PKUHEP-2021-001
2025 年 3 月 8 日

统计数据分析习题

杨振伟

北京大学高能物理研究中心

内部资料，仅供教学使用。

目录

第一章　基本概念

习题 1.1. 考虑某样本空间 S 以及给定子空间 B，并假设 $P(B) > 0$。证明条件概率

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.1}$$

满足概率的公理。

习题 1.2. 证明

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

*(*提示：将 $A \cup B$ 表示成 3 个不相交的子集的并。*)*

习题 1.3. 证明德摩根律（又见集合运算）的一般形式。设 $\{A_\alpha : \alpha \in \Gamma\}$ 是一列集合（可能是不可数多个），证明：

(a) $(\cup_\alpha A_\alpha)^c = \cap_\alpha A_\alpha^c$；

(b) $(\cap_\alpha A_\alpha)^c = \cup_\alpha A_\alpha^c$；

习题 1.4. 如果一个随机现象的样本空间 Ω 充满某个区域，其度量（长度、面积或体积等）大小可用 S_Ω 表示。任意一点落在度量相同的子区域内（可能位置不同）是等可能的。若事件 A 为 Ω 中的某个子区域，且其度量大小可以用 S_A 表示。则事件 A 的概率为

$$P(A) = \frac{S_A}{S_\Omega}$$

这个概率称为几何概率。试证明几何概率满足概率公理化定义。

习题 1.5. 某粒子束流包含 10^{-4} 的电子，其余为光子。粒子通过某双层探测器，可能在 *2* 层都给出信号，也可能只有一层给出信号或者没有任何信号。电子 (e) 和光子 (γ) 在穿过该双层探测器给出 0，1 或 2 个信号的概率如下

$$P(0|e) = 0.001 \qquad P(0|\gamma) = 0.99899$$
$$P(1|e) = 0.01 \qquad P(1|\gamma) = 0.001$$
$$P(2|e) = 0.989 \qquad P(2|\gamma) = 10^{-5}$$

(a) 如果只有一层给出信号，该粒子为光子的概率是多少？

(b) 如果两层都给出了信号，该粒子为电子的概率是多少？

习题 1.6. 假设随机变量 x 的概率密度函数为 $f(x)$。证明 $y = x^2$ 的概率密度函数为

$$g(y) = \frac{1}{2\sqrt{y}} f(\sqrt{y}) + \frac{1}{2\sqrt{y}} f(-\sqrt{y}). \tag{1.2}$$

习题 1.7. 假设两个独立的随机变量 x 和 y 都服从 *0* 到 *1* 之间的均匀分布，即概率密度函数 $g(x)$ 为

$$g(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{其它}, \end{cases} \tag{1.3}$$

概率密度函数 $h(y)$ 与 $g(x)$ 类似。

(a) 利用 *Statistical Data Analysis* 中的 *(1.35)* 式，证明，$z = xy$ 的概率密度函数 $f(z)$ 为

$$f(z) = \begin{cases} -\log z & 0 < z < 1, \\ 0 & \text{其它}. \end{cases} \tag{1.4}$$

(b) 利用 *Statistical Data Analysis* 的 *(1.37)* 和 *(1.38)* 式，通过另外定义一个函数 $u = x$，求 $z = xy$ 的概率密度函数。首先求 z 和 u 的联合概率密度函数，然后对 u 进行积分求出 z 的概率密度函数。

(c) 证明 z 的累积分布为

$$F(z) = z(1 - \log z). \tag{1.5}$$

习题 1.8. 考虑随机变量 x 与常数 α 和 β。证明

$$\begin{aligned} E[\alpha x + \beta] &= \alpha E[x] + \beta, \\ V[\alpha x + \beta] &= \alpha^2 V[x]. \end{aligned} \tag{1.6}$$

习题 1.9. 考虑两个随机变量 x 和 y。

(a) 证明 $\alpha x + y$ 的方差为

$$\begin{aligned} V[\alpha x + y] &= \alpha^2 V[x] + V[y] + 2\alpha \mathrm{cov}[x, y] \\ &= \alpha^2 V[x] + V[y] + 2\alpha\rho\sigma_x\sigma_y, \end{aligned} \tag{1.7}$$

其中 α 为任意常数，$\sigma_x^2 = V[x]$，$\sigma_y^2 = V[y]$，关联系数 $\rho = \mathrm{cov}[x, y]/\sigma_x\sigma_y$。

(b) 利用 *(a)* 的结果，证明关联系数总是位于区间 $-1 \le \rho \le 1$。*(*利用 $V[\alpha x + y]$ 的方差总是大于或等于零。*)*

习题 1.10. 假设随机变量 $\mathbf{x} = (x_1, \ldots, x_n)^T$ 用联合概率密度函数 $f(\mathbf{x})$ 描述，而变量 $\mathbf{y} = (y_1, \ldots, y_n)^T$ 由下面的线性变换定义

$$y_i = \sum_{j=1}^{n} A_{ij} x_j. \tag{1.8}$$

假设反变换 $\mathbf{x} = A^{-1}\mathbf{y}$ 存在。

(a) 证明 $\mathbf{y}$ 的联合概率密度函数为

$$g(\mathbf{y}) = f(A^{-1}\mathbf{y}) |\det(A^{-1})|. \tag{1.9}$$

(b) 当 A 为矩阵，即 $A^{-1} = A^T$ 时，求 $g(\mathbf{y})$。

习题 1.11. 若多元函数有连续偏导数，则求导的先后次序对求导结果没有影响。例如，对于二元函数 $f(x, y)$，下面两个三阶偏导相等:

$$\frac{\partial^3}{\partial x^2 \partial y} f(x, y) = \frac{\partial^3}{\partial y \partial x^2} f(x, y).$$

(a) 三元函数有多少个四阶偏导?

(b) 证明 n 元函数有 $\binom{n+r-1}{r}$ 个 r 阶偏导。

习题 1.12. A 和 B 两人分别轮流掷一枚硬币，最先掷得正面朝上的人胜出，假定 A 先开始掷。

(a) 如果所掷的硬币是公平硬币，A 胜出的概率是多少?

(b) 假设 $P(\text{正面朝上}) = p$，p 可能不等于 $\frac{1}{2}$。A 胜出的概率是多少?

(c) 证明：对任意的 p，$0 < p < 1$，$P(A \text{ 胜出}) > \frac{1}{2}$。

习题 1.13. 假设 X 和 Y 是两个连续的随机变量，且其方差有限。证明相关系数 $\rho \equiv \frac{\text{cov}(X,X)}{\sigma_X \sigma_Y} = \pm 1$ 的充要条件是 X 与 Y 几乎处处有线性关系，即存在常数 $a \neq 0$ 和 b 使得 $Y = aX + b$。

习题 1.14. 设 (X, Y) 是二维连续随机变量，且，且 $E(X)$ 存在。证明：$E(X) = E[E[X|Y]]$，其中 $E[X|Y]$ 是给定 Y 的条件下 X 的数学期望。

习题 1.15. 考虑两个连续随机变量 X 和 Y，其联合概率密度函数为

$$f(x,y) = \begin{cases} \frac{1}{\pi R^2}, & x^2 + y^2 \leq R^2, \\ 0, & \text{其他} \end{cases}$$

(a) 求边缘概率密度函数 $f_X(x)$ 和 $f_Y(y)$。

(b) 求条件概率密度函数 $f(x|Y = y)$ 和 $f(y|X = x)$。

(c) 验证所得到的条件概率密度函数和边缘概率密度函数满足贝叶斯定理。

(d) 判断 X 和 Y 是否相互独立。

第二章　常用概率密度函数

习题 2.1. 考虑 N 个服从多项分布的随机变量 $\mathbf{n} = (n_1, \ldots, n_N)$，概率为 $\mathbf{p} = (p_1, \ldots, n_N$，并且总试验次数为 $n_{\text{tot}} = \sum_{i=1}^{N} n_i$。假设变量 k 定义为前 M 个 n_i 之和，

$$k = \sum_{i=1}^{M} n_i, \qquad M \leq N. \tag{2.1}$$

利用误差传递以及多项分布的协方差

$$\text{cov}[n_i, n_j] = \delta_{ij} n_{\text{tot}} p_i (1 - p_i) + (\delta_{ij} - 1) p_i p_j n_{\text{tot}}, \tag{2.2}$$

求 k 的方差。证明该方差等于 $p = \sum_{i=1}^{M} p_i$ 并且总试验次数为 n_{tot} 的二项分布的方差。

习题 2.2. 假设随机变量 x 均匀分布于区间 $[\alpha, \beta]$，$\alpha, \beta > 0$。求 $1/x$ 的期待值，并将结果与 $1/E[x]$ 进行比较，取 $\alpha = 1, \beta = 2$。

习题 2.3. 考虑指数分布

$$f(x; \xi) = \frac{1}{\xi} e^{-x/\xi}, \qquad x \geq 0. \tag{2.3}$$

(a) 证明对应的累积分布函数为

$$F(x) = 1 - e^{-x/\xi}, \qquad x \geq 0. \tag{2.4}$$

(b) 证明给定 $x > x_0$ 时 x 处于 x_0 与 $x_0 + x'$ 之间的条件概率 等于 x 小于 x' 的概率 *(*非条件概率*)*，即

$$P(x \leq x_0 + x' | x \geq x_0) = P(x \leq x'). \tag{2.5}$$

(c) 产生于大气上层的宇宙线 μ 子进入海平面的探测器，其中的一部分在探测器中停止并衰变。进入探测器与衰变的时间差 t 服从指数分布，t 的均值等于 μ 子的平均寿命 *(*近似为 2.2μS*)*。解释为什么 μ 子进入探测器前存活的时间对确定平均寿命没有影响。

习题 2.4. 假设 y 服从均值为 μ，方差为 σ^2 的高斯分布。

(a) 证明

$$x = \frac{y - \mu}{\sigma} \tag{2.6}$$

服从标准高斯分布$\varphi(x)$ *(*即，均值为零，方差为 *1)*。

(b) 证明累积分布函数$F(y)$ 与 $\Phi(x)$ 相等，即

$$F(y) = \Phi(\frac{y - \mu}{\sigma}). \tag{2.7}$$

习题 2.5.

(a) 证明，如果 y 服从均值为 μ 方差为 σ^2 的高斯分布，则 $x = e^y$ 服从对数正态分布，

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}\exp\left(\frac{-(\log x-\mu)^2}{2\sigma^2}\right). \tag{2.8}$$

(b) 通过积分

$$\begin{aligned} E[x] &= \int x f(x;\mu,\sigma^2)dx, \\ V[x] &= \int (x-E[x])^2 f(x;\mu,\sigma^2)dx. \end{aligned} \tag{2.9}$$

求 x 的均值与方差。

(c) 将 *(b)* 中求得的方差与通过误差传递 *($V[y]=\sigma^2$)* 得到的近似结果进行比较。在什么条件下误差传递近似成立？*(*注意 y 是无量纲的，因而 σ^2 也是无量纲的。*)*

习题 2.6. 证明自由度为 n 的 χ^2 分布的累积分布函数可以表示为

$$F_{\chi^2}(x;n) = P(\frac{x}{2},\frac{n}{2}), \tag{2.10}$$

其中 P 为不完全伽马函数 *(incomplete gamma function)*

$$P(x,n) = \frac{1}{\Gamma(n)}\int_0^x e^{-t}t^{n-1}dt. \tag{2.11}$$

习题 2.7. 设随机变量 X 服从参数为 ν 的泊松分布，求 $E[X]$ 和 $V[X]$。

习题 2.8. 假设 X 和 Y 分别服从参数为 ν_x 和 ν_y 的泊松分布，且相互独立。求 $Z = X+Y$ 的概率质量函数。

第三章　蒙特卡罗方法

习题 3.1. 写一段小程序，利用 ROOT 中的随机数产生子产生 10^4 个 $(0,1]$ 之间均匀分布 的随机数，将并结果画到直方图中（分 100 个区间）。

习题 3.2. 修改习题 *3.1* 中的直方图，使其只有 *5* 个区间，事例数 $N = 100$。产生的直方图可以看做对矢量 $(n_1, \ldots, n_5)$ 的观测，该矢量满足参数 $N = 100$，$p_i = 0.2 (i = 1, \ldots, 5)$ 的多项分布。

(a) 将上面的程序代码放到循环中产生直方图，重复该 *MC* 实验 100 次，每次用不同的随机数种子。只要保证每次实验程序不中断，程序在下一次实验（即下一个循环）时自动更新种子。定义一个直方图，将每次实验第 i 个区间（比如 $i = 3$）的值填充其中。这应该服从均值为 $Np_i = 20$，标准偏差为 $\sqrt{Np_i(1-p_i)} = 4$ 的二项分布。

(b) 为任意两个区间的值 n_i 和 n_j 作出散点图（二维直方图）。其协方差的理论值为 $\mathrm{cov}[n_i, n_j] = -Np_ip_j = -4$，或者说关联系数 $\rho = -4/4^2 = -0.25$。

习题 3.3. 考虑锯齿分布

$$f(x) = \begin{cases} \frac{2x}{x_{\max}^2} & 0 < x < x_{\max}, \\ 0 & \text{其它。} \end{cases} \tag{3.1}$$

(a) 参考课本 *3.2* 节，利用变换方法寻找函数 $x(r)$，以产生服从 $f(x)$ 的随机数。用程序实现该方法，并生成一个直方图。（可以取 $x_{\max} = 1$。）

(b) 参考课本 *3.3* 节，编写一段程序，利用舍选法产生服从锯齿分布的随机数。画出相应的直方图。

习题 3.4. 该练习的目的是产生服从高斯分布的随机数。有很多算法可以产生高斯分布，一个最简单的适合教学目的的算法基于中心极限定理：当 n 很大时，n 个随机数之和趋向于高斯分布，只要其中任何一项不占绝对份额（参考课本第 *10* 章）。

(a) 假设 x 均匀分布于 $[0,1]$，考虑 n 个独立随机变量 x 之和，

$$y = \sum_{i=1}^{n} x_i. \tag{3.2}$$

证明 y 的期待值为 $n/2$，方差为 $n/12$。并证明变量 z

$$z = \frac{\sum_{i=1}^{n} x_i - \frac{n}{2}}{\sqrt{n/12}} \tag{3.3}$$

均值为零，标准偏差为 *1*。

(b) 写一段小程序产生 *(a)* 中定义的变量 z，n 可以为任意值。分别取 $n = 1, \ldots, 20$，生成直方图，每个直方图包含 10^4 个 z 的值。何时 z 的分布近似为高斯分布？作为简单的高斯分布产生子，可以取 $n = 12$。评论此算法的局限性。选作：对于 $n = 2$，推导出 z 的概率密度函数的明显形式。

习题 3.5. 变量 t 服从均值 $\tau = 1$ 的指数分布，x 服从均值 $\mu = 0$，标准偏差 $\sigma = 0.5$ 的高斯分布。写一段 *MC* 程序，产生变量

$$y = t + x. \tag{3.4}$$

这里 t 可以表示不稳定粒子的真实衰变时间，x 表示测量误差，所以 y 表示测量到的衰变时间。画出 y 的直方图。修改程序以研究 $\tau \ll \sigma$ 和 $\tau \gg \sigma$ 的情形。

习题 3.6. 考虑服从 *Cauchy(Breit-Wigner)* 分布的随机变量 x，

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}. \tag{3.5}$$

(a) 证明如果 r 均匀分布于 $[0,1]$，则

$$x(r) = \tan[\pi(r - \frac{1}{2})] \tag{3.6}$$

服从 *Cauchy* 分布。

(b) 利用 *(a)* 中的结果，写一段小程序产生 *Cauchy* 分布的随机数。产生 10^4 个事例并画出直方图。

(c) 修改 *(b)* 中的程序，重复进行实验，每次实验 n 个独立的柯西分布数值（如取 $n = 10$）。对每个样本，计算样本均值 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$。比较 $\overline{x}$ 的直方图与 x 的原始直方图。（参见习题 *10.8*。）

习题 3.7. 光电倍增管是可以探测单光子的设备，如图*3.1*所示。[1]

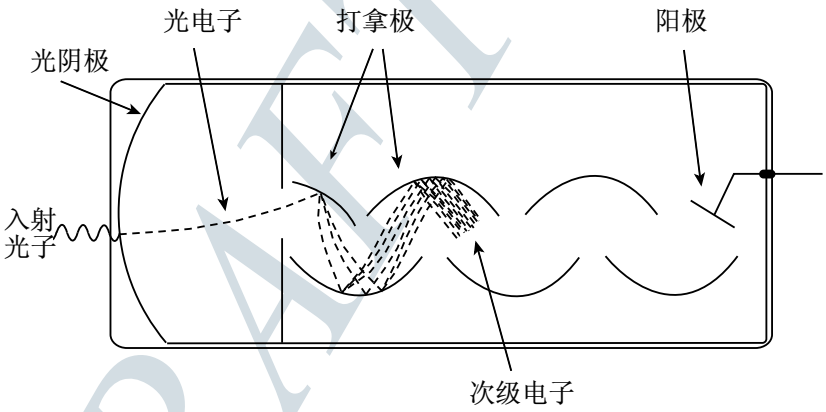

图 3.1: 光电倍增管的示意图。入射光子入射到光阴极上产生光电子，光电子被加速到打拿极上产生次级电子。

光子打到光阴极上，有一定的概率发射出一个电子（即光电子）。光电子在电场中向电极（称为打拿极）加速。当光电子撞到第一个打拿极时，光电子可以进一步释放出电子。这些电子又被加速到第二个打拿极，产生更多的电子。该过程一直持续，直到最后一个打拿极产生的电子被收集起来。

进入第 i 个打拿极的每个电子产生的次级电子数目可以看成均值为 ν_i 的泊松变量n_i，一般而言，每个打拿极的泊松均值 ν_i 不同。假设光电倍增管有 N 个打拿极，单个入射光电子最终产生的电子数目 n_{out} 的期待值（即光电倍增管的增益）为

$$\nu_{\text{out}} = E[n_{\text{out}}] = \prod_{i=1}^{N} \nu_i. \tag{3.7}$$

（*a*）写一段蒙特卡罗程序，用来计算 $N = 6$ 时单个光电子产生的电子数 n_{out} 的分布，对所有打拿极取 $\nu = 3.0$。（*ROOT* 中可以直接调用函数产生泊松分布。）利用你写的程序，重复 $M = 1000$ 次单个光电子事

[1] 更详细的描述请参见相关文献，例如，C. Grupen, A. Böhrer and L. Smolik, *Particle Detectors*, Cambridge University Press, Cambridge, 1996。

件，从而得到 M 个 n_{out}，作出这 M 个 n_{out} 值的直方图。（建议直方图分 *50* 个区间。）通过计算样本均值和样本方差

$$\overline{n}_{\text{out}} = \frac{1}{M}\sum_{i=1}^{M} n_{\text{out},i} \tag{3.8}$$

$$s^2_{\text{out}} = \frac{1}{M-1}\sum_{i=1}^{M}(n_{\text{out},i} - \overline{n}_{\text{out}})^2 \tag{3.9}$$

以估计均值 ν_{out} 和方差 $V[n_{\text{out}}] = \sigma^2_{\text{out}}$。（样本均值和样本方差详见 *Statistical Data Analysis* 第 *5* 章。）将样本均值与（*3.7*）式给出的值进行比较。比较样本方差（或标准差）与均值为 ν_{out} 的泊松变量的方差。定性解释为什么 n_{out} 的标准差远大于泊松变量的标准差。

（*b*）我们希望 n_{out} 的标准差尽量小，以便能够尽可能精确地确定初始光电子的数目（从而可以估计入射光子的数目）。在某些应用中，需要标准差小到可以区分到底是一个还是两个光子，因此希望使相对分辨率（即标准差与均值的比值）小于 1。减小方差的一个方法是提高第一个打拿极产生电子数的均值，这可以通过增大电压从而提高入射到打拿极的光电子的能量来实现，也可以通过选取较低功函数的金属（即提高发射次级电子的概率）来实现。

修改（*a*）中的程序，增大第一个打拿极的均值（比如增大至 6）。运行程序并估计 n_{out} 的标准差与均值的比值。定性解释为什么这么做比所有 ν_i 都相同时的分辨率更好。为什么提高后面的打拿极的增益对提高分辨率作用不大？

（*c*）尝试将程序改成模拟 $N = 12$ 个打拿极。你会发现模拟每个打拿极上每个电子的碰撞需要太长时间。可以考虑换个思路，取 $N = 6$ 运行足够多的事例从而获得足够精确的 n_{out} 的分布（例如至少 $M \approx 10^4$，分 *50* 个区间作出 $0 \le n_{\text{out}} \le 5000$ 之间的直方图）。用某种方法，比如舍选法，产生服从该分布的随机数。对前 6 个打拿极得到的每个电子，用同样方法模拟它在后面 6 个打拿极产生的电子数目。对前 6 个打拿极，取 $\nu_1 = 6$，其余 $\nu_i = 3$；对后 6 个打拿极，全部取 $\nu_i = 3$。

习题 3.8. 假设二维随机变量 (r, θ) 表示二维平面上的某点的极坐标。写一段程序，产生 1000 对 (r, θ)，使其代表的点在以圆点为圆心的单位圆内均匀分布。

习题 3.9. 写一段程序，用蒙特卡罗方法计算下面的定积分：

$$\int_0^1 \frac{e^{-x}}{\sqrt{x}}\mathrm{d}x$$

习题 3.10. 假设利用加速器产生了从原点出发沿 z 轴正向运动的单能 K^0_{S} 粒子，能量 $E_K = \frac{M_K^2 c^2}{2m_\pi}$。$K^0_{\text{S}}$ 粒子平均寿命为 τ，在实验室系飞行一段距离后衰变成 $\pi^+\pi^-$ 粒子对。在 K^0_{S} 质心系中，$\pi^\pm$ 的角分布各向同性。粒子束流前放置了一个圆盘状的探测器以记录末态粒子 $\pi^\pm$，圆盘半径 $R = 7\,\text{cm}$，轴线与 z 轴重合，距离原点 $D = 14\,\text{cm}$。见图*3.2*。

末态粒子对 $\pi^+\pi^-$ 同时击中探测器则表明探测到了 K^0_{S} 粒子的衰变。求探测器的接受效率。（已知质量 $M_K = 0.498\,\text{GeV}/c^2$，$m_\pi = 0.140\,\text{GeV}/c^2$，寿命 $\tau = 8.954 \times 10^{-11}\,\text{s}$，光速 $c = 3 \times 10^8\,\text{m/s}$。）

习题 3.11. 假设粒子在穿过气体时可以发生两种相互独立的过程 A 和 B。如果仅存在 A 过程，其平均自由程为 L_A，即粒子从发生上一次 A 过程到发生一下次 A 过程之间飞行的距离 $X \sim f_A(x) = \frac{1}{L_A}e^{-x/L_A}$。如果仅存在 B 过程，其平均自由程为 L_B，即粒子从发生上一次 B 过程到发生一下次 B 过程之间飞行的距离 $X \sim f_B(x) = \frac{1}{L_B}e^{-x/L_B}$。请问：在 A 过程和 B 过程同时存在的情况下，粒子的平均自由程是多少？可解析计算，也可用蒙特卡罗方法计算（取 $L_A = 1\,\text{cm}$，$L_B = 2\,\text{cm}$）。

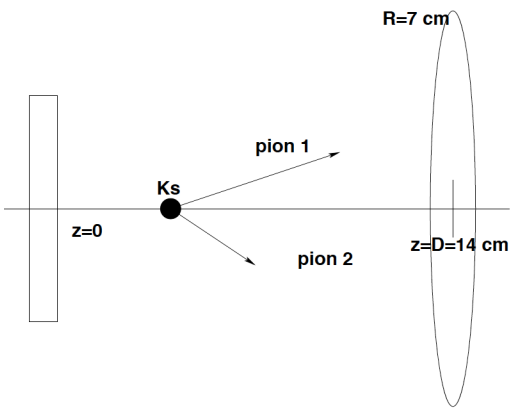

图 3.2: 关于探测效率的估计的示意图。

第四章　统计检验

习题 4.1. 带电粒子穿过气体会发生电离现象，电离的数目与入射粒子的类型有关。假设利用对电离的测量构造了某个检验统计量t，使其服从高斯分布：如果带电粒子是电子，则高斯分布的均值为 0，标准差为 1；如果带电粒子是 π 介子，则高斯分布的均值为 2，标准差为 1。构造一个检验，通过要求 $t < 1$ 选择出电子事件。

(a) 该检验的显著性水平（*significance level*）如何？（显著性水平即在拒绝域中接受电子的概率。）

(b) 该检验排除带电粒子为 π 介子的假设的功效（*power*）多大？有多大概率将 π 介子鉴别为电子？

(c) 假定已知样本中 π 介子和电子的比例分别为 99% 何 1%，求由 $t < 1$ 得到的电子样本的纯度（*purity*）。

(d) 假定要求所得电子样本的纯度不低于 95%，应当如何选择拒绝域（即判选条件）？此时该检验接受电子的效率和显著性水平如何？

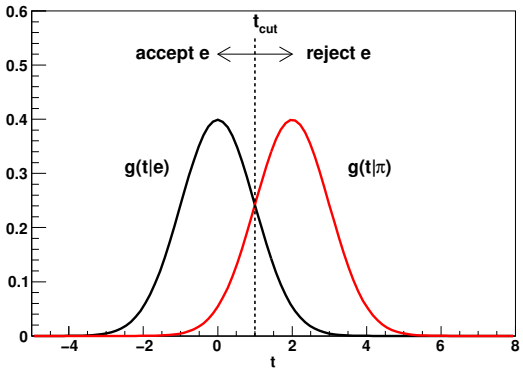

图 4.1: 用于判选电子 e 和 π 介子的检验统计量 t 的概率密度，判选条件为 $t_{\text{cut}} = 1$。

习题 4.2. 考虑某检验统计量 t 为输入变量 $\mathbf{x} = (x_1, \ldots, x_n)$ 的线性组合，系数为 $\mathbf{a} = (a_1, \ldots, a_n)$，即

$$t(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i = \mathbf{a}^T \mathbf{x}. \tag{4.1}$$

假定在 H_0 和 H_1 假设下，$\mathbf{x}$ 的均值分别为 $\boldsymbol{\mu}_0$ 和 $\boldsymbol{\mu}_1$，协方差矩阵分别为 V_0 和 V_1，检验统计量 t 的均值分别为 τ_0 和 τ_1，方差分别为 Σ_0^2 和 Σ_1^2（见 *Statistical Data Analysis* 第 *4.4.1* 节）。

(a) 证明：当系数为 $\mathbf{a} = W^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ 时（其中 $W \equiv V_0 + V_1$），下式定义的量（即费舍尔线性甄别量）达到最大值：

$$J(\mathbf{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}. \tag{4.2}$$

(b) 假定 $V_0 = V_1 = V$，且 $\mathbf{x}$ 在 H_0 和 H_1 假设下的概率密度函数 $f(\mathbf{x}|H_0)$ 和 $f(\mathbf{x}|H_1)$ 都是多维高斯分布，均值分别为 $\boldsymbol{\mu}_0$ 和 $\boldsymbol{\mu}_1$（参见 *Statistical Data Analysis* 的式 *4.26*）。记 H_0 和 H_1 假设下 $\mathbf{x}$ 的先验概率密度分别为 π_0 和 π_1。利用贝叶斯定理，求验后概率 $P(H_0|\mathbf{x})$ 和 $P(H_1|\mathbf{x})$ 作为 t 的函数。

(c) 推广该检验统计量，使其包含一个偏倚 a_0，即：

$$t(\mathbf{x}) = a_0 + \sum_{i=1}^{n} a_i x_i. \tag{4.3}$$

证明此时验后概率 $P(H_0|\mathbf{x})$ 可以表示为

$$P(H_0|\mathbf{x}) = \frac{1}{1+e^{-t}}, \tag{4.4}$$

其中偏倚 a_0 为

$$a_0 = -\frac{1}{2}\boldsymbol{\mu}_0^T V^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T V^{-1}\boldsymbol{\mu}_1 + \log\frac{\pi_0}{\pi_1}. \tag{4.5}$$

习题 4.3. 正负电子对撞中观测到具有特殊运动学性质的事件数可以被视为泊松变量。假定对于给定积分亮度（即束流强度对取数时间的积分），预期从某已知过程得到的事件数为 3.9，而实际上观测到 16 个事件。零假设为没有新物理过程对观测到的事件数有贡献；计算零假设的 P 值。在计算泊松分布求和时，可以利用关系式

$$\sum_{n=0}^{m} P(n;\nu) = 1 - F_{\chi^2}(2\nu; n_{\text{dof}}), \tag{4.6}$$

其中 $P(n;\nu$ 是均值为 ν 的泊松分布的概率，F_{χ^2} 是自由度数目为 $n_{\text{dof}} = 2(m+1)$ 的 χ^2 分布的累积分布。可以调用 *ROOT* 中的 `TMath` 名字空间下的函数 `TMath::Prob(double chi2, int ndf)` 计算，或者查表得到。

习题 4.4. 表*4.1*是实验获取的分区间数据和理论预言值。第二、三列是区间边界，第四列是对应区间的观测事件数 n_i（$i = 1, \ldots, 20$），服从泊松分布。第五、六列是两种理论对期待值 $\nu_i = E[n_i]$ 的预言，如图*4.2*所示。

(*a*)写一段程序，将表中 *20* 个区间的实验观测值和理论预期值画成直方图，画到一张图上，并根据 *"Statistical Data Analysis"* 的式（*4.39*）式计算 χ^2 统计量。

(*b*) 因为很多区间的事件数很小甚至为零，前面计算的统计量不太可能服从 χ^2 分布。写一段程序，根据两种理论假设（**theory1** 和 **theory2**）给出真实的 χ^2 分布。如果利用（*a*）中的数据计算统计检验，其 P-值是多少？如果利用正常的 χ^2 分布计算，P-值是多少？

习题 4.5. 在放射性实验中，卢瑟福和盖革记录了固定时间间隔内 α 衰变的次数。数据如表*4.2*所示。假定放射源中放射性核素的数目非常大，且任意一个核素在小时间间隔内发射一个 α 粒子的概率很小，则可以认为在时间间隔 Δt 内发生衰变的次数 m 服从泊松分布。如果观测结果与泊松分布的假设存在差异，则表明核素的 α 衰变不相互独立，比如某个核素发生 α 衰变可能会引发邻近核素也发生衰变，从而在短时间间隔内形成衰变簇团。

(*a*) 利用表*4.2*的数据，计算样本均值

$$\overline{m} = \frac{1}{n_{\text{tot}}}\sum_m n_m m, \tag{4.7}$$

以及样本方差

$$s^2 = \frac{1}{n_{\text{tot}} - 1}\sum_m n_m (m - \overline{m}), \tag{4.8}$$

序号	$x_{\min}$	$x_{\max}$	n (data)	ν (theory1)	ν (theory2)
1	0.0	0.5	1	0.2	0.2
2	0.5	1.0	0	1.2	0.7
3	1.0	1.5	3	1.9	1.1
4	1.5	2.0	4	3.2	1.6
5	2.0	2.5	6	4.0	1.9
6	2.5	3.0	3	4.5	2.2
7	3.0	3.5	3	4.7	2.7
8	3.5	4.0	4	4.8	3.3
9	4.0	4.5	5	4.8	3.6
10	4.5	5.0	7	4.5	3.9
11	5.0	5.5	4	4.1	4.0
12	5.5	6.0	5	3.5	4.0
13	6.0	6.5	2	3.0	3.9
14	6.5	7.0	0	2.4	3.5
15	7.0	7.5	1	1.6	3.2
16	7.5	8.0	0	0.9	2.8
17	8.0	8.5	0	0.5	2.2
18	8.5	9.0	1	0.3	1.5
19	9.0	9.5	0	0.2	1.0
20	9.5	10.0	0	0.1	0.5

表 4.1: 实验获取的分区间数据和理论预言值。第二、三列是区间边界。

m	n_m	m	n_m
0	57	8	45
1	203	9	27
2	383	10	10
3	525	11	4
4	532	12	0
5	408	13	1
6	273	14	1
7	139	> 14	0

表 4.2: 卢瑟福与盖革实验数据，即在时间间隔 $\Delta t = 7.5\,\mathrm{s}$ 内发生 m 次 α 衰变的次数。

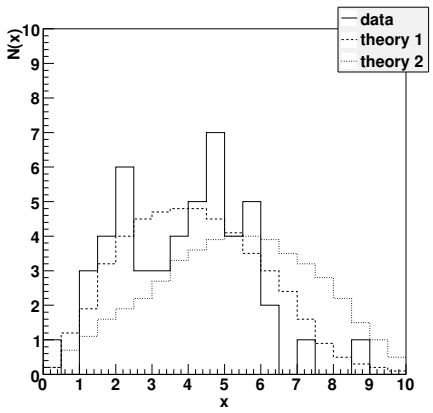

图 4.2: 从表中读取的观测数据和理论预言值的结果。

其中 n_m 是发生 m 个衰变的次数，$n_{\text{tot}} = \sum_m n_m = 2608$ 是测量时间内总衰变次数。求和从 $m = 0$ 一直到测量时间内最大的衰变次数 *(*次数为 $m = 14$*)*。利用 $\overline{m}$ 和 s^2，求分散度

$$t = \frac{s^2}{\overline{m}}. \tag{4.9}$$

$\overline{m}$ 和 s^2 分别为 m 的均值和方差的估计量（参见 *Statistical Data Analysis* 第 *5* 章）；如果 m 服从泊松分布，则 $\overline{m}$ 和 s^2 应该相等，于是可以预期 t 大约为 1。可以证明对于泊松分布，当 n_{tot} 很大时，$(n_{\text{tot}} - 1)t$ 服从自由度为 $n_{\text{tot}} - 1$ 的 χ^2 分布。而且，当 n_{tot} 很大时，该分布变成均值为 $n_{\text{tot}} - 1$，方差为 $2(n_{\text{tot}} - 1)$ 的高斯分布。

(b) m 服从泊松分布这一假设的 P-值为多少？为了表征 t 的观测值与泊松假设相符或不相符，应该选取什么样的 t 值（即 t 大表示相符还是 t 小表示相符）？

(c) 写一段蒙特卡罗程序产生很多组数据，每组数据包含 n_{tot} 个服从泊松分布的 m 值。(泊松分布的随机数可以在 ROOT 中直接调用，gRandom$->$ Possion(ν)。) 对于 m 的均值，可以取表*4.2*中数据的均值 $\overline{m}$。对于每组数据，计算 t 的值并填充至直方图。利用直方图和从卢瑟福实验数据得到的 t 值，计算泊松假设的 P-值。将该结果与 (a) 中的结果进行比较。*(*选作：将 $(n_{\text{tot}} - 1)t$ 记录至直方图，与均值为 $n_{\text{tot}} - 1$，方差为 $2(n_{\text{tot}} - 1)$ 的高斯分布进行比较。*)*

习题 4.6. 在宇称守恒的条件下某可观测量 x 的取值大于零和小于零的概率均为 0.5。现在对 x 作了 *1000* 次观测，其中 *560* 次 $x > 0$，*440* 次 $x < 0$。根据这组观测，试问宇称守恒的假设合理吗？请用假设检验给出显著性水平 $\alpha = 0.05$ 下的结论。

习题 4.7. 设 $x_1, x_2, \ldots, x_n$ 是来自 $N(\mu, 1^2)$ 的样本，考虑如下假设检验问题：

$$H_0 : \mu = 2 \qquad vs \qquad H_1 : \mu = 3.$$

检验的拒绝域选为 $W = \{\overline{x} \geq 2.6\}$。

(a) 当 $n = 20$ 时，求该检验犯第一类错误和第二类错误的概率；

(b) 如果要使得该检验犯第二类错误的概率 $\beta \leq 0.01$，则 n 最小应该取多少？

(c) 证明：当 $n \to \infty$ 时，$\alpha \to 0$ 且 $\beta \to 0$。

习题 4.8. 设 $x_1, x_2, \ldots, x_n$ 是来自正态总体 $N(\mu, 2^2)$ 的样本，考虑如下假设检验问题：

$$H_0: \mu = 6 \qquad vs \qquad H_1: \mu \neq 6.$$

检验的拒绝域取为 $W = \{|\overline{x} - 6| \geq c\}$。试求 c 使得检验的显著性水平为 0.05，并求该检验在 $\mu = 6.5$ 处犯第二类错误的概率。取 $n = 16$。

习题 4.9. 根据某理论，观测到流星表示幸运事件。根据以往的统计，某人每年平均观测到 *10* 颗流星。*2022* 年某人观测到 *5* 颗流星。我们能据此说 *2022* 年对于这个人来说不是幸运年吗？请在 $\alpha = 0.05$ 的显著性水平下给出结论。

习题 4.10. 如果对某个假设进行了几个独立的显著性检验，给出了显著性水平 P_1, $P_2, \ldots, P_n$，总的显著性水平不能通过将这些概率相乘得到。为什么呢？

如果 X 是在 0 和 1 之间均匀分布的随机变量，证明 $-2\ln X$ 是自由度为 2 的 χ^2 变量。我们可以利用这个结果来合并独立的显著性检验的结果。如果三个检验的显著性水平分别为 0.145、0.263 和 0.087，我们应当如何评估总的显著性？

第五章　参数估计的一般概念

习题 5.1. 考虑随机数 x，其均值和方差分别为 μ 和 σ^2。假设样本空间由 n 次观测结果 $x_1, x_2, \cdots, x_n$ 构成。本题的目的是证明样本均值

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{5.1}$$

为均值 μ 的一致性估计量。

(a) 第一步要证明 *Chebyshev* 不等式，即只要 x 的方差存在，对任意 $a > 0$，下面式子成立：

$$P(|x-\mu|) \geq a) \leq \frac{\sigma^2}{a^2}. \tag{5.2}$$

该式的证明需要用到方差的定义，

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx, \tag{5.3}$$

其中 $f(x)$ 为随机变量 x 的概率密度函数。利用如下事实：如果积分区域限定为 $|x-\mu| \geq a$，则积分 *(5.3)* 会变小，如果用 a^2 替换 $(x-\mu)^2$，则积分会更小。

(b) 利用 *Chebyshev* 不等式证明大数弱定理，即，对任意 $\epsilon > 0$，

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| \geq \epsilon\right) = 0. \tag{5.4}$$

这等价于，$\overline{x}$ 是 μ 的一致性估计量，只要 x 的方差存在。

习题 5.2. 考虑均值为 μ，方差为 σ^2 的随机变量 x，并得到样本值为 $x_1, x_2, \cdots, x_n$ 的样本空间。

(a) 假设均值 μ 已经利用样本均值 $\overline{x}$ 估计。证明样本方差

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}(\overline{x^2} - \overline{x}^2) \tag{5.5}$$

为方差 σ^2 的无偏估计量。（利用 $E[x_i x_j] = \mu^2 (i \neq j)$，$E[x_i^2] = \mu^2 + \sigma^2 (i = 1, 2, \ldots, n)$。

(b) 假设均值 μ 已知。证明 σ^2 的无偏估计量为

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \overline{x^2} - \mu^2. \tag{5.6}$$

习题 5.3.

(a) 证明样本方差 s^2*(5.5)* 的方差为

$$V[s^2] = E[s^4] - (E[s^2])^2 = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right), \tag{5.7}$$

其中 $\mu_k = E[(x-\mu)^k]$ 为 x 的 k 阶中心矩。为此，需要先证明 s^2 可以写为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)}\sum_{i,j=1}^{n} x_i x_j. \tag{5.8}$$

然后证明 s^4 的期待值为

$$E[s^4] = \frac{1}{(n-1)^2}\sum_{i,j=1}^{n} E[x_i^2 x_j^2] - \frac{2}{n(n-1)^2}\sum_{i,j,k=1}^{n} E[x_i x_j x_k^2] + \frac{1}{n^2(n-1)^2}\sum_{i,j,k,l=1}^{n} E[x_i x_j x_k x_l]. \tag{5.9}$$

计算每个求和给出多少项代数矩 μ_4' 或 $\mu_2'^2$。注意其余所有项都 μ 的一次项或高次项。令 $\mu = 0$，将结果表示为中心矩 μ_2 和 μ_4 的函数。从中减掉习题5.2得到的 $(E[s^2])^2$ 即可得到结果。

(b) 如果 x 服从高斯分布，计算 s^2 的方差。利用高斯分布的 *4* 阶中心矩为 $\mu_4 = 3\sigma^4$。

习题 5.4. 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，$\overline{X}$ 是样本均值，S^2 是样 > 本方差。证明

(1) $\overline{X} \sim N(\mu, \sigma^2/n)$。

(2) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$。

(3) $\overline{X}$ 与 S^2 相互独立。（选做，后面可以直接使用这个结论）

(4) $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$。

习题 5.5. 设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是来自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的样本，且这两个样本相互独立。设 $\overline{X}$ 和 $\overline{Y}$ 分别是这两个样本的样本均值，S_1^2 和 S_2^2 分别是这两个样本的样本方差。证明

(1) $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$。

(2) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时，

$$\frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2),$$

其中

$$S_w^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}, \quad S_w = \sqrt{S_w^2}.$$

习题 5.6. 假设 S_x^2 和 S_y^2 分别是来自 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的两个简单随机样本的样本方差。证明 $F = \frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2} \sim F(m-1, n-1)$。

第六章　最大似然法

习题 6.1.

(a) 给定一组数据高斯分布的数据样本 $x_1, \ldots, x_n$，求均值 μ 和方差 σ^2 的最大似然估计量。

(b) 将估计量 $\hat{\mu}$ 和 $\hat{\sigma^2}$ 同《统计数据分析》第 5 章定义的估计量 $\bar{x}$ 和 s^2 联系起来，计算 $\hat{\mu}$ 和 $\hat{\sigma^2}$ 的期待值和方差。

(c) 通过计算

$$(V^{-1})_{ij} = -E\left[\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\right], \tag{6.1}$$

近似求解协方差矩阵的逆 *(*大统计样本有效*)*。其中 θ_i 和 θ_j $i, j = 1, 2$ 分别为 μ 和 σ^2。对 V^{-1} 求逆，计算出协方差矩阵，并将对角元素 *(*方差*)* 与 *(b)* 中求得的精确值比较。注意到，在大样本极限下，*(b)* 和 *(c)* 的结果符合。

习题 6.2. 考虑某二项分布变量，即 N 次试验中成功的次数 n，单个实验试验成功的概率为 p。如果只进行一次观测得到 n，p 的最大似然估计量为多少？证明 $\hat{p}$ 是无偏的，并求其方差。证明 $\hat{p}$ 的方差等于最小方差边界 *(*见《统计数据分析》*(6.16)* 式*)*。

习题 6.3.

(a) 重新考虑二项分布的随机变量 n，每次试验结果 *(*成功或失败*)* 的概率为 p 和 $q = 1 - p$。利用习题6.2得到的 p 的估计量，构造不对称度

$$\alpha = p - q = 2p - 1 \tag{6.2}$$

的最大似然估计量 $\hat{\alpha}$，并求标准差 $\sigma_{\hat{\alpha}}$。

(b) 假设需要测量某个非常小的不对称度，预计大约是 $\alpha \approx 10^{-3}$ 水平。如果要求标准差不大于 α 的三分之一，至少需要多少次试验？

习题 6.4. 考虑泊松变量的单次观测值 n。均值 ν 的最大似然估计量为多少？证明该估计量是无偏的并求其方差。证明 $\hat{\nu}$ 的方差等于最小方差边界。

习题 6.5. 支持粒子物理标准模型的早期证据是观测到左手 *(R)* 和右手 *(L)* 极化的电子与氘靶的非弹散射截面 σ_R 和 σ_L 不同。对于给定积分亮度 L*(*正比于电子束流密度以及取数时间*)*，两种类型的事例数 n_R 和 n_L 均为泊松变量，平均值分别为 ν_R 和 ν_L。均值与散射截面的关系为 $\nu_R = \sigma_R L$ 和 $\nu_L = \sigma_L L$，并且实验中两种情形的亮度 L 相同。利用习题的结果，构造计划不对称度的估计量 $\hat{\alpha}$，

$$\alpha = \frac{\sigma_R - \sigma_L}{\sigma_R + \sigma_L}, \tag{6.3}$$

利用误差传递，求标准差 $\sigma_{\hat{\alpha}}$，用 α 和 $\nu_{tot} = \nu_R + \nu_L$ 表示。将此与习题6.3的结果进行比较。预计不对称度大约为 10^{-4} 水平，要想使 $\sigma_{\hat{\alpha}}$ 比不对称度小一个数量级，需要多少散射事例？*(*事例数非常大，以至于事例

不能单独记录下来，而是测量探测器输出电流。参见 *C.Y. Prescott et al., Parity non-conservation in inelastic electron scattering, Phys. Lett. B77(1978)347.)*

习题 6.6. 随机变量 x 服从分布 $f(x;\theta)$，其中 θ 为未知参数。考虑样本空间 $\mathbf{x}=(x_1,\ldots,x_n)$，以此构造 θ 的估计量 $\hat{\theta}(\mathbf{x})$*(不限于最大似然估计量)*。证明 *Rao-Cramér-Frechet(RCF)* 不等式

$$V[\hat{\theta}] \geq \frac{\left(1+\frac{\partial b}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]}, \tag{6.4}$$

其中 $b=E[\hat{\theta}]-\theta$ 为估计量的偏置。可分以下几步证明：

(a) 首先，证明 *Cauchy-Schwarz* 不等式，即对任意两个随机变量 u 和 v，

$$V[u]V[v] \geq (\mathrm{cov}[u,v])^2, \tag{6.5}$$

其中 $V[u]$ 和 $V[v]$ 为方差，$\mathrm{cov}[u,v]$ 为协方差。利用 $\alpha u+v$ 的方差必定大于等于零 *(*对任意 α*)*。然后考虑特殊情形 $\alpha=(V[v]/V[u])^{1/2}$。

(b) 利用 *Cauchy-Schwarz* 不等式，令

$$\begin{aligned} u &= \hat{\theta}, \\ v &= \frac{\partial}{\partial \theta}\log L, \end{aligned} \tag{6.6}$$

其中 $L=f_{\text{joint}}(\mathbf{x};\theta)$ 为似然函数，也是 $\mathbf{x}$ 的联合概率密度函数。代入 *(6.5)*，表示出 $V[\hat{\theta}]$ 的下界。这里要注意的是，我们将似然函数看成 $\mathbf{x}$ 的函数，即似然函数本身也是一个随机变量。

(c) 假设对 θ 的微分可以移到积分的外面，证明

$$E\left[\frac{\partial}{\partial \theta}\log L\right] = \int\cdots\int f_{\text{joint}}(\mathbf{x};\theta)\frac{\partial}{\partial \theta}\log f_{\text{joint}}(\mathbf{x};\theta)dx_1\ldots dx_n = 0. \tag{6.7}$$

我们将推导的 *RCF* 不等式的形式依赖于该假设，这在感兴趣的问题中一般都成立。*(*只要积分的极限不依赖于 θ，该假设总是成立的。*)* 利用 *(6.7)* 与 *(6.5)*、*(6.6)*，证明

$$V[\hat{\theta}] \geq \frac{\left(E\left[\hat{\theta}\frac{\partial \log L}{\partial \theta}\right]\right)^2}{E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right]}. \tag{6.8}$$

(d) 证明 *(6.8)* 的分子可以表示为

$$E\left[\hat{\theta}\frac{\partial \log L}{\partial \theta}\right] = 1+\frac{\partial b}{\partial \theta}, \tag{6.9}$$

分母可以表示为

$$E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]. \tag{6.10}$$

再次假设对 θ 的微分与对 $\mathbf{x}$ 的积分可以互换次序。将 *(c)* 和 *(d)* 的结果放到一起即可证明 *(6.4)*。

习题 6.7. 写一段程序，根据指数分布

$$f(t;\tau) = \frac{1}{\tau}e^{-t/\tau}, \quad t\geq 0, \tag{6.11}$$

产生样本容量为 n 的样本 $(t_1,\ldots,t_n)$。

(a) 证明 τ 的最大似然估计量由样本平均 $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ 给出。产生 *1000* 个 $\tau = 1$，$n = 10$ 的样本。对每个样本计算 $\hat{\tau}$ 并做直方图。比较 $\hat{\tau}$ 的均值与真值 $\tau = 1$。

(b) 假设 t 的概率密度函数的参数为 $\lambda = 1/\tau$，即

$$f(t;\lambda) = \lambda e^{-\lambda t}, \quad t \geq 0, \tag{6.12}$$

证明 λ 的最大似然估计量为 $\hat{\lambda} = 1/\sum_{i=1}^{n} t_i$。修改 *(a)* 中的程序，加入 $\hat{\lambda}$ 的直方图。比较 $\hat{\lambda}$ 的均值与真值 $\lambda = 1$。对 $n = 5, 10, 100$ 三种情况，分别数值计算偏置 $b = E[\hat{\lambda}] - \lambda$。

习题 6.8. 日内瓦的出租车牌照号是从 *1* 一直到总数目 N_{taxi}。对牌照的 N 次观测得到观测数字 $n_1, \ldots, n_N$。

(a) 构造出租车总数目的最大似然估计量。*(*这是最大似然估计量有偏并且不有效的常见例子。困难的根源在于数据的边界依赖于参数。*)*

(b) 提出出租车数目更好的估计量并给出均值和方差。

习题 6.9. 考虑 N 个独立的泊松变量 $n_1, \ldots, n_N$，均值为 $\nu_1, \ldots, \nu_N$。假设均值依赖于某可控变量 x

$$\nu(x) = \theta a(x), \tag{6.13}$$

其中 θ 为未知参数，$a(x)$ 为任意已知函数。N 个 ν_i 因而可以由 $\nu(x_i) = \theta a(x_i)$ 给出，其中假设 $x_1, \ldots, x_N$ 已知。证明 θ 的最大似然估计量为

$$\hat{\theta} = \frac{\sum_{i=1}^{N} n_i}{\sum_{i=1}^{N} a(x_i)}. \tag{6.14}$$

证明 $\hat{\theta}$ 为无偏的，并且其方差有最小方差边界给出 *(*参见习题6.6*)*。

习题 6.10. 习题6.9描述的状况的例子之一为 *(*反*)* 中微子-核子散射。根据夸克-部分子模型，反应 $\nu N \to \mu^- X$ 和 $\bar{\nu} N \to \mu^+ X$ 的截面为

$$\begin{aligned} \sigma(\nu N \to \mu^- X) &= \frac{G^2 M E}{\pi}\left(\langle q\rangle + \frac{1}{3}\langle \bar{q}\rangle\right) \equiv \theta_\nu E \\ \sigma(\bar{\nu} N \to \mu^+ X) &= \frac{G^2 M E}{\pi}\left(\frac{1}{3}\langle q\rangle + \langle \bar{q}\rangle\right) \equiv \theta_{\bar{\nu}} E \end{aligned} \tag{6.15}$$

其中 E 为入射 *(*反*)* 中微子的能量，$M = 0.938\,\text{GeV}$ 为靶核子的质量，$G = 1.16 \times 10^{-6}\,\text{GeV}^{-2}$ 为费米常数。(取自然单位制 $c = 1$。) 这里变量 x 对应于能量 E，式 *(*6.15*)* 右端的参数对应两个不同的参数 θ_ν 和 $\theta_{\bar{\nu}}$。

假设在 N 个不同能量值处收集了数据。每个能量点，事例数的期待值为

$$\nu_i = \sigma(E_i)\epsilon(E_i)\mathcal{L}_i, \tag{6.16}$$

其中 $\sigma(E_i)$ 为 *(*反*)* 中微子在能量为 E_i 时的截面，$\mathcal{L}_i$ 为积分亮度，$\epsilon(E_i)$ 为记录事例的概率 *(*效率*)*，效率通常为能量的函数。出于本习题的目的，我们假设能量 E_i，积分亮度 $\mathcal{L}_i$ 以及效率 $\epsilon(E_i)$ 精确已知没有不确定度。*(*再假设没有本底事例。*)* 确定参数 θ_ν 和 $\theta_{\bar{\nu}}$ 的最大似然估计量，并据此求 $\langle q\rangle$ 和 $\langle \bar{q}\rangle$ 的估计量。在夸克-部分子模型中，它们分别对应夸克和反夸克携带的动量占核子动量的份额。确定除了正反夸克外其它粒子 *(*即，胶子*)* 携带的动量份额 $\langle g\rangle = 1 - \langle q\rangle - \langle \bar{q}\rangle$。

习题 6.11. 确定阿伏伽德罗常数的最早实验之一是基于布朗运动，实验装置如图6.1所示。*Jean Perrin*[1] 用该装置观测悬浮在水中的乳香 *(*一种抛光材料*)* 颗粒。

[1] Jean Perrin, Mouvement brownien et réalité moléculaire, *Ann. Chimie et Physique,* 8^{ϵ} *série,* **18**(1909)1-114; *Les Atomes*, Flammarion, Paris, 1991(first edition, 1913); *Brownian Movement and Molecular Reality*, in Mary-Jo Nye, ed., *The Question of the Atom*, Tomash, Los Angeles, 1984.

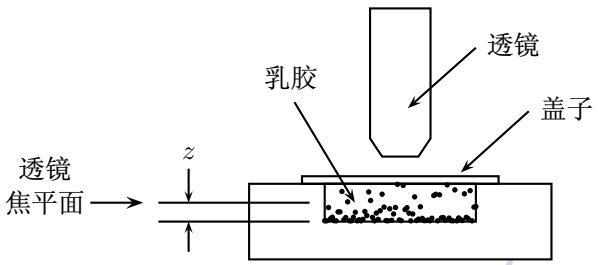

图 6.1: Jean Perrin 的实验装置，用于观测悬浮在水中的粒子数作为高度的函数。

粒子为半径 $r = 0.52\,\mu\text{m}$ 的球状颗粒，密度为 $1.063\,\text{g/cm}^3$，即，比水的密度大 $0.063\,\text{g/cm}^3$。在显微镜中观测这些颗粒，只有大约 $1\,\mu\text{m}$ 的一层在聚焦范围内，范围之外的粒子观测不到。通过调节显微镜的透镜，焦平面可以垂直移动。在 4 个不同高度 z 处拍摄了照片，*(*最低高度任意设为 $z = 0$*)*，并数出不同 z 处的粒子数 $n(z)$。数据如表*6.1*所示。

高度 z(μm)	粒子数 n
0	1880
6	940
12	530
18	305

表 6.1: Perrin 观测的数据，乳剂中不同高度 z 处乳香粒子的数目。

水中球状乳香粒子的引力势能为

$$E = \frac{4}{3}\pi r^3 \Delta g z, \tag{6.17}$$

其中 $\Delta\rho = \rho_{\text{乳香}} - \rho_{\text{水}} = 0.063\,\text{g/cm}^3$ 为密度差，$g = 980\,\text{cm/s}^2$ 为重力加速度。统计力学预言，粒子处在能量为 E 的态的概率正比于

$$P(E) \propto e^{-E/kT}, \tag{6.18}$$

其中 k 为 *Boltzmann* 常数，T 为绝对温度。因此，粒子数作为高度的函数服从指数规律，其中在 z 处观测到的粒子数 n 可以看作均值为 $\nu(z)$ 的泊松变量。结合式 *(6.17)* 和 *(6.18)* 得到

$$\nu(z) = \nu_0 \exp\left(-\frac{4\pi r^3 \Delta\rho g z}{3kT}\right), \tag{6.19}$$

其中 ν_0 为 $z = 0$ 时粒子数的期待值。

(a) 写程序用最大似然法计算参数 k 和 ν_0。利用表*6.1*中的数据按照泊松概率构造最大似然函数 *(*参见《统计数据分析》*6.10* 节*)*，

$$\log L(\nu_0, k) = \sum_{i=1}^{N} (n_i \log \nu_i - \nu_i), \tag{6.20}$$

其中 $N = 4$ 为测量次数。温度取 $T = 293\,\text{K}$。

(b) 利用得到的 k，通过下面的关系计算阿伏伽德罗常数

$$N_A = R/k, \tag{6.21}$$

其中 R 为气体常数。*Perrin* 计算时取值为 $R = 8.32 \times 10^7\,\text{erg/mol K}$。

(c) 不求解对数似然函数 *(6.20)* 的最大值，而是通过最小化

$$\chi^2_{\mathrm{P}}(\nu_0, k) = 2\sum_{i=1}^{N}\left(n_i \log\frac{n_i}{\nu_i} + \nu_i - n_i\right), \tag{6.22}$$

其中 $\nu_i = \nu(z_i)$ 通过式 *(6.19)* 依赖于 ν_0 和 k。利用 χ^2_{P} 的值计算拟合优度 *(*参见《统计数据分析》*6.11* 节*)*。讨论 *Perrin* 测量 N_A 中可能的系统不确定度。

第七章　最小二乘法

习题 7.1. *Galileo* 研究运动的实验之一是小球和斜坡的实验。在离开斜坡边缘之前，小球的轨迹变成水平，如图7.1所示。对于不同的高度 h，测量从斜坡边缘到落地点的水平距离 d. *1608* 年，*Galileo* 测量了 *5* 组数据，如表7.1所示[1]。假设高度 h 的误差可以忽略，水平距离 d 可以看做独立的标准差 $\sigma = 15\,\text{punti}$ 的高斯

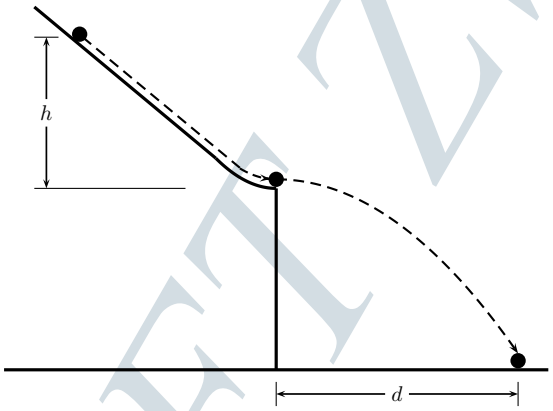

图 7.1: Galileo 小球和斜坡实验的示意图。

h	d
1000	1500
828	1340
800	1328
600	1172
300	800

表 7.1: Galileo 斜坡实验的 5 组数据。给定初始高度 h，d 为落地前的水平距离。单位为punti，1punto≃1mm。

随机变量。（实际上我们不清楚 *Galileo* 如何估计测量误差，但是 $1-2\%$ 的误差是可以接受的。）此外，我们知道如果 $h=0$，则水平距离 d 将为零，即，如果球从斜坡边缘出发，它将垂直落到地上。

(a) 考虑 h 和 d 的关系为如下形式

$$d = \alpha h \tag{7.1}$$

[1]参见 Stillman Drake and Maclachlan, Galileo's discovery of the parabolic trajectory, *Scientific American* **232** (March 1975) 102; Stillman Drake, *Galileo at Work*, University of Chicago Press, Chicago (1978).

以及

$$d = \alpha h + \beta h^2. \tag{7.2}$$

求参数 α 和 β 的最小二乘估计量。对应于这两个假设的最小 χ^2 和 P-值分别为多少？

(b) 假设 d 和 h 的关系为如下形式

$$d = \alpha h^\beta. \tag{7.3}$$

写一段程序对 α 和 β 进行最小二乘拟合。注意这是参数的非线性函数，必须数值求解。

(c)Galileo 认为运动是水平分量和垂直分量的叠加，其中水平运动是匀速运动，垂直速度在斜坡的最低处为零，随后随时间线性增加。证明这将导致关系式

$$d = \alpha\sqrt{h}. \tag{7.4}$$

求 α 的最小二乘估计量以及最小 χ^2。该假设的 P-值是多少？

习题 7.2. 考虑对直方图的最小二乘拟合。直方图对应区间 $i = 1, \ldots, N$ 的事例数为 y_i，理论预言值为

$$\lambda_i(\boldsymbol{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \boldsymbol{\theta}) dx, \tag{7.5}$$

其中 $f(x; \boldsymbol{\theta})$ 依赖于未知参数 $\boldsymbol{\theta}$。假设用参数 ν 代替总事例数 n，并且该参数在最小化

$$\chi^2(\boldsymbol{\theta}, \nu) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\boldsymbol{\theta}, \nu)^2}{\sigma_i^2} \tag{7.6}$$

时与其它参数同时调整。

(a) 证明如果 $\sigma_i^2 = \lambda_i$ 则总事例数的估计量为

$$\hat{\nu}_{LS} = n + \frac{\chi^2_{\min}}{2} \tag{7.7}$$

(b) 证明如果 $\sigma_i^2 = y_i$*(修正后的最小二乘)*，则估计量为

$$\hat{\nu}_{MLS} = n - \chi^2_{\min}. \tag{7.8}$$

习题 7.3. 考虑对直方图的最小二乘拟合，第 i 个区间的事例数为 y_i，$i = 1, \ldots, N$，对应的期待值为 $\lambda_i(\boldsymbol{\theta})$。假设总事例数 n 可以看作常数，于是 y_i 服从多项分布。

(a) 求协方差矩阵 $V_{ij} = \mathrm{cov}[y_i, y_j]$。为什么该矩阵的逆不存在？

(b) 考虑只使用前 $N - 1$ 个区间进行拟合。求协方差矩阵的逆，并证明这等价于对所有 N 个区间进行拟合但不考虑相关性。

习题 7.4. 假设样本容量为 n，利用该样本得到 N 个量的测量值：$y_1, \ldots, y_N$。这 N 个测量值将被用于最小二乘拟合以估计若干未知参数。如果这 N 个测量是相关的，在构造 χ^2 时需要用到协方差矩阵的逆 V^{-1}。一般来说，我们先通过某种办法（比如蒙特卡罗计算）得到相关系数矩阵 $\rho_{ij} = V_{ij}/(\sigma_i\sigma_j)$（其中 $i, j = 1, 2, \ldots, N$，然后再计算得到协方差矩阵的逆。

（a）注意到对于有效估计量，协方差矩阵的逆正比于样本容量 n。证明：在此条件下，相关系数矩阵与样本容量 n 无关。

(*b*) 证明协方差矩阵的逆为

$$(V^{-1})_{ij} = \frac{(\rho^{-1})_{ij}}{\sigma_i \sigma_j} \tag{7.9}$$

【提示：从下面等式出发

$$\delta_{ij} = \sum_k (V^{-1})_{ik} V_{kj} = \sum_k (V^{-1})_{ik} \rho_{kj} \sigma_k \sigma_j \tag{7.10}$$

对式（*7.10*）两边都乘以 ρ^{-1}，并对适当的指标求和即可得到式（*7.9*）。】

习题 7.5. 考虑随机变量 x 的两个部分重叠的样本，它们的样本容量分别为 n 和 m，共有部分的样本容量为 c。假设已知 x 的方差 $V[x] = \sigma^2$。考虑样本均值

$$y_1 = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7.11}$$

和

$$y_2 = \frac{1}{m} \sum_{i=1}^{m} x_i. \tag{7.12}$$

(*a*) 证明协方差为

$$\mathrm{cov}[y_1, y_2] = \frac{c\sigma^2}{nm}. \tag{7.13}$$

(*b*) 利用 *7.6* 节的结果，求 y_1 和 y_2 的加权平均和方差。

习题 7.6. 天文学家托勒密（*Claudius Ptolemy*）利用圆盘做过光折射的实验。他把圆盘的一半浸入水中，圆心正好位于水面处，如图*7.2*所示。大约公元 *140* 年，*Ptolemy* 对 *8* 组不同的入射角 θ_i 测量了相应的折射角

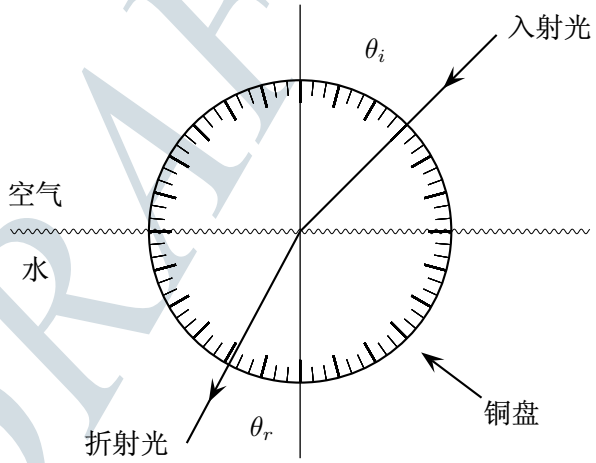

图 7.2: Ptolemy 用来研究光折射的设备。

θ_r，结果如表*7.2*所示[2]。本练习中，我们认为入射角已知且误差可以忽略，而把折射角看作标准差为 $\sigma = \frac{1}{2}^\circ$ 的高斯随机变量。（这是一个合理的假设，记录的角度精确到最邻近的半度。注意，我们可以将 θ_i 的误差吸收到 θ_r 的有效误差中。）

(*a*) 直到 *17* 世纪才发现正确的折射定律，在此之前，通常的假设是

$$\theta_r = \alpha \theta_i, \tag{7.14}$$

[2]取自 Pedersen and Mogens Pihl, *Early Physics and Astronomy: A Historical Introduction*, MacDonald and Janes, London, 1974

θ_i	θ_r
10	8
20	$15\frac{1}{2}$
30	$22\frac{1}{2}$
40	29
50	35
60	$40\frac{1}{2}$
70	$45\frac{1}{2}$
80	50

表 7.2: 入射角和折射角（单位：度）。

然而 *Ptolemy* 更喜欢用下面的形式

$$\theta_r = \alpha\theta_i - \beta\theta_i^2. \tag{7.15}$$

对这两种不同的假设，求参数的最小二乘估计量，并计算最小 χ^2 值。评论一下两个假设的拟合优度。是否可以相信所有的数据都是从实际测量得来的？[3]

(*b*) *1621* 年 *Snell* 发现了折射定律

$$\theta_r = \sin^{-1}\left(\frac{\sin\theta_i}{r}\right), \tag{7.16}$$

其中 $r = n_r/n_i$ 为两种介质的折射率之比。求 r 的最小二乘估计量并计算出最小 χ^2 值。评价对 θ_r 作 $\sigma = \frac{1}{2}^\circ$ 假设的合理性。

习题 7.7. 重新考虑练习 *6.9*：N 个独立的泊松变量 $\mathbf{n} = (n_1, \ldots, n_N)$，均值为 $\boldsymbol{\nu} = (\mu_1, \ldots, \nu_N)$，其中均值与某控制变量 x 有关，

$$\nu(x) = \theta a(x). \tag{7.17}$$

(a) 首先考虑最小二乘方法 *(LS)*，χ^2 的分母使用 $\sigma_i^2 = \nu_i$。证明 θ 的最小二乘估计量为

$$\hat{\theta} = \left(\frac{\sum_{i=1}^N \frac{n_i^2}{a(x_i)}}{\sum_{i=1}^N a(x_i)}\right)^{1/2}. \tag{7.18}$$

通过对 $\hat{\theta}(\mathbf{n})$ 在 $\boldsymbol{\nu}$ 处进行泰勒展开至第二阶，计算期待值，证明 *(7.18)* 的偏置为

$$b = \frac{N-1}{2\sum_{i=1}^N a(x_i)} + O(E[(n_i - \nu_i)^3]). \tag{7.19}$$

*(*利用独立泊松变量的协方差 $\mathrm{cov}[n_i, n_j] = \delta_{ij}\nu_j$*.)*

(b) 取 χ^2 的分母为 $\sigma_i^2 = n_i$，即把观测值作为方差，用修正的最小二乘法 *(MLS)* 重复 *(a)* 中的步骤。证明 θ 的最小二乘估计量为

$$\hat{\theta} = \frac{\sum_{i=1}^N a(x_i)}{\sum_{i=1}^N \frac{a(x_i)^2}{n_i}}, \tag{7.20}$$

并且偏置为

$$b = -\frac{N-1}{\sum_{i=1}^N a(x_i)} + O(E[(n_i - \nu_i)^3]). \tag{7.21}$$

将 *(a)* 和 *(b)* 得到的偏置与习题 *7.2* 进行比较。

[3]参见 R. Feynman, R. Leighton and M. Sands, *The Feynman Lectures on Physics,* Vol. I, Addison-Wesley, Menlo Park, 1963, Section 26-2.

(c) 利用误差传递，对 *LS* 和 *MLS* 两种情况估计 $\hat{\theta}$ 的方差。

注意，由于习题 *(6.9)* 已经证明了 θ 的最大似然估计量是无偏的并且方差最小，这里并不推荐最小二乘 *(LS)* 和修正的最小二乘 *(MLS)* 估计量。然而，对于足够大的数据样本，三个方法是类似的，参见习题 *(7.8)*。

习题 7.8. 重新考虑 *Perrin* 关于乳香粒子作为高度的函数 *(*习题 *6.11)*。通过最小化

$$\chi^2(k,\nu_0)=\sum_{i=1}^{N}\frac{(n_i-\nu_i(k,\nu_0))^2}{\sigma_i^2}, \tag{7.22}$$

求玻尔兹曼常数 k*(*或者等价于阿伏伽德罗常数 $N_A=R/k$*)* 和系数 ν_0 的最小二乘估计量。

(a) 取 n_i 的标准差 σ_i 为 $\sqrt{\nu_i}$*(*通常的最小二乘法*)*。

(b) 取 σ_i 为 $\sqrt{n_i}$*(*修正的最小二乘法*)*。

将 *(a)* 和 *(b)* 得到的估计量与习题 *(6.11)* 中最大似然估计量进行比较。

第八章　矩方法

习题 8.1. 考虑服从高斯分布的随机变量 x，均值 μ 和方差 σ^2 未知，并假设样本为 $x_1, \ldots, x_n$。

(a) 利用矩方法构造 μ 和 σ^2 的估计量。利用函数 $a_1 = x$，$a_2 = x^2$，使得期待值 $E[a_i(x)]$ 对应于 x 的一阶和二阶代数矩。

(b) 计算 *(a)* 中得到的估计量 $\hat{\mu}$ 与 $\hat{\sigma^2}$ 的期待值。这两个估计量是否是无偏的？

习题 8.2. 考虑粒子反应中产生的 ρ^0 介子衰变为两个带电 π 介子 $(\pi^+\pi^-)$。衰变角定义为，在 $\pi^+\pi^-$ 质心系中 π^+ 运动方向与 ρ 的原初方向的夹角，见图 *(8.1)*。由于 ρ^0 的自旋为 *1*，π 介子的自旋为 *0*，可以证明 $\cos\theta$

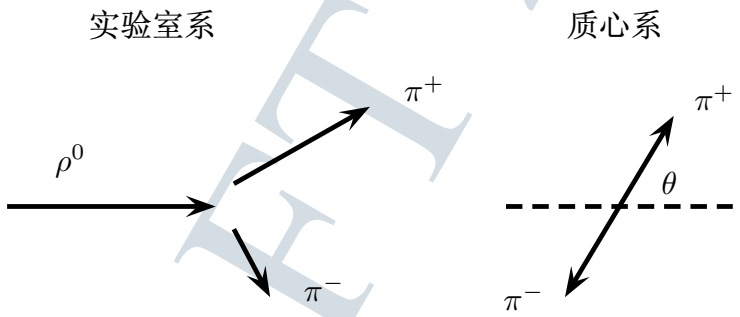

图 8.1: $\rho^0 \to \pi^+\pi^-$ 中衰变角的定义。

的分布具有如下形式

$$f(\cos\theta;\eta) = \frac{1}{2}(1-\eta) + \frac{3}{2}\eta\cos^2\theta, \tag{8.1}$$

其中自旋排列参数 η 的取值范围为 $-\frac{1}{2} \le \eta \le 1$。

(a) 假设某反应产生的 ρ^0 中，测量到 n 个 $\cos\theta$ 值。利用矩方法，取 $a = x^2$，构造自旋排列参数的估计量 $\hat{\eta}$。为什么不能用 $a = x$ 构造估计量？

(b) 计算 $\hat{\eta}$ 的期待值和方差。

第九章　统计误差、置信区间和极限

习题 9.1. 假设估计量 $\hat{\theta}$ 服从高斯分布，高斯分布的参数分别为 $\hat{\theta}$ 的真值 θ 和标准偏差 $\sigma_{\hat{\theta}}$。假设 $\sigma_{\hat{\theta}}$ 已知。

(a) 画出定义置信带的函数 $u_\alpha(\theta)$ 和 $v_\beta(\theta)$（参见 *Statistical Data Analysis* 第 *9.2* 节）。

(b) 证明置信水平为 $1-\gamma$ 时参数 θ 的中心置信区间 由下式给出

$$[\hat{\theta} - \sigma_{\hat{\theta}}\phi^{-1}(1-\gamma/2), \hat{\theta} + \sigma_{\hat{\theta}}\phi^{-1}(1-\gamma/2)], \tag{9.1}$$

其中 ϕ^{-1} 是标准高斯分布的分位数。

习题 9.2. 随机变量 x 服从均值为 ξ 的指数分布，考虑对 x 的 n 次观测。参数 ξ 的最大似然估计量 *(见 Statistical Data Analysis (6.6) 式)* 由下式给出

$$\hat{\xi} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{9.2}$$

并且 $\hat{\xi}$ 的概率密度函数（参见 *Statistical Data Analysis (10.25)* 式）为

$$g(\hat{\xi};\xi) = \frac{n^n}{(n-1)!}\frac{\hat{\xi}^{n-1}}{\xi^n}e^{-n\hat{\xi}/\xi}. \tag{9.3}$$

(a) 证明：定义置信带的曲线 $u_\alpha(\xi)$ 和 $v_\beta(\xi)$ 为

$$\begin{aligned} u_\alpha(\xi) &= \frac{\xi}{2n}F^{-1}_{\chi^2}(1-\alpha;2n), \\ v_\beta(\xi) &= \frac{\xi}{2n}F^{-1}_{\chi^2}(\beta;2n), \end{aligned} \tag{9.4}$$

其中 $F^{-1}_{\chi^2}$ 为 χ^2 分布的分位数。根据习题2.6，χ^2 分布的累积分布可以与不完全伽马函数 $P(x,n)$ 联系起来：

$$F_{\chi^2}(2x;2n) = P(x,n) \equiv \frac{1}{\Gamma(n)}\int_0^x e^{-t}t^{n-1}dt. \tag{9.5}$$

取 $\alpha=\beta=0.159$，$n=5$，画出 $u_\alpha(\xi)$ 和 $v_\beta(\xi)$。*(χ^2 分布的分位数可以从标准分布表中查出，或者在 ROOT 中调用* `TMath::ChisquareQuantile(Double_t p, Double_t ndf)` *函数得到。)*

(b) 求出置信区间 $[a,b]$ 作为估计值 $\hat{\xi}$、样本容量 n 以及置信水平 α 和 β 的函数。假设估计值为 $\hat{\xi}=1.0$，在 $u_\alpha(\xi)$ 和 $v_\beta(\xi)$ 的图上画出该估计量的值。取 $n=5$，$\alpha=\beta=0.159$，计算 a 和 b。将计算结果与估计值加减一倍标准差得到的区间进行比较。

习题 9.3. 证明二项分布的参数 p 的上限和下限为

$$\begin{aligned} p_{lo} &= \frac{nF^{-1}_F[\alpha;2n,2(N-n+1)]}{N-n+1+nF^{-1}_F[\alpha;2n,2(N-n+1)]} \\ p_{up} &= \frac{(n+1)F^{-1}_F[1-\beta;2(n+1),2(N-n)]}{(N-n)+(n+1)F^{-1}_F[1-\beta;2(n+1),2(N-n)]}. \end{aligned} \tag{9.6}$$

其中上下限的置信水平分别为 $1-\alpha$ 和 $1-\beta$，n 为 N 次试验中成功的次数，F_F^{-1} 为 F 分布的分位数，由 F 分布定义：

$$f(x;n_1,n_2)=\left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{\Gamma(\frac{1}{2}(n_1+n_2))}{\Gamma(\frac{1}{2}n_1)\Gamma(\frac{1}{2}n_2)}x^{n_1/2-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2}, \tag{9.7}$$

其中 $x>0$，参数 n_1 和 n_2 为整数 *(自由度)*。利用二项分布累积分布函数与自由度为 $n_1=2(n+1)$ 和 $n_2=2(N-n)$ 的累积分布函数 $F_F(x)$ 的关系[1]

$$\sum_{k=0}^{n}\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}=1-F_F\left[\frac{(N-n)p}{(n+1)(1-p)};2(n+1),2(N-n)\right]. \tag{9.8}$$

F 分布的分位数可以从标准分布表中查得，或者调用 *ROOT* 中的函数计算。

习题 9.4. *在 CERN 的 Gargamelle 气室中进行了反中微子-核子散射实验。选择的事例是只产生强子的事例 (通过中性流过程 $\overline{\nu}_\mu N\to\overline{\nu}_\mu X$)，或者是产生强子和一个 μ 子的事例 (通过带电流过程 $\overline{\nu}_\mu N\to\mu^+X$)。从 212 个事例样本中，64 个事例归类为中性流 (NC) 过程，148 个归类为带电流 (CC) 过程。估计事例为中性流 (NC) 的概率，并求出 68.3% 的中心置信区间。求 NC 与 CC 过程概率之比的估计量和置信区间。*[2]

习题 9.5. *在研究粒子碰撞的实验中，选择出来 10 个某种类型的事例，比如说某个量 x 具有较高值的事例。这 10 个高 x 值的事例中，发现有 2 个事例包含 μ 子。*

(a) 高 x 值事例中包含 μ 子的事例数目服从二项分布，求参数 p 的 68.3% 中心置信区间。将结果表示为 $p=\hat{p}^{+c}_{-d}$，其中 $\hat{p}$ 为 p 的最大似然估计，$[\hat{p}-c,\hat{p}+d]$ 为置信区间。

(b) 将 (a) 中的区间与 $\hat{p}\pm\hat{\sigma}_{\hat{p}}$ 进行比较，其中 $\hat{\sigma}_{\hat{p}}$ 为 $\hat{p}$ 的标准偏差的估计量。

(c) 经常犯的错误是将高 x 值的事例数 10 当作随机变量，并将其方差引入 $\hat{p}$ 的误差中 (例如通过误差传递)。为什么这种方法是不正确的？

习题 9.6. *假设为了产生习题 (9.5) 中的事例，收集的总数据对应于积分亮度为 $L=1\,\mathrm{pb}^{-1}$(误差可以忽略)。产生的给定类型事例总数可以看作均值为 $\nu=\sigma L$ 的泊松随机变量 n，其中 σ 为产生截面。(为什么是泊松分布？)*

(a) 对于高 x 值事例以及高 x 值的 μ 子事例，假设观测到的事例数分别为 $n_x=10$ 和 $n_{x\mu}=2$，求事例数期待值 ν_x 和 $\nu_{x\mu}$ 的 68.3% 中心置信区间。对应的产生截面 σ_x 和 $\sigma_{x\mu}$ 的置信区间为多少？

(b) 将 (a) 中得到的置信区间与通过加减一倍标准偏差得到的区间进行比较。

(c) 假设另外一个实验的积分亮度为 $L'=100\,\mathrm{pb}^{-1}$，观测到 n'_x 个高 x 值的事例。但是这个实验不能鉴别 μ 子。利用数据 n_x，$n_{x\mu}$ 和 n'_x 构造参数 σ_x 和 $\sigma_{x\mu}/\sigma_x$ 的对数似然函数。证明 p 的最大似然估计量与 n'_x 无关。这是否说明第二个实验的结果对 p 的估计没有影响？

(d) 假设最初的实验没有测量高 x 值的事例数，只是测得了包含 μ 子的高 x 值事例数。利用结果 $n_{x\mu}=2$ 和 n'_x，构造 σ_x 和 p 的对数似然函数，并求出最大似然估计量。利用误差传递估计 $\hat{p}$ 的标准偏差，并比较区间 $\hat{p}\pm\sigma_{\hat{p}}$ 与习题 (9.5) 中 (a) 和 (b) 得到的区间。选作：这种情况下，如何构造 p 的置信区间？

习题 9.7. *相互作用中产生的某粒子以相对于 z 的某一角度发射出来，如图9.1所示。探测器放在距离相互作用顶点为 d 处测量粒子垂直于 z 方向的位置 x。角度 θ 定义为 z 轴与粒子径迹在 (x,z) 平面上投影的夹角。假设测量值 x 可以看做以真值为中心值，标准偏差为 σ_x 的高斯变量。*

[1] 利用 F 分布计算二项分布的置信区间是 A. Hald 提出来的，见 *Statistical Theory with Engineering Applications*, John Wiley, New York, 1952.

[2] 实际实验中还考虑了小本底的修正，参见 F.J. Hasert et al., Observation of neutrino-like interactions without muon or electron in the Gargamelle neutrino experiment. *Phys. Lett.* **46B**(1973) 138.

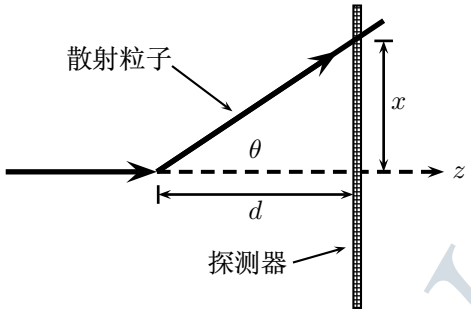

图 9.1: 粒子径迹投影到 (x, z) 平面上的散射角 θ 的定义。

(a) 求 $\cos\theta$ 置信水平为 $1-\gamma$ 的中心置信区间。

(b) 取 $d = 1\text{m}$，$\sigma_x = 1\,\text{mm}$，并假设测量值为 $x = 2.0\,\text{mm}$。求 $\cos\theta$ 置信水平分别为 68.3% 和 95% 的中心置信区间。

第十章　特征函数

习题 10.1. 证明高斯分布

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{10.1}$$

的特征函数为

$$\phi(k)=\exp(i\mu k-\frac{1}{2}\sigma^2k^2). \tag{10.2}$$

习题 10.2. 证明指数分布

$$f(x;\xi)=\frac{1}{\xi}e^{-x/\xi} \tag{10.3}$$

的特征函数为

$$\phi(k)=\frac{1}{1-ik\xi}. \tag{10.4}$$

习题 10.3. 证明自由度为 n 的 χ^2 分布

$$f(z;n)=\frac{1}{2^{n/2}\Gamma(n/2)}z^{n/2-1}e^{-z/2} \tag{10.5}$$

的特征函数为

$$\phi(k)=(1-2ik)^{-n/2}. \tag{10.6}$$

提示：证明中需要用到伽马函数，定义为

$$\Gamma(x)=\int_0^\infty e^{-t}t^{x-1}dt. \tag{10.7}$$

习题 10.4. 假定随机变量 $x_1,\ldots,x_n$ 相互独立，且都服从均值为 μ、方差为 σ^2 的高斯分布。第 *5* 章和第 *6* 章给出，可以用样本均值

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{10.8}$$

作为均值 μ 的估计量。

(a) 求样本均值的特征函数。

(b) 利用特征函数证明 $\overline{x}$ 服从高斯分布，并求其均值和方差。

习题 10.5. 利用特征函数证明高斯分布的前 *4* 阶代数矩为

$$\begin{aligned}E[x]&=\mu\\E[x^2]&=\mu^2+\sigma^2\\E[x^3]&=\mu^3+3\mu\sigma^2\\E[x^4]&=3(\mu^2+\sigma^2)^2-2\mu^4.\end{aligned} \tag{10.9}$$

习题 10.6.

(*a*) 利用特征函数证明自由度为 n 的 χ^2 分布的均值和方差分别为 n 和 $2n$。

(*b*) 假设 z 服从自由度为 n 的 χ^2 分布。证明：在大 n 极限下，χ^2 分布变为均值为 $\mu = n$、方差为 $\sigma^2 = 2n$ 的高斯分布。提示：证明中需要定义变量

$$y = \frac{z-n}{\sqrt{2n}}, \tag{10.10}$$

y 的均值为零，标准差为 *1*。证明 y 的特征函数为

$$\phi_y(k) = e^{-ik\sqrt{n/2}}\phi_z\left(\frac{k}{\sqrt{2n}}\right). \tag{10.11}$$

将 $\phi_y(k)$ 的对数展开，并只保留大 n 极限下不消失的项，然后再变换回变量 z 得到要证明的结果。

习题 10.7. 假设 n 个相互独立的随机变量 $x_1, \ldots, x_n$ 都服从标准高斯分布，即对所有 $i = 1, \ldots, n$，

$$\varphi(x_i) = \frac{1}{\sqrt{2\pi}}e^{-x_i^2/2}. \tag{10.12}$$

考虑下面这个变量的性质：

$$y = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}. \tag{10.13}$$

(*a*) 首先只考虑某一个 x_i。通过变量变换，证明 $u = x_i^2$ 的概率密度函数为

$$f(u) = \frac{1}{\sqrt{2\pi u}}e^{-u/2}. \tag{10.14}$$

这是自由度为 *1* 的 χ^2 分布。

(*b*) 证明 u 的特征函数为

$$\phi_u(k) = \frac{1}{\sqrt{1-2ik}}. \tag{10.15}$$

(*c*) 利用相加定理，求下面变量的特征函数

$$v = \sum_{i=1}^{n} x_i^2. \tag{10.16}$$

(*d*) 利用变量变换，证明 $y = (\sum_{i=1}^{n} x_i^2)^{1/2}$ 的概率密度函数为

$$h(y) = \frac{1}{2^{n/2-1}\Gamma(n/2)}y^{n-1}e^{-y^2/2}. \tag{10.17}$$

(*e*) 写出 $n = 3$ 时的概率密度函数。这是 *Maxwell-Boltzmann* 分布。假设气体中分子的速度分量 v_x，v_y 和 v_z 都服从均值为零、标准差为 σ 的高斯分布。写出分子速度 $v = (v_x^2 + v_y^2 + v_z^2)^{1/2}$ 的概率密度函数。

(*f*) 写出 $n = 1$ 时的概率密度函数。即，如果 x 服从标准高斯分布，则 $y = |x|$ 的概率密度函数是什么？

习题 10.8. 考虑服从柯西分布的随机变量 x，

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}. \tag{10.18}$$

(*a*) 证明其特征函数为

$$\phi(k) = e^{-|k|}. \tag{10.19}$$

(利用留数定理，$k > 0$ 时选取上半平面的路径，$k < 0$ 时选取下半平面的路径。)

(*b*) 考虑柯西随机变量 x 的某个样本，样本容量为 n。利用（*a*）中得到的特征函数并应用相加定理，证明样本均值 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 也服从柯西分布。这是一个特例，即 $\overline{x}$ 的概率密度函数不随样本容量的增加而改变，这与柯西分布的各阶矩不存在有关。

习题 10.9. 狄拉克 δ 函数

$$f(x;\mu) = \delta(x-\mu) \tag{10.20}$$

定义为

$$\begin{aligned} &\delta(x-\mu) = 0, x \neq \mu, \\ &\int_{-\infty}^{\infty} \delta(x-\mu)dx = 1. \end{aligned} \tag{10.21}$$

即，$\delta(x-\mu)$ 在 $x=\mu$ 处为无限尖锐的峰，但在其它地方都等于零。求 $\delta(x-\mu)$ 的特征函数，并据此得到 δ 函数的积分表示。

第十一章　解谱法

习题 11.1. 考虑图 *9.1* 的探测器设备。假设 x 的分辨率函数服从高斯分布

$$s(x|x') = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x-x')^2}{2\sigma_x^2}\right]. \tag{11.1}$$

求 $\cos\theta = a/\sqrt{x^2+a^2}$ 的分辨率函数。

习题 11.2. 对等区间宽度的直方图 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$，考虑 *k=1* 的 *Tikhonov* 正规化函数，

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-1} (\mu_i - \mu_{i+1})^2. \tag{11.2}$$

求 $M \times M$ 维矩阵 G，使得 $S(\boldsymbol{\mu})$ 可以表示成以下形式

$$S(\boldsymbol{\mu}) = -\sum_{i,j=1}^{M} G_{ij}\mu_i\mu_j = -\boldsymbol{\mu}^T G \boldsymbol{\mu}. \tag{11.3}$$

习题 11.3. 考虑期待值为 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ 的直方图，对应的概率为 $\mathbf{p} = \boldsymbol{\mu}/\mu_{\text{tot}}$，其中 $\mu_{\text{tot}} = \sum_{i=1}^{M} \mu_i$。

(a) 证明 *Shannon* 熵

$$H(\mathbf{p}) = -\sum_{i=1}^{M} p_i \log p_i, \tag{11.4}$$

在所有区间 i 的 $p_i = 1/M$ 时最大。（利用 *Lagrange* 乘子法引入限制条件 $\sum_{i=1}^{M} p_i = 1$。）

(b) 证明交叉熵

$$K(\mathbf{p};\mathbf{q}) = -\sum_{i=1}^{M} p_i \log \frac{p_i}{Mq_i}, \tag{11.5}$$

在概率 $\mathbf{p}$ 等于参考分布 $\mathbf{q}$ 时最大。

习题 11.4. 考虑观测到的直方图 $\mathbf{n} = (n_1, \ldots, n_N$，对应的期待值 $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ 与真值直方图 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ 的关系为 $\boldsymbol{\nu} = R\boldsymbol{\mu}$。假设协方差矩阵 V 和响应矩阵R 已知，并且直方图中没有本底。

(a) 通过最大化 $\Phi(\boldsymbol{\mu})$ 构造 $\boldsymbol{\mu}$ 的估计量

$$\Phi(\boldsymbol{\mu}) = -\frac{\alpha}{2}\chi^2(\boldsymbol{\mu}) + S(\boldsymbol{\mu}) \tag{11.6}$$

$$= -\frac{\alpha}{2}(\mathbf{n} - R\boldsymbol{\mu})^T V^{-1}(\mathbf{n} - R\boldsymbol{\mu}) - \boldsymbol{\mu}^T G \boldsymbol{\mu}, \tag{11.7}$$

其中 α 为正规化参数，$M \times M$ 对称矩阵 G 由已知常数确定 *(参考 Statistical Data Analysis 中 11.5.1 节)*。证明估计量 $\hat{\boldsymbol{\mu}}$ 为

$$\hat{\boldsymbol{\mu}} = (\alpha R^T V^{-1} R + 2G)^{-1} \alpha R^T V^{-1} \mathbf{n}, \tag{11.8}$$

并求协方差矩阵 $U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$.

(b) 考虑限制条件 $\nu_{\text{tot}} = \sum_{i=1}^{N} \nu_i = \sum_{i=1}^{N} \sum_{j=1}^{M} R_{ij}\mu_j$ 等于总观测事例数 $n_{\text{tot}} = \sum_{i=1}^{N} n_i$，通过对参数 $\boldsymbol{\mu}$ 和 *Lagrange* 乘子 λ 最大化 $\varphi(\boldsymbol{\mu})$ 求得结果

$$\varphi(\boldsymbol{\mu}) = -\frac{\alpha}{2}(\mathbf{n} - R\boldsymbol{\mu})^T V^{-1}(\mathbf{n} - R\boldsymbol{\mu}) - \boldsymbol{\mu}^T G\boldsymbol{\mu} + \lambda(n_{\text{tot}} - \nu_{\text{tot}}). \tag{11.9}$$

求估计量 $\hat{\boldsymbol{\mu}}$ 及其协方差。

(c) 利用 *Statistical Data Analysis* 中方程 *(11.76)*，对 *(a)* 和 *(b)* 两种情况分别构造偏置 $\mathbf{b} = E[\hat{\boldsymbol{\mu}}] - \boldsymbol{\mu}$ 的估计量 $\hat{\mathbf{b}}$。