相同。所以，极大似然解既是无偏的也是有效的，也就是说，极大似然解在所有无偏估计量中的方差最小。利用最小二乘法，可以得到同样的结果，这种情况下，高斯-马可夫 (Gauss-Markov) 定理保证了估计量的无偏性和有效性。

尽管图11.1(d) 中的解与真实分布几乎没有任何相似之处，这个解却有很多理想的性质。例如，估计量很容易构造，偏倚为零，方差等于 RCF 边界。然而，要想使用这个估计量，必须考虑相关性的行为。例如，可以通过构造 χ^2 统计量，检验估计量 $\hat{\boldsymbol{\mu}}$ 与 $\boldsymbol{\mu}_0$ 假设之间的一致性，这个卡方统计量为

$$\chi^2 = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T U^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0), \quad (11.27)$$

其中用到了估计量的完全协方差矩阵 U 。如果只是利用 U 的对角元素计算 χ^2 ，这个检验将毫无意义。我们还需要注意的是，尽管这里的例子的方差非常大，但是，如果区间宽度比分辨率函数的宽度还大，则方差将被大大减小。

尽管矩阵求逆法存在很大的缺点，这个方法给出了非常重要的经验，并为其它方法提供了出发点。由于矩阵求逆法的解没有偏倚，并且方差等于 RCF 不等式给出的最小方差，所以，如果想减小方差，必然要

引入一定的偏倚。解谱法的关键就在于如何构造有偏估计量 $\hat{\mu}$ ，使其能够保证，只要我们的先验观念（通常是与光滑性有关的一些假设）是正确的，则这个偏倚就很小。大致说来，其目标是在偏倚和方差之间寻求最佳平衡点，尽管在11.7节将看到，如何确定这个最优化存在一定的任意性。

如果需要与先验知识结合，建议使用贝叶斯方法，将验前概率与数据结合起来得到真实分布的验后概率（参见1.6节和6.13节）。在解谱法的文献中（用贝叶斯方法），通常都以此为出发点。然而，遇到的困难是，先验知识通常都是定性的或非常复杂，很难用验前概率进行表示。先验观念的主观性在这里不是什么真正的缺点，在经典统计中，对如何选择有偏估计量也存在某种主观性。后面我们将主要使用经典统计的概念，利用偏倚和方差作为判断结果好坏的判选条件。并且，只要有可能，我们会指出与贝叶斯统计的联系。

在响应矩阵求逆法的最后，考虑一下这种情况：被解谱直方图的区间数 M 不等于测量直方图的区间数 N 。如果 $M > N$ ，(11.12) 式的矢量方程 $\nu = R\mu + \beta$ 是欠定方程组，解不唯一。可以用11.4节的方法选择一个解作为估计量 $\hat{\mu}$ 。如果 $M < N$ ，(11.12) 是超定方程组，一般来说不存在精确解。这时可以利用极大似然法或最小二乘法等构造近似解，即问题等价于第五章到第八章讨论的参数估计。如果 M 很大，估计量的相关性可能导致很大的方差。这种情况下，可能有必要利用11.4节介绍的正规化方法，以引入偏倚为代价减小方差。

11.3 修正因子法

考虑真实分布 μ 与数据 \mathbf{n} 的区间数相同的情况。一种最简单也可能最常用的方法是取 μ_i 的估计量为

$$\hat{\mu}_i = C_i(n_i - \beta_i), \quad (11.28)$$

其中 β_i 为本底期望值， C_i 为相乘的**修正因子 (correction factor)**。修正因子可以通过蒙特卡罗程序确定，其中既包括待研究过程的模型，也包括对探测仪器的模拟。在有探测器模拟和没有探测器模拟的情况下分别运行蒙特卡罗程序，得到每个区间真值和观测值的模型预言 ν_i^{MC} 和 μ_i^{MC} ，以此确定修正因子 C_i 。这里， ν_i^{MC} 指的是纯信号过程，不包括本底。修正因子为 ν_i^{MC} 与 μ_i^{MC} 的比率，

$$C_i = \frac{\nu_i^{\text{MC}}}{\mu_i^{\text{MC}}}. \quad (11.29)$$

假设可以产生足够多的蒙特卡罗数据，使得修正因子的统计误差可以忽略不计。如果该假设不成立，则 C_i 的不确定度可以通过误差传递进入估计量 $\hat{\mu}_i$ 的不确定度。

如果分辨率效应可以忽略不计，则响应矩阵为对角的，即 $R_{ij} = \delta_{ij}\epsilon_j$ ，因此有

$$\nu_i^{\text{sig}} = \nu_i - \beta_i = \epsilon_i \mu_i, \quad (11.30)$$

其中 ν_i^{sig} 为没有本底时区间 i 事例数的期望值。因此修正因子变为 $C_i = 1/\epsilon_i$ ，而 $1/C_i$ 起到广义效率的作用。然而，如果响应矩阵有非对角项， $1/C_i$ 的值可以大于 1。也就是说，由于不同区间内事例的迁移，在某个区间观测到的事例数可能超过真实事例数。

估计量 $\hat{\mu}_i$ 的期望值为

$$\begin{aligned} E[\hat{\mu}_i] &= C_i E[n_i - \beta_i] = C_i(\nu_i - \beta_i) = \frac{\nu_i^{\text{MC}}}{\mu_i^{\text{MC}}} \nu_i^{\text{sig}} \\ &= \left(\frac{\nu_i^{\text{MC}}}{\mu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i^{\text{sig}}} \right) \nu_i^{\text{sig}} + \mu_i. \end{aligned} \quad (11.31)$$

因此， $\hat{\mu}_i$ 是有偏的，只有当蒙特卡罗模型和真实实验中比率 μ_i/ν_i^{sig} 相同时，偏倚才变为零。

估计量的协方差矩阵为

$$\begin{aligned}\text{cov}[\hat{\mu}_i, \hat{\mu}_j] &= C_i^2 \text{cov}[n_i, n_j] \\ &= C_i^2 \delta_{ij} \nu_i.\end{aligned}\quad (11.32)$$

最后一行假设 n_i 为独立的泊松变量，均值为 ν_i 。对于很多实际问题， C_i 的量级大约为 1，因此 $\hat{\mu}_i$ 的方差与完美分辨率时得到的方差大致相同。此外，修正因子法很容易实现，甚至不需要对矩阵求逆，但是付出的代价是引入了偏倚，

$$b_i = \left(\frac{\nu_i^{\text{MC}}}{\mu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i^{\text{sig}}} \right) \nu_i^{\text{sig}}. \quad (11.33)$$

由于偏差造成的系统不确定度，可以通过计算不同蒙特卡罗模型的修正因子进行大致估计。显然更好的模型产生的偏倚更小，因此，通常推荐用估计得到的分布 $\hat{\mu}$ 来调整蒙特卡罗模型。也就是说，通过调整蒙特卡罗模型参数，提高 ν^{MC} 与减除本底后的数据 $\mathbf{n} - \beta$ 的符合程度。于是，可以通过反复迭代，用调整的模型得到改进的修正因子。

修正因子法存在的危险是，偏倚通常将估计量 $\hat{\mu}$ 推向模型预言值 μ^{MC} 。这使得模型检验变得非常困难，而这可能是实验的首要目的。这种情况下，必须确保在估计的系统误差中已经考虑了由于修正因子法的模型依赖导致解谱结果的不确定度，并且确保这些都已经包含到各种模型检验中。

11.4 正规化解谱法的一般策略

尽管修正因子法既简单又有广泛的实际应用，它还是存在很多缺点，这些缺点主要跟结果的模型依赖有关。另外一种方法是，以某种方式将光滑度的度量强加给真实直方图 μ 的估计量。这就是常说的解谱分布的**正规化(regularization)**。

作为出发点，让我们回到11.2节用矩阵求逆法得到的振荡解。 μ 的这个估计是通过数似然函数 $\log L_{\max}$ 的最大值，或者 χ^2 的最小值得到的。后面，我们通常只提及对数似然函数，对应的 χ^2 可以通过 $\log L = -\chi^2/2$ 得到。

可以考虑 μ -空间中围绕极大似然 (或最小二乘) 解的某个区域，用来表示可接受的所有解。这里所说的可接受指的是预言的期望值 ν 与数据 \mathbf{n} 的符合程度可以接受。这个区域的范围可以通过要求 $\log L$ 处于最大值的某个限定区域来定义。也就是说， μ -空间可接受的区域，可以通过

$$\log L(\mu) \geq \log L_{\max} - \Delta \log L \quad (11.34)$$

来确定，或者，对于最小二乘法，通过

$$\chi^2(\mu) \leq \chi_{\min}^2 + \Delta \chi^2 \quad (11.35)$$

来确定。其中 $\Delta \log L$ 或 $\Delta \chi^2$ 为适当选择的值。 $\Delta \log L$ 或 $\Delta \chi^2$ 的值将确定解谱得到的直方图中偏倚和方差的平衡点，在11.7节我们将回到这一点详细讨论。

除了所得解的可接受度，我们需要通过引入函数 $S(\mu)$ 来定义解的光滑度， $S(\mu)$ 称为**正规化函数(regularization function)**。下一节将讨论 $S(\mu)$ 的几种可能形式。一般策略是，从不等式 (11.34) 或 (11.35) 定义的可接受解中选择出光滑度最高的一个。在要求 $\log L(\mu)$ 等于 $\log L_{\max} - \Delta \log L$ 的约束条件下最大化正规化函数 $S(\mu)$ ，等价于求下面这个量

$$\alpha [\log L(\mu) - (\log L_{\max} - \Delta \log L)] + S(\mu) \quad (11.36)$$

在不同 μ 和 α 下的最大值。这里 α 为拉格朗日乘子，称为**正规化参数(regularization parameter)**。选取不同的 α ，对应于不同的 $\Delta \log L$ 值。对于给定的 α ，则通过最大化 $\log L$ 和 $S(\mu)$ 的加权组合确定估计量的解，

$$\Phi(\mu) = \alpha \log L(\mu) + S(\mu). \quad (11.37)$$

如果取 $\alpha = 0$ ，则得到最光滑的分布，数据 \mathbf{n} 完全被忽略。 α 很大时，则得到响应矩阵求逆法的振荡解，对应于似然函数取最大值的解。

要使最大化 $\Phi(\mu)$ 的做法与前面所说的一般性策略实际上等价，应当保证 $\log L(\mu)$ 和 $S(\mu)$ 等于常数的曲面行为足够好，尤其是不能有凹凸变化或复杂的拓扑结构，以至于存在多个局域极大值。下面我们默认这个假设始终成立。

注意到，由于关系式 $\nu = R\mu + \beta$ 总是成立，所以可以将 $\log L$ 和 S 写作 μ 或 ν 的函数。类似地，我们总是取关系式

$$\hat{\nu} = R\hat{\mu} + \beta \quad (11.38)$$

来定义 ν 的估计量，所以，知道了 $\hat{\nu}$ 就等价于知道了估计量 $\hat{\mu}$ 。然而，需要注意的是，与11.2节的方法不同，这里将不再有 $\hat{\mathbf{n}}\mathbf{u} = \mathbf{n}$ 。还需要牢记的一点是， $\mu_{\text{tot}} = \sum_j \mu_j$ 和 $\nu_{\text{tot}} = \sum_i \nu_i = \sum_{i,j} R_{ij} \mu_j$ 也是 μ 的函数。

这里我们将只考虑估计量 $\hat{\mu}$ ，估计的总事例数 $\hat{\nu}_{\text{tot}}$ 等于实际观测到的事例数，

$$\hat{\nu}_{\text{tot}} = \sum_{i=1}^N \hat{\nu}_i = \sum_{i=1}^N \sum_{j=1}^M R_{ij} \hat{\mu}_j + \beta_i = n_{\text{tot}}. \quad (11.39)$$

一般来说，这个条件不能自动满足，可以通过修改 (11.37) 式将其引入，得到

$$\varphi(\mu, \lambda) = \alpha \log L(\mu) + S(\mu) + \lambda \left[n_{\text{tot}} - \sum_{i=1}^N \nu_i \right], \quad (11.40)$$

其中 λ 为拉格朗日乘子。令 $\partial \varphi / \partial \lambda = 0$ 则可得到 $\sum_i \nu_i = n_{\text{tot}}$ 。

注意到，从技术上来说，将正规化参数乘在正规化函数 $S(\mu)$ 上 (很多参考文献都如此定义) 还是乘在似然函数上，这实际上无关紧要。在11.9节给出的数值实现中，把 α 乘在似然函数上更方便。

11.5 正规化函数

11.5.1 Tikhonov 正规化

经常使用的一种光滑度是真实分布微分的均方值。该技术是由菲利普 (Phillips) [70] 和吉洪诺夫 (Tikhonov) [71, 72] 先后独立提出来的，通常称为 **Tikhonov 正规化(Tikhonov Regularization)**。如果考虑离散化成直方图之前的概率密度函数 $f_{\text{true}}(y)$ ，则正规化函数为

$$S[f_{\text{true}}(y)] = - \int \left(\frac{d^k f_{\text{true}}(y)}{dy^k} \right)^2 dy, \quad (11.41)$$

其中积分区域为所有可能的 y 。前面的负号是因为我们选取的惯例，即要最大化 (11.40) 式定义的 φ 。也就是说， S 越大，正规化函数越光滑。(等价地，当然也可以最小化取相反符号的正规化函数与对数似然函数，很多文献中也经常遇到这种惯例。)

原则上,可以采用不同阶导数项的线性组合,然而实际应用中,通常选定 k 取某个值。如果 $f_{\text{true}}(y)$ 用直方图表示,导数替换为有限差分。对于等区间宽度,如果取 $k = 1$,可以得到 ([22])

$$S(\boldsymbol{\mu}) = - \sum_{i=1}^{M-1} (\mu_i - \mu_{i+1})^2, \quad (11.42)$$

如果 $k = 2$, 则

$$S(\boldsymbol{\mu}) = - \sum_{i=1}^{M-2} (-\mu_i + 2\mu_{i+1} - \mu_{i+2})^2, \quad (11.43)$$

如果 $k = 3$, 则

$$S(\boldsymbol{\mu}) = - \sum_{i=1}^{M-3} (-\mu_i + 3\mu_{i+1} - 3\mu_{i+2} + \mu_{i+3})^2. \quad (11.44)$$

通常选择 $k = 2$ 阶导数,所以 $S(\boldsymbol{\mu})$ 与平均曲率有关。

如果所有区间宽度 Δy_i 都相等,则区间宽度在 (11.42)-(11.44) 式中可以忽略,它们只提供了一个正比因子,可以等效地吸收到正规化参数 α 中。如果 Δy_i 并不完全相等,则可以直接将它们放到有限差分中。例如,对 $k = 2$,可以假定在任意三个相邻区间 $f_{\text{true}}(y)$ 都具有抛物线的形式,

$$f_i(y) = a_{0i} + a_{1i}y + a_{2i}y^2. \quad (11.45)$$

共有 $M - 2$ 组这样的 3 个相邻区间的组合,中心分别为区间 $i = 2, \dots, M - 1$ 。每组的系数可以通过下面的方式确定,即令 $f_i(y)$ 在区间 $i - 1$ 、 i 和 $i + 1$ 上的积分分别等于对应的 μ_{i-1} 、 μ_i 和 μ_{i+1} 。于是,以区间 i 为中心的这一组的二阶导数为 $f_i'' = 2a_{2i}$, 正规化函数可以写为

$$S(\boldsymbol{\mu}) = - \sum_{i=2}^{M-1} f_i''^2 \Delta y_i. \quad (11.46)$$

需要注意的是,在第一个和最后一个区间,二阶导数无法确定。这里,它们没有包含在 (11.46) 式的求和中,也就是说,这两个区间的二阶导数被取为零。另外一种做法是,将这两个区间的二阶微分分别取为第 2 个区间和第 $M - 1$ 个区间的导数值。

对于任意的 k 阶导数,无论区间宽度是否相同,前面给出的正规化函数 $S(\boldsymbol{\mu})$ 可以表示为

$$S(\boldsymbol{\mu}) = - \sum_{i,j=1}^M G_{ij} \mu_i \mu_j = -\boldsymbol{\mu}^T G \boldsymbol{\mu}, \quad (11.47)$$

其中 G 为对称的常数矩阵。例如,对于 $k = 2$ 并且等区间宽度的 (11.43) 式, G 可以表示为

$$\left. \begin{aligned} G_{ii} &= 6 \\ G_{i,i\pm 1} &= G_{i\pm 1,i} = -4 \\ G_{i,i\pm 2} &= G_{i\pm 2,i} = 1 \end{aligned} \right\} \quad 3 \leq i \leq M - 2, \quad (11.48)$$

$$\begin{aligned} G_{11} &= G_{MM} = 1, \\ G_{22} &= G_{M-1,M-1} = 5, \\ G_{12} &= G_{21} = G_{M,M-1} = G_{M-1,M} = -2, \end{aligned}$$

所有其它 G_{ij} 都等于零。

为了得到估计量及其协方差矩阵(见11.6节),需要给出 S 的一阶和二阶导数,分别为

$$\frac{\partial S}{\partial \mu_i} = -2 \sum_{j=1}^M G_{ij} \mu_j \quad (11.49)$$

和

$$\frac{\partial^2 S}{\partial \mu_i \partial \mu_j} = -2G_{ij}. \quad (11.50)$$

$k = 2$ 的 Tikhonov 正规化在粒子物理的解谱问题中已经得到了广泛的应用, 例如对结构函数 (轻子-核子散射中运动学变量的分布) 的解谱。更深入的介绍和讨论请参见文献 [73–76]。

11.5.2 基于熵的正规化函数

另外一种常用的正规化函数基于概率分布 $\mathbf{p} = (p_1, \dots, p_M)$ 的熵(entropy) H , 定义为 [77]

$$H = - \sum_{i=1}^M p_i \log p_i. \quad (11.51)$$

这里的想法是把熵解释为直方图 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ 光滑性的度量, 并以

$$S(\boldsymbol{\mu}) = H(\boldsymbol{\mu}) = - \sum_{i=1}^M \frac{\mu_i}{\mu_{\text{tot}}} \log \frac{\mu_i}{\mu_{\text{tot}}} \quad (11.52)$$

作为正规化函数。基于 (11.52) 式的估计量称为根据最大熵原理 (principle of maximum entropy, 或 MaxEnt) 构造的估计量。要想了解熵是如何与光滑度联系起来的, 可以考虑在总事例数为 μ_{tot} 的情况下, 以不同方式构造直方图 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ 的数目 (先把 μ_j 当作整数), 这个数目为

$$\Omega(\boldsymbol{\mu}) = \frac{\mu_{\text{tot}}!}{\mu_1! \mu_2! \cdots \mu_M!}. \quad (11.53)$$

(注意到, 在多项分布(2.6)式出现过相同的因子。)取 (11.53) 式的对数, 并利用斯特令 (Stirling) 近似, 即 n 很大时满足 $\log n! \approx n(\log n - 1)$, 可以得到

$$\begin{aligned} \log \Omega &\approx \mu_{\text{tot}}(\log \mu_{\text{tot}} - 1) - \sum_{i=1}^M \mu_i(\log \mu_i - 1) \\ &= - \sum_{i=1}^M \mu_i \log \frac{\mu_i}{\mu_{\text{tot}}} \\ &= \mu_{\text{tot}} S(\boldsymbol{\mu}). \end{aligned} \quad (11.54)$$

我们将利用 (11.54) 式, 把 $\log \Omega$ 推广至 μ_i 不是整数的情形。

如果所有的事例都集中在一个区间内, 即直方图的光滑性最差, 则只有一种方式构造直方图, 因此熵也最小。另一个极端情形是所有 μ_i 都相等, 即直方图对应均匀分布, 可以证明这种情况下熵最大。考虑约束条件 $\sum_i p_i = 1$ 下最大化 H , 需要用到拉格朗日乘子。

为供后面参考, 我们先列出基于熵的正规化函数 $S(\boldsymbol{\mu})$ 的一阶和二阶微分:

$$\frac{\partial S}{\partial \mu_i} = - \frac{1}{\mu_{\text{tot}}} \log \frac{\mu_i}{\mu_{\text{tot}}} - \frac{S(\boldsymbol{\mu})}{\mu_{\text{tot}}} \quad (11.55)$$

以及

$$\frac{\partial^2 S}{\partial \mu_i \partial \mu_j} = \frac{1}{\mu_{\text{tot}}^2} \left[1 - \frac{\delta_{ij} \mu_{\text{tot}}}{\mu_i} + \log \left(\frac{\mu_i \mu_j}{\mu_{\text{tot}}^2} \right) + 2S(\boldsymbol{\mu}) \right]. \quad (11.56)$$

11.5.3 利用熵的贝叶斯动机

很多关于解谱问题的文献中，最大熵原理是在贝叶斯统计的框架下发展起来的。(例如，参见文献 [18, 22, 69]。)然而，我们在后面将看到，这种解谱方法存在很大的困难。不管怎样，将贝叶斯最大熵与前一小节介绍的经典方法进行一番对比会很有意义。

在贝叶斯方法中， μ 是主观概率意义上的随机变量 (参见 1.2 节)，联合概率密度 $f(\mu|\mathbf{n})$ 表示真实直方图为 μ 的可信度。要想根据数据 \mathbf{n} 修正我们对 μ 的认识，可以利用贝叶斯定理，

$$f(\mu|\mathbf{n}) \propto L(\mathbf{n}|\mu)\pi(\mu), \quad (11.57)$$

其中 $L(\mathbf{n}|\mu)$ 为似然函数 (给定 μ 时数据为 \mathbf{n} 的条件概率)， $\pi(\mu)$ 为验前概率密度。验前概率密度表示在得到数据 \mathbf{n} 之前我们对 μ 的认识。

这里，我们将总事例数 μ_{tot} 当作整数，然而在经典统计方法中则不然。经典统计中 μ_{tot} 表示整数随机变量的期望值，本身不一定为整数。假设我们对这 μ_{tot} 个事例如何在直方图中分布没有任何先验知识，那么可以认为，根据对称性，任何一种分布方式都是等几率的。因此，在没有任何先验信息的情况下，某个直方图 (μ_1, \dots, μ_M) 的概率应当正比于构成该直方图的所有方式的数目，即 (11.53) 式中的 Ω 。因此，事例分布的总方式数 $\Omega(\mu)$ 被解释为验前概率 $\pi(\mu)$ ，

$$\begin{aligned} \pi(\mu) &= \Omega(\mu) = \frac{\mu_{\text{tot}}!}{\mu_1! \mu_2! \cdots \mu_M!} \\ &= \exp(\mu_{\text{tot}} H), \end{aligned} \quad (11.58)$$

其中 H 为 (11.51) 式定义的熵。

从严格的贝叶斯观点，只要确定了 $f(\mu|\mathbf{n})$ ，问题就已经结束。然而，由于这个函数的变量数目与解谱分布的区间数 M 相同，实际上完整地报告 $f(\mu|\mathbf{n})$ 是不可行的。因此，必须寻找某种方法对 $f(\mu|\mathbf{n})$ 加以概括。为此，典型的做法是选择一个单独的矢量 $\hat{\mu}$ 作为贝叶斯估计量。通常的选择是，使概率 $f(\mu|\mathbf{n})$ 最大的 μ ，或者等价地，使 $f(\mu|\mathbf{n})$ 的对数取最大值的 $\hat{\mu}$ 。根据 (11.57) 式，这可以通过最大化下面的量来确定：

$$\begin{aligned} \log f(\mu|\mathbf{n}) &\propto \log L(\mu|\mathbf{n}) + \log \pi(\mu) \\ &= \log L(\mu|\mathbf{n}) + \mu_{\text{tot}} H(\mu). \end{aligned} \quad (11.59)$$

因此，贝叶斯方法实际上对应于使用正规化函数

$$S(\mu) = \mu_{\text{tot}} H(\mu) = - \sum_{i=1}^M \mu_i \log \frac{\mu_i}{\mu_{\text{tot}}}. \quad (11.60)$$

而且，正规化参数 α 不再是任意参数因子，而是被设定为 1。如果所有的效率 ϵ_i 都相等，则 $\nu_{\text{tot}} = n_{\text{tot}}$ 的要求意味着 μ_{tot} 也是常数。于是，这等价于前面介绍的使用 $S(\mu) = H$ 作为正规化函数并取 $\alpha = 1/\mu_{\text{tot}}$ 。

然而，如果效率不完全相等， ν_{tot} 为常数并不意味着 μ_{tot} 也是常数，结果导致最大 $S(\mu) = \mu_{\text{tot}} H(\mu)$ 的分布不再是均匀分布。这是因为 S 可以简单地通过提高 μ_{tot} 而增大，因此在最大 S 的分布中，效率低的区间被提高了。这种情况下，利用 H 和 $\mu_{\text{tot}} H$ 作为正规化函数将导致结果在一定程度上的差异，尽管实际中这种差异并不大，只要不同区间的效率在一个量级上。无论如何， $S = H$ 作为光滑性的度量很容易进行判别，因为最大 H 的分布总是均匀分布。

在 11.9 节我们将看到，(11.59) 式的贝叶斯估计量中熵的权重太大 (见图 11.3(a) 和文献 [68])。从经典统计的观点，贝叶斯方法没有很好地在偏倚与方差之间进行平衡，导致不太合理的较大偏倚。可以对贝叶斯

解释进行修正，将 (11.59) 式中的 μ_{tot} 替换为有效事例数 μ_{eff} ，并调整 μ_{eff} 使其小于 μ_{tot} 。于是，估计量可以通过最大化

$$\log L(\boldsymbol{\mu}|\mathbf{n}) + \mu_{\text{eff}} H(\boldsymbol{\mu}) \quad (11.61)$$

给出。这等效于前面介绍的使用 $S(\boldsymbol{\mu}) = H(\boldsymbol{\mu})$ 作为正规化函数，参数 μ_{eff} 起到正规化参数的作用。

贝叶斯原始解的问题，根源在于使用 $\Omega(\boldsymbol{\mu})$ 作为验前概率密度。无论从贝叶斯统计还是经典统计的角度， $\mathbf{p} = \boldsymbol{\mu}/\mu_{\text{tot}}$ 都是一组未知常数，例如某特殊类型 β 衰变的电子能量分布。无论哪种情况，我们关于分布形式的先验知识（即关于 \mathbf{p} 而不是 $\boldsymbol{\mu}$ 的分布形式），都应当与得到的数据样本的观测数无关。这实际上指向了使用 $\pi(\boldsymbol{\mu}) = \Omega(\boldsymbol{\mu})$ 的根本问题，因为随着 μ_{tot} 的增大， $\Omega(\boldsymbol{\mu})$ 变得更加集中于均匀分布（即所有的 p_i 都相等）。

经常出现的情况是，我们的确对概率分布 \mathbf{p} 存在一定的先验信念，但很难定量描述。例如，我们可以说具有大量结构的分布基本不太可能，因为很难想象物理理论会预言存在很多峰的分布。另一方面，完全平坦的分布看上去也不存在物理意义，所以， $\Omega(\boldsymbol{\mu})$ 并不能真实反映我们的先验信念。由于将 $\Omega(\boldsymbol{\mu})$ 解释为验前概率密度函数存在这些困难，我们将继续采用经典统计的方法，而只把熵当作正规化函数的一种可能选择。

11.5.4 基于交叉熵的正规化函数

注意到最大熵的分布是均匀的，所以 $\hat{\boldsymbol{\mu}}$ 中引入的偏倚将使结果更倾向于均匀分布。假设我们已知某分布 $\mathbf{q} = (q_1, \dots, q_M)$ ，并将其当作真实分布 $\mathbf{p} = \boldsymbol{\mu}/\mu_{\text{tot}}$ 最可能的先验形式。我们称 \mathbf{q} 为**参考分布 (reference distribution)**。然而，假设我们不知道如何定量描述对 \mathbf{q} 的信任程度，因此没有验前概率密度 $\pi(\boldsymbol{\mu})$ 可以用于贝叶斯定理。也就是说， \mathbf{q} 表示归一化的 $\boldsymbol{\mu}/\mu_{\text{tot}}$ ，使得验前概率密度最大，但它并不指定完整的验前概率密度。

这种情况下，正规化函数可以取为

$$S(\boldsymbol{\mu}) = K(\mathbf{p}; \mathbf{q}), \quad (11.62)$$

其中 $K(\mathbf{p}; \mathbf{q})$ 称为**交叉熵 (cross-entropy)** [78]，或**香农-杰恩斯 (Shannon-Jaynes) 熵** [79]，定义为

$$K(\mathbf{p}; \mathbf{q}) = - \sum_{i=1}^M p_i \log \frac{p_i}{M q_i}. \quad (11.63)$$

交叉熵的定义经常不包含 M 因子，有时也不包括前面的负号，这种情况下，最大熵原理变成最小交叉熵原理。我们将保留定义中的负号，使得 $K(\mathbf{p}; \mathbf{q})$ 与 (11.51) 式定义的 Shannon 熵 $H(\mathbf{p})$ 保持一致。需要注意的是，当参考分布为均匀分布，即对所有的 i 满足 $q_i = 1/M$ 时， $K(\mathbf{p}; \mathbf{q}) = H(\mathbf{p})$ 。

很容易证明，当概率分布 \mathbf{p} 等于参考分布 \mathbf{q} 时，交叉熵 $K(\mathbf{p}; \mathbf{q})$ 取最大值。使用 (11.62) 式的正规化函数的效果是，如果真实分布等于（或接近于）参考分布，则估计量 $\hat{\boldsymbol{\mu}}$ 的偏倚为零（或非常小）。

11.6 估计量的方差和偏倚

估计量 $\hat{\boldsymbol{\mu}}$ 是数据 \mathbf{n} 的函数，所以其自身也是随机变量。为了得到协方差矩阵 $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$ ，我们可以计算 $\hat{\boldsymbol{\mu}}$ 作为 \mathbf{n} 的函数的近似表达式，然后利用误差传递公式 (1.54) 将 U 与数据的协方差矩阵 $V_{ij} = \text{cov}[n_i, n_j]$ 联系起来。

估计量 $\hat{\boldsymbol{\mu}}$ 是通过最大化 (11.40) 式中的函数 $\varphi(\boldsymbol{\mu}, \lambda)$ 得到的，该函数由给定的对数似然函数（或 χ^2 函数）以及某种形式的正规化函数 $S(\boldsymbol{\mu})$ (Tikhonov 规则、最大熵规则，等等) 定义。因此，估计量 $\hat{\boldsymbol{\mu}}$ 和拉格朗

日乘子 λ 是 $M + 1$ 个方程构成的方程组的解

$$F_i(\boldsymbol{\mu}, \lambda, \mathbf{n}) = 0, \quad i = 1, \dots, M + 1, \quad (11.64)$$

其中

$$F_i(\boldsymbol{\mu}, \lambda, \mathbf{n}) = \begin{cases} \frac{\partial \varphi}{\partial \mu_i} & i = 1, \dots, M, \\ \frac{\partial \varphi}{\partial \lambda} & i = M + 1. \end{cases} \quad (11.65)$$

假设实际获得的数据为矢量 $\tilde{\mathbf{n}}$ ，对应的估计值为 $\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\tilde{\mathbf{n}})$ ，拉格朗日乘子 λ 的值为 $\tilde{\lambda}$ 。我们希望知道，如果数据变为不同值 \mathbf{n} ， $\hat{\boldsymbol{\mu}}$ 和 λ 将如何变化。将函数 $F_i(\boldsymbol{\mu}, \lambda, \mathbf{n})$ 围绕 $\tilde{\boldsymbol{\mu}}$ 、 $\tilde{\lambda}$ 和 $\tilde{\mathbf{n}}$ 泰勒展开到第一阶，得到

$$\begin{aligned} F_i(\boldsymbol{\mu}, \lambda, \mathbf{n}) \approx & F_i(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}) + \sum_{j=1}^M \left[\frac{\partial F_i}{\partial \mu_j} \right]_{\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}} (\mu_j - \tilde{\mu}_j) \\ & + \left[\frac{\partial F_i}{\partial \lambda} \right]_{\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}} (\lambda - \tilde{\lambda}) + \sum_{j=1}^M \left[\frac{\partial F_i}{\partial n_j} \right]_{\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}} (n_j - \tilde{n}_j). \end{aligned} \quad (11.66)$$

第一项 $F_i(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}})$ 以及整个表达式 $F_i(\boldsymbol{\mu}, \lambda, \mathbf{n})$ 都等于零，因为两组变量都是方程组的解。对方程 (11.66) 求解 $\boldsymbol{\mu}$ 得到

$$\hat{\boldsymbol{\mu}}(\mathbf{n}) \approx \tilde{\boldsymbol{\mu}} - A^{-1}B(\mathbf{n} - \tilde{\mathbf{n}}), \quad (11.67)$$

其中 $\boldsymbol{\mu}$ 的第 $M + 1$ 个分量为拉格朗日乘子 λ 。 $(M + 1) \times (M + 1)$ 的对称矩阵 A 为

$$A_{ij} = \begin{cases} \frac{\partial^2 \varphi}{\partial \mu_i \partial \mu_j}, & i, j = 1, \dots, M, \\ \frac{\partial^2 \varphi}{\partial \mu_i \partial \lambda} = -1, & i = 1, \dots, M, j = M + 1, \\ \frac{\partial^2 \varphi}{\partial \lambda^2} = 0, & i = M + 1, j = M + 1, \end{cases} \quad (11.68)$$

$(M + 1) \times N$ 的矩阵 B 为

$$B_{ij} = \begin{cases} \frac{\partial^2 \varphi}{\partial \mu_i \partial n_j}, & i = 1, \dots, M, j = 1, \dots, N \\ \frac{\partial^2 \varphi}{\partial \lambda \partial n_j}, & i = M + 1, j = 1, \dots, N. \end{cases} \quad (11.69)$$

利用 (1.54) 式的误差传递公式，可以由数据的协方差矩阵 $V_{ij} = \text{cov}[n_i, n_j]$ 得到估计量的协方差矩阵 $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$ ，

$$\text{cov}[\hat{\mu}_i, \hat{\mu}_j] = \sum_{k,l=1}^N \frac{\partial \hat{\mu}_i}{\partial n_k} \frac{\partial \hat{\mu}_j}{\partial n_l} \text{cov}[n_i, n_j]. \quad (11.70)$$

(11.70) 式中的导数可以利用 (11.67) 式计算得出

$$\frac{\partial \hat{\mu}_i}{\partial n_k} = -(A^{-1}B)_{ik} \equiv C_{ik}, \quad (11.71)$$

其中矩阵 A 和 B 由 (11.68) 和 (11.69) 式给出。我们这里要用的不是整个矩阵 C ，而只是 $M \times N$ 的子矩阵，第 $i = M + 1$ 行表示拉格朗日乘子 λ ，需要排除在外。因此，协方差矩阵 U 的最终表达式可以表示成更紧凑的形式

$$U = CVC^T. \quad (11.72)$$

(11.68) 式和 (11.69) 式中的导数依赖于正规化函数的选择，以及用来定义 $\varphi(\boldsymbol{\mu})$ 的特定的对数似然函数，例如泊松分布，高斯分布 ($\log L = -\chi^2/2$)，等等。比如，如果数据是独立的泊松变量，协方差矩阵为

$V_{ij} = \delta_{ij}\nu_i$ ，并且使用基于熵的正规化函数 (11.54)，则有

$$\begin{aligned} \frac{\partial^2 \varphi}{\partial \mu_i \partial \mu_j} &= -\alpha \sum_{k=1}^N R_{ki} R_{kj} \frac{n_k}{\nu_k^2} \\ &\quad + \frac{1}{\mu_{\text{tot}}} \left[1 - \frac{\delta_{ij} \mu_{\text{tot}}}{\mu_i} + \log \left(\frac{\mu_i \mu_j}{\mu_{\text{tot}}^2} + 2S(\boldsymbol{\mu}) \right) \right] \end{aligned} \quad (11.73)$$

以及

$$\frac{\partial^2 \varphi}{\partial \mu_i \partial n_j} = \frac{\alpha R_{ji}}{\nu_j}. \quad (11.74)$$

利用实际实验得到的 $\boldsymbol{\mu}$ 和 $\boldsymbol{\nu}$ 的估计值，矩阵 A 和 B 可以通过计算 (11.73) 和 (11.74) 式的导数确定 (因此也确定了矩阵 C)。表 11.1 总结了泊松和高斯对数似然函数所需要的要素 (各阶导数)。需要注意的是，对于高斯情形，即对于最小二乘法，表中的变量都是指 $\log L = -\frac{1}{2}\chi^2$ ，而不是 χ^2 本身。Tikhonov 正规化函数和基于熵的正规化函数的导数分别在 11.5.1 小节和 11.5.2 小节给出。

	泊松分布	高斯分布 (最小二乘)
$\log L$	$\sum_i (n_i \log \nu_i - \nu_i)$	$-\frac{1}{2} \sum_{i,j} (\nu_i - n_i) (V^{-1})_{ij} (\nu_j - n_j)$
$\frac{\partial \log L}{\partial \mu_i}$	$\sum_j \left(\frac{n_j}{\nu_j} - 1 \right) R_{ji}$	$-\sum_{j,k} R_{ji} (V^{-1})_{jk} (\nu_k - n_k)$
$\frac{\partial^2 \log L}{\partial \mu_i \partial \mu_j}$	$-\sum_k \frac{n_k R_{ki} R_{kj}}{\nu_k^2}$	$-(R^T V^{-1} R)_{ij}$
$\frac{\partial^2 \log L}{\partial n_i \partial \mu_j}$	$\frac{R_{ij}}{\nu_i}$	$(V^{-1} R)_{ij}$

表 11.1: 泊松和高斯变量的对数似然函数及其微分。

为了确定偏倚 $b_i = E[\hat{\mu}_i] - \mu_i$ ，我们可以利用 (11.67) 式的近似关系计算期望值 $E[\hat{\mu}_i]$ ，

$$b_i = E[\hat{\mu}_i] - \mu_i \approx \tilde{\mu}_i + \sum_{j=1}^N C_{ij} (\nu_j - \tilde{n}_j) - \mu_i. \quad (11.75)$$

这个偏倚的估计，可以通过将 μ_i 替换为 (11.67) 式中的估计量，并将 ν_j 替换为对应的估计量 $\hat{\nu}_j = \sum_k R_{jk} \hat{\mu}_k$ 来进行，从而得到偏倚的估计量为

$$\hat{b}_i = \sum_{j=1}^N C_{ij} (\hat{\nu}_j - n_j) = \sum_{j=1}^N \frac{\partial \hat{\mu}_i}{\partial n_j} (\hat{\nu}_j - n_j). \quad (11.76)$$

当 $(\hat{\nu}_j - n_j)$ 很小时，或者等效地说，当正规化参数 α 比较大时，构造估计量 \hat{b}_i 的近似是合理的。当 α 很小时，矩阵 C 实际上趋近于零，因为此时估计量 $\hat{\mu}_i$ 与测量数据 n_j 之间处于退耦状态，参见 (11.71) 式。这种情况下实际上偏倚最大。但是，由于我们只是用偏倚 \hat{b}_i 和方差一起确定正规化参数 α ，这个近似对我们的目的来说足够了。

通过误差传递 (忽略矩阵 C 的方差)，可以得到 \hat{b}_i 的协方差矩阵 W ，

$$W_{ij} = \text{cov}[\hat{b}_i, \hat{b}_j] = \sum_{k,l=1}^N C_{ik} C_{jl} \text{cov}[(\hat{\nu}_k - n_k), (\hat{\nu}_l - n_l)]. \quad (11.77)$$

可以通过利用 $\hat{\nu}_k = \sum_m R_{km} \hat{\mu}_m$, 将协方差矩阵 $\text{cov}[\hat{\nu}_k, \hat{\nu}_l]$ 与真实分布的估计量的协方差矩阵 $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$ 联系起来, 从而计算出 W_{ij} 。而真实分布估计量的协方差矩阵又通过 (11.72) 式与数据的协方差矩阵联系起来, $U = CVC^T$ 。将所有这些关系联合起来得到

$$\begin{aligned} W &= (CRC - C)V(CRC - C)^T \\ &= (CR - I)U(CR - I)^T, \end{aligned} \quad (11.78)$$

其中 I 为 $M \times M$ 单位矩阵。方差 $V[\hat{b}_i] = W_{ii}$ 可以用来判断估计的偏倚是否明显偏离零, 这反过来可以用来作为确定正规化参数的判选条件。

然而, 在进入正规化参数选择的问题之前, 很重要的是要注意到, 所有正规化解谱法的偏倚通常都不为零, 它们是真实分布的函数 (不处处为零的函数)。然而, 偏倚的数值对某些特殊的 μ 实际上可以为零。解谱的一个指导原则是, 如果 μ 拥有先验相信正确的性质, 则要选择一种方法使偏倚为零 (或者很小)。例如, 如果真实分布是均匀分布, 则 (11.42) 式中基于 $k = 1$ 的 Tikhonov 正规化得到的估计将具有零偏倚。如果真实分布是线性的, 则 (11.43) 式中基于 $k = 2$ 的 Tikhonov 正规化给出的估计具有零偏倚。如果真实分布等于参考分布 \mathbf{q} , 则利用 (11.63) 式交叉熵的解谱法将得到零偏倚的估计。

11.7 正规化参数的选择

正规化参数 α 的选择, 或者等价地, $\Delta \log L$ (或 $\Delta \chi^2$) 的选择, 确定了估计量 $\hat{\mu}$ 的偏倚与方差之间的平衡。如果 α 非常大, 结果几乎由似然函数决定, 可以得到 $\log L = \log L_{\max}$ (或者在最小二乘法中 $\chi^2 = \chi_{\min}^2$), 对应的方差非常大。另一个极端, 如果 $\alpha \rightarrow 0$, 则所有的权重都集中在正规化函数上, 得到的是最光滑的解。

很多最佳折衷的定义都是可行的, 这些定义可以结合协方差矩阵的估计值 $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$, 偏倚的估计值 \hat{b}_i , 以及偏倚估计量的协方差矩阵 $W_{ij} = \text{cov}[\hat{b}_i, \hat{b}_j]$ 。这里出现的 U 和 W 都指的是它们的估计值 \hat{U} 和 \hat{W} , 符号中的尖帽将不明显写出来。

最终结果优度的一种可能度量是对所有区间求平均的均方差, 参见 (5.5) 式,

$$\text{MSE} = \frac{1}{M} \sum_{i=1}^M (U_{ii} + \hat{b}_i^2). \quad (11.79)$$

确定 α 从而得到某个 MSE 值的方法依赖于数值实现。通常简单的做法是, 选择 α 的某个尝试值, 最大化 $\varphi(\mu, \lambda)$, 然后重复该过程进行迭代, 直到得到的解符合要求。

然而, 我们可以讨论说, 由于不同区间的测量精度可能不同, 不同区间均方差的贡献也应当不同。由于泊松变量的方差等于均值, 所以均值为 μ_i 的泊松变量的方差为 μ_i , 类似于最小二乘法中的 χ^2 , 我们可以定义加权的均方差 (MSE),

$$\text{MSE}' = \frac{1}{M} \sum_{i=1}^M \frac{U_{ii} + \hat{b}_i^2}{\hat{\mu}_i}. \quad (11.80)$$

如果数据服从泊松分布, 则 MSE' 表示由于有限分辨率导致均方差的增大。因此, 要求这个量很小是很自然的事情。

正规化参数选择的一种流行做法基于这样的思想, 平均来说, 每个区间对 χ^2 的贡献应当大致等于 1, 也就是说, 由 $\chi^2 = N$ 来确定 α 。如果 \mathbf{n} 服从高斯分布, 则有 $\log L = -\chi^2/2$, 由于对高斯分布的 \mathbf{n} 来说, $\log = -\chi^2/2$, 所以, 可以将前面的方法推广到对数似然函数的情形, 即 α 的选取要保证 $\Delta \log L = \log L_{\max} - \log L = N/2$ 。

我们可能会很自然地期望, χ^2 增加一个单位可以设定数据 \mathbf{n} 与估计值 $\hat{\mathbf{v}}$ 之间差异的适当标准。然而, 这种做法通常会导致结果出现不合理的巨大方差。问题的根源在于, 估计量 \hat{v}_i 不仅包含 n_i 的贡献, 还包含相邻区间的贡献。估计量 \hat{v}_i 与测量值 n_j 之间的耦合可以用矩阵表示

$$\frac{\partial \hat{v}_i}{\partial n_j} = \frac{\partial}{\partial n_j} \sum_{k=1}^M R_{ik} \hat{\mu}_k = (RC)_{ij}. \quad (11.81)$$

文献 [80] 建议对条件 $\Delta\chi^2 = 1$ 进行修改, 其想法是将有效 χ^2 增加一个单位,

$$\Delta\chi_{\text{eff}}^2 = (\hat{\mathbf{v}} - \mathbf{n})^T R C V^{-1} (RC)^T (\hat{\mathbf{v}} - \mathbf{n}) = 1, \quad (11.82)$$

其中矩阵 RC 等效地考虑了估计量 \hat{v}_i 与数据 n_i 之间的约化耦合。

作为另外一种选择, 可以考察偏倚及其方差的估计。如果偏倚明显不为零, 则显然需要将其减除。这等价于取更小的 $\Delta \log L$ 。作为偏倚偏离零的度量, 我们可以构造其平方的加权求和,

$$\chi_b^2 = \sum_{i=1}^M \frac{\hat{b}_i^2}{W_{ii}}. \quad (11.83)$$

因此, 我们的策略是减小 $\Delta \log L$ (即, 增大 α), 直到 χ_b^2 足够小, 比如 χ_b^2 等于区间数 M 。这种情况下, 偏倚的标准差大致等于偏倚本身, 此时, 进一步减小偏倚去掉一定误差的同时也会新引入同样大小的误差。

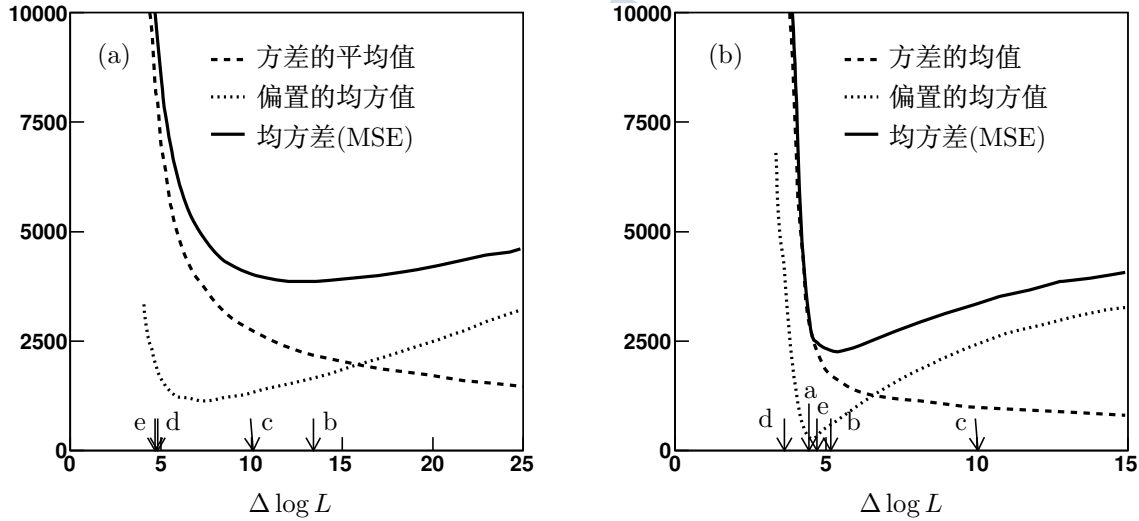


图 11.2: 平均方差、平均偏倚的平方以及二者之和 (均方差) 的估计值随 $\Delta \log L$ 的变化关系。图 (a) 和图 (b) 分别采用了最大熵和 $k = 2$ 的 Tikhonov 正规化规则。箭头指出了图 11.3 和图 11.4 的解: 箭头 (b)、(c)、(d)、(e) 分别对应最小均方差 (MSE)、 $\Delta \log L = N/2$ 、 $\Delta\chi_{\text{eff}}^2 = 1$ 和 $\chi_b^2 = M$ 。图中没有指出最大熵情形的贝叶斯解 $\Delta \log L = 970$ 。对于 Tikhonov 正规化, 箭头 (a) 指出了最小加权均方差 (MSE) 的解。

图 11.2 给出了偏倚的平方、方差以及二者之和 (即均方差 MSE) 作为 $\Delta \log L$ 的函数。使用的数据是图 11.1 的样本。当时采用响应矩阵求逆法进行解谱, 这里分别采用最大熵解谱法 ((a) 图) 和 Tikhonov 正规化解谱法 ((b) 图)。当 $\Delta \log L$ 比较小时, 估计的偏倚增加很快, 这反映了估计量 \hat{b}_i 的方差; 当 $\Delta \log L$ 趋于零时, 真实的偏倚应当变为零。箭头所指的地方表示前面介绍的各种判选方法的解, 下一节将对此进行深入讨论。

还有很多其它设定正规化参数的条件, 例如基于奇异值分析的方法 [74], 或者利用交叉验证过程的方法 [81]。然而, 正规化参数 α 的选择在某种程度上仍是尚未解决的问题。实际应用中, 随着 $\Delta \log L$ 的减

小，最终的估计相对稳定，直到达到某一点时方差突然暴增 (见图11.2)。方差突然快速增大的起点，大致给出了设定 α 的自然选择。

11.8 解谱法举例

图11.3和11.4 分别给出了基于最大熵和 Tikhonov 正规化的例子。 μ 、 ν 和 \mathbf{n} 的分布与图11.1的 (a)-(c) 相同，区间数为 $N = M = 20$ ，所有的效率 ϵ 都等于 1，本底 β 等于零。估计量 $\hat{\mu}$ 是通过最大化 $\varphi(\mu, \lambda)$ (11.40) 得到的，其中 $\varphi(\mu, \lambda)$ 是基于独立泊松分布的数据利用对数似然函数构造的。左图给出了原始的“真实”分布 μ 、解谱后的分布 $\hat{\mu}_i$ ，以及相应的误差棒 $\sqrt{U_{ii}}$ ，该误差对应于给定正规化参数 α (等价于给定的 $\Delta \log L$)。右图给出了对应的偏倚 \hat{b}_i 的估计及其标准差 $\sqrt{W_{ii}}$ 。不要将此与真值残差 $\hat{\mu}_i - \mu_i$ 混淆，如果不知道真实直方图 μ ，则该残差无法构造出来。然而， \hat{b}_i 的估计则由数据确定。

先考虑图11.3，其正规化函数为基于熵的正规化函数 $S(\mu) = H(\mu)$ 。图11.3(a) 对应于 $\alpha = 1/\mu_{\text{tot}}$ ，即贝叶斯方法 (11.59)，得到 $\Delta \log L = 970$ 。给出正规化参数的这一选择，我们只是想说明验前概率密度 $\pi(\mu) = \Omega(\mu)$ 无法得到合理解，尽管标准差 $\sqrt{U_{ii}}$ 非常小，但偏倚很大。右图给出的 \hat{b}_i 的估计的确非常大，而且从其误差棒可以看出这个偏倚明显不为零。需要注意的是，由于构造估计量 \hat{b}_i 时采用的近似 (参见 (11.67) 式)，这里估计的偏倚实际上与真值残差 $\hat{\mu}_i - \mu_i$ 并不能很好地吻合。随着 $\Delta \log L$ 的减小，这个近似变得越来越好，直到标准差 $\sqrt{W_{ii}}$ 与偏倚本身相当。

图11.3(b) 给出的是基于最小均方差 (11.79) 的结果，给出 $\Delta \log L = 13.3$ 和 $\chi_b^2 = 154$ 。尽管估计的偏倚远小于 $\alpha = 1/\mu_{\text{tot}}$ 时的偏倚，但该偏倚仍然明显不为零。

图11.3(c) 给出的是 $\Delta \log L = N/2 = 10$ 的解。这里给出的偏倚比最小均方差 (MSE) 给出的偏倚要稍微微小一些，但还是明显不为零，给出 $\chi_b^2 = 87$ 。在这个例子中，要求 (11.80) 式中的加权均方差最小给出 $\Delta \log L = 10.5$ ，因此与 $\Delta \log L = N/2 = 10$ 的结果差不多。

图11.3(d) 和 (e) 分别给出了 $\Delta \chi_{\text{eff}}^2 = 1$ 和 $\chi_b^2 = M$ 的结果。这两种方法给出的偏倚都接近于零，与 $\Delta \log L = N/2$ 或最小均方差的结果比较，其代价是方差更大。对于 $\Delta \chi_{\text{eff}}^2 = 1$ 的情形， $\chi_b^2 = 20.8$ ；对于 $\chi_b^2 = M$ 的情形， $\Delta \chi_{\text{eff}}^2 = 0.85$ 。所以，在这个例子中，这两种方法实际上非常相似。

再考虑图11.4，使用的分布与图11.3相同，但采用的正规化方法是 $k = 2$ 的 Tikhonov 正规化。其中 (a) 图对应于最小加权均方差 (MSE)，(b) 图对应最小均方差 (MSE)，(c) 图对应 $\Delta \log L = N/2$ ，(d) 图对应 $\Delta \chi_{\text{eff}}^2 = 1$ ，(e) 图对应 $\chi_b^2 = M$ 。这些结果中， $\Delta \log L = N/2$ 的解不是很成功，尽管其统计误差 $\sqrt{U_{ii}}$ 很小，但偏倚很大并且明显不为零 ($b_i^2 \gg W_{ii}$)。(a)、(b) 和 (e) 的结果比较合理，而要求 $\Delta \chi_{\text{eff}}^2 = 1$ (d) 给出的结果显得过于好了。(d) 中的偏倚与零偏倚一致，但不比 $\chi_b^2 = M$ 的结果更好，然而其统计误差却大很多。

Tikhonov 正规化的一个问题是，得到的估计可能为负，这一点可以在图11.4的区间看到。(只有图11.4(a) 所有区间的频数都为正。) 实际上，这个算法本身没有办法防止负值的出现。如果一定要避免负值的出现，则需要对算法进行修改，例如，在可能出现负值的地方，手动减小误差。最大熵解谱法不存在这种问题，因为如果任何 μ_i 趋近于零，则 $S(\mu)$ 的梯度发散，这使得所有的 ν_i 保持为正。

本章讨论的技术可以很容易推广到多维分布。例如，对二维情形，解谱法已经广泛应用于图像还原 (image restoration) [67, 82–84]，尤其是天文学 [85]、医学成像 [86]。对多维情况的完整讨论超出了本书的范畴，我们只利用简单的例子说明一些主要的思想。

图11.5的例子是用最大熵解谱法对一个 56×56 像素的测试图片进行解谱。图11.5(a) 为“真实”图像，表示矢量 μ 。在图11.5(b) 中，图像被高斯分辨率函数模糊化，高斯分布的标准差取为像素大小的 0.6 倍。

出于本练习的目的，设定了有效“光子”数 (或者根据成像系统类型的不同，也可能是卤化银颗粒数、光电子数等等)，使得最亮像素点的频数为 10^4 量级。因此，如果把像素点 i 的频数 n_i 看作期望值为 ν_i 的

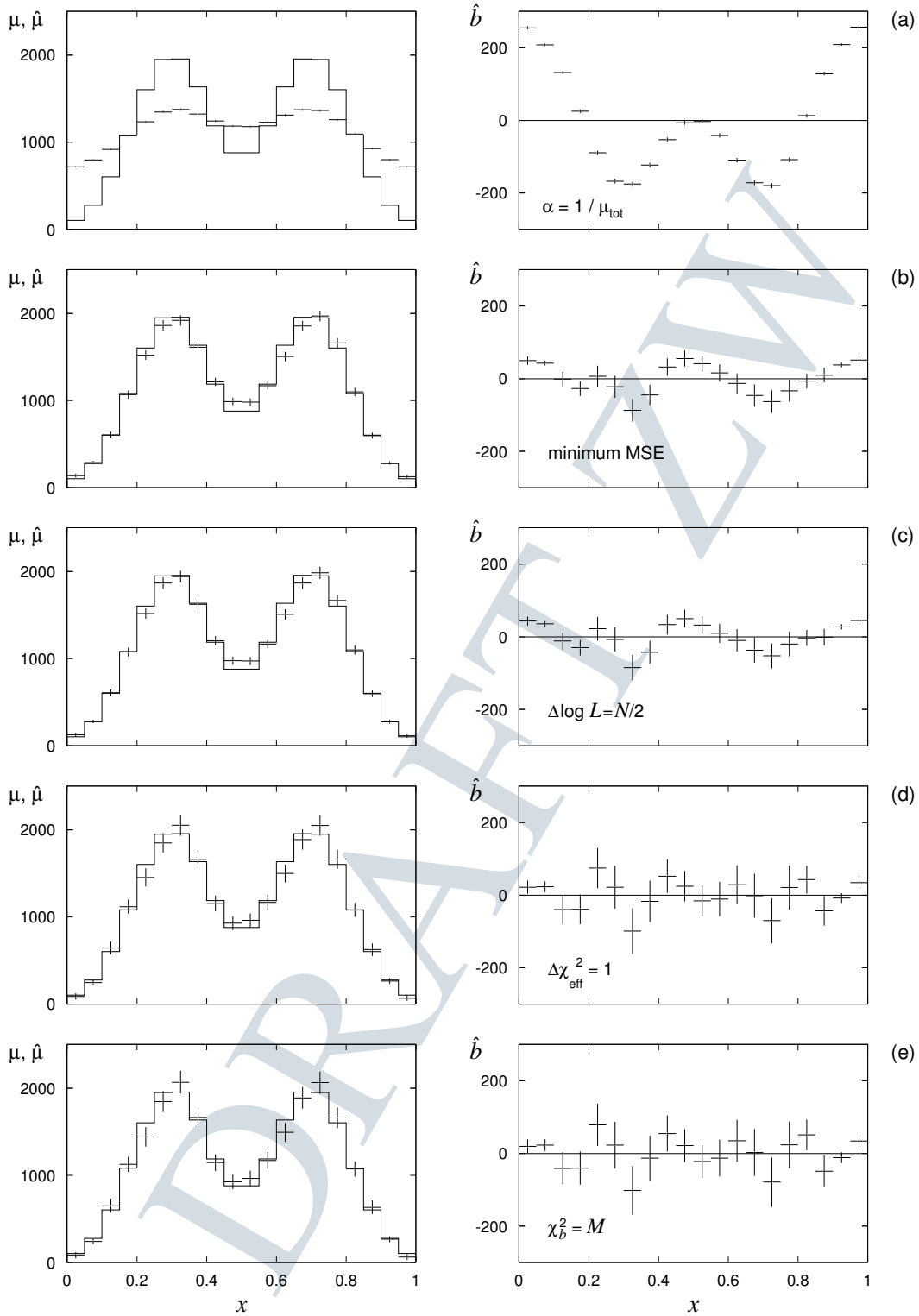


图 11.3: 左图中, 最大熵解谱法得到的分布用“点”表示, 真实分布用直方图表示。右图为正规化参数 α 取不同值时偏倚的估计值。其中, (a) 图对应贝叶斯规则 $\alpha = 1/\mu_{\text{tot}}$, (b) 图对应最小均方差 (MSE), (c) 图对应 $\Delta L = N/2$, (d) 图对应 $\Delta \chi_{\text{eff}}^2$, (e) 图对应 $\chi_b^2 = M$ 。在这个例子里, 最小加权均方差与 (c) 中 $\Delta \log L = N/2$ 类似。

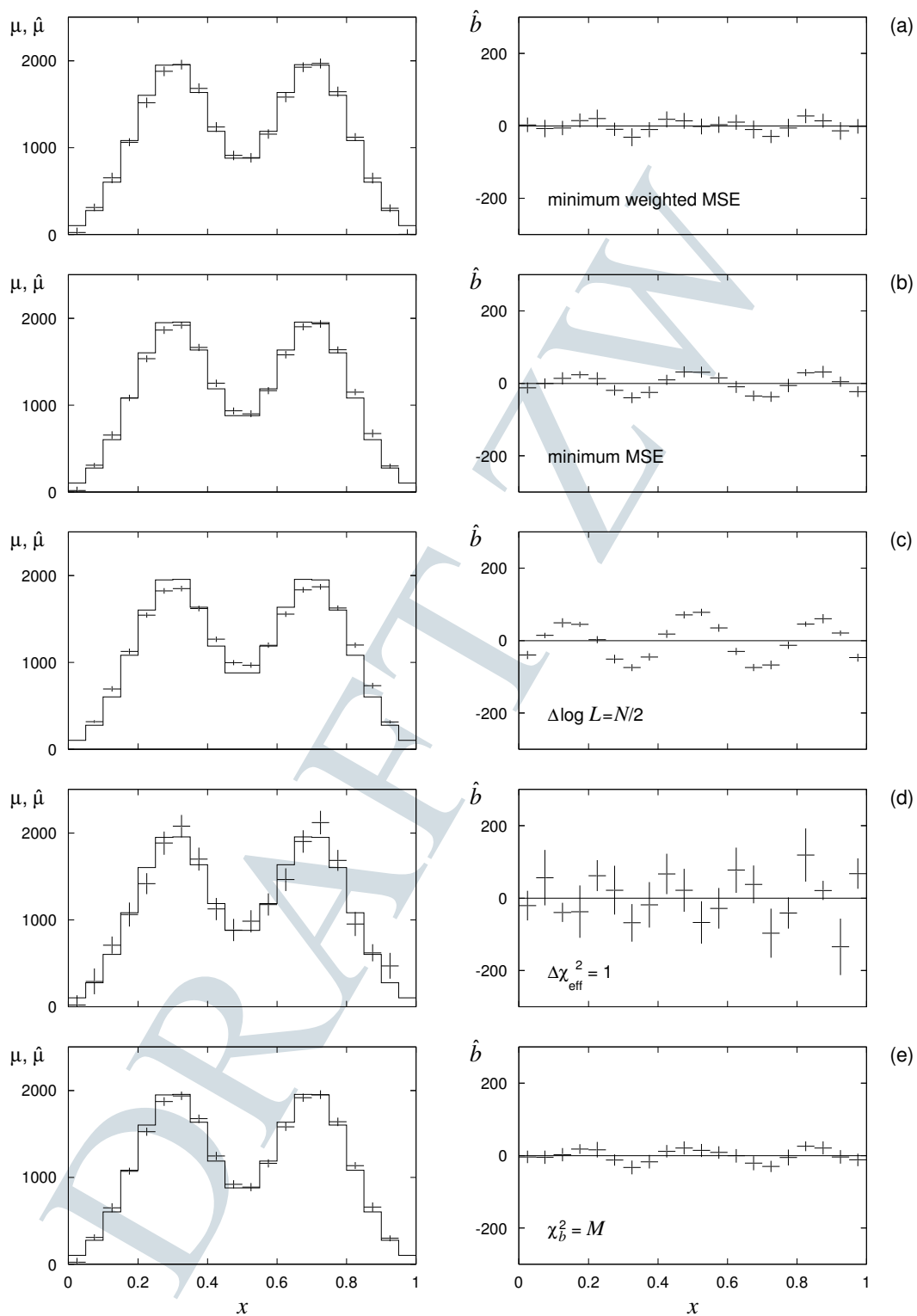


图 11.4: 左图中, 利用 Tikhonov 正规化解谱法 ($k=2$) 得到的分布用“点”表示, 真实分布用直方图表示。右图为正规化参数 α 取不同值时偏倚的估计值。其中, (a) 图对应最小加权均方差, (b) 图对应最小均方差 (MSE), (c) 图对应 $\Delta L = N/2$, (d) 图对应 $\Delta \chi^2_{\text{eff}}$, (e) 图对应 $\chi^2_b = M$ 。

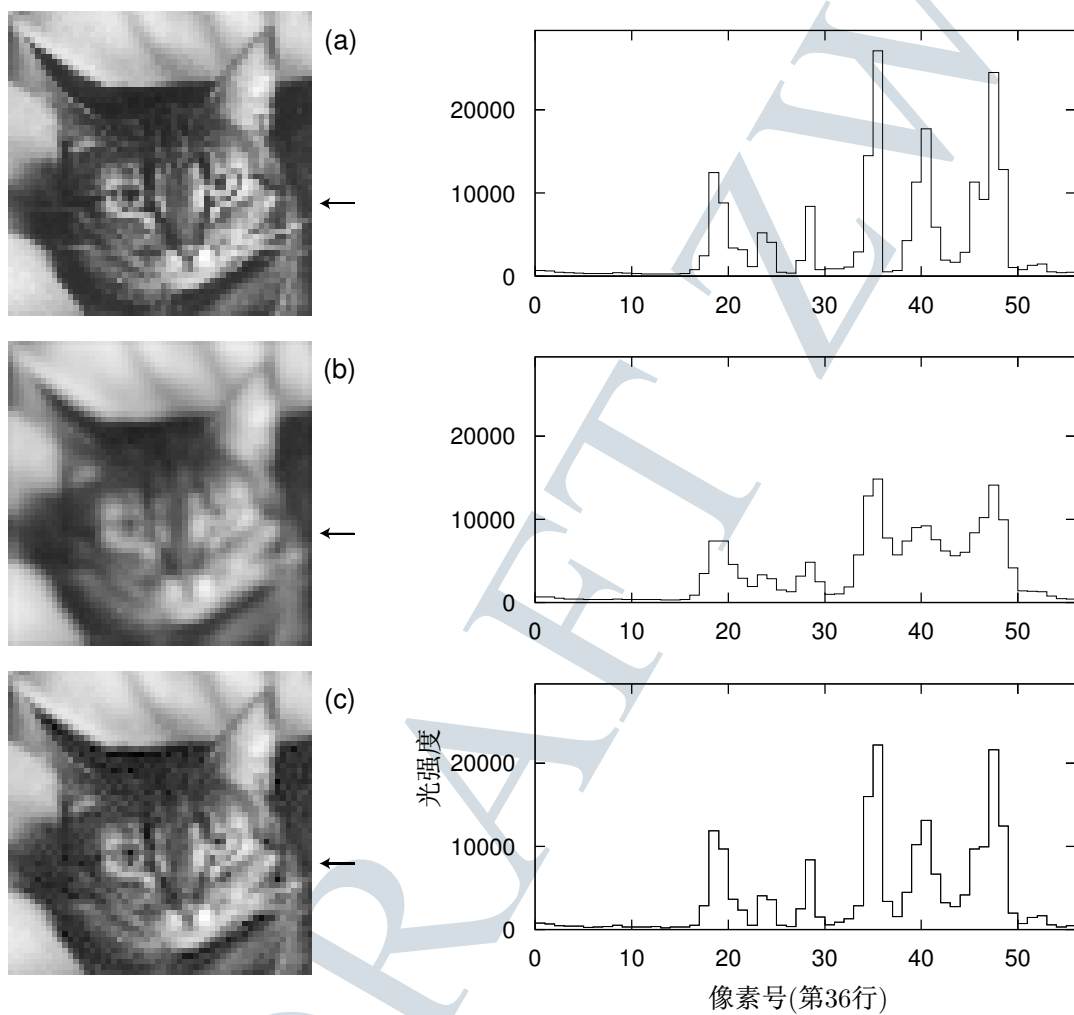


图 11.5: (a) 原始的“真实”图像 μ 。(b) 观测到的图像 \mathbf{n} , 经过标准差等于像素尺寸 60% 的高斯点扩散函数模糊化。(c) 最大熵解谱法得到的图像。右边的直方图给出了第 36 行像素的光强 (在左边用箭头指示的地方)。

泊松变量，较亮区域涨落的相对大小在 1% 的量级 ($\sigma_i/\nu_i = 1/\sqrt{\nu_i}$)。图11.5(c) 给出了 $\Delta \log L = N/2$ 时最大熵解谱法的结果，其中 $N = 3136$ 为像素点的数目。图11.5右边的直方图给出的是相应图片第 36 行像素的光强。

在这个特殊的例子中，最大熵正规化比 Tikhonov 正规化具有一定的优势。首先，前面已经提过，最大熵解谱法的一个性质是可以保证解谱后所有区间的频数都保持为正。除此之外，另外一个优点是，很容易将熵推广到多维分布。这一点也很明显，因为熵 $H = -\sum_j p_j \log p_j$ 只是对所有区间的求和，不需要关心相邻区间的相对值。对 Tikhonov 正规化，可以利用拉普拉斯算子的差分近似将正规化函数 $S(\mu)$ 推广到二维，参见文献 [22] 第 18 章。

由于熵与区间的相对位置无关，所以直方图中存在孤立峰的问题相对来说并不严重。天文学中，恒星会在图像中形成亮点，在直方图中对应尖峰，这就是天文学中喜欢用最大熵解谱法的原因。对于相对平滑的分布，如图11.3和图11.4，在给定偏倚的情况下，Tikhonov 正规化方法得到的方差要小很多。然而，如果分布中存在尖峰，如图11.5中的照片，情况就会有所不同。

最大熵方法的一个缺点是，它必然导致 μ 的非线性方程组。由于图片中的像素往往非常大，所以很难直接对矩阵求逆，最终都将采用数值迭代技术。

11.9 解谱法的数值实现

数值实现前面几节介绍的解谱方法并不容易。要想求函数

$$\varphi(\mu, \lambda) = \alpha \log L(\mu) + S(\mu) + \lambda \left[n_{\text{tot}} - \sum_{i=1}^N \nu_i \right], \quad (11.84)$$

对于 μ 和 λ 的最大值，意味着求解11.64式中的 $M+1$ 个方程。如果 φ 为 μ 的二次型函数，则 (11.64) 中的方程都是线性的。例如，如果数据服从高斯分布，则对数似然函数为 $\log L = -\chi^2/2$ ，并且采用 Tikhonov 正规化方法，那么对应的 φ 就是 μ 的二次型函数。这种情况下，基于奇异值分解的解谱法在 [74] 中讨论。如果 φ 中包含基于泊松分布的对数似然函数，或者基于熵的正规化函数，则得到的方程为非线性方程，必须通过数值迭代技术求解。

例如，考虑基于泊松分布的似然函数，参见11.21和11.22式，

$$\log L(\mu) = \sum_{i=1}^N (n_i \log \nu_i - \nu_i), \quad (11.85)$$

正规化函数采用 $S = H$ ，其中 H 为 (11.51) 式定义的熵。

图11.6说明了最大熵正规化求解的一种可能方法。图中的三个坐标轴表示 μ -空间的三个维度，对角平面是 $\sum_i \nu_i = n_{\text{tot}}$ 为常数的子空间。平面上标注的两个点分别表示最大熵 (所有 μ_i 都相等) 对应的点和极大似然对应的点。连接这两个点的曲线表示正规化参数 α 取不同值时 (11.40) 式的可能解，例如， $\alpha = 0$ 对应最大熵点， $\alpha \rightarrow \infty$ 对应极大似然点。这条曲线通过等熵线和等似然线接触的点。需要注意的是，极大似然点不在所有 $\nu_i > 0$ 的区域。这实际上是典型的极大似然振荡解，参见图11.1(d)。

图11.3和 (11.5) 展示的最大熵解的例子，其采用的计算机程序采用了下面的算法，该算法包含了文献 [18,67] 介绍的复杂方法的部分特性。极大似然点通常不能用作 μ 的初值，因为该点通常具有负值的 μ_i ，所以熵在该点没有定义。相反，最大熵点被当作 μ 的初值。最大熵点通过要求所有 μ_i 都相等并考虑下面

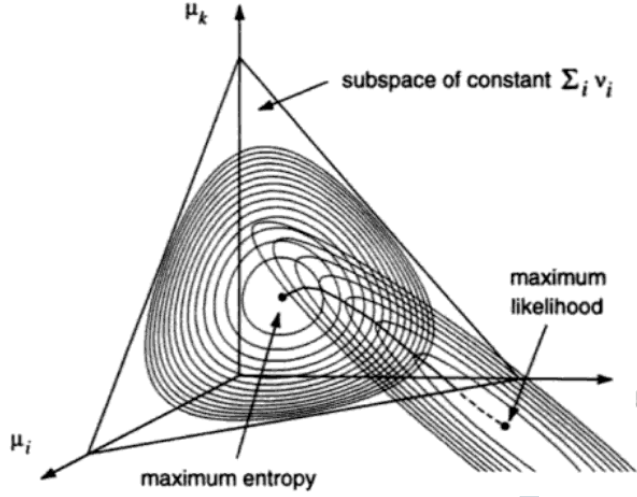


图 11.6: μ -空间的 3 个维度, 用来说明最大熵解谱法的数值实现 (详见本节内容)。

的约束条件来确定:

$$\begin{aligned}\nu_{\text{tot}} &= \sum_{i=1}^N \nu_i = \sum_{i=1}^N \sum_{j=1}^M R_{ij} \mu_j = \sum_{j=1}^M \epsilon_j \mu_j \\ &= n_{\text{tot}},\end{aligned}\quad (11.86)$$

其中 ϵ_j 为区间 j 的效率。因此, 最大熵点为

$$\mu_i = \frac{n_{\text{tot}}}{\sum_{j=1}^M \epsilon_j}.\quad (11.87)$$

如果利用 $S(\mu) = \mu_{\text{tot}} H$, 并且效率并不处处相等, 则最大的 $S(\mu)$ 不是均匀分布, 而是由 M 个方程的解给出,

$$\log \frac{\mu_i}{\mu_{\text{tot}}} + \frac{S(\mu) \epsilon_i}{n_{\text{tot}}} = 0, \quad i = 1, \dots, M.\quad (11.88)$$

以最大 $S(\mu)$ 的点为出发点, 沿着 ν_{tot} 为常数的子空间中 φ 最大的曲线走下去。只要保持在这个子空间中, 只需要最大化下面这个量,

$$\Phi(\mu) = \alpha \log L(\mu) + S(\mu),\quad (11.89)$$

即, 不考虑 $\varphi(\mu)$ 中的拉格朗日乘子项, 参见 (11.40) 式。然而, 简单地要求 $\nabla \Phi = 0$ 并不能给出期望的解。相反, 必须先将 $\nabla \Phi$ 投影到 ν_{tot} 为常数的子空间, 然后将得到的矢量分量设为零。采用这种方法, 拉格朗日乘子 λ 不会明显进入这个算法中。也就是说, 可以通过要求

$$D\Phi = \nabla \Phi - \mathbf{u}(\mathbf{u} \cdot \nabla \Phi) = 0\quad (11.90)$$

进行求解, 其中 \mathbf{u} 为 $\nabla \nu_{\text{tot}}$ 方向上的单位矢量, 由下式给出 (参见 (11.10) 式),

$$\frac{\partial \nu_{\text{tot}}}{\partial \mu_k} = \sum_{i=1}^N \sum_{j=1}^M R_{ij} \frac{\partial \mu_j}{\partial \mu_k} = \epsilon_k,\quad (11.91)$$

因此, 矢量 \mathbf{u} 只是由效率矢量归一化到单位长度给出,

$$\mathbf{u} = \frac{\boldsymbol{\epsilon}}{|\boldsymbol{\epsilon}|}.\quad (11.92)$$

如 (11.90) 式中定义，我们将用微分算子 D 表示梯度在 ν_{tot} 为常数的子空间的投影。

因此，从最大熵点出发，沿 $D \log L$ 方向取一个小步长，得到的 μ 通常不在 Φ 最大的曲线上，但是非常接近，只要选取的步长足够小。要考察 μ 离曲线的距离，可以计算 $|D\Phi|$ 。如果这个值超过一定的限制，则表明步长太大，需要重新选取一个更小的步长。

如果得到的 μ 实际上在最大 Φ 的曲线上，则应该满足 $\alpha D \log L + DS = 0$ ，对应的正规化参数为

$$\alpha = \frac{|DS|}{|D \log L|}. \quad (11.93)$$

参数 α 可以简单地取为 (11.93) 式的右手端，并选取一定的侧向步长使 μ 回到 $D\Phi = 0$ 的曲线。这可以通过求函数最大值的标准方法实现 (通常改成求极小值，参见文献 [3, 22])。这些侧向步长也要确保 μ 停留在 $\nu_{\text{tot}} = n_{\text{tot}}$ 的子空间中，也就是说，搜寻方向被投影到这个子空间。于是，以这种方式进行处理，通过 $D \log L$ 的向前步长增大 α ，并通过侧向步长移动到 $D\Phi = 0$ 的解上，直至达到 $\Delta \log L = \log L_{\text{max}} - \log L$ 的理想值。为了确定最佳终止点，中间结果可以保存下来留作检验。

尽管这个算法的基本思想也可以应用到 Tikhonov 正规化，但是在 Tikhonov 正规化中，情况有些复杂，因为 $S(\mu)$ 最大化的解不唯一。例如， $k = 2$ 时，任何线性函数都给出 $S = 0$ 。可以简单地从 $\mu_i = n_{\text{tot}}/M$ 出发，并取正规化参数 α 足够大，找到唯一的解。

Tikhonov 正规化也可能不使用 $\sum_i \nu_i = n_{\text{tot}}$ 的约束条件，因为正规化函数在这里并不倾向于将解拖向差别明显的总归一化。然而，如果忽略归一化条件，则无法严格得到 $\sum_i \hat{\nu}_i = n_{\text{tot}}$ 。我们可以讨论说， $\hat{\nu}_{\text{tot}}$ 应当是总事例数的无偏估计量，但是由于偏倚不大，所以通常不考虑这个约束条件。

DRAFT ZV

参考文献

- [1] A. N. Kolmogorov, Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer, Berlin (1933); English translation, Foundations of the Theory of Probability, 2nd edition, Chelsea, New York(1956).
- [2] G.R. Grimmett and D.R. Stirzaker, Probability and Random Processes, 2nd edition, Clarendon Press, Oxford(1992).
- [3] S. Brandt, Datenanalyse, 3rd edition, BI-Wissenschaftsverlag, Mannheim(1992); Statistical and Computational Methods in Data Analysis, Springer, New York (1997).
- [4] Geoffrey Grimmett and Dominic Welsh, Probability: an Introduction, Clarendon Press, Oxford(1986).
- [5] T. Bayes, An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc.**53**(1763)370. Reprinted in Biometrika,**45**(1958) 293.
- [6] R.A. Fisher, Statistical Methods, Experimental Design and Scientific Inference, a re-issue of Statistical Methods for Research Workers, The Design of Experiments, and Statistical Methods and Scientific Inference, Oxford University Press, Oxford(1990).
- [7] Alan Stuart and J. Keith Ord, Kendall's Advanced Theory of Statistics, Vol. I, Distribution Theory, 6th edition, Edward Arnold, London (1994); Vol. II, Classical Inference and Relationships, 5th edition, Edward Arnold, Londn (1991).
- [8] Harld Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, New Jersey (1946).
- [9] Richard von Mises, Wahrscheinlichkeit, Statistik und Wahrheit, 2nd edition, Springer, Vienna (1951); Probability, Statistics and Truth, Allen and Unwin, London (1957).
- [10] Richard von Mises, Mathematical Theory of Probability and Statistics, edited and complemented by Hilda Geiringer, Academic Press, New York (1964).
- [11] Harold Jeffreys, Theory of Probability, 2nd edition, Oxford University Press, London (1948).
- [12] Leonard J. Savage, The Foundations of Statistics, 2nd revised edition, Dovedr, New York (1972).
- [13] Bruno de Finetti, Theory of Probability: A Critical Introductory Treatment (2 volumes), John Wiley, New York (1974).
- [14] R.T. Cox, Probabilityl, frequency and reasonable expectation, Am. J. Phys. **14** (1946).
- [15] D.V. Lindley, Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 1: Probability, Part 2: Inference, Cambridge University Press, cambridge (1965).

- [16] A. O'hagan, Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference, Edward Arnold, London (1994).
- [17] Peter M. Lee, Bayesian Statistics: an Introduction, Edward Arnold, London (1989).
- [18] D.S. Sivia, Data Analysis, a Bayesian Tutorial, Clarendon Press, Oxford (1996).
- [19] M.D. Springer, The Algebra of Random Variables, John Wiley, New York (1979).
- [20] A.G. Frodesen, O. Skjeggstad and H. Tøfte, Probability and Statistics in Particle Physics, Universitetsforlaget, Oslo (1979).
- [21] George B. Arfken and Hans-Jurgen Weber, Mathematical Methods for Physicists, 4th edition, Academic Press, New York (1995).
- [22] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes, 2nd edition, Cambridge University Press, Cambridge (1992).
- [23] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, Statistical Methods in Experimental Physics, North-Holland, Amsterdam (1971).
- [24] Edward J. Dudewicz and Satya N. Mishra, Modern Mathematical Statistics, John Wiley, New York (1988).
- [25] CERN Program Library, CERN, Geneva, (1997). For software and documentation see <http://www.cern.ch/CERN/Computing.html> and links therein.
- [26] ROOT package, CERN, Geneva, (2009). For software and documentation see <http://root.cern.ch/drupal/> and links therein.
- [27] L. Landau, On the energy loss of fast particles by ionisation, J. Phys. USSR **8** (1944) 201.
- [28] K.S. Koelbig and B. Schorr, A program package for the Landau distribution, Comput. Phys. Commun. **31** (1984) 97.
- [29] H.D. Maccabee and D.G. Papworth, Correction to Landau's energy loss formula, Phys. Lett. **30A** (1969) 241.
- [30] W. Allison and J. Cobb, Relativistic Charged Particle Identification by Energy Loss, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.
- [31] G. Marsaglia and A. Zaman, A new class of random number generators, Ann. Appl. Probab. **1** (1991) 462.
- [32] Martin Lüscher, A portable high-quality random number generator for lattice field theory simulations, Comput. Phys. Commun. **79** (1994) 100.
- [33] F. James, RANLUX: A Fortran implementation of the high-quality pseudorandom number generator of Lüscher, Comput. Phys. Commun. **79** (1994) 111.
- [34] F. James, A review of pseudorandom number generators, Comput. Phys. Commun. **60** (1990) 329.
- [35] P.L. L'Ecuyer, Efficient and portable combined random number generators, Commun. ACM **31** (1988) 742.

- [36] John von Neumann, Various techniques used in connection with random digits, J.Res. NBS Appl. Math. Ser. **12** (1951) 36; reprinted in John von Neumann: Collected Works, A.H. Taub (ed.), Vol. V, Pergamon Press, Oxfor (1963).
- [37] Donald H. Perkins, Introduction to High Energy Physics, Addison-Wesley, Menlo Park, California (1987).
- [38] F. James, Monte Carlo theory and practice, Rep. Prog. Phys. **43** (1980) 1145.
- [39] Christopher M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford (1995).
- [40] John Hertz, Anders Krogh and Richard G. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, New York (1991).
- [41] Berndt Müller, Michael T. Strickland and Joachim Reinhardt, Neural Networks: and Introduction, 2nd edition, Springer, Berlin (1995).
- [42] Leif Lönnblad, Carsten Peterson and Thorsteinn Rognvaldsson, Pattern recognition in high energy physics with artificial neural networks-JETNET 2.0, Comput. Phys. Commun. **79** (1992) 167.
- [43] Carsten Peterson and Thorsteinn Rognvaldsson, An Introduction to Artificial Neural Networks, in Proceedings of the 1991 CERN School of Computing, C. Verkerk (ed.), CERN 92-02 (1992).
- [44] Wayne S. Babbage and Lee F. Thompson, The use of neural networks in $\gamma - \pi^0$ discrimination, Nucl. Instrum. Methods **A330** (1993) 482.
- [45] R. Gnanadesikan (ed.), Discriminant Analysis and Clustering, National Academy Press, Washington, DC (1988).
- [46] J.O. Berger and D.A. Berry, Statistical analysis and the illusion of objectivity, Am. Sci. **76** No. 2 (1988) 146.
- [47] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. **7** (1936) 179; reprinted in Contributions to Mathematical Statistics, John Wiley, New York (1950).
- [48] F. James, and M. Roos, CERN Program Library routine D506(long write-up), 1989; F. James, Interpretation of the Errors on Parameters as given by MINUIT, supplement to long write-up of routine D506(1978).
- [49] R.J. Muirhead, Aspects of Multivariate Statistical Theory, John Wiley, New York (1982).
- [50] Steve Baker and Robert D. Cousins, Clarification of the use of the chi-square and likelihood functions in fits to histograms, Nucl. Instrum. Methods **221** (1984) 437.
- [51] The Particle Data Group, Review of particle properties, Phys. Rev. **D54** (1996) 1.
- [52] L. Lyons, Statistics for Nuclear and Particle Physicists, Cambridge University Press, Cambridge (1986).
- [53] G. D'Agostini, On the use of the covariance matrix to fit correlated data, Nucl. Instrum. Methods **346** (1994) 306.
- [54] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of probability, Philos. Trans. **A 236** (1937) 333.
- [55] D.J. Hudson, Statistics Lectures II: Maximum Likelihood and Least Squares Theory, CERN 64-18 (1964).

- [56] V.L. Highland, Estimation of upper limits from experimental data, Temple University Note COO-3539-38 (1983).
- [57] F. James and M. Roos, Statistical notes on the problem of experimental observations near an unphysical region, Phys. Rev. **D44** (1991) 299.
- [58] O. Helene, Upper limited of peak area, Nucl. Instrum. Methods **212** (1983) 319.
- [59] Robert D. Cousins, Why isn't every physicist a Bayesian?, Am. J. Phys. **63** (1995) 398.
- [60] Robert D. Cousins and Virgil L. Highland, Incorporating systematic uncertainties into an upper limit, Nucl. Instrum. Methods **A212** (1992) 331.
- [61] Carsten Peterson, Thorsteinn Rögnvaldsson and Leif Lönnblad, JETNET3.0-A versatile artificial neural network package, Comput. Phys. Commun. **81** (1994) 185.
- [62] V.B. Anykeyev, A.A. Spiridonov and V.P. Zhigunov, Comparative investigation of unfolding methods, Nucl. Instrum. Methods **A303** (1991) 350.
- [63] V.B. Anykeyev, A.A. Spiridonov and V.P. Zhigunov, Correcting factors Method as an unfolding technique, Nucl. Instrum. Methods **A322** (1992) 280.
- [64] E.A. Belgorlov and V.P. Zhigunov, Interpretation of the solution to the inverse problem for the positive function and the reconstruction of neutron spectra, Nucl. Instrum. Methods **A235** (1985) 146.
- [65] V.P. Zhigunov, Improvement of resolution function as an inverse problem, Nucl. Instrum. Methods **A216** (1983) 183.
- [66] V.P. Zhigunov T.B. Kostkina and A.A. Spiridonov, On estimating distributions with the maximum entropy principle, Nucl. Instrum. Methods **A273** (1988) 362.
- [67] John Skilling and S.F. Gull, Algorithms and applications, in Maximum Entropy and Bayesian Methods in Inverse Problems, C. Ray Smith and W.T. Grady (ed.), D. Reidel, Dordrecht (1985) 83.
- [68] John Skilling, Theory of maximum entropy image reconstruction, in Maximum Entropy and Bayesian Methods in Applied Statistics, J.H. Justice (ed.), Cambridge University Press, Cambridge (1986) 156.
- [69] E.T. Jaynes, Monkeys, kangaroos and N , in Maximum Entropy and Bayesian Methods in Applied Statistics, J.H. Justice (ed.), Cambridge University Press, Cambridge (1986) 26.
- [70] D.L. Phillips, A technique for the numerical Solution of certain integral equations of the first kind, J. ACM **9** (1962) 84.
- [71] A.N. Tikhonov, On the solution of improperly posed problems and the method of regularization, Sov. Math. **5** (1963) 1035.
- [72] A.N. Tikhonov and V.Ya. Arsenin, Solutions of Ill-Posed Problems, John Wiley, New York (1977).
- [73] V. Blobel, Unfolding methods in high energy physics experiments, in Proceedings of the 1984 CERN School of Computing, CERN 85-09 (1985).

- [74] Andreas Höcker and Vakhtang Kartvelishvili, SVD approach to data unfolding, Nucl. Instrum. Methods **A372** (1996) 469.
- [75] Byron P. Roe, Probability and Statistics in Experimental Physics, Springer, New York (1992).
- [76] G. Zech, Comparing statistical data to Monte Carlo simulation - parameter fitting and unfolding, DESY 95-113 (1995).
- [77] C.E. Shannon, A mathematical theory of communication, Bell Sys. Tech. J. **27** (1948) 379, 623. Reprinted in C.E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana (1949).
- [78] S. Kullback, Information Theory and Statistics, John Wiley, New York (1964).
- [79] E.T. Jaynes, Prior probabilities, IEEE Trans. Syst. Sci. Cybern. **SSC-4** (1968) 227.
- [80] Michael Schmelling, The method of reduced cross-entropy. A general approach to unfold probability distributions, Nucl. Instrum. Methods **A340** (1994) 400.
- [81] G. Wahba, Smoothing and ill-posed problems, in Solution Methods for Integral Equations, Michael Goldberg (ed.), Plenum Press, New York (1979).
- [82] B.R. Frieden, Restoring with maximum likelihood and maximum entropy, J. Opt. Soc. Am. **62** (1972) 511.
- [83] B.R. Frieden, Statistical models for the image restoration problem, Comput. Graphics Image Process. **12** (1980) 40.
- [84] B.R. Frieden, Probability, Statistical Optics, and Data Testing, Springer, New York (1983).
- [85] R. Narayan and R. Nityananda, Maximum entropy image restoration in astronomy, Ann. Rev. Astron. Astrophys. **24** (1986) 127.
- [86] A.K. Louis, Medical imaging: state of the art and future development, Inverse Probl. **8** (1992) 709.

索引

- N 维高斯分布, 24, 26
- N 维高斯概率密度函数, 73
- P -值, 45, 46, 79, 94
- 备择假设, 37, 45
- 贝叶斯定理, 2, 3, 5, 10, 39, 42, 71, 105, 107, 108, 128, 130
- 贝叶斯方法, 107, 124, 128
- 贝叶斯概率, 4
- 贝叶斯估计量, 71, 129
- 贝叶斯假设, 72
- 贝叶斯区间, 105
- 贝叶斯统计, 5, 45, 71, 117
- 贝叶斯置信区间, 91
- 本底, 120
- 边缘概率密度函数, 8, 12, 62
- 变量变换法, 32
- 标准差, 13, 34, 58, 68, 91
- 标准高斯分布, 23
- 不完全伽马函数, 107
- 布莱特-魏格纳分布, 27
- 参考分布, 130, 133
- 参数变换, 72
- 参数估计, 69, 124
- 参数拟合, 51
- 残差, 135
- 乘法线性同余算法, 31
- 抽样分布, 52
- 纯度, 39
- 大数弱定律, 53
- 大样本极限, 59–61, 91, 95, 98, 100, 101
- 代数矩, 13, 110
- 单侧置信区间, 93, 95
- 单层感知器, 43
- 第二类误差, 38
- 第一类误差, 38
- 点扩散函数, 119
- 独立, 2, 10, 14
- 对数似然函数, 56, 123
- 对数正态分布, 25
- 多维高斯分布, 91
- 多维随机变量, 7
- 多项分布, 20, 67, 118, 128
- 二项分布, 19, 45, 109
- 反傅立叶变换, 110, 113
- 反卷积, 117
- 范围概率, 93
- 方差, 13, 56, 58
- 非正当验前概率, 72
- 费舍尔信息矩阵, 60
- 费舍尔甄别函数, 41
- 分辨率, 104, 117
- 分辨率函数, 119, 121, 135
- 分位数, 7, 24, 95, 100, 102–104
- 傅立叶变换, 12, 109
- 傅立叶卷积, 12
- 复合假设, 37, 80
- 负相关, 14, 122
- GEANT, 35
- 概率, 1
- 概率函数, 3
- 概率密度函数, 5, 10
- 高斯-马可夫定理, 74, 82, 123
- 高斯分布, 23, 58, 60, 62, 91, 95, 109, 110
- 高斯随机变量, 73

功效, 38, 40
估计量, 34, 51
估计值, 51
观测显著性水平, 45
光滑度, 125, 126, 128

激活函数, 43
积分变换, 12
极大似然法, 55, 71, 78
极大似然估计量, 55, 106
伽马分布, 113
伽马函数, 26
加权的均方差, 133
加权均方差, 135
加权平均, 71, 81, 83
假设, 4, 37
检验功效, 38
检验统计量, 37, 40, 69, 78, 93
简单假设, 37
渐进无偏估计量, 52
渐进正态性, 60
交叉熵, 130, 133
交叉熵原理, 130
交叉验证, 134
接受度, 118
接受域, 37
解谱, 117, 119
解谱法, 52
经典统计, 4, 117
经典置信区间, 91
矩, 110
矩方法, 87
矩方法估计量, 87
拒绝域, 37, 40
卷积, 12
均方差, 52, 133
均匀分布, 22, 109

卡方, 74
卡方分布, 26, 48, 70, 74, 79, 97, 102, 109, 111
柯西分布, 27, 53, 109, 110
可选终止, 45

扩展的似然函数, 64, 108

朗道分布, 27, 112
累积分布, 6, 22, 24, 32, 92, 95, 96
离散随机变量, 97
联合概率密度函数, 7, 12, 71
链接强度, 43
留数定理, 113

梅林变换, 12
梅林卷积, 12
蒙特卡罗, 62
蒙特卡罗方法, 31, 48, 59
模式识别, 43

奈曼-皮尔逊引理, 40
拟合优度, 37, 44, 76, 78, 79
拟合优度检验, 26, 44, 47, 69, 93
逆问题, 52, 117

判选条件, 38, 40, 124, 132
皮尔逊卡方统计量, 47
偏倚, 52, 56, 60, 67, 78, 114, 124, 130, 132, 135
频数, 6, 67
泊松变量, 64, 96, 106, 118
泊松分布, 21, 77, 97, 109

期望值, 12
奇异值分解, 17, 139
奇异值分析, 134
前馈网络, 43
区间, 6
去模糊化, 117
全概率定律, 2, 10, 119
权重, 43

RCF 边界, 75, 102, 123
RCF 不等式, 60, 100, 102, 123
人工神经网络, 43

Shannon-Jaynes 熵, 130
Shannon 熵, 128, 130
散点图, 7
熵, 128

上限, 104, 106, 107
舍选法, 32
神经网络, 43
事件, 4
事例产生子, 35
似然, 5
似然比, 40, 42
似然函数, 55, 60, 118, 129
似然区间, 100
输出节点, 43
输入层, 43
双层感知器, 43
斯特令近似, 128
算术平均, 87
随机变量, 1, 10
随机数产生子, 31
随机数序列, 31
随机性, 1

Tikhonov 解谱法, 134
Tikhonov 正规化, 126, 133–135, 141
探测效率, 65
特征函数, 109
条件概率, 2, 8, 119
条件概率密度函数, 8, 71
统计不确定度, 53, 59, 61, 91, 101
统计检验, 37, 94
统计量, 51, 123
统计误差, 80, 89, 91, 135
图像还原, 135

upper limits, 106

网络学习, 44
网络训练, 44
伪随机数, 31
无偏估计量, 52, 57, 114, 121
无偏性, 123
误差, 16
误差棒, 46, 91, 94, 101, 122, 135
误差传递, 12, 16, 17, 39, 75, 82, 88, 89, 91, 124, 130–132
误差函数, 44

误差矩阵, 13
误差逆传播, 44

系统误差, 120
下限, 104
先验概率, 3
显著水平, 79, 94
显著性水平, 37, 40
线性最小二乘估计量, 74
相对频率, 3, 5, 94
相对频率极限, 4
相对误差, 122
相关, 85
相关系数, 14, 25, 53, 63, 76, 82, 83, 89, 98, 122
相关系数矩阵, 85
相关性, 44, 84, 123
相合, 51
相合估计量, 51
相加定理, 111
相应矩阵, 117
响应函数, 119
响应矩阵, 119
响应矩阵求逆法, 134
协方差, 13, 53
协方差矩阵, 13, 16, 82, 123
信息不等式, 60
修正的最小二乘法, 77
修正因子, 117, 124
选择效率, 38, 40
学生分布, 96

雅可比行列式, 12
验后概率, 3, 5, 42, 124
验后概率密度, 71
验后概率密度函数, 104
验前概率, 3, 5, 42, 45, 108, 124
验前概率密度, 71, 129
验前概率密度函数, 105
样本, 51
样本标准差, 59, 63, 67
样本方差, 53
样本概率密度函数, 108

样本矩, 118
样本均值, 52, 56, 58, 59, 63, 85, 87, 112
样本空间, 1, 51
样本容量, 85, 100
样本相关系数, 63
样本协方差, 63
隐藏层, 43
有偏估计量, 58, 66, 124
有限分辨率, 117
有效估计量, 60, 85
有效性, 123
元输出, 4
原假设, 37

正规化, 125
正规化参数, 126, 129, 134, 140
正规化函数, 125, 128–131
正规化解谱, 117
正规化解谱法, 133
正交变换, 17, 76
正相关, 14, 62, 83
直方图, 6
指数分布, 23, 56, 109, 112
置信带, 92
置信区间, 59, 91, 93, 99, 106, 112, 114
置信区域, 75, 101, 102
置信水平, 45, 93, 95, 99, 103, 112
中位数, 7
中心极限定理, 24, 25, 73, 95, 111, 114
中心矩, 13, 54
中心置信区间, 61, 93–95, 100, 105, 114
种子, 31
众数, 7, 71, 72, 105
主观概率, 3, 4, 71
子样本, 71
总体方差, 13
总体平均, 12
最大熵解谱法, 134, 135
最大熵原理, 128
最大熵正规化, 134, 135
最小二乘法, 26, 71, 73, 118, 123, 132
最小二乘估计量, 74, 81

最小二乘拟合, 76
最小二乘原则, 42
最小方差边界, 78
最小方法边界, 60
最小加权均方差, 135
最小均方差, 135