第一章　数理统计的基本概念

内容提要

1. 数据与模型
2. 统计量及其抽样分布
3. 数据简化的原则（充分性、完备性、不变性等）
4. 统计推断与统计决策概述

§1 数据与模型

§1.1 问题与方法

典型的统计问题： 通过一项随机试验（或观测）收集到一批数据；采用合适的方法从数据中提取信息；根据实际背景对统计分析结果作出解释，给出结论。

统计学：研究数据的收集、分析及解释的一门学科。

统计学中的试验设计、抽样调查等分支主要研究如何收集数据。有些课程如回归分析、多元分析、时间序列分析、属性数据分析等等，研究某一类统计分析方法。有些课程如生存分析、可靠性统计、计量经济、保险统计、医药卫生统计等等，研究统计学在某一领域的具体应用。本课程讨论统计分析的一般原理与方法，而且主要关心基于概率模型的统计分析；不讨论数据收集的问题，也不深入讨论如何解决具体领域中的应用统计问题。

Lehmann and Casella (1998) 认为，“ The answer (统计分析的结论) depends not only on the data, on what is observed, but also on background knowledge of the situation; the latter is formalized in the assumptions with which the analysis is entered.” 他们认为，统计分析方法大致可分为三大类：

1. **描述性分析：**只对数据本身作分析，不作其他假定。主要目的在于对数据作整理、概括，揭示数据的特征与潜在结构。常用统计量（如均值、中位数、方差、四分位间距等）、图形、表格等工具。经典教材有：

   Tukey, J. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

   Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1985). *Exploring Data Tables, Trends and Shapes*. New York: Wiley.

2. **经典的推断与决策：**基于一定的概率模型的统计分析。通常把观测数据x视为随机变量(向量) X 的一个实现（一次观测结果）；假定X 具有某种类型的概率分布，但部分信息未知；根据观测数据对之进行推断或决策。

3. **Bayes分析：**除了观测数据之外，还将关于X 的概率分布的一些先验信息（不是来自于当前数据的信息）用于统计推断或决策。

本课程主要讨论第二、三类方法。

Efron (1998) 认为统计推断主要有三种范式：Bayesian, Fisherian, frequentist。

1. **Bayesian:** 视未知参数为随机变量；统计推断是根据Bayes公式、用观测数据将未知参数的先验分布修改为后验分布的形式统一的过程；用概率分布来表示主观的带有不确定性的观点。

2. **Fisherian:** 强调概率应有实验解释；提出重复抽样原则，统计推断应考虑由观测数据的随机性导致的结论的变动；提出似然函数、条件原则；引进最优化技术寻找统计推断方法。

3. **frequentist:** 由Neyman、E. S. Pearson、Wald、Lehmann等人在1930-40年代在Fisher的最优化思想的启发下发展起来。借助于数学工具研究统计推断问题；将统计推断视为决策问题而不仅仅是对数据作概括；在得到数据之前先研究用最优化方法寻找推断方法；基于重复抽样原则评价统计推断方法。

例1.1 (测量问题) 为了解某一未知的物理量b (如重量、距离、温度等)，进行了n 次重复测量，得到测量值$x_1, \cdots, x_n$。假如每次测量都有误差。通常用样本均值$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 作为b 的估计值。那么诸如

- 为什么用$\bar{x}$ 估计b 比用单次的观测值x_i 估计好？
- $\bar{x}$ 是否是b 的一个合适的估计？有没有更好的估计？
- 用$\bar{x}$ 估计b 误差多大？

等等问题，描述性统计都无法很好地回答，需要在更深入的层面上进行讨论。

§1.2 模型概述

模型的基本结构：把观测数据x 视为随机变量(向量) X 的一个实现；假定X 的模型为$(\mathscr{X}, \mathscr{B}_X, \mathscr{P})$，其中

- $\mathscr{X}$ 是包含X 一切取值的集合（通常为有限维实数空间或其子集），$\mathscr{B}_X$ 是由$\mathscr{X}$ 的某些子集构成的σ-域（通常为Borel 域）；
- 可测空间$(\mathscr{X}, \mathscr{B}_X)$ 称为**样本空间**；
- 设X 的概率分布为P，其中至少有部分信息未知，需要通过数据去推断。假定P 属于某个已知的分布族$\mathscr{P}$。常用θ 表示P 中的未知部分，称之为**参数**，而记P 为P_θ，记$\mathscr{P} = \{P_\theta : \theta \in \Theta\}$，称$\Theta$ 为**参数空间**。对Θ 的如下两个基本要求是显然的：

  1) 当θ 遍历Θ 时，P_θ 遍历$\mathscr{P}$。

  2) (可识别性) 对于$\forall \theta_1, \theta_2 \in \Theta$, $P_{\theta_1} \neq P_{\theta_2}$ 当且仅当$\theta_1 \neq \theta_2$。

可控分布族：若存在定义于样本空间$(\mathscr{X}, \mathscr{B}_X)$ 上的σ-有限测度ν，使$P_\theta \ll \nu$, $\forall \theta \in \Theta$，则称分布族$\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ 关于ν **可控**($\mathscr{P}$ is dominated by ν)，记作$\mathscr{P} \ll \nu$ 。可控族可等价地用其密度函数(Radon-Nikodym 导数)来描述：

$$\mathscr{P} = \{P_\theta : \ dP_\theta(x)/d\nu(x) = p_\theta(x), \ \theta \in \Theta\}.$$

对于定义在欧氏样本空间$(\mathbb{R}^d, \mathscr{B}_{\mathbb{R}^d})$ 上的概率分布，其中$\mathscr{B}_{\mathbb{R}^d}$ 表示$\mathbb{R}^d$ 上的Borel域，Lebesgue*测度*与*计数测度* 是两类最常用的控制测度。关于Lebesgue测度绝对连续的分布就是连续型分布，关于计数测度绝对连续的分布就是离散型分布。

定义在欧氏样本空间$(\mathbb{R}^d, \mathscr{B}_{\mathbb{R}^d})$ 上的概率分布，与其分布函数$F(x)$ 一一对应。相应的分布族也可等价地用其分布函数来描述：

$$\{P_\theta : \ P_\theta\text{的分布函数为 } F_\theta(x), \ \theta \in \Theta\}.$$

模型设定的依据： 抽样方法；专业知识；经验；历史资料；便于数学处理等。

统计分析目标： 根据观测数据x 去推断与P_θ 或与未知参数θ 有关的量。常见形式：点估计、区间估计、假设检验、预测、模型诊断等等。

(续例1.1)假定每次测量的误差是随机的，且具有可加、独立结构，即第i 次测量的可能结果

$$X_i = b + \varepsilon_i,\ i = 1, \cdots, n, \quad \text{且 } \varepsilon_1, \cdots, \varepsilon_n \text{ i.i.d.}$$

实际获得的测量值$x = (x_1, \cdots, x_n)$ 看作是随机向量$X = (X_1, \cdots, X_n)$ 的一个实现。则样本空间自然可取为$(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n})$。根据对误差分布的不同假定，可形成多种模型，例如：

模型1. 假定$\varepsilon_i \sim N(0, \sigma^2)$, $\sigma > 0$ 未知。则$X_1, \cdots, X_n$ i.i.d. $N(b, \sigma^2)$，X 的分布族

$$\mathscr{P}_1 = \left\{ P_\theta :\ \frac{dP_\theta(x)}{d\lambda(x)} = (\sqrt{2\pi}\sigma)^{-n} \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - b)^2 \Big], \theta = (b, \sigma) \in \Theta_1 \right\},$$

其中λ为n 维Lebesgue 测度。参数空间为

$$\Theta_1 = \{(b, \sigma) : -\infty < b < \infty, \sigma > 0\} = \mathbb{R} \times \mathbb{R}^+,$$

是一个2 维的点集。

模型2. 假定ε_i 的分布关于0 对称、关于一维Lebesgue 测度具有pdf (记为g)、二阶矩有限。则$X_1, \cdots, X_n$ 的分布族为

$$\mathscr{P}_2 = \{P : \frac{dP}{d\lambda} = \prod_{i=1}^{n} g(x_i - b),\ g(\cdot)\text{是二阶矩有限的pdf，且是偶函数}\}.$$

这里，参数空间

$$\Theta_2 = \{(b, g) : b \in \mathbb{R},\ g(\cdot)\text{是二阶矩有限的pdf，且是偶函数}\}$$

维数无穷。

此例中，分布族$\mathscr{P}_1$ 与$\mathscr{P}_2$ 都关于n 维Lebesgue 测度可控。统计分析的目的是推断模型参数b 的值。

例1.2 为估计一批产品的废品率θ，从中有放回地随机抽取n 个作检查，记其中的废品数为X。

如果抽样是随机的、等概率的，则可认为$X \sim B(n, \theta)$。此时，其分布族为$(\mathbb{R}, \mathscr{B}_{\mathbb{R}}, \mathscr{P})$，其中

$$\mathscr{P} = \left\{ P_\theta :\ dP_\theta(x)/d\nu(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} I_A(x),\ \theta \in [0, 1] \right\},$$

ν 为$(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$上基于$A = \{0, 1, \cdots, n\}$ 的计数测度，即$\nu(B) = \sum_{x \in A} I_B(x), \forall B \in \mathscr{B}_{\mathbb{R}}$。

参数模型与非参数模型

若模型的参数空间维数有限，则常称为**参数模型**(parametric model)。如上例中的模型1。

若参数空间维数无穷，则常称为**非参数模型**(nonparametric model)。有些非参数模型的参数空间可表示为一个有限维点集与一个无限维集合的乘积集，常称为**半参数模型**(semiparametric model)。如上例中的模型2。

相对于参数模型而言，非参数或半参数模型所作的数学假定较少，一定程度上能避免模型误设对统计分析结论的影响，在分析过程中能充分地让数据说话。但是，需要大量的观测数据才可能获得有效的结论。近年来，对于非参数、半参数模型统计分析方法的研究，是统计学领域的一个热点。

在设定非参数模型时，往往也需要根据实际应用背景、数学处理的可行性等方面的考虑对模型作一些限定。如上例中的模型2。

参数模型与非参数模型的分析方法之间是有一定联系的。有些分析非参数模型的方法也常用于分析参数模型，有些分析参数模型的方法也被用于分析半参数模型。

具有独立同分布结构的模型：在理论与应用中涉及最多的一类模型是具有i.i.d.结构的模型，如例1.1中的两个分布族。即$X=(X_1,\cdots,X_n)$，其中$X_1,\cdots,X_n$ i.i.d.。若假定X_1 的分布族为$(\mathbb{R}^k,\mathscr{B}_{\mathbb{R}^k},Q_\theta),\ \theta\in\Theta$，则$X$ 的分布族为

$$(\mathbb{R}^k,\mathscr{B}_{\mathbb{R}^k},Q_\theta)^n,\quad \theta\in\Theta.$$

此时，常称Q_θ 为**总体分布**，称n 为**样本容量**。

当然，非i.i.d.结构的模型也很多，比如回归模型、时间序列模型、无放回抽样模型等等。以下即是一例。

例1.3 (抽样调查问题) 设$\mathscr{Y}=\{y_1,\cdots,y_N\}$ 是一个由N 个元素构成的有限总体，其中每个元素用其下标来标识，$y_i\in\mathbb{R}^k$ 表示第i 个元素的k 个数值特征。通常，先按某种既定的抽样方法从$\mathscr{Y}$ 中随机抽取一个由n 个不同的元素组成的样本，其下标集为$\{J_1,\cdots,J_n\}$，它们是$\{1,\cdots,N\}$ 的某个含n 个元素的子集。然后对此样本作调查，获得数据$y_{J_i},i=1,\cdots,n$。再用这些样本数据对总体的一些特征作推断，如总体总和 $\eta=\sum_{i=1}^N y_i$ 等。

这个问题中，可视样本数据为随机向量$X = (X_1, \cdots, X_n)$ 的一个实现，其中$X_i = y_{J_i}, i = 1 \cdots, n$。$X$ 的分布取决于$(J_1, \cdots, J_n)$ 的联合分布及总体数据$y_1, \cdots, y_N$，前者由既定的抽样方法确定，后者是模型参数。

记$\{J_1, \cdots, J_n\}$ 所有可能的$c \hat{=} \binom{N}{n}$ 种取值形成的集合为$\mathcal{S}$。确定抽样方法就相当于确定$p(s) = P(\{J_1, \cdots, J_n\} = s), \forall s \in \mathcal{S}$。如果采用的是不放回的简单随机抽样方法，那么$p(s) \equiv c^{-1}$，$X$ 的概率分布为

$$P(X_1 = y_{j_1}, \cdots, X_n = y_{j_n}) = c^{-1},\ \{j_1, \cdots, j_n\} \in \mathcal{S}.$$

其他抽样方法，如分层抽样、整群抽样、系统抽样、不等概率抽样等，相应的模型亦可类似地描述。这些模型往往不具有i.i.d.结构。

支撑与共同支撑

支撑(Support)：对于$(\mathbb{R}^k, \mathscr{B}_{\mathbb{R}}^k)$ 上的概率分布P，称

$$S = \{x \in \mathbb{R}^k : \text{若 } A\text{为包含}x\text{的任一开矩形，则}P(A) > 0\}$$

为P 的支撑。

注：

(1) 显然$P(S) = 1$；反之不然，即满足$P(B) = 1$ 的B 不一定是P 的支撑。

(2) 更一般地，若$B \subset S$且$P(B) = 1$，则也称B 为P 的支撑。（这样定义的支撑不唯一，相差一个零测集，但便于使用。）

(3) 若$p(x)$ 是一连续型分布的pdf，则

$$A = \{x : p(x) > 0\text{且 } p(x)\text{在 } x \text{ 处连续}\}$$

为该分布的支撑。

(4) 若$P \ll Q$，$Q \ll P$，则P、Q 有共同支撑；反之不然。

例1.4 Poisson$(\lambda), \lambda > 0$ 的支撑：非负整数集；

$B(n, \theta), 0 < \theta < 1$ 的支撑：$\{0, 1, \ldots, n\}$；

$U[0, 1]$ 的支撑：$[0, 1]$ 或$[0, 1] \setminus N$，N 是任一Lebesgue 零测集；

Exp(θ)，pdf 为$\theta^{-1} e^{-x/\theta} I(x > 0)$，其支撑为：$[0, \infty)$ 或$[0, \infty) \setminus N$。

共同支撑： 若A是P_θ的支撑，$\forall\theta\in\Theta$，则称分布族$\mathscr{P}=\{P_\theta,\theta\in\Theta\}$有共同支撑。

例如，分布族$\{N(\mu,\sigma^2):\mu\in\mathbb{R},\sigma^2>0\}$有共同的支撑$\mathbb{R}$；分布族$\{U(0,\theta):\theta>0\}$没有共同支撑。

a.s. $\mathscr{P}$: 若某个有关x的命题在$x\in\mathscr{X}\backslash N$上成立，其中$P(N)=0,\forall P\in\mathscr{P}$，则称该命题a.s.$\mathscr{P}$成立。

“a.s. $\mathscr{P}$”等价于“a.s. $P,\forall P\in\mathscr{P}$”或者“a.s. $P_\theta,\forall\theta\in\Theta$”。

若$\mathscr{P}\ll\nu$，则“a.e. ν” $\Longrightarrow$ “a.s. $\mathscr{P}$”（反之不然）。

§1.3 指数族、位置-尺度族与正则族

指数族与位置-尺度族是统计理论中两类最重要的模型，涵盖了许多常用的分布族。这两类模型有良好的数学性质。

指数族(Exponential Families)

定义1.1 若定义在样本空间$(\mathscr{X},\mathscr{B}_X)$ 上的分布族$\mathscr{P}=\{P_\theta:\theta\in\Theta\}$ 关于某个σ有限测度ν 可控，P_θ 关于ν 有如下pdf

$$p_\theta(x)=\frac{dP_\theta(x)}{d\nu(x)}=B(\theta)\cdot\exp\left\{\sum_{j=1}^{s}\eta_j(\theta)T_j(x)\right\}h(x),\quad \theta\in\Theta, \tag{1.1}$$

其中$B(\theta)\in(0,\infty)$, $\eta_j(\theta),j=1,\dots,s$ 为θ 的s 个有限实值函数；$h(x),T_j(x),j=1,\dots,s$为与θ 无关的$\mathscr{B}_X$ 可测的有限实值函数，则称$\mathscr{P}$ 为**指数族**。

称$\eta=(\eta_1(\theta),\dots,\eta_s(\theta))$ 为 (1.1) 的*自然参数*，称

$$p_\eta(x)=A(\eta)\cdot\exp\left\{\sum_{j=1}^{s}\eta_jT_j(x)\right\}h(x),\quad \eta\in\tilde{\Theta}. \tag{1.2}$$

为 (1.1) 的*典则形式*。显然 $\tilde{\Theta}=\{(\eta_1(\theta),\dots,\eta_s(\theta)):\theta\in\Theta\}\subset\mathbb{R}^s$。

自然参数空间(natural parameter space): 称 (1.2)中η 最大的可能取值范围

$$\Xi=\left\{\eta:\ 0<\frac{1}{A(\eta)}=\int_{\mathscr{X}}\exp\Big[\sum_{j=1}^{s}\eta_jT_j(x)\Big]h(x)d\nu(x)<\infty\right\}$$

为典则形式指数族 (1.2) 的自然参数空间。

例1.5 指数族包含了正态分布、Gamma分布、二项分布、Poisson 分布等许多常用的分布族。

(1) 二项分布族$\{B(n,\theta):\ 0<\theta<1\}$ 是定义在$(\mathbb{R},\mathscr{B}_\mathbb{R})$ 上的指数族，P_θ 关于$(\mathbb{R},\mathscr{B}_\mathbb{R})$上基于$A=\{0,1,\cdots,n\}$ 的计数测度ν（即$\nu(B)=\sum_{x\in A}I_B(x),\forall B\in\mathscr{B}_\mathbb{R}$）的pdf 为

$$p_\theta(x)=\binom{n}{x}\Big(\frac{\theta}{1-\theta}\Big)^x(1-\theta)^nI_{\{0,1,\cdots,n\}}(x),$$

自然参数为$\eta=\frac{\theta}{1-\theta}$，自然参数空间为$\Xi=(0,\infty)$。

(2) $\{\text{Poisson}(\lambda):\ \lambda>0\}$ 也是$(\mathbb{R},\mathscr{B}_{\mathbb{R}})$ 上的指数族，自然参数为$\eta=\log(\lambda)$，自然参数空间为$\Xi=(-\infty,\infty)$。

(3) $\{N(\mu,\sigma^2):\ \mu\in\mathbb{R},\sigma^2>0\}$ 与$\{\Gamma(\alpha,\beta):\ \alpha>0,\beta>0\}$ 都是$(\mathbb{R},\mathscr{B}_{\mathbb{R}})$ 上的指数族。

(4) 设 $X_1,\ldots,X_n$i.i.d. $N(\mu,\sigma^2),\mu\in\mathbb{R},\sigma^2>0$，则 $X=(X_1,\cdots,X_n)$ 的分布族是$(\mathbb{R}^n,\mathscr{B}^n_{\mathbb{R}})$ 上的指数族。自然参数？自然参数空间？

(4) Weibull分布族

$$\left\{P_{m,\eta}:\frac{dP_{m,\eta}(x)}{d\lambda(x)}=(\frac{m}{\eta})(\frac{x}{\eta})^{m-1}e^{-(\frac{x}{\eta})^m}I_{(0,\infty)}(x),\ m>0,\eta>0\right\}$$

不是指数族，其中λ 为$(\mathbb{R},\mathscr{B}_{\mathbb{R}})$ 上的Lebesgue 测度。

注：

- 指数族的密度函数形式不唯一。
- 指数族必有共同支撑。
- 要求 $T_1(x),\ldots,T_s(x)$ 线性无关，且 $\eta_1(\theta),\ldots,\eta_s(\theta)$ 线性无关，否则模型 (1.1) 还可以简化。
- 设 $T_1(x),\ldots,T_s(x)$ 线性无关。若 $\tilde{\Theta}\subset\mathbb{R}^s$ 包含一个s 维的开矩形，则称模型 (1.1)是*满秩指数族*；否则称模型 (1.1) 为*弯曲指数族*(Curved Exponential Family)。

例1.6 （非指数族）均匀分布族 $\{U(0,\theta),\theta>0\}$ 无共同支撑，非指数族。

例1.7 满秩与弯曲指数族。

(1) **例1.5**(1)、(2)中的两个指数族都是满秩的，$\tilde{\Theta}\subset\mathbb{R}$，含内点。**例1.5**(3)、(4)中的指数族也都是满秩的，$\tilde{\Theta}\subset\mathbb{R}^2$，含内点。

(2) 正态分布族$\mathscr{P}=\{N(\theta,\theta^2):\ \theta\in\mathbb{R}\backslash\{0\}\}$ 也是一个指数族，pdf 为

$$p_\theta(x)=\frac{1}{\sqrt{2\pi}\theta}e^{-\frac{1}{2}}\exp\{-\frac{x^2}{2\theta^2}+\frac{1}{\theta}x\},$$

其中$\eta_1(\theta)=-\frac{1}{2\theta^2},\ \eta_2(\theta)=\frac{1}{\theta}$，满足$\eta_1=-\frac{1}{2}\eta_2^2$。因此$\tilde{\Theta}=\{(\eta_1(\theta),\eta_2(\theta)):\ \theta\in\Theta\}$ 是 $\mathbb{R}^2$中的一条曲线，不含内点，故该指数族非满秩，为弯曲指数族。

(3) (logisitic回归模型) 设 $X_i \sim B(n_i, \theta_i), i = 1, \ldots, m,\ X_1, \ldots, X_m$ 相互独立。因而$X_1, \ldots, X_m$的联合密度为

$$\begin{aligned} & p(x_1, \ldots, x_m) \\ = & \prod_{i=1}^{m} \binom{n_i}{x_i} \cdot \theta_i^{x_i} (1-\theta_i)^{n_i - x_i} I_{\{0,\cdots,n_i\}}(x_i) \\ = & \prod_{i=1}^{m} (1-\theta_i)^{n_i} \cdot \exp\left\{ \sum_{i=1}^{m} x_i \cdot \log \frac{\theta_i}{1-\theta_i} \right\} \cdot \prod_{i=1}^{m} \binom{n_i}{x_i} I_{\{0,\cdots,n_i\}}(x_i). \end{aligned}$$

假定θ_i 满足

$$\log \frac{\theta_i}{1-\theta_i} = \alpha + \beta z_i, \quad \alpha, \beta \in \mathbb{R}, i = 1, \cdots, m,$$

其中$z_1, \ldots, z_n$ 已知且非全等。若视模型的未知参数为(α, β)，则 $\tilde{\Theta} \subset \mathbb{R}^2$，含内点，故$(X_1, \ldots, X_m)$ 的分布族是满秩的指数族。若视模型的未知参数为$(\theta_1, \cdots, \theta_m)$，则 $\tilde{\Theta} \subset \mathbb{R}^m$ 不含内点，故$(X_1, \ldots, X_m)$ 的分布族是弯曲指数族。

指数族的性质

性质1.1 随机向量X, Y相互独立，各自的分布族都是指数族，则(X, Y)的分布族也是指数族。特别地，若$X_i, i = 1, \ldots, n$ *i.i.d.* (1.1)，则$(X_1, \ldots, X_n)$ 的*jpdf*为

$$B^n(\theta) \cdot \exp\left\{ \sum_{j=1}^{s} \eta_j(\theta) \sum_{i=1}^{n} T_j(x_i) \right\} \cdot \prod_{i=1}^{n} h(x_i), \quad \theta \in \Theta,$$

该分布族仍是指数族。

性质1.2 典则形式指数族的自然参数空间 Ξ 是凸集。

证明. 对 $\forall \eta_1, \eta_2 \in \Xi$ 及 $\forall \alpha \in (0,1)$，由Hölder 不等式知，

$$\begin{aligned} & \int_{\mathscr{X}} \exp\left[\sum_{j=1}^{s} \Big\{ \alpha\eta_{1j} T_j(x) + (1-\alpha)\eta_{2j} T_j(x) \Big\} \right] h(x) d\nu(x) \\ \leq & \left[\int_{\mathscr{X}} \exp\Big\{ \sum_{j=1}^{s} \eta_{1j} T_j(x) \Big\} h(x) d\nu(x) \right]^{\alpha} \left[\int_{\mathscr{X}} \exp\Big\{ \sum_{j=1}^{s} \eta_{2j} T_j(x) \Big\} h(x) d\nu(x) \right]^{1-\alpha} \\ < & \infty. \end{aligned}$$

即 $\alpha\eta_1 + (1-\alpha)\eta_2 \in \Xi$，故结论得证。 □

性质1.3 f 是$(\mathscr{X}, \mathscr{B}_X)$ 上的可测函数。记

$$G(\eta) = \int_{\mathscr{X}} f(x) \cdot \exp\Big\{\sum_{j=1}^{s} \eta_j T_j(x)\Big\} h(x) d\nu(x),$$

$$\Theta_f = \{\eta : \ |G(\eta)| < \infty\}.$$

则$G(\eta)$ 在Θ_f 的任一内点处连续、有任意阶偏导数、求导与积分次序可交换。

证明主要用到Lebesgue 控制收敛定理。具体参见Lehmann *et. al.* (2005)[TSH]，p49，Theorem 2.7.1 的证明。

若取$f(x) \equiv 1$，则

1. $A(\eta)$ 在Ξ 的任一内点处连续、任意阶可导；
2. $\frac{1}{A(\eta)} = \int_{\mathscr{X}} \exp\{\sum_{j=1}^{s} \eta_j T_j(x)\} h(x) d\nu(x)$。两边关于$\eta_k$求偏导

$$-\frac{1}{A^2(\eta)} \frac{\partial A(\eta)}{\partial \eta_k} = \int_{\mathscr{X}} T_k(x) \exp\{\sum_{j=1}^{s} \eta_j T_j(x)\} h(x) d\nu(x), \qquad (1.3)$$

得

$$\mathrm{E}_\eta[T_k(X)] = -\frac{\partial \log A(\eta)}{\partial \eta_k}.$$

在 (1.3)两边同乘以$A(\eta)$，再同时关于η_l 求偏导，则得

$$\mathrm{Cov}(T_k(X), T_l(X)) = -\frac{\partial^2 \log A(\eta)}{\partial \eta_k \partial \eta_l}.$$

位置-尺度族(Location-scale Families)

有些常用的分布族，可以看作是通过对一个分布已知的随机变量或者随机向量，施加一系列变换得到的。位置-尺度族就是其中的一类。一个已知的分布经一族平移、尺度变换，就可得到一个位置-尺度族。为使生成的分布族有较好的性质，往往需要对施行的变换有一定的要求。统计理论研究中，常要求生成分布族的变换构成一个变换群。这类分布族称为*群族*(group family)，有良好的“对称”结构。位置-尺度族是研究得最多的一类群族。

定义1.2 （一维的位置族、尺度族、位置- 尺度族）设$F_0(x)$是一个已知的一元分布函数。称

$$\mathscr{P}_1 = \{P : \ P\text{的分布函数为}F_0(x-a), \ a \in (-\infty, +\infty)\}$$

为位置族。称

$$\mathscr{P}_2 = \{P:\ P\text{的分布函数为}F_0(x/b),\ b \in (0,+\infty)\}$$

为尺度族。称

$$\mathscr{P}_3 = \{P:\ P\text{的分布函数为}F_0\big(\frac{x-a}{b}\big),\ a \in \mathbb{R}, b \in (0,+\infty)\}$$

为位置-尺度族。

若r.v. U 的分布函数为$F_0(x)$，那么显然，对U 施以平移变换$X = U + a$ 即得$\mathscr{P}_1$ 中的分布；施以尺度变换$X = bU, b > 0$，即得$\mathscr{P}_2$ 中的分布；施以两种变换的组合$X = a + bU, b > 0$，即得$\mathscr{P}_3$ 中的分布。

如果$F_0(x)$ 关于Lebesgue 测度具有pdf $f_0(x)$，则$\mathscr{P}_1 \sim \mathscr{P}_3$ 中元素的pdf 分别为

$$f(x) = f_0(x-a), \quad f(x) = \frac{1}{b} f_0(x/b), \quad f(x) = \frac{1}{b} f_0\left(\frac{x-a}{b}\right).$$

例1.8 取$f_0(x)$ 为

(1) 标准正态的pdf $\varphi(x)$，

(2) 指数分布的pdf $e^{-x} I_{(0,\infty)}(x)$，

(3) 双指数分布的pdf $\frac{1}{2} e^{-|x|}$，

(4) Cauchy 分布的pdf $\frac{1}{\pi(1+x^2)}$，

(5) U(0,1) 的pdf 等等。

对$f_0(x)$ 施加由所有$a \in \mathbb{R}$ 对应的位移变换，就可生成一些常用的位置族；施加由所有$b > 0$ 对应的尺度变换，就可生成一些常用的尺度族；施加由所有$a \in \mathbb{R}, b > 0$ 对应的位移-尺度变换，就可生成一些常用的位置-尺度族。

定义1.3 (一般的位置-尺度族) 设P 是$(\mathbb{R}^k, \mathscr{B}_{\mathbb{R}^k})$ 上的一个已知概率测度，F 是P 的分布函数。记$\mathcal{V} \subset \mathbb{R}^k$, $\mathcal{M}$ 是一些k 阶正定矩阵的集合。若下列$(\mathbb{R}^k, \mathscr{B}_{\mathbb{R}^k}) \to (\mathbb{R}^k, \mathscr{B}_{\mathbb{R}^k})$ 上的一一对应变换的集合

$$\big\{g(\cdot):\ g(x) = \Sigma^{1/2} x + \mu,\ \mu \in \mathcal{V}, \Sigma \in \mathcal{M}\big\}$$

关于复合运算构成一个变换群，则

$$\mathscr{P} = \{P_{(\mu,\Sigma)} :\ F_{(\mu,\Sigma)}(x) = F(\Sigma^{-1/2}(x-\mu)),\ \mu \in \mathcal{V}, \Sigma \in \mathcal{M}\}$$

称为$(\mathbb{R}^k, \mathscr{B}_{\mathbb{R}^k})$ 上的位置- 尺度族，其中$F_{(\mu,\Sigma)}(x)$ 是$P_{(\mu,\Sigma)}$ 的分布函数。若P 关于Lebesgue 测度具有pdf f，则

$$f_{(\mu,\Sigma)}(x) = |\Sigma|^{-1/2} f(\Sigma^{-1/2}(x-\mu)).$$

这类分布族包含了如多元正态分布族、线性模型族等许多常用的分布族。

正则族(Regular Families)

这也是统计理论中得较多的一类比较一般的分布族。这类分布族是根据Cramér-Rao 不等式所要求的正则条件来定义的。

定义1.4 设定义在样本空间$(\mathscr{X}, \mathscr{B}_X)$ 上的分布族$\mathscr{P} = \{P_\theta : \theta = (\theta_1, \cdots, \theta_s) \in \Theta\}$ 关于某个σ有限测度ν 可控，$p_\theta(x)$ 为P_θ 关于ν 的密度函数。若$\mathscr{P}$ 满足下列正则条件：

1'. Θ 为$\mathbb{R}^s$ 中的开凸集；

2'. P_θ 有共同支撑A；$\forall x \in A, \forall \theta \in \Theta, p_\theta(x) > 0$，且$\frac{\partial p_\theta(x)}{\partial \theta_i}$ 存在、有限，$i = 1, \ldots, s$；

3'. $\int_{\mathscr{X}} \frac{\partial}{\partial \theta_i} p_\theta(x) d\mu(x) = \frac{\partial}{\partial \theta_i} \int_{\mathscr{X}} p_\theta(x) d\mu(x) = 0,\ i = 1, \ldots, s$。

则称$\mathscr{P}$ 为**正则族**。

指数族、位置尺度族、正则族这三类分布族，互有交叉，又各有不同。有的分布族同时属于这三类，比如正态分布族；有的分布族同时属于其中的两类，比如Cauchy 分布族，既是位置族又是正则族。指数族一般都是正则族，但只有部分位置尺度族属正则族。统计理论更偏重于考察各类模型的共性。

常用分布

Casella and Berger (2002) p621-627 列举了若干一元分布及其相互关系，第四章中还涉及多元正态、多项分布等一些常用多元分布。这些分布的来源、主要特点及其应用，大家应当熟悉。

§1.4 凸(Convex)函数

一、一元凸函数

定义1.5 ϕ为定义在开区间(a,b)上的实值函数，$-\infty \le a < b \le \infty$，若对任意$a<x<y<b$及$\forall \gamma \in (0,1)$有

$$\phi(\gamma x + (1-\gamma)y) \le \gamma\phi(x) + (1-\gamma)\phi(y), \tag{1.4}$$

则称ϕ是凸的。若不等号严格成立，则称ϕ严凸。若$-\phi$是凸的，则称ϕ是凹的。

性质1.4 $\phi(\cdot)$是(a,b)上的凸函数，

1. ϕ在(a,b)上连续；
2. $\forall x \in (a,b), G_x(\varepsilon) = [\phi(x+\varepsilon)-\phi(x)]/\varepsilon$ 是ε的单调增函数；
3. $\forall x \in (a,b), \phi(x)$的左、右导数存在；
4. $f_i(\cdot), i \in I$都是(a,b) 上的凸函数，则$f(x) = \sup\limits_{i\in I} f_i(x)$是凸的；$\lambda_i \ge 0, f(x) = \sum\limits_{i=1}^{k} \lambda_i f_i(x)$是凸的。

凸性判断：

定理1.1 1. ϕ 定义在(a,b) 上可微，则ϕ 凸$\Longleftrightarrow \phi'(x) \le \phi'(y), \forall a<x<y<b$；

2. ϕ 二阶可导，则ϕ 凸$\Longleftrightarrow \phi''(x) \ge 0, \forall x \in (a,b)$；

   $\phi''(x) > 0 \Longrightarrow \phi$ 严凸，$\forall x \in (a,b)$。但ϕ 严凸$\not\Rightarrow \phi''(x) > 0, \forall x \in (a,b)$。

定理1.2 ϕ 为(a,b) 上的凸函数，$t \in (a,b)$，则存在过$(t,\phi(t))$ 的直线$L(x) = c\cdot(x-t)+\phi(t)$，使$L(x) \le \phi(x), \forall x \in (a,b)$

凸性可推广至有限个点的情形：

$x_1,\dots,x_m \in (a,b), 0<\gamma_i<1, \sum\limits_{i=1}^{m}\gamma_i = 1$，则$\phi(\sum\limits_{i=1}^{m}\gamma_i x_i) \le \sum\limits_{i=1}^{m}\gamma_i\phi(x_i)$；

设随机变量X 的分布列为$P(X=x_i)=\gamma_i,\ i=1,\cdots,m$，那么由上式即可得$\phi(EX) \le E\phi(X)$。

定理1.3 （*Jensen*不等式）ϕ 为(a,b) 上的凸函数，X 是(a,b) 上的随机变量且EX 有限，则$\phi(EX) \le E\phi(X)$。若ϕ 严凸且X *a.s.* 非常数，则不等号严格成立。

设X, Y 为随机变量。由凸性还可得其他一些常用的不等式，如：

(1) Hölder 不等式：若$p \ge 1, q \ge 1$ 且满足$1/p + 1/q = 1$（约定$1 + 1/\infty = 1$），则

$$E(|XY|) \le (E|X|^p)^{1/p}(E|Y|^q)^{1/q}.$$

(2) Liapounov 不等式：若$1 \le r \le s \le \infty$，则

$$(E|X|^r)^{1/r} \le (E|X|^s)^{1/s}.$$

(3) Minkowski不等式：若$p \ge 1$，则

$$(E|X+Y|^p)^{1/p} \le (E|X|^p)^{1/p} + (E|Y|^p)^{1/p}.$$

(4) 熵不等式：若$f(x), g(x)$ 是关于σ有限测度ν 的概率密度函数，则

$$E_f\big[\log(\frac{f(X)}{g(X)})\big] = \int_{\mathscr{X}} f(x)\log\big[\frac{f(x)}{g(x)}\big]d\nu(x) \ge 0,$$

其中$E_f\big[\log(\frac{f(X)}{g(X)})\big]$ 称为 f 与 g 之间相对于 f 的熵距离(Entropy distance)，或 g 与 f 之间的Kullback-Leibler 距离，或 g 在f 处的Kullback-Leibler 信息量。由熵不等式即可得

$$E_f \log[g(X)] \le E_f \log[f(X)].$$

在EM算法中，该不等式用于论证各轮迭代似然函数不减。

二、多元情况

定义1.6 $\mathbb{R}^k$ 中的非空子集S 满足：若$\forall x, y \in S$，有$\gamma x + (1-\gamma)y \in S, \forall \gamma \in (0,1)$，则称$S$ 为凸集。

定义1.7 ϕ 为定义在开凸集S 上的多元实函数，若 (1.4) 对$\forall x, y \in S$ 及$\forall \gamma \in (0,1)$ 成立，则称ϕ 为凸的；若 (1.4) 对$\forall x \ne y$ 及$\gamma \in (0,1)$，不等号严格成立，则称ϕ 严凸。

判别准则：

定理1.4 ϕ 定义于开凸集$S \subset \mathbb{R}^k$ 上两阶可微，则ϕ 凸$\Longleftrightarrow$ ϕ 的*Hessian* 矩阵$\left(\dfrac{\partial^2\phi}{\partial x_i \partial x_j}\right)$ 半正定。

*Hessian*矩阵正定$\Longrightarrow \phi$ 严凸；但ϕ 严凸$\nRightarrow$ *Hessian* 矩阵正定。

设$\phi_j(\cdot)$ 是定义在开区间I_j 上的凸函数，$j = 1, \ldots, k$。取$c_j > 0, j = 1, \ldots, k$，则$\phi(x_1, \ldots, x_k) = \sum_{j=1}^k c_j \phi_j(x_j)$ 是定义在$I_1 \times \cdots \times I_k$ 上的凸函数。

例如，在讨论参数向量$\gamma(\theta) = (\gamma_1(\theta), \ldots, \gamma_k(\theta))$ 的估计问题时，常取损失函数为：

$$L_1(\theta, a) = \sum_{j=1}^k c_j (a_j - \gamma_j(\theta))^2,$$

这是个凸函数。或者取损失函数为

$$L_2(\theta, a) = \sum_{i=1}^k \sum_{j=1}^k c_{ij}(a_i - \gamma_i(\theta))(a_j - \gamma_j(\theta)) = [a - \gamma(\theta)]' C [a - \gamma(\theta)],$$

其中矩阵$C = (c_{ij})_{k \times k}$。当Hessian矩阵$(C + C')$ 非负定时，这也是一个凸损失函数。

定理1.5 （定理 *1.2*的推广）ϕ 是定义于开凸集$S \subset \mathbb{R}^k$ 上的凸函数，$\forall t \in S$，存在一个过$(t, \phi(t))$ 的超平面$L(x) = c' \cdot (x - t) + \phi(t), \forall x \in S$，使得$L(x) \leq \phi(x), \forall x \in S$。

定理1.6 （*Jensen*不等式的推广）随机向量X 定义于开凸集$S \subset \mathbb{R}^k$ 上，且X 的期望向量$E(X) = (EX_1, \ldots, EX_n)'$ 存在、有限，ϕ 为S 上的凸函数，则$\phi(EX) \leq E\phi(X)$。

三、求形如$E\rho(X - a)$ 的极小值的问题

例如，$\rho(t) = |t|$，X 是离散型均匀分布的随机变量，分布列为$P(X = x_i) = \frac{1}{n},\ i = 1, \cdots, n$，那么

$$\text{argmin}_a E\rho(X-a) = \text{argmin}_a \frac{1}{n}\sum_{i=1}^n |x_i - a| = \begin{cases} n\text{为奇数，有唯一极小值点} x_{(\frac{n+1}{2})}, \\ n\text{为偶数，} [x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}], \end{cases}$$

引理1.1 ϕ 为$(-\infty,\infty)$ 上的凸函数，有下界且非单调，则ϕ 能达到其最小值，且最小值点的集合S 为一个闭区间；若ϕ 严凸，则最小值点唯一。

证明. ϕ 凸、非单调，则$x \to \mp\infty$ 时，$\phi(x) \to \infty$。

ϕ 连续，有下界，ϕ 必能达到下确界。

再由ϕ 的凸性可知，S 是一个区间。由ϕ 的连续性可证S 闭。 □

定理1.7 ρ 为定义在$(-\infty,\infty)$ 上的凸函数，随机变量X 使$\phi(a) = E\rho(X-a)$ 在某些a 处有限，若ρ 非单调，则$\phi(a)$ 能取得最小值，且最小值点是一个闭区间；若ρ 严凸，则最小值点唯一。

推论1.1 在定理 *1.7*的前提下，若ρ 是偶函数且X 的分布关于μ 对称，则$\phi(a)$ 在$a = \mu$ 处达到最小值。

§2 统计量及其抽样分布

统计分析的一个重要手段是对观测数据进行压缩、简化，以便从中提炼出有用的信息。统计量是数据加工、压缩的常用工具。若T 是$(\mathscr{X},\mathscr{B}_X)\to(\mathscr{T},\mathscr{B}_T)$ 的一个已知的可测映射，则称$T(X)$ 是X 的**统计量**(statistic)。

常用统计量的值域空间是欧氏的。能够起到简化作用的统计量，其维数往往低于原始数据X 的维数。也有一些特殊的统计量，比如经验分布函数、次序统计量等，它们一般没有降维的作用。

假定 S,T 是X 的两个统计量，值域空间都是欧氏的，且S 是T 的函数，T 也是S 的函数，那么这两个统计量的作用相当。例如，设样本空间为$(\mathscr{X},\mathscr{B}_X)=(\mathbb{R},\mathscr{B}_{\mathbb{R}})$。显然，$T_1(x)=x^2,T_2(x)=|x|,T_3(x)=\exp(-x^2)$ 都是作用相同的统计量。

通常，统计量$T(X)$ 是随机的，其分布常称为**抽样分布**(sampling distribution)。推导统计量的抽样分布，是统计理论研究中的一类重要问题。常用的方法有：

- 由变换法求得 T 的精确分布（少量参数型模型的场合）；
- 由中心极限定理(CLT)等工具获得大样本情况下T 的渐近分布；
- 用随机模拟、Bootstrap等方法获得T 的近似分布；
- 用Laplace 逼近等方法获得T 的近似分布；
- 通过若干阶矩等数字特征了解抽样分布的部分信息；等等。

一、有限样本下的精确分布

例2.1 设$X_1,X_2,\cdots,X_n$ i.i.d.。样本均值与样本方差是两个常用统计量：

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

它们的抽样分布依赖于X_i 的共同分布。下面是一些常用的结果：

(1) 若$X_1\sim N(\mu,\sigma^2)$，则$\overline{X}\sim N(\mu,\sigma^2/n)$，$(n-1)S^2/\sigma^2\sim\chi^2(n-1)$，且$\overline{X}$ 与S^2 相互独立。

(2) 若X_i 的共同分布是其他一些形式较简单的参数型分布，如$E(0,\theta)$ 等，则$\overline{X}$ 的精确分布有时还可导出，但S^2 的精确分布往往没有显式表达。

(3) 若已知$\mathrm{E}X_1 \hat{=} \mu$ 存在，则有$\mathrm{E}\overline{X} = \mu$；
若已知$\mathrm{Var}X_1 \hat{=} \sigma^2$ 存在，则有$\mathrm{E}\overline{X} = \mu, \mathrm{Var}\overline{X} = \sigma^2/n, \mathrm{E}S^2 = \sigma^2$；
若知道 $\mathrm{E}|X_1|^3, \mathrm{E}|X_1|^4$ 存在，则还可知$\mathrm{E}\overline{X}^3, \mathrm{E}\overline{X}^4, \mathrm{Var}(S^2), \mathrm{Cov}(\overline{X}, S^2)$。

例2.2 (次序统计量) 设$X_1, X_2, \cdots, X_n$ i.i.d.，其共同的分布函数F 关于一维Lebesgue 测度具有密度函数f。则其次序统计量$(X_{(1)}, \cdots, X_{(n)})$ 的jpdf为

$$g(x_1, \cdots, x_n) = n! f(x_1) f(x_2) \cdots f(x_n) I(x_1 < x_2 < \cdots < x_n).$$

对于$1 \le i < j \le n$, $(X_{(i)}, X_{(j)})$ 的jpdf为

$$\begin{aligned} g_{i,j}(x,y) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\ &\quad \cdot [F(x)]^{i-1} f(x) [F(y) - F(x)]^{j-i-1} f(y) [1 - F(y)]^{n-j} I(x < y), \end{aligned}$$

$X_{(i)}$ 的pdf为

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

二、大样本情况下统计量的渐近分布

当样本容量固定时，绝大多数的统计量的分布不易求得，而当样本容量趋于无穷时，统计量的分布往往趋于一个常见的简单分布。这为我们在样本很大时了解统计量的性质提供了依据。

在讨论统计量的大样本性质时，我们不便再把样本简单记为X 了，而应当将 $(X_1, \ldots, X_n), n = 1, 2, \ldots$ 视为一个样本序列。当说到某个统计量 $\delta_n = \delta(X_1, \ldots, X_n)$ 时，实际上指的是一串统计量 $\delta(X_1), \delta(X_1, X_2), \ldots, \delta(X_1, X_2, \cdots, X_n), \ldots$ 中的一个。统计量 δ_n 的极限性质，指的是统计量序列 $\{\delta_1, \delta_2, \ldots\}$ 的极限性质。

收敛性

设 $X_1, X_2, \cdots$ 为随机变量序列，X 是随机变量。

(1) 依概率收敛(Convergence in Probability)：

$$X_n \to_P X$$

if and only if (iff)

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

(2) L_r 收敛(Convergence in rth mean)：设常数 $r \in (0, \infty)$

$$X_n \to_{L_r} X$$

iff

$$\lim_{n\to\infty} E(|X_n - X|^r) = 0.$$

(3) 依分布收敛(Convergence in distribution, Convergence in law):

$$X_n \to_D X \quad \text{or} \quad X_n \to_D F(\cdot)$$

iff

$$\lim_{n\to\infty} F_n(x) = F(x),$$

对F 的所有连续点x 都成立，其中F_n 与F 分别是X_n 和X 的分布函数。另外，也称分布函数序列F_n 弱收敛(converge weakly) 到F，记为$F_n \to_w F$。

等价条件：

$$\begin{aligned} X_n \to_D X &\iff Ef(X_n) \to Ef(X), \text{对任意有界、连续(a.s.)、实函数 } f(\cdot); \\ &\iff M_{X_n}(t) \to M_X(t); \\ &\iff \Phi_{X_n}(t) \to \Phi_X(t). \end{aligned}$$

（Pólya 定理）若$F_n \to_w F$ 且F 在$\mathbb{R}^k$ 上连续，则

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}^k} |F_n(x) - F(x)| = 0.$$

(4) 一致可积(Uniformly Integrable)：$\{X_n\}$ 一致可积iff

$$\lim_{c\to\infty} \sup_n E[|X_n| I_{|X_n|>c}] = 0.$$

$\{X_n\}$ 一致可积的两个充分条件：

1) $\exists \delta > 0, \ni \sup_n E|X_n|^{1+\delta} < \infty$.

2) $P(|X_n| > c) \le P(|X| > c),\ \forall n,\ \forall c > 0$，且$E|X| < \infty$。

一个有用结果：

若$X_n \to_D X$，常数$r \in (0, \infty)$，且$|X_n|^r,\ n = 1, 2, \cdots$ 皆可积，则下列三者等价：

(i) $X_n \to_r X$;

(ii) $E|X_n|^r \to E|X|^r < \infty$

(iii) $\{X_n^r\}$ 一致可积。

(5) 随机向量的收敛性。设$X, X_1, X_2, \cdots$ 为维数相同的随机向量。

$$X_n \to_P X$$

iff X_n 的每个分量依概率收敛于X 的相应分量；

$$X_n \to_D X$$

iff X_n 的每一个线性组合依分布收敛于X 的相应线性组合(Cramér - Wold 定理)。

(6) (Slutsky 定理). 设$X_n \to_D X$，且$Y_n \to_P c$，c 是一个常数。则

1) $X_n + Y_n \to_D X + c$;

2) $X_n Y_n \to_D cX$;

3) $X_n / Y_n \to_D X/c$ 若$c \neq 0$。

(7) 随机向量序列函数的收敛性。设$X, X_1, X_2, \cdots$ 是k 维随机向量，g 是定义在$\mathbb{R}^k$ 上的向量值连续函数。则

1) $X_n \to_P X \quad \Longrightarrow \quad g(X_n) \to_P g(X)$;

2) $X_n \to_D X \quad \Longrightarrow \quad g(X_n) \to_D g(X)$.

依概率收敛的统计应用：估计量的相合性。基本判断方法：

1. 大数定律(例如，Chebyshev、Markov、Khinchin 大数定律等)

   $X_1, \ldots, X_n$ i.i.d.，$EX_1 = \xi$，则 $\bar{X} \to_P \xi$。

2. 由L_p 收敛来判断：若存在$p > 0$，使得$E|X_n - X|^p \to 0$，则$X_n \to_P X$。

3. 设$X_{jn} \to_P a_j,\ j = 1, \cdots, k$，函数$h(y_1, \cdots, y_k)$ 在点$(a_1, \cdots, a_k)$ 处连续。则$h(X_{1n}, \cdots, X_{kn}) \to_P h(a_1, \cdots, a_k)$。

例2.3 设$X_1, \ldots, X_n$ i.i.d.，$VarX_1 = \sigma^2 < \infty$。则由

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}\bar{X}\right) \end{aligned}$$

及上述判断方法1，3知，$S_n^2 \to_P \sigma^2,\ \forall \sigma^2 > 0$。

渐近正态性

渐近正态性是最典型的大样本性质。

设$\{X_n\}$ 是随机变量序列，若存在常数序列$\{\mu_n\}$ 与$\{\sigma_n\}$，使得

$$\frac{X_n - \mu_n}{\sigma_n} \to_D N(0, 1)$$

则称$\{X_n\}$ 渐近正态。记作$X_n \sim AN(\mu_n, \sigma_n)$。称$\mu_n$ 为渐近均值，σ_n^2 为渐近方差。

注: 1) μ_n，σ_n^2 不一定是X_n 的均值、方差。

2) 随机向量的渐近正态性也类似。

基本判断方法：

1. (中心极限定理). $X_1, \ldots, X_n\ i.i.d., EX_1 = \xi, VarX_1 = \sigma^2 < \infty$，则$\sqrt{n}(\bar{X} - \xi) \to_D N(0, \sigma^2)$。也即$\bar{X} \sim AN(\xi, \sigma^2/n)$.

2. (δ-方法). 若$\sqrt{n}(T_n - \theta) \to_D N(0, \tau^2(\theta))$，则当$h'(\theta)$ 存在且不为0时，有

$$\sqrt{n}[h(T_n) - h(\theta)] \to_D N(0, [h'(\theta)]^2\tau^2(\theta)).$$

证明. 对$h(T_n)$ 在θ 处Taylor 展开

$$h(T_n) = h(\theta) + (T_n - \theta)[h'(\theta) + R_n].$$

其中$R_n = O(T_n - \theta)$。当$T_n \to_P \theta$ 时，$R_n \to_P 0$。所以

$$\sqrt{n}[h(T_n) - h(\theta)] = \sqrt{n}(T_n - \theta)[h'(\theta) + R_n] \to_D N(0, [h'(\theta)]^2\tau^2(\theta)).$$

□

3. 若$\sqrt{n}(T_n - \theta) \to_D N(0, \tau^2(\theta))$，又$h'(\theta) = 0$，$h''(\theta)$ 存在且不等于0，则

$$n[h(T_n) - h(\theta)] \to_D \frac{1}{2}\tau^2(\theta)h''(\theta)\chi^2(1).$$

证明. 对$h(T_n)$ 在θ 处Taylor 展开

$$h(T_n) = h(\theta) + \frac{1}{2}(T_n - \theta)^2[h''(\theta) + R_n],$$

其中$R_n = O(T_n - \theta)$。当$T_n \to_P \theta$ 时，$R_n \to_P 0$。所以

$$n[h(T_n) - h(\theta)] = \frac{1}{2}n(T_n - \theta)^2[h''(\theta) + R_n] \to_D \frac{1}{2}\tau^2(\theta)h''(\theta)\chi^2(1).$$

□

例2.4 设$X_1, \ldots, X_n$ i.i.d. $B(1, p)$。

记$T_n = \bar{X}_n$，由中心极限定理，$\sqrt{n}(T_n - p) \to_D N(0, p(1-p))$。

考虑$T_n(1 - T_n)$，$T_n(1 - T_n) \cdot \dfrac{n}{n-1}$ 的渐近分布。

因为$h'(p) = 1 - 2p \begin{cases} \neq 0, & \text{if} \quad p \neq \frac{1}{2}, \\ = 0, & \text{if} \quad p = \frac{1}{2}. \end{cases}$

所以

$$\text{若 } p \neq \frac{1}{2}, \quad \sqrt{n}[T_n(1 - T_n) - p(1-p)] \to_D N(0, p(1-p)(1-2p)^2),$$

$$\text{若 } p = \frac{1}{2}, \quad n[T_n(1 - T_n) - p(1-p)] \to_D -p(1-p)\chi^2(1) = -\frac{1}{4}\chi_1^2.$$

$T_n(1 - T_n) \cdot \dfrac{n}{n-1}$ 有相同的渐近分布。

例2.5 设$X_1, X_2, \cdots, X_n$ i.i.d.。若已知$\mathrm{E}X_1 \hat{=} \mu, \mathrm{Var}X_1 \hat{=} \sigma^2$ 存在，则由CLT 知

$$\sqrt{n}(\overline{X} - \mu) \to_D N(0, \sigma^2),$$

即当n 较大时，有$\overline{X} \dot{\sim} N(\mu, \sigma^2/n)$。

例2.6 设$X_1, \ldots, X_n$ i.i.d. $N(\xi, 1)$。给定u，估计$p = P(X_1 \leq u)$。UMVUE为$\hat{p} = \Phi(\sqrt{\frac{n}{n-1}}(u - \bar{X}))$。其渐近分布是什么？渐近方差呢？

多变量情形

4. （多维中心极限定理）设$(X_{iv}, \ldots, X_{sv})', v = 1, \ldots, n,\ iid$，$EX_v = \xi = (\xi_1, \ldots, \xi_s)'$，$cov(X_v) = \Sigma$存在，记$\bar{X}_n = \frac{1}{n}\sum_{v=1}^n X_v$，则

$$\sqrt{n}(\bar{X}_n - \xi) \to_D N_s(0, \Sigma).$$

5. (δ-方法). 设h 是一个s 元实函数，可微，则

$$\begin{cases} \sqrt{n}[h(\bar{X}_n) - h(\xi)] \to_D N(0, v^2), & \text{if} \quad v^2 = \left(\frac{\partial h}{\partial \xi}\right)' \Sigma \left(\frac{\partial h}{\partial \xi}\right) > 0; \\ n[h(\bar{X}_n) - h(\xi)] \to_D ?, & \text{else.} \end{cases}$$

设$\sqrt{n}(T_n - \theta) \to_D N_s(0, \Sigma)$，$h = (h_1, \cdots, h_r)'$ 是$r \leq s$ 个θ 的实函数，在θ的一个邻域ω 内可微，偏微商$B = \left(\dfrac{\partial h_i}{\partial \theta_j}\right)_{r \times s}$ 在ω 中行满秩，则

$$\sqrt{n}[h(T_n) - h(\theta)] \to_D N_r(0, B\Sigma B').$$

例2.7 $(X_1, \ldots, X_n),\ iid$，X_1 的四阶矩存在，求$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ 的渐近分布。

记$\overline{X^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$。因为$S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = h(\bar{X}, \overline{X^2})$，其中$h(x, y) = y - x^2$，故由CLT知，

$$\left[\begin{pmatrix} \bar{X} \\ X^2 \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}\right] \to_D N_2\left[0, \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix}\right],$$

其中α_k 是X_1 的k 阶原点矩，记μ_k 是X_1 的k 阶中心矩。

因为$\frac{\partial h}{\partial (x\ y)} = (-2x\ 1)$，由$\delta$方法知，

$$\sqrt{n}(h(\bar{X}, \overline{X^2}) - h(\alpha_1, \alpha_2)) \to_D N(0, v^2),$$

其中$v^2 = \begin{pmatrix} -2\alpha_1 & 1 \end{pmatrix} \Sigma \begin{pmatrix} -2\alpha_1 \\ 1 \end{pmatrix}$。即$\sqrt{n}(S_n^2 - \mu_2) \to_D N(0, \mu_4 - \mu_2^2)$.

进一步还可证得

$$\sqrt{n}(\overline{X} - \mu, S^2 - \sigma^2) \to_D N_2(0, V),$$

其中

$$V = \begin{pmatrix} \mu_2^2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}.$$

注意： 有些统计推断问题中需要研究其他形式的极限分布，如极值统计量的渐近分布等，教材p235/eg5.5.11即是一例。这方面的内容本课程中不展开讨论。

矩的极限

用δ_n 去估计$g(\theta)$，若当$n \to \infty$ 时，有$E\delta_n \to g(\theta),\ \forall \theta \in \Theta$，则称$\delta_n$ 极限无偏。

研究中经常关心：偏倚$E\delta_n - g(\theta)$、方差Varδ_n 趋于0 的速度。

基本方法：

1. 设$X_1, \ldots, X_n$ i.i.d.，$EX_1 = \xi$，Var$X_1 = \sigma^2 < \infty$，$\mu_k = E(X_1 - \xi)^k, k = 3, 4$。则$E\bar{X} = \xi$，$E(\bar{X} - \xi)^2 = \sigma^2/n$，$E(\bar{X} - \xi)^3 = \dfrac{\mu_3}{n^2}$，$E(\bar{X} - \xi)^4 = \dfrac{\mu_4}{n^3} + \dfrac{(n-1)\sigma^4}{n^3}$。
2. 设$X_1, \ldots, X_n$ i.i.d.，I 为X_1 的支撑，$EX_1 = \xi$，Var$X_1 = \sigma^2$，X_1 的四阶矩有限，$h(x)$ 在I 上有4阶导数，$|h^{(4)}(x)| \le M < \infty, \forall x \in I$，则

$$Eh(\bar{X}) = h(\xi) + \frac{\sigma^2}{2n} h''(\xi) + R_n.$$

若h^2 的四阶导数也有界，则

$$\text{Var}h(\bar{X}) = \frac{\sigma^2}{n}[h'(\xi)]^2 + R_n.$$

其中$R_n = O(\frac{1}{n^2})$。

注：

- 对$k \geq 3$，若函数h 有k 阶导数，且第k 阶导数有界，则上述$Eh(\bar{X})$, $\mathrm{Var}h(\bar{X})$ 的公式成立。
- 相比渐近分布的δ方法，对于$h(\cdot)$ 的要求高，不适用于$h(x) = \frac{1}{x}$ 或$\sqrt{x}$ 等情况。
- 若$c_n = 1 + \frac{a}{n} + O(n^{-2})$，则$\delta_n = h(c_n\bar{X})$ 的方差为

$$\mathrm{Var}(\delta_n) = \frac{\sigma^2}{n}[h'(\xi)]^2 + O(\frac{1}{n^2}).$$

例2.8 1. 求例2.6 中估计量$\Phi\left(\sqrt{\frac{n}{n-1}}(u - \bar{X})\right)$ 的方差趋于0 的速度。

记$\delta_n = h(c_n \cdot \bar{Y}_n) = \Phi(c_n \cdot \bar{Y}_n)$，其中$\bar{Y}_n = u - \bar{X}_n$，$c_n = (1 - \frac{1}{n})^{-\frac{1}{2}} = 1 + \frac{1}{2n} + O(\frac{1}{n^2})$。

只需验证$h^2(\cdot)$ 的前四阶导存在、有界，即有

$$\mathrm{Var}\delta_n = \frac{1}{n}\phi^2(u - \xi) + O(\frac{1}{n^2}).$$

2. 设$X_1, \ldots, X_n$ i.i.d. $f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}I(x > 0)$，其中$\theta > 0$。考察$\sqrt{\bar{X}}$ 的方差趋于0 的速度。

显然，$EX_i = \theta$, $\mathrm{Var}X_i = \theta^2$。因为$h(x) = \sqrt{x}$，所以不能用方法 2求，但可精确计算得

$$\mathrm{Var}(\sqrt{\bar{X}}) = \left[1 - \frac{1}{n}\left(\frac{\Gamma(n + \frac{1}{2})}{\Gamma(n)}\right)^2\right]\theta,$$

因此

$$\lim_{n\to\infty} n \cdot \mathrm{Var}(\sqrt{\bar{X}}) = \frac{\theta}{4} = \theta^2[h'(\theta)]^2.$$

一般地，估计量经恰当正则化后，方差的极限大于或等于其极限分布的方差。下列引理说明了这一点。

引理2.1 *设$Y_n \to_D Y$，其中$EY = 0$, $VarY = EY^2 = v^2 < \infty$。对常数$A$ 定义$Y_{nA} = Y_nI(|Y_n| \leq A) + AI(|Y_n| > A)$。则*

1. $\lim_{A\to\infty}\lim_{n\to\infty} EY_{nA}^2 = v^2$;
2. *若$EY_n^2 \to w^2$，则$w^2 \geq v^2$。*

注意： 渐近分布的矩与矩的极限之间的区别。有些情况下，随机变量序列矩的极限不存在，但是其渐近分布的矩存在。（反例？）

三、随机模拟方法求抽样分布

基本原理：LLN，格里汶科定理等

伪随机数的生成（参见教材p245–255）：

1. $U(0,1)$ 的伪随机数是生成其他分布伪随机数的基础；
2. 直接法。常用变换。
3. 间接法。Accept-Reject 算法，MCMC（Metroplis-Hasting 算法，Gibbs sampler）