数理统计概要

管狐

2024.05.13

前言

本概要整理自2024年春季的数理统计课程，主要参考了张伟平老师的课件和韦来生老师的《数理统计（第二版）》教材，同时结合了张老师上课所举的实例、网上相关资料，以及舍友送的美时巧克力。

由于身体原因，本人在本学期请假时间较多，自学过程中遇到了诸多困难。为了系统整理这门课程的内容，帮助同学们复习，并为未来的学习者提供参考，我编写了这部分知识的大纲，希望对大家的学习有所帮助。

由于个人水平有限，对部分内容理解不深，因此保留了课件中的英文原文，希望大家谅解。不过，许多英文原文后添加了中文注解，以便大家在遇到理解障碍时参考。部分章节标题带有*号，表示这部分知识可能属于老师补充讲解的。对于大多数定理，本大纲未给出详细证明，因为本人不会，好在数理统计课程对证明的要求不高。

课程尚未全部结束，还有贝叶斯部分未讲解，但因本人时间有限，本学期可能无法更新。如果发现本概要中有任何疏漏或有任何建议，请发送邮件至1211596100@qq.com，如果想要某一部分的latex代码，也可以加我的QQ。感谢所有已经或将要为本大纲作出贡献的同学。

管狐

目录

Lec 1 Introduction

主要是与上学期所学的概率论进行的衔接，并进行一些基础概念的介绍。

1.1 Statistics

此处作出概念的统计学定义列举

定义 1.1.1. 样本函数

一些关于样本信息的函数，称之为样本函数。

例 1.1.2. 样本均值

设$X_1, X_2, \cdots, X_n$是来自总体的样本，样本均值定义为

$$T = T(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

样本均值函数显然是一个样本函数。

定义 1.1.3. 统计量

设$X_1, X_2, \cdots, X_n$是来自总体的样本，$T = T(X_1, X_2, \ldots, X_n)$是样本函数，如果$T$不依赖于总体的未知参数，即$T$是样本的函数，而不是总体的函数，那么$T$称为统计量。

注记. 或者也可以写地更分析一点，设$\{x_i\}_{i=1}^n$是取自总体的样本，完整的样本均值函数称之为

$$T : \mathbb{R}^n \to \mathbb{R}, (x_1, x_2, \ldots, x_n) \mapsto \frac{1}{n}\sum_{i=1}^{n} x_i$$

但这里是统计.jpg，所以请适应上面的写法。

例 1.1.4. 样本方差

设$X_1, X_2, \cdots, X_n$是来自总体的样本，样本方差定义为

$$S^2 = S^2(X_1, X_2, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

显然样本均值函数与样本方差函数都是统计量。

注记. 为了表述方便，统计里也将样本均值函数在之后省略称之为样本均值（或者均值）。也就是说很多熟知的概念，例如样本均值，样本方差，样本标准差等等，都是样本函数，都是统计量。

例 1.1.5. 样本中位数/分位数/次序统计量

设$\{X_i\}$为$i.i.d$样本，将其值按照大小排列为$\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$（$i.e.$ $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$）则$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$称为次序统计量。

注记. 次序统计量，对之后求均匀分布性质等内容时候可以简化问题。

定义 1.1.6. 样本向量

由n个随机变量组成的向量，称之为样本向量，或观测向量。通常记作$\mathbf{X} = (X_1, X_2, \cdots, X_n)'$。

定义 1.1.7. 分布族

分布族是一个包含所有可能的分布的集合。

定义 1.1.8. 统计模型

样本的分布族的集合，称之为统计模型。通常记为$\mathcal{P}$。

定义 1.1.9. 参数统计模型

设$\mathbf{X} = (X_1, X_2, \cdots, X_n)'$是来自总体的样本，$P_\theta$是总体分布族中的分布，$\theta$是参数空间中的参数，那么$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$称之为参数统计模型。

注记. 参数统计模型可以看作一个代数结构，其中$\mathcal{P}$是一个集合，Θ是一个参数空间，P_θ是一个分布族，θ是参数空间中的参数。

例 1.1.10. 正太分布与正太统计族正太分布式一种连续型的概率分布，一个随机变量 X 服从正态分布，记作 $X \sim N(\mu, \sigma^2)$，如果它的概率密度函数为：

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

正太分布族一组所有具有正态分布形式的概率分布的集合。这组分布由均值 μ 和方差 σ^2 这两个参数确定。正态分布族中的每个成员（即具体的正态分布）可以通过不同的参数值来表示。

1.2 Convergence of Random Variables

这部分给出了上学期所讲述过的几种收敛与相关的定理，我们没有那么关心具体的定理的证明细节，只需要了解这些定理的统计上的意义即可。

连续映射定理和Slktsky定理两个定理说明收敛的性质的良好，切比雪夫不等式给出了一个概率不等式，强大数律和中心极限定理则是说明了在一定条件下样本均值的收敛性质。

定义 1.2.1. 收敛

Convergence in Distribution: $X_n \xrightarrow{\mathcal{L}} X$,

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t) \quad \textit{at all } t \tag{1.1}$$

Convergence in Probability: $X_n \xrightarrow{P} X$,

$$\forall \epsilon > 0 : \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

Almost Sure Convergence: $X_n \xrightarrow{a.s.} X$,

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Convergence in r-th Mean: $X_n \xrightarrow{r} X$,

$$\lim_{n\to\infty} E|X_n - X|^r = 0, \quad r > 0.$$

定理 1.2.2. 连续映射定理

如果 $X_n \xrightarrow{a.s.,P,L} X$，且函数 g 存在一组不连续点 D_g，使得 $P[X \in D_g] = 0$，那么 $g(X_n) \xrightarrow{a.s.,P,L} g(X)$。

注记. 这个定理表明，只要函数 g 在 X 的不连续点上的概率为 0，那么 $g(X_n)$ 也会收敛到 $g(X)$。

定理 1.2.3. *Slutsky* 定理

如果 $X_n \xrightarrow{L} X$，且 $Y_n \xrightarrow{P} c$，其中 c 是一个常数，则 *(a).* $X_n + Y_n \xrightarrow{L} X + c$ *(b).* $X_n Y_n \xrightarrow{L} cX$ *(c).* $X_n/Y_n \xrightarrow{L} X/c$，前提是 c 是可逆的。

注记. 随机变量与另一个随机变量的和、积、商也会收敛到相应的值。

定理 1.2.4. 切比雪夫不等式

如果 X 是任意随机变量，则对于任意 $b > 0$，有

$$P(|X - E[X]| \geq b) \leq \frac{Var(X)}{b^2}$$

定理 1.2.5. 强大数律*(LLN)*

设 $X_1, X_2, \cdots$ 是独立同分布的随机变量序列，如果 $E|X| < \infty$，那么有

$$\bar{X}_n \equiv \frac{S_n}{n} \to EX \quad as \quad n \to \infty \quad w.p.1$$

定理 1.2.6. 中心极限定理*(CLT)*

设 $X_1, X_2, \cdots$ 是独立同分布的随机变量序列，且 $E[X_i] = \mu, Var(X_i) = \sigma^2$，那么有

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1) \quad as \quad n \to \infty$$

定理 1.2.7. 设 $X_i, i = 1, \ldots, n$ 为服从分布函数为F 的一列*i.i.d.* 其经验分布函数定义为：

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}$$

他有如下的渐进收敛性质：

1. *For any fixed* $x \in \mathbb{R}$, $nF_n(x) \sim B(n, F(x))$ *and*

$$\sqrt{n}\left(F_n(x) - F(x)\right) \xrightarrow{L} \mathcal{N}\left(0, F(x)(1 - F(x))\right)$$

2. ***Glivenko-Cantelli Theorem (Laws of Large Numbers):***

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

3. ***Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality:*** *for any* $\epsilon > 0$ *and* n,

$$P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \epsilon\right) \le 2e^{-2n\epsilon}$$

如上的收敛性质*1*说明了经验分布函数的分布，*2*说明了经验分布函数的一致收敛性，*3*说明了经验分布函数的概率收敛性。

Lec 2 Statistical model

主要讲了指数族和一些统计模型的基本概念。

2.1 Exponential Family

定义 2.1.1. 指数族

考虑参数分布族$\{P_\theta : \theta \in \Theta\}$，设 X 是一个随机变量，$f(x;\theta)$ 是 X 的概率密度函数，如果存在实数 θ，和函数 $h(x), \eta(\theta), B(\theta)$，使得

$$f_\theta(x) = h(x)e^{\eta^T(\theta)T(x)-B(\theta)}$$

那么称考虑参数分布族$\{P_\theta : \theta \in \Theta\}$为**指数族**。

其中$B(\theta)$称为**势函数**，显然，势函数可以写为

$$B(\theta) = log \int_\Omega e^{\eta^T(\theta)T(x)-B(\theta)}dx$$

注记. $f(x;\theta)$与$f_\theta(x)$是固定θ，仅与x有关的同一个函数。

性质 2.1.2. 指数族的支撑集，

$$\mathcal{G} = \{x : f(x,\theta) > 0\} = \{x : h(x) > 0\}$$

与参数θ无关。

例 2.1.3. 均匀分布集

$[-\theta, 1]$上的均匀分布族$\{U(0,\theta) : \theta > 0\}$不是指数分布族，因为他的支撑集与参数$\theta$有关。

例 2.1.4. 泊松分布

泊松分布是指数族，因为

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{(ln\lambda)x - \lambda}$$

定义 2.1.5. 指数族的自然形式

若分布族的密度函数可以写为

$$f_\theta(x) = h(x) e^{\theta' T(x) - B(\theta)}$$

也就是$\eta(\theta) = \theta$，则称其为T生成的**自然指数族**。自然参数空间定义为

$$\mathcal{E} = \{\theta : f(x,\theta) > 0\} = \{\theta : e^{B(\theta)} < \infty\}$$

性质 2.1.6. 自然参数空间$\mathcal{E}$必为凸集，势函数$B(\theta)$必为$\mathcal{E}$上的凸函数。

定义 2.1.7. 自然形式指数族的秩

设$P_\eta : \eta \in \mathcal{E}$是自然形式指数族，如果存在某$T(x)$的投影向量$\hat{T}(x) = (T_1(x), \ldots, T_k(x))$的维数为$k$，且$P\{1, T_1(x), \ldots, T_k(x)$线性无关$\} > 0$，则称$k$为指数族的秩。

称指数族满秩当且仅当$k = n$，n为样本的维数。

若指数族不满秩，则该指数族被称为曲线指族。一个常见的曲线指数族的例子为，考虑正态分布族$\{N(\mu, \mu^2) : \mu \in \mathbb{R}\}$。

性质 2.1.8.

$$E_\eta[T_i(x)] = \frac{\partial}{\partial \eta_i} B(\theta)$$

$$Cov_\eta[T_i(x), T_j(x)] = \frac{\partial^2}{\partial \eta_i \partial \eta_j} B(\theta)$$

性质 2.1.9. $T(x)$的矩母函数为

$$M_\eta(t) = E_\eta[e^{t^T T(x)}] = e^{B(\eta+t)-B(\eta)}$$

定理 2.1.10. *If the distribution of X belongs to a canonical exponential family and η is an interior point of $\mathcal{E}$, then for any integrable function $g(x)$, the following function satisfies:*

$$G(\eta) = \int \cdots \int g(x) \exp(\eta^T T(x)) h(x)\, dx$$

has the following asymptotic properties:

$$\frac{\partial^m G(\eta)}{\partial \eta_{m_1} \cdots \partial \eta_{m_k}} = \int \cdots \int T_{m^1}(x) \cdots T_{m^k}(x)[g(x) \exp\{\eta^T T(x)\} h(x)]\, dx$$

where $m_1 + \cdots + m_k = m$.

注记. 我也不知道这个定理有什么用，可能超出了我们学习这门课程所涉及的范围。但是据说在深入讨论MGF和CGF的时候会用到。相关的概念和定理包括：

1.指数族分布的充分统计量性质（Properties of sufficient statistics in exponential families）

2.矩生成函数和累积量生成函数的导数性质（Derivatives of moment generating functions and cumulant generating functions）

2.2 The Maximum Entropy Duality*

这一节给出了定义指数族的动机。

假设给定一个来自某个分布的随机样本 $\{X_1, \ldots, X_n\}$，并且计算某些函数的经验期望：

$$\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} T_i(X_j) \quad \text{对于 } i \in \{1, \ldots, k\}$$

仅基于这些经验期望，希望推断出样本的完整概率分布。如果分布 p 满足以下条件，则认为它与观察到的数据一致，

$$\hat{\mu}_i = \mathbb{E}_p[T_i(X)] \quad \text{对于 } i \in \{1, \ldots, k\}$$

定义 2.2.1. 熵

分布的熵定义为:

$$H(p) = -\int p(x) \log(p(x))\, dx$$

考虑最大熵原理时，建议选择具有最大（香农）熵的分布。满足约束条件的分布 p^*可以由此定义出来，即

$$p^* = \arg\max_p H(p)$$

其中约束条件为

$$\hat{\mu}_i = \mathbb{E}_p[T_i(X)] \quad \text{对于} i \in \{1, \ldots, k\}.$$

该问题的解可以显示为以下形式

$$p^*(x) = \exp\left[\sum_{i=1}^{k} \theta_i T_i(x) - A(\theta)\right] h(x).$$

这提供了指数族的动机。

2.3 Location and Scale Families

定义 2.3.1. 位置-尺度族

如果 X 有密度 f_X， Y 有密度

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{y-\sigma}{\mu}\right)$$

可以称所有的$f_Y(y)$为f_X的位置-尺度族。

性质 2.3.2. 其中X与Y有如下性质：

$Y \stackrel{d}{=} \sigma X + \mu$

$E[Y] = \sigma E[X] + \mu, Var[Y] = \sigma^2 Var[X]$

并且称位置变换族为$\{f_\mu(x) = f_X(x-\mu)\}$

尺度变换族为$\{f_\sigma(x) = \frac{1}{\sigma} f_X\left(\frac{x}{\sigma}\right)\}$。

Lec 3 Sampling distribution

举例一些特殊分布的$\bar{X}_n$和S_n^2

3.1 Preliminary:χ^2, t and F distributions

此节特殊函数的性质都易于证明，在此仅全部列举而出。

定义 3.1.1. 卡方分布

设$Z_1, Z_2, \cdots, Z_n$是独立同分布的标准正态随机变量，那么随机变量$Q = \sum_{i=1}^{n} Z_i^2$服从自由度为n的卡方分布，记作$Q \sim \chi^2(n)$。

$X \sim \chi_n^2$,卡方分布的密度函数为

$$g_n(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, & x > 0 \\ 0 & x \leq 0 \end{cases}$$

X的特征函数为$\phi(t) = (1-2it)^{-n/2}$。

性质 3.1.2. 卡方分布的数学期望和方差分别为$E(Q) = n$，$Var(Q) = 2n$。

性质 3.1.3. 设$Q_i \sim \chi_{n_i}^2, i = 1, 2, \ldots, k$，且$Q_i$相互独立，那么$\sum_{i=1}^{k} Q_i \sim \chi_{n_1+n_2+\cdots+n_k}^2$。

定义 3.1.4. t分布

设$Z \sim N(0,1)$，$Q \sim \chi^2(n)$，且Z与Q相互独立，那么随机变量$T = \frac{Z}{\sqrt{Q/n}}$服从自由度为n的t分布，记作$T \sim t(n)$。

t分布的密度函数为

$$f_n(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{t^2}{n})^{-\frac{n+1}{2}}$$

性质 3.1.5. 对于 $T \sim t_n$，$r < n$ 时的期望矩 $E(T^r)$（当 $n > 1$ 时）为

$$E(T^r) = \begin{cases} \frac{n^{\frac{r}{2}}\Gamma(\frac{r+1}{2})\Gamma(\frac{n-r}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}, & r \text{ 为偶数}, \\ 0, & r \text{ 为奇数}. \end{cases}$$

性质 3.1.6. $n \geq 2$时，$E(T) = 0$。 $n \geq 3$时，T 的方差为 $Var(T) = \frac{n}{n-2}$。

性质 3.1.7. 当 $n \to \infty$ 时，$t_n(x)$ 趋近于标准正态分布 $N(0,1)$。

定义 3.1.8. F分布

设$X \sim \chi^2_m, Y \sim \chi^2_n$，且$X$与$Y$相互独立，那么随机变量$F = \frac{X/m}{Y/n}$服从自由度为$(m,n)$的$F$分布，记作$F \sim F(m,n)$。

F分布的密度函数为

$$f_{m,n}(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}(\frac{m}{n})^{m/2}x^{m/2-1}(1+\frac{m}{n}x)^{-(m+n)/2}, & x > 0 \\ 0 & x \leq 0 \end{cases}$$

性质 3.1.9. 若 $Z \sim F_{m,n}$，则 $\frac{1}{Z} \sim F_{n,m}$

性质 3.1.10. 对于 $Z \sim F_{m,n}$ 且 $r > 0$ 时，

对于 $n > 2r$

$$E(Z^r) = \left(\frac{n}{m}\right)^r \frac{\Gamma\left(\frac{m}{2}+r\right)\Gamma\left(\frac{n}{2}-r\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

对于 $n > 2$，$E(Z) = \frac{n}{n-2}$。

对于 $n > 4$，Z 的方差为

$$Var(Z) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

性质 3.1.11. 若 $T \sim t_n$，则 $T^2 \sim F_{1,n}$

性质 3.1.12. 对于 $0 < \alpha < 1$，$F_{m,n}(1-\alpha) = \frac{1}{F_{n,m}(\alpha)}$

3.2 Normal population case

在概率论中已经细致地学习过高维正态分布的一些性质，此处仅列出其中重要的部分。

引理 3.2.1. 设 $X = (X_1, \ldots, X_n)^T$，$\mu = (\mu_1, \ldots, \mu_n)^T$，且 Σ 为正定矩阵。如果 $X \sim N_n(\mu, \Sigma)$，则：

1. X 是 $N_n(\mu, \Sigma)$ 当且仅当任何非零向量 a 的线性组合 $a^T X$ 都有正态分布。
2. 线性组合 $a^T X$ 的分布为 $N(a^T\mu, a^T\Sigma a)$。
3. 如果 a 是 $m \times 1$ 向量，B 是 $m \times n$ 矩阵，那么 $a + BX$ 的分布为 $N_m(a + B\mu, B\Sigma B^T)$。
4. BX 和 $X^T AX$ 独立当且仅当 $B\Sigma A = 0$。

定理 3.2.2. 设 $X_1, X_2, \ldots, X_n$ 为来自 $N(a, \sigma^2)$ 的独立同分布样本，则：

1. 样本均值 $\overline{X}$ 服从 $N(a, \sigma^2/n)$。
2. $(n-1)S^2/\sigma^2$ 服从自由度为 $n-1$ 的卡方分布。
3. 如果总体为正态分布，$\overline{X}$ 和 S^2 独立。
4. $T_n = \sqrt{n}(\overline{X} - a)/S$ 服从自由度为 $n-1$ 的 t 分布。

3.3 Asymptotic results for non-normal case

在有多个相同分布叠加的时候，注意到他们经常会服从卡方分布，t分布，F分布。

定理 3.3.1. *(CLT)* 设 $X_1, X_2, \ldots, X_n$ 为来自总体的独立同分布样本，$E(X_i) = \mu$，$Var(X_i) = \sigma^2$。当 $n \to \infty$时

$$\overline{X}_n \xrightarrow{P} \mu$$

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to N(0,1)$$

定理 3.3.2. 当样本来自具有四阶矩存在的分布时，样本方差的标准化形式会趋向于正态分布。

设 $X_1, X_2, \ldots, X_n$ 为来自总体的独立同分布样本，$E(X_i) = \mu$，$Var(X_i) = \sigma^2$。样本的方差为 $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$，那么:

$$S_n^2 \xrightarrow{P} \sigma^2$$

$$\sqrt{n}(S_n^2 - \sigma^2) \to N(0, \mu_4 - \sigma^4) \quad where \quad \mu_4 = E(X_1 - \mu)^4$$

推论 3.3.3. 设 $X_1, X_2, \ldots, X_n, i.i.d$ 服从密度函数为$f(x,\lambda) = \lambda e^{-\lambda x} I_{x>0}(x)$的指数分布$Exp(\lambda)$，那么:

$$2\lambda n\bar{X} = 2\lambda \sum_{i=1}^{n} X_i \sim \chi_{2n}^2$$

推论 3.3.4. 设$X_1, X_2, \cdots, X_m i.i.d \sim N(\mu_1, \sigma^2)$，$Y_1, Y_2, \cdots, Y_n i.i.d \sim N(\mu_2, \sigma^2)$，且$X_i$与$Y_j$相互独立，那么:

$$T = \frac{\bar{X} - \mu_1}{S_1}\sqrt{n} \sim t_{m-1} \tag{3.1}$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_\omega}\sqrt{\frac{mn}{m+n}} \sim t_{m+n-2} \tag{3.2}$$

其中$(n+m-2)S_\omega^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2$。

推论 3.3.5. 设$X_1, X_2, \cdots, X_m i.i.d \sim N(\mu_1, \sigma_1^2)$，$Y_1, Y_2, \cdots, Y_n i.i.d \sim N(\mu_2, \sigma_2^2)$，且$X_i$与$Y_j$相互独立，那么:

$$F = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1}$$

Lec 4 Sampling distributions of order statistics

本节仅列举次序统计量的分布性质，不作证明。

4.1 Order statistics

定义 4.1.1. 次序统计量

设$\{X_i\}$为$i.i.d$随机变量，将其值按照大小排列为$\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$（$i.e.\ X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$）则$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$称为次序统计量。

定理 4.1.2. 设$X_1, X_2, \cdots, X_n$是来自总体的样本，$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$是其次序统计量，$f(x)$是总体的概率密度函数，$F(x)$是总体的分布函数，则次序统计量的联合分布密度函数为

$$f_{X_{(1)},X_{(2)},\cdots,X_{(n)}}(x_1, x_2, \cdots, x_n) = n! f(x_1) f(x_2) \cdots f(x_n) I_{\{x_1 \leq x_2 \leq \cdots \leq x_n\}} \quad (4.1)$$

边际分布密度函数为

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1} [1 - F(x)]^{n-i} f(x)$$

而两个统计次序量的联合分布为

$$f_{X_{(i)},X_{(j)}}(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(x)^{i-1} [F(y) - F(x)]^{j-i-1} [1 - F(y)]^{n-j} f(x) f(y)$$

例 4.1.3. 均匀分布的次序统计量

设$X_1, X_2, \cdots, X_n$是来自$U(0,1)$的样本，$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$是其次序统计量，那么$X_{(i)}$的*pdf*为

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}x^{i-1}(1-x)^{n-i} \tag{4.2}$$

巧合地注意到，$X_{(i)}$的*pdf*与$Beta(i, n-i+1)$的*pdf*相同。

例 4.1.4. 二项分布的次序统计量

设$X_1, X_2, \cdots, X_n$是来自$B(1,p)$的样本，$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$是其次序统计量，那么$X_{(i)}$的*pdf*为

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}p^{i}(1-p)^{n-i}$$

巧合地注意到，$X_{(i)}$的*pdf*与$Beta(i, n-i+1)$的*pdf*相同。

注记. 意外地发现，二项分布与均匀分布的次序统计量的pdf是一样的，都可以看作是$Beta(i, n-i+1)$的pdf。

例 4.1.5. 样本极差

设$X_1, X_2, \cdots, X_n$是来自总体的样本，$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$是其次序统计量，那么样本极差$R = X_{(n)} - X_{(1)}$的*pdf*为

$$f_R(r) = n(n-1)[F(r)]^{n-2}f(r)$$

4.2 Properties of sample quantiles

定义 4.2.1. 样本分位数

设X的分布函数为$F(x)$，对于$0 < p < 1$，$\xi(p)$称为X的p分位数，如果

$$\xi(p) = F^{-1}(p) = inf\{x : F(x) \geq p\}$$

注记. 如果X的分布函数$F(x)$是连续的，那么$F(x)$的p分位数$\xi(p)$就是$F^{-1}(x)$，此时逆函数良定（$i.e. F^{-1}(x)$唯一）。

如果X的分布函数$F(x)$是不连续的，那么F的p分位数$\xi(p)$就是$F^{-1}(x)$的下确界，或者说存在x_0使得$F(x_0^-) < p < F(x_0)$，即$\xi(p) = x_0$。

性质 4.2.2. 均匀分布的分位数

设U服从$U(0,1)$,$X = F^{-1}(U)$，那么X的分布函数为$F(x)$

这给出了一个已知$F(x)$生成X=X的方法：取$U(0,1)$的随机变量U，然后取$X = F^{-1}(U)$。

性质 4.2.3. 平移-尺缩分布的分位数

设X服从$F(x)$，$Y = a + bX$，那么Y的分布函数为$F_Y(y) = F_X(\frac{y-a}{b})$，$Y$的$p$分位数为$\xi_Y(p) = a + b\xi_X(p)$。

定义 4.2.4. 样本分位数

设 $\{X_1, X_2, ..., X_n\}$ 是从 F 中抽取的样本。样本第 p 个分位数 $\hat{\xi}_{pn}$（根据 np 是否为整数而定，当 np 为整数时为 $X_{([np])}$，否则为 $X_{([np]+1)}$，这里 $[np]$ 表示 np 的整数部分）定义如下：

$$\hat{\xi}_{pn} = F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$$

其中 F_n 是相应的经验分布函数。具体而言，$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\}$。

注记. 形象的理解是，$F_n(x)$是样本中小于等于 x 的比例，$\hat{\xi}_{pn}$ 是样本中最先达到这一样本比例的 x。

事实上$\hat{\xi}_{pn}$写成$\hat{\xi}_n(p)$，与分位数$\xi(p)$使用类似的记法，或许看起来更直观一点。

例 4.2.5. 样本中位数

比如一列数$1,5,7,2,3$，那么$F_n(1)=\frac{1}{5},F_n(2)=\frac{2}{5},F_n(3)=\frac{3}{5},F_n(5)=\frac{4}{5},F_n(7)=\frac{5}{5}$，所以样本中位数为$\hat{\xi}_{0.5}=3$。

例子是很直观的，甚至说我们小学就已经接触过这些了，定义看起来有点唬人而已。

性质 4.2.6. 次序统计量与分位数

设$X_1,X_2,\cdots,X_n$是来自总体的样本，$X_{(1)},X_{(2)},\cdots,X_{(n)}$是其次序统计量，那么样本的$p$分位数为

$$\hat{\xi}_n(p)=\hat{\xi}_{pn}=X_{([np])}=F_n^{-1}(p) \tag{4.3}$$

也就是 $X_{(i)}=F^{-1}(\frac{i}{n})=\hat{\xi}_{\frac{i}{n}n}$。

定理 4.2.7. 样本分位数的渐进性质

设$X_1,X_2,\cdots,X_n$是来自总体的样本，$X_{(1)},X_{(2)},\cdots,X_{(n)}$是其次序统计量，那么有

$$\hat{\xi}_{pn}\xrightarrow{P}\xi(p)$$

事实上，样本的p分位数$\hat{\xi}_{pn}$是p的一致渐进分布，即

$$\sqrt{n}(\hat{\xi}_{pn}-\xi(p))\xrightarrow{d}N(0,\frac{p(1-p)}{f^2(\xi(p))})$$

例 4.2.8. 均匀分布的样本分位数的渐进性质

设$X_1,X_2,\cdots,X_n$是来自$U(0,1)$的样本。给定$p=i/n$，有

$$\hat{\xi}_{np}\overset{4.3}{=}X_{([np])}$$

其中服从的分布是

$$\hat{f}_{np}(x)=P(\{\hat{\xi}_{np}=x\})=f_{X_{([np])}}(x)=\frac{n!}{([np]-1)!(n-[np])!}x^{[np]-1}(1-x)^{n-[np]}$$

可以求得与p的分位数$\xi(p)=p$的密度分布函数$f_n=\begin{cases}1 & x=p\\ 0 & x\neq p\end{cases}$满足依概率收敛的条件

因此满足$\hat{\xi}_{pn} \xrightarrow{P} \xi(p)$ 再由中心极限定理，有

$$\sqrt{n}(\hat{\xi}_n(p) - \xi(p)) = \sqrt{n}(X_{([np])} - p) \xrightarrow{d} N(0, p(1-p))$$

证明. 不妨假设$\hat{\xi}_{np} - \xi_p = O_p(d_n)$。考虑

$$P(p \le F_n(\xi_p + d_n)) = P(\hat{\xi}_p \le \xi_p + d_n)$$

记$p_n = F(\xi_p + d_n)$，则$nF_n(\xi_p + d_n) = \sum_{i=1}^{n} I\{x_i \le \xi_p + d_n\} \sim Bin(n, p_n)$，由中心极限定理有：

$$\begin{aligned} P(\hat{\xi}_p \le \xi_p + d_n) &= P(p \le F_n(\xi_p + d_n)) \\ &= P(\frac{np - np_n}{\sqrt{np_n(1-p_n)}} \le \frac{nF_n(\xi_p + d_n) - np_n}{\sqrt{np_n(1-p_n)}}) \\ &\xrightarrow{d} 1 - \Phi(-\frac{\sqrt{n}d_n F'(\xi_p')}{\sqrt{p_n(1-p_n)}}) \\ &= \Phi(\frac{\sqrt{n}d_n F'(\xi_p')}{\sqrt{p_n(1-p_n)}}) \end{aligned}$$

其中Φ是标准正态分布的概率分布函数这是因为，由中值定理，

$$p - p_n = F(\xi_p) - F(\xi_p + d_n) = d_n f(\xi_p')$$

其中$\xi_p' \in (\xi_p,\ \xi_p + d_n)$。

我们希望能够选择合适的d_n，使得$\frac{\sqrt{n}d_n F'(\xi_p')}{\sqrt{p_n(1-p_n)}} = const$。可以令$d_n = \frac{t\sqrt{p(1-p)}}{\sqrt{n}F'(\xi_p)}$。此时$d_n \to 0, p_n \to 0, \xi_p' \to \xi_p$

于是由连续映射定理和$Slutsky$定理，上式可以写成

$$P(\hat{\xi}_p \le \xi_p + d_n) = P(\sqrt{n}(\hat{\xi}_p - \xi_p)\frac{F'(\xi_p)}{\sqrt{p(1-p)}} \le t) \xrightarrow{d} \Phi(t)$$

□

Lec 5 Sufficient and Complete Statistics

本节要讨论的是统计量的性质，这是统计推断的基础。数理统计研究的是如何以样本来推断总体的性质，而统计量是样本的函数。

统计量的性质可以分为两类：一类是关于统计量的无偏性、有效性、一致性等性质，另一类是关于统计量的充分性、完全性等性质。本节主要讨论后者。

5.1 Sufficient Statistics

对于总体的不同性质，一个自然的想法是希望找到一种能够充分说明总体的某一性质的统计量，不仅是为了简化庞大的数据集为统计量，也是为了减少数据集中异常值的干扰。

求充分统计量是重点

注记. 充分性原则

如果$T(X)$为参数θ的充分统计量，则依赖样本$\mathbf{X}$对θ的任何统计推断仅通过统计量$T(\mathbf{X})$进行，不会有关于θ信息的损失。

也就是说我们希望$T(X)$满足，当$\mathbf{x}$与$\mathbf{y}$为满足$T(\mathbf{x}) = T(\mathbf{y})$的两个样本点时，无论观察到$\mathbf{X} = \mathbf{x}$还是$\mathbf{X} = \mathbf{y}$，关于$\theta$的推断都是相同的。

上面的描述可能听起来很费解，但是可以看完定义下面的例子之后再

来对照着看上面的统计意义下的描述，可能会清晰许多。

定义 5.1.1. 充分统计量

设$X_1, X_2, \cdots, X_n$是来自总体$F(x;\theta)$的样本，$T(\mathbf{X})$是$X_1, X_2, \cdots, X_n$的函数，如果对于任意θ，给定$T(\mathbf{X})$的条件下，样本$\mathbf{X}$的条件分布不依赖于θ，则称$T(\mathbf{X})$是θ的充分统计量。

i.e.

$$P_\theta(X_1 < x_1, \cdots, X_n < x_n | T(\mathbf{X}) = t) = \psi(t)$$

$\psi(t)$与θ无关。

例 5.1.2. 二项分布的充分统计量

设$X_1, X_2, \cdots, X_n$是来自$B(1,p)$的样本，$T = \sum_{i=1}^n X_i$，那么T是p的充分统计量。

证明. 二项分布的概率密度函数为

$$f_p(x) = C_n^x p^x (1-p)^{n-x}$$

那么$X_1, X_2, \cdots, X_n$的联合概率密度函数为

$$f_p(x_1, x_2, \cdots, x_n) = C_n^{x_1} C_n^{x_2} \cdots C_n^{x_n} p^{\sum x_i} (1-p)^{n-\sum x_i}$$

由于$T = \sum_{i=1}^n X_i \sim B(1,p)$，所以$T$的概率密度函数为

$$f_p(t) = C_n^t p^t (1-p)^{n-t}$$

因此

$$\begin{aligned} P_p(X_1 = x_1, \cdots, X_n = x_n | T = t) =& \frac{P_p(X_1 = x_1, \cdots, X_n = x_n, T = t)}{P_p(T = t)} \\ =& \frac{f_p(x_1, x_2, \cdots, t - x - 1 - \cdots - x_{n-1})}{f_p(t)} \\ =& \frac{C_n^{x_1} C_n^{x_2} \cdots C_n^{t - x_1 - \cdots - x_{n-1}}}{C_n^t} = \psi(t) \end{aligned}$$

由于$\psi(t)$与p无关，所以T是p的充分统计量。 □

例 5.1.3. 泊松分布的充分统计量

设$X_1, X_2, \cdots, X_n$是来自$P(\lambda)$的样本，$T = \sum_{i=1}^{n} X_i$，那么T是λ的充分统计量。

证明. 泊松分布的概率密度函数为

$$f_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

那么$X_1, X_2, \cdots, X_n$的联合概率密度函数为

$$f_\lambda(x_1, x_2, \cdots, x_n) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{x_1!x_2!\cdots x_n!}$$

由于$T = \sum_{i=1}^{n} X_i \sim P(n\lambda)$，所以$T$的概率密度函数为

$$f_{n\lambda}(t) = \frac{e^{-n\lambda}(n\lambda)^t}{t!}$$

$$\begin{aligned} P_p(X_1 = x_1, \cdots, X_n = x_n | T = t) =& \frac{P_p(X_1 = x_1, \cdots, X_n = x_n, T = t)}{P_p(T = t)} \\ =& \frac{f_p(x_1, x_2, \cdots, t - x - 1 - \cdots - x_{n-1})}{f_p(t)} \\ =& \frac{t!}{n^t \prod_{i=1}^{n} x_i!} = \psi(t) \end{aligned}$$

由于$\psi(t)$与p无关，所以T是p的充分统计量。 □

定理 5.1.4. 因子分解定理

设$X_1, X_2, \cdots, X_n$是来自总体$F(x;\theta)$的样本，$T(\mathbf{X})$是$X_1, X_2, \cdots, X_n$的函数，那么$T(\mathbf{X})$是θ的充分统计量当且仅当存在两个函数$h(\mathbf{x}), g(t,\theta)$使得

$$f_\theta(\mathbf{x}) = h(\mathbf{x})g(T(\mathbf{x}), \theta)$$

简而言之我们只需要求出联合概率密度f，找出他与T, θ的因子，然后剩下的部分如果只与样本相关即可

例 5.1.5. 指数分布的充分统计量

设$X_1, X_2, \cdots, X_n$是来自$Exp(\lambda)$的样本

指数分布的概率密度为

$$f_\lambda(x) = \lambda e^{-\lambda x}$$

指数分布的联合概率密度为

$$f_\lambda(x_1, x_2, \cdots, x_n) = \lambda^n e^{-\lambda \sum x_i}$$

取$T = \sum_{i=1}^n X_i, g(t, \lambda) = \lambda^n e^{-\lambda t}$即可判断

例 5.1.6. 均匀分布的充分统计量

设$X_1, X_2, \cdots, X_n$是来自$U(a, b)$的样本，参数未知，求他的一个充分统计量

样本的联合概率密度函数为

$$p_{a,b}(\mathbf{x}) = \frac{1}{b-a} \prod_{i=1}^n I(a \le x_{(1)}) I(b \ge x_{(n)})$$

取$T = (x_{(1)}, x_{(n)})$即可判断

注记. 参数为二维的，充分统计量也为二维的，否则信息丢失过多。在多维情况下对$g(t, \theta)$的要求是表达成$g(t_1, \cdots, t_n, \theta_1, \cdots, \theta_n)$即可

例 5.1.7. 指数族分布的充分统计量

Let $X_1, X_2, \cdots, X_n$ *are i.i.d. from* $f(x; \theta) = h(x)c(\theta)e^{Q(\theta)T(x)}$

then $f_\theta(\mathbf{x}) = \prod_{i=1}^n h(x_i)c(\theta)e^{Q(\theta)T(x_i)} = h(\mathbf{x})c(\theta)^n e^{Q(\theta)\sum T(x_i)}$, *then by the factorisation ,*$T(\mathbf{x}) = \sum T(x_i)$ *is sufficient for* θ.

注记. 也就是对于所有的指数族，我们只需要求出$\sum T_i(x)$即可。

定义 5.1.8. 极小充分统计量

设$T(\mathbf{X})$是θ的充分统计量，如果对于任意θ的充分统计量$T'(\mathbf{X})$，存在映射φ使得$T'(\mathbf{X}) = \varphi(T(\mathbf{X}))$，则称$T(\mathbf{X})$是$\theta$的极小充分统计量。

注记. 从这里我们也可以发现，通过构造$\mathscr{T} = \varphi(T)|\varphi isbijective$就可以得到$\theta$的极小充分统计量的集合。不虑双射映射复合，极小充分统计量唯一。

特别的还有θ的充分统计量不唯一。

定理 5.1.9. 极小充分统计量判定

设 $\mathcal{X}$ 为样本空间，$\{p(x;\theta), \theta \in \Theta\}$ 为一族在 $\mathcal{X}$上的pdf。如果存在 T 满足，对于所有的 $x, y \in \mathcal{X}$

$$p(x;\theta) = C_{x,y}p(y;\theta) \iff T(x) = T(y)$$

对任意的 θ 和任意几个 $C_{x,y} \in \mathbb{R}$ 。这里我们使用下标强调此参数与 θ无关。那么 T 是一个极小统计量。

这里的$for\ some\ C_{x,y}$容易令人迷惑，因此给出证明以辅助理解。

证明. T is sufficient: Start with

$$T(\mathcal{X}) = \{t : t = T(x) \text{ for some } x \in \mathcal{X}\} = \text{range of } T.$$

For each $t \in T(\mathcal{X})$, consider the preimage $A_t = \{x : T(x) = t\}$ and select an arbitrary representative x_t from each A_t. Then, for any $y \in X$ we have $y \in A_{T(y)}$ and $x_{T(y)} \in A_{T(y)}$. By the definition of A_t this implies that $T(y) = T(x_{T(y)})$. From the assumption of the theorem,

$$p(y;\theta) = C_{y,x_{T(y)}}p(x_{T(y)};\theta) = h(y)g_\theta(T(y))$$

which yields sufficiency of T by the NFFC.

If $T(x) = T(y)$, such that $p(x;\theta)/p(y;\theta) = C_{x,y}$ does not depend on θ, then T is sufficient.

T is minimal: Consider another sufficient statistic T'. By the NFFC

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x).$$

Take any x, y such that $T'(x) = T'(y)$. Then

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x) = \tilde{g}_\theta(T'(y))\tilde{h}(y)\tilde{h}(x)/\tilde{h}(y) = p(y;\theta)C_{x,y}$$

Hence, $T(x) = T(y)$ by the assumption of the theorem. So, $T'(x) = T'(y)$ implies $T(x) = T(y)$ for any sufficient statistic T' and any x, y. As a result, T is a minimal sufficient statistic.

If $T(x)$ is sufficient, and $p(x;\theta)/p(y;\theta) = C_{x,y}$ not depending on θ implies $T(x) = T(y)$, then T is minimal sufficient. □

如果还是感到迷惑，这里给出例子的完整说明以辅助理解

例 5.1.10. 设 $Y_1, Y_2, \ldots, Y_n$ 是独立同分布的 *Poisson(θ)* 随机变量，其中 $\theta > 0$ 是参数。要找出 $Y_1, Y_2, \ldots, Y_n$ 的极小充分统计量。

首先， *Poisson* 分布的概率质量函数：

$$P(Y_i = y_i;\theta) = \frac{\theta^{y_i}e^{-\theta}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

联合概率质量函数为：

$$P(Y_1 = y_1, \ldots, Y_n = y_n;\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}e^{-\theta}}{y_i!} = \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!}$$

定义统计量 $T = \sum_{i=1}^n Y_i$。我们可以将联合概率质量函数重写为：

$$P(Y_1 = y_1, \ldots, Y_n = y_n;\theta) = \frac{\theta^T e^{-n\theta}}{\prod_{i=1}^n y_i!}$$

根据定理 *2*，如果存在统计量 T 使得对于每对 $x, y \in X$ 和任意参数 θ，我们有：

$$p(x;\theta) = C_{x,y}p(y;\theta) \iff T(x) = T(y)$$

则 T 是极小充分统计量。现在我们验证 $T = \sum_{i=1}^{n} Y_i$ 是否满足这个条件。

考虑两个样本 $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ 和 $\mathbf{z} = (z_1, z_2, \ldots, z_n)$，联合概率质量函数分别为：

$$P(Y_1 = y_1, \ldots, Y_n = y_n; \theta) = \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!}$$

和

$$P(Y_1 = z_1, \ldots, Y_n = z_n; \theta) = \frac{\theta^{\sum_{i=1}^{n} z_i} e^{-n\theta}}{\prod_{i=1}^{n} z_i!}$$

如果 $T(\mathbf{y}) = T(\mathbf{z})$，即 $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i$，则有：

$$\frac{P(Y_1 = y_1, \ldots, Y_n = y_n; \theta)}{P(Y_1 = z_1, \ldots, Y_n = z_n; \theta)} = \frac{\frac{\theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!}}{\frac{\theta^{\sum_{i=1}^{n} z_i} e^{-n\theta}}{\prod_{i=1}^{n} z_i!}} = \frac{\prod_{i=1}^{n} z_i!}{\prod_{i=1}^{n} y_i!} = C_{\mathbf{y},\mathbf{z}}$$

其中 $C_{\mathbf{y},\mathbf{z}}$ 不依赖于 θ。反之，如果

$$\frac{P(Y_1 = y_1, \ldots, Y_n = y_n; \theta)}{P(Y_1 = z_1, \ldots, Y_n = z_n; \theta)} = C_{\mathbf{y},\mathbf{z}}$$

且 $C_{\mathbf{y},\mathbf{z}}$ 不依赖于 θ，则 $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i$，即 $T(\mathbf{y}) = T(\mathbf{z})$。

因此，根据定理 *2*，我们可以确定 $T = \sum_{i=1}^{n} Y_i$ 是 $Y_1, Y_2, \ldots, Y_n$ 的极小充分统计量。

例 5.1.11. 指数族的极小充分统计量

如果样本X服从满秩指数族分布，那么$T(X) = (T_1(X), \ldots, T_k(X))$是极小充分统计量。

例 5.1.12. *Gamma*分布的极小充分统计量

设$X_1, X_2, \ldots, X_n$是来自$Gamma(\alpha, \beta)$的样本，其中$\alpha > 0$和$\beta > 0$是参数。我们要找出$X_1, X_2, \ldots, X_n$的极小充分统计量。

*Gamma*分布的联合概率密度函数为：

$$f_{\alpha,\beta}(x_1, x_2, \ldots, x_n) = \frac{\alpha^{\beta n} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\beta \sum_{i=1}^{n} x_i}}{\Gamma(\alpha)^n} I(x_{(1)} \geq 0)$$

$$\frac{f_{\alpha,\beta}(x_1,x_2,\ldots,x_n)}{f_{\alpha,\beta}(y_1,y_2,\ldots,y_n)}=\frac{\prod_{i=1}^n x_i^{\alpha-1}e^{-\beta\sum_{i=1}^n x_i}}{\prod_{i=1}^n y_i^{\alpha-1}e^{-\beta\sum_{i=1}^n y_i}}=C_{x,y}$$

令$T=(\sum_{i=1}^n X_i,\prod_{i=1}^n X_i)$，取$C_{x,y}=1$再由$\alpha,\beta$的任意性得知$T(x)=T(y)$，所以$T=(\sum_{i=1}^n X_i,\prod_{i=1}^n X_i)$是极小充分统计量。

5.2 Ancillary Statistics*

这部分对于本门课程后续的内容并不是必须的，但是对于理解统计量的性质有一定的帮助。

定义 5.2.1. *Ancillary Statistics*

A statistic S is called ancillary if the distribution of S does not depend on the parameter θ.

注记. *Ancillary statistics are useful in the sense that they provide information about the parameter θ without affecting the distribution of the sample.*

例 5.2.2. *Cauchy distribution*

Consider a sample $X_1,X_2,\ldots,X_n\overset{i.i.d}{\sim}Cauchy(\theta)$ whose pdf is given by

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

then$(X_{(1)},\cdots X_{(n)})$ is minimal sufficient for θ and $R=X_{(n)}-X_{(1)}$ is ancillary.

例 5.2.3. *Let $X_1,...,X_n$ be iid uniform observations on the interval $(\theta,\theta+1),-\infty<\theta<\infty$. Letting $X_{(1)}=min_i x_i$ and $X_{(n)}=max_i x_i$, we will see that the range statistic $R=X_{(n)}-X_{(1)}$ is an ancillary statistic. Recall the cdf of a uniform random variable :* $F(x|\theta)=\begin{cases}0 & x\le\theta\\ x-\theta & \theta<x<\theta+1\\ 1 & x\ge\theta+1\end{cases}$.

Now recall that in the passage Random Sample, we have discussed the joint pdf of two ordered variables : THEN, the joint pdf of $X_{(i)}$ *and* $X_{(j)}, 1 \leq i < j \leq n, -\infty < u < v < \infty$ *is :*

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n! f_X(u) f_X(v) [F_X(u)]^{i-1} [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}}{(i-1)!(j-1-i)!(n-j)!}$$

. Thus, we can write

$$g(x_{(1)}, x_{(n)}) = \begin{cases} n(n-1)(x_{(1)} - x_{(n)})^{n-2} & x_{(n)} - 1 < \theta < x_1 \\ 0 & otherwise \end{cases}$$

. Make the transformation

$$\begin{cases} R = X_{(n)} - X_{(1)} \\ M = \frac{X_{(1)} + X_{(n)}}{2} \end{cases}$$

, and we have

$$\begin{cases} X_{(1)} = \frac{2M - R}{2} \\ X_{(n)} = \frac{2M + R}{2} \\ |J| = 1 \end{cases}$$

. Then we have the joint pdf of the range statistic and median statistic and the marginal pdf of $\mathbb{R}$*:*

$$h(r, m|\theta) = \begin{cases} n(n-1) r^{n-2} & 0 < r < 1, \frac{2\theta + r}{2} < m < \frac{2\theta - r + 2}{2} \\ 0 & otherwise \end{cases}.$$

$$h(r|\theta) = \int_{\frac{2\theta + r}{2}}^{\frac{2\theta - r + 2}{2}} n(n-1) r^{n-2} dm = n(n-1) r^{n-2}(1-r) \sim beta(n-1, 2)$$

. Still we can see that the pdf is constant as a function of θ*. Thus,* R *is ancillary. Here we can see that* R*, as an ancillary statistic, also gives vital information about* θ *because the statistic* (R, M) *is a minimal sufficient statistic.*

极差在这里可以被视为一个辅助统计量，它提供了关于数据的范围和离散程度的信息，其中包括均值 μ 的一些信息。但是即使我们知道了样本的极差，样本仍然是来自同一正态分布，说明极差不改变样本的分布。

5.3 Complete Statistics

定义 5.3.1. 完全统计量

统计量 $T(X)$ 被称为完全统计量，如果对于所有可测函数 g 和所有参数 $\theta \in \Theta$，

$$E_\theta[g(T(X))] = 0 \Rightarrow P_\theta(g(T(X)) = 0) = 1$$

，即 $g(T(X))$ 几乎处处为零。

例 5.3.2. 均匀分布的完全统计量

设 $X_1, X_2, \ldots, X_n$ 是来自 $U(0,\theta)$ 的样本，其中 $\theta > 0$ 是参数。则 $T = X_{(n)}$ 是 θ 的完全统计量。

证明. 固定θ，则$X_{(n)}$的概率密度函数为

$$f_\theta(x) = \frac{n}{\theta^n} x^{n-1} I(0 \le x \le \theta)$$

我们考虑任意可测函数 g，并设：$E_\theta[g(X_{(n)})] = 0$ 因为 $X_{(n)}$ 是 $U(0,\theta)$ 中的最大值，其累积分布函数为：

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(\max(X_1, X_2, \ldots, X_n) \le x) = \left(\frac{x}{\theta}\right)^n, \quad 0 \le x \le \theta$$

假设 $E_\theta[g(X_{(n)})] = 0$ 对于所有 θ 成立，那么我们有：

$$\int_0^\theta g(x) n \frac{x^{n-1}}{\theta^n}\, dx = 0$$

通过变换变量 $u = \frac{x}{\theta}$，我们得到：

$$\int_0^1 g(u\theta) n u^{n-1}\, du = 0$$

由于这个积分对所有 θ 都为零，特别地，对于 $\theta = 1$，我们有：

$$\int_0^1 g(u) n u^{n-1} \, du = 0$$

由于 u^{n-1} 在 $[0,1]$ 上是非零的，唯一的可能性是 $g(u) = 0$ 几乎处处为零。因此，$g(X_{(n)}) = 0$ 几乎处处为零，说明 $X_{(n)}$ 是完备统计量。

因此，样本的最大值 $X_{(n)}$ 是 $U(0,\theta)$ 的完备统计量。 □

定理 5.3.3. 完备性

如果极小充分统计量存在，且 $T(X)$ 是完备且充分的，则 $T(X)$ 是极小充分的。

定理 5.3.4. 指数族分布的完备性

如果指数族满秩，那么自然生成空间包含一个开集，那么他的最小充分统计量$T(x) = (T_1(X), \cdots, T_n(X))$是完备的。

定理 5.3.5. *If T is a boundedly complete sufficient statistic for the family of distributions P_θ, then any ancillary statistic V is independent of T.*

推论 5.3.6. *For a statistic T, If a non-constant function of T, say $R(T)$, is ancillary, then T cannot be complete.*

例 5.3.7. 设$X_1, \cdots, X_n \overset{i.i.d}{\sim} U(\theta - 1/2, \theta + 1/2)$由前面已知$T = (X_{(1)}, X_{(n)})$是极小充分统计量。

但是$R = X_{(n)} - X_{(1)}$是一个辅助统计量，因为他不依赖于θ，因此T不是完备的，因为T的分布不是唯一的。

Lec 6 Point Estimation

常见的点估计方法包括最大似然估计和最小方差无偏估计等

6.1 Estimation principle

一些基础的定义介绍

定义 6.1.1. 估计

设$X = (X_1, X_2, \cdots, X_n)$是来自总体$F_\theta(x)$的观测向量，$\theta \in \Theta \subset R^d$，那么任意一个统计量$\hat{g}_n(X) = \hat{g}(X_1, \cdots, X_n)$都可以视为$g(\theta)$的一个估计

注记. 估计是一个随机变量，因为它是由样本得到的函数。任意一个统计量都可以视为一个估计，那么这些估计是显然不同的，我们需要一个标准来衡量估计的好坏。我们认为一个估计是更好的，当$\hat{\theta}_n(X)$更靠近θ。

定义 6.1.2. 无偏估计量

设$\hat{\theta}_n(X)$是θ的估计，如果对于$\forall\theta, E_\theta[\hat{\theta}_n(X)] = \theta$，则称$\hat{\theta}_n(X)$是$\theta$的无偏估计量。

还有渐进无偏性，即$\lim_{n\to\infty} \hat{\theta}_n(X) = \theta$

例 6.1.3. 设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$，那么$\bar{X}$是$\mu$的无偏估计量，$S^2$是$\sigma^2$的无偏估计量。

定义 6.1.4. 有效估计量

设$\hat{\theta}_n(X)$是θ的估计，如果对于$\forall\theta$，$Var_\theta(\hat{\theta}_n(X))$最小，则称$\hat{\theta}_n(X)$是$\theta$的有效估计量。

也可以说，当g,h是θ的估计，如果$Var_\theta(g) \leq Var_\theta(h)$，则称$g$比$h$更有效。

例 6.1.5. 设总体分布$X \sim P(\lambda)$，注意到$E(X) = \lambda$，$Var(X) = \lambda$，计算$\hat{\theta}_1(X) = X$和$\hat{\theta}_2(X) = \bar{X}$的方差

$$Var(\hat{\theta}_1(X)) = Var(X) = \lambda, Var(\hat{\theta}_2(X)) = Var(\bar{X}) = \frac{\lambda}{n}$$

因此$\hat{\theta}_2(X)$更有效。

定义 6.1.6. 相合估计量

设$\hat{\theta}_n(X)$是θ的估计，如果对于$\forall\theta$，$\hat{\theta}_n(X) \xrightarrow{P} \theta$，则称$\hat{\theta}_n(X)$是$\theta$的相合估计量。

强相合估计量是指$\hat{\theta}_n(X) \xrightarrow{a.s} \theta$

定义 6.1.7. 渐进正态性

设$\hat{\theta}_n(X)$是θ的估计，如果对于$\forall\theta$，$B_n^{-1}(\hat{\theta}_n(X)-\theta) \xrightarrow{d} N(0,I)$，则称$\hat{\theta}_n(X)$是$\theta$的渐进正态估计量。

6.2 Method of Moments

定义 6.2.1. 矩估计

如果$g(\theta) = G(\alpha_1, \cdots, \alpha_k; \mu_2, \cdots, \mu_s)$，*i.e.*为原点矩和中心矩的函数，那么我们称$g(\theta)$的矩估计为

$$\hat{g}_n(X) = G(a_n1, \cdots, a_nk; m_n2, \cdots, m_ns)$$

其中$a_n1, \cdots, a_nk; m_n2, \cdots, m_ns$为样本矩

例 6.2.2. 设X服从*Maxwell*分布，其概率密度函数为

$$f(x;\theta)=\frac{x^2}{\theta^3}e^{-x^2/(2\theta)}$$

X的原点矩为$E(X)=\theta$

$g(\theta)=G(\alpha_1)=1/\alpha_1$，那么$\hat{g}_n(X)=1/\bar{X}$是$1/\theta$的矩估计

定理 6.2.3. 矩估计的相合性

设$\hat{g}_n(X)=G(a_{n1},\cdots,a_{nk};m_{n2},\cdots,m_{ns})$是$g(\theta)$的矩估计，如果$g(\theta)$是连续的，那么$\hat{g}_n(X)$是$g(\theta)$的相合估计

定理 6.2.4. 矩估计的渐进正态性

设$\hat{g}_n(X)=G(a_{n1},\cdots,a_{nk};m_{n2},\cdots,m_{ns})$是$g(\theta)$的矩估计，如果$g(\theta)$是一阶连续的，那么$\hat{g}_n(X)$是$g(\theta)$的渐进正态估计

例 6.2.5. 均匀分布的矩估计

设$X_1,X_2,\cdots,X_n\overset{i.i.d}{\sim}U(a,b)$

有两个未知量，我们需要列出一阶矩和二阶矩，$\alpha_1=E(X)=\frac{a+b}{2},\mu_2=Var(X)=\frac{(b-a)^2}{12}$

解得

$$\begin{cases}a=\alpha_1-\sqrt{3\mu 2}\\ \alpha_1+\sqrt{3\mu 2}\end{cases}$$

得a,b的矩估计为$\bar{X}-\sqrt{3S^2},\bar{X}+\sqrt{3S^2}$

注记. 此处说明了多变量的情况

6.3 Maximum Likelihood Estimation

求解极大似然函数是重点

定义 6.3.1. 似然函数

设$X = (X_1, X_2, \cdots, X_n)$是来自总体$f(x, \theta)$的观测向量，当$X = x$已经由观测结果确定，那么将有

$$L_n(\theta) = L(\theta; x) = f(x, \theta)$$

称之为似然函数

对每个样本点，我们都有一个似然函数，我们可以定义该点的极大似然估计量为

$$\hat{\theta}_n(x) = \arg\max_{\theta \in \Theta} L_n(\theta)$$

例 6.3.2. 求估计量，验证无偏性、有效性

设X为从总体中抽取的样本，总体的概率分布为

X	0	1	2	3
P	$\theta/2$	θ	$3\theta/2$	$1-3\theta$

其中抽取的样本为$\mathbf{x} = (0, 3, 1, 1, 0, 2, 2, 2, 3, 2)$依次求矩估计量$\hat{\theta}_M(x)$和最大似然估计量$\hat{\theta}_L(x)$并作无偏性和有效性检验。

求解估计量

矩估计：由一阶矩的表达式$E(X) = 3 - 5\theta$,求得$\theta(\alpha_1) = \theta = \frac{3-E(X)}{5} = \frac{3-\alpha_1}{5}$

解得$\hat{\theta}_M(\mathbf{x}) = \frac{3-\bar{x}}{5} = 0.28$

极大似然估计：似然函数为$L(\theta; x) = (\theta/2)^2(\theta)^2(3\theta/2)^4(1-3\theta)^2$

取对数得

$$\begin{aligned}\log L(\theta; x) =& x_1 \log(\theta/2) + x_2 \log(\theta) + x_3 \log(3\theta/2) + x_4 \log(1-3\theta) \\ =& (x_1 + x_2 + x_3) log(\theta) + x_4 \log(1-3\theta) + C_{x_1, x_2, x_3, x_4}\end{aligned}$$

对θ求导得

$$\frac{x_1 + x_2 + x_3}{\theta} + \frac{-3x_4}{1-3\theta} = 0$$

解得

$$\hat{\theta}_L(x) = \frac{x_1 + x_2 + x_3}{3n} \approx 0.26$$

验证无偏性

矩估计

$$E(\hat{\theta}_M(X)) = E(\frac{3 - \bar{X}}{5}) = \frac{3 - E(\bar{X})}{5} = \frac{3 - E(X)}{5} = \theta$$

极大似然估计

$$\begin{aligned} E(\hat{\theta}_L(X)) &= \sum_{t=0}^{n} P(\hat{\theta}_L(X) = \frac{n-t}{3n}) \frac{n-t}{3n} \\ &= \sum_{t=0}^{n} C_n^t (1-3\theta)^t (3\theta)^{n-t} \frac{n-t}{3n} \\ &= \sum_{t=0}^{n-1} C_{n-1}^t (1-3\theta)^t (3\theta)^{n-t-1} \frac{3\theta n}{3n} \\ &= \theta \end{aligned}$$

比较有效性

矩估计

$$D(\hat{\theta}_M(X)) = E(X^2) - (E(X))^2 = (10\theta - 25\theta^2) \in [0, 1]$$

$$D(\hat{\theta}_M(X)) = \frac{3 - D(\bar{X})}{5} = \frac{3n - D(X)}{5n} \in [0.58, 0.6]$$

极大似然估计

$$\begin{aligned} E((\hat{\theta}_L(X))^2) &= \sum_{t=0}^{n} C_n^t (1-3\theta)^t (3\theta)^{n-t} (\frac{n-t}{3n})^2 \\ &= \frac{\theta}{9} \sum_{t=0}^{n-1} C_{n-1}^t (1-3\theta)^t (3\theta)^{n-t-1} \frac{n-t}{n} \\ &= \frac{\theta}{3} (\sum_{t=0}^{n-1} C_{n-1}^t (1-\theta)^t (3\theta)^{n-t} - \frac{n-1}{n} \sum_{t=0}^{n-1} C_{n-1}^t (1-\theta)^t (3\theta)^{n-t} \frac{t}{n-1}) \\ &= \frac{\theta(3n\theta + 1 - 3\theta)}{3n} \end{aligned}$$

因此

$$\begin{aligned} D(\hat{\theta}_L(X)) &= E((\hat{\theta}_L(X))^2) - (E(\hat{\theta}_L(X)))^2 \\ &= \frac{\theta(3n\theta + 1 - 3\theta)}{3n} - \theta^2 \\ &= \frac{\theta - 3\theta^2}{3n} \in [0, 0.0028] \end{aligned}$$

极大似然函数更有效

注记. 此例不仅说明了极大似然函数的有效性很强，也给出了计算极大似然函数以及计算有效性的方式，下面将用更多例题来熟悉这一过程。

例 6.3.3. 设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} U(0, \theta)$，求$\theta$的极大似然估计量

解. 似然函数为

$$L(\theta; x) = \frac{1}{\theta^n} I(\theta \geq \max x_i)$$

由于$\theta \geq \max x_i$，因此$\hat{\theta}_L(\mathbf{X}) = X_{(n)}$

有偏：

$$E(\hat{\theta}_L(X)) = E(X_{(n)}) \overset{4.2}{=} \int_0^\theta \frac{nx^{n-1}}{\theta^n} x dx = \frac{n}{n+1}\theta$$

如果为了修正有偏性，我们可以取

$$\hat{\theta}'_L(\mathbf{X}) = \frac{n+1}{n}\hat{\theta}_L(\mathbf{X})$$

□

当给出了观测向量$\mathbf{X}$被映射φ转换后的实际观测向量，我们仍然能求得参数的极大似然估计，这里我们给出一个例题。此例还说明了参数θ的函数$g(\theta)$的极大似然估计如何求解。

例 6.3.4. 指数分布的极大似然估计

设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} Exp(\lambda)$

密度函数为

$$f(x;\lambda) = \lambda e^{-\lambda x} I(x > 0)$$

我们仅观测到了$X_{(1)}, X_{(2)}, \cdots, X_{(r)}$，求$\lambda$和$g(\lambda) = 1/\lambda$的极大似然估计量

解. 为方便表述，设$t_i = x_{(i)}, i = 1, 2, \cdots, n$，则$t_1 < t_2 < \cdots < t_r$

$$\begin{aligned} p(t_1, t_2, \cdots, t_r; \lambda) &= \int\limits_{-t_r < t_{r+1} < \cdots < t_n < \infty} \cdots \int n! f(t_1;\lambda) f(t_2;\lambda) \cdots f(t_r;\lambda) d_{r+1} \cdots d_n \\ &= \frac{n!}{(n-r)!} f(t_1;\lambda) f(t_2;\lambda) \cdots f(t_r;\lambda) [1 - F(t_r;\lambda)]^{n-r} \\ &= \frac{n!}{(n-r)!} \lambda^r \exp(-\lambda(\sum_{i=1}^{r} t_i + (n-r)t_r)) \end{aligned}$$

记$T = \sum_{i=1}^{r} t_i + (n-r)t_r$，则似然函数为

$$L(\lambda; t_1, t_2, \cdots, t_r) = \frac{n!}{(n-r)!} \lambda^r e^{-\lambda T}$$

求对数后求导再取零解λ得

$$\hat{\lambda}^* = \frac{r}{T} = \frac{r}{\sum_{i=1}^{r} t_i + (n-r)t_r}$$

$$\hat{g}(\lambda^*) = \frac{1}{\hat{\lambda}^*} = \frac{\sum_{i=1}^{r} t_i + (n-r)t_r}{r}$$

□

注记. 此例给出了观测向量$\mathbf{X}$被映射φ转换的情况，我们在已知转换后的向量（即便维数降低，信息丢失）的情况下，可以仍然可以得到对参数的估计，当然如果映射φ过于复杂，求联合分布等过程可能会极其困难。

从这里我们整理一下极大似然估计的步骤：先求观测数据的联合分布，然后求对数似然函数，求导数，令导数为零，解得极大似然估计量。

我们还可以用类似的方法求一般指数族分布的极大似然估计量

6.4 EM Algorithm*

EM算法是一种迭代算法，用于估计包含隐变量的概率模型的参数。

In many problems, MLE based on observed data X would be greatly simplified if we had additionally observed another piece of data Y. Y is called the hidden or latent data.

Let $\ell(\theta) = \log f(x|\theta)$ be the observed log-likelihood and also define the complete data log-likelihood:

$$\ell_c(\theta) = \log f(x, y|\theta) = \log f(y|x, \theta) f(x|\theta) = \log f(y|x, \theta) + \log f(x|\theta) = \log f(y|x, \theta) + \ell(\theta)$$

Suppose our current guess of θ is $\theta^{(t)}$ and that we would like to improve this guess. Consider

$$\ell(\theta) - \ell(\theta^{(t)}) = \ell_c(\theta) - \ell_c(\theta^{(t)}) + \log \frac{f(y|x, \theta^{(t)})}{f(y|x, \theta)}$$

Now take the expectation of both sides with respect to $y \sim f(y|x, \theta^{(t)})$, we have:

$$\ell(\theta) - \ell(\theta^{(t)}) = E_y[\ell_c(\theta)] - E_y[\ell_c(\theta^{(t)})] + D(f(y|x, \theta^{(t)}) \| f(y|x, \theta))$$

Since $D(f(y|x, \theta^{(t)}) \| f(y|x, \theta)) \geq 0$, we have the following inequality:

$$\ell(\theta) - \ell(\theta^{(t)}) \geq E_y[\ell_c(\theta)] - E_y[\ell_c(\theta^{(t)})] = Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})$$

where

$$Q(\theta, \theta') := E_{f(y|x,\theta')}[\log f(x, y|\theta)]$$

is the expectation of the complete data log-likelihood.

We choose $\theta^{(t+1)}$ as the solution of the following optimization problem:

$$\theta^{(t+1)} = \text{argmax}_\theta Q(\theta, \theta^{(t)})$$

The process of EM algorithm is as follows:

- Init: $t = 0, \theta^{(0)} = 0$ or random value.
- E step: $Q(\theta, \theta^{(t)}) = E_{f(y|x,\theta^{(t)})}[\log f(x, y|\theta)]$
- M step: $\theta^{(t+1)} = \text{argmax}_\theta Q(\theta, \theta^{(t)})$

The E-step and M-step repeat until convergence.

The EM algorithm is an attractive option if the Q function is easily computed and optimized. The relationship between $\log f(x, \theta)$, $Q(\theta, \theta^{(t)})$, θ_t and $\theta^{(t+1)}$ are depicted in the following figure:

注记. 在许多问题中，基于观测数据 X 的MLE计算可以通过假设另外观测到一部分数据 Y 而简化。这个 Y 称为隐藏数据或潜在数据。

定义对数似然函数

观测数据的对数似然函数定义为$\ell(\theta) = \log f(x|\theta)$

而完全数据的对数似然函数定义为$\ell_c(\theta) = \log f(x, y|\theta)$

根据概率论，完全数据的对数似然可以拆分为两个部分：

$$\ell_c(\theta) = \log f(x, y|\theta) = \log f(y|x, \theta) f(x|\theta) = \log f(y|x, \theta) + \log f(x|\theta) = \log f(y|x, \theta) + \ell(\theta)$$

迭代更新参数 假设当前参数估计为 $\theta^{(t)}$，希望改进这个估计。考虑以下等式：

$$\ell(\theta) - \ell(\theta^{(t)}) = \ell_c(\theta) - \ell_c(\theta^{(t)}) + \log \frac{f(y|x, \theta^{(t)})}{f(y|x, \theta)}$$

期望步骤（E-step） 对等式两边取关于 $y \sim f(y|x, \theta^{(t)})$ 的期望：

$$\ell(\theta) - \ell(\theta^{(t)}) = E_y[\ell_c(\theta)] - E_y[\ell_c(\theta^{(t)})] + D(f(y|x, \theta^{(t)}) \| f(y|x, \theta))$$

其中 $D(f(y|x, \theta^{(t)}) \| f(y|x, \theta)) \geq 0$，因此有：

$$\ell(\theta) - \ell(\theta^{(t)}) \geq E_y[\ell_c(\theta)] - E_y[\ell_c(\theta^{(t)})] = Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})$$

这里定义

$$Q(\theta, \theta') := E_{f(y|x,\theta')}[\log f(x, y|\theta)]$$

最大化步骤（M-step） 选择 $\theta^{(t+1)}$ 作为以下优化问题的解：

$$\theta^{(t+1)} = \text{argmax}_\theta Q(\theta, \theta^{(t)})$$

EM算法的过程 EM算法通过以下步骤进行迭代，直到收敛：

- 初始化： $t = 0, \theta^{(0)} = 0$ 或随机值。
- E步：计算 $Q(\theta, \theta^{(t)}) = E_{f(y|x,\theta^{(t)})}[\log f(x, y|\theta)]$。
- M步：最大化 $Q(\theta, \theta^{(t)})$，得到新的参数估计 $\theta^{(t+1)}$。

由此我们给出了EM算法的操作步骤。如果 Q 函数容易计算和优化，那么EM算法是一个很好的选择。

EM算法的好处是，每一次迭代，对数似然值都会增加，因此EM算法是收敛的。

例 6.4.1. *Example: Mixture of Normals*

Suppose:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \sum_{j=1}^{m} \pi_j N(\mu_j, \sigma_j^2), \sum \pi_j = 1, 0 < \pi_j < 1$$

Find the MLE $(\hat{\mu}_j, \hat{\sigma}_j^2), j = 1, \ldots, m.$

Let $y_1, \ldots, y_n$ *be the latent member index, then*

$$f(x, y|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left\{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right\} I(y_i = j)$$

Thus,

$$\log f(x, y|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log\left(\pi_j \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left\{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right\}\right) I(y_i = j)$$

and

$$E_{f(y|x,\theta^{(t)})}[\log f(x, y|\theta)] = \sum_{i=1}^{n} \sum_{j=1}^{m} \log\left(\pi_j \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left\{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right\}\right) E_{f(y|x,\theta^{(t)})}[I(y_i = j)]$$

Denote

$$f^{(t)}(y_i = j) = \frac{\pi_j^{(t)} N(x_i; \mu_j^{(t)}, (\sigma_j^{(t)})^2)}{\sum_{l=1}^{m} \pi_l^{(t)} N(x_i; \mu_l^{(t)}, (\sigma_l^{(t)})^2)},$$

we have the expression of $Q(\theta, \theta^{(t)})$*:*

$$Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} f^{(t)}(y_i = j)\log(\pi_j N(x_i; \mu_j, \sigma_j^2)) = \sum_{i=1}^{n}\sum_{j=1}^{m} f^{(t)}(y_i = j)\log(N(x_i; \mu_j, \sigma_j^2)) + const$$

Set $\frac{\partial Q}{\partial \theta} = 0$*, we have:*

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} f^{(t)}(y_i = j) x_i}{\sum_{i=1}^{n} f^{(t)}(y_i = j)}$$

$$(\sigma_j^{(t+1)})^2 = \frac{\sum_{i=1}^{n} (x_i - \mu_j^{(t+1)})^2 f^{(t)}(y_i = j)}{\sum_{i=1}^{n} f^{(t)}(y_i = j)}$$

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} f^{(t)}(y_i = j)}{\sum_{j=1}^{m}\sum_{i=1}^{n} f^{(t)}(y_i = j)}$$

Using EM algorithm, fit a mixture of two normal distributions to the waiting time of Faithful data in R.

注记. 这个例子说明了如何使用对比函数和*EM*算法来构建估计方程，求解混合正态分布的参数的最大似然估计，并证明了估计方程估计的无偏性。具体步骤如下：

1. **模型假设：**给定独立同分布的数据 $X_1, X_2, \ldots, X_n$ 来自一个包含 m 个成分的混合正态分布：

$$X_i \overset{iid}{\sim} \sum_{j=1}^{m} \pi_j N(\mu_j, \sigma_j^2), \quad \sum \pi_j = 1, \quad 0 < \pi_j < 1$$

目标是找到每个成分的参数的最大似然估计 $(\hat{\mu}_j, \hat{\sigma}_j^2), j = 1, \ldots, m$。

2. **引入潜在变量：**定义潜在的成员索引 $y_1, \ldots, y_n$，表示每个观测值属于哪个成分。

3. **对数似然函数：**通过潜在变量构建完全数据的对数似然函数：

$$\log f(x, y|\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log\left(\pi_j \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left\{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right\}\right) I(y_i = j)$$

4. ***(E-step)*****：计算在当前参数估计下的期望对数似然函数：**

$$E_{f(y|x,\theta^{(t)})}[\log f(x,y|\theta)] = \sum_{i=1}^{n}\sum_{j=1}^{m}\log\left(\pi_j \frac{1}{\sqrt{2\pi\sigma_j}}\exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\}\right) E_{f(y|x,\theta^{(t)})}[I(y_i=j)]$$

5. ***(M-step)*****：通过最大化上一步得到的期望对数似然函数，更新参数：**

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} f^{(t)}(y_i=j)x_i}{\sum_{i=1}^{n} f^{(t)}(y_i=j)}$$

$$(\sigma_j^{(t+1)})^2 = \frac{\sum_{i=1}^{n}(x_i-\mu_j^{(t+1)})^2 f^{(t)}(y_i=j)}{\sum_{i=1}^{n} f^{(t)}(y_i=j)}$$

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} f^{(t)}(y_i=j)}{\sum_{j=1}^{m}\sum_{i=1}^{n} f^{(t)}(y_i=j)}$$

6.5 Properties of MLEs*

MLE的性质十分良好，我们依次说明MLE的映射不变性，充分性，相合性，渐进正态性，以及一些渐进有效性。

定理 6.5.1. *Invariance Principle of MLEs*

设$\hat{\theta}$是θ的MLE，当g为一个定义在Θ上的函数时，那么$g(\hat{\theta})$是$g(\theta)$的MLE。

证明. Define a new parameter $\phi = g(\theta)$, and suppose $\theta = h(\phi)$ is the inverse function of $g(\theta)$ (if it exists). Then

$$\hat{\phi} = g(\hat{\theta}).$$

The likelihood function in terms of ϕ is $L(h(\phi); X)$.

Since $\hat{\theta}$ is the MLE of θ, we have

$$L(\hat{\theta}; X) \geq L(\theta; X) \quad \text{for all } \theta \in \Theta.$$

Substituting $\theta = h(\phi)$, we get

$$L(h(g(\hat{\theta})); X) \geq L(h(\phi); X) \quad \text{for all } \phi \in g(\Theta).$$

Since $h(g(\hat{\theta})) = \hat{\theta}$, this simplifies to

$$L(\hat{\theta}; X) \geq L(h(\phi); X) \quad \text{for all } \phi \in g(\Theta).$$

Therefore, $\hat{\phi} = g(\hat{\theta})$ maximizes the likelihood function $L(h(\phi); X)$, which means

$$L(h(\hat{\phi}); X) \geq L(h(\phi); X) \quad \text{for all } \phi \in g(\Theta).$$

Thus, $\hat{\phi} = g(\hat{\theta})$ is the MLE of $\phi = g(\theta)$. This proves the invariance principle of MLEs.

□

定理 6.5.2. *Existence and uniqueness*

Let $X_n = (X_1, \ldots, X_n)$ be a simple sample from a population with density $f(x, \theta)$, $\theta \in \Theta$ open $\subset \mathbb{R}^d$. If further $\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(x_i, \theta)$ is strictly concave and $\ell_n(\theta) \to -\infty$ as $\theta \to \partial\Theta$ (the boundary of Θ), then the MLE $\hat{\theta}(x_n)$ exists and is unique.

注记. 这个定理比较复杂，从知识体系上比较重要，此处给出此定理的详细解释（但如果只是为了复习，知道存在性和唯一性的含义是什么，那么以下部分可以跳过）：

1. **样本和总体背景：** 设 $X_n = (X_1, \ldots, X_n)$ 是从一个总体中抽取的简单随机样本。这个总体的密度函数是 $f(x, \theta)$，其中 θ 是参数，取值范围在开集 $\Theta \subset \mathbb{R}^d$。

2. **对数似然函数：** $\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(x_i, \theta)$ 是参数 θ 的对数似然函数的平均值。

3. **严格凹性：** 假设对数似然函数 $\ell_n(\theta)$ 是严格凹的，这意味着对于任何两个不同的参数值 θ_1 和 θ_2，以及对任意 $\alpha \in (0, 1)$，都有：

$$\ell_n(\alpha\theta_1 + (1-\alpha)\theta_2) > \alpha\ell_n(\theta_1) + (1-\alpha)\ell_n(\theta_2)$$

严格凹性确保了对数似然函数没有多个局部最大值。

4. **边界行为：**参数 θ 趋向于参数空间的边界（$\partial\Theta$）时，对数似然函数 $\ell_n(\theta)$ 趋向于负无穷。

5. **最大似然估计（*MLE*）的存在性和唯一性：**基于上述条件，定理保证了最大似然估计 $\hat{\theta}(x_n)$ 存在且唯一。这意味着在给定的样本下，能够找到一个唯一的参数值 $\hat{\theta}(x_n)$，使得对数似然函数 $\ell_n(\theta)$ 达到最大值。

定理 6.5.3. 指数族的极大似然估计的存在性和唯一性

设 $X_1, X_2, \ldots, X_n$ 是取自指数族的样本，*i.e.*密度函数为

$$f(x;\theta) = h(x)\exp(\theta' T(x) - A(\theta)), for x \in \mathcal{X}$$

如果

$$\frac{\partial A(\theta)}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$$

在Θ^o内存在一解，那么该解即为θ的MLE且唯一。

注记. 这个定理是很自然的，证明起来也不是很困难。在这里给出一个例子替代证明。

例 6.5.4. 指数分布的MLE

设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} Exp(\lambda)$，求$\lambda$的$MLE$

解. 一般的推断

似然函数为

$$L(\lambda; x) = \lambda^n e^{-\lambda\sum_{i=1}^{n} x_i}$$

对数似然函数为

$$\ell(\lambda; x) = n\ln\lambda - \lambda\sum_{i=1}^{n} x_i$$

求导得

$$\frac{\partial \ell(\lambda; x)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

解得

$$\hat{\lambda}^* = \frac{n}{\sum_{i=1}^{n} x_i}$$

□

解. 使用定理

指数族的密度函数为

$$f(x; \lambda) = \lambda e^{-\lambda x}$$

方程为

$$\frac{\partial - In(\lambda)}{\partial \lambda} = \frac{1}{n} \sum_{i=1}^{n} -X_i$$

解得

$$\hat{\lambda}^* = \frac{n}{\sum_{i=1}^{n} x_i}$$

□

随感. 这两节已经给出太多极大求解似然函数的例子了，因此这里也省略了大量的步骤和说明。如果看不懂可以再重复看看前面的例子。

接下来给出MLE的相合性和渐进正态性的定理。

定理 6.5.5. *MLE consistency*

Let $X = (X_1, \ldots, X_n)$ be a random sample from a population with density $f(x, \theta)$, $\theta \in \Theta$ open $\subset \mathbb{R}^d$. Assume that

1. *θ is identifiable, i.e., for any $\theta' \neq \theta$, there exists x such that $f(x, \theta') \neq f(x, \theta)$;*

2. *the support of* $f(\cdot, \theta)$ *does not depend on* θ*, and* $f(x, \theta)$ *is differentiable with respect to* θ *in* Θ*; and*

3. *the true value* θ_0 *is an interior point of parameter space* Θ*.*

Then the likelihood has a maximizer $\hat{\theta}$ *and* $\hat{\theta} \xrightarrow{p} \theta_0$*.*

定理 6.5.6. *MLE asymptotic normality*

In the setting above, assume that conditions (1)-(3) in the MLE consistency theorem hold. In addition, assume that

1. $f(x, \theta)$ *is three times differentiable with respect to* θ *and we can interchange integration with respect to* x*; and*

2. $\frac{\partial^3 \log f(x,\theta)}{\partial \theta^3} \leq M(x)$ *and* $\mathbb{E}[M(X_i)] < \infty$*.*

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}(\theta_0)\right)$$

in distribution. Denote by $\ell_n(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta)$*.*

除了极大似然估计之外，还有着其他似然估计，列举定义如下：

定义 6.5.7. *Likelihood*

Maximum Likelihood (ML): *Given a statistical model* $\mathcal{M}$ *parameterized by* θ*, the maximum likelihood estimator* $\hat{\theta}_{ML}$ *is defined as*

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i; \theta),$$

where $f(x_i; \theta)$ *is the probability density function (or mass function) of the model given the observed data* $x_1, x_2, \ldots, x_n$*.*

Profile Likelihood: *For a parameter* θ *of interest, the profile likelihood function* $L_p(\theta)$ *is defined as*

$$L_p(\theta) = \max_{\theta_{nuisance}} \mathcal{L}(\theta, \theta_{nuisance}),$$

where $\mathcal{L}(\theta, \theta_{nuisance})$ is the likelihood function, and $\theta_{nuisance}$ represents the nuisance parameters.

Partial Likelihood: *In the context of censored or truncated data, the partial likelihood function is used. For an event time t_i and a censoring indicator δ_i, the partial likelihood function is defined as*

$$L_p(\theta) = \prod_{i=1}^{n} \frac{f(t_i;\theta)}{\sum_{j:t_j \geq t_i} f(t_j;\theta)},$$

where $f(t_i;\theta)$ is the probability density function of the event time.

Induced Likelihood: *In certain situations, an induced likelihood is considered, especially when the likelihood function is not directly available. It's defined as*

$$L_{ind}(\theta) = \sup_{\mathcal{M}_\theta} \mathbb{P}_\theta(X = x),$$

where $\mathcal{M}_\theta$ represents the family of probability distributions under consideration and $\mathbb{P}_\theta(X = x)$ is the probability of observing the data x given the parameter θ.

6.6 Minimal contrast estimation and estimating equations*

此节给出了极大似然由来的动机，极大似然估计实际上是一个最小对比估计。

定义 6.6.1. *Minimal Contrast Estimator*

Contrast function:

$$\rho : \mathcal{X} \times \Theta \to \mathbb{R}$$

and define $D(\theta_0, \theta) \equiv \mathbb{E}_{\theta_0}\rho(X, \theta)$

Let $X_i \overset{i.i.d}{\sim} p_{\theta_0}$. *Furthermore, let* ρ *be a real function such that*

$$D(\theta_0, \theta) > D(\theta_0, \theta_0) \forall \theta \neq \theta_0$$

then,

$$\hat{\theta} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, \theta)$$

is the minimal contrast estimator (MCE) of θ.

这一部分我没有完全搞懂，因此我用一段课件原文一段自己理解的方式来解释。

Usually, $Q_n(X, \theta) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \theta) \to D(\theta_0, \theta)$ as $n \to \infty$. It is natural to consider $\hat{\theta}(X)$ minimizing $D(\theta_0, \theta)$.

注记. 样本平均函数 $Q_n(X, \theta)$ 随着样本数量 n 的增加而收敛到一个期望值 $D(\theta_0, \theta)$。这个期望值 $D(\theta_0, \theta)$ 用来衡量估计值 θ 与真值 θ_0 之间的差异。希望通过最小化 $D(\theta_0, \theta)$ 来找到估计值 θ 。假设 θ_0 是参数空间 Θ 的内部点，并且 $D(\theta_0, \theta)$ 是光滑的函数。

Now suppose the true θ_0 is an interior point of Θ, and $\theta \to D(\theta_0, \theta)$ is smooth. Then we expect

$$\nabla_\theta D(\theta_0, \theta)\big|_{\theta=\theta_0} = 0$$

where ∇ denotes the gradient $\nabla_\theta = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_d}\right)'$.

注记. 期望 $D(\theta_0, \theta)$ 在 $\theta = \theta_0$ 处的梯度为零，这提供了一个用于估计的必要条件：

$$\nabla_\theta D(\theta_0, \theta)\big|_{\theta=\theta_0} = 0$$

梯度 ∇_θ 表示对各个参数的偏导数。

Arguing heuristically again we are led to estimates $\hat{\theta}$ that solve

$$\nabla_\theta Q_n(X, \hat{\theta}) = 0$$

which is a special form of the following estimating equations.

注记. 基于上述梯度条件，提出了一个特殊形式的估计方程：通过解 $\nabla_\theta Q_n(X,\hat{\theta})=0$ 来找到估计值 $\hat{\theta}$。

Estimation equation: More generally, suppose we are given a function $\Psi: X\times\mathbb{R}^d\to\mathbb{R}^d$, $\Psi\equiv(\psi_1,\ldots,\psi_d)^T$ and define

$$S(\theta_0,\theta)=E_{\theta_0}\Psi(X,\theta)$$

Suppose $S(\theta_0,\theta)=0$ has $\theta=\theta_0$ as its unique solution for all $\theta_0\in\Theta$.

Then we say $\hat{\theta}$ solving

$$\Psi(X,\hat{\theta})=0$$

is an estimating equation estimate.

注记. 更一般地，定义了估计方程的形式，假设给定一个函数 $\Psi: X\times\mathbb{R}^d\to\mathbb{R}^d$。定义 $S(\theta_0,\theta)=E_{\theta_0}\Psi(X,\theta)$，并假设 $S(\theta_0,\theta)=0$ 的唯一解是 θ_0。最终，提出通过解 $\Psi(X,\hat{\theta})=0$ 来找到估计值 $\hat{\theta}$，这被称为估计方程估计。

例 6.6.2. 最小二乘法

Let $(\mathbf{X}_i,Y_i)$ *be i.i.d. from*

$$Y_i=g(\mathbf{X}_i,\beta)+\epsilon_i=\mathbf{X}_i^T\beta+\epsilon_i,\quad \textit{if linear model}$$

By letting

$$\rho(\mathbf{X},\beta)=\sum_{i=1}^{n}[Y_i-g(\mathbf{X}_i,\beta)]^2$$

be a contrast function, we have

$$\begin{aligned}D(\beta_0,\beta)&=E_{\beta_0}\rho(\mathbf{X},\beta)=nE_{\beta_0}[Y-g(\mathbf{X},\beta)]^2\\&=nE\{g(\mathbf{X},\beta_0)-g(\mathbf{X},\beta)\}^2+n\sigma^2\end{aligned}$$

which is indeed minimized at $\beta = \beta_0$*. Hence, the minimum contrast estimator is*

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} [Y_i - g(\mathbf{X}_i, \beta)]^2 \quad \textit{(least-squares regression)}$$

It satisfies the system of equations

$$\sum_{i=1}^{n} \left(Y_i - g(\mathbf{X}_i, \hat{\beta}) \right) \frac{\partial g(\mathbf{X}_i, \beta)}{\partial \beta_j} = 0, \quad j = 1, \ldots, d,$$

under some mild regularity conditions. One can easily check that $\psi_j(\beta) = (Y_i - g(\mathbf{X}_i, \beta)) \frac{\partial g(\mathbf{X}_i, \beta)}{\partial \beta_j}$ *satisfies*

$$E_{\beta_0} \psi_j(\beta) \Big|_{\beta=\beta_0} = E\{g(\mathbf{X}_i, \beta_0) - g(\mathbf{X}_i, \beta_0)\} \frac{\partial g(\mathbf{X}_i, \beta_0)}{\partial \beta_j} = 0$$

Thus, it is also an estimator based on the estimating equations.

注记. 如果还是晕晕，不知道为什么这么对最小二乘的操作符合了上述的分析，那在此处给出更详细的解释，请对比着上述解释估计方程估计的动机来理解。

1. **定义对比函数：** 给定模型 $Y_i = g(\mathbf{X}_i, \beta) + \epsilon_i$，当模型是线性时，表示为 $Y_i = \mathbf{X}_i^T \beta + \epsilon_i$。选择对比函数为：

$$\rho(\mathbf{X}, \beta) = \sum_{i=1}^{n} [Y_i - g(\mathbf{X}_i, \beta)]^2$$

这个对比函数衡量了预测值 $g(\mathbf{X}_i, \beta)$ 与实际观测值 Y_i 之间的差异的平方和。

2. **构造期望对比函数：** 定义期望对比函数 $D(\beta_0, \beta)$ 为：

$$D(\beta_0, \beta) = E_{\beta_0} \rho(\mathbf{X}, \beta)$$

具体计算为：

$$D(\beta_0, \beta) = n E_{\beta_0} [Y - g(\mathbf{X}, \beta)]^2 = n E\{g(\mathbf{X}, \beta_0) - g(\mathbf{X}, \beta)\}^2 + n\sigma^2$$

这个公式表示在$\beta = \beta_0$处对比函数的期望值最小。

3. **最小对比估计量：**由上述期望对比函数的性质，我们定义最小对比估计量为：

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} [Y_i - g(\mathbf{X}_i, \beta)]^2$$

这实际上是最小二乘回归估计量。

4. **满足的方程系统：**该估计量满足以下方程系统：

$$\sum_{i=1}^{n} \left(Y_i - g(\mathbf{X}_i, \hat{\beta})\right) \frac{\partial g(\mathbf{X}_i, \beta)}{\partial \beta_j} = 0, \quad j = 1, \ldots, d,$$

在一定的正则条件下，这个方程系统成立。

5. **验证无偏性：**定义估计方程为：

$$\psi_j(\beta) = (Y_i - g(\mathbf{X}_i, \beta)) \frac{\partial g(\mathbf{X}_i, \beta)}{\partial \beta_j}$$

验证无偏性：

$$E_{\beta_0} \psi_j(\beta)\Big|_{\beta=\beta_0} = E\{g(\mathbf{X}_i, \beta_0) - g(\mathbf{X}_i, \beta_0)\} \frac{\partial g(\mathbf{X}_i, \beta_0)}{\partial \beta_j} = 0$$

由此证明了估计方程在$\beta = \beta_0$时的无偏性，因为上述期望值等于零。

此例给出了如何利用对比函数来构建估计方程，并证明了估计方程估计的无偏性。

例 6.6.3. *L_1-regression*

Consider

$$\rho(\mathbf{X}, Y, \beta) = |Y - \mathbf{X}^T \beta|.$$

Then,

$$D(\beta_0, \beta) = E_{\beta_0} |Y - \mathbf{X}^T \beta| = E|\mathbf{X}^T (\beta - \beta_0) + \epsilon|.$$

For any a, define

$$f(a) = E|\epsilon + a|.$$

Then,

$$f'(a) = E\, sgn(\epsilon + a) = P(\epsilon + a > 0) - P(\epsilon + a < 0) = 2P(\epsilon + a > 0) - 1.$$

If $med(\epsilon) = 0$*, then* $f'(0) = 0$*. In other words,* $f(a)$ *is minimized at* $a = 0$*, or* $D(\beta_0, \beta)$ *is minimized at* $\beta = \beta_0$*! Thus, if* $med(\epsilon) = 0$*, then*

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \mathbf{X}_i^T \beta|$$

is a minimum contrast estimator.

例 6.6.4. 极大似然估计

Let $L(\theta) = f(x, \theta)$ *be the likelihood function,* θ_0 *is the true parameter.*

Then, let $D(\theta_0, \theta) = -E_{\theta_0} \log f(X|\theta)$ *and to say that* $D(\theta_0, \theta)$ *is minimized uniquely when* $P_\theta = P_{\theta_0}$ *is equivalent to*

$$D(\theta_0, \theta) - D(\theta_0, \theta_0) = -(E_{\theta_0} \log f(X|\theta) - E_{\theta_0} \log f(X|\theta_0)) = -E_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)} > 0$$

unless $\theta = \theta_0$*. Here* $D(\theta_0, \theta_0)$ *is called the entropy of* X*. In the case of* $X_1, \ldots, X_n$ *i.i.d.*

$$\rho(X, \theta) = -\frac{1}{n} \sum \log f(X_i, \theta)$$

satisfies the condition of being a contrast function, and the MLE is a minimum contrast estimate.

Define the mutual entropy or Kullback-Leibler information divergence between $f(X|\theta_0)$ *and* $f(X|\theta)$ *by*

$$K(f(X|\theta_0), f(X|\theta)) = -E_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)}$$

Then maximizing likelihood is equivalent to minimizing Kullback-Leibler divergence, i.e., maximizing the likelihood of data is equal to minimizing the difference between the estimate and the real data distribution.

注记. 我查了以下资料，这里其实讲的是*Kullback-Leibler*散度，这是一个用来衡量两个分布之间差异的指标，这里的极大似然估计实际上是在最小化两个分布之间的差异，这个差异越小，说明估计的分布越接近真实分布，因此极大似然估计是一个最小对比估计。

同样的我们也给出详细的说明以辅助理解：

1. **定义对比函数** $D(\theta_0,\theta)$：对比函数 $D(\theta_0,\theta)$ 被定义为在参数 θ 下，对数似然函数的期望值与其在真实参数 θ_0 下的期望值之差，即

$$D(\theta_0,\theta) = -E_{\theta_0}\log f(X|\theta).$$

2. **对比函数的性质**：当 $\theta=\theta_0$ 时，对比函数 $D(\theta_0,\theta)$ 达到最小值。也就是说，对于所有 $\theta\neq\theta_0$，有

$$D(\theta_0,\theta)-D(\theta_0,\theta_0) = -(E_{\theta_0}\log f(X|\theta)-E_{\theta_0}\log f(X|\theta_0)) = -E_{\theta_0}\log\frac{f(X|\theta)}{f(X|\theta_0)} > 0.$$

这说明 $D(\theta_0,\theta)$ 在 $\theta=\theta_0$ 处取得唯一最小值。

3. **引入对比函数** $\rho(X,\theta)$：在样本 $X_1,\ldots,X_n$ 独立同分布的情况下，可以定义对比函数 $\rho(X,\theta)$ 为

$$\rho(X,\theta) = -\frac{1}{n}\sum\log f(X_i,\theta).$$

满足对比函数的条件，即对任意 $\theta\neq\theta_0$，有 $D(\theta_0,\theta) > D(\theta_0,\theta_0)$。

4. ***MLE* 是最小对比估计的证明**：根据定义，如果对比函数 $\rho(X,\theta)$ 满足上述条件，则最大似然估计可以被视为最小对比估计。

5. **互信息或*Kullback-Leibler*信息散度的定义**：定义 $f(X|\theta_0)$ 和 $f(X|\theta)$ 之间的互信息或*Kullback-Leibler*信息散度为

$$K(f(X|\theta_0),f(X|\theta)) = -E_{\theta_0}\log\frac{f(X|\theta)}{f(X|\theta_0)}.$$

由此我们得知了极大似然估计的动机，即最小化估计和真实数据分布之间的差异（衡量分布之间差异性用的是*Kullback-Leibler*散度），使得估计更接近真实分布。

Lec 7 Optimal Unbiased Estimation

7.1 Minimum variance unbiased estimator

定义 7.1.1. *Minimum Variance Unbiased Estimator*

设$\mathbf{X} = (X_1, X_2, \cdots, X_n)$是来自参数分布族$F_\theta, \theta \in \Theta$的一个样本，一个$g(\theta)$的无偏估计量$\hat{g}(\theta)$满足

$$Var_\theta(\hat{g}_n(\mathbf{X})) \leq Var_\theta(\hat{g}_n^*(\mathbf{X}))$$

对所有$\theta \in \Theta$成立，其中$\hat{g}_n^*(\mathbf{X})$是任意$g(\theta)$的无偏估计量。则称$\hat{g}_n(\mathbf{X})$是$g(\theta)$的最小方差无偏估计量（$MVUE$）。

定理 7.1.2. *Rao-Blackwell Theorem*

设$T(\mathbf{X})$是参数分布族$F_\theta, \theta \in \Theta$中$\theta$的充分统计量。如果$\hat{g}(\mathbf{X})$是$g(\theta)$的一个估计量，那么有以下结论：

1. $E(\hat{g}(\mathbf{X})|T) = h(T)$是一个统计量
2. 如果$\hat{g}(\mathbf{X})$是无偏的，则$h(T)$也是无偏的
3. $Var_\theta(\hat{g}(\mathbf{X})) \geq Var_\theta(h(T))\theta \in \Theta$ 等号成立当且仅当$\hat{g}((X)) = h(T(\mathbf{X}))a.s.P_\theta$成立。

注记. 这个定理说明了，如果我们有一个无偏估计，可以用其对某充分统计量取期望，这能导出一个新的无偏估计，且方差更小。

而且，当一个特殊的充分统计量使得无偏估计取期望的函数仍是他自身的时候，他就成为了*UMVUE*，我们将用若干例子说明这一定理。

例 7.1.3. 设$\mathbf{X} = (X_1, \cdots, X_n)$是从两点分布族$b(1,p) : 0 < p < 1$中获得的样本，依次说明$X_1$是$p$的一个无偏估计，$T(\mathbf{X}) = \sum_{i=1}^{n} X_i$是$p$的充分统计量，并以此构造一个比$X_1$方差更小的无偏估计。

解. X_1是p的一个无偏估计：$E(X_1) = EX = p$

$T(\mathbf{X}) = \sum_{i=1}^{n} X_i$是$p$的充分统计量(参考5.1.2)

构造新的无偏估计量

$$\begin{aligned} h(t) &= E(X_1|T=t) \\ &= 1P(X_1 = 1|T=t) + 0P(X_1 = 0|T=t) \\ &= \frac{P(X_1 = 1, T=t)}{P(T=t)} \\ &= \frac{p\binom{n-1}{t-1}p^{t-1}(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{t}{n} \end{aligned}$$

故 $h(T(\mathbf{X})) = \frac{\sum X_i}{n} = \bar{X}$ 为一个方差更小的无偏估计。

□

7.2 Lehmann-Scheffe theorem and Unbiased estimators of zero

定理 7.2.1. *Lehmann-Scheffe theorem*

若 $T(X)$ 是分布族$\{F_\theta, \theta \in \Theta\}$ 的一个完全充分统计量， $\hat{g}_n(T)$ 是$g(\theta)$的一个无偏估计。那么 $\hat{g}_n(T)$ 是$g(\theta)$唯一的 *UMVUE(a.s.)*。

注记. 这个定理说明我们只要有一个完全充分统计量$T(X)$ 和一个使其变为无偏估计的函数，那么这个完全充分统计量变为的无偏估计就是$UMVUE$,这里给出例子，注意这个$\hat{g}_n$是如何找到的。

例 7.2.2. 设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} U(0,\theta)$，求$\theta$的$MVUE$。

解. 由于 $T(\mathbf{X}) = X_{(n)}$ 是 θ 的完全充分统计量（参考5.1.6与5.3.2），我们可以考虑 $\dot{g}_n(T) = T = X_{(n)}$ 的修正，使得$\dot{g}_n$修正为无偏估计$\hat{g}_n$，为了知道如何修正，需要先求出$E(X_{(n)})$

首先，我们需要确定 $X_{(n)}$ 的分布函数和密度函数。

$X_{(n)}$ 的分布函数为：

$$F_{X_{(n)}}(x) = 1 - (1 - F(x))^n = 1 - (1 - \frac{x}{\theta})^n$$

然后，$X_{(n)}$ 的密度函数为：

$$f_{X_{(n)}}(x) = \frac{n}{\theta}\left(1 - \frac{x}{\theta}\right)^{n-1}$$

接下来，我们计算 $X_{(n)}$ 的期望值：

$$E(X_{(n)}) = \int_0^\theta x\frac{n}{\theta}\left(1 - \frac{x}{\theta}\right)^{n-1} dx$$

我们通过变量替换 $u = \frac{x}{\theta}$ 进行积分计算，得到：

$$E(X_{(n)}) = \theta \cdot \frac{n}{n+1}$$

因此，$\hat{g}_n(T) = \frac{n+1}{n}X_{(n)}$ 是 θ 的无偏估计量，并且因为 $X_{(n)}$ 是完全充分统计量，根据 Lehmann-Scheffé 定理，$\frac{n+1}{n}X_{(n)}$ 是 θ 的最小方差无偏估计量（UMVUE）。

□

例 7.2.3. 设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} Poi(\lambda)$，求出$(1)g_1(\lambda) = \lambda (2)g_2(\lambda) = \lambda^r, r > 0$为自然数$(3)g_3(\lambda) = P_\lambda(X_1 = x)$的$UMVUE$，给出求解。

解. $T=\sum_{i=1}^{n}X_i$是λ的充分完备统计量。

1. $E(\bar{X})=\lambda$，故$\bar{X}$是λ的无偏估计，由L-S定理知$h_1(T)=T/n$是λ的UMVUE。

2. $T\sim Poi(n\lambda)$,令$h_2(T)$为$g_2(\lambda)=\lambda^r$的无偏估计，故有$E_\lambda[h_2(T)]=\lambda^r$，即

$$\sum_{t=0}^{\infty}h_2(t)\frac{e^{-n\lambda}(n\lambda)^t}{t!}=\lambda^r\Leftrightarrow\sum_{t=0}^{\infty}h_2(t)\frac{n^t\lambda^t}{t!}=e^{n\lambda}\lambda^r=\sum_{t=r}^{\infty}\frac{n^{t-r}\lambda^t}{(t-r)!}$$

比较系数得

$$h_2(t)=0\qquad as\quad t=0,1,\cdots,r-1$$
$$h_2(t)=\frac{t!n^{t-r}}{(t-r)!n^t}\quad as\quad t=r,\cdots$$

综合得出

$$h_2(T)=\frac{T(T-1)\cdots(T-r+1)}{n^r}$$

由L-S定理知，$h_2(T)$为$g_2(\lambda)$的UMVUE

3. $g_3(\lambda)=P_\lambda(X_1=x)=\frac{e^{-\lambda}\lambda^x}{x!}$。令$\varphi(\mathbf{X})=I_{[X_1=x]}$有$E_\lambda[\varphi(\mathbf{X})]=P_\lambda(X=x)=g_3(\lambda)$，因此$\varphi(\mathbf{X})$是$g_3(\lambda)$的一个无偏估计。注意到$T\sim Poi(n\lambda)$，用$g_3(\lambda)$改进的无偏估计为

$$\begin{aligned}h_3(t)=&E(\varphi(\mathbf{X})|T=t)\\&=\frac{P(X_1=x,T=t)}{P(T=t)}\\&=\frac{P(X_1=x,\sum_{i=2}^{n}X_i=t-x)}{P(T=t)}\\&=\frac{(n-1)^{t-x}t!}{n^t(t-x)!x!}\\&=\binom{t}{x}(\frac{1}{n})^x(1-\frac{1}{n})^{t-x},t\geq x.\end{aligned}$$

由R-B定理知，$h(T)$是$g_3(\lambda)$的无偏估计量。他还是T的函数，由L-S定理知，$h_3(T)$是$g_3(\lambda)$的UMVUE。

□

在讨论无偏估计量时，若 $\hat{g}_n(X)$ 是 $g(\theta)$ 的无偏估计量，则 $\hat{g}_n(X) + aU(X)$ 也是 $g(\theta)$ 的无偏估计量，其中 a 是任意常数，$U(X)$ 是任何零的无偏估计量。我们可以分析方差：

$$\mathrm{Var}_\theta[\hat{g}_n(X)+aU(X)] = \mathrm{Var}_\theta[\hat{g}_n(X)]+a^2\mathrm{Var}_\theta[U(X)]+2a\mathrm{Cov}_\theta(U(X),\hat{g}_n(X))$$

若对于某些 θ ，$\mathrm{Cov}_\theta(U(X),\hat{g}_n(X)) \neq 0$，则存在一个 a 使得

$$\mathrm{Var}_\theta[\hat{g}_n(X) + aU(X)] < \mathrm{Var}_\theta[\hat{g}_n(X)]$$

因此，$\hat{g}_n(X)$ 不能是 UMVUE（统一最小方差无偏估计量）。这表明，与零的无偏估计量的协方差是确定 UMVU 估计量是否存在的关键。

定理 7.2.4. *Unbiased estimators of zero*

设 X 是来自 F_θ 的样本，$\hat{g}(X)$ 是 $G = \{\hat{g} : E_\theta \hat{g}^2(X) < \infty, \forall\theta \in \Theta\}$ 中的一个估计量，设 U 表示所有零的无偏估计量的集合。则 $\hat{g}(X)$ 是其期望 $g(\theta)$ 的 *UMVUE* 的充要条件是：

$$E_\theta(\hat{g}(X)U) = 0, \forall U \in \mathcal{U}, \forall\theta \in \Theta$$

（注意：因为 $E_\theta(U) = 0, \forall U \in \mathcal{U}$，所以 $E_\theta(\hat{g}(X)U) = cov_\theta(\hat{g}, U)$，这等价于 $\hat{g}(X)$ 与每个 $U \in \mathcal{U}$ 不相关的条件。）

例 7.2.5. 样本 $X_1, X_2, \cdots, X_n$ 来自 $B(1,p)$ ，用零无偏估计法证明样本均值是参数 p 的*UMVUE.*

解. 容易验证样本均值 $\bar{X}$ 是方差有限的无偏估计量，所以只需证明对所有零无偏估计量，它与 $\bar{X}$ 的协方差为0.

对任意零无偏估计量 $l = l(\tilde{X})$ ，它的期望自然为0，即

$$0 = \mathrm{E}_p(l) = \sum_{x_i=0,1;i=1,\ldots,n} l(\mathbf{x}) p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \quad \forall p \in (0,1)$$

令 $\theta = \frac{p}{1-p}$ 并在方程两边去掉系数 $(1-p)^n$ 得到

$$0 = \sum_{x_i=0,1;i=1,\ldots,n} l(\mathbf{x})\theta^{\sum_{i=1}^n x_i}, \quad \forall \theta > 0$$

由这个式子对 θ 求导得到

$$0 = \sum_{x_i=0,1;i=1,\ldots,n} \left(\sum_{i=1}^n x_i\right) \cdot l(\mathbf{x})\theta^{\sum_{i=1}^n x_i - 1}, \quad \forall \theta > 0$$

其中右边与 $\mathrm{E}_p(\bar{X} \cdot l(\mathbf{X}))$ 就差一个常数，所以 $\mathrm{E}_p(\bar{X} \cdot l(\mathbf{X})) = 0$ ，证毕. □

随感. 课件在此处的处理是很奇怪的，课件这里把零无偏放在L-S定理之后。但零无偏的的思想是对于一个估计，其他的无偏估计可以视为拆分该估计和一个零无偏估计。而当该估计与任意的零无偏无关时，其他的无偏估计的方差就会更大。

我们其实还有一个推论说明了零无偏估计的范围可以只限制在充分统计量之中。除此之外，我们会希望一个函数的零无偏估计量只有0，这样就不用考虑其他情况了，完备的估计满足了这个条件。

也就是说当估计T完全充分的时候，且这个估计T的函数g(T)是无偏的，那么这个估计就是UMVUE，这就是L-S定理的内容。

7.3 UMVUE

单开一章讲述如何求UMVUE，以显示求解UMVUE的重要性。

方法 7.3.1. *UMVUE*

完全充分统计量+无偏估计+条件期望作用（条件期望法） 找到一个完全充分统计量，接下来可以找一个无偏估计，然后用*R-B*定理改造以得到条件期望。

完全充分统计量+函数+无偏（统计量函数法） 找到一个完全充分统计量，然后再找一个关于他的函数，使得他是无偏的。

例 7.3.2. 二项分布的$UMVUE$（统计量函数法）

假设 $\tilde{X} = (X_1, X_2, \ldots, X_n)$ 是来自两点总体 $\{B(1,p); 0 < p < 1\}$ 的样本，已知$T = \sum_{i=1}^{n} X_i$是完全充分统统计量，求参数 p 和 p^2 的$UMVUE$。

参数 p 的$UMVUE$

我们求 T 的期望：

$$E(T) = E\left(\sum_{i=1}^{n} X_i\right) = np$$

由于 T 的期望关于 p 是线性函数，所以只需纠偏就可以得到 p 的$UMVUE$:

$$\hat{g}(T) = \frac{T}{n} = \bar{X}$$

$$E(\hat{g}(T)) = \frac{E(T)}{n} = p$$

参数 p^2 的$UMVUE$ 我们求 T^2 的期望：

$$E(T^2) = \sum_{i=1}^{n} E(X_i^2) + \sum_{i \neq j} E(X_i)E(X_j) = np + n(n-1)p^2$$

解出 p^2：

$$p^2 = \frac{E(T^2) - E(T)}{n(n-1)} = \frac{E(T^2 - T)}{n(n-1)}$$

因此，参数 p^2 的$UMVUE$为：

$$\frac{T^2 - T}{n(n-1)} = \frac{1}{n-1}\left(\frac{T^2}{n} - \bar{X}\right)$$

注记. 参数形式简单的时候，用统计量函数法，通过T的各种期望，与待估参数都有线性性，使用简单的加法乘法函数来纠偏，即可得到$UMVUE$。

例 7.3.3. 二项分布的*UMVUE*（条件期望法）

设$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} b(1,p)$，我们知道$T = \sum_{i=1}^{n} X_i$是$p$的充分完备统计量，使用*Lehmann-Scheffé*定理求解$g(p) = p(1-p)$的*UMVUE*。

证明. 令$\varphi(\mathbf{X}) = I_{[X_1=1,X_2=0]}$有$E[\varphi(\mathbf{X})] = P(X_1 = 1, X_2 = 0) = p(1-p)$，因此$\varphi(\mathbf{X})$是$g(p) = p(1-p)$的一个无偏估计。注意到$T \sim b(n,p)$，用$g(p) = p(1-p)$改进的无偏估计为

$$h(t) = E(\varphi(\mathbf{X})|T = t) = \frac{P(X_1 = 1, X_2 = 0, T = t)}{P(T = t)} = \frac{t(n-t)}{n(n-1)}$$

由R-B定理知，$h(T)$是$g(p)$的无偏估计量。他还是T的函数，由L-S定理知，$h(T)$是$p(1-p)$的UMVUE。 □

注记. 参数形式复杂，并且可以变化为一个概率的形式的时候，采用条件期望法。这里给出用条件期望法求*UMVUE*的步骤；

1. **找到充分完备统计量**：$T = \sum_{i=1}^{n} X_i$是p的充分完备统计量。
2. **构造初始无偏估计量**：通过指示函数$\varphi(\mathbf{X})$构造出一个初始无偏估计量。
3. **应用*Rao-Blackwell*定理**：将初始无偏估计量改进为基于完备充分统计量T的无偏估计量$h(T)$。
4. **应用*Lehmann-Scheffé*定理**：通过说明$h(T)$是充分完备统计量的函数，确认$h(T)$是*UMVUE*。

在这里我们意识到*R-B*定理降低方差的效果十分显著。

例 7.3.4. *Gamma*分布的*UMVUE*

设$X_i \overset{i.i.d}{\sim} \Gamma(\alpha, \lambda)$，其中$\alpha$已知，求$\lambda$的*UMVUE*。

解. 先找充分完备统计量，直接写出样本的联合密度

$$p(\mathbf{x};\lambda)=\frac{\lambda^{n\alpha}}{(\Gamma(\alpha))^n}\prod_{i=1}^{n}x_i^{\alpha-1}\exp\left\{-\lambda\sum_{i=1}^{n}x_i\right\}\cdot .I\{x_i>0,i=1,\cdots,n\}$$

显然 $-\lambda$ 的范围有非空内部，再结合因子分解定理，可知 $T=\sum_{i=1}^{n}X_i$ 为充分完备统计量.

接着求UMVUE，参数形式比较简单，用统计量函数法，先求充分完备统计量的期望：

$$E_\lambda(T)=\frac{n\alpha}{\lambda}$$

它关于参数不是线性的了，但是关于参数的倒数是线性的，自然想到求该统计量倒数的期望：

$$\begin{aligned}E_\lambda(\frac{1}{T})&=\int_0^\infty\frac{1}{t}p(t;\lambda)dt\\&=\frac{\lambda^{n\alpha}}{\Gamma(n\alpha)}\int_0^\infty t^{(n\alpha-1)-1}e^{-\lambda t}dt\\&=\frac{\lambda\Gamma(n\alpha-1)}{\Gamma(n\alpha)}\\&=\frac{\lambda}{n\alpha-1}\end{aligned}$$

这样就得到关于参数时线性的统计量，纠偏一下就得到UMVUE为

$$\widehat{\lambda}=\frac{n\alpha-1}{\sum_{i=1}^{n}X_i}=\frac{n\alpha-1}{n\bar{X}}$$

□

7.4 Cramer-Rao Lower Bound

定理 7.4.1. *Cramer-Rao Lower Bound*

设 $X_1,X_2,\ldots,X_n$ 是来自分布族 $\{F_\theta,\theta\in\Theta\}$ 的一个样本， $T(X)$ 是 θ 的一个充分统计量， $\hat{g}(T)$ 是 $g(\theta)$ 的一个无偏估计量。如果满足以下条件(称之为正则条件)：

1. $\hat{g}(T)$ 是连续的

2. $\frac{d}{d\theta}E_\theta\hat{g}(T) = E_\theta\frac{d}{d\theta}\hat{g}(T)$

3. $\frac{d}{d\theta}E_\theta\left[\frac{d}{d\theta}\hat{g}(T)\right] = E_\theta\left[\frac{d}{d\theta}\frac{d}{d\theta}\hat{g}(T)\right]$

那么，对于任意无偏估计 $\hat{g}(T)$，有：

$$Var_\theta(\hat{g}(T)) \geq \frac{(g'(\theta))^2}{nI(\theta)}$$

其中 $I(\theta)$ 是 *Fisher* 信息量，定义为：

$$I(\theta) = E_\theta\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]$$

注记. 这个定理说明了，对于任意无偏估计，其方差的下界是 $\frac{(g'(\theta))^2}{nI(\theta)}$，其中 $I(\theta)$ 是 *Fisher* 信息量。这个下界被称为 *Cramer-Rao* 下界。

定义 7.4.2. *Fisher Information*

对于一个由参数 θ 所确定的概率分布，其概率密度函数为 $f(x;\theta)$，*Fisher* 信息 $I(\theta)$ 定义为对数似然函数的一阶导数的方差。具体公式如下：

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)$$

或者等价地：

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]$$

其中，$\log f(X;\theta)$ 是对数似然函数，E 表示期望值。

注记. *Fisher*信息量越大，表示数据中关于参数 θ 的信息越多，即我们可以更精确地估计该参数。反之，*Fisher*信息量越小，表示数据中关于参数的信息越少，估计的准确性也越低。

性质 7.4.3. *Fisher*信息量越大，表示数据中关于参数 θ 的信息越多

设 $L(\theta; X) = \log f(X; \theta)$ 为对数似然函数，则*Fisher*信息可以写成：

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

当然在多变量情况下

$$I(\theta) = -E\left[\bigtriangledown_\theta \log f(X;\theta)\right]$$

这个公式利用了对数似然函数的二阶导数的期望值来计算*Fisher*信息。

例 7.4.4. 正态分布中的*Fisher*信息

假设 X 是来自正态分布 $N(\mu, \sigma^2)$ 的样本，其中 μ 是均值，σ^2 是已知的方差。对数似然函数为：

$$\log f(X;\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

对 μ 求一阶导数：

$$\frac{\partial}{\partial\mu}\log f(X;\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

对 μ 求二阶导数：

$$\frac{\partial^2}{\partial\mu^2}\log f(X;\mu) = -\frac{n}{\sigma^2}$$

因此，*Fisher*信息为：

$$I(\mu) = -E\left[-\frac{n}{\sigma^2}\right] = \frac{n}{\sigma^2}$$

定理 7.4.5. 在正则条件下，函数的无偏估计量达到 *Cramer-Rao* 下界的充分必要条件是，样本服从单参数指数族且$\hat{g}_n(X)$是充分统计量。

注记. 如果一个无偏估计量达到了*CR*下界，那么这个估计量必是充分统计量，且样本分布必须是单参数指数族。

定理 7.4.6. 设$X = (X_1, \cdots, X_n)$ 是从密度函数为 $f(x, \theta)$的总体中抽取的简单分布。如果 $g(\theta)$的某一估计$\hat{g}(\theta)$有效，那么该估计就是$g(\theta)$的MLE。

其中有效性定义为

$$e_{\hat{g}_n}(\theta) = \frac{[g'(\theta)]^2/(nI(\theta))}{Var_\theta(\hat{g}_n(X))} = 1$$

例 7.4.7. Poi分布的C-R下界

先求$Fish$信息$I(\lambda)$，密度函数函数为：

$$f(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

对对数密度函数求二阶导：

$$\frac{\partial^2 logp}{\partial \lambda^2} = -\frac{x}{\lambda^2}$$

求期望

$$I(\lambda) = -E_\lambda(-\frac{x}{\lambda^2}) = \frac{1}{\lambda^2} E_\lambda X = \frac{1}{\lambda}$$

C-R下界为

$$\frac{(g'(\theta))^2}{nI(\theta)} = \frac{\lambda}{n}$$

注记. 当有充分无偏估计量达到这个下界的时候，那他就是$UMVUE$。

Lec 8 Interval Estimation

8.1 Confidence Interval

定义 8.1.1. 区间估计

区间估计由两个函数 $L(x_1,\ldots,x_n)$ 和 $U(x_1,\ldots,x_n)$ 组成，它们依赖于样本数据。这些函数创建一个区间 $[L(x),U(x)]$，使得对于任何观察到的样本 x，区间满足 $L(x)\leq U(x)$。当观察到一个特定的样本 $X=x$ 时，区间 $[L(x),U(x)]$ 用于做出推断 $L(x)\leq\theta\leq U(x)$。这个区间是基于观测数据对参数 θ 的估计。

随机区间 $[L(X),U(X)]$，其中 X 表示随机样本，称为区间估计量。它提供了参数 θ 可能所在范围的估计，并具有一定的概率。

覆盖概率是指区间 $[L(X),U(X)]$ 包含真实参数 θ 的概率，记作 $P_\theta(\theta\in[L(X),U(X)])$.

区间估计的精度与区间的大小有关，可以通过不同的方法进行度量。一个常用的度量是区间的期望宽度 $E[U(X)-L(X)]$.

例 8.1.2. 正态分布对均值的区间估计量

假设我们有一个正态分布的总体，其均值为 μ，标准差为 σ（σ已知）。我们从该总体中抽取随机样本，样本量为 n。我们想要估计总体均值 μ 的一个区间估计量。

一个区间估计量的例子是：

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

其中，$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$是样本均值，$z_{\alpha/2}$ 是标准正态分布中，使得两尾概率之和为 α 的分位数。

（事实上$[X_{(1)}, X_{(2)}]$也可以是由随机样本构造的随机区间，故也是一个区间估计量，但可能在判断均值问题的时候使用起来不方便，故我们不使用这一随机区间。）

对于上述区间估计量，覆盖率为：

$$P\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha$$

该区间估计量的精度，我们按一般的区间的期望宽度来衡量：

$$E\left[\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right] = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

定义 8.1.3. 置信区间

设 $X = (X_1, \ldots, X_n)$ 是来自总体 $\{f(x \mid \theta), \theta \in \Theta\}$ 的一个样本，支撑集为 X。设 $L_n(X)$ 和 $U_n(X)$ 是统计量，使得 $L_n(x) \leq U_n(x)$，对于所有 $x \in X$ 均成立。那么，如果对于所有 $\theta \in \Theta$，

$$P_\theta(L_n(X) \leq \theta \leq U_n(X)) \geq 1 - \alpha$$

则 $[L_n(X), U_n(X)]$ 是参数 θ 的一个 $100(1-\alpha)\%$ 的置信区间（*CI*）。

对于大样本，如果对于所有 $\theta \in \Theta$，当 $n \to \infty$ 时，

$$P_\theta(L_n(X) < \theta < U_n(X)) \to 1 - \alpha$$

则区间 $(L_n(x), U_n(x))$ 是参数 θ 的一个 $100(1-\alpha)\%$ 的大样本的（渐近）置信区间。

置信水平：$1 - \alpha$

置信系数：$\inf_\theta P_\theta(L_n(X) \leq \theta \leq U_n(X))$

注记.

1. 有时使用开区间 $(L_n(x), U_n(x))$ 或者半开半闭区间会更加自然。

2. 区间也可以是单侧区间：$(-\infty, U_n(X)]$ 或者 $[L_n(X), \infty)$。

3. 该定义可以扩展到多维情况：二维为置信带（*confidence band*）；三维及以上为置信区域（*confidence region*）

4. 自然地，我们希望我们的区间具有较小的宽度和较大的覆盖概率，但这样的区间通常难以构造。

5. 一个典型的 θ 的置信区间形式是 $\hat{\theta} \pm sd \times F_q$，其中 F_q 是分位数，sd 是 $\hat{\theta}$ 的标准误差。值 $sd \times F_q$ 称为误差边界（*margin of error*）。

*CI*是一个随机区间，由随机变量构造的统计量来组成区间的两端。

8.2 The exact CI

枢轴法构造置信区间

定义 8.2.1. 如果随机变量 $g(\mathbf{X}, \theta)$ 的分布与参数 θ 独立，则称其为一个枢轴量或枢轴数量。

方法 8.2.2. 置信区间的一种构造

对于枢轴量 $g(\mathbf{X}, \theta)$，找到 $a < b$ 使得

$$P_\theta(a \le g(\mathbf{X}, \theta) \le b) \ge 1 - \alpha$$

将区间重写为

$$C(\mathbf{x}) = \{\theta : a \le g(x, \theta) \le b\} = \{\theta : L_n(x, a) \le \theta \le U_n(x, b)\}$$

则 $C(\mathbf{X})$ 是参数 θ 的一个 $1 - \alpha$ 置信区间。

例 8.2.3. 指数分布的置信区间

密度函数为

$$p(x;\theta) = \theta^{-1}e^{-x/\theta}, \quad x > 0$$

从中抽取简单随机样本 $X_1, X_2, \cdots, X_n$ ，求平均寿命 θ 的单侧置信下限.

解. 我们把过程总结为三步

1. 第一步，从充分统计量或点估计出发找枢轴量

   对于指数分布总体，充分统计量为 $T = \sum_{i=1}^{n} X_i$ ，由于指数分布是伽玛分布，结合伽玛分布的可加性得到 $T \sim \Gamma(n, 1/\theta)$ ，再利用伽玛分布关于尺度参数的伸缩性，得到

   $$2T/\theta \sim \Gamma(n, 1/2) = \chi^2(2n)$$

   它包含了待估参数且分布完全已知，所以是合适的枢轴量.

2. 第二步，确定常数

   因为是单侧情形，所以我们只需要确定一个常数。这里需要从最终想要得到的区间形式出发，由于求的是置信下限，所以区间形式为 θ 大于某个数，而在枢轴量中 θ 在分母上，所以应该是枢轴量小于某个数，也就是寻找 d.

   常数 d 可以用卡方分布的分位数表示，即

   $$\mathrm{P}_\theta\left\{\frac{2T}{\theta} \leq \chi^2_\alpha(2n)\right\} = 1 - \alpha$$

3. 第三步，等价改写得区间估计

   将上述概率中不等式改写，得到

$$\mathrm{P}_\theta\left\{\frac{2T}{\chi^2_\alpha(2n)} \le \theta\right\} = 1-\alpha$$

因此所求的单侧置信下限为

$$\widehat{\theta}_L = \frac{2T}{\chi^2_\alpha(2n)}$$

□

例 8.2.4. 正态分布总体均值和方差的置信区间的构造

1. σ已知，求μ

我们选择标准化的样本均值作为枢轴量:

$$T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

当样本量足够大时，Z 近似服从标准正态分布。我们找到标准正态分布中落在 $(-z_{\alpha/2}, z_{\alpha/2})$ 区间内的概率大于等于 $1-\alpha$，因此

$$[\bar{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}]$$

是均值的一个（渐进）$(1-\alpha)$置信区间

2. σ未知，求μ

我们选择枢轴量:

$$T = \frac{(\bar{X}-\mu)\sqrt{n}}{S}$$

由*3.1*得$T \sim t_{n-1}$故为枢轴量，因此

$$[\bar{X} - \frac{St_{(n-1,\alpha/2)}}{\sqrt{n}}, \bar{X} + \frac{St_{(n-1,\alpha/2)}}{\sqrt{n}}]$$

是均值的一个（渐进）$(1-\alpha)$置信区间

3. μ已知，求σ 枢轴量：

$$S_\mu^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}, T = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

令

$$c_1 : P_{\sigma^2}(\frac{nS_\mu^2}{\sigma^2} < c_1) = \frac{\alpha}{2}, c_2 : P_{\sigma^2}(\frac{nS_\mu^2}{\sigma^2} > c_2) = \frac{\alpha}{2}$$

置信区间：

$$\left[\frac{nS_\mu^2}{c_2}, \frac{nS_\mu^2}{c_1}\right]$$

4. μ未知，求σ 枢轴量：

$$T = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

令

$$d_1 : P_{\sigma^2}(\frac{nS^2}{\sigma^2} < d_1) = \frac{\alpha}{2}, d_2 : P_{\sigma^2}(\frac{nS^2}{\sigma^2} > d_2) = \frac{\alpha}{2}$$

置信区间：

$$\left[\frac{(n-1)S^2}{d_2}, \frac{(n-1)S^2}{d_1}\right]$$

8.3 Other methods*

查阅其他资料，摘抄讲义中的内容，仅作参考阅读。

例 8.3.1. ***Wilson Interval***

The Wilson interval is a method for constructing a confidence interval for a proportion in a binomial distribution. Unlike the traditional Wald interval, which can perform poorly for small sample sizes or proportions near 0 or 1, the Wilson interval is more reliable and provides better coverage properties.

The Wilson interval is derived by inverting a hypothesis test and can be represented as follows:

$$CI_W = \frac{\hat{p} + \frac{\kappa^2}{2n}}{1 + \frac{\kappa^2}{n}} \pm \frac{\kappa\sqrt{n\hat{p}(1-\hat{p}) + \frac{\kappa^2}{4}}}{n + \kappa^2}$$

where:

- $\hat{p}$ *is the sample proportion.*
- n *is the sample size.*
- κ *is the critical value from the standard normal distribution corresponding to the desired confidence level.*

例 8.3.2. ***Wald Interval***

The Wald interval is a method for constructing a confidence interval for a proportion in a binomial distribution. It is one of the simplest methods and is based on the normal approximation to the binomial distribution. The Wald interval is given by:

$$CI_S = \hat{p} \pm \kappa \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where:

- $\hat{p}$ *is the sample proportion.*
- $\hat{q} = 1 - \hat{p}$ *is the complement of the sample proportion.*
- n *is the sample size.*
- κ *is the critical value from the standard normal distribution corresponding to the desired confidence level.*

However, the Wald interval can perform poorly for small sample sizes or when the proportion is near 0 or 1. This is because the normal approximation to the binomial distribution may not be accurate in these cases.

例 8.3.3. ***Agresti-Coull Interval***

The Agresti-Coull interval is a method for constructing a confidence interval for a proportion in a binomial distribution. It is an improvement over the Wald interval and is known to perform better, especially for small sample sizes or proportions near 0 or 1. The Agresti-Coull interval adds a correction factor to the Wald interval, making it more reliable.

The Agresti-Coull interval is given by:

$$CI_{AC} = \tilde{p} \pm \kappa \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}$$

where:

- $\tilde{p} = \frac{\tilde{X}}{\tilde{n}}$ *is the adjusted sample proportion.*
- $\tilde{q} = 1 - \tilde{p}$ *is the complement of the adjusted sample proportion.*
- $\tilde{X} = X + \frac{\kappa^2}{2}$ *is the adjusted number of successes.*
- $\tilde{n} = n + \kappa^2$ *is the adjusted sample size.*
- κ *is the critical value from the standard normal distribution corresponding to the desired confidence level.*

例 8.3.4. ***Agresti-Coull Interval***

The Agresti-Coull interval is a method for constructing a confidence interval for a proportion in a binomial distribution. It is an improvement over the Wald interval and is known to perform better, especially for small sample sizes or proportions near 0 or 1. The Agresti-Coull interval adds a correction factor to the Wald interval, making it more reliable.

The Agresti-Coull interval is given by:

$$CI_{AC} = \tilde{p} \pm \kappa \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}$$

where:

- $\tilde{p} = \frac{\tilde{X}}{\tilde{n}}$ *is the adjusted sample proportion.*
- $\tilde{q} = 1 - \tilde{p}$ *is the complement of the adjusted sample proportion.*
- $\tilde{X} = X + \frac{\kappa^2}{2}$ *is the adjusted number of successes.*
- $\tilde{n} = n + \kappa^2$ *is the adjusted sample size.*
- κ *is the critical value from the standard normal distribution corresponding to the desired confidence level.*

例 8.3.5. ***Exact Interval***

Exact interval: $CI_{CP} = [L_{CP}(x), U_{CP}(x)]$, *where* $L_{CP}(x)$ *and* $U_{CP}(x)$ *are, respectively, the solutions in* p *to the equations:*

$$P_p(X \geq x) = \alpha/2$$
$$P_p(X \leq x) = \alpha/2$$

The exact interval is a method for constructing a confidence interval for a proportion in a binomial distribution that does not rely on approximations. Instead, it uses the exact distribution of the binomial variable. The exact interval is often based on the Clopper-Pearson method, which provides conservative confidence intervals by inverting the binomial test.

例 8.3.6. ***Bootstrap Confidence Interval***

Introduction:

Bootstrap confidence intervals are a non-parametric approach to estimate the confidence intervals of a parameter. This method was introduced by Brad Efron in 1979 and has become a powerful tool in statistical inference. The bootstrap method involves resampling the original data with replacement to create multiple bootstrap samples. From these samples, we can calculate the

bootstrap estimates and use their distribution to approximate the confidence interval of the parameter of interest.

Explanation:

The idea behind the bootstrap confidence interval is illustrated as follows:

Resampling: From the initial sample, generate multiple bootstrap samples by sampling with replacement.

Estimation: Calculate the parameter estimate (e.g., mean, median) for each bootstrap sample to obtain a distribution of the bootstrap estimates.

Approximation: Use the empirical distribution of the bootstrap estimates to approximate the unknown distribution of the parameter estimate from the original sample.

Confidence Interval: Determine the confidence interval from the distribution of bootstrap estimates.

For a percentile bootstrap confidence interval, we take the $2.5th$ *and* $97.5th$ *percentiles of the bootstrap estimates to form the lower and upper bounds of a* 95% *confidence interval.*

Steps for Basic Bootstrap Confidence Interval:

Original Sample: Consider an initial sample $X = (X_1, X_2, \ldots, X_n)$ *and compute the estimator* $\hat{\theta}$.

Bootstrap Samples: Generate M *bootstrap samples* $(X_1^*, X_2^*, \ldots, X_M^*)$ *from the original sample by resampling with replacement.*

Bootstrap Estimates: Calculate the estimator $\hat{\theta}_i^*$ *for each bootstrap sample* $i = 1, 2, \ldots, M$.

Percentile Interval: Find the $2.5th$ *percentile* $\hat{\theta}^{(0.025)*}$ *and* $97.5th$ *percentile* $\hat{\theta}^{(0.975)*}$ *of the bootstrap estimates.*

Adjust for Bias: If necessary, adjust for bias by using the relationship $\theta - \hat{\theta} \approx \hat{\theta} - \hat{\theta}^*$.

Mathematical Formulation:

Given the bootstrap estimates $\hat{\theta}_*^1, \hat{\theta}_*^2, \ldots, \hat{\theta}_*^M$*, the percentile confidence interval is:*

$L = \hat{\theta}^{(0.025)*}, U = \hat{\theta}^{(0.975)*}$

For a basic bootstrap confidence interval, we use the fact that the behavior of $\theta - \hat{\theta}$ *is approximately the same as the behavior of* $\hat{\theta} - \hat{\theta}^*$*. Thus:*

$L = 2\hat{\theta} - \hat{\theta}^{(0.975)*}, U = 2\hat{\theta} - \hat{\theta}^{(0.025)*}$

注记. 对课件里的这部分给出一个解释。

*1.*总体分布未知

在这个阶段，$\widehat{\theta} = \widehat{\theta}(\tilde{X})$ 的分布依赖于总体分布，但由于我们不知道总体分布，因此也无法确定这个统计量的分布。为了解决这个问题，我们利用样本数据得到的经验分布函数来代替总体分布。

2. 总体分布近似已知

现在，我们假设已经知道了总体的近似分布，我们可以通过模拟的方法从这个近似分布中抽取新的样本 $\widetilde{X}^* = (X_1^*, \cdots, X_m^*)$。利用这个新样本，我们采用相同的方法构造 $\hat{\theta}^* = \widehat{\theta}(\tilde{X}^*)$，这样我们就可以用 $\widehat{\theta}^*$ 来近似 $\widehat{\theta}$。

3. 总体分布与 $\widehat{\theta}$ 的分布近似已知

注意到 $\widehat{\theta} = \widehat{\theta}(\tilde{X})$ 是 θ 的估计量，因此我们可以用 $\widehat{\theta}^* - \widehat{\theta}(\tilde{x})$ 的分布来近似 $\widehat{\theta} - \theta$ 的分布，其中 $\widehat{\theta}^* - \widehat{\theta}(\tilde{x})$ 的分布是已知的。因此，我们可以通过 $\widehat{\theta}^* - \widehat{\theta}(\tilde{x})$ 的分布得到常数 c 和 d。

4. 常数近似已知

最后，我们将不等式中间项进行近似替换，得到 $\mathrm{P}^*(c \leq \widehat{\theta} - \theta \leq d) \approx 1 - \alpha$。进一步等价改写不等式，即可得到参数 θ 的置信区间近似为 $[\widehat{\theta}(\tilde{X}) - d, \widehat{\theta}(\tilde{X}) - c]$。

8.4 Methods of Evaluating Interval Estimators*

The shortest interval is a confidence interval that, for a given confidence level, has the minimum possible length. This method aims to improve the precision of the estimate by minimizing the interval length.

定义 8.4.1. *Uniformly Most Accurate (UMA) Confidence Set*

A uniformly most accurate confidence interval is one that provides the highest confidence level for all possible values of the parameter. It satisfies the following conditions:

1. *It is a* $1-\alpha$ *confidence interval that attains the confidence coefficient* $1-\alpha$*, i.e.,*

$$P_\theta(L^*(X) \leq \theta \leq U^*(X)) = 1-\alpha \quad \textit{for all } \theta \in \Theta.$$

2. *For any other* $1-\alpha$ *confidence interval* $[L(X), U(X)]$*, it satisfies*

$$P_\theta(L(X) \leq \theta \leq U(X)) \leq P_\theta(L^*(X) \leq \theta \leq U^*(X)) \quad \textit{for all } \theta \in \Theta.$$

Lec 9 Introduction to Parametric Tests

9.1 Statistical hypothesis testing

这部分很自然，不用过多解释，给出例子用于熟悉概念。

例 9.1.1. 假设检验的例子

假设我们有一个硬币，我们想要检验这个硬币是否是均匀的。我们抛了 100 次硬币，结果有 60 次正面朝上。我们想要检验这个硬币是否是均匀的，即正面朝上的概率是否为 0.5。

我们可以建立如下的假设检验问题：

- 零假设 H_0：硬币是均匀的，正面朝上的概率为 0.5。
- 备择假设 H_1：硬币不是均匀的，正面朝上的概率不为 0.5。

我们可以使用二项分布来建立检验统计量，假设硬币是均匀的，那么抛 100 次硬币正面朝上的次数 $X \sim B(100, 0.5)$。

通过计算，我们可以得到 $P(X \geq 60) \approx 0.028$，这个概率很小，因此我们可以拒绝零假设，认为硬币不是均匀的。

事实上，我们把这个规则称之为显著性检验，根据是否满足这个规则，我们将样本空间分为拒绝域和接受域。有时候为了方便可以引入检验函数，定义为接受域的示性函数。

在这个例子中我们计算 $P(X \in [50-a, 50+a]) \geq 0.95$得到得到$a > 7$，因此我们可以将接受域设为$[43, 57]$。

注记. *1.*在假设检验中，我们通常会设定一个显著性水平 α，如果 P 值小于 α，我们就拒绝零假设。常见的显著性水平有 0.05 和 0.01。 *2.*如果不是对参数的假设，称为非参数检验。(在这个问题里，对硬币是否均匀的假设实际上是对两点分布中的参数的假设，因此是参数检验。) *3.*假设检验有两种错误，第一类错误（拒真）和第二类错误（纳伪)。第一类的意思是拒绝了正确的零假设，第二类的意思是接受了错误的零假设。

例 9.1.2. *功效函数与检验水平*

设$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$，其中$\mu$未知，$\sigma$已知。设$H_0 : \mu = \mu_0 = 0$，$H_1 : \mu \neq \mu_0$问题的否定域为

$$\left\{(X_1, \cdots, X_n) : \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| \geq c\right\}$$

否定域的含义是，如果现实中抽样所得的$\bar{X}$属于上述集合，我们就拒绝原假设。

检验水平通常是我们事先给好的，比如$\alpha = 0.01, 0.05, \cdots$，检验水平是拒绝零假设的概率，即第一类错误的概率。

功效函数$\beta(\mu)$定义为$\beta(\mu) = P_\mu$(接受H_0)，是衡量第二类错误的概率。

在这个问题里面，我们做一些计算并给出说明：

1. 当 $\alpha = 0.1$ 时：

$$c = \Phi^{-1}(1 - 0.05) = \Phi^{-1}(0.95) \approx 1.645.$$

2. 当 $\alpha = 0.05$ 时：

$$c = \Phi^{-1}(1 - 0.025) = \Phi^{-1}(0.975) \approx 1.96.$$

3. 当 $\alpha = 0.01$ 时：

$$c = \Phi^{-1}(1 - 0.005) = \Phi^{-1}(0.995) \approx 2.575.$$

我们可以看出，当检验水平越小，拒绝域越大，因此功效函数越小。也就是我们要求第一类错误的概率越小时，我们希望不要排除正确的选择，因此会带来更大的拒绝域。

1. 当 $\alpha = 0.1$，$c \approx 1.645$：

$$\beta(\mu) = 1 - \Phi\left(1.645 - \frac{\mu}{0.4}\right) + \Phi\left(-1.645 - \frac{\mu}{0.4}\right).$$

2. 当 $\alpha = 0.05$，$c \approx 1.96$：

$$\beta(\mu) = 1 - \Phi\left(1.96 - \frac{\mu}{0.4}\right) + \Phi\left(-1.96 - \frac{\mu}{0.4}\right).$$

3. 当 $\alpha = 0.01$，$c \approx 2.575$：

$$\beta(\mu) = 1 - \Phi\left(2.575 - \frac{\mu}{0.4}\right) + \Phi\left(-2.575 - \frac{\mu}{0.4}\right).$$

固定μ进行比较的话，可以看出当拒绝域越大的时候，功效函数越小，也就是说第二类错误的概率越大。这也很直观，如果拒绝域更大，说明我们很容易接受了一个错误的零假设。

功效函数是会随着实际上的μ而变化，意思是说，如果真实的μ离我们的零假设越远，我们越容易拒绝零假设，也就是功效函数越大；如果真实的μ离我们的零假设越近，我们越容易接受零假设，也就是功效函数越小。

否定域通常是由题设或者经验决定的。如果否定域的选取导致零假设为真时，有超过设定检验水平的概率落入否定域；或者如果否定域的选取导致在备择假设为真时，功效函数的值较低，那么否定域的选取可能是不合理的。

上面也给出了一个计算顺序，先由经验假设否定域的形式，接着由检验水平去计算否定域的值，接着给出功效函数。

9.2 Tests about a Normal mean

让我们摆脱以下全符号的说明，全部都用一个现实例子来说明，我觉得这样很好玩。

例 9.2.1. 罐头厂（*1*）

食品厂用自动装罐机装罐头食品，每罐的标准重量为*500g*，每天开工需检查机器的工作状况，今抽得*10*罐，称得重量如下（单位：*g*）：

495，*510*，*505.498*，*503*，*492*，*502*，*512*，*497*，*506*

假定罐头重量服从正态分布，问该机器是否正常工作？（检验水平取*0.05*）

解. 此处给出相对完整的过程，与上面的说明相对应

先计算上述数据的样本均值和样本方差，得到$\bar{X} = 502$。

这个问题里我们检验的是均值，假设正常工作，即设$X_1, \cdots, X_10 \sim N(500, \sigma^2)$，其中$\mu_0$已知，$\sigma$未知。设$H_0 : \mu = \mu_0 = 500$，$H_1 : \mu \neq \mu_0$问题的否定域为

$$\left\{(X_1, \cdots, X_n) : \left|\frac{(\bar{X} - \mu_0)\sqrt{n}}{S}\right| \geq c\right\}$$

令

$$u = u(X) = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S}$$

其中$\mu_0 = 500$，$u \sim t_9$。

查表得$t_{0.025,9} = 2.262$，故否定域为$\{u : |u| \geq 2.262\}$，而计算知$u = 0.995$，因此在容许区间内，我们接受零假设，认为机器正常工作。 □

注记. 功效函数

在这个问题里面，我们使用的统计量是$u = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$，其中$\bar{X}$是样本均值，$\mu_0$是零假设中假定的均值，$s$是样本标准差，$n$是样本量。

对于实际的均值μ_a，我们定义统计量

$$Z = \frac{\bar{X} - \mu_a}{s/\sqrt{n}}$$

$$Z_{\mu_a} = \frac{\mu_a - \mu_0}{s/\sqrt{n}}$$

于是$u = Z + Z_{\mu_a}$，因此我们有

$$\beta(\mu_a) = P\left(Z > \frac{t_{\alpha/2,n-1} \cdot s/\sqrt{n} - (\mu_a - \mu_0)}{s/\sqrt{n}}\right) + P\left(Z < \frac{-t_{\alpha/2,n-1} \cdot s/\sqrt{n} - (\mu_a - \mu_0)}{s/\sqrt{n}}\right)$$

代入数值进行计算：

$$\beta(\mu_a) = 1 - \Phi(\frac{\mu_a - 500}{6.35/\sqrt{10}} - 2.262) + \Phi(-\frac{\mu_a - 500}{6.35/\sqrt{10}} - 2.262)$$

- $\mu_a = 495$

$$Power = 0.1189$$

- $\mu_a = 500$

$$Power = 0.0500$$

- $\mu_a = 505$

$$Power = 0.4107$$

这里给出了一些实际均值（也就是实际上的罐头重量变为的重量），所对应的概率，当罐头变为$505g$时，我们有41.07%的概率拒绝零假设，认为机器不正常工作。

注意到当实际均值 μ_a 与假设均值 μ_0 相同的时候，功效恰好等于显著性水平 α。这是因为在这种情况下，我们没有偏离零假设的情况，任何对零假设的拒绝完全是由于样本变异导致的“假阳性”结果，这正是显著性

水平的定义。显著性水平就是零假设为真时犯第一类错误（错误拒绝零假设）的概率，因此当 $\mu_a = \mu_0$ 时，功效函数值等于显著性水平 α。

要意识到的是，在计算功效函数的过程中，使用了样本数据，也就是说，功效函数是基于我们已经得到的样本数据，在实际均值与假设均值不同时，检验的灵敏度。

如果机器已经异常，实际平均重量为$505g$时，我们有58.93%的概率无法检验到这个差异。如果机器实际平均重量为$495g$时，我们有88.11%的概率无法检验到这个差异。

例 9.2.2. 罐头厂（*2*）

现在我们假设之前几年统计出来的罐头的重量的标准差为$5g$，但我们这几天因为神奇的原因忘记了每个罐头应该装多少肉，既不愿意看历史的统计资料也不愿意看包装上的标准重量，更不巧的是，我们还换了一个操作员。出了这么大乱子，我们保持沉着冷静，抽得10罐，称得重量如上（单位：g）。并假定罐头重量服从正态分布，问从标准差的角度看该机器是否正常工作？（检验水平取0.05）

解. 已知$\bar{X} = 502$.

这个问题中未知均值，检验标准差，令

$$T = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi_9^2$$

计算知$T = 15.2$，查表得χ_9尾部概率为0.05的点为16.92 $>$ 15.2因此也在容忍区间内，接受假设。 □

例 9.2.3. 罐头厂（*3*）

现在我们假设之前几年统计出来的罐头的重量的标准差为$5g$，填充量为$500g$，这两天天气百年难得一遇的热，不知道对机器有没有影响，抽取了10罐重量如上，问该机器是否正常工作？（检验水平取0.05）

解. 已知$\bar{X} = 502$，$S^2 = 38$。

这个问题里我们先检验的是均值，假设正常工作，令

$$u = u((X)) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

其中$\mu_0 = 500$，$u \sim N(0,1)$。

计算知$u = 1.265$，查表得$u_1.27 = 0.8980 < 0.9975$，因此在容许区间内，我们接受零假设.

再检验方差，令

$$T = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi^2_{10}$$

计算知$T = 16.8$，查表得χ_{10}尾部概率为0.05的点为18.31 > 16.8因此也在容忍区间内，接受假设。 □

9.3 P value

这部分书上写的异常复杂

定义 9.3.1. *P值*

P值是在零假设成立的情况下，观察到的统计量或更极端情况出现的概率。P值越小，我们就越有理由拒绝零假设。

P值的计算方法取决于检验问题的具体形式，但通常是通过查找分布表或计算分布函数得到的。

命题 9.3.2. *The P-value is the smallest significance level α at which the null hypothesis can be rejected. Because of this, the P-value is alternatively referred to as the observed significance level (OSL) for the data.*

$$p\ value = \sup_{\theta \in \Theta_0} P(Observed\ T\ or\ more\ extreme\ value)$$

Decision rule based on p-value: Select a significance level α (as before, the desired type I error probability). Then reject H_0 if P-value $\leq \alpha$; do not reject H_0 if P-value $> \alpha$.

注记. 我们之前是通过设定显著性水平（比如*0.05*）来计算检验的统计量的临界边界是多少，然后在比对是否拒绝零假设。现在我们可以通过计算*P*值，然后比对*P*值和显著性水平来决定是否拒绝零假设。

举个例子，上面那个判断硬币是否均匀的例子之中。我们选定显著性水平为*0.1*，然后得出容许区间是*[42,58]*，*60*不在其中因此矛盾，而*P*值检验说的是，我们计算得到大于等于*60*次，或者小于等于*40*此的概率为$0.028\times 2 = 0.056$，这个值小于*0.1*，因此我们拒绝零假设。

注意两点，当给出正面朝上是*60*次的时候，我们先做判断，大于*60*次还是小于*60*次的概率*¡0.5*（这里当然是前者）。然后要注意到比*60*次还极端的例子不仅是大于等于*60*次，还有小于等于*40*次的情况，因此我们要乘以*2*，这是由分布是双边带来的。

如果还是看的晕晕的话，可以看下面给出的例子。

例 9.3.3. 两台机床生产同一个型号的滚珠，从从甲机床生产的滚珠中抽取*8*个，从乙机床生产的滚珠中抽取*9*个，测得这些滚珠的直径，的方差之比有

$$\frac{S_1^2}{S_2} = 0.795$$

假设两者都服从正态分布，要求检验两者方差相等。

解. 这个问题中我们检验的是方差，假设两者方差相等，令

$$F = \frac{S_1^2}{S_2} \sim F_{7,8}$$

知$F = 0.795$，查表得$F_{0.025,7,8} = 0.268$，$F_{0.975,7,8} = 3.50$，因此在容许区间内，我们接受零假设。

但是我们也可以计算P值，计算得到$P(F \leq 0.795) = 0.39$，而这个检验是双边的，因此P值为0.78，大于0.05，因此我们接受零假设。 □

注记. 统计显著性与实际显著性

小的p值通常表示统计显著性，并建议拒绝零假设 H_0。然而，小的p值也可能是由于大样本量和从 H_0 的微小偏离所致，而这种偏离在实际意义上可能并不显著。举了一个例子，其中 H_0 断言总体均值 μ 为 100，而 H_a 断言 $\mu > 100$。如果真实均值 μ 为 101，那么观察到样本均值 $\bar{x} = 101$ 不应强烈主张拒绝 H_0，因为误差相对较小。随着样本量 n 的增加，即使结果的实际显著性可能仍然最小，p值也会变得更小。

9.4 The Duality between confidence intervals and tests*

说明了P值检验和置信区间检验存在着一个对应。

Recall the definition of $1-\alpha$ confidence region $S(X)$:

$$P_\theta(\theta \in S(X)) \geq 1-\alpha, \forall \theta \in \Theta$$

Next, consider the testing framework where we test the hypothesis $H_0 : \theta = \theta_0$ for some specified value θ_0. Suppose we have a test $\phi(x, \theta_0)$ with level α.

Then the acceptance region

$$A(\theta_0) = \{x : \phi(x, \theta_0) = 0\}$$

is a subset of sample space X with probability at least $1-\alpha$.

注记. 置信区间和假设检验之间的对偶性

定义 $1-\alpha$ 置信区域 $S(X)$，使得对于参数空间 Θ 中的所有 θ，有 $P_\theta(\theta \in S(X)) \geq 1-\alpha$。在假设检验中，我们检验零假设 $H_0 : \theta = \theta_0$。如果我们有一个水平为 α 的检验 $\phi(x, \theta_0)$，则接受区域 $A(\theta_0) = \{x : \phi(x, \theta_0) = 0\}$ 的概率至少为 $1-\alpha$。对偶性定理指出，如果 $S(X) = \{\theta_0 \in \Theta : x \in A(\theta_0)\}$ 是 θ 的 $1-\alpha$ 置信区域，则当且仅当 $P_{\theta_0}[X \in A(\theta_0)] \geq 1-\alpha$。

例 9.4.1. 设$X_1, X_n i.i.d \sim N(\mu, \sigma^2)$,μ, σ未知，利用假设检验的方法导出σ^2的置信系数为$1-\alpha$的置信区间.

首先，样本方差 S^2 的计算公式为：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

其中 $\bar{X}$ 为样本均值：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

已知 $\frac{(n-1)S^2}{\sigma^2}$ 服从卡方分布，具有 $n-1$ 自由度，即：

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

为了导出 σ^2 的置信区间，我们需要确定 χ^2 分布的关键值，使得其覆盖概率为 $1-\alpha$。

由此，我们可以得到 σ^2 的接受域：

$$P\left(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha$$

通过重新整理，我们可以得到 σ^2 的置信区间：

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha$$

因此，σ^2 的 $1-\alpha$ 置信区间为：

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right)$$

注记. 中间得到接受域，然后重新整理接受域就是置信区间，这就是对偶性的体现。

定理 9.4.2. *Let $S(x) = \{\theta_0 \in \Theta : x \in A(\theta_0)\}$, the set of θ_0 being accepted under sample x and test ϕ, then*

$$P_{\theta_0}[X \in A(\theta_0)] \geq 1 - \alpha$$

if and only if $S(X)$ is a level $1-\alpha$ confidence region for θ.

证明. For some specified θ_0, H_0 may be accepted, for other specified θ_0, H_0 may be rejected.

- Consider the set of θ_0 for which H_0 is accepted, that is, $S(x)$. By definition,

  $$P_{\theta_0}(\theta_0 \in S(X)) = P_{\theta_0}(X \in A(\theta_0)) \geq 1 - \alpha$$

  $S(X)$ is a level $1-\alpha$ confidence region for θ.

- Conversely, if $S(X)$ is a level $1-\alpha$ confidence region for θ, then define the test

  $$\phi(x, \theta_0) = \begin{cases} 1, & \text{if } \theta_0 \notin S(x) \\ 0, & \text{if } \theta_0 \in S(x) \end{cases}$$

  Then $\phi(X, \theta_0)$ is a level α test for H_0.

□

注记.

Fisher: *We do not need an alternative hypothesis; we can simply test a null hypothesis using a goodness-of-fit test. The outcome is a p-value, providing a measure of evidence for the null hypothesis.*

Neyman: *We must perform a hypothesis test between a null and an alternative. The test is such that it would result in type-1 errors at a fixed, pre-specified rate,* α*. The outcome is a decision - to reject or not reject the null hypothesis at the level* α*. We need an alternative from a decision-theoretic perspective - we are making a choice between two courses of action - and because we should report the power of the test* $1-\beta(Accept\ H_0|H_1)$*, we should seek the most powerful tests possible to have the best chance of rejecting* H_0 *when the alternative is true. To satisfy both these points, the alternative hypothesis cannot be the vague 'not* H_0*' one.*

*Fisher*关注于使用拟合优度检验来检验零假设，生成*p*值作为对零假设的证据。*Neyman*强调需要备择假设，并主张控制*I*类错误率在预定水平 α 的假设检验。（欸不是真的有人读这一段话吗，我只是上一段英文剩下了一点，这一面只有这些的话很不好看，所以把他翻译了一遍，没有什么重要的思想的。）*Neyman*的方法的结果是做出拒绝或不拒绝零假设的决定，并旨在使检验具有最强的效能，以最大化在备择假设为真时拒绝零假设的概率。

Lec 10 Uniformly Most Powerful Tests

10.1 Uniformly Most Powerful Test

定义 10.1.1. 设Φ_α表示所有显著性水平为α的检验。对于检验问题$H_0 : \theta \in \Theta_0$对比$H_1 : \theta \in \Theta_1$，如果$\phi \in \Phi_\alpha$是一个一致最强检验（$UMPT$），则

$$\beta_\phi(\theta) \geq \beta_{\phi_1}(\theta), \quad \forall \phi_1 \in \Phi_\alpha, \forall \theta \in \Theta_1$$

也就是说，对于$\alpha \in [0, 1]$，我们选择ϕ以最大化其功效，同时保证显著性水平α:

$$\arg\max_\phi \beta_\phi(\theta) = \mathbb{E}_\theta \phi(X), \quad \forall \theta \in \Theta_1$$

并且满足

$$\max_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha$$

注记. 假设检验问题的形式为：检验原假设 $H_0 : \theta \in \Theta_0$ 对比备择假设 $H_1 : \theta \in \Theta_1$。$\Theta_0$ 是原假设下参数 θ 的集合。Θ_1 是备择假设下参数 θ 的集合。

检验函数 $\phi(X)$ 定义为 $\phi : X \to P(\text{拒绝} H_0)$

β 是功效函数（*power function*）。功效函数 $\beta(\theta)$ 描述了检验在不同参数值下拒绝原假设的概率。具体来说，$\beta(\theta)$ 在参数 θ 下定义为：

$$\beta(\theta) = P(\text{拒绝 } H_0 \mid \theta)$$

这是检验函数 $\phi(X)$ 的期望值，其中 X 是观察到的数据：

$$\beta(\theta) = \mathbb{E}_\theta[\phi(X)]$$

显著性水平（*significance level*） α_ϕ 是检验在原假设为真时错误拒绝 H_0 的概率，即第一类错误的概率：

$$\alpha_\phi = \sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\phi(X)] = \sup_{\theta\in\Theta_0} P(\text{拒绝 } H_0 \mid \theta)$$

显著性水平为α的检验的全体$\Phi_\alpha = \{\phi|\alpha_\phi \geq \alpha\}$

例 10.1.2. 喜欢抛出正面的硬币

我们感觉某硬币更容易抛出正面（记正面向上为*0*，反面向上为*1*，这是个伯努利分布），我们想检验这是否正确，于是我们定义这枚硬币抛出正面的概率为p，接受域$P_0 = (0.5, 1]$，零假设为$H_0 : p \in P_0$,给定显著性水平为*0.99*，试验次数n=*7*.

现在我们提出两个检验函数

$$\phi_1(X) = \bar{X}, \qquad \phi_2(X) = X_{(n)}$$

（这两个函数都满足：当样本中有更多X_i的值为*1*的时候，反面越多，函数越大，拒绝零假设的概率也越大。）

我们可以计算得知

$$\beta_{\phi_1} = \mathbb{E}_\theta[\bar{X}] = \mathbb{E}_\theta[X_i] = P(X_i = 1) = 1 - p$$

$$\beta_{\phi_2} = \mathbb{E}_\theta[X_{(n)}] = P(X_{(n)} = 1) = 1 - p^n$$

对于所有的p都有$\beta_{\phi_2} \geq \beta_{\phi_1}$所以认为前者是一个更强的检验，从直观上来说，也是$\phi_2$只要等于*0*了，能代表这枚硬币抛出正面的能力是很强的。

但是注意到一个问题，那不是p的幂次越大这个函数会一致越强吗？这时候就要考虑显著性水平来制约了。

我们可以计算得知：

$$\underset{(0,5,1]}{\alpha_{\phi_1} = (1-p) = 0.5}$$

$$\underset{(0,5,1]}{\alpha_{\phi_1} = (1-p^n) = 0.992 > 0.990}$$

这说明在显著性*0.99*的制约下，如果一味的增加p的幂次，那么就会超过显著性。直观上也是，尽管ϕ_2只要等于*0*，就能说明很大问题，但是这种情况过于极端了，我们不需要保证每次抛出都是正面，大量抛出的时候只有一两次反面也能说明这枚硬币更容易抛到正面。

但是如果存在一个检验函数，他的$\beta(p)$的值在任意p下都是最大的（说明他发生的条件比较苛刻，可以一旦观察到结果就可以很好说明问题），而且显著性水平没有超过所给的要求（说明不会条件不会太苛刻，以至于难以观察到结果）那这个函数就就可以拿来对着样本检验假设是否成立了。这个检验函数就是*UMPT*。

定理 10.1.3. *Neyman-Pearson Lemma*

考虑检验$H_0 : \theta = \theta_0$对比$H_1 : \theta = \theta_1$（简单原假设与简单备择假设），其中与θ_i对应的样本概率密度函数或概率质量函数是$f(x, \theta_i)$，$i = 0, 1$。

1. ***存在性***：对于每个$\alpha \in (0, 1)$，存在一个检验ϕ和常数$0 \le c$和$0 \le r \le 1$，使得

$$\phi(x) = \begin{cases} 1, & f(x;\theta_1)/f(x;\theta_0) > c \\ r, & f(x;\theta_1)/f(x;\theta_0) = c \\ 0, & f(x;\theta_1)/f(x;\theta_0) < c \end{cases}$$

并且

$$\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$$

2. ***UMP*检验的充分条件**：若检验满足上述条件，则它是H_0对H_1在显著性水平α下的*UMP*检验。

3. ***UMP*检验的必要条件：**若$\tilde{\phi}$是H_0对H_1显著性水平α下的*UMP*检验，并且ϕ按上述定义，则几乎处处有$\tilde{\phi} = \phi$。

注记. 其中构造c和r的方法是有必要知道的：

由

$$E_{\theta_0}[\phi(X)] = P_{\theta_0}(\Lambda(X) > c) + rP_{\theta_0}(\Lambda(X) = c) = \alpha$$

并设$\Lambda(X)$的分布函数为$F(x)$，那么我们可以得到

$$1 - F(c) + r[F(c) - F(c-0)] = \alpha$$

令

$$c : max\{c|F(c) \geq 1 - \alpha\}$$
$$r : r = \frac{F(c) - (1-\alpha)}{F(c) - F(c-0)}$$

当分布函数连续的时候，r可以任意取。

方法 10.1.4. 找到*UMPT*的一般步骤

基于*N-P*引理，到 $H_0 : \theta \in \Theta_0$ 对比 $H_1 : \theta \in \Theta_1$ 的*UMP*检验的一般步骤如下：

1. 取 $\theta_0 \in \partial\Theta_0$，以及任意 $\theta_1 \in \Theta_1$。然后考虑简单假设

$$H_0' : \theta = \theta_0 \quad 对比 \quad H_1' : \theta = \theta_1$$

2. 通过*Neyman-Pearson*引理，找到 H_0' 对比 H_1' 的显著性水平为 α 的*UMP*检验 ϕ_{θ_1}。

3. 验证 ϕ_{θ_1} 与 θ_1 无关，因此 $\phi_{\theta_1} = \phi$。即，ϕ 是 H_0' 对比 H_1 的*UMP*检验。

4. 证明 ϕ 在 Θ_0 上的显著性水平为 α，那么 ϕ 就是 H_0 对比 H_1 的*UMP*检验。

例 10.1.5. 设 $X = (X_1, \cdots, X_n)$ 是来自 $N(\theta, 1)$（方差已知为*1*）的独立同分布样本。考虑以下假设检验问题

$$H_0 : \theta = 0 \quad \text{对比} \quad H_1 : \theta = \theta_1 \ (\theta_1 > 0)$$

求解该问题的显著性水平为 α 的*UMP*检验。

解. 根据Neyman-Pearson引理，检验基于似然比

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)} = \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right\}}{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\}}$$

其中 $f_0(x)$ 和 $f_1(x)$ 分别是 H_0 和 H_1 下的概率密度函数。简化后，我们得到

$$\Lambda(x) = \exp\left\{\theta_1 \sum_{i=1}^{n} x_i - \frac{n\theta_1^2}{2}\right\} = \exp\left\{n\theta_1(\bar{x} - \frac{\theta_1}{2})\right\}$$

UMPT的否定域为

$$D = \{X : \Lambda(X) > c\} = \{X : \sqrt{n}\bar{X} > c'\}$$

由$\sqrt{n}\bar{X} \sim N(0, 1)$解

$$E_0[\phi(X)] = P(\sqrt{n}\bar{X} > c' | H_0) = \alpha$$

也就是由否定域有

$$D = \{X : \sqrt{n}\bar{X} > c'\} = X : \bar{X} > \frac{u_\alpha}{\sqrt{n}}$$

所以UMPT的检验函数为

$$\varphi(x) = I\left(\bar{x} > \frac{u_\alpha}{\sqrt{n}}\right)$$

□

注记. 此处我们可以看出$\phi(x)$与θ_1的选取无关，故当$H_1 : \theta \neq 0$的时候，*UMPT*不变。

例 10.1.6. 假设我们有一个样本 $X = (X_1, \ldots, X_n)$，其中每个 X_i 都是从均匀分布 $U(0, \theta)$ 中独立同分布采样的，且 $\theta > 0$。我们想检验以下假设：

$$H_0 : \theta = \theta_0 \quad \text{对比} \quad H_1 : \theta = \theta_1 (\theta_1 > \theta_0 > 0)$$

求解显著性水平为 α 的*UMPT*。

解. 均匀分布的密度函数为

$$f(x;\theta) = \frac{1}{\theta^n} I(0 \le x \le \theta)(x_{(n)})$$

似然比为

$$\lambda(x) = \frac{f(x,\theta_1)}{f(x,\theta_2)} = \begin{cases} \left(\frac{\theta_0}{\theta_1}\right)^n & 0 < x_{(n)} < \theta_0 \\ \infty & \theta_0 < x_{(n)} < \infty \end{cases}$$

这里能够得到似然比关于$x_{(n)}$的单调性，因此我们可以设最终的检验函数有形式：

$$\varphi(x) = I(\{x_{(n)} > c\})$$

然后用期望解其中的参数c：

$$E_{\theta_0}[\varphi(X)] = P(X_{(n)} > c | H_0) = \alpha$$

$$E_{\theta_0}[\varphi(X)] = \int_0^\infty \varphi(t) \frac{n t^{n-1}}{\theta_0^n} I_{(0,\theta)}(t) dt = 1 - \frac{c^n}{\theta_0^n}$$

解得

$$\varphi(x) = I(\{x_{(n)} > \theta_0 \sqrt[n]{1-\alpha}\})$$

□

10.2 Monotone Likelihood Ratio family

注意到似然比经常能成为一个统计量的单调函数

定义 10.2.1. 如果模型族（一族*pdfs*或*pmfs*）$\{f(x,\theta):\theta\in\Theta\}$ 具有单调似然比性质，则称该模型族是统计量 T 的单调似然比（MLR）族，具体条件如下：

1. 对于 $\theta_1<\theta_2$，分布 $f(x,\theta_1)$ 和 $f(x,\theta_2)$ 是不同的，
2. 存在一个一维统计量 $T(x)$，使得比率 $f(x;\theta_2)/f(x;\theta_1)$ 是 $T(x)$ 的单调函数。

定理 10.2.2. 考虑检验问题 $H_0:\theta\le\theta_0$ 对比 $H_1:\theta>\theta_0$。假设 $T(X)$ 是参数 θ 的充分统计量，且充分统计量 T 的概率密度函数或概率质量函数族 $\{f(t;\theta):\theta\in\Theta\}$ 具有非递减的单调似然比。则对于任意 t_0，检验

$$\phi(T)=\begin{cases}1, & T>t_0\\ r, & T=t_0\\ 0, & T<t_0\end{cases}$$

是检验 H_0 对比 H_1 的显著性水平为 α $UMPT$，其中 $\alpha=E_{\theta_0}\phi(T)$。

推论 10.2.3. 在上面定理的条件下，存在检验 $H_0:\theta\ge\theta_0$ 对比 $H_1:\theta<\theta_0$ 的$UMPT$，其形式为

$$\phi(T(x))=\begin{cases}1, & T(x)<t_0\\ r, & T(x)=t_0\\ 0, & T(x)>t_0\end{cases}$$

其中 r 和 t_0 满足 $E_{\theta_0}\phi(T)=\alpha$。

例 10.2.4. 设 $X=(X_1,\cdots,X_n)$ 是一个来自 $Poisson$ 分布 $P(\lambda)$ 的样本，其中 $\lambda>0$。求：

$$H_0:\lambda\le\lambda_0 \quad \text{对比} \quad H_1:\lambda>\lambda_0$$

的显著性水平为 α 的UMP检验。

解. poi分布的概率质量函数为

$$f(x;\lambda)=\frac{e^{-n\lambda}\lambda^{T}(x)}{x_1!\cdots x_n!}$$

其中$T(X)=\sum_{i=1}^{n}X_i\sim P(n\lambda)$的似然比单调，因此我们可以假设检验函数有形式：

$$\varphi(x)=\begin{cases}1,T(x)>t_0\\ r,T(x)=t_0\\ 0,T(x)<t_0\end{cases}$$

其中c有以下不等式确定

$$E_{\lambda_0}[\varphi(T)]=P(T>c|\lambda_0)+rP(T=c|\lambda_0)=\alpha$$

$$c=inf\{c:P(T>c|\lambda_0)\leq\alpha\}$$
$$r=\frac{\alpha-P(T>c|\lambda_0)}{P(T=c|\lambda_0)}$$

即求得UMPT。　□

Lec 11 Likelihood Ratio Test

11.1 Likelihood Ratio Test

假设样本 $X = (X_1, \ldots, X_n)$ 具有概率密度函数（pdf）或概率质量函数（pmf） $f(x, \theta)$，并考虑如下假设检验：

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

其中 $\Theta = \Theta_0 \cup \Theta_1$ 是 $\mathbb{R}^k$ 的子集。

- 对于 $\Theta_0 = \{\theta_0\}$ 和 $\Theta_1 = \{\theta_1\}$，最优的检验统计量是

$$\frac{f(X, \theta_1)}{f(X, \theta_0)}$$

- 如果我们有复合假设，我们可以使用

$$\tilde{\Lambda}_n = \frac{\sup_{\theta \in \Theta_1} f(X, \theta)}{\sup_{\theta \in \Theta_0} f(X, \theta)}$$

由于临界值会大于1，所以如果我们将这个统计量替换为 $\tilde{\Lambda}_n \vee 1$，检验结果不会改变。

定义 11.1.1. *Likelihood Ratio*

记似然比（LR）为

$$\Lambda_n(x) = \left(\tilde{\Lambda}_n \vee 1\right) = \frac{\left(\sup_{\theta \in \Theta_1} f(x, \theta)\right) \vee \left(\sup_{\theta \in \Theta_0} f(x, \theta)\right)}{\sup_{\theta \in \Theta_0} f(x, \theta)} = \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} f(x, \theta)}{\sup_{\theta \in \Theta_0} f(x, \theta)} = \frac{L(\hat{\theta}_n)}{L(\hat{\theta}_R)},$$

其中 $L(\hat{\theta}_n)$ 和 $L(\hat{\theta}_R)$ 分别是完整参数空间和受限参数空间下的最大似然值。

然后，一个显著水平为 α 的似然比检验（LRT）由以下方式给出：

$$\phi(x) = \begin{cases} 1, & \Lambda_n(x) > c \\ r, & \Lambda_n(x) = c \\ 0, & \Lambda_n(x) < c \end{cases}$$

其中 c, r 满足 $\sup_{\theta \in \Theta_0} E_\theta[\phi(X)] = \alpha$。

方法 11.1.2. 一般步骤用于找到 $H_0 : \theta \in \Theta_0$ 对应的似然比检验（LRT）：

1. 确定 Θ_0 和 Θ 以及似然函数。
2. 计算 θ 的无约束最大似然估计（MLE） $\hat{\theta}_n$。
3. 计算在 Θ_0 下 θ 的有约束最大似然估计 $\hat{\theta}_R$。
4. 形成 $\Lambda_n(x)$。
5. 找到一个在 Λ_n 的范围内严格递增的函数 h，使得 $h(\Lambda_n(X))$ 在 H_0 下有一个简单形式和已知分布。因为 $h(\Lambda_n(X))$ 等价于 $\Lambda_n(X)$，所以我们通过检验统计量 $h(\Lambda_n(X))$ 来指定显著性水平为 α 的似然比检验。

例 11.1.3. 假设我们有一个样本 $X_1, X_2, \ldots, X_n$ 来自正态分布 $N(\mu, \sigma^2)$，其中 σ^2 已知，我们要检验均值 μ 是否等于某个特定值 μ_0：

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$$

解. $\Theta_0 = \{\mu_0\} \Theta = \mathbb{R}$（因为 μ 可以取任意实数）

似然函数为：

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

无约束最大似然估计 $\hat{\mu}_n$ 是样本均值：

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

在 Θ_0 下，有约束的最大似然估计就是 μ_0：

$$\hat{\mu}_R = \mu_0$$

$$L(\hat{\mu}_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \hat{\mu}_n)^2}{2\sigma^2}\right)$$
$$L(\hat{\mu}_R) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu_0)^2}{2\sigma^2}\right)$$

将这两个似然函数代入 $\Lambda_n(x)$：

$$\Lambda_n(x) = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 - \sum_{i=1}^{n}(X_i - \mu_0)^2\right)\right)$$

考虑函数 $h(x) = -2\log(x)$，其在 $\Lambda_n(x)$ 的范围内严格递增。因此我们可以使用$-2\log(\Lambda_n(X))$

将其转化为更简单的形式：

$$-2\log(\Lambda_n(X)) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2\right) = \frac{1}{\sigma^2}\left(n(\hat{\mu}_n - \mu_0)^2\right)$$

我们得到了一个具有 χ^2 分布的统计量：

$$\frac{n(\hat{\mu}_n - \mu_0)^2}{\sigma^2} \sim \chi^2(1)$$

所以，显著性水平为 α 的似然比检验为：

$$\phi(x) = \begin{cases} 1, & \frac{n(\hat{\mu}_n - \mu_0)^2}{\sigma^2} > \chi^2_{1,\alpha} \\ 0, & \text{otherwise} \end{cases}$$

其中 $\chi^2_{1,\alpha}$ 是 χ^2 分布在自由度为1，显著性水平为 α 下的临界值。 □

例 11.1.4. 设 $X_1, X_2, \ldots, X_m$ 来自 $N(\mu_1, \sigma^2)$ 和 $Y_1, Y_2, \ldots, Y_n$ 来自 $N(\mu_2, \sigma^2)$ 时，关于 $\mu_2 - \mu_1 = 0$ 的LRT。

解. 假设 H_0 和 H_1 分别为：

$$H_0 : \mu_2 - \mu_1 = 0$$

$$H_1 : \mu_2 - \mu_1 \neq 0$$

假设 $X_1, X_2, \ldots, X_m$ 来自 $N(\mu_1, \sigma^2)$，且 $Y_1, Y_2, \ldots, Y_n$ 来自 $N(\mu_2, \sigma^2)$。样本联合似然函数为：

$$L(\mu_1, \mu_2, \sigma^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu_1)^2}{2\sigma^2}\right) \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_j - \mu_2)^2}{2\sigma^2}\right)$$

取对数得到对数似然函数：

$$\ln L(\mu_1, \mu_2, \sigma^2) = -\frac{m}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(X_i - \mu_1)^2 - \frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j - \mu_2)^2$$

在原假设 H_0 下，$\mu_1 = \mu_2$。我们令 $\mu = \mu_1 = \mu_2$，对数似然函数变为：

$$\ln L(\mu, \mu, \sigma^2) = -\frac{m+n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{m}(X_i - \mu)^2 + \sum_{j=1}^{n}(Y_j - \mu)^2\right)$$

为了最大化对数似然函数，我们对 μ 和 σ^2 求导，并找到极值点：

μ 的最大似然估计：

$$\mu = \frac{\sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j}{m+n}$$

我们称之为 $\bar{Z}$，其中：

$$\bar{Z} = \frac{m\bar{X} + n\bar{Y}}{m+n}$$

σ^2 的最大似然估计：

$$\sigma^2 = \frac{1}{m+n}\left(\sum_{i=1}^{m}(X_i - \bar{Z})^2 + \sum_{j=1}^{n}(Y_j - \bar{Z})^2\right)$$

称之为 S_Z^2。

在备择假设 H_1 下，我们分别最大化 μ_1、μ_2 和 σ^2。

μ_1 和 μ_2 的最大似然估计：

$$\mu_1 = \bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

$$\mu_2 = \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

σ^2 的最大似然估计：

$$\sigma^2 = \frac{1}{m+n}\left(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2\right)$$

称之为 S_W^2。

我们现在计算似然比统计量 Λ：

$$\Lambda = \frac{\sup_{H_0} L(\mu, \mu, \sigma^2)}{\sup_{H_1} L(\mu_1, \mu_2, \sigma^2)}$$

取对数并乘以 -2：

$$-2\ln\Lambda = -2\left(\ln L(\bar{Z}, \bar{Z}, S_Z^2) - \ln L(\bar{X}, \bar{Y}, S_W^2)\right)$$

将 $\ln L(\mu, \mu, \sigma^2)$ 和 $\ln L(\mu_1, \mu_2, \sigma^2)$ 代入上式：

在 H_0 下：

$$\ln L(\bar{Z}, \bar{Z}, S_Z^2) = -\frac{m+n}{2}\ln(2\pi S_Z^2) - \frac{m+n}{2}$$

在 H_1 下：

$$\ln L(\bar{X}, \bar{Y}, S_W^2) = -\frac{m+n}{2}\ln(2\pi S_W^2) - \frac{m+n}{2}$$

因此，对数似然比为：

$$-2\ln\Lambda = -2\left(-\frac{m+n}{2}\ln(2\pi S_Z^2) - \frac{m+n}{2} - \left(-\frac{m+n}{2}\ln(2\pi S_W^2) - \frac{m+n}{2}\right)\right)$$

化简得到：

$$-2\ln\Lambda = (m+n)\left(\ln S_W^2 - \ln S_Z^2\right)$$

在大样本情况下，根据Wilks定理，$-2\ln\Lambda$ 在原假设 H_0 成立时近似服从自由度为 1 的卡方分布。因此，似然比检验的统计量是：

$$-2\ln\Lambda = (m+n)\left(\ln\left(\frac{S_W^2}{S_Z^2}\right)\right)$$

该统计量近似服从自由度为1的卡方分布，用于检验 $H_0 : \mu_2 - \mu_1 = 0$。

□

注记. 其中服从自由度为*1*的卡方分布是由被检验参数的数目差导致的。具体而言，在原假设下，有μ一个自由参数，在备择假设下有μ_1, μ_2两个自由参数，故被检验参数的数目差为*1*

11.2 Asymptotic distribution of LR*

我看不懂，仅翻译

定理 11.2.1. *Let $X_1, \ldots, X_n$ be a simple random sample from $f(x, \theta)$, $\theta \in \Theta$, where Θ is an open set of $\mathbb{R}^k$. Consider $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta - \Theta_0$, where Θ_0 is an open set of $\mathbb{R}^d$ with $d = k - s \geq 1$. Under regularity conditions (conditions for MLE to be consistent and asymptotically normal) and assuming H_0 is true, then*

$$2 \log \Lambda_n(X) \xrightarrow{L} \chi^2_{k-d} = \chi^2_s, \text{ as } n \to \infty.$$

Note: Wilks's theorem depends critically on the fact that not only is Θ open but that Θ_0 is open.

Specification on Θ_0:

1. *There exists a non-empty subset A in $\mathbb{R}^d$, $d = k - s \geq 1$, and a one-to-one mapping $h = (h_1, \ldots, h_k) : A \to \Theta_0$ where $\partial h / \partial \eta$ is full rank of d at the interior of A, that is,*

$$\Theta_0 = \{\theta = h(\eta), \eta \in A\}$$

such that,

$$\{f(x, \theta), \theta \in \Theta_0\} = \{f(x, h(\eta)), \eta \in A\}.$$

2. *Or, Θ_0 can be equivalently expressed (after some transformation) as*

$$\Theta_0 = \{\theta = (\theta_1, \ldots, \theta_k) \in \Theta : \theta_i = \theta_{i0}, i = 1, \ldots, s\}$$

and $(\theta_{10}, \ldots, \theta_{s0}, \theta_{s+1}, \ldots, \theta_k)$ is interior of Θ.

3. *Or the null can be expressed as s constraints in general:*

$$H_0 : g(\theta) = 0_{s \times 1}$$

where $g : \mathbb{R}^k \to \mathbb{R}^s$ and the rank of $\partial g(\theta) / \partial \theta$ is s.

4. *In a word, $\dim \Theta - \dim \Theta_0 = s \geq 1$.*

注记. 设 $X_1, \ldots, X_n$ 是从密度函数 $f(x, \theta)$ 中抽取的简单随机样本，其中 $\theta \in \Theta$，Θ 是 $\mathbb{R}^k$ 的开集。考虑假设检验 $H_0 : \theta \in \Theta_0$ 与 $H_1 : \theta \in \Theta - \Theta_0$，其中 Θ_0 是 $\mathbb{R}^d$ 的开集，且 $d = k - s \geq 1$。在满足正则条件（即最大似然估计是一致和渐近正态的条件）下，假设 H_0 成立，则

$$2 \log \Lambda_n(X) \xrightarrow{L} \chi^2_{k-d} = \chi^2_s, \quad \text{当 } n \to \infty.$$

注意：*Wilks*定理依赖于不仅 Θ 是开集，而且 Θ_0 也是开集这一事实。

1. 存在一个非空子集 A 在 $\mathbb{R}^d$ 中，$d = k - s \geq 1$，以及一个一一映射 $h = (h_1, \ldots, h_k) : A \to \Theta_0$，其中 $\partial h / \partial \eta$ 在 A 的内部是满秩的，即

   $$\Theta_0 = \{\theta = h(\eta), \eta \in A\}$$

   这样，

   $$\{f(x, \theta), \theta \in \Theta_0\} = \{f(x, h(\eta)), \eta \in A\}\Theta$$

2. 或者，Θ_0 可以等价地表示为（经过一些变换后）

   $$\Theta_0 = \{\theta = (\theta_1, \ldots, \theta_k) \in \Theta : \theta_i = \theta_{i0}, i = 1, \ldots, s\}$$

   且 $(\theta_{10}, \ldots, \theta_{s0}, \theta_{s+1}, \ldots, \theta_k)$ 在 Θ 的内部。

3. 或者，原假设可以表示为 s 个约束条件：

   $$H_0 : g(\theta) = 0_{s \times 1}$$

   其中 $g : \mathbb{R}^k \to \mathbb{R}^s$，且 $\partial g(\theta) / \partial \theta$ 的秩为 s。

4. 总之，$\dim \Theta - \dim \Theta_0 = s \geq 1$。

只是原封不动的翻译了一遍。

11.3 Wald Test*

Test on a single parameter

For the hypothesis $H_0 : \theta = \theta_0$, since the MLE satisfies $\sqrt{n}(\hat{\theta} - \theta) \to N(0, 1/I(\theta))$, the Wald test rejects H_0 when

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})} = n(\hat{\theta} - \theta_0)i(\hat{\theta})(\hat{\theta} - \theta_0) > \chi_1^2(\alpha).$$

Test on multiple parameters

For the hypothesis $H_0 : \theta = \theta_0, \theta \in \mathbb{R}^k$, since $\sqrt{n}(\hat{\theta}-\theta) \to N_k(0, I^{-1}(\theta))$, it follows that the Wald test rejects H_0 when

$$W_n(\theta_0) = n(\hat{\theta} - \theta_0)^T i(\hat{\theta})(\hat{\theta} - \theta_0) \geq \chi_k^2(\alpha).$$

Test on $H_0 : g(\theta) = 0_{s\times 1}$

For the hypothesis $H_0 : g(\theta) = 0_{s\times 1}$, the Wald test rejects H_0 when

$$n(g(\hat{\theta}))^T[\nabla g(\hat{\theta})i^{-1}(\hat{\theta})(\nabla g(\hat{\theta}))^T]^{-1}g(\hat{\theta}) > \chi_s^2(\alpha).$$

Note that $i(\hat{\theta})$ usually takes $-\frac{1}{n}\frac{\partial^2 \ell_n}{\partial\theta\partial\theta^T}\big|_{\theta=\hat{\theta}}$, and $I(\theta_0)$ would also do if easily computed.

单参数检验

对于假设 $H_0 : \theta = \theta_0$，由于最大似然估计满足 $\sqrt{n}(\hat{\theta}-\theta) \to N(0, 1/I(\theta))$，Wald 检验在以下情况拒绝 H_0：

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})} = n(\hat{\theta} - \theta_0)i(\hat{\theta})(\hat{\theta} - \theta_0) > \chi_1^2(\alpha).$$

多参数检验

对于假设 $H_0 : \theta = \theta_0, \theta \in \mathbb{R}^k$，由于 $\sqrt{n}(\hat{\theta} - \theta) \to N_k(0, I^{-1}(\theta))$，Wald 检验在以下情况拒绝 H_0：

$$W_n(\theta_0) = n(\hat{\theta} - \theta_0)^T i(\hat{\theta})(\hat{\theta} - \theta_0) \geq \chi_k^2(\alpha).$$

关于 $H_0 : g(\theta) = 0_{s\times 1}$ 的检验

对于假设 $H_0 : g(\theta) = 0_{s\times 1}$，Wald 检验在以下情况拒绝 H_0：

$$n(g(\hat{\theta}))^T[\nabla g(\hat{\theta})i^{-1}(\hat{\theta})(\nabla g(\hat{\theta}))^T]^{-1}g(\hat{\theta}) > \chi_s^2(\alpha).$$

注意 $i(\hat{\theta})$ 通常取 $-\frac{1}{n}\frac{\partial^2 \ell_n}{\partial\theta\partial\theta^T}\big|_{\theta=\hat{\theta}}$，并且如果容易计算，$I(\theta_0)$ 也可以使用。

例 11.3.1. *Find the Wald test for the hypothesis* $H_0 : \theta \in \Theta_0 = \{\theta : \theta_j = \theta_{0,j}, j = 1, \ldots, s\}$.

The H_0 *can be rewritten as*

$$H_0 : g(\theta) = \theta^{(1)} - \theta_0^{(1)} = 0,$$

where we write the MLE for $\theta \in \Theta$ *as* $\hat{\theta}_n = (\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)})$, *where* $\hat{\theta}_n^{(1)} = (\hat{\theta}_1, \ldots, \hat{\theta}_s)$ *and* $\hat{\theta}_n^{(2)} = (\hat{\theta}_{s+1}, \ldots, \hat{\theta}_k)$. *We define the Wald statistic as*

$$W_n(\theta_0^{(1)}) = n(\hat{\theta}_n^{(1)} - \theta_0^{(1)})^T[I_{11}(\hat{\theta}_n)]^{-1}(\hat{\theta}_n^{(1)} - \theta_0^{(1)}) \xrightarrow{L} \chi_s^2, \text{ under } H_0,$$

where $I_{11}(\theta)$ *is the upper diagonal block of* $I^{-1}(\theta)$ *written as*

$$I^{-1}(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}.$$

注记. 找到假设 $H_0 : \theta \in \Theta_0 = \{\theta : \theta_j = \theta_{0,j}, j = 1, \ldots, s\}$ 的 *Wald* 检验。

该 H_0 可以重写为

$$H_0 : g(\theta) = \theta^{(1)} - \theta_0^{(1)} = 0,$$

其中我们将 $\theta \in \Theta$ 的最大似然估计记为 $\hat{\theta}_n = (\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)})$，其中 $\hat{\theta}_n^{(1)} = (\hat{\theta}_1, \ldots, \hat{\theta}_s)$ 和 $\hat{\theta}_n^{(2)} = (\hat{\theta}_{s+1}, \ldots, \hat{\theta}_k)$。我们定义 *Wald* 统计量为

$$W_n(\theta_0^{(1)}) = n(\hat{\theta}_n^{(1)} - \theta_0^{(1)})^T[I_{11}(\hat{\theta}_n)]^{-1}(\hat{\theta}_n^{(1)} - \theta_0^{(1)}) \xrightarrow{L} \chi_s^2, \text{ 在 } H_0 \text{ 下},$$

其中 $I_{11}(\theta)$ 是 $I^{-1}(\theta)$ 的左上角块，表示为

$$I^{-1}(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}.$$

11.4 Score Test*

1. For the simple hypothesis $H_0 : \theta = \theta_0$, Rao' s score test or Lagrange Multipliers test is based on the observation that, letting H be the Lagrangian:

   $$H = \ell_n(\theta) - \lambda'(\theta - \theta_0),$$

   the first-order conditions are:

   $$\frac{\partial \ell_n}{\partial \theta} = \lambda; \quad \theta = \theta_0,$$

   so $\lambda/\sqrt{n} = U_n(\theta_0) \xrightarrow{L} N(0, I(\theta_0))$. Thus, $LM_n = \lambda^T I^{-1}(\theta_0)\lambda/n \xrightarrow{L} \chi^2_k$ under H_0.

2. Equivalently, it follows from this that under H_0, as $n \to \infty$,

   $$LM_n(\theta_0) = U_n^T(\theta_0) I^{-1}(\theta_0) U_n(\theta_0) \xrightarrow{L} \chi^2_k.$$

   The test rejects H_0 when $LM_n(\theta_0) \geq \chi^2_k(\alpha)$.

3. For test on $H_0 : g(\theta) = 0_{s\times 1}$, the statistic employed in the score test is based on the ML estimate $\hat{\theta}_n^R$ that is obtained from the solution of the constrained optimization problem

   $$\hat{\theta}_n^R = \arg \max_{\theta \in H_0} \ell_n(\theta).$$

The test statistic, called score statistic, is

$$LM_n = U_n(\hat{\theta}_n^R)^\top I^{-1}(\hat{\theta}_n^R)U_n(\hat{\theta}_n^R) \xrightarrow{L} \chi_s^2, \quad \text{under } H_0.$$

4. Note that $I^{-1}(\theta_0)$ depends on the unknown parameter value θ_0. We can evaluate the terms involved in θ_0 at the constrained MLE, $\hat{\theta}_n^R$ to get a usable statistic. We can approximate $I(\theta_0)$ with either $-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n^R)$ or $\frac{1}{n}\nabla\ell_n(\hat{\theta}_n^R)\nabla^T\ell_n(\hat{\theta}_n^R)$.

1. 对于简单假设 $H_0 : \theta = \theta_0$，Rao的得分检验或拉格朗日乘数检验基于以下观察：令 H 为拉格朗日函数：

$$H = \ell_n(\theta) - \lambda'(\theta - \theta_0),$$

一阶条件为：

$$\frac{\partial \ell_n}{\partial \theta} = \lambda; \quad \theta = \theta_0,$$

因此 $\lambda/\sqrt{n} = U_n(\theta_0) \xrightarrow{L} N(0, I(\theta_0))$。因此，在 H_0 下，$LM_n = \lambda^T I^{-1}(\theta_0)\lambda/n \xrightarrow{L} \chi_k^2$。

2. 等价地，由此可得在 H_0 下，当 $n \to \infty$ 时，

$$LM_n(\theta_0) = U_n^T(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) \xrightarrow{L} \chi_k^2\Theta$$

当 $LM_n(\theta_0) \geq \chi_k^2(\alpha)$ 时，检验拒绝 H_0。

3. 对于 $H_0 : g(\theta) = 0_{s\times 1}$ 的检验，得分检验中使用的统计量基于通过以下约束优化问题的解得到的最大似然估计 $\hat{\theta}_n^R$：

$$\hat{\theta}_n^R = \arg\max_{\theta \in H_0} \ell_n(\theta)\Theta$$

检验统计量称为得分统计量：

$$LM_n = U_n(\hat{\theta}_n^R)^\top I^{-1}(\hat{\theta}_n^R)U_n(\hat{\theta}_n^R) \xrightarrow{L} \chi_s^2\text{fi} \quad \text{在 } H_0 \text{ 下}\Theta$$

4. 注意 $I^{-1}(\theta_0)$ 依赖于未知参数值 θ_0。我们可以在约束最大似然估计 $\hat{\theta}_n^R$ 处评价涉及 θ_0 的项，以获得可用的统计量。我们可以用 $-\frac{1}{n}\nabla^2\ell_n(\hat{\theta}_n^R)$ 或 $\frac{1}{n}\nabla\ell_n(\hat{\theta}_n^R)\nabla^T\ell_n(\hat{\theta}_n^R)$ 来近似 $I(\theta_0)$。

11.5 Confidence Regions*

1. Any of the tests discussed above (LRT, Wald, score) can be inverted to find confidence regions for the values of l parametric functions $u = (u_1(\theta), u_2(\theta), \ldots, u_l(\theta))'$.

2. That is, for any vector of potential values for these functions $c = (c_1, c_2, \ldots, c_l)'$, one may define $g_{c,i}(\theta) = u_i(\theta) - c_i$ and a test of some type above for

$$H_0 : g_c(\theta) = u - c = 0$$

The set of all c for which an approximately α-level test does not reject constitutes an approximately $(1-\alpha)\times 100\%$ confidence set for the vector $(u_1(\theta), u_2(\theta), \ldots, u_l(\theta))'$.

3. When Wald tests are used, the regions will be ellipsoidal:

$$CR = \{c : [c - u(\hat{\theta})]' Var^{-1}(u(\hat{\theta}))[c - u(\hat{\theta})] \leq \chi_l^2(\alpha)\}$$

1. 以上讨论的任何检验（LRT，Wald，得分检验）都可以反转来找到 l 个参数函数 $u = (u_1(\theta), u_2(\theta), \ldots, u_l(\theta))'$ 的值的置信区域。

2. 即，对于这些函数的任何潜在值向量 $c = (c_1, c_2, \ldots, c_l)'$，可以定义 $g_{c,i}(\theta) = u_i(\theta) - c_i$ 并进行某种类型的检验

$$H_0 : g_c(\theta) = u - c = 0$$

对于所有使得一个近似的 α 水平检验不拒绝的 c 的集合，构成了向量 $(u_1(\theta), u_2(\theta), \ldots, u_l(\theta))'$ 的一个近似 $(1-\alpha)\times 100\%$ 的置信集。

3. 当使用 Wald 检验时，这些区域将是椭圆形的：

$$CR = \{c : [c - u(\hat{\theta})]' Var^{-1}(u(\hat{\theta}))[c - u(\hat{\theta})] \leq \chi_l^2(\alpha)\}$$

Lec 12 Goodness-of-fit Tests

“拟合优度检验”是一种确定一组分类数据是否来自声称的分布的方法。原假设是数据来自该分布，对立假设是数据不来自该分布。

- $H_0 : X \sim F$
  - 如果 F 是离散分布：使用Pearson χ^2 检验
  - 如果 F 是连续分布：使用Kolmogorov检验
  - 两个样本情况：使用Smirnov检验
- $H_0 : F_1(x) = F_2(x)$

12.1 Pearson χ^2 Test

方法 12.1.1. *1.* 制定原假设 H_0 和似然函数。

2. 在 H_0 下计算参数 η 的最大似然估计 $\hat{\eta}$。

3. 计算期望频数 $E_i = np_i(\hat{\eta})$，合并单元格以确保 $E \geq 5$。

4. 构造检验统计量

$$\chi^2 = \sum \frac{(O-E)^2}{E}\Theta$$

5. 将卡方值与临界值比较或计算 p 值，最终决定是否拒绝原假设。

方法 12.1.2. 分类数据的卡方检验

设总体X可分为有限r类：$A_1, A_2, \ldots, A_r$，需要检验的零假设为H_0 : $P(X = A_i) = p_i, i = 1, 2, \ldots, r$，备择假设为$H_1 : \exists i, P(X = A_i) \neq p_i$。其中$p_i$是已知的。

现抽取容量为n的样本，考察各类出现的频数$n_1, n_2, \ldots, n_r$. $\sum_{i=1}^{r} n_i = n$。检验概率是否符合要求。

自然的想法是，如果零假设成立，那么每一个频率n_i/n应该与p_i接近。我们可以使用卡方统计量来检验这个假设。

那么我们令

$$u = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{r-1}$$

这是依分布渐进收敛到后面的分布的，在大样本下，可以近似认为成立。

因此拒绝域可以取为

$$\{u > \chi^2_{r-1,\alpha}\}$$

例 12.1.3. 骰子：假设骰子是均匀的，抛了*600*次，各个面出现的次数为*100, 110, 90, 150, 120, 130*。检验骰子是否均匀，显著性水平设为*0.05*。

解.

$$H_0 : P(X = i) = 1/6, i = 1, 2, \ldots, 6. H_1 : \exists i, P(X = i) \neq 1/6$$

$$u = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = 10. \qquad \chi^2_{5,0.05} = 11.07 > 10.8$$

因此我们不能拒绝原假设。 □

方法 12.1.4. 分类数据的卡方检验（含未知参数）

设总体X可分为有限r类：$A_1, A_2, \ldots, A_r$，需要检验的零假设为H_0 : $P(X = A_i) = p_i, i = 1, 2, \ldots, r$，备择假设为$H_1 : \exists i, P(X = A_i) \neq p_i$。其中$p_i = p_i(\theta_1, \theta_2, \cdots, \theta_m)$依赖$m$个未知参数的未知概率。

现抽取容量为n的样本，考察各类出现的频数$n_1, n_2, \ldots, n_r$. $\sum_{i=1}^{r} n_i = n$。检验概率是否符合要求。

对于这种情况，我们可以将未知参数用其估计代替，然后使用卡方统计量进行检验。那么我们令

$$u = \sum_{i=1}^{r} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2_{r-1-m}$$

因此拒绝域可以取为

$$\{u > \chi^2_{r-1-m,\alpha}\}$$

例 12.1.5. 后代

一种动植物杂交后代的分布是A_1, A_2, A_3, A_4，其概率分别为$(p-1)^2, p^2, (1-p)p, (1-p)p$。现在抽取了*1000*个后代，各类出现的频数为*200, 300, 250, 250*。检验这个分布是否合理，显著性水平设为*0.05*。

解.

$$\hat{p} = \underset{p}{argmax}\{p^{1050}(1-p)^{750}\} = 0.55$$

$$u = \sum_{i=1}^{r} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2_{r-1-m} = 0.102 \qquad \chi^2_{2,0.05} = 0.103 > 0.102$$

因此我们不能拒绝原假设。 □

方法 12.1.6. 分布拟合的卡方检验

想要判断总体X是否服从某分布F，零假设为$H_0 : X \sim F$，备择假设为$H_1 : X \nsim F$。其中F是可以是已知的，也可以是$F = F(\theta_1, \cdots, \theta_m)$依赖$m$个未知参数的未知分布。

当分布是离散的时候，可以仍然借用上述分类然后计算区间的方法，如果分布是连续的，我们可以将分布函数的取值范围划分为数个区间：

$$A_1 = (-\infty, x_1], A_2 = (x_1, x_2], \cdots, A_r = (x_{r-1}, +\infty)$$

然后我们可以计算每个区间的频数n_i，然后使用卡方统计量进行检验。

12.2 Test of independence and homogeneity

方法 12.2.1. 独立性检验

设有两个分类变量X和Y，X有r个类别，Y有s个类别。我们需要检验X和Y是否独立，即$H_0 : P(X = A_i, Y = B_j) = P(X = A_i)P(Y = B_j)$。

我们可以使用卡方统计量来检验这个假设。我们令

$$\hat{K}^* == \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} \sim \chi^2_{(r-1)(s-1)}$$

因此拒绝域可以取为

$$\{u > \chi^2_{(r-1)(s-1),\alpha}\}$$

例 12.2.2. 一项调查中，有*1000*人参与，其中男性有*600*人，女性有*400*人。调查结果显示，男性中有*400*人喜欢足球，*200*人喜欢篮球；女性中有*200*人喜欢足球，*200*人喜欢篮球。检验性别和喜欢的运动是否独立，显著性水平设为*0.05*。

解.

$$\hat{K}^* == \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} \sim \chi^2_{(r-1)(s-1)} = 25.8. \qquad \chi^2_{1,0.05} = 0.004 < 25.8$$

因此我们拒绝原假设，性别和喜欢的运动不独立。 □

方法 12.2.3. 齐一性检验

设有c个分类变量X，有r个类别，我们需要检验这r个类别的分布是否相同，即$H_0 : p_{1j} = \cdots = p_{rj}, j = 1, \cdots, c$其中$p_{ij}$是第$j$个分类变量取第$i$类时候的概率。同时定义$n_{ij}$为第$j$个分类变量取第$i$类的频数。

我们可以使用卡方统计量来检验这个假设。我们令

$$\hat{K}^* = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \sim \chi^2_{(r-1)(s-1)}$$

12.3 Kolmogorov-Smirnov Test

上述卡方检验想要检验连续分布的时候分划了区间，但这不准确，因此我们可以使用Kolmogorov-Smirnov检验。

方法 12.3.1. *Kolmogorov-Smirnov*检验

设总体X的分布函数为$F(x)$，我们需要检验$H_0 : X \sim F(x)$，备择假设为$H_1 : X \nsim F(x)$。

我们可以使用*Kolmogorov-Smirnov*统计量来检验这个假设。我们令

$$D_n = \max_{|1\le i\le n}\{F_n(x_i) - F(x_i)|\}$$

其中$F_n(x_i)$是经验分布函数，$F(x_i)$是理论分布函数。

经验分布函数的定义为：

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

其拒绝域的形式为：

$$\{D_n > D_{(n,\alpha)}\}$$

其中$D_{(n,\alpha)} = \lambda_\alpha/\sqrt{n}$是临界值。

例 12.3.2. 一组数据为：*1, 2, 3, 4, 5, 6, 7, 8, 10, 10*。检验这组数据是否服从均匀分布，显著性水平设为*0.05*。

解.

$$D_n = \max_{1\le i\le n}\{|F_n(x_i) - F(x_i)|\} = 0.1$$

$$D_{(10,0.05)} = 0.409 > 0.1$$

因此我们不能拒绝原假设。　□

例 12.3.3. 一组数据为：（*1*到*100*，但是没有*90*到*99*，而是有十个*100*）。检验这组数据是否服从均匀分布，显著性水平设为*0.05*。

解.

$$D_n = \max_{1\le i\le n}\{|F_n(x_i) - F(x_i)|\} = 0.1$$

$$D_{(100,0.05)} = \lambda_0.95/10 = 0.136 > 0.1$$

因此我们不能拒绝原假设。 □

方法 12.3.4. *Smirnov*检验：两个样本的检验

设有两个总体X和Y，我们需要检验$H_0 : F_X(x) = F_Y(x)$，备择假设为$H_1 : F_X(x) \neq F_Y(x)$。

和上面的思路其实一样，我们设X和Y的经验分布函数为$G_X(x)$和$G_Y(x)$，我们可以使用*Smirnov*统计量来检验这个假设。我们令

$$D_{(X,Y)} = \max_{|1\le i\le n}\{|G_X(x_i) - G_Y(x_i)|\}$$

（如果是单边问题，那么此处检验统计量不加绝对值）

其拒绝域的形式为：

$$\{D_{(X,Y)} > D_{(n,\alpha)}\}$$

12.4 Coursewareflash

思来想去还是把课件敲了一份放在这里，以供观瞻。

12.4.1 Pearson' s χ^2 Test

Goodness-of-Fit in a Multinomial Model.

Let $X_1, \ldots, X_n$ be a simple sample from population

$$X \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_r \\ p_1 & p_2 & \cdots & p_r \end{pmatrix}$$

where $p_i \geq 0$, $p_1 + \cdots + p_r = 1$. Because $p_r = 1 - \sum_{j=1}^{r-1} p_j$, we consider the parameter $\theta = (p_1, \ldots, p_{r-1})^T$ and we test the following hypothesis

$$H_0 : p_j = p_{0j},\ j = 1, \ldots, r,\ \text{i.e.,}\ H_0 : \theta = \theta_0$$

for specified $\theta_0 = (p_{01}, \ldots, p_{0r-1})^T$.

The likelihood function is

$$L_n(\theta) = L_n(p_1, \ldots, p_{r-1}) = p_1^{n_1} \cdots p_{r-1}^{n_{r-1}} (1 - \cdots - p_{r-1})^{n_r}$$

and the MLE $\hat{p}_j = \frac{n_j}{n}$, where $n_j = \sum_{i=1}^{n} 1\{X_i = a_j\}$.

The Wald test: Let $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_{r-1})^T$, then the Wald test rejects $H_0 : \theta = \theta_0$ when

$$W_n(\theta_0) = n(\hat{\theta}_n - \theta_0)^T I(\theta_0)(\hat{\theta}_n - \theta_0) > \chi^2_{r-1}(\alpha)$$

has asymptotic level α under H_0.

The Fisher information $I(\theta) = (I_{ij})$ with

$$I_{ij} = -E\left[\frac{\partial^2}{\partial p_i \partial p_j} \sum_{k=1}^{r} I(X = a_k) \log p_k\right] = \begin{cases} \frac{1}{p_i} + \frac{1}{p_r}, & i = j \\ \frac{1}{p_r}, & i \neq j \end{cases}$$

Thus,

$$W_n(\theta_0) = n \sum_{j=1}^{r-1} \frac{[\hat{p}_j - p_{0j}]^2}{p_{0j}} + n \sum_{j=1}^{r-1} \sum_{i=1}^{r-1} \frac{(\hat{p}_i - p_{0i})(\hat{p}_j - p_{0j})}{p_{0r}} = n \sum_{j=1}^{r} \frac{[\hat{p}_j - p_{0j}]^2}{p_{0j}} = \sum_{j=1}^{r} \frac{[n_j - np_{0j}]^2}{np_{0j}}$$

The term on the right is called Pearson' s chi-square (χ^2) statistic and is the statistic that is typically used for this multinomial testing problem. It is easily remembered as

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

The LRT (called G-test)

$$G = 2\log\Lambda_n(X) = 2\sum_{j=1}^{r} n_j \log\left(\frac{n_j}{np_{0j}}\right) = \sum O_i \log\left(\frac{O_i}{E_i}\right) > \chi^2_{r-1}(1-\alpha)$$

The score test

$$LM_n(\theta_0) = U_n^T(\theta_0) I^{-1}(\theta_0) U_n(\theta_0) > \chi^2_{r-1}(\alpha)$$

where $U_n(\theta) = \frac{1}{\sqrt{n}}\nabla\ell_n(\theta)$. It follows that the Rao statistic equals Pearson's χ^2.

定理 12.4.1. ***Goodness-of-Fit to Composite Multinomial Models. Contingency Tables***

We test the hypothesis

$$H_0 : \theta = \theta(\eta) = (p_1(\eta), \ldots, p_{r-1}(\eta))^T$$

where $\eta = (\eta_1, \ldots, \eta_s)^T$ *with* $s < r-1$ *is the unknown parameter.*

LRT *If a maximizing value,* $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_s)$ *exists, the log likelihood ratio is given by*

$$G = 2\log\Lambda(n_1, \ldots, n_k) = \sum_{i=1}^{r} n_i \log\frac{n_i}{n\theta_i(\hat{\eta})} = \sum O_i \log\frac{O_i}{E_i}$$

then reject H_0 *when* $G > \chi^2_{r-1-s}(\alpha)$.

Score test

$$LM_n(\theta(\hat{\eta})) = \sum_{j=1}^{r} \frac{[n_j - np_j(\hat{\eta})]^2}{np_j(\hat{\eta})} > \chi^2_{r-1-s}(\alpha)$$

Notice that the score test statistic is (see Lecture 11)

$$LM_n(\theta(\hat{\eta})) = U_n^T(\hat{\theta}_R) I^{-1}(\hat{\theta}_R) U_n(\hat{\theta}_R)$$

and

$$U_n(\theta) = \frac{1}{\sqrt{n}}\left[\frac{n_1}{p_1} - \frac{n_r}{p_r}, \ldots, \frac{n_{r-1}}{p_{r-1}} - \frac{n_r}{p_r}\right]^T$$

$$I^{-1}(\theta) = [A_{r-1} + \frac{1}{p_r}11^T]^{-1} = diag(\theta) - \theta\theta^T, \text{ where } A_{r-1} = diag\left(\frac{1}{p_1}, \ldots, \frac{1}{p_{r-1}}\right).$$

(using fact $(A + bc^T)^{-1} = A^{-1} - A^{-1}bc^TA^{-1}/(1 + c^TA^{-1}b)$ *)*

Wald test *the Wald statistic is also equal to Pearson' s* χ^2 *test statistic.*

WLOG, we find a one-to-one mapping $(p_1, \ldots, p_{r-1})$ *to* $(\theta'_1, \ldots, \theta'_{r-1})$ *such that* H_0 *can be equivalently expressed as*

$$H'_0 : \theta'_j = 0, j = s+1, \ldots, r-1.$$

12.4.2 Test of independence

Suppose that variable A has r levels, and variable B has c levels. We consider the following hypothesis:

$$H_0 : A \text{ and } B \text{ are independent.}$$

$$H_a : A \text{ and } B \text{ are not independent.}$$

Equivalently,

$$H_0 : P(A = i, B = j) = p_{ij} = P(A = i)P(B = j) = u_iv_j$$

where $\{p_{ij}\}, \{u_i\}, \{v_j\}$ are pmfs. That is,

$$H_0 : p_{ij} = p_{ij}(\eta) = u_iv_j$$

where $\eta = (u_1, \ldots, u_{r-1}, v_1, \ldots, v_{c-1})$ by noting $\sum u_i = 1, \sum v_j = 1$.

Thus the Pearson Chi-square test is applicable with df $= rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1)$.

$$H_0 : A \text{ and } B \text{ are independent}$$

Data :

$A\backslash B$	1	$\cdots$	c	Total
1	n_{11}	$\cdots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rc}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot c}$	n

Test statistic:

$$T = \sum \frac{(O - E)^2}{E} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n}$$

12.4.3 Test of homogeneity

In the test of homogeneity, we consider two or more populations (P) and a single categorical variable (B). e.g.,

populations

$P\backslash B$	1	$\cdots$	c	sum
1	p_{11}	$\cdots$	p_{1c}	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	p_{r1}	$\cdots$	p_{rc}	1

Data

$P\backslash B$	1	$\cdots$	c	sum
1	n_{11}	$\cdots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rc}	$n_{r\cdot}$

and consider the hypothesis:

H_0 : The distributions of the r populations are the same., i.e.,

$$H_0 : p_{1j} = \cdots = p_{rj}, \quad j = 1, \ldots, c$$

If we treat P as a categorical variable, then the above problem is testing for independence.

例 12.4.2. ***Fisher' s exact test***

The tests discussed so far that use the chi-square approximation, and perform well when the contingency tables have a reasonable number of observations in each cell.

When samples are small, we can perform inference using an exact distribution (or estimates of exact distributions).

Fisher' s exact test: test association between two characteristics in 2×2 *contingency table. But in this case both row and column totals are assumed to be fixed - not random.*

Steps:

Work out all the 2×2 *tables that deviate from the expected values more than your sample.*

Use Fisher' s formula to calculate the probability of each of these 2×2 *tables.*

	Milk first	Tea first	Totals
Lady's guess: Milk	n_{11}	n_{12}	n_{1+}
Lady's guess: Tea	n_{21}	n_{22}	n_{2+}
Totals	n_{+1}	n_{+2}	n

表 12.1: Contingency table for Fisher's exact test

a	b	n_{1+}
c	d	n_{2+}
n_{+1}	n_{+2}	n

表 12.2: Another form of contingency table for Fisher's exact test

$$P(a=t)=\frac{\binom{n_{+1}}{t}\binom{n_{+2}}{n_{1+}-t}}{\binom{n}{n_{+1}}} \quad t=0,1,\ldots,n_{+1}$$

Add up the probabilities of the tables as extreme or more extreme than that observed, and (for a 2-tailed test) double that summed probability.

12.4.4 Kolmogorov-Smirnov Tests

Kolmogorov test

Let $X_1,\ldots,X_n$ be a simple sample from population $X\sim F$, we test the following hypothesis

$$H_0: F(x)=F_0(x)$$

where $F_0(x)$ is some known distribution.

Let F_n be the empirical distribution function based on $X_1,\ldots,X_n$. One form of the Kolmogorov test statistic for H_0 is

$$D_n := \sup_x |F_n(x)-F_0(x)|.$$

The p-value is

$$\text{p-value} = P(D_n \geq D_{\text{obs}}|H_0)$$

Theorem 1. If $X_1, X_2, \ldots,$ are i.i.d. with the continuous distribution function F_0 then 1. The null distribution of D_n does not depend on F_0; it depends only on n. 2. If $n \to \infty$ the distribution of $\sqrt{n}D_n$ is asymptotically Kolmogorov' s distribution with the c.d.f.

$$K(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}, \quad x > 0$$

that is

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq x) = K(x)$$

Smirnov test

Let $X_{11}, \ldots, X_{1n_1}$ be i.i.d with a c.d.f. F_1, and $X_{21}, \ldots, X_{2n_2}$ i.i.d with a c.d.f. F_2. We are interested in testing the null hypothesis of the form

$$H_0 : F_1(x) = F_2(x) \text{ for all } x$$

Let $F_{kn_k}(x) = \frac{1}{n_k}\sum_{i=1}^{n_k} I(X_{ki} \leq x)$, $k = 1, 2$, be the empirical distribution functions, then consider the following statistics

$$D_{n_1,n_2} = \sup_{-\infty<x<\infty} |F_{1n_1}(x) - F_{2n_2}(x)| = \min\{D^+_{n_1,n_2}, D^-_{n_1,n_2}\}$$

$$D^+_{n_1,n_2} = \sup_{-\infty<x<\infty} (F_{1n_1}(x) - F_{2n_2}(x))$$

$$D^-_{n_1,n_2} = \sup_{-\infty<x<\infty} (F_{2n_2}(x) - F_{1n_1}(x))$$

Theorem 2 (Smirnov). Under the null hypothesis H_0, we have

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D^+_{n_1,n_2} \leq x\right) = 1 - e^{-2x^2}, \quad x > 0$$

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D_{n_1,n_2} \leq x\right) = K(x)$$

where $K(x)$ is the Kolmogorov' s distribution given by (12.1).

Test for $H_0 : F_1 = F_2 \leftrightarrow H_0 : F_1 \neq F_2$: Reject H_0 if $D_{n_1,n_2} > \lambda\sqrt{\frac{n_1 n_2}{n_1+n_2}}$

Test for $H_0 : F_1 \leq F_2 \leftrightarrow H_0 : F_1 > F_2$: Reject H_0 if $D^+_{n_1,n_2} > \lambda_1\sqrt{\frac{n_1 n_2}{n_1+n_2}}$, $\lambda_1 = \sqrt{-\ln \alpha/2}$

Test for $H_0 : F_1 \geq F_2 \leftrightarrow H_0 : F_1 < F_2$: Reject H_0 if $D^-_{n_1,n_2} > \lambda_1\sqrt{\frac{n_1 n_2}{n_1+n_2}}$, $\lambda_1 = \sqrt{-\ln \alpha/2}$

Lec 13 Bayesian Inference

后验分布计算，贝叶斯点估计与区间估计，贝叶斯假设推断

13.1 Characteristics of Bayesian inference

了解后验分布怎么计算，以及先验分布的选择对后验分布的影响。可以直接看例题。

1. **概率的主观定义**：贝叶斯推断依赖于概率的主观定义。这意味着不同的人可以对一个给定事件 A 有各自的概率 $P(A)$。这被称为先验概率。

2. **利用贝叶斯定理更新信念**：当有新的证据（事件 B）出现时，我们使用贝叶斯定理来更新对事件 A 的信念：

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

3. **参数的先验分布**：对于参数空间 Θ 中的参数 θ，我们有一个先验分布 $\pi(\theta)$。

4. **用样本更新信念**：给定一个样本 x，我们使用贝叶斯定理更新对 θ 的信念：

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}$$

其中 $f(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$。这意味着：

$$\text{后验} \propto \text{似然} \times \text{先验}$$

5. **后验分布**：贝叶斯推断基于参数 θ 的后验分布 $\pi(\theta|x)$。

注记. 其中重要的是第4点，即后验分布的计算。下面的那个正比符号也相当重要，通常我们只用求出来分母对θ正比关系然后归一化就可以了，下面也会给出例题来进行说明。

其中可能遇到的两个困难：

指定先验分布：主要挑战之一是指定先验分布 $\pi(\theta)$。先验的选择会显著影响后验分布，尤其是在数据有限的情况下。

计算边际分布：另一个挑战是计算边际分布 $f(x)$，这涉及对整个参数空间 Θ 进行积分。这可能需要复杂的数值方法或近似方法。

当我们希望在没有先验知识的情况下保持客观时，可以使用以下几种非信息化先验分布：

- **均匀先验**： $\pi(\theta) \propto \frac{1}{|\Theta|}$。在有限维度问题中可以使用，但如果参数空间$\Theta$是无界的，那么这种先验分布是不适当的（即 $\int_\Theta \pi(\theta)d\theta = \infty$）。
- **不适当先验**： $\pi(\theta) \geq 0, \int_\Theta \pi(\theta)d\theta = \infty$，使得$\pi(\theta|x)$是适当的。
  - 对于位置参数：$\pi(\theta) \equiv 1$
  - 对于尺度参数：$\pi(\sigma) \propto \frac{1}{\sigma}$
- **Jeffreys先验**： 其密度函数与Fisher信息矩阵的行列式的平方根成正比，

$$\pi(\theta) \propto \sqrt{\det I(\theta)}$$

这种先验在重新参数化下是不变的。

- 其他非信息化先验

例 13.1.1. 先验分布的选择对后验分布的影响

假设有一枚硬币，正面向上概率为θ,未知，抛*10*次，正面向上*7*次，求在

（*1*）*Beta(2,2)*

（*2*）*U(0,1)*

的先验条件下，后验分布。

解. （1） Beta(2,2) 贝叶斯定理：

$$\pi(\theta|X) \propto L(X|\theta)\pi(\theta)$$

似然函数

$$L(X|\theta) = \binom{10}{7}\theta^7(1-\theta)^3 \propto \theta^7(1-\theta)^3$$

先验分布

$$\pi(\theta) = \frac{1}{B(2,2)}\theta(1-\theta) \propto \theta(1-\theta)$$

因此后验分布：

$$\pi(\theta|X) \propto \theta^8(1-\theta)^4$$

归一化：（因为与beta分布只差了一个系数，取正比然后归一化之后就得到了Beta分布)

$$\pi(\theta|X) = Beta(\theta; 9, 5)$$

（2）后验分布 $\pi(\theta \mid X)$ 为：

$$\pi(\theta \mid X) \propto L(X \mid \theta) \cdot \pi(\theta)$$

由于先验分布是均匀分布，即 $\pi(\theta) = 1$，似然函数

$$L(X \mid \theta) = \binom{10}{7}\theta^7(1-\theta)^3$$

所以后验分布为：

$$\pi(\theta \mid X) \propto \theta^7(1-\theta)^3$$

归一化后，后验分布是一个 Beta 分布：

$$\pi(\theta \mid X) = \text{Beta}(\theta; 8, 4)$$

□

注记. *1.*先验分布通常是我们对样本的初始信念，比如这枚硬币之前已经测试过很多次了，这枚硬币接近平衡我们就选择*Beta(2,2)*；当我们认为硬币平衡的时候，我们就选择均匀假设（这或许解释了课件里*13.2*节在说什么）

*2.*当有了数据之后，对θ的不确定性减少了，表示了我们对θ的信息增加了。其他检验方法没有将样本纳入考虑之中，比如*2*次有*1*次正面，和*1000*次有*500*次正面，用点估计算出来的θ是一个定值*0.5.*但后验分布算出来是一个分布，随着实验次数增加而越来越集中于*0.5*，这符合我们的认知，毕竟抛两次和抛*1000*次能说明θ的能力直观上是不同的。

*3.*似然函数写L，先验和后验函数写π，觉得挺不错的，可以方便自己不要搞混。

例 13.1.2. 假设$X_1, X_2, \cdots, X_n$ 为从总体$N(\theta, \sigma^2)$中取出的随机样本，θ的先验分布$\pi(\theta)$为（*1*）正态分布；（*2*）均匀分布；试求后验分布$\pi(\theta|X)$

解. (1)θ 的先验分布为正态分布

先验分布：假设先验分布为 $\theta \sim N(\mu_0, \tau_0^2)$。

观测数据：我们有 $X_1, X_2, \ldots, X_n$ 来自 $N(\theta, \sigma^2)$。

似然函数：

$$L(\theta; X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\theta)^2}{2\sigma^2}\right)$$

简化后为：

$$L(\theta; X) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2\right)$$

后验分布：

$$p(\theta \mid X) \propto p(X \mid \theta)\cdot p(\theta) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2\right)\cdot\exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

进一步整理：

$$p(\theta \mid X) \propto \exp\left(-\frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta\sum_{i=1}^{n}X_i\right) - \frac{1}{2\tau_0^2}(\theta^2 - 2\mu_0\theta + \mu_0^2)\right)$$

$$p(\theta \mid X) \propto \exp\left(-\frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)\theta^2 - 2\theta\left(\frac{\sum_{i=1}^{n}X_i}{\sigma^2}+\frac{\mu_0}{\tau_0^2}\right)\right)\right)$$

通过比较形式，可以看出这是一个正态分布：

$$p(\theta \mid X) \sim N\left(\frac{\frac{\sum_{i=1}^{n}X_i}{\sigma^2}+\frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}\right)$$

所以，后验分布为：

$$\theta \mid X \sim N\left(\frac{\frac{\sum_{i=1}^{n}X_i}{\sigma^2}+\frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}\right)$$

(2)θ 的先验分布为均匀分布

先验分布：假设先验分布为均匀分布 $\theta \sim U(a,b)$。

观测数据：同样，我们有 $X_1, X_2, \ldots, X_n$ 来自 $N(\theta, \sigma^2)$。

似然函数：与前面相同，总似然函数为：

$$L(\theta; X) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2\right)$$

后验分布：根据贝叶斯定理，后验分布 $p(\theta \mid X) \propto p(X \mid \theta)p(\theta)$。

由于先验分布 $p(\theta)$ 是均匀分布 $U(a,b)$，即 $p(\theta) = \frac{1}{b-a}I(a \le \theta \le b)$

$$p(\theta \mid X) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2\right)\cdot\frac{1}{b-a}I(a \le \theta \le b)$$

在 $a \leq \theta \leq b$ 区间上，后验分布的形状由似然函数决定。因此，后验分布在 $a \leq \theta \leq b$ 上为正态分布，但被截断在区间 $[a, b]$ 上：

$$p(\theta \mid X) \sim \text{TruncatedNormal}\left(\bar{X}, \frac{\sigma^2}{n}, a, b\right)$$

其中 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$。

所以，后验分布为截断正态分布：

$$\theta \mid X \sim \text{TruncatedNormal}\left(\bar{X}, \frac{\sigma^2}{n}, a, b\right)$$

□

方法 13.1.3. 共轭分布

从这里能发现一个有趣的事情，对于一些特殊的分布，假定待估参数服从某种分布，那么在抽取样本后，对待估参数服从的分布改变的只是参数，而不会改变参数服从的分布形式。

事实上，对于大部分我们接触的总体分布，都存在一个先验分布，可以使得后验分布与先验分布形式相同。这种先验分布被称为共轭分布。

我们给出共轭分布的对应关系：

1.样本服从$B(n,p)$，p的先验分布为$Beta(a,b)$

2.样本服从$N(\mu,\sigma^2)$，μ的先验分布为$N(\mu_0,\sigma_0^2)$

3.样本服从$N(\mu,\sigma^2)$，σ^2的先验分布为$IG(\alpha,\beta)$

4.样本服从$Poisson(\lambda)$，λ的先验分布为$Gamma(a,b)$

5.样本服从$Exp(\lambda)$，λ的先验分布为$Gamma(a,b)$

6.样本服从$U(0,\theta)$，θ的先验分布为$Pareto(\alpha,\beta)$

7.样本服从$N(\mu,\sigma^2)$，μ,σ^2的先验分布为$N(\mu_0,\sigma_0^2)IG(\alpha,\beta)$

有关这些分布的详细的性质见附录。

13.2 Point Estimationn

这里我有点没看懂，因此给出原文，并给出机翻。

For a Bayesian, estimation is treated as a decision problem. In a given situation, we should select an estimator in order to minimize the loss that we expect to incur (discussed later).

Based on the posterior distribution, a point estimator for θ can be

$$\hat{\theta}_B(x) = \begin{cases} \hat{\theta}_{\text{MAP}} = \arg\max_\theta \pi(\theta|x), & \text{Maximum A Posteriori} \\ \hat{\theta}_E = \mathbb{E}[\theta|x], & \text{Posterior Expectation} \\ \hat{\theta}_{\text{ME}} = \arg\max_a \mathbb{E}[|\theta - a||x], & \text{Posterior Median,}\theta \text{ is scalar} \end{cases}$$

$\hat{\theta}_E$ is usually more reasonable than $\hat{\theta}_{\text{MAP}}$ (see example 7.3.2).

The Posterior mean square error for $\hat{\theta}_B(x)$ is defined by

$$\text{PMSE}(\hat{\theta}_B(x)) = \mathbb{E}_{\theta|x}[(\theta - \hat{\theta}_B(x))^2]$$

When $\hat{\theta}_B(x) = \mathbb{E}(\theta|x)$, then the PMSE of $\hat{\theta}_B(x)$ is

$$\text{PMSE}(\hat{\theta}_B(x)) = \mathbb{E}_{\theta|x}[(\theta - \hat{\theta}_B(x))^2] = \text{Var}(\theta|x).$$

Denote by $\mathbb{E}(\theta|x) = \mu_\pi(x)$ the posterior mean for θ, and $V_\pi(x) = \text{Var}(\theta|x)$ the posterior variance, then

$$\text{PMSE}(\delta(x)) = \mathbb{E}_{\theta|x}[(\theta - \mu_\pi(x)) + (\mu_\pi(x) - \delta(x))]^2 = V_\pi(x) + (\mu_\pi(x) - \delta(x))^2 \geq V_\pi(x).$$

where the equality holds iff $\delta(x) = \mu_\pi(x)$. i.e., the posterior mean estimator $\hat{\theta}_E$ for θ minimizes the PMSE.

$\hat{\theta}_E$ is the optimal estimator under PMSE, and thus is recommended to use in application.

点估计

对于贝叶斯方法，估计被视为一个决策问题。在给定的情况下，我们应选择一个估计量，以最小化我们预期会遭受的损失（后面讨论）。

基于后验分布，参数 θ 的点估计量可以是

$$\hat{\theta}_B(x) = \begin{cases} \hat{\theta}_{\text{MAP}} = \arg\max_\theta \pi(\theta|x), & \text{最大后验估计} \\ \hat{\theta}_E = \mathbb{E}[\theta|x], & \text{后验期望} \\ \hat{\theta}_{\text{ME}} = \arg\max_a \mathbb{E}[|\theta - a||x], & \text{后验中位数,}\theta\text{ 为标量} \end{cases}$$

$\hat{\theta}_E$ 通常比 $\hat{\theta}_{\text{MAP}}$ 更合理（见例 7.3.2）。

$\hat{\theta}_B(x)$ 的后验均方误差定义为

$$\text{PMSE}(\hat{\theta}_B(x)) = \mathbb{E}_{\theta|x}[(\theta - \hat{\theta}_B(x))^2]$$

当 $\hat{\theta}_B(x) = \mathbb{E}(\theta|x)$ 时，$\hat{\theta}_B(x)$ 的 PMSE 为

$$\text{PMSE}(\hat{\theta}_B(x)) = \mathbb{E}_{\theta|x}[(\theta - \hat{\theta}_B(x))^2] = \text{Var}(\theta|x).$$

记 $\mathbb{E}(\theta|x) = \mu_\pi(x)$ 为 θ 的后验均值，$V_\pi(x) = \text{Var}(\theta|x)$ 为后验方差，则有

$$\text{PMSE}(\delta(x)) = \mathbb{E}_{\theta|x}[(\theta - \mu_\pi(x)) + (\mu_\pi(x) - \delta(x))]^2 = V_\pi(x) + (\mu_\pi(x) - \delta(x))^2 \geq V_\pi(x).$$

当且仅当 $\delta(x) = \mu_\pi(x)$ 时，上式取等号。即，参数 θ 的后验均值估计量 $\hat{\theta}_E$ 使得 PMSE 最小。

$\hat{\theta}_E$ 是在 PMSE 下的最优估计量，因此推荐在应用中使用。

方法 13.2.1. 由前面的充分统计量的因子定理，其实我们可以用充分统计量的似然函数$L(t:\theta)$替代$f(X:\theta)$，这可以简化问题。

直观上理解，充分统计量提供了关于参数的所有信息，因此用充分统计量的分布函数不会损失信息。

例 13.2.2. 从几何分布总体 $p(x \mid \theta) = \theta(1-\theta)^{x-1},\ x = 1, 2, 3, \ldots$，取出随机样本 $X_1, \ldots, X_n$，其中参数 θ 的先验分布为均匀分布 $U(0,1)$，求后验期望分布，后验期望估计。

解. 几何分布：

$$p(x \mid \theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, 3, \ldots$$

联合概率密度函数为：

$$p(X_1, X_2, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{X_i - 1}$$

可以简化为：

$$p(X_1, X_2, \ldots, X_n \mid \theta) = \theta^n (1-\theta)^{\sum_{i=1}^{n}(X_i - 1)} = \theta^n (1-\theta)^{n\bar{X} - n}$$

其中，$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 是样本均值。由此可知，$\sum_{i=1}^{n} X_i$ 是 θ 的充分统计量。

充分统计量 $\sum_{i=1}^{n} X_i = T$ 的似然函数为：

$$L(\theta \mid T) = \theta^n (1-\theta)^{T-n}$$

先验分布为均匀分布 $U(0,1)$：

$$\pi(\theta) = 1, \quad 0 \le \theta \le 1$$

后验分布化简为：

$$\pi(\theta \mid X) \propto L(\theta \mid T)\pi(\theta) = \theta^n (1-\theta)^{T-n} = \pi(\theta \mid X) \propto \theta^n (1-\theta)^{T-n}$$

可以看出，后验分布是 Beta 分布 $\mathrm{Beta}(n+1, T-n+1)$：

$$\theta \mid X \sim \mathrm{Beta}(n+1, T-n+1)$$

因此，后验期望估计为：

$$\hat{\theta}_E = \mathbb{E}[\theta \mid X] = \frac{n+1}{n+1+(T-n+1)} = \frac{n+1}{T+2}$$

顺便我们还可以求得最大后验估计MAP：

$$\hat{\theta}_{MAP} = \frac{(n+1)-1}{(n+1)+(T-n+1)-2} = \frac{n}{T}$$

（后验中位数ME我算不出来，咱们就不算了）

这一通说明反正后面能继续证明，得到后验期望估计有最小的均方损失PMSE。

$$Var(\theta|X) = \frac{(n+1)(T-n+1)}{((n+1)+(T-n+1)+1)((n+1)+(T-n+1)^2)} = \frac{(n+1)(T-n+1)}{(T+3)(T+2)^2}$$

(这里我本来想求一下各个估计的均方损失大小，说明上面的“期望估计是在PMSE下的最优估计量”但是发现太难求了，遂放弃) □

13.3 Intercal Estimation

定义 13.3.1. 设参数θ的后验分布为$\pi(\theta|x)$，对给定的样本$\mathbf{x}$和 $0 < \alpha < 1$其中α较小，若存在两个统计量$\hat{\theta}_1$和$\hat{\theta}_2$，使得

$$P(\hat{\theta}_1 \le \theta \le \hat{\theta}_2|x) \ge 1-\alpha$$

则称$[\hat{\theta}_1, \hat{\theta}_2]$为$\theta$的可信水平为$1-\alpha$的*Bayes*可信区间

若满足

$$P(\theta \ge \hat{\theta}_1|x) \ge 1-\alpha$$

则称为$\hat{\theta}_1$为θ的可信下限

若满足

$$P(\theta \le \hat{\theta}_2|x) \ge 1-\alpha$$

则称为$\hat{\theta}_2$为θ的可信上限

例 13.3.2. 设总体X服从指数分布，其密度函数为

$$f(x|\theta) = \theta e^{-\theta x}, x > 0$$

其中θ为未知参数，θ的先验分布为$\pi(\theta) = \frac{1}{\theta}, \theta > 0$，现从这个总体中抽取容量为$n$的样本，样本值为$x_1, x_2, \cdots, x_n$，求$\theta$的后验分布和$\theta$的$1-\alpha$的*Bayes*可信区间

解. 后验分布：

$$\pi(\theta|x) \propto \left(\prod_{i=1}^{n} f(x_i|\theta)\right) \pi(\theta) = \left(\prod_{i=1}^{n} \theta e^{-\theta x_i}\right) \frac{1}{\theta} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i} . \frac{1}{\theta} = \theta^{n-1} e^{-\theta \sum_{i=1}^{n} x_i}$$

这个形式是一个 Gamma 分布的形式，因此：

$$\theta|x \sim \text{Gamma}\left(n, \sum_{i=1}^{n} x_i\right)$$

后验分布的均值和方差分别为：

$$\mathbb{E}(\theta|x) = \frac{n}{\sum_{i=1}^{n} x_i}, \quad \text{Var}(\theta|x) = \frac{n}{\left(\sum_{i=1}^{n} x_i\right)^2}$$

由于 $\theta|x \sim \text{Gamma}(n, \sum_{i=1}^{n} x_i)$，可以使用 Gamma 分布的分位数来构造贝叶斯可信区间，而不是直接使用正态近似。

如果我们要找到 θ 的 $1-\alpha$ 的贝叶斯可信区间，我们需要找到 Gamma 分布的第 $\alpha/2$ 和 $1-\alpha/2$ 分位数：

设 θ_L 和 θ_U 分别是 Gamma 分布 $\text{Gamma}(n, \sum_{i=1}^{n} x_i)$ 的 $\alpha/2$ 和 $1-\alpha/2$ 分位数，那么：

$$P\left(\theta_L \leq \theta \leq \theta_U\right) = 1 - \alpha$$

因此，θ 的 $1-\alpha$ 贝叶斯可信区间为：

$$[\theta_L, \theta_U]$$

其中 θ_L 和 θ_U 是 Gamma 分布 $\text{Gamma}(n, \sum_{i=1}^{n} x_i)$ 的相应分位数。

由于 $\theta|x \sim \text{Gamma}(n, \sum_{i=1}^{n} x_i)$，可以使用 Gamma 分布的分位数来构造贝叶斯可信区间.

设 θ_L 和 θ_U 分别是 Gamma 分布 $\text{Gamma}(n, \sum_{i=1}^{n} x_i)$ 的 $\alpha/2$ 和 $1-\alpha/2$ 分位数，那么：

$$P\left(\theta_L \leq \theta \leq \theta_U\right) = 1 - \alpha$$

因此，θ 的 $1-\alpha$ 贝叶斯可信区间为：

$$[\theta_L, \theta_U]$$

其中 θ_L 和 θ_U 是 Gamma 分布 $\text{Gamma}(n, \sum_{i=1}^{n} x_i)$ 的相应分位数。

令样本均值为 5，样本容量为 10，则 $\sum_{i=1}^{n} x_i = n \cdot \bar{x} = 10 \cdot 5 = 50$。

后验分布为 $\theta|x \sim \text{Gamma}(10, 50)$。

根据查表或计算器得出的 $\text{Gamma}(10, 50)$ 分位数，$\theta_{0.025}$ 约为 0.066，$\theta_{0.975}$ 约为 0.164。

所以，θ 的 95% 贝叶斯可信区间约为：

$$[0.066, 0.164]$$

□

定义 13.3.3. *A Bayesian credible interval of size* $1-\alpha$ *is an interval* (a, b) *such that*

$$P(a \leq \theta \leq b|x) = 1 - \alpha.$$

i.e.,

$$\int_a^b \pi(\theta|x)\, d\theta = 1 - \alpha.$$

- Conceptually, probability comes into play in a frequentist confidence interval before collecting the data. Ex: there is a 95% probability that we will collect data that produces an interval that contains the true parameter value. Awkward!

- Meanwhile, probability comes into play in a Bayesian credible interval after collecting the data. Ex: based on the data, we now think there is a 95% probability that the true parameter value is in the interval.

It is clear that in general, we will have (infinitely) many credible intervals for θ. The shortest possible credible interval is called a highest posterior density (HPD) interval.

定义 13.3.4. *The* $100 \times (1-\alpha)\%$ *HPD interval is an interval of form*

$$C = \{\theta : \pi(\theta|x) \geq c(\alpha)\}$$

where $c(\alpha)$ *is the largest number such that* $P(\theta \in C|x) \geq 1-\alpha$.

*HPD*称为最高后验密度区间，意思是所有满足可信水平的可信区间中，*HPD*最短的区间，他包含了最高的后验概率密度。

这个概率通常不易求得，但是当分布函数是对称的时候，那么*HPD*的中轴应该恰好为后验分布的对称轴。

13.4 Bayesian Testing

后验分布概率大的，就认可检验。

- Assume now that we wish to test the hypothesis

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. In theory, this is straightforward.

- Given a sample of data, x, we can calculate the posterior probabilities $\alpha_0 = P(H_0 \mid x)$ and $\alpha_1 = P(H_1 \mid x)$, and we can reject the null hypothesis H_0 if $\alpha_1 > \alpha_0$.

- Or we can use the Bayes factor (the support of H_0):

$$BF_{01}(x) = \frac{\text{Posterior Odds}}{\text{Prior Odds}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

where $\pi_0 = P(H_0)$ and $\pi_1 = P(H_1)$. It tells us about the changes in our relative beliefs about the two models caused by the data. Jeffreys recommends to attain H_0 when $BF_{01}(x) > 3$.

这部分很直观，故只给出多重贝叶斯假设检验的例子，注意到我们选那个概率最大的假设。

例 13.4.1. 灯泡寿命的检验

假设灯泡的寿命服从期望为 λ 的指数分布。我们有以下测试数据：

$$3215, 5684, 3350, 4092, 2598, 3364, 3518, 2956, 3002, 3695$$

假设先验概率服从$\Gamma(3, 0.001)$，求下列多重检验问题：

$$H_0 : 2000 \leq \frac{1}{\lambda} < 2500$$

$$H_1 : 2500 \leq \frac{1}{\lambda} < 3000$$

$$H_2 : 3000 \leq \frac{1}{\lambda} < 3500$$

$$H_3 : 3500 \leq \frac{1}{\lambda} < 4000$$

解.

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 3547.4$$

假设灯泡寿命服从指数分布，则 λ 的似然函数为：

$$L(\lambda|x_1, x_2, \ldots, x_{10}) = \lambda^{10} e^{-\lambda\sum_{i=1}^{10} x_i}$$

结合先验分布 Gamma(α, β)，后验分布为：

$$\lambda|x_1, x_2, \ldots, x_{10} \sim \text{Gamma}(13, 35474.001)$$

$$P(H_0) = P\left(\frac{1}{2500} \leq \lambda < \frac{1}{2000}\right) = P(0.0004 \leq \lambda < 0.0005)$$

$$\approx \text{GammaCDF}(0.0005, 13, 35474.001) - \text{GammaCDF}(0.0004, 13, 35474.001)$$

$$P(H_1) = P\left(\frac{1}{3000} \leq \lambda < \frac{1}{2500}\right) = P(0.000333 \leq \lambda < 0.0004)$$

$$\approx \text{GammaCDF}(0.0004, 13, 35474.001) - \text{GammaCDF}(0.000333, 13, 35474.001)$$

$$P(H_2) = P\left(\frac{1}{3500} \leq \lambda < \frac{1}{3000}\right) = P(0.000286 \leq \lambda < 0.000333)$$

$$\approx \text{GammaCDF}(0.000333, 13, 35474.001) - \text{GammaCDF}(0.000286, 13, 35474.001)$$

$$P(H_3) = P\left(\frac{1}{4000} \leq \lambda < \frac{1}{3500}\right) = P(0.00025 \leq \lambda < 0.000286)$$

$$\approx \text{GammaCDF}(0.000286, 13, 35474.001) - \text{GammaCDF}(0.00025, 13, 35474.001)$$

$$P(H_0) = P\left(2000 \leq \frac{1}{\lambda} < 2500\right) = 0.238$$
$$P(H_1) = P\left(2500 \leq \frac{1}{\lambda} < 3000\right) = 0.257$$
$$P(H_2) = P\left(3000 \leq \frac{1}{\lambda} < 3500\right) = 0.180$$
$$P(H_3) = P\left(3500 \leq \frac{1}{\lambda} < 4000\right) = 0.108$$

根据计算的后验概率，我们可以得出灯泡寿命期望值最有可能落在区间 $H_1 : [2500, 3000)$ 中，因为这个区间的后验概率 $P(H_1)$ 最大，为 0.257。

我们还可以利用后验分布服从 $\Gamma(13, 35474)$ 来计算得出倒数（寿命）的期望：倒数服从逆伽马分布$Y \sim IGamma(13, 35747)$

$$E(Y) = \frac{35747}{12} \approx 2956$$

这个结果体现了两个问题。一是样本不够多，导致样本分布过于分散不足以显著区分假设（假设之间的概率区别不算特别大）；二是假设区间分划不合适，恰好分划到了期望寿命附近，这可能会导致H_1, H_2近似分属于均值两端。

□

13.5 Bayesian Prediction

- Let $X \sim f(x \mid \theta)$, given a sample x, we want to predict the value of a random variable Z which has density $g(z \mid \theta)$. It is commonly assumed that Z and X are independent given θ.

- Bayesian prediction logic: Let $\pi(\theta \mid x)$ be the posterior density for θ, then $g(z \mid \theta)\pi(\theta \mid x)$ is the joint density for (Z, θ) given $X = x$, thus $p(z \mid x)$ can be obtained by integrating out θ.

- Definition: Suppose X has density $f(x \mid \theta)$, and θ has prior $\pi(\theta)$. Given a sample $X = x$, the posterior predictive density of a random variable Z is defined as
$$p(z \mid x) = \int_{\Theta} g(z \mid \theta)\pi(\theta \mid x)d\theta.$$

- Point prediction: use the mean, median or mode of $p(z \mid x)$ to estimate z.

- Interval prediction: use the $1-\alpha$ credible interval $[a,b]$ based on $p(z \mid x)$:

$$P(a \leq Z \leq b \mid x) = \int_a^b p(z \mid x)dz = 1-\alpha.$$

例 13.5.1. 一枚未知概率的硬币，假设先验分布$\pi(\theta)$ 服从分布$Beta(2\text{fi}2)$，在*20*次抛掷中向上*16*次，求后验分布$\pi(\theta|x)$,后验预测密度$p(z|x)$，求出$z = 11, ..., 20$的值，并给出一个90%预测区间，并给出分布的众数

解. 参照前面的共轭分布的结论，我们直接得出：

$$\theta|X \sim Beta(18, 6)$$

后验预测密度$p(z|x)$为：

$$p(z|x) = \int_{-\infty}^{\infty} \binom{5}{z} \theta^z (1-\theta)^{5-z} \frac{1}{Beta(18,6)} \theta^{17}(1-\theta)^5 = \binom{5}{z} \frac{Beta(18+z, 11-z)}{Beta(18,6)}$$

计算得：

- $z = 0$: 0.0026 $z = 1$: 0.0231 $z = 2$: 0.0974
- $z = 3$: 0.2436 $z = 4$: 0.3654 $z = 5$: 0.2679

90%预测区间：[2,5]

后验预测密度的众数是 4

□

Lec 14 Statistical Decision Theory

损失函数与降低风险，Minimax

14.1 Basic Definitions

A general statistical decision problem has the following components:

- Statistical models $\{f(x|\theta), \theta \in \Theta\}$ for sample X, Θ is parameter space.
- A decision/action space $A = \{a = \delta(x)\}$
- A loss function $L(a, \theta)$ (the negative of utility, nonnegative)
- For Bayesian statistical decision, a prior $\pi(\theta)$ for θ is needed

The parameter θ influences the distribution of the sample X. The decision maker picks a decision a after observing X. She wants to pick a decision that minimizes loss $L(a, \theta)$, for the unknown θ. X is useful because it reveals some information about θ, at least if $f(x|\theta)$ does depend on θ.

The problem of statistical decision theory is to find decision functions δ which are good in the sense of making loss small.

Loss function

- In estimation, we want to find an a which is close to some function μ of θ, such as for instance $\mu(\theta) = E[X]$. Loss is larger if the difference between our estimate and the true value is larger. A commonly used loss in this case is the squared error,

$$L(a, \theta) = (a - \mu(\theta))^2.$$

- An alternative to squared error loss is absolute error loss,

$$L(a, \theta) = |a - \mu(\theta)|.$$

- In testing, we want to decide whether some statement $H_0 : \theta \in \Theta_0$ about the parameter is true. We might evaluate such a decision $a \in \{0, 1\}$ based on the loss

$$L(a, \theta) = \begin{cases} 1 & \text{if } a = 1, \theta \in \Theta_0 \\ c & \text{if } a = 0, \theta \notin \Theta_0 \\ 0 & \text{else.} \end{cases}$$

Risk function

An important intermediate object in evaluating a decision function δ is the risk function R. It measures the expected loss for a given true parameter θ,

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)].$$

Note the dependence of R on the parameter - a decision function δ might be good for some values of θ, but bad for other values. Examples:

- We observe $X \in \{0, \ldots, k\}$ multinomially distributed with $P(X = x) = f(x)$. We want to estimate $f(0)$ and loss is squared error loss. Taking the estimator $\delta(X) = I(X = 0)$, we get the risk function

$$R(\delta, f) = E[(\delta(X) - f(0))^2] = \mathrm{Var}(\delta(X)) = f(0)(1 - f(0)).$$

- We observe $X \sim N(\mu, 1)$, and want to estimate μ. Loss is again squared error loss. Consider the estimator

$$\delta(X) = a + b \cdot X.$$

We get the risk function

$$R(\delta, \mu) = E[(\delta(X)-\mu)^2] = \text{Var}(\delta(X))+\text{Bias}(\delta(X))^2 = b^2\text{Var}(X)+(a+bE[X]-E[X])^2 = b^2+(a+(b$$

Choosing a and b involves a trade-off of bias and variance, and this trade-off depends on μ.

例 14.1.1. *7.22*

接到船运来的一大批零件，从中抽检5件。假设其中不合格品数 $X \sim b(5, \theta)$，又从该批次的先验信息中确定 θ 的先验分布为 $Be(1, 9)$。若观察值为 $x = 0$，在下列损失函数下求检验问题：

$$H_0 : 0 \le \theta \le 0.15 \quad vs \quad H_1 : \theta > 0.15$$

设行动 a_i 表示接受 H_i，$i = 0, 1$。

(1) 损失函数

$$L(a_0, \theta) = \begin{cases} 0, & \theta \le 0.15, \\ 1, & \theta > 0.15. \end{cases} \qquad L(a_1, \theta) = \begin{cases} 2, & \theta \le 0.15, \\ 0, & \theta > 0.15. \end{cases}$$

(2) 损失函数

$$L(a_0, \theta) = \begin{cases} 0, & \theta \le 0.15, \\ 1, & \theta > 0.15. \end{cases} \qquad L(a_1, \theta) = \begin{cases} 0.15 - \theta, & \theta \le 0.15, \\ 0, & \theta > 0.15. \end{cases}$$

解. 后验分布为

$$f(\theta|X = 0) = \frac{P(X = 0|\theta)f(\theta)}{\int_0^1 P(X = 0|\theta)f(\theta)d\theta} = \frac{(1-\theta)^5 \cdot 9(1-\theta)^8}{\int_0^1 (1-\theta)^{13} \cdot 9d\theta} = \frac{9(1-\theta)^{13}}{9 \cdot B(1, 14)} = 14(1-\theta)^{13},$$

(1) 损失函数情况下的贝叶斯风险：

行动 a_0 的贝叶斯风险：

$$R(a_0) = \int_0^{0.15} 0{\cdot}14(1-\theta)^{13}d\theta + \int_{0.15}^1 1{\cdot}14(1-\theta)^{13}d\theta = \int_{0.15}^1 14(1-\theta)^{13}d\theta = 0.85^{14}.$$

行动 a_1 的贝叶斯风险：

$$R(a_1) = \int_0^{0.15} 2{\cdot}14(1-\theta)^{13}d\theta + \int_{0.15}^1 0{\cdot}14(1-\theta)^{13}d\theta = 2\int_0^{0.15} 14(1-\theta)^{13}d\theta = 2(1-0.85^{14}).$$

(2) 损失函数情况下的贝叶斯风险：

行动 a_0 的贝叶斯风险：

$$R(a_0) = \int_0^{0.15} 0{\cdot}14(1-\theta)^{13}d\theta + \int_{0.15}^1 1{\cdot}14(1-\theta)^{13}d\theta = \int_{0.15}^1 14(1-\theta)^{13}d\theta = 0.85^{14}.$$

行动 a_1 的贝叶斯风险：

$$R(a_1) = \int_0^{0.15} (0.15-\theta){\cdot}14(1-\theta)^{13}d\theta + \int_{0.15}^1 0{\cdot}14(1-\theta)^{13}d\theta = 14\int_0^{0.15} (0.15-\theta)(1-\theta)^{13}d\theta.$$

比较 $R(a_0)$ 和 $R(a_1)$，选择贝叶斯风险较小的行动。

□

14.2 Optimality criteria

Admissibility

A decision function δ is said to dominate another function δ' if

$$R(\delta,\theta) \le R(\delta',\theta) \text{ for all } \theta, \text{ and}$$

$$R(\delta,\theta) < R(\delta',\theta) \text{ for at least one } \theta.$$

- Note that dominance only generates a partial ordering of decision functions. In general, there will be many different admissible decision functions.

- It seems natural to only consider decisions functions which are not dominated. Such decision functions are called admissible, all other decision functions are inadmissible.

δ^* is a uniformly optimal decision if

$$R(\delta^*, \theta) \leq R(\delta, \theta), \forall \delta, \forall \theta \in \Theta$$

Such a decision does not exist in general.

- An approach which does not rely on the choice of a prior is minimaxity. This approach evaluates decision functions based on the worst-case risk,

$$\bar{R}(\delta) = \sup_{\theta} R(\delta, \theta)$$

- A minimax decision function, if it exists, solves the problem

$$\delta^* = \arg\min_{\delta} \bar{R}(\delta) = \arg\min_{\delta} \sup_{\theta} R(\delta, \theta).$$

Bayes optimality

An approach which comes naturally is to trade off risk across different θ by assigning weights (prior) $\pi(\theta)$ to each θ. Such an approach evaluates decision functions based on the integrated risk, called Bayes risk,

$$R_\pi(\delta) = \int R(\delta, \theta)\pi(\theta)d\theta$$

- A Bayes decision function minimizes the Bayes risk,

$$\delta^*_\pi = \arg\min_{\delta} R_\pi(\delta).$$

- Let π be a prior distribution, we can define posterior expected loss

$$R_\pi(\delta|x) := \int L(\delta(x), \theta)\pi(\theta|x)d\theta$$

where $\pi(\theta|x)$ is the posterior distribution

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

and

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta$$

is the normalizing constant.

- It is easy to see that any Bayes decision function δ^*_π can be obtained by minimizing $R_\pi(\delta|x)$ through choice of $\delta(x)$ for every x, since

$$R_\pi(\delta) = \int R_\pi(\delta(x)|x)m(x)dx.$$

14.3 Some relationships between these concepts

Much can be said about the relationship between the concepts of admissibility, Bayes optimality, and minimaxity.

Admissibility and Minimaxity If δ^* is admissible with constant risk, then it is a minimax decision function.

Proof: Suppose that δ' had smaller minimax risk than δ^*. Then

$$R(\delta', \theta') \leq \sup_\theta R(\delta', \theta) < \sup_\theta R(\delta^*, \theta) = R(\delta^*, \theta'),$$

where we used constant risk in the last equality. But this contradicts admissibility.

Admissibility of Bayes Decisions If the prior distribution $\pi(\theta)$ is strictly positive and the Bayes decision function δ^*_π has finite risk and risk is continuous in θ, then it is admissible.

Sketch of proof: Suppose it is not admissible. Then it is dominated by δ'. But then

$$R_\pi(\delta') = \int R(\delta', \theta)\pi(\theta)\, d\theta < \int R(\delta^*_\pi, \theta)\pi(\theta)\, d\theta = R_\pi(\delta^*_\pi)$$

since $R(\delta', \theta) \leq R(\delta^*_\pi, \theta)$ for all θ with strict inequality for some θ. This contradicts δ^*_π being a Bayes decision function.

Optimal Bayes Risk vs Minimax Risk The optimal Bayes risk $R(\pi) := \inf_\delta R_\pi(\delta)$ is always smaller than the minimax risk $\bar{R} := \inf_\delta \sup_\theta R(\delta, \theta)$.

Proof:

$$R(\pi) = \inf_\delta R_\pi(\delta) \leq \sup_\pi \inf_\delta R_\pi(\delta) \leq \inf_\delta \sup_\pi R_\pi(\delta) \leq \inf_\delta \sup_\theta R(\delta, \theta) = \bar{R}.$$

If there exists a prior π^* such that $\inf_\delta R_{\pi^*}(\delta) = \sup_\pi \inf_\delta R(\delta, \pi)$, it is called the least favorable distribution.