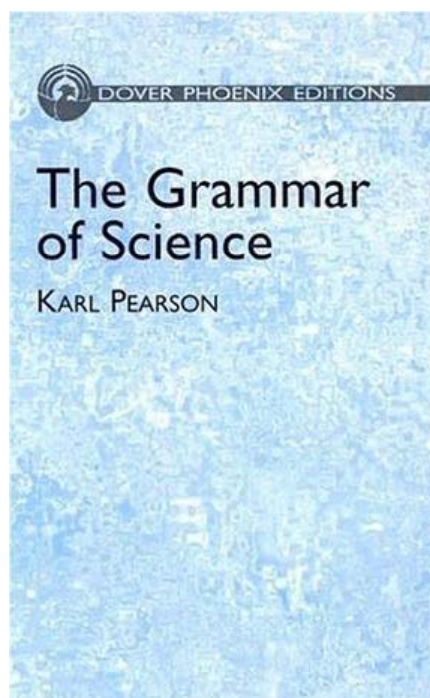




Zhu Huaqiu  
@Peking University



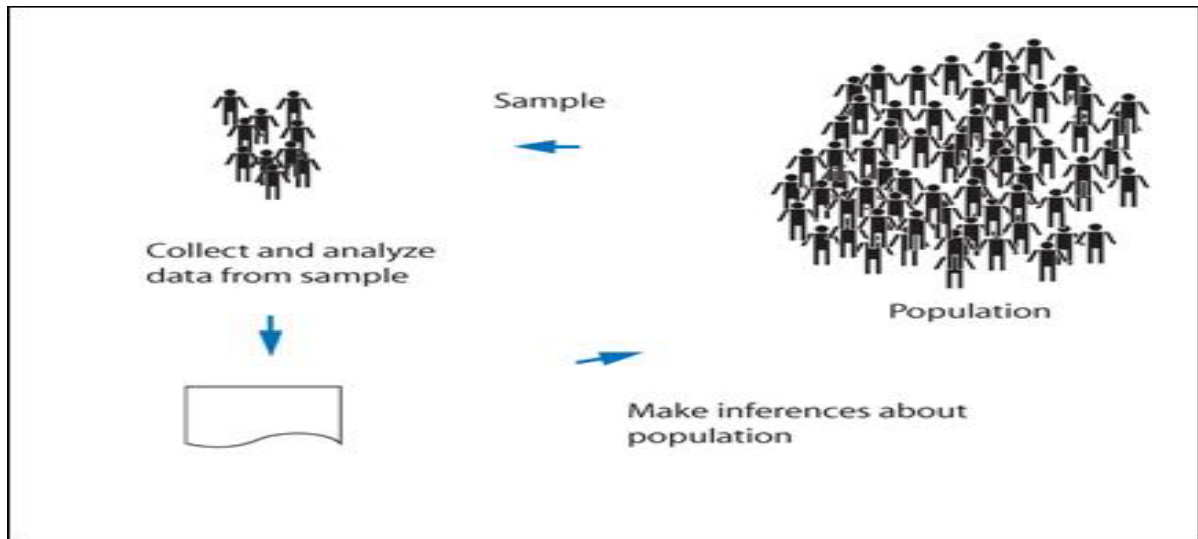
The scientific method has the following distinctive features:

- (a) careful and accurate measurement of data, and observation of their correlation and sequence;
- (b) discovery of scientific laws by aid of the creative imagination;
- (c) self-criticism;
- (d) final decisions having equal validity for all normally constituted minds.

—Karl Pearson (*The Grammar of Science* in 1892)

## 推断统计学 (Inferential statistics) 或统计推断 (Statistical inference)

Inference about a population from a random sample drawn from it or, more generally, about a random process from its observed behavior during a finite period of time.



统计学推断  
的重要理论

假设检验理论

参数估计理论

统计抽样理论

概率论

# 参数估计和假设检验

## (1) 参数估计 (Parameter estimation) :

当总体的分布函数类型已知，如何利用样本数据对其中的一个或多个未知参数进行估计的问题，以及如何对估计的参数进行评价。

## (2) 假设检验 (Hypothesis test) :

根据抽取的样本信息来判定总体是否具有某种性质。

## § 5.1 统计推断的经典例子

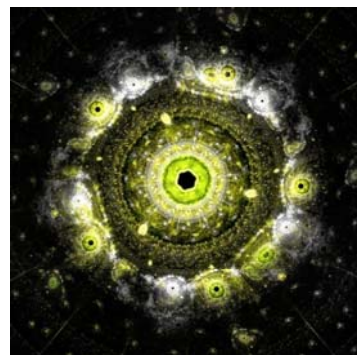
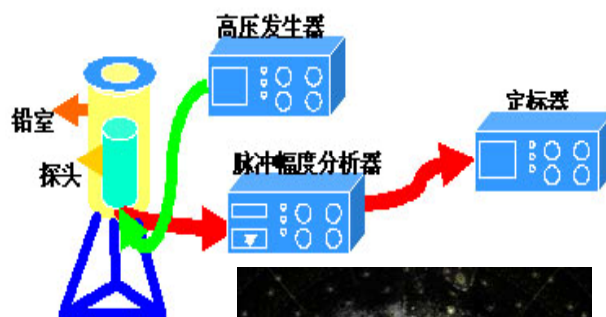
### ——放射性衰变的Poisson分布

#### 放射性衰变的 $\alpha$ 粒子数的Poisson分布:

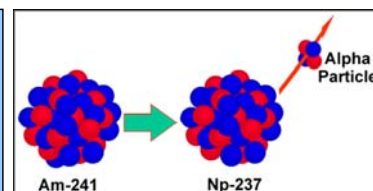
实验测量和统计研究表明，一段时间内某放射物质放射出的 $\alpha$ 粒子数具有明显的随机性，可用Poisson分布来很好地描述。

满足下列条件：

- (1) 在某一段时间或某一空间范围内，事件发生的速率是恒定的；
- (2) 在不同的时间段或空间范围，事件的发生是独立的；
- (3) 所有事件不能同时发生。



【Example 5.1】Berkson对美国国家标准局的一组数据进行分析。测量数据的放射源为镅（americium）241，每10s记录一次测得的 $\alpha$ 粒子数 $k$ ，共计1027次，平均每次的 $\alpha$ 粒子数为8.392。

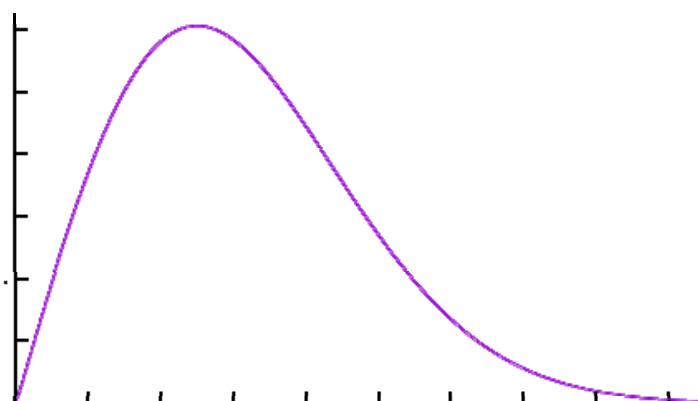


$k$	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17+
Observed	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5

**估计参数：**根据观察的数据来估计Poisson分布的参数 $\lambda$

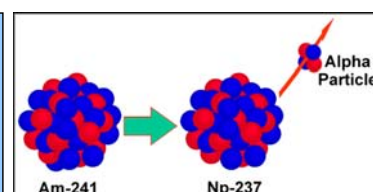
$$\pi_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\hat{\lambda} = 8.39 \text{ (估计值)}$$



(Berkson J. (1966). Examination of randomness of alpha particle emission. In *Research Papers in Statistics*, New York: Wiley)

【Example 5.1】Berkson对美国国家标准局的一组数据进行分析。测量数据的放射源为镅（americium）241，每10s记录一次测得的 $\alpha$ 粒子数 $k$ ，共计1027次，平均每次的 $\alpha$ 粒子数为8.392。



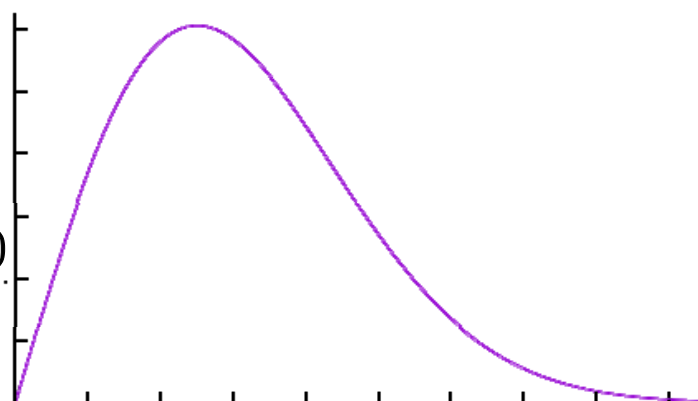
$k$	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17+
Observed	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5

**问题：**

**1 参数估计的数学方法**

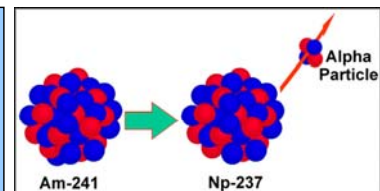
**2 估计值的概率分布性质**  
抽样分布(sampling distribution)

**3 总体分布假设的检验**  
非正态总体



(Berkson J. (1966). Examination of randomness of alpha particle emission. In *Research Papers in Statistics*, New York: Wiley)

【Example 5.1】Berkson对美国国家标准局的一组数据进行分析。测量数据的放射源为镅（americium）241，每10s记录一次测得的 $\alpha$ 粒子数 $k$ ，共计1027次，平均每次的 $\alpha$ 粒子数为8.392。



$k$	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17+
Observed	18	28	56	105	126	146	164	161	123	101	74	53	23	15	9	5
Expected	12.2	27.0	56.5	94.9	132.7	159.1	166.9	155.6	130.6	99.7	69.7	45.0	27.0	15.1	7.9	7.1
$\chi^2$	2.76	0.04	0.01	1.07	0.34	1.08	0.05	0.19	0.44	0.02	0.27	1.42	0.59	0.00	0.57	0.57

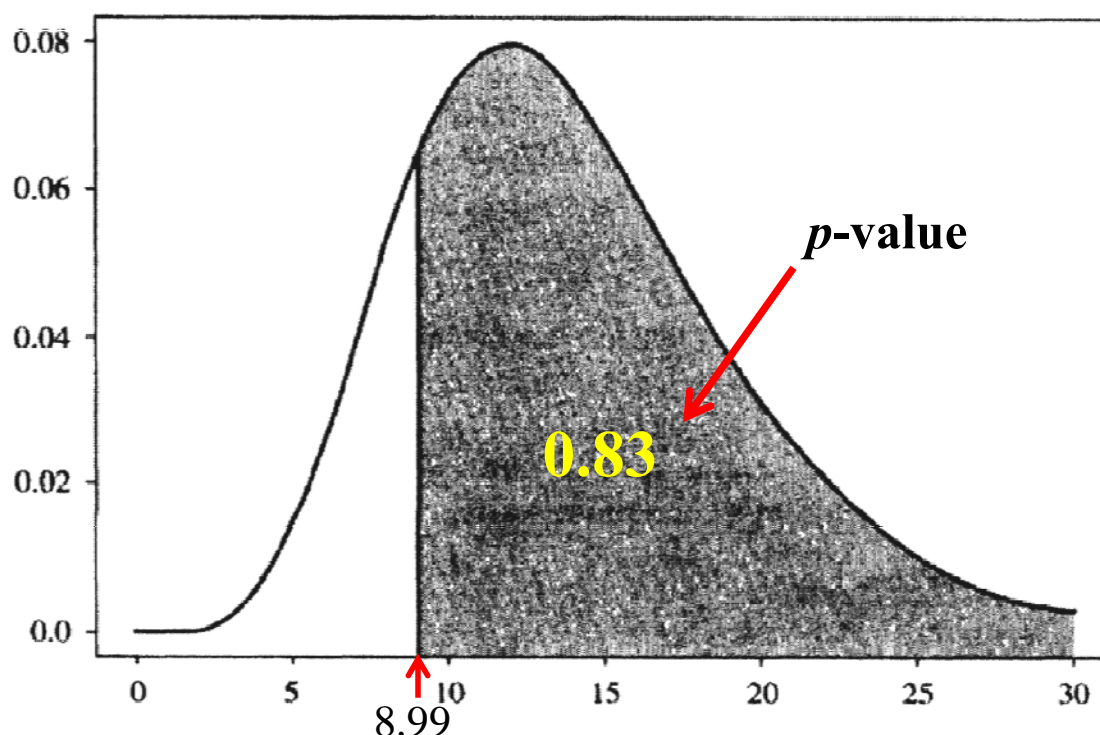
#### 4 总体分布的拟合检验

$$X^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i^2} \quad (\text{Pearson } \chi^2 \text{ 统计量})$$

$$= 8.99$$

(Berkson J. (1966). Examination of randomness of alpha particle emission. In *Research Papers in Statistics*, New York: Wiley)

#### $\chi^2(n=14)$ 分布的概率密度





## § 5.2 参数估计概述

当总体的分布函数类型已知时

- (1) 如何利用样本数据对其中的一个或多个未知参数进行估计
- (2) 如何对估计的参数进行评价

两个问题都很重要！

### § 5.2.1 正态分布总体的参数估计

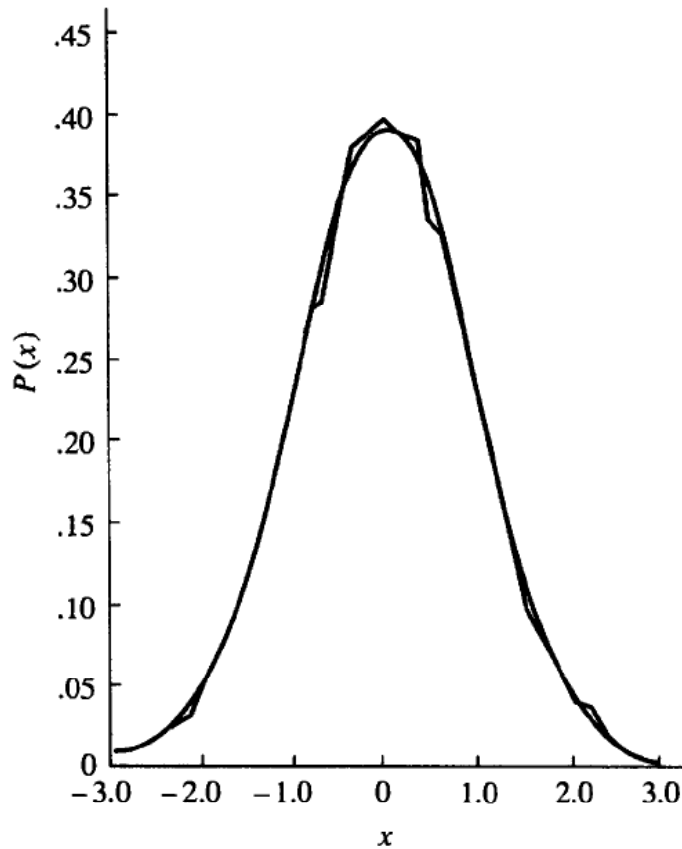
$X \sim N(\mu, \sigma^2)$

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (-\infty < x < +\infty),$$

其中 $\mu, \sigma$ 为待定的未知参数且 $\sigma > 0$ 。

**【Example 5.2】**Bevan等（1979）对细胞膜的离子通道电流量的实验测定值进行分析。肌肉和神经细胞的膜上有大量的离子通道供所需的离子穿过。一般认为在平衡状态下只有少部分的通道是打开的（数量仍然很大），每一通道的打开和闭合状态是随机且相互独立的。因此，某一时刻的电流即表现为大量相互独立发生的事件（离子穿过通道）的总和，随着各通道的随机的开合状态变化，电流也表现出随机的涨落。

Gaussian fit of current flow across a cell membrane to a frequency polygon.



(Bevan S., Kullberg R. and Rice J. (1979) An analysis of cell membrane noise. *Annals of Statistics*, 7: 237-257)

## § 5.2.2 Gamma分布总体的参数估计

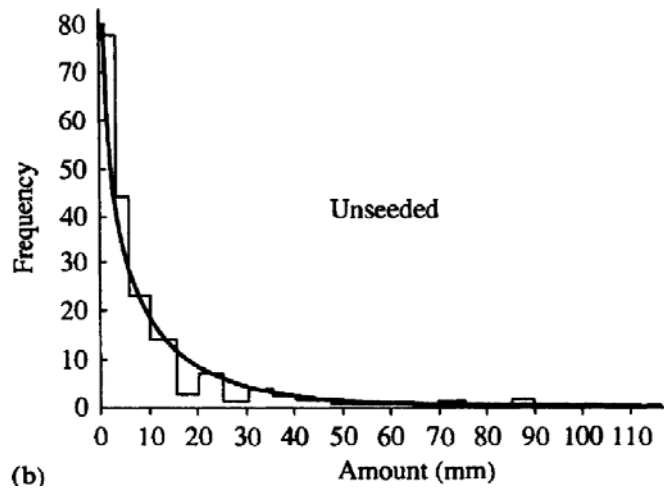
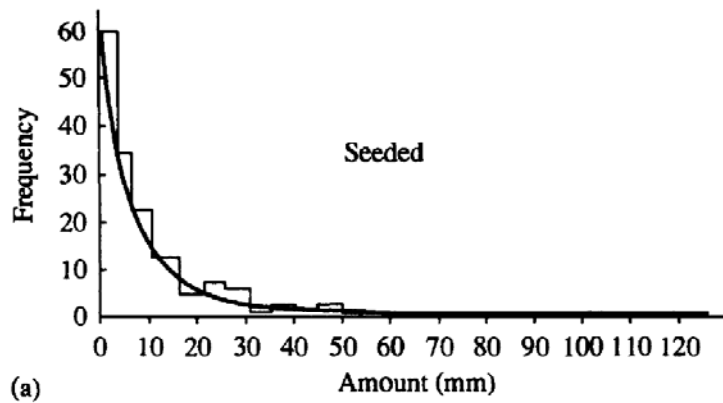
$$f(x | \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad \alpha, \lambda > 0 \text{ 为常数。}$$

包含两个待定的未知参数 $\alpha, \lambda$ 。

**【Example 5.3】** Le Cam等对一组人工催雨降雨量的数据和一组自然降雨的数据分别进行Gamma分布的拟合，两组数据的差别可以反映在估计的参数 $\alpha, \lambda$ 。

(Le Cam L. and Neyman J. (eds.) (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley: University of California Press )

Fit of gamma densities to amounts of rainfall for (a) seeded and (b) unseeded storms.



## § 5.2.3 参数估计的问题描述

$X_1, X_2, \dots, X_n$ : 观察得到的样本（随机变量），联合概率是未知参数 $\theta$ 的函数（ $\theta$ 可以是矢量）

**独立同分布** (I.I.D., independent and identically distribution)

$$f(x_1 | \theta) \cdot f(x_2 | \theta) \cdot \dots \cdot f(x_n | \theta)$$

构造 $X_1, X_2, \dots, X_n$ 的函数 $T(X_1, X_2, \dots, X_n)$ 作为 $\theta$ 的估计量 $\tilde{\theta}$ ，即

$$\tilde{\theta} = T(X_1, X_2, \dots, X_n)$$

显然 $\tilde{\theta}$ 也是**随机变量**，其概率分布被称为**抽样分布** (sampling distribution)。对估计量抽样分布的了解通常是求 $\theta$ 的期望值、方差或者标准差（也称标准误，standard error）或者其概率分布。



## 参数估计的任务

---

- (1) 求解 $\theta$ 的估计量 $\tilde{\theta}$
- (2) 估计量 $\tilde{\theta}$ 的抽样分布

## 参数估计的主要方法

- 矩方法 (method of moment)
- 极大似然方法 (method of maximum likelihood)
- Bayes估计方法

“最优估计”原则：

估计量 $\tilde{\theta}$ 最接近 $\theta$ 且分布最集中

比较：总体参数的估计问题

## § 5.3 矩方法 (The Method of Moment)

---

### 定义

设 $X$ 为随机变量，若 $E(X^k)$  ( $k=1, 2, \dots$ ) 存在，则称 $E(X^k)$ 是 $X$ 的**k阶原点矩**，简称**k阶矩**，记为 $\mu_k$ 。

若 $X_1, X_2, \dots, X_n$ 为取自 $X$ 的一组样本（即独立同分布），则定义

$$\tilde{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

为**k阶样本原点矩**，简称**k阶样本矩**。

## 基本思想（背景：大数定律）

将k阶样本矩作为k阶矩的估计量，从而得到待定参数的表达式。

由大数定律，若总体X的期望值 $E(\bar{X})$ 存在， $X_1, X_2, \dots, X_n$ 为取自X的一组样本，则当 $n \rightarrow \infty$ 时 $\bar{X}$ 依概率收敛于 $E(X)$ 。

更一般地，若总体X的k阶矩 $\mu_k$ 存在，当n充分大时，可以用k阶样本矩 $\tilde{\mu}_k$ 作为 $\mu_k$ 的估计，并由此得到未知参数的估计量。

$$\tilde{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{n \rightarrow \infty} \mu_k = E(X^k)$$

设总体X的概率密度函数为 $f(x | \theta_1, \theta_2, \dots, \theta_k)$ ，其中 $\theta_1, \theta_2, \dots, \theta_k$ 为待定的未知参数， $X_1, X_2, \dots, X_n$ 为取自X的一组样本（即独立同分布）， $x_1, x_2, \dots, x_n$ 为其观察值，且样本的前k阶样本矩 $\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_k$ 存在。

总体的前k阶矩可表达成如下封闭的方程组：

$$\begin{aligned}\mu_1 &= h_1(\theta_1, \theta_2, \dots, \theta_k) \\ \mu_2 &= h_2(\theta_1, \theta_2, \dots, \theta_k) \\ &\dots \\ \mu_k &= h_k(\theta_1, \theta_2, \dots, \theta_k)\end{aligned}$$

$$\theta_1 = g_1(\mu_1, \mu_2, \dots, \mu_k)$$

$$\theta_2 = g_2(\mu_1, \mu_2, \dots, \mu_k)$$

...

$$\theta_k = g_k(\mu_1, \mu_2, \dots, \mu_k)$$

方程的解



$$\tilde{\theta}_1 = g_1(\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_k)$$

$$\tilde{\theta}_2 = g_2(\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_k)$$

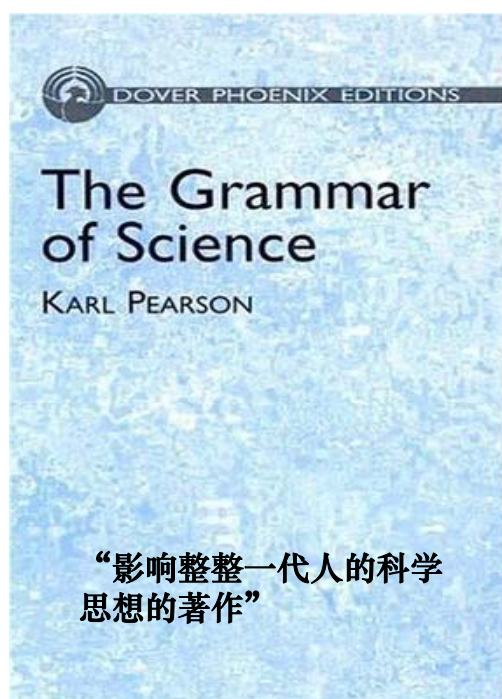
...

$$\tilde{\theta}_k = g_k(\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_k)$$

以样本矩表示的方程解  
表达式——  
估计量

**Karl Pearson (1857-1936)**

英国统计学家



“最重要、最具有创造性的统计学家之一”

## § 5.3.1 实例

### 一、Poisson分布

Poisson分布 $X \sim P(\lambda)$ 的分布律:

$$P(X = k | \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0, 1, 2, \dots, \lambda > 0$$

$\lambda$ 为待定的未知参数。

矩估计解:

$$\tilde{\lambda} = \overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

### 矩估计解的抽样分布:

直接得到估计值 $\tilde{\lambda}$ 的概率分布表达式:

设Poisson分布的总体 $X$ 的参数为 $\lambda_0$ ,  $X_1, X_2, \dots, X_n$ 为取自 $X$ 的一组独立同分布样本, 可以证明,  $S = \sum X_i$ 服从参数为 $(n\lambda_0)$ 的Poisson分布。

$$E(\tilde{\lambda}) = \frac{1}{n} E(S) = \lambda_0$$

$$D(\tilde{\lambda}) = \frac{1}{n^2} D(S) = \frac{\lambda_0}{n}$$

分布形式: (1) Poisson分布; (2) 近似服从正态分布

## 讨论：

(1) 对于期望值等于真实值 $\lambda_0$ 的估计值 $\tilde{\lambda}$ ，定义该估计是**无偏估计**。

(2) 估计值 $\tilde{\lambda}$ 的方差是总体方差的 $1/n$ ，更为集中；

(3) 定义估计值 $\tilde{\lambda}$ 的标准差为**标准误** (standard error) ，即：

$$\sigma_{\tilde{\lambda}} = \sqrt{\lambda_0/n}$$

大多数情况下，参数的真实值 $\lambda_0$ 未知，不妨用 $\lambda_0$ 的估计值 $\tilde{\lambda}$ 代替，即

$$s_{\tilde{\lambda}} = \sqrt{\tilde{\lambda}/n}$$

称为**估计标准误** (estimated standard error) 。

**【Example 5.4】**用电子显微镜测量样本中的纤维数目来计算石棉纤维材料的浓度。Steel等（1980）的分析表明测得的纤维数目能很好地服从Poisson分布。以一组数据为例：

31, 29, 19, 18, 31,  
28, 34, 27, 34, 30,  
16, 18, 26, 27, 27,  
18, 24, 22, 28, 24,  
21, 17, 24

求参数估计解及其抽样分布。

( Steel E., Small J., Leigh S., and Filliben J. (1980). Statistical consideration in the preparation of chrysotile filter standard reference materials. *NBS Technical Report (Washington, D. C.)* )

## 二、正态分布

---

正态分布 $X \sim N(\mu, \sigma^2)$ 的概率密度函数为

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (-\infty < x < +\infty),$$

其中 $\mu, \sigma$ 为待定的未知参数且 $\sigma > 0$ 。

矩估计解：

$$\begin{aligned}\tilde{\mu} &= \overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \\ \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2\end{aligned}$$

矩估计解的抽样分布：

---

$$\tilde{\mu} = \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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



### 三、Gamma分布

---

Gamma分布 $X \sim \Gamma(\alpha, \lambda)$ 的概率密度函数为

$$f(x | \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

其中 $\alpha, \lambda > 0$ 为待定的未知参数。

矩估计解：

---

$$\begin{array}{ccc} \lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} & \longrightarrow & \tilde{\lambda} = \frac{\overline{X}}{\tilde{\sigma}^2} \\ \alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2} & & \tilde{\alpha} = \frac{\overline{X^2}}{\tilde{\sigma}^2} \end{array}$$

矩估计解的抽样分布：

直接导出解析式：过于复杂

近似方法：计算机模拟——自举法（Bootstrap resampling method）



To pull oneself up by one's bootstraps

《吹牛大王历险记》（蒙乔森男爵历险记）Rudolph Erich Raspe

## 自举法（自助法） （ Bootstrap resampling ）

——以原始数据为基础的模拟抽样统计推断法

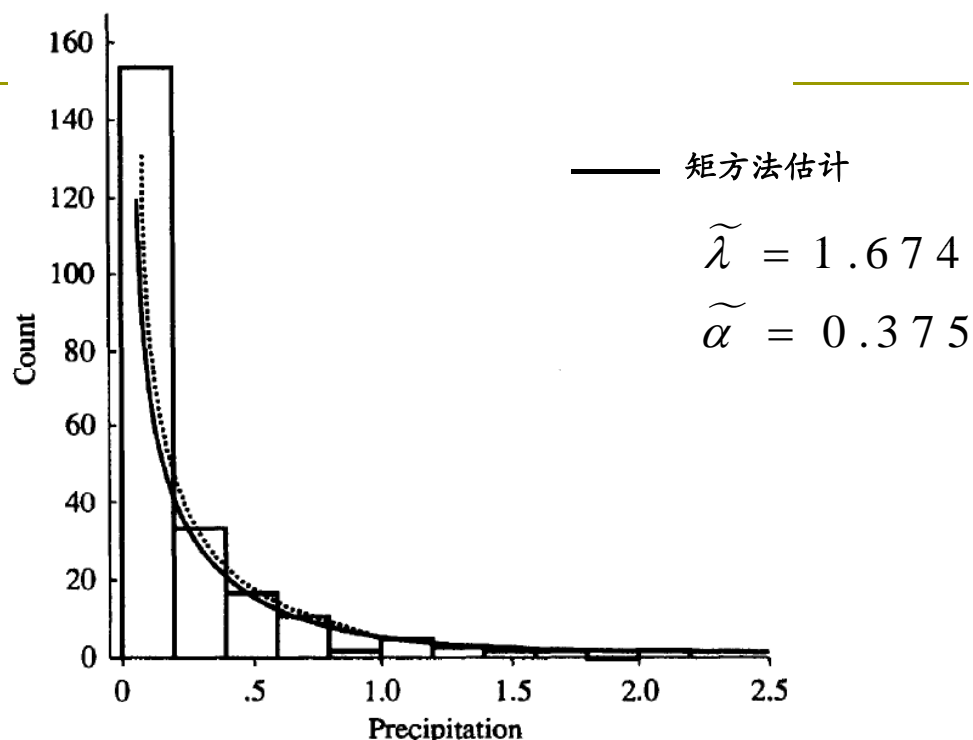
基本思想：

（1）在原始数据的范围内作有放回的再抽样，样本含量仍为 $n$ ，原始数据中每个观察单位每次被抽到的概率相等，为 $1/n$ ，所得样本称为bootstrap样本。

（2）根据bootstrap样本，可得到参数 $\theta$ 的估计值 $\tilde{\theta}^{(b)}$ ，重复 $B$ 次，得到该参数的 $B$ 个估计值。

（3）根据 $B$ 个估计值 $\tilde{\theta}^{(b)}$ ，研究其分布，此即估计量的抽样分布。

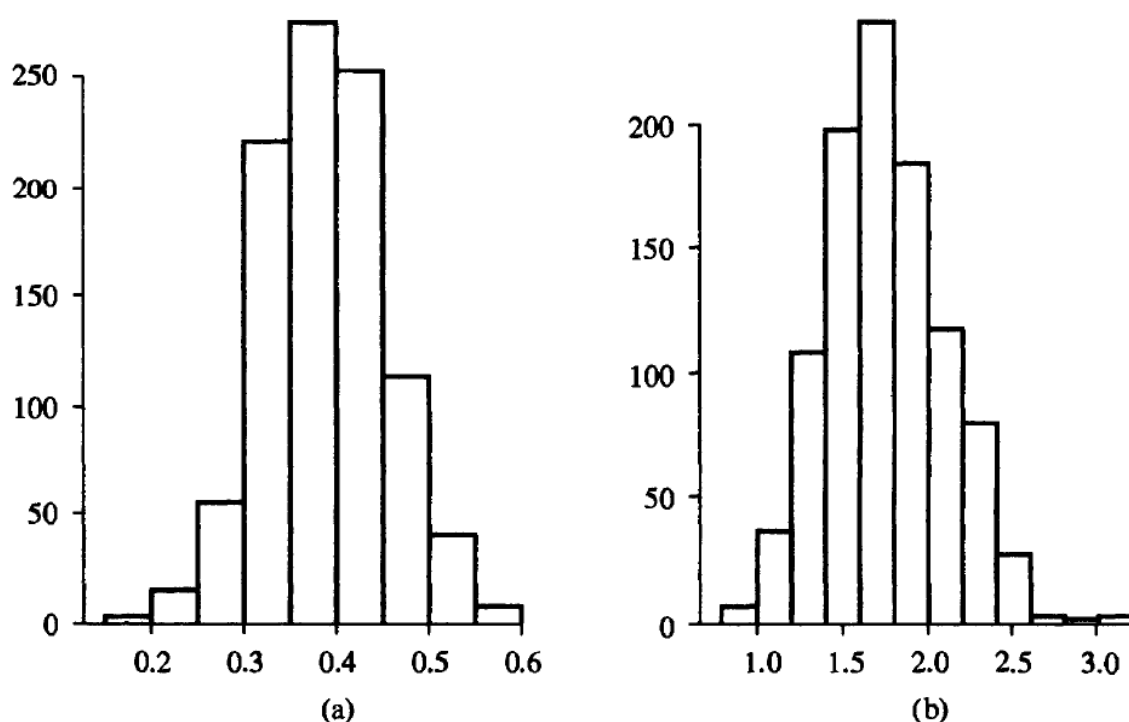
【Example 5.5】Le Cam等对美国Illinois州南部地区1960-64年期间的227场暴雨的降雨量数据进行Gamma分布的拟合。



(Le Cam L. and Neyman J. (eds.) (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley: University of California Press)

【Example 5.5】Le Cam等对美国Illinois州南部地区1960-64年期间的227场暴雨的降雨量数据进行Gamma分布的拟合。

Histogram of 1000 simulated method of moment estimates of (a)  $\alpha$  and (b)  $\lambda$ .



(Le Cam L. and Neyman J. (eds.) (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley: University of California Press)

## THE 1977 RIETZ LECTURE

### BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE

BY B. EFRON

Stanford University

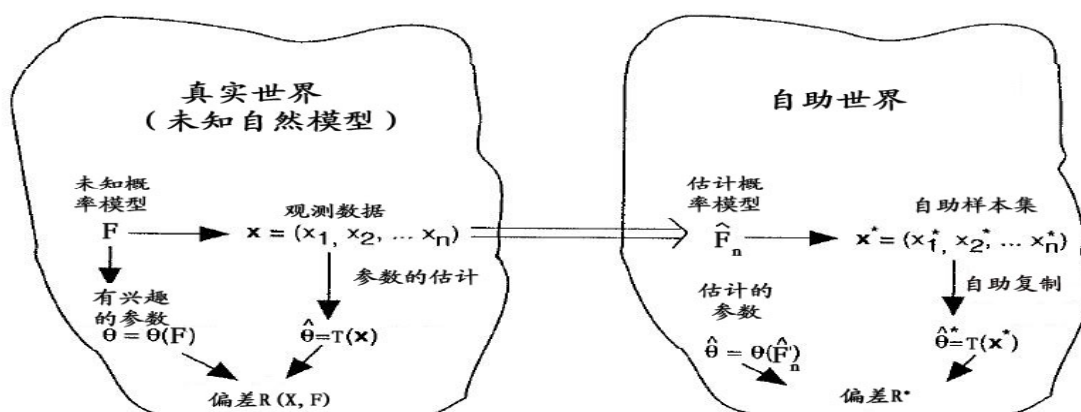
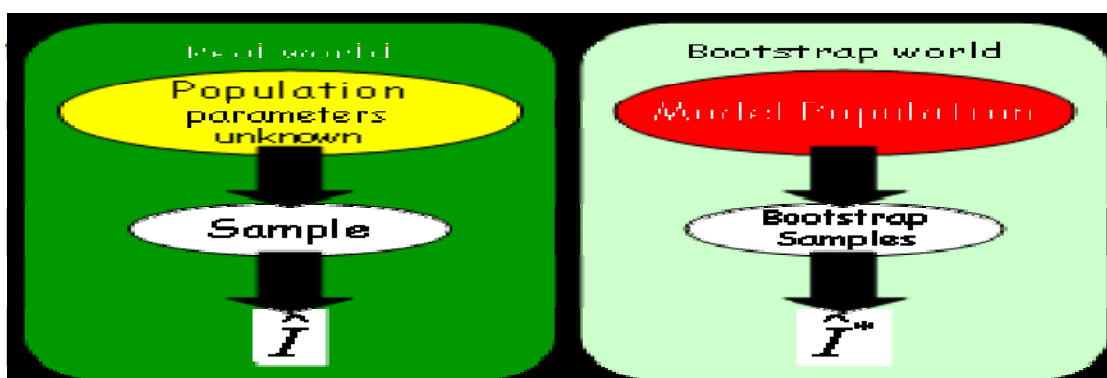
We discuss the following problem: given a random sample  $X = (X_1, X_2, \dots, X_n)$  from an unknown probability distribution  $F$ , estimate the sampling distribution of some prespecified random variable  $R(X, F)$ , on the basis of the observed data  $x$ . (Standard jackknife theory gives an approximate mean and variance in the case  $R(X, F) = \theta(\hat{F}) - \theta(F)$ ,  $\theta$  some parameter of interest.) A general method, called the "bootstrap," is introduced, and shown to work satisfactorily on a variety of estimation problems. The jackknife is shown to be a linear approximation method for the bootstrap. The exposition proceeds by a series of examples: variance of the sample median, error rates in a linear discriminant analysis, ratio estimation, estimating regression parameters, etc.

Brad Efron

Max H. Stein Professor and Professor of  
Statistics and of Health Research and Policy,  
Department of Statistics, Stanford University



样本的信息通过计算机模拟的方式可以反复加以利用，从而减少统计推断偏差，并依靠数据本身产生临界值。



## 关于Bootstrap方法的讨论：

---

- (1) 数学基础：原样本集是来自总体统计模型的独立同分布样本；
- (2) 目的：用Bootstrap样本集上的统计量的分布去逼近原样本集上统计量的分布；
- (3) Bootstrap方法通过经验分布函数构建了Bootstrap world，将不适定的估计概率分布的问题转化为从给定样本集中重抽样（resampling）问题；
- (4) Bootstrap方法可以解决不光滑（non-smooth）参数的问题。遇到不光滑参数估计时，可以有效地给出中位数的估计；

## 关于Bootstrap方法的讨论：

---

- (5) Bootstrap方法不需要对未知自然模型做任何假设，也无需事先推导出估计量的精确解析式，只需重抽样并计算估计值。本质上是一种非参数方法。

在实现过程中，如果没有计算机，Bootstrap方法理论只能是纸上谈兵。

（“计算机时代的Computer-intensive的统计方法”）

由于对于自然模型参数的一些估计无法得到明确的解析式，Bootstrap方法的出现使得人们绕过这种繁琐的理论推导。



## § 5.3.2 关于矩方法的进一步讨论

### 一、参数估计的一致性（或相合性，consistency）

大数定律

定义

令  $\tilde{\theta}_n$  为参数  $\theta$  基于容量为  $n$  的样本的估计值。如果  $n$  趋向无穷大时， $\tilde{\theta}_n$  依概率收敛到  $\theta$ ，则称  $\tilde{\theta}_n$  是  $\theta$  的一致性估计。即对任意的  $\varepsilon > 0$ ，当  $n \rightarrow \infty$  时，有

$$P\left(\left|\tilde{\theta}_n - \theta\right| > \varepsilon\right) \rightarrow 0$$

根据一致性的定义，还有：

$$P\left(\left|\sigma(\tilde{\theta}_n) - \sigma(\theta_0)\right| > \varepsilon\right) \rightarrow 0$$

其中  $\theta_0$  是  $\theta$  的真实值。

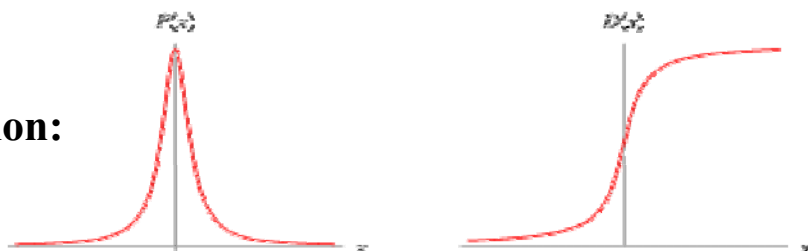
### 二、矩方法的优缺点

(1) 优点：直观、简单。

(2) 缺点：当总体矩不存在时，不能使用；对某些总体参数的估计值可能不唯一。

(3) 矩方法只是利用了样本矩的信息，没有充分利用总体分布函数  $F(x|\theta)$  的信息，因此可能不是最优的方法。

Cauchy distribution:

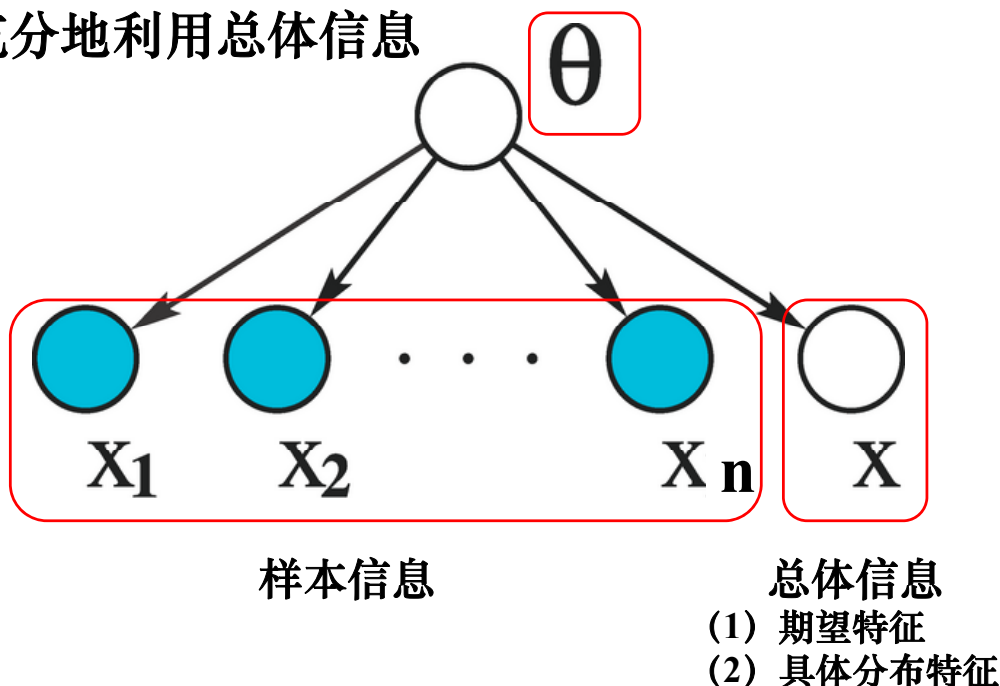




## § 5.4 最大似然估计方法

(MLE, Maximum Likelihood Estimation)

如何更充分地利用总体信息



**Likelihood: the chance that something will happen**

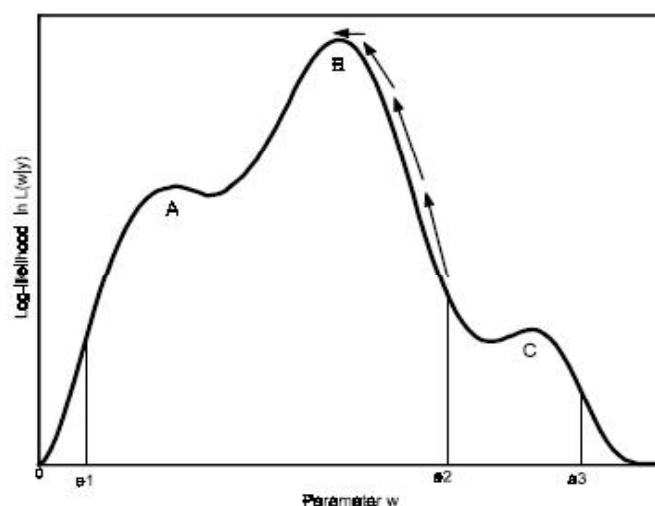
“似然”：可能性

—— “最大可能性估计”

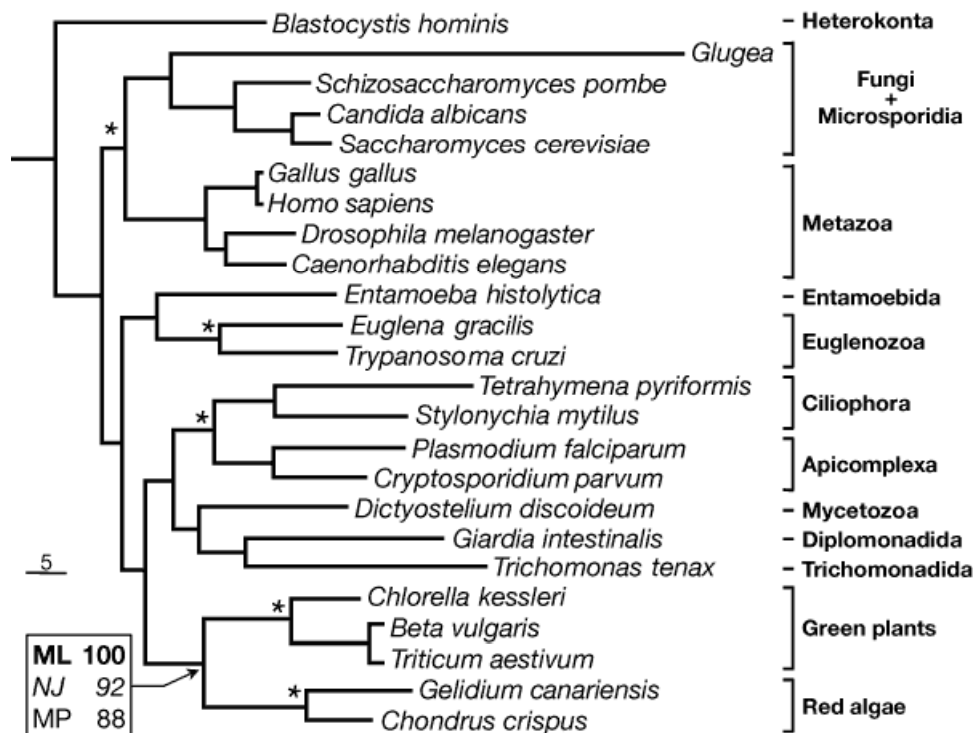
——通过对样本（观察）数据的最大似然估计，来寻找待拟合的模型参数的最优解

——MLE具有比矩估计方法更扎实的理论基础

（矩方法损失了随机变量 $X$ 总体的分布信息）



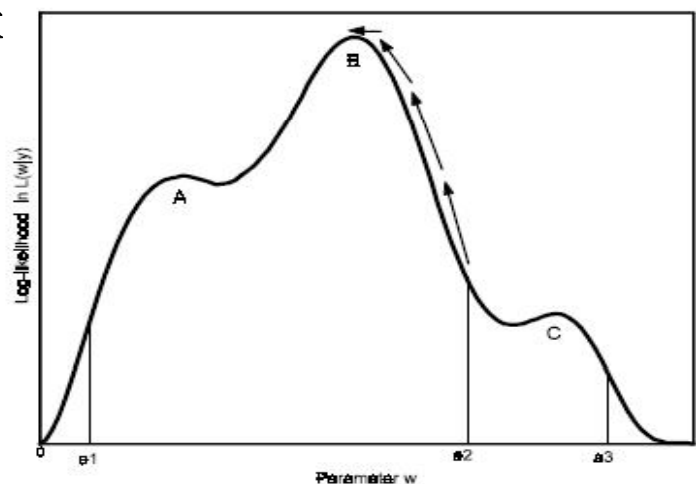
# MLE应用于生物分子系统发生树的构建



## § 5.4.1 MLE方法概述

——利用总体的分布函数表达式以及样本所提供的信息，来建立未知参数的基于最大似然原理的估计量

——**最大似然原理**：一个随机试验如有若干个可能的结果 $E_1, E_2, \dots, E_i, \dots$ ，若在一次试验中观察到结果 $E_i$ 发生，则一般认为 $E_i$ 发生的概率（可能性）最大



已知样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布（概率密度函数为 $f(x|\theta)$ ）的随机变量， $x_1, x_2, \dots, x_n$ 为其观察值，它们的联合概率为：

$$\prod_{i=1}^n f(x_i | \theta) = f(x_1 | \theta) \cdot f(x_2 | \theta) \cdot \dots \cdot f(x_n | \theta)$$

设 $\Theta$ 为 $\theta$ 的取值范围（即参数空间），定义 **$\theta$ 的似然函数** (likelihood function) 为：

$$lik(\theta) = \prod_{i=1}^n f(X_i | \theta), \quad \theta \in \Theta$$

则 **$\theta$ 的最大似然估计值就是使得似然函数最大的值**，也就是使得样本观察值 $x_1, x_2, \dots, x_n$ 为最大可能的值。

为计算方便，采用**对数似然函数** (log likelihood function)：

$$l(\theta) = \sum_{i=1}^n \log(f(X_i | \theta)), \quad \theta \in \Theta$$

$\theta$ 的估计量 $\tilde{\theta}$ 满足：

$$l(\tilde{\theta}) = \max_{\theta \in \Theta} [l(\theta)] = \max_{\theta \in \Theta} \left[ \sum_{i=1}^n \log(f(X_i | \theta)) \right]$$

若 $l(\theta)$ 对 $\theta$ 可导，则可用微积分求极值的方法计算估计值。令

$$\frac{d}{d\theta} l(\theta) = 0$$

解出上述方程（组）的解。

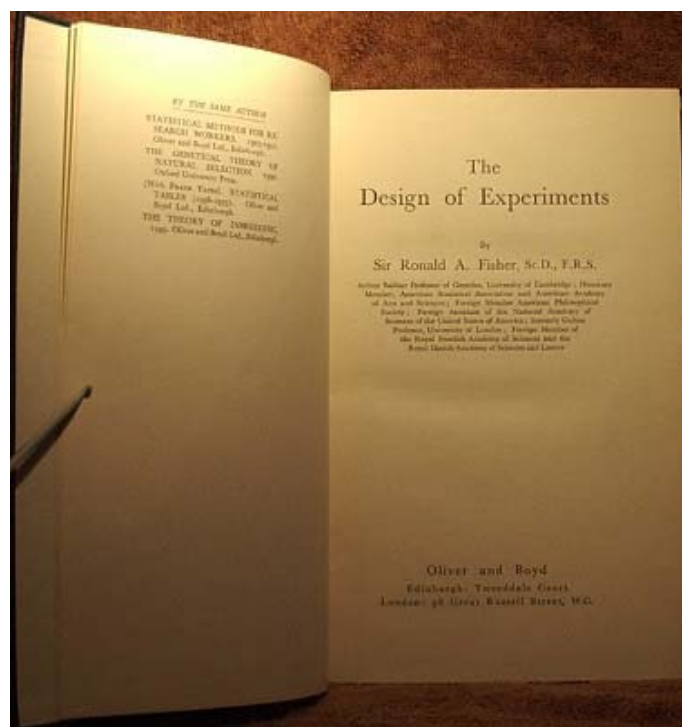
若为非线性方程组时，需采用数值解法。

**R. A. Fisher**  
**(1890-1962)**  
**英国统计学家、遗传学家**



剑桥大学Caius College宴会厅里的染色玻璃窗，上方的彩绘方格用以纪念拉丁方阵 (Latin square)，下方的白色文字则是为了纪念R. Fisher。

**R. A. Fisher**  
**(1890-1962)**  
**英国统计学家、遗传学家**



# 一、总体服从Poisson分布的MLE

Poisson分布 $X \sim P(\lambda)$ 的分布律:

$$P(X = k | \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0, 1, 2, \dots, \lambda > 0$$

$\lambda$ 为待定的未知参数。

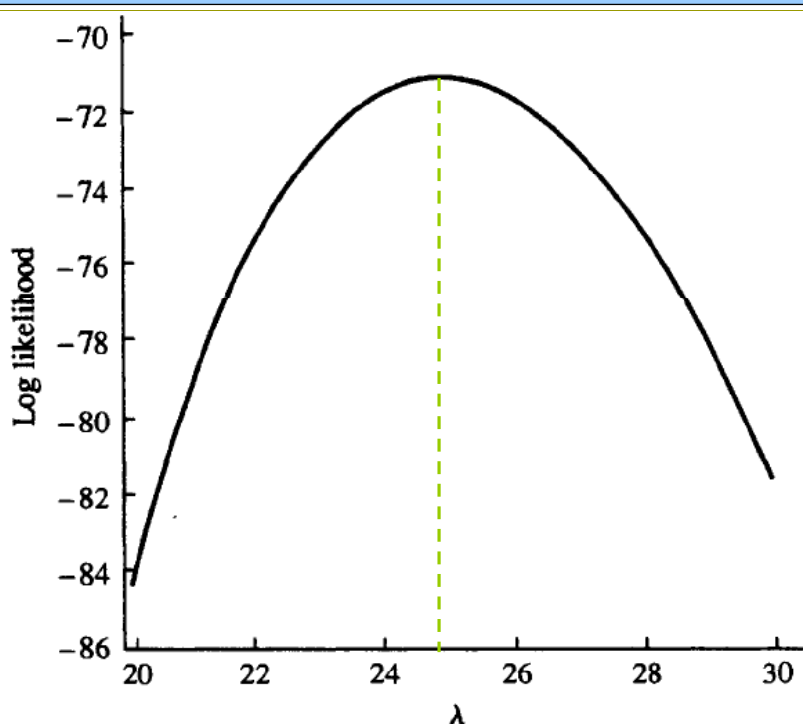
设样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布的随机变量。对数似然函数为:

$$l(\lambda) = \log \lambda \sum_{i=1}^n X_i - n\lambda - \sum_{i=1}^n (\log X_i!)$$
$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^n X_i - n = 0 \quad \longrightarrow \quad \tilde{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

抽样分布: ?

**【Example 5.4】**Steel等（1980）的分析表明测得的纤维数目能很好地服从Poisson分布。以一组数据为例:

31, 29, 19, 18, 31, 28, 34, 27, 34, 30, 16, 18, 26, 27, 27, 18, 24, 22, 28, 24, 21, 17, 24



Likelihood function of  $\lambda$  for asbestos data

## 二、总体服从正态分布的MLE


正态分布 $X \sim N(\mu, \sigma^2)$ 的概率密度函数为

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (-\infty < x < +\infty),$$

其中 $\mu, \sigma$ 为待定的未知参数且 $\sigma > 0$ 。

设样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布的随机变量。对数似然函数为：

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


$$\begin{aligned} \tilde{\mu} &= \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \\ \tilde{\sigma} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

抽样分布：？

## 三、总体服从Gamma分布的MLE

Gamma分布 $X \sim \Gamma(\alpha, \lambda)$ 的概率密度函数为

$$f(x | \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

其中 $\alpha, \lambda > 0$ 为待定的未知参数。

设样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布的随机变量。对数似然函数为：

$$l(\alpha, \lambda) = n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^n \log X_i - \lambda \sum_{i=1}^n X_i - n \log \Gamma(\alpha)$$



估计解的表达式：

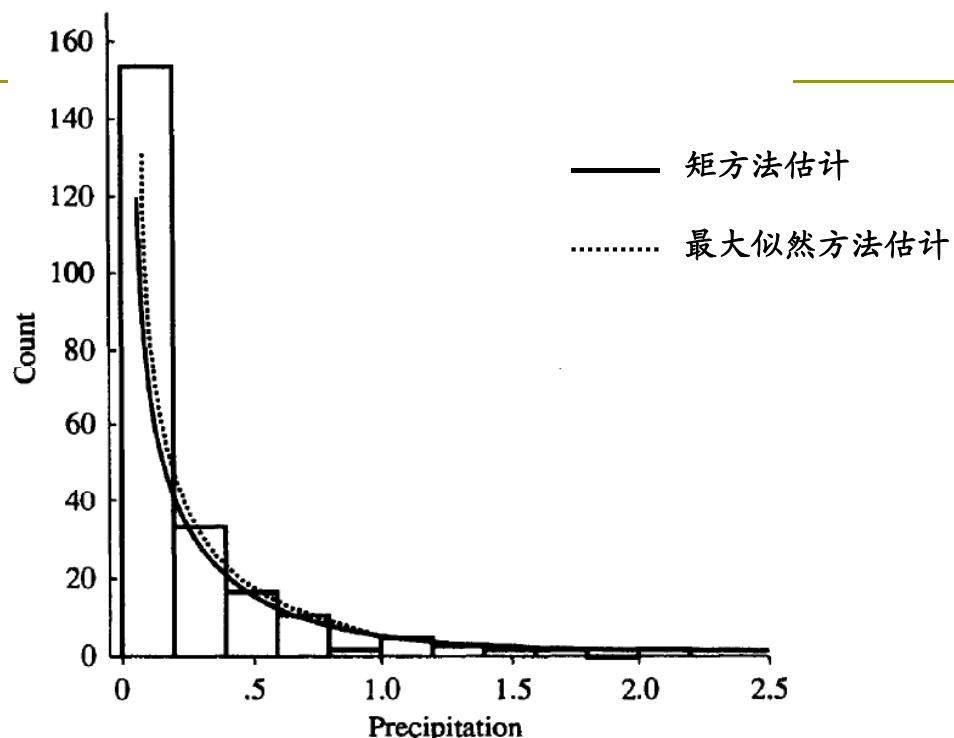
$$n \log \tilde{\alpha} - n \log \bar{X} + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\tilde{\alpha})}{\Gamma(\tilde{\alpha})} = 0$$

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

非线性方程的数值求解：迭代法

初始值？——可采用矩方法的解

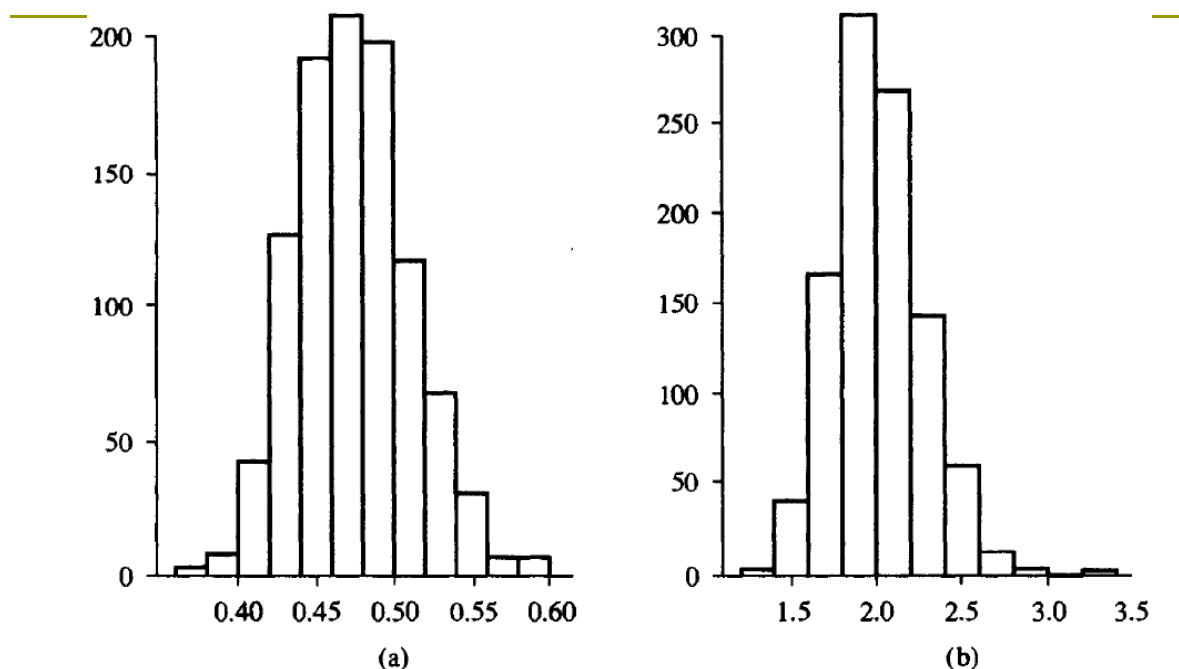
【Example 5.5】Le Cam等对美国Illinois州南部地区1960-64年期间的227场暴雨的降雨量数据进行Gamma分布的拟合。



(Le Cam L. and Neyman J. (eds.) (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley: University of California Press)

**【Example 5.5】** Le Cam等对美国Illinois州南部地区1960-64年期间的227场暴雨的降雨量数据进行Gamma分布的拟合。

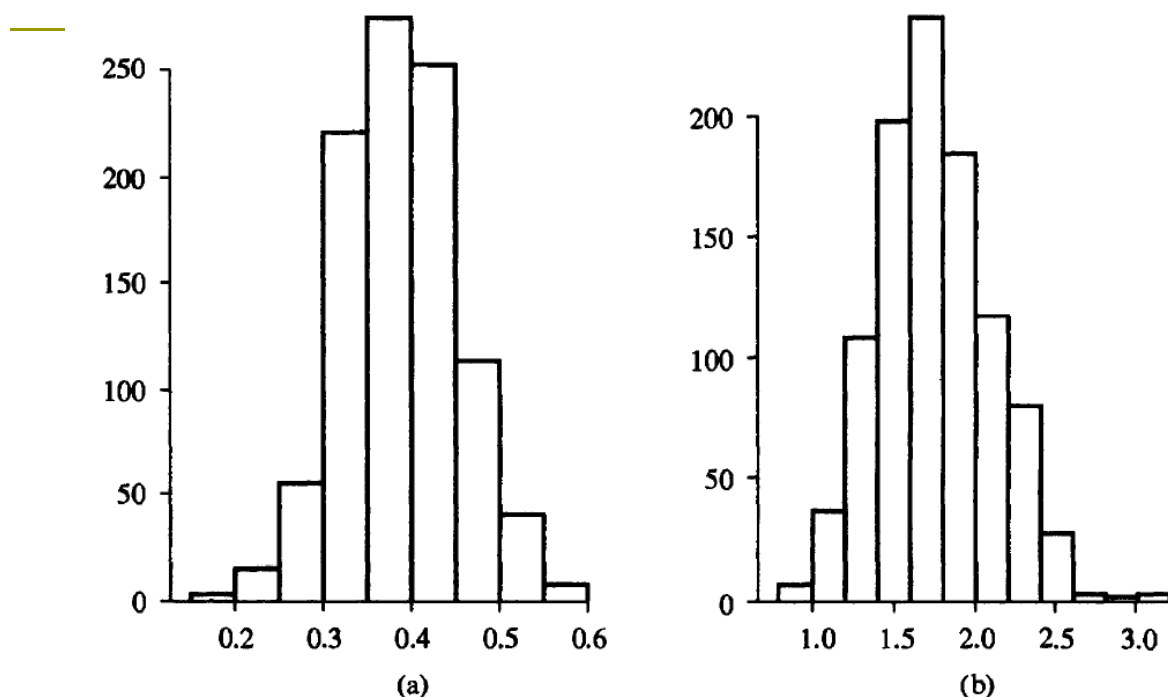
Histogram of 1000 simulated MLEs of (a)  $\alpha$  and (b)  $\lambda$ .



(Le Cam L. and Neyman J. (eds.) (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley: University of California Press)

**【Example 5.5】** Le Cam等对美国Illinois州南部地区1960-64年期间的227场暴雨的降雨量数据进行Gamma分布的拟合。

Histogram of 1000 simulated method of moment estimates of (a)  $\alpha$  and (b)  $\lambda$ .



(Le Cam L. and Neyman J. (eds.) (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume V: Weather Modification*. Berkeley: University of California Press)

## § 5.4.2 多项分布概率问题的MLE

多项分布概率 (Multinomial cell probability) 问题:

设  $X_1, X_2, \dots, X_m$  为分布在  $m$  个格子中、满足多项分布的采样计数:

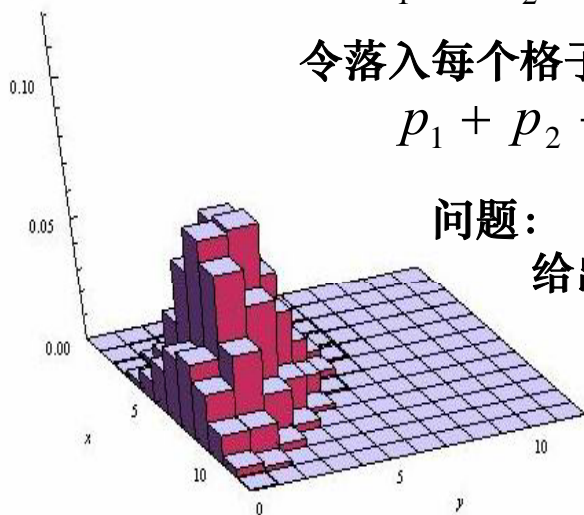
$$X_1 + X_2 + \dots + X_m = n$$

令落入每个格子的概率分别为  $p_1, p_2, \dots, p_m$ :

$$p_1 + p_2 + \dots + p_m = 1$$

问题:

给出  $p_1, p_2, \dots, p_m$  的估计值。



联合概率为:

$$f(x_1, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$

对数似然函数:

$$l(p_1, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

约束:

$$p_1 + p_2 + \dots + p_m = 1$$

↓  
 $\tilde{p}_j = ?$

联合概率为：

$$f(x_1, \dots, x_m \mid p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$

---

对数似然函数：

$$l(p_1, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

约束：

$$p_1 + p_2 + \dots + p_m = 1$$



$$\tilde{p}_j = \frac{x_j}{n}$$

抽样分布？

对于  $p_i = p_i(\theta)$  的情形，有：

---

$$l(\theta) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i(\theta)$$

**【Example 5.6】群体遗传学的Hardy-Weinberg Equilibrium (遗传平衡定律)**：主要用于描述群体中等位基因频率以及基因型 (genotypes) 频率之间的关系，即一个无穷大的群体在理想情况 (不受选择、迁移和突变影响) 下进行随机交配，经过多个世代，仍可保持基因频率与基因型频率处于稳定的平衡状态，亦即基因型AA, Aa和aa在群体中的比例应为  $(1-\theta)^2$ ,  $2\theta(1-\theta)$  和  $\theta^2$ 。

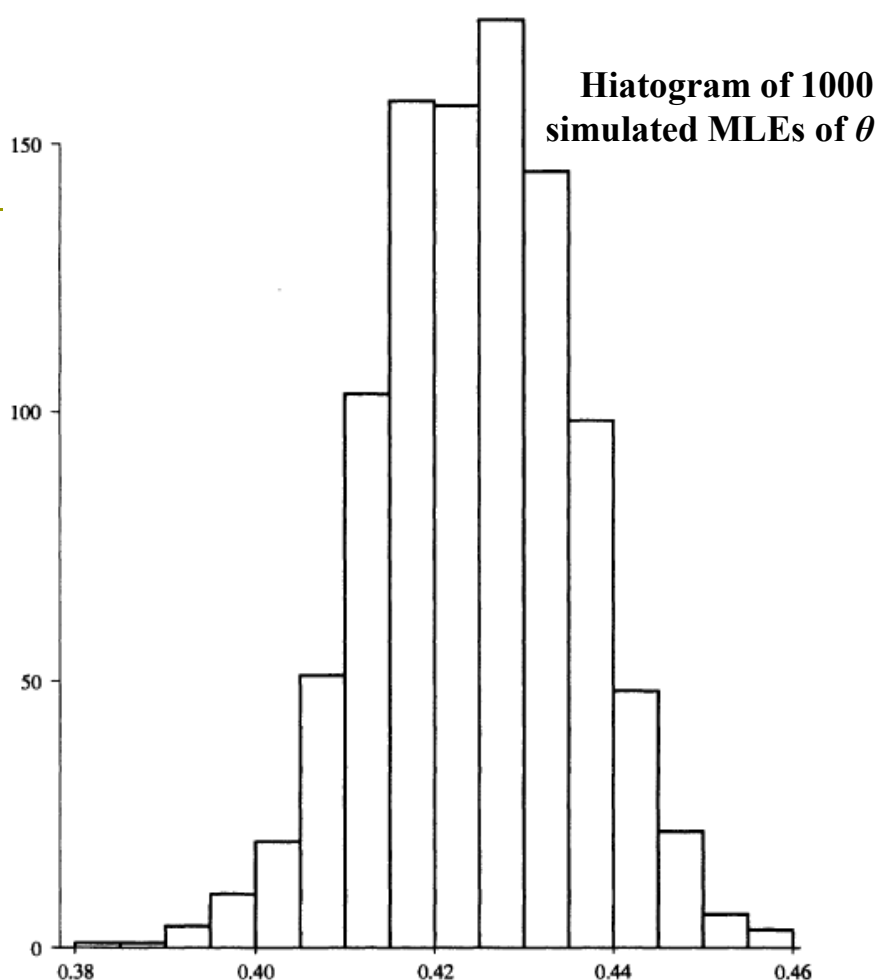
如表是1937年对香港地区中国人血型的一次抽样调查数据，M, N是红血细胞抗原 (erythrocyte antigens)。请估计N的比例 $\theta$ 。

	Blood Type			
	M	MN	N	Total
Frequency	342	500	187	1029

$$l(\theta) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i(\theta)$$

## 抽样分布

问题：如何实现  
Bootstrap模拟抽  
样？



## § 5.4.3 MLE的大样本理论

---

MLE的不变性:

如果 $\tilde{\theta}$ 是 $\theta$ 的MLE, 则在通常的情况下 $g(\tilde{\theta})$ 也是 $g(\theta)$ 的MLE。

当样本数目渐大时, MLE的抽样分布性质:

- (1) 在合适的条件下, MLE具有一致性;
- (2) 对于大容量样本, MLE的抽样分布近似服从正态分布。

问题描述:

---

已知样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布 (概率密度函数为 $f(x|\theta)$ ) 的随机变量,  $x_1, x_2, \dots, x_n$ 为其观察值,  $\theta$ 的对数似然函数为:

$$l(\theta) = \sum_{i=1}^n \log(f(X_i | \theta))$$

记 $\theta$ 的真实值为 $\theta_0$ 。



## 定理

---

在概率密度函数  $f(X|\theta)$  保证充分光滑的条件下，独立同分布样本的MLE为一致性估计。

## 引理（定义）

---

**Fisher信息量  $I(\theta)$ :**

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

在概率密度函数  $f(X|\theta)$  保证充分光滑的条件下， $I(\theta)$ 可表示成：

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

## 定理

(1) 在概率密度函数  $f(X|\theta)$  保证充分光滑的条件下，统计量

$$\sqrt{nI(\theta_0)}(\tilde{\theta} - \theta_0)$$

依概率服从标准正态分布，其中  $\theta_0$  为  $\theta$  的真实值。

(2) 当样本容量趋向无穷大时，MLE 的抽样分布趋向均值为  $\theta_0$ 、方差为  $\frac{1}{nI(\theta_0)}$  的正态分布。

——MLE 是 **渐近无偏估计** (asymptotically unbiased estimation)，其方差为 **渐近方差** (asymptotically variance)。

## § 5.5 参数置信区间(CI)的估计

### 目的

——基于真实值  $\theta$  的估计值  $\tilde{\theta}$  的抽样分布来讨论  $\theta$  的置信区间 (confidence interval)，置信区间从概率上表明了估计值  $\tilde{\theta}$  的与真实值  $\theta$  的关系，是一种不确定性的程度，亦即真实值  $\theta$  落在该区间的可能性大小

### 确定置信区间的方法

- (1) 解析式推导方法；
- (2) 根据MLE渐近性质的推导方法；
- (3) Bootstrap模拟方法。

# 一、解析式推导方法

正态分布 $X \sim N(\mu, \sigma^2)$ 的概率密度函数为

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (-\infty < x < +\infty),$$

其中 $\mu, \sigma$ 为待定的未知参数且 $\sigma > 0$ 。

设样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布的随机变量。参数的MLEs为：

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

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

## 1. $\mu$ 的CI估计：

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

### § 3.2.2 正态总体样本均值与样本方差的抽样分布

**定理** 设总体 $X$ 服从正态分布 $N(\mu, \sigma^2)$ ，其随机样本 $(X_1, X_2, \dots, X_n)$ 为有放回、相互独立的个体集合，即每个随机变量 $X_i$  ( $i=1, \dots, n$ ) 也符合正态分布。则有：

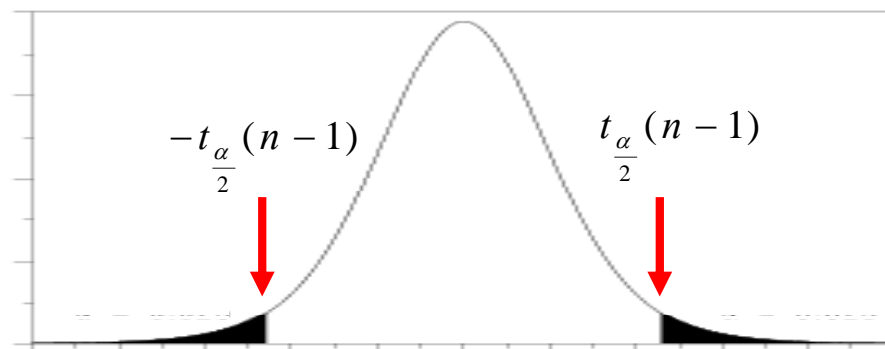
$$T = \frac{(\bar{X} - \mu) \sqrt{n}}{S} \sim t(n-1) \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{\frac{\alpha}{2}}(n-1)\right) = 1 - \alpha$$

$$P\left(\left[\bar{X} - \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right] \leq \mu \leq \left[\bar{X} + \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right]\right) = 1 - \alpha$$

$\mu$ 的 $100(1-\alpha)\%$ 的CI为:

$$\left[\bar{X} - \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1), \quad \bar{X} + \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right]$$



## CI的讨论:

- 1 置信水平 (confidence level)、可信范围
- 2 包含参数真实值的可信程度
- 3 构造CI的方法: 参数估计、假设检验

## 参数估计的任务

——点估计 (point estimate) : 求解估计量 $\tilde{\theta}$

——区间估计 (interval estimate) : 给出置信区间 $(\tilde{\theta}_1, \tilde{\theta}_2)$


## 2. $\sigma^2$ 的CI估计:

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

### § 3.2.2 正态总体样本均值与样本方差的抽样分布

**定理** 设总体 $X$ 服从正态分布 $N(\mu, \sigma^2)$ , 其随机样本 $(X_1, X_2, \dots, X_n)$ 为有放回、相互独立的个体集合, 即每个随机变量 $X_i$  ( $i=1, \dots, n$ ) 也符合标准正态分布。则有:

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1), \quad \text{where} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

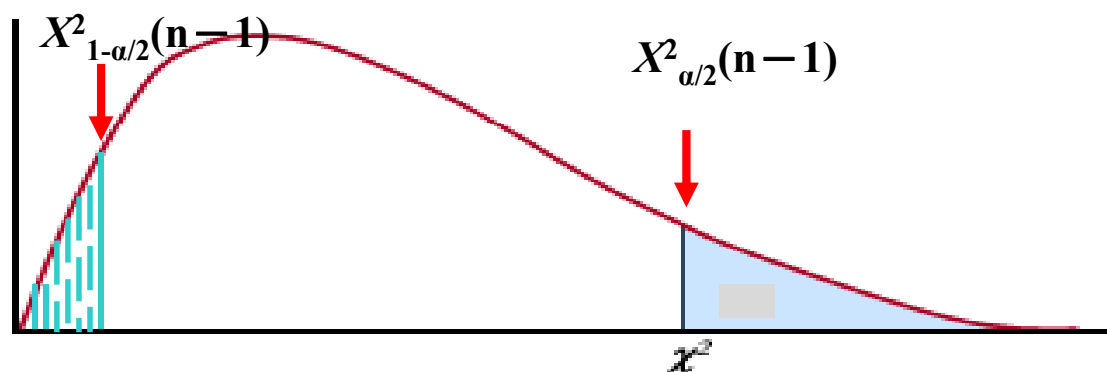

$$\frac{n}{\sigma^2} \tilde{\sigma}^2 \sim \chi^2(n-1), \quad \text{where} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

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

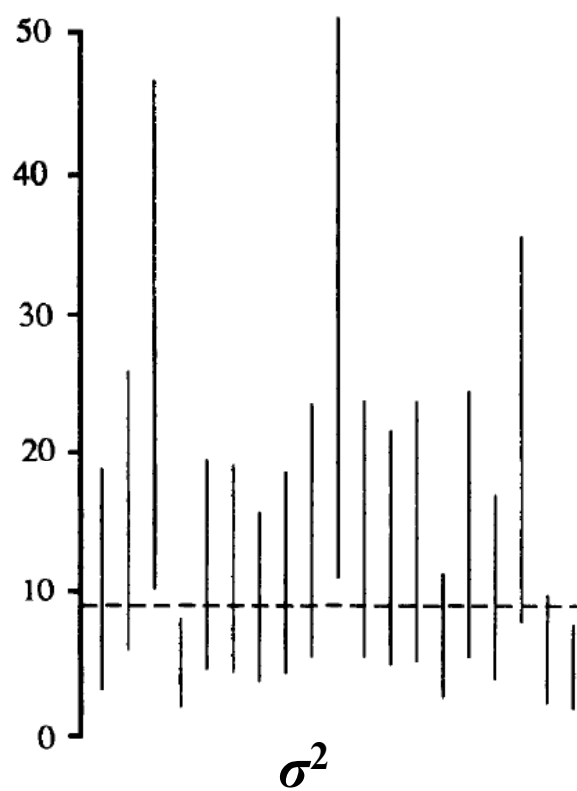
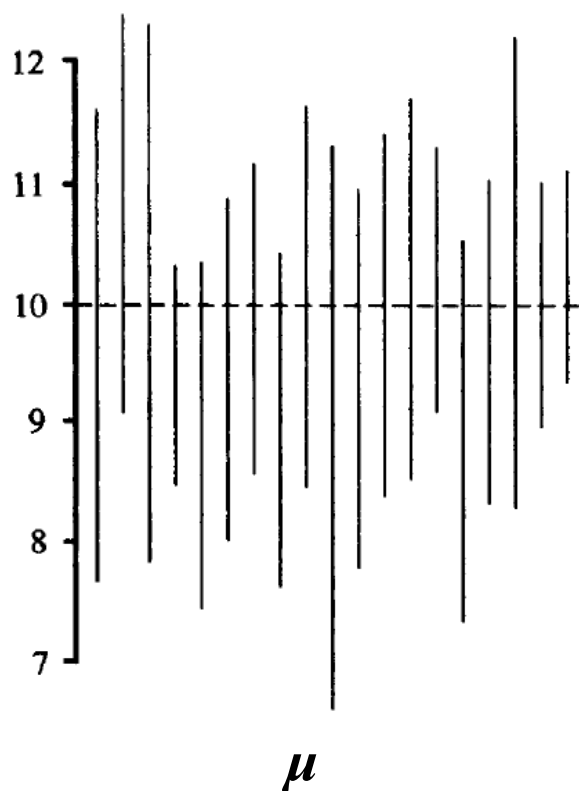
$$P\left(\frac{n\tilde{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{n\tilde{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) = 1 - \alpha$$

$\sigma^2$ 的 $100(1-\alpha)\%$ 的CI为:

$$\left[ \frac{n\tilde{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{n\tilde{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \right]$$



已知 $\mu=10, \sigma^2=9, n=11$ , 进行20次90%置信区间的计算



## 二、根据MLE渐近性质的推导方法

**定理** 在概率密度函数  $f(X|\theta)$  保证充分光滑的条件下,

$$\sqrt{nI(\theta_0)}(\tilde{\theta} - \theta_0)$$

依概率服从标准正态分布。

由于 $\theta_0$ 未知,不妨以 $I(\tilde{\theta})$ 代替 $I(\theta_0)$ ,故有:

$$P\left(-z_{\frac{\alpha}{2}} \leq \sqrt{nI(\tilde{\theta})}(\tilde{\theta} - \theta_0) \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

于是 $\theta$ 的 $100(1-\alpha)\%$ 置信区间为:

$$\left[ \tilde{\theta} - \frac{1}{\sqrt{nI(\tilde{\theta})}} z_{\frac{\alpha}{2}}, \quad \tilde{\theta} + \frac{1}{\sqrt{nI(\tilde{\theta})}} z_{\frac{\alpha}{2}} \right]$$

**Poisson分布** $X \sim P(\lambda)$ 的分布律:

$$P(X = k | \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0, 1, 2, \dots, \lambda > 0$$

$\lambda$ 为待定的未知参数。设样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布的随机变量。

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

$$I(\lambda) = -E\left[\frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda)\right] = \frac{1}{\lambda}$$

故 $\lambda$ 的 $100(1-\alpha)\%$ 置信区间为:

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



### 三、Bootstrap模拟方法


已知观察得到的样本数据 $X_1, X_2, \dots, X_n$ 为独立同分布的一组随机变量，未知参数 $\theta$ 的真实值为 $\theta_0$ （未知），它的估计量为 $\tilde{\theta}$ 。

若已知 $\theta$ 的真实值 $\theta_0$ 和 $\varepsilon = (\tilde{\theta} - \theta_0)$ 的分布，则可以定义两个临界点：

$$P\left((\tilde{\theta} - \theta_0) \leq \delta_{lower}\right) = \alpha/2, \quad P\left((\tilde{\theta} - \theta_0) \leq \delta_{upper}\right) = 1 - \alpha/2$$

$$P\left(\delta_{lower} \leq (\tilde{\theta} - \theta_0) \leq \delta_{upper}\right) = 1 - \alpha$$

$$P\left((\tilde{\theta} - \delta_{upper}) \leq \theta_0 \leq (\tilde{\theta} - \delta_{lower})\right) = 1 - \alpha$$


$$\left((\tilde{\theta} - \delta_{upper}), (\tilde{\theta} - \delta_{lower})\right)$$

但是， $\theta_0$ 和 $\varepsilon = (\tilde{\theta} - \theta_0)$ 的分布都未知

Bootstrap方法的解决方案：

用bootstrap的大量估计值 $\tilde{\theta}^{(b)}$ 来模拟真实值 $\theta_0$ 和 $\varepsilon = (\tilde{\theta} - \theta_0)$ 的分布，并推导出近似的置信区间。

根据bootstrap样本，可得到参数 $\theta$ 的估计值 $\tilde{\theta}^{(b)}$ ，这样重复若干次，记为 $B$ 。设 $B=1000$ ，就得到该参数的1000个估计值。

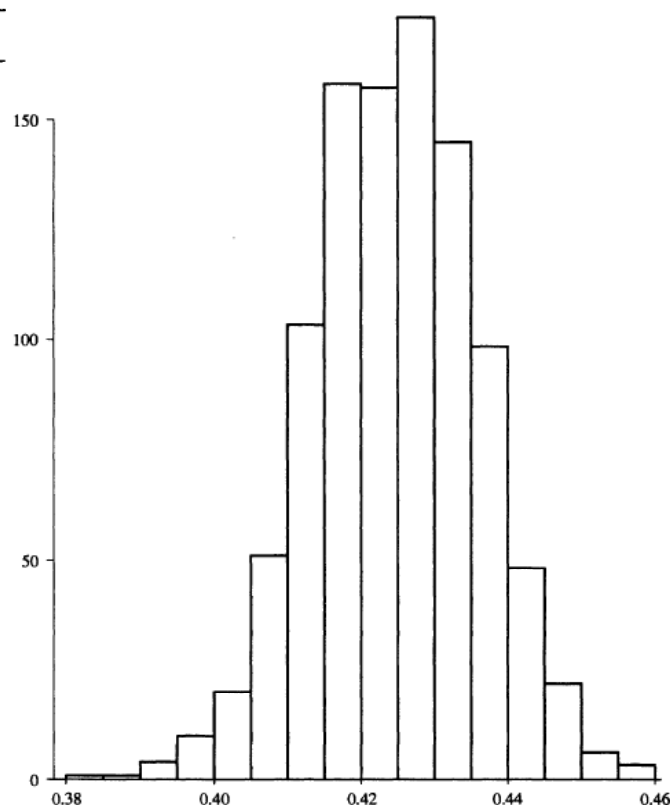
——当 $\tilde{\theta}^{(b)}$ 的频数分布近似正态分布时，以其均数作为 $\theta_0$ 的估计值，用正态原理估计置信区间；

——当 $\tilde{\theta}^{(b)}$ 的频数分布为偏态时，以其中位数作为 $\theta_0$ 的估计值，以上、下分位数作为其置信区间。

**【Example 5.6】Hardy-Weinberg Equilibrium (遗传平衡定律)：**  
如表是1937年对香港地区中国人血型的一次抽样调查数据，M, N是红细胞抗原 (erythrocyte antigens)。请估计N的比例 $\theta$ 的置信区间。

	Blood Type			Total
	M	MN	N	
Frequency	342	500	187	1029

Histogram of 1000  
simulated MLEs of  $\theta$



## § 5.6 参数估计的有效性和 Cramer-Rao下界定理

**“最优估计”原则：**

估计量 $\tilde{\theta}$ 最接近 $\theta$ 且分布最集中。

**定义 (有效性)**

设 $\tilde{\theta}_1(X_1, X_2, \dots, X_n)$ 、 $\tilde{\theta}_2(X_1, X_2, \dots, X_n)$  是参数 $\theta$ 的两个无偏估计量，若对任意的 $\theta \in \Theta$ ，都有：

$$D(\tilde{\theta}_1) \leq D(\tilde{\theta}_2)$$

则称估计量  $\tilde{\theta}_1$  比  $\tilde{\theta}_2$  更有效。

【Example 5.7】 设总体 $X \sim N(\mu, \sigma^2)$ ,  $X_1, X_2, \dots, X_n$ 为来自总体的样本, 证明:

(1)  $S_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$  是 $\sigma^2$ 的无偏估计;

(2) 令  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , 则估计量 $S_1^2$ 较 $S^2$ 有效。

**定义** (一致最小方差无偏估计, UMVUE, Uniformly minimum variance unbiased estimation)

设  $\tilde{\theta}(X_1, X_2, \dots, X_n)$  是参数 $\theta$ 的估计量, 若有:

(1) 对任意的 $\theta \in \Theta$ ,  $\tilde{\theta}$  是 $\theta$ 的无偏估计;

(2) 对 $\theta$ 的任一无偏估计  $\tilde{\theta}'$ , 都有  $D(\tilde{\theta}) \leq D(\tilde{\theta}')$ 。

则称  $\tilde{\theta}$  是 $\theta$ 的一致最小方差无偏估计。

## 定理 (C-R下界定理, *Cramer-Rao lower bound theorem*)

设总体X的密度函数为  $\{f(x|\theta), \theta \in \Theta\}$ ,  $X_1, X_2, \dots, X_n$  为来自总体X的一个样本, 其联合概率密度函数为

$$L(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

$T = T(X_1, X_2, \dots, X_n)$  是  $g(\theta)$  的无偏估计量。若下列条件成立:

(1)  $G = \{x: f(x|\theta) > 0\}$  与  $\theta$  无关;



(2)  $g'(\theta)$  和  $\frac{\partial f(x|\theta)}{\partial \theta}$  存在, 且对任一  $\theta \in \Theta$  都有:

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{+\infty} f(x|\theta) dx = \int_{-\infty}^{+\infty} \frac{\partial}{\partial \theta} f(x|\theta) dx,$$

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} T \cdot L dx_1 dx_2 \dots dx_n = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} T \cdot \frac{\partial L}{\partial \theta} dx_1 dx_2 \dots dx_n$$

(3)  $I(\theta) > 0$ ;

则对任一  $\theta \in \Theta$  都有

$$D(T) \geq \frac{[g'(\theta)]^2}{n I(\theta)}$$

并称  $\frac{[g'(\theta)]^2}{n I(\theta)}$  为  $g(\theta)$  的无偏估计T的 **C-R下界** (**Cramer-Rao lower bound**)。

Calyampudi Radhakrishna Rao

(1920 - )

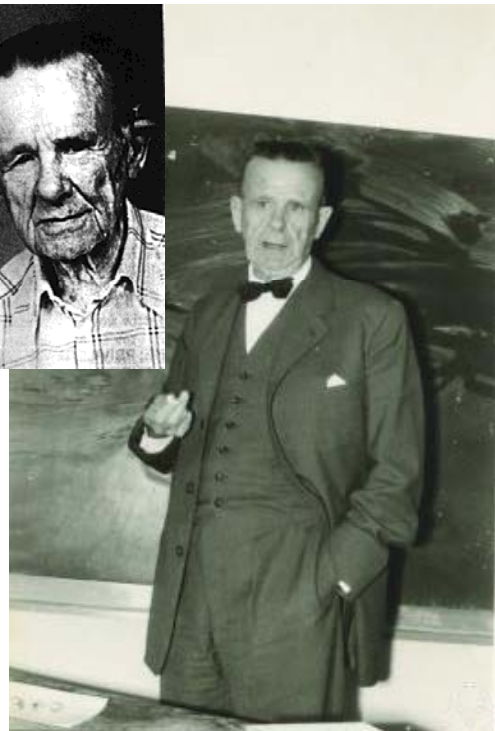
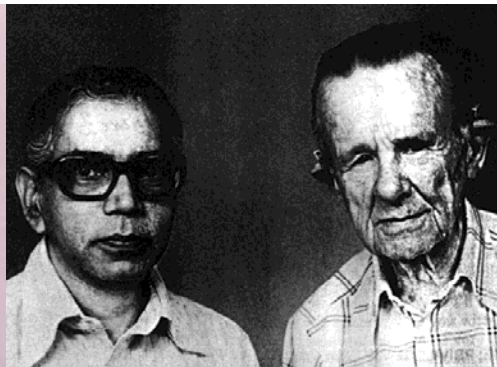
American Indian born statistician



Harald Cramér

(1893 – 1985)

Swedish mathematician



## C-R下界定理的讨论:

(1) 当 $g(\theta) = \theta$ 时, 结论简化成  $D(T) \geq \frac{1}{nI(\theta)}$ 。

(2) C-R下界定理给出了参数估计方法中无偏估计的下界。

(3) C-R下界定理的条件1、2通常称为**正则条件**。大多数分布都满足正则条件。

(4) 对于 $\theta$ 的无偏估计  $\tilde{\theta}$ , 若满足:

$$D(\tilde{\theta}) \geq \frac{1}{nI(\theta)}$$

则称  $\tilde{\theta}$  为 $\theta$ 的**有效估计**。

(5) 对于 $\theta$ 的无偏估计  $\tilde{\theta}_n$ ，若满足：

$$\lim_{n \rightarrow \infty} \frac{1}{n I(\theta)} \frac{1}{D(\tilde{\theta}_n)} = 1$$

则称  $\tilde{\theta}_n$  为  $\theta$  的渐近有效估计。

(6) 满足C-R下界定理条件时，有效估计一定是一致最小方差无偏估计（UMVUE），而一致最小方差无偏估计不一定是有效估计。

**【Example 5.8】** 设总体  $X \sim P(\lambda)$ ， $X_1, X_2, \dots, X_n$  为来自总体的样本，试求  $\lambda$  的无偏估计的C-R下界。

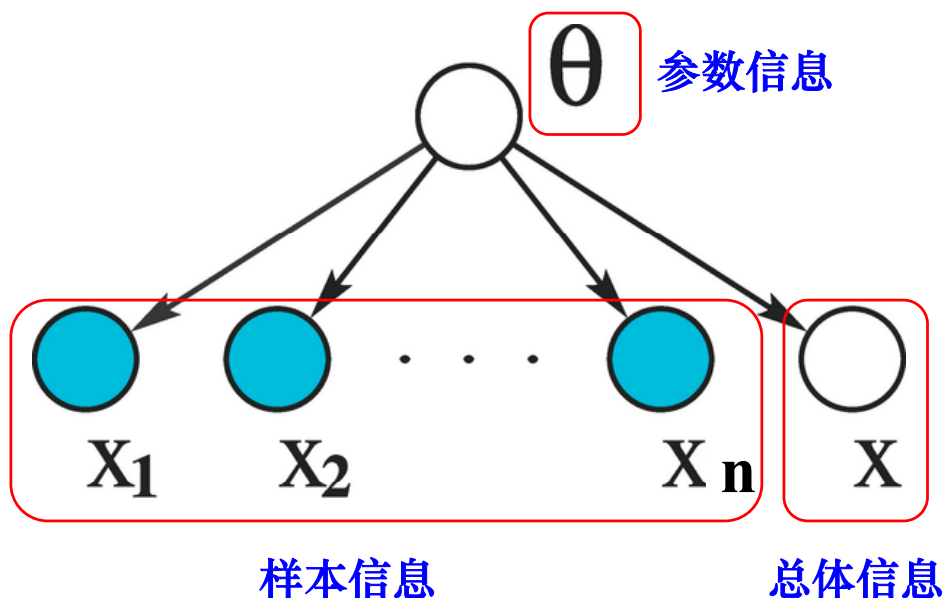
## § 5.7 基于Bayes推断的估计方法

“Years ago a statistician might have claimed that statistics deals with the processing of data... today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty.”

—— Chernoff, H. & Moses, L.E.(1959).

Elementary Decision Theory. New York: Wiley.

## Classic Inference and Bayesian Inference





## § 5.7.1 先验分布与后验分布

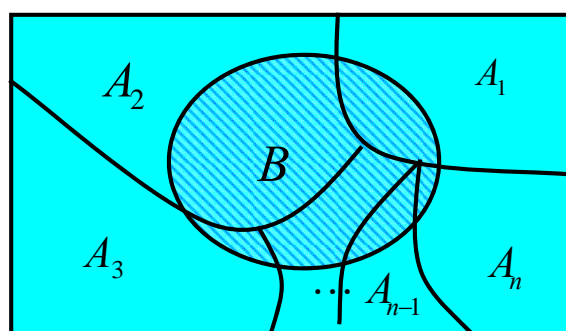
**Bayes概率公式** 设试验E的样本空间为 $\Omega$ ，事件 $A_1, A_2, \dots, A_n$ 构成样本空间 $\Omega$ 的完备事件组，且 $P(A_i) > 0$  ( $i=1, 2, \dots, n$ )，则对任一事件B， $P(B) > 0$ ，有

$$P(A_i | B) = \frac{P(A_i) P(B | A_i)}{\sum_{j=1}^n P(A_j) P(B | A_j)}$$

$P(A_i)$ : 完备事件组 $\{A_i\}$ 的先验概率 (prior probability)

$P(B)$ : 事件B的先验概率 (prior probability)

$P(A_i|B)$ : 后验概率 (posterior probability)



## 定义

假设定义在参数空间 $\Theta$ 上的待估参数 $\theta$ 为随机变量（或随机向量），则定义在 $\Theta$ 上的 $\theta$ 的概率分布 $\pi(\theta)$ 为 $\theta$ 的先验分布 (prior distribution)。

在给定样本 $X_1, X_2, \dots, X_n$ 后，则 $\theta$ 的条件分布 $f(\theta | (X_1, X_2, \dots, X_n))$ 为 $\theta$ 的后验分布 (posterior distribution)。

设总体 $X$ 分布的密度函数为 $f(x|\theta)$ ,  $\theta \in \Theta$ ,  $\theta$ 的先验分布为 $\pi(\theta)$ , 故 $X, \theta$ 的联合概率密度为  $\pi(\theta)f(x|\theta)$ 。

设 $X_1, X_2, \dots, X_n$ 为取自总体 $X$ 的一个样本, 则 $X_1, X_2, \dots, X_n, \theta$ 的联合概率密度为

$$\prod_{i=1}^n f(X_i | \theta) \pi(\theta)$$

由概率论知, 样本 $X_1, X_2, \dots, X_n$ 的边沿概率密度为:

$$f_X(X_1, X_2, \dots, X_n) = \int_{\Theta} \prod_{i=1}^n f(X_i | \theta) \pi(\theta) d\theta$$

在给定样本 $X_1, X_2, \dots, X_n$ 后的 $\theta$ 的条件概率密度为:

$$h(\theta | X_1, X_2, \dots, X_n) = \frac{\prod_{i=1}^n f(X_i | \theta) \pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i | \theta) \pi(\theta) d\theta}, \quad \theta \in \Theta$$

此即给定样本条件 $X_1, X_2, \dots, X_n$ 下 $\theta$ 的后验分布。

## Bayes统计学:

根据先验信息对未知参数 $\theta$ 获得其先验分布 $\pi(\theta)$ ；通过随机试验获得样本 $X_1, X_2, \dots, X_n$ ；然后通过

$$h(\theta | X_1, X_2, \dots, X_n) = \frac{\prod_{i=1}^n f(X_i | \theta) \pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i | \theta) \pi(\theta) d\theta}, \quad \theta \in \Theta$$

对 $\theta$ 的先验分布进行调整，调整的结果就是 $\theta$ 的后验分布，使人们的认识由先验分布 $\pi(\theta)$ 调整到 $h(\theta | X_1, X_2, \dots, X_n)$ 。

参数 $\theta$ 的后验分布包含了**总体信息**、**样本信息**和**先验信息**，对参数 $\theta$ 的统计推断是建立在后验分布的基础上。

**【Example 5.9】** 设总体 $X \sim N(a, \sigma^2)$ ， $a$ 未知而 $\sigma$ 已知。给定 $a$ 的先验分布为 $a \sim N(\mu, \tau^2)$ ，已知 $X_1, X_2, \dots, X_n$ 是取自总体 $X$ 的样本，求 $a$ 的后验分布。

$$h(a | X_1, X_2, \dots, X_n) \propto e^{-\frac{1}{2\eta^2}(a - t)^2}$$

$$t = \frac{\frac{n \overline{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

$$\eta^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

## § 5.7.2 Bayes估计方法简介

在给定样本 $X_1, X_2, \dots, X_n$ 后的 $\theta$ 的后验分布:

$$h(\theta | X_1, X_2, \dots, X_n) = \frac{\prod_{i=1}^n f(X_i | \theta) \pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i | \theta) \pi(\theta) d\theta}, \quad \theta \in \Theta$$

目标:

从 $h(\theta | X_1, X_2, \dots, X_n)$ 出发来确定参数 $\theta$ 的估计值。

### 一、最大后验(Maximu a posteriori, MAP) 估计

#### 基本思想:

取使后验概率密度函数（或后验概率函数） $h(\theta | X_1, X_2, \dots, X_n)$ 达到最大的参数值作为待估计参数的估计值，记为 $\tilde{\theta}_M$ 。

与最大似然估计方法有密切关系

规则化（Regularization）的最大似然估计

【Example 5.9】设总体 $X \sim N(a, \sigma^2)$ ， $a$ 未知而 $\sigma$ 已知。给定 $a$ 的先验分布为 $a \sim N(\mu, \tau^2)$ ，已知 $X_1, X_2, \dots, X_n$ 是取自总体 $X$ 的样本，求 $a$ 的最大后验估计值。

$$h(a | X_1, X_2, \dots, X_n) = C e^{-\frac{1}{2\eta^2}(a-t)^2}$$

$$t = \frac{\frac{n \bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \eta^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

$$\tilde{a}_M = t = \frac{\frac{n \bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

## 二、期望型估计

### 基本思想：

取后验分布的期望值

$$E \left[ h(a | X_1, X_2, \dots, X_n) \right]$$

作为待估计参数的估计值，记为  $\tilde{\theta}_E$ 。

直观的估计方法，即利用后验分布的概率分布期望特征来估计参数 $\theta$ 。

【Example 5.9】设总体 $X \sim N(a, \sigma^2)$ ， $a$ 未知而 $\sigma$ 已知。给定 $a$ 的先验分布为 $a \sim N(\mu, \tau^2)$ ，已知 $X_1, X_2, \dots, X_n$ 是取自总体 $X$ 的样本，求 $a$ 的期望型估计值。

$$h(a | X_1, X_2, \dots, X_n) = C e^{-\frac{1}{2\eta^2}(a-t)^2}$$

$$t = \frac{\frac{n \bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \eta^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

$$\tilde{a}_E = t = \frac{\frac{n \bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

### 三、最小Bayesian风险估计

#### 基本思想：

基于统计决策论（statistical decision）的思想，提出用以度量估计优劣的准则，并在可能的情况下寻找最优估计。主要是统计决策理论框架内的**损失**（lost）和**风险**（risk）概念的应用。

## Statistical Decision Theory



□ Abraham Wald (1902 – 1950)

- Wald's test
- Rigorous proof of the consistency of MLE

“Note on the consistency of the maximum likelihood estimate”, *Ann. Math. Statist.*, 20, 595-601.

## 定义

设  $\tilde{\theta} = \tilde{\theta}(X_1, X_2, \dots, X_n)$  为用样本  $X_1, X_2, \dots, X_n$  对总体  $X$  的待估计参数  $\theta$  的一个估计，非负二元实值函数  $L(\theta, \tilde{\theta})$  表示估计所带来的损失，称为**损失函数 (lost function)**，并称  $L(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$  为平方差损失函数。显然， $L(\theta, \tilde{\theta})$  也是随机变量，其期望值表示总体损失，称  $R_{\tilde{\theta}} = E(L(\theta, \tilde{\theta}))$  为**风险函数 (risk fuction)**。

讨论：

- (1) 一般而言，用  $\tilde{\theta}$  估计  $\theta$  的效果越差，损失函数值越大；
- (2) 风险的概念就是“平均损失”，其平均计算既对样本又对参数进行。风险函数值是用于评价估计优劣的准则，风险越大，估计越差。

## 定义

设  $\tilde{\theta} = \tilde{\theta}(X_1, X_2, \dots, X_n)$  为用样本  $X_1, X_2, \dots, X_n$  对总体  $X$  的待估计参数  $\theta$  的所有可能的估计， $L(\theta, \tilde{\theta})$  为损失函数， $R_{\tilde{\theta}} = E(L(\theta, \tilde{\theta}))$  为风险。若存在一个关于参数  $\theta$  的估计  $\tilde{\theta}^*$ ，对任意一个估计  $\tilde{\theta}$ ，满足：

$$R_{\tilde{\theta}^*} \leq R_{\tilde{\theta}}$$

则称  $\tilde{\theta}^*$  为参数  $\theta$  的**最小Bayes风险估计**。



## Bayes风险估计三要素：

$\theta$ 的先验分布为 $\pi(\theta)$

总体 $X$ 分布的密度函数为 $f(x|\theta)$

损失函数  $L(\theta, \tilde{\theta})$

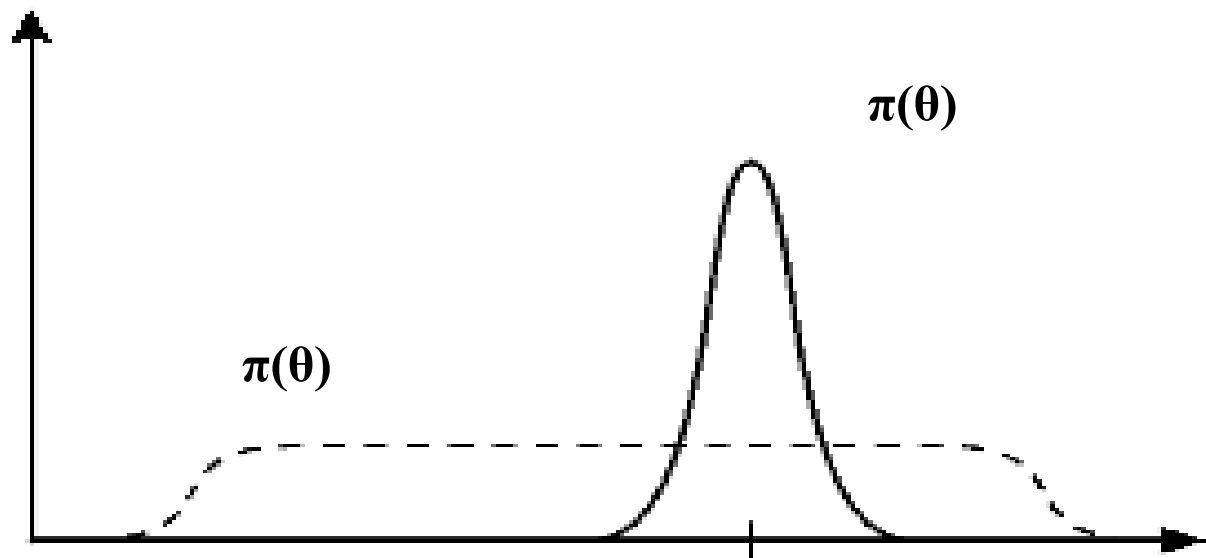
【Example 5.9】设总体 $X \sim N(a, \sigma^2)$ ， $a$ 未知而 $\sigma$ 已知。给定 $a$ 的先验分布为 $a \sim N(\mu, \tau^2)$ ，已知 $X_1, X_2, \dots, X_n$ 是取自总体 $X$ 的样本，求 $a$ 的最小Bayes风险估计值。

$$h(a | X_1, X_2, \dots, X_n) = C e^{-\frac{1}{2\eta^2}(a-t)^2}$$

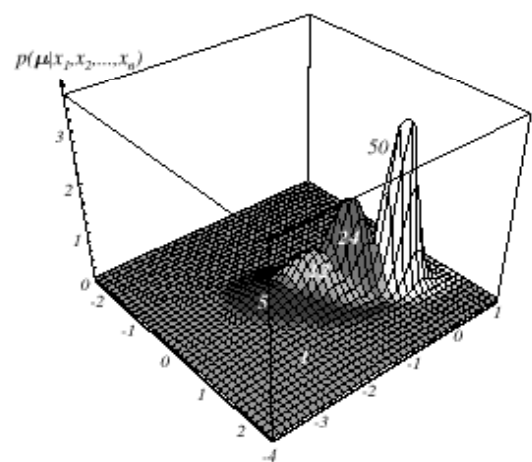
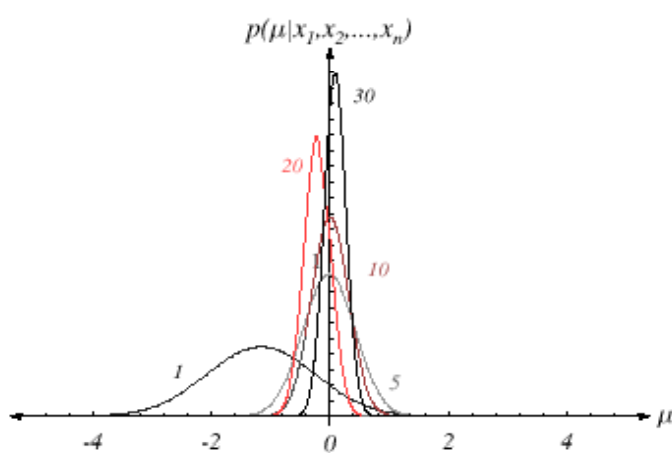
$$t = \frac{\frac{n \overline{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \eta^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

$$\tilde{a}_R = t = \frac{\frac{n \overline{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

## The prior distribution in the Bayesian inference



$$\tilde{a} = \frac{\frac{n \bar{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$



**FIGURE 3.2.** Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

## § 5.7.3 关于先验分布的简单讨论

---

讨论：

- (1) **客观法** 如果对参数本身的随机性有一定的认识，则可事先提出较准确的先验分布形式
- (2) **主观法** 根据个人对参数的经验和认识，结合已有的经验知识和理论知识，提出参数的先验分布
- (3) **同等未知原则** (Bayes假设) 假定参数在取值范围内均匀分布