第一章 基本概念

1.1 概率与随机变量

概率的定义

随机性的大小可以用概率的概念定量描述。
考虑某集合 S（称为样本空间）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（柯尔莫哥洛夫公理）定义：

$$
\begin{gathered}
P(A) \geq 0, \forall A \subset S \\
P(S)=1 \\
\text{若} A \cap B=\emptyset \quad P(A \cup B)=P(A)+P(B)
\end{gathered}
$$

由此可得如下性质

$$
\begin{gathered}
P(\bar{A})=1-P(A) \quad \text{其中} \bar{A} \text{为} A \text{的补集} \\
P(A \cup \bar{A})=1 \\
0 \leq P(A) \leq 1 \\
P(\emptyset)=0 \\
A \subset B \text{则} P(A) \leq P(B) \\
P(A \cup B)=P(A)+P(B)-P(A \cap B)
\end{gathered}
$$

条件概率与独立性

给定事件 B 的条件下，事件A发生的条件概率定义为

$$
P(A|B)=\frac{P(A \cap B)}{P(B)} \quad P(B) \neq 0
$$

此即贝叶斯定理，也可写为

$$
P(A \cap B)=P(A|B)P(B)=P(B|A)P(A)
$$

若 $P(A \cap B)=P(A)P(B)$，则称 A 与 B 相互独立，此时有

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

可证明条件概率本身也满足概率的三个公理；此外概率 $P(A)$ 可以看作给定 S 时 A 的条件概率 $P(A)=P(A|S)$

全概率公式

假设样本空间 S 可以划分成若干两两互斥的子集 A_i，即 $S=\cup_i A_i$ 且 $i \neq j$ 时 $A_i \cap A_j=\emptyset$，再假设 $\forall i, P(A_i) \neq 0$。
则 S 的任意子集 B 可表示为

$$
B=B \cap S=B \cap (\cup_i A_i)=\cup_i (B \cap A_i)
$$

则有全概率公式

$$
P(B)=P(\cup_i (B \cap A_i))=\sum_i P(B \cap A_i)=\sum_i P(B|A_i)P(A_i)
$$

此时贝叶斯定理也可写为

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

贝叶斯理论与主观概率

贝叶斯理论通常用于主观概率问题，有 $\text{后验概率}=\text{似然性} \cdot \text{先验概率}$，即

$$
P(\text{理论}|\text{实验})=\frac{P(\text{实验}|\text{理论})}{P(\text{实验})}P(\text{理论})
$$

如

$$
P(y|+)=\frac{P(+|y)P(y)}{P(+|y)P(y)+P(+|n)P(n)}
$$

1.2 概率的解释

相对频率解释

主观概率（贝叶斯概率）

1.3 概率密度函数

概率密度

观测值在无限小区间 $[x, x+\mathrm{d}x]$ 内的概率可由概率密度函数（PDF）$f(x)$给出

$$观测到x处于区间[x, x+\mathrm{d}x]的概率 = f(x)\mathrm{d}x \quad x \in \mathbb{R}$$

概率密度函数 $f(x)$ 是归一化的

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

概率密度函数 $f(x)$ 的累积分布（CDF）$F(x)$ 为

$$F(x) = \int_{-\infty}^{x} f(x')\mathrm{d}x'$$

概率密度函数可以定义为

$$f(x) = \frac{\partial F(x)}{\partial x}$$

设 X 为随机变量，$0 < \alpha < 1$，若 x 满足 $P(X \leq x_\alpha) = F(x_\alpha) = \alpha$，则称 x 为 X 的 α-分位数（下侧 α-分位数）

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

多维随机变量

联合概率密度函数（jPDF）$f(x,y)$ 定义为

$$P(A \cap B) = x处在[x, x+\mathrm{d}x]且y处在[y, y+\mathrm{d}y]的概率 = f(x,y)\mathrm{d}x\mathrm{d}y$$

归一化条件

$$\iint_S f(x,y)\mathrm{d}x\mathrm{d}y = 1$$

边缘概率密度

$$f_x(x) = \int f(x,y)\mathrm{d}y \quad f_y(y) = \int f(x,y)\mathrm{d}x$$

若 $f(x,y) = f_x(x)f_y(y)$，则随机变量 x, y 相互独立

而给定 x 时 y 的条件概率密度函数（cPDF）$h(y|x)$ 定义为

$$h(y|x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y\prime)\mathrm{d}y\prime}$$

1.4随机变量的函数

一维随机变量的函数

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 $f(x)$
假设函数 $a(x)$ 单调，$a \in [a, a+\mathrm{d}a]$ 的概率 $g(a)\mathrm{d}a$ 等于 $x \in \mathrm{d}S$ 的概率 $f(x)\mathrm{d}x$，则有

$$g(a) = f(x(a))\left|\frac{\mathrm{d}x}{\mathrm{d}a}\right|$$

如果函数 $a(x)$ 的逆不唯一，需要将 $\mathrm{d}S$ 包含的多段 $\mathrm{d}x$ 区间全部考虑进来。

多维随机变量的函数

考虑多维随机变量 $\vec{x} = (x_1, \cdots, x_n)$ 与函数 $a(\vec{x})$
已知 $\vec{x}$ 的概率密度 $f(\vec{x}) = f(x_1, \cdots, x_n)$，要求 a 的概率密度 $g(a)$

$$g(a')\,\mathrm{d}a' = \int \cdots \int_{\mathrm{d}S} f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

其中 $\mathrm{d}S$ 定义为在 $a(\vec{x}) = a'$ 和 $a(\vec{x}) = a' + \mathrm{d}a'$ 定义的两个曲面之间的 $\vec{x}$ 空间范围

特殊的，若如果两个随机变量 $x,y>0$，服从联合概率密度 $f(x,y)$，考虑函数 $z=xy$，则其概率密度函数 $g(z)$ 有

$$g(z)\mathrm{d}z=\iint f(x,y)\mathrm{d}x\mathrm{d}y=\int_0^\infty \mathrm{d}x\int_{\frac{z}{x}}^{\frac{z+\mathrm{d}z}{x}} f(x,y)\mathrm{d}y$$
$$\therefore g(z)=\int_0^\infty f(x,\frac{z}{x})\frac{\mathrm{d}x}{x}=\int_0^\infty f(\frac{z}{y},y)\frac{\mathrm{d}y}{y}$$

若如果两个随机变量 $x,y>0$，服从联合概率密度 $f(x,y)$，考虑函数 $z=x+y$，则其概率密度函数 $g(z)$ 有

$$g(z)\mathrm{d}z=\iint f(x,y)\mathrm{d}x\mathrm{d}y=\int_0^\infty \mathrm{d}x\int_{z-x}^{z+\mathrm{d}z-x} f(x,y)\mathrm{d}y$$
$$\therefore g(z)=\int_0^\infty f(x,z-x)\mathrm{d}x=\int_0^\infty f(z-y,y)\mathrm{d}y$$

若随机变量 x,y 相互独立，分别服从 $g(x)$ 与 $h(y)$ 分布，则函数 $z=xy$ 的概率密度函数 $f(z)$ 为梅林卷积，有

$$f(z)=\int_{-\infty}^\infty g(x)h(\frac{z}{x})\frac{\mathrm{d}x}{|x|}=\int_{-\infty}^\infty g(\frac{z}{y})h(y)\frac{\mathrm{d}y}{|y|}$$

函数 $z=x+y$ 的概率密度函数 $f(z)$ 为傅里叶卷积，有

$$f(z)=\int_{-\infty}^\infty g(x)h(z-x)\mathrm{d}x=\int_{-\infty}^\infty g(z-y)h(y)\mathrm{d}y$$

考虑随机矢量 $\vec{x}=(x_1,\dots,x_n)$，联合概率密度为 $f(\vec{x})$，构造 n 个线性独立的函数 $\vec{a}(\vec{x})=(a_1(\vec{x}),\dots,a_n(\vec{x}))$，并且其逆函数 $x_1(\vec{a}),\dots,x_n(\vec{a})$ 存在。

则 $\vec{a}$ 的联合概率密度为

$$\boldsymbol{g}(\overrightarrow{\boldsymbol{a}})=|\boldsymbol{J}|\boldsymbol{f}(\overrightarrow{\boldsymbol{x}})$$

其中 J 是雅可比行列式

$$J=\left|\frac{\partial\vec{x}}{\partial\vec{a}}\right|$$

对联合概率密度 $g(\vec{a})$ 积分掉其他不关心的变量，可以得到任意一个边缘概率密度 $g_i\left(a_i\right)$。这是数据分析中误差传递的基础。

1.5 期望值

期待值、方差、标准差

假设随机变量 x 服从概率密度函数为 $f(x)$，则 x 的期望值 $E[x]$（不是x的函数，而是 $f(x)$ 的泛函）定义为

$$E[x]=\int_{-\infty}^\infty xf(x)\mathrm{d}x=\mu$$

对于随机变量的函数 $a(x)$，其期望值

$$E[a]=\int_{-\infty}^\infty ag(a)\mathrm{d}a=\int_{-\infty}^\infty a(x)f(x)\mathrm{d}x=\mu$$

可考虑特殊的期望值，如 x 的 n 阶代数距

$$E[x^n]=\int_{-\infty}^\infty x^nf(x)\mathrm{d}x=\mu_n$$

x 的 n 阶中心距

$$E[(x-E[x])^n]=\int_{-\infty}^\infty (x-\mu)^nf(x)\mathrm{d}x=v_n$$

中心矩与代数矩存在多项式关系

$$v_k=\sum_{i=0}^k C_k^i\mu_i(-\mu_1)^{k-i}$$

一阶代数矩即为随机变量的期望值 $\mu=\mu_1$，而二阶中心矩

$$E[(x-E[x])^2]=E[x^2]-\mu^2=\int_{-\infty}^\infty (x-\mu)^2f(x)\mathrm{d}x=\sigma^2=V[x]$$

为 x 的总体方差，其平方根 σ 为标准差

协方差与相关系数

定义协方差 $\mathrm{cov}[x,y]$ (也可用矩阵 V_{xy} 表示)为

$$\mathrm{cov}[x,y]=E[(x-\mu_x)(y-\mu_y)]=E[xy]-\mu_x\mu_y$$

相关系数（皮尔逊相关系数）定义为无量纲数

$$\rho_{xy}=\frac{\mathrm{cov}[x,y]}{\sigma_x\sigma_y}$$

如果 x,y 相互独立，则

$$f(x,y)=f_x(x)f_y(y)$$
$$E[xy]=\iint xyf(x,y)\mathrm{d}x\mathrm{d}y=\mu_x\mu_y$$

则 $\mathrm{cov}[x,y]=0\quad \rho_{xy}=0$，即 x,y 不相关

特征函数

设 x 是一个随机变量，则称 e^{itx} 的数学期望值，即

$$\varphi(t)=\int_{-\infty}^{\infty}\mathrm{e}^{itx}f(x)\mathrm{d}x=E[\mathrm{e}^{itx}]\qquad -\infty<t<\infty$$

为随机变量 x 的特征函数。

特征函数有性质

$$|\varphi(t)|\le\varphi(0)=1$$
$$\varphi(-t)=\overline{\varphi(t)}$$
$$\varphi_{ax+b}(t)=\mathrm{e}^{ibt}\varphi_x(at)$$

若随机变量 x,y 独立，则

$$\varphi_{x+y}(t)=\varphi_x(t)\varphi_y(t)$$

此外，有级数展开

$$\varphi(t)=E\left[\sum_{n=0}^{\infty}\frac{(-it)^n}{n!}x^n\right]=\sum_{n=0}^{\infty}\frac{(-it)^n}{n!}E\left[x^n\right]$$
$$\mathrm{e}^{itx_0}\varphi(t)=E[\mathrm{e}^{it(x-x_0)}]=\sum_{n=0}^{\infty}\frac{(-it)^n}{n!}E\left[(x-x_0)^n\right]$$

1.6 误差传递

假设我们对某个量测量了一组值 $\vec{x}=(x_1,\dots,x_n)$ 并得到其协方差 $V_{ij}=\mathrm{cov}[x_i,x_j]$，考虑一函数 $y(\vec{x})$ 的方差 $V[y]$

假设已知 $\vec{\mu}=E[\vec{x}]$，对 y 在 $\vec{\mu}$ 处作一阶泰勒展开

$$y(\vec{x})\approx y(\mu)+\sum_{i=1}^{n}\left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}}(x_i-\mu_i)$$

由于$E[x_i-\mu_i]=0$，则近似到一阶项，y 的期望值为

$$E[y(\vec{x})]\approx y(\vec{\mu})$$

y^2 的期望值为

$$E[y^2(\vec{x})]\approx y^2(\vec{\mu})+2y(\vec{\mu})\sum_{i=1}^{n}\left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}}E[x_i-\mu_i]$$
$$+E\left[\left(\sum_{i=1}^{n}\left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}}(x_i-\mu_i)\right)\left(\sum_{j=1}^{n}\left(\frac{\partial y}{\partial x_j}\right)_{\vec{x}=\vec{\mu}}(x_j-\mu_j)\right)\right]$$
$$=y^2(\vec{\mu})+\sum_{i,j=1}^{n}\left(\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right)_{\vec{x}=\vec{\mu}}V_{ij}$$
$$\therefore\sigma_y^2=E[y^2]-E^2[y]\approx\sum_{i,j=1}^{n}\left(\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right)_{\vec{x}=\vec{\mu}}V_{ij}$$

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

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

也可用矩阵形式记为 $U=AVA^T$，其中导数矩阵 $A_{ij}=\left(\dfrac{\partial y_i}{\partial x_j}\right)_{\vec{x}=\vec{\mu}}$

此即误差传递。

特殊的，若 $y=x_1+x_2$，则

$$\sigma_y^2=\sigma_1^2+\sigma_2^2+2\mathrm{cov}[x_1,x_2]$$

若 $y=x_1x_2$，则

$$\frac{\sigma_y^2}{y^2}=\frac{\sigma_1^2}{x_1^2}+\frac{\sigma_2^2}{x_2^2}+\frac{2\mathrm{cov}[x_1,x_2]}{x_1x_2}$$

可用正交变换消除随机变量间的相关性。

若 $E[x]=\mu$ 小于标准差或者与之相当，则误差传递的近似不成立，此时需使用蒙特卡洛方法或置信区间处理。

第二章 常用概率函数

2.1 二项分布和多项分布

二项分布

N 次独立测量(伯努利试验)，每次只有成功（概率始终为 p）或失败（概率为 $1-p$）两种可能。
定义离散随机变量 n 为成功的次数，$1\leq n\ln N$，则 n 服从二项分布

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

可证明其满足归一化条件

$$\sum_{n=1}^{N}f(n;N,p)=\sum_{n=1}^{N}\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}=[p+(1-p)]^N=1$$

适用于衰变分支比、探测效率不确定度的计算

n 的均值

$$\begin{aligned}E[n]&=\sum_{n=0}^{N}nf(n;N,p)=\sum_{n=1}^{N}\frac{N!}{(n-1)!(N-n)!}p^n(1-p)^{N-n}\\&=Np\sum_{n=1}^{N}\frac{(N-1)!}{(n-1)!(N-n)!}p^{n-1}(1-p)^{N-n}=Np\end{aligned}$$

n 的方差

$$\begin{aligned}V[n]&=E[n^2]-E^2[n]=\sum_{n=0}^{N}n^2f(n;N,p)-N^2p^2\\&=N(N-1)p^2+Np-(Np)^2=Np(1-p)\end{aligned}$$

二项分布适用条件

伯努利试验；每次尝试仅有两种可能性；每次尝试的成功概率是一样的；不同次尝试的结果是独立的。

如，考虑效率和效率不确定度的估计

多层阻性板室 MRPC 的探测效率

$$\varepsilon=\frac{N'}{N}$$

其中 N' 为 MRPC 记录的粒子数，N 为穿过 MRPC 的粒子数（闪烁体 1 与 2 同时击中）

MRPC 的探测效率的不确定度

$$\Delta\varepsilon=\frac{\Delta N'}{N}=\frac{\sqrt{N\varepsilon(1-\varepsilon)}}{N}=\sqrt{\frac{\varepsilon(1-\varepsilon)}{N}}$$

多项分布

多项分布是二项分布的推广，即每次试验的输出结果存在 m 种不同的结果。对于某次试验，第 i 种输出的概率为 p_i（要求 $\sum_{i=1}^m p_i = 1$）

N 次试验，得到的结果可用 m 维矢量表示 $\vec{n} = (n_1, \cdots, n_m)$，$\sum_{i=1}^m n_i = N$

$\vec{n}$ 是是服从多项分布的随机变量

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}$$

若将 m 种输出分成两类：输出 i 和不输出 i，则与先前的二项分布一致，有

$$E[n_i] = Np_i$$
$$V[n_i] = Np_i(1 - p_i)$$

如果考虑有三种可能的输出：i, j 以及所有其他输出，则

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

则 $i \neq j$ 时协方差 $V_{ij} = \mathrm{cov}[n_i, n_j] = E[(n_i - E[n_i])(n_j - E[n_j])] = -Np_ip_j$，因此有

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

这表明任意两个区间的事例数都是负相关的。

2.2泊松分布

考虑二项分布随机变量 $n \in \mathbb{N}$ 在 $N \to \infty, p \to 0, E[n] = Np \to \nu$ 下的极限，有泊松分布

$$f(n; \nu) = \pi(\nu) = \frac{\nu^n}{n!} \mathrm{e}^{-\nu} \quad n \in \mathbb{N}$$

随机变量 n 的期望值

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n}{n!} \mathrm{e}^{-\nu} = \nu$$

方差

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

泊松分布是二项分布的近似。

2.3 均匀分布

连续变量 $x \in (-\infty, \infty)$ 的均匀分布定义为

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

x 的均值和方差分别为

$$E[x] = \int_\alpha^\beta \frac{x}{\beta - \alpha} \mathrm{d}x = \frac{1}{2}(\alpha + \beta)$$
$$V[x] = \int_\alpha^\beta \frac{(x - \frac{\alpha + \beta}{2})^2}{\beta - \alpha} \mathrm{d}x = \frac{1}{12}(\beta - \alpha)^2$$

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 $y = F(x)$，即满足 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 $y = F(x)$，有

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)$$

由此可以得到 y 的概率密度函数为

$$g(y) = f(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| = 1 \quad 0 \leq y \leq 1$$

均匀分布是用蒙特卡罗模拟随机现象的基础。

2.4 指数分布

连续随机变量 $x \in [0, \infty)$ 的指数分布定义为

$$f(x;\xi) = Exp(\xi) = \begin{cases} \frac{1}{\xi}\mathrm{e}^{-\frac{x}{\xi}} & x \geq 0 \\ 0 & \text{其他} \end{cases}$$

x 的均值和方差分别为

$$E[x] = \frac{1}{\xi}\int_0^\infty x\mathrm{e}^{-\frac{x}{\xi}}\mathrm{d}x = \xi$$
$$V[x] = \frac{1}{\xi}\int_0^\infty (x-\xi)^2\mathrm{e}^{-\frac{x}{\xi}}\mathrm{d}x = \xi^2$$

指数分布没有记忆性

$$f(x - x_0 | x \geq x_0) = f(x)$$

2.5 高斯分布

高斯分布

连续随机变量 x 的高斯分布定义为

$$f(x;\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$E[x] = \frac{1}{\xi}\int_{-\infty}^\infty x\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x = \mu$$
$$V[x] = \frac{1}{\xi}\int_{-\infty}^\infty (x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x = \sigma^2$$

特殊的，当 $\mu = 0, \sigma = 1$ 时所定义标准高斯分布的概率密度函数

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$$

对应的累积分布

$$\Phi(x) = \int_{-\infty}^x \varphi(x')\mathrm{d}x'$$

可证明若 x 服从均值为 μ 方差为 σ^2 的高斯分布，则变量 $x = \frac{y-\mu}{\sigma}$ 服从标准高斯分布 $\varphi(x)$

中心极限定理

对于 n 个独立的随机变量 x，均值和方差分别为 μ_i 和 σ_i^2，如果每个 x 的方差存在，那么这些变量之和构成的随机变量 $y = \sum_{i=1}^n x_i$ 在 $n \to \infty$ 的极限下，服从高斯分布 $N(\mu, \sigma^2)$，其中

$$\mu = \sum_{i=1}^n \mu_i \qquad \sigma^2 = \sum_{i=1}^n \sigma_i^2$$

此外，对于 n 有限的情况，如果这 n 个变量之和的涨落不是有一个或少数变量主导，那么中心极限定理近似成立。

二项分布在 $N \to \infty, p \to 0, Np = \nu = \mu$ 时为泊松分布，而二项分布与泊松分布分别在 $N \to \infty$ 与 $\nu \to \infty$ 时为高斯分布

多维高斯分布

随机变量 $\vec{x} = (x_1, \cdots, x_n)$ 的多维高斯函数概率密度为

$$f(\vec{x};\vec{\mu},V) = \frac{1}{(2\pi)^{\frac{n}{2}}|V|^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu})}$$

其期望值、方差和协方差分别为

$$E[x_i] = \mu_i \quad V[x_i] = V_{ii} \quad \mathrm{cov}[x_i, x_j] = V_{ij}$$

对二维情形，概率密度函数为

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

其中相关系数

$$\rho = \frac{\mathrm{cov}[x_1, x_2]}{\sigma_1 \sigma_2}$$

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 $x = e^y$ 服从对数正态分布，其概率密度函数

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}\mathrm{e}^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$
$$E[x] = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2}$$
$$V[x] = \mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2}-1)$$

2.7 卡方分布

如果 $x_1, \cdots, x_n$ 是相互独立的高斯随机变量，定义 $z = \sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2} \geq 0$ 服从自由度为 n 的卡方分布

$$f(z;n) = \chi^2(n) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}z^{\frac{n}{2}-1}\mathrm{e}^{-\frac{z}{2}}$$

其中 Γ 函数

$$\Gamma(x) = \int_0^\infty x^{r-1}\mathrm{e}^{-x}\mathrm{d}x$$

z 的均值和方差为

$$E[z] = n \quad V[z] = 2n$$

若 x_i 不独立但服从 N 维高斯分布，则变量 $z = (\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu})$ 服从自由度为 N 的卡方分布。

卡方分布通常用来检验假设与实际情况的符合程度，如最小二乘法拟合的拟合优度检验。

2.8 柯西(布莱特-魏格纳)分布

连续随机变量 $x \in (-\infty, \infty)$ 的柯西（布莱特-魏格纳）概率密度函数定义为

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

柯西分布的期望值没有定义，$E[|x|]$ 发散。

粒子物理中的布莱特-魏格纳分布的一般形式为

$$f(x) = \frac{1}{\pi}\frac{\frac{\Gamma}{2}}{\frac{\Gamma^2}{4}+(x-x_0)^2}$$

其中 x_0 为模，Γ 为半高全宽，常用于描述“共振态”粒子的不变质量分布。

2.9 朗道分布

电离能损

贝塔分布

效率的先验概率密度

伽马分布

指数分布随机变量的和

学生氏分布

尾部可调的分辨率函数

几何分布

在 n 次伯努利试验中，试验 k 次才得到第一次成功的机率

$$P(k) = (1-p)^{k-1}p \quad k \in \mathbb{Z}^+$$
$$E[k] = \frac{1}{p}$$

$$V[k]=\frac{1-p}{p^2}$$

第四章 统计检验

4.1 假设、检验统计量、显著性水平和功效

假设

统计检验的目的是表述观测数据与预期概率(即假设)之间的符合程度。

待考察的假设通常称为原假设或零假设 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；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。

给定 H 时 x 的概率又称作假设 H 的似然值 $L(x|H)$

检验

检验的目标是，根据观测数据 x 对可能的假设的正确性给出某种论断

考虑简单假设 H_0 和备择假设 H_1。对 H_0 的检验定义为：
对数据样本指定一个临界域 W，使得在 H_0 正确的情况下，观测到这个数据的概率不超过某个小概率 α，即

$$P(x\in W|H_0)\leq\alpha$$

其中 α 检验的显著性水平或检验的大小
如果在临界域观测到 x，则拒绝 H_0；临界域又称为拒绝域，其补集称为接受域。（拒绝假设 H_0 并不等价于我们相信 H_0 为假而 H_1 为真）

通常存在无穷多个可能的临界域可给出相同的显著性水平 α
所以，对 H_0 检验的临界域的选择需要考虑备择假设 H_1

临界域的选择应当满足：临界域内 H_0 为真的概率较小， H_1 为真的概率较大。

在贝叶斯统计中，假设的概率（信心度）可由贝叶斯定理给出：

$$P(H|x)=\frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H)\mathrm{d}H}$$

依赖于先验概率 $\pi(H)$

第一类错误和第二类错误

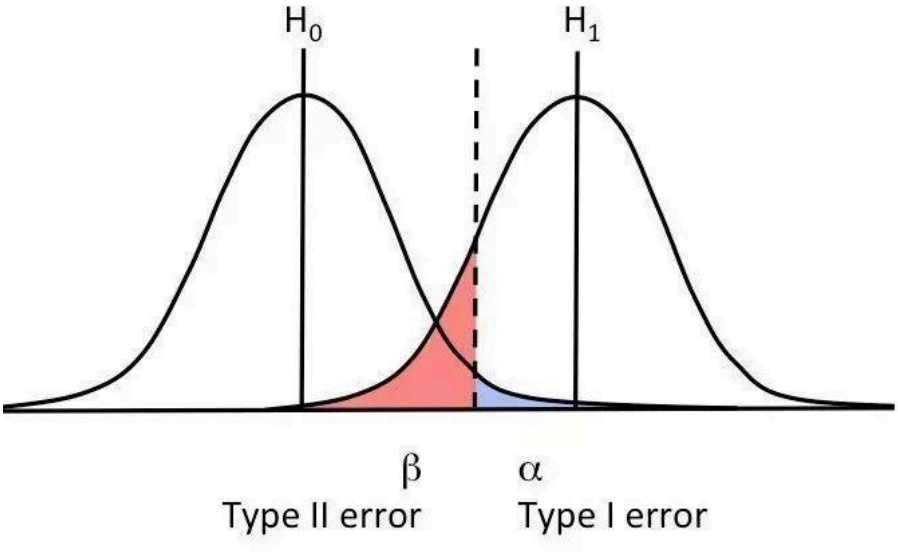
如果假设 H_0 为真而被拒绝，称为第一类错误，或弃真错误。第一类错误的最大概率等于检验的显著性水平

$$P(x\in W|H_0)\leq\alpha$$

如果 H_0 假设被接受，但真假设不是 H_0 而是某个备择假设 H_1，称为第二类错误，或取伪错误，概率为

$$P(x\in S-W|H_1)=\beta$$

$1-\beta$ 为相对于备择假设 H_1 的检验的功效



临界域的选择

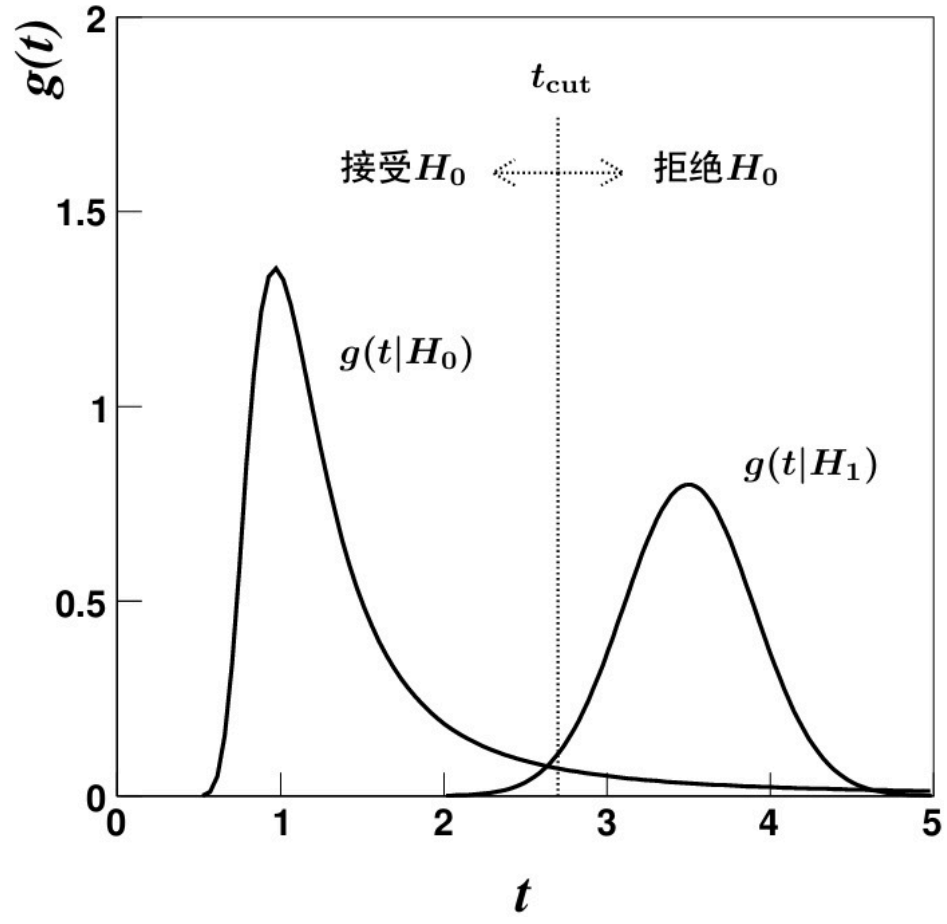
双侧检验和单侧检验

4.2 信号本底的甄别（粒子选择的统计检验）

检验统计量

考虑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。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 $H_0 = e$，$H_1 = \pi$。

希望根据判选条件 $t \leq t_{\text{cut}}$ 选出电子样本。（e 为信号假设，π 为本底假设）
（对于 n 维数据空间，临界域的边界可取为标量检验统计量 $t(x_1, \cdots, x_n) = t_{\text{cut}}$，以便将 n 维问题约化为一维问题）



H_0 为真却拒绝 H_0 的概率（弃真错误）

$$\alpha = \int_W g(\vec{x}|H_0)\mathrm{d}\vec{x}$$

H_1 为真却接受 H_0 的概率（取伪错误）

$$\beta = \int_{\bar{W}} g(\vec{x}|H_0)\mathrm{d}\vec{x}$$

则接受某给定类型粒子的概率（选择效率）分别为

$$\varepsilon_e = \int_{-\infty}^{t_{\mathrm{cut}}} g(t|e)\mathrm{d}t = 1-\alpha$$
$$\varepsilon_\pi = \int_{-\infty}^{t_{\mathrm{cut}}} g(t|\pi)\mathrm{d}t = \beta$$

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

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

其中 a_e 和 $a_\pi = 1-a_e$ 分别为电子和 π 介子的比率

纯度

利用贝叶斯定理，观测值为t时，粒子为电子或 π 介子的概率 $h(e|t)$ 和 $h(\pi|t)$

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

其中 a_e 和 $a_\pi = 1-a_e$ 分别为假设 e 和假设 π 的验前概率

事例选择的纯度信号为事例被正确分类的概率趣，即根据 $t \le t_{\mathrm{cut}}$ 选择出来的电子候选者样本中电子的比率

$$p_e = \frac{t \le t_{\mathrm{cut}}\text{的电子数}}{t \le t_{\mathrm{cut}}\text{的所有粒子数}}$$
$$= \frac{\int_{-\infty}^{t_{\mathrm{cut}}} a_e g(t|e)\mathrm{d}t}{\int_{-\infty}^{t_{\mathrm{cut}}} (a_e g(t|e) + (1-a_e)g(t|\pi))\mathrm{d}t}$$

此即电子概率 $h(e|t)$ 在 $(-\infty, t_{\mathrm{cut}}]$ 区间内的平均值

$$p_e = \frac{\int_{-\infty}^{t_{\mathrm{cut}}} h(e|t) f(t;a_e)\mathrm{d}t}{\int_{-\infty}^{t_{\mathrm{cut}}} f(t;a_e)\mathrm{d}t}$$

4.3 用奈曼皮尔逊引理选择拒绝域

多维检验统计量 $\vec{x} = (x_1, \cdots, x_m)$，原假设 H_0，备择假设 H_1，需选择一个最佳的临界域，有奈曼-皮尔逊引理：在给定效率条件下，要得到最高纯度的信号样本，或者在给定显著性水平下得到最高功效，可以选下列接受域来实现

$$\frac{g(\vec{x}|H_0)}{g(\vec{x}|H_1)} > c$$

利用奈曼皮尔逊引理确定接受域，实际上等价于左式的比值给出的一维检验统计量

$$r \equiv \frac{g(\vec{x}|H_0)}{g(\vec{x}|H_1)}$$

这个比值称为简单假设 H_0 和 H_1 的似然比，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 $\vec{x} = (x_1, \cdots, x_m)$，并且我们希望以此构造一维检验统计量 $t(x)$，用来区分简单假设 H_0 和 H_1。

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

$$t(\vec{x}) = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)}$$

利用多维直方图确定 $f(\vec{x}|H_0)$ 与 $f(\vec{x}|H_1)$ 以近似得到似然比在 $\vec{x}$ 的维数 n 很大时不可行，但仍可对 $t(x)$ 的函数形式做一个简单的拟设，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。

线性检验统计量、费舍尔甄别函数

为了简化问题，可以采用线性变换方法给出包含少量参数的检验统计量，并确定参数，最大限度地区分 H_0 与 H_1。有线性变换

$$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i = \vec{a}^T \vec{x}$$

给定变换系数 $\vec{a}$，可以得到相应的概率密度 $g(\vec{x}|H_0), g(\vec{x}|H_1)$
通过选择 $\vec{a}$，达到最大程度区分 $g(\vec{x}|H_0)$ 与 $g(\vec{x}|H_1)$ 的目的。

可考虑 $t(\vec{x})$ 在不同假设下的均值与方差

观测量 $\vec{x}$ 的均值与方差

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k) \mathrm{d}\vec{x}$$
$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(\vec{x}|H_k) \mathrm{d}\vec{x}$$

其中 $k = 0, 1 \qquad i, j = 1, \cdots, n$

类似地，可得 $t(\vec{x})$ 的均值与方差

$$\tau_k = \int t(\vec{x}) f(\vec{x}|H_k) \mathrm{d}\vec{x} = \vec{a}^T \vec{\mu}_k$$
$$\Sigma_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x}|H_k) \mathrm{d}\vec{x} = \vec{a}^T V_k \vec{a}$$

要求 $|\tau_0 - \tau_1|$ 大，而 Σ_0^2 和 Σ_1^2 小，使得概率密度分布 $f(\vec{x}|H_0)$ 与 $g(\vec{x}|H_1)$ 都集中在各自的均值附近且均值相差较大。

虑这两点的分离度可以由费舍尔甄别函数 $J(\vec{a})$ 给出

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2} = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}}$$

其中

$$B = (\vec{\mu}_0 - \vec{\mu}_1)(\vec{\mu}_0 - \vec{\mu}_1)^T$$
$$W = V_0 + V_1$$

令 $\dfrac{\partial J}{\partial a_i} = 0$，则使分离度最大的系数 $\vec{a}$ 有

$$\vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$

$t(\vec{x}) = \sum_{i=1}^n a_i x_i = \vec{a}^T \vec{x}$ 即为费舍尔线性甄别函数，为了确定系数 a_i，需要知道矩阵 W 和期望值 $\mu_{0,1}$

可以将 t 的定义推广为

$$t(\vec{x}) = a_0 + \sum_{i=1}^{n} a_i x_i$$

以利用偏移量 a_0 和标度（$\vec{a}$ 的常系数）的任意性，将期望值 τ_0 和 τ_1 固定为任意所需的值，此时最大化 $J(\vec{a})$ 等价于最小化方差之和

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

其中 E_k 表示 t 在 H_k 假设下的期望值，这类似于参数估计中的最小二乘原则。

特殊情形下，概率密度函数 $f(\vec{x}|H_0)$ 与 $g(\vec{x}|H_1)$ 都是多维高斯分布，具有相同的协方差矩阵 $V_0 = V_1 = V$，即

$$f(\vec{x}|H_k) = \frac{1}{(2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T V^{-1} (\vec{x}-\vec{\mu}_k)} \qquad k = 0, 1$$

此时，取包含偏移量的费舍尔甄别函数为

$$t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x}$$

则似然比

$$r = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} = \mathrm{e}^{(\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x} - \frac{1}{2}\vec{\mu}_0^T V^{-1} \vec{\mu}_0 + \frac{1}{2}\vec{\mu}_1^T V^{-1} \vec{\mu}_1} \propto \mathrm{e}^t$$

这表明检验统计量 t 为似然比 r 的单调函数，因此这种情况下费舍尔甄别量与似然比等价。

同时，如果多维变量 $\vec{x}$ 在不同假设下协方差相同，则验后概率有简单的表达式，如对于假设 H_0 有

$$P(H_0|\vec{x}) = \frac{\pi_0 f(\vec{x}|H_0)}{\pi_0 f(\vec{x}|H_0) + \pi_1 f(\vec{x}|H_1)} = \frac{1}{1 + \frac{\pi_1}{\pi_0 r}}$$

其中 π_k 为假设 H_k 的验前概率

代入 r 的表达式，并取

$$a_0 = \frac{1}{2}\vec{\mu}_0^T V^{-1}\vec{\mu}_0 + \frac{1}{2}\vec{\mu}_1^T V^{-1}\vec{\mu}_1 + \ln\frac{\pi_0}{\pi_1}$$

有

$$P(H_0|\vec{x}) = \frac{1}{1+\mathrm{e}^{-t}} \equiv s(t)$$

为逻辑 S 型函数的一个特例，其取值范围为 $(0,1)$

非线性检验统计量、神经网络

输入变量的选择

4.5 拟合优度检验，p 值定义与应用

拟合优度检验

拟合优度检验的结果可以由所谓的 P 值给出，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或置信水平。

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；而在假设检验中，显著性水平 α 是一个常数。

p 值

可用 p 值表示假设检验的拟合优度，定义为观测到数据 $\vec{x}$ 与假设 H 的符合程度不好于实际数据 $\vec{x}_{\mathrm{obs}}$ 与 H 的符合程度的概率，这不是 H 为真的概率。

经典统计不讨论 $P(H)$，除非 H 表示可重复观测；贝叶斯统计把 H 当成随机变量，并利用贝叶斯定理得到

$$p(H|t) = \frac{P(t|H)}{\int P(t|H)\pi(H)\mathrm{d}H}$$

其中 $\pi(H)$ 为 H 的先验概率

p 值是数据的函数，其本身也是有一定分布的随机变量。
从检验统计量 $t(\vec{x})$ 得到假设 H 的 p 值

$$p_H = \int_t^\infty f(t'|H)\mathrm{d}t'$$

在 H 的假设下，p 值的概率密度函数为

$$g(p_H|H) = \frac{f(t|H)}{|\frac{\partial p_H}{\partial t}|} = \frac{f(t|H)}{f(t|H)} = 1 \qquad 0 \le p_H \le 1$$

对于连续数据，在 H 的假设下，$p_H \sim U(0,1)$；在很多备择假设下，聚集于零附近。

因此 H_0 假设的 p值小于 α 的概率为

$$P(p_0 \le \alpha|H_0) = \alpha$$

形式上，p 值仅与 H_0 有关，但是得到的检验还与相对于给定的备择假设 H_1 的功效有关。

4.6 观测信号的显著性

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 $\nu = \nu_s + \nu_b$。则观测到 n 个事例的概率为

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

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

$$P(n \ge n_{\mathrm{obs}}) = \sum_{n=n_{\mathrm{obs}}}^{\infty} f(n;\nu_s=0,\nu_b)$$

$$= 1 - \sum_{n=0}^{n_{\text{obs}}-1} f(n;\nu_s=0,\nu_b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{\nu_b^n}{n!}\mathrm{e}^{-\nu_b}$$

4.7 皮尔逊卡方检验

将拟合优度检验应用于变量 x 的分布，可将观测到的 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。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。

皮尔逊卡方统计量是最常用的拟合优度检验

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i-\mu_i)^2}{\sigma_i^2} \qquad \sigma_i^2 = V[n_i]$$

若 $n_i \sim \pi(\nu_i)$，则 $V[n_i] = v_i$，皮尔逊 χ^2 统计量变为

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

如 $n_i \sim N(\nu_i, \sigma_i^2)$，则皮尔逊 χ^2 统计量（记为 z）服从自由度为 N 的卡方分布 $z \sim \chi^2(N)$

$$f_{\chi^2}(z;N) = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})} z^{\frac{N}{2}-1}\mathrm{e}^{-\frac{z}{2}}$$

可认为卡方检验与分布无关，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。

从数据得到的 χ^2 值可以给出对应的 p 值

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z;N)\mathrm{d}z$$

第五章 参数估计的一般概念

5.1 样本、估计量、偏倚

参数估计

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。可以定义新的样本空间为 n 维矢量 $\vec{x} = (x_1, \cdots, x_n)$ 所有可能取值的集合，即把包含 n 次测量看成一次随机测量。

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 $f_{\text{sample}}(x_1, \cdots, x_n)$ 可以表示为

$$f_{\text{sample}}(x_1, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

考虑对随机变量 x 进行了 n 次测量，而 x 服从的概率密度函数 $f(x)$ 未知，需要根据观测值 $x_1, \cdots, x_n$ 推断概率密度函数 $f(x)$ 的性质。即要构造 x_i 的函数以估计概率密度函数 $f(x)$ 的各种性质。通常，会假设概率密度函数为 $f(x;\theta)$，它依赖于未知参数 θ 或多维参数 $\vec{\theta} = (\theta_1, \cdots, \theta_n)$，希望构造观测值 x_i 的函数，以估计未知参数的值。

观测值 $x_1, \cdots, x_n$ 的函数（不包含未知参数）称为统计量。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。
某个量 θ 的估计量用 $\hat{\theta}$ 表示

估计量是数据样本的函数；对于给定数据样本，估计量的结果称为估计值。

估计量的性质

重复整个测量，每次得到的估计值将服从某个分布 $g(\hat{\theta};\theta)$，其期望

$$E[\hat{\theta}(\vec{x})] = \int \hat{\theta} g(\hat{\theta},\theta)\mathrm{d}\hat{\theta} = \int\cdots\int \hat{\theta}(\vec{x}) f(x_1;\theta)\cdots f(x_n;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

此过程即为参数拟合。

希望估计量满足：
偏倚 $b = E[\hat{\theta}] - \theta$ 为零或很小（系统不确定度小），即多次重复测量的均值应当趋于真值；
方差 $V[\hat{\theta}]$ 小（统计不确定度小），而偏倚小和方差小通常是相互矛盾的要求。

估计量好坏的三个标准：

一致性：
如果 $\hat{\theta}$ 在大 n 极限下收敛于 θ，则称该估计量为相合估计量（一致估计量）

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

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。

无偏性：
定义估计量 $\hat{\theta}$ 的偏倚为

$$b = E[\hat{\theta}] - \theta$$

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。
很多实际情况中，偏倚相比于统计误差（即标准差）非常小。

有效性：
定义均方差

$$\mathrm{MSE} = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E[\hat{\theta}])^2] + E^2[\hat{\theta} - \theta] = V[\hat{\theta}] + b^2$$

对于任意估计量 $\hat{\theta}'$，都有 $\lim_{n\to\infty} \frac{V[\hat{\theta}]}{V[\hat{\theta}']} \leq 1$

通常，如果估计量的偏倚为零并且方差最小，则认为这个估计量是最优的。

5.2 均值、方差和协方差的估计量

样本均值

假设有随机变量 x 的一个样本，样本容量为 n：$x_1, \cdots, x_n$。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。

希望构造 x_i 的函数作为 x 的期望值 μ 的估计量，考虑求 x_i 的代数平均值，定义样本元素的代数平均值为样本均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

x 的期望值记为 μ 或 $E[x]$，而样本均值 $\bar{x}$ 为一个估计量。

样本均值满足大数弱定律，若 x 的方差存在，则 x 为总体均值 μ 的相合估计量，即

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu| \geq \varepsilon) = 0$$

样本均值的期望值

$$E[\bar{x}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

其中

$$E[x_i] = \int \cdots \int x_i f(x_1) \cdots f(x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n = \mu$$

由此可见，样本均值 $\bar{x}$ 是总体均值 μ 的无偏估计量。

样本均值的方差

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

其中

$$\begin{aligned} E[x_i^2] &= \mu^2 + \sigma^2 \\ i \neq j \quad E[x_i x_j] &= E[x_i]E[x_j] = \mu^2 \end{aligned}$$

样本方差

样本方差 s^2 定义为

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

其中因子 $\frac{1}{n-1}$ 保证 s^2 无偏，即 $E[s^2]=\sigma^2$

若均值 $\mu=E[x]$ 先验已知，则

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

s^2 的方差

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

其中 k 阶中心矩

$$\mu_k=E[(x-E[x])^k]=\int_{-\infty}^{\infty}(x-\mu)^n f(x)\mathrm{d}x$$

可将其估计为

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

协方差与相关系数的估计量

同理可证明两个随机变量 x 和 y 的协方差 $V_{xy}=\mathrm{cov}[x,y]$ 的无偏估计量

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

相关系数 $\rho=\dfrac{V_{xy}}{\sigma_x\sigma_y}$ 的估计量

$$\hat{\rho}=r_{xy}=\frac{\hat{V}_{xy}}{s_x s_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2)^{\frac{1}{2}}}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sqrt{(\overline{x^2}-\bar{x}^2)(\overline{y^2}-\bar{y}^2)}}$$

r_{xy} 有偏移，但当 $n\to\infty$ 时，偏移趋近于 0

一般而言，概率密度 $g(r;\rho,n)$ 形式复杂；对于高斯变量 x,y

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

第六章 极大似然法

6.1 极大似然估计量

似然函数，最大似然估计量

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 $\theta=(\theta_1,\cdots,\theta_m)$）的取值未知，即 $f(x;\theta)$ 表示概率密度函数的一个复合假设。

极大似然法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 $x_1,\cdots,x_n$。若给定假设 $f(x;\theta)$ 以及参数值 θ，第一次测量结果处于区间 $[x_1,x_1+\mathrm{d}x_1]$ 的概率为 $f(x_1;\theta)\mathrm{d}x_1$，且假定所有测量都是独立的，则总概率

$$P(\forall i,x_i\in[x_i+x_i+\mathrm{d}x_i])=\prod_{i=1}^{n}f(x_i;\theta)\mathrm{d}x_i$$

若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。又由于 $\mathrm{d}x_i$ 与参数无关，可定义似然函数

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

其中，$L(\theta)$ 实为 x_i 的联合概率密度函数，且 x_i 视为实验结束后的测量值。

可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 $\theta_1,\cdots,\theta_m$ 可导，并最大值不在参数区间的边界，即可由下式给出估计量

$$\frac{\partial L}{\partial \theta_i}=0 \quad i=1,\cdots,m$$

在经典统计中，$L(\theta)$ 并不是 θ 的概率密度；在贝叶斯统计中，把 $L(\theta)=L(\vec{x}|\theta)$ 看作给定 θ 的情况下，$\vec{x}$ 的概率密度，然后利用贝叶斯定理得到验后概率密度 $p(\theta|\vec{x})$

又由于对数函数是单调函数，可取

$$\frac{\partial \ln L}{\partial \theta_i}=0$$

可证明最大似然估计值与参数选取无关，具有唯一性。

6.2 指数分布与高斯分布参数的最大似然估计

指数分布

考虑指数分布

$$f(t;\tau)=\frac{1}{\tau}\mathrm{e}^{-\frac{t}{\tau}}$$

n次测量值为 $t_1,\cdots,t_n$。则对数似然函数为

$$\ln L(\tau)=\sum_{i=1}^{n}\ln f(t_i;\tau)=\sum_{i=1}^{n}(-\ln\tau-\frac{t_i}{\tau})$$

令 $\dfrac{\partial \ln L}{\partial \tau}=0$ 可得最大似然估计值

$$\hat{\tau}=\frac{1}{n}\sum_{i=1}^{n}t_i$$

此处，极大似然估计量 $\hat{\tau}$ 即为测量值的样本均值。$\hat{\tau}$ 的期望值

$$\begin{aligned}
E\left[\hat{\tau}\left(t_1,\ldots,t_n\right)\right]&=\int\cdots\int\hat{\tau}(\vec{t})f_{\text{joint}}\left(\vec{t};\tau\right)\mathrm{d}t_1\ldots\mathrm{d}t_n\\
&=\int\cdots\int\left(\frac{1}{n}\sum_{i=1}^{n}t_i\right)\frac{1}{\tau}e^{-\frac{t_1}{\tau}}\ldots\frac{1}{\tau}e^{-\frac{t_n}{\tau}}\mathrm{d}t_1\ldots\mathrm{d}t_n\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\int t_i\frac{1}{\tau}e^{-\frac{t_i}{\tau}}\mathrm{d}t_i\prod_{j\neq i}\int\frac{1}{\tau}e^{-\frac{t_j}{\tau}}\mathrm{d}t_j\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\tau=\tau
\end{aligned}$$

τ 为 t 的无偏估计量（样本均值是任意概率密度函数期望值的无偏估计量）

若考虑衰变常数 $\lambda=\frac{1}{\tau}$ 的最大似然估计，由最大似然估计量的唯一性，有

$$\hat{\lambda}=\frac{1}{\hat{\tau}}=\frac{n}{\sum_{i=1}^{n}t_i}$$

利用特征函数方法，可证明

$$\begin{aligned}
E[\hat{\lambda}]&=\lambda\frac{n}{n-1}\\
b=E[\hat{\lambda}]-\lambda&=\frac{\lambda}{n-1}
\end{aligned}$$

即当 $n\to\infty$ 时，$\hat{\lambda}=\frac{1}{\hat{\tau}}$ 才为 $\frac{1}{\tau}$ 的无偏估计量，即为渐进无偏估计量。

高斯分布

考虑高斯变量，参数 μ 和 σ 未知，其对数似然函数为

$$\ln L(\mu,\sigma^2)=\sum_{i=1}^{n}\ln f(x_i;\mu,\sigma^2)=\sum_{i=1}^{n}(-\ln\sqrt{2\pi}-\frac{1}{2}\ln\sigma^2-\frac{(x_i-\mu)^2}{2\sigma^2})$$

对参数求导可得

$$\frac{\partial \ln L\left(\mu,\sigma^2\right)}{\partial\mu}=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}=0\qquad\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\frac{\partial \ln L\left(\mu, \sigma^2\right)}{\partial \sigma^2}=-\frac{n}{2 \sigma^2}+\sum_{i=1}^n \frac{\left(x_i-\mu\right)^2}{2 \sigma^4}=0 \qquad \hat{\sigma^2}=\frac{1}{n} \sum_{i=1}^n\left(x_i-\hat{\mu}\right)^2$$

$\hat{\mu}$ 是 μ 的无偏估计量，有 $E[\hat{\mu}]=\mu$

$\hat{\sigma^2}$ 的期望值有

$$E[\hat{\sigma^2}]=\frac{n-1}{n} \sigma^2$$

σ^2 的最大似然估计量 $\hat{\sigma^2}$ 是有偏估计量，但是在大样本极限下偏倚变为零。

而样本方差 s^2 对任何概率密度分布都是方差的无偏估计量，则

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

为高斯分布参数 σ^2 的无偏估计量。

6.3 极大似然估计量的方差：解析方法

某些情形下，我们可以解析计算极大似然估计量的方差。

如对于均值为 τ 的指数分布，τ 的极大似然估计量为 $\hat{\tau}=\frac{1}{n} \sum_{i=1}^n t_i$，可得

$$\begin{aligned} V[\hat{\tau}] & =E\left[\hat{\tau}^2\right]-(E[\hat{\tau}])^2 \\ & =\int \cdots \int\left(\frac{1}{n} \sum_{i=1}^n t_i\right)^2 \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \mathrm{d} t_1 \ldots \mathrm{d} t_n \\ & -\left(\int \cdots \int\left(\frac{1}{n} \sum_{i=1}^n t_i\right) \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \mathrm{d} t_1 \ldots \mathrm{d} t_n\right)^2 \\ & =\frac{\tau^2}{n} \end{aligned}$$

样本均值的方差等于 $\frac{1}{n}$ 乘以 t（单次测量结果）的概率密度函数的方差

6.4 极大似然估计量的方差：蒙特卡罗方法

6.5 极大似然估计量的方差：RCF边界方法

可采用RCF不等式（信息不等式）给出估计量方差的最小边界值，它适用于任何估计量，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，有

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

左侧即为最小方差界（MVB），b 为偏倚量

若 $b=0$ 且等式成立，则 $\hat{\theta}$ 为有效估计量。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。

以指数分布为例，有

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

同时有 $b=0, \frac{\partial b}{\partial \tau}=0$，则 $\hat{\tau}$ 的最小方差边界

$$V[\hat{\tau}]=\frac{1}{E[-\frac{n}{\tau^2}(1-\frac{2 \hat{\tau}}{\tau})]}=\frac{1}{-\frac{n}{\tau^2}(1-\frac{2 E[\hat{\tau}]}{\tau})}=\frac{\tau^2}{n}$$

因此 ML 估计量 τ 对任何样本容量 n 都是有效估计量。

对于 m 个参数 $\vec{\theta}=\left(\theta_1, \cdots, \theta_m\right)$，最小方差界由费舍尔信息矩阵给出

$$I_{i j}=E\left[-\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right]=-n \int f(x ; \vec{\theta}) \frac{\partial^2 \ln f(x ; \vec{\theta})}{\partial \theta_i \partial \theta_j} \mathrm{d} x$$

其中 $f(x ; \vec{\theta})$ 为随机变量 x 的概率密度函数，且观测次数为 n。

信息不等式表明 $V-I^{-1}$ 为半正定矩阵，其中 $V_{i j}=\operatorname{cov}[\hat{\theta}_i, \hat{\theta}_j]$

在数据样本足够大的情况下，可用观测数据和极大似然估计值 $\hat{\vec{\theta}}$ 计算二阶导数，以此估计 V^{-1}

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

6.6 极大似然估计量的方差：图解法，双参数的极大似然估计

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 $\hat{\theta}$作泰勒展开并忽略高阶项有

$$\begin{aligned}\ln L(\theta)&=\ln L(\hat{\theta})+\left[\frac{\partial\ln L}{\partial\theta}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta})+\frac{1}{2!}\left[\frac{\partial^2\ln L}{\partial\theta^2}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2+\dots\\&=\ln L_{\max}-\frac{(\theta-\hat{\theta})^2}{2\hat{\sigma_\theta^2}}\end{aligned}$$

即

$$\ln L(\hat{\theta}\pm\hat{\sigma_\theta})=\ln L_{\max}-\frac{1}{2}$$

考虑双参数的极大似然估计，对于大样本容量 n，$\ln L$ 在最大值处具有二次型的形式

$$\ln L(\alpha,\beta)\approx\ln L_{\max}-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)^2+\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)^2-2\rho\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\bar{\alpha}}}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)\right]$$

等高线

$$\ln L(\alpha,\beta)=\ln L_{\max}-\frac{1}{2}$$

是个椭圆

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

在蒙特卡罗样本中，每次的拟合结果对应于 $\beta-\alpha$ 平面上的一个点。

椭圆的倾角 ϕ 与相关性有关

$$\tan\phi=\frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2-\sigma_{\hat{\beta}}^2}$$

6.7 扩展的极大似然估计

先前只考虑了样本容量 n 固定的情形，某些情况下 n 被看作均值为 ν 的泊松变量（如期待事例数 $\nu(\vec{\theta})=\sigma(\vec{\theta})\int L\mathrm{d}t$，其中总截面 $\sigma(\vec{\theta})$ 预期为理论参数的函数。），则实验结果可定义为 $n,x_i,\cdots,x_n$，则扩展的似然函数

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

假设 ν 给定为 $\vec{\theta}$ 的函数 $\nu=\nu(\vec{\theta})$，则扩展的对数似然函数

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

此时扩展的最大似然法利用了更多信息，以使 $\hat{\vec{\theta}}$ 的不确定度更小

若 ν 和 $\vec{\theta}$ 相互独立，则

$$\begin{aligned}L(\nu,\vec{\theta})&=\frac{\nu^n}{n!}\mathrm{e}^{-\nu}\prod_{i=1}^n f(x_i;\vec{\theta})\\\frac{\partial L}{\partial\nu}&=(\frac{n}{\nu}-1)\frac{\nu^n}{n!}\mathrm{e}^{-\nu}\prod_{i=1}^n f(x_i;\vec{\theta})=0\qquad\hat{\nu}=n\end{aligned}$$

而 $\frac{\partial L}{\partial\vec{\theta}}=0$ 依旧可得最大似然估计

若联合概率密度函数可以表示为

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

根据概率的定义可知并非所有的 θ_i 独立

$$\sum_{i=1}^{m}\theta_i=1\quad\therefore\theta_m f_m(x)=\left(1-\sum_{i=1}^{m-1}\theta_i\right)f_m(x)$$

在扩展的最大似然法中

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

定义 $\mu_i=\nu\theta_i$，有

$$\ln L(\vec{\mu})=-\sum_{j=1}^{m}\mu_j+\sum_{i=1}^{n}\ln\left(\sum_{j=1}^{m}\mu_j f_j(x_i)\right)$$

μ_j 为类型 j 的事例数期待值，n 为观测事例总数。

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。假设有两类事例：信号 s 与本底 b

$$f\left(x;\mu_{\mathrm{s}},\mu_{\mathrm{b}}\right)=\frac{\mu_{\mathrm{s}}}{\mu_{\mathrm{s}}+\mu_{\mathrm{b}}}f_{\mathrm{s}}(x)+\frac{\mu_{\mathrm{b}}}{\mu_{\mathrm{s}}+\mu_{\mathrm{b}}}f_{\mathrm{b}}(x)$$

假设 $f_{\mathrm{s}}(x)$ 和 $f_{\mathrm{b}}(x)$ 已知，需要估计 μ_{s} 和 μ_{b}

$$\ln L\left(x;\mu_{\mathrm{s}},\mu_{\mathrm{b}}\right)=-\left(\mu_{\mathrm{s}}+\mu_{\mathrm{b}}\right)+\sum_{i=1}^{n}\ln\left[\left(\mu_{\mathrm{s}}+\mu_{\mathrm{b}}\right)f\left(x_i;\mu_{\mathrm{s}},\mu_{\mathrm{b}}\right)\right]$$

本底高低对拟合结果不确定度影响很大。这样的实验拟合可能拟合出负的信号估计值 $\hat{\mu}_{\mathrm{s}}<0$，为非物理结果。

6.7 分区间数据的极大似然估计

通常数据 $\vec{x}$ 在划分为 N 个区间的直方图中的频数为

$$\vec{n}=(n_1,\ldots,n_N),\quad n_{\mathrm{tot}}=\sum_{i=1}^{N}n_i$$

即对直方图拟合。

在某种假设下，频数期待值为

$$v_i(\vec{\theta})=v_{\mathrm{tot}}\int_{x_i^{\min}}^{x_i^{\max}}f(x;\vec{\theta})\mathrm{d}x\quad\therefore\vec{v}=(v_1,\ldots,v_N)\quad v_{\mathrm{tot}}=\sum_{i=1}^{N}v_i$$

如果用多项分布描述样本（n_{tot} 为常数）

$$f(\vec{n};\vec{v})=\frac{n_{\mathrm{tot}}\,!}{n_1!\cdots n_N!}\left(\frac{v_1}{n_{\mathrm{tot}}}\right)^{n_1}\cdots\left(\frac{v_N}{n_{\mathrm{tot}}}\right)^{n_N}$$

$$\therefore\ln L(\vec{\theta})=\sum_{i=1}^{N}n_i\ln v_i(\vec{\theta})$$

6.8 极大似然法的拟合优度检验

6.9 用极大似然法合并实验测量

考虑不等精度观测结果的合并
不等精度观测结果的合并
对某固定量 μ 作 n 次独立的不等精度测量，结果为 $x_i\pm\sigma_i$，方差已知，且 $x_i\sim N\left(\mu,\sigma_i^2\right)$。则该样本的似然函数为

$$L(\mu)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right]$$

取对数并解似然方程

$$\frac{\partial\ln L(\mu)}{\partial\mu}=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma_i}=0$$

$$\therefore \hat{\mu}=\frac{\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}=\frac{1}{\omega}\sum_{i=1}^{n}\omega_i x_i$$

其中权重因子

$$\omega_i=\frac{1}{\sigma_i^2}\quad \omega=\sum_{i=1}^{n}\omega_i$$

因此有

$$\hat{\sigma}_{\hat{\mu}}^2=\left(-\frac{1}{\frac{\partial^2 \ln L}{\partial \mu^2}}\right)\Bigg|_{\mu=\hat{\mu}}=\frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}=\frac{1}{\omega}$$

也可将权重因子记为

$$\omega_i=\frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\quad \sum_{i=1}^{n}\omega_i=1$$
$$\hat{\mu}=\sum_{i=1}^{n}\omega_i x_i$$

6.10 极大似然与贝叶斯估计量的关系

第七章 最小二乘法

7.1与极大似然的联系

似然函数，最大似然估计量

由中心极限定理可得很多情形下，测量值 y 可看作以真值 λ 为中心值的高斯随机变量。

考虑 N 个独立的高斯随机变量 $y_i\quad (i=1,\cdots,N)$，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。设不同 y_i 的均值 $E[y_i]=\lambda_i$ 不同并且未知，其方差 $V[y_i]=\sigma_i^2$ 也不同但是大小已知。
则对于独立高斯变量 y_i，联合概率密度

$$g(\vec{y};\vec{\lambda},\vec{\sigma}^2)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma_i^2}}\mathrm{e}^{-\frac{(y_i-\lambda_i)^2}{2\sigma_i^2}}$$

进一步假设真值 λ 为 x 的函数，即 $\lambda=\lambda(x;\vec{\theta})$，该函数依赖于未知参数 $\vec{\theta}=(\theta_1,...,\theta_m)$，最小二乘法需对参数 $\vec{\theta}$ 进行估计，且可给出函数 $\lambda=\lambda(x;\vec{\theta})$ 的拟合优度。

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

$$\ln L(\vec{\theta})=-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\lambda(x_i;\vec{\theta}))^2}{\sigma_i^2}$$

对数似然函数的最大化，等价于求使 $\chi^2(\vec{\theta})$ 最小化的参数 $\vec{\theta}$ 的值，其中

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

此即最小二乘法（LS）

如果测量不相互独立，y_i 是多维高斯变量，协方差矩阵为 V ，满足

$$g(\vec{y};\vec{\lambda},V)=\frac{1}{(2\pi)^{N/2}|V|^{1/2}}\exp\left[-\frac{1}{2}(\vec{y}-\vec{\lambda})^T V^{-1}(\vec{y}-\vec{\lambda})\right]$$

对数似然函数为

$$\ln L(\vec{\theta})=\frac{1}{2}\sum_{i=1}^{N}\left[y_i-\lambda\left(x_i;\theta\right)\right]\left(V^{-1}\right)_{ij}\left[y_j-\lambda\left(x_j;\theta\right)\right]$$

即应求下式的最小值

$$\chi^2(\vec{\theta})=\sum_{i,j=1}^{N}\left[y_i-\lambda\left(x_i;\theta\right)\right]\left(V^{-1}\right)_{ij}\left[y_j-\lambda\left(x_j;\theta\right)\right]$$

其最小值定义了最小二乘估计量 $\hat{\theta}$

即使 y_i 不是高斯变量，该定义依然适用。

7.2 线性最小二乘拟合

若 $\lambda(x;\vec{\theta})$ 是 $\vec{\theta}$ 的线性函数

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

其中 $a_i(x)$ 为 x 的任意线性独立函数

高斯马可夫定理可证明线性最小二乘估计量的偏倚为零，并且方差最小。

函数 $\lambda(x;\vec{\theta})$ 在 x_i 处有

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

其中 $A_{ij}=a_j(x_i)$，因此有矩阵形式

$$\chi^2(\vec{\theta})=\left(\vec{y}-\vec{\lambda}\right)^T V^{-1}\left(\vec{y}-\vec{\lambda}\right)=\left(\vec{y}-A\vec{\theta}\right)^T V^{-1}\left(\vec{y}-A\vec{\theta}\right)$$

对 θ_i 求微分，有

$$\nabla\chi^2=-2\left(A^TV^{-1}\vec{y}-A^TV^{-1}A\vec{\theta}\right)=0$$
$$\hat{\vec{\theta}}=(A^TV^{-1}A)^{-1}A^TV^{-1}\vec{y}\equiv B\vec{y}$$

解得的估计量 $\hat{\vec{\theta}}$ 是测量量 $\vec{y}$ 的线性函数

线性条件下协方差矩阵元 $U_{ij}=\mathrm{cov}[\theta_i,\theta_j]$可由误差传递得到

$$U=BVB^T=(A^TV^{-1}A)^{-1}$$

等价地，协方差矩阵的逆

$$(U^{-1})_{ij}=\frac{1}{2}\left[\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right]_{\vec{\theta}=\hat{\vec{\theta}}}=A_{ik}V_{kl}^{-1}A_{jl}$$

如果 y_i 是高斯变量，其方差与 RCF 边界一致。

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

$$\chi^2(\vec{\theta})=\chi^2(\hat{\vec{\theta}})+\frac{1}{2}\left[\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right]_{\vec{\theta}=\hat{\vec{\theta}}}(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)$$

有

$$\chi^2(\hat{\vec{\theta}}\pm\hat{\sigma}_{\hat{\vec{\theta}}})=\chi^2_{\min}+1$$

而等值线

$$\chi^2(\vec{\theta})=\chi^2(\hat{\vec{\theta}})+1$$

在参数空间定义了一个区域，可以解释为置信区域

7.3 非线性最小二乘法估计、约束情况下的最小二乘法

多项式的最小二乘拟合

作为 $\lambda(x;\vec{\theta})$ 的一个假设，可能尝试 m 次多项式（即 $m+1$ 个参数）

$$\lambda(x;\theta_0,\cdots,\theta_m)=\sum_{j=0}^{m}x^j\theta_j$$

非线性最小二乘法估计同理，不过最小二乘法没有参数的解析解，需要通过迭代法求 $\hat{\theta}$ 的近似解

最小二乘估计量的方差

多数情况下与最大似然法中方差估计类似。若数据服从高斯分布，则有

$$\chi^2(\theta) = -2\ln L(\theta)$$
$$\hat{\sigma^2_{\hat{\theta}}} \approx 2\left[\frac{\partial^2\chi^2}{\partial\theta^2}\right]^{-1}_{\theta=\hat{\theta}}$$

约束情况下的最小二乘法拟合

7.4 拟合优度最小二乘检验

假设 $y_i \quad (i=1,\cdots,N)$ 是独立的高斯变量（σ 已知），且 $\lambda(x;\vec{\theta})$ 是 $\vec{\theta}$ 的线性函数，所采用的函数形式也是正确的，则

$$\vec{\theta} = \hat{\vec{\theta}} \qquad \chi^2_{\min} = \sum_{i=1}^{N}\frac{(y-\lambda(x;\hat{\vec{\theta}}))^2}{\sigma_i^2}$$

其中 $\chi^2_{\min} \sim \chi^2(n_d)$，自由度的数目 $n_d = N-m$，m 为参数个数

$\chi^2_{\min}$ 可以用作拟合优度统计量，检验假设的函数形式 $\lambda(x;\vec{\theta})$ 的好坏。$\chi^2_{\min}$ 小（$\frac{\chi^2_{\min}}{n_d} \approx 1$），则表明假设的函数形式与数据相符；$\chi^2_{\min}$ 大（$\frac{\chi^2_{\min}}{n_d} \gg 1$），则表明假设的函数形式与数据不符；而若 $\frac{\chi^2_{\min}}{n_d} \ll 1$），则拟合好于预期（可能过拟合）

一般为给定的 $\chi^2_{\min}$ 提供一个显著水平，即 p 值

$$p = \int_{\chi^2_{\min}}^{\infty} f_{\chi^2}(z;N)\mathrm{d}z$$

7.5 最小二乘法处理分区数据

考虑直方图有 N 个区间，总频数 n 。假设的概率密度函数形式为 $f(x;\vec{\theta})$，y_i 为第 i 个区间的频数，有

$$\lambda_i(\vec{\theta}) = n\int_{x_i^{\min}}^{x_i^{\max}} f(x;\vec{\theta})\mathrm{d}x = np_i(\vec{\theta})$$

最小二乘法拟合使下式有最小值

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N}\frac{\left(y_i-\lambda_i(\vec{\theta})\right)^2}{\sigma_i^2}$$

其中 $\sigma_i^2 = V[y_i]$ 为先验未未知量

把 y_i 看作泊松变量，方差为

$$\sigma_i^2 = \lambda_i(\vec{\theta}) \quad (\text{最小二乘法LS})$$
$$\sigma_i^2 = y_i \quad (\text{改进的最小二乘法MLS})$$

改进的最小二乘法虽方便了计算，但对于有些区间频数太少时 $\chi^2_{\min}$ 不再服从最小二乘的概率密度分布函数(或无定义)。

最小二乘法拟合中，尽量避免拟合归一化常数，例如引入可调参数 ν 并与 $\vec{\theta}$ 一起拟合

$$\lambda_i(\vec{\theta},\nu) = \nu\int_{x_i^{\min}}^{x_i^{\max}} f(x;\vec{\theta})\mathrm{d}x = \nu p_i(\vec{\theta})$$

$\hat{\nu}$ 不是 n 的好估计量，可以证明

$$\hat{\nu}_{\mathrm{LS}} = n + \frac{\chi^2_{\min}}{2}$$
$$\hat{\nu}_{\mathrm{MLS}} = n - \chi^2_{\min}$$

LS 和 MLS 方法得到的估计量 $\hat{\nu}$ 都有偏倚，不过偏移大小可以知道

7.6 用最小二乘法合并实验测量

用最小二乘法并合各实验结果

已知 λ 的 N 个测量结果，需求其平均值

设 y_i 为第 i 个测量结果，$\sigma_i^2 = V[y_i]$假设已知，λ 为真值

若各测量量之间不相关，则需求下式最小值（结果与 ML 方法一样）

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

$$\therefore \hat{\lambda}=\sum_{i=1}^{N}\omega_i y_i$$

$$\omega_i=\frac{1/\sigma_i^2}{\sum_{j=1}^{N}1/\sigma_j^2}\qquad V[\hat{\lambda}]=\sum_{i=1}^{N}\omega_i^2\sigma_i^2$$

如果各测量量之间相关，$\operatorname{cov}[y_i,y_j]=V_{ij}$，则求下式最小值:

$$\chi^2(\lambda)=\sum_{i,j=1}^{N}(y_i-\lambda)\left(V^{-1}\right)_{ij}(y_j-\lambda)$$

$$\therefore \hat{\lambda}=\sum_{i=1}^{N}\omega_i y_i$$

$$\omega_i=\frac{\sum_{j=1}^{N}\left(V^{-1}\right)_{ij}}{\sum_{k,k=1}^{N}\left(V^{-1}\right)_{kl}}\qquad V[\hat{\lambda}]=\sum_{i,j=1}^{N}\omega_i V_{ij}\omega_j$$

LS 方法得到的 $\hat{\lambda}$ 是无偏的，且方差最小(高斯-马科夫定理)

两个相关实验的平均值

假设有两个相关的测量量 y_1 和 y_2 ，且

$$V=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$$

$$\therefore \hat{\lambda}=w_1y_1+(1-w_1)y_2,\quad w_1=\frac{\sigma_2^2-\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}]=\frac{\left(1-\rho^2\right)\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\equiv\sigma$$

$$\frac{1}{\sigma^2}=\frac{1}{1-\rho^2}\left[\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}-\frac{2\rho}{\sigma_1\sigma_2}\right]$$

由第二个测量导致方差倒数的增加量

$$\frac{1}{\sigma^2}-\frac{1}{\sigma_1^2}=\frac{1}{1-\rho}\left(\frac{\rho}{\sigma_1}-\frac{1}{\sigma_2}\right)^2>0$$

表明第二个测量结果对平均值总是有帮助

若 $\rho>\frac{\sigma_1}{\sigma_2}$，即$w_1<0$，则加权平均的结果将不在 y_1 和 y_2 之间

如果相关性由使用相同数据引起，不可能发生这种情况；如果相关性来自共同的随机效应，有可能发生这种情况。如果 ρ,σ_1,σ_2 不正确，结果将很不可信，需要检查

7.7 似然比与拟合优度

第八章 矩方法

有时管极大似然法和最小二乘法很难具体实现，可考虑另一种参数估计方法矩方法（MM）

设有随机变量 $x\sim f(x;\vec{\theta})$，其中包含 m 个未知参数 $\vec{\theta}=(\theta_1,\cdots,\theta_m)$，需要利用 n 个观测值 $x_1,\cdots,x_n$ 估计 $\vec{\theta}$

考虑构造 m 个线性独立的函数 $a_i(x)\quad i=1,\cdots,m$（$a_i(x)$ 本身也是随机变量），其数学期望值是真实参数的函数

$$E\left[a_i(x)\right]=\int a_i(x)f(x;\vec{\theta})\mathrm{d}x\equiv e_i(\vec{\theta})$$

样本均值为随机变量数学期望值的无偏估计量，因此可用 x 的观测值计算函数 $a_i(x)$，再用函数 $a_i(x)$ 的算术平均估计数学期望值 $e_i=E\left[a_i(x)\right]$

$$\hat{e}_i=\bar{a}_i=\frac{1}{n}\sum_{j=1}^{n}a_i\left(x_j\right)$$

参数 $\vec{\theta}$ 的矩方法估计量定义为如下方法：令期望值 $e_i(\vec{\theta})$ 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 $\hat{\theta}_1,\cdots,\hat{\theta}_m$ 的 m 个方程

$$e_1(\hat{\vec{\theta}})=\frac{1}{n}\sum_{i=1}^{n}a_1\left(x_i\right)$$

$$\cdots$$

$$e_m(\hat{\vec{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} a_m\left(x_i\right)$$

函数 $a_i(x)$ 可取为变量 x 的整数幂次方 $x^1, \cdots, x^m$，则数学期望值 $E\left[a_i(x)\right] = E\left[x^i\right]$ 为 x 的 i 阶矩。

还需要评估估计量 $\hat{\theta}_1, \cdots, \hat{\theta}_m$ 的协方差矩阵 $\operatorname{cov}\left[a_i(x), a_j(x)\right]$，有

$$\hat{\operatorname{cov}}\left[a_i(x), a_j(x)\right] = \frac{1}{n-1}\sum_{k=1}^{n}\left(a_i\left(x_k\right) - \bar{a}_i\right)\left(a_j\left(x_k\right) - \bar{a}_j\right)$$

该协方差可与函数的算数平均的协方差建立联系

$$\begin{aligned}\operatorname{cov}\left[\bar{a}_i, \bar{a}_j\right] &= \operatorname{cov}\left[\frac{1}{n}\sum_{k=1}^{n} a_i\left(x_k\right), \frac{1}{n}\sum_{l=1}^{n} a_j\left(x_l\right)\right] \\ &= \frac{1}{n^2}\sum_{k,l=1}^{n}\operatorname{cov}\left[a_i\left(x_k\right), a_j\left(x_l\right)\right] \\ &= \frac{1}{n}\operatorname{cov}\left[a_i, a_j\right]\end{aligned}$$

其中对 k 与 l 的求和只有 $k = l$ 的 n 项有贡献，并且每项贡献为 $\operatorname{cov}\left[a_i, a_j\right]$。

因此，期望值 $\hat{e}_i = \bar{a}_i$ 的估计量的协方差矩阵 $\operatorname{cov}\left[\hat{e}_i, \hat{e}_j\right]$ 可以估计为

$$\hat{\operatorname{cov}}\left[\hat{e}_i, \hat{e}_j\right] = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(a_i\left(x_k\right) - \bar{a}_i\right)\left(a_j\left(x_k\right) - \bar{a}_j\right)$$

为了得到参数估计量本身的协方差矩阵 $\operatorname{cov}\left[\hat{\theta}_i, \hat{\theta}_j\right]$，可以使用误差传递公式得到

$$\operatorname{cov}\left[\hat{\theta}_i, \hat{\theta}_j\right] = \sum_{k,l}\frac{\partial\hat{\theta}_i}{\partial\hat{e}_k}\frac{\partial\hat{\theta}_j}{\partial\hat{e}_l}\operatorname{cov}\left[\hat{e}_k, \hat{e}_l\right]$$

第九章 统计不确定度、置信区间和极限

9.1 标准差作为统计不确定度

统计误差：实验的结果是对某个参数的估计，这个估计量的方差

可使用均值与标准差描述一个分布的特征：

- 从样本数据 $x_1, x_2 \ldots, x_n$ 可以通过一些方法（例如极大似然法）构造函数 $\theta(x_1, x_2 \ldots, x_n)$ 作为参数 θ 的估计。
- 利用一些方法（例如解析法，RCF 边界，蒙特卡罗方法，图解法）估计 $\hat{\theta}$ 的标准差。

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。通过估计标准差或者协方差矩阵，我们可以获得一个分布的信息。

如果 $g(\hat{\theta})$ 不是高斯分布，这实际上不是常规的定义。通常使用经典置信区间，并且一般来说会导致不对称的误差棒。

9.2 经典置信区间

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 为未知，但只要给定 θ 的值，就知道 $\hat{\theta}$ 的概率密度函数。

定义 $\alpha, \beta, u_\alpha, v_\beta$ 为：

$$\begin{aligned}\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) &= \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}, \theta)\mathrm{d}\theta = 1 - G(u_\alpha(\theta); \theta) \\ \beta = P(\hat{\theta} \leq v_\beta(\theta)) &= \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}, \theta)\mathrm{d}\theta = G(v_\beta(\theta); \theta)\end{aligned}$$

由此可以在固定 α, β 时反解出隐函数 $u_\alpha(\theta), v_\beta(\theta)$

在 $\hat{\theta} - \theta$ 图中，$u_\alpha(\theta), v_\beta(\theta)$ 两条曲线中间的区域为**置信带**的 $(\theta, \hat{\theta})$ 满足 $P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta$

只要 $u_\alpha(\theta), v_\beta(\theta)$ 单调增，可以解出反函数 $\alpha(\hat{\theta}) := u_\alpha^{-1}(\theta), \beta(\hat{\theta}) := v_\beta^{-1}(\theta)$

所以 $v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta) \Leftrightarrow \alpha(\hat{\theta}) \leq \theta \leq \beta(\hat{\theta})$

把 $[a(\theta_{\mathrm{obs}}), b(\theta_{\mathrm{obs}})]$ 称作置信水平为 $1 - \alpha - \beta$ 时的**置信区间**。

单侧置信区间：$\alpha = 0$ 或 $\beta = 0$ 的情况，只关心上限或下限

中心置信区间：$\alpha=\beta=\gamma/2$

在报道测量结果时，置信区间 $[a,b]$经常表示成 $\hat{\theta}^{+d}_{-c}$，$c=\hat{\theta}-a$和$d=b-\hat{\theta}$ 称为**误差棒**

9.3 高斯分布估计量的置信区间

高斯分布的累计分布函数：

$$G(\hat{\theta},\theta,\sigma_{\hat{\theta}})=\int_{-\infty}^{\hat{\theta}}\frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}}\exp(\frac{-(\hat{\theta}'-\theta)^2}{2\sigma_{\hat{\theta}}^2})\mathrm{d}\hat{\theta}'$$

根据上一节的解法求 α,β

$$\alpha=1-\Phi\left(\frac{\hat{\theta}_{obs}-a}{\sigma_{\hat{\theta}}}\right)$$
$$\beta=\Phi\left(\frac{\hat{\theta}_{obs}-b}{\sigma_{\hat{\theta}}}\right)$$

反解得：

$$a=\hat{\theta}_{obs}-\sigma_{\hat{\theta}}\Phi^{-1}(1-\alpha)$$
$$b=\hat{\theta}_{obs}+\sigma_{\hat{\theta}}\Phi^{-1}(1-\beta)$$

9.4 泊松分布均值的置信区间

泊松分布

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

求α,β：

$$\alpha=P(\nu\geq\nu_{obs};a)=1-\sum_{n=0}^{n_{\mathrm{obs}}-1}\frac{a^n}{n!}e^{-a}$$
$$\beta=P(\nu\leq\nu_{obs};b)=\sum_{n=0}^{n_{\mathrm{obs}}}\frac{b^n}{n!}e^{-b}$$

反解得：

$$a=\frac{1}{2}F^{-1}_{\chi^2}(\alpha;n_d=2n_{\mathrm{obs}})$$
$$b=\frac{1}{2}F^{-1}_{\chi^2}(1-\beta;n_d=2n_{\mathrm{obs}}+2)$$

若 $n_{\mathrm{obs}}=0$ 则无法确定下限 a，此时 $\beta=e^{-b},b=-\ln\beta$。对置信水平 $1-\beta=95\%$ 得到 $b=2.996\approx3$

9.5 相关系数、参数变换的置信区间

相关系数 r 的分布函数 $g(r;\rho,n)$ 非对称（ρ 为真实相关系数）。至少需要 $n\geq500$ 才能认为 g 近似为高斯分布。

而统计量

$$z=\tanh^{-1}r=\frac{1}{2}\ln(\frac{1+r}{1-r})$$

的概率密度函数随样本n趋向于高斯分布的速度要快得多，故定义估计量：

$$\zeta=\tanh^{-1}\rho=\frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$$

可以证明：

$$E[z]=\frac{1}{2}\ln(\frac{1+\rho}{1-\rho})+\frac{\rho}{2(n-1)},V[z]=\frac{1}{n-3}$$

利用z近似为高斯分布可以套用9.3节方法计算置信区间。

9.6 用似然函数或卡方求置信区间

大样本极限下，可以证明，似然函数本身也是高斯形式：

$$L(\theta) = L_{max} \exp(\frac{-(\hat{\theta'} - \theta)^2}{2\sigma_{\hat{\theta}}^2})$$

故可知

$$\ln(L(\hat{\theta} \pm N\sigma_\theta)) = \ln L_{max} - \frac{N^2}{2}$$

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：

$$\log L(\hat{\theta}_{-c}^{+d}) = \log L_{max} - \frac{N^2}{2}$$

由于高斯分布的最小二拟合有$\log L = -\chi^2/2$，写出等价的式子：

$$\chi^2(\hat{\theta}_{-c}^{+d}) = \chi_{max}^2 + N^2$$

用上面两个式子即可以求出置信区间

9.6 从似然函数估计近似的置信区间

似然比与拟合优度

假设用依赖于 N 个参数 $\vec{\mu} = (\mu_1, \ldots, \mu_N)$ 的似然比 $L(\vec{\mu})$ 描述数据。定义统计量:

$$t_{\vec{\mu}} = -2 \ln \frac{L(\vec{\mu})}{L(\hat{\vec{\mu}})} \quad \hat{\vec{\mu}} : \vec{\mu} \text{ 的ML估计量}$$

$t_{\vec{\mu}}$ 的值可以反映假设的 $\vec{\mu}$ 与数据的符合程度:
符合较好意味着 $\hat{\vec{\mu}} \approx \vec{\mu} \to t_{\vec{\mu}}$ 的值较小;
$t_{\vec{\mu}}$ 值较大意味着数据与 $\vec{\mu}$ 不太一致。

用 p 值定量描述 “拟合优度” :

$$p_{\vec{\mu}} = \int_{t_{\vec{\mu},\,\mathrm{obs}}}^{\infty} f\left(t_{\vec{\mu}} \mid \vec{\mu}\right) \mathrm{d}t_{\vec{\mu}} \quad \text{需要 } f\left(t_{\vec{\mu}} \mid \vec{\mu}\right) \text{ 已知}$$

假设参数 $\vec{\mu} = (\mu_1, \ldots, \mu_N)$ 可以由另一组参数 $\vec{\theta} = (\theta_1, \ldots, \theta_M)$ 确定，$M < N$。
例如，在 LS 拟合中，$\mu_i = \mu\left(x_i; \vec{\theta}\right)$，$x$ 为控制变量。
定义统计量:

$$q_{\vec{\mu}} = -2 \ln \frac{L(\vec{\mu}(\hat{\vec{\theta}}))}{L(\hat{\vec{\mu}})}$$

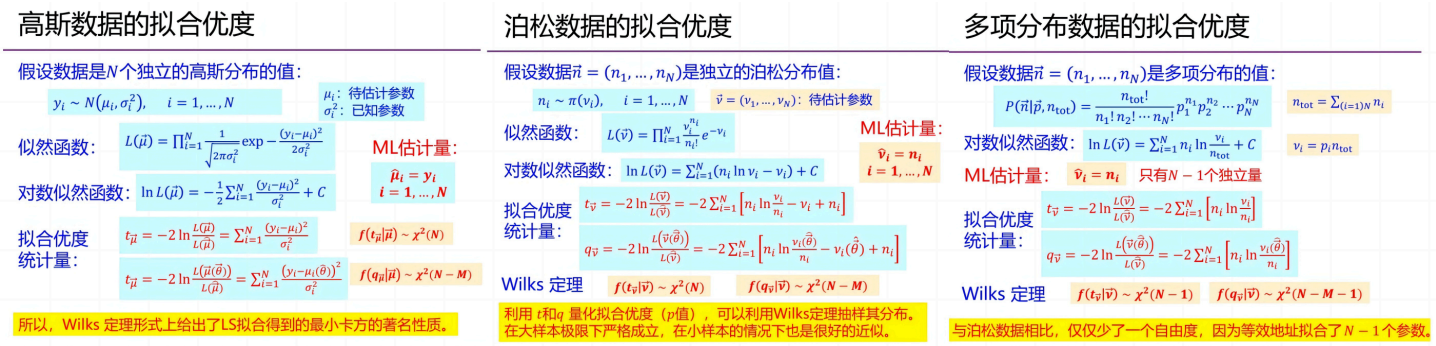
其中 $\hat{\vec{\theta}}$ 拟合 M 个参数；$\hat{\vec{\mu}}$ 拟合 N 个参数
用 $q_{\vec{\mu}}$ 检验假设的函数形式 $\mu(x; \vec{\theta})$。
要得到 p 值，需要知道 $f\left(t_{\vec{\mu}} \mid \vec{\mu}\right)$。

有Wilks定理：
如果假设的参数 $\vec{\mu} = (\mu_1, \ldots, \mu_N)$ 为真，那么在大数据样本极限下（同时满足其他一些条件），$t_{\vec{\mu}}$ 和 $q_{\vec{\mu}}$ 服从卡方分布。
当分子中的 $\vec{\mu} = (\mu_1, \ldots, \mu_N)$ 固定时：

$$t_{\vec{\mu}} = -2 \ln \frac{L(\vec{\mu})}{L(\hat{\vec{\mu}})} \quad f\left(t_{\vec{\mu}} \mid \vec{\mu}\right) \sim \chi^2(N)$$

如果分子中有 M 个可调参数:

$$q_{\vec{\mu}} = -2 \ln \frac{L(\vec{\mu}(\vec{\theta}))}{L(\hat{\mu})} \quad f\left(q_{\vec{\mu}} \mid \vec{\mu}\right) \sim \chi^2(N - M)$$



从Wilks定理估计置信区间

有Wilks定理（满足大样本极限和其他一些条件）

$$f\left(t_{\vec{\theta}} \mid \vec{\theta}\right) \sim \chi^2(n)$$

为分量个数 n 的卡方分布，其中自由度数目等于参数 $\vec{\theta} = (\theta_1, \dots, \theta_n)$

假设Wilks定理成立，p 值为

$$p_{\vec{\theta}} = 1 - F_{\chi^2(n)}\left(t_{\vec{\theta}}\right)$$

要得到置信区间的边界，令 $p_{\vec{\theta}} = \alpha$ 并求解 $t_{\vec{\theta}}$ 有

$$t_{\vec{\theta}} = F_{\chi^2(n)}^{-1}(1-\alpha)$$

注意，$t_{\vec{\theta}}$ 还可以表示为:

$$t_{\vec{\theta}} = -2 \ln \lambda(\vec{\theta}) = -2 \ln \frac{L(\vec{\theta})}{L(\hat{\theta})} \Longrightarrow \ln \vec{\theta} \text{ 可用 } F_{\chi^2(n)}^{-1} \text{ 表示}$$

在 $\vec{\theta}$ 空间，置信区域的边界为

$$\ln L(\vec{\theta}) = \ln L(\hat{\vec{\theta}}) - \frac{1}{2} F_{\chi^2(n)}^{-1}(1-\alpha)$$

例如，对于 $1-\alpha = 68.3\%$ 和 $n = 1$ 个参数

$$F_{\chi^2(n)}^{-1}(0.683) = 1$$

所以， 68.3% 置信水平的置信区间由下式确定:

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

这与求估计量标准差的方法一样，即

$$\left[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}\right] \text{ 是 CL} = 68.3\% \text{ 的置信区间。}$$

样本和抽样分布

统计量是仅依赖于样本的随机变量，所以它必有一个概率分布。

统计量 $T_n = g(X_1, X_2, \cdots, X_n)$ 的分布称为抽样分布