概率论与数理统计

Probability and Statistics

Dait

目　录

前　言

在唐宏岩老师的第一节概率论课程上，感觉这个老师比较凶险.

课程内容

1. 初等概率论 (8 次课)
2. 统计推断初步 (7 次课)

特点　内容多、作业多、退课的同学多.

参考书

- 陈希孺：概率论与数理统计
- D. Bertsekas & J. Tsitsiklis 概率导论
- J. Rice 数理统计与数据分析
- S. Ross 概率论基础教程
- G. Casella & R. Berger 统计推断

第一部分

统计概率论

第一章　概率

概率的发展史　概率的发展经历了如下几个过程：

- de Méré 问题
- Pascal, Fermat 首创概率的数学理论 (初等数学方法)
- Laplace 创立分析概率论 (微积分分析方法)
- Kolmogorov 发展现代理论 (测度论方法)

1.1　试验与事件

定义 1.1.1: 试验与样本空间

概率论研究试验 (experiment)，试验具有如下性质：

1. 不能预先确知结果；
2. 试验之前可预知所有可能结果，其集合构成样本空间 (sample space) Ω.

定义 1.1.2: 事件

事件 (event) 是样本空间的子集 (well-defined select). 试验的单一结果称为基本事件. 特别地，Ω 是必然事件，$\varnothing$ 是不可能事件.

事件的运算　借助集合语言以及 Venn 图

- 余 $A^{\mathrm{c}} = \Omega \backslash A$，对立 $A = B^{\mathrm{c}}$
- 和与差 $A + B, A - B$
- 积 $AB = A \cap B$，互斥 $AB = \varnothing$
- de Morgan 定律[I]

$$\left(\sum A_i\right)^{\mathrm{c}} = \prod A_i^{\mathrm{c}},$$

概率的解释　古典解释，等可能性，Bertrand 悖论；频率解释，主观解释.

1.2　公理化定义

定义样本空间 Ω 的幂集 2^{Ω} 为 Ω 所有子集构成的集合，则

[I]此处省略了上下标 $i = 1$ 和 ∞，后同.

定义 1.2.1: 事件集类

事件集类 $\mathcal{F} \subset 2^{\Omega}$ 是 Ω 的 σ-代数，满足事件运算的封闭性：

1. $\Omega \in \mathcal{F}$;
2. $\forall A \in \mathcal{F}, A^{\mathrm{c}} \in \mathcal{F}$;
3. 特别地，若 $\forall A_i \in \mathcal{F}$，则 $\sum A_i \in \mathcal{F}$.

例 1.2.1: σ-代数

若 $\Omega = \{a, b, c, d\}$，则一个平凡的 σ-代数

$$\mathcal{F}_1 = \{\varnothing, \Omega\},$$

另一个 σ-代数可以为

$$\mathcal{F}_2 = \{\varnothing, \{a\}, \{b, c, d\}, \Omega\}.$$

定义 1.2.2: 概率

定义概率 (probability) $\mathrm{P} : \mathcal{F} \to \mathbb{R}$，满足以下公理：

1. $\forall A \in \mathcal{F}, \mathrm{P}(A) \geqslant 0$;
2. $\mathrm{P}(\Omega) = 1$;
3. (**加法公理**) $\forall A_i \in \mathcal{F}, A_i A_j = \varnothing (i \neq j)$,

$$\mathrm{P}\left(\sum A_i\right) = \sum \mathrm{P}(A_i). \tag{1.1}$$

$(\Omega, \mathcal{F}, \mathrm{P})$ 构成了概率空间 (probability space).

由定义，可推出以下命题：

1. $\forall A \in \mathcal{F}, \mathrm{P}(A) \leqslant 1$;
2. $\mathrm{P}(\varnothing) = 0$;
3. $\forall A_i \in \mathcal{F}, A_i A_j = \varnothing (i \neq j)$,

$$\mathrm{P}\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i)$$

特别地，$\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = 1$;

4. $\forall A \subset B, \mathrm{P}(A) \leqslant \mathrm{P}(B)$;
5. $\mathrm{P}(A + B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB)$. 推广之，得到

定理 1.2.1: 容斥恒等式

容斥恒等式 (inclusion and exclusion identities)

$$\begin{aligned}\mathrm{P}\left(\sum_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n}\mathrm{P}(A_i) - \sum_{i<j}^{n}\mathrm{P}(A_iA_j) + \cdots + \mathrm{P}\left(\prod_{i=1}^{n} A_i\right) \\ &= \sum_{r=1}^{n}(-)^{r+1}\sum_{i_1<\cdots<i_r}\mathrm{P}\left(\prod_{k=1}^{r} A_{i_k}\right). \end{aligned} \tag{1.2}$$

例 1.2.2: 乱序

n 个人每人一个帽子，离开时随机取帽子，问无人拿到自己帽子的概率为多少？

记 A_i 表示第 i 个人拿到自己的帽子，则

$$\mathrm{P}(A_i) = \frac{1}{n} \equiv \frac{(n-1)!}{n!}.$$

至少一个人拿到自己的帽子：

$$\begin{aligned}\mathrm{P}\left(\sum_{i=1}^{n} A_i\right) &= \sum_{r=1}^{n}(-)^{r+1}\sum_{i_1<\cdots<i_r}\mathrm{P}\left(A_{i_1}\cdots A_{i_r}\right) \\ &= \sum_{r=1}^{n}(-)^{r+1}\binom{n}{r}\frac{(n-r)!}{n!} = \sum_{r=1}^{n}(-)^{r+1}\frac{1}{r!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} + \cdots + (-)^{n+1}\frac{1}{n!}.\end{aligned}$$

故所求概率即

$$\mathrm{P}(A_1^{\mathrm{c}}\cdots A_n^{\mathrm{c}}) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-)^{n}\frac{1}{n!} \to \frac{1}{\mathrm{e}}.$$

1.3　条件概率

定义 1.3.1: 条件概率

定义给定 B 事件发生的条件下，A 事件发生的条件概率 (conditional probability) 为

$$\mathrm{P}(A|B) := \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} \tag{1.3}$$

其中 $\mathrm{P}(B) > 0$.

条件概率可用于缩小样本空间.

事实上，给定 B 且 $\mathrm{P}(B) > 0$，则 $\mathrm{P}(\cdot|B) : \mathcal{F} \to \mathbb{R}$ 是概率函数，$\big(\Omega, \mathcal{F}, \mathrm{P}(\cdot|B)\big)$ 仍为概率空间.

定理 1.3.1: 乘法法则

由条件概率的定义可以直接导出：

$$\mathrm{P}(AB) = \mathrm{P}(B)\,\mathrm{P}(A|B) = \mathrm{P}(A)\,\mathrm{P}(B|A). \tag{1.4}$$

一般推广：

$$\mathrm{P}(A_1 \cdots A_n) = \mathrm{P}(A_1)\,\mathrm{P}(A_2|A_1) \cdots \mathrm{P}(A_n|A_1 \cdots A_{n-1}).$$

加多限制条件以后，算概率可能会变简单.

称 $\mathrm{P}(A)$ 为**先验概率** (priori probability)，$\mathrm{P}(A|B)$ 为**后验概率** (posterior probability).
已观测到 A 事件，等价于 $\mathrm{P}(A|A) \equiv 1$，绝非 $\mathrm{P}(A) = 1$.

1.4　独立事件

定义 1.4.1: 独立事件

定义 A, B 相互独立 (independent) 当

$$\mathrm{P}(AB) = \mathrm{P}(A)\,\mathrm{P}(B). \tag{1.5}$$

即，B 事件发生与否，不影响 A 事件发生的概率：

$$\mathrm{P}(A|B) = \mathrm{P}(A) \iff \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A\Omega)}{\mathrm{P}(\Omega)}.$$

易证，若 A, B 独立，则 A^{c}, B 独立，A, B^{c} 独立，$A^{\mathrm{c}}, B^{\mathrm{c}}$ 独立.

定义 1.4.2: 三个事件的独立

A, B, C 相互独立等价于：

1. A, B, C 两两独立；
2. $\mathrm{P}(ABC) = \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C)$

需要注意，只知其一不可推出另一个条件.
进而定义可数个事件 $A_1, A_2, \ldots$ 相互独立：任取有限个事件 $A_{i_1}, \ldots, A_{i_n}$ 都有

$$\mathrm{P}(A_{i_1} \cdots A_{i_n}) = \mathrm{P}(A_{i_1}) \cdots \mathrm{P}(A_{i_n}).$$

定义 1.4.3: 条件独立

若

$$\mathrm{P}(AB|E) = \mathrm{P}(A|E)\,\mathrm{P}(B|E),$$

则 A, B 关于事件 E 条件独立 (conditional independent).

A, B 关于 E 条件独立与 A, B 独立无关，二者之间既不充分也不必要.

1.5　Bayes 公式

定理 1.5.1: 全概率公式

定义 Ω 的一个分割 $\{B_i\}$，满足：

1. $\sum B_i = \Omega$;
2. $\forall i \neq j,\ B_iB_j = \varnothing$;
3. $\mathrm{P}(B_i) > 0$.

则 A 事件的概率

$$\mathrm{P}(A) = \sum \mathrm{P}(B_i)\,\mathrm{P}(A|B_i) \tag{1.6}$$

例 1.5.1: 假阳性悖论

$B =$ 患病，$A =$ 阳性，$\mathrm{P}(B) = 10^{-4}, \mathrm{P}(A|B) = 0.99, \mathrm{P}(A|B^{\mathrm{c}}) = 10^{-3}$

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(AB)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B)\,\mathrm{P}(A|B)}{\mathrm{P}(B)\,\mathrm{P}(A|B) + \mathrm{P}(B^{\mathrm{c}})\,\mathrm{P}(A|B^{\mathrm{c}})} \doteq 9\%.$$

例 1.5.2: 赌徒

两个人赌博．假设甲的赌本为 i 元，乙的赌本为 $n-i$．甲赢的概率为 p．每赌一局输家给赢家 1 元，其中一人输光游戏结束．求甲成为最终赢家的概率 Q_i．

$A =$ 甲最终赢，$B =$ 甲本局赢，由全概率公式

$$\mathrm{P}(A) = \mathrm{P}(B)\,\mathrm{P}(A|B) + \mathrm{P}(B^{\mathrm{c}})\,\mathrm{P}(A|B^{\mathrm{c}}).$$

即

$$Q_i = pQ_{i+1} + (1-p)Q_{i-1},$$

且有边界条件：$Q_0 = 0, Q_n = 1$，

$$Q_i = \begin{cases} \dfrac{i}{n}, & p = 1/2 \\ \dfrac{1-\lambda^i}{1-\lambda^n}, & p \neq 1/2,\ \lambda := \dfrac{1-p}{p}. \end{cases}$$

定理 1.5.2: Beyas 公式

Beyas 公式即

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(B_i)\,\mathrm{P}(A|B_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B_i)\,\mathrm{P}(A|B_i)}{\sum_j \mathrm{P}(B_j)\,\mathrm{P}(A|B_j)} \tag{1.7}$$

这个公式的重要性不仅在数学意义上，还在于先验概率 $\mathrm{P}(B_i)$ v.s. 后验概率 $\mathrm{P}(B_i|A)$

注. 计算正确的概率、正确计算概率、正确使用概率.

第二章　随机变量及其分布

定义 2.0.1: (一维) 随机变量

定义随机变量 (random variable) $X:\Omega\to\mathbb{R}$ 是样本空间上的实值函数. 有

- 离散型 (discrete)：至多可数个取值；
- 连续型 (continuous)：区间型取值 (不严格)；
- 其他

$\forall I\subset\mathbb{R}$ 可测，记原像集 $X^{-1}(I)\in\mathcal{F}$，定义[I]

$$\mathrm{P}_X(X\in I):=\mathrm{P}(X^{-1}(I)),\quad \forall I\subset\mathbb{R}\text{ 可测}$$

一般记 P_X 为 P.

定义 2.0.2: 累计分布函数

记 X 的累计分布函数 (cumulative distribution function, CDF)

$$\mathrm{F}(x):=\mathrm{P}(X\leqslant x),\quad \forall x\in\mathbb{R}$$

则 $\mathrm{P}(a<X\leqslant b)\equiv\mathrm{F}(b)-\mathrm{F}(a)$.

定理 2.0.1: PDF 的性质

1. $\mathrm{F}(x)$ 单调递增 (不严格单调)；
2. $\lim\limits_{x\to+\infty}\mathrm{F}(x)=1,\ \lim\limits_{x\to-\infty}\mathrm{F}(x)=0$;
3. $\mathrm{F}(x)$ 右连续，不一定左连续.

注.

1. 随机要素来自样本点 ω 的随机选择；
2. X,Y 同样本空间时，一般地，$aX+bY$ 等 $g(X,Y)$ 也是随机变量；
3. 随机变量同分布 $\iff$ CDF 相同；但不代表变量相同.

[I] 好看的鱼 v.s. 好吃的鱼

2.1　离散分布

定义 2.1.1: 离散分布

离散分布可由分布列 (probability distribution) 表示概率在所有的可能发生的情况中的分布

X	x_1	x_2	$\cdots$	x_n
p	p_1	p_2	$\cdots$	p_n

概率质量函数 (probability mass function, PMF)

$$f(x) = \mathrm{P}(X = x), \quad \forall x \in \mathbb{R}$$

离散分布的 CDF 为阶梯函数.

定义 2.1.2: 期望和方差

期望 (expectation) 即均值

$$\mathrm{E}(X) := \sum x_i p_i, \tag{2.1}$$

期望存在要求级数绝对收敛.

方差 (variance)

$$\mathrm{Var}(X) := \sum \big(x_i - \mathrm{E}(X)\big)^2 p_i \equiv \mathrm{E}(X^2) - \mathrm{E}(X)^2. \tag{2.2}$$

例 2.1.1

随机变量 X 的函数 $g(X)$ 的期望

$$\mathrm{E}(g(X)) = \sum g(x_i) p_i.$$

定义 2.1.3: Bernoulli 分布

Bernoulli 分布也称 0-1 分布，p 为成功概率，记作 $X \sim \mathrm{B}(p)$，其分布列为

X	0	1
p	$1-p$	p

定义 2.1.4: 二项分布

n 次独立 Bernoulli 试验的成功次数 X 服从二项分布 (binominal distribution)，记作

$X \sim \mathrm{B}(n,p)$，其分布列为：

$$\mathrm{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,\dots,n$$

$\mathrm{E}(X)=np,\quad \mathrm{Var}(X)=np(1-p).$

定义 2.1.5: Poisson 分布

一段时间或一定空间内出现的小概率事件次数 X 服从 Poisson 分布，记作 $X\sim\pi(\lambda)$，其分布列为：

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

$\mathrm{E}(X)=\lambda,\quad \mathrm{Var}(X)=\lambda.$

定理 2.1.1: 二项分布趋于 Poisson 分布

$n\to\infty$ 时，二项分布 $\mathrm{B}(n,\lambda/n)$ 趋于 Poisson 分布 $\pi(\lambda)$

$$\begin{aligned}\binom{n}{k}p^k(1-p)^{n-k}&=\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\&=\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\frac{n!}{n^k(n-k)!}\cancel{\left(1-\frac{\lambda}{n}\right)^{-k}}\to\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}.\end{aligned}$$

用 $\pi(np)$ 近似 $\mathrm{B}(n,p)$ 的误差最多为 $\min(p,np^2)$，证略.

若试验不独立，但满足弱相依条件下，Poisson 分布仍为较好近似.

例 2.1.2: 弱相依条件举例：配对问题

在例 1.2.2 中，尽管 A_i 和 A_j 并不独立，但弱相依

$$\mathrm{P}(A_i)=\frac{1}{n}\doteq\mathrm{P}(A_i|A_j)=\frac{1}{n-1}.$$

记 $X=$ 拿到自己帽子的人数，当 $n\to\infty$ 时，$X\sim\pi(1)$

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

下面用常规做法检验.

$E=$ 指定的 k 个人拿到自己的帽子，
$F=$ 余下 $n-k$ 个人都未拿到自己的帽子

$$\mathrm{P}(EF)=\mathrm{P}(E)\,\mathrm{P}(F|E)=\frac{(n-k)!}{n!}f(n-k),$$

其中 $f(n)$ 表示例 1.2.2 中的结果，故

$$\mathrm{P}(X=k)=\binom{n}{k}\mathrm{P}(EF)=\frac{1}{k!}f(n-k)\to\frac{\mathrm{e}^{-1}}{k!}.$$

2.2　连续分布

定义 2.2.1: 连续分布

若存在 $f(x) \geqslant 0$，使得 $\forall I \subset \mathbb{R}$ 可测，都有

$$\mathrm{P}(X \in I) = \int_I f(x)\,\mathrm{d}x$$

则称 X 为连续型随机变量，服从连续分布，$f(x)$ 为概率密度函数 (probability density function, PDF).

定理 2.2.1: 连续分布性质

- $\forall x \in \mathbb{R},\ \mathrm{P}(X = x) \equiv 0$;
- 归一性：

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x \equiv 1; \tag{2.3}$$

- 期望和方差

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x, \tag{2.4}$$

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathrm{E}(X))^2 f(x)\,\mathrm{d}x. \tag{2.5}$$

  若期望存在，则要求上积分式**绝对收敛**.
- 连续分布的 CDF $\mathrm{F}(x)$ 连续且可导，且

$$\mathrm{F}'(x) = f(x). \tag{2.6}$$

  若 $\mathrm{F}(x)$ 严格递增，则 $\mathrm{F}^{-1}(y)$ 存在；但若其不严格递增，也可 well define

$$\mathrm{F}^{-1}(y) := \inf\{x \,|\, \mathrm{F}(x) \geqslant y\}. \tag{2.7}$$

定义 2.2.2: 均匀分布

均匀分布 (uniform distribution) 记作 $X \sim \mathrm{U}(a, b)$，

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a, b) \\ 0, & \text{elsewhere} \end{cases} \tag{2.8}$$

$\mathrm{E}(X) = \dfrac{a+b}{2}$，　$\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$.

定义 2.2.3: 正态分布

正态分布 (normal distribution) 记作 $X \sim \mathrm{N}(\mu, \sigma^2)$，

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad x \in \mathbb{R} \tag{2.9}$$

$\mathrm{F}(x) =: \Phi(x)$, $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$;

定义 2.2.4: 指数分布

指数分布 (exponential distribution) 通常刻画寿命、等待时间，记作 $X \sim \mathrm{Exp}(\lambda)$，

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{2.10}$$

定义平均寿命

$$\beta := \mathrm{E}(X) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \beta^2.$$

也有教材因此记作 $\mathrm{Exp}(\beta)$.

指数分布的尾概率

$$\mathrm{P}(X > x) = \mathrm{e}^{-\lambda x},$$

与 Poisson 分布有关.

例 2.2.1: 指数分布的导出

失效率

$$\begin{aligned} \mathrm{P}(x < X < x + \mathrm{d}x | X > x) &= \frac{\mathrm{P}(x < X < x + \mathrm{d}x)}{\mathrm{P}(X > x)} \\ &= \frac{\mathrm{F}(x + \mathrm{d}x) - \mathrm{F}(x)}{1 - \mathrm{F}(x)} \doteq \frac{\mathrm{F}'(x)}{1 - \mathrm{F}(x)} \mathrm{d}x \end{aligned}$$

令失效率为 $\lambda(x)\,\mathrm{d}x$

$$\frac{\mathrm{F}'(x)}{1 - \mathrm{F}(x)} = \lambda(x), \implies \mathrm{F}(x) = 1 - \exp\left[-\int_0^x \lambda(t)\,\mathrm{d}t\right].$$

若假设无老化：$\lambda(t) \equiv \lambda$，则分布为指数分布：

$$\mathrm{F}(x) = 1 - \mathrm{e}^{-\lambda x}, \quad x > 0, \tag{2.11}$$

这体现出指数分布的无记忆性：

$$\mathrm{P}(X > t + \tau | X > \tau) = \frac{\mathrm{e}^{-\lambda(t+\tau)}}{\mathrm{e}^{-\lambda\tau}} = \mathrm{e}^{-\lambda t} = \mathrm{P}(X > t)$$

与 τ 无关.

改进 $\lambda(x)=\alpha x^{\alpha+1}/\beta^{\alpha}$，得到 Weibull 分布

$$\mathrm{F}(x)=1-\mathrm{e}^{-(x/\beta)^{\alpha}},\quad x>0.$$

2.3　随机变量的函数

设 $Y=g(X)$，X 离散可推出 Y 离散.

定理 2.3.1: 连续型随机变量的函数

若 g 处处可导且严格单调，则 $Y=g(X)$ 的 PDF 为

$$f_Y(y)=f_X\big(g^{-1}(y)\big)\left|\big[g^{-1}(y)\big]'\right|. \tag{2.12}$$

其本质是

$$\mathrm{F}_Y(y)=\mathrm{F}_X\big(g^{-1}(y)\big). \tag{2.13}$$

例 2.3.1: 生成随机变量

服从 CDF $\mathrm{F}(y)$ 的随机变量 Y 可由随机数 $X\sim\mathrm{U}(0,1)$ 生成：

$$Y=\mathrm{F}^{-1}(X). \tag{2.14}$$

Review

1. PMF/PDF, CDF
2. 期望 μ、标准差 σ；标准化 $\dfrac{X-\mu}{\sigma}$
3. 参数的意义：位置、尺度、形状
4. $Y=g(X)$

第三章　联合分布

3.1　随机向量

定义 3.1.1: 联合 CDF

$(X_1,\dots,X_n):\Omega\to\mathbb{R}^n$ 是 (n 维) 随机向量，$X_i(i=1,\dots,n)$ 为随机变量. 可定义联合 CDF：

$$\mathrm{F}(x_1,\dots,x_n):=\mathrm{P}(X_1\leqslant x_1,\dots,X_n\leqslant x_n),\quad (x_1,\dots,x_n)\in\mathbb{R}^n$$

特别地，当 $n=2$ 时，联合分布称为二元分布.

离散分布　若 X_i 均为离散型，则称 $(X_1,\dots,X_n)$ 为离散型随机向量.

例 3.1.1: 多项分布

给定 Ω 的一个分割 $\{B_i\}$，$\mathrm{P}(B_i)=:p_i$.
试验 n 次，记 X_i 是 B_i 发生的次数，则 $(X_1,\dots,X_n)$ 服从多项分布 (multinomial distribution)，记作 $(X_1,\dots,X_n)\sim\mathrm{B}(n,p_1,\dots,p_n)$，其联合 PDF

$$\mathrm{P}(X_1=k_1,\dots,X_n=k_n)=\binom{n}{k_1,\dots,k_n}p_1^{k_1}\cdots p_n^{k_n}.$$

其中多项式系数

$$\binom{n}{k_1,\dots,k_n}:=\frac{n!}{k_1!\cdots k_n!}.$$

连续分布　若存在 $f\geqslant 0$ 且 $\forall Q\subset\mathbb{R}^n$ 可测，都有

$$\mathrm{P}\big((X_1,\dots,X_n)\in Q\big)=\int_Q f(x_1,\dots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n$$

则称 $(X_1,\dots,X_n)$ 为连续型，且 f 为其 PDF.

定理 3.1.1: 连续分布相关性质

1. 归一性：

$$\int_{\mathbb{R}^n} f(x_1,\dots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n\equiv 1;$$

2. 以 $n=2$ 为例，其 CDF 为

$$\mathrm{F}(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(\xi,\eta)\,\mathrm{d}\xi\mathrm{d}\eta,\quad f(x,y)=\frac{\partial^2\mathrm{F}}{\partial x\partial y}(x,y).$$

例 3.1.2: 二元矩形均匀分布

矩形区域

$$f(x,y)=\begin{cases}\dfrac{1}{(b-a)(c-d)}, & (x,y)\in(a,b)\times(c,d)\\ 0, & \text{elsewhere}\end{cases}$$

例 3.1.3: 二元正态分布

$(X,Y)\sim\mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，$\forall(x,y)\in\mathbb{R}^2,|\rho|<1$

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\cdot\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2-2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2}\right]\right\}\tag{3.1}$$

记

$$X=\left[\frac{x-\mu_1}{\sigma_1},\frac{y-\mu_2}{\sigma_2}\right]^\top$$

则指数项部分构成了一个二次型：

$$-\frac{1}{2}X^\top WX,\quad W=\frac{1}{1-\rho^2}\begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix}$$

对正定的 W 进行 Cholesky 分解 $W=A^\top A$，可取

$$A=\frac{1}{\sqrt{1-\rho^2}}\begin{bmatrix}1 & -\rho\\ 0 & \pm\sqrt{1-\rho^2}\end{bmatrix}$$

进而就可将二次型转化为 $\mathrm{e}^{-(x^2+y^2)/2}$ 的标准形式.

3.2　边际分布

定义 3.2.1: 边际分布

$(X_1,\ldots,X_n)$ 中 X_i 的边际 CDF

$$\mathrm{F}_{X_i}(x):=\mathrm{P}(X_i\leqslant x,-\infty<X_j<+\infty,\forall j\neq i)\tag{3.2}$$

以 $n=2$ 为例，(X,Y) 中 X 的边际 CDF

$$\mathrm{F}_X(x)=\lim_{y\to+\infty}\mathrm{F}(x,y)$$

离散型边际概率

$$\mathrm{P}(X=x)=\sum_y \mathrm{P}(X=x,Y=y);$$

连续型边际 PDF

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y.$$

有

$$\mathrm{P}(X>a,Y>b)=1-\mathrm{F}_X(a)-\mathrm{F}_Y(b)+\mathrm{F}(a,b)$$

例 3.2.1: 二元正态的边际

边际密度

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y=\cdots=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right].$$

即 $X\sim\mathrm{N}(\mu_1,\sigma_1^2)$，同理 $Y\sim\mathrm{N}(\mu_2,\sigma_2^2)$.

联合分布可确定边际分布，反之不行.

例 3.2.2: Farlie-Gumbel-Morgenstern 族

随机变量 X,Y 的 CDF 分别为 $\mathrm{F}(x),\mathrm{G}(y)$，则 $\forall\alpha\in[-1,1]$,

$$\mathrm{H}(x,y)=\mathrm{F}(x)\mathrm{G}(y)\big[1+\alpha\big(1-\mathrm{F}(x)\big)\big(1-\mathrm{G}(y)\big)\big] \tag{3.3}$$

是 (X,Y) 的二元 CDF. 且 X,Y 的边际分布分别为 $\mathrm{F}(x),\mathrm{G}(y)$.

例 3.2.3: 连接函数

我们称边际分布为 $\mathrm{U}(0,1)$ 的联合 CDF $\mathrm{C}(u,v)$ 为连接 (Copula) 函数，则对于随机变量 X,Y，可用连接函数构造二元分布

$$\mathrm{H}(x,y)=\mathrm{C}\big(\mathrm{F}(x),\mathrm{G}(y)\big) \tag{3.4}$$

使其边际 CDF 为 $\mathrm{F}(x),\mathrm{G}(y)$.

3.3　条件分布

以 $n=2$ 为例，离散型

$$\mathrm{P}(X=x_i,Y=y_j)=:p_{ij}\geqslant 0,\quad \sum_{i,j}p_{ij}=1,$$

有条件概率

$$\mathrm{P}(X=x_i|Y=y_i)=\frac{\mathrm{P}(X=x_i,Y=y_i)}{\mathrm{P}(Y=y_i)}=\frac{p_{ij}}{\sum_k p_{kj}}=:\frac{p_{ij}}{p_{\cdot j}}. \tag{3.5}$$

连续型 PDF 为 $f(x,y)$，可定义条件 PDF[I]

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}. \tag{3.6}$$

条件 CDF

$$\mathrm{F}(a|y)=\mathrm{P}(X\leqslant a|Y=y)=\int_{-\infty}^{a}f_{X|Y}(x|y)\,\mathrm{d}x.$$

乘法法则

$$f(x,y)=f_Y(y)f_{X|Y}(x|y)=f_X(x)f_{Y|X}(y|x); \tag{3.7}$$

全概率公式

$$f_X(x)=\int_{-\infty}^{+\infty}f_{X|Y}(x|y)f_Y(y)\,\mathrm{d}y; \tag{3.8}$$

Bayes 公式

$$f_{Y|X}(y|x)=\frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}. \tag{3.9}$$

例 3.3.1: 二元正态的条件密度

$X,Y\sim\mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$

$$f_{Y|X}(y|x)=\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)^2\right],$$

当 $X=x$ 时，$Y\sim\mathrm{N}\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),(1-\rho^2)\sigma_2^2\right)$

3.4　独立性

定义 3.4.1: 联合分布的独立性

定义 (X,Y) 的 CDF 为 $\mathrm{F}(x,y)$，定义 X,Y 独立满足

$$\mathrm{F}(x,y)=\mathrm{F}_X(x)\,\mathrm{F}_X(y),\quad\forall x,y\in\mathbb{R} \tag{3.10}$$

进而定义 $X_1,\ldots,X_n$ 独立：

$$\mathrm{F}(x_1,\ldots,x_n)=\mathrm{F}_1(x_1)\cdots\mathrm{F}(x_n),\quad\forall x_1,\ldots,x_n\in\mathbb{R}$$

离散型中，即

$$p_{ij}=p_{i\cdot}p_{\cdot j},\quad\forall i,j$$

连续型中，即

$$f(x,y)=f_X(x)f_Y(y),\quad\forall x,y\in\mathbb{R}$$

[I]唐宏岩老师记作 $f_X(x|y)$

例 3.4.1: 二元正态分布的独立性

若 $(X,Y)\sim \mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则 X,Y 独立等价于 $\rho=0$.

定理 3.4.1: 独立性相关定理

1. 若 $X_1,\dots,X_n$ 独立，则
$$Y_1=g_1(X_1,\dots,X_m),\quad Y_2=g_2(X_{m+1},\dots,X_n)$$
独立；
2. 若联合 PDF 可以分离变量
$$f(x_1,\dots,x_n)=g_1(x_1)\cdots g_n(x_n),\quad x_1,\dots,x_n\in\mathbb{R}$$
则 $X_1,\dots,X_n$ 相互独立，且 g_i 与 f_i 相差常数因子.

3.5　随机向量的函数

密度函数变换　(X,Y) 为二维连续型随机变量，其 PDF 为 $f(x,y)$. 由此可导出随机变量 $U=u(X,Y),V=v(X,Y)$，求其 PDF.

若 $u=u(x,y)$ 和 $v=v(x,y)$ 可微可逆
$$\begin{cases}x=x(u,v)\\ y=y(u,v)\end{cases}$$
则该变换的 Jacobi 行列式
$$\det J=\left|\frac{\partial(x,y)}{\partial(u,v)}\right|=\begin{vmatrix}x_u & y_u\\ x_v & y_v\end{vmatrix}\neq 0$$
则 U,V 的联合分布 PDF 为
$$f_{U,V}(u,v)=f_{X,Y}\left[x(u,v),y(u,v)\right]|\det J|. \tag{3.11}$$
特别需要注意是否有“一对多”的情况，比如 $(U,V)=(X^2+Y^2,X/Y)$.

例 3.5.1: $X+Y$

(X,Y) 的和 $Z=X+Y$ 的 PDF 为
$$f_{X+Y}(z)=\int_{-\infty}^{+\infty}f(x,z-x)\,\mathrm{d}x\equiv\int_{-\infty}^{+\infty}f(z-y,y)\,\mathrm{d}y \tag{3.12}$$
特别地，当 X,Y 相互独立时，
$$f_{X+Y}(z)=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)\,\mathrm{d}x\equiv\int_{-\infty}^{+\infty}f_X(z-y)f_Y(y)\,\mathrm{d}y$$
这是 f_X,f_Y 的 Fourier 卷积，记作 f_X*f_Y.

可构造辅助随机变量 $(Z,W)=(X+Y,X)$，用式 (3.11) 得出.

例 3.5.2: $Y/X, XY$

$Z=Y/X, XY$ 的 PDF 为

$$f_{Y/X}(z)=\int_{-\infty}^{+\infty} f(x,xz)\,|x|\,\mathrm{d}x \tag{3.13}$$

$$f_{XY}(z)=\int_{-\infty}^{+\infty} f\left(x,\frac{z}{x}\right)\frac{1}{|x|}\,\mathrm{d}x \tag{3.14}$$

特别地，当 X,Y 相互独立时，

$$f_{Y/X}(z)=\int_{-\infty}^{+\infty} f_X(x)f_Y(xz)\,|x|\,\mathrm{d}x$$

$$f_{XY}(z)=\int_{-\infty}^{+\infty} f_X(x)f_Y\left(\frac{z}{x}\right)\frac{1}{|x|}\,\mathrm{d}x$$

第二个公式是 f_X, f_Y 的 Mellin 卷积.

例 3.5.3: $\max(X,Y), \min(X,Y)$

若 (X,Y) 独立，则 $\max(X,Y), \min(X,Y)$ 的分布为

$$\mathrm{F}_{\max}(z)=\mathrm{F}_X(x)\,\mathrm{F}_Y(y), \tag{3.15}$$

$$\mathrm{F}_{\min}(z)=1-[1-\mathrm{F}_X(x)][1-\mathrm{F}_Y(y)]. \tag{3.16}$$

上式很容易推导：

$$\mathrm{P}(\max(X,Y)<z)=\mathrm{P}(X<z,Y<z)=\mathrm{P}(X<z)\,\mathrm{P}(Y<z),$$

例 3.5.4: 离散分布的和

$X_i\sim \mathrm{B}(n_i,p)$ 独立，$Y=X_1+X_2$

$$\begin{aligned}\mathrm{P}(Y=k)&=\sum_{k_1=0}^{k}\mathrm{P}(X_1=k_1,X_2=k-k_1)\\&=\sum_{k_1=0}^{k}\binom{n_1}{k_1}\binom{n_2}{k-k_1}p^k(1-p)^{n_1+n_2-k}\\&=\binom{n_1+n_2}{k}p^k(1-p)^{n_1+n_2-k}\end{aligned}$$

故 $Y\sim \mathrm{B}(n_1+n_2,p)$.

$X_i \sim \pi(\lambda_i)$ 独立，$Y = X_1 + X_2$

$$\begin{aligned}\mathrm{P}(Y=k) &= \sum_{k_1=0}^{k} \frac{\lambda_1^{k_1}}{k_1!}\mathrm{e}^{-\lambda_1}\frac{\lambda_2^{k-k_1}}{(k-k_1)!}\mathrm{e}^{-\lambda_2} \\ &= \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)}}{k!}\sum_{k_1=0}^{k}\binom{k}{k_1}\lambda_i^{k_1}\lambda_2^{k-k_1} \\ &= \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)}}{k!}(\lambda_1+\lambda_2)^k\end{aligned}$$

故 $Y \sim \pi(\lambda_1 + \lambda_2)$.

在学习到矩母函数和特征函数后，由定理 4.6.2，这个结论就是显然的.

例 3.5.5: 连续分布的和

$(X_1, X_2) \sim \mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，$Y = X_1 + X_2$，

$$f(y) = \int_{-\infty}^{+\infty} f(x, y-x)\,\mathrm{d}x = \cdots$$

故 $Y = X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_1^2 + 2\rho\sigma_1\sigma_2)$.

例 3.5.6: 指数分布的最小值

若 $X_1 \sim \mathrm{Exp}(\lambda_1), X_2 \sim \mathrm{Exp}(\lambda_2)$，则

$$\mathrm{F}_{\min}(x) = 1 - \big[1 - \mathrm{F}_{X_1}(x)\big]\big[1 - \mathrm{F}_{X_2}(x)\big] = 1 - \mathrm{e}^{-(\lambda_1+\lambda_2)x}.$$

$\min(X_1, X_2) \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$.

例 3.5.7: max, min 的联合分布

显然，$Y := \min(X_1, \ldots, X_n)$ 和 $Z := \max(X_1, \ldots, X_n)$ 不是独立的，

$$\mathrm{P}(y \leqslant Y, Z \leqslant z) = \big[\mathrm{F}(z) - \mathrm{F}(y)\big]^n.$$

故

$$\begin{aligned}f_{Y,Z}(y,z) &= -\frac{\partial^2}{\partial y \partial z}\mathrm{P}(y \leqslant Y, Z \leqslant z) \\ &= n(n-1)f(y)f(z)\big[\mathrm{F}(z) - \mathrm{F}(y)\big]^{n-2}.\end{aligned}$$

3.6　三大分布

定义 3.6.1: 卡方分布

若 $X_1,\dots,X_n$ 独立且都服从 $\mathrm{N}(0,1)$ (独立同分布，independent identically distribution, iid)，则 $X:=X_1^2+\cdots+X_n^2$ 服从自由度为 n 的卡方分布，记作 $X\sim\chi^2(n)$，由数学归纳法可知

$$f(x)=\frac{1}{\Gamma(n/2)2^{n/2}}x^{n/2-1}\,\mathrm{e}^{-x/2},\quad x>0. \tag{3.17}$$

$\mathrm{E}(X)=n,\ \mathrm{Var}(X)=2n.$

定理 3.6.1: 卡方分布的性质

1. 若 $X_1\sim\chi^2(n_1),X_2\sim\chi^2(n_2)$ 独立，则 $X_1+X_2\sim\chi^2(n_1+n_2)$;
2. 若 $X\sim\mathrm{Exp}(\lambda)$，则 $2\lambda X\sim\chi^2(2)$.

定义 3.6.2: Student's t 分布

$X_1\sim\chi^2(n),X_2\sim\mathrm{N}(0,1)$ 独立，则

$$X:=\frac{X_2}{\sqrt{X_1/n}}\sim\mathrm{t}(n).$$

其 PDF 为

$$f(x)=\frac{1}{\sqrt{n}\mathrm{B}(n/2,1/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2},\quad x\in\mathbb{R}. \tag{3.18}$$

由 PDF，特别地，当 $n=1$ 时，是 Cauchy 分布；$n\to\infty$ 时，趋于 $\mathrm{N}(0,1)$

$\mathrm{E}(X)=0,\ \mathrm{Var}(X)=\dfrac{n}{n-2}$. 当 $n\leqslant 2$ 时，方差不存在.

定义 3.6.3: F 分布

$X_1\sim\chi^2(n),X_2\sim\chi^2(m)$ 独立，则

$$X:=\frac{X_1/n}{X_2/m}\sim\mathrm{F}(n,m).$$

其 PDF 为

$$f(x)=\frac{1}{\mathrm{B}(n/2,m/2)}\left(\frac{n}{m}\right)^{n/2}x^{n/2-1}\left(1+\frac{n}{m}x\right)^{-(n+m)/2},\quad x>0. \tag{3.19}$$

定理 3.6.2: 正态总体统计量的分布

$X_1 \dots, X_n$ iid $\sim \mathrm{N}(\mu, \sigma^2)$，则

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1); \tag{3.20}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1); \tag{3.21}$$

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}(n-1); \tag{3.22}$$

$$\frac{S^2/S'^2}{\sigma^2/\sigma'^2} \sim \mathrm{F}(n-1, n'-1). \tag{3.23}$$

证明. 式 (3.20) 是 trivial 的，因为 $\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$.

式 (3.22) 可由式 (3.21), (3.20) 结合 t 分布的定义推导出来

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}} \sim \mathrm{t}(n-1),$$

式 (3.23) 也可由式 (3.21) 结合 F 分布的定义推导而来，因此证明的关键在于证明式 (3.21).

找一个正交方阵 A，第一行各元均为 $1/\sqrt{n}$，作正交变换

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = A \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

则 $(X_1, \dots, X_n)$ 的 PDF 为

$$f(x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2\right)\right].$$

由于这是正交变换，$\det A = 1$，且有

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n X_i^2, \quad Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \equiv \sqrt{n}\,\overline{X}.$$

从而，$(Y_1, \dots, Y_n)$ 的 PDF

$$\begin{aligned} f(y_1, \dots, y_n) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\mu\sqrt{n} y_1 + n\mu^2\right)\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(y_1 - \sqrt{n}\mu)^2/2\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-y_2^2/2\sigma^2} \cdots \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-y_n^2/2\sigma^2}. \end{aligned}$$

故 $Y_1, \dots, Y_n$ 独立，且

$$Y_1 \sim \mathrm{N}(\sqrt{n}\mu, \sigma^2), \quad Y_i \sim \mathrm{N}(0, \sigma^2), \ i = 2, \dots, n.$$

从而 Y_1 与 $Y_2^2 + \cdots + Y_n^2$ 独立，又

$$\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \sum_{i=1}^n (X_i - \overline{X})^2.$$

故 $\overline{X}$ 与 S^2 独立，且

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=2}^{n}Y_i^2 \sim \chi^2(n-1).$$

□

第四章　随机变量的数字特征

4.1　期望

期望是集中趋势的一种刻画，定义 2.1.2 已经定义过

$$\mathrm{E}(X) := \begin{cases} \sum_{x\in\mathbb{R}} x f(x), & \text{离散} \\ \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x, & \text{连续} \end{cases} \tag{4.1}$$

期望存在要求**绝对收敛**．期望的一般定义

$$\mathrm{E}(X) = \int_0^1 x\,\mathrm{d}\,\mathrm{F}$$

其中 F 是 CDF，这是一个 Lebesque-Stieltjes 积分．

定理 4.1.1: 期望的性质

1. 多元：$\mathrm{E}\big[(X_1,\dots,X_n)\big] = \big(\mathrm{E}(X_1),\dots,\mathrm{E}(X_n)\big)$;
2. 函数：$\mathrm{E}\big[g(X_1,\dots,X_n)\big] = \int_{\mathbb{R}^n} g(x_1,\dots,x_n) f(x_1,\dots,x_n)\,\mathrm{d}x$;
3. 线性：$\mathrm{E}(aX+bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$;
4. 若 $X_1,\dots,X_n$ 独立[I]，则 $\mathrm{E}(X_1\cdots X_n) = \mathrm{E}(X_1)\cdots\mathrm{E}(X_n)$.

[I]后面很快会看到，有一个更弱的条件：$X_1,\dots,X_n$ 不相关.

4.2　分位数、众数

定义 4.2.1: 分位数

定义 $\forall\alpha\in(0,1)$ 若

$$\mathrm{P}(X\leqslant a)\geqslant\alpha,\quad \mathrm{P}(X\geqslant a)\geqslant 1-\alpha,$$

则称 a 为 X 的下 α-分位数 (lower α-quantile)．$a = \mathrm{F}^{-1}(\alpha)$.

特别地，$\alpha = 0.5$ 时，a 对应中位数，也是集中趋势的一种刻画．

众数　众数也是一种集中趋势的刻画，方便定义 $\arg\max f(x)$.

4.3　方差

定义 2.1.2 中已经定义过：

$$\mathrm{Var}(X) := \mathrm{E}\big[(X-\mathrm{E}(X))^2\big] \equiv \mathrm{E}(X^2) - \mathrm{E}^2(X) \tag{4.2}$$

标准差 (standard deviation, SD) $\mathrm{SD}(X) := \sqrt{\mathrm{Var}(X)}$.

若 X 期望为 μ，标准差为 σ，标准化

$$\check{X} := \frac{X-\mu}{\sigma}, \tag{4.3}$$

可使 $\mathrm{E}(\check{X}) = 0, \mathrm{Var}(\check{X}) = 1$.

定理 4.3.1: 方差的性质

1. $\mathrm{Var}(c) = 0,\ \mathrm{Var}(X+c) = \mathrm{Var}(X),\ \mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$;
2. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$;
3. X, Y 独立时，可推出 $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$，
   $\mathrm{Var}(XY) = \mathrm{Var}(X)\,\mathrm{Var}(Y) + \underline{\mathrm{E}^2(X)\,\mathrm{Var}(Y) + \mathrm{E}^2(Y)\,\mathrm{Var}(X)}$.

例 4.3.1: 期望在均方误差下的意义

期望是使得均方误差最小的常数估计，即 $\forall c$ 均有

$$\mathrm{E}\big[(X-c)^2\big] \geqslant \mathrm{Var}(X),$$

取 $=$ 当且仅当 $c = \mathrm{E}(X)$. 因为

$$\mathrm{E}\big[(X-c)^2\big] = \mathrm{E}(X^2) - 2\,\mathrm{E}(X)c + c^2 =: g(c)$$

当且仅当 $c = \mathrm{E}(X)$ 时取到 $g_{\min} = \mathrm{E}(X^2) - \mathrm{E}^2(X) = \mathrm{Var}(X)$.

例 4.3.2: 中位数在绝对值误差下的意义

m 为中位数，$\forall c$ 均有

$$\mathrm{E}(|X-c|) \geqslant \mathrm{E}(|X-m|)$$

取 $=$ 当且仅当 $c = m$.

4.4　协方差及相关系数

定义 4.4.1: 协方差

定义协方差 (covariance)

$$\mathrm{Cov}(X,Y) := \mathrm{E}\big[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\big] \equiv \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \tag{4.4}$$

显然，$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

协方差是可交换且双线性的.

定义 4.4.2: 协方差矩阵

定义 $X=(X_1,\dots,X_n), Y=(Y_1,\dots,Y_n)$ 的协方差矩阵 (covariance matrix)

$$\operatorname{Cov}(X,Y):=\Big[\operatorname{Cov}(X_i,Y_j)\Big]_{n\times n}, \tag{4.5}$$

相应的，$\operatorname{Cov}(X,X)$ 就是向量 X 的方差矩阵.

由协方差矩阵可以得到

$$\operatorname{Var}\Big(\sum_{i=1}^n X_i\Big)=\sum_{i=1}^n \operatorname{Var}(X_i)+2\sum_{i<j}\operatorname{Cov}(X_i,X_j)\equiv\sum_{i,j}\operatorname{Cov}(X_i,X_j). \tag{4.6}$$

例 4.4.1: 多项分布的协方差

在例 3.1.1 中给过多项分布 $(X_1,\dots,X_n)\sim \mathrm{B}(n,p_1,\dots,p_n)$，显然，其边际分布 $X_i\sim \mathrm{B}(n,p_i)$，下面给出 X_i,X_j 的协方差

$$\operatorname{Cov}(X_i,X_j)=-np_ip_j,\quad i\neq j.$$

得到这个结论的过程见条件期望部分的例 4.7.2.

例 4.4.2: 配对问题的期望和方差

尽管例 2.1.2 中 X 的分布列较复杂，我们依然可以通过简单的计算得到其期望和方差. 记 $X_i\sim \mathrm{B}\big(\mathrm{P}(A_i)\big)$ 表示第 i 个人是否拿到自己的帽子，从而

$$X=X_1+\cdots+X_n,$$

尽管 X_i 间并不独立，但其期望是相同的 $(1/n)$. 进而可以得到 X 的期望就是期望的和 (这并不要求独立性)

$$\mathrm{E}(X)=\sum_{i=1}^n \mathrm{E}(X_i)=n\cdot\frac{1}{n}=1.$$

而 $X_iX_j(i\neq j)$ 的期望

$$\mathrm{E}(X_iX_j)=\mathrm{P}(X_iX_j=1)+0=\frac{1}{n}\cdot\frac{1}{n-1},$$

从而协方差

$$\begin{aligned}\operatorname{Cov}(X_i,X_j)&=\mathrm{E}(X_iX_j)-\mathrm{E}(X_i)\,\mathrm{E}(X_j)\\&=\frac{1}{n}\cdot\frac{1}{n-1}-\frac{1}{n}\cdot\frac{1}{n}=\frac{1}{n^2(n-1)}.\end{aligned}$$

因此方差

$$\begin{aligned}\operatorname{Var}(X) &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i\neq j} \operatorname{Cov}(X_i, X_j) \\ &= n \cdot \frac{1}{n}\left(1-\frac{1}{n}\right) + n(n-1)\cdot \frac{1}{n^2(n-1)} = 1.\end{aligned}$$

定义 4.4.3: 相关系数

定义相关系数 (correlation coefficient)

$$\operatorname{Corr}(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \equiv \operatorname{E}\left(\frac{X-\operatorname{E}(X)}{\sqrt{\operatorname{Var}(X)}}, \frac{Y-\operatorname{E}(Y)}{\sqrt{\operatorname{Var}(Y)}}\right), \tag{4.7}$$

若 $\operatorname{Corr}(X,Y)=0$，称为 X,Y 不 (线性) 相关.

X,Y 不相关与

$$\operatorname{E}(XY) = \operatorname{E}(X)\operatorname{E}(Y), \quad \operatorname{Var}(X+Y) = \operatorname{Var}(X)+\operatorname{Var}(Y).$$

是等价的. 易见，独立是不相关的充分不必要条件.

定理 4.4.1: 相关系数的性质

$|\operatorname{Corr}(X,Y)| \leqslant 1$，取等号当且仅当

$$\operatorname{P}(Y = aX+b) = 1, \ \text{(almost surely, a.s.)}$$

只需考虑 Cauchy-Schwartz 不等式

$$\operatorname{E}^2(UV) \leqslant \operatorname{E}(U)\operatorname{E}(V), \tag{4.8}$$

这可以用 $\forall t\in\mathbb{R}, \operatorname{E}\left[(U-tV)^2\right] \geqslant 0$ 证明.

例 4.4.3: 二元正态分布的相关系数

取 $\xi := \dfrac{x-\mu_1}{\sigma_1}, \eta := \dfrac{y-\mu_2}{\sigma_2}$，则

$$\operatorname{Corr}(X,Y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \iint_{\mathbb{R}^2} \xi\eta \exp\left[-\frac{(\xi-\rho\eta)^2}{2(1-\rho^2)} - \frac{\eta^2}{2}\right] \mathrm{d}\xi\mathrm{d}\eta$$

由

$$\int_{-\infty}^{+\infty} x\,\mathrm{e}^{-a(x-b)^2+c}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}}\,b\,\mathrm{e}^{c}.$$

可得

$$\begin{aligned}\operatorname{Corr}(X,Y) &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\eta\sqrt{2\pi(1-\rho^2)}\cdot\rho\eta\,\mathrm{e}^{-\eta^2/2}\,\mathrm{d}\eta \\ &= \frac{\rho}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\eta^2\,\mathrm{e}^{-\eta^2/2}\,\mathrm{d}\eta = \rho.\end{aligned}$$

4.5　矩

定义 4.5.1: 矩

X 的关于 c 点的 k 阶矩 (moment)

$$\mathrm{E}\big[(X-c)^k\big], \tag{4.9}$$

特别地，当 $c=0$ 时，称为原点矩；当 $c=\mathrm{E}(X)$ 时，称为中心矩.

故 $\mathrm{E}(X)$ 为 1 阶原点矩，$\operatorname{Var}(X)$ 为 2 阶中心矩.

定义 4.5.2: 偏度系数

定义 3 阶标准矩为偏度系数 (skewness)

$$\operatorname{skew}(X) := \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]. \tag{4.10}$$

1 阶标准矩为 0，2 阶标准矩为 1. 当中心往右偏时，偏度系数 <0.

定义 4.5.3: 峰度系数

定义 4 阶标准矩为峰度系数 (kurtosis)

$$\operatorname{kurt}(X) := \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]. \tag{4.11}$$

正态分布的峰度系数为 3. 一般地，尖峰厚尾的峰度系数 >3.

4.6　矩母函数

定义 4.6.1: 矩母函数

若下面期望在 t 的某含 0 邻域内存在，

$$\mathrm{M}_X(t) := \mathrm{E}\left(\mathrm{e}^{tX}\right) \tag{4.12}$$

则称 $\mathrm{M}_X(t)$ 为 X 的矩母函数 (moment generating function, MGF).

矩母函数的意义是用来确定矩

$$\mathrm{E}(X^n) = \mathrm{M}_X^{(n)}(0) \tag{4.13}$$

例 4.6.1: 指数分布和正态分布的矩母函数

若 $X \sim \mathrm{Exp}(\lambda)$，则

$$\mathrm{M}_X(t) = \int_0^{+\infty} \mathrm{e}^{tx}\lambda\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{\lambda}{\lambda - t}, \quad t \in (-\lambda, \lambda).$$

若 $X \sim \mathrm{N}(\mu, \sigma^2)$，则

$$\mathrm{M}_X(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} \mathrm{e}^{tx}\,\mathrm{e}^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x = \exp\Big(\mu t + \frac{1}{2}\sigma^2 t^2\Big).$$

特别地，若 $X \sim \mathrm{N}(0,1)$，则

$$\mathrm{M}_X(t) = \mathrm{e}^{-t^2/2} = \sum_{n=0}^{\infty} \frac{(2n)!}{2^n n!}\frac{t^{2n}}{(2n)!}.$$

故 $\mathrm{E}(X^{2k+1}) = 0$，$\mathrm{E}(X^{2k}) = (2k-1)!!$

定理 4.6.1: 矩母函数确定分布

若存在 $a > 0$，使得

$$\mathrm{M}_X(t) = \mathrm{M}_Y(t), \quad \forall t \in (-a, a)$$

则 X, Y 同分布.

证明超过了本节课的范围.

例 4.6.2: 离散分布的矩母函数

从 X 的矩母函数，比如

$$\mathrm{M}_X(t) = \frac{1}{4}\mathrm{e}^{-t} + \frac{1}{2} + \frac{1}{8}\mathrm{e}^{4t} + \frac{1}{8}\mathrm{e}^{5t},$$

可直接看出 X 的分布列.

例 4.6.3: 同矩不一定同分布

对数正态分布 $\ln X_1 \sim \mathrm{N}(0,1)$，则 X_1 的 PDF

$$f_1(x) = \frac{1}{\sqrt{2\pi}x}\mathrm{e}^{-\ln^2 x/2}, \quad x > 0$$

构造 X_2，其 PDF

$$f_2(x) = f_1(x)\big[1 + \sin(2\pi \ln x)\big] \geqslant 0.$$

可以证明，X_1, X_2 所有的矩均相等

$$\mathrm{E}(X_2^n) = \mathrm{E}(X_1^n) + \int_0^{+\infty} x^n f_1(x) \sin(2\pi \ln x)\,\mathrm{d}x$$

后面的积分做换元 $y := \ln x - n$ 可积得结果为 0.

从上面构造的例子可以看出：同矩不一定同分布，因此矩母函数相同是比所有矩相同更强的条件.

定理 4.6.2: 独立随机变量和的分布

若 X, Y 独立，则

$$\mathrm{M}_{X+Y}(t) = \mathrm{M}_X(t)\,\mathrm{M}_Y(t).$$

因此，n 个独立的正态分布 $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ 的和仍然服从正态分布，且

$$\sum_{i=1}^n X_i \sim \mathrm{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

定义 4.6.2: 联合分布的矩母函数

联合分布的矩母函数定义为

$$\mathrm{M}_{X_1,\ldots,X_n}(t_1,\ldots,t_n) := \mathrm{E}\left(\mathrm{e}^{t_1X_1+\cdots+t_nX_n}\right). \tag{4.14}$$

也可以通过矩母函数确定分布.

定义 4.6.3: 概率母函数

离散型随机变量 X 取值为非负整数

$$p_k = \mathrm{P}(X = k), \quad k = 0, 1, 2, \ldots$$

则概率母函数 (probability generating function, PGF)

$$g(t) := \mathrm{E}(t^X) = \sum_{k=0}^{\infty} t^k p_k.$$

概率母函数给出了各点的概率

$$\mathrm{P}(X = k) = \frac{g^{(k)}(0)}{k!}$$

也给出了期望 $\mathrm{E}(X) = g'(1)$.

定义 4.6.4: 特征函数

定义特征函数 (characteristic function, CF)

$$\psi(t) := \mathrm{E}\left(\mathrm{e}^{\mathrm{i}tX}\right). \tag{4.15}$$

与矩母函数不同，特征函数总是存在的.

4.7　条件期望

定义 4.7.1: 条件期望

定义条件期望

$$\mathrm{E}(Y|X\in A) := \begin{cases} \sum\limits_i y_i \,\mathrm{P}(Y=y_i|X\in A), & \text{离散} \\ \int_{\mathbb{R}} y f(y|X\in A)\,\mathrm{d}y, & \text{连续} \end{cases} \tag{4.16}$$

若 $\mathrm{P}(X\in A)>0$，条件期望可由下式计算

$$\mathrm{E}(Y|X\in A) = \frac{\mathrm{E}(Y, X\in A)}{\mathrm{P}(X\in A)}. \tag{4.17}$$

$\mathrm{E}(Y|X)$ 是 X 的函数，是一个新的随机变量 (Y 对 X 的回归函数).

例 4.7.1: 二元正态分布的条件期望

$(X,Y)\sim \mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，由例 3.3.1

$$\mathrm{E}(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X-\mu_1)^2.$$

定理 4.7.1: 全期望公式

可由全概率公式推出

$$\mathrm{E}(Y) = \mathrm{E}\big[\mathrm{E}(Y|X)\big]. \tag{4.18}$$

定理 4.7.2: 条件期望的性质

1. $\mathrm{E}\big[g(X)h(Y)|X\big] = g(X)\,\mathrm{E}\big[h(Y)|X\big]$;
2. $\mathrm{E}\big[\mathrm{E}\big[g(X,Y)|X\big]\big] = \mathrm{E}\big[g(X,Y)\big]$;
3. 若 X,Y 独立，则 $\mathrm{E}(Y|X)=\mathrm{E}(Y)$ 为常数，与 X 无关.

注意第 3 条，若 $\mathrm{E}(Y|X)=\mathrm{E}(Y)$，则 X,Y 不相关：

$$\mathrm{E}(XY) = \mathrm{E}\big[\mathrm{E}(XY|X)\big] = \mathrm{E}\big[X\,\mathrm{E}(Y|X)\big] = \mathrm{E}\big[X\,\mathrm{E}(Y)\big] = \mathrm{E}(X)\,\mathrm{E}(Y).$$

推不出 X,Y 独立.

例 4.7.2: 多项分布的协方差

接例 4.4.1，由全期望公式

$$\mathrm{E}(X_iX_j)=\mathrm{E}\big[\mathrm{E}(X_iX_j|X_j)\big]=\mathrm{E}\big[X_j\,\mathrm{E}(X_i|X_j)\big].$$

而 $\mathrm{E}(X_i|X_j)$ 可以理解成先从 n 份中抽出 X_j 个 B_j，剩下变成新的多项分布

$$(\ldots,X_i,\ldots)_{i\neq j}\sim\mathrm{B}\Big(n-X_j,\ldots,\frac{p_i}{1-p_j},\ldots\Big)_{i\neq j}$$

继而

$$\mathrm{E}(X_i|X_j)=(n-X_j)\frac{p_i}{1-p_j}.$$

$$\begin{aligned}\mathrm{E}(X_iX_j)&=\frac{p_i}{1-p_j}\big[n\,\mathrm{E}(X_j)-\mathrm{E}(X_j^2)\big]\\&=\frac{p_i}{1-p_j}\big[n\cdot np_j-\big(np_j(1-p_j)+(np_j)^2\big)\big]\\&=n(n-1)p_ip_j.\end{aligned}$$

协方差

$$\mathrm{Cov}(X_i,X_j)=n(n-1)p_ip_j-np_i\cdot np_j=-np_ip_j.$$

定理 4.7.3: 最优预测

用 $g(X)$ 估计 Y 的均方误差

$$\mathrm{E}\big[(Y-g(X))^2\big]\geqslant\mathrm{E}\big[(Y-\mathrm{E}(Y|X))^2\big].\tag{4.19}$$

这源于例 4.3.1 所给出的：

$$\mathrm{E}\big[(Y-g(X))^2|X\big]\geqslant\mathrm{E}\big[(Y-\mathrm{E}(Y|X))^2|X\big].$$

事实上这也给出了条件期望 $\mathrm{E}(Y|X)$ 的几何意义：Y 在 X 张成的线性空间上的投影.

例 4.7.3: 误差

Y 对 X 的回归函数 $\hat{Y}:=\mathrm{E}(Y|X)$，误差 $\tilde{Y}:=Y-\hat{Y}$，易得

$$\mathrm{E}(\hat{Y})=\mathrm{E}(Y),\quad\mathrm{E}(\tilde{Y})=0,$$

且 $\hat{Y},\tilde{Y}$ 不相关，也就是说

$$\mathrm{Var}(Y)=\mathrm{Var}(\hat{Y})+\mathrm{Var}(\tilde{Y}).$$

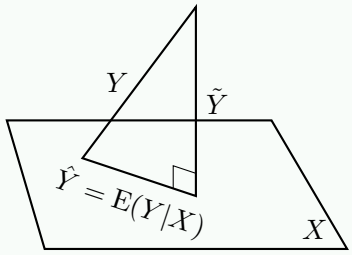


图 4.1: $Y, \hat{Y}, \tilde{Y}$ 关系示意图

定义 4.7.2: 条件方差

定义条件方差

$$\mathrm{Var}(Y|X) := \mathrm{E}\big[\big(X - \mathrm{E}(Y|X)\big)^2|X\big] \tag{4.20}$$

自然

$$\mathrm{Var}(Y|X) = \mathrm{E}(Y^2|X) - \mathrm{E}^2(Y|X). \tag{4.21}$$

定理 4.7.4: 全方差公式

由例 4.7.3

$$\mathrm{Var}(Y) = \mathrm{E}\big[\mathrm{Var}(Y|X)\big] + \mathrm{Var}\big[\mathrm{E}(Y|X)\big]. \tag{4.22}$$

例 4.7.4: 随机变量个随机变量的和

随机变量序列 $X_1, X_2, \ldots$ iid，N 是取值为正整数的随机变量，且与 X_i 独立，令 $Y = X_1 + \cdots + X_N$，则

$$\mathrm{E}(Y) = \mathrm{E}\big[\mathrm{E}(Y|N)\big] = \mathrm{E}\big[N\,\mathrm{E}(X)\big] = \mathrm{E}(N)\,\mathrm{E}(X); \tag{4.23}$$

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}\big[\mathrm{Var}(Y|N)\big] + \mathrm{Var}\big[\mathrm{E}(Y|N)\big] \\ &= \mathrm{E}\big[N\,\mathrm{Var}(X)\big] + \mathrm{Var}\big[N\,\mathrm{E}(X)\big] \\ &= \mathrm{Var}(X)\,\mathrm{E}(N) + \mathrm{E}^2(X)\,\mathrm{Var}(N).\end{aligned} \tag{4.24}$$

特别地，$N \sim \pi(\lambda)$ 时，称 Y 为复合 Poisson 随机变量

$$\mathrm{Var}(Y) = \lambda\big[\mathrm{Var}(X) + \mathrm{E}^2(X)\big] = \lambda\,\mathrm{E}(X^2).$$

第五章　不等式与极限定理

5.1　概率不等式

定理 5.1.1: Markov 不等式

若随机变量 $X>0$，则 $\forall \varepsilon>0$，有

$$\mathrm{P}(X \geqslant \varepsilon) \leqslant \frac{\mathrm{E}(X)}{\varepsilon}. \tag{5.1}$$

证明. 引入示性函数 (characteristic function)

$$I(X \geqslant \varepsilon) := \begin{cases} 1, & X \geqslant \varepsilon \\ 0, & X < \varepsilon \end{cases}$$

则 $I(X \geqslant \varepsilon) \leqslant X/\varepsilon$，两边取期望，即证. □

定理 5.1.2: Chebyshev 不等式

若随机变量 X 方差 $\mathrm{Var}(X)$ 存在，则 $\forall \varepsilon>0$

$$\mathrm{P}(|X-\mathrm{E}(X)| \geqslant \varepsilon) \leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2}. \tag{5.2}$$

证明. 由 Markov 不等式

$$\mathrm{P}\left[(X-\mathrm{E}(X))^2 \geqslant \varepsilon^2\right] \leqslant \frac{\mathrm{E}\left[(X-\mathrm{E}(X))^2\right]}{\varepsilon^2} \equiv \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

□

定理 5.1.3: Chernoff 不等式

对随机变量 X，$\forall \varepsilon>0, \forall t>0$，有

$$\mathrm{P}(X \geqslant \varepsilon) \leqslant \frac{\mathrm{E}(\mathrm{e}^{tX})}{\mathrm{e}^{t\varepsilon}}. \tag{5.3}$$

即使 X 矩母函数不存在，不等式也成立.

证明. 由 Markov 不等式，

$$\mathrm{P}(X \geqslant \varepsilon) = \mathrm{P}(\mathrm{e}^{tX} \geqslant \mathrm{e}^{t\varepsilon}) \leqslant \frac{\mathrm{E}(\mathrm{e}^{tX})}{\mathrm{e}^{t\varepsilon}}.$$

□

例 5.1.1: 正态分布的尾概率估计

$X \sim \mathrm{N}(0,1)$，则三个不等式给出

$$\mathrm{P}(|X| \geqslant 3) \leqslant \begin{cases} \dfrac{1}{3}\mathrm{E}(X) = \dfrac{1}{3}\sqrt{\dfrac{2}{\pi}} \doteq 0.27, & \text{(Markov)} \\ \dfrac{1}{9}\mathrm{Var}(X) = \dfrac{1}{9} \doteq 0.11, & \text{(Chebyshev)} \\ \dfrac{2}{\mathrm{e}^{3t}}\,\mathrm{e}^{t^2/2} \leqslant 2\,\mathrm{e}^{-9/2} \doteq 0.022. & \text{(Chernoff)} \end{cases}$$

比较而言，Markov 仅用到了均值的信息，Chebyshev 用到了方差的信息，Chernoff 用到了各阶矩的信息．而结果也是越来越好的．

5.2　大数定律

为何能以某事件发生的频率作为该事件的概率的估计？

定义 5.2.1: 依概率收敛

$X_1, X_2, \ldots$ 为随机变量序列，X 为随机变量，若 $\forall \varepsilon > 0$，均有

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| \geqslant \varepsilon) = 0,$$

则称序列 X_n 依概率收敛 (converge in probability) 于 X，记作

$$X_n \xrightarrow{\mathrm{P}} X. \tag{5.4}$$

定理 5.2.1: Khinchin 弱大数定律

随机变量序列 $X_1, X_2, \ldots$ iid，且 $\mathrm{E}(X_i) = \mu$，则

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mu \tag{5.5}$$

这就是 Khinchin 弱大数定律 (weak law of large numbers, WLLN).

Khinchin LLN 指出：数学期望可以由 n 个 iid 的随机变量的算术平均值近似．最初的大数定律应用于 $X_i \sim \mathrm{B}(p)$ 中用频率 $\overline{X}$ 估计 p，这称为 Bernoulli LLN.

例 5.2.1: 用 Chebyshev 不等式证明 Khinchin 大数定律

若方差 $\mathrm{Var}(X_i) = \sigma^2$ 存在[I]，则

$$\mathrm{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \mu, \tag{5.6}$$

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \tag{5.7}$$

由 Chebyshev 不等式

$$\mathrm{P}(|\overline{X}_n - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

[I]事实上 Khinchin LLN 并没有要求 $\mathrm{Var}(X_i)$ 存在.

可对 Kinchin LLN 进行推广：当条件为 X_i 两两不相关，$\mathrm{Var}(X_i)$ 一致有界时，是 Chebyshev LLN；还有 Markov LLN.

定义 5.2.2: 以概率 1 收敛

若随机变量序列 $X_1, X_2, \ldots$ 极限存在，且

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1, \tag{5.8}$$

则称序列 X_n 以概率 1 收敛 (converge almost surely) 于 X，记作

$$X_n \xrightarrow{\text{a. s.}} X. \tag{5.9}$$

可以证明，以概率 1 收敛是比依概率收敛更强的结论.

定理 5.2.2: Kolmogorov 强大数定律

若 $X_1, X_2, \ldots$ iid，$\mathrm{E}(X_i) = \mu$，则

$$\overline{X}_n \xrightarrow{\text{a. s.}} \mu. \tag{5.10}$$

芝士 Kolmogorov 强大数定律 (~~powerful law of number~~, strong law of large numbers, SLLN).

SLLN 是包含 Khinchin WLLN 的，但是除了 Khinchin，还有其他形式的 WLLN.

SLLN 说明：概率的频率解释是合理的. 也是 Monte Carlo 积分的原理.

例 5.2.2: 以概率 1 收敛强于依概率收敛

$\Omega=[0,1]$ 上的均匀分布，从而有 $(\Omega,\mathcal{F},\mathrm{P})$，定义

$$
\begin{aligned}
&Y_1(\omega)=I_{[0,1]}(\omega),\quad \omega\in[0,1]\\
&Y_2(\omega)=I_{[0,1/2]}(\omega),\ Y_3(\omega)=I_{[1/2,1]}(\omega),\\
&Y_4(\omega)=I_{[0,1/3]}(\omega),\ Y_5(\omega)=I_{[1/3,2/3]}(\omega),\ Y_6(\omega)=I_{[2/3,1]}(\omega),\\
&\cdots
\end{aligned}
$$

显然，Y_n 依概率收敛到 0：

$$\lim_{n\to\infty}\mathrm{P}(|Y_n|\geqslant\varepsilon)=0.$$

但 $\lim\limits_{n\to\infty}Y_n$ 不存在 (总有破坏的区间)，也就不存在以概率 1 收敛.

事实上以概率 1 收敛可以推出依概率收敛，但我就不证了.

5.3　中心极限定理

定理 5.3.1: Lindberg-Levi 中心极限定理

随机变量序列 $X_1,X_2,\ldots$ iid，且期望 μ 和方差 σ^2 存在. $\overline{X}$ 的标准化变量为

$$Z_n:=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\equiv\sum_{i=1}^{n}\frac{X_i-\mu}{\sqrt{n}\sigma}. \tag{5.11}$$

则 Z_n 的 CDF $\mathrm{F}_n(z)$ 极限存在，且

$$\lim_{n\to\infty}\mathrm{F}_n(z)=\Phi(z).\quad\forall z\in\mathbb{R} \tag{5.12}$$

其中 $\Phi(z)$ 是 $\mathrm{N}(0,1)$ 的 CDF，这就是 Lindberg-Levi 中心极限定理 (central limit theorem, CLT).

证明. 标准正态分布的 CF $\psi(t)=\mathrm{e}^{-t^2/2}$，

$$
\begin{aligned}
\psi_{Z_n}(t)&=\left[\psi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n\\
&=\left[1+0+\frac{1}{2}\psi''_{X_i-\mu}(0)\cdot\left(\frac{t}{\sqrt{n}\sigma}\right)^2+o\left(\frac{t^2}{n}\right)\right]^n\\
&=\left[1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right]^n\to\mathrm{e}^{-t^2/2}.
\end{aligned}
$$

□

由 CLT，我们可以近似的认为 $\overline{X}\overset{\cdot}{\sim}\mathrm{N}(\mu,\sigma^2/n)$.

例 5.3.1: DeMoivre-Laplace 中心极限定理

特别地，当 $X_i \sim \mathrm{B}(p)$ 时，$Y := X_1 + \cdots + X_n \sim \mathrm{B}(n,p)$ 二项分布近似于正态分布 $\mathrm{N}\big(np, np(1-p)\big)$

$$\mathrm{P}\left(Y \leqslant t\right) \doteq \Phi\left(\frac{t-np}{\sqrt{np(1-p)}}\right).$$

因为正态分布是连续的，对上式进行连续型修正

$$\mathrm{P}(t_1 \leqslant Y \leqslant t_2) \doteq \Phi\left(\frac{t_2-np+1/2}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{t_1-np-1/2}{\sqrt{np(1-p)}}\right).$$

这称为 DeMoivre-Laplace CLT.

例 5.3.2: 选举问题

为统计选民支持度 p (未知)，随机调查 n 个人，支持比例 $P_n = \overline{X}$，$N \gg n$ 近似的，可视为有放回 $X_i \sim \mathrm{B}(p)$，且

$$\frac{P_n - p}{\sqrt{p(1-p)/n}} \sim \mathrm{N}(0,1)$$

要求精度 $\varepsilon = 0.03$，置信度为 95% ($\alpha = 0.05$)，CLT 给出

$$\mathrm{P}(|P_n - p| \geqslant \varepsilon) \doteq 2\left[1 - \Phi\left(\frac{\varepsilon}{\sqrt{p(1-p)/n}}\right)\right] \leqslant \alpha$$

当 $p = 1/2$ 时 $n_{\min}$ 最大：

$$n_{\min} = \left(\frac{z_{1-\alpha/2}}{2\varepsilon}\right)^2 \doteq 1067.11.$$

与 N 无关.

定理 5.3.2: Lyapunov 中心极限定理

随机变量序列 $X_1, X_2, \ldots$ 是独立的随机变量序列，且具有期望和方差

$$\mathrm{E}(X_k) = \mu_k, \quad \mathrm{Var}(X_k) = \sigma_k^2; \quad B_n^2 := \sum_{k=1}^{n} \sigma_k^2.$$

若存在 $\delta > 0$ 满足 Lyapunov 条件：

$$\lim_{n\to\infty} \frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} \mathrm{E}\big(|X_k - \mu_k|\big) = 0.$$

则随机变量之和的标准化变量

$$\hat{X}_n := \frac{1}{B_n} \sum_{k=1}^{n} (X_k - \mu_k)$$

的 CDF

$$\lim_{n\to\infty} \mathrm{F}_n(x) = \Phi(x), \quad \forall x \in \mathbb{R}.$$

这说明，当 N 很大时，

$$\sum_{i=1}^{n} X_i \dot{\sim} \mathrm{N}\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2\right). \tag{5.13}$$

定理 5.3.3: Markov 中心极限定理

随机变量序列 $X_1, X_2, \ldots$ 不独立，但具有 Markov 性：

$$\mathrm{P}(X_i|X_{i-1}X_{i-2}\cdots) = \mathrm{P}(X_i|X_{i-1}).$$

加上可逆性和可达性的条件，则序列构成一个 Markov 链，则

$$\overline{X}_n \xrightarrow{\mathrm{P}} Z \dot{\sim} \mathrm{N}(\mu, \sigma^2/n) \tag{5.14}$$

注.

1. 尾部概率控制；
2. LLN, CLT；
3. CLT 的应用：$\overline{X} \dot{\sim} \mathrm{N}(\mu, \sigma^2/n)$, $\sum X_i \dot{\sim} \mathrm{N}(n\mu, n\sigma^2)$；
4. 三种收敛：$\xrightarrow{\text{a. s.}}, \xrightarrow{\mathrm{P}}, \xrightarrow{\mathrm{d}}$ (依分布收敛).

第二部分

统计推断

> *It is better to have an approximate answer to the right question than an exact answer to the wrong one.*
>
> -John Tukey

数理统计是从数据到信息的学科

$$\text{数理统计}\begin{cases}\text{数据收集}\\ \text{数据分析}\begin{cases}\text{描述}\\ \text{推断：样本}\rightarrow\text{总体}\end{cases}\end{cases}$$

第六章　参数估计

定义 6.0.1: 总体

总体 (population) 是指与所研究的问题有关的对象 (个体) 的全体的某个数值特征 X 的概率分布.

总体分为有限总体和无限总体.

定义 6.0.2: 统计模型

统计模型 (model) 是一族概率分布.

可以通过若干个参数表达出来的模型是参数模型，否则是非参数模型.

定义 6.0.3: 样本

样本 (sample) 是随机变量序列 $X_1, X_2, \ldots, X_n$，X_i 取自总体，n 称为样本容量.

样本的获取方式：试验、观测.

定义 6.0.4: 简单随机抽样

简单随机抽样 (simple random sampling) 总体个数 N 有限，无放回，任意容量相同的样本都有相同的发生概率.

$$p = 1 \Big/ \binom{N}{n}.$$

定义 6.0.5: 随机样本

$X_1, \ldots, X_n$ iid，比如有放回抽样

不当抽样.

定义 6.0.6: 统计量

统计量 (statistic) 是完全由样本决定的量，比如样本期望 $\overline{X}$ 和样本方差 S^2

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

总体决定样本，用样本来推断总体，这就是统计推断.

6.1　矩估计

随机样本 $X_1, X_2, \dots$ iid

定义 6.1.1: 样本矩

样本 k 阶原点矩

$$a_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{\text{P}} \mathrm{E}(X^k),$$

k 阶中心矩

$$m_k := \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \xrightarrow{\text{P}} \mathrm{E}\big[(X-\mu)^k\big],$$

矩估计就是认为总体矩 $\doteq$ 样本矩，从而得到参数的估计.

例 6.1.1: 正态分布和 Poisson 分布的矩估计

iid $X_i \sim \mathrm{N}(\mu, \sigma^2)$ 用样本均值 $\overline{X}$ 和样本 2 阶中心矩 m_2 估计

$$\begin{cases} \mu = \overline{X}, \\ \sigma^2 = m_2 \end{cases}$$

iid $X_i \sim \mathrm{Exp}(\lambda)$ 的参数 λ 有不止一种估计方法

$$\lambda = \frac{1}{\overline{X}} \text{ or } \lambda = \frac{1}{\sqrt{m_2}}.$$

但是我们采用低阶矩 $\overline{X}$，因为其受噪声的影响更小.

6.2　极大似然估计

假设 $X_1, \dots, X_n$ 的联合分布为

$$f(x_1, \dots, x_n; \theta),$$

其中 θ 可为标量，也可为向量.

定义 6.2.1: 似然函数

对于观测 $(X_1, \dots, X_n)$ 的似然函数 (likelihood function)

$$L(\theta) := f(X_1, \dots, X_n; \theta). \tag{6.1}$$

观测数据通常记为 $x_1, \dots, x_n$，称为随机变量 $X_1, \dots, X_n$ 的一个实现. 离散情形，$L(\theta)$ 即为实现 $X_1, \dots, X_n$ 的概率.

若 $X_1,\dots,X_n$ iid，总体分布为 $g(x;\theta)$，则

$$L(\theta)=g(X_1;\theta)\cdots g(X_n;\theta).$$

定义 6.2.2: 极大似然估计

θ 的极大似然估计 (maximum likelihood estimation, MLE) 为

$$\theta^*:=\arg\max_{\theta} L(\theta). \tag{6.2}$$

例 6.2.1: 正态分布的 MLE

iid $X_i\sim \mathrm{N}(\mu,\sigma^2)$，$\mu,\sigma^2$ 未知，则其似然函数为

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

要求 L 的最大值点，可对 L 求导，然而更方便的做法是求对数似然：

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

最大值点处满足

$$\begin{cases}\dfrac{\partial \ln L}{\partial \mu}=0\\[2mm] \dfrac{\partial \ln L}{\partial (\sigma^2)}=0\end{cases}\quad\Longrightarrow\quad\begin{cases}\mu^*=\overline{X}\\ (\sigma^2)^*=\dfrac{1}{n}\sum(X_i-\overline{X})^2.\end{cases}$$

经验证，$\mu^*,(\sigma^2)^*$ 为所求 MLE.

注. MLE 具有不变性：

$$\big(g(\theta)\big)^*=g(\theta^*). \tag{6.3}$$

例 6.2.2: 均匀分布的 MLE

iid $X_i\sim \mathrm{U}(0,\theta)$，$\theta$ 未知

$$L(\theta)=\begin{cases}\dfrac{1}{\theta^n}, & \theta\geqslant X_i>0\\ 0, & \text{elsewhere}\end{cases}$$

$\theta^*=\max(X_1,\dots,X_n)$

例 6.2.3: Cauchy 分布的估计

估计 Cauchy 分布的参数 θ

$$f(x;\theta)=\frac{1}{\pi}\frac{1}{1+(x-\theta)^2},\quad x\in\mathbb{R}$$

期望不存在，无矩.

似然方程

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\theta}=0,\implies\sum_{i=1}^{n}\frac{X_i-\theta}{1+(X_i-\theta)^2}=0.$$

当 $n>2$ 时很难解．回到参数的意义，θ 可用样本中位数估计．

估计方法不唯一，即使 MLE 也不一定唯一，MLE 需要 f (分布的参数形式)，算法．

6.3　Bayes 估计

矩估计和 MLE 都是点估计，将参数 θ 被视为未知的数 (组)．若对参数 θ 有先验知识，可用随机变量 Θ 的概率分布来刻画，称之为先验分布 (priori distribution)

$$f_\Theta(\theta)$$

θ 是 Θ 的实现值；X 是试验 (观测) 结果，样本分布为

$$f_{X|\Theta}(x|\theta)$$

由 Bayes 公式，后验分布 (posterior distribution) 为

$$f_{\Theta|X}(\theta|X)=\frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}=\frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\int_\theta f_\Theta(\theta)f_{X|\Theta}(x|\theta)\,\mathrm{d}\theta}.\tag{6.4}$$

若 $\Theta\sim\mathrm{U},f_\Theta(\theta)\propto 1$，称为 Bayes 法则，也叫同等无知原则．

例 6.3.1: 二项分布的 Bayes 估计

$X=n$ 次掷币正面向上次数，当 $\Theta=\theta$ 时，$X\sim\mathrm{B}(n,\theta)$

$$f_{X|\Theta}(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad x=0,1,\ldots,n.$$

采用 Bayes 法则，(X,Θ) 的联合分布 PDF 为

$$f_{X,\Theta}(x,\theta)=f_\Theta(\theta)f_{X|\Theta}(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad\theta\in(0,1).$$

所以

$$\begin{aligned}f_X(x)&=\binom{n}{x}\int_0^1\theta^x(1-\theta)^{n-x}\,\mathrm{d}\theta\\&=\binom{n}{x}\mathrm{B}(x+1,n-x+1)=\frac{1}{n+1}.\end{aligned}$$

故后验概率

$$f_{\Theta|X}(\theta|x)=\frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}.$$

故 $\theta\sim\mathrm{Be}(x+1,n-x+1)$，

$$\hat\theta=\mathrm{E}(\Theta|x)=\frac{x+1}{n+2}.$$

或

$$\theta^* = \arg\max f_{\Theta|X}(\theta|x) = \frac{x}{n}.$$

上面取得先验分布为 $\mathrm{U}(0,1) = \mathrm{Be}(1,1)$，若先验分布取为 $\mathrm{Be}(a,b)$，则后验分布为 $\mathrm{Be}(a+x, b+n-x)$

需要注意，这里的先验分布并不是唯一指定的.

6.4　无偏性

定义 6.4.1: 偏差

统计量

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n).$$

是对 θ 的一个估计，定义偏差 (bias)

$$\mathrm{E}(\hat{\theta} - \theta) = \mathrm{E}(\hat{\theta}) - \theta. \tag{6.5}$$

若 $\forall\theta$, $\mathrm{E}(\hat{\theta}) = \theta$，则 $\hat{\theta}$ 是 θ 的一个无偏估计 (unbiased estimation).

$\hat{g}(X_1, \ldots, X_n)$ 是 $g(\theta)$ 的无偏估计当

$$\mathrm{E}(\hat{g}) = g(\theta), \quad \forall\theta.$$

若无偏，由 LLN

$$\frac{1}{N}\sum_{n=1}^{N} \hat{g}(X_1^{(n)}, \ldots, X_n^{(n)}) \xrightarrow{\text{a. s.}} \mathrm{E}(\hat{g}) = g(\theta).$$

例 6.4.1: 样本期望、方差的无偏性

样本均值的期望和方差

$$\mathrm{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \mu, \tag{6.6}$$

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \tag{6.7}$$

样本方差，由定义

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

的均值

$$\begin{aligned}\mathrm{E}(S^2) &= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i^2) - \mathrm{E}(\overline{X}^2)\right] \\ &= \frac{n}{n-1}\left[\sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \sigma^2.\end{aligned} \tag{6.8}$$

例 6.4.2: 均匀分布的无偏估计

在例 6.2.2 中给出了均匀分布 $\mathrm{U}(0,\theta)$ 的 MLE

$$\theta^* = \max(X_1,\dots,X_n).$$

这种估计是有偏的，可以证明

$$\mathrm{E}(\theta^*) = \frac{n}{n+1}\theta.$$

以下四种估计都是无偏的：

$$\begin{aligned}
\hat{\theta}_1 &:= \frac{n+1}{n}\max(X_1,\dots,X_n).\\
\hat{\theta}_2 &:= \max(X_1,\dots,X_n) + \min(X_1,\dots,X_n),\\
\hat{\theta}_3 &:= 2\overline{X},\quad \text{(矩估计)}\\
\hat{\theta}_4 &:= (n+1)\min(X_1,\dots,X_n).
\end{aligned}$$

但其精确度越来越差.

6.5　均方误差准则

定义 6.5.1: 均方误差

定义均方误差 (mean square error, MSE)

$$\mathrm{E}\big[(\hat{\theta}-\theta)^2\big] = \mathrm{Var}(\hat{\theta}) + \mathrm{E}^2(\hat{\theta}-\theta). \tag{6.9}$$

MSE 由两部分构成，$\mathrm{Var}(\hat{\theta})$ 表示 $\hat{\theta}$ 的精确度 (precision)，$\mathrm{E}^2(\hat{\theta}-\theta)$ 代表准确度 (accuracy). 无偏估计后者为 0.

定义 6.5.2: 有效性

假定 $\hat{\theta}_1,\hat{\theta}_2$ 是 θ 的无偏估计，若

$$\mathrm{Var}(\hat{\theta}_1) \leqslant \mathrm{Var}(\hat{\theta}_2),\quad \forall\theta$$

且存在一个分布 (i.e. 一个 θ 值) 使得 $<$ 成立，则称在 MSE 意义下，$\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效.

如果所有无偏估计中存在最小的方差，则称为一致最小方差无偏估计 (uniformly minimum variance unbiased estimator, UMVUE).

定理 6.5.1: Cramer-Rao 下限

任何 θ 的无偏估计量满足

$$\mathrm{Var}(\hat{\theta}) \geqslant \frac{1}{I_n(\theta)}. \tag{6.10}$$

其中 $I_n(\theta)$ 是 Fisher 信息量，见定义 6.6.4

例 6.5.1: 低偏倚换方差

总体 $X \sim \mathrm{N}(\mu, \sigma^2)$，样本方差 S^2 是无偏的，而样本二阶矩 m_2 是有偏的，但 MSE

$$\mathrm{E}\big[(m_2 - \sigma^2)^2\big] < \mathrm{E}\big[(S^2 - \sigma^2)^2\big].$$

6.6　大样本性质

$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ 当样本容量 $n \to \infty$ 时的性质，

定义 6.6.1: 渐进无偏性

$n \to \infty$ 时，

$$\mathrm{E}(\hat{\theta}) \to \theta. \tag{6.11}$$

定义 6.6.2: 相合性

称 $\hat{\theta}$ 为 θ 的 (弱) 相合估计 (consistent estimate) 当

$$\hat{\theta} \xrightarrow{\mathrm{P}} \theta. \tag{6.12}$$

WLLN 指出，$\overline{X}$ 为 μ 的一个相合估计．相合性为良好点估计的自然要求．

例 6.6.1: 二阶中心矩

二阶中心矩 m_2 是总体方差 σ^2 的相合估计．

$$\begin{aligned} m_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2 \xrightarrow{\mathrm{P}} \sigma^2 + 0. \end{aligned}$$

定义 6.6.3: 渐进正态性

若 $\exists \sigma_n > 0$ 使得

1. $\lim\limits_{n\to\infty} \sigma_n = 0$;
2. $\lim\limits_{n\to\infty} \mathrm{P}\left(\dfrac{\hat{\theta} - \theta}{\sigma_n} \leqslant x\right) = \Phi(x)$.

则称 $\hat{\theta}$ 为 θ 的相合近似正态估计．

1. 有时，取 $\sigma_n^2 = \mathrm{Var}(\hat{\theta})$，比如用 $\overline{X}$ 估计 μ;

2. $n \gg 1$ 时，$\hat{\theta} \dot{\sim} \mathrm{N}(\theta, \sigma_n^2)$.；
3. 原式分布 (即 X_i 的分布) f 满足光滑性条件，则 $\exists \sigma_n > 0$ 使得

$$\frac{\theta^* - \theta}{\sigma_n} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{6.13}$$

定义 6.6.4: Fisher 信息量

$X_1, \dots, X_n$ iid，对数似然函数

$$\ell(\theta) := \ln f(X;\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta).$$

定义 Fisher 信息量

$$I_n(\theta) := \mathrm{E}(\ell'^2). \tag{6.14}$$

Fisher 信息量是衡量观测所得的随机变量 X 携带的关于未知参数 θ 的信息量，其中 X 的概率分布依赖于参数 θ.

$\ell(\theta)$ 导数的期望

$$\mathrm{E}(\ell') = \int \frac{1}{f}\frac{\partial f}{\partial \theta} \cdot f \,\mathrm{d}x = \frac{\partial}{\partial \theta}\int f \,\mathrm{d}x \equiv 0.$$

二阶导期望

$$\mathrm{E}(\ell'') = \mathrm{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial \theta^2} - \left(\frac{1}{f}\frac{\partial f}{\partial \theta}\right)^2\right] = 0 - \mathrm{E}(\ell'^2) = -\operatorname{Var}(\ell').$$

Fisher 信息量与样本容量有关：

$$I_n(\theta) = \sum_{i=1}^{n} \mathrm{E}\left[\left(\frac{\partial}{\partial \theta}\ln f(X_i;\theta)\right)^2\right] = nI_1(\theta) =: nI(\theta).$$

例 6.6.2: Fisher 信息量证明式 (6.13)

由 MLE 的性质，$\ell'(\theta^*) = 0$，做 Talyor 展开，

$$0 = \ell'(\theta^*) = \ell'(\theta) + (\theta^* - \theta)\ell''(\theta) + \cdots$$

故

$$\theta^* - \theta \doteq -\frac{\ell'(\theta)}{\ell''(\theta)}.$$

设 $\ell'(\theta) = n\overline{Y}$，

$$Y_i := \frac{\partial}{\partial \theta}\ln f(X_i;\theta), \quad \mathrm{E}(Y_i) = 0, \ \operatorname{Var}(Y_i) = I(\theta).$$

CLT 说明，$\overline{Y}$ 是渐进正态的，故分子

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \sqrt{n}\,\overline{Y} \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, I(\theta)\big).$$

分母

$$\frac{1}{n}\ell''(\theta) \xrightarrow{\mathrm{P}} I(\theta).$$

故 $\theta^* - \theta$ 是渐进正态的：

$$\sqrt{n}(\theta^* - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \frac{1}{I(\theta)}\right).$$

式 (6.13) 中的

$$\sigma_n = \frac{1}{\sqrt{nI(\theta)}} \text{ 或 } \frac{1}{\sqrt{nI(\theta^*)}}$$

Review

1. 参数估计：样本 $X_1, \ldots, X_n$，通常假设 iid
   样本分布 $f(x_1, \ldots, x_n; \theta)$ 即 $X_1, \ldots, X_n$ 的联合分布
   估计 $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ 是统计量，样本的函数
   标准误差：$\hat{\theta}$ 的标准差

   $$se(\hat{\theta}) := \sqrt{\mathrm{Var}(\hat{\theta})}. \tag{6.15}$$

   标准误差的估计 $\hat{se} = \hat{se}(\hat{\theta})$
2. 经典估计的优良性：
   n 固定：无偏性、有效性；
   $n \to \infty$：渐进无偏性、相合性、渐进正态性.
3. Bayes 估计：将 θ 理解成一个随机变量 Θ 的实现值. Θ 刻画了对 θ 的认知

   $$f_\Theta(\theta) \quad \to \quad f_{\Theta|X}(\theta|x),$$

   后验众数 v.s. MLE. Bayes 假设二者相等.

6.7　区间估计

定义 6.7.1: 置信区间

给定 $\alpha \in (0, 1)$，$\forall \theta$ 可能值，有 $\hat{\theta}_1, \hat{\theta}_2$ 使得

$$\mathrm{P}(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geqslant 1 - \alpha. \tag{6.16}$$

这里 $\hat{\theta}_i = \hat{\theta}_i(X_1, \ldots, X_n)$，则称 $(\hat{\theta}_1, \hat{\theta}_2)$ 为 θ 的 $(1 - \alpha)$ 置信的区间估计.

α 通常取 0.05, 0.01, 0.1，可靠性优先.

定义 6.7.2: 枢轴变量

仅含一个待估参数的样本的连续函数，且分布不依赖于未知参数.

方法 6.7.1: 枢轴变量法

1. 确定 $\hat{\theta}$;
2. 找到枢轴变量 $H(\hat{\theta},\theta)$ 的分布可用;
3. 确定 $(\hat{\theta}_1,\hat{\theta}_2)$

例 6.7.1: 正态分布的区间估计

$X\sim \mathrm{N}(\mu,\sigma^2)$ 的 σ^2 已知，μ 未知，由 $\overline{X}\sim \mathrm{N}(\mu,\sigma^2/n)$，取枢轴变量

$$Z:=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim \mathrm{N}(0,1)$$

取对称的上下限

$$\mathrm{P}\left(|Z|\geqslant z_{\alpha/2}\right)=\alpha,\implies \Phi(z_{\alpha/2})=1-\alpha/2$$

其中 $z_{\alpha/2}$ 是 $\mathrm{N}(0,1)$ 的上 $\alpha/2$ 分位数. 故 μ 的 $(1-\alpha)$ 置信区间为

$$\left(\overline{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \tag{6.17}$$

若 μ,σ^2 均未知，需要用 S^2 估计 σ^2，由式 (3.21)

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

卡方分布并不对称，我们选择等尾置信区间，即

$$\mathrm{P}\big(\chi^2<\chi^2_{\alpha/2}(n-1)\big)=\mathrm{P}\big(\chi^2>\chi^2_{1-\alpha/2}(n-1)\big)=\frac{\alpha}{2},$$

σ^2 的 $(1-\alpha)$ 置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right); \tag{6.18}$$

对 μ 的估计，由式 (3.22)

$$t:=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim \mathrm{t}(n-1).$$

可得 μ 的 $(1-\alpha)$ 置信区间为

$$\left(\overline{X}\pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right). \tag{6.19}$$

表 6.1: 区间估计

条件	估计	枢轴量	置信区间
σ^2 已知	μ	$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$\left(\overline{X}\pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$
σ^2 未知	μ	$\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$	$\left(\overline{X}\pm t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}\right)$
μ 未知	σ^2	$\dfrac{(n-1)S^2}{\sigma^2}$	$\left(\dfrac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$

同理可求单侧置信限，如 σ^2 已知时 μ 的上、下单侧置信限分别为

$$\overline{\mu}=\overline{X}+z_\alpha\frac{\sigma}{\sqrt{n}},\quad \underline{\mu}=\overline{X}-z_\alpha\frac{\sigma}{\sqrt{n}}.$$

例 6.7.2: 两个正态总体

$X\sim \mathrm{N}(\mu_1,\sigma^2), Y\sim \mathrm{N}(\mu_2,\sigma^2)$ 独立，μ_1,μ_2,σ^2 未知，估计 $\mu_1-\mu_2$

$$\overline{X}-\mu_1-(\overline{Y}-\mu_2)\sim \mathrm{N}\Big(0,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}\Big);$$
$$\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2).$$

利用 t 分布的定义，构造枢轴变量：

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

其中

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

$\mu_1-\mu_2$ 的 $(1-\alpha)$ 置信区间为

$$\left(\overline{X}-\overline{Y}\pm t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

大样本方法　渐进置信区间

例 6.7.3: 选举问题的置信区间估计

真实支持率 p 未知，$n=1200$，观测比例

$$\frac{684}{1200}\doteq 0.57$$

给出 p 的一个 95% 置信区间.

近似有放回：$X_i\sim \mathrm{B}(p)$ iid，$P_n=\overline{X}$，故

$$\mathrm{E}(P_n)=p,\quad \mathrm{Var}(P_n)=\frac{p(1-p)}{n}$$

由 CLT，

$$\frac{P_n-p}{\sqrt{p(1-p)/n}}\dot{\sim}\mathrm{N}(0,1).$$

由于 p 未知，这个方法不能直接用，需要进行一个好的估计.

I 用 S^2 估计 $\sigma^2 = p(1-p)$，则标准误差

$$\hat{se} = \sqrt{\frac{S^2}{n}} = 0.2475.$$

可认为

$$\frac{P_n - p}{\hat{se}} \dot{\sim} \mathrm{N}(0,1).$$

p 的 95% 置信区间

$$(P_n \pm z_{\alpha/2}\hat{se}) = (0.542, 0598)$$

II 用 m_2 估计 σ^2，

$$\begin{aligned} m_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= P_n(1-P_n)^2 + (1-P_n)(0-P_n)^2 = P_n(1-P_n). \end{aligned}$$

相当于用 P_n 估计 p，估计结果与 I 相同.

III (Naïve) 用 $1/4$ 估计 $p(1-p)$，此时 $\hat{se} = 1/\sqrt{4n}$，估计结果为 $(0.542, 0.599)$.

IV (不具有推广性) 由于

$$|P_n - p| = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

可以解出 $p_\pm$，区间 $(p_-, p_+) = (0.542, 0598)$.

例 6.7.4: 相合估计

上例，极大似然估计 $P^* = P_n$，由 MLE 的渐进正态性

$$\frac{P^* - P}{\sigma_n^2} \dot{\sim} \mathrm{N}(0,1)$$

$f(X_i; p) = p^{X_i}(1-p)^{1-X_i}$，故

$$\frac{\partial}{\partial p}\ln f(X_i; p) = \frac{X_i}{p} - \frac{1-X_i}{1-p} = \frac{X_i - p}{p(1-p)}.$$

Fisher 信息量

$$I(p) = \mathrm{E}\left[\left(\frac{X_i - p}{p(1-p)}\right)^2\right] = \frac{1}{p(1-p)}.$$

则

$$\frac{P^* - p}{1/\sqrt{nI(p)}} \dot{\sim} \mathrm{N}(0,1).$$

可以用 $I(P^*)$ 估计 $I(p)$，即

$$\frac{P_n - p}{\sqrt{P_n(1-P_n)/n}} \dot{\sim} \mathrm{N}(0,1)$$

例 6.7.5: 两个正态总体的相合估计

$X \sim \mathrm{N}(\mu_1, \sigma_1^2), Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ 独立，自然

$$\frac{\overline{X} - \mu_1 - (\overline{Y} - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1).$$

可估计

$$\frac{\overline{X} - \mu_1 - (\overline{Y} - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} \mathrm{N}(0,1).$$

6.8　Beyas 区间估计

有了 Θ 的后验分布 $f_{\Theta|X}(\theta|x)$

$$\mathrm{P}(a < \Theta < b|x) = 1 - \alpha.$$

最大后验区间满足

$$f_{\Theta|X}(a|x) = f_{\Theta|X}(b|x).$$

等尾可信区间满足

$$\mathrm{P}(\Theta < a|x) = \mathrm{P}(\Theta > b|x) = \frac{\alpha}{2}.$$

Review

1. 不管是连续还是离散分布，它们的参数都是连续的.
2. 置信区间 v.s. Bayes 区间 (可信区间)

$$\mathrm{P}(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geqslant 1 - \alpha, \quad \hat{\theta}_i = \hat{\theta}_i(X_1, \ldots, X_n)\text{——统计量}$$

$(\hat{\theta}_1, \hat{\theta}_2)$——随机区间. 当得到样本具体观测值 $(x_1, \ldots, x_n)$ 代入 $\hat{\theta}_1, \hat{\theta}_2$ 得到具体区间.

$$\mathrm{P}(a < \Theta < b|x) = 1 - \alpha.$$

后验分布

3. 小样本方法 v.s. 大样本方法：精确分布 v.s. 近似分布

第七章　假设检验

例 7.0.1: 零件检测

一大批电子元件寿命 X，样本 $X_1, \ldots, X_n$

1. 若假设 $X \sim \mathrm{Exp}(\lambda)$，那么 $\lambda = ?$
   回答：参数估计
2. 若合格标准为 $\mathrm{E}(X) \geqslant 5000$，那么如何判断这一批是否合格？
   尝试回答：建立执行标准，比如样本 $\overline{X} > \ell$，那么 $\ell = ?$

参数估计是对参数一无所知；假设检验是对参数有所了解，但有怀疑猜测需要证实.

7.1　基本概念

定义 7.1.1: 统计假设

假设 (hypothesis) 是对一个或多个总体的某种推断或猜测.

- 原假设 (null hypothesis) H_0：被检验的假设；
- 备择假设 (alternative ~) H_1：拒绝 H_0 后可供选择的假设.

简单假设：只对应一个总体；复合假设：对应多个总体

若假设可表示为参数形式，则

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$$

其中 $\Theta_0 \cap \Theta_1 = \varnothing$, $\Theta_0 \cup \Theta_1 = \{\theta \text{ 的所有可能取值 (应用角度)}\}$.

例 7.1.1

接例 7.0.1，

$$H_0 : \mu \geqslant \mu_0, \quad H_1 : \mu < \mu_0,$$

H_0 是复合假设，单侧检验 (看 H_1).

$X \sim \mathrm{N}(\mu, \sigma_0^2)$，

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

H_0 是简单假设 (σ_0^2 已知的情况下)，双侧检验 (看 H_1).

H_0 往往是受保护的，无充分证据不能拒绝；H_1 往往是真正感兴趣的.

定理 7.1.1: 小概率原理

在假设 H_0 为真的情况下，所观测的样本出现的概率如果很小，意味着样本提供的证据拒绝 H_0.

在统计学中，小概率又叫**显著性水平**，因此，假设检验又称为**显著性检验**. 一般来说，我们都希望拒绝 H_0 得到所谓的显著性差异.

定义 7.1.2: 假设检验

检验 (准则)：做出决策的一个具体法则.
R：临界域 (或拒绝域，critical region)，形式上可抽象为

$$R = \{(X_1, \ldots, X_n) \,|\, T(X_1, \ldots, X_n) \geqslant c\}.$$

c 称为临界值.

定义 7.1.3: 两类错误

I 弃真错误：H_0 为真时拒绝 H_0，概率

$$\alpha(R) := \mathrm{P}_\theta\big((X_1, \ldots, X_n) \in R\big), \quad \theta \in \Theta_0 \tag{7.1}$$

II 取伪错误：H_0 为假时接受 H_0，概率

$$\beta(R) := \mathrm{P}_\theta\big((X_1, \ldots, X_n) \notin R\big), \quad \theta \in \Theta_1 \tag{7.2}$$

依据样本决策，错误不可避免. 当 n 固定时，α, β 不能同时变小.

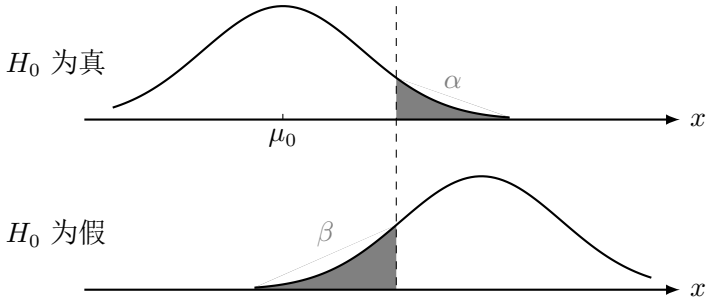


图 7.1: I 类错误和 II 类错误示意图

定义 7.1.4: 功效函数

定义功效函数 (power function) 是样本落在 R 上的概率

$$\mathrm{P}_\theta\big((X_1, \ldots, X_n) \in R\big) = \begin{cases} \alpha(R), & \theta \in \Theta_0 \\ 1 - \beta(R), & \theta \in \Theta_1 \end{cases} \tag{7.3}$$

功效指在 H_1 成立时拒绝 H_0 的概率，即 $1-\beta(R)$，越大越好.

定理 7.1.2: Neyman-Pearson 范式

(n 固定) 预先给定检验水平 $\alpha>0$，控制

$$\alpha(R)\leqslant\alpha,\ \forall\theta\in\Theta_0,$$

再在这个限制下使 $\beta(R)$ 尽可能小.

α 固定时，使 $\beta(R)$ 最小的检验称为水平 α 下的**一致最优检验** (不一定存在，即使存在一般也不易求)

7.2　临界值检验法

方法 7.2.1: 临界值检验法

1. 提出假设 H_0, H_1
2. 给定 $\alpha>0$
3. 选择检验统计量，确定拒绝域形状 (由 H_1 决定)
4. 构建检验：$\alpha(R)\leqslant\alpha\rightsquigarrow R$
5. 采样，计算检验统计量的值
6. 代入检验准则，进行决策

这个过程缺少对于功效 $1-\beta(R)$ 的讨论.

例 7.2.1: Z 检验法和 t 检验法

$X\sim\mathrm{N}(\mu,\sigma^2)$，$\sigma^2$ 已知，

$$H_0:\mu=\mu_0,\quad H_1:\mu\neq\mu_0,$$

当 H_0 为真时，由式 (3.20)，选择统计量

$$Z:=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim\mathrm{N}(0,1),$$

显著性水平 $\alpha>0$ 给定，检验的拒绝域为

$$C=\left\{Z\,\middle|\,|Z|\geqslant z_{\alpha/2}\right\}.$$

这被称为 Z 检验法.

当 σ^2 未知时，就需要选择统计量

$$t:=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim\mathrm{t}(n-1),$$

相应的，检验的拒绝域为

$$C = \left\{ Z \,\middle|\, |t| \geqslant t_{\alpha/2}(n-1) \right\}.$$

这被称为 t 检验法.

表 7.1: 假设检验

条件	H_0	统计量	拒绝域
σ^2 已知	$\mu = \mu_0$	$Z := \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$	$\lvert Z\rvert \geqslant z_{\alpha/2}$
σ^2 未知	$\mu = \mu_0$	$t := \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim \mathrm{t}(n-1)$	$\lvert t\rvert \geqslant t_{\alpha/2}(n-1)$
μ 未知	$\sigma^2 = \sigma_0^2$	$\chi^2 := \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2 \geqslant \chi^2_{\alpha/2}(n-1)$ 或 $\chi^2 \leqslant \chi^2_{1-\alpha/2}(n-1)$

当 H_0 形如 $\mu \geqslant \mu_0$ 时，进行单侧检验.

Review

1. H_0 v.s. H_1，二者天然不对等
   若认可某一组样本，则用它来证实或证伪某理论 (推测)；
2. 决策：拒绝 H_0 或不拒绝 H_0
   检验 (准则) = 决策准则，拒绝域 R 的划分
3. 统计学中没有绝对的证实或证伪，
   α, β 属于检验程序的属性，而非样本的属性.
   $\alpha(R) \leqslant \alpha, \beta(R) \leqslant \beta$ 预先指定的可接受的长期错误率.
4. 功效函数 v.s. 功效，
   不拒绝 v.s. 接受：$\beta(R)$ 小 (功效大)，当样本支持 H_0，才能接受 H_0；通常人们忽略了对 II 类错误的系统控制，将导致对结果意义以及下一步工作方向的误判.
5. 临界值检验法. 缺少对功效的讨论

7.3　临界值检验与置信区间的对偶关系

例 7.3.1: 正态分布的临界值检验与置信区间

$X \sim \mathrm{N}(\mu, \sigma^2)$，$\sigma^2$ 已知，$\alpha \in (0,1)$ 给定，$X_1, \ldots, X_n$ iid
(双侧) 置信区间

$$\mu \in \left(\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

假设检验：考虑

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

检验：接受域

$$R^{\mathrm{c}}=\left\{(X_1,\ldots,X_n)\,\middle|\,|\overline{X}-\mu_0|<z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}.$$

定理 7.3.1: 临界值检验与置信区间的对偶关系

置信区间包含 μ_0

$\Updownarrow$

用 $\overline{X}$ 做检验统计量，建设检验 $H_0:\mu=\mu_0$ $H_1:\mu\neq\mu_0$ 不拒绝 H_0

注. 区间估计的信息更丰富，可作为证据强弱的体现.

7.4　P 值检验法

例 7.4.1

$X\sim\mathrm{N}(\mu,\sigma^2)$，$\sigma^2=25$，考虑

$$H_0:\mu=10,\quad H_1:\mu\neq 10,$$

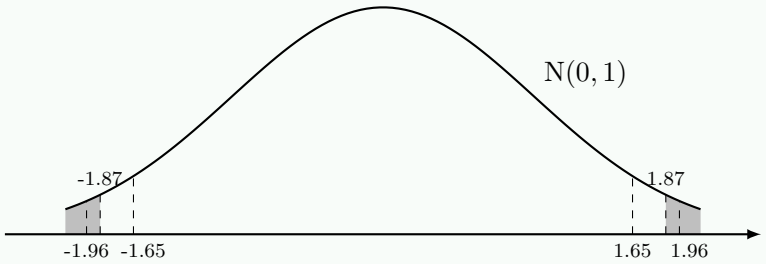
$n=100$，观测到的均值 $\overline{x}=10.935$
检验：若给定 $\alpha=0.05$，则

$$|\overline{x}-10|=0.935<1.96\times\frac{5}{\sqrt{100}}\implies\text{不拒绝 }H_0;$$

若给定 $\alpha=0.1$，则

$$|\overline{x}-10|=0.935>1.65\times\frac{5}{\sqrt{100}}\implies\text{拒绝 }H_0.$$

$$\mathrm{P}(|\overline{X}-10|\geqslant|\overline{x}-10|)=\mathrm{P}(|Z|\geqslant 1.87)=0.0614.$$



定义 7.4.1: 检验 P 值

定义当 H_0 为真时，检验统计量的观测值以及更极端的观测出现的概率称为检验的 P 值.

当 P 值 $\leqslant\alpha$ 时拒绝 H_0，也称为观测值显著.

1. P 值 $=\mathrm{P}($观测值及更极端的观测$|H_0) \neq \mathrm{P}(H_0|$观测值$)$
2. 若 P 值不小，则不拒绝 H_0，原因可能为：(a) H_0 为真；(b) H_0 为假，但检验功效低.

方法 7.4.1: P 值检验法

1. 提出假设 H_0, H_1
2. 给定 $\alpha > 0$
3. 选择检验统计量，确定“极端”的形式 (由 H_1 决定)
4. 采样，计算检验统计量的值
5. 计算 P 值
6. 比较 P 值与 α，进行决策

例 7.4.2: 选举问题

$n = 1200$，调查到的支持比例为

$$\frac{684}{1200} \doteq 0.57 =: p_n,$$

考虑

$$H_0 : p = p_0, \quad H_1 : p > p_0.$$

检验统计量 P_n，由 CLT

$$\frac{P_n - p}{\sqrt{p(1-p)/n}} \dot{\sim} \mathrm{N}(0,1)$$

当 H_0 为真时，

$$Z := \frac{P_n - p_0}{se(P_0)} \dot{\sim} \mathrm{N}(0,1), \quad se(P_n) = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

则 P 值

$$\mathrm{P}_{p=p_0}(P_n \geqslant p_n) \doteq \mathrm{P}(Z \geqslant z_0), \quad z_0 := \frac{p_n - p_0}{se(P_n)}$$

若 $p_0 = 0.55$，则 P 值 $= 0.081$；若 $p_0 = 0.5$，则 P 值 $\ll 0.01$.

考虑复合假设

$$H_0 : p \leqslant p_0, \quad H_1 : p > p_0.$$

可以取

$$\hat{se}(P_n) = \sqrt{\frac{p_n(1-p_n)}{n}} = 0.014$$

当原假设为真时

$$\mathrm{P}_{p\leqslant p_0}(P_n \geqslant p_n) \doteq \mathrm{P}_{p\leqslant p_0}\left(Z \geqslant \frac{p_n - p}{\hat{se}(P_n)}\right)$$

若拒绝 $H_0 : \theta \in \Theta_0 \iff T(X_1, \ldots, X_n) \geqslant c$，则

$$\text{检验的 } P \text{ 值} := \sup_{\theta\in\Theta_0} \mathrm{P}_\theta\Big[T(X_1, \ldots, X_n) \geqslant T(x_1, \ldots, x_n)\Big]$$

故

$$P\text{ 值}=\sup_{p\leqslant p_0}\mathrm{P}\left(Z\geqslant\frac{p_n-p}{\hat{se}(P_n)}\right)=\mathrm{P}\left(Z\geqslant\frac{p_n-p_0}{\hat{se}(P_n)}\right).$$

同 $p=p_0$ 的情况.

7.5　Bayes 假设检验

例 7.5.1

掷 10 次硬币，观测到正面朝上 x 次

$$H_0:p=0.5,\quad H_1:p=0.7,$$

则

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)}=\frac{\mathrm{P}(H_0)\,\mathrm{P}(x|H_0)}{\mathrm{P}(H_1)\,\mathrm{P}(x|H_1)}<1\ (\text{或}\ll 1)\tag{7.4}$$

时拒绝 H_0

若 $H_0:\theta=\theta_0$，Θ 连续，则 $\mathrm{P}(H_0|x)=\mathrm{P}(\Theta=\theta_0|x)\equiv 0$，此时需要技术性处理，可参考陈 5.2.8

Review

1. Neyman-Pearson 假设检验：临界值、P 值检验法 Bayes 假设检验，主观解释概率，引入随机变量认知参数
2. 临界值：拒绝域形状；P 值：更极端形式

7.6　拟合优度检验

定理 7.6.1: χ^2 检验法

设总体 X 的分布未知，检验假设

$$H_0:\text{总体分布函数为 }F(x)$$
$$H_1:\text{总体分布函数不是 }F(x)$$

将在 H_0 下 X 可能取值的全体 Ω 分成互不相交的子集 $A_1,\dots,A_k$，$p_i:=\mathrm{P}(A_i)$. 观测到样本值 $x_1,\dots,x_n$ 落在 A_i 的频数为 f_i，相应的期望频数为 np_i，Pearson 证明：

$$\chi^2:=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}\dot{\sim}\chi^2(k-1).\tag{7.5}$$

拒绝域 $\chi^2\geqslant\chi^2_\alpha(k-1)$.

推导过程见 Pearson's chi-squared test．应用中需要 $n \geqslant 50$，$np_i \geqslant 5$，才能较好运用这个定理．

例 7.6.1: 均匀的骰子

点数	1	2	3	4	5	6	total
观测频数 f_i	4	6	17	16	8	9	60

检验假设

$$H_0 : \text{均匀}(p_1 = p_2 = \cdots = p_6), \quad H_1 : \text{不均匀}$$

Pearson χ^2 统计量的观测值

$$\chi^2 := \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = 14.2.$$

检验 P 值 $= \mathrm{P}_{H_0}(\chi^2(5) \geqslant \chi^2) = 0.014$

连续情况，划分区间，分别计算参数 θ 的 MLE θ^*，卡方统计量分布近似 $\chi^2(k-1-r)$，其中 $r = \dim(\theta)$．但分别计算 MLE 非常麻烦．

实践中先计算 θ 的 MLE θ^*，再划分区间，计算 p_i，卡方统计量分布不是近似 $\chi^2(k-1-r)$，但是检验 P 值介于分布 $\chi^2(k-1-r)$ 和 $\chi^2(k-1)$ 算得的 P 值之间．

事实：独立的卡方统计量可以合并，自由度相应的相加．

例 7.6.2: Mendel 豌豆试验

Mendel 的试验全部独立 (不同作物组)，Fisher 计算其每个卡方统计量并且合并，得到的卡方统计量 $\chi^2 = 41.6056$，自由度为 84，P 值

$$\mathrm{P}(\chi^2(84) \geqslant \chi^2) = 0.99993$$

were far too perfect，Fisher 对这个 extraordinary result 没有发表任何评论，只是指出 *the bias seems to pervade the whole of the data.*

7.7　列联表检验

列联表 (conttingency table) 是一种按两个属性作双向分类的表．

表 7.2: $a \times b$ 列联表

	A_1	$\cdots$	A_a	total
B_1	n_{11}	$\cdots$	n_{a1}	$n_{\cdot 1}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_b	n_{1b}	$\cdots$	n_{ab}	$n_{\cdot b}$
total	$n_{1\cdot}$	$\cdots$	$n_{a\cdot}$	n

例 7.7.1: 独立性检验

样本写在列联表中

$$H_0:\ 独立\ (p_{ij}=p_{i\cdot}p_{\cdot j})\quad H_1:\ 不独立$$

在 H_0 为真的前提下，估计 p_{ij}，MLE 解得

$$p_{ij}^*=p_{i\cdot}^*p_{\cdot j}^*=\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}.$$

卡方统计量

$$\chi^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(nn_{ij}-n_{i\cdot}n_{\cdot j})^2}{nn_{i\cdot}n_{\cdot j}}. \tag{7.6}$$

自由度 $=ab-1-(a-1+b-1)=(a-1)(b-1)$.

特别地，当 $a=b=2$ 时，自由度为 1

$$\chi^2=\frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}.$$

例 7.7.2: 齐性检验

Jane Austen 作家，仿写[I]

表 7.3: Jane Austen 作品单词频数

	A	B	C	total	c
a	147	186	101	434	83
an	25	26	11	62	29
this	32	39	15	86	15
that	94	105	37	236	22
with	59	74	28	161	43
without	18	10	10	38	4
total	375	440	202	1017	196

(1) 检验 Austen 不同作品中单词用法的一致性

$$H_0:\ 具有一致性\ (p_{i1}=p_{i2}=p_{i3})$$

在 H_0 为真的前提下，$p_1^*=434/1017$ 等，卡方统计量观测值

$$\chi^2=12.27,$$

自由度为 10，P 值 ~ 0.3，故不拒绝 H_0.

(2) 检验仿写者与 Austen 不同单词的用法的一致性

$$H_0:\ 具有一致性\ (p_{i\cdot}=p_{ic})$$

可得 $\chi^2 = 32.81$，自由度 5，P 值 $< 10^{-3}$，拒绝 H_0.

[I] A: Sense and Sensibility. B: Emma. C: Sadition (unfinished). c: Sadition (imitation)

7.8　似然比检验

例 7.8.1

接例 7.5.1，式 (7.4) 中，考虑

$$\underbrace{\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)}}_{\text{后验比}} = \underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{先验比}} \underbrace{\frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)}}_{\text{似然比}} < 1 \ (\text{或} \ll 1)$$

时拒绝 H_0；似然比检验就是当

$$\frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)} \leqslant c$$

时拒绝 H_0.

可以证明，当 H_0, H_1 均为简单假设时，似然比检验是最优的.

当 H_0, H_1 不是简单假设时，

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$$

得到随机样本 $X_1, \ldots, X_n$，定义广义似然比

$$\Lambda^* := \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta_1} L(\theta)}.$$

基于技术原因，检验统计量为

$$\Lambda := \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta_0\cup\Theta_1} L(\theta)} \equiv \frac{\sup_{\theta\in\Theta_0} L(\theta)}{L(\theta^*)}. \tag{7.7}$$

易得，当 $\theta^* \in \Theta_0$ 时，$\Lambda = 1 > \Lambda^*$，当 $\theta^* \in \Theta_1 - \Theta_0$ 时，$\Lambda = \Lambda^*$，即

$$\Lambda = \min(\Lambda^*, 1).$$

Λ 越小则样本越不支持 H_0.

方法 7.8.1: 临界值检验

选择 λ_0 使得 $\mathrm{P}(\Lambda \leqslant \lambda_0|H_0) \leqslant \alpha$，若 H_0 为真时 Λ 的分布方便求，则可直接计算 $\mathrm{P}(\Lambda \leqslant \lambda_0|H_0)$.

定理 7.8.1

在一定光滑性的条件下，在 H_0 为真的前提下，

$$-2\ln\Lambda \xrightarrow{\mathrm{d}} \chi^2(d). \tag{7.8}$$

其自由度 $d=\dim(\Theta_0\cup\Theta_1)-\dim(\Theta_0)$，dim 表示自由参数的个数.

例 7.8.2: 多项分布的似然比估计

多项分布

$$H_0: p_1=p_1^0,\dots,p_k=p_k^0,\ \ (p_1^0+\cdots+p_k^0=1)$$

观测频数 $n_1,\dots,n_k$，$n_1+\cdots+n_k=n$，似然函数

$$L(p_1,\dots,p_k)=\binom{n}{n_1,\dots,n_k}p_1^{n_1}\cdots p_k^{n_k},$$

故

$$\Lambda=\frac{L(p_1^0,\dots,p_k^0)}{L(p_1^*,\dots,p_k^*)},\quad p_i^*=\frac{n_i}{n}.$$

则

$$\begin{aligned}-2\ln\Lambda&=-2\sum_{i=1}^k n_i\ln\left(\frac{p_i^0}{p_i^*}\right)=2\sum_{i=1}^k n_i\ln\left(\frac{n_i}{np_i^0}\right)\\&=2\cancel{\sum_{i=1}^k(n_i-np_i^0)}+\sum_{i=1}^k\frac{(n_i-np_i^0)^2}{np_i^0}+\cdots\end{aligned}$$

$\dim(\Theta_0)=0,\ \ \dim(\Theta_0\cup\Theta_1)=k-1$

7.9　两总体的比较

两独立总体 X,Y 比较，均值方差分别为 $\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$

例 7.9.1: 比较成功率

阿司匹林对于降低心脏病发病率的有效性 (历时 5 年)，样本信息：

表 7.4: 阿司匹林

	心脏病发作	未发作	合计	发作率
安慰剂	239	10795	11034	0.0217
阿司匹林	129	10898	11037	0.0126

安慰剂和阿司匹林组的发作率为 p_1,p_2，假设检验

$$H_0: p_1=p_2\,(\text{无效}),\quad H_1: p_1>p_2\,(\text{有效})$$

故

$$\mathrm{E}(P_1-P_2)=p_1-p_2,\ \ \mathrm{Var}(P_1-P_2)=\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}.$$

且

$$\frac{(P_1-P_2)-(p_1-p_2)}{\hat{se}}\dot{\sim}\,\mathrm{N}(0,1),$$

在 H_0 为真的前提下，$p_1 = p_2 =: p$，

$$p^* = \frac{k_1 + k_2}{n_1 + n_2},\ \hat{se} = \sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.00175$$

检验 P 值 $= \mathrm{P}(Z \geqslant 5.20) \ll 10^{-3}$，拒绝 H_0.

注. 随机分组，双盲试验，n 充分大.

7.10　显著性思考

假设检验不能解释原因，不能检验试验设计，试验者需要对样本负责.

统计显著 $\neq$ 实际显著 (陈 P236)

第八章　方差分析和回归分析

8.1　方差分析

方差分析 (Analysis of Variance, AoV or ANOVA) 由著名英国统计学家 R. A. Fisher 在 1923 年提出，又称 F 检验，用于两个及两个以上样本均数差别的显著性检验.

定义 8.1.1

试验指标：要考察的指标；
因素：影响试验指标的条件；
因素的水平：因素所处的状态；
单因素试验：在一项试验中只有一个因素在改变，
多因素试验：多于一个因素在改变.

基本假设：

1. 样本相互对立，来自正态总体
2. 各总体方差相等

变差分解　组数 s，总例数 $N=\sum_{j=1}^{s} n_j$

定义 8.1.2: 总变差

总变差的大小可用总偏差平方和表示，这反映了所有观测值之间总的变差程度.

$$\begin{aligned} S_T &:= \sum_{j=1}^{s}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2 \\ &= \sum_{i,j} X_{ij}^2 - \frac{1}{N}\left(\sum_{i,j} X_{ij}\right)^2, \end{aligned} \tag{8.1}$$

总自由度 $\nu_T = N-1$，总均方 $MS_T = S_T/\nu_T$.

变异程度除与偏差平方和 S 有关外，还与其自由度 ν 有关，由于各部分自由度不相等，因此各部分偏差平方和不能直接比较，须将各部分偏差平方和除以相应自由度，其比值称为均方差，简称均方 (mean square, MS).

定义 8.1.3: 组间变差

各处理组的样本均值大小不等, 这种变差称为组间变差

$$S_A := \sum_{j=1}^{s} n_j(\overline{X}_{\cdot j} - \overline{X})^2 \tag{8.2}$$
$$= \sum_{j=1}^{s} \frac{1}{n_j}\left(\sum_{i=1}^{n_j} X_{ij}\right)^2 - \frac{1}{N}\left(\sum_{i,j} X_{ij}\right)^2,$$

组间自由度 $\nu_A = s-1$，组间均方 $MS_A = S_A/\nu_A$.

组间变差存在的原因

- 随机误差，包括个体变差和测量误差
- 处理因素的不同水平可能对试验结果有影响.

定义 8.1.4: 组内变异

在同一处理组内，虽然各受试对象接受的处理相同，但测量值之间仍不同，这种变异称为组内变异 (误差)

$$S_E := \sum_{i,j}(X_{ij} - \overline{X}_{\cdot j})^2 \tag{8.3}$$

反映了随机误差的影响.
组内自由度 $\nu_E = N - s$，组内均方 $MS_E = S_E/\nu_E$.

偏差平方和与自由度具有可加性

$$S_T = S_A + S_E,$$
$$\nu_T = \nu_A + \nu_E.$$

假设检验

$$H_0 : \mu_1 = \mu_2 = \cdots, \quad H_1 : \ldots$$

检验统计量

$$F := \frac{S_A}{S_E} \sim \mathrm{F}(\nu_A, \nu_E). \tag{8.4}$$

当 $F > F_\alpha(\nu_A, \nu_E)$ 时拒绝 H_0.

双因素检验　设有两个因素 A, B 作用于试验的指标. 因素 A 有 r 个水平 $A_1, A_2, \ldots, A_r$，因素 B 有 s 个水平 $B_1, B_2, \ldots, B_s$. 现对因素 A, B 的所有组合 (A_i, B_j) 都作 t 次试验，称为等重复试验.

设 $X_{ijk} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$，$i, j, k$ 分别为因素 A, B 和重复试验的下标. 则总平均

$$\mu := \frac{1}{rs}\sum_{i,j}\mu_{ij},$$

A_i, B_j 的效应分别为

$$\alpha_i = \mu_{i\cdot} - \mu, \quad \beta_j = \mu_{\cdot j} - \mu,$$

交互效应

$$\gamma_{ij} = \mu_{ij} - \mu_{\cdot j} - \mu_{i\cdot} + \mu.$$

可见

$$\sum_{i=1}^{r} \alpha_i = \sum_{i=1}^{r} \gamma_i, \quad \sum_{j=1}^{s} \beta_j = \sum_{j=1}^{s} \gamma_j.$$

- S_T 为总变差，自由度 $rst - 1$
- S_E 误差平方和，自由度 $rs(t-1)$
- S_A, S_B 为因素 A, B 的效应平方和，自由度为 $r-1, s-1$
- $S_{A\times B}$ 为 A, B 的交互效应平方和，自由度为 $(r-1)(s-1)$

$$\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk} - \overline{X})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}[(X_{ijk} - \overline{X}_{ij\cdot}) + (\overline{X}_{i\cdot\cdot} - \overline{X}) + (\overline{X}_{\cdot j\cdot} - \overline{X}) + \\
&\qquad (\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X})]^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk} - \overline{X}_{ij\cdot})^2 + st\sum_{i=1}^{r}(\overline{X}_{i\cdot\cdot} - \overline{X})^2 + \\
&\qquad rt\sum_{j=1}^{s}(\overline{X}_{\cdot j\cdot} - \overline{X})^2 + t\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X})^2 \\
&= S_E + S_A + S_B + S_{A\times B}.
\end{aligned} \tag{8.5}$$

比如，对 A 做假设检验

$$H_0 : \alpha_1 = \cdots = \alpha_r = 0, \quad H_1 : \ldots$$

检验统计量

$$F := \frac{S_A}{r-1} \Big/ \frac{S_E}{rs(t-1)} \sim \mathrm{F}\big(r-1, rs(t-1)\big).$$

8.2　回归分析

用 $g(X)$ 估计 Y，离差 $Y - g(X)$，均方误差就是离差平方的期望

$$Q := \mathrm{E}\big[(Y - g(X))^2\big].$$

最小二乘法所求的是均方误差意义下的最优线性预测

$$(\alpha, \beta) = \arg\min_{a,b} \mathrm{E}\big[(Y - (a + bX))^2\big]$$

实际上就是求二元函数

$$\begin{aligned}
f(a,b) &:= \mathrm{E}\big[(Y - (a+bX))^2\big] \\
&= \mathrm{E}(X^2)b^2 + 2\,\mathrm{E}(X)ab + a^2 - 2\,\mathrm{E}(XY)b - 2\,\mathrm{E}(Y)a + \mathrm{E}(Y^2).
\end{aligned}$$

的最小值点，则在最小值点处有

$$\frac{\partial f}{\partial a}=0,\ \ \frac{\partial f}{\partial b}=0.$$

解得

$$\begin{cases}\beta=\dfrac{\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)}{\mathrm{E}(X^2)-\mathrm{E}^2(X)}\equiv\dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)},\\ \alpha=\mathrm{E}(Y)-\beta\,\mathrm{E}(X)\end{cases}\tag{8.6}$$

误差的期望和方差

$$\begin{aligned}\mathrm{E}[Y-(\alpha+\beta X)]&=\mathrm{E}(Y)-\alpha-\beta\,\mathrm{E}(X)=0;\\ \mathrm{Var}[Y-(\alpha+\beta X)]&=\mathrm{Var}(Y-\beta X)\\ &=\mathrm{Var}(Y)+\beta^2\,\mathrm{Var}(X)-2\beta\,\mathrm{Cov}(Y,X)\\ &=\left[1-\mathrm{Corr}^2(X,Y)\right]\mathrm{Var}(Y).\end{aligned}$$

相关性检验　对任意两个变量的一组观察值 (x_i,y_i) 都可以用最小二乘法形式上求得 y 对 x 的回归方程，如果 y 与 x 没有线性相关关系, 这种形式的回归方程就没有意义．因此需要考察 y 与 x 间是否确有线性相关关系，这就是回归效果的检验问题．

一元线性回归模型

$$y=a+bx+\epsilon,\ \ \epsilon\sim\mathrm{N}(0,\sigma^2).$$

因随机因素引起的误差称为残差平方和

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

则

$$\frac{Q_e}{\sigma^2}\sim\chi^2(n-2),\implies\hat{\sigma}^2=\frac{Q_e}{n-2},$$

检验假设

$$H_0:b=0,\quad H_1:b\neq0,$$

$\hat{b}$ 满足

$$\hat{b}\sim\mathrm{N}\left(b,\frac{\sigma^2}{L_{xx}}\right),\quad L_{xx}:=\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

如果原假设成立，

$$\frac{\hat{b}\sqrt{L_{xx}}}{\sigma}\sim\mathrm{N}(0,1),\quad\frac{Q_e}{\sigma^2}\sim\chi^2(n-2).$$

检验统计量

$$t:=\hat{b}\sqrt{\frac{(n-2)L_{xx}}{Q_e}}\sim\mathrm{t}(n-2).\tag{8.7}$$

拒绝域 $|t|\geqslant t_{\alpha/2}(n-2)$.

线性化回归　在许多实际问题中，两个变量之间并不一定是线性关系，而是某种曲线关系，应该用曲线来拟合．可以进行适当的变量代换，把它线性化，这样就把一个非线性回归问题化为线性回归问题而得以解决．

多线性回归　随机变量 Y 往往与多个变量 $x1, x2, \ldots, xp$ 有关．对自变量 $x1, x2, \ldots, xp$ 的一组确定的值，Y 有它的分布

$$Y = b_0 + \sum_{i=1}^{p} b_i x_i + \epsilon, \ \ \epsilon \sim \mathrm{N}(0, \sigma^2).$$

第九章　随机抽样

9.1　Monte Carlo 方法

Monte Carlo 方法是用随机数解决计算问题方法的统称.

定理 9.1.1: von Neumann 舍选法

1. 选取容易抽样的分布 $h(x)$，使得曲线 $Ch(x)$ 可以覆盖待抽样的分布 $f(x)$；
2. 从 $h(x)$ 抽样得 x_i；
3. 从 $\mathrm{U}(0,1)$ 抽样得 r_i；
4. 若 $r_i Ch(x_i) \leqslant f(x_i)$，则保留 x_i，否则舍弃；
5. 被接受的 x_i 序列，服从以 $f(x)$ 为概率密度的分布；
6. 使用 x_i 序列的算术平均，估算数学期望

第 4 步是舍选法的关键步骤，但在高维问题中，无法知道 $f(x)$ 的绝对数值，只知道相对值.

定理 9.1.2: Markov 链法

把 $f(x_i)$ 与上一个采样点 $f(x_{i-1})$ 比较，决定是否保留，以概率

$$\min\left[1, \frac{f(x_i)}{f(x_{i-1})}\right]$$

的概率接受 x_i，否则令 $x_i = x_{i-1}$.

9.2　Bootstrap 自助法

Why can not a man lift himself by pulling up on his bootstraps?

方法 9.2.1: bootstrap 自助法

1. 把样本看作总体的估计
2. 从样本抽样
3. 统计量

定义 9.2.1: 经验分布函数

$x_1,\ldots,x_n$ 是来自分布函数为 $\mathrm{F}(x)$ 总体 X 的样本观察值．X 的经验分布函数 $\mathrm{F}_n(x)$ 定义为

$$\mathrm{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}I(x\geqslant x_i). \tag{9.1}$$

用经验分布函数 F_n 估计总体分布函数 F.

从 F_n 抽样可以看作是对样本 $x_1,\ldots,x_n$ 进行可放回的重新抽样，称作 bootstrap 样本，记为 $x_1^*,\ldots,$ 或 $\boldsymbol{x}^*$.

使用一系列 $\boldsymbol{x}^*$，得到相应的 $\hat{\theta}_b^*=\hat{\theta}(\boldsymbol{x}_b^*)$．定义

$$\overline{\theta}^*=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_b^*.$$

为 θ 的 bootstrap 估计.

定理 9.2.1: bootstrap 中心极限定理

当总体 F 换为经验分布 F_n 时，$\mathrm{E}(\hat{\theta})\to\overline{\theta}^*$，$\theta\to\hat{\theta}$

$$\overline{\theta}^*-\hat{\theta}\xrightarrow{\mathrm{P}}\mathrm{E}(\hat{\theta})-\theta. \tag{9.2}$$

收敛速度比正态分布 CLT 预测得更快，称为自助法的二次修正.

计算 $\mathrm{E}_{\mathrm{F}_n}(\hat{\theta})$ 需要考虑所有的 n^n 种重抽样可能性．计算量过大，改良为使用 Monte Carlo 方法近似从 $x_1,\ldots,x_n$ 进行随机重抽样

第三部分

其他部分

第十章　随机过程

随机过程的研究对象是随时间演变的随机现象.

不能用随机变量或多维随机变量来合理表达，而需要用一族 (无限多个) 随机变量来描述.

定义 10.0.1: 随机过程

随机过程 (stochastic process) 是一族随机变量 $\{X(t)\,|\,t\in T\}$，其中 t 是参数，T 称为参数集.

对随机过程进行一次试验 (即在 T 上进行一次全程观测)，其结果是 t 的函数，记作 $\{x(t)\,|\,t\in T\}$，称为随机过程的一个样本函数.

可将随机过程写成

$$\{X(\omega,t)\,|\,\omega\in\Omega,t\in T\}$$

的形式，其中 ω,Ω 分别是随机试验的样本点和样本空间.

统计学描述

定义 10.0.2: 一维分布函数

对随机过程 $\{X(t)\,|\,t\in T\}$，对每个固定的 t

$$\mathrm{F}_X(x,t):=\mathrm{P}\big(X(t)\leqslant x\big),\quad x\in\mathbb{R},\tag{10.1}$$

称为随机过程的一维分布函数，$\{\mathrm{F}_X(x,t)\,|\,t\in T\}$ 称为一维分布函数族.

类似地，可以推广定义 n 维分布函数 (族). 有限维分布函数可以完全确定随机过程的统计特性.

定义 10.0.3: 随机过程的数字特征

均值函数 $\mu_X(t):=\mathrm{E}\big(X(t)\big)$，均方值函数 $\Psi_X^2(t):=\mathrm{E}\big(X^2(t)\big)$
方差函数 $\sigma_X^2(t):=\mathrm{Var}\big(X(t)\big)$，标准差函数 $\sigma_X(t):=\sqrt{\sigma_X^2(t)}$
相关函数 $R_X(s,t):=\mathrm{E}\big(X(s)X(t)\big)$
协方差函数 $C_X(s,t):=\mathrm{Cov}\big(X(s),X(t)\big)$

不难得到

$$\sigma_X^2(t) = R_X(t,t) - \mu_X^2(t),$$
$$C_X(s,t) = R_X(s,t) - \mu_X(s)\mu_X(t).$$

10.1　独立增量过程

定义 10.1.1: 独立增量过程

若 $\forall t_1 < \cdots < t_n (n \geqslant 2, t_i \in T)$，诸增量

$$X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1}),$$

相互独立，则称 $\{X(t) \,|\, t \in T\}$ 是一个独立增量过程.

定理 10.1.1: 独立增量过程的性质

若 $\{X(t) \,|\, t \in T\}$ 是一个独立增量过程，且 $X(0) = 0$，则：

- $X(t)$ 的有限维分布函数族可以由增量 $X(t_2) - X(t_1)(t_2 \geqslant t_1 \geqslant 0)$ 的分布所确定
- 设 $\sigma_X^2(t)$ 已知，则协方差函数

$$C_X(s,t) = \sigma_X^2\big(\min(s,t)\big). \tag{10.2}$$

定义 10.1.2: Poisson 过程

计数过程 $\{N(t) \,|\, t \geqslant 0\}$ 称为强度 $\lambda > 0$ 的时齐 Poisson 过程 (homogeneous Poisson process)，若满足：

1. $N(0) = 0$
2. $\{N(t) \,|\, t \geqslant 0\}$ 是时齐的独立增量过程
3. $\forall t > 0, \Delta t > 0$ 有
   - $\mathrm{P}\big(N(t+\Delta t) - N(t) = 1\big) = \lambda \Delta t + o(\Delta t)$
   - $\mathrm{P}\big(N(t+\Delta t) - N(t) \geqslant 2\big) = o(\Delta t)$

Poisson 过程中任意长度为 t 的区间中事件个数 $\sim \pi(\lambda t)$.

定理 10.1.2: Poisson 过程的性质

$N(t) \sim \pi(\lambda t),\ \mu_N(t) = \lambda t,\ \sigma_N^2(t) = \lambda t,\ R_N(s,t) = \lambda \min(s,t).$

二项分布视角　将区间 $[0,t]$ 分为 k 份，当 $k \to \infty$ 时，$N(t) \sim \mathrm{B}(k,p)$

$$\mu_N(t) = \lim_{k\to\infty} kp = \lim_{k\to\infty} k\left[\frac{\lambda t}{k} + o\left(\frac{t}{k}\right)\right] = \lambda t + o(t).$$

指数分布视角　将 Poisson 过程第 n 个事件发生时刻记为 T_n. 记 $\Delta T_n = T_n - T_{n-1}$ 为第 n 个事件的等待时间，特别地 $\Delta T_1 = T_1$.

定理 10.1.3: Poisson 过程的指数间隔

$\Delta T_1, \Delta T_2, \ldots$ iid $\sim \mathrm{Exp}(\lambda)$

$$\mathrm{P}(\Delta T_1 > t) = \mathrm{P}\big(N(t) = 0\big) = \mathrm{e}^{-\lambda t};$$
$$\mathrm{P}(\Delta T_2 > t | T_1 = s) = \mathrm{P}\big(N(t+s) - N(t) = 0 | \cancel{T_1 = s}\big) = \mathrm{e}^{-\lambda t}.$$

由全概率公式，$\mathrm{P}(\Delta T_2 > t) = \mathrm{E}_{T_1}\big[\mathrm{P}(\Delta T_2 > t | T_1)\big] = \mathrm{e}^{-\lambda t}$.

定理 10.1.4: Poisson 过程判定

给定 iid 随机变量列 $\{\Delta T_i \sim \mathrm{Exp}(\lambda)\}$. 称第 n 个事件在时间 $T_n = \Delta T_1 + \cdots + \Delta T_n$ 发生，得到 $\{N(t) = \max(n) \,|\, t \geqslant T_n\}$ 是强度为 λ 的时齐 Poisson 过程.

定义 10.1.3: Wiener 过程

满足一下条件的随机过程 $\{W(t) \,|\, t \geqslant 0\}$ 称为 Wiener 过程若

- $W(0) = 0$
- $\{W(t) \,|\, t \geqslant 0\}$ 是时齐的独立增量过程
- 增量服从正态分布，即 $\forall t \geqslant s \geqslant 0$

$$W(t) - W(s) \sim \mathrm{N}\big(0, (t-s)\sigma^2\big).$$

若 $\sigma = 1$，则称为标准 Wiener 过程.

Brown 运动是 Wiener 过程.

定理 10.1.5: Wiener 过程的性质

$W(t) \sim \mathrm{N}(0, \sigma^2 t)$, $\mu_W(t) = 0$, $\sigma_W^2(t) = \sigma^2 t, \ldots$

任意 Wiener 过程 $W(t)$，可令 $W(t)/\sigma$ 转化为标准 Wiener 过程.
对 $t_1 < \cdots < t_n$，标准 Wiener 过程的各个增量

$$W(t_1) = w_1 \sim \mathrm{N}(0, t_1),$$
$$W(t_2) - W(t_1) = w_2 - w_1 \sim \mathrm{N}(0, t_2 - t_1),$$
$$\vdots$$
$$W(t_n) - W(t_{n-1}) = w_n - w_{n-1} \sim \mathrm{N}(0, t_n - t_{n-1}).$$

独立，则 $W(t_1), \ldots, W(t_n)$ 的联合 PDF 为

$$f_W(w_1, \ldots, w_n; t_1, \ldots, t_n) = \frac{\exp\left\{-\frac{1}{2}\left[\frac{w_1^2}{t_1} + \frac{(w_2 - w_1)^2}{t_2 - t_1} + \cdots + \frac{(w_n - w_{n-1})^2}{t_n - t_{n-1}}\right]\right\}}{\sqrt{2\pi t_1 (t_2 - t_1) \cdots (t_n - t_{n-1})}}.$$

例 10.1.1: 白噪声

设 $\{W(t)\,|\,t\geqslant 0\}$ 为标准 Wiener 过程，令 f 为区间 $[a,b]$ 有连续导数的函数，定义随机积分

$$\int_a^b f(t)\,\mathrm{d}W(t)=\lim_{n\to\infty}\sum_{i=1}^{n}f(t_{i-1})\big(W(t_i)-W(t_{i-1})\big),$$

其中 $a=t_1<t_2<\cdots<t_n=b$ 是区间 $[a,b]$ 的一个划分．有

$$\int_a^b f(t)\,\mathrm{d}W(t)=f(b)W(b)-f(a)W(a)-\int_a^b W(t)\,\mathrm{d}f(t).$$

$\{\mathrm{d}W(t)\,|\,t\geqslant 0\}$ 称为白噪声，即一个时变函数 f 在白噪声的介质中传播导致输出 $\int f(t)\,\mathrm{d}W(t)$.

10.2　Markov 过程

定义 10.2.1: Markov 过程

随机过程 $\{X(t)\,|\,t\in T\}$ 为 Markov 过程若

$$\begin{aligned}&\mathrm{P}\big(X(t_n)\leqslant x_n|X(t_1)=x_1,\ldots,X(t_{n-1})=x_{n-1}\big)\\=\;&\mathrm{P}\big(X(t_n)\leqslant x_n|X(t_{n-1})=x_{n-1}\big)\end{aligned}$$

时间和状态都离散 Markov 过程称为 Markov 链，记为：

$$\{X_n=X(n)\,|\,n=0,1,2,\ldots\}.$$

定义 10.2.2: 转移概率

条件概率

$$\mathrm{P}(X_{t+n}=j|X_t=i)=:p_{ij}(t,t+n)$$

称为 Markov 链在时间 t 处于状态 i 条件下，在时间 $t+n$ 转移到状态 j 的转移概率.

显然，

$$\sum_{j=1}^{\infty}p_{ij}(t,t+n)=1,\ \ \forall i.$$

n 步转移概率矩阵为

$$P(n)=\begin{bmatrix}p_{11}(t,t+n) & p_{12}(t,t+n) & \cdots\\ p_{21}(t,t+n) & p_{22}(t,t+n) & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}\tag{10.3}$$

每一行元素概率为 1.

定义 10.2.3: 时齐 Markov 链

若 Markov 链是时齐的，即 $p_{ij}(t,t+n)=p_{ij}(n)$ 与 t 无关，则概率转移矩阵为

$$P(n)=\begin{bmatrix} p_{11}(n) & p_{12}(n) & \cdots \\ p_{21}(n) & p_{22}(n) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

一步转移概率矩阵 $P(1)=:P$.

定理 10.2.1: Chapman-Kolomogorov 方程

时齐 Markov 链，有

$$p_{ij}(m+n)=\sum_{k=1}^{\infty}p_{ik}(m)p_{kj}(n),\quad i,j=1,2,\ldots$$

可以简写成矩阵形式

$$P(m+n)=P(m)P(n).$$

自然，$P(n)=P^n$，时齐的 Markov 链的有限维分布由初始分布和一步转移概率完全确定.

定义 10.2.4: 遍历性和极限分布

若时齐 Markov 链的转移概率 $p_{ij}(n)$ 存在极限

$$\pi_j:=\lim_{n\to\infty}p_{ij}(n),$$

则称该链具有遍历性，又若 $\sum_j\pi_j=1$，则称行向量

$$\pi=[\pi_0,\pi_1,\ldots]$$

为该链的极限分布.

定理 10.2.2: 遍历性的充分条件

若存在正整数 m 使得 $\forall i,j\in I$ 有 $p_{ij}(m)>0$，则此链具有遍历性，且极限分布是矩阵方程

$$\pi=\pi P \tag{10.4}$$

满足 $\pi_j>0,\sum_j\pi=1$ 的唯一解.

满足上述条件的 Markov 链的极限分布是该链的平稳分布，即若将 π 作为初始分布，那么每一时刻该链的分布都是 π.

例 10.2.1: 雨伞问题

我有 2 把雨伞用于往返于家和办公室之间．若我从学校出发回家的时候正在下雨，我就会带一把雨伞去学校．从家去学校同理．假定我出发时下雨的概率是 p (不依赖于过去)．

令 $X(t)$ 为我所在地雨伞的数量，状态空间为 $\{0,1,2\}$．

一步转移概率矩阵

$$P=\begin{bmatrix}0 & 0 & 1\\ 0 & 1-p & p\\ 1-p & p & 0\end{bmatrix}$$

$P_{ij}(4)>0$，故具有遍历性．极限分布

$$\left[\frac{1-p}{3-p},\frac{1}{3-p},\frac{1}{3-p}\right]$$

我被淋湿这件事等价于“处于状态 0 并且下雨”，即

$$p\cdot\frac{1-p}{3-p}$$

例 10.2.2: 赌徒 (续)

在例 1.5.2 中，我们用全概率公式解决了赌徒问题，下面我们用 Markov 链解决，我们可以写出一步转移矩阵

$$\begin{bmatrix}1 & 0 & \cdots & & & \\ q & 0 & p & \cdots & & \\ & q & 0 & p & \cdots & \\ & & q & 0 & \cdots & \\ & & & \ddots & \ddots & \\ & & & & \cdots & 0 & p\\ & & & & \cdots & 0 & 1\end{bmatrix}$$

解得 P_i，显然现实中 p 均 $<1/2$，则当 $N\to\infty$ 时，$P_i\equiv 0$，因此若不收手，终将破产．

10.3　平稳随机过程

随机过程作为“随机的函数”太一般了，太难了．如果给它加一些限制，可以使它更有用．迄今有三类限制在随机过程的研究中取得了突破，加深了我们对随机过程的理解

- Markov 过程
- 鞅 (martingales)
  未来给现有已知状态的增量期望为 0．与 stochastic calculus 联系紧密，在博弈论和金融中使用．
- 平稳随机过程 (stationary stochastic process)
  变量一般指时间，当变量为空间时为 homogeneous stochastic process．

定义 10.3.1: 严平稳过程

统计性质不随机时间变化的随机过程，即

$$\mathrm{F}_X(t_1,\dots,t_n)=\mathrm{F}_X(t_1+\tau,\dots,t_n+\tau)$$

称为严平稳随机过程，简称严平稳过程.

一个平稳的 Gauss 过程同时也是严平稳的，因为 Gauss 过程的有限分布函数由一二阶矩完全确定.

定义 10.3.2: 平稳过程

一二阶矩不随机时间变化的随机过程，即

$$\mu_X(t)=\mu_X,\quad R_X(t,t+\tau)=R_X(\tau) \tag{10.5}$$

那么该随机过程称为宽平稳过程 (weak stationary stochastic process)，广义平稳过程或平稳过程.

独立增量过程与平稳过程无关，至少从二阶矩的角度，

$$R_X(t,t+\tau)=\sigma_X^2(t)+\mu_X(t)\mu_X(t+\tau)$$

与 t 相关. 易证，Poisson 过程和 Wiener 过程都不是平稳过程.

但是它们的均值与自方差函数对时间平移的导数都是常数，可以通过微分或差分转化成平稳随机过程，称为广义 (差分) 平稳随机过程.

相关函数　平稳过程的均值函数为常数，因此相关函数 $R_X(\tau)$ 包含了平稳过程的主要信息. 但相关函数 $R_X(\tau)$ 并不能完整刻画随机过程的所有统计学性质.

定理 10.3.1: 平稳随机过程的性质

- $R_X(\tau)$ 是偶函数，且在 0 处取最大值
- 若 $\exists\tau\neq 0,|R_X(\tau)|=R_X(0)$，则 $X(t)$ 是周期的
- 若 $R_X(\tau)$ 在 0 处连续，则处处连续
- $R_X(\tau)$ 非负定，即 $\forall t_i\in T,g(t)$

$$\sum_{i,j=1}^{n}R_X(t_i-t_j)g(t_i)g(t_j)\geqslant 0.$$

任意连续的非负定函数都是某平稳过程的自相关函数.

标准自协方差函数

$$\rho_X(\tau)=\frac{C_X(\tau)}{C_X(0)}.$$

特别地，当 $\mu_X(t)\equiv 0$ 时，

$$\rho_X(\tau)=\frac{R_X(\tau)}{R_X(0)}.$$

例 10.3.1: 随机相位正弦波

随机过程 $X(t)=A\cos(\omega t+\theta)$

对于随机相位正弦波 $\{X(t)\}$，$R_X(0)=\mathrm{E}\big(X^2(t)\big)$ 具有功率的意义.

- 多个随机相位正弦波叠加的信号，总功率为各个分量功率之和.
- 借助 Fourier 分解推广，可定义功率谱函数 $S_X(\omega)$，其加和为 $R_X(0)$，与 $R_X(\tau)$ 互为 Fourier 变换.

遍历性与估计　均值函数

$$\mu_X(t):=\mathrm{E}\big(X(t)\big)$$

定义为随机过程的期望，称为系综平均.

- 如果事件能在系综里假想发生，那么它就一定会在一次足够长的观测中发生. (Landau 统计力学)
- 系综平均 $\mu_X(t)$ 指在假想的多个随机过程的样本轨迹下，t 时刻观测值的平均.
- 对于平稳过程，如果观测足够长时间，是否可以与多个样本轨迹等价？即平稳过程相距较大的观测值 $X(t)$ 与 $X(t+\tau)$ 之间，是否可看作不相关？

定义 10.3.3: 时间均值

记 $x_1:=x(t_1)$，定义时间均值为

$$\hat{m}_n:=\frac{x_1+\cdots+x_n}{n}\to m$$

若

$$\sum_{\tau=0}^{\infty}R_X(\tau)<\infty$$

那么

$$\hat{m}_n\xrightarrow{\mathrm{P}}\mu_X.$$

此时称过程 $X(t)$ 的均值具有各态历经性 (ergodicity) 或遍历性.

- $\hat{m}_n$ 是 μ_X 的无偏相合估计量.
- 估计量的方差

$$\lim_{n\to\infty}n\operatorname{Var}(\hat{m}_n)=\sum_{\tau=-\infty}^{+\infty}R_X(\tau)$$

区间估计：如果相关性较弱，适用 CLT，可使用正态近似给出区间估计.

自相关函数的估计

$$\hat{R}_n(\tau)=\frac{1}{n}\sum_{t=1}^{n-\tau}(x_t-\hat{m})(x_{t+\tau}-\hat{m}).$$

10.4　时间序列

时间序列是离散平稳过程的特例，与滤波理论联系紧密.

定义 10.4.1: 平稳时间序列

若时间序列 $\{X_t\}$ 是平稳过程，称之为平稳时间序列.

$$\mathrm{E}(X_t)=\mu,\quad \mathrm{E}(X_tX_{t+\tau})=f(\tau).$$

定义

$$\gamma_k:=\mathrm{E}\big[(X_t-\mu)(X_{t+k}-\mu)\big]=f(\tau)-\mu^2$$

是平稳过程的协方差函数 $R_X(\tau)$ 的离散版本.

与 $R_X(\tau)$ 性质类似，γ_k 偶函数、非负定、γ_0 最大，定义自相关函数

$$\rho_k:=\frac{\gamma_k}{\gamma_0}.$$

定义 10.4.2: 偏相关函数

用 X_t 的前 k 个时刻的值 $X_{t-1},\dots,X_{t-k}$ 对 X_t 做最小二乘估计，即

$$a_{k1},\dots,a_{kk}=\arg\min \mathrm{E}\left[\left(X_t-\sum_{i=1}^{k}a_{ki}X_{t-i}\right)^2\right]$$

其中 a_{kk} 称作 X_t 的偏相关函数.

$k\geqslant 3$，偏相关函数值变得不显著.

AR(p) 模型

定义 10.4.3: 白噪声序列

时间序列 $\{\epsilon_t\}$ 是白噪声序列若

$$\mathrm{E}(\epsilon_t)=0,\quad \mathrm{E}(\epsilon_s\epsilon_t)=\sigma^2\delta_{st}.$$

$\gamma_k=\sigma^2\delta_k$

一般还可以再假设 $\epsilon_t\sim \mathrm{N}(0,\sigma^2)$ 为 Gauss 白噪声序列.

定义 10.4.4: AR(p) 过程

平稳过程 $\{X_t\}$ 被叫作 p 阶自回归 (auto-regression) 过程，简记为 AR(p) 过程，如果

$$X_t-\varphi_1X_{t-1}-\cdots-\varphi_pX_{t-p}=\epsilon_t. \tag{10.6}$$

白噪声 ϵ_t 是 AR 过程的创新 (innovation).

在 Gauss 过程中，创新也是 Gauss 的．取名 innovation 代表给随机过程带来新的熵.

定义 10.4.5: 生成多项式

记延迟算子 B 为

$$BX_t = X_{t-1}$$

则 AR(p) 过程的条件可记为 $\Phi(B)X_t = \epsilon_t$，其中

$$\Phi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p \tag{10.7}$$

称为 AR(p) 过程的生成多项式.

只有 X_t 与 ϵ_t 相关，即

$$\mathrm{Cov}(\epsilon_t, X_s) = \begin{cases} 0, & t > s \\ \sigma^2, & t = s \\ < \sigma^2, & t < s \end{cases}$$

定理 10.4.1: AR(p) 的性质

- 生成多项式为 $\Phi(\cdot)$ 的 AR(p) 过程的功率谱为

  $$S_X(\omega) = \frac{\sigma^2}{|\Phi(\mathrm{e}^{-\mathrm{i}\omega})|^2}.$$

  因此它的自相关函数 γ_k 作为 $S_X(\omega)$ 的 Fourier 变换，有无穷多非零项，称为拖尾的.
- 由 AR(p) 的定义知，当 $k > p$ 时，它们偏相关函数 $a_{kk} = 0$，称为截尾的.

由于 ϵ_t 是白噪声，即作为误差独立同分布，可以使用多线性回归模型来估计生成多项式中的系数 $\varphi_1, \ldots, \varphi_p$

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \epsilon_t, \ \ \epsilon_t \sim \mathrm{N}(0, \sigma^2).$$

ARMA(p, q)

定义 10.4.6: ARMA(p, q)

随机过程

$$X_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}$$

是有创新 $\{\epsilon_t\}$ 和生成多项式

$$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q \tag{10.8}$$

的滑动平均 (moving-average) 过程 MA(q).

第十一章　测量不确定度

定义 11.0.1: 测量的术语

- 误差 (error) = 测量值 − 参考值
- 不确定度 (uncertainty)：根据所用到的信息，表征赋予被测量量值分散性的非负参数.
- 系统误差 (systematic error)：在重复测量中保持不变或按可预见方式变化的测量误差的分量.
  参考值是 (约定) 真值，
- 随机误差 (random error)：在重复测量中按不可预见方式变化的测量误差的分量.
  系统误差 = 误差 − 随机误差
- A 类评定 (type A evaluation)：对在规定测量条件下测得的量值用统计分析的方法进行测量不确定度分量的评定.
- B 类评定：用不同于测量不确定度 A 类评定的方法对测量不确定度分量的评定.
- 合成标准不确定度：由在一个测量模型中各输入量的标准测量不确定度获得输出量的标准测量不确定度.

不确定度的传递　一个被测量 y 可能是通过对一些输入变量 $x_1, \ldots, x_n$ 的测量而间接得到的. 如果被测量 y 和输入变量之间满足关系式

$$y = f(x_1, \ldots, x_n)$$

则 y 的标准不确定度 $u(y)$ 可以由输入变量的标准不确定度 $u(x_1), \ldots, u(x_n)$ 通过下式计算得到：

$$u^2(y) = \sum_{i=1}^{n} \left(\frac{\partial y}{\partial x_i} \right)^2 u^2(x_i) \tag{11.1}$$

要求 x_i 之间互不相关. 输入变量的标准不确定度可以是 A 类，也可以是 B 类.

若具有相关性，$\rho(x_i, x_j) \neq 0$

$$u^2(y) = \sum_{i,j} \rho(x_i, x_j) \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} u(x_i) u(x_j). \tag{11.2}$$

扩展不确定度　样本均值 $\bar{x}$ 的置信区间

$$\left(\mu \pm t_\alpha(\nu) \frac{s}{\sqrt{n}} \right)$$

α 为置信度，ν 为自由度．$k = t_\alpha(\nu)$ 是置信因子．把一个估计值的标准差乘以 k 便得到这个估计值在特定置信度下 (一般 95%) 的扩展不确定度．

不确定度的不确定度　样本方差更一般的定义

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{n}\epsilon_i^2$$

其中 ν 为自由度，ϵ_i 为残差．

s^2 常用作不确定度的 A 类估计，它自身的不确定度由 χ^2 分布估计

$$\frac{\nu}{\sigma^2}s^2 \sim \chi^2(\nu), \implies u^2(s^2) = \frac{2\sigma^4}{\nu}. \tag{$*$}$$

如果 s 表示测量值的标准不确定度，那么 s 自身的不确定度 $u(s)$ 为

$$u^2(s) = \frac{1}{(2s)^2}u^2(s^2), \implies u(s) = \frac{1}{\sqrt{2\nu}}\frac{\sigma^2}{s} \doteq \frac{s}{\sqrt{2\nu}}.$$

自由度 ν 等于测量值的个数 n 减去用这些测量值所决定的特征量的个数．s 的相对不确定度表示为

$$\frac{u(s)}{s} \sim \frac{1}{\sqrt{2\nu}}.$$

如果 $\nu < 4$，那么 s 的相对不确定度高达 35%；而如果 $\nu > 50$，那么 s 的相对不确定度会降到 $< 10\%$．

定理 11.0.1: Welch-Satterthwaite 公式

设两个互不相关的输入变量 x_1, x_2，且输出变量 $y = f(x_1, x_2)$，则有

$$u^2(y) = c_1^2u^2(x_1) + c_2^2u^2(x_2),$$

y 对应的不确定度的不确定度为

$$u^2\big(u^2(y)\big) = c_1^4u^2\big(u^2(x_1)\big) + c_2^4u^2\big(u^2(x_2)\big),$$

由 χ^2 分布，利用“不确定度的不确定度”和自由度 ν 的关系式 $(*)$

$$\frac{2u^4(y)}{\nu_{\text{eff}}} = \frac{2c_1^4u^4(x_1)}{\nu_1} + \frac{2c_2^4u^4(x_2)}{\nu_2}.$$

ν_{eff} 为 y 的等效自由度 (不必为整数)，则

$$\frac{[c_1^2u^2(x_1) + c_2^2u^2(x_2)]^2}{\nu_{\text{eff}}} = \frac{c_1^4u^4(x_1)}{\nu_1} + \frac{c_2^4u^4(x_2)}{\nu_2}.$$